Diseño SY

Daniel Czarnievicz

2017

Estrategia de selección

Se fija un número a llamado *intervalo de muestreo*, de forma tal que $n = \left[{}^{N}\!/_{a} \right] \Rightarrow N = na + c$ con $0 \leq c < a$. Luego se sortea $r \sim Unifd(1; \ldots ; a)$ llamado *arranque aleatorio*. La muestra queda conformada por:

$$s = \big\{ k : k = r + (j-1)a \leq N; \;\; j = 1; \ldots ; n_S \big\}$$

El espacio muestral será: $\mathscr{S}_{SY} = \{ s_1; \ldots ; s_a \}$ donde $s_i \cap s_j = \emptyset \, \forall i \neq j$ y $\bigcup\limits_{i=1}^{a} s_i = U$.

El diseño muestral será entonces:

$$p(s) = \begin{cases} {}^{1}\!/_{a} & \text{si } s \in \mathscr{S}_{SY} \\ 0 & \text{en otro caso} \end{cases}$$

Dada la estrategia de selección, las muestras posibles serán:

Muestra	s_1	$\ldots$	s_r	$\ldots$	s_a
U	y_1	$\ldots$	y_k	$\ldots$	y_a
	y_{1+a}	$\ldots$	y_{r+a}	$\ldots$	y_{2a}
	$\vdots$		$\vdots$		$\vdots$
	$y_{1+(n-1)a}$	$\ldots$	$y_{r+(n-1)a}$	$\ldots$	y_{na}
Total	t_{s_1}	$\ldots$	t_{s_r}	$\ldots$	t_{s_a}
Media	$\bar{y}_{s_1}$	$\ldots$	$\bar{y}_{s_r}$	$\ldots$	$\bar{y}_{s_a}$

Tamaño muestral

El tamaño muestral será entonces:

$$n_S = \begin{cases} n+1 & \text{si } 0 < r \leq c \\ n & \text{si } c < r \leq a \end{cases}$$

Con probabilidades:

$$p(n_S) = \begin{cases} P(n_S = n+1) & = & P(r \leq c) & = & {}^{c}\!/_{a} \\ P(n_S = n) & = & P(r > c) & = & 1 - {}^{c}\!/_{a} \end{cases}$$

Control del tamaño muestral

- **Intervalo de muestreo fraccionario:**

  Sea $a = {}^{N}\!/_{n}$ con $n \in \mathbb{N}$ el tamaño deseado, y sea $\varepsilon \sim Unif(0; a)$ el arranque aleatorio. La muestra se forma con las etiquetas k que cumplen:

  $$s = \{ k : k - 1 < \varepsilon + (j-1)a \leq k; \;\; j = 1; \ldots ; n \}$$

  De igual forma, si $r = \varepsilon n \sim Unif(0; n)$, la muestra se forma por las etiquetas:

  $$s = \{ k : (k-1)n < r + (j-1)N \leq k\, n; \;\; j = 1; \ldots ; n \}$$

  Las probabilidades de inclusión serán: $\pi_k = P(k \in s) = n\,\dfrac{1}{N} = \dfrac{n}{N} = \dfrac{1}{a} \;\; \forall k \in U$

- **Muestreo sistemático circular**

  La población se ordena de forma circular, de forma que al elemento N le sigue el elemento 1. Se sortea $r \sim Unifd(1;N)$. Sea $a = \left[{}^{N}\!/_{n} \right]$. La muestra será:

$$S = \left\{ k : k = \left\{ \begin{array}{lll} r+(j-1)a & \text{si} & r+(j-1)a \leq N \\ r+(j-1)a-N & \text{si} & r+(j-1)a > N \end{array} \right. ; \; j = 1;\ldots;n \right\}$$

Probabilidades de inclusión

$$\star \; \pi_k = P(k \in s) = P(\text{“seleccionar la muestra } s_r\text{”}) = {}^{1}\!/_{a} \;\; \forall k \in U$$

$$\star \; \pi_{kl} = P(k;l \in s) = \left\{ \begin{array}{cc} {}^{1}\!/_{a} & \text{si } k;l \in s_r \in \mathscr{S}_{SY} \\ 0 & \text{en otro caso} \end{array} \right.$$

Por lo tanto, el diseño no es medible por lo que no existen estimadores insesgados de $V_{SY}(\hat{t}_\pi)$.

El estimador $\hat{t}_\pi$

$$\star \; t_y = \sum\nolimits_U y_k = \sum_{r=1}^{a} t_{s_r} = \sum_{r=1}^{a} \sum\nolimits_{s_r} y_k$$

$$\star \; \hat{t}_\pi = \sum\nolimits_s y_k^{\checkmark} = \sum\nolimits_s \frac{y_k}{{}^{1}\!/_{a}} = a \sum\nolimits_s y_k = a\, t_s = \frac{N}{n} t_s = N\, \bar{y}_s$$

Esto puede pensarse como si tuviéramos una población $U_t = \{t_{s_1};\ldots;t_{s_r};\ldots;t_{s_a}\}$ y se quiere estimar $t = \sum_{U_t} t_{s_r}$. Se toma una muestra bajo un diseño SI de tamaño $n=1$, así $\hat{t}_\pi = t_s^{\checkmark} = \dfrac{t_s}{{}^{1}\!/_{a}} = at_s$.

