

Muestreo Estratificado

Ejercicio 2^{*}

García Prado, Sergio
sergio@garciparedes.me

12 de noviembre de 2017

Resumen

En este trabajo se desarrollan las expresiones del tamaño de muestra n global necesario para poder asegurar que el error de estimación se aproxima a B con un nivel de confianza k para los estimadores del total poblacional \hat{T} y la proporción poblacional \hat{P} .

1. Introducción

Denotaremos por $U = U_1 \cup \dots \cup U_h \cup \dots \cup U_L = \{1, \dots, k, \dots, N\}$ a la población, para la cual tenemos una división en L estratos denotando por U_h al estrato $h \in \{1, \dots, L\}$. Sea I_h el conjunto de índices de las observaciones seleccionadas en el estrato U_h y s_h la muestra extraída de dicho estrato. Por tanto, podemos denotar a la muestra global por $s = s_1 \cup \dots \cup s_h \cup \dots \cup s_L$.

En este caso, tal y como se ha indicado anteriormente, se va a presuponer la utilización de *muestreo aleatorio simple (m.a.s.)* sobre cada estrato. Este método de muestreo se caracteriza por la fijación *a-priori* del tamaño de muestras y la selección de observaciones *sin reemplazamiento*. Esto es equivalente a decir que una vez seleccionada una observación esta desaparece del conjunto de candidatas a aparecer en la muestra. Por tanto, no hay observaciones repetidas en la muestra.

Denotaremos por \mathcal{T} al total poblacional de una determinada variable de interés Y denotando como $y_k \quad \forall k \in U$ al k -ésimo valor de Y . Es fácil entender por tanto que el total poblacional se define como $\mathcal{T} = \sum_U y_k$.

Denotaremos por P a la proporción poblacional de una determinada variable de interés Y de carácter binario denotando como $y_k \quad \forall k \in U$ al k -ésimo valor de Y . Es fácil entender por tanto que la proporción poblacional se define como $P = \frac{\sum_U y_k}{N}$.

Bajo la hipótesis de *muestreo aleatorio simple (m.a.s.)*, a partir del cual se pretende obtener una aproximación lo más precisa posible tanto del total poblacional \mathcal{T} como de la proporción poblacional P . Para ello, es necesario apoyarse en los valores del tamaño de la población N , el del estrato U_h como N_h y el de la muestra s_h denotado por n_h . El tamaño relativo del estrato se define como $W_h = \frac{N_h}{N}$. También se define el tamaño relativo de la muestra como $f_h = \frac{n_h}{N_h}$. Entonces, en este caso un buen estimador del total poblacional es el π -estimador $\hat{T} = \sum_{s_h} \frac{y_k}{W_h}$. Para el caso del estimador de la proporción poblacional, un buen estimador es $\hat{P} = \frac{\sum_{s_h} \frac{y_k}{W_h}}{N}$.

En las ecuaciones (1) y (3) se definen respectivamente las varianzas del estimador del total poblacional \hat{T} así como la proporción poblacional \hat{P} . Estas se han desarrollado de tal manera que n_h quede lo menos relacionada posible con el resto de variables, lo cual será útil para las demostraciones de las secciones 2 y 3.

^{*}URL: <https://github.com/garciparedes/statistical-sampling-stratified>

$$\begin{aligned}
Var(\hat{T}) &= \\
&= \sum_{h=1}^L \frac{N_h^2(1-f_h)\sigma_h^{*2}}{n_h} \\
&= \sum_{h=1}^L \frac{N_h^2(1-\frac{n_h}{N_h})\sigma_h^{*2}}{n_h} \\
&= \sum_{h=1}^L \frac{(N_h^2 - N_h n_h)\sigma_h^{*2}}{n_h} \\
&= \sum_{h=1}^L \left(\frac{N_h^2}{n_h} - N_h \right) \sigma_h^{*2}
\end{aligned} \tag{1}$$

$$\begin{aligned}
Var(\hat{P}) &= \\
&= \sum_{h=1}^L W_h^2 \frac{1-f_h}{n_h} \frac{N_h}{N_h-1} P_h(1-P_h) \\
&= \sum_{h=1}^L W_h^2 \frac{1-\frac{n_h}{N_h}}{n_h} \frac{N_h}{N_h-1} P_h(1-P_h) \\
&= \sum_{h=1}^L \frac{W_h^2 - \frac{W_h^2 n_h}{N_h}}{n_h} \frac{N_h}{N_h-1} P_h(1-P_h) \\
&= \sum_{h=1}^L \frac{W_h^2 - \frac{n_h * N_h}{N^2}}{n_h} \frac{N_h}{N_h-1} P_h(1-P_h) \\
&= \sum_{h=1}^L \left(\frac{W_h^2}{n_h} - \frac{N_h}{N^2} \right) \frac{N_h}{N_h-1} P_h(1-P_h)
\end{aligned} \tag{3}$$

Tal y como se ha indicado anteriormente, n_h se refiere al tamaño de la muestra h -ésima. Entoces, este puede definirse como el tamaño de la muestra global ponderado por un determinado peso w_h dependiente de la estrategia de afijación escogida. Esto se puede definir matemáticamente como:

$$n_h = n * w_h \tag{4}$$

Donde w_h se define tal y como se indica en las ecuaciones (5) y (6) para los casos de *afijación proporcional* y *mínima varianza* respectivamente.

$$w_h = W_h \tag{afijación proporcional} \tag{5}$$

$$w_h = \frac{N_h \sigma_h^*}{\sum_{i=1}^L N_i \sigma_i^*} \tag{afijación mínima varianza} \tag{6}$$

Para tamaños de estrato N_h grandes es fácil comprobar que se cumple la propiedad $\frac{N_h}{N_h-1} \simeq 1$, lo cual permite simplificar la ecuación del tamaño de la muestra.

