11 Distribuciones bidimensionales

EJERCICIOS PROPUESTOS

1. En una muestra de 8 ciudades, con aproximadamente la misma población, se ha calculado la temperatura media anual en grados centígrados (*X*) y el gasto medio anual en calefacción por habitante en euros (*Y*). Los resultados se recogen en la tabla siguiente:

Gasto medio (€)	75	140	20	170	52	90	100	125
Temperatura media (ºC)	22	12	24	10	20	18	20	16

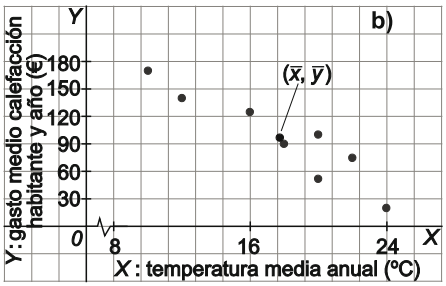
a) Representa gráficamente el conjunto de datos.

b) ¿Observas alguna tendencia en la nube de puntos? ¿De qué tipo?

c) Calcula las medias de la temperatura y del gasto por habitante en las 8 ciudades y sitúalas en el gráfico anterior.

d) ¿Cuál de las dos variables presenta mayor variabilidad? Razona la respuesta.

a)



Se observa que si aumenta la temperatura, disminuye el gasto en calefacción. La tendencia es, por tanto, decreciente en Y.

c) Para los cálculos de este apartado y del siguiente se añaden a la tabla las columnas necesarias

	Y	X	Y^2	X^2	$Y \cdot X$
	75	22	5625	484	1650
	140	12	19600	144	1680
	20	24	400	576	480
	170	10	28900	100	1700
	52	20	2704	400	1040
	90	18	8100	324	1620
	100	20	10000	400	2000
	125	16	15625	256	2000
SUMA	772	142	90954	2684	12170

$$\bar{Y} = \frac{772}{8} = 96{,}6\ €;\ \bar{X} = \frac{142}{8} = 17{,}75\ €$$

d) Se deben calcular los coeficientes de variación de ambas variables.

$$s_X^2 = \frac{2684}{8} - 17{,}75^2 = 20{,}438 \Rightarrow s_x = 4{,}521\ ºC \Rightarrow CV(X) = \frac{4{,}521}{17{,}75} = 0{,}2547$$

$$s_Y^2 = \frac{90\,954}{8} - 96{,}5^2 = 2057 \Rightarrow s_Y = 45{,}354\ € \Rightarrow CV(Y) = \frac{45{,}354}{96{,}5} = 0{,}47$$

El gasto medio en calefacción por habitante y año tiene mayor variabilidad que la temperatura media anual.

2. La siguiente tabla de contingencia da la distribución conjunta del sexo (*X*) y el hábito de correr (*Y*) de una muestra de 250 personas.

X \ *Y*	Corredor	No corredor	Totales *X*
Hombre	**	54	107
Mujer	62	**	**
Totales *Y*	**	**	250

a) Copia y completa la tabla.

b) Indica el porcentaje de no corredores.

c) ¿Cuál es el porcentaje de mujeres corredoras?

d) Dentro de los no corredores, ¿cuál es el porcentaje de mujeres?

e) Dentro de los hombres, ¿cuál es el porcentaje de corredores?

f) ¿Cuál sería el modo más adecuado para representar gráficamente estos datos?

a)

Y \ *X*	Corredor	No corredor	Totales *X*
Hombre	53	54	107
Mujer	62	81	143
Totales *Y*	115	135	250

b) $p_{NC} = \frac{135}{250} \cdot 100 = 54\%$

c) $p_{MC} = \frac{62}{250} \cdot 100 = 24,8\%$

d) $p(M \mid NC) = \frac{81}{135} \cdot 100 = 60\%$

e) $p(C \mid H) = \frac{53}{107} \cdot 100 = 49,53\%$

f) Lo más apropiado sería un diagrama de barras acumulados o un diagrama de sectores más adecuados para variables cualitativas.

3. La tabla adjunta recoge las calificaciones obtenidas por 10 alumnos en la parte escrita (*X*) y en la parte oral (*Y*) de un examen de inglés de la Escuela Oficial de Idiomas.

X	1,6	7,8	7,1	2,3	5,8	4,2	7,6	9,8	6,4	7,6
Y	2,0	4,0	5,0	5,5	6,0	6,5	7,1	7,6	8,4	9,3

a) Calcula las medias y las varianzas marginales.

b) Representa la nube de puntos.

c) A partir de la nube de puntos, comenta la relación entre las variables y su tendencia.

a) Se amplía la tabla con las columnas necesarias para los cálculos de las medias y las varianzas marginales:

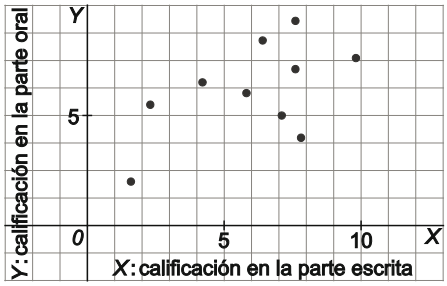
X	*Y*	X^2	Y^2
1,6	2,0	2,7	4,0
7,8	4,0	61,0	16,0
7,1	5,0	50,8	25,0
2,3	5,5	5,2	30,3
5,8	6,0	33,3	36,0
4,2	6,5	17,3	42,3
7,6	7,1	57,0	50,4
9,8	7,6	95,8	58,2
6,4	8,4	40,8	70,7
7,6	9,3	57,0	85,6
60,1	61,4	421,0	418,4

De esta manera, las medias y las varianzas de la parte escrita (*X*) y la oral (*Y*) son:

$$\overline{X} = \frac{60,1}{10} = 6,01 \qquad s_X^2 = \frac{421}{10} - 6,01^2 = 6,0235$$

$$\overline{Y} = \frac{61,4}{10} = 6,14 \qquad s_Y^2 = \frac{418,4}{10} - 6,14^2 = 4,1544$$

b) La nube de puntos de la variable estadística bidimensional es



c) La relación es directa, a mayor calificación en la parte escrita mayor calificación en la parte oral. Pero la relación no es muy fuerte ya que los puntos no se sitúan próximos a una línea recta.

4. La tabla muestra la puntuación (*Y*) obtenida por 1000 personas en función de su edad (*X*), en un test de nociones básicas de aritmética.

X \ Y	0 – 175	176 – 225	226 – 275	276 – 325	326 – 500
16 – 35	23	62	163	94	28
36 – 55	24	55	159	80	22
55 – 65	33	65	127	53	12

a) Obtén las distribuciones marginales.

b) Halla las medias y las varianzas marginales. Utiliza en los cálculos la marca de clase de cada intervalo.

a) Se amplía la tabla con las filas y columnas de las frecuencias marginales de *X* (f_X) e *Y* (f_Y). También se han añadido las filas y columnas necesarias para los cálculos del apartado b).

X \ Y	0 – 175	176 – 225	226 – 275	276 – 325	326 – 500	f_{xj}	x_j	$f_{Xj}x_j$	$f_{Xj}x_j^2$
16 – 35	23	62	163	94	28	370	25,5	9435	240 592,5
36 – 55	24	55	159	80	22	340	45	15 300	688 500
56 – 75	33	65	127	53	12	290	65	18 850	1 225 250
f_{Yj}	80	182	449	227	62	1000		43 585	2 154 342,5
y_j	87,5	200	250	300	412,5				
$f_{yj}y_j$	7000	36 400	112 250	68 100	25 575	249 325			
$f_{yj}y_j^2$	612 500	7 280 000	28 062 500	20 430 000	10 549 687,5	66 934 687,5			

b) Las medias y varianzas marginales se obtienen a partir de los datos de la tabla de la siguiente manera:

$$\overline{X} = \frac{43585}{1000} = 43,585 \text{ años} \quad ; \quad s_x^2 = \frac{2154342,5}{1000} - 43,585^2 = 254,690$$

$$\overline{Y} = \frac{249325}{1000} = 249,325 \text{ años} \; ; \; s_Y^2 = \frac{66934687,5}{1000} - 254,59^2 = 4771,732$$

5. Ejercicio resuelto.

6. La distribución de 1163 fumadores según sexo (X) y grupo de edad de 15 a 54 años (Y), se recoge en la tabla siguiente.

X \ Y	[15, 24]	[25, 34]	[35, 44]	[45, 54]
Hombres	112	178	164	172
Mujeres	105	141	141	150

a) Escribe las distribuciones marginales.

b) Halla les distribuciones de frecuencias relativas de Y condicionadas por cada valor de X.

c) Halla la media y la varianza de Y | hombres.

d) ¿Son dependientes estas variables?

a) La tabla con las distribuciones marginales es:

X \ Y	[15, 24]	[25, 34]	[35, 44]	[45, 54]	f_{x_i}
Hombres	112	178	164	172	626
Mujeres	105	141	141	150	537
f_{y_j}	217	319	305	322	1163

b) Para hallar las marcas de clase se puede usar el mismo criterio que en el ejercicio 4, la media entre el extremos superior de un intervalo y el superior del siguiente, en este ejercicio se calcula la media entre los extremos del intervalo.

X	$h_{i\|1}$	$h_{i\|2}$	$h_{i\|3}$	$h_{i\|4}$	
Hombres	$h_{1\|1} = \frac{112}{626} \simeq 0,1789$	$h_{1\|2} = \frac{178}{626} \simeq 0,2843$	$h_{1\|3} = \frac{164}{626} \simeq 0,2620$	$h_{1\|4} = \frac{172}{626} \simeq 0,2748$	1
Mujeres	$h_{2\|1} = \frac{105}{537} \simeq 0,1955$	$h_{2\|2} = \frac{141}{537} \simeq 0,2626$	$h_{2\|3} = \frac{141}{537} \simeq 0,2626$	$h_{2\|4} = \frac{150}{537} \simeq 0,2793$	1

c) Para el cálculo de la media y varianza de $Y|_{X=\text{hombres}}$, consideramos la tabla:

$Y\|_{X=\text{hombres}}$	[15 , 24]	[25 , 34]	[35 , 44]	[45 , 54]	
f_j	112	178	164	172	626
$f_j \cdot Y\|_{X=h}$	2184	5251	6478	8514	22 427
$f_j \cdot Y\|_{X=h^2}$	42 588	154 904,5	255 881	421 443	874 816,5

$$\overline{Y}\,|_{X=h} = \frac{22427}{626} = 35,826 \text{ años} \; ; \; s^2_{Y|X=h} = \frac{874816,5}{626} - 35,826^2 = 113,98$$

d) La tabla de distribuciones conjuntas y marginales es:

h_{ij}	[15, 24]	[25, 34]	[35, 44]	[45, 54]	h_i
Hombres	0,0963	0,1531	0,1410	0,1479	0,5383
Mujeres	0,0903	0,1212	0,1212	0,1290	0,4617
h_j	0,1866	0,2743	0,2622	0,2769	1

Las variables son estadísticamente dependientes. Se puede comprobar que, por ejemplo $h_{14} \neq h_1 \cdot h_4$ $(0,1479 \neq 0,5383 \cdot 0,2769 = 0,1491)$.

7. Ejercicio resuelto.

8. Un concesionario de coches contrata empleados para los fines de semana. La tabla muestra los coches vendidos (*Y*) y los vendedores (*X*) en una muestra de 8 fines de semana.

X	6	5	4	4	6	3	5	8
Y	18	20	10	14	15	7	13	21

a) Representa la nube de puntos de la distribución.

b) Escribe la ecuación de la recta de regresión de *Y* sobre *X*

c) Calcula la varianza residual.

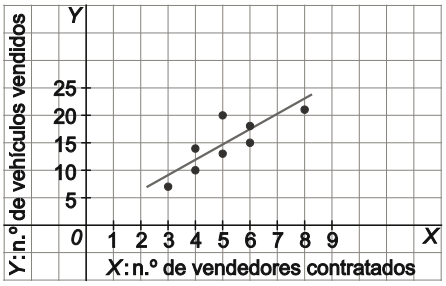
d) Si la empresa decide contratar 8 vendedores, ¿cuántos coches se estima que podría vender?

e) Si contrata 8 vendedores, calcula el residuo correspondiente y valora el resultado obtenido.

a) Se representa la nube de puntos junto con la recta de regresión ajustada de *Y* sobre *X* que se obtiene en el apartado b)

b) Para obtener la recta de regresión de *Y* sobre *X*, se amplía la tabla dada con las columnas que se necesitan para calcular las medias, las varianzas y la covarianza:

x_j	y_j	x_j^2	y_j^2	$x_j y_j$
6	18	36	324	108
5	20	25	400	100
4	10	16	100	40
4	14	16	196	56
6	15	36	225	90
3	7	9	49	21
5	13	25	169	65
8	21	64	441	168
41	118	227	1904	648



Entonces

$\overline{X} = \frac{41}{8} = 5,125$ vendedores ; $s_X^2 = \frac{227}{8} - 5,125^2 = 2,11$; $s_X = 1,452$ vendedores

$\overline{Y} = \frac{118}{8} = 14,75$ vehículos ; $s_Y^2 = \frac{1904}{8} - 14,75^2 = 20,44$; $s_Y = 4,521$ vehículos

$s_{XY} = \frac{648}{8} - 5,125 \cdot 14,75 = 5,406$

De esta forma, la pendiente y la ordenada en el origen de la recta de regresión son respectivamente:

$a = 14,75 - 2,563 \cdot 5,125 = 1,615$; $b = \frac{5,406}{2,11} = 2,563$

La ecuación de la recta de regresión de *Y* sobre *X* es $y = 1,615 + 2,563x$.

c) La varianza residual o Error Cuadrático Medio es

$$\text{ECM}_{Y|X} = s_Y^2\left(1 - \frac{s_{XY}^2}{s_X^2 s_Y^2}\right) = 20,44 \cdot \left(1 - \frac{5,406^2}{2,11 \cdot 20,44}\right) = 6,851$$

d) Sustituyendo $x = 8$ en la ecuación de la recta de regresión obtenida en el apartado b), resulta

$$y(8) = 1,615 + 2,563 \cdot 8 = 22,12$$

Se estima que podría vender alrededor de 22 coches.

e) El residuo correspondiente se obtiene mediante la diferencia entre el valor observado para $x = 8$, que es 21, y el valor pronosticado, que es 22,12. Esto es –1,12.
La diferencia entre el valor observado y el estimado (pronosticado) es apenas de 1 vehículo de los más de 20 vendidos. Luego puede considerarse una buena estimación.

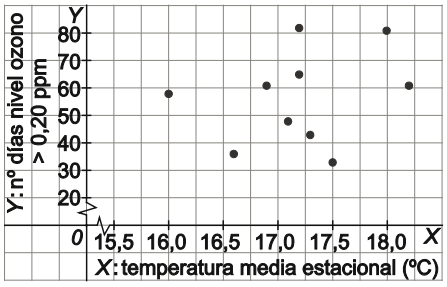
9. La tabla recoge la temperatura media invernal (*X*) en ºC en una ciudad costera y el número de días (*Y*) en que el nivel de ozono superó los 0,20 ppm (partes por millón) durante 10 años.

X	16,0	17,2	18,0	17,2	16,9	17,1	18,2	17,3	17,5	16,6
Y	58	82	81	65	61	48	61	43	33	36

a) Dibuja el diagrama de dispersión.

b) Estima el número de días en los que se superará el nivel de ozono estándar (0,20 ppm) si la temperatura media estacional es de 16º y analiza la precisión de la predicción en función del ECM.

a) El diagrama de dispersión del número de días en que se superó el nivel límite de ozono (*Y*) en función de la temperatura media estacional (*X*) se muestra a la derecha.



b) Para llevar a cabo la estimación pedida, se debe obtener la recta de regresión del número de días (*Y*) sobre la temperatura media estacional (*X*). Para ello, se amplía la tabla de datos con las filas correspondientes para el cálculo de los valores medios, las varianzas y la covarianza:

x_j	16	17,2	18	17,2	16,9	17,1	18,2	17,3	17,5	16,6	172
y_j	58	82	81	65	61	48	61	43	33	36	568
x_j^2	256	295,8	324	295,8	285,6	292,4	331,2	299,3	306,3	275,6	2962,04
y_j^2	3364	6724	6561	4225	3721	2304	3721	1849	1089	1296	34854
$y_j x_j$	928	1410	1458	1118	1031	820,8	1110	743,9	577,5	597,6	9795,3

$$\overline{X} = \frac{172}{10} = 17,2\,ºC \quad ; \quad s_X^2 = \frac{2962,04}{10} - 17,2^2 = 0,364$$

$$\overline{Y} = \frac{568}{10} = 56,8 \text{ días} \quad ; \quad s_Y^2 = \frac{34854}{10} - 56,8^2 = 259,16$$

$$s_{XY} = \frac{9795,3}{10} - 56,8 \cdot 17,2 = 2,57$$

Los coeficientes de la recta de regresión son $a = 56,8 - 7,06 \cdot 17,2 = -64,64$; $b = \frac{2,57}{0,364} = 7,06$.

