Cálculo de la distancia Mahalanobis

Maisy Samai Vázquez Sánchez

2022-06-05

Ejercicio 1

```
# Cargar los datos
ventas= c( 1054, 1057, 1058, 1060, 1061, 1060, 1061,
           1062, 1062, 1064, 1062, 1062, 1064, 1056,
           1066, 1070)
clientes= c(63, 66, 68, 69, 68, 71, 70, 70, 71, 72, 72,
            73, 73, 75, 76, 78)
# Utilizamos la función data.frame() para crear
# un juego de datos en R
datos <- data.frame(ventas ,clientes)
```

Exploración de los datos

```
library(knitr)
dim(datos)
```

```
## [1] 16  2
```

```
str(datos)
```

```
## 'data.frame':    16 obs. of  2 variables:
##  $ ventas  : num  1054 1057 1058 1060 1061 ...
##  $ clientes: num  63 66 68 69 68 71 70 70 71 72 ...
```

```
kable(summary(datos))
```

ventas	clientes
Min. :1054	Min. :63.00
1st Qu.:1060	1st Qu.:68.75
Median :1062	Median :71.00
Mean :1061	Mean :70.94
3rd Qu.:1062	3rd Qu.:73.00
Max. :1070	Max. :78.00

Hay 16 observaciones y 2 variables numéricas.

Calculo de la distancia de Mahalanobis

El método de distancia Mahalanobis mejora el método clásico de distancia de Gauss eliminando el efecto que pueden producir la correlación entre las variables a analizar

1.Determinar el número de outlier que queremos encontrar.

```
num.outliers <- 2
```

2.Ordenar los datos de mayor a menor distancia, según la métrica de Mahalanobis.

```
mah.ordenacion <- order(mahalanobis(datos, colMeans(datos), cov(datos)), decreasing=TRUE)
mah.ordenacion
```

```
##  [1] 14 16  1 15  2  5  3 10 13  8 12  4  6  7  9 11
```

3.Generar un vector boleano los dos valores más alejados segun la distancia Mahalanobis.

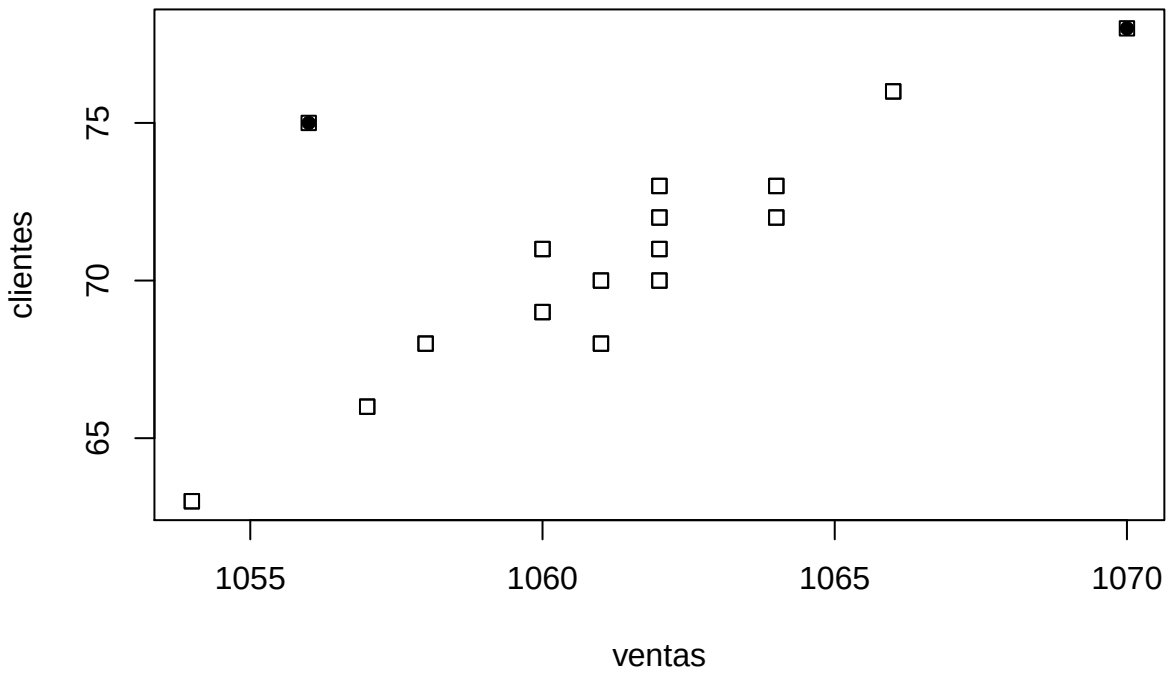
```
outlier2 <- rep(FALSE , nrow(datos))
outlier2[mah.ordenacion[1:num.outliers]] <- TRUE
```

