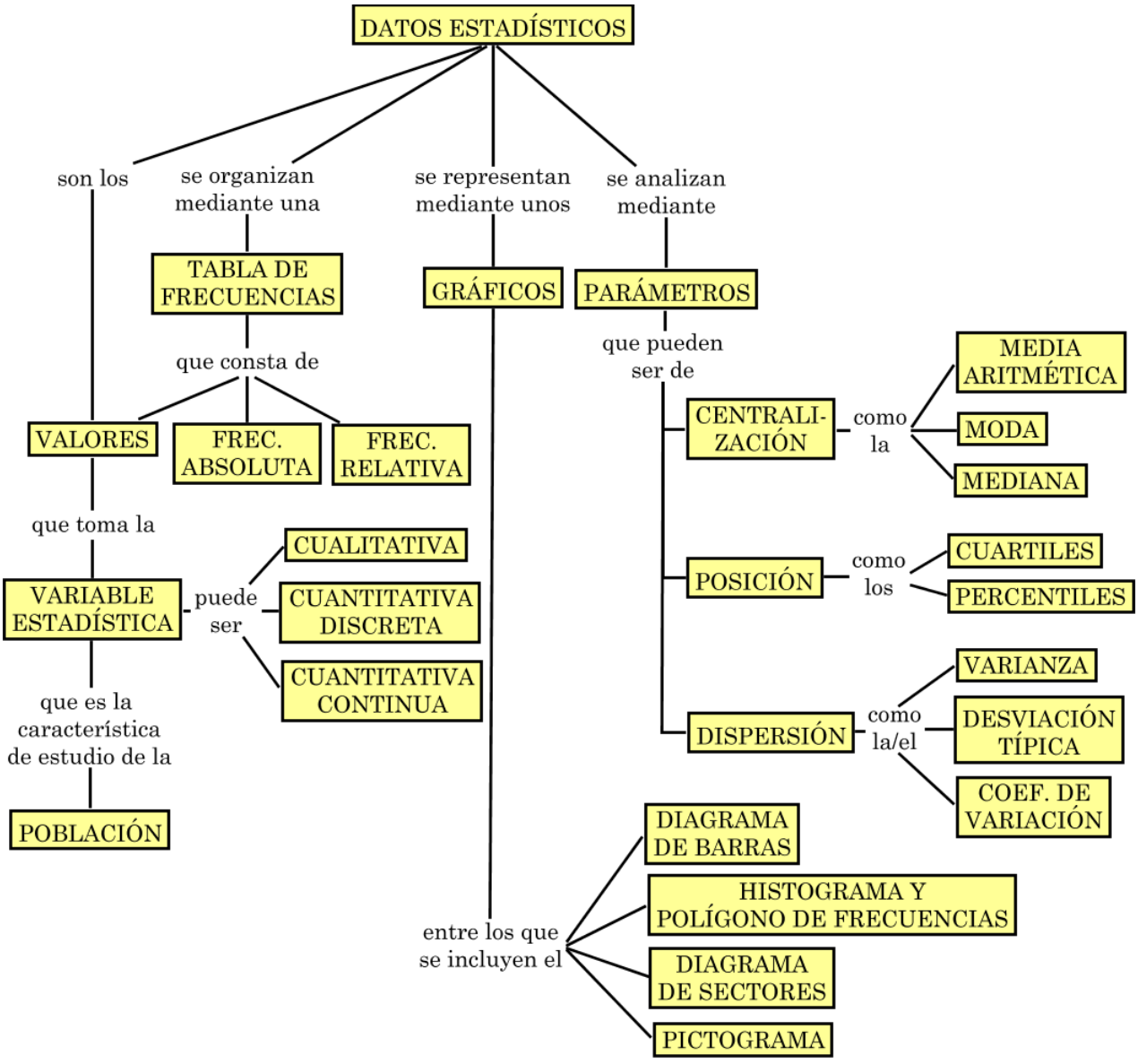
DEPARTAMENTO DE MATEMÁTICAS.

ESTADÍSTICA DE BACHILLERATO.

UNIDAD 1. ESTADÍSTICA UNIDIMENSIONAL.

IES LA BAHIA

MAPA CONCEPTUAL DE LA UNIDAD



1. Población, muestra y variable estadística.

La **Estadística Descriptiva** se ocupa de la recogida, recuento y ordenación de datos sobre una población y del cálculo de ciertos valores (**parámetros**) que nos informan de manera global sobre la característica objeto de estudio de la población. No hace uso del cálculo de probabilidades.

La **Estadística Inferencial** se ocupa de inferir* conclusiones válidas sobre la característica objeto de estudio de una población a partir de los resultados obtenidos de una muestra. Sí hace uso del cálculo de probabilidades.
**Inferir = llegar a una conclusión por medio de un razonamiento.*

Población: conjunto de elementos de los cuales se extraen los datos sobre la característica objeto de estudio. Cada uno de estos elementos se llama **individuo**. Un individuo estadístico no tiene por qué ser necesariamente una persona.

Muestra: es una parte representativa de la población que se toma para realizar el estudio.
A veces no es posible estudiar todos los individuos de una población por razones económicas, por la cantidad de tiempo o porque el estudio requiere su destrucción. De ahí la necesidad de tomar una muestra e inferir conclusiones sobre la variable objeto de estudio de la población a partir de la muestra.
Para que las conclusiones del estudio sean fiables, es necesario que la muestra elegida sea **representativa** de la población. Al proceso seguido para elegir una muestra se le llama **muestreo**.

Variable estadística: es la característica objeto de estudio de la población. Puede ser cualitativa, cuantitativa discreta o cuantitativa continua:

– **Cualitativa** si toma valores no numéricos. Por ejemplo, el lugar de nacimiento, deporte favorito, localidad de residencia…

– **Cuantitativa discreta** si solo puede tomar valores numéricos aislados. Por ejemplo, el número de hermanos, número de goles marcados, número de bombillas fabricadas…

– **Cuantitativa continua** si puede tomar todos los valores numéricos posibles dentro de un intervalo. Por ejemplo, el peso, la estatura, el tiempo, la temperatura…

Ejemplo: si se realiza un estudio sobre el tiempo de duración de las bombillas producidas en una fábrica, la población estaría formada por todas las bombillas producidas y la variable estadística sería el tiempo de duración de las bombillas, de tipo cuantitativa continua.

2. Frecuencias. Tabla de frecuencias.

Dada una variable estadística que toma valores $x_1, x_2, \ldots, x_n$, con N el número total de datos:

Se llama **frecuencia absoluta** f_i del valor x_i al número de veces que se repite dicho valor.

Se llama **frecuencia relativa** h_i del valor x_i al cociente $h_i = \frac{f_i}{N}$

Se llama **frecuencia porcentual** p_i del valor x_i al producto $p_i = 100 \cdot h_i$

Se llama **frecuencia absoluta acumulada** F_i del valor x_i a la suma de todas las frecuencias absolutas desde la primera hasta la f_i incluida. Es decir, $F_i = f_1 + f_2 + \ldots + f_i$

Se llama **frecuencia relativa acumulada** H_i del valor x_i a la suma de todas las frecuencias relativas desde la primera hasta la h_i incluida. Es decir, $H_i = h_1 + h_2 + \ldots + h_i$

Para su cálculo en una tabla de frecuencias se puede usar que $H_i = \frac{F_i}{N}$

Se llama **frecuencia porcentual acumulada** P_i del valor x_i a la suma de todas las frecuencias porcentuales desde la primera hasta la p_i incluida. Es decir, $P_i = p_1 + p_2 + ... + p_i$
Para su cálculo en una tabla de frecuencias se puede usar que $P_i = 100 \cdot H_i$

Se llama **tabla de frecuencias** de una distribución de datos a la disposición en columnas de los datos de una variable y sus correspondientes frecuencias.

Para construir una tabla de frecuencias, se indican cinco columnas: Valores que toma la variable (x_i), Frecuencias absolutas (f_i), Frecuencias relativas (h_i), Frecuencias absolutas acumuladas (F_i), Frecuencia relativas acumuladas (H_i).
A tener en cuenta:

– Los valores x_i de la primera columna deben aparecer ordenados de menor a mayor.

– Las frecuencias relativas h_i de la 3ª columna deben aparecer en su expresión decimal.

Ejemplo: en un grupo de 20 personas se cuentan las faltas de ortografía cometidas por cada una de ellas en un dictado. Los datos son: 4, 5, 7, 8, 6, 5, 4, 5, 6, 7, 5, 6, 5, 6, 6, 3, 8, 7, 5, 7

En la primera columna se escriben los números del 3 al 8, que son los valores que toma la variable estadística "**Faltas de ortografía**".

Las frecuencias relativas h_i de la 3ª columna se calculan así:

$h_1 = \frac{f_1}{N} = \frac{1}{20} = 0{,}05$ $\quad$ $h_2 = \frac{f_2}{N} = \frac{2}{20} = 0{,}1$

$h_3 = \frac{f_3}{N} = \frac{6}{20} = 0{,}3$ $\quad$ $h_4 = \frac{f_4}{N} = \frac{5}{20} = 0{,}25$

$h_5 = \frac{f_5}{N} = \frac{4}{20} = 0{,}2$ $\quad$ $h_6 = \frac{f_6}{N} = \frac{2}{20} = 0{,}1$

x_i	f_i	h_i	F_i	H_i
3	1	0,05	1	0,05
4	2	0,1	3	0,15
5	6	0,3	9	0,45
6	5	0,25	14	0,7
7	4	0,2	18	0,9
8	2	0,1	20	1

De la tabla se puede extraer información:

a) ¿Cuántas personas han cometido 7 faltas de ortografía? 4 personas

b) ¿Qué porcentaje de personas ha cometido 7 faltas de ortografía? 20%

c) ¿Cuántas personas han cometido 5 o menos de 5 faltas de ortografía? 9 personas

d) ¿Qué porcentaje de personas ha cometido 5 o menos de 5 faltas de ortografía? 45%

3. Intervalos y marcas de clase en variables cuantitativas continuas.

Todas las variables continuas o algunas discretas suelen presentar un alto número de valores diferentes. Con el objeto de simplificar la tabla, los datos se suelen agrupar en intervalos. Se pierde algo de precisión a cambio de agilizar los cálculos.

Se llama **intervalo de clase** a cada uno de los intervalos en los que pueden agruparse los datos de una variable estadística.

La **marca de clase** del intervalo es el valor medio (semisuma) de los extremos de dicho intervalo y se representa por x_i.

El tamaño o longitud de cada intervalo se puede tomar como el entero mayor más cercano al resultado de la operación $\frac{M-m}{n}$, donde:

M y m son los valores mayor y menor, respectivamente de la lista ordenada de valores que toma la variable, mientras que n es el número de intervalos fijado de antemano.

Ejemplo: se quiere realizar un estudio sobre el peso de un grupo de personas. Al recoger los datos y ordenarlos, se obtiene la siguiente lista de datos, expresados en kg:

49	50,2	50,8	51,4	52,3	52,6	52,7	53,1	53,1	53,8	54,4	54,5	54,7	55	55,1
55,6	56	56,2	56,5	57	57,9	58,6	59,7	59,8	59,8	59,9	60	61,8	63,4	66

La variable es **Peso de las personas**, de tipo cuantitativa continua. Toma hasta 28 valores distintos, ya que solo dos de ellos se repiten. Si se intenta elaborar una tabla indicando los valores uno a uno, sería de considerable tamaño.

Conviene pues, agrupar los datos en intervalos. Si se decide agrupar los datos en **6** intervalos de igual amplitud, ésta se determina de la manera siguiente:

(1º) Se calcula la diferencia entre los valores extremos: 66 – 49 = 17. (Diferencia = **17**)

(2º) Se calcula la amplitud de cada intervalo: **17** : **6** = 2,83…, redondeado a **3**. (Amplitud = **3**)

(3º) Se toma como origen del primer intervalo el menor de los valores que toma la variable si es entero, o el valor entero menor o igual más cercano al mismo: en este caso sería **49**.
(Origen del primer intervalo = **49**)

(4º) Se escriben los intervalos, de menor a mayor. Cada intervalo es cerrado por la izquierda y abierto por la derecha, excepto el último que es cerrado por ambos extremos.

[49, 52) [52, 55) [55, 58) [58, 61) [61, 64) [64, 67]

(5º) Se determina la marca de clase de cada intervalo:

Marca de clase de [49, 52): $\frac{49+52}{2} = 50{,}5$ y así sucesivamente.

Así quedaría la tabla de frecuencias:

Peso	Marca de clase x_i	Frecuencia absoluta f_i	Frecuencia relativa h_i	Frecuencia absoluta acumulada F_i	Frecuencia relativa acumulada H_i
[49, 52)	50,5	4	0,13	4	0,13
[52, 55)	53,5	9	0,3	13	0,43
[55, 58)	56,5	8	0,27	21	0,70
[58, 61)	59,5	6	0,20	27	0,90
[61, 64)	62,5	2	0,07	29	0,97
[64, 67]	65,5	1	0,03	30	1

De la tabla se puede extraer información:

a) ¿Cuántas personas pesan entre 55 y 58 kg? 8 personas

b) ¿Qué porcentaje de personas pesan entre 55 y 58 kg? 27%

c) ¿Cuántas personas pesan menos de 58 kg? 21 personas

d) ¿Qué porcentaje de personas pesan menos de 58 kg? 70%

4. Gráficos estadísticos.

A) Diagrama de barras

Un gráfico muy utilizado para variables cualitativas o cuantitativas discretas con pocos valores es el diagrama de barras.

Para representar un diagrama de barras se aplican los siguientes pasos:

Paso 1. Se dibujan dos ejes perpendiculares, uno horizontal y otro vertical.

Paso 2. En el eje horizontal se escriben los valores x_i que toma la variable estadística y el nombre de la variable estadística.

Paso 3. En el eje vertical se escriben las frecuencias absolutas f_i correspondientes a cada x_i.

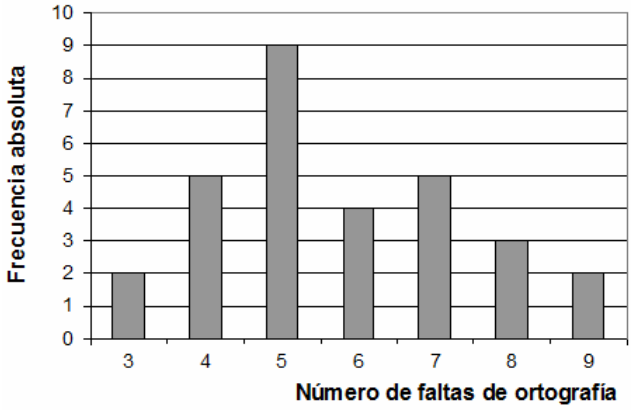
Paso 4. En la posición de cada valor x_i se representa una barra de altura proporcional a la correspondiente f_i. Estas barras deben estar separadas unas de otras.

Ejemplo: en un grupo de 30 estudiantes se cuentan las faltas de ortografía cometidas por cada uno de ellos en un dictado.

Los datos son:

4, 5, 7, 8, 9, 5, 4, 5, 6, 7, 5, 6, 5, 5, 4,
3, 8, 7, 5, 8, 9, 4, 3, 5, 7, 6, 7, 5, 4, 6

x_i	3	4	5	6	7	8	9
f_i	2	5	9	4	5	3	2



B) Diagrama de sectores

El diagrama de sectores es otro tipo de gráfico que sirve para representar variables de cualquier tipo. Para representar un diagrama de sectores se aplican los siguientes pasos:

Paso 1. Se dibuja un círculo.

Paso 2. Se multiplican las frecuencias relativas h_i en su expresión fraccionaria por 360º, obteniendo así las amplitudes de cada sector en grados sexagesimales.

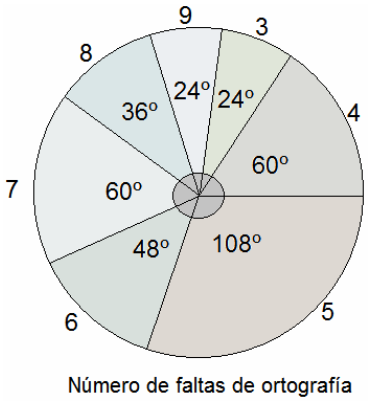
Paso 3. Se divide el círculo en sectores cada uno con las amplitudes calculadas en el paso anterior y se colorea de forma diferente cada sector.

Paso 4. Se representan rectángulos, cada uno con el color correspondiente a cada sector.

Ejemplo: en un grupo de 30 estudiantes se cuentan las faltas de ortografía cometidas por cada uno de ellos en un dictado. Los datos son:

4, 5, 7, 8, 9, 5, 4, 5, 6, 7, 5, 6, 5, 5, 4, 3, 8, 7, 5, 8, 9, 4, 3, 5, 7, 6, 7, 5, 4, 6

x_i	f_i	h_i	$h_i \cdot 360^o$
3	2	2/30	24º
4	5	5/30	60º
5	9	9/30	108º
6	4	4/30	48º
7	5	5/30	60º
8	3	3/30	36º
9	2	2/30	24º



C) Pictograma

Otro tipo de gráfico muy utilizado es el pictograma. Para representar un pictograma se procede del siguiente modo:

Paso 1. Se dibuja una línea (vertical u horizontal) y se indican los valores de la variable.

Paso 2. Se inventa un dibujo representativo y se le asigna un valor adecuado.

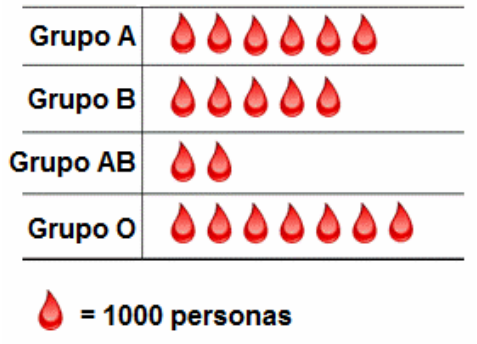
Paso 3. Se coloca, en cada valor de la variable, tantos símbolos como indique su frecuencia absoluta.

Ejemplo: en una población formada por 20 000 personas se analiza el grupo sanguíneo de cada una.

Los datos aproximados son:

6 000 del grupo A 5 000 del grupo B

2 000 del grupo AB 7 000 del grupo O



D) Histograma y polígono de frecuencias

El **histograma** es un tipo de gráfico que suele utilizarse para variables continuas o para variables discretas con datos agrupados por intervalos.
Para representar un histograma se aplican los siguientes pasos:

Paso 1. Se dibujan dos ejes perpendiculares, uno horizontal y otro vertical.

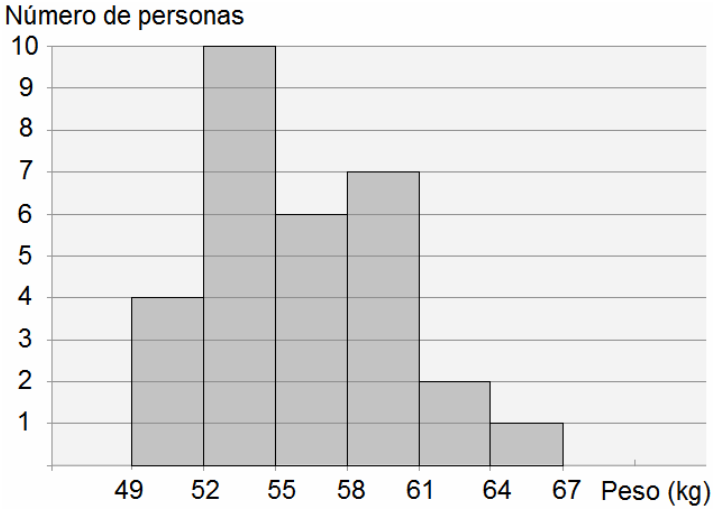
Paso 2. En el eje horizontal se representan los intervalos uno a continuación de otro.

Paso 3. En el eje vertical se escriben las frecuencias absolutas f_i correspondientes a cada intervalo, utilizando una escala adecuada.

Paso 4. En la posición de cada intervalo se representa un rectángulo de altura proporcional a su correspondiente f_i . Estos rectángulos deben estar contiguos, al igual que los intervalos.

Ejemplo: se quiere realizar un estudio sobre los pesos de un grupo de 30 personas. La lista ordenada de datos agrupados por intervalos y expresada en kg es la siguiente:

Pesos (kg)	f_i
[49, 52)	4
[52, 55)	10
[55, 58)	6
[58, 61)	7
[61, 64)	2
[64, 67]	1



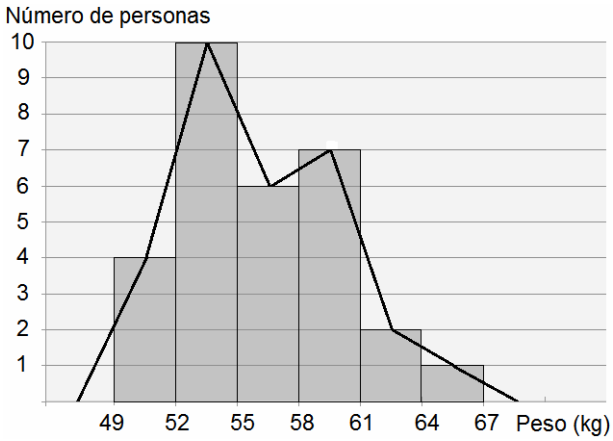
Polígono de frecuencias

El polígono de frecuencias es un gráfico que se obtiene a partir de un histograma. Para representar un polígono de frecuencias basta con aplicar los siguientes pasos:

Paso 1. En el histograma que se haya construido previamente, se unen mediante una línea poligonal los puntos medios de los lados superiores de cada rectángulo.

Paso 2. Se prolonga dicha línea poligonal, suponiendo que existen dos intervalos de frecuencia cero en cada extremo.

Ejemplo: el polígono de frecuencias asociado al estudio anterior sería el siguiente:



5. Parámetros de centralización.

Dada una variable estadística, que toma valores x_1, x_2, ..., x_n, siendo N el número total de datos.

Los **parámetros de centralización** son números reales que en general, tienden a situarse hacia el centro del conjunto de valores ordenados. Los parámetros de centralización son: la **media aritmética**, la **moda** y la **mediana**.

A) Media aritmética

Dada una variable estadística que toma valores $x_1, x_2, \ldots, x_n$, con N el número total de datos, la **media aritmética** es la suma de todos los valores de la variable dividida entre el número total de datos.

$$\text{Media} = \overline{x} = \frac{x_1 \cdot f_1 + x_2 \cdot f_2 + \ldots + x_n \cdot f_n}{N} = \frac{\sum x_i f_i}{N}$$

Ejemplo: en un grupo de 20 alumnos se cuentan las faltas de ortografía cometidas por cada uno de ellos en un dictado. Los datos son: 4, 5, 7, 8, 6, 5, 4, 5, 6, 7, 5, 6, 5, 6, 6, 3, 8, 7, 5, 7

Para calcular la media aritmética se construye una tabla como ésta:

x_i	f_i	$x_i \cdot f_i$
3	1	3
4	2	8
5	6	30
6	5	30
7	4	28
8	2	16
	20	115

$$\text{Media} = \overline{x} = \frac{3 \cdot 1 + 4 \cdot 2 + 5 \cdot 6 + 6 \cdot 5 + 7 \cdot 4 + 8 \cdot 2}{20} = \frac{115}{20} = 5{,}75$$

Interpretación: es como si cada alumno "hubiera tenido 5,75 faltas".
Como 5,75 está más cerca de 6 que de 5, se puede decir que en este grupo, la media está en algo menos de 6 faltas por alumno.

Propiedades

P1. Que una distribución con N datos tenga media aritmética $\overline{x}$, es "como si todo el conjunto de datos estuviera formada por N valores todos ellos iguales a $\overline{x}$".

P2. Si todo el conjunto de datos posee valores atípicos, excepcionalmente raros, éstos producen una distorsión sobre el valor de la media, alterando su representatividad. Es decir, la media no es resistente a las alteraciones producidas por los valores atípicos.

Por ejemplo, si a la lista de datos: 4,5 5 5,5 6,25 7,25 7,5 cuya media es $\overline{x} = 6$, se le añade un valor muy alejado, como es 300, entonces la nueva lista:
4,5 5 5,5 6,25 7,25 7,5 300 tendría media $\overline{x} = 48$, para nada representativa de todo el conjunto.

B) Moda

La moda es el valor de la variable que tiene mayor frecuencia absoluta. Se representa por M_o.

La moda no tiene por qué ser única: puede haber dos, tres, … modas.

Si todas las frecuencias son iguales, el conjunto de datos no tiene moda.

Si los valores están agrupados por intervalos, el intervalo modal es el que tiene mayor frecuencia absoluta y se puede tomar como moda su marca de clase.

Ejemplo 1: en la distribución 2, 5, 5, 7, 7, 9, 9, 9, 10, 10, 11, 12, 12 la moda es 9.

Ejemplo 2: en la distribución 2, 4, 5, 6, 7, 7, 7, 9, 11, 11, 11, 15 las modas son 7 y 11.

Ejemplo 3: en la distribución 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15 no existe la moda.

Propiedades

P1. La moda representa el valor dominante de todo el conjunto.

P2. La moda no siempre se sitúa en la zona central. Puede estar cerca de los valores extremos.

P3. La moda es menos representativa que la media aritmética, pero hay ocasiones en que es más útil. Por ejemplo, cuando los datos corresponden a una variable cualitativa.

C) Mediana

Ordenada de menor a mayor la lista de datos, la mediana es un número real que deja al 50% de los datos por arriba y al otro 50% por debajo. Se representa por M_e.

Cálculo de la mediana

Caso 1. Si los datos no están agrupados en intervalos, la mediana es el valor x_i de la variable cuya F_i supera por primera vez a la mitad de N.
Si la mitad de N coincidiera con la F_i de algún valor, entonces se toma como mediana la semisuma entre ese valor y el siguiente.

Ejemplo 1: en la lista de datos 2, 3, 5, 6, 9, 11, 12 la mediana es 6.

Ejemplo 2: en la lista de datos 2, 3, 5, 6, 9, 11, 12, 13 la mediana es $\frac{6+9}{2} = 7{,}5$

Ejemplo 3: hallar la media aritmética, la moda y la mediana de la lista de datos:

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

Se elabora una tabla como ésta:

x_i	f_i	F_i	$x_i \cdot f_i$
1	2	2	2
2	2	4	4
3	5	9	15
4	5	14	20
5	8	22	40
6	9	31	54
7	3	34	21
8	3	37	24
9	3	40	27
	40		207

$\text{Media} = \bar{x} = \frac{207}{40} = 5{,}175$

$M_o = 6$ porque es el valor con mayor $f_i = 9$

$M_e = 5$ porque es el valor cuya F_i supera por primera vez al número $\frac{N}{2} = \frac{40}{2} = 20$

Caso 2. Si los datos están agrupados en intervalos, entonces $M_e = a + \frac{\frac{N}{2} - F_{I-1}}{f_I} \cdot c$ donde

I = intervalo cuya F_i supera por primera vez al número $\frac{N}{2}$ (I contiene a la mediana)

a = extremo inferior de I c = amplitud de I f_I = frecuencia absoluta de I

F_{I-1} = frecuencia absoluta acumulada del intervalo inmediatamente anterior a I

Ejemplo: hallar la media aritmética, la moda y la mediana de la siguiente distribución de frecuencias:

Intervalos	[38, 44)	[44, 50)	[50, 56)	[56, 62)	[62, 68)	[68, 74)	[74, 80]
f_i	7	8	15	25	18	9	18

Se elabora una tabla como ésta:

Intervalos	x_i	f_i	$x_i \cdot f_i$	F_i
[38, 44)	41	7	287	7
[44, 50)	47	8	376	15
[50, 56)	53	15	795	30
[56, 62)	59	25	1 475	55
[62, 68)	65	18	1 170	73
[68, 74)	71	9	639	82
[74, 80]	77	18	1 386	100
		100	6 128	

$\text{Media} = \bar{x} = \frac{6128}{100} = 61{,}28$

$M_o = 59$ porque es el valor con mayor $f_i = 25$

Intervalo modal = [56, 62)

Mediana

I = [56, 62) porque su $F_I = 55$ supera por primera vez al número $\frac{N}{2} = \frac{100}{2} = 50$

$a = 56 \quad c = 6 \quad f_I = 25 \quad F_{I-1} = 30 \quad M_e = a + \frac{\frac{N}{2} - F_{I-1}}{f_I} \cdot c = 56 + \frac{50 - 30}{25} \cdot 6 = 56 + 4{,}8 = 60{,}8$

6. Parámetros de dispersión.

Antes de definir cada uno de los parámetros de dispersión, obsérvese el siguiente ejemplo:

Se realiza un examen a dos grupos de ocho alumnos, obteniéndose los siguientes resultados:

Grupo A: 4,6 4,8 4,9 5 5 5,1 5,2 5,4 Grupo B: 1 1,8 3 5 5 7 8,2 9

En ambos grupos, tanto la media, la moda como la mediana son iguales a 5. Sin embargo, la lectura en detalle de los datos refleja dos grupos bien distintos: mientras que las puntuaciones de A están concentradas en torno a la media, las puntuaciones de B se alejan de la media.

Por ello, un análisis estadístico queda incompleto si sólo se estudian las parámetros de centralización. **Es imprescindible conocer si los datos numéricos están agrupados o no alrededor de los valores centrales.** Esta característica recibe el nombre de **dispersión**: en el ejemplo anterior, las puntuaciones del grupo B están más dispersas que las del grupo A.

A los valores que miden esta desviación respecto a la media se les llama **parámetros de dispersión**. Éstos son **rango**, **desviación media**, **varianza**, **desviación típica** y **coeficiente de variación**.

Dada una variable estadística que toma valores x_1, x_2, ..., x_n siendo N el número total de datos.

A) Rango

El rango es la diferencia entre el mayor y el menor valor de los datos. Cuanto menor es el rango del conjunto de datos, más representativos son los parámetros de centralización.

Ejemplo: el rango del grupo A anterior es $5,4 - 4,6 = 0,8$ y el rango del grupo B es $9 - 1 = 8$

B) Desviación media

La desviación media es el promedio de las diferencias entre cada valor de la variable y la media aritmética. Se representa por DM.

$$DM = \frac{\sum |x_i - \overline{x}| \cdot f_i}{N}$$

C) Varianza

La varianza es el promedio de los cuadrados de las diferencias entre cada valor de la variable y la media aritmética. Se representa por σ^2.

$$\sigma^2 = \frac{\sum (x_i - \overline{x})^2 \cdot f_i}{N} = \frac{\sum x_i^2 \cdot f_i}{N} - \overline{x}^2$$

D) Desviación típica

La desviación típica es la raíz cuadrada positiva de la varianza. Se representa por σ.

$$\sigma = \sqrt{\frac{\sum (x_i - \overline{x})^2 \cdot f_i}{N}} = \sqrt{\frac{\sum x_i^2 \cdot f_i}{N} - \overline{x}^2}$$

Propiedades de la desviación típica

P1. La desviación típica es una medida de la dispersión: indica cuánto de alejados están los datos respecto de la media.

P2. La desviación típica permite comparar las dispersiones de dos poblaciones con medias muy parecidas. En el ejemplo anterior, en el grupo A: 4,6 4,8 4,9 5 5 5,1 5,2 5,4 la desviación típica es $\sigma = 0{,}23$, mientras que en el grupo B: 1 1,8 3 5 5 7 8,2 9 la desviación típica es $\sigma = 2{,}75$
Esto indica que las puntuaciones del grupo B tienen mayor dispersión que las del grupo A.

P3. La varianza no se expresa en las mismas unidades que los datos, la desviación típica sí.

Ejemplo 1: hallar el rango y la desviación típica de la distribución de datos:

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

Se elabora una tabla como ésta:

x_i	f_i	$x_i \cdot f_i$	$x_i^2 \cdot fi$
1	2	2	2
2	2	4	8
3	5	15	45
4	5	20	80
5	8	40	200
6	9	54	324
7	3	21	147
8	3	24	192
9	3	27	243
	40	207	1 241

Rango = 9 – 1 = 8

$$\text{Media} = \bar{x} = \frac{207}{40} = 5{,}175$$

Varianza

$$\sigma^2 = \frac{\sum x_i^2 \cdot f_i}{N} - \bar{x}^2 = \frac{1241}{40} - (5{,}175)^2 = 4{,}244375$$

Desviación típica $\sigma = \sqrt{4{,}244375} = 2{,}06$

Ejemplo 2: hallar el rango y la desviación típica de la distribución de frecuencias:

Intervalos	[38, 44)	[44, 50)	[50, 56)	[56, 62)	[62, 68)	[68, 74)	[74, 80]
f_i	7	8	15	25	18	9	18

Se elabora una tabla como ésta:

Intervalos	x_i	f_i	$x_i \cdot f_i$	$x_i^2 \cdot fi$
[38, 44)	41	7	287	11 767
[44, 50)	47	8	376	17 672
[50, 56)	53	15	795	42 135
[56, 62)	59	25	1 475	87 025
[62, 68)	65	18	1 170	76 050
[68, 74)	71	9	639	45 369
[74, 80]	77	18	1 386	106 722
		100	6 128	386 740

Rango = 80 – 38 = 42

$$\text{Media} = \bar{x} = \frac{6128}{100} = 61{,}28$$

Varianza

$$\sigma^2 = \frac{\sum x_i^2 \cdot f_i}{N} - \bar{x}^2 = \frac{386740}{100} - (61{,}28)^2 = 112{,}1616$$

Desviación típica $\sigma = \sqrt{112{,}1616} = 10{,}59$

E) Coeficiente de variación

Se dice que un conjunto de datos es más homogéneo que otro si los valores del primer conjunto guardan más parecido entre ellos que los del segundo.
La expresión "más homogéneo que" es equivalente a la expresión "menos disperso que".

Para comparar la dispersión de dos conjuntos de datos con medias aritméticas parecidas basta con recurrir a la desviación típica. ¿Y si las medias son valores muy distintos? Para comparar la dispersión de dos conjuntos de datos con medias muy distintas es necesario recurrir a un nuevo parámetro de dispersión llamado coeficiente de variación.

El **coeficiente de variación** se define como el cociente entre la desviación típica y la media aritmética. Se representa por CV.

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

Así pues, el CV mide la dispersión (u homogeneidad) de un conjunto de datos. Dados dos conjuntos de datos (tengan la media que tengan), es menos disperso (más homogéneo) aquel cuyo CV es menor.

Por otra parte, si un conjunto de valores se analiza de forma aislada, sin comparar su CV con el de otro conjunto, se suele considerar que un CV mayor o igual que 0,30 indica un grado elevado de dispersión.

Ejemplo 1: se analizan los pesos de dos poblaciones muy diferentes: toros y perros.

El peso medio de los toros de una ganadería es 510 kg con una desviación típica de 25 kg.

El peso medio de los perros de una perrera es 19 kg con una desviación típica de 10 kg.

¿Cuál de las dos poblaciones es más homogénea?

$$\text{Toros: } CV = \frac{25}{510} = 0{,}049 \qquad \text{Perros: } CV = \frac{10}{19} = 0{,}526$$

Los pesos de los toros son más homogéneos (menos dispersos) que los de los perros.

En la población de toros, la mayoría de ellos tiene un peso muy cercano a la media, es decir, sus pesos guardan un gran parecido. Sin embargo, en la población de perros habrá desde perros con un peso muy bajo (caniche, bichón maltés, chihuahua, etc) hasta perros con un alto peso (san bernardo, mastín, gran danés…), pasando por todos los pesos intermedios.

Ejemplo 2: se analizan las calificaciones de un examen de dos grupos de diez estudiantes.

Grupo A	1	1	2	3	5	5	6	7	8	9
Grupo B	2	3	6	6	6	7	7	9	10	10

Grupo A: $\overline{x} = 4{,}7$ $\sigma = 2{,}72$ $CV = 0{,}579$ Grupo B: $\overline{x} = 6{,}6$ $\sigma = 2{,}54$ $CV = 0{,}385$

Las calificaciones del grupo B son menos dispersas (más homogéneas) que las del grupo A.

7. Parámetros de posición.

Los parámetros de posición son los **percentiles**. Dentro de los percentiles se encuentran los **cuartiles** y dentro de éstos, la **mediana**.

Ordenada de menor a mayor la distribución de los datos, los **percentiles** son todos y cada uno de los 99 números que dividen a la serie ordenada de datos en 100 partes iguales, cada una de ellas conteniendo el mismo número de datos. Se representan por P_1, P_2, ... P_{99}

El percentil P_k es un número real que deja el k% de los datos de la distribución por debajo suya y el otro (100 – k)% restante por encima.

En particular, los percentiles P_{25} y P_{75} se llaman cuartil primero Q_1 y cuartil tercero Q_3, respectivamente. El percentil P_{50} se llama cuartil segundo Q_2 o mediana M_e.

Ejemplo 1: en la distribución 2, 3, 5, 6, 9, 11, 12 los cuartiles son $Q_1 = 3$, $Q_2 = 6$, $Q_3 = 11$

Ejemplo 2: en la distribución 2, 3, 5, 6, 9, 11, 12, 13 los cuartiles son $Q_1 = 4$, $Q_2 = 7{,}5$, $Q_3 = 11{,}5$

Cálculo del percentil P_k

Caso 1. Si los datos no están agrupados en intervalos, el percentil P_k es el valor x_i de la variable cuya F_i supera por primera vez al número $\frac{kN}{100}$.

Si este número coincidiera con la F_i de algún valor, entonces el percentil P_k es la semisuma entre ese valor y el siguiente.

Ejemplo: hallar los cuartiles Q_1 y Q_3 y los percentiles P_{35} y P_{90} de la distribución de datos:

x_i	1	2	3	4	5	6	7	8	9
f_i	2	2	5	5	8	9	3	3	3

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

Se elabora una tabla como ésta:

x_i	f_i	F_i
1	2	2
2	2	4
3	5	9
4	5	14
5	8	22
6	9	31
7	3	34
8	3	37
9	3	40
	40	

$Q_1 = P_{25} = 4$ porque es el valor cuya $F_i = 14$ supera por primera vez al número $\frac{k \cdot N}{100} = \frac{25 \cdot 40}{100} = 10$

$Q_3 = P_{75} = 6$ porque es el valor cuya $F_i = 31$ supera por primera vez al número $\frac{k \cdot N}{100} = \frac{75 \cdot 40}{100} = 30$

$P_{35} = 4{,}5$ porque el número $\frac{k \cdot N}{100} = \frac{35 \cdot 40}{100} = 14$ coincide con la $F_i = 14$ de 4. Así que se toma la semisuma entre 4 y el valor siguiente 5.

