1 Variable estadística bidimensional

Una **variable estadística bidimensional** resulta al estudiar dos características diferentes de los individuos de una población, y está formada por las dos variables estadísticas unidimensionales. La variable bidimensional (X, Y) queda determinada por los pares de datos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Ejemplo

1 Realizamos un estudio sobre el uso del autobús en una ciudad, y se pregunta a los usuarios por el número de autobuses que utilizan al día y el tiempo aproximado, en minutos, que tardan en llegar a su destino. ¿Qué variables estadísticas se van a estudiar?

$X \rightarrow$ Número de autobuses $\quad Y \rightarrow$ Tiempo necesario para llegar a su destino

Para cada persona encuestada tenemos un par de datos: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Este conjunto de pares de datos forma una variable estadística bidimensional.

1.1. Frecuencias

- **Frecuencia absoluta conjunta**, f_{ij}, es el número de veces que aparece cada par de datos (x_i, y_j) de la variable bidimensional.
- **Frecuencia relativa conjunta**, h_{ij}, es el cociente de la frecuencia absoluta conjunta de cada par (x_i, y_j) y el número total de pares de datos.
- **Frecuencia absoluta marginal** es el número de veces que aparece cada dato al estudiar por separado las dos variables unidimensionales.
- **Frecuencia relativa marginal** es el cociente de la frecuencia absoluta marginal de cada dato y el número total de datos.

Date cuenta

Las frecuencias marginales de una variable bidimensional son las frecuencias de las variables unidimensionales correspondientes.

Ejemplo

2 Diez personas encuestadas sobre el número de autobuses que utilizan y el tiempo que tardan en llegar a su destino, responden:

(1, 20) (1, 30) (3, 10) (2, 30) (2, 40) (1, 30) (2, 10) (3, 30) (3, 40) (1, 30)

a) ¿Cuál es la frecuencia absoluta conjunta del par (1, 30)? ¿Qué significa?

b) ¿Y las frecuencias absolutas marginales de los datos 1 y 30?

a) El par (1, 30) aparece 3 veces, así su frecuencia absoluta conjunta es 3. Significa que hay 3 personas que utilizan 1 autobús y tardan 30 minutos en llegar a su destino.

b) Al considerar por separado las dos variables, comprobamos que hay 4 personas que utilizan 1 autobús y 5 que tardan 30 minutos en llegar a su destino. Por tanto, 4 y 5 son las frecuencias absolutas marginales correspondientes.

ACTIVIDADES

1 Considera estas variables bidimensionales, y escribe las variables unidimensionales correspondientes y tres pares de datos que las determinan.

a) Edad y sexo de los asistentes a un concierto.

b) Tamaño de un archivo informático y tiempo que se tarda en copiarlo.

2 En un estudio estadístico se han obtenido estos datos.

(1, 4) (1, 8) (2, 8) (3, 6) (3, 6)
(2, 6) (2, 4) (1, 8) (1, 6) (2, 8)

a) ¿Cuáles son las frecuencias absolutas conjuntas? ¿Y las marginales?

b) Determina las frecuencias relativas.

1.2. Tablas de doble entrada

Para organizar los datos en las variables bidimensionales, los podemos ordenar en una **tabla de doble entrada**.

En ellas establecemos dos entradas, una para los datos de cada variable, y colocamos las frecuencias absolutas de cada par de datos en las casillas centrales.

Frecuencia absoluta conjunta de (x_2, y_2)

Y \ X	x_1	x_2	...	x_n	Total
y_1	f_{11}	f_{21}	...	f_{n1}	$\sum_{i=1}^{n} f_{i1}$
y_2	f_{12}	f_{22}	...	f_{n2}	$\sum_{i=1}^{n} f_{i2}$
...	...	...	...	...	...
y_m	f_{1m}	f_{2m}	...	f_{nm}	$\sum_{i=1}^{n} f_{im}$
Total	$\sum_{i=1}^{m} f_{1i}$	$\sum_{i=1}^{m} f_{2i}$	...	$\sum_{i=1}^{m} f_{ni}$	

Frecuencia absoluta marginal de y_2

Frecuencia absoluta marginal de x_2

Se escribe así

Para expresar las frecuencias absolutas marginales utilizamos sumatorios. Así, la frecuencia absoluta marginal de y_1 es:

$$\sum_{i=1}^{n} f_{i1} = f_{11} + f_{21} + \ldots + f_{n1}$$

Ejemplo

3 Esta tabla de doble entrada muestra los resultados al estudiar la variable bidimensional (X, Y) siendo X, el número de autobuses que utilizan al día las personas encuestadas e Y, el tiempo que emplean en llegar a su destino.

Y \ X	1	2	3	Total
10	12	4	1	17
20	8	16	3	27
30	14	3	5	22
40	7	5	4	16
50	4	2	2	8
Total	45	30	15	90

a) ¿Cuántos usuarios utilizan 3 autobuses al día?

b) ¿Cuántos emplean 40 minutos en llegar a su destino?

c) ¿Cuántos usuarios utilizan un autobús y tardan 30 minutos en llegar?

a) Hay 15 usuarios que utilizan 3 autobuses al día.

b) Emplean 40 minutos en llegar a su destino 16 personas.

c) Hay 14 personas que utilizan un solo autobús y tardan 30 minutos.

ACTIVIDADES

3 Observa esta tabla de doble entrada.

Y \ X	5	10	15	Total
100	3	2	5	10
200	1	8	6	15
300	2	1	2	5
Total	6	11	13	30

a) ¿Cuál es la frecuencia absoluta conjunta del par (10, 200)? ¿Y la relativa conjunta de este par?

b) Indica las frecuencias marginales de 5 y 300.

4 Ordena estos datos en una tabla de doble entrada.

X	Y
0	18
0	12
1	7
2	8

X	Y
1	14
2	23
1	17
2	8

a) ¿Hay pares de datos que tengan la misma frecuencia absoluta conjunta?

b) Indica las frecuencias marginales de la variable X.

1.3. Tablas de frecuencias marginales

Las **tablas de frecuencias marginales** se obtienen al estudiar por separado cada una de las variables que forman la variable bidimensional.

Ejemplo

4 Determina las tablas de frecuencias marginales de esta tabla de doble entrada.

Y \ X	1	2	3	Total
10	14	4	1	19
20	10	16	3	29
30	14	4	6	24
40	7	6	5	18
Total	45	30	15	90

Consideramos los valores de las variables X e Y y sus frecuencias marginales.

Tabla de frecuencias marginales de X

X	Frecuencias
1	45
2	30
3	15
Total	90

Tabla de frecuencias marginales de Y

Y	Frecuencias
10	19
20	29
30	24
40	18
Total	90

1.4. Covarianza

La **covarianza**, σ_{XY}, de una variable bidimensional (X, Y) es una medida estadística y se calcula utilizando las siguientes expresiones.

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

No olvides

La covarianza se puede calcular con cualquiera de las dos fórmulas de la definición. Sin embargo, la fórmula más sencilla es:

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

Hazlo así

CÓMO CALCULAMOS LA COVARIANZA

Calcula la covarianza de estos datos.

Y \ X	1	2	3	Total
10	14	4	1	19
20	10	16	3	29
30	14	4	6	24
40	7	6	5	18
Total	45	30	15	90

PRIMERO. A partir de las tablas marginales, determinamos la media de cada una de las variables.

$$\bar{x} = \frac{45 \cdot 1 + 30 \cdot 2 + 15 \cdot 3}{90} = 1,\widehat{6}$$

$$\bar{y} = \frac{19 \cdot 10 + 29 \cdot 20 + 24 \cdot 30 + 18 \cdot 40}{90} = 24,\widehat{5}$$

SEGUNDO. Hallamos la covarianza. $\sigma_{XY} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n} x_i \cdot y_j \cdot f_{ij}}{N} - \bar{x} \cdot \bar{y} = \frac{3.830}{90} - 1,\widehat{6} \cdot 24,\widehat{5} = 1,63$

ACTIVIDADES

5 Construye la tabla de doble entrada y las tablas marginales correspondientes.

X	16	17	18	16	14	17	14	13	14	15
Y	5	4	6	6	8	3	5	4	8	8

6 Determina la covarianza para los datos que aparecen en la siguiente tabla.

X	8	10	11	9	13	12	9	14
Y	20	18	16	22	10	10	21	9

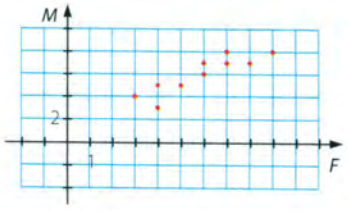
2 Diagrama de dispersión

Se llama **diagrama de dispersión** o **nube de puntos** al gráfico que se obtiene al representar, en unos ejes de coordenadas, los N pares de datos que toma la variable bidimensional.

Ejemplo

5 Esta tabla refleja las notas de una evaluación de un grupo de alumnos en Física y Matemáticas. Dibuja su diagrama de dispersión.

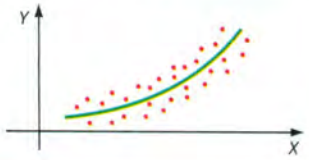
Física (F)	6	7	9	6	4	7	3	8	4	5
Matemáticas (M)	7	8	8	6	3	7	4	7	5	5



Representamos por un punto del plano cada dato de la variable, siendo la primera coordenada la *nota obtenida en Física*, y la segunda, la *nota obtenida en Matemáticas*.

