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

Andrés Gutiérrez[1]

2019-07-17

[1]Experto regional en estadísticas sociales - Unidad de Estadística Social - Comisión Económica para América Latina y el Caribe (CEPAL) - andres.gutierrez@cepal.org

Índice general

Índice de cuadros

Índice de figuras

Resumen

Las encuestas de hogares son un instrumento necesario para realizar seguimiento a un conjunto amplio de indicadores requeridos para el diseño y evaluación de las políticas públicas. Las encuestas de hogares que se implementan en América Latina son de tipo y características diversas. Aunque los conceptos y procesos para su diseño y análisis guardan similitudes, este documento se enfoca principalmente en los procesos referidos a las encuestas de empleo y de propósitos múltiples, con las que los países estiman los principales indicadores relacionados con el mercado laboral, el nivel y distribución de ingresos y la condición de pobreza y las principales características sociodemográficas de la población. Se realiza un recorrido por los diferentes diseños de muestreo, las metodologías más usadas en la selección de las muestras y las estrategias de estimación de los parámetros de interés. También se revisan las técnicas utilizadas para medir el error de muestreo y los métodos disponibles para encarar desafíos como la ausencia de respuesta y la desactualización de los marcos de muestreo.

UNBIS Keywords. Encuestas por muestreo, encuestas de hogares, indicadores socioeconómicos.

Capítulo 1

Introducción

Las encuestas de hogares son un caso particular de investigación social que indaga acerca de características específicas a nivel del individuo, de la vivienda o del hogar 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 y 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. Sin embargo, existe una tendencia creciente de extender las encuestas a constructos más diversos. Es así como cada vez tienen más espacio las encuestas de propósitos múltiples como una fuente relevante de información que permite monitorear indicadores sociales.

En este tipo de encuestas, el hogar es la unidad de análisis, la cual ha sido definida por la División de Estadística de la Organización de las Naciones Unidas (ONU 2011) como:

> a. Un grupo de dos o más personas que se combinan para ocupar la totalidad o parte de una vivienda y para 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 sus ingresos.
> b. 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 una hogar de múltiples personas.

Nótese que la anterior definición refleja la dinámica natural del cambio en las poblaciones de hogares, por lo cual se deben tener distintos enfoques para abordar el problema de la medición de indicadores sociales. No todas las encuestas se diseñan de la misma forma y no por ende 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 del tiempo con el mismo diseño metodológico en donde la selección de hogares se hace de forma independiente para cada aplicación; *encuestas tipo panel*, para las cuales los datos son recolectados en diferentes momentos del tiempo utilizando la misma muestra de hogares en el tiempo; *encuestas rotativas*, en donde un porcentaje de hogares se mantiene en un periodo de tiempo respondiendo la encuesta y en cada aplicación salen algunos hogares que son reemplazados por nuevos hogares. En América Latina, existen una gran variedad de encuestas que abordan diferentes problemáticas sociales. Todas y cada una de ellas han sido diseñadas cuidadosamente para que respondan a las necesidades de la sociedad. Este documento plantea una recopilación de las técnicas usadas tanto en su diseño, como en su análisis.

El diseño de la encuesta dependerá sistemáticamente del objetivo de la medición. Kalton (2009) afirma que es prudente hacer un buena inversión en el desarrollo e implementación de un buen diseño para amortizar los costos de todo estudio. Por lo tanto, lo que se quiere al diseñar una encuesta de hogares es que sea un instrumento confiable, que brinde estimaciones exactas y precisas, puesto que de lo contrario no se podrían monitorear las políticas públicas y los indicadores de interés de forma consistente. Por ejemplo, uno de los indicadores sociales con mayor impacto en la actualidad es la tasa de desocupación, que mide la razón entre la cantidad de personas que se encuentran desocupados, pero que forman parte del mercado de trabajo. Duncan y Kalton (1987) mencionan que las encuestas de hogares pueden proveer estimaciones de los parámetros poblacionales en distintos puntos del tiempo, por ejemplo, la estimación de la tasa de desempleo mensual; proveer estimaciones del cambio neto de los parámetros poblacionales entre periodos de tiempo, por ejemplo, el cambio en la tasa de desempleo entre dos meses; 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 rotativo.

Por supuesto, la medición de los indicadores en el mercado de trabajo es sólo un pequeño componente en el basto universo de posibilidades de medición que brindan las encuestas de hogares. Por esta razón, este tipo de levantamientos se ha convertido en una herramienta fundamental para medir indicadores sociales en todo el mundo y que, en particular, permiten que las naciones de América Latina puedan hacer seguimiento a su desarrollo económico y social. Sin embargo, este tipo de instrumentos puede ser utilizado como herramienta para monitorear el progreso de los países en términos de metas y objetivos comunes. Es así como en 2015, la Asamblea General de la Organización de las Naciones Unidas aprobó una resolución que plantea un plan de acción en favor de las personas, el planeta y la prosperidad (ONU 2015). Esa resolución 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 a los ODS es posible utilizar diferentes fuentes de información, como censos, registros administrativos, registros estadísticos, proyecciones demográfi-

cas y también las encuestas de hogares (ONU 2016). En particular, el seguimiento a los ODS se realiza a través de indicadores, muchos de los cuales no pudieran ser estimados de no ser por la información disponible en las encuestas de hogares. Por ejemplo, el objetivo 8 busca *promover el crecimiento económico sostenido, inclusivo y sostenible, el empleo pleno y productivo y el trabajo decente para todos*. Claramente de este objetivo se desprenden indicadores que permiten conocer la evolución del país en la consecución de las metas.

Desde otra perspectiva, en el marco de la decimotercera conferencia internacional de estadísticos del trabajo en 1982, la Organización Internacional del Trabajo (OIT) adoptó algunas directrices concernientes con 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, la resolución afirma que las encuestas de hogares deberían:

1. Brindar datos de la población económicamente activa.
2. Proveer estadísticas básicas de sus actividades durante el año, así como las relaciones entre el empleo, ingreso y otras características económicas y sociales.
3. Proveer datos sobre otros temas particulares para responder a las necesidades a largo plazo y de índole permanente.

En el año 2013, la OIT decidió revisar esta resolución y propuso algunos cambios en el marco de la 19 Conferencia Internacional de Estadísticos del Trabajo en donde se acogieron algunas modificaciones en términos de 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, métodos de recopilación de datos, entre otras (OIT 2013). Es así como los Institutos Nacionales de Estadística (INE) de América Latina no sólo planean las encuestas de hogares de tal forma que puedan responder a los nuevos retos en términos de la estimación de los parámetros de interés en cuento al trabajo remunerado o no remunerado para mantener la comparabilidad de las estadísticas laborales entre los países proporcionando nuevos y mejores indicadores para contribuir al análisis de la dinámica del mercado laboral, sino que se actualizan paulatinamente para poder brindar la información que la sociedad necesita a medida que evoluciona el constructo social de interés.

Es importante resaltar que los indicadores de bienestar (en términos de ingresos y gastos) también hacen parte del conjunto de parámetros que se pueden estimar desde las encuestas de hogares. Medir el ingreso a partir de las encuestas de hogares se constituye en un reto metodológico para los institutos nacionales de estadística en el mundo, y particularmente en América Latina. Es recomendable seguir las directrices de la Comisión Económica para Europa que revisten una actualización de los estándares internacio-

nales, recomendaciones y buenas prácticas en la medición del ingreso en los hogares. Por ejemplo, el llamado Grupo de Canberra ha revisado exhaustivamente el tópico de la estimación del ingreso estudiando 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 en el hogar (ONU 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 el 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 reviste retos metodológicos mayores puesto que los entrevistados deben responder con precisión cuando se les indague por este constructo que contiene los ingresos personales de cada individuo en el hogar, como sueldos y salarios, ganancias, ingresos del empleo, pensiones, etc. y también los ingresos del hogar, incluidas las rentas por alquiler y los ingresos generados por el comercio. Por lo tanto, el diseño de la encuesta 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 un 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 información de interés, como los rubros de seguridad social hechos por su empleador. Por otro lado, si se requiere que el respondiente brinde información acerca de un determinado periodo de tiempo, el planteamiento de la pregunta, la forma de indagar y el entrenamiento de los encuestadores pueden sesgar sistemáticamente la respuesta y por consiguiente inducir estimaciones poco confiables. Mucho se ha investigado al respecto de cómo realizar preguntas certeras en este tipo de levantamientos y el lector interesado puede consultar los trabajos de Biemer y Lyberg (2003), Presser y col. (2004), y Groves y col. (2009).

Este documento pretende revisar algunas de las metodologías más usadas por los INE 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 involucrados en los proceso técnicos de este tipo de encuestas. De la misma forma, este documento considera 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 en campo y parten la realización de la encuesta en dos. Los lectores que están familiarizados con la investigación social a través de las encuestas de hogares encontrarán que las encuestas se planean teniendo en cuenta muchos pormenores que podrían suceder en campo, pero que este diseño en la mayoría de ocasiones toma distancia de la información que se recolecta en campo. Es por esto que el trabajo de las encuestas asciende cuando se logra plasmar la información en forma de base de datos. En este segundo momento es cuando se debe asegurar que lo que se planificó efectivamente sea incorporado en el

análisis de esta información. Desde esta perspectiva, este documento puede verse en dos partes: los capítulos 2, 3 y 4 abordan el diseño, mientras que los capítulos 4, 5 y 6 y 7 abordan el análisis. Esta distinción así como los procesos que la componen se presenta en la Figura 1.

En el primer capítulo se considera una breve introducción a la problemática de las encuestas de hogares. En el capítulo dos se aborda con más detalle los elementos básicos que se consideran por lo regular en los diseños de las encuestas de hogares. Un aspecto relevante de este documento es que, si bien considera que las encuestas de hogares tienen muchos elementos en común, diferencia de forma cuidadosa las particularidades de cada encuesta. Por ejemplo, en este capítulo 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, describe las características metodológicas que se deben considerar al momento de diseñar la encuesta y revisa los conceptos esenciales que determinarán el tipo de aplicación que se debe considerar. El capítulo tres describe los principales diseños de muestreo que se utilizan en este tipo de estudios y expone de forma estándar los conceptos de estratificación y aglomeración de las poblaciones. El capítulo cuatro complementa estos conceptos con varias aplicaciones prácticas para determinar el tamaño de muestra adecuado para lograr los objetivos de la investigación. 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.

Pasando a la parte del análisis de las encuestas, el capítulo cinco revisa los procesos de imputación y ponderación en la encuesta. Los procesos de imputación tratan de recuperación tanta información como sea posible para que el investigador pueda contar con una base de datos rectangular y completa. Luego de esto, es necesario aplicar los factores de expansión a la información contenida en la base de datos para que se puedan realizar inferencias a nivel nacional o regional. Sin embargo, en aquellos casos en donde 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 una muestra representativa de toda la población. El capítulo seis analiza las principales metodologías de estimación, tanto de los parámetros de interés como de sus errores de muestreo. 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 están basadas en el diseño de muestreo y no en supuestos metodológicos ligados a algún modelo estocástico. Es por esto que 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. El capítulo siete presenta de forma detallada los procesos que se surten cuando se agregan encuestas a lo largo de un periodo de tiempo. Acudiendo a la perspectiva del autor, el capítulo ocho presenta los criterios de calidad que se deberían tener en cuenta

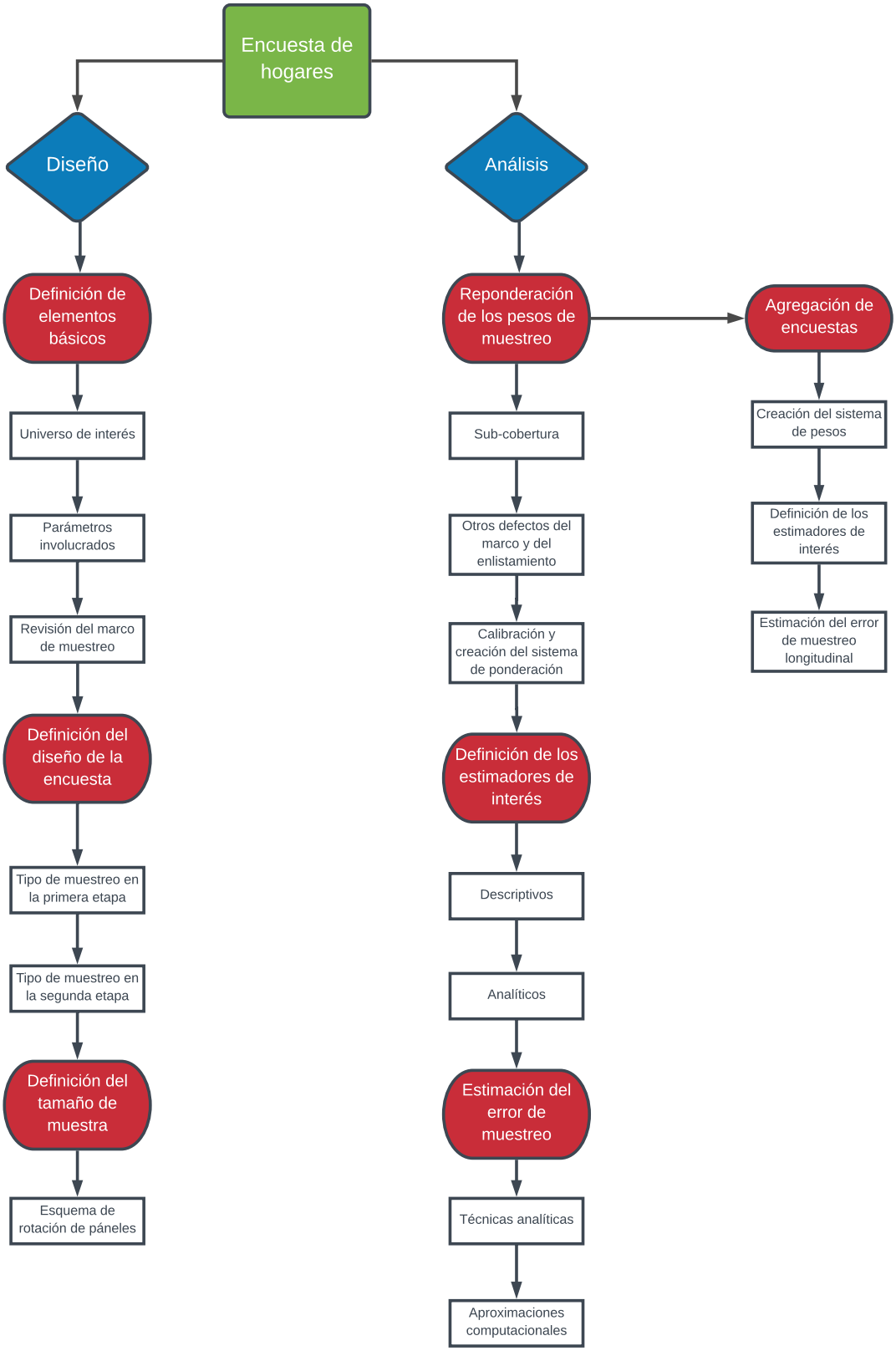

Figura 1.1: Esquema de procesos en el análisis y diseño de una encuesta de hogares.

para decidir si una cifra, resultante de un proceso de estimación estadística basada en encuestas de hogares, debería ser o no publicada a la sociedad. Por úlitmo, el capítulo nueve presenta una discusión acerca del uso presente de las encuestas de hogares y los retos que depara el futuro en materia de la medición de indicadores sociales a través de las encuestas de hogares. Asimismo, en los anexos se contempla una revisión del software que se utiliza actualmente en los INE para llevar a cabo esta ardua tarea de diseñar y analizar las encuestas de hogares, una revisión rápida de algunas de las encuestas de la región, así como algunas directrices que se deberían considerar al momento de documentar los procesos asociados a las encuestas de hogares.

Capítulo 2

Elementos básicos y planeación

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

I Universo, muestra y unidades

El término encuesta se encuentra directamente relacionado con una población finita compuesta de individuos a los cuales es necesario observar y medir. Este proceso muchas veces es realizado por medio de una entrevista. 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* de Ecuador define su población objetivo como todas las personas mayores de 10 años residentes en viviendas particulares en Ecuador.

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, la *Encuesta Nacional de Ingresos y Gastos de los Hogares* define como unidades de análisis el ámbito al que pertenece la vivienda: urbano alto, complemento urbano y rural. Por otro lado, la *Gran Encuesta Integrada de Hogres* de Colombia tiene cobertura nacional y sus unidades de análisis están definidas por 13 grandes ciudades junto con sus áreas metropolitanas.

Como se explicará más adelante, es muy difícil contar con una lista actualizada de todos

los hogares del país; por lo tanto, para recolectar la información de la población objetivo, el diseño de una encuesta de hogares en América Latina plantea la necesidad de seleccionar en varias etapas ciertas *unidades de muestreo* que sirven como medio para seleccionar finalmente a los hogares y personas que participarán de la muestra. Cuando se requiere seleccionar personas, se hace necesario seleccionar un subconjunto de zonas geográficas; para cada zona seleccionada, se procede a seleccionar a su vez un subconjunto de secciones cartográficas, que antecede a la selección de hogares y, finalmente, para cada hogar se seleccionan las personas; siendo estas las unidades de observación. Por ejemplo, se puede citar la experiencia de Brasil con la *Pesquisa Nacional por Amostra de Domicilios* que se realiza por medio de una muestra de viviendas en tres etapas: las unidades primarias de muestreo (UPM) son los municipios, mientras que las unidades secundarias de muestreo (USM) son los sectores censales, que conforman una malla territorial conformada en el último Censo Demográfico. Por último, las unidades finales en ser seleccionadas son las viviendas.

Duncan y Kalton (1987, pág. 105) afirman que la composición de la población de interés en las encuestas de hogares cambia durante el tiempo, puesto que lo individuos nacen, mueren, migran, e incluso pasan a ser parte de organizaciones que hacen que pierdan su estatus de unidad de observación. Nótese que la población objetivo de la mayoría de encuestas de hogares en América Latina se refiere a la población civil excluyendo miembros de organizaciones militares, personas en cárceles, hospitales, etc. De igual forma, se debe tener en cuenta que existen nuevos hogares que pueden crearse o desintegrarse. Por ende, los equipos técnicos de los INE 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 en el tiempo, sí lo hace la población objetivo y se deben plantear esquemas de seguimiento y actualización que den cuenta de esta realidad.

II Periodicidad en el tiempo

Los institutos nacionales de estadística - que son los entes encargados de administrar, diseñar, analizar y difundir los resultados de las encuestas - no realizan este tipo de levantamientos de manera aislada; de hecho una característica fundamental de estas operaciones estadísticas 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 en el tiempo. Es por esto que la planeación de la encuesta debe contemplar esta periodicidad para que el levantamiento de la información primaria en campo se haga de manera más eficiente y, de la misma forma, que la estimación de estos indicadores se pueda realizar ajustándose a los recursos de la operación. Como se mencionó anteriormente, dado que la población es dinámica en el tiempo, la planeación y análisis de este tipo de encuestas es desafiante, puesto que si la composición de la población y las características de los elementos se

considerara fija, una encuesta transversal (realizada una sola vez en un periodo de tiempo largo) sería suficiente para realizar estimaciones precisas que resuelvan los objetivos del estudio.

En algunas ocasiones, basta con realizar un medición simple en un punto específico del tiempo para completar los objetivos de la investigación. Este es el caso de las encuestas de ingresos y gastos cuya periodicidad es en promedio es cada 10 años y las cuales son utilizadas para, entre muchos otros propósitos, actualizar la canasta básica familiar, de la cual se derivan los insumos básicos para la medición de la pobreza (CEPAL 2018a). Para otro tipo de problemáticas, como por ejemplo el seguimiento al mercado de trabajo, es necesario recurrir a la medición periódica a través de encuestas de hogares, en donde los cambios naturales en las características de la población hacen que realizar una medición simple en un punto del tiempo sea inviable a la luz del seguimiento y monitoreo de los indicadores de interés.

Por consiguiente, al momento de realizar la planeación de una encuesta continua o periódica se debe tener en cuenta que, a pesar de que crezca la dificultad en el diseño, es posible 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 la constituyen individuos, hogares, familias, negocios, etc., las unidades de observación deben ser consideradas como parte de la población de interés cuando nacen, inmigran o alcanzan un umbral predefinido de edad. Asimismo, las unidades salen de la población de interés cuando mueren, emigran, o se involucran en instituciones (como el servicio militar). Por ejemplo, si las unidades son los hogares, es evidente que la población no es la misma en diferentes puntos del tiempo (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 donde un hogar se divide en dos. Además, los hogares en donde todos sus miembros han fallecido dejan de ser parte de la población objetivo. De la misma forma, dos hogares dejan de ser parte de la población objetivo cuando se unen a través de un matrimonio o algún tipo de unión civil. Teniendo en cuenta el papel dinámico de las poblaciones y los objetivos de investigación es posible plantear diferentes tipos de levantamientos; a continuación enumeramos algunas categorías de encuestas que las oficinas nacionales de estadística realizan en la región.

A Encuestas transversales

Este tipo de encuestas son diseñadas para recolectar información únicamente en un punto específico del tiempo, o sobre un periodo de referencia, y proveen toda la información pertinente acerca de la población particular restringida a un tiempo y periodo de recolección específico. Puesto que el propósito fundamental de este tipo de encuestas no se centra en las comparaciones intertemporales, no es posible estimar cambios de ningún tipo, a no ser que se realicen indagaciones retrospectivas. El siguiente cuadro

muestra un esquema de este tipo de operaciones estadísticas en donde se observa una muestra de una población específica en un periodo de tiempo específico (Tiempo 2). Dado que es una muestra transversal, no hay un patrón de repetición en los restantes periodos.

Muestra	Tiempo 1	Tiempo 2	Tiempo 3	Tiempo 4	Tiempo 5	Tiempo 6
1		**x**				
2		**x**				
3		**x**				
4		**x**				
5		**x**				
6		**x**				

B Encuestas repetidas

Cuando existe interés en realizar un seguimiento del fenómeno en observación durante el tiempo, entonces se utilizan encuestas repetidas que recolectan información de manera periódica. Este tipo de encuestas proveen información acerca de la dinámica de la composición de la población en el tiempo. De esta forma, en cada levantamiento se observa una muestra de la población en un tiempo determinado. Por ejemplo, el siguiente cuadro muestra un acercamiento gráfico a este tipo de encuestas en donde se evidencia el carácter sistemático de estas operaciones estadísticas; además de mostrar que no es posible medir cambios individuales porque las muestras son independientes en el tiempo.

Muestra	Tiempo 1	Tiempo 2	Tiempo 3	Tiempo 4	Tiempo 5	Tiempo 6
1	**x**					
2		**x**				
3			**x**			
4				**x**		
5					**x**	
6						**x**

C Encuestas panel

Las encuestas en panel están diseñadas para recolectar información periódica sobre la misma muestra en diferentes puntos del tiempo. Por definición, las unidades de muestreo son las mismas en los diferentes periodos de tiempo y, de manera general, se miden las mismas variables en cada levantamiento. Por la caracterización propia de este tipo

de encuestas, sí es posible medir los cambios individuales y cambios netos sobre la población de la que se seleccionó la muestra. Sin embargo, como la muestra no cambia en ningún momento del tiempo, las inferencias que se realicen estarán supeditadas a la población de la cual se seleccionó la muestra en un principio. Si la población cambia su estructura, no será posible captar este cambio puesto que las inferencias resultantes de este tipo de encuestas no son representativas de la población actual. El siguiente cuadro muestra un esquema propio de las encuestas de panel en donde los individuos que fueron seleccionados la primera vez son observados a lo largo del tiempo.

Muestra	Tiempo 1	Tiempo 2	Tiempo 3	Tiempo 4	Tiempo 5	Tiempo 6
1	**x**	**x**	**x**	**x**	**x**	**x**
2	**x**	**x**	**x**	**x**	**x**	**x**
3	**x**	**x**	**x**	**x**	**x**	**x**
4						
5						
6						

D Encuestas de panel dividido

Para hacerle 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 crearon 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 inferencias precisas acerca de los cambios de una cohorte a través del tiempo y, al mismo tiempo, del cambio en estructura de la población actual. De esta forma, se realiza el seguimiento continuo, periódico y sistemático de una muestra a través del tiempo, pero en cada levantamiento se incluyen nuevos elementos seleccionados de la población actual. Como se señalará más adelante, este tipo de encuestas cubre con eficiencia la mayoría de indicadores de interés en un estudio de investigación social. El siguiente cuadro muestra una caracterización de estos levantamientos que fijan una muestra de panel a lo largo del tiempo, a la vez que se añaden nuevas observaciones.

Muestra	Tiempo 1	Tiempo 2	Tiempo 3	Tiempo 4	Tiempo 5	Tiempo 6
1	**x**	**x**	**x**	**x**	**x**	**x**
2	**x**					
3		**x**				
4			**x**			
5				**x**		
6					**x**	
7						**x**

E Encuestas de panel rotativo

Mantener una muestra de panel es un proceso costoso desde una perspectiva económica, pero también desde el desgaste de la fuente. Es evidente que a medida que el tiempo transcurra la propensión a responder será más baja, puesto que el entrevistado se sentirá agotado al ser visitado una y otra vez. Por lo tanto, se definen las encuestas de panel rotativo para poder realizar inferencias parciales - restringidas a periodos 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 a la vez mantienen elementos comunes con mediciones anteriores. Obviando las dificultades que acarrea la ausencia de respuesta, las encuestas panel definen 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 del desgaste del panel (sobre la población inicial) y el efecto de la pérdida de muestra. Además, la inclusión de nuevos elementos en la muestra provee información pertinente del cambio en la composición estructural de la población. El siguiente cuadro ejemplifica el diseño de las encuestas rotativas.

Muestra	Tiempo 1	Tiempo 2	Tiempo 3	Tiempo 4	Tiempo 5	Tiempo 6
1	**x**					
2	**x**	**x**				
3	**x**	**x**	**x**			
4		**x**	**x**	**x**		
5			**x**	**x**	**x**	
6				**x**	**x**	**x**
8					**x**	**x**
9						**x**

III Rotación de páneles

Algunas encuestas de hogares en América Latina permiten que un hogar sea visitado en más de una ocasión con el fin de tener estimaciones precisas acerca de los cambios de estado que el hogar o las personas que lo habitan puedan sufrir. Por ejemplo, un hogar que en un periodo estuvo en condición de pobreza extrema, puede estar en otro periodo en condición de pobreza o inclusive puede pasar a estar fuera de la pobreza; en las encuestas de fuerza laboral, una persona puede pasar de estar empleada en un periodo a desempleada en otro periodo. Estos cambios y la dinámica propia que conllevan son de interés para los investigadores y deben ser contemplados desde una perspectiva más amplia en cuanto a su diseño. Nótese que este tipo de variaciones sobre los individuos necesariamente tiene que ser captada a través de 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 (ver anexo). Por ejemplo, la *Encuesta Permanente de Hogares* en Argentina renueva periódicamente el conjunto de hogares que serán entrevistados mediante un esquema[1] de rotación 2(2)2 que selecciona a las viviendas para ser entrevistadas en dos periodos consecutivos, luego los siguientes dos periodos esas viviendas salen de la selección, para finalmente volver a ser encuestadas en los siguiente dos periodos. De esta forma, un hogar es seguido a lo largo de 18 meses y esto permite cumplir con los objetivos de la encuesta. Este esquema induce algunas propiedades interesantes, que pueden ser ejemplificadas usando el siguiente cuadro definido para los cuatro trimestres de los años 2016, 2017, 2018 en cuatro grupos de muestra A, B, C y D.

- Entre el primer y el segundo periodo de medición hay un traslape del 50% de hogares. En particular, nótese que entre 2016-T1 y 2016-T2, la muestra se conserva en un 50%, puesto que $a1$ y $d1$ se repiten. Esto mismo sucede en cada trimestre del esquema rotacional.
- En el tercer periodo no habrá traslape con el primer periodo. Nótese que 2016-T1 y 2016-T3 no existe ningún elemento en común. De la misma manera, entre 2016-T2 y 2016-T4, no existe ningún elemento en común. Este mismo patrón se encuentra a lo largo del esquema rotacional.
- En el cuarto periodo se tendrá un 25% de traslape con el primer periodo. Nótese, por ejemplo, que entre 2017-T1 y 2017-T4, $d3$ se repite; de la misma manera, entre 2017-T4 y 2018-T3, $d4$ se repite.
- Finalmente en el quinto periodo se volverá a tener un 50% de traslape con respecto al primer periodo. Por ejemplo, 2016-T1 y 2017-T1 comparten el 50% de la muestra $a1$ y $b1$; asimismo, 2017-T1 y 2018-T1 comparten el 50% de la muestra $c3$ y $b3$.

Cuadro 2.6: *Rotación de páneles 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*

[1]Un esquema de rotación $x(y)z$, se define como aquel en donde la vivienda entra al panel por x meses, se excluye por los siguientes y meses y este patrón se repite z veces en el tiempo.

Año	Trimestre	A	B	C	D
	T2	*a3*	*b4*	*c4*	*d3*
	T3	*a4*	*b4*	*c4*	*d4*
	T4	*a4*	*b3*	*c5*	*d4*

Otro ejemplo de una encuesta que utilizada rotación de paneles es la *Encuesta Continua de Empleo* que, aplicada por el Instituto Nacional de Estadística de Bolivia, hace uso de una metodología mixta que permite el seguimiento continuo y transversal a la tasa de desempleo y a la tasa de subocupación, así como el seguimiento a los cambios que se presentan entre los periodos de interés (trimestres y semestres), a través del análisis longitudinal de los datos en el sector urbano (pues el diseño no es rotativo en el sector rural, debido a la baja incidencia de desempleo en esta zona). En este esquema rotacional 4(0)1 una vivienda es entrevistada durante cuatro trimestres consecutivos, y luego sale del panel definitivamente. Un ejemplo de este tipo de esquemas se presenta en el siguiente cuadro. - Nótese que entre el primer y el segundo periodo de medición hay un traslape del 75% de hogares. En particular, entre 2016-T1 y 2016-T2, la muestra se conserva en tres cuartas partes puesto que $a1$, $c1$ y $d1$ se repiten. Esto mismo sucede en cada trimestre del esquema rotacional. - Por otro lado, entre el primer y el tercer periodo habrá un traslape del 50%. Nótese que entre 2016-T1 y 2016-T3, la mitad de la muestra se conserva puesto que $a1$ y $c1$ se repiten. Este mismo patrón se encuentra a lo largo del esquema rotacional. - Entre el primer y el cuarto periodo se tendrá un 25% de traslape. Nótese, por ejemplo, que entre 2017-T1 y 2017-T4, $a2$ se repite; de la misma manera, entre 2017-T4 y 2018-T3, $d3$ se repite. - Finalmente entre el primer y quinto periodo no se tiene ningún tipo de traslape.

Cuadro 2.7: *Rotación de páneles 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*

Los diseños de las encuestas de hogares deben tener en cuenta la rotación de los paneles y el número de veces que es visitado un hogar. Esta caracterización depende directamente de los indicadores a los cuales la encuesta debe responder. Por ejemplo, el diseño de rotación debe ser diferente si el interés se centra en indicadores de cambio trimestral, a si se requieren indicadores de cambio anual. Por ejemplo, si el objetivo está en comparar las estimaciones de la tasa de desocupación el mismo mes entre diferentes años, el diseño 4(0)1 no es conveniente puesto que el traslape anual es nulo. En cualquier caso debe existir un esquema longitudinal, pero la diferencia principal radica en el tiempo en el que un hogar pertenecerá al panel. Por supuesto, hay que tener en cuenta que la tasa de ausencia de respuesta y pérdida de muestra por desgaste crecerá en la medida en que se le pida a un hogar una participación más duradera en el tiempo.

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 se centra en la estimación del cambio del indicador en dos periodos de tiempo, entonces el cálculo de la precisión de las estimaciones debe tener en cuenta que las muestras no son independientes y por lo tanto se debe 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 deben intervenir en el cálculo de los coeficientes de variación, así como en la determinación del tamaño de muestra en cada ronda. En efecto, como lo afirma McLaren y D. G. 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 donde la vivienda entra en un primer mes en el panel, se excluye por los siguientes dos meses y este patrón se repite m veces consecutivas. A partir de allí la vivienda ya no vuelve a ser incluida en el estudio. En resumen, por la naturaleza de las encuestas de hogares en la región, al momento de pensar en incluir o cambiar la estructura rotacional en el sistema de encuestas de hogares, se debería considerar en primer lugar el esquema de repartición de paneles mensual. Una mirada más profunda de este tipo de análisis longitudinales se encuestar presente en el último capítulo de este documento.

IV Parámetros de interés

Las encuestas son usadas para producir estimaciones de parámetros que describen la situación de una población, respondiendo a los objetivos de la investigación. Por lo general, y como se describirá más adelante, el conocimiento de la población a cualquier nivel está reflejado en forma de totales, o de funciones de totales. Es por esta razón que este documento se enfoca y profundiza 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, notada como y, en la población. Se calcula mediante la siguiente ecuación:

$$t_y = \sum_U y_k$$

En donde U hace referencia al universo de estudio e 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 de 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 notada como y_k.

Un caso particular de este parámetro es el **tamaño poblacional** que designa la cantidad de unidades que conforman una población y se denota como N. Por lo general, este parámetro es regularmente conocido, o al menos se tiene una aproximación de esta cantidad. En una encuesta de hogares, este parámetro podría denotar el número de hogares en el país - el cual no es conocido literalmente, aunque sí se conocen aproximaciones (o proyecciones) a esta cantidad con base en los resultados de los censos de población y vivienda - o el número de habitantes del país - el cual tampoco es conocido exactamente, aunque sí se cuente con proyecciones poblacionales.

Tal vez el parámetro más relevante en la investigación social lo constituye el **promedio poblacional** que describe la cantidad que debería ser asignada 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 de la población, definido como el total de los ingresos de todos los hogares del país dividido entre el número de habitantes del país. En este caso la variable de interés y es el ingreso per cápita. De la misma forma, también se podría estimar el gasto promedio de los hogares en educación; en donde la variable de interés y es el gasto de todos lo miembros del hogar en este concepto (sin importar la edad ni el nivel propedéutico) y N sería el número de hogares del país.

La incidencia de los fenómenos sociales sobre los hogares o personas puede ser medida a través de la **proporción poblacional**, que es un parámetro definido como un promedio sobre una variable dicotómica z_k que toma el valor de 1 si el k-ésimo individuo tiene el atributo de interés y de 0 en otro caso. Se calcula mediante la siguiente ecuación:

$$P_U = \frac{\sum_U z_k}{N}$$

Por ejemplo, la proporción de personas en condición de pobreza es una proporción sobre toda la población, en donde la variable de interés z indica si el ingreso per cápita de un individuo es menor que la línea de pobreza; CEPAL (2018a) presenta los pormenores metodológicos del cálculo de la pobreza en los países de América Latina y el Caribe. Por otro lado, la **razón poblacional** se calcula como el cociente entre dos totales, el primer total asociado a una variable de interés y, el segundo total asociado a una variable de interés z. Este parámetro se calcula mediante la siguiente expresión:

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

en donde t_y es el total poblacional asociado a la variable y, t_z es el total asociado a la variable z. 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. Nótese que para clasificar a una persona como desocupada, ocupada, activa o inactiva, es necesario realizar una indagación en la encuesta a cada uno de los miembros del hogar; por lo tanto ambas cantidades, numerador y denominador, corresponden a cantidades desconocidas de antemano. Es más, la condición de ocupación de las personas puede variar entre los periodos de observación. Con respecto a indicadores de pobreza, otro tipo de parámetros pueden expresarse como razones poblacionales; es el caso de la brecha de pobreza y de la incidencia de la pobreza expresada en términos de un umbral de poder adquisitivo (Foster, Greer y Thorbecke 1984). Este tipo de indicadores complejos se pueden expresar mediante la siguiente relación

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

En donde y_k determina el ingreso del individuo k, u se refiere al umbral que establece la línea de pobreza y α pertenece al intervalo $(0, 1)$. Por ejemplo, en el caso en el que $\alpha = 0$, este indicador calcula la tasa de pobreza, que es la incidencia de este fenómeno en la población; si $\alpha = 1$, este indicador calcula la brecha de la pobreza, que es la cantidad de dinero que se necesitaría en promedio para que un país no tuviera personas en situación de pobreza. En este punto vale la pena resaltar que en la definición de los parámetros básicos que se quieren estimar en una encuesta, el papel de los totales poblacionales es absolutamente relevante. De igual manera, existen otros parámetros que pueden ser considerados complejos - no por su forma funcional, sino por los procesos complejos que hay detrás del levantamiento de la información primaria - pero que al igual que los mencionados anteriormente resultan ser también una función de totales poblacionales. Por ejemplo, considere el **cambio neto** de los totales de la variable de interés y en dos periodos de tiempo (t_1 y t_2) dado por la siguiente expresión:

$$\Delta_y = t_y^2 - t_y^1$$

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, la siguiente tabla muestra la composición del mercado de trabajo en una población observada en dos periodos de interés (las cifras están en millones). De esta forma, los totales marginales de la tabla corresponden a los **cambios netos** que permiten una comparación simple con el periodo anterior. Específicamente, es posible observar que hay 313 mil empleados menos, 80 mil desempleados menos y 393 mil inactivos más en el segundo periodo, en comparación al primero.

Cuadro 2.8: *Composición del mercado de trabajo en dos periodos de tiempo*

Condición	Ocupado	Desocupado	Inactivo	Total
Ocupado	9222	128	662	10012
Desocupado	221	322	151	694
Inactivo	256	164	5941	6361
Total	9699	614	6754	17067

Una comparación más profunda está dada en términos de los **cambios brutos**, que corresponden a las entradas de la tabla cruzada. De esta manera, los cambios en la fuerza de trabajo de un periodo a otro, se explican porque el 92.1%=(9222/10012)% de los empleados conservó su empleo; el 31.8% de los desempleados y el 4.0% de los inactivos consiguió un nuevo empleo; el 6.6% de los empleados es ahora inactivo en la fuerza laboral y el 1.3% de los empleados perdió su empleo. Así mismo, el 46.4% de los desempleados conservó su clasificación; el 1.3% de los empleados perdió su empleo y el 2.6% de los inactivos entró a la fuerza laboral como desempleado; el 31.8% de los desempleados es ahora empleado y el 21.8% de los desempleados es ahora inactivo.

A Ejemplos de indicadores de interés y su relación con los tipos de encuestas

En esta sección se relacionan algunos de los parámetros anteriormente mencionados con los tipos más comunes de encuestas. Estos ejemplos nos presentan algunas indicaciones del tipo de encuestas que se encuentran en América Latina y examinan el raciocinio detrás de estos levantamientos. Tomando en consideración las características generales de las encuesta de hogares, Duncan y Kalton (1987) mencionan las siguientes situaciones, ejemplificadas a continuación.

- **Estimación de parámetros poblacionales en un punto del tiempo**. Por ejemplo, suponga que se quiere estimar el *ingreso per cápita promedio por área (rural - urbano) en las regiones de un país*. En este tipo de estudios, la encuestas

aptas serían las transversales, las repetidas, las de panel rotativo y las de panel dividido. Nótese que las encuestas de panel puro no son aptas para captar 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 la cual 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 segundo trimestre de 2018 y el primer trimestre de 2018 en un país*, entonces las encuestas aptas serían las repetidas, las de panel rotativo y las de panel dividido. Una encuesta transversal no sería apta para lograr esta estimación, puesto que su frecuencia de realización no es trimestral. De la misma forma que en el parámetro anterior, las encuestas de panel puro no son aptas 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 2018 que estuvieron desocupadas en el primer trimestre de 2018 en un país* es necesario que la encuesta tenga algún patrón de selección de los mismos individuos en los dos periodos. De esta forma, las únicas encuestas aptas para estimar este cambio bruto son las de panel, panel rotativo y panel dividido. Las encuestas transversales o repetidas no podrían arrojar este tipo de estimativas puesto que su diseño no considera a los mismos individuos en la muestra en dos periodos de tiempo.

- **Estimación de la incidencia de eventos en un periodo de tiempo**. Suponga que se quiere estimar la *proporción de mujeres que fueron víctimas de un evento de violencia en los úlitmos seis meses en un país*. En este caso todas las encuestas resultarían aptas mediante ligeras modificaciones en el diseño. Por ejemplo, la encuesta transversal debería preguntar de forma retrospectiva; las encuestas repetidas podrían ser agregadas en los últimos seis meses, las encuestas de tipo panel rotativo y divididas deberían preguntar en cada medición de los últimos seis meses por este evento.

- **Estimación de la incidencia de eventos raros en el 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 aptas. En el primer caso, dado que el evento es raro por definición, los requerimientos de tamaño de muestra en una encuesta transversal sobrepasarían el presupuesto y los costos de la encuesta; en el segundo caso, además de las consideraciones anteriormente planteadas del tamaño de muestra, por la misma definición de evento raro, tampoco sería plausible que en el panel se presentaran estos eventos en los individuos a través del tiempo. Por otro lado, al agregar las encuestas repetidas, las de panel rotativas y la parte nueva del panel dividido, sería posible llegar al tamaño de muestra adecuado para poder captar esta incidencia de forma precisa y eficiente.

Estos últimos ejemplos muestran la importancia de contar con procedimientos adecuados de acumulación de datos y encuestas a lo largo de un periodo de interés, por ejemplo de forma anual o semestral. La acumulación de datos genera todo tipo de parámetros en una ventana más amplia del tiempo. Es posible acumular datos eficientemente 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 levantamiento aparecen nuevos individuos, dado que el diseño de las encuestas repetidas selecciona diferentes individuos en cada punto del tiempo. Este es el caso de la *Gran Encuesta Integrada de Hogares de Colombia* que está diseñada para tener representatividad a niveles de desagregación mayores, juntando los individuos observados en los doce levantamientos continuos en un año. Por otro lado, las encuestas de panel permiten un tipo diferente de agregación, no basado en individuos, sino en variables en el tiempo. A diferencia de las encuestas repetidas, las encuestas de panel, panel rotativo o panel dividido permiten observar a los individuos en diferentes periodos de tiempo y la agregación puede hacerse de forma horizontal, manteniendo a los individuos en las filas y añadiendo columnas cada vez que se observe una nueva medición en un periodo de tiempo diferente.

Capítulo 3

Marcos de muestreo

Todo procedimiento de muestreo probabilístico requiere de un dis-po-si-ti-vo que permita identificar, seleccionar y ubicar a todos y cada uno de los objetos per-te-ne-cien-tes a la población objetivo y que participarán en la selección aleatoria. Este dispositivo se conoce con el nombre de **marco de muestreo**.

Cuando se dispone de un marco de elementos, se puede aplicar un diseño de muestreo de elementos; en muchas 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 construirlo) se debe recurrir a diseños de muestreo en conglomerados; es decir, que se utilizan marcos de conglomerados. Por ejemplo, al realizar una encuesta cuya unidad de observación sean las personas que viven en una ciudad, es muy difícil poder acceder a un marco de muestreo de las personas. Sin embargo, se puede tener acceso a la división sociodemográfica de la ciudad y así seleccionar algunos barrios de la ciudad, en una primera instancia y luego, seleccionar a las personas de los barrios en una segunda instancia. En el ejemplo anterior, los barrios son un ejemplo claro de conglomerados. Estas agrupaciones de elementos tienen la características de aparecer en el estado de la naturaleza. De esta forma, si se dispone de un marco de elementos, por ejemplo, el listado de empleados de una entidad, es posible aplicar un diseño de muestreo de elementos, realizar la selección aleatoria y de acuerdo a ese mismo diseño realizar las estimaciones necesarias. Cuando no existe un marco de muestreo disponible es necesario construirlo. Existen dos tipos de marcos de muestreo, a saber:

- **De Lista:** listados físicos o magnéticos, ficheros, archivos de expedientes, historias clínicas que permiten identificar y ubicar a los objetos que participarán en el sorteo aleatorio.
- **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 en forma tal que su identificación y su ubicación sobre el terreno sea posible.

Es una virtud del marco si contiene **información auxiliar** que permite aplicar diseños muestrales y/o estimadores que conduzcan a estrategias más eficientes con respecto a la precisión de los resultados. O también si la información auxiliar[1] está organizada por órdenes deseables. Se llama información auxiliar **discreta**, si el marco de muestreo permite la desagregación de la población objetivo en ca-te-go-rías o grupos poblacionales más pequeños. Por ejemplo nivel socioeconómico, grupo industrial, etc. Se llama información auxiliar **continua** si existe una o varias características de interés de tipo continuo y positivas. Es deseable que la información auxiliar continua esté altamente relacionada con la característica de interés. Por otra parte, un marco de muestreo es defectuoso si presenta alguno o varios de los siguientes casos:

- **Sobre-cobertura:** se presenta si en el dispositivo aparecen objetos que no pertenecen a la población objetivo. *No son todos los que están.*
- **Sub-cobertura:** 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. *No están todos los que son.*
- **Duplicación:** Ll duplicación en un marco de muestreo se presenta si en el dispositivo aparecen varios registros para 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 generar las correspondientes estimaciones, generando sesgo, pérdida de precisión y, en algunos casos, que los resultados del estudio pierdan toda validez.

En resumen, el marco de muestreo es cualquier dispositivo o mecanismo usado para obtener acceso observacional a la población de interés.

- Identificar y seleccionar una muestra de manera que respete un esquema de muestreo probabilístico.
- Establecer contacto con los elementos seleccionados (mail, CAPI, CATI, personal)

Recuerde que la población objetivo constituye el conjunto de elementos sobre la cual se desea información y se requieren estimaciones de parámetros. La población del marco es el conjunto de todos los elementos que son enlistados directamente como unidades en el marco o identificados mediante un marco más complejo, tal 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 construirlo. Luego, las siguientes características son deseables para un marco de muestreo son:

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

- Las unidades en el marco son identificados con un serial
- Cualquier unidad puede ser ubicada (dirección, teléfono)
- Está ordenado sistemáticamente (geografía, tamaño)
- Puede contener información adicional para cada unidad
- El marco especifica el dominio a donde pertenece cada unidad
- Cada elemento de la población está presente sólo una vez
- No contiene elementos que no estén en la población
- Todos los elementos de la población de interés están en el marco muestral

La calidad del marco puede ser medida mediante la relación que existe entre la población objetivo y la población del marco. Algunas imperfecciones comunes de los marcos de muestreo son:

- Subcobertura: Cuando algunos elementos de la población objetivo no están en la población del marco
- Sobrecobertura: Cuando los elementos que no están en la población objetivo están en la población del marco
- Duplicación: Cuando un elemento de la población objetivo se encuentra listado dentro del marco en más de una ocasión

Para realizar el proceso de selección sistemática de los hogares es necesario contar con un marco de muestreo que sirva de vínculo entre los hogares y las unidades de muestreo y que permita tener acceso a la población de interés. Como regla general, el marco de muestreo debe permitir *identificar* y *ubicar* a todos los hogares que conforman la población objetivo. Como lo afirma Gutiérrez (2016a), el marco de muestreo más utilizado en este tipo de encuestas es de áreas geográficas que vinculan directamente a los hogares o personas. 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 podemos realizar un diseño de muestreo con base en esta forma de aglomeración? Pues bien, las personas se aglomeran en hogares, los cuales a su vez se aglomeran en comunidades más grandes: barrios, comunas, segmentos. Estas comunidades forman ciudades y la reunión de estas divisiones es el país. Por lo tanto, a pesar de que ningún país tiene a disposición una lista actualizada de todos los hogares junto con su ubicación e identificación, sí existe en todos los países una lista actualizada de los segmentos cartográficos presentes en las zonas urbanas y rurales. 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, entonces de forma indirecta estaremos seleccionando una muestra de hogares que puede representar la realidad de todo un país.

I Conceptos básicos

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 debe surtir varias etapas, de acuerdo a las agrupaciones definidas en los marcos de muestreo, que usualmente son marcos de área obtenidos de la división geográfica del país, región o municipio en áreas menores mutuamente excluyentes. Los institutos de estadística en América Latina hacen grandes esfuerzos para mantener actualizados sus marcos de muestreo. Por ejemplo, la *Encuesta Nacional de Hogares* de Costa Rica utiliza un marco muestral construido a partir de los censos nacionales de población y vivienda de 2011 y corresponde a un marco de áreas en donde sus unidades son superficies geográficas asociadas con las viviendas. Este marco en particular permite la definición de UPM con 150 viviendas en las zonas urbanas y 100 viviendas en las zonas rurales. En general, el marco está conformado por 10461 UPM (64.5% urbanas y 35.5% rurales). Gambino y Silva (2009) mencionan que, en la práctica, la consecución de los marcos de lista de lo hogares en la última etapa del muestreo puede tornarse difícil puesto que dentro del conglomerado no es obvio observar de manera exhaustiva los hogares, especialmente cuando la frontera del conglomerado es una línea imaginaria. Por ejemplo en lo urbano, la distinción entre dos conglomerados está demarcada claramente por las calles que conforman la ciudad; sin embargo, en la ruralidad, difícilmente existen caminos que puedan servir para delimitar los conglomerados. De la misma manera, esta delimitación se torna compleja cuando han ocurrido cambios en la infraestructura del área y aparecen nuevas construcciones.

Observe que en general, ante el estudio de un fenómeno social, las desagregaciones geográficas más amplias constituyen un interés natural para los usuarios de las encuestas; es así como los investigadores que planean las encuestas quisieran poder desagregar la información por las regiones geográficas más grandes, que a su vez tienen cierta independencia política y administra. Las estadísticas nacionales que se publican a partir de las encuestas de hogares, cobran mayor relevancia a nivel de regiones, estados o departamentos. Este tipo de desagregaciones geográficas se conocen con el nombre de estratos. Inclusive los diseños de las encuestas de hogares han ido evolucionando para permitir que este tipo de agregaciones tenga representatividad en la encuesta. Aunado a lo anterior, si la característica de interés con la cual se planea la encuesta hace que la distribución de la población sea altamente sesgada, como en el caso de los ingresos o gastos, es recomendable crear un estrato de inclusión forzosa con las unidades más importantes en la población. Esta práctica asegura que el error de muestreo sea más bajo. Algunos países hacen uso de la información censal para definir una estratificación socio económica sobre los segmentos cartográficos del marco de muestreo utilizando para tal fin algunas preguntas del censo de población más reciente. Esta práctica representa una ventaja metodológica porque, en la mayoría de encuestas, los parámetros de interés tienen un comportamiento estructural diferente en cada uno de los subgrupos poblacionales creados, tendiendo a tener una mayor precisión en la estimación de los

parámetros de interés. Por ejemplo, utilizando las preguntas del censo, es posible crear un índice de condiciones de vivienda y con este definir algunos grupos de viviendas mutuamente excluyentes que contengan viviendas parecidas dentro de ellos, pero que entre ellos sean muy disimiles. De esta forma, es posible estratificar los sectores cartográficos de todo un país y generar estimaciones más precisas de los indicadores sociales (como pobreza, ingreso medio, etc.) si el diseño de muestreo contempla este tipo de estratificación.

Para el caso de la *Gran Encuesta Integrada de Hogares* en Colombia, los criterios de estratificación forman dos grupos: el primero correspondiente a las 24 capitales junto con sus correspondientes áreas metropolitanas y el segundo correspondiente al resto de cabeceras municipales, centros poblados y la ruralidad dispersa. Además, la encuesta también contempla criterios de estratificacion 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, el diseño de la muestra maestra del Instituto Nacional de Estadística y Geografía de México contempla este tipo de estratificación basada en los indicadores generados con la información del Censo de Población y Vivienda 2010. Previo al proceso de estratificación sociodemográfica, fue necesario construir 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).

De la misma manera, 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 dice complejo puesto que involucra varias etapas de selección y estratificación. Por ende, los marcos de muestreo están conformados por unidades primarias de muestreo (UPM) que se definen como segmentos cartográficos individuales, como una agrupación de segmentos o incluso como una división de segmentos masivos. Por ejemplo, tomemos en consideración el estrato urbana, en 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, salvo en algunos países, las UPM no tienen el mismo tamaño dentro de los estratos; es decir no están constituidas por un número igual de viviendas. El caso es más evidente es la ruralidad, en donde podría ocurrir que una única UPM agrupe un conjunto de viviendas con demasiada heterogeneidad y una alta dispersión geográfica. Es así como 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 la encuestas de hogares y por esto algunos países están considerando la re-definición de las UPM como unidades con un número uniforme de viviendas.

Es usual que tras el levantamiento de un nuevo censo se actualice el marco de muestreo

con el que se seleccionarán las viviendas y hogares para todas las encuestas subsiguientes. Por la naturaleza de los censos, los INE deben recorrer la geografía de los países produciendo una nueva cartografía que derivará en la actualización de los marcos de muestreo. Por ejemplo, considere un país que cuente con un marco de muestreo que consta de cien mil UPM y, para cada una de estas, se ha logrado construir una estratificación socioeconómica que estuvo basada 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 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. Lo anterior aumenta la incertidumbre en el costo final del operativo, pues si se seleccionan UPM con pocas viviendas, se deberán seleccionar nuevas UPM para cumplir con la cuota de viviendas o bien colapsar algunas otras UPM para mantener la selección inicial.

Con base en lo anterior, se esperaría que la actualización de la cartografía y de los marcos de muestreo se realizara cada diez años. Es importante que estas actualizaciones conlleven a una definición de los marcos de muestreo que permitan tener mayor fluidez en los procesos logísticos de selección de hogares y que induzcan una mejora en la precisión de las estimaciones de los parámetros de interés. Por ejemplo, una forma muy conveniente de abordar este desafío es creando UPM que contengan, en la medida de lo posible, un mismo número de viviendas y, de esta manera, mantener una distribución uniforme en cada estrato. Siguiendo el consejo de Valliant, Dever y Kreuter (2013, pág. 212), si el equipo de planeación de la encuesta tiene la flexibilidad de definir las UPM, como usualmente es el caso en las encuestas de hogares, entonces las UPM definitivamente deberían estar conformadas por una cantidad igual de viviendas.

Los censos de población y vivienda se constituyen en el primer insumo principal de las encuestas de hogares. En promedio, los países de la región realizan los censos cada diez años, y en este levantamiento masivo de información se enlistan todos los hogares del país, se enumeran todos los habitantes del país y se observan algunas variables de interés que servirán a su vez para asentar las bases de comparación de las cifras en los siguientes diez años. El periodo que existe entre la realización de dos censos se denomina *periodo intercensal* y en este se realizan encuestas de hogares de diferentes constructos económicos y sociales. Los Institutos Nacionales de Estadística (INE) utilizan las particiones geográficas y cartográficas generadas en el levantamiento del censo con el fin de seleccionar, mediante diseños en varias etapas, muestras de hogares. Comúnmente, estas particiones reciben el nombre de secciones cartográficas y están formadas por un número determinado de hogares contiguos. En adelante nos referiremos a estas particiones como Unidades Primarias de Muestreo (UPM), la cuales en el área urbana, pueden ser manzanas o agregaciones de manzanas, y en área rural pueden ser veredas o sectores censales definidos de antemano.

