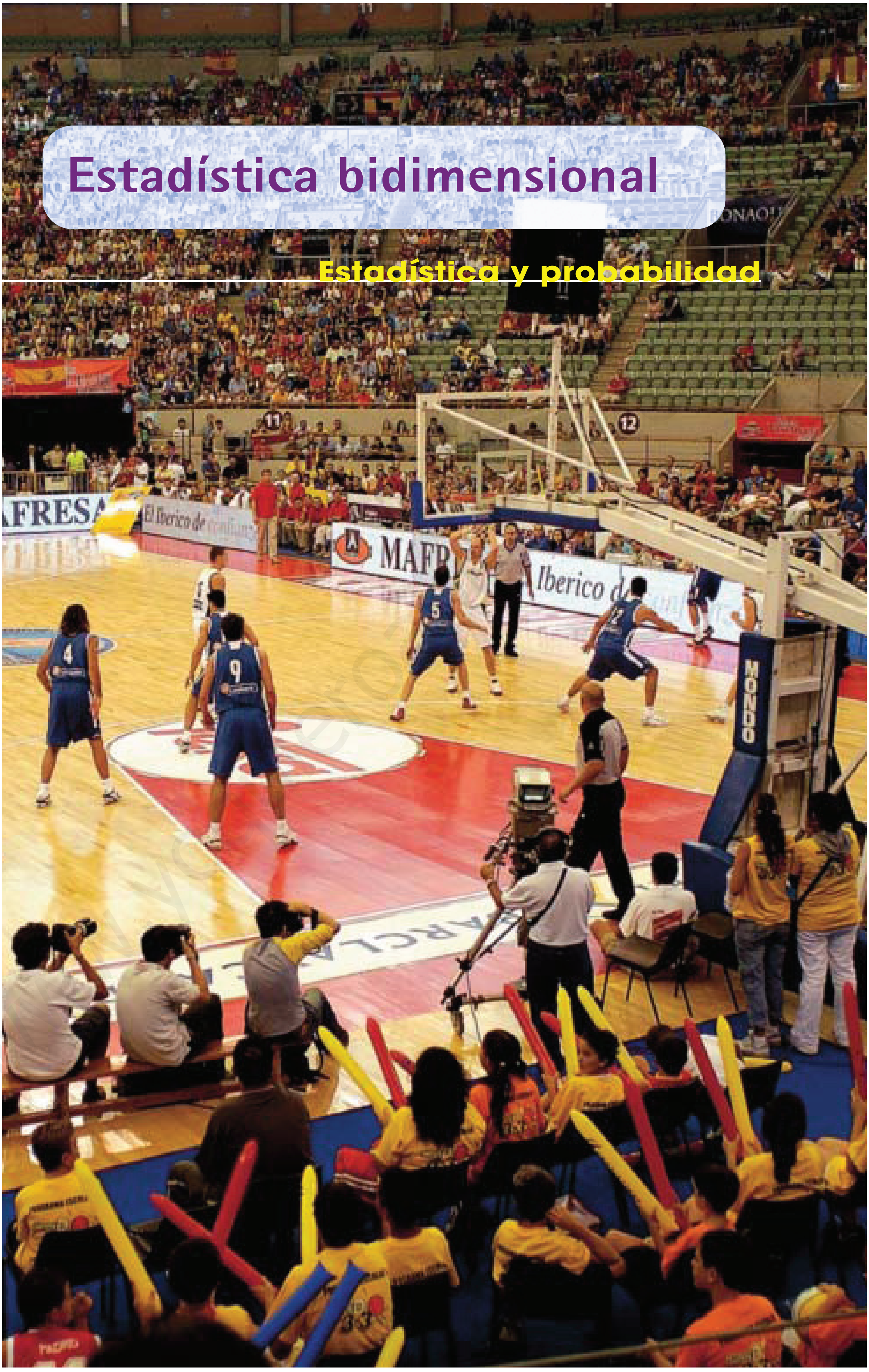
12

Estadística bidimensional

Estadística y probabilidad



Introducción

Las variables estadísticas pueden estudiarse de forma conjunta cuando se desea ver la relación entre dos o más caracteres del mismo individuo. Por ejemplo, se puede estudiar la relación que hay entre la estatura y el peso de unos jugadores de un equipo de baloncesto, o la relación que hay entre la cantidad de un medicamento y el tiempo que tarda en hacer reacción, etcétera.

