

11 Distribucións estatísticas dobres

ÍNDICE DE CONTIDOS

1. VARIABLES ESTADÍSTICAS DOBRES.	2
2. DIAGRAMA DE DISPERSIÓN OU NUBE DE PUNTOS.	3
3. AXUSTE DA NUBE DE PUNTOS POR UNHA RECTA. RECTA DE REGRESIÓN.	5
4. CONCEPTO DE CORRELACIÓN.	9
4.1. Covarianza.	9
4.2. Coeficiente de correlación.	10

En moitos campos do coñecemento xorde a necesidade de establecer relación entre dous conxuntos de datos, ou dúas variables estatísticas, aínda sabendo que tal relación non pode ser funcional, é dicir, que non existe unha fórmula que permita obter os datos dun dos conxuntos, ou dunha das variables, a partir dos do outro, ou da outra variable.

Hai dous problemas fundamentais no estudo das relacións entre dúas variables estatísticas. O primeiro consiste en considerar unha das variables, a mellor coñecida, como variable independente e atopar unha función, no noso caso só falaremos da función lineal, que ilustre de modo aproximado a relación entre as dúas variables e permita facer predicións para algúns datos descoñecidos. Este problema coñécese como Análise da Regresión ou simplemente axuste dos datos pola recta de regresión. O segundo dos problemas conduce ao cálculo do coeficiente de correlación lineal que mide o grao de interdependencia lineal entre dúas variables estatísticas, cando os datos de ambas as dúas teñen a mesma fiabilidade e non ten moito sentido tomar unha das variables como variable independente.

O propósito desta Sección é, en primeiro lugar, atopar a recta de regresión entre dúas variables estatísticas e a continuación, mediante o emprego do coeficiente de correlación, descubrir se o grao de relación entre as variables é o suficientemente grande como para que a recta de regresión teña algunha utilidade.

1. Variables estatísticas dobres

Nunha poboación estudaremos dúas variables estatísticas: unha variable que denominamos X e outra que denominamos Y , de modo que cada individuo da poboación estará determinado por un par de datos (x_i, y_j) , no que x_i representa os valores ou marcas de clase da variable X e y_j representa os valores ou marcas de clase da variable Y .

Ao estudo conxunto de dúas características ou variables estatísticas unidimensionais X e Y sobre unha mesma poboación acostúmase chamarlle **variable estatística bidimensional**.

Por exemplo, nunha avaliación de 30 alumnos rexistrouse o número de suspensos e o número de horas diarias que dedica cada un ao estudo, obténdose os seguintes resultados:

(0, 2) (2, 2) (5, 0) (2, 1) (1, 2) (1, 3) (0, 4) (4, 0) (2, 2) (2, 1) (1, 2) (0, 4) (1, 3)
(4, 2) (1, 2) (2, 1) (1, 2) (0, 2) (0, 3) (2, 3) (2, 2) (2, 2) (1, 2) (6, 0) (3, 1) (2, 2)
(1, 2) (3, 1) (4, 1) (1, 2)

Estamos ante dúas variables. A variable X , a máis fiable, conta o número de suspensos e serve para explicar a variable Y , as horas diarias de estudo. O par (x_i, y_j) , rexistra o número de suspensos, x_i , e o número de horas de estudo, y_j .

Os datos dunha variable estatística bidimensional distribúense en táboas de frecuencias de dobre entrada, así:

$X \backslash Y$	0	1	2	3	4	Totais
0	0	0	2	1	2	5
1	0	0	7	2	0	9
2	0	3	5	1	0	9
3	0	2	0	0	0	2
4	1	1	1	0	0	3
5	1	0	0	0	0	1
6	1	0	0	0	0	1
Totais	3	6	15	4	2	30

Na primeira columna da táboa puxemos os valores da variable X e na primeira fila os valores da variable Y , e en cada casa figura a frecuencia absoluta f_{ij} do par (x_i, y_j) . A última fila e a última columna presentan as chamadas **distribucións marxinais**. Na última fila figuran as frecuencias da variable Y e na última columna as frecuencias da variable X . As distribucións de frecuencias bidimensionais reflíctense en táboas de dobre entrada, que no caso xeral sería así:

