estimación de la varianza

Muestreo II

Licenciatura en Estadística

2023

- una vez finalizada la recolección de los datos, es una buena práctica computar los errores estándar (raíz cuadrada de la varianza) para medir la precisión de las estimaciones, ya sea, por medio de CV y/o IC
- poder captar todos los componentes que influyen en el SE de la estimación puede llegar a ser difícil o imposible, por lo tanto, la idea es obtener una estimación del SE que capte la mayor parte de la variabilidad de las estimaciones.

componentes del SE

los componentes que afectan el SE de una estimación son:

1. estructura de la población
2. el tamaño de la muestra n. Obviamente!
3. la estrategia de selección (diseño muestral)
4. ajustes por no respuesta
5. método de estimación (e.g HT, GREG, etc.)
6. tipo de parámetro (e.g. total, media, ratio, etc.)

métodos para computar los SE

- exactos
- método del ultimo conglomerado junto con linearización de taylor
- réplicas o remuestreo (e.g. JK1, JKn, Boostrap)

requerimientos para un "buen" estimador

un estimador de Var($\hat{\theta}$) debería:

1. ser (al menos) aproximadamente insesgado
2. estable
3. producir intervalos de confianza con el nivel de cobertura esperado

la varianza del estimador HT para un total Y bajo un diseño muestral cualquiera $p(s)$ viene dada por:

$$\hat{V}(\hat{Y}^{HT}) = \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

cuál es el problema?

- según el tipo de diseño, las probabilidades de inclusión de segundo orden (π_{ij}) pueden ser difícil o **imposibles** de calcular.

estrategias para estimación de la varianza

en una primera instancia vamos a:

- asumir que el muestreo se realizó sin remplazo

esto permite flexibilizar el conocimiento de las probabilidades de inclusión de segundo orden.

ejemplo 1 - SI vs SIR

bajo un SI (muestreo aleatorio simple sin reposición)

$$\hat{Y} = \sum_{i \in s} w_i \times y_i = N \times \bar{y}$$

$$\hat{V}(\hat{Y}) = N^2(1 - f)\frac{\text{var}[y]}{n}$$

ejemplo 1 - SI vs SIR

si la tasa de muestreo $f = n/N$ es cercara a cero, podemos utilizar una estimación sesgada (pero simplificada) asumiendo un SIR (m.a.s c/r)

$$\hat{V}_o = N^2 \frac{\text{var}[y]}{n}$$

cuál es el sesgo del estimador?

$$\frac{E(\hat{V}_0) - V}{V} = \frac{f}{1-f}$$

si $f \approx 0$ el sesgo es despreciable

ejemplo 2

- Asumimos que la muestra fue seleccionada bajo un diseño cualquiera que asigna distinta probabilidad de inclusión (e.g. PPS s/rep).
- Problema? π_{ij} difícil o imposibles de calcular

una aproximación puede ser:

$$\hat{V}_0 = \frac{1}{n(n-1)} \times \sum_{i \in s} (y_i w_i n - \hat{Y})^2$$

conclusiones

- asumir que el muestreo se hizo con remplazo, nos ahorra computar los π_{ij} (si es que podemos)
- si bien el estimador es sesgado, lo es del "lado bueno" ya que sobrestima la varianza y por ende, estamos siendo conservadores.

ejemplo de aproximación de la varianza (sistemático)

NO existe un estimador insesgado de la varianza debido a que algunas probabilidades de inclusión de segundo orden $\pi_{ij} = 0$

dos escenarios:

- si la población no tiene un orden, podemos asumir un muestreo aleatorio simple.
- si la población tiene un orden (utilizando una variable auxiliar x correlacionada con la variable y), podemos crear pseudo-estratos agrupando de a 2 unidades para no sobre-estimar la varianza.

método del ultimo conglomerado

- asume que la mayor variabilidad en la estimación proviene de la primera etapa de muestreo (i.e. selección de las UPM)
- asume que las UPM son seleccionadas al igual que en los casos anteriores con reposición
- en la práctica los diseños en varios etapas incluye algún tipo de estratificación de las UPMs

en este caso V_O toma la siguiente forma:

$$\hat{V}_0(\hat{Y}) = \sum_{h=1}^{H} \frac{1}{m_h(m_h-1)} \times \sum_{j \in s_h} (\hat{Y}_{jh} - \hat{Y}_h)^2$$

donde:

- m_h es la cantidad de UPM seleccionadas en el estrato h
- $\hat{Y}_{jh} = \sum_{s_{jh}} w_i \times y_i$ es la estimación del total de y en la UPM j perteneciente al estrato h
- $\hat{Y}_h$ es el total de la variable y en el estrato h

método del ultimo conglomerado con GREG/CAL

si se utilizan estimadores GREG/CAL se sustituye en la ecuación anterior:

- la variable y por los errores $e = y - \hat{y}$
- los ponderadores originales $w_i = \pi_i^{-1}$ por los ponderadores calibrados $w_i^* = g_i \times w_i$

si el objetivo es estimar el total de la variable y pero para un dominio o área de estimación cualquiera d, el SE se obtiene reemplazando la variable y por la variable extendida y_d la cual, toma el valor y para los casos de la muestra incluidos en el dominio d y cero en otro caso.

Esta técnica se utiliza para añadir una variabilidad extra a las estimaciones, producto de que el tamaño de muestra en el dominio (por ejemplo, personas entre 15 y 24 años) es aleatorio.

linearización de Taylor

- método ampliamente utilizado para aproximar SE de parámetros no lineales (e.g. un ratio)
- la idea es aproximar un estimador no lineal por medio de una función lineal.
- una vez realizado lo anterior y teniendo en cuenta el diseño muestral utilizado, se obtiene una aproximación del SE del estimador.

para el caso del estimador de una tasa o razón $R = Y/Z$, es aproximado utilizando el desarrollo de taylor de primer orden y dicha aproximación queda definida como:

$$\hat{R} \approx R + \frac{1}{Z} \times \sum_{i \in s} w_i \times (y_i - R z_i)$$

el estimador de la varianza se obtiene reemplazando la variable y por una variable $b_i = y_i - \hat{R}z_i$ y añadiendo el término $1/\hat{Z}$ al cuadrado

$$\widehat{\mathrm{SE}}^2(\hat{R}) = \hat{V}(\hat{R}) = \frac{1}{\hat{Z}^2} \times \sum_{h=1}^{H} \frac{1}{m_h(m_h-1)} \times \sum_{j \in s_h} (\hat{B}_{jh} \times m_h - \hat{B}_h)^2$$

donde

- $\hat{Z} = \sum_{i \in s} w_i \times z_i$ es la estimación del total de la variable z.
- $\hat{B}_{jh} = \sum_{s_{jh}} w_i \times b_i$ es la estimación del total de b en la UMP j perteneciente al estrato h.
- $\hat{B}_h = \sum_{h=1}^{H} \hat{B}_{jh}$ es la estimación del total de la variable b en el estrato h.

Al igual que para el caso de un total, el SE de la estimación de un ratio R para un dominio cualquiera d se obtiene de reemplazando en la ecuación anterior las variables y y z por sus correspondientes variables extendidas en el dominio, y_d y z_d.

réplicas o remuestreo

Los métodos de remuestreo o replicación son utilizados en su mayoría para obtener estimaciones de los SE de estimadores no lineales.

La idea es seleccionar un cantidad suficiente de submuestras (réplicas) de la muestra original s y luego resumir las propiedades de un estadístico a través de todas las réplicas.

existen una amplia gama de técnicas:

- Jackknife
- Balanced Repeated Replication (BRR)
- Bootsrap

- todos los métodos seleccionan únicamente submuestras o replica dentro de las UPM incluidas en la muestra original (s)
- **NO** se seleccionan submuestras para unidades (e.g. hogares o personas) dentro de la UPM a la que pertenecen.
- todas las unidades dentro de una UPM son retenidas.
- los pesos w para las unidades incluidas en una réplica son ajustados de distintas formas dependiendo del método de remuestreo

Jackknife

El método de Jackknife para muestreos directos y **sin estratificar** es realizado creando réplicas eliminando una unidad muestral a la vez y "reponderando" las unidades remanentes para producir las estimaciones de la población en cada una de las réplicas.

Jackknife

si la unidad i de una muestra aleatoria de tamaño n es eliminada (dejada afuera), el ponderador para la unidad j incluida en las "retenidas" es:

$$w_{j(i)} = \frac{n}{n-1} w_j$$

Luego, la estimación del total de la variable y computada con la réplica i es:

$$\hat{Y}_{(i)} = \sum_{j \in s_{(i)}} w_{j(i)} \times y_j$$

donde $s_{(i)}$ es la muestra original excluyendo la unidad i

Jackknife

Este proceso se realiza para cada una de las n unidades de la muestra, por lo tanto, conformamos n réplicas.