$$\star \; Rec(\hat{t}_\pi) = \{a\, t_{s_1};\ldots;a\, t_{s_a}\}$$

$$\star \; P(\hat{t}_\pi = a\, t_{s_r}) = {}^{1}\!/_{a} \;\; \forall r = 1;\ldots;a$$

$$= a \sum\nolimits_U E_{SY}(I_k) y_k = a \sum\nolimits_U \frac{1}{a} y_k = \sum\nolimits_U y_k = t_y$$

$$\star \; E_{SY}(\hat{t}_\pi) = E_{SY}(a\, t_S) = a\, E_{SY}(t_s) = a\, E_{SY}\left(\sum\nolimits_s y_k\right) = a\, E_{SY}\left(\sum\nolimits_U I_k y_k\right) =$$

$$= a \sum\nolimits_U E(I_k) y_k = a \sum\nolimits_U \frac{1}{a} y_k = \sum\nolimits_U y_k = t_y$$

$$\star \; E_{SY}(\hat{t}_\pi) = \sum_{r=1}^{a} \frac{\hat{t}_\pi}{a} = \sum_{r=1}^{a} a \frac{t_{s_r}}{a} = \sum_{r=1}^{a} t_{s_r} = t_y$$

$$\star \; V_{SY}(\hat{t}_\pi) = \sum\sum\nolimits_U \Delta_{kl}\, y_k^{\checkmark}\, y_l^{\checkmark} = \sum\sum\nolimits_U (\pi_{kl} - \pi_k\, \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} =$$

$$= \sum\sum\nolimits_U \left(\frac{\pi_{kl}}{\pi_k\, \pi_l} \right) y_k\, y_l - \sum\sum\nolimits_U y_k\, y_l =$$

Luego como $\pi_{kl} = 0$ si k y l no pertenecen a la misma muestra, $\sum\sum_U$ puede cambiarse por $\sum_{r=1}^{a} \left(\sum\sum_{s_r} \right)$.

$$= \sum_{r=1}^{a} \left(\sum\sum\nolimits_{s_r} a\, y_k\, y_l \right) - \Big(\underbrace{\sum\nolimits_U y_k}_{t_y} \Big)^2 = a \sum_{r=1}^{a} \Big(\underbrace{\sum\nolimits_{s_r} y_k}_{t_{s_r}} \Big)^2 =$$

$$= a\sum_{r=1}^{a} t_{s_r}^2 - (a\bar{t})^2 = a\left(\sum_{r=1}^{a} t_{s_r}^2 - a\,\bar{t}^2\right) = a\sum_{r=1}^{a}\left(t_{s_r} - \bar{t}\right)^2 = a\left(\frac{a-1}{a-1}\right)\sum_{r=1}^{a}\left(t_{s_r} - \bar{t}\right)^2 =$$

$$= a(a-1)\underbrace{\frac{1}{a-1}\sum_{r=1}^{a}\left(t_{s_r} - \bar{t}\right)^2}_{S_t^2} = a(a-1)S_t^2$$

Si $t_{s_1} = t_{s_2} = \ldots = t_{s_a} = {}^{t}\!/_{a} \Rightarrow V_{SY}(\hat{t}_\pi) = 0$. $V_{SY}(\hat{t}_\pi)$ depende de cómo se ordene la población.

$$\star\ V_{SY}(\hat{t}_\pi) = a\sum_{r=1}^{m}\left(t_{s_r} - \bar{t}\right)^2 = a\sum_{r=1}^{m}\left(n\bar{y}_{s_r} - \frac{t}{a}\right)^2 =$$

$$= a\sum_{r=1}^{m}\left(\frac{N}{a}\bar{y}_{s_r} - \frac{N}{a}\bar{y}_U\right)^2 = \frac{N^2}{a}\sum_{r=1}^{m}\left(\bar{y}_{s_r} - \bar{y}_U\right)^2 = N\,n\sum_{r=1}^{m}\left(\bar{y}_{s_r} - \bar{y}_U\right)^2 = N \times SSB$$

Como SST es fija, aumentar SSW implica disminuir SSB. Conviene que cada una de las muestras posibles sean muy heterogéneas, de forma de que SSW sea grande. Medimos la homogeneidad mediante:

$$\star\ \delta = 1 - \frac{N-1}{N-a}\,\frac{SSW}{SST} = 1 - \frac{S_{yW}^2}{S_{yU}^2} \text{ donde } S_{yW}^2 = \frac{SSW}{N-a} \text{ y } S_{yU}^2 = \frac{SST}{N-1}$$

- $\delta_{\text{máx}} = 1 \Leftrightarrow S_{yW}^2 = 0 \Rightarrow SSW = 0 \Rightarrow$ los grupos son lo más homogéneos posible $\Rightarrow V_{SY}(\hat{t}_\pi)$ es la mayor posible.
- $\delta_{\text{mín}} = -\dfrac{a-1}{N-a} \Leftrightarrow SSW = SST \Rightarrow SSB = 0 \Rightarrow$ los grupos son lo más heterogéneos posible $\Rightarrow V_{SY}(\hat{t}_\pi) = 0$.

