Estadística descriptiva con datos cuantitativos

Ramon Ceballos

30/1/2021

Funciones para estudiar datos cuantitativos de forma estadística

1. Función summary()

La función **summary()** aplicada a un vector numérico o a una variable cuantitativa nos devuelve un resumen estadístico con los valores mínimo y máximo del vector, sus tres cuartiles y su media.

Al aplicar esta función a un data frame, esta se aplica a todas sus variables de forma simultánea. De este modo, podemos observar rápidamente si hay diferencias notables entre sus variables numéricas.

Ejemplo 1

```
#Cargamos el data frame
cangrejos = read.table("../../../data/datacrab.txt", header = TRUE)

#Eliminamos la primera columna
cangrejos = cangrejos[-1]

#Aplicamos la función summary
summary(cangrejos)
```

```
##      color           spine           width          satell           weight
##  Min.   :2.000   Min.   :1.000   Min.   :21.0   Min.   : 0.000   Min.   :1200
##  1st Qu.:3.000   1st Qu.:2.000   1st Qu.:24.9   1st Qu.: 0.000   1st Qu.:2000
##  Median :3.000   Median :3.000   Median :26.1   Median : 2.000   Median :2350
##  Mean   :3.439   Mean   :2.486   Mean   :26.3   Mean   : 2.919   Mean   :2437
##  3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:27.7   3rd Qu.: 5.000   3rd Qu.:2850
##  Max.   :5.000   Max.   :3.000   Max.   :33.5   Max.   :15.000   Max.   :5200
```

Si nos interesase comparar numéricamente los pesos (weight) y las anchuras (width) de los cangrejos con 3 colores con los que tienen 5 colores, utilizaríamos las siguientes instrucciones:

```
#summarize() para cangrejos de tres colores
summary(subset(cangrejos, color == 3,c("weight","width")))
```

```
##      weight         width
##  Min.   :1300   Min.   :22.5
##  1st Qu.:2100   1st Qu.:25.1
```

```
##  Median :2500   Median :26.5
##  Mean   :2538   Mean   :26.7
##  3rd Qu.:3000   3rd Qu.:28.2
##  Max.   :5200   Max.   :33.5
```

```
#summarize() para cangrejos de cinco colores
summary(subset(cangrejos, color == 5,c("weight","width")))
```

```
##      weight         width
##  Min.   :1300   Min.   :21.00
##  1st Qu.:1900   1st Qu.:23.90
##  Median :2125   Median :25.50
##  Mean   :2174   Mean   :25.28
##  3rd Qu.:2400   3rd Qu.:26.57
##  Max.   :3225   Max.   :29.30
```

Se deduce así que los cangrejos con 5 colores pesan ligeramente menos y tienen menos anchura que los que tienen 3 colores.

2. Función by()

La función **by()** se utiliza para aplicar una determinada función a algunas columnas de un data frame segmentándolas según los niveles de un factor.

La sintaxis de esta función es **by(columnas, factor, FUN = función)**.

Con lo cual, haciendo uso de la función **by** y especificando **FUN = summary**, podremos calcular el resumen estadístico anteriormente comentado a subpoblaciones definidas por los niveles de un factor.

Ejemplo 2

Para este ejemplo, haremos uso del famoso dataset iris.

Si nos interesase calcular de forma rápida y sencilla las longitudes de sépalos y petalos en función de la especie, necesitaríamos hacer uso de la instrucción mostrada a continuación.

Por motivos de espacio, no se muestran los resultados proporcionados por R.

```
#Selección de las columnas 1 (sepal.length) y 3 (petal.length)
#Agrupo por especies (iris$species)
#Aplico la función summary()
by(iris[,c(1,3)], iris$Species, FUN = summary)
```

```
## iris$Species: setosa
##   Sepal.Length    Petal.Length
##  Min.   :4.300   Min.   :1.000
##  1st Qu.:4.800   1st Qu.:1.400
##  Median :5.000   Median :1.500
##  Mean   :5.006   Mean   :1.462
##  3rd Qu.:5.200   3rd Qu.:1.575
##  Max.   :5.800   Max.   :1.900
## ------------------------------------------------------------
## iris$Species: versicolor
```

```
##   Sepal.Length    Petal.Length
##  Min.   :4.900   Min.   :3.00
##  1st Qu.:5.600   1st Qu.:4.00
##  Median :5.900   Median :4.35
##  Mean   :5.936   Mean   :4.26
##  3rd Qu.:6.300   3rd Qu.:4.60
##  Max.   :7.000   Max.   :5.10
## ------------------------------------------------------------
## iris$Species: virginica
##   Sepal.Length    Petal.Length
##  Min.   :4.900   Min.   :4.500
##  1st Qu.:6.225   1st Qu.:5.100
##  Median :6.500   Median :5.550
##  Mean   :6.588   Mean   :5.552
##  3rd Qu.:6.900   3rd Qu.:5.875
##  Max.   :7.900   Max.   :6.900
```

3. Función aggregate()

Tanto la función **by** como la función **aggregate** son equivalentes. No obstante, los resultados se muestran de forma diferente en función de cual utilicemos.

En el caso del ejemplo anterior, convenía más hacer uso de la función **by**.

Podéis comprobarlo introduciendo por consola la siguiente instrucción:

```
# Agrego sepal.length y petal.length a las especies de iris
#tomo los datos recogidos en iris
#aplico summary()
aggregate(cbind(iris$Sepal.Length,iris$Petal.Length)~iris$Species,
          data = iris, FUN = summary)
```

```
##   iris$Species V1.Min. V1.1st Qu. V1.Median V1.Mean V1.3rd Qu. V1.Max. V2.Min.
## 1       setosa   4.300      4.800     5.000   5.006      5.200   5.800   1.000
## 2   versicolor   4.900      5.600     5.900   5.936      6.300   7.000   3.000
## 3    virginica   4.900      6.225     6.500   6.588      6.900   7.900   4.500
##   V2.1st Qu. V2.Median V2.Mean V2.3rd Qu. V2.Max.
## 1      1.400     1.500   1.462      1.575   1.900
## 2      4.000     4.350   4.260      4.600   5.100
## 3      5.100     5.550   5.552      5.875   6.900
```

4. Valores NA

La mayoría de las funciones vistas a lo largo de este tema no funcionan bien con valores **NA**.

Para no tenerlos en cuenta a la hora de aplicar estas funciones, hay que especificar el parámetro **na.rm = TRUE** en el argumento de la función.

```
x = c(1,2,3,NA)
sum(x)
```

```
## [1] NA
```

```r
sum(x, na.rm = TRUE)
```

```
## [1] 6
```

```r
mean(x)
```

```
## [1] NA
```

```r
mean(x, na.rm = TRUE)
```

```
## [1] 2
```

```r
var(x)
```

```
## [1] NA
```

```r
var(x, na.rm = TRUE)
```

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

```r
sd(x)
```

```
## [1] NA
```

```r
sd(x, na.rm = TRUE)
```

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

Ejemplo 3

Importa especificar na.rm =TRUE en las diversas funciones que empleemos para obtener medidas estadísticas.

```r
#Defino una semilla
set.seed(0)

#creo la variable dado2
dados2 = sample(1:6,15, replace = TRUE)
dados2
```

```
##  [1] 6 1 4 1 2 5 3 6 2 3 3 1 5 5 2
```

```r
#anulo la semilla
set.seed(NULL)
```

```r
dadosNA = c(dados2,NA)
dadosNA
```

```
##  [1]  6  1  4  1  2  5  3  6  2  3  3  1  5  5  2 NA
```

```r
mean(dadosNA)
```

```
## [1] NA
```

```r
mean(dadosNA, na.rm = TRUE)
```

```
## [1] 3.266667
```