

UNIVERSIDAD NACIONAL DE COLOMBIA
Sede-Medellín

Facultad de Ciencias
Introducción al análisis multivariado
Estadística

Trabajo #2

Integrantes:

Juan David García Zapata
C.C.1001.660.790
jgarciaza@unal.edu.co

Juan Diego Espinosa Hernández
C.C.1000.192.466
jespinosah@unal.edu.co

21/10/2023



La cedula a utilizar es la correspondiente al estudiante Juan Diego Espinosa Hernandez, la cual es:
1000192466

Estas son las variables que contiene la base de datos:

Table 1: Variables base de datos

N_variable	Descripcion
SEXO	Hom, Muj
P1	Masa, en kg
P7	Perímetro muslo mayor, en cm
P16	Perímetro abdominal cintura, en cm)
P22	Anchura caderas, en cm
P25	Distancia Nalga a fosa poplítea, en cm
P27	Longitud promedio de los pies, en cm
P29	Longitud promedio de las manos, en cm
P38	Altura, en cm
CAT_IMC	Delgado, Normal, Obeso

El taller decia que generaramos una muestra de **200** para trajar y que no modificaramos el codigo solo cambiariamos la semilla, y siguiendo estas pautas el codigo genero una muestra de **199**, asi que trabajaremos con eso ya que estamos siguiendo los lineamientos.

Parte A

En el caso de el numero de cedula elegido, los primeros 5 digitos no nulos son: **11924**, donde cada digito corresponde a un d_i .

Sea $X = (X_1, X_2, X_3, X_4, X_5)$ un vector aleatorio tal que $X \sim N_5(\mu, \Sigma)$, donde $\mu = (1, 1, 9, 3, 5)'$ es el vector de medias y Σ esta dado por:

$$\Sigma = \begin{pmatrix} \mathbf{1} & 4 & 6 & 1 & 6 \\ 4 & \mathbf{1} & 9 & 7 & 3 \\ 6 & 9 & \mathbf{9} & 10 & 5 \\ 1 & 7 & 10 & \mathbf{2} & 8 \\ 6 & 3 & 5 & 8 & \mathbf{4} \end{pmatrix}$$

Punto 1

Defina el vector $\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 - 2X_5 + 1 \\ X_2 - X_3 + 3X_4 + 2 \end{bmatrix}$. Halle la distribucion del vector Y

Tenemos:

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}; \vec{a} = (1, 1, 9, 3, 5); \Sigma = \begin{bmatrix} 1 & 4 & 6 & 1 & 6 \\ 4 & 1 & 9 & 7 & 3 \\ 6 & 9 & 9 & 10 & 5 \\ 1 & 7 & 10 & 2 & 8 \\ 6 & 3 & 5 & 8 & 4 \end{bmatrix}$$

Sea $\vec{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} x_1 - 2x_5 + 1 \\ x_2 - x_3 + 3x_4 + 2 \end{bmatrix}$

notemos que $\vec{Y} = A\vec{x} + b$ con

- $A = \begin{pmatrix} 1 & 0 & 0 & 0 & -2 \\ 0 & 1 & -1 & 3 & 0 \end{pmatrix}$
- $b = (1, 2)^T$
- de este manera $\vec{Y} \sim N_2(\vec{\mu}_Y, \Sigma_Y)$
- Por propiedades si: $\vec{Y} \sim N_2(\vec{\mu}, \Sigma)$, $\vec{Y} = A\vec{x} + b \sim N_2(A\vec{\mu} + b, A\Sigma A^T)$

$\Rightarrow \Sigma_Y = A\Sigma A^T =$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & -2 \\ 0 & 1 & -1 & 3 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 4 & 6 & 1 & 6 \\ 4 & 1 & 9 & 7 & 3 \\ 6 & 9 & 9 & 10 & 5 \\ 1 & 7 & 10 & 2 & 8 \\ 6 & 3 & 5 & 8 & 4 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 3 \\ -2 & 0 \\ 5 & 2 \end{pmatrix} = \begin{pmatrix} -7 & -43 \\ -43 & -8 \end{pmatrix}$$

Ayuda Con R

```
# Definir las matrices
matriz_A <- matrix(c(1, 0, 0, 0, -2, 0, 1, -1, 3, 0), nrow = 2, byrow = TRUE)
vector_B <- matrix(c(1, 1, 9, 3, 5), nrow = 5, byrow = TRUE)
vector_C <- matrix(c(1, 2), nrow = 2)

# Realizar la multiplicación de matrices y la suma
resultado <- matriz_A %*% vector_B + vector_C
```

```
# Definir las matrices
matriz_A <- matrix(c(1, 0, 0, 0, -2, 0, 1, -1, 3, 0),
nrow = 2, byrow = TRUE)
matriz_B <- matrix(c(1, 4, 6, 1, 6, 4, 1, 9, 7, 3, 6, 9,
9, 10, 5, 1, 7, 10, 2, 8, 6, 3, 5, 8, 4), nrow = 5, byrow =
TRUE)
matriz_C <- matrix(c(1, 0, 0, 1, 0, -1, 0, 3, -2),
nrow = 5, byrow = TRUE)

# Realizar la multiplicación de matrices
resultado_intermedio <- matriz_A %*% matriz_B
resultado <- resultado_intermedio %*% matriz_C
```

Como resultado tenemos:

$$\vec{Y} \sim N_2 \left(\begin{bmatrix} -8 \\ 3 \end{bmatrix}, \begin{bmatrix} -7 & -43 \\ -43 & -8 \end{bmatrix} \right)$$

Punto 2

Considere los sub-vectores: $\mathbf{X}^{(1)} = \begin{bmatrix} X_2 \\ X_4 \\ X_5 \end{bmatrix}; \mathbf{X}^{(2)} = \begin{bmatrix} X_1 \\ X_3 \end{bmatrix}$

a)

