INTRODUCCIÓN A LA ESTADÍSTICA

La estadística es una rama de las matemáticas que estudia los conjuntos de datos para sacar conclusiones a partir de ellos (y hacer inferencias basadas en el cálculo de probabilidades, pero esto se verá en bachillerato).

Al realizar un estudio estadístico, lo primero que hay que hacer es saber qué tipo de variable estadística estamos estudiando.

a) Variables cualitativas: aquellas que NO se pueden expresar con números. Ej: color de pelo, equipo favorito, etc.
b) Variables cuantitativas: aquellas que se pueden expresar numéricamente. Hay que distinguir entre dos tipos.
    - Discretas: solo pueden tomar valores de los números naturales (1, 2, 3, etc.). Ej: nº de hijos, piso en el que vives, etc.
    - Continuas: pueden tomar cualquier valor. Ej: altura, peso, precio, etc.

Lo siguiente que podemos hacer a partir de un conjunto de datos es ordenarlos y clasificarlos utilizando tablas de frecuencias. A partir de estas tablas podremos hacer representaciones gráficas que nos permitan entender los datos de una forma más sencilla y visual. Veamos un ejemplo de todo esto.

EJEMPLO 1: Hemos entrevistado a 20 personas y les hemos preguntado su color favorito. Hemos obtenido las siguientes respuestas:

Negro, azul, amarillo, rojo, azul, azul, rojo, negro, amarillo, rojo, rojo, amarillo, amarillo, azul, rojo, negro, azul, rojo, negro, amarillo.

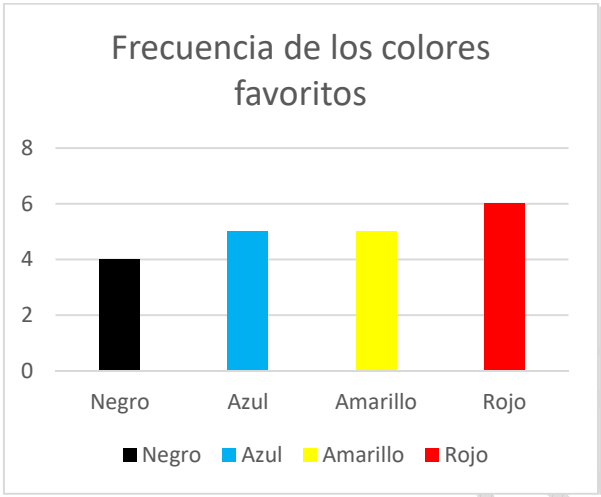
Realizar una tabla de frecuencias para estos datos.

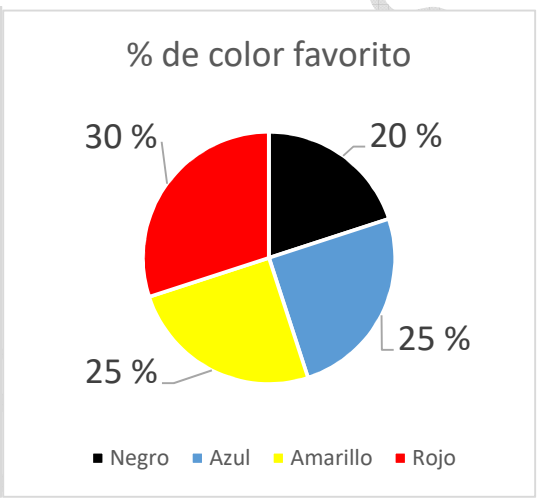
Antes de ponernos manos a la obra vamos a especificar qué columnas queremos incluir en la tabla de frecuencias:

a) Valores de la variable (X_i): Son los diferentes valores que toma la variable de estudio. En nuestro caso la variable es “color favorito”, que por cierto es una variable cualitativa, y puede tomar los valores: negro, azul, amarillo y rojo.
b) Frecuencia absoluta (n_i): Es la cantidad de veces que aparece cada uno de los valores que puede tomar la variable. Es evidente que si sumo esa columna me tiene que salir el número total de datos, que en este ejemplo es de 20.
c) Frecuencia absoluta acumulada (N_i): Es el acumulado o suma de las frecuencias absolutas. El último valor de esta columna debe ser igual al número de datos.
d) Frecuencia relativa (f_i): Es la fracción o proporción que representa uno de los valores de la variable respecto del total de datos. Se calcula dividiendo la frecuencia absoluta del dato en cuestión entre el número total de datos. La suma de esta columna debe dar 1.
e) Frecuencia relativa acumulada (F_i): Es el acumulado o suma de las frecuencias relativas. El último valor de esta columna debe ser igual a 1.
f) Porcentaje: Es el tanto por ciento que representa cada uno de los valores de la variable respecto del total de datos. Se obtiene multiplicando la frecuencia relativa por 100.
g) Tanto por ciento acumulado: Es el acumulado o la suma de los pocentajes.

Ya podemos realizar la correspondiente tabla de frecuencias para nuestro estudio.