Para aumentar la eficiencia de la inferencia en las encuestas de hogares, es de particular interés que el marco de muestreo permita clasificar a las UPM de acuerdo con su

nivel socio-económico con el fin de poder realizar selecciones independientes en cada categoría de la clasificación. De esta forma se garantiza la homogeneidad entre grupos y se disminuye la incertidumbre de la estimación. Este proceso se conoce con el nombre de estratificación. En la literatura especializada, es posible encontrar varias metodologías que pretenden clasificar las UPM. Este documento realiza un resumen no exhaustivo de las principales técnicas utilizadas por los INE de la región, propone un algoritmo para encontrar la mejor estratificación basada en los datos de los censos e ilustra los procedimientos computacionales necesarios para implementar esta metodología en el software estadístico `R`.

Esta sección recoge algunas metodologías de estratificación vigentes en la literatura y aplicadas por otros institutos de estadística de la región, así como el seguimiento a la ejecución de la partición del marco por parte de los funcionarios del INE. Luego de un breve introducción, la sección dos establece las diferencias entre la estratificación de las bases de datos a nivel personas y UPMs; la sección tres resume de forma no exhaustiva algunas de las metodologías más usadas para la estratificación de marcos de muestreo, con dos enfoques: univariados sobre medidas de resumen, y multivariados sobre toda la matriz de clasificación; la sección cuatro presenta los criterios de evaluación de los métodos de estratificación; la sección cinco presenta un pequeño resumen de los resultados finales de la estratificación del marco realizada por el INE; finalmente, la sección seis expone algunas conclusiones y recomendaciones en el corto plazo.

II Metodologías de estratificación

A Información a nivel de UPM

En primer lugar, es necesario tomar en consideración que la estratificación que se pretende realizar es a nivel de las UPM. Esto implica que una vez que la UPM esté categorizada en algún estrato, todos sus componentes también estarán clasificados en la misma categoría; por consiguiente las personas y los hogares de la UPM pertenecerán al estrato en el cual la UPM fue clasificado. Cabe resaltar que, tomando en consideración la información recolectada en el censo, es posible también clasificar a las personas o a los hogares en una primera instancia y después agregarlos hasta llegar al nivel de la UPM; sin embargo, en la práctica este proceso puede resultar un poco más complejo y no son claras sus ventajas. Por lo anteriormente mencionado, este documento estará enfocado en la clasificación de las UPM a partir de una matriz de información a nivel de esta misma agregación.

Con la información del censo se deben seleccionar o crear las variables estén relacionadas directamente con los fenómenos que se estudiarán por las diferentes encuestas de hogares a lo largo del periodo intercensal. En general, es posible clasificar estas variables en los siguientes constructos:

- *Demografía y estructura de la población*: sexo, edad, parentesco, origen extranjero, pertenencia a grupos indígenas, número de hijos, número de dependientes, etc.
- *Educación*: analfabetismo, asistencia, años de estudios, grado de escolaridad, etc.
- *Mercado de trabajo*: población en edad de trabajar, pertenencia a la fuerza de trabajo por sexo, condición de ocupación por sexo, rama de actividad, etc.
- *Características de la vivienda*: tipo de vivienda, materiales de construcción, hacinamiento, equipamiento, etc.
- *Acceso a servicios básicos*: fuente de agua, alcantarillado, acceso a salud, acceso a seguridad social, etc.

Debido a que las UPM tienen, en estricto rigor, tamaños diferentes, la escala y el nivel en el que se midan los indicadores puede afectar los procesos de clasificación. Luego, si la matriz de información con la cual se realiza la estratificación se construye con base en el número de personas (con determinadas características) dentro de la UPM, al no tener en cuenta el tamaño de esta, es muy probable que las metodologías de estratificación no logren agrupar de forma homogénea a las UPM. Por ejemplo, asuma que hay dos UPM con tamaño 100 y 300 hogares, que agrupan a 200 y 400 personas en la fuerza de trabajo, y además suponga que una de las variables de la matriz de información se define como el número de personas ocupadas. A su vez, asuma que la primera UPM pertenece a un sector acaudalado y la segunda UPM pertenece a un sector marginal. Es posible que el número de personas ocupadas en ambas UPM sea de 150 y que por esta razón queden erróneamente clasificadas en el mismo grupo. Por ende, definir la matriz de información en términos relativos (porcentaje de ocurrencia de cada variable) es una mejor alternativa 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 que se ha definido el conjunto de variables que entrarán 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 realizar un proceso de refinamiento sobre esta matriz para eliminar aquellas variables que puedan estar altamente correlacionadas con el resto de las variables o que puedan expresarse como combinación lineal de otras variables. De esta manera, se evitan los problemas de multicolinealidad y se asegura una estratificación parsimoniosa. Al final se supone que se debe contar con una matriz de información $\mathbf{X}$ compuesta por P variables, y n_I filas; en donde cada fila de la matriz de información representará la observación de las UPM a novel censal para cada una de las P variables.

La teoría estadística ha definido que la mejor estratificación es aquella que minimice los errores de muestreo de los estimadores, expresados en forma de varianzas o errores estándar. Además, una particularidad de los procesos de estratificación es que las varianzas de estos estimadores dependen a su vez de la variación de los microdatos a nivel poblacional, que se han observado en el censo. Sin embargo, lo que podría resultar

ser una estratificación óptima para una variable tal vez sea, al mismo tiempo, una estratificación pésima para otras variables. Más aun, sabiendo que no todas las variables de interés que se observarán en las encuestas durante el periodo intercensal han sido medidas y observadas en el censo, se debe estudiar muy bien, por medio del estudio de numerosos escenarios, qué estratificación utilizar.

Todos los INE de la región tienen, en mayor o menor grado, evidencia empírica de que la mayoría de los fenómenos que se observan en las encuestas de hogares están supeditados a la distribución de la población en las UPM. Por ejemplo, si lo que se quiere medir es la informalidad en el mercado de trabajo, seguramente nos encontraremos con que este fenómeno está mucho más presente en aquellas UPM marginales, en donde también estarán presenten otros fenómenos como menos años de educación, menores tasas de acceso a la salud, menores ingresos y gastos, mayores tasas de embarazo adolescente, entre otros. Lo anterior conlleva a afirmar que existe una alta correlación entre la UPM que se habita y la incidencia de fenómenos sociales y económicos. Por lo tanto, los ejercicios de estratificación que se deben estudiar tendrán una alta consistencia interna, de tal manera que al escoger la mejor estratificación se garantiza que el INE dispondrá de una clasificación óptima en el periodo intercensal para todas las encuestas de hogares que se ejecuten.

En general, hay dos grandes escenarios que deben ser revisados al momento de 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). Para cualquiera de estas, se recalca que el objetivo es encontrar la mejor partición que asegure que la varianza de los estimadores de muestreo sea mínima. A continuación se presentan algunas técnicas que se pueden considerar. Estas técnicas están disponibles en el software estadístico `R` mediante las librerías `stratification` (Baillargeon y Rivest 2011) y `SamplingStrata` (Barcaroli 2014). En ambos casos existe documentación disponible acerca de cómo utilizar las funciones de estratificación.

B Univariadas sobre medidas de resumen

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

Por ejemplo, si se utiliza la técnica de componentes principales, entonces se tomaría como medida de resumen el primer componente, que resulta ser función del vector pro-

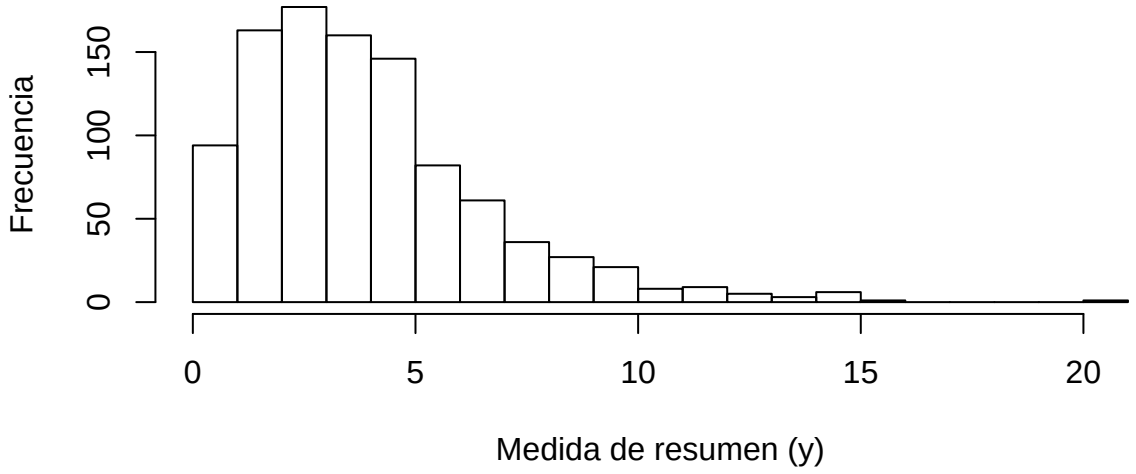

Figura 3.1: *Histograma de la medida de resumen (y) sobre las UPM*

pio asociado al mayor valor propio de la matriz de covarianzas asociada a la matriz de información. Por otro lado, 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 es una parte importante en la aplicación de las técnicas de estratificación. Nótese que la matriz de información está construida por cinco constructos censales (*demografía y estructura de la población*, *educación*, *mercado de trabajo*, *características de la vivienda* y *acceso a servicios básicos*) que deberían ser resumidos en una medida de bienestar de la UPM, que a su vez debe tener sentido en cuanto a la relación (o contribución) de las variables al componente o factor. 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:

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

Nótese que se esperaría que esta variable de resumen, al estar definida como una medida de bienestar sobre las UPM, tuviera un comportamiento sesgado, tal como se puede observar en la siguiente figura.

En particular, el INE de Chile ha realizado un valioso estudio de modelación del *ingreso autónomo* con algunos indicadores homologados, disponibles en la encuesta CASEN y haciendo uso de la técnica de *random forest*. Con base en este modelo, es posible determinar una predicción del ingreso total de los hogares en el censo. Es posible considerar a este *ingreso autónomo imputado* como una medida de resumen del bienestar de la UPM y realizar los mismos análisis de estratificación sobre esta predicción. Para que este escenario sea comparable con los demás, este modelo debe volver a ajustarse teniendo en cuenta exactamente las mismas variables incluidas en la matriz de información. Además, valdría la pena volver a preguntarse si esta componente del ingreso es la que mejor resumiría el bienestar de las UPM.

Partición en cuantiles (Q)

Este método divide la población de UPM en grupos creados a partir de la división en intervalos regulares de la distribución de la medida de resumen. Los cuantiles más usados son los cuartiles (que dividen la población en cuatro grupos), los quintiles (que dividen la población en cinco grupos) y los deciles (que dividen la población en 10 grupos); sin embargo, con los propósitos de estratificación, también es útil considerar la partición en terciles (que dividen la población en tres grupos).

Método de raíz de frecuencia acumulada (DH)

Dalenius y Hodges (1959) propusieron esta técnica de estratificación basada en la acumulación de la raíz cuadrada de las frecuencias acumuladas de la medida de resumen sobre las UPM. Esta técnica es exacta y no requiere de algún procedimiento iterativo. La idea principal de esta técnica es encontrar grupos que minimicen la siguiente función:

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

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

Estratificación óptima (LH)

Lavallée e Hidiroglou (1988) propusieron por primera vez la construcción de una estratificación óptima para poblaciones de encuestas reales, basada en la minimización de una expresión ligada al tamaño de muestra. Más adelante Kozak (2004) definió un algoritmo iterativo mediante arranques aleatorios para optimizar el proceso de minimización.

Estratificación geométrica (GH)

Gunning y Horgan (2004) desarrollaron este método con el objetivo de que los coeficientes de variación de la medida de resumen tiendan a ser iguales dentro de los estratos y, de esta forma, encontraron que los límites que definían estos grupos estaban conformados en progresión geométrica.

C Multivariadas sobre la matriz de información

Partiendo de la matriz de información $\mathbf{X}$ a nivel de las UPM, es posible considerar algunos procedimientos que no necesitan de la reducción a una sola dimensión, sino que admiten tantas dimensiones como indicadores definidos en las columnas de la matriz $\mathbf{X}$. Teniendo en cuenta que en el periodo intercensal se realizarán encuestas que miden variables que están fuertemente ligadas a las observadas en el censo, entonces encontrar una estratificación que sea óptima para todo el conjunto de variables de la matriz de información asegurará una partición óptima para todas las encuestas realizadas en el

periodo intercensal. Las siguientes metodologías permiten minimizar conjuntamente la varianza de los P estimadores de muestreo en un diseño estratificado.

K-medias (KmJ)

Jarque (1981) propone utilizar una versión modificada del algoritmo de K-medias (Macqueen 1967), cuya objetivo es la minimización de la siguiente función de distancia:

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

En donde $\mathbf{x}_j$ corresponde a la medición de las P variables de la matriz de información en la k-ésima UPM, $\bar{\mathbf{x}}_h$ es el vector de medias de la matriz de información en el estrato h y $\mathbf{\Lambda}$ es una matriz diagonal de tamaño $P \times P$ cuyas entradas se definen como la varianza de las P variables de la matriz $\mathbf{X}$, es decir $\mathbf{\Lambda}[p,p] = S^2_{x_p}$. Esta modificación tiene como objetivo minimizar la relación entre la varianza de un estimador de muestreo estratificado con asignación proporcional y la de un muestreo aleatorio simple. Como se puede ver en el anexo, cuando $\mathbf{\Lambda} = \mathbf{I}$, el algoritmo resultante es idéntico al algoritmo clásico de K-medias.

Partición genética (BB)

Ballin y Barcaroli (2013) argumentan que la mejor estratificación es aquella partición del marco de muestreo que asegura el mínimo costo muestral que satisfaga algunas restricciones de precisión; o, que maximice la precisión de los indicadores de interés bajo las restricciones. Haciendo uso de algoritmos genéticos evolutivos, esta la 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 particiones (estratos atómicos). Este universo de posibles estratificaciones evoluciona, 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.

III Evaluación y escogencia de la mejor estratificación

En la evaluación de los escenarios de estratificación entran las técnicas univariadas y multivariadas. Al final, el resultado de aplicar una u otra técnica es simplemente una clasificación de las UPM. Por lo tanto, cada una de las posibles estratificaciones debe ser evaluada con base en 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 lo tanto, la evaluación

de la estratificación debe estar supeditada también a esta medida, que para la variable $p = 1, \ldots, P$, está dada por:

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

En 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 otro lado, Gutiérrez (2016a, página 184) demuestra que, cuando la asignación es proporcional, esta relación se puede escribir de la siguiente manera:

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

En 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. Nótese que una ventaja de expresar el efecto de diseño como en la ecuación anterior es que no dependerá del tamaño de 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. Siguiendo las ideas de 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:

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

Ante una estratificación pertinente, se esperaría que $Var_{ST}(\bar{x}_p) < Var_{SI}(\bar{x}_p)$, por lo tanto $0 < DEFF_p < 1$, lo que conlleva a que $0 < F(S) < P$. Luego, se debería escoger el escenario para el cual $F(S)$ fuera mínimo. Nótese que, para cada uno de los escenarios en estudio, es necesario fijar el número de estratos; en general se propende porque el número de grupos esté entre tres y cinco. Esta escogencia del número de grupos debe ser discutida al interior del INE con los equipos que determinan la rotación de las UPM en cada periodo de levantamiento de las encuestas de hogares. Escoger un número alto de estratos reducirá la varianza, pero a su vez 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 lo anterior, se recomienda restringir los escenarios de evaluación a la consideración de $H = 3$ y $H = 4$ estratos.

El siguiente cuadro ejemplifica la evaluación de estas técnicas para dos escenarios de estratificación (tres y cuatro estratos) en una matriz de información que contiene 8

variables. De la tabla se puede deducir varias conclusiones interesantes. Por ejemplo, para el primer indicador, la mejor estratificación es DH con cuatro estratos; para el segundo indicador, la mejor estratificación es BB con cuatro estratos; mientras que para el último indicador, la mejor estratificación es LH con cuatro estratos. Como se puede notar, para cada indicador existirá un método que induzca una mayor eficiencia, pero que para otros indicadores puede ser deficiente. Esto claramente muestra que la estratificación con respecto a un solo indicador puede ser un procedimiento inadecuado. Por lo tanto, basados en este ejemplo, el mejor método sería DH con cuatro estratos puesto que induce una mayor eficiencia conjunta al reducir el efecto de diseño generalizado.

Cuadro 3.1: *Efectos de diseño $DEFF_p$ y efecto de diseño generalizado $G(S)$ considerando tres ($H = 3$) y cuatro ($H = 4$) estratos para ocho variables.*

DEFF	Q (H=3)	DH (H=3)	LH (H=3)	GH (H=3)	KmJ (H=3)	BB (H=3)	Q (H=4)	DH (H=4)	LH (H=4)	GH (H=4)	KmJ (H=4)	BB (H=4)
$\bar{x}_1$	0.87	0.85	0.81	0.82	1	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	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

XXXXXXXXXXX XXXXXXXXXXX

El INE de Chile culminó satisfactoriamente la estratificación del marco de muestreo 2017. Durante este proceso se consideraron más de 180 escenarios de estratificación junto con varias metodologías apropiadas y acordes tanto a la experiencia internacional como a la literatura especializada. Es importante aclarar que el proceso de estratificación del marco de muestreo tiene por objetivo coadyuvar a los equipos responsables de las estrategias de muestreo en la consecución de diseños de muestreo que induzcan una mayor eficiencia y precisión a la hora de la estimación de las estadísticas oficiales.

De hecho, la propuesta de Inho Park (2003) apunta en este mismo sentido, pues propone que el efecto de diseño de cualquier encuesta se puede descomponer 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, $deff^S$; y por último se tiene el efecto debido al muestreo en varias etapas, $deff^C$. Por lo tanto:

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

Al encontrar la mejor estratificación, el equipo del INE ha permitido que la segunda componente de esta descomposición sea mínima, en cuanto se han seguido los pasos necesarios para ello y se ha escogido la estratificación en tres particiones que aseguró un menor $deff^S$ para los indicadores estudiados. Ahora, cuando el marco se entregue al área de producción estadísticas, será tarea del INE asegurar que los efectos de diseño dados por el efecto de conglomeración y el uso del muestreo en varias etapas $deff^C$ sea mínimo, así como el efecto debido al uso de factores de ponderación desiguales, $deff^W$. En el primer caso, se deberá estudiar, para cada encuesta y operación estadística que haga uso del marco de muestreo estratificado, la relación entre UPMs y hogares a la luz de los indicadores de interés; en particular, es necesario decidir cuántos hogares serán seleccionados en cada UPM y cuántas UPMs serán seleccionadas dentro de cada estrato. De la mima manera, en el segundo caso, también se debe decidir, a la luz de la correlación entre los indicadores particulares de cada encuesta de hogares, cuáles variables de control serán utilizadas en la calibración de los estimadores. De esta forma, en esta estrategia tripartita, se asegura que el efecto de diseño de una encuesta levantada por el INE sea pequeño.

Por otro lado, los escenarios de estratificación que se consideraron tuvieron en cuenta la correlación con el ingreso modelado mediante una técnica de predicción que hizo uso de información auxiliar externa. Los resultados encontrados resaltan la congruencia de la estratificación con el comportamiento de esta variable ilustrativa. Por consiguiente, cabe resaltar esta innovación que seguramente redundará en mejores prácticas de validación, tanto para el INE como para la región. En este sentido, y tendiendo en cuenta que la predicción del ingreso no estuvo presente en la estratificación final, se puede concluir que la estrategia de partición es apropiada para resumir un constructo de poder adquisitivo sobre las UPMs, que al final está altamente correlacionado con la incidencia de los fenómenos sociales de interés, como la pobreza, las necesidades básicas insatisfechas, los patrones de consumo, el ingreso medio, la condición de ocupación, la incidencia de criminalidad y de violencia, los indicadores educativos, entre mucho otros.

Finalmente, ante la rápida y evidente expansión demográfica, resulta apropiado que se realicen pruebas internas acerca del impacto de la actualización del marco de muestreo con el uso de registros administrativos (por ejemplo, de avalúos catastrales). En particular, se debe establecer la existencia de una relación incontrovertible entre el registro administrativo y las categorías de estratificación. Se alienta a los funcionarios del INE a que realicen las pruebas y ejercicios que permitan establecer la eventual viabilidad de estas estrategias de actualización, y su posterior utilización en el siguiente censo de población y vivienda. Asimismo, se aconseja que este marco se establezca en el corto plazo sin ninguna modificación y que los cambios y hallazgos que se evidencien en el levantamiento de las encuestas sean incorporados en los procesos de adjudicación de pesos de muestreo y no en la modificación de las probabilidades de inclusión de las UPMs dentro del marco.

Capítulo 4

Selección de la muestra

Todas las encuestas de hogares en la región comparten el mismo principio inferencial: la selección de una pequeña muestra que puede representar 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. Tal vez en este momento de la historia, la práctica de estos procedimientos ya no genere ningún tipo de asombro, pero el lector podría animarse a contemplar todas los posibles escenarios que una sociedad enfrentaría ante la ausencia de las encuestas de hogares y sus repercusiones en materia del desarrollo social. Es innegable la potencia y el poder que hay detrás de estas operaciones estadísticas que están sustentadas en el muestreo probabilístico que induce una inferencia que procede 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 en estudio, en la forma en que justamente las variables se incorporan en el formulario de la encuesta. Gutiérrez (2016a) afirma que el muestreo es un procedimiento que responde a la necesidad de información estadística precisa sobre la población y los conjuntos de elementos que la conforman; el muestreo probabilístico trata con investigaciones parciales sobre la población que apuntan a inferir a la población completa y en general está basado en los siguientes principios:

- *Aleatorización*: las unidades incluidas en la muestra son seleccionadas 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 esquema de muestreo definido.
- *Inclusión*: todas las unidades de la población tienen una probabilidad no nula de ser incluidas en la muestra. Lo anterior quiere decir que el procedimiento de selección le da chance de ser seleccionado a todas las unidades que componen la población. De esta manera, la muestra final puede estar compuesta por cualquier combinación plausible de hogares o individuos.

Por supuesto, para que los anteriores principios se cumplan a cabalidad, es necesario contar con un instrumento que permita seleccionar a los hogares del país de forma exhaustiva y completa; esto quiere decir que el instrumento debería contener todos y cada uno de los hogares de la población. Dado que no existe una lista que permita identificar y ubicar cada uno de los hogares de la población, entonces se deben contemplar otras posibilidades que permitan lograr el objetivo. Debido al principio natural de la aglomeración social de las poblaciones humanas, es posible lograr este cometido de manera indirecta a través de la definición de marcos de muestreo de áreas.

I Muestras representativas

Las encuestas han tenido una gran trascendencia en las evaluaciones que el gobierno realiza para medir los resultados o el impacto de las diferentes intervenciones del gobierno en la sociedad. El sector gubernamental está interesado en investigar, no sólo, la efectividad de la intervención, sino su implementación y desarrollo; también querrá determinar las mejoras de las condiciones sociales y/o económicas de los distintos actores sociales.

Aunque no constituyen el único método existente, sí es posible afirmar que las encuestas son una de las más poderosas herramientas para la evaluación de las políticas públicas ya a través de ellas es posible realizar evaluaciones que indaguen acerca del impacto de una política, en subgrupos específicos de la población, mediante la estimación de sus efectos en términos de lo que se esperaba al principio de la intervención, comparada con otra intervención, o comparada con un escenario contra factual. De la misma forma, también es posible realizar evaluaciones que indaguen cómo, por qué y bajo qué condiciones una política es exitosa o no.

Tal como lo afirma Gutiérrez (2016a), el muestreo es un procedimiento que responde a la necesidad de información estadística precisa sobre la población y los conjuntos de elementos que la conforman; el muestreo trata con investigaciones parciales sobre la población que apuntan a inferir a la población completa. Es así como en las últimas décadas ha tenido bastante desarrollo en diferentes campos principalmente en el sector gubernamental con la publicación de las estadísticas oficiales que permiten realizar un seguimiento a las metas del gobierno. Cada vez es más notorio un aumento en el uso de las técnicas de muestreo en la evaluación de políticas públicas puesto que se trata de un procedimiento que cuesta menos dinero, consume menos tiempo y porque sus niveles de precisión son muy altos. De esta forma, una muestra bien seleccionada de unos cuantos miles de individuos puede representar con gran precisión una población de millones de personas.

Uno de los objetivos de este documento es traer a discusión algunos conceptos estadísticos involucrados en las encuestas por muestreo y profundizar no sólo en su significado sino también en su interpretación. Existe la creencia de que una muestra representativa

es un modelo reducido de la población y de aquí 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". Nada más falso que esta creencia. En algunos casos es fundamental "sobrerepresentar" algunas categorías o incluso seleccionar unidades con probabilidades desiguales.

La muestra no debe ser un modelo reducido de la población; debe ser una herramienta usada para obtener estimaciones. Es así como el concepto de muestra representativa pierde peso cuando se usa para afirmar que la muestra se parece a la población. La teoría de muestreo se ha ocupado de estudiar estrategias óptimas que permitan asegurar la calidad de las estimaciones. Entonces, en términos de encuestas por muestres, el concepto de representatividad debe estar asociado con la estrategias de muestreo y no sólo con las muestras.

Consecuentemente, la muestra como subconjunto de la población es una herramienta que no admite el calificativo de representativa, puesto que su objetivo no es parecerse a la población sino permitir que, mediante la correcta caracterización de una estrategia de muestreo, el proceso de inferencia logre reproducir la estructura de la población. Lo anterior no indica que debamos abandonar del todo este adjetivo (representativo) en los proceso de muestreo. Por el contrario, el objetivo del equipo técnico experto en la selección de muestras debe estar supeditado a lograr que efectivamente este adjetivo se pueda aplicar 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, acercamiento a modelos estadísticos para hacer frente a la ausencia de respuesta, entre otros. De esta forma, uno de los objetivos de este documento es hacer precisión sobre la estructura de este proceso conjunto para que al final, sea 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.

En resumen, no sería aceptable que un informe final de una encuesta base sus conclusiones en lo observado en la muestra. Y, por tanto, las decisiones de políticas públicas que se tomen con base en la encuesta no deberíasn estar basadas en frases como las siguientes:

- «En la muestra, el 65% de los beneficiarios del programa han aumentado su calidad de vida.»
- «El promedio de ingresos mensuales en los beneficiarios de la muestra del grupo tratamiento es de 685mil pesos.»

La razón de la inadmisibilidad es que los países está en la obligación de proveer cifras confiables, precisas y exactas acerca de las características de la intervención en la población (no en una muestra reducida). Por lo tanto, la muestra debe ser un vehículo que lleve al investigador y al estado a tomar decisiones de carácter poblacional. Luego, en un informe técnico de una encuesta, se espera que las decisiones se tomen con base

en frases como las siguientes:

1. «En la población, el 65% (error del 3.2%) de los beneficiarios del programa han aumentado su calidad de vida.»
2. «El promedio de ingresos mensuales en los beneficiarios de la muestra del grupo tratamiento es de 685mil pesos (error del 2.1%).»

Nótese la gran diferencia que existe en reportar una estadística de la muestra y una estimación de la población. En esta última, se debe reflejar el comportamiento estructural en toda la población de interés. Ahora, debido a que estas estimaciones están basadas en una investigación parcial, entonces todas las cifras de estimación deben adjuntar un error de muestreo que refleje la certidumbre y el nivel de confianza sobre la cifra.

Es posible pensar que, dado que la muestra es un subconjunto parecido a la población, entonces lo que suceda en la muestra puede también suceder en la población. Sin embargo, como se verá más adelante, dado que en este tipo de evaluaciones, el diseño de muestreo es casi siempre complejo, entonces, lo que se percibe en la muestra, resulta tener una gran diferencia con la inferencia en la población. Todo lo anterior está basado en el principio de representatividad que da origen a los factores de expansión, que a su vez soportan el proceso de inferencia en la población.

II Diseños de muestreo

Una vez que los marcos de muestreo se han refinado y se ha definido una estratificación apropiada para las UPM que las componen, es necesario realizar el proceso de muestreo en etapas para la selección final de los hogares. Este proceso de selección debe inducir insesgamiento, además de ser eficiente; por tanto, el procedimiento de muestreo le asigna una probabilidad de selección conocida a cada posible muestra. Por supuesto, esta asignación de probabilidades se realiza de manera teórica. Luego de establecer este conjunto de probabilidades, una única muestra es escogida mediante un mecanismo aleatorio que siga a cabalidad esta configuración estocástica inducida por estas probabilidades que se conocen antes de comenzar el operativo de campo. 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). Nótese que, basado en el anterior principio, estas probabilidades deben ser distintas de cero puesto que, de lo contrario, no se podría garantizar una inferencia insesgada, puesto que estaría excluyendo 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 más adelante.

Existen muchas formas de seleccionar una muestra de hogares y cada una de ellas induce una medida de probabilidad sobre los elementos que conforman la población de interés.

En general, asociado a cada esquema particular de muestreo se define una única función que asocia a cada hogar k con una probabilidad de inclusión en la muestra s, definida de la siguiente manera:

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

Estas probabilidades de inclusión de los hogares, inducidas por los diseños de muestreo asociados a cada encuesta, cumplen con las siguientes propiedades

1. $\pi_k > 0$
2. $\sum_U \pi_k = n$

Observe que la primera propiedad garantiza que ningún hogar será excluido de la selección inicial. Si bien no todos lo hogares serán seleccionados para pertenecer a la muestra s, todos tendrán un chance de ser escogidos por el mecanismo de selección aleatoria. En segunda medida, el tamaño de la muestra de hogares estará inducido por la magnitud de las probabilidades de inclusión. Por esta razón, una encuesta con una tamaño de muestra grande asignará una mayor probabilidad de inclusión a todos los hogares, que una encuestas 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 toman las probabilidades de inclusión en cada esquema.

Existe una clara diferenciación 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. Y el último se define como el proceso de selección que una única muestra que respeta las probabilidades del diseño de muestreo. En una evaluación de política pública es indispensable la definición de los dos componentes. Es decir, si se ha decidido que el diseño de muestreo sea en etapas, es menester del consultor definir exhaustamente cada etapa de muestreo, junto con sus correspondientes unidades de muestreo y diseño de muestreo en cada etapa. Luego, es igual de importante explicar qué algoritmos de selección serán utilizados en cada etapa de muestreo. De esta forma habrá total transparencia en la selección de las unidades y esto redunda en la obtención de cifras oficiales confiables y precisas.

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 de muestreo es necesario tener información suficiente y exhaustiva de la ubicación e identificación de todas las unidades de interés. Su uso es común en las etapas finales de selección de las encuestas, en donde los hogares o personas se seleccionan con las misma probabilidad.

Nótese que una vez se ha escogido la UPM, una parte del operativo de campo deberá estar dedicada al enlistamiento de todas las viviendas en le conglomerado. Cuando se haya realizado este empadronamiento, entonces es posible asignarle la misma probabilidad de inclusión a cada hogar en la UPM. Por ende, las probabilidades de inclusión en el muestreo aleatorio simple sin reemplazo son todas iguales y dadas por la siguiente expresión:

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

Una variante de este tipo de esquemas de selección de muestras de hogares dentro de la UPM es el muestreo sistemático, en 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 siguiente factor $a = N/n$, conocido como el intervalo de salto. Por ejemplo, una muestra sistemática podría ser:

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

.

En donde el primer hogar elegido en la UPM fue el segundo y con saltos sistemáticos de diez hogares se va encuestando los restantes hogares en 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

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

Muestreo proporcional al tamaño

Este tipo de muestreo utiliza como insumo una característica de información auxiliar cuantitativa, también conocida como medida de tamaño (MOS, por sus siglas en inglés). Para la ejecución de este diseño, es menester que el marco de muestreo contenga el valor correspondiente a la medida de tamaño. Este muestreo es utilizado con frecuencia en las etapas iniciales de selección de las encuestas, particularmente en la selección de las UPM que harán parte de la muestra. De esta forma, 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 en la muestra para las UPM serán desiguales y proporcionales a la medida de tamaño. Observe que la cantidad de individuos en las UPM es una cifra conocida, puesto que es el resultado directo de los censos de población y vivienda.

Una de las ventajas de este tipo de muestreos es que hace más eficiente la estimación de los indicadores de interés. Para que esto ocurra, la medida de tamaño debe estar linealmente relacionada con la característica de interés. Esto a menudo ocurre en las

problemáticas sociales, puesto que a mayor número de hogares, se observa una mayor incidencia de estas problemáticas. Por ejemplo, restringidos a un estrato particular, es evidente que en las UPM con mas hogares se observarán mayor número de personas pobres, o de ingresos bajos, o desempleados, etc. La razón detrás de esto es muy simple y poderosa. 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 la raíz cuadrada de este conteo o como una función compuesta de conteos de subpoblaciones. Siendo N_i la medida de tamaño de la i-ésima UPM U_i, es decir el número de hogares que componen esa UPM; n_I 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) se tiene que las probabilidades de inclusión a la muestra s_I están dadas por la siguientes expresión:

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

Por último, no es cierto que la asignación de probabilidades desiguales en las unidades de muestreo induzca sesgo en la encuesta. Esta frase es cierta, siempre y cuando el estimador que se utilice no sea el adecuado. Por ejemplo, la frase pierde su validez cuando se utiliza el estimador de expansión (Hansen-Hurwitz, para el caso de muestreos con reemplazo - Horvitz-Thompson, en muestreos sin reemplazo). Ahora, lo natural es que si el diseño es con probabilidades desiguales, éstas se utilicen dentro de un estimador que considere esta desigualdad y lo menos usual es que se utilice el estimador de expansión en donde habría que corregir el sesgo causado por la omisión de las probabilidades desiguales.

Muestreo estratificado

Esta familia de diseños de muestreo permite realizar inferencias precisas en subgrupos poblacionales de interés, usualmente definidos como agregaciones geográficas grandes. Por ejemplo, si se quieren estimaciones de la incidencia de la pobreza en las regiones geográficas de un país específico, entonces 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. Además, una consecuencia directa de la estratificación es que cada subgrupo tendrá un marco de muestreo de UPM independiente, disyunto y mutuamente excluyente. Esta última caracterización induce una de las mayores ventajas del muestreo estratificado puesto que hay independencia entre los estratos. Esto significa que, al interior de cada estrato, se pueden ejecutar distintas estrategias de muestreo de forma independiente. Es común que en los países de América Latina las áreas geográficas más grandes conformen los estratos (regiones, departamentos y ciudades), asimismo una desagregación común

en investigación social es la división territorial del país: urbano y rural. Evidentemente, la realidad social del entorno urbano difiere tanto del entorno rural que bien vale la pena considerar esta escisión 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. Por ejemplo, si en cada estrato ($h = 1, \ldots, H$) se hubiese planeado un diseño aleatorio simple, entonces las probabilidades de inclusión estarían dadas por la siguiente expresión

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

Por supuesto, es posible que la estrategia de muestreo cambie dependiendo de los estratos. Por ejemplo, en la planificación de las encuestas de uso de tiempo, una de las características de interés por las cuales se quiere indagar es la cantidad de horas que hombres y mujeres dedican a actividades de trabajo no remuneradas. Esta realidad cambia dramáticamente entre zonas rurales y urbanas, tanto que los operativos de campo difieren estructuralmente entre zonas haciendo que, a su vez, los diseños de muestreo cambien radicalmente. Para este tipo de encuestas de hogares, la flexibilidad que tienen los diseños estratificados es un baluarte valioso que permite que los operativos de campo lleguen a buen término.

Se ha dicho que el muestreo estratificado es el método de muestreo por excelencia. Esta afirmación sugiere que el muestreo estratificado es el mejor, lo cual no es verdadero necesariamente. Aunque en muchas ocasiones, la opción de estratificar es adecuada e inclusive conveniente, no es cierto estrictamente que el muestreo estratificado sea el mejor diseño de muestreo. 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. Si, por el contrario, el comportamiento estructural es homogéneo dentro de los estratos y heterogéneo entre los estratos, entonces resultaría adecuado utilizar este diseño de muestreo. Sin embargo, el lector debe notar que hay una brecha muy grande entre el adjetivo adecuado y la calificación de mejor.

Muestreo de conglomerados

Este diseño de muestreo surge como contraparte a la imposibilidad de generar una muestra de hogares directamente 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 generados por 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, regiones, etc.)

la idea general detrás de este diseño es la conformación de unidades homogéneas entre sí (conglomerados), de las cuales se extraerá una muestra y para cada elemento del conglomerado se realizará un proceso exhaustivo de medición censal. De esta forma, es natural definir a las UPM como los conglomerados. Luego de seleccionar una muestra de estas UPM se realiza un censo de hogares sobre cada una de las UPM seleccionadas. Nótese que este proceso logístico induce un esquema económico en términos presupuestales, puesto que limita el operativo de campo a un cierto número de UPM que se deben medir exhaustivamente.

A pesar de que esta estrategia resulte conveniente desde el punto de vista económico, logístico y operativo, ciertamente no lo es desde el punto de vista de la eficiencia estadística, puesto que los errores de muestreo que se producen al utilizar esta metodología son bastante más elevados que la estrategia simple, puesto que 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, generando mayor incertidumbre en la inferencia de la encuesta. Para superar estos inconvenientes se podría pensar en un esquema 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 esquema inviable en la práctica, pero que da paso al esquema de muestreo más común en las encuestas de hogares: la selección por etapas.

Note que los conglomerados de tamaño desigual aumentan la varianza de estimador; es por esto que, en encuestas probabilísticas, se crean conglomerados pequeños, a nivel de manzana, o subsección cartográfica, e incluso hogares. Esta es una práctica muy pertinente, siempre y cuando el muestreo de los conglomerados sea aleatorio simple sin reemplazo. Por supuesto, como la varianza del estimador de expansión está en función de la varianza de los totales de los conglomerados, entonces si hay harta variación en los tamaños, habrá harta variación en los totales y por consiguiente la varianza del estimador será alta. De otra forma, si se tiene conocimiento de una característica de información auxiliar a nivel de conglomerados (también llamadas medidas de tamaño), es posible definir un diseño de muestreo de conglomerados muy desiguales en tamaño, pero que al final induzcan una muy pequeña varianza en el estimador (que considere estas variaciones en las probabilidades), incluso más pequeña que la del muestreo aleatorio simple con conglomerados iguales en tamaño.

Muestreo en varias etapas

En este esquema de muestreo, la idea general es retomar los principios del muestreo de conglomerados y realizar un submuestreo de hogares dentro de los conglomerados o UPM seleccionados inicialmente. Este submuestreo puede ser tan incluyente como sea necesario. En general, en América Latina son muy comunes los esquemas de selección en dos etapas: en la primera etapa se selecciona una muestra de UPM y en la segunda etapa

se selecciona una muestra de hogares en aquellas UPM seleccionadas en la primera etapa. Sin pérdida de generalidad, es posible encontrar en algunos cuantos países esquemas 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 segunda etapa se selecciona una muestra de hogares en aquellas UPM seleccionadas en la segunda etapa. Si un municipio es incluido en la muestra es posible realizar un proceso de aglomeración continúa sistemática, 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. En América Latina todas las encuestas de hogares seleccionan sus muestras haciendo uso de esta técnica.

Si el esquema de muestreo incluye la selección de municipios en la primera etapa, el diseño de muestreo apropiado en esta instancia deberá ser proporcional a una medida de tamaño, que puede ser 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) serán siempre parte del estudio. Por otro lado, es posible que en algunas encuestas exista un submuetreo de personas dentro del hogar. En este caso, Clark y D. G. Steel (2007) aclaran que la escogencia de las personas dentro de los hogares no debería ser aleatoria simple puesto que ciertos grupos poblacionales podrían estar sub-representados o sobre-presentados. En general, el muestreo en varias etapas tiene dos características esenciales que lo hacen robusto, en términos estadísticos, y eficiente al momento de planear la logística del levantamiento de información; estas son:

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

Un esquema de selección bastante usado en las encuestas de hogares de América Latina es el relacionado con los diseños auto-ponderados, lo cuales, en la primera etapa de muestreo seleccionan n_I UPM con probabilidad proporcional al número de hogares que la habitan; es decir:

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

En la segunda etapa de muestreo se seleccionan hogares dentro de las UPM que fueron incluidas en la etapa anterior. Esta selección de hogares se hace con 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 más pequeña que las otras, el número de hogares que serán 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:

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

En los esquemas auto-ponderados, a pesar de tener 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 $n = n_0 * n_I$ es el número total de hogares que serán seleccionados, puesto que resulta ser la multiplicación del número de UPM que fueron seleccionadas en la primera etapa por el número de hogares que serán submuestreados en cada UPM en la segunda etapa. Este tipo de esquemas 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. La razón de esto es que, en la mayoría de encuestas, el submuestreo de las personas es exhaustivo (censo en el hogar) y por ende, la probabilidad de inclusión en el submuestreo es forzosa.

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

Por lo anterior, se tiene que la probabilidad de inclusión de las personas en la muestra es idéntica a la del hogar:

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

Muestreo en dos fases

En algunos casos en donde el marco de muestreo contiene poca o deficiente información para proponer un diseño de muestreo eficiente, el investigador puede obtener información acerca de 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*. Para cada uno de los elementos en esa muestra se debe obtener información

sobre una o más variables auxiliares con el fin de estratificar de mejor manera o simplemente para 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 fase, se selecciona una submuestra mediante un diseño de muestreo conveniente.

Un ejemplo de este tipo de diseños de muestreo se da en el caso de México, en donde el INEGI ha planteado la construcción 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 <<a partir de la construcción del Marco Maestro de Muestreo 2012, se diseñó la Muestra Maestra para lograr mantener actualizada de forma continua la información de las viviendas particulares dentro de esta muestra. El diseño de la muestra maestra consideró y respetó las UPM formadas y la estratificación con que fue construido el marco de muestreo por lo que heredó la mayoría de sus propiedades. El diseño de la Muestra Maestra está basado en la cobertura, tamaño y distribución de las encuestas continuas y periódicas del INEGI. Los tamaños de muestra en viviendas para estas encuestas junto con el promedio óptimo de viviendas a seleccionar dentro de una UPM determinaron el número de UPM a seleccionar para la Muestra Maestra 2012>>. De esta forma, la muestra maestra constituye un elemento esencial para el levantamiento de la Encuesta Nacional de Ocupación y Empleo, la Encuesta Nacional sobre la Confianza del Consumidor, la Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública, la Encuesta Nacional de Gasto de los Hogares, entre algunas otras.

III El diseño de muestreo estándar en una encuesta de hogares

A continuación se describe de manera genérico cómo es un diseño de muestreo típico de una encuesta de hogares en la región. Por supuesto, en la vida práctica existen variantes que se pueden alejar un poco de esta generalización, pero que en general mantienen la misma estructura. Se debe mencionar también que el diseño de muestreo de muchas de las encuestas de hogares que se realizan actualmente mantienen el mismo espíritu de los diseños que anteriormente sirvieron para levantar la información primaria. Es decir, el nivel de innovación en este campo no se da de forma intempestiva, y más bien se podría afirmar que cada vez que se rediseña una encuesta de hogares, el punto de partida será el diseño anterior de la encuesta, lo cual es oportuno si es que se quiere mantener la comparabilidad de las cifras entre los levantamientos periódicos. De esta forma, en general el diseño de muestreo de un encuesta de hogares es probabilístico estratificado y bietápico:

- Se realiza una estratificación por zona: urbano/rural y en algunos casos por región.

- Dentro de cada estrato se realiza un muestreo bietápico. En la primera etapa se seleccionan los conglomerados cartográficos, conocidos como unidades primarias de muestreo (UPM) siguiendo un diseño de muestreo proporcional al número de viviendas, hogares o personas del conglomerado; y en la segunda etapa se escoge aleatoriamente un número fijo de hogares dentro de cada UPM siguiendo un diseño de muestreo aleatorio simple.

Este tipo de esquemas tienen una consecuencia importante en cuanto a la eficiencia estadística. Nótese que en la segunda etapa de muestreo, la variación que se pueda presentar entre los hogares seleccionados en una misma UPM es muy baja con respecto a la variación 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 particular, estos hogares tendrán similares realidades en cuanto a su ingreso, gasto, desocupación, analfabetismo, educación, etc. No es de esperarse encontrar un hogar con altos niveles de ingreso y gasto, cuyos integrantes tienen un nivel de educación muy alto, habitando una vivienda que se encuentre en un sector marginal o deprimido de la ciudad, en donde no hay acceso alcantarillado, servicio de electricidad o agua potable; aunque podría suceder, no es lo que se esperaría. De la misma forma, no es de esperar que un hogar pobre, cuyo ingreso per cápita es bastante bajo y no alcanza para cubrir las necesidades básicas de sus habitantes, ocupe una vivienda ubicada en un sector acaudalado. De la misma manera, en este tipo de investigaciones sociales, la varianza existente entre los conglomerados es inmensa al compararla con la variación dentro de los conglomerados. Por esta razón, es de esperarse que existan diferencias significativos entre las UPM que componen la muestra, puesto que la realidad de una UPM en un sector deprimido no es la misma que la de una UPM en un sector opulento. Esta es una realidad en América Latina que ha ocupado la agenda política y legislativa de las últimas décadas y que en general hace que los diseños de muestreo tengan esta caracterización. Retomaremos esta particularidad en los posteriores capítulos, cuando se aborde el tema de la eficiencia estadística y la medición del error de muestreo.

A continuación se definirán todos los elementos involucrados en 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 sub-muestra seleccionada desde el marco de muestreo compuesto por los sectores cartográficos definidos en el último censo. Suponga que la población de hogares U se divide en N_I UPM, que definen una partición de la población, llamados también **conglomerados** y denotadas 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). Note que la i-ésima UPM U_i $i = 1, \ldots, N_I$ contiene N_i hogares. Luego, el proceso de selección se surte de la siguiente manera:

- Una muestra s_I de UPM es seleccionada de U_I de acuerdo a 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 a la

medida de probabilidad $p_I(s_I)$.

- Para cada UPM U_i $i = 1, \ldots, n_I$ 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 a 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 a la medida de probabilidad $p_i(s_i)$.

Por lo tanto, en la primera etapa se han identificado todos los sectores cartográficos de país y se ha generado el marco de muestreo de las UPM que se separan en grupos mutuamente excluyentes, según las variables de estratificación explícita previamente definidas; dentro de cada estrato se selecciona la muestra de UPM en 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; evidentemente las regiones con más habitantes tendrán una muestra de UPM más grande, aunque esta relación no siempre es lineal. 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 igual probabilidad de ser elegido en la muestra dentro del estrato. Es pertinente observar que, para la segunda etapa se requiere contar con un listado exhaustivo de todos los hogares dentro de todas las UPM seleccionadas. Este proceso de selección requerirá de un empadronamiento previo que, no solo actualice el número de hogares, sino que permita 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 UPM.

IV Recomendaciones en la planeación del diseño de muestreo

- El diseño de muestreo debe ser tan simple como sea posible. Nótese que los esquemas de estimación se van volviendo más complejos a medida que el diseño de muestra agrega más etapas o más fases.
- Si la encuesta se realiza de forma periodica, es necesario actualizar los marcos de muestreo y los tamaños poblacionales a través de tiempo. Si es necesario, el investigador puede apoyarse en las proyecciones demográficas (nacimientos esperados, muertes esperadas y población proyectada) disponibles en fuentes oficiales.
- La mayoría de enuestas son de naturaleza multipropósito. Esto quiere decir que existen múltiples variables de interés. Por lo anterior, el investigador debe definir las varibales más importantes de la evaluación y sobre estas planear el diseño de muestreo. Esta directriz implica que para obtener simultáneamente la precisión

requerida en todas las estimaciones, el tamaño de muestra será un poco más exigente.

- La escogencia de los estratos debe estar directamente determinada por los objetivos de la encuesta y por la definición de las unidades de muestreo.
- Aunque las estrategias de estratificación están generalmente supeditadas a la ubicación e identificación de las unidades de muestreo dentro de los estratos, a veces esta información no se encuentra disponible. Luego, si la encuesta requiera de precisión en estos subgrupos poblacionales de interés, se deben utilizar técnicas de modelación en áreas pequeñas.
- Siempre que no haya un marco de muestreo de elemenetos, es posible utilizar los principios del muestreo en varias etapas, mediante la selección de diferentes unidades de muestreo que continen a los elementos, para lograr una muestra de los elemenetos de interés.
- Si los estratos están conformados por unidades homogéneas que, a su vez, crean categorías heterogéneas entre sí, entonces se dice que el proceso de estratificación es eficiente y el error de meustreo se verá reducido significativamente.
- Si la característica de interés hace que la población sea altamente sesgada, es recomendable crear un estrato de inclusión forzosa con las unidades más importantes en la población. Esta práctica asegura que el error de muestreo para este estrato sea nulo.
- Si la contribución de algunas unidades al total poblacional es no significativa, y además esas unidades son de difícil acceso, es posible redefinir el universo y crear un estrato de exclusión forsoza. En este estrato no se realiza ninguna encuesta y sus respectivas estimaciones deben ser realizadas con modelos de áreas pequeñas.
- Si se requieren estimativos precisos para distintos subgrupos poblacionales, pero no existe un marco de muestreo confiable o actualizado, que permita diseñar un muestreo estratificado, entonces es necesario realizar un esquema de muestreo en dos fases. De esta forma, se selecciona uma muestra aleatoria simple de tamaño moderado. Luego, se realiza un empadronamiento de los individuos en la muestra, a los cuales se eles pregunta acerca de su memebresía a los subgrupos poblacionales de interés. Luego, en una segunda fase, con ayuda de la información recolectada en la primera fase, se realiza un diseño estratificado.
- Cuando se planea un diseño de muestreo estratificado, se debe tener en cuenta las posibles clasificaciones erróneas del marco de muestreo. En evaluación de políticas públicas es común encontrar que algunos encuestados, que se supone que han recibido la intervención del gobierno, reportan que no son beneficiarios del programa. Más aún, cuando se planean evaluaciones de impacto, es posible que los no beneficiarios en una primera medición, que se seleccionaron al levantar la línea de base y que pertenecen al grupo de control, hayan recibido la intervención

en el tiempo intermedio entre el levantamiento de la línea de base y las posteriores mediciones.

- Algunas encuestas se realizan de manera continua en el tiempo. En este tipo de circunstanacias es necesario definir, antes de la primera medición, el tiempo que las unidades van a permanecer en el panel, el esquema de rotación y el esquema de atrición (durante cuánto tiempo las unidades van a ser excluidas de la muestra).
- Se recomienda docuemntar los códigos computacionales que utilizaron para la selección de la muestra probabilística. En cualquier caso, es necesario que los resultados puedan ser replicados por lo que debe fijar una semilla aleatoria al comienzo del código computacional.

Capítulo 5

Tamaño de muestra

Uno de los tópicos que a menudo se dan por sentado en la literatura del diseño y análisis de encuestas de hogares es el tamaño de muestra. De hecho, en los libros de estadística y muestreo se establecen las características principales de los esquemas de muestreo y las propiedades estocásticas de los estimadores sin profundizar en que la muestra debe seleccionarse y que esta selección depende de cuántos hogares se necesiten en el estudio. De hecho, al hablar del tamaño de muestra en una encuesta de hogares, no solo se debe hacer referencia a los hogares, sino también a las personas. En efecto, la determinación del tamaño de muestra también depende del propósito de la encuesta. Por ejemplo, considere una encuesta de propósito múltiples que se levanta cada año con el fin de indagar acerca de múltiples fenómenos demográficos, sociales, educativos, y de condiciones de vida; en este contexto, se debe tener en cuenta que el tamaño de muestra definido debe ser útil, pertinente y apropiado para todas los indicadores que se desean medir. En este capítulo, el lector podrá encontrar una guía útil para identificar la mejor ruta a la hora de abordar el cálculo del tamaño de muestra en las encuestas de hogares.

Confiabilidad y precisión

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

$$IC(1-\alpha) = \left[\hat{\theta} - z_{1-\alpha/2}\sqrt{Var(\hat{\theta})}, \hat{\theta} + z_{1-\alpha/2}\sqrt{Var(\hat{\theta})}\right] \tag{5.1}$$

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 presentir de la distribución normal estándar por el presentir de una distribución $t - student$ con $N_I - H$ grados de libertad, suponiendo que hay N_I unidades primarias de muestreo y H estratos. En este orden de ideas, nótese que

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

donde Q_0 es el conjunto de todas las posible muestras cuyo intervalo de confianza contiene al total poblacional t_y. 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

$$ME = z_{1-\alpha/2}\sqrt{Var(\hat{\theta})} \tag{5.2}$$

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

$$EE = \sqrt{Var(\hat{\theta})} \tag{5.3}$$

Las anteriores medidas sólo 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

$$MER = z_{1-\alpha/2}\frac{\sqrt{Var(\hat{\theta})}}{E(\hat{\theta})} \tag{5.4}$$

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

$$CV = \frac{\sqrt{Var(\hat{\theta})}}{E(\hat{\theta})} \tag{5.5}$$

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

Uno de los primeros paradigmas con el que se debe lidiar es el de la independencia entre las observaciones. Este es un supuesto que gobierna gran parte de la teoría de análisis estadístico, pero que infortunadamente no se aplica en el contexto de las encuestas de hogares. Evidentemente, ante los retos que se debe enfrentar y las diversas estrategias

de recolección de información, las fórmulas que se desprenden del supuesto de que los observaciones corresponden a una muestra de variables independientes e idénticamente distribuidas no son plausibles. La estratificación, las múltiples etapas y la aglomeración de las unidades de muestreo hacen que este supuesto no se cumpla en la práctica y por tanto, utilizar las expresiones tradicionales que se encuentran en los libros introductorios de estadística guiará a tamaños de muestra insuficientes. El problema del tamaño de muestra en encuestas de hogares ha sido abordado por diferentes autores con diferentes enfoques. Quizás uno de los más aceptados es aquel que define un factor de ajuste, llamado efecto de diseño (DEFF), en función de la correlación que hay entre la variable de interés con 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 utilizando un diseño de muestreo de conglomerados o en varias etapas, entonces es imposible afirmar que existe independencia entre las observaciones. Lo anterior hace 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 la da la siguiente relación, denotada como efecto de diseño:

$$DEFF(\hat{\theta}) = \frac{Var_p(\hat{\theta})}{Var_{MAS}(\hat{\theta})} \tag{5.6}$$

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 un total, una proporción, una razón, un coeficiente de regresión, etc.). Por lo anterior, es posible escribir la varianza del estimador bajo el diseño de muestreo complejo como

$$Var_p(\hat{\theta}) = DEFF(\hat{\theta})\ Var_{MAS}(\hat{\theta}) \tag{5.7}$$

$$= DEFF(\hat{\theta})\ \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{y_U} \tag{5.8}$$

En este orden de ideas, dado que en evaluación de políticas públicas es muy factible encontrar información relacionada con las características de interés puesto que, dado que el gobierno debe cumplir con su obligación legal de evaluar las intervenciones, hay encuestas pasadas similares que están disponibles. Por lo tanto, si al implementar un muestreo aleatorio simple y, para un tamaño de muestra n_0, es posible conseguir la precisión deseada, entonces, el valor del tamaño de muestra que tendrá en cuenta el efecto de aglomeración para un diseño complejo estará cercano a $n = n_0 * DEFF$.