La recta de regresión de *Y* sobre *X* es: $y = -64,64 + 7,06x$.

Entonces, si x =16ºC, se estima que el número de días en que se superará el límite de ozono es:

$$y = 7,06 \cdot 16 - 64,4 = 48,33 \text{ días}$$

es decir aproximadamente entre 48 y 49 días.
El Error Cuadrático Medio es:

$$ECM_{Y|X} = s_Y^2\left(1 - \frac{s_{XY}^2}{s_X^2 s_Y^2}\right) = 259,16\left(1 - \frac{2,57^2}{0,364 \cdot 259,16}\right) = 259,16\left(1 - 0,07\right) = 241,01$$

El Error Cuadrático Medio es alto e indica que el ajuste de la recta a la nube de puntos no es bueno.

El coeficiente de determinación $\frac{s_{XY}^2}{s_X^2 s_Y^2} = 0,07$ y puede decirse que solo el 7 % de la variabilidad observada en el número de días en que se superó el nivel de ozono de 0,20 ppm se explica por la temperatura media estacional.
En definitiva, la relación lineal entre ambas variables es muy débil y, por tanto, las predicciones que se puedan hacer con la recta de regresión estimada en el apartado anterior no son fiables.

10. Ejercicio interactivo.

11. La empresa que distribuye una conocida marca de refrescos ha tomado al azar 10 semanas del pasado año, recogiendo los siguientes datos:

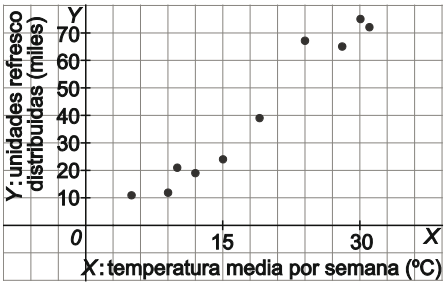
***X*: "temperatura media de cada semana en ºC"**

***Y*:"unidades de refrescos distribuidas en miles"**

X	10	28	12	31	30	19	24	5	9	15
Y	21	65	19	72	75	39	67	11	12	24

Representa la nube de puntos y calcula los coeficientes de determinación y de correlación, interpretando los resultados.

El diagrama de dispersión o nube de puntos de la distribución se muestra a la derecha.



Añadimos las columnas necesarias a la tabla para realizar los cálculos que se piden:

x_j	y_j	x_j^2	y_j^2	$y_j x_j$
10	21	100	441	210
28	65	784	4225	1820
12	19	144	361	228
31	72	961	5184	2232
30	75	900	5625	2250
19	39	361	1521	741
24	67	576	4489	1608
5	11	25	121	55
9	12	81	144	108
15	24	225	576	360
183	405	4157	22 687	9612

$$\overline{X} = \frac{183}{10} = 18,3\,ºC \quad ; \quad s_X^2 = \frac{4157}{10} - 18,3^2 = 80,81$$

$$\overline{Y} = \frac{405}{10} = 40,5\,\text{unidades} \quad ; \quad s_Y^2 = \frac{22687}{10} - 40,5^2 = 628,45$$

$$s_{XY} = \frac{9612}{10} - 40,5 \cdot 18,3 = 220,05$$

Los coeficientes de determinación y correlación son respectivamente:

$$R^2 = \frac{s_{XY}^2}{s_X^2 s_Y^2} = \frac{220,05^2}{80,81 \cdot 628,45} = 0,9535 \Rightarrow r = \sqrt{0,9535} = 0,9765$$

Esto indica que un alto porcentaje, el 95,35 %, de la variabilidad observada en la venta de refrescos es explicada por la variabilidad de la temperatura.

El valor del coeficiente de correlación, $r = 0,9765$, indica que la relación entre la temperatura media semanal y las unidades de refrescos vendidas es lineal y directa y con un elevado nivel de fiabilidad, (alta correlación, próxima a 1).

12. Ejercicio resuelto.

13. Los datos de la siguiente tabla se refieren a una muestra de 10 viviendas en las que se han observado el número de habitaciones (*X*) y el de personas que habitan en la vivienda (*Y*)

X	2	2	3	3	4	4	4	4	5	5
Y	1	2	2	3	2	4	5	6	4	6

a) Calcula la recta de regresión de *Y* sobre *X*.

b) Calcula los coeficientes de determinación y de correlación y valora el ajuste de la recta a la nube de puntos.

c) ¿Cuál es el porcentaje de la variabilidad del número de habitantes por vivienda explicado por el número de habitaciones?

d) ¿Cuál es el número estimado de personas que habitan en una vivienda de 3 habitaciones?

a) Con la ayuda de los datos de la tabla siguiente, se calculan las medias, las varianzas y la covarianza de las dos variables

x_i	y_j	x_i^2	y_j^2	x_iy_j
2	1	4	1	2
2	2	4	4	4
3	2	9	4	6
3	3	9	9	9
4	2	16	4	8
4	4	16	16	16
4	5	16	25	20
4	6	16	36	24
5	4	25	16	20
5	6	25	36	30
36	35	140	151	139

$$\overline{X} = \frac{36}{10} = 3,6 \text{ habitaciones/vivienda}; \quad s_X^2 = \frac{140}{10} - 3,6^2 = 1,04$$

$$\overline{Y} = \frac{35}{10} = 3,5 \text{ personas/vivienda}; \quad s_Y^2 = \frac{151}{10} - 3,5^2 = 2,85$$

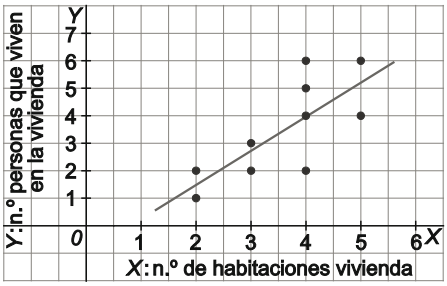
$$s_{XY} = \frac{139}{10} - 3,5 \cdot 3,6 = 1,3$$

La pendiente y la ordenada en el origen de la recta de regresión son respectivamente

$$b = \frac{1,3}{1,04} = 1,25 \quad ; \quad a = 3,5 - 1,25 \cdot 3,6 = -1$$

Luego la ecuación de la recta de regresión del número de personas que habitan una vivienda (*Y*) sobre el número de habitaciones (*X*) es: $y = -1 + 1,25x$.

b) La representación gráfica de la recta de regresión estimada sobre las distribución de datos se muestra a la derecha.



A partir de los cálculos efectuados en el apartado anterior, se obtienen los coeficientes de determinación y de correlación:

$$R^2 = \frac{1,3^2}{1,04 \cdot 2,85} = 0,5702 \Rightarrow r = \sqrt{0,5702} = 0,7551$$

El ajuste de la recta de regresión a la nube de puntos es razonablemente aceptable a la vista de la representación gráfica y del valor del coeficiente de correlación lineal.

c) El 57,02 % de la variabilidad observada en el número de personas (*Y*) que habitan en una vivienda es explicado por el número de habitaciones (*X*) que tiene la misma.

d) $y(3) = -1 + 1,25 \cdot 3 = 2,75$ personas.

14. La media de las calificaciones globales (*Y*), obtenidas por 10 alumnos fue 6,8 puntos y sus horas semanales de estudio (*X*) suman 120. Se sabe que el coeficiente de correlación es 0,8 y que las desviaciones típicas de *X* e *Y* coinciden. Con estos datos, ¿puedes estimar la calificación de un alumno que ha estudiado 10 horas semanales?

Sabemos que: $\overline{X} = \frac{120}{10} = 12 \quad ; \quad \overline{Y} = 6,8$.

La pendiente y la ordenada en el origen de la recta de

regresión: $b = r \cdot \frac{s_Y}{s_X} = 0,8 \cdot 1 = 0,8 \quad ; \; a = \overline{Y} - b \cdot \overline{X} = 6,8 - 0,8 \cdot 12 = -2,8$, y la recta de regresión es: $y = -2,8 + 0,8x$.

$y(10) = -2,8 + 0,8 \cdot 10 = 5,2$ puntos. Se trata de un valor fiable, ya que se encuentra dentro del rango de valores empíricos y el coeficiente de correlación es alto (0,8).

15. Los datos de la tabla siguiente se refieren a una muestra de 8 parejas hermano – hermana adultos, en las que se ha observado la estatura del hombre (*Y*) y de la mujer (*X*) en centímetros.

X	164	162	167	168	167	152	160	170
Y	172	171	178	183	166	164	180	169

a) Representa la nube de puntos.

b) Escribe la recta de regresión de *Y* sobre *X*.

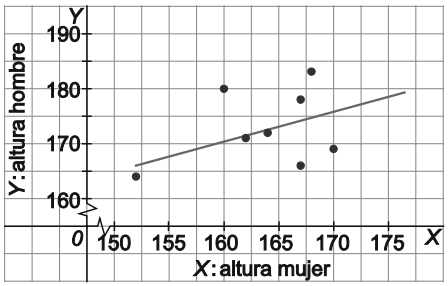
c) Calcula los coeficientes de determinación y de correlación y valora el ajuste de la recta a la nube de puntos

d) ¿Cuál es el porcentaje de variabilidad de la estatura de los hombres explicado por la estatura de los hombres?

e) ¿Cuál es la estatura estimada de un hombre, si su hermana mide 165 cm?

f) ¿Cuál es la estatura estimada de una mujer, si su hermano mide 175 cm?

a) La gráfica de dispersión de la distribución y la recta de regresión calculada en el siguiente apartado se muestra a la derecha.



b) Los cálculos necesarios para obtener la recta de regresión de *Y* (estatura hombres) sobre *X* (estatura mujer) son:

x_j	y_j	x_j^2	y_j^2	$y_j x_j$
164	172	26 896	29 584	28 208
162	171	26 244	29 241	27 702
167	178	27 889	31 684	29 726
168	183	28 224	33 489	30 744
167	166	27 889	27 556	27 722
152	164	23 104	26 896	24 928
160	180	25 600	32 400	28 800
170	169	28 900	28 561	28 730
1310	1383	214 746	239 411	226 560

$$\overline{X} = \frac{1310}{8} = 163,75\,\text{cm} \quad ; \quad s_x^2 = \frac{214746}{8} - 163,75^2 = 29,1875$$

$$\overline{Y} = \frac{1383}{8} = 172,875\text{cm} \quad ; \quad s_y^2 = \frac{239411}{8} - 172,875^2 = 40,6094$$

$$s_{xy} = \frac{226560}{8} - 172,875 \cdot 163,75 = 11,7188$$

$$b = \frac{s_{xy}}{s_x^2} = \frac{11,7188}{29,1875} = 0,4015 \quad ; \quad a = \overline{Y} - b \cdot \overline{X} = 172,875 - 0,4015 \cdot 163,75 = 107,13$$

Por tanto, la ecuación de la recta de regresión es: $y = 107,13 + 0,4015x$

c) $R^2 = \dfrac{s_{xy}^2}{s_x^2 s_y^2} = \dfrac{11,7188^2}{29,1875 \cdot 40,60094} = 0,1159 \Rightarrow r = \sqrt{0,1159} = 0,3404$

El valor $r = 0,3404$ indica que la relación lineal entre ambas variables es débil, como se puede observar en el gráfico del aparatado a).

d) La simetría de los cálculos indica que solo un 11,59 % de la variabilidad de la estatura de las mujeres es explicada por la variabilidad de la estatura de sus hermanos.

e) $y(165) = 107,13 + 0,4015 \cdot 165 = 173,38$ cm

f) Se sustituye el valor esperado y=175 en la ecuación de la recta de regresión de *Y* sobre *X* y se despeja el valor de x:

$$175 = 107,13 + 0,4015 \cdot x \Rightarrow x = 169,04\,\text{cm}$$

16. Ejercicio interactivo.

17 a 22. Ejercicios resueltos.

EJERCICIOS

Variables bidimensionales: nube de puntos y distribuciones marginales.

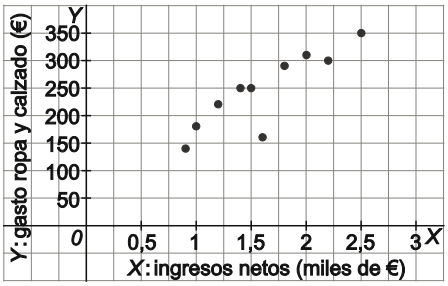
23. En la tabla se dan los ingresos netos en miles de euros (*X*) y el gasto en ropa y calzado (*Y*) en euros de 10 familias en el mes de julio.

X	1,2	2,2	0,9	1,5	1,8	1,4	1,0	2,0	2,5	1,6
Y	220	300	140	250	290	250	180	310	350	160

a) Representa la nube de puntos y comenta cuál es la tendencia observada en la relación entre las dos variables.

b) Calcula las medias y las varianzas marginales.

a) El diagrama de dispersión o nube de puntos de la distribución conjunta de *X* e *Y* se muestra a la derecha..



b)

x_i	y_j	x_i^2	y_j^2
1,2	220	1,44	48 400
2,2	300	4,84	90 000
0,9	140	0,81	19 600
1,5	250	2,25	62 500
1,8	290	3,24	84 100
1,4	250	1,96	62 500
1	180	1	32 400
2	310	4	96 100
2,5	350	6,25	122 500
1,6	160	2,56	25 600
16,1	2450	28,35	643 700

De esta forma, la media y la varianza de cada una de las variables es:

$$\overline{X} = \frac{16,1}{10} = 1,61 \text{miles de euros} \quad ; \quad s_x^2 = \frac{28,35}{10} - 1,61^2 = 0,2429$$

$$\overline{Y} = \frac{2450}{10} = 245 \text{ euros} \quad ; \quad s_y^2 = \frac{643700}{10} - 245^2 = 4345$$

24. La tabla muestra el número de hijos (*Y*) que tienen 50 mujeres en función de su edad (*X*).

X \ *Y*	0	1	2	3	*Total Y*
20 – 25	7	2	1	**	10
26 – 30	**	5	3	**	**
31–35	3	**	7	2	**
36–40	1	1	**	2	**
Total X	16	**	14	6	50

a) Copia y completa la tabla.

b) Obtén las distribuciones marginales y sus medias y varianzas.

c) Escribe la distribución del número de hijos si la madre está entre 31 y 35 años. Calcula su media y su varianza.

a) Se completa la tabla con los datos que faltan (en negrita).

X \ *Y*	0	1	2	3	*Total Y*
20 – 25	7	2	1	**0**	10
26 – 30	**5**	5	3	**2**	**15**
31–35	3	**6**	7	2	**18**
36–40	1	1	**3**	2	**7**
Total X	16	**14**	14	6	50

b) Las distribuciones marginales se pueden ver en la última fila (*Y*) y en la última columna (*X*) de la tabla del apartado anterior, no obstante, se pueden escribir:

X	f_X
20 – 25	10
26 – 30	15
31 – 35	18
36 – 40	7

Y	f_Y
0	16
1	14
2	14
3	6

Completando las tablas anteriores con las columnas necesarias y la fila de las sumas tenemos:

Clases	f_j	x_j	$f_j \cdot x_j$	$f_j \cdot x_j^2$
20 – 25	10	23	230	5290
26 – 30	15	28	420	11 760
31 – 35	18	33	594	19 602
36 – 40	7	38	266	10 108
	50		1510	46 760

y_j	f_j	$f_j \cdot yj$	$f_j \cdot y_j^2$
0	16	0	0
1	14	14	14
2	14	28	56
3	6	18	54,0
	50	60	124

Entonces, las medias y las varianzas marginales son:

$$\overline{X} = \frac{1510}{50} = 30,2 \text{ años} \; ; s_x^2 = \frac{46760}{50} - 30,2^2 = 23,16$$

$$\overline{Y} = \frac{60}{50} = 1,2 \text{ hijos} \qquad ; s_Y^2 = \frac{124}{50} - 1,2^2 = 1,04$$

c) En el caso en que la madre esté entre 31 y 35 años la tabla ampliada es:

Y	0	1	2	3	$Total\ Y\vert_{X=[31,35]}$
$X = [31,35]$	3	6	7	2	18
$f_j \cdot Y_j\vert_{X=[31,35]}$	0	6	14	6	26
$f_j \cdot Y_j^2\vert_{X=[31,35]}$	0	6	28	18	52

$$\overline{Y}\,\big|_{X=[31,35]} = \frac{26}{18} = 1,44 \text{ hijos} \qquad ; s^2_{Y|X=[1,35]} = \frac{52}{18} - 1,44^2 = 0,8025$$

25. El número de unidades producidas al mes en miles (*X*) por una empresa y el número de unidades defectuosas (*Y*) en 6 meses es:

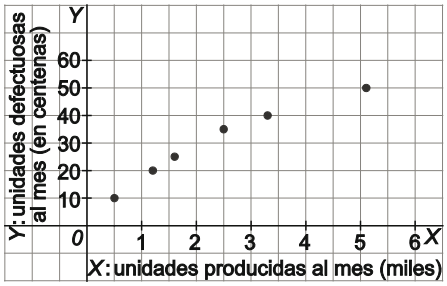
X	0,5	1,2	1,6	2,5	3,3	5,1
Y	10	20	25	35	40	50

a) Representa gráficamente los datos.

b) Calcula los parámetros media y varianza de las distribuciones marginales de *Y* y *X*.

c) ¿Se puede afirmar que la nube de puntos puede ajustarse por una recta? Justifica la respuesta.

a) La nube de puntos de la distribución conjunta se muestra a la derecha.



b)

x_j	y_j	x_j^2	y_j^2
0,5	10	0,25	100
1,2	20	1,44	400
1,6	25	2,56	625
2,5	35	6,25	1225
3,3	40	10,89	1600
5,1	50	26,01	2500
14,2	180	47,4	6450

$\overline{X} = \frac{14,2}{6} = 2,367$ miles de uds./mes ; $s_x^2 = \frac{47,4}{6} - 2,367^2 = 2,2989$

$\overline{Y} = \frac{180}{6} = 30$ defectuosas/mes ; $s_y^2 = \frac{6450}{6} - 30^2 = 175$

c) A la vista del diagrama de dispersión, parece claro que el número de unidades defectuosas (*Y*) aumenta con el aumento de la producción (*X*) y que esa relación es aproximadamente lineal. Por ello puede afirmarse que la recta de regresión de *Y* sobre *X* será un buen ajuste lineal a la nube de puntos de la distribución.