COLOR (X_i)	Frecuencia absoluta (n_i)	Frec. Absoluta acumulada (N_i)	Frecuencia relativa (f_i)	Frec. Relativa acumulada (F_i)	Porcentaje (%)	Porcentaje Acumulado (%)
Negro	4	4	4/20 = 0,2	0,2	20 %	20 %
Azul	5	9	5/20 = 0,25	0,45	25 %	45 %
Amarillo	5	14	5/20 = 0,25	0,7	25 %	70 %
Rojo	6	20	6/20 = 0,3	1	30 %	100 %
	20		1		100 %	





Al primer gráfico se le llama diagrama de barras, mientras que al segundo se le llama diagrama de sectores. La realización de ambos es muy sencilla. En el primero hemos puesto en el eje X los diferentes valores que puede tomar nuestra variable mientras que en el eje Y hemos puesto la frecuencia absoluta de cada uno de ellos.

En el segundo sector en lugar de representarlo con la frecuencia absoluta lo hemos hecho con el tanto por ciento. La única dificultad que podemos encontrar es decidir cuantos grados le corresponden a cada uno de los "quesitos". Para ello simplemente resolvemos una regla de tres. La vuelta entera son 360º y eso correspondería al 100 % de los datos, así que si queremos ver cuantos grados le corresponden al 25 % de los datos resolvemos la regla de tres correspondiente.

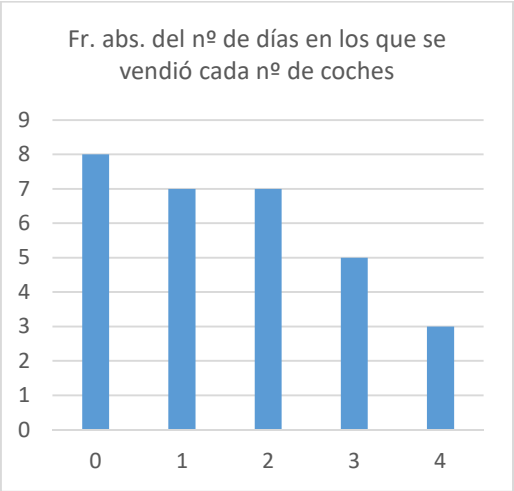
EJEMPLO 2: En una tienda de coches de segunda mano hacemos un registro del número de coches de la marca "Toyota" que se han vendido cada día del mes de septiembre, obteniendo los siguientes valores:

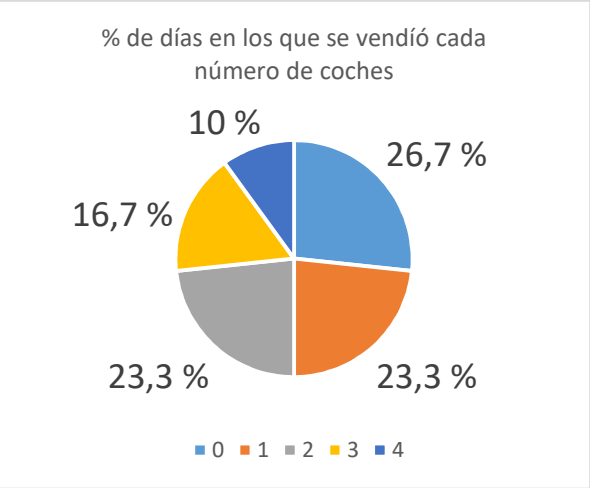
0, 1, 2, 1, 2, 0, 3, 2, 4, 0, 4, 2, 1, 0, 3, 0, 0, 3, 4, 2, 0, 1, 1, 3, 0, 1, 2, 1, 2, 3

Realizar una tabla de frecuencia para esos datos.

En primer lugar observamos que en esta ocasión la variable de nuestro estudio es cuantitativa discreta, ya que toma valores numéricos pero solo números naturales o 0.

Coches vendidos (X_i)	Frecuencia absoluta (n_i)	Frec. Absoluta acumulada (N_i)	Frecuencia relativa (f_i)	Frec. Relativa acumulada (F_i)	Porcentaje (%)	Porcentaje Acumulado (%)
0	8	8	8/30 = 0,267	0,267	26,7 %	26,7 %
1	7	15	7/30 = 0,233	0,5	23,3 %	50 %
2	7	22	7/30 = 0,233	0,733	23,3 %	73,3 %
3	5	27	5/30 = 0,167	0,9	16,7 %	90 %
4	3	30	3/30 = 0,1	1	10 %	100 %
	30		1		100 %	





EJEMPLO 3: En un centro comercial se consultó la edad a todas las personas que entraban entre las 12:00 y las 12:30. Los resultados obtenidos fueron los siguientes.

