Práctica 1 Estadística II

Alberto Parramón Castillo

Introducimos en una variable los datos de la tabla Iris. Sólo las 50 primeras filas, menos la quinta columna: longitud del sépalo - anchura del sépalo - longitud del pétalo - anchura del pétalo

```
datos <- iris[1:50,-5]
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
```

Ejercicio 1

Calcula el vector de medias muestral y las matrices de covarianzas y de correlaciones (cor) muestrales. ¿Entre qué par de variables es más alta la correlación? ¿Qué variable tiene la mayor varianza?

A) Vector de medias:

```
mediasIris <- colMeans(datos)
mediasIris
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##        5.006        3.428        1.462        0.246
```

B) Matriz de covarianzas:

```
covIris <- cov(datos)
covIris
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length   0.12424898 0.099216327  0.016355102 0.010330612
## Sepal.Width    0.09921633 0.143689796  0.011697959 0.009297959
## Petal.Length   0.01635510 0.011697959  0.030159184 0.006069388
## Petal.Width    0.01033061 0.009297959  0.006069388 0.011106122
```

C) Matriz de correlaciones:

```
corIris <- cor(datos)
corIris
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000   0.7425467    0.2671758   0.2780984
## Sepal.Width     0.7425467   1.0000000    0.1777000   0.2327520
## Petal.Length    0.2671758   0.1777000    1.0000000   0.3316300
## Petal.Width     0.2780984   0.2327520    0.3316300   1.0000000
```

D) ¿Entre qué par de variables es más alta la correlación?

Entre longitud de sepalos y anchura de sepalos: 0.7425467

E) ¿Qué variable tiene la mayor varianza?

La anchura de sepalos

Ejercicio 2

Calcula las distancias de Mahalanobis entre cada uno de los lirios y el vector de medias. Representa los datos, usando el color rojo para el 25 % de los lirios más lejanos al vector de medias.

A) Utilizamos la función de Mahalanobis con parámetros: los datos, el vector de medias, y la matriz de covarianzas:

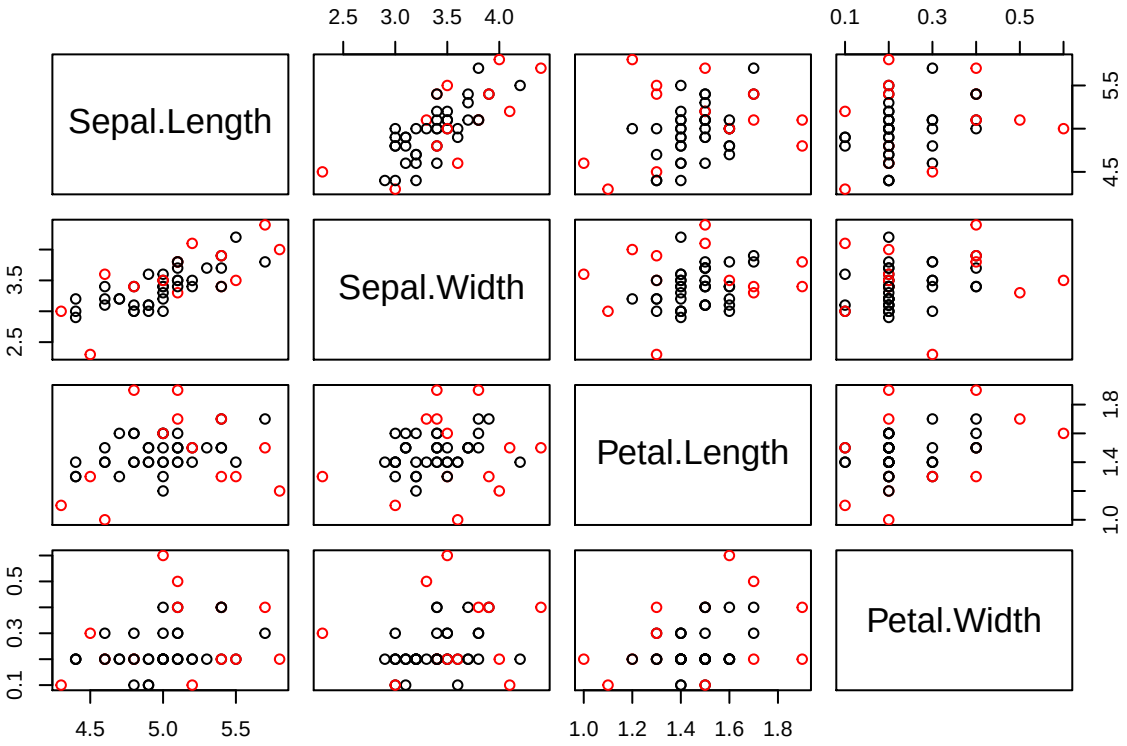
```
distancias <- mahalanobis(datos, mediasIris, covIris)
```

