Muestreo Estadístico: π-Estimador *

García Prado, Sergio
sergio@garciparedes.me

26 de septiembre de 2017

1. Demostración de insesgadez

Para una población población finita U formada por los individuos $\{1, ..., i, ..., N\}$ cuya variable de interés se denota por y, siendo y_k el valor que toma dicha variable en el individuo k. Se define el total poblacional como sigue:

$$t = \sum_{k \in U} y_k \tag{1}$$

Sea s una muestra extraida de la población U (por tanto $s \subset U$), se define la variable aleatoria I_k como se indica a continuación:

$$I_k = \begin{cases} 1 & \text{si } k \in s \\ 0 & \text{si } k \notin s \end{cases} \tag{2}$$

De esta manera, se puede utilizar el valor π_k para modelizar la probabilidad de que I_k tome el valor 1, es decir, de que el elemento k-ésimo esté en la muestra s. Por tanto, se puede afirmar que $I_k \sim B(\pi_k)$, una distribución de *Bernoulli* de parámetro π_k. A partir de dicha distribución se tiene que $E[I_K] = \pi_k$. Además, $pi_k > 0 \forall k \in U$ para que se cumpla la restricción de muestreo probabilístico.

Se define el π-estimador como:

$$\widehat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in U} I_k \frac{y_k}{\pi_k} \tag{3}$$

La demostración acerca de la insesgadez de este estimador deriva de la esperanza de la variable I_k, que toma el valor π_k, tal y como se indicaba anteriormente, lo cual produce que se anule dicho valor pi_k con el pi_k del divisor del estimador. Además, la restricción de muestreo probabilístico $pi_k > 0 \forall k \in U$ hace que el valor $\widehat{t}_\pi$ exista siempre (no hay división entre 0). Por estas razones se puede decir que el estimador es insesgado. De manera matemática esto se describe a continuación:

$$E[t] = \sum_{k \in U} y_k \tag{4}$$

$$E[\widehat{t}_\pi] = \sum_{k \in U} E[I_k] \frac{y_k}{\pi_k} = \sum_{k \in U} \pi_k \frac{y_k}{\pi_k} = \sum_{k \in U} y_k \tag{5}$$

$$bias(t, \widehat{t}_\pi) = \left| E[t] - E[\widehat{t}_\pi] \right| = \left| \sum_{k \in U} y_k - \sum_{k \in U} y_k \right| = 0 \tag{6}$$

*URL: https://github.com/garciparedes/statistical-sampling-pi-estimator

Nótese por tanto, que el funcionamiento del π-estimador se basa en la "expansión" del conjunto de individuos seleccionados en la muestra s, de tal manera que se siga cumpliendo de la forma más precisa posible la distribución de probabilidades de la población U.

2. Demostración del intervalo de confianza para la varianza

En este caso se pretende probar el intervalo de confianza de nivel $1-\alpha$ para la distribución de probabilidades referida a la variable aleatoria t, es decir, a la suma del total poblacional para la variable Y. Es decir, se pretende probar la igualdad descrita en la ecuación (7). Dicha demostración se llevará a cabo siguiendo la notación de valores de probabilidad.

$$\left[\widehat{t}_\pi \pm z_{1-\alpha/2}\sqrt{\widehat{Var}[\widehat{t}_\pi]}\right] \equiv \Pr\left(t \in \left[\widehat{t}_\pi \pm z_{1-\alpha/2}\sqrt{Var[\widehat{t}_\pi]}\right]\right) = 1-\alpha \tag{7}$$

Un paso previo para la demostración del intervalo de confianza es la demostración de la ecuación (8). Por el *Teorema Central del Límite* se puede asumir que la distribución t sigue una distribución normal con media $\widehat{t}$ y varianza $Var[\widehat{t}_\pi]$. Por tanto, este puede ser tipificado a una distribución normal estándar. Esta propiedad se indica en la ecuación (8)

$$\frac{t-\widehat{t}_\pi}{\sqrt{Var[\widehat{t}_\pi]}} \sim N(0,1) \tag{8}$$

Por tanto, para demostrar que el intervalo de confianza sigue un nivel $1-\alpha$ basta con aplicar distintas propiedades algebraicas sobre este, hasta llegar a la distribución normal estándar tal y como se indicó en la ecuación (8), que debido a su simetría permite sustituir el intervalo multiplicando por 2 el valor de α. Por último, debido a las propiedades de la distribución normal estándar, queda demostrado que en hasta el percentil $z_{1-\alpha}$ se localiza la probabilidad $1-\alpha$.

$$\begin{aligned}
\Pr\left(t \in \left[\widehat{t}_\pi \pm z_{1-\alpha/2}\sqrt{Var[\widehat{t}_\pi]}\right]\right) &= && (9)\\
&= \Pr\left(t-\widehat{t}_\pi \in \left[\pm z_{1-\alpha/2}\sqrt{Var[\widehat{t}_\pi]}\right]\right) && (10)\\
&= \Pr\left(\frac{t-\widehat{t}_\pi}{\sqrt{Var[\widehat{t}_\pi]}} \in \left[\pm z_{1-\alpha/2}\right]\right) && (11)\\
&= \Pr\left(N(0,1) \in [\pm z_{1-\alpha/2}]\right) && (12)\\
&= \Pr\left(N(0,1) \leq z_{1-\alpha}]\right) && (13)\\
&= 1-\alpha && (14)
\end{aligned}$$

Referencias

[SSW03] Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model assisted survey sampling.* Springer Science & Business Media, 2003.

[TG18] Jesús Alberto Tapia García. Muestreo Estadístico 1, 2017/18.