Fundamentos estadísticos para el análisis de las encuestas postcensales

Andrés Gutiérrez[1], Giovany Babativa, Stalyn Guerrero

2025-02-20

[1]Comisión Económica para América Latina y el Caribe (CEPAL) - andres.gutierrez@cepal.org

Contents

List of Figures

List of Tables

Abstract

Este es el repositorio inicial para la Serie de Estudios Estadísticos sobre el análisis de las encuestas post-censales para la medición de la cobertura en los censos de población

Chapter 1

El sistema de estimación dual

Para poder hacer un análisis estadístico apropiado de las encuestas de cobertura como instrumentos que pretenden medir la omisión de un censo, es necesario remitirse a los rudimentos originales de su proceso inferencial, el cual está basado en el sistema de estimación dual. Este enfoque fue primeramente utilizado en modelos de captura y recaptura que se originaron en el siglo XVII, con desarrollos modernos a partir de Petersen (1896), Lincoln (1930) y Schnabel (1938). La aplicación a eventos vitales humanos fue iniciada por el trabajo de Sekar and Deming (1949).

Este capítulo pretende establecer las condiciones iniciales bajo las cuales es apropiado utilizar este enfoque, así como los supuestos que se deben cumplir para que esta metodología produzca estimadores insesgados y precisos.

1.1 Planteamiento del problema

Wolter (1986) considera una población humana U, de tamaño N, el cual es fijo pero desconocido y es precisamente el parámetro de interés sobre el cual se requiere una inferencia precisa. En una primera instancia, se supone que se realiza un censo de la población en un momento específico en el tiempo, y que el censo intenta enumerar a cada individuo. Sin embargo, por diversas razones, algunos individuos no son contados en el censo. La diferencia entre el conteo censal y N se denotará como el *error de cobertura*.

Una de las principales complicaciones del error de cobertura es que su magnitud no puede determinarse únicamente a partir de los datos del censo. Para cuantificar este error, es imprescindible disponer de información adicional, la cual se obtiene generalmente mediante una encuesta por muestreo aplicada a la misma población objetivo. Esta encuesta (conocida como encuesta de postenumeración o encuesta de cobertura) se realiza habitualmente después del censo, utilizando el mismo período de referencia temporal. La encuesta permite estimar la magnitud del error de cobertura al comparar

los resultados obtenidos con los datos del censo, proporcionando así una medida más precisa y ajustada de la población real.

Inicialmente, considérese que la encuesta representa una enumeración completa de toda la población, lo cual no es cierto en la práctica, pero este paso es necesario para esclarecer las propiedades estadísticas del sistema de estimación dual. Por supuesto, en los próximos capítulos de este documento se abordarán los acercamientos necesarios para ajustar la inferencia al caso real en el que la encuesta únicamente llega a una fracción de la población.

1.2 Condiciones y supuestos

El modelo del error de cobertura descansa bajo un número de supuestos que son imprescindibles a la hora de utilizar una encuesta de postenumeración como instrumento fiable para la medición del error de cobertura en un censo. A continuación se realiza un listado exhaustivo de ellos.

1.2.1 Estabilidad poblacional

Se supone que la población U es cerrada y de tamaño fijo N. En la práctica, esto implica que:

1. El período de referencia del censo está bien definido; es decir que el censo se lleva a cabo en un intervalo de tiempo específico y claramente establecido. Este período es crucial para garantizar que todos los datos recolectados se refieran a la misma fecha o intervalo de tiempo, evitando así inconsistencias y errores en la estimación de la población.

2. No existen incorporaciones durante el período de referencia; es decir que se asume que no ocurren **nacimientos** ni **inmigraciones**. Esto significa que no se agregan nuevos individuos a la población censada.

3. No existen pérdidas; es decir que se asume que no ocurren **defunciones** ni **emigraciones** durante el período de referencia. Esto asegura que no se restan individuos de la población censada.

1.2.2 Estructura multinomial

El evento conjunto de que un individuo esté o no esté en el censo y esté o no en la encuesta se puede modelar correctamente usando una distrución multinomial con los siguientes parámetros:

	En la encuesta	Fuera de la encuesta	Total
En el censo	p_{11}	p_{12}	p_{1+}
Fuera del censo	p_{21}	p_{22}	p_{2+}
Total	p_{+1}	p_{+2}	1

En donde:

- p_{11} denota la probabilidad de que un individuo sea encontrado en el censo y en la encuesta.
- p_{12} denota la probabilidad de que un individuo sea encontrado en el censo, pero no en la encuesta.
- p_{21} denota la probabilidad de que un individuo no sea encontrado en el censo, pero sí en la encuesta.
- p_{22} denota la probabilidad de que un individuo no sea encontrado ni en el censo, ni en la encuesta.