$X \backslash Y$	y_1	y_2	\vdots	y_m	FRECUENCIAS VARIABLE X
x_1	f_{11}	f_{12}	\vdots	f_{1m}	$\sum f_{1i}$
x_2	f_{21}	f_{22}	\vdots	f_{2m}	$\sum f_{2i}$
\dots	\dots	\dots	\vdots	\dots	\dots
x_n	f_{n1}	f_{n2}	\vdots	f_{nm}	$\sum f_{ni}$
FRECUENCIAS VARIABLE Y	$\sum f_{j1}$	$\sum f_{j2}$	\vdots	$\sum f_{jm}$	N

Non obstante, cando o número de datos ou observacións é pequeno, en vez de táboas de dobre entrada, empregaremos táboas simples de dúas filas, de modo que en cada columna figuren os valores, (x_i, y_j) , correspondentes ás dúas variables. No sucesivo só empregaremos táboas de dúas filas (ou de dúas columnas, se as táboas as poñemos de pé).

Por exemplo, as cualificacións de 12 alumnos en Matemáticas e Lingua son as seguintes:

(2, 2), (4, 7), (4, 4), (6, 2), (4, 5), (6, 5), (3, 6), (6, 4), (5, 8), (7, 1), (3, 7), (7, 6).

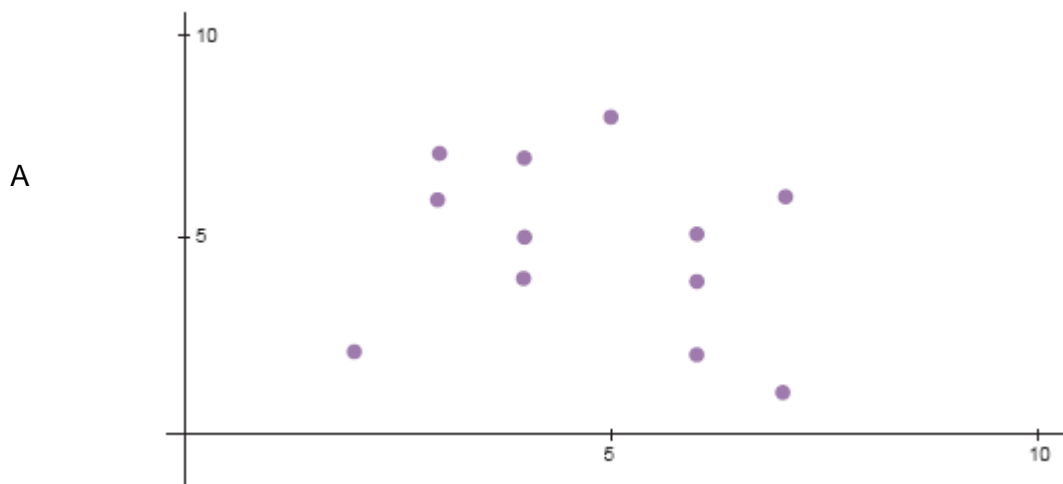
Estes datos dispóñense nunha táboa simple de dúas filas así:

Matemáticas	2	4	4	6	4	6	3	6	5	7	3	7
Lengua	2	7	4	2	5	5	6	4	8	1	7	6

2. Diagrama de dispersión ou nube de puntos

Cando as variables X e Y dunha distribución bidimensional son cuantitativas podemos representar os datos por puntos sobre uns eixes de coordenadas. No eixe de abscisas levamos os valores da variable X , que consideramos como variable independente, e sobre o eixe de ordenadas levamos os valores da variable Y , que consideramos como dependente. Debe quedar claro que as dúas variables non xogan o mesmo papel, a que denominamos independente é a que permite explicar o comportamento da outra, a denominada variable Y .

No caso das notas de Matemáticas e Lingua de 12 alumnos, do apartado anterior, se levamos as cualificacións de Matemáticas sobre o eixe de abscisas e as de Lingua sobre o eixe de ordenadas obtemos o seguinte gráfico:



La representación gráfica dunha distribución bidimensional denomínase **diagrama de dispersión ou nube de puntos**. Cada punto ten por coordenadas os valores que en cada individuo teñen as variables X e Y. A nube de puntos permítenos apreciar se existe unha posible relación entre as variables.