No olvides

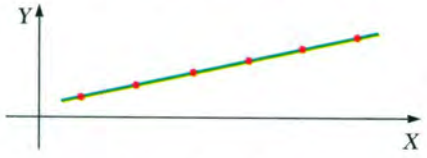
La nube de puntos se puede ajustar también a una función que no sea una recta. A este tipo de dependencia se le denomina **dependencia funcional**.

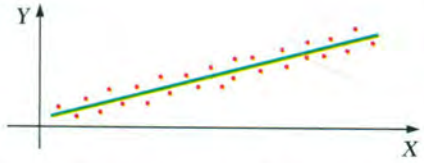


Dependencia lineal

Si los puntos siguen, aunque sea aproximadamente, una configuración rectilínea, diremos que hay **dependencia** entre ellos. Esta dependencia puede ser:

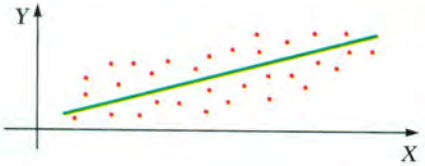
Dependencia lineal exacta, si los datos de la variable se ajustan completamente a una recta.





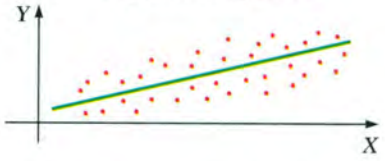
Dependencia lineal fuerte, si los datos no se ajustan a una recta, pero se encuentran muy próximos. En este caso, la nube de puntos es estrecha.

Dependencia lineal débil, si los puntos se encuentran alejados de la recta que se puede trazar entre ellos. La nube de puntos es más ancha.

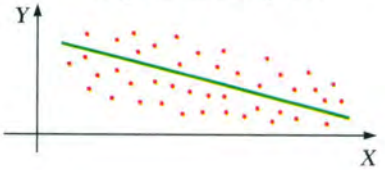


La dependencia puede ser **positiva** o **negativa**, según sea el signo de la pendiente de la recta a la que se aproximan los valores de la variable.

Dependencia positiva



Dependencia negativa

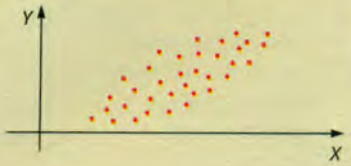


ACTIVIDADES

7 Representa la nube de puntos correspondiente a la siguiente variable estadística bidimensional.

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

8 Indica la dependencia entre estas variables.



3 Correlación

El **coeficiente de correlación**, r, es una medida de la variable (X, Y) que determina el grado de dependencia lineal entre las variables X e Y.

Se calcula utilizando esta expresión: $r = \dfrac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$

siendo σ_{XY} la covarianza de la variable (X, Y), σ_X la desviación típica de X y σ_Y la desviación típica de Y.

El coeficiente de correlación indica la aproximación de los valores de la variable a una línea recta. Su valor está comprendido entre -1 y 1.

- Si el coeficiente de correlación, r, es positivo, la dependencia será positiva, y si es negativo, la dependencia será negativa.
- Si $r = 1$ o $r = -1$, la dependencia es exacta. Será positiva cuando $r = 1$, y será negativa cuando $r = -1$.
- Cuanto más se acerque r a 1 o -1, la dependencia es más fuerte.
- Cuanto más se acerque r a 0, la dependencia es más débil.

Date cuenta

Como σ_X y σ_Y son siempre positivos, el signo del coeficiente de correlación viene determinado por el signo de σ_{XY}.

- Si $\sigma_{XY} < 0$, la dependencia es negativa.
- Si $\sigma_{XY} > 0$, la dependencia es positiva.

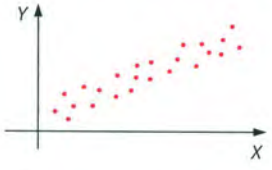
Ejemplo

6 **Describe el grado de correlación que existe entre las variables representadas y el posible valor de su coeficiente de correlación.**

La nube de puntos coincide con una recta cuya pendiente es negativa. La dependencia lineal es exacta y negativa.

El coeficiente de correlación es -1.

La nube de puntos se aproxima poco a una recta, y a medida que la variable X crece, la variable Y decrece. La dependencia lineal es débil y negativa. El coeficiente de correlación tomará un valor cercano a 0.



La nube de puntos se aproxima bastante a una recta con pendiente positiva. La correlación es fuerte y positiva. El coeficiente de correlación estará cercano a 1.

La nube de puntos se aproxima poco a una recta, y a medida que la variable X crece, la variable Y también crece. La correlación es débil y positiva. El coeficiente de correlación tomará un valor cercano a 0.

ACTIVIDADES

9 Describe el grado de correlación entre las dos variables representadas.

10 Si el signo de la covarianza entre dos variables es negativa, ¿qué podemos decir del signo del coeficiente de correlación?

¿Y si la covarianza es positiva?

Hazlo así

CÓMO CALCULAMOS E INTERPRETAMOS EL COEFICIENTE DE CORRELACIÓN

Se quiere determinar si existe algún tipo de relación entre la altura de un grupo de personas y su peso o el número de libros que lee anualmente. Para ello se ha hecho una encuesta y se han reflejado los datos obtenidos en esta tabla.

Altura (cm)	176	182	167	172	169	191	177	161	173	168
Peso (kg)	68	76	61	65	70	79	69	56	61	64
N.º de libros leídos	16	9	12	11	14	14	8	5	12	7

A la vista de estos datos, ¿crees que hay alguna relación entre la altura y el peso de estos individuos? ¿Y entre la altura y el número de libros que leen anualmente?

PRIMERO. Calculamos las desviaciones típicas de cada una de las variables.

$X \rightarrow$ Altura

$\overline{x} = 173{,}6$ cm $\qquad \sigma_X = 8{,}05$

$Y \rightarrow$ Peso

$\overline{y} = 66{,}9$ kg $\qquad \sigma_Y = 6{,}67$

$Z \rightarrow$ Libros leídos

$\overline{z} = 10{,}8$ libros $\qquad \sigma_Z = 3{,}31$

SEGUNDO. Calculamos la covarianza.

$$\sigma_{XY} = \frac{\sum_{i=1}^{N} x_i \cdot y_i}{N} - \overline{x} \cdot \overline{y} = 48{,}16$$

$$\sigma_{XZ} = \frac{\sum_{i=1}^{N} x_i \cdot z_i}{N} - \overline{x} \cdot \overline{z} = 11{,}42$$

TERCERO. Determinamos el coeficiente de correlación.

$$r_{XY} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} = 0{,}897 \qquad r_{XZ} = \frac{\sigma_{XZ}}{\sigma_X \cdot \sigma_Z} = 0{,}428$$

CUARTO. Interpretamos el valor obtenido para el coeficiente de correlación.

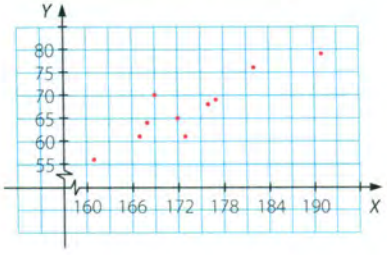
Para comparar las variables X e Y tenemos que interpretar $r_{XY} = 0{,}897$.

- Por ser próximo a 1, existe una dependencia lineal fuerte entre las variables, es decir, la altura y el peso en una persona de ese grupo están muy relacionados.
- Por ser positivo, la dependencia es positiva, es decir, a medida que aumenta la altura de una persona de ese grupo, aumenta su peso.

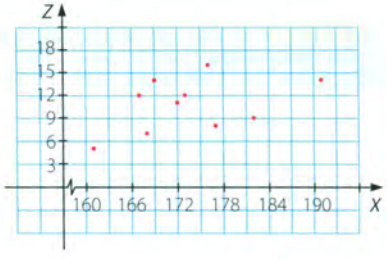
Para comparar las variables X e Z tenemos que interpretar $r_{XZ} = 0{,}428$.

- Por estar alejado de 1, existe una dependencia lineal débil entre las variables, es decir, la altura y el número de libros que lee anualmente una persona tienen poca relación.
- Aunque la dependencia es positiva, al ser débil, no se pueden extraer conclusiones de su comportamiento.

Dependencia lineal fuerte positiva



Dependencia lineal débil positiva



ACTIVIDADES

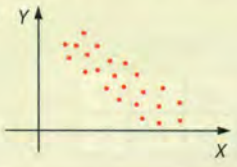
11 Representa el diagrama de dispersión y halla el coeficiente de correlación de esta variable.

X	39	43	40	40	42	41	42	38	39	44
Y	167	184	177	168	185	173	180	164	170	194

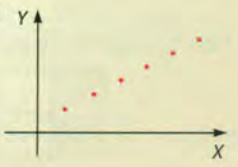
¿Qué relación puedes describir entre ellos?

12 Razona qué valor tomará el coeficiente de correlación.

a)

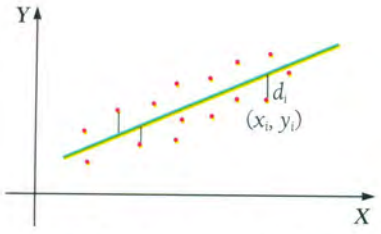


b)



4 Rectas de regresión

4.1. Recta de regresión de *Y* sobre *X*



d_i es la distancia entre la ordenada y la recta.

