UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

FACULTAD DE CIENCIAS FISÍCO MATEMÁTICAS

Maestría en Ciencia de Datos

Materia

Métodos Estadísticos Multivariados

Profesora

M.E.T. Rosa Isela Hernández Zamora

Tarea 3

Alumna

I.M. María Luisa Argáez Salcido

Matrícula

2173261

Fecha

06 de febrero de 2023

Tarea 3 de Métodos Estadísticos Multivariados

Instrucciones: Contesta cada uno de los ejercicios en un archivo en Word o en hojas blancas. Puedes usar R o Excel, favor de anexar el código de R o el archivo de Excel. Al finalizar sube tus evidencias en el lugar correspondiente en Teams en formato PDF.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats
```

Ejercicio 1.-

Considera un vector aleatorio X con distribución $N_5(\mu, \Sigma)$ donde:

$$\mu' = (100, 95, 230, 400, 86)$$

y

$$\Sigma = \begin{bmatrix} 10 & -2 & 1 & 0 & 3 \\ -2 & 9 & -3 & 4 & 5 \\ 1 & -3 & 15 & 7 & -2 \\ 0 & 4 & 7 & 20 & 2 \\ 3 & 5 & -2 & 2 & 5 \end{bmatrix}$$

Procedimiento 1 a)

a) Obtener $P(90 < X_2 < 100)$

- Se declara que $X_2 \sim N(95, 9)$.
- $z = \frac{x-\mu}{\sigma}$ ...(1)
- Se separa el ejercicio en dos condiciones: $1.P(90 < X_2)$ y $2.P(X_2 < 100)$

1. $P(90 < X_2)$ de aqui se conoce que x = 90 , μ_2 = 90 y Σ_2 = 3. , al sustituir en (1), se conoce que $z = \frac{90-95}{3} = -1.66$ y $P(z < -1.66) = 0.0484$
2. $P(X_2 < 100)$ de aqui se conoce que x = 100 , μ_2 = 90 y Σ_2 = 3. , al sustituir en (1), se conoce que $z = \frac{100-95}{3} = 1.66$ y $P(z < -1.66) = 0.9515$

Por último, se realiza la resta de 0.9515 - 0.0484 = 0.9031.

Solución 1 a)

Por tanto, $P(90 < X_2 < 100) = 0.9031$

Procedimiento 1 b)

b) Obtener $P(3X_1 + 4X_3 - 5X_5 > 800)$

Si $P(3X_1 + 4X_3 - 5X_5 > 800)$ entonces:

- $3X_1 + 4X_3 - 5X_5 \sim N((100, 230, 86), (10, 15, 5))$
- $E(3X_1 + 4X_3 - 5X_5) = 3E(X_1) + 4E(X_3) - 5E(x_5) = 3(100) + 4(230) - 5(86) = 790$
- $Var(y) = A\Sigma_x A^T$ ...(2)
- Al sustituir en (2) $Var(3X_1 + 4X_3 - 5X_5)$,se obtiene:

```
coef_1 = np.matrix([[3,0,4,0,-5]])
sigma_x = np.matrix([[10,-2,1,0,3],
                    [-2,9,-3,4,5],
                    [1,-3,15,7,-2],
                    [0,4,7,20,2],
                    [3,5,-2,2,5]])
```

```
var_y = coef_1 * sigma_x * coef_1.T
var_y
```

```
matrix([[469]])
```

- Sustituyendo en (1) , se obtiene $z = \frac{800-790}{\sqrt{469}} = 0.4617$
- Entonces, $P(0.4617 > Z) = 0.3221$

Solución 1 b)

- Entonces, $P(0.4617 > Z) = 0.3221$

Procedimiento 1 c)

Sea el vector aleatorio

$$Y = \begin{bmatrix} X_1 + 3X_2 - 4X_3 + 6X_4 + X_5 \\ 2X_1 + 9X_2 - 10X_3 + X_4 - X_5 \\ X_2 + X_4 - X_5 \end{bmatrix}$$

- Se observa que $Y = AX$.
- Con distribución $N_5(\mu_y \Sigma_y)$.
- $\mu' = (100, 95, 230, 400, 86)$
- $\mu'_y = A\mu_x$

```
coef_y = np.matrix([[1,3,-4,6,1],
                    [2,9,-10,1,-1],
                    [0,1,0,1,-1]])
mu_x = np.matrix([100,95,230,400,86])
mu_y = coef_y*mu_x.T
mu_y
```

