MUEST_T13. MÉTODOS SIMPLIFICADOS de ESTIMACIÓN de VARIANZAS en ENCUESTAS COMPLEJAS.

MT. de los GRUPOS ALEATORIOS.

MT. de los CONGLOMERADOS ÚLTIMOS.

MT. de SEMIMUESTRAS REITERADAS.

MT. JACKNIFE.

MT. BOOTSTRAP.

1_ MT. SIMPLIFICADOS de ESTIMACIÓN de VARIANZAS en ENCUESTAS COMPLEJAS

En la práctica, suelen utilizarse esquemas de muestreo polietápico complejos, en los que se estiman muchas características poblacionales. En estos casos, las fórmulas ordinarias de estimación de varianzas son de difícil aplicación, pues conducen a tediosos cálculos.

Este hecho ha llevado al desarrollo de técnicas más sencillas de estimación de varianzas, aunque resulten menos precisas, son de más fácil aplicación.

De entre estas técnicas ya conocemos el mt. de las muestras interpenetrantes, que se utiliza en muestreo sistemático y el mt. de los conglomerados últimos, que volveremos a ver.

A continuación veremos:

Ø_ INTRODUCCIÓN

Muestreo
- Concepto
- Tipos

Muestreo polietápico
- Concepto
- Estimación
- Tma. Madow.
- Tma. Durbin.

1_ MTS_ SIMPLIFICADOS

Caso general → Tma Durbin

Casos particulares → aquí

→ mirar atrás

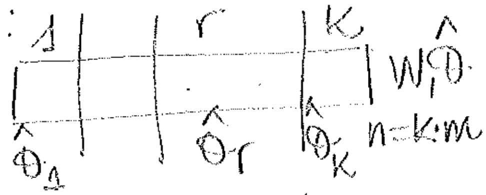
2_ MT. de los GRUPOS ALEATORIOS

A partir de una muestra de tamaño n obtenida de una población finita de N unidades, el mt. consiste en subdividir la muestra en K submuestras de tamaño m, de modo que $n = K \cdot m$. Así, cada grupo aleatorio es una submuestra de la muestra y a su vez una muestra de la población completa, de menor tamaño pero con las mismas propiedades probabilísticas que la muestra completa.

La formación de los K grupos aleatorios de tamaño m dentro de una muestra W de tamaño n puede realizarse considerando una permutación aleatoria de los nº $1, 2, \dots, n$, y eligiendo el primer nº aleatorio formado por los elementos de la muestra que ocupan los lugares definidos por los m primeros números de la permutación. Así sucesivamente.

Si $\hat{\theta}$ es un estimador insesgado de la característica poblacional θ basado en la muestra completa W, y si $\hat{\theta}_r$ es un estimador insesgado de θ basado en el r-ésimo grupo aleatorio, un estimador insesgado de la varianza de $\hat{\theta}$ en muestreo CR es:

$$(CR)\quad \hat{V}(\hat{\theta}) = \frac{1}{K(K-1)} \sum_{r=1}^{K} (\hat{\theta}_r - \hat{\theta})^2$$



(SR) $(1-f)$

El caso más sencillo de aplicación es el de muestreo aleatorio simple con reposición:

El estimador insesgado del total poblacional X es $\hat{X} = N\bar{x}$ con $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$, media muestral de la muestra completa W

$$V(\hat{X}) = V(N\bar{x}) = N^2 V(\bar{x}) = N^2 \cdot \frac{\sigma^2}{n}, \quad \text{con } \sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (X_i - \bar{X})^2$$

Como la r-ésima muestra, o r-ésimo grupo aleatorio $(X_{r1} \dots X_{rm})$ es también una muestra de la poblac. total, podemos estimar el total X_r:

$$\hat{X}_r = N\bar{x}_r = N \cdot \frac{x_r}{m}$$

n = km

$$V(\hat{X}_r) = N^2 V(\bar{x}_r) = N^2 \cdot \frac{\sigma^2}{m} = N^2 \cdot \frac{\sigma^2}{\frac{n}{k}} = k \cdot N^2 \frac{\sigma^2}{n} = kV(\hat{X})$$

