Segundo trabajo

Estudiantes

Cristian Alberto Cortes Zarate

Pablo Montoya Granada

Andrea Vallejo Cardona

Docente

Rene Iral Palomino

Asignatura

Introducción al Análisis Multivariado

Sede Medellín

9 de octubre de 2022

Índice

Índice de figuras

1. Parte A.

Sea $X = (X_1, X_2, X_3, X_4)'$, $N_4(\mu, \Sigma)$ donde $\mu = (d_1, d_2, d_3, d_4)'$ y $\Sigma = \begin{pmatrix} 4 & 3 & 2 & 1 \\ 3 & 8 & -1 & 3 \\ 2 & -1 & 3 & 1 \\ 1 & 3 & 1 & 6 \end{pmatrix}$

(d_1, d_2, d_3, d_4) corresponden a los cuatro últimos dígitos no nulos del número de su cédula o documento de identidad.

Sea $X^1 = \begin{pmatrix} X_2 \\ X_4 \end{pmatrix}$ y $X^1 = \begin{pmatrix} X_1 \\ X_3 \end{pmatrix}$

Se usa la cédula de la compañera Andrea Vallejo para escoger los valores que debe llevar nuestro vector de medias $\mu = (9, 9, 7, 9)'$.

1.1. (10 pts.)

Calcule $P(a'X < 2)$, donde $a' = (1, -1, 1, -1)$. Explique paso a paso cómo realizar dicho cálculo.

Dada la siguiente propiedad de la distribución normal multivariada:

Si $X \sim N_p(\mu, \Sigma)$, entonces

$$a'X = a_1X_1 + ... + a_pX_p \sim N(a'\mu, a'\Sigma a)$$

Análogamente, si $\forall a \in \mathbb{R}^p, a'X$ se distribuye normal univariada, entonces X se distribuye normal multivariada (Caracterización de Rao).

Se procede a aplicar la anterior propiedad a nuestros valores, obteniendo así:

$$Y \sim \left((1, -1, 1, -1) \begin{pmatrix} 9 \\ 9 \\ 7 \\ 9 \end{pmatrix}, (1, -1, 1, -1) \begin{pmatrix} 4 & 3 & 2 & 1 \\ 3 & 8 & -1 & 3 \\ 2 & -1 & 3 & 1 \\ 1 & 3 & 1 & 6 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} \right)$$

Las respectivas operaciones entre los vectores y la matriz dan como resultado:

$$\begin{aligned} & (1, -1, 1, -1) \cdot \begin{pmatrix} 9 \\ 9 \\ 7 \\ 9 \end{pmatrix} \\ & = (1 \cdot 9 + (-1) \cdot 9 + 1 \cdot 7 + (-1) \cdot 9) \\ & = 2 \end{aligned}$$

$$(1,-1,1,-1)\begin{pmatrix} 4 & 3 & 2 & 1 \\ 3 & 8 & -1 & 3 \\ 2 & -1 & 3 & 1 \\ 1 & 3 & 1 & 6 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}$$

$$(1,-1,1,-1)\begin{pmatrix} 4 & 3 & 2 & 1 \\ 3 & 8 & -1 & 3 \\ 2 & -1 & 3 & 1 \\ 1 & 3 & 1 & 6 \end{pmatrix} = (2,-9,5,-7)$$

$$(2,-9,5,-7)\begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}$$

$$= (2\cdot 1 + (-9)(-1) + 5\cdot 1 + (-7)(-1))$$

$$= 23$$

Teniendo así que X se distribuye como una normal con $\mu = 2, \sigma^2 = 23$.

$$Y \sim N(-2, 23)$$

Ya teniendo cómo se distribuye Y, ahora si se calcula la $P(a'X < 2)$ por medio del R, para calcularla en R se usa la función pnorm, la "p" permite calcular la f.d.a. para un valor a dado, es decir, $F(a) = P(X \leq a)$, el "norm" representa la distribución normal, a la función se le ingresan los parámetros $\mu = 2, \sigma^2 = 23$ y se usa el método lower.tail=TRUE para indicar que se quiere calcular una probabilidad menor a 2 en nuestro caso.

```
pnorm(2,-2,sqrt(23),lower.tail=TRUE)
```

```
## [1] 0.7978758
```

Se obtiene así una probabilidad del 79.78 % de que $a'X < 2$.

1.2. (15 pts.)

Sea $Z = MX + d$, con $M = \begin{pmatrix} -2 & 1 & 2 & 5 \\ 1 & -2 & 1 & 3 \end{pmatrix}$ y $d = (2,1)'$.

Halle la distribución de Z, (indicando cual es el vector de medias y la matriz de covarianzas de Z). Explique paso a paso cómo obtener la distribución pedida. ¿Son las variables aleatorias asociadas al vector Z, estadísticamente independientes? Justifique su respuesta.

Dada la siguiente propiedad de la distribución normal multivariada:

Si $X \sim N_p(\mu, \Sigma)$, entonces:

- El vector $Y = AX + b \sim N_q(A\mu + b, A\Sigma A')$; donde $A_{q\times p}$ y $b_{q\times 1}$.

