Estadística

(Se corresponde con el tema 14 del libro de Oxford de 4ºESO Opc. B)

1. Conceptos Básicos

La Estadística es la ciencia que se encarga de recopilar y ordenar datos referidos a diversos fenómenos para su posterior análisis e interpretación.

Ejemplos:
1) Analizar el consumo energético o el número de turistas que visitaron el país, así como realizar sondeos electorales son algunas de las aplicaciones de la Estadística.
2) Hacer encuestas de opinión, estudios de consumo y censos de población son también aplicaciones de la Estadística.

1.1. Población y Muestra

Llamamos **población** al conjunto de todos los elementos objeto de estudio. Cada uno de los elementos que forman la población se denomina **individuo**. En muchos casos no se puede trabajar con todos los elementos de una población y hacemos el estudio con una parte de ellos. A este conjunto de elementos se le llama muestra y al número de elementos de ésta **tamaño de la muestra**.

Ejemplo:
Los alumnos que cursan 3º ESO en cierta ciudad son 6578.
- Los 6578 alumnos constituyen la población objeto de estudio.
- Los alumnos de 3º ESO del IES "Cervantes" de dicha ciudad son una muestra de la población. El número de alumnos de 3º ESO de dicho instituto, 74, es el tamaño de la muestra.
- Cada uno de los 6578 alumnos es un individuo de la población.
- En esta población podemos estudiar el número de hermanos que tiene cada uno, la estatura, el peso, la edad, aficiones musicales, etc.

1.2. Variables estadísticas

Se llama variable estadística a cada una de las propiedades o características que podemos estudiar de un conjunto de datos.

Ejemplo: la estatura, el peso, la edad, aficiones, etc, que podemos estudiar en una población.

Las variables estadísticas se clasifican en:

Tipos	*Propiedades*	*Ejemplos*
Cualitativas	Los valores de la variable no son números sino cualidades.	- Sexo (Varón, Hembra) - Color de ojos
Cuantitativas	Los valores que toma la variable son números.	- Peso. - Número de hermanos.

A su vez las variables cuantitativas se pueden clasificar en:

- Discreta: la variable sólo toma entre cada dos posibles valores un número determinado de valores. Por ejemplo: en el número de amigos entre 2 y 5 sólo puede tomar los valores 3 y 4 pero no 3'5.

- Continua: la variable puede tomar entre cada dos posibles valores tantos valores como queramos. Por ejemplo: en la estatura entre 170 y 172 cm pueden existir infinidad de valores posibles: 170'8, 170'83, 171'2, etc.

2. Frecuencias

Después de recopilar los datos para el estudio se procede a su recuento para expresarlos de manera ordenada, generalmente en forma de tabla.

Si la variable es continua, los datos se agrupan en intervalos o clases, usualmente de la misma amplitud. Al efectuar cálculos estadísticos usando intervalos, a todos los datos de un mismo intervalo se les da el mismo valor, que se llama **marca de clase** y es el punto medio del intervalo.

Ejemplos:
1) Anotamos los libros leídos por 23 alumnos durante el último año:

Datos: 1, 3, 4, 2, 2, 3, 2, 2, 1, 3, 3, 1, 1, 2, 4, 4, 2, 3, 3, 2, 3, 3, 3.

Agrupamos los datos en una tabla (tabla de frecuencias):

Número de libros leídos	***Recuento (Frecuencias)***
1	4
2	7
3	9
4	3

2) El peso en kg de 20 alumnos es: 66'5, 59'2, 60'1, 64'2, 70, 50, 41'6, 47'9, 42'8, 55, 52'2, 50'3, 42'2, 61'9, 52'4, 49'2, 41'6, 38'7, 36'5 y 45. Construyamos la tabla de frecuencias.

La amplitud de cada intervalo viene dada por la fórmula: $\frac{Val.\,Max. - Val.\,Min.}{\sqrt{N}}$, donde Val. Max. es el valor máximo que toma la variable, Val. Min. es el valor mínimo que toma y N es el número total de datos.