Como el r-ésimo grupo aleatorio es también una submuestra de la muestra completa $W = (X_1 \dots X_n)$:

$$E_W(\hat{X}_r) = E_W\left(N \cdot \frac{x_r}{m}\right) = E_W(N\bar{x}_r) = N\bar{x} = \hat{X}$$

luego

$$E[\hat{X}_r] = E E_W(\hat{X}_r) = E[\hat{X}] = X.$$

$\hat{X}_r$ es una copia de $\hat{X}$ para el grupo aleatorio r-ésimo de W, y es un estimador insesgado de X.

$$V(\hat{X}_r) = E(\hat{X}_r - X)^2 = E(\hat{X}_r - \hat{X} + \hat{X} - X)^2 = E(\hat{X}_r - \hat{X})^2 + E(\hat{X} - X)^2$$

$$\Rightarrow E(\hat{X}_r - \hat{X})^2 = V(\hat{X}_r) - V(\hat{X}) = kV(\hat{X}) - V(\hat{X}) = (k-1)V(\hat{X})$$

Sumando en $r = 1 \dots k$ los dos términos de la igualdad:

$$\sum_{r=1}^{k} E(\hat{X}_r - \hat{X})^2 = \sum_{r=1}^{k} (k-1)V(\hat{X}) = k(k-1)V(\hat{X}) \Rightarrow$$

~~por lo que~~ $\Rightarrow V(\hat{X}) = E\left(\frac{\sum_{r=1}^{k} (\hat{X}_r - \hat{X})^2}{k(k-1)}\right)$

luego $\frac{1}{k(k-1)} \sum_{r=1}^{k} (\hat{X}_r - \hat{X})^2$ es un estimador insesgado de $V(\hat{X})$.

– Este mét. también se puede utilizar con grupos aleatorios de distinto tamaño m_r / $\sum_{r=1}^{k} m_r = n$

$$\hat{V}(\hat{\theta}) = \frac{1}{k-1} \sum_{r=1}^{k} \frac{m_r}{n} (\hat{\theta}_r - \hat{\theta})^2$$ estim. insesgado de $V(\hat{\theta})$

– El mt. también puede aplicarse a:

- Muestreo polietápico, formando los K grupos aleatorios con las n unidades primarias que constituyen la muestra completa.
- Muestreo estratificado, tomando K grupos aleatorios en cada estrato
- Estimadores de razón, siempre que sea insesgados.

En el caso de probabilidades iguales y muestreo sin reposición:

$$V(\hat{X}_r) = N^2(1-f) \cdot \frac{S^2}{m} = N^2(1-f) \cdot \frac{S^2}{n/k} = kN^2(1-f)\frac{S^2}{n} = k(1-f)V(\hat{X})$$

luego

$$\hat{V}(\hat{X}_r) = \frac{1-f}{k(k-1)} \sum_{r=1}^{k} (\hat{X}_r - \hat{X})^2$$ es un estim. insesgado de $V(\hat{X})$

La precisión será mayor cuanto mayor sea el nº de grupos aleatorios

3_ MÉTODO de los CONGLOMERADOS ÚLTIMOS

El método de los conglomerados últimos, estudiado en el tema anterior, es un caso particular del método de los grupos aleatorios. Basta hacer $k = n$, es decir tomar cada una de las n observaciones como un grupo aleatorio de tamaño $m = 1$.

$$\hat{V}(\hat{\theta}) = \frac{1}{n(n-1)} \sum_{r=1}^{K} (\hat{\theta}_r - \hat{\theta})^2$$ estim. insesgado de $V(\hat{\theta})$

siendo $\hat{\theta} = \frac{1}{n}\sum_{r=1}^{n} \hat{\theta}_r$

$V(\hat{\theta}_r) = nV(\hat{\theta}) \rightarrow$ se cumple en muestr. con repos.

$E_W(\hat{\theta}_r) = \hat{\theta} \rightarrow \hat{\theta}_r$ copia de de $\hat{\theta}$ para W_r.

$\rightarrow E[\hat{\theta}_r] = \theta$

Este mt. es más preciso que el de grupos aleatorios, porque $n > k$.

Significa que es preferible utilizar n estimadores $\hat{\theta}_i$ individuales que K estimadores $\hat{\theta}_r$ basados en m unidades cada uno.

El mt. de los conglomerados últimos debe su nombre a su aplicación en muestreo polietápico, para considerarlo como un caso particular del muestreo monoetápico de conglom.

