

## **Trabajo 2**

**Ivan Santiago Rojas Martinez**

Estudiante de Pregrado en Estadística

Docente

**Rene Iral Palomino**

Asignatura

**Introducción al Análisis Multivariado**



Sede Medellín  
Octubre 21 de 2023

# 1 Parte A

## 2 Parte B

Para todos los efectos el vector  $X = (P_1, P_7, P_{16}, P_{22}, P_{25}, P_{27}, P_{29}, P_{38})$  contiene las variables continuas de su base de datos. Por notación sea el respectivo  $\mu = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6, \mu_7, \mu_8)$  vector de medias y  $\Sigma$  su matriz de covarianzas.

Se procede a tomar la muestra aleatoria con la cedula **1020479466** y a seleccionar las variables numéricas.

```
library(splitstackshape)
uno <- read.table("Data/base.txt", header = TRUE)
genera <- function(cedula){
  set.seed(cedula)
  aux <- stratified(uno, "CAT_IMC", 200/2100)
  aux
}

datos <- genera(1020479466)
x <- datos %>% select(P1, P7, P16, P22, P25, P27, P29, P38)
```

- 1. (10 pts.)** Sea  $\mu_0 = (66.1, 58, 81.6, 37, 47, 25, 19.2, 167)'$ . Pruebe la hipótesis:  
 $H_0: \mu = \mu_0$ . Debe especificar todas las condiciones y elementos para probar esta hipótesis.  
 Anexe los códigos en R usados.

Primero procederemos a verificar si el vector de variables se distribuye normal por medio de la prueba estadística Shapiro-Wilk de normalidad multivariada.

```
library(mvnormtest)
mu_0 <- c(66.1, 58, 81.6, 37, 47, 25, 19.2, 167)
mshapiro.test(t(as.matrix(x)))
```

```
##
## Shapiro-Wilk normality test
##
## data: Z
## W = 0.9307, p-value = 4.09e-08
```

Observando un  $ValorP = 4.09 \times 10^{-8}$ , podemos rechazar la hipótesis nula con un nivel de significancia de  $\alpha = 0.05$  lo que nos permite concluir que  $X$  no cumple normalidad multivariada. Basado en el **Teorema del limite central** sabemos que si se tiene:

$$\bar{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \text{ y } \mathbf{Z}_n = \sqrt{n} (\bar{\mathbf{X}}_n - \boldsymbol{\mu})$$

Luego:

$$\mathbf{Z}_n \xrightarrow{d} N_p(\mathbf{0}, \Sigma).$$

con  $\Sigma > 0$ :

$$\tilde{\mathbf{Z}}_n = \Sigma^{-\frac{1}{2}} \sqrt{n} (\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} N_p(\mathbf{0}, I_p)$$

y

$$n (\bar{\mathbf{X}}_n - \boldsymbol{\mu})' S^{-1} (\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} \chi^2(p)$$

Bajo  $H_0$  cierta. El estadístico de prueba es:

$$\chi_0^2 = n (\bar{X} - \mu_0)' S^{-1} (\bar{X} - \mu_0)$$

Se define las siguientes pruebas de hipótesis:

$$H_0 : \mu = \mu_0 \text{ vs } H_a : \mu \neq \mu_0$$

Se procede a hallar el vector de medias muestral  $\bar{X}$  y la matriz de covarianzas muestral  $S$  en R.

```
xbar <- colMeans(x)
s <- cov(x)
n <- nrow(x)
p <- length(x)
mu_0 <- as.matrix(c(mu_0))
chi_0 <- as.numeric(n*(t(xbar-mu_0)) %*%solve(s) %*%(xbar-mu_0))
chi_0
```

```
## [1] 1603.695
```

Se plantea una región de rechazo de  $H_0$  dada por:

$$X_0^2 > X_{\alpha}^2(p)$$

```
qchisq(0.05, p, lower.tail = F)
```

```
## [1] 15.50731
```

Dado que la prueba nos arroja  $X_0^2 = 1603.695 > X_{0.05}^2(8) = 15.50731$  con un nivel de significancia de  $\alpha = 0.05$ . Se rechaza  $H_0$ , lo que nos permite concluir que existen diferencia entre el vector  $\mu$  y  $\mu_0$ .