En este tema se trabajan las distribuciones bidimensionales. En concreto, se estudia la forma de organizar la información en tablas de frecuencia, los parámetros que permiten interpretar dicha información, la correlación, que es la medida del grado de relación entre las variables, y la regresión, que estudia una variable condicionada al comportamiento de la otra.

Organiza tus ideas

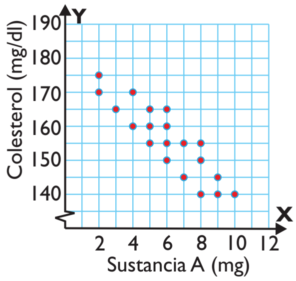
Distribuciones bidimensionales

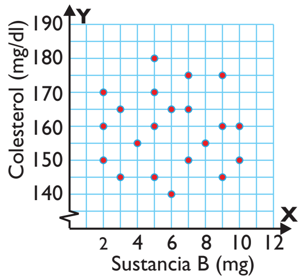
- se organizan → en tablas de frecuencias:
  - simples
  - doble entrada
- se calculan las →
  - medias marginales
  - desviaciones típicas marginales
  - covarianza
- se estudia la →
  - correlación → coeficiente de correlación de Pearson
  - regresión → recta de mínimos cuadrados

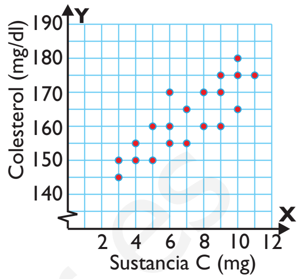
1. Distribuciones bidimensionales

Piensa y calcula

Se ha administrado una sustancia A, otra B y otra C a 20 individuos para estudiar su relación con los niveles de colesterol. Observando las gráficas, indica qué sustancia tiene mayor relación con la subida o bajada de colesterol.







1.1. Distribución bidimensional

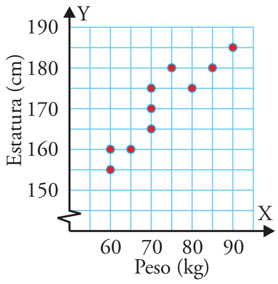
Una **distribución bidimensional** es la que se obtiene al estudiar un fenómeno respecto de dos variables estadísticas unidimensionales X e Y

Los **datos** de una distribución bidimensional son pares $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, donde $x_1, x_2, \ldots, x_n$ son los valores de la variable X, y donde $y_1, y_2, \ldots, y_n$ son los valores de la variable Y

Ejemplo

La distribución bidimensional que se obtiene al estudiar la estatura y el peso de 10 personas es:

Peso (kg)	70	65	85	60	70	75	90	80	60	70
Estatura (cm)	175	160	180	155	165	180	185	175	160	170

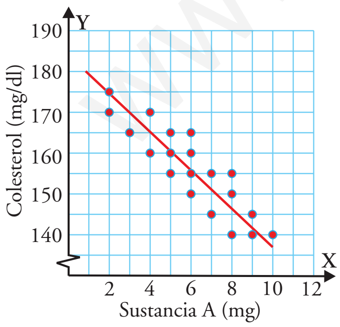


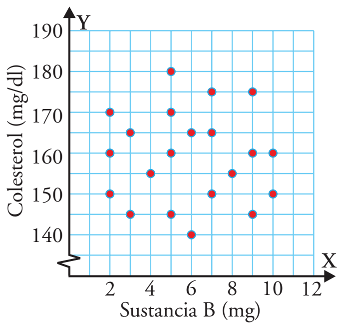
Nube de puntos o diagrama de dispersión

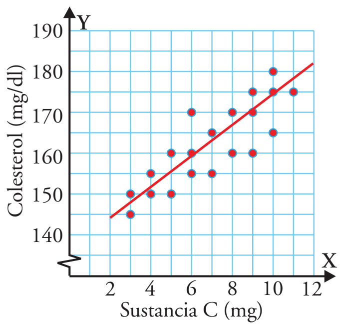
Una **nube de puntos** o **diagrama de dispersión** es la representación en unos ejes cartesianos de los datos (x_i, y_i) de una distribución bidimensional.

