Estimación del error de muestreo

Andrés Gutiérrez

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

Tabla de contenidos I

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 $e_k = (1/\hat{t}_{y_2})(y_{1k} - \hat{\theta}\ y_{2k})$.

Pesos replicados

Pesos replicados

- La estrategia de pesos replicados es una técnica computacional aproximada utilizada para estimar la varianza del estimador de interés cuando el proceso de linealización puede resultar complicado.
  1. Dividir la toda la muestra en pequeños subconjutnos (réplicas).
  2. Repetir los mismos procesos de ajuste de ponderadores en cada réplica.
  3. Hacer la estimación en cada subgrupo.
  4. La varianza del estimador se calcula de manera simple como la varianza muestral de todas las estimaciones en cada réplica.
- Esta técnica es útil cuando no se dispone de fórmulas adecuadas para calcular la varianza y se han desarrollado métodos empíricos que ofrecen resultados satisfactorios para propósitos prácticos.

Pesos replicados

- La metodología de pesos replicados permite estimar el error de muestreo sin necesidad de información detallada sobre estratos o UPM en las bases de datos públicas, protegiendo la anonimización de los respondientes.
- Es eficiente y precisa para estimar la varianza de parámetros complejos, utilizada en encuestas como la American Community Survey en EE. UU. y la PNADC de Brasil en América Latina.
- Las tres metodologías principales son los pesos replicados repetidos balanceados, el Jackknife y el Bootstrap, todas ellas basadas en la selección de réplicas de UPMs y reponderación de pesos para mantener la representatividad y obtener estimaciones precisas.
- Los investigadores pueden automatizar el cálculo del error de muestreo sin necesidad de fórmulas matemáticas complejas, facilitando el proceso de análisis de datos en encuestas complejas.

Pesos replicados

- La técnica de Jackknife es efectiva para estimar parámetros lineales, pero no es adecuada para estimar percentiles o funciones de distribución.
- Las réplicas repetidas balanceadas son útiles para estimar parámetros lineales y no lineales, pero pueden tener problemas con dominios pequeños que resultan en estimaciones nulas en la configuración de los pesos.
- El ajuste de Fay a las réplicas repetidas balanceadas soluciona estos problemas, limitando el número de conjuntos de pesos replicados para evitar sobrecargar la publicación de la base de datos.
- El Bootstrap debe utilizarse con precaución, ya que requiere replicar exactamente el diseño de muestreo y construir una población a partir de los pesos de muestreo.

La técnica de Jackknife

- El Jackknife es útil para estimar parámetros lineales y no lineales, excepto para los percentiles.
- Se divide la muestra en grupos iguales y se calcula la varianza del estimador a partir de estas divisiones.
- Calcula $\hat{\theta}_{(a)}$ para cada grupo, que es una estimación del parámetro θ basada en la información restante después de eliminar el grupo a.
- Define $\hat{\theta}_a$ como un pseudovalor de θ utilizando la fórmula

$$A\hat{\theta} - (A-1)\hat{\theta}_{(a)}$$

para $a = 1, 2, ..., A$.

La técnica de Jackknife

- El estimador alternativo de θ, $\hat{\theta}_{JK}$, se obtiene como la media de los pseudovalores $\hat{\theta}_a$ para todos los grupos.

$$\hat{\theta}_{JK} = \frac{1}{A}\sum_{a=1}^{A}\hat{\theta}_a$$

- El estimador de la varianza obtenido mediante Jackknife se obtiene como:

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

- Es posible utilizar como estimador alternativo:

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

La técnica de Jackknife

Para diseños estratificados y multietápicos, el estimador de varianza de Jackknife esta dado por:

$$\widehat{Var}_{JK}(\hat{\theta}) = \sum_{h=1}^{H} \frac{n_{Ih}-1}{n_{Ih}} \sum_{i=1}^{n_{Ih}} (\hat{\theta}_{h(i)} - \hat{\theta})^2$$

donde $\hat{\theta}_{h(i)}$ es la estimación de θ usando los datos de la muestra excluyendo las observaciones en la i-ésima unidad primaria de muestreo (Korn y Graubard 1999, pg. 29 – 30)

