Estadística descriptiva con datos cuantitativos

Ramon Ceballos

30/1/2021

Diagrama de Caja y Bigotes (Box plot)

1. Definición y visión general

El conocido **diagrama de caja** o **box plot** es un tipo de gráfico que básicamente, remarca o resume 5 valores estadísticos de variables cuantitativas (se lee de abajo a arriba) que son:

- La **mediana** (2^o cuartil), representada por la línea gruesa que divide la caja.
- El **primer y tercer cuartil**, que son los lados inferior y superior, respectivamente. De este modo, la altura de la caja es el **rango intercuantílico**.
- Los extremos, los valores b_{inf}, b_{sup}, son los **bigotes** (**whiskers**) del gráfico. Si m y M son el **mínimo y máximo** de la variable cuantitativa, entonces los extremos se calculan del siguiente modo:

$$b_{inf} = \max\{m, Q_{0.25} - 1.5(Q_{0.75} - Q_{0.25})\}$$

$$b_{sup} = \min\{M, Q_{0.75} + 1.5(Q_{0.75} - Q_{0.25})\}$$

- **Valores atípicos** o **outliers**, que son los que están más allá de los bigotes. Se marcan como puntos aislados.

Por su definición, concluimos que los bigotes marcan el mínimo y máximo de la variable cuantitativa, a no ser que haya datos muy alejados de la caja intercuantílica.

En tal caso, el bigote inferior marca el valor 1.5 veces el rango intercuantílico por debajo de $Q_{0.25}$, mientras que el superior marca el valor 1.5 veces el rango intercuantílico por encima de $Q_{0.75}$.

Da una referencia muy bena sobre la dispersión de la población que constituye la variable, ya que si hay muchos **outliers** la dispersión será grande y viceversa.

1.2 Función boxplot()

La instrucción `boxplot()` dibuja diagramas de caja en R. En el interior de la función se indica bien un vector, o bien la columna de un Data Frame.

Vamos a aplicar dicha función para la variable dados2.

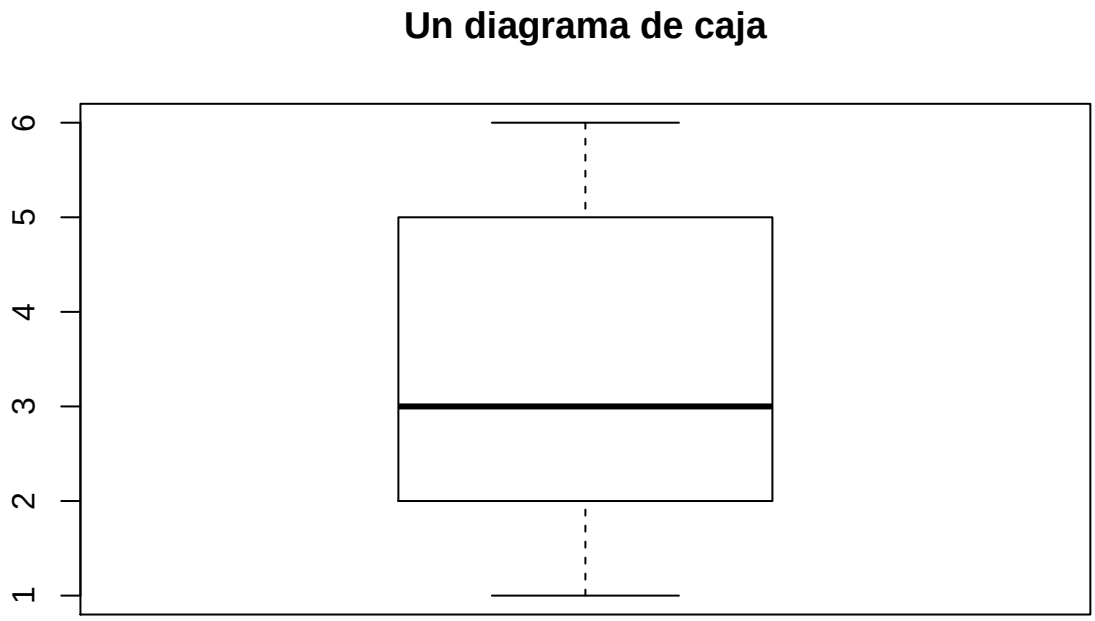
```
#Defino una semilla
set.seed(0)

#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
```