4.Resaltar con un punto relleno los 2 valores outliers.

```
colorear.outlier <- outlier2 *16
```

5.Visualizar el gráfico con los datos destacando sus outlier.

```
plot(datos , pch=0)
points(datos , pch=colorear.outlier)
```



Ejercicio 2

Se generan datos, su matriz de varianzas y su distancia de mahalanbis

```
require(graphics)
ma <- cbind(1:6, 1:3)
(S <-  var(ma))
```

```
##      [,1] [,2]
## [1,]  3.5  0.8
## [2,]  0.8  0.8
```

```
mahalanobis(c(0, 0), 1:2, S)
```

```
## [1] 5.37037
```

```
x <- matrix(rnorm(100*3), ncol = 3)
stopifnot(mahalanobis(x, 0,
                      diag(ncol(x))) == rowSums(x*x))
```

Se usa D^2 como la distancia Euclidea comun

```
Sx <- cov(x)
D2 <- mahalanobis(x, colMeans(x), Sx)
```

Gráfico de la densidad de las distancias de Mahalanobis

```
plot(density(D2, bw = 0.5),
     main="Squared Mahalanobis distances,
     n=100, p=3") ; rug(D2)
```

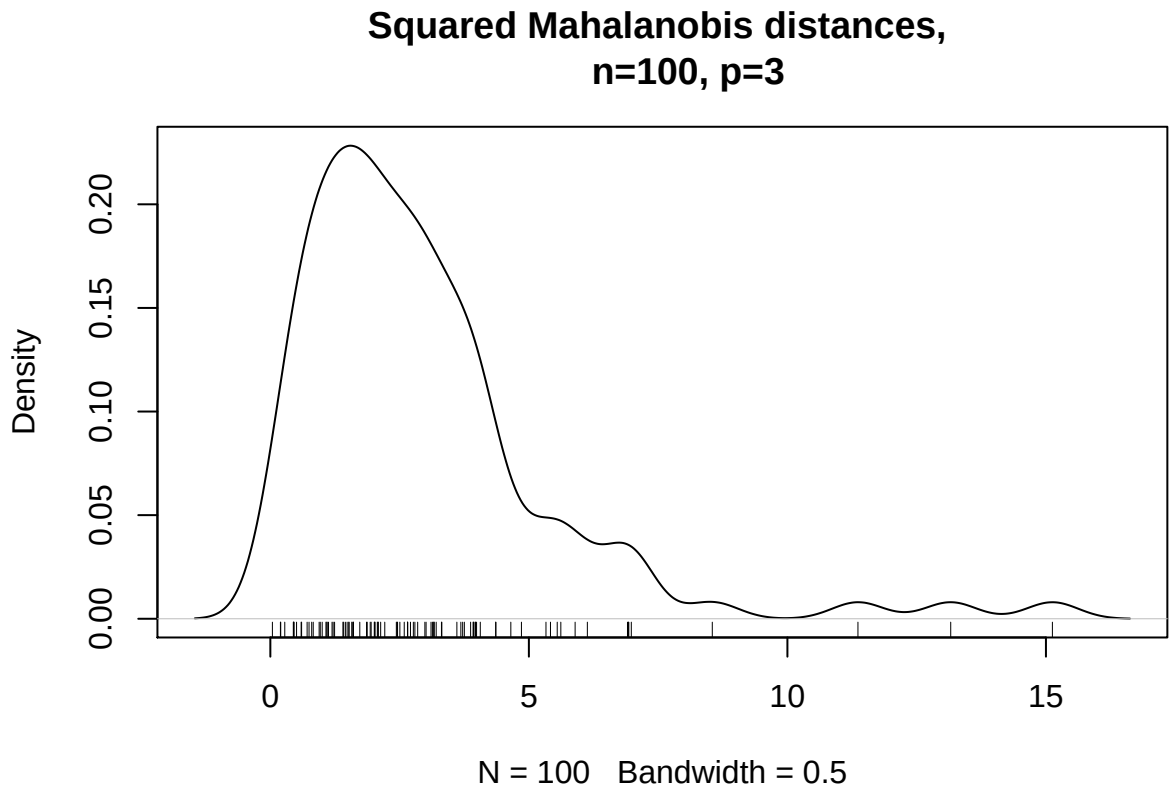
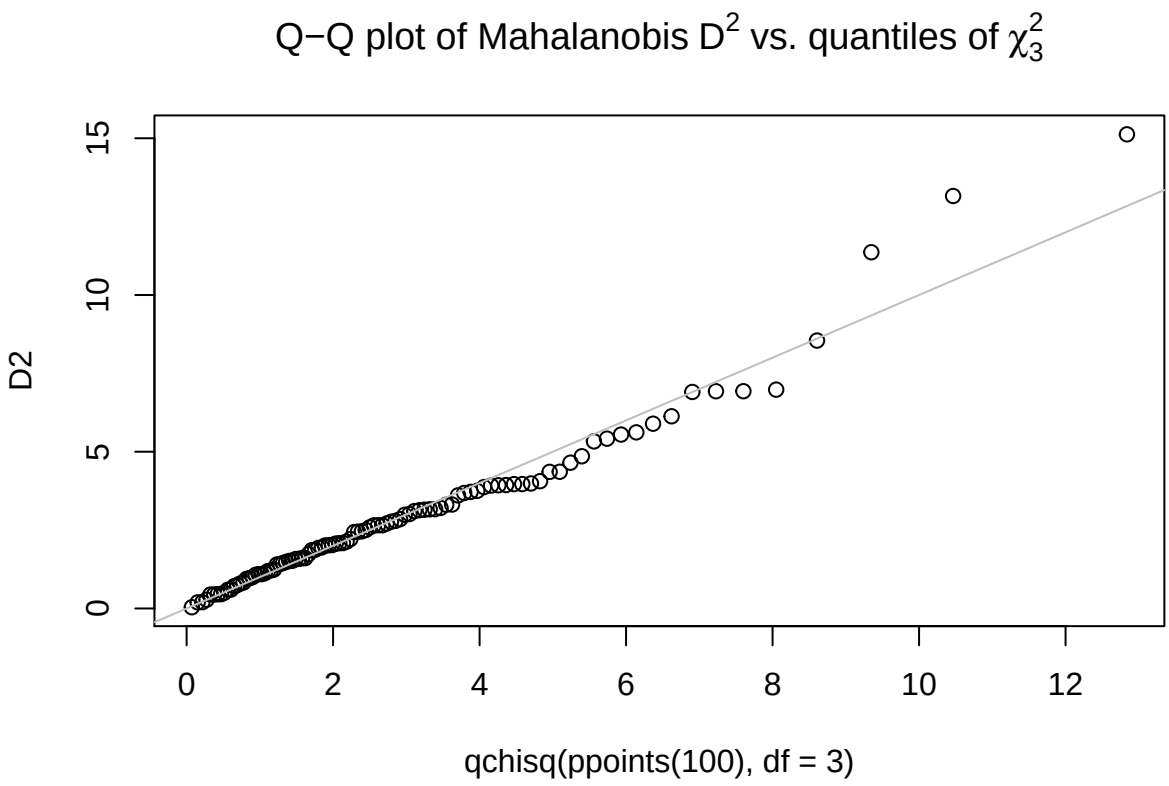


Gráfico qqplot sobre los datos

```
qqplot(qchisq(ppoints(100), df = 3), D2,
       main = expression("Q-Q plot of Mahalanobis" * ~D^2 *
                         " vs. quantiles of" * ~ chi[3]^2))
abline(0, 1, col = 'gray')
```