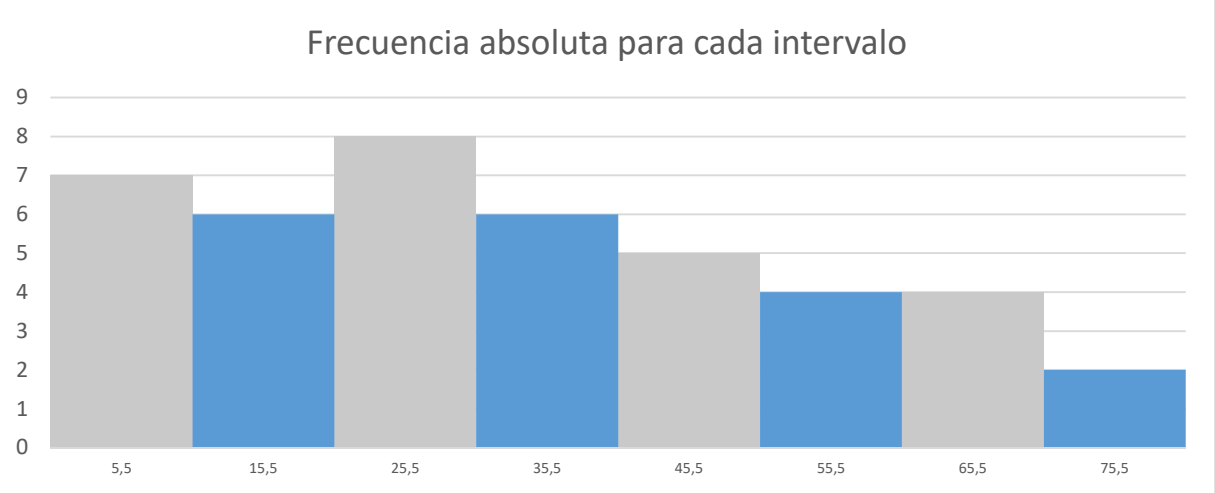
15, 73, 1, 65, 16, 3, 42, 36, 42, 3, 61, 19, 36, 47, 30, 45, 29, 73, 69, 34, 23, 22, 21, 33, 27, 55, 58, 17, 4, 17, 48, 25, 36, 11, 4, 54, 70, 51, 3, 34, 26, 10

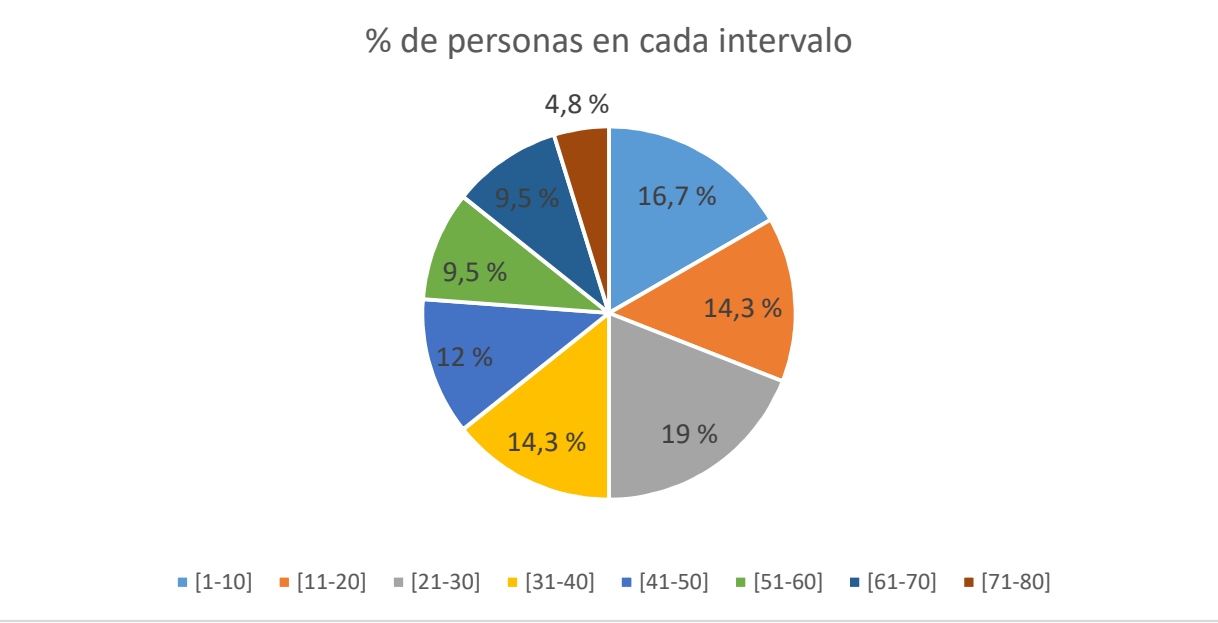
Haz una tabla de frecuencias agrupando los datos en intervalos de longitud 10.

En primer lugar vemos que esta vez la variable del estudio es cuantitativa continua. Podría parecer discreta porque todos los datos son números naturales, pero eso es así porque lo hemos hecho para simplificar, ya que en realidad si diésemos la edad en segundos en lugar de en años ninguno de esos valores coincidiría. Para las variables cuantitativas continuas se suelen usar tablas de frecuencias en las que agrupamos los datos en intervalos ya que si no habría demasiados valores diferentes en la primera columna. Cuando hagamos una tabla de frecuencias con datos agrupados, añadiremos una columna justa a la derecha de la de los intervalos que llamaremos MARCA DE CLASE. En ella lo que haremos será escribir la media de los límites de cada intervalo. En los cálculos que veremos más adelante para calcular los parámetros estadísticos de centralización usaremos los datos de esa columna.