Covarianza, regresión lineal y correlación.

26. De las 100 observaciones de la distribución conjunta de la variable bidimensional (*X*,*Y*) se obtiene la siguiente información:

- **Las medias muestrales de *X* e *Y* son 0,9 y 1,2 respectivamente.**
- **Las varianzas muestrales de *X* e *Y* son 10,09 y 13,96 respectivamente.**
- **La covarianza es 8,12.**

Con estos datos:

a) Calcula el coeficiente de correlación lineal y, a la vista del resultado obtenido, razona si es correcto un ajuste lineal entre las dos variables.

b) Calcula el Error Cuadrático Medio del ajuste lineal.

c) Escribe la recta de regresión de *Y* sobre *X* y estima el valor esperado de *Y* cuando la variable *X* tome el valor 1.

a) $R^2 = \frac{s_{XY}^2}{s_X^2 s_Y^2} = \frac{8,12^2}{10,09 \cdot 13,96} = 0,4333 \Rightarrow r = \sqrt{0,4333} = 0,6583$

Apenas el 43,3 % de la variabilidad de *Y* es explicada por la variabilidad de *X*, lo que arroja un coeficiente de correlación de 0,6583. Puede ajustarse una recta de regresión entre *X* e *Y*, con precauciones si se utiliza la recta para realizar predicciones de una variable en función de los valores de la otra, valorando el error cometido en el valor estimado.

b) $\text{ECM} = s_y^2(1 - R^2) = 13,96 \cdot (1 - 0,4333) = 7,911$

c) La pendiente y la ordenada en el origen de la recta de regresión de *Y* sobre *X* son, respectivamente:

$b = \frac{s_{XY}}{s_x^2} = \frac{8,12}{10,9} = 0,745$; $a = \overline{Y} - b \cdot \overline{X} = 1,2 - 0,745 \cdot 0,9 = 0,53$

La recta de regresión de *Y* sobre *X* es: $y = 0,53 + 0,745x$

Y si $x = 1$, el valor esperado de *y* es: $y(1) = 0,53 + 0,745 \cdot 1 = 1,275$

27. En una determinada región vinícola se han evaluado las pérdidas en cientos de miles de euros (*Y*) en la producción en función del número de días de lluvia (*X*) de la campaña.

X	87	80	77	75	63	71	76	74
Y	19,5	18,5	14,6	15,5	12,7	14	18,1	15

a) A la vista de la representación gráfica, ¿se puede afirmar que existe una relación lineal?

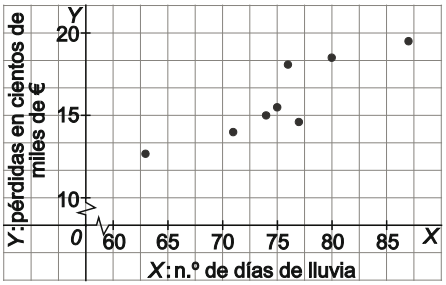
b) Calcula el coeficiente de correlación, ¿Se confirma la impresión anterior?

c) Escribe la ecuación de la recta de regresión de *Y* sobre *X*.

d) ¿A cuánto ascenderán las pérdidas un año en el que ha habido 83 días de lluvia?

e) Calcula el Error Cuadrático Medio e interpreta el resultado.

a) El diagrama de dispersión se observa a la derecha, (observar la escala de los ejes).
A la vista de la nube de puntos puede afirmarse que existe una relación lineal estadística entre las variables *X* e *Y*.



b)

x_j	y_j	x_j^2	y_j^2	$x_j y_j$
87	19,5	7569	380,25	1696,5
80	18,5	6400	342,25	1480
77	14,6	5929	213,16	1124,2
75	15,5	5625	240,25	1162,5
63	12,7	3969	161,29	800,1
71	14	5041	196	994
76	18,1	5776	327,61	1375,6
74	15	5476	225	1110
603	127,9	45785	2085,81	9742,9

$\overline{X} = \frac{603}{8} = 75,375$ días de lluvia ; $s_x^2 = \frac{45785}{8} - 75375^2 = 41,7344$

$\overline{Y} = \frac{127,9}{8} = 15,9875$ cientos de miles de euros ; $s_Y^2 = \frac{2085,81}{8} - 15,9875^2 = 5,1261$

$s_{XY} = \frac{9742,9}{8} - 73,375 \cdot 15,9875 = 12,8047$

El coeficiente de correlación es: $r = \frac{12,8047}{\sqrt{41,7344 \cdot 5,1261}} = 0,8754$, que confirma la impresión del apartado a), respecto a la relación lineal entre las variables *X* e *Y*.

c) La pendiente y la ordenada en el origen de la recta de regresión de *Y* sobre *X* son respectivamente:

$b = \frac{12,8047}{41,7344} = 0,3044$; $a = 15,9875 - 0,3068 \cdot 75,375 = -7,1386$

De manera que la recta de regresión lineal de *Y* sobre *X* es: $y = -7,1386 + 0,3068x$

d) Se calcula la pérdida esperada sustituyendo $x = 83$ en la ecuación de la recta de regresión de *Y* sobre *X*:

$$y(83) = -7,1386 + 0,3068 \cdot 83 = 18,3258 \text{ cientos de miles de euros}$$

e) El ECM de la recta de regresión de *Y* sobre *X* es: $ECM = 5,1261 \cdot (1 - 0,8754^2) = 1,1974$

Que es la media de los cuadrados de los residuos (distancia de cada valor observado a cada valor predicho por la recta de regresión) y que, en este caso es pequeño dado el buen ajuste de la recta de regresión a la distribución conjunta.

28. En un estudio de rendimiento de automóviles, se estudia la relación entre la potencia del motor en CV (*X*) y los kilómetros recorridos por litro de combustible (*Y*). Se recogieron datos de 10 vehículos:

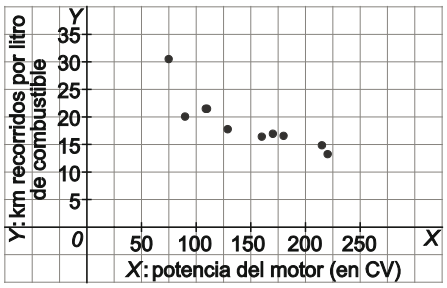
X	170	90	110	75	109	129	215	180	220	160
Y	17,0	20,1	21,5	30,4	21,5	17,8	14,9	16,6	13,3	16,5

a) Representa gráficamente los datos y explica razonadamente si existe relación entre las variables y, caso afirmativo, de qué tipo es.

b) Cuantifica la relación entre la potencia del motor y los km recorridos y explica, justificadamente, el resultado obtenido.

c) Escribe la ecuación de la recta de regresión que explica los km recorridos en función de la potencia del motor. ¿Ajusta bien la recta a la nube de puntos?

a) El diagrama de dispersión de la distribución conjunta de los datos se muestra a la derecha.



Parece que existe una buena relación lineal entre las variables y que es de tendencia decreciente: a más potencia del motor, menos km recorre el vehículo por litro de combustible.

b)

x_j	y_j	x_j^2	y_j^2	x_jy_j
170	17,0	28 900	289	2890
90	20,1	8100	404,01	1809
110	21,5	12 100	462,25	2365
75	30,4	5625	924,16	2280
109	21,5	11 881	462,25	2343,5
129	17,8	16 641	316,84	2296,2
215	14,9	46 225	222,01	3201,35
180	16,6	32 400	275,56	2986,2
220	13,3	48 400	176,89	2919,4
160	16,5	25 600	272,25	2640
1458,0	189,6	235 872,0	3805,22	25 730,7

$$\overline{X} = \frac{1458}{10} = 145,8\,CV \quad ; \quad s_x^2 = \frac{235872}{10} - 145,8^2 = 2329,56$$

$$\overline{Y} = \frac{189,6}{10} = 18,96 \text{ km/L} \quad ; \quad s_y^2 = \frac{3805,22}{10} - 18,96^2 = 21,04$$

$$s_{xy} = \frac{25730,7}{10} - 18,96 \cdot 145,8 = -190,574$$

$$R^2 = \frac{(-190,574)^2}{21,04 \cdot 2329,56} = 0,7410 \Rightarrow r = -\sqrt{0,7393} = -0,8608$$

Es decir, el 73,93 % de la variabilidad observada en los km recorridos por litro de combustible es explicado por la potencia del motor.

c) La pendiente y la ordenada en el origen de la recta de regresión de *Y* (km recorridos) sobre *X* (potencia del motor) son respectivamente

$$b = \frac{-190,754}{2329,54} = -0,082 \qquad a = 18,96 + 0,082 \cdot 145,8 = 30,882$$

La recta de regresión que se pide es: $y = 30,882 - 0,082x$

El coeficiente de correlación obtenido en el apartado anterior, señala un alto nivel de correlación lineal inversa entre ambas variables.

29. La tabla muestra el número de calzado (*Y*) y el peso en kg (*X*) de 8 chicos elegidos al azar en un centro educativo.

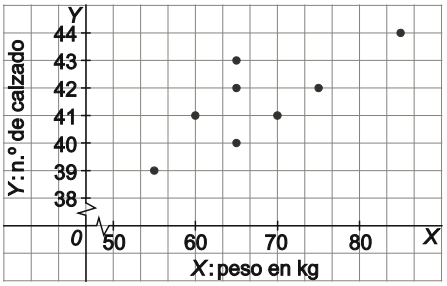
Y	39	40	41	41	42	42	43	44
X	55	65	60	70	65	75	65	85

a) Dibuja la nube de puntos.

b) Calcula los coeficientes de determinación y correlación

c) Valora la relación lineal que explica el número del calzado en función del peso.

a) La nube de puntos o diagrama de dispersión se muestra a la derecha, (obsérvese la escala de los ejes).



b)

x_j	y_j	x_j^2	y_j^2	x_jy_j
55	39	3025	1521	2145
65	40	4225	1600	2600
60	41	3600	1681	2460
70	41	4900	1681	2870
65	42	4225	1764	2730
75	42	5625	1764	3150
65	43	4225	1849	2795
85	44	7225	1936	3740
540	332	37 050	13 796	22 490

$$\overline{X} = \frac{540}{8} = 67,5\,\text{kg} \quad ; \quad s_x^2 = \frac{37050}{8} - 67,5^2 = 75$$

$$\overline{Y} = \frac{332}{8} = 41,5 \quad ; \quad s_y^2 = \frac{13796}{8} - 41,5^2 = 2,25$$

$$s_{XY} = \frac{22490}{8} - 41,5 \cdot 67,5 = 10$$

$$R^2 = \frac{10^2}{2,25 \cdot 75} = 0,5926 \Rightarrow r = \sqrt{0,5926} = 0,7698$$

c) De los resultados del apartado anterior se infiere que el 59,26 % de la variabilidad de los datos observados en el número del calzado es explicado por la variabilidad en el peso de los individuos.
Ello implica una correlación positiva entre el peso y el número de calzado, cuya intensidad se mide por el coeficiente de correlación lineal $r = 0,7698$; que puede considerarse medio – alto.
En conclusión, se observa una relación lineal moderada – alta entre el peso (*X*) y el número de calzado.

30. La tabla siguiente recoge el diámetro en cm (*Y*) del tronco de una determinada especie de árbol en 5 años (*X*) consecutivos.

X	1	2	3	4	5
Y	6,5	7,5	7,9	8,3	9,1

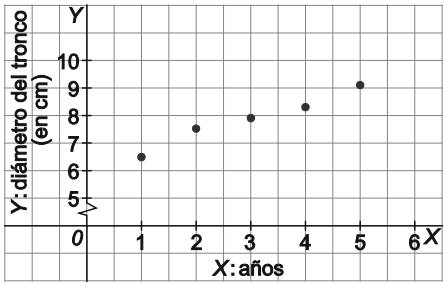
a) Dibuja el diagrama de dispersión y razona la tendencia que observas.

b) Escribe la recta de regresión de *Y* sobre *X* y estima el diámetro del tronco en un árbol en el sexto año.

c) ¿Qué porcentaje de la variabilidad observada en el diámetro se explica por la variabilidad en *X*?

d) Halla el Error Cuadrático Medio e interpreta el resultado.

a) El diagrama de dispersión o nube de puntos de la distribución bidimensional se muestra a la derecha.
Se observa con claridad una tendencia creciente con una asociación lineal fuerte entre ambas variables.



b)

x_j	y_j	x_j^2	y_j^2	$x_j y_j$
1	6,5	1	42,25	6,5
2	7,5	4	56,25	15,0
3	7,9	9	62,41	23,7
4	8,3	16	68,89	33,2
5	9,1	25	82,81	45,5
15	39,3	55	312,61	123,9

$$\overline{X} = \frac{15}{5} = 3 \text{ años} \quad ; \quad s_X^2 = \frac{55}{5} - 3^2 = 2$$

$$\overline{Y} = \frac{39,3}{5} = 7,86 \text{ cm} \quad ; \quad s_Y^2 = \frac{312,61}{5} - 7,86^2 = 0,7424$$

$$s_{XY} = \frac{123,9}{5} - 7,86 \cdot 3 = 1,20$$

Los coeficientes de la recta de regresión son:

$$b = \frac{1,20}{2} = 0,6 \quad ; \quad a = 7,86 - 0,6 \cdot 3 = 7,06$$

Y la recta de regresión tiene por ecuación: $y = 7,06 + 0,6x$.

Se calcula el diámetro esperado sustituyendo el valor $x = 6$ en la ecuación: $Y(6) = 7,06 + 0,6 \cdot 6 = 10,66$ cm.

c) Los coeficientes de determinación de correlación son:

$$R^2 = \frac{s_{XY}^2}{s_Y^2 s_X^2} = \frac{1,2^2}{0,7424 \cdot 2} = 0,9698 \Rightarrow r = \sqrt{0,9698} = 0,9848$$

Es decir, el 96,98 % de la variabilidad observada en el diámetro del tronco de los árboles viene explicada por la edad del árbol. Un porcentaje muy alto que confirma la fuerte relación observada entre ambas variables.

d) Para concluir, el Error Cuadrático Medio vendrá dado por:

$$\text{ECM} = s_Y^2 \left(1 - R^2\right) = 0,7424 \cdot \left(1 - 0,9698\right) = 0,0224$$

Su pequeño valor indica un buen ajuste de la recta a la nube de puntos.

31. El número de partidos ganados por 8 equipos de fútbol (*Y*) y el número de goles encajados por cada uno de ellos (*X*) se recoge en la tabla siguiente.

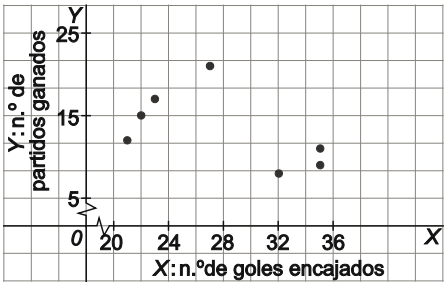
X	32	22	21	27	23	35	35	21
Y	8	15	12	21	17	9	11	12

a) Representa la nube de puntos y comenta la tendencia que observas justificando la respuesta.

b) Calcula el coeficiente de correlación e interpreta el resultado.

a) La nube de puntos o diagrama de dispersión se muestra a la derecha.

