10 Estadística unidimensional

EJERCICIOS PROPUESTOS

1. **Clasifica justificadamente las siguientes variables en cualitativas, cuantitativas discretas o continuas.**

a) Número de defectos en un modelo de automóvil.

b) Origen, por continentes, de los ciudadanos extranjeros residentes en España.

c) Edad de los habitantes de una determinada ciudad.

d) Cociente intelectual de los alumnos de una escuela.

e) Distribución del PIB de un país en 2012 por sectores económicos (agricultura, industria, servicios).

f) Temperatura registrada el día 1 de julio en cada una de las capitales de provincia de España.

g) Color de ojos de los alumnos y alumnas de un centro.

h) Toneladas de cereal producidas en una determinada región en los últimos 50 años.

a) Variable cuantitativa discreta

b) Variable cualitativa

c) Variable cuantitativa discreta

d) Variable cuantitativa continua

e) Variable cualitativa

f) Variable cuantitativa continua

g) Variable cualitativa

h) Variable cuantitativa continua

2. **La tabla recoge la percepción de la situación económica de España, según el barómetro del CIS (enero 2013), sobre una muestra de 2483 personas**

x_i	Muy buena	Buena	Regular	Mala	Muy mala	NS/NC	Total
f_i	1	16	207	942	1313	4	2483

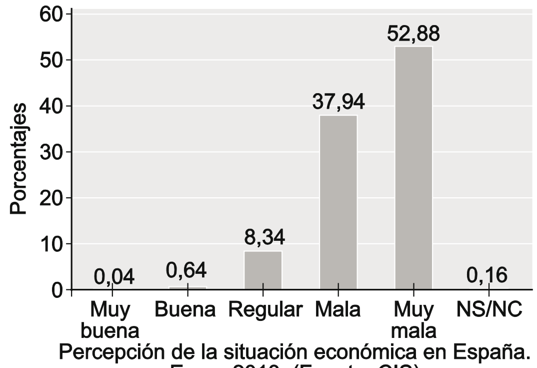
a) Halla las frecuencias relativas y los porcentajes.

b) Representa gráficamente los porcentajes mediante un diagrama de barras y uno de sectores

a) Las frecuencias relativas y los porcentajes para las categorías de la variable se recogen en la tabla siguiente

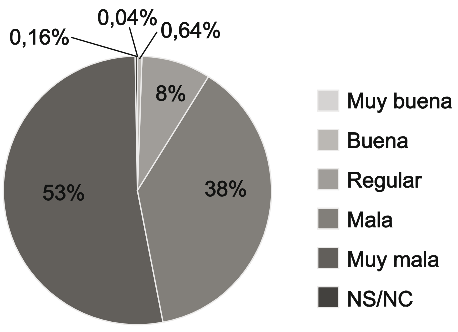
X	f_i	h_i	%
Muy buena	1	0,00040	0,04
Buena	16	0,00644	0,64
Regular	207	0,08337	8,34
Mala	942	0,37938	37,94
Muy mala	1313	0,52880	52,88
NS/NC	4	0,00161	0,16
Total	2483	1	100

b) Los diagramas de barras y de sectores:



Percepción de la situación económica en España. Enero 2013. (Fuente: CIS)

Percepción de la situación económica en España. Enero 2013. (Fuente: CIS)



3. A lo largo del último mes, las urgencias atendidas en un centro de salud han sido las siguientes

0 1 3 1 2 7 4 1 2 1 4 3 6 5 3

1 2 2 5 8 6 3 0 1 1 3 5 4 2 2

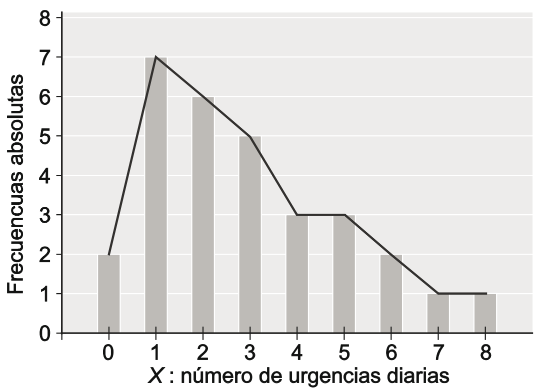
a) Construye la tabla de frecuencias.

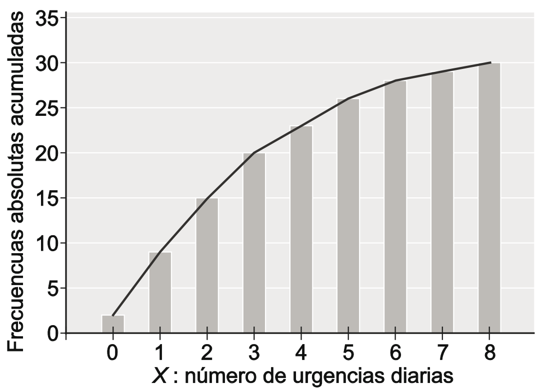
b) Representa gráficamente los datos mediante un diagrama de barras de frecuencias absolutas y otro de frecuencias absolutas acumuladas. Dibuja los correspondientes polígonos de frecuencias.

a) Efectuado el recuento, la tabla de frecuencias absolutas y absolutas acumuladas queda:

x_i	f_i	F_i
0	2	2
1	7	9
2	6	15
3	5	20
4	3	23
5	3	26
6	2	28
7	1	29
8	1	30
	30	

b) Los diagramas de barras de frecuencias absolutas y absolutas acumuladas son:





4. Ejercicio resuelto.

5. En una encuesta sobre el precio mensual del alquiler en euros (X), se han obtenido los datos de 1000 viviendas repartidas en una amplia región.

Clases	f_i
[100, 200)	42
[200, 300)	55
[300, 500)	135
[500, 700)	300
[700, 1000)	283
[1000, 1500)	110
[1500, 2000]	75

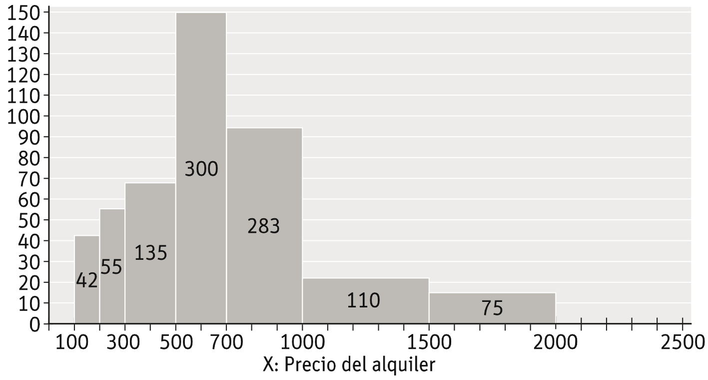
a) Representa gráficamente los datos mediante el histograma de frecuencias absolutas.

b) Dibuja el histograma de frecuencias relativas acumuladas

c) Representa los polígonos de frecuencias en los histogramas anteriores.

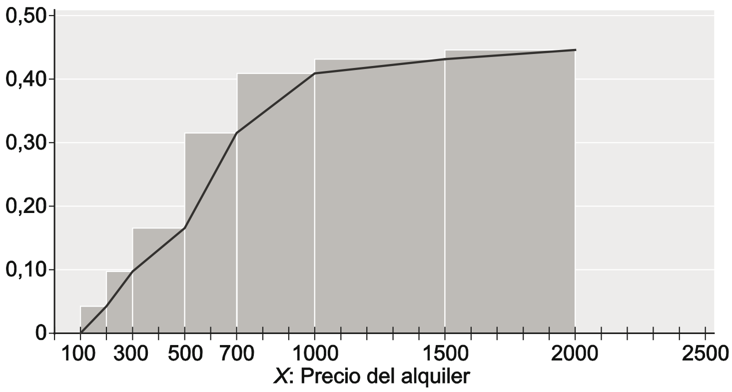
a) Como los intervalos no tienen la misma longitud, para dibujar los rectángulos del histograma se debe calcular su altura en función de la densidad de frecuencias correspondiente a cada clase. Para ello se elige una unidad, en este caso unidad=100 y la altura del rectángulo se establece para que el área represente la frecuencia absoluta:

X	f_i	Altura
[100, 200)	42	42
[200, 300)	55	55
[300, 500)	135	67,5
[500, 700)	300	150
[700, 1000)	283	94,3
[1000, 1500)	110	22
[1500, 2000]	75	15
	1000	



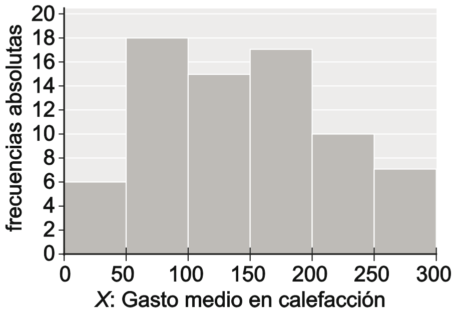
b) Para representar el histograma de frecuencias relativas acumuladas, se debe calcular la altura de cada rectángulo, teniendo en cuenta la longitud de cada clase y la frecuencia relativa que le corresponde:

X	f_i	h_i	H_i	altura	altura acumulada
[100, 200)	42	0,042	0,042	0,042	0,042
[200, 300)	55	0,055	0,097	0,055	0,097
[300, 500)	135	0,135	0,232	0,068	0,165
[500, 700)	300	0,300	0,532	0,150	0,315
[700, 1000)	283	0,283	0,815	0,094	0,409
[1000, 1500)	110	0,110	0,925	0,022	0,431
[1500, 2000]	75	0,075	1	0,015	0,446
	1000	1			



c) Los polígonos de frecuencias de los histogramas anteriores ya se han dibujado en los mismos.

6. **El gasto medio mensual en calefacción (*X*), en euros, de 73 viviendas se muestra en el siguiente gráfico .**



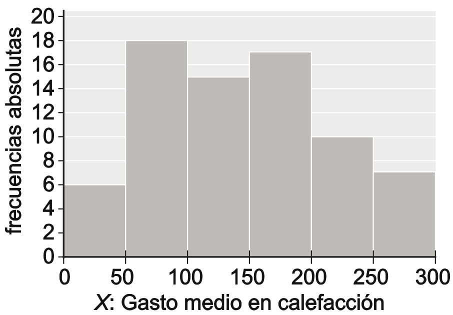
a) Completa la tabla de frecuencias.

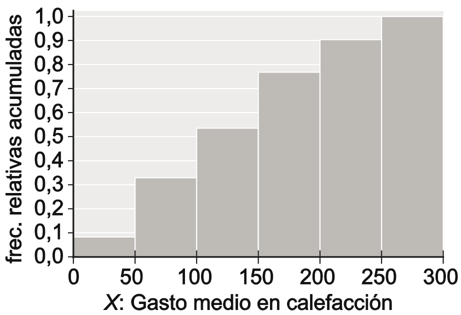
b) Dibuja los histogramas de frecuencias absolutas y de frecuencias relativas acumuladas.

a) Se completa la tabla con las frecuencias relativas (h_j) y las frecuencias acumuladas, tanto las absolutas (F_j) como las relativas (H_j):

Clases	f_j	h_j	F_j	H_j
[0, 50)	6	0,08219	6	0,08219
[50, 100)	18	0,24658	24	0,32877
[100, 150)	15	0,20548	39	0,53425
[150, 200)	17	0,23288	56	0,76712
[200, 250)	10	0,13699	66	0,90411
[250, 300]	7	0,09589	73	1,00000
	73	1		

b) Se representan los histogramas de frecuencias absolutas, sin acumular y acumuladas:





7. **Ejercicio interactivo.**

8 y 9. **Ejercicios resueltos.**

10. Considera la siguiente distribución de frecuencias de una variable cuantitativa discreta.

x_i	2	3	4	5	6	7
f_i	8	7	10	8	9	6

a) Calcula la media aritmética.

b) Determina la moda y la mediana

c) Halla los cuartiles y los percentiles 5 y 95.

a) La tabla proporciona la distribución de frecuencias absolutas y absolutas acumuladas necesarias para contestar las preguntas.

x_i	f_j	F_j	f_jx_j
2	8	8	16
3	7	15	21
4	10	25	40
5	8	33	40
6	9	42	54
7	6	48	42
	48		213

La media es entonces: $\overline{x} = \dfrac{213}{48} = 4{,}4375$.

b) Usando la tabla anterior tenemos que la modal es $M_o = 4$, al ser este el valor de frecuencia absoluta

Para la mediana se tiene en cuenta que el 50 % de 48 es 24, y ordenados los datos de menor a mayor, el valor de la variable que ocupa los lugares 24 y 25 es 4, de acuerdo a la columna de frecuencias absolutas acumuladas. Por tanto es $M = 4$.

c) Utilizando la columna de las frecuencias acumuladas de la tabla, los cuartiles Q_1 y Q_3 se calculan de forma similar a la mediana:

El 25 % de 48 es 12 y por tanto el primer cuartil corresponde al valor 3, por tanto $Q_1 = 3$.

De igual forma , el 75% de 48 es 36, con lo que se llega a $Q_1 = 6$.

Los percentiles 5 y 95 se calculan por el mismo procedimiento que los cuartiles:

El 5 % de 48 es 2,4; de manera que el percentil 5 corresponde al valor 2, esto es, $p_5 = 2$.

El 95 % de 48 es 45,6; con los que finalmente, se tiene $p_{95} = 7$.

11. La tabla recoge la distribución acumulada de las edades de 120 habitantes de una aldea.

a) Completa la tabla de frecuencias.

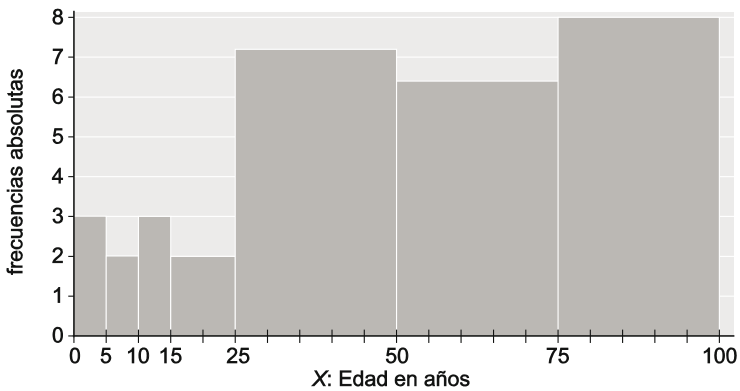
b) Dibuja el histograma de frecuencias absolutas.

c) Calcula la media, la moda y la mediana y los deciles 1 y 9.

a) La tabla proporciona la distribución de las frecuencias absolutas acumuladas. Se incluyen las clases, las frecuencias absolutas y relativas y las frecuencias relativas acumuladas y la columna $f_i \cdot x_i$.

Clases	f_j	F_j	h_j	H_j	$f_j x_j$
[0, 5)	3	3	0,0250	0,0250	7,5
[5, 10)	2	5	0,0167	0,0417	15,0
[10, 15)	3	8	0,0250	0,0667	37,5
[15, 25)	4	12	0,0333	0,1000	80,0
[25, 50)	36	48	0,3000	0,4000	1350,0
[50, 75)	32	80	0,2667	0,6667	2000,0
[75, 100]	40	120	0,3333	1,0000	3500,0
	120		1		6990,0

b) Para dibujar el histograma debe tenerse en cuenta que las clases tienen distinta amplitud. En la tabla se ha calculado la altura de los rectángulos del histograma tomando 5 años como unidad:



Clases	f_i	altura
[0, 5)	3	3
[5, 10)	2	2
[10, 15)	3	3
[15, 25)	4	2
[25, 50)	36	7,2
[50, 75)	32	6,4
[75, 100]	40	8
	120	

c) La media es $\bar{X} = \dfrac{6990}{120} = 58{,}25$ años;

La clase modal es [75, 100] ya que presenta la mayor densidad de frecuencias por unidad de medida (5 años) (el rectángulo es el de más altura en el histograma)

Para la mediana se tiene en cuenta que el 50% de 120 es 60, y por tanto está en el intervalo [50, 75), intervalo de amplitud 25, con 32 observaciones. Antes de este intervalo se acumulan 48 observaciones, luego la mediana es:

$$M = 50 + \frac{(60-48)\cdot 25}{32} = 59{,}38$$

Los deciles 1 y 9 se calculan por el mismo procedimiento:
El 10 % de 120 es 12 y, como se ve en la tabla de las frecuencias acumuladas, $D_1 = 25$.
El 90 % de 120 es 108, de manera que el percentil 90 se encuentra en el intervalo [75, 100], intervalo de amplitud 25 con 40 observaciones. Antes de este intervalo se tienen 80 observaciones:

$$D_9 = 75 + \frac{(108-80)\cdot 25}{40} = 92{,}5$$

12. Ejercicio resuelto

13. En un test de aptitudes, la puntuación obtenida (*X*) por 260 alumnos se distribuye como sigue:

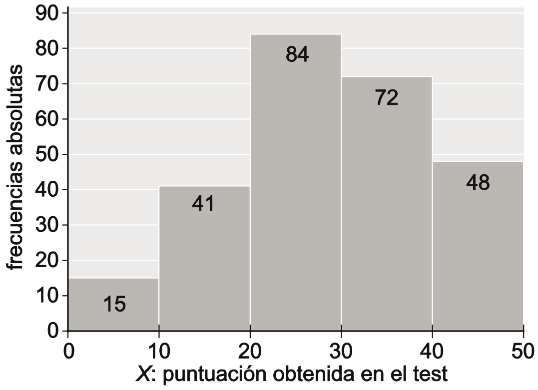
Clases	[0, 10)	[10, 20)	[20, 30)	[30, 40)	[40, 50]
f_i	15	41	84	72	48

a) Dibuja el histograma de frecuencias absolutas.

b) Calcula la media, la mediana y la moda.

c) Halla las desviaciones absoluta media y típica y el CV.

a)



b) Para los cálculos se construye la tabla siguiente:

Clases	f_i	x_i	$f_i \cdot x_i$	$f_i \cdot x_i^2$	F_i	$f_i\left\|x_i - \bar{X}\right\|$
[0, 10)	15	5	75	375	15	355,96
[10, 20)	41	15	615	9225	56	562,96
[20, 30)	84	25	2100	52 500	140	313,38
[30, 40)	72	35	2520	88 200	212	451,38
[40, 50]	48	45	2160	97 200	260	780,92
	260		7470	247 500		2464,62

La media es: $\bar{X} = \dfrac{7470}{260} = 28{,}73$,

La clase modal es [20, 30), que contiene 84 observaciones.

