

---

# ESTADÍSTICA DESCRIPTIVA

---

# FRECUENCIAS

Una **variable estadística** es una característica medible en los individuos o elementos de una *población*. Por ejemplo, la altura de los habitantes de Bilbao mayores de 21 años, el diámetro de los ejes producidos por una máquina, etc. Las variables estadísticas pueden ser:

a) **Cualitativas.** Son las clasificables en categorías, como el sexo, color de pelo, etc.

b) **Ordinales.** Los valores que toman este tipo de variables pueden ser ordenados de menor a mayor. Por ejemplo, el grado de satisfacción de un producto de consumo.

c) **Cuantitativas** que a su vez se clasifican en:

- **Discretas.** Son las variables *numerables* como el número de hijos, el número de piezas de un lote, etc.

- **Continuas.** Son las variables *medibles* mediante un número real, como el peso de un individuo.

Los datos de un estudio estadístico suelen presentarse generalmente de un modo desorganizado. Es conveniente, por tanto, agruparlos en una tabla que muestre las **frecuencias absolutas**  $F_i$ , o número de veces que la variable toma el valor  $x_i$ , y las frecuencias relativas  $f_i$ , que es el cociente entre las frecuencias absolutas y el número total de observaciones:

$$f_i = F_i / N.$$

También se suelen considerar las **frecuencias acumuladas**  $Fac_i$  que son el número de veces que aparece un valor igual o inferior a uno dado y las **frecuencias relativas acumuladas**  $fac_i$ .

## ■ Ejemplo 2.1

Las calificaciones finales entre 0 y 5 puntos, de un grupo de alumnos han sido:

5,5,5,0,0,5,3,0,1,5,5,3,3,4,5,5,0,3,0,4,4,2,1,2,5.

Representar estos datos mediante una tabla de frecuencias.

$x_i$	$F_i$	$Fac_i$	$f_i$	$f_{aci}$
$x_1=0$	5	5	5/25	5/25
$x_2=1$	2	7	2/25	7/25
$x_3=2$	2	9	2/25	9/25
$x_4=3$	4	13	4/25	13/25
$x_5=4$	3	16	3/25	16/25
$x_6=5$	9	25	9/25	1
Totales	25		1	

■

Las variables continuas suelen agruparse en intervalos. Algunos autores recomiendan que el número de estos no debe exceder de  $\sqrt{N}$ . Otros proponen construir un número de intervalos igual al entero más cercano a  $1 + 3,332 \log_{10} N$  (regla de Sturges). Cuando se agrupan los valores en intervalos, a efectos de cálculo de índices de tendencia central y de dispersión, se toma como valor la **marca de clase** o punto medio del intervalo.

## ■ Ejemplo 2.2

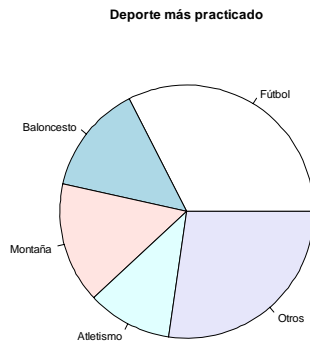
La tabla siguiente muestra la distribución de frecuencias de los salarios mensuales, medidos en euros, de 65 empleados de la empresa S&R.

Salarios	Marca de clase $x_i$	$F_i$	$F_{aci}$	$f_i$	$f_{aci}$
[2500,2600)	2550	8	8	0.123	0.123
[2600,2700)	2650	10	18	0.154	0.277
[2700,2800)	2750	16	34	0.246	0.523
[2800,2900)	2850	14	48	0.215	0.738
[2900,3000)	2950	10	58	0.154	0.892
[3000,3100)	3050	5	63	0.077	0.969
[3100,3200)	3150	2	65	0.031	1.000
		65		1.000	

■

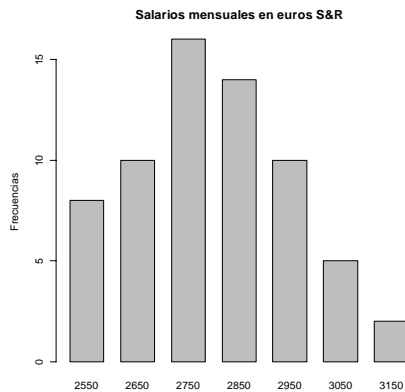
# **REPRESENTACIONES GRÁFICAS**

- a) **Diagrama de sectores** (figura 2.1): Utilizado principalmente en variables cualitativas.



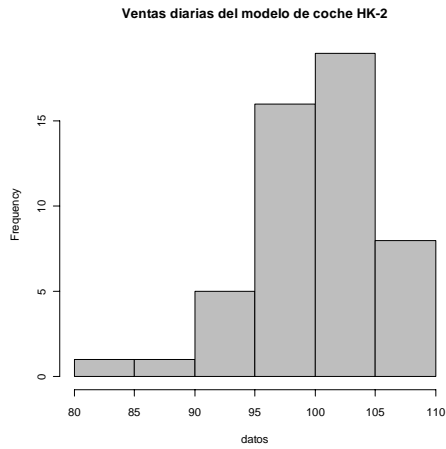
**Figura 2.1**

- b) **Diagrama de barras** (figura 2.2): Este gráfico es muy sencillo pero muy intuitivo. Muy apropiado para variables cuantitativas discretas.



**Figura 2.2**

c) **Histograma** (figura 2.3): Similar al gráfico anterior pero para variables cuantitativas continuas.



**Figura 2.3**

- d) **Diagrama de tallos y hojas** (figura 2.4):  
Es una alternativa al histograma que permite hacer una representación gráfica global de la distribución de frecuencias manteniendo la individualidad de los datos. Para construir este gráfico se hace: 1) Redondear los datos a 2 ó 3 cifras significativas. 2) Cada observación se divide en el **tallo**, formado por todos los dígitos excepto el último de la derecha, y la **hoja** que es el dígito final. 3) Se escriben los tallos en vertical empezando por el menor. 4) Se escribe cada hoja en una fila a la derecha de su tallo.

Ventas diarias del modelo de coche HK-2

The decimal point is at the |

```
84 | 5
86 |
88 | 9
90 | 9
92 | 266
94 | 5227
96 | 2889
98 | 355677089
100 | 35568067788
102 | 11478
104 | 19013
106 | 0581
108 | 70
```

**Figura 2.4**

# ÍNDICES DE TENDENCIA CENTRAL

- **Moda:** es el valor cuya frecuencia es mayor. Puede haber varias modas (distribución multimodal). Es la medida adecuada para describir variables cualitativas.

- **Media:** es la media aritmética de los valores que toma la variable:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{N}$$

Cuando los datos se agrupan por frecuencias resulta:

$$\bar{x} = \frac{\sum_{i=1}^n x_i F_i}{N} = \sum_{i=1}^n x_i f_i$$

- **Mediana:** es el valor  $Me$  que divide a la distribución en dos partes iguales, estando por debajo el 50% de las observaciones y por encima el otro 50%. Para calcular la mediana se ordenan los valores de menor a mayor; si hay un número impar de valores se toma el del centro, en caso contrario se toma como mediana el promedio de los dos valores centrales. Cuando los datos están agrupados por intervalos se procede del siguiente modo: se busca en la tabla el índice  $j$  tal que  $Fac_{j-1} < \frac{N}{2} \leq Fac_j$  y se identifica el intervalo  $[a_j, b_j)$ , siendo la mediana el valor

$$Me = a_j + \frac{\frac{N}{2} - Fac_{j-1}}{Fac_j - Fac_{j-1}}(b_j - a_j)$$

La media es el índice más utilizado por sus propiedades matemáticas, aunque a veces no resulta muy descriptiva, como por ejemplo cuando la distribución es muy asimétrica. En la serie de datos 2,5,8,11,99,  $\bar{x} = 25$ ,  $Me=8$ , siendo este último valor más representativo.

### ■ Ejemplo 2.3

A) Calcular la mediana para los datos del ejemplo 2.1.

En este caso la mediana será el valor que ocupa el lugar 13:  $Me=3$ .

B) Calcular la mediana para los datos del ejemplo 2.2.

Como  $N/2=65/2=32,5$  la mediana estará en el intervalo  $[2700,2800)$  y su valor será:

$$Me = 2700 + \frac{\frac{65}{2} - 18}{34 - 18} (2800 - 2700) = 2790,63$$

■

# ÍNDICES DE DISPERSIÓN

- **Recorrido:** Es la diferencia entre la observación máxima y la observación mínima.

- **Varianza:** Es la media de las desviaciones respecto de la media, elevada al cuadrado (no negativa, por tanto):

$$\begin{aligned}
 s^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2)}{n} = \frac{\sum_{i=1}^n x_i^2}{n} - 2\bar{x} \frac{\sum_{i=1}^n x_i}{n} + \bar{x}^2 \frac{\sum_{i=1}^n 1}{n} = \\
 &= \frac{\sum_{i=1}^n x_i^2}{n} - 2\bar{x} \cdot \bar{x} + \bar{x}^2 = \frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2
 \end{aligned}$$

Esta última expresión puede leerse del siguiente modo: *La varianza es igual a la media de los cuadrados menos el cuadrado de la media.*

Cuando los datos se agrupan por frecuencias resulta:

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2 F_i}{n} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 f_i}{n} = \sum_{i=1}^n (x_i^2 f_i - 2x_i \bar{x} f_i + \bar{x}^2 f_i) = \\ &= \sum_{i=1}^n x_i^2 f_i - 2\bar{x} \sum_{i=1}^n x_i f_i + \bar{x}^2 \sum_{i=1}^n f_i = \sum_{i=1}^n x_i^2 f_i - \bar{x}^2 \end{aligned}$$

- **Desviación típica:** Es la raíz cuadrada de la varianza

$$s = \sqrt{s^2}$$

La ventaja de este índice respecto de la varianza es que la desviación típica viene expresada en las mismas unidades que la variable.

En ocasiones resulta interesante tipificar un dato:

$$z - score \text{ de } x_i = \frac{x_i - \bar{x}}{s},$$

sobre todo si se trata de comparar valores correspondientes a diferentes conjuntos de observaciones.

- **Coeficiente de variación:** es la relación entre la desviación típica y la media:

$$CV = \frac{s}{|\bar{x}|}$$

Este índice suele utilizarse para comparar la dispersión de variables cuando se miden en unidades diferentes.

## ■ Ejemplo 2.4

A) Calcular el coeficiente de variación para los datos del ejemplo 2.1. Obtener la puntuación tipificada correspondiente a una nota igual a 4. Para ello añadimos dos nuevas columnas:

$x_i$	$F_i$	$Fac_i$	$f_i$	$f_{aci}$	$x_i f_i$	$x_i^2 f_i$
$x_1=0$	5	5	5/25	5/25	0	0
$x_2=1$	2	7	2/25	7/25	0,08	0,08
$x_3=2$	2	9	2/25	9/25	0,16	0,32
$x_4=3$	4	13	4/25	13/25	0,48	1,44
$x_5=4$	3	16	3/25	16/25	0,48	1,92
$x_6=5$	9	25	9/25	1	1,8	9
Totales	25		1		3	12,76

$$\bar{x} = \sum x_i f = 3;$$

$$s^2 = \sum x_i^2 f_i - \bar{x}^2 = 12,76 - 3^2 = 3,76;$$

$$s = \sqrt{3,76} = 1,94;$$

$$CV = \frac{1,94}{3} = 0,6467 = 64,67\%;$$

$$z - score(4) = \frac{4 - 3}{1,94} = 0,52$$

B) Calcular el coeficiente de variación para los datos del ejemplo 2.2. Obtener la puntuación tipificada correspondiente a un sueldo de 3100€.

Salarios	Marca de clase	$F_i$	$F_{aci}$	$f_i$	$f_{aci}$	$x_i f_i$	$x_i^2 f_i$
[2500-2600)	2550	8	8	0.123	0.123	313.65	799807.5
[2600-2700)	2650	10	18	0.154	0.277	408.1	1081465
[2700-2800)	2750	16	34	0.246	0.523	676.5	1860375
[2800-2900)	2850	14	48	0.215	0.738	612.75	1746338
[2900-3000)	2950	10	58	0.154	0.892	454.3	1340185
[3000-3100)	3050	5	63	0.077	0.969	234.85	716292.5
[3100-3200)	3150	2	65	0.031	1.000	97.65	307597.5
		65		1.000		2797.8	7852061

$$\bar{x} = \sum x_i f = 2797,8:$$

$$s^2 = \sum x_i^2 f_i - \bar{x}^2 = 7852061 - 2797,8^2 = 24376,16;$$

$$s = \sqrt{24376,16} = 156,13;$$

$$CV = \frac{156,13}{2797,8} = 0,0558 = 5,58\%;$$

$$z\text{-score}(3100) = \frac{3100 - 2797,8}{156,13} = 1,94$$

■

# **REGLA EMPÍRICA**

La información conjunta que proporcionan la media y la desviación típica puede precisarse, en una gran variedad de situaciones, de acuerdo a la **regla empírica** siguiente:

- 1) Aproximadamente el 68% de las observaciones están en el intervalo  $(\bar{x} - s, \bar{x} + s)$ . En valores tipificados: (-1,1).
  
- 2) Aproximadamente el 95% de las observaciones están en el intervalo  $(\bar{x} - 2s, \bar{x} + 2s)$ . En valores tipificados: (-2,2).
  
- 3) Aproximadamente el 99,7% de las observaciones están en el intervalo  $(\bar{x} - 3s, \bar{x} + 3s)$ . En valores tipificados: (-3,3).

# ÍNDICES DE POSICIÓN

Los índices siguientes corresponden a valores que ocupan determinadas posiciones en el conjunto de observaciones:

- **Percentiles:** El percentil  $p$ -ésimo  $P_p$  es el valor que, una vez ordenadas las observaciones de menor a mayor, deja por debajo el  $p\%$  de las observaciones y por encima el  $(100-p)\%$  restante. Para valores no agrupados por intervalos el cálculo es análogo al visto para la mediana. Cuando los datos están agrupados por intervalos se procede así: se busca en la tabla el índice  $j$  tal que  $Fac_{j-1} < \frac{N}{100}p \leq Fac_j$  y se identifica el intervalo  $[e_{j-1}, e_j)$ , siendo  $P_p$  el valor

$$P_p = e_{j-1} + \frac{\frac{N}{100}p - Fac_{j-1}}{Fac_j - Fac_{j-1}}(e_j - e_{j-1})$$

- **Cuartiles:** El cuartil inferior  $Q_1$  es el percentil 25,  $P_{25}$ ; el cuartil medio  $Q_2$  es el percentil 50,  $P_{50}$ , o sea la mediana, y el cuartil superior  $Q_3$  es el percentil 75,  $P_{75}$ .

- **Recorrido intercuartílico:** Es el valor  $RIQ=Q_3-Q_1$ .

## ■ Ejemplo 2.5

A) Calcular RIQ para los datos del ejemplo 2.1.

Como  $\frac{N}{100}p = \frac{25}{100}25 = 6,25 \Rightarrow Q_1$  será el valor promedio entre el 6° y 7°

$$\text{o sea, } Q_1 = \frac{1+1}{2} = 1$$

Como  $\frac{N}{100}p = \frac{25}{100}75 = 18,75 \Rightarrow Q_3$  será el valor promedio entre el 18° y 19°

$$\text{o sea, } Q_3 = \frac{5+5}{2} = 5; \quad RIQ = Q_3 - Q_1 = 5 - 1 = 4$$

B) Calcular RIQ para los datos del ejemplo 2.2.

Como  $\frac{N}{100}p = \frac{65}{100}25 = 16,25 \Rightarrow Q_1$  será el valor:

$$2600 + \frac{16,25 - 8}{18 - 8}(2700 - 2600) = 2682,5$$

Como  $\frac{N}{100}p = \frac{65}{100}75 = 48,75 \Rightarrow Q_3$  será el valor:

$$2900 + \frac{48,75 - 34}{48 - 34}(2900 - 2800) = 3005,36$$

$$RIQ = Q_3 - Q_1 = 3005,36 - 2682,5 = 322,86$$

■

# DETECCIÓN DE DATOS ATÍPICOS

Los datos atípicos o *outliers* suelen aparecer normalmente cuando se recogen datos. Es importante el estudio de estos datos con objeto de corregirlos, eliminarlos o simplemente tenerlos en cuenta. Para su detección calcularemos los intervalos siguientes:

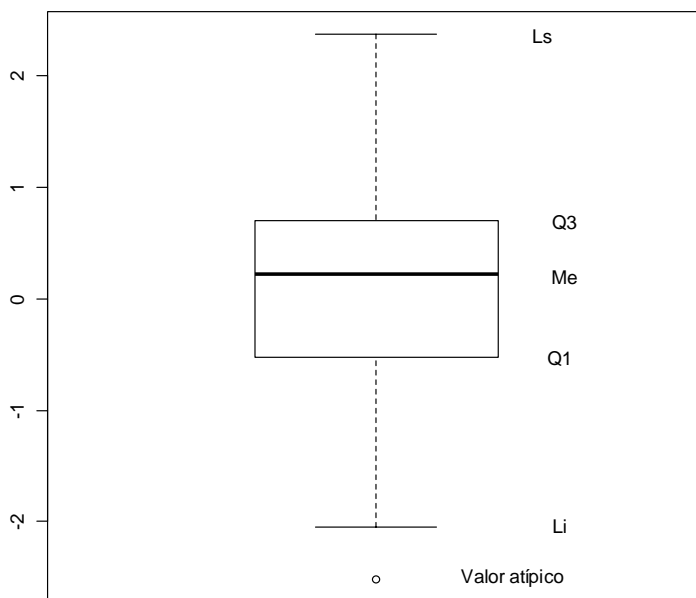
- Se consideran datos sospechosos de atípicos:

$$x_i \notin \left( Q_1 - \frac{3}{2}RIQ, Q_3 + \frac{3}{2}RIQ \right)$$

- Se consideran datos sospechosos de atípicos:

$$x_i \notin (Q_1 - 3RIQ, Q_3 + 3RIQ)$$

La regla anterior se puede expresar de forma gráfica en el **diagrama box-plot** (véase figura 2.5).  $L_i$  es el valor más pequeño no atípico (es decir, que entra dentro del primer intervalo) y  $L_s$  es el valor más grande no atípico (es decir, que entra dentro del primer intervalo).



**Figura 2.5**

# PROBLEMAS

## Ejercicios resueltos

■1) Los datos siguientes son los pesos en gramos de 72 muestras de una determinada sustancia:

64, 51, 55, 42, 53, 46, 60, 29, 56, 20, 52, 51, 33, 61, 57, 55, 59, 38, 56, 41, 47, 68, 24, 67, 52, 64, 69, 43, 47, 42, 65, 96, 21, 48, 47, 25, 82, 37, 60, 12, 77, 56, 97, 28, 45, 63, 28, 45, 63, 28, 52, 60, 51, 61, 62, 52, 97, 73, 45, 69, 67, 29, 75, 63, 30, 17, 69, 68, 74, 16, 83, 47.

Agrupar los datos en una tabla de frecuencias y calcular la media, la mediana y la varianza.

Al ser  $N = 72$ , un número conveniente de intervalos puede ser 8.

Intervalos	Marcas de clase $x_i$	$F_i$	$F_{aci}$	$f_i$	$f_{aci}$	$x_i^2$	$x_i f_i$	$x_i^2 f_i$
11-21	16	6	6	0.08	0.08	256	1.28	20.48
22-32	27	8	14	0.11	0.19	729	2.97	80.19
33-43	38	7	21	0.10	0.29	1444	3.80	144.40
44-54	49	17	38	0.24	0.53	2401	11.76	576.24
55-65	60	18	56	0.25	0.78	3600	15.00	900.00
66-76	71	10	66	0.14	0.92	5041	9.94	705.74
77-87	82	3	69	0.04	0.96	6724	3.28	268.96
88-98	93	3	72	0.04	1	8649	3.72	345.96
Totales		72		1			51.75	3041.97

$$\bar{x} = 51.71$$

El intervalo de clase donde se encontrará la mediana es el que contiene la frecuencia acumulada  $N/2 = 36$ ; es decir, el intervalo 44 -54. Para calcular la mediana se hace:

$$Me = 43.5 + \frac{(72/2) - 21}{17} = 44.38$$

$$s^2 = 3041.97 - 51.71^2 = 368.05$$

■2) En la tabla siguiente se muestra la facturación en miles de euros de 110 empresas:

Facturación	[0,10)	[10,20)	[20,30)	[30,40)	[40,50)
Núm. de empresas	15	25	32	23	15

a) Calcular la media y la mediana. b) Obtener la desviación típica y el coeficiente de variación. c) ¿Cuál es el porcentaje estimado de empresas que han facturado menos de 32.000 euros?

a)

	$X_i$	$X_i^2$	$R_i$	$f_{ac}$	$f_i$	$K_i f_i$	$X_i^2 f_i$
[0,10)	5	25	15	15	0,137	0,615	3,425
[10,20)	15	225	25	40	0,227	3,405	51,075
[20,30)	25	625	32	72	0,291	7,275	181,875
[30,40)	35	1225	23	95	0,209	7,315	256,025
[40,50)	45	2025	15	110	0,136	6,112	225,4
			110			24,8	767,8

$$\bar{X} = \sum K_i \cdot f_i = 24,8$$

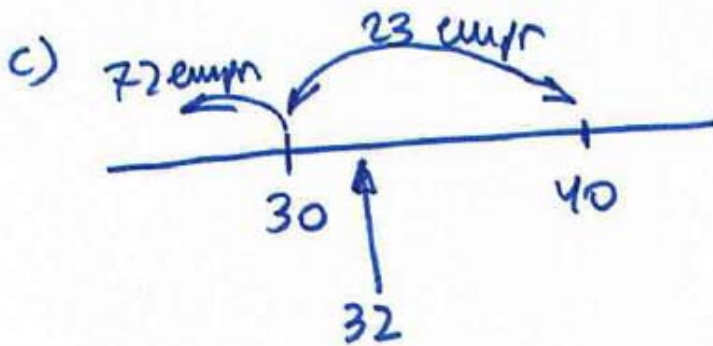
Como  $\frac{110}{2} = 55$  Me está en [20,30)

$$Me = 20 + \frac{55 - 40}{72 - 40} (30 - 20) = 24,69$$

$$b) s^2 = \sum X_i^2 f_i - \bar{X}^2 = 767,8 - 24,8^2 = 152,76$$

$$s = \sqrt{152,76} = 12,36$$

$$CV = \frac{s}{|\bar{X}|} = \frac{12,36}{24,8} = 0,4984 = 49,84\%$$



$$40 - 30 = 10 \text{ miles } \text{€} \quad \text{---} \quad 23 \text{ km/hr}$$

$$32 - 30 = 2 \text{ miles } \text{€} \quad \text{---} \quad x$$

$$x = \frac{2 \cdot 23}{10} = 4,6$$

$$72 + 4,6 = 76,6 \text{ km/hr.}$$

$$\frac{76,6}{110} = 0,6964 = 69,64\%$$

## Ejercicios propuestos

■1) Las puntuaciones obtenidas por un grupo de sujetos en una cierta prueba han sido: 21, 36, 19, 23, 32, 25, 28, 20, 34, 33, 31. Determinar la media, la mediana y la varianza.

■2) Lanzar 50 veces un dado al aire y formar la tabla de frecuencias correspondiente. Representar los datos gráficamente. Calcular la media y la varianza.

■3) Se lanzan cinco monedas 1000 veces. El número de lanzamientos en los que han salido 0, 1, 2, 3, 4 y 5 caras se indican en la tabla 1.3. Se pide: a) Representar gráficamente los datos. b) Construir una tabla que muestre los porcentajes de tiradas que han dado un número de caras menor que 0, 1, 2, 3, 4, 5 ó 6, c) Representar los datos de la tabla del apartado anterior.

Número de caras	Frecuencia
0	38
1	144
2	342
3	287
4	164
5	25

■4) La clasificación de los equipos de primera división tras la disputa de 11 jornadas de la liga de fútbol española 1997-98 fue:

Puesto	Equipo	Puntos	Goles
1	Barcelona	25	24
2	R. Madrid	24	19
3	Celta	24	20
4	Espanyol	22	20
5	At. Madrid	21	27
6	R. Sociedad	21	16
7	Mallorca	19	20

8	Oviedo	17	14
9	Athletic	17	14
10	Betis	14	14
11	Mérida	14	9
12	Deportivo	12	12
13	Zaragoza	11	16
14	Racing	11	12
15	Tenerife	11	11
16	Compostela	10	18
17	Valladolid	9	8
18	Valencia	8	9
19	Salamanca	6	5
20	Sporting	1	7

Clasificar los equipos de acuerdo al criterio rentabilidad de los goles, medio en puntos/gol. Calcular la mediana y la moda. ¿Cuál de los dos índices es más significativo? Hallar también la desviación y el coeficiente de variación.

■5) Se ha medido la longitud de 20 caimanes machos capturados en un cierto lago obteniéndose los

valores: 118, 132, 132, 140, 142, 142, 147, 147, 150, 152, 163, 165, 165, 165, 165, 168, 170, 173, 178, 190.  
 1º) Calcular el coeficiente de variación de la variable longitud. 2º) Estudiar la existencia de datos atípicos.

■6) Con objeto de estudiar la distribución de las edades en una cierta población se ha extraído una muestra de 200 individuos, obteniéndose los siguientes datos:

Edad	Mujeres	Hombres
[0-20)	13	15
[20-40)	22	27
[40-60)	38	24
[60-80)	28	12
[80-100)	14	7

a) ¿Quiénes son más jóvenes en promedio? Razonar la respuesta. b) Estimar la proporción de hombres mayores de 75 años. c) ¿La variabilidad de la edad de las mujeres es mayor que la de los hombres? Razonar la respuesta

■7) Se ha recogido una muestra de datos de una cierta variable que toma valores en el intervalo  $[30,90]$  y se ha obtenido el siguiente diagrama de tallos y hojas:

3 | 00259

4 | 2256

5 | 556

6 | 03

7 | 0128

8 | 34

Se pide: 1º) Tamaño de la muestra. 2º) Media y varianza muestrales. 3º) Cuartiles y recorrido intercuartílico. 4º) Detectar datos atípicos, si los hay, razonando la respuesta.

■8) En un cierto estudio se han recogido los siguientes datos: 15, 16, 21, 23, 23, 26, 26, 30, 32, 41, 42, 51, 53, 53, 53, 53, 55, 60, 60, 69. Se pide: 1º) Diagrama de tallos y hojas. 2º) Coeficiente de variación. 3º) Recorrido intercuartílico.