ÍNDICE

1. VARIABLES BIDIMENSIONALES

En el análisis estadístico es muy frecuente analizar o investigar las relaciones entre varias variables. Si se limita el estudio al caso de dos variables, se habla entonces de variables bidimensionales.

La variable estadística bidimensional se representa por el símbolo (X, Y) y cada uno de los individuos de la población por (x_i, y_i).

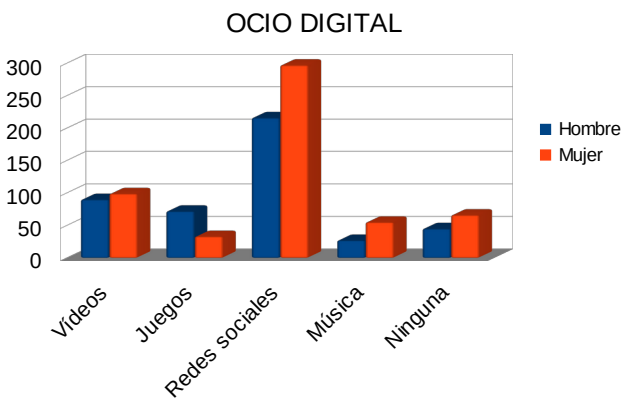
Ejemplo 1: El tiempo, en minutos, hasta ser atendido en una oficina depende del número de clientes que haya en la sala de espera. Durante un día se registraron estos datos:

Nº de clientes (x_i)	2	3	5	6	9	4	2	1
Tiempo de espera en minutos (y_i)	4	5	10	14	20	10	5	2

Podemos calcular la media de los clientes, $\bar{x}$ = (2+3+5+6+9+4+2+1) / 8 = 32 / 8 = 4 clientes, y la media del tiempo de espera, $\bar{y}$ = (4+5+10+14+20+10+5+2) / 8 = 70 / 8 = 8,75 minutos.
Pero también nos interesa saber la relación entre el número de clientes y el tiempo de espera.

Ejemplo 2: Un estudio realizado a 1000 jóvenes de entre 15 y 29 años sobre su preferencia diaria en ocio digital ha arrojado los siguientes resultados, separados por sexo:

Sexo / Ocio digital	**Hombre**	**Mujer**
Vídeos	90	99
Juegos	72	33
Redes sociales	216	297
Música	27	55
Ninguna	45	66

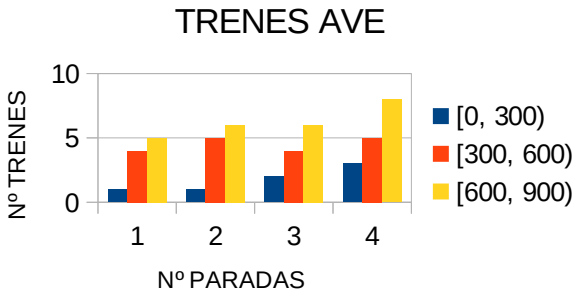


De los 450 hombres del estudio 216 prefieren las redes sociales, un 48%. Mientras que entre las mujeres son 297 de 550, un 54%. En ambos casos es la moda. Son dos variables cualitativas.

Ejemplo 3: Los datos obtenidos para 50 trenes AVE sobre el número de paradas intermedias, x, y la distancia recorrida por trayecto, y, en kilómetros están recogidos en la tabla:

	Distancia (km)		
Nº paradas	**[0, 300)**	**[300, 600)**	**[600, 900)**
1	1	4	5
2	1	5	6
3	2	4	6
4	3	5	8

La variable, x, número de paradas, es discreta y la variable y, distancia por trayecto (en km), es continua. Podemos comprobar que hay 8 trenes que realizan un trayecto entre 600 y 900 km con 4 paradas, el 16%.



1.1. Distribución conjunta

En general los datos para variables bidimensionales se pueden presentar en dos formatos en los que figuran las frecuencias absolutas:

Tabla bidimensional simple

Variable X x_i	Variable Y y_i	f_i
x_1	y_1	f_1
x_2	y_2	f_2
...	...	...
x_n	y_n	f_n

En los casos más simples $f_i = 1$

Tabla bidimensional de doble entrada

X \ Y	y_1	y_2	...	y_m
x_1	f_{11}	f_{12}		f_{1m}
x_2	f_{21}	f_{22}		f_{2m}
...	...	...		...
x_n	f_{n1}	f_{n2}		f_{nm}

$\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij} = N$ N = total de datos

También hay tablas bidimensionales con frecuencias relativas $h_{ij} = \frac{f_{ij}}{N}$, donde $\sum_{i=1}^{n}\sum_{j=1}^{m} h_{ij} = 1$.

En el ejemplo 2 sobre ocio digital podemos calcular las frecuencias relativas dividiendo todas las frecuencias absolutas entre 1000:

Sexo / Ocio digital	Hombre	Mujer
Vídeos	0,090	0,099
Juegos	0,072	0,033
Redes sociales	0,216	0,297
Música	0,027	0,055
Ninguna	0,045	0,066

1.2. Distribución condicionada

En algunas ocasiones a partir de una distribución bidimensional, resulta interesante analizar cómo se comporta una de las variables para ciertos valores de la otra. Para expresar la variable X condicionada a un valor y_i de la variable Y, escribiremos $(X/ Y = y_i)$.

En el ejemplo 2 sobre ocio digital podemos estudiar los datos condicionados:

Ocio digital (X \ Y = "Mujer")	Mujer
Vídeos	99
Juegos	33
Redes sociales	297
Música	55
Ninguna	66

Sexo (Y \ X = "Redes sociales"	Hombre	Mujer
Redes sociales	216	297

1.3. Distribución marginal

La suma de las frecuencias absolutas por filas es la frecuencia absoluta de la variable X y la suma por columna la de la variable Y.

X \ Y	y_1	y_2	...	y_m	Frecuencia absoluta de la variable X
x_1	f_{11}	f_{12}	...	f_{1m}	$\sum_{j=1}^{m} f_{1j}$
x_2	f_{21}	f_{22}	...	f_{2m}	$\sum_{j=1}^{m} f_{2j}$
...	...	...	...	...	...
x_n	f_{n1}	f_{n2}	...	f_{nm}	$\sum_{j=1}^{m} f_{nj}$
Frecuencia absoluta de la variable Y	$\sum_{i=1}^{n} f_{i1}$	$\sum_{i=1}^{n} f_{i2}$	...	$\sum_{i=1}^{n} f_{im}$	$\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij} = N$

En el ejemplo 2 sobre ocio digital obtenemos la distribución de las variables X e Y:

Y=Sexo X=Ocio digital	Hombre	Mujer	
Vídeos	90	99	189
Juegos	72	33	105
Redes sociales	216	297	513
Música	27	55	82
Ninguna	45	66	111
	450	550	1000

X=Ocio digital	f_i
Vídeos	189
Juegos	105
Redes sociales	513
Música	82
Ninguna	111
	1000

Y=Sexo	f_i
Hombre	450
Mujer	550
	1000

Ahora podemos calcular los parámetros estadísticos de cada variable por separado, en este caso la moda al ser variables cualitativas. La moda de X es "Redes digitales" y la moda de Y es "Mujer".