```
matrix([[1951],
        [-931],
        [ 409]])
```

- $\Sigma_y = A\Sigma_x A'$

```
sigma_x = np.matrix([[10,-2,1,0,3],
                    [-2,9,-3,4,5],
                    [1,-3,15,7,-2],
                    [0,4,7,20,2],
                    [3,5,-2,2,5]])
sigma_y = coef_y * sigma_x * coef_y.T
sigma_y
```

```
matrix([[ 992,  943,  129],
        [ 943, 2508,   22],
        [ 129,   22,   28]])
```

Solución 1 c)

- $\mu'_y = A\mu_x$

```
mu_y
```

```
matrix([[1951],
        [-931],
        [ 409]])
```

- $\Sigma_y = A\Sigma_x A'$

```
sigma_y
```

```
matrix([[ 992,  943,  129],
        [ 943, 2508,   22],
        [ 129,   22,   28]])
```

Procedimiento 1 d)

Obtener la distancia estadística de $X_1'(110, 97, 230, 396, 85)$ a $X_2'(96, 93, 237, 408, 90)$

Se conoce que la distancia estadística esta dada por:

$$d^2(x_1, x_2) = (x_1 - x_2)'\Sigma^{-1}(x_1 - x_2)$$

```
sigma_x
```

```
matrix([[10, -2,  1,  0,  3],
        [-2,  9, -3,  4,  5],
        [ 1, -3, 15,  7, -2],
        [ 0,  4,  7, 20,  2],
        [ 3,  5, -2,  2,  5]])
```

```
x_1 = np.matrix([110,97,230,396,85]).T
x_2 =  np.matrix([96,93,237,408,90]).T
dif_x1_x2 = x_1 - x_2
distancia_cuad = dif_x1_x2.T * np.linalg.inv(sigma_x) * dif_x1_x2
distancia_x1_x2 = np.sqrt(distancia_cuad)
distancia_x1_x2
```

```
matrix([[22.40339456]])
```

Solución 1 d)

La distancia estadística esta dada por:

```
distancia_x1_x2
```

```
matrix([[22.40339456]])
```

Procedimiento 1 e)

Indicar que componentes de X son independientes

```
sigma_x
```

```
matrix([[10, -2,  1,  0,  3],
        [-2,  9, -3,  4,  5],
        [ 1, -3, 15,  7, -2],
        [ 0,  4,  7, 20,  2],
        [ 3,  5, -2,  2,  5]])
```

Dado que se estipulo que la covarianza cero implica que los componentes correspondientes son independientes. Se observa que x_1 y x_4 son independientes ya que tienen covarianza igual con cero.

Solución 1 e)

Componentes x_1 y x_4

Ejercicio 2. -

Considerando el problema 1, se toma una muestra aleatoria de tamaño 40.

Procedimiento 2 a)

Obtener $P(\bar{X}_3 < 229)$

```
x =229
x_3_mean = 230
sigma_x3 = np.sqrt(1/40 * 15)
from statistics import NormalDist
```

```

NormalDist(mu=x_3_mean, sigma=sigma_x3).cdf(x)
```

0.05123521742987469

Solución 2 a)

La $P(\bar{X}_3 < 229)$ es :

```
NormalDist(mu=x_3_mean, sigma=sigma_x3).cdf(x)
```

```
0.05123521742987469
```

Procedimiento 2 b)

Obtener $P(4\bar{X}_1 + 3\bar{X}_3 - \bar{X}_4 < 687)$

- Se conoce que $\Sigma_z = Cov(z) = Cov(CX) = C\Sigma_x C'$, si $z = 4\bar{X}_1 + 3\bar{X}_3 - \bar{X}_4$, entonces

```
coef_z  = np.matrix([4,0,3,-1,0])
sigma_z = coef_z * sigma_x * coef_z.T
sigma_z

```

```
matrix([[297]])
```

- Se conoce que $E(z) = 4E(\bar{X}_1) + 3E(\bar{X}_3) - E(\bar{X}_4) = 4(100) + 3(230) - 400 = 690$

```
mu_z = (4*100) + (3*230) -400
var_z = np.sqrt(sigma_z)
NormalDist(mu=mu_z, sigma=var_z).cdf(687)
```

```
0.43090221652450544
```

Solución 2 b)

Por tanto, $P(4\bar{X}_1 + 3\bar{X}_3 - \bar{X}_4 < 687) = 0.4309$

