12 Distribuciones bidimensionales

EJERCICIOS PROPUESTOS

1 y 2. Ejercicios resueltos.

3. En un periodo de fuertes lluvias se midió cada hora, durante 48 horas, el caudal (*X*, en m^3/s) del río Ebro. Los datos agrupados se muestran en la tabla siguiente:

X	[26, 38)	[38, 50)	[50, 62)	[62, 74)	[74, 86)	[86,98)	[98,110]
f_i	6	2	8	10	10	4	8

a) Halla la media, la mediana y la clase modal. **b)** Calcula la varianza y la desviación típica.

Ampliamos la tabla con las columnas necesarias para los cálculos que se piden:

X	f_j	x_j	$f_j x_j$	$f_j x_j^2$	F_j
[26, 38)	6	32	192	6144	6
[38, 50)	2	44	88	3872	8
[50, 62)	8	56	448	25 088	16
[62, 74)	10	68	680	46 240	26
[74, 86)	10	80	800	64 000	36
[86, 98)	4	92	368	33 856	40
[98, 110)	8	104	832	86 528	48
	48		3408	265 728	

a) El caudal medio en las 48 horas ha sido $\overline{x} = \frac{3408}{48} = 71$ m^3/s.

Mediana, la mitad del conjunto de datos es 24. Antes de la clase [62, 74) se acumulan 16 datos, por lo que faltan 8 para llegar al 50%. La citada clase contiene 10 datos y su longitud es 12. De manera que por interpolación lineal se obtiene $M = 62 + \frac{(24-16)\cdot 10}{12} = 68{,}67$ m^3/s.

En cuanto a la moda, se observan dos clases modales: [62, 74) y [74, 86)

b) La varianza y la desviación típica son respectivamente: $s^2 = \frac{265\,728}{48} - 71^2 = 495 \Rightarrow s = \sqrt{495} = 22{,}249$ m^3/s

4. En una muestra de 150 familias, el número de hijos (*X*) por familia se recoge en la tabla siguiente:

X	0	1	2	3	4	5
f_j	20	32	59	28	8	3

a) Calcula la media y señala la moda y la mediana. **c)** Halla la varianza y la desviación típica.

b) Calcula los cuartiles.

Se construye la tabla para los cálculos añadiendo las columnas correspondientes

x_i	f_i	$x_i f_i$	$x_i^2 f_i$	F_i
0	20	0	0	20
1	32	32	32	52
2	59	118	236	111
3	28	84	252	139
4	8	32	128	147
5	3	15	75	150
	150	281	723	

a) El número medio de hijos por familia es $\overline{x} = \frac{281}{150} = 1{,}873$.

La moda es $M_o = 2$, que es el número de hijos que más familias tienen en la muestra.

La mediana es $M = 2$, que es valor de variable que deja debajo (encima) el 50% de las observaciones

b) Observando la columna de las frecuencias acumuladas F_j de la tabla anterior, se tiene que:

El 25% de 150 es 37,5. El valor de la variable que ocupa el lugar 37 (o 38) es el primer cuartil $Q_1 = 1$.

El cuartil Q_2 es la mediana, luego $Q_2 = M = 2$.

El 75% de 150 es 112,5. El valor de la variable que ocupa el lugar 112 (113) es el tercer cuartil $Q_3 = 3$.

c) De la tabla anterior: $s^2 = \frac{723}{150} - (1{,}873)^2 = 1{,}3106 \Rightarrow s = \sqrt{1{,}3106} = 1{,}1448$

5. Ejercicio resuelto.

6. Con el fin de relacionar el índice de flúor en el agua (*X* en ppm) con la tasa de caries (*Y* en %) se han tomado muestras en 10 ciudades, los datos obtenidos son:

a) Calcula las medias y varianzas de las distribuciones marginales.

b) Representa la nube de puntos.

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

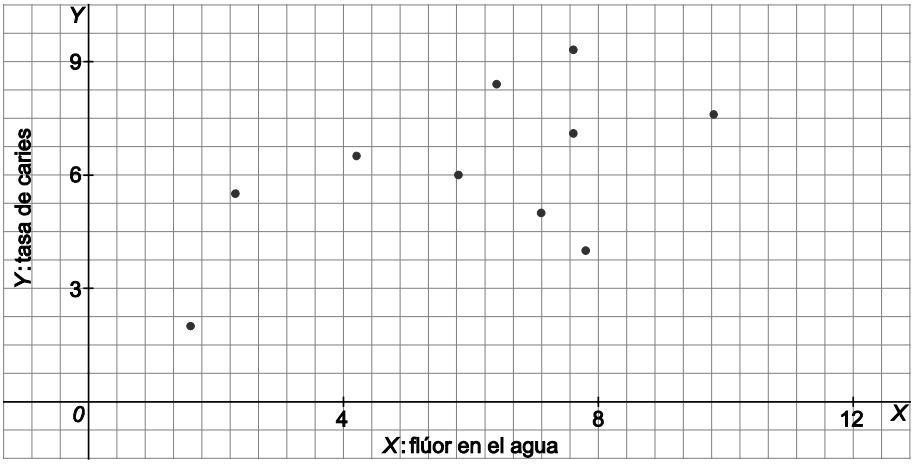
d) Halla la covarianza.

Se completa la tabla con las columnas necesarias para calcular las varianzas del índice de flúor (*X*) y de la tasa de caries (*Y*), así como la covarianza de las variables *X* e *Y*.

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1,6	2	2,56	4	3,2
7,8	4	60,84	16	31,2
7,1	5	50,41	25	35,5
2,3	5,5	5,29	30,25	12,65
5,8	6	33,64	36	34,8
4,2	6,5	17,64	42,25	27,3
7,6	7,1	57,76	50,41	53,96
9,8	7,6	96,04	57,76	74,48
6,4	8,4	40,96	70,56	53,76
7,6	9,3	57,76	86,49	70,68
60,2	61,4	422,9	418,72	397,53

a) El índice medio de flúor en ppm es $\overline{x} = \frac{60,2}{10} = 6,02$ y la varianza $s_x^2 = \frac{422,9}{10} - (6,02)^2 = 6,05$. El porcentaje medio de tasa de caries es: $\overline{y} = \frac{61,4}{10} = 6,14$ y la varianza $s_y^2 = \frac{418,72}{10} - (6,14)^2 = 4,172$.

b) La nube de puntos es:



c) La relación entre el índice de flúor en el agua y la tasa de caries es directa hasta una tasa de 4 ppm aproximadamente. A partir de ese índice la relación entre las dos variables es inversa.

d) La covarianza es: $s_{xy} = \frac{397,53}{10} - 6,14 \cdot 6,02 = 2,790$

7. El número de libros de consulta solicitados (*X*) en una biblioteca municipal, junto con el número de revistas (*Y*) se recoge en la siguiente tabla:

X \ *Y*	[10, 15)	[15, 20)	[20, 25)	[25, 30)
[20, 30)	23	62	163	28
[30, 40)	24	55	159	22
[40, 50)	33	65	127	12

a) Obtén las distribuciones marginales.

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

c) Cuantifica la relación entre las variables.

a) Distribución marginal de *X*:

X	x_i
[20, 30)	25
[30, 40)	35
[40, 50)	45

Distribución marginal de *Y*:

Y	y_i
[10, 15)	12,5
[15, 20)	17,5
[20, 25)	22,5
[25, 30)	27,5

b) Se completan las tablas anteriores para realizar los cálculos de las medias y las varianzas marginales calculando las marcas de clase.

X	x_i	f_i	$f_i x_i$	$f_i x_i^2$
[20, 30)	25	276	6900	172 500
[30, 40)	35	260	9100	318 500
[40, 50)	45	237	10 665	479 925
		773	26 665	970 925

Y	y_i	f_i	$y_i f_i$	$f_i y_i^2$
[10, 15)	12,5	80	1000	12 500
[15, 20)	17,5	182	3185	55 737,5
[20, 25)	22,5	449	10 102,5	227 306,25
[25, 30)	27,5	62	1705	46887,5
		773	15 992,5	342 431,25

La media de *X* es: $\overline{x} = \frac{26\,665}{105} = 34,495$ y la varianza de *X* es $s_x^2 = \frac{970\,925}{773} - 34,495^2 = 66,143$

La media de *Y* es: $\overline{y} = \frac{15\,992,5}{773} = 20,689$ y la varianza de *Y* es $s_y^2 = \frac{342\,431,25}{773} - 20,689^2 = 14,955$

c) Se calcula la covarianza $s_{xy} = -3,436 \Rightarrow$ La relación entre las variables es inversa.

8. Ejercicio resuelto.

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

	[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 las distribuciones de frecuencias relativas de Y condicionadas por cada valor de *X*.

c) Halla la media y la varianza de *Y*| mujeres.

d) ¿Son independientes estas variables?

a) Distribución marginal de *X*

X	f_i
hombres	626
mujeres	537
	1163

Distribución marginal de *Y:*

Y	f_i
[15, 24]	217
[25, 34]	319
[35, 44]	305
[45, 54]	322
	1163

b) Frecuencias relativas de *Y* condicionadas por cada valor de *X:*

	Y				
X	[15, 24]	[25, 34]	[35, 44]	[45, 54]	
hombres	0,17891374	0,28434505	0,26198083	0,27476038	1
mujeres	0,19553073	0,26256983	0,26256983	0,27932961	1

c) Se halla la distribución Y| mujeres y se completa la tabla con los datos necesarios para calcular la media y la varianza.

Y| mujeres	f_i	x_i	$x_i f_i$	$f_i x_i^2$
[15, 24]	105	19,5	2047,5	39 926,25
[25, 34]	141	29,5	4159,5	122 705,25
[35, 44]	141	39,5	5569,5	219 995,25
[45, 54]	150	49,5	7425	367 537,5
	537	138	19 201,5	750 164,25

Media: $\frac{19\,201}{537} = 35,757$ y la varianza es $\frac{750\,164,25}{537} - 35,757^2 = 118,391$

d) Para ver si *X* e *Y* son independientes se construye la tabla de las distribuciones relativas conjunta y marginales.

	[15, 24]	[25, 34]	[35, 44]	[45, 54]	$h_{j.}$
hombres	0,096 302 67	0,153 052 45	0,141 014 62	0,147 893 38	0,538 263 11
mujeres	0,090 283 75	0,121 238 18	0,121 238 18	0,128 976 78	0,461 736 89
$h_{.i}$	0,186 586 41	0,274 290 63	0,262 252 79	0,276 870 16	1

Como $0,274\,290\,63 \cdot 0,538\,263\,11 \neq 0,153\,052\,45$ entonces las variables son dependientes.

10. Ejercicio resuelto.

11. En la tabla se recogen para el periodo 2004 – 2013 la temperatura media invernal (*X*) en ºC en una región de la costa sur de California y el número de días (*Y*) en que el nivel de ozono superó los 0,20 ppm (partes por millón).

Año	04	05	06	07	08	09	10	11	12	13
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) Estima el número de días en los que se superarán 0,20 ppm de ozono si la temperatura media es de 16 ºC.

b) Analiza la precisión de la predicción en función de ECM y R^2.

a) 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*).

Se amplía la tabla de datos con las filas correspondientes para el cálculo de los valores medios, las varianzas y la covarianza.

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

$$\overline{x} = \frac{172}{10} = 17,2 \qquad s_x^2 = \frac{2962,04}{10} - 17,2^2 = 0,364$$

$$\overline{y} = \frac{568}{10} = 56,8 \qquad s_y^2 = \frac{34\,854}{10} - 56,8^2 = 259,16$$

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

De modo que la recta de regresión de *Y* sobre *X* es:

$$y = 56,8 + \frac{2,57}{0,364}(x - 17,2) \Rightarrow y = 7,06x - 64,64$$

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

$y = 7,06 \cdot 16 - 64,64 = 48,56$ .Es decir, entre 48 y 49 días.

b) Se calcula el Error Cuadrático Medio y el coeficiente de determinación R^2.

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{2,57^2}{(0,364) \cdot (259,6)} = 0,07$$

$$\text{ECM} = s_y^2 \left(1 - R^2\right) = 259,16\left(1 - 0,07\right) = 241,01$$

El Error Cuadrático Medio es alto e informa de que el ajuste de la recta a la nube de puntos no es bueno, aunque, como depende de las unidades de medida de la variable *Y*, no sirve para realizar comparaciones.

Sin embargo, el coeficiente de determinación es 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 son poco precisas.

12. Para analizar la relación entre las semanas de gestación (*X*) y el peso del feto (*Y*) en gramos, se tomó una muestra de 8 embarazadas. Los datos fueron:

X	28	30	33	32	28	34	38	36
Y	1150	1350	1925	1850	1040	2125	2950	2475

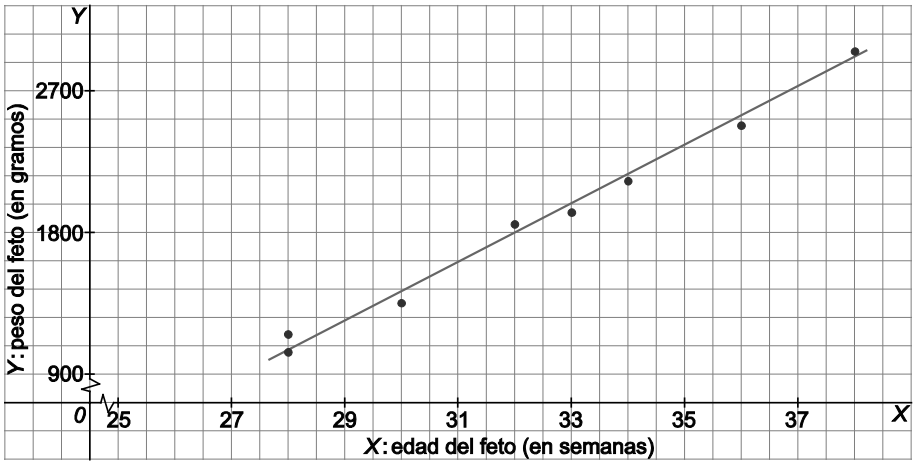
a) Representa los datos gráficamente.

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

c) Calcula el ECM y el coeficiente de determinación y analiza la bondad del ajuste.

d) ¿Qué peso tendría un feto con 31 semanas?

a) La nube de puntos del peso en gramos (*Y*) frente a las semanas de gestación (*X*) es, con la recta de regresión ya dibujada:



Puede observarse una fuerte relación lineal directa entre ambas variables.

b) Para escribir la recta de regresión se amplía la tabla con las columnas correspondientes para el cálculo de los valores medios y de las varianzas y la covarianza.

X	*Y*	x_i^2	y_i^2	x_iy_i
28	1150	784	1 322 500	32 200
30	1350	900	1 822 500	40 500
33	1925	1089	3 705 625	63 525
32	1850	1024	3 422 500	59 200
28	1040	784	1 081 600	29 120
34	2125	1156	4 515 625	72 250
38	2950	1444	8 702 500	112 100
36	2475	1296	6 125 625	89 100
259	14 865	8477	30 698 475	497 995

De modo que:

$$\overline{x} = \frac{259}{8} = 32{,}375 \qquad s_x^2 = \frac{8477}{8} - 32{,}375^2 = 11{,}484$$

$$\overline{y} = \frac{14\,865}{8} = 1858{,}125$$

$$s_y^2 = \frac{30\,698\,475}{8} - 1858{,}125^2 = 384\,680{,}859$$

$$s_{xy} = \frac{497995}{8} - 32{,}375 \cdot 1858{,}125 = 2092{,}578$$

La recta de regresión de *Y* sobre *X* es: $y = 1858{,}125 + \frac{2092{,}578}{11{,}484}(x - 32{,}375) \Rightarrow y = 182{,}11x - 4040{,}952$

c) El Error Cuadrático medio es $\text{ECM} = s_y^2\left(1 - R^2\right) = 384\,680{,}859\left(1 - 0{,}9911\right) = 3390{,}349$ y mide la bondad del ajuste pero como depende de las unidades de *Y*, en este caso gramos al cuadrado, no da idea de si el ajuste es bueno o no y, por ello, se recurre al coeficiente de determinación $R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = 0{,}9911$, que es muy próximo a 1 y, por tanto, permite concluir que el ajuste de la recta de regresión de *Y* sobre *X* a la nube de puntos es muy bueno, como se puede apreciar en la representación gráfica. Además R^2 informa que el 99,11 % de la variabilidad de *Y* viene explicada por el regresor *X*.

d) Para predecir el peso que tendría un feto de 31 semanas, se utiliza la recta de regresión obtenida en el apartado b), sustituyendo $x = 31$ en la ecuación, resulta:

$$y = 182{,}11 \cdot 31 - 4040{,}952 = 1607{,}589 \text{ g}$$

13. Ejercicio resuelto.

14. Al analizar si la productividad (*Y*) de cierto cultivo, en tm, viene explicada por las horas de regadío (*X*), se tomó una muestra en 6 fincas diferentes:

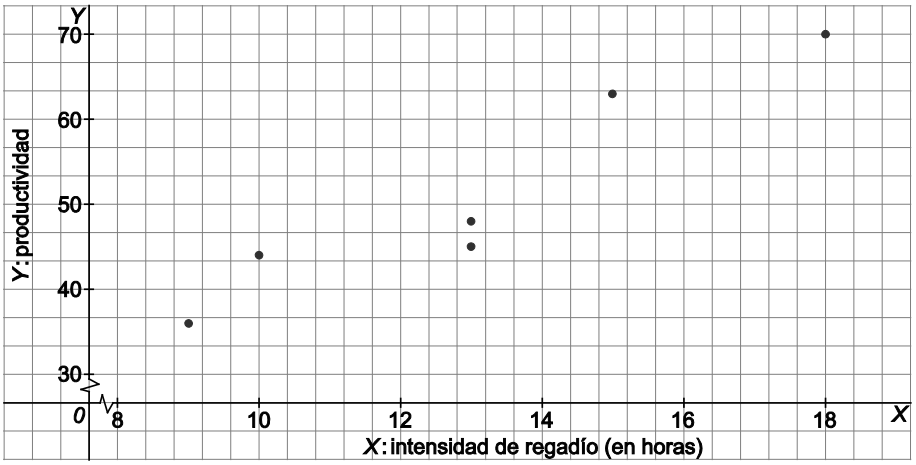
X	9	10	13	15	18	13
Y	36	44	48	63	70	45

a) Dibuja el diagrama de dispersión y calcula R^2 y r.

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

c) ¿Es fiable el ajuste lineal en este caso?

a) El diagrama de dispersión o nube de puntos de las observaciones correspondientes a las 6 fincas es:



Para calcular los coeficientes de determinación y correlación es preciso ampliar la tabla con las columnas necesarias para calcular las varianzas de X e Yy la covarianza.

X	Y	x_i^2	y_i^2	x_iy_i
9	36	81	1296	324
10	44	100	1936	440
13	48	169	2304	624
15	63	225	3969	945
18	70	324	4900	1260
13	45	169	2025	585
78	306	1068	16430	4178

$$\overline{x} = \frac{78}{6} = 13 \qquad s_x^2 = \frac{1068}{6} - 13^2 = 9$$

$$\overline{y} = \frac{306}{6} = 51 \qquad s_y^2 = \frac{16\,340}{6} - 51^2 = 137{,}33$$

$$s_{xy} = \frac{4178}{6} - 13 \cdot 51 = 33{,}33$$

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{33{,}33^2}{9 \cdot 137{,}33} = 0{,}8990 \Rightarrow r = 0{,}9481$$

b) Con los datos obtenidos en el apartado anterior se construye la recta de regresión de la productividad (Y) sobre las horas de regadío (X).

$$y = 51 + \frac{33{,}33}{9}(x - 13) \Rightarrow y = 3{,}7x + 2{,}86$$

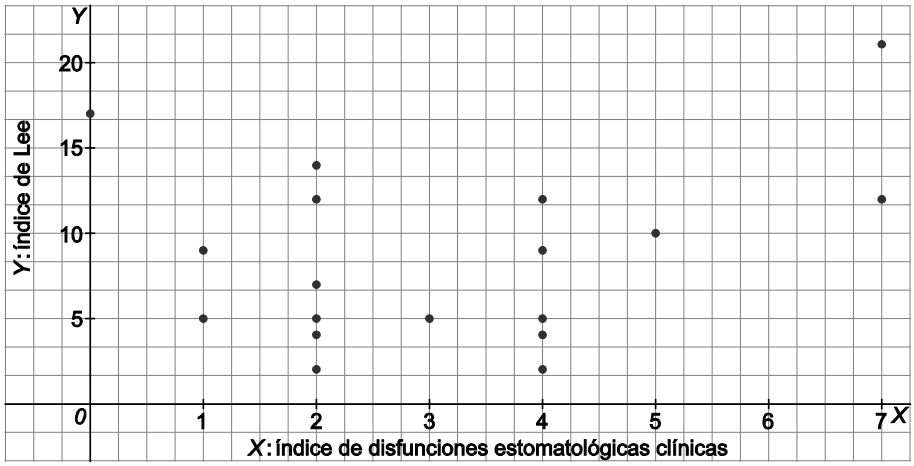
c) A la vista de los valores de los coeficientes de determinación y correlación se puede afirmar que el ajuste lineal a los datos es muy bueno puesto que el coeficiente de correlación lineal está próximo a 1.

Además, el 89,9% de la variabilidad observada en la productividad se puede explicar por el modelo de regresión, es decir por la intensidad, en horas, de regadío.

15. Los datos de la tabla recogen los índices de disfunciones estomatológicas clínicas (*X*) y de Lee (*Y*) que mide el avance de artritis reumatoide. Estudia si hay relación lineal entre dichos índices.

X	0	1	1	2	2	2	2	2	2
Y	17	9	5	5	7	12	2	4	14
X	3	4	4	4	4	4	5	7	7
Y	5	4	9	2	5	12	10	12	21

Para estudiar la existencia de la relación lineal, se representala nube de puntos y se calculan los coeficientes de determinación y de correlación.



A la vista del diagrama de dispersión, puede apreciarse que la relación lineal es débil.

Se calcula el coeficiente de correlación, añadiendo a la tabla las filas necesarias para el cálculo de las varianzas y la covarianza.

x_i	0	1	2	2	2	3	4	4	5	1	2	2	2	4	4	4	7	7	56
y_i	17	9	5	7	12	5	4	9	10	5	2	4	14	2	5	12	12	21	155
x_i^2	0	1	4	4	4	9	16	16	25	1	4	4	4	16	16	16	49	49	238
y_i^2	289	81	25	49	144	25	16	81	100	25	4	16	196	4	25	144	144	441	1809
x_iy_i	0	9	10	14	24	15	16	36	50	5	4	8	28	8	20	48	84	147	526

$$\bar{x} = \frac{56}{18} = 3,11 \qquad s_x^2 = \frac{238}{18} - 3,11^2 = 3,543$$

$$\bar{y} = \frac{155}{18} = 8,61 \qquad s_y^2 = \frac{1809}{18} - 8,61^2 = 26,349$$

$$s_{xy} = \frac{526}{18} - 3,11 \cdot 8,61 = 2,432$$

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{2,432^2}{3,543 \cdot 26,349} = 0,0634 \Rightarrow r = 0,2517$$

De modo que se confirma que la relación lineal entre el índice de Lee (*Y*) y el índice de disfunciones estomatológicas (*X*) es débil, ya que el coeficiente de correlación es pequeño. Por otro lado, el índice de disfunciones apenas explica un poco más del 6% de la variabilidad observada en el índice de Lee.

16. Ejercicio resuelto.

17. Para evaluar la relación de la tasa agua/cemento (*X*) con la resistencia (*Y*) del material de construcción resultante, se han tomado los siguientes datos:

X	1,21	1,29	1,37	1,46	1,62	1,79
Y	1,302	1,231	1,061	1,040	0,803	0,711

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

b) ¿Cuál será la resistencia esperada del material si la tasa fuera 1,4?

a) Ampliamos la tabla con las columnas correspondientes para calcular la recta de regresión de *Y* sobre *X*:

X	*Y*	x_i^2	y_i^2	x_iy_i
1,21	1,302	1,4641	1,6952	1,5754
1,29	1,231	1,6641	1,5154	1,5880
1,37	1,061	1,8769	1,1257	1,4536
1,46	1,04	2,1316	1,0816	1,5184
1,62	0,803	2,6244	0,6448	1,3009
1,79	0,711	3,2041	0,5055	1,2727
8,74	6,148	12,9652	6,5682	8,7089

$$\bar{x} = \frac{8,74}{6} = 1,4567 \quad s_x^2 = \frac{12,9652}{6} - 1,4567^2 = 0,0390$$

$$\bar{y} = \frac{6,148}{6} = 1,0247 \quad s_y^2 = \frac{6,5682}{6} - 1,0247^2 = 0,0448$$

$$s_{xy} = \frac{8,7089}{6} - 1,4567 \cdot 1,0247 = -0,0411$$

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{-0,0411^2}{0,0390 \cdot 0,0448} = 0,9634 \Rightarrow r = -0,9841$$

Calculamos la recta de regresión de *Y* sobre *X*:

$$y = 1,0247 + \frac{-0,0411}{0,0390}(x - 1,4567) \Rightarrow y = -1,0538x + 2,5598$$

b) Si la tasa agua/cemento es de 1,4 para obtener la resistencia esperada del material resultante basta con sustituir $x = 1,4$ en la ecuación obtenida en el apartado anterior $y = -1,0538 \cdot 1,4x + 2,5598 = 1,0845$

18. La edad (*X*) en años y el peso (*Y*) en kilos de 5 niños se recogen en la tabla.

X	2	6	8	7	4
Y	15	25	34	33	19

a) ¿Cuál es porcentaje de variabilidad del peso explicado por la edad?

b) ¿Qué peso se espera que tenga un niño de 5 años?

a) Debemos calcular el coeficiente de determinación, añadiendo a la tabla las columnas necesarias para ello.

X	*Y*	x_i^2	y_i^2	x_iy_i
2	15	4	225	30
6	25	36	625	150
8	34	64	1156	272
7	33	49	1089	231
4	19	16	361	76
27	126	169	3456	759

$$\bar{x} = \frac{27}{5} = 5,4 \qquad s_x^2 = \frac{169}{5} - 5,4^2 = 4,64$$

$$\bar{y} = \frac{126}{5} = 25,2 \qquad s_y^2 = \frac{3456}{5} - 25,2^2 = 56,16$$

$$s_{xy} = \frac{759}{5} - 5,4 \cdot 25,2 = 15,72 \; ; \; R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{15,72^2}{4,64 \cdot 56,16} = 0,9483$$

Por tanto, el 94,83 % de la variabilidad del peso de los niños (*Y*) viene explicada por su edad (*X*)

b) Con los datos obtenidos en el apartado anterior, se confecciona la ecuación de la recta de regresión del peso (*Y*) sobre la edad (*X*).

$$y = 25,2 + \frac{15,72}{4,64}(x - 5,4) \Rightarrow y = 3,39x + 6,89$$

Para calcular el peso esperado de un niño de 5 años, basta con sustituir $x = 5$ en la ecuación de regresión:

$$y = 3,39 \cdot 5 + 6,89 = 23,84 \text{ kg}$$

19. Ejercicio interactivo.

20. Ejercicio resuelto.

21. La tabla siguiente recoge el peso (*X*, kg) y la presión sistólica (*Y*, mmHg) de 10 hombres.

X	75	76	70	96	86	95	91	68	72	77
Y	130	133	128	151	150	140	148	125	133	135

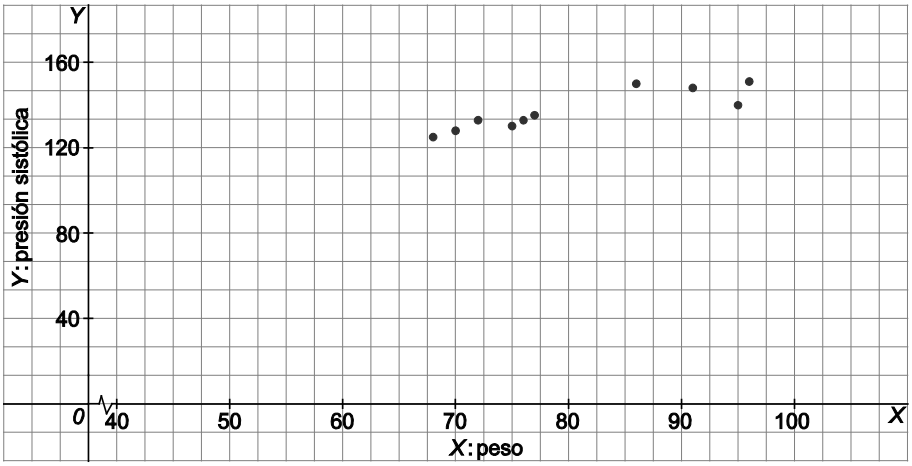
a) Representa la nube de puntos.

b) Escribe la recta de regresión completa de la presión sanguínea sobre el peso.

c) Ajusta una recta de regresión de *Y* sobre *X* que pase por el origen.

d) Compara los dos modelos ajustados ¿cuál es mejor?

a)



b) Ampliando la tabla con las columnas correspondientes para calcular la recta de regresión completa de *Y* sobre *X*:

$\overline{x} = \frac{806}{10} = 80,6 \qquad s_x^2 = \frac{65\,956}{10} - 80,6^2 = 99,24 \qquad \overline{y} = \frac{1373}{10} = 137,3$

$s_y^2 = \frac{189\,317}{10} - 137,3^2 = 80,41$

$s_{xy} = \frac{111453}{10} - 80,6 \cdot 137,3 = 78,92$

$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{78,92^2}{99,24 \cdot 80,41} = 0,78051$

X	*Y*	x_i^2	y_i^2	$x_i y_i$
75	130	5625	16 900	9750
76	133	5776	17 689	10 108
70	128	4900	16 384	8960
96	151	9216	22 801	14 496
86	150	7396	22 500	12 900
95	140	9025	19 600	13 300
91	148	8281	21 904	13 468
68	125	4624	15 625	8500
72	133	5184	17 689	9576
77	135	5929	18 225	10 395
806	1373	65 956	189 317	111 453

La ecuación de la recta de regresión completa de *Y* sobre *X*:

$y = 137,3 + \frac{78,92}{99,24}(x - 80,6) \Rightarrow y = 0,795x + 73,203$

c) Calculamos $b = \frac{111453}{65956} = 1,69$. La recta de regresión que pasa por el origen es: $y = 1,69x$.

d) Como $\text{ECM} = 80,41\left(1 - \frac{78,92^2}{99,24 \cdot 80,41}\right) = 17,649$, $\text{ECM}_0 = \frac{189\,317}{10}\left(1 - \frac{111453^2}{65\,956 \cdot 189\,317}\right) = 98,279$

$R^2 = 0,78051$ y $R_0^2 = 0,995$. Entonces, aunque $R_0^2 > R^2$, no es razonable ajustar una recta que pase por el origen ya que $\text{ECM} < \text{ECM}_0$.

22. Ejercicio resuelto.

23. La tabla siguiente muestra los datos del número medio de bacterias (*Y*) en un alimento en conserva dependiendo del tiempo (*X*, en minutos) de exposición a una fuente de calor:

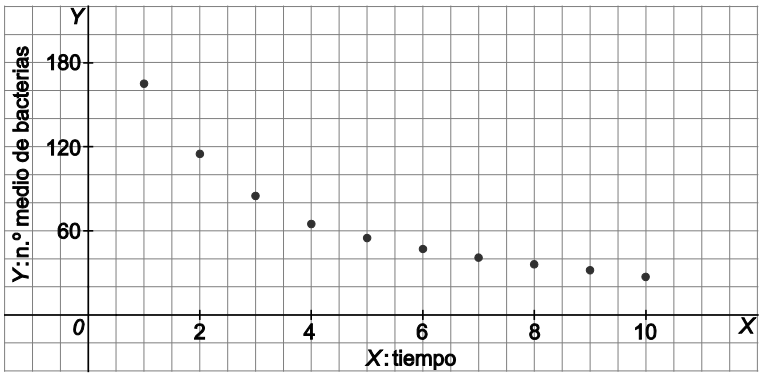
X	1	2	3	4	5	6	7	8	9	10
Y	165	115	85	65	55	47	41	36	32	27

a) Dibujar el diagrama de dispersión y razona si sería adecuado ajustar un modelo lineal.

b) Ajusta un modelo de regresión lineal de *Y* sobre *X* y valora la adecuación del ajuste.

c) Transforma la variable *Y* en $Z=\ln Y$, halla la recta de regresión de *Z* sobre *X* y compara este ajuste con el anterior.

a)



El diagrama de dispersión parece indicar que un ajuste lineal no es el más adecuado

b) Completando la tabla con las columnas correspondientes para calcular la recta de regresión de *Y* sobre *X*:

X	Y	x_i^2	y_i^2	x_iy_i
1	165	1	27 225	165
2	115	4	13 225	230
3	85	9	7225	255
4	65	16	4225	260
5	55	25	3025	275
6	47	36	2209	282
7	41	49	1681	287
8	36	64	1296	288
9	32	81	1024	288
10	27	100	729	270
55	668	385	61 864	2600

$$\overline{x}=\frac{55}{10}=5,5 \qquad s_x^2=\frac{385}{10}-5,5^2=8,25$$

$$\overline{y}=\frac{668}{10}=66,8 \qquad s_y^2=\frac{61864}{10}-66,8^2=1724,16$$

$$s_{xy}=\frac{2600}{10}-5,5\cdot 66,8=-107,4 \quad R^2=\frac{s_{xy}^2}{s_x^2 s_y^2}=\frac{-107,4^2}{8,25\cdot 1724,16}=0,811$$

La recta de regresión de *Y* sobre *X* es: $y=-13,02x+138,4$ con $R^2=0,811$, que permite disponer de un buen ajuste lineal.

c) Los nuevos datos quedarían:

X	$Z=\ln Y$	x_i^2	z_i^2	x_iz_i
1	5,105 945 47	1	26,070 6792	5,105 945 47
2	4,744 932 13	4	22,514 380 9	9,489 864 26
3	4,442 651 26	9	19,737 1502	13,327 953 8
4	4,174 387 27	16	17,425 509 1	16,697 5491
5	4,007 333 19	25	16,058 719 3	20,036 665 9
6	3,850 147 6	36	14,823 636 6	23,100 885 6
7	3,713 572 07	49	13,790 617 5	25,995 004 5
8	3,583 518 94	64	12,841 608	28,668 151 5
9	3,465 735 9	81	12,011 325 3	31,191 623 1
10	3,295 836 87	100	10,862 540 6	32,958 368 7
55	40,384 060 7	385	166,136167	206,572 012

$$\overline{z}=\frac{40,384}{10}=4,0384$$

$$s_z^2=\frac{166,136}{10}-4,0384^2=0,3049$$

$$s_{xz}=\frac{206,572}{10}-5,5\cdot 4,0384=-1,554$$

$$R^2=\frac{s_{xy}^2}{s_x^2 s_y^2}=\frac{(-1,554)^2}{8,25\cdot 0,3049}=0,96$$

La recta de regresión de *Z* sobre *X* es: $z=-0,1884x+5,0744$

$R^2=0,96$, que permite mejorar la proporción de variabilidad explicada por la variable regresora.