Asimismo, en términos de las probabilidades marginales, se definen las siguientes cantidades:

- p_{1+} es la probabilidad de que un individuo sea correctamente encontrado en el censo.
- p_{+1} es la probabilidad de que un individuo sea correctamente encontrado en la encuesta.

Esto quiere decir que el individuo tiene chance de ser clasificado en cualquiera de los cuatro estados definidos por las entradas de la tabla anterior; pero que al momento de la recolección de los datos, el individuo sólo puede pertenecer a uno y solo uno de estos estados.

1.2.3 Independencia autónoma

El censo y la encuesta se generan como resultado de N ensayos mutuamente independientes. Cada ensayo representa a un individuo de la población real U. A partir de la recolección de los datos, se obtiene la siguiente clasificación:

	En la encuesta	Fuera de la encuesta	Total
En el censo	N_{11}	N_{12}	N_{1+}
Fuera del censo	N_{21}	N_{22}	N_{2+}
Total	N_{+1}	N_{+2}	$N_{++} = N$

Note que $N_{ab} = \sum_{k \in U} x_{kab}$ y x_{kab} es una variable indicadora que señala si el individuo k pertenece a la celda (a, b) de la tabla $(a, b = 1, 2, +)$. Bajo este esquema inicial, se tiene que:

1. El conteo del censo N_{1+} se considera observable.
2. Los valores N_{11}, N_{12} y N_{21} se consideran observables con base en los datos de la encuesta y el emparejamiento con el censo.
3. Los valores N_{22}, y el tamaño de la población de interés $N = N_{++}$, se consideran desconocidos y deben estimarse con base en el modelo.
4. Bajo este modelo, el conteo del censo N_{1+} define una variable aleatoria con media $E(N_{1+}) = Np_{1+}$ y varianza $V(N_{1+}) = Np_{1+}(1 - p_{1+})$.

1.2.4 Independencia causal

Se supone que el evento de ser incluido en el censo es independiente del evento de ser incluido en la encuesta. Como resultado de este supuesto, la razón de productos cruzados de probabilidades, conocida comúnmente como la Razón de Odds, satisface la siguiente relación:

$$\frac{p_{11} \cdot p_{22}}{p_{12} \cdot p_{21}} = 1$$

El resultado anterior se tiene, puesto que la probabilidad conjunta de un individuo en una celda específica de la tabla de contingencia se factoriza como:

$$p_{11} = P(\text{individuo está en el censo y en la encuesta}) = p_{1+} \cdot p_{+1}$$

Similarmente, se tiene que

$$p_{12} = p_{1+} \cdot (1 - p_{+1})$$
$$p_{21} = (1 - p_{i1+}) \cdot p_{i+1}p$$

$$p_{22} = (1 - p_{i1+}) \cdot (1 - p_{i+1})$$

Sustituyendo adecuadamente en la razón de productos cruzados, entonces se tiene que

$$\frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{p_{1+}p_{+1}(1 - p_{1+})(1 - p_{+1})}{p_{1+}(1 - p_{+1})(1 - p_{1+})p_{+1}} = 1$$

La dependencia causal, como señala el Bureau (2022), es un fenómeno que ocurre cuando la inclusión o exclusión de un individuo en el censo influye en su probabilidad de ser incluido en la encuesta. Este tipo de dependencia puede generar sesgos en los datos y afectar la calidad de las estimaciones estadísticas, lo que a su vez puede comprometer la validez de las conclusiones derivadas de estos estudios. Por ello, es fundamental

implementar estrategias que mitiguen este riesgo y aseguren la independencia operativa entre ambos sistemas.

Una de las medidas clave para lograr esta independencia operativa es garantizar que el personal involucrado en la recolección de datos de la encuesta no participe en las mismas áreas geográficas o comunidades donde trabajaron durante el censo. Esto reduce la posibilidad de que los encuestadores influyan en las respuestas de los individuos basándose en interacciones previas o en información recopilada durante el censo. Además, al evitar la superposición de personal, se minimiza el riesgo de que los encuestados asocien ambas actividades, lo que podría alterar su disposición a participar o la veracidad de sus respuestas.

Otra estrategia importante es asegurar que las entrevistas de la encuesta se realicen después de que las operaciones del censo hayan concluido en un área específica. Esto permite que los procesos de recolección de datos no se solapen temporalmente, lo que reduce la posibilidad de que los resultados de una actividad afecten directa o indirectamente a la otra. Por ejemplo, si un individuo ha sido contactado recientemente para el censo, podría sentirse menos motivado a participar en la encuesta, o viceversa. Separar prudencialmente ambas operaciones ayuda a mantener la independencia de las respuestas.