Se procede a aplicar la anterior propiedad a nuestros valores, en busca de hallar la distribución de Z:

$$Z = \begin{pmatrix} -2 & 1 & 2 & 5 \\ 1 & -2 & 1 & 3 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

Vector de medias

$$\begin{aligned} \mu_z = M\mu + d &= \begin{pmatrix} -2 & 1 & 2 & 5 \\ 1 & -2 & 1 & 3 \end{pmatrix} \begin{pmatrix} 9 \\ 9 \\ 7 \\ 9 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} (-2)\cdot 9 + 1\cdot 9 + 2\cdot 7 + 5\cdot 9 \\ 1\cdot 9 + (-2)\cdot 9 + 1\cdot 7 + 3\cdot 9 \end{pmatrix} = \begin{pmatrix} 50 \\ 25 \end{pmatrix} \\ &= \begin{pmatrix} 50 \\ 25 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 52 \\ 26 \end{pmatrix} \end{aligned}$$

Matriz de covarianzas

$$\begin{aligned} \Sigma_z = M\Sigma M' &= \begin{pmatrix} -2 & 1 & 2 & 5 \\ 1 & -2 & 1 & 3 \end{pmatrix} \begin{pmatrix} 4 & 3 & 2 & 1 \\ 3 & 8 & -1 & 3 \\ 2 & -1 & 3 & 1 \\ 1 & 3 & 1 & 6 \end{pmatrix} \begin{pmatrix} -2 & 1 \\ 1 & -2 \\ 2 & 1 \\ 5 & 3 \end{pmatrix} \\ &= \begin{pmatrix} 184 & 79 \\ 79 & 65 \end{pmatrix} \end{aligned}$$

Teniendo asi que Z se distribuye como una normal con parámetros $\mu = (52, 26)', \Sigma = \begin{pmatrix} 184 & 79 \\ 79 & 65 \end{pmatrix}$.

$$Z \sim N_2((52, 26)', \begin{pmatrix} 184 & 79 \\ 79 & 65 \end{pmatrix})$$

Ahora en busca de analizar si las variables aleatorias asociadas al vector Z son estadísticamente independientes; sea ρ la correlación de las variables, de tal forma que:

$$\rho^2 = \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}}$$

Sabiendo que:

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} 184 & 79 \\ 79 & 65 \end{pmatrix}$$

Al reemplazar los valores se obtiene que:

$$\begin{aligned} \rho &= \sqrt{\frac{79^2}{184 * 65}} \\ &= 0.722 \end{aligned}$$

Como se obtuvo un valor diferente a 0, en este caso 0.722 se interpreta como que existe una correlación fuerte positiva entre las variables aleatorias asociadas al vector Z, obteniendo así que no son estadísticamente independientes dichas variables.

1.3. (15 pts.)

Halle la distribución condicional de X^1 dado $X^2 = (1, 2)'$. Explique de manera detallada cada paso para llegar a la distribución pedida.

Dada la siguiente propiedad de la distribución normal multivariada:

La distribución condicional de $X^{(1)}$ dado $X^{(2)} = x^{(2)} = (1, 2)'$ es una normal multivariada con vector de medias:

$$\mu_{\mathbf{X}^{(1)}|x^{(2)}} = \boldsymbol{\mu}^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}\left(\mathrm{x}^{(2)} - \boldsymbol{\mu}^{(2)}\right)$$

y matriz de covarianzas

$$\Sigma_{X^{(1)}|x^{(2)}} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Se procede a aplicar la anterior propiedad a nuestros valores, en busca de hallar la distribución condicional de $X^{(1)}$ dado $X^{(2)} = (1, 2)'$:

Se define Y como $Y = (X^{(1)} \mid X^{(2)}) = (X_2, X_4 \mid X_1, X_3)'$

Vector de medias

-$X^{(1)} = \begin{pmatrix} X_2 \\ X_4 \end{pmatrix} = \mu^{(1)} = \begin{pmatrix} 9 \\ 9 \end{pmatrix}$

-$X^{(2)} = \begin{pmatrix} X_1 \\ X_3 \end{pmatrix} = \mu^{(2)} = \begin{pmatrix} 9 \\ 7 \end{pmatrix}$

Dado así el vector de medias de Y es:

$$\mu_Y = \left(\mu^{(1)} \mid \mu^{(2)}\right)' = (9, 9 \mid 9, 7)'$$

Matriz de covarianzas

La matriz de covarianzas de Σ_Y seria:

$$\Sigma_{11} = \begin{pmatrix} 8 & 3 \\ 3 & 6 \end{pmatrix}, \Sigma_{12} = \begin{pmatrix} 3 & -1 \\ 1 & 1 \end{pmatrix}, \Sigma_{21} = \begin{pmatrix} 3 & 1 \\ -1 & 1 \end{pmatrix}, \Sigma_{22} = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}$$

$$\Sigma = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right) = \left(\begin{array}{cc|cc} 8 & 3 & 3 & -1 \\ 3 & 6 & 1 & 1 \\ \hline 3 & 1 & 4 & 2 \\ -1 & 1 & 2 & 3 \end{array}\right)$$

Dado que existe una matriz A tal que $Y = AX$, donde A es la matriz $A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$ y haciendo uso de la propiedad $Y \sim N_4(A\mu Y, A\Sigma Y A')$ se obtiene que μ_x y Σ_x dan resultados iguales a μ_y y Σ_y.

