8. ANÁLISIS MULTIVARIADO

El análisis de datos multivariantes comprende el estudio estadístico de varias variables medidas en elementos de una población, con los siguientes objetivos: 1) Resumir los datos mediante un pequeño conjunto de nuevas variables, construidas como transformaciones de las originales, con la mínima pérdida de información; 2) Encontrar grupos en los datos, si existen; 3) Clasificar nuevas observaciones en grupos definidos, y 4) Relacionar dos conjuntos de variables (Peña, 2002).

8.1 MATRIZ DE DATOS

Si se tiene **p** variables numéricas en un conjunto de **n** elementos, cada una de estas **p** se denomina variable univariante, y el conjunto de las **p** variables se denomina una variable multivariante. Para construir esta matriz en R existen varias formas; una es mediante el comando **cbin()**, visto en la sección correspondiente a matrices.

Una matriz de datos también se puede construir mediante las hojas de datos (**data frames**), que son estructuras similares a una matriz, en donde cada columna puede ser de un tipo distinto a las otras. Las hojas de datos son apropiadas para describir "matrices de datos" donde cada fila representa a un individuo, y cada columna, una variable, variables que pueden ser numéricas o categóricas. Considere las siguientes variables: género, peso, edad y estatura, medidas en 5 personas, a partir de las cuales se construye la siguiente matriz de datos.

```
>
> # Variables
> Género=c("Hombre","Mujer","Hombre","Mujer","Mujer")
> Estatura=c(170,160,162,168,160)
> Peso=c(70,50,65,60,62)
> Edad=c(27,26,32,40,21)
>
> # matriz de datos
> D=data.frame(Género,Estatura,Peso,Edad)
> D
   Género Estatura Peso Edad
1 Hombre      170   70   27
2  Mujer      160   50   26
3 Hombre      162   65   32
4  Mujer      168   60   40
5  Mujer      160   62   21
>
```

Imagen 127. Salida R Construcción de un data frame

Para describir los datos multivariantes primero se debe realizar una descripción de cada una de las variables por separado y luego las relaciones que se presentan entre ellas. Es así como para calcular algunas estadísticas sobre cada variable se usa el comando **summary()**.

```
>
> # Estadísticas descriptivas por variable
> summary(D)
   Género      Estatura        Peso           Edad
 Hombre:2   Min.   :160   Min.   :50.0   Min.   :21.0
 Mujer :3   1st Qu.:160   1st Qu.:60.0   1st Qu.:26.0
            Median :162   Mean   :62.0   Median :27.0
            Mean   :164   Mean   :61.4   Mean   :29.2
            3rd Qu.:168   3rd Qu.:65.0   3rd Qu.:32.0
            Max.   :170   Max.   :70.0   Max.   :40.0
>
```

Imagen 128. Salida R Resumen data frame

A continuación se presentan algunos conceptos y comandos útiles para realizar el análisis exploratorio de observaciones. Existen varias formas para calcular medidas de interés en este tipo de análisis, las cuales dependen de la presentación de la matriz de datos.

8.2 VECTOR DE MEDIAS

Es la medida de centralización más usada para describir datos multivariantes; es el vector constituido por los promedios de cada una de las variables. Cuando la matriz de datos es construida con el comando **cbind()** se cuenta con dos procedimientos para calcular el vector de medias. El primero es de forma matricial, tal como se muestra a continuación:

$$vecmedias = \frac{1}{n} X^t 1$$

Ecuación 15. Vector de medias

donde 1 representa un vector de unos con dimensión igual al tamaño de la muestra **n**, y **X** es la matriz de datos (para el caso es la matriz de datos traspuesta).

Ejemplo: Dadas las variables x1, x2 y x3, determinar el vector de medias.

```
>
> # Variables
> x1=c(23,15,46,25,32)
> x2=c(65,70,59,71,68)
> x3=c(173,168,159,150,154)
>
> # Matriz de datos
> X= cbind(x1,x2,x3)
> X
     X1 x2  x3
[1,] 23 65 173
[2,] 15 70 168
[3,] 46 59 159
[4,] 25 71 150
[5,] 32 68 154
>
```

```
>
> # Tamaño muestral
> n = 5
>
> # Vector de unos
> unos = c(rep(1,n))
>
> # Vector de medias
> vecmedias = (1/n)*t(X)%*%unos
> vecmedias
     [,1]
x1   28.2
x2   66.6
x3  160.8
>
```

Imágenes 129 y 130. Salida R Vector de medias

El segundo procedimiento es mediante el comando **apply(datos, 1 ó 2, mean)**; recuerde que si se deja 1 se calcula la función por filas, y si se deja 2 se calcula la función por columnas.

```
>
> vecmedias=apply(X,2,mean)
> vecmedias
   x1    x2     x3
 28,2 66.6 160.8
>
```

