PARÁMETROS ESTADÍSTICOS

1. PARÁMETROS DE CENTRALIZACIÓN

La información recogida en una tabla o gráfica estadística suele resumirse en unos pocos valores que nos informan del comportamiento de todos los individuos del colectivo estudiado. Estos valores, representativos de todos los de una distribución, se llaman ***parámetros*** o ***medidas de centralización.*** Estos parámetros tienden a situarse hacia el centro del conjunto de datos ordenados.

1.1. Media aritmética

Media aritmética de una variable estadística es el cociente entre la suma de todos los valores de dicha variable y el número total de éstos. Se representa por $\overline{x}$.

Su cálculo se realiza, según las expresiones que siguen, atendiendo a la presentación de los datos.

- Para datos sin frecuencias. Si la variable toma los N valores $x_1, x_2, \ldots, x_N$, la media aritmética se calcula mediante la expresión:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

- Para datos con frecuencias. Si la variable toma los valores o marcas de clase $x_1, x_2, \ldots, x_k$, con frecuencias absolutas $n_1, n_2, \ldots, n_k$, la media aritmética se calcula mediante la expresión:

$$\overline{x} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{N} = \frac{\sum_{i=1}^{k} n_i x_i}{N}$$

Ejemplo.- Calcula la media aritmética de la superficie de los Parques Nacionales.

Parques Nacionales	*ha*
Picos de Europa	16.925
Ordesa y Monte Perdido	15.608
El Teide	13.571
La Caldera Taburiente	4.690
Timanfaya	5.107
Doñana	50.720
La Tablas de Daimiel	1.928
El Archipiélago de Cabrera	10.025
Garajonay	3.984
Total	**122.558**

Los datos aparecen sin frecuencias. Debemos sumar todas las superficies y dividir esta suma por el número de Parques Nacionales.

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{122.558}{9} = 13.617'6 \text{ hectáreas}$$

Ejemplo.- Las edades de los alumnos de una clase son las que se reflejan en la tabla. Hallamos la edad media de los alumnos.

Edad (x_i)	n_i	$n_i x_i$
13	6	78
14	7	98
15	4	60
16	3	48
Total	**20**	**284**

La variable es discreta. Para hallar la media debemos sumar todos los datos, lo que equivale a multiplicar cada valor por su frecuencia absoluta y sumar todos los productos.

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{N} = \frac{284}{20} = 14'2 \text{ años}$$

Ejemplo.- Hemos tallado los treinta alumnos de una clase. Con los datos obtenidos y agrupando las tallas por intervalos hemos calculado la talla media de éstos.

La variable es continua y los datos están agrupados en clases. La media de una variable continua se obtiene sumando los productos de las marcas de clase de los intervalos por su frecuencia absoluta y dividiendo ese producto entre el número total de datos.

Talla (cm)	*Marcas de clase* (x_i)	n_i	$n_i x_i$
[150, 155)	152'5	1	152'5
[155, 160)	157'5	3	472'5
[160, 165)	162'5	10	1.625
[165, 170)	167'5	12	2.010
[170, 175]	172'5	4	690
Total		**30**	**4.950**

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{N} = \frac{4.950}{30} = 165 \text{ cm}$$

1.2. Moda

Moda de una variable estadística es el valor de dicha variable que tiene mayor frecuencia absoluta (es decir, el valor de la variable que más se repite). Se representa por ***Mo.***

Ejemplo.- Para las edades de los alumnos del ejemplo anterior, la moda es *Mo* = 14, ya que 14 es la edad que tienen un mayor número de alumnos.
Esta distribución, como sólo tiene una moda, se dice que es una *distribución unimodal.*

Ejemplo.- Dada una distribución de frecuencias, su tabla es la siguiente.

x_i	1	2	3	4	5	6
n_i	6	7	14	10	14	9

Las modas son *Mo* = 3 y *Mo* = 5, por ser estos dos valores de la variable los que tienen mayor frecuencia. Diremos en estos casos que se trata de una *distribución bimodal.*

En el caso de que los datos estén agrupados en intervalos llamamos ***clase*** o ***intervalo modal*** a la clase que presenta mayor frecuencia absoluta. Si no necesitamos mucha precisión en el cálculo de la moda, podemos tomar como valor aproximado de la misma la marca de clase del intervalo modal. Cuando es necesaria mayor precisión en el cálculo recurrimos a la siguiente expresión que nos da su valor exacto.

$$Mo = e_i + \frac{n_{Mo} - n_{Mo-1}}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})} \cdot a$$

- e_i = extremo inferior del intervalo modal
- a = amplitud del intervalo modal
- n_{Mo}, n_{Mo-1}, n_{Mo+1} = frecuencias absolutas del intervalo modal, del intervalo anterior y del posterior, respectivamente

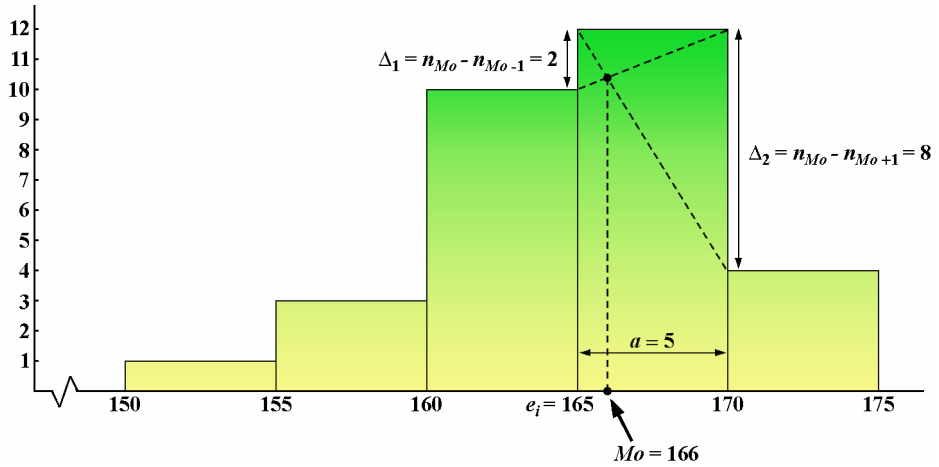
Ejemplo.- En la distribución estadística vista con anterioridad que nos proporciona la talla de los treinta alumnos de una clase, el mayor valor de la frecuencia absoluta 12 da como intervalo modal [165, 170). El valor aproximado de la moda es 167'5 cm.

No obstante, el valor exacto de está es:

$$Mo = 165 + \frac{12-10}{(12-10)+(12-4)} \cdot 5 = 165 + \frac{2}{2+8} \cdot 5 = 165 + 1 = 166 \text{ cm}$$

Por tanto, la talla moda de esta clase es Mo = 166 cm, que como podemos ver se aproxima mucho a la marca de clase del intervalo modal, 167'5 cm.

Observa como gráficamente, a través del histograma, también se puede calcular la moda.



$$Mo = e_i + \frac{n_{Mo} - n_{Mo-1}}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})} \cdot a = e_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot a = 165 + \frac{2}{2+8} \cdot 5 = 165 + 1 = 166 \text{ cm}$$

1.3. Mediana

Las calificaciones que han obtenido 7 alumnos en Matemáticas y 8 alumnos en Lengua han sido las siguientes:

Matemáticas: 2 4 5 **6** 6 7 7 ; Lengua: 2 2 4 **4** **6** 8 8 8

Observamos que en Matemáticas la nota 6 deja tres alumnos a su izquierda y tres a su derecha. En las de Lengua, como no hay una nota central, tomamos la media aritmética de las dos notas centrales: (4 + 6)/2 = 5. Decimos que la nota mediana en Matemáticas es 6 y en Lengua 5.

> ***Mediana*** de una variable estadística es el valor (no es siempre un valor de la variable) que, tras ordenar los datos de forma creciente, deja a su izquierda un número de datos iguales a los que deja a su derecha; es decir, es el valor tal que la mitad de los datos son menores o iguales que él y la otra mitad iguales o mayores. Se denota por ***Me.***

Ejemplo.- Dada la serie estadística 11, 3, 5, 9, 12, 2, 6, calcula la mediana.

Ordenamos los datos: 2, 3, 5, 6, 9, 11, 12 $\Rightarrow$ la mediana es Me = 6, por ser éste el valor central.

Ejemplo.- Dada la serie estadística 12, 5, 3, 9, 11, 13, 2, 6, calcula la mediana.

Ordenamos los datos: 2, 3, 5, 6, 9, 11, 12, 13 $\Rightarrow$ en este caso hay dos valores centrales, que son 6 y 9; la mediana es Me = (6 + 9)/2 = 7'5.

El proceso anterior, para calcular la mediana, es útil cuando disponemos de pocos datos, pero cuando el número de éstos es grande este procedimiento resulta muy laborioso, siendo necesario construir una tabla estadística con frecuencias absolutas acumuladas.

De esta forma, para el cálculo de la mediana de una variable estadística discreta debemos distinguir dos casos:

- Que no exista ningún valor de la variable, x_i, cuya frecuencia absoluta acumulada, N_i, sea igual que la mitad del número de individuos, $N/2$. E este caso, la mediana es el primer valor de la variable cuya frecuencia absoluta acumulada sea mayor que la mitad del número de individuos.

Ejemplo.- Las calificaciones (x_i) que obtuvieron los 32 alumnos de una clase en la asignatura de Inglés fueron las que proporciona la siguiente tabla. Halla la calificación mediana.

x_i	n_i	N_i
1	2	2
2	2	4
3	3	7
4	5	12
5	7	19
6	5	24
7	3	27
8	2	29
9	2	31
10	1	32
Total	**32**	

La mitad del número total de individuos es $N/2 = 16$

La calificación mediana es $Me = 5$, dado que es el primer valor de la variable cuya frecuencia absoluta acumulada, 19, excede a la mitad del número de datos, 16.

- En el caso de que exista un valor de la variable, x_i, cuya frecuencia absoluta acumulada sea igual que la mitad del número de individuos, es decir, $N_i = N/2$, la mediana ha de ser la media aritmética entre dicho valor de la variable y el siguiente.

Ejemplo.- En el examen de evaluación, las calificaciones que obtuvieron fueron muy parecidas. Hallemos nuevamente la calificación mediana.

x_i	n_i	N_i
1	2	2
2	2	4
3	3	7
4	5	12
5	4	16
6	7	23
7	3	26
8	2	28
9	3	31
10	1	32
Total	**32**	

En este caso, el valor $x = 5$ tiene por frecuencia absoluta acumulada 16, que es precisamente la mitad del número total de individuos: $N/2 = 16$.

La calificación mediana es ahora $Me = (5 + 6)/2 = 5'5$

En el caso de que los datos estén agrupados en intervalos, llamamos ***intervalo*** o ***clase mediana*** a la primera clase o intervalo cuya frecuencia absoluta acumulada sobrepase estrictamente a la mitad del número de individuos. Si no necesitamos mucha precisión podemos tomar como valor aproximado de la mediana la marca de clase correspondiente a la clase mediana. Cuando es necesaria mayor precisión en el cálculo de la mediana, para variables agrupadas en intervalos, utilizamos la siguiente expresión que nos da su valor exacto.

$$Me = e_i + \frac{\frac{N}{2} - N_{Me-1}}{n_{Me}} \cdot a$$

- e_i = extremo inferior de la clase mediana
- a = amplitud de la clase mediana
- n_{Me} = frecuencia absoluta de la clase mediana
- N_{Me-1} = frecuencia absoluta acumulada de la clase anterior a la clase mediana

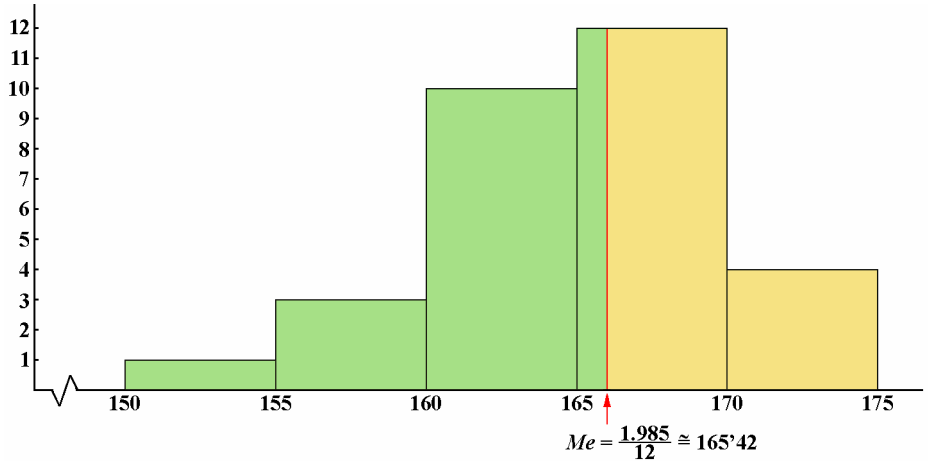
Ejemplo.- Encuentra la talla mediana en la distribución estadística vista en el ejemplo del epígrafe 1.1.

