Tema 7: Estadística.

7.1 Conceptos estadísticos.

- La **Estadística** es una parte importante de las Matemáticas que se encarga de recoger, organizar, analizar e interpretar datos con el objetivo de obtener información útil y sacar conclusiones útiles a partir de ellos.
- La **población**, también conocida como universo o colectivo, es el conjunto de referencia objeto de estudio en el que se realizan las observaciones y se recogen los datos.
- Un **individuo** es cada uno de los elementos que componen la población estadística. Se trata de un ente observable y no tiene por qué ser una persona: puede ser un objeto, un ser vivo o incluso algo abstracto.
- La **muestra** es un subconjunto de elementos o individuos de la población estadística. Se toman muestras cuando resulta difícil, costoso o imposible observar todos los elementos de la población estadística.
- Los **caracteres estadísticos**: son cualidades o propiedades inherentes al individuo. La observación y estudio del individuo se centran en recopilar datos de uno o más caracteres estadísticos. Pueden ser de tres tipos:
  - **Cualitativos:** que describen cualidades o categorías y no se expresan con números. Por ejemplo: el color de ojos, la profesión o la marca de coche.
  - **Cuantitativos discretos:** toman valores aislados, normalmente números naturales, y no admiten valores intermedios. Por ejemplo: el número de goles marcados, el número de hijos, el número de discos comprados o el número de pulsaciones.
  - **Cuantitativos continuos:** pueden tomar cualquier valor dentro de un intervalo de números reales, por lo que suelen expresarse mediante números decimales. Ejemplos de este tipo de variables son la altura, el peso, el volumen, la presión sanguínea o la temperatura. En las variables continuas es habitual agrupar los datos en intervalos de clase contiguos, con el fin de facilitar su análisis sin perder demasiada información. Los extremos de cada intervalo se denominan extremos de clase, mientras que el punto medio recibe el nombre de marca de clase x_i, que es valor que representará a todo el intervalo. Si procuramos que todas las clases tengan la misma amplitud y los límites de cada clase sean números redondos (por ejemplo, múltiplos de 5), conseguiremos simplificar mucho los cálculos.
- Se llama **tamaño muestral** N al número de observaciones realizadas o, lo que es lo mismo, al número total de datos recogidos.
- La **frecuencia absoluta** f_i de un valor x_i de la variable es el número de veces que dicho valor se repite en el conjunto de las observaciones realizadas.
- La **frecuencia relativa** h_i de un valor x_i de la variable es el cociente, o tanto por 1, entre la frecuencia absoluta y el número de observaciones realizadas.
- La **frecuencia absoluta acumulada** F_i del valor x_i de la variable es la suma de las frecuencias absolutas de los valores inferiores o iguales a él. Para calcularla, los valores de la variable deben ordenarse previamente de menor a mayor. La frecuencia absoluta acumulada del último valor será siempre N.

- La **frecuencia relativa acumulada** H_i del valor x_i es el cociente entre la frecuencia absoluta acumulada y el número de observaciones realizadas N. La frecuencia relativa acumulada del último valor será siempre 1

  Ejemplo: consideremos la variable estadística discreta asociada al experimento consistente en anotar las calificaciones de matemáticas de un colectivo de 50 alumnos. Realiza con los datos una tabla de frecuencias.
  Los resultados han sido: 1-6-8-8-2-2-3-4-5-10-3-4-5-6-7-8-9-7-7-6-5-5-5-4-4-5-6-7-10-4-1-2-5-5-6-6-7-4-5-6-5-4-6-7-6-5-4-3-4-5.

x_i	f_i	h_i	%	F_i	H_i	%Ac
1	2	0,04	4	2	0,04	4
2	3	0,06	6	5	0,10	10
3	3	0,06	6	8	0,16	16
4	9	0,18	18	17	0,34	34
5	12	0,24	24	29	0,58	58
6	9	0,18	18	38	0,76	76
7	6	0,12	12	44	0,88	88
8	3	0,06	6	47	0,94	94
9	1	0,02	2	48	0,96	96
10	2	0,04	4	**50**	**1**	**100**
Total	**50**	**1**	**100**			

Otro ejemplo: consideremos la variable estadística continua asociada al experimento consistente en medir la estatura (en cm) de un grupo de personas. Los datos obtenidos vienen ahora recogidos mediante intervalos en esta tabla:

