Tema 8: Estadística.

8.1 Conceptos estadísticos.

- La **Estadística** es una parte importante de las Matemáticas que estudia métodos para recopilar, analizar e interpretar datos con el fin de extraer conclusiones precisas y 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 que 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 es difícil o costoso observar a todos los elementos de la población estadística.
- **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:** son categóricos y no numéricos. Por ejemplo el color de los ojos, la profesión, la marca de coche, etc.
  - **Cuantitativos discretos:** aquellos que toman valores aislados (números naturales), y que no pueden tomar ningún valor intermedio con decimales. Por ejemplo, el número de goles, el número de hijos, el número de discos comprados, el número de pulsaciones, etc.
  - **Cuantitativos continuos:** son aquellos que se expresan con números decimales y pueden tomar, en teoría, todos los infinitos números reales de un rango o intervalo. Por ejemplo, la altura, el peso, el volumen, la presión sanguínea, la temperatura, etc. En estos casos, es aconsejable agrupar los datos de la variable estadística en intervalos de clase contiguos elegidos convenientemente para no perder mucha información. Los extremos de los intervalos de clase se denominan "extremos de clase" y el punto medio se llama "marca de la 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 al número total de datos.
- 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.

Evidentemente, los valores x_i deben estar ordenados de manera creciente y 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.

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:

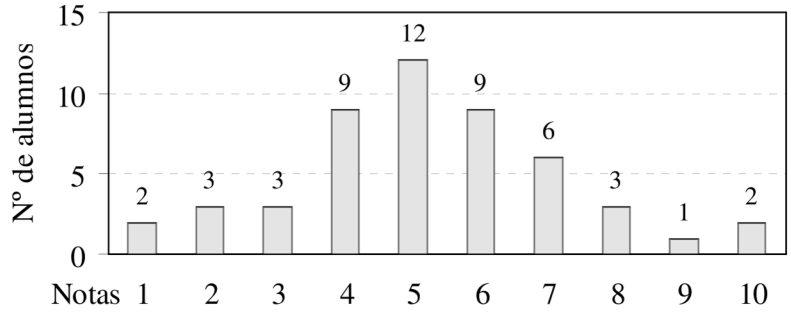
Estatura	[140,150)	[150,160)	[160,170)	[170,180)	[180,190)
Nº personas	6	9	14	15	6

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. Luego, los valores de x_i son: 145, 155, 165, 175 y 185.
La tabla de frecuencias es ahora la siguiente:

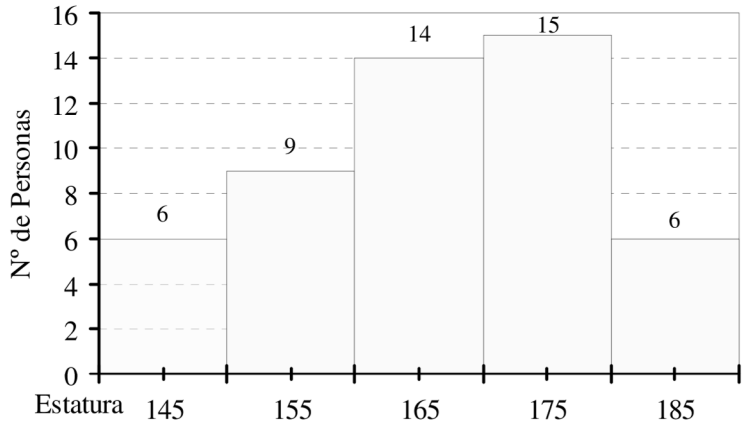
Intervalo	x_i	f_i	h_i	%	F_i	H_i	%Ac
[140,150)	145	6	0,12	12	6	0,12	12
[150,160)	155	9	0,18	18	15	0,30	30
[160,170)	165	14	0,28	28	29	0,58	58
[170,180)	175	15	0,30	30	44	0,88	88
[180,190)	185	6	0,12	12	**50**	**1**	**100**
	Total	**50**	**1**	**100**			

