Fundamentos estadísticos para el análisis de las encuestas postcensales

Andrés Gutiérrez[1], Giovany Babativa, Stalyn Guerrero

2025-02-13

[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 encontrado en el censo.
- p_{+1} es la probabilidad de que un individuo sea 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 El estimador

Planteamiento del problema: queremos estimar el tamaño total de una población N_{++} usando dos fuentes de información. En primer lugar, el **censo**, que captura a N_{+1} individuos de la población correctamente capturados por el censo; luego, la **encuesta de cobertura**, que captura correctamente a N_{1+} individuos de la población. El número de individuos que fueron capturados en ambas fuentes se denota como N_{11}.

Este esquema de captura se puede representar en la siguiente tabla de dos entradas:

Censo - Encuesta	Sí	No	Total
Sí	N_{11}	N_{12}	N_{1+}
No	N_{21}	N_{22}	N_{2+}
Total	N_{+1}	N_{+2}	N_{++}

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.

En esta instancia, denotamos a p_{1+} como la probabilidad de que una persona sea correctamente encontrada por el censo y p_{+1}, la probabilidad de que una persona sea correctamente encontrada en la encuesta de cobertura. El sistema de estimación dual también supone que ambas operaciones estadísticas son independientes una de la otra; es decir que la probabilidad de que una persona sea correctamente encontrada tanto en el censo como en la encuesta será:

$$p_{11} = p_{1+} \cdot p_{+1}$$

Bajo este modelo, las variables aleatorias siguen distribuciones binomiales:

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

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

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

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

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

1.4 Insesgamiento del estimador

El estimador $\hat{N}_{++}$, 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 $E[\hat{N}_{++}] = N_{++}$. 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:

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

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:

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

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

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

1.5 Supuestos imprescindibles del estimador

El estimador dual es insesgado bajo el supuesto de independencia entre el censo y la encuesta de cobertura. Si esta independencia no se cumple, el estimador puede estar sesgado.

Chapter 2

La encuesta de cobertura como lista B

Si la encuesta que se utiliza para medir el error de cobertura tiene el formato de una encuesta de hogares, entonces la población de interés únicamente podrá ser definida en términos de las personas civiles no institucionalizadas y el error de cobertura únicamente podrá ser estimado en esta subpoblación.

Chapter 3

La muestra E y la muestra P

La muestra E y la muestra P desempeñan roles críticos en la estimación de la cobertura poblacional y la corrección de errores en los conteos del censo. Estas muestras se utilizan en el contexto de una Encuesta de Cobertura para evaluar la precisión del censo y estimar el número de personas omitidas o incluidas erróneamente.

3.1 La muestra E (Muestra de Enumeración)

La muestra E consiste en registros del censo que se revisan de manera independiente para determinar su exactitud. Su propósito es identificar errores en el censo, tales como:

1. **Inclusiones erróneas**: 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 muesrtas antes del censo, migrantes, entradas ficticias, entre otros).
2. **Exclusiones erróneas**: Personas que fueron omitidas por el censo pero que sí deberían haber sido enumeradas (por ejemplo, xxxx xxxxxx xxxxx)

La muestra E se compone de una muestra parcial de registros del censo, los cuales se verifican de manera independiente (por ejemplo, mediante trabajo de campo o comparación con otra fuente) para determinar si son correctos. 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.

3.2 La muestra P (Muestra de Población)

La muestra P consiste en registros de una **encuesta de cobertura** independiente, en la que se comparan los registros del censo para determinar errores de cobertura. Su

propósito es identificar a las personas que fueron

1. **Contadas correctamente**: Personas que fueron contadas tanto en el censo como en la EPE
2. **Omitidas por el censo**: Personas que no fueron contadas en el censo pero deberían haber sido parte de la enumeración.

La muestra P ayuda a estimar la proporción de personas omitidas por el censo y proporciona una base para ajustar el conteo del censo agregando personas omitidas.

La muestra E y la muestra P se utilizan juntas en el marco del sistema dual de estimación para estimar la población total ($\widehat{N}$) al corregir el número de personas omitidas por el censo (subconteos estimados con la muestra P) y el número de inclusiones erróneas en el censo (sobreconteos estimados con la muestra E).

3.3 Clasificación de los errores

Bibliography

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.

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.