Estadística

Estadística

- La estadística es la ciencia que se ocupa de la recogida de datos, su organización y análisis.
- Según el método de estudio y el problema a resolver se puede distinguir:
  - **Estadística descriptiva**: Se encarga de recolectar datos, organizarlos y en el cálculo de valores que describan el conjunto objeto de estudio.
  - **Estadística inferencial**: se encarga de elaborar conclusiones para el conjunto estudiado a partir del estudio de una muestra.

Conceptos estadísticos I

- Población: conjunto de elementos a estudiar.
- Individuo: cada uno de los elementos de la población.
- Muestra: Subconjunto de la población que se toma para representar a la población en el estudio.
- Tamaño de la muestra: Número de elementos que componen la muestra.

Conceptos estadísticos II

- Carácter estadístico: propiedad de los individuos de una población que permite clasificarlos.
  - Cualitativo: no pueden expresarse numéricamente.
  - Cuantitativo: puede expresarse numéricamente.
- Variable estadística: conjunto de valores que puede tomar un carácter estadístico cuantitativo.
  - Discreta: la variable puede tomar un número finito de valores
  - Continua: la variable puede tomar un valor que se encuentre dentro de un intervalo de números reales.

Estudio estadístico

Para realizar un estudio estadístico se siguen las siguientes pautas:

- Extracción de una muestra de la población

  Debe ser representativa de la población

- Recuento de datos

  Proporciona la frecuencia absoluta de la tabla

- Elaboración de una tabla estadística

Recuento de datos

- El recuento de datos proporciona la frecuencia absoluta para cada valor de la variable estadística
- En la primera columna se disponen los distintos valores que puede tomar la variable estadística
- En la siguiente, tras contar los casos en los que aparece cada valor de la variable, se calcula la frecuencia absoluta.

Ejemplo I

A las 50 familias que viven en una calle se les ha realizado una encuesta preguntando el número de personas que viven en su vivienda.

3	3	5	8	3
7	2	4	4	3
2	6	1	5	4
4	3	3	4	1
4	6	4	7	8
7	4	2	2	7
3	2	2	2	2
6	3	3	3	4
2	4	6	8	1
3	5	5	3	6

El **carácter estadístico** se corresponde con el número de personas que viven juntas.
El **tamaño de la muestra** es 50.

Ejemplo I (continuación)

Valor de la variable	Frecuencia absoluta
x_i	f_i
1	3
2	9
3	12
4	10
5	4
6	5
7	4
8	3
Total	50

El rango de valores de la variable va de 1 a 8. La variable es **discreta**.

La frecuencia absoluta establece en este caso el número de viviendas que está ocupada por el número de inquilinos que le corresponde al valor de la variable.

Hay 4 viviendas de las 50 ocupadas por 7 personas.

La suma de las frecuencias absolutas es el tamaño de la muestra.

Intervalos o clases

- Cuando una variable estadística cuantitativa puede tomar una gran cantidad de valores distintos, se agruparán en intervalos.
- Normalmente se toman intervalos de igual amplitud.
- Para realizar cálculos se utilizará la marca de clase que no es más que el punto medio del intervalo.

Ejemplo II

A los 100 empleados de una empresa , se les ha realizado una prueba sobre resolución de problemas. En una escala de 0 a 100 se ha obtenido las siguientes puntuaciones:

2	58	69	62	17	53	5	30	76	33
84	76	69	70	63	65	46	99	65	2
5	82	83	80	72	74	39	21	84	48
68	5	15	90	48	99	32	17	27	62
21	54	11	4	7	33	90	76	4	18
27	53	47	11	88	66	64	29	59	99
46	61	20	27	49	44	37	90	72	35
84	7	85	14	28	71	89	76	48	52
70	41	25	85	90	4	81	78	55	36
73	8	2	94	84	68	11	13	37	9

Ejemplo II (continuación)

Se ha decidido agrupar los valores en 5 clases, como el menor valor es 2 y el mayor 100, la amplitud del intervalo será:

$$\frac{Máximo\ valor - Mínimo\ valor}{Número\ de\ clases} = \frac{100 - 2}{5} = 19{,}6$$

Tomaremos para la longitud del intervalo 20, pues es más fácil de realizar cálculos:

Intervalo	Marca de clase	Frecuencia Absoluta
$(x_i, x_{i+1}]$	x_i	f_i
(0,20]	10	23
(20,40]	30	17
(40,60]	50	16
(60,80]	70	25
(80,100]	90	19
	Total	100

Frecuencias

Frecuencia absoluta: Número de individuos de la muestra que presenta un valor de la variable estadística. Su suma debe ser el tamaño de la muestra.