Estatura	[140,150)	[150,160)	[160,170)	[170,180)	[180,190)
Nº personas	6	9	14	15	6

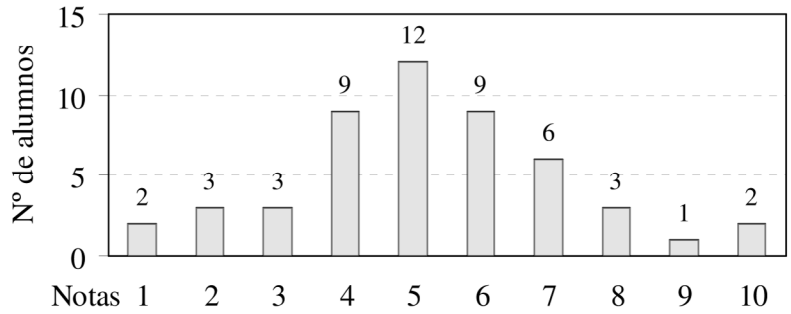
Los posibles valores x_i de la variable estadística son ahora las marcas de la clase de cada intervalo, que se obtienen sumando los extremos del intervalo y dividiendo por 2. Por tanto, los valores de x_i son: 145, 155, 165, 175 y 185.

La tabla de frecuencias es ahora la siguiente:

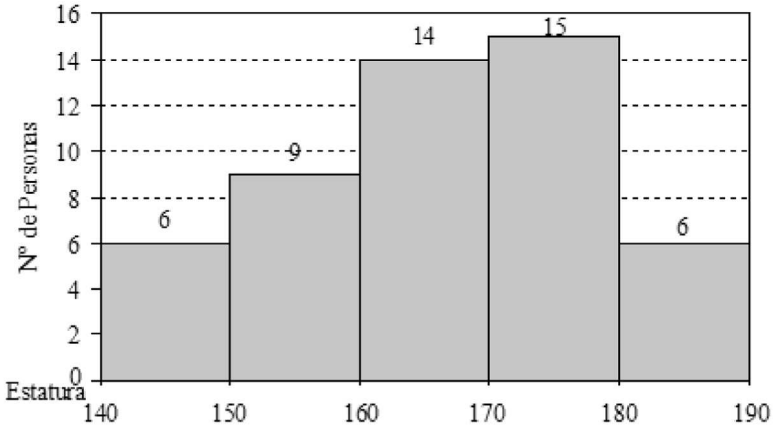
Intervalo	x_i	f_i	h_i	%	F_i	H_i	%Ac
[140,150)	145	6	0,12	12	6	0,12	12
[150,160)	155	9	0,18	18	15	0,30	30
[160,170)	165	14	0,28	28	29	0,58	58
[170,180)	175	15	0,30	30	44	0,88	88
[180,190)	185	6	0,12	12	**50**	**1**	**100**
	Total	**50**	**1**	**100**			

7.2 Gráficos.

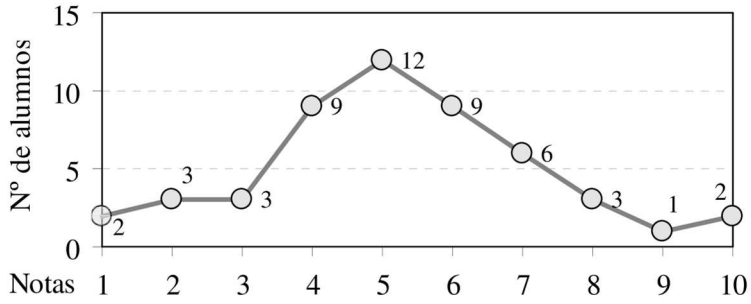
- **Diagrama de barras**: colocamos en el eje de abscisas los valores de la variable x_i, y en el eje de ordenadas los valores de las frecuencias. Dibujamos barras de igual anchura cuya altura sea exactamente la frecuencia. Así, en el ejemplo de las notas, tenemos::



