T2. ESTADÍSTICA DESCRIPTIVA UNIDIMENSIONAL

- MEDIDAS DE POSICIÓN
  - CENTRADAS
    - $\bar{X}$ (Media)
    - Me (Mediana)
    - Mo (Moda)
  -
    - CUARTILES (Q_1, Q_2, Q_3)
    - OCTILES
    - DECILES
    - PERCENTILES ($P \dots P_{25}, \dots P_{95} \dots$)
- MEDIDAS DE DISPERSIÓN
  - ABSOLUTAS
    - Varianza
    - Desviación
    - Rango
    - Rango intercuartílico
  - RELATIVAS → Coeficiente de Variación
    - → variable estandarizada
- MEDIDAS DE FORMA
  - SIMETRÍA / ASIMETRÍA
  - CURTOSIS → apuntamiento respecto a la Normal

2.1 MEDIDAS DE POSICIÓN

La **moda** es el valor más repetido, es decir, el valor que tiene una mayor frecuencia.

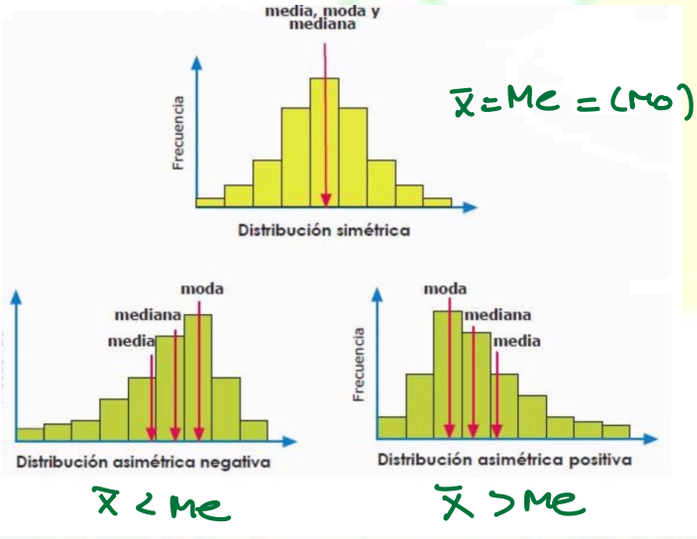
La **mediana** es un valor numérico que se sitúa en el medio una vez ordenadas las observaciones de más pequeñas a más grandes. Para calcularlo se hará:

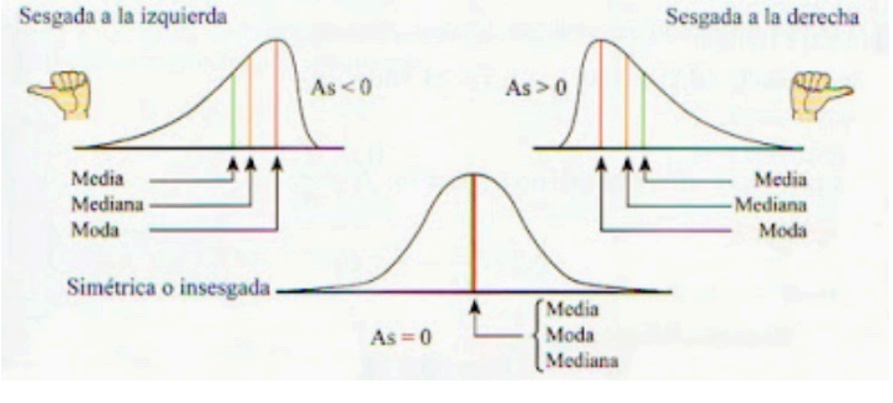
1) Ordenar los valores de menor a mayor.
2) Si N es impar, la media es la observación que ocupa el lugar $(N+1)/2$.
3) Si N es par, calculamos el valor $(N+1)/2$, y la mediana será el valor medio de las dos observaciones que ocupan los lugares más cercanos a $(N+1)/2$.