En general, en 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 UPM y contienen a su vez a los hogares en 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 las Oficinas Nacionales de Estadística (ONE) es una lista de UPM que agrupa a toda la población del país, se hace necesario más allá de calcular el tamaño de muestra de las personas, también calcular el tamaño de muestra de UPM y hogares en la muestra. Por lo tanto, en este documento se pretende sintetizar los mecanismos de asignación de muestra en tres escenarios que son comunes en la práctica estadística del diseño de encuestas de hogares:

1. Primer escenario: asignación del tamaño de muestra en problemas de inferencia que tienen que ver con la estimación de parámetros de personas. En este escenario se presenta la metodología apropiada para calcular el tamaño de muestra de UPM, hogares y finalmente personas.

2. Segundo escenario: cuando la variable de diseño y en general, las variables más importantes de la encuestas están presentes a nivel de hogar, entonces no es necesario realizar un submuestreo de personas. Partiendo de la lógica presentada en el escenario anterior, se presenta la metodología adecuada para calcular el tamaño de muestra de UPM y de hogares.

3. Tercer escenario: un caso menos común en los países de América Latina se presenta cuando el marco de muestreo empadrona las personas dentro de las UPM y además la encuesta sólo pretende observar características asociadas a los habitantes del hogar (y por tanto no 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 muestra, es menester ser cuidadoso con la notación, para lo cual suponemos una población U de N elementos sobre la que se desea seleccionar una muestra s de n elementos en 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 de personas promedio que se van a seleccionar en cada UPM.

- $\bar{n}_{II}$ es el número de hogares promedio que se van a seleccionar en cada UPM.
- ρ 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 de personas con la característica de interés.
- z_α es el percentil $(1 - \alpha/2)$ asociado a una distribución normal estándar y a la confianza que se requiera en la inferencia.

Para introducir las metodologías apropiadas, junto con las expresiones adecuadas, en cada escenario se definirán las cantidades de interés, se dará una breve introducción al problema y se realizarán los cálculos detenidamente con muchos ejemplos de encuestas reales. Para mantener la uniformidad en los cálculos, todos los ejemplos suponen una población de tamaño $N = 50$ millones, con $N_{II} = 12$ millones de hogares, para el cual se desea obtener una muestra con una confianza del 90%. En cada escenario se supone que el país está dividido en $N_I = 30$ mil UPM, conformadas por segmentos cartográficos (agregaciones de manzanas).

I Tamaño de 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 hogar, será necesario siempre basar nuestros cálculos en el tamaño de muestra de las personas. Por ejemplo, para tener una inferencia apropiada al estimar el ingreso medio percápita, el porcentaje de personas pobres o el porcentaje de personas con una característica particular es necesario definir a la población objetivo como todas las personas que componen un hogar para posteriormente medir la variable de interés que será observada para todas ellas.

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 debe fijar los valores para r y b. Si la unidad de análisis son todas las personas del hogar, entonces 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 de la región o estrato en la 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

ser repetido de forma iterativa en los pasos subsiguientes para poder evaluar la calidad del diseño. De las varias escogencias 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 escogencia 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, entonces $\bar{n}$ se puede descomponer manteniendo la relación con r y b, de la siguiente manera:

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

- **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. Luego de esto se debe calcular el efecto de diseño $DEFF$ como función de ρ y de $\bar{n}$. Ahora, el efecto de diseño $DEFF$, definido como una función de la correlación existente entre la variable de interés y la conformación de las UPM, está dado por la siguiente expresión

$$DEFF \approx 1 + (\bar{n} - 1)\rho$$

- **Calcular el tamaño de muestra de personas**. A partir de las expresiones de tamaño de muestra para diseños de muestreo complejos, 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 error relativo máximo de δ % es de

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

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

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

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

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

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

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

Ejemplo: proporción de personas pobres

Suponga que el parámetro de interés es el porcentaje de personas pobres (cuyo ingreso está por debajo de un umbral preestablecido) y se quiere hacer inferencia con un error relativo máximo del 5%. Por estudios anteriores en este país, 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 hogar puesto que $r = 100\%$. En este país se ha estimado que el tamaño promedio del hogar es de alrededor de $b = 3.5$ personas. Por último, según levantamientos anteriores, la correlación intraclase de la característica de interés con las unidades primarias de muestre es $\rho = 0.034$.

La siguiente tabla resume los resultados del ejercicio para $\bar{n}_{II} = 10$ hogares por UPM, los cuales implican que por cada UPM se entrevistarían en promedio a $\bar{n}$ 35 personas. Con lo anterior se obtendría un efecto de diseño $DEFF = 2.2$, para un total de personas en la muestra de $n = 55936$ que serán observados a partir de la selección de $n_{II} = 15982$ hogares en $n_I = 1598$ UPM.

Hogares promedio por UPM	Personas promedio por UPM	DEFF	Muestra de UPM	Muestra de hogares	Muestra de personas
10	35	2.2	1598	15982	55936

Por supuesto que es posible plantear otros escenarios a medida que se evalúe el efecto que conlleva el cambio del número de hogares que se seleccionan en cada UPM. Por ejemplo, el investigador podría proponer que se seleccionarán en promedio 5 hogares por UPM, lo cual cambiaría el número de UPM que serían seleccionadas en la muestra de la primera etapa, así como también 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 a seleccionar en las UPM, esta cifra también variará. A continuación se muestran algunos resultados que permiten establecer estos escenarios cuando se varía el tamaño de muestra promedio de hogares por UPM. La escogencia del escenario ideal se debe dar en términos de la conveniencia logística y presupuestal en el estudio.

Hogares promedio por UPM	Personas promedio por UPM	DEFF	Muestra de UPM	Muestra de hogares	Muestra de personas
5	18	1.6	2315	11575	40512
10	35	2.2	1598	15982	55936
15	52	2.8	1359	20386	71351
20	70	3.4	1239	24787	86756
25	88	3.9	1167	29186	102152
30	105	4.5	1119	33582	117538
35	122	5.1	1085	37976	132915
40	140	5.7	1059	42366	148282
45	158	6.3	1039	46754	163640

Ejemplo: ingreso promedio por persona

Suponga que se desea estimar el ingreso promedio por hogar con un error relativo máximo del 2%. La variable de interés (ingreso) es continua y se estima que la media oscila alrededor de $\bar{y}_U = 1180$ dólares con una varianza de $S^2_{y_U} = 1845.94^2$. En este caso, la población objetivo son todos los habitantes del hogar por lo cual $r = 100\%$. La composición del hogar se calcula en $b = 3.79$ personas por hogar. Por último, según levantamientos anteriores, la correlación intraclase de la característica de interés es $\rho = 0.035$. Nótese que la correlación intraclase cambia con respecto a la característica que se desee medir.

La siguiente tabla muestra los resultados del ejercicio al seleccionar $\bar{n}_{II} = 15$ hogares por UPM, que a su vez implica que por cada UPM se encontrarían en promedio $\bar{n}$ 57 personas por UPM. Con lo anterior se obtendría un efecto de diseño $DEFF = 3$, para un total de personas en la muestra de $n = 48861$ que serán observados a partir de la selección de $n_{II} = 12892$ hogares en $n_I = 859$ UPM.

Hogares promedio por UPM	Personas promedio por UPM	DEFF	Muestra de UPM	Muestra de hogares	Muestra de personas
15	57	3	859	12892	48861

A continuación se muestran algunos resultados que permiten establecer otros escenarios de muestreo cuando se varía el tamaño de muestra promedio de hogares por UPM. Recuérdese que cualquiera de estos escenarios es válido, desde el punto de vista de la eficiencia estadística, aunque no todos serán válidos si se tienen en cuenta otros aspectos como los logísticos o presupuestales. Por ejemplo, si se escogiera el penúltimo escenario,

entonces para 50 hogares por UPM, se debería encuestar en promedio a 190 personas, lo cual reduciría el número de UPM a 662, pero aumentaría el tamaño de muestra general a 33098 personas, lo cual involucraría mayores costos de contratación de encuestadores, supervisores y seguramente un operativo de campo de más días de duración.

Hogares promedio por UPM	Personas promedio por UPM	DEFF	Muestra de UPM	Muestra de hogares	Muestra de personas
5	19	1.6	1422	7108	26938
10	38	2.3	1000	10001	37902
15	57	3.0	859	12892	48861
20	76	3.6	789	15783	59816
25	95	4.3	747	18672	70766
30	114	4.9	719	21560	81711
50	190	7.6	662	33098	125443
100	379	14.2	619	61857	234439

Ejemplo: proporción de adultos mayores en condición de discriminación

Suponga que la incidencia del fenómeno está alrededor de $P = 5.5\%$ y que la población objetivo son los adultos mayores y se estima que en promedio hay $r = 4.6\%$ de adultos mayores por hogar, cuyo tamaño promedio es de alrededor de $b = 5$ personas. Además, se quiere hacer inferencia con un error relativo máximo del 15%. Por último, según levantamientos anteriores, la correlación intraclase de la característica de interés es $\rho = 0.7$.

La siguiente tabla muestra los resultados del ejercicio, que implica que seleccionar $\bar{n}_{II} = 20$ hogares por UPM implicaría un promedio de $\bar{n}$ 4.6 adultos mayores (casos de interés) por UPM. Con lo anterior se obtendría un efecto de diseño $DEFF = 3.5$, para un total de adultos mayores en la muestra de $n = 7272$ que serán observados a partir de la selección de $n_{II} = 31617$ hogares en $n_I = 1581$ UPM.

Hogares promedio por UPM	Personas promedio por UPM	DEFF	Muestra de UPM	Muestra de hogares	Muestra de casos
20	4.6	3.5	1581	31617	7272

Nótese que en este caso la muestra en los 31617 hogares induce un operativo muy grande que implicaría la observación de $31617 * 5 = 158085$ personas, de las cuales

$n = 7272$, serían los casos de interés. Como se ha visto en los anteriores ejemplos, es posible plantear otros escenarios a medida que se evalúe el efecto que conlleva el cambio del número de hogares que se seleccionan en cada UPM. A continuación se muestran algunos resultados que permite establecer estos escenarios cuando se varía el tamaño de muestra promedio de hogares por UPM.

Hogares promedio por UPM	Personas promedio por UPM	DEFF	Muestra de UPM	Muestra de hogares	Muestra de casos
5	1.1	1.1	1985	9926	2283
10	2.3	1.9	1716	17157	3946
15	3.5	2.7	1626	24387	5609
20	4.6	3.5	1581	31617	7272
25	5.8	4.3	1554	38848	8935
30	6.9	5.1	1536	46074	10597
50	11.5	8.3	1500	74983	17246
100	23.0	16.4	1472	147222	33861

II Tamaño de muestra para UPM y hogares

En algunas ocasiones la variable de interés y la unidad de observación están a nivel de hogar. Por ejemplo, cuando todas las variables de interés se miden en la vivienda. En este caso es posible modificar el algoritmo de la sección anterior para seleccionar únicamente a los hogares en la muestra, sin necesidad de realizar un submuestreo de personas. En este caso algunas cantidades desaparecen porque no son objeto de la población de hogares, por ejemplo r y b; algunas otras expresiones deben ser redefinidas al contexto de los hogares, como por ejemplo, el coeficiente de correlación intraclase ρ, el efecto de diseño y todas las expresiones de tamaños de 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 sigue siendo el insumo más importante del algoritmo y se propone crear escenarios de muestreo a partir de su modificación y 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_{II}de la variable de interés *a nivel del hogar* con el agrupamiento por UPM definido por la división cartográfica del último censo. Además de los anterior, el efecto de diseño $DEFF$ será función del tamaño de muestra promedio de hogares por UPM $\bar{n}_{II}$, como se muestra a

continuación:

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

- **Tamaño de muestra de hogares**. Partiendo de las expresiones de tamaño de muestra generales para muestreos complejos y teniendo en cuenta que la población de interés son lo hogares y que la variable de interés y_{II} está a nivel de hogar, entonces el tamaño de muestra necesario para alcanzar un error relativo máximo de δ % es de

$$n_{II} \geq \frac{S^2_{y_{II}} DEFF}{\frac{\delta^2 \bar{y}^2_{II}}{z^2_\alpha} + \frac{S^2_{y_{II}} DEFF}{N_{II}}}$$

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

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

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

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

A Ejemplo: gasto promedio del hogar

Suponga que se desea estimar el promedio de gasto en dólares en los hogares del país con un error relativo máximo admisible del 3.5%. La variable de interés (gasto) es continua y se estima que la media oscila alrededor de $\bar{y}_U = 1407$ dólares con una varianza de $S^2_{y_U} = 2228^2$. En este ejemplo se supone que el país está dividido en $N_I = 10$ mil UPM y la correlación intraclase de la característica de interés, medida a nivel del hogar, con las UPM es de $\rho_{II} = 0.173$.

La siguiente tabla muestra los resultados del ejercicio para $\bar{n}_{II} = 12$ hogares por UPM, que serán observados a partir de la selección de $n_I = 1338$ UPM y $n_I = 16056$ hogares que inducen un efecto de diseño $DEFF = 2.9$.

Hogares promedio por UPM	DEFF	Muestra de UPM	Muestra de hogares
12	2.9	1338	16056

A continuación se muestran algunos resultados que permiten establecer otros escenarios de muestreo al variar el tamaño de muestra promedio de hogares por UPM. Nótese que,

por ejemplo, en el caso de seleccionar 20 hogares por UPM, se debería seleccionar una muestra de 23695 hogares en tan solo 1185 UPM. Por otro lado, si sólo se seleccionaran 2 hogares por UPM, se tendrían que visitar 3246 UPM en todo el país, aunque el número de encuestas totales descendería a 6493. Para este tipo de operativos, en donde los cuestionarios de gasto de los hogares están acompañados de un operativo exhaustivo que le permite al investigador conocer los hábitos de consumo del hogar de forma desagregada, y en donde se visita el hogar durante un periodo de tiempo determinado, tal vez sea más conveniente seleccionar más hogares por UPM y menos UPM para que el operativo de campo no exija la contratación de demasiado personal en campo. Al estar agrupados en menos UPM, se podrían definir un mejor levantamiento 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.

Hogares promedio por UPM	DEFF	Muestra de UPM	Muestra de hogares
2	1.2	3246	6493
4	1.5	2102	8407
6	1.9	1720	10320
8	2.2	1529	12233
10	2.6	1414	14145
12	2.9	1338	16056
14	3.2	1283	17967
16	3.6	1242	19877
18	3.9	1210	21787
20	4.3	1185	23695

B Ejemplo: proporción de hogares sin agua potable

Suponga que se desea obtener una muestra con un error relativo máximo admisible del 10% sobre la variable de interés (necesidades básicas insatisfechas en agua) y el parámetro de interés es el porcentaje de hogares con esta carencia. En este país, se estima que la proporción de hogares con esta 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 $\rho_{II} = 0.045$.

La siguiente tabla muestra los resultados del ejercicio para $\bar{n}_{II} = 10$ hogares por UPM, que serán observados a partir de la selección de $n_I = 436$ UPM y $n_{II} = 4357$ hogares que inducen un efecto de diseño $DEFF = 1.3$.

Hogares promedio por UPM	DEFF	Muestra de UPM	Muestra de hogares
10	1.3	436	4357

A continuación se muestran algunos resultados que permiten establecer otros escenarios de muestreo al variar el tamaño de muestra promedio de hogares por UPM. Observe que el efecto de diseño DEFF es directamente proporcional al número de hogares entrevistados por UPM y al tamaño de muestra de hogares final; de la misma manera, es inversamente proporcional al número de UPM.

Hogares promedio por UPM	DEFF	Muestra de UPM	Muestra de hogares
5	1.1	758	3790
10	1.3	436	4357
15	1.5	328	4924
20	1.6	274	5490
25	1.8	242	6057
30	2.0	221	6624
35	2.2	205	7190
40	2.3	194	7757
45	2.5	185	8323

III Tamaño de muestra para UPM y personas

En algunos casos en donde la variable de interés esté a nivel de persona y el cuestionario de la encuesta no induzca preguntas acerca del hogar, y además exista un inventario detallado de las personas que residen en cada UPM, es posible evadir la selección de los hogares e ir directamente a la selección de personas. En este caso, la lógica del cálculo del tamaño de muestra se mantiene modificando en cierta manera el algoritmo de las secciones anteriores, tal 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 que se desea encuestar en cada una de las 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. Nótese que, debido al efecto de r, la siguiente relación se mantiene

  $$\bar{n}^e = \bar{n} * r$$

  En donde $\bar{n}^e$ denota el número promedio de encuestas realizadas en cada UPMS sobre la población de interés.
- **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 ρ^e, está dado por la siguiente expresión

$$DEFF^e \approx 1 + (\bar{n}^e - 1)\rho^e$$

- **Tamaño de muestra de personas**. A partir de las expresiones de tamaño de muestra para muestreos complejos, calcular el tamaño de muestra necesario para lograr una precisión adecuada en la inferencia. En primer lugar, el tamaño de muestra necesario para alcanzar un error relativo máximo de δ % es de

$$n^e \geq \frac{S^2_{y_U} DEFF^e}{\frac{\delta^2 \bar{y}^2_U}{z^2_\alpha} + \frac{S^2_{y_U} DEFF^e}{N}}$$

  Si lo que se quiere estimar es una proporción P, entonces la expresión apropiada para calcular el tamaño de muestra estará dada por

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

- **Tamaño de muestra final**. Es necesario calcular el número total de personas que deben ser seleccionados para lograr observar a quienes hacen parte de la población objetivo. Esta cantidad está dada por la siguiente expresión.

$$n = n^e / r$$

- **Cálculo del número de UPMs**. Finalmente, las personas se observan a partir de las UPM. En este paso final es necesario calcular el número de UPM que deben ser seleccionadas en el muestreo a partir de la relación.

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

A Ejemplo: ingreso promedio en personas empleadas

Suponga que se desea estimar el ingreso promedio en las personas empleadas con un error relativo máximo admisible del 2%. La variable de interés (ingreso) es continua y se estima que la media oscila alrededor de $\bar{y}_U = 1458$ dólares con una varianza de $S^2_{y_U} = 2191^2$. Nótese que la población objetivo son todas las personas empleadas, cuya proporción se estima en $r = 46\%$. La correlación intraclase de la característica de interés es $\rho = 0.038$.

La siguiente tabla muestra los resultados del ejercicio al seleccionar $\bar{n}_{II} = 50$ personas por UPM, que a su vez implica que por cada UPM se encontrarían en promedio $\bar{n}$ 23 personas ocupadas por UPM. Con lo anterior se obtendría un efecto de diseño $DEFF = 1.8$, para un total de personas en la muestra de $n = 60933$ de las cuales habrían $n_e = 28029$ ocupados repartidos en $n_I = 1219$ UPM.

Personas promedio por UPM	Casos promedio por UPM	DEFF	Muestra de UPM	Muestra de Casos	Muestra de personas
50	23	1.8	1219	28029	60933

A continuación se muestran algunos resultados que permiten establecer otros escenarios de muestreo cuando se varía el tamaño de muestra promedio de hogares por UPM.

Personas promedio por UPM	Casos promedio por UPM	DEFF	Muestra de UPM	Muestra de Casos	Muestra de personas
25	12	1.4	1857	21360	46435
50	23	1.8	1219	28029	60933
75	34	2.3	1006	34695	75424
100	46	2.7	899	41360	89913
125	58	3.1	835	48023	104398

B Ejemplo: proporción de analfabetas pobres

Suponga que se quiere estimar la proporción de incidencia de la pobreza sobre la población analfabeta con un 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, tal 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 $\rho^e = 0.045$.

La siguiente tabla muestra los resultados del ejercicio al seleccionar $\bar{n}_{II} = 100$ personas por UPM, que a su vez implica que por cada UPM se encontrarían en promedio $\bar{n}$ 14 personas que no saben leer ni escribir por UPM. Con lo anterior se obtendría un efecto de diseño $DEFF = 1.6$, para un total de personas en la muestra de $n = 32671$ de las cuales habrían $n_e = 4574$ ocupados repartidos en $n_I = 327$ UPM.

Personas promedio por UPM	Casos promedio por UPM	DEFF	Muestra de UPM	Muestra de Casos	Muestra de personas
100	14	1.6	327	4574	32671

Es posible plantear otros escenarios a medida que se evalúe el efecto que conlleva el cambio del número de hogares que se seleccionan en cada UPM. A continuación se muestran algunos resultados que permite establecer estos escenarios cuando se varía el

tamaño de muestra promedio de hogares por UPM.

Personas promedio por UPM	Casos promedio por UPM	DEFF	Muestra de UPM	Muestra de Casos	Muestra de personas
25	3.5	1.1	917	3211	22936
50	7.0	1.3	524	3665	26179
75	10.5	1.4	392	4120	29429
100	14.0	1.6	327	4574	32671
125	17.5	1.7	287	5029	35921

IV Otros escenarios de interés

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

A Tamaño de muestra para la estimación de la diferencia de proporciones

Suponga una población U, que se encuentra particionada en dos subpoblaciones U_1 de tamaño N_1 y U_2, de tamaño N_2. En el levantamiento de una línea de base, una subpoblación corresponde al grupo que no recibe la intervención (grupo control) y la otra al grupo susceptible de recibir la intervención (grupo tratamiento). El interés del investigador está en conocer la diferencia de algunas proporciones entre estos grupos.

Por ejemplo, suponga que se quiere conocer la diferencia entre las proporciones de niños desnutridos. Se espera que la proporción de niños desnutridos del grupo tratamiento no supere la proporción de niños desnutridos en el grupo control. Por lo tanto, el parámetro de interés se escribe como:

$$\theta = P_1 - P_2 = \frac{N_{d1}}{N_1} - \frac{N_{d2}}{N_2} \tag{5.9}$$

En donde $N_{di} = \sum_{k \in U_i} Z_{dik}$, para $i = 1, 2$. Ahora, Z_{dik} es una característica dicotómica que indica si el niño k-ésimo de la subpoblación U_i está en estado de desnutrición.

Por supuesto, bajo muestreo aleatorio simple, un estimador insesgado para θ es

$$\hat{\theta} = \hat{P}_1 - \hat{P}_2 = \frac{\hat{N}_{d1}}{N_1} - \frac{\hat{N}_{d2}}{N_2} \tag{5.10}$$

En donde, $\hat{N}_{di} = \frac{N_i}{n_i} \sum_{k \in S_i} Z_{dik}$, para $i = 1, 2$. La varianza del anterior estimador es:

$$Var(\hat{P}_1 - \hat{P}_2) = Var\left(\frac{\hat{N}_{d1}}{N_1}\right) + Var\left(\frac{\hat{N}_{d2}}{N_2}\right)$$
$$= \frac{1}{n_1}\left(1 - \frac{n_1}{N_1}\right) P_1Q_1 + \frac{1}{n_2}\left(1 - \frac{n_2}{N_2}\right) P_2Q_2$$

Nótese que no hay covarianzas puesto que las selecciones son independientes en cada subgrupo. Ahora, para encontrar el tamaño de muestra óptimo, es útil realizar los **siguientes supuestos**:

1. Asumir que las subpoblaciones son grandes y por ende $N_1 = N_2 = N$
2. Por lo anterior, asumir que los tamaños de muestra pueden ser iguales y tales que $n_1 = n_2 = n$.
3. Esto conlleva a que el tamaño de toda la población es $2N$ y el tamaño de la muestra total (para el grupo tratamiento y el control) es de $2n$.

Luego, la varianza se simplifica de la siguiente manera:

$$Var(\hat{P}_1 - \hat{P}_2) = \frac{1}{n}\left(1 - \frac{n}{N}\right)(P_1Q_1 + P_2Q_2) \tag{5.11}$$

Nótese a su vez que, si el levantamiento de las observaciones no puede ser realizado utilizando muestreo aleatorio simple, sino que por el contrario, la muestra aleatoria fue seleccionada mediante un diseño de muestreo complejo con un efecto de diseño[1] ($DEFF$) no ignorable y mayor a uno, entonces la varianza tomaría la siguiente forma

$$Var(\hat{P}_1 - \hat{P}_2) = \frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1Q_1 + P_2Q_2) \tag{5.12}$$

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

$$IC(95\%)_{P_1-P_2} = (\hat{P}_1 - \hat{P}_2) \pm Z_{1-\alpha/2}\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1Q_1 + P_2Q_2)} \tag{5.13}$$

Lo anterior quiere decir que el margen de error, e, de la encuesta debe ser tal que:

$$e < Z_{1-\alpha/2}\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1Q_1 + P_2Q_2)} \tag{5.14}$$

[1]Recuerde que si el muestreo es aleatorio simple, el efecto de diseño es $DEFF = 1$.

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

$$n > \frac{DEFF(P_1Q_1 + P_2Q_2)}{\frac{e^2}{Z^2_{1-\alpha/2}} + \frac{DEFF(P_1Q_1 + P_2Q_2)}{N}} \tag{5.15}$$

Para el levantamiento de una línea de base, que quiera estimar una diferencia de proporciones $P_1 - P_2$ (donde $Q_1 = 1 - P_1$ y $Q_2 = 1 - P_2$), entonces el tamaño de muestra mínimo[2] necesario para lograr una estimación confiable de esta diferencia, con menos de $\varepsilon\%$ de error, es:

$$n \geq \frac{(P_1Q_1 + P_2Q_2)DEFF}{\frac{\varepsilon^2}{z^2_\alpha} + \frac{(P_1Q_1 + P_2Q_2)DEFF}{N}} \tag{5.16}$$

Por ejemplo, en una encuesta de fuerza laboral que quiera estimar la diferencia entre el porcentaje de personas desocupadas en dos periodos diferentes CITAR CPS del Census Bureau.

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

B Diferencia de proporciones para contraste de hipótesis

Suponga que el investigador desea realizar el contraste de una hipótesis de interés. En particular, suponga que hay dos grupos de interés en la población finita[3] y que la hipótesis está inducida por 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 tamaño del efecto que el investigador desea detectar.

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

$$H_o : P_1 - P_2 = 0 \qquad vs. \qquad H_a : P_1 - P_2 = D > 0$$

Nótese que, acudiendo a la distribución normal de los estimadores de las proporciones,

[2]Note que el tamaño de muestra de toda la evaluación es $2n$, puesto que se debe seleccionar n elementos en el grupo control y n elementos en el grupo tratamiento.

[3]En este apartado, se deben considerar los mismos tres supuestos de la sección 5.3.1.

la regla de decisión en este caso induce el rechazo de 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 $DEFF > 1$, entonces esta regla de decisión rechaza la hipótesis nula si

$$\frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1Q_1 + P_2Q_2)}} > Z_{1-\alpha}$$

En este caso, es necesario controlar la probabilidad de cometer el error tipo 2 (aceptar una hipótesis nula, dado que ésta es falsa). A esta probabilidad se le conoce como *potencia* y, suponiendo que nuestro interés está en $P_1 - P_2 = D$, está dada por

$$\begin{aligned}
\beta &\leq Pr\left(\frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1Q_1 + P_2Q_2)}} > Z_{1-\alpha} \mid P_1 - P_2 = D\right) \\
&= Pr\left(\frac{(\hat{P}_1 - \hat{P}_2) - D}{\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1Q_1 + P_2Q_2)}} > Z_{1-\alpha} - \frac{D}{\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1Q_1 + P_2Q_2)}}\right) \\
&= 1 - \Phi\left(Z_{1-\alpha} - \frac{D}{\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1Q_1 + P_2Q_2)}}\right)
\end{aligned}$$

Por lo anterior, se tiene que

$$1 - \beta \geq \Phi\left(Z_{1-\alpha} - \frac{D}{\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1Q_1 + P_2Q_2)}}\right)$$

Entonces,

$$Z_{1-\beta} \leq Z_{1-\alpha} - \frac{D}{\sqrt{\frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1Q_1 + P_2Q_2)}}$$

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

$$n \geq \frac{DEFF(P_1Q_1 + P_2Q_2)}{\frac{D^2}{(Z_{1-\alpha} + Z_\beta)^2} + \frac{DEFF(P_1Q_1 + P_2Q_2)}{N}} \tag{5.17}$$

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

Para el seguimiento respectivo en una segunda oleada (SO) evaluación de impacto, que quiera estimar el impacto de la intervención, definido como la diferencia en diferencias de las proporciones de interés teniendo en cuenta el grupo y el tiempo en el cual fue realizada la encuesta. De esta forma, el impacto se define como:

$$Imp = (P_{1,SO} - P_{1,LB}) - (P_{2,SO} - P_{2,LB}) \tag{5.18}$$

En donde $Q_{1,SO} = 1 - P_{1,SO}$, $Q_{1,LB} = 1 - P_{1,LB}$, $Q_{2,SO} = 1 - P_{2,SO}$, $Q_{2,LB} = 1 - P_{2,LB}$). Nótese que el índice 1 denota el grupo tratamiento y el índice 2 denota el grupo control, el índice SO denota la segunda oleada y el índice LB denota la línea de base. En este caso, $P_{1,SO}$ describe la proporción de interés en el grupo tratamiento para la segunda oleada y $P_{2,LB}$ denota la proporción de interés en el grupo control para la línea de base, y así sucesivamente. Entonces el tamaño de muestra mínimo[4] necesario para lograr una estimación confiable de esta diferencia, con menos de $\varepsilon\%$ de error, es:

$$n \geq \frac{(P_{1,SO}Q_{1,SO} + P_{1,LB}Q_{1,LB} + P_{2,SO}Q_{2,SO} + P_{2,LB}Q_{2,LB})(1 - TR)DEFF}{\frac{\varepsilon^2}{z_\alpha^2} + \frac{(P_{1,SO}Q_{1,SO} + P_{1,LB}Q_{1,LB} + P_{2,SO}Q_{2,SO} + P_{2,LB}Q_{2,LB})(1 - TR)DEFF}{N}} \tag{5.19}$$

En donde T corresponde a la tasa de traslape ($T = 1$ corresponde a un panel completo, $T = 0.5$ a un semi-panel con traslape del 50% y el caso extremo $T = 0$ a una encuesta en donde no hay traslape) y R se define como la correlación entre las dos mediciones - línea de base y segunda oleada - ($R = 0$ 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 evaluación de impacto de a una política de asistencia laboral a la población por debajo de la línea de pobreza, esta expresión se utilizaría para estimar la diferencia en diferencias de la proporción de desempleados entre los grupos control y tratamiento y en dos periodos de tiempo.

[4]Note que el tamaño de muestra de toda la evaluación es $4n$, en las dos oleadas, puesto que se debe seleccionar n elementos en el grupo control para la línea base, n elementos en el grupo control para la segunda oleada, n elementos en el grupo tratamiento para la línea de base y, finalmente, n elementos en el grupo tratamiento para la segunda oleada.

V Algunas consideraciones adicionales

Cuando la encuesta se ha planeado para que tenga representatividad para algún conjunto de estratos, es necesario replicar 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 de UPM.

A El efecto de diseño en los estratos

La estructura de cálculo del tamaño de muestra tiene como insumo central al efecto de diseño. Una expresión generalizada que da cuenta del efecto de aglomeración en los diseños de muestreo en varias etapas (I. Park y Lee 2006) es la siguiente:

$$DEFF \approx 1 + (\bar{m} - 1)\rho$$

En donde $\bar{m}$ representa el número promedio de hogares seleccionados dentro de cada UPM y ρ es el coeficiente de correlación intraclase, que representa el grado de homogeneidad de la variable de interés dentro de cada hogar. Sin embargo, esta cifra cambia dependiendo de si la inferencia de la encuesta de hogares se quiere realizar a nivel nacional o a nivel regional. Por ejemplo, UN (2005, capítulo 7) presenta el comportamiento de esta medida a lo largo de tres encuestas de hogares en Brasil: la *Pesquisa Nacional por Amostra de Domicílios* (PNAD), la *Pesquisa Mensal de Emprego* (PME) y la *Pesquisa de Padrões de Vida* (PPV). En general, estas encuestas utilizan estratificación y selección de UPM con probabilidades desiguales; además, el tamaño promedio de las UPM es de 250 viviendas, de las cuales son seleccionadas 13 por la PNAD, 20 en la PME y 16 y 8 viviendas en la PPV en la zona rural y urbana, respectivamente.

Basado en lo anterior, se nota que los efectos de diseño no solo son diferentes para cada parámetro que se desea estimar sino que varían de acuerdo a la subpoblación en la que se realice la estimación. Por ejemplo, considere el parámetro *proporción de hogares con electricidad*; en la PNAD se ha estimado que el efecto de diseño para este parámetro es 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 lo anterior, y basado en la expresión que define el efecto de diseño, se observa que, fijando $\bar{m} = 10$, el coeficiente de correlación intraclase varía dependiendo de la zona. En efecto, $\rho = 0.76$ a nivel nacional, $\rho = 0.0033$ en las zonas metropolitanas, $\rho = 0.38$ en las ciudades grandes y $\rho = 0.69$ en las áreas rurales. Lo anterior implica que hay una mayor heterogeneidad de los hogares con electricidad entre las UPM a nivel nacional y en las áreas rurales, es decir algunos hogares tienen electricidad y otros no entre las UPM. Sin embargo, en las zonas metropolitanas la variación de esta variable entre las UPM es casi nula, es decir que todos lo hogares tienen electricidad entre las UPM de estas zonas.

Por otro lado, para la misma encuesta PNAD, los efectos de diseño para el número promedio de cuartos usados como dormitorios es 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 $\bar{m} = 10$, el coeficiente de correlación intraclase es de $\rho = 0.12$ a nivel nacional, $\rho = 0.15$ en las zonas metropolitanas, $\rho = 0.08$ en las ciudades grandes y $\rho = 0.12$ en las áreas rurales. Lo anterior implica que hay una mayor homogeneidad del número de cuartos utilizados como dormitorio entre las UPM del país y de las zonas que lo componen. 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 permitan establecer el tamaño de muestra en la planeación de las encuestas de hogares. Las anteriores expresiones corresponden al número de individuos que deberían ser seleccionados en cada subgrupo de interés. Por lo tanto, estos cálculos deben ser hechos tantas veces como subgrupos de interés exista en la encuesta, adecuando cada expresión a su contraparte poblacional. Por ejemplo, si el interés está en hacer inferencia en dos estratos: el rural y el urbano, entonces se debe calcular estas expresiones dos veces, una para cada estrato. Al final, el tamaño de muestra nacional será la sumatoria de los tamaños de muestra en cada uno de los estratos del país.

B Ajustes de subcobertura

Debido a las características propias de las encuestas de hogares, siempre se presentará un fenómeno que puede ser descrito como una realidad: *existirá ausencia de respuesta en las encuestas de hogares*. En estos términos, los institutos nacionales de estadística deben tomar medidas preventivas al momento de adjudicar los tamaños de muestra en cada estrato, puesto que contar con un tamaño efectivo de muestra mucho menor al planeado inicialmente puede conllevar problemas de sesgo y de precisión en las estimaciones de las cifras nacionales o regionales, con las cuales se aborda la política económica y de desarrollo de los países de la región.

En encuestas de hogares cuyo diseño es longitudinal, no solamente se debe abordar el problema de la ausencia de respuesta al momento de la aplicación de la encuesta, sino que debe ser visto 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. Es así como la atrición se convierte en un problema que enmarca la ausencia de respuesta como un fenómeno al cual se debe prestar atención para evitar problemas de sesgo y baja confiabilidad.

Kalton (2009) advierte que el diseño de la encuesta debe tener en cuenta el ajuste de submuestras; por ejemplo, para estimar el cambio de la condición de pobreza o indigencia en los hogares es necesario realizar un ajuste al tamaño de muestra inicial para que al final de la aplicación de la encuesta el tamaño de muestra efectivo cumpla con los requerimientos de precisión de la inferencia estadística. Los INE pueden estimar, con base en su basta 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 $\phi = Pr(k \in s_r)$, en donde s_r denota el subconjunto de respondientes efectivos, entonces los tamaños de muestra de individuos y hogares serán ajustados al dividirlos por ϕ. De esta forma, si esta probabilidad fue estimada en $\phi = 0.8$, entonces todos los tamaños de muestra calculados en los pasos anteriores deberán ser ajustados como

$$n_{final} = \frac{n_{inicial}}{0.8}$$

Por último, si la información auxiliar lo permite, este ajuste debería realizarse 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 lo rural que en lo urbano, entonces este ajuste debería tenerse en cuenta de forma diferenciada.

C Sustituciones y reemplazos

Una práctica común en los operativos de campo de las encuestas de hogares en Latinoamérica es sustituir las UPM y viviendas para las cuales no se ha obtenido respuesta. Por ejemplo, se consideraría el reemplazo de las UPM cuando no se puede acceder al sitio geográfico por diferentes razones, como 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, no se puede tampoco acceder a ninguno de los hogares que la integran. Los esquemas de sustituciones y reemplazos en las encuestas de hogares utiliza, por lo general, la metodología de *estratificación implícita* que permite seleccionar de manera automática a los reemplazos adecuados de acuerdo a la conformación de subgrupos poblacionales similares.

La estratificación implícita es usada 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 UN (2008, pág. 46), en la mayoría de países la secuencia podría empezar con el estrato urbano, desagregado por departamento, a su vez desagregado por municipio; el estrato rural, de forma similar, es desagregado por departamento, a su vez desagregado por comuna o vereda. Observe que la selección sistemática de UPM está condicionada a la medida de tamaño utilizada en la primera etapa, es decir el número de viviendas que la componen. De esta forma, la estratificación implícita consiste en que, para cada estrato explícito (urbano, rural, regiones, etc.) se crea una lista ordenada de UPM. Esta lista estará ordenada por los estratos implícitos definidas en la planeación de la encuesta (departamento, municipio) y dentro de cada subgrupo se ordenan 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 originalmente, su reemplazo será

la inmediatamente anterior (o posterior) en la lista estratificada implícitamente. Este procedimiento seleccionará como reemplazo a 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 considera un procedimiento similar a la estratificación implícita, los reemplazos de las UPM podrían ser seleccionados aleatoriamente en otro departamento y con un número de viviendas mucho más grande o mucho más pequeño, añadiendo sesgo a la selección inicial.

Aunque la estratificación implícita permite acotar el sesgo generado por la ausencia de respuesta de las UPM, Vehovar (1999, págs. 348 - 349) advierte que se debe tener precaución en cuanto a los usos de esta práctica puesto que también puede conllevar sesgos importantes en las estimaciones de interés. Lo anterior se desprende del hecho de que los individuos ubicados en zonas donde sí es posible acceder diferirán significativamente de aquellos individuos ubicados en las zonas de difícil acceso; es evidente que se trata de 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 levantamiento que permita clasificar el esquema de recolección de información primaria y se valore su impacto en la precisión de los estimadores resultantes.

Capítulo 6

Ausencia de respuesta: imputación y reponderación

En el diseño y puesta en marcha de una encuesta puede ocurrir cierto tipo de situaciones que pueden sesgar las estimaciones finales. Este tipo de sesgos puede ocurrir antes, durante y después de la recolección de los datos. Es tarea del estadístico advertir ante todas las posibles instancias de los problemas que causan los sesgos y procurar que, en todas las etapas de la encuesta, se minimice el error humano y el error estadístico para que al final los resultados del estudio sean tan confiables como sea posible.

I Sesgos generados en las encuestas

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Sesgo de selección

Este tipo de sesgo ocurre cuando parte de la población objetivo no está en el marco de muestreo. Una muestra a conveniencia[1] es sesgada pues 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:

1. La selección de la muestra depende de cierta característica asociada a las propiedades de interés. Por ejemplo: Frecuencia con que los adolescentes hablan con los padres acerca del SIDA.

[1] A pesar de que las muestras por conveniencia o por juicio no pueden ser utilizadas para estimar parámetros de la población, éstas sí pueden proporcionar información valiosa en las primeras etapas de una investigación o cuando no es necesario generalizar los resultados a la población.

2. La muestra se realiza mediante elección deliberada o mediante 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 está reflejando el comportamiento promedio de las compras.

3. Existen errores en la especificación de la población objetivo. Por ejemplo, en encuestas electorales, cuando la población objetivo contiene a personas que no están registradas como votantes ante la organización electoral de su país.

4. Existe sustitución deliberada de unidades no disponibles en la muestra. Si, por alguna razón, no fue posible obtener la medición y consecuente observación de la característica de interés para algún individuo en la población, la sustitución de este elemento debe hacerse bajo estrictos procedimientos estadísticos y no debe ser subjetiva en ningún modo.

5. Existe ausencia de respuesta. Este fenómeno puede causar distorsión de los resultados cuando los que no responden a la encuesta difieren críticamente de los que si respondieron.

6. La muestra está compuesta por respondientes voluntarios. Los foros radiales, las encuestas de televisión y los estudios de portales de internet no proporcionan información confiable.

Sesgo de medición

Este tipo de sesgo ocurre cuando el instrumento con el que se realiza la medición tiene una tendencia a diferir del valor verdadero que se desea averiguar. Éste sesgo debe ser considerado y minimizado en la etapa de diseñó de la encuesta. Nótese que ningún análisis estadístico puede revelar que una pesa añadió a cada persona 2Kg de más en un estudio de salud. Lohr (2000) cita algunas situaciones en donde se presenta este sesgo de medición:

1. Cuando el respondiente miente. Esta situación se presenta a menudo en encuestas que pregunta acerca del ingreso salarial, alcoholismo y drogadicción, nivel socioeconómico e incluso edad.

2. Difícil comprensión de las preguntas. Por ejemplo: ¿No cree que no este es un buen momento para invertir? La doble negación en la pregunta es muy confusa para el respondiente.

3. Las personas tienden a olvidar. Es bien sabido que las malas experiencias suelen ser olvidadas; esta situación debe acotarse si se está trabajando en una encuesta de criminalidad.

4. Distintas respuestas a distintos entrevistadores. En algunas regiones es muy probable que la raza, edad o género del encuestador afecte directamente la respuesta del entrevistado.

5. Leer mal las preguntas o polemizar con el respondiente. El encuestador puede influir notablemente en las respuestas. Por lo anterior, es muy importante que el proceso de entrenamiento del entrevistador sea riguroso y completo.

6. La muestra está compuesta por respondientes voluntarios. Los foros radiales, las encuestas de televisión y los estudios de portales de internet no proporcionan información confiable.

II Tipos de ausencia de respuesta

Todas las encuestas de hogares sufren del fenómeno de la ausencia de respuesta en algunas de las variables de interés. En algunas ocasiones y aún después de un diseño cuidadoso y una planificación logística exhaustiva, esta problemática puede ser tan grande que los resultados de la encuesta pueden quedar en entredicho. Por esta razón, este problema debe ser considerado como una faceta normal, si bien no deseable, en el desarrollo de este tipo de estudios y la planificación y el diseño de todos los levantamientos de información a través de encuestas debe contemplar varios ajustes que prevean las consecuencias de este fenómeno. Es por esto que la última sección del capítulo cuatro abordó el tema del ajuste de subcobertura, que garantiza que el tamaño de muestra efectivo sea el adecuado para realizar un inferencia precisa. De otra forma, si el diseño de la enceusta no tiene en cuenta estos ajustes, el tamaño de muestra final se verá reducido puesto que muchos hogares no contestarán algunas preguntas del cuestionarios, y en algunos casos, muchos hogares no contestarán la totalidad del cuestionario. Existe un consenso completo de que la ausencia de respuesta puede perjudicar severamente la calidad de las estadísticas calculadas y publicadas en una encuesta. Little y Rubin (2002) establecen tres tipos de mecanismos de ausencia de respuesta.

- **Completamente aleatoria** *(Missing At Random - MAR)*: cuando la probabilidad de que un individuo responda no depende de la característica de interés, ni de alguna otra covariable auxiliar. Por ejemplo, en una encuesta del mercado de trabajo, podría existir ausencia de respuesta que no dependa del estado actual de ocupación del respondiente, ni de su edad, ni de la UPM en donde está la vivienda, ni de su género, ni de su nivel educativo, ni de ninguna otra característica auxiliar. De esta forma, se puede considerar que esta ausencia de respuesta está dispersa de manera uniforme sobre toda la población. Es decir que cuando el investigador produzca estadísticas descriptivas sobre las personas que respondieron la encuestas, ese procentaje de personas sea muy similar y tenga un comportamiento uniforme sobre todas las posibles covariables que afecten al individuo. El siguiente gráfico podría mostrar algunos indicios de que el patrón de ausencia de respuesta podría ser MCAR puesto que el porcentaje de respuesta es similar en las variables auxiliares.

- **Aleatoria dependiente de covariables** *(Missing At Random - MAR)*: cuan-

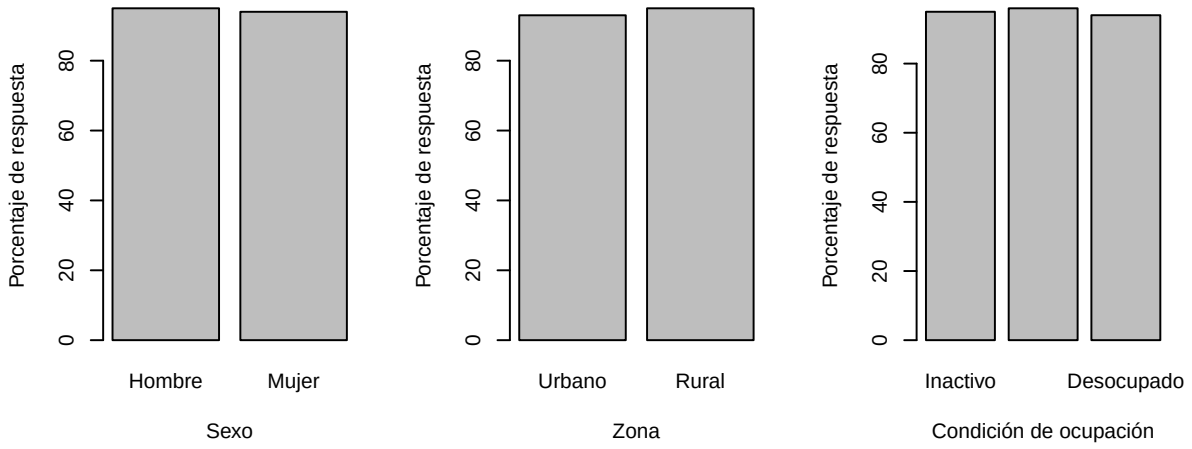

Figura 6.1: Patrón de respuesta MCAR

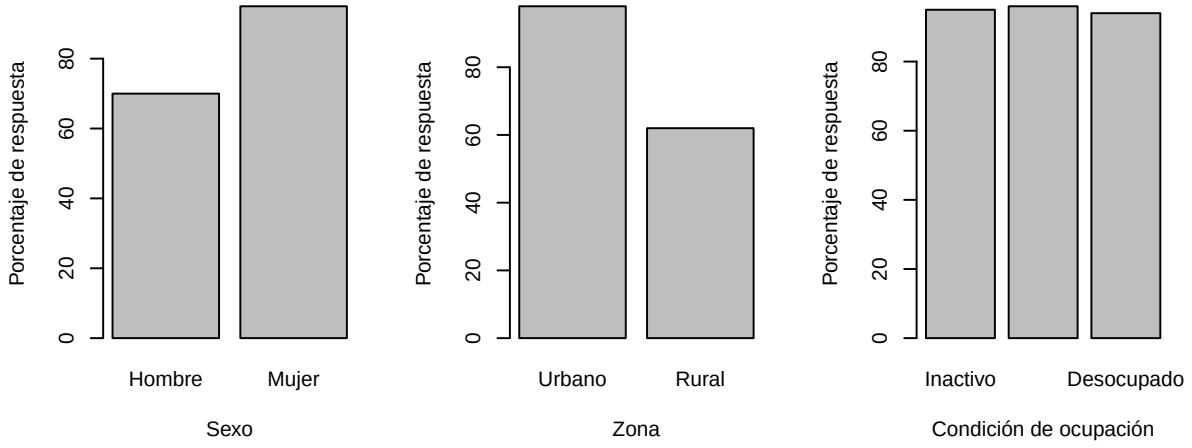

Figura 6.2: Patrón de respuesta MAR

do la probabilidad de que un individuo responda depende de algunas covariables auxiliares, pero no depende de la característica de interés. Por ejemplo, en una encuesta de fuerza laboral, la ausencia de respuesta puede depender de la edad del respondientes, o del sexo, o incluso del nivel económico del individuo, pero no depende de su clasificación laboral. El siguiente gráfico muestra que el patrón de ausencia de respuesta podría ser MAR puesto que el sexo y la zona del respondiente están influenciando el porcentaje de respuesta, aunque no el estado de ocupación.

- **No aleatoria** *(Missing Not At Random - MNAR)*: cuando la ausencia de respuesta depende de la característica de interés. Por ejemplo, en la encuesta de fuerza laboral, es posible que los no respondientes dependan de su clasificación laboral. En este caso puede suceder que los desocupados sean los que sistemáticamente no respondan la encuesta. El siguiente gráfico muestra indicios de que el patrón

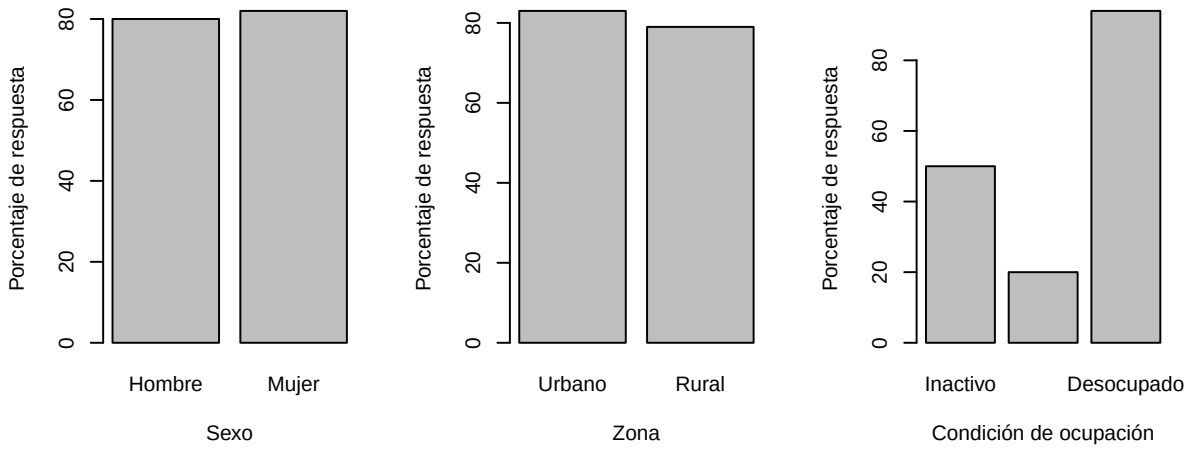

Figura 6.3: Patrón de respuesta MNAR

de respuesta es MNAR puesto que la condición de ocupación es la que influencia el porcentaje de respuesta.

Nótese que a pesar de que se hayan tomado las medidas de ajuste necesarias en el diseño de la encuesta, cuando ya ha terminado el proceso de recolleción de infomración, se debe lidiar con la ausencia de respuesta para evitar sesgos y aumentar la precisión de la encuesta. La literatura especializada examina dos metodologías diferentes pero complementarias en el ejercicio de una encuesta: la prevención de la ausencia de respuesta (antes de que ocurra) y las técnicas de estimación necesarias para tener en cuenta la ausencia de respuesta de manera apropiada en el proceso de inferencia, después de la recolección de los datos. Si el mecanismo de ausencia de respuesta se asume MCAR, es posible contemplar en el proceso de inferencia únicamente a aquellas unidades que tienen registros completos y eliminar de la base de datos a aquellas unidades que no contestaron (*list-wise deletion*). A pesar de que este tipo de análisis es simple, para evitar subestimaciones de los parámetros de interés, se debe realizar un ajuste de los factores de expansión inducidos por el diseño muestral, que originalmente fue planeado con un tamaño de muestra más grande que el efectivo. De esta forma, es posible suponer que la muestra de respondientes corresponde a una submuestra completamente aleatoria de la población y utilizar los principios de los diseños en dos fases. Heeringa, West y Berglund (2010, capítulo 11) afirman que este tipo de análisis, además de inducir posibles sesgos si el supuesto MCAR no se cumple, reduce la eficiencia de la inferencia debido al decremento del tamaño de muestra efectivo. Por lo tanto, en la mayoría de encuestas, este supuesto no se asume y se realiza un ajuste adicional después de que ha ocurrido la ausencia de respuesta. Es por esto que Särndal, Swensson y Wretman (2003, sección 15.5) afirman que las principales técnicas para tratar la ausencia de respuesta son el ajuste a los pesos de muestreo y la imputación. El ajuste por ponderación implica aumentar los pesos aplicados en la estimación de los valores y de los encuestados para compensar los valores que se pierden debido a la ausencia de respuesta, mientras que

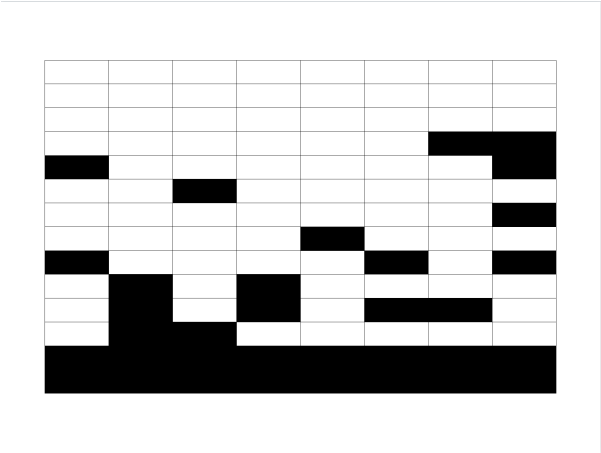

Figura 6.4: *Un conjunto de datos después del proceso de observación.*

la imputación implica la sustitución de los valores faltantes por valores artificiales.

