

Bloque 4. Estadística y Probabilidad

1. Estadística unidimensional

1. Definición de estadística

La **Estadística** es la rama de las matemáticas que trata del recuento, ordenación y clasificación de los datos obtenidos a partir de observaciones (por ejemplo, un sondeo sobre la edad de las personas encuestadas, o el grupo sanguíneo de los pacientes de un Hospital), para poder hacer comparaciones y sacar conclusiones.

Un **estudio estadístico** consta de las siguientes fases:

- Recogida de datos
- Organización y representación de datos.
- Análisis de datos.
- Obtención de conclusiones

2. Población y muestra

Al conjunto total de personas o de objetos de los que nos interesa conocer una determinada opinión o característica en la observación realizada es a lo que llamaremos **POBLACIÓN**. Cada una de esas personas u objetos es un **individuo** de la población.

- Preguntar a toda la población normalmente es imposible, así que habrá que **elegir un grupo que represente toda la población.**

Al grupo elegido para que responda al cuestionario se le denomina **MUESTRA**.

- Cuanto mayor sea el número de personas que forman la muestra más fiable será el estudio estadístico.
- Si elegimos mal la muestra los resultados no serán reales.

¿Cómo elegir entonces la muestra? En la mayoría de los casos debe ser lo más representativa posible en relación a los datos que se desean estudiar. Hay que tener en cuenta que:

- ☉ La elección de la muestra puede ser.
 - **Aleatoria:** los encuestados se eligen al azar.
 - Ventaja: suelen arrojar datos más objetivos
 - Inconveniente: es posible que la muestra no sea representativa.
 - **Intencional:** el encuestador elige a los que quiere.
 - Inconveniente: la subjetividad del encuestador

Ejemplo: tomar datos de la altura, en centímetros, de las primeras diez personas que pasan por la calle, para calcular la estatura media de los castellano-manchegos.

- ***Muestra aleatoria:*** si se toman datos tomados en una calle cualquiera de una ciudad de Castilla La Mancha.
 - Datos 1: 167, 169, 165, 178, 177, 169, 181, 176, 168 y 175
- ***Muestra Intencionada/no adecuada:*** datos tomados en la puerta de un pabellón polideportivo a la hora en la que salen de su entrenamiento unos jugadores de un equipo de baloncesto. Lo normal es que nos salgan datos más elevados de los normales.
 - Datos 2: 174, 199, 197, 187, 206, 189, 188, 203, 188 y 178

3. Variables estadísticas

Una variable estadística es una característica que se quiere estudiar en una población. Se dividen en:

a) CUALITATIVAS: No son números. Pueden seguir una escala de orden (**cualitativas ordinales**), o no tener un criterio de ordenación (**cualitativas nominales**). Ejemplos:

- a. Color preferido de un grupo de gente (nominal)
- b. Partido al que votarás en las siguientes elecciones (nominal)
- c. Intensidad del dolor en pacientes recién operados con respuestas muy leve, leve, moderado, intenso y muy intenso (ordinal).

b) CUANTITATIVAS: Son números, y pueden ser **discretas**, si representan valores numéricos aislados, o **continuas**, si representan valores numéricos en una escala de números real.

- Altura de un grupo de personas (continua)
- Gasto mensual de las familias de una ciudad en hipoteca (continua)
- Número de hermanos (discreta)

4. Organización de datos en tablas de frecuencias

Es muy importante la organización de los datos en forma de tabla, ya que los hace más comprensibles y facilita los cálculos.

En el caso de variables discretas o cuando hay poca cantidad de datos, se pone en la primera columna cada valor de la variable, de forma ordenada.

VALORES x_i	FRECUENCIAS ABSOLUTAS n_i	FRECUENCIAS ABSOLUTAS ACUMULADAS N_i	FRECUENCIAS RELATIVAS f_i	FRECUENCIAS RELATIVAS ACUMULADAS F_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
x_1	n_1	$N_1 = n_1$	$f_1 = n_1/n$	$F_1 = f_1$	$x_1 \cdot n_1$	$x_1^2 \cdot n_1$
x_2	n_2	$N_2 = N_1 + n_2$	$f_2 = n_2/n$	$F_2 = F_1 + f_2$	$x_2 \cdot n_2$	$x_2^2 \cdot n_2$
x_3	n_3	$N_3 = N_2 + n_3$	$f_3 = n_3/n$	$F_3 = F_2 + f_3$	$x_3 \cdot n_3$	$x_3^2 \cdot n_3$
.
.
.
x_k	n_k	$N_k = N_{k-1} + n_k = n$	$f_k = n_k/n$	$F_k = F_{k-1} + f_k = 1$	$x_k \cdot n_k$	$x_k^2 \cdot n_k$
$\sum_{i=1}^k n_i = n$			$\sum_{i=1}^k f_i = 1$		$\sum_{i=1}^k x_i n_i$	$\sum_{i=1}^k x_i^2 n_i$

- ⊙ **Frecuencias absolutas:** repetición de cada valor de la variable.
- ⊙ **Frecuencias absolutas acumuladas:** en cada fila, es la suma de la frecuencia absoluta más la frecuencia acumulada de la fila anterior.
- ⊙ **Frecuencias relativas:** las frecuencias relativas para cada valor es el resultado de dividir su frecuencia absoluta entre el número de datos n .
- ⊙ **Frecuencias relativas acumuladas:** en cada fila, es la suma de la frecuencia relativa más la frecuencia relativa de la fila anterior.
- ⊙ **$x_i n_i$:** producto del valor de la variable por su frecuencia absoluta
- ⊙ **$x_i^2 n_i$:** producto del cuadrado del valor de la variable por su frecuencia absoluta.

Ejemplo: Preguntamos a 30 personas el número de hermanos que tienen, y obtenemos los siguientes resultados.