La técnica de Jackknife

Los pesos d_{hk}^{i} para cada unidad k en la UPM U_i del estrato U_h se definen de manera específica dependiendo de si la unidad k está o no en la UPM U_h, es decir:

$$d_{hk}^{i} = \begin{cases} 0, \text{ si } U_i \in U_h \text{ y } k \in U_i \\ d_k, \text{ si } k \notin U_h \\ \frac{n_{Ih}}{n_{Ih}-1} d_k, \text{ si } U_i \in U_h \text{ y } k \notin U_i \end{cases}$$

En donde n_{Ih} es el número de UPM seleccionadas en el estrato U_h.

La técnica de Jackknife

- Para reducir el número de pesos replicados, se pueden formar *unidades de varianza* y *estratos de varianza*.
- Las unidades de varianza unen varias UPM dentro de un mismo estrato, emparejándolas según la medida de tamaño.
- El estimador de varianza resultante se calcula como

$$\widehat{Var}_{JK}(\hat{\theta}) = \sum_{h} \frac{n_{Ih} - n_{Ihg}}{n_{Ih}} \sum_{i \in s_{hg}} (\hat{\theta}_{h(g)} - \hat{\theta})^2$$

En donde $\hat{\theta}_{hg}$ es el estimador del parámetro retirando el g-ésimo subgrupo del estrato U_h y n_{Ihg} es el tamaño del subgrupo en la muestra denotado como s_{hg}.

Ejemplo

Tabla 1: Ejemplo reducido de la creación de pesos replicados con la técnica de Jackknife.

k	Estrato	UPM	$d_k^{(1)}$	$d_k^{(2)}$	$d_k^{(3)}$	$d_k^{(4)}$
1	Estrato1	UPM1	0	1,03	1,03	1,03
2	Estrato1	UPM1	0	1,03	1,03	1,03
3	Estrato1	UPM2	1,03	0	1,03	1,03
4	Estrato1	UPM2	1,03	0	1,03	1,03
5	Estrato2	UPM3	1,03	1,03	0	1,03
6	Estrato2	UPM3	1,03	1,03	0	1,03
7	Estrato2	UPM4	1,03	1,03	1,03	0
8	Estrato2	UPM4	1,03	1,03	1,03	0

El método de las réplicas repetidas balanceadas

- El método de las réplicas repetidas balanceadas (BRR) fue desarrollado para diseños donde dos UPM son seleccionadas por estrato.
- Funciona consistentemente para estimaciones de parámetros lineales y no lineales, incluyendo percentiles.
- Asegura máxima dispersión de las UPM a través de las regiones geográficas (estratos).
- Evita la complejidad computacional al definir réplicas al seleccionar aleatoriamente una UPM en cada estrato.

El método de las réplicas repetidas balanceadas

Es posible lograr la misma eficiencia reduciendo el número de pesos replicados utilizando un enfoque ortogonal con matrices de Hadamard. Por ejemplo,

$$\begin{pmatrix} +1 & +1 & +1 & +1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & -1 & +1 \end{pmatrix}$$

El valor $+1$ implica que la primera UPM se mantiene como parte de la réplica y la segunda UPM es retirada de la réplica; el valor -1 implica que la segunda UPM se mantiene como parte de la réplica y la primera UPM es retirada de la réplica

El método de las réplicas repetidas balanceadas

- Las UPM que se mantienen en cada réplica se conocen como *half-samples*.
- El peso de los individuos en la UPM que se mantiene se multiplica por un factor de 2. Entonces, se tiene que

$$d_k = \begin{cases} 0, \text{ si } k \text{ pertence a la UPM que fue retirada} \\ 2d_k, \text{ en otro caso.} \end{cases}$$

El método de las réplicas repetidas balanceadas

Bajo esta metodología BRR, el estimador de la varianza toma la siguiente forma:

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

En donde $\hat{\theta}_a$ es el estimador del parámetro de interés en la réplica a.