Intervalos	Marca de clase (X_i)	Frecuencia absoluta (n_i)	Frec. Absoluta acumulada (N_i)	Frecuencia relativa (f_i)	Frec. Relativa acumulada (F_i)	Porcentaje (%)	Porcentaje Acumulado (%)
1-10	5,5	7	7	7/42 = 0,167	0,167	16,7 %	16,7 %
11-20	15,5	6	13	6/42 = 0,143	0,309	14,3 %	30,9 %
21-30	25,5	8	21	8/42 = 0,19	0,5	19 %	50 %
31-40	35,5	6	27	6/42 = 0,143	0,64	14,3 %	64 %
41-50	45,5	5	32	5/42 = 0,12	0,76	12 %	76 %
51-60	55,5	4	36	4/42 = 0,095	0,86	9,5 %	86 %
61-70	65,5	4	40	4/42 = 0,095	0,95	9,5 %	95 %
71-80	75,5	2	42	2/42 = 0,048	1	4,8 %	100 %
		42		1		100 %	

Cuando tenemos una variable cuantitativa continua y hemos agrupado en intervalos, en lugar de un diagrama de barras haremos lo que se denomina un HISTOGRAMA. Es básicamente lo mismo, solo que ahora los rectangulitos que dibujaremos estarán juntos y sus extremos serán los límites del intervalo. En el centro de cada uno de los rectángulos se encontraría la marca de clase.





Ya hemos visto lo suficiente acerca de la clasificación y representación gráfica de una serie de datos estadísticos. Vamos ahora a estudiar lo que conocemos como **PARÁMETROS ESTADÍSTICOS**. Un parámetro estadístico no es más que un número que resume de una sola vez las características de una serie de datos, de modo que si estamos hablando de miles y miles de datos nos serán de mucha utilidad para hacernos una idea de lo que estamos viendo sin tener que detenernos estudiar todos los datos.

Dentro de los parámetros estadísticos estudiaremos de dos tipos:

a) Parámetros de centralización:
   - Media
   - Mediana
   - Moda

b) Parámetros de dispersión:
   - Desviación típica y varianza

A) PARÁMETROS DE CENTARLIZACIÓN:

a) Media: la media de un conjunto de datos se calcula sumándolos todos y dividiendo ese resultado entre el número total de datos. La escribimos de la siguiente forma:

$$\bar{x} = \frac{\sum_i x_i}{N}$$

Si tenemos los datos ya ordenados en una tabla de frecuencias, no será necesario sumarlos todos, porque ya sabremos cuantos hay de cada clase y para eso se ha inventado la multiplicación. Lo expresamos entonces de la siguiente forma:

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

La fórmula de arriba lo que te dice es que si sabes que hay 200 personas que tienen 2 hermanos no te pongas a sumar el número 2 doscientas veces, sino que para eso está la multiplicación y sumar doscientas veces el número 2 es lo mismo que multiplicar el 2 por doscientos. Así cada dato de nuestro estudio lo multiplicamos por su frecuencia absoluta y vamos sumando esos resultados. Al final dividiremos el resultado de esa suma entre el número total de datos.

La media, como es obvio, no se puede usar para datos cualitativos. Para datos cuantitativos es el parámetro de centralización por excelencia.

Ej: Calcula la media de las notas en una clase de matemáticas en la que se han obtenido las siguientes calificaciones:

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

$$\bar{x} = \frac{8+5+6+3+7+7+4+5+5+5+6+6+7+4+4+3+3+2+2+5+4+7+8+8}{24}$$

$$\bar{x} = 5{,}167$$

Si nos hemos molestado previamente en hacer una tabla de frecuencias o directamente nos lo dan ya clasificado:

X_i	n_i
2	2
3	3
4	4
5	5
6	3
7	4
8	3
	TOTAL: 24

$$\bar{x} = \frac{2 \cdot 2 + 3 \cdot 3 + 4 \cdot 4 + 5 \cdot 5 + 6 \cdot 3 + 7 \cdot 4 + 8 \cdot 3}{24} = 5{,}167$$

b) Mediana: la mediana es simplemente el dato que ocupa la posición central de nuestro conjunto de datos. Como tenemos que ordenar los datos de menor a mayor, queda claro que es un parámetro que no sirve para variables cualitativas, porque no se puede ordenar, por ejemplo, el color de pelo.

Ej: Calcula la mediana de las notas de matemáticas de la siguiente clase:

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

Primero ordenamos los datos de menor a mayor: 1, 2, 2, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 7, 8

En total hay 15 datos, así que el que ocupa la posición central es el dato número 8 (deja 7 datos más pequeños a su izquierda y 7 datos mayores a su derecha). Contamos desde la izquierda y la mediana será aquel dato que ocupe la octava posición:

1, 2, 2, 3, 3, 4, 4, **4**, 5, 5, 5, 6, 6, 7, 8.

Por tanto la mediana es 4.

