MUEST_T12. ESTIMACIÓN de VARIANZAS de estim. lineales en el muestreo de conglomerados con submuestreo.
TEOREMAS I y II de DURBIN.
APLICACIÓN AL MUESTREO SIN REPOSICIÓN y PROBAB. DESIGUALES en PRIMERA ETAPA.
ESTIMACIÓN de la VARIANZA en el MUESTREO CON REPOSICIÓN y PROBAB. DESIGUALES

1- MUESTREO de CONGLOMERADOS CON SUBMUESTREO

El muestreo de conglomerados con submuestreo ó muestreo de conglomerados bietápico surge cuando al realizar muestreo de conglomerados monoetápico ~~y~~ las unidades elegidas para la muestra presentan homogeneidad dentro de ellas, es decir, dentro de cada conglomerado basta con ~~elegir~~ observar un pequeño nº de unidades elementales para obtener una muestra representativa $\longrightarrow$ + barato y + rápido.

Las muestras así obtenidas se llaman bietápicas, por haberse obtenido en dos etapas:

Etapa 1: Selección de unidades primarias ó conglomerados.

N conglom $\longrightarrow$ n conglom. de tamaño M_i

La selección puede ser CON o SIN reposición.

Etapa 2: Selección de unidades elementales.

Para cada conglomerado obtenido en la etapa 1, se seleccionan de manera aleatoria m_i unidades, ~~inde~~ de forma indep. para cada conglom.

Congl. i $(i=1 \dots n)$: M_i unid $\longrightarrow$ m_i unid.

La selección puede ser con o sin reposición, pero normalmente se utiliza m.a.s. (sin reposición y con probab. iguales).

Según el tipo de muestreo considerado en la etapa 1, se obtienen los estimadores lineales insesgados del parámetro poblacional. Sus varianzas se pueden calcular ~~a po~~ utilizando el teorema de Madow.

a) Etapa 1 SIN reposición:

Si consideramos la unidad muestral primaria i-ésima de muestreo como una población, siendo $\hat{X}_i$ una estim. de su total al considerar el submuestreo, y representamos por $\bar{x}_i$ un estimador insesgado, podemos aplicar la expresión general de estimador de Horvitz y Thompson:

$$\hat{\hat{X}}_{HT} = \sum_{i=1}^{n} \frac{\hat{X}_i}{\pi_i} = \sum_{i=1}^{n} \frac{M_i \bar{x}_i}{\pi_i}$$

Es insesgado:

$$E[\hat{\hat{X}}_{HT}] = E_1 \sum_{i=1}^{n} \frac{M_i E_2[\bar{x}_i]}{\pi_i} = E_1 \sum_{i=1}^{n} \frac{M_i \bar{X}_i}{\pi_i} = E_1[\hat{X}_{HT}] = X$$

Casos particulares serían:

$$M_i = \bar{M}, \forall i \longrightarrow \hat{\hat{X}}_{HT} = \bar{M} \sum_{i=1}^{n} \frac{\bar{x}_i}{\pi_i}$$

$$\pi_i = \frac{nM_i}{M} \longrightarrow \hat{\hat{X}}_{HT} = \frac{M}{n} \sum_{i=1}^{n} \frac{M_i \bar{x}_i}{M_i} = \frac{M}{n} \sum_{i=1}^{n} \bar{x}_i$$

$$\pi_i = \frac{n}{N}, \forall i \longrightarrow \hat{\hat{X}}_{HT} = \frac{N}{n} \sum_{i=1}^{n} M_i \bar{x}_i$$

Para hallar la varianza del estimador general del total en muestreo bietápico de conglomerados, se utiliza el teorema de Madow:

$$V(\hat{\hat{X}}_{HT}) = V_1 E_2(\hat{\hat{X}}_{HT}) + E_1 V_2(\hat{\hat{X}}_{HT})$$

$$V_1 E_2(\hat{\hat{X}}_{HT}) = V_1 E_2\left(\sum_{i=1}^{n} \frac{\hat{X}_i}{\pi_i}\right) = V_1\left(\sum_{i=1}^{n} \frac{E_2(\hat{X}_i)}{\pi_i}\right) = V_1\left(\sum_{i=1}^{n} \frac{X_i}{\pi_i}\right) =$$

$$= V_1(\hat{X}_{HT}) \overset{SR}{=} \sum_{i=1}^{N} \frac{X_i^2}{\pi_i^2} \pi_i(1-\pi_i) + \sum_{i \neq j}^{N} \frac{X_i}{\pi_i} \cdot \frac{X_j}{\pi_j} (\pi_{ij} - \pi_i \pi_j)$$