Se denomina conglom. último al conjunto de unidades muestrales de ~~pr~~ última etapa seleccionadas en la misma unidad primaria.

Cada conglomerado ~~último~~ primer etapa es un grupo aleatorio, y a partir de su conglomerado último se construye un estimador insesgado $\hat{\theta}_i$ para θ, de modo que el estimador insesgado de θ basado en la muestra completa es:

$$\hat{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i \qquad \text{insesgado de } \theta = \frac{1}{N}\sum_{i=1}^{N}\theta_i$$

En el caso de muestreo polietápico con reposición en 1ªetapa y probabilidades desiguales, el estimador insesgado del total poblacional X es:

$$\hat{\hat{X}}_{HH} = \sum_{i=1}^{n}\frac{\hat{X}_i}{nP_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{X}_i}{P_i} = \frac{1}{n}\sum_{i=1}^{n}\hat{X}_{(i)}$$

donde $\hat{X}_i \rightarrow$ estim. insesgado de X_i

$\hat{X}_{(i)} \rightarrow$ estim. insesgado de X basado en el conglom. último i-ésimo.

$\hat{X}_{(i)}$ son indep. por serlo las muestras de las que proceden

$$\hat{V}(\hat{\hat{X}}_{HH}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{\hat{X}_i}{P_i} - \hat{\hat{X}}_{HH}\right)^2 \text{ es insesgado de } V(\hat{\hat{X}}_{HH})$$

por ser:

$$V(\hat{\hat{X}}_{HH}) = V\left(\sum^{n}\frac{X_i}{nP_i}\right) = \frac{1}{n^2}\sum_{i=1}^{n}V\left(\frac{X_i}{P_i}\right) = \frac{1}{n}V(X_{(i)}) \Rightarrow V(X_{(i)}) = nV(\hat{\hat{X}}_{HH})$$

$$E_W(\hat{X}_{(i)}) = E_W\left(\frac{\hat{X}_i}{P_i}\right) = \sum_{i=1}^{n}\frac{\hat{X}_i}{P_i}\cdot P(\hat{X}_i) = \sum_{i=1}^{n}\frac{\hat{X}_i}{P_i}\cdot\frac{1}{n} = \hat{\hat{X}}_{HH}$$

Aunque este método se utiliza siempre para muestreo CON reposición en primera etapa, se puede aplicar de forma aprox. al muestreo SIN reposición, multiplicando por $(1-f)$.

4_ MT. de las SEMIMUESTRAS REITERADAS

Supongamos una muestra completa W de tamaño n, de la cual extraemos una submuestra aleatoria de tamaño $n/2$ (sp. n par), denominada semimuestra.

Reponiendo la semimuestra, repetimos de manera indep. k veces la selección $\longrightarrow$ se obtienen K semimuestras reiteradas de W, y K estimadores insesgados $\hat{\theta}_r$ de θ.

No se trata de "grupos" como en el mt. grupos aleatorios, pues la unión de las semimuestras reiteradas no coincide con la muestra completa.

~~Si se cumplen las condiciones~~:

Si se verifican las condiciones:

(1) $E_W(\hat{\theta}_r) = \hat{\theta} \quad (\Rightarrow E(\hat{\theta}_r) = \theta)$

(2) $V(\hat{\theta}_r) = 2V(\hat{\theta})$

entonces

$$\hat{V}(\hat{\theta}) = \frac{1}{K}\sum_{r=1}^{K}(\hat{\theta}_r - \hat{\theta})^2$$ es un estim. insesgado de $V(\hat{\theta})$.

- La 2ª condición es obvia en muestreo con reposición, puesto que cada reiteración es de tamaño $n/2$.
- La 1ª condición tb es inmediata, por ser cada reiteración una muestra aleatoria de la muestra completa W.