Imagen 131. Salida R Vector de medias

8.3 MATRIZ DE VARIANZAS Y COVARIANZAS

Esta matriz permite determinar, por una parte, la variabilidad respecto a la media de cada una de las variables, y, por otra, la relación lineal por pares de variables, si esta existe. Al igual que en el cálculo del vector de medias, existen varias opciones. Cuando la matriz de datos es construida con el comando **cbind()**, se cuenta con dos procedimientos para calcular la matriz de varianzas y covarianzas; el primero es de forma matricial, tal como se indica en la expresión siguiente:

$$\widehat{S} = \frac{1}{n-1} X^t P X$$

$$P = I - \frac{1}{n} 1 1^t$$

Ecuación 16. Matriz de varianzas

Con I matriz identidad de orden n y el vector 1 de dimensión n.

Ejemplo: Considere las variables *x1, x2, x3 y x4* y construya la matriz de varianzas y covarianzas.

```
>
> # Variables
> x1=c(23,15,46,25,32)
> x2=c(65,70,59,71,68)
> x3=c(173,168,159,150,154)
>
> # Matriz de datos
> X= cbind(x1,x2,x3)
>
> # Tamaño de la muestra
> n = 5
>
> # Vector de unos
> unos = c(rep(1,5))
>
> # Matriz identidad
> I = diag(5)
>
> # Matriz P
> P = I-(1/n)*unos%*%t(unos)
>
> # Matriz de varianzas y covarianzas
> cov = (1/(n-1))*t(X)%*%P%*%X
> cov
       x1      x2      x3
x1 135.70 -45.15 -45.45
x2 -45.15  23.30  -9.60
x3 -45.45  -9.60  91.70
>
```

Imagen 132. Salida R Matriz de varianzas

El segundo procedimiento es mediante el comando **cov(nombre de la matriz de datos).**

```
>
> # Matriz de varianzas y covarianzas
> cov(X)
       x1      x2      x3
x1 135.70 -45.15 -45.45
x2 -45.15  23.30  -9.60
x3 -45.45  -9.60  91.70
>
```

Imagen 133. Salida R Matriz de varianzas

8.4 MATRIZ DE CORRELACIONES

La dependencia por pares entre las variables se mide por la matriz de correlación R, matriz cuadrada y simétrica con unos en su diagonal principal y fuera de ella los coeficientes de correlación lineal entre pares de variables. Esta matriz se puede calcular de forma matricial de la siguiente forma:

$$R = D^{-1/2} S D^{-1/2}$$

Ecuación 17. Matriz de correlaciones

donde D corresponde a la matriz diagonal formada por los elementos de la diagonal principal de la matriz de varianzas y covarianzas muestrales S.

Ejemplo: Para la ilustración considere los mismos datos utilizados en los ejemplos del vector de medias y matriz de varianzas y covarianzas.

```
>
> # Matriz de varianzas y covarianzas
> S = cov(X)
>
> # Elementos de la diagonal de S
> E = diag(S)
>
> Matriz diagonal "diagonal(S)"
> D = diag(E)
>
> # Matriz de correlaciones
> R = solve(sqrt(D))%*%S%*%solve(sqrt(D))
> R
           [,1]       [,2]       [,3]
[1,]  1.0000000 -0.8029525 -0.4074359
[2,] -0.8029525  1.0000000 -0.2076867
[3,] -0.4074359 -0.2076867  1.0000000
>
```

Imagen 134. Salida R Matriz de correlaciones

Para realizar el cálculo directo de la matriz de correlaciones se recurre al comando **cor(matriz de datos)**.

```
>
> # Calculo directo de la matriz de correlaciones
> R = cor(X)
> R
           x1         x2         x3
x1  1.0000000 -0.8029525 -0.4074359
x2 -0.8029525  1.0000000 -0.2076867
x3 -0.4074359 -0.2076867  1.0000000
>
```

Imagen 135. Salida R Matriz de correlaciones

8.5 CÁLCULOS A PARTIR DE UN DATA FRAME

Un data frame puede estar formado tanto de variables cualitativas como cuantitativas; por esta razón se hace necesario que las variables cualitativas sean excluidas al momento de determinar el vector de medias, la matriz de varianzas y la matriz de correlaciones de cada matriz de datos.