Además, es crucial restringir el acceso del personal del censo a la información sobre la muestra de la encuesta. De manera similar, el personal de la encuesta no debería tener acceso a los resultados del censo durante la fase de recolección de datos, ya que esta información podría sesgar su enfoque o interpretación de las respuestas.

1.2.5 Emparejamiento

Se asume que es posible realizar un emparejamiento preciso entre los resultados de la encuesta y los del censo. Esto significa que se puede identificar de manera exacta y sin errores:

1. Cuáles individuos registrados en la encuesta también aparecen en los registros del censo.
2. Cuáles individuos de la encuesta no están presentes en los datos del censo.

Este emparejamiento correcto es crucial para evaluar la cobertura del censo y para ajustar las estimaciones de la población total, asegurando que los datos sean lo más precisos y completos posible.

Por otro lado, se asume que inevitablemente habrá algún grado de no respuesta en el censo y en la encuesta. Esto significa que algunos individuos no serán contactados o no proporcionarán la información solicitada. Para abordar este problema, es fundamental recopilar suficiente información auxiliar sobre los no respondientes. Esta información

puede incluir datos como nombres, direcciones, fechas de nacimiento y otros identificadores únicos que permitan una correcta identificación de los individuos.

En la práctica, se implementan procedimientos específicos para asegurar que la información recopilada sea lo suficientemente detallada y precisa para permitir un emparejamiento exacto entre los datos del censo y los de la encuesta. Este emparejamiento es crucial para evaluar la cobertura del censo y ajustar las estimaciones de la población total.

1.2.6 Eventos espurios

Se asume que tanto el censo como la encuesta están libres de incidencias espurias o falsas, o que estas han sido eliminadas antes de realizar las estimaciones. Esto implica que se han tomado medidas para evitar cualquier tipo de error en el registro de los resultados tanto del censo como de la encuesta. En la práctica, esto significa que se han implementado procedimientos rigurosos para identificar y corregir cualquier anomalía en los datos. Algunos de los eventos espurios más importantes que pueden ocurrir incluyen:

1. Duplicaciones en la lista del censo. Esto ocurre cuando un individuo es contado más de una vez en el censo, lo que puede inflar artificialmente el tamaño de la población.
2. Registros de casos inexistentes. Estos son registros de individuos que no existen en realidad, pero que han sido incluidos erróneamente en el censo o en la encuesta. Esto puede suceder debido a errores de entrada de datos o malentendidos durante la recolección de información.
3. Casos no pertinentes. Estos son individuos que no deberían haber sido incluidos en el censo debido a que no cumplen con los criterios del período de referencia. Un ejemplo común es el registro de un individuo que nació después del período de referencia del censo, lo que resulta en una inclusión incorrecta en los datos.

Para asegurar la precisión de las estimaciones, es crucial que estos eventos espurios sean identificados y eliminados antes de proceder con el análisis de los datos.

1.2.7 Postestratificación

Es frecuente y beneficioso aplicar algún tipo de postestratificación en la estimación del tamaño real de la población. La postestratificación es una técnica estadística que permite ajustar las estimaciones de la población dividiéndola en subgrupos homogéneos, basados en variables categóricas. Esta técnica mejora la precisión y la validez de las estimaciones al considerar las diferencias dentro de la población.

Por ejemplo, una forma común de postestratificación es por edad. En este caso, la población se divide en diferentes grupos de edad, como niños, adolescentes, adultos jóvenes, adultos de mediana edad y personas mayores. Para cada uno de estos grupos de edad, se realizan estimaciones específicas de la población. Estas estimaciones se basan en los datos recolectados tanto en el censo como en la encuesta. Una vez obtenidas las estimaciones específicas por edad, se agregan para calcular una estimación total de la población, denotada como N.

La postestratificación no se limita solo a la edad; también se puede aplicar a otras variables demográficas y socioeconómicas, como sexo, etnia, nivel educativo, región geográfica, entre otras. Es fundamental que cualquier variable utilizada para la postestratificación esté correctamente registrada para todos los individuos tanto en el censo como en la encuesta.

1.3 Inferencia

Nuestro objetivo es estimar el tamaño total de una población, denotado como N_{++}, utilizando dos fuentes de información complementarias. La primera fuente es el censo, el cual logra capturar correctamente a N_{+1} individuos de la población. La segunda fuente es la encuesta, que captura de manera precisa a N_{1+} individuos.

Uno de los supuestos del sistema de estimación dual es que el evento de que una persona sea encontrada se puede modelar como un proceso estocástico de tipo Bernoulli. Esto quiere decir que N_{11}, N_{1+} y N_{+1} se asumen como variables aleatorias binomiales al ser sumas de eventos Bernoulli.