- 2. (15 pts.)** Repita la hipótesis anterior, pero discriminando por Género. ¿Observa algún cambio en la conclusión? Debe especificar todas las condiciones y elementos para probar esta hipótesis. Anexe los códigos en R usados.

## 2.1 Análisis discriminado sexo Masculino:

Se procede a filtra los datos por el genero masculino.

```
hom <- datos %>% filter(SEXO == "Hom") %>% select(P1, P7, P16, P22, P25, P27, P29, P38)
```

Se plantea sus hipótesis de normalidad multivariada discriminando por el genero de hombres.

$$H_o : X_H \sim N_p(\mu, \Sigma) \text{ VS } H_a : X_H \not\sim N_p(\mu, \Sigma)$$

```
mshapiro.test(t(as.matrix(hom)))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.92315, p-value = 1.91e-06
```

Observando un  $ValorP = 1.91 \times 10^{-6}$ , podemos rechazar la hipótesis nula con un nivel de significancia de  $\alpha = 0.05$  lo que nos permite concluir que el vector de variables  $X$  para los hombres no se distribuye normal, por tanto basado en el teorema del limite central tenemos que el estadístico de prueba es:

$$\chi_0^2 = n \left( \bar{X}_{hom} - \mu_0 \right)' S_{hom}^{-1} \left( \bar{X}_{hom} - \mu_0 \right) \xrightarrow{d} \chi_{\alpha}^2(\mathbf{P})$$

Con sus respectivas hipotesis:

$$H_0 : \mu_{homb} = \mu_0 \text{ vs } H_a : \mu_{homb} \neq \mu_0$$

```
xbar <- colMeans(hom)
s <- cov(hom)
n <- nrow(hom)
p <- length(hom)
mu_0 <- as.matrix(c(mu_0))
chi_0 <- as.numeric(n*(t(xbar-mu_0)) %*%solve(s) %*%(xbar-mu_0))
chi_0
```

```
## [1] 1403.088
```

Se plantea una región de rechazo de  $H_0$  dada por:

$$X_0^2 > X_{\alpha}^2(p)$$

```
qchisq(0.05, p, lower.tail = F)
```

```
## [1] 15.50731
```

Dado que la prueba nos arroja  $X_0^2 = 1403.088 > X_{0.05}^2(8) = 15.50731$  con un nivel de significancia de  $\alpha = 0.05$ . Se rechaza  $H_0$ , lo que nos permite concluir que existen diferencia entre el vector  $\mu_{homb}$  y  $\mu_0$ .

## 2.2 Análisis discriminado sexo Femenino:

Se procede a filtra los datos por el genero masculino.

```
muj <- datos %>% filter(SEX0 == "Muj") %>% select(P1, P7, P16, P22, P25, P27, P29, P38)
```

Se plantea sus hipótesis de normalidad multivariada discriminando por el genero de mujeres.

$$H_o : X_M \sim N_p(\mu, \Sigma) \text{ VS } H_a : X_M \not\sim N_p(\mu, \Sigma)$$

```
mshapiro.test(t(as.matrix(muj)))
```

```
##
## Shapiro-Wilk normality test
##
## data:  Z
## W = 0.93582, p-value = 0.001285
```

Observando un  $ValorP = 0.001285$ , podemos rechazar la hipótesis nula con un nivel de significancia de  $\alpha = 0.05$  lo que nos permite concluir que el vector de variables  $X$  para las mujeres no se distribuye normal, por tanto basado en el teorema del limite central tenemos que el estadístico de prueba es:

$$\chi_0^2 = n \left( \bar{X}_{muj} - \mu_0 \right)' S_{muj}^{-1} \left( \bar{X}_{hom} - \mu_0 \right) \xrightarrow{d} \chi_{\alpha}^2(P)$$