En el ejemplo 3 de los trenes AVE obtenemos las distribuciones marginales:

X \ Y	Distancia (km)			
Nº paradas	**[0, 300)**	**[300, 600)**	**[600, 900)**	
1	1	4	5	10
2	1	5	6	12
3	2	4	6	12
4	3	5	8	16
	7	18	25	50

X	f_i
1	10
2	12
3	12
4	16
	50

Y	f_i
[0, 300)	7
[300, 600)	18
[600, 900)	25
	50

Podemos calcular por separado los parámetros estadísticos de X (variable cuantitativa discreta) e Y (variable cuantitativa continua):

$X=x_i$	f_i	x_if_i	$x_i^2f_i$
1	10	10	10
2	12	24	48
3	12	36	108
4	16	64	256
	50	134	422

Y	y_i	f_i	x_if_i	$x_i^2f_i$
[0, 300)	**150**	7	1050	157500
[300, 600)	**450**	18	8100	3645000
[600, 900)	**750**	25	18750	14062500
		50	27900	17865000

Parámetros variable X: $\bar{x}=\frac{134}{50}=2{,}68$ $\sigma^2_x=\frac{422}{50}-2{,}68^2=1{,}2576$ $\sigma_x=\sqrt{1{,}2576}=1{,}1214$

Parámetros variable Y: $\bar{y}=\frac{27900}{50}=558$

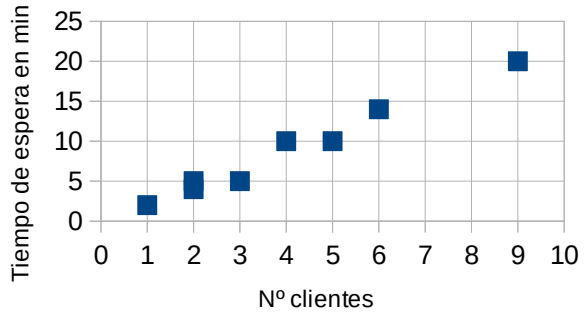
$\sigma^2_y=\frac{17865000}{50}-558^2=45936$ $\sigma_y=\sqrt{45936}=214{,}3269$

2. DIAGRAMA DE DISPERSIÓN O NUBE DE PUNTOS

Una forma sencilla y habitual de representar gráficamente una distribución de dos variables cuantitativas es el **diagrama de dispersión** o **nube de puntos**, que proporciona una buena descripción de la relación entre dos variables.

En el ejemplo 1 sobre el tiempo de espera de unos clientes teníamos los datos:

Nº de clientes (x_i)	2	3	5	6	9	4	2	1
Tiempo de espera en minutos (y_i)	4	5	10	14	20	10	5	2



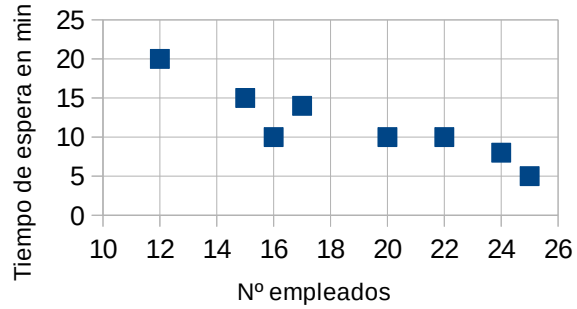
Representados en un diagrama de dispersión podemos observar la relación entre las dos variables:

A mayor número de clientes, mayor tiempo de espera, hay una **relación positiva** entre ambas variables

Otros ejemplos de relación entre dos variables son los siguientes:

Ejemplo 4:

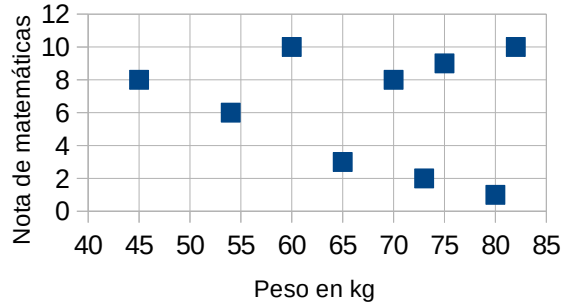
Nº de empleados (x_i)	12	15	16	17	20	22	24	25
Tiempo de espera en minutos (y_i)	20	15	10	14	10	10	8	5



A mayor número de empleados, menor tiempo de espera, hay una **relación negativa** entre ambas variables

Ejemplo 5:

Peso en kg estudiantes (x_i)	45	54	60	65	70	73	75	80	82
Nota de matemáticas (y_i)	8	6	10	3	8	2	9	1	10

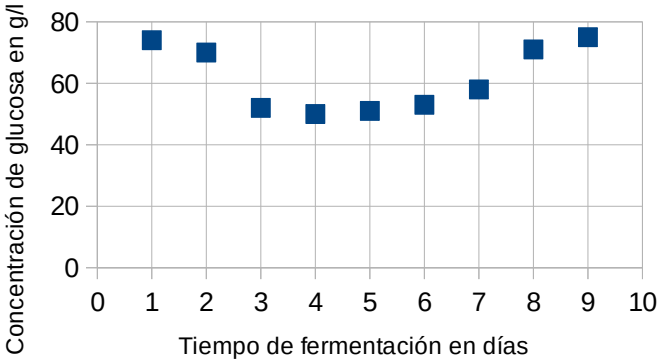


Como era de esperar no se observa **ninguna relación** entre las dos variables. Las dos variables son **independientes**.

Es decir, conocer el peso de un estudiante no nos proporciona información sobre la nota de matemáticas.

Ejemplo 6: Estudiamos la concentración de glucosa en g/l de un determinado licor según el tiempo de fermentación:

Tiempo de fermentación en días (x_i)	1	2	3	4	5	6	7	8	9
Concentración de glucosa en g/l (y_i)	74	70	52	50	51	53	58	71	75



Se observa una relación entre ambas variables, pero en este caso no es una relación lineal como en los dos primeros diagramas.

La nube de puntos se podría aproximar por una función cuadrática, una parábola.

Nuestro objetivo en este tema será estudiar la **relación lineal entre dos variables**. Para ello definiremos dos parámetros estadísticos, **la covarianza y la correlación lineal**, que midan el grado de relación lineal de dos variables.

3. DEPENDENCIA O CORRELACIÓN LINEAL.

El diagrama de puntos ofrece una aproximación al estudio de la dependencia lineal de dos variables, sin embargo necesitamos cuantificar esta relación lineal.

3.1. Covarianza

La covarianza fue introducida por Karl Pearson para medir la relación lineal de dos variables, X e Y que toman N valores (x_i, y_i) con una frecuencia f_i:

$$\sigma_{xy}=\frac{\sum_{i=1}^{N}(x_i-\bar{x})\cdot(y_i-\bar{y})f_i}{N}=\frac{\sum_{i=1}^{N}x_i\cdot y_i\cdot f_i}{N}-\bar{x}\cdot\bar{y}$$

El **signo de la covarianza** nos va a indicar si la **relación entre las variables es positiva o negativa**.