1.3.1 Los estimadores del sistema dual

Bajo este modelo, las variables aleatorias siguen distribuciones binomiales condicionales:

$$N_{1+} \sim \text{Bin}(N_{++}, p_{1+}), \quad N_{+1} \sim \text{Bin}(N_{++}, p_{+1}), \quad N_{11} \sim \text{Bin}(N_{++}, p_{11})$$

Una vez que los datos hayan sido recolectados y clasificados bajo este esquema, es bien sabido en la literatura estadística, que los estimadores para las probabilidades de interés toman la siguiente forma:

$$\tilde{p}_{11} = \frac{N_{11}}{N_{++}}, \quad \tilde{p}_{1+} = \frac{N_{1+}}{N_{++}}, \quad \tilde{p}_{+1} = \frac{N_{+1}}{N_{++}}$$

Al asumir independencia entre la captura en el censo y la captura en la encuesta, entonces $\tilde{p}_{11} = \tilde{p}_{1+} \cdot \tilde{p}_{+1}$, y por ende:

$$\frac{N_{11}}{N_{++}} = \frac{N_{1+}}{N_{++}} \cdot \frac{N_{+1}}{N_{++}}$$

Luego, al despejar convenientemente, se encuentra que el estimador del sistema dual para el total poblacional N_{++} está dado por

$$\tilde{N}_{++} = \frac{N_{1+} \cdot N_{+1}}{N_{11}}$$

A partir de este resultado, podemos reemplazar en las expresiones $\tilde{p}_{11}$, $\tilde{p}_{1+}$ y $\tilde{p}_{+1}$ para obtener estimadores de máxima verosimilitud para las probabilidades de interés son los siguientes:

$$\tilde{p}_{11} = \frac{N_{11}}{\tilde{N}_{++}} = \frac{N_{11}^2}{N_{1+} \cdot N_{+1}}$$

$$\tilde{p}_{1+} = \frac{N_{1+}}{\tilde{N}_{++}} = \frac{N_{11}}{N_{+1}}$$

$$\tilde{p}_{+1} = \frac{N_{+1}}{\tilde{N}_{++}} = \frac{N_{11}}{N_{1+}}$$

Wolter (1986, sección 2.4) plantea un esquema conjunto que induce estos mismos estimadores a partir de la función de verosimilitud asociada al modelo, la cual está dada por la siguiente expresión:

$$L(N, p_{i+}, p_{+i}) = \binom{N}{x_{11}, x_{12}, x_{21}} p_{1+}^{x_{1+}} (1 - p_{1+})^{N - x_{1+}} p_{+1}^{x_{+1}} (1 - p_{+1})^{N - x_{+1}}.$$

Los estimadores de máxima verosimilitud de los parámetros de interés se encuentran maximizando la anterior expresión sujeta a las restricciones pertinentes sobre las sumas de las probabilidades.

1.3.2 Propiedades del estimador

El estimador $\tilde{N}_{++}$, es conocido como el método de Petersen, y es utilizado en estudios de captura y recaptura para estimar el tamaño de una población. Este método fue desarrollado por el biólogo danés Carl Georg Johannes Petersen (Petersen, 1896) y más tarde popularizado por C. Chandra Sekar y W. Edwards Deming en 1949 para estimar tasas de nacimientos y defunciones, así como la cobertura de los registros vitales (Sekar and Deming, 1949).

Para demostrar que este estimador es insesgado, se debe verificar que $E[\tilde{N}_{++}] = N_{++}$. En primer lugar, por la propiedad de la esperanza en distribuciones binomiales, se tiene que:

$$E[N_{1+}] = N_{++}p_{1+}, \quad E[N_{+1}] = N_{++}p_{+1}, \quad E[N_{11}] = N_{++}p_{11}$$

Ahora, la esperanza del estimador toma la siguiente forma:

$$E[\tilde{N}_{++}] = E\left[\frac{N_{1+} \cdot N_{+1}}{N_{11}}\right]$$

En primera instancia como N_{1+} y N_{+1} son variables aleatorias, es necesario apelar a las propiedades de la esperanza condicional, de la siguiente manera:

$$E[\tilde{N}_{++}] = E\left[E\left(\frac{N_{1+} \cdot N_{+1}}{N_{11}}\middle| N_{1+}, N_{+1}\right)\right]$$

Además, como N_{11} también es una variable aleatoria, entonces bajo condiciones de regularidad que permitan utilizar la expansión de Taylor, es posible aproximar la esperanza de este cociente al cociente de las esperanzas (Casella and Berger, 2002). De esta forma, se tiene que:

$$E\left(\frac{N_{1+} \cdot N_{+1}}{N_{11}}\middle| N_{1+}, N_{+1}\right) = \frac{E(N_{1+} \cdot N_{+1}|N_{1+}, N_{+1})}{E(N_{11}|N_{1+}, N_{+1})}$$