En una nube de puntos se puede apreciar de forma general la relación que existe entre las variables.

Ejemplo







En la gráfica de la izquierda se observa que a mayor cantidad de la sustancia A, los niveles de colesterol bajan. La gráfica de la derecha muestra lo contrario. Se puede decir que la sustancia A es buena para bajar el colesterol y que la sustancia C es perjudicial. En la gráfica central no hay relación.

1.2. Tabla de frecuencias

Cuando el número de datos de una distribución bidimensional es pequeño, se trabaja con los datos ordenados, pero cuando el número de datos es grande, se trabaja con tablas de frecuencias. Dichas tablas pueden darse de dos maneras:

a) **Tablas simples:** se recogen en fila o columna las frecuencias de los datos.

b) **Tablas de doble entrada:** se recoge en cada casilla la frecuencia correspondiente a cada fila y cada columna de los valores de cada variable.

Ejemplo

Dada la siguiente distribución bidimensional:

Variable X	1	1	1	1	1	1	1	1	1	2	2	2	2	2	3	3	4	4	4	4
Variable Y	1	1	3	3	3	3	4	4	4	1	1	2	4	4	2	2	1	3	3	3

se obtienen las tablas de frecuencias.

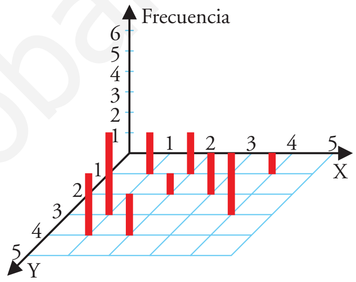
Tabla simple

X	Y	Frecuencia
1	1	2
1	3	4
1	4	3
2	1	2
2	2	1
2	4	2
3	2	2
4	1	1
4	3	3

Tabla de doble entrada

Y \ X	1	2	3	4	
1	2	2	0	1	**5**
2	0	1	2	0	**3**
3	4	0	0	3	**7**
4	3	2	0	0	**5**
	9	**5**	**2**	**4**	**20**

Diagrama de barras



Aplica la teoría

1. Las calificaciones de 30 estudiantes en dos exámenes han sido las siguientes:

1^er^ examen	4	5	6	7	7	9	10
2° examen	5	5	7	6	7	8	10
N° estudiantes	5	10	4	2	4	3	2

Haz la tabla de frecuencia de doble entrada.

2. Dibuja el diagrama de barras correspondiente a la siguiente distribución bidimensional:

Y \ X	1	2	3	4	5
1	3	0	0	0	0
2	0	1	0	0	0
3	2	1	6	0	0
4	0	3	0	3	2
5	0	0	0	0	4

3. Dibuja la nube de puntos de la siguiente distribución bidimensional:

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

2. Parámetros

Piensa y calcula

La siguiente distribución recoge las calificaciones de Matemáticas y de Lengua de un grupo de 6 alumnos. Calcula mentalmente la media de cada asignatura:

Matemáticas	2	3	5	5	6	9
Lengua	4	4	5	6	7	10

2.1. Parámetros

Las **medias marginales** son las medias de las variables X e Y:

$$\bar{x} = \frac{\Sigma\, n_i \cdot x_i}{N} \qquad \bar{y} = \frac{\Sigma\, n_i \cdot y_i}{N}$$

El **centro de gravedad** es el par de valores de las medias marginales: $G(\bar{x}, \bar{y})$