Siguiendo la notación en Särndal y Lundström (2006), consideramos una muestra seleccionada s de unidades de interés; como resultado del proceso de observación se denota a r como el conjunto de unidades que respondieron a uno o más de los I variables de interés. Una unidad que no respondió a ninguna de las variables del estudio pertenecerá entonces al conjunto sr. Así mismo, el conjunto de unidades que respondieron a una variable particular es denotado por r_i. Observe que, en general, todos los r_i tendrán tamaños diferentes y además estarán contenidos en r; de esta forma, se tiene que:

$$r_i \subseteq r \subseteq s$$

La siguiente figura ilustra cómo, después de la recolección de datos, hay individuos que no respondieron a una o todas las variables de la encuesta. En esta ilustración, las unidades están representadas por las filas y las variables por las columnas. Observe que lo primeros tres individuos contestaron a todas las preguntas del cuestionario; el cuarto individuo no contestó las últimas dos preguntas; el quinto individuo no contestó ni la primera ni la última pregunta; el sexto individuo no contesto a la tercera pregunta; y así sucesivamente, hasta llegar a los últimos dos individuos quienes no contestaron a ninguna pregunta del cuestionario. Para este ejemplo particular, se observa que $I = 8$, $n = \#(s) = 14$, $\#(r_1) = 10$, $\#(r_1) = 10$, $\#(r_2) = 9$, $\#(r_3) = 10$, $\#(r_4) = 10$, $\#(r_5) = 11$, $\#(r_6) = 10$, $\#(r_7) = 10$, $\#(r_8) = 8$ y $\#(r_8) = 12$.

Al lidiar con la ausencia de respuesta podemos distinguir varios enfoques para la estimación. El más extremo se conoce como el **enfoque de eliminación**, en donde se conservan para el análisis únicamente aquellas unidades que han respondido a todas y cada una de las pregutnas del cuestionario; es decir que aqueelas unidades que no respondierona al menos una pregunta serán eliminadas del análisis. Note que en este enfoque sólo las unidades del conjunto r se consideran para el análisis posterior. Por

Figura 6.5: *Enfoque de eliminación: todas las unidades que no pertencen a r con eliminadas.*

supuesto, en general, esto no es aconsejable puesto que trae problemas de sesgo, dado que las unidades que contestaron todo el cuestionario generalmente difieren de forma estructural de las unidades que no contestaron; además trae 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. La siguiente figura muestra que, teneindo en cuenta el ejemplo anterior, solo tres unidades serían tenidas en cuenta para el análisis de la información, mientras que nueve unidades, que no contestaron al menos una pregunta, más las tres unidades que no contestaron ninguna pregunta, serían eliminadas del análisis estadístico. Es decir, la mayoría de unidades de la muestra incial serían descartadas.

El segundo enfoque se denomina **imputación total**, y se trata de imputar todos los valores faltantes del conjunto de observaciones. Es decir, los valores de las personas que no respondieron al menos a una variable en todo el cuestionario serán imputados. En este enfoque, la imputación se utiliza para tratar la ausencia de respuesta del ítem y la ausencia de respuesta de la unidad al mismo tiempo. La siguiente figura muestra un ejemplo de las unidades que serían consideradas para el análisis después de la imputación. Nótese entonces que las tres unidades que respondieron todas las preguntas del cuestionario entran al análisis sin nigun ajuste; mientras que las nueve unidades que no respondieron a todo el cuestionario entran al análisis habiendose imputado las celdas correspondientes a la ausencia de respuesta; además, las dos unidades que no respondieron ninguna pregunta del cuestionario también entran al análisis puesto que todas sus respuestas fueron imputadas. Luego, en este enfoque todas las unidades en el conjunto s se consideran para el análisis posterior.

Observe que si y_k es una observación faltante y es imputada, entonces $\hat{y}_k$ denotará el valor imputado obtenido por cualquiera de los métodos de imputación que se describirán más adelante. Luego, como en este enfoque se imputan valores para todas las

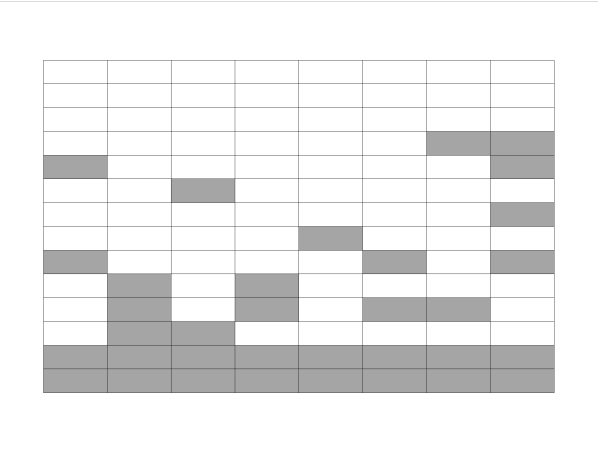

Figura 6.6: *Imputation total: todas las unidades que no están en $s - r$ son imputadas (las celdas en gris indican los valores que fueron imputados).*

observaciones faltantes, ya sea por unidad o por no respuesta del ítem, se tendrá un conjunto de datos completos con los valores

$$y_{\circ\ k} = \begin{cases} y_k, \text{ for } k \in r_i \\ \hat{y}_k, \text{ for } k \in s - r_i \end{cases}$$

El tercer enfoque se denomina **ponderación total** y se utiliza para cada variable de interés en el estudio, una a la vez. En este enfoque nunca se utiliza la imputación, puesto que existirán tantos conjuntos de pesos y ponderaciones como variables con valores faltantes. Para este esquema, se utilizan pesos $w_k^{(i)}$ para cada variable $i \in I$ que compensan la ausencia de respuesta de la unidad. Si todos los r_i son diferentes, cada variable de estudio requirá un peso diferente. Siguiendo con los ejemplos de las ilustraciones, se nota que la primera variable del cuestionario fue respondida por 10 personas, y cuatro personas no respondieron esta pregunta. Por lo tanto, en este enfoque se crearán pesos $w_k^{(1)}$ para cada $k \in s$ que ponderen satisfactoriamente la información recolectada en esta variable. Sin emabrgo, este conjunto de pesos no será único, puesto que, en particular, la segunda variable del cuestionario fue respondida por nueve personas, y tres personas no respondieron esta pregunta. Por lo tanto, en este enfoque se crearán pesos $w_k^{(2)}$ para cada $k \in s$ que ponderen esta información recolectada en esta variable. Nótese que 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.

Por último, el enfoque recomendado es una combinación de los procediemintos de imputación y ponderación y se conoce como **enfoque combinado** que imputa los valores de las celdas ausentes para los individuos que tienen al menos un valor faltante, exceptuando a aquellos que tienen todos los valores faltantes. De esta forma, los individuos que no contestaron ninguna pregunta del cuestionario son eliminados del análisis, mientras

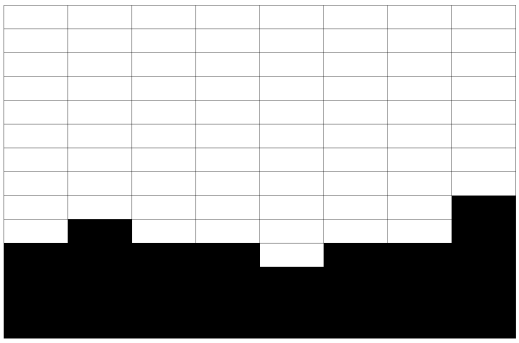

Figura 6.7: *Ponderación total: cada variable tendrá un conjunto de pesos diferente. No se utiliza ningún método de imputación.*

que los restantes serán considerado en el análisis con sus respuestas originales o con la imputación de las celdas vacías. En resumen, este enfoque implica que la imputación se utilice para tratar sólo la ausencia de respuesta del ítem, mientras que la ponderación se utilice únicamente para tratar la ausencia de respuesta del individuo. Retomando el ejemplo, el siguiente gráfico permite observar que los dos últimos individuos de la muestra fueron totalmente descartados puesto que no contestaron ninguna pregunta del cuestionario; además, para la primera variable, los valores del quinto y noveno individuo fueron imputados. De la misma manera, para la segunda variable, los valores de lso individuos 10, 11 y 12, fueron imputados; y así sucesivamente, hasta llegar a la última variable en donde los valores de los indivioduos tres, cuatro, seis y ocho fueron imputados.

En la realidad de las encuestas de hogares, lo más común es encontrar que se presente ausencia de respuesta tanto de unidad como de ítem. En estos casos, se aconseja imputar primero y luego descartar los individuos que no contestaron a niguna pregunta, de esta forma al final se obtendrá una matriz de datos rectangular completa. Por tanto, el conjunto de datos completados (originales o imputados) para la variable y es $\{y_{\circ\ k} : k \in r\}$, en donde

$$y_{\circ\ k} = \begin{cases} y_k, & \text{for } k \in r_i \\ \hat{y}_k, & \text{for } k \in r - r_i \end{cases}$$

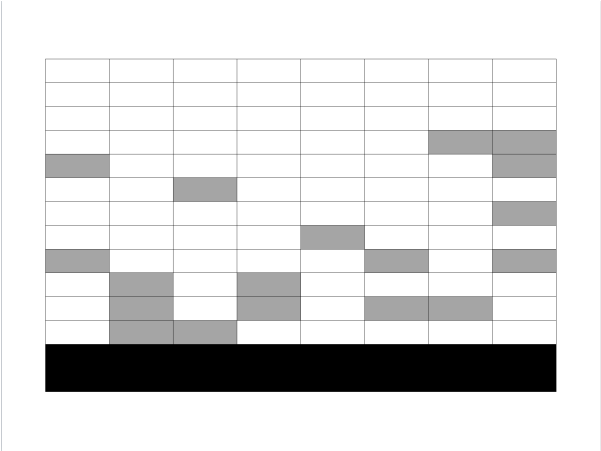

Figura 6.8: *Enfoque combinado: las unidades que no respondieron a ningún ítem son eliminadas del análisis y los respondientes parciales son imputados.*

III Imputación

El término imputación se refiere al conjunto de técnicas por las cuales los valores faltantes en una o más variables se reemplazan con infrmación plausible con el objetivo de lograr valores sustitutivos en una base de datos que pueda ser analizada posteriormente. Este proceso introduce un nuevo elemento de error, conocido como el error de imputación, debido a la incertidumbre que introducen los valores no observados. Cuando se tiene ausencia de respuesta por ítem, las técnicas de imputación se prefieren antes que la utilización de los esquemas 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 un individuo contengan información y no exista ningún vacío. Para lograr la sutitución de los valores faltantes con infomración plausible, es posible encontrar donantes apropiados, en la misma muestra que se ha conseguido, definidos como respondientes que comparten características demográficas similares con el individuo que no respondió. Por lo tanto, la información del respondiente donante (o una función de estos valores) se copiará en las celdas vacías del no respondiente. Para encontrar los donantes es posible realizar un análisis estadístico con base en métodos de clasificación. Dentro de los métodos de imputación más usados en encuestas de hogares se encuentran los siguientes:

- Imputación promedio (*mean value imputation*) que utiliza la media de la variable (dentro de las UPM o en un subconjunto apropiado de datos). En este caso, si se encuentra un valor faltante, inmediantamente será reemplazado por el promedio de los datos de los respondientes en esta variable.
- Imputación por paquete caliente (*hot deck imputation*) que reemplaza los valores faltantes por los valores de un donante que es un respondiente de la encuesta

en el mismo levantamiento. En este caso, el valor faltante es reemplazado por la información del individuo escogido de antemano.
- Imputación por paquete frío (*cold deck imputation*) que reemplaza los valores faltantes por los valores de un donante que es un respondiente de encuesta en un levantamiento anterior. En este caso, el valor faltante es reemplazado por la información auxiliar de un individuo escogido de encuestas anteriores.
- Imputación estadística que se apoya en un modelo (de regresión, generalmente) en donde la variable dependiente es aquella que se quiere imputar y las covariables se derivan del restante conjunto de datos. En este caso, el valor faltante es reemplazado por la predicción (o una función) del modelo ajustado con la información en la muestra.

Como se mencionó anteriormente, cuando se trata de imputación, se pueden definir dos tipos de métodos. La imputación de la unidad completa, que se produce cuando toda la información de un individuo es imputada, y la imputación del ítem, que se da cuando un único valor de un individuo es imputado. Observe que la imputación de la unidad se utiliza para hacerle frente a la ausencia de respuesta de la unidad, cuando no hay datos para el individuo, mientras que la imputación del ítem se utiliza para la no respuesta del ítem, cuando no todos los valores se proporcionan para un individuo, pero algunos sí.

La imputación se realiza a menudo en grupos no traslapados $g = 1, \ldots, G$, donde la unión de $s_1, \ldots, s_G$ equivale a la muestra completa s. Se pueden utilizar diferentes métodos para cada grupo, pero dentro de cada grupo se debe utilizar el mismo método de imputación. Esto se debe a que pueden existir diferentes covariables disponibles para cada grupo. Cuando la disponibilidad de las variables auxiliares (covariables) es limitada, es posible considerar una jerarquía de métodos de imputación. De esta forma, para los grupos con más información disponible, es posible utilizar métodos más sofisticados de imputación; mientras que para grupos con menos información auxiliar disponbible, se deben usar métodos de imputación más simples. Särndal y Lundström (2006) presentan una discusión acerca del uso de esta técnica en combinación con los estimadores utilizados en las encuestas de hogares que proveen estadísticas oficiales. A continuación, se presenta una compilación no exhaustiva de algunos de los principales métodos de imputación que se utilizan en las encuestas de hogares.

Imputación por regresión

En este método determinístico, el valor imputado para el valor faltate y_k se calcula utilizando una regresión lineal.

$$\hat{y}_k = \mathbf{x}_k \hat{\boldsymbol{\beta}}_i$$

Donde,

$$\hat{\boldsymbol{\beta}}_i = (\sum_{r_i} a_k \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_{r_i} a_k \mathbf{x}_k y_k$$

El vector de coeficientes de regresión $\hat{\boldsymbol{\beta}}_i$ se produce a partir de un ajuste de regresión múltiple utilizando los datos $(y_k, \mathbf{x}_k)$ disponibles para cada unidad $k \in r_i$ con pesos a_k especificados adecuadamente. Nótese que, en general, las predicciones del modelo de regresión no necesariamente serán valores observados en algún otro individuo de la muestra. Por lo tanto, este método inducirá valores imputados que no han sido observados en la encuesta. Además, se deberán generar tantos modelos de regresión como variables con valores faltantes existan.

Imputación de razón

Un caso especial del anterior método se da cuando solo se tiene acceso a una sola covariable (positiva) $\mathbf{x}_k = x_k$, y definiendo $a_k = \frac{1}{x_k}$. En este caso, la estimación del coeficiente de regresión será

$$\hat{\beta}_i = \frac{\sum_{r_i} y_k}{\sum_{r_i} x_k} = R_i$$

Y por 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, entonces el coeficiente utilizado para la imputación es la relación entre los dos puntos en el tiempo.

Imputación de 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 respondientes. Nótese que no se requiere de ninguna información adicional en este método.

El vecino más cercano

Si asumimos que valores similares de x producirán valores similares de y, podemos "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

$$\hat{y}_k = y_{l(k)}$$

Dónde $l(k)$ es el "elemento donante", determinado al minimizar una ecuación de distancia. Enn el caso más simple, para una sola covariable de imputación x_k, la distancia entre los posibles donantes l a la unidad k es:

$$D_{lk} = |x_k - x_l|$$

El donante l al elemento k es aquel individuo en el conjunto r con la menor distancia D_{lk} entre todos los posibles elementos $l \in r$. Para el caso en donde se contemple más de una covariable de imputación, es posible considerar la siguiente distancia

$$D_{lk} = \left(\sum_{j=1}^{J} h_j (x_{jk} - x_{jl})^2 \right)$$

En donde h_j se utiliza para ponderar adecuadamente cada una de las J covariables de la matriz de imputación.

Imputación Hot Deck

La imputación por regresión y el vécino más cercano son métodos que asumen una fuerte relación entre la variable de interés y y las covariables $\mathbf{x}$. Sin embargo, en algunas aplicaciones esta relación no se puede establecer fácilmente, y no es plausible 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, ya que no se cuenta con información auxiliar para determinar un buen sustituto.

Imputación múltiple

Cuando existe información auxiliar que permita relacionar las covariables con la variable de interés, es posible establecer mejores modelos que no solo mantienen el insesgamiento de la inferencia, sino que estiman con bastante precisión el error de muestreo. Con respecto a esta última categoría de imputación, es posible completar el conjunto de datos utilizando información auxiliar de los respondientes en la encuesta (o encuestas anteriores, si se trata de un diseño rotativo) y la información disponible a nivel de la población para predecir los valores faltantes usando 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). Suponga que este modelo es de la forma

$$y_k = f(\mathbf{x_k}, \boldsymbol{\beta}) + \varepsilon_k$$

En donde ε_k es un término de error aleatorio. Una vez formulado el modelo, y debido a la naturaleza estocástica de ε_k, es posible generar $Q > 1$ realizaciones de la variable de interés para los registros faltantes; esto se logra de manera muy sencilla, simulando Q valores del término de error. De esta forma, se generan Q conjuntos de datos completos. Para cada conjunto de datos, se generarán Q estimaciones de interés que luego se promedian para obtener una estimación puntual.

A Ejemplo: imputación en una encuesta de ingresos y gastos

Una vez que se ha discutido acerca de los propósitos de la imputación en una encuesta de hogares, se debe escoger un método (o métodos) de imputación y una vez establecido el mecanismo de imputación, generar el conjunto de datos rectangular y completo. En esta sección analizaremos, a la luz de las particularidades de una encuesta de hogares de ingresos y gastos, los pasos que se deben surtir para completar un proceso de imputación. Por sus características, este tipo de encuestas presenta tasas elevadas de ausencia de respuesta de ítem, aunque también de individuo.

En general, el levantamiento común de este tipo de encuestas se centra en un trabajo de campo masivo que visita al hogar en varias ocasiones, pidiéndole al respondiente que diligencie sendos cuestionarios, y registre toda la información asociada al gasto y a los ingresos del hogar, durante un periodo 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 habrá visitado varias veces el domicilio del hogar en el periodo de observación y tendrá un formulario totalmente diligenciado. Sin embargo, en muchas otras ocasiones, a pesar del seguimiento exhaustivo del encuestados, no se obtendrá el gasto de la totalidad de las categorías de la encuesta, sino que se obtendrá infomración parcial que se transformará en celdas

vacías por la ausencia de respuesta. En el peor de los casos se obtendrán cuestionarios diligenciados en porcentaje tan bajo, que al final serán declarados como faltantes, lo cual se transforma en ausencia de respuesta de ese hogar.

El siguiente ejemplo trata de ilustrar de manera escueta cómo se debería realizar el procedimiento de imputación en una encuesta de ingresos y gastos. El lector encontrará varios pasos en esta metodología, puesto que antes de imputar las variables de interés, es ncesario conocer qué covariables se relacionan fuertemente con las variables que se quieren imputar. Además de eso, es necesario primero imputar todas las covariables en primer lugar y reemplazar sus valores faltantes con información plausible que pueda ser utilziada en los modelos que se ajusten. Suponga que, para el conjunto de hogares que se consideró con fines de imputación, se observaron al menos las siguientes variables:

- Tamaño del hogar.
- Número de hombres y mujeres dentro del hogar.
- Número de niños y adultos en el hogar.
- Edad del jefe de hogar.
- Estado de ocupación del jefe de hogar.
- Grado educativo más alto del jefe de hogar.
- Número de personas empleadas en el hogar.

El camino que se seguirá en este ejemplo será primero la imputación de los ingresos, como principal covariable del gasto y del consumo. Una vez que se imputaron las covariables, el segundo paso de este proceso se relaciona con la imputación de los filtros, que son las preguntas que se realizan para conocer si un hogar ha adquirido un bien o servicio específico. El resultado de este paso produjo el tercer paso dedicado a la imputación de los valores de gasto anualizados en cada unidad. Esta serie de pasos metodológicos ha sido recomendados por diferentes agencias de estadística, incluyendo institutos y oficinas nacinoales de estadística. Por ejemplo, Hayes y Watson (2009) y Sun (2010) siguen esta metodología en el *Australian Bureau of Statistics* para imputación en la encuesta *Household, Income and Labour Dynamics in Australia (HILDA)*

Imputación del ingreso

En primer lugar, debe ser notado que tanto teórica como empíricamente, los ingresos han demostrado ser un potente predictor de los gastos (Starick y Watson 2011). Si la base de datos contiene hogares que reportaron un ingreso nulo en todo el año, es posible que esos valores se consideren como faltantes porque se asume que los hogares no deben tener ingresos nulos durante todo un año. Además, los hogares con ingresos superiores a un límite también pueden ser considerados como valores atípicos y luego ser imputados.

La imputación del ingreso está basada 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 pronosticado que se computa para las unidades que faltan. Así, para cada unidad con información faltante en el ingreso, se identifica un solo donante que corresponderá al hogar cuyo ingreso disponible es más cercano a la predicción del modelo de regresión. Por ende, todos los componentes de los ingresos son imputados por el donante. El modelo lineal se describe como se indica a continuación y la predicción de los ingresos para los hogares faltantes se calcula utilizando una regresión lineal.

$$\tilde{y}_k = \mathbf{x}_k \hat{\boldsymbol{\beta}}_i$$

Donde, $\tilde{y}_k$ es el valor pronosticado del ingreso disponible para el hogar k, $\mathbf{x}_k$ es el vector de las covariables del modelo, y los coeficientes de regresión estimados están dados por:

$$\hat{\boldsymbol{\beta}}_i = (\sum_{r_i} a_k \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_{r_i} a_k \mathbf{x}_k y_k$$

Este vector de coeficiente de regresión $\hat{\boldsymbol{\beta}}_i$ se produce a partir de un ajuste de regresión múltiple utilizando los datos $(y_k, \mathbf{x}_k)$ disponibles para cada unidad $k \in r_i$ con pesos a_k especificados adecuadamente. De aquí en adelante, esta covariable fue imputada a nivel de hogar y la información necesaria (incluida en $\mathbf{x}_k$) para hacerlo se resume de la siguiente manera:

- **Composición del hogar**: número de adultos, número de niños, número de hombres, número de mujeres, edad adulta media, edad media de los niños, edad de la persona más joven, edad de la persona mayor, edad del jefe de hogar, grado educativo más alto del jefe de hogar.
- **Ocupación y fuerza de trabajo**: situación laboral del jefe de hogar, número de personas empleadas, número de desempleados en el hogar.
- **Calidad de la vivienda**: creada creado a partir de la sección de calidad de la vivienda, incluye por ejemplo, un índice de hacinamiento (como la relación entre 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 en el hogar.
- **Ubicación del hogar**: municipalidad y provincia, como primera y segunda desagregación cartográfica del país.

Asumiendo que valores similares de las predicciones del modelo lineal $\tilde{y}$ producirán valores similares en las observaciones del ingreso y, podríamos pedir prestado un valor real de ingreso y para imputar el valor faltante con la infmración de este vecino que tiene valores similares en las predicciones $\tilde{y}$ del modelo lineal. Así, el valor imputado para la unidad k es dada por

$$\hat{y}_k = y_{l(k)}$$

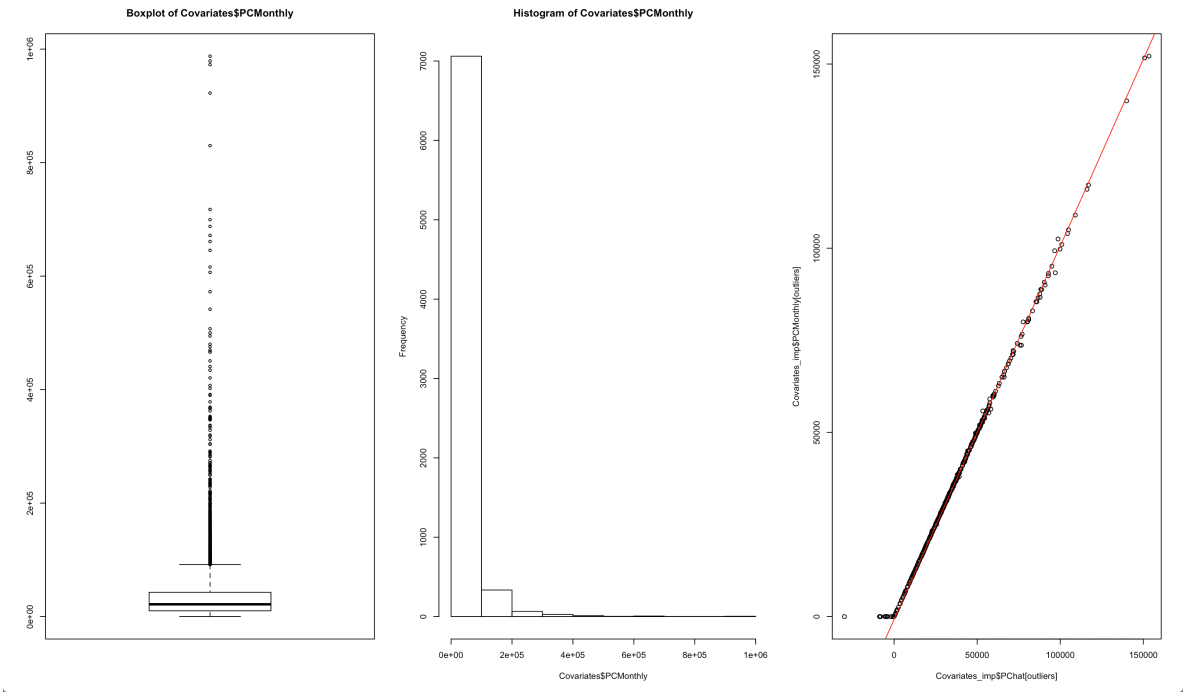

Figura 6.9: *Distribución de los ingresos (izquierda y centro) y Relación entre los valores predichos e imputados para los hogares con datos de ingresos faltantes (derecha).*

Donde $l(k)$ es el "elemento donante", determinado por minimizar una medida simple de distancia entre todos posibles donantes l a la unidad k. Esta distancia está dada por:

$$D_{lk} = |\tilde{y}_k - y_l|$$

El donante l al elemento k será aquel hogar en el conjunto r_i con la menor distancia D_{lk} entre todos los posibles hogares $l \in r_i$. Como regla general, todo los donantes deben estar ubicados en la misma provincia que la unidad faltante. La siguiente figura muestra un diagrame de caja junto con el histograma de los ingresos (antes de la imputación), así como la relación lineal entre los valores pronosticados derivados del modelo y los valores imputados tomados de los donantes.

Imputación del filtro

El siguiente paso, luego de haber logrado imputar con éxito las covariables determinantes del gasto es precisamente utilizarlas para lograr imputar el gasto en bienes o servicios. Por lo general, las encuestas de ingresos y gastos preguntan si el hogar consumió o adquirió cierto bien o servicio específico. En caso de responder afirmativamente, se pregunta por la cantidad de dinero gastado en el bien o servicio y por la cantidad de

artículos adquiridos en el periodo de referencia; en caso de responder negativamente, se procede a preguntar por el siguiente bien o servicio. Por supuesto, diferentes artículos tiene diferentes tasas de respuesta en sus filtros. De aquí en adelante, el valor a ser imputado en esta etapa es dicotóminco: sí o no. Si el valor imputado hubiera sido no, eso significaría que el hogar no debería tener ningún gasto asociado a ese ítem. Debido a la naturaleza del filtro, un modelo de regresión logística es conveniente para modelar la ausencia de respuesta en el filtro. De esta manera, la probabilidad de consumo (o compra)a un artículo i en particular es $p_k = Pr(Filter_i = 1)$ y puede ser estimada por medio del siguiente modelo:

$$\tilde{p}_k = logit^{-1}(\mathbf{x}_k\hat{\boldsymbol{\beta}}_i) = \frac{exp(\mathbf{x}_k\hat{\boldsymbol{\beta}}_i)}{1 + exp(\mathbf{x}_k\hat{\boldsymbol{\beta}}_i)}$$

Las covariables incluidas en la matriz $\mathbf{x}$ podrían ser las mismas utilizados 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 ocupación y fuerza de trabajo de los miembros del hogar, la calidad de la vivienda, la ubicación del hogar y los ingresos del hogar. Asumiendo que los similares valores de $\tilde{p}$ producirán valores de filtro similares, podemos "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 es dada por

$$\hat{y}_k = y_{l(k)}$$

Donde $l(k)$ es el "elemento donante", determinado por la minimización de una distancia simple entre los posibles donantes l a la unidad k, dada por:

$$D_{lk} = |\tilde{p}_k - p_l|$$

Nótese que el donante l al elemento k es el elemento en el conjunto r_i con el valor más pequeño de la distancia D_{lk} entre todos los posibles $l \in r_i$. Por regla general, todo los donantes deben estar en la misma provincia que la unidad con el valor faltante. Por ejemplo, considere el artículo arroz, para el cual algunos hogares no proveyeron ninguna respuesta asociada al filtro de compra. Como este es un artículo de consumo masivo en nuestra región, se supondría que la mayoría de hogares respondiera que efectivamente ha comprado arroz en el periodo de refernecia. De esta manera, al utilizar la regresión logística como modelo para la ausencia de respuesta del filtro del arroz, es posible encontrar que la distribución de las probabilidades estimadas de compra de arroz esté sesgada hacia el valor uno y alejada del valor cero, como lo muestra la siguiente figura. Está claro que la distribución de estas los valores imputados también debería estar cargada hacia el uno, reflejando la realidad de la compra de un artículo esencial como el arroz.

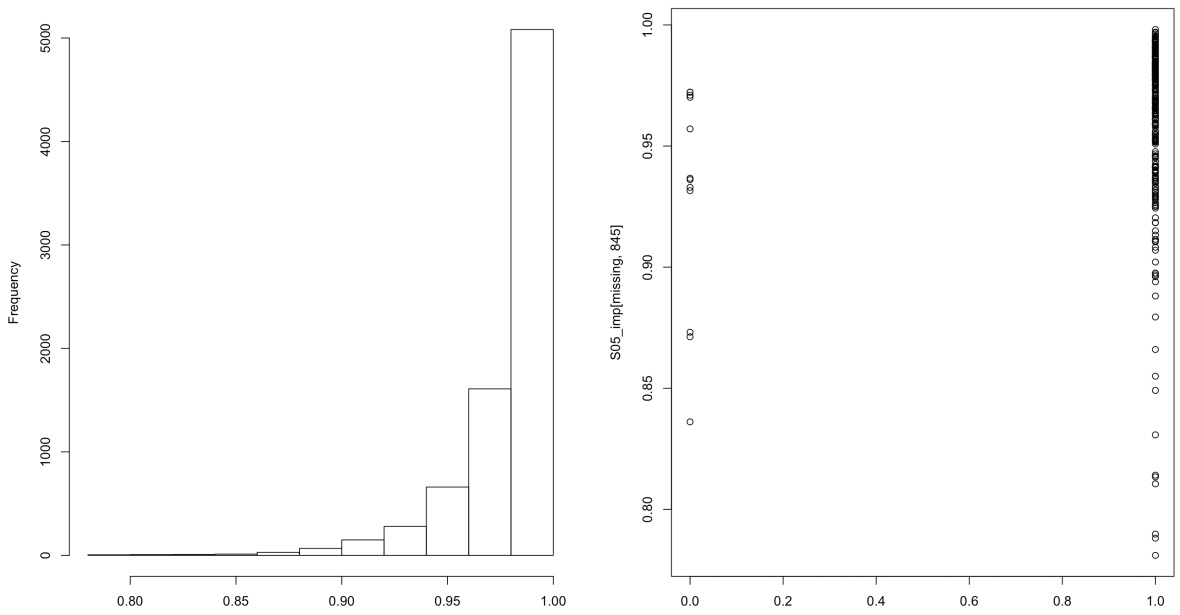

Figura 6.10: *Distribución de las probabilidades estimadas de compra de arroz (izquierda) y valores imputados para los hogares con valores faltantes en el filtro (derecha).*

Por otro lado, el filtro para algunos artículos de bajo consumo estará más sesgado hacia el valor cero. La siguiente figura muestra la distribución de las probabilidades estimadas de compra de un artículo de bajo consumo, así como los valores imputados.

Imputación del gasto

Éste es el paso final del proceso de imputación y está fuertemente influenciado por los resultados de la imputación de la pregunta de filtro. En este paso, los hogares cuyo valor imputado de filtro es cero automáticamente tendrá un cero imputado como la cantidad de dinero gastado en ese artículo. Es decir, si el resultado de la imputación en el filtro es cero, esto implica directamente que el hogar no compró (o produjo) el artículo en el periodo de referencia, y por tanto la frecuencia de compra, la cantidad que ítems comprados y la cantidad de dinero gastado en ese artículo debe ser cero. Las unidades restantes deben tener un valor observado o imputado de uno en el filtro, y por lo tanto los valores faltantes del gasto deben ser imputados.

Observe que el grupo de donantes está restringido a los que tienen un valor de gasto distinto de cero en el artículo específico. Es decir, para aquellas unidades con un valor de filtro distinto de cero, un donante debe ser identificado. Para la imputación del gasto, la técnica del vecino más cercano con el método de regresión puede considerarse en el mismo sentido que fue implementado en la imputación de los ingresos. Por lo tanto, se considera un modelo lineal en donde las covariables incluidas en la matriz $\mathbf{x}$ son

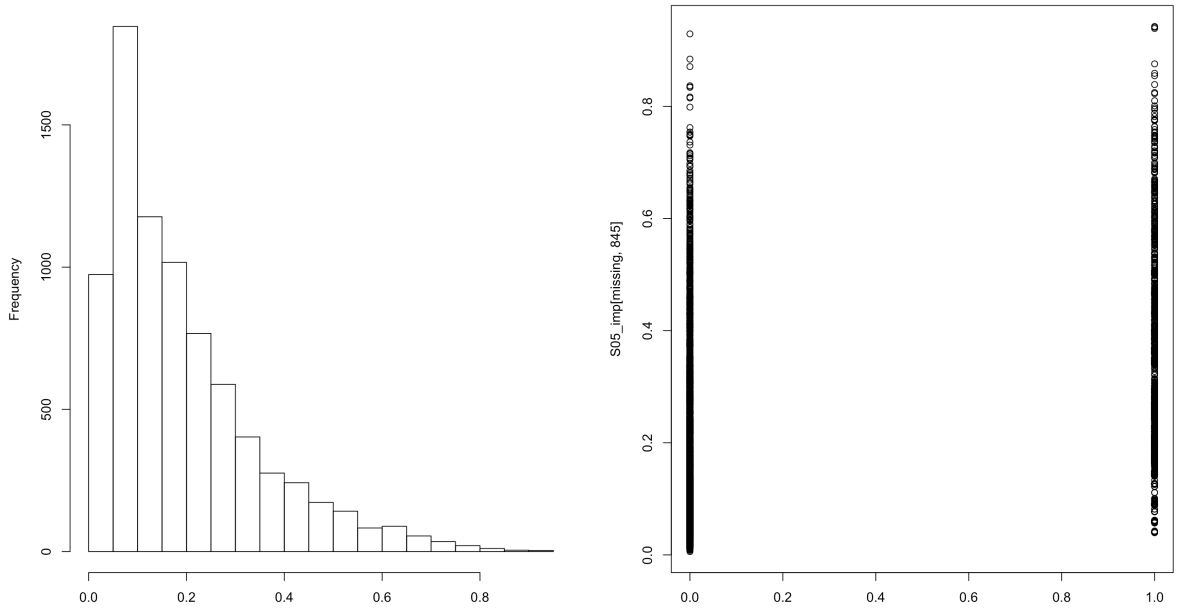

Figura 6.11: *Distribución de la probabilidad estimada de compra de un artículo de bajo consumo (izquierda) y sus valores imputados para los hogares que no respondieron el filtro (derecha).*

la composición del hogar, el estado de ocupación y fuerza de trabajo, la calidad de la vivienda, la ubicación del hogar y los ingresos.

Volviendo a los ejemplos anteriores,la siguiente figura muestra la distribución de los gastos imputados sobre el arroz. Se nota que la cantidad de dinero gastado en este artículo es baja y que la relación entre los valores pronosticados del modelo y los valores imputados es fuertemente lineal.

IV Reponderación de los pesos de muestreo

Los procesos de inferencia estadistica establecidos en cualquier encuesta de hogares descansan sobre el principio de representatividad que afirma 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 encuestas de hogares descansan sobre las probabilidades de inclusión generadas por el diseño de muestreo que se implementó en la encuesta. En general el peso de muestreo w_k asociado a un individuo k en la muestra s es una función del inverso de la probabilidad de inclusión del individuo, así

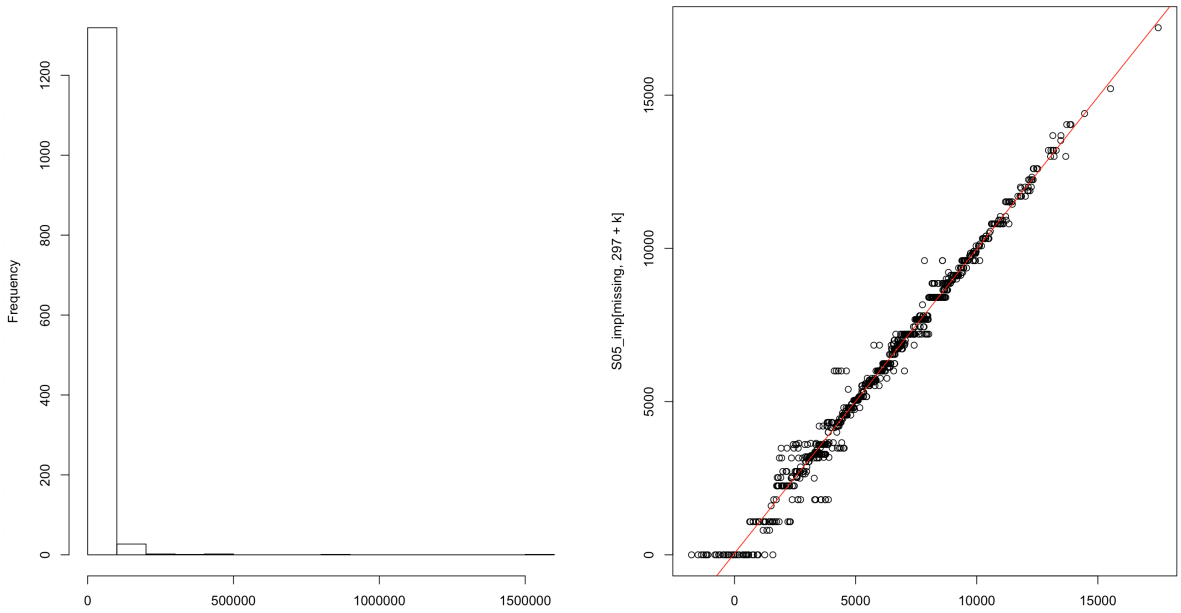

Figura 6.12: *Distribución de los gastos imputados sobre el arroz (izquierda) y relación entre los valores predichos e imputados para los hogares con valores faltantes en el gasto (derecha).*

$$w_k \propto \frac{1}{Pr(k \in s)}$$

Como se mencionó anteriormente, para conservar estabilidad en los pesos de muestreo, es posible definir diseños de muestreo auto-ponderados, en donde las unidades finales de muestreo tengan las misma probabilidad de inclusión, sin importar el tamaño de la unidad primaria de muestreo que la contiene. Este tipo de diseños es útil porque induce mayor control sobre las estimaciones finales. Es de notar que la conformación de los pesos de muestreo se transforma en un reto metodológico para el investigador, puesto que debe ajustarse a la realidad de la región en donde 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. Varias soluciones a este problema han sido planteadas (Gambino y Silva 2009) y todas ellas requieren de esfuerzos económicos, logísticos y técnicos. Por ende, los equipos de los 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 emabargo, en general, no es posible contar con una lista de todos los elementos de la población y, en el contexto de las encuestas a hogares, no existe una lista que enumere todos los hogares de un país de manera actualizada, por lo que la práctica estándar es construir el marco de muestreo en varias etapas, seleccionando

una muestra de áreas geográficas, realizando un empadronamiento exhaustivo de todos los hogares en las áreas seleccionadas y luego seleccionando hogares. Este esquema de muestreo hace que el marco de muestreo de las encuestas a hogares presente imperfecciones. El siguiente gráfico, adaptado de Valliant y Dever (2017), describe los problemas inferenciales que se deben surtir al trabajar con marcos de muestreo imperfectos y su relación con los pesos de muestreo originales.

XXXXXX incluir gráfico acá XXXXXXXXX

Para hacerle frente a las imperfeccciones del marco, la AsociacEl esquema de ponderación *American Association for Public Opinion Research* (AAPOR) recomienda tratar la ausencia de respuesta de manera diferenciada y clasificar a cada unidad en la muestra en algunas de las siguientes categorías:

1. ER (*unidades elegibles que fueron respondientes efectivos*): casos elegibles para los cuales se ha recolectado una cantidad suficiente de información.
2. ENR (*unidades eligibles no respondientes*): casos elegibles para los cuales no se recolectó ningún dato o la información fue parcialmente recolectada.
3. IN (*unidades no elegibles*): casos de miembros no elegibles que no hacen parte de la población de interés.
4. UNK (*unidades con elegibilidad desconocida*): casos en donde no se puede conocer si la unidad es elegible o no.

Para construir los factores de expansión de una encuesta se recomienda seguir en este orden los siguientes procesos:

1. Creación de los pesos básicos.
2. Descarte de las unidades no elegibles.
3. Ajuste por elegibilidad desconocida.
4. Ajuste por ausencia de respuesta.
5. Calibración por proyecciones poblacionales y variables auxiliares.
6. Preparación de la base de datos de respondientes.

Creación de los pesos básicos

Este primer paso ya ha sido explicado de forma detallada en la sección dedicada a la selección de la muestra. Observe que, asociado a cada esquema particular de muestreo, existe una única función que víncula a 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

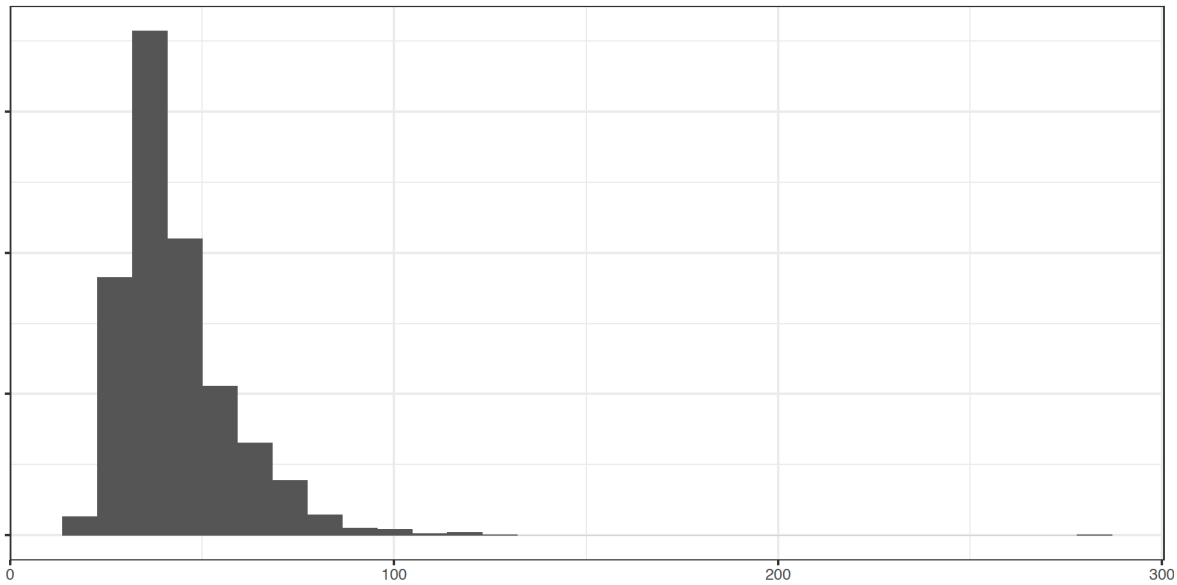

Figura 6.13: *Distribución de los pesos básicos de muestreo en una encuesta de hogares.*

probabilidad de inclusión

$$d_{1k} = \frac{1}{\pi_k}$$

Estos pesos son creados incluso para aquellas unidades que serán excluidas de la muestra porque son no elegibles o porque no proveyeron ninguna información y luego serán modificados convenientemente. La siguiente figura muestra la distribución típica de los pesos originales en una encuesta de hogares. A través de las modificaciones posteriores, esta distribución irá sufriendo algunos cambios. Si la distribución original de los pesos básicos difiere estructuralmente con la distribución final de los ponderadores, resultante de todos los ajustes debidos a las imperfecciones del marco, entonces las propiedades estadísticas de insesgamiento, consistencia y precisión podrían desvanecerse. Lo anterior implica que el nivel de desactaulización del marco de muestreo tiene implicaciones directas en la calidad de la inferencia. Por tanto, si el marco de meustreo es muy imperfecto, los ponderadores finales no inducirán una inferecnia precisa.

Descarte de las unidades no elegibles

Si hay viviendas seleccionadas desde el marco de muestreo que han cambiado su estado de ocupación y ahora no contienen ningún hogar, entonces el segundo paso consiste en ajustar su peso básico de la siguiente manera:

$$d_{2k} = \begin{cases} 0, & \text{si la unidad } k \text{ no pertenece a la población objetivo} \\ d_{1k}, & \text{en otro caso} \end{cases}$$

Ajuste por elegibilidad desconocida

El tercer paso consiste en redistribuir el peso de las unidades cuyo estado de elegibilidad es desconocido. Por ejemplo, si la encuesta está enfocada en la población mayor de 15 años y hay personas que no proveen ninguna información acerca de su edad, entonces es necesario distribuir estos pesos. Esta situación también se puede presentar a nivel de hogar cuando no puede ser contactado porque nadie nunca atendió el llamado del encuestador (*nadie en casa*). Se acostumbra a redistribuir los pesos de los UNK entre las unidades que sí disponen de su estatus de elegibilidad (ER, ENR, IN).

Luego, 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 conjunto de las unidades *elegibles* en la muestra s_e, el conjunto de las unidades *no elegibles* en la muestra s_n y el conjunto de las unidades con *elegibilidad desconocidad* s_u. En este último caso, la elegibilidad de estos casos es 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 imputar su estado de elegibilidad.

Se recomienda formar B ($b = 1, \ldots, B$) categorías[2] basadas en la información del marco de muestreo. Estas categorías pueden ser estratos o cruces de subpoblaciones. Siendo s_b la muestra de unidades en la categoría b (que incluye a ER, ENR y UNK), se define el factor de ajuste por elegibilidad como:

$$a_b = \frac{\sum_{s_b} d_{2k}}{\sum_{s_b \cap s_e} d_{2k}}$$

Para la categoría b, los pesos ajustados por elegibilidad desconocida para aquellas unidades cuya elegibilidad si pudo ser establecida (independientemente de su estado de respuesta) estarán dados por la siguiente expresión:

$$d_{3k} = a_b * d_{2k}$$

Ajuste por ausencia de respuesta

En este paso los pesos básicos de los ER se ajustan para tener en cuenta a los ENR. Al final del proceso, los pesos de los ER se incrementan para compensar el hecho de

[2]Se acostumbra a formar categorías con al menos 50 casos.

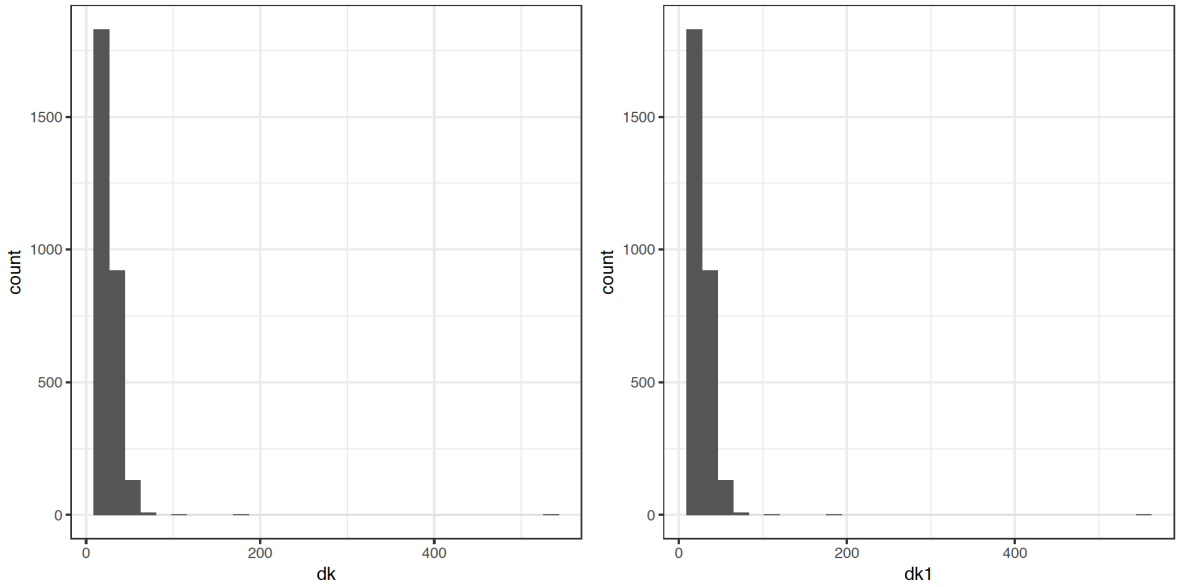

Figura 6.14: *Comparación de la distribución de los pesos básicos de muestreo (izquierda) con los pesos ajustados por el estado de elegibilidad (derecha) en una encuesta de hogares.*

que algunas unidades elegibles no proveyeron información. Para el manejo efectivo de la ausencia de respuesta se consideran las siguientes variables aleatorias:

$$I_k = \begin{cases} 1, & \text{si } k \text{ pertenece a la muestra } s \\ 0, & \text{en otro caso.} \end{cases}$$

$$D_k = \begin{cases} 1, & \text{si } k \text{ pertenece al conjunto de respondientes } s_r \\ 0, & \text{en otro caso.} \end{cases}$$

Al suponer que la distribución de las respuestas puede ser estimada, entonces la probabilidad de respuesta (*propensity score*) está dada por

$$Pr(k \in s_r | k \in s) = Pr(D_k = 1 | I_k = 1) = \phi_k$$

Si el patrón de ausencia de respuesta es completamente aleatorio (en donde la no respuesta no sigue ningún patrón específico) o aleatorio (en donde el patrón de la no respuesta puede ser explicado por covariables $\mathbf{x}$ de la encuesta), entonces

$$\phi_k = f(\mathbf{x}_k, \boldsymbol{\beta})$$

Ahora, si es posible tener acceso a las covariables $\mathbf{x}$, entonces es posible estimar el patrón de ausencia de respuesta mediante

$$\hat{\phi}_k = f(\mathbf{x}_k, \hat{\boldsymbol{\beta}})$$

Por otro lado, si el patrón de ausencia de respuesta es no aleatorio (en donde el patrón de la no respuesta es explicado por la variable de interés; por ejemplo cuando los desempleados son los que no responden), entonces

$$\phi_k = f(\mathbf{y}_k, \beta)$$

Luego, como no es posible tener acceso a la variables de interés para todos los individuos en la muestra (porque no todos respondieron), entonces no es posible estimar el patrón de ausencia de respuesta y por ende hay sesgo. Bajo los dos primeros escenarios, es posible definir el siguiente estimador insesgado

$$\hat{t}_y = \sum_{k \in s_r} d_{4k} y_k$$

En donde

$$d_{4k} = \frac{d_{3k}}{\hat{\phi}_k}$$

Nótese que el sesgo se anula puesto que

$$E(I_k D_k) = EE(I_k D_k | I_k) = E(I_k) E(D_k | I_k) = \pi_k \phi_k$$

XXXXXXXX

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 genera la ausencia de respuesta en la encuesta de hogares es MAR, es posible ajustar un modelo para la ausencia de respuesta (en donde la variable dependiente es una variable indicadora de la respuesta del individuo por lo general supeditado a una distribución Bernoulli o Binomial). Kim y Riddles (2012) muestran que es posible utilizar un modelo basado en el *propensity score* de las respuestas. Si la muestra de los respondientes se denota como s_r entonces la probabilidad de que un individuo conteste es $\phi_k = Pr(k \in s_r)$. Al suponer que existe un vector de información auxiliar $\mathbf{z}_k$ conocido para todo $k \in s$ es posible estimarla por medio de un modelo de regresión logística; esto es,

$$\hat{\phi}_k = \frac{\exp\{\mathbf{z}_k' \hat{\boldsymbol{\beta}}\}}{1 + \exp\{\mathbf{z}_k' \hat{\boldsymbol{\beta}}\}}$$

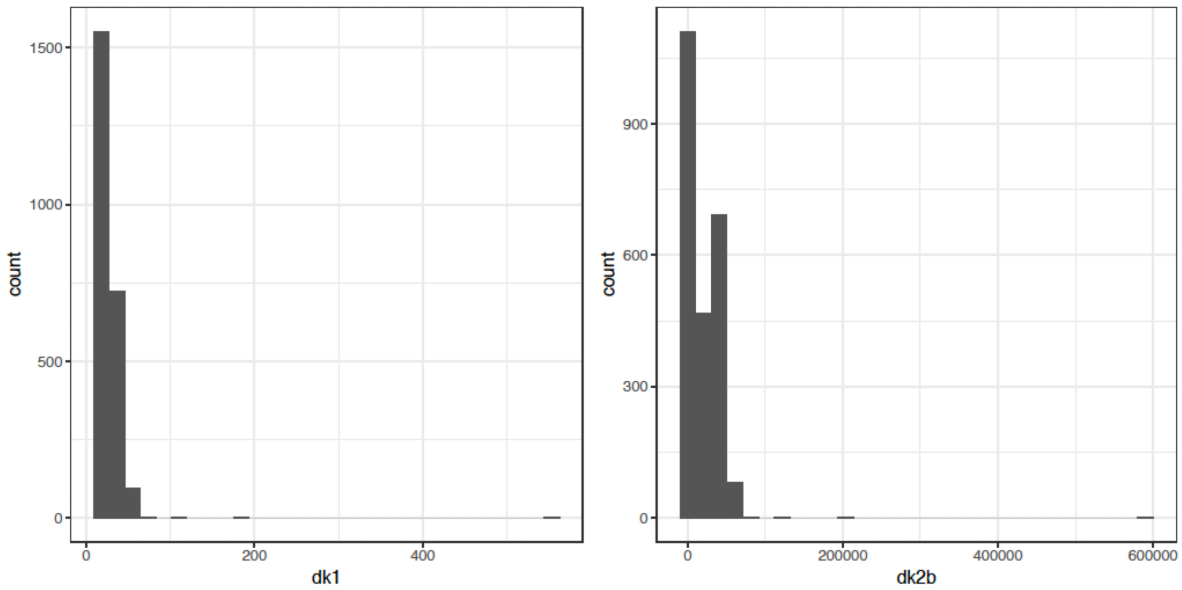

Figura 6.15: *Comparación de la distribución de los pesos básicos de muestreo (izquierda) con los pesos ajustados por ausencia de respuesta (derecha) en una encuesta de hogares.*

donde $\hat{\beta}$ es el vector de coeficientes estimado de la regresión logística. Finalmente, el nuevo peso estimador para un total poblacional, con el ajuste debido a la ausencia de respuesta no ignorable, queda expresado como

$$\hat{t}_y^{(adj)} = \sum_{k \in s_r} \frac{w_k}{\hat{\phi}_k} y_k$$

XXXXXX

Calibración de los pesos

Después de conformar el sistema de ponderación de pesos de muestreo en la encuesta, es posible calibrar estos pesos con la información auxiliar disponible para cada país, a nivel nacional, por estratos de interés, e incluso por variable continuas sobre las que se tenga interés. Särndal y Lundström (2006) afirman que cuando los estudios por muestreo están afectados por la ausencia de respuesta, es deseable tener las siguientes propiedades en la estructura inferencial que sustenta el muestreo:

1. Sesgo pequeño o nulo.
2. Errores estándares pequeños.
3. Un sistema de ponderación que reproduzca la información auxiliar disponible[3].