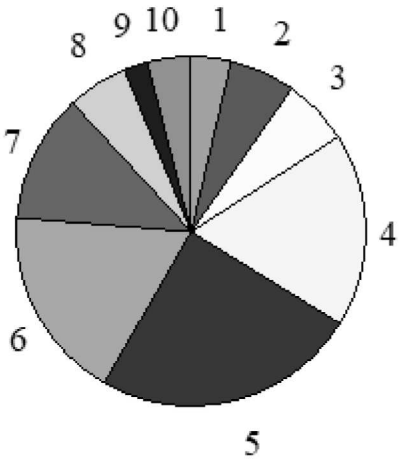
- **Histograma**: igual que el diagrama de barras, pero se utiliza en variables estadísticas continuas. Por lo tanto, las barras están unidas entre sí y cada una se sitúa sobre el intervalo de la clase. En el ejemplo de las estaturas, tenemos:

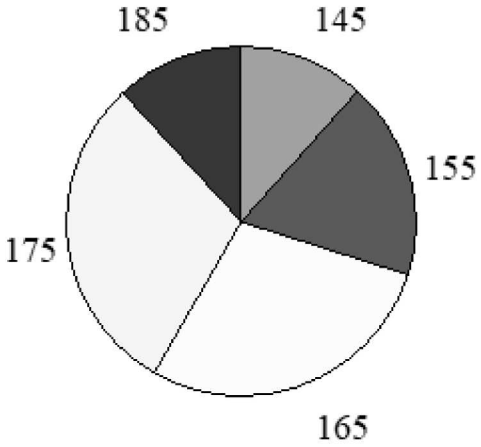


- **Polígono de frecuencias**: se unen los centros de las bases superiores de las barras tanto del diagrama de barras como del histograma. De esta manera, se puede observar con claridad la forma de la distribución.



- **Diagrama de sectores**: se obtiene dividiendo el círculo en sectores cuyo área sea proporcional a cada frecuencia respectiva. Para ello, se obtiene el ángulo central que ocupará cada sector mediante una proporcionalidad directa. Por ejemplo, si al total de la suma de frecuencias 50 le corresponden 360º, a la frecuencia f_i le corresponderán xº, es decir, $x^\circ = \frac{f_i \cdot 360^\circ}{50}$. En los ejemplos de las notas y las estaturas resultan los siguientes gráficos:





- **Pictogramas**: se construyen a partir del diagrama de barras, donde se sustituyen éstas por un dibujo de altura proporcional a las frecuencias, lo que hace más intuitiva la interpretación de los resultados. Por ejemplo, podríamos sustituir las barras por dibujos de libros.

7.3 Parámetros estadísticos.

Los parámetros estadísticos son valores numéricos que resumen y describen las características principales en un estudio de la población.

La **media aritmética** se obtiene al sumar todos los datos obtenidos de la variable y dividir por el número total de observaciones. Para una tabla de frecuencias en la que a cada valor de la variable x_i le corresponda una

frecuencia absoluta f_i, se puede calcular la media (que se representa por $\overline{x}$) de la siguiente manera: $\overline{x} = \dfrac{\sum_{i=1}^{n} x_i \cdot f_i}{\sum_{i=1}^{n} f_i} = \dfrac{\sum_{i=1}^{n} x_i \cdot f_i}{N}$

En el ejemplo de las notas de matemáticas, la media se obtiene así:

x_i	f_i	$x_i \cdot f_i$
1	2	2
2	3	6
3	3	9
4	9	36
5	12	60
6	9	54
7	6	42
8	3	24
9	1	9
10	2	20
	$N = 50$	$\sum_{i=1}^{n} x_i \cdot f_i = 262$

La media aritmética de las notas es: $\overline{x} = \dfrac{262}{50} = 5{,}24$.

En el ejemplo de las estaturas la media se obtiene de forma análoga, utilizando las marcas de clase de cada intervalo:

Intervalo	x_i	f_i	$x_i \cdot f_i$
$[140,150)$	145	6	870
$[150,160)$	155	9	1395
$[160,170)$	165	14	2310
$[170,180)$	175	15	2625
$[180,190)$	185	6	1110
	Total	50	8310

La media aritmética de las alturas es: $\overline{x} = \dfrac{8310}{50} = 166{,}2$.

- La **moda** es el valor que tiene mayor frecuencia absoluta. En el ejemplo de las notas, la moda es $M_0 = 5$, ya que esta nota le corresponde la mayor frecuencia (12). Si a dos o más valores les corresponde la misma frecuencia máxima, se dice que la distribución es bimodal o multimodal.