Las **desviaciones típicas marginales** son las desviaciones típicas de las variables X e Y:

$$s_x = \sqrt{\frac{\Sigma\, n_i \cdot x_i^2}{N} - \bar{x}^2}$$

$$s_y = \sqrt{\frac{\Sigma\, n_i \cdot y_i^2}{N} - \bar{y}^2}$$

$$s_x = \sqrt{\frac{\Sigma\, n_i \cdot (x_i - \bar{x})^2}{N}}$$

$$s_y = \sqrt{\frac{\Sigma\, n_i \cdot (y_i - \bar{y})^2}{N}}$$

Covarianza

La **covarianza** de una variable bidimensional (X, Y) es:

$$s_{xy} = \frac{\Sigma\, n_i x_i y_i}{N} - \bar{x} \cdot \bar{y}$$

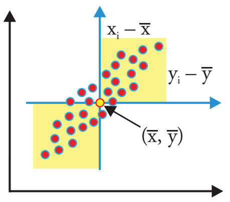
$$s_{xy} = \frac{\Sigma\, n_i(x_i - \bar{x})(y_i - \bar{y})}{N}$$

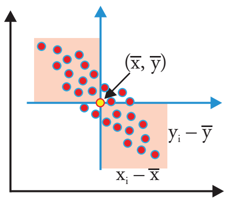
Interpretación de la covarianza

Según sea el signo de la covarianza, se interpreta:

a) **Covarianza positiva:** al aumentar los valores de la variable X, aumentan los valores de la variable Y. La nube de puntos se orienta a la derecha y hacia arriba.

b) **Covarianza negativa:** al aumentar los valores de la variable X, disminuyen los valores de la variable Y. La nube de puntos se orienta a la derecha y hacia abajo.





Si se calcula el centro de gravedad $G(\bar{x}, \bar{y})$ y se toman unos ejes con el origen en este centro, se observa:

- Si los puntos están en el 1^er^ y 3^er^ cuadrantes, mayoritariamente los productos $(x_i - \bar{x})(y_i - \bar{y})$ son positivos.
- Si los puntos están en el 2º y 4º cuadrantes, mayoritariamente los productos $(x_i - \bar{x})(y_i - \bar{y})$ son negativos.

Ejemplo

Calcula los parámetros de la distribución del número de años de antigüedad en una empresa y el salario diario que tienen 40 trabajadores. Interpreta los resultados.

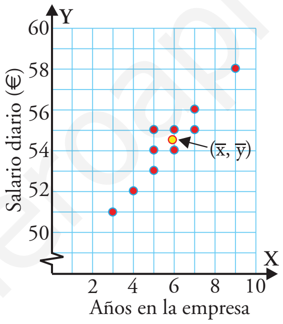
Años: x_i	Salario: y_i	n_i	$n_i \cdot x_i$	$n_i \cdot x_i^2$	$n_i \cdot y_i$	$n_i \cdot y_i^2$	$n_i \cdot x_i \cdot y_i$
4	52	3	12	48	156	8 112	624
5	54	4	20	100	216	11 664	1 080
7	55	5	35	245	275	15 125	1 925
6	54	7	42	252	378	20 412	2 268
5	53	3	15	75	159	8 427	795
7	56	5	35	245	280	15 680	1 960
5	55	4	20	100	220	12 100	1 100
9	58	3	27	243	174	10 092	1 566
3	51	2	6	18	102	5 202	306
6	55	4	24	144	220	12 100	1 320
Total		**40**	**236**	**1 470**	**2 180**	**118 914**	**12 944**

$\bar{x} = \dfrac{236}{40} = 5{,}9 \qquad \bar{y} = \dfrac{2\,180}{40} = 54{,}5$

$$s_x = \sqrt{\frac{1\,470}{40} - 5{,}9^2} = 1{,}39$$

$$s_y = \sqrt{\frac{118\,914}{40} - 54{,}5^2} = 1{,}61$$

$$s_{xy} = \frac{12\,944}{40} - 5{,}9 \cdot 54{,}5 = 2{,}05$$



Interpretación de los resultados: al ser la covarianza positiva, la nube de puntos se orienta a la derecha y hacia arriba; es decir, al aumentar la antigüedad, aumenta el salario.

Calculadora

Se hace para la **CASIO fx-82MS.** Para otras es análogo; ver el manual.

a) Se selecciona el modo **REG/Lin**
[MODE] (REG) [3] (Lin) [1]

b) Se borran todos los datos:
[SHIFT] [CLR] (Scl) [1] [=]

c) Se escribe el 1^er dato de X, se pulsa una coma [,], se escribe el 1^er dato de Y y se pulsa [DT]
Si la frecuencia es mayor que uno, se pulsa antes de la frecuencia [;] como se indica:
[4] [,] [52] [;] [3] [DT] [...]

d) Si se introduce un dato erróneo, se puede modificar utilizando [▲] o [▼] para buscar el dato o la frecuencia; se introduce y se pulsa la tecla [=]

e) Se obtienen los resultados:

- Media $\bar{x}$
[SHIFT] [S-VAR] ($\bar{x}$) [1] [=] **5,9**
- Media $\bar{y}$
[SHIFT] [S-VAR] [▶] ($\bar{y}$) [1] [=] **54,5**
- Desviación típica: s_x
[SHIFT] [S-VAR] ($x\sigma n$) [2] [=] **1,39**
- Desviación típica: s_y
[SHIFT] [S-VAR] [▶] ($y\sigma n$) [2] [=] **1,61**
- Covarianza: s_{xy}
[Σxy] [÷] [n] [–] [$\bar{x}$] [×] [$\bar{y}$] [=] **2,05**

Para obtener: Σxy
[SHIFT] [S-SUM] [▶] (Σxy) [3]

Para obtener: n
[SHIFT] [S-SUM] (n) [3]

Volver a modo normal la calculadora

Se pulsan las teclas:
[SHIFT] [CLR] (Mode) [2] [=]

Aplica la teoría

4. Calcula la covarianza de la siguiente distribución bidimensional:

x_i	8	7	6	5	7	8	6	5
y_i	5	4	7	4	3	6	5	5
n_i	2	4	3	5	3	4	2	2

5. Calcula la covarianza de la siguiente distribución bidimensional:

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

6. La siguiente tabla recoge la distribución de la cilindrada de un motor y la velocidad máxima que puede generar:

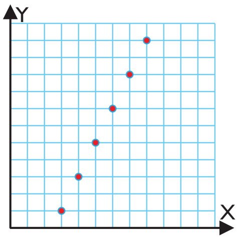
Cilindrada (cm^3)	Velocidad (km/h)
1 000	125
1 200	130
1 400	140
1 600	145
1 600	150
1 800	170
2 000	190
2 000	195

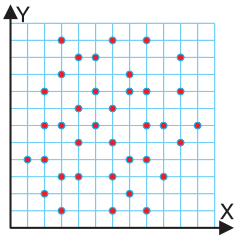
a) Representa la nube de puntos.
b) Representa el centro de gravedad.
c) Calcula e interpreta la covarianza.

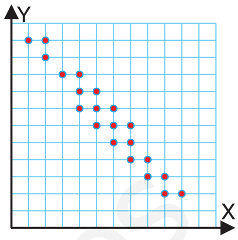
3. Correlación

Piensa y calcula

Indica el signo de la covarianza y si la relación entre las variables es funcional, fuerte o nula en los siguientes casos:





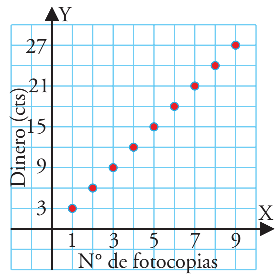


3.1. Correlación

La **covarianza** indica cómo es la relación entre dos variables; es decir, cómo se orienta la nube de puntos, pero este parámetro no indica de una forma precisa la medida de esa relación. Para resolver este problema, se definen los conceptos de **correlación** y **coeficiente de correlación**.

Correlación es la relación que existe entre las dos variables que intervienen en una distribución bidimensional.

a) Correlación funcional: todos los puntos están situados sobre una recta o una curva. Existe una relación funcional entre las variables X e Y



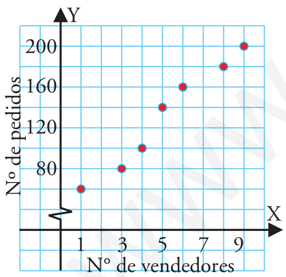
Ejemplo

El precio de las fotocopias de una copistería es:

Nº copias: x_i	1	2	3	4	5	6
Dinero (cts): y_i	3	6	9	12	15	18

La función y = 3x da la relación entre las dos variables.

b) Correlación directa: al aumentar una variable, aumenta la otra.

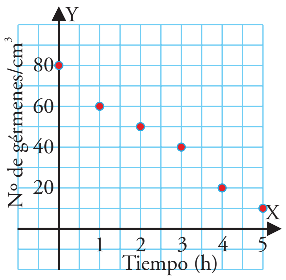


Ejemplo

El número de pedidos que sirve un almacén y el número de vendedores que tiene contratados dicho almacén son:

Nº de vendedores: x_i	1	3	4	5	6	8	9
Nº de pedidos: y_i	60	80	100	140	160	180	200

c) Correlación inversa: al aumentar una variable, la otra disminuye.



Ejemplo

El número de gérmenes por cm^3 y el tiempo transcurrido con un tratamiento específico son:

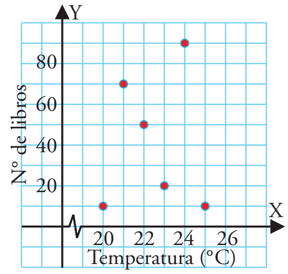
Tiempo (h): x_i	0	1	2	3	4	5
Nº de gérmenes: y_i	80	60	50	40	20	10

d) Correlación nula: no existe relación entre las variables.

Ejemplo

El número de libros vendidos en una librería y la temperatura del día es:

Temperatura (°C): x_i	20	21	22	23	24	25
Nº de libros: y_i	10	70	50	20	90	10



3.2. Coeficiente de correlación de Pearson

El **coeficiente de correlación de Pearson** es:

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

Propiedades del coeficiente de correlación

a) El coeficiente de correlación de Pearson es un número. No depende de las unidades en las que están expresadas las variables **x** e **y**

b) Está comprendido entre – 1 y 1

- Si r = – 1 o r = 1, la **correlación es perfecta** o **funcional.**
- Si **r** está próximo a – 1 o a 1, la **correlación es fuerte.**
- Si **r** está próximo a cero, la **correlación es débil.**
- El signo, $r > 0$ o $r < 0$, indica si la correlación es directa o inversa, respectivamente.

Ejemplo

Calcula el coeficiente de correlación entre el número de pedidos que sirve un almacén y el número de vendedores que tiene contratados dicho almacén.

Nº de vendedores: x_i	2	4	5	6	7	9	10
Nº de pedidos: y_i	70	90	110	150	170	190	210

El coeficiente de correlación es: $r = \frac{s_{xy}}{s_x s_y} = \frac{124,08}{2,59 \cdot 48,82} = 0,98$

La correlación es fuerte y directa.

Interpretación

El coeficiente de correlación de Pearson indica la correlación que existe entre las dos variables; es decir, si los puntos están muy próximos o alejados del centro de gravedad.

Correlación fuerte

Se considera que la correlación es fuerte si $|r| > 0,85$

Calculadora

Se introducen los datos y se pulsa:

SHIFT S-VAR ▶ ▶ (r) 3 = **0,98**

Aplica la teoría

7. Calcula el coeficiente de correlación e indica el tipo de correlación para la siguiente distribución bidimensional:

x_i	1	4	4	2	5	3	1
y_i	5	2	3	6	3	2	4

8. La temperatura media en los meses de invierno en varias ciudades y el gasto medio por habitante en calefacción han sido:

Temperatura (°C)	10	12	14	15	17	20
Gasto (€)	150	120	102	90	50	18

Calcula el coeficiente de correlación e interpreta el resultado.

9. Calcula el coeficiente de correlación e indica el tipo de correlación para la siguiente distribución bidimensional:

Y \ X	1	2	3	4
1	1	2	0	0
2	2	1	0	0
3	0	1	2	3
4	0	4	3	1