La **media** $(\bar{x})$ se obtiene dividiendo la suma de todas las observaciones por el número de individuos.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{N}$$

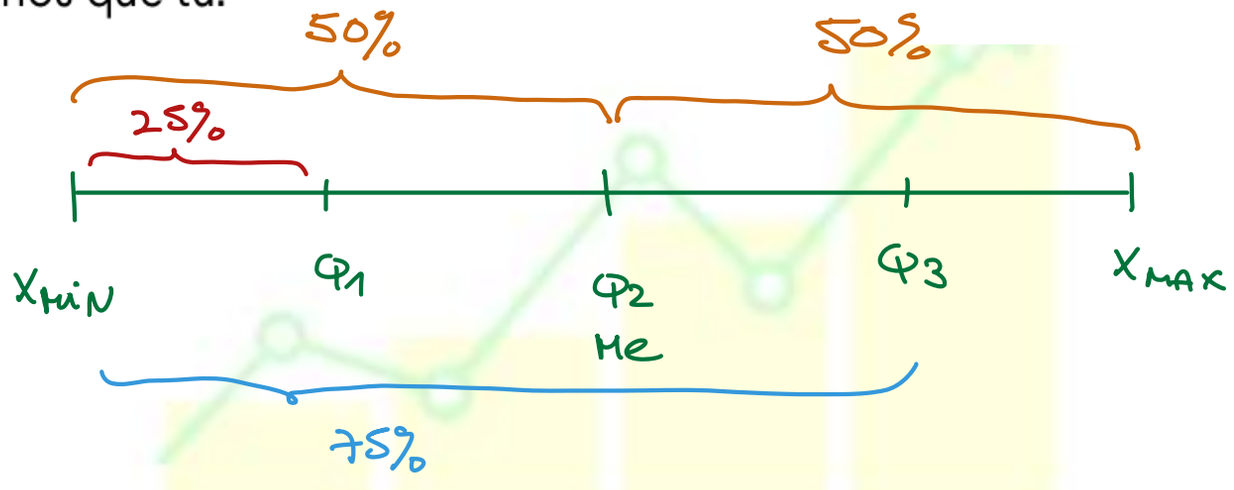
Estos tres estadísticos nos dan una idea del **centro de la distribución** (de hecho, técnicamente son estadísticos de tendencia central), con diferencias entre ellos: la media es POCO ROBUSTA, es decir, si añadimos un valor extremo por ser muy grande o muy pequeño, la media cambia mucho. Y eso a la mediana no le pasa. Por eso, cuando los datos tienen valores extremos por un lado (que no son compensados por el otro), o lo que es lo mismo, cuando la variable es muy **ASIMÉTRICA**, la media no es buena para determinar el centro de la distribución. ¿Y qué significa que una variable sea **SIMÉTRICA**? Fácil, que, al representarla en un histograma, obtienes algo parecido a la primera figura, mientras que las otras dos son asimétricas (sesgadas):

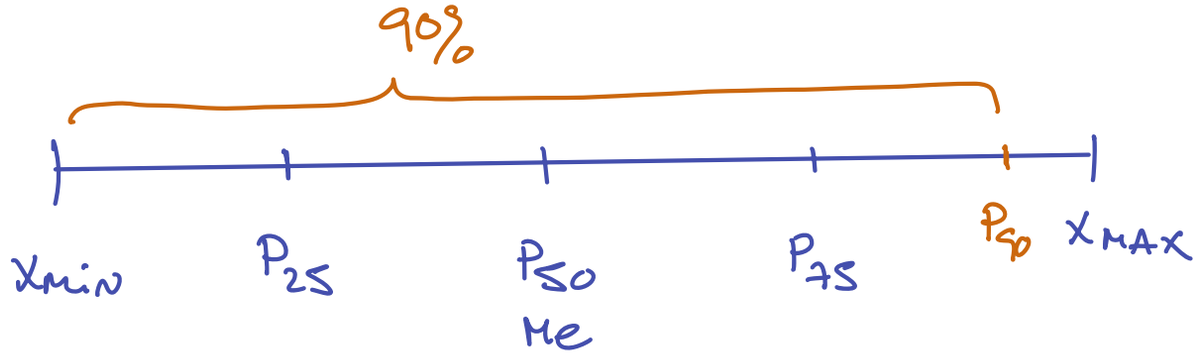




Cuartiles: igual que el resto, pero dividen a la muestra en 4 partes iguales, y por tanto hay 3 cuartiles: el cuartil 1 (o percentil 25), el cuartil 2 (o percentil 50), y el cuartil 3 (o percentil 75).

Percentiles. Un percentil nos indica qué porcentaje de valores están por debajo de ese percentil. Por ejemplo, si tu salario está en el percentil 90 quiere decir que el 90% de la población (o de la muestra, depende) está por debajo de tu salario. O lo que es lo mismo, el 90% de la población cobra menos que tú.