Ejemplo

Tabla 2: Ejemplo reducido de la creación de pesos replicados con la técnica de las réplicas repetidas balanceadas.

k	Estrato	UPM	$d_k^{(1)}$	$d_k^{(2)}$
1	Estrato1	UPM1	2	0
2	Estrato1	UPM1	2	0
3	Estrato1	UPM2	0	2
4	Estrato1	UPM2	0	2
5	Estrato2	UPM3	2	0
6	Estrato2	UPM3	2	0
7	Estrato2	UPM4	0	2
8	Estrato2	UPM4	0	2

El método de las réplicas repetidas balanceadas

- En el método BRR, las unidades en dominios con muestra pequeña pueden estar ausentes en algunas combinaciones de pesos replicados, lo que causa una pérdida de precisión en el cálculo del error estándar.
- Para la aplicación de la Réplicas Repetidas Balanceadas es recomendable usar el método de Fay.

$$d_k^a = \begin{cases} \rho * d_k, & \text{si } k \text{ pertence a la UPM que fue retirada} \\ (2-\rho)d_k, & \text{en otro caso.} \end{cases}$$

En donde $0 < \rho < 1$.

El método de las réplicas repetidas balanceadas

Algunos estudios por simulación han mostrado una buena eficiencia para valores de ρ iguales a 0.3, 0.5 o 0.7. Bajo la metodología BRR con el ajuste de Fay, el estimador de la varianza toma la siguiente forma:

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

En donde $\hat{\theta}_a$ es el estimador del parámetro de interés en la réplica a

Ejemplo

Tabla 3: Ejemplo reducido de la creación de pesos replicados con el ajuste de Fay.

k	Estrato	UPM	$d_k^{(1)}$	$d_k^{(2)}$
1	Estrato1	UPM1	1.5	0.5
2	Estrato1	UPM1	1.5	0.5
3	Estrato1	UPM2	0.5	1.5
4	Estrato1	UPM2	0.5	1.5
5	Estrato2	UPM3	1.5	0.5
6	Estrato2	UPM3	1.5	0.5
7	Estrato2	UPM4	0.5	1.5
8	Estrato2	UPM4	0.5	1.5

Método de Bootstrap

- El método de Bootstrap es ampliamente utilizado debido a su facilidad de implementación y flexibilidad en la generación de pesos replicados para el cálculo de errores estándar.
- Consiste en remuestrear las unidades primarias de muestreo seleccionadas desde el marco de áreas, permitiendo calcular estimaciones de indicadores junto con las estimaciones de las varianzas.
- Es eficiente tanto en la estimación de parámetros lineales como no lineales, y funciona bien incluso con tamaños de muestra pequeños, a diferencia del método BRR que requiere al menos dos UPM por estrato.
- El Bootstrap requiere una cantidad significativa de pesos replicados, típicamente más de 200, para obtener resultados confiables.

Método de Bootstrap

Siendo s_{BS} la submuestra Bootstrap, el peso replicado del individuo k perteneciente a la UPM i del estrato h sigue la siguiente expresión:

$$d_k^b = \begin{cases} 0, \text{ si la UPM } i \text{ no pertence a } s_{BS} \\ d_k \left[1 - \sqrt{\frac{n_{Ih}^*}{n_{Ih}-1}} + \sqrt{\frac{n_{Ih}^*}{n_{Ih}-1}} \frac{n_{Ih}}{n_{Ih}^*} n_{Ihi}^*\right], \text{ en otro caso} \end{cases}$$

En donde n_{Ih} es el úmero de UPM en la muestra original del estrato h, n_{Ih}^* es el número de UPM en la muestra Bootstrap y n_{Ihi}^* es el número de veces que la UPM i fue seleccionada en la muestra Bootstrap.

