Estimación del error de muestreo

Andrés Gutiérrez

Asesor Regional en Estadísticas Sociales - Comisión Económica para América Latina y el Caribe (CEPAL) - andres.gutierrez@cepal.org

Tabla de contenidos I

Estimación del error de muestreo

Fórmulas exactas y linealización de Taylor

La técnica del último conglomerado

Linealización de Taylor

Estimación del error de muestreo

Introducción

- Después de la selección de la muestra y el proceso de medición, es esencial estimar los parámetros junto con sus errores estándar, que son la raíz cuadrada de la varianza.
- La estimación del error estándar es crucial en la inferencia estadística y depende de la complejidad del diseño de muestreo y del tipo de estimador utilizado.
- Hay tres alternativas para calcular el error estándar:
  1. *Fórmulas exactas* basadas en el diseño de muestreo.
  2. *Linealización de Taylor* para estimadores no lineales.
  3. Métodos computacionales modernos como los *pesos replicados*.

Introducción

- Los softwares estadísticos modernos ofrecen procedimientos para la estimación de la varianza en diseños de muestreo complejos.
- Una forma sencilla de usarlos es siguiendo estos pasos en una base de datos agregada:
  1. Modificar los pesos, de tal forma que cumplan las restricciones poblacionales básicas.
  2. Definir los estratos de interés en donde el diseño de muestreo se realiza de forma independiente.
  3. Definir estrictamente las UPM como aglomerados poblacionales que incluyen a los hogares y personas (con sus múltiples entrevistas).

Fórmulas exactas y linealización de Taylor

Fórmulas exactas y linealización de Taylor

- Las fórmulas exactas para calcular la estimación de errores en diseños de muestreo pueden ser complejas, especialmente en diseños multietápicos y con estimadores complejos.
- La estimación de la varianza en una estrategia de muestreo se basa en probabilidades de inclusión de primer y segundo orden.
- La fórmula exacta para la varianza del estimador de Horvitz-Thompson en un diseño sin reemplazo es dada por

$$\sum_U \sum_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

, donde $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$.

Fórmulas exactas y linealización de Taylor

- La probabilidad de inclusión de segundo orden π_{kl} define la probabilidad de que los elementos k y l pertenezcan a la muestra al mismo tiempo.

$$\pi_{kl} = Pr(k \in s,\ l \in s) = Pr(I_k\ I_l = 1) = \sum_{s \ni k,l} p(s).$$

En donde el subíndice $s \ni k, l$ se refiere a la suma sobre todas las muestras que contienen a los elementos k-ésimo y l-ésimo.

- Calcular estas fórmulas exactas es inviable en la práctica debido a razones computacionales y a la imposibilidad de acceder a registros sobre toda la población finita.

Estimador Insesgado de la Varianza.

- Gutiérrez (2016) afrima que un estimador insesgado para esta varianza está dada por la siguiente expresión:

$$\widehat{Var}_1(\hat{t}_{y,\pi}) = \sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

- Si el diseño es de tamaño de muestra fijo, un estimador insesgado está dado por

$$\widehat{Var}_2(\hat{t}_{y,\pi}) = -\frac{1}{2} \sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2$$

Intervalo de Confianza.

Un intervalo de confianza de nivel $(1-\alpha)$ para el total poblacional t_y

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

donde $z_{1-\alpha/2}$ se refiere al percentil $(1-\alpha/2)$ de una variable aleatoria con distribución normal estándar.

Ejemplo: Muestreo Aleatorio Simple

- Para un diseño de muestreo aleatorio simple y un estimador de Horvitz-Thompson, la fórmula de la estimación de la varianza es

$$\widehat{V}ar(\hat{t}_{y\pi}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{y_s}$$

, donde $S^2_{y_s}$ es la varianza de los valores de la característica de interés en la muestra aleatoria s, dada por

$$S^2_{y_s} = \frac{1}{n-1}\sum_{k \in S}(y_k - \bar{y}_s)^2$$

Ejemplo: Muestreo Aleatorio Estratificado

En un diseño de muestreo aleatorio estratificado para una media, la fórmula del estimador de Horvitz-Thompson es

$$\bar{y}_\pi = \frac{1}{N}\sum_s d_k y_k = \sum_{h=1}^{H} W_h \bar{y}_h$$

, con

$$\widehat{V}ar(\bar{y}_\pi) = \sum_{h=1}^{H} w_h^2 \frac{1-f_h}{n_h} S_{yh}^2$$

para la estimación de la varianza.

