**PRIMER TRABAJO INTRODUCCION AL ANALISIS MULTIVARIADO**

**SEM 02 – 2023 Septiembre 1 de 2023**

En este trabajo se pretende realizar un análisis descriptivo a nivel univariado y multivariado de las variables contenidas en su base de datos. Los gráficos, medidas descriptivas calculadas y tablas elaboradas deben contener una breve interpretación, acorde con lo que se desea mostrara o evidenciar.

Considere la base de datos que corresponde a las medidas antropométricas de la población laboral colombiana (ACOPLA). Las variables de interés son: **SEXO** (Hom, Muj), **P1**(Masa corporal, en kg), **P7**(Perímetro muslo mayor, en cm), **P16**(Perímetro abdominal cintura, en cm), **P22**(Anchura de las caderas, en cm), **P27**(Longitud promedio de los pies, en cm), **P29**(Longitud promedio de las manos, en cm), **P38**(Estatura, en cm), **CAT**\_**IMC** (DELGADO, NORMAL Y OBESO). Recuerde que el objetivo de esta primera parte es detectar patrones, datos atípicos, datos faltantes, tendencias, relaciones, comportamientos y posiblemente realizar comparaciones entre variables.

Cada pareja, deberá seleccionar una muestra aleatoria de 200 registros. Para ello debe seguir las siguientes instrucciones.

Los datos originales están en el archivo: “data.txt”, el cual está disponible en Moodle, pestaña *Vectores Aleatorios y Distribuciones Muestrales, Distancia Estadística*. Para hacerlo debe seguir los siguientes pasos:

# Para leer el archivo de datos, darle click al archivo “data.txt”, aparecerá una nueva pestaña. Copie el contenido y guárdelo en un block de notas, con el mismo nombre: “data.txt”. En R ejecute el siguiente código:

**uno <- read.table(file.choose(), header=T, sep=",")**

Se habilitará un explorador y se escoge el archivo “data.txt”

Copiar el siguiente código en R, sin modificar ninguna de sus líneas

**library(splitstackshape)**

**genera <- function(cedula){**

**set.seed(cedula)**

**aux <- stratified(uno, "CAT\_IMC", 200/2100, bothSets=T)**

**mue <- aux$SAMP1**

**mue**

**}**

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

**datos <- genera(cedula)**

Con la base de datos obtenida responda a las siguientes preguntas

1. (**15 pts.**) Para todas sus variables realice un análisis exploratorio gráfico e identifique posibles valores atípicos u otro tipo de anomalías. (Para las variables Categóricas diagramas de barras, para las continuas o discretas, use Histogramas y/o Boxplot). Comente brevemente.
2. (**15 pts.**) Realice el respectivo proceso de imputación para los datos faltantes en su base de datos. Explique cómo realiza dicha imputación, cuál criterio utiliza y muestre un par de ejemplos ilustrativos.
3. (**20 pts.**) Considere las variables P1, P29 y P38. ¿Se puede afirmar que cada variable por separado permitiría discriminar entre Hombres y Mujeres? Elabore los resúmenes numéricos y gráficos que considere pertinentes para responder la pregunta.
4. (**20 pts.**) Usando las variables continuas, realice un gráfico de dispersión para identificar posibles relaciones entre sus variables. Explique si lo que se observa gráficamente tiene sentido o es coherente a la luz de sus datos. Corrobore lo observado con el cálculo de la matriz de correlaciones. Comente. Repita el proceso discriminando por SEXO. ¿Hay cambios en las estructuras de Covarianzas para ambos grupos? Comente
5. (**15 pts.**) Elabore una tabla de porcentajes de doble entrada con las variables CAT\_IMC y SEXO. Luego presente la información gráficamente. ¿Se puede afirmar que la distribución porcentual de la variable CAT\_IMC es diferente para hombre y mujeres? Justifique su respuesta.
6. (**15 pts.**) Se tienen los siguientes datos de 5 personas, de las cuales se desconoce su CAT\_IMC.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **P1** | **P7** | **P16** | **P22** | **P27** | **P29** | **P38** | **CAT\_IMC** |
| 66.1 | 53.9 | 73.8 | 34.7 | 27.6 | 20.9 | 181.6 |  |
| 55.8 | 50.1 | 76.9 | 39.5 | 24.7 | 17.3 | 154.5 |  |
| 62.8 | 54.3 | 80.4 | 37.5 | 23.5 | 16.5 | 156.6 |  |
| 63.9 | 50.6 | 75.6 | 31.5 | 24.9 | 18.6 | 173.1 |  |
| 50.7 | 46.3 | 72.7 | 30.4 | 23.5 | 16.7 | 159.5 |  |

Usando la distancia estadística, determine a cuál de las tres categorías pertenece cada sujeto. Explique claramente el proceso empleado para clasificar los sujetos. Anexe el código empleado.

El trabajo debe ser cargado en el curso en Moodle, pestaña *Vectores Aleatorios y Distribuciones Muestrales, Distancia Estadística*, en la carpeta *Soporte primer trabajo*, **SOLO** en formato pdf.