Con sus respectivas hipotesis:

$$H_0 : \mu_{muj} = \mu_0 \text{ vs } H_a : \mu_{muj} \neq \mu_0$$

```
xbar <- colMeans(muj)
s <- cov(muj)
n <- nrow(muj)
p <- length(muj)
mu_0 <- as.matrix(c(mu_0))
chi_0 <- as.numeric(n*(t(xbar-mu_0))%*%solve(s)%*%(xbar-mu_0))
chi_0
```

```
## [1] 2773.412
```

Se plantea una región de rechazo de  $H_0$  dada por:

$$X_0^2 > X_\alpha^2(p)$$

```
qchisq(0.05, p, lower.tail = F)
```

```
## [1] 15.50731
```

Dado que la prueba nos arroja  $X_0^2 = 2773.412 > X_{0.05}^2(8) = 15.50731$  con un nivel de significancia de  $\alpha = 0.05$ . Se rechaza  $H_0$ , lo que nos permite concluir que existen diferencia entre el vector  $\mu_{muj}$  y  $\mu_0$ .

- 3. (15 pts.)** Para el mismo vector  $\mathbf{X}$ , definido anteriormente, se sabe que  $\mathbf{X} \sim N_8(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Pruebe la hipótesis:  $H_0: 2\mu_1 - \mu_2 + 3\mu_4 + \mu_7 - \mu_6 = \mu_8$  y  $3\mu_2 - 4\mu_3 + 2\mu_5 + 2\mu_6 = 0$ . Debe especificar todas las condiciones y elementos para probar esta hipótesis. Anexe los códigos en R usados.

Primero se procede a plantear las hipótesis para los contrastes, como:

$$H_0 : C\boldsymbol{\mu} = \boldsymbol{\gamma} \quad VS \quad C\boldsymbol{\mu} \neq \boldsymbol{\gamma}$$

Escritas de otra manera como:

$$H_0 : 2\mu_1 - \mu_2 + 3\mu_4 + \mu_7 - \mu_6 = \mu_8 \quad y \quad 3\mu_2 - 4\mu_3 + 2\mu_5 + 2\mu_6 = 0$$

Donde:

$$C = \begin{pmatrix} 2 & -1 & 0 & 3 & 0 & -1 & 1 & -1 \\ 0 & 3 & -4 & 0 & 2 & 2 & 0 & 0 \end{pmatrix}$$

Estadístico de prueba:

$$T_0^2 = n(C\bar{\mathbf{X}} - \boldsymbol{\gamma})' (CSC')^{-1} (C\bar{\mathbf{X}} - \boldsymbol{\gamma})$$

Bajo  $H_0$  cierto. Se rechaza si  $\frac{n-k}{(n-1)k} T_0^2 > f_\alpha(k, n-k)$

Luego tenemos un  $\frac{199-2}{2(199-1)} T_0^2 = 50.57493 > f_{0.05}(2, 197) = 3.041753$  y con una significancia de  $\alpha = 0.05$  lo que nos permite rechazar  $H_0$  concluyendo que  $2\mu_1 - \mu_2 + 3\mu_4 + \mu_7 - \mu_6 \neq \mu_8$  y  $3\mu_2 - 4\mu_3 + 2\mu_5 + 2\mu_6 \neq 0$

- 4. (15 pts.)** Para el mismo vector  $\mathbf{X}$ , definido anteriormente, sean  $\boldsymbol{\Sigma}_D$ ,  $\boldsymbol{\Sigma}_N$  y  $\boldsymbol{\Sigma}_O$  las respectivas matrices de covarianzas para el grupo de Delgados, Normales y Obesos respectivamente. Suponga que el vector  $\mathbf{X}$  tiene una distribución Normal multivariada, para los tres grupos definidos en la variable **CAT\_IMC**. Determine si la estructura de covarianzas es similar en los tres grupos. Debe especificar todas las condiciones y elementos para probar esta hipótesis. Anexe los códigos en R usados.