Método de Bootstrap

En este caso se selecciona una muestra Bootstrap con $m_h^* = m_h - 1$, y los pesos toman la siguiente forma

$$d_k^a = \begin{cases} 0, & \text{si la UPM } i \text{ no pertence a } s_{BS} \\ d_k \left[1 - \sqrt{\frac{n_{Ih}^*}{n_{Ih}-1}} + \sqrt{\frac{n_{Ih}^*}{n_{Ih}-1}} \frac{n_{Ih}}{n_{Ih}^*} n_{Ihi}^*\right], & \text{en otro caso} \end{cases}$$

Bajo la metodología Bootstrap (BS), el estimador de la varianza toma la siguiente forma:

$$\widehat{Var}_{BS}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b - \hat{\theta})^2$$

En donde $\hat{\theta}_b$ es el estimador del parámetro de interés en la réplica b inducida por la muestra Bootstrap.

Ejemplo

Tabla 4: Ejemplo reducido de la creación de pesos replicados con la técnica de Bootstrap.

k	Estrato	UPM	$d_k^{(1)}$	$d_k^{(2)}$	$d_k^{(3)}$	$d_k^{(4)}$
1	Estrato1	UPM1	2	0	1	1
2	Estrato1	UPM1	2	0	1	1
3	Estrato1	UPM2	0	2	1	1
4	Estrato1	UPM2	0	2	1	1
5	Estrato2	UPM3	1	1	2	0
6	Estrato2	UPM3	1	1	2	0
7	Estrato2	UPM4	1	1	0	2
8	Estrato2	UPM4	1	1	0	2

Función generalizada de varianzas

- La función generalizada de varianzas (GVF) simplifica el cálculo de las varianzas de los estimadores de muestreo al definir un modelo sobre una medida de probabilidad m en lugar de p.
- Si los parámetros del modelo pueden ser estimados con encuestas pasadas o un conjunto reducido de datos, las varianzas pueden obtenerse evaluando el modelo en los datos actuales.
- En el contexto de estimaciones subnacionales, especialmente cuando hay un gran número de celdas en los resultados, la GVF ofrece una alternativa computacionalmente eficiente.

Función generalizada de varianzas

- Las GVF también son útiles cuando el tamaño de muestra de las subpoblaciones es pequeño o la heterogeneidad en la muestra es insuficiente, evitando estimaciones imprecisas o varianzas iguales a cero.
- El personal técnico de las ONE debe evaluar cuidadosamente las estimaciones de varianza problemáticas y considerar la aplicación de GVF para mejorar la aproximación.
- Las GVF proporcionan una herramienta valiosa para mejorar la precisión de las estimaciones de varianza en situaciones donde los métodos tradicionales pueden fallar.

Función generalizada de varianzas

Reconocer la naturaleza positiva de las varianzas de los estimadores es crucial al modelarlas, y un enfoque log-lineal puede ser beneficioso para manejar esta estructura de datos.

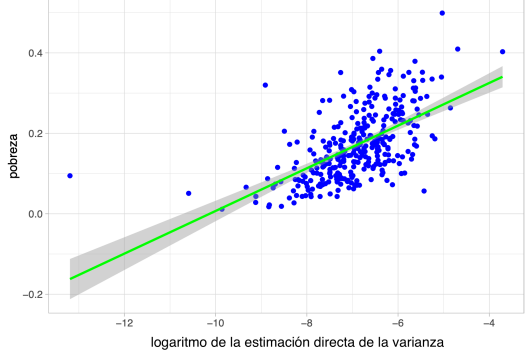

Figura 1: *Relación entre un estimador de la tasa de pobreza estimada y el logaritmo de la estimación directa de su varianza. Fuente: elaboración propia.*

