Diagrama de caja y bigotes

Victor Lopez

2023-01-24

Diagrama de caja

Representa el min, 1er cuartil, mediana, 3er cuartil, el mas y los outlieres

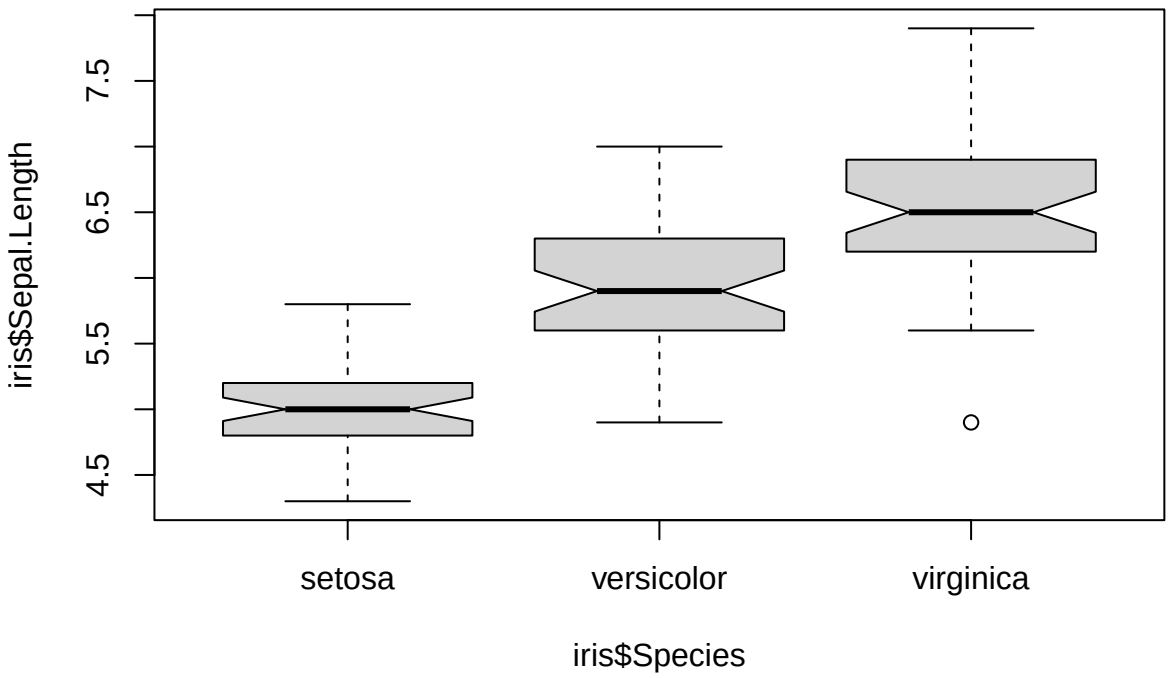
$$bigoteinferior = Q_{0.25} - 1.5 \cdot (Q_{0.75} - Q_{0.25})$$
$$bigotesuperior = Q_{0.75} + 1.5 \cdot (Q_{0.75} - Q_{0.25})$$

En pocas palabras los bigotes estan a 1.5 veces mas de distancia que el rango intercuartilico. Los cuales marcan los valores maximos y minimos, a no ser que haya datos muy alejados de la caja intercuartilica