***Talla* (cm)**	x_i	n_i	N_i
[150, 155)	152'5	1	1
[155, 160)	155'5	3	4
[160, 165)	162'5	10	14
[165, 170)	167'5	12	26
[170, 175]	172'5	4	30
Total		**30**	

El intervalo o clase mediana es [165, 170) ya que es el primer intervalo cuya frecuencia absoluta acumulada, 26, sobrepasa a la mitad del número de individuos, $N/2 = 15$.
El valor aproximado de la mediana es entonces 167'5 cm.
Para obtener la mediana exacta utilizamos la expresión indicada anteriormente:

$$Me = 165 + \frac{\frac{30}{2} - 14}{12} \cdot 5 = 165 + \frac{5}{12} = \frac{1.985}{12} \cong 165'42 \text{ cm}$$

Cuando los datos están agrupados en intervalos, la mediana puede interpretarse geométricamente como el punto del eje de abscisas que permite dividir el histograma de frecuencias absolutas en dos partes de igual área. Lo comprobamos a continuación con el ejemplo anterior.



Área de la izquierda: $5 \cdot 1 + 5 \cdot 3 + 5 \cdot 10 + \frac{5}{12} \cdot 12 = 5 + 15 + 50 + 5 = 75$

Área de la derecha: $\frac{55}{12} \cdot 12 + 5 \cdot 4 = 55 + 20 = 75$

Ejemplo.- En este ejemplo puedes ver que no se presenta ningún problema cuando se encuentra un intervalo cuya frecuencia absoluta acumulada es igual que la mitad del número total de individuos. Calculamos la talla mediana de esta distribución (ligeramente modificada de la anterior).

***Talla* (cm)**	x_i	n_i	N_i
[150, 155)	152'5	1	1
[155, 160)	155'5	3	4
[160, 165)	162'5	11	15
[165, 170)	167'5	10	25
[170, 175]	172'5	5	30
Total		**30**	

El intervalo o clase mediana es el mismo, [165, 170), pues es el primero cuya frecuencia absoluta acumulada, 25, sobrepasa a la mitad del número de individuos, $N/2 = 15$.

El valor aproximado de la mediana es 167'5 cm.

Obtenemos el valor exacto:

$$Me = 165 + \frac{\frac{30}{2} - 15}{10} \cdot 5 = 165 + 0 = 165 \text{ cm; lógicamente } Me = e_i.$$

2. CUARTILES

Hemos visto anteriormente que la mediana separa los datos, ordenados de menor a mayor, en dos partes con el mismo número de datos. Pero en ocasiones necesitamos saber más acerca de la distribución de los datos, por lo que se hace necesario introducir otras medidas como son los ***cuartiles***.

Así como la mediana separa los datos en dos grupos, los cuartiles separan los datos en cuatro grupos de la siguiente manera:

$$2 \quad 2 \quad 3 \quad \mathbf{3} \quad 3 \quad 4 \quad 4 \quad \mathbf{5} \quad 6 \quad 6 \quad 7 \quad \mathbf{7} \quad 8 \quad 8 \quad 9$$

$$Q_1 = 3 \qquad Q_2 = Me = 5 \qquad Q_3 = 7$$

$$2 \quad 2 \quad 3 \quad \mathbf{3} \quad \mathbf{3} \quad 4 \quad 4 \quad \mathbf{5} \quad \mathbf{6} \quad 6 \quad 7 \quad \mathbf{7} \quad \mathbf{8} \quad 8 \quad 9 \quad 9$$

$$Q_1 = \frac{3+3}{2} = 3 \qquad Q_2 = Me = \frac{5+6}{2} = 5'5 \qquad Q_3 = \frac{7+8}{2} = 7'5$$

Primer cuartil Q_1: es el menor valor que supera a la cuarta parte de los datos.

Segundo cuartil Q_2: es el menor valor que supera a la mitad de los datos, es decir, la mediana.

Tercer cuartil Q_3: es el menor valor que supera las tres cuartas partes de los datos.

El proceso para hallar estos parámetros es análogo al cálculo de la mediana. Veamos, a través de los siguientes ejemplos, cómo se hallan los cuartiles de una distribución estadística discreta o continua.

- Variable estadística discreta

x_i	n_i	N_i
1	2	2
2	2	4
3	3	7
4	5	12
5	7	19
6	5	24
7	3	27
8	2	29
9	2	31
10	1	32
Total	**32**	

La cuarta parte del número total de datos es $N/4 = 8$. El primer cuartil es $Q_1 = 4$, dado que es el primer valor de la variable cuya frecuencia absoluta acumulada, 12, excede a la cuarta parte del número de datos, 8.

La mitad del número total de datos es $N/2 = 16$. E segundo cuartil o mediana es, por tanto, $Q_2 = Me = 5$, dado que es el primer valor de la variable cuya frecuencia absoluta acumulada, 19, excede a la mitad del número de datos, 16.

En este caso, el valor $x = 6$ tiene por frecuencia absoluta acumulada 24, que es precisamente las tres cuartas partes del número total de datos: $3N/4 = 24$. El tercer cuartil es ahora $Q_3 = (6 + 7)/2 = 6'5$

- Variable estadística continua

En el caso de que los datos estén agrupados en intervalos, consideraremos el *intervalo* o *clase* cuya frecuencia absoluta acumulada sobrepase estrictamente al número de datos en cuestión. Podemos tomar como valor aproximado de los distintos cuartiles las marcas de clases correspondientes; cuando es necesaria mayor precisión en sus cálculos, usaremos las siguientes expresiones que nos proporcionan los valores exactos.

$$\boldsymbol{Q_1 = e_i + \frac{\frac{N}{4} - N_{Q_1-1}}{n_{Q_1}} \cdot a}$$

$$\boldsymbol{Q_2 = Me = e_i + \frac{\frac{N}{2} - N_{Me-1}}{n_{Me}} \cdot a}$$

$$\boldsymbol{Q_3 = e_i + \frac{\frac{3N}{4} - N_{Q_3-1}}{n_{Q_3}} \cdot a}$$

Hallemos los cuartiles de la distribución estadística continua correspondiente a los datos obtenidos de las alturas de 32 personas.

***Talla* (cm)**	x_i	n_i	N_i
[150, 155)	152'5	2	2
[155, 160)	155'5	3	5
[160, 165)	162'5	10	15
[165, 170)	167'5	12	37
[170, 175]	172'5	5	32
Total		**32**	

$N/4 = 8$, luego el intervalo correspondiente al primer cuartil es [160, 165) ya que es el primer intervalo cuya frecuencia absoluta acumulada, 15, sobrepasa a la cuarta parte del número de individuos.
El valor aproximado del primer cuartil es entonces 162'5 cm.
Obtengamos su valor exacto:

$$Q_1 = e_i + \frac{\frac{N}{4} - N_{Q_1-1}}{n_{Q_1}} \cdot a = 160 + \frac{\frac{32}{4} - 5}{10} \cdot 5 = 160 + 1'5 = 161'5 \text{ cm}$$

$N/2 = 16$, con lo que el intervalo [165, 170) es el intervalo mediano o intervalo correspondiente al segundo cuartil, ya que es el primero cuya frecuencia absoluta acumulada, 37, sobrepasa a la mitad del número de individuos.
El valor aproximado del segundo cuartil o mediana es 167'5 cm.
Hallamos su valor exacto:

$$Q_2 = Me = e_i + \frac{\frac{N}{2} - N_{Me-1}}{n_{Me}} \cdot a = 165 + \frac{\frac{32}{2} - 15}{12} \cdot 5 = 165 + 0'42 = 165'42 \text{ cm}$$

$3N/4 = 24$, por lo que [165, 170) es también el intervalo correspondiente al tercer cuartil, pues es el primer intervalo cuya frecuencia absoluta acumulada, 37, sobrepasa a las tres cuartas partes del número de individuos.
Consecuentemente, el valor aproximado del tercer cuartil es también 167'5 cm.
Calculamos su valor exacto:

$$Q_3 = e_i + \frac{\frac{3N}{4} - N_{Q_3-1}}{n_{Q_3}} \cdot a = 165 + \frac{\frac{3 \cdot 32}{4} - 15}{12} \cdot 5 = 165 + 3'75 = 168'75 \text{ cm, valor bastante superior a la mediana}$$

EJERCICIOS

1. La temperatura que ha marcado un termómetro en los diferentes días de la semana, ha sido (en grados centígrados) los que pueden verse en la tabla.

	Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
Mínima	4	–2	–3	1	4	0	3
Máxima	19	18	21	13	12	14	22

   a) Calcula la temperatura media mínima.
   b) Calcula la temperatura media máxima.
   c) Calcula la media de las oscilaciones extremas diarias.

2. Dada la distribución estadística siguiente: 3, 2, 5, 7, 6, 4, 2, 1, 9, 5, 7, 6, 4. Calcula la media aritmética, la moda, la mediana y los cuartiles.

3. Halla la media, la mediana, la moda y los cuartiles de la distribución cuya tabla de frecuencias es la siguiente.

x_i	3	6	7	8	10	12
n_i	6	9	7	8	17	13

4. Las edades de los componentes de una peña de aficionados al fútbol son:

18, 16, 21, 20, 18, 16, 21, 18, 21, 18, 20, 19, 36, 24, 18, 20, 18, 19, 20

a) Calcula la edad media, la edad moda y la edad mediana, así como los cuartiles.
b) Representa gráficamente los datos de esta distribución.

5. La siguiente tabla muestra la distribución, a lo largo de un mes, del número de camiones que circulan diariamente por un cruce de carreteras.

Nº de camiones por día	[350, 400)	[400, 450)	[450, 500)	[500, 550)	[550, 600]
Nº de días	2	5	11	9	4

Calcula la media, la moda, la mediana y los cuartiles de esta distribución.

6. Las respuestas correctas a un test de 80 preguntas realizado por 600 personas son las que se recogen a continuación.

Respuestas	[0, 10)	[10, 20)	[20, 30)	[30, 40)	[40, 50)	[50, 60)	[60, 70)	[70, 80]
Nº de personas	40	60	75	90	105	85	80	65

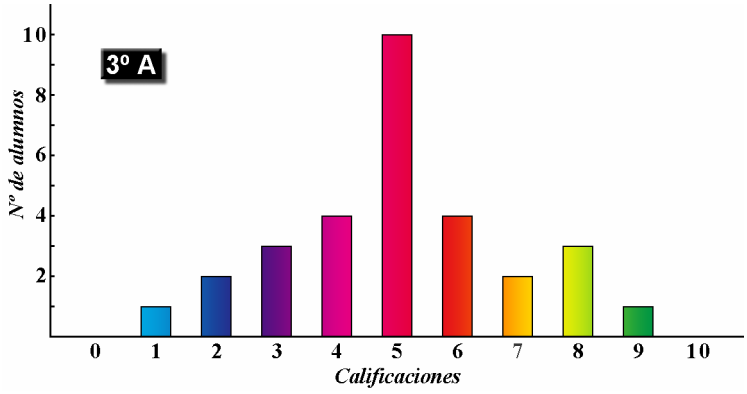
Calcula el número medio de respuestas correctas, la moda y la mediana. Halla los cuartiles. Interpreta gráficamente el cálculo de la moda y de la mediana, y comprueba que la mediana es el punto del eje de abscisas que divide el histograma de frecuencias absolutas en dos partes de igual área.

7. La media de x, $4x - 3$, $x + 4$, -16, 9 y $x - 5$ es 4. ¿Cuánto vale la mediana de esta serie de números?

8. La siguiente serie de datos: 18, 21, 24, a, 36, 37, b, está ordenada y tiene de mediana 30 y de media 32. Encuentra el valor de a y b.

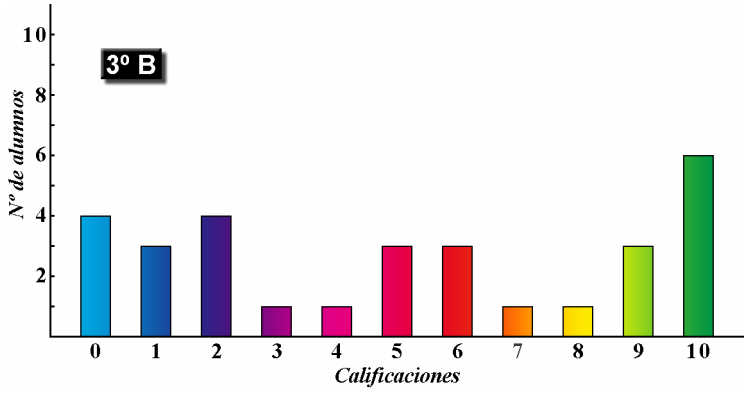
3. Parámetros de dispersión

Este año hay dos cursos muy desiguales en cuanto al rendimiento en Matemáticas. Observa sus calificaciones.



¡Poca dispersión!

En 3º A hay pocas notas bajas, pocas altas y casi todas se sitúan en torno al 5.



¡Mucha dispersión!

En 3º B hay bastantes alumnos con muy bajo rendimiento, bastantes con muy buen rendimiento y pocas calificaciones en torno al 5.

Aunque estas distribuciones de notas tienen aspecto diferente, sus medias son parecidas, $\bar{x}_A = 5'03$ y $\bar{x}_B = 5'1$; es decir, lo que diferencia a ambos cursos es su comportamiento respecto a la media.

Es necesario, pues, conocer en qué medida los datos numéricos están agrupados o no alrededor de los valores centrales. A esto es a lo que se llama ***dispersión,*** y los parámetros que nos informan de cómo se separan los datos se llaman *parámetros* o *medidas de dispersión.*

> Los ***parámetros de dispersión*** son valores numéricos que nos informan de las desviaciones que sufren los datos de una distribución estadística respecto de los parámetros centrales, en particular respecto a la media aritmética.