En el ejemplo de las estaturas el intervalo o clase modal es $[170,180)$, ya que en él se encuentra la mayor frecuencia (15). Sin embargo, si se desea hallar la moda con exactitud, aplicaremos la siguiente fórmula: $\boxed{M_0 = L_i + \frac{D_1}{D_1 + D_2} a}$.

Donde:

- L_i es el extremo inferior de la clase modal.
- D_1 es la diferencia entre la frecuencia absoluta modal y la frecuencia absoluta del intervalo previo a la clase modal.
- D_2 es la diferencia entre la frecuencia absoluta modal y la frecuencia absoluta del intervalo posterior a la clase modal.
- a es la amplitud de los intervalos.

En el ejemplo: $M_0 = 170 + \frac{1}{1+9} 10 \Rightarrow M_0 = 170 + \frac{1}{10} 10 \Rightarrow M_0 = 170 + 1 = 171$.

- La **mediana** M_e es el valor de la variable que excede al 50% de los datos. Es decir, al menos la mitad de los valores de la distribución es inferior o igual a M_e, y al menos la mitad es superior o igual a M_e.

  Para calcular la mediana en una variable discreta, se ordenan los datos de menor a mayor. Si hay un número impar de datos, la mediana es el que ocupa el lugar central. Si su número es par, se hace la media aritmética de los dos valores centrales. En el ejemplo de las notas, con un total de $N = 50$ valores, que es un número par, los dos valores centrales se encuentran en las posiciones 25 y 26. Al revisar la tabla de frecuencias absolutas acumuladas, vemos que ambos corresponden al valor 5 (ya que hay 29 valores menores o iguales a él), por tanto $M_e = 5$.

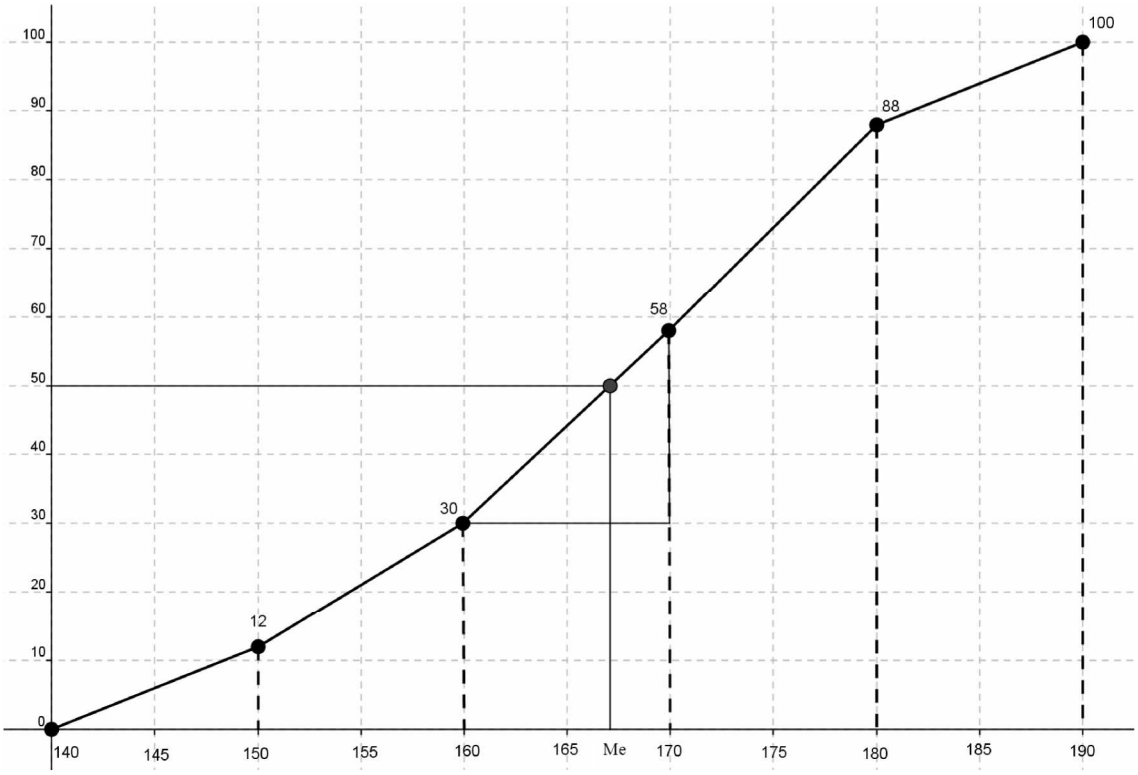
  Para calcular la mediana en una variable continua, como en el ejemplo de las estaturas, primero se identifica el intervalo mediano donde se supera el 50% de los datos. En este caso, el intervalo es $[160,170)$. Pero para calcular el valor exacto de la mediana, debemos seguir el siguiente proceso gráfico:

  - En primer lugar, consideramos la tabla de frecuencias y nos fijamos en la columna de intervalos y en los porcentajes de frecuencias relativas acumuladas.

