# Capítulo 1

# Estadística descriptiva univariante

**1.** *Una encuesta entre fumadores sobre el número de cigarrillos que consumen al día ha dado lugar a los resultados siguientes:*

| *Número de cigarrillos* | *4,5-9,5* | *9,5-14,5* | *14,5-19,5* | *19,5-24,5* | *24,5-29,5* |
|---|---|---|---|---|---|
| *Número de fumadores* | *10* | *15* | *25* | *18* | *22* |

*Determina* **a)** *la media, la mediana y la moda de la distribución e interpreta los resultados,* **b)** *la desviación típica y valora la representatividad de la media,* **c)** *el rango intercuartílico e interpreta el resultado,* **d)** *el porcentaje de individuos que fuman entre 12 y 22 cigarrillos diarios,* **e)** *el consumo medio diario de cigarrillos para una población de 1000 individuos, sabiendo que el porcentaje de fumadores es del 30%.*

Primeramente, construiremos una tabla estadística en la que apoyarnos para dar respuesta a las distintas cuestiones:

| $I_i$ | $n_i$ | $x_i$ | $x_i n_i$ | $N_i$ | $x_i^2 n_i$ |
|---|---|---|---|---|---|
| 4,5 - 9,5 | 10 | 7 | 70 | 10 | 490 |
| 9,5 - 14,5 | 15 | 12 | 180 | 25 | 2160 |
| 14,5 - 19,5 | 25 | 17 | 425 | 50 | 7225 |
| 19,5 - 24,5 | 18 | 22 | 396 | 68 | 8712 |
| 24,5 - 29,5 | 22 | 27 | 594 | 90 | 16038 |
| | 90 | | 1665 | | 34625 |

**a)** La media es de $\bar{x} = 1665/90 = 18{,}5$ cigarrillos diarios por individuo. Por otra parte, teniendo en cuenta que el primer intervalo cuya frecuencia absoluta acumulada supera o iguala la mitad del tamaño de la población es el tercero, la mediana estará en el intervalo de extremos 14,5 y 19,5, así que será

$$Me = 14{,}5 + \frac{45 - 25}{25} \times 5 = 18{,}5 \text{ cigarrillos}$$

lo que quiere decir que los que más fuman consumen, como mínimo, 19 cigarrillos al día o, equivalentemente, el 50 % que menos fuma, consume, a lo sumo, 18 cigarrillos al día. Finalmente, la moda, que es el valor más frecuente de la variable se encuentra en el intervalo modal, que es el de extremos 14,5 y 19,5, ya que es el que tiene mayor frecuencia absoluta (obsérvese que todos los intervalos tienen la misma amplitud). Así, la moda se puede hallar como

$$Mo = 14{,}5 + \frac{25-15}{(25-15)+(25-18)} \times 5 = 17{,}4 \text{ cigarrillos}$$

esto es, lo más habitual entre estos fumadores es consumir 17,4 cigarrillos al día.

**b)** La varianza es

$$s^2 = \frac{34625}{90} - (18{,}5)^2 = 42{,}47 \text{ cigarrillos}^2 \quad \Rightarrow \quad s = 6{,}52 \text{ cigarrillos}$$

Así, el coeficiente de variación de Pearson es

$$CV = \frac{6{,}52}{18{,}5} = 0{,}3524 \quad \longrightarrow \quad 35{,}24\,\%$$

lo que significa que la media es poco representativa.

**c)** Para el cálculo del rango intercuartílico se requieren los cuartiles primero y tercero de la distribución:

Puesto que $0{,}25 \times 90 = 22{,}5$, entonces $Q_1$ estará en el intervalo de extremos 9,5 y 14,5, pues es el primero cuya frecuencia absoluta acumulada supera 22,5, así que

$$Q_1 = 9{,}5 + \frac{22{,}5-10}{15} \times 5 = 13{,}67$$

Además, como $0{,}75 \times 90 = 67{,}5$, entonces $Q_3$ estará en el intervalo de extremos 19,5 y 24,5, pues es el primero cuya frecuencia absoluta acumulada supera 67,5, por lo que

$$Q_3 = 19{,}5 + \frac{67{,}5-50}{18} \times 5 = 24{,}36$$

Así, el rango intercuartílico es $Q_3 - Q_1 = 10{,}69$, lo que significa que el 50 % correspondiente a los valores centrales distan entre sí, como mucho, 10,69 unidades (cigarrillos).

**d)** Si 12 es el percentil $p$, $P_p = 12$, y 22 es el percentil $q$, $P_q = 22$, entonces el $p\,\%$ fuman, a lo sumo, 12 cigarrillos y el $q\,\%$ fuman, a lo sumo, 22 cigarrillos, lo que significa que $(q-p)\,\%$ es el porcentaje de los que fuman entre 12 y 22 cigarrillos. Por tanto,

$$P_p = 12 = 9{,}5 + \frac{p\frac{90}{100} - 10}{15} \times 5 \quad \Rightarrow \quad p = 19{,}44$$

$$P_q = 22 = 19{,}5 + \frac{q \times 0{,}90 - 50}{18} \times 5 \quad \Rightarrow \quad q = 65{,}56$$

y así, $q - p = 46{,}12$, esto es, el 46,12 % fuman entre 12 y 22 cigarrillos.

**e)** El 30 % de dicha población se corresponde con 300 fumadores, y puesto que cada uno fuma una media de 18,5 cigarrillos, el consumo medio diario es de $300 \times 18{,}5 = 5550$ cigarrillos.

**2.** *Se ha efectuado un recopilación de datos sobre el número de guantes desechables usados al cabo de un determinado día por los 50 sanitarios que pasan consulta durante la mañana en un centro de salud, $X$, y el correspondiente análisis estadístico arroja los siguientes resultados:*

$$\bar{x} = 10 \quad Q_1 = 5{,}5 \quad Q_3 = 12 \quad s^2 = 36 \quad Mo = 9$$

*Determina* **a)** *el número de guantes desechables usados al cabo de ese día,* **b)** *la representatividad del promedio,* **c)** *el número mínimo de guantes que usa el 75 % formado por los empleados que más guantes usan,* **d)** *lo más frecuente en cuanto a empleo de guantes desechables se refiere,* **e)** *el número medio de guantes empleados entre ese día y el siguiente, en el que 40 sanitarios pasaron consulta y el promedio de guantes empleados fue de 12.*

**a)** Dado que

$$\bar{x} = \frac{\sum_{i=1}^{50} x_i}{50}$$

se deduce que $\sum_{i=1}^{50} = 500$ es el número de guantes usados entre todos los sanitarios que pasaron consulta esa mañana.

**b)** Para valorar la representatividad de $\bar{x}$, recurrimos al coeficiente de variación de Pearson,

$$CV = \frac{6}{10} \times 100 = 60\,\%$$

lo que implica una baja representatividad de la media.

**c)** La respuesta a esta pregunta se formula en base al primer cuartil de la distribución, que es 5,5, pero, como la variable "número de guantes" es discreta y toma, como valores, números naturales, podemos decir que el mínimo de guantes empleados por el 75 % que más guantes gasta es 6.

**d)** La respuesta a esta pregunta la proporciona la moda de la distribución, que es 9 guantes.

**e)** El número medio de guantes usados entre los sanitarios que pasaron consulta ambos días será

$$\frac{10 \times 50 + 12 \times 40}{50 + 40} = 10{,}89 \text{ guantes}$$

**3.** *Una vacuna antitetánica se ha administrado a 42 personas. A las 5 horas se les ha tomado la temperatura, obteniéndose los datos siguientes:*

| $T^{\underline{a}}$ (°C) | [37, 37,5] | (37,5, 38] | (38, 38,5] | (38,5, 39] | (39, 39,5] | (39,5, 40] |
|---|---|---|---|---|---|---|
| $N^{\underline{o}}$ de personas | 1 | 5 | 15 | 6 | 10 | 5 |

**a)** *¿Cuántas personas han tenido una temperatura de, como máximo, 38 °C?, ¿qué porcentaje de individuos ha tenido una fiebre superior a 38 °C pero de, a lo sumo, 39 °C?* **b)** *¿Cuál ha sido la temperatura promedio al cabo de 5 horas?* **c)** *¿Cuál ha sido la temperatura más frecuente al cabo de 5 horas?* **d)** *¿Cuál ha sido la temperatura mínima de los individuos que han tenido más fiebre?*

Primeramente, construiremos una tabla estadística en la que apoyarnos para dar respuesta a las distintas cuestiones, denotando por $T$ a la variable estadística "temperatura" (en °C):

| $I_i$ | $t_i$ | $n_i$ | $f_i$ | $t_i n_i$ | $N_i$ |
|---|---|---|---|---|---|
| [37, 37,5] | 37,25 | 1 | 1/42 | 37,25 | 1 |
| (37,5, 38] | 37,75 | 5 | 5/42 | 188,75 | 6 |
| (38, 38,5] | 38,25 | 15 | 15/42 | 573,75 | 21 |
| (38,5, 39] | 38,75 | 6 | 6/42 | 232,50 | 27 |
| (39, 39,5] | 39,25 | 10 | 10/42 | 392,50 | 37 |
| (39,5, 40] | 39,75 | 5 | 5/42 | 198,75 | 42 |
| | | 42 | 1 | 1623,5 | |

**a)** El número de personas con, a lo sumo, 38°C de fiebre ha sido $n_1 + n_2 = N_2 = 6$, mientras que el porcentaje correspondiente a los que han tenido más de 38°C y, como mucho, 39°C, ha sido $(f_3 + f_4) \times 100 = 50\,\%$.

**b)** La temperatura media al cabo de 5 horas ha sido $\bar{t} = \frac{1623,5}{42} = 38,65$°C.

**c)** Se trata de la moda. Dado que todos los intervalos presentan la misma amplitud, el intervalo modal será el de mayor frecuencia (absoluta o relativa), es decir, $Mo \in (38, 38,5]$, por lo que

$$Mo = 38 + \frac{15 - 5}{(15 - 5) + (15 - 6)} \times 0,5 = 38,26°\text{C}$$

**d)** Se trata de la temperatura "mediana". Dado que el primer intervalo cuya frecuencia absoluta acumulada supera la mitad del tamaño de la población es $(38, 38,5]$, se tendrá que $Me \in (38, 38,5]$ y, concretamente, $Me$ es el extremo superior de dicho intervalo, ya que

$$Me = 38 + \frac{21 - 6}{15} \times 0,5 = 38,5°\text{C}$$

**4.** *Las alturas (cm) de 30 alumnos de una clase son las siguientes:*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *174* | *185* | *166* | *176* | *145* | *166* | *191* | *177* | *164* | *171* |
| *175* | *158* | *156* | *156* | *187* | *162* | *172* | *193* | *183* | *173* |
| *197* | *181* | *151* | *161* | *153* | *172* | *162* | *179* | *188* | *179* |

**a)** *Agrupa los datos en intervalos de amplitud 10.* **b)** *Construye la tabla de frecuencias.* **c)** *¿Cuál es la altura más habitual entre los alumnos de esa clase?* **d)** *¿Qué percentil le corresponde a un alumno cuya altura es de 181 cm?* **e)** *¿Qué altura mínima debe tener un alumno para que pueda considerarse entre el 20% de los alumnos más altos de la clase?*

**a)**, **b)** El dato máximo es 197 y el mínimo 145. Por tanto, el rango de la variable es 52 y si la amplitud ha de ser 10, el número de intervalos es 5,2, por lo que debemos considerar 6, ya que si consideramos 5, la amplitud que suman entre todos es de 50, que no sería suficiente para completar el rango de la variable. La tabla con las frecuencias absolutas y absolutas acumuladas queda

| $I_i$ | $n_i$ | $N_i$ |
|---|---|---|
| (140, 150] | 1 | 1 |
| (150, 160] | 5 | 6 |
| (160, 170] | 6 | 12 |
| (170, 180] | 10 | 22 |
| (180, 190] | 5 | 27 |
| (190, 200] | 3 | 30 |
| | 30 | |

**c)** Puesto que los intervalos tienen la misma amplitud, la moda está en el intervalo de mayor frecuencia, $Mo \in (170, 180]$, y su valor es

$$Mo = 170 + \frac{10-6}{(10-6)+(10-5)} \times 10 = 174{,}44 \text{ cm}$$

**d)** Si $P_x = 181$, entones $P_x \in (180, 190]$, con lo cual

$$P_x = 180 + \frac{\frac{x}{100}30 - 22}{5} \times 10 \quad \Rightarrow \quad x = 75$$

esto es, 181 es el percentil 75 o, lo que es igual, el tercer cuartil.

**e)** Se trata del percentil 80. Por tanto, $0{,}80 \times 30 = 24$, con lo que $P_{80} \in (180, 190]$ y

$$P_{80} = 180 + \frac{24-22}{5} \times 10 = 184 \text{ cm}$$

**5.** *El peso (en Kg) de los bebés al cumplir el primer mes de vida, observados en una consulta varían de acuerdo a la tabla siguiente:*

| *Peso (Kg)* | *(3, 3,7]* | *(3,7, 4]* | *(4, 4,2]* | *(4,2, 4,5]* | *(4,5, 4,7]* | *(4,7, 5]* | *(5, 5,4]* |
|---|---|---|---|---|---|---|---|
| *Número de bebés* | *2* | *5* | *18* | *91* | *15* | *6* | *3* |

**a)** *¿Cuántos bebés han pesado, como mucho, 4 Kg?, ¿qué porcentaje de bebés han pesado más de 3,5 y, a lo sumo, 4,5 Kg?* **b)** *¿Cuál ha sido el peso medio de los bebés atendidos?* **c)** *¿Cuál ha sido el peso más frecuente?* **d)** *¿Qué debe pesar un bebé para pertenecer al grupo de los más pesados?*

La tabla estadística en la que apoyarnos es

| $I_i$ | $x_i$ | $n_i$ | $x_i n_i$ | $a_i$ | $h_i$ | $N_i$ |
|---|---|---|---|---|---|---|
| (3, 3,7] | 3,35 | 2 | 6,70 | 0,7 | 2,86 | 2 |
| (3,7, 4] | 3,85 | 5 | 19,25 | 0,3 | 16,67 | 7 |
| (4, 4,2] | 4,10 | 18 | 73,80 | 0,2 | 90 | 25 |
| (4,2, 4,5] | 4,35 | 91 | 395,85 | 0,3 | 303,33 | 116 |
| (4,5, 4,7] | 4,60 | 15 | 69 | 0,2 | 75 | 131 |
| (4,7, 5] | 4,85 | 6 | 29,10 | 0,3 | 20 | 137 |
| (5, 5,4] | 5,20 | 3 | 15,60 | 0,4 | 7,5 | 140 |
| | | 140 | 609,3 | | | |

**a)** Hay 7 bebés que, como mucho, han pesado 4 Kg ya que $N_2 = 7$. El porcentaje de bebés que han pesado, como mucho, 4,5 Kg es $\frac{N_4}{N} \times 100 = 116/140 \times 100 = 82{,}86\,\%$, mientras que el porcentaje de bebés que han pesado más de 3,5 Kg lo determinaremos recurriendo a la fórmula para hallar los percentiles:

$$3{,}5 = P_x = 3 + \frac{1{,}4x - 0}{2} \times 0{,}7 \quad \Rightarrow \quad x = 1{,}02$$

Así, el porcentaje de bebés que han pesado más de 3,5 Kg pero menos de 4,5 Kg es $82{,}86 - 1{,}02 = 81{,}84\,\%$.

**b)** El peso medio es de $609{,}3/140 = 4{,}35$ Kg.

**c)** El peso más frecuente está en el intervalo (4,2 4,5], puesto que es el de mayor densidad de frecuencia ($h_i$), y su valor es

$$Mo = 4{,}2 + \frac{303{,}33 - 90}{303{,}33 - 90 + 303{,}33 - 75} \times 0{,}3 = 4{,}34 \text{ Kg}$$

**d)** El peso mínimo que corresponde al grupo de los más pesados es la mediana, que se sitúa también en dicho intervalo por ser el primero cuya frecuencia absoluta acumulada

supera la mitad del tamaño de la población, y su valor es

$$Me = 4{,}2 + \frac{70 - 25}{91} \times 0{,}3 = 4{,}35 \text{ Kg}$$

**6.** *(***Supervivencia de los conejillos de Indias***) Los datos siguientes se refieren a los días de supervivencia de 72 conejillos de Indias después de que se les inyectara el bacilo de la tuberculosis en un experimento médico:*

| *Días de supervivencia* | *(10, 28]* | *(28, 46]* | *(46, 64]* | *(64, 82]* | *(82, 100]* |
|---|---|---|---|---|---|
| *Número de conejillos* | *16* | *17* | *11* | *13* | *15* |

**a)** *¿Cuál es el número medio de días de supervivencia?* **b)** *¿Cuál es el número más frecuente de días de supervivencia?* **c)** *En el grupo de los conejillos que han muerto antes, ¿cuál es el número máximo de días que han sobrevivido?* **d)** *¿Cuál es la desviación típica del número de días de superviviencia?* **e)** *Analiza la simetría de la distribución de frecuencias.*

Nos apoyaremos en la tabla estadística siguiente:

| $I_i$ | $x_i$ | $n_i$ | $x_i n_i$ | $N_i$ | $(x_i - \bar{x})^2 n_i$ |
|---|---|---|---|---|---|
| (10, 28] | 19 | 16 | 304 | 16 | 19044 |
| (28, 46] | 37 | 17 | 629 | 33 | 4628,25 |
| (46, 64] | 55 | 11 | 605 | 44 | 24,75 |
| (64, 82] | 73 | 13 | 949 | 57 | 4943,25 |
| (82, 100] | 91 | 15 | 1365 | 72 | 21093,75 |
| | | 72 | 3852 | | 49734 |

**a)** El promedio de días de supervivencia es $\bar{x} = 53{,}50$ días.

**b)** El número más frecuente de días de supervivencia ha sido

$$Mo \in (28,\ 46] \quad \Rightarrow \quad Mo = 28 + \frac{17 - 16}{1 + 6} \times 18 = 30{,}57 \text{ días}$$

**c)** El máximo de días de vida de los que han muerto antes es equivalente a la mediana de la distribución. Se tiene que

$$Me \in (46,\ 64] \quad \Rightarrow \quad Me = 46 + \frac{36 - 33}{11} \times 18 = 50{,}91 \text{ días}$$

**d)** La varianza es $s^2 = 49734/72 = 690{,}75$ días$^2$, con lo cual $s = 26{,}28$ días es la desviación típica.

**e)** Puesto que $Mo < Me < \bar{x}$, la distribución es asimétrica positiva (de cola derecha).

**7.** *Preguntamos a 20 personas sobre el número de veces que han acudido a su centro de salud durante el pasado año. Los resultados son los siguientes:*

*0 3 3 4 4 6 1 2 2 3*

*3 3 6 5 5 3 6 4 4 4*

**a)** *Elabora una tabla de frecuencias absolutas, relativas y acumuladas.* **b)** *¿Cuántas personas acudieron al centro de salud 5 veces durante el pasado año?* **c)** *¿Cuántas personas fueron al centro de salud 4 veces o más?* **d)** *¿Qué porcentaje de personas ha acudido al centro de salud 4 veces?* **e)** *¿Qué porcentaje de personas ha asistido al centro de salud, a lo sumo, 2 veces?* **f)** *Calcula la media aritmética, la moda y la mediana e interpreta los resultados.* **g)** *Calcula el tercer cuartil y el octavo decil e interpreta los resultados.* **h)** *Valora la representatividad de la media aritmética de la distribución.*

**a)** La tabla de frecuencias es, añadiéndole dos columnas adicionales:

| $x_i$ | $n_i$ | $N_i$ | $f_i$ | $F_i$ | $x_i n_i$ | $x_i^2 n_i$ |
|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 0,05 | 0,05 | 0 | 0 |
| 1 | 1 | 2 | 0,05 | 0,1 | 1 | 1 |
| 2 | 2 | 4 | 0,1 | 0,2 | 4 | 8 |
| 3 | 6 | 10 | 0,3 | 0,5 | 18 | 54 |
| 4 | 5 | 15 | 0,25 | 0,75 | 20 | 80 |
| 5 | 2 | 17 | 0,1 | 0,85 | 10 | 50 |
| 6 | 3 | 20 | 0,15 | 1 | 18 | 108 |
| | 20 | | 1 | | 71 | 301 |

**b)** Dos personas, y se corresponde con $n_6$ en la tabla.

**c)** Nos piden $n - N_4 = 20 - 10 = 10$. Otra forma de verlo es $n_5 + n_6 = 10$.

**d)** Nos piden $f_5 \times 100$, que será 25 %.

**e)** Se nos pide $F_3 \times 100 = (f_1 + f_2 + f_3) \times 100 = 20\,\%$.

**f)** La media es $\bar{x} = \frac{71}{20} = 3{,}55$ veces. La moda es $x_4 = 3$ veces. La mediana es $M_e = \frac{3+4}{2} = 3{,}5$ veces ya que $N_4 = \frac{n}{2} = 10$. Por tanto, el grupo encuestado acude, en promedio, 3,55 veces al centro de salud durante el pasado año. Además, lo más frecuente es asistir en 3 ocasiones, mientras que el 50 % que más asiste, lo hace un mínimo de 4 veces y el 50 % que menos un máximo de 3.

**g)** Dado $\frac{3}{4}n = 15$ y $x_5 = 4$ es el primer valor con $N_i \geq 15$, entonces $Q_3 = \frac{4+5}{2} = 4{,}5$, lo que significa que el 25 % que más va al centro de salud lo hace un mínimo de 5 veces.

Dado que $\frac{8}{10}n = 16$ y $x_6 = 5$ es el primer valor con $N_i \geq 16$, entonces $D_8 = x_6 = 5$, lo que significa que el 20 % que más va al centro de salud lo hace un mínimo de 5 veces.

**h)** La varianza es $s^2 = \frac{301}{20} - 3{,}55^2 = 2{,}45$ veces$^2$ $\Rightarrow$ $s = 1{,}56$ veces y, por tanto, el coeficiente de variación de Pearson queda $CV = \frac{1{,}56}{3{,}55} \times 100 = 43{,}94\,\%$, lo que implica una baja representatividad de la media aritmética.

**8.** *Se estudian los salarios que perciben los sanitarios de un hospital. El menor de los salarios es de 600 euros/mes y el mayor de 2400 euros/mes. ¿Cuál de los siguientes resultados puede ser cierto?* **a)** $\bar{x}$ *=1200 euros, s =0 euros,* **b)** $\bar{x}$ *=1000 euros, s =200 euros,* **c)** $\bar{x}$ *=500 euros, s =200 euros,* **d)** $\bar{x}$ *=1200 euros, s =-150 euros.*

La única opción posible es la **b)**. El motivo por el que podemos descartar **a)** tiene que ver con que la desviación típica no es 0 en el momento en que hay dos valores distintos de la variable. Además, **c)** no es posible, ya que la media aritmética de los valores de la variable debe estar comprendida entre el mínimo y el máximo de esta y, finalmente, podemos descartar **d)** porque la desviación típica nunca puede ser negativa (recuérdese que, de hecho, se define como la raíz cuadrada positiva de la varianza).

**9.** *La siguiente tabla recoge el tiempo (en segundos) que han durado las últimas 100 llamadas al 112.*

| *Duración de la llamada* | *[0,20)* | *[20,60)* | *[60,120)* | *[120,200]* |
|---|---|---|---|---|
| *Frecuencia relativa acumulada* | *0,1* | *0,44* | *0,8* | *1* |

*Determina* **a)** *el porcentaje de llamadas han durado menos de 2 minutos,* **b)** *el número llamadas que tienen una duración de, como mínimo, 1 minuto,* **c)** *la duración media de una llamada y evalúa su representatividad,* **d)** *la duración más habitual de estas llamadas,* **e)** *el tiempo que duran, como mínimo, las llamadas más duraderas,* **f)** *el porcentaje de llamadas tiene una duración mínima de medio minuto.*

La tabla estadística en la que apoyarnos para dar respuesta a las diferentes cuestiones planteadas es

| $I_i$ | $c_i$ | $F_i$ | $f_i$ | $n_i$ | $N_i$ | $c_i n_i$ | $c_i^2 n_i$ | $h_i$ |
|---|---|---|---|---|---|---|---|---|
| $[0, 20)$ | 10 | 0,1 | 0,1 | 10 | 10 | 100 | 1000 | 0,5 |
| $[20, 60)$ | 40 | 0,44 | 0,34 | 34 | 44 | 1360 | 54400 | 0,85 |
| $[60, 120)$ | 90 | 0,8 | 0,36 | 36 | 80 | 3240 | 291600 | 0,6 |
| $[120, 200]$ | 160 | 1 | 0,2 | 20 | 100 | 3200 | 512000 | 0,25 |
| | | | 1 | 100 | | 7900 | 859000 | |

**a)** El porcentaje de llamadas que duran menos de 2 minutos se corresponde con $F_3 \times 100 = 80\,\%$.

**b)** El número llamadas que tienen una duración de, como mínimo, 1 minuto es $n_3+n_4 = n - N_2 = 56$.

**c)** La duración media de una llamada es $\bar{x} = 7900/100 = 79$ segundos. Con objeto de evaluar su representatividad, hemos de hallar el coeficiente de variación de Pearson, apoyándonos, para ello, en el cálculo de la varianza:

$$s^2 = \frac{859000}{100} - 79^2 = 2349 \text{ llamadas}^2 \Rightarrow s = 48{,}47 \text{ llamadas}$$

Luego,

$$CV = \frac{48{,}47}{79} = 0{,}6135 \longrightarrow 61{,}35\,\%$$

lo que implica una baja representatividad de la media aritmética.

**d)** La duración más habitual se corresponde con la moda. Dado que los intervalos no tienen la misma amplitud, el intervalo modal será el de mayor densidad de frecuencia, $h_i$, hallada como el cociente entre la frecuencia absoluta y la amplitud del correspondiente intervalo. Así, $Mo \in [20, 60)$, y se tiene que

$$M_o = 20 + \frac{0{,}85 - 0{,}5}{0{,}85 - 0{,}5 + 0{,}85 - 0{,}6} \times 40 = 43{,}33 \text{ segundos}$$

**e)** El valor que divide las llamadas entre las más y las menos duraderas es la mediana. Dado que el primer intervalo cuya frecuencia absoluta acumulada supera o iguala 50 es $[60, 120)$, se tiene que $Me \in [60, 120)$ y, concretamente,

$$Me = 60 + \frac{50 - 44}{36} \times 60 = 70 \text{ segundos}$$

es el tiempo mínimo de las llamadas más duraderas.

**f)** El porcentaje de llamadas que tienen una duración mínima de medio minuto será $x\,\%$, donde $x$ es tal que $P_x = 30$. Dicho percentil está, como es natural, en el segundo intervalo, así que

$$P_x = 20 + \frac{\frac{x}{100}100 - 10}{34} \times 40 = 30 \Leftrightarrow x = 18{,}5$$

Así, el $18{,}5\,\%$ de las llamadas duran menos de medio minuto.

**10.** *Se han medido los niveles de colinesterasa (en unidades por mililitro) en un recuento de eritrocitos de 34 agricultores expuestos a insecticidas agrícolas, obteniéndose los resultados siguientes:*

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *10,6* | *12,5* | *11,1* | *9,2* | *11,5* | *9,9* | *11,9* | *11,6* | *14,9* | *12,5* | *10,1* | *10,2* |
| *12,5* | *12,3* | *12,2* | *10,8* | *16,5* | *15,0* | *10,3* | *12,4* | *9,1* | *7,8* | *12,4* | |
| *11,3* | *12,3* | *9,7* | *12,0* | *11,8* | *12,7* | *11,4* | *9,3* | *8,6* | *8,5* | *11,1* | |

*Agrupando los datos en 6 intervalos de amplitud constante de tal forma que el extremo inferior sea 7,5, determina* **a)** *el nivel de colinesterasa más frecuente,* **b)** *el nivel de colinesterasa que debe tener un agricultor para que se le pueda considerar entre el 50% de mayor nivel,* **c)** *el nivel de colinesterasa que debe tener un agricultor para que se le pueda considerar entre el 25% de mayor nivel,* **d)** *el nivel de colinesterasa promedio con los datos agrupados y también con los datos sin agrupar, ¿cuál de los promedios es más preciso?, ¿por qué?,* **e)** *el número de agricultores que han tenido un nivel de colinesterasa superior a la media.*

Tenemos a nuestra disposición $n = 34$ datos, siendo el máximo 16,5 y el mínimo 7,8, con lo cual el rango de la variable es $R = 16{,}5 - 7{,}8 = 8{,}7$ y si el número de intervalos $k$ ha de ser 6, entonces $\frac{R}{k} = \frac{8{,}7}{6} = 1{,}45 \approx a$, donde $a$ denota la amplitud. Si $a = 1{,}45$ y $L_1 = 7{,}5$, entonces el último intervalo no contiene al dato máximo pues $L_1 + k\,a = 16{,}2$. Ahora bien, si $a = 1{,}6$ y $L_1 = 7{,}5$, entonces $L_1 + k\,a = 17{,}1$. Así pues, la tabla queda

| $I_i$ | $c_i$ | $n_i$ | $N_i$ | $c_i n_i$ |
|---|---|---|---|---|
| (7,5, 9,1] | 8,3 | 4 | 4 | 33,2 |
| (9,1, 10,7] | 9,9 | 8 | 12 | 79,2 |
| (10,7, 12,3] | 11,5 | 13 | 25 | 149,5 |
| (12,3, 13,9] | 13,1 | 6 | 31 | 78,6 |
| (13,9, 15,5] | 14,7 | 2 | 33 | 29,4 |
| (15,5, 17,1] | 16,3 | 1 | 34 | 16,3 |
| | | 34 | | 386,2 |

**a)** El valor más frecuente se corresponde con la moda. Dado que disponemos de una agrupación de datos en intervalos, la determinamos a partir del intervalo modal, que es el de mayor frecuencia, pues todos los intervalos tienen la misma amplitud. Por tanto,

$$Mo \in (10{,}7, 12{,}3] \quad \Rightarrow \quad Mo = 10{,}7 + \frac{13-8}{(13-8)+(13-6)} \times 1{,}6 = 11{,}37 \text{ ud/mL}$$

**b)** $0{,}5 \times 34 = 17$, con lo que $Me \in (10{,}7,\ 12{,}3]$. Más específicamente,

$$Me = 10{,}7 + \frac{17-12}{13} \times 1{,}6 = 11{,}315 \text{ ud/mL}$$

**c)** $0{,}75 \times 34 = 25{,}5$, con lo que $Q_3 \in (12{,}3, 13{,}9]$. Más específicamente,

$$Me = 12{,}3 + \frac{25{,}5-25}{6} \times 1{,}6 = 12{,}43 \text{ ud/mL}$$

**d)** La media sin agrupar es 11,35 ud/mL, mientras que la media agrupada es

$$\bar{x} = \frac{386{,}2}{34} = 11{,}36 \text{ ud/mL}$$

Es más preciso el primer promedio, pues tiene en cuenta los valores que ha tomado la variable, mientras que la media aritmética para datos agrupados siempre considera como valores las marcas de clase, en lugar de los datos exactos.

**e)** El número de agricultores que han tenido un nivel de colinesterasa superior a la media es 17, si revisamos los datos proporcionados al principio.

**11.** *En una clínica odontológica, se ha observado el número de dientes permanentes cariados en 50 niños de 12 años, y los resultados han sido:*

| *Número de dientes cariados* | *0* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* |
|---|---|---|---|---|---|---|---|---|---|
| *Número de niños* | *8* | *12* | *10* | *6* | *4* | *4* | *4* | *0* | *2* |

**a)** *¿Cuántos niños tienen al menos 6 dientes cariados?, ¿qué porcentaje de niños tienen menos de 3 dientes cariados?* **b)** *¿Cuál es el número medio de dientes cariados de estos niños?* **c)** *¿Cuál es el número de dientes cariados más frecuente en estos niños?* **d)** *Si los niños se dividen en dos grupos: los de boca más sana y los de boca menos sana, ¿cuál es el número máximo de dientes cariados de los niños con una boca más sana?*

Primeramente, construyamos una tabla estadística con las correspondientes columnas para dar respuesta a las distintas cuestiones:

| $x_i$ | $n_i$ | $N_i$ | $f_i$ | $F_i$ | $x_i n_i$ |
|---|---|---|---|---|---|
| 0 | 8 | 8 | 0,16 | 0,16 | 0 |
| 1 | 12 | 20 | 0,24 | 0,4 | 12 |
| 2 | 10 | 30 | 0,2 | 0,6 | 20 |
| 3 | 6 | 36 | 0,12 | 0,72 | 18 |
| 4 | 4 | 40 | 0,08 | 0,8 | 16 |
| 5 | 4 | 44 | 0,08 | 0,88 | 20 |
| 6 | 4 | 48 | 0,08 | 0,96 | 24 |
| 8 | 2 | 50 | 0,04 | 1 | 16 |
| | 50 | | 1 | | 126 |

**a)** Hay $n - N_6 = 50 - 44 = 6$ niños con al menos 6 dientes cariados, y el porcentaje de niños con menos de 3 dientes cariados es del $F_3 \times 100 = 60\,\%$.

**b)** El número medio de dientes cariados es $\bar{x} = 126/50 = 2{,}52$.

**c)** El número más frecuente de dientes cariados (moda) es 1, que es el valor con mayor frecuencia.

**d)** Se trata de la mediana y su valor es 2, ya que dicho valor es el primero cuya frecuencia absoluta acumulada supera la mitad del tamaño de la población.

# Capítulo 2

# Estadística descriptiva bivariante

**12.** *Se ha realizado un estudio sobre los niños que acuden a un centro de fisioterapia, observándose el número de sesiones de rehabilitación (X) a las que tienen que asistir para su recuperación y su edad (Y) y los resultados quedan recogidos en la tabla que sigue:*

| | ***Edad*** | | | |
|---|---|---|---|---|
| ***Nº sesiones de rehabilitación*** | *3* | *4* | *5* | *6* |
| *2* | *1* | *0* | *0* | *0* |
| *3* | *0* | *1* | *2* | *0* |
| *4* | *0* | *1* | *0* | *1* |

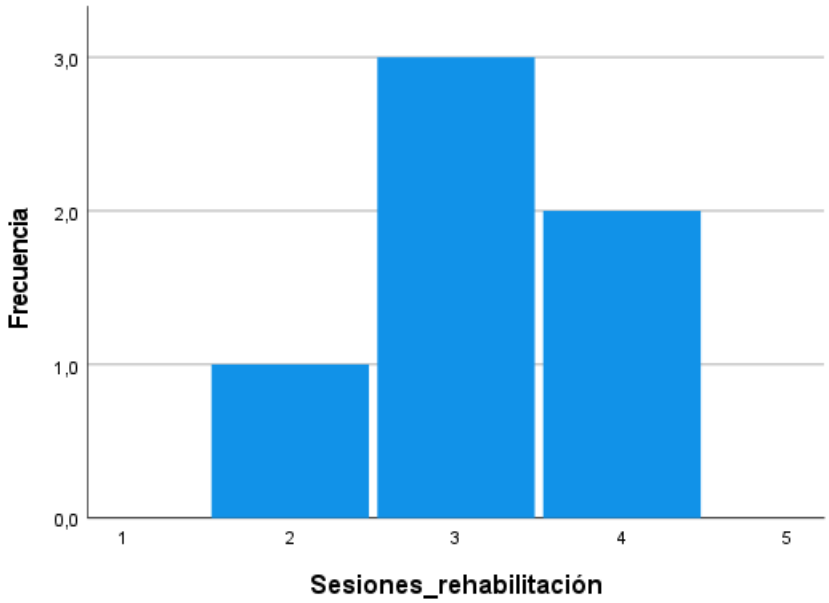
**a)** *¿Cuantos niños acuden a este centro?* **b)** *Construye el gráfico adecuado correspondiente a la variable X.* **c)** *¿Son independientes las variables observadas?* **d)** *¿Podemos afirmar que con el aumento de la edad de los niños, disminuye el número de sesiones de rehabilitación?* **e)** *¿Cuál es la edad más habitual de los niños que reciben más de 2 sesiones?* **f)** *Calcula e interpreta la mediana de la variable X.*

Comenzaremos confeccionando la siguiente tabla, con objeto de recopilar distintos cálculos para dar respuesta al resto de apartados con posterioridad más cómodamente:

| $x_i$ | $y_j$ | $n_{i,j}$ | $x_i n_{i,j}$ | $y_j n_{i,j}$ | $x_i y_j n_{i,j}$ |
|---|---|---|---|---|---|
| 2 | 3 | 1 | 2 | 3 | 6 |
| 3 | 4 | 1 | 3 | 4 | 12 |
| 3 | 5 | 2 | 6 | 10 | 30 |
| 4 | 4 | 1 | 4 | 4 | 16 |
| 4 | 6 | 1 | 4 | 6 | 24 |
| | | 6 | 19 | 27 | 88 |

**a)** Basta con sumar las frecuencias absolutas conjuntas presentes en la tabla. Se trata, pues, de 6 niños.

**b)** El gráfico más apropiado para la variable $X$ es el diagrama de barras, representado a continuación:




**c)** Dos variables estadísticas son independientes si, y solo si para cada par de valores de la distribución conjunta se tiene que $\frac{n_{i,j}}{n} = \frac{n_{i,\cdot}}{n} \times \frac{n_{\cdot,j}}{n}$. Las variables objeto de estudio no son independientes, puesto que, por ejemplo, $n_{1,1} = 1$, $n_{1,\cdot} = 1$, $n_{\cdot,1} = 1$, $n = 6$ y se tiene que

$$\frac{1}{6} \neq \frac{1}{6} \times \frac{1}{6}$$

**d)** Nótese que, en virtud de los valores de la tabla previa, los promedios son

$$\bar{x} = \frac{19}{4} = 3{,}17 \text{ sesiones de rehabilitación} \quad \bar{y} = \frac{27}{6} = 4{,}5 \text{ años}$$

y, por tanto, la covarianza queda

$$s_{XY} = \frac{88}{6} - 3{,}17 \times 4{,}5 = 0{,}4 \text{ sesiones de rehabilitación} \times \text{años}$$

que, al ser positiva, prueba una relación positiva entre edad y número de rehabilitaciones y la respuesta a la pregunta es negativa.

**e)** Se corresponde con la moda de la distribución condicionada de $Y$ dado que $X > 2$. Esta distribución es bimodal, y los valores con mayor frecuencia son 4 y 5 años.

**f)** El primer valor de la variable $X$ cuya frecuencia absoluta acumulada supera o iguala 3 es el valor $x_2 = 3$. Como su frecuencia absoluta acumulada supera estrictamente 3, podemos afirmar que $Me = 3$.