R básico

10-2022

Sección 1

Medidas de dispersión

Medidas de dispersión

Las medidas de dispersión evalúan lo dispersos que están los datos. Algunas de las más importantes son:

- El rango o recorrido, que es la diferencia entre el máximo y el mínimo de las observaciones.
- El rango intercuartílico, que es la diferencia entre el tercer y primer cuartil, $Q_{0.75} - Q_{0.25}$.
- La varianza, a la que denotaremos por s^2, es la media aritmética de las diferencias al cuadrado entre los datos x_i y la media aritmética de las observaciones, $\bar{x}$.

$$s^2 = \frac{\sum_{j=1}^{n}(x_j - \bar{x})^2}{n} = \frac{\sum_{j=1}^{k} n_j(X_j - \bar{x})^2}{n} = \sum_{j=1}^{k} f_j(X_j - \bar{x})^2.$$

Medidas de dispersión

- La desviación típica es la raíz cuadrada positiva de la varianza, $s = \sqrt{s^2}$.
- La varianza muestral es la corrección de la varianza. La denotamos por $\tilde{s}^2$ y se corresponde con

$$\tilde{s}^2 = \frac{n}{n-1} s^2 = \frac{\sum_{j=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- La desviación típica muestral, que es la raíz cuadrada positiva de la varianza muestral, $\tilde{s} = \sqrt{\tilde{s}^2}$

Propiedades de la varianza

Propiedades de la varianza.}

- $s^2 \geq 0$. Esto se debe a que, por definición, es una suma de cuadrados de números reales.
- $s^2 = 0 \Longrightarrow x_j - \bar{x} = 0 \ \forall j = 1, \ldots, n$. En consecuencia, si $s^2 = 0$, entonces todos los datos son iguales.
- $s^2 = \frac{\sum_{j=1}^{n} x_j^2}{n} - \bar{x}^2$. Es decir, la varianza es la media de los cuadrados de los datos menos el cuadrado de la media aritmética de estos.

Varianza y varianza muestral

La diferencia entre ambas definiciones viene por la interrelación entre la estadística descriptiva y la inferencial.

Por un lado, es normal medir cómo varían los datos cuantitativos mediante su varianza definida como la media aritmética de las distancias al cuadrado de los datos a su valor medio. No obstante, por otro lado, el conjunto de nuestras observaciones, por lo normal, será una muestra de una población mucho mayor y nos interesará estimar entre otras muchas cosas su variabilidad.

La varianza de una muestra suele dar valores más pequeños que la varianza de la población, mientras que la varianza muestral tiende a dar valores alrededor de la varianza de la población.

Varianza y varianza muestral

Esta corrección, para el caso de una muestra grande no es notable. Dividir n entre $n-1$ en el caso de n ser grande no significa una gran diferencia y aún menos si tenemos en cuenta que lo que tratamos es de estimar la varianza de la población, no de calcularla de forma exacta.

En cambio, si la muestra es relativamente pequeña (digamos $n < 30$), entonces la varianza muestral de la muestra aproxima significativamente mejor la varianza de la población que la varianza.

La diferencia entre desviación típica y desviación típica muestral es análoga.

Con R, calcularemos la varianza y la desviación típica **muestrales**. Con lo cual, si queremos calcular las que no son muestrales, tendremos que multiplicarlas por $\frac{n-1}{n}$, donde n es el tamaño de la muestra. Lo veremos a continuación.

Varianza y desviación típica

Nótese que tanto la varianza como la desviación típica dan una información equivalente. Entonces, es comprensible preguntarse por qué se definen ambas medidas si con una basta. Pues bien, las unidades de la varianza (metros, litros, años...), ya sea muestral o no, están al cuadrado, mientras que las de la desviación típica no.

Medidas de dispersión con R

Medida de dispersión	Instrucción
Valores mínimo y máximo	`range(x)`
Rango	`diff(range(x))`
Rango intercuartílico	`IQR(x, type = ...)`
Varianza muestral	`var(x)`
Desviación típica muestral	`sd(x)`
Varianza	`var(x)*(length(x)-1)/length(x)`
Desviación típica	`sd(x)*sqrt((length(x)-1)/length(x))`

Ejemplo 4

```
dados2
```

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

```
diff(range(dados2))
```

```
## [1] 5
```

```
IQR(dados2)
```

```
## [1] 3
```

```
var(dados2)
```

```
## [1] 3.209524
```

Ejemplo 4

```
sd(dados2)
```

```
## [1] 1.791514
```

```
n = length(dados2)
var(dados2)*(n-1)/n
```

```
## [1] 2.995556
```

```
sd(dados2)*sqrt((n-1)/n)
```

```
## [1] 1.730767
```

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 5

```
cangrejos = read.table("../data/datacrab.txt", header = TRUE)
#Cargamos el data frame
cangrejos = cangrejos[-1] #Eliminamos la primera columna
summary(cangrejos) #Aplicamos la función summary
```

```
##      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
```

Ejemplo 5

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

```
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
```

Ejemplo 5

```
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
```

Y deducimos así que los cangrejos con 5 colores pesan ligeramente menos y tienen menos anchura que los que tienen 3 colores.