En nuestro caso Val. Max = 70, Val. Min = 36'5 y N = 20. Por tanto, la amplitud de cada intervalo será:

$\frac{70-36'5}{\sqrt{20}}=\frac{33'5}{\sqrt{20}}=7'4908277$, que al no ser entero la aproximamos por exceso al siguiente entero, 8.

33'5 es la diferencia entre el valor mayor y el menor. Dividimos dicha diferencia entre 8 y al no ser entero dicho cociente aproximamos el resultado por exceso al siguiente entero. En este caso sale 5, luego:

Tomaremos 5 intervalos de amplitud 8, empezando desde la parte entera del valor menor : 36.

36+8 = 44 ----------> [36, 44) (es el primer intervalo)
44+8 = 52 ----------> [44, 52) (es el segundo intervalo)
52+8 = 60 ----------> [52, 60) (es el tercer intervalo)
60+8 = 68 ----------> [60, 68) (es el cuarto intervalo)
68+8 = 76 ----------> [68, 76) (es el quinto intervalo)

La marca de clase es el punto medio de cada intervalo (se halla sumando los extremos de cada intervalo y dividiendo el resultado obtenido entre dos.

Debemos poner todos estos datos en una tabla:

Intervalo de clase	***Marca de clase***	***Recuento (Frecuencia)***
[36, 44)	40	6
[44, 52)	48	5
[52, 60)	56	4
[60, 68)	64	4
[68, 76)	72	1

Llamaremos:

- Frecuencia absoluta de un dato de la variable, x_i, es el número de veces que aparece. Se representa por n_i. La suma de frecuencias absolutas es igual al número total de datos, N.
- Frecuencia relativa de un valor de la variable, x_i, es el cociente entre la frecuencia absoluta, n_i y el número total de datos, N. Se representa por f_i. Así pues: $f_i = \frac{n_i}{N}$. La suma de todas las frecuencias relativas es igual a la unidad.
- Frecuencia absoluta acumulada de un dato, x_i, es la suma de todas las frecuencias absolutas de los valores menores o iguales que él. Se suele representar como N_i. Es decir: $N_i = \sum_{j=1}^{i} n_j$. La última es igual a N.
- Frecuencia relativa acumulada de un dato, x_i, es la suma de todas las frecuencias relativas de los valores menores o iguales que él. Se suele representar como F_i. Es decir: $F_i = \sum_{j=1}^{i} f_j = \sum_{j=1}^{i} \frac{n_j}{N}$. La última es igual a 1.
- Además damos la frecuencia porcentual de una modalidad, $\mathbf{p_i}$ y también la frecuencia acumulada relativa porcentual $\mathbf{\%F_i}$ o $\mathbf{P_i}$. Es decir:

$\boldsymbol{p_i = 100 \cdot f_i}$ (expresada en %) y $\boldsymbol{P_i = \sum_{1}^{i} p_j = 100 \cdot F_i}$

Ejemplos:

1) Las alturas de un grupo de niños, en cm, son:

130, 128, 141, 139, 137, 126, 135, 136, 134, 131, 143, 140, 129, 128, 137, 136, 142, 138, 144, 136, 139,

126, 127, 143, 135, 139, 139, 141.

La variable estadística "altura" es cuantitativa continua por lo que podremos agrupar los datos en intervalos.

N = 28 y la amplitud del intervalo viene dada por la fórmula: $\frac{Val.\ Max. - Val.\ Min.}{\sqrt{N}}$.

Val. Max. = 144 y Val. Min. = 126, con lo que tendremos: $\frac{144-126}{\sqrt{28}} = 3'40168...$, que aproximándola por exceso a una cantidad entera será de 4.

Así pues, la amplitud de cada intervalo será de 4.

144-126 = 18

18/4 = 4'5. Como 4'5 intervalos no puedo tomar elijo 5.

