Muestreo por clusters

Daniel Czarnievicz

2017

En el muestreo por clusters el problema principal es que no se tiene ni se puede construir un marco muestral que permita realizar muestreo directo de elementos. La población $U = \{1; \dots; k; \dots; N\}$ se particiona en N_I subpoblaciones (clusters), $\{U_1; \dots; U_i; \dots; U_{N_I}\}$, donde $U = \bigcup\limits_{i \in U_I} U_i$. El set de clusters lo representamos como $U_I = \{1; \dots; i; \dots; N_I\}$.

Llamamos N_i a la cantidad de elementos poblacionales en el i-ésimo cluster. De esta forma,

$$N = \sum_{i \in U_I} N_i = \sum\nolimits_{U_I} N_i$$

Estrategia de selección

Una muestra s_I se selecciona de U_I de acuerdo con el diseño $p_I(.)$. Es decir, se seleccionan clusters. Luego todos los elementos dentro de los clusters seleccionados son relevados. Es decir, se censa los clusters. La muestra queda entonces conformada por $s = \bigcup\limits_{i \in s_I} U_i$. El tamaño de s_I se denota como: n_I si la muestra es de tamaño fijo, o n_{s_I} si la muestra es de tamaño aleatorio. Téngase presente que si N_i varía, entonces n_s variará. Dado que los clusters seleccionados son censados, $n_s = \sum\nolimits_{s_I} N_i$.

Probabilidades de inclusión

Probabilidades de seleccionar los distintos clusters:

$$\star \; \pi_{I_i} = P(\text{“selecicionar el cluster } i\text{”}) = P(i \in s_I) = \sum_{s_I \ni i} p_I(s_I)$$

$$\star \; \pi_{I_{ij}} = P(\text{“selecicionar los clusters } i \text{ y } j\text{”}) = \begin{cases} P(i; j \in s_I) = \displaystyle\sum_{s_I \ni i;j} p_I(s_I) & \text{si} \quad i \neq \\ \pi_{I_{ii}} = \pi_{I_i} & \text{si} \quad i = j \end{cases}$$

$$\star \; \Delta_{I_{ij}} = \pi_{I_{ij}} - \pi_{I_i} \pi_{I_j} \qquad \star \; \Delta^{\checkmark}_{I_{ij}} = \frac{\Delta_{I_{ij}}}{\pi_{I_{ij}}}$$

Probabilidades de inclusión de las unidades poblacionales:

$$\star \; \pi_k = P(k \in s) = P(i \in s_I) = \pi_{I_i}$$

$$\star \; \pi_{kl} = P(k; l \in s) = \begin{cases} P(i \in s_I) & = & \pi_{I_i} & \text{si} & k; l \in U_i \\ P(i; j \in s_I) & = & \pi_{I_{ij}} & \text{si} & k \in U_i; \; l \in U_j \end{cases}$$

$$\star \; \pi_{kk} = \pi_k$$

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

$$\star\ t_y = \sum\nolimits_U y_k = \sum\nolimits_{U_I} t_{y_i}$$

$$\star\ \hat{t}_\pi = \sum\nolimits_{s_I} t_{y_i}^{\checkmark} = \sum\nolimits_{s_I} \frac{t_{y_i}}{\pi_{I_i}}$$

$$\star\ E\left(\hat{t}_\pi\right) = E\left(\sum\nolimits_{s_I} t_{y_i}^{\checkmark}\right) = \sum\nolimits_{U_I} E(I_k)\,\frac{t_{y_i}}{\pi_{I_i}} = \sum\nolimits_{U_I} \pi_k\,\frac{t_{y_i}}{\pi_{I_i}} = \sum\nolimits_{U_I} \pi_{I_i}\,\frac{t_{y_i}}{\pi_{I_i}} = \sum\nolimits_{U_I} t_{y_i} = t_y$$

$$\star\ V_{p_I(s_I)}\left(\hat{t}_\pi\right) = \sum\sum\nolimits_{U_I} \Delta_{I_{ij}}\, t_{y_i}^{\checkmark}\, t_{y_j}^{\checkmark}$$

$$\star\ \hat{V}_{p_I(s_I)}\left(\hat{t}_\pi\right) = \sum\sum\nolimits_{s_I} \Delta_{I_{ij}}^{\checkmark}\, t_{y_i}^{\checkmark}\, t_{y_j}^{\checkmark}$$

Si $p_I(.)$ es de tamaño fijo, entonces se cumple que:

$$\star\ V_{p_I(s_I)}\left(\hat{t}_\pi\right) = -\frac{1}{2}\sum\sum\nolimits_{U_I} \Delta_{I_{ij}}\left(t_{y_i}^{\checkmark} - t_{y_j}^{\checkmark}\right)^2$$

$$\star\ V_{p_I(s_I)}\left(\hat{t}_\pi\right) = -\frac{1}{2}\sum\sum\nolimits_{s_I} \Delta_{I_{ij}}^{\checkmark}\left(t_{y_i}^{\checkmark} - t_{y_j}^{\checkmark}\right)^2$$

Si $t_{y_i}^{\checkmark} = \dfrac{t_{y_i}}{\pi_{I_i}}$ es constante para todos los i clusters, entones $V_{p_I(s_I)}(\hat{t}_\pi) = 0$. Por lo tanto, si podemos elegir $\pi_{I_i} \dot{\propto}\, t_{y_i}$ el muestreo por clusters será eficiente. Si lo N_i son conocidos, entonces $\pi_{I_i} \propto N_i$. Dado que $t_{y_i} = N_i\,\bar{y}_{U_I} = \sum\nolimits_{U_I} y_k$, esta será una buena elección si hay poca variación entre los $\bar{y}_{U_I}$. Si todos los $\bar{y}_{U_I}$ son iguales, entonces $V_{p_I(s_I)}(\hat{t}_\pi) = 0$.

Tomar π_{I_i} constante para todos los clusters es una elección pobre si los N_i varían mucho.