```
NormalDist(mu=mu_z, sigma=var_z).cdf(687)
```

```
0.43090221652450544
```

Procedimiento 2 c)

Obtener la distancia estadística de $\bar{X}' = (99.5, 96, 231, 400, 86.2)$ a μ

```
x_3 = np.matrix([99.5,96,231,400,86.2])
dif_x3_mux = (x_3 - mu_x ).T
dist_estad =np.sqrt( dif_x3_mux.T * np.linalg.inv(sigma_x) * dif_x3_mux)
dist_estad
```

```
matrix([[0.6874447]])
```

Solución 2 c)

Por tanto, la distancia estadística de $\bar{X}' = (99.5, 96, 231, 400, 86.2)$ a μ es

```
dist_estad
```

```
matrix([[0.6874447]])
```

Procedimiento 2 d)

Obtener la distribución de:

$$W = \begin{bmatrix} \bar{X}_1 + 2\bar{X}_3 - \bar{X}_4 \\ \bar{X}_2 + \bar{X}_5 \end{bmatrix}$$

```
coef_w = np.matrix([[1,0,2,-1,0],
                    [0,1,0,0,1]])
mu_y2 = coef_w * mu_x.T
mu_y2
```

```
matrix([[160],
        [181]])
```

```
sigma_y2 = coef_w * sigma_x * coef_w.T
sigma_y2
```

```
matrix([[ 66, -15],
        [-15,  24]])
```

Solución 2 d)

La Distribución de W esta dada por:

```
mu_y2
```

```
matrix([[160],
        [181]])
```

```
sigma_y2
```

```
matrix([[ 66, -15],
        [-15,  24]])
```

```

```

```r
library(MVN)
library(nortest)
```

Lectura de datos

```r
path = "C:/Users/Maria Luisa/OneDrive/Documentos/MasterDataScience/MEM/datostarea3.csv"
data<- read.csv(path)
data
```

```
##       X1    X2     X3     X4
## 1  79.76 42.00 104.02 124.18
## 2  83.41 39.46 101.32 117.34
## 3  80.41 39.72  99.83 119.47
## 4  82.94 43.16 104.23 123.88
## 5  83.75 43.81 105.23 126.46
## 6  79.45 44.01 104.64 125.49
## 7  79.04 41.79 102.72 122.48
## 8  75.54 39.89 100.06 125.42
## 9  81.86 39.34  99.24 117.18
## 10 82.97 43.60 104.81 123.44
## 11 80.09 38.59  96.91 118.25
## 12 79.89 40.15  98.87 120.10
## 13 82.64 38.60  97.61 112.78
## 14 76.95 37.69 100.04 118.81
## 15 83.85 40.48  99.88 119.61
## 16 77.65 38.53  99.95 118.25
## 17 79.25 39.02  98.59 118.55
## 18 77.69 38.02  97.69 118.58
## 19 78.88 39.00 101.96 117.00
## 20 75.22 39.81 100.86 121.24
## 21 80.15 39.41  99.00 118.59
## 22 83.94 42.69 101.65 122.60
## 23 80.00 42.67 101.54 123.52
## 24 83.13 40.16  99.63 117.82
## 25 86.95 41.49  97.67 121.29
## 26 75.80 38.74 100.89 117.57
## 27 84.75 42.21 101.03 119.13
## 28 78.30 41.71 100.05 124.46
## 29 83.20 39.68 101.52 118.79
## 30 82.58 42.91 103.09 126.39
## 31 76.47 35.80  98.03 113.33
## 32 80.08 38.95  98.74 119.86
## 33 80.86 40.13  98.63 120.82
## 34 85.99 45.17 107.22 124.49
## 35 75.66 36.78  97.59 117.76
## 36 75.62 40.03 102.62 123.14
## 37 83.21 42.25 101.44 119.87
## 38 76.07 40.73 101.02 125.52
## 39 79.71 41.09 102.12 124.76
## 40 76.88 38.21  99.90 116.58
## 41 76.23 37.93  99.74 121.09
```

```
## 42 80.20 40.02  99.96 120.99
## 43 75.33 39.30  99.19 121.55
## 44 79.30 42.81 106.11 130.32
## 45 75.75 39.37 100.64 121.49
## 46 81.21 39.05  97.96 117.26
## 47 74.22 40.26  99.65 121.47
## 48 76.91 36.44  94.87 112.84
## 49 79.47 39.21 100.29 118.49
## 50 84.64 40.33  99.66 119.05
```

Descriptivos de datos

