ESTADÍSTICA I **Curso 2020/2021**
Doble Grado en Ing. Informática-Matemáticas

Relación 1 de problemas

1. Realiza un análisis descriptivo de los datos británicos de ingresos familiares en 1975 (reescalados dividiendo por la media) contenidos en el fichero Datos-ingresos.txt. En concreto, calcula los estadísticos de posición o tendencia central, las medidas de dispersión, representa un diagrama de cajas y un estimador del núcleo de la función de densidad. Comenta los resultados.

2. Demuestra que

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \min_{a\in\mathbb{R}} \sum_{i=1}^{n}(x_i - a)^2.$$

¿Qué significa esto en relación con la interpretación intuitiva de la media muestral?

3. Representa en el mismo gráfico los diagramas de cajas correspondientes a la variable *Largo* del fichero tortugas.txt para los ejemplares hembra y para los ejemplares macho. Emplea colores distintos para los dos diagramas.

4. Los datos del fichero kevlar.txt corresponden al tiempo hasta el fallo (en horas) de 101 barras de un material utilizado en los transbordadores espaciales, llamado Kevlar49/epoxy, sometidas a un cierto nivel de esfuerzo. Los datos han sido tomados de Barlow et al. (1984).

(a) Calcula las principales medidas descriptivas numéricas de estos datos.
(b) Representa un diagrama de cajas.
(c) Representa un histograma con un número de clases apropiado.
(d) Estudia la presencia de datos atípicos en la muestra. Si hay datos atípicos, suprímelos y repite todos los apartados anteriores. Compara los resultados obtenidos.

5. El paquete *gapminder* contiene un fichero de datos de población, esperanza de vida y renta per cápita de los países del mundo entre 1952 y 2007. Instala el paquete y lleva a cabo los siguientes gráficos:

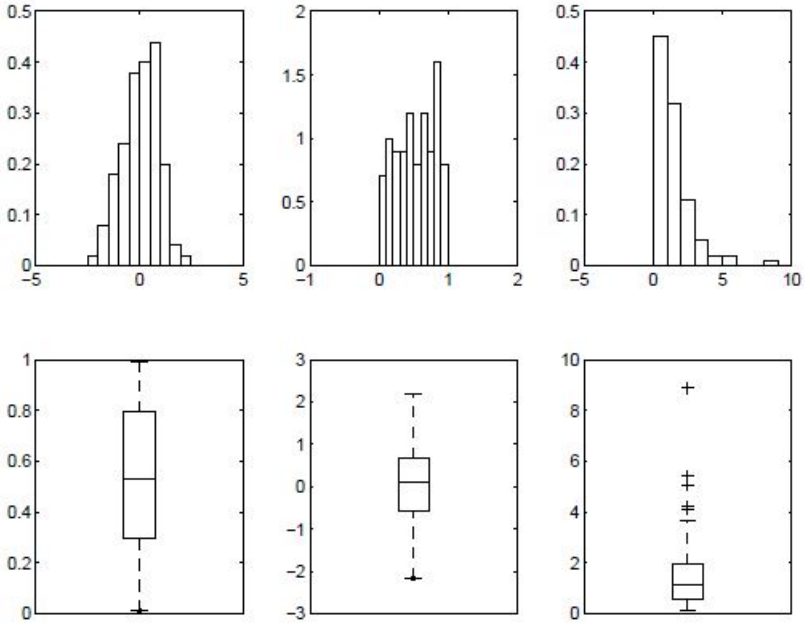
(a) Un histograma de la esperanza de vida en 2007 de los países de Europa.
(b) Diagramas de cajas con las esperanzas de vida de cada continente en el año 1952.
(c) Un diagrama de dispersión de la renta per cápita y la esperanza de vida de cada país en el año 2007.
(d) Mejora el gráfico anterior representando cada punto de un color diferente en función del continente al que pertenece cada país y representando la renta per cápita en una escala logarítmica (mira esta documentación).

6. Determina razonadamente si las siguientes afirmaciones son verdaderas o falsas:

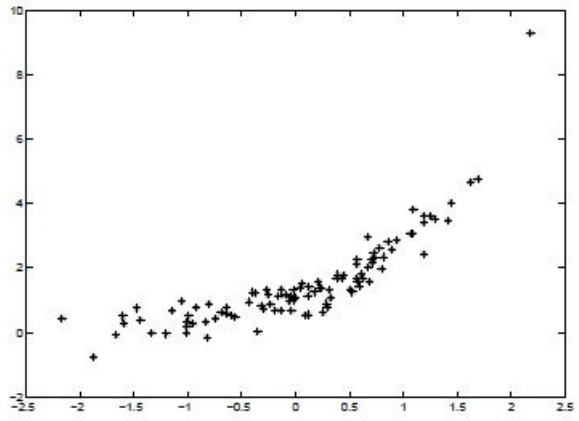
(a) Si añadimos 7 a todos los datos de un conjunto, el primer cuartil aumenta en 7 unidades y el rango intercuartílico no cambia.
(b) Si todos los datos de un conjunto se multiplican por -2, la desviación típica se dobla.
(c) Si todos los datos de un conjunto se multiplican por 2, la varianza se dohbla.
(d) Si cambiamos el signo de todos los datos de un conjunto, el coeficiente de asimetría también cambia de signo.
(e) Al multiplicar por tres todos los datos de un conjunto, el coeficiente de asimetría no varía.
(f) Si el coeficiente de correlación entre dos variables vale -0.8, los valores por debajo del promedio de una variable están asociados con valores por debajo del promedio de la otra.
(g) Si para todo i, se cumple $y_i < x_i$, el coeficiente de correlación entre x e y es negativo.
(h) Al restar una unidad a cada dato de un conjunto, la desviación típica siempre disminuye.
(i) Si a un conjunto de datos con media $\bar{x}$ se le añade un nuevo dato que coincide con $\bar{x}$, la media no cambia y la desviación típica disminuye.

