Muestreo y asignación aleatoria

Domingo Martínez

11/1/2021

En esta práctica veremos como hacer un muestreo aletorio simple y uno estratificado, así como realizar la asignación aleatoria de sujetos a un grupo control y a uno experimental.

Objetivo 1: Lectura de la base de datos.

Trabajaremos con la base de datos YRBSS, la cual contiene información de 13583 estudiantes referente al Youth Risk Behavior Surveillance System https://www.cdc.gov/healthyyouth/data/yrbs/data.htm citado en OpenIntro Statistics, David Diez, Mine Cetinkaya-Rundel, Christopher Barr, and OpenIntro, tercera edición, capítulo 4 https://www.openintro.org/book/os/.

IMPORTANTE: Debes asegurarte que el archivo *yrbss.csv* se encuentre en tu directorio de trabajo. Para identificar tu directorio de trabajo ejecuta el siguiente código:

```
getwd()
```

```
## [1] "/home/domingo/Documentos/CURSOS_ENES/Curso_Diseño_Experimental/Introducción_a_R"
```

```
# A continuación leemos la base de datos con el comando read.csv()
# y lo guardamos en la variable "bd"
bd<-read.csv("yrbss.csv")
# Con el comando View() echamos una ojeada a la base de datos.
View(bd)
```

Notarás que tenemos información de las siguientes variables, y otras más.

age: Edad del estudiante.

gender: sexo del estudiante.

grade: Grado en el High school.

height: Estatura del estudiante en metros.

weight: Peso del estudiante en kilogramos.

hekmet: Frecuencia con que el estudiante usó casco al andar en bici en los últimos 12 meses.

active: Número de días en los que relizó actividad física mayor a 60 minutos en los últimos siete días.

lifting: Número de días en lo que realizó entrenamiento vigoroso (e.g. levantar pesas) durante los últimos siete días.

```
# Con el comando names() podemos ver todas nuestras variables.
names(bd)
```

```
##  [1] "age"                    "gender"
##  [3] "grade"                  "hispanic"
##  [5] "race"                   "height"
##  [7] "weight"                 "helmet_12m"
```

```
##  [9] "text_while_driving_30d"   "physically_active_7d"
## [11] "hours_tv_per_school_day"  "strength_training_7d"
## [13] "school_night_hours_sleep"
```

Objetivo 2: Exploración descriptiva de los datos.

En primer lugar vamos a pedirle a R que nos muestre que tipo de datos tenemos.

```
# Usamos el comando str() para conocer la estructura de nuestras variables,
# es decir, el tipo de variables que tenemos.
str(bd)
```

