# Actividad 1.5 Combinaciones lineales y Validación de la Normal Multivariada

Franco Mendoza Muraira A01383399

2023-11-10

## 1. Matriz X

```r
X = matrix(c(1,4,3,6,2,6,8,3,3),nrow=3,ncol=3,byrow=TRUE)
X
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    3
## [2,]    6    2    6
## [3,]    8    3    3
```

$b'X = X_1 + X_2 + X_3$

$c'X = X_1 + 2X_2 - 3X_3$

### a) Hallar la media, varianza y covarianza de X

```
## Media de X: 5 3 4
```

```
##
##
## Varianza de X:
```

```
##      [,1] [,2] [,3]
## [1,] 13.0 -2.5  1.5
## [2,] -2.5  1.0 -1.5
## [3,]  1.5 -1.5  3.0
```

```
##
##
## Covarianza de X:
```

```
##      [,1] [,2] [,3]
## [1,] 13.0 -2.5  1.5
## [2,] -2.5  1.0 -1.5
## [3,]  1.5 -1.5  3.0
```

## b) Hallar la media, varianza y covarianza de b'X y c'X

### Media

```
bc = matrix(c(1,1,1,1,2,-3),nrow=2,ncol=3,byrow=TRUE)
mediabc = bc%*%mediaX
mediabc
```

```
##      [,1]
## [1,]   12
## [2,]   -1
```

### Matriz de Varianzas y covarianzas

```
varianzabc = bc%*%varianzaX%*%t(bc)
varianzabc
```

```
##      [,1] [,2]
## [1,]   12   -3
## [2,]   -3   43
```

## c) Hallar el determinante de S

```
## c) Determinante de S (matriz de var-covarianzas de X):  0
```

## d) Hallar los valores y vectores propios de S

```
## d) Valores propios de S:
##  13.7915 3.208497 -7.859007e-17
```

```
##
##
## d) Vectores propios de S:
```

```
##            [,1]       [,2]       [,3]
## [1,]  0.9645458 -0.2295697 -0.1301889
## [2,] -0.2076189 -0.3555080 -0.9113224
## [3,]  0.1629288  0.9060418 -0.3905667
```

## e) Argumentar si hay independencia entre b'X y c'X , ¿y qué ocurre con X1, X2 y X3? ¿son independientes?

```
bx = matrix(c(1,1,1),nrow=1,ncol=3)
cx=matrix(c(1,2,-3),nrow=1,ncol=3)
cor.test(X[,1],X[,2])
```

```
##
##  Pearson's product-moment correlation
##
## data:  X[, 1] and X[, 2]
## t = -0.96225, df = 1, p-value = 0.5122
## alternative hypothesis: true correlation is not equal to 0
## sample estimates:
##        cor
## -0.6933752
```

```r
cor.test(X[,1],X[,3])
```

```
##
##  Pearson's product-moment correlation
##
## data:  X[, 1] and X[, 3]
## t = 0.24744, df = 1, p-value = 0.8456
## alternative hypothesis: true correlation is not equal to 0
## sample estimates:
##       cor
## 0.2401922
```

```r
cor.test(X[,2],X[,3])
```

```
##
##  Pearson's product-moment correlation
##
## data:  X[, 2] and X[, 3]
## t = -1.7321, df = 1, p-value = 0.3333
## alternative hypothesis: true correlation is not equal to 0
## sample estimates:
##        cor
## -0.8660254
```

$H_0$ : Correlación $= 0$

$H_1$ : Correlación $\neq 0$

Debido a que los p valores de todas las combinaciones de columnas en la matriz X son mayor a el nivel de significancia de 0.05, no podemos rechazar la hipótesis nula, por lo que se concluye que la correlación entre las variables es $= 0$. Como la matriz original es independiente, se sabe que todas las combinaciones serán independientes.

## Hallar la varianza generalizada de S. Explicar el comportamiento de los datos de X basándose en los la varianza generalizada, en los valores y vectores propios de S.

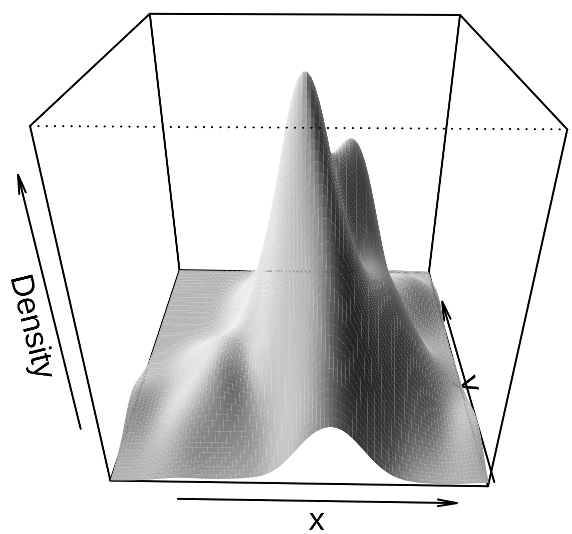
```
## La varianza generalizada de S es: 17
```