Comprobamos con en el ejemplo 1 que la covarianza es positiva y en el ejemplo 4 que es negativa. Fíjate que en estos ejemplos la frecuencia de cada par de datos (x_i, y_i) es 1:

Ejemplo 1:

Nº de clientes (x_i)	Tiempo de espera en minutos (y_i)	x_iy_i
2	4	8
3	5	15
5	10	50
6	14	84
9	20	180
4	10	40
2	5	10
1	2	5
$\sum_{i=1}^{N}x_i=32$	$\sum_{i=1}^{N}y_i=70$	$\sum_{i=1}^{N}x_i\cdot y_i=389$

$$\bar{x}=\frac{\sum_{i=1}^{N}x_i}{N}=\frac{32}{8}=4 \text{ clientes}$$

$$\bar{y}=\frac{\sum_{i=1}^{N}y_i}{N}=\frac{70}{8}=8{,}75 \text{ minutos}$$

$$\sigma_{xy}=\frac{\sum_{i=1}^{N}x_i\cdot y_i}{N}-\bar{x}\cdot\bar{y}=\frac{389}{8}-4\cdot 8{,}75=13{,}625$$

Ejemplo 4:

Nº de empleados (x_i)	Tiempo de espera en minutos (y_i)	x_iy_i
12	20	240
15	15	225
16	10	160
17	14	238
20	10	200
22	10	220
24	8	192
25	5	125
$\sum_{i=1}^{N}x_i=151$	$\sum_{i=1}^{N}y_i=92$	$\sum_{i=1}^{N}x_i\cdot y_i=1600$

$$\bar{x}=\frac{\sum_{i=1}^{N}x_i}{N}=\frac{151}{8}=18{,}875 \text{ empleados}$$

$$\bar{y}=\frac{\sum_{i=1}^{N}y_i}{N}=\frac{92}{8}=11{,}5 \text{ minutos}$$

$$\sigma_{xy}=\frac{\sum_{i=1}^{N}x_i\cdot y_i}{N}-\bar{x}\cdot\bar{y}=$$

$$\frac{1600}{8}-18{,}875\cdot 11{,}5=-17{,}0625$$

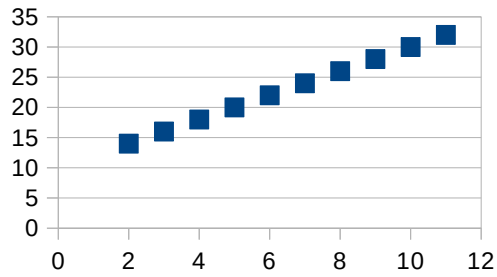
3.2. Correlación lineal

Con el fin de evitar la dependencia de las unidades se define el coeficiente de correlación lineal:

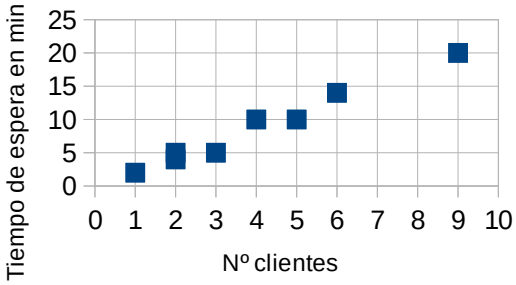
$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

Las propiedades más importantes del coeficiente de correlación lineal son:

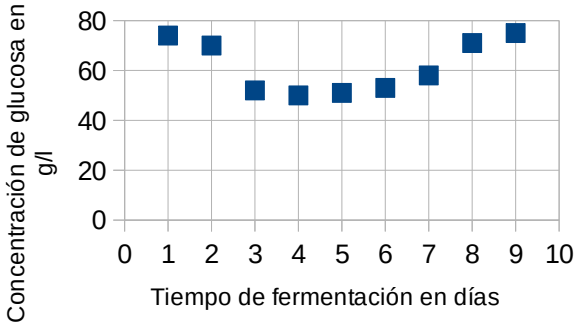
1. Tiene el mismo signo que la covarianza, ya que las desviaciones típicas son siempre positivas.
2. Su valor es independiente de las unidades de las variables.
3. Su valor está entre 1 y -1, **$-1 \le r \le 1$**.
4. Si r = -1 o r = -1, los valores de las variables X e Y están en una línea recta. Cuanto más alejado está el valor de r de estos valores y más próximo a 0 menos relación lineal hay entre las variables.



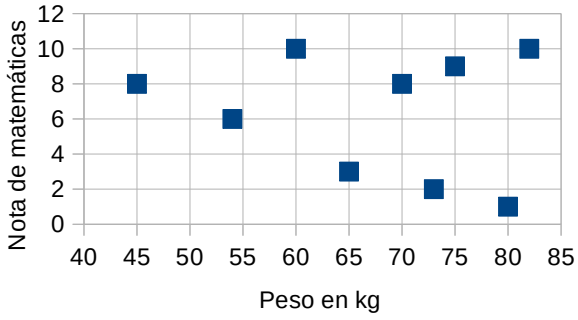
r = 1 Relación funcional (pendiente positiva)



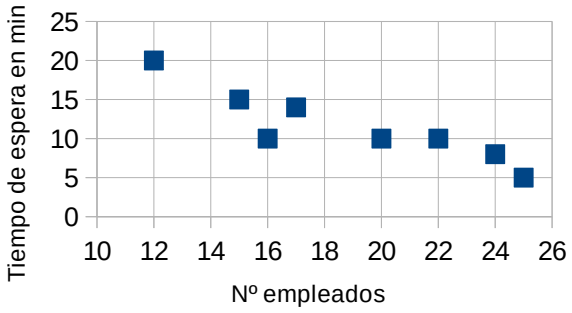
r = 0,9881 Relación lineal positiva muy fuerte



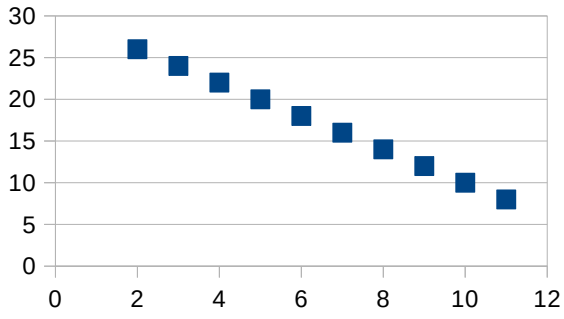
r = 0,0937 Ausencia de relación lineal



r = -0,1898 Relación lineal negativa muy débil



r = -0,9074 Relación lineal negativa muy fuerte

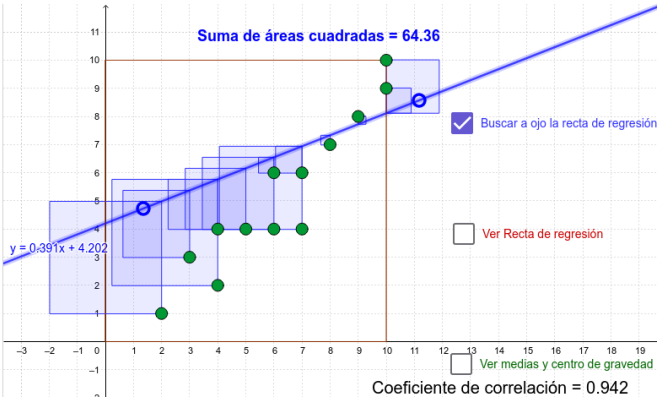


r = -1 Relación funcional (pendiente negativa)

	$-0,5 \le r \le 0,5$	Correlación débil
$-0,75 \le r \le -0,5$	o $0,5 \le r \le 0,75$	Correlación moderada
$-1 \le r \le -0,75$	o $0,75 \le r \le 1$	Correlación fuerte

4. RECTAS DE REGRESIÓN

En caso de relación lineal entre las variables X e Y, calcularemos la recta y = mx + n que aproxime los datos (x_i, y_i) de tal manera que la diferencia de los datos reales yi con los estimados $\hat{y}_i$ sea la mínima posible:



Minimizar $\sum_{i=1}^{N}(y_i-\hat{y}_i)^2\cdot f_i$

Es el llamado **método de mínimos cuadrados**.

La solución es la recta de pendiente

$m=\frac{\sigma_{xy}}{\sigma^2_x}$ que pasa por $(\bar{x},\bar{y})$

Recta de regresión de la variable Y sobre la variable X: $y-\bar{y}=\frac{\sigma_{xy}}{\sigma_x^2}(x-\bar{x})$

De forma análoga si minimizamos la diferencia entre los datos x_i y los estimados $\hat{x}_i$ obtenemos:

Recta de regresión de la variable X sobre la variable Y: $x-\bar{x}=\frac{\sigma_{xy}}{\sigma_y^2}(y-\bar{y})$

Características de las rectas de regresión:

1) Se cortan en el punto de las medias $(\bar{x},\bar{y})$.

2) Al multiplicar la pendiente de la primera recta por la inversa de la pendiente de la segunda se obtiene r^2:

$$\frac{\sigma_{xy}}{\sigma^2_x}\cdot\frac{\sigma_{xy}}{\sigma^2_y}=\left(\frac{\sigma_{xy}}{\sigma_x\cdot\sigma_y}\right)^2=r^2$$

Ejemplo 1: Calculamos el coeficiente de correlación y las rectas de regresión en este ejemplo.

Nº de clientes (x_i)	Tiempo de espera en minutos (y_i)	x_i^2	y_i^2	x_iy_i
2	4	4	16	8
3	5	9	25	15
5	10	25	100	50
6	14	36	196	84
9	20	81	400	180
4	10	16	100	40
2	5	4	25	10
1	2	1	4	2
$\sum_{i=1}^{N}x_i=32$	$\sum_{i=1}^{N}y_i=70$	$\sum_{i=1}^{N}x_i^2=176$	$\sum_{i=1}^{N}y_i^2=866$	$\sum_{i=1}^{N}x_i\cdot y_i=389$

$\bar{x}=4$ clientes

$\bar{y}=8{,}75$ min

$\sigma_{xy}=13{,}625$

$\sigma_x^2=6$

$\sigma_x=2{,}4495$

$\sigma_y^2=31{,}6875$

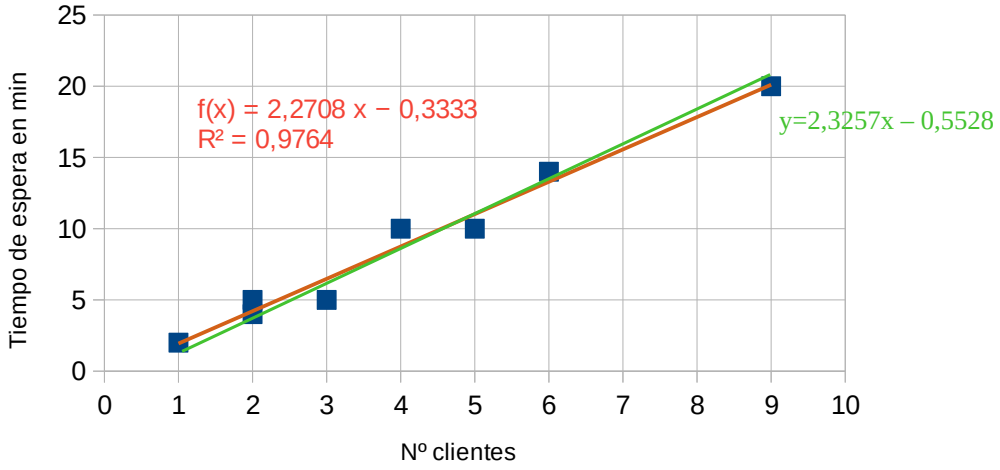
$\sigma_y=5{,}6292$

Coeficiente de correlación: $r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{13{,}625}{2{,}4495 \cdot 5{,}6292} = 0{,}9881$

Recta de regresión de Y sobre X: $y - 8{,}75 = \frac{13{,}625}{6}(x - 4)$ –> $\boxed{y = 2{,}2708x - 0{,}3333}$

Recta de regresión de X sobre Y: $x - 4 = \frac{13{,}625}{31{,}6875}(y - 8{,}75)$ –> $\boxed{y = 2{,}3257x - 0{,}5528}$

Representando las dos rectas sobre el diagrama de dispersión:



Las dos rectas pasan por el punto de las medias (4, 8,75), centro de gravedad de la nube de puntos.

4.1. Estimaciones

El objetivo principal de calcular la recta de regresión es hacer estimaciones o predicciones, que serán más fiables cuanto más se acerque el coeficiente de correlación, r, a 1 o -1.

Ejemplo 1: podemos hacer estimaciones puesto que r es muy próximo a 1. Podemos estimar el tiempo de espera en caso de haya 7 clientes:

x = 7 clientes –> $\hat{y}$ (30) = 2,2708·7 – 0,3333 = 15,5623 minutos de espera

Si queremos estimar el número de clientes para un tiempo de espera de 12 minutos:

y = 12 minutos –> $\hat{x}$ (12) = 0,43·(12 – 8,75) +4 = 5,3975 –> aproximadamente 5 clientes

Sin embargo, estas predicciones o estimaciones serán válidas únicamente en el rango de valores de las variables, X e Y, o en sus proximidades.

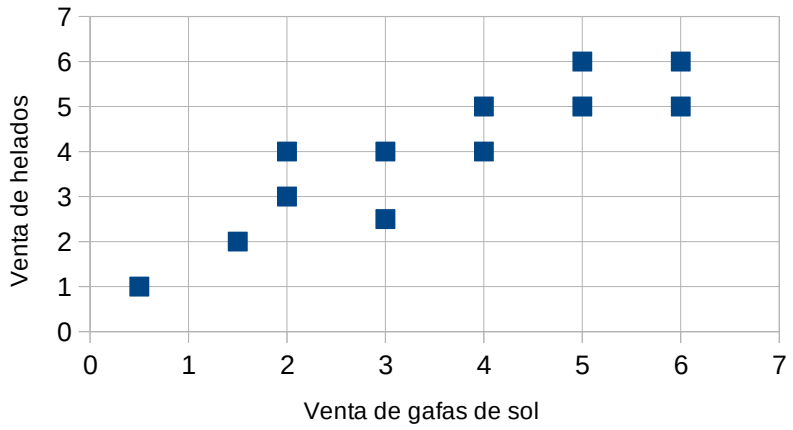
4.2. Correlación y causalidad

Conviene tener cuidado a la hora de interpretar el coeficiente de correlación y el diagrama de dispersión en términos de causa y efecto entre ambas variables. Véase el siguiente ejemplo:

Ejemplo 7: Se estudia la venta de gafas de sol (X) y la venta de helados (Y) durante 12 días del mes de julio en una óptica y en una heladería en miles de euros.

Venta de gafas de sol (x_i)	0,5	1,5	2	2	3	3	4	4	5	5	6	6
Venta de helados (y_i)	1	2	3	4	2,5	4	4	5	5	6	5	6

Vamos a comprobar por el diagrama de dispersión y el coeficiente de correlación que hay una relación lineal positiva fuerte entre ambas variables.



r = 0,89

Es evidente que no podemos afirmar que exista una relación de causa y efecto entre el gasto en gafas de sol y helados. Hay otra variable causante de esta relación entre X e Y que es la temperatura en el mes de julio.

5. EJERCICIOS RESUELTOS

EJERCICIO MODELO DE ESTADÍSTICA BIDIMENSIONAL

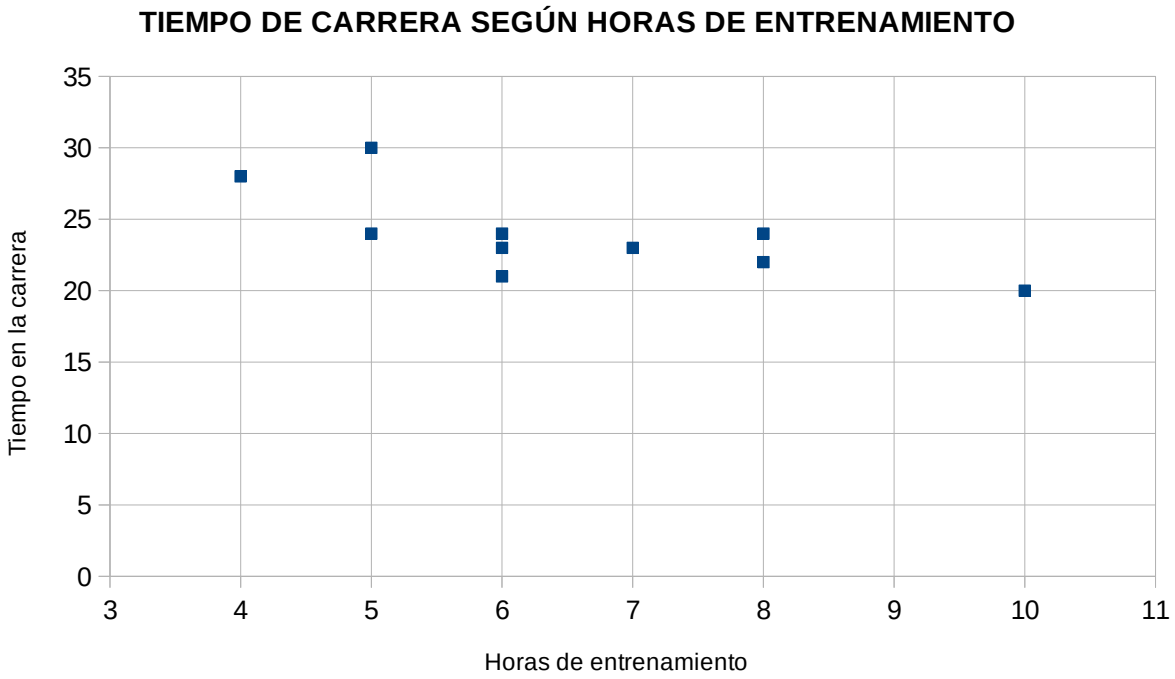
Se ha medido el número medio de horas de entrenamiento a la semana de un grupo de 10 atletas y el tiempo, en minutos, que han hecho en una carrera, obteniendo los siguientes resultados:

Horas de entrenamiento	5	6	6	5	8	6	8	10	7	4
Tiempo carrera	30	23	24	24	22	21	24	20	23	28

a) Estudia la dependencia de estas variables a partir de la nube de puntos.
b) Calcula la covarianza y el coeficiente de correlación.
c) Halla la ecuación de la recta de regresión de Y sobre X y represéntala en el diagrama de dispersión.
d) Halla la ecuación de la recta de regresión de X sobre Y y represéntala en el diagrama de dispersión.
e) Haz una estimación del tiempo que hará en la carrera un atleta que entrene 9 horas de media a la semana.
f) Haz una estimación del número de horas que había dedicado un atleta a entrenar si su tiempo en la carrera fue de 25 minutos.

SOLUCIÓN:

a) La dependencia en variables cuantitativas se estudia comprobando si la nube de puntos se ajusta a una recta (lineal) o a una curva (no lineal) o bien si son independientes (el valor de una no depende del valor de la otra, no se ajustan a un patrón).



En este caso se observa una relación de dependencia lineal entre las horas de entrenamiento y el tiempo en la carrera.
La dependencia es negativa, a más horas de entrenamiento en menos tiempo corrió la carrera.

b)

Horas de entrenamiento xi	**Tiempo de carrera yi**	$\mathbf{x_i^2}$	$\mathbf{y_i^2}$	$\mathbf{x_i\, y_i}$
5	30	25	900	150
6	23	36	529	138
6	24	36	576	144
5	24	25	576	120
8	22	64	484	176
6	21	36	441	126
8	24	64	576	192
10	20	100	400	200
7	23	49	529	161
4	28	16	784	112
65	**239**	**451**	**5795**	**1519**

Media de las x: $\bar{x}=\frac{\sum_{i=1}^{N} x_i}{N}=\frac{65}{10}=6{,}5$ Media de las y: $\bar{y}=\frac{\sum_{i=1}^{N} y_i}{N}=\frac{239}{10}=23{,}9$

Varianza de las x: $s_x^2 = \dfrac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2 = \dfrac{451}{10} - 6{,}5^2 = 2{,}85 \quad \Rightarrow \quad \sigma_x = \sqrt{\sigma_x^2} = 1{,}6882$

Varianza de las y: $s_y^2 = \dfrac{\sum_{i=1}^{N} y_i^2}{N} - \bar{y}^2 = \dfrac{5795}{10} - 23{,}9^2 = 8{,}29 \quad \Rightarrow \quad \sigma_y = \sqrt{\sigma_y^2} = 2{,}8792$

Covarianza: $s_{xy} = \dfrac{\sum_{i=1}^{N} x_i y_i}{N} - \bar{x}\,\bar{y} = \dfrac{1519}{10} - 6{,}5 \cdot 23{,}9 = -3{,}45$ dependencia negativa

Coeficiente de correlación: $r = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y} = \dfrac{-3{,}45}{1{,}688 \cdot 2{,}879} = -0{,}7099 \approx -0{,}71$

El coeficiente de correlación, es una medida que determina el grado de **dependencia lineal** entre las variables X e Y. No tiene unidades y su valor está entre -1 y 1. Un valor próximo a 0 indica la ausencia de dependencia lineal.

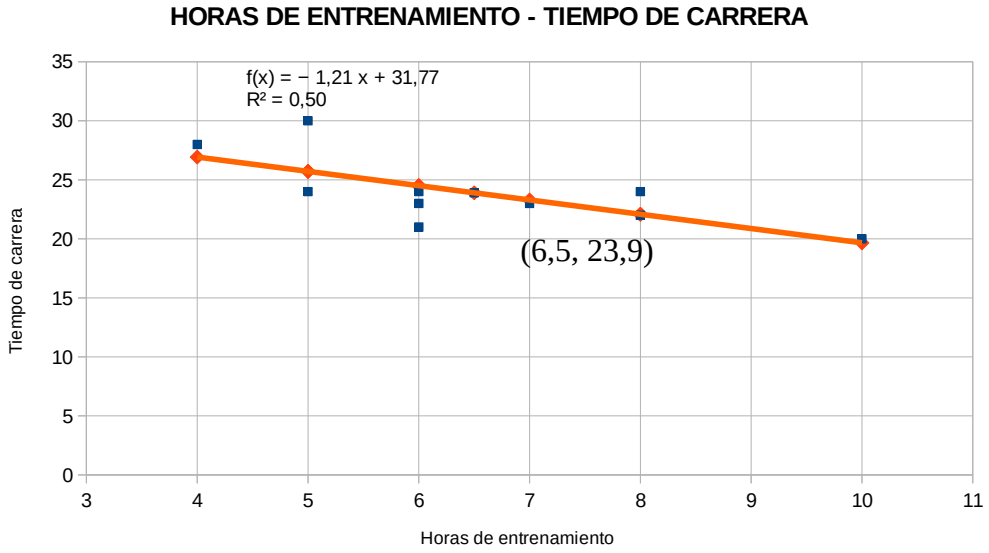
El valor de -0,71 de este ejemplo muestra una dependencia lineal negativa moderada. Los puntos de la nube se pueden aproximar por una recta aunque no se ajustarán por completo a ella.

c) De las rectas posibles que podemos ajustar a la nube de puntos elegimos la que hace mínima la suma de las distancias entre las ordenadas de cada punto y las de la recta. A esta recta se le llama recta de regresión de Y sobre X:

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$$

En este ejemplo, $y - 23{,}9 = \dfrac{-3{,}45}{2{,}85}(x - 6{,}5) \quad \Rightarrow \quad y = -1{,}21(x - 6{,}5) + 23{,}9$

$$\Rightarrow \quad y = -1{,}21\,x + 31{,}765$$



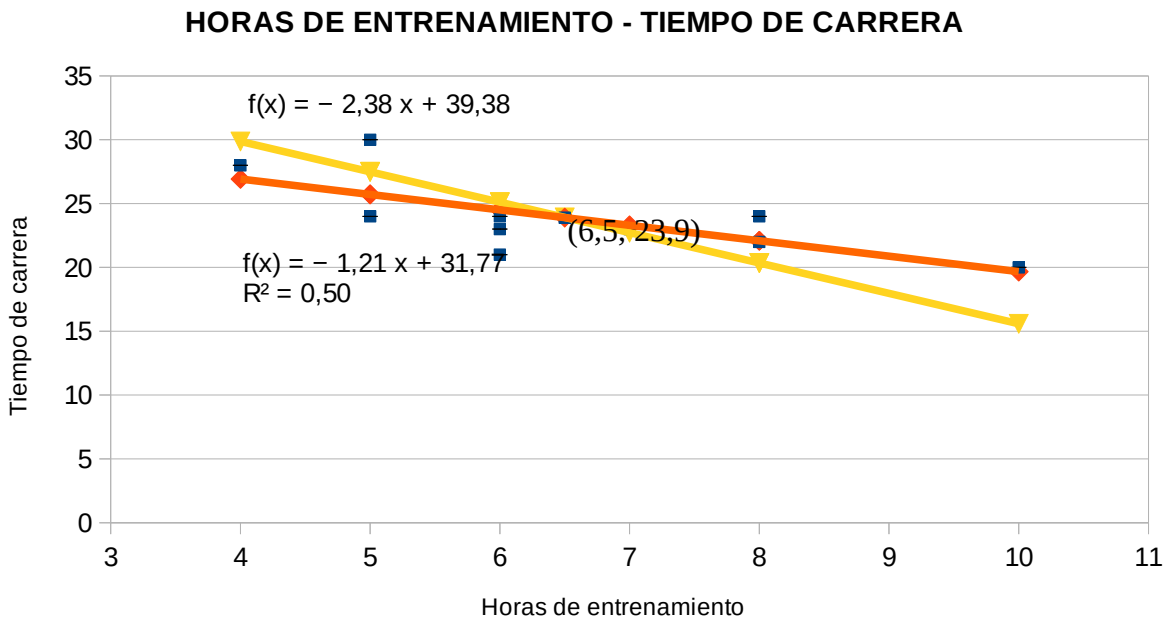
La recta de regresión pasa siempre por el punto $(\bar{x}, \bar{y}) = (6{,}5, 23{,}9)$, centro de gravedad de la nube de puntos.

d) Al minimizar las distancias entre las abscisas de cada punto y las de la recta , obtenemos la recta de regresión de X sobre Y:

$$x-\bar{x} = \frac{\sigma_{xy}}{\sigma_y^2}(y-\bar{y})$$

En este ejemplo, $x-6{,}5 = \frac{-3{,}45}{8{,}29}(y-23{,}9) \quad \Rightarrow \quad x = -0{,}42(y-23{,}9)+6{,}5$

$$\Rightarrow \quad x = -0{,}42\,y + 16{,}538$$



Las dos rectas pasan por el centro de gravedad (6,5, 23,9) y el ángulo de ambas rectas no es muy grande debido a la dependencia lineal de las variables.

e) Utilizamos la recta de regresión de Y sobre X: $\Rightarrow \quad y = -1{,}21\,x + 31{,}765$

$\Rightarrow y = -1{,}21 \cdot 9 + 31{,}765 = 20{,}875 \approx 21$ minutos

f) Utilizamos la recta de regresión de X sobre Y: $\Rightarrow \quad x = -0{,}42\,y + 16{,}538$

$\Rightarrow x = -0{,}42 \cdot 25 + 16{,}538 = 6{,}038 \approx 6$ horas de media a la semana

6. EJERCICIOS

1. En una investigación biomédica sobre la relación entre el hábito de fumar (X) y la hipertensión (Y) se tomaron estos datos:

	Con hipertensión	**Sin hipertensión**	
No fumador	20	50	
Fumador moderado	40	25	
Fumador empedernido	60	5	

a) Calcula la distribución conjunta de frecuencias relativas en porcentaje.
b) Representa estos datos en un diagrama de barras.
c) Halla la distribución de Y \ X = "Fumador empedernido". Halla la moda.
d) Halla la distribución de X \ Y = "Con hipertensión". Halla la moda.
e) Halla las distribuciones marginales de X e Y.

2. Se quiere analizar el gasto mensual en telefonía de 30 hogares, para ello se ha estudiado el número de móviles (Y) y el coste mensual en € (X):

Coste (€) \ Nº móviles	**1**	**2**	**3**	**4**	
[0, 20)	5	3	1	0	
[20, 40)	1	3	3	1	
[40, 60)	0	1	4	5	
[60, 80)	0	0	1	2	

a) Halla la distribución del número de móviles condicionado a un gasto entre 40€ y 60€. Halla la moda.
b) Halla la distribución del gasto entre los hogares que tienen 2 móviles. ¿Qué porcentaje de estos hogares tienen un gasto entre 20€ y 40€?
c) Halla las distribuciones marginales de X e Y.
d) Halla la media y la desviación típica de las distribuciones marginales.

3. En una competición de atletismo se ha estudiado la relación entre las horas de entrenamiento diarias (Y) y el puesto conseguido (X) por 20 atletas sabiendo que el 20% obtuvieron un primer puesto. Completa esta tabla de frecuencias con los datos que faltan:

Horas (Y) / Puesto (X)	**6**	**8**	**10**	**12**	
1	1	2	1		
2	1	3			11
3	2			1	
		6	8		

a) Calcula la distribución conjunta de frecuencias relativas en porcentaje. ¿Qué porcentaje de atletas entrenó 10 horas o más?

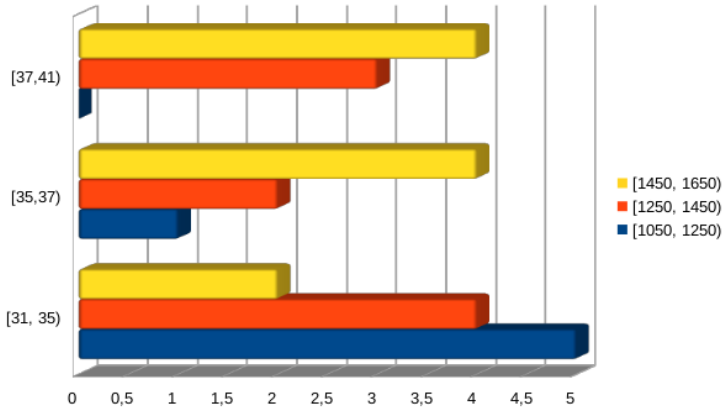
b) Halla la distribución del número de horas de entrenamiento condicionado a un primer puesto. Halla la moda.

c) Halla la distribución de puestos entre los que entrenaron 8 horas. ¿Qué porcentaje de estos atletas tienen obtuvieron un primer o segundo puesto?

d) Halla las distribuciones marginales de X e Y.

e) Halla la media y la desviación típica de las distribuciones marginales.

4. El número de horas trabajadas semanalmente (X) y el salario bruto semanal de un grupo de trabajadores de una empresa se recoge en la gráfica:



a) Pasa los datos a una tabla de doble entrada.

b) ¿A cuántos trabajadores se estudió?

c) Halla el número medio de horas trabajadas semanalmente.

d) ¿Cuántas horas trabaja la mayoría de las personas mejor pagadas?

e) Calcula el número de horas trabajadas por las personas que tienen un salario de menos de 1450€.

5. Dados los siguientes ejemplos dibuja su diagrama de dispersión y analiza si observas alguna relación lineal entre ellos:

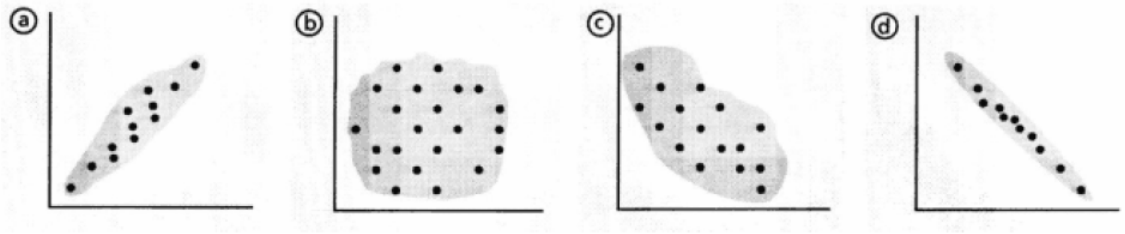
a)

X	2	3	5	5	6	7	8	8
Y	20	18	15	13	10	8	5	6

b)

X	5	9	12	15	17	20	25	30
Y	50	32	4	55	20	2	45	12

6. Los números 0,1; 0,99; 0,6 y 0,89 son los valores absolutos del coeficiente de correlación de las distribuciones bidimensionales cuyas nubes de puntos adjuntamos. Asigna a cada diagrama su coeficiente de correlación cambiando el signo cuando sea necesario.



Indica en cada caso la relación entre las variables.

7. Calcula la covarianza y la correlación lineal de los ejemplos del ejercicio 5. ¿Encuentras relación entre los parámetros calculados y el análisis del diagrama de puntos?

8. Se ha realizado un estudio a ocho empleados de una multinacional sobre el nivel de estrés (escala del 0 al 40) y la satisfacción laboral (escala del 0 al 10):

Nivel de estrés (X)	28	25	16	12	8	31	20	21
Satisfacción laboral (Y)	6	7	8	9	8	4	6	5

¿Se puede afirmar que, cuanto mayor es el nivel de estrés, menor es la satisfacción laboral? Razónalo mediante el diagrama de puntos primero y mediante el coeficiente de correlación después.

9. Los salarios brutos mensuales (Y), en euros, y la antigüedad (X), en años, de seis empleados de una empresa fueron:

Antigüedad (X)	0	1	2	3	4	5
Salario bruto (Y)	1200	1375	1600	1780	1900	2200

a) Dibuja la nube de puntos y analiza si se observa relación lineal entre las variables.

b) Halla la covarianza y el coeficiente de correlación lineal y compara el resultado con el apartado a)

c) Halla las rectas de regresión de Y sobre X y de X sobre Y.

d) Cuando se recogieron los datos había un empleado enfermo con 6 años de antigüedad, ¿podrías estimar cuánto cobra?

e) El empleado más antiguo de la empresa tiene 20 años de antigüedad, ¿podrías estimar también su salario?

f) Un empleado dice que cobra 2000€ al mes, ¿qué antigüedad dirías que tiene?

10. Una empresa tiene ocho tiendas y cada una realiza su propia compaña publicitaria. Los gastos de cada tienda en publicidad (X) y sus ventas anuales (Y), en miles de euros, vienen dados en la siguiente tabla:

Publicidad (X)	23	34	21	19	18	27
Ventas (Y)	1600	1700	1800	1200	1300	1700

a) Analiza si la publicidad influye favorablemente en las ventas de forma lineal dibujando un diagrama de dispersión.

b) La empresa está pensando en abrir otra tienda y se propone invertir 20000 € en publicidad. ¿Qué ingresos por ventas podría esperar?

c) ¿Qué cantidad debería invertir en publicidad si quiere alcanzar unas ventas anuales de 1,5 millones de euros? ¿Es fiable esta estimación?

11. Midiendo la potencia en CV y el consumo en l/100 km en seis modelos de diferentes de coches, hemos obtenido los siguientes resultados:

Potencia (X)	95	95	110	115	120	145
Consumo (Y)	4,8	5,1	5,2	6	6,2	7

a) Halla la recta de regresión de Y sobre X.
b) Calcula el consumo estimado de un coche de 190 CV. ¿Es fiable esta estimación?