Ej: Calcula la mediana de las notas de matemáticas de la siguiente clase:

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

Primero ordenamos los datos de menor a mayor: 1, 2, 2, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 7, 8

En total hay 16 datos, así que ahora no hay uno que ocupe la posición central. En ese caso cogemos los dos centrales y calculamos la media entre ellos.

1, 2, 2, 3, 3, 4, 4, **4**, **5**, 5, 5, 5, 6, 6, 7, 8. $Mediana = \frac{4+5}{2} = 4{,}5$

c) Moda: la moda es el dato más repetido de un conjunto de datos. Este parámetro es que se usa cuando tenemos datos cualitativos ya que no podemos hacer ninguno de los anteriores.

Ej: En una clase de 1º de ESO se ha preguntado a los alumnos cuál es su equipo de fútbol favorito obteniendo las siguientes respuestas:

Madrid, Madrid, Atlético, Barcelona, Betis, Sevilla, Barcelona, Madrid, Atlético, Atlético, Valencia, Sevilla, Madrid, Barcelona, Valencia, Atlético, Madrid, Madrid, Atlético, Betis.

Lo único que tenemos que hacer es contar los datos para averiguar cuál es el que más se repite. En este caso el dato que más se repite es Madrid. Por lo tanto la moda del conjunto de datos es: Madrid.

EJEMPLO 4: Realizamos un estudio acerca del número de coches de la marca "Toyota" que se han vendido a lo largo de los días del mes de septiembre. Ya hemos hecho el trabajo previo y tenemos la siguiente tabla de frecuencias:

Coches vendidos (X_i)	Frecuencia absoluta (n_i)	Frec. Absoluta acumulada (N_i)	Frecuencia relativa (f_i)	Frec. Relativa acumulada (F_i)	Porcentaje (%)	Porcentaje Acumulado (%)
0	8	8	8/30 = 0,267	0,267	26,7 %	20 %
1	7	15	7/30 = 0,233	0,5	23,3 %	50 %
2	7	22	7/30 = 0,233	0,733	23,3 %	73,3 %
3	5	27	5/30 = 0,167	0,9	16,7 %	90 %
4	3	30	3/30 = 0,1	1	10 %	100 %
	30		1		100 %	

a) **¿Cuál es la media del número de coches vendidos en un día?**
b) **¿Cuál es la mediana del número de coches vendidos en un día?**

Como tenemos la tabla de frecuencias calcularemos la media usando las frecuencias absolutas para ahorrarnos trabajo. La media será por tanto:

$$\bar{x} = \frac{0 \cdot 8 + 1 \cdot 7 + 2 \cdot 7 + 3 \cdot 5 + 4 \cdot 3}{30} = 1{,}6$$

Esto quiere decir que por término medio cada día se han vendido 1,6 coches. Está claro que no se pueden vender un número decimal de coches, pero nos sirve como resumen del número de coches que se han vendido cada día.

Para calcular la mediana tenemos que ver el cuál es el dato que se me queda en medio si ordeno los datos de menor a mayor. En este caso tenemos un número par de datos, así que tendremos que proceder como en el ejemplo anterior.

Podríamos ordenar los datos para verlo:

0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, **1, 2**, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4

Como hay dos datos centrales la mediana será la media de esos valores: $Mediana = \frac{1+2}{2} = 1{,}5$

Pero esto no era necesario porque hemos incorporado a la tabla la columna de % acumulado. En ella podemos observar que contando los 0 y los 1 tenemos exactamente el 50 % de los datos, así que de ahí en adelante habrá también el 50 % de los datos. Esto nos hace ver directamente que los datos centrales son el 1 y el 2 y procedemos igual.

EJEMPLO 5: Vamos ahora a reutilizar la tabla de frecuencias que hicimos en el ejemplo 3 de datos agrupados en intervalos.

Intervalos	Marca de clase (X_i)	Frecuencia absoluta (n_i)	Frec. Absoluta acumulada (N_i)	Frecuencia relativa (f_i)	Frec. Relativa acumulada (F_i)	Porcentaje (%)	Porcentaje Acumulado (%)
1-10	5,5	7	7	7/42 = 0,167	0,167	16,7 %	16,7 %
11-20	15,5	6	13	6/42 = 0,143	0,309	14,3 %	30,9 %
21-30	25,5	8	21	8/42 = 0,19	0,51	19 %	51 %
31-40	35,5	6	27	6/42 = 0,143	0,64	14,3 %	64 %
41-50	45,5	5	32	5/42 = 0,12	0,76	12 %	76 %
51-60	55,5	4	36	4/42 = 0,095	0,86	9,5 %	86 %
61-70	65,5	4	40	4/42 = 0,095	0,95	9,5 %	95 %
71-80	75,5	2	42	2/42 = 0,048	1	4,8 %	100 %
		42		1		100 %	

a) **¿Cuál es la edad media de la gente que acudió al centro comercial a esas horas?**
b) **¿Cuál es el intervalo mediano?**

Supongamos que no tenemos como en el ejemplo anterior los datos desarrollados y que solo tenemos la tabla de frecuencias. ¿Cómo calculamos entonces la media? Pues vamos a suponer que todas las personas que están clasificadas en el primer intervalo (aquellos que tienen entre 1 y 10 años) tienen 5,5 años. Para eso añadimos la columna de la marca de clase. Es cierto que estamos haciendo un poco de trampa, pero el error será muy pequeño, así que lo asumimos.