Entonces ya contando con el vector de medias y la matriz de covarianzas se puede aplicar la propiedad nombrada al inicio.

Vector de medias aplicando la propiedad

$$\mu_{\mathbf{X}^{(1)}|x^{(2)}} = \boldsymbol{\mu}^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}\left(\mathrm{x}^{(2)} - \boldsymbol{\mu}^{(2)}\right)$$

$$\mu_{\mathbf{X}^{(1)}|x^{(2)}} = \begin{pmatrix} 9 \\ 9 \end{pmatrix} + \begin{pmatrix} 3 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix} - 1\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 9 \\ 7 \end{pmatrix}\right)$$

$$= \begin{pmatrix} \frac{17}{4} \\ \frac{27}{4} \end{pmatrix} = \begin{pmatrix} 4.25 \\ 6.75 \end{pmatrix}$$

Matriz de covarianzas aplicando la propiedad

$$\Sigma_{X^{(1)}|x^{(2)}} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$$\Sigma_{X^{(1)}|x^{(2)}} = \begin{pmatrix} 8 & 3 \\ 3 & 6 \end{pmatrix} - \begin{pmatrix} 3 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix} - 1\begin{pmatrix} 3 & 1 \\ -1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{21}{8} & \frac{23}{8} \\ \frac{23}{8} & \frac{45}{8} \end{pmatrix} = \begin{pmatrix} 2.625 & 2.875 \\ 2.875 & 5.625 \end{pmatrix}$$

Teniendo así que $X^{(1)} \mid X^{(2)}$ se distribuye como una normal con $\mu_{\mathbf{X}^{(1)}|x^{(2)}} = \begin{pmatrix} 4.25 \\ 6.75 \end{pmatrix}, \Sigma_{X^{(1)}|x^{(2)}} = \begin{pmatrix} 2.625 & 2.875 \\ 2.875 & 5.625 \end{pmatrix}$

$$X^{(1)} \mid X^{(2)} \sim N(\begin{pmatrix} 4.25 \\ 6.75 \end{pmatrix}, \begin{pmatrix} 2.625 & 2.875 \\ 2.875 & 5.625 \end{pmatrix})$$

2. Parte B.

2.1. (10 pts.)

Elabore un gráfico de dispersión con todas las variables continuas. Comente sobre posibles estructuras que den indicio de normalidad bivariada. Seleccione un par de variables con tal comportamiento y realice una prueba para verificar normalidad bivariada. Anexe los códigos en R usados.