$$E_1 V_2(\hat{\hat{X}}_{HT}) = E_1 V_2\left(\sum_{i=1}^{n} \frac{\hat{X}_i}{\pi_i}\right) = E_1\left(\sum_{i=1}^{n} \frac{V_2(\hat{X}_i)}{\pi_i^2}\right) = E_1\left(\sum_{i=1}^{n} \frac{1}{\pi_i^2}(1-f_{2i}) M_i^2 \cdot \frac{S_i^2}{m_i}\right) =$$

$$= \sum_{i=1}^{N} \frac{1}{\pi_i^2} (1-f_{2i}) \cdot M_i^2 \cdot \frac{S_i^2}{m_i} \underbrace{E[e_i]}_{\substack{\pi_i \\ (SR)}} = \sum_{i=1}^{N} \frac{1}{\pi_i} (1-f_{2i}) M_i^2 \cdot \frac{S_i^2}{m_i}$$

luego $V(\hat{\hat{X}}_{HT}) = \sum_{i=1}^{N} \frac{X_i^2}{\pi_i^2}\pi_i(1-\pi_i) + \sum_{j\neq i}^{N} \frac{X_i}{\pi_i}\frac{X_j}{\pi_j}(\pi_{ij}-\pi_i\pi_j) +$

$$+ \sum_{i=1}^{N} \frac{(1-f_{2i})\cdot M_i^2 S_i^2}{\pi_i \; m_i}$$

$V(\hat{\hat{X}}_{HT}) = V_1(\hat{X}_{HT})$ + penalización debida al submuestreo.

Para prob. iguales: $\pi_i = \frac{n}{N}$, $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$

$$V(\hat{\hat{X}}_{HT}) = \dots = N^2(1-f_1)\cdot\frac{S_1^2}{n} + \sum_{i=1}^{N} \frac{(1-f_{2i})M_i^2 S_i^2}{m_i \frac{n}{N}}$$

La estimación de la varianza se obtiene a partir del teo. de Durbin.

b) <u>Etapa 1 CON reposición</u>:

Si consideramos la unidad muestral primaria i-ésima como una población, siendo $\hat{X}_i$ una estimación del total considerando el submuestreo, y representamos por $\bar{x}_i$ un estimador insesgado de su media, podemos aplicar la expresión del estimador general de Hansen y Hurwitz $\hat{\hat{X}}_{HH}$:

$$\hat{\hat{X}}_{HH} = \sum_{i=1}^{n} \frac{\hat{X}_i}{nP_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{X}_i}{P_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{M_i\bar{x}_i}{P_i}$$

Es insesgado:

$$E[\hat{\hat{X}}_{HH}] = E_1\left(\sum_{i=1}^{n}\frac{1}{n}\frac{M_i E_2(\bar{x}_i)}{P_i}\right) = E_1\left(\sum_{i=1}^{n}\frac{1}{n}\cdot\frac{M_i\bar{X}_i}{P_i}\right) = E_1\left(\sum_{i=1}^{n}\frac{X_i}{nP_i}\right) = E_1(\hat{X}_{HH}) = X$$

Casos particulares:

$M_i = \bar{M}, \forall i \longrightarrow \hat{\hat{X}}_{HH} = \frac{\bar{M}}{n}\sum_{i=1}^{n}\frac{\bar{x}_i}{P_i}$

$P_i = \frac{M_i}{M}$ $\longrightarrow \hat{\hat{X}}_{HH} = \frac{M}{n}\sum_{i=1}^{n}\bar{x}_i$

$P_i = \frac{1}{N}, \forall i \longrightarrow \hat{\hat{X}}_{HH} = \frac{N}{n}\sum_{i=1}^{n} M_i\bar{x}_i$

Para calcular la varianza del estimador general del total, se utiliza tb. el teo. de Madow:

$$V(\hat{\hat{X}}_{HH}) = V_1 E_2(\hat{\hat{X}}_{HH}) + E_1 V_2(\hat{\hat{X}}_{HH})$$

$$V_1 E_2(\hat{\hat{X}}_{HH}) = V_1 E_2\left(\sum_{i=1}^{n} \frac{\hat{X}_i}{nP_i}\right) = V_1\left(\sum_{i=1}^{n} \frac{E_2(\hat{X}_i)}{nP_i}\right) = V_1\left(\sum_{i=1}^{n} \frac{X_i}{nP_i}\right) =$$

$$= V_1(\hat{X}_{HH}) = \frac{1}{n}\left(\sum_{i=1}^{N} \frac{X_i^2}{P_i} - X^2\right) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{X_i}{P_i} - X\right)^2 P_i$$