12. En una academia para aprender a conducir se han estudiado las semanas de asistencia a clase de sus alumnos y las semanas que tardan en aprobar el examen teórico (desde que se apuntaron a la autoescuela). Los datos correspondientes a seis alumnos son:

Asistencia (X)	6	1	4	3	5	8
Aprobado (Y)	6	5	5	6	5	10

a) Halla las dos rectas de regresión y represéntalas.

b) Observando el grado de proximidad entre las dos rectas, ¿cómo crees que será la correlación entre las dos variables?

13. La talla media de una muestra de padres es de 1,68 m. con una desviación típica de 5 cm. y la talla media de una muestra de sus hijos es de 1,70 m. con una desviación típica de 7,5 cm. El coeficiente de correlación entre las tallas de hijos y padres es 0,7. Estimar la talla de dos hijos si la talla de sus padres fuera de 1,80 y 1,60 respectivamente.

14. En una prueba de natación de 100 m libres un conjunto de 6 nadadores obtienen las siguientes marcas:

55,3 s 54,9 s 58,1 s 52,8 s 56,4 s 57,3 s

a) Calcula la media y desviación típica del conjunto de tiempos.

b) Los mismos nadadores obtienen en la prueba de 100 m mariposa las siguientes marcas:

56,8 s 55,4 s 57,3 s 54,0 s 57,1 s 57,2 s

Calcular el coeficiente de correlación entre ambas pruebas y dar una interpretación. ¿Qué marca obtendría en 100 m mariposa un nadador con una marca de 55 s en 100 m libres?

15. Un examen de cierta asignatura consta de dos partes, una teórica (X) y otra práctica (Y). El profesor de la misma quiere ver si existe algún tipo de correlación entre las notas de teoría y práctica. Obtiene que la recta de regresión de Y sobre X es 4x – 3y = 0 y la de X sobre Y es 3x–2y=1.

a) Calcular el coeficiente de correlación y decir si las variables están o no correlacionadas.

b) Calcular la media de las notas de teoría y práctica.

16. Un conjunto de datos bidimensionales (x_i , y_i) tiene coeficiente de correlación r = -0,9 siendo las medias marginales 1 y 2, respectivamente. Se sabe que una de las cuatro ecuaciones siguientes corresponde a la recta de regresión de Y sobre X. Selecciona razonadamente dicha recta:

a) y = -x + 2 b) y = x + 1 c) 3x – y = 1 d) 2x + y = 4

17. Sea y = 3x – 10 la recta de regresión de Y sobre X. Sabiendo que $\bar{x}$ = 4, σ_{xy} = 3 y σ_y^2 = 16, encuentra la media de Y, la varianza de X, el coeficiente de correlación y la recta de X sobre Y.

18. Halla las rectas de regresión de esta distribución bidimensional:

X \ Y	8	9	10
3	4	13	3
5	6	7	7

19. - Las calificaciones de 40 alumnos en matemáticas II (X) y en métodos estadísticos (Y) han sido las siguientes:

Nota de matemáticas II (X)	3	4	5	6	6	7	7	8	9	10
Nota de métodos estadísticos (Y)	4	5	7	8	6	8	9	10	9	9
Nº alumnos	4	4	2	12	4	5	4	2	1	2

a) Obtener la ecuación de la recta de regresión de calificaciones de estadística respecto de las calificaciones de psicología.

b) ¿Cuál será la nota esperada en estadística para un alumno que obtuvo un 4,5 en psicología?

6.1. EJERCICIOS REPASO PARA EXAMEN

20. La siguiente tabla muestra los ingresos familiares mensuales de una familia en cientos de euros, X, y los metros cuadrados de la vivienda familiar, Y.

Y \ X	[0, 5)	[5, 10)	[10, 15)	[15, 20)	[20, 25)
[0, 50)	20	18	2	1	0
[50, 100)	25	40	30	2	1
[100, 200)	5	10	15	25	3
[200, 250)	0	5	15	20	8
[250, 300)	0	1	2	7	10

a) Calcula la distribución conjunta de frecuencias relativas en porcentaje.

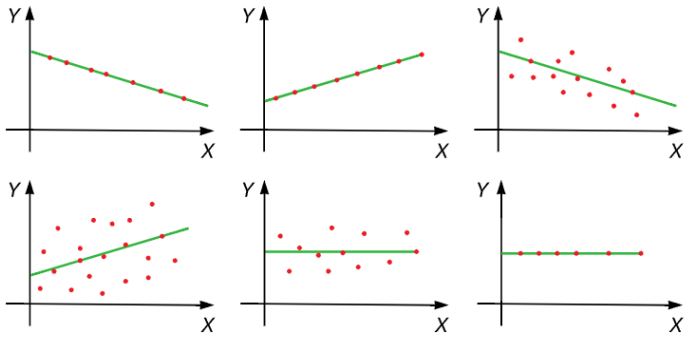
b) Representa estos datos en un diagrama de barras.

c) Halla la distribución de Y condicionado a las familias de ingresos entre 1000€ y 1500€ mensuales. Halla la moda.

d) Halla la distribución de X condicionada a las familias que viven en pisos de 250 m² o más. Halla la media de ingresos mensuales de estas familias.

e) Halla las distribuciones marginales de X e Y. Halla la varianza de ambas variables.

21. Relaciona cada diagrama de dispersión con su coeficiente de correlación:



a) 0 b) 0,3 c) −1 d) 1 e) No existe. f) −0,6

22. Un modelo muy conocido en las teorías de control de especies es la evolución de las poblaciones de zorros y conejos en función de su interacción. La población de conejos de una zona suele tener oscilaciones relacionadas con la cantidad de zorros que hay en esa misma área.

En una zona, en los últimos años, se han realizado ocho censos de animales.

Número de zorros	20	32	16	18	25	30	14	15
Número de conejos	320	500	260	300	400	470	210	240

Si la correlación es fuerte:

a) Determina las dos rectas de regresión.

b) Estima los conejos que habría si hubiera 10 zorros.

c) ¿Cuántos zorros habría, aproximadamente, si hubiéramos contado 350 conejos?

d) ¿Cuál de las dos estimaciones es más fiable?

23. Se ha medido el peso, X, y la estatura, Y, de los estudiantes de una clase. Su peso medio ha sido de 56 kg, con una desviación típica de 2,5 kg. La ecuación de la recta de regresión que relaciona la estatura y el peso es $y = 1{,}8x + 62$.

a) ¿Qué estatura puede estimarse en un estudiante que pesa 64 kg?

b) Y si un estudiante pesara 44 kg, ¿cuál sería su altura?

c) ¿Cuál es la estatura media de los estudiantes de esa clase?

d) La pendiente de esa recta es positiva, ¿qué significa esto?

24. Encuentra el coeficiente de correlación de la variable bidimensional cuyas rectas de regresión son:

- Recta de regresión de Y sobre X: $2x - y - 1 = 0$
- Recta de regresión de X sobre Y: $9x - 4y - 9 = 0$

a) Halla la media aritmética de cada una de las variables.

b) ¿Podrías calcular la desviación típica de Y sabiendo que la de la variable X es $\sqrt{2}$?