Ejemplo: Considere las siguientes cuatro variables: edad (x1), peso (x2), estatura en cm (x3) y genero (x4, H=hombre y M=mujer).

```
> # Variables
> x1=c(23,15,46,25,32
> x2=c(65,70,59,71,68)
> x3=c(173,168,159,150,154)
> x4=c("H","M","M","H","H")
>
> # Matriz de datos
> X=data.frame(x1,x2,x3,x4)
> X
   X1  X2  X3  x4
1  23  65  173  H
2  15  70  168  M
3  46  59  159  M
4  25  71  150  H
5  32  68  154  H
```

Imagen 136. Salida R Matriz de datos data frame

Con esta matriz se puede determinar el vector de medias, la matriz de varianzas y la de correlaciones; se debe tener en cuenta que para realizar estos cálculos se debe obviar la variable x_4.

```
> # Vector de medias
> vecmed = mean(X[,-c(4)])
> vecmed
   x1    x2    x3
 28.2  66.6 160.8
>
> # Matriz de varianzas
> mvar = var(X[,-c(4)])
> mvar
      x1      x2      x3
x1 135.70 -45.15 -45.45
x2 -45.15  23.30  -9.60
x3 -45.45  -9.60  91.70
>
> # matriz de correlaciones
> mcor = cor(X[,-c(4)])
> mcor
           x1         x2         x3
x1  1.0000000 -0.8029525 -0.4074359
x2 -0.8029525  1.0000000 -0.2076867
x3 -0.4074359 -0.2076867  1.0000000
```

Imagen 137: Salida R Matriz de datos data frame

Como se puede apreciar en el ejemplo anterior, se hizo necesario eliminar la cuarta variable para el cálculo de algunas medidas numéricas de interés para la matriz de datos; de ser necesario se puede eliminar más de una columna; esto se hace escribiendo el número de la columna por eliminar dentro del argumento X[,-c(***variables a ser eliminadas***)]; las variables deben estar separadas por comas.

8.6 DISTANCIA DE MAHALANOBIS

Se define la distancia de mahalanobis entre un punto y su vector de medias por:

$$d_i^2 = \left[(x_i - \hat{x})^t S^{-1}(x_i - \hat{x})\right]$$

Ecuación 18. Distancia de mahalanobis

La distancia de mahalanobis se calcula a través del comando **mahalanobis(x, center, cov)**, donde **x**, matriz de datos; **center**, vector de medias, y **cov**, matriz de varianzas y covarianzas.

Ejemplo: En la siguiente tabla se presentan medidas antropométricas tomadas a 15 trabajadores del sector alfarero del municipio de Ráquira (Boyacá); las variables en estudio son: estatura (EST), alcance lateral con asimiento (ALA), alcance frontal con asimiento (AFA), altura vertical con asimiento (AVA) y piso-codo (PC).

Tabla 7. Datos alfareros
Tomado de estudio de alfareros de Boyacá Grupo Taller 11

Observaciones	EST	ALA	AFA	AVA	PC
1	148	74	74	185	92
2	160	81	81	200	102
3	140	72	71	176	92
4	176	84	84	213	113
5	160	80	82	198	105
6	162	80	71	196	99
7	166	89	86	207	105
8	144	73	72.5	180.5	92
9	160	84	83	201	98
10	163	82	86	204	103
11	150	74	76	184	95
12	172	86	84	215	110
13	158	82	79	202	101
14	158	76	77	194	105
15	158	82	80	197	100

Inicialmente se procede a construir la matriz de datos y a determinar el vector de medias y la matriz de varianzas y covarianzas.

```
>
> # Variables
> Est=c(148,160,140,176,160,162,166,144,160,163,150,172,158,158,158)
> ALA=c(74,81,72,84,80,80,89,73,84,82,74,86,82,76,82)
> AFA=c(74,81,71,84,82,71,86,72.5,83,86,76,84,79,77,80)
> AVA=c(185,200,176,213,198,196,207,180.5,201,204,184,215,202,194,197)
> PC=c(92,102,92,113,105,99,105,92,98,103,95,110,101,105,100)
>
> # Matriz de datos
> X = cbind(EST,ALA,AFA,AVA,PC)
>
> # Vector de medias
> Vecmed = apply(X,2,mean)
>
> # Matriz de varianzas y covarianzas
> S = cov(X)
```

Imagen 138. Salida T Distancia de mahalanobis

Luego de esto se procede a calcular la distancia de mahalanobis para cada observación, y por último se presenta una matriz con la información de cada individuo con su correspondiente distancia.

```
>
> # distancia de mahalanobis
> di = mahalanobis(X, Vecmed, S)
> # Observaciones y distancias correspondientes
> Xdi = cbind(X,di)
> Xdi
      EST ALA  AFA   AVA  PC         di
 [1,] 148  74 74.0 185.0  92  3.9692477
 [2,] 160  81 81.0 200.0 102  0.3449645
 [3,] 140  72 71.0 176.0  92  5.6556945
 [4,] 176  84 84.0 213.0 113  6.0571089
 [5,] 160  80 82.0 198.0 105  2.6372598
 [6,] 162  80 71.0 196.0  99 10.5634604
 [7,] 166  89 86.0 207.0 105  8.5471896
 [8,] 144  73 72.5 180.5  92  2.2789303
 [9,] 160  84 83.0 201.0  98  4.1267558
[10,] 163  82 86.0 204.0 103  4.7876563
[11,] 150  74 76.0 184.0  95  3.9927099
[12,] 172  86 84.0 215.0 110  4.6368720
[13,] 158  82 79.0 202.0 101  6.7518356
[14,] 158  76 77.0 194.0 105  4.5135475
[15,] 158  82 80.0 197.0 100  1.1367675
>
```