Ejercicio Propuesto

Descripcion de la matriz de datos

```
#Reviso la dimension de la matriz procesada
datos = as.data.frame(iris)
dim(datos)
```

```
## [1] 150   5
```

```
names(datos)
```

```
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
```

```
str(datos)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

La matriz de datos esta compuesta por 150 observaciones y 4 variables.

Acercamiento Exploratorio en 3 dimensiones

```
datos = datos[is.na(datos$Petal.Length)==F,]
datos = datos[datos$Species!="Versicolor" ,]
dim(datos)
```

```
## [1] 150   5
```

```
row.names(datos) = 1:nrow(datos)
datos = datos[,c(1,2,3,4)]
head(datos)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1          5.1         3.5          1.4         0.2
## 2          4.9         3.0          1.4         0.2
## 3          4.7         3.2          1.3         0.2
## 4          4.6         3.1          1.5         0.2
## 5          5.0         3.6          1.4         0.2
## 6          5.4         3.9          1.7         0.4
```

```
names(datos) = c("L.Sepalo","A.Sepalo","L.Petalo","A.Petalo")

dim(datos)
```

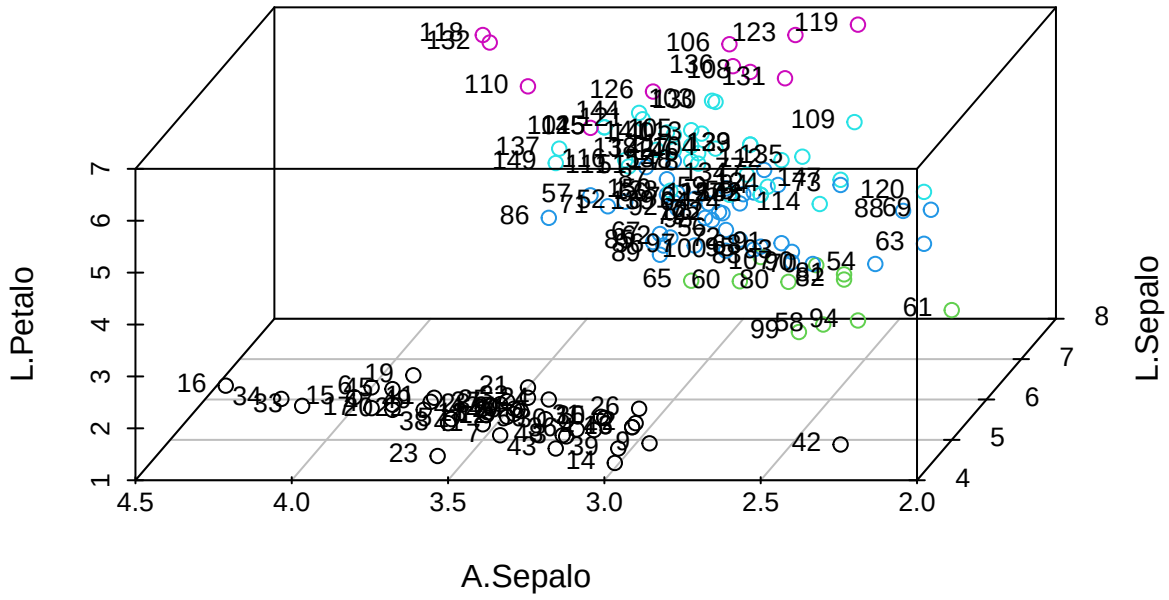
```
## [1] 150   4
```

```
names(datos)
```

```
## [1] "L.Sepalo" "A.Sepalo" "L.Petalo" "A.Petalo"
```

Se hace un gráfico de dispersión en 3d para observar el comportamiento de los sujetos:

```
library(scatterplot3d)
zz <- scatterplot3d(x = datos[,1], y = datos[,2],
                    z = datos[,3],
                    xlab = "L.Sepalo", ylab = "A.Sepalo", zlab = "L.Petalo",
                    pch = 1, color = as.numeric(datos$L.Petalo), grid = TRUE,
                    angle = 250)
zz.coords <- zz$xyz.convert(datos[,1], datos[,2], datos[,3])
text(zz.coords$x,
     zz.coords$y,
     labels = row.names(datos),
     cex = .8,
     pos = 2)
```