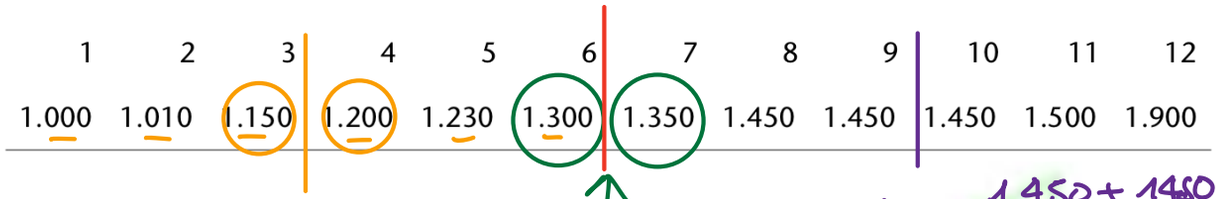


EJEMPLO 1

1. Los sueldos en € de los analistas contratados por una empresa informática son:

1.150	1.200	1.300	1.450	1.000	1.350	1.900	1.230	1.450	1.010	1.500	1.450

Calculad el sueldo medio y la mediana de los sueldos.



$Q_1 = \frac{1.150 + 1200}{2} = 1.175$

$Me = 1.325$

$Q_3 = \frac{1450 + 1450}{2} = 1450$

$\bar{x} = \frac{\sum x_i}{N} = \frac{15990}{12} = 1.332'5$

- tiene en cuenta todos los datos
- Poco Robusta a errores

Me $N = 12$ $\frac{N+1}{2} = \frac{13}{2} = 6'5$ → (6) (7)

$Me = \frac{x_6 + x_7}{2} = \frac{1.300 + 1.350}{2} = 1325$

↳
- No utiliza todos los datos → solo los centrales
- Robusta a errores

EJEMPLO 2

2. Un estudio de la universidad ha recopilado el tiempo en minutos que sus estudiantes dedican semanalmente a los *chats*, y se ha obtenido la tabla siguiente:

c_i	x_i	f_i
5	[0,10)	1%
20	[10,30)	15%
115	[30,200)	44%
300	[200,400)	40%

$$\bar{x} = \frac{\sum x_i \cdot n_i}{N} = \sum x_i \cdot f_i$$

Calculad la media del tiempo que los estudiantes dedican a chatear y representad gráficamente la distribución de la variable.

$$\bar{x} = \sum c_i \cdot f_i$$

$$\bar{x} = 5 \cdot 0'01 + 20 \cdot 0'15 + 115 \cdot 0'44 + 300 \cdot 0'4 = 173'65 \text{ min}$$

2.2 COMPARACIÓN MEDIA Y MEDIANA

- Media aritmética
  - Utiliza todos los datos
  - Poco Robusta a errores
  - Si hay datos altos atípicos $\bar{x} \uparrow\uparrow$
  - Si hay datos bajos atípicos $\bar{x} \downarrow\downarrow$

  } la $\bar{x}$ no es representativa

- Mediana
  - Utiliza solo los valores centrales
  - Más robusta a errores que la $\bar{x}$
  - Se utiliza en distribuciones asimétricas

2.3 PROPIEDADES DE LA MEDIA

- $\bar{x} = \frac{\Sigma x_i}{N}$ $\qquad \bar{x} = \frac{\Sigma x_i \cdot n_i}{N} = \Sigma x_i \cdot f_i$

$\Sigma x_i = N \cdot \bar{x}$

- $\Sigma (x_i - \bar{x}) = 0$

- Transformaciones lineales

$x_i \longrightarrow x_i + a$ $\qquad \bar{y} = \bar{x} + a$

$x_i \longrightarrow bx_i$ $\qquad \bar{y} = b\bar{x}$

$x \longrightarrow y = a + bx$

$\bar{x} \longrightarrow \bar{y} = a + b\bar{x}$

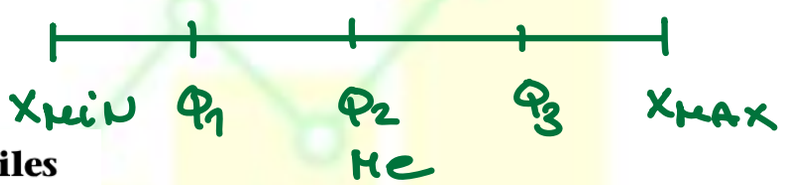
2.4 MEDIDAS DISPERSIÓN

CUARTILES Y MEDIANA

a) Primer cuartil (Q1): es aquel valor numérico tal que al menos el 25% de las observaciones son menores o iguales que aquél, y al menos el 75%, mayores o iguales.

b) Segundo cuartil (Q2): es la mediana.