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

La tabla se construye de la siguiente forma:

VALORES x_i	FRECUENCIAS ABSOLUTAS n_i	FRECUENCIAS ABSOLUTAS ACUMULADAS N_i	FRECUENCIAS RELATIVAS f_i	FRECUENCIAS RELATIVAS ACUMULADAS F_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
0	6	6	0,200	0,200	0	0
1	10	16	0,333	0,533	10	10
2	6	22	0,200	0,733	12	24
3	4	26	0,133	0,866	12	36
4	2	28	0,067	0,933	8	32
5	1	29	0,033	0,966	5	25
6	1	30	0,033	1	6	36
	30		1		53	163

Por ejemplo, para la segunda fila, el valor de la variable es 1 (1 hermano).

- En la columna correspondiente a frecuencias absolutas el valor es 10, ya que 10 personas respondieron que tenían sólo 1 hermano.
- En la columna correspondiente a frecuencias absolutas acumuladas, el valor es 16, ya que se suman la frecuencia absoluta de esta fila (10) con el valor de la frecuencia absoluta acumulada de la fila anterior (6).
- En la columna frecuencias relativas, el valor es 0,333, ya que es el resultado de dividir la frecuencia absoluta para el valor 1 (10) entre el número total de datos que es 30.
- En la columna correspondiente a frecuencias relativas acumuladas, el valor es 0.533, ya que se suman la frecuencia relativa de esta fila (0.333) con el valor de la frecuencia relativa acumulada de la fila anterior (0.200).
- En la columna $x_i \cdot n_i$ el valor resulta de multiplicar el valor de la variable (1) por su frecuencia absoluta (10).
- En la columna $x_i^2 \cdot n_i$ el valor resulta de multiplicar el valor de la variable al cuadrado (1^2) por su frecuencia absoluta (10).

5. Organización de los datos en tablas con intervalos o clases

A veces, cuando trabajamos con variables cuantitativas, la cantidad de resultados distintos puede ser muy elevada (por ejemplo, en un sondeo sobre la edad podemos encontrarnos con 30 ó 40 edades distintas). En estos casos, a la hora de crear la tabla, se agrupan los datos por intervalos (también llamados clases). Para agrupar por intervalos hay que seguir los siguientes pasos.

1. Se obtiene el número de clases (intervalos) en los que vamos a agrupar los datos. Para ello podemos emplear uno de estos métodos:
 - a) Se realiza la raíz cuadrada del número total de datos y se redondea al entero más próximo. Así obtenemos el número de intervalos.
 - b) Utilizar la regla de Sturges que obtiene el número de clases aplicando la fórmula $1 + 3,322 \cdot \log N$, siendo N el número total de datos. El valor resultante se redondea al entero más próximo.
2. A continuación hay que determinar la amplitud de los intervalos. Esto se hace tomando restando al valor máximo de los datos el menor (lo que se conoce como recorrido de los datos), y dividiéndolo entre el número de clases. La amplitud se obtendrá redondeando al ALZA este resultado.

$$\text{Amplitud intervalo} = (\text{max} - \text{min}) / n^{\circ} \text{ de clases}$$

3. Tras esto, es posible que haya un exceso. Para calcular el posible exceso se aplica la siguiente fórmula:

$$\text{Exceso} = (\text{Amplitud intervalo} \times N^{\circ} \text{ de clases}) - \text{Recorrido}$$

$$\text{Donde el Recorrido} = \text{max} - \text{min}$$

4. Si el exceso es inexistente, el valor mínimo del primer intervalo coincidirá con el valor mínimo de los datos y el valor máximo del último intervalo coincidirá con el valor máximo de los datos. Si el exceso es mayor que 0, tendremos:

$$\text{Valor mínimo del primer intervalo} = \text{Min} - \frac{\text{Exceso}}{2}$$

$$\text{Valor máximo del último intervalo} = \text{Max} + \frac{\text{Exceso}}{2}$$

Ejemplo: Se preguntan las edades de 20 personas, arrojando los siguientes datos:

15 17 20 65 34 23 76 54 45 45
46 54 21 67 89 32 41 32 32 19

Para agruparlos en clases, primero tenemos que determinar el número de clases que tendremos. Vamos a aplicar el criterio de la raíz cuadrada: como hay 20 datos, si el número de clases sería el resultado de redondear al entero más próximo $\sqrt{20} = 4,47$, es decir, 4 clases. (Si aplicáramos la regla de Sturges, el número de clases sería el resultado de redondear al entero más próximo $1 + 3,322 \cdot \log 20 = 5,32$, es decir, 5 clases).

El valor máximo de todos los datos es 89, y el más pequeño 15, luego el recorrido es $89 - 15 = 74$.

Ahora dividimos el recorrido entre el número de clases, y redondeamos al alza para obtener la Amplitud del Intervalo:

$$\text{Amplitud del intervalo} = 74 / 4 = 18,5 \rightarrow \text{Amplitud} = 19$$

$$\text{Calculamos el Exceso} = (19 \times 4) - 74 = 76 - 74 = 2$$

Y ahora:

$$\text{Valor mínimo del primer intervalo} = 15 - \frac{2}{2} = 14$$

Por tanto, el intervalo de la primera clase empezará en 14 y al sumarle su amplitud, obtenemos que llegará hasta $14 + 19 = 33$. El segundo intervalo empezará en 33 y al sumarle la amplitud, acabará en $33 + 19 = 52$. Y así sucesivamente construimos todos los intervalos, pudiendo obtener una de las dos siguientes soluciones (ambas válidas):

CLASES O INTERVALOS (Solución 1)	CLASES O INTERVALOS (Solución 2)
[14,33)	[14,33]
[33,52)	(33,52]
[52,71)	(52,71]
[71,90]	(71,90]

Para construir la tabla, añadimos una columna más a la izquierda, con los intervalos, y la columna de valores (xi) ahora se llamará marca de la clase. La marca de la clase corresponde al valor medio de los extremos de cada intervalo.

