

# Ciencia de datos, práctica 5

Juan Casado Ballesteros, Samuel García Gonzalez, Iván Anaya Martín

December 3, 2019

## **Abstract**

Pendiente

## 1 Analisis de los datos

En esta parte hablaremos de los datos que vamos a tratar durante la elaboracion de este documento. Los datos que tenemos son dos tablas de datos que usaremos en la primera parte del documento.

```
> datos1 <- read.table("datos1.txt")
> datos1 <- data.frame(datos1)
> datos1
```

	Teoria	Laboratorio
1	4	4
2	4	3
3	5	5
4	1	1
5	5	4

```
> plot(datos1,main="Tabla 1")
```

Estos datos representan la nota de laboratorio y de teorías de unos posibles alumnos. Observaremos si los datos tienen relación entre ellos y si existe un posible outlier.

```
> datos2 <- read.table("datos2.txt")
> datos2 <- data.frame(datos2)
> datos2
```

	Resistencia	Densidad
1	3.0	2.0
2	3.5	12.0
3	4.7	4.1
4	5.2	4.9
5	7.1	6.1
6	6.2	5.2
7	14.0	5.3

```
> plot(datos2,main="Tabla 2")
```

Estos datos representan la densidad y la resistencia de unos materiales. Usaremos técnicas para ver si todos los materiales están relacionados cercanos unos a otros.

## 2 K-vecinos sobre la muestra proporcionada para obtener outliers

Aplicaremos el algoritmo k-vecinos sobre la muestra que tenemos. Este algoritmo identificará de forma supervisada en la muestra datos anómalos, para poder obtener los outliers. En primer lugar deberemos cargar esta desde un archivo .txt.

En segundo lugar calculamos las distancias euclídeas entre todos los puntos

```
> distancias <- as.matrix(dist(datos1))
> distancias
```

	1	2	3	4	5
1	0.000000	1.000000	1.414214	4.242641	1.000000
2	1.000000	0.000000	2.236068	3.605551	1.414214
3	1.414214	2.236068	0.000000	5.656854	1.000000
4	4.242641	3.605551	5.656854	0.000000	5.000000
5	1.000000	1.414214	1.000000	5.000000	0.000000

Posteriormente, ordenamos las distancias de cada punto a todos los demás.

```
> for(i in 1:length(distanciasordenadas[,1])){
+   distancias[,i] <- sort(distancias[,i])
+ }
> distanciasordenadas <- distancias
```

Como último paso, reordenamos la matriz para organizarla en función de la distancia de cada punto a su vecino número 1,2,3...etc. Tras haber organizado la matriz, buscamos en el tercer vecino, que es el valor k que hemos usado en nuestro análisis para poder identificar los outliers.

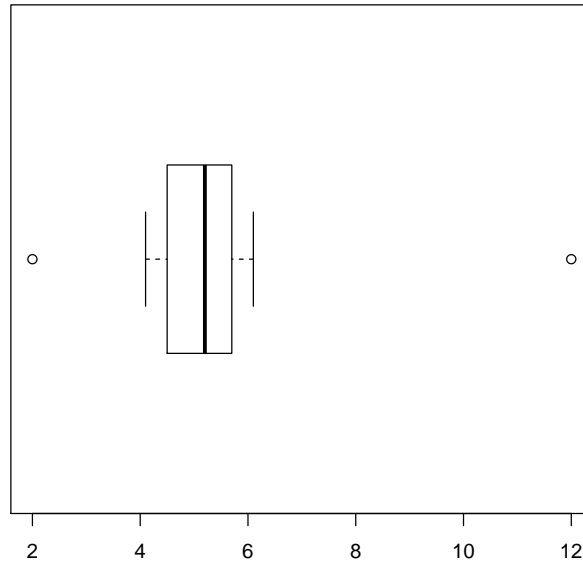
```
> outliers_kvecinos = list()
> for(i in 1:length(distanciasordenadas[,1])){
+   if(distanciasordenadas[4,i]>2.5){
+     outliers_kvecinos[[length(outliers_kvecinos)+1]] <- datos1[i,]
+   }
+ }
> outliers_kvecinos
```

```
[[1]]
      Teoria Laboratorio
4      1              1
```

Como podemos visualizar la siguiente gráfica, los resultados obtenidos son totalmente coherentes con el problema. Representamos en la gráfica los puntos rodeados por un radio de 2,5, y podemos observar que todos poseen a su tercer vecino dentro de su radio menos el punto(1,1)

## 3 Deteccion de datos anómalos sobre la resistencia

```
> boxplot(datos2$Resistencia, range=1.5, horizontal = TRUE)
```



Usamos la funcion para ver los datos en una caja de bigotes.