La 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 6

Ejemplo 6

Para este ejemplo, haremos uso del famoso dataset iris.

Si nos interesase calcular de forma rápida y sencilla las longitudes de sépalos y pétalos 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.

```
by(iris[,c(1,3)], iris$Species, FUN = summary)
```

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:

```
aggregate(cbind(Sepal.Length,Petal.Length)~Species, data=iris, FUN=summary)
```

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.

Ejemplo 7

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

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

```
mean(dadosNA)
```

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

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

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

Sección 2

Diagramas de caja

Diagramas de caja

El conocido diagrama de caja o box plot es un tipo de gráfico que básicamente, remarca 5 valores estadísticos:

- La mediana, 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.

Más sobre los bigotes

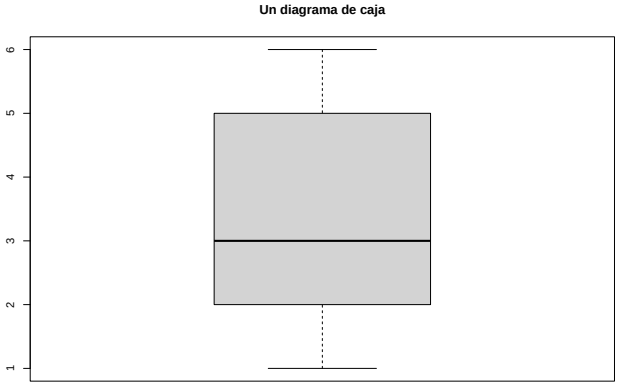
Por su definición, concluimos que los bigotes marcan el mínimo y máximo de la variable cuantitativa, a no ser que haya datos muy alejados de la caja intercuantílica.

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 $Q_{0.75}$

La función boxplot

La instrucción `boxplot()` dibuja diagramas de caja en R.

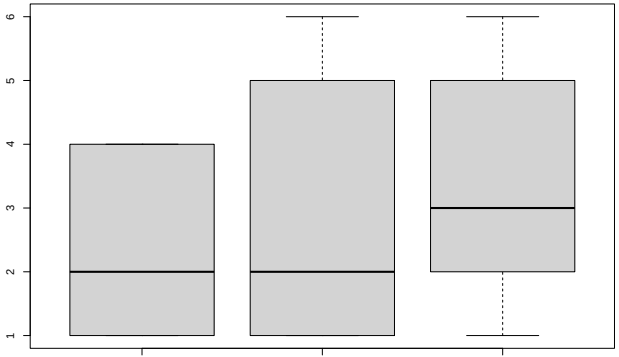
```
boxplot(dados2, main = "Un diagrama de caja")
```



La función boxplot

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

```
boxplot(dado,dados,dados2)
```



La función boxplot

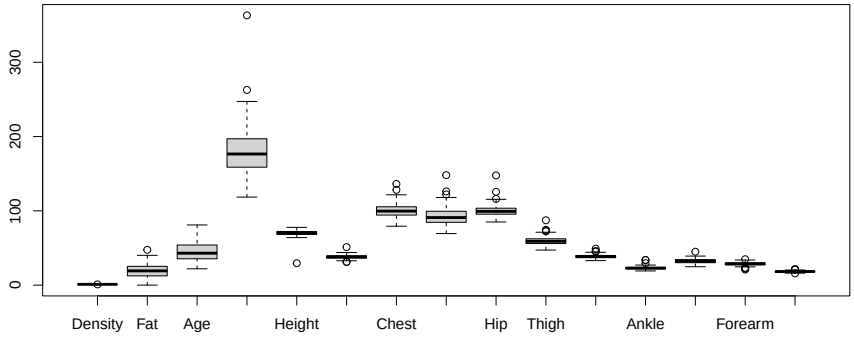
Además, podemos dibujar el diagrama de caja de todas las variables de un data frame en un solo paso aplicando la instrucción `boxplot(data.frame)`.