Dado que N_{1+} y N_{+1} son independientes, entonces $E[N_{1+} \cdot N_{+1}] = E[N_{1+}]E[N_{+1}]$. Reemplazando convenientemente, se tiene que

$$E[\tilde{N}_{++}] = \frac{N_{++}^2 p_{1+}p_{+1}}{N_{++}p_{1+}p_{+1}} = N_{++} = N$$

Por otro lado, Wolter (1986) afirma que la varianza del estimador puede ser estimada mediante la siguiente expresión:

$$\tilde{V}[\tilde{N}_{++}] = \frac{N_{1+} \cdot N_{+1} \cdot N_{12} \cdot N_{21}}{N_{11}^3}$$

Chapter 2

Estimación dual con la muestra de la encuesta

En el capítulo anterior, se partió de un supuesto simplificador: que todos los N miembros de la población tenían la posibilidad de ser incluidos tanto en el censo como en la encuesta. Esta suposición, aunque útil para establecer un marco teórico inicial, no refleja la realidad en la mayoría de los estudios de error de cobertura. En la práctica, es poco común que todos los individuos de una población estén expuestos a ser captados por ambas fuentes de información. Por ello, es necesario ajustar este enfoque para abordar situaciones más realistas.

En este contexto, ahora consideraremos un escenario más plausible: mientras que todos los miembros de la población están expuestos a ser incluidos en el censo (es decir, el censo intenta cubrir a toda la población), solo una muestra de la población tendrá la posibilidad de ser incluida en la encuesta. Esta distinción es fundamental, ya que introduce una asimetría en la forma en que ambas fuentes de datos interactúan con la población. El censo, al ser un esfuerzo exhaustivo, busca contar a todos los individuos dentro de un territorio o grupo definido. Sin embargo, la encuesta, por su naturaleza muestral, solo abarca una fracción de la población.

Este cambio en los supuestos implica una modificación significativa en los métodos de análisis, ya que se altera la estructura de la información disponible y las cantidades que se consideran conocidas o desconocidas. Anteriormente, se podía asumir que ciertos totales poblacionales eran observables o directamente medibles, pero bajo este nuevo enfoque, solo el total del censo, denotado como N_{1+}, se considera observable; sin emabrgo, no es directamente conocido, puesto que el censo está expuesto a errores de enumeración y duplicaciones. Esto significa que el número de individuos capturados correctamente por el censo no es una cantidad que se toma como dada y debe ser corregida con la muestra.

Por otro lado, el total de la población capturado por la encuesta, representado como N_{+1}, ahora se considera no observable.Además, otras cantidades clave, como N_{11} (el

número de individuos capturados por ambas fuentes), N_{12} (individuos capturados por el censo pero no por la encuesta), y N_{21} (individuos capturados por la encuesta pero no por el censo), también se consideran desconocidas. Sin embargo, todas estas cantidades pueden estimarse indirectamente a partir de los datos de la encuesta, utilizando los métodos estadísticos adecuados. En este documento se utilizará el superescrito $\hat{\cdot}$ para denotar una cantidad estimada directa o indirectamente haciendo uso de la muestra. De esta forma, la estructura de los datos y estimaciones necesarias para realizar la medición del error de cobertura usando ambas operaciones estadísticas puede ser descrita de la siguiente manera:

	En la encuesta	Fuera de la encuesta	Total
En el censo	$\hat{N}_{11}$	$\hat{N}_{12} = \hat{N}_{1+} - \hat{N}_{11}$	$\hat{N}_{1+}$
Fuera del censo	$\hat{N}_{21} = \hat{N}_{+1} - \hat{N}_{11}$		
Total	$\hat{N}_{+1}$		$\hat{N}_{++} = \hat{N}$

2.1 El diseño de muestreo

Por lo general, el diseño de muestreo para una encuesta postcensal sigue una estructura compleja que contempla al menos dos procesos: el primero es la estratificación y el segundo es la selección de conglomerados. Estos dos procesos introducen un efecto de diseño que, por lo general, aumenta el error estándar de los estimadores debido a la alta correlación intra-clase de los conglomerados en los estratos:

1. En el caso de la estratificación, este es un procedimiento que divide la población en grupos homogéneos (casi siempre supeditados a divisiones geográficas). Esta división pretende reducir la varianza de los estimadores, asegurando un tamaño de muestra óptimo para la representación de zonas o regiones.
2. Las unidades primarias de muestreo (UPM) son pequeños conglomerados geográficos, como manzanas o sectores censales, que en la mayoría de casos se seleccionan mediante probabilidad proporcional al número de viviendas, hogares o personas. Por lo general, en las UPM seleccionadas, se realiza un barrido total de todas sus estructuras y en cada vivienda se enlista a cada una de las personas de cada una de las viviendas. Este muestreo se conoce como muestreo de conglomerados. En otras ocasiones, es posible hacer un submuestreo de viviendas en las UPM seleccionadas.