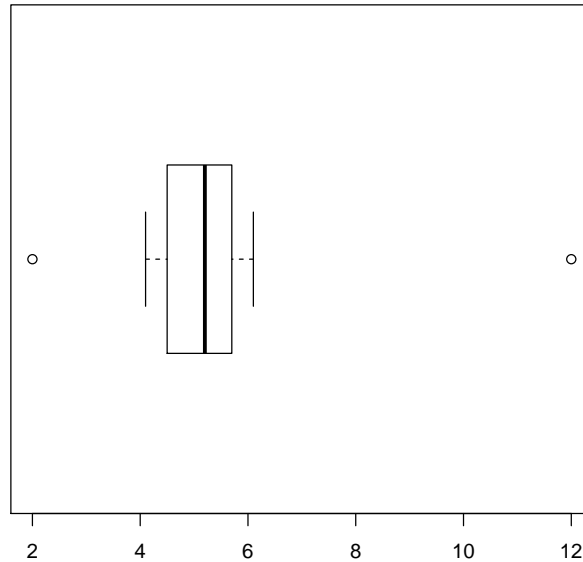
```
> cuart1_res<-quantile(datos2$Resistencia,0.25)
> cuart3_res<-quantile(datos2$Resistencia,0.75)
> int=c(cuart1_res-1.5*(cuart3_res-cuart1_res), cuart3_res+1.5*(cuart3_res-cuart1_res))
> outliers_cuartiles_resistencia = list()
> for(i in 1:length(datos2$Resistencia)){
+   if(datos2$Resistencia[i]<int[1]||datos2$Resistencia[i]>int[2]){
+     outliers_cuartiles_resistencia[[length(outliers_cuartiles_resistencia)+1]] <- t(matr
+   }
+ }
> outliers_cuartiles_resistencia

[[1]]
      Indice Resistencia
[1,]      7          14
```

Mediante el calculo de los primer cuartil y del tercero aplicamos la formula para ver si existe algun outlier en la muestra. Esto nos da un rango todo los valores fuera de este rango son outlier. Luego comprobamos todo los valores de los datos para comprobar si existe, y por lo tanto mostramos.

Como vemos el valor 12 es un outlier de los datos

```
> boxplot(datos2$Densidad, range=1.5, horizontal = TRUE)
```



```
> cuart1_den<-quantile(datos2$Densidad,0.25)
> cuart3_den<-quantile(datos2$Densidad,0.75)
> int=c(cuart1_den-1.5*(cuart3_den-cuart1_den), cuart3_den+1.5*(cuart3_den-cuart1_den))
> outliers_cuartiles_densidad <- list()
> for(i in 1:length(datos2$Densidad)){
+   if(datos2$Densidad[i]<int[1]||datos2$Densidad[i]>int[2]){
+     outliers_cuartiles_densidad[[length(outliers_cuartiles_densidad)+1]] <- t(matrix(c(i
+   })
+ }
> outliers_cuartiles_densidad
```

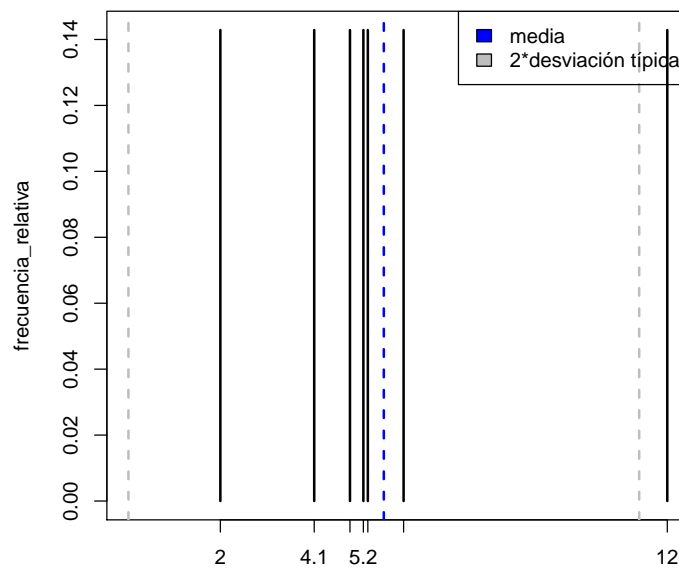
```
[[1]]
      Indice Densidad
[1,]         1         2
```

```
[[2]]
      Indice Densidad
[1,]         2        12
```

Realizamos, vemos que esta vez hay dos outlier uno es el 2 y el otro 12 de la densidad

