Muestreo en dos etapas

Daniel Czarnievicz

2017

En un muestreo en dos etapas la población $U = \{1; \dots; k; \dots; N\}$ es particionada en N_I grupos o conglomerados llamados PSU (primary sampling units), de forma tal que: $U_I = \{U_1; \dots; U_i; \dots; U_{N_I}\}$. Cada conglomerado U_i es de tamaño N_i. Luego entonces $N = \sum_{U_I} N_i$.

Mecanismo de selección

En una primera etapa se toma una muestra S_I de U_I según el diseño $p_I(.)$. El número de PSUs seleccionado lo anotamos como n_{S_I} si el diseño de primera etapa es de tamaño aleatorio, o n_I si el diseño de primera etapa es de tamaño fijo.

Luego, en una segunda etapa, para cada uno de los i conglomerados seleccionados en la primera etapa, se toma una muestra S_i de U_i según el diseño $p_i(.|s_I)$. El número de elementos en S_i se anota como n_{S_i} si el diseño de segunda etapa es de tamaño aleatorio, o n_i si el diseño de segunda etapa es de tamaño fijo. A los elementos seleccionados en la segunda etapa se los llama SSU (secondary sampling units).

Como resultado se obtiene una muestra s de U tal que $s = \bigcup\limits_{i \in S_I} s_i$. El número total de elementos seleccionados será $n_S = \sum\limits_{i \in S_I} n_i$.

Invarianza e Independencia

En lo que sigue se asume que se cumplen las siguientes propiedades:

1. **Invarianza**: $\forall i \in U_I$ y $\forall s_I \ni i$, se tiene que $p_i(.|s_I) = p_i(.)$. Esto implica que, sea cual sea la muestra de PSU en la primera etapa, si sale el conglomerado i, el diseño de muestreo de segunda etapa para dicho conglomerado será siempre $p_i(.)$.
2. **Independencia**: $p\left(\bigcup\limits_{i \in s_I} s_I | s_I \right) = p(s|s_I) = \prod\limits_{i \in s_I} p_i(.|s_I)$. Esto implica que el diseño de muestreo llevado a cabo en una PSU es independiente del llevado a cabo en cualquier otra PSU.

Asumiendo que se cumplen las propiedades de invarianza e independencia, y asumiendo que las SSU son elementos y no clusters, tenemos que:

$$p(s) = p\left(\bigcup_{i \in s_I} s_I | s_I \right) = p(s|s_I) = \prod_{i \in s_I} p_i(s_i|s_I) = \prod_{i \in s_I} p_i(.)$$

Probabilidades de inclusión

Probabilidades de inclusión para la primera etapa

$$\star \ \pi_{I_i} = P(\text{“el cluster } i \text{ fue seleccionado en la primera etapa”}) = P(i \in s_I) = P(U_i \in s_I)$$

$$\star \ \pi_{I_{ij}} = P(\text{“los clusters } i \text{ y } j \text{ fueron seleccionados en la primera etapa”}) =$$
$$= P(i; j \in s_I) = P(U_i; U_j \in s_I)$$

Luego entonces:

$$\star\ \Delta_{I_{ij}} = \left\{ \begin{array}{lll} \pi_{I_{ij}} - \pi_{I_i}\,\pi_{I_j} & \text{si} & i \neq j \\ \pi_{I_i}(1-\pi_{I_i}) & \text{si} & i = j \end{array} \right. \qquad \star\ \Delta^{\checkmark}_{I_{ij}} = \frac{\Delta_{I_{ij}}}{\pi_{I_{ij}}}$$

Probabilidades de inclusión para la segunda etapa

$$\star\ \pi_{k|i} = P(\text{“seleccionar el elemento } k\text{, dado que se seleccionó el cluster } i\text{”}) = P(k \in s | i \in s_I)$$

$$\star\ \pi_{kl|i} = P(\text{“seleccionar los elementos } k \text{ y } l\text{, dado que se seleccionó el cluster } i\text{”}) = P(k; l \in s | i \in s_I)$$

Luego entonces:

$$\star\ \Delta_{kl|i} = \left\{ \begin{array}{lll} \pi_{kl|i} - \pi_{k|i}\,\pi_{l|i} & \text{si} & k \neq l \\ \pi_{k|i}(1-\pi_{k|i}) & \text{si} & k = l \end{array} \right. \qquad \star\ \Delta^{\checkmark}_{kl|i} = \frac{\Delta_{kl|i}}{\pi_{kl|i}}$$