3.1. Rango o recorrido

Una manera muy sencilla de determinar el grado de dispersión de los datos es observar la separación entre el dato más grande y el más pequeño de la distribución estadística.

> ***Rango*** o ***recorrido*** de una variable estadística es la diferencia entre el mayor y el menor valor de la variable estadística. Se representa por ***R.***

Ejemplo.- Halla el rango de la siguiente distribución estadística.

Puntuación	[38, 44)	[44, 50)	[50, 56)	[56, 62)	[62, 68)	[68, 74)	[74, 80]
Nº de alumnos	7	8	15	25	18	9	6

$R = 80 - 38 = 42$ puntos

El rango es un parámetro fácil de calcular, pero que ofrece una información muy limitada. Así, nos da una idea de la amplitud del conjunto de datos, pero está muy influido por los valores extremos.

3.2. Desviación media

La distancia entre cualquier dato y la media aritmética, $|x_i - \bar{x}|$, recibe el nombre de *desviación* de dicho dato. Una manera de observar la dispersión de la distribución estadística es calcular la media aritmética de todas las desviaciones.

> ***Desviación media*** de una variable estadística es la media aritmética de las desviaciones de todos los datos respecto a su media aritmética. Se representa por $\boldsymbol{d_m}$.

- Si la variable toma los N valores $\boldsymbol{x_1, x_2, \ldots, x_N}$ (datos sin frecuencia) la desviación media se puede calcular mediante la siguiente expresión:

$$d_m = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \ldots + |x_N - \bar{x}|}{N} = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N}$$

- Si la variable toma los valores o marcas de clase $\boldsymbol{x_1, x_2, \ldots, x_k}$, con frecuencias absolutas $\boldsymbol{n_1, n_2, \ldots, n_k}$, la desviación media se calcula con la expresión siguiente:

$$d_m = \frac{n_1 |x_1 - \bar{x}| + n_2 |x_2 - \bar{x}| + \ldots + n_k |x_k - \bar{x}|}{N} = \frac{\sum_{i=1}^{k} n_i |x_i - \bar{x}|}{N}$$

Ejemplo.- Los siguientes datos corresponden al número de faltas de ortografía cometidas por dos alumnos en siete dictados.

Alumno A: 1, 2, 5, 5, 5, 8, 9 Alumno B: 4, 4, 5, 5, 5, 6, 6

Hallamos la desviación media de ambas series de datos:

$$\bar{x}_A = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{1+2+5+5+5+8+9}{7} = \frac{35}{7} = 5 \text{ faltas}$$

$$d_{m\,A} = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N} = \frac{|1-5|+|2-5|+|5-5|+|5-5|+|5-5|+|8-5|+|9-5|}{7} = \frac{14}{7} = 2 \text{ faltas}$$

Para la otra serie, agrupamos los datos en una tabla:

x_i	n_i	$n_i x_i$	$\lvert x_i - \bar{x} \rvert$	$n_i \lvert x_i - \bar{x} \rvert$
4	2	8	1	2
5	3	15	0	0
6	2	12	1	2
Total	**7**	**35**		**4**

$$\bar{x}_B = \frac{\sum_{i=1}^{k} n_i x_i}{N} = \frac{35}{7} = 5 \text{ faltas} \Rightarrow d_{m\,B} = \frac{\sum_{i=1}^{k} n_i |x_i - \bar{x}|}{N} = \frac{4}{7} = 0'5714 \text{ faltas}$$

Observamos que, aunque ambos tienen igual media aritmética, el número de faltas de ortografía está menos disperso en el segundo alumno (es decir, el alumno B es mucho más regular que el alumno A).

3.3. Varianza y desviación típica

Otro parámetro estadístico importante es el que mide la dispersión a partir de los cuadrados de las desviaciones.

Varianza de una variable estadística es la media aritmética de los cuadrados de las desviaciones de todos los datos respecto a su media aritmética. Se representa por σ^2.

- Si la variable toma los N valores $x_1, x_2, \ldots, x_N$ (datos sin frecuencia) la varianza se puede calcular mediante alguna de las siguientes expresiones:

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2}{N} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$$

o bien mediante $$\sigma^2 = \frac{x_1^2 + x_2^2 + \ldots + x_N^2}{N} - \bar{x}^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2$$

- Si la variable toma los valores o marcas de clase $x_1, x_2, \ldots, x_k$, con frecuencias absolutas $n_1, n_2, \ldots, n_k$, la varianza se calcula mediante las expresiones siguientes:

$$\sigma^2 = \frac{n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \ldots + n_k(x_k - \bar{x})^2}{N} = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}{N}$$

o bien mediante $\boxed{\sigma^2 = \frac{n_1 x_1^2 + n_2 x_2^2 + \ldots + n_k x_k^2}{N} - \bar{x}^2 = \frac{\sum_{i=1}^{k} n_i x_i^2}{N} - \bar{x}^2}$

Desviación típica de una variable estadística es la raíz cuadrada positiva de la varianza. Se denota por **σ.**

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

3.4. Coeficiente de variación

Una dispersión de un metro en la medida de una longitud de 100 km es muchísimo más pequeña que una dispersión de un metro en una medida de 10 km.

El cociente entre la desviación típica y la media de una variable estadística se denomina ***coeficiente de variación*** y es muy útil para comparar las dispersiones de dos variables estadísticas de diferente media o de diferente naturaleza. Se suele expresar en % y se representa por C_{var}.

$$C_{var} = \frac{\sigma}{\bar{x}}$$

La media $\bar{x}$, así como la desviación típica σ se expresan en la misma unidad que la variable *X*. El coeficiente de variación es una cantidad sin dimensión, independientemente de las unidades elegidas.

Ejemplo.- ¿Qué serie de números te parece más dispersa de las siguientes?

Serie A: 1, 3, 5, 7, 9 Serie B: 1, 4, 8, 8

Hallamos la media y la desviación típica de ambas series para calcular sus coeficientes de variación:

Serie A: $\bar{x}_A = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{1+3+5+7+9}{5} = \frac{25}{5} = 5$

$$\sigma_A^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2 = \frac{1^2+3^2+5^2+7^2+9^2}{5} - 5^2 = \frac{165}{5} - 25 = 8\,;$$

luego $\sigma_A = \sqrt{8} \cong 2'8284$

Serie B: $\bar{x}_B = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{1+4+8+8}{4} = \frac{21}{4} = 5'25$

$$\sigma_B^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2 = \frac{1^2+4^2+8^2+8^2}{4} - 5'25^2 = \frac{145}{4} - 27'5625 = 8'6875\,;$$

con lo que $\sigma_B = \sqrt{8'6875} \cong 2'9475$

Con los datos anteriores, los coeficientes de variación de las respectivas series son:

$$C_{\text{var}\,A} = \frac{\sigma_A}{\bar{x}_A} = \frac{2'8284}{5} \cdot 100 = 56'57\ \% \qquad C_{\text{var}\,B} = \frac{\sigma_B}{\bar{x}_B} = \frac{2'9475}{5'25} \cdot 100 = 56'14\ \%$$

Por tanto, la segunda serie es algo menos dispersa que la primera (aunque tenga mayor desviación típica).

Ejemplo.- Analicemos los parámetros de dispersión de las distribuciones estadísticas vistas anteriormente relativas a las calificaciones de los cursos 3º A y 3º B.

- **Distribución estadística de 3º B**

x_i	n_i	$n_i x_i$	$\lvert x_i - \bar{x} \rvert$	$n_i \lvert x_i - \bar{x} \rvert$	$(x_i - \bar{x})^2$	$n_i (x_i - \bar{x})^2$
0	4	0	5'1	20'4	26'01	104'04
1	3	3	4'1	12'3	16'81	50'43
2	4	8	3'1	12'4	9'61	38'44
3	1	3	2'1	2'1	4'41	4'41
4	1	4	1'1	1'1	1'21	1'21
5	3	15	0'1	0'3	0'01	0'03
6	3	18	0'9	2'7	0'81	2'43
7	1	7	1'9	1'9	3'61	3'61
8	1	8	2'9	2'9	8'41	8'41
9	3	27	3'9	11'7	15'21	45'63
10	6	60	4'9	29'4	24'01	144'06
Total	**30**	**153**		**97'2**		**402'7**

$$\bar{x}_B = \frac{\sum_{i=1}^{k} n_i x_i}{N} = \frac{153}{30} = 5'1 \text{ puntos} \qquad \sigma_B^2 = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}{N} = \frac{402'7}{30} = 13'42$$

$$R_B = 10 - 0 = 10 \text{ puntos} \qquad \sigma_B = \sqrt{\sigma^2} = \sqrt{13'42} = 3'66 \text{ puntos}$$

$$d_{m\,B} = \frac{\sum_{i=1}^{k} n_i \lvert x_i - \bar{x} \rvert}{N} = \frac{97'2}{30} = 3'24 \text{ puntos} \qquad C_{\text{var}\,B} = \frac{\sigma_B}{\bar{x}_B} = \frac{3'66}{5'1} \cdot 100 = 71'76\ \%$$

- **Distribución estadística de 3º A**

x_i	n_i	$n_i x_i$	$\lvert x_i - \bar{x} \rvert$	$n_i \lvert x_i - \bar{x} \rvert$	$n_i x_i^2$
1	1	1	4'03	4'03	1
2	2	4	3'03	6'06	8
3	3	9	2'03	6'09	27
4	4	16	1'03	4'12	64
5	10	50	0'03	0'30	250
6	4	24	0'97	3'88	144
7	2	14	1'97	3'94	98
8	3	24	2'97	8'91	192
9	1	9	3'97	3'97	81
Total	**30**	**151**		**41'30**	**865**

$$\bar{x}_A = \frac{\sum_{i=1}^{k} n_i x_i}{N} = \frac{151}{30} = 5'03 \text{ puntos} \qquad \sigma_A^2 = \frac{\sum_{i=1}^{k} n_i x_i^2}{N} - \bar{x}^2 = \frac{865}{30} - 5'03^2 = 3'53$$

$$R_A = 9 - 1 = 8 \text{ puntos} \qquad \sigma_A = \sqrt{\sigma^2} = \sqrt{3'53} = 1'89 \text{ puntos}$$

$$d_{m\,A} = \frac{\sum_{i=1}^{k} n_i \lvert x_i - \bar{x} \rvert}{N} = \frac{41'30}{30} = 1'38 \text{ puntos} \qquad C_{\text{var}\,A} = \frac{\sigma_A}{\bar{x}_A} = \frac{1'89}{5'03} \cdot 100 = 37'57\ \%$$

Conclusiones:

Parámetros de dispersión	*3º A*	*3º B*
Rango	8	10
Desviación media	1'38	3'24
Desviación típica	1'89	3'66
Coeficiente de variación	37'57 %	71'76 %

Como podemos apreciar, todos los parámetros de dispersión del grupo 3º A son menores que los del grupo 3º B, incluido el coeficiente de variación. Por tanto, podemos afirmar rotundamente que la distribución estadística del grupo 3º A es menos dispersa que la de 3º B.

EJERCICIOS

9. Las calificaciones de Juan en seis pruebas fueron: 87, 64, 92, 86, 69 y 71. Halla la media, la mediana y todos los parámetros de dispersión.

10. Fíjate que para hallar la varianza hay que elevar al cuadrado las desviaciones respecto a la media; por ello, la varianza no se expresa en las mismas unidades que los datos. De manera que si los datos se expresan en metros, ¿en qué unidades se expresará la varianza? ¿Y la desviación típica y el coeficiente de variación?

11. Los siguientes datos son calificaciones obtenidas en cierto examen de Lengua.

    2, 5, 3, 4, 7, 9, 5, 2, 7, 4, 8, 3, 5, 8, 7, 9, 3, 2, 4, 1, 10, 9, 4, 8, 6, 9, 3, 3, 7, 1, 2, 8, 6, 7, 3, 6, 4, 7, 4, 8, 2, 3, 7, 5, 4, 6, 7, 5, 6, 7, 8, 4, 3, 7, 5, 6, 9, 5, 7, 2

    a) Elabora una tabla en la que aparezcan las diferentes frecuencias simples.
    b) Calcula los parámetros de centralización de las calificaciones.
    c) Calcula todos los parámetros de dispersión.

12. En la fabricación de cierto tipo de bombillas se han detectado algunas defectuosas. Se han estudiado 200 lotes de 500 piezas cada uno, obteniéndose los datos de la tabla adjunta.

Defectuosas	1	2	3	4	5	6	7	8
Nº de lotes	5	15	38	42	49	32	17	2

    Calcula los parámetros de centralización y de dispersión.

13. En un hospital se quiere estimar el peso de los niños recién nacidos. Para ello se seleccionan, de forma aleatoria, 100 de éstos, obteniéndose los siguientes resultados.

***Peso* (kg)**	[1, 1'5)	[1'5, 2)	[2, 2'5)	[2'5, 3)	[3, 3'5)	[3'5, 4)	[4, 4'5)	[4'5, 5]
Nº de niños	1	2	5	20	40	26	5	1

    a) Calcula los pesos medio, mediano y moda de la distribución anterior.
    b) Determina el rango, la desviación media y la desviación típica de la variable.