## 4 Dispersión sobre la densidad, desviación típica

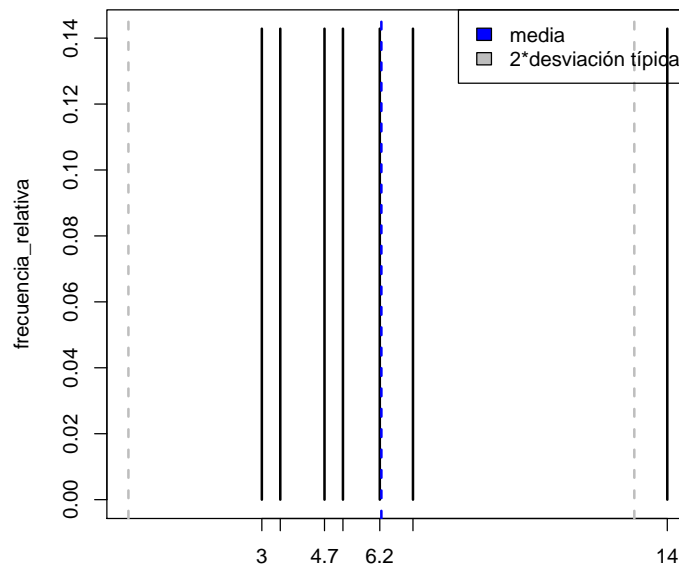
```
> datos2 <- read.table("datos2.txt")
> plotFrequencyData(datos2$Densidad)
```



```
> int <- c(mediaAritmetica(datos2$Densidad) - 2*desviacionTipica(datos2$Densidad), mediaAr
> outliers_desviacion = list()
> for(i in 1:length(datos2$Densidad)) {
+   if ((datos2$Densidad[i]<int[1]) || (datos2$Densidad[i]>int[2])) {
+     outliers_desviacion[[length(outliers_desviacion)+1]] <- t(matrix(c(i, datos2[i,]$Den
+   })
+ }
> outliers_desviacion

[[1]]
      Indice Densidad
[1,]      2      12

> plotFrequencyData(datos2$Resistencia)
```



```
> int <- c(mediaAritmetica(datos2$Resistencia) - 2*desviacionTipica(datos2$Resistencia), m
> outliers_desviacion = list()
> for(i in 1:length(datos2$Resistencia)) {
+   if ((datos2$Resistencia[i]<int[1]) || (datos2$Resistencia[i]>int[2])) {
+     outliers_desviacion[[length(outliers_desviacion)+1]] <- t(matrix(c(i, datos2[i,]$Res
+   })
+ }
> outliers_desviacion

[[1]]
      Indice Resistencia
[1,]          7          14
```

## 5 detección de datos anómalos sobre la regresión de la densidad en función de la resistencia

En este análisis detectamos los outliers utilizando la recta de regresión y el error estándar de los residuos. Como primer paso leemos la tabla de los datos.

```
> datos2 <- read.table("datos2.txt")
> datos2
```

```
      Resistencia Densidad
1             3.0        2.0
```

2	3.5	12.0
3	4.7	4.1
4	5.2	4.9
5	7.1	6.1
6	6.2	5.2
7	14.0	5.3

Seguidamente calculamos la recta de regresión sobre los datos.

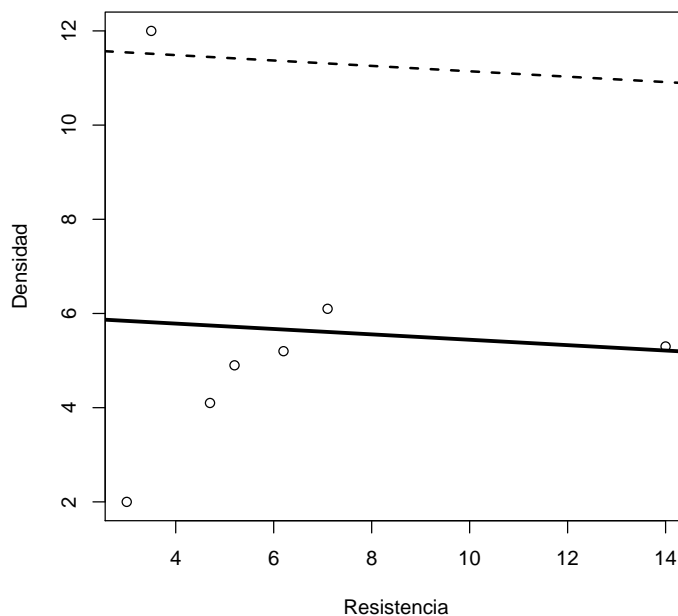
```
> dFr=lm(datos2$Densidad~datos2$Resistencia)
```

Posteriormente, obtenemos los residuos calculados a partir de la recta de regresión.

```
> res=summary(dFr)$residuals
```

Para finalizar, calculamos el error estándar de los residuos, y a partir de él comparamos cada uno para comprobar si algún residuo es 2 veces mayor que el error estándar. Si se da el caso, podemos considerar ese punto como un outlier.

```
> regPlot(datos2$Resistencia, datos2$Densidad, dFr, 2*sr, "Resistencia", "Densidad")
```



```
> sr=sqrt(sum(res^2)/length(res))
> outliers_regression = list()
> for(i in 1:length(res)){
+   if(abs(res[i])>2*sr){
+     outliers_regression[[length(outliers_regression)+1]] <- datos2[i,]
+   }
+ }
> outliers_regression
```



```
[[1]]
  Resistencia Densidad
2          3.5        12
```