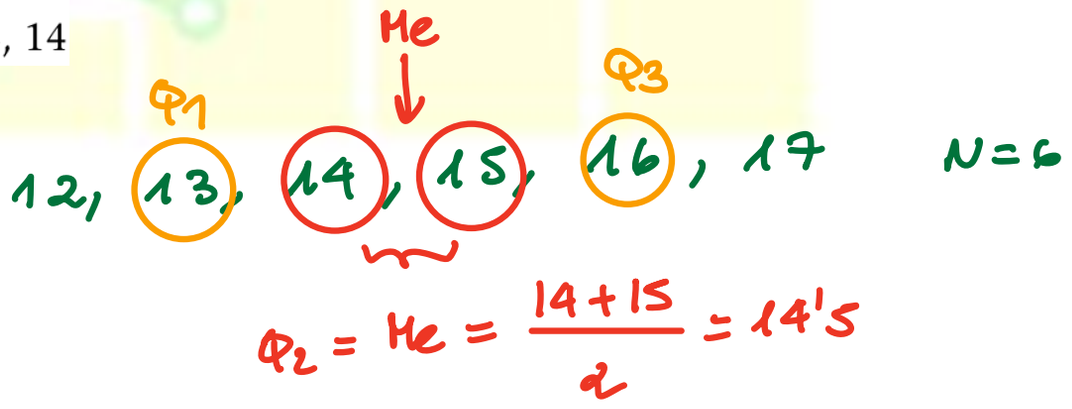
c) Tercer cuartil (Q3): es aquel valor numérico tal que al menos el 75% de las observaciones son menores o iguales que aquél, y al menos el 25%, mayores o iguales.