Imagen 139. Salida R Distancia de mahalanobis

8.7 ANÁLISIS GRÁFICO DE OBSERVACIONES MULTIVARIANTES

Un primer paso en el análisis multivariante es representar gráficamente las variables individualmente; en segundo lugar es conveniente construir diagramas de dispersión de las variables por parejas; esto se puede realizar mediante el comando **pairs(datos,...)**. A continuación se presenta un ejemplo con los datos de los trabajadores alfareros.

```
>
> # Diagrama de dispersión
> pairs(x, pch=5, col="blue",
+ main="Gráficos de dispersión bivariante")
>
```

Imagen 140. Salida R Construcción de diagrama bivariante

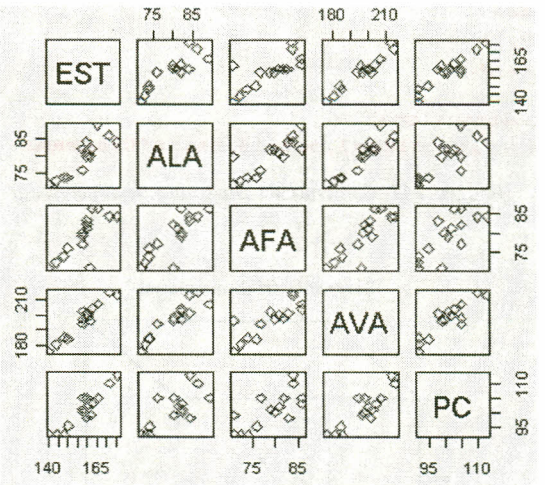

Imagen 141. Salida R Diagrama de dispersión bivariante

Este tipo de gráfico permite observar relaciones existentes entre las variables y la presencia de datos atípicos. Cuando se trabaja con tres o cuatro variables, la función **coplot()** puede ser más apropiada. Si **a** y **b** son vectores numéricos y **c** es un vector numérico o un factor (todos de la misma longitud), entonces, la orden **coplot(a ~ b / c)** produce diagramas de dispersión de **a** sobre **b** para cada valor de **c**. Si **c** es un factor, esto significa que **a** se representa sobre **b** para cada nivel de **c**. Si **c** es un vector numérico, entonces se agrupa en intervalos, y para cada intervalo se representa **a** sobre **b** para los valores de **c** dentro del intervalo. El número y tamaño de los intervalos puede controlarse con el argumento **given.values** de la función **coplot()**. La función **co.intervals()** también es útil para seleccionar intervalos. Asimismo, es posible utilizar dos variables condicionantes con una orden como **coplot(a ~ b / c +d)**, que produce diagramas de **a** sobre **b** para cada intervalo de condicionamiento de **c** y **d**.

Gráficos de dispersión 3 - variante: Cuando se tienen tres variables numéricas es posible realizar un diagrama de dispersión con ellas mediante el siguiente comando:

scatterplot3d(x, y=NULL, z=NULL, color=par("col"), pch=NULL, main=NULL, sub=NULL, xlim=NULL, ylim=NULL, zlim=NULL, xlab=NULL, ylab=NULL, zlab=NULL,...)

En el anterior comando sólo se presentan algunos argumentos; para mayor información se puede consultar la ayuda interactiva. Para realizar este diagrama es necesario que previamente se cargue el paquete **scatterplot3d**.

Ejemplo: Considere las variables x, y, z para realizar un diagrama de dispersión tri-dimensional.

```
>
> # Variables
> x = c(1,5,7,9,12)
> y = c(12,3,4,20,7)
> z = c(5,21,16,2,13)
>
> # Diagrama tri-dimensional
> scatterplot3d(x,y,z,color=3,pch=15,main="Diagrama tri-dimensional")
```

Imagen 142. Construcción Diagrama tri-dimensional

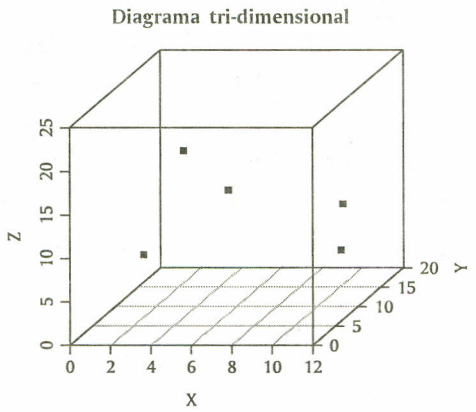

Imagen 143. Salida R Diagrama tri-dimensional

8.8 DISTRIBUCIÓN NORMAL MULTIVARIADA

El vector aleatorio p-dimensional **x** tiene distribución normal p-variante con vector de medias **ì** = ($ì_1$, $ì_2$, $ì_3$,...,$ì_p$) y matriz de covarianzas **Ó** de tamaño **pxp**, por ello tiene como función de densidad conjunta a:

$$f_x(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp^{\left[-\frac{1}{2}(X-\mu)^t \Sigma^{-1}(X-\mu)\right]}$$