```
head(datos)
```

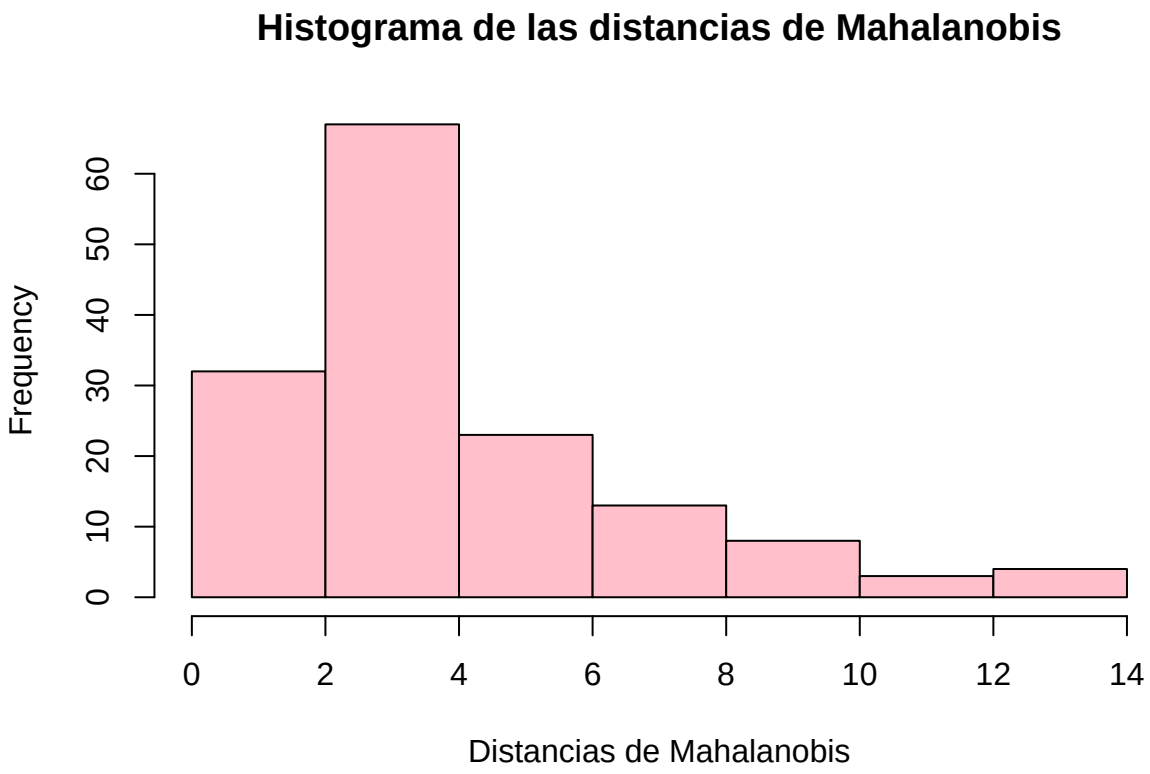
```
##   L.Sepalo A.Sepalo L.Petalo A.Petalo
## 1      5.1      3.5      1.4      0.2
## 2      4.9      3.0      1.4      0.2
## 3      4.7      3.2      1.3      0.2
## 4      4.6      3.1      1.5      0.2
## 5      5.0      3.6      1.4      0.2
## 6      5.4      3.9      1.7      0.4
```

Obtención de las distancias de mahalanobis

```
datosmahal = mahalanobis(datos,center = colMeans(datos), cov = cov(datos))
```

Observo la distribución de las distancias mediante un histograma

```
hist(datosmahal,col = "Pink",main = "Histograma de las distancias de Mahalanobis",
     xlab = "Distancias de Mahalanobis")
```



El histograma revela la presencia de valores muy alejado del resto.

Gráfico de cajas para detectar valores atipicos