8.2 Gráficos.

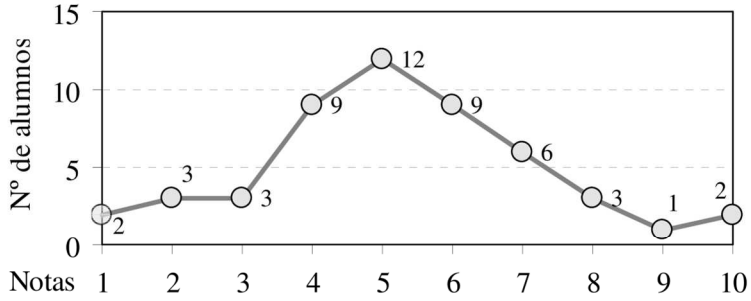
- **Diagrama de barras**: colocamos en el eje de abscisas los valores de la variable x_i y en el eje de ordenadas los valores de las frecuencias. Dibujamos barras de igual anchura cuya altura sea exactamente la frecuencia. Así, en el ejemplo de las notas, tenemos:



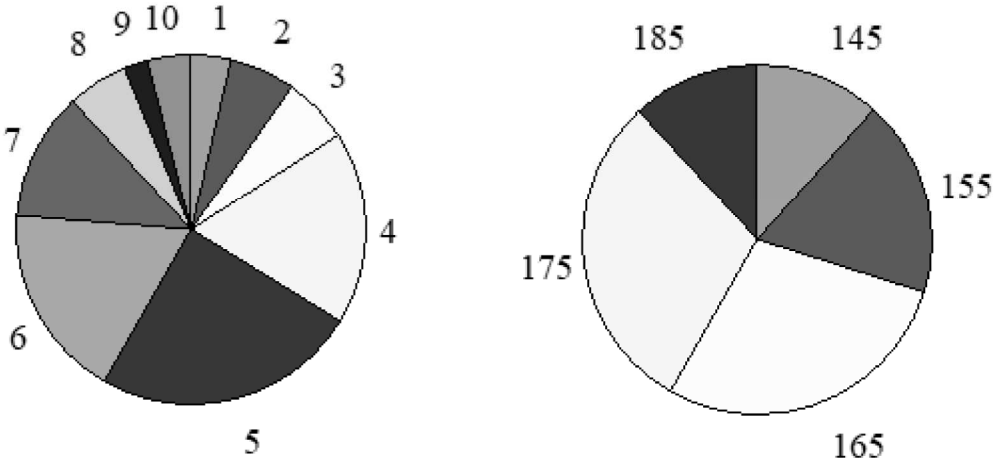
- **Histograma**: igual que el diagrama de barras, pero se utiliza en variables estadísticas continuas. Por lo tanto, las barras están unidas entre sí y cada una se sitúa sobre el intervalo de la clase. En el ejemplo de las estaturas, tenemos:



- **Polígono de frecuencias**: se unen los puntos medios de las bases superiores de las barras tanto del diagrama de barras como del histograma. De esta manera, se obtiene una representación gráfica de las frecuencias en función de los valores de la variable estadística.



- **Diagrama de sectores**: se obtiene dividiendo la circunferencia en tantas partes como valores tenga la variable, de manera que el área de cada sector sea proporcional a la respectiva frecuencia. Para ello, se obtiene el ángulo central que ocupará cada sector mediante una proporcionalidad directa. Por ejemplo, si a una frecuencia de 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.

8.3 Parámetros estadísticos.

La **media aritmética** se obtiene como la suma de todos los valores de una variable, dividido por el número total de dichos valores. 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: $$\overline{x} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{N}$$

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 es pues $\overline{x} = \frac{262}{50} = 5{,}24$

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 es pues $\overline{x} = \frac{8310}{50} = 166{,}2$

- La **moda** es el valor que tiene mayor frecuencia absoluta. En el ejemplo de las notas, la moda es $M_0 = 5$, ya que a 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 $[170,180)$, donde se encuentra la mayor frecuencia, 15. Sin embargo, si deseamos 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** es el valor M_e 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 ellos, la mediana es el que ocupa el lugar central. Si su número es par, se toma 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 $M_e = 5$.