Efecto diseño

Una forma alternativa de escribir la varianza del estimador π es:

$$\star\ V_{SY}(\hat{t}_\pi) = N \times SSB = N\big(SST - SSW\big) =$$

$$= N\left[(N-1)S_{yU}^2 - SSW\,\frac{SST}{SST}\,\frac{N-a}{N-a}\right] = N\left[(N-1)S_{yU}^2 - SSW\,\frac{(N-1)S_{yU}^2}{SST}\,\frac{N-a}{N-a}\right] =$$

$$= N\left[(N-1)S_{yU}^2 - (N-a)S_{yU}^2\underbrace{\frac{SSW}{SST}\,\frac{N-1}{N-a}}_{1-\delta}\right] = N\left[(N-1)S_{yU}^2 + (\delta-1)\underbrace{(N-a)}_{N-\frac{N}{n}}S_{yU}^2\right] =$$

$$= N\left[(N-1)S_{yU}^2 + (\delta-1)\underbrace{\left(N - \frac{N}{n}\right)}_{\frac{N}{n}(n-1)}S_{yU}^2\right] = N\left[(N-1)S_{yU}^2 + (\delta-1)\left(\frac{N}{n}(n-1)\right)S_{yU}^2\right] =$$

$$= N\left[(N-1)S_{yU}^2 + \delta\,\frac{N}{n}(n-1)S_{yU}^2 - \frac{N}{n}(n-1)S_{yU}^2\right] =$$

$$= N\left[\left(N - 1 - \frac{N}{n}(n-1)\right)S_{yU}^2 + \delta\,\frac{N}{n}(n-1)S_{yU}^2\right] =$$

$$= N\left[\left(N - 1 - N + \frac{N}{n}\right)S_{yU}^2 + \delta\,\frac{N}{n}(n-1)S_{yU}^2\right] =$$

$$= N\left[\frac{N}{n}\left(1-\frac{n}{N}\right)S^2_{y_U} + \delta\,\frac{N}{n}(n-1)S^2_{y_U}\right] =$$

$$= N\left[\frac{N}{n}(1-f)S^2_{y_U} + \delta\,\frac{N}{n}(n-1)S^2_{y_U}\right] =$$

$$= \frac{N^2}{n}S^2_{y_U}\Big[(1-f)+\delta(n-1)\Big] \text{ con } f=\frac{n}{N}=\frac{1}{a}$$

$$\star\ Deff(SY;\hat{t}_\pi) = \frac{\frac{N^2}{n}S^2_{y_U}\big[(1-f)+\delta(n-1)\big]}{\frac{N^2}{n}(1-f)S^2_{y_U}} = \frac{(1-f)+\delta(n-1)}{1-f} = 1+\frac{\delta(n-1)}{1-f}$$

$$\frac{n-1}{1-f} > 0 \Rightarrow \begin{cases} \text{si} \quad \delta > 0 & \Rightarrow SI \text{ es más eficietne que } SY \\ \text{si} \quad \delta = 0 & \Rightarrow SI \text{ y } SY \text{ son igualmente eficientes} \\ \text{si} \quad \delta < 0 & \Rightarrow SY \text{ es más eficietne que } SI \end{cases}$$

$\delta < 0 \Leftrightarrow S^2_{y_W} > S^2_{y_U}$, esto ocurre cuando los grupos son suficientemente heterogéneos.

Estimación de la varianza

Dado que no existe un estimador insesgado para $V_{SY}(\hat{t}_\pi$, se emplean las siguientes tácticas:

1. Si existe razón para creer que $V_{SY} \leq V_{SI}$, se utiliza la varianza del SI, por lo que:

$$\hat{V}_{SY}(\hat{t}_\pi) = \frac{N^2}{n}(1-f)S^2_{y_{s_r}} \text{ donde } S^2_{y_{s_r}} = \frac{1}{n-1}\sum\nolimits_{s_r}\left(y_k - \bar{y}_{s_r}\right)^2$$

2. Tomar más de un arranque aleatorio, $m > 1$, con intervalo de muestreo ma. Ahora cada SY contribuye una fracción ${}^{n}/_{m}$ de la muestra.

   Sean $r_1; \ldots; r_m$ los diferentes arranques y, por simplicidad, se supone que ${}^{n}/_{m}$ y a son enteros. La muestra será:

$$s = \left\{k : k = r_i + (j-1)ma;\ \ i = 1; \ldots; m;\ \ j = 1; \ldots; {}^{n}/_{m}\right\}$$

   Las probabilidades de inclusión serán:

$$\star\ \pi_k = \frac{m}{ma} = \frac{n}{N}\ \ \forall k \in U$$

$$\star\ \pi_{kl} = \begin{cases} \dfrac{n}{N} & \text{si} \quad k;l \text{ pertenecen a la misma muestra} \\[2ex] \dfrac{n}{N}\,\dfrac{m-1}{ma-1} & \text{si} \quad k;l \text{ pertenecen a distintas muestras} \end{cases}$$