Intervalo	MARCA DE LA CLASE x_i	FRECUENCIAS ABSOLUTAS n_i	FRECUENCIAS ABSOLUTAS ACUMULADAS N_i	FRECUENCIAS RELATIVAS f_i	FRECUENCIAS RELATIVAS ACUMULADAS F_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
[14,33)	23'5	9	9	0,450	0,450	211,5	4970,25
[33,52)	42'5	5	14	0,250	0,700	212,5	9031,25
[52,71)	61'5	4	18	0,200	0,900	246	15129
[71,90]	80'5	2	20	0,100	1	161	12960,5
		20		1		831	42091

6. Parámetros estadísticos centrales: media

La **media** es el valor medio de todos los datos de la muestra. Si tenemos acceso a todos los datos de la muestra, la forma de calcularla es sumarlos todos y dividir el resultado entre el número total de datos.

Ejemplo: Se pregunta la altura a diez personas obteniendo los datos que se muestran a continuación. ¿Cuál será el valor de su media?

167 169 165 178 177 169 181 176 168 175

La media por tanto se calculará como:

$$\frac{167+169+165+178+177+169+181+176+168+175}{10} = 172'5$$

Cuando trabajamos con tablas, para calcular la media hay que dividir el sumatorio de valores $x_i n_i$ entre el número de valores n .

$$\text{Media: } \bar{x} = \frac{\sum x_i \cdot n_i}{n}$$

Por ejemplo, en la siguiente tabla....

Intervalo	MARCA DE LA CLASE x_i	FRECUENCIAS ABSOLUTAS n_i	FRECUENCIAS ABSOLUTAS ACUMULADAS N_i	FRECUENCIAS RELATIVAS f_i	FRECUENCIAS RELATIVAS ACUMULADAS F_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
[14,33)	23'5	9	9	0,450	0,450	211,5	4970,25
[33,52)	42'5	5	14	0,250	0,700	212,5	9031,25
[52,71)	61'5	4	18	0,200	0,900	246	15129
[71,90]	80'5	2	20	0,100	1	161	12960,5
		20		1		831	42091

... la media será $831/20 = 41,55$

7. Parámetros estadísticos centrales: la mediana y los cuartiles

La mediana viene a ser el valor que está en el centro una vez ordenados los datos. Si trabajamos con todos los datos porque su número es pequeño, para calcular la mediana, primero los ordenamos de menor a mayor, y entonces habrá dos casos:

- Si el número de datos es impar, el dato central de la ordenación será la mediana.

Datos ordenados: 2, 2, 3, 5, 5, 7, 9

Mediana = 5

- Si el número de datos es par, la media de los dos datos centrales será la mediana.

Datos ordenados: 2, 2, 3, 4, 5, 5, 7, 9

Mediana = $(4+5) / 2 = 4,5$

Cuando hay que calcular la mediana a partir de datos agrupados en una tabla, dividimos el número total de datos entre 2 y se busca el resultado en la columna de las Frecuencias Absolutas Acumuladas. Si se encuentra en esta columna, lo tomaremos como indicador y si no está, se coge como indicador el número mayor más cercano. Después nos fijamos en el valor de la variable (o marca de la clase) que le corresponde a ese indicador, dicho valor es la mediana.

Por ejemplo, en la siguiente tabla la mediana será 42,5

Intervalo	MARCA DE LA CLASE x_i	FRECUENCIAS ABSOLUTAS n_i	FRECUENCIAS ABSOLUTAS ACUMULADAS N_i	FRECUENCIAS RELATIVAS f_i	FRECUENCIAS RELATIVAS ACUMULADAS F_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
[14,33)	23'5	9	9	0,450	0,450	211,5	4970,25
[33,52)	42'5	5	14	0,250	0,700	212,5	9031,25
[52,71)	61'5	4	18	0,200	0,900	246	15129
[71,90]	80'5	2	20	0,100	1	161	12960,5
		20		1		831	42091

- *En primer lugar, dividimos el número total de datos (20) entre 2, lo que nos da 10.*
- *Después, en la columna de frecuencias absolutas acumuladas, buscamos el primer número mayor o igual a 10 (14). Ese será el indicador.*
- *Finalmente, la mediana será el valor de la variable o marca de la clase en la misma fila donde encontramos el indicador.*

Para calcular los cuartiles se procede de forma análoga a la mediana. Los **cuartiles** son los tres valores de la variable que dividen a un conjunto de datos ordenados en cuatro partes iguales. Los **cuartiles Q_1 , Q_2 y Q_3** determinan los valores correspondientes al **25%, al 50% y al 75%** de los **datos**, de forma que **Q_2** coincide con la **mediana**. Es decir, el cuartil Q_1 es el valor para el cual quedan por debajo de él el 25% de los datos, y el **cuartil Q_3** es el valor para el cual quedan por debajo de él el 75% de los datos.

Para calcular los cuartiles en datos agrupados por valores:

$$\frac{k \cdot N}{4}, k = 1, 2, 3$$

1. En primer lugar calculamos para cada cuartil $\frac{k \cdot N}{4}, k = 1, 2, 3$
2. Si coinciden con alguna frecuencia absoluta acumulada, el cuartil es la media aritmética entre el dato al que corresponde esa frecuencia y el siguiente
3. Si no coincide con ninguna frecuencia absoluta acumulada, el cuartil será el primer dato cuya frecuencia acumulada es mayor.

Por ejemplo, para la siguiente tabla:

VALORES x_i	FRECUENCIAS ABSOLUTAS n_i	FRECUENCIAS ABSOLUTAS ACUMULADAS N_i	FRECUENCIAS RELATIVAS f_i	FRECUENCIAS RELATIVAS ACUMULADAS F_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
0	6	6	0,200	0,200	0	0
1	10	16	0,333	0,533	10	10
2	6	22	0,200	0,733	12	24
3	4	26	0,133	0,866	12	36
4	2	28	0,067	0,933	8	32
5	1	29	0,033	0,966	5	25
6	1	30	0,033	1	6	36
	30		1		53	163

- **Cuartil Q_1 :** calculamos $30/4 = 7,5$. Como encontramos la primera frecuencia acumulada mayor o igual que 7,5 que es 16 en la segunda fila, el cuartil será $Q_1 = 1$.
- **Cuartil Q_2 :** calculamos $(2 \cdot 30)/4 = 15$. Como encontramos la primera frecuencia acumulada mayor o igual que 15 que es 16 en la segunda fila, el cuartil será $Q_2 = 1$.
- **Cuartil Q_3 :** calculamos $(3 \cdot 30)/4 = 22,5$. Como encontramos la primera frecuencia acumulada mayor o igual que 22,5 que es 26 en la cuarta fila, el cuartil será $Q_3 = 3$.