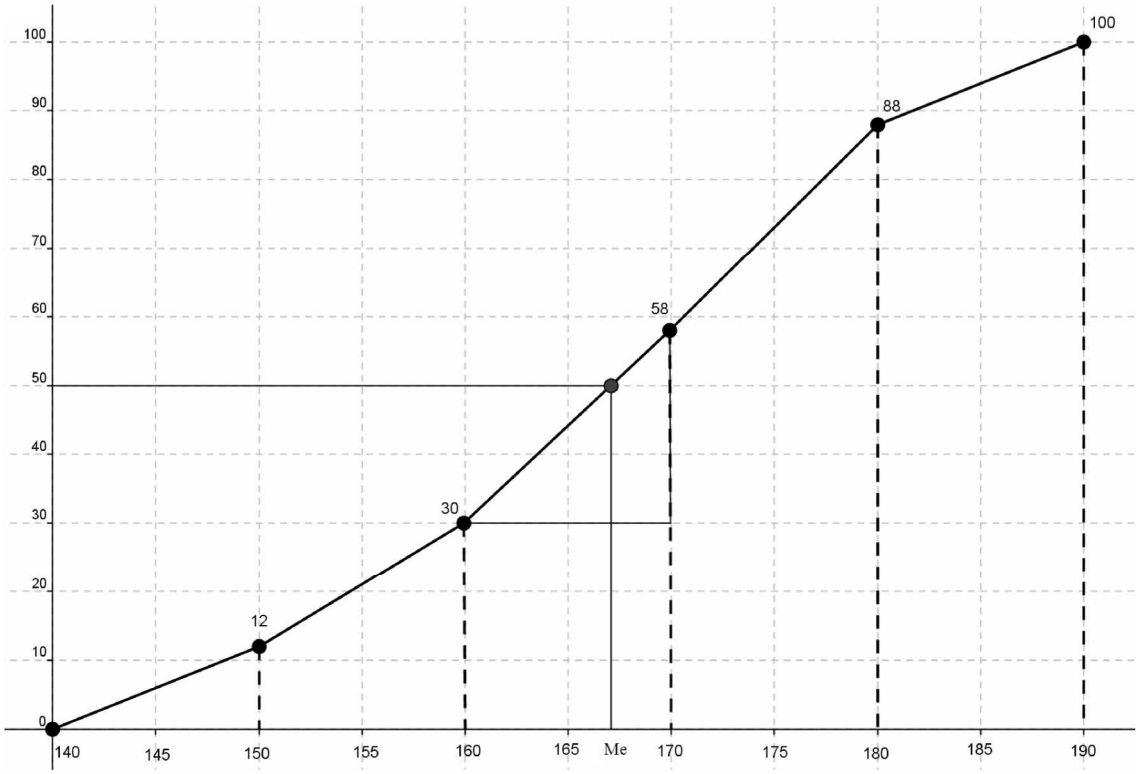
  Para calcular la mediana en una variable continua, como en el ejemplo de las estaturas, está claro el 50% se supera en el intervalo mediano $[160,170)$. Pero para calcular el valor exacto de la mediana, debemos seguir el siguiente proceso gráfico:

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

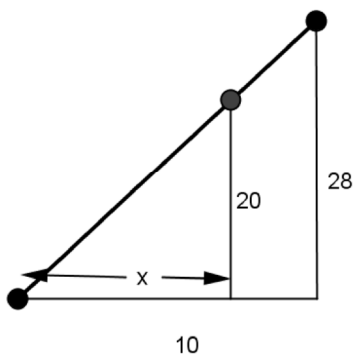
Intervalo	x_i	f_i	h_i	%	F_i	H_i	**% Ac**
$[140,150)$	145	6	0,12	12	6	0,12	**12**
$[150,160)$	155	9	0,18	18	15	0,30	**30**
$[160,170)$	165	14	0,28	28	29	0,58	**58**
$[170,180)$	175	15	0,30	30	44	0,88	**88**
$[180,190)$	185	6	0,12	12	50	1	**100**
	Total	50	1	100			

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

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.