Probabilidades de inclusión de los elementos

De las propiedades de invarianza e independencia se desprende que:

$$P(k \in s) = P(i \in s_I \cap k \in s_i) = P(k \in s_i | i \in s_I)\, P(i \in s_I) = \pi_{I_i}\, \pi_{k|i}$$

Por lo tanto

$$\star\ \pi_k = \pi_{I_i}\, \pi_{k|i}\ \ \forall k \in U_i$$

Luego entonces

$$P(k; l \in s) = P(i; j \in s_I \cap k \in s_i \cap l \in s_j) =$$

$$= P(k \in s_i | i; j \in s_I)\, P(l \in s_j | i; j \in s_I)\, P(i; j \in s_I) = \pi_{I_{ij}}\, \pi_{k|i}\, \pi_{l|j}$$

Por lo tanto:

$$\star\ \pi_{kl} = \left\{ \begin{array}{lll} \pi_{I_i}\, \pi_{k|i} & \text{si} & k = l \in U_i \\ \pi_{I_i}\, \pi_{kl|i} & \text{si} & k; l \in U_i \\ \pi_{I_{ij}}\, \pi_{k|i}\, \pi_{l|j} & \text{si} & k \in U_i \text{ y } l \in U_j \end{array} \right.$$

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

$$\star\ \hat{t}_\pi = \sum\nolimits_s y^{\checkmark}_k = \sum\nolimits_{S_I} \sum\nolimits_{s_i} \frac{y_{k|i}}{\pi_k} = \sum\nolimits_{s_I} \sum\nolimits_{s_i} \frac{y_{k|i}}{\pi_{I_i}\, \pi_{k|i}} = \sum\nolimits_{s_I} \frac{1}{\pi_{I_i}} \sum\nolimits_{s_i} \frac{y_{k|i}}{\pi_{k|i}} =$$

$$= \sum\nolimits_{s_I} \frac{1}{\pi_{I_i}} \sum\nolimits_{s_i} y^{\checkmark}_{k|i} = \sum\nolimits_{s_I} \frac{\hat{t}_{\pi_i}}{\pi_{I_i}}$$

$$\text{donde } \hat{t}_{\pi_i} = \sum\nolimits_{s_i} y^{\checkmark}_{k|i} \text{ estima } t_{y_i} = \sum\nolimits_{U_i} y_{k|i}$$

$$\star\ E_{p_i(.|s_I)}(\hat{t}_{\pi_i}) = E_{p_i(.|s_I)} \left(\sum\nolimits_{s_i} y^{\checkmark}_{k|i} \right) = \sum\nolimits_{U_i} E_{p_i(.|s_I)}(I_{k|i})\, \frac{y_{k|i}}{\pi_{k|i}} = \sum\nolimits_{U_i} \pi_{k|i}\, \frac{y_{k|i}}{\pi_{k|i}} = \sum\nolimits_{U_i} y_{k|i} = t_{y_i}$$

$$\star\ V_i(\hat{t}_{\pi_i}) = \sum \sum\nolimits_{U_i} \Delta_{kl|i}\, y^{\checkmark}_{k|i}\, y^{\checkmark}_{l|i}$$

$$\star\ \hat{V}_i(\hat{t}_{\pi_i}) = \sum \sum\nolimits_{s_i} \Delta^{\checkmark}_{kl|i}\, y^{\checkmark}_{k|i}\, y^{\checkmark}_{l|i}$$

Si $p_i(.)$ es de tamaño fijo:

$$\star\ V_i(\hat{t}_{\pi_i}) = -\frac{1}{2} \sum \sum\nolimits_{U_i} \Delta_{kl|i} \left(y^{\checkmark}_{k|i} - y^{\checkmark}_{l|i} \right)^2$$

$$\star\ \hat{V}_i(\hat{t}_{\pi_i}) = -\frac{1}{2} \sum \sum\nolimits_{s_i} \Delta^{\checkmark}_{kl|i} \left(y^{\checkmark}_{k|i} - y^{\checkmark}_{l|i} \right)^2$$

$$\star\ V_{2ST}(\hat{t}_\pi) = V_{p_I(.)}\Big[\underbrace{E_{p_i(.|s_I)}(\hat{t}_\pi|s_I)}_{\sum_{s_I} t_{y_i}^{\checkmark}}\Big] + E_{p_I(.)}\Big[\underbrace{V_{p_i(.|s_I)}(\hat{t}_\pi|s_I)}_{\sum_{s_I}\frac{V_i}{\pi_{I_i}^2}}\Big] =$$