```
summary(data)
```

```
##        X1              X2              X3               X4
##  Min.   :74.22   Min.   :35.80   Min.   : 94.87   Min.   :112.8
##  1st Qu.:76.92   1st Qu.:39.01   1st Qu.: 99.05   1st Qu.:118.3
##  Median :79.83   Median :39.95   Median :100.05   Median :120.0
##  Mean   :79.88   Mean   :40.24   Mean   :100.60   Mean   :120.6
##  3rd Qu.:82.86   3rd Qu.:41.77   3rd Qu.:101.62   3rd Qu.:123.4
##  Max.   :86.95   Max.   :45.17   Max.   :107.22   Max.   :130.3
```

Se observa que la media y la mediana de las cuatro variables en cuestión son cercanas.

Calculo del vector de medias, covarianza, correlación y distancia de Mahalanobis

Se calcula el vector de medias, la covarianza, correlación y la distancia de Mahalanobis.

```
mean_vector <- colMeans(data)
cov <- cov(data)
correlation <- cor(data)
distancia <- mahalanobis(data, mean_vector, cov)

print("Vector de medias")
```

```
## [1] "Vector de medias"
```

```
print( mean_vector )
```

```
##       X1       X2       X3       X4
##  79.8770  40.2440 100.5982 120.5870
```

```
print("\n")
```

```
## [1] "\n"
```

```r
print("Covarianza")
```

```
## [1] "Covarianza"
```

```r
print( cov )
```

```
##            X1       X2       X3         X4
## X1 10.4284214 3.604590 2.088756  0.6677827
## X2  3.6045898 4.171776 4.081607  5.6929673
## X3  2.0887557 4.081607 6.138733  6.5843067
## X4  0.6677827 5.692967 6.584307 13.1265602
```