Halle la distribución de $\mathbf{X}^{(1)}$ y de $\mathbf{X}^{(2)}$

2) Sea $\vec{x} \xrightarrow{(1)} = \begin{pmatrix} x_2 \\ x_4 \\ x_5 \end{pmatrix}$; $\vec{x} \xrightarrow{(2)} = \begin{pmatrix} x_1 \\ x_3 \end{pmatrix}$

a) Notemos que $\vec{x} \xrightarrow{(1)} = A \vec{x} + \vec{b} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} x_2 \\ x_4 \\ x_5 \end{pmatrix}$

$$\vec{x} \xrightarrow{(1)} \sim N_3(\vec{\mu}, \Sigma)$$

Con $\vec{\mu} \xrightarrow{(1)} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 9 \\ 3 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix}$

Con $\Sigma \xrightarrow{(1)} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{matrix} \begin{bmatrix} 1 & 4 & 6 & 1 & 6 \\ 4 & 1 & 9 & 7 & 3 \\ 6 & 9 & 9 & 10 & 5 \\ 2 & 7 & 10 & 2 & 8 \\ 6 & 3 & 5 & 8 & 4 \end{bmatrix} \\ 3 \times 5 \end{matrix} \times \begin{matrix} \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ 5 \times 3 \end{matrix} = \begin{bmatrix} 1 & 7 & 3 \\ 7 & 2 & 8 \\ 3 & 8 & 4 \end{bmatrix}$

Entonces:

$$\vec{x} \xrightarrow{(1)} \sim N_3\left(\begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}, \begin{bmatrix} 1 & 7 & 3 \\ 7 & 2 & 8 \\ 3 & 8 & 4 \end{bmatrix}\right)$$

Recordemos

$$\mu = \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix}$$

Definir las matrices

```
matriz_A <- matrix(c(0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1), nrow = 3, byrow = TRUE)
matriz_B <- matrix(c(1, 4, 6, 1, 6, 4, 1, 9, 7, 3, 6, 9, 9, 10, 5, 1, 7, 10, 2, 8, 6, 3, 5, 8, 4), nrow = 5, byrow = TRUE)
matriz_C <- matrix(c(0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1), nrow = 5, byrow = TRUE)
```

Realizar la multiplicación de matrices

```
resultado_intermedio <- matriz_A %*% matriz_B
resultado <- resultado_intermedio %*% matriz_C
```

Para $\vec{x} \xrightarrow{(2)} = \begin{pmatrix} x_1 \\ x_3 \end{pmatrix}$

Notemos que $\vec{x} \xrightarrow{(2)} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} : A \vec{x} + \vec{b}$

$$\Rightarrow \vec{x} \xrightarrow{(2)} \sim N_2(\vec{\mu}, \Sigma)$$

con

$$\vec{\mu} \xrightarrow{(2)} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} * \begin{pmatrix} 1 \\ 1 \\ 9 \\ 3 \\ 3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 9 \end{pmatrix}$$

$$\Sigma \xrightarrow{(2)} = A * \Sigma * A'$$

$$\Sigma \xrightarrow{(2)} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} * \begin{bmatrix} 1 & 4 & 6 & 1 & 6 \\ 4 & 1 & 9 & 7 & 3 \\ 6 & 9 & 9 & 10 & 5 \\ 2 & 7 & 10 & 2 & 8 \\ 6 & 3 & 5 & 8 & 4 \end{bmatrix} * \begin{bmatrix} 1 & 6 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 6 \\ 6 & 9 \end{bmatrix}$$

$$\vec{x} \xrightarrow{(2)} \sim N_2\left(\begin{bmatrix} 1 \\ 9 \end{bmatrix}, \begin{bmatrix} 1 & 6 \\ 6 & 9 \end{bmatrix}\right)$$

Recordemos

$$\mu = \begin{pmatrix} 1 \\ 9 \\ 3 \\ 5 \end{pmatrix}$$

Definir las matrices

```
matriz_A <- matrix(c(1, 0, 0, 0, 0, 0, 1, 0, 0, 0), nrow = 2, byrow = TRUE)
matriz_B <- matrix(c(1, 4, 6, 1, 6, 4, 1, 9, 7, 3, 6, 9, 9, 10, 5, 1, 7, 10, 2, 8, 6, 3, 5, 8, 4), nrow = 5, byrow = TRUE)
matriz_C <- matrix(c(1, 0, 0, 0, 1, 0, 0, 0), nrow = 5, byrow = TRUE)
```

Realizar la multiplicación de matrices

```
resultado_intermedio <- matriz_A %*% matriz_B
resultado <- resultado_intermedio %*% matriz_C
```

b)

Halle la distribución condicional de $\mathbf{X}^{(1)}$ dado $\mathbf{X}^{(2)} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$

b) Sea la distribución condicional de $\vec{x}^{(1)} \text{ dado } \vec{x}^{(2)} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$

$T \sim N_3(\vec{\mu}_T, \Sigma_T)$ ← utilizando las propiedades de la normal multivariada

Para hallar $\vec{\mu}_T$ es necesario descomponer Σ_T . Para llevar a cabo este proceso, se introducirán las siguientes variables

$$H = \begin{bmatrix} x_2 \\ x_4 \\ x_5 \\ x_1 \\ x_3 \end{bmatrix} = \begin{bmatrix} \vec{x}^{(1)} \\ \vec{x}^{(2)} \end{bmatrix} = \begin{bmatrix} x_2 \\ x_4 \\ x_5 \\ x_1 \\ x_3 \end{bmatrix} \quad \text{como } H = A \vec{x} + \vec{b}$$

$$= \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}; \quad H \sim N_5(\vec{\mu}_H, \Sigma_H)$$

$$\text{con } \vec{\mu}_H = \begin{pmatrix} \vec{\mu}^{(1)} \\ \vec{\mu}^{(2)} \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \\ 9 \end{pmatrix} \quad \text{y con } \Sigma_H = A_H \cdot \Sigma \cdot A_H^T$$

$$\begin{array}{|c|c|c|c|} \hline 6_{22} & 6_{24} & 6_{25} & 6_{21} & 6_{23} \\ \hline 6_{42} & 6_{44} & 6_{45} & 6_{41} & 6_{43} \\ \hline 6_{52} & 6_{54} & 6_{55} & 6_{51} & 6_{53} \\ \hline 6_{12} & 6_{14} & 6_{15} & 6_{11} & 6_{13} \\ \hline 6_{32} & 6_{34} & 6_{35} & 6_{31} & 6_{33} \\ \hline \end{array}$$

$$\begin{array}{|c|c|c|c|} \hline 1 & 7 & 3 & 4 & 9 \\ \hline 7 & 2 & 8 & 1 & 10 \\ \hline 3 & 8 & 4 & 6 & 5 \\ \hline 4 & 1 & 6 & 1 & 6 \\ \hline 9 & 10 & 5 & 6 & 9 \\ \hline \end{array}$$

$$\text{Luego } \bar{\Sigma}_H = \begin{bmatrix} \bar{\Sigma}_{11} & \bar{\Sigma}_{12} \\ \bar{\Sigma}_{21} & \bar{\Sigma}_{22} \end{bmatrix}$$

con:

$$\begin{array}{|c|c|c|c|} \hline \bar{\Sigma}_{11} & \bar{\Sigma}_{12} & \bar{\Sigma}_{21} & \bar{\Sigma}_{22} \\ \hline \begin{bmatrix} 1 & 7 & 3 \\ 7 & 2 & 8 \\ 3 & 8 & 4 \end{bmatrix} & \begin{bmatrix} 4 & 9 \\ 1 & 10 \\ 6 & 5 \end{bmatrix} & \begin{bmatrix} 4 & 1 & 6 \\ 9 & 10 & 5 \end{bmatrix} & \begin{bmatrix} 1 & 6 \\ 6 & 9 \end{bmatrix} \\ \hline \end{array}$$

Tenemos la fórmula

$$\vec{\mu}_{\vec{x}^{(1)} | \vec{x}^{(2)}} = \vec{\mu}^{(1)} + \bar{\Sigma}_{12} \bar{\Sigma}_{22}^{-1} (\vec{x}^{(2)} - \vec{\mu}^{(2)})$$

$$\vec{\mu}_{\vec{x}^{(1)} | \vec{x}^{(2)}} = \begin{pmatrix} 3 \\ 5 \end{pmatrix} + \begin{pmatrix} 4 & 9 \\ 1 & 10 \\ 6 & 5 \end{pmatrix}_{3x2} \begin{pmatrix} 1 & 6 \\ 6 & 9 \end{pmatrix}_{2x2}^{-1} \left(\begin{bmatrix} 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 1 \\ 9 \end{bmatrix} \right) = \begin{bmatrix} -26/9 \\ 109/27 \\ -82/27 \end{bmatrix}$$

En R: μ

```
install.packages("Matrix")
# Cargar el paquete Matrix
library(Matrix)

# Definir las matrices y vectores
v <- c(1, 3, 5)
A <- matrix(c(4, 9, 1, 10, 6, 5), nrow = 3,
byrow = TRUE)
B_inv <- solve(matrix(c(1, 6, 6, 9), nrow = 2,
byrow = TRUE))
u <- c(0, -7)

# Realizar la operación
result <- v + A %*% B_inv %*% u
```

```
A <- matrix(c(1, 7, 3, 7, 2, 8, 3, 8, 4), nrow = 3, byrow = TRUE)
B <- matrix(c(4, 9, 1, 10, 6, 5), nrow = 3, byrow = TRUE)
C <- matrix(c(1, 6, 6, 9), nrow = 2, byrow = TRUE)
D <- matrix(c(4, 1, 6, 9, 10, 5), nrow = 2, byrow = TRUE)

# Calcular la inversa de la matriz C
C_inv <- solve(C)

# Realizar la operación matricial
result <- A - B %*% C_inv %*% D
```

ahora

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

$$\begin{bmatrix} 1 & 7 & 3 \\ 7 & 2 & 8 \\ 3 & 8 & 4 \end{bmatrix}_{3 \times 3} - \begin{bmatrix} 4 & 9 \\ 1 & 10 \\ 6 & 5 \end{bmatrix}_{3 \times 2} \begin{bmatrix} 1 & 6 \\ 6 & 9 \end{bmatrix}_{2 \times 2}^{-1} \begin{bmatrix} 4 & 1 & 6 \\ 9 & 10 & 5 \end{bmatrix}_{2 \times 3}$$

$$= \begin{bmatrix} -20/3 & 7/9 & -34/9 \\ 7/9 & 43/27 & -70/27 \\ -34/9 & -70/27 & 97/27 \end{bmatrix}_{3 \times 3}$$

Notemos que se está trabajando

con un supuesto que no se cumple para $\text{up}(\vec{\mu}, \Sigma)$

Porque notemos que $\Sigma_{Y^{(1)} X^{(2)}}$ no es definida positiva ni semipositiva

Parte B

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

Hacemos prueba de normalidad multivariada con Royston's test

Table 2: Normalidad

Test	H	p value	MVN
Royston	29.1891	5.46e-05	NO

Punto 1

Primero debemos tener en cuenta que nuestros datos no provienen de una distribución normal y para hacer la prueba de hipótesis debemos de tener en cuenta que desconocemos tanto el vector de medias μ como la matriz de varianzas y covarianzas, por lo tanto debemos estimar ambos a partir de los datos, ya que tenemos conocimiento acerca de esto procederemos a definir los siguientes elementos para hacer la prueba de hipótesis $\mu = \mu_0$, la cual se refiere al siguiente juego de hipótesis:

$$H_0 : \mu = \mu_0 \quad vs \quad H_1 : \mu \neq \mu_0$$

Por la **ley debil de los grandes numeros** se garantiza que $\overline{X} \xrightarrow{p} \mu$ y $S \xrightarrow{p} \Sigma$, entonces:

$$\overline{X} = \begin{bmatrix} 64.87186 \\ 55.55025 \\ 79.94422 \\ 36.02864 \\ 46.38543 \\ 24.20503 \\ 17.6392 \\ 162.6829 \end{bmatrix} \text{ y } \mu_0 = \begin{bmatrix} 66.1 \\ 58 \\ 81.6 \\ 37 \\ 47 \\ 25 \\ 19.2 \\ 167 \end{bmatrix}$$

$$S = \begin{bmatrix} 122.890416 & 40.2635425 & 90.537362 & 16.8143960 & 13.401962 & 9.5883239 & 6.6404015 & 54.5494158 \\ 40.263542 & 24.6902903 & 25.364837 & 12.2737049 & 5.639524 & 1.1121704 & 0.6505457 & 6.2108629 \\ 90.537362 & 25.3648373 & 88.527833 & 9.3252421 & 6.606405 & 4.9877565 & 3.7227529 & 27.2633351 \\ 16.814396 & 12.2737049 & 9.325242 & 9.7819532 & 2.742894 & -0.0940841 & -0.3135526 & -0.5799627 \\ 13.401962 & 5.6395239 & 6.606405 & 2.7428943 & 6.402766 & 1.8217403 & 1.2889579 & 11.2003053 \\ 9.588324 & 1.1121704 & 4.987756 & -0.0940841 & 1.821740 & 2.4818938 & 1.6014182 & 11.8454903 \\ 6.640402 & 0.6505457 & 3.722753 & -0.3135526 & 1.288958 & 1.6014182 & 1.2667387 & 8.2208751 \\ 54.549416 & 6.2108629 & 27.263335 & -0.5799627 & 11.200305 & 11.8454903 & 8.2208751 & 71.1616258 \end{bmatrix}$$

Los calculos en R para calcular el estadistico T_0^2 son:

$$T_0^2 = n(\overline{X} - \mu_0)'S^{-1}(\overline{X} - \mu_0) \sim \frac{(n-1)p}{n-p} f(p, n-p)$$

Table 3: T_0^2

$$\begin{array}{c} \hline \text{Estadistico} \\ \hline 1810.498 \\ \hline \end{array}$$

$$T_0^2 = 1810.498; f(0.05, 8, 199 - 8) = 1.987138$$

Ahora:

$$\frac{199 - 8}{(199 - 1)8} 1810.498 > 1.987138$$

$$218.3113 > 1.987138$$

Por lo tanto se rechaza la hipotesis nula, entonces nuestro vector de medias estimado es diferente a el vector de medias planteado para la hipotesis nula, o en otras palabras: $\overline{X} \neq \mu_0$

Punto 2

Para realizar este punto filtramos la base de datos por variable **SEXO** la cual tiene las categorias: Hom(para hombres) y Muj(para mujeres), por lo tanto generamos 2 sub bases de datos filtradas por cada genero a partir de la muestra.

Hombres:

Primero aplicaremos prueba de normalidad multivariada con Royston's test:

Table 4: Normalidad hombres

Test	H	p value	MVN
Royston	26.66311	0.0001657	NO

Con un $\alpha = 0.05$ rechazamos la hipótesis nula y por lo tanto nuestros datos no se distribuyen normales multivariados.

Tambien debemos tener en cuenta que nuestros datos no provienen de una distribucion normal y para hacer la prueba de hipotesis debemos de tener en cuenta que desconocemos tanto el vector de medias μ_H como la matriz de varianzas y covarianzas, por lo tanto debemos estimar ambos a partir de los datos, ya que tenemos conocimiento acerca de esto procederemos a definir los siguientes elementos para hacer la prueba de hipotesis $\mu_H = \mu_0$, la cual se refiere al siguiente juego de hipotesis:

$$H_0 : \mu_H = \mu_0 \quad vs \quad H_1 : \mu_H \neq \mu_0$$

Por la **ley debil de los grandes numeros** se garantiza que $\bar{X}_H \xrightarrow{p} \mu_H$ y $S_H \xrightarrow{p} \Sigma_H$, entonces:

$$\bar{X}_H = \begin{bmatrix} 68.92130 \\ 54.86759 \\ 83.27222 \\ 34.85370 \\ 46.52870 \\ 25.18056 \\ 18.35093 \\ 168.06852 \end{bmatrix} \quad y \quad \mu_0 = \begin{bmatrix} 66.1 \\ 58 \\ 81.6 \\ 37 \\ 47 \\ 25 \\ 19.2 \\ 167 \end{bmatrix}$$

$$S_H = \begin{bmatrix} 121.134963 & 49.528454 & 83.6152700 & 24.8939858 & 11.934990 & 5.4282684 & 3.6393726 & 33.972546 \\ 49.528454 & 26.010996 & 31.5256334 & 11.1421305 & 4.823182 & 1.7561864 & 1.3634415 & 12.046634 \\ 83.615270 & 31.525633 & 79.1674455 & 16.9058048 & 4.777814 & 1.2761838 & 0.9776895 & 7.911547 \\ 24.893986 & 11.142130 & 16.9058048 & 7.3731637 & 2.339285 & 1.1490914 & 0.7065853 & 7.200959 \\ 11.934990 & 4.823182 & 4.7778141 & 2.3392852 & 5.901505 & 1.7554232 & 1.3002068 & 11.186426 \\ 5.428268 & 1.756186 & 1.2761838 & 1.1490914 & 1.755423 & 1.4699922 & 0.8590369 & 6.525457 \\ 3.639373 & 1.363442 & 0.9776895 & 0.7065853 & 1.300207 & 0.8590369 & 0.7305599 & 4.247132 \\ 33.972546 & 12.046634 & 7.9115472 & 7.2009588 & 11.186426 & 6.5254569 & 4.2471322 & 41.026289 \end{bmatrix}$$

Los calculos en R para calcular el estadistico T_0^2 son:

$$T_0^2 = n(\bar{X}_H - \mu_0)' S_H^{-1} (\bar{X}_H - \mu_0) \sim \frac{(n-1)p}{n-p} f(p, n-p)$$

Table 5: T_0^2

Estadistico
1058.521

$$T_0^2 = 1058.521; f(0.05, 8, 108 - 8) = 2.032328$$

Ahora:

$$\frac{108 - 8}{(108 - 1)8} 1058.521 > 2.032328$$

$$132.3151 > 2.032328$$

Por lo tanto se rechaza la hipotesis nula, entonces nuestro vector de medias estimado es diferente a el vector de medias planteado para la hipotesis nula, o en otras palabras: $\bar{X}_H \neq \mu_0$

Mujeres:

Primero aplicaremos prueba de normalidad multivariada con Royston's test

Table 6: Normalidad mujeres

Test	H	p value	MVN
Royston	16.05122	0.0217645	NO

Tambien debemos tener en cuenta que nuestros datos no provienen de una distribucion normal y para hacer la prueba de hipotesis debemos de tener en cuenta que desconocemos tanto el vector de medias μ_M como la matriz de varianzas y covarianzas, por lo tanto debemos estimar ambos a partir de los datos, ya que tenemos conocimiento acerca de esto procederemos a definir los siguientes elementos para hacer la prueba de hipotesis $\mu_M = \mu_0$, la cual se refiere al siguiente juego de hipotesis:

$$H_0 : \mu_M = \mu_0 \quad vs \quad H_1 : \mu_M \neq \mu_0$$

Por la **ley debil de los grandes numeros** se garantiza que $\bar{X}_M \xrightarrow{p} \mu_M$ y $S_M \xrightarrow{p} \Sigma_M$, entonces:

$$\bar{X}_M = \begin{bmatrix} 60.06593 \\ 56.36044 \\ 75.99451 \\ 37.42308 \\ 46.21538 \\ 23.04725 \\ 16.79451 \\ 156.29121 \end{bmatrix} \quad y \quad \mu_0 = \begin{bmatrix} 66.1 \\ 58 \\ 81.6 \\ 37 \\ 47 \\ 25 \\ 19.2 \\ 167 \end{bmatrix}$$

$$S_M = \begin{bmatrix} 83.311827 & 36.950193 & 64.4081441 & 19.8809060 & 13.772419 & 4.2742943 & 2.7189219 & 22.389364 \\ 36.950193 & 22.171529 & 24.2840024 & 11.6505897 & 6.929393 & 2.1064457 & 1.0852247 & 8.989648 \\ 64.408144 & 24.284002 & 71.5756361 & 10.6774615 & 7.602530 & 0.9362625 & 0.8119695 & 3.539507 \\ 19.880906 & 11.650590 & 10.6774615 & 9.1317949 & 3.694974 & 1.4346752 & 0.6645726 & 6.768094 \\ 13.772419 & 6.929393 & 7.6025299 & 3.6949744 & 7.015983 & 1.5540427 & 1.0223077 & 9.316359 \\ 4.274294 & 2.106446 & 0.9362625 & 1.4346752 & 1.554043 & 1.2151868 & 0.6798181 & 4.515087 \\ 2.718922 & 1.085225 & 0.8119695 & 0.6645726 & 1.022308 & 0.6798181 & 0.5889695 & 2.977840 \\ 22.389364 & 8.989648 & 3.5395067 & 6.7680940 & 9.316359 & 4.5150867 & 2.9778400 & 31.666366 \end{bmatrix}$$

Los calculos en R para calcular el estadistico T_0^2 son:

$$T_0^2 = n(\bar{X}_M - \mu_0)' S_M^{-1} (\bar{X}_H - \mu_0) \sim \frac{(n-1)p}{n-p} f(p, n-p)$$

Table 7: T_0^2

Estadistico
1583.959

$$T_0^2 = 1583.959; f(0.05, 8, 91-8) = 2.05201$$

Ahora:

$$\frac{91-8}{(91-1)8} 1583.959 > 2.05201$$

$$197.9949 > 2.05201$$

Por lo tanto se rechaza la hipotesis nula, entonces nuestro vector de medias estimado es diferente a el vector de medias planteado para la hipotesis nula, o en otras palabras: $\bar{X}_M \neq \mu_0$

Punto 3

Se sabe que $\mathbf{X} \sim N_8(\mu, \Sigma)$. Se quiere probar la hipotesis:

$$H_0 : \begin{cases} 2\mu_1 - \mu_2 + 3\mu_4 - \mu_6 + \mu_7 - \mu_8 = 0 \\ 3\mu_2 - 4\mu_3 + 2\mu_5 + 2\mu_6 = 0 \end{cases}$$

La cual se puede escribir como:

$$\begin{bmatrix} 2 & -1 & 0 & 3 & 0 & -1 & 1 & -1 \\ 0 & 3 & -4 & 0 & 2 & 2 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Como \bar{X} es un estimador insesgado para μ , por esta propiedad: $E[C\bar{X}] = CE[\bar{X}] = C\mu$ se tiene que $C\bar{X}$ es un estimador insesgado para $C\mu$. Tambien se tiene que:

$$\bar{X} \sim N_8(\mu, \frac{1}{199}\Sigma), \text{ lo que implica que } C\bar{X} \sim N_2(C\mu, \frac{1}{199}C\Sigma C')$$

Donde C es una matriz de dimension $k \times p$, donde $k=2$ y $p=8$. Como Σ es desconocida, se estima usando S

El estadistico bajo, H_0 cierta, es:

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

Donde se rechaza H_0 , con un α fijado, si $\frac{(n-k)}{(n-1)k}T_0^2 > f_{\alpha}(k, n-k)$

Haciendo el computo en R, se tiene que $T_0^2 = 79.14568$, $\frac{(n-k)}{(n-1)k}T_0^2 = 39.37298$. Asi, fijando un $\alpha = 0.05$, $f_{0.05}(2, 199 - 2) = 4.39842$

Para concluir, podemos ver que $39.37298 > 4.39842$ por lo tanto se rechaza la hipotesis nula (la planteada) y se concluye que la evidencia muestral no parece ser coherente con la hipótesis planteada

Punto 4

Para realizar este punto filtramos la base de datos por variable **CAT_IMC** la cual tiene las categorias: Delgado, Normal y Obeso, por lo tanto generamos 3 sub bases de datos filtradas por cada categoria de IMC a partir de la muestra.

Tenemos un vector **X** el cual tiene distribucion normal multivariada.

Para tres grupos, en nuestro caso delgado, normal y obeso, asi que lo que se plantea es: $\Sigma_D = \Sigma_N = \Sigma_O = \Sigma$, entonces definimos $U_1 \sim N_8(\mu_1, \Sigma_1)$ como la distribucion para las poblacion de personas delgadas, $U_2 \sim N_8(\mu_2, \Sigma_2)$ para personas IMC normal y $U_3 \sim N_8(\mu_3, \Sigma_3)$, defnimos la hipotesis nula:

$$H_0 : \Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma$$

Para probar esta hipotesis necesitamos el siguiente estadistico de razon de verosimilitud:

$$\lambda = \prod_{i=1}^3 \left(\frac{|S_i|}{|S_p|} \right)^{\frac{(n_i-1)}{2}}$$

Donde la expresion $S_i; i = 1, 2, 3$ representa la matriz de varianzas y covarianzas muestrals para las 3 categorias de la variables **CAT_IMC**, y S_p la matriz ponderada de estas poblaciones dada por:

$$S_p = \frac{1}{\sum_{i=1}^3 (n_i - 1)} [(n_1 - 1)S_1 + (n_2 - 1)S_2 + (n_3 - 1)S_3]$$

$$S_1 = \text{Delgado}, S_2 = \text{Normal}, S_3 = \text{Obeso}$$

Ahora necesitamos un estadistico para probar H_0 , asi que tenemos, teniendo ya un criterio para evaluar la hipotesis, fijando un nivel de significancia $\alpha = 0.05$, se rechaza H_0 si:

$$C = (1 - u)M > \chi_{0.05}^2 \left(\frac{1}{2} p(p+1)(g-1) \right)$$

En nuestro caso **g = 3** y **p = 8**, teniendo en cuenta esto.

Ahora iniciemos definiendo y calculando **M**

$$M = -2\ln(\lambda)$$

entonces:

$$M = [\sum_{i=1}^3 (n_i - 1)] * \ln|S_p| - \sum_{i=1}^3 (n_i - 1) * \ln|S_i|$$

Ahora **u**:

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

Teniendo ya un criterio para evaluar la hipótesis, fijando un nivel de significancia $\alpha = 0.05$, se rechaza H_0 si:

$$C = 122.6367 \text{ y } \chi^2_{0.05}(\frac{1}{2}8(8+1)(3-1)) = 53.46233$$

$$122.6367 > 53.46233$$

Como se **C** es mayor a el cuantil de la chi cuadrado, se rechaza H_0 y por lo tanto se tiene suficiente evidencia estadística para afirmar que la estructura de covarianzas para personas con la categoría IMC delgado, normal y obeso son diferentes.

Punto 5

Para realizar este punto filtramos la base de datos por variable **SEXO** la cual tiene las categorías: Hom(para hombres) y Muj(para mujeres), por lo tanto generamos 2 sub bases de datos filtradas por cada género a partir de la muestra.

Usaremos los mismos vectores de medias y la matrices de varianzas y covarianzas que estimamos en el punto 2 para ambos géneros.

La prueba a realizar es:

$$H_0 : \mu_H - \mu_M = \delta \text{ vs } H_1 : \mu_H - \mu_M \neq \delta$$

donde $\delta = (0, 0, 0, 0, 0, 0, 0, 0)'$

De los resultados muestrales obtenidos se tiene que:

$$\bar{X}_H - \bar{X}_M = \begin{bmatrix} 8.855 \\ -1.493 \\ 7.278 \\ -2.569 \\ 0.313 \\ 2.133 \\ 1.556 \\ 11.777 \end{bmatrix}$$

Tambien tenemos que:

$$S = \frac{1}{n_H} S_H + \frac{1}{n_M} S_M = \frac{1}{108} S_H + \frac{1}{91} S_M$$

$$S = \begin{bmatrix} 2.037 & 0.865 & 1.482 & 0.449 & 0.262 & 0.097 & 0.064 & 0.561 \\ 0.865 & 0.484 & 0.559 & 0.231 & 0.121 & 0.039 & 0.025 & 0.210 \\ 1.482 & 0.559 & 1.520 & 0.274 & 0.128 & 0.022 & 0.018 & 0.112 \\ 0.449 & 0.231 & 0.274 & 0.169 & 0.062 & 0.026 & 0.014 & 0.141 \\ 0.262 & 0.121 & 0.128 & 0.062 & 0.132 & 0.033 & 0.023 & 0.206 \\ 0.097 & 0.039 & 0.022 & 0.026 & 0.033 & 0.027 & 0.015 & 0.110 \\ 0.064 & 0.025 & 0.018 & 0.014 & 0.023 & 0.015 & 0.013 & 0.072 \\ 0.561 & 0.210 & 0.112 & 0.141 & 0.206 & 0.110 & 0.072 & 0.728 \end{bmatrix}$$

Bajo H_0 , el estadistico de prueba se calcula como:

$$X_c = [\bar{X}_H - \bar{X}_M - \delta_0]' \left[\frac{1}{n_H} S_H + \frac{1}{n_M} S_M \right]^{-1} [\bar{X}_H - \bar{X}_M - \delta_0] \xrightarrow{d} \chi^2(p)$$

$$X_c = 735.7922$$

Fijando un $\alpha = 0.05$, se tiene que nuestro valor critico es:

$$\chi^2_{0.05}(8) = 15.50731$$

$$735.7922 > 15.50731$$

Como X_c es mayor que el cuantil de la chi-cuadrado, se rechaza la hipotesis nula y por lo tanto se concluye que el vector \mathbf{X} es suficiente para discriminar entre hombres y mujeres.

Punto 6

Al calcular los valores propios obtenemos los siguientes resultados:

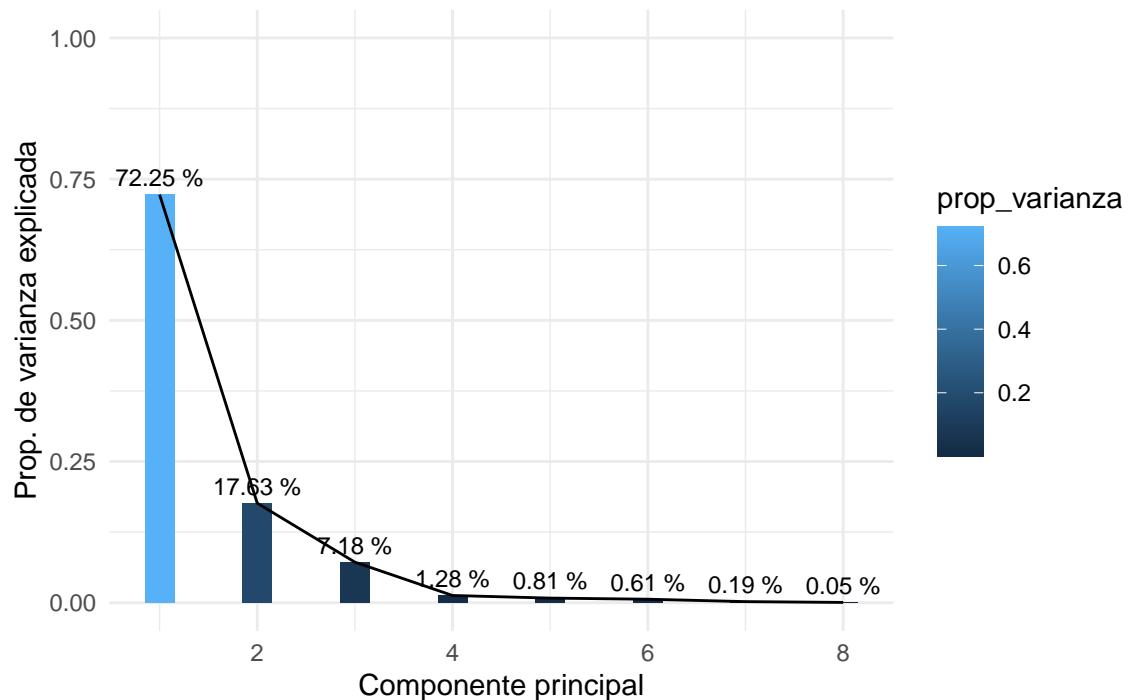
$$(236.409, 57.674, 23.497, 4.176, 2.666, 2.006, 0.626, 0.151)$$

Table 8: Y vectores propios

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
P1	-0.713	0.068	-0.202	-0.424	-0.217	0.467	0.022	-0.004
P7	-0.219	0.241	-0.614	0.063	-0.219	-0.681	-0.004	0.015
P16	-0.551	0.367	0.599	0.327	0.172	-0.258	-0.005	0.007
P22	-0.087	0.163	-0.430	0.200	0.829	0.225	-0.044	-0.046
P25	-0.081	-0.095	-0.194	0.815	-0.385	0.367	-0.002	0.018
P27	-0.060	-0.141	0.006	-0.012	0.021	-0.027	-0.826	0.541
P29	-0.042	-0.097	0.019	0.005	-0.042	-0.039	-0.529	-0.840
P38	-0.346	-0.859	0.004	0.073	0.196	-0.254	0.186	0.003

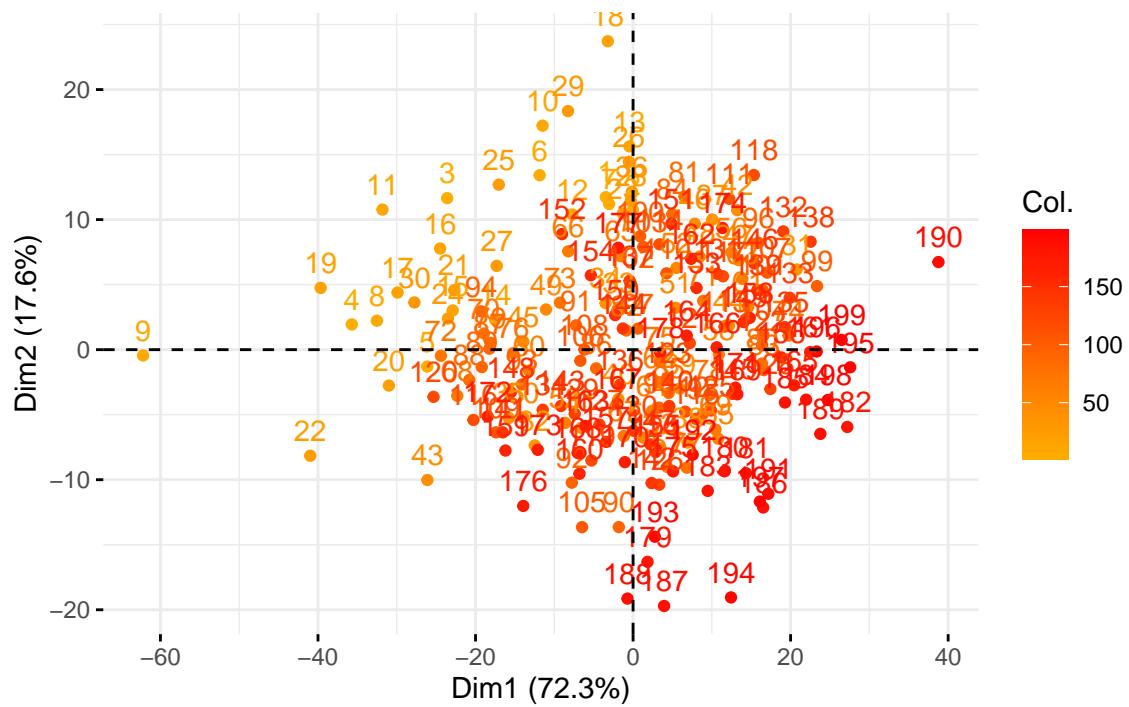
Cada columna es un vector propio el cual esta relacionado a su respectivo valor propio el cual mostramos al principio de este punto.