$$= \underbrace{V_{p_I(.)}\left(\sum\nolimits_{s_I} t_{y_i}^{\checkmark}\right)}_{\substack{\text{varianza del estimador}\\ \pi \text{ de } t_{y_i} \text{ dado } s_I}} + \underbrace{E_{p_I(.)}\left(\sum\nolimits_{s_I}\frac{V_i}{\pi_{I_i}^2}\right)}_{\substack{\text{valor esperado de la}\\ \text{varianza del estimador}\\ \pi \text{ de } t_{y_i} \text{ dado } s_I}} = \sum\sum\nolimits_{U_I}\Delta_{I_{ij}}\, t_{y_i}^{\checkmark}\, t_{y_j}^{\checkmark} + \frac{1}{\pi_{I_i}^2}\sum\nolimits_{U_I} E_{p_I(.)}(I_i)\, V_i =$$

$$= \sum\sum\nolimits_{U_I}\Delta_{I_{ij}}\, t_{y_i}^{\checkmark}\, t_{y_j}^{\checkmark} + \frac{1}{\pi_{I_i}^2}\sum\nolimits_{U_I}\pi_{I_i}\, V_i = \sum\sum\nolimits_{U_I}\Delta_{I_{ij}}\, t_{y_i}^{\checkmark}\, t_{y_j}^{\checkmark} + \sum\nolimits_{U_I}\frac{V_i}{\pi_{I_i}}$$

Debido a la propiedad de independencia:

$$E\left(\hat{t}_{\pi_i}\,\hat{t}_{\pi_j}\right) = \begin{cases} t_{y_i}^2 + V_i & \text{si} \quad i = j \\ t_{y_i}\, t_{y_j} & \text{si} \quad i \neq j \end{cases}$$

$$\star\ V_{PSU} = \sum\sum\nolimits_{U_I}\Delta_{I_{ij}}\,\hat{t}_{\pi_i}^{\checkmark}\,\hat{t}_{\pi_j}^{\checkmark}$$

$$\star\ \hat{V}_{PSU} = \sum\sum\nolimits_{U_I}\Delta_{I_{ij}}^{\checkmark}\,\hat{t}_{\pi_i}^{\checkmark}\,\hat{t}_{\pi_j}^{\checkmark} - \sum\nolimits_{s_I}\frac{1}{\pi_{I_i}}\left(\frac{1}{\pi_{I_i}} - 1\right)\hat{V}_i$$

$$\star\ E_{2ST}\left(\sum\sum\nolimits_{s_I}\Delta_{I_{ij}}^{\checkmark}\,\hat{t}_{\pi_i}^{\checkmark}\,\hat{t}_{\pi_j}^{\checkmark}\right) = E_{2ST}\left(\sum\sum\nolimits_{s_I}\Delta_{I_{ij}}^{\checkmark}\,\frac{\hat{t}_{\pi_i}\,\hat{t}_{\pi_j}}{\pi_{I_i}\,\pi_{I_j}}\right) =$$

$$= E_{p_I(.)}\left[E_{p_i(.|s_I)}\left(\sum\sum\nolimits_{s_I}\Delta_{I_{ij}}^{\checkmark}\,\frac{\hat{t}_{\pi_i}\,\hat{t}_{\pi_j}}{\pi_{I_i}\,\pi_{I_j}}\middle| s_I\right)\right] = E_{p_I(.)}\left[\sum\sum\nolimits_{s_I}\Delta_{I_{ij}}^{\checkmark}\,\frac{1}{\pi_{I_i}\,\pi_{I_j}} E_{p_i(.|s_I)}\left(\hat{t}_{\pi_i}\,\hat{t}_{\pi_j}|s_I\right)\right] =$$

$$= E_{p_I(.)}\left(\sum\sum\nolimits_{s_I}\Delta_{I_{ij}}^{\checkmark}\,\frac{t_{y_i}}{\pi_{I_i}}\,\frac{t_{y_j}}{\pi_{I_j}}\right) + E_{p_I(.)}\left(\sum\nolimits_{s_I}\Delta_{I_{ij}}^{\checkmark}\,\frac{V_i}{\pi_{I_i}^2}\right) =$$

$$= \sum\sum\nolimits_{s_I}\Delta_{I_{ij}}\,\frac{t_{y_i}}{\pi_{I_i}}\,\frac{t_{y_j}}{\pi_{I_j}} + \sum\nolimits_{s_I}\left(\frac{1}{\pi_{I_i}} - 1\right)V_i = V_{PSU} + \sum\nolimits_{s_I}\left(\frac{1}{\pi_{I_i}} - 1\right)V_i$$