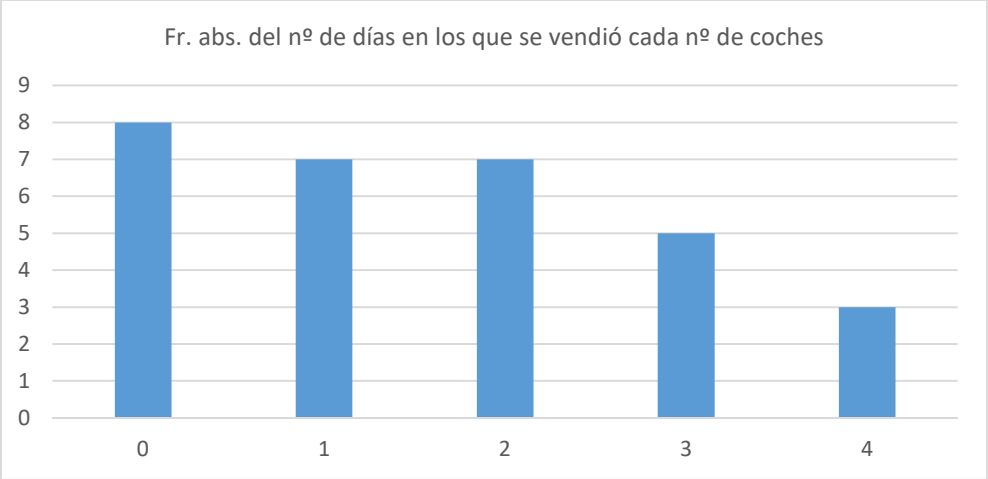
$$\bar{x} = \frac{5{,}5 \cdot 7 + 15{,}5 \cdot 6 + 25{,}5 \cdot 8 + 35{,}5 \cdot 6 + 45{,}5 \cdot 5 + 55{,}5 \cdot 4 + 65{,}5 \cdot 4 + 75{,}5 \cdot 2}{42} = 33{,}6\ años$$

Para ver cuál es el intervalo mediano vamos a usar directamente la columna de los porcentajes acumulados. El intervalo mediano es aquel en el que se supera por primera vez el 50 %. ¿Por qué? Porque eso quiere decir que hasta ese intervalo (incluido) se encuentran la mitad de los datos, que es exactamente lo que quiere decir la mediana.

En intervalo mediano será entonces: [21-30]. Esto quiere decir que desde la edad de 1 año hasta 31 se encuentran el 50 % de las personas que entraron al centro comercial a las horas en cuestión.

EJEMPLO 6: La siguiente gráfica muestra el número de coches de la marca "Toyota" vendidos al día durante el mes de septiembre en una tienda de coches de segunda mano.

a) ¿Cuál es la media del número de coches vendidos al día?



En realidad es como si nos hubiesen dado la tabla de frecuencias, porque mirando a la gráfica se puede ver fácilmente que se han vendido 0 coches un total de 8 días, se ha 1 coche un total de 7 días, etc. Por tanto, no tenemos ningún problema para calcular la media del mismo modo que hicimos cuando nos dieron la tabla de frecuencias:

$$\bar{x} = \frac{0 \cdot 8 + 1 \cdot 7 + 2 \cdot 7 + 3 \cdot 5 + 4 \cdot 3}{30} = 1{,}6\ coches$$

B) PARÁMETROS DE DISPERSIÓN:

Los parámetros de centralización están bien como resumen de un conjunto de datos, pero muchas veces queremos más información. Para eso utilizaremos los parámetros de dispersión. Imaginemos que tenemos las notas de dos clases de 10 alumnos de un examen de matemáticas:

Clase A: 5, 5, 5, 5, 5, 5, 5, 5, 5, 5

Clase B: 0, 10, 0, 10, 0, 10, 0, 10, 0, 10

Podemos calcular fácilmente la nota media de ambas clases y obtenemos:

$$\bar{x}(clase\ A) = \frac{5 \cdot 10}{10} = 5$$

$$\bar{x}(clase\ B) = \frac{0 \cdot 5 + 10 \cdot 5}{10} = 5$$

Resulta que las dos clases tienen la misma nota media, ¿podemos deducir entonces que las dos clases son similares? Para nada. En una de ellas todos los alumnos obtuvieron un 5, mientras que en la otra clase la mitad de ellos sacaron un 0 y la otra mitad obtuvieron un 10. ¿Cómo podemos complementar la media para obtener un mejor resumen de la información? (NOTA: Imagina que son las notas de los 10.000 alumnos de 1º de ESO de Madrid, no voy a poner a

leerme todas las notas para ver la situación, leeré el resumen de los datos que me ofrecen los parámetros de centralización y dispersión).