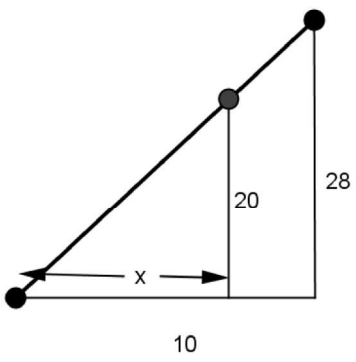
Intervalo	x_i	f_i	h_i	%	F_i	H_i	% Ac
$[140,150)$	145	6	0,12	12	6	0,12	12
$[150,160)$	155	9	0,18	18	15	0,30	30
$[160,170)$	165	14	0,28	28	29	0,58	58
$[170,180)$	175	15	0,30	30	44	0,88	88
$[180,190)$	185	6	0,12	12	50	1	100
	Total	50	1	100			

- A continuación, con los datos indicados anteriormente, construimos una tabla de valores que representaremos gráficamente y que se denomina polígono de frecuencias relativas acumuladas (en porcentaje).

Estat.	Frec.
140	0
150	12
160	30
170	58
180	88
190	100



- Se trata ahora de hallar un número M_e en el que su ordenada valga 50. Para ello, utilizaremos la semejanza de triángulos.



$$\frac{x}{10} = \frac{20}{28} \Rightarrow 28x = 200$$

Luego: $x = \frac{200}{28} = 7{,}1429$

Y el valor final de M_e es:

$$M_e = 160 + 7{,}1429 = 167{,}1429\,.$$

- Los **cuartiles** son tres valores que dividen a la serie en cuatro partes iguales.

  El cuartil Q_1 es el primer valor de la variable que excede al 25% de los datos.

  El cuartil Q_2 es el primer valor de la variable que excede al 50% de los datos (también conocido como mediana).

  El cuartil Q_3 es el primer valor de la variable que excede al 75% de los datos.

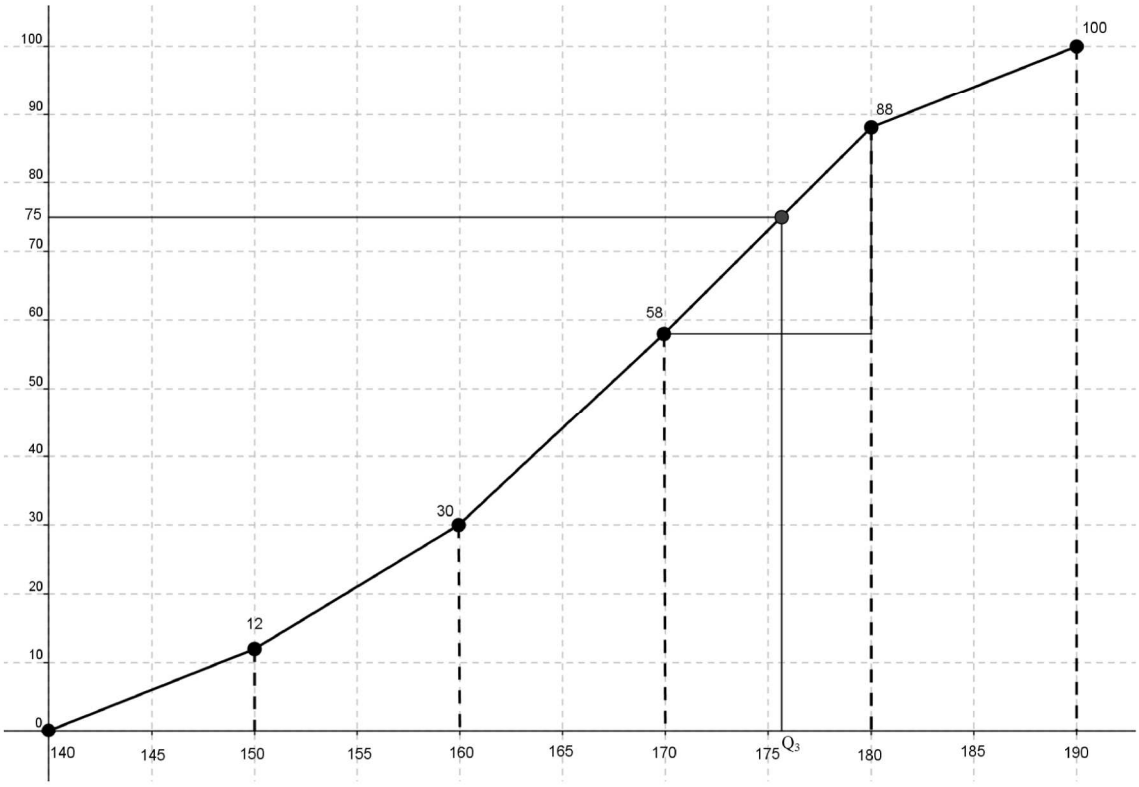
  Ejemplo: para calcular el tercer cuartil (Q_3) en el ejemplo de las notas, nos fijamos en el porcentaje de frecuencias acumuladas y buscamos el dato en el que se supera el valor 75%. En este caso, sucede para el 6, por lo que $Q_3 = 6$.