Para calcular los cuartiles en datos agrupados en intervalos:

1. En primer lugar buscamos la clase donde se encuentra $\frac{k \cdot N}{4}$, $k = 1, 2, 3$, en la tabla de las frecuencias acumuladas.

2. Después se calcula como:

$$Q_1 = L_i + \frac{\frac{N}{4} - N_{i-1}}{n_i} \cdot a_i \quad Q_2 = L_i + \frac{\frac{2N}{4} - N_{i-1}}{n_i} \cdot a_i \quad Q_3 = L_i + \frac{\frac{3N}{4} - N_{i-1}}{n_i} \cdot a_i$$

donde

L_i es el límite inferior de la clase donde se encuentra el cuartil.

N es la suma de las frecuencias absolutas.

N_{i-1} es la **frecuencia acumulada anterior** a la clase del cuartil.

a_i es la amplitud de la clase.

Por ejemplo, en la siguiente tabla el cálculo de los cuartiles se hace de la siguiente forma:

Intervalo	MARCA DE LA CLASE x_i	FRECUENCIAS ABSOLUTAS n_i	FRECUENCIAS ABSOLUTAS ACUMULADAS N_i	FRECUENCIAS RELATIVAS f_i	FRECUENCIAS RELATIVAS ACUMULADAS F_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
[14,33)	23'5	9	9	0,450	0,450	211,5	4970,25
[33,52)	42'5	5	14	0,250	0,700	212,5	9031,25
[52,71)	61'5	4	18	0,200	0,900	246	15129
[71,90]	80'5	2	20	0,100	1	161	12960,5
		20		1		831	42091

Cuartil Q_1 : calculamos $20/4 = 5$. Como encontramos la primera frecuencia acumulada mayor o igual que 5 que es 9, el primer cuartil está en la primera clase, y aplicamos la fórmula:

$$Q_1 = 14 + \frac{\frac{20}{4} - 0}{9} \cdot 19 = 24'55$$

Cuartil Q_2 : calculamos $40/4 = 10$. Como encontramos la primera frecuencia acumulada mayor o igual que 5 que es 14, el segundo cuartil está en la segunda clase, y aplicamos la fórmula:

$$Q_2 = 33 + \frac{\frac{40}{4} - 9}{5} \cdot 19 = 36'8$$

Cuartil Q_3 : calculamos $60/4 = 15$. Como encontramos la primera frecuencia acumulada mayor o igual que 5 que es 18, el tercer cuartil está en la tercera clase, y aplicamos la fórmula:

$$Q_3 = 52 + \frac{\frac{60}{4} - 14}{4} \cdot 19 = 56'75$$

8. Parámetros estadísticos centrales: la moda

La moda es el valor de la variable que más se repite. Si tenemos todos los datos, aquel o aquellos valores que más se repitan serán la moda o modas (puede haber más de una).

Datos: 2, 2, 3, 5, 5, 7, 9

Modas = 2 y 5 (ambos se repiten 2 veces)

Cuando se trabaja con tablas, buscamos el mayor valor (o los mayores) en la tabla de Frecuencias Absolutas, y el valor que tome en esa fila la variable o marca de la clase será una Moda.

Por ejemplo, en la siguiente tabla la moda será $Mo = 23,5$.

Intervalo	MARCA DE LA CLASE x_i	FRECUENCIAS ABSOLUTAS n_i	FRECUENCIAS ABSOLUTAS ACUMULADAS N_i	FRECUENCIAS RELATIVAS f_i	FRECUENCIAS RELATIVAS ACUMULADAS F_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
[14,33)	23'5	9	9	0,450	0,450	211,5	4970,25
[33,52)	42'5	5	14	0,250	0,700	212,5	9031,25
[52,71)	61'5	4	18	0,200	0,900	246	15129
[71,90]	80'5	2	20	0,100	1	161	12960,5
		20		1		831	42091

- *Se observa que el valor más alto en la columna de frecuencias absolutas es 9*
- *La moda será el valor de la variable o marca de la clase en la misma fila donde encontramos el indicador.*

9. Parámetros estadísticos de dispersión

Las medidas de dispersión, variabilidad o variación nos indican si esos datos están próximos entre sí o si están dispersos, es decir, nos indican cuán esparcidos se encuentran los datos. Estas medidas de dispersión nos permiten apreciar la distancia que existe entre los datos a un cierto valor central.

- Ejemplo: Si hacemos una encuesta de edades y tenemos de resultados 11, 15, 16, 84, 76 y 69, la media sería aproximadamente 45 años. Si nos fijamos, los datos están muy dispersos con respecto a esa media, ya que la diferencia de edades con respecto a la media, en cada caso, sería 34, 30, 29, 39, 31 y 24 años. Realmente, ninguna edad es cercana a la media.

10. Parámetros estadísticos de dispersión: desviación media

La **desviación media** es la **media aritmética** de los **valores absolutos de las desviaciones respecto a la media**. La **desviación media** se representa por

$$D_{\bar{x}} = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \dots + |x_n - \bar{x}|}{N}$$

$$D_{\bar{x}} = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{N}$$

Ejemplo: Calcular la desviación media de la distribución 9, 3, 8, 8, 9, 8, 9, 18

$$\bar{x} = \frac{9 + 3 + 8 + 8 + 9 + 8 + 9 + 18}{8} = 9$$

$$D_{\bar{x}} = \frac{|9 - 9| + |3 - 9| + |8 - 9| + |8 - 9| + |9 - 9| + |8 - 9| + |9 - 9| + |18 - 9|}{8} = 2.25$$

Si los datos vienen agrupados en una tabla de frecuencias, la expresión de la **desviación media** es:

$$D_{\bar{x}} = \frac{\sum_{i=1}^n |x_i - \bar{x}| n_i}{N}$$

Ejemplo: En la tabla anterior teníamos...

