

# Muestreo Estratificado

## Ejercicio 1<sup>\*</sup>

García Prado, Sergio  
sergio@garciparedes.me

2 de noviembre de 2017

### 1. Introducción

En este trabajo se realizarán distintas demostraciones relacionadas con la estimación del estadístico *total poblacional* denotado por  $\mathcal{T}$  sobre una metodología muestral basada en *muestreo estratificado* en el que en cada estrato se extrae una *muestra aleatoria simple (m.a.s.)*.

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 del método *m.a.s.* sobre cada estrato, caracterizado porque el tamaño de la muestra se fija *a-priori* y se seleccionan las observaciones *sin reemplazamiento*, es decir, una vez seleccionada una observación, esta desaparece del conjunto de candidatos 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  $\mathcal{T} = \sum_U y_k$ .

En el caso del *muestreo aleatorio simple (m.a.s.)*, a partir del cual se pretende obtener una aproximación lo más precisa posible del total poblacional  $\mathcal{T}$ , 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  $\pi$ -estimador  $\hat{\mathcal{T}} = \sum_{s_h} \frac{y_k}{W_h}$ .

La varianza del estimador del total sobre cada estrato se define a continuación:

$$Var(\hat{\mathcal{T}}_h) = \frac{N_h^2 \sigma_h^2}{n_h} \quad (1)$$

En las siguientes secciones se realiza la demostración acerca de los tamaños óptimos para la muestra de cada estrato suponiendo conocido el tamaño  $n$  de muestra global (sección 2) y fijado un presupuesto  $C$  (sección 3)

### 2. Tamaño de muestra en cada estrato, conocido el tamaño $n$ de muestra global

El problema se presenta como la obtención del tamaño óptimo  $n_h \quad h \in \{1, \dots, L\}$  que minimice la varianza global (**afijación de mínima varianza**) de la estimación, suponiendo como valor conocido  $N$  para la obtención del mejor estimador de un determinado estadístico. Definiremos la función  $\phi(n_1, \dots, n_L)$  como:

---

<sup>\*</sup>URL: <https://github.com/garciparedes/statistical-sampling-stratified>

$$\phi(n_1, \dots, n_L) = \text{Var}(\hat{\theta}) + \lambda \left( \sum_{h=1}^L n_h - n \right) \quad (2)$$

Entonces el objetivo es minimizar dicha función, de tal manera que se minimiza la varianza global de la estimación  $\hat{\theta}$ . Esto es equivalente a buscar los valores que hagan mínima dicha función, es decir:

$$\min\{\phi(n_1, \dots, n_L)\} \quad (3)$$

Para obtener dicho mínimo se pueden utilizar distintas técnicas, sin embargo, en este caso basta con la búsqueda del punto que hace nulo el valor de la derivada, tal y como se verá a continuación.

$$\phi(n_1, \dots, n_L) = \text{Var}(\hat{\theta}) + \lambda \left( \sum_{h=1}^L n_h - n \right) \quad (4)$$

El primer paso es derivar la función  $\phi(n_1, \dots, n_L)$  respecto del valor que se pretende minimizar:

$$\begin{aligned} \frac{\partial \phi(n_1, \dots, n_L)}{\partial n_h} &= \\ &= \frac{\partial \left( \frac{N_h^2 \sigma_h^2}{n_h} + \lambda \left( \sum_{h=1}^L n_h - n \right) \right)}{\partial n_h} \\ &= \frac{-N_h^2 \sigma_h^2}{n_h^2} + \lambda \\ &= \left( \frac{-N_h \sigma_h}{n_h} \right)^2 + \lambda \\ &= 0 \end{aligned} \quad (5)$$

Puesto que  $n_1 + \dots + n_L = n$  se obtiene que:

$$\sqrt{\lambda} = \frac{\sum_{h=1}^L N_h \sigma_h}{n} \quad (6)$$

Por último, se despeja el valor  $n_h$ , es decir, el tamaño de cada estrato:

$$\begin{aligned} n_h &= \\ &= \frac{N_h \sigma_h}{\sqrt{\lambda}} \\ &= \frac{N_h \sigma_h}{\frac{\sum_{h=1}^L N_h \sigma_h}{n}} \\ &= \frac{n N_h \sigma_h}{\sum_{h=1}^L N_h \sigma_h} \end{aligned} \quad (7)$$

### 3. Tamaño de muestra en cada estrato, fijado un presupuesto C

También se puede considerarear el mismo problema apoyandonos en una función coste en lugar del valor del tamaño poblacional  $n$ . Para ello, se define el coste como  $C = C_0 + \sum_{h=1}^L C_h n_h$  de tal manera que se presupone un coste constante  $C_0$  y un coste para cada estrato  $C_h$ . Suponemos el valor  $C$  como conocido y lo denominaremos presupuesto. Entonces ahora la función a minimizar se transforma en:

$$\phi(n_1, \dots, n_L) = \text{Var}(\hat{\theta}) + \lambda \left( C_0 \sum_{h=1}^L C_h n_h - C \right) \quad (8)$$

El primer paso es derivar la función  $\phi(n_1, \dots, n_L)$  respecto del valor que se pretende minimizar:

$$\frac{\partial \phi(n_1, \dots, n_L)}{\partial n_h} = 0 \quad (9)$$

[TODO ]

$$\sqrt{\lambda} = \quad (10)$$

Por último, se despeja el valor  $n_h$ , es decir, el tamaño de cada estrato:

$$n_h = \quad (11)$$

## 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.