## 2. Explore los resultados del siguiente código y dé una interpretación.

```
library(MVN)
```

```
## Warning: package 'MVN' was built under R version 4.2.3
```

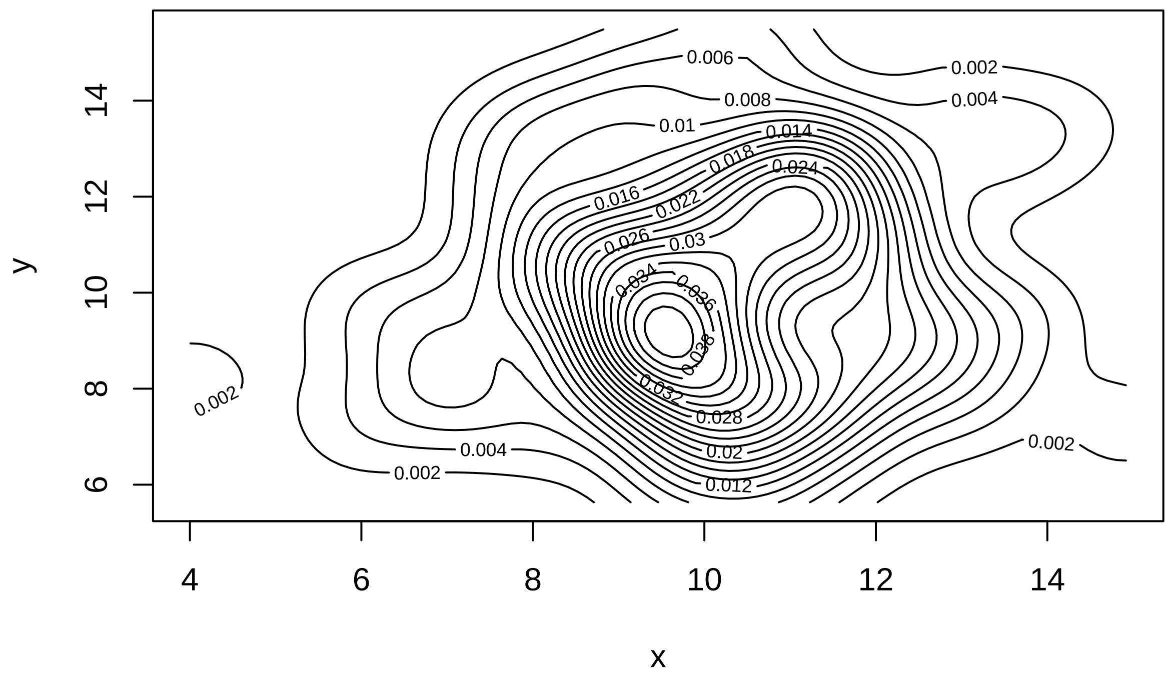
```
x = rnorm(100, 10, 2)
y = rnorm(100, 10, 2)
datos = data.frame(x,y)
mvn(datos, mvnTest = "hz", multivariatePlot = "persp")
```




```
## $multivariateNormality
##            Test       HZ  p value MVN
## 1 Henze-Zirkler 0.572713 0.442751 YES
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling     x        0.4542    0.2645    YES
## 2 Anderson-Darling     y        0.4675    0.2455    YES
##
## $Descriptives
##     n      Mean   Std.Dev    Median      Min      Max      25th      75th        Skew
```

```
## x 100 10.13709 1.794613 10.16735 4.006121 14.91636 9.263648 11.29666 -0.2891665
## y 100 10.00203 2.140028  9.80975 5.633720 15.48489 8.381147 11.37678  0.2730528
##     Kurtosis
## x  0.9095614
## y -0.6286527
```

```r
mvn(datos, mvnTest = "hz", multivariatePlot = "contour")
```




```
## $multivariateNormality
##           Test       HZ  p value MVN
## 1 Henze-Zirkler 0.572713 0.442751 YES
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling     x        0.4542    0.2645    YES
## 2 Anderson-Darling     y        0.4675    0.2455    YES
##
## $Descriptives
##     n     Mean  Std.Dev   Median      Min      Max     25th     75th       Skew
## x 100 10.13709 1.794613 10.16735 4.006121 14.91636 9.263648 11.29666 -0.2891665
## y 100 10.00203 2.140028  9.80975 5.633720 15.48489 8.381147 11.37678  0.2730528
##     Kurtosis
## x  0.9095614
## y -0.6286527
```

### Henze-Zirkler

$H_0$ : Los datos siguen una distribución normal multivariada.