24 a 29. Ejercicios resueltos.

EJERCICIOS

Distribuciones unidimensionales

30. En las dos últimas semanas, el tiempo en espera, en minutos, en una parada de autobús ha sido:

10 3 3 7 1 5 9 12 11 6 8 2 9 2

a) Halla el tiempo medio de espera de los catorce días.

b) Determinar la mediana y razona si, en este caso, es un valor más o menos informativo que la media.

c) Calcula los cuartiles.

d) Halla el rango y la desviación típica de los datos.

a) Basta sumar los datos y dividir por 14, para obtener el tiempo medio de espera:

$$\overline{x} = \frac{10+3+3+7+1+5+9+12+11+6+8+2+9+2}{14} = 6,286 \text{ min}$$

b) Se ordenan los datos de menor a mayor: 1 2 2 3 3 5 6 7 8 9 9 10 11 12

Como el número de datos es par, la mediana es cualquier valor M_e del intervalo [6, 7], ya que en cualquier caso deja el 50% de los datos es mayor o igual (menor o igual) que cualquier valor en el intervalo.En este caso, al estar muy próximas, la media y la mediana tienen el mismo valor informativo en cuanto a la localización de la variable.

c) Con los datos ordenados en el apartado b), se calculan el primer y el tercer cuartil, el segundo es la mediana y se ha calculado ya.

El 25% de 14 es 3,5; luego Q_1 es cualquier valor en el intervalo [2,3].

El 75% de 14 es 10,5; luego $Q_3 = 9$.

d) Rango = 12 – 1 = 11

La varianza: $s_x^2 = \frac{10^2+3^2+3^2+7^2+1^2+5^2+9^2+12^2+11^2+6^2+8^2+2^2+9^2+2^2}{14} - 6,286^2 = 12,49$

La desviación típica: $s_x = \sqrt{12,49} = 3,534$ min

31. En una empresa de ITV (Inspección Técnica de Vehículos) se elige una muestra de 50 vehículos en los que se mide el nivel de emisión de CO_2, en g/km. Los resultados se presentan agrupados en la siguiente tabla:

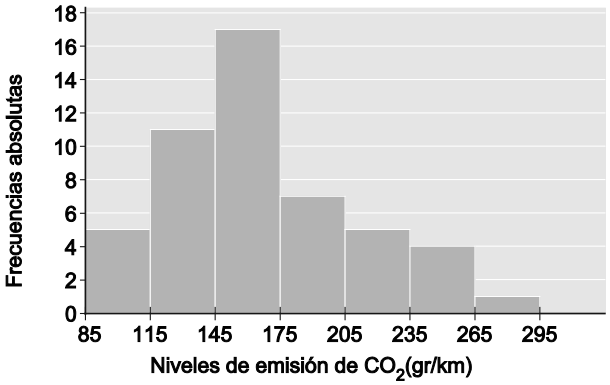
Clases	[85, 115)	[115, 145)	[145, 175)	[175, 205)	[205, 235)	[235, 265)	[265, 295]
f_i	5	11	17	7	5	4	1

a) Dibuja el histograma de frecuencias.

b) Calcula la media, la moda y los cuartiles.

c) Determina la varianza y la desviación típica.

a) El histograma de frecuencias absolutas es:



b) Se añaden a la tabla la columna con la marca de clase y las necesarias para el cálculo de los parámetros de localización y dispersión:

Clases	f_i	x_i	f_jx_j	$f_jx_j^2$	F_i
[85, 115)	5	100	500	50 000	5
[115, 145)	11	130	1430	185 900	16
[145, 175)	17	160	2720	435 200	33
[175, 205)	7	190	1330	252 700	40
[205, 235)	5	220	1100	242 000	45
[235, 265)	4	250	1000	250 000	49
[265, 295]	1	280	280	78 400	50
	50		8360	1 494 200	

Media: $\overline{x} = \dfrac{8360}{50} = 167,2$

Intervalo modal: [145, 175)

Los cuartiles se obtienen por interpolación lineal:

$$Q_1 = 115 + \frac{30(12,5-5)}{11} = 135,46$$

$$Q_2 = M = 145 + \frac{30(25-16)}{17} = 160,88$$

$$Q_3 = 175 + \frac{30(37,5-33)}{7} = 194,29$$

c) La varianza se obtiene a partir de la tabla y la desviación típica es su raíz cuadrada positiva:

$$s_x^2 = \frac{1494\,200}{50} - 167,2^2 = 1928,16 \Rightarrow s_x = \sqrt{1928,16} = 43,911$$

32. Una máquina produce piezas redondas de 1 pulgada de diámetro. En un control de calidad, los diámetros de las 10 piezas han sido:

1,11 1,12 1,09 0,98 1,03 1,09 0,97 1,2 1,15 1,07

a) Calcular la media y la desviación típica.

b) Se considera que la máquina no pasa el control de calidad si la medida observada no se encuentra en el intervalo $(\overline{x} - 1,96s_x, \overline{x} + 1,96s_x)$. ¿Hay alguna pieza en la muestra que no supera el control de calidad?

a) La media de las observaciones es: $\overline{x} = 1,081$

La deviación típica: $s_x^2 = \dfrac{11,7323}{10} - 1,081^2 = 0,004669 \Rightarrow s_x = \sqrt{0,004669} = 0,06833$

b) El intervalo $(\overline{x} - 1,96s_x, \overline{x} + 1,96s_x) = (1,081 - 1,96 \cdot 0,068\,33; 1,081 + 1,96 \cdot 0,068\,33) = (0,947; 1,215)$. Por lo tanto, todas las piezas observadas superan el control de calidad.

Distribuciones bidimensionales

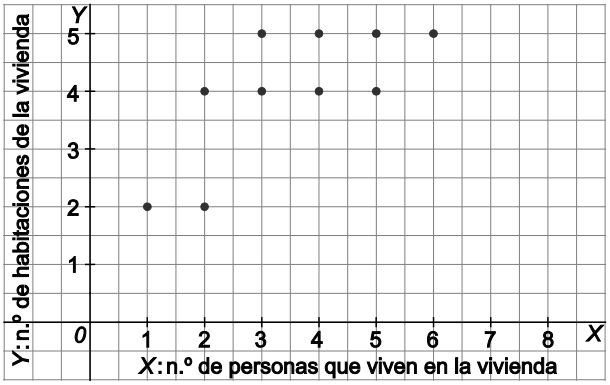
33. En una muestra de diez viviendas de una urbanización se han contabilizado el número de personas (*X*) que viven en cada una y el número de habitaciones (*Y*) que tiene. La tabla siguiente contiene los datos recogidos:

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

a) Dibuja el diagrama de dispersión. ¿Existe relación lineal entre las variables?

b) Calcula la covarianza y R^2. Explica los resultados.

a)



El diagrama de dispersión muestra una cierta relación lineal moderada entre las variables.

b) Para el cálculo de las varianzas y de la covarianza, es preciso ampliar la tabla con las columnas correspondientes:

$$\overline{x} = \frac{35}{10} = 3,5 \qquad s_x^2 = \frac{145}{10} - 3,5^2 = 2,25$$

$$\overline{y} = \frac{40}{10} = 4 \qquad s_y^2 = \frac{172}{10} - 4^2 = 1,2$$

$$s_{xy} = \frac{152}{10} - 3,5 \cdot 4 = 1,2$$

X	*Y*	x_i^2	y_i^2	$x_i y_i$
5	5	25	25	25
3	4	9	16	12
2	2	4	4	4
4	4	16	16	16
1	2	1	4	2
3	5	9	25	15
6	5	36	25	30
2	4	4	16	8
5	4	25	16	20
4	5	16	25	20
35	40	145	172	152

Entonces el coeficiente de determinación es:

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{1,2^2}{2,25 \cdot 1,2} = 0,533$$

Que confirma la relación moderada entre las variables, el 53,3 % de la variabilidad observada en el número de habitaciones es explicada por el número de personas que habitan en las viviendas.

34. En una fábrica se quiere probar la resistencia al calor de una determinada clase de cerámica. Se elige, para ello, una muestra de 8 parejas de piezas idénticas (de la misma hornada). De cada pareja, una de las piezas fue sometida a pruebas de dureza antes del proceso térmico y se anotó su resistencia a la rotura (*X*, en kg) y la otra después del horneado (*Y*, en kg)

X	148	213	380	180	200	190	240	198
Y	138	161	323	190	210	191	215	190

a) Calcula la media y la mediana de la resistencia antes y después del proceso térmico.

b) ¿En qué caso hay mayor variabilidad, antes o después del proceso térmico?

c) ¿Se puede afirmar que existe relación lineal entre las dos variables? Justifica la respuesta.

Para responder las cuestiones planteadas se construye la tabla siguiente:

X	*Y*	x_i^2	y_i^2	x_iy_i
148	138	21 904	19 044	20 424
213	161	45 369	25 921	34 293
380	323	144 400	104 329	122 740
180	190	32 400	36 100	34 200
200	210	40 000	44 100	42 000
190	191	36 100	36 481	36 290
240	215	57 600	46 225	51 600
198	190	39 204	36 100	37 620
1749	1618	416 977	348 300	379 167

a) Los valores medios de la resistencia a la rotura (en kg) antes y después del proceso térmico son, respectivamente:

$$\overline{x} = \frac{1749}{8} = 218{,}625 \qquad \overline{y} = \frac{1618}{8} = 202{,}25$$

Mediana de la resistencia a la rotura antes del proceso térmico (*X*):

Es cualquier valor del intervalo [198, 200], ya que se trata de un número par de observaciones, que en orden creciente resultan 148, 180, 190, 198, 200, 213, 240 y 380. Por tanto cualquier valor entre las observaciones cuarta y quinta puede ser elegido como mediana, puesto que es mayor o igual (menor o igual) que el 50% de los datos.

De forma análoga, cualquier valor del intervalo [190, 191] es mediana de la resistencia a la rotura después del proceso térmico (*Y*).

b) Las varianzas de *X* e *Y* son:

$$s_x^2 = \frac{416\,977}{8} - 218{,}625^2 = 4325{,}23 \qquad s_y^2 = \frac{348\,300}{8} - 202{,}25^2 = 2632{,}44$$

Por tanto, presentan mayor variabilidad las observaciones de *X*: resistencia a la rotura antes del proceso térmico de horneado.

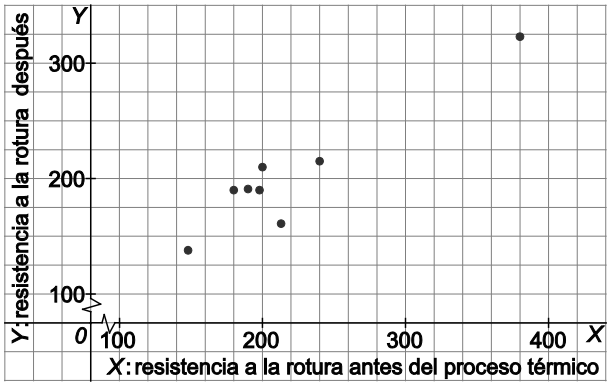
c) Para responder a esta cuestión se puede representar la nube de puntos de la distribución conjunta y calcular el coeficiente de correlación lineal.

$$s_{xy} = \frac{379\,167}{8} - 218{,}625 \cdot 202{,}25 = 3178{,}97$$

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{3178{,}972^2}{4325{,}23 \cdot 2632{,}44} = 0{,}88757 \Rightarrow$$

$$\Rightarrow r = 0{,}94211$$

En el diagrama se puede observar una fuerte relación lineal, que es confirmada por el valorpróximo a 1 del coeficiente de correlación lineal, *r*,con una observación muy influyente (*x*=380, *y*=323)en el modelo de regresión lineal.



35. Para compararel rendimiento de los trabajadores de una cadena de producción y el del trabajo automatizado mediante robots se planifican 5 tareas y se mide la tasa de rendimiento de los empleados (*X*) y del proceso automatizado (*Y*).

X	185	175	240	254	185
Y	180	186	269	250	216

a) Compara ambos rendimientos, en función de su media y su varianza o su desviación típica.

b) Calcula la covarianza y razona si puede existir relación entre ambos rendimientos.

Se amplía la tabla con las columnas necesarias para realizar los cálculos:

x	y	x^2	y^2	$x\,y$
185	180	34 225	32 400	33 300
175	186	30 625	34 596	32 550
240	269	57 600	72 361	64 560
254	250	64 516	62 500	63 500
185	216	34 225	46 656	39 960
1039	1101	221 191	248 513	233 870

a) La media y varianza de cada tasa de rendimiento son, respectivamente:

$$\overline{x} = \frac{1039}{5} = 207,8 \qquad s_x^2 = \frac{221191}{5} - 207,8^2 = 1057,36 \Rightarrow s_x = 32,52$$

$$\overline{y} = \frac{1101}{5} = 220,2 \qquad s_y^2 = \frac{248513}{5} - 220,2^2 = 1214,56 \Rightarrow s_y = 34,85$$

La tasa media de rendimiento automatizado es 12,4 puntos mayor que la manual (de los empleados) y también la dispersión es mayor en el proceso automatizado, si bien no en exceso, como se puede ver al comparar las desviaciones típicas.

b) La covarianza se obtiene a partir de los datos de la tabla:

$$s_{xy} = \frac{233870}{5} - 207,8 \cdot 220,2 = 1016,44 \qquad R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{1016,44^2}{1057,36 \cdot 1214,56} = 0,8045 \Rightarrow r = 0,897$$

Existe relación directa entre ambas variables y como el valor de la covarianza no está muy alejado del de ambas varianzas, la relación es relativamente fuerte.

36. La tabla siguiente recoge la puntuación obtenida por 25 trabajadores de una empresa en un test psicotécnico (*Y*) y las ausencias al trabajo en el último mes (*X*):

Y / *X*	[40, 60)	[60, 80)	[80, 100)
0	2	4	5
1	2	4	2
2	1	3	2

a) Halla la media y la varianza de las distribuciones marginales.

b) Calcula la media y la varianza de la distribución de *Y* condicionada al valor $X = 0$.

c) Obtén la covarianza e interpreta el resultado.

a) Se hallan las distribuciones marginales y se completa la tabla con las columnas necesarias para calcular la media y la varianza.

Marginal de X:

x_i	f_i	$f_i x_i$	$f_i x_i^2$
0	11	0	0
1	8	8	8
2	6	12	24
	25	20	32

Marginal de Y:

Y	f_i	y_i	$f_i y_i$	$f_i y_i^2$
[40, 60)	5	50	250	12 500
[60, 80)	11	70	770	53 900
[80, 100)	9	90	810	72 900
	25		1830	139 300

Media de *X*: $\overline{x} = \frac{20}{25} = 0,8$

Media de *Y*: $\overline{y} = \frac{1830}{25} = 73,2$

Varianza de *Y*: $s_x^2 = \frac{32}{25} - 0,8^2 = 0,64$

Varianza de *Y*: $s_y^2 = \frac{139\,300}{25} - 73,2^2 = 213,76$

b) Se halla la distribución de *Y* condicionada a $X = 0$ y se completa la tabla con las columnas necesarias para calcular la media y la varianza.

Y|*x*= 0	f_{0j}	x_{0j}	$f_{0j} x_{0j}$	$f_{0j} x_{0j}^2$
[40, 60)	2	50	100	5000
[60, 80)	4	70	280	19 600
[80, 100)	5	90	450	40 500
	11		830	65 100

$\overline{y}\,|_{x=0} = \frac{830}{11} = 75,45$

$s^2_{y|_{x=0}} = \frac{65\,100}{25} - 75,45^2 = 25,479$

c) Se completa la tabla para calcular la covarianza.

x_i	y_i	f_{ij}	$f_{ij} x_i y_j$
0	50	2	0
0	70	4	0
0	90	5	0
1	50	2	100
1	70	4	280
1	90	2	180
2	50	1	100
2	70	3	420
2	90	2	360
		25	1440

$s_{xy} = \frac{1440}{25} - 0,8 \cdot 73,2 = -0,96 \Rightarrow$ La relación entre las variables es inversa.

Regresión lineal y correlación

37. Los datos de la tabla siguiente se refieren al consumo mensual de energía eléctrica (*X*, en kWh) y al pico de demanda en una hora (*Y*, kW) en una empresa.

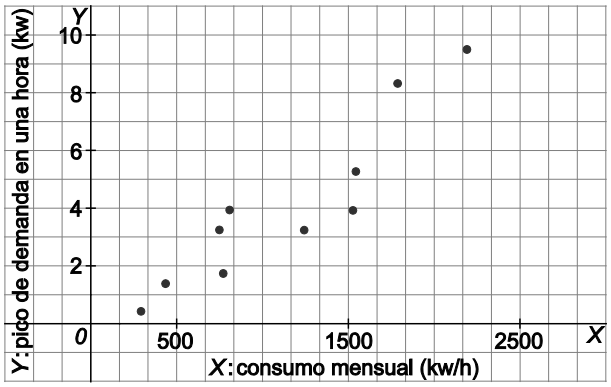
X	292	2189	747	435	1543	770	808	1242	1787	1526
Y	0,44	9,5	3,25	1,39	5,28	1,74	3,94	3,24	8,38	3,93

a) Dibuja el diagrama de dispersión. Valora la conveniencia de un ajuste lineal.

b) Escribe la recta de regresión del pico de demanda en función del consumo mensual y halla la varianza residual.

c) ¿Qué porcentaje de la variabilidad del pico de demanda es explicada por el consumo mensual?

a) El diagrama de dispersión se muestra al margen y a la vista de la nube de puntos parece razonable ajustar una recta de regresión que explique el pico de demanda en una hora (*Y*) en función del consumo mensual (*X*) en esta empresa.



b) Construimos la tabla con los datos necesarios para el cálculo de las medias, las varianzas y covarianzas, a partir de las cuales se obtiene la recta de regresión pedida y la varianza residual (ECM).

X	*Y*	x_i^2	y_i^2	x_iy_i
292	0,44	85264	0,1936	128,48
2189	9,50	4 791 721	90,2500	20 795,50
747	3,25	558 009	10,5625	2427,75
435	1,39	189 225	1,9321	604,65
1543	5,28	2 380 849	27,8784	8147,04
770	1,74	592 900	3,0276	1339,80
808	3,94	652 864	15,523 6	3183,52
1242	3,24	1 542 564	10,4976	4024,08
1787	8,38	3 193 369	70,2244	14 975,06
1526	3,93	2 328 676	15,4449	5997,18
11 339	41,04	16 315 441	245,5347	61 533,71

$$\overline{x} = \frac{11\,339}{10} = 1133,9$$

$$s_x^2 = \frac{16\,315\,441}{10} - 1133,9^2 = 345\,814,9$$

$$\overline{y} = \frac{41,04}{10} = 4,109$$

$$s_y^2 = \frac{245,5347}{10} - 4,109^2 = 7,6696$$

$$s_{xy} = \frac{61\,623,06}{10} - 1133,9 \cdot 4,109 = 1503,111$$

De modo que la recta de regresión de *Y* sobre *X* es:

$$y = 4,109 + \frac{1503,111}{345\,814,9}(x - 1133,9) \Rightarrow y = 0,004\,35x - 0,8196$$

El ECM o varianza residual es: $\text{ECM} = s_y^2\left(1 - \frac{s_{xy}^2}{s_x^2 s_y^2}\right) = 7,6696\left(1 - \frac{1503,111^2}{345\,814,9 \cdot 7,6696}\right) = 1,136$

c) Calculamos el coeficiente de determinación: $R^2 = \frac{s_{xy}^2}{s_x s_y} = \frac{1503,111^2}{345\,814,9 \cdot 7,6696} = 0,8519$ .De manera que el 85,19 % de la variabilidad observada en el pico de demanda horaria es explicada por el consumo mensual.

38. A una muestra de 10 ratones de laboratorio se les suministró diferentes cantidades diarias (*X*, en mg) de un combinado de vitaminas, anotándose el peso ganado (*Y*, en g) por cada uno tras una semana.

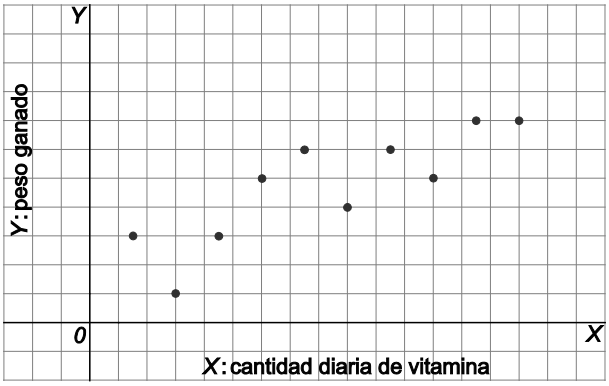
X	4	2	5	1	10	8	7	3	9	6
Y	5	1	6	3	7	5	6	3	7	4

a) Representa gráficamente la distribución y valora la viabilidad de un ajuste lineal.

b) Escribe la recta de regresión del peso ganado por semana en función de la dosis de vitamina diaria.

c) Determina la varianza residual y valora el resultado.

a) La representación gráfica de la distribución conjunta permite ver que es posible, con cierta fiabilidad, ajustar una recta a la nube de puntos.



b) La tabla para los cálculos se muestra a continuación:

X	*Y*	x_i^2	y_i^2	x_iy_i
4	5	16	25	20
2	1	4	1	2
5	6	25	36	30
1	3	1	9	3
10	7	100	49	70
8	5	64	25	40
7	6	49	36	42
3	3	9	9	9
9	7	81	49	63
6	4	36	16	24
55	47	385	255	303

$$\overline{x} = \frac{55}{10} = 5,5$$

$$s_x^2 = \frac{385}{10} - 5,5^2 = 8,25$$

$$\overline{y} = \frac{47}{10} = 4,7$$

$$s_y^2 = \frac{255}{10} - 4,7^2 = 3,41$$

$$s_{xy} = \frac{303}{10} - 5,5 \cdot 4,7 = 4,45$$

De modo que la recta de regresión de *Y* sobre *X* es: $y = 4,7 + \frac{4,45}{8,25}(x - 5,5) \Rightarrow y = 1,733x - 0,539$.

c) La varianza residual o ECM es: $\text{ECM} = s_y^2\left(1 - \frac{s_{xy}^2}{s_x^2 s_y^2}\right) = 3,41\left(1 - \frac{4,45^2}{8,25 \cdot 3,41}\right) = 1,0097$

Por tanto, la varianza residual es pequeña y permite afirmar que el ajuste lineal es bueno, aunque con las debidas reservas dado que el ECM depende de las unidades de media de *Y*.

39. En una muestra de 8 días elegidos durante el último mes, un viajante ha anotado los kilómetros recorridos (*X*) y los litros de combustible (*Y*) consumidos por su vehículo. Los datos son:

X	120	80	100	150	35	200	130	60
Y	6,5	6,5	7	8	4	9	7,5	5

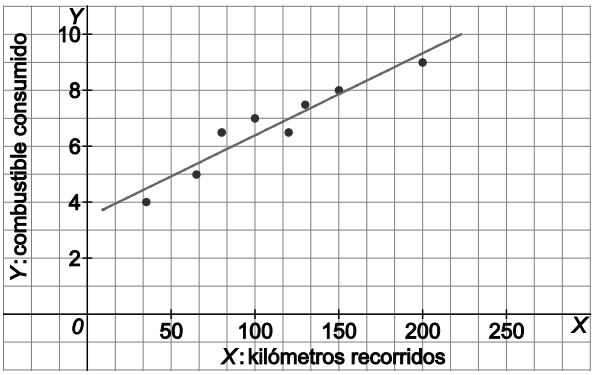
a) Dibuja la nube de puntos.

b) Halla la recta de regresión del consumo en función del kilometraje realizado.

c) Evalúa la precisión del ajuste de la nube puntos por la recta de regresión hallada.

d) ¿Cuál será el consumo esperado si un día tiene previsto viajar 115 km?

a) La nube de puntos se representa en el gráfico de dispersión y puede observarse que se presta a un ajuste lineal.



b) Para escribir la recta de regresión es preciso ampliar la tabla con las columnas correspondientes:

X	*Y*	x_i^2	y_i^2	x_iy_i
120	6,5	14 400	42,25	780
80	6,5	6400	42,25	520
100	7	10 000	49	700
150	8	22 500	64	1200
35	4	1225	16	140
200	9	40 000	81	1800
130	7,5	16 900	56,25	975
65	5	4225	25	325
880	53,5	115 650	375,75	6440

$$\overline{x} = \frac{880}{8} = 110 \qquad s_x^2 = \frac{115\,650}{8} - 110^2 = 2356{,}25$$

$$\overline{y} = \frac{53{,}5}{8} = 6{,}687\,5 \qquad s_y^2 = \frac{375{,}75}{8} - 6{,}687\,5^2 = 2{,}246\,1$$

$$s_{xy} = \frac{6440}{8} - 110 \cdot 6{,}6875 = 69{,}375$$

Por tanto la recta de regresión de *Y* sobre *X* es: $y = 6{,}687\,5 + \frac{69{,}375}{2356{,}25}(x - 110) \Rightarrow y = 0{,}0294x + 3{,}4488$.

c) En el gráfico se representa, también, la recta de regresión hallada en el apartado b) que, como se puede observar se ajusta muy bien a la nube de puntos. Este buen ajuste se confirma con el cálculo de los coeficientes de determinación y de correlación, ambos con valores próximos a 1:

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{69{,}375^2}{2{,}246\,1 \cdot 2356{,}25} = 0{,}909\,4 \Rightarrow r = \sqrt{0{,}909\,4} = 0{,}953\,6$$

d) El valor $x = 115$ km está dentro del rango de valores de la variable kilometraje recorrido, por tanto se puede utilizar la recta de regresión calculada para predecir, con fiabilidad, el consumo estimado en este caso:

$$y = 0{,}0294 \cdot 115 + 3{,}4488 = 6{,}83 \text{ L}$$

40. La temperatura media en ºC (*X*) y el gasto medio en calefacción en euros (*Y*) en el mes de diciembre en 7 ciudades se recoge en la tabla siguiente:

X	10	8	12	5	16	0	20
Y	145	140	110	170	80	200	40

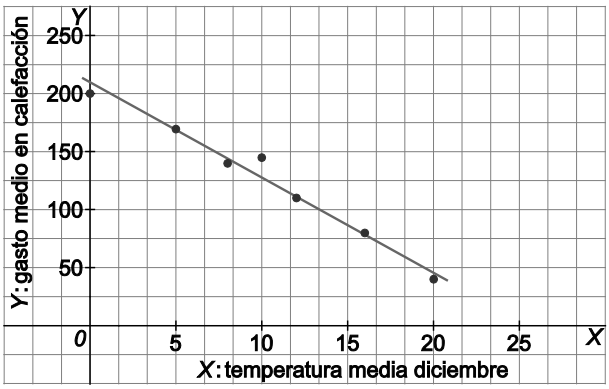
a) Dibuja el diagrama de dispersión y razona si se puede intuir una relación lineal entre las variables y de qué tipo.

b) Escribe la recta de regresión del gasto en calefacción en función de la temperatura y halla la varianza residual.

c) Valora el ajuste lineal a la nube de puntos.

d) Si la temperatura media en una ciudad en diciembre es de 14ºC, ¿a cuánto ascenderá el gasto esperado en calefacción?

a) La nube de puntos permite observar una relación lineal fuerte e inversa entre la temperatura media en diciembre y el gasto medio en calefacción en las 7 ciudades: a mayor temperatura media menor gasto en calefacción. Relación que, además, se corresponde con lo esperado



b) Para hallar la recta de regresión, se necesita ampliar la tabla:

X	*Y*	x_i^2	y_i^2	x_iy_i
10	145	100	21 025	1450
8	140	64	19 600	1120
12	110	144	12 100	1320
5	170	25	28 900	850
16	80	256	6400	1280
0	200	0	40 000	0
20	40	400	1600	800
71	885	989	129 625	6820

$$\overline{x} = \frac{72}{7} = 10{,}14 \qquad \overline{y} = \frac{885}{7} = 126{,}43$$

$$s_x^2 = \frac{989}{7} - 10{,}14^2 = 38{,}41$$

$$s_y^2 = \frac{129\,625}{7} - 126{,}43^2 = 2533{,}67$$

$$s_{xy} = \frac{6820}{7} - 126{,}43 \cdot 10{,}14 = -308{,}06$$

Por tanto la recta de regresión de *Y* sobre *X* es: $y = 126{,}43 + \frac{-308{,}06}{38{,}41}(x - 110) \Rightarrow y = -8{,}02x + 207{,}78$.

La varianza residual o ECM es: $\text{ECM} = s_y^2\left(1 - \frac{s_{xy}^2}{s_x^2 s_y^2}\right) = 2533{,}67\left(1 - \frac{(-308{,}06)^2}{38{,}41 \cdot 2533{,}67}\right) = 62{,}8$

c) En el gráfico se observa que el ajuste es muy bueno y ello se confirma con el cálculo de los coeficientes de determinación y correlación.

$$R^2 = \frac{(-308{,}06)^2}{38{,}41 \cdot 2533{,}67} = 0{,}975\,2 \Rightarrow r = -\sqrt{0{,}975\,2} = -0987\,5$$

El primero, muy próximo a 1, indica que el 97,52 % de la variabilidad observada en el gasto se debe a la temperatura media y el segundo, muy próximo a –1, que la correlación lineal es excelente.

d) Si la temperatura media es *x*= 14 ºC, puede predecirse el gasto medio en calefacción en el mes de diciembre, al ser un valor que está dentro del rango de valores del regresor.

$$y = -8{,}02 \cdot 14 + 207{,}78 = 95{,}5 \ €$$

41. La tabla siguiente muestra siete observaciones realizadas en laboratorio de un índice de rendimiento químico (*Y*) dependiendo de la concentración (*X*, en %) del catalizador de la reacción.

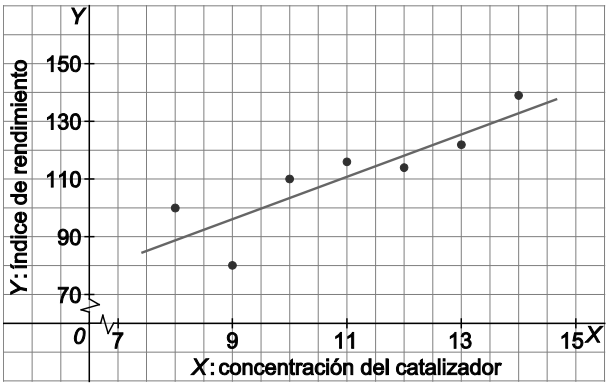
X	8	9	10	11	12	13	14
Y	100	80	110	116	114	122	139

a) Dibuja la nube de puntos.Halla la recta de regresión del índice de rendimiento en función de la concentración.

b) Calcula la varianza residual. Valora el resultado.

c) Un aumento de una unidad en la concentración del catalizador, ¿cuánto incrementa, en media, el rendimiento del proceso?

a) La nube de puntos se representa junto con la recta de regresión ajustada.



Para el cálculo de los coeficientes de la recta de regresión es preciso ampliar la tabla con las columnas correspondientes:

X	Y	x_i^2	y_i^2	x_iy_i
8	100	64	10 000	800
9	80	81	6400	720
10	110	100	12 100	1100
11	116	121	13 456	1276
12	114	144	12 996	1368
13	122	169	14 884	1586
14	139	196	19 321	1946
77	781	875	89 157	8796

A partir de la tabla, se obtienen las medias, las varianzas y la covarianza:

$$\overline{x} = \frac{77}{7} = 11\% \qquad \overline{y} = \frac{781}{7} = 111{,}57$$

$$s_x^2 = \frac{875}{7} - 11^2 = 4 \qquad s_y^2 = \frac{89\,157}{7} - 111{,}57^2 = 288{,}53$$

$$s_{xy} = \frac{8796}{7} - 111{,}57 \cdot 11 = 29{,}29$$

De manera que la recta de regresión de Y sobre X es:

$$y = 111{,}57 + \frac{29{,}29}{4}(x - 11) \Rightarrow y = 7{,}32x + 31{,}04$$

b) La varianza residual es:

$$\text{ECM} = s_y^2\left(1 - \frac{s_{xy}^2}{s_x^2}\right) = 288{,}53\left(1 - \frac{29{,}29^2}{4 \cdot 288{,}53}\right) = 74{,}12$$

Que puede considerarse un valor moderadamente bajo dados los valores que toma la variable Y, indicando un ajuste razonablemente bueno de la recta de regresión a la nube de puntos.

c) En el rango de los valores de X, por cada aumento en una unidad de la concentración del catalizador, el índice de rendimiento medio del proceso aumenta en 7,32 puntos, que es el valor de la pendiente de la recta de regresión.

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

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

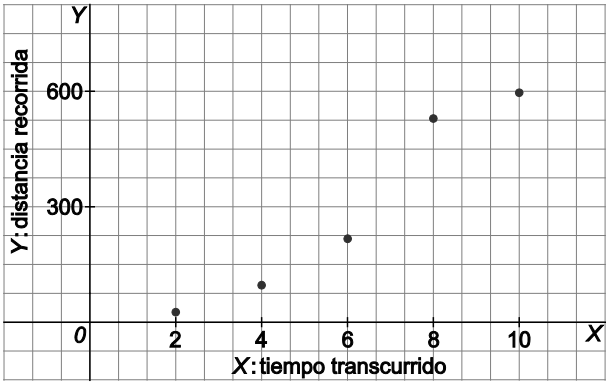
a) Dibuja la nube de puntos y calcula el coeficiente de correlación.

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

c) Transforma la variable *X* en $Z=X^2$ y realiza el ajuste de *Y* en función de *Z*, calculando el nuevo coeficiente de correlación. Valora los resultados.

d) Compara y valora los resultados obtenidos en los apartados a y c.

a) La nube de puntos es:



X	*Y*	x_i^2	y_i^2	x_iy_i
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

Completamos la tabla para calcular el coeficiente de correlación:

$$\overline{x}=\frac{30}{5}=6 \qquad \overline{y}=\frac{1464}{5}=292{,}8$$

$$s_x^2=\frac{220}{5}-6^2=8 \qquad s_y^2=\frac{694\,810}{5}-292{,}8^2=52\,230{,}16$$

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

Por tanto, el coeficiente de correlación *r* es: $R^2=\dfrac{s_{xy}^2}{s_x^2 s_y^2}=\dfrac{632{,}4^2}{8\cdot 52\,230{,}16}=0{,}957 \Rightarrow r=\sqrt{0{,}957}=0{,}978$

b) Con los datos calculados en el apartado a) se tiene que la recta de regresión de *Y* sobre *X* es:

$$y=292{,}8+\frac{632{,}4}{8}(x-6)\Rightarrow y=79{,}05x-181{,}5$$

c) Se calculan los nuevos datos y se completa la tabla para calcular el nuevo coeficiente de correlación.

Y	*X*	$Z=x^2$	z_i^2	z_iy_i
25	2	4	16	100
95	4	16	256	1520
216	6	36	1296	7776
530	8	64	4096	33 920
598	10	100	10 000	59 800
1464	30	220	15 664	103 116

$$\overline{z}=\frac{220}{5}=44 \qquad s_z^2=\frac{15\,664}{5}-44^2=1196{,}8$$

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

El nuevo coeficiente de correlación *r* es:

$$R^2=\frac{s_{yz}^2}{s_y^2 s_z^2}=\frac{7740^2}{52\,230{,}16\cdot 1196{,}8}=0{,}958\Rightarrow r=\sqrt{0{,}958}=0{,}979.$$

La recta de regresión de *Y* sobre *Z* es: $z=44+\dfrac{7740}{52\,230{,}16}(y-292{,}8)\Rightarrow y=0{,}148x+0{,}61$.

d) Aunque casi coinciden los dos coeficientes de determinación, el ECMen el apartado a) es $ECM=2245{,}9$ que es mayor que en b), que vale $ECM=50{,}26$. Por tanto la transformación mejora la aproximación lineal.

Síntesis

43. Con la finalidad de estudiar la relación del peso (*X*, en kg) con la medida de la presión sanguínea sistólica (*Y*), en una empresa se elige una muestra de 9 empleados y los datos recogidos fueron los siguientes:

X	74,8	70,3	86,2	78,0	78,9	88,5	67,6	77,1	87,1
Y	120	118	140	143	139	153	115	140	150

a) Calcula el peso medio, la presión sanguínea media y las desviaciones típicas de ambas variables.

b) Representa gráficamente los datos y razona si existe o no relación lineal entre las dos variables.

c) Calcula la covarianza. ¿Se puede afirmar que a mayor peso le corresponde mayor presión sanguínea?

d) Valora si existe relación lineal entre las variables calculando el coeficiente de correlación.

X	*Y*	x_i^2	y_i^2	x_iy_i
74,8	120	5595,04	14 400	8976
70,3	118	4942,09	13 924	8295,4
86,2	140	7430,44	19 600	12 068
78,0	143	6084	20 449	11 154
78,9	139	6225,21	19 321	10 967,1
88,5	153	7832,25	23 409	13 540,5
67,6	115	4569,76	13 225	7774
77,1	140	5944,41	19 600	10 794
87,1	150	7586,41	22 500	13 065
708,5	1218	56 209,61	166428	96 634

a) Se completa la tabla para calcular las medias y desviaciones típicas.

Peso medio: $\overline{x} = \frac{708}{9} = 78,72$

Presión sanguínea media: $\overline{y} = \frac{1218}{9} = 135,33$

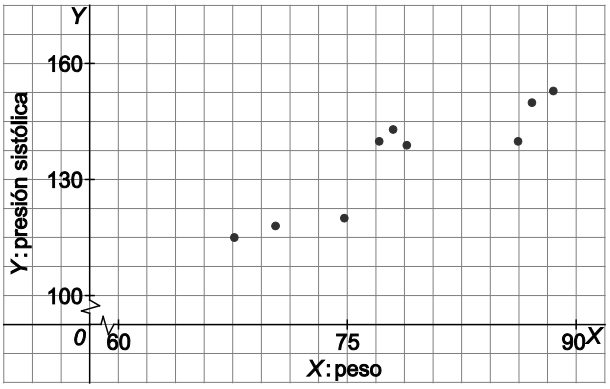
Varianza del peso y desviación típica:

$$s_x^2 = \frac{56209,61}{9} - 78,72^2 = 48,324 \Rightarrow s_x = \sqrt{48,324} = 6,952$$

Varianza de la presión sanguínea y desviación típica:

$$s_y^2 = \frac{166\,428}{9} - 135,33^2 = 176,89 \Rightarrow s_y = \sqrt{176,89} = 13,3$$

b) El diagrama de dispersión se representa en el gráfico y puede observarse que es razonable un ajuste lineal.



c) La covarianza es: $s_{xy} = \frac{96\,634}{9} - 78,72 \cdot 135,33 = 83,37 \Rightarrow$ la relación es directa. Se puede afirmar que a mayor peso le corresponde mayor presión sanguínea.

d) Coeficiente de correlación:

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{83,37^2}{48,324 \cdot 176,89} = 0,813 \Rightarrow r = \sqrt{0,813} = 0,902 \Rightarrow$$ Existe una relación lineal directa entre las variables.

44. Se conoce que el consumo de energía anual por habitante (*Y*, miles de kWh) está relacionado con la renta per cápita (*X*, miles de $).

Para estudiar cómo funciona esta relación en la región centroamericana 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

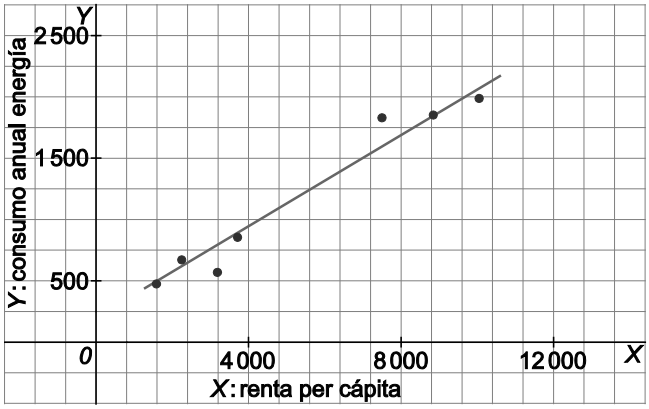
a) ¿Puede aproximarse, razonablemente, la nube de puntos por una recta?

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

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

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

a) La gráfica de dispersión muestra dos grupos de observaciones. Suponiendo que en la zona intermedia en la que no se dispone de datos, el comportamiento sea similar, se puede afirmar que la nube de puntos se puede ajustar razonablemente por una recta, que también se ha representado en el gráfico.



b) La tabla siguiente muestra los cálculos necesarios para la estimación de los coeficientes de la recta de regresión:

X	*Y*	x_i^2	y_i^2	x_iy_i
8647	1855	74 770 609	3 441 025	16 040 185
3708	855	13749 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

$$\overline{x} = \frac{36\,902}{7} = 5271{,}71 \qquad \overline{y} = \frac{8243}{7} = 1177{,}57$$

$$s_x^2 = \frac{263\,316\,370}{7} - 5271{,}71^2 = 9\,825\,652{,}776$$

$$s_y^2 = \frac{12\,483\,833}{5} - 1177{,}57^2 = 396\,730{,}245$$

$$s_{xy} = \frac{5\,699\,577\,711\,946}{7} - 177{,}57 \cdot 5271{,}71 = 1\,934\,433{,}735$$

De manera que la recta de regresión de *Y* sobre *X* es:

$$y = 1177{,}57 + \frac{1\,934\,433{,}735}{9\,825\,652{,}776}(x - 5271{,}71) \Rightarrow y = 0{,}1969x + 139{,}698$$

c) Se calcula el coeficiente de determinación: $R^2 = \dfrac{s_{xy}^2}{s_x^2 s_y^2} = \dfrac{1\,934\,433{,}735^2}{9\,825\,652{,}776 \cdot 396\,730{,}2449} = 0{,}959\,96$

De manera que el modelo de regresión explica prácticamente el 96% de la variabilidad observada en el consumo de energía anual por habitante. No obstante, conviene insistir en que no se tienen observaciones entre los dos grupos de datos mencionados antes y, por tanto, esta valoración debe realizarse con las debidas precauciones.

d) Si $x = 5000$ $, sustituyendo en la recta de regresión, resulta $y = 0{,}1969 \cdot 5000 + 139{,}698 = 1124{,}077$ kWh/hab.

Predicción que, con las precauciones señaladas en el apartado anterior, tiene alta fiabilidad puesto que el valor de *X* está dentro de su rango de valores experimentales y el nivel de correlación es muy alto ($r = 0{,}979\,77$).

CUESTIONES

45. 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?

La ecuación de la recta de regresión es $y = a + bx$, con $b = \frac{s_{xy}}{s_x^2}$ y $a = \bar{y} - b\bar{x}$. Por tanto:

a) El valor medio de *Y* es 3, puesto que en este caso *y*= 3 representa el valor "esperado" de *Y* cualquiera que sea el valor de *X*.

b) Como la pendiente de la recta de regresión es cero, la covarianza es cero, puesto que $b = \frac{s_{xy}}{s_x^2} = 0 \Rightarrow s_{xy} = 0$.

c) Dado que la covarianza es cero, el coeficiente de correlación también es cero,ya que $r = \frac{s_{xy}}{s_x s_y} = 0$.

d) La recta de regresión es horizontal y no existe relación lineal entre las variables *X* e *Y*, salvo que sea cual sea el valor de *X*, la observación de *Y* sea constante e igual a 3.

46. Dadas dos variable estadísticas *X* e *Y*, halla la recta de regresión de *Y* sobre *X* y el ECM, si 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).

Como las varianzas de *X* e *Y* son iguales, a partir del coeficiente de correlación se puede calcular la varianza de *X* (y la de *Y*): $r = \frac{s_{xy}}{s_x s_y} = \frac{s_{xy}}{s_x^2} \Rightarrow s_x^2 = \frac{s_{xy}}{r} = \frac{2,8}{0,8} = 3,5$

Entonces, la pendiente de la recta de regresión se calcula: $a = \frac{s_{xy}}{s_x^2} = \frac{2,8}{3,5} = 0,8$

Con lo que la recta de regresión de *Y* sobre *X* queda: $y = 0,8x + b$

Como el punto $(3,7)$ pertenece a la recta: $b = 7 - 0,8 \cdot 3 = 4,6$

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

47. 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) El coeficiente de correlación será 0,15 solo en caso de que las varianzas de *X* e *Y* sean iguales, ya que $b = r\frac{s_y}{s_x}$. Por lo que solo en el caso $s_y = s_x$ el coeficiente de correlación es 0,15.

b) Es cierto, porque el signo de la covarianza es igual que el de la pendiente de la recta, ya que: $a = \frac{s_{xy}}{s_x^2}$.

c) Para poder valorar esta afirmación sería preciso obtener el coeficiente de determinación, y no se tiene información suficiente para ello, salvo que se suponga que las varianzas de *X* e *Y* son iguales, en cuyo caso la afirmación es cierta, ya que entonces $R^2 = 0,15^2 = 0,0225$. Y solo el 2,25% de la variabilidad de *Y* es explicada por la variabilidad de *X*.

d) Si las varianzas de *X* e *Y* fueran iguales, entonces $r = 0,15$ y podría afirmarse que *X* e *Y* están débilmente correlacionadas. Pero, en general nada se puede afirmar al respecto.

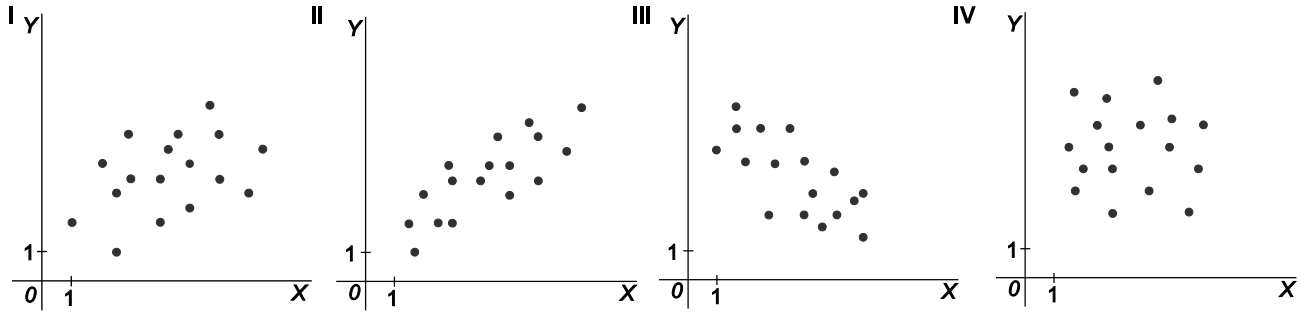
48. Una variable estadística bidimensional (*X*,*Y*), en la que *X* toma los valores 8, 13, 15, 16, 18, 20, 21 y 24, tiene como recta de regresión de *Y* sobre *X* a $y = 3,2 + 1,25x$. Calcula el valor estimado de *Y* cuando *X* tome los valores 1, 10, 20 y 100.

Si $x = 10$, entonces $y = 3,2 + 1,25 \cdot 10 = 15,7$

Si $x = 20$, entonces $y = 3,2 + 1,25 \cdot 20 = 28,2$

En los casos $x = 1$ y $x = 100$, no se puede estimar el valor utilizando la recta de regresión, puesto que ambos valores están lejos del rango de valores del regresor utilizados para calcular la recta de regresión.

49. 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) Al ser $r = -0,04$ la relación entre las dos variables es inversa y el ajuste entre la recta de regresión lineal y la nube de puntos es muy débil se tiene que corresponde a IV.

b) Puesto que $r = 0,4$ la relación es directa entre las dos variables pero el ajuste entre la recta de regresión y la nube de puntos es débil. Luego corresponde a I.

c) Dado que $r = -0,7$ la relación es inversa entre las dos variables y el ajuste entre la recta de regresión y la nube de puntos es fuerte. Por tanto, corresponde a III.

d) Como $r = 0,8$ la relación es directa entre las dos variables y el ajuste entre la recta de regresión y la nube de puntos es fuerte, entonces corresponde a II.

50. Dada la siguiente tabla de datos

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

a) Calcula el coeficiente de correlación.

b) Si a los valores de *X* se les multiplica por 3 y a los de *Y* por 2, ¿Cuál será ahora el coeficiente de correlación? Justifica la respuesta.

c) A los valores de *X* se les suma 2 y a los de *Y* se les resta 1. Razona cuál será entonces el valor del coeficiente de correlación.

a) Se completa la tabla para calcular las varianzas y la covarianza, que permiten obtener el coeficiente de correlación:

X	Y	x_i^2	y_i^2	$x_i y_i$
1	3	1	9	3
2	2	4	4	4
4	3	16	9	12
5	4	25	16	20
6	6	36	36	36
18	18	82	74	75

$$\bar{x} = \frac{18}{5} = 3,6 \qquad \bar{y} = \frac{18}{5} = 3,6$$

$$s_x^2 = \frac{82}{5} - 3,6^2 = 3,44 \qquad s_y^2 = \frac{74}{5} - 3,6^2 = 1,84$$

$$s_{xy} = \frac{75}{5} - 3,6 \cdot 3,6 = 2,04$$

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2}\sqrt{s_y^2}} = \frac{2,04}{\sqrt{3,44}\sqrt{1,84}} = 0,810\,85$$

b) El coeficiente de correlación es el mismo que en el apartado a), como se puede comprobar fácilmente, puesto que:

$$s_{3x2y} = 6s_{xy} \qquad s_{3x}^2 = 9s_x^2 \qquad s_{2y}^2 = 4s_y^2$$

Y, entonces:

$$r_{3x2y} = \frac{s_{3x2y}}{s_{3x}s_{2y}} = \frac{6s_{xy}}{3s_x \cdot 2s_y} = r_{xy}$$

c) Si a los valores de X se les suma 3, la varianza de $X + 3$ es igual que la de X, ya que lo único que se ha hecho es "trasladar" los valores de la variable. Lo mismo sucede con $Y - 1$ e Y. Luego:

$$s_{x+3}^2 = s_x^2 \qquad s_{y-1}^2 = s_y^2$$

Por idénticos motivos, tampoco cambia la covarianza, es decir:

$$s_{(x+3)(y-1)} = s_{xy}$$

Por lo que el coeficiente de correlación entre $X + 3$ e $Y-1$ es el mismo que entre X e Y.

PROBLEMAS

51. El índice de actividad de una sustancia radiactiva se miden en Becquerel por metro cúbico (Bq/m^3). Para investigar si en una determinada zona geográfica los niveles del isótopo radiactivo del radio, ^{226}Ra, superan el nivel máximo de exposición establecido por Sanidad, que es de (148 Bq/m^3), se toman 26 muestras de terreno. Los datos recogidos, en Bq/m^3, son:

54,02	**21,46**	**159,47**	**37,74**	**6,66**	**108,04**	**33,67**
15,91	**33,67**	**68,08**	**129,87**	**52,17**	**304,51**	**62,9**
74,74	**61,05**	**51,8**	**27,75**	**48,1**	**219,04**	**68,82**
155,77	**166,13**	**53,28**	**52,91**	**254,19**		

a) Calcula la media y la varianza sin agrupar los datos.

b) Agrupa los datos en 5 clases de igual longitud y calcula la media y la desviación de los datos agrupados. Compara los resultados con los del apartado anterior.

c) Se establece que si la media más dos veces la desviación típica supera el valor máximo establecido existe riesgo de contaminación radiactiva. ¿Es este el caso?

a) Para calcular la media, se suman todos los valores y se divide por 26, el resultado es:

$$\overline{x} = \frac{54,02 + 21,46 + \ldots + 254,19}{26} = 82,298$$

El cálculo de la desviación típica requiere obtener la varianza:

$$s_x^2 = \frac{54,02^2 + 21,46^2 + \ldots + 254,19^2}{26} - 89,298^2 = 5699,178 \Rightarrow s_x = 75,493$$

b) Se agrupan los datos en 5 clases de longitud 60, puesto que el rango es $304,88 - 6,29 = 298,59$. Empezando en 6 y terminando en 306:

Clases	f_i	x_i	f_ix_i	$f_ix_i^2$
[6, 66)	15	36	540	19 440
[66, 126)	4	96	384	36 864
[126, 186)	4	156	624	97 344
[186, 246)	1	216	216	46 656
[246, 306)	2	276	552	152 352
	26		2316	352 656

Con los datos de la tabla, la media y la desviación típica son: $\overline{x} = \frac{2316}{26} = 89,077$

$$s_x^2 = \frac{352\,656}{26} - 89,077^2 = 5628,98 \Rightarrow s_x = 75,027$$

Los valores son muy próximos a los obtenidos con los datos sin agrupar.

c) Tomando la media y la desviación típica de los datos sin agrupar (con los datos agrupados los resultados son equivalentes), el valor de la media más dos veces la desviación típica es 240,284, valor que supera claramente el valor de 148 Bq/m^3. Por lo que se concluye que existe riesgo de contaminación radiactiva.

52. La distribución de frecuencias de 200 personas adultas, entrevistadas en una ciudad pequeña, según su situación profesional (*X*) y su nivel de estudios (*Y*) se recoge en la tabla:

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

a) Escribe las distribuciones marginales y condicionadas.

b) De los que tienen estudios medios, ¿qué porcentaje son autónomos? ¿Y empleados?

c) ¿Son independientes estas variables?

a) Distribución marginal de *X*

X	f_i
Empleado fijo	54
Empleado temporal	70
Autónomo	30
Sin empleo	46
	200

Distribución marginal de *Y*

Y	f_i
Básicos	67
Medios	75
Altos	58
	200

Condicionada *X*|*Y*:

	Y		
X	Básicos	Medios	Altos
Empleado fijo	0,208 955	0,293 333	0,310 344
Empleado temporal	0,268 656	0,413 333	0,362 068
Autónomo	0,179 104	0,106 666	0,172 413
Sin empleo	0,343 283 58	0,186 666 67	0,155 172
	1	1	1

Condicionada *Y*|*X*:

	Y			
X	Básicos	Medios	Altos	
Empleado fijo	0,259 259	0,407 407	0,333 333	1
Empleado temporal	0,257 142	0,442 857	0,3	1
Autónomo	0,4	0,266 666	0,333 333 33	1
Sin empleo	0,5	0,304 347	0,195 652	1

b) Como hay 75 personas con estudios medios de los que 8 son autónomos y 31 + 22 = 53 son empleados, entonces se tiene que aproximadamente de los que tienen estudios medios el 10,7 % son autónomos y el 70,7 % son empleados.

c) Se construye la tabla de las distribuciones relativas conjunta y marginales.

h_{ij}	Básicos	Medios	Altos	$h_{i.}$
Empleado fijo	0,07	0,11	0,09	0,27
Empleado temporal	0,09	0,155	0,105	0,35
Autónomo	0,06	0,04	0,05	0,15
Sin empleo	0,115	0,07	0,045	0,23
$h_{.j}$	0,335	0,375	0,29	1

Como $h_{ij} \neq h_i h_j$, para algún *i*, *j*, por ejemplo, $0,09 \neq 0,335 \cdot 0,35$. Entonces las variables *X* e *Y* son dependientes.

53. Galton (1877) analizó la relación entre el diámetro de los guisantes (*X*) y el diámetro medio de sus descendientes (*Y*) (datos en centésimas de pulgada):

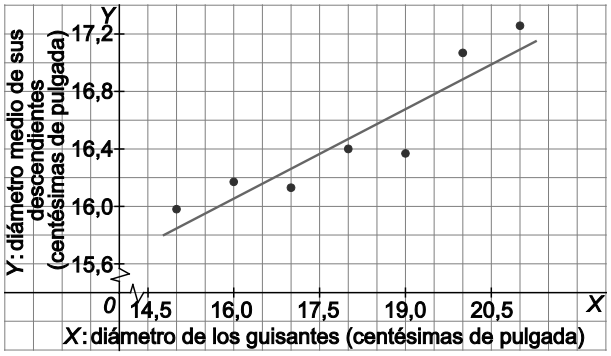
X	21	20	19	18	17	16	15
Y	17,26	17,07	16,37	16,4	16,13	16,17	15,98

a) Dibuja el diagrama de dispersión.

b) Escribe la recta de regresión de *Y* sobre *X*. ¿Qué conclusiones pueden extraerse?

c) Si un guisante tiene 4,5 milímetros de diámetro (0,177 pulgadas), ¿cuál será el diámetro esperado de sus descendientes?

a) En el gráfico se ha representado la nube de puntos junto con la recta de regresión de *Y* sobre *X* ajustada en el apartado siguiente.



b) Se completa la tabla para calcular los coeficientes de la recta de regresión de *Y* sobre *X*:

X	*Y*	x_i^2	y_i^2	x_iy_i
21	17,26	441	297,908	362,46
20	17,07	400	291,385	341,4
19	16,37	361	267,977	311,03
18	16,4	324	268,960	295,2
17	16,13	289	260,177	274,21
16	16,17	256	261,469	258,72
15	15,98	225	255,360	239,7
126	115,38	2296	1903,236	2082,72

$$\bar{x} = \frac{126}{7}18 \qquad s_x^2 = \frac{2296}{7} - 18^2 = 4$$

$$\bar{y} = \frac{115,38}{7} = 16,48 \qquad s_y^2 = \frac{1903,236}{7} - 16,48^2 = 0,2062$$

$$s_{xy} = \frac{2082,72}{7} - 18 \cdot 16,48 = 0,84$$

Por tanto, la recta de regresión del diámetro medio de los descendientes (*Y*) respecto al diámetro de sus predecesores (*X*) es:

$$y = 16,48 + \frac{0,84}{4}(x - 18) \Rightarrow y = 12,703 + ,021x$$

A la vista de los datos y de los cálculos realizados se pueden extraer, entre otras, algunas conclusiones:

- El diámetro medio de los guisantes cuyos predecesores son más "grandes" es menor que el de estos, mientras que el de aquellos cuyos predecesores son más "pequeños", es mayor (regresión a la media).
- Cuanto mayor es el tamaño de los guisantes, mayor es el tamaño medio de sus descendientes.
- La variabilidad observada en el tamaño medio de los descendientes es mucho menor que la de sus predecesores.
- Por cada centésima de aumento en el tamaño del guisante de siembra, aumenta 0,21 centésimas el tamaño medio de sus descendientes.

c) En este caso, con los resultados obtenidos en el apartado b), puede obtenerse el coeficiente de correlación:

$r = \frac{0,84}{\sqrt{4} \cdot \sqrt{0,2062}} = 0,9249$. Que confirma el alto nivel de correlación entre el diámetro de los guisantes sembrados y el diámetro medio de sus descendientes. De esta forma, se pueden realizar predicciones acerca del tamaño medio de los descendientes para diámetros de guisantes en el rango de valores de la tabla, como es el caso de $x = 17,7$ centésimas de pulgada.

$$y = 12,703 + ,021 \cdot 17,7$$

Es decir, el tamaño medio esperado para los descendientes de un predecesor de 0,177 pulgadas es de0,1642 pulgadas.

54. La tabla siguiente muestra el PIB per cápita (*X*, en miles de $) y el índice de natalidad (*Y*, en nacimientos/mil habitantes) en once países de Iberoamérica.

X	4,9	12,7	8,6	10,2	11,9	9,6	17,7	10,4	5,5	17,4	15,3
Y	24	20	20	19	17	17	17	17	17	14	13

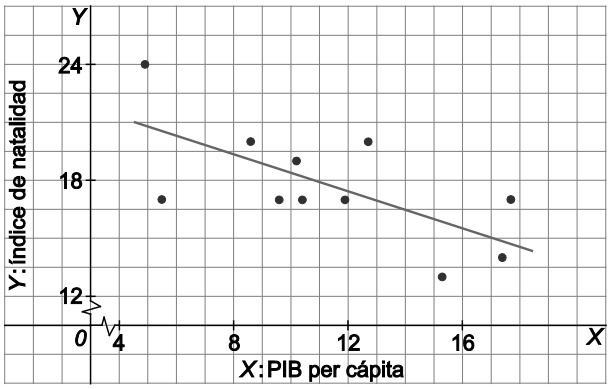
a) Representa gráficamente la distribución.

b) Calcula el coeficiente de correlación y señala la relación entre el PIB per cápita y el índice de natalidad.

c) Escribe la recta de regresión del índice de natalidad en función del PIB per cápita.

d) Si un país tiene un PIB per cápita de 14 mil dólares ¿Cuál será su índice de natalidad esperado?

a) En gráfico se representa la nube de puntos y la recta de regresión obtenida en el aparatado c).Se puede observar una tendencia decreciente, si bien el ajuste lineal a la nube de puntos no es muy bueno.



X	Y	x_i^2	y_i^2	x_iy_i
4,9	24	24,01	576	117,6
12,7	20	161,29	400	254
8,6	20	73,96	400	172
10,2	19	104,04	361	193,8
11,9	17	141,61	289	202,3
9,6	17	92,16	289	163,2
17,7	17	313,29	289	300,9
10,4	17	108,16	289	176,8
5,5	17	30,25	289	93,5
17,4	14	302,76	196	243,6
15,3	13	234,09	169	198,9
124,2	195	1585,62	3547	2116,6

b) Se completa la tabla con las columnas necesarias para los cálculos de este apartado y del siguiente

$$\overline{x} = \frac{12,24}{11} 11,29 \qquad s_x^2 = \frac{1585,62}{11} - 11,29^2 = 16,6626$$

$$\overline{y} = \frac{195}{11} = 17,73 \qquad s_y^2 = \frac{3547}{11} - 17,73^2 = 8,1983$$

$$s_{xy} = \frac{2116,6}{11} - 11,29 \cdot 17,73 = -7,7388$$

Por tanto, el coeficiente de correlación es:

$$r = \frac{-7,7388}{\sqrt{16,6626}\ \sqrt{8,1983}} = -0,6621$$

Que indica que a mayor PIB per cápita, menor índice de natalidad con una relación lineal moderada, tal como se intuía con la observación de la nube de puntos en el primer apartado.

c) La recta de regresión de *Y* sobre *X*, es decir, del índice de nacimientos respecto al PIB es:

$$y = 17,73 + \frac{-7,7387}{16,6626}(x - 11,29) \Rightarrow y = -0,4644x + 22,971$$

d) Si $x = 14$, entonces el número esperado de nacimientos por cada mil habitantes es:

$$y = -0,4644 \cdot 14 + 22,971 = 16,475$$

55. Se supone que el alargamiento (*Y*, en cm) de un cable está relacionado con la intensidad de la fuerza (*X*, en N) que se le aplica. Para estudiar el tipo de relación se toma una muestra de cinco cables de la misma clase y longitud y se les aplica distintas fuerzas. Los datos del alargamiento son:

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

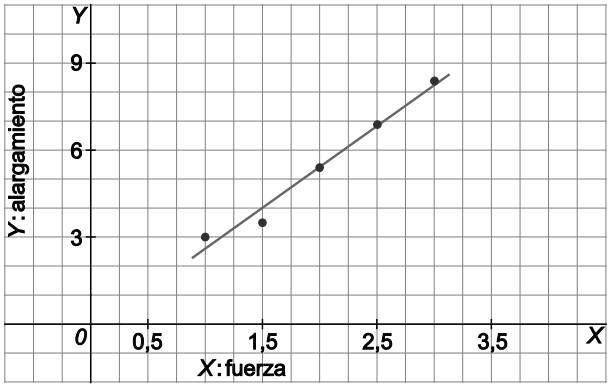
a) Dibuja la nube de puntos y comenta la relación que se observa.

b) Escribe la recta de regresión del alargamiento en función de la fuerza aplicada. ¿Cuánto aumentará el alargamiento por cada unidad de aumento en la fuerza?

c) ¿Cuál es el porcentaje de variabilidad del alargamiento que viene explicada por la variación en la fuerza que se aplica?

d) Si se aplica una fuerza de 1,2 N, ¿Cuál será el alargamiento esperado del cable? ¿Y si la fuerza que se aplica es de 2,2 N?

a) La nube de puntos junto con la recta de regresión ajustada en el aparatado b) se han representado en el gráfico y se observa una fuerte relación lineal directa entre la fuerza aplicada al cable y el alargamiento que se produce en el mismo: a mayor fuerza aplicada, mayor alargamiento. Situación que coincide con la intuición, por lo que parece razonable a pesar de contar con solo cinco observaciones.



b) Se amplía la tabla de datos con las columnas que se necesitan para efectuar los cálculos:

X	*Y*	x_i^2	y_i^2	x_iy_i
2	5,4	4,00	29,16	10,80
1,5	3,5	2,25	12,25	5,25
3	8,4	9,00	70,56	25,20
1	3	1,00	9,00	3,00
2,5	6,9	6,25	47,61	17,25
10	27,2	22,50	168,58	61,50

$$\bar{x} = \frac{10}{5} = 2 \qquad \bar{y} = \frac{27,2}{5} = 5,44$$

$$s_x^2 = \frac{22,5}{5} - 2^2 = 0,5$$

$$s_y^2 = \frac{168,58}{5} - 5,44^2 = 4,1224$$

$$s_{xy} = \frac{61,5}{115} - 2 \cdot 5,44 = 1,42$$

De manera que la recta de regresión del alargamiento (*Y*) en función de la fuerza aplicada (*X*) es:

$$y = 5,44 + \frac{1,42}{0,5}(x-2) \Rightarrow y = 2,84x - 0,24$$

De modo que por cada unidad (N) de aumento en la fuerza, el cable alargará 2,84 cm.

