Introducción a la Estadística y Ciencia de Datos

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 campus).

   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 campus').

   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. A partir de esta tabla estimar la probabilidad de sobrevivir dada la clase para los distintos valores de la variable 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) Armar un archivo que se llame `salchihas.txt` que contenga toda la información registrada en los tres archivos mencionados agregando una columna que indique el tipo de salchicha en cada caso.

   b) 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?

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?

d) Repetir con la cantidad de sodio.

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 `data_credit_card.csv` tiene información de `n=500` clientes de un banco, para las siguientes variables: `purchases` es el monto total de compras en el último año, `credit_limit` es el límite de crédito disponible para el cliente, `purchases_freq` es la proporción de semanas del año en las que el cliente realizó compras y `tenure` es la cantidad de meses que restan al cliente para cancelar el crédito. Se pide:

   a) Para todas las variables, graficar la función de distribución empírica. Discutir sobre el tipo de variable aleatoria que utilizaría para modelar en cada caso.

b) Para la variable `credit_limit` hacer un histograma y un gráfico de densidad usando la funcion `density`, ¿Qué observa? ¿Le parece adecuado realizar estos gráficos para las variables `purchases` y `tenure`?

c) Para la variable `tenure` hacer un *barplot* con las frecuencias relativas de cada valor. ¿Qué observa?

d) Para todas las variables, calcular la media, la mediana y la media $\alpha-$podada (con α =0.1). Comparar los resultados y justificar. ¿Qué medida de posición del centro de los datos le parece más adecuada en cada caso?

e) Para todas las variables, obtener los cuantiles de nivel 0.25 y 0.75 de los datos. Calcular el rango inter-cuartílico y la MAD muestrales. Graficar *boxplots*. ¿Qué observa?

f) Calcular el desvío estándar, el coeficiente de asimetría y el coeficiente de curtosis muestrales. Interpretar los resultados en relación a las distribuciones vistas.

g) Identificar datos atípicos. ¿Deberían excluirse? ¿Cómo se modifican las medidas obtenidas anteriormente si se los excluye?

8. En el archivo `ciclocombinado.xlsx` hay datos de la potencia entregada por una central térmica de ciclo combinado. Se registraron datos diarios de la potencia máxima entregada (PE, en MW) por la planta funcionando en capacidad máxima. La variable `HighTemp` vale 1 si la temperatura media diaria fue superior a 20°C en el día en el que se tomó el dato y vale 0 en caso contrario.

   a) Realizar un histograma y un gráfico `density` con los datos de PE, ¿Qué se observa?

   b) Clasificar los datos en dos vectores según la variable `HighTemp` y realizar gráficos `density` separados. Visualizar simultáneamente los gráficos en la misma escala. ¿Qué se observa?

   c) Estimar $P(\texttt{PE} < 450|\texttt{HighTemp} = 0)$ y $P(\texttt{PE} < 300|\texttt{HighTemp} = 1)$.

   d) Estimar $P(\texttt{PE} < 450)$.

   e) Estimar la potencia mínima garantizada con probabilidad 0.9 para un cierto día con `Hightemp = 1`.

   f) Estimar la potencia mínima garantizada con probabilidad 0.9 para un cierto día.

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.

e) Repetir a) y d) para el logaritmo de LYVE1.

10. **Boxplot para la distribución normal**. Sea $X \sim N(\mu, \sigma^2)$. Escribir, en términos de μ, σ y los cuantiles de la distribución normal estándar,

a) la mediana.

b) El cuartil superior (Q_S), o tercer cuartil, es decir, el cuantil 0.75 de la distribución, $Q_S = F_X^{-1}(0.75)$.

c) El cuartil inferior (Q_I), o primer cuartil, es decir, el cuantil 0.25 de la distribución de X.

d) El rango intercuartil o distancia intercuartil, $\text{RIQ} = Q_S - Q_I$.

e) El máximo posible valor del "bigote" superior, dado por $Q_S + 1.5\text{RIQ}$.

f) El mínimo posible valor del "bigote" inferior.

g) ¿Cuál es la probabilidad de que una observación de la variable X caiga en el intervalo comprendido entre el Q_I y el Q_S?

h) ¿Cuál es la probabilidad de que una observación de la variable X caiga en el intervalo comprendido entre el Q_S y el $Q_S + 1.5\text{RIQ}$?

i) ¿Cuál es la probabilidad de que una observación de la variable X caiga fuera del intervalo comprendido entre el $Q_I - 1.5\text{RIQ}$ y el $Q_S + 1.5\text{RIQ}$? Es decir, ¿cuál es la probabilidad de que una observación que proviene de una distribución normal sea etiquetada como *outlier* o atípica por el boxplot? Observemos que esto es lo mismo que preguntar, ¿qué proporción de observaciones provenientes de una distribución normal esperaría que fueran marcados con asterisco (como *outliers*) en un boxplot?