Siguiendo la notación de la litera consideremos un diseño estándar estratificado con selección de conglomerados en una sola etapa. La población se agrupa en M UPM y se asume que se selecciona una muestra aleatoria simple sin reemplazo de m UPM. Asumimos que la población de la encuesta se enumera completamente dentro de los conglomerados seleccionados. Además, se supone que la lista de conglomerados es

completa. Cada miembro de la población pertenece a uno y solo un conglomerado, y no hay miembros de la población que no estén cubiertos por uno de los M conglomerados.

En algunas ocasiones, el diseño de muestreo de la encuesta contempla un formato de encuesta de hogares en el que la selección de las viviendas se realiza en dos etapas. Por lo general, en la segunda etapa se seleccionan viviendas ocupadas por hogares al momento de la recolección de datos. Sin embargo, esta selección de viviendas ocupadas durante el trabajo de campo introduce limitaciones críticas, como las siguientes:

1. Limitación en la definición de la población de interés: la segunda etapa del muestreo (selección de viviendas ocupadas) inmediatamente restringe la población objetivo a las **personas civiles no institucionalizadas**, lo que genera sesgos en la medición de cobertura, puesto que se excluyen poblaciones no cubiertas como las personas en cárceles, hospitales, residencias de ancianos o bases militares (población institucionalizada). Todas estas personas quedan fuera del marco muestral, ya que estas viviendas colectivas no se incluyen en la selección de hogares tradicionales. Asimismo, los individuos en situación de calle, migrantes temporales o trabajadores itinerantes no tienen una "vivienda ocupada" fija durante el trabajo de campo (población móvil o sin techo).

2. Desfase temporal entre el censo y la encuesta: si hay un intervalo prolongado (meses o años) entre el censo y la encuesta postcensal, se violan algunos supuestos clave. Supongamos que, durante el censo, una vivienda estaba ocupada, pero al momento de la encuesta está deshabitada (ej.: migración, desastres naturales). Esta vivienda tendrá probabilidad nula de ser seleccionada en la encuesta, a pesar de haber albergado a un hogar censado. Asimismo, las viviendas construidas después del censo podrían contener hogares no censados.

2.2 La muestra E y la muestra P

A partir del diseño de muestreo para la encuesta, se seleccionan dos muestras. La primera, conocida como muestra de la población o muestra P, consiste en áreas que serán enumeradas después de la realización del censo. Su objetivo es estimar directamente los valores de N_{11} y N_{+1}. La segunda, denominada muestra de la enumeración o muestra E, es una muestra de registros del censo que serán examinados para estimar indirectamente el valor de N_{1+}. La muestra P y la muestra E desempeñan roles críticos en la estimación de la cobertura poblacional y la corrección de errores en los conteos del censo.

Generalmente, la muestra E y la muestra P provienen de las mismas áreas geográficas, lo que garantiza una base común para la comparación y el análisis de los datos. En los siguientes capítulos ampliaremos los conceptos sobre las reglas de emparejamiento que deben ser definidas a partir de la muestra P y sobre los conceptos que deberán utilizarse para encontrar errores de enumeración en el censo en la muestra E. en resumen:

1. La muestra E corregir la presencia de eventos espurios para que este supuesto se pueda utilizar en el sistema de estimación dual. En particular permite obtener una estimación sobre el número de personas que fueron contadas en el censo pero que no deberían haber sido parte de la enumeración (por ejemplo, duplicados, personas nacidas después del censo, personas muertas antes del censo, migrantes, entradas ficticias, entre otros). Con base en esta muestra se estima la proporción de inclusiones erróneas en el censo y se proporciona una base para ajustar el conteo del censo eliminando estas imprecisiones.

2. La muestra P en los registros de la encuesta de cobertura que se comparan con los registros del censo para obtener una estimación directa del número de personas que fueron contadas correctamente tanto en el censo como en la encuesta. Asimismo, permite obtener una estimación indirecta del número de personas que no fueron contadas en el censo pero deberían haber sido parte de la enumeración.

2.3 Los estimadores de muestreo

Como la encuesta representa una muestra de la población que viene de una medida de probabilidad, y a su vez, existe un modelo multinomial, entonces se introduce una complejidad metodológica clave: la necesidad de establecer las bases inferenciales para incluir dos fuentes de incertidumbre: el modelo y el muestreo (Binder, 2011). Wolter (1986) afirma que este cambio de enfoque implica que la estimación del error de cobertura debe considerar dos fuentes principales de incertidumbre: (1) la variabilidad debida a la selección muestral de la encuesta, y (2) la variabilidad del modelo asociada con el modelo de error de cobertura.