De las rectas posibles que podemos ajustar al diagrama de dispersión elegimos la que hace mínima la suma de las distancias entre las ordenadas de cada punto y la recta. A esta recta se le llama **recta de regresión de *Y* sobre *X***.

La recta de regresión de *Y* sobre *X* es una recta que se ajusta a los datos de una variable bidimensional. Su ecuación es:

$$y - \overline{y} = \frac{\sigma_{XY}}{\sigma_X^2}(x - \overline{x})$$

Hazlo así

CÓMO DETERMINAMOS Y REPRESENTAMOS LA RECTA DE REGRESIÓN

En una fábrica se ha medido la concentración, en gramos por litro, de uno de los componentes de una pintura y el tiempo que tarda en secarse.

Concentración (g/ℓ)	5	10	20	30
Tiempo (min)	16	17	18	19

Halla la recta de regresión y represéntala con el diagrama de dispersión.

PRIMERO. Construimos una tabla de frecuencias con las columnas necesarias para calcular las medidas estadísticas.

	x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
	5	16	25	256	80
	10	17	100	289	170
	20	18	400	324	360
	30	19	900	361	570
Total	**65**	**70**	**1.425**	**1.230**	**1.180**

SEGUNDO. Calculamos la media de cada variable, la varianza de *X* y la covarianza.

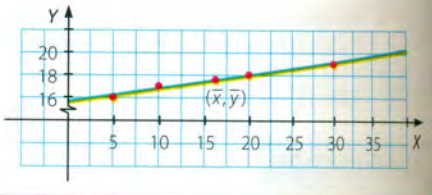
$$\overline{x} = \frac{65}{10} = 16{,}25 \qquad \overline{y} = \frac{70}{4} = 17{,}5 \qquad \sigma_X^2 = \frac{1.425}{4} - 264{,}06 = 92{,}19$$

$$\sigma_{XY} = \frac{1.180}{4} - 16{,}25 \cdot 17{,}5 = 10{,}63$$

TERCERO. Determinamos la recta de regresión de *Y* sobre *X* a partir de su ecuación.

$$y - \overline{y} = \frac{\sigma_{XY}}{\sigma_X^2}(x - \overline{x}) \rightarrow y - 17{,}5 = \frac{10{,}63}{92{,}19}(x - 16{,}25) \rightarrow y = 0{,}12x + 15{,}63$$

CUARTO. Representamos el diagrama de dispersión y trazamos la recta de regresión que hemos obtenido.



No olvides

La recta de regresión siempre pasa por el punto $(\overline{x}, \overline{y})$.
A este punto se le llama **centro de gravedad** de la nube de puntos.

ACTIVIDADES

13 Halla la recta de regresión de *Y* sobre *X*.

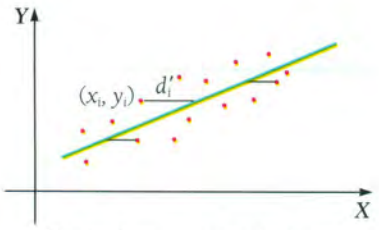
X	2	5	6	8	9
Y	4	13	16	22	25

14 Determina la recta de regresión correspondiente.

X	39	40	40	42	43	38	39	44	42	40
Y	167	168	180	164	177	154	185	195	183	172

4.2. Recta de regresión de *X* sobre *Y*

Si para ajustar la recta a los puntos del diagrama de dispersión hacemos mínima la suma de las distancias entre las abscisas de cada punto y de la recta, obtenemos la **recta de regresión de *X* sobre *Y***.



d'_i es la distancia entre la abscisa y la recta.

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

$$x - \overline{x} = \frac{\sigma_{XY}}{\sigma_Y^2}(y - \overline{y})$$

Ejemplo

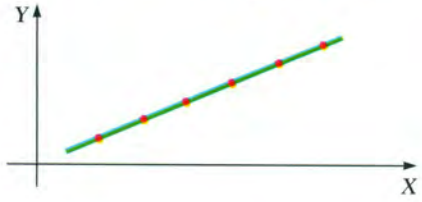
7 Determina la recta de regresión de *X* sobre *Y*, sabiendo que $\overline{x} = 16{,}25$; $\overline{y} = 17{,}5$; $\sigma_X^2 = 92{,}19$; $\sigma_Y^2 = 1{,}25$ y $\sigma_{XY} = 10{,}63$.

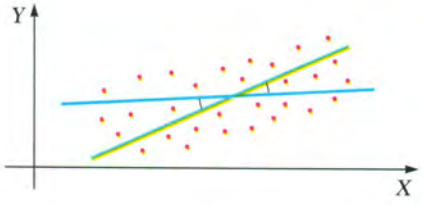
Como conocemos todas las medidas estadísticas que necesitamos, determinamos la recta de regresión de *X* sobre *Y* a partir de su ecuación.

$$x - \overline{x} = \frac{\sigma_{XY}}{\sigma_Y^2}(y - \overline{y}) \rightarrow x - 16{,}25 = \frac{10{,}63}{1{,}25}(y - 17{,}5) \rightarrow x = 8{,}5y - 132{,}57$$

7.3. Posiciones relativas de las dos rectas de regresión

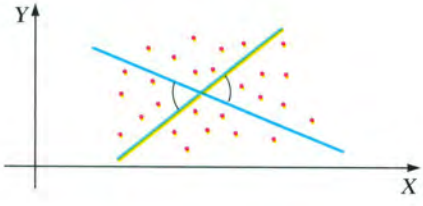
- Si $r = 1$ o $r = -1$, es decir, si la dependencia entre las dos variables es lineal exacta, las dos rectas son coincidentes.





- En general, cuanto más se acerque r a 1 o -1, es decir, cuanto más fuerte sea la dependencia, menor será el ángulo que forman ambas rectas.

- Si la dependencia es débil, es decir, si r se acerca a 0, las dos rectas forman un ángulo que se aproximará a 90°. Cuanto más se acerque r a 0, más se acercará el ángulo a 90°.



No olvides

Las dos rectas de regresión son siempre secantes y se cortan en el punto $(\overline{x}, \overline{y})$.

ACTIVIDADES

15 Determina las dos rectas de regresión, e indica la relación que hay entre las variables.

a)

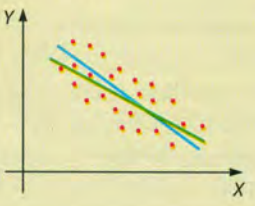
X	10	10	13	15	12
Y	6	5	2	3	5

b)

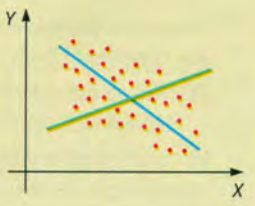
X	8	10	11	12	16	13	12	17	13	13
Y	15	10	15	10	20	15	10	25	10	15

16 Razona cuál es el grado de dependencia entre las variables en cada caso.

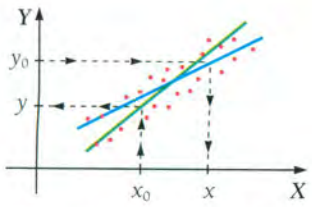
a)



b)



5 Estimación de resultados



Conocido x_0 se puede hallar y; y a partir de y_0 es posible calcular x.

Las rectas de regresión nos permiten obtener, de forma aproximada, los valores de una de las variables, conocidos los valores de la otra. Estos valores se llaman **estimaciones** de la variable.

Para hacer estimaciones de resultados debemos tener en cuenta que:

- Si queremos estimar y a partir de un valor de x, utilizaremos la recta de regresión de Y sobre X. En caso contrario, para estimar x a partir de un valor de y, utilizaremos la recta de regresión de X sobre Y.
- La estimación será más fiable si el coeficiente de correlación toma valores próximos a 1 o -1.
- Si el coeficiente de correlación es un valor próximo a 0, las estimaciones carecen de validez.
- Las estimaciones deben hacerse dentro del intervalo de valores obtenidos en el estudio estadístico, o muy próximos a él.

Hazlo así

CÓMO ESTIMAMOS VALORES UTILIZANDO LAS RECTAS DE REGRESIÓN

Esta tabla muestra la concentración de uno de los componentes de una pintura y el tiempo que tarda en secarse.

Concentración (g/ℓ)	5	10	20	30
Tiempo (min)	16	17	18	19

a) Haz una estimación del tiempo que tardaría en secarse la pintura si la concentración del componente es de 16 g/ℓ.

b) Si ha tardado en secarse 20 minutos, ¿cuál crees que será la concentración?

PRIMERO. Definimos las variables.

$X \rightarrow$ Concentración del componente $\qquad Y \rightarrow$ Tiempo que tarda en secarse

SEGUNDO. Calculamos las rectas de regresión y sustituimos el valor conocido en ella.

- Dado x, para estimar y, calculamos la recta de regresión de Y sobre X.

  a) $y = 0{,}12x + 15{,}63 \xrightarrow{x = 16} y = 0{,}12 \cdot 16 + 15{,}63 \rightarrow y = 17{,}55$

- Dado y, para estimar x, calculamos la recta de regresión de X sobre Y.

  b) $x = 8{,}5y - 132{,}57 \xrightarrow{y = 20} x = 8{,}5 \cdot 20 - 132{,}57 \rightarrow x = 37{,}43$

TERCERO. Interpretamos la solución.

a) Si la concentración del componente de la pintura es de 16 g/ℓ, la pintura tardará en secarse alrededor de 17 minutos y medio.

b) Si la pintura tarda en secarse 20 minutos la concentración será de 37 g/ℓ.