Ejemplo: Muestreo Aleatorio Estratificado Bietápico

En un diseño de muestreo estratificado y bietápico, la fórmula final del estimador de la varianza del estimador de Horvitz-Thompson para el total poblacional es más compleja. La fórmula es

$$\widehat{Var}(\hat{t}_{y,\pi}) = \sum_{h=1}^{H} \left[\frac{N_{Ih}^2}{n_{Ih}} \left(1 - \frac{n_{Ih}}{N_{Ih}}\right) S^2_{\hat{t}_{yh} S_I} + \frac{N_{Ih}}{n_{Ih}} \sum_{i \in S_{Ih}} \frac{N_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right) S^2_{y_{S_i}} \right]$$

En donde $S^2_{\hat{t}_y S_I}$ y $S^2_{y_{S_i}}$ son, respectivamente, las varianzas muestrales de los totales estimados en las UPM seleccionadas y las varianzas muestrales de los hogares incluidos en la submuestra dentro de las UPM seleccionadas en la muestra de la primera etapa.

La técnica del último conglomerado

La técnica del último conglomerado

- Estimar la varianza en encuestas complejas es difícil y costoso debido a las complicaciones algebraicas y computacionales.
- La técnica del último conglomerado (*ultimate cluster*) es una aproximación eficiente utilizada en encuestas multietápicas.
- Esta técnica solo considera la varianza en la primera etapa del muestreo, asumiendo selección con reemplazo.
- Es una opción viable cuando las etapas posteriores de muestreo no afectan significativamente la varianza de los estimadores.

La técnica del último conglomerado

- Considere cualquier estimador del total poblacional dado por la siguiente combinación lineal

$$\hat{t}_{y,\pi} = \sum_{k \in s} d_k y_k = \sum_{k \in U} I_k d_k y_k$$

En donde I_k son variables indicadoras de la pertenencia del elemento k a la muestra s.

La técnica del último conglomerado

- Suponiendo un diseño de muestreo en varias etapas con selección de una muestra s_I de m_I unidades primarias de muestreo (UPM) U_i.
- Si la selección se realizó con reeemplazo, la i-ésima UPM tiene probabilidad de selección p_{I_i}.
- Si la selección se realizó sin reeemplazo, la i-ésima UPM tiene probabilidad de inclusión π_{I_i}.

La técnica del último conglomerado

- En las etapas posteriores de muestreo se selecciona una muestra de elementos para cada UPM seleccionada en la primera etapa.
- La probabilidad condicional $\pi_{k|i}$ representa la probabilidad de que el k-ésimo elemento pertenezca a la muestra dada que la UPM que lo contiene fue seleccionada en la primera etapa.

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

- Se definen factores de expansión como:
  1. $d_{I_i} = \frac{1}{\pi_{I_i}}$ para la UPM,
  2. $d_{k|i} = \frac{1}{\pi_{k|i}}$ para el elemento dentro de la UPM, y
  3. $d_k = d_{I_i} \times d_{k|i}$ para el elemento en toda la población.

La técnica del último conglomerado

- El estimador de Hansen-Hurwitz (HH) es un estimador insesgado que puede ser considerado junto con el estimador HT en un diseño de muestreo con reemplazo.
- La expresión del estimador HH es más sencilla de calcular y proporciona estimaciones de varianza más manejables desde el punto de vista computacional en comparación con el estimador HT.

Ejemplo: Estimador de Hansen-Hurwitz (HH)

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

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

donde p_{Ii} representa la probabilidad de selección de la unidad i, mientras que m_I es el tamaño de la muestra (con reemplazo) en la primera etapa del muestreo.

Ejemplo: Estimador de Hansen-Hurwitz (HH)

La varianza estimada del estimador HH se calcula utilizando la fórmula

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

En donde las cantidades $\hat{t}_{y_i}$ representan lo totales estimados de la variable de interés en la i-ésima UPM y están dados por:

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

La técnica del último conglomerado

- La técnica del último conglomerado utiliza la expresión de la varianza del estimador HH en lugar de la expresión exacta en diseños de muestreo complejos sin reemplazo en la primera etapa. Esto se logra al equiparar las probabilidades de inclusión y selección en la primera etapa.