x_i	f_i	h_i	%	F_i	H_i	%Ac
1	2	0,04	4	2	0,04	4
2	3	0,06	6	5	0,1	10
3	3	0,06	6	8	0,16	16
4	9	0,18	18	17	0,34	34
5	12	0,24	24	29	0,58	58
6	**9**	**0,18**	**18**	**38**	**0,76**	**76**
7	6	0,12	12	44	0,88	88
8	3	0,06	6	47	0,94	94
9	1	0,02	2	48	0,96	96
10	2	0,04	4	50	1	100
Total	50	1	100			

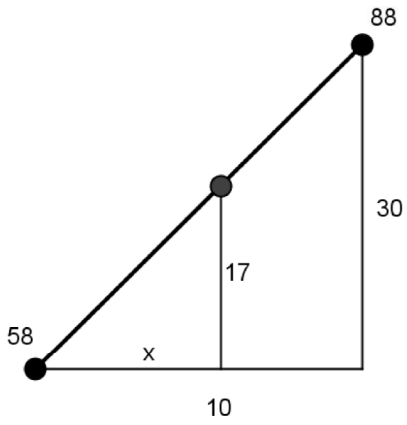
Para calcular un cuartil en una variable continua, debemos seguir un proceso gráfico análogo al estudiado con la mediana.

Por ejemplo, para calcular el cuartil Q_3 en el ejemplo de las estaturas, utilizamos la tabla de frecuencias, fijándonos en los porcentajes de frecuencias relativas acumuladas. A partir de estos datos, construimos una tabla de valores que darán lugar a los puntos del polígono de frecuencias relativas acumuladas (en porcentaje).

Estat.	Frec.
140	0
150	12
160	30
170	58
180	88
190	100



Se trata ahora de hallar un número Q_3 en el que su ordenada vale 75. Para ello, utilizaremos la semejanza de triángulos en el intervalo $[170,180)$.



$$\frac{x}{10}=\frac{17}{30}\Rightarrow 30x=170$$

Luego: $x=\frac{170}{30}=5{,}6667$

Y el valor final de Q_3 es:

$$Q_3=170+5{,}6667=175{,}6667\,.$$

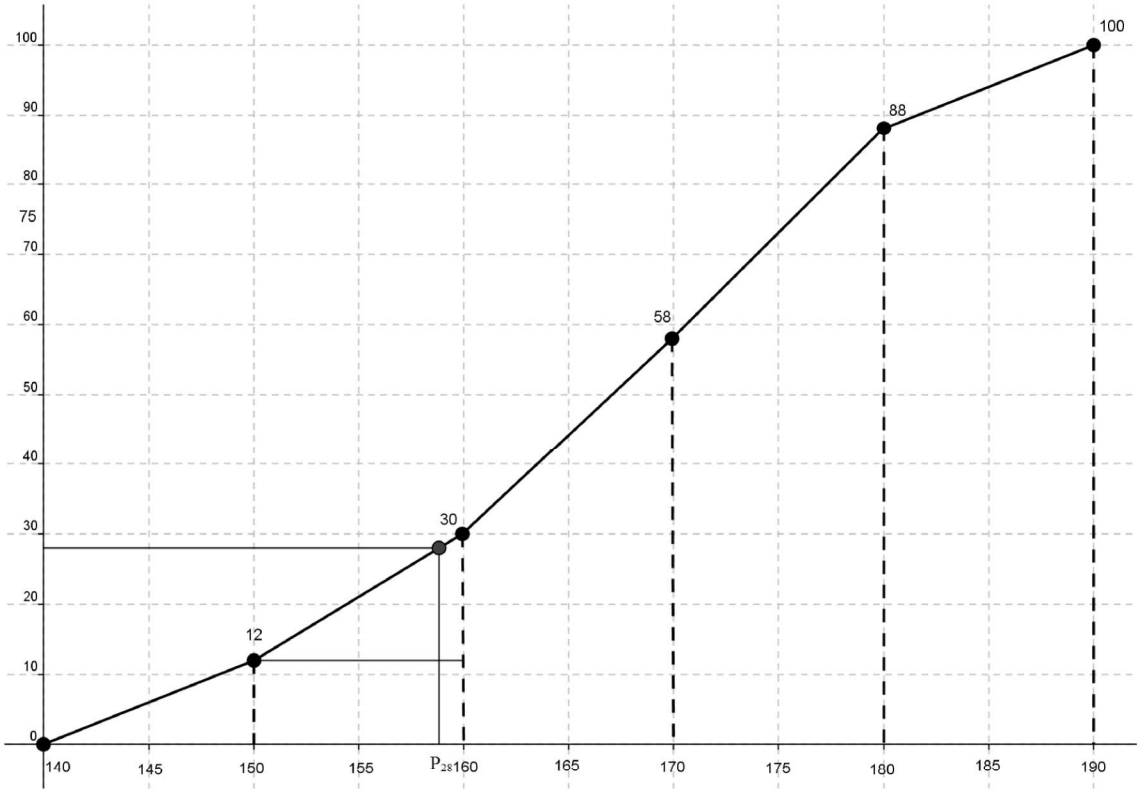
- Los **percentiles** son los valores que dividen la serie de datos en 100 partes iguales. Los percentiles dan los valores correspondientes al 1%, al 2%, ... , y así sucesivamente hasta el valor que indica el 99%. Claramente el valor P_{50} coincide con la mediana, y P_{25} y P_{75} coinciden con los cuartiles Q_1 y Q_3.
  Ejemplo: para calcular el percentil P_{28} en el ejemplo de las notas, nos fijamos en el porcentaje de frecuencias acumuladas y buscamos el dato en el que se supera el valor 28%, que sucede para el 4. Por tanto, $P_{28}=4$.

x_i	f_i	h_i	%	F_i	H_i	%Ac
1	2	0,04	4	2	0,04	4
2	3	0,06	6	5	0,1	10
3	3	0,06	6	8	0,16	16
4	**9**	**0,18**	**18**	**17**	**0,34**	**34**
5	12	0,24	24	29	0,58	58
6	9	0,18	18	38	0,76	76
7	6	0,12	12	44	0,88	88
8	3	0,06	6	47	0,94	94
9	1	0,02	2	48	0,96	96
10	2	0,04	4	50	1	100
Total	50	1	100			

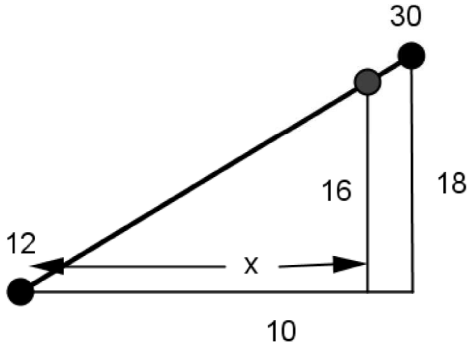
Para calcular un percentil en una variable continua, debemos seguir otra vez el proceso gráfico ya estudiado con la mediana y los cuartiles.
Ejemplo: calcular el percentil P_{28} en el ejemplo de las estaturas.