Para ver la insesgadez del estimador de la varianza:

$$V(\hat{\theta}_r) = E(\hat{\theta}_r - \theta)^2 = E(\hat{\theta}_r - \hat{\theta} + \hat{\theta} - \theta)^2 = E(\hat{\theta}_r - \hat{\theta})^2 + E(\hat{\theta} - \theta)^2$$

porque $E(\hat{\theta}_r - \hat{\theta})(\hat{\theta} - \theta) = E(\hat{\theta}_r \hat{\theta}) - E(\hat{\theta}_r \theta) - E(\hat{\theta}^2) + E(\hat{\theta}\theta) =$

$$= E(\hat{\theta}^2) - \theta^2 - E(\hat{\theta}^2) + \theta^2 = 0$$

luego $V(\hat{\theta}_r) = E(\hat{\theta}_r - \hat{\theta})^2 + E(\hat{\theta} - \theta)^2 \Rightarrow$

$$\Rightarrow E(\hat{\theta}_r - \hat{\theta})^2 = V(\hat{\theta}_r) - E(\hat{\theta} - \theta)^2 = V(\hat{\theta}_r) - V(\hat{\theta}) =$$

$$= 2V(\hat{\theta}) - V(\hat{\theta}) = V(\hat{\theta}).$$

Por lo que

$$\sum_{r=1}^{K} E(\hat{\theta}_r - \hat{\theta})^2 = \sum_{r=1}^{K} V(\hat{\theta}) \Rightarrow E\left(\sum_{r=1}^{K} (\hat{\theta}_r - \hat{\theta})^2\right) = K V(\hat{\theta}) \Rightarrow$$

$$\Rightarrow E\left(\frac{1}{K} \sum_{r=1}^{K} (\hat{\theta}_r - \hat{\theta})^2\right) = V(\hat{\theta})$$

y llegamos a comprobar que el estimador $\frac{1}{K} \sum_{r=1}^{K} (\hat{\theta}_r - \hat{\theta})^2$ es insesgado de $V(\hat{\theta})$.

- La precisión se puede mejorar tomando las K semimuestras reiteradas SIN reposición (para eliminar repetidas), y de modo que no aparezcan semimuestras complementarias (unión = W), pues proporcionan la misma información sobre $(\hat{\theta}_r - \hat{\theta})^2$.

En consecuencia, tomando solamente $\frac{1}{2}\binom{n}{n/2}$ semimuestras también se obtiene $V_W(\hat{\theta}_r)$.

5_ MT. JACKNIFE ó de los estimadores herramentales

Jacknife = Jack-knife = navaja multiusos.

Los estimadores herramentales son estimadores de múltiples usos y de fácil manejo.

El mt. fue desarrollado por Quenouille (1949) para eliminar o reducir el sesgo de ciertos estimadores. Su generalización y el nombre se debe a Tukey (1958), que lo desarrolló como método general de estimación.

Se parte de un estimador $\hat{\theta}_n$ sesgado, cuyo valor numérico se obtiene a partir de una muestra completa de tamaño n, $W=(X_1 \dots X_n)$.

A partir de W se consideran n muestras de tamaño n-1, suprimiendo para muestra i el dato i-ésimo.

$$\left.\begin{array}{lcl} W = (X_1 \dots X_n) & \longrightarrow & \hat{\theta}_n \\ W^{(1)}_{n-1} = (\square\, X_2 \dots X_n) & \longrightarrow & \hat{\theta}^{(1)}_{n-1} \\ W^{(2)}_{n-1} = (X_1 \square \dots X_n) & \longrightarrow & \hat{\theta}^{(2)}_{n-1} \\ \text{---} & \text{---} & \text{---} \\ W^{(n)}_{n-1} = (X_1 X_2 \dots \square) & \longrightarrow & \hat{\theta}^{(n)}_{n-1} \end{array}\right\} \longrightarrow \bar{\theta}(i) \longrightarrow \bar{\theta}_n$$

A partir de los estimadores obtenidos de cada muestra, se definen los n pseudovalores:

$$\bar{\theta}^{(i)} = n\hat{\theta}_n - (n-1)\hat{\theta}^{(i)}_{n-1} \quad , i=1\dots n$$

y el estimador herramental:

$$\bar{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\bar{\theta}^{(i)} = n\hat{\theta}_n - \frac{n-1}{n}\sum_{i=1}^{n}\hat{\theta}^{(i)}_{n-1}$$