14. Si has realizado los ejercicios 12 y 13 anteriores podrás comprobar que las desviaciones típicas son, respectivamente, 1'5254 y 0'5679. ¿Cuál de las dos distribuciones es menos dispersa?

15. Si a los números 10, 12, 14, 16, 18 y 20, los multiplicamos por 4 se obtiene 40, 48, 56, 64, 72 y 80. ¿Qué puedes decir de las medias, las varianzas y las desviaciones típicas de ambas series estadísticas?

16. Si a los números 10, 12, 14, 16, 18 y 20, les sumamos 9 se obtiene 19, 21, 23, 25, 27 y 29. Compara las medias, las varianzas y las desviaciones típicas de ambas series estadísticas.

4. Estudio conjunto de media y desviación típica: distribuciones normales

Las medidas de centralización media, mediana y moda a veces coinciden. Cuando esto ocurre se dice que la distribución es **simétrica.** Su gráfico, bien diagrama de barras o histograma, toma la forma que puedes ver a continuación, en donde las frecuencias correspondientes a valores de la variable equidistantes de un valor central son iguales.

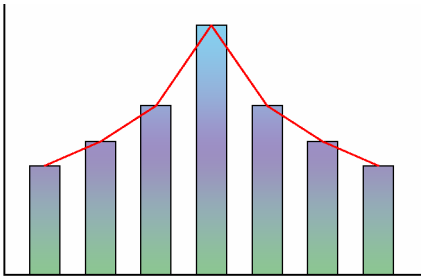
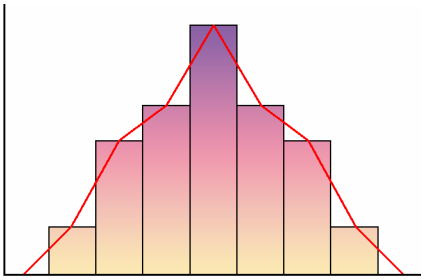


Gráfico de barras de una distribución simétrica



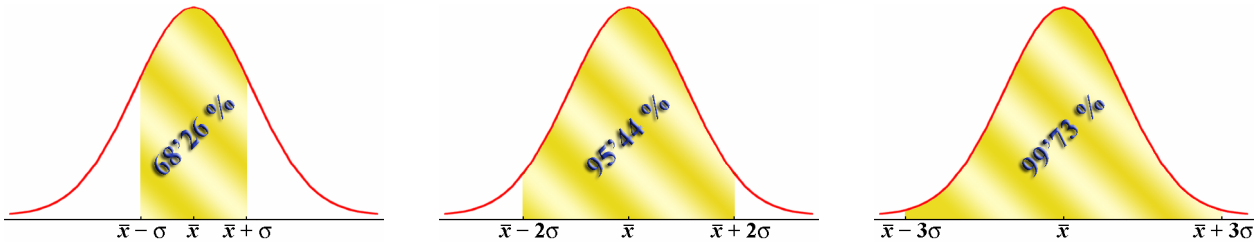
Histograma simétrico

La representación gráfica simétrica más conocida es la ***campana de Gauss,*** que corresponde a una ***distribución normal*** llamada así porque muchos fenómenos se distribuyen de esta manera. El punto más alto corresponde a la media aritmética, siendo los valores centrales más frecuentes que los alejados, cuya frecuencia disminuye. La media aritmética y la desviación típica son los parámetros estadísticos más utilizados. En toda distribución estadística, el estudio del comportamiento conjunto de estos parámetros nos aporta numerosa información sobre la distribución de frecuencias estudiada. Que la campana se encuentre más o menos aplastada depende del valor de la desviación típica σ. Cuando la campana es muy puntiaguda es porque hay poca dispersión, y cuando está muy aplastada la dispersión es mucho mayor.

Teorema de Chebyshev

En una distribución normal se considera que el 100 % de los datos es el área comprendida en la campana.

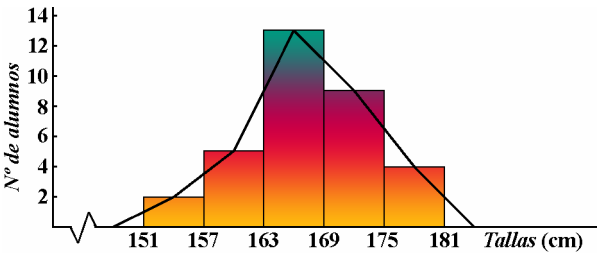
- En el intervalo $(\bar{x}-\sigma,\ \bar{x}+\sigma)$ se encuentra el 68'26 % del total de los datos.
- En el intervalo $(\bar{x}-2\sigma,\ \bar{x}+2\sigma)$ se encuentra el 95'44 % del total de los datos.
- En el intervalo $(\bar{x}-3\sigma,\ \bar{x}+3\sigma)$ se encuentra el 99'73 % del total de los datos.



Campana de Gauss y Teorema de Chebyshev

Ejemplo.- La siguiente tabla refleja la estatura, en centímetros, de 33 alumnos.

***Talla* (cm)**	x_i	n_i	$n_i x_i$	$n_i x_i^2$
[151, 157)	154	2	308	47.432
[157, 163)	160	5	800	128.000
[163, 169)	166	13	2.158	358.228
[169, 175)	172	9	1.548	266.256
[175, 181]	178	4	712	126.736
Total		**33**	**5.526**	**926.652**



Observando el histograma y polígono de frecuencias observamos que la distribución es unimodal y bastante simétrica. Calculamos la media y la desviación típica:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{N} = \frac{5.526}{33} = 167'45 \text{ cm}$$

$$\sigma^2 = \frac{\sum_{i=1}^{k} n_i x_i^2}{N} - \bar{x}^2 = \frac{926.652}{33} - 167'45^2 = 39'34 \Rightarrow \sigma = \sqrt{39'34} = 6'27 \text{ cm}$$

Vamos a calcular el porcentaje de personas con estaturas en los siguientes intervalos:

- En el intervalo $(\bar{x} - \sigma,\ \bar{x} + \sigma)$ = (161'18, 173'72) hay 13 + 9 = 22 individuos, que representan al 66'67 % de los datos.
- En el intervalo $(\bar{x} - 2\sigma,\ \bar{x} + 2\sigma)$ = (154'91, 179'99) hay 5 + 13 + 9 + 4 = 31 individuos, que representan al 93'94 % de los datos.
- En el intervalo $(\bar{x} - 3\sigma,\ \bar{x} + 3\sigma)$ = (148'64, 186'26) están los 33 individuos, que representan lógicamente al 100 % de los datos.

Los resultados del ejemplo anterior se dan de forma parecida en distribuciones estadísticas bastante simétricas respecto de un valor central de la variable estadística. Decimos que estas distribuciones tienen un ***comportamiento normal.***

Ejemplo.- Cierto Ayuntamiento va a construir un parque y quiere contar con la opinión de los vecinos. A una muestra de éstos se les ha preguntado sobre el grado de aceptación del proyecto y los resultados han sido:

Aceptación **(x_i)**	1	2	3	4	5	6	7	8	9
Frecuencia **(n_i)**	1	3	15	25	30	24	16	2	1

El ayuntamiento se pregunta por el comportamiento normal de las respuestas, por lo que debemos hallar el número de casos que hay en cada intervalo $(\bar{x} - \sigma,\ \bar{x} + \sigma)$, $(\bar{x} - 2\sigma,\ \bar{x} + 2\sigma)$ y $(\bar{x} - 3\sigma,\ \bar{x} + 3\sigma)$.

x_i	n_i	$n_i x_i$	$n_i x_i^2$
1	1	1	1
2	3	6	12
3	15	45	135
4	25	100	400
5	30	150	750
6	24	144	864
7	16	112	784
8	2	16	128
9	1	9	81
Total	**117**	**583**	**3.155**

Hallamos la media aritmética y la desviación típica:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{N} = \frac{583}{117} = 4'983 \text{ puntos}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} n_i x_i^2}{N} - \bar{x}^2} = \sqrt{\frac{3.155}{117} - 4'983^2} =$$

$$= \sqrt{2'136} = 1'462 \text{ puntos}$$

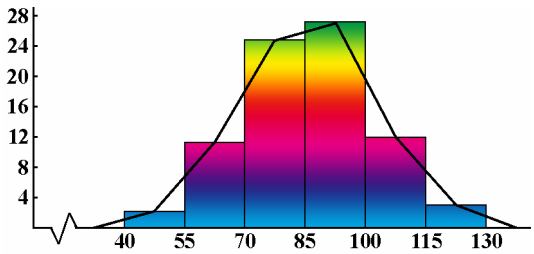
- En el primer intervalo (3'521, 6'445) hay 25 + 30 + 24 = 79 casos, que suponen el 67'52 % del total.
- En el segundo intervalo (2'059, 7'907) hay 15 + 25 + 30 + 24 + 16 = 110 casos, esto es, el 94'02 % del total.
- En el tercer intervalo (0'597, 9'369) hay 117 casos, que suponen el 100 % del total.

Se observa que los porcentajes obtenidos se corresponden, aproximadamente, con lo que hemos denominado como comportamiento normal.

Ejemplo.- En una consulta médica se ha medido durante una jornada la frecuencia cardiaca de 80 personas, en latidos por minuto, y se han obtenido los resultados recogidos en la tabla. Estudia el comportamiento normal de esta distribución estadística.

Intervalo	[40, 55)	[55, 70)	[70, 85)	[85, 100)	[100, 115)	[115, 130]
x_i	47'5	62'5	77'5	92'5	107'5	122'5
n_i	2	11	25	27	12	3

Representamos el polígono de frecuencias correspondiente a la distribución estadística *"latidos por minuto"*, el cual presenta bastante simetría.



Los parámetros estadísticos de la frecuencia cardiaca, una vez calculados, son:

$$\bar{x} = 85'94 \text{ latidos/min} \quad \text{y} \quad \sigma = 16'40 \text{ latidos/min}$$

Veamos cómo se distribuyen los pacientes en los intervalos mencionados:

- En el intervalo $(\bar{x} - \sigma,\ \bar{x} + \sigma) = (69'54, 102'34)$ hay 25 + 27 = 52 individuos, que representan al 65 % del total.
- En el intervalo $(\bar{x} - 2\sigma,\ \bar{x} + 2\sigma) = (53'14, 118'74)$ hay 11 + 25 + 27 + 12 = 75 individuos, que representan al 93'75 % del total.
- Por último, en el intervalo $(\bar{x} - 3\sigma,\ \bar{x} + 3\sigma) = (36'74, 135'14)$ se encuentran 80 que, lógicamente, son el 100 %.

Los porcentajes obtenidos nos indican que esta distribución estadística tiene, en buena medida, un comportamiento normal.

4.1. Variables normalizadas

Para poder comparar dos datos correspondientes a dos distribuciones distintas, hay que ***tipificar*** o ***normalizar*** dichos valores, es decir, calcular los valores $\boldsymbol{z = \dfrac{x - \bar{x}}{\sigma}}$ y, después, comparar los resultados.

Ejemplo.- En una prueba de Idioma, Juan obtiene una nota de 7 y en el conjunto de la clase se tiene $\bar{x} = 5'5$ y $\sigma = 2$. En otra prueba saca 6'8, siendo las calificaciones de la clase $\bar{x} = 6$ y $\sigma = 1$. ¿Cuál de las calificaciones es mejor respecto de la clase?

Para poder comparar ambas calificaciones, lo hacemos a través de sus correspondientes puntuaciones normalizadas.

Para $x = 7$, tenemos: $z_7 = \dfrac{7 - 5'5}{2} = 0'75$ puntos

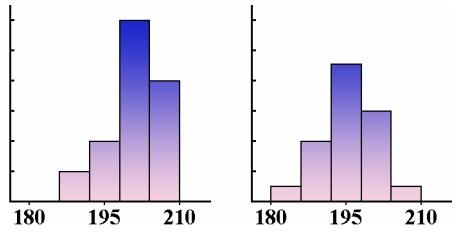
Para $x = 6'8$, tenemos: $z_{6'8} = \dfrac{6'8 - 6}{1} = 0'8$ puntos

Al ser mayor la segunda puntuación normalizada, la segunda calificación puede considerarse mejor que la primera.

EJERCICIOS

17. En una distribución intervienen 600 personas. Se sabe que es unimodal y bastante simétrica. Se tiene que la media aritmética es 50 y la desviación típica es 7. ¿Cuántas personas se distribuirán en el intervalo (43, 57)?

18. En el gráfico están representadas las distribuciones de la variable estadística talla (en centímetros) de dos equipos A y B de baloncesto. Uno de los equipos tiene $\bar{x}_A = 199$ y $\sigma_A = 4$; el otro $\bar{x}_B = 193'5$ y $\sigma_B = 4'5$.



a) Asocia cada uno de estos gráficos al equipo correspondiente. Razónalo.
b) Un nuevo jugador con una talla de 205 cm, ¿en cuál de los dos equipos sería «más» alto?

19. Se desea comparar la duración de dos marcas de lámparas halógenas. Para ello, elegimos dos muestras, compuestas por 10 lámparas de cada una de las marcas. La duración en semanas de cada una de ellas se refleja a continuación.

Marca A	23	26	24	32	28	26	22	25	20	21
Marca B	22	29	24	27	30	29	25	27	22	30

a) Calcula la media y la desviación típica de las duraciones de cada marca de lámparas.
b) ¿Qué marca sería aconsejable elegir? ¿Cuál de las dos distribuciones tiene menor dispersión?