11. **Relación entre distintas medidas de dispersión para la distribución normal**. Sea $X \sim F = N(\mu, \sigma^2)$. Escribir, en términos de μ y σ

a) El desvío estándar poblacional, es decir,

$$sd_F(X) = \sqrt{Var_F(X)} = \sqrt{E_F\left((X - E_F(X))^2\right)},$$

que estudiamos en Proba (M). (En este ítem no hay que hacer cuentas).

b) La MAD (*Median Absolute Deviation*) en su versión poblacional, es decir, la

$$MAD_F(X) = Med_F\left\{|X - Med_F(X)|\right\}$$

donde $Med_F(W)$ es la notación para indicar la mediana poblacional de la variable aleatoria W que tiene distribución F, es decir Es decir, $Med_F(W) = F^{-1}(1/2)$.

i. Escribir a $Med_F(X)$ en términos de μ y σ.

ii. Escribir la función de distribución acumulada de la v.a. $W = |X - Med_F(X)|$ en términos de μ, σ y la función de distribución acumulada de la normal estándar, que notaremos Φ. Chequear que vale

$$F_W(t) = 2\left[\Phi(t/\sigma) - 1\right] I_{(0,+\infty)}(t).$$

iii. Deducir del ítem anterior que la MAD de X resulta ser $\Phi^{-1}(0.75)\sigma$.

c) El rango intercuartil de X. Verificar que

$$RIQ_F(X) = 2\Phi^{-1}(0.75)\sigma.$$

d) Explique por qué el RIQ y la MAD se dividen por 1.35 y 0.675, respectivamente, para poder ser comparados con el desvío estándar en el caso de tener una muestra normal.

12. **QQ-plot normal**. En el QQ-plot se grafican las observaciones ordenadas versus los percentiles de una distribución teórica de interés.

a) Sean $X_1, \ldots, X_n$ v.a.i.i.d. con distribución $N(0,1)$. Una realización de estas variables aleatorias da lugar a las observaciones (o datos observados) $x_1, \ldots, x_n$. A los datos ordenados de forma creciente los notaremos por

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

Para $\alpha \in (0,1)$ definimos el $\alpha-$cuantil muestral como la observación que ocupa el lugar $[\alpha(n+1)]$ en la muestra ordenada, es decir $x_{([(n+1)\alpha])}$, donde [ ] indica la parte entera.

i ¿Para qué valor de α resulta ser $x_{(1)}$ el $\alpha-$cuantil muestral? ¿Y $x_{(2)}$? ¿Y en general $x_{(i)}$?

ii Utilizando a Φ, la función de distribución acumulada de una normal estándar, exprese el $\alpha-$cuantil poblacional correspondiente a $\alpha = \frac{i}{n+1}$.

El QQ-plot es el gráfico de los puntos $\left(\Phi^{-1}\left(\frac{i}{n+1}\right), x_{(i)}\right)$, con $i = 1, \ldots, n$. Si $X_1, \ldots, X_n \sim F$, para cada $t \in \mathbb{R}$, la ley de los Grandes Números aplicada a la distribución Bernoulli garantiza que la función de distribución empírica evaluada en t, $\widehat{F}_n(t) = \sum_{i=1}^n I_{(-\infty,t]}(X_i)$ converge casi seguramente y en probabilidad a $F(t)$ cuando $n \to \infty$. Como $\widehat{F}_n(t)$ aproxima a $F(t)$ para todo t, bajo los supuestos de este ejercicio, los puntos del QQ-plot se ubicarán cerca de la recta identidad (de pendiente 1 y ordenada al origen 0).

b) ¿Qué cambiará en el QQ-plot cuando las variables $Y_1, \ldots, Y_n$ sean i.i.d. con distribución $N(\mu, \sigma^2)$ pero seguimos graficando en el eje horizontal los cuantiles teóricos de la $N(0,1)$? Para responder a esto, sea $y_1, \ldots, y_n$ una realización de $Y_1, \ldots, Y_n$

i Relacionar a $\left\{y_{(i)}\right\}_{1 \leq i \leq n}$ con $\left\{z_{(i)}\right\}_{1 \leq i \leq n}$, siendo $z_i = \frac{y_i - \mu}{\sigma}$.

ii Por el ítem anterior, ¿dónde se ubicarán aproximadamente los puntos de la forma $\left(\Phi^{-1}\left(\frac{i}{n+1}\right), z_{(i)}\right)$?

iii Deducir de los incisos anteriores donde se ubicarán aproximadamente los puntos del QQ-plot de las observaciones $y_1, \ldots, y_n$.