[3]Por ejemplo, el número de hogares o habitantes en el país.

4. Un sistema de ponderación que sea eficiente al momento de estimar cualquier característica de interés en un estudio multipropósito.

A pesar de que cada vez es más extendido el uso de los pesos calibrados en las encuestas de América Latina, es necesario que se disemine más esta práctica metodológica que induce estimadores muestrales que reproducen exactamente la información auxiliar conocida a nivel poblacional. Debido a la construcción teórica de los estimadores de calibración, los pesos finales responden a la siguiente restricción

$$\sum_s w_k^* \boldsymbol{x}_k = \sum_U \boldsymbol{x}_k = \boldsymbol{t}_{\boldsymbol{x}}$$

El ejemplo más básico se encuentra cuando se desea que los pesos de muestreo deberían reproducir con exactitud el tamaño de las regiones N_h y el tamaño del país N. Es así como, 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 lo estratos explícitos

$$\sum_{s_h} w_k^* = N_h$$

Esta coherencia entre las cifras oficiales y las que la encuesta puede producir hace que sea preferible el uso de los estimadores de calibración. Las anteriores características son satisfechas al usar el enfoque de calibración que induce una estructura inferencial robusta en presencia de información disponible puesto que reduce tanto el error de muestreo como el error debido a la ausencia de respuesta. Los estimadores de calibración son **aproximandamente insesgados**, pero la magnitud del sesgo está dada por la siguiente expresión:

$$Bias(\hat{t}_{y,cal}) = E_p \left[\sum_{k \in s} (w_k - d_k) y_k \right]$$

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 (inducidas por todos los cruces y celdas considerados) sea menor del 10% para asegurar que el sesgo de los estimadores de calibración sea 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 deba ajustar los pesos originales. Esto induce sesgo en cada subgrupo ajustado. Si aún así se decide optar por múltiples aumentar las variables de calibración, es necesario hacer un chequeo empírico

del ajuste que cada modelo pueda tener con todas las variables de la encuesta, aunque se advierte que este chequeo a veces puede ser demorado e ir en contravía de las apretadas agendas de producción estadísticas que se manejan en el INE.

Silva (2004) presenta algunas consideraciones al respecto del sesgo que puede generarse al usar esta metodología en las encuestas de hogares y aborda algunos criterios para evaluar la calidad de la calibración.

La idea general del proceso de calibración es encontrar un número de restricciones parsimonioso que permita tener estimaciones aproximadamente insesgadas con una varianza menor a la generada con los factores de expansión originales. En general los INE podrán clasificar sus procesos de calibración en una de las siguientes tres categorías:

1. Calibración con variables continuas, que es el caso en donde la calibración se realiza con los totales de variables continuas como ingreso, gasto, entre otras.
2. Post-estratificación con variables categóricas, que representa el caso en donde la calibración se realiza con los tamaños poblacionales (basados en proyecciones demográficas o registros administrativos) de subgrupos de interés.
3. *Raking* con variables categóricas, que se define como una calibración sobre los tamaños marginales de tablas de contingencia de subgrupos de interés. A diferencia del caso anterior, esta calibración no tiene en cuenta los tamaños de los cruces, sino solo los tamaños marginales; por ende, este método induce menos restricciones.

En un encuesta de hogares las restricciones de calibración pueden establecerse sobre características de hogares y características de personas al mismo tiempo. De esta forma, por ejemplo, es posible calibrar sobre las proyecciones demográficas de personas y al mismo tiempo controlar las estimaciones del número de hogares en el país de manera conjunta. Estevao y Särndal (2006) discuten una amplia variedad de casos en donde se calibra conjuntamente en distintos niveles de desagregación sobre diferentes esquema de muestreo. Por ejemplo, para la *Encuesta Continua de Empleo* de Bolivia la calibración está inducida por una post-estratificación sobre los tamaños poblacionales de los cruces resultantes entre las variable Departamento (hay 9 departamentos), Zona (rural y urbano) y PET (con dos categorías: mayor o igual a 10 años y menor de 10 años).

El asesor regional dio una charla acerca de la metodología de calibración al equipo del Departamento de Investigación y Desarrollo, en donde se presentaron algunas medidas de calidad como protección en contra del sesgo generado por considerar demasiadas restricciones y se resaltó la importancia de que las variables utilizadas para la calibración sean estimadas de manera precisa por los estimadores clásicos de muestreo. Por ejemplo, si el número de personas por hogar es utilizada como una variable de calibración (utilizando como total auxiliar las proyecciones demográficas), entonces el coeficiente de variación del estimador de Horvitz-Thompson sobre esta variable debería ser menor al 10%.

La teoría afirma que entre más variables de calibración se tengan menor será la varianza

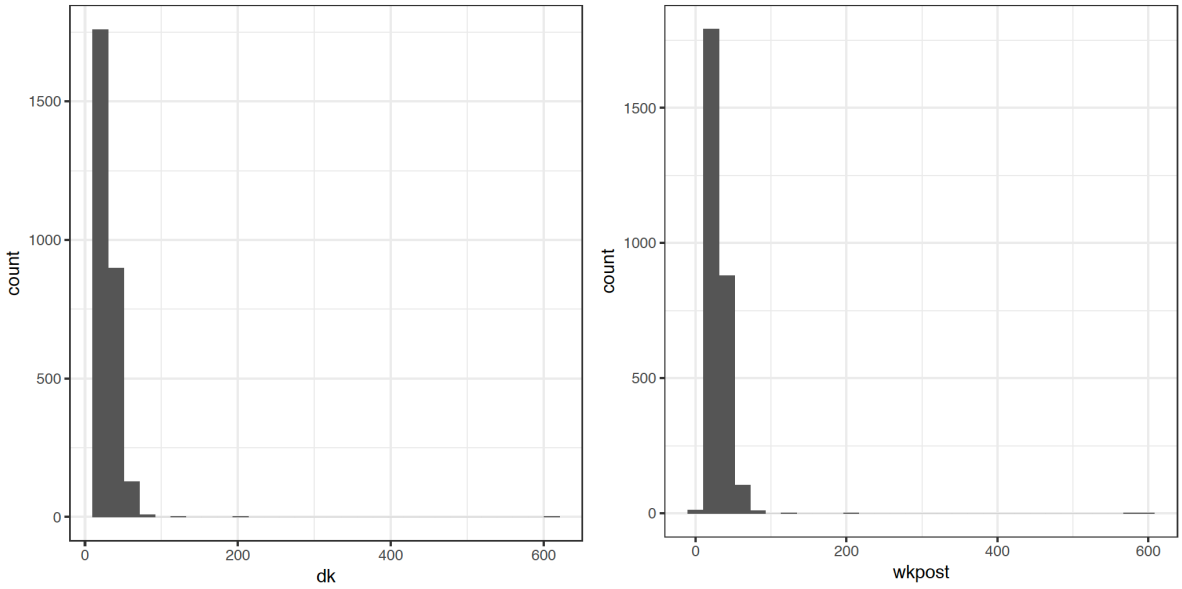

Figura 6.16: *Comparación de la distribución de los pesos básicos de muestreo (izquierda) con los pesos ajustados por ausencia de respuesta (derecha) en una encuesta de hogares.*

asociada a las estimaciones (no así el sesgo). Sin embargo, existen problemas computacionales cuando crecen las restricciones que se deben satisfacer son demasiadas. Una primera opción es verificar que no se tengan variables que puedan tener codependencia lineal con otras. Al descartar estas variables es posible conservar una varianza pequeña puesto que se descartan combinaciones lineales de otras variables.

Si los pesos de calibración resultan ser menores que uno su interpretación puede tornarse difícil (aunque no reviste un problema teórico). El usuario común entiende al factor de expansión como un factor de representatividad: *es la cantidad de veces que una persona se represnta a sí misma y a algunas otras más 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 a estimaciones negativas para algunos dominios en donde el tamaño de muestra es pequeño, lo cual resulta ser problemático en un contexto en donde todas las variables de estudio son no negativas.

Para garantizar que los pesos se ubiquen en un intervalo determinado, se debe minimizar una distancia que a su vez debe inducir pesos restringidos a este intervalo y que respete 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. Nótese que los estimadores de calibración se pueden escribir como

$$w_k = g_k * d_k$$

Con base en lo anterior, es necesario analizar los pesos g_k en perspectiva en cada dominio,

estrato y postestrato de interés. Una buena idea puede ser identificar aquellos g_k que resulten potencialmente grandes o influyentes. Se recomienda postestratificar la muestra, y aplicar la calibración a aquellas unidades en los que los g_k sean estables y usar los pesos originales en el restante conjunto.

Es posible hacer 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 $w_k \geq 1$ implica $g_k \geq 1$ y por tanto $L = 1$. Se acostumbra a tomar $U > Q_3 + 1.5 * (Q_3 - Q_1)$ en 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 ausencia de respuesta no es aleatorio (MAR) o completamente aleatorio (MCAR), es posible que los ponderadores de calibración induzcan sesgo en las estimaciones finales. En general, cuando hay ausencia de respuesta es más probable que aparezcan pesos de calibración negativos y que los pesos de calibración no convergieran a los pesos originales. Además, la varianza de los estimadores de calibración no convergerá a los resultados usuales de los estimadores de regresión.

Silva (2004) presenta algunas medidas que permiten decidir cuáles escenarios de calibración son los mejores. A continuación se citan tales medidas.

- Error relativo promedio sobre las variables auxiliares

$$M1 = \frac{1}{p} \sum_{j=1}^{p} \frac{|\hat{t}_{xc} - t_x|}{t_x}$$

- Coeficiente de variación HT relativo promedio

$$M2 = \frac{1}{p} \sum_{j=1}^{p} \frac{(Var(\hat{t}_{x\pi}))^{1/2}}{t_x}$$

- Proporción de pesos extremos (límite inferior)

$$M3 = \frac{1}{n} \sum_{k \in s} I(g_k < L)$$

- Proporción de pesos extremos (límite superior)

$$M4 = \frac{1}{n} \sum_{k \in s} I(g_k > U)$$

- Coeficiente de variación de los g_k

$$M5 = \frac{\sigma(g)}{\bar{g}}$$

- Distancia entre los pesos de calibración y los pesos originales

$$M6 = \frac{1}{n}\sum_{k\in s}\frac{(w_k - d_k)^2}{d_k} = \frac{1}{n}\sum_{k\in s} d_k(g_k - 1)^2$$

- Eficiencia de los estimadores de calibración sobre la estimación del diseño de muestreo

$$M7 = \frac{1}{J}\sum_{j=1}^{J}\frac{Var(\hat{t}_{y_j c})}{Var(\hat{t}_{y_j \pi})}$$

- Efecto de la calibración sobre la dispersión de los ponderadores (DEFFk)

$$M8 = 1 + \frac{\sigma_w^2}{\bar{w}^2}$$

Calibración de personas con bases de datos de hogares

Una de las preguntas recurrentes en la calibración de encuestas de hogares es el nivel al cual se debería realizar este ajuste. En prinicpio, es posible realizar la calibración al nivel de las personas, o al nivel de los hogares. Cada una de estas opciones trae algunas ventajas y consideraciones que se deben tener en cuenta.

Calibrar al nivel de los hogares implica que el hogar tendrá unos nuevos pesos que cumplen con las restricciones de calibración, y esos pesos los heredará a las personas que habitan el hogar. De esta forma todas las personas pertenecientes a un mismo hogar tendrán el mismo peso de muestreo, sin importar sus diferencias en composición demográfica. Es decir que, hombres, mujeres, menores y mayores de 15 años tendrán el mismo peso de muestreo. Esta propiedad es atractiva puesto que emula el diseño de muestreo que se definió en la fase de planeación. Por otro lado, realizar la calibración a nivel de los hogares hace que dentro de las unidades primarias de muestreo (UPMs) los hogares no tengan un peso homogeneo, lo que se distancia de las propiedades del diseño sistemático simple que se usa para la selección de los hogares dentro de las UPMs.

Por otro lado, calibrar a nivel de personas implica que los pesos de muestreo de los hogares también pueden verse alterados, y que los pesos finales de muestreo de las personas sean diferentes dentro de los hogares. De esta forma, de acuerdo a las características de las personas se tendrá un peso diferente. Es decir que hombres, mujeres, menores y mayores de 15 años **no** tendrían el mismo peso de muestreo. Por consiguiente cuando se calibra por personas y se utiliza un filtro sobre esa base de personas para crear una base de hogares, las características observadas de los jefes de hogares influenciarían los pesos d emuestreo resultantes en dicha base de datos de hogares.

En principio, se supone que se selecciona una muestra de unidades de una población finita $U = \{1, 2, \ldots, k, \ldots, M\}$ la cual está agrupada en conglomerados $U_I =$

$\{1, 2, \ldots, i, \ldots, N_I\}$. El proceso de selección se puede resumir de la siguiente manera (Gutiérrez 2016a):

- Se selecciona una muestra S_I de conglomerados de U_I con probabilidades de inclusión π_{I_i} para $i \in U_I$. Los pesos muestrales de la primera etapa son $d_{I_i} = \frac{1}{\pi_{I_i}}$
- En la segunda etapa se seleccionan unidades para $i \in S_I$. Se enumeran las unidades en U_i y se selecciona una muestra S_i con probabilidades de inclusión condicionales $\pi_{k|i}$.
- De esta forma, $d_{k|i} = \frac{1}{\pi_{k|i}}$ son los pesos condicionales y $d_k = d_{I_i} d_{k|i}$ es el peso de muestreo general para la k-ésima unidad, siendo $s = \bigcup_{i \in s_I} S_i$ la muestra de unidades.

Para calibrar sobre la información auxiliar a nivel de hogar, se deben satisfacer las siguientes ecuaciones:

$$\sum_{i \in S_I} w_{I_i} x_{ci} = \sum_{i \in U_I} x_{ci} = t_{x_c}$$

En donde x_{ci} denota el vector de variables auxiliares a nivel del hogar, que siempre será definido como un conteo de individuos con cierta característica en el hogar. Por ejemplo, en su caso más básico x_{ci} puede ser el número de individuos en el hogar, mientras que el total auxiliar $t_{x_c} = \sum_{i \in U_I} x_{ci}$ será el número de individuos en todos los hogares de Chile Luego de que se ha calibrado la base de hogares, se construyen los pesos a nivel de persona recurriendo a la siguiente expresión:

$$w_k = d_{k|i} w_{I_i} \quad \forall k \in S_i$$

Y además, se preserva la siguiente propiedad

$$\frac{w_k}{w_{I_i}} = d_{k|i}$$

Lo anterior imita la propiedad del muestreo en dos etapas de que

$$\frac{d_k}{d_{I_i}} = d_{k|i}$$

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 $w_k = w_{I_i} \quad \forall k \in S_i$. Por lo anterior, dado que el muestreo es de conglomerados en la última etapa y todos los individuos del hogar son seleccionados, entonces el peso de muestreo del hogar será el promedio de los pesos individuales.

$$w_{I_i} = \sum_{k \in S_i} \frac{w_k}{N_i} = \sum_{k \in U_i} \frac{w_k}{N_i} = \bar{w}_i$$

Por último, nótese que la matriz de calibración toma la siguiente forma:

$$X_{[i]} = X = \begin{bmatrix} x_{c11} & x_{c12} & \cdots & x_{c1p} \\ x_{c21} & x_{c22} & \cdots & x_{c2p} \\ x_{c31} & x_{c32} & \cdots & x_{c3p} \\ \vdots & \vdots & \ddots & \\ x_{cN_I1} & x_{cN_I2} & \cdots & x_{cN_Ip} \end{bmatrix}$$

Por otro lado, la calibracion se puede realizar a nivel de las personas o a nivel de los hogares. Por ejemplo, si la calibración se realiza a nivel de 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, puesto 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 hogares, en 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 principio, y debido al diseño de las encuesta, los pesos de muestreo originales son idénticos para todos los miembros de un mismo hogar. Sin embargo, cuando la calibración trata de ajustar los totales de las restricciones, y debido a que la población no está equitativamente distribuida, entonces de igual manera se presenta un reajuste en los factores de calibración. Por otro lado, calibrar a nivel de hogar hace que se cree un único factor de expansión para el hogar, que será heredado por todos sus miembros. Sin embargo, en este acercamiento y debido a que la composición de los hogares es diferente, entonces dentro de una misma UPM se tendrán hogares con diferentes pesos de muestreo.

En el caso particular en que haya información auxiliar disponible a nivel de personas y hogares (al mismo tiempo), es posible calibrar conjuntamente ambas variables en un sólo procedieminto de calibración. Estevao y Särndal (2006, sec. 5) recrea la calibración conjunta para hogares y personas, en donde se genera una variable indicadora para el jefe de hogar y sobre esta se crea una nueva restricción de calibración que utilice los totales auxiliares de los hogares. Con esta nueva calibración, se generan unos nuevos pesos de calibración en la base de datos de personas. Al filtrar esta base de datos por el jefe de hogar, se crea inmediatamente una base de hogares (puesto que solo hay un único jefe de hogar) que puede ser utilizada para combinarla con la información de los hogares. De esta forma, los pesos que venían de la base de datos de personas serán los que se utilicen en la base de datos de hogares obteniendo estimaciones consistentes.

Capítulo 7

Estimadores y error de muestreo

I Estimadores puntuales

La mayoría de indicadores sociales a nivel nacional pueden verse como funciones de totales de una o más variables de interés. Por ejemplo, si el interés está en estimar un total $t_y = \sum_U y_k$, el estimador de expansión provee una metodología que induce insesgamiento.

$$\hat{t}_y = \sum_s w_k y_k$$

En donde la muestra s hace referencia al subconjunto de la población que fue seleccionado siguiendo un diseño de muestreo probabilístico que induce los pesos de muestreo w_k que expanden el valor de la variable de interés y_k para el k-ésimo individuo. Nótese que w_k es función de la probabilidad de inclusión de k-ésimo individuo en la muestra $\pi_k = Pr(k \in s)$. En presencia de esquemas de estratificación y selección de conglomerados y varias etapas, esta probabilidad resulta ser el producto de las probabilidades condicionales que surgen en los subsecuentes procesos de selección probabilística. Por tanto, el peso final de muestreo resulta ser por lo general una multiplicación de factores de expansión en cada etapa del esquema de muestreo.

Por ejemplo, si el diseño de la encuesta es estratificado por regiones h (agrupaciones de municipios), con tres etapas de selección dentro de cada estrato (la primera etapa con selección de municipios i dentro del estrato, la segunda con selección de segmentos cartográficos j y la última con selección de hogares k), entonces el peso de muestreo final y el estimador del total estará dado por la siguiente expresión

$$\hat{t}_y = \sum_s w_j y_k = \sum_h \sum_{i \in s_h} \sum_{j \in s_{hi}} \sum_{k \in s_{hij}} w_{hijk} y_{hijk}$$

En presencia de información auxiliar es posible mejorar la eficiencia de la estimación acudiendo a diferentes formas funcionales que estiman el total, por ejemplo, con el estimador de Hájek:

$$\hat{t}_y = t_x \frac{\sum_s w_k y_k}{\sum_s w_k x_k}$$

En donde t_x denota el total poblacional, que se supone conocido para toda la población, de una variable auxiliar x que es preguntada 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 $R_U = t_{y_1}/t_{y_2}$ cuya estimación se lleva a cabo estimando ambos componentes de la fracción

$$\hat{R} = \frac{\hat{t}_{y_1}}{\hat{t}_{y_2}} = \frac{\sum_s w_k y_{1k}}{\sum_s w_k y_{2k}}$$

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

$$\hat{\bar{y}}_s = \frac{\hat{t}_y}{\hat{N}} = \frac{\sum_s w_k y_k}{\sum_s w_k}$$

Las encuestas de hogares con diseños panel o rotativos, tienen un mayor interés en la estimación del cambio de indicadores en dos periodos tiempo $\Delta = t_y^{(t)} - t_y^{(t-1)}$. Nótese que un estimador de este parámetro está dado por

$$\hat{\Delta} = \hat{t}_{y^{(t)}} - \hat{t}_{y^{(t-1)}}$$

Además, es posible mejorar la estimación del total actual $t_y^{(t)}$ al tener en cuenta la información inducida por el traslape de la encuesta en el segundo periodo, así:

$$\tilde{t}_{y^{(t)}} = \alpha \hat{t}_{y^{(t)}} + (1-\alpha)(\tilde{t}_{y^{(t-1)}} - \hat{\Delta})$$

En donde $0 < \alpha < 1$. Por otro lado, si el interés está en 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, entonces el siguiente estimador puede ser utilizado

$$\hat{F}_\alpha = \frac{1}{N} \sum_{k \in s} w_k \left(\frac{l - y_k}{l} \right)^\alpha I(y_k < l)$$

En donde $I(y_k < l)$ es una variable indicadora que toma el valor uno si $y_k < l$ o cero, en cualquier otro caso. Note que si $\alpha = 0$, se tiene una estimación de la incidencia de la pobreza y si $\alpha = 1$, se obtiene un estimación de la brecha de la pobreza (Foster, Greer y Thorbecke 1984).

La selección del estimador está altamente relacionada con el diseño de la encuesta. Por ejemplo, si se pretende estimar un indicador para un periodo de tiempo definido, el diseño de la encuesta no debería inducir un esquema de rotación que tenga traslape de hogares, puesto que la correlación del indicador induciría un aumento de su varianza y por ende pérdida de eficiencia. Sin embargo, si se desea estimar el cambio del indicador entre dos periodos de tiempos, es necesario contar con un esquema de rotación que asegure un tamaño de muestra suficiente para estimar con precisión este cambio. Cochran (1977, sección 12.13) afirma que, cuando el interés se centra tanto en la estimación del indicador en el periodo actual como en la estimación del cambio entre periodos, es recomendable tener una tasa de traslape de 2/3, 3/4 o 4/5 de una ronda a otra.

II Estimación del error de muestreo

Aunque la escogencia del diseño de muestreo y el estimador sean de libre elección para los investigadores, no lo es el cálculo de las medidas de confiabilidad y precisión. Dado que la base científica sobre la cual descansa el muestreo es la inferencia estadística, se deben respetar las normas básicas para la asignación y posterior cálculo del margen de error, que constituye una medida unificada del error total de muestreo que cuantifica la incertidumbre acerca de las estimaciones en una encuesta.

A La técnica del último conglomerado

Debido a las dificultades algebricas y computacionales, estimar la varianza en encuestas complejas que contemplan esquemas de conglomeración, selección en varias etapas y estratifcación, puede tornarse bastante tedioso, costoso y además muy demorado. En este documento se explica por qué la técnica del último conglomerado resulta ser una buena opción a la hora de aproximar la varianza en una encuesta compleja.

Para la estimación de la varianza de los estimadores de interés en encuestas multietápicas, los programas computacionales existentes utilizan una aproximación conocida como la técnica del último conglomerado. Esta aproximación, que sólo tiene en cuenta la varianza de los estimadores en la primera etapa, supone que ese muestreo fue realizado con reemplazo. Los procedimientos de muestreo en etapas posteriores de la selección son ignorados a menos que el factor de corrección para poblaciones finitas sea importante a nivel municipal.

En particular considere cualquier estimador del total poblacional dado por la siguiente combinación lineal

$$\hat{t}_y = \sum_{k \in s} d_k y_k = \sum_{k \in U} I_k d_k y_k \tag{7.1}$$

En donde I_k son variables indicadoras de la pertenecia del elemento k a la muestra s. Ahora, asumamos que el factor de expansión de la encuesta d_k cumple con los supuestos básicos de un ponderador que hace insesgado a $\hat{t}_y$, es decir:

$$E_p(I_k d_k) = 1$$

Se supone un diseño de muestreo en varias etapas (dos o más) en donde la primera etapa supone la selección de una muestra s_I de m_I unidades primarias de muestreo (UPM) U_i $(i \in s_I)$ de tal forma que

- Si la selección se realizó con reeemplazo, la i-ésima UPM tiene probabhilidad de selección p_{I_i}.
- Si la selección se realizó sin reeemplazo, 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 pertenzca a la muestra dada que la UPM que la contiene ha sido seleccionada en la muestra de la primera etapa está dada por la siguiente expresión:

$$\pi_{k|i} = Pr(k \in s_i | i \in s_I)$$

Por ejemplo, si el muestreo es sin reemplazo en todas sus etapas, la probabilidad de inclusión del k-ésimo elemento a la muestra s está dada por

$$\begin{aligned} \pi_k &= Pr(k \in s) \\ &= Pr(k \in s_i, i \in s_I) \\ &= Pr(k \in s_i | i \in s_I) Pr(i \in s_I) = \pi_{k|i} \times \pi_{I_i} \end{aligned}$$

Dado que el inverso de las probabilidades de inclusión son un ponderador natural, entonces se definen las siguientes cantidades:

1. $d_{I_i} = \frac{1}{\pi_I}$, 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 para 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.

Resultado: *Bajo un diseño de muestreo en varias etapas, el estimador de Hansen-Hurwitz para el total poblacional está dada por:*

$$\hat{t}_{y,p} = \frac{1}{m_I} \sum_{i=1}^{m_I} \frac{\hat{t}_{y_i}}{p_{I_i}} \tag{7.2}$$

Y su varianza estimada es:

$$\widehat{Var}(\hat{t}_{y,p}) = \frac{1}{m_I(m_I - 1)} \sum_{i=1}^{m_I} \left(\frac{\hat{t}_{y_i}}{p_{I_i}} - \hat{t}_{y,p} \right)^2 \tag{7.3}$$

Supongamos ahora que la encuesta tiene un diseño complejo $p(s)$ que no contempla reeemplazo en la primera etapa. Por lo tanto, algunas cantidades deben ser equiparadas para poder utilizar esta aproximación. En principio, nótese que las cantidades $\hat{t}_{y_i}$ representan lo totales estimados de la variable de intereés en la i-ésima UPM y están dados por:

$$\hat{t}_{y_i} = \sum_{k \in s_i} \frac{y_k}{\pi_{k|i}} = \sum_{k \in s_i} d_{k|i} y_k \tag{7.4}$$

Utilizar la aproximación de la varianza requiere equiparar los términos de manera apropiada. En primer lugar, fijémonos en los estimadores dados por (9.2) y (9.1). Para realizar esta comparación, se requiere que se asuma la siguiente igualdad en las probabilidades de inclusión de la primera etapa:

$$\pi_{I_i} = p_{I_i} \times m_I \tag{7.5}$$

Por lo tanto, el estimador del total poblacional quedaría definido desde (9.1) como un estimador tipo Hanwen-Hurwitz.

$$\hat{t}_y = \sum_{k \in s} d_k y_k = \sum_{i=1}^{m_I} \sum_{k \in s_i} d_k y_k = \sum_{i=1}^{m_I} \sum_{k \in s_i} \frac{1}{\pi_{I_i} \pi_{k|i}} y_k = \sum_{i=1}^{m_I} \frac{\hat{t}_{y_i}}{\pi_{I_i}} \approx \frac{1}{m_I} \sum_{i=1}^{m_I} \frac{\hat{t}_{y_i}}{p_{I_i}}$$

Ahora, dado que la forma del estimador ha sido equiparada con un estimador tipo Hanwen-Hurwitz, es posible utilizar su estimación de varianza. Aún más, después de un poco de álgebra y tuilizando la equiparación dada por (9.5), es posible tener la siguiente

aproximación, cuya gran ventaja es que sólo hace uso de los factores de expansión finales d_k, que suelen ser reportados por los Institutos Nacionales de Estadística cuando liberan los microdatos de sus encuestas, en vez de los factores de expansión de la primera etapa o los factores de expansión condicionales dentro de las UPM.

$$\begin{aligned}
\widehat{Var}(\hat{t}_{y,p}) &= \frac{1}{m_I(m_I-1)}\sum_{i=1}^{m_I}\left(\frac{\hat{t}_{y_i}}{p_{I_i}} - \hat{t}_y\right)^2 \\
&= \frac{m_I}{m_I-1}\sum_{i=1}^{m_I}\frac{1}{m_I^2}\left(\frac{\sum_{k\in s_i} d_{k|i}y_k}{p_{I_i}} - \sum_{i=1}^{m_I}\sum_{k\in s_i} d_k y_k\right)^2 \\
&= \frac{m_I}{m_I-1}\sum_{i=1}^{m_I}\left(\frac{\sum_{k\in s_i} d_{k|i}y_k}{m_I p_{I_i}} - \frac{1}{m_I}\sum_{i=1}^{m_I}\sum_{k\in s_i} d_k y_k\right)^2 \\
&= \frac{m_I}{m_I-1}\sum_{i=1}^{m_I}\left(\frac{\sum_{k\in s_i} d_{k|i}y_k}{\pi_{I_i}} - \frac{1}{m_I}\sum_{i=1}^{m_I}\sum_{k\in s_i} d_k y_k\right)^2 \\
&= \frac{m_I}{m_I-1}\sum_{i=1}^{m_I}\left(\sum_{k\in s_i} d_k y_k - \frac{1}{m_I}\sum_{i=1}^{m_I}\sum_{k\in s_i} d_k y_k\right)^2
\end{aligned}$$

Basado en lo anterior, al definir $\breve{t}_{y_i} = \sum_{k\in s_i} d_k y_k$ como la contribución[1] de la i-ésima UPM a la estimación del total poblacional y $\bar{\breve{t}}_y = \frac{1}{m_I}\sum_{i=1}^{m_I}\breve{t}_{y_i}$ como la contribución promedio en el muestreo de la primera etapa, entonces el estimador de varianza toma la siguiente forma, conocida como el estimador de varianza del **último conglomerado**.

$$\widehat{Var}(\hat{t}_{y,p}) = \frac{m_I}{m_I-1}\sum_{i=1}^{m_I}\left(\breve{t}_{y_i} - \frac{1}{m_I}\sum_{i=1}^{m_I}\breve{t}_{y_i}\right)^2 = \frac{m_I}{m_I-1}\sum_{i=1}^{m_I}\left(\breve{t}_{y_i} - \bar{\breve{t}}_y\right)^2 \tag{7.6}$$

Siguiendo con el escenario de muestreo planteado en las secciones anteriores, si el diseño de la encuesta es estratificado por regiones h, con tres etapas de selección dentro de cada estrato, entonces al utilizar la técnica del último conglomerado, el estimador de la varianza de $\hat{t}_y$ estaría dado por

$$\hat{V}(\hat{t}_y) = \sum_h \frac{n_h}{n_h-1}\sum_{i\in s_h}\left(\hat{t}_{y_i} - \bar{\hat{t}}_{y_h}\right)^2$$

En donde $\hat{t}_{y_i} = \sum_{k\in s_{hi}} w_k y_k$, $\bar{\hat{t}}_{y_h} = (1/n_h)\sum_{i\in s_h}\hat{t}_{y_i}$ y n_h es el número de UPMs seleccionadas en el estrato h. Este procedimiento, propuesto por Morris H Hansen, William N

[1]Note que la suma de estas contribuciones en la muestra de la primera etapa da como resultado la estimación $\hat{t}_y$.

Hurwitz y William G Madow (1953) tiende a sobrestimar la varianza verdadera, aunque resulta ser una técnica apetecida por los investigadores puesto que utiliza directamente los pesos finales de muestreo o factores de expansión que son publicados por los INE.

Utilizar la técnica del **último conglomerado** es una salida práctica al problema de la estimación de la varianza que, para la mayoría de encuestas que brindan estadísticas oficiales a los países, puede tornarse bastante complejo. Si bien, la expresión (9.6) no brinda estimaciones de varianza estrictamente insesgadas, sí constituye una aproximación bastante precisa.

¿Qué es un **último conglomerado**? Es la primera unidad de muestreo en un diseño complejo. Por ejemplo, considere el siguiente diseño de muestreo en cuatro etapas:

$$\underbrace{\textbf{Municipio}}_{\text{UPM}} \Rrightarrow \underbrace{\textbf{Sector}}_{\text{USM}} \Rrightarrow \underbrace{\textbf{Vivienda}}_{\text{UTM}} \Rrightarrow \underbrace{\textbf{Hogar}}_{\text{UFM}}$$

En la primera las unidades primarias de muestreo (UPM) son los municipios; dentro de cada municipio, se seleccionan unidades secundarias de muestreo (USM) que corresponden a sectores cartográficos; de esta forma, el submuestreo continua hasta seleccionar las unidades finales de muestreo (UFM) que son los hogares.

Ahora, por lo general, la primera etapa de muestreo de una encuesta está inducida por dos tipos de diseños: estratificado o con probabilidad de selección proporcional al tamaño del municipio. En cualquiera de los dos casos, se cren subgrupos de inclusión forzosa. En el muestreo estratificado serán las ciudades grandes y en el muestreo proporcional también, puesto que la medida de tamaño inducira probabilidades de inclusión mayores a uno. Luego, los municipios pertenecientes a este subgrupo de inclusión forzosa no pueden ser condierados como UPM, sino como un estrato de ciudades grandes. En cada ciudad de este estrato se realizará un muestreo de la siguiente manera:

$$\underbrace{\textbf{Sector}}_{\text{UPM}} \Rrightarrow \underbrace{\textbf{Vivienda}}_{\text{USM}} \Rrightarrow \underbrace{\textbf{Hogar}}_{\text{UFM}}$$

Es necesario tener en cuenta esta particularidad de las encuestas para poder aplicar correctamente esta técnica de aproximación de varianzas. En resumen, para aquellas ciudades que pertenecen al estrato de inclusión forsoza, 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 menor a uno.

B Linealización de Taylor

Cuando se trata de estimar parámetros que tienen una forma no lineal, es posible recurrir al uso de las herramientas del análisis matemático para aproximar sus varianzas

con el fin de publicar las cifras oficiales con sus respectivos errores estándar. Valliant, Dever y Kreuter (2013) mencionan que esta técnica se basa en expresar el estimador como función de estimadores lineales de totales. Por ejemplo, si el interés recae en estimar un parámetro poblacional θ que a su vez depende de p estimadores lineales, entonces su estimador de muestreo se debe expresar como

$$\hat{\theta} = f(\hat{t}_1, \ldots, \hat{t}_p)$$

En donde $\hat{t}_j = \sum_{k \in s} w_k y_{jk}$ es un estimador del j-ésimo total. La linealización de Taylor supone que es posible definir una aproximación lineal de $\hat{\theta}$ así

$$\hat{\theta} - \theta \approx \sum_{j=1}^{p} \frac{\partial f(\hat{t}_1, \ldots, \hat{t}_p)}{\partial \hat{t}_j} (\hat{t}_j - t_j) = \sum_{k \in s} w_k z_k + c$$

En donde $z_k = \sum_{j=1}^{p} \frac{\partial f(\hat{t}_1, \ldots, \hat{t}_p)}{\partial \hat{t}_j} y_{jk}$ son variables linealizadas y c son constantes determinísticas que por consiguiente no aportan a la varianza de $\hat{\theta}$. Nótese lo conveniente de expresar esta aproximación de esta manera puesto que al final, las cantidades que intervienen en la varianza se pueden expresar como una suma ponderada de las variables z_k y por consiguiente es posible aplicar todos los principios establecidos anteriormente. De esta forma, asumiendo el escenario de muestreo planteado en las secciones anteriores, el estimador de la varianza de la aproximación lineal de $\hat{\theta}$ está dado por

$$\hat{V}(\hat{\theta}) = \sum_{h} \frac{n_h}{n_h - 1} \sum_{i \in s_h} \left(\hat{t}_{z_i} - \bar{\hat{t}}_{z_h} \right)^2$$

En donde $\hat{t}_{z_i} = \sum_{k \in s_{hi}} w_k z_k$ y $\bar{\hat{t}}_{z_h} = (1/n_h) \sum_{i \in s_h} \hat{t}_{z_i}$. Por ejemplo, si el interés estuviera en estimar una razón, entonces las nuevas variables linealizadas son $z_k = (1/\hat{t}_{y_2})(y_{1k} - \hat{\theta} y_{2k})$.

En la región la *Pesquisa Nacional por Amostra de Domicilios Continua*, en Brasil, y la *Encuesta de Caracterización Socioeconómica Nacional*, en Chile, utilizan esquemas de linealización de Taylor en conjunción con el acercamiento del último conglomerado.

C Réplicas

Las complicaciones en el cálculo de los errores de muestreo pueden ser mayores dependiendo de la escogencia del estimador y del diseño de muestreo asumido para la recolección de la información primaria. En algunas ocasiones, 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 idea de la selección sistemática de *submuestras* que son utilizadas para estimar el parámetro de interés, utilizando los mismos principios de estimación que con la muestra completa. Por lo anterior, se obtienen estimaciones puntuales para cada réplica, las cuales son utilizadas para estimar la varianza del estimador de interés.

En particular, hay tres metodologías que abordan este problema: las réplicas repetidas balanceadas (McCarthy 1969; Judkins 1990), el Jackknife (Krewski y Rao 1981) y el Bootstrap (Rao y Wu 1988). La idea general detrás de estos métodos es que, partiendo de la muestra completa, en cada réplica se seleccione un conjunto de UPMs manteniendo todas las unidades que hayan sido seleccionadas dentro de esas UPMs. Luego, es necesario reponderar los pesos de muestreo para que se conserve la representatividad; de esta manera, para cada réplica se obtendrá un nuevo conjunto de pesos de muestreo. Con estos pesos, se calcula la estimación de interés, obteniendo tantas estimaciones como réplicas definidas. Wolter (2007) provee todos los detalles teóricos referentes al problema de la estimación de la varianza utilizando réplicas.

Por ejemplo, al utilizar la técnica de Jackknife, y asumiendo el escenario de muestreo planteado en las secciones anteriores (estratificado con varias etapas de muestreo), si se quiere estimar el error asociado a una medida de pobreza estimada, entonces se eliminará la primera UPMs del primer estrato para formar la primera réplica, luego la segunda UPMs del primer estrato para formar la segunda réplica y así sucesivamente hasta que se hayan formado todas las réplicas necesarias. De esta forma, si la UPMs i del estrato h fue eliminada, se deben crear los siguientes pesos de muestreo para las unidades que permanecen en la submuestra

$$w_k^{(hi)} = \begin{cases} 0, & \text{si } k \text{ pertenece a la UPMs } i \text{ del estrato } h. \\ \frac{n_h}{n_h - 1} w_k, & \text{si } k \text{ pertenece a la UPMs } i \text{ del estrato } h. \\ w_k, & \text{si } k \text{ no pertenece al estrato } h. \end{cases}$$

Con cada nuevo conjunto de pesos, es necesario estimar la medida de pobreza requerida utilizando la siguiente expresión:

$$\hat{F}_\alpha^{(hi)} = \frac{1}{N} \sum_{k \in s^{(hi)}} w_k^{(hi)} \left(\frac{l - y_k}{l} \right)^\alpha I(y_k < l)$$

En donde $s^{(hi)}$ hace referencia al subconjunto de la muestra inducida por la réplica al eliminar la UPMs i del estrato h. Luego, de obtener cada una de las estimaciones, la varianza de $\hat{F}_\alpha$ será estimada de la siguiente manera

$$\hat{V}_J(\hat{F}_\alpha) = \sum_h \frac{n_h - 1}{n_h} \sum_{i \in s_h} \left(\hat{F}_\alpha^{(hi)} - \hat{F}_\alpha \right)$$

En lo concerniente con las técnicas de remuestreo y la utilización de las réplicas para el cálculo de los errores de muestreo se recalca que la técnica de *Jackknife* es útil para estimar parámetros lineales, pero no tiene un buen comportamiento 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 cuando se tienen dominios pequeños que pueden inducir estimaciones nulas en la configuración de los pesos. Sin embargo, el ajuste de Fay a la técnica anterior resulta palear todos los anteriores inconvenientes. En este caso es importante utilizar una matriz de Hadammard que induzca no más de 120 réplicas para que la publicación de la base de datos no se sobrecargue. Por último, el *bootstrap* debe ser utilizado con con detenimiento porque debe replicar el diseño de muestreo exacto y esto se hace construyendo una población a partir de los pesos de muestreo.

XXXX

Las fórmulas computacionales requeridas para estimar la varianza de estadísticas descriptivas como la media muestral están disponibles 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, tales como coeficientes de correlación y coeficientes de regresión, no se encuentra fácilmente las fórmulas específicas en diseños muestrales que se aparten del muestreo aleatorio simple. Estas fórmulas son enormemente complicadas o, en última instancia, se resisten al análisis matemático (Frankel, 1971).

En ausencia de fórmulas adecuadas, en los últimos años han aparecido una variedad de técnicas empíricas que proporcionan *varianzas aproximadas que parecen satisfactorias para fines prácticos* (Kish, 1995). 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, así como intervalos de confianza para esa estimación.

Entre los métodos de muestreo, se encuentran principalmente la técnica de Jackknife, el método de las Réplicas Repetidas Balanceadas (RRB) y el método de Bootstrap.

La técnica de Jackknife

Es posible estimar la varianza de los estimadores de interés usando la técnica de Jackknife. El desarrollo del procedimiento de Jackknife se remonta a un método utilizado por Quenouille (1956) para reducir el sesgo de las estimaciones. El refinamiento ulterior del método (Mosteller & Tukey, 1968) llevó a su aplicación en una serie de situaciones de las ciencias sociales en las que las fórmulas no están fácilmente disponibles para el cálculo de errores de muestreo.

Este procedimiento ofrece los siguientes beneficios:

1. *Mayor flexibilidad*: el Jackknife puede implementarse en una amplia variedad de diseños muestrales.
2. *Facilidad de uso*: el Jackknife no requiere de software especializado.

El concepto principal de esta técnica parte de una muestra de tamaño n, la cual 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 (luego de 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 pseudovalor de θ. El estimador obtenido mediante 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 Jackknife se obtiene como:

$$\hat{V}_{JK1} = \frac{1}{A(A-1)}\sum_{a=1}^{A}\left(\hat{\theta}_a - \hat{\theta}_{JK}\right)^2$$

También es posible utilizar como estimador alternativo:

$$\hat{V}_{JK2} = \frac{1}{A(A-1)}\sum_{a=1}^{A}\left(\hat{\theta}_a - \hat{\theta}\right)^2$$

Para diseños estratificados y multietápicos en los cuales unidades primarias de muestreo han sido seleccionadas 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

$$\hat{V}_{JK}(\hat{\theta}) = \sum_{h=1}^{H}\frac{n_h - 1}{n_h}\sum_{i=1}^{n_h}(\hat{\theta}_{(hi)} - \hat{\theta})^2$$

donde $\hat{\theta}_{(hi)}$ es la estimación de θ usando los datos de la muestra excluyendo las observaciones en la i-ésima unidad primaria de muestreo (Korn & Graubard, 1999, pg. 29 – 30). Shao & Tu (1995, Teorema 6.2) garantiza la convergencia en probabilidad de este estimador hacia la varianza teórica, de donde se puede concluir que es un estimador

X1	X2	X3	X4	X5	X6	X7	X8
0.00	1.03	1.03	1.03	1.03	1.03	1.03	1.03
1.03	0.00	1.03	1.03	1.03	1.03	1.03	1.03
1.03	1.03	0.00	1.03	1.03	1.03	1.03	1.03
1.03	1.03	1.03	0.00	1.03	1.03	1.03	1.03
1.03	0.00	1.03	1.03	1.03	1.03	1.03	1.03
1.03	1.03	1.03	1.03	0.00	1.03	1.03	1.03

Figura 7.1: *Primeras ocho réplicas del método de JacKnife.*

aproximadamente insesgado para la varianza teórica. Computacionalmente, se puede obtener la estimación Jackknife por medio de la creación de la base de datos omitiendo las observaciones necesarias usando comando `as.svrepdesign` de la librería `survey` del software estadístico 'R para calcular $\hat{V}_{JK}(\hat{\theta})$, y posteriormente calcular el valor de la estimación Jackknife.

En el contexto de este estudio, se tiene un número relativamente grande de unidades primarias de muestreo: 289 escuelas para el grado tercero y 287 para el grado sexto, por lo que el método de Jackknife puede ser un poco ineficiente puesto que implica la creación de 289 réplicas para el grado tercero y 287 réplicas para el sexto, por lo cual se recurrirá otro método de remuestreo, presentado más adelante, donde el usuario define el número de réplicas.

El método de las Réplicas Repetidas Balanceadas

Las varianzas de muestreo tabmién pueden ser calculadas haciendo uso del método conocido como Réplicas Repetidas Balanceadas, el cual permite explicar la varianza que se obtiene en las estimaciones debido al muestreo. Este método es el que utilizan pruebas internacionales como PISA para realizar los análisis de datos. Bajo la metodología BRR, el estimador de la varianza toma la siguiente forma:

$$Var(\hat{\theta}) = \frac{1}{A}\sum_{a=1}^{A}(\hat{\theta}_a - \hat{\theta})^2$$

Para la aplicación de la Réplicas Repetidas Balanceadas es recomendable usar el método de Fay, el cual es similar al método Jackknife, pero es más apropiado cuando hay funcio-

X1	X2	X3	X4	X5	X6	X7	X8
2	2	0	2	0	2	2	0
2	2	0	0	2	0	2	2
2	0	0	0	0	0	0	2
2	0	2	0	0	2	0	0
2	2	0	0	2	0	2	2
2	0	0	0	2	0	2	0

Figura 7.2: *Primeras ocho réplicas del método de las Réplicas Repetidas Balanceadas.*

nes no diferenciables en el estudio. En PISA, por ejemplo, el método de Fay es preferido porque el método Jackknife no proporciona un estimador de varianza estadísticamente consistente para los cuantiles. Por otro lado, la Réplicas Repetidas Balanceadas brinda estimadores lineales simples que son imparciales y consistentes. Además, tiene una consistencia asintótica deseable para un conjunto amplio de estimadores, bajo diseños complejos y estudios de simulación empírica. Bajo la metodología BRR con el ajuste de Fay, el estimador de la varianza toma la siguiente forma:

$$Var(\hat{\theta}) = \frac{1}{A(1-\rho)^2} \sum_{a=1}^{A} (\hat{\theta}_a - \hat{\theta})^2$$

En donde $0 < \rho < 1$ es el modificador del peso de muestreo de las UPM de la siguiente manera:

$$d_k^a = \begin{cases} \rho * d_k \\ \text{si } k \text{ no pertenece a la *half-sample*} (2-\rho) d_k \text{en otro caso} \end{cases}$$

Algunos estudios por simulación han mostrado una buena eficiencia para valores de ρ iguales a 0.3, 0.5 o 0.7.

Para la aplicación de este método, los pesos de muestreo se ajustan para generar los pesos de repetición y, posteriormente, se repiten los ajustes por ausencia de respuesta

X1	X2	X3	X4	X5	X6	X7	X8
1.7	1.7	0.3	1.7	0.3	1.7	1.7	0.3
1.7	1.7	0.3	0.3	1.7	0.3	1.7	1.7
1.7	0.3	0.3	0.3	0.3	0.3	0.3	1.7
1.7	0.3	1.7	0.3	0.3	1.7	0.3	0.3
1.7	1.7	0.3	0.3	1.7	0.3	1.7	1.7
1.7	0.3	0.3	0.3	1.7	0.3	1.7	0.3

Figura 7.3: *Primeras ocho réplicas del método de las Réplicas Repetidas Balanceadas con el ajuste de Fay.*

de escuelas y estudiantes para estos nuevos pesos. Con estos pesos de repetición se estiman los errores de muestreo y la varianza de muestreo, incluyendo el impacto de la ausencia de respuesta, el cual se espera que sea pequeño, pero relevante en el momento de calcular estimadores más precisos.

Para este estudio, si se quiere aplicar esta metodología, debe seguir los siguientes pasos: primero, las unidades de muestreo (escuelas y estudiantes) deben ser agrupadas en los estratos definidos por las tres variables: departamento, sector y zona; segundo, dentro de cada estrato se eliminó una de las unidades (siguiendo una matriz de Hadamard) y se recalcula el peso (peso replicado) para la otra; y tercero, para cada conjunto de pesos replicados se calcula el indicador de interés y se determina su error estándar.

Retomando observaciones hechas anteriormente, hay estratos donde se encuetra solo una escuela, por lo que el método de las réplicas repetidas balanceadas no es aplicable puesto que al eliminar una unidad, algunos estratos quedarán vacíos.

Método de Bootstrap

En este apartado se presenta el método de Bootstrap, introducido por Bradley Efron (1979). Este método de remuestreo es el utilizado para este estudio, pues es un método que se implementa fácil, es flexible en términos del número de réplicas que se crean.

Teniendo los pesos muestrales calibrados (denotados por w_k^{cal}), se procede a crear las réplicas con el método de remuestreo con el fin de poder calcular estimaciones de indicadores junto con las varianzas de estimación. En el contexto de este estudio, se trata de realizar un remuestreo a las unidades primarias de muestreo (escuelas). Se toma a las n_I escuelas de la muestra como si fuera la población, y de esta población se selecciona una muestra con reemplazo de n_I selecciones teniendo en cuenta la probabilidad de selección del diseño π-PT de las escuelas. Dado que la selección es con reemplazo, una escuela puede quedar seleccionada más de una vez en esta nueva muestra. Al terminar la selección, se obtiene un vector de tamaño n (tamaño muestral de los estudiantes) indicando el número de veces que la escuela correspondiente queda seleccionada, posteriormente se multiplica este vector con los pesos muestrales de los estudiantes w_k^c. El anterior proceso se repite un número grande de veces, que para este estudio, se escogió usar 200 réplicas (se usó la función `as.svrepdesign` de la librería `survey` para la creación de estas réplicas).

En resumen, para la i-ésima réplica, se tiene los pesos muestrales $w_1^i, w_2^i, \cdots, w_n^i$, con $i = 1, \cdots, 200$, y estos pesos serán utilizados para calcular las estimaciones de totales, proporciones, promedios y razones y sus respectivas varianzas o desviaciones. Por ejemplo, si para una variable y_k se requiere estimar el total poblacional t_y, se proce a calcular el total en cada una de las 200 réplicas

$$\hat{t}_y^i = \sum_{k \in s} y_k w_k^i$$

con $i = 1, \cdots, 200$. La estimación de t_y es

$$\hat{t}_y = \frac{1}{200} \sum_{i=1}^{200} \hat{t}_y^i$$

y varianza estimada de la estimación es

$$\hat{Var}(\hat{t}_y) = \frac{1}{199} \sum_{i=1}^{200} (\hat{t}_y^i - \hat{t}_y)^2$$

Usando la anterior expresión de la varianza se puede calcular medidas de precisión para las estimaciones, tales como: coeficiente de variación, margen de error relativo, margen de error absoluto e intervalo de confianza.

Error de muestreo utilizando imputación múltiple

Por ejemplo, si se quiere realizar mediciones de pobreza utilizando la imputación múltiple es necesario primero establecer un modelo sobre los ingresos y_k y luego generar Q posibles valores y_k^q $(q = 1, \ldots, Q)$ para cada individuo que no respondió. Luego, utilizando los Q conjuntos de datos completos, es necesario estimar la siguientes cantidades

X1	X2	X3	X4	X5	X6	X7	X8
2	0	2	3	2	2	0	1
1	1	3	1	0	1	0	1
0	0	1	1	1	1	0	0
0	1	1	1	0	1	0	1
1	1	3	1	0	1	0	1
0	1	2	2	2	1	0	0

Figura 7.4: *Primeras ocho réplicas del método de Bootstrap.*

$$\hat{F}_\alpha^q = \frac{1}{N}\sum_{k\in s} w_k \left(\frac{l - y_k}{l}\right)^\alpha I(y_k < l) \qquad q = 1, \ldots, Q.$$

El estimador final basado en la técnica de imputación múltiple será el promedio simple de las anteriores estimaciones, dado por

$$\tilde{F}_\alpha = \frac{1}{Q}\sum_{q=1}^{Q} \hat{F}_\alpha^q$$

La varianza de esta metodología se puede descomponer en dos componentes, el primero correspondiente a la variación dentro de cada conjunto de datos creado, y el segundo correspondiente a la variación entre cada estimación resultante. Por lo tanto, la varianza asociada a $\tilde{F}_\alpha$ es

$$\hat{V}(\tilde{F}_\alpha) = \frac{1}{Q}\sum_{q=1}^{Q} \hat{V}(\hat{F}_\alpha^q) + \left(1 + \frac{1}{Q}\right)\frac{1}{Q-1}\sum_{q=1}^{Q}(\hat{F}_\alpha^q - \tilde{F}_\alpha)^2$$

Nótese que, una vez se tienen los conjuntos de datos completos, es posible estimar $\hat{V}(\hat{F}_\alpha^q)$ utilizando las técnicas del último conglomerado en conjunción con el Jackkinfe. Por último, existen otras formas de imputación no probabilística, tales como el vecino

más cercano, técnicas de *hot-deck*, imputación lógica, entre otras. En resumen, la característica principal del proceso imputación es utilizar la información auxiliar para aproximar con precisión los valores faltantes. De esta forma, las estimaciones poblacionales de los parámetros de interés tendrán sesgo nulo o despreciable y la confiabilidad de la estrategia de muestreo se mantendrá como se planeó en una primera instancia.