Graficar

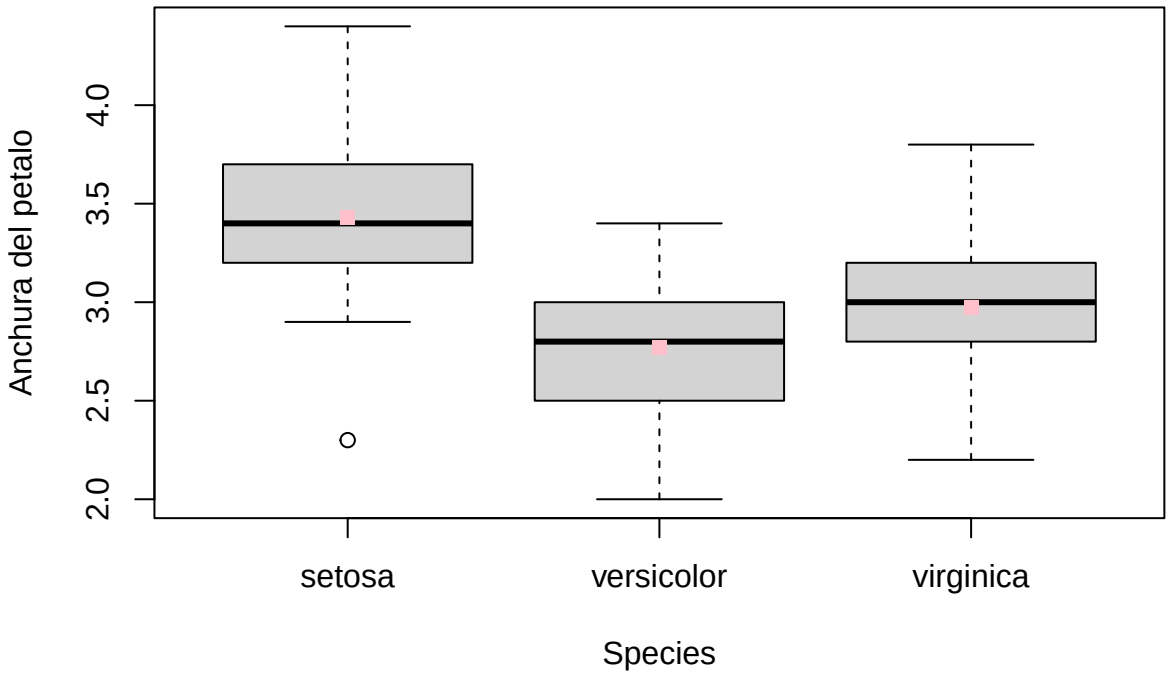
```
boxplot(vector, vector2, vector3) # Cada vector es un boxplot

boxplot(dataframe, names=  c("nombre1", "nombre2")) # Crea un boxplot de cada variable, incluyendo vari
```

```
# Graficar variable agrupada en niveles de otra variable, es decir, un boxplot por cada nivel
boxplot(iris$Sepal.Length~iris$Species, notch = TRUE)
```




```
# El parametro notch añade una muesca en la mediana de la caja. Nos sirven para ver si las
# medianas de dos boxplots se solapan o no y ver si las medianas son o no diferentes
```

```
boxplot(Sepal.Width~Species, data = iris, ylab = "Anchura del petalo")
medias <- aggregate(Sepal.Width~Species, data = iris, mean)
points(x=medias$Species, y=medias$Sepal.Width, col = "pink", pch = 15)
```