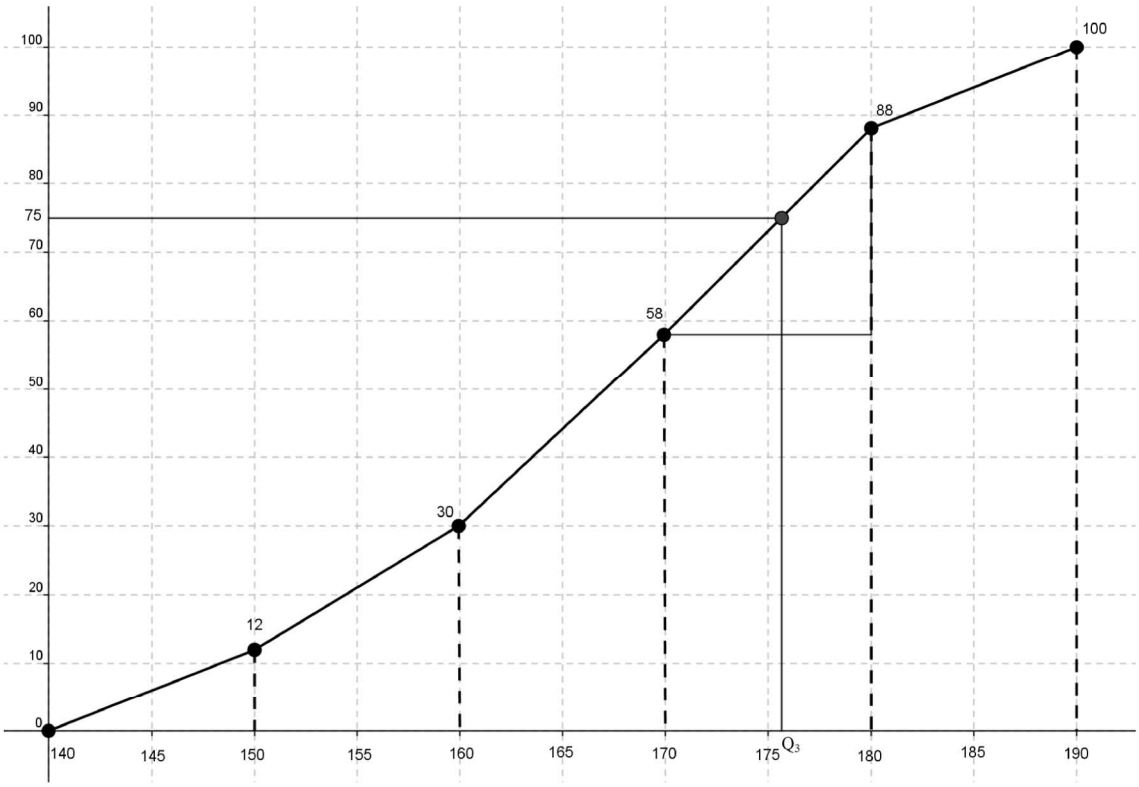
  Ejemplo: para calcular el tercer cuartil (Q_3) en el ejemplo de las notas, nos fijamos en el porcentaje de frecuencias acumuladas y buscamos el dato en el que se supera el valor 75%. En este caso, sucede para el 6, por lo que $Q_3 = 6$.

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			

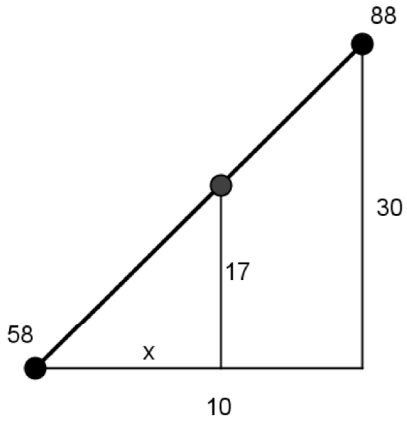
Para calcular un cuartil en una variable continua, debemos seguir un proceso gráfico análogo al estudiado con la mediana.

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 acumuladas. A partir de estos datos, construimos una tabla de valores que darán lugar a los puntos del polígono de frecuencias acumuladas.

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.



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

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.