Función generalizada de varianzas

- $Var_{GVF}(\hat{\theta})$ representa la varianza suavizada del estimador directo $\hat{\theta}$.
- La varianza $\widehat{Var}(\hat{\theta})$ no es un valor fijo y no depende estrictamente de las covariables auxiliares.
- Se usa un estimador insesgado $\widehat{Var}(\hat{\theta})$ para suavizar la varianza del estimador directo.

$$E_{mp}\left(\widehat{Var}(\hat{\theta})\right) = E_m\left(E_p\left(\widehat{Var}(\hat{\theta})\right)\right) = E_m(Var(\hat{\theta})) = Var_{GVF}(\hat{\theta})$$

- La igualdad mencionada incluye subíndices m y p que representan la medida de probabilidad del modelo y del diseño de muestreo, respectivamente.

Función generalizada de varianzas

- Los estimadores de varianza del diseño de muestreo son insesgados pero pueden ser inestables con muestras pequeñas, comunes en la desagregación de estimaciones.
- Los modelos de suavizamiento para la estimación de varianzas directas se definen mediante

$$\log(\widehat{Var}(\hat{\theta})) = \mathbf{z}_d'\alpha + \varepsilon_d$$

, donde $\mathbf{z}_d$ son covariables, α son parámetros a estimar, y ε_d son errores aleatorios.

- La estimación suavizada de la varianza de muestreo se obtiene como

$$Var_{GVF}(\hat{\theta}) = \exp(\mathbf{z}_d'\alpha) \cdot \Delta$$

.En donde, $E_{mp}(\varepsilon_d) = \Delta$.

Función generalizada de varianzas

- El estimador insesgado para Δ:

$$\hat{\Delta} = \frac{\sum_{d=1}^{D} \widehat{Var}(\hat{\theta})}{\sum_{d=1}^{D} \exp(\mathbf{z}_d' \alpha)}$$

- La estimación del coeficiente de parámetros de regresión está dada por la siguiente expresión:

$$\hat{\alpha} = \left(\sum_{d=1}^{D} \mathbf{z}_d \mathbf{z}_d' \right)^{-1} \sum_{d=1}^{D} \mathbf{z}_d \log(\widehat{Var}(\hat{\theta}))$$

- El estimador suavizado de la varianza muestral está definido por:

$$\widehat{Var}_{GVF}(\hat{\theta}) = \exp(\mathbf{z}_d' \hat{\alpha}) \hat{\Delta}$$

Función generalizada de varianzas

La conclusión de Rivest y Belmonte (2000) fue que el estimador suavizado de varianza no tiene sesgo, ya que el promedio de las estimaciones suavizadas $\widehat{Var}_{GVF}(\hat{\theta})$ es igual al promedio de las varianzas directas $\widehat{Var}(\hat{\theta})$. Esto se expresa como:

$$\frac{\sum_{d=1}^{D} \widehat{Var}_{GVF}(\hat{\theta})}{D} = \frac{\sum_{d=1}^{D} \widehat{Var}(\hat{\theta})}{D}$$

Función generalizada de varianzas

- Statcan utiliza modelos de regresión para informar sobre el mercado laboral en 149 áreas censales, excluyendo áreas con menos de 10 personas en la fuerza laboral y con varianza directa igual a cero. La estimación de la varianza directa está basada en el diseño de muestreo complejo.
- El modelo de regresión para la varianza suavizada es:

$$\log(\widehat{Var}(\hat{\theta})) = \mathbf{z}_d'\alpha + \varepsilon_d$$

donde

$$\mathbf{z}_d' = \left(1, \log\left(\frac{N_d^{EIB}}{N_d^{15+}}\right), \log\left(1 - \frac{N_d^{EIB}}{N_d^{15+}}\right), \log\left(N_d^{15+}\right)\right)'$$

N_d^{EIB} es el número de beneficiarios del seguro de desempleo en el área d, y N_d^{15+} es el número de personas en la fuerza laboral.