20. Una fábrica de yogures empaqueta éstos en cajas de cien unidades cada una. Para probar la eficacia de la producción se han analizado 80 cajas comprobando los yogures defectuosos que contiene cada una y se han obtenido los resultados de la tabla.

Nº de yogures defectuosos	0	1	2	3	4	5	6
Nº de cajas	40	15	10	9	3	2	1

Define cuáles son los individuos de esta muestra y la variable estadística. Después calcula para esta distribución estadística los parámetros de centralización y los parámetros de dispersión. Por último, representa gráficamente la distribución y halla el número de cajas que están en los intervalos $(\bar{x} - \sigma,\ \bar{x} + \sigma)$, $(\bar{x} - 2\sigma,\ \bar{x} + 2\sigma)$ y $(\bar{x} - 3\sigma,\ \bar{x} + 3\sigma)$. A la vista de los resultados, ¿puede calificarse la distribución de normal?

21. El presupuesto del Insalud, por Comunidades Autónomas y en miles de millones de pesetas, del año 1992 fue el siguiente.

Comunidad	Cataluña	Navarra	Andalucía	Galicia	C. Valenciana	País Vasco	Gestión directa
Presupuesto	379	29	417	146	243	132	1.040

a) Construye el diagrama de sectores correspondiente a esta distribución de frecuencias.
b) ¿Qué datos de los anteriores se encuentran en el intervalo $(\bar{x} - \sigma,\ \bar{x} + \sigma)$?

22. Durante el mes de julio, en una determinada ciudad, se han registrado las siguientes temperaturas máximas.

32, 31, 28, 29, 29, 33, 32, 31, 30, 31, 31, 27, 28, 29, 29, 30, 32, 31, 31, 30, 30, 29, 29, 30, 30, 31, 30, 31, 34, 33, 33

a) Halla los parámetros de centralización.
b) Calcula el rango y la desviación típica.
c) Comprueba si en el intervalo $(\bar{x} - 2\sigma,\ \bar{x} + 2\sigma)$ se encuentra aproximadamente el 95'44 % de los datos.

23. De una muestra de 75 pilas eléctricas, se han obtenido los datos de la tabla adjunta sobre su duración en horas.

***Duración* (horas)**	[25, 30)	[30, 35)	[35, 40)	[40, 45)	[45, 50)	[50 ,55]
Nº de pilas	3	5	21	28	12	6

a) Realiza la representación gráfica de la distribución.
b) Calcula la media y la desviación típica.
c) ¿Qué porcentaje de pilas tienen su duración comprendida en los intervalos $(\bar{x} - \sigma,\ \bar{x} + \sigma)$, $(\bar{x} - 2\sigma,\ \bar{x} + 2\sigma)$ y $(\bar{x} - 3\sigma,\ \bar{x} + 3\sigma)$? ¿Puede considerarse que la distribución tiene un comportamiento normal?

5. Cálculo estadístico

5.1. Estadística y calculadora

Recordamos, una vez más, que cada modelo de **calculadora** tiene una forma particular de actuación, tanto en la introducción de datos y operaciones como en la realización de los cálculos. Por ello, se debe consultar y leer con atención el manual de instrucciones.

La mayoría de las calculadoras tiene funciones estadísticas y en estos casos en el teclado figuran los símbolos:

$$\sum x \qquad \sum x^2 \qquad N \qquad \bar{x} \qquad \sigma_n$$

5.2. Estadística y hojas de cálculo

La Estadística, que por su propia naturaleza maneja una gran cantidad de datos, ha encontrado un aliado muy valioso en la Informática, ya que los ordenadores son capaces de realizar, sin equivocarse, miles de operaciones en un segundo.

Aparte de muchos programas especializados en Estadística, una de las aplicaciones más utilizadas en los cálculos estadísticos es la **hoja de cálculo,** que permite introducir los datos de una distribución y obtener rápidamente los parámetros de una forma sencilla y elegante.

Cada hoja de cálculo se presenta como una pantalla cuadriculada formada por **filas** (horizontales) y **columnas** (verticales). La intersección de una fila con una columna se llama **celda.** Cada celda se identifica con la letra de su columna y el número de su fila (A1, D5, …).

Para introducir una información en una celda primero hay que activarla y luego escribir el **texto, dato** o **fórmula** que nos interesa. Las fórmulas son operaciones matemáticas que se realizan con los datos de otras celdas. Cuando se activa una celda que contiene una fórmula, en ella se ve el resultado de la operación correspondiente, pero en la parte superior de la hoja se visualiza su expresión.

Calcular la puntuación media, la varianza y la desviación típica de la distribución obtenida al aplicar un test a 88 alumnos, cuyos resultados se muestran en la tabla siguiente.

Puntuaciones	[38, 44)	[44, 50)	[50, 56)	[56, 62)	[62, 68)	[68, 74)	[74, 80]
Nº de alumnos	7	8	15	25	18	9	6

La pantalla que nos muestra la hoja de cálculo es similar a la siguiente:

	A	B	C	D	E	F	G	H	I
1	***Puntuaciones***		x_i	n_i	$n_i x_i$	$\lvert x_i - \bar{x} \rvert$	$n_i \lvert x_i - \bar{x} \rvert$	x_i^2	$n_i x_i^2$
2	38	44	41	7	287	18,13636	126,9545	1.681	11.767
3	44	50	47	8	376	12,13636	97,0909	2.209	17.672
4	50	56	53	15	795	6,13636	92,0455	2.809	42.135
5	56	62	59	25	1.475	0,13636	3,4091	3.481	87.025
6	62	68	65	18	1.170	5,86364	105,5455	4.225	76.050
7	68	74	71	9	639	11,86364	106,7727	5.041	45.369
8	74	80	77	6	462	17,86364	107,1818	5.929	35.574
10			**Total**	**88**	**5.204**		**639**		**315.592**
11									
12		**Media =**	59,1364	puntos			**Varianza =**	89,1632	
13		**Rango =**	42	puntos			**Desv. típica =**	9,4426	puntos
14		**Desv. media =**	7,2614	puntos			**Coef. de variación =**	0,1597	

Soluciones a los ejercicios propuestos

Los cálculos necesarios para la resolución de los ejercicios han sido elaborados con una hoja de cálculo tomando redondeos con distintas cifras decimales. Si has usado calculadora para ello es probable que, en algunos casos, no coincidan exactamente los resultados obtenidos.

1. La temperatura que ha marcado un termómetro en los diferentes días de la semana, ha sido (en grados centígrados) los que pueden verse en la tabla.

	Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
Mínima	4	−2	−3	1	4	0	3
Máxima	19	18	21	13	12	14	22

a) Calcula la temperatura media mínima.
b) Calcula la temperatura media máxima.
c) Calcula la media de las oscilaciones extremas diarias.

Solución.-

a) $\bar{x}_{mínima} = \dfrac{\sum_{i=1}^{N} x_i}{N} = \dfrac{4+(-2)+(-3)+1+4+0+3}{7} = \dfrac{7}{7} = 1$ grado centígrado.

b) $\bar{x}_{máxima} = \dfrac{\sum_{i=1}^{N} x_i}{N} = \dfrac{19+18+21+13+12+14+22}{7} = \dfrac{119}{7} = 17$ grados centígrados.

c) Hallemos las oscilaciones extremas diarias, es decir, la diferencia entre la *máxima* y la *mínima* diaria.

	Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
Oscilaciones	15	20	24	12	8	14	19

$\bar{x}_{oscilaciones} = \dfrac{\sum_{i=1}^{N} x_i}{N} = \dfrac{15+20+24+12+8+14+19}{7} = \dfrac{112}{7} = 16$ grados centígrados.

Observa que $\bar{x}_{oscilaciones} = \bar{x}_{máxima} - \bar{x}_{mínima} = 17 - 1 = 16$ grados centígrados.

2. Dada la distribución estadística siguiente: 3, 2, 5, 7, 6, 4, 2, 1, 9, 5, 7, 6, 4. Calcula la media aritmética, la moda, la mediana y los cuartiles.

Solución.-

Ordenemos los 13 datos de forma creciente: 1, 2, 2, 3, 4, 4, 5, 5, 6, 6, 7, 7, 9. Tenemos entonces:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{1+2+2+3+4+4+5+5+6+6+7+7+9}{13} = \frac{61}{13} = 4'6923$$

La distribución tiene cinco modas: $Mo = 2$, $Mo = 4$, $Mo = 5$, $Mo = 6$ y $Mo = 7$.

Mediana y cuartiles:

1 2 2 3 4 4 5 5 6 6 7 7 9

$Q_1 = 3$ $Q_2 = Me = 5$ $Q_3 = 6$

3. Halla la media, la mediana, la moda y los cuartiles de la distribución cuya tabla de frecuencias es la siguiente.

x_i	3	6	7	8	10	12
n_i	6	9	7	8	17	13

Solución.-

x_i	n_i	N_i	$n_i x_i$
3	6	6	18
6	9	15	54
7	7	22	49
8	8	30	64
10	17	47	170
12	13	60	156
Total	**60**		**511**

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{N} = \frac{511}{60} = 8'52$$

Se trata de una distribución unimodal, siendo $Mo = 10$

$N/4 = 15$; precisamente el valor $x = 6$ de la variable tiene por frecuencia absoluta acumulada 15, por lo que el primer cuartil es $Q_1 = (6 + 7)/2 = 6'5$

De forma análoga, $N/2 = 30 \Rightarrow Me = Q_2 = (8 + 10)/2 = 9$

$3N/4 = 45$ luego $Q_3 = 10$, por ser éste el primer valor de la variable cuya frecuencia absoluta acumulada, 47, es mayor que las tres cuartas partes del número total de datos.

4. Las edades de los componentes de una peña de aficionados al fútbol son:

18, 16, 21, 20, 18, 16, 21, 18, 21, 18, 20, 19, 36, 24, 18, 20, 18, 19, 20

a) Calcula la edad media, la edad moda y la edad mediana, así como los cuartiles.
b) Representa gráficamente los datos de esta distribución.

Solución.-

a) Construyamos la correspondiente tabla de frecuencias:

x_i	n_i	N_i	$n_i x_i$
16	2	2	32
18	6	8	108
19	2	10	38
20	4	14	80
21	3	17	63
24	1	18	24
36	1	19	36
Total	**19**		**381**

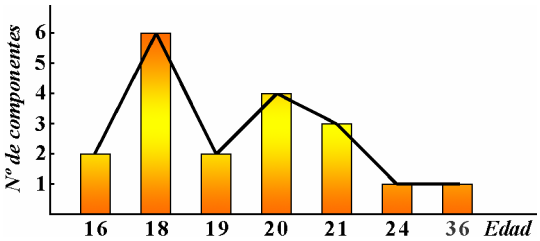
$$\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{N} = \frac{381}{19} = 20'05 \text{ años.}$$

La edad moda es $Mo = 18$ (distribución unimodal) ya que es la edad que tienen un mayor número de componentes.

$N/2 = 9'5 \Rightarrow$ la edad mediana es $Me = Q_2 = 19$ (primer valor de la variable cuya frecuencia absoluta acumulada, 10, es mayor que la mitad del número de individuos).

$N/4 = 4'75 \Rightarrow Q_1 = 18$; $3N/4 = 14'25 \Rightarrow Q_3 = 21$

b) Representamos la distribución mediante un diagrama de barras y polígono de frecuencias:



5. La siguiente tabla muestra la distribución, a lo largo de un mes, del número de camiones que circulan diariamente por un cruce de carreteras.

Nº de camiones por día	[350, 400)	[400, 450)	[450, 500)	[500, 550)	[550, 600]
Nº de días	2	5	11	9	4

Calcula la media, la moda, la mediana y los cuartiles de esta distribución.

Solución.-

Nº de camiones	x_i	n_i	N_i	$n_i x_i$
[350, 400)	375	2	2	750
[400, 450)	425	5	7	2.125
[450, 500)	475	11	18	5.225
[500, 550)	525	9	27	4.725
[550, 600]	575	4	31	2.300
Total		**31**		**15.125**

El valor de la media es $\bar{x} = \dfrac{\sum_{i=1}^{k} n_i x_i}{N} = \dfrac{15.125}{31} = 487'90$ camiones por día.