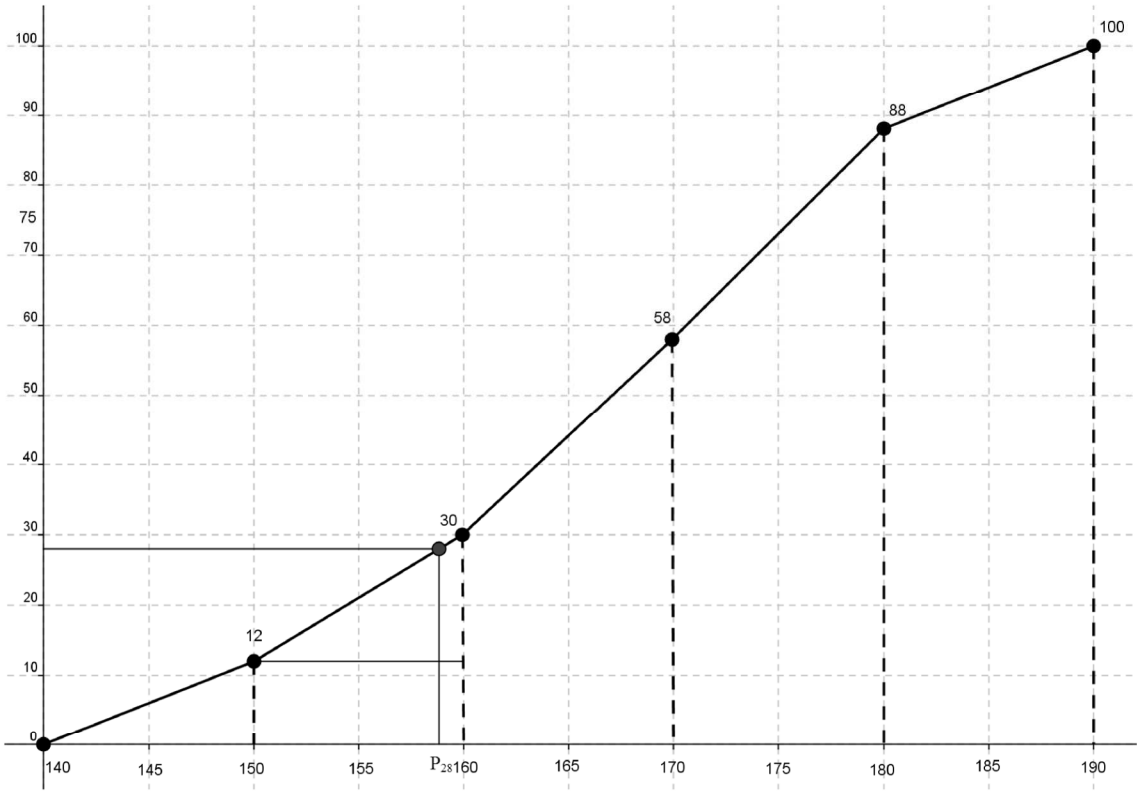
Ejemplo: para calcular el percentil P_{28} en el ejemplo de las notas, nos fijamos en el porcentaje de frecuencias acumuladas y buscamos el dato en el que se supera el valor 28%, que sucede para el 4. Por tanto, $P_{28} = 4$.

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			

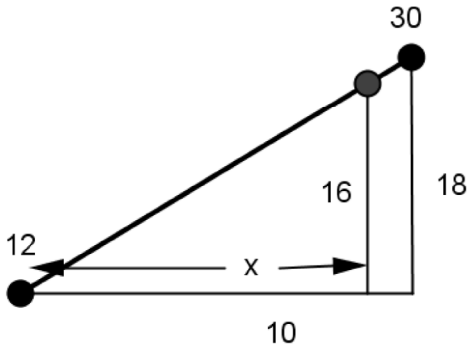
Para calcular un percentil en una variable continua, debemos seguir un proceso gráfico análogo al estudiado con la mediana y los cuartiles.
Ejemplo: calcular el percentil P_{28} en el ejemplo de las estaturas.

Igual que en la mediana y en los cuartiles, utilizamos la tabla de frecuencias, fijándonos en los porcentajes de frecuencias acumuladas, y a partir de estos datos construimos una tabla de valores que darán lugar a los puntos del polígono de frecuencias acumuladas:

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.



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

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:

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

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 de la distribución de su valor medio, es decir, mayor es el error que se comete al sustituirlos todos por su media aritmética.

Para el ejemplo de las notas de matemáticas, calcularíamos la desviación típica de la siguiente manera:

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

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**, cuanto mayor es, más dispersos están los datos. En nuestro ejemplo, el coeficiente de variación es $\frac{2,0155}{5,24} = 0,3846$, que indica una dispersión elevada (38%).

  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%).