No olvides

- Las estimaciones siempre se realizan de forma aproximada, pues su grado de fiabilidad depende de la dependencia entre las variables.
- Las estimaciones se llaman **interpolaciones** cuando el valor estimado pertenece al intervalo de los valores obtenidos en el estudio estadístico. Si el valor no pertenece a ese intervalo, se denomina **extrapolación**.

ACTIVIDADES

17 En un estudio sobre los ingresos mensuales, X, y la superficie de las viviendas, Y, resulta: $y = 0{,}02x + 47{,}96$.

a) Halla la estimación de la superficie de la vivienda de una familia cuyos ingresos mensuales son de 3.200 €.

b) Si una familia vive en una casa de 90 m^2, ¿cuáles serán sus ingresos mensuales?

18 En un estudio estadístico, el coeficiente de correlación entre dos variables X e Y es $-0{,}8$. Se sabe que $\overline{x} = 20$; $\sigma_x = 4$; $\overline{y} = 8$ y $\sigma_y = 1$.

a) Determina las dos rectas de regresión, represéntalas y analiza la correlación que existe entre las variables.

b) Si $x = 30$, ¿cuál es la estimación de y?

6 Estadística con calculadora

Algunas calculadoras científicas nos permiten obtener los resultados que aparecen en Estadística bidimensional.

Hazlo así

CÓMO TRABAJAMOS LA ESTADÍSTICA BIDIMENSIONAL CON LA CALCULADORA

En una ciudad se han registrado estas temperaturas y precipitaciones durante los seis primeros meses del año.

Temperatura (°C)	11	14,3	14,7	15,7	19,6	23,3
Precipitaciones (mm)	46,9	0	13,7	113,4	44,3	1,2

Estudia la correlación que existe entre las dos variables. A partir de datos conocidos, ¿tiene sentido hacer estimaciones?

PRIMERO. Establecemos el Modo de Regresión Lineal en la calculadora; aunque el más frecuente es el Modo LR, esto puede variar según el modelo de calculadora que utilicemos.

MODE (LR)

SEGUNDO. Antes de comenzar a realizar cálculos nos aseguramos de que no hay datos almacenados en la memoria.

SHIFT (AC)

TERCERO. Introducimos los datos para cada una de las variables, así como las frecuencias, si hay datos que se repiten. La forma más habitual es hacerlo utilizando las teclas (x_0, y_0) y (DATA).

x_i (x_0, y_0) y_j (×) f_{ij} (DATA)

CUARTO. Tras almacenar los datos, la calculadora nos da directamente las sumas, los productos y las medidas estadísticas, pulsando las teclas correspondientes.

$\overline{x} = 16{,}43$ $\quad \Sigma x = 98{,}6$ $\quad \Sigma x^2 = 1.715{,}12$

$\overline{y} = 36{,}58$ $\quad \Sigma y = 219{,}5$ $\quad \Sigma y^2 = 17.210{,}79$

$\sigma_X = 3{,}97$ $\quad \sigma_Y = 39{,}12$ $\quad \Sigma xy = 3.393{,}91$

QUINTO. Determinamos el coeficiente de correlación.

En este caso, el coeficiente de correlación, $r = -0{,}23$, es un valor próximo a 0; por tanto, la correlación entre las variables es casi nula.

SEXTO. Hallamos los coeficientes de la recta de regresión. La calculadora la presenta de la forma $y = ax + b$.

En este ejemplo, los coeficientes son: $a = -2{,}25$ y $b = 73{,}54$.

La calculadora nos permite hacer estimaciones; sin embargo, en este caso debemos observar que no tiene sentido realizarlas, ya que la correlación entre las variables es casi nula.

Date cuenta

Cada modelo de calculadora puede tener formas diferentes de introducir los datos. Por ello conviene consultar su manual de instrucciones antes de utilizarla.

ACTIVIDADES

19 Utiliza la calculadora para determinar todas las medidas estadísticas.

a)

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

b)

X	24	27	22	23	24	26	27	28	22	23
Y	2	1	2	4	5	2	3	4	1	2

20 Estudia la correlación entre estas variables, utilizando la calculadora para realizar las operaciones.

X	14	16	17	14	15	12	13	13	14	16
Y	32	34	36	34	32	34	31	36	38	32

Determina la recta de regresión y razona si tiene sentido estimar el valor de Y si la variable X toma el valor 18.

PROBLEMAS RESUELTOS

Variables bidimensionales

1. CÓMO SE ANALIZA EL TIPO DE RELACIÓN QUE EXISTE ENTRE LAS DOS VARIABLES QUE FORMAN UNA VARIABLE BIDIMENSIONAL

8 Las tablas siguientes muestran los datos correspondientes a varias personas que se acercan al cine para reservar sus entradas.

N.º de entradas	3	10	5	2	4	7
Precio (€)	22,50	75	37,50	15	30	52,50

Altura (cm)	150	155	161	172	175	177
Peso (kg)	53	58	60	72	70	80

Edad	21	45	18	30	40	25
N.º de entradas	3	10	5	2	4	7

Decide si existe dependencia entre las variables y encuentra, si es posible, una expresión algebraica que las relacione.

a) ¿Cuánto pagará una persona que compre 15 entradas?

b) ¿Cuál será el peso de una persona que mide 1,70 metros?

c) ¿Cuántas entradas comprará una persona de 28 años?

SOLUCIÓN

PRIMERO. Se analiza la relación entre las dos variables. Es necesario tener en cuenta que puede haber variables que aunque no están relacionadas mediante una expresión algebraica, influyen una en la otra y por tanto, existe dependencia entre ellas.

a) A mayor número de entradas le corresponde mayor importe. Las variables están relacionadas mediante la expresión $y = 7{,}5x$.

b) Aunque no hay una dependencia fuerte entre las dos variables, sí influyen una en la otra. Se puede estimar el peso de una persona a partir de su altura.

c) La edad de una persona no tiene ninguna relación con el número de entradas que compra.

SEGUNDO. Se determinan, si es posible, los datos a partir de la dependencia entre las variables.

a) Se tiene que: $y = 7{,}5x$. Por tanto, una persona que compre 15 entradas pagará:

$$7{,}5 \cdot 15 = 112{,}50\ €$$

b) Se puede estimar el peso de una persona a partir de su altura puesto que a partir de la altura de un grupo de personas se puede determinar su peso mediante relaciones estadísticas.

c) No es posible saber el número de entradas que comprará una persona de 28 años, ya que las variables no están relacionadas.

2. CÓMO SE AGRUPAN DATOS DE VARIABLES BIDIMENSIONALES EN INTERVALOS

9 Una prueba a la que se han presentado 20 personas consta de un test de inteligencia y un test de conocimientos.

Las puntuaciones obtenidas por cada persona han sido, respectivamente, las siguientes.

Test de inteligencia:

90 102 110 91 100 115 93
104 116 95 107 116 96 109
103 111 92 97 104 99

Test de conocimientos:

0,4 2 4 0 2,4 4,6 0,8
3 5,4 1 3,4 5,9 1,6 3,6
2,2 5 0,6 1,4 3,8 1,8

Agrupa las puntuaciones del test de inteligencia en intervalos de 10 puntos, y las puntuaciones del test de conocimientos, en intervalos de 2 puntos. Construye la tabla de frecuencias simple para datos agrupados.

SOLUCIÓN

PRIMERO. Se determinan los intervalos para la primera variable y las marcas de clase correspondientes.

Como las puntuaciones del test de inteligencia van de 90 a 116 puntos, se puede hacer la división en tres intervalos:

[90, 100) [100, 110) [110, 120)

Las marcas de clase son los puntos medios de cada intervalo, esto es:

95 105 115

SEGUNDO. Se determinan los intervalos para la segunda variable y las marcas de clase correspondientes.

Como las puntuaciones del test de conocimientos van de 0 a 5,9 puntos, podemos hacer la división en tres intervalos:

[0, 2) [2, 4) [4, 6)

En este caso, las marcas de clase son:

1 3 5

TERCERO. Se construye la tabla de frecuencias simple para datos agrupados.

Para ello se cuentan los datos que pertenecen a cada intervalo, para determinar las frecuencias absolutas, y se representa cada intervalo mediante su marca de clase.

Puntuación del test de inteligencia	95	105	115
Puntuación del test de conocimientos	1	3	5
Frecuencias absolutas	8	7	5

Tablas de frecuencias y gráficos

1. CÓMO SE CONSTRUYE Y SE LEE UNA TABLA DE DOBLE ENTRADA

10 Una agencia de viajes ha preparado una encuesta para preguntar sobre el número de viajes realizados durante el último año (Y), tanto de trabajo como en período de vacaciones, dependiendo de la edad (X). Se ha encuestado a un grupo de personas y las respuestas han sido:

(23, 2) (34, 1) (21, 2) (29, 5) (22, 4) (27, 0)
(36, 4) (38, 5) (47, 3) (49, 3) (42, 4) (35, 5)
(38, 7) (36, 5) (48, 8) (39, 2) (44, 5) (37, 7)
(21, 2) (24, 5) (26, 2) (27, 0) (42, 7) (43, 8)
(36, 5) (37, 5) (37, 3) (53, 5) (52, 6) (64, 7)

Agrupa los resultados en intervalos y construye una tabla de doble entrada.

a) ¿Cuántas personas menores de 35 años han sido encuestadas? ¿Y mayores de 50 años?

b) ¿Cuántas personas han viajado de 3 a 5 veces?