Cálculo de la mediana:

El 50 % de 260 es 130, de forma que la mediana está en el intervalo [20, 30), de amplitud 10 y 84 observaciones. Hasta llegar a esta clase se han acumulado 56 observaciones, luego:

$$M = 20 + \frac{(130 - 56) \cdot 10}{84} = 28{,}81$$

c) La desviación absoluta media es: $D_{\bar{x}} = \dfrac{2464{,}62}{260} = 9{,}479$

Se calculan la varianza y la desviación típica:

$$s^2 = \frac{247500}{260} - 28{,}73^2 = 126{,}466 \Rightarrow s = \sqrt{126{,}466} = 11{,}246$$

Finalmente, el coeficiente de variación, que mide la variabilidad relativa de las puntuaciones del test:

$$CV = \frac{11{,}246}{28{,}73} = 0{,}3914$$

14. Ejercicio interactivo.

15 a 20 Ejercicios resueltos.

EJERCICIOS

Tablas de frecuencia y gráficos

21. Para determinar si un dado es equilibrado o no, se lanza 100 veces y se anota el número obtenido en cada lanzamiento

6	**1**	**1**	**2**	**5**	**1**	**6**	**4**	**4**	**3**	**4**	**3**	**6**	**5**	**1**	**6**	**1**	**2**	**3**	**1**
3	**3**	**4**	**5**	**6**	**4**	**2**	**5**	**4**	**4**	**3**	**5**	**2**	**3**	**6**	**5**	**6**	**5**	**1**	**3**
6	**1**	**5**	**6**	**5**	**6**	**5**	**2**	**1**	**2**	**6**	**3**	**2**	**5**	**2**	**4**	**3**	**4**	**3**	**3**
6	**4**	**2**	**5**	**5**	**4**	**3**	**5**	**2**	**3**	**1**	**3**	**6**	**2**	**6**	**4**	**1**	**1**	**4**	**2**
5	**2**	**1**	**6**	**6**	**1**	**4**	**6**	**1**	**4**	**2**	**1**	**3**	**4**	**3**	**1**	**6**	**1**	**5**	**2**

a) Forma la tabla de frecuencias absolutas y relativas.

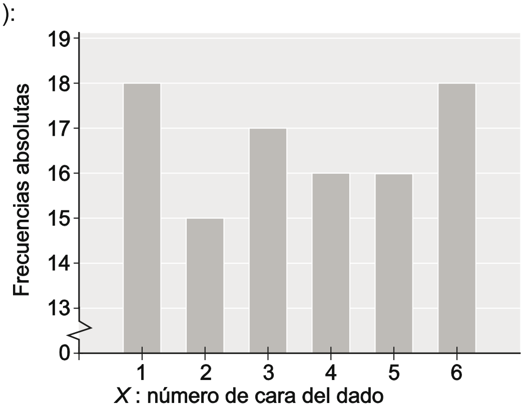
b) Representa gráficamente la distribución.

c) A la vista de la tabla y el gráfico, ¿se puede afirmar que el dado está equilibrado?

a) Una vez realizado el recuento, la distribución de frecuencias absolutas y relativas es:

x_i	f_i	h_i
1	18	0,18
2	15	0,15
3	17	0,17
4	16	0,16
5	16	0,16
6	18	0,18
	100	1

b) La distribución de las frecuencias absolutas (o relativas) se puede representar mediante un diagrama de barras (observa la escala del eje vertical):



c) A la vista del diagrama de barras y de la tabla de frecuencias, aunque hay diferencias entre las frecuencias de los distintos resultados, no hay suficiente información para decir que el dado está sesgado. Habría que aumentar el número de ensayos para poder llegar a una conclusión. Si el dado está equilibrado, la frecuencia relativa de cada resultado debería estabilizarse en torno a $\frac{1}{6}$.

22. En la tabla siguiente, se muestra la distribución de frecuencias de una variable estadística *X*.

x_i	1	3	5	7	9	11
f_i	14	10	8	30	32	21

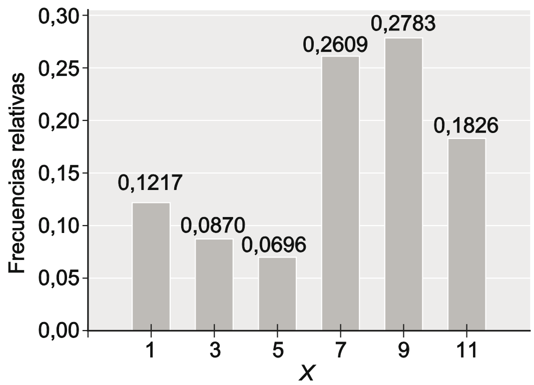
a) Halla la distribución de frecuencias absolutas acumuladas y de frecuencias relativas absolutas y acumuladas.

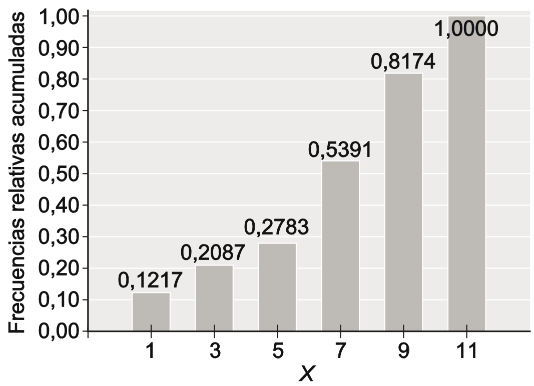
b) Representa los diagramas de frecuencias relativas y de frecuencias relativas acumuladas.

a) La tabla se amplía con las columnas de las frecuencias absolutas y las frecuencias relativas.

x_i	f_i	h_i	F_i	H_i
1	14	0,1217	14	0,1217
3	10	0,0870	24	0,2087
5	8	0,0696	32	0,2783
7	30	0,2609	62	0,5391
9	32	0,2783	94	0,8174
11	21	0,1826	115	1,0000
	115	1		

b) Los diagramas de frecuencias relativas y frecuencias relativas acumuladas son respectivamente:





23. *La primera estrofa y el estribillo de la canción del pirata, de José Espronceda (1808–1842) dicen así:

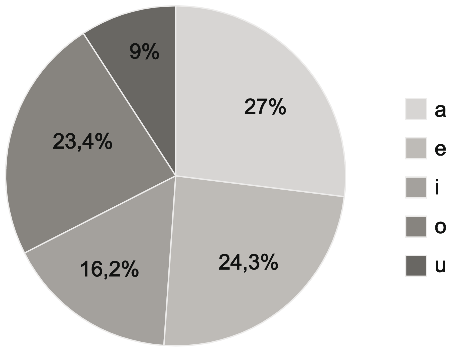
Con cien cañones por banda,
viento en popa a toda vela,
no corta el mar, sino vuela,
un velero bergantín;
bajel pirata que llaman
por su bravura, el Temido
en todo el mar conocido
del uno al otro confín
Que es mi barco mi tesoro,
que es mi Dios la libertad;
mi ley, la fuerza y el viento,
mi única patria, la mar

Obtén la distribución de frecuencias de las vocales, tantos absolutas como relativas y representa la última gráficamente.

Efectuado el recuento de vocales, la tabla con las frecuencias absolutas y relativas es:

X	f_i	h_i	Ángulo (grados sexagesimales)
a	30	0,2703	97,30
e	27	0,2432	87,57
i	18	0,1622	58,38
o	26	0,2342	84,32
u	10	0,0901	32,43
	111	1	360

Para representar gráficamente con un diagrama de sectores las frecuencias relativas, se ha añadido a la tabla la columna del ángulo que le corresponde a cada letra. Así, dicha representación queda como sigue.



24. Las altitudes máximas de cada una de las 50 provincias de España son:

898	**2083**	**1558**	**2606**	**1482**	**2648**
2591	**1110**	**2531**	**1482**	**2131**	**2399**
1654	**2613**	**1813**	**1371**	**1570**	**1839**
1544	**2913**	**3479**	**2262**	**912**	**3404**
1445	**2167**	**2262**	**1949**	**2648**	**3143**
1821	**2430**	**2065**	**2001**	**2438**	**2124**
2536	**1177**	**2425**	**3715**	**2430**	**1129**
2313	**1447**	**2024**	**1447**	**1832**	**964**
2124	**2313**				

Fuente: Instituto Geográfico Nacional

a) Obtén la distribución de frecuencias absolutas agrupando los datos en 7 clases.

b) Representa el histograma de frecuencias absolutas.

a) La altura máxima es 3715 metros y la mínima es 898,

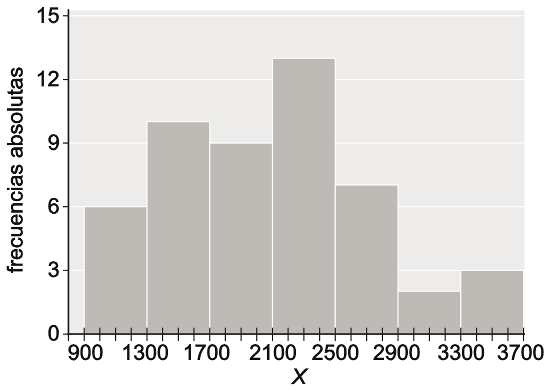
El rango de datos es $3715 - 898 = 2817$.

Si buscamos clases de similar amplitud esta debería ser próxima a $\frac{2817}{7} = 402,43$ m.

Así, si tomamos 900 m como inicio del rango y, por simplificar, con intervalos de amplitud 400, podemos organizar los datos en una tabla de frecuencias así aumentando la amplitud del último intervalo en 100 m.

Clases	x_i	f_i
[900, 1300)	1100	6
[1300, 1700)	1500	10
[1700, 2100)	1900	9
[2100, 2500)	2300	13
[2500, 2900)	2700	7
[2900, 3300)	3100	2
[3300, 3700)	3500	3
		50

b) El histograma correspondiente puede ser este:



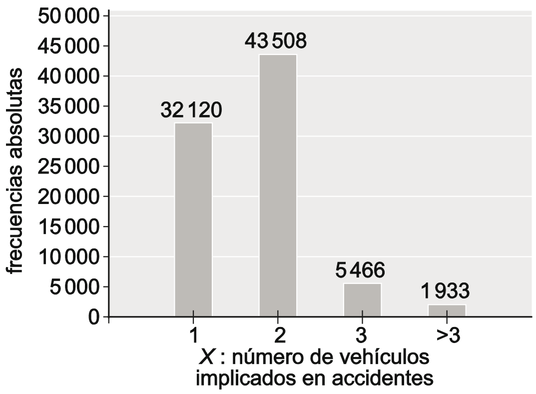
25. El número de vehículos implicados en accidentes con víctimas mortales y el número total de accidentes se dan en esta tabla:

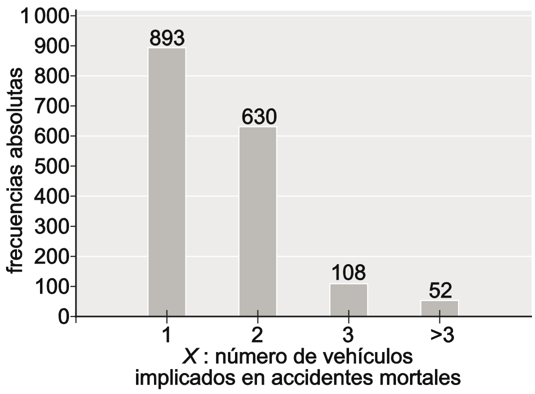
x_i	Mortales	Total
1	893	32 120
2	630	43 508
3	108	5466
> 3	52	1933

a) Representa las dos distribuciones mediante un diagrama de barras.

b) Dibuja el diagrama de sectores para cada distribución con los porcentajes correspondientes.

a) Los diagramas de barras de frecuencias absolutas son:

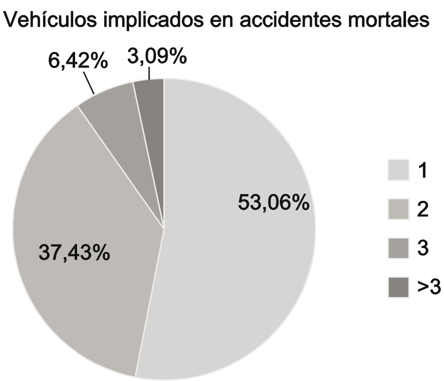


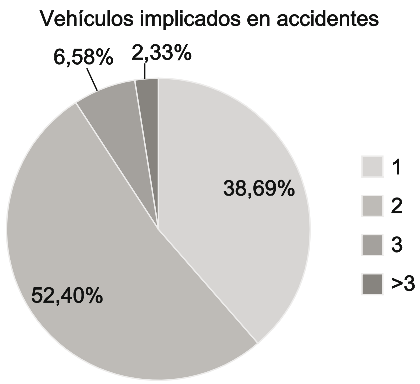


b) A la tabla se le añaden las columnas correspondientes a los porcentajes del número de accidentes con victimas mortales y del número total de accidentes, además de las columnas con los grados sexagesimales que corresponden a cada valor de la variable.

x_i	mortales	Total	% mortales	% Total	Ángulo mortales	Ángulo total
1	893	32 120	53,06 %	38,69 %	191,02	139,27
2	630	43 508	37,43 %	52,40 %	134,76	188,65
3	108	5466	6,42 %	6,58 %	23,10	23,70
> 3	52	1933	3,09 %	2,33 %	11,12	8,38
	1683	83 027	100 %	100 %	360	360

Y los diagramas de sectores correspondientes son:





26. El porcentaje de población activa, por sectores económicos, que se desprende de la Encuesta de Población Activa (EPA), de finales de 2012, incluyendo el total y distinguiendo entre hombres y mujeres, se muestra en la tabla siguiente.

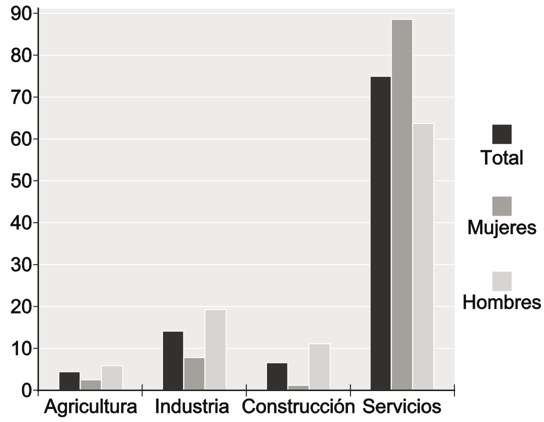
Sector	Total	Mujeres	Hombres
Agricultura	4,4	2,5	5,9
Industria	14,1	7,8	19,3
Construcción	6,6	1,2	11,1
Servicios	74,9	88,5	63,7

a) Representa las tres distribuciones en un mismo diagrama de barras.

b) ¿Qué diferencias se observan entre los hombres y las mujeres?

c) Dibuja un diagrama de sectores para cada distribución.

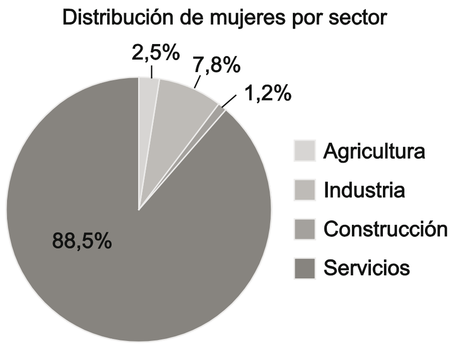
a) El diagrama de barras es el dela derecha.

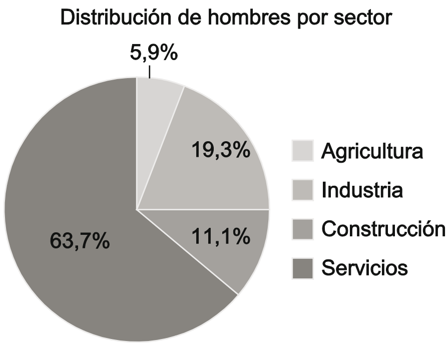


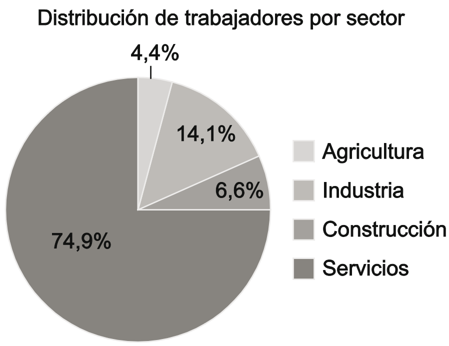
b) Entre otras observaciones, se puede ver que los hombres son mayoría en los sectores de agricultura, industria y construcción, mientras que en el sector servicios sucede lo contrario.

c) Para hacer los diagramas de sectores se incuye en la tabla una columna con los grados sexagesimales correspondientes a cada distribución:

Sector	Total	Ángulo total	Mujeres	Ángulos mujeres	Hombres	Ángulo hombres
Agricultura	4,4	15,84	2,5	9	5,9	21,24
Industria	14,1	50,76	7,8	28,08	19,3	69,48
Construcción	6,6	23,76	1,2	4,32	11,1	39,96
Servicios	74,9	269,64	88,5	318,6	63,7	229,32
		360		360		360







Medidas de localización y de dispersión

27. En cierta ocasión, se reunieron 5 miembros de una misma familia, de edades 23, 28, 32, 35 y 40 años.

a) Calcula la media y la varianza de las edades.

b) Cuatro años más tarde vuelven a reunirse los mismos familiares, ¿cuál será entonces su edad media? ¿y la varianza?

a) La media aritmética es: $\bar{X} = \dfrac{23+28+32+35+40}{5} = 31{,}6$

Y la varianza: $s^2 = \dfrac{23^2+28^2+32^2+35^2+40^2}{5} - 31{,}6^2 = 33{,}84$

b) Cuatro años más tarde, la nueva variable es $Y = X + 4$. La media será cuatro años más y la varianza la misma, es decir,

$\bar{Y} = 35{,}6$

$s_Y^2 = 33{,}84$

Si los datos se "trasladan" uniformemente, la dispersión respecto a la media no varía.

28. Sabiendo que 3 es la media del conjunto de datos:

2, 3, 2, 5, *x*, 6, 4, 0

Encuentra el valor del dato que falta.

Se escribe la media del conjunto de datos y se iguala a 3:

$$\frac{2+3+2+5+x+6+4+0}{8} = 3 \Rightarrow x+22 = 24 \Rightarrow x = 2$$

29. El salario medio (*X*) en cuatro empresas ubicadas en una misma población y el número de empleados que actualmente tiene cada una de ellas viene dado en la tabla siguiente.