Ecuación 19. Distribución Normal

Es posible generar datos aleatorios de una distribución p-variante con el comando **mvrnorm(n=#, mu, Sigma)**, donde **n** indica el número de observaciones que se desean; **mu** es el vector de medias, y **sigma**, la matriz de varianzas y covarianzas.

Ejemplo: Si se quiere generar 6 observaciones de una distribución 5-variante con vector de medias $\mu = (2, 3, 4, 5, 6)$ y matriz de varianzas y covarianzas igual a la identidad, se procede así:

```
>
> # Número de observaciones
> n = 6
>
> # Vector de medias
> mu = c(2,3,4,5,6)
>
> # Matriz de varianzas y covarianzas
> sigma = diag(5)
>
> # Normal 5-variante
> mvrnorm(n, mu, sigma)
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 2.7439461 2.195412 4.160518 2.546877 5.532961
[2,] 0.8959809 2.289943 4.077959 6.917811 6.342687
[3,] 1.2166718 4.192020 2.870380 3.957813 6.321503
[4,] 2.4007936 2.104596 4.764110 5.553305 6.567440
[5,] 2.5109468 4.529655 3.716143 6.777010 6.213452
[6,] 3.3561570 1.673395 3.469901 5.473750 5.037210
>
```

Imagen 144. Salida R Normal multivariante

8.9 ELIPSES DE CONFIANZA

Un caso particular de la distribución normal multivariante se presenta cuando p=2, con lo que se genera la distribución normal bivariada, utilizada en muchas aplicaciones de la vida cotidiana; a continuación se muestra cómo construir las coordenadas de elipses de confianza del (1-a)100% para un conjunto de n observaciones de una distribución normal bivariada; previamente se debe haber cargado el paquete **ellipse**; el comando utilizado es:

ellipse(x, centre, level = 0.95, npoits =)

Los argumentos utilizados son: **x**, matriz de correlaciones; **centre**, vector con las coordenadas del centro de la elipse (vector de medias); **level**, indica el nivel de confianza para la región, y

npoints indica el número de parejas ordenadas (puntos de la elipse). Para graficar esta elipse, el comando anterior se escribe dentro del comando **plot()**, así:

plot(ellipse(x, centre, level = 0.95, npoits =))

Si la matriz de correlaciones es igual a la matriz identidad, entonces, la gráfica corresponderá a una circunferencia; a continuación se presentan ejemplos de los casos expuestos anteriormente.

Ejemplo: Dada una distribución normal bivariante con vector de medias μ y matriz de correlaciones R, construir una elipse de confianza del 92% para la distribución, donde,

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ y } R = \begin{bmatrix} 1 & 0.35 \\ 0.35 & 1 \end{bmatrix}$$

Imagen 145. Vector de medias y matriz de varianzas

```
>
> # centro de la elipse
> cen = c(0,0)
> # matriz de correlaciones
> R = matrix(c(1,0.35,0.35,1),2)
>
> # Elipse
> plot(ellipse(R,center=cen,level=0.92,npoints=100))
```

Imagen 146. Creación de elipse de confianza

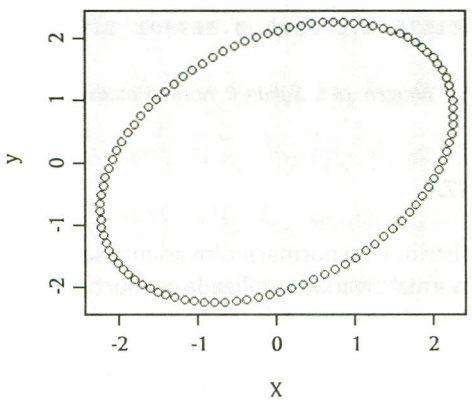

Imagen 147. Salida R elipse de confianza

Ejemplo: Dada una distribución normal bivariante con vector de medias μ y matriz de correlaciones R, construir una elipse de confianza del 96% para la distribución.