SOLUCIÓN

PRIMERO. Se determinan los intervalos para la primera variable.

Como la edad de las personas encuestadas varía de 21 a 64 años, $64 - 21 = 44$, podemos establecer tres intervalos de amplitud 15:

[20, 35) [35, 50) [50, 65)

SEGUNDO. Se determinan los intervalos para la segunda variable.

Como el número de viajes oscila de 0 a 8, $8 - 0 = 8$, podemos establecer tres intervalos de amplitud 3:

[0, 3) [3, 6) [6, 9)

TERCERO. Se construye la tabla de doble entrada, teniendo en cuenta que en cada casilla aparecerá la frecuencia absoluta conjunta correspondiente a los pares de datos que cumplen las dos condiciones que fijan los intervalos elegidos.

Y \ X	[20, 35)	[35, 50)	[50, 65)	Total
[0, 3)	7	1	0	8
[3, 6)	3	11	1	15
[6, 9)	0	5	2	7
Total	10	17	3	$N = 30$

CUARTO. Se observan las columnas de los totales y se extraen las conclusiones.

a) Se ha encuestado a 10 personas menores de 35 años y a 3 personas mayores de 50 años.

b) Las personas que han viajado de 3, 4 o 5 veces han sido 15.

2. CÓMO SE CONSTRUYE Y SE LEE UNA TABLA DE FRECUENCIAS RELATIVAS CONJUNTAS A PARTIR DE LA TABLA DE DOBLE ENTRADA

11 La tabla muestra los resultados de una encuesta sobre el número de viajes que se han realizado durante el último año (Y), tanto de trabajo como en período de vacaciones, dependiendo de su edad (X).

Y \ X	[20, 35)	[35, 50)	[50, 65)	Total
[0, 3)	7	1	0	8
[3, 6)	3	11	1	15
[6, 9)	0	5	2	7
Total	10	17	3	$N = 30$

Construye la tabla de frecuencias relativas conjuntas correspondiente.

a) ¿Qué porcentaje de personas menores de 35 años han sido encuestadas? ¿Y qué porcentaje de personas ha viajado de 3 a 5 veces?

b) ¿Qué porcentaje de personas entre 35 y 50 años ha viajado de 6 a 9 veces?

SOLUCIÓN

PRIMERO. Se divide cada una de las frecuencias absolutas conjuntas por el número total de datos.

En este caso, se divide en toda la tabla por 30.

Y \ X	[20, 35)	[35, 50)	[50, 65)	Total
[0, 3)	0,23	0,03	0	0,26
[3, 6)	0,10	0,37	0,03	0,50
[6, 9)	0	0,17	0,07	0,24
Total	0,33	0,57	0,10	1,00

SEGUNDO. Se elige la frecuencia relativa por la que se pregunta.

a) Personas menores de 35 años → [20, 35)
Frecuencia relativa → 0,33

Personas que han viajado de 3 a 5 → [3, 6)
Frecuencia relativa → 0,5

b) Personas de 35 a 50 años que han viajado de 6 a 9 veces → [35, 50) y [6, 9)
Frecuencia relativa → 0,17

TERCERO. Se multiplica cada una de las frecuencias relativas por 100 para obtener el porcentaje.

a) $0{,}33 \cdot 100 = 33\,\%$ es el porcentaje de personas encuestadas menores de 35 años.
El $0{,}5 \cdot 100 = 50\,\%$ ha viajado de 3 a 5 veces.

b) El $0{,}17 \cdot 100 = 17\,\%$ están entre 35 y 50 años y han viajado de 6 a 9 veces.

PROBLEMAS RESUELTOS

3. CÓMO SE REPRESENTAN VARIABLES BIDIMENSIONALES TENIENDO EN CUENTA LA FRECUENCIA DE LOS DATOS

12 Se estudia la relación entre la antigüedad, X, de los trabajadores de una empresa y el número de productos defectuosos, Y, que han elaborado en el último año.

Los resultados se recogen en esta tabla de doble entrada.

Y \ X	1	2	3	4	Total
3	10	5	5	0	20
4	20	10	15	5	50
Total	30	15	20	5	70

Representa los datos en un gráfico que refleje la frecuencia de cada uno de ellos.

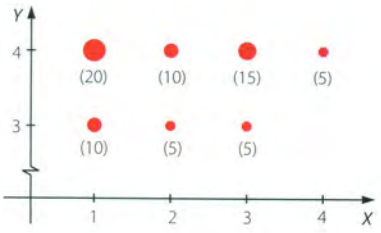
SOLUCIÓN

PRIMERO. Se interpreta la tabla de doble entrada.

Cada celda representa el número de trabajadores que cumple ambas condiciones. Por ejemplo, hay 20 empleados que llevan 1 año en la empresa y que han elaborado 4 productos defectuosos en el último año.

Así, al construir el diagrama de dispersión se debe reflejar que el par de datos (1, 4) aparece 20 veces.

SEGUNDO. Se dibuja la nube de puntos teniendo en cuenta la frecuencia de los datos. Se suelen representar los datos con puntos de diferente grosor dependiendo de su frecuencia que se indica entre paréntesis.



TERCERO. También se puede representar la distribución en el espacio, dibujando paralelepípedos de volúmenes proporcionales a las frecuencias. Este gráfico se llama **estereograma**.

Correlación

1. CÓMO SE CALCULA EL COEFICIENTE DE CORRELACIÓN EN UNA TABLA DE DOBLE ENTRADA

13 Esta tabla muestra los datos tomados sobre los ingresos mensuales, en euros, de 25 familias (X) y la superficie de la vivienda que habitan, en metros cuadrados (Y).

Estudia la correlación existente entre las dos variables.

Y \ X	[1.000, 2.000)	[2.000, 3.000)	[3.000, 4.000)	Total
[40, 70)	3	0	0	3
[70, 100)	4	2	0	6
[100, 130)	2	3	4	9
[130, 160)	0	2	5	7
Total	9	9	7	25

SOLUCIÓN

PRIMERO. Se construyen las tablas de frecuencias marginales.

Ingresos (X)	[1.000, 2.000)	[2.000, 3.000)	[3.000, 4.000)
Marca de clase	1.500	2.500	3.500
Frecuencia	9	9	7

Superficie (Y)	[40, 70)	[70, 100)	[100, 130)	[130, 160)
Marca de clase	55	85	115	145
Frecuencia	3	6	9	7

SEGUNDO. Se hallan las medias y las desviaciones típicas de cada una de las variables.

$\bar{x} = 2.500$ € $\sigma_X = 848{,}53$

$\bar{y} = 109$ m^2 $\sigma_Y = 29{,}39$

TERCERO. Se calcula la covarianza.

$$\sum_{i=1}^{3}\sum_{j=1}^{4} x_i \cdot y_j \cdot f_{ij} = 7.262.500$$

$$\sigma_{XY} = \frac{\sum_{i=1}^{3}\sum_{j=1}^{4} x_i \cdot y_j \cdot f_{ij}}{25} - \bar{x} \cdot \bar{y} = 18.000$$

CUARTO. Se determina el coeficiente de correlación.

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} = 0{,}721$$

QUINTO. Se interpreta el coeficiente de correlación.

Como el coeficiente de correlación está muy próximo a 1, se puede afirmar que hay cierta relación entre los ingresos mensuales de una familia (X) y la superficie de su vivienda (Y).

Rectas de regresión

1. CÓMO SE DETERMINAN LA MEDIA DE UNA DE LAS VARIABLES Y EL SIGNO DEL COEFICIENTE DE CORRELACIÓN, A PARTIR DE LA RECTA DE REGRESIÓN

14 Se ha realizado un estudio estadístico a un grupo de 100 alumnos . Con los datos recogidos se ha obtenido que la estatura media del grupo es de 155 cm, con una desviación típica de 15,5 cm.

Además, la recta de regresión que relaciona el peso de los alumnos, X, con su estatura, Y, es:

$$y = 80 + 1{,}5\,x$$

a) ¿Cuál es el peso medio del grupo de alumnos?

b) ¿Cuál será el signo de la covarianza?

c) ¿Se puede afirmar, en este grupo de alumnos, que cuanto mayor sea el peso hay mayor altura?

SOLUCIÓN

PRIMERO. Se despeja x en la recta de regresión y se determina el valor de la media para la variable X a partir de la media de la variable Y.

a) Al despejar x resulta:

$$y = 80 + 1{,}5x \rightarrow x = \frac{y - 80}{1{,}5}$$

Como $\bar{y} = 155$, se tiene que:

$$\bar{x} = \frac{155 - 80}{1{,}5} = 50$$

El peso medio del grupo de alumnos encuestados es de 50 kg.

SEGUNDO. Se estudia la pendiente de la recta para determinar el signo de la covarianza.

b) La pendiente de la recta de X sobre Y es:

$$1{,}5 = \frac{\sigma_{XY}}{\sigma_X^2}$$

Como σ_X^2 es positivo, la covarianza σ_{XY} será también positiva.

TERCERO. Se determina el signo del coeficiente de correlación.

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

c) Como σ_X y σ_Y son positivos y σ_{XY} es también positiva, entonces el coeficiente de correlación, r, es positivo.

Por tanto, la dependencia es positiva, es decir, cuando crece la variable X: el peso de los alumnos, crece la variable Y, su estatura.