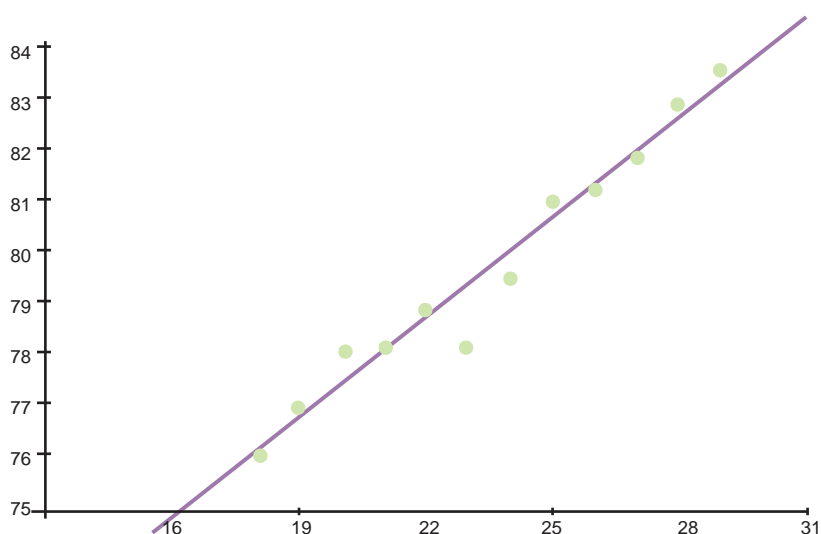
No diagrama anterior non parece que exista ningunha relación entre as dúas variables, pero isto non sempre é así. Vexamos outros exemplos.

Exemplo

Un pediatra anotou as idades, en meses, e a altura en cm de 12 nenos obtendo os seguintes resultados:

meses	18	19	20	21	22	23	24	25	26	27	28	29
altura	76,1	77	78,1	78,2	78,8	78,2	79,5	81	81,2	81,8	82,8	83,5

A nube de puntos desta distribución sería:

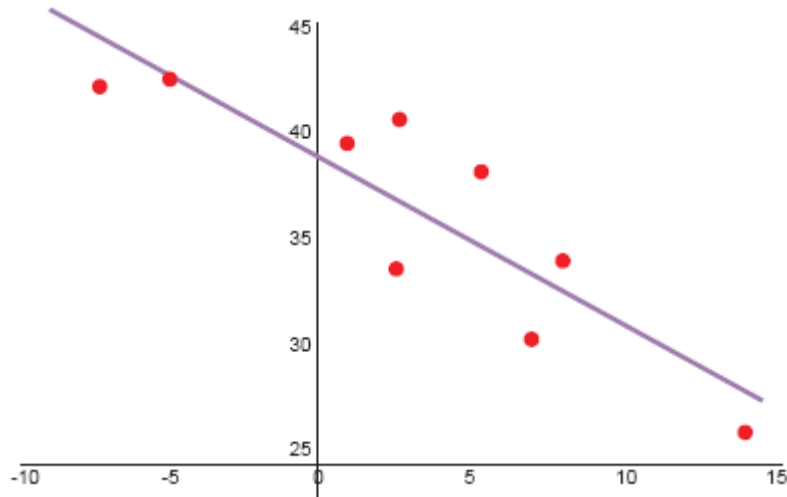


Debuxamos unha recta entre os puntos porque todo suxire que a relación entre as variables idades e alturas aproxímase a unha relación lineal.

A táboa seguinte indica a media das temperaturas mínimas no mes de xaneiro e as latitudes dalgunhas cidades de Estados Unidos

	Temperatura	Latitud
Os Ánxeles	8.3	34.3
San Francisco	5.5	38.4
Washington	1	39.7
Miami	14.4	26.3
Atlanta	2.7	33.9
Chicago	-7.2	42.3
Nova Orleans	7.2	30.8
Nova York	2.7	40.8
Boston	-5	42.7

A nube de puntos correspondente a esta distribución é a seguinte:



Tamén debuxamos unha recta que suxire a existencia dunha relación lineal, aínda que non tan forte como no caso anterior.

A nube de puntos permite apreciar se hai ou non unha relación entre as dúas variables. O problema que se nos formula agora é o seguinte: **se a nube de puntos suxire unha relación lineal entre as variables, como podemos atopar a recta que mellor se axusta á nube de puntos?** Porque, evidentemente, podemos trazar varias rectas que pasen a través dos puntos do diagrama de dispersión. A resposta a esta pregunta verémola en apartado seguinte.

3. Axuste da nube de puntos por unha recta. Recta de regresión

Pretendemos atopar unha recta $e = ax + b$ que estea o máis próxima posible aos puntos da nube. Podiamos calcular a pendente, a , e a ordenada na orixe, b , de modo que a suma das distancias dos puntos á recta sexa mínima, pero iso obrigáranos a empregar a función valor absoluto e é un pouco incómoda.