$$E_1 V_2(\hat{\hat{X}}_{HH}) = E_1 V_2\left(\sum_{i=1}^{n} \frac{\hat{X}_i}{nP_i}\right) = E_1\left(\sum_{i=1}^{n} \frac{V_2(\hat{X}_i)}{n^2 P_i^2}\right) = E_1 \sum_{i=1}^{n} \frac{1}{n^2 P_i^2} \cdot V_2(\hat{X}_i) =$$

$$= E_1\left(\sum_{i=1}^{n} \frac{1}{n^2 P_i^2} \cdot M_i^2 (1 - f_{2i}) \cdot \frac{S_i^2}{m_i}\right) = E_1\left[\sum_{i=1}^{N} \frac{1}{n^2 P_i^2} M_i^2 (1 - f_{2i}) \cdot \frac{S_i^2}{m_i} \cdot e_i\right] =$$

$$= \sum_{i=1}^{N} \frac{1}{n^2 P_i^2} \cdot M_i^2 (1 - f_{2i}) \cdot \frac{S_i^2}{m_i} \underbrace{E_1[e_i]}_{nP_i} \underset{\substack{\uparrow \\ SR}}{=} \sum_{i=1}^{N} \frac{1}{nP_i} M_i^2 (1 - f_{2i}) \frac{S_i^2}{m_i}$$

por lo que:

$$V(\hat{\hat{X}}_{HH}) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{X_i}{P_i} - X\right)^2 P_i + \sum_{i=1}^{N} \frac{M_i^2 (1 - f_{2i})}{nP_i} \cdot \frac{S_i^2}{m_i}$$

$V(\hat{\hat{X}}_{HH}) = V_1(\hat{X}_{HH})$ + penalización por submuestreo

Para prob. iguales: $P_i = \frac{1}{N}$

Para prob. proporcionales a los tamaños: $P_i = \frac{M_i}{M}$, $M = \sum_{i=1}^{N} M_i$

La estimación de la varianza se hace utilizando el teor. de Durbin

2_ TEOREMAS I y II de DURBIN.

El teorema I de Durbin proporciona una expresión general para la varianza de un estimador lineal insesgado en muestreo bietápico, siendo válido el resultado para muestreo con y sin reposición.

El tma I de Durbin tb. es aplicable en muestreo polietápico con sucesivas submuestras en varias etapas; y además es uno de las hipótesis básicas del tma II de Durbin.

El teorema II de Durbin proporciona una expresión general para la estimación ~~general~~ insesgada de la varianza de un estimador lineal insesgado en muestreo bietápico con selección SIN reposición en la primera etapa.

Cuando el muestreo de la primera etapa es CON reposic. existen otros mts para ~~calcular~~ estimar la varianza de los estimadores → tema sgte.

Por coherencia, cambiar $u_i \to c_i$

- En la primera etapa, sobre una población finita de N unidades $\{u_1 \dots u_N\}$ definimos la v.a. auxiliar e_i como el nº de veces que aparece u_i en la muestra

(SR) → $e_i \to Ber(\pi_i)$ / $E_1[e_i] = \pi_i$

(CR) → $e_i \to B(n, P_i)$ / $E_1[e_i] = nP_i$

~~El parámetro poblacional $\theta =$~~

Para generalizar llamamos $\pi_i = E_1[e_i] = \begin{cases} \pi_i & (SR) \\ nP_i & (CR) \end{cases}$

Podemos expresar el parámetro poblacional de forma general como:

$$\cancel{\theta = \sum_{i=1}^{N} Y_i \pi_i} \qquad \theta = \sum_{i=1}^{N} Y_i$$

cuyo estimador lineal insesgado es

$$\cancel{\hat{\theta} = \sum_{i=1}^{n} Y_i} \qquad \hat{\theta} = \sum^{n} \frac{Y_i}{\pi_i}$$

$E_1(\hat{\theta}) = \theta$.