La variabilidad inducida por la selección de la muestra de la encuesta implica que las estimaciones derivadas de ella (como N_{+1} o N_{11}) están afectadas por la aleatoriedad inherente a la selección de unidades en la muestra. Si la encuesta utiliza un diseño complejo (como estratificación o conglomerados), la variabilidad aumenta debido a los efectos de diseño. Este tipo de variabilidad se mide con los métodos clásicos de inferencia estadística en encuestas de hogares. En segundo lugar, está la variabilidad derivada del modelo multinomial. En esta instancia, la novedad radica en integrar estas incertidumbres por medio de una inferencia doble, usando los resultados bien conocidos de las esperanzas y varianzas condicionales.

Si denotamos por π_k la probabilidad de inclusión del elemento k en la muestra s_P, la cual está determinada por su selección probabilística, entonces el peso de muestreo del elemento k-ésimo en la muestra P se define como $w_k = \pi_k^{-1}$. Este peso refleja la inversa de la probabilidad de inclusión y se utiliza para ajustar las estimaciones en función del diseño de muestreo. De manera similar, los pesos de muestreo se definirán para la muestra s_E. Para simplificar la notación, vincularemos la muestra correspondiente a través de los subíndices en las sumas. Por ejemplo, al referirnos a la muestra s_P,

utilizaremos el subíndice P en las sumas, y para la muestra s_E, emplearemos el subíndice E.

Asumiendo que $x_{k,11}$ representa una variable aleatoria dicotómica que toma el valor de uno si el individuo k fue encontrado tanto en la muestra como en el censo y, cero, en otro caso, entonces los estimadores de muestreo de N_{+1} y N_{11}, serán respectivamente:

$$\hat{N}_{+1} = \sum_{k \in s_P} w_k$$
$$\hat{N}_{11} = \sum_{k \in s_P} w_k \; x_{k,11}$$

Asimismo, si z_k representa una variable aleatoria dicotómica que toma el valor de uno si el individuo k fue correctamente incluido en el censo y, cero, en otro caso, entonces el estimador de muestreo de N_{1+} será:

$$\hat{N}_{1+} = N^0_{1+} - \sum_{k \in s_E} w_k(1 - \; z_k)$$

En donde N^0_{1+} denota el conteo no corregido de personas en el censo. Esta cifra debe basarse exclusivamente en los datos recopilados durante el operativo censal, sin incluir imputaciones, proyecciones ni ningún otro tipo de ajustes estadísticos. Esto garantiza que los resultados reflejen fielmente la información obtenida en el campo. Para los anteriores estimadores, es claro que $x_{k,11}$ es una variable aletaoria que se define en la muestra s_P, mientras que z_k es una variable aleatoria que se define en la muestra s_E. Por otro lado, Bureau (2022) propone un estimador directo alternativo para N_{1+}, que se define a partir de la muestra E, y que corresponde a un conteo ponderado de enumeraciones correctas. Este estimador toma la siguiente forma:

$$\hat{N}_{1+} = \sum_{k \in s_E} w_k \; z_k$$

Recordando que el estimador del modelo para N es $\tilde{N} = \frac{N_{1+} \cdot N_{+1}}{N_{11}}$; entonces, su estimador insesgado bajo el diseño de muestreo se encuentra reemplazando N_{1+}, N_{+1} y N_{11} por sus respectivos estimadores insesgados en la muestra. Por consiguiente, se tiene que el estimador de muestreo del tamaño poblacional N tomará la siguiente forma:

$$\hat{N}_{++} = \hat{N} = \frac{\hat{N}_{1+} \cdot \hat{N}_{+1}}{\hat{N}_{11}}$$

Nótese que los estimadores de muestreo para N_{12} y N_{21} toman la siguiente forma:

$$\hat{N}_{12} = \hat{N}_{1+} - \hat{N}_{11}$$
$$\hat{N}_{21} = \hat{N}_{+1} - \hat{N}_{11}$$

La existencia de individuos que no fueron capturados en ninguno de los dos listados representa un desafío significativo, ya que su número solo puede ser estimado indirectamente a partir de la superposición observada entre la encuesta y el censo. Por otro lado, Wolter (1986) establece las condiciones sobre las cuales estos estimadores son insesgados y además propone el siguiente estimador aproximadamente insesgado de su varianza:

$$\tilde{V}(\hat{N}) = \tilde{V}_m(\tilde{N}) + \tilde{V}_p(\hat{N})$$

En donde $\tilde{V}_m(\tilde{N})$ es el estimador de la varianza de $\tilde{N}$ bajo el modelo multinomial, que usa las contrapartes muestrales en lugar de las poblacionales, de la siguiente forma:

$$\tilde{V}_m(\tilde{N}) = \frac{\hat{N}_{1+} \cdot \hat{N}_{+1} \cdot (\hat{N}_{1+} - \hat{N}_{11}) \cdot (\hat{N}_{+1} - \hat{N}_{11})}{\hat{N}_{11}^3}$$