XXXX

III El efecto de diseño y el error de muestreo

Cuando se selecciona una muestra utilizando un diseño de muestreo complejo es muy improbable que exista independencia entre las observaciones. Una forma sencilla de incorporar el efecto de aglomeración en las expresiones de tamaño de muestra está dada por la siguiente relación, denotada como efecto de diseño (Kish 1965):

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

En donde $Var_p(\hat{\theta})$ denota la varianza de un estimador $\hat{\theta}$ bajo un diseño de muestreo complejo p y $Var_{MAS}(\hat{\theta})$ denota la varianza del mismo estimador $\hat{\theta}$ bajo 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 un total, una proporción, una razón, un coeficiente de regresión, etc.

Por ejemplo, suponiendo que el parámetro de interés es la media poblacional ($\bar{y}_U$) de una variable de interés y, como el ingreso mensual, es posible escribir la varianza del estimador bajo el diseño de muestreo complejo como

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

En donde $S^2_{y_U}$ corresponde a la varianza de la características de interés, N es el tamaño de la población de interés y n el tamaño de la muestra de individuos. Ahora, al partir de la anterior expresión, es posible mostrar que el tamaño de muestra requerido para estimar este parámetro de interés en una encuesta repetida, con un error de muestreo relativo menor a δ y una confianza estadística mayor a $1-\alpha/2$, está dado por la siguiente expresión.

$$n \geq \frac{S_{y_U}^2 DEFF}{\frac{\delta^2 \bar{y}_U^2}{z_\alpha^2} + \frac{S_{y_U}^2 DEFF}{N}}$$

En donde z_α es el percentil $(1 - \alpha/2)$ asociado a una distribución normal estándar. Nótese que si ρ es grande, entonces el valor del efecto de diseño $DEFF$ también lo será y por consiguiente el tamaño de 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 el tipo de hogar está altamente asociado con el ingreso de los individuos. Esto quiere decir que los ingresos no están uniformemente dispersos a través de todos los hogares, y por ende el coeficiente de correlación intraclase será alto.

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

$$n \geq \frac{P\ (1-P)\ DEFF}{\frac{\delta^2}{z_\alpha^2} + \frac{P\ (1-P)\ DEFF}{N}}$$

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

$$DEFF_h = \frac{Var_p(\hat{\theta}_h)}{Var_{MAS}^h(\hat{\theta}_h)}$$

En donde $Var_{MAS}^h(\hat{\theta}_h)$ es la varianza restringida al estrato $h(h = 1, \ldots, H)$, cuyo valor es el siguiente:

$$Var_{MAS}^h(\hat{\theta}_h) = \frac{1}{n_h}\left(1 - \frac{n_h}{N_h}\right) S_{z_h}^2$$

Siendo n_h el tamaño de la muestra en el estrato h, N_h el tamaño poblacional del estrato h y $S_{z_h}^2$ la varianza muestral de la variable linealizada restringida al subgrupo h.

Por otro lado, cuando la subpoblación de interés no es un estrato sino un subgrupo aleatorio - como por ejemplo las personas pobres, las personas menores de 25 años, o cualquier otro subgrupo no planeado en el diseño de la encuesta - cuyo tamaño de muestra es aleatorio entonces la estimación correcta del efecto de diseño es la siguiente:

$$DEFF_U = \frac{Var_p(\hat{\theta}_h)}{Var_{MAS}^U(\hat{\theta}_h)}$$

En donde $Var^{U}_{MAS}(\hat{\theta}_h)$ es la varianza poblacional del estimador de interés, cuyo valor es el siguiente:

$$Var^{U}_{MAS}(\hat{\theta}_h) = \frac{1}{n}\left(1 - \frac{n}{N}\right) S^2_{z_{hU}}$$

En donde $S^2_{z_{hU}}$ es la varianza muestral de la variable linealizada calculada en toda la población. Por lo tanto, en ambos efectos de diseño, la estimación de la varianza del diseño de muestreo complejo $Var_p(\hat{\theta}_h)$ es la misma, pero el denominador cambia dependiendo de si el subgrupo es un estrato o no. Es por esta razón que en los tres softwares las cifras relacionadas con la estimación puntual, errores estándar, intervalos de confianza y coeficientes de variación coinciden plenamente en los tres softwares.

Ahora, tanto los software Stata como SPSS estiman por defecto el $DEFF_U$. Nótese que, en este caso, las estimaciones de $Var^{U}_{MAS}(\hat{\theta}_h)$ y $Var^{h}_{MAS}(\hat{\theta}_h)$ serán diferentes, puesto que la primera involucra a toda la muestra, mientras que la segunda involucra únicamente a la muestra del estrato. Retomando el ejemplo, debido a que los subgrupos de interés son agregaciones de los estratos de diseño, no es correcto utilizar el enfoque que Stata trae por defecto.

Capítulo 8

Agregación de encuestas y análisis longitudinal

Para producir indicadores sociales de forma anual, es común recurrir a la agregación de las bases de datos provenientes de las encuestas de hogares, cuya periodicidad puede ser mensual, trimestral o semestral. En este docuemnto se exploran algunas estrategias de estimación ligadas al tratamiento de los pesos inducidos por el diseño de muestreo complejo y al tratamiento de las unidades que se repiten en algún periodo debido al caracter rotativo de la medición.

Uno de los primeros acercamientos al problema de la estimación conjunta de indicadores sociales utilizando varios periodos de recolección se presenta en Gurney y Daly (1965), en donde se examina cómo mejorar el estimador puntual por medio de la correlación natural que se tiene con periodos anteriores, siguiendo un enfoque inferencail basado en modelos estocásticos. Lent, Miller y Duff (1999) definen una aproximación a un estimador para las distintas clasificaciones de la fuerza de trabajo que está basado en la optimización de los coeficientes de un estimador compuesto. Por su parte, **Fuller_1990** provee una discusión acerca de los sesgos que se pueden generar en el análisis de encuestas repetidas debido a errores de medición y revisa detalladamente algunos modelos estimados con mínimos cuadrados. Además, Bell (2001) examina varios acercamientos al problema de estimar indicadores sociales, específicamente relacionados con la fuerza de trabajo, provenientes de encuestas de hogares que tienen definido un esquema de rotación y traslape entre distintos periodos de tiempo.

Recientemente, D. Steel y McLaren (2008) revisaron las principales dificualtades al momento de diseñar y analizar encuestas repetidas. Teniendo en cuenta los patrones de rotación en la estimación de los indicadores de nivel y de camnbio, examinan su efecto en la estrategia de estimación de las varianzas de los estimadores de interés. Luego, Lewis (2017) definieron algunos procediemientos que se deben seguir al momento de combinar dos o más conjuntos de datos con el propósito de implementar eficicientemente pruebas

de significación estadística sobre indicadores de cambio sobre el tiempo, además de incrementar el tamaño de meustra para realizar inferencias de subgrupos poblacionales que están insuficientemente representados en una sola medición.

I Esquemas de acumulación de muestras

Antes de entrar en los detalles técnicos involucrados en este tipo de procedeimitnos, tomemos una situación ejemplificante específica para ilustrar la problemática que se quiere abordad. Para esto, suponga que una Oficina Nacional de Estadística en América Latina ha previsto una nueva forma de análisis de su encuesta de empleo. Con el fin de tener representatividad a nivel más desagregado (provinicial, por ejemplo), y poder estimar con mayor precisión, ha decidido realizar una agregación anual de todos los levantamientos de su encuesta de empleo. Por ejemplo, suponga que en los meses de marzo, junio, septiembre y diciembre se planean levantamientos trimestrales y que este esquema tenía representatividad nacional, en el área urbana y rural, pero no tenía representatividad provincial, ni de las ciudades principales del país. Con la metodología de agregación de muestras podría ser posible asegurar la representatividad en las provincias desagregadas por área (urbano o rural).

Los procesos de acumulación de muestras son realizados con frecuencia en las encuestas continuas con publicación trimestral. Por ejemplo, se pueden planear levantamientos mensuales y acumular tres meses para realizar la publicación trimestral de la cifra de desempleo. De hecho, algunos países han decidio publicar cifras mensuales del desempleo teniendo en cuenta la acumulación de los úlitmos tres levantamientos, lo que es conocido cono trimiestres móviles. Teniendo en cuenta el diseño rotativo de la mayoría de encuestas en América Latina, una de las bondades de estos esquemas de agregación de muestras en los trimestres móviles es el panel original se mantiene y además, por diseño, la misma vivienda no es entrevistada dos veces en el trimestre movil. En este tipo de diseños, inclusive es posible que, al final de cada año en diciembre, se contemple la publicación de un gran agregado anual que contemple la agregación de los doce meses anteriores. En este escenario sí existen viviendas que han sido entrevistadas dos veces y este porcentaje, dependiendo del diseño rotativo, puese no ser bajo. Por ejemplo, en un panel 2(2)2, el diseño rotativo induce un traslape natural del 50% entre trimestres.

Korn y Graubard (1999, capítulos 7 y 8) proveen un recuento exhaustivo sobre las opciones de ponderación y otros temas a considerar 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 UPM por diseño. Es muy importante que la identificación de las UPMs y de los estratos de muestreo se debe realizar de manera inequívoca y se debe asegurar que los siguientes principios se cumplas a cabalidad:

1. *Cuando se combinan dos o más oleadas del mismo panel es importante*

asegurarse que las UPMs sean emparejadas correctamente, de tal forma que el software las reconozca como iguales.

2. Cuando se combinan dos o más muestras independientes es importante asegurarse que las UPMs estén codificada de tal 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 complicados. **Train_Cahoon_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 manera 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 *Jackknife*, *Bootstrap* o *BRR*. Nótese que el uso apropiado de tales métodos, también dependerá del origen de la encuesta y de sus objetivos. Por ejemplo, los insumos de aplicación de los métodos serían unos si la encuesta está orientada a medir el desempleo (por ejemplo, porque es una encuesta nacional de la fuerza de trabajo llevada a cabo por un instituto nacional de estadística), y serían diferentes si la encuesta está diseñada para estimar los cambios brutos entre dos periodos de tiempo.

Actualmente, los softwares estadísticos más comunmente utilizados incluyen procedimientos para la estimación de la varianza teniendo en cuenta diseños de muestreo complejos. Una forma sencilla de usarlos es siguiendo estos pasos:

1. Modificar los pesos, de tal forma que cumplan las restricciones poblacionales básicas.
2. Definir los estratos de interés en donde el diseño de meustreo se realiza de forma independiente.
3. Definir estrictamente las UPM como aglomerados poblacionales que incluyen a los hogares y personas (con sus múltiples entrevistas).
4. Realizar una estimación de la razón entre el número de personas desempleadas, como el numerador, y el núemro de personas que pertenecen a la fuerza de trabajo, como el denominador.

Suponiendo una muestra grande, la varianza resultante, calculada de esta manera, será un poco conservadora (sobreestimada).

II Factores de expansión y estimadores de muestreo

Si el investigador está interesado en estimar la tasa de desempleo anual de una encuesta rotativa, que se lleva a cabo cuatro veces al año, es posible usar los cuatro conjuntos de datos y unir los cuatro trimestres para estimar la tasa de desempleo anual. Una solución inicial a este problema consiste en unir los cuatro conjuntos de datos y dividir los

pesos de muestreo de cada periodo por un factor de cuatro. El anterior procedimiento induce estimadores puntuales aproximadamente insesgados, aunque las estimaciones de los errores estándar se tornan un poco más complicadas, puesto que se debe concatenar exahustivamente las UPM (o incluso crear algunas pseudo-UPM), si el archivo no contiene códigos de grupo.

Por supuesto, las encuestas que utilizan diseños rotativos, en donde un hogar es entrevistado en varias ocasiones, deben adjuntar dos clases de pesos de muestreo: los transversales y los longitudinales. Los pesos transversales, discutidos en las secciones anteriores, son aquellos inducidos por el diseño de muestreo de la encuesta en cada aplicación y que permiten obtener estimaciones de los parámetros de interés de forma periódica (mensual, trimestral o semestral). De esta forma, por ejemplo en una encuesta de fuerza de trabjo, los datos transversales se usarán para producir estimaciones periódicas de la participación en la fuerza de trabajo, o de la tasa de pobreza, o de la tasa de desempleo, de manera periódica. Por ejemplo, la estimación de la tasa de desempleo usa un promedio ponderado de la siguiente forma

$$\hat{\theta} = \frac{\sum_s w_k y_k}{\sum_s w_k z_k}$$

En donde, para la persona k-ésima, w_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 fuerza de trabajo). Esta estrategia de estimación asume que cada persona se representa a sí misma y a otras más en la población. Nótese que los pesos transversales asignados estarán determinados por la probabilidad de selección de las UPMs, la probabilidad de selección del hogar dentro de la UPMs, el ajuste por ausencia de respuesta en ese mismo mes, ajustes de submuestreo, calibración, entre otros. Por tales razones, aunadas a la incorporación de la nueva muestra en un diseño rotativo, además de la ausencia de respuesta y también por los cambios en el tamaño de la población de interés, el peso de un individuo puede cambiar de un periodo a otro. De esta forma, si w_k^{t-1} y w_k^t representan el peso de muestreo del individuo k en los periodos $t-1$ y t, respectivamente, es casi seguro que

$$w_k^{t-1} \neq w_k^t$$

Es posible considerar diferentes acercamientos ante esta situación. Feinberg y Stasny (1983) asumen que las diferencias en los dos pesos ocurren solamente como resultado de los flujos naturales de entrada y salida de la población de interés. Por ejemplo, si el individuo es clasificado como empleado en ambos tiempos y $w_k^{t-1} = 300$ y $w_k^t = 305$, entonces 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 $w_k^t = 300$, entonces 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. Aunque es posible considerar otros supuestos, en cualquier caso es necesario crear un nuevo conjunto de pesos (llamados pesos longitudinales) que resulten del traslape de las bases de datos en los tiempos $t-1$ y t. Por ejemplo, siguiendo la metodología de la encuesta *Survey of Labour and Income Dynamics* (Naud 2002; LaRoche 2003), estos pesos pueden ser ajustados por un factor γ (que resulta ser el inverso de la probabilidad de traslape). De esta forma un primer paso para crear los pesos longitudinales es mediante la siguiente expresión

$$w_k^{longitudinal} = \gamma \; w_k^t$$

En conjunción con lo anterior y teniendo en cuenta que es posible que los INE ya hayan publicado las cifras oficiales del parámetros de interés en los tiempos $t-1$ y t, es necesario que los totales fila coincidan con las estimaciones publicadas en el periodo $t-1$ y que los totales columna coincidan con las estimaciones publicadas en el periodo t. Para mantener esta consistencia es posible utilizar el método de *Raking* que se incializa con un factor de expansión inicial para el análisis de los trimestres móviles que debe ser modificado de manera proporcional a los pesos originales de los levantamientos mensuales teniendo en cuenta la siguiente relación

$$\hat{t}_y = \sum_{s1 \cup s2 \cup s3} w_k y_k = \sum_{s1} d_{1k} y_k + \sum_{s2} d_{2k} y_k + \sum_{s3} d_{3k} y_k$$

Por lo tanto, en la agregación trimestral el factor de expansión de cada individuo y hogar debe ser multiplicado por el siguiente factor:

$$a_i = \frac{\sum_{k \in s_i} d_{ik}}{\sum_{i=1}^{3} \sum_{k \in s_i} d_{ik}}$$

El factor de expansión para el análisis de la agregación anual debe ser modificado de manera proporcional a los pesos originales de los levantamientos mensuales y trimestrales teniendo en cuenta la siguiente relación

$$\hat{t}_y = \sum_{s_1 \cup \ldots \cup s_{12}} w_k y_k = \sum_{s_1} d_{1k} y_k + \cdots + \sum_{s_{12}} d_{12k} y_k$$

Por lo tanto, en la agregación anual el factor de expansión de cada individuo y hogar debe ser multiplicado por el siguiente factor:

$$b_i = \frac{\sum_{k \in s_i} d_{ik}}{\sum_{i=1}^{12} \sum_{k \in s_i} d_{ik}}$$

La nueva estructura longitudinal de los pesos debe garantizar que la la suma de los pesos esté acorde con la población a la cual quiere representar.

$$\sum_{i}^{3}\sum_{s_i} a_i d_{ik} \approx N$$

$$\sum_{i}^{12}\sum_{s_i} b_i d_{ik} \approx N$$

Una vez se han construido los pesos longitudinales, es necesario realizar los siguientes chequeos para verificar que la ponderación es correcta:

- Suma de pesos en las cinco ciudades
- Suma de pesos a nivel nacional
- Suma de pesos por rural/urbano
- Suma de pesos en cada una de las provincias (en el caso de la agregación anual)

Una vez que se haya generado la base de datos agregada (trimestral o anual) es necesario decidir acerca de qué totales auxiliares serán utilizados en esta calibración. Por ejemplo en la agregación trimestral, es posible escoger el mes intermedio o el promedio de los tres meses. Se espera que este ajuste final de los pesos sea minúsculo y no afecte la estructura de los pesos mensuales puesto que se trata de calibrar unos pesos que orignalmente fueron calibrados en las publicaciones mensuales. Por otro lado, debido a que este último paso se realiza con propósitos de mantener la consistencia de 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 contemplar a todas las viviendas que fueron parte del trimestre móvil. Mientras que las agregaciones anuales deben contemplar las viviendas que han sido seleccionadas más de una vez (debido al esquema de rotación del panel) y por ende todas sus mediciones deben aparecer en la base de datos tantas veces como fueron visitadas. Por otro lado, las agregaciones trimestrales no deberían contemplar ninguna vivienda con mediciones repetidas puestos que el esquema de panel actual no lo contempla. 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 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 muy cercana a cero. Nótese que el estimador de interés toma la siguiente forma:

$$\hat{t}_y = \sum_{s1} d_{1k} y_k + \sum_{s2} d_{2k} y_k + \sum_{s3} d_{3k} y_k = \hat{t}_y^1 + \hat{t}_y^2 + \hat{t}_y^3$$

En este caso, la varianza del estimador está dada por

$$Var(\hat{t}_y) = Var(\hat{t}_y^1) + Var(\hat{t}_y^2) + Var(\hat{t}_y^3)$$

En la estimación del error de muestreo para las agregaciones anuales se debe considerar que el muestreo no es independiente en los doce meses. Nótese que el estimador de interés toma la siguiente forma:

$$\hat{t}_y = \sum_{i=1}^{12} \sum_{s_i} d_{ik} y_k = \sum_{i=1}^{12} \hat{t}_y^i$$

En este caso, la varianza del estimador está dada por

$$Var(\hat{t}_y) = \sum_{i=1}^{12} Var(\hat{t}_y^i) + 2 \sum_{i=1}^{12} \sum_{j<i} Cov(\hat{t}_y^i, \hat{t}_y^j)$$

Una vez que se ha llevado a cabo el proceso de computo de los nuevos pesos longitudinales en las agregaciones (trimestrales o anuales) es necesario que se realice nuevamente un proceso de calibración sobre las variables involucradas en la calibración mensual.

III Efecto del tipo de encuesta en la eficiencia de los indicadores

Lograr una estimación adecuada del error de muestreo en las comparaciones de múltiples periodos de tiempo, ya sea con la agregación de 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.

Cambios netos

Por ejemplo, considere el cambio neto de la media de la variable de interés x en dos periodos de tiempo (t_2 y t_1)

$$\Delta = \bar{x}_2 - \bar{x}_1$$

Este parámetro de cambio en los dos periodos de tiempo es estimado de forma aproximadamente insesgada mediante la siguiente expresión:

$$\hat{\Delta} = \hat{\bar{x}}_2 - \hat{\bar{x}}_1 = \frac{\sum_{k \in s_2} \frac{x_k}{\pi_k}}{\sum_{k \in s_2} \frac{1}{\pi_k}} - \frac{\sum_{k \in s_1} \frac{x_k}{\pi_k}}{\sum_{k \in s_1} \frac{1}{\pi_k}}$$

En donde s_2 y s_1 representan las muestras seleccionadas en los periodos 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:

$$Var(\hat{\Delta}) = Var(\hat{\bar{x}}_2) + Var(\hat{\bar{x}}_1) - 2Cov(\hat{\bar{x}}_2, \hat{\bar{x}}_1)$$

En general, el último término se puede expresar como

$$2Cov(\hat{\bar{x}}_2, \hat{\bar{x}}_1) = 2\sqrt{Var(\hat{\bar{x}}_2)}\sqrt{Var(\hat{\bar{x}}_1)}\sqrt{T_2}\sqrt{T_1}R_{12}$$

En donde T_2 y T_1 representan el porcentaje de muestra común que se traslapa en ambos levantamientos y R_{12} representa la correlación de la variable de interés x en los periodos observados. Suponiendo que la variación de la variable de interés es homogénea en ambos periodos y que el traslape es común por diseño, entonces la expresión de la varianza se reduce de la siguiente manera:

$$Var(\hat{\Delta}) = 2Var(\hat{\bar{x}}) - 2\sqrt{Var(\hat{\bar{x}})}TR = 2Var(\hat{\bar{x}})(1 - TR)$$

Nótese que la varianza de este indicador cambiará de acuerdo al tipo de encuesta que se elija:

- Encuesta repetida: en donde $T = 0$ y

$$Var(\hat{\Delta}) = 2Var(\hat{\bar{x}})$$

.

- Encuesta de panel, en donde $T = 1$, $R > 0$ y

$$Var(\hat{\Delta}) = 2Var(\hat{\bar{x}})(1 - R)$$

.

- Encuesta rotativa: en donde $T \neq 0$, $R > 0$ y

$$Var(\hat{\Delta}) = 2Var(\hat{\bar{x}})(1 - TR)$$

.

Si se supone que la **correlación es positiva** para la variable en los dos periodos de tiempo, entonces se tiene la siguientes conclusión:

$$2Var(\hat{\bar{x}})(1-R) < 2Var(\hat{\bar{x}})(1-TR) < 2Var(\hat{\bar{x}})$$

Es decir que se necesita **menos** tamaño de muestra para medir los cambios netos usando un diseño panel que un diseño sin traslape. Un camino medio es el diseño rotativo.

Promedio trimestral

Considere una encuesta continua y mensual en donde se quiere estimar el promedio trimestral de la variable de interés x en tres periodos de tiempo (t_3, t_2 y t_1)

$$\Theta = \frac{\bar{x}_3 + \bar{x}_2 + \bar{x}_1}{3}$$

Un estimador del promedio trimestral que es estimado de forma aproximadamente insesgada mediante la siguiente expresión:

$$\hat{\Theta} = \frac{1}{3}\left(\hat{\bar{x}}_3 + \hat{\bar{x}}_2 + \hat{\bar{x}}_1\right) = \frac{1}{3}\left(\frac{\sum_{k\in s_3}\frac{x_k}{\pi_k}}{\sum_{k\in s_3}\frac{1}{\pi_k}} + \frac{\sum_{k\in s_2}\frac{x_k}{\pi_k}}{\sum_{k\in s_2}\frac{1}{\pi_k}} + \frac{\sum_{k\in s_1}\frac{x_k}{\pi_k}}{\sum_{k\in s_1}\frac{1}{\pi_k}}\right)$$

En donde s_3, s_2 y s_1 representan las muestras seleccionadas en los periodos 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:

$$\begin{aligned} Var(\hat{\Theta}) = \frac{1}{9}[&Var(\hat{\bar{x}}_3) + Var(\hat{\bar{x}}_2) + Var(\hat{\bar{x}}_2) + \\ &2Cov(\hat{\bar{x}}_3, \hat{\bar{x}}_2)) + 2Cov(\hat{\bar{x}}_3, \hat{\bar{x}}_1)) + 2Cov(\hat{\bar{x}}_2, \hat{\bar{x}}_1)] \end{aligned}$$

Suponiendo que la variación de la variable de interés es homogénea en ambos periodos y que el traslape es común por diseño y que los errores de muestreo son débilmente estacionarios entre dos y tres meses, entonces la expresión de la varianza se reduce de la siguiente manera:

$$Var(\hat{\Theta}) = \frac{1}{9}Var(\hat{\bar{x}})[3 + 6TR]$$

En donde R es la correlación de la variable de interés en dos y tres meses (asumida homogénea). Nótese que la varianza de este indicador cambiará de acuerdo al tipo de encuesta que se elija:

- Encuesta repetida: en donde $T = 0$ y

$$Var(\hat{\Theta}) = \frac{1}{3}Var(\hat{\bar{x}})$$

.

- Encuesta de panel, en donde $T = 1$, $R > 0$ y

$$Var(\hat{\Theta}) = \frac{1}{3}Var(\hat{\bar{x}})[3 + 6R]$$

.

- Encuesta rotativa: en donde $T \neq 0$, $R > 0$ y

$$Var(\hat{\Theta}) = \frac{1}{3}Var(\hat{\bar{x}})[3 + 6TR]$$

.

De esta forma, si se supone que la **correlación es positiva** para la variable en los tres periodos de tiempo, entonces se tiene la siguiente conclusión:

$$\frac{1}{3}Var(\hat{\bar{x}})[3 + 6R] > \frac{1}{3}Var(\hat{\bar{x}})[3 + 6TR] > \frac{1}{3}Var(\hat{\bar{x}})$$

Es decir que se necesita **más** tamaño de muestra para estimar un promedio trimestral usando un diseño panel que un diseño sin traslape.

IV Pruebas de hipótesis sobre indicadores longitudinales

Para decidir si un cambio en la dinámica de los parámetros de interés es significativo entre dos periodos de tiempo es necesario llevar a cabo una prueba de hipóstesis. 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 conocer, por ejemplo, si hay un cambio significativo e importante en la reducción de la desocupación (entre grupos y en distintos periodos del 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 UPMs y que el tamaño de muestra de hombres y mujeres es aleatorio. Para realizar comparaciones nacionales o regionales en dos periodos de tiempo (por ejemplo comparar la situación laboral de un paíse 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 enceustas que contemplan diseños de panel rotativo. Considere 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 trabajamos con el siguiente estimador de diferencias:

$$\hat{d} = \hat{\theta}_2 - \hat{\theta}_1$$

La varianza asociada a este estimador está dada por

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

Y por último, el término de covarianza se puede escribir como

$$Cov(\hat{\theta}_1, \hat{\theta}_2) = \sqrt{Var(\hat{\theta}_1)}\sqrt{Var(\hat{\theta}_2)}\sqrt{P_1}\sqrt{P_2}R_{1,2}$$

Existen muchos escenarios de comparación que son de interés cuando se analizan datos de una encuesta de empleo. Estas comparaciones se hacen mas complejas cuando se incluye en el análisis el diseño en 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 UPMs, el término de covarianza no será nulo.*

A Ejemplos

Covarianza en comparaciones mensuales

Suponga que se quiere comparar la tasa de desempleo nacional entre dos meses consecutivos. En este escenario existe independencia en el muestreo de los dos meses consecutivos y por lo tanto el porcentaje de traslape de muestra entre los dos meses (que por diseño es nulo) es igual a cero.

Por lo tanto, $P_1 = P_2 = 0$. Luego, el término de la covarianza se anula. En resumen, la varianza del estimador en este caso sería igual a:

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

Covarianza en comparaciones trimestrales o anuales

Suponga 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 puestos que la estructura del panel garantiza un traslape del 50%. En este caso $P_1 = P_2 \approx 0.5$.

Por otro lado, existe una correlación natural entre las viviendas comunes en el panel que se midieron en los periodos de interés, por lo tanto $R_{1,2} \neq 0$. Note que 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 fuerza de trabajo + estado de ocupación).

En resumen, el término de covarianza en este caso sería igual a:

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

Covarianza en comparaciones de un mismo mes

Suponga que se quiere comparar la tasa de desempleo entre hombres y mujeres en un mismo mes. A pesar de que la estructura de panel no se considere, 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 P_1 es la proporción de hombres en la muestra y P_2 es la proporción de mujeres en la muestra. Nótese que $P_1 \neq P_2$.

Por otro lado, existe una correlación natural entre las UPMs que fueron seleccionadas y que contienen tanto a hombres como a mujeres, por lo tanto $R_{12} \neq 0$. Note que esta correlación se calcula sobre todos los individuos de la muestra 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:

$$Cov(\hat{\theta}_1, \hat{\theta}_2) = \sqrt{Var(\hat{\theta}_1)}\sqrt{Var(\hat{\theta}_2)}\sqrt{P_1}\sqrt{P_2}R_{1,2}$$

Covarianza en comparaciones de un mismo mes

Suponga que se quiere comparar la tasa de desempleo entre Quito y Guayaquil en un mismo mes. En este escenario existe independencia en el muestreo de las dos ciudades porque la selección es independiente en cada ciudad. Esta independencia se tiene por definición del diseño de muestreo puesto que ambas ciudades son estratos de muestreo. En este caso P_1 es la proporción de personas de Quito en la muestra y P_2 es la proporción de personas de Guayaquil en la muestra. Nótese que $P_1 \neq P_2$.

Por otro lado, no existe una correlación entre las UPMs que fueron seleccionadas en Quito y Guayaquil 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:

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

B Estadístico de prueba

Una vez se ha definido la estructura de varianza del estimador de interés, el siguiente paso es definir si la siguiente cantidad es distinta de cero para determinar si el parámetro ha cambiado entre grupos o a lo largo del tiempo.

$$t = \frac{\hat{d}}{\sqrt{Var(\hat{d})}}$$

Este estadístico de prueba sigue una distribución *t-student* con *df* grados de libertad que están dados por la resta entre el número de UPMs seleccionadas menos el número de estratos de muestreo considerados. De esta forma, se tiene que:

$$df = \sum_{h=1}^{H}(n_{Ih} - 1) = \sum_{h=1}^{H} n_{Ih} - H = \#UPMs - \#Estratos$$

Los grados de libertad permiten tener una inferencia precisa a medida que crecen. Por ejemplo, considere por ejemplo 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:

$$df_{subpoblacin} = \sum_{h=1}^{H} v_h(n_{Ih} - 1)$$

En donde v_h es una variable indicadora que toma el valor uno si el estrato h contiene uno o mas casos de las subpoblaciones de interés y toma el valor cero en otro caso.

Capítulo 9

Calidad de las estimaciones

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

La División de Estadísticas de la Comisión Económica para américa Latina y el Caribe provee estimaciones de indicadores sociales para cada país de la región. A través de convenios de cooperación, las oficinas nacionales de estadística (ONE) proveen anualmente las bases de datos de las encuestas de hogares y estas son sistematizadas en un repositorio interno llamado `BADEHOG`.

Las encuestas de hogares tienen un diseño complejo, probabilístico, estratificado, multietápico y con probabilidades de inclusión no uniformes. Por ende, las estimaciones elaboradas a partir de estas operaciones estadísticas están sujetas al error muestral, y se requiere evaluar su validez estadística mediante diversos indicadores de calidad que describen su precisión y confiabilidad y que a su vez alertan al usuario cuando la precisión de la estimación no es 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 con base en la información sobre el diseño muestral complejo, resumida en el factor de expansión, los estratos y las unidades primarias de muestreo (UPM), mediante el comando `svy` en `Stata` y `srvyr` en `R`. Cuando las encuestas de hogares sólo proveen los pesos de muestreo se utiliza el ajuste de West y McCabe (2012). Los criterios de calidad estimados para cada cifra están comprendidos por:

- *Intervalos de confianza*: describe un conjunto valores en donde es posible que el indicador de interés se encuentre. Este intervalo está determinado por el estimador de muestreo para el indicador, su error estándar y el percentil 0,975 de una distribución $t - student$ con grados de libertad igual a la diferencia entre estratos y UPM.
- *Tamaño de muestra*: el número de unidades que comprenden el diseño de muestreo afecta indirectamente la amplitud del intervalo de confianza mediante el error estándar, el cual depende de manera inversamente proporcional al tamaño de

muestra. Con un mayor tamaño de muestra tendremos estimaciones más precisas y un intervalo de confianza más angosto.

- *Efecto de diseño*: da cuenta de la correlación entre la variable de interés y la distribución de los hogares en las unidades primarias de muestreo. Esta medida depende del promedio de hogares seleccionados por UPM y del coeficiente de correlación intraclase entre la variable de interés y las UPM.
- *Tamaño de muestra efectivo*: se calcula como la razón entre el tamaño de muestra y el efecto de diseño. El tamaño de muestra efectivo tiene por objeto deflactar el tamaño de muestra inicial por medio de la correlación intraclase de las UPM seleccionadas; de esta forma se evita contabilizar el exceso de información repetida debido a la aglomeración de los hogares en las UPM.
- *Coeficiente de variación*: este indicador modela el error de muestreo de un estimador. Se calcula como la razón entre el error estándar de la estimación y su estimación puntual. Esta medida es ampliamente utilizada para validar la precisión de una estimación, un coeficiente de variación elevado puede implicar que el error estándar de la estimación es relativamente grande.
- *Grados de libertad*: se definen como la diferencia entre el número de UPM y el número de estratos. Estos definen el valor del percentil de la distribución $t - student$ a la hora de calcular el intervalo de confianza; mientras más grados de libertad, el intervalo de confianza será más estrecho y la estimación será más precisa.
- *Coeficiente de variación logarítmico*: como el coeficiente de variación no es simétrico alrededor de 0.5, esta medida se define como una transformación logarítmica sobre la proporción, la cual evita que las estimaciones cercanas a cero sean castigadas con un coeficiente de variación alto aun cuando su variación sea pequeña.
- *Conteo de casos no ponderado*: está determinado por el número de casos en la muestra afectados por el fenómeno de estudio, sin considerar el factor de expansión, ni el diseño complejo de la encuesta.

El proceso de estimación de las medidas de calidad produce una *alerta*, que establece una señal en caso de que la estimación no sea confiable para hacer inferencias dado el diseño de muestreo complejo de cada encuesta en particular. Una estimación no confiable arrojará una señal de alerta en los siguientes casos:

1. Tamaño de muestra menor a 100 unidades.
2. Tamaño de muestra efectivo menor a 68 unidades.
3. Conteo de casos no ponderado menor a 50 unidades.
4. Coeficiente de variación mayor a 20%.
5. Grados de libertad menor a 10.
6. Coeficiente de variación logarítmico mayor a 17.5%.

I Introducción

Una gran cantidad de indicadores sociales son estimados a partir de las encuestas de hogares. En particular, la Unidad de Estadísticas Sociales de la CEPAL utiliza el repositorio de bases de datos de encuestas de hogares BADEHOG que agrupa un total de 18 países y con el cual se calculan estimaciones de pobreza, distribución del ingreso, gastos, mercado laboral, entre otros. En este documento, se pretende generar una metodología de procesamiento de las bases de este repositorio, que le permita al analista decidir acerca de la pertinencia de una estimación con base en algunos criterios de calidad en términos de la precisión de las estimaciones.

El resultado de los procesamientos corresponderá a una tabla simple que puede alimentar los procesos de análisis de información propios en la División de Estadísticas, así como dar apoyo a otras Divisiones que usen el repositorio de datos BADEHOG. El procesamiento computacional de este procedimiento se realiza conjuntamente con el software estadístico `R` (utilizando las librerías `dplyr`, `srvyr`, `survey`, `snow` y `TeachingSampling`) y con `Stata`. Como ejemplo particular, el cuadro 9.1 presenta la estimación de la proporción de la variable `Pobreza` (considerando las categorías `Pobre` y `No pobre`) para las subpoblaciones `Migrante` y `No migrante` en un país Latinoamericano para el año 2017. Note que la naturaleza de los resultados no es compleja y este documento describen los detalles del procesamiento, así como la definición de cada columna resultante.

Cuadro 9.1: *Estimación de la proporción de personas pobres desagregada por estado migratorio, junto con algunas medidas de precisión, para un país Latinoamericano.*

Subpoblación	Proporción estimada	Límite inferior	Límite superior	Coeficiente de variación	Efecto de diseño	Tamaño de muestra	Tamaño de muestra efectivo	Grados de libertad	Número de casos	Coeficiente de variación logarítmico	Alerta
Migrante	0.15	0.06	0.30	40.0	3.5	127	37	43	23	20.8	*
No migrante	0.35	0.33	0.37	2.6	14.5	38074	2620	929	12698	2.5	
Total	0.35	0.33	0.37	2.6	14.5	38201	2632	929	12721	2.5	

En el caso anterior, se presenta un procesamiento que proviene de una base de datos del repositorio BADEHOG, el cual presenta la proporción estimada de personas pobres desagregada en la subpoblación definida por el estado migratorio. Esta tabla contiene además algunas medidas de calidad que pueden ser utilizadas para concluir acerca de la precisión de la estimación puntual y, por consiguiente, la pertinencia de utilizarla para hacer recomendaciones de política pública. Como se puede notar, cuando una estadística se clasifica como no confiable, la tabla lo muestra mediante la incorporación de un asterisco en la columna llamada `alerta`.

Basados en la anterior tabla, se puede concluir que la estimación del porcentaje de

pobreza en las personas migrantes no es precisa. La razón por la cual esta cifra no es precisa es simplemente porque la encuesta no fue planeada con estos propósitos y por ende una o más medidas de calidad están evidenciándolo. Nótese que la implicación entre precisión y diseño va en una sola dirección: *si la precisión es deficiente significa que esta estimación no fue considerada en el diseño de la encuesta.* Sin embargo, es posible encontrar en muchas ocasiones que a pesar de que una estimación no fue considerada inicialmente por el diseño de la encuesta, está pueda ser considerada como precisa y confiable. En este documento se definen y establecen algunos criterios que son considerados actualmente por algunas oficinas nacionales de estadística, así como por entidades dedicadas a la investigación social.

II Principios básicos de estimación

En primer lugar, se supone que cada base de datos del repositorio BADEHOG representa el levantamiento de una encuesta de hogares para algún país de la región. Esta encuesta está inducida por la selección de una muestra s de tamaño n sobre la población de interés U de tamaño N. Para realizar la medición sobre las unidades pertenecientes a la muestra, se plantea un diseño de muestreo $P(s)$, que se asume probabilístico, estratificado y multietápico (Gutiérrez 2016a).

El procesamiento de las bases de datos que vienen de encuestas de hogares debe tener en cuenta los criterios del diseño que se estableció para realizar el levantamiento de la información primaria. En particular, la base de datos debe contener como mínimo la siguiente información para cada individuo:

- *Estratos*: son particiones geográficas de la población para las cuales se definen selecciones independientes. Estas agrupaciones son mutuamente excluyentes e inducen H sub-grupos poblacionales. En las bases de datos de `BADEHOG`, los estratos están definidos por la variable categórica `_estrato`.
- *Unidades primarias de muestreo*: son agregaciones de hogares definidas por un límite cartográfico proveniente del censo. Corresponden a la primera subdivisión de la población de hogares y están anidadas dentro de los estratos. En las bases de datos de `BADEHOG`, las UPM están definidas por la variable categórica `_upm`.
- *Pesos de muestreo*: corresponden a las ponderaciones utilizadas para representar a la población nacional a partir de los elementos de la muestra. En las bases de datos de `BADEHOG`, los pesos de muestreo están definidos por la variable continua `_fep`.

A Análisis en subpoblaciones

Aunque el marco de referencia de la teoría de muestreo es la estimación de un parámetro de interés sobre alguna característica de interés, lo cierto es que en la práctica no solo se necesitan estimaciones que cobijen la población entera, sino que también son indispensables estimaciones que involucren subgrupos poblacionales, puesto que éstos inducen una partición de la población de interés. En general, es bien sabido que cuando se habla de subgrupos poblacionales se está haciendo referencia a dominios de interés, estratos o post-estratos. Cuando el investigador se enfrenta a una encuesta que tiene en cuenta subgrupos poblacionales - es decir, siempre - es indispensable conocer en qué se diferencian cada uno de ellos, pues de esto depende que la investigación arroje resultados confiables mediante el planteamiento de la mejor estrategia de muestreo.

En términos de notación, sean $U_1, \ldots, U_g, \ldots, U_G$ los subgrupos poblacionales de interés, lo cuales constituyen una partición de la población; además, N_g denota el tamaño absoluto del subgrupo U_g, y por ende, se tiene que $\sum_{g=1}^{G} N_g = N$. A partir de las anteriores definiciones, se han planteado algunas diferencias y similitudes entre cada uno de ellos. A continuación se resumen rápidamente:

- **Dominios de interés**: Este tipo de subgrupos poblacionales son aquellos para los cuales se requieren estimaciones separadas. Estos requerimientos se planean en la etapa de diseño para asegurar que el diseño de la muestra sea tal que al momento de la recolección de la información exista una buena cobertura en cada uno de los dominios de interés. Lo anterior sólo se puede lograr ampliando el tamaño de muestra n puesto que el marco de muestreo no informa acerca de la pertenencia de los individuos a los dominios de interés. Un aspecto importante de esta clase de subgrupos poblacionales es que el número de individuos en la muestra que pertenecen a un dominio n_g de interés es siempre aleatorio, y para algunos dominios particulares puede llegar a ser muy pequeño. Por otro lado el tamaño absoluto de cada dominio N_g no se conoce ni antes de la etapa de diseño ni después de la etapa de estimación. Un ejemplo claro de estos subgrupos es la condición de ocupación, la condición de pobreza, la rama de actividad, entre otros.

- **Estratos**: Cuando el marco de muestreo permite conocer la pertenencia de todos los individuos de la población a un subgrupo poblacional, se dice que esta clase de subgrupos se llaman estratos. Más aún, 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 los individuos a los estratos se incorpora en la etapa de diseño de la muestra. Nótese que a diferencia de los dominios, en los estrato se conoce tanto N_g como n_g antes de la etapa de estimación. Un ejemplo

claro de estos subgrupos son las zonas urbanas o rurales, regiones y municipios.

- **Post-estratos**: La propiedad que caracteriza a este tipo de subgrupos poblacionales es que aunque en la etapa de diseño el tamaño del post-estrato N_g es conocido, se desconoce el número de individuos que pertenecerán al post-estrato n_g en la muestra realizada. Un ejemplo claro de estos subgrupos son los grupos etarios, el sexo o la etnia que, si bien no son utilizadas en la fase de diseño, sí se utilizan sus proyecciones demográficas en la fase de análisis para analizar mejorar la eficiencia de los estimadores. Al respecto Särndal, Swensson y Wretman (2003) afirman que existen dos situaciones en las cuales se presenta esta situación, llamada comúnmente post-estratificación:

  - Cuando el marco de muestreo es tal que se conoce la pertenencia de todos los elementos a los subgrupos poblacionales pero el investigador decide no utilizar esta información en la etapa de diseño. Las razones para esto son diversas, pero principalmente se decide obviar este tipo de información por facilidad logística. El investigador decide utilizar la información auxiliar de pertenencia a los post-estratos en la etapa de estimación para mejorar la eficiencia de la estrategia de muestreo, en particular del estimador propuesto.
  - Mediante alguna fuente de información confiable se conocen los tamaños absolutos N_g de cada subgrupo poblacional aunque se desconoce la pertenencia de los individuos a los subgrupos, pues el marco de muestreo presenta esta deficiencia. Después de la etapa de diseño, se observa la característica de interés y se pregunta acerca de la pertenencia de los individuos seleccionados en los post-estratos de tal forma que en la etapa estimación se utiliza esta información para mejorar la eficiencia de los estimadores de los parámetros de interés.

El diseño y re-diseño de las encuestas se basan fundamentalmente en la búsqueda de estos subgrupos en la población. En todas las encuestas de hogares que se planean en América Latina se busca investigar fenómenos asociados a subgrupos poblacionales que se encuentran dispersos en la geografía de los países y podría especificar los siguientes aspectos:

1. El tamaño de muestra de una encuesta siempre se basa en la incidencia de un fenómeno que clasifica a la población en algún dominio de interés.
2. Este tamaño de muestra siempre se reparte entre los diferentes estratos geográficos para mejorar la eficiencia del levantamiento de la información y del diseño de muestreo.
3. En muchas ocasiones, las proyecciones demográficas sobre los post-estratos son utilizadas en la fase de estimación para mejorar la precisión de los estimadores.

B Estimación puntual de promedios y proporciones

Teniendo en cuenta las anteriores definiciones, se asume que la variable de interés y, que puede ser de naturaleza continua o discreta, es observada sobre un individuo $k \in s$, el cual se encuentra clasificado como miembro de una subpoblación U_g $(g = 1, \ldots, G)$. Por lo tanto, se tiene que:

$$y_{g_k} = y_k * z_{g_k} = \begin{cases} y_k, \text{ Si } k \in U_g \\ 0, \text{ en otro caso} \end{cases}$$

En donde y_k es el resultado de la observación de la variable de interés y en el individuo k y z_k es la variable indicadora de la subpoblación de interés U_g, definida de la siguiente manera:

$$z_{w_k} = \begin{cases} 1, \text{ Si } k \in U_g \\ 0, \text{ en otro caso} \end{cases}$$

De esta forma, la estimación de la proporción de personas (que sería un promedio en el caso de que y fuese de naturaleza continua) con la característica de interés en el dominio U_g es:

$$\hat{P}_g = \frac{\sum_h \sum_i \sum_k w_k y_{g_k}}{\sum_h \sum_i \sum_k w_k z_{g_k}} = \frac{\hat{t}_{y_g}}{\hat{t}_{z_g}}$$

Nótese que las tres sumatorias corresponden a los estratos $(h = 1 \ldots, H)$, las UPM $(i = 1, \ldots, n_{Ih})$ y los individuos $(k = 1, \ldots, n_i)$; en donde H representa el número de estratos, n_{Ih} denota el número de UPM en el estrato h y n_i representa el número de individuos (hogares o personas) en la UPM i. Este estimador corresponde a una razón, puesto que tanto el numerador como el denominador son aleatorios, y el ponderador w_k corresponde al peso de muestreo. La estimación para la proporción (o promedio, si es una variable continua) definida en toda la población está dada por la siguiente expresión:

$$\hat{P} = \frac{\sum_h \sum_i \sum_k w_k y_k}{\sum_h \sum_i \sum_k w_k} = \frac{\hat{t}_y}{\hat{t}_z}$$

C Aproximación de la varianza

Debido a las dificultades algebraicas y computacionales, estimar la varianza del estimador - utilizando las expresiones matemáticas exactas - en encuestas complejas que contemplan esquemas de selección en varias etapas, estratificación y uso de pesos desiguales, puede tornarse bastante costoso e ineficiente (Wolter 2007). Por lo anterior, el acercamiento recomendado a la estimación de las varianzas es por medio del uso de aproximaciones. En esta sección se explican las ventajas de utilizar dicho recurso con el objetivo de establecer y definir los criterios de precisión.

Para la estimación de la varianza de los estimadores de interés en encuestas multietápicas es posible utilizar una aproximación conocida como la técnica del **último conglomerado**. Morris H. Hansen, William N. Hurwitz y William G. Madow (1953) propusieron este procedimiento y desde entonces esta aproximación se ha posicionado en las oficinas nacionales de estadística (West 2012), como la técnica estándar para la estimación de varianzas. Esta aproximación es utilizada por todos los software computacionales usados en las oficinas nacionales de estadística (como `SAS`, `SPSS`, `R`, `Stata`, `Wesvar`, `SUDAAN`, entre otros).

En general, esta aproximación, que sólo tiene en cuenta la varianza de los estimadores en la primera etapa de muestreo, supone que esa selección fue realizada con reemplazo. Los procedimientos de muestreo en etapas posteriores de la selección son ignorados, a menos que la fracción de muestreo sea importante en la primera etapa de muestreo. ¿Qué es un **último conglomerado**? Es la primera unidad de muestreo en un diseño complejo. Por ejemplo, considere el siguiente diseño de muestreo en cuatro etapas:

$$\underbrace{\textbf{Municipio}}_{\text{UPM}} \Rrightarrow \underbrace{\textbf{Sector}}_{\text{USM}} \Rrightarrow \underbrace{\textbf{Vivienda}}_{\text{UTM}} \Rrightarrow \underbrace{\textbf{Hogar}}_{\text{UFM}}$$

En la primera las unidades primarias de muestreo son los municipios; dentro de cada municipio, se seleccionan unidades secundarias de muestreo (USM) que corresponden a sectores cartográficos; de esta forma, el submuestreo continua hasta seleccionar las unidades finales de muestreo (UFM) que son los hogares. Ahora, por lo general, la primera etapa de muestreo de una encuesta está inducida por dos tipos de diseños: estratificado o con probabilidad de selección proporcional al tamaño del municipio. En cualquiera de los dos casos, se crean subgrupos de inclusión forzosa. En el muestreo estratificado serán las ciudades grandes y en el muestreo proporcional también, puesto que la medida de tamaño es el número hogares en cada municipio y podría inducir probabilidades de inclusión mayores a uno. Luego, los municipios pertenecientes a este subgrupo de inclusión forzosa no pueden ser considerados como UPM, sino como un estrato de ciudades grandes. En cada ciudad de este estrato se realizará un muestreo de la siguiente manera:

$$\underbrace{\textbf{Sector}}_{\text{UPM}} \Rrightarrow \underbrace{\textbf{Vivienda}}_{\text{USM}} \Rrightarrow \underbrace{\textbf{Hogar}}_{\text{UFM}}$$

Es necesario tener en cuenta esta particularidad de las encuestas para poder 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 menor a uno. En general, la idea es que el software reconozca tres elementos importantes en la base de datos: los estratos

$(h = 1, \ldots, H)$, las UPM $(i \in s_{I_h})$ y los pesos finales de muestreo (w_k). En las siguientes secciones, se exploran los detalles de esta técnica de aproximación de varianza.

D Razonamiento probabilístico

En particular considere cualquier estimador del total poblacional dado por la siguiente combinación lineal

$$\hat{t}_y = \sum_{k \in s} w_k y_k = \sum_{k \in U} I_k w_k y_k \tag{9.1}$$

En donde I_k son variables indicadoras de la pertenencia del elemento k a la muestra s. Ahora, asumamos que el factor de expansión de la encuesta w_k cumple con los supuestos básicos de un ponderador que hace insesgado a $\hat{t}_y$, es decir:

$$E(I_k w_k) = 1$$

Se supone un diseño de muestreo en varias etapas (dos o más) en donde la primera etapa supone la selección de una muestra s_{I_h} de n_{I_h} unidades primarias de muestreo $(i \in s_{I_h})$ en el estrato h de tal forma 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, dado que la UPM que la contiene ha sido seleccionada en la muestra de la primera etapa, está dada por la siguiente expresión:

$$\pi_{k|i} = Pr(k \in s_i | i \in s_{I_h})$$

Por ejemplo, dentro del estrato h, si el muestreo es sin reemplazo en todas sus etapas, la probabilidad de inclusión del k-ésimo elemento a la muestra s está dada por

$$\begin{aligned} \pi_k &= Pr(k \in s) \\ &= Pr(k \in s_i, i \in s_{I_h}) \\ &= Pr(k \in s_i | i \in s_{I_h}) Pr(i \in s_{I_h}) = \pi_{k|i} \times \pi_{I_i} \end{aligned}$$

Dado que el inverso de las probabilidades de inclusión son un ponderador natural, entonces se definen las siguientes cantidades:

1. $w_{I_i} = \frac{1}{\pi_{I_i}}$, que es el factor de expansión de la i-ésima UPM.
2. $w_{k|i} = \frac{1}{\pi_{k|i}}$, que es el factor de expansión del k-ésimo elemento dentro para la i-ésima UPM.
3. $w_k = w_{I_i} \times w_{k|i}$, que es el factor de expansión final del k-ésimo elemento para toda la población U.

Por lo tanto, cuando la muestra de UPM de la primera etapa fue seleccionada con reemplazo, entonces el estimador insesgado (conocido como el estimador de Hansen-Hurwitz) para el total poblacional está dado por la siguiente expresión.

$$\hat{t}_{y,p} = \sum_{h=1}^{H} \sum_{i \in sI_h} \frac{1}{n_{I_h}} \frac{\hat{t}_{y_i}}{p_{I_i}} \tag{9.2}$$

Y una estimación insesgada de su varianza es:

$$\widehat{Var}(\hat{t}_{y,p}) = \sum_{h=1}^{H} \sum_{i \in sI_h} \frac{1}{n_{I_h}(n_{I_h} - 1)} \left(\frac{\hat{t}_{y_i}}{p_{I_i}} - \hat{t}_{y,p} \right)^2 \tag{9.3}$$

Suponga ahora que la encuesta tiene un diseño bietápico $p(s)$ que no contempla reemplazo en la primera etapa. Por lo tanto, algunas cantidades deben ser equiparadas para poder utilizar esta aproximación. En principio, nótese que las cantidades $\hat{t}_{y_i}$ representan los totales estimados de la variable de interés en la i-ésima UPM y están dados por la siguiente expresión:

$$\hat{t}_{y_i} = \sum_{k \in s_i} \frac{y_k}{\pi_{k|i}} = \sum_{k \in s_i} w_{k|i} y_k \tag{9.4}$$

Utilizar la aproximación de la varianza requiere equiparar los términos de manera apropiada. En primer lugar, fijémonos en los estimadores dados por (9.2) y (9.1). Para realizar esta comparación, se requiere que se asuma la siguiente igualdad en las probabilidades de inclusión de la primera etapa:

$$\pi_{I_i} = p_{I_i} \times n_{I_h} \tag{9.5}$$

Por lo tanto, el estimador del total poblacional quedaría definido desde (9.1) como un estimador de tipo Hansen-Hurwitz.

$$\hat{t}_y = \sum_{k \in s} w_k y_k = \sum_{h=1}^{H} \sum_{i \in sI_h} \sum_{k \in s_i} w_k y_k = \sum_{h=1}^{H} \sum_{i \in sI_h} \sum_{k \in s_i} \frac{1}{\pi_{I_i} \pi_{k|i}} y_k = \sum_{h=1}^{H} \sum_{i \in sI_h} \frac{\hat{t}_{y_i}}{\pi_{I_i}} \approx \sum_{h=1}^{H} \sum_{i \in sI_h} \frac{1}{n_{I_h}} \frac{\hat{t}_{y_i}}{p_{I_i}}$$

Ahora, dado que la forma del estimador ha sido equiparada con un estimador tipo Hansen-Hurwitz, es posible utilizar su estimación de varianza. Aún más, después de un poco de álgebra y utilizando la equiparación dada por (9.5), es posible tener la siguiente aproximación, cuya gran ventaja es que sólo hace uso de los factores de expansión finales w_k, los estratos y las UPM, que suelen ser reportados por las oficinas nacionales de estadística cuando liberan los micro-datos de sus encuestas, en vez de los factores de expansión de la primera etapa o los factores de expansión condicionales dentro de las UPM. Basado en lo anterior, al definir $\breve{t}_{y_i} = \sum_{k\in s_i} w_k y_k$ como la contribución[1] de la i-ésima UPM a la estimación del total poblacional y $\bar{\breve{t}}_y = \frac{1}{n_{I_h}} \sum_{i=1}^{n_{I_h}} \breve{t}_{y_i}$ como la contribución promedio en el muestreo de la primera etapa, entonces el estimador de varianza toma la siguiente forma, conocida como el estimador de varianza del **último conglomerado**.

$$\widehat{Var}(\hat{t}_{y,p}) = \sum_{h=1}^{H} \sum_{i\in sI_h} \frac{n_{I_h}}{n_{I_h}-1} \left(\breve{t}_{y_i} - \frac{1}{n_{I_h}} \sum_{i=1}^{n_{I_h}} \breve{t}_{y_i} \right)^2 = \sum_{h=1}^{H} \sum_{i\in sI_h} \frac{n_{I_h}}{n_{I_h}-1} \left(\breve{t}_{y_i} - \bar{\breve{t}}_y \right)^2 \tag{9.6}$$

Utilizar la técnica del **último conglomerado** es una salida práctica al problema de la estimación de la varianza que, para la mayoría de las encuestas que brindan estadísticas oficiales a los países, puede tornarse bastante complejo. Si bien, la expresión (9.6) no brinda estimaciones de varianza estrictamente insesgadas, sí constituye una aproximación bastante precisa, que en el peor de los casos sobre-estima ligeramente este parámetro (Wolter 2007). La expresión (9.6) se obtiene teniendo en cuenta el siguiente desarrollo algebraico.

$$\begin{aligned}
\widehat{Var}(\hat{t}_{y,p}) &= \frac{1}{n_{I_h}(n_{I_h}-1)} \sum_{i=1}^{n_{I_h}} \left(\frac{\hat{t}_{y_i}}{p_{I_i}} - \hat{t}_y \right)^2 \\
&= \frac{n_{I_h}}{n_{I_h}-1} \sum_{i=1}^{n_{I_h}} \frac{1}{n_{I_h}^2} \left(\frac{\sum_{k\in s_i} w_{k|i} y_k}{p_{I_i}} - \sum_{i=1}^{n_{I_h}} \sum_{k\in s_i} w_k y_k \right)^2 \\
&= \frac{n_{I_h}}{n_{I_h}-1} \sum_{i=1}^{n_{I_h}} \left(\frac{\sum_{k\in s_i} w_{k|i} y_k}{n_{I_h} p_{I_i}} - \frac{1}{n_{I_h}} \sum_{i=1}^{n_{I_h}} \sum_{k\in s_i} w_k y_k \right)^2 \\
&= \frac{n_{I_h}}{n_{I_h}-1} \sum_{i=1}^{n_{I_h}} \left(\frac{\sum_{k\in s_i} w_{k|i} y_k}{\pi_{I_i}} - \frac{1}{n_{I_h}} \sum_{i=1}^{n_{I_h}} \sum_{k\in s_i} w_k y_k \right)^2 \\
&= \frac{n_{I_h}}{n_{I_h}-1} \sum_{i=1}^{n_{I_h}} \left(\sum_{k\in s_i} w_k y_k - \frac{1}{n_{I_h}} \sum_{i=1}^{n_{I_h}} \sum_{k\in s_i} w_k y_k \right)^2
\end{aligned}$$

[1] Note que la suma de estas contribuciones en la muestra de la primera etapa da como resultado la estimación $\hat{t}_y$.