Grafico de dispersión general

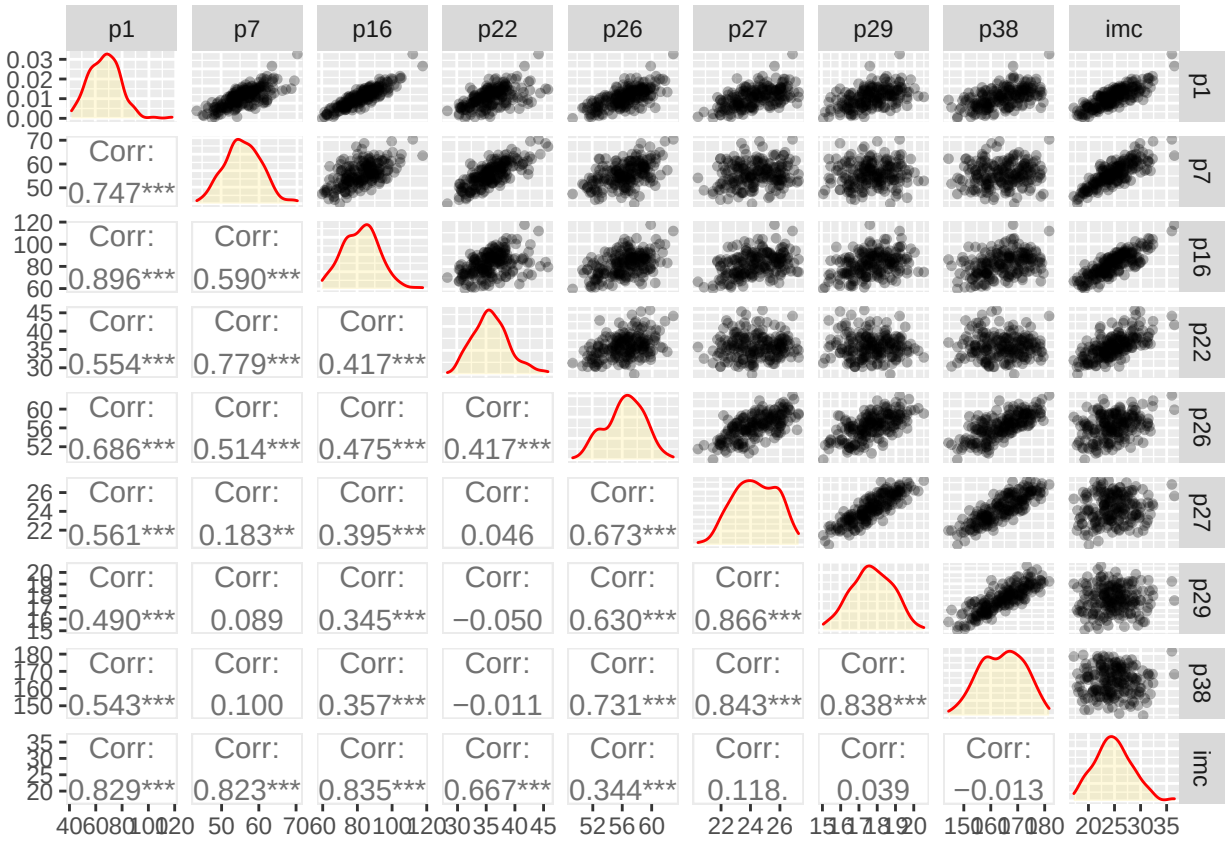

Figura 1: Grafico de dispersión general

En el grafico anterior se observa, la correlación entre las variables, sus debidas densidades y los graficos de dispersión. Visualizando las densidades, las variables perimetro abdominal mayor (p7), anchura de caderas (p22),longitud promedio de las manos (p29) y el IMC podrian distribuirse normal univariada. En los graficos de dispersión se busca una forma de elipse entre las variables, debido a que esto podria indicar normalidad, ademas, los graficos que tienden a ser mas lineales, darían indicios de no normalidad, teniendo encuenta lo anterior vemos que las variables p27 vs p38 y p22 vs p38 podrian tender a una normalidad bivariada, tambien vemos que las correlaciones entre estos pares de variables son 0.843 y

-0.011 respectivamente lo cual no nos proporciona mayor información para dar una idea de normalidad o no.

Luego de entender un poco el comportamiento de las variables basandonos en el gráfico anterior y de saber cuales de ellas pueden tener tendencia a una normalidad bivariada según lo explicado, se seleccionó la siguiente estructuras de variables: P16-P38 Perimetro abdominal cintura - Altura P7-P38 Perimetro muslo mayor - Altura P22-P38 Anchura de caderas - Altura p27-P38 Longitud de los pies - Altura imc-p1 Índice de masa corporal - Pes De las cuales se decició usar las variables p27 (Longitud de los pies) y p38(altura) para verificar su normalidad bivariada, empezando por la realización de un gráfico qqplot.