$E_1(\hat{\theta}) = E_1\left[\sum^N Y_i e_i\right] = \sum^N Y_i E_1(e_i) = \sum^N Y_i \pi_i$

⟹ MIRAR ATRÁS

Etapa 1 : $\{C_1 - C_N\} \xrightarrow[CR]{SR} \{C_1 - C_n\}$

$$\theta = \sum_{i=1}^{N} Y_i \longrightarrow \hat{\theta} = \sum_{i=1}^{n} \frac{Y_i}{\pi_i}, \text{ donde } \begin{cases} \pi_i = \pi_i \ (SR) \\ \pi_i = nP_i \ (CR) \end{cases}$$

$$E_1[\hat{\theta}] = E_1\left[\sum^{n} \frac{Y_i}{\pi_i}\right] = E_1\left[\sum^{N} \frac{Y_i}{\pi_i} e_i\right] = \sum^{N} \frac{Y_i}{\pi_i} E_1(e_i) = \sum^{N} Y_i = \theta$$

porque $e_i \longrightarrow \begin{cases} B(1, \pi_i) \ SR \ / \ E_1[e_i] = \pi_i \\ B(n, P_i) \ CR \ / \ E_1[e_i] = nP_i \end{cases}$

Etapa 2 : Para cada $C_i \xrightarrow{\text{mas sr}} \{u_{i1} - u_{iM_i}\}$

$\theta_i = Y_i \longrightarrow \hat{\theta}_i = \hat{Y}_i$ estim. lineal insesgado

Et 1 + Et 2 :

$$\theta = \sum_{i=1}^{N} Y_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij} \longrightarrow \hat{\hat{\theta}} = \sum_{i=1}^{n} \frac{\hat{Y}_i}{\pi_i} \ / \ E[\hat{\hat{\theta}}] = \theta$$

$$E[\hat{\hat{\theta}}] = E_1 E_2(\hat{\hat{\theta}}) = E_1\left[\sum^{n} \frac{1}{\pi_i} E_2(\hat{Y}_i)\right] = E_1\left[\sum^{n} \frac{1}{\pi_i} Y_i\right] =$$

$$= E_1\left[\sum^{N} \frac{1}{\pi_i} Y_i e_i\right] = \sum^{N} \frac{Y_i}{\pi_i} \underbrace{E_1[e_i]}_{= \pi_i} = \sum^{N} Y_i = \theta.$$

- En la segunda etapa, ~~si~~ $\hat{Y}_i$ es un estimador insesgado de Y_i obtenido al aplicar submuestreo en la unidad primaria u_i. (Se puede extender a muestreo polietápico).

$$E_2(\hat{Y}_i) = Y_i$$

Entonces $\hat{\hat{\Theta}} = \sum_{i=1}^{n} \hat{Y}_i / \pi_i$ es un estimador lineal insesgado de Θ.

$$E[\hat{\hat{\Theta}}] = E_1E_2(\hat{\hat{\Theta}}) = E_1E_2\left(\sum^{n} \hat{Y}_i\right) = E_1\left(\sum^{n} E_2(\hat{Y}_i)\right) = E_1\left(\sum^{n} Y_i\right) =$$
$$= E_1\left[\sum^{N} Y_i e_i\right] = \sum^{N} Y_i E_1(e_i) = \sum^{N} Y_i \pi_i = \Theta.$$

Teorema I de Durbin

El teorema I de Durbin asegura que la varianza de $\hat{\hat{\Theta}}$ tiene dos componentes:

- la varianza en primera etapa de $\hat{\Theta}$
- la suma poblac. ponderada de las varianzas en segunda etapa de cada unidad primaria, ponderada por $E_1(e_i)$

$$\text{TID} \longrightarrow V(\hat{\hat{\Theta}}) = V_1(\hat{\Theta}) + \sum_{i=1}^{N} V_2(\hat{Y}_i)\cdot \pi_i$$

Dem: Utilizando la descomp. de la varianza del teorema de Madow:

$$V(\hat{\hat{\Theta}}) = E_1V_2[\hat{\hat{\Theta}}] + V_1E_2[\hat{\hat{\Theta}}]$$

son indep.

$$\rightarrow \begin{cases} V_2(\hat{\hat{\Theta}}) = V_2\left(\sum_{i=1}^{n} \frac{\hat{Y}_i}{\pi_i}\right) = \sum_{i=1}^{n} \frac{1}{\pi_i^2} V_2(\hat{Y}_i) = \sum_{i=1}^{N} V_2(\hat{Y}_i)\cdot e_i \\ E_1V_2(\hat{\hat{\Theta}}) = E_1\left(\sum_{i=1}^{N} V_2(\hat{Y}_i)e_i\right) = \sum_{i=1}^{N} V_2(\hat{Y}_i)E_1(e_i) = \sum_{i=1}^{N} V_2(\hat{Y}_i)\pi_i \end{cases}$$

$$E_2[\hat{\hat{\Theta}}] = E_2\left(\sum_{i=1}^{n} \hat{Y}_i\right) = \sum_{i=1}^{n} E_2(\hat{Y}_i) = \sum_{i=1}^{n} Y_i = \hat{\Theta}$$

$$V_1E_2[\hat{\hat{\Theta}}] = V_1(\hat{\Theta})$$

luego $V(\hat{\hat{\Theta}}) = V_1(\hat{\Theta}) + \sum_{i=1}^{N} V_2(\hat{Y}_i)\pi_i$ cqd.