```
boxplot(datosmahal,col = "pink")
```

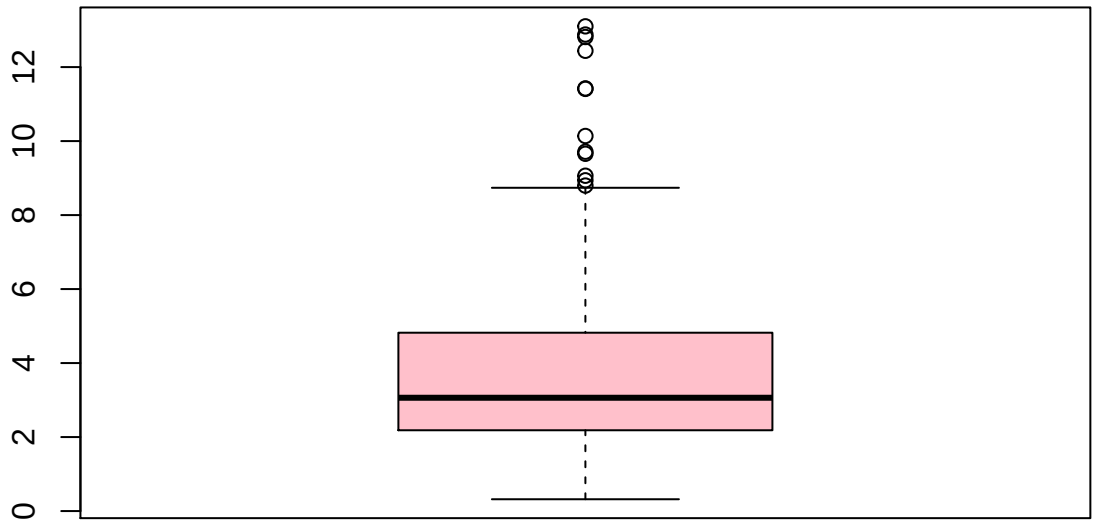
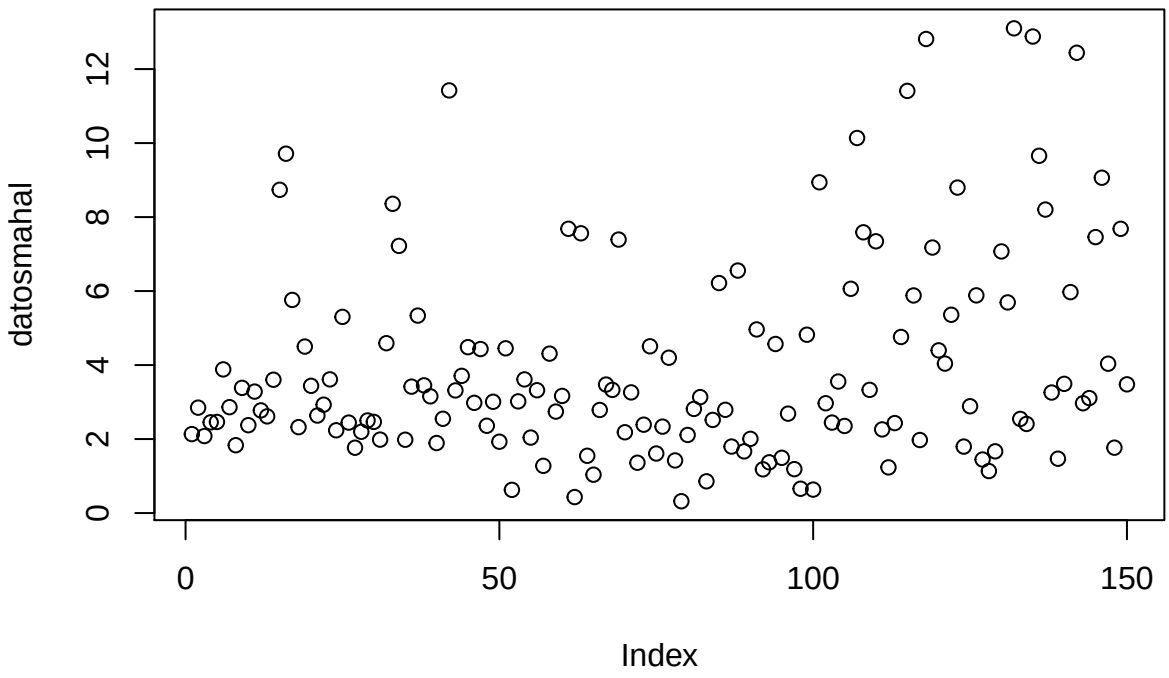


Gráfico de las distancias de Mahalanobis