$P_{90} = 8$ porque es el valor cuya $F_i = 37$ supera por primera vez al número $\frac{k \cdot N}{100} = \frac{90 \cdot 40}{100} = 36$

Caso 2. Si los datos están agrupados en intervalos, entonces $P_k = a + \dfrac{\frac{k \cdot N}{100} - F_{I-1}}{f_I} \cdot c$ donde

I = intervalo cuya F_i supera por primera vez al número $\dfrac{k \cdot N}{100}$ (I contiene a P_k)

a = extremo inferior de I c = amplitud de I f_I = frecuencia absoluta de I

F_{I-1} = frecuencia absoluta acumulada del intervalo inmediatamente anterior a I

Ejemplo: hallar los cuartiles Q_1 y Q_3 y los percentiles P_{40} y P_{90} de la siguiente distribución:

Intervalos	[38, 44)	[44, 50)	[50, 56)	[56, 62)	[62, 68)	[68, 74)	[74, 80]
f_i	7	8	15	25	18	9	18

Se elabora una tabla como ésta:

Intervalos	x_i	f_i	F_i
[38, 44)	41	7	7
[44, 50)	47	8	15
[50, 56)	53	15	30
[56, 62)	59	25	55
[62, 68)	65	18	73
[68, 74)	71	9	82
[74, 80]	77	18	100
		100	

Cuartil $Q_1 = P_{25}$

I = [50, 56) porque su $F_I = 30$ supera por primera vez al número $\dfrac{k \cdot N}{100} = \dfrac{25 \cdot 100}{100} = 25$

a = 50 c = 6 $f_I = 15$ $F_{I-1} = 15$

$$Q_1 = a + \frac{\frac{k \cdot N}{100} - F_{I-1}}{f_I} \cdot c = 50 + \frac{25 - 15}{15} \cdot 6 = 50 + 4 = 54$$

Cuartil $Q_3 = P_{75}$

I = [68, 74) porque su $F_I = 82$ supera por primera vez al número $\dfrac{k \cdot N}{100} = \dfrac{75 \cdot 100}{100} = 75$

a = 68 c = 6 $f_I = 9$ $F_{I-1} = 73$

$$Q_3 = a + \frac{\frac{k \cdot N}{100} - F_{I-1}}{f_I} \cdot c = 68 + \frac{75 - 73}{9} \cdot 6 = 68 + 1{,}33 = 69{,}33$$

Percentil P_{40}

I = [56, 62) porque su $F_I = 55$ supera por primera vez al número $\dfrac{k \cdot N}{100} = \dfrac{40 \cdot 100}{100} = 40$

a = 56 c = 6 $f_I = 25$ $F_{I-1} = 30$

$$P_{40} = a + \frac{\frac{k \cdot N}{100} - F_{I-1}}{f_I} \cdot c = 56 + \frac{40 - 30}{25} \cdot 6 = 56 + 2{,}4 = 58{,}4$$

Percentil P_{90}

I = [74, 80] porque su $F_I = 100$ supera por primera vez al número $\dfrac{k \cdot N}{100} = \dfrac{90 \cdot 100}{100} = 90$

a = 74 c = 6 $f_I = 18$ $F_{I-1} = 82$

$$P_{90} = a + \frac{\frac{k \cdot N}{100} - F_{I-1}}{f_I} \cdot c = 74 + \frac{90 - 82}{18} \cdot 6 = 74 + 2{,}66 = 76{,}66$$

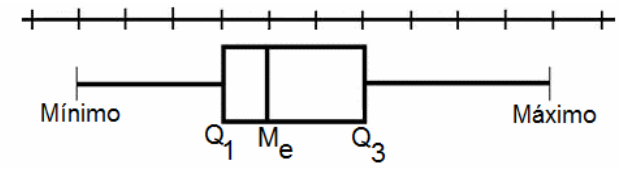
Diagrama de caja y bigotes

El diagrama de caja y bigotes es un gráfico utilizado para representar el comportamiento de una variable cuantitativa, a través del rango y los cuartiles.

Se compone de:

– Un rectángulo o **caja** delimitado por el primer y tercer cuartiles, Q_1 y Q_3. Dentro de la caja, una línea indica dónde se encuentra la mediana o segundo cuartil, M_e.

– Dos brazos o **bigotes**: uno a la izquierda, que empieza en el valor mínimo del rango y acaba al inicio de la caja (en Q_1); y otro a la derecha, que empieza al final de la caja (en Q_3) y acaba en el valor máximo del rango.



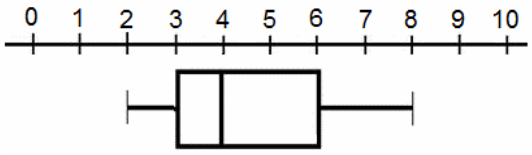
Ejemplo: a un grupo de veinte personas se les pregunta por el número de libros que han leído durante el último año, obteniéndose los siguientes resultados:

2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 6, 6, 6, 6, 6, 7, 7, 8, 8

El rango es 8 – 2 = 6. El valor mínimo es 2 y el valor máximo es 8.

Se calculan los cuartiles, obteniéndose: $Q_1 = 3$ $M_e = 4$ $Q_3 = 6$

Éste es el diagrama de cajas y bigotes:



Interpretación de un diagrama de caja

El diagrama de caja y bigotes solo aporta información numérica exacta de los cuartiles y del rango. No ofrece con exactitud los valores de la media aritmética y de la desviación típica.

Sin embargo, un análisis adecuado del **tamaño de la caja** (y de los **bigotes**) y de su **orientación** hacia la izquierda o derecha, ofrece pistas sobre los posibles valores de la media y de la desviación típica.

El **tamaño de la caja** (y de los **bigotes**) no tiene nada que ver con el **número de datos**, sino con el **porcentaje de datos** que contiene. Que una caja (o bigote) sea más larga que otra no significa que contenga más datos. Cada sección de la caja y cada bigote contiene exactamente un 25% de los datos, sea cual sea el tamaño de la muestra. De lo cual se deduce lo siguiente:

a) Si una **caja** (o **bigote**) es **más larga** que otra, significa que el rango de los datos es mayor, que los datos están más dispersos. Cajas (o bigotes) más largos indican **mayor dispersión** de los datos, menor cercanía o parecido entre los mismos, por lo tanto, mayor desviación típica.

b) Si un diagrama tiene la caja y el segundo bigote **más orientados hacia la derecha** que otro (especialmente si es a partir del comienzo de la caja, primer cuartil), entonces debe tener **mayor media aritmética**, ya que sitúa al 75% de los datos en la zona de mayores valores del rango.