8.4 Conceptos de Estadística bidimensional.

- Variables estadísticas bidimensionales: ahora se trata de estudiar un fenómeno en el que se consideran dos variables X e Y, en lugar de una sola, como hasta ahora. Ejemplo: pulso y temperatura de los pacientes de un hospital, ingresos y gastos de las familias de un colectivo, edad y número de días que faltan al trabajo los trabajadores de una fábrica, etc.
- Tipos de distribuciones bidimensionales:
  - Cualitativa – cualitativa.
  - Cualitativa – cuantitativa (discreta o continua).
  - Cuantitativa (discreta o continua) – cuantitativa (discreta o continua).
- Tipos de tablas:
  - Tabla de dos columnas (x_i , y_i) para pocos datos.
  - Tabla de tres columnas (x_i , y_i , f_i) para muchos datos y pocos valores posibles.
  - Tablas de doble entrada para muchos datos y muchos valores posibles. Ejemplo: las notas de Matemáticas y Física de 20 alumnos:

Notas Mat.	Notas Fís.	Frecuencia
1	2	2
1	3	1
2	3	1
3	2	1
3	5	1
4	3	1
5	1	1
5	2	1
6	1	1
6	2	1
6	5	2
7	6	1
7	7	2
8	2	1
9	8	1
10	9	2
	Total	20

Otro ejemplo: se ha medido el volumen, en litros, y el peso, en kilogramos, de distintos tipos de maletas, obteniéndose los siguientes resultados:

Volumen	97	102	94	107	92	98
Peso	6,9	7,1	6,7	7,4	5,8	6,1

8.5 Cálculo de parámetros bidimensionales.

- Ya conocemos los parámetros unidimensionales para cada variable:

$$\boxed{\bar{x} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{N}} \qquad \boxed{s_x^2 = \frac{\sum_{i=1}^{n} f_i \cdot (x_i - \bar{x})^2}{N} = \frac{\sum_{i=1}^{n} x_i^2 \cdot f_i}{N} - \bar{x}^2}$$

$$\boxed{\bar{y} = \frac{\sum_{i=1}^{n} y_i \cdot f_i}{N}} \qquad \boxed{s_y^2 = \frac{\sum_{i=1}^{n} f_i \cdot (y_i - \bar{y})^2}{N} = \frac{\sum_{i=1}^{n} y_i^2 \cdot f_i}{N} - \bar{y}^2}$$

- Ahora aparece un parámetro nuevo: la **covarianza**, que es la media aritmética de las desviaciones de cada una de las variables respecto a sus medias respectivas. $\boxed{s_{xy} = \frac{\sum_{i=1}^{n} f_i \cdot (y_i - \bar{y})(x_i - \bar{x})}{N} = \frac{\sum_{i=1}^{n} f_i \cdot x_i \cdot y_i}{N} - \bar{x} \cdot \bar{y}}$
- El **coeficiente de correlación lineal** (coef. de Pearson = r) es una forma de cuantificar de forma más precisa el tipo de correlación que hay entre las dos variables: $\boxed{r = \frac{s_{xy}}{s_x s_y}}$

8.6 Correlación o dependencia.

- Es la teoría que trata de estudiar la relación o dependencia entre las dos variables que intervienen en una distribución bidimensional, según sean los diagramas de dispersión, también llamados nube de puntos, podemos establecer los siguientes casos:
  - Independencia funcional o correlación nula: cuando no existe ninguna relación entre las variables ($r = 0$).
  - Dependencia funcional o correlación funcional: cuando existe una función tal que todos los valores de la variable la satisfacen (a cada valor de x le corresponde uno solo de y o a la inversa) ($r = \pm 1$).
  - Dependencia aleatoria o correlación curvilínea (ó lineal): cuando los puntos del diagrama se ajustan a una línea recta o a una curva, puede ser positiva o directa, o negativa o inversa ($0 < r < 1$ ó $-1 < r < 0$)

8.7 Regresión lineal.

- Consiste en ajustar lo más posible la nube de puntos de un diagrama de dispersión a una curva. Cuando ésta es una recta, se obtiene la recta de regresión lineal. Cuando es una parábola se obtiene una regresión parabólica. Y cuando es una exponencial, se obtiene una regresión exponencial. (Es importante tener en cuenta que en todos los casos, r debe ser distinto de 0).
- Al valor $\boxed{\frac{s_{xy}}{s_x^2}}$ se le llama **coeficiente de regresión de y sobre x,** y nos proporciona la pendiente de la recta de regresión.