Scree-plot



En este grafico podemos ver el porcentaje de variabilidad total que explica cada una de las componentes principales, en este caso la primera explica un **72.25%** y las primeras dos componentes explican un total de **89.88%** de la variabilidad total y ademas se observa que las observaciones tienen un peso mayor en las componentes principales antes vistas.

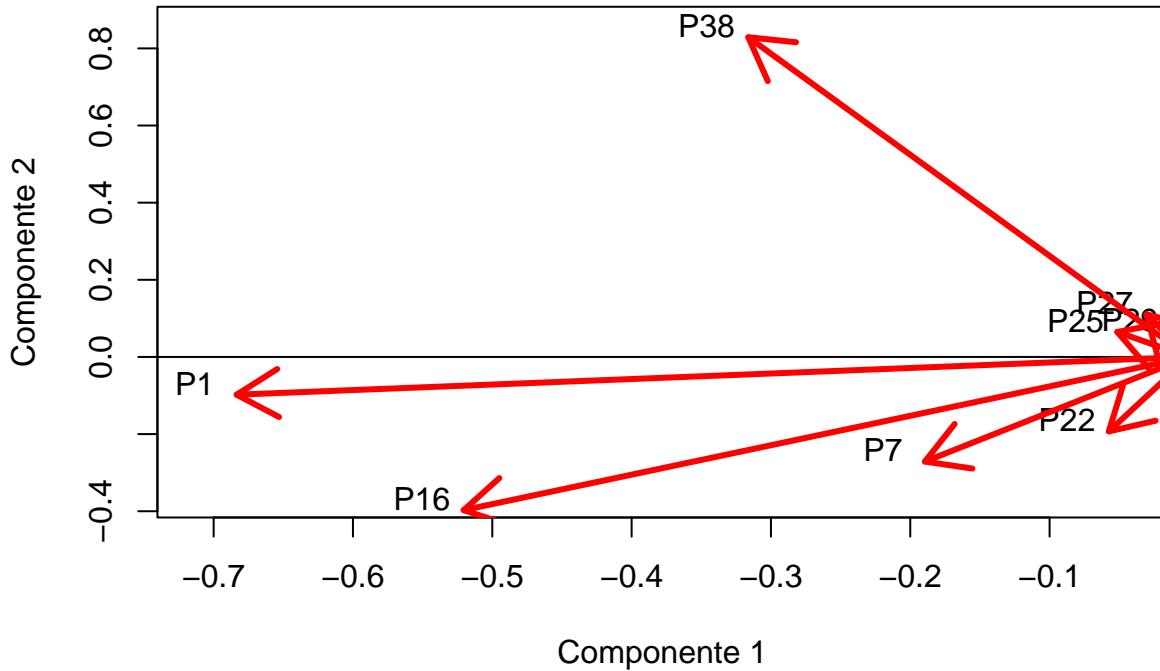
Individuals – PCA



En el gráfico, cada punto representa un individuo, y el color del punto representa la proporción de varianza explicada por la primera componente principal. Los individuos de color rojo representan a los individuos que tienen una mayor proporción de varianza explicada por la primera componente principal, mientras que los individuos de color amarillo representan a los individuos que tienen una menor proporción de varianza explicada por la primera componente principal.

Table 9: Coeficientes primera componente

P1	P7	P16	P22	P25	P27	P29	P38
-0.713	-0.219	-0.551	-0.087	-0.081	-0.06	-0.042	-0.346



Con estos valores podemos ver que las variables que tienen un mayor aporte a la primera componente son: P1(Masa), P16(Perímetro abdominal cintura) y P38 (altura) tambien tiene un peso importante.

Se observa que todos los coeficientes en esta combinación son negativos, lo que sugiere que esto puede servir como un indicador de individuos con mediciones antropométricas muy pequeñas o muy grandes. Para clasificar a las personas, simplemente se realiza el producto escalar entre un conjunto de datos y el vector de coeficientes de la primera componente principal. A partir de este resultado, se genera un score, cuanto menor sea este valor, se infiere que la persona posee características antropométricas más altas, mientras que un valor mayor indica cualidades antropométricas más bajas.

Debemos probar que nuestro valor mas pequeño el cual es $\lambda_8 = 0.151$ es significativamente mayor que 1, y esto lo haremos usando el siguiente juego de hipótesis:

$$H_0 : \lambda_8 = 0 \text{ vs } H_1 : \lambda_8 \neq 0$$

Así con un $\alpha = 0.05$, un intervalo de confianza aproximado al $100(1 - \alpha)$ para λ_8 está dado por:

$$I = \left(\frac{\hat{\lambda}_8}{1+Z_{0.025}\sqrt{\frac{2}{199}}}, \frac{\hat{\lambda}_8}{1-Z_{0.025}\sqrt{\frac{2}{199}}} \right) = \left(\frac{0.151}{1+1.96\sqrt{\frac{2}{199}}}, \frac{0.151}{1-1.96\sqrt{\frac{2}{199}}} \right)$$

$$I = (0.126, 0.188)$$

Como el intervalo de confianza no contiene el 0, esto quiere decir que se rechaza la hipótesis nula H_0 y esto quiere decir que los valores propios son significativamente diferentes de 0.

Anexos

A través del link y/o el código QR podrá acceder al código con el cual se hizo todo este trabajo



LINK: https://drive.google.com/file/d/1rjXSY1OMPcF_GfvLPDkYDHciyeulwueR/view?usp=sharing