El estimador Jackknife de la varianza es calculado de la siguiente forma:

$$v_J = \frac{n-1}{n} \times \sum_{i=1}^{n} (\hat{Y}_{(i)} - \hat{Y})^2$$

el estimador anterior es aplicable a diseños no estratificados y es denominado **JK1**

Jackknife

Por ejemplo, bajo un m.a.s (SI) la varianza del estimador $\hat{Y} = N\bar{y}$ queda como:

$$\frac{N^2}{n(n-1)} \times \sum_{i \in s} (y_i - \bar{y})^2$$

Jackknife

importante: cuando estamos bajo diseños en varias etapas, **"eliminar una unidad"** significa **"eliminar una UPM"**.

Al eliminar una UPM, implica que todas las unidades pertenecientes a la misma, son eliminadas de la réplica.

Eso quiere decir que el ponderador de las unidades retenidas, queda definido como su ponderador original w_i multiplicado por el factor $m/(m-1)$

Jackknife

Cuando el diseño es estratificado, el método Jackknife es aplicado dentro de cada estrato.

si la unidad i perteneciente al estrato h de una muestra ade tamaño n_h es eliminada (dejada afuera), el ponderador para la unidad j incluida en las "retenidas" en el estrato h es:

$$w_{j(hi)} = \frac{n_h}{n_h - 1} w_{ih}$$

este método se denomina **JKn**

Bootstrap

- el método de Bootstrap fue introducido por Efron (1982) y el mismo es utilizado en una amplia variedad de campos de la estadística.
- existen muchas variaciones del método original de Efron, pero en general se utiliza la variación de **Rao-Wu**.
- el método de Boootsrap de Rao-Wu se aplica a diseños aleatorio, estratificados y en varia etapas de selección

Bootstrap Rao-Wu

- en cada estrato h se selecciona una muestra aleatoria simple con reposición de tamaño m_h de las UPM entre las UPM seleccionadas originalmente.
- Sea m^*_{hj} la cantidad de veces que la UPM j es seleccionada en el estrato h de tal forma que:

$$\sum_{j=1}^{n_h} m^*_{hj} = m_h$$

- Se crea un peso replicado para cada una de las unidades i que se encuentran incluidos en la UPMs de la muestra inicial, de la siguiente manera:

$$w^*_i = w_i \times (1 - \sqrt{\frac{m_h}{n_h - 1}} + \sqrt{\frac{m_h}{n_h - 1}} \frac{n_h}{m_h} m^*_{hj})$$

Bootstrap Rao-Wu

- los pesos replicados w_i^* son computados para todos las unidades seleccionadas en la muestra original de las UPM independientemente de que se encuentren incluidos en la réplica Bootstrap.
- utilizando los pesos replicados w_i^* se procede a computar la estimación del indicador de interés $\hat{\theta}$.
- se repite el proceso de los puntos anteriores B veces, por lo que se obtienen $\hat{\theta}_{(1)}, \hat{\theta}_{(2)}, ..., \hat{\theta}_{(B)}$ estimaciones.

Bootstrap Rao-Wu

Finalmente, la varianza estimada utilizando Bootstrap Rao-Wu viene dada por:

$$\hat{V}_{\text{boot}}(\hat{\theta}) = \frac{1}{B} \times \sum_{b=1}^{B} (\hat{\theta}_{(b)} - \hat{\theta})^2$$

estimadores GREG/CAL en las réplicas

independientemente del método de remuestreo, para añadir la fuente de la variabilidad introducida por el método de estimación, se debe realizar la calibración para cada una de las réplicas de la misma forma que la calibración con la muestra original s.

no respuesta en el calculo de la varianza

- en los métodos anteriores se parte siempre de la muestra efectiva (ER) en donde los ponderadores ya fueron ajustados por NR.
- la NR tiende a aumentar los SE de los estimadores debido al aumento en la variabilidad de los ponderadores
- para incluir el efecto de la NR, el remuestreo (e.g. Boostrap) se debe realizar sobre la muestra original
- posteriormente se realizan las estimaciones de las propensiones $\hat{\phi}_i$ para cada una de las réplicas.
- generamente, en la práctica, la NR no es incluida en el remuestreo.