Igual que en otras ocasiones, utilizamos la tabla de frecuencias, fijándonos en los porcentajes de frecuencias relativas acumuladas, y a partir de estos datos construimos una tabla de valores que darán lugar a los puntos del polígono de frecuencias acumuladas (en porcentaje):

Estat.	Frec.
140	0
150	12
160	30
170	58
180	88
190	100



Se trata ahora de hallar un número P_{28} en el que su ordenada vale 28. Para ello, utilizaremos la semejanza de triángulos en el intervalo $[150,160)$.



$$\frac{x}{10}=\frac{16}{18}\Rightarrow 18x=160$$

Luego: $x=\frac{160}{18}=8,8889$

Y el valor final de P_{28} es:

$$P_{28}=150+8,8889=158,8889\,.$$

- **Varianza y desviación típica**. Se define la varianza de una distribución de frecuencias como el número obtenido de la siguiente expresión:

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

La raíz cuadrada de la varianza se llama desviación típica:

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

Cuanto mayor es la desviación típica, más alejados están los valores respecto de su valor medio. Es decir, mayor es el error que se comete al sustituirlos todos por su media aritmética.

Para aplicar la fórmula a los datos del ejemplo de las notas de matemáticas, realizamos la siguiente tabla:

x_i	f_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^2 \cdot f_i$
1	2	$-4{,}24$	17,9776	35,9552
2	3	$-3{,}24$	10,4976	31,4928
3	3	$-2{,}24$	5,0176	15,0528
4	9	$-1{,}24$	1,5376	13,8384
5	12	$-0{,}24$	0,0576	0,6912
6	9	0,76	0,5776	5,1984
7	6	1,76	3,0976	18,5856
8	3	2,76	7,6176	22,8528
9	1	3,76	14,1376	14,1376
10	2	4,76	22,6576	45,3152
	N=50			$\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot f_i = 203{,}12$

Con lo que se tiene: $s^2 = \dfrac{203{,}12}{50} = 4{,}0624$ y $s = \sqrt{4{,}0624} = 2{,}01554$.

OTRA FORMA: $$s^2 = \frac{\sum_{I=1}^{n} x_i^2 \cdot f_i}{N} - \overline{x}^2$$

x_i	x_i^2	f_i	$x_i^2 \cdot f_i$
1	1	2	2
2	4	3	12
3	9	3	27
4	16	9	144
5	25	12	300
6	36	9	324
7	49	6	294
8	64	3	192
9	81	1	81
10	100	2	200
	Total	N=50	1576

Así pues será: $s^2 = \dfrac{1576}{50} - (5{,}24)^2 = 4{,}0624$ y $s = \sqrt{4{,}0624} = 2{,}01554$.

- El número $\frac{s}{x}$ se llama **coeficiente de variación**. Mide la dispersión relativa y, cuanto mayor es, más dispersos están los datos. En nuestro ejemplo, el coeficiente de variación es: $\frac{2,0155}{5,24} = 0,3846$ (38,46%).

  Para el ejemplo de las estaturas, calcularíamos la desviación típica así:

Intervalo	x_i	f_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^2 \cdot f_i$
$[140,150)$	145	6	$-21,2$	449,44	2696,64
$[150,160)$	155	9	$-11,2$	125,44	1128,96
$[160,170)$	165	14	$-1,2$	1,44	20,16
$[170,180)$	175	15	8,8	77,44	1161,60
$[180,190)$	185	6	18,8	353,44	2120,64
Total		50			$\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot f_i = 7128$

Con lo que se tiene: $s^2 = \frac{7128}{50} = 142,56$ y $s = \sqrt{142,56} = 11,93985$.

OTRA FORMA: $\boxed{s^2 = \frac{\sum_{I=1}^{n} x_i^2 \cdot f_i}{N} - \overline{x}^2}$

Intervalo	x_i	x_i^2	f_i	$x_i^2 \cdot f_i$
$[140,150)$	145	21025	6	126150
$[150,160)$	155	24025	9	216225
$[160,170)$	165	27225	14	381150
$[170,180)$	175	30625	15	459375
$[180,190)$	185	34225	6	205350
		Total	N=50	1388250

Así pues, será: $s^2 = \frac{1388250}{50} - (166,2)^2 = 142,56$ y $s = \sqrt{142,56} = 11,93985$.

Ahora el coeficiente de variación es $\frac{s}{x} = \frac{11,93985}{166,2} = 0,07184$, que indica una dispersión baja (7%).