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ y } R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Imagen 148. Vector de medias y matriz de varianzas

```
>
> # centro de la elipse
> cen = c(0,0)
> # matriz de correlaciones
> R = diag(2)
>
> # Elipse
> plot(ellipse(R,center=cen,level=0.96,npoints=100))
```

Imagen 149. Construcción elipse de confianza

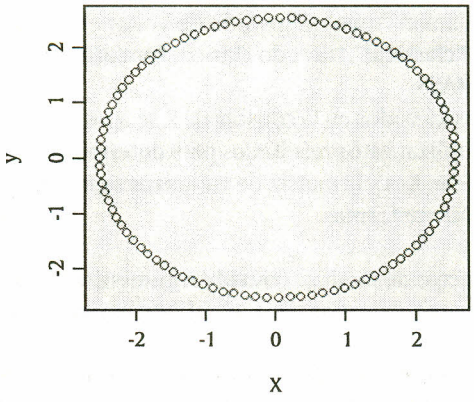

Imagen 150. Salida R elipse de confianza

Las gráficas anteriores pueden ser modificadas por medio de los parámetros gráficos anteriormente descritos (títulos y colores, entre otros).

8.10 EVALUACIÓN DE LA MULTINORMALIDAD

Un primer paso para probar la multinormalidad de un conjunto de observaciones es analizar cada una de las variables por separado, advirtiendo que esto no es suficiente, puesto que si solo se hiciera esto se estaría dejando de lado la asociación lineal entre las variables.

Datos atípicos: son aquellas observaciones que parecen haberse generado de manera distinta a las demás. Un primer procedimiento para identificar este tipo de observaciones es mediante gráficos y cálculo de distancias entre observaciones (distancia de mahalanobis) a fin de verificar si algún punto está alejado del conjunto de observaciones. Las consecuencias de una sola observación atípica pueden ser graves, entre estas se encuentran distorsión en promedios y

desviaciones estándar de las variables; por tanto, y como la distancia de mahalanobis está directamente relacionada con el vector de medias y la matriz de varianzas y covarianzas, puede no llegar a reflejar correctamente las observaciones atípicas (efecto de enmascaramiento). Una propuesta para obviar este problema es utilizar estimadores robustos, que son diseñados para verse poco afectados por cierta contaminación de atípicos (Peña, 2002).

Los estimadores robustos permiten realizar estimaciones para el vector de medias y la matriz de varianzas y covarianzas; estas estimaciones no se ven tan afectadas por la presencia de datos atípicos, y al utilizarlas para determinar la distancia de mahalanobis, esta refleja realmente los posibles alejamientos de un dato o un conjunto de datos de la población bajo estudio. El comando que permite realizar dichas estimaciones es:

cov.rob(x, cor=FALSE, method=c("mve", "mcd", "classical"))

Los argumentos utilizados en este comando son: **x**, matriz de datos; **cor=** función lógica por defecto FALSE, si es TRUE devuelve junto con los resultados la matriz de correlaciones; **method=** se refiere al método por el cual se realizan las estimaciones, en este caso los métodos implementados en R se llaman **"mve"** (Elipsoide de Volumen Mínimo); **"mcd"**, Covarianza de Determinante Mínimo, y **"classical"**, método clásico. Para utilizar este comando se debe cargar previamente el paquete **MASS.**
Al aplicar cualquiera de los métodos en la consola de R se aprecian todos los resultados como un solo objeto, y si se desea utilizar estos resultados para determinar la distancia de mahalanobis se necesita que el vector de medias y la matriz de varianzas sean objetos independientes, para lo cual se procede de la siguiente forma:

vector de medias: **cov.rob(argumentos)$center**
matriz de varianzas: **cov.rob(argumentos)$cov**

Ejemplo: Consiste en la generación de observaciones provenientes de dos distribuciones multinormales con distintos parámetros, con el fin de comparar los estimadores robustos frente a los estimadores usuales (vector de medias y matriz de varianzas y covarianzas muestrales).

Generación de muestras:

M1: muestra aleatoria de tamaño n = 25 de una distribución normal 3-variante con

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \text{ y } \Sigma_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

```
> mu1 = c(0,0,0)
> sigma1 = diag(3)
> x1 = mvrnorm(n=25,mu1,sigma1)
```

Imágenes 151 y 152. Construcción de muestras aleatorias M1

M2: muestra aleatoria de tamaño n=5 de una distribución normal 3-variante con

$$\mu_2 = \begin{bmatrix} 15 \\ 85 \\ 70 \end{bmatrix} \text{ y } \Sigma_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

```
> mu2 = c(15,85,70)
> sigma2 = diag(3)
> x2 = mvrnorm(n=5,mu2,sigma2)
```