Tomaremos 5 intervalos de amplitud 4. Podemos empezar en el 125 (valor entero menor que el primero).

Entonces formamos la siguiente tabla:

Intervalos de clase	***Marca de clase (x_i)***	***Frecuencias Absolutas (n_i)***	***Frecuencias Relativas (f_i)***	***Frecuencias Absolutas Acumuladas (N_i)***	***Frecuencias Relativas Acumuladas (F_i)***
[125, 129)	127	5	0'1786	5	0'1786
[129, 133)	131	3	0'1071	8	0'2857
[133, 137)	135	6	0'2143	14	0'5
[137, 141)	139	8	0'2857	22	0'7857
[141, 145)	143	6	0'2143	28	1
Suma		28	1		

2) La talla de calzado en una clase de 26 alumnos es: 43, 42, 41, 39, 41, 37, 40, 43, 44, 40, 39, 39, 38, 41, 40, 39, 38, 39, 39, 40, 37, 42, 41, 37, 42, 38. Construye su tabla de frecuencias y de porcentajes asociada.

Como el número de calzado es una variable cuantitativa discreta no hemos de agrupar los datos en intervalos de clase. Contamos el número de veces que aparece cada valor x_i, f_i. Dividiendo cada uno de estos valores entre N obtendremos h_i. Si ahora multiplicamos la frecuencia relativa por 100 obtendremos los porcentajes de las frecuencias absolutas. Ya hemos visto como se calculan las frecuencias absolutas acumuladas. Para calcular la columna % F_i podremos utilizar la columna % f_i y las frecuencias relativas acumuladas, H_i, se pueden calcular a partir de % F_i sin más que dividir entre 100 cada resultado.

Valores (x_i)	***Frecuencias Absolutas (n_i)***	***Frecuencias Relativas (f_i)***	***Porcentajes Frecuencias Absolutas (p_i)***	***Frecuencias Absolutas Acumuladas (N_i)***	***Porcentajes Frecuencias Absolutas Acumuladas (P_i)***	***Frecuencias Relativas Acumuladas (F_i)***
37	3	0'1154	11'54	3	11'54	0'1154
38	3	0'1154	11'54	6	23'08	0'2308
39	6	0'2308	23'08	12	46'15	0'4616
40	4	0'1538	15'38	16	61'54	0'6154
41	4	0'1538	15'38	20	76'92	0'7692
42	3	0'1154	11'54	23	88'46	0'8846

Valores (x_i)	*Frecuencias Absolutas* (n_i)	*Frecuencias Relativas* (f_i)	*Porcentajes Frecuencias Absolutas* (p_i)	*Frecuencias Absolutas Acumuladas* (N_i)	*Porcentajes Frecuencias Absolutas Acumuladas* (P_i)	*Frecuencias Relativas Acumuladas* (F_i)
43	2	0'0769	7'69	25	96'15	0'9615
44	1	0'0385	3'85	26	100	1
Suma	26	1	100			

Ejercicio:

1) Los pesos, en kg, de un grupo de personas son: 68'5, 34'2, 47'5, 39'2, 80, 63'4, 58'6, 50'2, 60'5, 70'8, 30'5, 42'7, 59'4, 39'3, 48'6, 56'8, 72.
   a) Obtén la tabla de frecuencias asociada y de porcentajes.
   b) Observando dicha tabla, ¿cuántas personas pesan menos de 54 kg?.
   c) Observando dicha tabla, ¿qué porcentajes de personas pesan menos de 64 kg?. ¿Y más de 44 kg?.

3. Gráficos estadísticos

Los datos estadísticos se suelen expresar de forma gráfica ya que, de este modo, la información se muestra de una manera más intuitiva. En función del tipo de variable usaremos un tipo de gráfico u otro.