Tal y como lo hago yo no sale, pero sólo es cuestión de notación

$\Rightarrow$ DEJARLO COMO ESTÁ

Teorema II de Durbin: Para E1SR

~~Si se cumplen~~:
Si $\hat{\hat{\theta}}$ verifica el tma I de Durbin, una estimación insesgada de la varianza de $\hat{\hat{\theta}}$ es:

$$\hat{V}(\hat{\hat{\theta}}) = \hat{V}_c(\hat{\hat{\theta}}) + \sum_{i=1}^{n} \hat{V}_2(\hat{Y}_i) \cdot \pi_i$$

donde:

$\hat{V}_c(\hat{\hat{\theta}}) \equiv$ copia de la varianza $\hat{V}_1(\hat{\theta})$, estimación insesgada de $V_1(\hat{\theta})$, cambiando Y_i por $\hat{Y}_i$.

$\hat{V}_2(\hat{Y}_i) \equiv$ estimador insesgado de $V_2(\hat{Y}_i)$, $i = 1 \ldots n$.

Dem:

En el caso de muestreo sin reposición (TIID no es aplicable CR)

$$\hat{\theta} = \sum_{i=1}^{n} Y_i \Rightarrow \boxed{V_1(\hat{\theta})} = V_1\left(\sum_{i=1}^{n} Y_i\right) = V_1\left(\sum_{i=1}^{N} Y_i e_i\right) =$$
$$= \sum_{i=1}^{N} Y_i^2 V(e_i) + \sum_{i \neq j}^{N} Y_i Y_j \operatorname{cov}(e_i, e_j) =$$
$$= \sum_{i=1}^{N} Y_i^2 \pi_i (1-\pi_i) + \sum_{i \neq j}^{N} Y_i \cdot Y_j (\pi_{ij} - \pi_i \pi_j)$$

$$\boxed{\hat{V}_1(\hat{\theta})} = \sum_{i=1}^{n} Y_i^2 \cdot \frac{\pi_i(1-\pi_i)}{\pi_i} + \sum_{j \neq i}^{n} Y_i Y_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}}$$ es un estimador insesgado de $V_1(\hat{\theta})$ [se demuestra directamente].

Cambiando Y_i por $\hat{Y}_i$ obtenemos la copia:

$$\boxed{\hat{V}_c(\hat{\hat{\theta}})} = \sum_{i=1}^{n} \hat{Y}_i^2 \cdot \frac{\pi_i(1-\pi_i)}{\pi_i} + \sum_{j \neq i}^{n} \hat{Y}_i \hat{Y}_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}}$$

Si le sumamos $\sum_{i=1}^{n} \hat{V}_2(\hat{Y}_i) \pi_i$ obtenemos una estimación insesgada de $V(\hat{\hat{\theta}})$.

$$\hat{V}(\hat{\Theta}) = \sum_{i=1}^{n} \hat{Y}_i^2 (1-\pi_i) + \sum_{j\neq i}^{n} \hat{Y}_i \hat{Y}_j \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}} + \sum_{i=1}^{n} \hat{V}_2(\hat{Y}_i)\pi_i$$

estimación insesgado de

$$V(\hat{\Theta}) = V_1(\hat{\Theta}) + \sum_{i=1}^{n} V_2(\hat{Y}_i)\pi_i$$

Lo demostremos por partes:

$$E[\hat{V}(\hat{\Theta})] = E\Big[\underbrace{\sum_{i=1}^{n} \hat{Y}_i^2(1-\pi_i) + \sum_{j\neq i}^{n} \hat{Y}_i\hat{Y}_j \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}}_{①}\Big] + E\Big[\underbrace{\sum^{n} \hat{V}_2(\hat{Y}_i)\pi_i}_{②}\Big]$$

$$= \sum_{i=1}^{n} Y_i \qquad V_1(\hat{\Theta}) \qquad + \sum_{i=1}^{n} V_2(\hat{Y}_i)\pi_i$$