El intervalo modal es [450, 500), siendo 475 camiones por día el valor aproximado de la moda. Su valor exacto es:

$$Mo = e_i + \frac{n_{Mo} - n_{Mo-1}}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})} \cdot a = 450 + \frac{11-5}{(11-5)+(11-9)} \cdot 50 = 450 + 37'5 = 487'5 \text{ camiones por día.}$$

$N/4 = 7'75$, luego el intervalo correspondiente al primer cuartil es [450, 500); el valor aproximado del primer cuartil es 475 camiones por día. Hallamos su valor exacto:

$$Q_1 = e_i + \frac{\frac{N}{4} - N_{Q_1-1}}{n_{Q_1}} \cdot a = 450 + \frac{\frac{31}{4} - 7}{11} \cdot 50 = 450 + 3'41 = 453'41 \text{ camiones por día.}$$

También [450, 500) es el intervalo o clase mediana, ya que es el primer intervalo cuya frecuencia absoluta acumulada, 18, sobrepasa a la mitad del número de individuos, $N/2 = 15'5$; el valor aproximado de la mediana es 475 camiones por día. Hallamos, no obstante, la mediana exacta de la distribución:

$$Me = Q_2 = e_i + \frac{\frac{N}{2} - N_{Me-1}}{n_{Me}} \cdot a = 450 + \frac{\frac{31}{2} - 7}{11} \cdot 50 = 450 + 38'64 = 488'64 \text{ camiones por día.}$$

$3N/4 = 23'25$, luego [500, 550) es el intervalo correspondiente al tercer cuartil; el valor aproximado de éste es 525 camiones por día. Hallamos su valor exacto:

$$Q_3 = e_i + \frac{\frac{3N}{4} - N_{Q_3-1}}{n_{Q_3}} \cdot a = 500 + \frac{\frac{3 \cdot 31}{4} - 18}{9} \cdot 50 = 500 + 29'17 = 529'17 \text{ camiones por día.}$$

6. Las respuestas correctas a un test de 80 preguntas realizado por 600 personas son las que se recogen a continuación.

Respuestas	[0, 10)	[10, 20)	[20, 30)	[30, 40)	[40, 50)	[50, 60)	[60, 70)	[70, 80]
Nº de personas	40	60	75	90	105	85	80	65

Calcula el número medio de respuestas correctas, la moda y la mediana. Halla los cuartiles. Interpreta gráficamente el cálculo de la moda y de la mediana, y comprueba que la mediana es el punto del eje de abscisas que divide el histograma de frecuencias absolutas en dos partes de igual área.

Solución.-

Respuestas	x_i	n_i	N_i	$n_i x_i$
[0, 10)	5	40	40	200
[10, 20)	15	60	100	900
[20, 30)	25	75	175	1.875
[30, 40)	35	90	265	3.150
[40, 50)	45	105	370	4.725
[50, 60)	55	85	455	4.675
[60, 70)	65	80	535	5.200
[70, 80]	75	65	600	4.875
Total		**600**		**25.600**

Número medio de respuestas correctas: $\bar{x} = \dfrac{\sum_{i=1}^{k} n_i x_i}{N} = \dfrac{25.600}{600} = 42'67$ respuestas.

El intervalo modal es [40, 50) y 45 respuestas el valor aproximado de la moda. El valor exacto de la misma es:

$$Mo = e_i + \frac{n_{Mo} - n_{Mo-1}}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})} \cdot a = 40 + \frac{105 - 90}{(105 - 90) + (105 - 85)} \cdot 10 = 40 + \frac{150}{35} = 44'29 \text{ respuestas.}$$

$N/4 = 150$, con lo que [20, 30) es el intervalo correspondiente al primer cuartil; el valor aproximado de éste es 25 respuestas. Hallamos su valor exacto:

$$Q_1 = e_i + \frac{\frac{N}{4} - N_{Q_1 - 1}}{n_{Q_1}} \cdot a = 20 + \frac{\frac{600}{4} - 100}{75} \cdot 10 = 20 + \frac{500}{75} = 26'67 \text{ respuestas.}$$

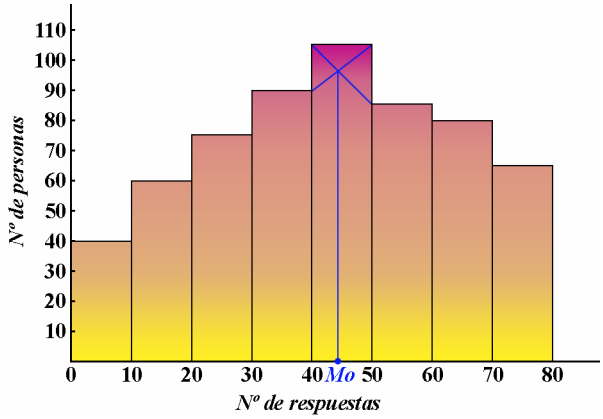
La clase mediana e [40, 50) pues es el primer intervalo cuya frecuencia absoluta acumulada, 370, sobrepasa a la mitad del número de individuos, $N/2 = 300$; el valor aproximado de la mediana es 45 respuestas, y el valor exacto es:

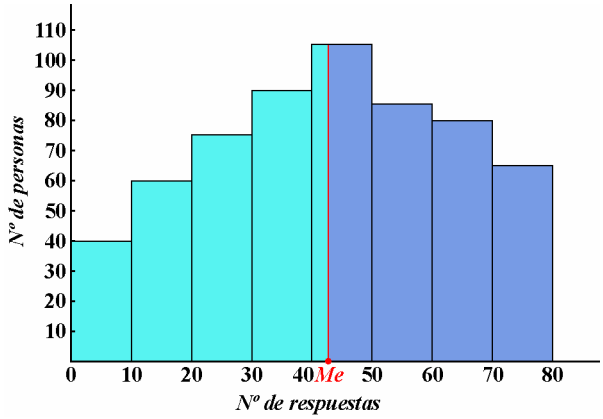
$$Me = Q_2 = e_i + \frac{\frac{N}{2} - N_{Me-1}}{n_{Me}} \cdot a = 40 + \frac{\frac{600}{2} - 265}{105} \cdot 10 = 40 + \frac{350}{105} = 43'33 \text{ respuestas.}$$

$3N/4 = 450$, luego [50, 60) es el intervalo correspondiente al tercer cuartil; el valor aproximado de éste es 55 respuestas. Hallamos su valor exacto:

$$Q_3 = e_i + \frac{\frac{3 \cdot N}{4} - N_{Q_3 - 1}}{n_{Q_3}} \cdot a = 50 + \frac{\frac{3 \cdot 600}{4} - 370}{85} \cdot 10 = 50 + \frac{800}{85} = 59'41 \text{ respuestas.}$$

Interpretación gráfica de los resultados:





Por último, comprobemos que ambas zonas en que la mediana divide al histograma tienen igual área:

Área de la izquierda: $10\cdot 40+10\cdot 60+10\cdot 75+10\cdot 90+\frac{10}{3}\cdot 105=400+600+750+900+350=3.000$

Área de la derecha: $\frac{20}{3}\cdot 105+10\cdot 85+10\cdot 80+10\cdot 65=700+850+800+650=3.000$

7. La media de x, $4x-3$, $x+4$, -16, 9 y $x-5$ es 4. ¿Cuánto vale la mediana de esta serie de números?

Solución.-

A través de la media hallamos el valor de x:

$$\bar{x}=\frac{x+(4x-3)+(x+4)+(-16)+9+(x-5)}{6}=4 \Rightarrow \frac{7x-11}{6}=4 \Rightarrow x=5$$

La serie estadística obtenida es: 5, 17, 9, −16, 9, 0.

Para hallar la mediana ordenamos los datos: −16, 0, 5, 9, 9, 17 $\Rightarrow$ como hay dos valores centrales, 5 y 9, la mediana es $Me = (5 + 9)/2 = 7$.

8. La siguiente serie de datos: 18, 21, 24, a, 36, 37, b, está ordenada y tiene de mediana 30 y de media 32. Encuentra el valor de a y b.

Solución.-

El valor central a es la mediana 30, por tanto, $a = Me = 30$. Como la media es 32, obtenemos así el valor de b:

$$\bar{x}=32 \Rightarrow \frac{18+21+24+30+36+37+b}{7}=32 \Rightarrow \frac{166+b}{7}=32 \Rightarrow b=58$$

9. Las calificaciones de Juan en seis pruebas fueron: 87, 64, 92, 86, 69 y 71. Halla la media, la mediana y todos los parámetros de dispersión.

Solución.-

Observa que son datos sin frecuencia o, equivalentemente, con frecuencia absoluta 1. Primeramente ordenamos los datos: 64, 69, 71, 86, 87, 92.

Media: $\bar{x} = \dfrac{\sum_{i=1}^{N} x_i}{N} = \dfrac{64+69+71+86+87+92}{6} = \dfrac{469}{6} = 78'17$ puntos

Como hay dos valores centrales, la mediana resulta ser $Me = (71 + 86)/2 = 78'5$ puntos

Fácilmente, el rango es $R = 92 - 64 = 28$ puntos

Calculamos la desviación media:

$$d_m = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N} = \frac{|64-78'17| + |69-78'17| + |71-78'17| + |86-78'17| + |87-78'17| + |92-78'17|}{6}$$

$$= \frac{61}{6} = 10'17 \text{ puntos}$$

La varianza la podemos hallar con dos expresiones distintas:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N} = \frac{(64-78'17)^2 + (69-78'17)^2 + (71-78'17)^2 + (86-78'17)^2 + (87-78'17)^2 + (92-78'17)^2}{6} =$$

$$= \frac{666'8334}{6} = 111'14$$

O bien utilizando la siguiente expresión:

$$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2 = \frac{64^2 + 69^2 + 71^2 + 86^2 + 87^2 + 92^2}{6} - \left(\frac{469}{6}\right)^2 = \frac{37.327}{6} - \frac{219.961}{36} = \frac{4.001}{36} = 111'14$$

Hallamos la desviación típica: $\sigma = \sqrt{\sigma^2} = \sqrt{111'14} = 10'54$ puntos

Por último, calculamos el coeficiente de variación: $C_{var} = \dfrac{\sigma}{\bar{x}} = \dfrac{10'54}{78'17} \cdot 100 = 13'48\ \%$

No obstante, si te resulta muy engorrosa esta notación, siempre puedes recurrir a organizar los datos en una tabla:

	x_i	$\lvert x_i - \bar{x} \rvert$	$(x_i - \bar{x})^2$	x_i^2
	64	14'17	200'7889	4.096
	69	9'17	84'0889	4.761
	71	7'17	51'4089	5.041
	86	7'83	61'3089	7.396
	87	8'83	77'9689	7.569
	92	13'83	191'2689	8.464
Total	**469**	**61**	**666'8334**	**37.327**

De esta forma:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{469}{6} = 78'17 \text{ puntos}$$

$$d_m = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N} = \frac{61}{6} = 10'17 \text{ puntos}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N} = \frac{666'8334}{6} = 111'14\ ;\ \text{o bien, } \sigma^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2 = \frac{37.327}{6} - \left(\frac{469}{6}\right)^2 = \frac{4.001}{36} = 111'14$$

10. Fíjate que para hallar la varianza hay que elevar al cuadrado las desviaciones respecto a la media; por ello, la varianza no se expresa en las mismas unidades que los datos. De manera que si los datos se expresan en metros, ¿en qué unidades se expresará la varianza? ¿Y la desviación típica y el coeficiente de variación?

Solución.-

Si los datos se expresan en metros, entonces la varianza se expresará en metros cuadrados.

La desviación típica es la raíz cuadrada de la varianza, por ello, ésta vendrá expresada en metros. La media aritmética se expresa también en metros, al igual que la desviación típica, pero el coeficiente de variación no se expresa en ninguna medida.

11. Los siguientes datos son calificaciones obtenidas en cierto examen de Lengua.

2, 5, 3, 4, 7, 9, 5, 2, 7, 4, 8, 3, 5, 8, 7, 9, 3, 2, 4, 1, 10, 9, 4, 8, 6, 9, 3, 3, 7, 1, 2, 8, 6, 7, 3, 6, 4, 7, 4, 8, 2, 3, 7, 5, 4, 6, 7, 5, 6, 7, 8, 4, 3, 7, 5, 6, 9, 5, 7, 2

a) Elabora una tabla en la que aparezcan las diferentes frecuencias simples.
b) Calcula los parámetros de centralización de las calificaciones.
c) Calcula todos los parámetros de dispersión.

Solución.-

a) Hallamos las frecuencias simples absolutas, relativas y porcentuales.

x_i	1	2	3	4	5	6	7	8	9	10	**Total**
n_i	2	6	8	8	7	6	11	6	5	1	**60**
f_i	0’0333	0’1	0’1333	0’1333	0’1167	0’1	0’1833	0’1	0’0833	0’0167	**1**
p_i	3’33	10	13’33	13’33	11’67	10	18’33	10	8’33	1’67	**100**

b) Aprovechamos la siguiente tabla, en la que aparecen también los cálculos necesarios para el siguiente apartado.