2. CÓMO SE REALIZAN ESTIMACIONES PARA LAS DOS VARIABLES, A PARTIR DE LAS DOS RECTAS DE REGRESIÓN

15 La siguiente tabla recoge las notas en Matemáticas, X, y las notas medias de todas las asignaturas, Y, de 10 alumnos.

X	4	6	8	5	6	3	5	6	8	3
Y	5	7	8	6	6	4	6	7	8	4

a) Si un alumno obtiene un 7 en Matemáticas, ¿qué nota media se podría estimar?

b) Si un alumno tuviera un 3 de nota media ¿qué nota tendría en Matemáticas?

c) ¿Son fiables ambas estimaciones? Razona la respuesta.

SOLUCIÓN

PRIMERO. Se calculan las medidas estadísticas necesarias para determinar las rectas.

$\bar{x} = 5{,}4$ $\quad \sigma_X = 1{,}69$

$\bar{y} = 6{,}1$ $\quad \sigma_Y = 1{,}37$

$\sigma_{XY} = 2{,}26$

SEGUNDO. Se escriben las dos rectas de regresión.

$$\frac{\sigma_{XY}}{\sigma_X^2} = \frac{2{,}26}{1{,}69^2} = 0{,}8$$

$$\frac{\sigma_{XY}}{\sigma_Y^2} = \frac{2{,}26}{1{,}37^2} = 1{,}2$$

- La recta de regresión de Y sobre X es:

$$y - \bar{y} = \frac{\sigma_{XY}}{\sigma_X^2}(x - \bar{x}) \rightarrow y - 6{,}1 = 0{,}8(x - 5{,}4)$$

$$y = 0{,}8x + 1{,}78$$

- La recta de regresión de X sobre Y es:

$$x - \bar{x} = \frac{\sigma_{XY}}{\sigma_Y^2}(y - \bar{y}) \rightarrow x - 5{,}4 = 1{,}2(y - 6{,}1)$$

$$x = 1{,}2y + 1{,}92$$

TERCERO. Se calculan las estimaciones a partir de la recta de regresión correspondiente.

Para estimar la nota media (Y) a partir de una nota conocida de Matemáticas (X), se sustituye en la ecuación de la recta de regresión de Y sobre X.

$$y = 0{,}8 \cdot 7 + 1{,}78 = 7{,}38$$

Para estimar la nota de Matemáticas (X), conocida la nota media (Y), se sustituye en la ecuación de la recta de regresión de X sobre Y.

$$x = 1{,}2 \cdot 3 + 1{,}92 = 5{,}52$$

CUARTO. Se calcula el coeficiente de correlación para establecer la fiabilidad de las estimaciones.

$$r = 0{,}98$$

Ambas estimaciones son buenas ya que el coeficiente está próximo a 1. Por tanto, la dependencia lineal entre las variables es fuerte.

ACTIVIDADES

Diagramas de dispersión

21 ●○○ Representa la nube de puntos asociada a las siguientes distribuciones bidimensiones.

a) (2, 2) (3, 6) (5, 10) (6, 14) (8, 19) (9, 23) (10, 25)

b) (5, 2) (6, 0) (8, −2) (10, −7) (11, −9) (13, −13) (15, −17)

c) (120, 60) (122, 75) (126, 60) (128, 90) (130, 50) (132, 100) (136, 70)

d) (7, 3) (8, 9) (9, 2) (10, 8) (11, 5) (12, 1) (13, 7)

Decide si existe dependencia entre las variables y de qué tipo es.

22 ●○○ Representa la nube de puntos asociada a estas variables bidimensionales, y decide si hay dependencia entre las variables que las forman.

En caso afirmativo, califícala.

A	6	8	9	11	13	15	16	18
B	8	13	13	16	21	26	28	33

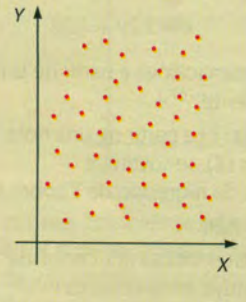
C	1	3	6	7	10	13	17	18
D	25	21	18	20	12	15	8	6

E	110	112	115	116	118	120	121	124
F	40	45	35	40	60	70	45	33

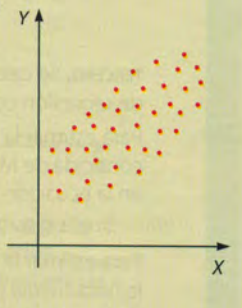
G	26	24	23	22	18	15	14	12
H	8	12	14	7	10	11	9	13

23 ●○○ A partir de los diagramas de dispersión, decide si hay o no dependencia lineal y, en su caso, si es fuerte o débil, y si es positiva o negativa.

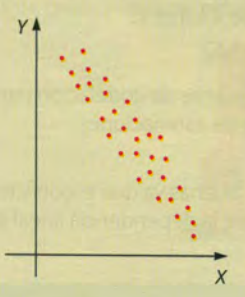
a)



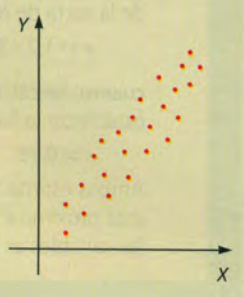
c)



b)



d)



24 ●○○ Representa las nubes de puntos correspondientes a las variables bidimensionales definidas por estas fórmulas.

a) $y = 2x + 5$

b) $y = x^2 + 3x$

¿Qué tipo de dependencia presentan?

25 ●○○ La tabla muestra el número de cuadros que han pintadolos alumnos de un taller sobre paisajes y bodegones.

Bodegones \ Paisajes	4	5	6	7	8
4	2	1	0	0	0
5	4	4	3	0	1
6	2	5	4	2	0
8	0	0	3	2	1

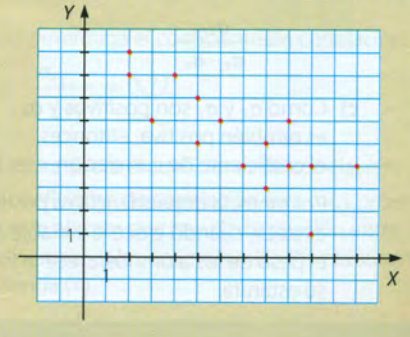
a) Determina las tablas de frecuencias marginales de paisajes y bodegones.

b) Calcula las medias y las desviaciones típicas de cada una de las variables.

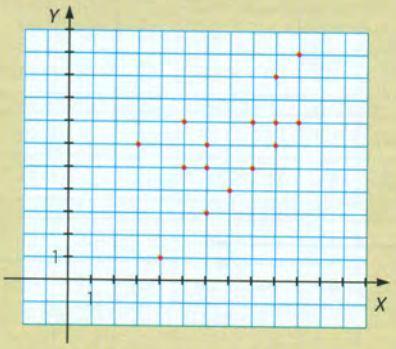
c) Usa el coeficiente de variación para decidir cuál de las dos variables es más dispersa.

d) Realiza el diagrama de dispersión correspondiente a la variable bidimensional.

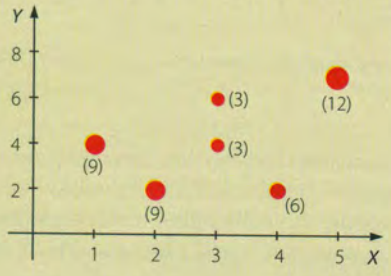
26 ●○○ Construye la tabla de doble entrada que corresponde a esta variable bidimensional, representada mediante el diagrama de dispersión.



27 A partir de este diagrama de dispersión, construye la tabla ●○○ de doble entrada correspondiente.



28 Construye la tabla de doble entrada correspondiente, ●○○ a partir del diagrama de dispersión, teniendo en cuenta la frecuencia de los datos que figura entre paréntesis.



Coeficiente de correlación

29 Calcula la covarianza y el coeficiente de correlación para ●○○ las variables bidimensionales indicadas en las siguientes tablas.

P	0	1	2	3	4	5	6	7
Q	20	18	17	15	12	10	7	4

R	90	80	70	60	50	40	30
S	−5	−7	−8	−11	−13	−16	−17

30 Halla la covarianza y el coeficiente de correlación ●○○ correspondientes a estas variables estadísticas.

T	−12	−14	−15	−16	−18	−20	−22
U	8	5	3	12	20	10	6

V	2,4	2,8	3,2	3,6	4	4,4	4,8	5,2
W	100	150	220	270	340	400	460	520

31 Representa la variable bidimensional cuyos pares ●●○ de datos son:

(8, 2) (12, 6) (10, 4) (12, 2) (8, 6)

a) Calcula su covarianza y razona el resultado.

b) Elimina un punto de manera que se mantenga la correlación.

32 Construye el diagrama de dispersión correspondiente ●●○ a la variable bidimensional determinada por los siguientes datos.

(10, 20) (16, 30) (10, 30) (16, 20)

a) Calcula su covarianza y explica a qué se debe el resultado.

b) Añade un punto de manera que se mantenga la correlación.

33 En la tabla se presentan datos climatológicos referidos ●●○ a una ciudad: la temperatura, en °C; la humedad relativa del aire, en %, y la velocidad del viento, en km/h.

Días	L	M	X	J	V	S	D
Temperatura	22	24	25	24	23	21	20
Humedad	78	90	80	92	88	74	80
Velocidad del viento	1	3	6	4	4	1	0

Determina la covarianza y el coeficiente de correlación de las siguientes variables bidimensionales.

a) *Temperatura–Humedad.*

b) *Temperatura–Velocidad del viento.*

c) *Humedad–Velocidad del viento.*

34 Se ha hecho una encuesta a personas que han tenido ●●○ un accidente de tráfico preguntando por el número de meses transcurridos e incluyendo el grupo de edad.

Las respuestas han sido:

Carmen, 35: [60, 70)
Teresa, 15: [50, 60)
Pilar, 12: [50, 60)
Esther, 6: [20, 30)
Juan, 8: [40, 50)
Jacinto, 15: [30, 40)
Jesús, 24: [50, 60)
Marta, 12: [30, 40)
José, 28: [40, 50)
Andrés, 3: [20, 30)
María Jesús, 20: [40, 50)
Beatriz, 16: [30, 40)

a) Construye la tabla correspondiente a la variable bidimensional.

b) Representa el diagrama de dispersión.

c) Estudia si hay correlación entre ambas variables y determina su coeficiente de correlación lineal.

ACTIVIDADES

35 En la siguiente tabla se han perdido dos datos.

●●○

	23	24	25	27	28	29	33	34	36
2	4	3	5		6	7	9	6	8

Se sabe que la media de la primera variable es 28 y la media de la segunda variable es 5,8. Completa la tabla y determina el coeficiente de correlación.

36 Se está estudiando imponer un impuesto a las empresas ●●○ químicas que sea proporcional a sus emisiones de azufre a la atmósfera. Se ha experimentado con varios procedimientos para medir dichas emisiones, pero no se ha encontrado ninguno fiable. Finalmente, se ha decidido investigar algún método indirecto.

Se cree que la emisión de azufre puede estar relacionada con el consumo eléctrico, con el consumo de agua o con el volumen de las chimeneas de las fábricas. Para valorarlo se ha realizado un estudio en un medio controlado.

Los resultados pueden verse en la tabla.

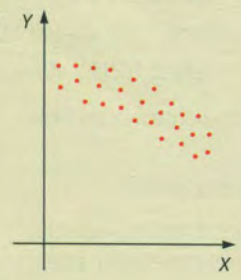
Cantidad de azufre (t)	2,3	1,8	1	0,4	0,6	3	0,5
Consumo eléctrico (kWh)	1.400	1.250	1.850	600	300	3.400	400
Consumo de agua (ℓ)	100	230	45	50	10	540	22
Volumen de las chimeneas (m^3)	18	16	12	5	6	21	4

¿Cuál de las medidas estadísticas se relaciona de forma más evidente con las emisiones de azufre? Justifica la respuesta.

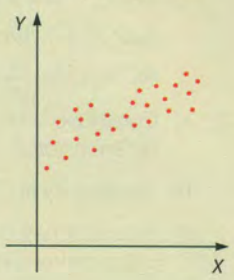
Rectas de regresión

37 Traza a mano alzada, y sin realizar cálculos, la recta ●○○ de regresión de las siguientes variables bidimensionales.

a)

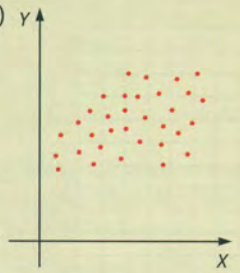


b)

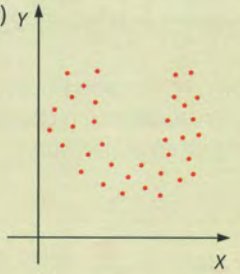


38 Representa, sin hallar su ecuación, la recta de regresión ●○○ correspondiente a estas variables.

a)

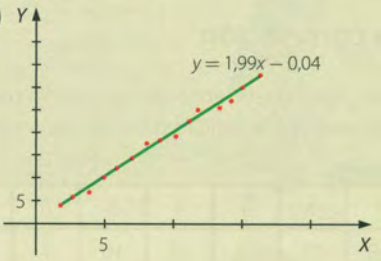


b)

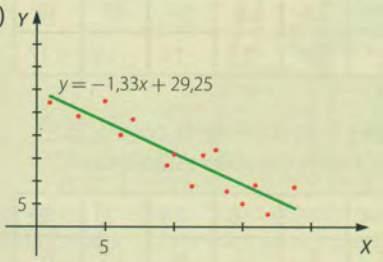


39 Para las variables bidimensionales representadas a ●○○ continuación, hemos ajustado las nubes de puntos correspondientes a diferentes rectas de regresión. Estima el valor que tendrá y en cada una de ellas para un valor de $x = 12$.

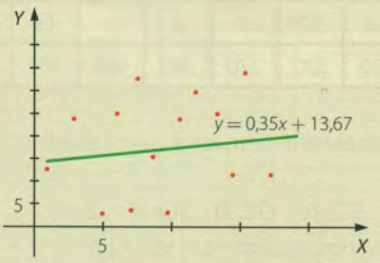
a)



b)



c)



¿Cuál de las estimaciones te parece más fiable?

40 Determina la recta de regresión de Y sobre X y la recta de regresión de X sobre Y correspondientes a estas tablas.

a)

X	10	11	12	13	14	15	16	17
Y	20	24	28	30	36	32	42	40

b)

X	60	70	80	90	100	110	120
Y	−5	−8	−12	−15	−16	−24	−20

c)

X	−3	−4	−5	−6	−9	−10	−13
Y	80	92	100	88	76	70	60

d)

X	0,2	0,4	0,5	0,7	0,8	0,9	1	1,2
Y	40	50	120	70	40	40	60	50

41 Encuentra cinco puntos que pertenecen a la recta.

$$y = 4x + 6$$

a) Calcula el coeficiente de correlación correspondiente y explica el resultado.

b) Halla las dos rectas de regresión.

42 Obtén cinco puntos que pertenecen a la recta.

$$y = -20x + 10$$

a) Calcula el coeficiente de correlación y explica el resultado.

b) Halla las dos rectas de regresión.

Razona los resultados obtenidos.

43 Se cree que el número de zorros en una finca está relacionado con el número de conejos.

En los últimos años se han realizado ocho censos de ambos animales, resultando estos datos.

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

Si la correlación es fuerte:

a) Determina las dos rectas de regresión.

b) Estima la cantidad de conejos que habría si hubiera 10 zorros.

c) ¿Cuántos zorros serían si hubiéramos contado 350 conejos?

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

44 A lo largo de un día se han medido la tensión y el pulso cardíaco de una persona, tratando de decidir si ambas variables tienen alguna relación.

Los datos obtenidos se han reflejado en la tabla.

Nivel mínimo de tensión	6	5	9	4	10	8	6	9
N.º de pulsaciones por minuto	60	55	80	40	95	75	55	90

a) Calcula la covarianza, el coeficiente de correlación y las dos rectas de regresión.

b) Si la correlación es fuerte, estima las pulsaciones que tendrá la persona cuando su nivel mínimo de tensión sea 15.

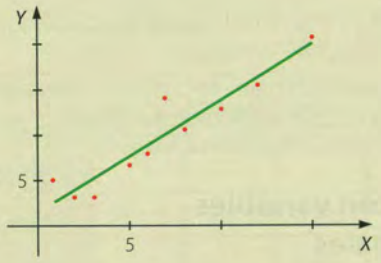
c) ¿Qué nivel mínimo de tensión se estima cuando las pulsaciones cardíacas por minuto son 70?

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

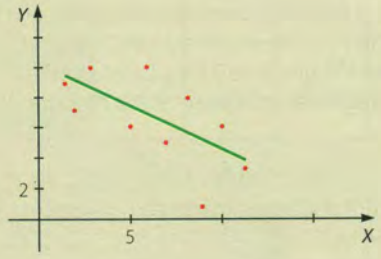
e) Dibuja la nube de puntos y la recta de regresión correspondientes.

45 Tenemos dos variables bidimensionales representadas por estas nubes de puntos.

(I)



(II)



a) Elige los coeficientes de correlación de ambas y razónalo.

−0,92 0,95

0,6 −0,65

b) Ahora decide cuáles son las ecuaciones de las dos rectas de regresión correspondientes.

$y = 3x + 0,2$

$y = 1,3x + 0,9$

$y = -0,6x + 10$

$y = -2x + 12,6$

Justifica la respuesta.

ACTIVIDADES

46 ●●○ Una empresa está investigando la relación entre sus gastos en publicidad y sus beneficios (en millones de euros).

Este es un resumen del estudio.

Año	98	99	00	01	02	03	04	05	06	07
Gastos	2	2,4	2	2,8	3	3,2	3,2	3,3	3,5	4
Beneficios	12	15	13	15	18	19	19	20	20	22

a) Comprueba si existe relación entre las magnitudes y, si es posible, estima los beneficios que se obtendrán en el año 2008, si se va a invertir 4,2 millones de euros en publicidad.

b) ¿Qué inversión sería necesaria para alcanzar 30 millones de euros de beneficios?

47 ●●○ María y Diego viven en la misma calle, pero en aceras opuestas. Los dos tienen un termómetro en su balcón y, como María cree que el suyo está estropeado, deciden tomar la temperatura exterior, en °C, durante una semana y a la misma hora del día.

Han anotado los resultados en una tabla.

Diego	22	24	25	27	18	20	21
María	18	20	18	17	20	21	16

a) ¿Crees que las dos variables están relacionadas? ¿Y opinas que deberían estarlo?

b) Razona si con estos datos se puede obtener alguna conclusión sobre el termómetro de María.