E Aproximación de la varianza en las subpoblaciones

Existe un precio que debe pagarse cuando se desconoce la membrecía de los individuos al subgrupo poblacional; es decir, cuando se realiza una estimación en dominios, la varianza tiende a ser más grande, y cuando se realizan estimaciones en estratos la varianza es mucho menor. Por otro lado, esas diferencias teóricas se encuentran matizadas en la práctica, puesto que en la realidad de las oficinas nacionales de estadística se utilizan programas computacionales que hacen uso de aproximaciones en la estimación de los errores estándar de los estimadores de interés. En general, cuando se trata de estimar la varianza de un estimador, la estimación dentro de los estratos arroja menores coeficientes de variación que la estimación dentro de los dominios, y este es el precio que se debe pagar ante el desconocimiento de la membrecía de las unidades del marco de muestreo a los subgrupos poblacionales. En particular, el estimador de la varianza del último conglomerado cumple con esta regla. El autor de la librería `survey` de `R` (Lumley 2010), plantea que cuando se trata de la estimación en dominios, la varianza involucra una cierta cantidad de ceros para todas aquellas unidades que no pertenecen al dominio. En efecto, operacionalmente, la estimación de la varianza en un subconjunto de la encuesta de hogares debe incluir varias contribuciones iguales a cero en la ecuación de estimación. Esto implica que las observaciones fuera del dominio se descartan con un filtro y las contribuciones nulas se agregan en el momento de la estimación de la varianza.

Cuando se requiere una estimación sobre un subgrupo poblacional que coincide con un estrato h' (o es una agregación de estratos), y teniendo en cuenta que los individuos de las UPM que no son de este estrato tomarán un valor nulo, tanto para la estimación del total, como para la aproximación de la varianza, entonces la forma de la aproximación se escribe de la siguiente manera:

$$\begin{aligned}
\widehat{Var}(\hat{t}_y) &= \sum_{h=1}^{H} \sum_{i \in sI_h} \frac{n_{I_h}}{n_{I_h}-1} \left(\breve{t}_{y_i} - \bar{\breve{t}}_y\right)^2 \\
&= \frac{n_{Ih'}}{n_{Ih'}-1} \sum_{i \in sI_{h'}} \left(\breve{t}_{y_i} - \bar{\breve{t}}_{y'_h}\right)^2 + \sum_{h \neq h'} \frac{n_{Ih}}{n_{Ih}-1} \sum_{i \in sI_h} \left(\breve{t}_{y_i} - \bar{\breve{t}}_{y_h}\right)^2 \\
&= \frac{n_{Ih'}}{n_{Ih'}-1} \sum_{i \in sI_{h'}} \left(\breve{t}_{y_i} - \bar{\breve{t}}_{y'_h}\right)^2
\end{aligned}$$

La anterior igualdad se tiene puesto que $\sum_{h \neq h'} \frac{n_{Ih}}{n_{Ih}-1} \sum_{i \in sI_h} \left(\breve{t}_{y_i} - \bar{\breve{t}}_{y_h}\right)^2 = 0$, dado que $\breve{t}_{y_i} = 0$ en estas UPM que no pertenecen al estrato h'. Sin embargo, esta reducción en la varianza no se tiene cuando los subgrupos poblacionales son dominios, puesto que $\breve{t}_{y_i} \neq 0$. Es decir, la cuantificación del precio que se debe pagar ante el desconocimiento de la membrecía de las unidades al subgrupo es del orden de $\sum_{h \neq h'} \frac{n_{Ih}}{n_{Ih}-1} \sum_{i \in sI_h} \left(\breve{t}_{y_i} - \bar{\breve{t}}_{y_h}\right)^2$.

III Criterios de calidad

Los criterios que aparecen en esta sección pueden ser tenidos en cuenta para determinar si una estadística debe ser considerada como precisa y confiable. En el caso particular de las proporciones, el resultado final del proceso inducirá un tabla con la estructura evidenciada en el cuadro 9.2.

Cuadro 9.2: *Estructura de resultados del procesamiento de un proporción para un país del repositorio BADEHOG.*

Subpoblaciones	Proporción estimada	Límite inferior	Límite superior	Coeficiente de variación	Efecto de diseño	Tamaño de muestra	Tamaño de muestra efectivo	Grados de libertad	Número de casos	Coeficiente de variación logarítmico
U_1	$\hat{P}_1$	LI_1	LS_1	CV_1	$DEFF_1$	n_1	n_1^*	gl_1	n_1^y	CVL_1
U_2	$\hat{P}_2$	LI_2	LS_2	CV_2	$DEFF_2$	n_2	n_2^*	gl_2	n_2^y	CVL_2
...	...	...	...	...	...	...	...	...	...	...
U_G	$\hat{P}_G$	LI_G	LS_G	CV_G	$DEFF_G$	n_G	n_G^*	gl_G	n_G^y	CVL_G
U	$\hat{P}$	LI	LS	CV	$DEFF$	n	n^*	gl	n^y	CVL

A 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 U_g, entonces un intervalo del 95% de confianza sobre esa subpoblación está dado por la siguiente expresión (Heeringa, West y Berglund 2010):

$$(\hat{\theta} - t_{0{,}975,gl} * se(\hat{\theta}), \quad \hat{\theta} + t_{0{,}975,gl} * se(\hat{\theta}))$$

En donde $\hat{\theta}$ es un estimador por muestreo para el parámetro de interés θ, $t_{0{,}975,gl}$ es el percentil 0.975 de una distribución *t-student* con gl grados de libertad, que están dados por la resta entre el número de UPM seleccionadas menos el número de estratos de muestreo considerados y $se(\hat{\theta})$ es el error estándar de la estimación, definido por la raíz cuadrada de la varianza 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 menor a cero, o mayor a uno, 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 al estimador. De esta manera, si $\hat{P}$ es una estimación de la proporción, se define la transformación Logit de la proporción.

$$\hat{L} = \log\left(\frac{\hat{P}}{1-\hat{P}}\right) = logit(\hat{P}) \tag{9.7}$$

Note que la aproximación de Taylor de primer orden para $\hat{L}$ es:

$$\hat{L} \cong L(P) + \left.\frac{\partial \hat{L}}{\partial \hat{P}}\right|_{\hat{P}=P} (\hat{P} - P) = L(P) + \left(\frac{-1}{P(1-P)}\right)(\hat{P} - P)$$

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

$$\left(\hat{L} - t_{0{,}975,gl}\sqrt{Var(\hat{L})}, \quad \hat{L} + t_{0{,}975,gl}\sqrt{Var(\hat{L})}\right) = (\hat{L}_1, \quad \hat{L}_2)$$

Finalmente, de (9.7) 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 por

$$\Big(logit(\hat{L}_1), \quad logit(\hat{L}_2)\Big) = \left(\frac{\exp(\hat{L}_1)}{1+\exp(\hat{L}_1)}, \quad \frac{\exp(\hat{L}_2)}{1+\exp(\hat{L}_2)}\right) \subseteq (0,1)$$

B Coeficientes de variación

Esta medida configura un acercamiento al error de muestreo que permite verificar si la inferencia es válida, su definición es 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 las ONE en la publicación de estadísticas oficiales. Su uso es transversal puesto que, por su definición, tiene una naturaleza relativa, liberando al usuario de la unidad de medida inducida por la varaible de interés. Además, es posible reformular los intervalos de confianza en términos del coeficiente de varaición, 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 lo afirman Singh, Westlake y Feder (2004), esta es una medida de fácil interpretación, proporcional a la amplitud del intervalo de confianza, que provee una medida estadarizada y relativa de la precisión alrededor de la estimación puntual, que permite comparar dos estimaciones del mismo indicador en diferentes sub-poblaciones, y además que es utilizada en el diseño y a re-diseño de las encuestas, entre otras cualidades. Por ejemplo, desde el punto de vista teórico, Särndal, Swensson y Wretman (2003) expresan que un estadístico puede expresar su opinión de "que un valor del coeficiente de variación del 2% es bueno, considerando las restricciones de la encuesta, mientras que un valor del coeficiente de variación de 9% puede ser considerado inaceptable." De esta forma, muchos institutos nacionales de estadística alrededor del mundo han considerado que las precisiones de las estadísticas resultantes de una encuesta estén 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 en la validación de la confiabilidad y precisión de las cifras que provienen de estudios por muestreo.

Nótese que, cuando se están estimando proporciones, esta medida tiene algunas consideraciones importantes. En primer lugar, fijar un umbral para el coeficiente de variación tiene una interpretación directa sobre la amplitud relativa del intervalo de confianza. Por ejemplo, si la ONE 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 alrededor de 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 un proporción, la varianza y el error estándar de la proporción obtiene 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, bajo este paradigma, la precisión de una proporción $P = 0,9$, es la misma que la de una proporción $P = 0,1$; de la misma manera, 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 es la varianza

y el error estándar y, por su definición, cuando la proporción es pequeña, el coeficiente de variación tiende a ser muy grande, indicando que la precisión es baja.

C El efecto de diseño DEFF

Cuando se selecciona una muestra utilizando 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 lo anterior, cuando se analizan datos que provienen de encuestas de hogares la inferencia correcta debe tener en cuenta estas grandes desviaciones con respecto al análisis estadístico clásico, que considera muestras aleatorias simples. Por ello, en la mayoría de ocasiones se necesita aumentar el tamaño de muestra para obtener la precisión deseada. Una forma sencilla de incorporar el efecto del diseño complejo está dada por la siguiente relación, denotada como efecto de diseño (Kish 1965):

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

En donde $Var(\hat{\theta})$ denota la varianza de un estimador $\hat{\theta}$ bajo un diseño de muestreo complejo P y $Var_{MAS}(\hat{\theta})$ denota la varianza del este estimador $\hat{\theta}$ bajo 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 un total, un promedio, una proporción, una razón, un percentil, etc.). Por ejemplo, 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 bajo el diseño de muestreo complejo como

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

En donde S_y^2 corresponde a la varianza de la características de interés, N es el tamaño de la población de interés 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 de bajo de la línea de pobreza en un país), es posible escribir la varianza del estimador bajo el diseño de muestreo complejo como

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

Por otro lado, en un diseño de muestreo bietápico, el efecto de diseño también se puede expresar de la siguiente manera

$$DEFF = 1 + (\bar{m} - 1)\rho_y$$

donde $\bar{m}$ corresponde al promedio de hogares seleccionados por UPM y ρ_y es el coeficiente de correlación intraclase entre la variable de interés y las UPM. En general, nótese que el efecto de diseño será mayor cuando:

1. El coeficiente de correlación crezca, lo cual no puede ser controlado 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 en el 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, entre más segregación haya, mayor será su valor.
2. El promedio de hogares seleccionados por UPM ascienda. Esto es controlado de antemano en la etapa de diseño y será un número fijo y transversal en la encuesta.

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

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

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

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

Siendo n_h el tamaño de la muestra en el estrato h, N_h el tamaño poblacional del estrato h y $S^2_{y_h}$ la varianza muestral de la variable de interés restringida al subgrupo h. Cuando la subpoblación de interés no es un estrato o un post-estrato sino un subgrupo aleatorio - como 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 - cuyo tamaño de muestra no es fijo (o condicionalmente fijo por la calibración) sino aleatorio, entonces la estimación correcta del efecto de diseño es la siguiente:

$$DEFF = \frac{Var(\hat{\theta}_h)}{Var^U_{MAS}(\hat{\theta}_h)}$$

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

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

En donde $S^2_{y_{hU}}$ es la varianza muestral de la variable de interés calculada en toda la población. Por lo tanto, en ambos efectos de diseño, la estimación de la varianza del diseño de muestreo complejo $Var(\hat{\theta}_h)$ es la misma, pero el denominador cambia dependiendo de si el subgrupo es un estrato o no. Es por esta razón que en los software computacionales las cifras relacionadas con la estimación puntual, errores estándar, intervalos de confianza y coeficientes de variación coinciden plenamente. Sin embargo, tanto los software `Stata` como `SPSS` estiman por defecto el DEFF. Nótese que, en este caso, las estimaciones de $Var^U_{MAS}(\hat{\theta}_h)$ y $Var^h_{MAS}(\hat{\theta}_h)$ serán diferentes, puesto que la primera involucra a toda la muestra, mientras que la segunda involucra únicamente a la muestra del estrato. Retomando el ejemplo, cuando los subgrupos de interés son agregaciones de los estratos de diseño, no es correcto utilizar el enfoque que `Stata` trae por defecto. En efecto, Lumley (2010) afirma que el efecto del 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 re-escalado o no son recíprocos 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.

D Tamaño de muestra

El tamaño de 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 muestra se hace más grande. Un adecuado tamaño de muestra garantiza la convergencia en distribución de los estimadores a la distribución teórica de donde se calculan los percentiles en el cálculo del intervalo de confianza. En la fase de diseño, es posible mostrar que 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 menor a $\delta \in (0, 1)$ y una confianza estadística mayor a $1 - \alpha$, está dado por la siguiente expresión.

$$n \geq \frac{S^2_y \; DEFF}{\dfrac{\delta^2 \; \bar{y}^2}{z^2_{1-\alpha/2}} + \dfrac{S^2_y \; DEFF}{N}}$$

En donde $z_{1-\alpha/2}$ es el percentil $(1-\alpha/2)$ asociado a una distribución normal estándar. Nótese que si ρ_y es grande, entonces el valor del efecto de diseño DEFF también lo será y por consiguiente el tamaño de 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 el tipo de hogar está altamente asociado con el ingreso de los individuos. Esto quiere decir que los ingresos no están uniformemente dispersos a través de todos los hogares, y por ende el coeficiente de correlación intraclase será alto. Por otro lado, si lo que se quiere estimar es una proporción P, entonces la expresión apropiada para calcular el tamaño de muestra estará dada por

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

Como se puede apreciar, el tamaño de muestra es un indicador de la calidad de la encuesta, el cual resulta ser muy importante en la etapa de planeación y diseño. Sin embargo se tiene que considerar que:

- Si el parámetro de interés **sí** fue tenido en cuenta en la planeación de la encuesta con el propósito de tener representatividad sobre una subpoblación, entonces el tamaño de muestra será apropiado y, por ende, el error de muestreo estará controlado, al igual que el coeficiente de variación, el intervalo de confianza y la precisión de la inferencia será óptima.
- Si el parámetro de interés **sí** fue tenido en cuenta en la planeación de la encuesta, pero hubo una alta tasa de no respuesta, entonces el tamaño de muestra será mucho menor al 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 ancho, haciendo que la precisión de la inferencia no sea apropiada.
- Si el parámetro de interés **no** fue contemplado en la planeación y diseño de la encuesta de hogares, entonces es posible que el tamaño de muestra sea menor al necesario y, por ende, el error de muestreo será mayor, el coeficiente de variación será mayor, el intervalo de confianza será más amplio y la precisión de la inferencia será deficiente.

E Tamaño de muestra efectivo

El principio general detrás de esta medida está supeditado a que en la inferencia propia de las encuestas de hogares con diseños de muestreo complejos no existe una sucesión de variables que sean independientes e idénticamente distribuidas. Por lo tanto, si se piensa en la muestra $(y_1, \ldots, y_n)$ como un vector en el espacio n-dimensional, el estándar clásico de la teoría estadística asumiría 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 asociados 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 (UN 2005, capítulo 6) 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 en que, si la variable de interés tiene una alta correlación intraclase, entonces la realidad de las personas y de los hogares dentro de una misma UPM será muy homogénea, tanto que se podría interpretar como que la información estuviese repetida, y que los individuos u hogares de una misma UPM no estuvieran aportando de manera diferenciada. Por lo tanto, debido a los efectos del diseño de muestreo complejo, la cantidad de individuos que están aportando a la inferencia del indicador no es el número de personas, ni el número de hogares en la muestra, sino el tamaño de muestra efectivo n_{eff}, que deflacta los efectos de aglomeración.

F Grados de libertad

La amplitud del intervalo de confianza de un indicador no sólo está supeditada al error estándar, sino también al percentil de la distribución $t - student$ con sus correspondientes grados de libertad. De esta manera, entre 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 donde la subpoblación sea toda la población objetivo, los grados de libertad se reducen 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 componen el conglomerado, el número de unidades independientes será únicamente el número de UPM seleccionadas en al primera etapa de muestreo puesto 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 donde 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. Nótese la importancia de utilizar la distribución $t - student$ como base inferencial para la construcción de los intervalos de confianza. Por ejemplo, considere el

percentil 0.975 para el cual los valores críticos de la distribución t varían con respecto a sus grados de libertad; en este contexto se tiene que $t - student_{gl=1} = 12,7$, $t - student_{gl=2} = 4,30$, $t - student_{gl=5} = 2,57$, $t - student_{gl=40} = 2,02$ y $t - student_{gl=\infty} = Z = 1,96$.

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, p. 209) proponen el siguiente método de cálculo sobre los grados de libertad en una subpoblación U_g:

$$gl_g = \sum_{h=1}^{H} v_h * (n_{Ih}^g - 1)$$

En donde v_h es una variable indicadora que toma el valor uno si el estrato h contiene uno o mas casos de la subpoblación de interés y toma el valor cero en otro caso, n_{Ih}^g es el número de unidades primarias de muestreo en el estrato h $(h = 1, \ldots, H)$ con uno o más casos de la subpoblación.

G Coeficiente de variación logarítmico

El coeficiente de variación es una medida que define la precisión de un indicador, pero para el caso de las proporciones no constituye una medida simétrica, como sí lo es el error estándar o la varianza. Por ejemplo, suponga que se está estimando una proporción P, si la estimación del parámetro de interés es muy cercana a cero, sin importar que tan pequeña 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 se traduce en una paradoja, puesto que el mismo fenómeno está siendo medido, pero los coeficiente de variación son contradictorios. Debido a lo anterior, 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 al rededor de $P = 0,5$, que es donde se maximiza la variabilidad de la proporción (Barnett-Walker y col. 2003). Por ende, si $P \leq 0,5$, se define $\hat{L} = -\log(\hat{P})$. En este caso, la aproximación de Taylor de primer orden es:

$$\hat{L} \cong L + \left.\frac{\partial \hat{L}}{\partial \hat{P}}\right|_{\hat{P}=P} (\hat{P} - P) = L + \left(\frac{-1}{P}\right)(\hat{P} - P)$$

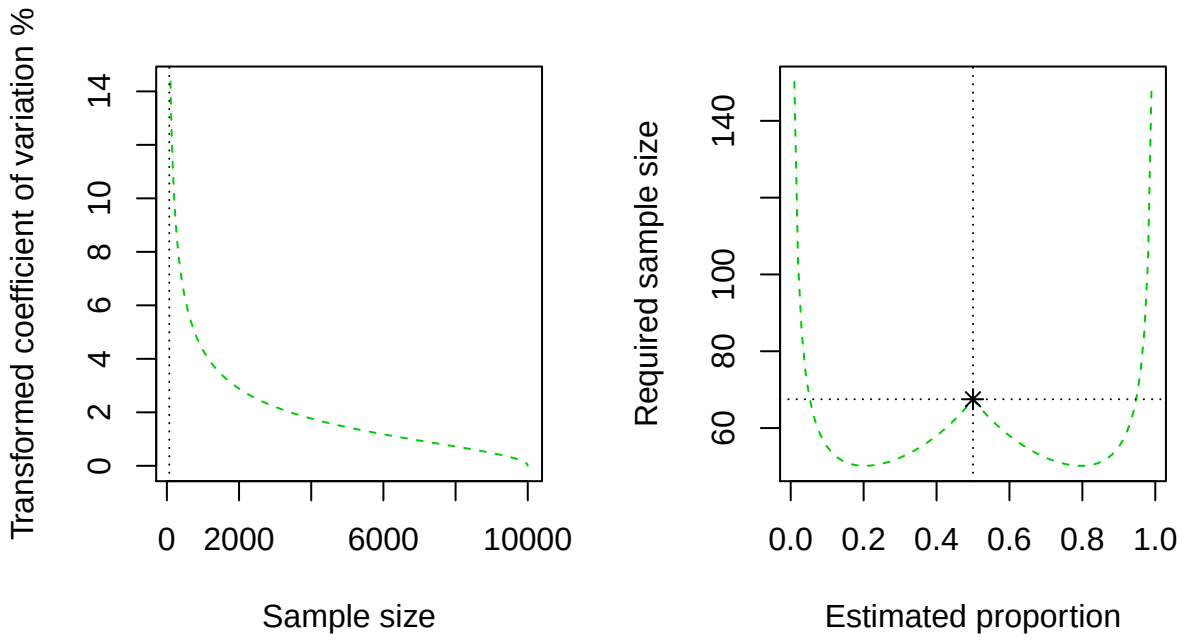

Figura 9.1: *Relación entre el tamaño de muestra y la precisión de un indicador utilizando la transformación Logit.*

Luego, la varianza de $\hat{L}$ será $Var(\hat{L}) \cong AV(\hat{L}) = \dfrac{Var(\hat{P})}{P^2}$, 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, podemos definir una medida de suavizamiento 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 uno cuando $P \in (0,2 \ \ , \ \ 0,8)$, y en este caso el $CV(\hat{L})$ será similar a $CV(\hat{P})$.

La figura 9.1 muestra que, al igual que con el coeficiente de variación original, el tamaño de muestra aumentará a medida que se requiera mayor precisión en la estimación; pero 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 eficientemente la proporción P con una precisión mayor a un determinado umbral del coeficiente de variación δ es:

$$n \geq \frac{P\ (1-P)\ DEFF}{\frac{P\ (1-P)\ DEFF}{N} + \log^2\ (P)\ P^2\ \delta^2} \tag{9.8}$$

La expresión (9.8) se obtiene teniendo en cuenta el siguiente desarrollo algebraico. En particular, cuando $P > 0,5$, se desea que el coeficiente de variación logarítmico sea menor a un umbral δ y, por lo tanto, habiendo definido $S^2 = P\ (1-P)\ DEFF$, se tiene el siguiente conjunto de implicaciones.

$$\begin{aligned}
CV(\hat{L}) \leq \delta \Longrightarrow \quad & \frac{SE(\hat{P})}{-\log(1-\hat{P})(1-\hat{P})} \leq \delta \\
\Longrightarrow \quad & \frac{\sqrt{\frac{1}{n}(1-\frac{n}{N})S^2}}{-\log(1-\hat{P})(1-\hat{P})} \leq \delta \\
\Longrightarrow \quad & \frac{1}{n}(1-\frac{n}{N})S^2 \leq \delta^2(1-\hat{P})^2(-\log(1-\hat{P}))^2 \\
\Longrightarrow \quad & \frac{1}{n} - \frac{1}{N} \leq \frac{\delta^2(1-\hat{P})^2(\log(1-\hat{P}))^2}{S^2} \\
\Longrightarrow \quad & \frac{1}{n} \leq \frac{\delta^2(1-\hat{P})^2\log^2(1-\hat{P})}{S^2} + \frac{1}{N} \\
\Longrightarrow \quad & \frac{1}{n} \leq \frac{N\ \delta^2(1-\hat{P})^2\log^2(1-\hat{P}) + S^2}{S^2 N} \\
\Longrightarrow \quad & n \geq \frac{S^2 N}{N\ \delta^2(1-\hat{P})^2\log^2(1-\hat{P}) + S^2} \\
\Longrightarrow \quad & n \geq \frac{S^2}{\delta^2(1-\hat{P})^2\log^2(1-\hat{P}) + \frac{S^2}{N}}
\end{aligned}$$

Análogamente, cuando $P \leq 0,5$, se tiene que

$$
\begin{aligned}
CV(\hat{L}) \leq \delta \implies & \quad \frac{SE(\hat{P})}{-\log(\hat{P})\hat{P}} \leq \delta \\
\implies & \quad \frac{\sqrt{\frac{1}{n}(1-\frac{n}{N})S^2}}{-\log(\hat{P})\hat{P}} \leq \delta \\
\implies & \quad (-\log(\hat{P})\hat{P}\delta)^2 \leq \frac{1}{n}(1-\frac{n}{N})S^2 \\
\implies & \quad (\log(\hat{P}))^2\hat{P}^2\delta^2 \leq (\frac{1}{n}-\frac{1}{N})S^2 \\
\implies & \quad \frac{\log^2(\hat{P})\hat{P}^2\delta^2}{S^2} \leq \frac{1}{n}-\frac{1}{N} \\
\implies & \quad \frac{\log^2(\hat{P})\hat{P}^2\delta^2}{S^2}+\frac{1}{N} \leq \frac{1}{n} \\
\implies & \quad n \geq \frac{S^2N}{\log^2(\hat{P})\hat{P}^2\delta^2N+S^2} \\
\implies & \quad n \geq \frac{S^2}{\frac{S^2}{N}+\log^2(\hat{P})\hat{P}^2\delta^2}
\end{aligned}
$$

H 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 son afectados por un fenómeno de interés en estudio. Esta cifra está supeditada únicamente a razones y proporciones y tiene un efecto indirecto en la determinación de la precisión del estimador de interés y está determinada por la siguiente expresión.

$$n_y = \sum_s \delta_k^y$$

En donde δ_k^y es una variable indicadora sobre cada individuo k de la muestra s que toma el valor de uno si el individuo está afectado por el fenómeno inducido por la variable de interés y. Nótese que esta es una cantidad aleatoria por definición, y 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 cero), tanto el coeficiente de variación original y 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 usadas 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 susceptible 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 presupuestal para atender a esta población. Por ejemplo, si la estimación del total de personas afectadas por el fenómeno es del orden de 5% y su margen de error es 5%, entonces el coeficiente de variación será de 100% y el intervalo de confianza de la proporción será $(0\%, 10\%)$, 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.

IV Consideraciones adicionales

En esta sección se abordan temáticas referentes al procesamiento apropiado de los indicadores de interés en las subpoblaciones y, habiendo realizado las estimaciones y calculado sus respectivos criterios de calidad, se plantea la utilización de umbrales apropiados para la supresión, revisión o publicación de cifras.

A Variables y subpoblaciones

El análisis apropiado de las estadísticas generadas a partir de las encuestas de hogares debe pasar por una definición clara tanto de las subpoblaciones sobre las cuales 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, por ende, es posible que se dé lugar a confusiones. Para aclarar esto, se proponen a continuación algunos ejemplos que permiten dilucidar el cálculo de las medidas de calidad sobre un conjunto no exhaustivo de indicadores de interés.

Promedio del ingreso per cápita en el país

En este caso la variable de interés es una característica continua $y_k \geq 0 \quad (\forall k \in U)$ definida sobre toda la población del país, mientras que el indicador se escribe como una razón.

$$\hat{\bar{y}}_{nacional} = \frac{\hat{t}_y}{\hat{N}} = \frac{\sum_h \sum_i \sum_k w_k y_k}{\sum_h \sum_i \sum_k w_k}$$

En donde los subíndices h, i y k, se refieren a los estratos, las UPM y los individuos, respectivamente. 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 lo individuos de la muestra. En este caso se tiene que los grados de libertad corresponden a todas las UPM menos todos los estratos de la encuesta en el país.

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.

$$\hat{\bar{y}}_{ciudad} = \frac{\hat{t}_{y_g}}{\hat{N}_g} = \frac{\sum_h \sum_i \sum_k w_k z_{g_k} y_k}{\sum_h \sum_i \sum_k w_k z_{g_k}}$$

Nótese que la variable que define la subpoblación es dicotómica dada por

$$z_{g_k} = \begin{cases} 1, \text{ Si } k \text{ reside en la ciudad } U_g \\ 0, \text{ en otro caso} \end{cases}$$

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 en la ciudad menos todos los estratos en la ciudad.

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.

$$\hat{P}_{urbano} = \frac{\hat{t}_{y_g}}{\hat{N}_g} = \frac{\sum_h \sum_i \sum_k w_k z_{g_k} y_k}{\sum_h \sum_i \sum_k w_k z_{g_k}}$$

En donde y_k es la variable de interés que define una característica dicotómica de la siguiente manera

$$y_k = \begin{cases} 1, & \text{si el ingreso per cápita de la persona está por debajo de la línea de pobreza} \\ 0, & \text{en otro caso.} \end{cases}$$

Las mediciones se realizan sobre la subpoblación definida por la siguiente variable

$$z_{g_k} = \begin{cases} 1, & \text{si la persona reside en el área urbana } U_g \\ 0, & \text{en otro caso.} \end{cases}$$

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.

Tasa de desocupación nacional

Este indicador está definido como la división entre el total de personas desocupadas sobre 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:

$$\widehat{TD}_{nacional} = \frac{\hat{t}_y}{\hat{t}_z} = \frac{\sum_h \sum_i \sum_k w_k z_k y_k}{\sum_h \sum_i \sum_k w_k z_k}$$

En donde la variables de interés toman la siguiente forma

$$y_k = \begin{cases} 1, & \text{si el individuo es desocupado,} \\ 0, & \text{si el individuo no es desocupado,} \\ NA, & \text{si no está en edad de trabajar.} \end{cases}$$

Y la variable que define la subpoblación es

$$z_k = \begin{cases} 1, & \text{si el individuo es activo,} \\ 0, & \text{si el individuo es inactivo,} \\ NA, & \text{si el individuo no está en edad de trabajar.} \end{cases}$$

En este caso el tamaño de muestra es $n = \sum_s z_k$, es decir el número de personas en la muestra que están en edad de trabajar y son 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 activas. Nótese que estos son los mismos grados de libertad inducidos por la tasa de ocupación. Además, el conteo de casos no ponderado corresponde al número de individuos desocupados en la muestra.

Tasa de desocupación masculina en migrantes

Este indicador está definido como la división entre el total de hombres migrantes desocupados sobre el total de hombres migrantes activos. El estimador del indicador está definido como una razón de dos estimadores de totales poblacionales:

$$\widehat{TD}_{hombre-migrante} = \frac{\hat{t}_{y_g}}{\hat{t}_{z_g}} = \frac{\sum_h \sum_i \sum_k w_k z_{g_k} y_k}{\sum_h \sum_i \sum_k w_k z_{g_k}}$$

En donde la variables de interés toman la siguiente forma

$$y_k = \begin{cases} 1, & \text{si el individuo es desocupado,} \\ 0, & \text{si el individuo no es desocupado,} \\ NA, & \text{si no está en edad de trabajar.} \end{cases}$$

Y la variable que define la subpoblación es

$$z_{g_k} = \begin{cases} 1, & \text{si el individuo es activo, hombre y migrante,} \\ 0, & \text{si el individuo es inactivo, hombre y migrante} \\ NA, & \text{si el individuo no está en edad de trabajar, o es mujer o es no migrante.} \end{cases}$$

En este caso el tamaño de muestra es $n = \sum_s z_{g_k}$, es decir el número de personas en 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 en 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.

B Secuencia lógica para crear reglas de supresión

En este documento se ha querido enfatizar el hecho de que la precisión de una estimación recae directamente en los intervalos de confianza, los cuales pueden ser descompuestos en elementos fundamentales que permiten crear una secuencia lógica de revisión, publicación o supresión de cifras. Nótese que lo anterior se basa en que la longitud de los intervalos de confianza induce la seguridad de que un estimador es o no es preciso. Considere los siguientes ejemplos prácticos:

- La incidencia de la pobreza en un departamento de un país se estimó en $5,2\%$, con un intervalo de confianza de $(5,15\%,\ 5,25\%)$.
- La tasa de desocupación en el país para los hombres se ubicó en $7,5\%$, con un intervalo de confianza de $(7,1\%,\ 7,9\%)$; mientras que para las mujeres se ubicó en $9,2\%$, con intervalo de confianza de $(8,8\%,\ 9,6\%)$.
- La tasa de asistencia neta estudiantil en primaria para el último quintil de ingreso se estimó en 85%, con un intervalo de confianza de $(48,2\%,\ 100,0\%)$.

Claramente, en la última situación ejemplificada, el intervalo de confianza no brinda la precisión adecuada para que una Oficina Nacional de Estadística publique esta cifra confiadamente, o para que un gobierno pueda realizar algún tipo de política pública educativa, y mucho menos para estimar los recursos que una de intervención estatal sobre la población de interés. Como se ha descrito a lo largo de este documento, utilizar únicamente el coeficiente de variación como estándar para la supresión de cifras es un

criterio que no tiene en cuenta toda las variantes asociadas a la inferencia en un muestreo complejo. A continuación se incorporan algunas recomendaciones internacionales que incorporan otros criterios adicionales a este.

- **Coeficiente de variación**: CEPAL (2018b) realizó una revisión de las experiencias internacionales, con base en la información publicada en las páginas web de las ONE, para determinar cómo son usados los criterios de supresión de información y los umbrales que las oficinas nacionales de estadística definen para la validación de las cifras. Para encuestas de hogares, se encontró que Estados Unidos y los países del Mercosur utilizan un umbral de $CV > 30\%$, Canadá y México usan como referencia un umbral del $CV > 25\%$, Chile y Costa Rica utilizan un umbral del $CV > 20\%$, Ecuador y Perú utilizan un umbral del $CV > 15\%$, mientras que Colombia usa un umbral del $CV > 10\%\%$. De esta forma, cualquier cifra estimada cuyo coeficiente de variación sea mayor al umbral predefinido es suprimida o marcada como una cifra poco confiable.
- **Tamaño de muestra**: este criterio debe ser considerado como uno de los más importantes a la hora de decidir la ruta 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 tanto el tamaño de la subpoblación como su tamaño de muestra asociado no sean pequeños. En este espíritu, Barnett-Walker y col. (2003) proponen que todas las estimaciones basadas en un tamaño de muestra menor a 100 unidades deberían ser suprimidas o marcadas como no confiables.
- **Tamaño de muestra efectivo**: al igual que con el anterior criterio, el tamaño de muestra efectivo induce que las aproximaciones teóricas, en términos de convergencia de las distribuciones de los estimadores y la cobertura de los intervalos de confianza, se cumplan. Hornik y col. (2002) consideran que si el tamaño de muestra efectivo no es mayor a 140, entonces la cifra no debería ser considerada para publicación. Por otro lado, teniendo en cuenta el tamaño de muestra inducido por la transformación logarítmica, Barnett-Walker y col. (2003) afirman que cuando la proporción se encuentra entre $0,05$ y $0,95$, entonces el tamaño de muestra efectivo, dado por la expresión (9.8), es máximo cuando $P = 0,5$, siendo su valor $n_{eff} = 68$, tal como se puede ver en la figura 9.1.
- **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, entonces es posible que las estimaciones asociadas a tamaños, totales y proporciones sobre este fenómeno no sean confiables. En particular, para las proporciones es posible restringir las estimaciones tales que $\hat{P} < 0,001$, pero es más expedito crear una regla a partir del conteo de casos en la muestra. Por ejemplo, National Research Council (2015) plantea que si el número de casos no ponderados es menor a 50 unidades entonces la estimación no es publicada.
- **Grados de libertad**: este criterio apunta a aislar el efecto inflacionario del tamaño de 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 inducidos por la subpoblación son menos de ocho, la cifra debería ser suprimida.

- **Coeficiente de variación logarítmico**: Esta medida de suavizamiento toma valores altos cuando las proporciones estimadas están demasiado cercanas a cero o a uno. Barnett-Walker y col. (2003) proponen que la cifra debe ser suprimida si el coeficiente de variación logarítmico es mayor que 17.5%.

Nótese que los criterios mencionados en este documento no deberían ser aplicados de manera independiente, sino que tendrían que seguir cierta lógica, puesto que 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 cual 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 debería ser suprimida por tener un tamaño de muestra insuficiente, pero a la vez, que la cifra debería ser publicada, por tener un tamaño de muestra efectivo suficiente. Lo anterior, podría llevar a contradicciones por parte de las ONE y malas interpretaciones por parte los usuarios finales de los datos.

De manera general, se recomienda que las ONE estudien a profundidad sus políticas de supresión, revisión y publicación de cifras en cada una de las encuestas que realiza y, de manera independiente, defina las reglas apropiadas para cada caso y que los criterios de supresión sean plasmados en forma de diagrama de flujo en la documentación de las encuestas. Además, cada encuesta debería considerar un algoritmo de forma particular; es decir, los criterios de supresión no necesariamente deben coincidir para cada operación estadística.

Por ejemplo, la figura **??** muestra una propuesta preliminar, para la estimación de proporciones o razones, en cuanto a los criterios de supresión de cifras. En una primera instancia se realiza la estimación clásica de los parámetros de interés y se genera una tabla que adjunte el cálculo de todos los criterios descritos anteriormente. Luego, dependiendo 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 es decidir, para 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. Por ejemplo, en el diagrama propuesto se definen 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 los primeros cuatro criterios no se satisface, entonces la cifra es suprimida.

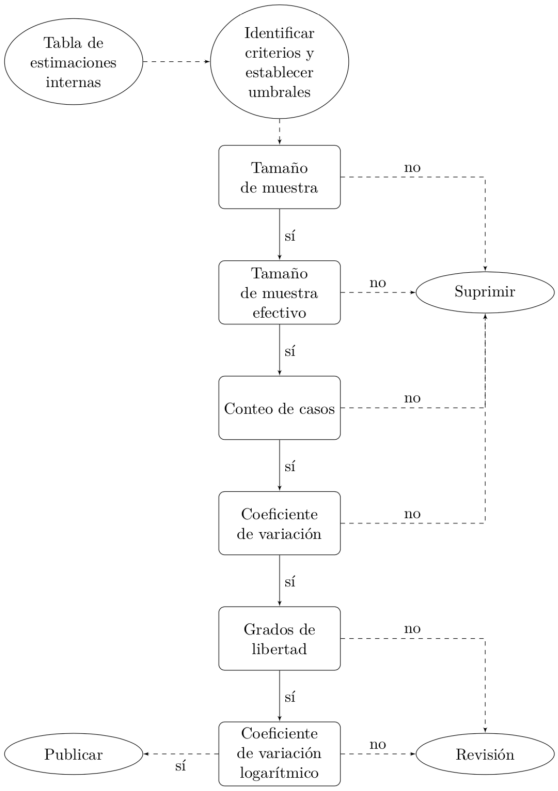

Figura 9.2: *Diagrama de flujo propuesto para la publicación, supresión y revisión de estimaciones de proporciones o razones en encuestas de hogares.*

Capítulo 10

Desafios futuros

La forma de medición de los indicadores sociales debe estar alineada con el diseño de la encuesta. Los equipos técnicos de los países deben ajustar sus metodologías a los requerimientos de la encuesta para proveer estadísticas oficiales que sean no sólo confiables y precisas sino eficientes en términos de los recursos que se destinan a la recolección de la información primaria, máxime cuando estos muchas veces se deben recortar por limitaciones en el presupuesto de los países.

En Latinomérica se observa un incremento progresivo de la tasa de ausencia de respuesta debido al aumento de 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 que se supone que deberían entrevistar, o en vez de una vivienda ahora encuentran un edificio de apartamentos. Con el uso, cada vez más frecuente de los dispositivos de almacenamiento electrónico y sistemas CAPI, es posible realizar un análisis mejor estructurado de los instrumentos de recolección de información. Por ejemplo, es posible programar saltos más complejos, estimar el tiempo promedio de respuesta en las preguntas del cuestionario, el tiempo promedio de respuesta en los bloques de preguntas, entre otras.

En el seguimiento a las metas de la Agenda 2030 y en la búsqueda del cumplimiento de los ODS, las Naciones Unidas ha expresado la necesidad de contar con estadísticas oficiales, no sólo a nivel nacional, sino a nivel de desagregaciones geográficas o de categorías demográficas de interés (ONU 2017). Es así como se plantea la siguiente necesidad a los países:

> *La disponibilidad de datos desagregados, oportunos y de alta calidad es vital para la toma de decisiones basada en la evidencia y para garantizar la responsabilidad de la implementación de la Agenda 2030. El seguimiento del progreso en los Objetivos de Desarrollo Sostenible requiere una cantidad sin precedentes de datos y estadísticas en todos los niveles, lo que plantea un*

> *desafío importante para los sistemas estadísticos nacionales e internacionales. La comunidad estadística mundial está trabajando para modernizar y fortalecer los sistemas estadísticos para abordar todos los aspectos de la producción y el uso de datos para el desarrollo sostenible.*

Por supuesto, la región no es ajena a este reto y debe desde ya plantearse la necesidad de crear capacidad en los equipos técnicos de los INE que deberán utilizar como insumo las encuestas de hogares y los registros administrativos para poder establecer metodologías de estimación en las desagregaciones de interés, que como mínimo deberán clasificarse por sexo, grupos etarios, ingreso, raza, etnia, estado migratorio, discapacidad y localización geográfica (ONU 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 Sinngh, Gambino y Mantel (1994) en donde se plantean algunas consideraciones para la estimación de indicadores sociales a nivel de dominios pequeños. Sin embargo, más allá de considerar el diseño metodológico, ahora es posible considerar otro tipo de acercamientos inferenciales que pretenden estimar indicadores a nivel de estos dominios pequeños. Rao y Molina (2014) proveen un resumen exhaustivo de las técnicas más usadas en la diseminación de estadísticas oficiales en desagregaciones pequeñas. En América Latina, este tipo de metodologías, que utilizan como insumo 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 de Bolivia utilizando los datos del censo poblacional del 2001; Araujo (2007) resume la experiencia ecuatoriana de la estimación de la pobreza en los municipios, cantones y provincias; Lopez-Calva, Rodríguez-Chamussy y Székely (2007) presentan la estimación de indicadores de desarrollo humano utilizando estimación de áreas pequeñas en las municipalidades de México a través de la *Encuesta Nacional de Ingresos y Gastos de los Hogares*. Finalmente, Casas-Cordero Valencia, Encina y Lahiri (2016) presentan un ejercicio de estimación de la pobreza en Chile utilizando la encuesta *Encuesta de Caracterización Socioeconómica Nacional*.

Finalmente, vale la pena mencionar que 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 evitar que las UPMs no tengan el mismo tamaño dentro de los estratos. Por ejemplo, en la ruralidad se pueden presentar casos en donde una única UPMs agrupa un conjunto de viviendas con demasiada heterogeneidad. Es así como es posible encontrar UPMs con pocas viviendas o UPMs 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 UPMs trae varias consecuencias negativas. Por ejemplo, las estimaciones de las varianzas son mucho más grandes y por ende las cifras oficiales serán menos precisas, necesitándose un tamaño de muestra más amplio para satisfacer un umbral de error de muestreo. Lo anterior se basa en que,

en particular para la estimación de un total poblacional, el error de muestreo está en función de la siguiente expresión

$$Var(\hat{t}_y) \approx \frac{N_I^2}{n_I}\left(1 - \frac{n_I}{N_I}\right) S_{t_{U_I}}^2$$

En donde N_I es número de UPMs en la población, n_I es el número de UPMs seleccionadas en la muestra de la primera etapa y $S_{t_{U_I}}^2$ es la variación poblacional entre los totales de las UPMs. En particular **SSW_2003** afirman que esta varianza se incrementa cuando la variación de los tamaños (número de viviendas) de las UPMs es alta.

Béland y col. (2005) describen los principales elementos del diseño de una encuesta de hogares y esta es una tarea que deben afrontar los equipos técnicos de los INE en términos de aprender de las experiencias del pasado para mejorar los procesos operativos, metodológicos y logísticos en las siguientes aplicaciones de las encuestas. Es así como ante la nueva oleada de censos que se avecina en la próxima década, será natural actualizar los marcos de muestreo y con ello se viene un reto para los equipos técnicos encargados de la encuestas de hogares en América Latina que consiste en evaluar el impacto del cambio de los marcos de muestreo y sus efectos en la comparabilidad de las cifras oficiales.

Capítulo 11

Documentación de los diseños de muestreo en las encuestas de hogares

El presente documento presenta las acciones que se deben tomar por parte de los técnicos dentro de las oficinas nacionales de estadística para el diseño de una encuesta de hogares. Se trata de una guía que recoje las experiencias internacionales encontradas por el asesor regional que puede ser mejorada o complementada con otras experiencias o actualizaciones teóricas. Este documento pretende definir un proceso que debe ser seguido en estricto orden para documentar correctamente el diseño de muestreo de las encuestas de hogares. Por su naturaleza, el docuemnto no pretende abordar detalles teóricos, ni tampoco discusiones epistemológicas de los métodos.

Características y objetivos de la encuesta

Una de las tareas más importantes en una encuesta es formular sus objetivos. Esto establece no solo las necesidades de información de la encuesta, sino también las definiciones operativas que se utilizarán, los temas específicos que se abordarán y el plan de análisis. Este paso de la encuesta determina qué se debe incluir en la encuesta y qué debe excluirse.

1. Especificar los objetivos de la encuesta tan minuciosa y claramente como sea posible.
    - Detallar el objetivo general y los objetivos específicos.
2. Defina los principales conceptos (pobreza, ingreso, empleo, ocupación, consumo, hogar, vivienda, etc.).
3. Hacer una breve descripción de los módulos de la encuesta.
4. Establezca el período de referencia y la cobertura geográfica para la encuesta.

   - Es posible que la encuesta esté compuesta por módulos que resultan ser representativos para algunas desagregaciones geográficas y para otras no.
5. Determine los usuarios y usos de la información que producirá la encuesta.
6. Determine el periodo de levantamiento de la encuesta y el método de recolección que se quiere utilizar.

Definición de la población objetivo

La población objetivo debe ser definida de forma clara, concisa y escueta en función de los objetivos de la encuesta. Se deben establecer las definiciones pertinentes que estén involucradas en la medición y observación de las unidades de interés. En este sentido es pertinente definir quiénes son los sujetos objeto de medición. Una primera definición que vale la pena aclarar es la de **subpoblaciones de interés** que denota aquellos subconjuntos de la población objetivo para los que se quiere una cierta representatividad. Por ejemplo, **INE_ES_2002** define como subpoblaciones de interés a los subestratos definidos por condición de ocpuàción y sector (agricultores, trabajadores cuenta propia, directivos y profesionales, etc).

1. Describa la población objetivo.
   - En algunas encuestas la unidad objetivo es el hogar, en otras, como encuestas de empleo la población objetivo son las personas mayores de 15 años.
2. Defina las unidades de observación y su ubicación geográfica.
3. Establezca quién contestará la encuesta en el hogar En algunas enceustas se usa un informante adecuado que cumple ciertas características.
4. Descripción de las subpoblaciones de interés en donde se quiere representatividad estadística.
   - las subpoblaciones generalmente son geográficos (estados, departamentos, ciudades grandes), de zona (urbano, rural), aunque también pueden depender de la definición de la variable de interés (porcentaje de pobreza, gasto promedio del hogar en vestimenta, alimentos, licor, etc.).
   - Definir las áreas en términos del número de habitantes.
   - Es de interés tener representatividad en los cruces o sólo en los marginales de las subpoblaciones.
   - Enumerar los municipios (o agregacioenes geográficas) que deben ser autorrepresentados (siempre deben ser parte de) en la enceusta.

Definición de los parámetros y variables de diseño

1. Descripción de las variables de interés que serán medidas en la encuesta.
2. Definición de los parámetros claves que resuelven el objetivo general de la encuesta.

    - Definir si los parámetros de interés son de nivel o de cambio, o ambos.
3. Descripción y presentación de las expresiones matemáticas de los parámetros de interés.
4. Haga un primer resumen de los cruces necesarios para establecer el plan de tabulación de la encuesta.

Consolidación del marco de muestreo

En esta etapa viene a colación la definición de **estrato** que representa un subconjunto de la población de interés para el cual se ha decidido realizar un muestreo independiente de los otros estratos. En algunas ocasiones los estratos coinciden con las subpoblaciones de interés. Sin embargo, la creación de los estratos no necesariamente garantiza representativad, mientras que las subpoblaciones de interés sí lo hacen.

1. Descripción general del plan de recolección de información.
    - Dependiendo de los parámetros de interés los hogares se visitarían sólo una vez o podría haber algún tipo de rotación en la muestra.
2. Descripción de las unidades de muestreo en cada etapa del plan.
    - Conformación de las UPMs en las subpoblaciones de interés. Por ejemplo, en el sector urbano una UPM contiene entre 80 y 160 viviendas; en el complemento urbano, entre 80 y 160 viviendas; en el sector rural entre 160 y 300 viviendas.
3. Identificación de las áreas que serán excluidas de la población objetivo.
    - Áreas de dificil acceso, UPMs con un número de viviendas o hbaitantes muy pequeño (por ejemplo, menos de 10 viviendas), UPMs que estén siendo visitadas por otros operativos de campo, zonas de conflicto, etc.
    - Presentar una tabla en donde se nombre la zona, su ubicación geográfica (en qué departamento o estado se encuentra), enumerar el número de hogares descartados y el porcentaje poblacional.
4. Definición y descripción de los marcos de muestreo que se utilizarán para la selección de todas las unidades de muestreo.
    - Para los marcos de áreas cartográficas hacer un resumen de cómo fue construido, asociado a qué levantamiento censal y la última vez que fue actualizado.
    - Calcular la cobertura del marco de muestreo (diferencia porcentual entre la población de interés y la población enmarcada).
    - Recuento del número de unidades primarias de muestreo, viviendas, hogares y personas a nivel nacional.
    - Recuento del número total y proporción de unidades primarias de muestreo, viviendas, hogares y personas por cada subpoblación de interés.
        - La suma de las proporciones debería ser igual a 100% si las subpoblaciones son mutuamente excluyentes.

  - En caso contrario, cuando las subpoblaciones no están anidados, presentar tablas separadas y en cada tabla la suma porcentual debe ser 100%.

5. Descripción de las medidas de tamaño en cada marco de muestreo.
   - En general las medidas de tamaño están asociadas al número de habitantes por UPM, pero también pueden estar definidas como el número de viviendas por UPM.
   - En el caso de que la medida de tamaño se el número de vivienda, es necesario verificar que la vivienda se encuentre habitada.
   - Es necesario justificar la creación de la medida de tamaño si resulta ser función de dos o más variables del marco.
6. Descripción de la metodología de estratificación explícita de las unidades primarias de muestreo (por nivel socioeconómico, zona, etc.).
   - Verificar que todas las subpoblaciones sean estratos (pero no todos los estratos son subpoblaciones).
   - Docuemntar la creación de estratos de UPMs que no coinciden necesariamente con las subpoblaciones de interés, pero para los que se hace un muestreo independiente para reducir la varianza.
   - Definición de las agregaciones geográficas (ciudad, centro poblado, rural, resto rural) en términos del número de habitantes por zona.
     - Definición de zona (generalmente rural está supeditada a activades de autoconsumo).
     - Cuando el presupuesto es limitado, es posible estratificar implícitamente por el tamaño de la localidad (por ejemplo, localidades con más de 10mil habitantes, entre 15mil y 99mil, entre 2.5mil y 15mil, menos de 2.5mil habitantes).
     - Delimitación del área urbana (por ejemplo, centros poblados con más de 2mil habitantes)
     - Presentación de las variables de estratificaión socioeconómica (habitualmente de tenencia de bienes muebles y de necesidades básicas).
     - Presentación del método utilzado (Dalenius, Jarque, *k-means*, etc) y de las medidas de distancia utilizadas para la creación de estratos estadísticos (generalmente son función del número de viviendas en las UPM, agregaciones sociodemosgráficas y de otras variables de interés)
       * La estratificación de las UPMs debe llevarse a cabo para todo el país.
       * Presentación de escenarios de estratificación (varaindo el núemro de estratos) y justificación del escenario escogido.
       * Presentación de tabla con el número de estratos diferentes en el cruce de subpoblación de interés (Departamento) con tamaño de localidad.
       * Presentación de tablas con el número de UPM por estrato socioeconómico y subpoblación de interés (Departamento).

