I. ANÁLISIS DESCRIPTIVO DE UN CONJUNTO DE DATOS

1 Series Estadísticas. Distribuciones de frecuencias.

1.1 Definición de Estadística.

1.1.1 Conceptos generales.

1.1.2 Tipo de caracteres.

1.2 Escalas de medida.

1.3 Resúmenes numéricos.

1.4 Diagrama de tronco y hojas.

1.1 Definición de estadística.

Estadística: Ciencia que trata de la teoría y aplicación de métodos apropiados para coleccionar, representar, resumir y analizar datos para hacer inferencias a partir de ellos.

Clasificaciones de estadística

- **Estadística descriptiva o deductiva**. Se encarga de recoger y resumir las características de una población o muestra deduciendo de esta descripción conclusiones sobre su estructura, además de las relaciones existentes entre otros colectivos distintos con los cuales se compara.
- **Estadística inductiva o inferencial.** Basándose en los resultados del análisis de la muestra de la población induce o estima las leyes generales de comportamiento de la población.

1.1.1 Conceptos generales.

Población: Es un conjunto de elementos al que está referida la investigación y de la que se extraen los datos.
Individuo o unidad estadística: Cada uno de los elementos que componen la población. Es un ente observable que no tiene por qué ser una persona, puede ser algo abstracto.
Muestra: Subconjunto de elementos de la población, a partir del cual se realiza el estudio estadístico en caso de que no sea posible recopilar toda la información de la población.
Variable o serie estadística: Es una representación numérica de los caracteres, o una función que a cada modalidad asigna un valor.
Caracteres: Cada uno de las unidades estadísticas se describen mediante cualidades llamadas caracteres.

Existen dos tipos:

- **Caracteres cualitativos**: Aquellos caracteres que no se pueden medir (no cuantificables).
- **Caracteres cuantitativos**: Aquellos en los que se puede establecer una escala de medida, y se pueden subdividir a su vez en:
  - **Variables discretas**: Las que no pueden tomar valores entre dos consecutivos:
  - **Variables contínuas**: Las que admiten modalidades intermedias, es decir, puede haber infinitos valores entre dos.

A su vez las variables se pueden clasificar en agrupadas en intervalos y no agrupadas en intervalos.

Modalidad: Cada una de las diferentes situaciones posibles del carácter.

Ejemplo:

- Población: "Alumnos de la Universidad de Sevilla".
- Individuo: "Alumno".
- Muestra: "Subconjunto representativo de los alumnos de la Universidad de Sevilla".
- Caracteres:
  - Carrera: (Cualitativo)
    - Ingeniería.
    - Matemáticas.
    - Física.
    - Derecho....
  - Año de nacimiento (Cuantitativo discreto)
  - Nº de Hermanos (Cuantitativo discreto)
  - Lugar de nacimiento (Cualitativo)
  - Altura (Cualitativo contínuo)

1.2 Escalas de medida.

La regla es aceptar sólo como relaciones válidas entre los números aquellas que sean verificables empíricamente entre las correspondientes modalidades.

Escala: Conjunto de modalidades distintas y conjunto de números distintos relacionados biunívocamente. Las escalas se clasifican según Stevens en:

① **Escala NOMINAL:** Cuando solamente es posible establecer una relación de igualdad o desigualdad entre las modalidades.

② **Escala ORDINAL:** No sólo una relación de igualdad o desigualdad, sino también un orden.

③ **Escala de INTERVALOS:** Unidad de medida y sirve para comprobar cuantas veces está contenida esa unidad en la diferencia de 2 modalidades.

④ **Escala de RAZÓN:** Podemos decir cuántas veces una modalidad es mayor que otra.

1.3 Resúmenes numéricos.

Supongamos que tenemos n individuos u observaciones y $x_1,\ldots,x_k$ modalidades distintas. Normalmente se dará que $x_1 < x_2 < \ldots < x_k$.

- **Frecuencia absoluta** (de la modalidad x_i): Es el número de individuos que presenta dicha modalidad del carácter x.

$$\sum_{i=1}^{k} n_i = n$$

- **Frecuencia relativa** (de la modalidad x_i): Es la proporción de individuos que presenta dicha modalidad del carácter *x*. Se denota por f_i.

$$\sum_{i=1}^{k} f_i = 1 \qquad f_i = \frac{n_i}{n}$$

- **Porcentaje** (de la modalidad x_i): Viene dada por el producto $f_i \cdot 100$

$$p_i = f_i \times 100$$

Tipos de frecuencias para variables cuantitativas.

- **Frecuencia absoluta acumulada** (de la modalidad x_i): Nº de individuos de la población que presentan una modalidad $\leq x_i$. Se denota por N_i.

$$N_i = \sum_{j=1}^{i} n_j$$

- **Frecuencia relativa acumulada** (de la modalidad x_i): Proporción de individuos de la población que presentan una modalidad $\leq x_i$. Se denota por F_i .

$$F_i = \frac{N_i}{n}$$

- **Porcentaje acumulado**: $P_i = F_i \times 100$

Para el caso de variables cuantitativas contínuas o en intervalos definimos...

- **Marca de clase**: Valor que representa al intervalo y viene dado por el punto medio de ese intervalo.

$$x_i = \frac{e_{i-1} + e_i}{2} \qquad I_i = (e_{i-1}, e_i]$$

Siendo I_i el intervalo abierto por la izquierda y cerrado por la derecha.