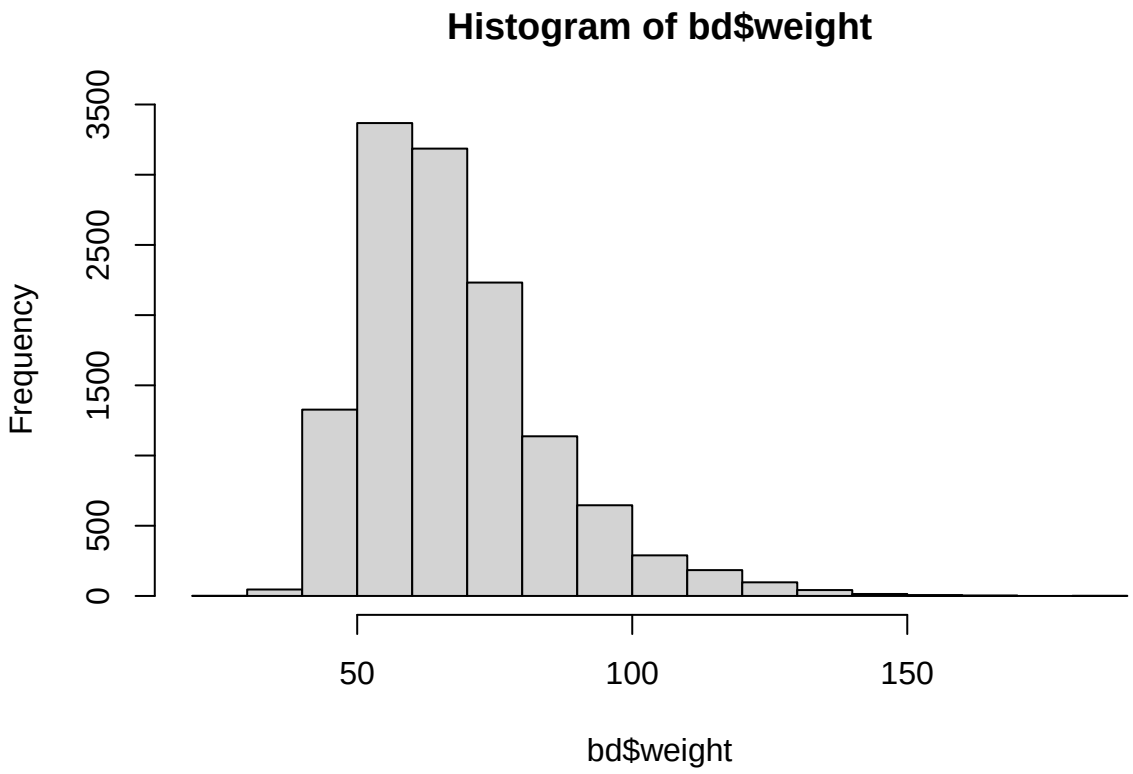
```
## 'data.frame':    13583 obs. of  13 variables:
##  $ age                     : int  14 14 15 15 15 15 15 14 15 15 ...
##  $ gender                  : chr  "female" "female" "female" "female" ...
##  $ grade                   : chr  "9" "9" "9" "9" ...
##  $ hispanic                : chr  "not" "not" "hispanic" "not" ...
##  $ race                    : chr  "Black or African American" "Black or African American" "Native Hau
##  $ height                  : num  NA NA 1.73 1.6 1.5 1.57 1.65 1.88 1.75 1.37 ...
##  $ weight                  : num  NA NA 84.4 55.8 46.7 ...
##  $ helmet_12m              : chr  "never" "never" "never" "never" ...
##  $ text_while_driving_30d  : chr  "0" NA "30" "0" ...
##  $ physically_active_7d    : int  4 2 7 0 2 1 4 4 5 0 ...
##  $ hours_tv_per_school_day : chr  "5+" "5+" "5+" "2" ...
##  $ strength_training_7d    : int  0 0 0 0 1 0 2 0 3 0 ...
##  $ school_night_hours_sleep: chr  "8" "6" "<5" "6" ...
```

```
# Y el comando summary() para ver un resumen de la estadística descriptiva.
summary(bd)
```

```
##       age           gender             grade              hispanic
##  Min.   :12.00   Length:13583       Length:13583       Length:13583
##  1st Qu.:15.00   Class :character   Class :character   Class :character
##  Median :16.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :16.16
##  3rd Qu.:17.00
##  Max.   :18.00
##  NA's   :77
##      race               height          weight        helmet_12m
##  Length:13583       Min.   :1.270   Min.   : 29.94   Length:13583
##  Class :character   1st Qu.:1.600   1st Qu.: 56.25   Class :character
##  Mode  :character   Median :1.680   Median : 64.41   Mode  :character
##                     Mean   :1.691   Mean   : 67.91
##                     3rd Qu.:1.780   3rd Qu.: 76.20
##                     Max.   :2.110   Max.   :180.99
##                     NA's   :1004    NA's   :1004
##  text_while_driving_30d physically_active_7d hours_tv_per_school_day
##  Length:13583           Min.   :0.000        Length:13583
##  Class :character       1st Qu.:2.000        Class :character
##  Mode  :character       Median :4.000        Mode  :character
##                         Mean   :3.903
##                         3rd Qu.:7.000
##                         Max.   :7.000
##                         NA's   :273
##  strength_training_7d school_night_hours_sleep
##  Min.   :0.00         Length:13583
```

```
##  1st Qu.:0.00         Class :character
##  Median :3.00         Mode  :character
##  Mean   :2.95
##  3rd Qu.:5.00
##  Max.   :7.00
##  NA's   :1176
```

```r
# Si queremos visualizar como se distribuye una variable
# usamos el comando hist()
hist(bd$weight)
```