La **recta de regresión de y sobre x** es entonces: $\boxed{y-\overline{y}=\frac{s_{xy}}{s_x^2}(x-\overline{x})}$.

- Análogamente, el valor $\boxed{\frac{s_{xy}}{s_y^2}}$ se llama **coeficiente de regresión de x sobre y**.

  Y la **recta de regresión de x sobre y** es entonces: $\boxed{x-\overline{x}=\frac{s_{xy}}{s_y^2}(y-\overline{y})}$.

- En el ejemplo de las notas de Matemáticas y Física.

Notas Mat	Notas Fís	Frecuencia	$x_i \cdot f_i$	$y_i \cdot f_i$	$x_i^2 \cdot f_i$	$y_i^2 \cdot f_i$	$x_i \cdot y_i \cdot f_i$
1	2	2	2	4	2	8	4
1	3	1	1	3	1	9	3
2	3	1	2	3	4	9	6
3	2	1	3	2	9	4	6
3	5	1	3	5	9	25	15
4	3	1	4	3	16	9	12
5	1	1	5	1	25	1	5
5	2	1	5	2	25	4	10
6	1	1	6	1	36	1	6
6	2	1	6	2	36	4	12
6	5	2	12	10	72	50	60
7	6	1	7	6	49	36	42
7	7	2	14	14	98	98	98
8	2	1	8	2	64	4	16
9	8	1	9	8	81	64	72
10	9	2	20	18	200	162	180
	Total	20	107	84	727	488	547

Media x	5,35	media y	4,2
desviación s_x	2,7798381	desviación s_y	2,6
covarianza s_{xy}	4,88	coef. corr. lin. r	0,6751915
Coeficiente de Regresión de y sobre x	0,6315108	Coeficiente de Reg. x sobre y	0,7218935

Las medias son: $\overline{x}=\frac{\sum_{i=1}^{n} x_i \cdot f_i}{N}=\frac{107}{20}=5{,}35$ e $\overline{y}=\frac{\sum_{i=1}^{n} y_i \cdot f_i}{N}=\frac{84}{20}=4{,}2$.

Las desviaciones típicas son las siguientes:

$$s_x=\sqrt{\frac{\sum_{i=1}^{n} f_i \cdot (x_i-\overline{x})^2}{N}}=\sqrt{\frac{\sum_{i=1}^{n} f_i \cdot x_i^2}{N}-\overline{x}^2}=\sqrt{\frac{727}{20}-5{,}35^2}=\sqrt{7{,}7275}=2{,}7798381.$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} f_i \cdot (y_i - \overline{y})^2}{N}} = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 \cdot f_i}{N} - \overline{y}^2} = \sqrt{\frac{488}{20} - 4{,}2^2} = \sqrt{6{,}76} = 2{,}6\,.$$

La covarianza es:

$$s_{xy} = \frac{\sum_{i=1}^{n} f_i \cdot (y_i - \overline{y})(x_i - \overline{x})}{N} = \frac{\sum_{i=1}^{n} f_i \cdot x_i \cdot y_i}{N} - \overline{x} \cdot \overline{y} = \frac{547}{20} - 5{,}35 \cdot 4{,}2 = 4{,}88\,.$$

El coeficiente de correlación lineal es: $r = \frac{s_{xy}}{s_x s_y} = \frac{4{,}88}{2{,}779838 \cdot 2{,}6} = 0{,}6751915\,.$

Que indica en este ejemplo que la correlación entre los datos es fuerte.

Los coeficientes de regresión, son los siguientes:

Coeficiente de regresión de y sobre x: $\frac{s_{xy}}{s_x^2} = \frac{4{,}88}{(2{,}7798381)^2} = 0{,}6315108\,.$

Coeficiente de regresión de x sobre y: $\frac{s_{xy}}{s_y^2} = \frac{4{,}88}{(2{,}6)^2} = 0{,}7218935\,.$

Recta de regresión de y sobre x : $y - 4{,}2 = 0{,}6315108 \cdot (x - 5{,}35)$

Recta de regresión de x sobre y: $x - 5{,}35 = 0{,}7218935 \cdot (y - 4{,}2)$

Por ejemplo, si un alumno tiene un 7 en Matemáticas, ¿qué nota se espera que obtenga en Física?