- Para áreas con más de 400 casos efectivos, la estimación suavizada por GVF es igual a la estimación directa; es decir, $\widehat{Var}_{GVF}(\hat{\theta}) = \widehat{Var}(\hat{\theta})$.

Función generalizada de varianzas

- Fuquene et al. (2019) utilizan el enfoque GVF para estimar la prevalencia de migrantes internacionales en los municipios de Colombia. Utilizan un modelo log-lineal con el siguiente vector de covariables auxiliares:

$$\mathbf{z}_d' = \left(1,\ \hat{\theta}_d,\ \sqrt{\hat{\theta}_d},\ n_d,\ \sqrt{n_d},\ \sqrt{\hat{\theta}_d \times n_d}\ \right)'$$

- Fay y Herriot (1979) presenta una aplicación destacada de modelos de estimación en áreas pequeñas.
- El *United States Census Bureau* utilizó un GVF de varianzas para estimar el ingreso percápita a nivel desagregado, tomando resultados de ocho estados y generalizándolos para el resto del país.

Función generalizada de varianzas

- MDSF y CEPAL (2021) utilizó un modelo GVF para estimar las varianzas de las tasas de pobreza comunal a partir de la CASEN 2020 en la región.
- Las comunas con tasa nula de pobreza y, por ende, estimación nula de la varianza del estimador directo, fueron excluidas del ajuste del modelo, pero se obtuvieron las predicciones de sus varianzas.
- En el reporte se justificó la inclusión de las covariables y las relaciones establecidas en el modelo mediante esquemas descriptivos. Además, se mencionó que el factor de ajuste $\hat{\Delta}$ estuvo cercano a 1.2 en todas las series estudiadas.

Consideraciones adicionales sobre la estimación de la varianza de los estimadores de muestreo

Consideraciones adicionales sobre la estimación de la varianza de los estimadores de muestreo

La varianza del estimador HT se calcula como:

$$Var(\hat{t}_{y,\pi}) = \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

donde $\Delta_{kl} = (\pi_{kl} - \pi_k \pi_l)$ y π_{kl} es la probabilidad de inclusión conjunta de los elementos k y l en la muestra s.

Estimadores de la varianza

Bajo diseños de muestreo de tamaño fijo, existen dos estimadores insesgados para esta varianza. El primero, propuesto por Horvitz y Thompson (1952), se define como:

$$\widehat{Var}_1(\hat{t}_{y,\pi}) = \sum_{k \in s} \sum_{l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

El segundo estimador propuesto por Sen (1953) y Yates y Grundy (1953), está dado por la siguiente expresión:

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

Estimaciones negativas de varianza

Estimaciones negativas de varianza

- En teoría, las estimaciones negativas de varianza no tienen sentido y plantean problemas de interpretación.
- En la práctica del diseño de muestreo, especialmente en estructuras poblacionales particulares, es posible obtener estimaciones negativas de varianza.
- Es crucial contar con un equipo experimentado en muestreo que pueda identificar y abordar estas situaciones de manera adecuada.

Ejemplo

Considere el siguiente diseño de muestreo de tamaño fijo e igual a $n = 2$, el cual induce seis posibles muestras.

s	I_1	I_2	I_3	I_4	$p(s)$
s_1	1	1	0	0	0.31
s_2	1	0	1	0	0.20
s_3	1	0	0	1	0.14
s_4	0	1	1	0	0.03
s_5	0	1	0	1	0.01
s_6	0	0	1	1	0.31