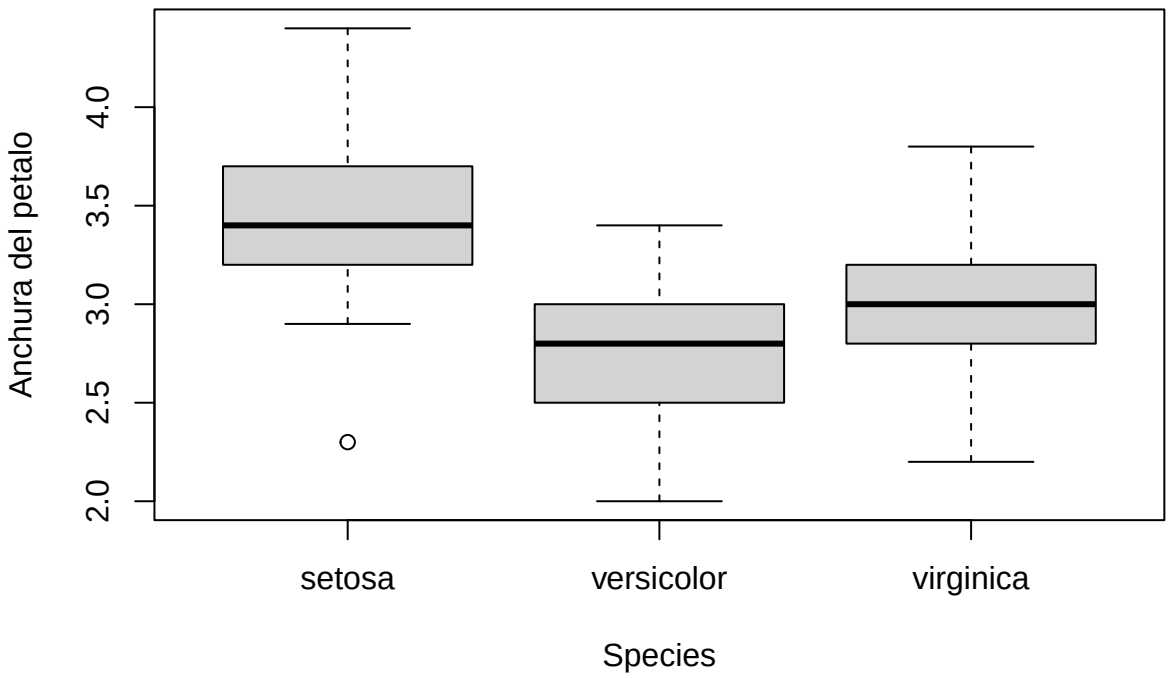



```
# En el pdf si lo pinta bien
```

```
str(boxplot(Sepal.Width~Species, data = iris, ylab = "Anchura del petalo"))
```

```
## List of 6
##  $ stats: num [1:5, 1:3] 2.9 3.2 3.4 3.7 4.4 2 2.5 2.8 3 3.4 ...
##  $ n    : num [1:3] 50 50 50
##  $ conf : num [1:2, 1:3] 3.29 3.51 2.69 2.91 2.91 ...
##  $ out  : num 2.3
##  $ group: num 1
##  $ names: chr [1:3] "setosa" "versicolor" "virginica"
```

```
# stats nos devuelve binf, Q0.25, mediana, Q0.75, bsup y por lo tanto van de 5 en 5
# n indica la cantidad de observaciones utilizadas para cada boxplot
# out nos devuelve los valores atipicos en caso de haber diversos diagramas en un plot
# group nos indica a que diagramas pertenecen esos outliers

boxplot(Sepal.Width~Species, data = iris, ylab = "Anchura del petalo")$stats
```




```
##      [,1] [,2] [,3]
## [1,]  2.9  2.0  2.2
## [2,]  3.2  2.5  2.8
## [3,]  3.4  2.8  3.0
## [4,]  3.7  3.0  3.2
## [5,]  4.4  3.4  3.8
```

```
aggregate(Sepal.Width~Species, data = iris, FUN = sd)
```

```
##      Species Sepal.Width
## 1     setosa   0.3790644
## 2 versicolor   0.3137983
## 3  virginica   0.3224966
```

Ejercicios

Ejercicio 1

Calcula la media, la mediana y la moda redondeada a dos cifras decimales de las notas numéricas de los exámenes tanto de forma conjunta como por estudio universitario. ¿En qué grupo observamos la nota más alta? ¿Qué grupo está por encima de la media de estudiantes?

```r
grades = read.table("../../data/grades.txt")

mod = function(col){
  round(as.numeric(names(which(table(col) == max(table(col))))[[1]], 2)
}

round(median(grades$nota), 2)
```

```
## [1] 3.5
```

```r
round(mean(grades$nota, na.rm = TRUE), 2)
```

```
## [1] 3.87
```

```r
round(mod(grades$nota))
```

```
## [1] 0
```

```r
aggregate(nota ~ estudio, data = grades,
          FUN = function(col){round(median(col, na.rm = TRUE), 2)})
```

```
##       estudio nota
## 1  Industrial 3.44
## 2 Informatica 3.48
## 3       Mates 4.10
## 4  Telematica 3.41
```

```r
aggregate(nota ~ estudio, data = grades,
          FUN = function(col){round(mean(col, na.rm = TRUE), 2)})
```

```
##       estudio nota
## 1  Industrial 3.58
## 2 Informatica 3.95
## 3       Mates 4.22
## 4  Telematica 3.70
```

```r
aggregate(nota ~ estudio, data = grades,
          FUN = mod)
```

```
##       estudio nota
## 1  Industrial 0.00
## 2 Informatica 1.85
## 3       Mates 0.00
## 4  Telematica 0.00
```

```r
aggregate(nota ~ estudio, data = grades,
          FUN = function(col){round(max(col, na.rm = TRUE),3)})
```

```
##       estudio   nota
## 1  Industrial 10.025
## 2 Informatica  8.516
## 3       Mates  8.540
## 4  Telematica  9.290
```

```
# La fun itera sobre las columnas
# Recuerda que es mejor usar aggregate que by()

# En industrial observamos la nota mas alta: 10.025
# Mates esta por encima de la media de estudiantes
```

Ejercicio 2

```
aggregate(nota ~ estudio, data = grades, FUN = sd)
```

```
##       estudio     nota
## 1  Industrial 2.010948
## 2 Informatica 1.537978
## 3       Mates 2.195236
## 4  Telematica 1.909706
```

```
aggregate(nota ~ estudio, data = grades, FUN = var)
```

```
##       estudio     nota
## 1  Industrial 4.043913
## 2 Informatica 2.365377
## 3       Mates 4.819062
## 4  Telematica 3.646977
```

```
# En el de Mates, ya que es el que tiene una mayor varianza y desviacion estandar
```

Ejercicio 3

¿Hay mucha diferencia entre el grupo que has respondido en el ejercicio 1 y en el ejercicio 2? Intenta dar una explicación objetiva del suceso.