Problemas con variables bidimensionales

48 ●●○ Se ha medido el peso, X, y la estatura, Y, de los alumnos de una clase. Su peso medio ha sido de 56 kg, con una desviación típica de 2,5 kg. La ecuación de la recta de regresión que relaciona la estatura y el peso es:

$$y = 1{,}8x + 62$$

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

b) ¿Y si pesara 44 kg?

c) ¿Cuál es la estatura media de esos alumnos?

d) La pendiente de esa recta es positiva. ¿Qué significa esto?

49 ●●○ Daniel afirma que si una nube de puntos es de una recta, el coeficiente de correlación siempre vale 1 o −1. Como Eva no está de acuerdo, Daniel prueba con los puntos de la recta cuya ecuación es $y = -5x + 20$, y Eva hace lo mismo con los puntos de $y = 2x - x^2$.

a) ¿Quién tiene razón? ¿Por qué?

b) Si la hipótesis de Daniel no resulta cierta, ¿podrías formularla de forma que se verifique siempre?

50 ●●○ Un equipo de alpinistas que escaló una montaña, midió la altitud y la temperatura cada 200 metros de ascensión. Luego reflejó los datos en estas tablas.

Altitud (m)	800	1.000	1.200	1.400	1.600	1.800	2.000
Temperatura (°C)	22	20	17	15	11	9	8

Altitud (m)	2.200	2.400	2.600	2.800	3.000	3.200
Temperatura (°C)	5	3	2	2	2	1

a) Toma las diez primeras mediciones y, si la correlación es fuerte, calcula la recta de regresión de la temperatura sobre la altitud.

b) Estima la temperatura que habrá a los 1.900 metros de altitud.

c) ¿Qué temperatura se estima a los 3.200 metros? ¿Cómo explicas las diferencias?

51 ●○○ El alcalde de un pueblo ha constatado una reducción del número de nacimientos de niños, y ha encargado realizar un estudio.

Año	86	89	92	95	98	01	04	07
Nacimientos	50	54	40	33	34	23	21	17

a) ¿Puede establecerse, de forma fiable, una fórmula que relacione el año con el número de nacimientos?

b) ¿Cuántos nacimientos pueden estimarse en 2008? ¿Y en 2010? ¿Qué puede estimarse para 2050?

c) ¿Es fiable esta última estimación? Razona la respuesta.

52 ●○○ En una empresa se está estudiando el número de días de baja por enfermedad, Y, de cada uno de sus empleados en el último año. Para compararlo con la antigüedad, X, de los empleados dentro de la empresa, se ha elaborado la siguiente tabla.

Y \ X	1	2	3	4	5
0	6	12	8	3	0
2	4	5	3	2	1
3	0	1	3	2	0
5	0	0	2	2	1
9	0	0	0	0	1

a) Calcula las medias y las desviaciones típicas de las variables unidimensionales.

b) Determina la covarianza y el coeficiente de correlación.

c) Halla la recta de regresión de Y sobre X y estima, si es fiable, el número de días de baja que puede esperarse en un empleado con 6 años de antigüedad en la empresa.

53 Un inversor bursátil quiere predecir la evolución ●●○ que va a tener el Índice de la Bolsa de Madrid (IBEX).

Ha concluido que lo que sucede con el IBEX un día es lo que le sucede a la cotización de la empresa AW&B el día anterior.

Investiga si esto es correcto, a partir de sus cotizaciones durante una semana y los valores alcanzados por el IBEX al día siguiente.

Día	1.º	2.º	3.º	4.º	5.º	6.º	7.º
AW&B	21,8	23,4	19,6	19,4	18,4	19,9	19,2

Día	2.º	3.º	4.º	5.º	6.º	7.º	8.º
IBEX	12.560	12.720	11.580	11.420	10.930	11.450	11.480

a) ¿Qué cotización tendrá AW&B el día anterior al día en que el IBEX alcance los 14.000 puntos?

b) Si un día AW&B tiene una cotización de 24 euros, ¿qué valor podemos esperar que alcance el IBEX al día siguiente?

54 Encuentra el coeficiente de correlación de la variable ●●● bidimensional cuyas rectas de regresión son:

- Recta de Y sobre X:

$$2x - y - 1 = 0$$

- Recta de X sobre Y:

$$9x - 4y - 9 = 0$$

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

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

55 Se tiene la siguiente variable bidimensional. ●○○

X	3	5	8	9	10	12	15
Y	2	3	7	4	8	5	8

Investiga lo que sucede con la covarianza y el coeficiente de correlación en cada caso.

a) Sumamos 10 a todos los valores de la variable X.

b) Sumamos 10 a todos los valores de la variable X y de la variable Y.

c) Multiplicamos por 4 todos los valores de la variable X.

d) Multiplicamos por 4 todos los valores de la variable X y de la variable Y.

56 Investiga sobre las siguientes cuestiones. ●○○

a) ¿Es cierto que el signo de las pendientes de las dos rectas de regresión de una variable bidimensional es siempre igual?

b) ¿Qué sucede si las dos rectas de regresión tienen la misma pendiente? ¿Cómo es la correlación?

57 El ángulo que forman las dos rectas de regresión ●○○ de una distribución bidimensional es mayor cuanto menor sea el coeficiente de correlación.

Vamos a comprobarlo estudiando las dos magnitudes en estas distribuciones.

10	12	14	16	18
3	8	1	9	2

10	12	14	16	18
3	6	8	6	7

10	12	14	16	18
5	6	6,5	8,5	9

58 Se ha realizado un test de memoria, X, y otro test ●●○ de atención, Y, a varios alumnos y se han reflejado los resultados en esta tabla.

Y \ X	[0, 10)	[10, 20)	[20, 30)	[30, 40)	[40, 50)
[0, 10)					
[10, 20)	Beatriz	Jesús	Marta		
[20, 30)		Daniel	María Esther	Miguel	
[30, 40)			Elena	Jacinto Carmen	Inés
[40, 50)				Diego	

a) Calcula la covarianza y el coeficiente de correlación.

b) Determina las dos rectas de regresión.

c) Si es factible, estima qué puntuación obtendrá Andrés en memoria, si ha obtenido 33 en atención.

d) Si es factible, estima qué puntuación obtendrá Eva en atención, si ha obtenido 27 en memoria.

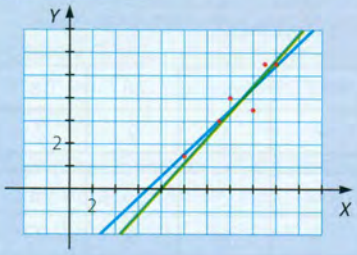
PARA FINALIZAR...

Reflexiona sobre la teoría

59 Halla la relación existente entre el coeficiente de correlación lineal de una distribución bidimensional y las pendientes de sus rectas de regresión.

Comprueba el resultado obtenido para estos datos.

X	10	13	16	14	17	18
Y	3	6	7	8	11	11



60 Discute si es posible que la recta de regresión de X sobre Y y la recta de regresión de Y sobre X sean paralelas. ¿Y perpendiculares?

61 Investiga sobre cómo varía el coeficiente de correlación entre dos variables estadísticas cuando multiplicamos los datos relativos a una de ellas por una cantidad constante, k.

¿Y si las multiplicamos por la misma constante? ¿Qué sucedería si multiplicamos cada variable por una constante distinta?

IDEA CLAVE

Comprueba lo que ocurre, en cada caso, con la covarianza y las desviaciones típicas, teniendo en cuenta que $r = \dfrac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$.

62 Demuestra que el coeficiente de correlación de dos variables estadísticas no varía si a cada valor de las dos variables se les suma o resta un mismo número.

Utiliza esta propiedad para calcular el coeficiente de correlación de las siguientes variables estadísticas.

X	2.001	2.002	2.003	2.004	2.005
Y	7.390	7.350	7.240	7.210	7.110

Piensa un poco más

63 En dos estudios realizados sobre los datos de una variable bidimensional, las rectas de regresión fueron las siguientes.

En el primer estudio, la recta de regresión de Y sobre X es: $8x - 3y - 61 = 0$ y la recta de X sobre Y es: $x - y + 18 = 0$.

Y en el otro estudio, las rectas de regresión son respectivamente:

$$8x - 5y + 20 = 0 \qquad 5x - 2y - 10 = 0$$

Si conocemos $\bar{x} = 23$, $\bar{y} = 41$ y $r = 0{,}8$, comprueba cuál de los estudios es válido.

Análisis del enunciado

Tenemos dos estudios relativos a los mismos datos que presentan resultados diferentes. Queremos averiguar si son o no válidos.

Diseño de la resolución

Comprobamos si, conocidas las medias y el coeficiente de correlación, es posible obtener esas rectas.

Clave para resolver el problema

Halla el punto de corte de las dos rectas.

64 Sean dos variables estadísticas X e Y. Sabemos que:

- La recta de regresión de Y sobre X pasa por los puntos (1, 3) y (2, 5).
- La recta de regresión de X sobre Y tiene pendiente $m = 3$ y su ordenada en el origen es 2.
- La varianza de Y es 3.

Calcula las medidas estadísticas de cada una de las variables estadísticas y el coeficiente de correlación.

Análisis del enunciado

Conocidas las dos rectas de regresión, hay que calcular la media, la varianza y la desviación típica de cada variable, así como la covarianza y el coeficiente de correlación.

Diseño de la resolución

Se determinan las rectas de regresión y el punto de corte de ambas.

Clave para resolver el problema

Con el punto de corte de ambas rectas hallamos las medias de las variables, y con los coeficientes de las rectas de regresión se forma un sistema de ecuaciones no lineal, que nos permitirá calcular las demás medidas estadísticas.