Ejemplo

Para esta configuración se obtienen las estimaciones puntuales para cada una de las seis posibles muestras, así como las dos posibles estimaciones de la varianza

s	I_1	I_2	I_3	I_4	$p(s)$	$\hat{t}_{y,\pi}$	$\widehat{Var}_1(\hat{t}_{y,\pi})$	$\widehat{Var}_2(\hat{t}_{y,\pi})$
s_1	1	1	0	0	0.31	9.560440	38.099984	-0.9287681
s_2	1	0	1	0	0.20	5.883191	-4.744190	2.4710422
s_3	1	0	0	1	0.14	4.933110	-3.680428	8.6463858
s_4	0	1	1	0	0.03	7.751323	-100.252974	71.6674365
s_5	0	1	0	1	0.01	6.801242	-165.715154	323.3238494
s_6	0	0	1	1	0.31	3.123994	3.426730	-0.1793659

Estimación de δ

Para evitar estas estimativas negativas, Gutiérrez (2016) afirma que es necesario garantizar que la covarianza (Δ_{kl}) sea negativa para cada par de elementos en la población ($k \neq l$), lo cual no sucede con este esquema de muestreo, puesto que:

$$\Delta_{kl} = \begin{bmatrix} 0.2275 & 0.0825 & -0.1510 & -0.1590 \\ 0.0825 & 0.2275 & -0.1590 & -0.1510 \\ -0.1510 & -0.1590 & 0.2484 & 0.0616 \\ -0.1590 & -0.1510 & 0.0616 & 0.2484 \end{bmatrix}$$

Disminución de la varianza ante el aumento del tamaño de muestra

- Aumentar el tamaño de la muestra generalmente reduce la varianza del estimador, pero hay excepciones.
- Algunas estrategias de muestreo pueden aumentar la varianza a medida que crece el tamaño de la muestra.
- Se presentará un ejemplo para ilustrar esta situación utilizando un diseño de muestreo con tamaños de muestra fijos de $n = 1$ y $n = 2$.

s	I_1	I_2	I_3	$p(s)$
s_1	1	0	0	0.5
s_2	0	1	0	0.1
s_3	0	0	1	0.4

La varianza del estimador de Horvitz-Thompson es igual a $Var(\hat{t}_{y,\pi}) = 1.5$.

Disminución de la varianza ante el aumento del tamaño de muestra

En un segundo caso, considere el siguiente diseño de muestreo de tamaño de muestra $n = 2$:

s	I_1	I_2	I_3	$p(s)$
s_1	1	1	0	0.7
s_2	1	0	1	0.2
s_3	0	1	1	0.1

La varianza del estimador de Horvitz-Thompson es igual a $Var(\hat{t}_{y,\pi}) = 2.3$.

¡Gracias!

Email: andres.gutierrez@cepal.org

Referencias

Fay, R . E., y R. A. Herriot. 1979. «Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data». *Journal of the American Statistical Association* 74 (366a): 269-77. https://doi.org/10.1080/01621459.1979.10482505.

Fuquene, Jairo, Cesar Cristancho, Mariana Ospina, y Domingo Morales. 2019. «Prevalence of international migration: an alternative for small area estimation». *arXiv:1905.00353 [stat]*, abril. http://arxiv.org/abs/1905.00353.

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

Horvitz, D. G., y D. J. Thompson. 1952. «A generalization of sampling without replacement from a finite universe». *Journal of the American Statistical Association* 47: 663-85.

Korn, Edward Lee, y Barry I. Graubard. 1999. *Analysis of health surveys*. Wiley.

MDSF y CEPAL. 2021. *Estimaciones Comunales de Pobreza por ingresos en Chile Mediante Métodos de Estimación en Áreas Pequeñas*. Santiago de Chile: División de Estadísticas de la CEPAL y Observatorio Social del MDSF.

Rivest, Louis-Paul, y Eve Belmonte. 2000. «A conditional mean squared error of small area estimators». *Survey Methodology* 26 (1): 67-78.

Sen, A. R. 1953. «On the estimate of the variance in sampling with varying probabilities». *Journal of the Indian Society of Agricultural Statistics* 5: 119-27.