Q-Q plot de p27 vs p38

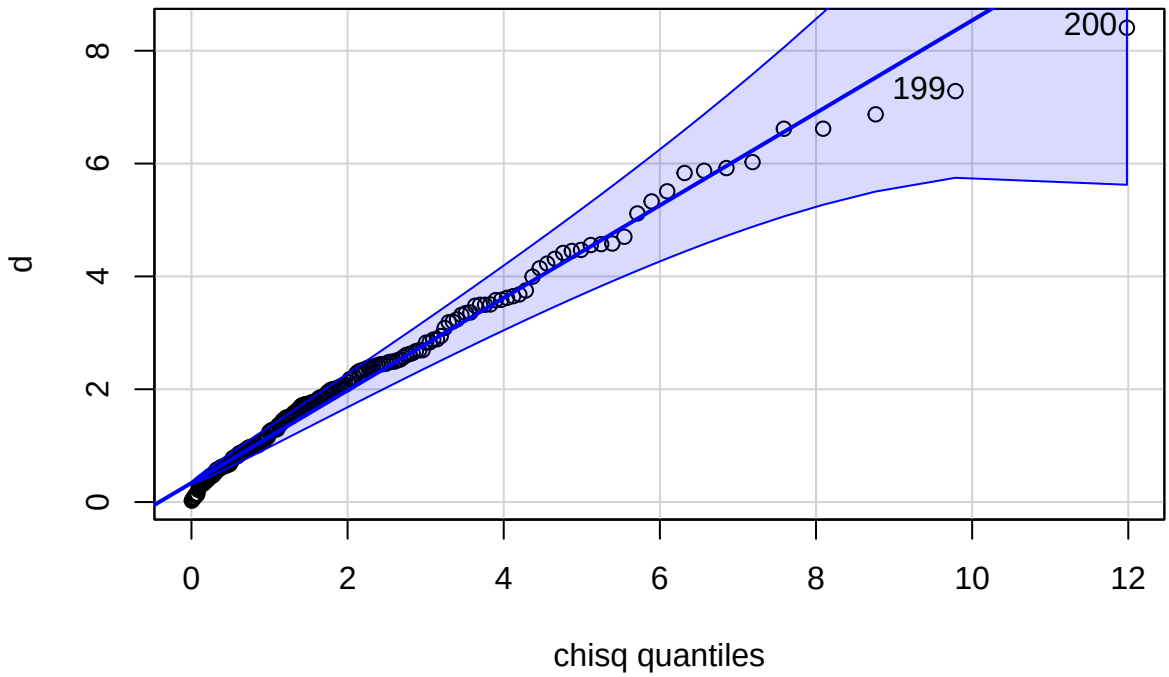

Figura 2: Q-Q plot de p27 vs p38

```
## [1] 200 199
```

Observando el qqplot se puede decir que los datos parecerían normales debido a que la mayoría de los datos se encuentran en la franja menos el datos 399 y 400 que serían las observaciones más anormales y podríamos pensar que son outliers, sin embargo generalmente tienen un buen ajuste, no obstante procedemos a realizar un test más confiable como lo es el de Shapiro-Wilk para estar seguros de la normalidad de los datos.

Prueba de normalidad bivariada (p27 vs p38)

```
mat <- as.matrix(datos2[,c("p27","p38")])
mshapiro.test(t(mat))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.99406, p-value = 0.6097
```

Luego de realizar el test de Shapiro-Wilk y al observar el p-valor = 0.1494 se concluye que no sé tiene evidencia suficiente para rechazar H0, por lo tanto se acepta normalidad bivariada entre las variables p27 y p38.

2.2. (20 pts.)

Del Gráfico anterior seleccione dos variables que a su juicio no muestren indicios de normalidad multivariada. Verifique si en efecto no hay normalidad bivariada. Posteriormente encuentre una transformación de Box y Cox para normalizar y después de hacer la respectiva transformación, verifique si se logró la normalidad bivariada en los datos transformados. Explique detalladamente cada paso del proceso y anexe el respectivo código en R usado.

Al igual que lo hicimos anteriormente lo primero que observamos fue nuestro gráfico de dispersión general, para de éste tener las posibles variables que aparentemente no serían normales bivariadas, al tener en cuenta lo anterior se busca en los graficos de disperción una tendencia más lineal con una amplitud más achatada y se obtuvo la siguiente estructura de posibles variables: p27-p38 Longitud de los pies - Altura p27-p29 Longitud de los pies - Longitud promedio pies p1-p16 Masa - Perimetro abdominal cintura De las cuales se consideró realizar el qqplot con las variables p1 y p16 ya que éstas tiene una tendencia lineal bastante pronunciada, a continuación el qqplot:

Grafico Chi-cuadrado p1 vs p16

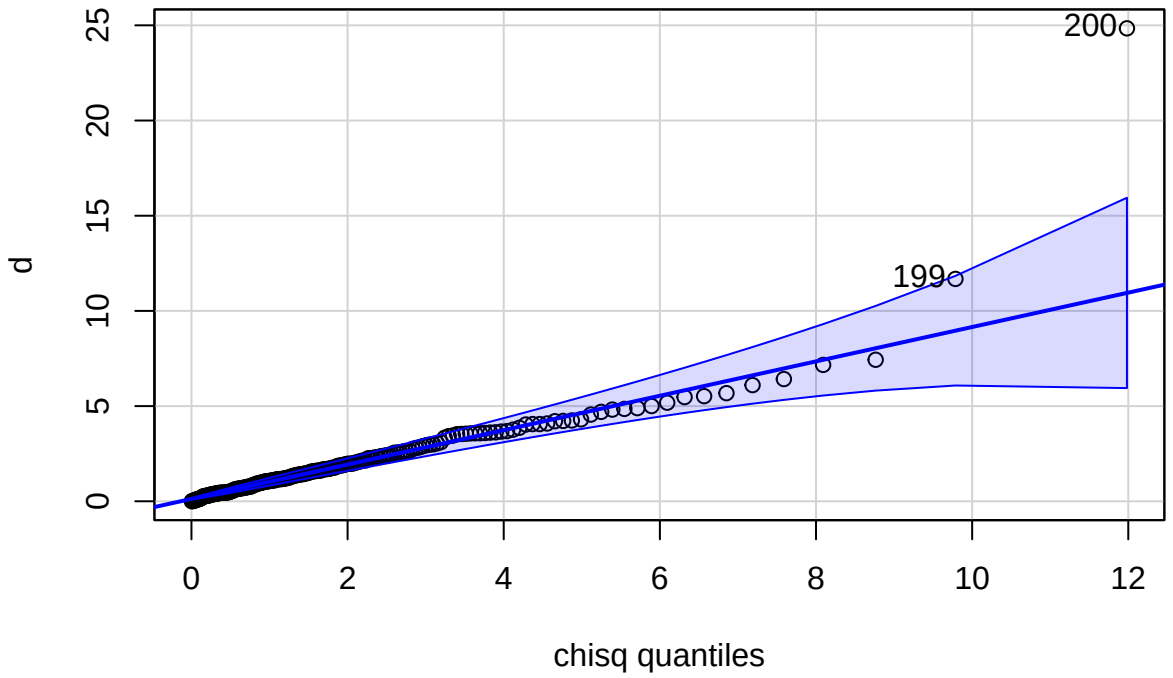

Figura 3: Q-Q plot de p1 vs p16

```
## [1] 200 199
```

Según el qqplot observado, éste presenta un ajuste bastante menor al de las variables anteriores (p27-p38), tiene algunas barrigas y presenta más datos fuera de la franja, lo que nos podría llevar a pensar que las variables p1 y p16 no serían candidatas a una normalidad bivariada, pero para verificar ésto necesitamos un test más robusto, para ésto utilizaremos de nuevo el test de Shapiro-Wilk.

Prueba de normalidad bivariada (p1 vs p16)

```
mat <- as.matrix(datos2[,c("p1","p7")])
mshapiro.test(t(mat))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.97097, p-value = 0.0003724
```

Luego de realizar el test de Shapiro se verifica que con un p-valor mucho menor a 0.05 que es el alpha propuesto, tenemos evidencia suficiente para rechazar H0, por lo tanto se concluye que las variables p1 y p16 no presentan una normalidad bivariada.

Transformación box cox

Como anteriormente se evidencio, las variables no son normales bivariadas. Para normalizar dichas variables, realizaremos una transformación de Box-Cox, la cual se define como:

$$x^{(\lambda)} = \begin{cases} \frac{x^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \ln x & \lambda = 0 \end{cases}$$

En donde $x^{(\lambda)}$ representa los datos transformados.

Con la funcón powerTransform de la libreria Car podremos hallar el mejor lambda para la transformación, la cual utiliza el enfoque de máxima verosimilitud de Box y Cox para seleccionar una transformación de una respuesta univariada o multivariada para normalidad

Para transformaciones a nivel multivariado, basta con realizar la misma transformación de Box-Cox para componente del vector. Por lo tanto, obtendremos un Lambda para cada una de las variables (p1 y p16)

```
a <- powerTransform(datos2[,c(1,3)],family="bcPower")
a
```

```
## Estimated transformation parameters
##        p1       p16
## 0.1188571 0.1822505
```

Luego de obtener el mejor Lambda, para p1 sera igual a 0.1188571 y para la variable p16 sera 0.1822505, luego realizamos la transformación para cuando $\lambda \neq 0$, obteniendo asi los datos transformados.

Función para la transformación de potencia Box-Cox

```
tranp1 <-as.matrix((dp1^0.1188571)-1)/0.1188571
tranp16 <- as.matrix((dp16^0.1822505)-1)/0.1822505
tran <- cbind(tranp1,tranp16)
```

Luego de que se le realiza la transformación box cox a los datos con las respectivas variables p1 y p26 lo que resta es verificar el supuesto de normalidad y para ésto se procede como anteriormente verificamos normalidad, empezando por un analisis e impresiones del grafica que brinda el qqplot, a continuación el qqplot de los datos que se les aplicó la transformación:

Grafico Chi-cuadrado p1 vs p16 para las variables transformadas

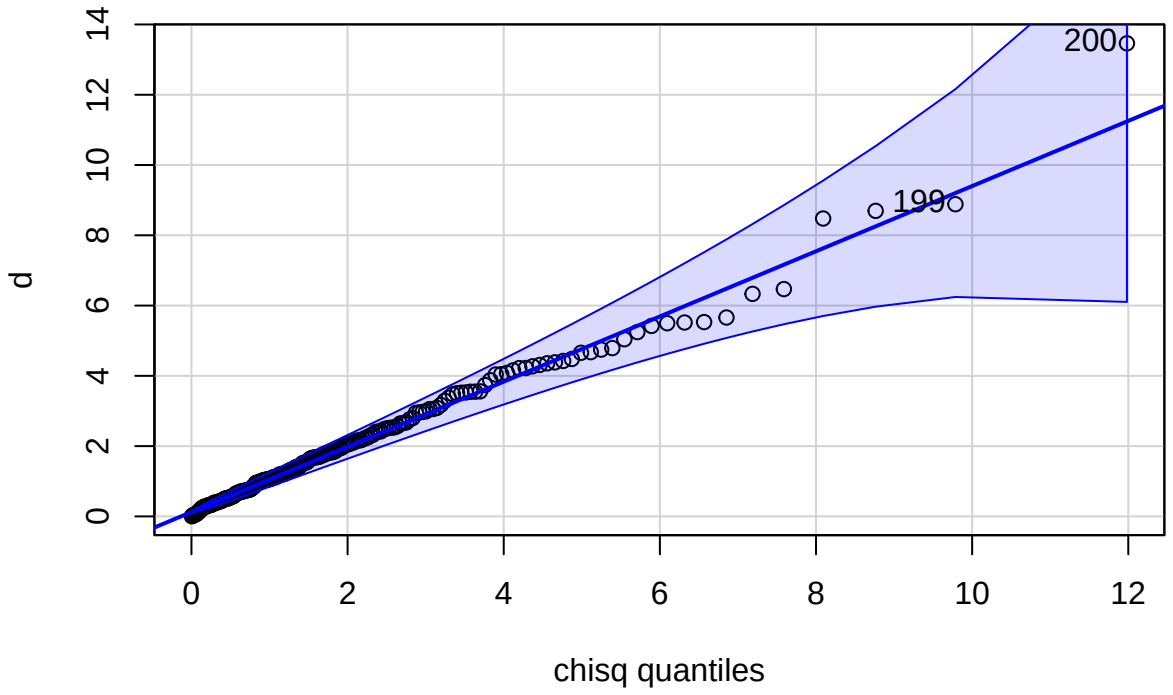

Figura 4: Q-Q plot de p1 vs p16 para las variables transformadas

```
## [1] 200 199
```

Observando la gráfica anterior vemos grandes diferencias con respecto a los datos normales debido a que se observa que todos los datos estan contenidos en la franja, no presenta tantas barrigas y además parecería que tiene un buen ajuste, no obstante no se podría asegurar normalidad bivariada, Y para ésto se procede a realizar una prueba de Shapiro-Wilk.

Prueba de normalidad bivariada para p1 vs p16 transformadas

```
mat <- as.matrix(tran)
mshapiro.test(t(mat))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.98813, p-value = 0.09437
```

Observando el test de Shapiro-Wilk realizado se observa un P-valor= 0.09437, debido a ésto se concluye que no se tiene evidencia suficiente para rechazar H0 y por lo tanto se comprueba normalidad bivariada entre las variables p1 y p16 despues de haber sido transformadas.

2.3. (10 pts.)

Considere el vector $\mathbf{X} = (P_1, P_7, P_{26}, P_{27}, P_{29}, P_{38}, IMC)'$ Realice una prueba para verificar si el vector X tiene una distribución Normal multivariada. Escriba las respectivas hipótesis y la conclusión obtenida. Anexe el respectivo código en R usado. Si no se logra la normalidad multivariada, repita el proceso discriminando por SEXO. Comente.

Creacción de vector X

```
x <- t(datos1[,c("p1","p7","p26","p27","p29","p38", "imc")])
```

Para evaluar Normalidad multivariada, se tienen las siguientes hipotesis a contrastar.

$$\begin{cases} H_0 : \boldsymbol{X} \sim N_7\left(\boldsymbol{\mu X}, \boldsymbol{\Sigma X}\right) \\ H_1 : \boldsymbol{X} \nsim N_7\left(\boldsymbol{\mu} X, \Sigma \boldsymbol{X}\right) \end{cases}$$

con un $\alpha = 0.05$

Prueba de normalidad multivariada para el vector X

```
mshapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.83417, p-value = 7.795e-14
```

Prueba de normalidad multivariada discriminada por la variable Sexo

Sean H el vector que contiene los datos discriminados para los Hombres y M el vector para las observaciones discriminadas para las Mujeres

```
sx <- datos1[,c("Sexo","p1","p7","p26","p27","p29","p38", "imc")]
H <- subset(sx, Sexo == "Hom", select = -c(Sexo))
M <- subset(sx, Sexo == "Muj", select = -c(Sexo))
```

Se realiza una prueba de normalidad multivariada para cada uno de los vectores (H y M), bajo las siguientes hipotesis. En el caso de del vector H

$$\begin{cases} H_0 : \boldsymbol{H} \sim N_7\left(\boldsymbol{\mu H}, \boldsymbol{\Sigma H}\right) \\ H_1 : \boldsymbol{H} \nsim N_7\left(\boldsymbol{\mu} H, \Sigma \boldsymbol{H}\right) \end{cases}$$

Prueba de normalidad multivariada para el vector H

```
mat <- as.matrix(H)
mshapiro.test(t(mat))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.74966, p-value = 1.319e-13
```

y la hipotesis para el vector M

$$\begin{cases} H_0 : \boldsymbol{M} \sim N_7\left(\boldsymbol{\mu M}, \boldsymbol{\Sigma M}\right) \\ H_1 : \boldsymbol{M} \nsim N_7\left(\boldsymbol{\mu} M, \Sigma \boldsymbol{M}\right) \end{cases}$$

Prueba de normalidad multivariada para el vector M

```
mat <- as.matrix(M)
mshapiro.test(t(mat))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.91369, p-value = 0.0001408
```

Basados en el test de Shapiro-Wilk realizado anteriormente para los vectores H y M se observa un P-valor de 1.319e-13 y de 0.0001408 respectivamente, se observa que estos valores son inferiores al nivel de significancia de $\alpha = 0.05$ y se tiene evidencia suficiente para rechazar H0, se puede concluir que los vectores H y M no presentan normalidad multivariada y se podria decir que discriminar el vector X por la variable Sexo no presenta un cambio significativo.