Intervalo	MARCA DE LA CLASE x_i	FRECUENCIAS ABSOLUTAS n_i	FRECUENCIAS ABSOLUTAS ACUMULADAS N_i	FRECUENCIAS RELATIVAS f_i	FRECUENCIAS RELATIVAS ACUMULADAS F_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
[14,33)	23'5	9	9	0,450	0,450	211,5	4970,25
[33,52)	42'5	5	14	0,250	0,700	212,5	9031,25
[52,71)	61'5	4	18	0,200	0,900	246	15129
[71,90]	80'5	2	20	0,100	1	161	12960,5
		20		1		831	42091

...cuya media era $831/20 = 41,55$

Para calcular la desviación media hacemos:

Intervalo	MARCA DE LA CLASE x_i	FRECUENCIAS ABSOLUTAS n_i	$ x_i - \text{media} $	$ x_i - \text{media} n_i$
[14,33)	23'5	9	$ 23'5 - 41'55 = 18'05$	$18'05 \cdot 9 = 162'45$
[33,52)	42'5	5	$ 42'5 - 41'55 = 0'95$	$0,95 \cdot 5 = 4'75$
[52,71)	61'5	4	$ 61'5 - 41'55 = 19'95$	$19'95 \cdot 4 = 79'8$
[71,90]	80'5	2	$ 80'5 - 41'55 = 38'95$	$38'95 \cdot 2 = 77'9$
		20		TOTAL = 324'9

... Y la desviación media será $324'9/20 = 16,25$

11. Parámetros estadísticos de dispersión: la varianza

La varianza, al ser un parámetro de dispersión, sirve para identificar si los datos están cercanos a la media o no. Su valor mínimo es 0, cuando todos los datos sean iguales a la media. Cuando los datos más se acercan a la media, más pequeño será su valor.

Se calcula sumando los valores que se obtienen de elevar al cuadrado la diferencia de cada dato con la media, y dividiendo este valor entre el número de datos. Para representar este parámetro se utilizan los símbolos S^2 o σ^2 .

$$S^2 = \frac{1}{n} \sum x_i^2 n_i - \bar{x}^2$$

En las tablas, se calcula de la siguiente manera:

Intervalo	MARCA DE LA CLASE x_i	FRECUENCIAS ABSOLUTAS n_i	FRECUENCIAS ABSOLUTAS ACUMULADAS N_i	FRECUENCIAS RELATIVAS f_i	FRECUENCIAS RELATIVAS ACUMULADAS F_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
[14,33)	23'5	9	9	0,450	0,450	211,5	4970,25
[33,52)	42'5	5	14	0,250	0,700	212,5	9031,25
[52,71)	61'5	4	18	0,200	0,900	246	15129
[71,90]	80'5	2	20	0,100	1	161	12960,5
		20		1		831	42091

Siendo la media 41,55

$$S^2 = \frac{1}{20} \cdot 42091 - (41,55)^2 = 378,15$$

12. Parámetros estadísticos de dispersión: la desviación típica

La desviación típica da un valor de las diferencias de los valores con respecto a la media que se obtiene haciendo la raíz cuadrada de la varianza, lo que hace que el valor sea más comprensible y manejable que el obtenido con la propia varianza.

Varianza y Desviación típica:

$$S^2 = \frac{1}{n} \sum x_i^2 n_i - \bar{x}^2 = \frac{203}{30} - (2'1)^2 = 6'7666 - 4'41 = 2'3566$$

$$S = +\sqrt{S^2} = +\sqrt{2'3566} = 1'5351$$

Por ejemplo, si

$$S^2 = \frac{1}{20} \cdot 42091 - (41,55)^2 = 378,15$$

Entonces

$$S = +\sqrt{S^2} = +\sqrt{378,15} = 19,45$$

13. Parámetros estadísticos de dispersión: el coeficiente de variación

Es el cociente entre la desviación típica y la media. Su fórmula expresa la desviación estándar como porcentaje de la media aritmética, mostrando una mejor interpretación porcentual del grado de variabilidad que la desviación típica o estándar. A mayor valor del coeficiente de variación mayor heterogeneidad de los valores de la variable; y a menor C.V., mayor homogeneidad en los valores de la variable. Suele representarse por medio de las siglas **C.V.**, y se calcula como:

$$C_V = \frac{\sigma}{|\bar{x}|}$$

Por ejemplo, si

$$\text{Media} = 41,55$$

$$S^2 = \frac{1}{20} \cdot 42091 - (41,55)^2 = 378,15$$

$$S = +\sqrt{S^2} = +\sqrt{378,15} = 19,45$$

Entonces

$$Cv = \frac{\text{Desviación típica}}{\text{Media}} = \frac{19,45}{41,55} = 0,47$$

Si se multiplica por cien, se expresa en tanto por ciento (en este ejemplo 47%)

14. Representación gráfica: Diagramas de barras

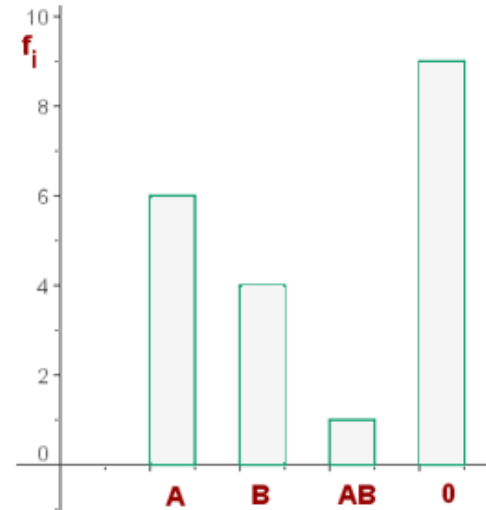
Una gráfica estadística es la mejor forma de disponer de toda la información que se haya recogido con una simple “ojeada” y que además permite distinguir, sin dificultad alguna, que opción es la preferida por los encuestados.



Los diagramas de barras se caracterizan por:

- ⦿ **Son los indicados para variables estadísticas cualitativas o cuantitativas discretas (sin intervalos)**
- ⦿ Se representan sobre el eje de abscisas los valores de la variable y sobre el de ordenadas las frecuencias asociadas a cada valor
- ⦿ Se levanta sobre cada valor de la variable un segmento vertical de altura igual a la frecuencia con que se ha observado dicho valor.

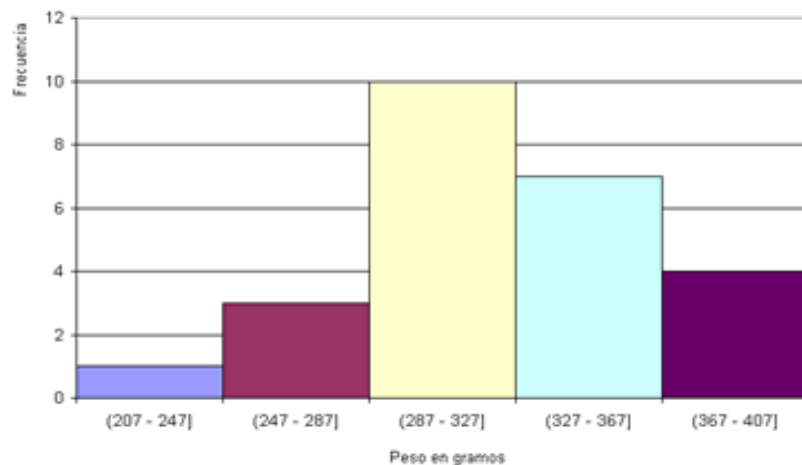
Grupo sanguíneo	f_i
A	6
B	4
AB	1
O	9
	20



15. Representación gráfica: Histogramas

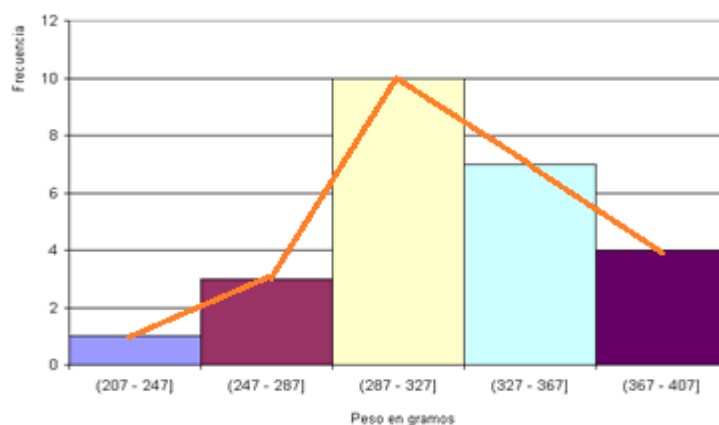
Los histogramas se caracterizan por:

- Se utilizan para variables estadísticas cuantitativas expresadas en intervalos
- Sobre el eje de abcisas se representan las distintas clases o intervalos en los que se han agrupado los valores de la variable, y sobre cada clase se construye un rectángulo cuya base sea el intervalo y la altura la frecuencia absoluta de dicha clase. Las barras quedaran gráficamente unidas unas a otras.



16. Representación gráfica: Polígonos de frecuencia

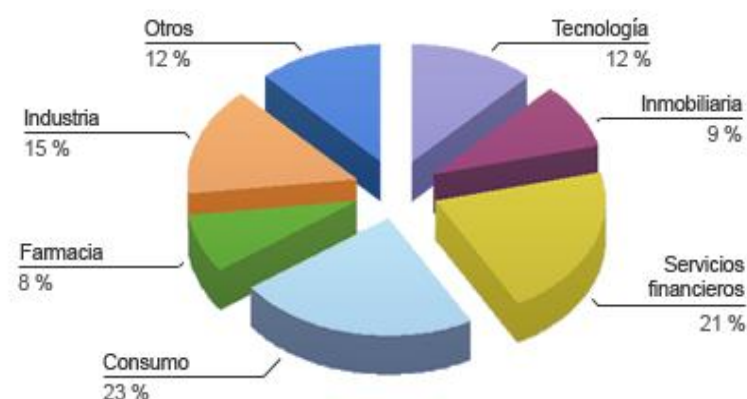
Uniendo los puntos medios de las bases superiores de los rectángulos de un histograma se dibuja lo que se conoce como polígono de frecuencias.



17. Representación gráfica: Diagramas de sectores

Los diagramas de sectores se caracterizan por:

- Se utilizan para caracteres cualitativos y cuantitativos.
- Consiste en repartir el área del círculo en sectores de tamaño proporcional a la frecuencia de cada valor que ha presentado un determinado carácter.



Para hacer el diagrama de sectores, hay que hacer unos pequeños cálculos que nos dicen el ángulo de cada porción del diagrama. En concreto, para cada valor de la variable o marca de la clase, el ángulo que tomará del círculo vendrá definido por:

$$\text{Ángulo de la variable o clase} = \frac{\text{frecuencia absoluta de la clase}}{\text{número total de datos}} \cdot 360^\circ$$

Ejemplo:

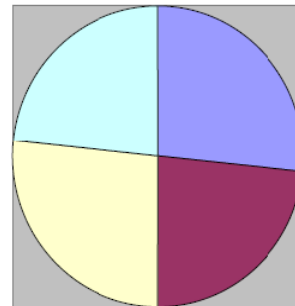
INTERVALOS $L_{i-1} - L_i$	MARCA DE CLASE (x_i)	FRECUENCIAS ABSOLUTAS (n_i)	FRECUENCIAS ABSOLUTAS ACUMULADAS (N_i)	$x_i n_i$	$x_i^2 n_i$
[33,43)	38	8	8	304	11552
[43,53)	48	7	18	336	16128
[53,63)	58	8	23	464	26912
[63,73)	68	7	30	476	32368
		30		1580	86960

$$1^a. - \frac{360^\circ}{30} = \frac{x_1^\circ}{8} \Rightarrow x_1^\circ = \frac{8}{30} 360^\circ = 96^\circ$$

$$2^a. - \frac{360^\circ}{30} = \frac{x_2^\circ}{7} \Rightarrow x_2^\circ = \frac{7}{30} 360^\circ = 84^\circ$$

$$3^a. - \frac{360^\circ}{30} = \frac{x_3^\circ}{8} \Rightarrow x_3^\circ = \frac{8}{30} 360^\circ = 96^\circ$$

$$4^a. - \frac{360^\circ}{30} = \frac{x_4^\circ}{7} \Rightarrow x_4^\circ = \frac{7}{30} 360^\circ = 84^\circ$$



Ejercicios –

1. ¿Cuál de las siguientes informaciones te parece claramente manipulada o errónea?

- a. Según un estudio estadístico, realizado a dos personas en un club náutico, se determina que a todos los españoles les encanta el buceo deportivo.
- b. Según un estudio estadístico, realizado por una compañía eléctrica, se sabe que los andaluces no aprecian que haya contaminación generada por las centrales térmicas en nuestro territorio.
- c. Un estudio estadístico determina que el cien por cien de los encuestados respiran cada día.
- d. Todas las opciones anteriores son estudios manipulados o sin sentido.

2. Se quiere conocer la cantidad de CO₂ que hay en el aire en una determinada población. ¿Cuál sería la opción más adecuada para llevar a cabo este estudio?

- a. Crear un cuestionario abierto preguntando por la cantidad de CO₂ que hay en el aire
- b. Crear un cuestionario cerrado con las respuestas: 20 mg/m³, 10 mg/m³ y otra cantidad.
- c. Instalar un aparato medidor en algún punto de la ciudad que registre los datos de cantidad de CO₂ que hay en el aire a lo largo de un periodo determinado de tiempo.

3. Indica si las siguientes variables aleatorias son cualitativas o cuantitativas

- A. Energía aportada por distintas marcas de muesli
- B. Sistema de calefacción utilizado en el invierno por familias de Madrid
- C. Volumen de basura generado por las familias de una barriada de Toledo
- D. Soluciones al problema de la contaminación de las aguas

4. Se quiere estudiar el nivel de contaminación del agua de un determinado río. Elige la opción más adecuada para elegir la muestra:

- a. Se cogería una muestra de agua al azar de cualquier zona del cauce del río.
- b. Se tomarían varias muestras de agua al azar de distintas zonas a lo largo del cauce del río y en distintos períodos de tiempo.
- c. Se tomaría una muestra de agua al lado de una fábrica que vierte sus residuos directamente al cauce del río.
- d. Se tomaría una muestra de agua en el lugar de nacimiento del río.


5. Estás realizando un estudio estadístico para conocer la satisfacción de la gente del barrio con el nuevo polideportivo. ¿Qué forma de elegir la muestra crees que es mejor?

- a. Preguntar a 50 personas que estén en el polideportivo.
- b. Preguntar a 50 personas de tus amistades.
- c. Elegir al azar 50 números de teléfono de casas del barrio, llamar y preguntar.
- d. Preguntar a 50 personas que estén por la mañana comprando en el mercado.

6. En un determinado paraje se ha medido la altura de 10 olivos, siendo sus alturas 3,5 m; 3,8 m; 3,4 m; 3,1 m; 3,6 m; 3,8 m; 3 m; 3,7 m; 2,8 m; 3,3 m. La altura media de los diez olivos del paraje es de:

- a. 3 m
- b. 3,4 m
- c. 4 m

7. Se realiza una encuesta a 100 personas preguntando si separan o no los residuos para reciclarlos, siendo los resultados los recogidos en esta tabla:

	Nº de respuestas
Siempre, clasificando en las categorías: orgánica, vidrio, envases y papel.	10
Siempre, pero sólo papel y vidrio.	15
Casi siempre el papel	23
Casi siempre el vidrio	18
Normalmente no	16
Nunca	10
Otras opciones	8

La Moda es:

- a. Casi siempre el papel.
- b. Siempre, clasificando en las categorías: orgánica, vidrio, envases y papel.
- c. Casi siempre el vidrio.

8. En una recogida de datos sobre los metros cuadrados ocupados por las distintas zonas verdes en dos localidades datos:

Localidad 1

	m ² zona verde
Zona1	780
Zona2	1080
Zona3	2200
Zona4	2800
Zona5	5600
Zona6	950
Zona7	4200
Zona8	2600
Zona9	4100
Zona10	3500

Localidad 2

	m ² zona verde
Zona1	4500
Zona2	600
Zona3	1800
Zona4	5400
Zona5	1000
Zona6	700
Zona7	1900
Zona8	6100

¿Cuál de las dos localidades presenta una distribución de zonas verdes más “dispersa”? (Haría falta calcular el coeficiente de variación de los metros cuadrados destinados a zona verde de ambas localidades)

9. Hemos consultado, en diferentes comercios, el precio (en euros) de un determinado modelo de impresora, obteniendo los datos siguientes:

146 - 150 - 141 - 143 - 139 - 144 - 133 - 153

- Calcula el precio medio.
- ¿Cuál es la mediana?
- Halla el recorrido.
- Halla la desviación típica.

10. En la familia Fernández, el salario mensual del padre es de 950 €, y el salario de la madre, 1 600 €. En la familia Torres, el padre gana 1 800 € al mes, y la madre 750 €.

- ¿Cuál es el sueldo medio de cada familia?
- ¿En cuál de ellas es mayor la dispersión?
- ¿Cuál es el rango en cada familia?

11. En un control de velocidad en carretera se obtuvieron los siguientes datos:

VELOCIDAD (km/h)	N.º DE COCHES
60 - 70	5
70 - 80	15
80 - 90	27
90 - 100	38
100 - 110	23
110 - 120	17

- Haz una tabla reflejando las marcas de clase y las frecuencias.
- Calcula la media y la desviación típica.
- ¿Qué porcentaje circula a más de 90 km/h?

