1. Ejemplo de estimación en MAS con R

Muestreo Aleatorio Simple y...

Objetivos:

- Puesta en práctica de un muestreo aleatorio simple (m.a.s.): selección de las unidades, tratamiento computacional,...
- Cálculo de estimadores, errores de muestreo e intervalos de confianza.

De la población formada por 270 manzanas de edificios [?], seleccionar una muestra de tamaño 20 mediante:

- *Un muestreo aleatorio simple*

A partir de dichas muestras calcular, en cada caso:

1. *Una estimación del número total de viviendas y del número total de viviendas alquiladas*
2. *Una estimación del error de muestreo de los estimadores anteriores.*
3. *Una estimación de la proporción de viviendas alquiladas.*

La población (marco):

Población1. Manzanas de Edificios

Unidades elementales de muestreo: *Manzanas de edificios*
Tamaño de la población: $N = 270$
Variables en estudio:

- $\boldsymbol{x}$ *Número de viviendas*
- $\boldsymbol{y}$ *Número de viviendas alquiladas*

Parámetros a estimar:

- *Número total de viviendas (X) y su error de muestreo*
- *Número total de viviendas alquiladas (Y) y su error de muestreo*
- *Proporción de viviendas alquiladas* (P_Y)

Esquema muestral

En este caso se verifica que la probabilidad de selección de cada unidad en cada extracción es $p_i = \frac{1}{270-t}, \quad t = 0, 1, ..., n-1; \quad i = 1, 2, ..., N$. *Por otro lado, cada una de las unidades de la población tiene una probabilidad* $\pi_i = n/N = 20/270, \quad i = 1, 2, ..., N$ *de pertenecer a la muestra final de observaciones.*

En primer lugar tenemos que leer los datos de la población

```
manzanas = read.table("manzanas.txt", header = TRUE)
```

Vamos a visualizar las primeras líneas del conjunto de datos para asegurarnos de que la lectura se ha efectuado correctamente.

```
head (manzanas)
  manzana   x   y
1       1 149 131
2       2  10   6
3       3  30  23
4       4  90  79
5       5  56  47
6       6  42  34
```

1. *Una estimación del número total de viviendas y del número total de viviendas alquiladas*

Lo primero que debemos hacer es definir los tamaños de la población y de la muestra

```
N = 270
n = 20
```

A continuación, vamos a utilizar la función `S.SI`*, contenida en el paquete* ***TeachingSampling****, para extraer la muestra aleatoria simple. Esta función tiene dos parámetros:* `S.SI(N, n)`*, donde*

- `N` *indica el tamaño de la población*
- `n` *indica el tamaño de la muestra*

De manera que, podemos utilizar esta función en nuestro ejemplo mediante la siguiente llamada

```
library(TeachingSampling)
ind = S.SI(N, n)
```

Nota: *Para poder utilizar un paquete por primera vez tenemos que instalarlo*

y cargarlo. Una vez instalado, cada vez que iniciemos una sesión de R, para poder utilizar ese paquete sólo hace falta cargarlo.

Como resultado, la función `S.SI` *devuelve una matriz con tantas filas como elementos tenga la población y una única columna. Cada elemento de la matriz indica si la unidad ha sido seleccionada en la muestra o no. Así, a las unidades no incluidas en la muestra les corresponde el valor* 0*, mientras que a aquellas otras que sí han sido incluidas en la muestra se le asigna la posición que ocupan dentro de la población. Por tanto,* `ind` *incluye un total de* $n = 20$ *valores distintos de* 0*.*

El siguiente paso consiste en extraer de la población de manzanas la información relativa a las 20 que han sido seleccionadas en la muestra.

```
s = manzanas[ind,]
head(s)
   manzana  x  y
13      13 33 25
42      42 24 13
64      64 68 52
67      67 48 46
72      72 37 27
77      77 81 60
```

Ya tenemos la información muestral de las dos variables principales, de manera que es posible calcular las estimaciones para el total de viviendas y de viviendas alquiladas mediante el estimador de Horvitz-Thompson:

$$\hat{X} = \sum_{i \in s} \frac{x_i}{\pi_i} \qquad\qquad \hat{Y} = \sum_{i \in s} \frac{y_i}{\pi_i}$$

El paquete ***TeachingSampling*** *incluye la función* `E.SI` *que calcula el estimador de Horvitz-Thompson. Esta función tiene los siguientes parámetros:*

- `N` *el tamaño de la población*
- `n` *el tamaño de la muestra*
- `y` *información de la(s) variable(s) de interés.*

Como se indica, el parámetro `y` *puede contener información sobre una o varias variables. De manera que este parámetro puede ser un vector, una matriz o incluso un data frame. Teniendo esto en cuenta, calculemos las estimaciones que nos piden:*

```
var_s = data.frame(s$x, s$y)
result = E.SI(N, n, var_s)
```

```
result
                  N        s.x        s.y
Estimation      270 7816.50000 5130.00000
Standard Error    0 1581.41270 1155.14866
CVE               0   20.23172   22.51752
DEFF            NaN    1.00000    1.00000
```

La función devuelve las estimaciones del total de cada una de las variables y de los errores de muestreo y los coeficientes de variación correspondientes (en porcentaje). Basándonos en la muestra que hemos seleccionado, podemos concluir que, aproximadamente, hay un total de 7817 viviendas en la población, de las cuales 5130 se encuentran alquiladas.

2. *Una estimación del error de muestreo de los estimadores anteriores.*

El error de muestreo estimado para el estimador de un total viene recogido dentro de la salida de la función `E.SI` *que hemos obtenido en el apartado anterior, de manera que podemos afirmar que los errores muestrales estimados son 1581.4127 para la estimación del total de viviendas y 1155.14866 para la estimación del total de viviendas alquiladas.*

3. *Una estimación de la proporción de viviendas alquiladas*

Para estimar la proporción de viviendas alquiladas dentro de la población a partir de los datos de la muestra que hemos seleccionado, basta con dividir el número total estimado de viviendas alquiladas entre el número total de viviendas, es decir, la estimación del total de la variable Y entre la estimación del total de la variable X.

```
py = result[1,3] / result[1,2]
py
[1] 0.656304
```

Se puede concluir que, aproximadamente, un 65.63 % de las viviendas de la población se encuentran alquiladas.