Imágenes 153 y 154. Construcción de muestras aleatorias M2

Creación de la matriz de datos. Se unen las dos muestras aleatorias dentro de un mismo arreglo mediante el comando **rbind()**.

```
> X=rbind(x1,x2)
> X
             [,1]        [,2]        [,3]
 [1,]  0.430314213 -2.46499251  1.48551076
 [2,] -0.812518531 -0.33530859  1.01089672
 [3,]  1.536175155 -0.44371158  0.04336103
 [4,] -2.103800963 -0.11578560 -1.02276904
 [5,]  0.206273353  0.06576601  0.36081157
 [6,] -1.381900544 -0.07797751  0.24924408
 [7,] -0.871928839  0.88283780  0.17276662
 [8,] -1.213167012  1.39217469  1.29961514
 [9,] -1.005834566  0.39247735  0.60899808
[10,] -1.381900544 -0.07797751  0.24924408
[11,]  0.003762816  0.83687569  0.57117769
[12,]  1.147156783 -1.05345334 -0.82417008
[13,] -1.926147757 -1.20268938  1.78945174
[14,]  0.237389932 -0.22885753 -0.55974433
[15,]  0.139029616  1.22639374 -1.78587836
[16,] -0.531334715  0.72616816 -0.05103377
[17,]  0.213814969  0.97092876  0.7420808
[18,] -0.555377515  0.66231270 -1.15219277
[19,]  1.046889659 -0.78080671  0.26134942
[20,] -0.070029371  1.12885487 -0.51858915
[21,] -0.636398694 -1.61744701  0.48155900
[22,] -0.058228944  0.23079044  1.08871860
[23,] -1.036050912  0.85473871  0.91852479
[24,]  0.588420690  0.01574196  0.53243638
[25,] -1.409747460 -0.05213258  1.18323715
>
```

Imagen 155. Construcción matriz comando rbind

Estimadores usuales vector de medias, matriz de varianzas y covarianzas:

```
>
> vmedusual=apply(X,2,mean)
> vmedusual
[1]   2.25199  14.17923  12.06242
>
> covusual=cov(X)
> covusual
             [,1]       [,2]      [,3]
 [1,]   35.04836   187.4725  155.9492
 [2,]  187.47248  1029.9764  857.0561
 [3,]  155.94925   857.0561  715.3208
>
```

Imagen 156. Salida R estimadores usuales

Estimador robusto (Elipsoide de Volumen Mínimo):

```
>
> vmedMVE=cov.rob(X,method="mve")$center
> vmedMVE
[1] -0.3000781  0.1959372  0.2842068
>
> covMVE=cov.rob(X,method="mve")$cov
> covMVE
             [,1]       [,2]       [,3]
 [1,]   0.7664138 -0.1609999 -0.3356907
 [2,] -0.1609999  0.9497991 -0.2514190
 [3,] -0.3356907 -0.2514190  0.7692446
>
```

Imagen 157. Salida R estimador mve

Estimador robusto (Covarianza de Determinante Mínimo):

```
>
> vmedMCD=cov.rob(X,method="mcd")$center
> vmedMCD
[1] -0.2732624  0.1934830  0.3164735
>
> covMCD=cov.rob(X,method="mcd")$cov
> covMCD
             [,1]        [,2]       [,3]
 [1,]  0.77968740 -0.08684056 -0.3867783
 [2,] -0.08684056  0.68510753 -0.1274671
 [3,] -0.38677827 -0.12746706  0.7446783
>
```

Imagen 158. Salida R estimadores mcd

Cálculo de los cuadrados de la distancia de mahalanobis para cada uno de los estimadores:

```
>
> # Distancias estimador usual
> diusual=mahalanobis(X,vmedusual,covusual)
>
> # Distancias estimador MVE
> diMVE=mahalanobis(X,vmedMVE,covMVE)
>
> # Distancias estimador MCD
> diMCD=mahalanobis(X,vmedMCD,covMCD)
```

Imagen 159. Salida R Distancia de mahalanobis para los estimadores

Gráficas para las distancias calculadas con cada uno de los estimadores

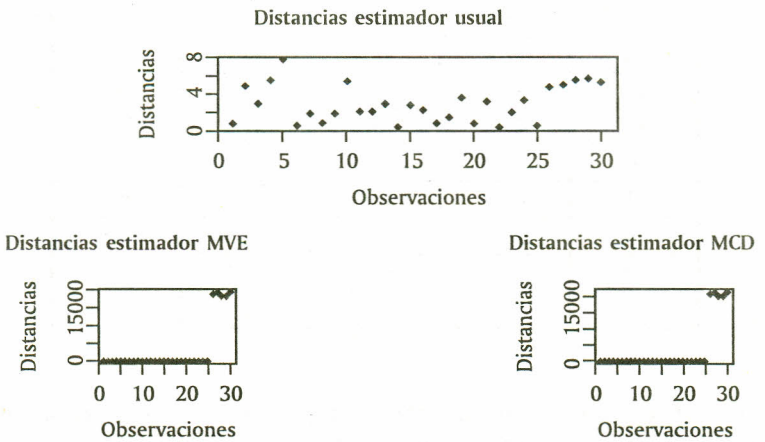

Imagen 160. Salida R Gráfica de las distancias de mahalanobis

Se observa claramente en los gráficos de las distancias de mahalanobis para los estimadores MVE y MCD que las observaciones con las que se contaminó el primer conjunto de datos están alejadas de este, mientras que en el gráfico para las distancias calculadas con el estimador usual estas observaciones se pueden llegar a confundir dentro del conjunto. El ejemplo anterior permitió verificar la eficacia de los estimadores robustos en la detección de datos atípicos cuando la matriz de datos es contaminada a propósito con datos provenientes de una distribución diferente a los datos iniciales de la matriz.

