Muestreo con probabilidad proporcional al tamaño

Daniel Czarnievicz

2017

En los casos en que y_k es proporcional a una variable auxiliar x_k, la varianza puede reducirse si se utiliza dicha información.

Muestreos sin reposición de tamaño fijo y el estimador $\hat{t}_\pi$

Considérese el estimador $\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k}$ para el total de la variable y. Supongamos que es posible construir un mecanismo de selección de elementos sin reposición y que produzca una muestra de tamaño fijo n. Supongamos también que la variable y es tal que ${}^{y_k}/_{\pi_k} = c \ \forall k \in U$. En estos casos entonces, $\hat{t}_\pi = n \, c \ \forall s \in \mathscr{S}$. Esto implica que la varianza del estimador π será cero.

Claramente este diseño no puede construirse dado que requeriría conocer todos los valores de y_k de antemano. Pero si se cuenta con una variable auxiliar x_k que sea aproximadamente proporcional a la variable y_k, entonces podía utilizarse para construir el diseño, y esto reduciría significativamente la varianza del estimador π. Esto se lograría tomando:

$$\pi_k = n \, \frac{x_k}{\sum_U x_k} \ \ \forall k \in U$$

Si la variable x presenta mucha variabilidad, podría ocurrir que, para algún(os) k, $\pi_k \geq 1$. Esto claramente no es aceptable dado que π_k es una probabilidad. En estos casos, π_k es fijado en 1 para dichos k, y el rsto se construyen de forma proporcional:

$$\pi_k = (n - n_A) \, \frac{x_k}{\sum_{U-A} x_k} \ \ \forall k \in U - A$$

donde A es el conjunto de elementos para los cuales $\pi_k = 1$.

Muestreos con reposición y el estimador $\hat{t}_{pwr}$

Considérese el estimador $\hat{t}_{pwr} = \frac{1}{m} \sum_{i=1}^{m} \frac{y_k}{p_k}$. De nuevo, la idea es encontrar un diseño en el cual ${}^{y_k}/_{p_k} = c \forall k \in U$. Si esto fuera posible, entonces $\hat{t}_{pwr} = c \ \forall s \in \mathscr{S}$, y el estimador tendría varianza cero. Al igual que antes, este tipo de diseño no puede implementarse directamente ya que no se conocen de antemano los valores y_k. La solución es trabajar con una variable auxiliar x_k, de forma tal que $p_k \propto x_k$. Por lo tanto, los p_k se elijen de forma tal que:

$$p_k = \frac{x_k}{\sum_U x_k} \ \ \forall k \in U$$

El estimador $\hat{t}_{pwr}$ del total será entonces:

$$\star \ \hat{t}_{pwr} = \frac{1}{m} \sum_{i=1}^{m} \frac{y_k}{p_k} = \frac{1}{m} \sum_{i=1}^{m} \frac{y_k}{\frac{x_k}{\sum_U x_k}} = \frac{1}{m} \left(\sum_U x_k \right) \left(\sum_{i=1}^{m} \frac{y_k}{x_k} \right)$$

$$\star \ \hat{V} \hat{t}_{pwr} = \left(\sum_U x_k \right)^2 \frac{1}{m(m-1)} \left[\sum_{i=1}^{m} \left(\frac{y_k}{x_k} \right)^2 - \frac{1}{m} \left(\sum_{i=1}^{m} \frac{y_k}{x_k} \right)^2 \right]$$

Podrían combinarse los beneficios de utilizar el estimador π (el cual es más eficiente) y el estimador pwr (el cual tiene una varianza más sencilla de calcular) de la siguiente forma:

1. Se utiliza un diseño πps de tamaño fijo m en el cual $\pi_k = m\, p_k = \dfrac{m\, x_k}{\sum_U x_k}$.
2. Se utiliza el estimador π para calcular el total t_y.
3. Se utiliza la fórmula del muestreo pps para calcular el estimador de la varianza:

$$v = \frac{1}{m(m-1)} \sum\nolimits_s \left(\frac{y_k}{p_k} - \frac{1}{m} \sum\nolimits_s \frac{y_k}{p_k} \right)^2$$

4. El sesgo de la varianza viene dado por: $B(v) = E(v) - V(\hat{t}_\pi) = \dfrac{m}{m-1} \Big[V(\hat{t}_{pwr}) - V(\hat{t}_\pi) \Big]$

Diseños πps

En un diseño πps las probabilidades de inclusión deben ser tales que $\pi_k \propto x_k \ \forall k \in U$, siendo $x_1; \ldots ; x_N$ números positivos y conocidos de antemano.

Tamaño de muestra: $n = 1$

Se utiliza el método conocido como “cumulative total method”. Este se implementa de la siguiente forma:

1. Sean $T = 0$, $T_k = T_{k-1} + x_k \ \forall k \in U$.
2. Se sortea $\varepsilon \sim Unif(0; 1)$. Si $T_{k-1} < \varepsilon\, T_N \leq T_k$, entonces el elemento k es seleccionado.

El diseño es πps dado que: $\pi_k = P(T_{k-1} < \varepsilon\, T_N \leq T_k) = 0 \dfrac{T_k - T_{k-1}}{T_N} = \dfrac{x_k}{\sum_U x_k}$

Nótese que $\pi_{kl} = 0 \ \forall k \neq l$, por lo que no existen estimadores insesgados de la varianza.