$H_1$ : Los datos no siguen una distribución normal multivariada.

Debido a el alto valor p de los tests de Henze Zirkler, con un valor de 0.6316, no podemos rechazar la hipótesis nula, y se concluye qye los datos usados tienen una distribución normal multivariada.

### Anderson-Darling

$H_0$ : Los datos siguen una distribución normal univariada.

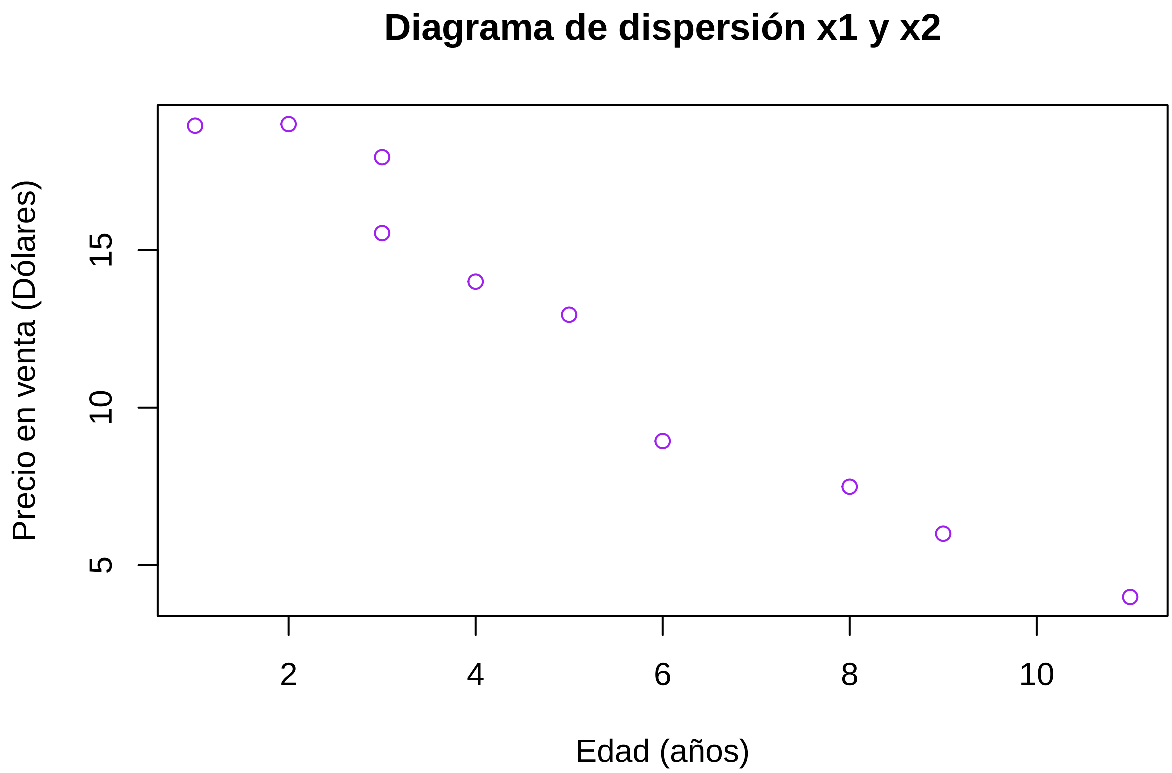
$H_1$ : Los datos no siguen una distribución normal univariada.

Viendo que los pvalue de la prueba de Anderson-Darling es de 0.5405 en la primer variable, y 0.9752 en la segunda variable, y siendo que son mayores que el $\alpha$ de 0.05, no podemos rechazar la hipótesis nula, y se concluye que se presenta una distribución normal univariada en las 2 variables de los datos.

## 3. Un periódico matutino enumera los siguientes precios de autos usados para un compacto extranjero con edad medida en años y precio en venta medido en miles de dólares.

### a) Diagrama de dispersión

```
x1 =c(1, 2, 3, 3, 4, 5, 6, 8, 9, 11)
x2=c(18.95, 19.00, 17.95, 15.54, 14.00, 12.95, 8.94, 7.49, 6.00, 3.99)
plot(x1,x2,col="purple",main="Diagrama de dispersión x1 y x2",xlab = "Edad (años)",ylab = "Precio en ve
```




### b)Inferir el signo de la covarianza a partir del gráfico

Viendo la gráfica se puede inferir que el signo de la covarianza muestral es negativo, esto porque las variables son inversamente proporcional entre ellas.