Nota: Hemos hallado los datos necesarios para calcular la varianza usando las dos expresiones estudiadas.

x_i	n_i	N_i	$n_i x_i$	$\lvert x_i - \bar{x} \rvert$	$n_i \lvert x_i - \bar{x} \rvert$	$(x_i - \bar{x})^2$	$n_i (x_i - \bar{x})^2$	x_i^2	$n_i x_i^2$
1	2	2	2	4,35	8’70	18’9225	37’8450	1	2
2	6	8	12	3,35	20’10	11’2225	67’3350	4	24
3	8	16	24	2,35	18’80	5’5225	44’1800	9	72
4	8	24	32	1,35	10’80	1’8225	14’5800	16	128
5	7	31	35	0,35	2’45	0’1225	0’8575	25	175
6	6	37	36	0,65	3’90	0’4225	2’5350	36	216
7	11	48	77	1,65	18’15	2’7225	29’9475	49	539
8	6	54	48	2,65	15’90	7’0225	42’1350	64	384
9	5	59	45	3,65	18’25	13’3225	66’6125	81	405
10	1	60	10	4,65	4’65	21’6225	21’6225	100	100
Total	**60**		**321**		**121’70**		**327’6500**		**2.045**

Media: $\bar{x} = \dfrac{\sum_{i=1}^{k} n_i x_i}{N} = \dfrac{321}{60} = 5'35$ puntos ; Moda: $Mo = 7$ puntos ; Mediana: $N/2 = 30 \Rightarrow Me = 5$ puntos

c) Hallamos ahora los parámetros de dispersión:

Rango: $R = 10 - 1 = 9$ puntos

Desviación media: $d_m = \dfrac{\sum_{i=1}^{k} n_i \lvert x_i - \bar{x} \rvert}{N} = \dfrac{121'7}{60} = 2'0283$ puntos

Varianza: $\sigma^2 = \dfrac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}{N} = \dfrac{327'65}{60} = 5'4608$; o bien, $\sigma^2 = \dfrac{\sum_{i=1}^{k} n_i x_i^2}{N} - \bar{x}^2 = \dfrac{2.045}{60} - 5'35^2 = 5'4608$

Desviación típica: $\sigma = \sqrt{\sigma^2} = \sqrt{5'4608} = 2'3368$ puntos

$C_{var} = \dfrac{\sigma}{\bar{x}} = \dfrac{2'3368}{5'35} \cdot 100 = 43'68\ \%$

12. En la fabricación de cierto tipo de bombillas se han detectado algunas defectuosas. Se han estudiado 200 lotes de 500 piezas cada uno, obteniéndose los datos de la tabla adjunta.

Defectuosas	1	2	3	4	5	6	7	8
Nº de lotes	5	15	38	42	49	32	17	2

Calcula los parámetros de centralización y de dispersión.

Solución.-

x_i	n_i	N_i	$n_i x_i$	$n_i \mid x_i - \bar{x} \mid$	$n_i x_i^2$
1	5	5	5	17’225	5
2	15	20	30	36’675	60
3	38	58	114	54’910	342
4	42	100	168	18’690	672
5	49	149	245	27’195	1.225
6	32	181	192	49’760	1.152
7	17	198	119	43’435	833
8	2	200	16	7’110	128
Total	**200**		**889**	**255**	**4.417**

Media: $\bar{x} = \dfrac{\sum_{i=1}^{k} n_i x_i}{N} = \dfrac{889}{200} = 4'445$ bombillas defectuosas

Moda : $Mo = 5$ bombillas defectuosas

Mediana: $N/2 = 100 \Rightarrow Me = (4 + 5)/2 = 4’5$ bombillas defectuosas

Rango: $R = 8 - 1 = 7$ bombillas defectuosas

Desviación media: $d_m = \dfrac{\sum_{i=1}^{k} n_i \mid x_i - \bar{x} \mid}{N} = \dfrac{255}{200} = 1'275$ bombillas defectuosas

Varianza: $\sigma^2 = \dfrac{\sum_{i=1}^{k} n_i x_i^2}{N} - \bar{x}^2 = \dfrac{4.417}{200} - 4'445^2 = 2'326975$

Desviación típica: $\sigma = \sqrt{\sigma^2} = \sqrt{2'326975} = 1'5254$ bombillas defectuosas

Coeficiente de variación: $C_{var} = \dfrac{\sigma}{\bar{x}} = \dfrac{1'5254}{4'445} \cdot 100 = 34'32\ \%$

13. En un hospital se quiere estimar el peso de los niños recién nacidos. Para ello se seleccionan, de forma aleatoria, 100 de éstos, obteniéndose los siguientes resultados.

***Peso* (kg)**	[1, 1'5)	[1'5, 2)	[2, 2'5)	[2'5, 3)	[3, 3'5)	[3'5, 4)	[4, 4'5)	[4'5, 5]
Nº de niños	1	2	5	20	40	26	5	1

a) Calcula los pesos medio, mediano y moda de la distribución anterior.
b) Determina el rango, la desviación media y la desviación típica de la variable.

Solución.-

***Peso* (kg)**	x_i	n_i	N_i	$n_i x_i$	$n_i \mid x_i - \bar{x} \mid$	$n_i x_i^2$
[1, 1'5)	1'25	1	1	1'25	1'995	1'5625
[1'5, 2)	1'75	2	3	3'50	2'990	6'1250
[2, 2'5)	2'25	5	8	11'25	4'975	25'3125
[2'5, 3)	2'75	20	28	55'00	9'900	151'2500
[3, 3'5)	3'25	40	68	130'00	0'200	422'5000
[3'5, 4)	3'75	26	94	97'50	13'130	365'6250
[4, 4'5)	4'25	5	99	21'25	5'025	90'3125
[4'5, 5]	4'75	1	100	4'75	1'505	22'5625
	Total	**100**		**324'50**	**39'720**	**1.085'2500**

a) $\bar{x} = \dfrac{\sum_{i=1}^{k} n_i x_i}{N} = \dfrac{324'5}{100} = 3'245$ kg.

$N/2 = 50 \Rightarrow$ [3, 3'5) es el intervalo mediano, siendo 3'25 el peso mediano aproximado. El valor exacto del peso mediano es:

$$Me = e_i + \frac{\frac{N}{2} - N_{Me-1}}{n_{Me}} \cdot a = 3 + \frac{\frac{100}{2} - 28}{40} \cdot 0'5 = 3 + \frac{11}{40} = 3'275 \text{ kg.}$$

[3, 3'5) es también el intervalo modal, y 3'25 el valor aproximado de la moda. Su valor exacto es:

$$Mo = e_i + \frac{n_{Mo} - n_{Mo-1}}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})} \cdot a = 3 + \frac{40 - 20}{(40 - 20) + (40 - 26)} \cdot 0'5 = 3 + \frac{10}{34} = 3'294 \text{ kg.}$$

b) $R = 5 - 1 = 4$ kg.

$$d_m = \frac{\sum_{i=1}^{k} n_i \mid x_i - \bar{x} \mid}{N} = \frac{39'72}{100} = 0'3972 \text{ kg.}$$

$$\sigma^2 = \frac{\sum_{i=1}^{k} n_i x_i^2}{N} - \bar{x}^2 = \frac{1.085'25}{100} - 3'245^2 = 0{,}322475 \Rightarrow \sigma = \sqrt{\sigma^2} = \sqrt{0'322475} = 0'5679 \text{ kg.}$$

14. Si has realizado los ejercicios 12 y 13 anteriores podrás comprobar que las desviaciones típicas son, respectivamente, 1'5254 y 0'5679. ¿Cuál de las dos distribuciones es menos dispersa?

Solución.-

Para comparar las dispersiones de dos variables estadísticas de diferente media o de diferente naturaleza se utiliza el coeficiente de variación. En los ejercicios anteriores tenemos que:

$C_{var(12)} = 34'32$ % y $C_{var(13)} = 17'50$ %, por lo que la distribución del ejercicio 13 es menos dispersa.

15. Si a los números 10, 12, 14, 16, 18 y 20, los multiplicamos por 4 se obtiene 40, 48, 56, 64, 72 y 80. ¿Qué puedes decir de las medias, las varianzas y las desviaciones típicas de ambas series estadísticas?

Solución.-

Hallemos los parámetros mencionados de dichas series estadísticas:

Serie A

	x_i	x_i^2
	10	100
	12	144
	14	196
	16	256
	18	324
	20	400
Total	**90**	**1.420**

Serie B

	x_i	x_i^2
	40	1.600
	48	2.304
	56	3.136
	64	4.096
	72	5.184
	80	6.400
Total	**360**	**22.720**

Serie A: $\bar{x}_A = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{90}{6} = 15$ $\quad \sigma_A^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}_A^2 = \frac{1.420}{6} - 15^2 = \frac{35}{3}$ $\quad \sigma_A = \sqrt{\sigma_A^2} = \sqrt{\frac{35}{3}}$

Serie B: $\bar{x}_B = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{360}{6} = 60$ $\quad \sigma_B^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}_B^2 = \frac{22.720}{6} - 60^2 = \frac{560}{3}$ $\quad \sigma_B = \sqrt{\sigma_B^2} = \sqrt{\frac{560}{3}}$

Podemos observar que: $\bar{x}_B = 60 = 4 \cdot 15 = 4 \cdot \bar{x}_A \Rightarrow \boxed{\bar{x}_B = 4 \cdot \bar{x}_A}$

$$\sigma_B^2 = \frac{560}{3} = 16 \cdot \frac{35}{3} = 4^2 \cdot \frac{35}{3} = 4^2 \cdot \sigma_A^2 \Rightarrow \boxed{\sigma_B^2 = 4^2 \cdot \sigma_A^2}$$

$$\sigma_B = \sqrt{\frac{560}{3}} = \sqrt{\frac{4^2 \cdot 35}{3}} = 4\sqrt{\frac{35}{3}} = 4 \cdot \sigma_A \Rightarrow \boxed{\sigma_B = 4 \cdot \sigma_A}$$

16. Si a los números 10, 12, 14, 16, 18 y 20, les sumamos 9 se obtiene 19, 21, 23, 25, 27 y 29. Compara las medias, las varianzas y las desviaciones típicas de ambas series estadísticas.

Solución.-

Calculamos los parámetros de la segunda serie (los de la primera figuran en el ejercicio anterior):

Serie C

	x_i	x_i^2
	19	361
	21	441
	23	529
	25	625
	27	729
	29	841
Total	**144**	**3.526**

$$\bar{x}_C = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{144}{6} = 24$$

$$\sigma_C^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}_C^2 = \frac{3.526}{6} - 24^2 = \frac{35}{3}$$

$$\sigma_C = \sqrt{\sigma_C^2} = \sqrt{\frac{35}{3}}$$

Tenemos ahora que:

$\bar{x}_C = 24 = 15 + 9 = \bar{x}_A + 9 \Rightarrow \boxed{\bar{x}_C = \bar{x}_A + 9}$

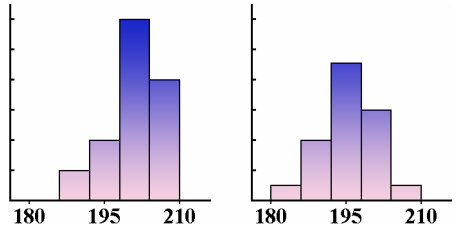
Sin embargo, las varianzas y, por tanto, las desviaciones típicas no varían: $\boxed{\sigma_C^2 = \sigma_A^2} \Rightarrow \boxed{\sigma_C = \sigma_A}$

17. En una distribución intervienen 600 personas. Se sabe que es unimodal y bastante simétrica. Se tiene que la media aritmética es 50 y la desviación típica es 7. ¿Cuántas personas se distribuirán en el intervalo (43, 57)?

Solución.-

Por ser la distribución unimodal y bastante simétrica tendrá un comportamiento normal, de ahí que en el intervalo (43, 57) = ($\bar{x} - \sigma$, $\bar{x} + \sigma$) se encuentre aproximadamente el 68'26 % del total de individuos, esto es, unas 410 personas.

18. En el gráfico están representadas las distribuciones de la variable estadística talla (en centímetros) de dos equipos A y B de baloncesto. Uno de los equipos tiene $\bar{x}_A = 199$ y $\sigma_A = 4$; el otro $\bar{x}_B = 193'5$ y $\sigma_B = 4'5$.



a) Asocia cada uno de estos gráficos al equipo correspondiente. Razónalo.
b) Un nuevo jugador con una talla de 205 cm, ¿en cuál de los dos equipos sería «más» alto?

Solución.-

a) El primer histograma se asocia al equipo A, pues los datos están más agrupados en torno al intervalo (198, 204) que contiene a la media $\bar{x}_A = 199$ y la desviación típica $\sigma_A = 4$ es menor.

b) Para hacer la comparación debemos normalizar la altura del jugador en ambos equipos; así, para $x = 205$ cm tenemos:

$$z_A = \frac{205-199}{4} = 1'5 \qquad z_B = \frac{205-193'5}{4'5} = 2'55$$

Por tanto, este jugador se consideraría más alto en el equipo B.

19. Se desea comparar la duración de dos marcas de lámparas halógenas. Para ello, elegimos dos muestras, compuestas por 10 lámparas de cada una de las marcas. La duración en semanas de cada una de ellas se refleja a continuación.

Marca A	23	26	24	32	28	26	22	25	20	21
Marca B	22	29	24	27	30	29	25	27	22	30

a) Calcula la media y la desviación típica de las duraciones de cada marca de lámparas.
b) ¿Qué marca sería aconsejable elegir? ¿Cuál de los dos distribuciones tiene menor dispersión?

Solución.-

a) Hallamos la media y la desviación típica:

Marca A

	x_i	x_i^2
	23	529
	26	676
	24	576
	32	1.024
	28	784
	26	676
	22	484
	25	625
	20	400
	21	441
Total	**247**	**6.215**

Marca B

x_i	x_i^2
22	484
29	841
24	576
27	729
30	900
29	841
25	625
27	729
22	484
30	900
Total	**265**	**7.109**

Marca A: $\bar{x}_A = \dfrac{\sum_{i=1}^{N} x_i}{N} = \dfrac{247}{10} = 24'7$ semanas; $\sigma_A = \sqrt{\dfrac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}_A^2} = \sqrt{\dfrac{6.215}{10} - 24'7^2} = \sqrt{11'41} = 3'38$ semanas.