$y - 4{,}2 = 0{,}6315108 \cdot (7 - 5{,}35) \Rightarrow y = 4{,}2 + 0{,}6315108 \cdot 1{,}65 = 4{,}2 + 1{,}04199282$

Luego, se espera que obtenga una nota aproximada de 5,24 en Física.

- En el ejemplo de las maletas: se ha medido el volumen, en litros, y el peso, en kilogramos, de distintos tipos de maletas, obteniendo los resultados que se recogen en esta tabla:

Volumen	97	102	94	107	92	98
Peso	6,9	7,1	6,7	7,4	5,8	6,1

Vol.	Peso	Frecuencia	$x_i \cdot f_i$	$y_i \cdot f_i$	$x_i^2 \cdot f_i$	$y_i^2 \cdot f_i$	$x_i \cdot y_i \cdot f_i$
97	6,9	1	97	6,9	9409	47,61	669,3
102	7,1	1	102	7,1	10404	50,41	724,2
94	6,7	1	94	6,7	8836	44,89	629,8
107	7,4	1	107	7,4	11449	54,76	791,8
92	5,8	1	92	5,8	8464	33,64	533,6
98	6,1	1	98	6,1	9604	37,21	597,8
	Total	6	590	40	58166	268,52	3946,5

media x	98,333333	media y	6,666666
desviación s_x	4,988876	desviación s_y	0,555777
covarianza	2,194444	coef. corr. lin.	0,791445
coef. Regres. de y sobre x	0,088159	coef. Regresión de x sobre y	7,104316

Las medias son: $\overline{x} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{N} = \frac{590}{6} = 98{,}333333$ $\overline{y} = \frac{\sum_{i=1}^{n} y_i \cdot f_i}{N} = \frac{40}{6} = 6{,}666666$.

Las desviaciones típicas son las siguientes:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} f_i \cdot (x_i - \overline{x})^2}{N}} = \sqrt{\frac{\sum_{i=1}^{n} f_i \cdot x_i^2}{N} - \overline{x}^2} = \sqrt{\frac{58166}{6} - 98{,}333333^2} = 4{,}988876.$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} f_i \cdot (y_i - \overline{y})^2}{N}} = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 \cdot f_i}{N} - \overline{y}^2} == \sqrt{\frac{268{,}52}{6} - 6{,}666666^2} = 0{,}555777.$$

La covarianza es:

$$s_{xy} = \frac{\sum_{i=1}^{n} f_i \cdot (y_i - \overline{y})(x_i - \overline{x})}{N} = \frac{\sum_{i=1}^{n} f_i \cdot x_i \cdot y_i}{N} - \overline{x} \cdot \overline{y} = \frac{3946{,}5}{6} - 98{,}333333 \cdot 6{,}666666 = 2{,}194444.$$

El coeficiente de correlación lineal es:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{2{,}194444}{4{,}988876 \cdot 0{,}555777} = 0{,}791445.$$

Este coeficiente indica que la correlación entre los datos es muy fuerte.

Los coeficientes de regresión, son los siguientes:

Coeficiente de y sobre x: $\frac{s_{xy}}{s_x^2} = \frac{2{,}194444}{(4{,}988876)^2} = 0{,}08816$.

Coeficiente de x sobre y: $\frac{s_{xy}}{s_y^2} = \frac{2{,}194444}{(0{,}555777)^2} = 7{,}104316$.

Recta de regresión de y sobre x : $y - 6{,}666666 = 0{,}08816 \cdot (x - 98{,}333333)$

Recta de regresión de x sobre y: $x - 98{,}333333 = 7{,}104316 \cdot (y - 6{,}666666)$

Por ejemplo, si una maleta tiene un volumen de 120 litros, ¿qué peso se espera que tenga?

$y - 6{,}666666 = 0{,}08816 \cdot (120 - 98{,}333333)$

$y - 6{,}666666 = 0{,}08816 \cdot 21{,}666667$

$y - 6{,}666666 = 1{,}91034226$

$y = 6{,}66666666 + 1{,}91034226$

$y = 8{,}57700892$

Luego, se espera que la maleta pese 8577 gramos.