```r
library(tidyverse)
library(knitr)
library(mvnormtest)
```

Trabajo 2

Ivan Santiago Rojas Martinez

Estudiante de Pregrado en Estadística

Docente

Rene Iral Palomino

Asignatura

Introducción al Análisis Multivariado

UNIVERSIDAD NACIONAL DE COLOMBIA

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: $H_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0$. 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:

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

```r
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.

```r
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.

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

```r
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.