```
plot(datosmahal)
```



Deteccion de los valores mas alejados

```
dmsa = order(datosmahal,decreasing = T)[1:4]
dmsa
```

```
## [1] 132 135 118 142
```

Se extraen los valores

```
limpios = datosmahal[-dmsa]
```

Histograma de la nueva base

```
hist(limpios,col = "Pink", main = "Histograma de los datos limpios",xlab =  "Distancias de Mahalanobis")
```

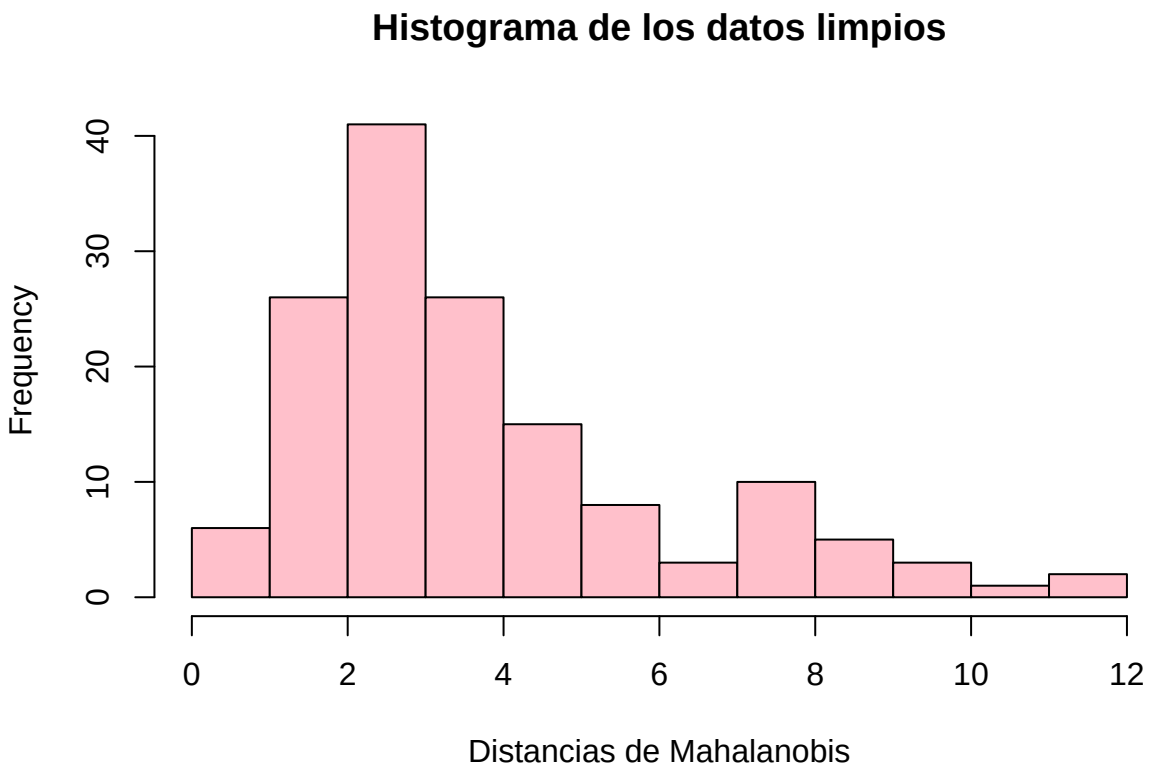
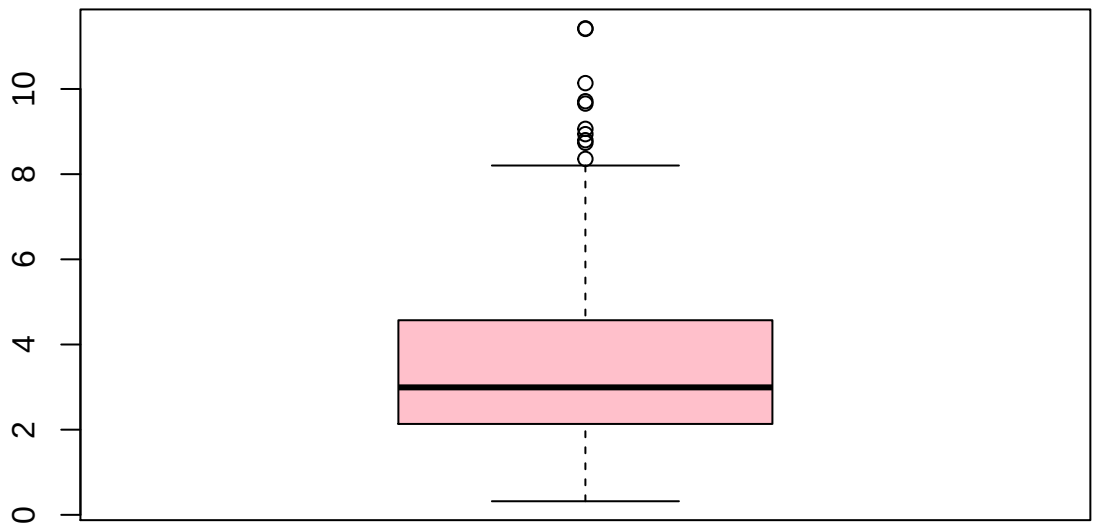


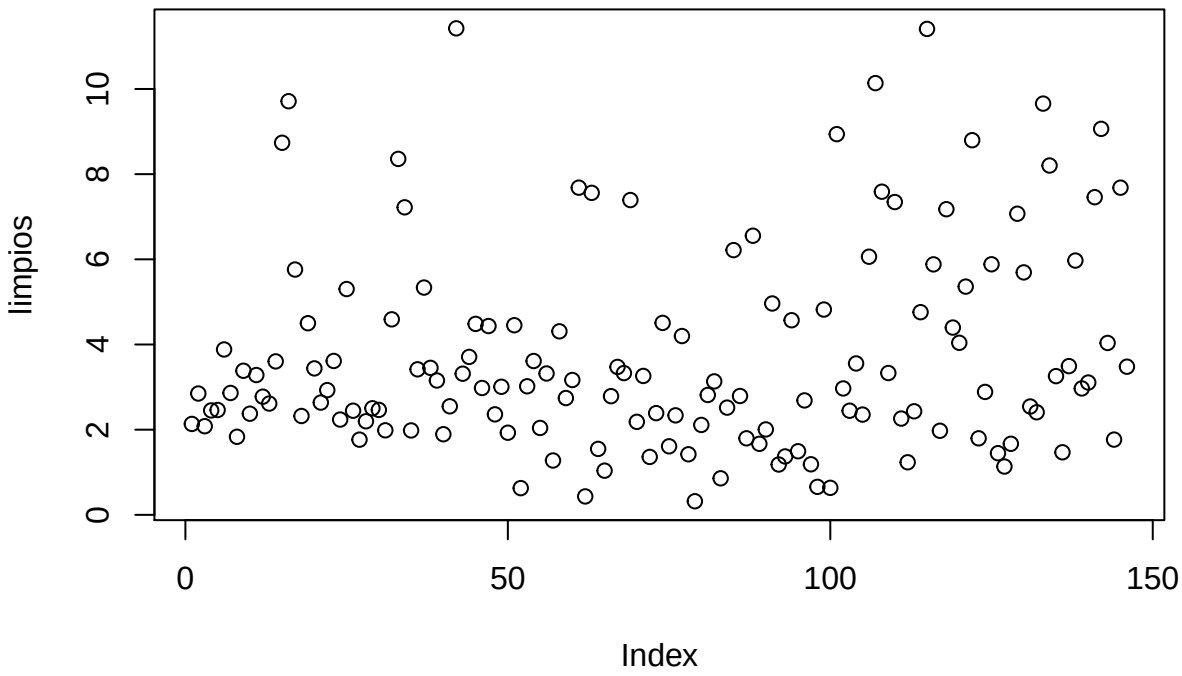
Gráfico de cajas de la nueva base

```
boxplot(limpios, col = "pink")
```



Revisamos el gráfico de dispersion de las distancias de Mahalanobis

```
plot(limpios)
```


limpios
0
2
4
6
8
10
0
50
100
150
Index