Empresa	A	B	C	D
x_i	1500	1625	1450	1275
f_i	115	60	38	77

a) Calcula el ingreso medio de todos los empleados.

b) Elige y dibuja un gráfico que represente la importancia relativa de estas cuatro empresas en cuanto al número de empleados.

c) Halla la varianza y la desviación típica.

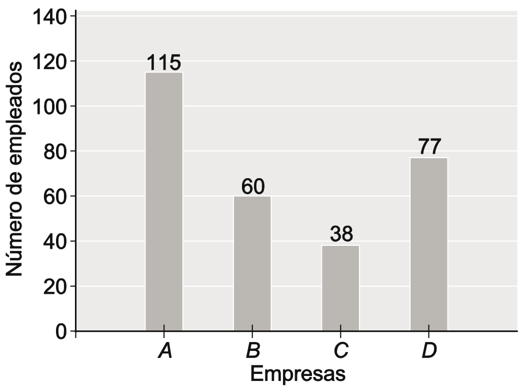
d) Calcula el coeficiente de variación y comenta el resultado que se obtiene.

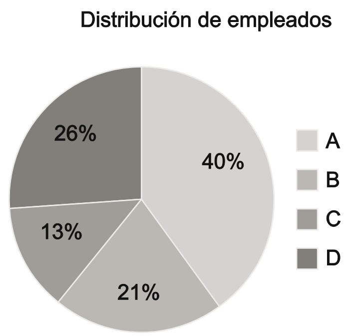
a) Para los cálculos de los apartados a), c) y d) se construye la tabla siguiente:

Empresa	x_i	f_i	f_ix_i	$f_ix_i^2$
A	1500	115	172 500	258 750 000
B	1625	60	97 500	158 437 500
C	1450	38	55 100	79 895 000
D	1275	77	98 175	125 173 125
		290	423 275	622 255 625

El salario medio de los empleados de las cuatro empresas es: $\overline{X} = \dfrac{1}{290}\sum_{i=1}^{4} f_i x_i = \dfrac{423\,275}{290} = 1459{,}57$

b) Se puede representar el número de empleados de cada empresa mediante un diagrama de barras o uno de sectores, utilizando las frecuencias absolutas o relativas (porcentajes). En el diagrama de barras se ha representado cada empresa con su número de empleados (frecuencia absoluta) y en el de sectores cada empresa con el porcentaje de empleados del total que le corresponde (frecuencia relativa).





c) Calculamos la varianza y la desviación típica usando los datos de la tabla anterior:

$$s_x^2 = \frac{1}{290}\sum_{i=1}^{4} f_i x_i^2 - \overline{X}^2 = \frac{622\,255\,625}{290} - 1459{,}57^2 = 15\,367{,}49 \Rightarrow s_x = 123{,}97$$

d) El coeficiente de variaciones, por tanto, $CV = \dfrac{s_x}{\overline{x}} = \dfrac{123{,}97}{1459{,}57} = 0{,}0849$.

El valor obtenido indica que los salarios son relativamente homogéneos entre las empresas, al presentar una baja variabilidad.

30. En un determinado mes la media aritmética de los salarios abonados por una empresa a sus empleados ascendió a 1400 euros. La media de los salarios pagados a los hombres ascendió a 1600 euros, mientras que la media de los pagados a las mujeres fue de 1350. Con esta información, ¿cuáles son los porcentajes de mujeres y hombres empleados en esta empresa?

Si *p* es la proporción de hombres que trabaja en esta empresa, 1 – *p* es la proporción de mujeres. Entonces, con la información proporcionada se tiene que:

$$1600\,p + 1350\,(1-p) = 1400$$

De donde se obtiene que $p = 0{,}2$. Es decir el 20 % de los empleados de esta empresa son hombres y el 80 % mujeres.

31. El número de faltas de asistencia (*X*), en un grupo de 35 alumnos, a la clase de Matemáticas Aplicadas a la Ciencias Sociales I, se ha agrupado en la tabla siguiente

x_i	0	1	2	3	4	5
f_i	12	8	7	5	3	2

a) Representa gráficamente la distribución de frecuencias absolutas y absolutas acumuladas.

b) Dibuja los polígonos de frecuencias absolutas y absolutas acumuladas.

c) Calcula el número medio de faltas de asistencia en esta clase. También la moda y la mediana.

d) Calcula las desviaciones absoluta media y típica.

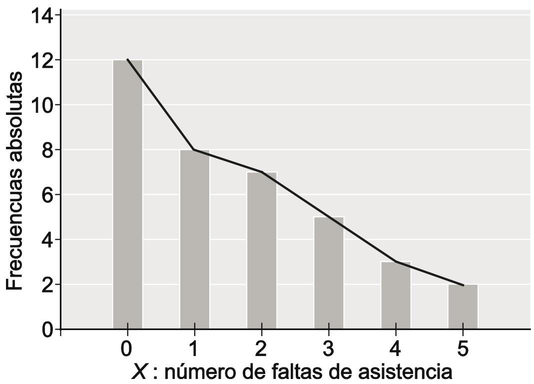
e) ¿Cuántas faltas tiene como mínimo un alumno que se encuentra entre el 25% que más falta?

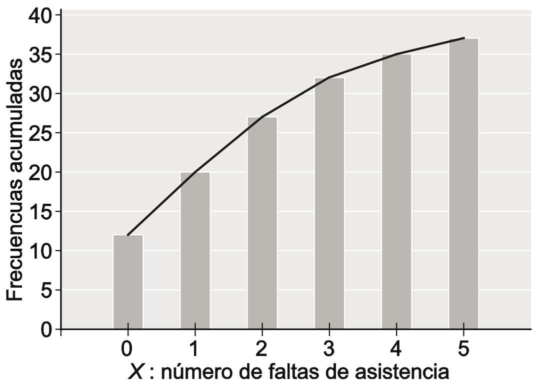
f) Determina los percentiles 34 y 67.

a) Para las representaciones gráficas y los cálculos de media y mediana se añaden a la tabla las columnas de frecuencias absolutas acumuladas y la de los productos necesarios para los apartados siguientes.

Los diagramas de barras y las gráficas de los polígonos de frecuencias absolutas y absolutas acumuladas son:

x_i	f_i	F_i	f_ix_i	$f_ix_i^2$	$\left\|x_i-\bar{X}\right\|$	$f_i\cdot\left\|x_i-\bar{X}\right\|$
0	12	12	0	0	1,59	19,08
1	8	20	8	8	0,59	4,72
2	7	27	14	28	0,41	5,74
3	5	32	15	45	1,41	7,05
4	3	35	12	48	2,41	7,23
5	2	37	10	50	3,41	6,82
	37		59	179	9,82	50,64





b) Los polígonos de frecuencias se incluyen en los gráficos anteriores.

c) El número medio de faltas viene dado por: $\bar{X}=\frac{59}{37}=1{,}59$ **faltas de asistencia**

La moda es $M_o=0$, ya que es el valor de la variable que presenta mayor frecuencia (12).

Como se tiene un número impar de datos (37), la mediana es el valor de la variable que ocupe la posición 19 que, observando la columna de las frecuencias acumuladas, es $M=1$.

d) La varianza y desviación típica vienen dadas por:

$$s_x^2=\frac{1}{37}\sum_{i=1}^{6}f_ix_i^2-\bar{X}^2=\frac{179}{37}-1{,}59^2=2{,}295\Rightarrow s_x=1{,}515$$

Para el cálculo de la desviación absoluta media usamos la última columna:

$$D_x=\frac{1}{37}\sum_{i=1}^{6}f_i\left|x_i-\bar{X}\right|=\frac{50{,}64}{37}=1{,}379$$

e) Un alumno que está entre el 25% de alumnos que más falta se sitúa por encima del tercer cuartil. Como el 75% de 37 es 27,75, mirando la columna de frecuencias absolutas acumuladas, vemos que debe tener al menos 3 faltas.

f) Para identificar los percentiles 34 y 67, vemos que:

El 34 % de 37 es 12,58. La primera frecuencia absoluta acumulada que alcanza esta cifra corresponde a $p_{34}=1$.

El 67 % de 37 es 24,79. La primera frecuencia absoluta acumulada que alcanza esta cifra corresponde a $p_{67}=1$.

32. De una muestra de 100 hogares, seleccionados aleatoriamente en una ciudad pequeña, se contabiliza el número de personas empleadas. Los datos se recogen agrupados en la tabla siguiente

x_i	0	1	2	3	4
f_i	11	35	32	13	9

a) Calcula el número medio de empleados por hogar y su desviación típica.

b) Representa los datos. ¿Puede considerarse asimétrica esta distribución?

c) Determina los percentiles 5 y 95. ¿Cuántas unidades de la muestra se encuentran entre esos dos percentiles?

a) Se amplía la tabla con las columnas necesarias para calcular los valores que se pide

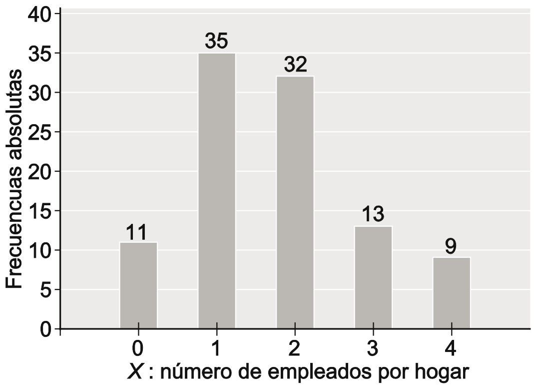
x_i	f_i	$f_i x_i$	$f_i x_i^2$
0	11	0	0
1	35	35	35
2	32	64	128
3	13	39	117
4	9	36	144
	100	174	424

Entonces, el número medio de empleados (la media de la distribución), su varianza y desviación típica son:

Media $\bar{X} = \frac{174}{100} = 1{,}74$ **empleados**

Varianza: $s_x^2 = \frac{424}{100} - 1{,}74^2 = 1{,}2124 \Rightarrow s_x = \sqrt{1{,}2124} = 1{,}1011$

b) Se representa la gráfica mediante un diagrama de barras de frecuencias absolutas:



La distribución de frecuencias es claramente asimétrica. Para ser simétrica, las frecuencias a izquierda y derecha del valor $x = 2$ deberían ser "parecidas".

c) Para el cálculo del percentil 5, p_5, se ordenan los datos de menor a mayor y el valor que deja al menos el 5 % de los datos por debajo (y como mucho el 95 % de los datos por encima) (el 5 % de 100 es 5) es $p_5 = 0$.

El percentil 95, p_{95}, el que deja como mucho 5% de los datos encima (y al menos el 95% por debajo) es $p_{95} = 4$.

Entre estos dos valores se encuentra el 90% de los datos de la muestra. Como se tienen 100 observaciones, 90 de ellas están entre estos dos valores. Ahora bien, si descontamos los valores $x = 0$ (11 observaciones) y $x = 4$ (9 observaciones), entre estos dos valores tenemos 80 observaciones distintas de 0 y 4.

33. En la tabla siguiente se muestra el número de alumnos f_i, la calificación media, x_i, y la varianza, s^2, en tres grupos de alumnos que cursan Matemáticas aplicadas a las Ciencias Sociales, de 1º de Bachillerato.

Grupo	f_i	x_i	s^2
A	31	6,2	2,8
B	35	5,8	3
C	32	5	2

a) Calcula la nota media global de los tres grupos.

b) Halla el coeficiente de variación para cada grupo y ordena los grupos por homogeneidad.

a) Para responder a las cuestiones planteadas, se añaden a la tabla las columnas de los productos $f_i x_i$ y la de los coeficientes de variación de cada grupo, CV_i

Grupo	f_i	x_i	s^2	$f_i x_i$	CV_i
A	31	6,2	2,8	192,2	0,2699
B	35	5,8	3	203	0,2986
C	32	5	2	160	0,2828
	98			555,2	

La nota media global se obtiene ponderando la media de cada grupo por su número de alumnos. Es decir:

$$\overline{X} = \frac{1}{98}\sum_{i=1}^{3} f_i x_i = \frac{552{,}2}{98} = 5{,}67$$

b) El coeficiente de variación de cada grupo se ha incluido en la última columna de la tabla

$CV_i = \dfrac{s_i}{\overline{X}}$, con lo que resulta, ordenados de menor a mayor variabilidad

$CV_A = 0{,}2699$, $CV_C = 0{,}2828$ y $CV_B = 0{,}2986$.

34. *La producción de remolacha azucarera dada en toneladas (X), en 4 fincas con distintos tipos de cultivo y distintas superficies dadas en hectáreas se da en la tabla siguiente:

Fincas	ha	x_i
A	6	42
B	10	60
C	4	32
D	7	40

a) Calcular el rendimiento por ha en cada una de las cuatro fincas.

b) Hallar el rendimiento medio global y la variabilidad por ha de terreno.

a) Se añade a la tabla la columna de los rendimientos medios por ha en cada una de las cuatro fincas, obtenido dividiendo la producción total de cada finca por su superficie.

Fincas	ha	x_i	RM/ha
A	6	42	7,000
B	10	60	6,000
C	4	32	8,000
D	7	40	5,714
	27	174	

Puede observarse que el rendimiento medio más alto se da en la finca C, 8 toneladas por ha, y el más bajo en la finca D, 5,714 toneladas por ha.

b) El rendimiento global se obtiene dividiendo la producción total 174 toneladas, entre las 27 ha de cultivo que hay entre las tres fincas. Es decir:

$$\overline{X} = \frac{174}{27} = 6{,}444 \text{ toneladas por ha.}$$

35. La distribución de las temperaturas máximas en ºC (*X*) alcanzadas en una localidad en 60 días consecutivos viene dada en la tabla de la derecha.

X	f_i
[–10,0)	11
[0, 5)	12
[5, 10)	14
[10, 15)	11
[15, 25)	12

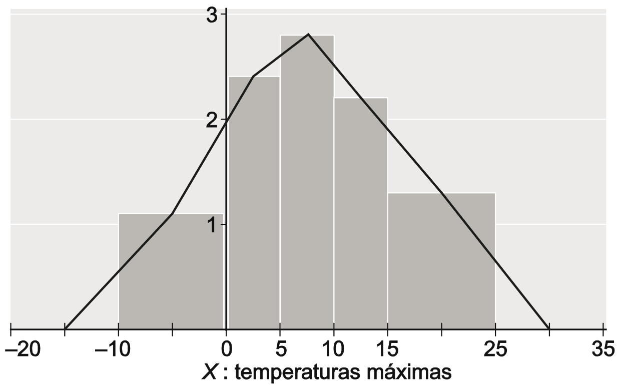
a) Representa gráficamente los datos y dibuja el polígono de frecuencias.

b) Calcula la temperatura media en la localidad de los 60 días y su desviación típica.

c) Determina la temperatura que solo fue superada el 5 % de los días.

a) Se representa la distribución de las temperaturas mediante un diagrama de barras. Se ha tomado 5 como unidad para la base de los rectángulos, de tal manera que las frecuencias de cada clase se corresponden con el área de los rectángulos.

Para dibujar el polígono de frecuencias se añaden dos clases con frecuencia cero al principio y al final de la distribución. La longitud de estas clases es la misma que la de las clases que le siguen o que le preceden respectivamente



b) Se añaden a la tabla las columnas necesarias para el cálculo de la media y de la desviación típica:

Clases	f_i	x_i	$f_i x_i$	$f_i x_i^2$	F_i
[–10,0)	11	–5	–55	275	11
[0, 5)	12	2,5	30	75	23
[5, 10)	14	7,5	105	787,5	37
[10, 15)	11	12,5	137,5	1718,75	48
[15, 25)	12	20	240	4800	60
	60		457,5	7656,25	

De esta manera, la media es: $\bar{X} = \frac{457,5}{60} = 7,625\ ºC$

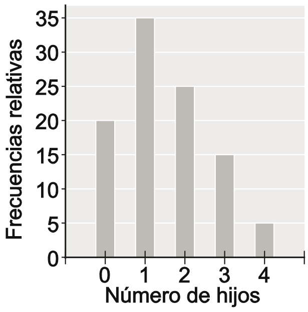
y la varianza y desviación típica son: $s^2 = \frac{7656,25}{60} - 7,625^2 = 69,4635 \Rightarrow s = 8,3345\ ºC$

c) Para determinar esta temperatura, se incluyó en la tabla anterior la columna de las frecuencias absolutas acumuladas. Se trata de encontrar el percentil 95, p_{95}, Como el 95% de 60 observaciones son 57, el percentil 95 se encuentra en el intervalo [15, 25), cuya longitud es 10 y que contiene 12 observaciones. Además antes de este intervalo se tienen acumuladas 48 observaciones. Luego:

$$p_{95} = 15 + \frac{(57-48)10}{12} = 22,5\ ºC$$

Síntesis

36. El número de hijos (*X*) de una muestra de 100 familias se recoge en el siguiente diagrama de barras.



a) Escribe la tabla de frecuencias y representa la distribución mediante el diagrama de frecuencias relativas acumuladas.

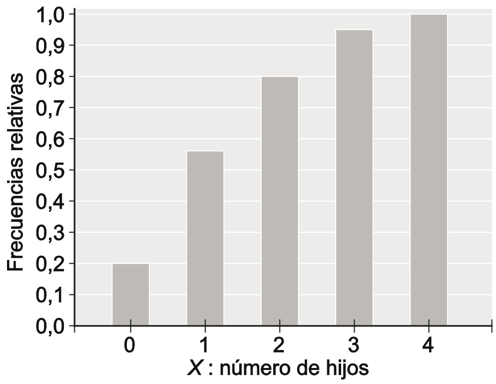
b) Calcula la media, la moda y la mediana.

c) Determina los cuartiles Q1 y Q3.

a) Construimos la tabla a la que se le añaden las columnas de las frecuencias relativas (h_i) y las frecuencias relativas acumuladas (H_i) y las columnas con los productos necesarios para los cálculos posteriores.

x_i	f_i	F_i	h_i	H_i	$f_i x_i$
0	20	20	0,20	0,20	0
1	35	55	0,35	0,55	35
2	25	80	0,25	0,80	50
3	15	95	0,15	0,95	45
4	5	100	0,05	1	20
	100		1		150

Y el diagrama de barras correspondientes es:



b) Para calcular la media y la mediana recurrimos a la tabla de manera que la media es

$$\overline{X} = \frac{1}{100}\sum_{i=1}^{5} f_i x_i = \frac{150}{100} = 1{,}5$$

La mediana, se obtiene de la siguiente manera: el 50 % de 100 es 50 y el valor de la variable *X* que, una vez ordenados los datos de menor a mayor, ocupa el lugar 50–51 (observa que el número de datos es par) es 1.