① $E\Big[\sum_{i=1}^{n} \hat{Y}_i^2 (1-\pi_i) + \sum_{j\neq i}^{n} \hat{Y}_i\hat{Y}_j \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\Big] = E_1 E_2 [\qquad] =$

$$= E_1\Big[\sum_{i=1}^{n}(1-\pi_i)\underbrace{E_2(\hat{Y}_i^2)}_{V_2(\hat{Y}_i) + (\underbrace{E_2(\hat{Y}_i)}_{Y_i})^2} + \sum_{j\neq i}^{n} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}} \underbrace{E_2(\hat{Y}_i\hat{Y}_j)}_{\text{indep.}\; E_2(\hat{Y}_i)E_2(\hat{Y}_j)}\Big] =$$

$$= E_1\Big[\sum_{i=1}^{N}(1-\pi_i)(V_2(\hat{Y}_i) + Y_i^2)e_i + \sum_{j\neq i}^{N} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}} Y_iY_j \underbrace{e_ie_j}_{\pi_{ij}}\Big] =$$

$$= \sum_{i=1}^{N}(V_2(\hat{Y}_i) + Y_i^2)\pi_i(1-\pi_i) + \sum_{j\neq i} Y_iY_j(\pi_{ij}-\pi_i\pi_j)) =$$

② $E\Big[\sum_{i=1}^{n}\hat{V}_2(\hat{Y}_i)\pi_i\Big] = E_1E_2\Big[\sum_{i=1}^{n}\hat{V}_2(\hat{Y}_i)\pi_i\Big] = E_1\Big[\sum_{i=1}^{n}\pi_i E_2[\hat{V}_2(\hat{Y}_i)]\Big] =$

$$= E_1\Big(\sum_{i=1}^{n}\pi_i V_2(\hat{Y}_i)\Big) = E_1\Big[\sum_{i=1}^{N} V_2(\hat{Y}_i)\pi_i e_i\Big] =$$

$$= \sum_{i=1}^{N} V_2(\hat{Y}_i)\pi_i \underbrace{E_1(e_i)}_{\pi_i} = \sum_{i=1}^{N} V_2(\hat{Y}_i)\pi_i^2$$

luego $E[\hat{V}(\hat{\Theta})] = \sum_{i=1}^{N} Y_i^2\pi_i(1-\pi_i) + \sum_{j\neq i}^{N} Y_iY_j(\pi_{ij}-\pi_i\pi_j) +$ ← $V_1(\hat{\Theta})$

$$+ \sum_{i=1}^{N} V_2(\hat{Y}_i)\underbrace{\pi_i(1-\pi_i)}_{\pi_i-\pi_i^2} + \sum_{i=1}^{N} V_2(\hat{Y}_i)\pi_i^2 =$$

$$= V_1(\hat{\Theta}) + \sum_{i=1}^{N} V_2(\hat{Y}_i)\pi_i = V(\hat{\Theta}) \quad \text{cqd.}$$

3. APLICACIÓN al muestreo SIN REPOSICIÓN y con PROBAB. DESIGUALES EN PRIMERA ETAPA

En nuestro desarrollo anterior, $\Theta = \sum_{i=1}^{N} Y_i \pi_i$

El total poblacional es $X = \sum_{i=1}^{N} X_i \longrightarrow Y_i = \frac{X_i}{\pi_i}$

cuyo estimador es $\hat{\hat{\Theta}} = \sum_{i=1}^{n} Y_i \longrightarrow \hat{\hat{X}} = \sum_{i=1}^{n} \frac{\hat{X}_i}{\pi_i} = \hat{\hat{X}}_{HT}$

Aplicando el tma I de Durbin resulta:

$V(\hat{\hat{\Theta}}) = V_1(\hat{\Theta}) + \sum_{i=1}^{N} V_2(\hat{Y}_i)\cdot\pi_i$

TID $\longrightarrow V(\hat{\hat{X}}_{HT}) = V_1(\hat{X}_{HT}) + \sum_{i=1}^{N} \frac{V_2(\hat{X}_i)}{\pi_i}$

ya que $V_2(\hat{Y}_i) = V_2\left(\frac{\hat{X}_i}{\pi_i}\right) = \frac{1}{\pi_i^2} V_2(\hat{X}_i) \Rightarrow \pi_i V_2(\hat{Y}_i) \longrightarrow \frac{V_2(\hat{X}_i)}{\pi_i}$

con $V_1(\hat{X}_{HT}) = \sum_{i=1}^{N} \frac{X_i^2}{\pi_i^2}\pi_i(1-\pi_i) + \sum_{j\neq i}^{N} \frac{X_i}{\pi_i}\cdot\frac{X_j}{\pi_j}(\pi_{ij} - \pi_i\pi_j)$