3. Parte C.

3.1. (20 pts.)

Con la base de datos Acopla, se seleccionaron de manera conveniente 10 variables. Usando un criterio de discriminación para cada variable, el experto clasificaba el sujeto en 0 o 1 (0 indica que no cumple la condición y 1 que la cumple). Los resultados para 7 sujetos se muestran a continuación:

Sujeto	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
1	0	0	0	0	0	1	0	1	0	0
2	0	0	1	1	0	1	0	0	0	0
3	1	0	1	1	0	1	1	1	1	1
4	1	0	1	1	0	0	1	1	1	1
5	1	1	1	1	0	1	1	1	1	1
6	0	1	0	1	0	1	0	1	1	0
7	1	0	0	1	1	0	1	1	0	0

Usando esta información halle la matriz de similaridades para los 7 sujetos usando el índice de Sokal y Michener. Indique cuales son los dos sujetos más parecidos y por qué. Explique paso a paso cómo obtiene dicha matriz.

R/ Para la realización de la matriz de similaridades usando el índice de Sokal y Michener primero debemos obtener cada componente de la matriz de similaridad y esto se hace con otra matriz 2x2 con coeficientes a, b, c, d en la cual se comparan dos sujetos i y j, además el coeficiente a es el número donde tanto el sujeto i como el j no cumplen la condición en cada variable, el coeficiente b es el número donde el i cumple la condición, pero el j no, el coeficiente c sería el contrario del b y por último el coeficiente d es el número de variables donde tanto el sujeto i como el j cumplen la condición. la matriz para la obtención de cada una de las entradas de la matriz de similaridad sería la siguiente:

	Sujeto i		
		0	1
Sujeto j	0	a	b
	1	c	d

Ahora que sabemos que significa cada uno de los componentes de la matriz 2x2 veremos cómo obtener cada una de las entradas, para no confundirnos los denotaremos cómo S(i, j) el cual se calcula cómo S(i, j) = (a+b)/p, dónde p es el número de variables, en éste caso p=10.

Ahora que sabemos cómo realizar cada entrada de la matriz de similaridad procedemos a hacerlo, cómo tenemos 7 sujetos sabemos que obtendremos una matriz 7x7, haremos el procedimiento de la S(1, 2) que sería lo mismo que S(2, 1)

La matriz de matches sería la siguiente:

$S(1,2) =$	1	1
	2	6

Y la entrada (1, 2) de la matriz de similaridades sería S(1, 2)= (1+6)/10= 7/10 = 0.7

Así mismo lo hacemos con las otras entradas, las cuales serían las siguientes: S(1, 3)= (2+2)/10= 4/10= 0.4 S(1, 4)= (1+2)/10= 3/10= 0.3 S(1, 5)= (2+1)/10= 3/10= 0.3 S(1,6)= (2+5)/10= 7/10= 0.7 S(1,7)= (1+4)/10= 5/10= 0.5 S(2,3)= (3+2)/10= 5/10= 0.5 S(2,4)= (2+2)/10= 4/10= 0.4 S(2,5)= (3+1)/10= 4/10= 0.4 S(2,6)= (2+4)/10= 6/10= 0.6 S(2,7)= (2+4)/10= 6/10= 0.6 S(3,4)= (7+2)/10= 9/10= 0.9 S(3,5)= (8+1)/10= 9/10= 0.9 S(3,6)= (4+1)/10= 5/10= 0.5 S(3,7)= (4+1)/10= 5/10= 0.5 S(4,5)= (7+1)/10= 8/10= 0.8 S(4,6)=

(3+1)/10= 4/10= 0.4 S(4,7)= (4+2)/10= 6/10= 0.6 S(5,6)= (5+1)/10= 6/10= 0.6 S(5,7)= (4+0)/10= 4/10= 0.4 S(6,7)= (2+2)/10= 4/10= 0.4

Ahora la matriz de similaridades:

Matriz de similaridades						
1	0.7	0.4	0.3	0.3	0.7	0.5
0.7	1	0.5	0.4	0.4	0.6	0.6
0.4	0.5	1	0.9	0.9	0.5	0.5
0.3	0.4	0.9	1	0.8	0.4	0.6
0.3	0.4	0.9	0.8	1	0.6	0.4
0.7	0.6	0.5	0.4	0.6	1	0.4
0.5	0.6	0.5	0.6	0.4	0.4	1

Conclusión: Según la matriz obtenida se observa que se encuentra una gran similitud entre el sujeto 3 con el 4 y el mismo sujeto 3 con el 5, los dos con un valor de 0.9, se presenta ésta gran similitud debido a que de las 10 variables brindadas hacen match en 9 de ellas lo que representa un 90 por ciento de similitud entre los sujetos 3 y 4, además del 3 y 5.

4. Anexos

En el siguiente link se redireccionara a un repositorio donde encuentra todo el trabajo y los codigos empleados para su solución: