1.- Terminología estadística

Estadística descriptiva

Es la ciencia que estudia conjuntos de datos obtenidos de la realidad.
Estos datos son interpretados mediante tablas, gráficas y otros parámetros como la media, moda, varianza, etc.

Población

Es el conjunto formado por todos los elementos que queremos estudiar.
Por ejemplo, si vamos a estudiar el peso de los jóvenes de 16 años nacidos en España, la población sería precisamente el conjunto formado dichos jóvenes

Variable estadística

Es la característica que queremos estudiar de la población.
Hay distintos tipos de variables estadísticas

<table>
<tr><td colspan="2">Cualitativa
Si los valores son cualidades. Por ejemplo, partido político preferido, color del pelo, etc.</td></tr>
<tr><td rowspan="2">Cuantitativa
Si los valores son números. Por ejemplo, nº de hermanos, estatura, peso, edad, temperatura, etc.</td><td>Discreta
Cuando los valores son aislados.
Por ejemplo, nº de hermanos, edad, etc.</td></tr>
<tr><td>Continua
Cuando entre dos valores, aunque estén muy próximos entre sí, siempre es posible tomar otro valor.
Por ejemplo, la temperatura, el peso, etc.</td></tr>
</table>

2.- Tablas de frecuencias

Los datos obtenidos en estadística se organizan en unas tablas, llamadas tablas de frecuencias.

Tabla de frecuencias para datos aislados
Ejemplo: Edades de un grupo de alumnos de alumnos

x_i	f_i	F_i	h_i	H_i
13	6	6	30%	30%
14	5	11	25%	55%
15	7	18	35%	90%
16	1	19	5%	95%
18	1	20	5%	100%
Suma total	20 = n	-	100%	-

Tabla de frecuencias para datos agrupados
Ejemplo: Notas en un examen de un grupo

Clases	f_i	F_i	h_i	H_i
[2,3)	3	3	15%	15%
[3,4)	2	5	10%	25%
[4,5)	3	8	15%	40%
[5,6)	5	13	25%	65%
[6,7)	3	16	15%	80%
[7,8)	4	20	20%	100%
Total	20 = n	-	100%	-

x_i representa los valores que hay en los datos. En el caso de datos agrupados, las clases son los intervalos

f_i se llama frecuencia absoluta y representa las veces que aparece cada valor en los datos
En el caso de datos agrupados, fi representa el nº de datos que hay en el intervalo o clase

F_i es la frecuencia absoluta acumulada y se calcula sumando uno a uno los valores de la columna f_i.

h_i se llama frecuencia relativa y se calcula dividiendo cada valor fi entre el nº total de datos y se expresa en %

H_i es la frecuencia relativa acumulada y se calcula sumando uno a uno los valores de la columna h_i.

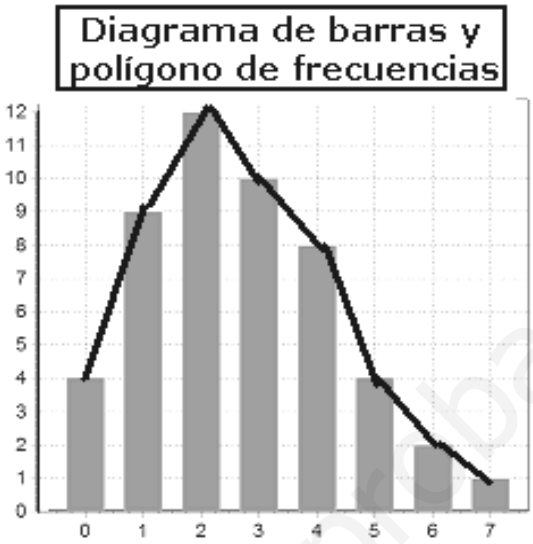
3.- Gráficos estadísticos

Diagrama de barras

Se representan los valores x_i en un eje horizontal y para cada valor x_i se dibuja una barra cuya altura sea la frecuencia de x_i que se quiera representar. Las barras deben ser de la misma anchura y debemos dibujarlas separadas.
Uniendo los extremos superiores de las barras por su punto medio, se obtiene una línea quebrada llamada **polígono de frecuencias**

Ejemplo: Número de hijos de un grupo de matrimonios

x_i	f_i
0	4
1	9
2	12
3	10
4	8
5	4
6	2
7	1
Total	50 = n



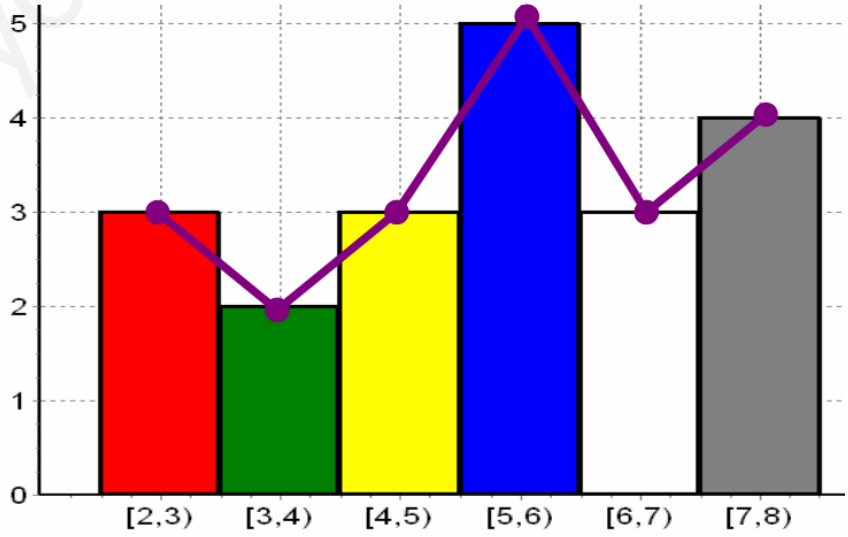
El diagrama de barras se suele utilizar para variables discretas con "pocos" valores y para variables cualitativas

Histograma

Es similar al diagrama de barras, sólo que la base de cada barra es el intervalo de la tabla de frecuencias y por tanto no hay espacios entre las barras.

Ejemplo:
Notas de 20 alumnos en un examen:

clases	f_i
$2 \leq x < 3$	3
$3 \leq x < 4$	2
$4 \leq x < 5$	3
$5 \leq x < 6$	5
$6 \leq x < 7$	3
$7 \leq x < 8$	4
Total	20 = n



Uniendo los extremos superiores de las barras por su punto medio, se obtiene la línea quebrada llamada **polígono de frecuencias**.

Los histogramas se utilizan cuando los datos los agrupamos en intervalos

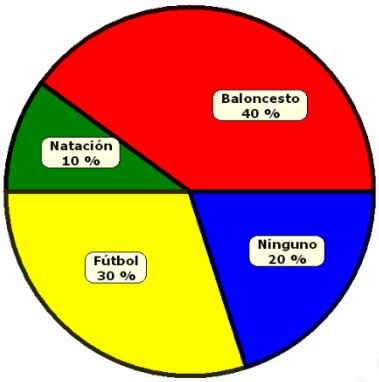
Diagrama de sectores

Para dibujar el diagrama de sectores se dibuja un círculo y se divide en tantos sectores (quesitos) como valores haya en los datos.

Ejemplo:

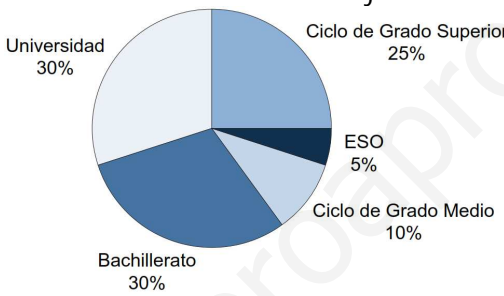
Deporte preferido por un grupo de 30 alumnos

Deporte	f_i	h_i (en %)	Ángulo del sector
Baloncesto	12	40	40% de 360º = 144º
Natación	3	10	10% de 360º = 36º
Fútbol	9	30	30% de 360º = 108º
Ninguno	6	20	20% de 360º = 72º
Total	30	100%	360º



El diagrama de sectores se suele utilizar para variables discretas con "pocos" valores y para variables cualitativas

Ejercicio 1 A los 200 alumnos y alumnas de 2º y 3º de E.S.O. de un Instituto les preguntamos sobre el nivel máximo de estudios que esperan realizar. El resultado es el reflejado en el siguiente gráfico de sectores:



Pasa esta información a una tabla de frecuencias y a un diagrama de barras

4.- Parámetros estadísticos

La media aritmética

Es la suma de todos los datos dividida entre el número total de datos, n.

Se calcula por la fórmula $\bar{x} = \frac{\sum(x_i\ f_i)}{n}$, donde $\sum$ significa suma.

Ejemplo:

Notas en un examen de un grupo de amigos

x_i	f_i	$x_i\ f_i$
4	1	4
5	2	10
6	4	24
7	3	21
Total	10 = n	59

$$\bar{x} = \frac{\sum(x_i\ f_i)}{n} = \frac{59}{10} = 5{,}9$$

Si los datos están agrupados en intervalos, se toma como x_i el punto medio del intervalo.

Este valor se llama **marca de clase**

Ejemplo:

Gasto mensual en €, en teléfono móvil, de un grupo de jóvenes

clases	x_i	f_i	$x_i\ f_i$
[10, 11)	10,5	4	42
[11, 12)	11,5	6	69
[12, 13)	12,5	7	87,5
[13, 14)	13,5	3	40,5
Total		20 = n	239

$$\bar{x} = \frac{\sum(x_i\ f_i)}{n} = \frac{239}{20} = 11{,}95\ €$$

La media ponderada

Se calcula cuando los datos tienen distinto peso o importancia

Ejemplo:

Tres exámenes tienen distinto peso: el primero vale 1, el segundo 2, y el tercero 3.
Un alumno obtiene calificaciones de 9, 4 y 8, respectivamente.
¿Qué nota le debe poner el profesor?

Se multiplica cada nota por su peso y se divide entre la suma de los pesos.

Media ponderada: $\frac{9.1+4.2+8.3}{1+2+3}=\frac{41}{6}=6{,}8$. Luego, le debe poner un 6,8

La moda (Mo)

Es el valor que más se repite en los datos. La moda es el valor xi que tiene mayor frecuencia absoluta.

Si los datos están agrupados en intervalos se toma el intervalo de mayor frecuencia (intervalo o clase modal).

Puede haber más de una moda o puede que no haya moda porque todos los valores tengan la misma frecuencia absoluta.

Ejemplo:

x_i = Equipo de fútbol preferido	Nº de personas
Madrid	12
Granada	7
Barcelona	12
Málaga	6

Hay dos modas, Madrid y Barcelona.

La mediana (Me)

Es el dato que está justamente en medio, cuando tenemos todos los datos ordenados de menor a mayor

Cálculo de la mediana cuando hay "pocos" datos

- Si el **nº de datos es impar** , la mediana es el dato central

Ejemplo:

Edades de 9 personas: 15 , 12 , 17 , 15 , 14 , 14 , 17 , 15 , 15
Ordenando los datos: 12, 14, 14, 15, 15, 15, 15, 17, 17 → Me = 15

- Si el **nº de datos es par** , la mediana es la media aritmética de los 2 datos centrales

Ejemplo:

Notas de 12 alumnos: 7 , 4 , 6 , 5 , 7 , 7 , 8 , 5 , 8 , 4 , 4 , 5
Ordenando los datos: 4, 4, 4, 5, 5, 5, 6, 7, 7, 7, 8, 8 → Me = 5,5

Cálculo de la mediana cuando hay "muchos" datos

En este caso, la mediana es el primer valor x_i cuya H_i es mayor que el **50%**

Ejemplo:

Notas en Inglés de 20 alumnos:

clases	x_i	H_i
[2,3)	2,5	15
[3,4)	3,5	25
[4,5)	4,5	40
[5,6)	5,5	**65**
[6,7)	6,5	80
[7,8)	7,5	100

Me = 5,5

Los cuartiles

Cuando los datos están ordenados de menor a mayor, los cuartiles son tres valores Q_1 , Q_2 , Q_3 que dividen a los datos en 4 partes iguales

El primer cuartil, Q_1 , es el primer valor x_i cuya H_i es mayor que el 25%

El segundo cuartil, Q_2 , es el primer valor x_i cuya H_i es mayor que el 50%. Es decir, Q_2 = Me

El tercer cuartil, Q_3 , es el primer valor x_i cuya H_i es mayor que el 75%

El ***rango intercuartílico*** (RI) es la distancia entre Q_1 y Q_3 → RI: $Q_3 - Q_1$

Si los datos estuviesen agrupados en intervalos se toma como x_i la marca de clase

Ejemplo:

Notas en Inglés de 20 alumnos:

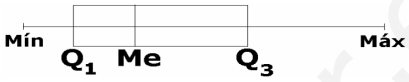
clases	x_i	H_i
[2,3)	2,5	15
[3,4)	3,5	25
[4,5)	4,5	40
[5,6)	5,5	65
[6,7)	6,5	80
[7,8)	7,5	100

$Q_1 = 4,5$ $Q_2 = Me = 5,5$ $Q_3 = 6,5$ $RI = 6,5 - 4,5 = 2$

Diagrama de caja

Los cuartiles se suelen representar en un diagrama, llamado diagrama de caja

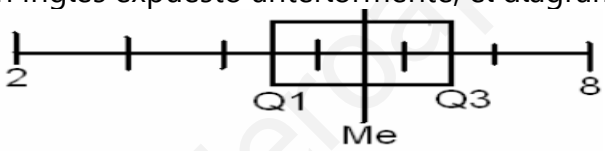
Para dibujar el diagrama de caja, se calculan los valores mínimo y máximo de x_i así como los cuartiles. Después se dibuja una caja, cuyos extremos son Q_1 y Q_3, que indica donde se concentran el 50% de los datos y una línea central que marca la mediana.



Esta representación nos permite saber qué datos son atípicos y si la distribución de datos es simétrica respecto de la mediana: si la caja está desplazada hacía la izquierda o hacía la derecha respecto de la mediana significa que la distribución de datos es asimétrica.

Ejemplo:

Para las notas en Inglés expuesto anteriormente, el diagrama de caja sería:



Los percentiles

Cuando los datos están ordenados de menor a mayor, los percentiles son 99 valores P_1, P_2, ... P_{99} que dividen a los datos en 100 partes iguales.

Por ejemplo, P_1 es el primer valor x_i cuya H_i supera el 1%, P_2 es el primer valor x_i cuya H_i supera el 2%, etc.

Ejemplos:

P_{25} es el primer valor x_i cuya H_i supera el 25%. Luego $P_{25} = Q_1$

P_{50} es el primer valor x_i cuya H_i supera el 50%. $P_{50} = Q_2 = Me$

P_{75} es el primer valor x_i cuya H_i supera el 75%. Luego $P_{75} = Q_3$

Los deciles

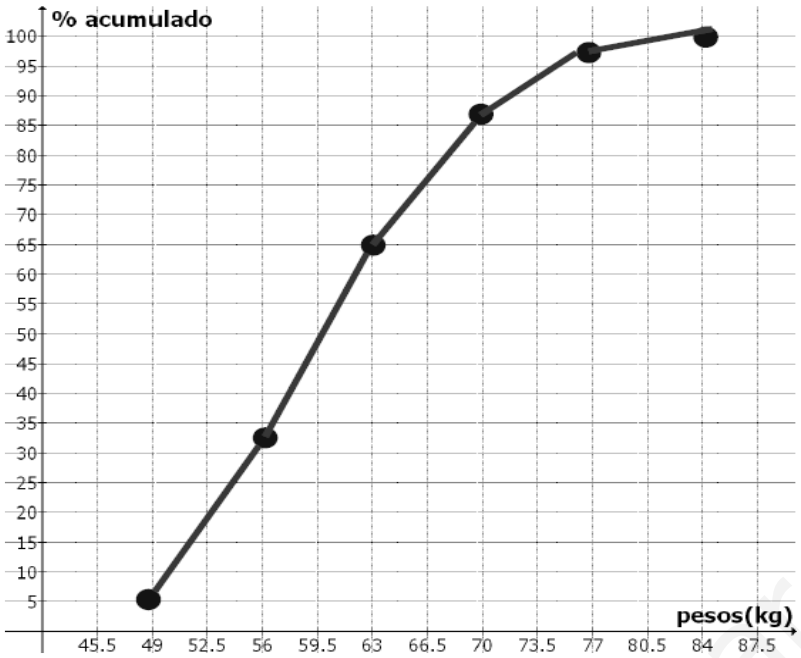
Cuando los datos están ordenados de menor a mayor, los deciles son 9 valores D_1, D_2, ... D_9 que dividen a los datos en 10 partes iguales.

Por ejemplo, D_1 es el primer valor x_i cuya H_i supera el 10%, D_2 es el primer valor x_i cuya H_i supera el 20%, etc.

Ejemplo:

D_5 es el primer valor x_i cuya H_i supera el 50%. Luego $D_5 = Q_2 = Me$

Ejercicio 2 El siguiente gráfico representa el polígono de frecuencias relativas acumuladas, H_i, de los pesos de 40 personas



Determina a partir del gráfico, de forma aproximada: a) La mediana b) P_{20}

c) El percentil que corresponde a un peso de 63 kg d) El número de personas que pesan más de 66,5 kg

Parámetros de dispersión

El *recorrido* o rango, R, es la diferencia entre el mayor y el menor valor de x_i

La *varianza*, s^2, se calcula con la fórmula: $\boxed{s^2 = \frac{\sum x_i^2 \cdot f_i}{n} - \bar{x}^2}$

La *desviación típica*, s, es la raíz cuadrada de la varianza: $\boxed{s = \sqrt{s^2}}$

El *coeficiente de variación*, CV, se calcula con la fórmula: $\boxed{CV = \frac{s}{\bar{x}}}$

En el intervalo $(\bar{x} - s\ ,\ \bar{x} + s)$ se encuentra, aproximadamente el 68% de todos los datos

En el intervalo $(\bar{x} - 2s\ ,\ \bar{x} + 2s)$ se encuentra, aproximadamente el 95% de todos los datos

En el intervalo $(\bar{x} - 3s\ ,\ \bar{x} + 3s)$ se encuentra, aproximadamente el 99% de todos los datos

Ejemplo:

Se han agrupado los hoteles de una región por el número de habitaciones, obteniéndose la siguiente tabla:

Habitaciones	número de hoteles
0 – 100	20
100 – 200	50
200 – 300	60
300 – 400	30
400 – 500	40

clases	x_i	f_i	$x_i f_i$	$x_i^2 f_i$
[0,100)	50	20	1 000	50 000
[100,200)	150	50	7 500	1 125 000
[200,300)	250	60	15 000	3 750 000
[300,400)	350	30	10 500	3 675 000
[400,500)	450	40	18 000	8 100 000
Total		n = 200	52 000	16 700 000

Rango: 500 – 0 = 500 $\qquad \bar{x} = \frac{52\,000}{200} = 260$

$s^2 = \frac{16\,700\,000}{200} - 260^2 = 15\,900 \qquad s = \sqrt{15\,900} \cong 126{,}0952 \qquad C.V. = \frac{126{,}0952}{260} = 0{,}485$

Ejercicio 3 Se realiza una estadística en dos centros de enseñanza, uno público y otro privado, referente a la nota global del bachillerato de cada uno de los alumnos que van a acudir a los exámenes de selectividad.

Las distribuciones de frecuencias son las siguientes:

Público	
Nota	Alumnos
[5,6)	60
[6,7)	70
[7,8)	40
[8,9)	20
[9,10)	10

Privado	
Nota	Alumnos
5,5	8
6,5	12
7,5	20
8,5	30
9,5	10

a) Un alumno del centro privado tiene una nota global de un 8,5 y otro del centro público una nota de un 7. ¿Cuál de los dos es mejor alumno comparándolo con la media de su Centro?

b) Calcula el coeficiente de variación de las dos distribuciones.

5.- Distribuciones bidimensionales

Concepto de distribución bidimensional

Cuando se quieren estudiar dos características X e Y de una misma población, los datos que se obtienen son parejas de valores (x_i, y_i). El conjunto de datos (x_i,y_i) se llama *distribución bidimensional*

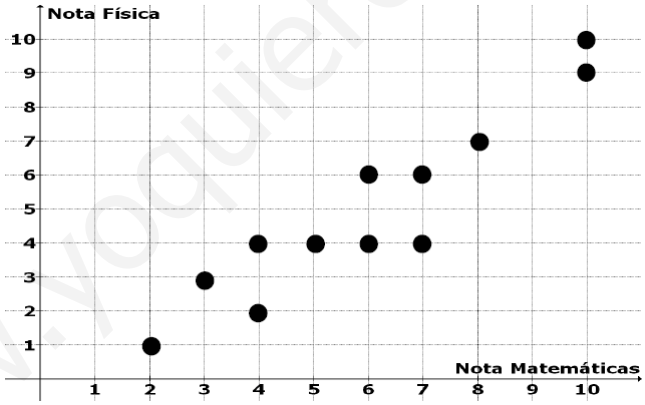
Diagrama de dispersión o nube de puntos

Es la representación gráfica de los puntos (x_i, y_i)

Ejemplo:

Notas de 12 alumnos en Matemáticas y Física

Alumno	a	b	c	d	e	f	g	h	i	j	k	l
Matemáticas	2	3	4	4	5	6	6	7	7	8	10	10
Física	1	3	2	4	4	4	6	4	6	7	9	10

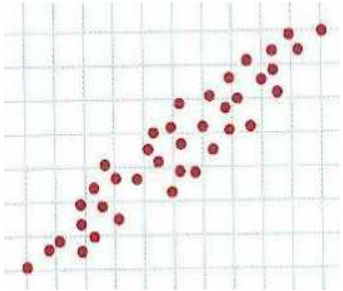


Interpretación del diagrama de dispersión: Correlación

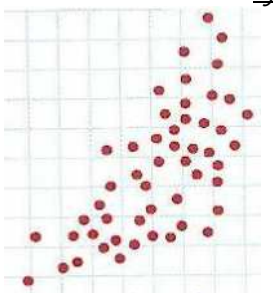
Si la nube de puntos se concentra en torno a una línea se dice que hay correlación entre las dos variables.

Se dice que hay *correlación lineal* si la nube de puntos se concentra en torno a una recta.
La correlación será *positiva o directa* si la línea es creciente y *negativa o inversa* si es decreciente y será más *fuerte* cuanto mayor sea la concentración de los puntos entorno a esa línea.

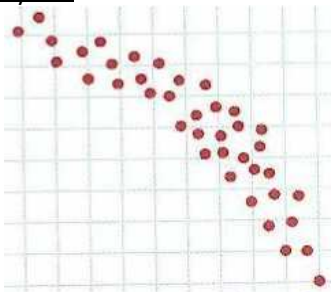
Ejemplos:



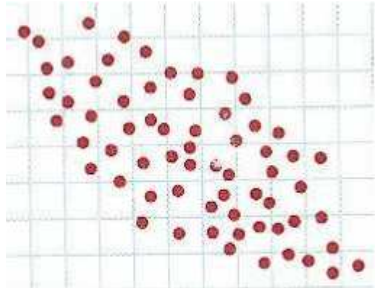
Correlación lineal directa fuerte



Correlación directa débil

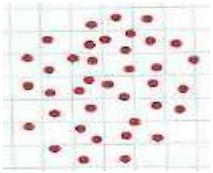


Correlación inversa fuerte



Correlación lineal inversa débil

Si los puntos están esparcidos sin concentrarse en torno a ninguna línea, se dice que no hay relación entre las variables o que la correlación es nula.



Correlación nula

6.- Parámetros estadísticos bidimensionales

Medias aritméticas de X y de Y

$$\bar{x} = \frac{\sum x_i . f_i}{n} \qquad \bar{y} = \frac{\sum y_i . f_i}{n}$$

Centro de gravedad de la distribución: $(\bar{x}, \bar{y})$

Varianza y desviación típica de X y de Y

X: varianza: $s_x^2 = \frac{\sum x_i^2 . f_i}{n} - \bar{x}^2$; Desviación típica: $s_x = \sqrt{s_x^2}$

Y: varianza: $s_y^2 = \frac{\sum y_i^2 . f_i}{n} - \bar{y}^2$; Desviación típica: $s_y = \sqrt{s_y^2}$

Covarianza entre X e Y

$$s_{xy} = \frac{\sum x_i y_i f_i}{n} - \bar{x}\,\bar{y}$$

También se puede usar la fórmula $s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$

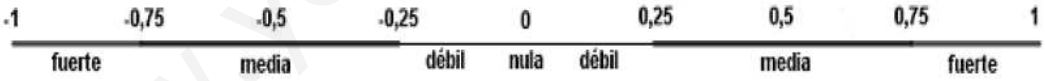
Coeficiente de correlación lineal de Pearson

$$r = \frac{s_{xy}}{s_x\, s_y}$$

Propiedades del coeficiente de correlación

1) El coeficiente de correlación, r, tiene el mismo signo que la covarianza y nos sirve para medir el grado de relación o dependencia entre las variables X e Y
2) $-1 \le r \le 1$
3) Si r = 1 ó r = –1, la nube de puntos se ajusta perfectamente a una recta
4) Si r es positivo, la correlación es positiva y si r es negativo, la correlación es negativa
5) Cuanto más próximo esté r al 0 más débil es la correlación
6) Cuánto más próximo esté | r | al 1, más fuerte es la correlación

Interpretación de la correlación según el valor del coeficiente de correlación



Tablas de doble entrada

Cuando las datos $(x_i y_i)$ se repiten se suele utilizar una tabla de doble entrada para evitar escribir la misma pareja varias veces.