Por tanto, $M = 1$.

La moda es $M_o = 1$, al ser el valor que aparece más veces (36) en la distribución.

c) Los cuartiles primero y tercero se calculan de forma similar a la mediana (cuartil 2).

Q_1, cuartil 1º. El 25 % de 100 es 25. El valor de la variable que ocupa el lugar 25, ordenados los datos de menor a mayor, es $Q_1 = 1$. Coincide, por tanto con la mediana.

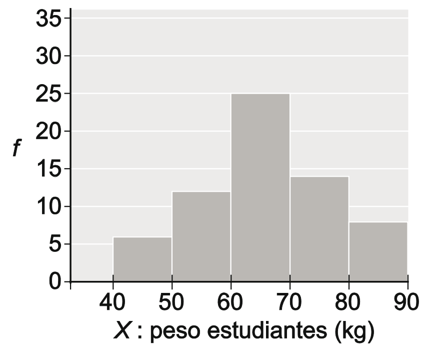
El 75 % de 100 es 75. Por tanto el cuartil $Q_3 = 2$, que es el valor que ocupa el lugar 75 una vez ordenados los datos de menor a mayor.

37. La distribución del peso en kg (X) de una muestra de 65 estudiantes de un centro educativo se muestra en el siguiente gráfico:

a) Escribe la tabla de frecuencias y dibuja el polígono de frecuencias relativas acumuladas.

b) Obtén el peso medio y desviación típica.

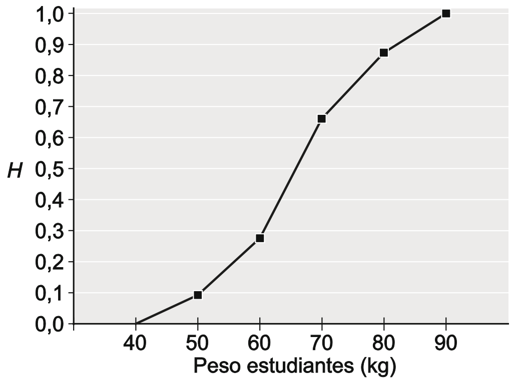
c) Calcula la mediana y los cuartiles primero y tercero.



a) A partir del histograma construimos la tabla de frecuencias ampliada para los cálculos posteriores:

X	f_i	F_i	h_i	H_i	x_i	$f_i x_i$	$f_i x_i^2$
[40, 50)	6	6	0,092	0,092	45	270	12 150
[50, 60)	12	18	0,185	0,277	55	660	36 300
[60, 70)	25	43	0,385	0,662	65	1625	105 625
[70, 80)	14	57	0,215	0,877	75	1050	78 750
[80, 90)	8	65	0,123	1	85	680	57 800
	65		1			4285	290 625

A partir de los datos de la tabla se construye el polígono de frecuencias relativas acumuladas.



b) El peso medio de los estudiantes y su desviación típica se calculan a partir de los datos de la tabla

$$\bar{X} = \frac{4285}{65} = 65{,}923\ kg \qquad s^2 = \frac{290625}{65} - 65{,}923^2 = 125{,}302 \Rightarrow s = \sqrt{125{,}302} = 11{,}194\ kg$$

c) Para el cálculo de los cuartiles se utiliza la columna de las frecuencias absolutas acumuladas.
Cuartil Q_1. El 25 % de 65 es 16,25, que se acumula en el intervalo [50, 60), de longitud 10 y que contiene 12 observaciones. Antes de éste intervalo se tienen acumuladas 6 observaciones, luego:

$$Q_1 = 50 + \frac{10 \cdot (16{,}25 - 6)}{12} = 58{,}542 kg$$

Cuartil Q_2. Es la mediana. El 50 % de 65 es 32,5, que se acumula en el intervalo [60, 70), de longitud 10 y que contiene 25 observaciones. Antes de este intervalo se tienen acumuladas 18 observaciones. Luego

$$M = Q_2 = 60 + \frac{10 \cdot (32{,}5 - 18)}{25} = 65{,}8\ kg$$

Cuartil Q_3. El 75 % de 65 es 48,75, que se acumula en el intervalo [70, 80), de longitud 10 y que contiene 14 observaciones. Antes de este intervalo se tienen acumuladas 43 observaciones. Luego:

$$Q_3 = 70 + \frac{10 \cdot (48{,}75 - 43)}{14} = 74{,}11\ \text{kg}$$

38. La distribución del número de materias suspendidas (*X*) por 38 alumnos de 1º de bachillerato es:

x_i	0	1	2	3	4	5
f_i	14	10	6	4	3	1

a) Determina la moda, la mediana y la media ¿Cuál de ellas representa mejor la distribución de los datos?

b) Calcula la desviación absoluta media.

c) ¿Qué porcentaje de datos se encuentran en el intervalo $(\bar{x}-2s, \bar{x}+2s)$, siendo *s* la desviación típica.

a) Para los cálculos posteriores, se añaden a la tabla las columnas de las frecuencias absolutas acumuladas y las necesarias para hallar la media y la varianza.

x_i	f_i	F_i	$f_i x_i$	$f_i x_i^2$
0	14	14	0	0
1	10	24	10	10
2	6	30	12	24
3	4	34	12	36
4	3	37	12	48
5	1	38	5	25
	38		51	143

Con los datos y resultados de la tabla, se tiene que:

La moda es $M_o = 0$, que es el número de materias suspendidas que aparece con más frecuencia.

La mediana es $M = 1$, ya que ordenadas de menor a mayor es la calificación que ocupa los lugares 19 y 20.

La media aritmética: $\bar{X} = \frac{51}{38} = 1{,}342$ **asignaturas suspensas**

El valor que mejor representa la distribución, dada su asimetría, es la mediana, aunque en este caso el valor de la media no está muy alejado de la mediana.

b) La desviación absoluta media se calcula a partir de su definición:

$$D_x = \frac{1}{38}\sum_{i=1}^{6} f_i\left|x_i - \bar{X}\right| = \frac{44{,}421}{38} = 1{,}169\ .$$

c) Se obtiene la desviación típica de la distribución a partir de la última columna de la tabla:

$$s^2 = \frac{143}{38} - 1{,}342^2 = 1{,}9619 \Rightarrow s = \sqrt{1{,}9619} = 1{,}4007$$

d) De modo que el intervalo:

$$(\bar{X}-2s,\ \bar{X}+2s) = (1{,}342-2{,}8014;\ 1{,}342+2{,}8014) = (-1{,}4594;\ 4{,}1434)$$

incluye 37 observaciones, que representan el 97,37 % de las mismas.

39. En los diez primeros partidos de liga, los goles marcados por dos equipos de fútbol rivales fueron:

Equipo *A*: 0, 2, 1, 5, 1, 4, 3, 0, 2, 1

Equipo *B*: 2, 1, 0, 0, 6, 1, 4, 2, 1, 1

a) Determina la media, moda y mediana del número de goles marcados por cada equipo y compáralas.

b) Halla los cuartiles del número de goles de cada equipo.

c) Dibuja el diagrama de caja de las dos distribuciones. Compáralos.

d) Calcula el coeficiente de variación. ¿Cuál de los dos equipos muestra mayor regularidad?

a) De los datos se obtiene que la media de goles de cada equipo es

$$\bar{X}_A = \frac{0+2+1+5+1+4+3+0+2+1}{10} = 1{,}9 \qquad \bar{X}_B = \frac{2+1+0+0+6+1+4+2+1+1}{10} = 1{,}8$$

Ambos equipos presentan la misma moda $Mo_A = Mo_B = 1$. En el caso del equipo *A*, marcó esa cifra de goles en tres partidos y en el caso del equipo B en cuatro partidos.

Si se ordena de menor a mayor la distribución de los goles en ambos equipos:

Equipo *A*: 0, 0, 1, 1, 1, 2, 2, 3, 4, 5 Equipo *B*: 0, 0, 1, 1, 1, 1, 2, 2, 4, 6

Al tener 10 observaciones (número par), la mediana es cualquier valor comprendido entre los que ocupen las posiciones 5 y 6. Luego la mediana del equipo A es 1 ó 2 (cualquier valor entre estos dos puede considerarse mediana), mientras que la mediana del equipo B es 1. Si se comparan estos valores, se deduce que ambos equipos presentan cifras muy similares.

b) Se deben calcular, para ambas distribuciones, los cuartiles primero y tercero. El segundo cuartil, la mediana, ya ha sido calculado.

Cuartiles de la distribución del equipo *A*: $Q_1 = 1$, $Q_2 = 1{,}5$ y $Q_3 = 3$ (Para la mediana se ha tomado el punto medio entre 1 y 2).

Luego el rango intercuartílico del equipo *A* es $RIC_A = 3 - 1 = 2$.

Análogamente para la distribución de goles del equipo *B*:

Cuartiles de la distribución del equipo *B*: $Q_1 = 1$, $Q_2 = 1$ y $Q_3 = 2$.

El rango intercuartílico del equipo *B* es $RIC_B = 3 - 1 = 2$.

c) De los resultado del apartado b) resulta que para el equipo *A*, los extremos del diagrama (los bigotes) son:

$Q_3 + 1{,}5\ RIC_A = 3 + 1{,}5 \cdot 2 = 6$. De donde LS_A (mayor de los valores que son menores o iguales que $Q_3 + 1{,}5\ RIC_A$ es 5.

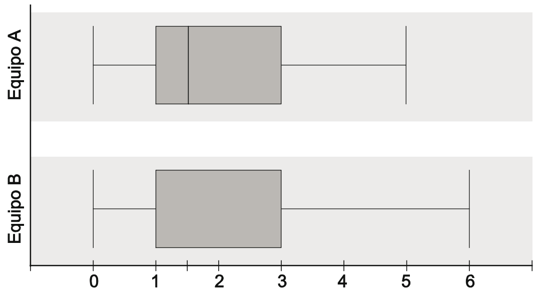
$Q_1 - 1{,}5\ RIC_A = 1 - 1{,}5 \cdot 2 = -2$. De donde LI_A (menor de los valores que son mayores o iguales que $Q_1 - 1{,}5\ RIC_B = 0$.

Y para el equipo B:

$Q_3 + 1{,}5$ RIC $= 3 + 1{,}5 \cdot 2 = 6$. De donde LS_B (mayor de los valores que son menores o iguales que $Q_3 + 1{,}5$ RIC) $= 6$.

$Q_1 - 1{,}5$ RIC $= 1 - 1{,}5 \cdot 2 = -2$. De donde LI_B (menor de los valores que son mayores o iguales que $Q_1 - 1{,}5$ RIC) $= 0$.

Los diagramas de cajas quedan, entonces, así:



d) Para calcular los coeficientes de variación necesitamos las desviaciones típicas:

$$s_A^2 = \frac{(0-1{,}9)^2 \cdot 2 + (1-1{,}9)^2 \cdot 3 + (2-1{,}9)^2 \cdot 2 + (3-1{,}9)^2 + (4-1{,}9)^2 + (5-1{,}9)^2}{10} = 2{,}49 \Rightarrow s_A = 1{,}578$$

$$s_B^2 = \frac{(0-1{,}8)^2 \cdot 2 + (1-1{,}8)^2 \cdot 4 + (2-1{,}8)^2 \cdot 2 + (4-1{,}8)^2 + (6-1{,}8)^2}{10} = 3{,}16 \Rightarrow s_B = 1{,}778$$

De donde:

$$CV_A = \frac{1{,}578}{1{,}9} = 0{,}831; \quad CV_B = \frac{1{,}778}{1{,}8} = 0{,}988$$

Parece que el equipo *A* presenta algo menos de variabilidad. Esto también se puede observar en el diagrama de cajas.

CUESTIONES.

40. El valor más pequeño observado de una variable estadística cuantitativa continua es 34,2 y el mayor 43,3. Se dispone de 110 observaciones.

Haz, de forma razonada, al menos dos propuestas para agrupar las observaciones de esta variable en clases.

Como el rango de las observaciones es 43,3 – 34,2 = 9,1; se puede proponer:

a) Dividir el recorrido de la variable en 7 clases de longitud 1,3 cada clase:

[34,2; 35,5); [35,5; 36,8); [36,8; 38,1); [38,1; 39,4); [39,4; 40,7); [40,7; 42,0) y [42,0; 43,3]

Es la propuesta más ajustada y, tal vez, la más recomendable en este caso.

b) Si se propone utilizar 10 clases (por tener más de 100 observaciones), puede hacerse con clases de longitud 1, desde 34 a 44:

[34; 35); [35; 36); [36; 37); [37; 38); [38; 39); [39; 40); [40; 41); [41; 42); [42; 43) y [43; 44]

Aunque en esta propuesta los intervalos inicial y final han quedado algo desequilibrados, ya que se restó 0,2 al primero y se añadió 0,7 al último.

c) Una tercera posibilidad, con 10 intervalos de longitud 1 sería equilibrar lo que se añade y se resta al máximo y al mínimo, respectivamente, de las observaciones:

[33,8; 34,8); [34,8; 35,8); [35,8; 36,8); [36,8; 37,8); [37,8; 38,8); [38,8; 39,8); [39,8; 40,8); [40,8; 41,8); [41,8; 42,8); [42,8; 43,8)

41. Sea una variable estadística *X* que toma valores $x_1, x_2, ..., x_n$, con frecuencias $f_1, f_2, ..., f_n$ y cuya media es $\bar{X}$ Considera la variable aleatoria *Y*, de valores los de la tabla y con frecuencias absolutas $f_1, f_2, ..., f_n$ iguales a las de los valores correspondientes de *X*.

Calcula la media de *Y*.

Y	f_i
$y_1 = x_1 - \bar{X}$	f_1
$y_2 = x_2 - \bar{X}$	f_2
…	…
$y_n = x_n - \bar{X}$	f_n

Se calcula la media de *Y* sustituyendo sus valores en función de *X* y utilizando que $N = \sum_i f_i$ y $\bar{X} = \frac{1}{N}\sum_i f_i x_i$, se obtiene

$$\bar{Y} = \frac{1}{N}\sum_i f_i y_i = \frac{1}{N}\sum_i f_i\left(x_i - \bar{X}\right) = \frac{1}{N}\sum_i f_i x_i - \frac{\bar{X}}{N}\sum_i f_i = \frac{1}{N}\sum_i f_i x_i - \frac{\bar{X}}{\cancel{N}}\cancel{N} = \frac{1}{N}\sum_i f_i x_i - \bar{X} = 0$$

Por tanto, $\bar{Y} = 0$.

42. De una característica *X* se obtienen las observaciones 1, 5, 7, 3, 7, 11, 1 y 3

a) Calcula la media y la varianza de X.

b) Multiplica los valores anteriores por 2 y obtén ahora su media y su varianza.

a) La media es $\bar{X} = \dfrac{1+5+7+3+7+11+1+3}{8} = \dfrac{38}{8} = 4{,}75$ y la varianza:

$$s_x^2 = \frac{1^2+5^2+7^2+3^2+7^2+11^2+1^2+3^2}{8} - 4{,}75^2 = 33 - 4{,}75^2 = 10{,}4375$$

b) Para la variable $Y = 2X$ se obtienen los valores 2, 10, 14, 6, 14, 22, 2, 6.

Entonces: $\bar{Y} = \dfrac{2+10+14+6+14+22+2+6}{8} = \dfrac{76}{8} = 9{,}5$; es decir, $\bar{Y} = 2\bar{X}$.

La varianza de *Y*: $s_Y^2 = \dfrac{2^2 \cdot 2 + 6^2 \cdot 2 + 10^2 + 14^2 \cdot 2 + 22^2}{8} - 9{,}5^2 = 41{,}75$; es decir, $s_Y^2 = 2^2 s_X^2$.

43. Sea *X* una variable estadística cuya media aritmética es $\bar{X}$, y su varianza s^2.

A los valores de la variable *X* se les suma la constante *k*, obteniéndose una nueva variable $Y = X + k$ ¿Cuáles serán la media y la varianza de *Y*?

Si *X* es una variable estadística cuya media aritmética es $\bar{X}$, la media aritmética de la variable $Y = X + k$ es:

$$\bar{Y} = \frac{1}{N}\sum_1^n f_i \cdot y_i = \frac{1}{N}\sum_1^n f_i \cdot (x_i + k) = \frac{1}{N}\sum_1^n f_i \cdot x_i + \frac{k}{N}\sum_1^n f_i k = \bar{X} + k$$

y la varianza es: $s_Y^2 = \dfrac{1}{N}\sum_1^n f_i \cdot (y_i - \bar{Y})^2 = \dfrac{1}{N}\sum_1^n f_i \cdot (x_i + k - \bar{X} - k)^2 = \dfrac{1}{N}\sum_1^n f_i \cdot (x_i - \bar{X})^2 = s_X^2$.

Es decir, la varianza no cambia pero la media se ve trasladada *k* unidades, por tanto la traslación de los datos no influye en su dispersión alrededor de la media.

44. Sea *X* una variable estadística cuya media aritmética es $\bar{X}$, y su varianza s^2.

Considera ahora la variable *Y*, resultado de multiplicar los valores de *X* por una constante *c*. Calcula la media y la varianza de *Y*.

Si *X* es una variable estadística cuya media aritmética es $\bar{X}$, la media aritmética de la variable $Y = c - X$ es:

$$\bar{Y} = \frac{1}{N}\sum_1^n f_i \cdot y_i = \frac{1}{N}\sum_1^n f_i \cdot c \cdot x_i = \frac{c}{N}\sum_1^n f_i \cdot x_i = c \cdot \bar{X}$$

y la varianza es: $s_Y^2 = \dfrac{1}{N}\sum_1^n f_i \cdot (y_i - \bar{Y})^2 = \dfrac{1}{N}\sum_1^n f_i \cdot (c \cdot x_i - c \cdot \bar{X})^2 = \dfrac{c^2}{N}\sum_1^n f_i \cdot (x_i - \bar{X})^2 = c^2 \cdot s_X^2$

Es decir, la media queda multiplicada por *c* y la varianza por c^2, por tanto la dilatación de los datos influye en su dispersión alrededor de la media.

45. Si la media, obtenida a partir de una muestra, de una variable estadística cuantitativa es 3 y su varianza 25, ¿se puede decir que la media es representativa del conjunto de datos?