### c) Calcular el cuadrado de las distancias de Mahalanobis

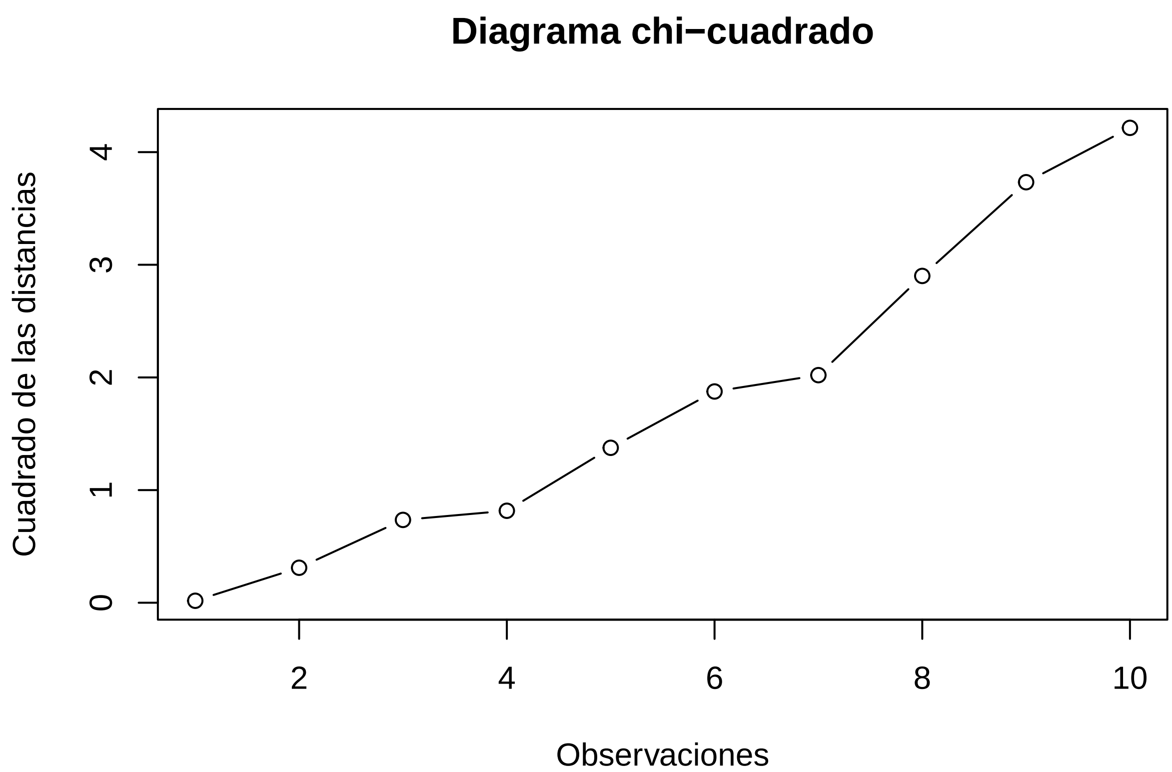
```
mx = cbind(x1,x2)
meanx = colMeans(mx)
varx=var(mx)
M = mahalanobis(mx,meanx,varx)
M
```

```
##  [1] 1.8753045 2.0203262 2.9009088 0.7352659 0.3105192 0.0176162 3.7329012
##  [8] 0.8165401 1.3753379 4.2152799
```

### d) Usando las anteriores distancias, determine la proporción de las observaciones que caen dentro del contorno de probabilidad estimado del 50% de una distribución normal bivariada.

```
## Proporción de datos: 0.5
```

**e) Ordene el cuadrado de las distancias del inciso c y construya un diagrama chi-cuadrado**




**f) Dados los resultados anteriores, serían argumentos para decir que son aproximadamente normales bivariados?**

Sí, acorde a como se distribuyen los datos, y como se ve la recta se puede creer que son aproximadamente normales bivariadas, aunque no siempre es seguro. Para asegurar esto, se tiene que hacer una prueba de shapiro:

```
shapiro.test(mx)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mx
## W = 0.92094, p-value = 0.1033
```

```
shapiro.test(M)
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  M
## W = 0.93759, p-value = 0.5265
```

$H_0$ : Los datos siguen una distribución normal.

$H_1$ : Los datos no siguen una distribución normal.

Con como se ven estas pruebas, y debido a que ambas pruebas nos dan valores de p mayores que 0.05, no se tiene suficiente evidencia para rechazar la hipótesis nula, y se concluye que los datos se distribuyen de forma normal, al igual que las distancias cuadradas.