EAC 1

Enfermedad coronaria

La prevención es la mejor manera de combatir la enfermedad cardíaca coronaria. Los factores potenciales que influyen en su desarrollo son una combinación que incluye (pero no es exclusiva) factores biológicos, hereditarios y elecciones de estilo de vida.

El archivo **heart2022.txt** tiene 455 filas y 5 columnas con datos que corresponden a un estudio sobre dicha enfermedad y analizaremos algunas de las variables presentes.

Las variables medidas son (correspondiente a cada una de las 5 columnas del archivo)

• ***sbp***: presión arterial sistólica (ésta es la que se denomina “la máxima” en mm de Hg)

• ***ldl***: colesterol de baja densidad en Kg/litro (en esta unidad una medida habitual de 7 o menos suele ser un nivel muy recomendable)

• ***famhist***: antecedentes familiares de enfermedad cardíaca. Es una variable categórica con dos niveles: {Ausente, Presente} codificada por los números {1, 0}.

• ***tipoA***: personalidad y comportamiento tipo A (en una escala convencional). La Personalidad Tipo A o Patrón de Conducta Tipo A es la tendencia de las personas a mostrar ambición, competitividad e implicación laboral. Los aspectos que rodean a la personalidad tipo A o patrón de conducta tipo A ha hecho que se hayan realizado muchas investigaciones sobre su relación con los problemas de salud, en concreto las enfermedades cardiovasculares o la hipertensión, y con las respuestas de ansiedad.

• ***chd***: enfermedad cardíaca coronaria {sí, no} codificada por los números {1, 0}

a) Represente los valores de las dos variables **sbp** y **ldl** en las dos formas indicadas en la clase (el estilo *serie de tiempo* y el estilo *constante vs valor medido*). Comente una primera impresión sobre la simetría de la distribución de los datos a la vista de estos gráficos.

b) Considere nuevamente esas variables **sbp** y **ldl** medidas en las 455 personas. Muestre en una tabla los valores de la media muestral, mediana, desvío estándar, cuartiles primero y tercero, rango intercuartílico, y los coeficientes de simetría y kurtosis redondeando las respuestas en dos decimales.

c) Realice los diagramas boxplot para cada una de estas dos variables y comente sobre la presencia de datos extremos o outliers. Calcule el porcentaje de cada tipo de dato extremo en ambas variables.

d) Realice los polígonos de frecuencias relativas de estas dos variables a partir de considerar la agrupación de datos en intervalos de clase de igual amplitud.

e) Muestre un gráfico de dispersión donde cada punto tenga coordenadas (sbp, ldl). Calcule la proporción de datos para los que sbp < 130 y ldl < 8. Esa zona corresponde a una zona que muchos médicos consideran adecuada.

f) Para cada una de las variables sbp, ldl, tipoA, cada una de ellas separadas según la variable chd, obtenga los gráficos de **frecuencias acumuladas (empíricas) con datos no agrupados.** Viendo el gráfico, una persona indica que los valores de ldl tienden a ser más altos entre quienes tienen la enfermedad coronaria que entre quienes no la padecen, ¿Es esto cierto?

g) Considerando cada una de las variables: sbp, ldl, y tipoA, separe las observaciones en tres grupos: A) No tienen enfermedad coronaria y no tienen antecedentes familiares, B) Tienen una de las dos condiciones, C) tienen ambas condiciones. Obtener los cuartiles 1 y 3 de los tres grupos.

h) Para cada una de las variables: sbp, ldl, tipoA haga boxplots paralelos considerando la historia familiar (famhist) y la presencia de enfermedad coronaria (chd) (grupos A, B y C). Alguien dice: en términos generales, resulta evidente que el colesterol de baja densidad tiende a ser mayor entre quienes tienen antecedentes familiares de enfermedad cardíaca coronaria y padecen enfermedad cardíaca coronaria que en los demás individuos. ¿Verdadero o falso? Justificar.
Entonces son tres grupos de boxplots: un grupo para cada variable y en ese grupo un boxplot para cada grupo de personas A, ó B ó C.
Para generar cada grupo tenga en cuenta que cada boxplot de una variable tendrá diferente cantidad de casos (ver ejemplo en Octave al finalizar el enunciado del trabajo).

i) Realizar e interpretar los histogramas de las variables: sbp, ldl, tipoA separando los datos en los tres grupos indicados (A,B y C).

j) Calcular las medias muestrales de las tres variables, separando los datos en los tres grupos. Para cada variable realizar un perfil de comportamiento de la media muestral según los grupos. Calcular e interpretar el coeficiente de simetría muestral de las variables sbp, ldl, tipoA separando según grupos (A,B y C).

Ejemplo en Octave de generación de boxplots en una misma figura.

Observe el uso de los { } para generar la estructura de tres vectores con diferente número de componentes.

```
boxplot({rand(100,1) exp(rand(300,1)) -log(rand(600,1))})
```