B) Utilizamos la función *summary*, que nos devuelve un vector cuyo quinto elemento es el tercer cuartil de los datos que le hayas pasado por argumento, en este caso las distancias.

```
cuartil3 <- summary(distancias)[5]
```

Creamos el vector de colores y pintamos con plot:

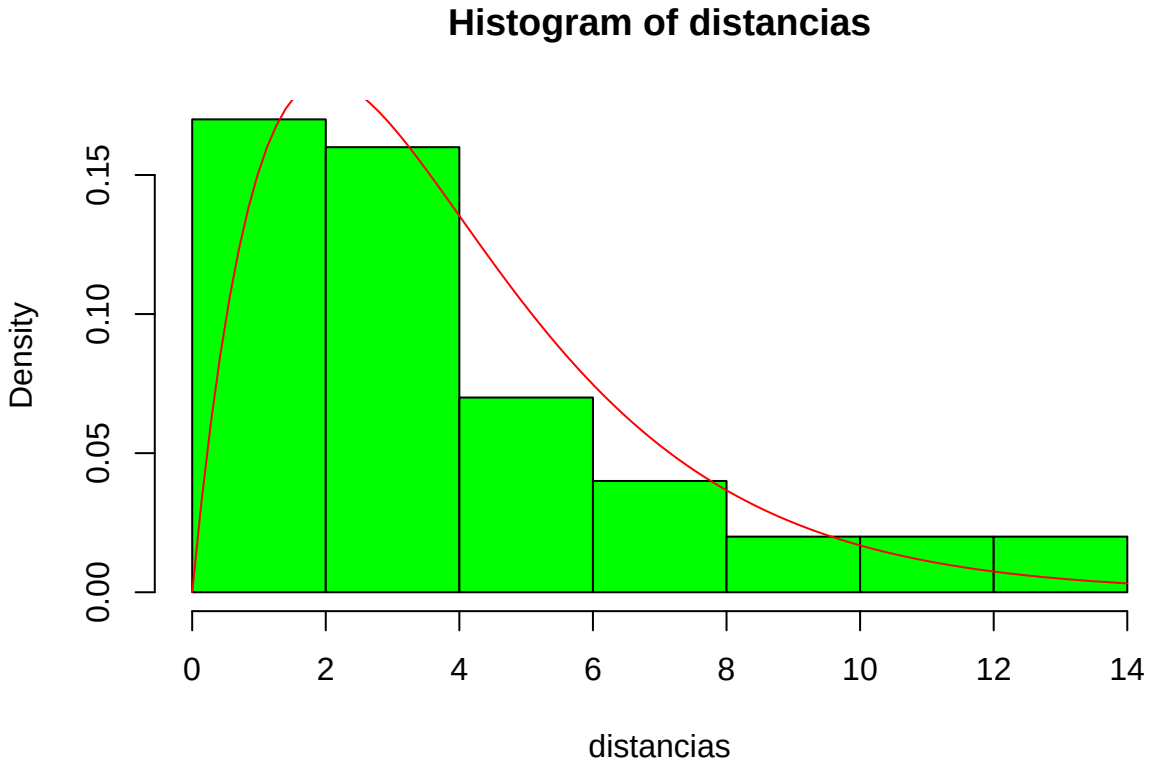
```
colores <- vector('character', length=50)
for(i in 1:50){
  if(distancias[i]>cuartil3){
    colores[i] <- 'red'
  }else{
    colores[i] <- 'black'
  }
}
pairs(datos, col=colores)
```



Ejercicio 3

Representa un histograma de las distancias y compáralo con la función de densidad de una variable χ^2 con 4 grados de libertad.

```
hist(distancias, col = "green", breaks = 8, freq=FALSE)
curve( dchisq(x, df=4), col='red', add=TRUE)
```



Ejercicio 4

Genera 100 observaciones con distribución normal bidimensional con vector de medias el origen y matriz de covarianzas:

$$\Sigma = \begin{pmatrix} 10 & 3 \\ 3 & 1 \end{pmatrix}$$

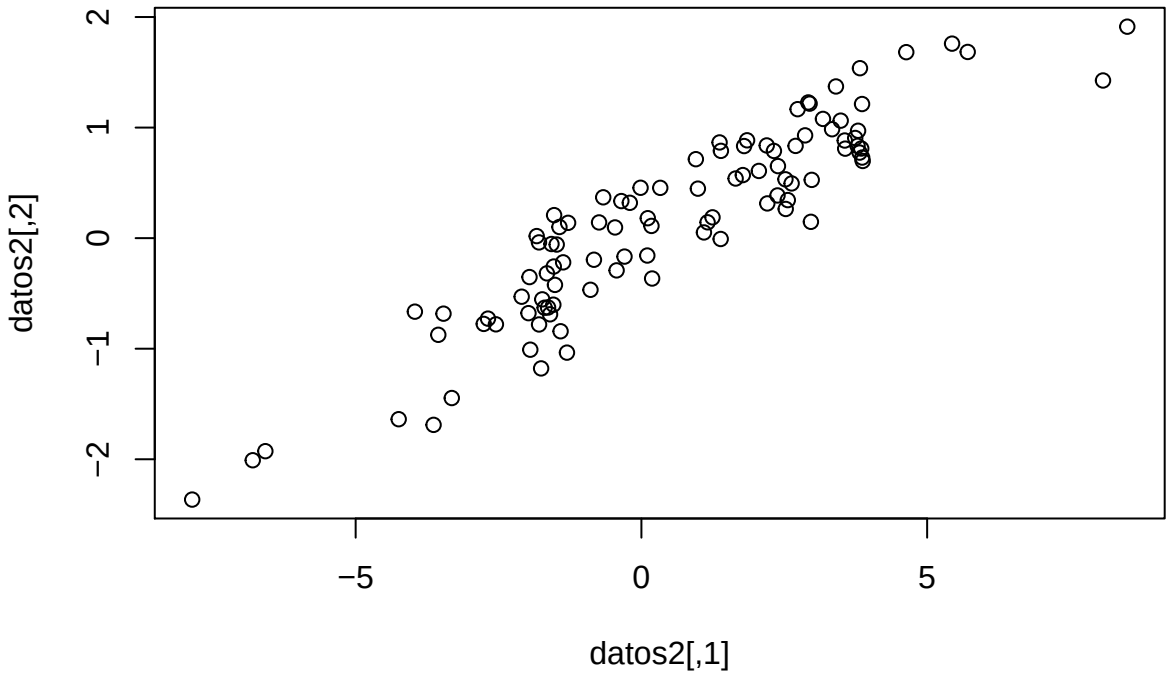
Representa la nube de puntos generados, su vector de medias y su matriz de covarianzas.

A) Obtenemos las 100 observaciones a partir de los datos del enunciado, siendo *mu* el vector de medias, *sigma* la matriz de covarianzas y n el número de observaciones:

```
set.seed(9111) #Esto establece una semilla para que siempre salgan los mismos datos aleatorios
library(MASS) #paquete necesario
n <-100
mu <- c(0,0)
sigma <-matrix(c(10,3,3,1),2)
datos2 <- mvrnorm(n,mu,sigma)
```

Representamos la nube de puntos:

```
plot(datos2)
```



B) Calculamos y representamos su vector de medias obtenido con los datos generados

```
medias = colMeans(datos2)
medias
```

```
## [1] 0.5300716 0.1524980
```

C) Calculamos y representamos la matriz de covarianza obtenida con los datos generados

```
covarianza = cov(datos2)
covarianza
```

```
##          [,1]      [,2]
## [1,] 8.669332 2.3585971
## [2,] 2.358597 0.7536912
```

Ejercicio 5

Para la misma distribución del apartado anterior, calcula el valor esperado teórico de la segunda coordenada respecto de la primera. Si no lo conocieras y solo dispusieras de los datos generados. ¿Cómo lo estimarías? Calcula el valor resultante para el estimador que has propuesto.

Si suponemos que queremos calcular el valor esperado de $X_2|X_1$. Utilizaremos las siguientes fórmulas generales.

$$\mu_{2.1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1)$$

$$\Sigma_{2.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

A) Valor esperado teórico para $X_2|X_1$, tenemos el vector de medias y la matriz de covarianzas siguiente:

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 10 & 3 \\ 3 & 1 \end{pmatrix}$$

Obtenemos:

$$\mu_{2.1} = 0 + \frac{3}{10}(X_1)$$
$$\Sigma_{2.1} = 1 - \frac{3}{10}3 = \frac{1}{10}$$

B) Valor esperado estimado a partir de las observaciones para $X_2|X_1$, tenemos el vector de medias y la matriz de covarianzas siguiente:

$$\mu = \begin{pmatrix} 0.53 \\ 0.15 \end{pmatrix}, \Sigma = \begin{pmatrix} 8.66 & 2.35 \\ 2.35 & 0.75 \end{pmatrix}$$

Obtenemos:

$$\mu_{2.1} = 0.15 + \frac{2.35}{8.66}(X_1 - 0.53) = 0.006 + 0.27X_1$$
$$\Sigma_{2.1} = 0.75 - \frac{2.35}{8.66}2.35 = 0.11$$