Asimismo, $\tilde{V}_p(\hat{N})$ corresponde con un estimador tradicional de varianzas para estimadores de muestreo (CEPAL, 2023). De esta forma, Wolter (1986, sección 3.1.) afirma que

$$\tilde{V}_p(\hat{N}) \approx \frac{M^2}{m}(1-f)S_d^2$$

Definiendo a $\tilde{N}_{i,+1}$ como la estimación del tamaño del i-ésimo conglomerado a partir de la muestra s_P, se tiene que $S_d^2 = \frac{1}{m-1}\sum_{i=1}^{m} d_i^2$ y además:

$$d_i = \frac{\hat{N}_{1+}}{\hat{N}_{11}}\left(\tilde{N}_{k,+1} - \frac{\hat{N}_{+1}}{\hat{N}_{11}} x_{k,11}\right)$$

Finalmente, es posible combinar los diferentes estimadores en las muestras E y P, junto con la información recolectada en el censo para crear otro tipo de estimadores. Siendo $\hat{N}_{1+}^0 = \sum_{k \in s_E} w_k$ un estimador de muestreo del número de enumeraciones en el censo (correctas o erroneas), es posible ajustar el número de enumeraciones en el censo con su contraparte muestral, y definir el siguiente estimador de razón:

$$\hat{N}_{++}^{ratio} = \frac{N_{1+}^0}{\hat{N}_{1+}^0} \frac{\hat{N}_{1+} \cdot \hat{N}_{+1}}{\hat{N}_{11}}$$

De la misma manera, es posible refinar el estimador usando la postestratificación (Gutiérrez, 2016). Esta es una técnica que particiona la población en subgrupos homogéneos y que permite minimizar el impacto del sesgo de correlación (que los individuos que no fueron enumerados en el censo serán más propensos a no ser incluidos en la encuesta). Como se mencionó anteriormente, es usual utilizar al menos las divisiones administrativas mayores, los grupos de edad y el sexo. Cada una de las particiones inducidas por el cruce de estas variables se conoce como post-estratos. Suponiendo que existen G postestratos, entonces el estimador de razón post-estratificada toma la siguiente forma:

$$\hat{N}_{++}^{post} = \sum_{g=1}^{G} \left[\frac{N_{g1+}^{0}}{\hat{N}_{g1+}^{0}} \frac{\hat{N}_{g1+} \cdot \hat{N}_{g+1}}{\hat{N}_{g11}} \right] = \sum_{g=1}^{G} \left[N_{g1+}^{0} \frac{\hat{p}_{g1+}}{\hat{p}_{g11}} \right]$$

En donde $\hat{p}_{g1+} = \frac{\hat{N}_{g+1}}{\hat{N}_{g1+}^{0}}$ y $\hat{p}_{g11} = \frac{\hat{N}_{g11}}{\hat{N}_{g+1}}$ son respectivamente estimadores directos de la proporción de individuos correctamente enumerados y de la proporción de emparejamiento en el post-estrato g.

Chapter 3

Clasificación de los errores

Bibliography

Binder, D. A. (2011). Estimating model parameters from a complex survey under a model-design randomization framework. *Pakistan Journal of Statistics*, 27(4):371–390.

Bureau, U. C. (2022). 2020 post-enumeration survey estimation design.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Duxbury Press, 2nd edition. See section on the Delta Method for approximations involving expectations of functions of random variables.

CEPAL (2023). *Diseño y análisis estadístico de las encuestas de hogares de América Latina*. Metodologías de la CEPAL.

Gutiérrez, H. A. (2016). *Estrategias de muestreo: diseño de encuestas y estimación de parámetros*. Ediciones de la U, segunda edición edition. Google-Books-ID: UlVmE5pkRwIC.

Lincoln, F. C. (1930). Calculating waterfowl abundance on the basis of banding returns. *Circular*, 118:1–4.

Petersen, C. G. J. (1896). The yearly immigration of young plaice into the limfjord from the german sea. *Report of the Danish Biological Station*, 6:1–48.

Schnabel, Z. E. (1938). The estimation of the total fish population of a lake. *American Mathematical Monthly*, 45:348–352.

Sekar, C. C. and Deming, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44(245):101–115.

Wolter, K. M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81(394):338–346.