Para estudiar la variabilidad hay que calcular el coeficiente de variación. Si la variable es *X* se tiene:

$$CV(X) = \frac{s}{|\bar{X}|} = \frac{\sqrt{25}}{3} = \frac{5}{3} = 1{,}667$$

Esto indica que los datos no son muy homogéneos y, por tanto, es fácil que la media resulte poco representativa.

46. Si a los valores obtenidos de una variable estadística cuantitativa se les multiplica por una constante k, ¿en qué medida cambia el coeficiente de variación? Nota: prueba con los valores 1, 2, 5, 7 y 8; y con $k = 3$.

Si X toma los valores 1, 2, 5, 7 y 8, entonces $Y = 3X$ tomará los valores 3, 6, 15, 21 y 24, de donde:

$$\bar{X} = \frac{1+2+5+7+8}{5} = 4{,}6 \quad ; \quad \bar{Y} = \frac{3+6+15+21+24}{5} = 13{,}8 \Rightarrow \bar{Y} = 3 \cdot \bar{X}$$

Las varianzas y desviaciones típicas son:

$$s_X^2 = \frac{1^2+2^2+5^2+7^2+8^2}{5} - 4{,}6^2 = 7{,}44 \Rightarrow s_X = 2{,}73 \text{ y } s_Y^2 = \frac{3^2+6^2+15^2+21^2+64^{2'}}{5} - 13{,}8^2 = 66{,}96 \Rightarrow s_Y = 8{,}183\,.$$

Por tanto los coeficientes de variación son: $CV_X = \frac{s_x}{\bar{X}} = \frac{2{,}73}{4{,}6} = 0{,}593$ y $CV_Y = \frac{s_y}{\bar{Y}} = \frac{8{,}183}{13{,}8} = 0{,}593$.

Es decir, los coeficientes de variación no cambian ante una dilatación de los datos.

Para el caso general, se pueden aplicar los resultados obtenidos en la cuestión 44. Así, al multiplicar una variable estadística X por una constante k, se obtiene la variable $Y = kX$, cuya media y desviación típica se relacionan con la de X en la forma: $\bar{Y} = k\bar{X}$ y $s_Y = k^2 s_x$. Por tanto, los respectivos coeficientes de variación se relacionan en la forma: $\frac{CV_Y}{CV_X} = \frac{s_Y}{\bar{Y}} : \frac{s_x}{\bar{X}} = \frac{s_Y \cdot \bar{X}}{s_X \cdot \bar{Y}} = \frac{ks_X \cdot \bar{X}}{s_X \cdot k\bar{Y}} = 1 \Rightarrow CV_Y = CV_X$, que confirma el resultado obtenido anteriormente.

PROBLEMAS.

47. En la encuesta del CIS (centro de Investigaciones Sociológicas) correspondiente al primer trimestre de 2013 sobre el nivel de estudios, se obtuvo

Nivel de estudios	f_i
Sin estudios	147
Primaria	1055
Secundaria	319
FP	451
Universitarios Medios	210
Universitarios Superiores	293
No contesta	8
	2483

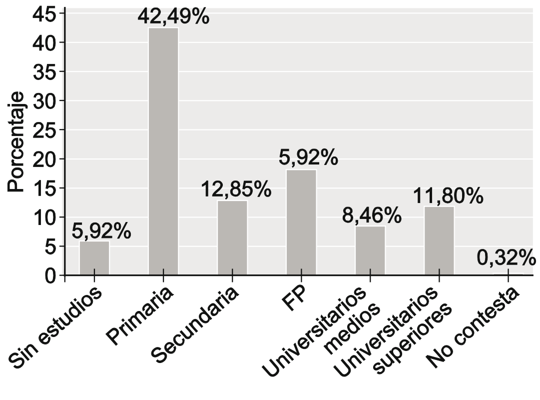
a) Calcula la distribución de frecuencias relativas y de porcentajes.

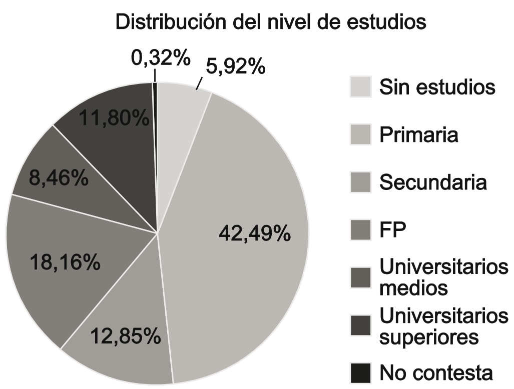
b) Utiliza el gráfico más adecuado para representar la distribución de los porcentajes.

a) A la tabla se le añaden las columnas de las frecuencias relativas y de los porcentajes:

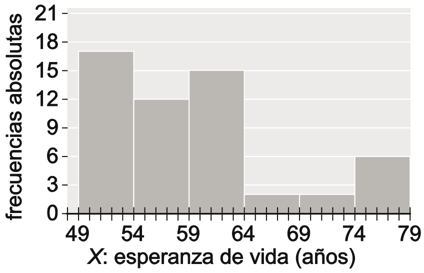
Nivel de estudios	f_i	h_i	Porcentajes
Sin estudios	147	0,0592	5,92%
Primaria	1055	0,4249	42,49%
Secundaria	319	0,1285	12,85%
FP	451	0,1816	18,16%
Universitarios Medios	210	0,0846	8,46%
Universitarios Superiores	293	0,1180	11,80%
No contesta	8	0,0032	0,32%
	2483	1	100%

b) Se propone un gráfico de barras o un diagrama de sectores:





48. La esperanza de vida, con base en 2012, en los países de África se muestra en el siguiente histograma con los países agrupados por intervalos de edad (Fuente: *CIA WorldFactbook*).



a) Construye la tabla de frecuencias.

b) Dibuja el polígono de frecuencias absolutas acumuladas.

c) Calcula la media, mediana y la moda.

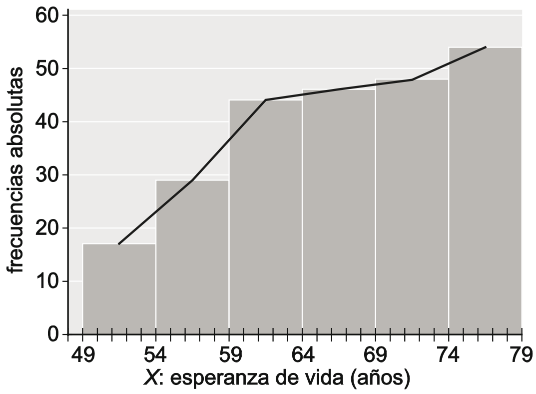
d) Determina la varianza y las desviaciones típica y absoluta media. Compara estas dos últimas.

a) A partir del histograma se recupera y amplía la tabla de frecuencias para los cálculos posteriores:

La última columna se ha escrito aquí por comodidad, aunque para su cálculo se utiliza la media hallada en el apartado c).

Clases	f_i	F_i	x_i	$f_i x_i$	$f_i x_i^2$	$f_i \left\| x_i - \bar{X} \right\|$
[49,54)	17	17	51,5	875,5	45 088,25	135,370
[54,59)	12	29	56,5	678	38 307	35,556
[59,64)	15	44	61,5	922,5	56 733,75	30,556
[64,69)	2	46	66,5	133	8844,5	14,074
[69,74)	2	48	71,5	143	10 224,5	24,074
[74,79]	6	54	76,5	459	35 113,5	102,222
	54			3211	194 311,5	341,852

b) El polígono de frecuencias absolutas acumuladas es:



c) Para los cálculos recurrimos a la tabla:

Media aritmética: $\bar{X} = \frac{3211}{54} = 59{,}46$ *años*

La mediana: el 50% de las 54 observaciones son 27. Ordenadas de menor a mayor y mirando en la tabla la columna de las frecuencias absolutas acumuladas (F_i), la observación que ocupa el lugar 27 se encuentra en el intervalo [54,59), cuya longitud es 5 y que contiene 12 observaciones y antes de este intervalo se han acumulado 17 observaciones. Luego la mediana es: $M = 54 + \frac{(27-17)5}{12} = 58{,}17$ *años*.

El intervalo modal es [49, 54), que contiene 17 observaciones.

d) Con ayuda de la tabla del apartado anterior se obtienen la varianza y las desviación típica y absoluta media.

$$s^2 = \frac{194\,311{,}5}{54} - 59{,}46^2 = 62{,}87 \Rightarrow s = 7{,}93 \text{ años}; \quad D_X = \frac{341{,}852}{54} = 6{,}33 \text{ años.}$$

Se observa que la desviación típica tiene un valor superior al de la desviación absoluta media.

49. En la siguiente tabla se han agrupado las provincias españolas según su tasa de natalidad (*X*: nacidos por 1000 habitantes) en el año 2011 (Fuente INE).

X	f_i
[5,7; 8)	8
[8; 10,3)	24
[10,3; 12,6)	18
[12,6; 14,9)	0
[14,9; 17,2)	1
[17,2; 19,5]	1

a) Halla la tasa media de nacimientos por provincia en el año 2011.

b) Determina la moda y la mediana.

c) Representa los datos.

d) Calcula la varianza y la desviación típica.

e) Determina el coeficiente de variación.

a) Se añaden a la tabla las columnas necesarias para efectuar los cálculos que se piden

X	f_i	F_i	x_i	$f_i x_i$	$f_i x_i^2$
[5,7; 8)	8	8	6,85	54,8	375,38
[8; 10,3)	24	32	9,15	219,6	2009,34
[10,3; 12,6)	18	50	11,45	206,1	2359,845
[12,6; 14,9)	0	50	13,75	0	0
[14,9; 17,2)	1	51	16,05	16,05	257,6025
[17,2; 19,5]	1	52	18,35	18,35	336,7225
	52			514,9	5338,89

La tasa media de natalidad en el año 2011 fue

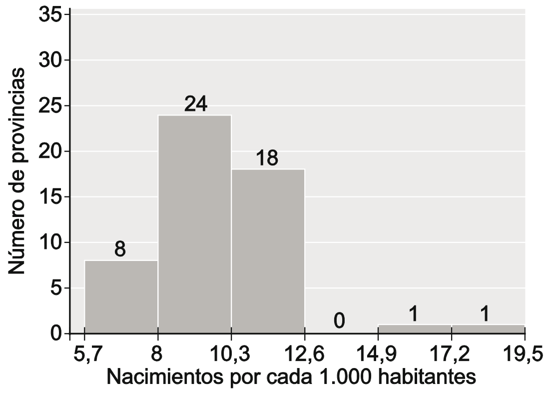
$$\bar{X} = \frac{514{,}9}{52} = 9{,}902 \text{ nacimientos por cada 1000 habitantes}$$

b) El intervalo modal es [8; 10,3) y contiene 24 observaciones.

Mediana: el 50% del total 52 son 26 observaciones. De la columna de las frecuencias acumuladas se observa que la observación número 26 (una vez ordenadas de menor a mayor) se encuentra en el intervalo [8; 10,3), de longitud 2,3 y que contiene 24 observaciones. Antes de este intervalo se han acumulado 8 observaciones. Luego

$$M = 8 + \frac{(26-8)2{,}3}{24} = 9{,}725 \text{ nacimientos por cada 1000 habitantes}$$

c) El histograma de frecuencias absolutas es:



d) La varianza se obtiene a partir de los resultados de la tabla:

$$s^2 = \frac{5338{,}89}{52} - 9{,}902^2 = 4{,}6229 \Rightarrow s = \sqrt{4{,}6229} = 2{,}1501$$

e) El coeficiente de variación es:

$$CV = \frac{s}{|\bar{X}|} = \frac{2{,}1501}{9{,}902} = 0{,}2171$$

50. La tabla siguiente incluye los porcentajes de gastos de administración (*X*) calculados sobre el total de primas recaudadas (f_i), en millones de euros, por cinco empresas en seguros de hogar.

Empresa	X	f_i
A	11	220
B	15	130
C	14	145
D	12	180
E	11	150

a) Determina el porcentaje medio de gastos.

b) Si se supone que estas cinco empresas cubren todo el mercado de seguros de hogar, calcula el coeficiente de variación y haz una valoración del resultado.

a) Se completa la tabla con las columnas necesarias para el cálculo del porcentaje medio y la varianza de los porcentajes:

Empresa	X	f_i	f_ix_i	$f_ix_i^2$
A	11	220	2420	26 620
B	15	130	1950	29 250
C	14	145	2030	28 420
D	12	180	2160	25 920
E	11	150	1650	18 150
		825	10 210	128 360

El porcentaje medio de gastos de administración se obtiene a partir de los cálculos de la tabla:

$$\bar{X} = \frac{10210}{825} = 12{,}376\ \%$$

b) Para calcular el coeficiente de variación se necesita la desviación típica:

$$s^2 = \frac{128360}{825} - 12{,}376^2 = 2{,}4285 \Rightarrow s = 1{,}5584\% \text{ y de aquí resulta } CV = \frac{1{,}5584}{12{,}376} = 0{,}1259$$

Que informa de la variabilidad en el porcentaje de gastos de administración cobrados por las distintas empresas y, en este caso puede verse que esa variabilidad es relativamente pequeña.

51. Los tiempos de vida (*X,* en horas) de 500 bombillas de una cierta marca se han agrupado en la tabla.

X	f_i
[0, 500)	15
[500, 1000)	56
[1000, 1500)	178
[1500, 2000)	115
[2000, 2500)	84
[2500, 3000)	52

a) Representa el histograma correspondiente, junto con el polígono de frecuencias.

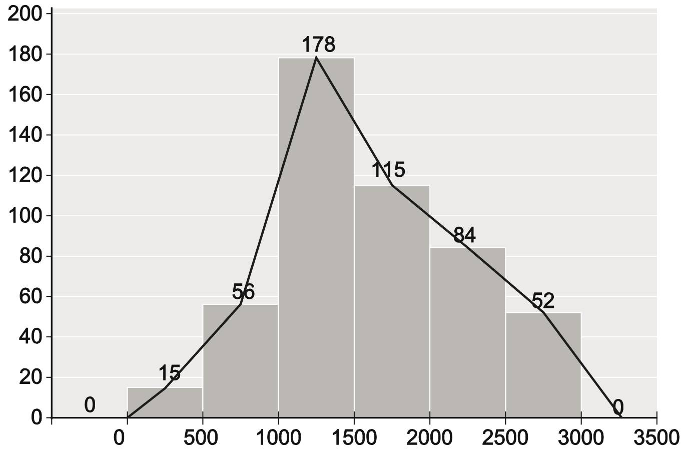
b) Calcula la media, la mediana y el intervalo modal.

c) Determina la desviación absoluta media, la varianza y la desviación típica.

d) Halla los cuartiles.

e) Estudia la variabilidad de la distribución de frecuencias.

a) El histograma y el polígono de frecuencias son:



b) El intervalo modal es el [1000, 1500), que contiene 178 observaciones

c)

Clases	f_i	F_i	x_i	$f_i x_i$	$f_i x_i^2$	$f_i\left\|x_i - \bar{X}\right\|$
[0, 500)	15	15	250	3750	937 500	20 295
[500, 1000)	56	71	750	42 000	31 500 000	47 768
[1000, 1500)	178	249	1250	222 500	278 125 000	62 834
[1500, 2000)	115	364	1750	201 250	352 187 500	16 905
[2000, 2500)	84	448	2250	189 000	425 250 000	54 348
[2500, 3000]	52	500	2750	143 000	393 250 000	59 644
	500			801 500	1 481 250 000	261 794

Para el cálculo de la media y la mediana se añaden a la tabla las columnas necesarias. En la tabla se incluye también la columna necesaria para el cálculo de la varianza.

El tiempo medio de vida de las bombillas es: $\bar{X} = \frac{801500}{500} = 1603 \text{ horas}$.

Mediana del tiempo de vida de las bombillas. El 50 % de 500 es 250 y la observación 250, una vez ordenados los datos queda incluida en el intervalo [1500, 2000), de longitud 500 y que incluye 115 observaciones. Hasta este intervalo se han acumulado 249 observaciones, luego:

$$M = 1500 + \frac{(250 - 249)500}{115} = 1504{,}3 \text{ horas}$$

d) La desviación absoluta media viene dada por: $D_x = \frac{261794}{500} = \mathbf{523{,}588\ horas}$

La varianza del tiempo de vida de las bombillas y su desviación típica son:

$$s^2 = \frac{1\,481\,250\,000}{500} - 1603^2 = 392\,891 \Rightarrow s = 626{,}81 \text{ horas.}$$

Como se puede observar, la desviación típica resulta ser más algo superior que la desviación absoluta media.

e) Para determinar los cuartiles se procede como sigue:

El 25% de 500 es 125 y la observación 125, una vez ordenados los datos queda incluida en el intervalo [1000, 1500), de longitud 500 y que incluye 178 observaciones. Hasta este intervalo se han acumulado 71 observaciones, luego:

$$Q_1 = 1000 + \frac{(125-71)500}{178} = 1151{,}69 \text{ horas}$$

Por otro lado $Q_2 = M = 1504{,}3$ **horas**.

Para el tercer cuartil se tiene en cuenta que el 75 % de 500 es 375 y la observación 375, una vez ordenados los datos queda incluida en el intervalo [2000, 2500), de longitud 500 y que incluye 84 observaciones. Hasta este intervalo se han acumulado 364 observaciones, luego:

$$Q_3 = 2000 + \frac{(375-364)500}{84} = 2065{,}48 \text{ horas}$$

f) Para estudiar la variabilidad se determina el coeficiente de variación que resulta:

$$CV = \frac{626{,}81}{1603} = 0{,}391$$

e indica una variabilidad media de los datos.

52. Las temperaturas medias (*X*) registradas a lo largo del mes de mayo en 50 ciudades se presentan agrupadas en la siguiente tabla

X	f_i
[15, 18)	7
[18, 21)	10
[21, 24)	15
[24, 27)	13
[27, 30]	5

a) Representa gráficamente los datos, mediante un histograma y añade el polígono de frecuencias.

b) Determina la media, la mediana y el intervalo modal.

