**ESTADÍSTICA I**
**Grado en Matemáticas (2023/24)**

## Tema 1: Estadística descriptiva

**1.1.** Descarga el fichero de datos `data_banknote_authentication.txt` del repositorio de aprendizaje automático de la UC Irvine. Se han obtenido de imágenes digitalizadas de billetes de banco auténticos y falsificados, pero la información del repositorio no incluye cuál de las etiquetas (0 ó 1) corresponde a los auténticos. De cada imagen se obtuvieron cuatro variables continuas:

1. Varianza de la transformada por ondículas de la imagen.
2. Asimetría de la transformada por ondículas de la imagen.
3. Curtosis de la transformada por ondículas de la imagen.
4. Entropía de la imagen.

Realiza un análisis exploratorio de los datos y compara los dos grupos. Calcula estadísticos de posición y dispersión y realiza representaciones gráficas. Extrae conclusiones.

**1.2.** 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?

**1.3.** Utilizando el fichero de datos `worldbank,` disponible en el paquete de R llamado `ks`, analiza los datos de emisiones de CO2 y de Producto Interior Bruto per cápita (GDP.capita) que se incluyen en el fichero para diferentes países del mundo. ¿Hay mucha diferencia entre las medias y las medianas? ¿Dirías que estos datos han sido extraídos de variables aleatorias con distribución normal?
Nota.- Observa que hay bastantes "missing values" en el fichero (denotados con NA). El comando `na.omit` permite descartarlos sin necesidad de suprimirlos manualmente.

**1.4.** Consideramos los pingüinos de la especie `Gentoo` que aparecen en el conjunto de datos `penguins` de la librería `palmerpenguins` de `R`. Para información detallada de estos datos escribir `help(penguins)` en la consola de `R`.

**a)** Realizar un análisis descriptivo univariante de la longitud del pico: calcular los estadísticos de tendencia central, las medidas de dispersión y representar un diagrama de cajas y un estimador núcleo de la función de densidad. ¿Qué se observa en relación con las modas? Estudia la presencia de datos atípicos en la muestra.
**b)** Dibuja el diagrama de dispersión de la longitud del pico y de la aleta. Selecciona una de las dos como variable respuesta y la otra como explicativa. Calcula la recta de regresión y el coeficiente de correlación. Interpreta los resultados.

**1.5.** 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.

**1.6.** 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 dobla.
**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.

**1.7.** Explica brevemente, línea a línea, qué hace el siguiente código de R:

```
X <- rnorm(30,-1,2)
summary(X)
quantile(X,0.9)
var(X)
library(moments)
skewness(X)
XO <- sort(X)
XO[1]
length(X)
f <- density(X)
```

De manera razonada, haz corresponder cada uno de los resultados que se ven debajo con el comando del código anterior que lo generó:

| | |
|---|---|
| `[1] 30` | `90%`<br>`1.210152` |
| `[1] 0.2894119` | `[1] -4.293553` |
| `Min. 1st Qu. Median Mean 3rd Qu. Max.`<br>`-4.2936 -2.2853 -1.1732 -1.1023 0.1571 3.2101` | `[1] 3.47647` |

**1.8.** Ejecuta el siguiente código en R:

```
n <- 50
x <- rnorm(n)
y <- exp(x) + rnorm(n,sd=0.5)
plot(x,y,type="p")
```

**a)** ¿Existe relación entre las variables? ¿Es lineal?
**b)** ¿Es linealizable? Si es así, ajusta la recta de regresión a los datos linealizados.
**c)** ¿Hay algún dato atípico?
**d)** De los tres valores siguientes: 0.01, 0.80 y -0.73, ¿cuál crees que podría corresponder al coeficiente de correlación entre $x$ e $y$?

**1.9.** 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?

**1.10.** 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.

**1.11.** 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 $\tilde{s}_n^2$?

**1.12.** 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 $\hat{b}$.

**1.13.** 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.

**1.14.** Sea $\hat{f}_n$ un estimador núcleo, con parámetro de suavizado $h_n$, construido a partir de una muestra $x_1, \ldots, x_n$. Definimos una nueva variable $X^0$ como $X^0 = X^* + h_n Z$, donde $X^*$ es la variable que toma los valores $x_1, \ldots x_n$ con probabilidad $1/n$ cada uno de ellos, y $Z$ es una variable con distribución $N(0, 1)$. Demuestra que $X^0$ es una v.a. con distribución absolutamente continua (es decir, tiene una función de densidad) y que la correspondiente densidad es $\hat{f}_n$, con núcleo $K$ normal estándar (N(0,1)).
Indicación: Calcula las distribuciones condicionadas de $X^0 | X^* = x_i$ y utiliza la fórmula de las probabilidades totales.