Puesto que lo que se pretende demostrar en este trabajo es la ecuación para estimar el tamaño n de la muestra global, fijando un error de estimación B a un nivel de confianza k , esto es equivalente a despejar el valor n en la ecuación (7).

$$\frac{B^2}{k^2} = Var(\hat{\theta}) \tag{7}$$

Una vez desarrollada la ecuación que a partir de la cual determinar el tamaño de muestra global n , fijado un erro de estimación b a una confianza de nivel k , el siguiente paso es particularizar esto para estimadores concretos. Esto consiste simplemente en la substitución del valor de la varianza por el del estimador en cuestión. En la sección 2 se realiza dicha demostración para el caso del total poblacional \widehat{T}_h y en la sección 3 se realiza para el caso del estimador de proporción poblacional \widehat{P}_h .

2. Demostración para estimador del total poblacional \widehat{T}

Para la demostración basta con desarrollar la función (7) utilizando la varianza del estimador del total poblacional \widehat{T} definida en la ecuación (1) para posteriormente dejar la fórmula en función de n .

$$\frac{B^2}{k^2} = Var(\widehat{T}) \quad (8)$$

$$\frac{B^2}{k^2} = \sum_{h=1}^L \left(\frac{N_h^2}{n_h} - N_h \right) \sigma_h^{*2} \quad (9)$$

$$\frac{B^2}{k^2} = \sum_{h=1}^L \left(\frac{N_h^2}{n * w_h} - N_h \right) \sigma_h^{*2} \quad (10)$$

$$\frac{B^2}{k^2} + \sum_{h=1}^L N_h \sigma_h^{*2} = \sum_{h=1}^L \frac{N_h^2}{n * w_h} \sigma_h^{*2} \quad (11)$$

$$n = \frac{\sum_{h=1}^L \frac{N_h^2}{w_h} \sigma_h^{*2}}{\frac{B^2}{k^2} + \sum_{h=1}^L N_h \sigma_h^{*2}} \quad (12)$$

$$(13)$$

3. Demostración para estimador de la proporción poblacional \widehat{P}

Al igual que para la demostración anterior, en este caso también basta con desarrollar la función (7), pero en este caso utilizando la varianza del estimador de la proporción poblacional \widehat{P} definida en la ecuación (3) para posteriormente dejar la fórmula en función de n .

$$\frac{B^2}{k^2} = Var(\widehat{P}) \quad (14)$$

$$\frac{B^2}{k^2} = \sum_{h=1}^L \frac{W_h^2}{n_h} - \frac{N_h}{N^2} \frac{N_h}{N_h - 1} P_h (1 - P_h) \quad (15)$$

$$\frac{B^2}{k^2} = \sum_{h=1}^L \frac{W_h^2}{n * w_h} - \frac{N_h}{N^2} \frac{N_h}{N_h - 1} P_h (1 - P_h) \quad (16)$$

$$\frac{B^2}{k^2} + \sum_{h=1}^L \frac{N_h}{N^2} \frac{N_h}{N_h - 1} P_h (1 - P_h) = \sum_{h=1}^L \frac{W_h^2}{n * w_h} \frac{N_h}{N_h - 1} P_h (1 - P_h) \quad (17)$$

$$n = \frac{\sum_{h=1}^L \frac{W_h^2}{w_h} \frac{N_h}{N_h - 1} P_h (1 - P_h)}{\frac{B^2}{k^2} + \sum_{h=1}^L \frac{N_h}{N^2} \frac{N_h}{N_h - 1} P_h (1 - P_h)} \quad (18)$$

$$n = \frac{\sum_{h=1}^L \frac{W_h^2}{w_h} \frac{N_h}{N_h - 1} P_h (1 - P_h)}{\frac{B^2}{k^2} + \sum_{h=1}^L \frac{W_h}{N} \frac{N_h}{N_h - 1} P_h (1 - P_h)} \quad (19)$$

$$(20)$$

Cabe destacar que tal y como se ha indicado al principio del trabajo, en ambos casos se puede simplificar el operando $\frac{N_h}{N_h-1}$ cuando N_h toma valores suficientemente grandes. Además, el valor w_h se debe fijar según las ecuaciones (5) y (6) según corresponda.

Referencias

- [1] SÄRNDAL, C.-E., SWENSSON, B., AND WRETMAN, J. *Model assisted survey sampling*. Springer Science & Business Media, 2003.
- [2] TAPIA GARCÍA, J. A. Muestreo Estadístico 1, 2017/18. Facultad de Ciencias: Departamento de Estadística.