Se plantean las respectivas Hipótesis como:

$$H_0 : \Sigma_{obeso} = \Sigma_{normal} = \Sigma_{delgado} \text{ vs } H_a : \Sigma_{obeso} \neq \Sigma_{normal} \neq \Sigma_{delgado}$$

```
delg <- datos %>% filter(CAT_IMC == 'Delgado')
norm <- datos %>% filter(CAT_IMC == 'Normal')
obes <- datos %>% filter(CAT_IMC == 'Obeso')

delg <- delg[, 2:9]
norm <- norm[, 2:9]
obes <- obes[, 2:9]

n1 <- nrow(delg)
n2 <- nrow(norm)
n3 <- nrow(obes)
p <- ncol(delg)
g <- 3

# Varianzas muestrales
s1 <- matrix(var(delg), ncol=8)
s2 <- matrix(var(norm), ncol=8)
s3 <- matrix(var(obes), ncol=8)

# Matriz ponderada
sum_ni <- ((n1-1)+(n2-1)+(n3-1))
sum_inv_ni <- (1/(n1-1))+(1/(n2-1))+(1/(n3-1))
k <- (2*p^2 + 3*p-1)/(6*(p+1)*(g+1))

sp <- ((n1-1)*s1+(n2-1)*s2+(n3-1)*s3)/sum_ni

# Estadístico M
M <- sum_ni*log(det(sp))-((n1-1)*log(det(s1))+(n2-1)*log(det(s2))+(n3-1)*log(det(s3)))
u <- (sum_inv_ni -(1/sum_ni))*k
C <- (1-u)*M

#Chi^2
alpha <- 0.05
gl <- p*(p+1)*(g-1)/2
q <- qchisq(alpha,gl)
```

Se obtiene un  $C = 121.4472 > \chi_{0.05}^2 = 53.46233$  con una significancia de  $\alpha = 0.05$  nos permite rechazar  $H_0$  y concluir que la estructura de covarianza es diferente para los tres grupos  $\Sigma_{obeso} \neq \Sigma_{normal} \neq \Sigma_{delgado}$ .

**5. (20 pts.)** Para el mismo vector  $\mathbf{X}$ , definido anteriormente, sean  $\boldsymbol{\mu}_H$  y  $\boldsymbol{\mu}_M$  los respectivos vectores de medias para Hombres y mujeres, respectivamente y sean  $\boldsymbol{\Sigma}_H$  y  $\boldsymbol{\Sigma}_M$  las respectivas matrices de covarianzas para Hombres y Mujeres, respectivamente. Determine si el vector  $\mathbf{X}$  es suficiente para poder discriminar entre Hombres y Mujeres. Debe especificar todas las condiciones y elementos para probar esta hipótesis. Anexe los códigos en R usados.

Como el vector  $X$  no se distribuye normal multivariado, entonces un estimador insesgado para  $\mu_h - \mu_m$  es:  $\bar{X} - \bar{Y}$ . Si  $n - p$  y  $m - p$  son grandes el TLC nos garantiza que:

$$\Sigma_H \sqrt{n}(\bar{X} - \mu_H) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{I}_p) \text{ y } \Sigma_M \sqrt{n}(\bar{Y} - \mu_M) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{I}_p)$$

Por Slutsky tenemos:

$$S^{-\frac{1}{2}}[(\bar{X}_n - \bar{Y}_m) - (\mu_H - \mu_M)] \xrightarrow{d} N_p(\mathbf{0}, \mathbf{I}_p)$$

Bajo  $H_0$  cierta se tiene que:

$$X_C = [(\bar{X}_n - \bar{Y}_m) - \delta_0] \left[ \frac{1}{n} S_1 + \frac{1}{m} S_2 \right]^{-1} [(\bar{X}_n - \bar{Y}_m) - \delta_0] \xrightarrow{d} \chi^2_\alpha(p)$$