7. Calcula el diagrama de dispersión de las dos variables correspondientes al peso y a la circunferencia de abdomen que aparecen en el fichero Datos-bodyfat.txt. Calcula la recta de regresión y el coeficiente de correlación. Comenta los resultados. Análogas preguntas para las dos variables X e Y del fichero Datos-geyser.txt, que corresponden a la duración de las erupciones y el tiempo hasta la siguiente erupción de un geyser.

8. Para tres conjuntos de datos se han representado los correspondientes histogramas y diagramas de cajas. Relaciona cada histograma con el diagrama de cajas que le corresponde:



9. Se presenta a continuación el diagrama de dispersión correspondiente a dos variables:



Contesta a las siguientes preguntas:

(a) ¿Existe relación entre las variables?
(b) ¿Hay algún dato atípico?
(c) De los tres valores siguientes: 0.01, 0.83 y -0.73, ¿cuál crees que podría corresponder al coeficiente de correlación entre x e y?

9. Un estudio sobre el efecto de la temperatura en el rendimiento de un proceso químico proporciona los siguientes resultados:

Temperatura (x)	-5	-4	-3	-2	-1	0	1	2	3	4	5
Rendimiento (y)	1	5	4	7	10	8	9	13	14	13	18

(a) Representa el diagrama de dispersión de los datos anteriores y calcula el coeficiente de correlación entre las dos variables. ¿Se puede admitir que existe una relación lineal aproximada entre ambas, es decir, $y_i \approx a + bx_i$?
(a) Calcula el término independiente y la pendiente de la recta de mínimos cuadrados.
(a) ¿Qué rendimiento predecirías para un nuevo proceso realizado a temperatura $x = 3{,}5$?

10. Disponemos de un conjunto de observaciones $x_1, \ldots, x_{100}$, ya ordenadas de menor a mayor, cuya media muestral es $\bar{x}$. Creamos una nueva muestra añadiendo a la anterior los valores x_1 y x_{100}. ¿Qué condición se debe cumplir para que la media muestral de la nueva muestra coincida con $\bar{x}$, la media muestral de la muestra original?

11. Sea $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ una muestra de datos bivariantes, con media muestral $(\bar{x}, \bar{y})$. Añadimos a la muestra el punto $(\bar{x}, \bar{y})$. Determina si la covarianza de la nueva muestra es mayor, menor o igual que la de la muestra original.

12. Tenemos una muestra $x_1, \ldots, x_n$ cuya media es $\bar{x}$ y cuya varianza muestral es s_n^2. Duplicamos ahora el tamaño muestral añadiendo los valores de signo opuesto a los originales:

$$x_1, \ldots, x_n, -x_1, \ldots, -x_n.$$

Llamamos $\tilde{s}_n^2$ a la varianza muestral de esta segunda muestra. ¿Cuál es mayor, s_n^2 ó $\tilde{s}_n^2$?

13. Dada una muestra $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, con $n \geq 2$, se pide obtener la recta $\hat{y} = \hat{b}\, x$, que pasa por el origen $(0, 0)$ y minimiza la suma de los residuos al cuadrado entre todas las rectas de ecuación $y = b\, x$. Escribe la fórmula de $\hat{b}$ y la expresión de la suma de los residuos al cuadrado en función de los valores muestrales.

14. Sea una muestra $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, con $n \geq 2$. Para realizar el ajuste lineal al conjunto de datos, suponemos que las observaciones tienen importancias relativas diferentes, cuantificadas mediante unos pesos $\omega_1, \ldots, \omega_n$ (es decir, $0 \leq \omega_i \leq 1$ para todo $1 \leq i \leq n$ y $\sum_{i=1}^{n} \omega_i = 1$). El error cuadrático medio ponderado (ECMP) obtenido al aproximar la muestra mediante la recta de ecuación $y = a + bx$ se define como

$$\mathrm{ECMP}(a, b) := \sum_{i=1}^{n} \omega_i (y_i - (a + bx_i))^2.$$

Halla los valores de a y b que minimizan $\mathrm{ECMP}(a, b)$.

15. El fichero star.txt contiene la temperatura y la intensidad de la luz en un conjunto de estrellas. Calcula y representa la recta de mínimos cuadrados para explicar la temperatura en función de la intensidad de la luz. Comenta el resultado.