Diseño ordenado SIR

Daniel Czarnievicz

2017

Estrategia de selección

Se realizan m extracciones con reposición de una población de tamaño N. En cada extracción, el k-ésimo elemento tiene probabilidad $p_k = {}^1/_N$ de ser seleccionado. Dado que se muestrea con reposición, los elementos pueden ser extraidos más de una vez. El mecanismo de extracción genera una *muestra ordenada*: $os = \{k_1; \dots; k_i; \dots; k_m\}$.

Se define la variable aleatoria r_k, la cual mide la cantidad de veces que el elemento k es sorteado. Por lo tanto, $r_k \sim Bin(m; {}^1/_N)$. Si N es lo suficientemente grande, esta probabilidad puede aproximarse mediante una distribución Poisson: $r_k \overset{a}{\sim} Poisson({}^m/_N)$. Dada la distribución Binomial de r_k,

$$P(\text{“extraer } r_k \text{ veces el elemento } k\text{”}) = \binom{m}{r} \left(\frac{1}{N}\right)^r \left(1 - \frac{1}{N}\right)^{m-r}$$

$$\star \; E(r_k) = \frac{m}{N} \qquad \star \; V(r_k) = \frac{m}{N}\left(1 - \frac{1}{N}\right) \doteq \frac{m}{N}$$

La probabilidad de que el elemento k nunca sea seleccionado en las m extracciones es:

$$P(\text{“no selecionar } k \text{ en ninguna de las } m \text{ extraciciones”}) = \left(1 - \frac{1}{N}\right)^m$$

Mientras que:

$$P(\text{“el elemento } k \text{ sea extraido”}) = P(r_k \geq 1) = 1 - P(r_k < 1) = 1 - P(r_k = 0) = 1 - \left(1 - \frac{1}{N}\right)^m$$

En cada una de las extracciones, la probabilidad de que el elemento k sea seleccionado es p_k, donde $\sum_U p_k = 1$. De esta forma $p(os) = \prod_{i=1}^{m} p_{k_i}$. En el diseño ordenado SIR, los p_k se eligen de forma que todos los elementos tengan la misma probabilidad de ser seleccionados, por lo tanto: $p_k = {}^1/_N$.

Dado que existen N^m muestras ordenadas de largo m, y las mismas son equiprobables, la probabilidad de seleccionar una muestra ordenada cualquiera es:

$$p(os) = \begin{cases} N^{-m} & \text{si la muestra ordenada es de tamanaño } m \\ 0 & \text{en otro caso} \end{cases}$$

Cada muestra ordenada, os, induce una muestra, s, tal que

$$s = \{k : k = k_i \text{ para algun } i = 1; \dots; m\}$$

Probabilidades de inclusión

Las probabilidades de inclusión de primer orden son:

$$\star \; \pi_k = P(k \in s) = P(r_k \geq 1) = 1 - \left(1 - \frac{1}{N}\right)^m \quad \forall k \in U$$

Las probabilidades de inclusión de segundo orden son:

$$\star \; \pi_{kl} = P(k; l \in s) = 1 - \left[2\left(1 - \frac{1}{N}\right)^m - \left(1 - \frac{1}{N}\right)^{2m}\right] \;\; \forall k \neq l \in U$$

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

$$\star \; \hat{t}_{pwr} = \frac{1}{m}\sum_{i=1}^{m}\frac{y_{k_i}}{p_{k_i}} = \frac{N}{m}\sum_{i=1}^{m} y_{k_i} = N\bar{y}_{os}$$

$$\star \; V_1 = \sum\nolimits_U \left(\frac{y_k}{p_k} - t_y\right)^2 p_k = \frac{1}{N}\sum\nolimits_U \left(Ny_k - N\bar{y}_U\right)^2 = N(N-1)S^2_{y_U}$$

$$\star \; V_{SIR}(\hat{t}_{pwr}) = \frac{V_1}{m} = \frac{N}{m}(N-1)S^2_{y_U}$$

$$\star \; \hat{V}_1 = \frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{y_{k_i}}{p_{k_i}} - \hat{t}_{pwr}\right)^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(Ny_{k_i} - N\bar{y}_{os}\right)^2 = N^2 S^2_{y_{os}}$$

$$\star \; \hat{V}_{SIR}(\hat{t}_{pwr}) = \frac{\hat{V}_1}{m} = \frac{N^2}{m}S^2_{y_{os}} \quad \text{donde} \quad S^2_{y_{os}} = \frac{1}{m-1}\sum_{i=1}^{m}\left(y_{k_i} - \bar{y}_{os}\right)^2$$

El estimador $\hat{\bar{y}}_U$

$$\star \; \hat{\bar{y}}_U = \frac{\hat{t}_{pwr}}{N}$$

$$\star \; V_{SIR}(\hat{\bar{y}}_U) = V_{SIR}\left(\frac{\hat{t}_{pwr}}{N}\right) = \frac{1}{N^2}V_{SIR}(\hat{t}_{pwr}) = \frac{1}{N^2}\frac{V_1}{m} = \frac{N-1}{Nm}S^2_{y_U}$$

$$\star \; \hat{V}_{SIR}(\hat{\bar{y}}_U) = \hat{V}_{SIR}\left(\frac{\hat{t}_{pwr}}{N}\right) = \frac{1}{N^2}\hat{V}_{SIR}(\hat{t}_{pwr}) = \frac{1}{N^2}\frac{\hat{V}_1}{m} = \frac{N^2 S^2_{y_{os}}}{N^2 m} = \frac{S^2_{y_{os}}}{m}$$

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

Para derivar el estimador π del total se debe trabajar con la muestra (s) inducida por el diseño ordenado. Dado que los elementos podrían repetirse, n_S es aleatoria. Para $m \geq 2$, $\hat{t}_{pwr}$ y $\hat{t}_\pi$ no tienen por qué coincidir.

$$\star \; \hat{t}_\pi = \sum\nolimits_s y_k^{\checkmark} = \sum\nolimits_s \frac{y_k}{\pi_k} = \sum\nolimits_s \frac{y_k}{1 - (1 - \frac{1}{N})^m}$$

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

Otro estimador insesgado pasado en el muestra s es el estimador $\hat{t}_{alt}$, donde:

$$\star \; \hat{t}_{alt} = N\,\bar{y}_s = \frac{N}{n_s}\sum\nolimits_s y_k$$

Si $E_{SIR}(n_S) = n = N\left[1 - \left(1 - \dfrac{1}{N}\right)^m\right]$, entonces $\hat{t}_{alt} = \dfrac{n}{n_s}\,\hat{t}_\pi$

Efecto diseño

Si $m = n$:

$$\star \ \frac{V_{SIR}(\hat{t}_{pwr})}{V_{SI}(\hat{t}_{\pi})} = \frac{\frac{N}{m}(N-1)S^2_{y_U}}{\frac{N^2}{n}(1-f)S^2_{y_U}} = \frac{N-1}{N}\frac{1}{1-f} \doteq \frac{1}{1-f}$$

Tamaño muestral

En caso que se quiera estimar t_y utilizando $\hat{t}_{pwr}$ bajo un diseño SIR con m extracciones:

$$\star \ \varepsilon^2 \doteq z^2_{1-\alpha/2} V_{SIR}(\hat{t}_y) = z^2_{1-\alpha/2} N(N-1)\frac{S^2_{y_U}}{m} \Rightarrow \boxed{m = \frac{1}{\varepsilon^2}\left[z^2_{1-\alpha/2} N(N-1)S^2_{y_U}\right]}$$