EJEMPLO 3

Ejemplos de cálculo de cuartiles

17, 16, 12, 13, 15, 14



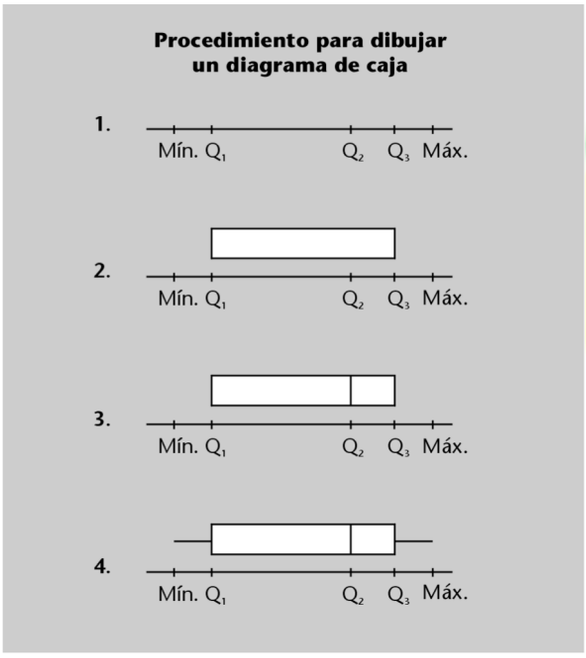
2.5 RANGO INTERCUARTÍLICO

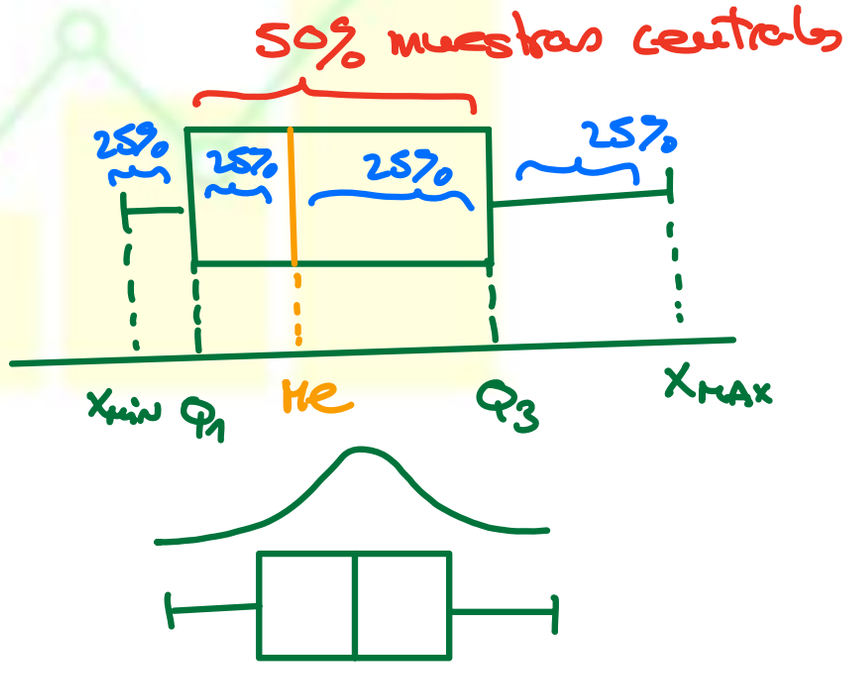
El rango intercuartílico es la diferencia entre el tercer y el primer cuartil, es decir:

Rango intercuartílico = Q3 – Q1 $IQR = Q_3 - Q_1$

2.6 DIAGRAMA DE CAJA

Los cinco números resumen de la distribución de una variable son: el mínimo, Q1, la mediana, Q3 y el máximo.





EJEMPLO 4

1. Se ha medido el tiempo en segundos que tarda en arrancar la última versión del programa Macrohard Phrase en los ordenadores de nuestra empresa según el sistema operativo con el que funcionan. Los resultados han sido los siguientes:

- En los ordenadores equipados con Doors95: 27, 25, 50, 33, 25, 86, 28, 31, 34, 36, 37, 44, 20, 59 y 85 segundos. $\bar{x} = \frac{\sum x_i}{N} = 41'33$
- En los ordenadores equipados con Doors98: 33, 7, 25, 14, 5, 31, 19, 10, 29 y 18 segundos. $\bar{y} = 19'1$

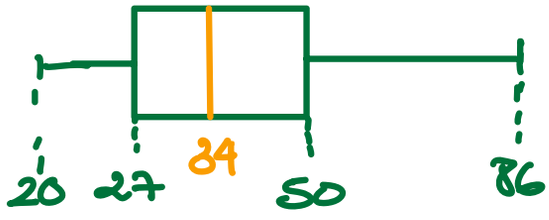
Calculad los cinco números resumen y la media de la distribución correspondiente al tiempo que el programa tarda en arrancar y dibujad algunos gráficos que os parezcan relevantes para comparar el tiempo que el programa tarda en arrancar según el sistema operativo. A partir de estos gráficos, comparad el comportamiento del programa según el sistema operativo y explicad si creéis que hay diferencia entre utilizar Doors95 y Doors98.

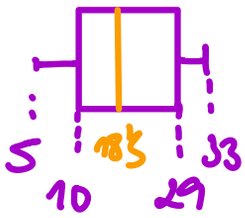
X 20, 25, 25, (27), 28, 31, 33, (34), 36, 37, 44, (50), 59, 85, 86

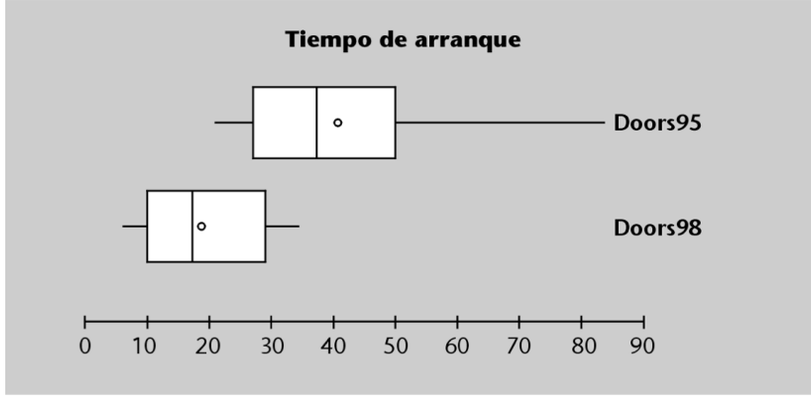
Q_1 Q_2 Q_3

Y 5, 7, (10), 14, (18), (19), 25, (29), 31, 33

Q_1 $Q_2 = 18'5$ Q_3







2.7 VARIANZA Y DESVIACIÓN

$Sx^2 \geq 0$

$Sx \geq 0$

La **varianza** (también llamada **varianza poblacional**) y que se denota por s^2 se calcula como la suma del cuadrado de las desviaciones con respecto a la media dividido por N, donde N es el número de las observaciones, es decir:

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

unidades^2 $€^2$

$$s_x^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - (\bar{x})^2$$

La **desviación típica** (denotada por s o s_x) se define como la raíz cuadrada positiva de la varianza, es decir:

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

unidades $€$

Ejemplo:

2, 5, 7, 9

$\bar{x} = 5'75$

$$Sx^2 = \frac{2^2+5^2+7^2+9^2}{4} - (5'75)^2$$

2.8 COEFICIENTE DE VARIACIÓN

→ medida dispersión relativa

↳ permite comparar distribuciones

$$CV = \frac{Sx}{\bar{x}} \qquad CV = \frac{Sx}{\bar{x}} \cdot 100$$

↳ adimensional

CV ↑↑ → mayor dispersión → $\bar{x}$ no es representativa

2.9 PROPIEDADES DE LA VARIANZA Y DESVIACIÓN

$xi \longrightarrow xi + a$ ← no le afecta la constante

Var(xi)
D(xi)

$Var(xi+a) = Var(xi) + Var(a)$ (0)

$D(xi+a) = D(xi)$

$xi \longrightarrow bxi$

Var(xi)
D(xi)

$Var(bxi) = b^2 Var(xi)$

$D(bxi) = b\,D(xi)$

EJEMPLO 5

Ejemplo de cálculo de la desviación típica

Consideremos los valores 7, 4, 6, 5, 5. Para calcular su varianza, organizamos los cálculos en una tabla como ésta:

$$\bar{x} = \frac{7+4+6+5+5}{5} = 5'4$$

$$S_x^2 = \frac{7^2+4^2+6^2+5^2+5^2}{5} - (5'4)^2 = 1'04$$

$$S_x = \sqrt{1'04} = 1'02$$

2.10 LA REGLA DE TCHEBICHEV

La regla de Tchebichev afirma que, dado cualquier conjunto de datos $x_1, x_2, \ldots, x_n$ con media $\bar{x}$ y desviación típica s_x, si m es un número cualquiera, entonces la proporción de datos que pertenecen al intervalo $(\bar{x} - ms_x, \bar{x} + ms_x)$ es, como mínimo:

$$1 - \frac{1}{m^2}$$

$$p(\bar{x} - m s_x < \bar{x} < \bar{x} + m \cdot s_x) \geq 1 - \frac{1}{m^2}$$

2.11 DATOS ESTANDARIZADOS

$$z = \frac{x - \text{Media de las } x}{\text{Desviación típica de las } x} = \frac{x - \bar{x}}{s_x}$$

$$z = \frac{x - \bar{x}}{Sx}$$ adimensional → Permite comparar

2.12 DATOS TABULADOS

$$s^2 = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}{N} = \sum_{i=1}^{k} f_i (x_i - \bar{x})^2 \text{ o bien } s^2 = \frac{\sum_{i=1}^{k} n_i (m_i - \bar{x})^2}{N} = \sum_{i=1}^{k} f_i (m_i - \bar{x})^2$$

$$S_x^2 = \frac{\Sigma x_i^2 \cdot n_i}{N} - (\bar{x})^2$$

EJEMPLO 6

Ejemplo de cálculo de la varianza a partir de datos tabulados

En el ejemplo de los virus atacantes (I), en el que se calcula el número de virus que han atacado los diferentes ordenadores de nuestra empresa durante el año 2000, la media vale 4,16. Para calcular la varianza, nos puede ayudar la tabla siguiente:

x_i	n_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$n_i \cdot (x_i - \bar{x})^2$
0	10	−4'16	17'3	173'056
5	15	0'84	0'705	10'58
9	6	4'84	23'42	140'55
Totales	31			Σ → 324'19

$$\bar{x} = \frac{0 \cdot 10 + 5 \cdot 15 + 9 \cdot 6}{31} = 4'16$$

$$Sx^2 = \frac{\Sigma (x_i - \bar{x}) \cdot n_i}{N} = \frac{324'19}{31} = 10'45$$

$$Sx^2 = \frac{0^2 \cdot 10 + 5^2 \cdot 15 + 9^2 \cdot 6}{31} - (4'16)^2 = 10'45$$

$$Sx = \sqrt{10'45} = 3'23$$