Tema 5. Modelado y Predicción Estadística. Diseño muestral para una población finita

Beatriz Coronado Sanz

12 de febrero de 2019

1. Datos del trabajo

Como datos tenemos un listado poblacional de las titulaciones que se imparten en la Universidad de Sevilla (anuario estadístico 2017-18).Los datos se encuentran en el fichero *Titulaciones.xlsx*. Mostramos a continuación la cabecera de los datos:

```
# A tibble: 6 x 6
  Rama         Centro                  Titulacion              Hombre Mujer Total
  <chr>        <chr>                   <chr>                    <dbl> <dbl> <dbl>
1 CIENCIAS DE~ F. DE MEDICINA          Ldo. en Medicina             1     1     2
2 CIENCIAS DE~ F. DE MEDICINA          Grado en Biomedicina B~     46   139   185
3 CIENCIAS DE~ F. DE ENFERMERÍA, FIS~ Grado en Enfermería        169   725   894
4 CIENCIAS DE~ F. DE FARMACIA          Grado en Farmacia         539  1254  1793
5 CIENCIAS DE~ F. DE ENFERMERÍA, FIS~ Grado en Fisioterapia      191   192   383
6 CIENCIAS DE~ F. DE MEDICINA          Grado en Medicina         706  1229  1935
```

Para cada elemento se recoge la rama de enseñanza (Rama), la facultad o escuela donde se imparte (Centro), el nombre de la tituación (Titulacion) y el número de alumnos matriculados por sexo y total (Hombre, Mujer y Total).

2. Diseño polietápico propuesto

Vamos a realizar un diseño polietápico estratificado en dos etapas. Los estratos serán las distintas ramas de enseñanza.

En la primera etapa, para cada rama vamos a seleccionar una facultad o escuela según un MAS.

En la segunda etapa, para cada facultad realizaremos un diseño IIPS según el número de alumnos para elegir un determinado número de titulaciones.

El número de titulaciones de cada facultad será proporcional al número de titulaciones de cada rama. En otras palabras, usaremos afijación proporcional.

Para realizar este diseño, lo primero que haremos será ordenar las titulaciones por rama. Seleccionamos $n = 10$ como tamaño muestral (número de titulaciones que queremos obtener al final en

nuestro estudio) y calculamos el número de titulaciones que debemos tener por rama en el resultado final.

Vemos a continuación cuantas titulaciones hay por rama y cuantas tenemos que elegir en el resultado final:

```
                                  N_estrat n_estrat
[1,] "ARTES Y HUMANIDADES"        "17"     "2"
[2,] "CIENCIAS"                   "10"     "1"
[3,] "CIENCIAS DE LA SALUD"       "12"     "1"
[4,] "CIENCIAS SOCIALES Y JURÍDICAS" "33"  "3"
[5,] "INGENIERÍA Y ARQUITECTURA"  "33"     "3"
```

Lo siguiente que haremos será calcular las propiedades de inclusión para cada titulación:

```
> #Propiedades de inclusion
> pikk=10*as.vector(tapply(datos$Total,datos$Titulacion,sum))/sum(datos$Total)
> head(pikk)

[1] 0.023774324 0.118072482 0.036360730 0.048347784 0.064330523 0.007791585
```

Aplicamos el diseño muestral y obtenemos la siguiente muestra:

```
STAGE 1
Number of strata: 5
STAGE  2
Number of selected clusters: 1
Number of units in the population and number of selected units: 17 9
Number of selected clusters: 1
Number of units in the population and number of selected units: 10 2
Number of selected clusters: 1
Number of units in the population and number of selected units: 12 1
Number of selected clusters: 1
Number of units in the population and number of selected units: 33 7
Number of selected clusters: 1
Number of units in the population and number of selected units: 33 3
STAGE  3

Population total and number of selected units: 9 2

Population total and number of selected units: 2 1

Population total and number of selected units: 1 1

Population total and number of selected units: 7 3

Population total and number of selected units: 3 3
```

Observamos que a veces el número de titulaciones escogidas es menor que el tamaño muestral $n = 10$ que hemos seleccionado. Esto es porque, al realizar un MAS sobre las facultades, se escoge alguna facultad que no cumple el mínimo de titulaciones para el tamaño de su rama.

Las titulaciones escogidas son:

```
 [1] "Grado en Filología Hispánica"
 [2] "Grado en Estudios Franceses"
 [3] "Grado en Biología"
 [4] "Grado en Psicología"
 [5] "Grado en Pedagogía"
 [6] "Grado en Educación Infantil"
 [7] "Grado en Educación Primaria"
 [8] "Arquitecto"
 [9] "Grado en Fundamentos de Arquitectura"
[10] "Grado en Arquitectura"
```

3. Estudio de una variable de interés

La variable de interés que hemos escogido es el número de alumos repetidores en cada titulación.

Para generar esta variable elegimos un número aleatorio entre 0 y el máximo número de alumnos de cada titulación de la muestra. De esta forma escogemos un número de alumnos repetidores para cada titulación.

```
> #Generar variable aleatoria Y
> AlumRep=round(runif(length(xx$Titulacion), min=0, max=xx$Total),digits=0)
```

A continuación mostramos para cada titulación el número de alumnos repetidores, el número total de alumnos y la proporción de alumnos repetidores por titulación:

```
      Titulacion                            AlumRep Total  Proporcion
 [1,] "Grado en Filología Hispánica"        "201"   "551"  "36.48"
 [2,] "Grado en Estudios Franceses"         "239"   "285"  "83.86"
 [3,] "Grado en Biología"                   "340"   "1104" "30.8"
 [4,] "Grado en Psicología"                 "853"   "1226" "69.58"
 [5,] "Grado en Pedagogía"                  "324"   "745"  "43.49"
 [6,] "Grado en Educación Infantil"         "283"   "756"  "37.43"
 [7,] "Grado en Educación Primaria"         "836"   "2456" "34.04"
 [8,] "Arquitecto"                          "34"    "119"  "28.57"
 [9,] "Grado en Fundamentos de Arquitectura" "127"   "1554" "8.17"
[10,] "Grado en Arquitectura"               "25"    "325"  "7.69"
```

Guardamos la muestra obtenida de las titulaciones escogidas y de la variable que hemos creado en *muestra.xlsx*.

3.1. Estimaciones e intervalos de confianza para la variable de interés

Vamos a obtener los estimadores e intervalos de confianza para el total y la media poblacional de alumnos repetidos por titulación.

Para ello tenemos que usar la función *svydesign* para definir el diseño muestral que queremos aplicar:

```
> dis.MC2= svydesign(ids=~Rama+1, probs=xx$Prob , data=xx)
```

Para estimar la variable total usamos la función *svytotal*. El resultado obtenido es:

```
        total    SE
AlumRep 38458 16713
```

El primer valor es la estimación obtenida para el total y el segundo la desviación típica. El intervalo de confianza asociado es:

```
           2.5 %  97.5 %
AlumRep 5700.602 71214.5
```

Para estimar la variable media usamos la función *svymean*. El resultado obtenido es:

```
          mean     SE
AlumRep 315.85 96.193
```

Y su intervalo de confianza asociado:

```
           2.5 %   97.5 %
AlumRep 127.3143 504.3823
```