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

- Para usar esta aproximación, se requiere equiparar las probabilidades de inclusión y selección en la primera etapa, lo que lleva a definir el estimador del total poblacional como un estimador tipo Hansen-Hurwitz.

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

La técnica del último conglomerado

- La forma del estimador ha sido equiparada con el estimador tipo Hansen-Hurwitz, lo que permite utilizar su estimación de varianza.
- La ventaja de esta aproximación es que utiliza los factores de expansión finales d_k, disponibles en los microdatos de las encuestas.
- La estimación de la varianza del estimador HH, bajo un diseño de muestreo en varias etapas, tiene una expresión más manejable computacionalmente, como se muestra en las ecuaciones presentadas.

La técnica del último conglomerado

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

La técnica del último conglomerado

Al definir $\breve{t}_{y_i}$ como la contribución de la i-ésima UPM a la estimación del total poblacional y $\bar{\breve{t}}_y$ como la contribución promedio en el muestreo de la primera etapa, el estimador de varianza toma la forma conocida como el estimador de varianza del *último conglomerado*.

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

La técnica del último conglomerado

En un escenario de muestreo estratificado con tres etapas de selección dentro de cada estrato, la técnica del último conglomerado permite aproximar el estimador de la varianza de la siguiente manera:

$$\widehat{Var}(\hat{t}_{y,p}) = \sum_h \frac{n_h}{n_h - 1} \sum_{i \in s_h} \left(\hat{t}_{y_i} - \bar{\hat{t}}_{y_h} \right)^2$$

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.

La técnica del último conglomerado

- Técnica del último conglomerado: sobrestima la varianza pero utiliza directamente los pesos finales de muestreo, apreciada por investigadores.
- La técnica del último conglomerado es una salida práctica al problema de la estimación de la varianza
- La expresión del estimador de la varianza no constituye un estimador estrictamente insesgado, sí se considera 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:

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

- Diseño de muestreo en cascada: inicio con unidades primarias (municipios), seguido de unidades secundarias (sectores cartográficos) y finalmente unidades finales (hogares).

Linealización de Taylor

Linealización de Taylor

- La linealización de Taylor es utilizada para aproximar las varianzas de parámetros no lineales.
- Esta técnica se basa en expresar el estimador como una función de estimadores lineales de totales, donde cada estimador $\hat{t}_j$ representa la suma ponderada de las observaciones.
- Cuando el estimador de interés no es lineal, las propiedades estadísticas como el sesgo, la eficiencia y la precisión deben aproximarse.
- La aproximación lineal de primer orden utilizando la linealización de Taylor permite estimar la varianza de manera más precisa.

Pasos para la linealización de Taylor

1. Expresar el estimador del parámetro de interés $\hat{\theta}$ como una función de estimadores de totales insesgados. Así,

$$\hat{\theta} = f(\hat{t}_1, \hat{t}_2, ..., \hat{t}_Q).$$

2. Determinar todas las derivadas parciales de f con respecto a cada total estimado $\hat{t}_q$ y evaluar el resultado en las cantidades poblacionales t_q. Así

$$a_q = \left. \frac{\partial f(\hat{t}_1, ..., \hat{t}_Q)}{\partial \hat{t}_q} \right|_{\hat{t}_1 = t_1, ..., \hat{t}_Q = t_Q}$$

Pasos para la linealización de Taylor

3. Aplicar el teorema de Taylor para funciones vectoriales para linealizar la estimación $\hat{\theta}$ con $\mathbf{a} = (t_1, t_2, \cdots, t_Q)'$. En el paso anterior, se vio que $\bigtriangledown\hat{\theta} = (a_1, \cdots, a_Q)$. Por consiguiente se tiene que

$$\hat{\theta} = f(\hat{t}_1, ..., \hat{t}_Q) \cong \theta + \sum_{q=1}^{Q} a_q(\hat{t}_q - t_q)$$