$\bar{\theta}_n$ = media muestral de los pseudovalores muestrales

De este modo, si el sesgo de $\hat{\theta}_n$ depende del tamaño de la muestra y es del tipo $\frac{b}{n}$ ó $\frac{b_1}{n}+\frac{b_2}{n^2}+\dots$, puede eliminado o al menos reducido en orden, como resultado de la aplicación de la herramienta. En efecto:

$$\left.\begin{array}{l} E[\hat{\theta}_n] = \theta + \frac{b}{n} \\ E[\hat{\theta}^{(i)}_{n-1}] = \theta + \frac{b}{n-1} \end{array}\right\} E(\bar{\theta}_n) = n\,E[\hat{\theta}_n] - \frac{n-1}{n}\cdot\sum_{i=1}^{n} E[\hat{\theta}^{(i)}_{n-1}] =$$

$$= n(\theta + b/n) - \frac{n-1}{n}\cdot n\left(\theta + \frac{b}{n-1}\right) =$$

$$= n\theta + b - (n-1)\theta - b = \theta$$

Una estimación de la varianza de $\bar{\theta}_n$ es:

$$\hat{V}(\bar{\theta}_n) = \frac{n-1}{n}\sum_{i=1}^{n}\left(\bar{\theta}^{(i)}_{n-1} - \bar{\theta}_n\right)^2$$

- Las expresiones anteriores pueden generalizarse suprimiendo dos o más elementos consecutivos en la muestra inicial.

- El mt. Jacknife tb se puede utilizar para mejorar la varianza muestral como estimador de la var. poblacional ó para mejorar los estimadores de razón, que son sesgados.

6_ Métodos BOOTSTRAP ó de autogeneración

Bootstrap ≡ correa que sirve para ayudar a ponerse las botas.

El mt. de autogeneración (bootstrap) se emplea para la estimación aproximada de sesgos, precisiones, intervalos de confianza, regresiones, etc., generalmente a partir de los datos de una sola muestra.

Se debe a Efron (1982) y se utiliza cuando se desconoce la distribución poblacional.

A partir de una muestra aleatoria con reposición de tamaño n, $W = (X_1 \dots X_n)$, donde X_i se consideran v.a.i.i.d, se construye la función de distrib. empírica de la muestra F_n, se obtiene $\hat{\theta}(F_n)$, estimador del parámetro θ con distrib. de probabilidad desconocida.

Como F_n asigna la frecuencia $\frac{1}{n}$ a cada valor muestral, podemos considerar la muestra W como una población donde $X_1 \dots X_n$ son los valores que toma la variable con probab $\frac{1}{n}$, de modo que $\hat{\theta}(F_n)$ es su parámetro.

A partir de la muestra inicial W se toma una muestra de tamaño n con reposición $W^* = (X_1^* \dots X_n^*)$ y se obtiene la función de distrib. empírica autogenerada F_n^* y el estimador $\hat{\theta}^*$.

Este proceso se repite, de manera independiente, un gran nº de veces, $M \longrightarrow \hat{\theta}_1^* \dots \hat{\theta}_M^*$

El estimador bootstrap se obtiene como promedio de los estimadores de las muestras autogeneradas:

$$\hat{\theta}_{BOOT} = \frac{1}{M}\sum_{j=1}^{M}\hat{\theta}_j^*$$

$$V(\hat{\theta}_{BOOT}) = \frac{1}{M}\sum_{j=1}^{M}(\hat{\theta}_j^* - \hat{\theta}_{BOOT})^2$$

que se puede estimar como:

$$\hat{V}(\hat{\theta}_{BOOT}) = \frac{1}{M-1}\sum_{j=1}^{M}(\hat{\theta}_j^* - \hat{\theta}_{BOOT})^2 = \frac{1}{M-1}\left[\sum_{j=1}^{M}\hat{\theta}_j^{*2} - M\left(\sum_{j=1}^{M}\frac{\hat{\theta}_j^*}{M}\right)^2\right].$$

Los experimentos efectuados por simulación y por mét. computación, ponen de manifiesto que la autogeneración presenta propiedades que la hacen deseable en cuanto a precisión y rapidez de cálculo, pero da malos resultados para un pequeño porcentaje de muestras posibles.

Los estimadores bootstrap, al igual que los estimadores jacknife, se basan en la obtención de datos ficticios a partir de datos originales, y estiman la variabilidad de un estimador basándose en su variabilidad sobre los conjuntos de datos ficticios, por lo que dan además una idea de la distrib. en el muestreo del estadístico en estudio.