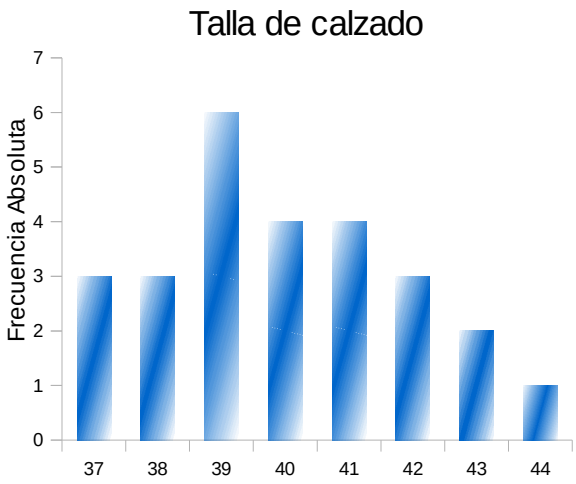
Los gráficos que más se suelen utilizar son los siguientes:

3.1 Diagrama de barras

Se utiliza para representar **variables cualitativas o cuantitativas discretas**. Sobre el eje horizontal se indican los valores de la variable y, en esos puntos, se levantan barras verticales de altura igual a las frecuencias que queramos representar.

Ejemplo:
Si recordamos los datos de la talla de calzado (realizado en un ejercicio anterior) cuya tabla de frecuencias absolutas se indican más abajo podemos obtener fácilmente el diagrama de barras.

Valores (x_i)	*Frecuencias Absolutas* (n_i)
37	3
38	3
39	6
40	4
41	4
42	3
43	2
44	1

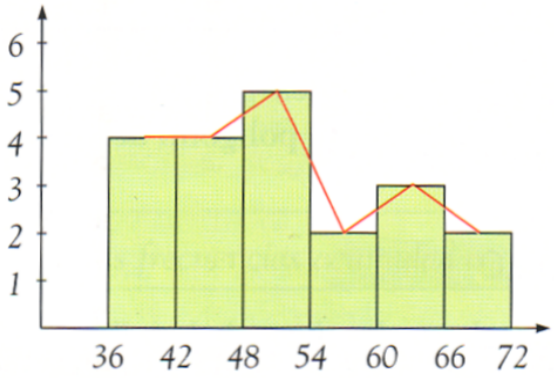


3.2 Histograma

Se utiliza para representar **variables cuantitativas continuas** cuando los datos se agrupan en intervalos. Sobre el eje horizontal se indican los extremos de los intervalos y se levantan rectángulos de base la amplitud del intervalo y altura la frecuencia.

Ejemplo:
Para los datos de la siguiente tabla el histograma queda así:

Intervalo	n_i
[36, 42)	4
[42, 48)	4
[48, 54)	5
[54, 60)	2
[60, 66)	3
[66, 72)	2



3.3 Polígono de frecuencia

Se determina uniendo los extremos superiores de las barras de un diagrama de barras o los puntos medios de las partes superiores de los rectángulos de un histograma.

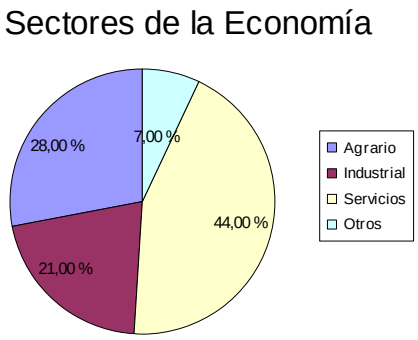
Ejemplo:
Ver ejemplo anterior.

3.4 Gráfico de sectores

El gráfico de sectores se usa para cualquier tipo de variable. Los datos se representan en un círculo, en forma de sectores circulares de amplitudes proporcionales a la frecuencias respectivas de cada valor.

Ejemplo:
Para los datos de la siguiente tabla utilizaremos un gráfico de sectores.

