

7. MUESTREO

El muestreo estadístico es un procedimiento empleado para obtener una o más muestras de una población, a fin de realizar estimaciones a algunos de sus parámetros. Una vez establecido un marco muestral representativo de la población, se realiza el muestreo, es decir, se procede a la selección de los elementos de la muestra. Existen diferentes técnicas para obtener una muestra.

Para realizar algunos procesos relacionados con muestreo en R es necesario cargar los paquetes PASWR, e1071, class, MASS. Las siguientes rutinas se realizan luego de haber cargado estos paquetes.

7.1 TAMAÑO DE MUESTRA

En estadística el tamaño de la muestra es el número de sujetos extraídos de una población que la componen, necesarios para que los datos obtenidos sean representativos de la población. Bajo ciertas condiciones, en R es posible determinar el tamaño muestral; el comando que se presenta a continuación sirve para determinarlo:

```
nsize(b, sigma = NULL, p = 0.5, conf.level= , type="pi")
```

donde **b** es el error máximo permisible; **sigma** se refiere a la desviación estándar poblacional (de ser conocida); **p** es la estimación de la proporción poblacional, que no se requiere en caso de que se use el tipo **mu**; **conf.level** se refiere al nivel de confianza deseado, y **type** indica el parámetro (media "mu" o proporción "pi") bajo el cual se estima el tamaño de la muestra.

```
> # Error máximo permitido 5%
> b = 0.05
>
> # Confiabilidad del 94%
> conf.level = 0.94
>
> # Estimación de la proporción
> p = 0.048
>
> # Tamaño de muestra
> nsize(b, p, conf.level,type="pi")
```

```
The required sample size (n) to estimate the population
proportion of successes with a 0.95 confidence interval
so that the margin of error is no more than 0.05 is 87 .
```

Imagen 120. Salida R Tamaño de muestra

7.2 MUESTREO ALEATORIO SIMPLE

Uno de los objetivos del muestreo estadístico es precisamente tomar muestras aleatorias de una población; el siguiente comando es útil para seleccionar muestras mediante la técnica denominada Muestreo Aleatorio Simple (MAS), una vez definido el tamaño de muestra:

```
sample(x, size, replace = FALSE, prob = NULL)
```

donde *x* es el vector del cual se seleccionan los elementos, que puede ser de tipo numérico o puede contener los nombres de los caracteres, en este último caso cada uno de los elementos debe ir entre comillas; *size* es un entero positivo que indica el número de elementos por seleccionar del vector *x*; *replace* es la indicación lógica para realizar un muestreo con reemplazamiento (TRUE) o sin reemplazamiento (FALSE), y *prob* es un vector con las probabilidades correspondientes a cada uno de los elementos del vector *x*.

Ejemplo: Se tienen 100 empresas del sector industrial; de estas se necesita tomar una muestra aleatoria de tamaño 17, con el fin de realizar un estudio de factibilidad por parte de una proveedora de insumos; tener en cuenta que una empresa no puede aparecer en más de una oportunidad y que todas las empresas tienen la misma probabilidad de salir en la muestra; considere que las 100 empresas están numeradas de 1 a 100.

```
>
> # Numeración de las empresas
> x = c ( 1:100)
>
> sample (x, size=17, replace=FALSE, prob=NULL )
[1,] 34  2 65 37 49  4 56 16 64 99 80 33 85 19 28 72 74
>
```

Imagen 121. Salida R MAS

Ejemplo: Se desea seleccionar, de los 6 ingenieros que tiene la empresa, 2 para cargos directivos; se cree que los ingenieros Pedro y Luis tienen una probabilidad de ser seleccionados de 0.3, mientras que los demás tienen una probabilidad de 0.1; determinar aleatoriamente cuáles ingenieros son seleccionados.

```
>
> # Ingenieros
> x=c("Pedro","Luis","Edwin","Juan","Diego","Ronal")
>
> # Selección de ingenieros
> sample (x, size=2, prob=c(0.3,0.3,0.1,0.1,0.1,0.1 ) )
[1,] "Luis"    "Pedro"
>
```

Imagen 122. Salida R MAS

En el ejercicio anterior, si se quisiera observar todas las posibles muestras, se utilizaría el comando `SRS(x, n)`, donde *n* es el tamaño de la muestra; esto se muestra a continuación:

```

>
> # Ingenieros
> x=c("Pedro","Luis","Edwin","Diego","Ronal")
>
> # Posibles muestras
> SRS(x,2)
      [,1] [,2]
[1,] "Pedro" "Luis"
[2,] "Pedro" "Edwin"
[3,] "Luis" "Edwin"
[4,] "Pedro" "Diego"
[5,] "Luis" "Diego"
[6,] "Edwin" "Diego"
[7,] "Pedro" "Ronal"
[8,] "Luis" "Ronal"
[9,] "Edwin" "Ronal"
[10,] "Diego" "Ronal"
>

```

Imagen 123. Salida R Posibles muestras

Para utilizar algunas otras técnicas de muestreo mediante comandos específicos es necesario cargar los paquetes **sampling**, **Ipsolve** y **MASS**.