La primera idea que se nos viene a la cabeza es medir la distancia que hay de cada dato a la media que hemos calculado y luego hacer la media de esas distancias. Veámoslo en una tabla:

CLASE A ($\bar{x} = 5$)			**CLASE B ($\bar{x} = 5$)**	
Notas	Distancia a la media		Notas	Distancia a la media
5	5 – 5 = 0		10	10 – 5 = 5
5	5 – 5 = 0		0	0 – 5 = -5
5	5 – 5 = 0		10	10 – 5 = 5
5	5 – 5 = 0		0	0 – 5 = -5
5	5 – 5 = 0		10	10 – 5 = 5
5	5 – 5 = 0		0	0 – 5 = -5
5	5 – 5 = 0		10	10 – 5 = 5
5	5 – 5 = 0		0	0 – 5 = -5
5	5 – 5 = 0		10	10 – 5 = 5
$\boldsymbol{Media\ de\ las\ distancias}$ $\boldsymbol{\frac{10 \cdot 0}{10} = 0}$			$Media\ de\ las\ distancias$ $\frac{5 \cdot 5 + 5 \cdot (-5)}{10} = 0$	

¿Qué ha ocurrido? En la clase A no ha pasado nada raro ya que todas las notas son 5, así que la distancia de cada una de ellas a la media (que es 5 también) siempre da 0 y, por tanto, la media de esas distancias es 0. Sin embargo, en la clase B no esperábamos que nos diese lo mismo, ¿qué ha pasado? Que como unas distancias son positivas y otras negativas se han compensado al hacer la media dando como resultado el mismo.

Para solucionar esto podemos hacer varias cosas, la más inmediata sería calcular esas distancias pero en valor absoluto, haciendo que los valores positivos se conviertan en positivos.

CLASE A ($\bar{x} = 5$)			**CLASE B ($\bar{x} = 5$)**	
Notas	Distancia a la media en valor absoluto		Notas	Distancia a la media en valor absoluto
5	|5 – 5| = 0		10	|10 – 5 |= 5
5	|5 – 5| = 0		0	|0 – 5| = 5
5	|5 – 5| = 0		10	|10 – 5 |= 5
5	|5 – 5| = 0		0	|0 – 5| = 5
5	|5 – 5| = 0		10	|10 – 5 |= 5
5	|5 – 5| = 0		0	|0 – 5| = 5
5	|5 – 5| = 0		10	|10 – 5 |= 5
5	|5 – 5| = 0		0	|0 – 5| = 5
5	|5 – 5| = 0		10	|10 – 5 |= 5
$\boldsymbol{Media\ del\ valor\ absoluto\ de\ las\ distancias}$ $\boldsymbol{\frac{10 \cdot 0}{10} = 0}$			$Media\ del\ valor\ absoluto\ de\ las\ distancias$ $\frac{10 \cdot 5}{10} = 5$	

A esto parámetro se le llama **Desviación Media**. Así que tendríamos como resumen:

Clase A: $\bar{x} = 5$ Desviación media = 0

Clase B: $\bar{x} = 5$ Desviación media = 5

De un vistazo y sin ver todos los datos sabemos que en la primera clase la media es 5 y además sabemos que todas las notas son un 5 ya que la desviación media es 0. Del mismo modo vemos que en la segunda clase la nota media es igual pero observamos que de media las notas están a 5 de distancia de la media, por lo que serán o bien 0, o bien dieces.

Sin embargo, por motivos que no son para nada abordables en este curso, a los matemáticos no nos gusta nada este parámetro de dispersión. Veamos entonces como solucionar de otra forma el problema de que las desviaciones negativas se me compensen con las negativas. ¿De qué otra forma hacemos que algo negativo se convierta en positivo? Elevándolo al cuadrado. Vamos a formar la misma tabla y lo que haremos ahora será calcular la distancia a la media de cada dato y elevar cada uno de esos resultados al cuadrado.

CLASE A ($\bar{x} = 5$)			**CLASE B ($\bar{x} = 5$)**	
Notas	Distancia a la media al cuadrado		Notas	Distancia a la media al cuadrado
5	$(5-5)^2 = 0$		10	$(10-5)^2 = 5^2 = 25$
5	$(5-5)^2 = 0$		0	$(0-5)^2 = (-5)^2 = 25$
5	$(5-5)^2 = 0$		10	$(10-5)^2 = 5^2 = 25$
5	$(5-5)^2 = 0$		0	$(0-5)^2 = (-5)^2 = 25$
5	$(5-5)^2 = 0$		10	$(10-5)^2 = 5^2 = 25$
5	$(5-5)^2 = 0$		0	$(0-5)^2 = (-5)^2 = 25$
5	$(5-5)^2 = 0$		10	$(10-5)^2 = 5^2 = 25$
5	$(5-5)^2 = 0$		0	$(0-5)^2 = (-5)^2 = 25$
5	$(5-5)^2 = 0$		10	$(10-5)^2 = 5^2 = 25$
Media de las distancias al cuadrado $\frac{\mathbf{10 \cdot 0}}{\mathbf{10}} = \mathbf{0}$			*Media de las distancias al cuadrado* $\frac{10 \cdot 25}{10} = 25$	

A este nuevo parámetro de dispersión que acabamos de calcular se le llama **VARIANZA** y a los matemáticos nos gusta bastante más. Pero tiene también un problema, y es que nos está dando un valor de media de las distancias mucho más grande de lo que realmente deberían ser ya que hemos elevado al cuadrado cada una de las mismas. Ahora mismo como resumen de los datos tendríamos:

Clase A: $\bar{x} = 5$ Varianza = 0

Clase B: $\bar{x} = 5$ Varianza = 25

Llegamos ya al parámetro de dispersión por excelencia. La **DESVIACIÓN TÍPICA**. Para solucionar el problema anterior lo que vamos a hacer es calcular la varianza de esta misma forma pero cuando la tengamos calculada tomaremos la **RAÍZ CUADRADA POSITIVA** de ese dato.