```r
print("\n")
```

```
## [1] "\n"
```

```r
print("Correlación")
```

```
## [1] "Correlación"
```

```r
print( correlation )
```

```
##            X1        X2        X3         X4
## X1 1.00000000 0.5464949 0.2610592 0.05707559
## X2 0.54649488 1.0000000 0.8065501 0.76931245
## X3 0.26105918 0.8065501 1.0000000 0.73349169
## X4 0.05707559 0.7693124 0.7334917 1.00000000
```

```r
print("\n")
```

```
## [1] "\n"
```

```r
print("Distancia de Mahalanobis")
```

```
## [1] "Distancia de Mahalanobis"
```

```r
print( distancia )
```

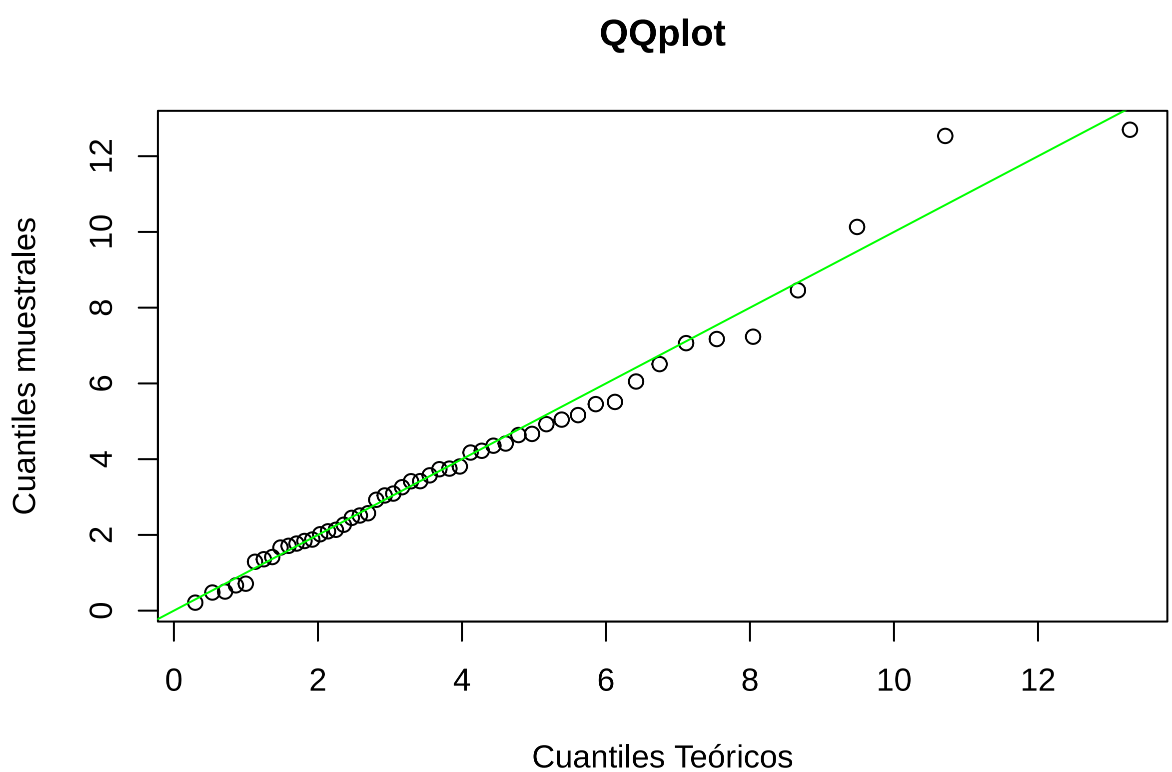
```
##  [1]  2.1349158  5.4532117  0.2116344  2.5719903  4.6672058  7.2335813
##  [7]  2.0172882  6.0508637  1.3577012  4.2216541  3.0416373  1.4160309
## [13]  6.5100188  3.8069507  2.2685610  1.2908890  0.6682071  1.8394120
## [19]  5.5112933  3.0912066  0.5042822  2.4500483  4.3559729  1.7121103
## [25] 12.6978543  4.9247648  4.4128614  5.1635944  5.0444965  3.7347456
## [31]  7.0635973  1.8773257  2.0950997 10.1318430  3.5730102  3.4174342
## [37]  3.7507100  4.1747125  2.5121483  3.2593581  4.6376477  0.4790773
## [43]  2.9277733 12.5348525  1.7714507  1.6665517  8.4586139  7.1722754
## [49]  0.7116826  3.4198525
```

Construcción de QQ Plot

Se construye el QQplot