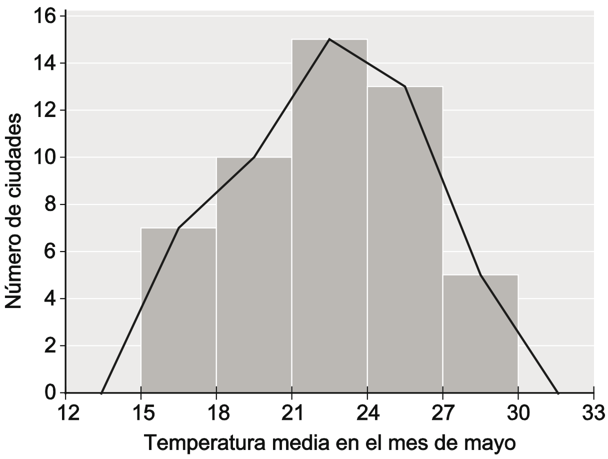
c) Halla los deciles 3 y 8.

d) Calcula las desviaciones absoluta media y la desviación típica. Compáralas.

e) Determina la proporción de variación de la distribución de las temperaturas.

f) ¿Por debajo de qué temperatura se encuentra el 35% de las ciudades con temperatura media más baja?

a) El histograma y el polígono de frecuencias son:



b) Para los cálculos de este apartado y siguientes se amplía la tabla con las columnas necesarias

X	f_i	F_i	x_i	$f_i x_i$	$f_i\lvert x_i - \bar{X}\rvert$	$f_i x_i^2$
[15, 18)	7	7	16,5	115,5	41,58	1905,75
[18, 21)	10	17	19,5	195	29,4	3802,5
[21, 24)	15	32	22,5	337,5	0,9	7593,75
[24, 27)	13	45	25,5	331,5	39,78	8453,25
[27, 30]	5	50	28,5	142,5	30,3	4061,25
a	50			1122	141,96	25 816,5

que la temperatura media es $\bar{X} = \dfrac{1122}{50} = 22{,}44\ °C$

Temperatura mediana: el 50% de las 50 observaciones es 25. La observación que ocupa el lugar 25, ordenadas de menor a mayor, está incluida en el intervalo [21, 24), de longitud 3 y que contiene 15 observaciones. Antes de este intervalo se acumulan 17 observaciones, luego la mediana es:

$$M = 21 + \frac{(25-17)\,3}{15} = 22{,}6\ °C$$

El intervalo modal es [21, 24), ya que es el que contiene mayor número de observaciones,15.

c) Para calcular el decil 3, D_3, se observa que el 30 % de las 50 observaciones es 15. La observación que ocupa el lugar 15, ordenadas de menor a mayor, está incluida en el intervalo [18, 21), de longitud 3 y que contiene 10 observaciones. Antes de este intervalo se acumulan 10 observaciones, luego $D_3 = 18 + \dfrac{(15-7)3}{10} = 20{,}4$ ºC.

Razonando de la misma forma se calcula el decil 8: $D_8 = 24 + \dfrac{(40-32)3}{13} = 25{,}85$ ºC

d) La desviación absoluta media se obtiene a partir de los cálculos de la tabla: $D_X = \dfrac{141{,}96}{50} = 2{,}8392$ °C

Y la varianza y desviación típica: $s^2 = \dfrac{25816{,}5}{50} - 22{,}442^2 = 12{,}7764 \Rightarrow s = \sqrt{12{,}7764} = 3{,}5744\ °C$.

Se observa que la desviación típica es mayor que la desviación absoluta media.

e) El coeficiente de variación de la distribución de temperaturas es $CV = \dfrac{3{,}5744}{22{,}44} = 0{,}1593$.

f) Para responder a esta cuestión, hay que calcular el percentil 35. Como el 35% de las 50 observaciones es 17,5 que está por encima de 17. La observación que ocupa el lugar 18, una vez ordenadas de menor a mayor, está incluida en el intervalo [21, 24), de longitud 3 y que contiene 15 observaciones. Antes de este intervalo se acumulan 17 observaciones, luego este percentil es: $p_{35} = 21 + \dfrac{(17{,}5-17)\,3}{15} = 21{,}1\ °C$.

De donde el 35 % de las ciudades con temperatura más baja no superaron los 21,1 ºC.

53. De una muestra de 100 recién nacidos en una clínica de maternidad, se ha obtenido la siguiente tabla de pesos en kg (*X*) para bebés de entre 3 y 4 kg de peso.

X	f_i
[3,0; 3,2)	27
[3,2; 3,4)	38
[3,4; 3,6)	16
[3,6; 3,8)	10
[3,8; 4,0]	9

a) Dibuja el histograma de la distribución de frecuencias relativas.

b) Encuentra la media, la mediana y el intervalo modal.

c) Determina la varianza y la desviación típica.

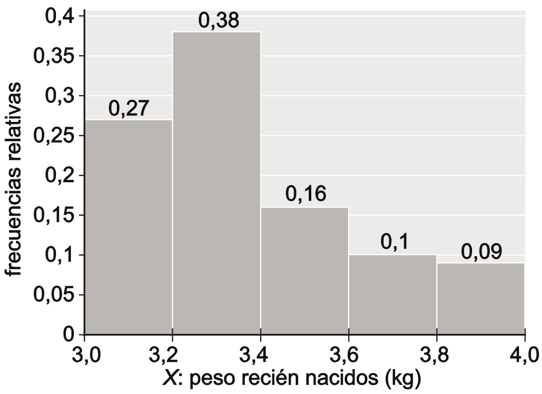
d) ¿Es simétrica la distribución? Justifica la respuesta.

e) Calcula los percentiles 5 y 95. ¿Cuántos bebés se encuentran por encima del 90 % del peso?

Se amplía la tabla con las columnas necesarias para responder a los apartados siguientes

Clases	f_i	F_i	h_i	x_i	f_ix_i	$f_ix_i^2$
[3,0; 3,2)	27	27	0,27	3,1	83,7	259,47
[3,2; 3,4)	38	65	0,38	3,3	125,4	413,82
[3,4; 3,6)	16	81	0,16	3,5	56	196
[3,6; 3,8)	10	91	0,1	3,7	37	136,9
[3,8; 4,0]	9	100	0,09	3,9	35,1	136,89
	100		1		337,2	1143,08

a) El histograma de frecuencias relativas se construye con la columna de las h_i:



b) El intervalo modal es [3,2; 3,4), ya que contiene más observaciones (todos los intervalos son de igual amplitud).

El peso medio de los bebés es $\bar{X} = \frac{337,2}{100} = 3,372\ kg$

La mediana del peso de los bebés se encuentra en el intervalo [3,2; 3,4), puesto que una vez ordenadas las observaciones, buscamos la que ocupa el lugar 50 (50 % de 100 observaciones). En intervalo tiene longitud 0,2 kg e incluye 38 observaciones. Además antes de llegar a este intervalo se acumulan 27 observaciones. Luego:

$$M = 3,2 + \frac{(50-27)0,3}{38} = 3,382\ kg$$

c) La varianza y la desviación típica vienen dadas por $s^2 = \frac{1143,08}{100} - 3,372^2 = 0,0604 \Rightarrow s = \sqrt{0,0604} = 0,2458$.

d) A la vista del histograma se puede afirmar que la distribución no es simétrica, puesto que las observaciones no se reparten de forma equilibrada a la izquierda y a la derecha de la media..

e) Para calcular los percentiles 5 y 95 se sigue el mismo procedimiento que con la mediana:

El 5 % de 100 observaciones es 5, de modo que el percentil 5 se encuentra en el primer intervalo [3,0; 3,2), que tiene longitud 0,2 y contiene 27 observaciones, luego $p_5 = 3,0 + \frac{5 \cdot 0,2}{27} = 3,037\ kg$.

El 95 % de 100 es 95, de manera que el percentil 95 se encuentra en el intervalo [3,8; 4,0), de longitud 0,2 y que contiene 9 observaciones. Como antes de este intervalo se acumulan 91 observaciones, resulta

$$p_{95} = 3,8 + \frac{(95-91) \cdot 0,2}{9} = 3,889\ kg.$$

Por encima del 90 % del peso se encuentra el 10 % de los bebés, es decir 10 bebés.

54. Dado el siguiente conjunto de datos, relativos a los vecinos de un inmueble mayores de 40 años.

61	**69**	**42**	**49**	**62**	**66**	**41**	**48**	**43**	**54**
51	**43**	**49**	**42**	**43**	**53**	**44**	**41**	**51**	**51**
54	**59**	**56**	**58**	**64**	**63**	**46**	**52**	**42**	**66**
69	**57**	**48**	**44**	**67**	**69**	**58**	**54**	**66**	**65**
42	**57**	**55**	**53**	**50**	**48**	**63**	**68**	**41**	**70**

a) Para los datos sin agrupar dibuja el diagrama de caja

b) Agrupa los datos en intervalos de longitud 5, por un lado, y de longitud 10 por otro.

c) Representa gráficamente las dos distribuciones

d) Calcula la media, la mediana y el intervalo modal en cada una de las dos distribuciones.

e) Calcula el coeficiente de variación de ambas distribuciones y comenta las diferencias encontradas.

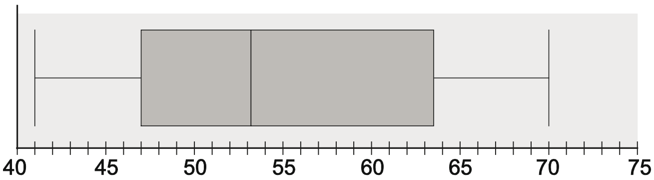
a) Para la representación del diagrama de cajas, ordenamos los 50 datos de menor a mayor.

El valor mínimo es 41 y el máximo es 70.

Como el 25 % de 50 es 12,5; el primer cuartil vendrá dado por la media de los que ocupen las posiciones 12 y 13, estos son los valores 46 y 48 y por tanto $Q_1 = 47$

Del mismo modo, como el 75 % de 50 es 37,5; el tercer cuartil vendrá dado por la media de los que ocupen las posiciones 37 y 38, estos son los valores 62 y 63 y por tanto $Q_3 = 62,5$.

La mediana viene dada por el valor intermedio de los que ocupan las posiciones 25 y 26, que son 53 y 54, de donde $M = 53,5$.



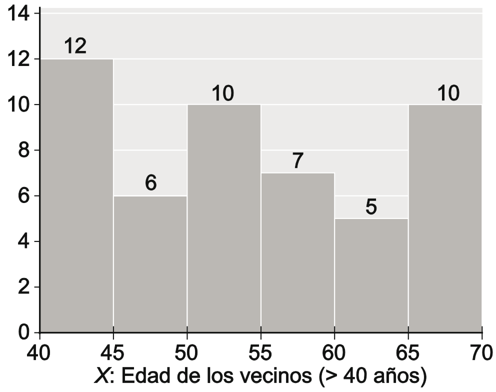
b) Como el máximo de las observaciones es 70 y el mínimo 41, elegimos seis intervalos de longitud 5 y 3 intervalos de longitud 10, empezando en 40 y finalizando en 70. Efectuado el recuento en ambos casos, resulta:

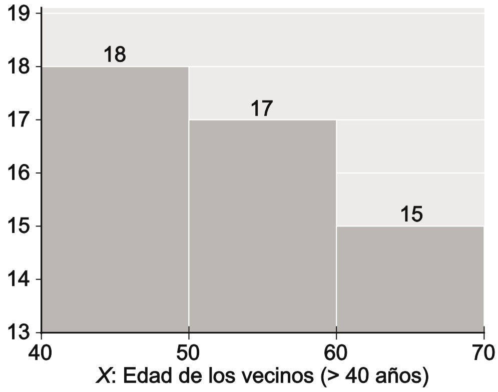
Clases	f_i	x_i	f_ix_i	f_ix_i2	F_i
[40, 45)	12	42,5	510	21 675	12
[45, 50)	6	47,5	285	13 537,5	18
[50, 55)	10	52,5	525	27 562,5	28
[55, 60)	7	57,5	402,5	23 143,75	35
[60, 65)	5	62,5	312,5	19 531,25	40
[65, 70]	10	67,5	675	45 562,5	50
	50		2710	151 012,5	

Clases	f_i	x_i	f_ix_i	$f_ix_i^2$	F_i
[40, 50)	18	45	810	36450	18
[50, 60)	17	55	935	51425	35
[60, 70]	15	65	975	63375	50
	50		2720	151250	

A las dos tablas se les han añadido las columnas necesarias para los cálculos posteriores.

c) Se dibuja el histograma correspondiente a las dos distribuciones:





Que son dos formas diferentes de representar la misma distribución del número de vecinos mayores de 40 años. La apariencia, como se ve en los histogramas, es claramente distinta.

d) Para cada distribución se calcula la media, la mediana y el intervalo modal

De la primera tabla, la de las clases de longitud 5, se obtiene que la media aritmética es

$$\bar{X} = \frac{2710}{50} = 54{,}2 \text{ años}$$

La mediana se encuentra en el intervalo [50,55), de longitud 5 y que contiene 10 observaciones. Antes de este intervalo se ha acumulado 18 observaciones, luego la mediana es $M = 50 + \frac{(25-18)5}{10} = 53{,}5$ años muy próxima a la media.

El intervalo modal en este caso es [40,45), ya que es el que contiene el mayor número de observaciones (12).

En el caso de la segunda tabla, donde se han agrupado los datos en clases de longitud 10, se tiene que la media aritmética es : $\bar{X} = \frac{2720}{50} = 54{,}4$ *años*, muy similar a la obtenida con las clases de longitud 5.

La mediana se encuentra en el intervalo [50, 60), de longitud 10 y que contiene 17 observaciones. Hasta llegar a este intervalo se acumulan 18 observaciones, de modo que ahora la mediana es $M = 50 + \frac{(25-18)10}{17} = 54{,}12$ años, también cercana a la mediana obtenida con clases de longitud 5.

En cuanto al intervalo modal, en este caso es [40, 50), que incluye el intervalo modal anterior.

e) Se debe obtener la desviación típica de cada una de las agrupaciones:

Para las clases de longitud 5: $s_x^2 = \frac{151012{,}5}{50} - 54{,}2^2 = 82{,}61 \Rightarrow s_x = \sqrt{82{,}61} = 9{,}089$ *años*

Para las clases de longitud 10: $s_y^2 = \frac{151250}{50} - 54{,}4^2 = 65{,}64 \Rightarrow s_y = \sqrt{65{,}64} = 8{,}102$ *años*

Entonces, $CV(x) = \frac{9{,}089}{54{,}2} = 0{,}1677$ $\qquad CV(y) = \frac{8{,}102}{54{,}4} = 0{,}1489$

La agrupación en clases de longitud 10 presenta menor variabilidad que la agrupación en clases de longitud 5, como era de esperar, puesto que en el primer caso, en realidad solo se eligen 3 valores distintos para realizar los cálculos (las marcas de clase de las tres clases), mientras que en el segundo caso son 6 los valores elegidos.

55. La siguiente tabla recoge la edad, agrupada en intervalos, de una muestra de 1000 conductores de 14 o más años involucrados en accidentes con víctimas en vía urbana.

Edad	f_i
[14, 18)	16
[18, 25)	149
[25, 45)	526
[45, 65)	255
[65, 75)	36
[75, 90]	18

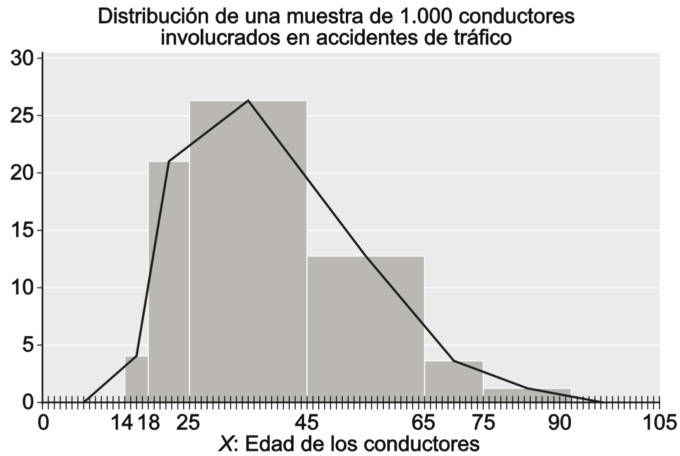
a) Representa la distribución mediante el diagrama de barras y el polígono de frecuencias absolutas.

b) Calcula la media, la mediana y las desviaciones típica y absoluta media.

c) Calcula el coeficiente de variación.

d) Suponiendo que la muestra es representativa de la población objeto de estudio, ¿Qué porcentaje de la misma se encuentra en el intervalo $\left(\bar{x}-2s, \bar{x}+2s\right)$.

a) El histograma y el polígono de frecuencias con el número de conductores por intervalos de edad de la muestra de 1000 conductores es



b) Se amplía la tabla con las columnas necesarias para los cálculos de los dos apartados siguientes:

Edad	f_i	F_i	x_i	$f_i x_i$	$f_i x_i^2$	$f_i\left\|x_i-\bar{X}\right\|$
[14, 18)	16	16	16	256	4096	382,4
[18, 25)	149	165	21,5	3203,5	68 875,25	2741,6
[25, 45)	526	691	35	18 410	644 350	2577,4
[45, 65)	255	946	55	14 025	771 375	3850,5
[65, 75)	36	982	70	2520	176 400	1083,6
[75, 90]	18	1000	82,5	1485	122 512,5	766,8
	1000			39 899,5	1 787 608,8	11 402,3

La media de la edad de los conductores involucrados en accidentes de tráfico es $\bar{X}=\frac{39899{,}5}{1000}=39{,}9$ *años*.

Es decir, aproximadamente 40 años de edad.

El 50 % de 1000 observaciones es 500. La observación que ocuparía el lugar 500 una vez ordenados los datos de menor a mayor está, por tanto, en el intervalo [25, 45) según se desprende de la columna F_i de la tabla. El intervalo tiene longitud 20 y 526 conductores se encuentran en este intervalo de edad. Las observaciones acumuladas hasta llegar al intervalo [25, 45) son 165. De esta forma, la mediana es:

$$M=25+\frac{(500-165)\cdot 20}{526}=37{,}7 \text{ años}$$

A partir de los datos de la tabla se calculan la varianza y la desviación típica.

$$s^2 = \frac{1787608.8}{1000} - 39{,}9^2 = 195{,}639 \Rightarrow s = 13{,}987 \text{ años}$$

La desviación absoluta media viene dada por:

$$D_x = \frac{1402{,}3}{1000} = 1{,}4023$$

El coeficiente de variación es una medida de la variabilidad de los datos observados:

$$CV = \frac{13{,}987}{39{,}9} = 0{,}3506$$

c) El intervalo $\left(\bar{X} - 2s, \bar{X} + 2s\right) = \left(39{,}9 - 2 \cdot 13{,}987; 39{,}9 + 2 \cdot 13{,}987\right) = (11{,}93; 67{,}87)$, que contiene todas las observaciones desde 14 hasta 67,87. Antes del intervalo [65, 75) hay 946 observaciones acumuladas. El intervalo [65, 75), de longitud 10, contiene 36 observaciones. Como 67,87–65 = 2,87, resulta que, suponiendo distribuidas de manera uniforme las 36 observaciones en el intervalo, en el intervalo [65; 67,87) se incluyen $\frac{2{,}87 \cdot 36}{10} = 10{,}33$, es decir, 10 observaciones. Luego desde 14 hasta 67,87 se incluyen 946 + 10 = 956 observaciones, que representan el 95,6 % de la población estudiada.