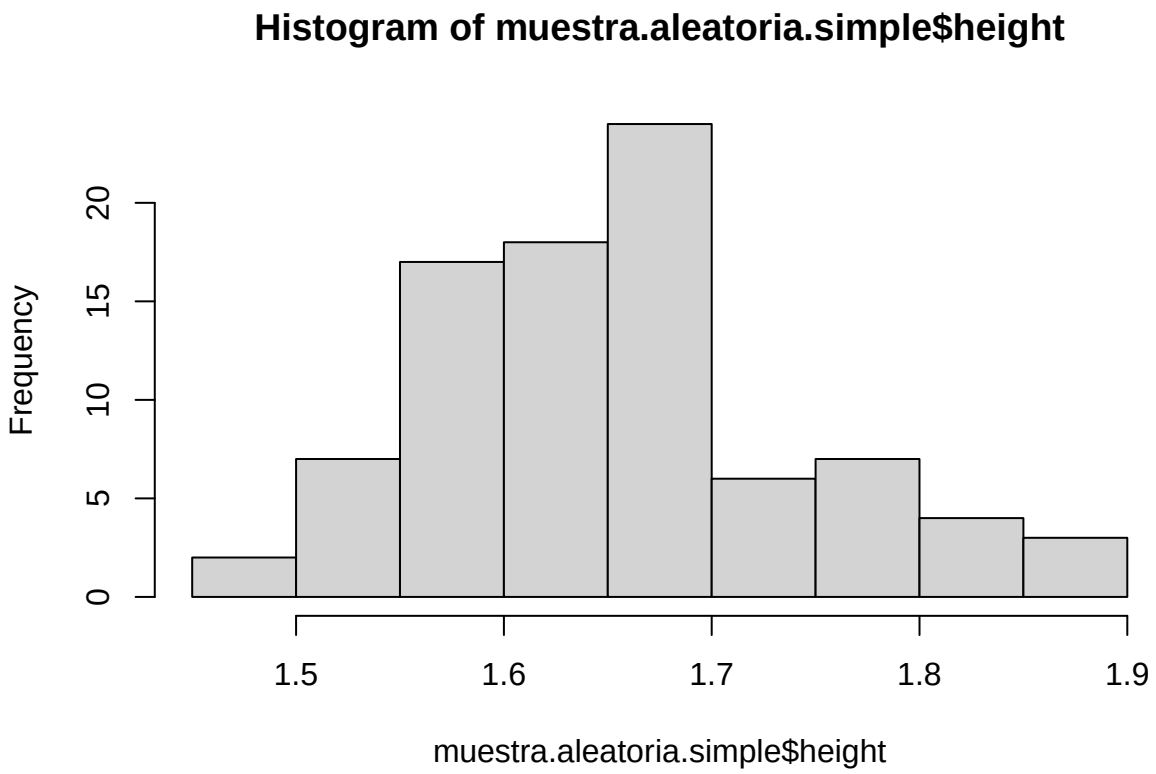
Objetivo 3: Muestro aleatorio simple.

En este apartado vamos a extraer una muestra de 100 sujetos, vamos a observar alguna propiedad de alguna variable númerica y, finalmente, vamos a observar la distribución de dicha variable.

```r
# Con el operador corchetes [], la función nrow() y la función sample
# obtengo una muestra aleatoria simple de tamaño n=100
muestra.aleatoria.simple <- bd[sample(1:nrow(bd), 100),]
View(muestra.aleatoria.simple) # Echo un ojo a los datos de mi muestra
# A continuación calculo la media de la estatura en mi muestra
media.de.la.estatura <- mean(na.omit(muestra.aleatoria.simple$height))
# Con el comando na.omit() quito las observaciones con valores perdidos.
media.de.la.estatura # Calculo la media de mi muestra.
```

```
## [1] 1.667386
```

```r
# Finalmente obserso su histograma
hist(muestra.aleatoria.simple$height)
```



Objetivo 4: Muestreo estratificado.