```
x1 <- (1:50 - 0.5) / 50

plot(qchisq((x1), df = 4), sort(distancia) , xlab = "Cuantiles Teóricos",
    ylab = "Cuantiles muestrales", main = "QQplot")
abline(a =0 , b=1 , col="green")
```



Se observa que la mayoria de los datos se ajustan a la linea, sin embargo, a partir del cuartil teórico 6 empiezan a tener una mayor dispersión respecto a la recta.

Prueba de Kolmogorov-Smirnov

Se calcula la prueba de Kolmogorov-Smirnov, en donde se planten las siguientes pruebas de hipótesis:

H_0 : Los datos provienen de una distribución normal

H_1 : Los datos no provienen de una distribución normal

Bajo la premisa de que se rechaza H_0 si $pval < \alpha$ con un $\alpha = 0.05$

```
ks.test(distancia, "pchisq", df=3)
```

```
##
```

```
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  distancia
## D = 0.1971, p-value = 0.0354
## alternative hypothesis: two-sided
```

Como $pval = 0.0354 < \alpha = 0.05$ Entonces se rechaza H_0, los datos no provienen de una distribución normal.

Prueba de Anderson-Darling.

Las prueba de hipótesis son las mismas que en el ejercicio anterior.

H_0 : Los datos provienen de una distribución normal

H_1 : Los datos no provienen de una distribución normal

Bajo la premisa de que se rechaza H_0 si $pval < \alpha$ con un $\alpha = 0.05$

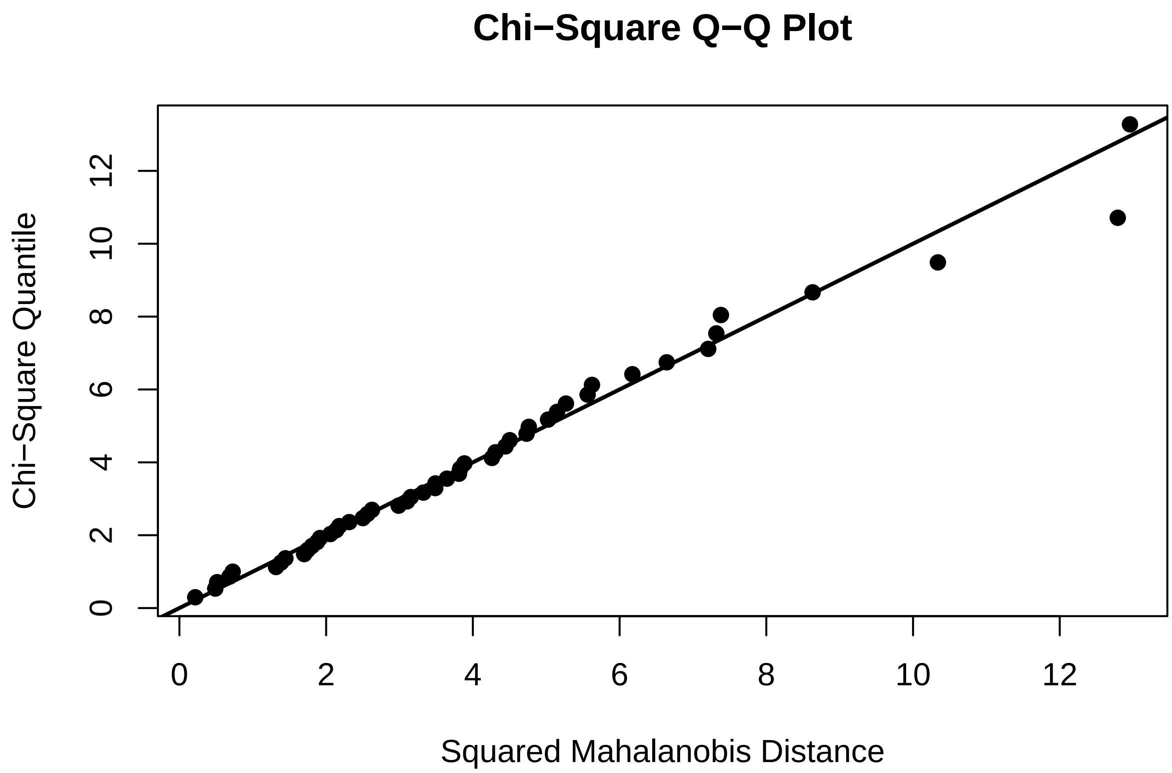
```
ad.test(distancia )
```

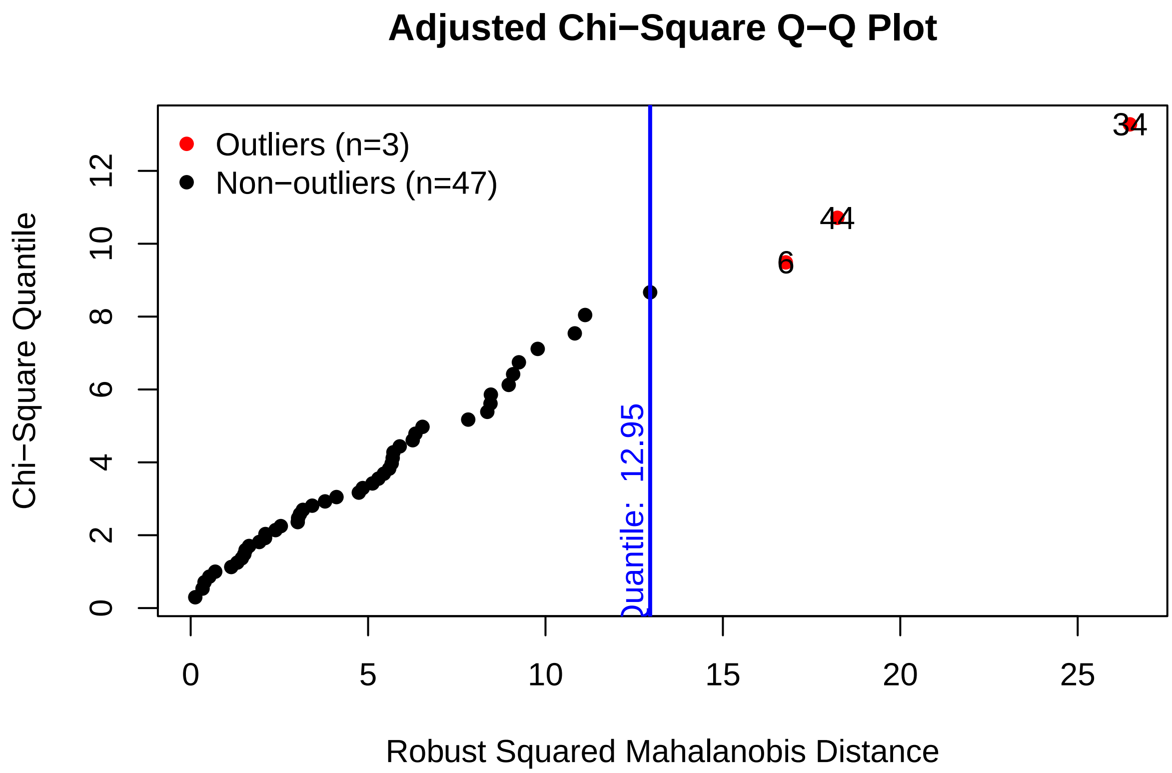
```
##
##  Anderson-Darling normality test
##
## data:  distancia
## A = 1.4118, p-value = 0.001053
```

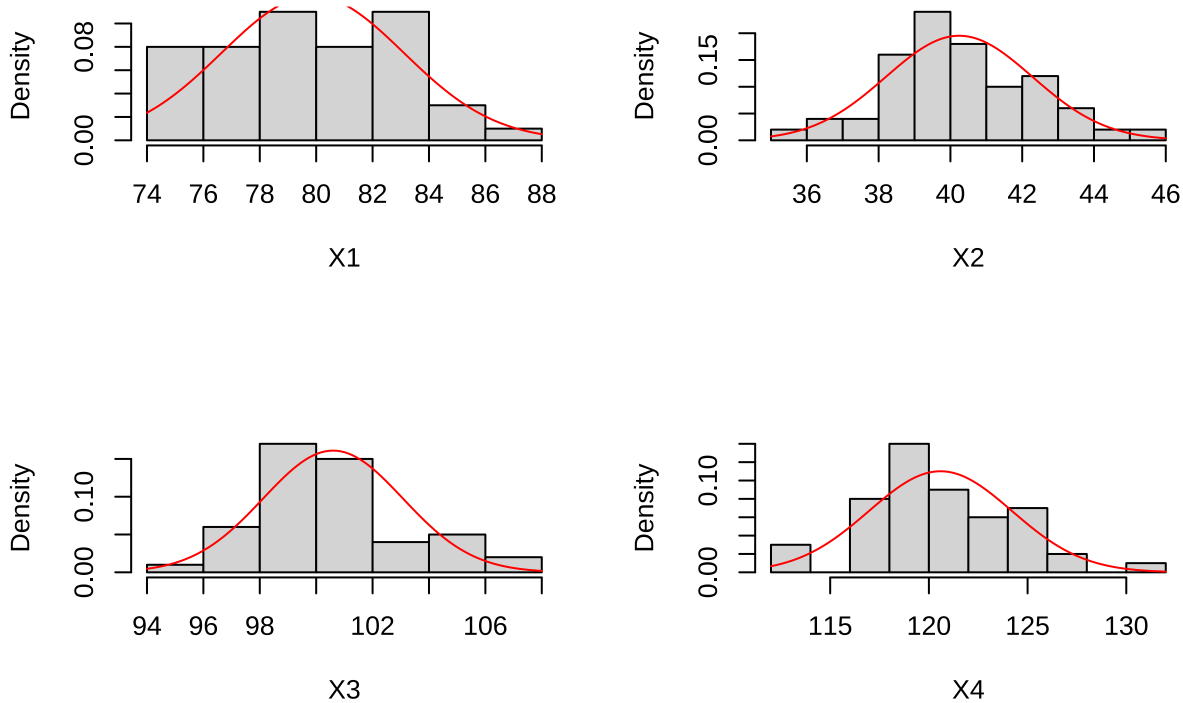
Como $pval = 0.001053 < \alpha = 0.05$ Entonces se rechaza H_0, los datos no provienen de una distribución normal.

Test multivariado