Se rechaza  $H_0$  si  $X_c > \chi^2_\alpha(2p)$

Las respectivas hipótesis son:

$$H_0 : \mu_h = \mu_m \text{ vs } H_a : \mu_h \neq \mu_m$$

Que es equivalente a probar:

$$H_0 : \mu_h - \mu_m = \delta_0 \text{ vs } H_a : \mu_h - \mu_m \neq \delta_0$$

con  $\delta_0 = (0, 0, 0, 0, 0, 0, 0, 0)$

```
muh <- colMeans(muj)
mum <- colMeans(hom)
varh <- cov(hom)
varm <- cov(muj)
nh <- nrow(hom)
nm <- nrow(muj)
r_0 <- cbind(c(0,0,0,0,0,0,0,0))
dif <- muh-mum
S <- (1/nm)*varm+(1/nh)*varh
Xc <- as.numeric(t(dif-r_0)%*%solve(S)%*%(dif-r_0))
f_alpha <- qchisq(0.05,8,lower.tail = F)
```

Se obtiene un  $X_C = 1050.692 > \chi^2_{0.05}(8) = 16.91898$  con una significancia de  $\alpha = 0.05$  nos permite rechazar  $H_0$  y concluir que el vector  $X$  no es suficiente para poder discriminar entre Hombres y Mujeres.



- 6. (25 pts.)** Usando la matriz de covarianzas muestral, calcule los vectores y valores propios de dicha matriz. Elabore el respectivo Scree-plot y comente sobre la variabilidad explicada por las componentes principales. Considere la primera componente principal, ¿Cuáles variables tienen mayor peso en su definición? ¿Puede dar alguna interpretación a dicha componente? Comente. Determine si el valor propio más pequeño de  $\Sigma$  es significativamente diferente de cero. Justifique su respuesta.

```
var_m <- cov(x) %>% as.matrix()
eigen_vec <- eigen(var_m)
```

### Vectores Propios

```
eigen_vec$vectors
```

```
##           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## [1,] -0.69429935  0.12777297  0.27596100  0.5575169474 -0.293845146 -0.16395845
## [2,] -0.17962978  0.21255066  0.61708787 -0.1393165671  0.627941437  0.35426663
## [3,] -0.53137074  0.45370428 -0.58395712 -0.3801672357  0.130019459  0.09624298
## [4,] -0.06166021  0.13463489  0.41250307 -0.6300457101 -0.287055930 -0.57271477
## [5,] -0.09881243 -0.12512280  0.15828236 -0.2823070789 -0.605159637  0.70931763
## [6,] -0.07308015 -0.11120975 -0.03730454 -0.0006601938  0.018550648 -0.03565882
## [7,] -0.05013086 -0.08680225 -0.02802493  0.0121589554 -0.006365356 -0.00912413
## [8,] -0.42650557 -0.82399284 -0.06821325 -0.2200805778  0.231161384 -0.07656085
##           [,7]      [,8]
## [1,]  0.037294644  0.004962479
## [2,] -0.049751603 -0.007635169
## [3,]  0.005949071 -0.007832278
## [4,] -0.021108164 -0.010306038
## [5,] -0.013830754  0.016994832
## [6,] -0.784465002  0.603240086
## [7,] -0.594462125 -0.797188838
## [8,]  0.163374095 -0.007214529
```

Los vectores propios indican la influencia de cada variable en la componente. Las variables con un valor absoluto más alto en el vector propio tienen una mayor influencia en la componente principal. En nuestro caso, las variables con mayor influencia son:

P1(masa), P7(Perímetro del muslo), P16(Perímetro abdominal cintura), P38 (Altura),

### Valores Propios

Como tenemos 8 variables de interés nuestra matriz de covarianza es de  $8 \times 8$  por lo tanto tendremos 9 valores propios los cuales son:

$\hat{\lambda}_1 = 271.1646218$ ;  $\hat{\lambda}_2 = 61.6813232$ ;  $\hat{\lambda}_3 = 22.3281270$ ;  $\hat{\lambda}_4 = 4.1034057$ ;  $\hat{\lambda}_5 = 2.8346277$ ;  $\hat{\lambda}_6 = 2.3134986$ ;