12. Los puntos conseguidos por Teresa y por Rosa en una semana de entrenamiento, jugando al baloncesto, han sido los siguientes:

TERESA	16	25	20	24	22	29	18
ROSA	23	24	22	25	21	20	19

- Halla la media de cada una de las dos.
- Calcula la desviación típica y el coeficiente de variación. ¿Cuál de las dos es más regular?

13. A la pregunta: ¿cuántas personas forman tu hogar familiar?, 40 personas respondieron esto:

4 5 3 6 3 5 4 6 3 2
 2 4 6 3 5 3 4 5 3 6
 4 5 7 4 6 2 3 4 4 3
 4 4 5 3 2 6 3 7 4 3

- Haz la tabla de frecuencias y el diagrama correspondiente.
- Calcula la media, la mediana, la moda y la desviación típica.

14. En un test de inteligencia realizado a una muestra de 200 personas, se han obtenido los resultados siguientes:

PUNTUACION	N.º DE PERSONAS
30 - 40	6
40 - 50	18
50 - 60	76
60 - 70	70
70 - 80	22
80 - 90	8

a) Dibuja un histograma para representar gráficamente los datos y haz también el polígono de frecuencias.

b) Calcula la media y la desviación típica.

15. Al medir el peso al nacer en una determinada especie de animales, hemos obtenido los datos siguientes:

PESO (kg)	N.º ANIMALES
3,5 - 4,5	1
4,5 - 5,5	8
5,5 - 6,5	28
6,5 - 7,5	26
7,5 - 8,5	16
8,5 - 9,5	1

- Representa estos datos con el gráfico adecuado.
- Calcula la media y la desviación típica.
- ¿Qué porcentaje de animales pesó entre 5,5 kg y 6,5 kg?
- ¿Y entre 4,5 kg y 8,5 kg?

16. Estas son las horas de estudio semanal de un grupo de alumnas y alumnos:

14	9	9	20	18	12	14	6	14	8
15	10	18	20	2	7	18	8	12	10
20	16	18	15	24	10	12	25	24	17
10	4	8	20	10	12	16	5	4	1
									3

- Reparte estos datos en los intervalos: 1,5-6,5; 6,5-11,5; 11,5-16,5; 16,5-21,5; 21,5-26,5
- Haz la tabla de frecuencias y el histograma.
- Calcula la media y la desviación típica.

17. Los gastos mensuales de una empresa A tienen una media de 60 000 € y una desviación típica de 7 500 €. En otra empresa más pequeña B, la media es 9 000 €, y la desviación típica, 1 500 €. Calcula, mediante el coeficiente de variación, cuál de las dos tiene más variación relativa.

18. Se ha estudiado el grupo sanguíneo de 200 personas, así como el Rh. Los resultados se nos han dado en esta tabla, denominada TABLA DE CONTINGENCIA, que está incompleta.

	GRUPO A	GRUPO B	GRUPO AB	GRUPO O	TOTALES
RH+	74		6	70	162
RH-		3	1		
TOTALES				86	200

- Completa la tabla
- ¿Qué porcentaje de la población estudiada tiene Rh negativo?
- ¿Qué porcentaje tiene grupo B+?
- Dentro de los de RH+, ¿qué porcentaje tiene el grupo A?
- Haz un diagrama de sectores

- 18.** Al observar las notas de un mismo examen en dos grupos de tercero de ESO, se comprueba que en una clase hay siete personas que han tenido un 1 y cinco que han tenido un 10, mientras que en otra hay sólo 2 personas que han sacado un 1 y tres que han tenido un 10. Sabemos, además, las medias y desviación típica en cada una de las clases, que son:

	NOTA MEDIA	DESVIACIÓN TÍPICA
Tercero A	5,43	3,01
Tercero B	5,56	1,35

¿En qué clase hay mayor número de dieces, en Tercero A o en Tercero B?

Ejercicios de las pruebas de acceso

- 19.** El servicio de urgencias de un centro de salud ha atendido en los últimos 20 días, en horario de 0:00 horas a 8:00 horas, las siguientes urgencias:

2, 3, 1, 0, 2, 4, 5, 4, 1, 2, 1, 0, 2, 1, 3, 4, 5, 4, 2 y 2

- Construya la tabla de frecuencias de la distribución
- Determine moda, mediana y media aritmética de la distribución.
- Calcula la varianza, la desviación típica.

- 20.** El porcentaje de población activa dedicada a la agricultura en 30 países africanos es:

47	24	70	63	91	61	63	75	56	57
68	74	77	69	68	70	75	64	37	36
65	91	62	14	66	81	24	66	63	43

- Agrupa estos datos en cinco intervalos de igual amplitud
- Calcula la media, moda y mediana
- Calcula la varianza, la desviación típica y el coeficiente de variación.

21. La edad de los asistentes a dos congresos se distribuye según esta tabla:

Edad	(28,34]	(34,40]	(40,46]	(46,52]	(52,58]	(58,64]
Congreso A	10	20	30	40	30	20
Congreso B	30	20	30	20	30	20

a) Calcula la media de asistentes a cada uno de los congresos

b) Calcula en cada caso la desviación típica.

c) Comenta los resultados obtenidos en los apartados anteriores comparando la distribución de las edades de los asistentes a cada uno de los congresos.

22. Los expertos en baloncesto quieren hacer estudios comparativos sobre las estaturas de los jugadores de 1ª división. Las estaturas de los jugadores de dos equipos (A y B) son:

EqA	180	186	193	196	202	206	210	184	199	203	207	189	188	183
EqB	186	192	198	204	208	188	193	199	209	194	199	194	181	205

Compara, a partir de estos datos, la altura de los dos equipos, llevando a cabo las siguientes cuestiones, para cada equipo:

a) Agrupa estos datos en seis intervalos de igual amplitud.

b) Calcula la media, la moda y la mediana

c) Calcula la varianza, desviación típica y coeficiente de variación

23. Dada la distribución estadística de la siguiente tabla:

x_i	[0,5]	[5,10]	[10,15]	[15,20]	[20,25]	[25,30]
f_i	5	7	9	10	4	7

a) Calcula la media, moda, mediana y cuartiles Q_1 y Q_3

b) Halla la varianza y la desviación típica