Como $\hat{V}_1(\hat{X}_{HT})$ es un estimador insesgado de $V_1(\hat{X}_{HT})$

$\hat{V}_1(\hat{X}_{HT}) = \sum_{i=1}^{n} \frac{X_i^2}{\pi_i^2}(1-\pi_i) + \sum_{j\neq i}^{n} \frac{X_i}{\pi_i}\cdot\frac{X_j}{\pi_j}\cdot\frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}}$

y $\hat{X}_i$ es un estimador insesgado de X_i, podemos definir la copia del estimador como:

$\hat{V}_c(\hat{\hat{X}}_{HT}) = \sum_{i=1}^{n} \frac{\hat{X}_i^2}{\pi_i^2}(1-\pi_i) + \sum_{j\neq i}^{n} \frac{\hat{X}_i}{\pi_i}\frac{\hat{X}_j}{\pi_j}\frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}}$

Aplicando el tma II de Durbin:

$\hat{V}(\hat{\hat{\Theta}}) = \hat{V}_c(\hat{\Theta}) + \sum_{i=1}^{n} \hat{V}_2(\hat{Y}_i)\pi_i$ es un estimador insesgado de $V(\hat{\hat{\Theta}})$

TIID $\longrightarrow \hat{V}(\hat{\hat{X}}_{HT}) = \hat{V}_c(\hat{X}_{HT}) + \sum_{i=1}^{n} \frac{\hat{V}_2(\hat{X}_i)}{\pi_i}$ es insesgado de $V(\hat{\hat{X}}_{HT})$

donde $\hat{V}_2(\hat{X}_i)$ estimador insesgado de $V_2(\hat{X}_i)$.

En el caso de muestreo bietápico, con muestreo SIN repos. ~~en UP~~ y probab. desiguales en primera etapa y submuestreo aleatorio simple sin reposición en 2ª etapa.

Total poblacional: $X = \sum_{i=1}^{N} X_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} X_{ij}$ / $X_i = \sum_{j=1}^{M_i} X_{ij}$ total de congl. i

su estimador lineal insesgado es:

$$\hat{\hat{X}}_{HT} = \sum_{i=1}^{n} \frac{\hat{X}_i}{\pi_i} \quad \text{con } \hat{X}_i = M_i \bar{x}_i = \frac{M_i}{m_i} \sum_{j=1}^{m_i} X_{ij} \text{ estim. del total del congl.}$$

El tma I de Durbin da una expresión de la varianza del estimador para el total poblacional;

$$V(\hat{\hat{X}}_{HT}) = V_1(\hat{X}_{HT}) + \sum_{i=1}^{N} \frac{V_2(\hat{X}_i)}{\pi_i}$$

$$\left\{ \begin{array}{l} \hat{X}_{HT} = \sum_{i=1}^{n} \frac{X_i}{\pi_i} \Rightarrow V_1(\hat{X}_{HT}) = \sum_{i=1}^{N} \frac{X_i^2}{\pi_i^2} \pi_i (1-\pi_i) + \sum_{j \neq i}^{N} \frac{X_i}{\pi_i} \frac{X_j}{\pi_j} (\pi_{ij} - \pi_i \pi_j) \\ V_2(\hat{X}_i) = V_2(M_i \bar{x}_i) = M_i^2 V_2(\bar{x}_i) = M_i^2 \cdot (1-f_{2i}) \cdot \frac{S_i^2}{m_i} \end{array} \right.$$

donde $f_{2i} = \frac{m_i}{M_i}$ = fracción muestreo en 2ª etapa

$$S_i^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (X_{ij} - \bar{X}_i)^2$$

$$\text{TID} \rightarrow V(\hat{\hat{X}}_{HT}) = V_1(\hat{X}_{HT}) + \sum_{i=1}^{N} \frac{M_i^2}{\pi_i} (1-f_{2i}) \frac{S_i^2}{m_i}$$

expresión que coincide con la de la varianza del tma. Madow.

Para aplicar el TIID necesitamos el estimador insesgado de $V_1(\hat{X}_{HT})$ y su copia*, y el estimador insesgado de $V_2(\hat{X}_i)$

$$\hat{V}_1(\hat{X}_{HT}) = \sum_{i=1}^{n} \frac{X_i^2}{\pi_i^2} \frac{\pi_i(1-\pi_i)}{\pi_i} + \sum_{j \neq i}^{n} \frac{X_i}{\pi_i} \frac{X_j}{\pi_j} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}}$$

$$\hat{V}_c(\hat{X}_{HT}) = \sum_{i=1}^{n} \frac{\hat{X}_i^2}{\pi_i^2} \frac{\pi_i(1-\pi_i)}{\pi_i} + \sum_{j \neq i}^{n} \frac{\hat{X}_i}{\pi_i} \frac{\hat{X}_j}{\pi_j} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}}$$

