SEGUNDO TRABAJO INTRODUCCIÓN AL ANÁLISIS MULTIVARIADO
SEM 02 – 2023 Octubre 16 de 2023

PARTE A. (30%)

Sea $\boldsymbol{X} = (X_1,\ X_2,\ X_3,\ X_4,\ X_5)$, un vector aleatorio tal que $\boldsymbol{X} \sim N_5(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, donde $\boldsymbol{\mu} = (d_1,\ d_2,\ d_3,\ 3,\ 5)'$ es el vector de medias y $\boldsymbol{\Sigma}$ está dada por:

$$\Sigma = \begin{pmatrix} d_1 & 4 & 6 & 1 & 6 \\ 4 & d_2 & 9 & 7 & 3 \\ 6 & 9 & d_3 & 10 & 5 \\ 1 & 7 & 10 & d_4 & 8 \\ 6 & 3 & 5 & 8 & d_5 \end{pmatrix}$$

En este caso $(d_1,\ d_2,\ d_3,\ d_4,\ d_5,)$ son los primeros 5 dígitos NO Nulos de su documento de identidad.

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

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

a) Halle la distribución de $\boldsymbol{X^{(1)}}$ y de $\boldsymbol{X^{(2)}}$.

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

Esta primera parte debe hacerse manualmente. Para los cálculos matriciales puede ayudarse de R, pero en todo caso debe mostrar el proceso PASO a PASO. Esto quiere decir que el desarrollo se debe escanear y anexar en formato pdf.

PARTE B. (70%)

Considere la base de datos que corresponde a la información sobre parámetros antropométricos de la población laboral Colombiana. De esta base estamos interesados solo en las variables: **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) Y **CAT_IMC** (Delgado, Normal, Obeso). La base de datos se encuentra en el curso de Moodle, pestaña "General", con el nombre "Acopla".

Cada grupo debe generar una muestra aleatoria de 200 registros usando los siguientes comandos:

```
uno <- read.table(file.choose(), header=T)
```

Copiar el siguiente código en R sin modificar nada

```
library(splitstackshape)

genera <- function(cedula){
set.seed(cedula)
aux <- stratified(uno, "CAT_IMC", 200/2100)
aux
}
```

Para crear la base de datos con la cual trabajara, debe ejecutar la siguiente línea:

datos <- genera(cedula)

El campo cédula corresponde a los últimos 5 dígitos del número de documento de identidad de alguno de los integrantes del grupo. Debe especificarse cual usaron.

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

Con esta muestra generada responda a las siguientes preguntas:

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.

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.

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.

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.

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

El trabajo debe ser cargado en el curso en Moodle antes de las 6:00 pm del 21 de octubre de 2023, en la carpeta Soporte Segundo Trabajo, dentro de la pestaña Inferencia Multivariada, **SOLO** en formato pdf. El número máximo de páginas, incluyendo los códigos en R, es de 15.