Parece una tendencia decreciente, a medida que aumenta el número de goles encajados (*X*) disminuye el número de partidos ganados (*Y*), si bien la relación que se observa es débil. La nube está claramente dividida en dos grupos y si se separaran las conclusiones no serían las mismas que con el conjunto total.



b)

x_j	y_j	x_j^2	y_j^2	$x_j y_j$
32	8	1024	64	256
22	15	484	225	330
21	12	441	144	252
27	21	729	441	567
23	17	529	289	391
35	9	1225	81	315
35	11	1225	121	385
21	12	441	144	252
216	105	6098	1509	2748

$$\overline{X} = \frac{216}{8} = 27 \text{ goles encajados} \quad ; \quad s_X^2 = \frac{6098}{8} - 27^2 = 33,25$$

$$\overline{Y} = \frac{105}{8} = 13,125 \text{ partidos ganados} \quad ; \quad s_Y^2 = \frac{1509}{8} - 13,125^2 = 16,3594$$

$$s_{XY} = \frac{2748}{8} - 13,125 \cdot 27 = -10,875$$

El coeficiente de correlación es:

$$r = \frac{-10,875}{\sqrt{16,3594 \cdot 33,25}} = -0,4663$$

Que confirma la relación negativa (cuando *X* crece, *Y* decrece) entre *X* e *Y*, si bien se trata de una relación lineal débil.

32. La tabla siguiente muestra la edad, en meses (*X*) y la altura en cm (*Y*) de una muestra de 10 niños tomada en una escuela.

X	126	132	116	140	115	135	141	137	119	121
Y	149	148	141	160	137	149	160	151	133	135

a) Representa la nube de puntos.

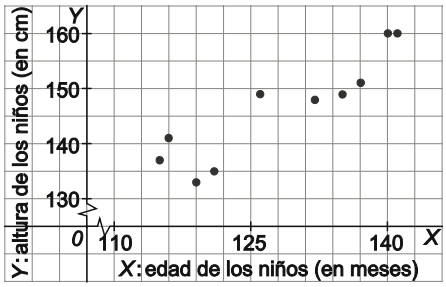
b) Halla la recta de regresión de la altura en función de la edad.

c) Calcula e interpreta el Error Cuadrático Medio.

d) Razona si la recta de regresión obtenida en b) representa adecuadamente los datos.

e) ¿Cuál sería la estatura de un niño de 12 meses?

a) En el diagrama de dispersión de esta distribución bidimensional puede observarse una tendencia creciente que puede ajustarse razonablemente por una recta.



b)

x_j	y_j	x_j^2	y_j^2	x_jy_j
126	149	15 876	22 201	18 774
132	148	17 424	21 904	19 536
116	141	13 456	19 881	16 356
140	160	19 600	25 600	22 400
115	137	13 225	18 769	15 755
135	149	18 225	22 201	20 115
141	160	19 881	25 600	22 560
137	151	18 769	22 801	20 687
119	133	14 161	17 689	15 827
121	135	14 641	18 225	16 335
1282	1463	165 258	214 871	188 345

$$\overline{X} = \frac{1282}{10} = 128,2 \text{ meses} \; ; \; s_x^2 = \frac{165258}{10} - 128,2^2 = 90,56$$

$$\overline{Y} = \frac{1463}{10} = 146,3 \text{ cm} \quad ; \; s_Y^2 = \frac{214871}{10} - 146,3^2 = 83,41$$

$$s_{XY} = \frac{188345}{10} - 146,3 \cdot 128,2 = 78,84$$

Los coeficientes de la recta de regresión de *Y* sobre *X*:

$$b = \frac{s_{XY}}{s_x^2} = \frac{78,84}{90,56} = 0,8709 \quad ; \quad a = \overline{Y} - b\overline{X} = 146,3 - 0,8709 \cdot 128,2 = 34,6913$$

De modo que la recta de regresión de la altura (cm) sobre la edad (meses) es: $y = 34,69 + 0,87x$.

c) El Error cuadrático medio de la regresión lineal de *Y* sobre *X* viene dado por:

$$\text{ECM} = s_Y^2\left(1 - \frac{s_{XY}^2}{s_Y^2 s_X^2}\right) = 83,41\left(1 - \frac{78,84^2}{83,41 \cdot 90,56}\right) = 14,7732$$

d) Los coeficientes de determinación y de correlación lineal son:

$$R^2 = \frac{s_{XY}^2}{s_Y^2 s_X^2} = \frac{78,84^2}{83,41 \cdot 90,56} = 0,8228 \Rightarrow r = \sqrt{0,8228} = 0,9071$$

Un porcentaje alto (82,28 %) de la variabilidad de la altura de los niños viene explicado por su edad, lo que supone un coeficiente de correlación lineal superior a 0,9. Por tanto, se puede afirmar que la recta de regresión se ajusta bastante bien a la nube de puntos y representa adecuadamente los datos.

e) El valor $x = 12$ meses no se encuentra dentro del rango de valores observados de la edad, sino que está muy alejado del mismo, por lo que no se puede hacer predicción de la altura correspondiente a este valor con la recta de regresión obtenida en el apartado b).

33. Como parte de un estudio sociológico, en un barrio de una gran ciudad, se recogieron en una muestra de 8 hogares los porcentajes del presupuesto familiar dedicados a gastos de alojamiento (*X*) y de ocio (*Y*).

X	13	15	19	24	14	17	21	17
Y	20	18	16	15	20	17	15	18

a) Determina la recta de regresión de *Y* sobre *X*.

b) Calcula el ECM y el coeficiente de correlación e interpreta los resultados.

c) Si se sabe que en un hogar el gasto en alojamiento es del 18 %, ¿Cuál sería el porcentaje de gasto esperado en ocio? Razona la fiabilidad de la predicción.

a) Para obtener la recta de regresión se construye la tabla siguiente:

x_j	y_j	x_j^2	y_j^2	$x_j y_j$
13	20	169	400	260
15	18	225	324	270
19	16	361	256	304
24	15	576	225	360
14	20	196	400	280
17	17	289	289	289
21	15	441	225	315
17	18	289	324	306
140	139	2546	2443	2384

$$\overline{X} = \frac{140}{8} = 17,5\,\% \text{ en alojamiento} \quad ; \quad s_X^2 = \frac{2546}{8} - 17,5^2 = 12$$

$$\overline{Y} = \frac{139}{8} = 17,375\,\% \text{ en ocio} \quad ; \quad s_Y^2 = \frac{2443}{8} - 17,375^2 = 3,484$$

$$s_{XY} = \frac{2384}{8} - 17,5 \cdot 17,375 = -6,0625$$

L coeficientes de la recta de regresión de *Y* sobre *X*

$$b = \frac{s_{XY}}{s_X^2} = \frac{-6,0625}{12} = -0,5052 \quad ; \quad a = \overline{Y} - b\overline{X} = 17,375 + 0,5052 \cdot 17,5 = 26,2161$$

De modo que la recta de regresión de la altura (cm) sobre la edad (meses) viene dada por:
$y = 18,924 - 0,0885x$.

b) El Error cuadrático medio de la regresión lineal de *Y* sobre *X* viene dado por

$$\text{ECM} = s_Y^2\left(1 - \frac{s_{XY}^2}{s_Y^2 s_X^2}\right) = 3,484\left(1 - \frac{(-6,0625)^2}{12 \cdot 3,484}\right) = 0,4212$$

Los coeficientes de determinación y de correlación lineal son:

$$R^2 = \frac{s_{XY}^2}{s_Y^2 s_X^2} = \frac{(-6,0625)^2}{12 \cdot 3,484} = 0,8791 \Rightarrow r = -\sqrt{0,0270} = -0,9373$$

Un alto porcentaje (87,91 %) de la variabilidad del gasto en ocio viene explicado por el gasto en alojamiento, lo que supone un coeficiente de correlación lineal inversa superior a 0,9. Por tanto, se puede afirmar que la recta de regresión se ajusta bastante bien a la nube de puntos y representa adecuadamente los datos.

c) El gasto esperado en ocio se calcula sustituyendo el valor $x = 18$ % en la ecuación de la recta:

$$y(18) = 18,924 - 0,0885 \cdot 18 = 17,331\,\% \text{ de gasto esperado en ocio}$$

Según lo expuesto en el apartado anterior se trata de una estimación altamente fiable.

34. En un control para prevenir la hipertensión en varones jóvenes, se elige una muestra de 8 con edades comprendidas entre 25 y 30 años y se mide su peso en kg (*X*) y su tensión sistólica en Hg/mg (*Y*).

X	74,8	81,6	70,3	95,3	67,6	78,2	76,2	88,5
Y	130	150	128	140	125	153	132	163

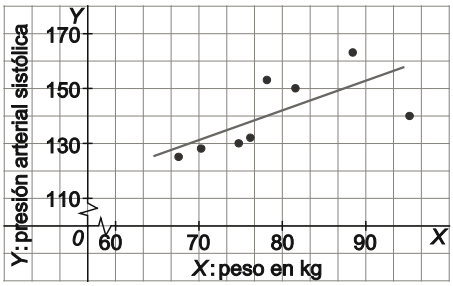
a) Representa gráficamente los datos.

b) Justifica, a la vista de la nube de puntos, que es razonable ajustar una recta regresión de *Y* sobre *X* y calcula los coeficientes de determinación y de correlación.

c) Escribe la ecuación de la recta de regresión de *Y* sobre *X*.

¿Cuál es la tensión sistólica que se estima que pueda tener un joven con 80 kg de peso?

a) La nube de puntos de la distribución bidimensional junto con la recta de regresión ajustada en el apartado c).



b) A la vista de la nube de puntos, y de la tendencia creciente observada, podrá ajustarse una recta de regresión, si bien el ajuste no es muy bueno, debido sobre todo a la observación (95,3; 140).

x_j	y_j	x_j^2	y_j^2	$x_j y_j$
74,8	130	5595,04	16 900	9724,0
81,6	150	6658,56	22 500	12 240,0
70,3	128	4942,09	16 384	8998,4
95,3	140	9082,09	19 600	13 342,0
67,6	125	4569,76	15 625	8450,0
78,2	153	6115,24	23 409	11 964,6
76,2	132	5806,44	17 424	10 058,4
88,5	163	7832,25	26 569	14 425,5
632,5	1121	50 601,47	158 411	89 202,9

$$\overline{X} = \frac{632,5}{8} = 79,063\,\text{kg} \quad ; \quad \overline{Y} = \frac{1121}{8} = 140,125$$

$$s_X^2 = \frac{50601,47}{8} - 79,063^2 = 74,3048 \quad ; \quad s_Y^2 = \frac{158411}{8} - 140,125^2 = 166,3594$$

$$s_{XY} = \frac{89202,9}{8} - 140,125 \cdot 79,063 = 71,7297$$

Los coeficientes de determinación y correlación son:

$$R^2 = \frac{s_{XY}^2}{s_Y^2 s_Y^2} = \frac{71,7297^2}{166,3594 \cdot 74,3048} = 0,4162 \Rightarrow r = \sqrt{0,4162} = 0,6452$$

que confirman una relación lineal positiva (tendencia creciente), aunque solo moderada.

c) Con los resultados obtenidos en el apartado anterior, se pueden obtener los coeficientes de la recta de regresión de *Y* (tensión arterial) sobre *X* (peso en kg).

$$b = \frac{s_{XY}}{s_X^2} = \frac{71,7297}{74,3048} = 0,9653 \quad ; \quad a = 140,125 - 0,9653 \cdot 79,0625 = 63,8025$$

La ecuación de la recta de regresión de *Y* sobre *X* es: $y = 63,8025 + 0,9653x$.

d) Si $x = 80$ kg, como está dentro del rango de valores observados del peso (*X*), se puede utilizar la recta anterior para predecir la tensión arterial correspondiente:

$$y(80) = 63,8025 + 0,9653 \cdot 80 = 141,03$$

35. En la siguiente tabla se muestran las distancias recorridas por un vehículo (*Y*) que se ha movido con aceleración constante durante 10 segundos, en función del tiempo transcurrido (*X*).

X	2	4	6	8	10
Y	25	95	216	530	598

a) Representa gráficamente los datos.

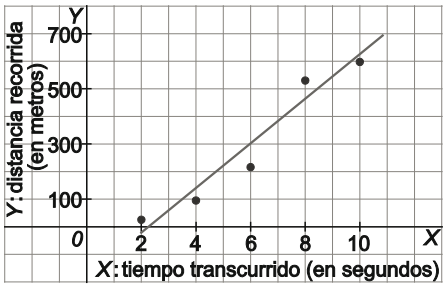
b) Calcula la recta de regresión lineal, el coeficiente de correlación y el Error Cuadrático Medio.

c) Como en el instante cero (X=0) el espacio que ha recorrido el vehículo es cero (Y=0), ajusta la nube de puntos a una recta que pasa por el origen.

d) Define una nueva variable $Z = X^2$, y realiza ahora el ajuste lineal de Y en función de Z, calculando el nuevo coeficiente de correlación.

e) Compara y valora los resultados obtenidos en los apartados b, c y d.

a) Se representa el diagrama de dispersión junto con la recta de regresión que se calcula en el apartado b)



Puede observarse la relación con tendencia creciente entre ambas variables.

b)

x_j (s)	y_j (m)	x_j^2	y_j^2	$x_j y_j$
2	25	4	625	50
4	95	16	9025	380
6	216	36	46 656	1296
8	530	64	280 900	4240
10	598	100	357 604	5980
30	1464	220	694 810	11 946

$$\overline{X} = \frac{30}{5} = 6 \text{ s.} \quad ; \quad s_x^2 = \frac{220}{5} - 6^2 = 8$$

$$\overline{Y} = \frac{1464}{5} = 292,8 \text{ m} \quad ; \quad s_Y^2 = \frac{694\,810}{5} - 292,8^2 = 53\,230,16$$

$$s_{XY} = \frac{11946}{5} - 292,8 \cdot 6 = 632,4$$

Los coeficientes de la ecuación de regresión son:

$$b = \frac{s_{XY}}{s_x^2} = \frac{632,4}{8} = 79,05 \quad ; \quad a = \overline{Y} - b\overline{X} = 292,8 - 79,05 \cdot 6 = -181,5$$

De manera que la ecuación de la recta de regresión de Y sobre X es: $y = -181,5 + 79,05x$.

Debe observarse que esta ecuación de la recta de regresión estima un valor negativo para la distancia recorrida cuando x=2, lo que obviamente no es una buena estimación.
El coeficiente de correlación entre X e Y es:

$$r_{XY} = \frac{s_{XY}}{\sqrt{s_x^2 s_Y^2}} = \frac{632,4}{\sqrt{8 \cdot 53230,16}} = 0,9691$$

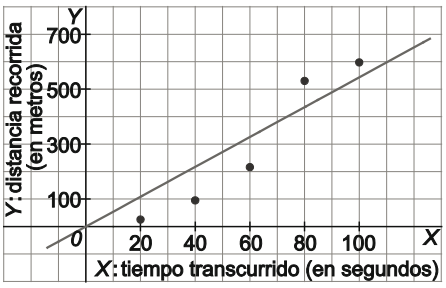
que indica una fuerte relación lineal entre las variables, con el inconveniente señalado antes.

c) El coeficiente de la recta de regresión que pasa por el origen viene dado por:

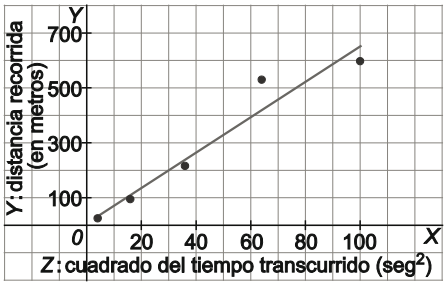
$$b = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{11946}{220} = 54,3$$

de donde la ecuación de la recta es:

$y = 54,3x$



d) Sea $Z = X^2$, entonces el diagrama de dispersión de las variables Z e *Y* junto con la recta de regresión de *Y* sobre Z señala que este ajuste mejora el anterior (aunque ligeramente, porque el otro ya es muy bueno).



La tabla con los valores de *Y* y *Z*, junto con las columnas necesarias para facilitar los cálculos de varianzas y covarianza es:

y (m)	Z	y^2	z^2	*yz*
25	4	625	16	100
95	16	9025	256	1520
216	36	46 656	1296	7776
530	64	280 900	4096	33 920
598	100	357 604	10 000	59 800
1464	220	694 810	15 664	103 116

$$\overline{Z} = \frac{220}{5} = 44 \quad ; \quad s_Z^2 = \frac{15664}{5} - 44^2 = 1196,8$$

$$\overline{Y} = \frac{1464}{5} = 292,8 \; ; \quad s_Y^2 = \frac{694\,810}{5} - 292,8^2 = 53\,230,16$$

$$s_{ZY} = \frac{103\,116}{5} - 292,8 \cdot 44 = 7740$$

Los coeficientes de la recta de regresión son:

$$b = \frac{s_{ZY}}{s_Z^2} = \frac{7740}{1196,8} = 6,467 \quad ; \quad a = \overline{Y} - b\overline{Z} = 292,8 - 6,467 \cdot 44 = 8,241$$

De manera que la ecuación de la recta de regresión de *Y* sobre *Z* es: $y = 8,241 + 6,467z$.

Y el coeficiente de correlación entre *Z* e *Y* es:

$$r_{ZY} = \frac{s_{ZY}}{\sqrt{s_Z^2 s_Y^2}} = \frac{7740}{\sqrt{1196,8 \cdot 53230,16}} = 0,9697$$ que es prácticamente igual que el de *X* e *Y*.

e) Los diagramas de dispersión muestran una buena relación lineal en los tres casos, ligeramente mejor en el caso de la variables Z e *Y*. Sin embargo, el coeficiente de correlación en el caso de la variables Z e *Y*, (0,9697), es prácticamente igual que en el caso de las variables *X* e *Y*, (0,9691), lo que significa que en los dos caso el ajuste lineal es similar y muy bueno.

Síntesis

36. Con los datos de la tabla siguiente:

X	2	3	6	10	12
Y	150	130	125	120	100

a) Dibuja el diagrama de dispersión.

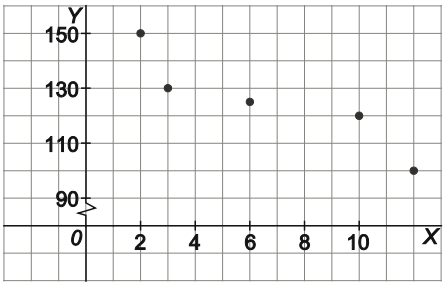
b) Halla la media y la varianza de las distribuciones X e Y.

c) Obtén la recta de regresión de Y sobre X y valora la bondad del ajuste.

d) ¿Qué porcentaje de la variabilidad de Y viene explicado por la variabilidad de X?

e) ¿Qué valor se espera en la variable Y si la variable X toma el valor $x=5$?

a) Se representa el diagrama de dispersión a la derecha.



b)

x_j	y_j	x_j^2	y_j^2	x_jy_j
2	150	4	22500	300
3	130	9	16900	390
6	125	36	15625	750
10	120	100	14400	1200
12	100	144	10000	1200
33	625	293	79425	3840

$$\overline{X} = \frac{33}{5} = 6,6 \;\; ; \;\; s_X^2 = \frac{293}{5} - 6,6^2 = 15,04$$

$$\overline{Y} = \frac{625}{5} = 125; \;\; s_Y^2 = \frac{79425}{5} - 125^2 = 260$$

c) Se obtiene, en primer lugar, la covarianza: $s_{XY} = \frac{3840}{5} - 125 \cdot 6,6 = -57$, que junto con los resultados del apartado b) proporcionan los coeficientes de la recta de regresión de Y sobre X.

$$b = \frac{s_{XY}}{s_X^2} = \frac{-57}{15,04} = -3,79 \;\; ; \;\; a = \overline{Y} - b\,\overline{X} = 125 + 3,79 \cdot 6,6 = 150,01$$

La recta de regresión de Y sobre X viene dada por la ecuación: $y = 150,01 - 3,79x$

La bondad del ajuste se puede valorar mediante el cálculo del coeficiente de correlación lineal

Los coeficientes de determinación y de correlación lineal son:

$$R^2 = \frac{s_{XY}^2}{s_Y^2 s_X^2} = \frac{(-57)^2}{260 \cdot 15,04} = 0,8309 \Rightarrow r = -\sqrt{0,8309} = -0,9115$$ (idéntico signo que s_{XY})

Se trata, por tanto de un muy buen ajuste lineal, que confirma la observación del diagrama de dispersión del apartado a).

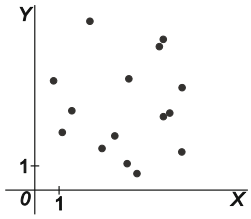
d) Observando el coeficiente de determinación, el 83,09% de la variabilidad observada en la variable Y viene dada por la variabilidad de X.

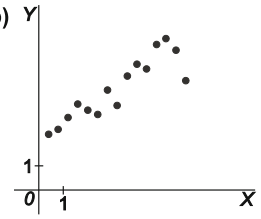
e) Por último, para estimar el valor esperado cuando X toma el valor $x = 5$, se sustituye este valor en la ecuación de la recta de regresión:

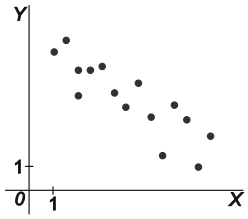
$$y = 150,01 - 3,79 \cdot 5 = 131,05$$

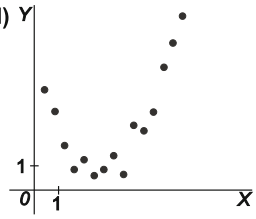
CUESTIONES

37. En los siguientes casos, se representa la nube de puntos de una variable bidimensional.

a) 

b) 

c) 

d) 

En cada caso indica si existe relación lineal entre las variables y, en caso afirmativo, ¿cuál es el signo de la covarianza y del coeficiente de correlación?

a) En este caso no parece que exista relación lineal entre las variables representadas.

b) Se puede observar una relación lineal directa, cuando una variable crece la otra también, y por lo tanto los signos de la covarianza y del coeficiente de correlación son positivos.

c) Puede observarse una relación lineal con tendencia inversa (cuando una variable crece la otra decrece). En este caso tanto la covarianza como el coeficiente de correlación lineal tienen signo negativo.

d) Parece que existe relación entre ambas variables, pero que esta no es lineal.

38. De una variable bidimensional (*X*,*Y*) se sabe que la ecuación de la recta de regresión de *Y* sobre *X*. es *y* = 3. Contesta razonadamente

a) ¿Cuál es la media de *Y*?

b) ¿Cuál es el valor de la covarianza?

c) ¿Cuánto vale el coeficiente de correlación?

d) ¿Qué conclusiones se pueden extraer?

a) Dado que la ecuación de la recta de regresión es $y = \overline{Y} + \frac{s_{XY}}{s_X^2}(x - \overline{X})$ y comparando con $y = 3 \Rightarrow \overline{Y} = 3$.

b) La covarianza es $s_{XY} = 0$, porque la pendiente de la recta es $b = \frac{s_{XY}}{s_X^2} = 0 \Rightarrow s_{XY} = 0$.

c) El coeficiente de correlación es cero, porque la covarianza es cero y $r = \frac{s_{XY}}{\sqrt{s_X^2 \; s_X^2}} = 0$.

d) No existe relación lineal entre las variables *X* e *Y*.

39. De dos variable estadísticas, *X* e *Y*, se sabe que:

- **Tienen las varianzas iguales.**
- **El coeficiente de correlación toma el valor 0,8.**
- **La covarianza es 2,8.**
- **La recta de regresión contiene al punto (3, 7).**

Halla la recta de regresión de *Y* sobre *X* y el Error Cuadrático Medio.

Sea $y = a + bx$ la ecuación de la recta de regresión de *Y* sobre *X*. El objetivo es calcular los coeficientes *a* y *b*.

Como las varianzas de *X* e *Y* son iguales, se tiene que: $s_X^2 = s_Y^2 \Rightarrow r = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}} = \frac{s_{XY}}{\sqrt{s_Y^2 s_Y^2}} = \frac{s_{XY}}{s_Y^2} \Rightarrow s_Y^2 = \frac{2,8}{0,8} = 3,5$

Además, en este caso: $b = \frac{s_{XY}}{s_X^2} = \frac{s_{XY}}{s_Y^2} = r = 0,8$

Para calcular la ordenada en el origen, se tiene en cuenta que $y(3) = 7 \Rightarrow 7 = a + 0,8 \cdot 3 \Rightarrow a = 4,6$.

Por lo que la ecuación de la recta de regresión de *Y* sobre *X* es: $y = 4,6 + 0,8x$.

Para terminar, el ECM viene determinado por: $\text{ECM} = s_Y^2\left(1 - R^2\right) = 3,5\left(1 - 0,8^2\right) = 1,26$.

40. La recta de regresión de una variable *Y* sobre otra variable *X* está dada por la ecuación $y = -2{,}3 + 0{,}15x$. Señala, de forma razonada, cuál o cuáles de las siguientes afirmaciones son ciertas o falsas

a) El coeficiente de correlación es 0,15.

b) La covarianza entre *X* e *Y* es positiva.

c) La variable *X* no explica en absoluto el comportamiento de la variable *Y*.

d) Estas dos variables están débilmente correlacionadas.

a) De la información obtenida de la recta de regresión: $b = \frac{s_{XY}}{s_X^2} = 0{,}15 \Rightarrow r = \frac{s_{XY}}{\sqrt{s_Y^2 s_X^2}} = \frac{b s_X^2}{\sqrt{s_Y^2 s_X^2}} = 0{,}15\sqrt{\frac{s_X^2}{s_Y^2}}$

Por tanto la afirmación es falsa, salvo que las varianzas de *X* e *Y* sean iguales.

b) Sí, ya que el signo del coeficiente de correlación coincide con el de la covarianza y con el de *b*.

c) Para valorar esta afirmación hay que calcular el coeficiente de determinación (o el de correlación) y, con la información disponible, no se puede obtener. Por lo tanto, la afirmación no es correcta en general.

d) La respuesta es similar a la del apartado c).

41. De la variable estadística bidimensional (*X*, *Y*), se sabe que:

- **Las medias son $\overline{X} = 9{,}2$ e $\overline{Y} = 7{,}5$.**
- **La desviación típica de la variable *Y* es el doble que la de la variable *X*.**
- **El coeficiente de correlación toma el valor 0,7.**

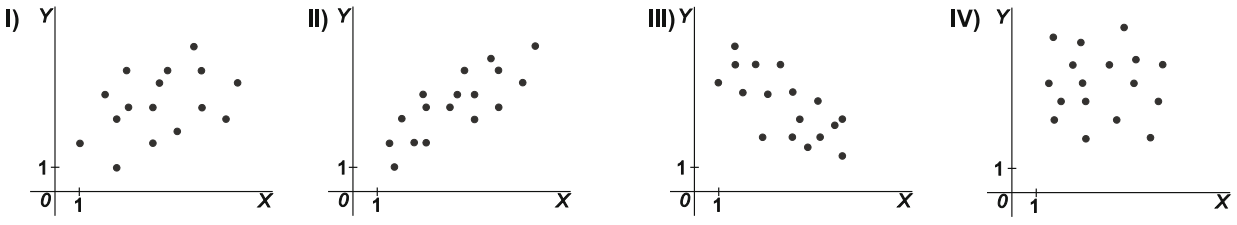
Con estos datos, determina la recta de regresión de *Y* sobre *X*.

De la información proporcionada se sabe que $s_Y^2 = 2s_X^2 \Rightarrow r = \frac{s_{XY}}{\sqrt{s_Y^2 s_X^2}} = \frac{s_{XY}}{\sqrt{s_X^2 s_X^2}} = \frac{s_{XY}}{s_X^2\sqrt{2}} \Rightarrow \frac{s_{XY}}{s_X^2} = \sqrt{2}\, r$

De donde: $b = \frac{s_{XY}}{s_X^2} = \sqrt{2} \cdot r = 0{,}7 \cdot \sqrt{2} \simeq 0{,}9899$; $a = \overline{Y} - \frac{s_{XY}}{s_x^2} \cdot \overline{X} = 7{,}5 - 0{,}7 \cdot \sqrt{2} \cdot 9{,}2 \simeq -1{,}6075$

Por tanto la ecuación de la recta de regresión es: $y = -1{,}6075 + 0{,}9899x$

42. A la vista de las siguientes nubes de puntos de dos distribuciones conjuntas bidimensionales, asigna el coeficiente de correlación que mejor se aproxime a cada una de las distribuciones siguientes:



a) $r = -0{,}04$ **b)** $r = 0{,}4$ **c)** $r = -0{,}7$ **d)** $r = 0{,}8$

a) $r = -0{,}04$ se asigna al diagrama de dispersión IV, puesto que no se observa tendencia en la nube de puntos.

b) $r = 0{,}4$ se asigna al diagrama I, ya que se observa un tendencia creciente, con relación lineal débil entre las variables.

c) $r = -0{,}7$ se asigna al diagrama III, en el que se observa una tendencia decreciente, con relación lineal moderada entre las variables.

d) $r = 0{,}8$ se asigna al diagrama de dispersión II, en el que se observa una tendencia lineal creciente con relación lineal moderada–alta.

PROBLEMAS

43. La tabla recoge la distribución de los alumnos de primero de bachillerato según sexo (*X*) y grupo (*Y*).

X \ *Y*	Grupo *A*	Grupo *B*	Grupo *C*	Totales *X*
Chicos	**	18	**	49
Chicas	21	**	16	**
Totales *Y*	**	**	32	103

a) Copia y completa la tabla.

b) Dentro del grupo *B*, ¿cuál es el porcentaje de mujeres?

c) Escribe las distribuciones marginales de frecuencias absolutas y relativas.

a) Completamos la tabla con los datos que faltan:

X \ *Y*	Grupo *A*	Grupo *B*	Grupo *C*	Totales *X*
Chicos	**15**	18	**16**	49
Chicas	21	**17**	16	**54**
Totales *Y*	**36**	**35**	32	103

b) $p(chicas \mid B) = \frac{17}{35} \cdot 100 = 48{,}57\ \%$

c) Las distribuciones marginales de frecuencias absolutas y relativas se recogen en la tabla

X \ *Y*	Grupo *A*	Grupo *B*	Grupo *C*	f_{Xi}	h_{Xi}
Chicos	15	18	16	49	0,4757
Chicas	21	17	16	54	0,5243
f_{Yj}	36	35	32	103	
h_{Yj}	0,3495	0,3398	0,3107		

44. La tabla muestra los datos de la tasa de mortalidad infantil en tanto por mil (*Y*), y el porcentaje del Producto Interior Bruto dedicado la salud (*X*). La muestra corresponde a 10 países de diferentes continentes y riqueza.

X	5,2	4,5	5,2	6,3	5,4	6,9	3,7	6,7	11,0	9,5
Y	110	94	77	56	36	16	12	8	5	3

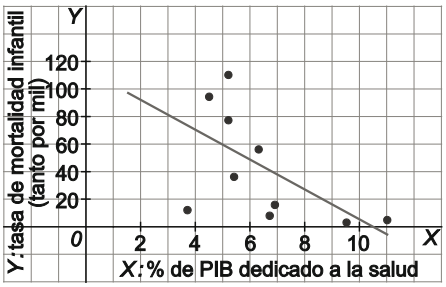
a) Representa la nube de puntos. ¿Existe relación entre ambas variables? ¿De qué tipo?

b) Cuantifica la relación existente y coméntala.

c) Escribe la ecuación de la recta de regresión de la tasa de mortalidad infantil en función del porcentaje del PIB dedicado a sanidad.

d) En un país que invierta un 5 % del PIB en sanidad, ¿Cuál será la tasa de mortalidad infantil esperada? Comenta la fiabilidad de esta estimación.

a) A la derecha se muestra el diagrama de dispersión de la distribución conjunta de ambas variables, junto con la recta de regresión de *Y* sobre *X* obtenida en el apartado c):
Puede verse que existe una relación lineal débil de tendencia decreciente: a mayor porcentaje del PIB dedicado a la salud, menor es la tasa de mortalidad infantil.



b)

x_j	y_j	x_j^2	y_j^2	$x_j y_j$
5,2	110	27,0	12100,0	572
4,5	94	20,3	8836,0	423
5,2	77	27,0	5929,0	400,4
6,3	56	39,7	3136,0	352,8
5,4	36	29,2	1296,0	194,4
6,9	16	47,6	256,0	110,4
3,7	12	13,7	144,0	44,4
6,7	8	44,9	64,0	53,6
11,0	5	121,0	25,0	55
9,5	3	90,3	9,0	28,5
64,4	417,0	460,6	31795,0	2234,5

$$\overline{X} = \frac{64,4}{10} = 6,44 \text{ \% del PIB} \quad ; \quad s_x^2 = \frac{460,6}{10} - 6,44^2 = 4,5884$$

$$\overline{Y} = \frac{417}{10} = 41,7 \text{ por mil de los habitantes} \quad ; \quad s_Y^2 = \frac{31795}{10} - 41,7^2 = 1440,61$$

$$s_{XY} = \frac{2234,5}{10} - 41,7 \cdot 6,44 = -45,098$$

Los coeficientes de determinación y correlación son:

$$R^2 = \frac{s_{XY}^2}{s_Y^2 s_X^2} = \frac{(-45,098)^2}{1440,61 \cdot 4,5884} = 0,3077 \Rightarrow r = -\sqrt{0,3077} = -0,5547$$

Lo que significa que el 30,77% de la variabilidad observada en la tasa de mortalidad infantil viene explicada por el porcentaje de PIB dedicado a salud (*X*). Se confirma la relación lineal (moderada) con tendencia negativa (inversa) entre ambas variables

c) Con los datos obtenidos en el apartado b), se estiman los coeficiente de la recta de regresión de la tasa de mortalidad infantil (*Y*) en función del porcentaje del PIB dedicado a la salud:

$$b = \frac{s_{XY}}{s_x^2} = \frac{-45,098}{4,5884} = -9,829 \quad ; \quad a = \overline{Y} - b\overline{X} = 41,7 + 9,829 \cdot 6,44 = 104,997$$

De modo que la recta de regresión de *Y* sobre *X* es: $y = 104,997 - 9,829x$

d) Si el porcentaje del PIB dedicado a salud es $x = 5$, entonces, la tasa esperada de mortalidad infantil es

$$y(5) = 104,997 - 9,829 \cdot 5 = 55,85 \text{ por mil habitantes}$$

Predicción de relativa fiabilidad dado que el coeficiente de correlación no es muy alto.