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

```
boxplot(dados2, main = "Un diagrama de caja")
```



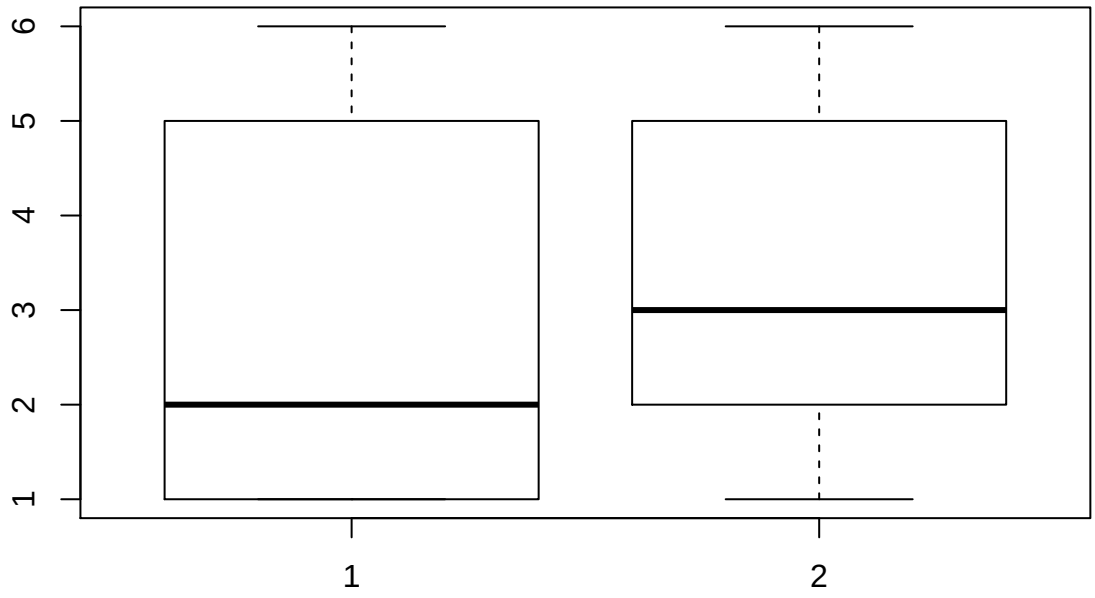
También podemos dibujar diversos diagramas de caja en un mismo gráfico. De este modo, se pueden comparar con mayor facilidad:

```
#Fijo una semilla
set.seed(162017)

#Vector determinado para el ejemplo
dados = sample(1:6,25,replace = TRUE)

set.seed(NULL)
```

```
#boxplot de dos cajas
boxplot(dados,dados2)
```



Los parámetros que se pueden emplear en el boxplot() son los mismos que se podian emplear en la función plot().

1.2.1 Función boxplot() para un data frame

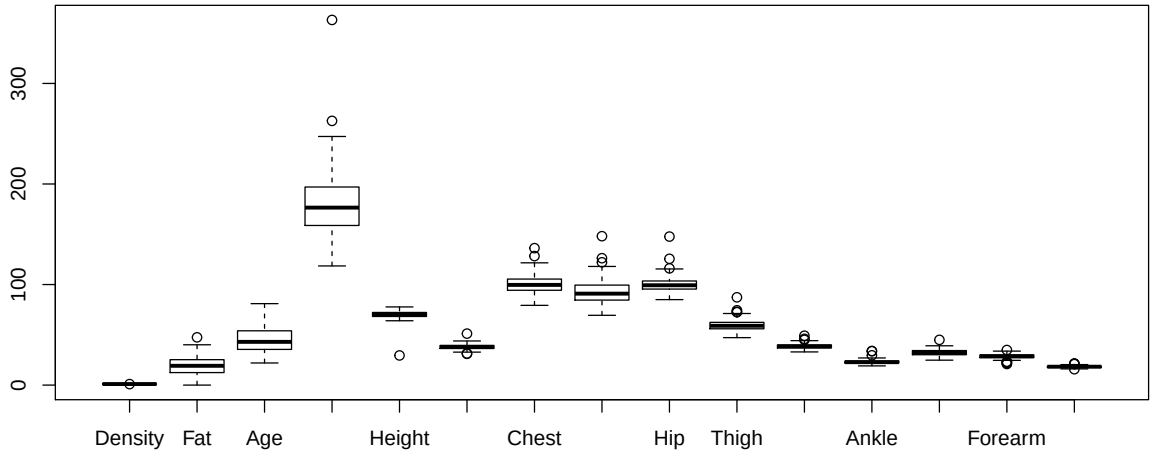
Además, podemos dibujar el diagrama de caja de todas las variables de un data frame en un solo paso aplicando la instrucción **`boxplot(data.frame)`**.