Marca B: $\bar{x}_B = \dfrac{\sum_{i=1}^{N} x_i}{N} = \dfrac{265}{10} = 26'5$ semanas; $\sigma_B = \sqrt{\dfrac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}_A^2} = \sqrt{\dfrac{7.109}{10} - 26'5^2} = \sqrt{8'65} = 2'94$ semanas.

b) Lógicamente es aconsejable elegir la marca B, pues tiene una duración media mayor (26'5 semanas) y menor desviación típica –variabilidad respecto de la media– (2'94 semanas).

Para comparar la dispersión de las distribuciones hallamos sus coeficientes de variación:

$$C_{\text{var}\,A} = \frac{\sigma_A}{\bar{x}_A} = \frac{3'38}{24'7}\cdot 100 = 13'68\ \% \qquad C_{\text{var}\,B} = \frac{\sigma_B}{\bar{x}_B} = \frac{2'94}{26'5}\cdot 100 = 11'09\ \%$$

Como vemos, según los coeficientes de variación, la distribución de la marca B tiene menor dispersión.

20. Una fábrica de yogures empaqueta éstos en cajas de cien unidades cada una. Para probar la eficacia de la producción se han analizado 80 cajas comprobando los yogures defectuosos que contiene cada una y se han obtenido los resultados de la tabla.

Nº de yogures defectuosos	0	1	2	3	4	5	6
Nº de cajas	40	15	10	9	3	2	1

Define cuáles son los individuos de esta muestra y la variable estadística. Después calcula para esta distribución estadística los parámetros de centralización y los parámetros de dispersión. Por último, representa gráficamente la distribución y halla el número de cajas que están en los intervalos $(\bar{x}-\sigma,\ \bar{x}+\sigma)$, $(\bar{x}-2\sigma,\ \bar{x}+2\sigma)$ y $(\bar{x}-3\sigma,\ \bar{x}+3\sigma)$. A la vista de los resultados, ¿puede calificarse la distribución de normal?

Solución.-

Los individuos de esta muestra son *cajas de cien unidades de yogures,* siendo la variable estadística el *número de yogures defectuosos que contiene cada caja.*

Hallemos los parámetros de centralización y de dispersión:

x_i	n_i	N_i	$n_i x_i$	$n_i \lvert x_i - \bar{x} \rvert$	$n_i x_i^2$
0	40	40	0	45'000	0
1	15	55	15	1'875	15
2	10	65	20	8'750	40
3	9	74	27	16'875	81
4	3	77	12	8'625	48
5	2	79	10	7'750	50
6	1	80	6	4'875	36
Total	**80**		**90**	**93'750**	**270**

Parámetros de centralización:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{N} = \frac{90}{80} = 1'125 \text{ yogures}$$

$Mo = 0$ yogures

$N/2 = 40 \Rightarrow Me = (0 + 1)/2 = 0'5$ yogures

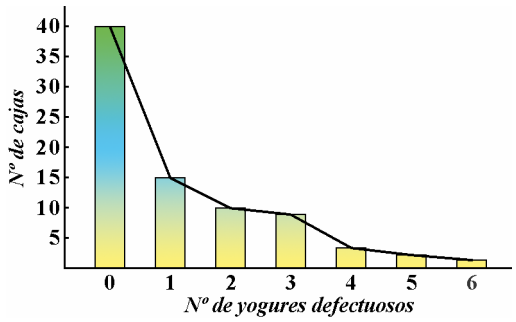
Parámetros de dispersión:

$R = 6 - 0 = 6$ yogures; $d_m = \dfrac{\sum_{i=1}^{k} n_i \lvert x_i - \bar{x} \rvert}{N} = \dfrac{93'75}{90} = 1'171875$ yogures

$$\sigma^2 = \frac{\sum_{i=1}^{k} n_i x_i^2}{N} - \bar{x}^2 = \frac{270}{90} - 1'125^2 = 2'1094 \quad \Rightarrow \quad \sigma = \sqrt{\sigma^2} = \sqrt{2'1094} = 1'4524 \text{ yogures}$$

$$C_{var} = \frac{\sigma}{\bar{x}} = \frac{1'4524}{1'125}\cdot 100 = 129'10\ \%$$

Representamos gráficamente la distribución mediante un diagrama de barras y un polígono de frecuencias:



En el intervalo $(\bar{x}-\sigma,\ \bar{x}+\sigma) = (-0'325, 2'575)$ hay 40 + 15 + 10 = 65 cajas, el 81'25 % del total.

En el intervalo $(\bar{x}-2\sigma,\ \bar{x}+2\sigma) = (-1'775, 4'025)$ hay 40 + 15 + 10 + 9 + 3 = 77 cajas, el 96'25 % del total.

En el intervalo $(\bar{x}-3\sigma,\ \bar{x}+3\sigma) = (-3'225, 5'475)$ hay 40 + 15 + 10 + 9 + 3 + 2 = 79 cajas, el 98'75 % del total.

En base a los porcentajes obtenidos, principalmente en el intervalo $(\bar{x}-\sigma,\ \bar{x}+\sigma)$, no se puede decir que la distribución tenga un comportamiento normal. Observa también que la distribución no es, en absoluto, simétrica respecto de ningún valor central de la variable estadística ($\bar{x} = 1'125$, $Mo = 0$ y $Me = 0'5$).

21. El presupuesto del Insalud, por Comunidades Autónomas y en miles de millones de pesetas, del año 1992 fue el siguiente.

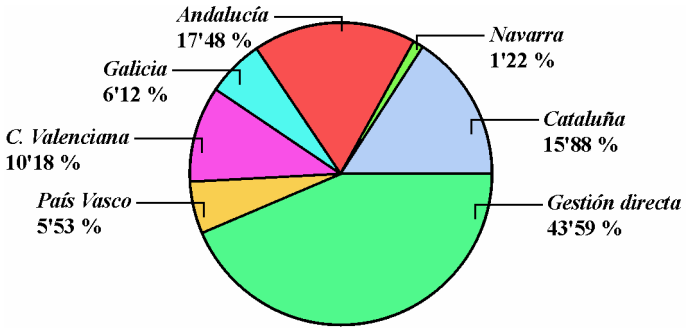
Comunidad	Cataluña	Navarra	Andalucía	Galicia	C. Valenciana	País Vasco	Gestión directa
Presupuesto	379	29	417	146	243	132	1.040

a) Construye el diagrama de sectores correspondiente a esta distribución de frecuencias.
b) ¿Qué datos de los anteriores se encuentran en el intervalo $(\bar{x}-\sigma,\ \bar{x}+\sigma)$?

Solución.-

a) Calculamos la amplitud y el porcentaje sobre el total para cada Comunidad Autónoma y completamos la tabla con los datos para el cálculo de los parámetros del apartado b.

Comunidad	Cataluña	Navarra	Andalucía	Galicia	C. Valenc.	P. Vasco	G. directa	**Total**
Presupuesto	379	29	417	146	243	132	1.040	**2.386**
Amplitud	57	4	63	22	37	20	157	**360**
%	15'88	1'22	17'48	6'12	10'18	5'53	43'59	**100**
x_i^2	143.641	841	173.889	21.316	59.049	17.424	1.081.600	**1.497.760**



b) La media y la desviación típica son:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{2.386}{7} = 340'86 \text{ miles de millones.}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2} = \sqrt{\frac{1.497.760}{7} - 340'86^2} = \sqrt{97.780'12} = 312'70 \text{ miles de millones.}$$

En el intervalo $(\bar{x} - \sigma,\ \bar{x} + \sigma) = (28'16, 653'56)$ se encuentran los presupuestos de todas las Comunidades Autónomas anteriores, lo que supone un total de 379 + 29 + 417 + 146 + 243 + 132 = 1.346 miles de millones de pesetas, esto es, un 56'41 % del presupuesto total.

22. Durante el mes de julio, en una determinada ciudad, se han registrado las siguientes temperaturas máximas.

32, 31, 28, 29, 29, 33, 32, 31, 30, 31, 31, 27, 28, 29, 29, 30, 32, 31, 31, 30, 30, 29, 29, 30, 30, 31, 30, 31, 34, 33, 33

a) Halla los parámetros de centralización.
b) Calcula el rango y la desviación típica.
c) Comprueba si en el intervalo $(\bar{x} - 2\sigma,\ \bar{x} + 2\sigma)$ se encuentra aproximadamente el 95'44 % de los datos.

Solución.-

a) Realizamos el recuento y hacemos la correspondiente tabla para el cálculo de los parámetros:

x_i	n_i	N_i	$n_i x_i$	x_i^2	$n_i x_i^2$
27	1	1	27	729	729
28	2	3	56	784	1.568
29	6	9	174	841	5.046
30	7	16	210	900	6.300
31	8	24	248	961	7.688
32	3	27	96	1.024	3.072
33	3	30	99	1.089	3.267
34	1	31	34	1.156	1.156
Total	**31**		**944**		**28.826**

Media aritmética: $\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{N} = \frac{944}{31} = 30'45\ ºC$

Moda: $Mo = 31$ ºC

Mediana: $N/2 = 15'5 \Rightarrow Me = 30$ ºC

b) Rango: $R = 34 - 27 = 7$ ºC

Desviación típica: $\sigma = \sqrt{\frac{\sum_{i=1}^{k} n_i x_i^2}{N} - \bar{x}^2} = \sqrt{\frac{28.826}{31} - 30'45^2} = \sqrt{2'57} = 1'60\ ºC$

c) En el intervalo $(\bar{x} - 2\sigma,\ \bar{x} + 2\sigma) = (27'25, 33'65)$ hay 2 + 6 + 7 + 8 + 3 + 3 = 29 datos, esto supone un 93'55 % del total de los datos.

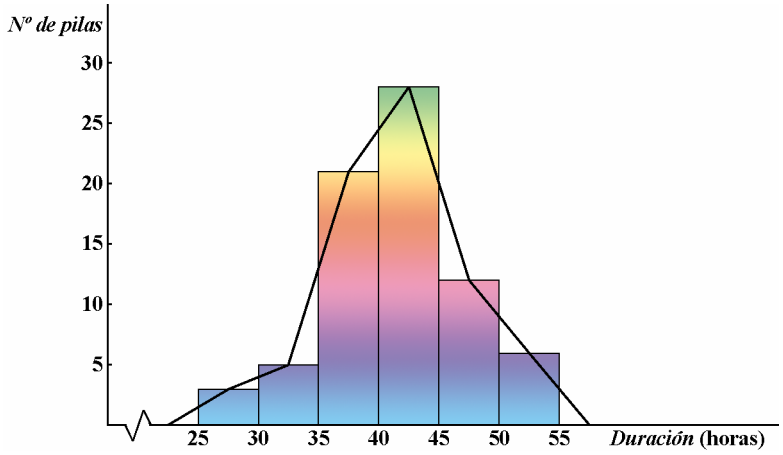
23. De una muestra de 75 pilas eléctricas, se han obtenido los datos de la tabla adjunta sobre su duración en horas.

***Duración* (horas)**	[25, 30)	[30, 35)	[35, 40)	[40, 45)	[45, 50)	[50 ,55]
Nº de pilas	3	5	21	28	12	6

a) Realiza la representación gráfica de la distribución.
b) Calcula la media y la desviación típica.
c) ¿Qué porcentaje de pilas tienen su duración comprendida en los intervalos $(\bar{x}-\sigma,\ \bar{x}+\sigma)$, $(\bar{x}-2\sigma,\ \bar{x}+2\sigma)$ y $(\bar{x}-3\sigma,\ \bar{x}+3\sigma)$? ¿Puede considerarse que la distribución tiene un comportamiento normal?

Solución.-

a) La representamos mediante un histograma y un polígono de frecuencias.



b) Hallamos ambos parámetros.

***Duración* (horas)**	x_i	n_i	$n_i x_i$	x_i^2	$n_i x_i^2$
[25, 30)	27’5	3	82’5	756’25	2.268’75
[30, 35)	32’5	5	162’5	1.056’25	5.281’25
[35, 40)	37’5	21	787’5	1.406’25	29.531’25
[40, 45)	42’5	28	1.190	1.806’25	50.575
[45, 50)	47’5	12	570	2.256’25	27.075
[50, 55]	52’5	6	315	2.756’25	16.537’5
Total		**75**	**3.107’5**		**131.268’75**

$$\bar{x}=\frac{\sum_{i=1}^{k} n_i x_i}{N}=\frac{3.107'5}{75}=41'43 \text{ horas};\qquad \sigma=\sqrt{\frac{\sum_{i=1}^{k} n_i x_i^2}{N}-\bar{x}^2}=\sqrt{\frac{131.268'75}{75}-41'43^2}=\sqrt{33'53}=5'79 \text{ horas}$$

c) En el intervalo $(\bar{x}-\sigma,\ \bar{x}+\sigma)$ = (35’54, 47’12) hay 21 + 28 = 49 pilas, un 65’33 % del total.

En el intervalo $(\bar{x}-2\sigma,\ \bar{x}+2\sigma)$ = (29’75, 52’91) hay 5 + 21 + 28 + 12 + 6 = 72 pilas, un 96 % del total.

En el intervalo $(\bar{x}-3\sigma,\ \bar{x}+3\sigma)$ = (23’96, 58’7) hay 75 pilas, lógicamente el 100 % del total.

Teniendo en cuenta los anteriores porcentajes, si podemos afirmar que la distribución tiene un comportamiento normal. Observa también en la representación gráfica anterior, que la distribución es unimodal y el polígono de frecuencias es bastante simétrico respecto de la media (41’43 horas).