$\hat{\lambda}_7 = 0.6964915$ ;  $\hat{\lambda}_8 = 0.1826296$ ;

```
eigen_values <- eigen_vec$values
eigen_values
```

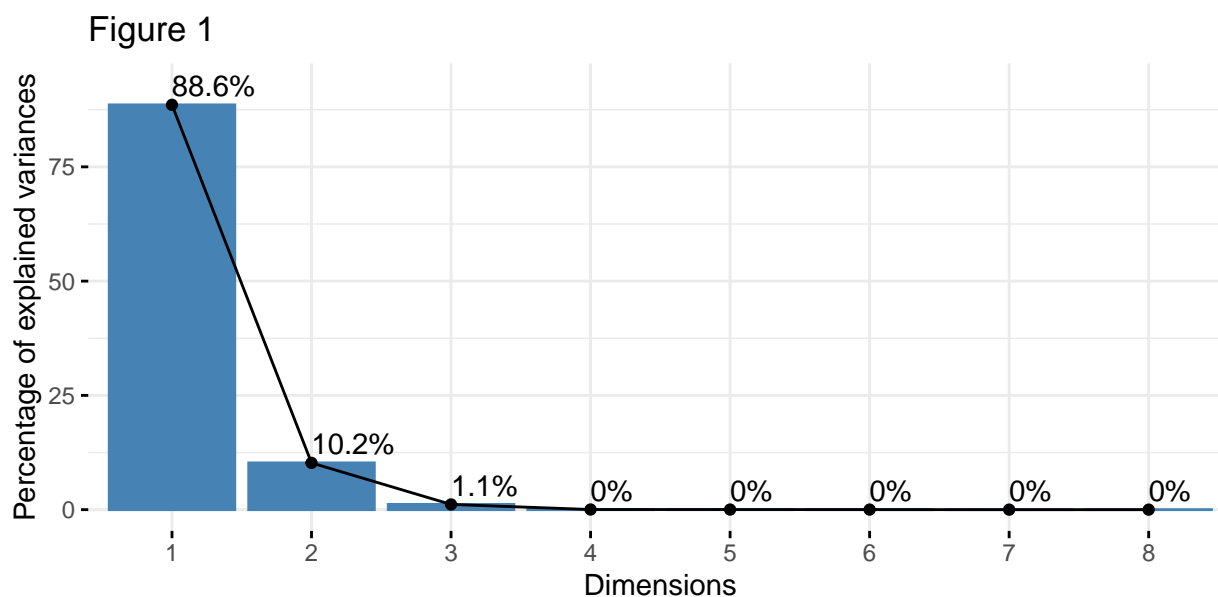
```
## [1] 271.1646218 61.6813232 22.3281270 4.1034057 2.8346277 2.3134986
## [7] 0.6964915 0.1826296
```

### Componentes Principales

```
pca <- prcomp(var_m)
summary(pca)
```

```
## Importance of components:
##
##          PC1      PC2      PC3      PC4      PC5      PC6      PC7
## Standard deviation 68.1144 23.1479 7.73823 1.23232 1.05365 0.84392 0.07997
## Proportion of Variance 0.8857 0.1023 0.01143 0.00029 0.00021 0.00014 0.00000
## Cumulative Proportion 0.8857 0.9879 0.99936 0.99965 0.99986 1.00000 1.00000
##
##          PC8
## Standard deviation 5.712e-17
## Proportion of Variance 0.000e+00
## Cumulative Proportion 1.000e+00
```

```
fviz_eig(pca, addlabels = TRUE, ylim = c(0, 93), main="Figure 1")
```



Como se puede observar en el gráfico anterior y en el resumen numérico de las componentes principales, la primera componente principal explica el 88.6%, y la segunda componente explica el 10.2% de la variabilidad total. En conjunto, las dos primeras componentes principales explican el 98.8% de la variabilidad total. La primera componente principal se relaciona con las variables **P1(masa)**, **P7(Perímetro del muslo)**, **P16(Perímetro abdominal cintura)** **P38 (Altura)**,