La mayoría de veces, dicho gráfico no será del todo satisfactorio. Dibujar diagramas de factores no tiene sentido alguno. Estos gráficos se pueden manipular incluyendo solo las variables de interés, cambiando los nombres...

Veamos un ejemplo:

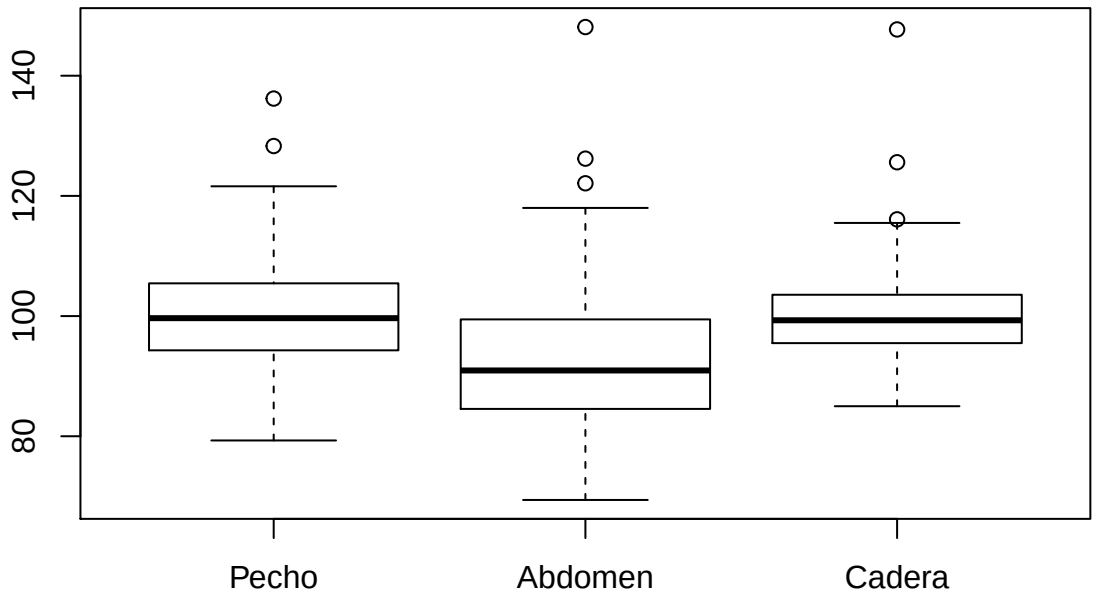
Ejemplo data frame

```
body = read.table("../../../data/bodyfat.txt", header = TRUE)
#Para todo el DF se hace el boxplot
boxplot(body)
```



No tiene mucho sentido hacerlo todo de golpe. Es una mejor idea seleccionar una serie de columnas que se pueda comparar entre sí.

```
#medidas de pecho, abdomen y caderas
boxplot(body[,7:9], names = c("Pecho", "Abdomen", "Cadera"))
```



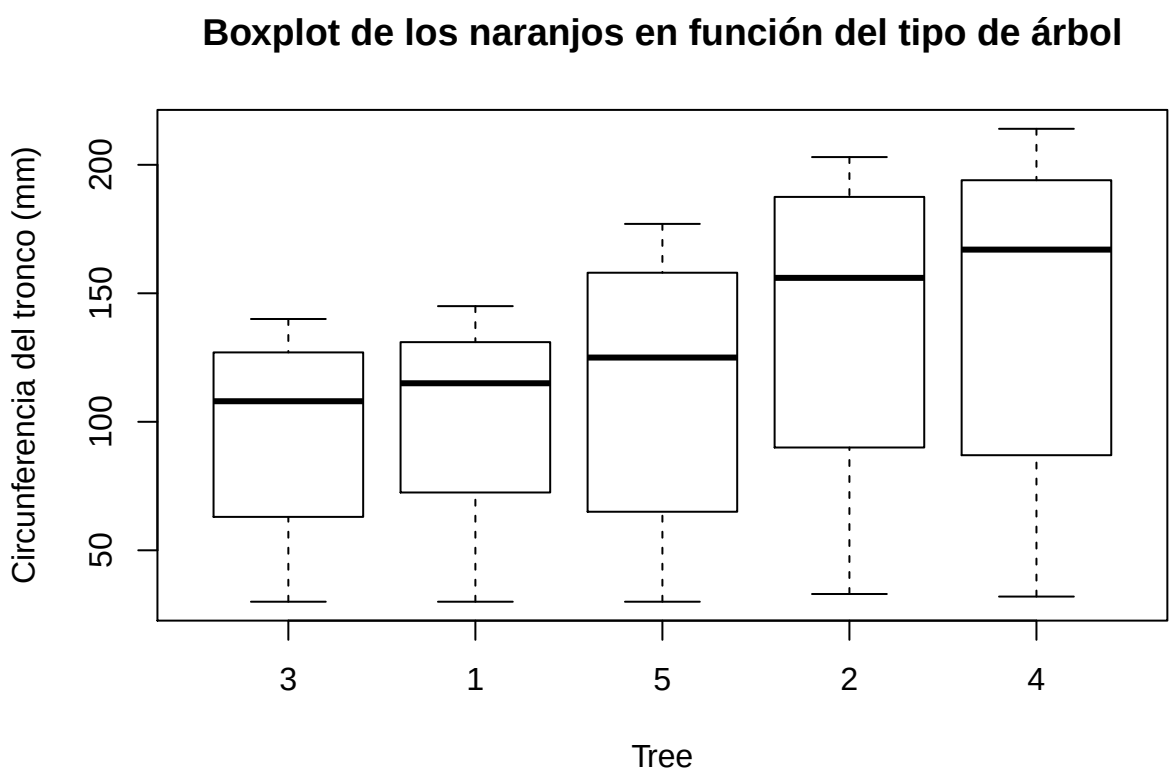
1.2.2 Configuración de los diagramas de cajas

Agrupar varios diagramas de caja en un solo gráfico tiene por objetivo poder compararlos visualmente, lo cual tiene sentido cuando las variables tienen significados parecidos o cuando comparamos una misma variable de poblaciones distintas.

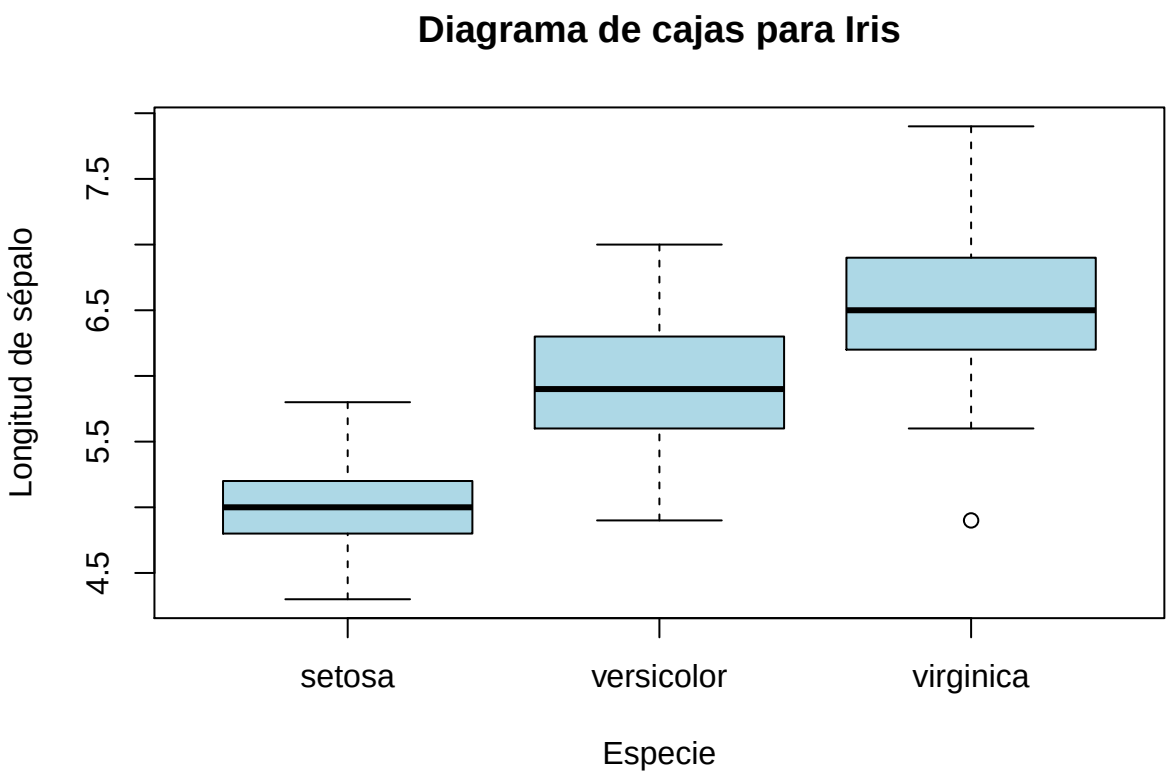
La mayoría de las veces, querremos comparar diagramas de cajas de una misma variable cuantitativa segmentada por los niveles de un factor.

La sintaxis de la instrucción para dibujar en un único gráfico los diagramas de caja de una variable numérica de un data frame en función de los niveles de un factor del mismo data frame es **`boxplot(var.numérica~factor, data = data frame)`**.

```
#Compara el tamaño de la circunferencia para cada tipo de naranjo
boxplot(circumference~Tree, data = Orange, ylab = "Circunferencia del tronco (mm)",
        main = "Boxplot de los naranjos en función del tipo de árbol")
```




```
boxplot(iris$Sepal.Length~iris$Species, ylab = "Longitud de sépalo", xlab = "Especie", main = "Diagrama
```



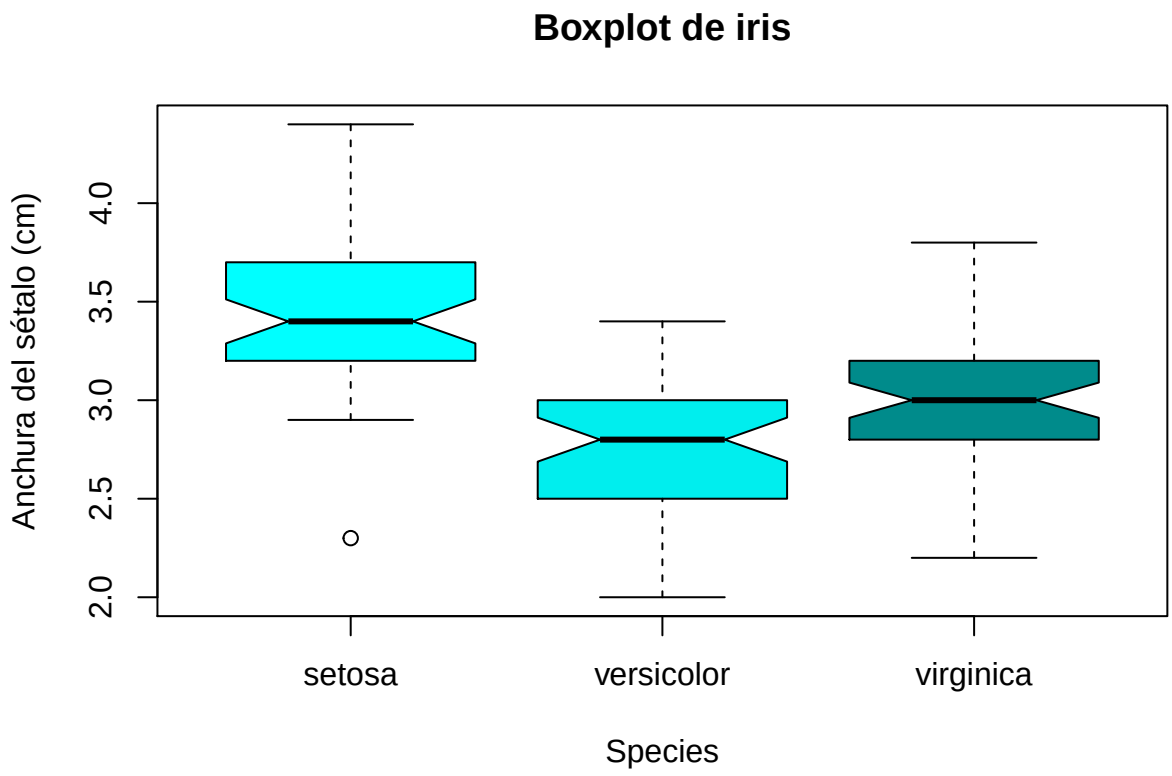
Todos los parámetros de la función **plot()** que tengan sentido pueden ser utilizados en los argumentos de la función **boxplot()**.

Aparte, la función **boxplot()** dispone de algunos parámetros específicos, de los cuales mencionaremos:

- **notch** igualado a **TRUE** añade una muesca en la mediana de la caja. Si se da el caso en que las muescas de dos diagramas de cajas no se solapan, entonces con alto grado de confianza, concluimos que las medianas de las poblaciones correspondientes son diferentes.

```
#Ejemplo en iris para notch = TRUE
boxplot(Sepal.Width~Species,
        data = iris,
        ylab = "Anchura del sétalo (cm)",
        notch = TRUE, col = c("cyan","cyan2","cyan4"),
        main = "Boxplot de iris")
```

Ejemplo 1

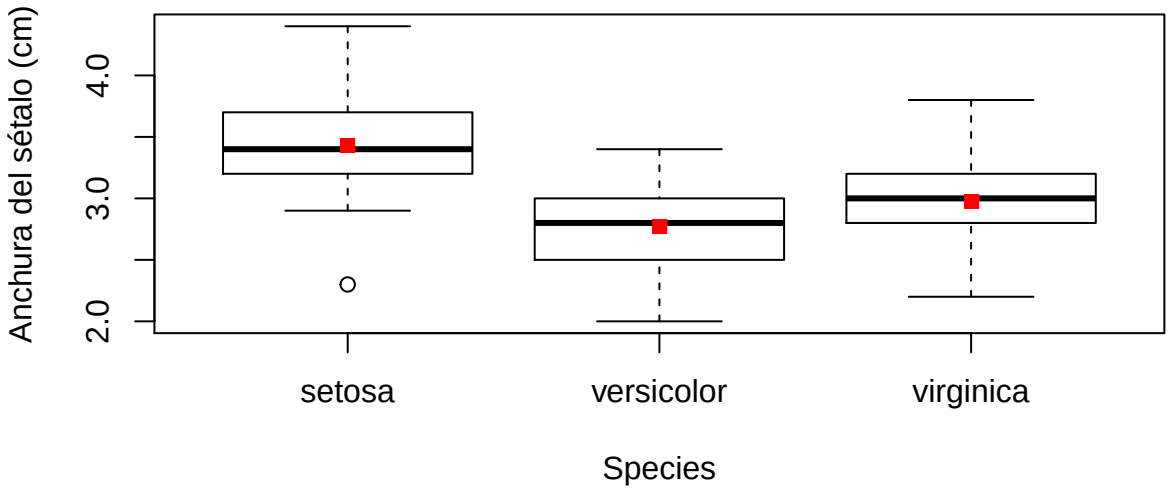


Si quisiéramos marcar de alguna forma en un diagrama de caja, cosa que puede ser muy útil en ocasiones, la media aritmética de la variable correspondiente, podríamos hacerlo mediante la función **`points()`**:

```
#determino el boxplot de estudio
boxplot(Sepal.Width~Species,
        data = iris,
        ylab = "Anchura del sétalo (cm)")

#calculo la media por especies para la variable estudiada
medias = aggregate(Sepal.Width~Species,
                   data = iris,
                   FUN = mean)

#Pinto la media en el boxplot con points() mediante puntos
points(medias, col = "red", pch = 15)
```



La primera instrucción del chunk anterior genera el diagrama de cajas de las anchuras de los sépalos en función de la especie. Por su parte, la segunda instrucción lo que hace es calcular las medias aritméticas de las anchuras según la especie. Finalmente, la tercera instrucción lo que hace es añadir al diagrama un punto cuadrado a cada caja en la ordenada correspondiente a su media aritmética.