c) La respuesta la encontramos en el valor del coeficiente de determinación:

$$R^2 = \frac{1,42^2}{0,5 \cdot 4,1224} = 0,9783$$

Es decir, el 97,83% de la variabilidad observada en el alargamiento viene explicada por la variación en la fuerza aplicada.

d) Si x= 1,2 N, entonces $y = 2,84 \cdot 1,2 - 0,24 = 3,168$ cm

Si x= 2,2 N, entonces $y = 2,84 \cdot 2,2 - 0,24 = 6,008$ cm

56. En la tabla se dan los datos observados de la estatura de un bebé (*Y*, en cm) según el número de semanas de embarazo (*X*) a partir de la semana 32. Los datos se dan en la tabla siguiente:

X	32	33	34	35	36	37	38	39	40
Y	42	45	47	48,5	49	51	52	54	55

a) Representa gráficamente los datos.

b) Halla la recta de regresión de la estatura del bebé en función del número de semanas transcurridas y justifica la bondad del ajuste obtenido.

c) ¿Cuál es el porcentaje de variabilidad de la estatura explicada por el modelo de regresión?

a) La gráfica de dispersión de los datos muestra una fuerte relación lineal entre el número de semanas de embarazo y la estatura del bebé, dado el buen ajuste que se observa de la recta de regresión a la nube de puntos.

b) Se añaden a la tabla las columnas correspondientes:

X	*Y*	x_i^2	y_i^2	x_iy_i
32	42	1024	1764	1344
33	45	1089	2025	1485
34	47	1156	2209	1598
35	48,5	1225	2352,25	1697,5
36	49	1296	2401	1764
37	51	1369	2601	1887
38	52	1444	2704	1976
39	54	1521	2916	2106
40	55	1600	3025	2200
324	443,5	11 724	21 997,25	16 057,5

$$\overline{x} = \frac{324}{9} = 36 \qquad s_x^2 = \frac{11724}{9} - 36^2 = 6,67$$

$$\overline{y} = \frac{443,5}{9} = 49,3 \qquad s_y^2 = \frac{21997,25}{9} - 49,3^2 = 15,84$$

$$s_{xy} = \frac{16057}{9} - 36 \cdot 49,3 = 10,17$$

La recta de regresión de la estatura del bebé (*Y*) en función del número de semanas de embarazo (*X*) es:

$$y = 49,3 + \frac{10,17}{6,67}(x - 36) \Rightarrow y = 1,525x - 5,6$$

Que se ajusta muy bien a la nube de puntos como se puede ver en el gráfico del apartado a). Además, el coeficiente de correlación lineal.

$$r = \frac{10,17}{\sqrt{6,67}\ \sqrt{15,84}} = 0,9894$$

es muy próximo a 1 y, por tanto, señalando una fuerte relación lineal directa entre ambas variables.

c) La respuesta la proporciona el coeficiente de determinación, obtenido directamente con la covarianza y las varianzas o el cuadrado del coeficiente de correlación:

$$R^2 = 0,9788$$

Así, el 97,88% de la variabilidad observada en la estatura viene explicada por el modelo de regresión.

57. La tabla siguiente muestra los datos del número atómico (X) y la densidad (Y, kg/m^3) de los metales alcalinotérreos (Grupo 2) de la tabla periódica.

X	4	12	20	38	56	88
Y	1848	1738	1550	2540	3594	5000

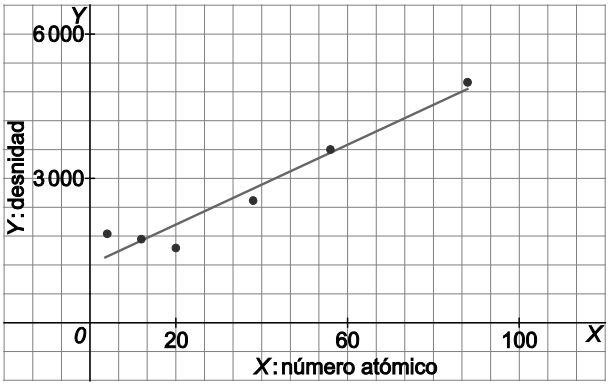
a) ¿Puede afirmarse que existe relación lineal entre el número atómico y la densidad? Razona la respuesta.

b) Escribe la recta de regresión lineal de la densidad en función del número atómico.

c) Calcula el coeficiente de determinación y valora la bondad del ajuste lineal.

a) Para visualizar si existe relación lineal entre las variables, se representan los datos en un gráfico de dispersión y se observa que pueden ajustarse, en conjunto, mediante una recta, si bien con las debidas precauciones, ya que por ejemplo, debe tenerse en cuenta que:

- Solo tenemos 6 observaciones.
- Las tres primeras parece que siguen una tendencia distinta a las otras tres.



b)

X	Y	x_i^2	y_i^2	x_iy_i
4	1848	16	3 415 104	7392
12	1738	144	3 020 644	20 856
20	1550	400	2 402 500	31 000
38	2540	1444	6 451 600	96 520
56	3594	3136	12 916 836	201 264
88	5000	7744	25 000 000	440 000
218	16 270	12 884	53 206 684	797 032

Se efectúan los cálculos con la ayuda de las columnas añadidas a la tabla de datos:

$$\overline{x} = \frac{218}{6} = 36,33 \quad \overline{y} = \frac{16\,270}{6} = 2711,67$$

$$s_x^2 = \frac{12\,884}{6} - 36,33^2 = 827,222$$

$$s_y^2 = \frac{53\,206\,684}{6} - 2711,64^2 = 1514\,644,556$$

$$s_{xy} = \frac{797\,032}{6} - 36,33 \cdot 2711,67 = 34\,314,778$$

La recta de regresión de la densidad Y en función del número atómico X es:

$$y = 2711,67 + \frac{34\,314,778}{827,222}(x - 36,33) \Rightarrow y = 41,482x + 1204,49$$

c) Con los resultados del apartado anterior, se calcula el coeficiente de determinación:

$$R^2 = \frac{34\,314,778^2}{827,222 \cdot 15\,514\,644,556} = 0,9398$$

Es decir, el 93,98% de la variabilidad observada en la densidad se debe al número atómico. Por lo tanto, el ajuste de la distribución (X, Y) mediante la recta de regresión obtenida en muy bueno, con las precauciones apuntadas en el apartado a).

58. La siguiente tabla refleja la distribución de una muestra de viviendas nuevas en una zona residencial, según el número de habitaciones (X) y su superficie (Y, en m^2).

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

a) Escribe la distribución de la superficie condicionada a que el número de habitaciones sea 3 y calcula la media y la varianza de esta distribución.

b) ¿Es independiente la superficie del número de habitaciones? Comenta el resultado obtenido.

a) La siguiente tabla recoge la distribución de frecuencias de la variable $Y|X=3$ junto con las columnas necesarias para hallar la media y la varianza.

$Y\|X=3$	f_{ij}	x_i	x_i^2	$x_i f_i$	$x_i^2 f_i$
[60, 70)	464	65	4225	30 160	1 960 400
[70, 80)	217	75	5625	16 275	1 220 625
[80, 90)	89	85	7225	7565	643 025
[90, 100)	26	95	9025	2470	234 650
	796		26 100	56 470	4 058 700

$$\text{Media de } Y|X=3\text{: } \frac{56\,470}{796} = 70{,}94$$

$$\text{Varianza de } Y|X=3\text{: } \frac{4\,058\,700}{796} - 70{,}94^2 = 66{,}386$$

b) Para ver si X e Y son independientes se construye la tabla de las distribuciones relativas y marginales.

h_{ij}	[60, 70)	[70, 80)	[80, 90)	[90, 100)	$h_{i.}$
2	0,0371 967 7	0,006 469	0,001 078 17	0,000 539 08	0,045 283 02
3	0,250 134 77	0,116 981 13	0,047 978 44	0,014 016 17	0,429 110 51
4	0,094 339 62	0,242 587 6	0,114 285 71	0,074 393 53	0,525 606 47
$h_{.j}$	0,381 671 16	0,366 037 74	0,163 342 32	0,088 948 79	1

Como la conjunta no es el producto de las marginales entonces X e Y no son independientes.

59. En un cultivo de laboratorio se ha medido el crecimiento (*Y*, en miles) de una colonia de bacterias, en función del número de días (*X*) que se mantiene el cultivo. Los datos se recogen en la siguiente tabla:

X	3	6	9	12	15	18
Y	115	147	239	356	579	864

a) Cuantifica la correlación lineal entre los días transcurridos y el número de bacterias presentes en el cultivo.

b) Escribe la recta de regresión del número de bacterias en función de los días transcurridos.

c) Define una nueva variable $Z = \ln Y$ y realiza ahora el ajuste de *Z* en función de *X* y valora la bondad del ajuste, comparándolo con el anterior.

a) Para calcular el coeficiente de correlación lineal, se necesitan las varianzas y la covarianza de las dos variables.

X	Y	x_i^2	y_i^2	x_iy_i
3	115	9	13 225	345
6	147	36	21 609	882
9	239	81	57 121	2151
12	356	144	126 736	4272
15	579	225	335 241	8685
18	864	324	746 496	15 552
63	2300	819	1 300 428	31 887

$$\overline{x} = \frac{63}{6} = 10{,}5 \qquad s_x^2 = \frac{819}{6} - 10{,}5^2 = 26{,}25$$

$$\overline{y} = \frac{2300}{6} = 383{,}33 \qquad s_y^2 = \frac{1300\,428}{6} - 383{,}33^2 = 69\,793{,}555\,6$$

$$s_{xy} = \frac{31887}{6} - 10{,}5 \cdot 383{,}33 = 1289{,}5$$

De manera que el coeficiente de correlación lineal entre el número de días y la cantidad de bacterias es:

$$r = \frac{1289{,}5}{\sqrt{26{,}25}\sqrt{69\,793{,}5556}} = 0{,}952\,69$$

Queda próximo a 1, por tanto, se tiene una relación lineal directa y fuerte.

b) Con los resultados del apartado a) se tiene que la recta de regresión de *Y* sobre *X* es:

$$y = 383{,}33 + \frac{1289{,}5}{26{,}25}(x - 10{,}5) \Rightarrow y = 49{,}123x - 132{,}47$$

c) Los nuevos datos son:

X	Y	$Z = \ln Y$	z_i^2	x_iz_i
3	115	4,744 932 13	22,514 380 9	14,234 796 4
6	147	4,990 432 59	24,904 417 4	29,942 595 5
9	239	5,476 463 55	29,991 653	49,288 172
12	356	5,874 930 73	34,514 811 1	70,499 168 8
15	579	6,361 302 48	40,466 169 2	95,419 537 2
18	864	6,761 572 77	45,718 866 3	121,708 31
63	2300	34,209 634 2	198,110 297 9	381,092 58

$$\overline{z} = \frac{34{,}21}{6} = 5{,}706 \quad s_z^2 = \frac{198{,}110}{6} - 5{,}7016^2 = 0{,}5101$$

$$s_{xz} = \frac{381{,}093}{6} - 10{,}5 \cdot 5{,}702 = 3{,}6487$$

La recta de regresión de *Z* sobre *X* es:

$$z = 5{,}702 + \frac{3{,}645}{26{,}25}(x - 10{,}5) \Rightarrow z = 0{,}139x + 5{,}563$$

El coeficiente de determinación ahora es: $R^2 = \dfrac{3{,}6487^2}{26{,}25 \cdot 0{,}5101} = 0{,}99424$

En el modelo anterior $R^2 = \dfrac{1289{,}5^2}{26{,}25 \cdot 69\,793{,}555\,6} = 0{,}908$ que es menor que el coeficiente de determinación que acabamos de calcular. Ahora bien, el error cuadrático medio para este modelo es menor $\text{ECM} = 0{,}5101(1 - 0{,}994\,24) = 0{,}002938$ que en el anterior modelo que era $\text{ECM} = 6448{,}40$. Por tanto, la transformación proporciona un mejor ajuste lineal.

ENTORNO MATEMÁTICO

Busco médico

1. Al padre de Sofía le han trasladado en su trabajo y toda su familia se ha mudado con él. Una de las cosas que han tenido que hacer al llegar a su nueva residencia es actualizar sus datos en el Sistema Nacional de Salud y solicitar que les asignen un nuevo médico de familia.

Al buscar por internet la información sobre los pasos que tenían que seguir para actualizar sus datos, Sofía se encontró con algunos datos estadísticos que llamaron su atención y que desconocía. Interesada, decidió investigar un poco y estas son algunas de las conclusiones que sacó:

- **La atención médica a la población difiere extraordinariamente de un país a otro, dependiendo de su grado de desarrollo, pero nosolo de éste sino de otras muchas circunstancias.**
- **La OMS (Organización mundial de la salud) estima que menos de 230 trabajadores de la salud (sólo médicos, enfermeras y comadronas) por cada 100 000 habitantes serían insuficientes para alcanzar la cobertura de las necesidades de atención primaria de salud.**
- **Muchos países africanos no llegan a tener un trabajador sanitario por cada 100 000 habitantes, mientras que en España, en el año 2012 se alcanzó la proporción de 498 médicos y 577 enfermeros.**

En Internet encontró los datos referentes a distintos países para analizarlos.

Analiza, al igual que Sofía, las tablas y responde a estas preguntas.

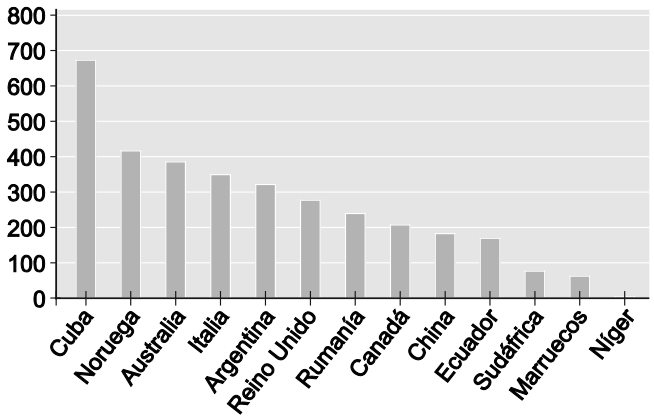
N.º	País	Médicos
2	Cuba	672
6	Rusia	431
8	Noruega	416
10	España	396
13	Australia	385
21	Alemania	369
26	Italia	349
30	Francia	338
35	Argentina	321
43	Egipto	283
46	Reino Unido	277
57	Estados Unidos	242
59	Rumania	239
60	Japón	214
62	Canadá	207
68	México	196
76	China	182
79	Brasil	176
83	Ecuador	169
109	Arabia Saudí	94
116	Sudáfrica	76
118	India	65
119	Marruecos	62
172	Senegal	6
190	Níger	2
193	Tanzania	1

a) Dibuja un diagrama de barras con los datos y calcula la media y la desviación típica.

b) Separa los países por continentes (considera América del Norte y del Sur por separado) y calcula la media para cada continente. Identifica los valores atípicos.

c) De los datos que has calculado en los apartados anteriores, ¿qué conclusiones puedes sacar?

a)



La media se obtiene sumando las frecuencias absolutas y dividiendo por el número de países (26):

Médicos por cada 1000 habitantes.

Y su desviación típica:

$$s_x^2 = \frac{672^2 + 431^2 + \ldots + 6^2 + 2^2 + 1^2}{26} - 237,23^2 = 24\,707,562 \Rightarrow s_x = \sqrt{24\,707,562} = 157,186$$

b)

Europa	M.
Rusia	431
Noruega	416
España	396
Alemania	369
Italia	349
Francia	338
Reino Unido	277
Rumanía	239

$\overline{x} = \frac{2815}{8} = 237,23$

$s_x^2 = 48\,212,92 \Rightarrow$

$\Rightarrow s_x = 219,57$

América del Norte	M.
Cuba	672
EEUU	242
Canadá	207
México	296

$\overline{x} = \frac{1417}{4} = 354,25$

$s_x^2 = 34\,660,19 \Rightarrow$

$\Rightarrow s_x = 186,172$

África	M.
Egipto	283
Sudáfrica	76
Marruecos	62
Senegal	6
Níger	2
Tanzania	1

$\overline{x} = \frac{430}{6} = 71,67$

$s_x^2 = 9821,74 \Rightarrow$

$\Rightarrow s_x = 99,104$

América del Sur	M.
Argentina	321
Brasil	176
Ecuador	169

$\overline{x} = \frac{666}{3} = 222$

$s_x^2 = 4908,67 \Rightarrow$

$\Rightarrow s_x = 70,62$

Asia	M.
Japón	214
China	182
Arabia Saudí	94
India	65

$\overline{x} = \frac{555}{4} = 138,75$

$s_x^2 = 3743,69 \Rightarrow$

$\Rightarrow s_x = 61,19$

Oceanía	M.
Australia	385

En cada tabla aparecen los países de la primera lista agrupados por continente y al lado de cada tabla el valor medio (número de medio de médicos por cada 100 000 habitantes) por continente y su desviación típica excepto para Oceanía ya que solo se dispone del dato de Australia.