Ejemplo: Ahora se aplicarán los estimadores robustos a un conjunto de datos trabajados por Díaz (2002, p. 74), en un ejercicio en el que mediante diferentes procedimientos determina que las observaciones 9, 12 y 20 son datos potencialmente atípicos. En la siguiente tabla se muestran los datos sobre longitud de huesos registrados en 20 jóvenes a los 8, 8.5, 9 y 9.5 años, respectivamente (Rencher, 1995, p. 90, citado por Díaz):

Tabla 8. Datos sobre longitud de huesos

# obs	8 años(x_1)	8.5 años(x_2)	9 años(x_3)	9.5 años(x_4)
1	47.5	48.8	49.0	49.7
2	46.4	47.3	47.7	48.4
3	46.3	46.8	47.8	48.5
4	45.1	45.3	46.1	47.2
5	47.6	48.5	48.9	49.3
6	52.5	53.2	53.3	53.7
7	51.2	53.0	54.3	54.5
8	49.8	50.0	50.3	52.7
9	48.1	50.8	52.3	54.4
10	45.0	47.0	47.3	48.3
11	51.2	51.4	51.6	51.9
12	48.5	49.2	53.0	55.5
13	52.1	52.8	53.7	55.0
14	48.2	48.9	49.3	49.8
15	49.6	50.4	51.2	51.8
16	50.7	51.7	52.7	53.3
17	47.2	47.7	48.4	49.5
18	53.3	54.6	55.1	55.3
19	46.2	47.5	48.1	48.4
20	46.3	47.6	51.3	51.8

En la siguiente gráfica se observa un diagrama de dispersión de las distancias de mahalanobis tanto con los estimadores usuales como con los robustos (MVE, MCD):

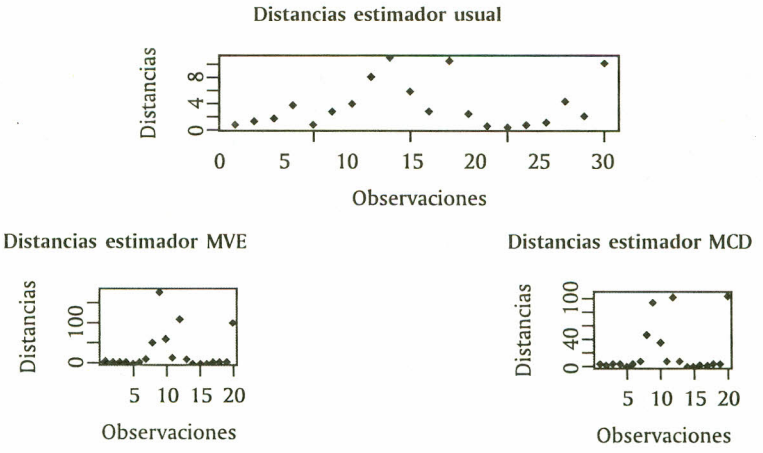

Imagen 161. Salida R Gráfica de las distancias de mahalanobis para los jóvenes

En los gráficos correspondientes a los estimadores robustos se identifican 5 posibles valores atípicos: los tres encontrados por Díaz (observaciones 9, 12 y 20) y dos observaciones adicionales (8 y 10) que surgen al utilizar los dos estimadores robustos.

8.11 EJERCICIOS

8.11.1 Los siguientes datos hacen referencia al seguimiento que la Secretaría de Salud viene realizando a 15 niños de una zona marginal de la ciudad.

Nombre	Edad (años)	Estatura (m)	Peso (kg)
José	12	1.4	48
Pedro	14	1.8	77
María	14	1.32	35
Carlos	16	1.6	40
Lucia	8	1.2	35
Maritza	9	1.4	35
Mariela	17	1.51	48
Mariana	15	1.56	52
Gabriela	12	1.3	45
Jesús	16	1.65	60
Oscar	15	1.7	62
David	9	1.2	30
Tania	12	1.4	40
Liliana	15	1.6	48
Lina	17	1.56	57

a. Introduzca estos datos en R como un data frame.
b. Construya: Vector de medias, matriz de varianzas y covarianzas, matriz de correlaciones.
c. Calcule la distancia de mahalanobis para cada observación.
d. Realice el análisis gráfico multivariante para el ejercicio.

8.11.2 Genere 10 observaciones de una distribución 6–variante con vector de medias $\mu = (5, 8, 2, 11, 3, 20)$ y matriz de varianzas y covarianzas igual a la identidad.