        * Presentación de tablas con el número de viviendas por estrato socioeconómico, subpoblación de interés (Departamento) y tamaño de localidad.
    - Presentación de un gráfico (ver Figura 1) que muestre la división del país en términos de zona, agregaciones geográficas y UPMs.
7. Descripción de la metodología de estratificación implícita

Diseño de muestreo

1. Definición del diseño de muestreo probabilístico general.
2. Definición del diseño de muestreo en cada una de las etapas o fases de la encuesta.
    - En caso de que existan variables auxiliares en el marco de muestreo de las UPMs, es posible utilizar un muestreo balanceado para equilibrar la muestra.
        – Descripción de las variables de balanceo o de las medidas de tamaño (usualmente la medida de tamaño es el número de hogares, pero las varaibles de balanceo pueden ser número de hombres, mujeres y edades por grupos quinquenales).
    - El esquema de Pareto (πPT, sin reemplazo y de tamaño de muestra fijo) induce buenas propiedades logísticas (coordinación de muestras) cuando hay varios opertaivos de campo, es óptimo y produce estimaciones de varianza válidas (a diferencia del muestreo sistemático πPT).
3. Definición de los algoritmos de selección que serán utilizados en todas las etapas, estratos o fases.
    - Citar la bibliografía de los algoritmos de muestreo y hacer un esquema general del algoritmo presentando el proceso de selección.
    - En cada etapa de selección el esquema deberá contextualizar al lector acerca del proceso de muestreo de las unidades que se están seleccionando.

Cálculo del tamaño de muestra

A nivel nacional

1. Definición del nivel de precisión a nivel nacional en términos de la variable de diseño, el parámetro de interés, su margen de error relativo, el margen de error absoluto, nivel de confianza y de los correspondientes intervalos de confianza.
    - En algunas ocasiones, dependiendo del comportamiento del parámetro de interés en las subpoblaciones, es necesario definir diferentes niveles de precisión. Por ejemplo, **Caceres_2015** considera cuatro categorias de pobreza (<2%, 2%-5%, 5%-10%, >10%) con diferentes errores relativos en cada caso (80%, 65%, 50%, 36%).

2. Definición y justificación del efecto de diseño (DEFF) y del coeficiente de correlación intraclase de la variable de diseño con las UPM del marco de áreas.
    - Por lo general, las encuestas demográficas y de migraciones normalmente requieren muestras más grandes dentro de cada UPM, ya que las características demográficas tienen una correlación intraclase más baja que las económicas.
3. Cálculo y justificación del tamaño de muestra de las unidades de observación, hogares, viviendas y unidades primarias de muestreo a nivel nacional.
4. Presentación de escenarios de muestreo variando el número promedio de hogares a nivel nacional
    - Escogencia y justificación del escenario de muestreo que mejor se ajuste a los intereses de la encuesta

A nivel de subpoblación

1. Definición del nivel de precisión a nivel nacional en términos de la variable de diseño, el margen de error relativo, el margen de error absoluto, nivel de confianza y de los correspondientes intervalos de confianza.
2. Definición y justificación del efecto de diseño (DEFF) y del coeficiente de correlación intraclase de la variable de diseño con las UPM del marco de áreas.
3. Cálculo y justificación del tamaño de muestra de las unidades de observación, hogares, viviendas y unidades primarias de muestreo por subpoblación.
    - Presentar de manera clara y exhaustiva todas las expresiones matemáticas involucradas en este cálculo.
4. Presentación de escenarios de muestreo variando el número promedio de hogares a nivel nacional
    - Incluir la fracción de muestreo esperada en cada subpoblación de interés.
    - La fracción de muestreo no debería superar el 10% o 20%.

Asignación de la muestra final

El insumo más importante para el cálculo del tamaño de muestra es el número promedio de hogares por UPM. Sin emabrgo, no todas las UPM tienen el mismo número de hogares.

1. Definir el porcentaje de sobremuestra sobre el tamaño de muestra inicial.
    - Este ajuste debería ser diferenciado cuando hay información auxiliar a nivel de subpoblación.
2. Estratificación de las UPMs por composición en términos del número de hogares.
    - Presentación de una tabla que indique el número promedio de viviendas en la población y en la muestra de las agregaciones geográficas (urbano alto, complemento urbano, rural).
    - En la tabla anterior verificar que el promedio ponderado de hogares seleccionados coincida con la cifra usada para la determinación del tamaño de muestra nacional.
3. Para los estratos que no coinciden con las subpoblaciones de interés docuemntar la

asignación de tamaño de muestra que se utilizó (Neyman, proporcional, potencia, Lavallée-Hidiroglou, Kish, etc.)

- El número de UPMs n_{Ih} que se deben seleccionar en cada estrato h está dado por la siguiente relación entre el número de hogares n_{IIh} que se debn seleccionar en el estrato h y el núemro promedio de hogares que se seleccionan en cada UPM $\bar{n}_{IIh}$

$$n_{Ih} = \frac{n_{IIh}}{\bar{n}_{IIh}}$$

  - El número de hogares n_{IIh} que se debn seleccionar en el estrato h se calcula de manera proporcional al número de hogares N_{IIh} resportados en el marco de muestreo.
- Incluir una tabla de resumen de los estratos de interés que presente el número de UPMs en el marco y el porcentaje, el núemro de UPMs seleccionadas y el porcentaje, el total de viviendas u hogares y su porcentaje.
  - Para garantizar que no haya problemas posteriores en la estimación de los errores de muestreo, es necesario que por cada estrato hayan mínimo dos UPMs.
  - En los casos en que sólo se puede seleccionar una UPM por estrato, es necesario colapsar.
- Incluir una tabla de resumen de las subpoblaciones de interés que presente el valor esperado del estimador, el tamaño de muestra de hogares y personas, el efecto de diseño, el error absoluto, el error relativo, el coeficiente de variación, el margen de error y los límites de los intervalos de confianza

4. Verificar que la ditribución (de habitantes y viviendas) en la muestra a nivel de estratos (o subpoblaciones) coincida con la distribución poblacional.
   - En algunas ocasiones las distruibuciones poblacionales y muestrales no necesariamente coincidirán. En este caso es necesario argumentar por qué.
     - La principal razón debiera ser que las subpoblaciones requieren un tamaño de muestra mínimo e independiente para garantizar la representatividad.

Definición de los factores de expansión

El siguiente diagrama presenta una visión general de la creación, tratamiento, ajuste y corrección de los factores de expansión a través del levantamiento de la ifnroación.

1. Presentación de los factores de expansión básicos inducidos por el diseño de muestreo en cada etapa y estrato.
   - Presentar los ponderadores con la notación anidada hi que quiere decir que la vivienda (o individuo) está presente en la UPM i del estrato h.
   - Presentar una tabla con la información poblacional del marco de muestreo del número de UPMs y número de viviendas por subpoblación de interés junto con la suma de los factores de expansión de UPMs y viviendas.

- Para cada etapa, verificar que la suma de los factores de expansión coincida con su contraparte poblacional.
- Presentar una tabla con suma, mínimo, máximo, media, mediana y desviación estandar de los ponderadores finales en cada subpoblación de interés.

2. Ajuste por exclusión
   - Si se ha definido que algunas áreas geográficas no serán encuestadas (o algunas UPM con un número de viviendas muy bajo) es necesario calcular un ajuste a los ponderadores de la segunda etapa de muestreo (selección de viviendas).
   - Este ajuste se realiza por estratos dividiendo el número de viviendas en el marco de muestreo para el estrato h sobre la suma ponderada por el factor de expansión de la primera etapa del número de viviendas seleccionadas en el estrato h.
3. Corrección por elegibilidad
   - Identificación de viviendas no elegibles
   - Identificación de viviendas con estatus de elegibilidad desconocida
4. Ajuste por la ausencia de respuesta
   - Creación de modelos que expliquen la probabilidad de respuesta.
   - Creación de subgrupos que expliquen el fenómeno de ausencia de respuesta (deciles a pertir de las probabilidades de predicción o clases a partir de árboles de clasificación).
5. Presentación de las variables de calibración de los factores de expansión.
   - Presentación del método escogido para calibrar (postestratificación, raking, calibración mixta) + Descripción de las pseudo-distancias de calibración y de los umbrales L y U utilizados para acotar los pesos de muestreo. + El factor de expansión calibrado w_k* se puede ver como $w_k* = w_k g_k$. Es pertinente revisar el comportamiento de los factores g_k con los mínimos, máximos, medias y medianas.
   - Descripción de las proyecciones poblacionales utilizadas para calibrar. + Presentación de las tablas con las correspondientes proyecciones demográficas en los post-estratos de interés.
   - Descripción de las variables continuas utilizadas para calibrar (por ejemplo en encuestas panel para la creación de pesos compuestos longitudinales).
6. Compilado de los factores de expansión + Presentar gráficos de boxplot para las subpoblaciones de interés de los factores de expansión en cada etapa. + Presentar histogramas suavizados de los factores de expansión nacionales.

Esquema de rotación (si aplica)

Para las encuestas que han sido planeadas de forma que sigan a los hogares a lo largo del tiempo (panel, split panel, rotativas). Ver por ejemplo **IBGE_2014**

1. Descripción y justificación del esquema de rotación
    - Definir el número de veces que un hogar es encuestado en el ciclo de vida de la encuesta.
    - Describir el porcentaje traslape de la encuesta en los diferntes levantamientos del año (mensual, trimestral, semestral y anual).
    - Incluir el esquema de rotación de forma gráfica.
2. Selección de las UPMs y hogares en los turnos de rotación
    - Descripción del algortimo que respalda el proceso de selección rotativa en línea con el esquema de rotación propuesta para la encuesta.
    - Tabla que indique el número de UPMs a visitar por semana en un trimestre agrupada para las subpoblaciones (geográficas) de interés.
3. Construcción de pesos longitudinales
4. Agregación de bases periódicamente
    - Construcción de factores de expansión agregados

Definición de los estimadores

1. Definición de expresiones teóricas de estimadores de nivel.
2. Definición de expresiones teóricas de estimadores para cambios (en encuestas rotativas).

Cálculo del error de muestreo

1. Definición de expresiones teóricas para estimar insesgadamente la varianza de los estimadores presentados en la sección anterior.
    - Escribir las fórmulas del último conglomerado para totaltes y medias.
        - Si la fracción de muestreo es significativa, utilizar la aproximación con el factor de corrección por finitud.
    - Adecuar las fórmulas del último conglomerado para variables linealizadas de razones o estimadores no lineales.
2. Descripción de los detalles computacionales para métodos de remuestreo
    - Descripción de la creación de estratos de varianza (*varstrat*) y unidades de varianza (*varunits*) en la aplicación de Jackknife o BRR.
        - Por ejemplo, optar por hacer *varstrat* agrupando UPMs en las mismas subpoblaciones de interés.
        - Por ejemplo, optar por hacer *varunits* agrupando hasta 30 viviendas en rural y hasta 50 viviendas en urbano.
    - Si se opta por BRR con ajuste de Fay, presentar la matriz de Hadamard que se utilizará y el coeficiente de Fay.
    - Presentar tabla con resumen por subpoblaciones de interés del número de estratos y UPMs originales y del número de *varstrat* y *varunits* creados.

Capítulo 12

Algunas encuestas de hogares en la región

Esta sección presenta una breve descripción del estado de la situación de las encuestas de hogares en la región. Aunque no se pretende hacer un resumen exhaustivo de cada encuesta y de sus componentes metodológicos, con seguridad el lector podrá enterarse de las características principales de la encuestas de hogares y sus condiciones de aplicación.

Argentina

El Instituto Nacional de Estadística y Censo lleva a cabo de forma trimestral la *Encuesta Permanente de Hogares* la cual 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 revela información sobre características demográficas básicas de los miembros del hogar y su situación laboral, ingresos de los individuos, 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 el relevamiento de sus gastos e ingresos. Sus resultados contribuyen con la elaboración de la canasta de bienes y servicios que se utiliza para medir el índice de precios al consumidor, así como aportan información para la estimación de la pobreza y la producción de indicadores de la economía nacional (INDEC 2018a).

Bolivia

El objetivo principal de la *Encuesta Continua de Hogares* aplicada anualmente por el Instituto Nacional de Estadística es suministrar información sobre las condiciones de vida de los hogares, a partir de la recopilación de información de variables económicas y demográficas de la población para le diseño de programas sociales y formulación, evaluación y seguimiento de las políticas públicas. Dentro de 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, los servicios básicos, la educación, entre otros. Mediante la EH, el INE obtiene estadísticas e indicadores socioeconómicos y demográficos de la población que son necesarias para la formulación, evaluación, monitoreo y seguimiento de las políticas del estado (INE 2018b).

Brasil

La *Pesquisa Nacional por Amostra de Domicílios* es implementada anualmente por el Instituto Brasileiro de Geografia e Estatística. Esta encuesta tiene como objetivo producir información básica para el estudio de la evolución económica de Brasil y la publicación continua de indicadores demográficos. Los constructos de ingreso, gastos y empleo son evaluados de forma continua, mientras que anualmente se abordan otros módulos de interés. Otros temas adicionales investigados en la encuesta están relacionados con las características de la vivienda, migración de los individuos del hogar, trabajo infantil, fecundidad, salud y seguridad alimentaria, uso de las tecnologías de información, transferencias de renta, uso del tiempo, entre otros (IBGE 2018b).

La *Pesquisa de Orçamentos Familiares* tiene como propósito obtener informaciones generales sobre domicilios, familias y personas, hábitos de consumo, gastos y recibos de las familias encuestadas, teniendo como unidad de recolección los domicilios. Permite actualizar la canasta básica de consumo y obtiene 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)

Chile

El Ministerio de Desarrollo Social aplica la *Encuesta de Caracterización Socioeconómica Nacional* de forma bianual y su objetivo es conocer periódicamente la situación de

los hogares y de la población, especialmente de aquella 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; identificar 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 de los módulos de registro, que incluye información de identificación de los hogares; educación, que indaga por la situación educacional de los miembros del hogar y la cobertura del sistema educativo; trabajo, que permite conocer la evolución de la situación laboral y ocupacional para formular y evaluar políticas públicas; ingresos, que permite investigar las condiciones de vida de los miembros del hogar; salud, en donde se indaga por la cobertura de los programas públicos (MDS 2015).

La *Encuesta de Presupuestos Familiares* es una encuesta socioeconómica aplicada a hogares, cuyo propósito es recopilar información sobre gastos en los que estos incurren y los ingresos que perciben en un período de tiempo determinado. La información que recoge la EPF es la base para elaborar la canasta de bienes y servicios con la cual se calcula el Índice de Precios al Consumidor y también se utiliza para actualizar las líneas de pobreza extrema y pobreza empleadas en las estadísticas oficiales de Chile (INE 2018c).

Colombia

El Departamento Nacional de Estadística aplica la *Gran Encuesta Integrada de Hogares* de forma mensual. Esta encuesta tiene como objetivo general proporcionar información económica básica con énfasis en las características de la fuerza de trabajo. Además indaga por constructos sociales y económicos. Dentro de la temática social, se pregunta por el acceso a la educación formal, condiciones de calidad de vida, ingresos y gastos, trabajo infantil y aspectos de seguridad y convivencia ciudadana. La temática económica indaga aspectos relacionados con industria, comercio, servicios y transporte. El instrumento de recolección de la encuesta está dividido 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, estableciendo así una caracterización general de la población. Por lo demás también indaga por el acceso a la seguridad social en salud, las características educativas de las población mayor de tres años y 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 indagan en forma detallada todos los ingresos de los miembros del hogar de 10 años y más (ingresos por trabajo, ingresos de capital, subsidios,

transferencias, ingresos ocasionales, etc) así como todos los posibles gastos en que puede incurrir un hogar, captados con diferentes periodicidades (semanal, mensual, trimestral y anual). Dentro de sus objetivos específicos está el obtener información para realizar actualizaciones del IPC, estimar líneas de indigencia y pobreza, caracterizar la distribución del ingreso del hogar con características demográficas, educativas y económicas, etc. (DANE 2018)

Costa Rica

La *Encuesta Nacional de Hogares*, llevada a cabo por el Instituto Nacional de Estadística y Censos de forma anual, tiene como objetivo producir estimaciones del nivel de bienestar de la población, especialmente centrados en la conformación del ingreso de los hogares, su distribución y características de los hogares y la población en situación de pobreza. El constructo principal y la motivación de esta encuesta está referido a la pobreza multidimensional y la desigualdad midiendo el ingreso promedio de los hogares por fuente y su distribución, su incidencia y severidad, 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 éstos 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 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 Índice de Precios al Consumidor (IPC); entre otros estudios sobre la estructura de gastos de los hogares y la distribución del ingreso (INEC 2018c).

Cuba

La Oficina Nacional de Estadística e Información aplica anualmente la *Encuesta Nacional de Ocupación* y la *Encuesta de Situación Económica de los Hogares*. Lo objetivos de estas investigaciones se centran en la caracterización de la población en edad de trabajar y su ocupación, así como en obtener indicadores para la toma de decisiones en materia de políticas sociales y económicas (ONE 2018c).

Estas encuestas investigan temas relacionados como las características de las viviendas, la educación en los hogares, las características económicas de los miembros del hogar (para las personas mayores de 15 años), la movilidad laboral y los ingresos en el hogar.

Ecuador

El Instituto Nacional de Estadística y Censos cuenta con el Sistema Integrado de Encuestas a Hogares con el cual se produce información de las características demográficas y económicas de los hogares y personas. Entre otras, el sistema cuenta con la *Encuesta de Empleo, Desempleo y Subempleo en Área Urbana y Rural*, 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. Las anteriores encuestas cuentan con temas en común, como condiciones de vivienda, caracterización de los hogares, educación de los miembros del hogar, condición de la 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 relacionando varios factores como educación, salud, pobreza e inequidad para la aplicación de política pública. La ECV 2013 – 2014 incluye 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).

El Salvador

La *Encuesta de Hogares de Propósitos Múltiples* es implementada anualmente 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 focalizar las políticas públicas del gobierno para elevar el bienestar de la población. Esta encuesta indaga por la información general de los miembros del hogar, su situación educacional (analfabetismo, escolaridad y asistencia), también pregunta por 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 recolecta información acerca de la tenencia de la tierra, superficie cultivada, y la actividad agropecuaria del entrevistado. Por último también pregunta acerca de variables de salud, dinámica de las remesas y los gastos del hogar (DIGESTYC 2018a).

La *Encuesta de Ingresos y Gastos de los Hogares* permite determinar la canasta de mercado para el desarrollo del Índice de Precios al Consumidor (IPC), el consumo privado en las Cuentas Nacionales (CN) y para 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 al hogar (DIGESTYC 2018b).

Guatemala

El Instituto Nacional de Estadística aplica 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 evolución del empleo, el desempleo, las características, composición, estructura y funcionamiento del mercado de trabajo. Además de investigar aspectos generales del mercado laboral, esta encuesta indaga las características de informalidad, ocupación y formas de contratación. Tiene un componente de ingresos, así como algunos aspectos demográficos y de educación en los hogares de Guatemala. Algunos cambios recientes han incluido módulos de uso del tiempo y uso de las tecnologías de información (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 caracterizando a la población pobre t no pobre del país, brindando resultados a nivel nacional, regional y departamental (INE 2018d).

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 recolectar información sobre las características generales de la población hondureña, en términos de vivienda, tasas de ocupación, desocupación y subempleo, ingreso en los hogares y acceso a las tecnologías de la información (INE 2018g).

La *Encuesta de Condiciones de Vida de los Hogares* es una investigacion 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 este 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 2018f).

México

La *Encuesta Nacional de Ocupación y Empleo* aplicada cada dos años por el Instituto Nacional de Estadística y Geografía 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; adicionalmente, ofrece información sobre las características

ocupacionales y demográficas de los integrantes del hogar, así como las características de la infraestructura de la vivienda y el equipamiento del hogar. Además de lo anterior, la encuesta cubre algunos constructos como percepciones y erogaciones financieras y de capital de los hogares y sus integrantes, características de la vivienda, características demográficas de los residentes de la vivienda, condición de actividad y características ocupacionales de los integrantes del hogar de 12 y más años, equipamiento del hogar y acceso a servicios (INEGI 2019).

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; adicionalmente, 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).

Nicaragua

Nicaragua lleva a cabo de forma trimestral la *Encuesta Nacional de Hogares Sobre la Medición de Niveles de Vida*, a través del Instituto Nacional de Información de Desarrollo, cuyo objetivo general es producir información continua sobre características ocupacionales, demográficas y 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 por el estado de la vivienda y sus características. Esta encuesta también dispone de las variables necesarias para la construcción de otras medidas de bienestar, como agregado de ingreso, necesidades básicas insatisfechas, etc. (INIDE 2018).

Panamá

El Instituto Nacional de Estadística y Censo de Panamá aplica anualmente 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 por la condición de 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 colaboración con actividades de protección y conservación de los recursos naturales, dinámica de turismo en los hogares, 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 2018d).

La *Encuesta de Ingresos y Gastos de los Hogares* se realiza para actualizar la información de los ingresos de los hogares en panamá y como estos distribuyen los presupuestos para obtener diferentes bienes y servicios. La información recopilada tiene como principales objetivos obtener coeficientes de ponderación y canastas de consumo que serán utilizadas para el cálculo del IPC y la CBFA. Por otra parte permite estructurar la demanda de los hogares en bienes de consumo privado (INEC 2018b).

Paraguay

La *Encuesta Permanente de Hogares* ejecutada anualmente por la Dirección General de Estadística, Encuestas y Censos tiene como objetivo la generación de indicadores anuales sobre las principales características de las condiciones de vida de la población y sus resultados son utilizados para las estimaciones de pobreza. Algunos de los constructos que investiga esta encuesta se relacionan con las características de las viviendas, educación de los miembros del hogar, salud, empleo e ingresos, condición de ocupación, 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 constituyen las líneas de pobreza, así como el de caracterizar y analizar las Condiciones de Vida de la Población del Paraguay. Esta información es recopilada 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, etc. (DGEEC 2018a).

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 para la medición de aspectos económicos y sociales; y conocer y explicar los determinantes o factores causales del comportamiento de dichos aspectos para el diseño, monitoreo y medición de resultados de las políticas públicas. Dentro de los módulos que la encuesta aborda se encuentran la caracterización de la vivienda y el hogar, la educación de los miembros del hogar, así como su estado de salud, la condición de actividad de empleo, ingresos y gastos, acceso a programas sociales, participación ciudadana, así como la percepción de la gobernabilidad y algunos tópicos concernientes al fenómeno de la discriminación (INEI 2016).

República Dominicana

La Oficina Nacional de Estadística cuenta con un Sistema Integrado de Encuestas de Hogares que agrupa, entre otras, a la *Encuesta Nacional de Hogares de Propósitos Múltiples*, 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; mientras que 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 tecnologías de 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 10 años para conocer la distribución del gasto en bienes y servicios de consumo de los hogares, así como los ingresos que éstos obtienen por diferentes fuentes para financiar su consumo. Dentro de sus objetivos se encuentra el obtener información para conocer el nivel y la estructura de los gastos de consumo de los hogares y su distribución en rubros de: alimentos y bebidas no alcohólicas; bebidas alcohólicas, tabaco y estupefacientes; prendas de vestir y calzado; alojamiento, agua, electricidad, gas, y otros combustibles; muebles, artículos para el hogar y para la conservación ordinaria del hogar; salud; transporte; comunicaciones; recreación y cultura; educación; restaurantes y hoteles y, bienes y servicios diversos (ONE 2018b).

Uruguay

El Instituto Nacional de Estadística aplica mensualmente la *Encuesta Continua de Hogares* que brinda los indicadores oficiales del mercado laboral (actividad, empleo y desempleo) y de ingresos de los hogares y las personas. Además esta encuesta permite estimar la proporción de hogares y personas por debajo de la línea de pobreza y de indigencia de forma anual. El cuestionario de la encuesta indaga por 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 tecnologías de información. De la misma forma, la encuesta aborda 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 10 años y permite conocer la realidad económica y social del país. A partir de los resultados obtenidos, se podrá elaborar una canasta actualizada para el Índice de Precios al Consumo (IPC) y también determinar las líneas de indigencia y de pobreza nacionales. Esta es 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 son los de inflación y pobreza (INE 2016b).

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, así como de las características económicas y demográficas de la población. Algunos de las temáticas más relevantes de esta encuesta se centran en la actividad económica de los miembros del hogar y su estado de empleo, así como la caracterización de las viviendas y de los hogares y algunas variables educativas, que dan origen a indicadores de analfabetismo. Asimismo, 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, características de las viviendas que habitan, composición de los hogares y otras variables económicas y sociales de sus miembros. Dentro de sus objetivos se encuentra conocer los cambios ocurridos en los patrones de consumo de los hogares, actualizar la canasta de bienes y servicios y las ponderaciones del IPC, etc. (INE 2018e).

Cuadro 12.1: *Características de las algunas encuestas repetidas en América Latina.*

País	Nombre de la encuesta	Tipo	Periodicidad	Rotación	Muestra de viviendas
Argentina	Encuesta Permanente de Hogares	Panel	Trimestral	50%	25000
Bolivia	Encuesta Continua de Hogares	Panel	Anual	25%	10000
Brasil	Pesquisa Nacional por Amostra de Domicilios	Repetida	Anual	-	115000
Brasil	Pesquisa Nacional por Amostra de Domicilios Continua	Panel	Mensual	20%	70000
Chile	Encuesta de Caracterización Socioeconómica Nacional	Repetida	Bienal	-	84000
Colombia	Gran Encuesta Integrada de Hogares	Repetida	Mensual	-	20000
Costa Rica	Encuesta Nacional de Hogares	Repetida	Anual	-	13000
Cuba	Encuesta Nacional de Ocupación	Panel	Anual	33%	63000
Ecuador	Encuesta de Empleo, Desempleo y Subempleo en Área Urbana y Rural	Panel	Trimestral	50%	16000
El Salvador	Encuesta de Hogares de Propósitos Múltiples	Repetida	Anual	-	20000
Guatemala	Encuesta Nacional de Empleo e Ingresos	Repetida	Semestral	-	6000
Honduras	Encuesta Permanente de Hogares de Propósitos Múltiples	Repetida	Semestral	-	7200
México	Encuesta Nacional de Ingresos y Gastos de los Hogares	Repetida	Bienal	-	20000
Nicaragua	Encuesta Nacional de Hogares Sobre la Medición de Niveles de Vida	Panel	Trimestral	20%	7500
Panamá	Encuesta de Propósitos Múltiples	Repetida	Anual	-	15000
Paraguay	Encuesta Permanente de Hogares	Repetida	Anual	-	6000
Perú	Encuesta Nacional de Hogares sobre Condiciones de Vida y Pobreza	Panel	Anual	20%	32000

País	Nombre de la encuesta	Tipo	Periodicidad	Rotación	Muestra de viviendas
República Dominicana	Encuesta Nacional de Hogares de Propósitos Múltiples	Repetida	Anual	-	34000
República Dominicana	Encuesta Nacional de Fuerza de Trabajo	Panel	Semestral	25%	10000
Uruguay	Encuesta Continua de Hogares	Repetida	Mensual	-	53000
Venezuela	Encuesta de Hogares por Muestreo	Repetida	Semestral	-	45000

Cuadro 12.2: *Características de las algunas encuestas transversales en América Latina.*

País	Nombre	Año	Tamaño de Muestra
Argentina	Encuesta Nacional de Gastos de los Hogares	2017-2018	45000
Bolivia	Encuesta de Hogares	2017	11136
Brasil	Encuesta de Presupuestos familiares	2008-2009	53154
Chile	VIII Encuesta de Presupuestos Familiares	2016-2017	15239
Colombia	Encuesta Nacional de Presupuestos de los Hogares	2016-2017	87201
Costa Rica	Encuesta Nacional de Ingresos y Gastos de los hogares	2018-2019	9828
Ecuador	Encuesta de Condiciones de Vida	2013-2014	29052
El Salvador	Encuesta de Ingresos y Gastos de Hogares	2005-2006	4576
Guatemala	Encuesta Nacional de Condiciones de Vida	2014	11536
Honduras	Encuesta de Condiciones de Vida de los Hogares	2004	8155
México	Encuesta Nacional de Ingresos y Gastos de los Hogares	2016	81515
Nicaragua	Encuesta Nacional de Hogares sobre Medición de Nivel de Vida	2014	6851
Panamá	Encuesta de Ingresos y Gastos de los Hogares	2007-2008	10152
Paraguay	Encuesta de Ingresos y Gastos de Condiciones de Vida	2011-2012	6000

País	Nombre	Año	Tamaño de Muestra
Perú	Encuesta Nacional de Hogares sobre Condiciones de Vida y Pobreza	2017	36996
República Dominicana	Encuesta Nacional de Ingresos y Gastos de los hogares	2006-2007	8358
Uruguay	Encuesta Nacional de Gastos e Ingresos de los Hogares	2016-2017	7500
Venezuela	IV Encuesta Nacional de Presupuestos Familiares	2008-2009	45768

Capítulo 13

Software

El diseño y análisis de la información proveniente de las encuestas de hogares debe contemplar el uso exhaustivo de las herramientas computacionales existentes. Esta sección revisa con detalle las aproximaciones computacionales del software estadístico utilizado para realizar cada uno de los procesos estadísticos que se necesitan para lograr el cometido de la publicación de cifras oficiales con altos niveles de precisión y confiabilidad. En particular, para los siguientes procesos:

1. Selección de muestras acorde al diseño de muestreo definido
2. Generación de pesos de muestreo para cada individuo y hogar.
3. Modelación de la ausencia de respuesta e imputación estadística.
4. Calibración de los pesos de muestreo y ajustes por ausencia de respuesta.
5. Estimación de los errores de muestreo para cada indicador de interés en los cuadros de producción estadística.
6. Análisis de las relaciones multivariantes entre las variables de la encuesta.
7. Modelación de las estimaciones para la predicción del parámetro de interés en dominios pequeños.

UN (2005, sección 7.8) 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 y advierten con un ejemplo empírico que de no hacerlo, es posible que las estimaciones resultantes sean sesgadas y además sus errores de muestreo se vean subestimados. A continuación se muestran algunos de las características más importantes que los paquetes estadísticos computacionales incorporan en el manejo de datos que provienen de estructuras de muestreo complejas como las encontradas en las encuestas de hogares. Una revisión más exhaustiva y detallada que adjunta sintaxis y código computacional puede encontrarse en Heeringa, West y Berglund (2010, Apéndice A).

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 replicación para obtener los estimativos de varianza inducidos por el diseño de muestreo (Westat 2007). Algunos de estos softwares son de uso libre, aunque la mayoría corresponde a productos licenciados cuya licencia debe ser pagada. En general estos productos, además de proveer estadísticas descriptivas (como medias, totales, proporciones, percentiles y razones), permiten ajustar modelos de regresión lineales y logísticos. Todas las estadísticas resultantes están basadas en el diseño de muestreo de la encuesta.

R

R es un software de uso libre cuyo uso es cada vez más frecuente en la investigación social, puesto que es muy probable encontrar los más recientes hallazgos científicos programados en este software (R Core Team 2017). 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 2016b) permite determinar el tamaño de muestra de individuos y hogares en encuestas de hogares repetidas, tipo panel y con rotación. Los paquetes `sampling` (Tillé y Matei 2016) y `TeachingSampling` (Gutiérrez 2015) permiten seleccionar muestras probabilísticas desde los marcos de muestreo bajo un gran variedad de diseños y algoritmos de muestreo. El paquete `survey` (Lumley 2016), una vez que el diseño de muestreo ha sido predefinido mediante la función `svydesign()`, permite analizar datos provenientes de encuestas de hogares y obtener estimaciones apropiadas de los errores estándar.

SPSS

El módulo `complex samples` de `SPSS` (IBM 2017) incorpora la selección de muestras complejas mediante la definición de un esquema de muestreo establecido por el usuario. Luego, 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 regresiones de Cox.

SAS

Este software estadístico incluye un procedimiento para la selección de muestras probabilísticas llamado `SURVEYSELECT` que integra los métodos de selección más comunes como

muestreo aleatorio simple, muestreo sistemático, muestreo con probabilidad proporcional al tamaño, y algunas herramientas de afijación en los estratos. Para analizar los datos provenientes de muestras complejas se han programado algunos procedimientos (SAS 2010). `SURVEYMEANS`, que estima totales, medias, proporciones, y percentiles, junto con sus respectivos errores estándar, límites de los intervalos de confianza y pruebas de hipótesis. `SURVEYFREQ`, estima las estadísticas descriptivas (como totales y proporciones) de interés en tablas de una y dos vías, brinda las estimaciones del error de muestreo, y realiza un análisis de la bondad del ajuste de las estimaciones, independencia, riesgos y razones de *odds*. `SURVEYREG` y `SURVEYLOGISTIC` ajustan modelos de regresión lineal y logísticas, respectivamente. Estos procedimientos estiman los coeficientes de regresión, con sus respectivos errores, y adjunta un análisis exhaustivo de las propiedades de los modelos. Por último, `SURVEYPHREG` ajusta modelos de riesgos utilizando técnicas de máxima-pseudo verosimilitud.

STATA

El entorno `svy` provee 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 identifican 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 determinada etapa 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.

Bibliografía

Araujo, Maria Caridad (2007). "The 1990 and 2001 Ecuador Poverty Maps". En: *More Than A Prety Picture: Using Poverty Maps to Design Beter Policies and Interventions, Washington DC: The World Bank*.

Arias, Omar y Marcos Robles (2007). "The Geography of Monetary Poverty in Bolivia". En: *More Than A*.

Baillargeon, Sophie y Louis-Paul Rivest (2011). "The construction of stratified designs in R with the package stratification". En: 37.1, págs. 53-65.

Ballin, Marco y Giulio Barcaroli (2013). "Joint determination of optimal stratification and sample allocation using genetic algorithm". En: *Survey Methodology* 39.2, págs. 369-393.

Barcaroli, Giulio (2014). "SamplingStrata: An R Package for the Optimization of Stratified Sampling". En: *Journal of Statistical Software* 61.1, págs. 1-24. ISSN: 1548-7660. DOI: 10.18637/jss.v061.i04.

Barnett-Walker, Kortnee C. y col. (2003). "2001 National Household Survey on Drug Abuse". En:

Béland, Y. y col. (2005). "The Canadian Community Health Survey: Building on the success from the past". En: *Proceedings of the American Statistical Association Joint Statistical Meetings 2005, Section on Survey Research Methods, August 2005*.

Bell, Phillip (2001). "Comparison of Alternative Labour Force Survey Estimators". En: *Survey Methodology* 27.1, págs. 53-63.

Biemer, Paul P. y Lars E. Lyberg (2003). *Introduction to survey quality*. Wiley series in survey methodology. Wiley-Interscience. ISBN: 978-0-471-19375-3.

Casas-Cordero Valencia, Carolina, Jenny Encina y Partha Lahiri (2016). "Poverty Mapping for the Chilean Comunas". En: *Analysis of Poverty Data by Small Area Estimation*. Ed. por MonicaEditor Pratesi. John Wiley y Sons, Ltd, págs. 379-404. ISBN: 978-1-118-81496-3. DOI: 10.1002/9781118814963.ch20. URL: http://onlinelibrary.wiley.com/doi/10.1002/9781118814963.ch20/summary.

CEPAL (2018a). *Medición de la pobreza por ingresos - Actualización metodológica y resultados*. Metodologías de la CEPAL. URL: 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:

Clark, R. G. y D. G. Steel (2007). "Sampling within households in household surveys". En: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170.1, págs. 63-82.

Cochran, W. G. (1977). *Sampling Techniques*. Third Edition. Wiley.

Dalenius, Tore y JosrEPH L Hodges (1959). "Minimum Variance Stratification". En: *Journal of the American Statistical Association* 54.285, pág. 15.

DANE (2017). *Gran Encuesta Integrada de Hogares - - Departamento Administrativo Nacional de Estadística*. URL: http://formularios.dane.gov.co/Anda_4_1/index.php/catalog/458.

— (2018). *Encuesta Nacional de Presupuestos de los Hogares (ENPH) - Departamento Administrativo Nacional de Estadística*. URL: https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/encuesta-nacional-de-presupuestos-de-los-hogares-enph.

Deville, Jean-Claude y Carl-Erik Särndal (1992). "Calibration Estimators in Survey Sampling". En: *Journal of the American Statistical Association* 87.418, págs. 376-382. ISSN: 0162-1459. DOI: 10.1080/01621459.1992.10475217.

DGEEC (2018a). *Aspectos Metodológicos de la Encuesta de Ingresos y Gastos y de Condiciones de Vida (EIGyCV)*. URL: http://www.dgeec.gov.py/microdatos/register/eig/Metodologia%20EIG%20y%20CV.pdf.

— (2018b). *Encuesta Permanente de Hogares - Dirección General de Estadística, Encuestas y Censos*. URL: http://www.dgeec.gov.py/Publicaciones/Biblioteca/eph2016/Boletin-de-pobreza-2016.pdf.

DIGESTYC (2018a). *Encuesta de Hogares de Propósitos Múltiples - Dirección General de Estadística y Censos*. URL: http://www.digestyc.gob.sv/index.php/temas/des/ehpm.html.

— (2018b). *Encuesta de Ingresos y Gastos de los Hogares - Dirección General de Estadística y Censos*. URL: http://www.censos.gob.sv/enigh/descargas/ENIGH_Publicacion.pdf.

Duncan, G. J. y G. Kalton (1987). "Issues of Design and Analysis of Surveys across Time". En: *International Statistical Review / Revue Internationale de Statistique* 55.1, pág. 97. ISSN: 03067734. DOI: 10.2307/1403273.

Estevao, Victor y Carl-Erik Särndal (2006). "Survey Estimates by Calibration on Complex Auxiliary Information". En: *International Statistical Review / Revue Internationale de Statistique* 74.2, págs. 127-147.

Feinberg, Stephen y Elizabeth Stasny (1983). "Estimating monthly gross flows in labour force participation". En: *Survey Methodology* 9.1, págs. 77-102.

Foster, James, Joel Greer y Erik Thorbecke (1984). "A Class of Decomposable Poverty Measures". En: *Econometrica* 52.3, págs. 761-766. ISSN: 0012-9682. DOI: 10.2307/1913475.

Gambino, J. G. y PL. d N. Silva (2009). "Chapter 16 - Sampling and Estimation in Household Surveys". En: *Handbook of Statistics*. Vol. 29. Handbook of Statistics. Elsevier, págs. 407-439. DOI: 10.1016/S0169-7161(08)00016-3.

Groves, Robert y col. (2009). *Survey Methodology*. John Wiley y Sons.

Gunning, Patricia y Jane M Horgan (2004). "A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations". En: *Survey Methodology* 30.2, págs. 159-166.

Gurney, M. y J. Daly (1965). "A Multivariate Approach to Estimation in Periodic Sample Surveys". En: *Proceedings of the Social Statistics Section, American Statistical Association*, págs. 242-257.

Gutiérrez, Hugo Andrés (2015). *TeachingSampling: Selection of Samples and Parameter Estimation in Finite Population*. R package version 3.2.2. URL: https://CRAN.R-project.org/package=TeachingSampling.

— (2016a). *Estrategias de muestreo: diseño de encuestas y estimación de parámetros*. Segunda edición. Google-Books-ID: UlVmE5pkRwIC. Ediciones de la U. ISBN: 978-958-762-586-8.

— (2016b). *samplesize4surveys: Sample Size Calculations for Complex Surveys*. R package version 3.1.2.400. URL: https://CRAN.R-project.org/package=samplesize4surveys.

Hansen, Morris H., William N. Hurwitz y William G. Madow (1953). *Sample Survey Methods And Theory*.

Hansen, Morris H, William N Hurwitz y William G Madow (1953). *Sample survey methods and theory*. Vol. 1. Wiley New York.

Hayes, Clinton y Nicole Watson (2009). "HILDA Imputation methods". En: *Working paper*.

Heeringa, Steven G., Brady T. West y Patricia A. Berglund (2010). *Applied survey data analysis*. Chapman and Hall/CRC statistics in the social and behavioral sciences series. CRC Press. ISBN: 978-1-4200-8066-7.

Hornik, Robert y col. (2002). "Evaluation of the National Youth Anti-Drug Media Campaign: Fourth Semi-Annual Report of Findings". En: pág. 740.

IBGE (2018a). *Pesquisa de Orçamentos Familiares - Instituto Brasileiro de Geografia e Estatística*. URL: https://ww2.ibge.gov.br/home/estatistica/pesquisas/pesquisa_resultados.php?id_pesquisa=25.

— (2018b). *Pesquisa Nacional por Amostra de Domicílios Contínua - Instituto Brasileiro de Geografia e Estatística*. URL: https://www.ibge.gov.br/estatisticas-novoportal/sociais/trabalho/9173-pesquisa-nacional-por-amostra-de-domicilios-continua-trimestral.html?redirect=1.

IBM (2017). *IBM SPSS Complex Samples*. URL: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/23.0/en/client/Manuals/IBM_SPSS_Complex_Samples.pdf.

INDEC (2018a). *Encuesta Nacional de Gastos de los Hogares - Instituto Nacional de Estadística y Censos*. URL: https://www.indec.gob.ar/engho/.

— (2018b). *Encuesta Permanente de Hogares - Instituto Nacional de Estadística y Censos*. URL: https://www.indec.gov.ar/bases-de-datos.asp.

INE (2016a). *Encuesta Continua de Hogares (ECH) - Instituto Nacional de Estadística*. URL: http://ine.gub.uy/encuesta-continua-de-hogares1.

— (2016b). *Encuesta de Gastos e Ingresos de los Hogares - ENGIH 2016/2017 - Instituto Nacional de Estadística*. URL: http://www.ine.gub.uy/engih2016.
— (2018a). *Empleo e Ingresos - Instituto Nacional de Estadística - Guatemala*. URL: https://www.ine.gob.gt/index.php/encuestas/empleo-e-ingresos.
— (2018b). *Encuesta de Hogares - INE*. URL: http://www.ine.gob.bo/sitio_EH/Encuesta_Hogares.html.
— (2018c). *Encuesta de Presupuestos Familiares (EPF) - Instituto Nacional de Estadística - Chile*. URL: https://www.ine.cl/estadisticas/ingresos-y-gastos/epf.
— (2018d). *Encuesta Nacional de Condiciones de Vida - Instituto Nacional de Estadística - Guatemala*. URL: https://www.ine.gob.gt/index.php/encuestas-de-hogares-y-personas/condiciones-de-vida.
— (2018e). *Ficha Técnica de Encuesta de Hogares por Muestreo - Instituto Nacional de Estadística*. URL: http://www.ine.gov.ve/index.php?option=com_content&id=333&Itemid=103.
— (2018f). *Honduras - Encuesta de Condiciones de Vida de los Hogares*. URL: http://170.238.108.229/index.php/catalog/76/overview.
— (2018g). *Instituto Nacional de Estadística*. http://www.ine.gob.hn.
INEC (2017). *Encuesta Nacional de Hogares - Instituto Nacional de Estadística y Censos*. URL: http://www.inec.go.cr/encuestas/encuesta-nacional-de-hogares.
— (2018a). *Encuesta de Condiciones de Vida (ECV) - Instituto Nacional de Estadística y Censos*. URL: http://www.ecuadorencifras.gob.ec/encuesta-de-condiciones-de-vida-ecv/.
— (2018b). *Encuesta de Ingresos y Gastos de los Hogares - Instituto Nacional de Estadística y Censo - Panamá*. URL: http://www.contraloria.gob.pa/inec/Aplicaciones/EIGH2008/intro.html.
— (2018c). *Encuesta Nacional de Ingresos y Gastos de los Hogares - Instituto Nacional de Estadística y Censos*. URL: http://www.inec.go.cr/encuestas/encuesta-nacional-de-ingresos-y-gastos-de-los-hogares.
— (2018d). *Instituto Nacional de Estadística y Censo - Panamá*. URL: https://www.contraloria.gob.pa/inec/Publicaciones/Publicaciones.aspx?ID_SUBCATEGORIA=38&ID_PUBLICACION=91&ID_IDIOMA=1&ID_CATEGORIA=5.
— (2018e). *Instituto Nacional de Estadística y Censos*. URL: http://www.ilo.org/surveydata/index.php/catalog/1393/study-description.
INEGI (2012). "Metodología de la Construcción del Marco Maestro de Muestreo 2012 y del Diseño de la Muestra Maestra 2012". En:
— (2016). *Encuesta Nacional de Ingresos y Gastos de los Hogares 2016 Nueva serie*. URL: https://www.inegi.org.mx/programas/enigh/nc/2016/.
— (2019). *Encuesta Nacional de Ocupación y Empleo (ENOE), población de 15 años y más de edad*. URL: https://www.inegi.org.mx/programas/enoe/15ymas/.
INEI (2016). *Encuesta Nacional de Hogares sobre Condiciones de Vida y Pobreza - Instituto Nacional de Estadística e Informática*. URL: http://webinei.inei.gob.pe/anda_inei/index.php/catalog/543.

INIDE (2018). *Instituto Nacional de Información de Desarrollo- INIDE de Nicaragua*. URL: http://www.inide.gob.ni/.

Jarque, Carlos M. (1981). "A Solution to the Problem of Optimum Stratification in Multivariate Sampling". En: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 30.2, págs. 163-169. ISSN: 0035-9254. DOI: 10.2307/2346387.

Judkins, David R. (1990). "Fay's Method for Variance Estimation". En: *Journal of Official Statistics; Stockholm* 6.3, pág. 223. ISSN: 0282423X.

Kalton, G. (2009). "Some Issues in the Design and Analysis of Longitudinal Surveys". En:

Kalton, G. y C. F. Citro (1993). "Panel surveys: adding the fourth dimension". En: *Survey Methodology* 19.2, págs. 205-215.

Kim, Jae Kwang y Minsun Kim Riddles (2012). "Some theory for propensity-score-adjustment estimators in survey sampling". En: *Survey Methodology* 38.2, págs. 157-165.

Kish, Leslie (1965). *Survey Sampling*. John Wiley y Sons.

Korn, Edward Lee y Barry I. Graubard (1999). *Analysis of health surveys*. Wiley. ISBN: 978-0-471-13773-3.

Kozak, Marcin (2004). "Optimal Stratification Using Random Search Method in Agricultural Surveys". En: *Statistic in Transition* 6.5, págs. 797-806.

Krewski, D. y J. N. K. Rao (1981). "Inference From Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods". En: *The Annals of Statistics* 9.5, págs. 1010-1019. ISSN: 0090-5364.

LaRoche, Silvia (2003). *Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics*.

Lavallée, Pierre y Michael A. Hidiroglou (1988). "On the Stratification of Skewed Populations". En: *Survey Methodology* 14.1, págs. 33-43.

Lent, Janice, Stephen M. Miller y Martha Duff (1999). "Effects of Composite Weights on Some Estimates from the Current Population Survey". En: *Journal of Official Statistics; Stockholm* 15.3, pág. 431. ISSN: 0282423X.

Lewis, Taylor (2017). "Estimation Strategies Involving Pooled Survey Data". En: *SAS Global Forum*.

Little, Roderick J. A. y Donald B. Rubin (2002). *Statistical analysis with missing data*. 2. ed. Wiley series in probability and statistics. Wiley. ISBN: 978-0-471-18386-0.

Lohr, S. (2000). *Sampling: Design and Analysis*. Thompson.

Lopez-Calva, Luis F, Lourdes Rodríguez-Chamussy y Miguel Székely (2007). "Poverty maps and public policy in Mexico". En: *More Than A Pretty Picture. Using Poverty Maps to Design Better Policies and Interventions. Washington DC: WorldBank*. Citeseer.

Lumley, Thomas (2010). *Complex surveys: a guide to analysis using R*. Wiley series in survey methodology. Wiley. ISBN: 978-0-470-28430-8.

— (2016). *survey: analysis of complex survey samples*. R package version 3.32.

Macqueen, J (1967). "Some methods for classification and analysis of multivariate observations". En: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, págs. 281-297.

McCarthy, P. J. (1969). "Pseudo-Replication: Half Samples". En: *Revue de l'Institut International de Statistique / Review of the International Statistical Institute* 37.3, págs. 239-264. ISSN: 0373-1138. DOI: 10.2307/1402116.

McLaren, C. y D. G. Steel (2001). "Rotation Patterns and Trend Estimation for Repeated Surveys Using Rotation Group Estimates". En: *Statistica Neerlandica* 55.2, págs. 221-238.

MDS (2015). *Observatorio Social - Ministerio de Desarrollo Social - Gobierno de Chile*. URL: http://observatorio.ministeriodesarrollosocial.gob.cl/casen/casen_obj.php.

National Research Council (2015). *Realizing the Potential of the American Community Survey: Challenges, Tradeoffs, and Opportunities*. National Academies Press. ISBN: 978-0-309-36678-6. DOI: 10.17226/21653. URL: http://www.nap.edu/catalog/21653.

Naud, Jean-Francois (2002). *Combined-panel longitudinal weighting - Survey of Labour and Income Dynamics*.

OIT (1982). *Resolución sobre estadísticas de la población económicamente activa, del empleo, del desempleo y del subempleo*. URL: http://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/normativeinstrument/wcms_087483.pdf.

— (2013). *Estadísticas del trabajo, el empleo y la subutilización de la fuerza de trabajo*. URL: http://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/publication/wcms_220537.pdf.

ONE (2018a). *Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) - Oficina Nacional de Estadística*. URL: https://www.one.gob.do/enhogar.

— (2018b). *Oficina Nacional de Estadística (ONE)*. URL: https://www.one.gob.do/encuestas/enigh.

— (2018c). *Oficina Nacional de Estadísticas. Cuba*. URL: http://www.one.cu/sitioone2006.asp.

ONU (2011). *Canberra Group Handbook on Household Income Statistics*. Second edition. United Nations Economic Comission for Europe. URL: https://www.unece.org/fileadmin/DAM/stats/groups/cgh/Canbera_Handbook_2011_WEB.pdf.

— (2015). *Transformar nuestro mundo: la Agenda 2030 para el Desarrollo Sostenible*. URL: http://unctad.org/meetings/es/SessionalDocuments/ares70d1_es.pdf.

— (2016). *Global Sustainable Development Report 2016*. URL: https://sustainabledevelopment.un.org/globalsdreport/2016.

— (2017). *Progress towards the Sustainable Development Goals*.

Park, I. y H. Lee (2006). "Design effects for the weighted mean and total estimators under complex survey sampling". En: *Quality control and applied statistics* 51.4, pág. 381.

Park, Inho (2003). "Design Effects and Survey Planning". En: pág. 8.

Parker, JD, M Talih y DJ Malec (2017). "National Center for Health Statistics Data Presentation Standards for Proportions". En: *Vital Health Stat* 2.175.

Presser, Stanley y col. (2004). *Methods for Testing and Evaluating Survey Questionnaires*. John Wiley y Sons.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rao, J. N. K. e Isabel Molina (2014). *Small-Area Estimation*. John Wiley y Sons, Ltd. ISBN: 978-1-118-44511-2. DOI: 10.1002/9781118445112.stat03310.pub2. URL: http://onlinelibrary.wiley.com/doi/10.1002/9781118445112.stat03310.pub2/abstract.

Rao, J. N. K. y C. F. J. Wu (1988). "Resampling Inference with Complex Survey Data". En: *Journal of the American Statistical Association* 83.401, págs. 231-241. ISSN: 0162-1459. DOI: 10.1080/01621459.1988.10478591.

Rubin, Donald B. (1987). *Multiple Imputation for nonresponse in surveys*. Wiley series in probability and mathematical statistics Applied probability and statistics. Wiley. ISBN: 978-0-471-08705-2.

Särndal, Carl-Erik y Sixten Lundström (2006). *Estimation in surveys with nonresponse*. Repr. Wiley series in survey methodology. Wiley. ISBN: 978-0-470-01133-1.

Särndal, Carl-Erik, Bengt Swensson y Jan Wretman (2003). *Model Assisted Survey Sampling*. Google-Books-ID: ufdONK3E1TcC. Springer Science y Business Media. ISBN: 978-0-387-40620-6.

SAS (2010). *SAS/STAT 9.22 User's Guide - Survey Sampling and Analysis Procedures*. URL: https://support.sas.com/documentation/cdl/en/statugsurveysamp/63778/PDF/default/statugsurveysamp.pdf.

Silva, PL. d N. (2004). "Calibration estimation: when and why, how much and how". En: *Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística*.

Singh, A. C., M. Westlake y M. Feder (2004). "A generalization of the Coefficient of variation with application to suppression of imprecise estimates". En:

Sinngh, M. P., J. G. Gambino y H. J. Mantel (1994). "Issues and strategies for small area data". En: *Survey Methodology* 20.1, págs. 3-22.

Starick, Rosslyn y Nicole Watson (2011). "Evaluation of Alternative Income Imputation Methods for the HILDA Survey". En: *Working paper*, pág. 38.

STATA (2013). *STATA Survey Data*. URL: https://www.stata.com/manuals13/svy.pdf.

Steel, D. y C. McLaren (2008). "Design and Analysis of Repeated Surveys". En:

Sun, Claire (2010). "HILDA Expenditure imputation". En: *Working paper*.

Tillé, Yves y Alina Matei (2016). *sampling: Survey Sampling*. R package version 2.8. URL: https://CRAN.R-project.org/package=sampling.

UN (2005). *Household surveys in developing and transition countries*. Studies in methods / United Nations, Department of Economic and Social Affairs, Statistics Division Series F. ISBN: 978-92-1-161481-7.

— (2008). *Designing household survey samples: practical guidelines*. Studies in methods / United Nations, Department of Economic and Social Affairs, Statistics Division Series F. United Nations. ISBN: 978-92-1-161495-4.

Valliant, Richard y Jill A. Dever (2017). *Survey Weights: A Step-by-step Guide to Calculation*. 1 edition. Stata Press.

Valliant, Richard, Jill A. Dever y Frauke Kreuter (2013). *Practical Tools for Designing and Weighting Survey Samples*. Springer New York. ISBN: 978-1-4614-6448-8. DOI: 10.1007/978-1-4614-6449-5. URL: http://link.springer.com/10.1007/978-1-4614-6449-5.

Vehovar, Vasja (1999). "Field Substitution and Unit Nonresponse". En: *Journal of Official Statistics* 15.2, págs. 335-350.

West, Brady T. (2012). "Accounting for Multi-stage Sample Designs in Complex Sample Variance Estimation". En:

West, Brady T. y Sean Esteban McCabe (2012). "Incorporating complex sample design effects when only final survey weights are available". En: *The Stata Journal* 12.4, págs. 718-725.

Westat (2007). *WesVar 4.3. Users guide.* URL: http://users.nber.org/~jroth/chap1.pdf.

Wolter, Kirk (2007). *Introduction to Variance Estimation*. Springer Science y Business Media. ISBN: 978-0-387-35099-8.