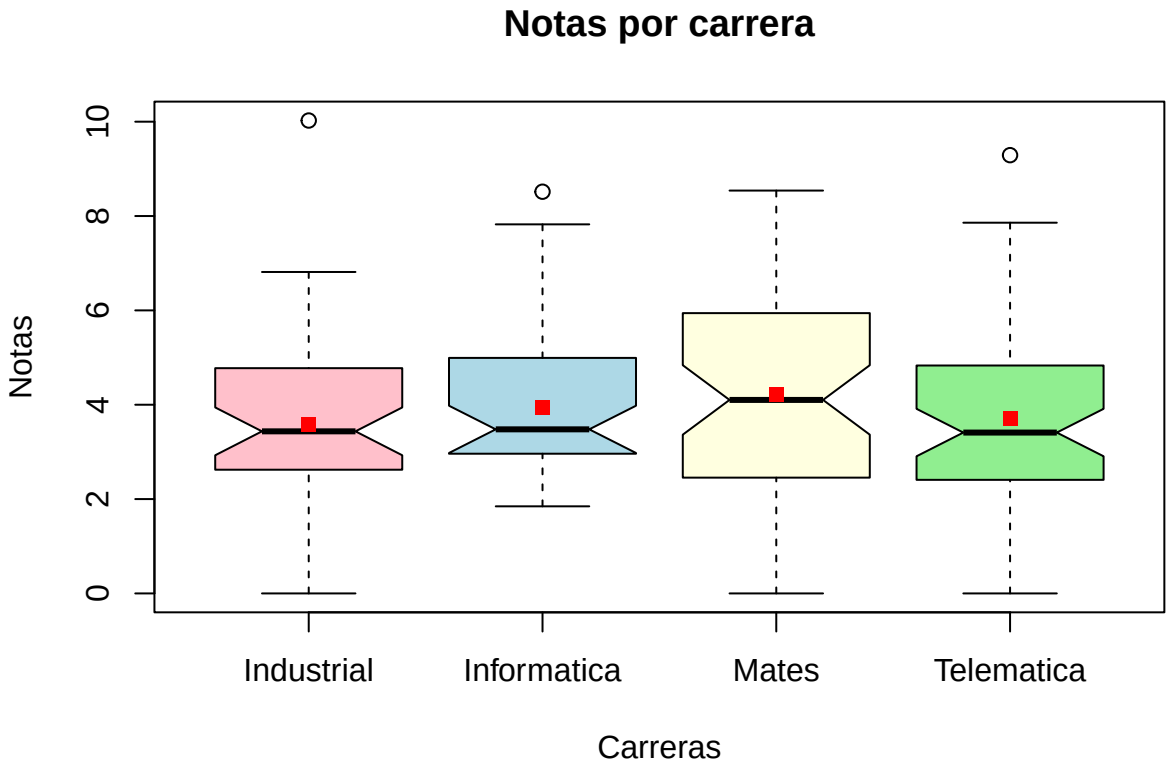
```
# Industrial tiene un promedio maximo de 10, mientras que mates tiene un promedio maximo de 8.54, y tam
```

Ejercicio 4

Dibuja en un único gráfico los cuatro diagramas de caja, uno para cada grupo. Añade también la nota media, pon nombre, título y color al gráfico para documentarlo correctamente.

```
boxplot(grades$nota ~ grades$estudio,
        notch = TRUE,
        main = "Notas por carrera",
```

```
        col = c("pink", "lightblue", "lightyellow", "lightgreen"),
        xlab = "Carreras",
        ylab = "Notas")
medias = aggregate(nota ~ estudio, data = grades,
          FUN = function(col){round(mean(col, na.rm = TRUE), 2)})
points(medias, pch = 15, col = "red")
```



Ejercicio 5

¿Observas algunos valores atípicos en el boxplot anterior? ¿A qué grupo pertence?