Introducción a la Estadística y Ciencia de Datos - Segundo cuatrimestre 2021

Práctica 1 - Estadística Descriptiva

---

1. El archivo `Debernardi.csv` contiene los datos referentes a un estudio acerca del cáncer de páncreas (más información en el archivo *Acerca de los datos*, en el Aula Virtual).

   a) Construir una tabla con los valores observados para la variable DIAGNOSIS y su frecuencia relativa.

   b) Realizar un gráfico de barras usando la tabla del ítem anterior.

2. El archivo `datos_titanic.csv` contiene información sobre una muestra seleccionada al azar de las personas, no tripulantes, que viajaban en el barco tristemente célebre *Titanic*, al momento de su hundimiento en el Océano Atlántico (más información en el archivo *Acerca de los datos*, en el Aula Virtual).

   a) Estimar la probabilidad de ser mujer sabiendo que sobrevivió y comparar con la estimación de ser mujer a bordo del *Titanic*.

   b) Hacer una tabla de contingencia entre las variables categóricas SURVIVED y PCLASS.

   c) Realizar un gráfico de barras que vincule a las variables categóricas SURVIVED y PCLASS.

3. En un experimento se midió la temperatura de sublimación del iridio y del rodio. En los archivos `iridio.txt` y `rodio.txt` se encuentran los datos recabados en el experimento.

   a) Comparar los dos conjuntos de datos mediante histogramas y boxplots, graficando los boxplots en paralelo.

   b) Hallar las medias, las medianas y las medias podadas al 10 % y 20 % muestrales. Comparar.

   c) Hallar los desvíos estándares, las distancias intercuartiles y las MAD muestrales como medidas de dispersión.

   d) Hallar los cuantiles 0,90, 0,75, 0,50, 0,25 y 0,10.

4. En un estudio nutricional se consideran las calorías y el contenido de sodio de tres tipos de salchichas y se obtuvieron los datos que se encuentran en los archivos `salchichas_A.txt`, `salchichas_B.txt` y `salchichas_C.txt`.

   a) Realizar un histograma para las calorías de cada tipo de salchichas. ¿Observa grupos en algún gráfico? ¿Cuántos grupos observa? ¿Observa algún candidato a dato atípico? ¿Alguno de los histogramas tiene una característica particular?

   b) Repetir con la cantidad de sodio.

   c) Realizar los boxplots paralelos para las calorías. ¿Observa la misma cantidad de grupos que antes? ¿A cuál conclusión llega? De acuerdo con los boxplots graficados, ¿cómo caracterizaría la diferencia entre los tres tipos de salchichas desde el punto de vista de las calorías?

5. El conjunto de datos que figura en el archivo `estudiantes.txt` corresponde a 100 determinaciones repetidas de la concentración de ion nitrato (en μg/l), 50 de ellas corresponden a un grupo de estudiantes (Grupo 1) y las restantes 50 a otro grupo (Grupo 2).

   a) Estudiar si la distribución de los conjuntos de datos para ambos grupos es normal, realizando los correspondientes histogramas y superponiendo la curva normal. Además dibujar los qqplots para cada conjunto de datos superponiendo, en otro color, la recta mediante el comando qqline.

   b) ¿Le parece a partir de estos datos que ambos grupos están midiendo lo mismo? Responder comparando medidas de centralidad y de dispersión de los datos. Hacer boxplots paralelos.

6. Con la finalidad de incrementar las lluvias en zonas desérticas, se desarrolló un método que consiste en el bombardeo de la nube con átomos. Para evaluar la efectividad del método se realizó el siguiente experimento:

   - Para cada nube que se podía bombardear se decidió al azar si se la trataba o no.
   - Las nubes no tratadas fueron denominadas nubes controles.

   En el archivo `nubes.txt` se presentan la cantidad de agua caída de 26 nubes tratadas y 26 nubes controles.

   a) Realizar boxplots paralelos. ¿Le parece que el método produce algún efecto?

   b) Analizar la normalidad realizando qqplots e histogramas (de densidad) para ambos conjuntos de datos y superponiendo la curva normal.

   c) Realizar la transformación logaritmo natural a los datos (log en R) y repetir *b*) para los datos transformados.

   d) Realizar boxplots paralelos habiendo transformado las variables con el logaritmo natural. Observar cómo se modificaron los datos atípicos respecto del ítem *a*).

7. El archivo `cpu.txt` contiene tiempos, en segundos, de CPUs correspondientes a 1000 trabajos enviados por una consultora. Para este conjunto de datos:

   a) Calcular la media muestral, la mediana muestral y la media α-podada muestral con $\alpha = 0{,}10$ (10 %).

   b) Calcular el desvío estándar, la distancia intercuartil y la MAD muestrales.

   c) Realizar un histograma, un gráfico de la densidad estimada usando la función density y un boxplot. ¿Cuáles son las características más sobresalientes? ¿Se observan datos atípicos?

   d) ¿Cree que los datos tienen distribución normal? Hacer un qqplot para constatar su conjetura o ponerla en duda.

   e) ¿Qué medida de posición considera más apropiada para describir el centro de los datos?

8. El archivo `Islander_data.csv` contiene los datos referentes a un experimento acerca de los efectos de un medicamento contra la ansiedad en un test de memoria cuando se expone a la persona ante recuerdos felices y tristes (más información en el archivo *Acerca de los datos*, en el Aula Virtual). Se quiere estudiar la diferencia, en segundos, entre el tiempo logrado en un test de memoria antes y después de tomar el medicamento.

   a) Realizar un histograma con las realizaciones de la variable aleatoria DIFF, la diferencia de tiempos.

   b) Estimar $\mathbb{P}(\text{DIFF} \leq 1)$.

   c) Graficar la función de distribución empírica de la variable DIFF.

   d) Estimar la densidad de DIFF usando estimadores basados en núcleos, utilizando diferentes ventanas ($h = 0{,}5, 1{,}5$ y $2{,}5$) y núcleos (rectangular, gaussiano y de Epanechnikov). ¿Qué observa?

9. Considerar nuevamente el conjunto de datos del ejercicio 1.

   a) Realizar histogramas para la variable LYVE1 basados en los datos brindados para las observaciones que cumplen DIAGNOSIS=1, DIAGNOSIS=2 y DIAGNOSIS=3. Es decir efectuar histogramas según los niveles de la variable factor DIAGNOSIS. Indicar las características más sobresalientes de los histogramas y aquellas que los diferencian.

   b) Graficar, en distintos colores y superpuestas, las funciones de distribución empíricas de la variable LYVE1 según los niveles de la variable factor DIAGNOSIS. Decidir si la siguiente afirmación es verdadera o falsa y justificar: "los valores de la variable LYVE1 tienden a ser más altos entre quienes tienen cáncer de páncreas que entre quienes sufren otras enfermedades asociadas al páncreas".

   c) Realizar boxplots paralelos para la variable LYVE1 según los niveles de la variable factor DIAGNOSIS, considerando el sexo de los pacientes (variable SEX). Decidir si la siguiente afirmación es verdadera o falsa y justificar: "en términos generales, el sexo del paciente no afecta los niveles de la proteína que se mide en la variable LYVE1".

   d) Graficar superpuestas las densidades estimadas, que brinda la función density, para la variable LYVE1 según los niveles de la variable factor DIAGNOSIS. Describir las características más sobresalientes de las densidades estimadas y aquellas que las diferencian.