Claramente, el número medio de médicos por cada 100 000 habitantes es muy bajo en África, a pesar de la cifra de Egipto, ya que países como Senegal, Níger o Tanzania presentan unas cifras ínfimas.

En América del Norte, es significativa la cifra de médicos por cada 100 000 habitantes de Cuba (672), muy superior a la de los demás países incluidos en esta lista.

c) De los datos se pueden extraer algunas conclusiones, entre otras:

- Las cifras de médicos por cada cien mil habitantes, varían extraordinariamente de unos países a otros.
- También es muy elevada la variación de unos continentes a otros.
- Países muy desarrollados, como Japón, EEUU o Canadá presentan cifras inferiores a países con menor grado de desarrollo, como es el caso de varios países europeos o de por ejemplo Argentina y Egipto.
- Las cifras de África, Asia y América del Sur son claramente inferiores a las de Europa y de A del Norte, aunque en este último caso, que el promedio sea tan elevado se debe exclusivamente al dato que aporta Cuba.

No ha llovido…,¿va a subir precio del pan?

2. A Héctor no le gusta el pan. Desde pequeño se comía el relleno del bocadillo en el recreo y se "deshacía" hábilmente del resto. Hoy se ha quedado pálido cuando ha oído en la radio que el precio en origen del trigo va a bajar hasta un 30 % respecto al año anterior, debido a la gran cosecha obtenida gracias a las lluvias primaverales. Ya se ve a sí mismo comiendo enormes bocadillos, sopas de ajo y todo tipo de horrores gastronómicos basados en el pan, que, a buen seguro, se le ocurrirán a la imaginativa mente de su madre.

Antes de rendirse, decide comprobar si la noticia es cierta. Así, se propone estudiar si hay relación entre las precipitaciones (cantidad de lluvia caída) en primavera y el precio que se paga a los agricultores por tonelada (unidad habitual, pero puedes usar cualquier otra) de cereal (trigo y cebada sobre todo). Los datos, los ha obtenido de la página del INE, en el enlace de Agricultura (proporcionadas por el Ministerio de Agricultura) y en Climatología.

Al igual que Héctor, y con ayuda de una hoja de cálculo, intenta ver si la relación es cierta o no. Para ello responde:

a) Escribe en una tabla las parejas de datos correspondientes a precipitaciones y precios (no olvides concretar las unidades) y realiza un análisis de regresión de los precios sobre las precipitaciones (al revés no tendría sentido).

b) ¿Se puede hacer el estudio por regiones? ¿Por cuencas hidrográficas? ¿Para toda España?

c) ¿Te sientes capaz de hacerlo para toda la península ibérica?

Es un problema abierto que se ha de considerar más como un trabajo pudiendo consultar los datos facilitados por el INE. Vamos a proponer un estudio para el apartado a).

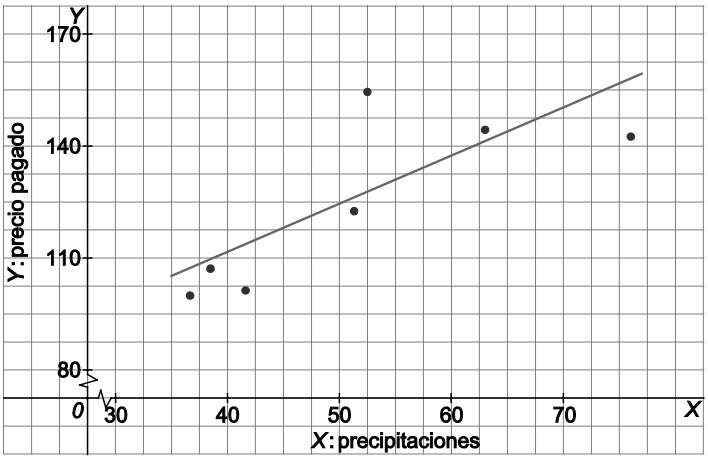
a) La tabla muestra las precipitaciones (X) en mm y el precio pagado a los agricultores por tonelada (Y) desde el año 2005, considerado como base (precio =100), hasta el 2011.

Se tienen en cuenta que:

- Los precios entre 2000 y 2004 se consideran con base en 2000 y, por ello no se han incluido aquí.
- Las precipitaciones son valores medios obtenidos a partir de los valores proporcionados en las tablas de INE para todas las regiones de España, de los meses de marzo, abril y mayo de cada año, desde 2005 hasta 2011.

Año	Y: precio pagado	X: precipitaciones	y_i^2	x_i^2	x_iy_i
2005	100	36,644	10 000,000	1342,773	3664,387
2006	101,21	41,621	10 243,464	1732,287	4212,437
2007	144,36	63,033	20 839,810	3973,201	9099,492
2008	142,54	76,029	20 317,652	5780,385	10 837,152
2009	107,18	38,514	11 487,552	1483,336	4127,942
2010	122,52	51,308	15 011,150	2632,474	6286,213
2011	154,51	52,482	23 873,340	2754,408	8109,064
	872,32	359,631	111 772,968	19 698,866	46 336,686

La nube de puntos junto con la recta de regresión ajustada, indican una tendencia creciente (contrariamente a lo que podría pensarse) y una relación lineal moderadamente aceptable.



Las medias, varianzas y covarianza de X e Y se obtienen a partir de los datos de la tabla:

$$\overline{x} = \frac{359,631}{7} = 51,376 \qquad \overline{y} = \frac{872,32}{7} = 124,617$$

$$s_x^2 = \frac{19\,698,866}{7} - 51,376^2 = 174,645$$

$$s_y^2 = \frac{111772,968}{7} - 124,617^2 = 438,135$$

$$s_{xy} = \frac{46\,336,686}{7} - (51,376)\cdot(124,617) = 217,214$$

Y, a partir de estos resultados se puede obtener la recta de regresión del precio pagado (Y) sobre las precipitaciones (X):

$$y = 124,617 + \frac{217,214}{174,645}(x - 51,376) \Rightarrow y = 1,244x + 60,719$$

Cuya bondad se mide a través de los coeficientes de determinación y de correlación:

$$R^2 = \frac{217,214^2}{174,645 \cdot 438,135} = 0,6166 \Rightarrow r = 0,7852$$

El valor del coeficiente de correlación (0,7852) confirma una relación lineal directa moderadamente alta entre las precipitaciones anuales en primavera y el precio pagado a los agricultores por el cereal, detectada en el diagrama de dispersión. Además, El 61,66% de la variabilidad observada en los precios viene explicada por la variación en las precipitaciones.

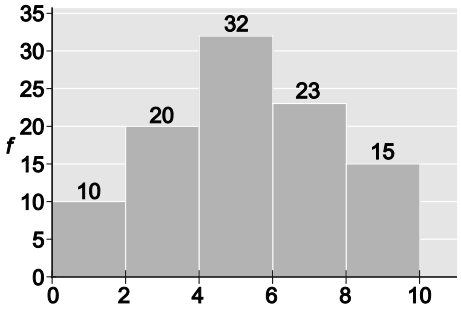
b) Se puede realizar un estudio parecido ya que se tienen datos por regiones, cuencas hidrográficas y para toda España. Para acceder a los datos se puede consultar la página del INE y los enlaces de smSaviadigital.com.

c) Del mismo modo se puede realizar un estudio similar a los anteriores para la península ibérica.

AUTOEVALUACIÓN

Comprueba qué has aprendido

1. El gráfico siguiente corresponde al histograma de una variable continua.



a) Escribe la tabla de frecuencias.

b) Calcula la media, $\overline{x}$, la moda y la mediana.

c) Halla la varianza, s^2, y la desviación típica, s.

a) Del diagrama de barras se obtiene directamente la tabla de frecuencias absolutas:

Clases	[0,2)	[2,4)	[4,6)	[6,8)	[8,10]
f_j	10	20	32	23	15

b) Para calcular la media, la mediana, así como la varianza y desviación típica del siguiente apartado, se amplía la tabla con las columnas necesarias:

Clases	f_j	x_j	$f_j x_j$	$f_j x_j^2$	F_j
[0,2)	10	1	10	10	10
[2,4)	20	3	60	180	30
[4,6)	32	5	160	800	62
[6,8)	23	7	161	1127	85
[8,10]	15	9	135	1215	100
	100		526	3332	

La media es: $\overline{x} = \dfrac{526}{100} = 5{,}26$

La clase modal es [4, 6), porque tiene la mayor frecuencia absoluta.

La mitad de las observaciones es 50 y hasta el intervalo [4, 6) se acumulan 30 observaciones, para las 20 restantes se procede por interpolación lineal: si el intervalo [4, 6), de longitud 2, contiene 32 observaciones, a 20 observaciones le corresponde una longitud:

$$L = \frac{20(6-4)}{32} = 1{,}25$$

Luego la mediana es $M = 4 + 1{,}25 = 5{,}25$.

c) De la tabla del apartado b) se obtiene la varianza y de esta la desviación típica

$$s_x^2 = \frac{3332}{100} - 5{,}26^2 = 5{,}6524 \Rightarrow s_x = 2{,}3775$$

2. 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	1

a) Representa gráficamente la distribución.

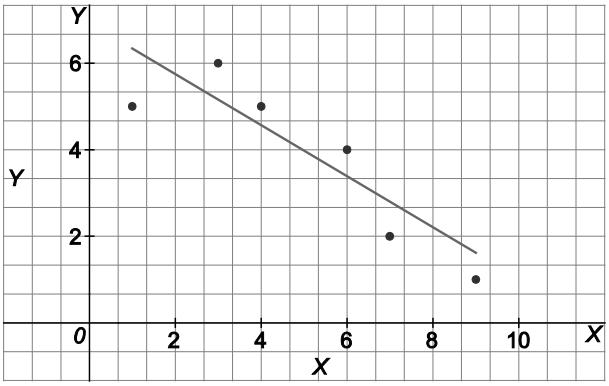
b) Calcula el coeficiente de correlación y evalúa el ajuste lineal a la distribución.

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

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

e) Si $x = 5$, ¿cuál es el valor esperado de *y*? ¿y si $x = 15$? Comenta la fiabilidad de ambas predicciones.

a) Se representan la nube de puntos de la variable estadística (*X*,*Y*), junto con la recta de regresión de *Y* sobre *X* obtenida en el apartado c), y se puede observar una relación lineal inversa fuerte entre ambas variables.



b) El cálculo del coeficiente de correlación precisa añadir a la tabla columnas con los cuadrados de las variables y el producto de las mismas.

X	*Y*	x_i^2	y_i^2	x_iy_i
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	1	81	1	9
30	23	192	107	90

Las medias, las varianzas y la covarianza de *X* e *Y* son:

$$\overline{x} = \frac{30}{6} = 5 \qquad \overline{y} = \frac{23}{6} = 3,83$$

$$s_x^2 = \frac{192}{6} - 5^2 = 7$$

$$s_y^2 = \frac{107}{6} - 3,83^2 = 3,1389$$

$$s_{xy} = \frac{90}{6} - 5 \cdot 3,83 = -4,1667$$

De modo que el coeficiente de correlación lineal es: $r = \dfrac{-4,1667}{\sqrt{7 \cdot 3,1389}} = -0,8889$

que confirma la intuición señalada en el apartado a). El coeficiente de correlación lineal toma un valor relativamente próximo a -1 y, por tanto, el ajuste de la distribución mediante una recta es aceptable.

c) La recta de regresión *Y* sobre *X*, se obtienen con los datos del apartado b):

$$y = 3,83 + \frac{(-4,1667)}{7}(x - 5) \Rightarrow y = -0,595x + 6,81$$

d) Este porcentaje se obtiene mediante el coeficiente de determinación: $R^2 = \dfrac{(-4,1667)^2}{7 \cdot 3,1389} = 0,7901$

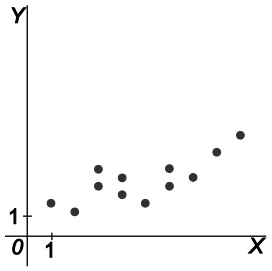
Esto es, el 79,01% de la variabilidad observada en *Y* viene explicada por el modelo de regresión (por la variable *X*).

e) Si $x = 5$, entonces $y = -0,595 \cdot 5 + 6,81 = 3,835$, predicción fiable por encontrase el valor $x = 5$, en el rango de valores de *X* y, además, muy próximo a la media de la variable.

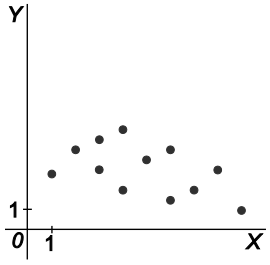
En cambio, no se puede hacer predicción para $x = 15$, por no estar dentro del rango de valores de la variable regresora.

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

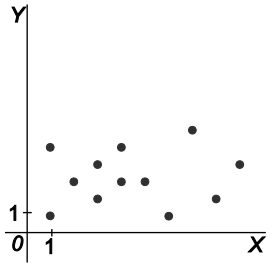
I



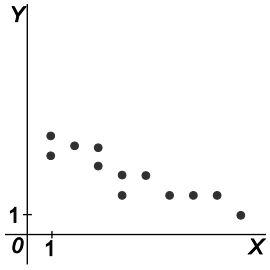
III



II



IV



a) r=0,102 **b)** r=–0,903 **c)** r= 0,776 **d)** r=–0,501

I→ r= 0,776 II→ r= 0,102 III→ r=–0,501 IV→ r=–0,903

4. Si las puntuaciones otorgadas a 7 alumnos en un examen de matemáticas son 4, 7, 6, 8, 3, 9 y 5.

a) Calcula la media y la varianza de las calificaciones.

b) Si se multiplican por 2, ¿cuáles son ahora la media y la varianza?

c) El coeficiente de correlación entre las notas de matemáticas y las de química para estos 7 alumnos es r= 0,78. Si tanto las calificaciones de matemáticas como las de química se multiplican por 2 ¿Cuál es ahora el valor del coeficiente de correlación?

a) La media es: $\overline{x} = \dfrac{4+7+6+8+3+9+5}{7} = 6$

La varianza es: $s_x^2 = \dfrac{4^2+7^2+6^2+8^2+3^2+9^2+5^2}{7} - 6^2 = 4$

b) Si se multiplican los valores de la variable por 2, la media queda multiplicada por 2 ya que, en general, si $y_j = 2x_j$:

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} 2x_i = \frac{2}{n}\sum_{i=1}^{n} x_i = 2\overline{x}$$

Mientras que la varianza queda multiplicada por 2^2= 4, puesto que en general:

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2 = \frac{1}{n}\sum_{i=1}^{n} (2x_i)^2 - (2\overline{x})^2 = 4\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2\right) = 4s_x^2$$

Por tanto, en este caso, la media y la varianza de la variable $Y = 2X$ son:

$$\overline{y} = 2\overline{x} = 12 \qquad s_y^2 = 4s_x^2 = 16$$

c) El coeficiente de correlación no cambia, puesto que

$$r_{(2x)(2z)} = \frac{s_{(2x)(2z)}}{\sqrt{s_{2x}^2 s_{2z}^2}} = \frac{4s_{xz}}{\sqrt{4s_x^2 4s_z^2}} = \frac{s_{xz}}{\sqrt{s_x^2 s_z^2}} = r_{xz}$$

Relaciona y contesta

Elige la única respuesta correcta en cada caso

1. La variable estadística *X* toma los valores 2, 3, 4, 5 y 6 con frecuencias respectivas 4, 5, *f*, 3 y 1. Si se sabe que la media aritmética de *X* es 3,6, el valor de *f* y la mediana de *X* son:

A. $f = 8, M = 3$ **B.** $f = 7, M = 4$ **C.** $f = 6, M = 4$ **D.** $f = 7, M = 5$

Solución:B

2. En la regresión lineal 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 ese dato, no hay relación lineal entre *X* e *Y*.

Solución: C

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. $\bar{y} = 2{,}5, r = -0{,}9$ **B.** $\bar{y} = 2{,}5, r = 0{,}9$ **C.** $\bar{y} = 1{,}1, r = -0{,}9$ **D.** $\bar{y} = 1{,}1, r = 0{,}9$

Solución: C

4. De la distribución conjunta de dos variables estadísticas *X* e *Y* se sabe que $s_x = 2$, $s_{xy} = -2$, $\bar{x} = 8$, $\bar{y} = 10$. En este caso:

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

Solución: B

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.** $R^2_{zt} = 9R^2_{xy}$

Solución: B y C

6. Con el modelo de regresiónlineal 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 su media.

C. Solo para valores pequeños de *X*

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

Soluciones: B y D

Elige la relación correcta entre las dos afirmaciones

7. 1. El coeficiente de correlación es $r = -0{,}7$ **2. La recta de regresión es $y = 50 - 0{,}7x$**

A. $1 \Rightarrow 2$ **B.** $2 \Rightarrow 1$ **C.** $1 \Leftrightarrow 2$ **D.** $1 \not\Leftrightarrow 2$

Solución: D