Sector	*Porcentaje de trabajadores*
Agrario	28%
Industrial	21%
Servicios	44%
Otros	7%



Ejercicio:

En una clase de 28 alumnos, 14 de ellos han suspendido el examen de Matemáticas, 3 han sacado Suficiente, un 25% ha obtenido Bien o Notable y el resto ha obtenido Sobresaliente.
a) Construye la tabla de frecuencias asociada a los datos.
b) Represéntalos mediante un diagrama o gráfico de sectores.

4. Medidas estadísticas

Las medidas estadísticas son valores que se calculan a partir de los datos y sus frecuencias. Sirven para resumir la información.

4.1 . Medidas de centralización

Las medidas de centralización más utilizadas son la media aritmética, la mediana y la moda.

La media aritmética, $\bar{x}$, es el cociente entre la suma de todos los valores de la variable multiplicados por sus correspondientes frecuencias absolutas y el número total de datos. En el caso de variables continuas tomaremos como valor de x_i la marca de clase de cada intervalo.

Se corresponde con la fórmula: $\bar{x}=\dfrac{\sum_{i=1}^{n} n_i \cdot x_i}{N}$

La moda, Mo, es el valor que tiene mayor frecuencia absoluta. En el caso de una variable continua hablaremos de intervalo modal.

La mediana, Me, es el valor que ocupa la posición central de los datos, después de ordenarlos, o la media de los valores centrales, en el caso de que el número de datos sea par. Si la variable es continua, hablaremos de intervalo mediano.

Nota: En el caso de que la variable sea continua y para que la moda y la mediana tomen un valor, se suele

considerar la marca de clase del intervalo modal y mediano, respectivamente.

Ejemplos:

1) Una encuesta realizada a 1000 parejas en la que se preguntaba sobre el número de hijos que tienen presenta los siguientes datos:

Número de hijos (x_i)	n_i
0	96
1	418
2	310
3	143
4	33

Calcula la media aritmética, la moda y la mediana.

2) La siguiente tabla indica el número de habitaciones libres que hay en los hoteles de una comunidad autónoma. Calcula las medidas de centralización.

Nº habitaciones	[0,10)	[10,20)	[20,30)	[30,40)	[40,50)	[50,60]
Nº hoteles	10	12	15	8	7	4

Ejercicio:

Organiza estos datos relativos al peso de 20 personas en una tabla de frecuencias y calcula sus medidas de centralización: 42, 51, 56, 66, 75, 47, 51, 45, 63, 79, 69, 59, 50, 70, 59, 62, 54, 60, 63 y 58.

4.2 . Medidas de posición

Una medida de posición es un valor de la variable que informa del lugar que ocupa un dato dentro del conjunto ordenado de valores.
Las medidas de posición más importantes son:

a) **Los cuartiles, Q_1, Q_2 y Q_3**, que son medidas que dividen todos los datos en 4 partes iguales, es decir, en cada tramo está el 25% de los datos recogidos en el estudio.

b) **Los percentiles o centiles, P_k**, que son medidas que dividen la distribución de datos en 100 partes iguales. P_k es un valor tal que el k % de los datos es menor que él y el (100- k %) es mayor.

Para el cálculo de estas medidas la variable debe ser cuantitativa. Además, para calcularlas trabajaremos con P_i . Y si la variable es continua, tomamos como valor la marca de clase.

Nótese que los cuartiles Q_1, Q_2 y Q_3 coinciden respectivamente con los percentiles P_{25} , P_{50} y P_{75}. Además la mediana, Me = $Q_2 = P_{50}$.

Ejemplo:

Para comprar zapatillas a los miembros de una peña de bolos se les ha preguntado por el número de calzado que usan y los resultados aparecen en la siguiente tabla. Calcula los tres cuartiles y los percentiles 90 y 37.

Número de calzado	***Frecuencia***
39	14
40	15
41	23
42	27
43	28
44	12
45	5
46	4

Ejercicios:

1) Con los datos de la tabla anterior calcula los percentiles P_7, P_{22} y P_{65}.
2) Salen 23 plazas a concurso por oposición y se presentan 325 personas, según la siguiente tabla:

Notas	0	1	2	3	4	5	6	7	8	9
Nº Opositores	18	32	46	51	62	47	31	21	13	4

a) ¿Con qué nota mínima se obtiene una de las 23 plazas?.
b) ¿Qué percentil es la nota 5?.
c) P_{15} y Q_3 .
d) Calcula el porcentaje de opositores cuya nota es superior a 6.
e) Calcula la media aritmética de las notas.

4.3. Medidas de dispersión

Las medidas de dispersión se utilizan para conocer el grado de agrupamiento de los datos. Las más utilizadas son:

a) **La varianza**, también conocida como V(X) o σ^2 : es la media de los cuadrados de las desviaciones de los datos respecto de la media. Se corresponde con la fórmula: $\sigma^2 = \dfrac{\sum_{i=1}^{n} n_i \cdot (x_i - \bar{x})^2}{N}$. A efectos prácticos, la fórmula que emplearemos para el cálculo de la varianza es $\sigma^2 = \dfrac{\sum_{i=1}^{n} n_i \cdot x_i^2}{N} - \bar{x}^2$. Se lee: "la varianza es la media de los cuadrados menos el cuadrado de la media".

b) **La desviación típica**, también conocida como σ: es la raiz cuadrada positiva de la varianza.
Se corresponde con la fórmula: σ = . A efectos prácticos, la fórmula que emplearemos para el cálculo de la desviación típica es:

$$\sigma = \sqrt{\frac{n_i \cdot x_i^{\,2}}{N} - (\bar{x})^2}$$

c) **Coeficiente de variación**, también conocido como CV: es el cociente entre la desviación típica y la media. No tiene unidades y se utiliza para comparar la dispersión entre las distintas variables estadísticas.

Es decir: $$CV = \frac{\sigma}{\bar{x}}$$

Ejemplo:

Las notas de 28 alumnos de 4º A ESO en Matemáticas vienen dadas por la tabla de la izquierda y las de 4º B por las de la tabla de la derecha. Calcula las medidas de dispersión de cada grupo y compara la dispersión entre ambas.

Calificaciones 4ºA	***Nº de alumnos***
0	5
1	4
2	4
3	5
4	3
5	3
6	3
7	1

Calificaciones 4ºB	***Nº de alumnos***
0	2
1	2
3	2
4	3
6	2
7	1
8	2

Ejercicios:

1) El número de preguntas acertadas por 100 alumnos en un test de 30 preguntas se presentan agrupadas en la siguiente tabla. Halla el percentil P_{30}, la mediana y las medidas de dispersión.

Notas	n_i
[0, 5)	3
[5, 10)	10
[10, 15)	25
[15, 20)	38
[20, 25)	16
[25, 30]	8

2) Un corredor entrena, de lunes a viernes, recorriendo las siguientes distancias: 2, 5, 5, 7 y 3 km, respectivamente. Si el sábado también sale a entrenar:
   a) ¿Cuántos kilómetros debe recorrer para que la media sea la misma?.
   b) ¿Y para que la mediana no varíe?.
   c) ¿Y para que la moda no varíe?.
   d) Con el dato del apartado "a", halla el CV.

3) El número de discos vendidos durante un año (en millones) en USA, según el Wall Street Journal, para los 20 discos más vendidos fue el siguiente:

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

a) Efectúa el recuento formando intervalos de clase de amplitud 0'5, siendo 2 el extremo inferior de la primera clase.
b) Representa el histograma de frecuencias absolutas.
c) Calcula la media, la mediana, la moda y el cuartil tercero.
d) Calcula el coeficiente de variación de la variable y el porcentaje de datos que se halla en los intervalos $(\bar{x}-2\cdot\sigma\,,\bar{x}+2\cdot\sigma)$ y $(\bar{x}-3\cdot\sigma\,,\bar{x}+3\cdot\sigma)$?.