45. Se conoce que el consumo de energía anual por habitante (*Y*, en miles de kWh) está relacionado con la renta per cápita (*X*, en miles de $). Para estudiar cómo funciona esta relación en Centroamérica se han recogido los datos en la siguiente tabla:

X	8647	3708	3178	2246	10047	1578	7498
Y	1855	855	567	671	1990	473	1832

a) Representa el diagrama de dispersión.

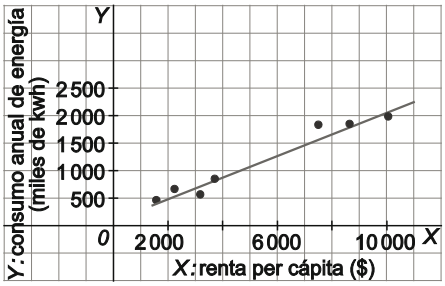
b) ¿Puede aproximarse, razonablemente, la nube de puntos por una recta?

c) Escribe la ecuación de la recta de regresión del consumo de energía sobre la renta per cápita.

d) Calcula el porcentaje de variabilidad en el consumo de energía explicada por la renta per cápita. Valora el resultado.

e) Calcula el consumo esperado de energía en un país cuya renta per cápita sea de 5000 $. Justifica la fiabilidad de la predicción.

a) Se dibuja el diagrama de dispersión de la distribución conjunta de ambas variables, junto con la recta de regresión de *Y* sobre *X* obtenida en el apartado c):



b) La nube de puntos parece mostrar una fuerte relación lineal creciente entre la renta per cápita y el consumo de energía.

c)

x_j	y_j	x_j^2	y_j^2	$x_j y_j$
8647	1855	74 770 609	3 441 025	16 040 185
3708	855	13 749 264	731 025	3 170 340
3178	567	10 099 684	321 489	1 801 926
2246	671	5 044 516	450 241	1 507 066
10047	1990	100 942 209	3 960 100	19 993 530
1578	473	2 490 084	223 729	746 394
7498	1832	56 220 004	3 356 224	13 736 336
36 902	8243	263 316 370	12 483 833	56 995 777

$$\overline{X} = \frac{36\ 902}{7} = 5271,71\ \$ \quad ; \quad s_x^2 = \frac{263\ 316\ 370}{7} - 5271,71^2 = 9\ 825\ 652,775$$

$$\overline{Y} = \frac{8243}{7} = 1177,57 \text{ miles de kWh} \quad ; \quad s_Y^2 = \frac{12\ 483\ 833}{7} - 1177,57^2 = 396\ 739,2449$$

$$s_{XY} = \frac{56\ 995\ 777}{7} - 5271,71 \cdot 1177,57 = 1\ 934\ 446,312$$

Los coeficientes de la recta de regresión son:

$$b = \frac{s_{XY}}{s_x^2} = \frac{1\ 934\ 446,312}{9\ 825\ 652,775} = 0,1969 \quad ; \quad a = \overline{Y} - b\overline{X} = 1177,57 - 0,1969 \cdot 5271,71 = 139,57$$

De modo que la recta de regresión de *Y* sobre *X* es: $y = 139,57 + 0,1969x$

d) A continuación, se obtienen los coeficientes de determinación y correlación:

$$R^2 = \frac{s_{XY}^2}{s_Y^2 s_X^2} = \frac{1\ 934\ 446,312^2}{396\ 730,245 \cdot 9\ 825\ 652,775} = 0,95997 \Rightarrow r = 0,9798$$

Lo que significa que el 95,99 % de la variabilidad observada en el consumo de energía viene explicada por la renta per cápita. Se confirma así la fuerte relación lineal entre ambas variables.

e) Para estimar el consumo esperado de energía en un país cuya renta per cápita sea 5000 $, dato que se encuentra dentro del rango de estudio, evaluamos en la ecuación:

$$y(5000) = 139,57 + 0,1969 \cdot 5000 = 1124,07 \text{ miles de kWh.}$$

Predicción con una alta fiabilidad dado que el coeficiente de correlación es muy alto.

46. La tabla siguiente muestra los datos relativos al número de salas de cine (*X*) y de espectadores en millones (*Y*), en España entre el 2004 y el 2011.

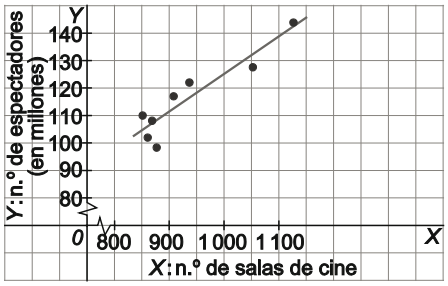
Años	2004	2005	2006	2007	2008	2009	2010	2011
X	1126	1052	936	907	868	851	860	876
Y	143,9	127,6	122	117	108	110	102	98,3

a) Mediante la representación gráfica ¿qué relación se observa entre las dos variables?

b) ¿Cuál de las dos variables presenta mayor variabilidad?

c) ¿Cuántos espectadores se esperan si hubiera 1000 salas de cine. Valora la precisión de esta estimación.

a) Se muestra a la derecha el diagrama de dispersión de la distribución bidimensional, junto con la recta de regresión calculada en el apartado c). Se observa un tendencia creciente con fuerte relación lineal.



b)

x_j	y_j	x_j^2	y_j^2	$x_j y_j$
1126	143,9	1 267 876	20 707,21	162 031,4
1052	127,6	1 106 704	16 281,76	134 235,2
936	122,0	876 096	14 884,00	114 192,0
907	117,0	822 649	13 689,00	106 119,0
868	108,0	753 424	11 664,00	93 744,0
851	110,0	724 201	12 100,00	93 610,0
860	102,0	739 600	10 404,00	87 720,0
876	98,3	767 376	9662,89	86 110,8
7476,0	928,8	7 057 926,0	109 392,9	877 762,4

$\overline{X} = \dfrac{7476}{8} = 934,5$ salas de cine ; $s_X^2 = \dfrac{7057926}{8} - 934,5^2 = 8950,5 \Rightarrow s_X = 94,6071$

$\overline{Y} = \dfrac{928,8}{8} = 116,1$millones de espectadores ; $s_Y^2 = \dfrac{109392,9}{8} - 116,1^2 = 194,8975 \Rightarrow s_Y = 13,9606$

Luego, los coeficientes de variación de *X* e *Y* son:

$CV(Y) = \dfrac{s_Y}{\overline{Y}} = \dfrac{13,9606}{116,1} = 0,1202$ $\qquad$ $CV(X) = \dfrac{s_X}{\overline{X}} = \dfrac{94,6071}{934,5} = 0,1012$

Se concluye que el número de espectadores (en millones) presenta una variabilidad ligeramente mayor (aprox. un 18,78 % más) que la del número de salas de cine (*X*).

c) Para estimar el número de espectadores esperado, se calcula la recta de regresión de *Y* sobre *X*:

$s_{XY} = \dfrac{877762,4}{8} - 116,1 \cdot 934,5 = 1224,85$ $\quad b = \dfrac{1224,85}{8950,5} = 0,1368$ $\quad a = 116,1 - 0,1368 \cdot 934,5 = -11,7836$

Luego la ecuación de la recta de regresión del número de espectadores (*Y*) en función del número de salas de cine (*X*) es: $y = -11,7836 + 0,1368x$.

Haciendo uso de ella se puede estimar el número de espectadores (*Y*, en millones) si el número de salas de cine fuera x=1000 (dentro del rango de valores observados de *X*):

$$y(1000) = -11,7836 + 0,1368 \cdot 1000 = 125,0635 \text{ millones de espectadores.}$$

Para estimar la precisión calculamos el coeficiente de correlación: $r = \dfrac{s_{XY}}{\sqrt{s_Y^2 \, s_X^2}} = \dfrac{1224,85}{94,6071 \cdot 13,9606} = 0,9274$

Por tanto se trata de una estimación bastante fiable.

47. La distribución conjunta de la superficie en metros cuadrados (*Y*) de una vivienda el número de habitaciones (*X*) viene dada en la tabla siguiente:

		Y: Superficie, en m^2			
		[60, 70)	[70, 80)	[80, 90)	[90,100)
X: Número habitaciones	2	69	12	2	1
	3	464	217	89	26
	4	175	450	212	138

a) Halla las distribuciones marginales.

b) Escribe la distribución de la superficie sabiendo que la vivienda dispone de tres habitaciones. Halla su media y su varianza.

c) Calcula la covarianza e interpreta el resultado.

a) Las tablas con las distribuciones marginales son:

X_j	f_{Xj}	h_{Xj}	$f_{Yj} \cdot X_j$
2	84	0,0453	168
3	796	0,4291	2388
4	975	0,5256	3900
	1855	1	6456

	Y_j	f_{Yj}	h_{Yj}	$f_{Yj} \cdot Y_j$
[60, 70)	65	708	0,3817	46 020
[70, 80)	75	679	0,3660	50 925
[80, 90)	85	303	0,1633	25 755
[90,100)	95	165	0,0890	15 675
		1855	1	138 375

b) La distribución para una vivienda de 3 habitaciones, ampliada para los cálculos posteriores es:

	[60, 70)	[70, 80)	[80, 90)	[90,100)	
Y_j	65	75	85	95	
$f_{Y\|X=3}$	464	217	89	26	796
$f_{Yj} \cdot Y_j$	30 160	16 275	7565	2470	56 470
$f_{Yj} \cdot Y_j^2$	1 960 400	1 220 625	643 025	234 650	4 058 700

Por tanto, la media y la varianza serán:

$$\overline{Y}\,|_{X=3} = \frac{56470}{796} = 70,94\,m^2 \;;\; s^2_{Y|X=3} = \frac{4058700}{796} - 70,94^2 = 66,07$$

c) Construimos una tabla auxiliar con los productos necesarios $f_{ij} \cdot X_i Y_j$:

Y_j	65	65	65	75	75	75	85	85	85	95	95	95	
X_i	2	3	4	2	3	4	2	3	4	2	3	45	
f_{ij}	69	464	175	12	217	450	2	89	212	1	26	138	
$f_{ij}X_iY_j$	8970	90 480	45 500	1800	48 825	135 000	340	22 695	72 080	190	7410	52 440	485 730

Las medias marginales son:

$$\overline{X} = \frac{6456}{1855} = 3,48 \quad ; \quad \overline{Y} = \frac{138375}{1855} = 74,60$$

de donde la covarianza será:

$$s_{XY} = \frac{485\,730}{1855} - 3,48 \cdot 74,60 = 2,241$$

Al tratarse de una covarianza positiva la relación existente entre ambas variables es directa, es decir, a mayor número de habitaciones, mayor superficie.

48. Las calificaciones (*Y*) de 8 alumnos en Econometría I en el primer curso de Grado en CC. Económicas y las obtenidas en la materia de economía de la empresa en la PAU (*X*), han sido:

Y	5,3	6,2	6,8	7,2	10	5,1	3,8	7,5
X	7,3	5,2	6	7	8,8	4	5,7	8,2

a) ¿Cuál es el grado de correlación entre las dos variables? Valora el resultado.

b) Si un alumno obtuvo 6,5 puntos en el examen de la PAU ¿Qué calificación se espera que obtenga en Econometría I? Comenta la precisión de esta estimación.

a) Se construye la tabla para facilitar los cálculos correspondientes:

x_j	y_j	x_j^2	y_j^2	$x_j y_j$
7,3	5,3	53,29	28,09	38,69
5,2	6,2	27,04	38,44	32,24
6,0	6,8	36,00	46,24	40,80
7,0	7,2	49,00	51,84	50,40
8,8	10,0	77,44	100,00	88,00
4,0	5,1	16,00	26,01	20,40
5,7	3,8	32,49	14,44	21,66
8,2	7,5	67,24	56,25	61,50
52,2	51,9	358,50	361,31	353,69

Se calculan las medias, las varianzas y la covarianza de ambas variables:

$$\overline{X} = \frac{52,2}{8} = 6,53 \quad ; \quad s_x^2 = \frac{358,5}{8} - 6,53^2 = 2,2369$$

$$\overline{Y} = \frac{51,9}{8} = 6,49 \quad ; \quad s_Y^2 = \frac{361,31}{8} - 6,49^2 = 3,0761$$

$$s_{XY} = \frac{353,69}{8} - 6,49 \cdot 6,53 = 1,8803$$

De donde el coeficiente de correlación lineal es:

$$r = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}} = \frac{1,8803}{\sqrt{2,2369 \cdot 3,0761}} = 0,7168$$

que informa de una correlación lineal moderada – alta entre las variables *X* e *Y*.

b) Se debe estimar la recta de regresión de la calificación en Econometría I (*Y*) en función de la calificación en economía de la empresa en la PAU (*X*).

$$b = \frac{s_{XY}}{s_X^2} = \frac{1,8803}{2,2369} = 0,8406 \quad ; \quad a = \overline{Y} - b\overline{X} = 6,4875 - 0,8406 \cdot 6,525 = 1,0026$$

De modo que la recta de regresión de *Y* sobre *X* es: $y = 1,0026 + 0,8406x$.

Luego: $y(6,5) = 1,0026 + 0,8406 \cdot 6,5 = 6,47$

Que, dado el valor del coeficiente de correlación, puede considerarse una buena estimación de la calificación en Econometría I.

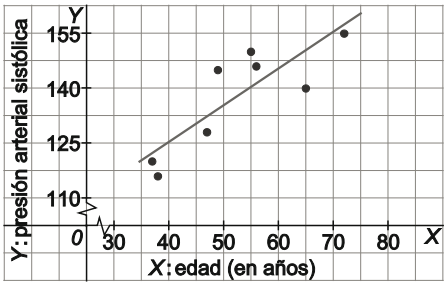
49. La siguiente tabla da la edad (*X*) en años y la tensión sistólica (*Y*) de una muestra de 8 mujeres tomada entre las pacientes de un centro de salud:

X	56	72	37	65	47	55	49	38
Y	146	155	120	140	128	150	145	116

a) Representa gráficamente los datos y la recta de regresión que da la tensión en función de la edad. Valora la tendencia observada.

b) ¿Cuál se espera que sea la tensión de una mujer de 50 años?

a) A la derecha se muestra la nube de puntos junto con la recta de regresión cuya ecuación determinaremos posteriormente.
Se puede observar una relación con tendencia creciente y lo que parece un buen ajuste de la recta de regresión a la nube de puntos.



x_j	y_j	x_j^2	y_j^2	$x_i y_j$
56	146	3136	21 316	8176
72	155	5184	24 025	11 160
37	120	1369	14 400	4440
65	140	4225	19 600	9100
47	128	2209	16 384	6016
55	150	3025	22 500	8250
49	145	2401	21 025	7105
38	116	1444	13 456	4408
419	1100	22993	152 706	58 655

$$\overline{X} = \frac{419}{8} = 52,375 \text{ años} \quad ; \quad s_x^2 = \frac{22993}{8} - 52,375^2 = 130,9844$$

$$\overline{Y} = \frac{1100}{8} = 137,5 \quad ; \quad s_Y^2 = \frac{152706}{8} - 137,5^2 = 182,0$$

$$s_{XY} = \frac{58655}{8} - 52,375 \cdot 137,5 = 130,3125$$

Con estos resultados se estiman los coeficientes de la ecuación de regresión:

$$b = \frac{s_{XY}}{s_x^2} = \frac{130,3125}{130,9844} = 0,9949 \quad ; \quad a = \overline{Y} - b\overline{X} = 137,5 - 0,9949 \cdot 52,375 = 85,3937$$

De manera que la ecuación de la recta de regresión de *Y* sobre *X* es: $y = 85,3937 + 0,9949x$.

b) Utilizando la ecuación estimada en al apartado anterior, el valor esperado de la tensión arterial sistólica para una persona de 50 años es :

$$y(50) = 85,3937 + 0,9949 \cdot 50 = 135,14$$

50. En la tabla se presentan los datos de la renta per cápita del año 2012 en miles de dólares (*Y*) y el porcentaje del PIB en el año 2009 destinado a educación (*X*) en diez países de la Unión Europea

X	5,0	5,8	5,9	4,7	5,1	5,6	6,5	8,7	6,8	7,3
Y	30,2	20,7	42,8	33,9	42,6	38,9	45,9	57,6	47,5	57,9

a) Escribe la recta de regresión de *Y* sobre *X*. ¿Qué porcentaje de la variabilidad de la renta per cápita se explica por el gasto en educación?

b) ¿Qué nivel de renta per cápita en 2012 se puede estimar que tendría un país que en 2009 invirtió en educación el 6 % de su PIB? Valora la fiabilidad de la predicción.

a) Se construye la tabla siguiente para facilitar la realización de los cálculos:

x_j	y_j	x_j^2	y_j^2	$x_j y_j$
5,0	30,2	25,0	909,0	150,75
5,8	20,7	33,6	426,9	119,8338
5,9	42,8	34,8	1831,2	252,4787
4,7	33,9	22,1	1152,1	159,5274
5,1	42,6	26,0	1816,9	217,3875
5,6	38,9	31,4	1512,5	217,7896
6,5	45,9	42,3	2102,5	298,0445
8,7	57,6	75,7	3314,5	500,8764
6,8	47,5	46,2	2255,9	322,9728
7,3	57,9	53,3	3358,0	423,0204
61,4	417,9	390,4	18679,5	2662,7

Se calculan las medias y varianzas marginales y la covarianza de las variables *X* e *Y*:

$$\overline{X} = \frac{61,4}{10} = 6,14 \text{ \% del PIB a educación} \quad ; \quad s_X^2 = \frac{390,4}{10} - 6,14^2 = 1,3384$$

$$\overline{Y} = \frac{418}{10} = 41,8 \text{ miles de euros} \quad ; \quad s_Y^2 = \frac{188682,78}{10} - 41,8^2 = 121,0380$$

$$s_{XY} = \frac{2663,15}{10} - 41,8 \cdot 6,14 = 9,6630$$

Con lo que los coeficientes de la ecuación de regresión de *Y* sobre *X* son:

$$b = \frac{9,6630}{1,3384} = 7,22 \qquad a = 41,8 - 7,22 \cdot 6,14 = -2,53$$

Luego, la ecuación de la recta de regresión del precio del alquiler (*Y*) en función del número de habitaciones (*X*) es: $y = -2,5297 + 7,2198x$.

El coeficiente de determinación es: $R^2 = \frac{s_{XY}}{s_Y^2 s_X^2} = \frac{9,6630}{121,0380 \cdot 1,3384} = 0,5764$.

Es decir, el 57,64 % de la variabilidad de la renta per cápita viene explicada por el porcentaje del PIB que se destina a educación.

b) Como $x = 6$, se encuentra dentro del rango de observaciones de la variable *X*, basta con sustituir $x = 6$ en la ecuación obtenida en el aparatado b) para estimar el valor de *Y*:

$$y(6) = -2,5297 + 7,2198 \cdot 6 = 40,789 \text{ miles de euros}$$

51. La evolución de los porcentajes de estudiantes en las áreas de Ciencias Sociales (*Y*) y de Ciencias de la Salud y Servicios Sociales (*X*), desde 2002 hasta 2009, viene recogida en la tabla siguiente.

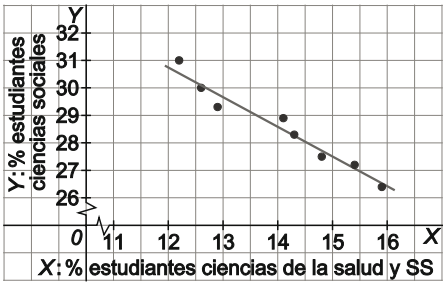
Año	2002	2003	2004	2005	2006	2007	2008	2009
X	12,2	12,6	12,9	14,1	14,3	14,8	15,4	15,9
Y	31	30	29,3	28,9	28,3	27,5	27,2	26,4

a) Representa la nube de puntos y comenta la relación que observas entre las dos variables.

b) Escribe la recta de regresión de *Y* sobre *X*.

c) Calcula el coeficiente de correlación y el Error Cuadrático Medio y valora los resultados.

a) Se representa la nube de puntos junto con la recta de regresión calculada en el apartado b)
Se observa una fuerte relación lineal con tendencia decreciente.



b)

x_j	y_j	x_j^2	y_j^2	$x_j y_j$
12,2	31,0	148,84	961,00	378,20
12,6	30,0	158,76	900,00	378,00
12,9	29,3	166,41	858,49	377,97
14,1	28,9	198,81	835,21	407,49
14,3	28,3	204,49	800,89	404,69
14,8	27,5	219,04	756,25	407,00
15,4	27,2	237,16	739,84	418,88
15,9	26,4	252,81	696,96	419,76
112,2	228,6	1586,32	6548,64	3191,99

$$\overline{X} = \frac{112,2}{8} = 14,025 \text{ \% de CC de la salud} \quad ; \quad s_X^2 = \frac{1586,32}{8} - 14,025^2 = 1,5894$$

$$\overline{Y} = \frac{228,6}{8} = 28,575 \text{ \% de CCSS} \quad ; \quad s_Y^2 = \frac{6548,64}{8} - 28,575^2 = 2,0494$$

$$s_{XY} = \frac{3191,99}{8} - 28,575 \cdot 14,025 = -1,7656$$

De modo que los coeficientes de la recta de regresión de *Y* sobre *X* son:

$$b = \frac{s_{XY}}{s_X^2} = \frac{-1,7656}{1,5894} = -1,1109 \quad ; \quad a = \overline{Y} - b\overline{X} = 28,575 + 1,1109 \cdot 14,025 = 44,1553$$

La recta de regresión del porcentaje de estudiantes de Ciencias Sociales (*Y*) en función del porcentaje de estudiantes de Ciencias de la Salud y Servicios Sociales (*X*) es:

$$y = 44,1553 - 1,1109x$$

Con los resultado obtenidos en el apartado b), se tiene que el coeficiente de correlación y el error cuadrático medio son:

$$r = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}} = \frac{-1,7656}{\sqrt{1,5859 \cdot 2,0494}} = -0,9783$$

$$\text{ECM} = s_Y^2\left(1 - \frac{s_{XY}^2}{s_X^2 s_Y^2}\right) = 2,0494\left(1 - \frac{1,7656^2}{1,5859 \cdot 2,0494}\right) = 0,0880$$

El valor del coeficiente de correlación, próximo a –1, indica una fuerte relación inversa entre ambas variables y que el ajuste lineal es muy bueno, confirmado por el pequeño valor del error cuadrático medio.

52. La tabla siguiente muestra la distribución de la situación profesional (*X*) y el nivel de estudios (*Y*) en una determinada población.

		Nivel de estudios		
		Básicos	Medios	Altos
Situación profesional (*X*)	Empleado fijo	14	22	18
	Empleado temporal	18	31	21
	Autónomo	12	8	10
	Sin empleo	23	14	9

a) Escribe las distribuciones marginales de frecuencias absolutas y relativas.

b) Halla la distribución de la situación profesional *X*, condicionada al nivel de estudios altos.

c) ¿Son independientes estas variables?

a) Las tablas con las distribuciones de frecuencias marginales son:

	X_i	f_i	h_i
Situación profesional	Empleado fijo	54	0,27
	Empleado temporal	70	0,35
	Autónomo	30	0,15
	Sin empleo	46	0,23
		200	1

	Y_j	f_j	h_j
Nivel de estudios	Básicos	67	0,335
	Medios	75	0,375
	Altos	58	0,29
		200	1

b) La tabla para la distribución de la situación profesional *X*, condicionada al nivel de estudios altos es:

$X\|_{Y=\text{altos}}$	$f_i\|_{Y=\text{altos}}$	$h_i\|_{Y=\text{altos}}$
X_1 = Empleado fijo	18	0,3103
X_2 = Empleado temporal	21	0,3621
X_3 = Autónomo	10	0,1724
X_4 = Sin empleo	9	0,1552
	58	1

c) Para comprobar que NO son independientes basta encontrar un caso en el que $h_{ij} \neq h_i \cdot h_j$, por ejemplo:

$$h_{X=\text{fijo},\ Y=\text{altos}} = 0,3103 \neq h_{X=\text{fijo}}\ h_{Y=\text{altos}} = 0,27 \cdot 0,29 = 0,0783$$

53. Cuantos más coches circulan por una carretera, menor es la velocidad del tráfico. Con el fin de mejorar el transporte, a la entrada de una ciudad se ha tomado una muestra de 10 observaciones de la densidad del tráfico (*X*, n.º de vehículos por km) y de la velocidad en ese instante (*Y*, km/h).

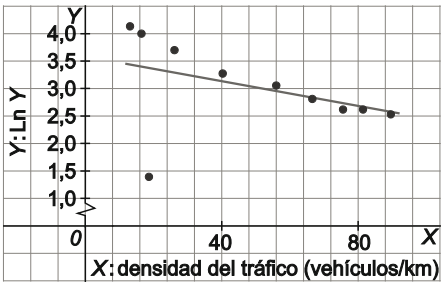
X	13	75,6	81,4	89,6	40,2	26,1	16,5	18,6	56	66,6
Y	62	13,7	13,8	12,6	26,3	40	54,7	4	21,2	16,6

a) Dibuja la nube de puntos *y* calcula el coeficiente de correlación.

b) Transforma la variable *Y* en $Z = \ln Y$. Dibuja la nube de puntos, calcula en nuevo coeficiente de correlación.

c) Escribe la recta de regresión de *Z* sobre *X*.

a) Se representa la nube de puntos con los datos del enunciado. Se observa una relación de tipo inverso, pero también se aprecia una fuerte relación funcional que no parece lineal.



Para determinar el coeficiente de correlación se construye la tabla ampliada:

x_j	y_j	x_j^2	y_j^2	$x_j y_j$
13	62	169	3844	806
75,6	13,7	5715,36	187,69	1035,72
81,4	13,8	6625,96	190,44	1123,32
89,6	12,6	8028,16	158,76	1128,96
40,2	26,3	1616,04	691,69	1057,26
26,1	40	681,21	1600	1044
16,5	54,7	272,25	2992,09	902,55
18,6	4	345,96	16	74,4
56	21,2	3136	449,44	1187,2
66,6	16,6	4435,56	275,56	1105,56
483,6	264,9	31 025,5	10 405,67	9464,97

$$\overline{X} = \frac{483,6}{10} = 48,36 \; ; \; \overline{Y} = \frac{264,9}{10} = 26,49$$

$$s_x^2 = \frac{31025,5}{10} - 48,36^2 = 763,86$$

$$s_Y^2 = \frac{10405,67}{10} - 26,49^2 = 338,847$$

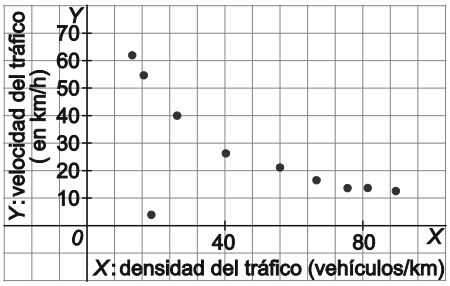
$$s_{XY} = \frac{9464}{10} - 26,49 \cdot 48,36 = -334,656$$

El coeficiente de correlación lineal es: $r = \dfrac{s_{XY}}{\sqrt{s_x^2 s_Y^2}} = \dfrac{-334,656}{\sqrt{763,86 \cdot 338,847}} = 0,6578$.

Informa de una correlación lineal moderada – alta entre las variables *X* e *Y*.

b) Si hacemos la transformación $Z = \text{Ln } Y$ la distribución que obtenemos es la siguiente:

x_j	$z_i = \text{Ln}(y_j)$	x_j^2	z_j^2	$x_j z_j$
13	4,13	169	17,03	53,65
75,6	2,62	5715,36	6,85	197,88
81,4	2,62	6625,96	6,89	213,65
89,6	2,53	8028,16	6,42	227,02
40,2	3,27	1616,04	10,69	131,44
26,1	3,69	681,21	13,61	96,28
16,5	4,00	272,25	16,01	66,03
18,6	1,39	345,96	1,92	25,79
56	3,05	3136	9,33	171,02
66,6	2,81	4435,56	7,89	187,11
483,60	30,11	31 025,50	96,65	1369,86



$$\overline{Z} = \frac{30,11}{10} 3,011$$

$$s_z^2 = \frac{96,65}{10} - 3,011^2 = 0,5999 \; ; \; s_{xz} = \frac{1369,86}{10} - 3,011 \cdot 48,36 = -8,6260 \; ; \; r = \frac{s_{xz}}{\sqrt{s_x^2 s_z^2}} = \frac{-8,6260}{\sqrt{763,86 \cdot 0,5999}} = 0,1638$$

c) Los coeficientes de la recta de regresión son:

$$b = \frac{s_{xz}}{s_x^2} = \frac{-8,6260}{763,86} = -0,0113 \; ; \; a = \overline{Z} - b\overline{X} = 3,011 + 0,0113 \cdot 48,36 = 3,5571$$

De modo que la recta de regresión de Z sobre *X* es : $z = 3,5571 - 0,0113x$.

ENTORNO MATEMÁTICO

Si eres chico y joven, tienes más probabilidad de sufrir un accidente

Las compañías aseguradoras de autos cobran primas más elevadas por las pólizas de seguros que incluyen conductores menores de 25 años, pues afirman que estos tienen más riesgo de sufrir un accidente que el resto. Además, por la misma razón, los varones pagan más que las mujeres. Esto es lo que encontraron Nicolás y Mercedes cuando, recién sacado el carnet de conducir, fueron a contratar el seguro para su coche. Indignados, decidieron comprobar si la información era cierta. Tras buscar información en la Dirección General de Tráfico (DGT) obtuvieron los siguientes datos del anuario estadístico de accidentes correspondiente a 2013.

Ayuda a Nicolás y Mercedes y, usando una hoja de cálculo, responde:

a) ¿Se puede establecer una relación lineal entre la edad del conductor y el número de accidentes?

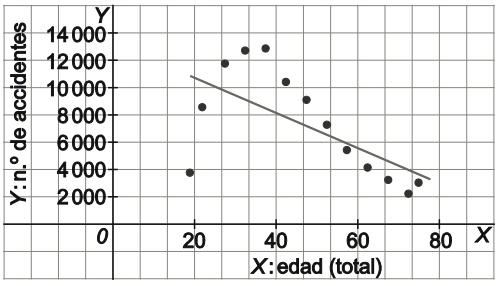
b) Si el conductor es varón, ¿esa relación es distinta que cuando es mujer?

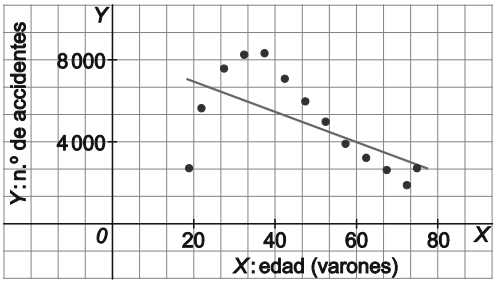
Haz los cálculos para el total de conductores y luego para hombres y mujeres por separado. ¿Cuáles son las conclusiones? ¿Están acertadas las compañías de seguros?

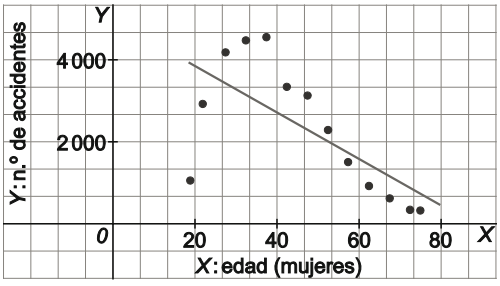
A partir de la tabla de datos podemos dibujar las nubes de puntos correspondientes al total de conductores, a los conductores varones y a las conductoras.

	Varones		total	Mujeres		Total	TOTAL
Edad (años)	En carretera	En zona Urbana	Varones	En carretera	En zona Urbana	Mujeres	
18 a 20	1277	1442	2719	530	532	1062	3781
21 a 24	2529	3110	5639	1460	1467	2927	8566
25 a 29	3429	4151	7580	2061	2128	4189	11769
30 a 34	3663	4586	8249	2193	2281	4474	12723
35 a 39	3816	4506	8322	2150	2408	4558	12880
40 a 44	3228	3849	7077	1246	2095	3341	10418
45 a 49	2680	3299	5979	1400	1738	3138	9117
50 a 54	2255	2736	4991	1024	1274	2298	7289
55 a 59	1743	2175	3918	687	822	1509	5427
60 a 64	1498	1724	3222	413	513	926	4148
65 a 69	1323	1310	2633	283	344	627	3260
70 a 74	993	895	1888	162	184	346	2234
75 ó más	1470	1240	2710	140	191	331	3041
TOTALES	29 904	35 023	64 927	13 749	15 977	29 726	94 653

En los tres casos se observa una dependencia similar del número de accidentes en función de la edad.
Llama la atención el hecho de que entre los 20 y los 30 años la relación es de tipo directo mientras que entre los 40 y los 70 años la relación es de tipo inverso.







Los resultados globales obtenidos en cada caso son:

Considerando todos los conductores:

$\overline{X} = 47,15$; $\overline{Y} = 7281$; $s_{XY} = -576480$; $R^2 = 0,4139$; $r = -0,6434$

recta de regresión: $y = -130 - 13412x$

Considerando solo a los varones:

$\overline{X} = 47,15$; $\overline{Y} = 4994,4$; $s_{XY} = -327411$; $R^2 = 0,3778$; $r = -0,6147$

recta de regresión: $y = 8477 - 74x$

Considerando solo a las mujeres:

$\overline{X} = 47,2$; $\overline{Y} = 2286,6$; $s_{XY} = -249069,6$; $R^2 = 0,4647$; $r = -0,6817$

recta de regresión: $y = 4936 - 56x$

En los tres casos se observa una correlación moderada entre la edad y el número de accidentes considerando todo el rango de edades si bien es mayor en el caso de las mujeres.
Si hacemos el estudio diferenciando dos rangos, uno entre 18 y 40 y otro entre 40 y 80 los resultados son notablemente diferentes:

Entre 18 y 40 años:

Considerando todos los conductores:

$R^2 = 0,7789$; $r = 0,8826$; recta de regresión: $y = -2837,2 + 466,5x$

Considerando solo a los varones:

$R^2 = 0,7860$; $r = 0,8865$; recta de regresión: $y = -1404 + 288,5x$

Considerando solo a las mujeres:

$R^2 = 0,7672$; $r = 0,8759$; recta de regresión: $y = -1433 + 178x$

De esta forma, se detecta una mayor correlación, así como una mayor pendiente en el caso de los varones que en el de las mujeres.

A la luz de estos datos, parece que las compañías no están muy acertadas ya que no es en el rango de edades menores donde se concentran los accidentes sino en la población de mediana edad.

En el siguiente rango de edades, la correlación negativa entre ambas variables asciende hasta el 98,7% en el caso general, 99,7 % en el de los varones y hasta el 98,7 % en el caso de las mujeres.

¿Cuándo nos vamos de viaje?

Aunque los padres de Nicolás y Mercedes no son muy dados a este tipo de premios, sus abuelos han decidido premiarles con un viaje por haber terminado con una buena media académica el Bachillerato.

Los chicos han decidido viajar a Roma, y tienen permiso para hacerlo al mes que deseen durante su primer año de carrera siempre y cuando no se salten ninguna clase.

Con el propósito de asegurarse unas vacaciones con el mejor tiempo posible, ya que no están dispuestos a asarse de calor ni tampoco a estar todo el día con el paraguas, han estado consultando el servicio meteorológico italiano y han obtenido una tabla con la relación entre temperaturas (máxima y mínima de cada mes) y la cantidad de lluvia caída (precipitación) en Roma.

Mientras buscaban la información encontraron la siguiente afirmación: "cuanto más frío es un mes, más lluvioso resulta, y que cuanto más caluroso es el mes, más seco resulta".

Con ayuda de la tabla, responde:

a) ¿Es cierta la afirmación en Roma?

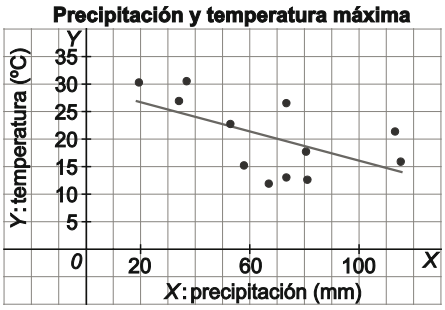
b) ¿Se puede establecer una relación lineal fiable entre la temperatura máxima (o mínima) y la cantidad de lluvia caída en un mes?

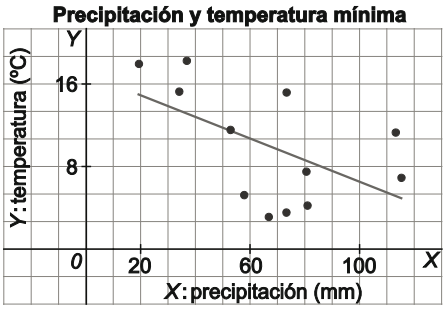
c) ¿Qué porcentaje de la variabilidad observada en las temperaturas a lo largo de un año es explicada por la cantidad de precipitaciones?

d) Busca datos y prueba con una tabla similar en tu ciudad o región.

Para responder a estas preguntas, se puede analizar la relación entre las precipitaciones a lo largo de los doce meses y la temperatura máxima (o mínima) observada en cada uno de esos doce meses. Empezamos por la temperatura máxima:

A la derecha se muestra la gráfica de dispersión de la variable bidimensional (precipitación total; temperatura diaria máxima), junto con la ecuación de la recta de regresión de la temperatura máxima respecto a la precipitación total y se incluye el coeficiente de determinación. Los cálculos se dejan para el estudiante.
Puede observarse una tendencia decreciente, a más precipitación, menor temperatura máxima, aunque no es una relación fuerte.
El coeficiente de determinación señala que el 33,8% de la variabilidad observada en las temperaturas máximas viene explicada por las precipitaciones.





La relación (en Roma) entre las precipitaciones y las temperaturas mínimas mensuales, se muestra en la gráfica de la izquierda.
Se observa que en este caso, la relación también muestra una tendencia decreciente, aún más débil que en el caso anterior, con una coeficiente de determinación que indica que apenas el 29,56% de la variabilidad observada en las temperaturas mínimas mensuales viene explicada por las precipitaciones.

Dada la correlación existente (débil) las ecuaciones de regresión obtenidas no deberían utilizarse para realizar predicciones, y si se usan para este fin, debe hacerse con precaución dada su baja fiabilidad.

AUTOEVALUACIÓN

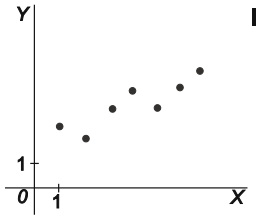
Comprueba qué has aprendido

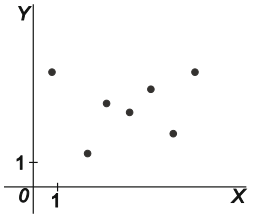
1. El coeficiente de correlación de una variable estadística bidimensional es $r = -0,8$ y las medias aritméticas de cada una de las distribuciones marginales son $\bar{x} = 1$ e $\bar{y} = 3$ y=3. Razona cuál de las siguientes cuatro rectas puede ser la recta de regresión de Y sobre X:

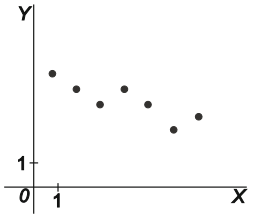
a) $y = -x + 6$ **b)** $y = 2x + 1$ **c)** $y = x + 2$ **d)** $y = -3x + 6$

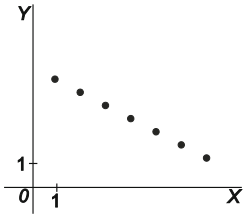
La recta de regresión debe contener el punto (1, 3), por lo tanto queda descartada la recta del apartado a). Las otras tres sí contienen este punto.
El valor negativo del coeficiente de correlación indica que la relación entre las variables es inversa, es decir, que si una crece la otra decrece. Esta situación solo describe la ecuación del apartado d) $y = -3x + 6$.

2. Asigna razonadamente a estos diagramas de dispersión el coeficiente de correlación adecuado.

I) 

II) 

III) 

IV) 

a) $r = -0,885$ **b)** $r = -1$ **c)** $r = 0,885$ **d)** $r = 0,119$

I) Correlación positiva: se asigna c).
II) Correlación próxima a cero: se asigna d).
III) correlación negativa: se asigna a).
IV) correlación negativa, buen ajuste lineal: se asigna b).

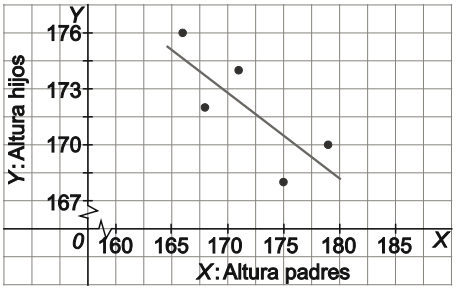
3. Las alturas de 5 padres (X) y las de sus hijos (Y) se recogen en la tabla siguiente:

X	171	168	175	166	179
Y	174	172	168	176	170

a) Representa gráficamente la nube de puntos. Cuantifica el tipo de relación entre las variables.

b) Escribe la recta de regresión de Y (estatura hijos) sobre X (estatura padres) y calcula el error cuadrático medio, explicando su significado.

a) La nube de puntos junto con la recta de regresión del apartado b) se muestra en el gráfico de la derecha. Se observa una relación inversa entre las dos variables.



x_j	y_j	x_j^2	y_j^2	$x_j y_j$
171	174	29 241	30 276	29 754
168	172	28 224	29 584	28 896
175	168	30 625	28 224	29 400
166	176	27 556	30 976	29 216
179	170	32 041	28 900	30 430
859	860	147 687	147 960	147 696

$$\bar{X} = \frac{859}{5} = 171,8 \quad ; \quad s_X^2 = \frac{147687}{5} - 171,8^2 = 22,16 \quad ; \quad \bar{Y} = \frac{860}{5} = 172 \quad ; \quad s_Y^2 = \frac{147960}{5} - 172^2 = 8$$

$$s_{XY} = \frac{147696}{5} - 171,8 \cdot 172 = -10,4 \quad ; \quad r = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}} = \frac{-10,4}{\sqrt{22,16 \cdot 8}} = -0,7811$$

b) $\text{ECM} = s_Y^2\left(1 - R^2\right) = 8 \cdot \left(1 - (-0,7811)^2\right) = 3,119$

$$b = \frac{s_{XY}}{s_X^2} = \frac{-10,4}{22,16} = -0,4693 \; ; \; a = \bar{Y} - b\bar{X} = 172 + 0,4693 \cdot 171,8 = 252,626$$

La recta de regresión de Y sobre X es: $y = 14871,05 - 0,4915x$. Puede considerarse una buena estimación.

4. La distribución de la variable bidimensional (X,Y) viene dada en la siguiente tabla:

X	6	3	4	1	7	9
Y	4	6	5	5	2	7

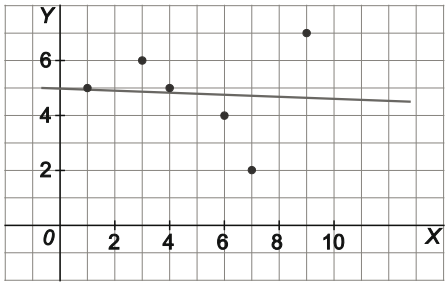
a) Representa gráficamente la distribución.

b) Halla la recta de regresión de Y sobre X.

c) ¿Qué porcentaje de la variabilidad de la variable Y es explicada por el modelo de regresión?

d) Si $X = 5$, ¿cuál es el valor esperado de Y? ¿Y si $X = 15$? Comenta la fiabilidad de ambas predicciones.

a) La nube de puntos junto con la recta de regresión del apartado b) es la del margen, que resulta ser aparentemente muy dispersa y con escasa correlación.



b)

x_j	y_j	x_j^2	y_j^2	$x_j y_j$
6	4	36	16	24
3	6	9	36	18
4	5	16	25	20
1	5	1	25	5
7	2	49	4	14
9	7	81	49	63
30	29	192	155	144

$$\overline{X} = \frac{30}{6} = 5 \quad ; \quad s_x^2 = \frac{192}{6} - 5^2 = 7$$

$$\overline{Y} = \frac{29}{6} = 4,83 \quad ; \quad s_y^2 = \frac{139}{6} - 4,83^2 = 2,5$$

$$s_{xy} = \frac{144}{6} - 5 \cdot 4,83 = -0,15$$

Los coeficientes de la recta de regresión son:

$$b = \frac{s_{xy}}{s_x^2} = \frac{-0,15}{7} = -0,0214 \quad ; \quad a = \overline{Y} - b\overline{X} = 4,83 + 0,0214 \cdot 5 = 4,9371$$

La recta de regresión de Y sobre X es: $y = 4,9371 - 0,0214x$.

c) Los coeficientes de correlación y de determinación son:

$$r = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{-0,15}{\sqrt{7 \cdot 2,5}} = -0,0359 \Rightarrow R^2 = (-0,0359)^2 = 0,0013$$

Lo que indica que tan solo el 0,13 % de la variabilidad de Y está explicada por la recta de regresión.

d) Para el valor $X = 5$, el valor esperado de Y a través de la recta de regresión, se calcula:
$y(5) = 4,9371 - 0,0214 \cdot 5 = 4,8301$.
La fiabilidad de esta estimación es escasa en virtud del valor del coeficiente de correlación.
En el caso de $X = 15$ se trata de un valor que está fuera del rango del estudio, luego la fiabilidad del cálculo es nula.

Relaciona y contesta

Elige la única respuesta correcta en cada caso

1. En la recta de regresión de *Y* sobre *X*, se ha obtenido un coeficiente de determinación $R^2 = 0{,}82$, entonces:

A. La relación entre *X* e *Y* es directa.

B. La pendiente de la recta de regresión es 0,82.

C. El 18 % de la variabilidad de *Y* queda sin explicar por el modelo de regresión.

D. Con este dato, no hay relación entre *X* e *Y*.

La respuesta correcta es la C: solo el 82 % de la variabilidad de *Y* queda explicado por el modelo de regresión.

2. Si los datos de una variable estadística bidimensional se multiplican por 3, el coeficiente de correlación:

A. Queda multiplicado por 3.

B. Es igual al anterior elevado al cubo.

C. Es el mismo.

D. Queda dividido por 3.

La respuesta correcta es la C, ya que $r_{3X3Y} = \dfrac{s_{3X3Y}}{\sqrt{s^2_{3X} s^2_{3Y}}} = \dfrac{3^2 s_{XY}}{\sqrt{3^2 s_x^2 3^2 s_Y^2}} = \dfrac{\cancel{3^2} s_{XY}}{\cancel{3^2}\sqrt{s_x^2 s_Y^2}} = \dfrac{s_{XY}}{\sqrt{s_x^2 s_Y^2}} = r_{XY}$.

3. El 81 % de la variabilidad de *Y* viene explicado por el modelo de regresión. Si la media de la variable *X* es 1 y la recta de regresión de *Y* sobre *X* es $y = 2{,}5 - 1{,}4x$, entonces:

A. $\overline{Y} = 2{,}5,\ r = -0{,}9$

B. $\overline{Y} = 2{,}5,\ r = 0{,}9$

C. $\overline{Y} = 1{,}1,\ r = -0{,}9$

D. $\overline{Y} = 1{,}1,\ r = 0{,}9$

La respuesta correcta es la C. Por ser $R^2 = 0{,}81$ y tener pendiente negativa se tiene que $r = -\sqrt{0{,}81} = -0{,}9$, luego podrían ser A o C. Por otra parte, $\overline{Y} = 2{,}5 - 1{,}4 \cdot 1 = 1{,}1$.

4. De la distribución (*X*,*Y*) se sabe que $S_X = 2$, $S_{XY} = -2$, $\overline{X} = 8$, $\overline{Y} = 10$. Entonces la recta de regresión de *Y* sobre *X* es:

A. $y = 10 - 2x$ **B.** $y = 14 - 0{,}5x$ **C.** $y = 10 + 0{,}5x$ **D.** $y = 8 - 2x$

La respuesta correcta es la B: es suficiente calcular $b = \dfrac{s_{XY}}{s_x^2} = \dfrac{-2}{2^2}$; $a = \overline{Y} - b\overline{X} = 10 + 0{,}5 \cdot 8 = 14$.

Señala, en cada caso, las respuestas correctas

5. De la variable (*X*, *Y*) se sabe que $s_{XY} = 2{,}5$ y que $R^2 = 0{,}75$. Si $Z = 3X$ y $T = Y + 3$, entonces:

A. $s_{ZT} = s_{XY} + 3$

B. $R^2_{ZT} = R^2_{XY}$

C. $s_{ZT} = 3s_{XY}$

D. $R^2_{ZT} = 9R^2_{XY}$

Las respuestas correctas son B. y C.

6. Con el modelo de regresión lineal de *Y* sobre *X*, se pueden realizar predicciones razonables sobre *Y*.

A. En cualquier caso.

B. Si el valor dado a *X* se encuentra cerca de la media de *X*.

C. Solo para valores pequeños de *X*.

D. Si el valor de *X* está en el rango de valores de la muestra.

La respuesta correcta es la D, siempre que el coeficiente de correlación sea próximo a 1 o –1

Elige la relación correcta entre las dos afirmaciones

7. 1. La recta de regresión es $y = 2 - x$. **2. Las medias marginales son $\bar{X} = 1$ e $\bar{Y} = 1$.**

A. $1 \Rightarrow 2$

B. $2 \Rightarrow 1$

C. $1 \Leftrightarrow 2$

D. $1 \not\Leftrightarrow 2$

La relación correcta es la D. A no es cierta porque con esta recta de regresión las medias de *X* e *Y* podrían ser, por ejemplo, 0 y 2. B no es cierta porque con esas medias, la recta de regresión podría ser, por ejemplo $y = x$. Por lo anterior C no puede ser cierta.