Tendremos entonces como resumen lo siguiente (a la desviación típica se la nombra con la letra griega “sigma” en minúscula “σ”):

Clase A: $\bar{x} = 5$ $\sigma = \sqrt{varianza} = \sqrt{0} = 0$

Clase B: $\bar{x} = 5$ $\sigma = \sqrt{varianza} = \sqrt{25} = 5$

Para calcular la desviación típica de un conjunto de datos procederemos como hemos hecho anteriormente, pero vamos a poner un ejemplo en el que me den los datos ya ordenados en una tabla de frecuencias:

EJEMPLO 7: En un colegio de línea 4 las notas de matemáticas en 1º de ESO de la segunda evaluación están recogidas en la siguiente tabla de frecuencias:

Notas (X_i)	**Frecuencia absoluta (n_i)**
0	5
1	10
2	12
3	15
4	20
5	30
6	25
7	20
8	15
9	10
10	5
	TOTAL: 167

Lo primero que tenemos que hacer es calcular la media para poder luego calcular las distancias respecto de ella. Como tenemos las frecuencias absolutas el trabajo es menos complicado:

$$\bar{x} = \frac{5 \cdot 0 + 10 \cdot 1 + 12 \cdot 2 + 15 \cdot 3 + 20 \cdot 4 + 30 \cdot 5 + 25 \cdot 6 + 20 \cdot 7 + 15 \cdot 8 + 10 \cdot 9 + 5 \cdot 10}{167}$$

$$\bar{x} = 5{,}144$$

Para calcular la desviación típica nos vamos a ayudar de esa misma tabla a la que vamos a añadir una columna más.

Notas (X_i)	Frecuencia absoluta (n_i)	Distancia del dato a la media al cuadrado
0	5	$(0 - 5{,}144)^2 = 26{,}46$
1	10	$(1 - 5{,}144)^2 = 17{,}17$
2	12	$(2 - 5{,}144)^2 = 9{,}88$
3	15	$(3 - 5{,}144)^2 = 4{,}59$
4	20	$(4 - 5{,}144)^2 = 2{,}08$
5	30	$(5 - 5{,}144)^2 = 0{,}02$
6	25	$(6 - 5{,}144)^2 = 0{,}73$
7	20	$(7 - 5{,}144)^2 = 3{,}44$
8	15	$(8 - 5{,}144)^2 = 8{,}16$
9	10	$(9 - 5{,}144)^2 = 14{,}87$
10	5	$(10 - 5{,}144)^2 = 23{,}58$
	TOTAL: 167	

Podemos calcular ya la **VARIANZA** que sabemos que es la media de los cuadrados de las desviaciones de cada dato respecto de la media, pero OJO, el dato 0 aparece 5 veces, así que tendré que sumar 5 veces esa distancia o lo que es lo mismo, multiplicarla por 5, etc.

$$VAR = \frac{26{,}46 \cdot 5 + 17{,}17 \cdot 10 \ldots\ldots\ldots\ldots\ldots\ldots + 14{,}87 \cdot 9 + 23{,}58 \cdot 10}{167} = 5{,}96$$

Ya tenemos la **VARIANZA** pero nos pidieron la **DESVIACIÓN TÍPICA**, que no es otra cosa que la raíz cuadrada de la varianza.

$$\sigma = \sqrt{VAR} = \sqrt{5{,}96} = 2{,}44$$

Así que como resumen final tenemos que en esos 6 cursos de 1º de ESO la nota media fue de un 5,144 con una desviación típica de 2,44.

La desviación típica es un cálculo laborioso, pero la calculadora lo hace sin mayores problemas, así que aprenderemos a calcularla con ella, pero sí que hay que tener claras varias cosas:

1.- LA DESVIACIÓN TÍPICA SIEMPRE ES POSITIVA, porque la calculamos quedándonos con la raíz positiva de la varianza.

2.- LA DESVIACIÓN TÍCICA NUNCA VALE 0, salvo que todos los datos de nuestro estudio sean iguales, pero eso rara vez se da en la vida real.

3.- LA DESVIACIÓN TÍPICA NO PUEDE VALER MÁS QUE LA DISTANCIA QUE HAYA ENTRE LA MEDIA Y CUALQUIERA DE LOS DATOS (O BIEN MÁS GRANDE O BIEN MÁS PEQUEÑO), porque es la media de las distancias respecto de la media, así que ese caso no se puede dar.