ENTORNO MATEMÁTICO

A la búsqueda de un trabajo.

Mariela y Fernando son dos hermanos a punto de terminar sus estudios universitarios e intentar acceder al mercado laboral, o, como ellos dicen, ponerse a buscar curro.

Están preocupados por el tema, y no hacen más que leer sobre tasas de empleo, tasas de empleo por sexo, o por grupos de edad… y demás términos indescifrables que complican la información de los telediarios.

Sus amigos les dicen que no van a tener las mismas posibilidades aunque han estudiado lo mismo y han sacado notas parecidas.

Y aunque Fernando cree que eso es una leyenda urbana, Mariela piensa que su hermano es un ingenuo y que, posiblemente, sus amigos tengan razón.

Para hacer que su hermano "se caiga del guindo", le propone que investiguen un poco y ya de paso sueñen con lo que pueden llegar a ganar, y dejar así de depender de sus padres. Investigando en el Instituto Nacional de Estadística obtienen la tabla de la Tasa de Paro en España, de los últimos años (en %).

Año	Total	Hombres	Mujeres
2002	11,45	8,30	16,15
2003	11,49	8,50	15,84
2004	10,97	8,26	14,82
2005	9,15	7,14	11,99
2006	8,45	6,35	11,35
2007	8,23	6,41	10,70
2008	11,25	10,05	12,82
2009	17,86	17,65	18,13
2010	19,86	19,57	20,22
2011	21,39	21,05	21,81
2012	24,79	24,58	25.04
2013	26,10	25,60	26,67
2014	24,69	23,87	25,65

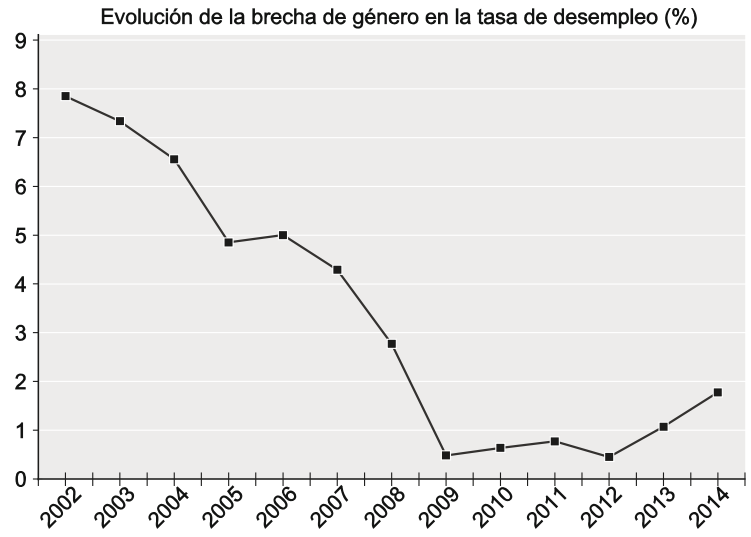
a) ¿Quién tiene razón Mariela o su hermano? Estudia cómo ha evolucionado la brecha de género (diferencia en puntos porcentuales entre la tasa de empleo de los hombres y de las mujeres) a lo largo de los últimos años. ¿Ha aumentado? ¿Ha disminuido?

b) ¿Crees que ha influido la crisis económica en la brecha de género en el empleo?

a) Ampliamos la tabla incluyendo la columna "Brecha de género" que muestra la diferencia entre las tasas de desempleo de mujeres menos la de hombres. En efecto, se observa cómo dicha brecha se redujo de forma significativa hasta 2009 y, desde entonces se ha mantenido hasta un repunte en los últimos años.

Año	Total	Hombres	Mujeres	Brecha de género (M–H)
2002	11,45	8,30	16,15	7,85
2003	11,49	8,50	15,84	7,34
2004	10,97	8,26	14,82	6,56
2005	9,15	7,14	11,99	4,85
2006	8,45	6,35	11,35	5
2007	8,23	6,41	10,70	4,29
2008	11,25	10,05	12,82	2,77
2009	17,86	17,65	18,13	0,48
2010	19,86	19,57	20,22	0,65
2011	21,39	21,05	21,81	0,76
2012	24,79	24,58	25.04	0,46
2013	26,10	25,60	26,67	1,07
2014	24,69	23,87	25,65	1,78

b) Los datos de la tabla parecen confirmar que la crisis ha vuelto a hacer crecer la brecha de género, desde 2012. Esto también se puede observar en el gráfico de la derecha.



El incendio en el pueblo de Anxo

Los campos alrededor del pueblo de Anxo en Galicia se han visto devastados por un feroz incendio. Tan grave fue, que él y sus padres tuvieron que ser evacuados durante los dos días que duraron las labores de extinción.

Anxo está enojado con el mundo porque piensa que han tenido muy mala suerte ya que les ha tenido que tocar precisamente a ellos. Y encima los rumores en el pueblo apuntan a viejas rencillas entre familias, otros dicen que no, que ha sido un accidente, y otros, los menos, que cosas de meigas… bueno, esto último se lo ha oído Anxo solo a uno y Anxo piensa que no está muy bien de la cabeza.

Por desgracia, la realidad es que Anxo y su familia no son los únicos. En las últimas décadas, los incendios forestales se han convertido en uno de los problemas medioambientales más importantes a escala mundial y constituyen un grave problema en España que, debido a los efectos del cambio climático, puede verse considerablemente agravado.

Con el revuelo que se ha montado con el incendio y la cantidad de informaciones que han salido en la prensa acerca de él y de los incendios en general, Anxo se ha enterado de que desde que en España se inicia la recogida de datos en 1961 el número de incendios se ha ido incrementando de manera alarmante. Y lo peor, más del 80 % son provocados por la mano del hombre.

a) Estudia cuál ha sido la distribución de los incendios forestales por comunidades autónomas y represéntala en un diagrama de barras. También puedes representar la distribución de los grandes incendios por año desde el año 1995, por ejemplo, hasta el último año disponible, calcular su número medio y variabilidad.

b) La superficie forestal, en hectáreas, afectada por incendio es otro dato que proporcionan las tablas. ¿Cuál ha sido la evolución a lo largo del tiempo?¿Por término medio cuántas hectáreas se han visto afectadas al año?¿Y en total?¿Ha habido mucha variación en los últimos diez años?

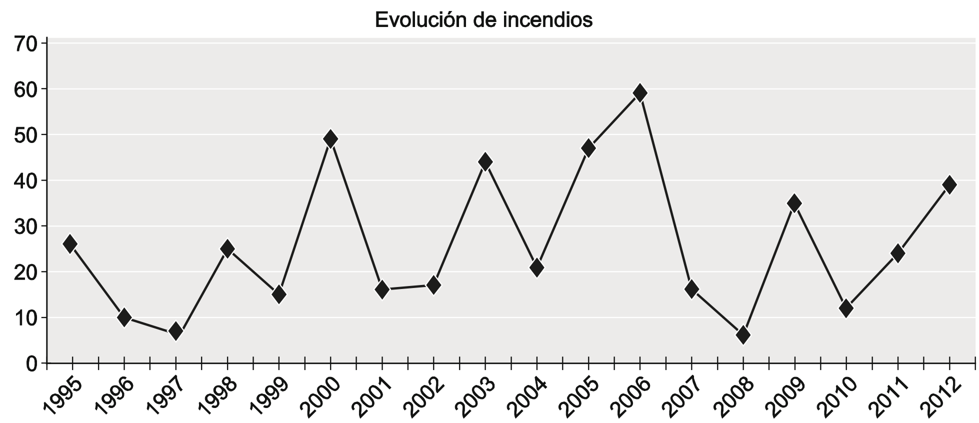
c) ¿Afectan por igual los incendios a los distintos tipos de arbolado?

Los datos se pueden encontrar en el Instituto Nacional de Estadística, en el Ministerio de Agricultura, Alimentación y Medio Ambiente a la que se puede acceder desde smSaviadigital.com

a) En primer lugar se representa la distribución del número de grandes incendios (aquellos en los que la superficie afectada es de más de 500 hectáreas) desde 1995 hasta 2012, teniendo en cuenta que los datos de 2011 y 2012 son provisionales en el momento de la elaboración de esta respuesta. La tabla es:

	1995	1996	1997	1998	1999	2000	2001	2002	2003
Grandes incendios (1)	26	10	7	25	15	49	16	17	44
	2004	2005	2006	2007	2008	2009	2010	2011 (2)	2012 (2)
	21	47	59	16	6	35	12	24	39

Y el gráfico que informa de la evolución del número de grandes incendios forestales en España es:



De la tabla se puede obtener el número medio de incendios de los 18 años disponible (recuerda que los dos últimos son provisionales) y su variabilidad:

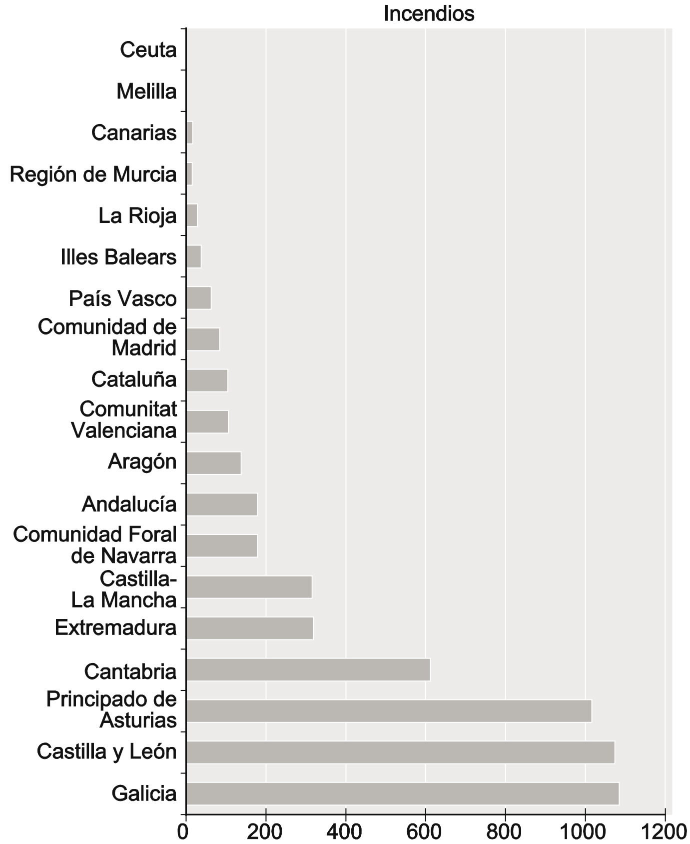
$$\bar{X} = \frac{26+10+\ldots+24+29}{18} = 26, \qquad s_X^2 = \frac{26^2+10^2+\ldots+24^2+29^2}{18} - 26^2 = 237{,}667 \Rightarrow s_X = 15{,}416$$

De forma que la variación relativa del número de grandes incendios en estos 18 años es:

$$CV(X)=\frac{15{,}416}{26}=0{,}59294$$

En segundo lugar se representa en un diagrama de barras la distribución del número de incendios (aquellos en los que la superficie afectada es de más de 1 hectárea) por comunidades autónomas. Se ha elegido el año 2012, el último del que se dispone de datos definitivos en el momento de elaborar esta respuesta.

	2012	
	Total siniestros	Incendios(1)
TOTAL	15 895	5375
Andalucía	888	179
Aragón	527	138
Asturias, Principado de	2220	1017
Balears, Illes	149	38
Canarias	119	16
Cantabria	728	612
Castilla y León	2611	1074
Castilla–La Mancha	1134	316
Cataluña	730	105
Comunitat Valenciana	502	106
Extremadura	1091	319
Galicia	3798	1085
Madrid, Comunidad de	391	84
Murcia, Región de	128	16
Navarra, Comunidad Foral de	598	179
País Vasco	176	63
Rioja, La	105	28
Ceuta	0	0
Melilla	0	0



b) Respecto a la superficie forestal afectada por los incendios, se tienen datos desde 1995 (INE) y se pueden realizar diferentes tipos de estudios. En la tabla siguiente se recoge la información disponible de la superficie total quemada en España y también desglosada según el tipo de masa forestal afectada (Vegetación leñosa, arbolado, matorral y monte abierto o herbáceos)

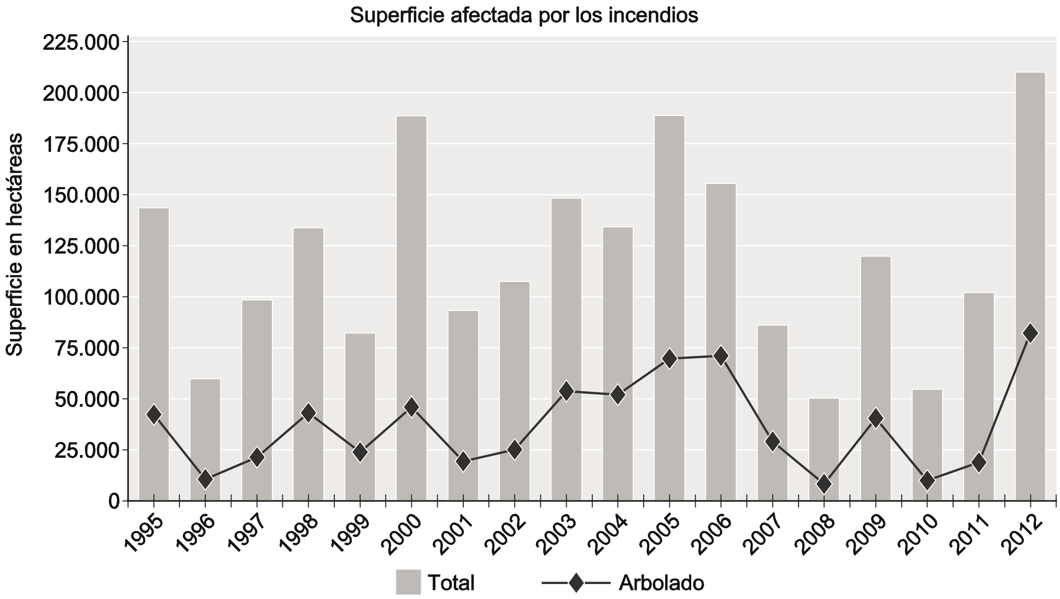
	1995	1996	1997	1998	1999	2000	2001	2002	2003
Superficie forestal total	143468	59813,7	98503,2	133643	82217,4	188586	93297,5	107464	148173
Vegetación leñosa	136921	53039,3	94207	126082	76996,5	170532	75710,8	89007,8	124141
Arbolado	42380,3	10530,9	21326,2	42959,3	24034,3	46138,2	19363,4	25196,9	53673
Matorral y monte abierto (3)	94540,7	42508,4	72880,8	83123,1	52962,2	124394	56347,4	63810,9	70467,8
Herbáceos (pastos y dehesas)	6546,8	6774,3	4296,1	7560,3	5221	18053,4	17586,8	18456,2	24031,6
Porcentaje superficie afectada	0,55	0,229	0,378	0,512	0,315	0,723	0,358	0,412	0,568
	2004	2005	2006	2007	2008	2009	2010	2011 (2)	2012 (2)
Superficie forestal total	134193	188698	155363	86112,5	50321,3	119892	54769,9	102161	209855
Vegetación leñosa	108338	175674	143136	71794,1	41289,7	107696	49464,2	91235,3	198937
Arbolado	51732,2	69396,8	71082,9	29402,6	8443,1	40393,3	10184,9	18847,5	82201,4
Matorral y monte abierto (3)	56606,1	106277	72053,3	42391,6	32846,6	67302,7	39279,3	72387,8	116735
Herbáceos (pastos y dehesas)	25854,3	13073,6	12226,7	14318,4	9031,6	12195,7	5305,7	10926	10918,6
Porcentaje superficie afectada	0,515	0,724	0,572	0,333	0,195	0,428	0,198	0,369	0,759

En la tabla debe observarse que:

Superficie forestal total = vegetación leñosa + herbáceos

Vegetación leñosa = arbolado + matorral y monte abierto

A continuación se representa conjuntamente en un diagrama la superficie total y la que corresponde a arbolado (parte de vegetación leñosa):



Puede verse que de unos años a otros, tanto al superficie total afectada como la de arbolado han sufrido notable variaciones desde las 50 321 hectáreas totales del año 2008 a las más de cuatro veces más del año 2012: 209 855 ha (provisional). También la superficie de arbolado ha variado notablemente, en línea con la total afectada.

En la tabla siguiente se presentan los valores medios de superficie quemada desde 1995 hasta 2012 y en los últimos diez años junto con su desviación típica y su coeficiente de variación:

	Media total 1995–2012	Media 10 últimos años 2002–2012	Desviación típica total	Desviación típica últimos 10 años	*CV* total	*CV* últimos 10 años
Superficie forestal total	119807,2	124953,7	45605,82	50368,88	0,3807	0,4031
Vegetación leñosa	107455,7	111170,5	43752,42	48695,60	0,4072	0,4380
Arbolado	37071,5	43535,8	21596,44	25011,04	0,5826	0,5745
Matorral y monte abierto	70384,2	67634,8	25596,20	25937,31	0,3637	0,3835
Herbáceos (pastos y dehesas)	12354,3	13788,2	6185,14	6060,44	0,5006	0,4395

Puede comprobarse que las cifras de las últimos 10 años son en todos los casos más altas que en el total excepto el caso de matorral y monte abierto. También puede verse que la variabilidad más alta corresponde a la superficie arbolada quemada y la más baja a la superficie total.

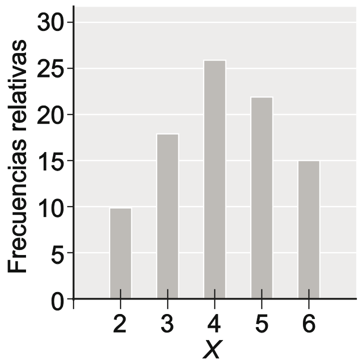
c) Los incendios no afectan por igual a los distintos tipos de superficie. Es claro que la superficie de matorral y monte abierto es la más afectada por los incendios, debido también a que dentro de la superficie forestal es la que ocupa la mayor parte.

