Diseño PO

Daniel Czarnievicz

2017

Estrategia de selección

El diseño PO es un diseño que permite probabilidades de inclusión distintas para cada elemento de la población, y puede verse como una generalización del diseño BE. Dada una población $U = \{1; \ldots; k; \ldots; N\}$, consideremos $\pi_1; \ldots; \pi_k; \ldots; \pi_N$ valores predetermiados y no necesariamente iguales, tales que $0 < \pi_k \leq 1 \ \forall k \in U$. Luego, sean $\varepsilon_1; \ldots; \varepsilon_k; \ldots; \varepsilon_N$ iid $Unif(0;1)$. La muestra se conforma de la siguiente manera:

$$s = \{k : \varepsilon_k < \pi_k; \ \ k \in U\}$$

El diseño PO está dado por:

$$p(s) = \prod_{k \in S} \pi_k \prod_{k \in U-S} (1 - \pi_k) \ \forall s \in \mathscr{S}_{PO}$$

Probabilidades de inclusión

Como las indicadores son independientes e $I_k \sim Ber(\pi_k) \ \forall k \in U$, las probabilidades de inclusión que induce el diseño son:

$$\star \ P(k \in s) = \pi_k \ \forall k \in U$$

$$\star \ P(k; l \in s) = \pi_{kl} = \pi_k \, \pi_l \ \forall k \neq l \in U$$

Además:

$$\star \ \Delta_{kl} = \pi_{kl} - \pi_k \, \pi_l = \begin{cases} 0 & \text{si} \quad k \neq l \\ \pi_k(1 - \pi_l) & \text{si} \quad k = l \end{cases}$$

Elección de los π_k

Un criterio razonable sería elegir los π_k de forma tal que minimicen la variarianza del estimador $\hat{t}_\pi$, sujetos a un tamaño de muestra esperado fijo, n. Se debe resolver el siguiente problema de optimización:

$$\begin{gathered} \min_{\pi_k} \left\{ V_{PO}(\hat{t}_\pi) \right\} \\ \text{s.a.} \ \sum\nolimits_U \pi_k = n \end{gathered}$$

Como $V_{PO}(\hat{t}_\pi) = \sum\nolimits_U \left(\frac{1}{\pi_k} - 1 \right) y_k^2 = \sum\nolimits_U \frac{y-k^2}{\pi_k} - \sum\nolimits_U y_k^2$ y, además, $\sum\nolimits_U y_k^2$ y $\sum\nolimits_U \pi_k = n$ están dadas, el problema es equivalente a resolver:

$$\min_{\pi_k} \left\{ \left(\sum\nolimits_U \frac{y-k^2}{\pi_k} - \sum\nolimits_U y_k^2 \right) \left(\sum\nolimits_U \pi_k \right) \right\} \approx \min_{\pi_k} \left\{ \left(\sum\nolimits_U \frac{y-k^2}{\pi_k} \right) \left(\sum\nolimits_U \pi_k \right) \right\}$$

Por la desigualdad de Cauchy-Schwartz:[1]

$$\left(\sum\nolimits_U \frac{y-k^2}{\pi_k} \right) \left(\sum\nolimits_U \pi_k \right) = \left[\sum\nolimits_U \left(\frac{y_k}{\sqrt{\pi_k}} \right)^2 \right] \left[\sum\nolimits_U \left(\sqrt{\pi_k} \right)^2 \right] \geq \left[\sum\nolimits_U \left(\frac{y_k}{\sqrt{\pi_k}} \left(\sqrt{\pi_k} \right) \right) \right]^2 = \left(\sum\nolimits_U y_k \right)^2$$

[1] Desigualdad de Cauchy-Schwartz:

$\left(\sum\nolimits_U x_k \, z_k \right)^2 \leq \left(\sum\nolimits_U x_k \right) \left(\sum\nolimits_U z_k \right)$ y la igualdad se cumple $\Leftrightarrow x_k = \lambda \, z_k \ \forall k \in U, \ \lambda$ fijo.

La igualdad se cumple $\Leftrightarrow \dfrac{y_k}{\sqrt{\pi_k}} = \lambda\sqrt{\pi_k} \Rightarrow \pi_k = \dfrac{y_k}{\lambda}\ \ \forall k \in U$.

Ahora bien, como $\sum_U \pi_k = \sum_U \dfrac{y_k}{\lambda} = \dfrac{t_y}{\lambda} = n \Rightarrow \lambda = \dfrac{t_y}{n}$, por lo que, la mejor elección de los π_k, sujeto a que $0 < \pi_k \leq 1\ \forall k \in U$, viene dada por:

$$\star\ \pi_k = n\,\frac{y_k}{t_y} = \frac{n\,y_k}{\sum_U y_k}\ \ \forall k \in U$$

Lo anterior no es asequible dado que y_k no es conocido para toda la población. Por lo tanto, se debe trabajar con una variable auxiliar que cumpla que:

- x_k es conocida para toda la población.
- $x_k > 0$ para toda la población.
- $x_k \doteq c\,y_k$ para toda la población.

En este caso, $\pi_k = n\,\dfrac{x_k}{t_x} = \dfrac{n\,x_k}{\sum_U x_k}\ \ \forall k \in U$

No tener información auxiliar podría pensarse como una situación en la que $x_k = 1\ \forall k \in U$, con lo que se obtendría que $\pi_k = \dfrac{n}{N}$. Esto último es lo que justifica el uso de diseños con probabilidades de inclusión iguales para todos los elementos de la población, cuando no se dispone de información auxiliar.

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

$$\star\ \hat{t}_\pi = \sum_s y_k^{\checkmark} = \sum_s \frac{y_k}{\pi_k}$$

$$\star\ V_{PO}(\hat{t}_\pi) = \sum\sum_U \Delta_{kl}\, y_k^{\checkmark}\, y_l^{\checkmark} = \sum_U \pi_k(1-\pi_k)\, {y_k^{\checkmark}}^2 = \sum_U \left(\frac{1}{\pi_k} - 1\right) y_k^2$$

$$\star\ \hat{V}_{PO}(\hat{t}_\pi) = \sum_s (1-\pi_k)\frac{y_k^2}{\pi_k^2} = \sum_s \frac{1}{\pi_k}\left(\frac{1}{\pi_k} - 1\right) y_k^2$$

Si los π_k fueron elegidos de forma óptima, entonces tendremos que:

$$\star\ \hat{t}_\pi = \sum_s \frac{y_k}{\pi_k} = \sum_s \frac{y_k}{\frac{n\,y_k}{\sum_s y_k}} = \sum_s \left(\sum_U \frac{y_k}{n}\right) = \left(\sum_U \frac{y_k}{n}\right)\left(\sum_s 1\right) = \frac{n_S}{n}\, t_y$$

Con lo que la varianza de $\hat{t}_\pi$ dependerá únicamente de n_S.

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

$$\star\ \hat{t}_{alt} = N\,\frac{\sum_s y_k^{\checkmark}}{\sum_s \frac{1}{\pi_k}} = N\,\frac{\hat{t}_\pi}{\hat{N}}$$

Tamaño muestral

El tamaño de muestra en un diseño PO es aleatorio. Dado que $n_S = \sum_U I_k$:

$$\star\ E_{PO}(n_S) = \sum_U \pi_k \qquad\qquad \star\ V_{PO}(n_S) = \sum_U \pi_k(1-\pi_k)$$