1.2.3 La estructura interna de boxplot

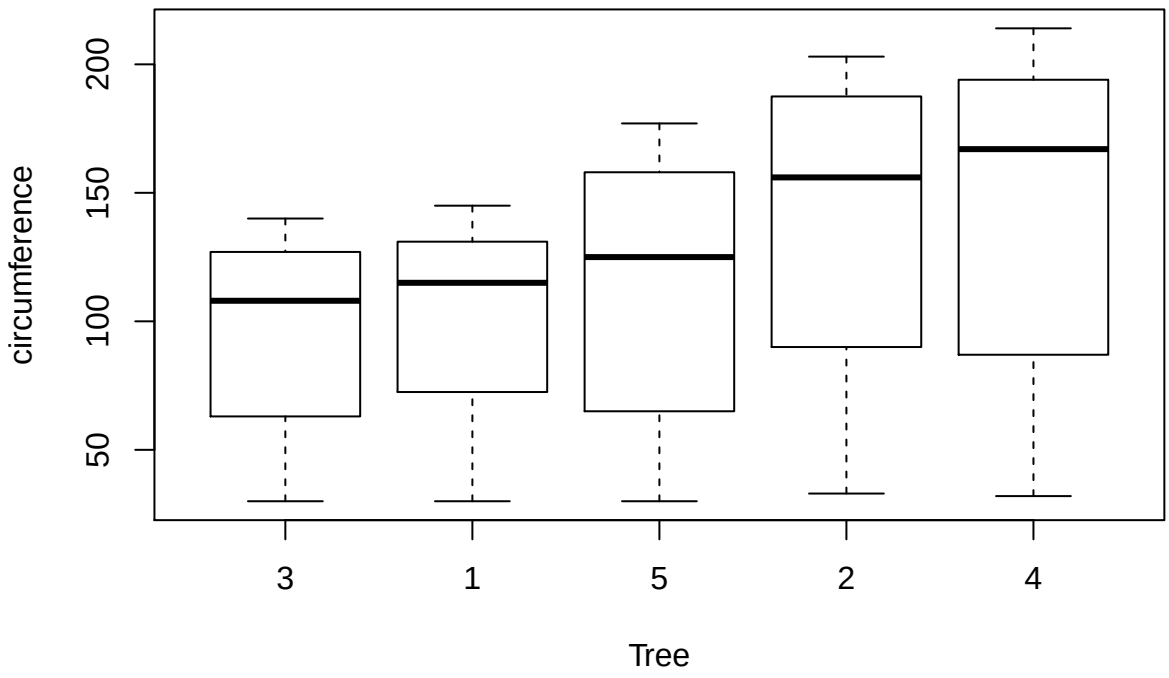
Como ya sabemos, podemos estudiar la función interna de algunos objetos con la función **`str`**.

Dicha función aplicada a un boxplot, nos produce una list. Podéis ver esta list si introducís por consola la siguiente instrucción: **`str(boxplot(circumference~Tree, data = Orange))`**

Destacaremos dos de sus componenetes aquí:

- **`stats`** nos devuelve los valores b_{inf}, $Q_{0.25}$, $Q_{0.5}$, $Q_{0.75}$, b_{sup}
- **`out`** nos retorna los valores atípicos. En caso de haber diversos diagramas en un plot, la componente **`group`** nos indica a qué diagramas pertenecen estos ouliers.

```
str(boxplot(circumference~Tree, data = Orange))
```




```
## List of 6
##  $ stats: num [1:5, 1:5] 30 63 108 127 140 30 72.5 115 131 145 ...
##  $ n    : num [1:5] 7 7 7 7 7
##  $ conf : num [1:2, 1:5] 69.8 146.2 80.1 149.9 69.5 ...
##  $ out  : num(0)
##  $ group: num(0)
##  $ names: chr [1:5] "3" "1" "5" "2" ...
```