NOTA: Los gráficos y resultados presentados son solo una pequeña parte de las posibilidades de análisis que pueden hacerse con los datos disponibles. Quedas invitado a analizar lo que ha pasado en tu comunidad autónoma e incluso en tu provincia y compararlo con lo que ha sucedido en el total del país o con otras comunidades autónomas.

AUTOEVALUACIÓN

Comprueba qué has aprendido

1. Dado el siguiente diagrama de barras,



a) Clasifica la variable.

b) Escribe la tabla de frecuencias absolutas, relativas y relativas acumuladas.

c) Calcula la media, la moda y la mediana

d) Calcula la varianza y la desviación típica.

a) La variable es cuantitativa discreta y toma los valores 2, 3, 4, 5 y 6.

b) La tabla que se deduce del diagrama es la siguiente, con las frecuencias absolutas (f_i), las relativas (h_i) y las relativas acumuladas (H_i).

X	f_i	h_i	H_i
2	10	0,11	0,11
3	17	0,19	0,30
4	26	0,29	0,59
5	22	0,24	0,83
6	15	0,17	1
	90	1	

c) Para calcular los valores de este apartado y del siguiente, se construye la tabla con los valores de la variable, las frecuencias absolutas, las absolutas a acumuladas y los productos necesarios:

X	f_i	F_i	f_ix_i	$f_ix_i^2$
2	10	10	20	40
3	17	27	51	153
4	26	53	104	416
5	22	75	110	550
6	15	90	90	540
	90		375	1699

De manera que la media aritmética es $\overline{X} = \dfrac{375}{90} = 4,167$.

La moda es $M_o = 4$, ya que es el valor de la variable que más veces aparece en la muestra.

La mediana es $M = 4$, ya que una vez ordenados los 90 datos de menor a mayor, los lugares 45 y 46 están ocupados por el valor 4.

d) La varianza y la desviación típica se obtienen a partir de los resultados de la tabla del apartado anterior:

$$s^2 = \frac{1699}{90} - 4,167^2 = 1,514 \Rightarrow s = 1,23$$

2. El número de hijos de 7 familias es 1, 1, 2, 2, 2, 6, 7. Calcula la media, la moda y la mediana ¿Cuál de las tres es más representativa?

La media, la moda y la mediana son respectivamente $\overline{X} = \dfrac{1+1+2+2+2+6+7}{7} = 3;\quad M_o = 2 \quad \text{y} \quad M = 2.$

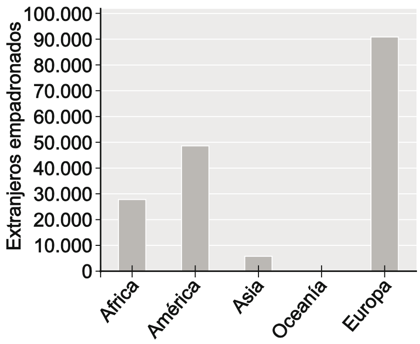
Parece claro que, en este caso, la más representativa de las tres es la mediana (o la moda), puesto que ninguna de las familias tiene 3 hijos. La media es muy sensible a datos extremos.

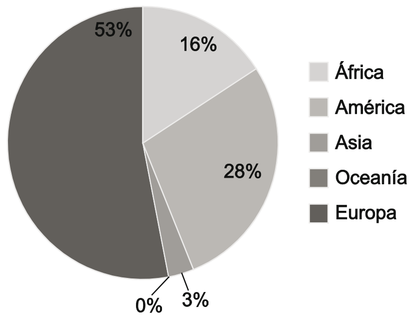
3. El número de extranjeros empadronados en una ciudad, según el continente de procedencia viene dado en la tabla. Describe el tipo de variable y representa adecuadamente los datos

Continente	f_i
África	26 990
América	49 101
Asia	5635
Oceanía	73
Europa	91 000

Se trata de una variable cualitativa.

El tipo de gráfico más apropiado es, bien un diagrama de barras o bien uno de sectores:





4. Una variable estadística *X* tiene la siguiente distribución de frecuencias:

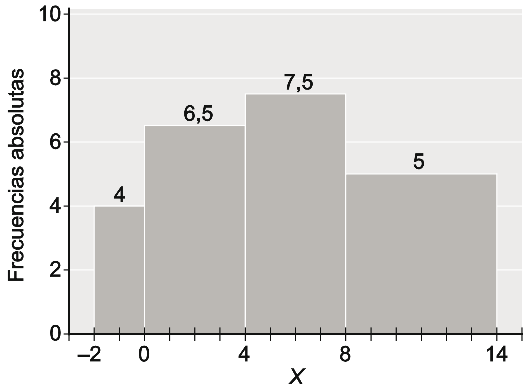
Clase	[–2, 0)	[0, 4)	[4, 8)	[8, 14]
f_i	8	26	30	30

a) Dibuja el histograma

b) Calcula la media, la moda y la mediana

c) Calcula la varianza.

a) Debe tenerse en cuenta que las clases tienen distinta longitud. Tomamos como unidad de base 2; así, la primera clase "mide" 1, la segunda clase "mide" 2, la tercera "mide" 2 y la cuarta "mide" 3.



b) Para calcular las medidas de este apartado y del siguiente se construye la tabla:

Clase	f_i	F_i	x_i	f_ix_i	$f_ix_i^2$
[–2, 0)	8	8	–1	–8	8
[0, 4)	26	34	2	52	104
[4, 8)	30	64	6	180	1080
[8, 14)	30	94	11	330	3630
	94			554	4822

De esta manera, la media aritmética es $\bar{X} = \frac{554}{94} = 5{,}8936$.

El intervalo modal es [4,8) como se puede ver en el histograma, ya que es el que tiene mayor densidad de observaciones, puesto que su longitud es 4, mientras que la del intervalo [8, 14) es 6.

El 50 % de 94 es 47, por lo que, observando la columna de las frecuencias acumuladas, la mediana se encuentra en el intervalo [4,8), que tiene longitud 4 y contiene 30 observaciones. Antes de este intervalo se han acumulado 34 observaciones. Luego $M = 4 + \frac{(47-34)\cdot 4}{30} = 5{,}733$.

c) La varianza se obtiene a partir de los resultados de la última columna de la tabla del apartado anterior

$$s^2 = \frac{4822}{94} - 5{,}8936^2 = 16{,}563$$

5. Los dos conjuntos siguientes de datos se refieren a la estatura de un grupo de 20 estudiantes, en cm, y a las puntuaciones obtenidas por un grupo de 25 personas en un test psicotécnico.

Estatura (X):	**174**	**178**	**165**	**167**	**182**
	172	**185**	**178**	**205**	**180**
	157	**166**	**172**	**190**	**161**
	170	**178**	**183**	**169**	**176**

Puntuación test (Y)	**56**	**86**	**70**	**67**	**68**
	76	**68**	**45**	**58**	**74**
	68	**87**	**51**	**27**	**67**
	51	**30**	**58**	**97**	**70**
	47	**76**	**57**	**71**	**56**

a) Determina cuál de los dos conjuntos de datos presenta mayor variabilidad relativa.

b) Calcula los cuartiles de ambos conjuntos de datos.

c) Dibuja el diagrama de caja de cada distribución de datos y señala si existen valores extremos.

a) Para poder comparar la variabilidad hay que conocer los coeficientes de variación.
En el caso de las estaturas, se pueden agrupar los datos en intervalos de amplitud 10.

Estatura	fi	F_i	x_i	$f_i x_i$	$f_i x_i^2$
[155, 165)	2	2	160	320	51 200
[165, 175)	8	10	170	1360	231 200
[175, 185)	7	17	180	1260	226 800
[185, 195)	2	19	190	380	72 200
[195, 205]	1	20	200	200	40 000
	20			3520	621 400

De donde la media es: $\bar{X} = \frac{3520}{20} = 176$.

La varianza y la desviación típica son: $s^2 = \frac{621\,400}{20} - 176^2 = 94 \Rightarrow s = \sqrt{94} = 9{,}695$.

El coeficiente de variación es $CV_X = \frac{9{,}685}{176} = 0{,}055$.

En el caso de las puntuaciones, los datos se pueden agrupar en intervalos de amplitud 15.

Puntos	f_i	F_i	y_i	$f_i y_i$	$f_i y_i^2$
[25, 40)	2	2	32,5	65	2112,5
[40, 55)	4	6	47,5	190	9025
[55, 70)	10	16	62,5	625	39 062,5
[70, 85)	6	22	77,5	465	36 037,5
[85, 100]	3	25	92,5	277,5	25 668,75
	25			1622,5	111 906,25

De donde la media es: $\bar{Y} = \frac{1622{,}5}{25} = 64{,}9$.

La varianza y la desviación típica son: $s^2 = \frac{111906{,}25}{25} - 176^2 = 264{,}24 \Rightarrow s = \sqrt{264{,}24} = 16{,}26$.

El coeficiente de variación es $CV_Y = \frac{16{,}26}{64{,}9} = 0{,}25$.

Por tanto las puntuaciones del test presentan mayor variabilidad.

b) Para calcular los cuartiles y los datos necesarios en el siguiente apartado se opera como sigue:

Estaturas

Cuartil Q_1. El 25 % de 20 es 5 que se acumula en el intervalo [165, 175), de longitud 10 y que contiene 8 observaciones. Antes de éste intervalo se tienen acumuladas 2 observaciones, luego:

$$Q_1(X) = 165 + \frac{(5-2)\cdot 10}{8} = 168,75cm$$

Cuartil Q_2, la mediana. El 50 % de 20 es 10, que se acumula en el intervalo [165, 175), de longitud 15 y que contiene 10 observaciones. Antes de éste intervalo se tienen acumuladas 2 observaciones, luego:

$$M_X = Q_2(X) = 165 + \frac{(10-2)\cdot 10}{8} = 175\ cm$$

Cuartil Q_3. El 75 % de 20 es 15, que se acumula en el intervalo [175, 185), de longitud 10 y que contiene 7 observaciones. Antes de éste intervalo se tienen acumuladas 10 observaciones, luego:

$$Q_3(X) = 175 + \frac{(15-10)\cdot 10}{7} = 182,14cm$$

El rango intercuartílico es $RIC_x = Q_3(X) - Q_1(X) = 182,14 - 168,75 = 13,39$

De donde los límites inferior y superior son:

$LI_X = Q_1(X) - 1,5\cdot RIC_x = 168,75 - 1,5\cdot 13,39 = 148,66$

$LS_X = Q_3(X) + 1,5\cdot RIC_x = 182,14 + 1,5\cdot 13,39 = 202,23$

Tenemos así un valor atípico que es 205 y quedan como máximo y mínimo respectivamente 190 y 157.

Puntos

Cuartil Q_1. El 25 % de 25 es 6,25 que se acumula en el intervalo [55, 70), de longitud 15 y que contiene 10 observaciones. Antes de éste intervalo se tienen acumuladas 6 observaciones, luego:

$$Q_1(Y) = 55 + \frac{(6,25-6)15}{10} = 55,375$$

Cuartil Q_2, la mediana. El 50 % de 25 es 12,5, que se acumula en el intervalo [55, 70), de longitud 15 y que contiene 10 observaciones. Antes de éste intervalo se tienen acumuladas 6 observaciones, luego:

$$M_Y = Q_2(Y) = 55 + \frac{(12,5-6)\cdot 15}{10} = 64,75$$

Cuartil Q_3. El 75 % de 25 es 18,75, que se acumula en el intervalo [70, 85), de longitud 15 y que contiene 6 observaciones. Antes de éste intervalo se tienen acumuladas 16 observaciones, luego:

$$Q_3(Y) = 70 + \frac{(18,75-16)\cdot 15}{6} = 76,875$$

El rango inter cuartílico es $RIC_Y = Q_3(Y) - Q_1(Y) = 76,785 - 55,375 = 21,5$

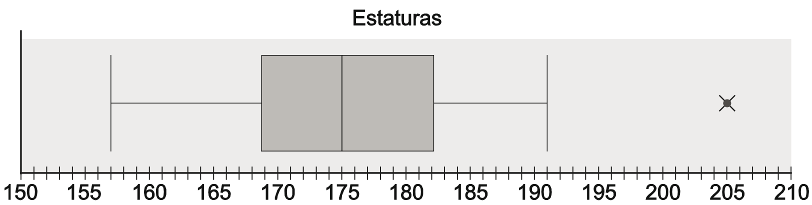
De donde los límites inferior y superior son:

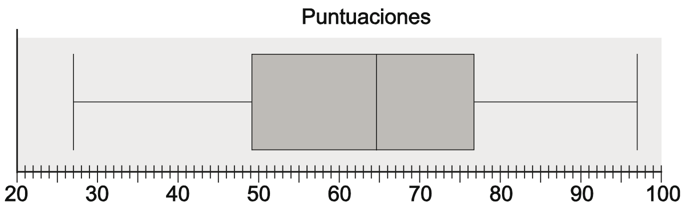
$LI_Y = Q_1(Y) - 1,5\cdot RIC_Y = 55,375 - 1,5\cdot 21,54 = 23,125$

$LS_Y = Q_3(Y) + 1,5\cdot RIC_Y = 76,875 + 1,5\cdot 21,5 = 109,125$

Por tanto no hay valores atípicos.

c) Con estos datos los diagramas de caja son:





Relaciona y contesta

Elije la única respuesta correcta en cada caso

1. En una población de 5200 habitantes mayores de 18 años, 780 son inmigrantes. Una muestra de 200 personas en la que esté representada la población inmigrante, debe constar de:

A. 20 inmigrantes. **B.** 150 nativos **C.** 30 inmigrantes **D.** 100 nativos

Como la proporción de inmigrantes en la muestra debe ser igual que en toda la población, ha de ser:

$\frac{5200}{780} = \frac{200}{x} \Rightarrow x = 30$ inmigrantes debe haber en la muestra. Respuesta C

2. Los datos recogidos de una variable estadística cuantitativa indican que $Q_1 = 14$, $RIC = 9$ y que la distancia entre la mediana y el Q_1 es el doble que entre la mediana y Q_3. Entonces, los valores de la mediana y del Q_3 son:

A. $M = 17, Q_3 = 20$ **B.** $M = 20, Q_3 = 23$ **C.** $M = 23, Q_3 = 20$ **D.** $M = 19, Q_3 = 23$

Por definición es $Q_3 = Q_1 + RIC = 14 + 9 = 23$. Por otra parte el enunciado afirma que $M - Q_1 = 2(Q_3 - M)$, de donde se deduce que $M = 20$. Respuesta B

3. Las observaciones de una variable estadística continua tienen media 5 y varianza 5. Si todas las observaciones se dividen por 2, la nueva media y varianza son:

A. $\bar{Y} = 2{,}5,\ s_Y^2 = 1{,}25$ **B.** $\bar{Y} = 1{,}25,\ s_Y^2 = 2{,}5$ **C.** $\bar{Y} = 1{,}25,\ s_Y^2 = 1{,}25$ **D.** $\bar{Y} = 2{,}5,\ s_Y^2 = 2{,}5$

La media se dividen por 2 y la varianza por $2^2 = 4$ (cuestión 44). Respuesta A

4. El coeficiente de variación de la variable *X* es el triple que el de la variable *Y*. Si la media de *Y* es seis veces mayor que la de *X*,

A. $s_Y^2 = 6s_X^2$ **B.** $s_Y^2 = 2s_X^2$ **C.** $s_Y^2 = 2s_X^2$ **D.** $s_Y^2 = 4s_X^2$

$CV_X = 3CV_Y \Rightarrow \frac{s_X}{\bar{X}} = 3\frac{s_Y}{\bar{Y}} \Rightarrow \frac{s_X}{\bar{X}} = \frac{3s_Y}{6\bar{X}} \Rightarrow s_Y = 2s_X \Rightarrow s_Y^2 = 4s_X^2$ Respuesta D

Señala, en cada caso, las respuestas correctas

5. En una distribución, $Q_3 = 2Q_1$, entonces:

A. El rango intercuartílico coincide con el primer cuartil.

B. La mediana es la media aritmética de los cuartiles primero y tercero.

C. El límite inferior necesario para calcular el diagrama de caja es negativo.

D. La distancia entre los límites superior e inferior (del diagrama de caja) es cuatro veces el primer cuartil.

A. Cierta. $RIC = Q_3 - Q_1 = 2Q_1 - Q_1 = Q_1$

B. Falso. Por ejemplo en la serie 1, 1, 1, 2, 2, 2, 9, 9, 9, 9 es $Q_1 = 1$; $Q_3 = 9$; $M = 2 \neq \frac{Q_1 + Q_3}{2} = \frac{1+9}{2} = 5$.

C. Cierto. $LI = Q_1 - 1{,}5RIC = Q_1 - 1{,}5Q_1 = -0{,}5Q_1$.

D. Cierto. $LS - LI = Q_3 + 1{,}5RIC - (Q_1 - 1{,}5RIC) = 2Q_1 - Q_1 + 3Q_1 = 4Q_1$ .Respuesta A, C y D

6. **Para medir la dispersión de los datos en torno a la media, se utiliza**

A. Los percentiles de la distribución.

B. La varianza o la desviación típica.

C. La desviación absoluta media.

D. El rango o recorrido de la variable.

La dispersión de los datos se evalúa con las llamadas "medidas de dispersión". De las magnitudes dadas en las soluciones solo los percentiles (A) no pertenecen a a esa categoría. Por tanto las respuestas correctas son B, C y D.

Elige la relación correcta entre las dos afirmaciones dadas

7. **1. La media de un conjunto de datos es 2 y su varianza 35.**

2. El conjunto de datos es heterogéneo.

A. $1 \Rightarrow 2$ pero $2 \not\Rightarrow 1$

B. $2 \Rightarrow 1$ pero $1 \not\Rightarrow 2$

C. $1 \Leftrightarrow 2$

D. No hay relación entre 1 y 2.

Si por heterogéneo se entiende disperso, entonces la afirmación 1 implica que este conjunto de dato es bastante disperso ya que su coeficiente de variación tiene un valor elevado:

$$CV = \frac{s_X}{\overline{X}} = \frac{\sqrt{35}}{2} = 2{,}96$$

Por tanto 1 ⇒ 2. La relación inversa, 2 ⇒ 1, no es cierta porque saber que un conjunto de datos es más o menos disperso no dice nada respecto a qué valores concretos puedan tener su media y su desviación típica.

En resumen, la respuesta es la A.