- **Amplitud**: La diferencia entre el extremo derecho y el izquierdo.

$$a_i = e_i - e_{i-1}$$

- **Tabla de frecuencias:**

Ejemplo: *Tenemos una empresa con 10 empleados y vamos a hacer un estudio de los trabajadores de una empresa..*

Salario	n_i	f_i	P_i	N_i	F_i	P_i	x_i	a_i
(140,150]	1	0,1	10%	1	0,1	10%	145	10
(150,160]	1	0,1	10%	2	0,2	20%	155	10
(160,170]	2	0,2	20%	4	0,4	40%	165	10
(170,180]	3	0,3	30%	7	0,7	70%	175	10
(180,190]	2	0,2	20%	9	0,9	90%	185	10
(190,200]	1	0,1	10%	10	1	100%	195	10
	10	1	100%					

Decimos que la amplitud es constante cuando todos los intervalos son iguales.

Cuando construimos una tabla de frecuencias agrupada en intervalos se está perdiendo información. Si está sin agrupar en intervalos no se pierde información.

Si empleamos un $k \geq 10$, entonces podemos agrupar en intervalos.

Todos los intervalos van a tener la misma amplitud: $\boldsymbol{a} \approx 10^{p}$, o múltiplos de 5 o de 2.

1.- L = "nº máx. de intervalos permitidos" = $[10 \cdot \log_{10} n]$

Ej: X= "alturas en cm." $n = 10$

150, 160, 162, 182, 185, 192, 192, 194, 194, 197

$L = [10 \cdot \log_{10} 10] = 10$ será el número máximo de intervalos permitidos.

2.- $\frac{R}{L} = \frac{\max x_i - \min x_i}{L} = \frac{197 - 150}{10} = \frac{47}{10} = 4'7$

3.- Determinar m: $\frac{R}{L} < 10^{m}$, con m mínimo. En este caso 4'7<10

4.a.- $\frac{R}{L} < 2 \times 10^{m-1}$ { Sí: Comprobar que no se supera el nº máx. de intervalos al considerar $a = 2 \times 10^{m-1}$ { No $\Rightarrow$ FIN; Sí supera $\Rightarrow$ b) } ; No $\Rightarrow$ b) }

4.b.- $\frac{R}{L} < 5 \times 10^{m-1}$ { Sí: Comprobar que no se supera el nº máx. de intervalos al considerar $a = 5 \times 10^{m-1}$ { No $\Rightarrow$ FIN; Sí supera $\Rightarrow$ c) } ; No $\Rightarrow$ c) }

4.c.- $a = 10^{m}$ **FIN**

Ej:

Intervalo	X_i	n_i	f_i	N_i	F_i
(140, 150]	145	1	0.1	1	0.1
(150, 160]	155	1	0.1	2	0.2
(160, 170]	165	1	0.1	3	0.3
(170, 180]	175	0	0	3	0.3
(180, 190]	185	2	0.2	5	0.5
(190, 200]	195	5	0.5	10	1
		10			

1.4 Diagrama de tronco y hojas (STEAM-AND-LEAF)

Se debe al estadístico TUKEY. Nos permiten ver la simetría, dispersión, así como datos extraños (outleer). Para el caso de variables discretas procederemos así:

1. Calcularemos el número máximo de ramas (L):
$$L = [10 \times \log_{10} n]$$
2. Identificar los dígitos más significativos. Será la potencia de 10 más cercana por exceso al cociente R/L, siendo R = Máx - Min

Ejemplo:

150, 160, 162, 182, 185, 192, 193, 194, 194,197 (altura en cms.)

(1) $L = [10 \cdot \log_{10} 10] = 10$

(2) R =197-150 =47; $\frac{R}{L} = \frac{47}{10} = 4{,}7 < 10 \Rightarrow$ decenas; Los dígitos más significativos serán decenas.

(Ramas)	(Hojas)
15	0
16	0 2
17	
18	2 5
19	2 3 4 4 7

En el caso de variables contínuas podemos proceder del siguiente modo:

(1) Nº Máximo de Intervalos: $L = [10 \cdot \log_{10} n]$

(2) Amplitud $R/_L = 10^n$

$$R = \max - \min;\ \left.\begin{array}{l} R/_L < 5 \times 10^{n-1} \\ R/_L < 2 \times 10^{n-1} \end{array}\right\}$$ Siempre que no sobrapasen L

Ejemplo:

150, 160, 162, 182, 185, 192, 193, 194, 197 $n = 10$

· $L = [10 \times \log_{10} n] = \lfloor 10 \rfloor = 10$

· $\frac{R}{L} = \frac{197-150}{10} = \frac{47}{10} = 4{,}7$

· Determinar m: 4'7 < $10^m \rightarrow m = 1$

· Hojas: $10^{m-1} = 10^0 = 1$

$$L = 10\ ;\ R = 47\ \left.\begin{array}{l} R/_L = 4{,}7 < 10 \\ R/_L = 4{,}5 < \underbrace{5 \times 10^0}_{5} \end{array}\right\}$$ El de abajo no es menor, así que no nos vale.

* ≡ 0, 1, 2, 3, 4

0 ≡ 5, 6, 7, 8, 9

TRONCO	HOJA
5*	0
50	
6*	02
60	
7*	
70	
8*	2
80	5
9*	2344
90	7
n=10, udad=1, 5×10≡150	