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:

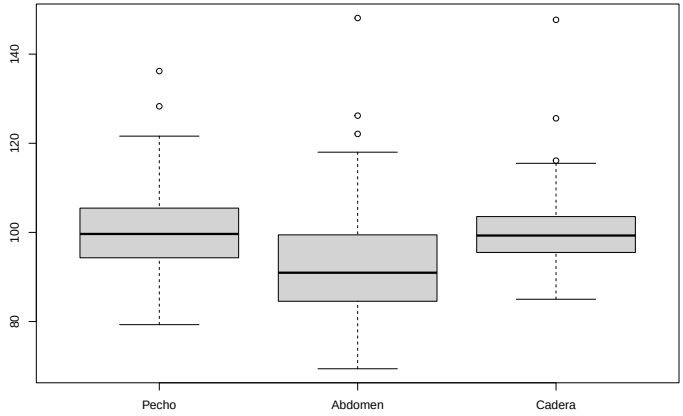
Ejemplo 8

```
body = read.table("../data/bodyfat.txt", header = TRUE)
boxplot(body)
```



Ejemplo 8

```
boxplot(body[,7:9], names = c("Pecho", "Abdomen", "Cadera"))
```



La función boxplot

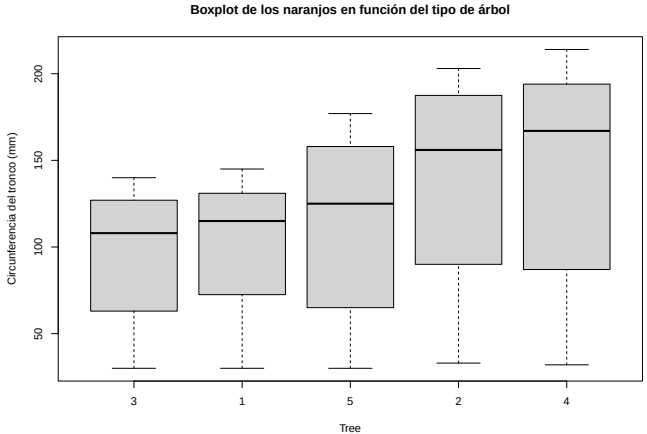
Agrupar varios diagramas de caja en un solo gráfico tiene por objetivo poder compararlos visualmente, lo cual tiene sentido cuando las variables tienen significados parecidos o cuando comparamos una misma variable de poblaciones distintas.

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)`

Ejemplo 9

```
boxplot(circumference~Tree, data = Orange, ylab = "Circunferencia del tronco
        main = "Boxplot de los naranjos en función del tipo de árbol")
```



Parámetros de la función boxplot

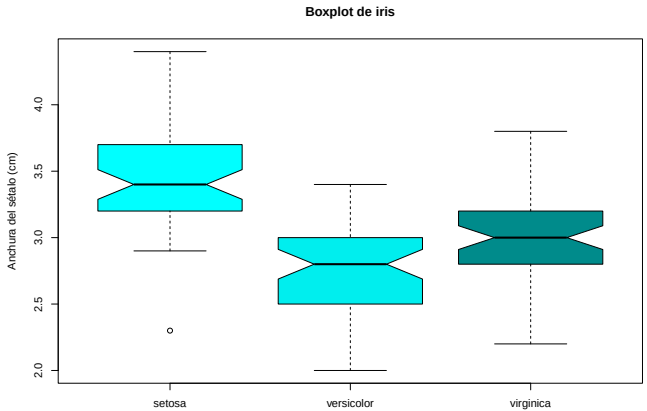
Todos los parámetros de la función `plot()` que tengan sentido pueden ser utilizados en los argumentos de la función `boxplot()`.

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 10

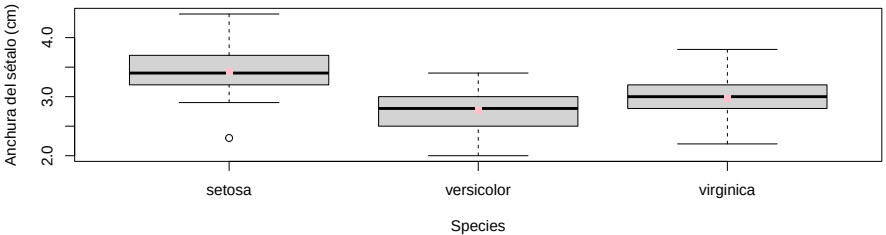
```
boxplot(Sepal.Width~Species, data = iris, ylab = "Anchura del sétalo (cm)",
        notch = TRUE, col = c("cyan","cyan2","cyan4"),
        main = "Boxplot de iris")
```



Ejemplo 10

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`:

```
boxplot(Sepal.Width~Species, data = iris, ylab = "Anchura del sétalo (cm)")
medias = aggregate(Sepal.Width~Species, data = iris, FUN = mean)
points(medias, col = "pink", pch = 15)
```



Ejemplo 10

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.

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 componentes 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.