$$\hat{V}_2(\hat{X}_i) = M_i^2 \cdot (1-f_{2i}) \cdot \frac{\hat{S}_i^2}{m_i}, \quad \text{con } \hat{S}_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (X_{ij} - \bar{x}_i)^2$$

$$\text{TIID} \rightarrow \hat{V}(\hat{\hat{X}}_{HT}) = \hat{V}_c(\hat{X}_{HT}) + \sum_{i=1}^{n} \frac{\hat{V}_2(\hat{X}_i)}{\pi_i} =$$

$$= \sum_{i=1}^{n} \frac{\hat{X}_i^2}{\pi_i^2} (1-\pi_i) + \sum_{j \neq i}^{n} \frac{\hat{X}_i}{\pi_i} \frac{\hat{X}_j}{\pi_j} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} + \sum_{i=1}^{n} \frac{M_i^2}{\pi_i} (1-f_{2i}) \frac{\hat{S}_i^2}{m_i}$$

es un estimador insesgado de $V(\hat{\hat{X}}_{HT})$.

4 – ESTIMACIÓN de la VARIANZA en el ~~ESTIMADOR~~ MUESTREO CON REPOSICIÓN y PROBAB. DESIGUALES

Como alternativa al TII D, existen otros métodos para estimar varianzas en un muestreo polietápico, cuando el muestreo de unidades primarias (conglomerados) se efectuó con reposición y probab. desiguales, y en cada unidad primaria se realizan submuestreos probabilísticos independientes sucesivos.

Si la muestra está constituida por al menos 2 unidades primarias ($n >> 2$), se puede emplear el mt. de los conglomerados últimos.

Hansen, Hurwitz y Madow idearon el concepto de "conglom. últimos" para considerar el muestreo polietápico como un caso especial del muestreo monoetápico de conglom.

Se denomina conglom. último al conjunto de unidades muestrales de última etapa que pertenecen a una unidad primaria, cualquiera que sea el nº de etapas efectuadas en ella.

La aplicación del mt. de conglomerados últimos es muy simple y conveniente cuando no se necesitan estimac. separadas de las contribuciones a la varianza debidas a las distintas etapas de submuestreo.

Si el muestreo es con reposición en primera etapa, se obtienen estimadores insesgados.

El método consiste en construir un estim. insesgado $\hat{\theta}_i$ del parámetro θ basado en cada conglomerado último y, a partir de los n estim. insesgados e independientes $\hat{\theta}_1 \dots \hat{\theta}_n$, construir un estimador insesgado $\hat{\hat{\theta}}$ de θ, basado en la muestra completa.

$\hat{\theta}_i$, $i = 1 \dots n$, estimador insesgado de θ

definimos

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

Por ser muestreo con reposic. en primera etapa:

$$V(\hat{\hat{\theta}}) = V\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i\right) = \frac{E(\hat{\theta}_i - \theta)^2}{n} = \frac{1}{n}\sum_{i=1}^{N}(\hat{\theta}_i - \theta)^2/N \; ?$$

$V(\bar{x}) = \frac{\sigma^2}{n}$

que tiene como estimador insesgado:

$$\hat{V}(\hat{\hat{\theta}}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{\theta}_i - \hat{\theta})^2$$

En particular, para estimar el total poblacional:

$$X = \sum_{i=1}^{N} X_i$$

se utiliza el estimador insesgado de Hansen y Hurwitz

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

(*) Mirar atrás

donde $\frac{\hat{X}_i}{P_i}$ es el estimador $\boxed{\text{insesgado}}$ de X que se obtiene con el congl. último i-ésimo.

Luego el estim. insesgado de la varianza es:

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

El mt. de los conglom. últimos tb. es aplicable a muestreo sin reposición, siempre que el nº de conglom. sea grande, añadiendo el factor de corrección $1-f$.

$$E\left[\frac{\hat{X}_i}{p_i}\right] = E_1 E_{>1}\left(\frac{\hat{X}_i}{p_i}\right) = E_1\left(\frac{X_i}{p_i}\right) = E_1\left[\sum^N \frac{X_i}{p_i} e_i\right] =$$

$$= \sum^N \frac{X_i}{p_i} E_1[e_i] = \sum^N X_i = X$$

$$V[\hat{X}_{HH}] = V\left[\frac{1}{n}\sum^n \frac{\hat{X}_i}{p_i}\right] = \frac{1}{n^2}\sum_{i=1}^{n} V\left(\frac{\hat{X}_i}{p_i}\right)$$