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

Índice

Índice de figuras

Índice de cuadros

Trabajo 2

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 Σ 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 $\boldsymbol{\mu}_0 = (66.1, 58, 81.6, 37, 47, 25, 19.2, 167)'$. Pruebe la hipótesis: $H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$. 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:
$\overline{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i$ y $\mathbf{Z}_n = \sqrt{n}\left(\overline{\mathbf{X}}_n - \boldsymbol{\mu}\right)$
Luego:

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

con $\Sigma > 0$:

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

y

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

Bajo H_0 cierta. El estadístico de prueba es:

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

Se define las siguientes pruebas de hipótesis:

$$H_0 : \mu = \mu_0 \; vs \; H_a : \mu \neq \mu_0$$

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 μ y μ_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.

$$H_o : X_H \sim N_p(\mu, \Sigma) \; VS \; H_a : X_H \nsim N_p(\mu, \Sigma)$$

```
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:

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

Con sus respectivas hipotesis:

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

```
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 μ_{homb} y μ_0.

2.2 Análisis discriminado sexo Femenino:

Se procede a filtra los datos por el genero masculino.

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

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

$$H_o : X_M \sim N_p(\mu, \Sigma) \; VS \; H_a : X_M \nsim N_p(\mu, \Sigma)$$

```
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:

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

Con sus respectivas hipotesis:

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

```r
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:

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

```r
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 μ_{muj} y μ_0.

3. **(15 pts.)** Para el mismo vector $\boldsymbol{X}$, definido anteriormente, se sabe que $\boldsymbol{X} \sim N_8(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. 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:

$$H_0 : C\mu = \gamma \quad VS \quad C\mu \neq \gamma$$

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:

$$T_0^2 = n(C\overline{\mathbf{X}} - \gamma)' \left(CSC'\right)^{-1} (C\overline{\mathbf{X}} - \gamma)$$

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

4. (**15 pts**.) Para el mismo vector $\boldsymbol{X}$, 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 $\boldsymbol{X}$ 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.

5. (**20 pts**.) Para el mismo vector $\boldsymbol{X}$, 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 $\boldsymbol{X}$ 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.

El estadístico de razón de verosimilitud que nos permite probar H_0 se define como:

$$\lambda = \prod_{i=1}^{g} \left(\frac{|S_i|}{|S_p|} \right)^{\frac{(n_i-1)}{2}}$$

Donde S_i son la matrices de covarianzas muéstrales para el genero masculino y el genero femenino. Y $S_p = \frac{1}{\sum_{i=1}^{g}(n_i-1)} \left[(n_1 - 1) S_1 + n_2 - 1) S_2 + \cdots + (n_g - 1) S_g \right]$

$$M = -2ln\lambda$$

$$M = \left[\sum_{i=1}^{g} (n_i - 1) \right] \ln(|\ S_p\ |) - \sum_{i=1}^{g} (n_i - 1) \ln(|S_i|)$$

$$u = \left[\sum_{i=1}^{g} \frac{1}{(n_i - 1)} - \frac{1}{\sum_{i=1}^{g}(n_i - 1)} \right] \left(\frac{2p^2 + 3p - 1}{6(p+1)(g+1)} \right)$$

Bajo H_0 verdadero se cumple:

$$C = (1-u)M \xrightarrow{d} \chi^2 \left(\frac{1}{2} p(p+1)(g-1) \right)$$

Con un $\alpha = 0.05$ se rechaza H_0 si $C > \chi^2_{0.05} \left(\frac{1}{2} p(p+1)(g-1) \right)$

Su respectiva hipótesis:

$$H_0 : \Sigma_{hom} = \Sigma_{muj}\ vs\ H_a : \Sigma_{hom} \neq \Sigma_{muj}$$

```r
nh <- nrow(hom)
nm <- nrow(muj)
varh<- cov(hom)
varm<- cov(muj)
sum_n <- (nh-1)*varh+(nm-1)*varm
sum_d <- (nh-1)+(nm-1)
sp <- sum_n/sum_d
#Estadistico de prueba
p <- 8
M <- sum_d*log(det(sp))-((nh-1)*log(det(varh))+(nm-1)*log(det(varm)))
sum_inv_ni <- (1/(nm-1))+(1/(nh-1))
k <- ((2*(p^2)+3*p-1)/(6*(p+1)*(2+1)))
u <- (sum_inv_ni-(1/sum_d))*k
C <- (1-u)*M

gl <- (p*(p+1)*(2-1))/2
alpha <- 0.05
q_alpha <- qchisq(0.05,gl,lower.tail = T)
```

Luego tenemos un $C = 88.22488 > \chi^2_{0.05} = 23.26861$ y con una significancia de $\alpha = 0.05$ lo que nos permite rechazar H_0 concluyendo que $\Sigma_{hom} \neq \Sigma_{muj}$.

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 $\mathbf{\Sigma}$ es significativamente diferente de cero. Justifique su respuesta.