Frecuencia relativa: Fracción que representa la frecuencia absoluta respecto del tamaño de la muestra. Se calcula dividiendo la frecuencia absoluta entre el tamaño de la muestra. La suma total debe ser 1.

Frecuencia porcentual: Porcentaje que representa cada frecuencia absoluta respecto del tamaño de la muestra. La suma total debe ser 100.

Frecuencias acumuladas: se corresponde con la suma de las frecuencias anteriores a las de un valor de la variable más la suya propia.

Ejemplo

Valor de la variable	Frecuencia absoluta	Frecuencia relativa	Frecuencia porcentual	Frecuencia acumulada
x_i	f_i	h_i	p_i	F_i
1	3	0,06	6	3
2	9	0,18	18	12
3	12	0,24	24	24
4	10	0,2	20	34
5	4	0,08	8	38
6	5	0,1	10	43
7	4	0,08	8	47
8	3	0,06	6	50
Total	50	1	100	

MÉTODOS GRÁFICOS PARA DESCRIBIR DATOS

Diagrama de barras

El diagrama de barras se utiliza para representar variables cualitativas o cuantitativas discretas.

A cada valor de la variable estadística o categoría, (que se corresponde con en el eje OX), se le asocia un rectángulo de área proporcional a la frecuencia absoluta o relativa que le corresponde a la categoría.

El polígono de frecuencias se crea uniendo con segmentos los puntos medios consecutivos del diagrama de barras.

Ejemplo: diagrama de barras

x_i	f_i
Lunes	12
Martes	15
Miércoles	17
Jueves	20
Viernes	15
Sábado	18
Domingo	22

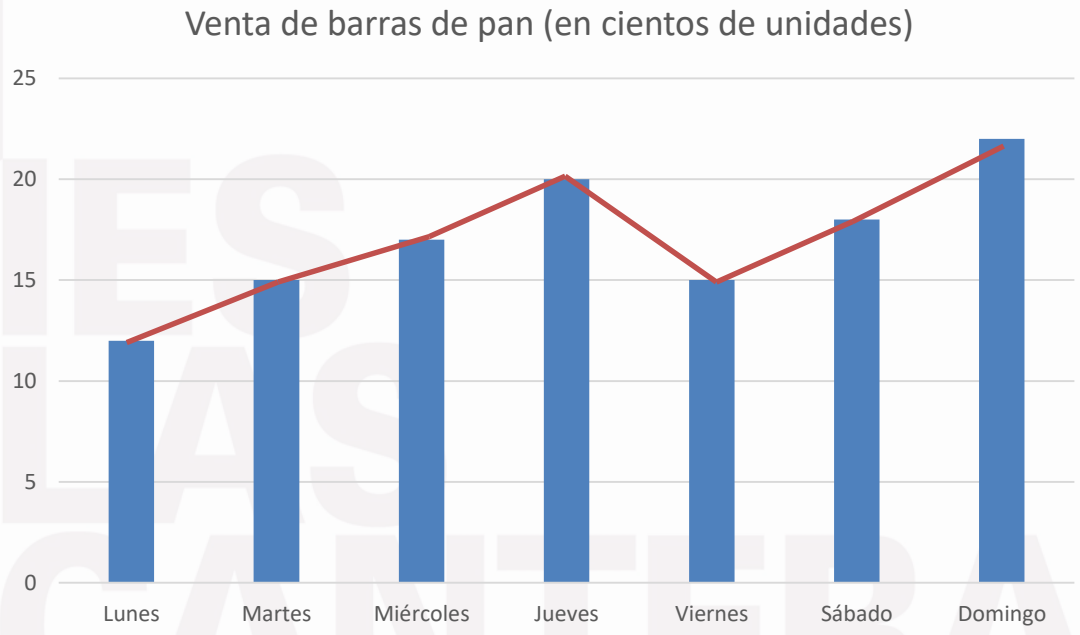


Diagrama de sectores

Consiste en representar cada una de los valores de la variable estadística mediante un sector circular proporcional a su valor.

Se utilizan cuando los valores de la variable estadística no son numerosos.

Para calcular el ángulo del sector que le corresponde a cada frecuencia absoluta se utiliza la siguiente proporción:

$$\frac{Grados\ del\ sector}{360^0} = \frac{Frecuencia\ absoluta}{Tamaño\ de\ la\ muestra}$$

Si se desea utilizar la frecuencia relativa, la proporción es:

$$\frac{Grados\ del\ sector}{360^0} = Frecuencia\ relativa$$

Ejemplo: diagrama de sectores

x_i	f_i	h_i
Lunes	38	0,1402214
Martes	56	0,20664207
Miércoles	120	0,44280443
Jueves	12	0,04428044
Viernes	45	0,16605166

Total: 271

El ángulo que le corresponde al miércoles es:

$$\frac{Grados\ del\ sector}{360^0} = \frac{120}{271}$$ es decir aproximadamente 159^0

Diagrama de tallo y hojas

Para realizar este tipo de diagramas se debe partir cada uno de los datos en dos partes: un tallo y una hoja. Por ejemplo, 213 puede ser dividido en 2 y 13 o en 21 y 3 (tallo y hoja respectivamente).

Cada tallo dará lugar a una fila que estará alineada en horizontal con todas las hojas ordenadas que comparten el mismo tallo. Estarán separadas por un línea vertical.

Cada tallo, en orden se encontrará alineado en vertical y en orden.

Este tipo de gráfico permite mostrar la misma información que un diagrama de barras a la vez que permite observar la distribución de los valores en cada uno de los tallos.

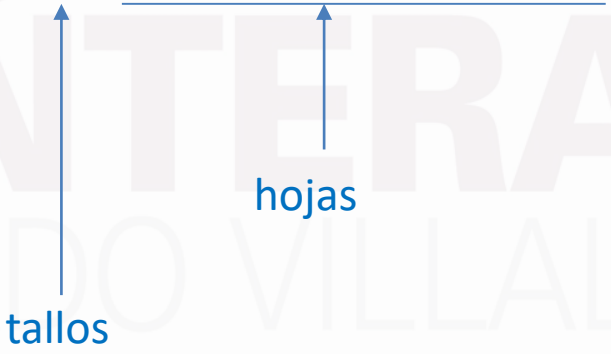
Ejemplo: diagrama tallo/hoja

Supongamos que la tabla incluye la medida, en centímetros de los alumnos de una clase. El tallo serán los dos dígitos más significativos y las hojas el último número.

160	181	190
178	162	152
181	165	140
154	158	182
154	157	180
151	174	158
173	145	150
168	152	141
148	184	176
170	186	181
161	149	162
187	154	186
154	144	154
143	172	141
154	167	141
178	182	175
164	142	164
149	174	149
148	155	140

Diagrama tallo / hoja

14 | 0 0 1 1 1 2 3 4 5 8 8 9 9 9
15 | 0 1 2 2 4 4 4 4 4 5 7 8 8
16 | 0 1 2 2 4 4 5 7 8
17 | 0 2 3 4 4 5 6 8 8
18 | 0 1 1 1 2 2 4 6 6 7
19 | 0



Histogramas

Un histograma para datos numéricos discretos es un gráfico de la frecuencia absoluta o relativa, similar al diagrama de barras.

Cada frecuencia es representada por un rectángulo centrado sobre el valor correspondiente.

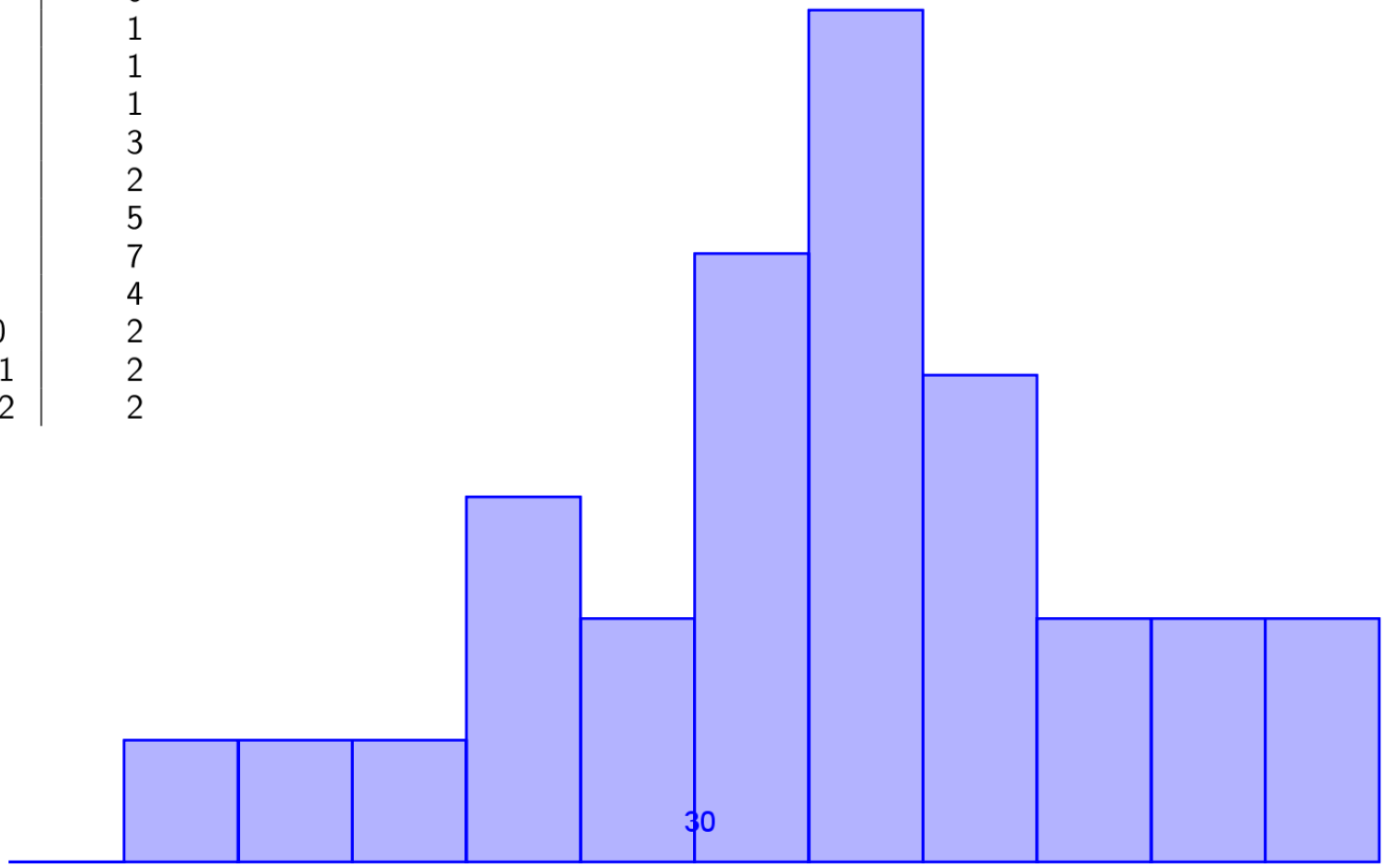
El área del rectángulo es proporcional a la frecuencia que se desea representar.

Se puede mostrar fácilmente:

- El centro o valor típico
- La variación de los valores
- Su forma general, la presencia de huecos y las partes aisladas

Ejemplo: Histograma

Intervalo	Frecuencia
0 – 1	0
1 – 2	1
2 – 3	1
3 – 4	1
4 – 5	3
5 – 6	2
6 – 7	5
7 – 8	7
8 – 9	4
9 – 10	2
10 – 11	2
11 – 12	2



Distribuciones unidimensionales

PARÁMETROS DE CENTRALIZACIÓN

Parámetros estadísticos

Permiten describir de forma resumida el comportamiento de la variable estadística

Existen dos grandes grupos:

- Parámetros de **centralización**
  - Moda, mediana y media
- Parámetros de **dispersión**
  - Rango, varianza, desviación típica y coeficiente de variación

Parámetros de centralización

La moda (*Mo*): de los valores de la variable estadística el de mayor frecuencia.

La mediana (*Me*): tras haber ordenado los valores de la variable (junto con sus frecuencias), aquel valor de la variable que ocupa el lugar central.

La media ($\overline{x}$): valor que se obtiene al sumar los productos de los valores de la variable estadística por la frecuencia absoluta, dividido por el tamaño de la muestra.

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{N}$$

Ejemplo

Valor de la variable	Frecuencia absoluta
x_i	f_i
1	3
2	9
3	12
4	10
5	4
6	5
7	4
8	3
Total	50

La moda es 3, pues es el valor de la variable que más frecuencia absoluta tiene

La mediana es 4, pues contiene el elemento 25 y 26 (mitad y mitad mas uno del tamaño de la población)

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{N} = \frac{3 \cdot 1 + 2 \cdot 9 + 3 \cdot 12 + \cdots .. + 8 \cdot 3}{50} = 3,98$$

A tener en cuenta para calcular la mediana

Cuando la variable estadística es discreta y el número de elementos es par, para calcular la mediana se utilizan los dos valores centrales y se realiza la media aritmética de los valores centrales de la variable estadística.

Si la variable estadística se encuentra agrupada en intervalos el que contiene el valor medio es el intervalo mediano pero la mediana se calculará mediante la siguiente fórmula:

$$M_e = L_i + a\frac{\frac{N}{2} - F_{i-1}}{f_i}$$

Donde:

- L_i Límite inferior del intervalo mediano
- a Amplitud del intervalo
- f_i Frecuencia absoluta de la clase mediana
- N Tamaño de la muestra
- F_{i-1} Frecuencia absoluta acumulada de la clase anterior

Ejemplo

Intervalo	Marca de clase	Frecuencia Absoluta
$(x_i,x_{i+1}]$	x_i	f_i
(0,20]	10	23
(20,40]	30	17
(40,60]	50	16
(60,80]	70	25
(80,100]	90	19
	Total	100

El intervalo mediano es [40,60] que acumula los datos 50 y 51

$$M_e = L_i + a\frac{\frac{N}{2} - F_{i-1}}{f_i} = 40 + 20\frac{\frac{100}{2} - 40}{16} = 52{,}5$$

Medidas de orden o posición

Estas medidas indican el orden o posición de una observación entre los valores de una variable cuantitativa.
Para realizar su cálculo es conveniente ordenar los valores de la muestra.
El **percentil** al **p%** es el valor que cumple que el p% de las observaciones de la muestra son inferiores a él.
Para el cálculo de la posición que ocupa el percentil, se utiliza la fórmula:

$$Posición = (n+1)\frac{p}{100}$$

Esta fórmula será válida cuando el valor obtenido sea entero.

Ejemplo

Calculemos el percentil a 25% y 75% de la siguiente muestra ordenada:

1.48 1.56 1.56 1.59 1.60 1.61 1.63 1.64 1.64 1.67 1.68
1.68 1.68 1.68 1.69 1.70 1.71 1.72 1.72 1.75 1.76 1.76
1.77 1.77 1.79 1.81 1.81 1.84 1.84 1.88 1.94

Percentil a 25%, $Posición = (31+1)\frac{25}{100} = \frac{32}{4} = 8$ Valor 1.64

Percentil a 75%, $Posición = (31+1)\frac{75}{100} = \frac{2400}{100} = 24$ Valor 1.77

Percentiles (posición decimal)

Cuando al calcular un percentil no obtenemos una posición entera, utilizaremos el valor entero obtenido:

$$Posición = (n+1)\frac{p}{100}$$

Y después, utilizaremos el siguiente valor para calcular el percentil:

$$Dec(Pos) \cdot X_{[Pos]+1} + \big(1 - Dec(Pos)\big) \cdot X_{[Pos]}$$

Donde:

$Dec(Pos)$ es la parte decimal de la posición obtenida inicialmente

$X_{[Pos]+1}$ es el valor que ocupa el lugar posición mas uno

$X_{[Pos]}$ es el valor que ocupa la posición obtenida

Ejemplo

Calculemos el percentil a 5% de la siguiente muestra ordenada:

1.48 1.56 1.56 1.59 1.60 1.61 1.63 1.64 1.64 1.67 1.68
1.68 1.68 1.68 1.69 1.70 1.71 1.72 1.72 1.75 1.76 1.76
1.77 1.77 1.79 1.81 1.81 1.84 1.84 1.88 1.94

Percentil a 5%, $Posición = (31 + 1)\frac{5}{100} = \frac{160}{100} = 1{,}6$

Utilizamos la siguiente fórmula, pues el valor obtenido es decimal

$$Dec(Pos) \cdot X_{[Pos]+1} + \big(1 - Dec(Pos)\big) \cdot X_{[Pos]}$$

$$0{,}6 \cdot 1{,}56 + (\,1 - 0{,}6\,) \cdot 1{,}48 = 0{,}936 + 0{,}592 = 1{,}528$$

El percentil a 5% es 1,528

Cuartiles

Los percentiles al 25%, 50% y 75% se denominan primer, segundo y tercer cuartil respectivamente.

Los cuartiles son importantes pues dividen la muestra en cuatro partes que contienen el mismo número de individuos.

El cálculo de los cuartiles se realiza con los respectivos percentiles.

La mediana es tanto el percentil al 50% como el segundo cuartil.

Los cuartiles permiten representar la información de un conjunto de datos de una forma muy resumida: los diagramas de caja.

Diagrama de caja

Se compone de un rectángulo, cuyo lado superior representa el primer cuartil y el lado inferior por el tercer cuartil.

En el centro del rectángulo se representa una línea que se corresponde con la mediana (segundo cuartil).

Y dos segmentos verticales (llamados “bigotes”):

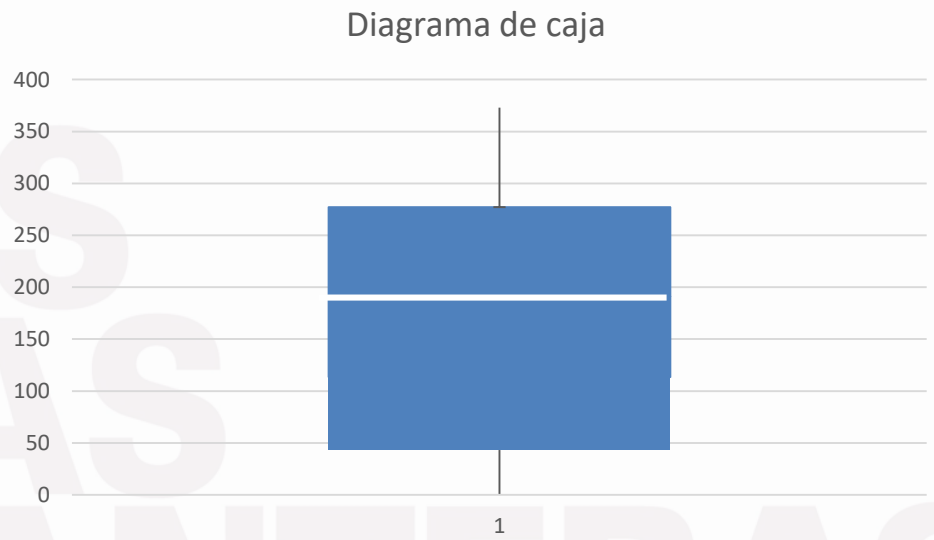
Uno nace del valor mínimo de los datos y termina en el lado superior del rectángulo.

El otro nace del lado inferior del rectángulo y termina en el mayor valor de los datos.

Ejemplo: diagrama de caja

El siguiente gráfico representa un diagrama de caja para los siguientes valores obtenidos de una muestra:

Mínimo	43
Primer cuartil	71
Mediana	78
Tercer cuartil	86
Máximo	96



Parámetros estadísticos de dispersión I

El rango (amplitud o recorrido): diferencia entre el mayor y menor valor de la variable estadística.

La varianza: mide la distancia entre la media y los valores de la variable estadística.

Las siguientes fórmulas permiten el cálculo de la varianza.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 f_i}{N} \qquad \sigma^2 = \frac{\sum_{i=1}^{n} x_i^2 f_i}{N} - \overline{x}^2$$

Parámetros estadísticos de dispersión II

La desviación típica: mide si los datos se encuentran agrupados en las proximidades de la media. Se mide en las mismas unidades que los datos

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 f_i}{N}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 f_i}{N} - \bar{x}^2}$$

La desviación típica es el parámetro de dispersión más utilizado

Si se suma una constante a cada uno de los valores de la variable, la desviación típica no varía

Si se multiplica una constante a cada uno de los valores de la variable, la desviación típica queda multiplicada por él.

Coeficiente de variación

Cuando se desea comparar dos variables aleatorias se utiliza el **coeficiente de variación,** que no es más que la razón entre la desviación típica y la media de la variable

$$CV = \frac{\sigma}{\overline{x}}$$

El coeficiente de variación siempre es positivo y no tiene unidades.

Cuanto más pequeño es, los valores de la variable se encuentran más concentrados alrededor de la media

Permite comparar dos variables aleatorias heterogéneas.