Ejemplo:

A un grupo de padres se les ha preguntado por el número de hijos que tienen y el número de horas que ven diariamente la televisión. Los resultados se han recogido en la siguiente tabla de doble entrada:

X = número de hijos, Y = número de horas que ven la televisión.

X / Y	0	1	2
0	2	1	0
1	3	4	1
2	0	5	3

Por ejemplo, la pareja de valores (1,1) aparece 4 veces lo que significa que hay 4 padres que tienen 1 hijo y ven la televisión 1 hora

Ejercicio 4 En los siguientes casos halla las medidas bidimensionales e indica el tipo de correlación que hay entre X e Y.

a) Las tallas y los pesos de 10 personas vienen recogidos en la siguiente tabla:

X = talla (en metros)	1,60	1,65	1,70	1,80	1,85	1,90	1,92	1,75	1,82	1,72
Y = peso (en kg)	58	61	65	73	80	85	83	68	74	67

b) Se han recogido una serie de datos y se ha hecho la siguiente tabla de doble entrada

X / Y	2	4	6
1	1	3	0
2	2	3	1

7.- Rectas de regresión

Son las rectas que mejor se ajustan a la nube de puntos de forma que la nube de puntos está muy concentrada en torno a ellas. Hay dos rectas de regresión:

Recta de regresión de Y sobre X

$$r_{yx}: \; y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

$\frac{s_{xy}}{s_x^2}$ es la pendiente de la recta y se llama *coeficiente de regresión de Y sobre X*

La recta de regresión de Y sobre X se puede usar para estimar lo que vale "y" para un valor dado de "x".

La estimación es más fiable cuanto más fuerte sea la correlación entre las variables y más cerca esté el valor "x" de los valores "x_i" de la distribución.

Recta de regresión de X sobre Y

$$r_{xy}: \; x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y})$$

$\frac{s_{xy}}{s_y^2}$ se llama *coeficiente de regresión de X sobre Y*

La recta de regresión de X sobre Y se puede usar para estimar lo que vale "x" para un valor dado de "y".

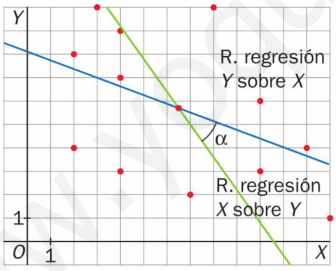
La estimación es más fiable cuanto más fuerte sea la correlación entre las variables y más cerca esté el valor "y" de los valores "y_i" de la distribución.

Propiedades de las rectas de regresión

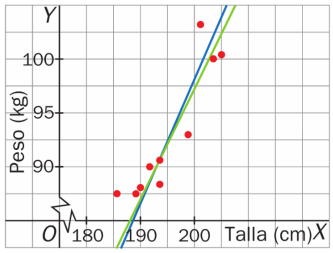
1) Las dos rectas se cortan en el centro de gravedad $(\bar{x}, \bar{y})$

2) Cuánto más fuerte es la correlación menor es el ángulo que forman entre sí ambas rectas.

Ejemplos:



Correlación débil



Correlación fuerte

Ejercicio 5 En distintos modelos de aspiradores se ha medido el peso, en kilogramos, y la capacidad útil de la bolsa, en litros, obteniendo los siguientes resultados:

***X:* Peso**	6,1	7	5,8	5,4	7	6,4
***Y:* Capacidad**	1,9	4,3	1,5	1,7	2,9	3,2

a) Halla la recta de regresión de Y sobre X y de X sobre Y.

b) Usando la recta que corresponda haz las siguientes estimaciones e indica si son fiables:

b1) La capacidad para un peso de 6,5 kg b2) El peso para una capacidad de 2 litros

b3) La capacidad para un peso de 10 kg

Ejercicio 6 La media de las estaturas X de los habitantes de una ciudad es 170 cm y la media de sus pesos Y es 65 kg. Las desviaciones típicas son 10 cm y 5 kg y la covarianza de ambas variables es 40.

a) Halla el coeficiente de correlación

b) Calcula la recta de regresión de Y respecto de X y de X respecto de Y

c) Estima el peso de un individuo de 180 cm de estatura y la estatura de un individuo de 60 kg de peso. ¿Será buena la estimación? Razónalo

d) Si quisiéramos estimar el peso de un niño de 50 cm mediante la recta de regresión ¿sería buena la predicción?