7.3 MUESTREO SISTEMÁTICO

Se utiliza cuando el universo o población es de gran tamaño o se extiende en el tiempo. Primero hay que identificar las unidades y relacionarlas con el calendario (cuando proceda) o en lista, numeradas en orden. El comando que permite extraer una muestra mediante este método es **UPsystematic(pik,eps=1e-6)**, donde **pik** es el vector de probabilidades de inclusión; **pik = c(rep(n/N,N))**, donde **n** es el tamaño de la muestra, **N** es el tamaño de la población, y **eps** es un valor de control que por defecto es $1e-6$.

Ejemplo: En el archivo de la biblioteca se encuentran 150 tesis de grado numeradas en orden, de las cuales se desea revisar si cuentan con artículo; para ello se extrae una muestra de tamaño 15 mediante el método sistemático. Determinar cuáles son las tesis seleccionadas.

```

>
> # Tamaño de la población
> N = 150
>
> # Tamaño de la muestra
> n = 15
>
> # Probabilidad de inclusión
> pik = c(rep(n/N, N))
>
> # Selección de la muestra
> s=UPsystematic(pik)
>
> # Elementos seleccionados
> (1:length(pik))[s==1]
[1] 8 18 28 38 48 58 68 78 88 98 108 118 128 138 148
>

```

Imagen 124. Salida R Muestreo sistemático

El comando `UPsystematic()` genera un vector con 1 y 0; 1 significa que el objeto es seleccionado, y 0, que no fue seleccionado; el comando `(1:length(pik))[s==1]` permite ver los elementos seleccionados dentro del vector que fue creado (los unos del vector).

7.4 MUESTREO ESTRATIFICADO

Consiste en la división previa de la población de estudio en grupos o clases que se suponen homogéneos respecto a la característica que se va a estudiar. Para poder realizar un muestreo estratificado en R es necesario que los estratos sean del mismo tamaño. El comando que permite tomar muestras mediante esta técnica es: `balancedstratification(X, strata, pik)`, donde `X` es el vector con los elementos que conforman la población, `X = cbind(c(elementos))`; `strata` es el vector que especifica los estratos, y `pik` es el vector de probabilidades de inclusión; `pik = rep(n/m, N)`, donde `n` es el tamaño de la muestra por estrato, `m` es el tamaño del estrato y `N` es el tamaño de la población.

Ejemplo: Se tienen 15 elementos clasificados en 3 estratos diferentes; determinar una muestra aleatoria de tamaño 6.

```
>
> # Vector de estratificación
> strata=c(1,1,1,1,1,2,2,2,2,2,3,3,3,3,3)
>
> # Elementos de la población numerados en orden
> # y ordenados de acuerdo al estrato.
> X=cbind(c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15))
>
> # Probabilidades de inclusión
> pik=rep(2/5, 15)
>
> # Selección de la muestra
> s=balancedstratification(X,strata,pik)
```

Imagen 125. Salida R Muestreo Estratificado

Luego de este último comando, en la consola aparece una descripción de la selección de los elementos de la muestra en cada estrato y en conglomerado; para visualizar los elementos seleccionados se escribe el comando `(1:length(pik))[s==1]` en la consola de R.

```
QUALITY OF BALANCING
TOTALS HorvitzThompson_estimators Relative_deviation
1      2                          2      0.000000e+00
2      2                          2      2.220446e+14
3      2                          2      -2.220446e+14
4     120                         115     -4.166667e+00
>
> # Elementos de la muestra
> (1:length(pik))[s==1]
[1] 3 4 6 9 11 13
>
```

Imagen 126. Salida R Muestreo Estratificado

Así se tiene que los elementos 3 y 4 pertenecen al estrato 1; los elementos 6 y 9, al estrato 2, y los elementos 11 y 13, al estrato 3. Es de notar que este comando sólo se puede utilizar cuando el tamaño de los estratos es igual.

7.5 EJERCICIOS

7.5.1 Determine el tamaño de muestra bajo las siguientes condiciones: error máximo permitido (4%), confiabilidad (93%), con estimación de la media (25) y desviación estándar (1.2).

7.5.2 La Dirección de la universidad está interesada en conocer la opinión de los estudiantes frente al cambio de profesores, para lo cual decide tomar una muestra aleatoria de tamaño 35; considere que los 223 estudiantes tienen asignados códigos entre 1 y 223. Determine cuáles códigos (estudiantes) aparecen en la muestra.

7.5.3 En la empresa JB quieren seleccionar 2 administradores, de los cinco que hay, para asumir cargos fuera del país; se cree que por la experiencia los tres primeros administradores tienen una probabilidad de ser seleccionados de 0.25; el cuarto, una de 0.15, y el quinto, una de 0.1. Asigne un nombre a cada uno de los administradores y determine aleatoriamente cuáles son seleccionados.

7.5.4 En la alcaldía se encuentran las 240 hojas de vida de los trabajadores, numeradas en orden; se quiere revisar si estas cuentan con toda la documentación requerida, para lo cual se extrae una muestra de tamaño 48 mediante el método sistemático. Determinar cuáles son las hojas de vida seleccionadas.