4. Definir una nueva variable E_k con $k \in s$ al nivel de cada elemento observado en la muestra aleatoria, así:

$$E_k = \sum_{q=1}^{Q} a_q y_{qk}$$

Pasos para la linealización de Taylor

5. Si los estimadores $\hat{t}_q$ son estimadores de Horvitz-Thompson, una expresión que aproxima la varianza de $\hat{\theta}$ está dada por

$$AVar(\hat{\theta}) = Var\left(\sum_{q=1}^{Q} a_q \hat{t}_{q,\pi}\right) = Var\left(\sum_{S} \frac{E_k}{\pi_k}\right) = \sum_{U}\sum_{U} \Delta_{kl} \frac{E_k}{\pi_k}\frac{E_l}{\pi_l}.$$

Linealización de Taylor

- Se aproximan los valores E_k reemplazando los totales desconocidos por los estimadores de los mismos, obteniendo

$$e_k = \sum_{q=1}^{Q} \hat{a}_q y_{qk}$$

.

- Se utiliza la aproximación de Taylor para encontrar la varianza del estimador de Horvitz-Thompson para un total, expresada como

$$\widehat{Var}(\hat{t}_{y,\pi}) = \sum\sum_{S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}$$

.

Linealización de Taylor

La estimación de la aproximación de la varianza del estimador de la tasa de desocupación se define en términos de las variables linealizadas como:

1. Se obtiene la variable linealizada e_k utilizando la fórmula $e_k = \frac{1}{\hat{t}_{z,\pi}}(y_k - \hat{\theta} z_k)$.
2. Bajo un diseño bietápico con selección aleatoria simple sin reemplazo en cada etapa, la estimación de la varianza $\widehat{Var}(\hat{\theta})$ toma la forma:

$$\widehat{Var}(\hat{\theta}) = \frac{N_I^2}{n_I}\left(1 - \frac{n_I}{N_I}\right) S^2_{\hat{t}_e S_I} + \frac{N_I}{n_I}\sum_{i \in S_I} \frac{N_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right) S^2_{e_{S_i}}$$

En donde $S^2_{\hat{t}_e S_I}$ es la varianza muestral de los totales estimados t_{ei} de las UPM seleccionadas en la primera etapa del muestreo y $S^2_{e_{S_i}}$ es la varianza muestral entre los valores e_k para los elementos incluidos en la submuestra dentro de cada UPM seleccionada en la primera etapa.

Linealización de Taylor

- La varianza estimada del estimador de calibración utilizando la técnica de linealización de Taylor se basa en variables linealizadas

$$e_k = y_k - \mathbf{x}_k' \hat{\theta}$$

.

- Las variables $\mathbf{x}_k$ están relacionadas con el vector de totales auxiliares $\mathbf{t}_\mathbf{x}$ y se miden en la misma encuesta.
- El vector $\hat{\theta}$ es el estimado de coeficientes de regresión entre y_k y $\mathbf{x}_k$.

Linealización de Taylor

- La *Pesquisa Nacional por Amostra de Domicilios Continua* (PNADC) en Brasil y la *Encuesta de Caracterización Socioeconómica Nacional* (CASEN) en Chile utilizan la linealización de Taylor junto con el enfoque del último conglomerado en sus esquemas de muestreo.

- La linealización de Taylor proporciona una aproximación lineal de $\hat{\theta}$, expresada como

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

, donde e_k son variables linealizadas y c es una constante determinística.

- Esta aproximación facilita la expresión del estimador de la varianza de la aproximación lineal de $\hat{\theta}$ como

$$\widehat{Var}(\hat{\theta}) = \sum_{h} \frac{n_h}{n_h - 1} \sum_{i \in s_h} \left(\hat{t}_{e_i} - \bar{\hat{t}}_{e_h} \right)^2$$

.

- Por ejemplo, si se desea estimar una razón, las nuevas variables linealizadas son $e_k = (1/\hat{t}_{y_2})(y_{1k} - \hat{\theta}\ y_{2k})$.

¡Gracias!

Email: andres.gutierrez@cepal.org

Referencias

Gutiérrez, H. A. 2016. *Estrategias de muestreo: diseño de encuestas y estimación de parámetros*. Segunda edición. Ediciones de la U.