

Práctico 5.

Análisis exploratorio de datos

Ejercicio 1

Los datos siguientes corresponden al tiempo (seg.) necesario para procesar 25 trabajos en una CPU.

| | | | | |
|------|------|------|------|------|
| 1.17 | 1.61 | 1.16 | 1.38 | 3.53 |
| 1.23 | 0.82 | 0.96 | 2.01 | 0.15 |
| 2.11 | 0.71 | 0.02 | 1.59 | 0.19 |
| 1.91 | 2.16 | 0.92 | 0.75 | 2.59 |
| 3.07 | 1.1 | 3.76 | 0.47 | 4.75 |

- ¿Cuál es la unidad de análisis, la variable en estudio y de qué tipo es?.
- Identificar la población y la muestra en estudio.
- Realizar una tabla de frecuencias organizando los datos en 5 intervalos y graficar adecuadamente.
- ¿Qué porcentaje de trabajos fueron procesados en menos de 3.80 segundos?
- ¿Qué cantidad de trabajos fueron procesados en al menos 1.92 segundos?
- ¿Cómo se interpreta la frecuencia relativa del tercer intervalo?
- Obtener un resumen descriptivo de los tiempos necesarios para procesar los 25 trabajos.
- Realizar un histograma y un diagrama de cajas en R.

Ejercicio 2

Utilizaremos la base de datos del archivo StudentSurvey.csv, usando R.

- ¿Cuál es la unidad de análisis, qué variables se han medido y de qué tipo son? ¿cuál es la dimensión de la base de datos?
- Renombrar las variables con nombres en castellano.
- Convertir los valores de la variable Altura de pulgadas a centímetros.
- Representar gráficamente la variable Año.
- ¿Existe algún tipo de asociación entre Género y Fuma?
- ¿Hay valores atípicos de Altura?

- g) ¿Cambia la distribución de Altura según Año?
- h) ¿Los datos sugieren algún tipo de relación entre Peso y Altura?

Ejercicio 3

El siguiente conjunto de datos representa el número de nuevas cuentas de email registradas durante diez días consecutivos:

| | | | | | | | | | |
|----|----|----|----|----|-----|----|----|----|----|
| 43 | 37 | 50 | 51 | 58 | 105 | 52 | 45 | 45 | 10 |
|----|----|----|----|----|-----|----|----|----|----|

- a) ¿Cuál es la unidad de análisis, cuál es la variable y a qué tipo corresponde?
- b) Calcular la media, la mediana, los cuartiles y la desviación estándar.
- c) Hallar los valores atípicos.
- d) ¿Cómo afecta la presencia de valores atípicos en las medidas descriptivas básicas (media, mediana, cuartiles y desvío estándar).

Ejercicio 4

Un proveedor de red investiga la velocidad de descarga de su red. El número de usuarios conectados es registrado en cincuenta localidades (miles de personas),

| | | | | | | | | | |
|------|------|------|------|------|------|------|------|------|------|
| 17.2 | 22.1 | 18.5 | 17.2 | 18.6 | 14.8 | 21.7 | 15.8 | 16.3 | 22.8 |
| 24.1 | 13.3 | 16.2 | 17.5 | 19 | 23.9 | 14.8 | 22.2 | 21.7 | 20.7 |
| 13.5 | 15.8 | 13.1 | 16.1 | 21.9 | 23.9 | 19.3 | 12.0 | 19.9 | 19.4 |
| 15.4 | 16.7 | 19.5 | 16.2 | 16.9 | 17.1 | 20.2 | 13.4 | 19.8 | 17.7 |
| 19.7 | 18.7 | 17.6 | 15.9 | 15.2 | 17.1 | 15.0 | 18.8 | 21.6 | 11.9 |

- a) Identificar población, muestra, unidad de análisis y variable.
- b) Calcular la media, la varianza y el desvío estándar del número de usuarios conectados.
- c) Hallar el resumen de cinco números y construir un diagrama de cajas.
- d) Calcular el rango intercuartílico. ¿Hay algún valor atípico?
- e) Se supone que el número de usuarios conectados sigue una distribución normal. ¿El histograma apoya esa hipótesis?

Ejercicio 5 *Tiempos de espera entre pulsos nerviosos*

Wasserman (2004) menciona el trabajo de Cox y Lewis (1966) en el cual se analizan 799 tiempos de espera entre pulsos sucesivos a lo largo de una fibra nerviosa.

En la biblioteca ACSWR se encuentran los datos **The Nerve Data**. Ver el enlace ACSWR para una descripción de los mismos.

- a) Obtener un resumen de 5 números (valor mínimo, valor máximo, cuartiles y mediana) e interpretar dichos valores.
- b) Construir un diagrama de cajas y describir lo que informa sobre la distribución de los datos.
- c) ¿Es simétrica la distribución empírica?

Ejercicio 6

A continuación se presentan los pesos (kg) y las cantidades de combustible consumidas en rutas (km/naf) de 7 marcas de automóvil elegidas al azar:

| | | | | | | | |
|------------------------|------|------|------|------|------|------|------|
| Peso | 1443 | 1568 | 1465 | 1811 | 1109 | 1136 | 1040 |
| Consumo de combustible | 43 | 46 | 44 | 39 | 59 | 55 | 59 |

- a) Calcular los estadísticos de posición y de dispersión que sean posibles para cada variable.
- b) ¿Cuál de las dos variables presenta más variabilidad? Justificar.
- c) Realizar el diagrama de dispersión para las variables y describir la relación existente entre ellas.
- d) ¿Qué sugiere el resultado de un plan nacional para reducir el consumo de combustible importado?