$$\star\ E_{2ST}\left[-\sum\nolimits_{s_I}\frac{1}{\pi_{I_i}}\left(\frac{1}{\pi_{I_i}} - 1\right)\hat{V}_i\right] = -E_{p_I(.)}\left[E_{p_i(.|s_I)}\left(\sum\nolimits_{s_I}\frac{1}{\pi_{I_i}}\left(\frac{1}{\pi_{I_i}} - 1\right)\hat{V}_i\middle| s_I\right)\right] =$$

$$= -E_{p_I(.)}\left[\sum\nolimits_{s_I}\frac{1}{\pi_{I_i}}\left(\frac{1}{\pi_{I_i}} - 1\right)E_{p_i(.|s_I)}\left(\hat{V}_i\middle| s_I\right)\right] = -E_{p_I(.)}\left[\sum\nolimits_{s_I}\frac{1}{\pi_{I_i}}\left(\frac{1}{\pi_{I_i}} - 1\right)V_i\right] =$$

$$= -\sum\nolimits_{U_I}\left(\frac{1}{\pi_{I_i}} - 1\right)V_i$$

Por lo tanto:

$$E(\hat{V}_{PSU}) = V_{PSU} + \sum\nolimits_{s_I}\left(\frac{1}{\pi_{I_i}} - 1\right)V_i - \sum\nolimits_{U_I}\left(\frac{1}{\pi_{I_i}} - 1\right)V_i = V_{PSU}$$

$$\star\ E_{2ST}(\hat{V}_{SSU}) = E_{2ST}\left(\sum\nolimits_{s_I}\frac{\hat{V}_i}{\pi_{I_i}^2}\right) = E_{p_I(.)}\left[E_{p_i(.|s_I)}\left(\sum\nolimits_{s_I}\frac{\hat{V}_i}{\pi_{I_i}^2}\middle| s_I\right)\right] =$$

$$= E_{p_I(.)}\left[\sum\nolimits_{s_I}\frac{1}{\pi_{I_i}^2}E_{p_i(.|s_I)}\left(\hat{V}_i\middle| s_I\right)\right] =_{p_I(.)}\left[\sum\nolimits_{s_I}\frac{1}{\pi_{I_i}^2}V_i\right] = \sum\nolimits_{U_I}\frac{V_i}{\pi_{I_i}} = V_{SSU}$$

En conclusión:

$$E_{2ST}\big(\hat{V}_{2ST}(\hat{t}_\pi)\big) = E_{2ST}\big(\hat{V}_{PSU} + \hat{V}_{SSU}\big) = E_{2ST}\big(\hat{V}_{PSU}\big) + E_{2ST}\big(\hat{V}_{SSU}\big) = V_{PSU} + V_{SSU} = V_{2ST}(\hat{t}_\pi)$$

Un estimador computacionalmente más sencillo viene dado por:

$$\star \; \hat{V}^* = \sum\sum\nolimits_{s_I} \Delta^{\checkmark}_{I_{ij}} \frac{\hat{t}_{\pi_i}}{\pi_{I_i}} \frac{\hat{t}_{\pi_j}}{\pi_{I_j}}$$

$$\star \; E_{2ST}(\hat{V}^*) = V_{PSU} + \sum\nolimits_{U_I} \left(\frac{1}{\pi_{I_i}} - 1\right) V_i = V_{PSU} + \underbrace{\sum\nolimits_{U_I} \frac{V_i}{\pi_{I_i}}}_{V_{SSU}} - \sum\nolimits_{U_I} V_i =$$

$$= \underbrace{V_{PSU} + V_{SSU}}_{V_{2ST}(\hat{t}_\pi)} - \sum\nolimits_{U_I} V_i = V_{2ST}(\hat{t}_\pi) - \sum\nolimits_{U_I} V_i$$

Por lo tanto, el sesgo de $\hat{V}^* = -\sum_{U_I} V_i$, por lo que su sesgo relativo está dado por:

$$\star \; \frac{B(\hat{V}^*)}{V_{2ST}(\hat{t}_\pi)} = -\frac{-\sum_{U_I} V_i}{\sum\sum_{U_I} \Delta_{I_{ij}}\, t^{\checkmark}_{y_i}\, t^{\checkmark}_{y_j} + \sum_{U_I} \frac{V_i}{\pi_{I_i}}}$$