En este tipo de muestreo quiero obtener una muestra de 100 sujetos, en el que 50 sean mujeres y 50 sean hombres.

```
# Lo primero que hago es filtrar mi bd para que me muestre solamente
# a quienes tienen sexo femenino. Utilizo el comando subset()
bd.mujeres <- subset(bd, gender=="female")
# Ahora obtengo una muestra aleatoria de 50 mujeres de bd.mujeres
muestra.estrato.mujeres<-bd.mujeres[ sample(1:nrow(bd.mujeres),50) ,]
View(muestra.estrato.mujeres)
##########################################################################
########## RETO: OBTEN LA MUESTRA N=50 DEL ESTRATO DE HOMBRES ##########
##########################################################################
```

Objetivo 5: Asignación aleatoria de sujetos.

Supongamos que queremos a realizar un experimento social con los alumnos de la materia de Diseño de Experimentos, nuestro diseño incluye un grupo control y un grupo bajo tratamiento experimental. Asignemos aleatoriamente a nuestros volunatarios en alguno de los dos grupos.

```
# En primer lugar leemos la base de datos.
sujetos<- read.csv("sujetos.csv")
View(sujetos) # Echo un vistazo
# Obtengo una muestra de n=22 a la que asignaré al grupo experimental
grupo.experimental <- sujetos[sample(1:nrow(sujetos),22),]
View(grupo.experimental) # Echemos un vistazo.
# Ahora creo un vector lógico que me diga TRUE si el sujeto pertence al
# grupo experimental y FALSE si no pertenece.
Pertenece.al.grupo.experimental<-sujetos$Id_Sujeto %in% grupo.experimental
```

```
Pertenece.al.grupo.experimental
```

```
##  [1]  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE
## [13]  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE
## [25] FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE
## [37]  TRUE FALSE FALSE  TRUE  TRUE  TRUE FALSE  TRUE
```

```
# Ahora agrego este vector lógico a mi base de datos de sujetos.
sujetos$Pertenece.al.grupo.experimental <- Pertenece.al.grupo.experimental
View(sujetos) # Veo cómo quedó.
# Por último, puedo hacer un filtro con cada grupo.
grupo.control <- subset(sujetos, Pertenece.al.grupo.experimental=="FALSE")
View(grupo.control)
grupo.experimental <- subset(sujetos, Pertenece.al.grupo.experimental=="TRUE")
View(grupo.experimental)
```

Adicionalmente, puedo verificar que ningún sujeto esté en ambos grupos

```
# Con un operador lógico verifico que no se dupliquen sujetos en los grupos
grupo.control==grupo.experimental
```

```
##    Id_Sujeto Pertenece.al.grupo.experimental
## 3      FALSE                           FALSE
## 5      FALSE                           FALSE
## 6      FALSE                           FALSE
## 8      FALSE                           FALSE
## 9      FALSE                           FALSE
## 12     FALSE                           FALSE
## 14     FALSE                           FALSE
## 15     FALSE                           FALSE
## 16     FALSE                           FALSE
## 17     FALSE                           FALSE
## 21     FALSE                           FALSE
## 22     FALSE                           FALSE
## 24     FALSE                           FALSE
## 25     FALSE                           FALSE
## 26     FALSE                           FALSE
## 28     FALSE                           FALSE
## 32     FALSE                           FALSE
## 34     FALSE                           FALSE
## 35     FALSE                           FALSE
## 38     FALSE                           FALSE
## 39     FALSE                           FALSE
## 43     FALSE                           FALSE
```