```
mvn(data, mvnTest = "royston", univariateTest = "CVM", univariatePlot = "histogram",
multivariatePlot = "qq", multivariateOutlierMethod = "adj",
showOutliers = TRUE, showNewData = TRUE)
```


Chi−Square Q−Q Plot
Chi−Square Quantile
Squared Mahalanobis Distance
0
2
4
6
8
10
12


Adjusted Chi−Square Q−Q Plot
Outliers (n=3)
Non−outliers (n=47)
Quantile: 12.95
34
44
6
Chi−Square Quantile
0
2
4
6
8
10
12
0
5
10
15
20
25
Robust Squared Mahalanobis Distance




```
## $multivariateNormality
##      Test        H  p value MVN
## 1 Royston 5.877263 0.170757 YES
##
## $univariateNormality
##              Test  Variable Statistic   p value Normality
## 1 Cramer-von Mises    X1        0.0909    0.1458    YES
## 2 Cramer-von Mises    X2        0.1164    0.0648    YES
## 3 Cramer-von Mises    X3        0.1244    0.0504    YES
## 4 Cramer-von Mises    X4        0.0699    0.2750    YES
##
## $Descriptives
##     n     Mean  Std.Dev  Median    Min    Max     25th     75th      Skew
## X1 50  79.8770 3.229307  79.825  74.22  86.95  76.9200  82.8650 0.1543652
## X2 50  40.2440 2.042492  39.955  35.80  45.17  39.0050  41.7700 0.2663509
## X3 50 100.5982 2.477647 100.045  94.87 107.22  99.0475 101.6225 0.5391567
## X4 50 120.5870 3.623060 119.985 112.78 130.32 118.3100 123.3650 0.1389550
##       Kurtosis
## X1 -1.00576687
## X2 -0.39414257
## X3  0.19347555
## X4 -0.03762874
##
## $multivariateOutliers
##    Observation Mahalanobis Distance Outlier
## 34          34               26.474    TRUE
```

```
## 44          44               18.227    TRUE
## 6            6               16.773    TRUE
##
## $newData
##       X1    X2     X3     X4
## 1  79.76 42.00 104.02 124.18
## 10 82.97 43.60 104.81 123.44
## 11 80.09 38.59  96.91 118.25
## 12 79.89 40.15  98.87 120.10
## 13 82.64 38.60  97.61 112.78
## 14 76.95 37.69 100.04 118.81
## 15 83.85 40.48  99.88 119.61
## 16 77.65 38.53  99.95 118.25
## 17 79.25 39.02  98.59 118.55
## 18 77.69 38.02  97.69 118.58
## 19 78.88 39.00 101.96 117.00
## 2  83.41 39.46 101.32 117.34
## 20 75.22 39.81 100.86 121.24
## 21 80.15 39.41  99.00 118.59
## 22 83.94 42.69 101.65 122.60
## 23 80.00 42.67 101.54 123.52
## 24 83.13 40.16  99.63 117.82
## 25 86.95 41.49  97.67 121.29
## 26 75.80 38.74 100.89 117.57
## 27 84.75 42.21 101.03 119.13
## 28 78.30 41.71 100.05 124.46
## 29 83.20 39.68 101.52 118.79
## 3  80.41 39.72  99.83 119.47
## 30 82.58 42.91 103.09 126.39
## 31 76.47 35.80  98.03 113.33
## 32 80.08 38.95  98.74 119.86
## 33 80.86 40.13  98.63 120.82
## 35 75.66 36.78  97.59 117.76
## 36 75.62 40.03 102.62 123.14
## 37 83.21 42.25 101.44 119.87
## 38 76.07 40.73 101.02 125.52
## 39 79.71 41.09 102.12 124.76
## 4  82.94 43.16 104.23 123.88
## 40 76.88 38.21  99.90 116.58
## 41 76.23 37.93  99.74 121.09
## 42 80.20 40.02  99.96 120.99
## 43 75.33 39.30  99.19 121.55
## 45 75.75 39.37 100.64 121.49
## 46 81.21 39.05  97.96 117.26
## 47 74.22 40.26  99.65 121.47
## 48 76.91 36.44  94.87 112.84
## 49 79.47 39.21 100.29 118.49
## 5  83.75 43.81 105.23 126.46
## 50 84.64 40.33  99.66 119.05
## 7  79.04 41.79 102.72 122.48
## 8  75.54 39.89 100.06 125.42
## 9  81.86 39.34  99.24 117.18
```

Conclusiones generales

Se concluye que respecto al gráfico "QQ plot ajustado de Chi Cuadrado" (Adjusted Chi-Square Q-Q plot) existen 3 datos atípicos a partir del cuartil 12.95, lo cual indica que el resto de la muestra, 47 datos, si siguen una distribución normal.

Respecto a la prueba de Cramen - Von Mises todos los p-valores son mayores a $\alpha = 0.05$, lo cual indica que no se rechaza H_0 , por tanto los datos tienen una distribución normal univariada.

Conforme a la prueba de normalidad multivariada de Royston se concluye que el $pvalor = 0.170757$ es mayor a alfa, por tanto los datos si provienen de una normal multivariada.

También se observa los histogramas de las cuatro variables y es posible decir que si tienden a una distribución normal, aun que presentan sesgo.