Determinaremos **a** e **b** impoñendo como condición que a suma dos cadrados das distancias dos puntos á recta sexa mínima:

$$\sum_{i=1}^n d_i^2 = \sum_{i=1}^n [y_i - (ax_i + b)]^2$$

Para calcular o mínimo desta función hai que realizar uns calculos que non estan entre os obxectivos deste curso, polo que pasaremos deles dando a solución directamente. As solucións do sistema veñen dadas por:

$$a = \frac{\frac{\sum x_i y_i}{n} - \bar{x} \bar{y}}{s_x^2} \quad b = \frac{\sum y_i}{n} - a \frac{\sum x_i}{n} = \bar{y} - a \bar{x}$$

e polo tanto a recta de regresión pasa polo (\bar{x}, \bar{y}) punto chamado **centro de gravidade** da nube de puntos.

Sabendo que a recta que buscamos pasa polo punto (\bar{x}, \bar{y}) e ten como pendente a entón a súa ecuación é:

$$y - \bar{y} = \frac{\frac{\sum x_i y_i}{n} - \bar{x} \bar{y}}{s_x^2} \cdot (x - \bar{x})$$

Á recta que mellor se axusta á nube de puntos chámase **recta de regresión**. Veremos agora, nos exemplos, que os ingredientes da recta de regresión son moi doados de atopar cunha calculadora científica sinxela.

Exemplos

1. Atopar a ecuación da recta de regresión correspondente á táboa das idades e alturas de 12 nenos rexistrados por un pediatra

meses	18	19	20	21	22	23	24	25	26	27	28	29
altura	76,1	77	78,1	78,2	78,8	78,2	79,5	81	81,2	81,8	82,8	83,5

Solución. Temos que atopar os elementos da ecuación:

$$y - \bar{y} = \frac{\frac{\sum x_i y_i}{n} - \bar{x} \bar{y}}{s_x^2} \cdot (x - \bar{x})$$

1. Coas teclas MODE • poñemos a calculadora en modo estatístico, na pantalla aparece SD, e xa están activas as teclas escritas en azul. Borramos os datos da memoria coas teclas SHIFT SAC e introducimos os datos da variable X:

18 DATA 19 DATA 20 DATA... 29 DATA

Unha vez introducidos os datos, coas teclas SHIFT \bar{x} e as teclas SHIFT σ_n

obtemos $\bar{x} = 23,5$ y $s_x = 3,452$ que elevando ao cadrado resulta $s_x^2 = 11,91$

2º. Despois de borrar a memoria, introducimos os valores de Y

76.1 DATA 77 DATA 78.1 DATA .. 83.5 DATA

e coas teclas SHIFT \bar{y} atopamos $\bar{y} = 79,683$

3º. Por último, despois de borrar a memoria, introducimos $\sum x_i y_i$

18x76.1 DATA 19x77 DATA 20x78.1 DATA... 29x83.5 DATA

E coas teclas SHIFT $\sum x_i$ obtemos que, $\sum x_i y_i = 22561,9$

4º. Escribimos a recta de regresión

$$y - 79,683 = \frac{\frac{22561,9}{12} - 23,5 \cdot 79,683}{11,916} \cdot (x - 23,5)$$

Facendo operacións $y - 79,683 = 0,638 \cdot (x - 23,5)$

$$y = 0,638x + 64,679$$

Hai calculadoras científicas, máis completas, que dan directamente a pendente e a ordenada na orixe da recta de regresión.

2. Atopar a ecuación da recta de regresión correspondente á distribución das temperaturas mínimas medias no mes de xaneiro e as latitudes de varias cidades de Estados Unidos

	Temperatura	Latitude
Os Ánxeles	8.3	34.3
San Francisco	5.5	38.4
Washington	1	39.7
Miami	14.4	26.3
Atlanta	2.7	33.9
Chicago	-7.2	42.3
Nova Orleans	7.2	30.8
Nova York	2.7	40.8
Boston	-5	42.7

Solución: Temos que atopar os elementos da ecuación:

$$y - \bar{y} = \frac{\frac{\sum x_i y_i}{n} - \bar{x} \bar{y}}{s_x^2} \cdot (x - \bar{x})$$

1º. Introducimos os datos da variable X:

8.3 DATA 5.5 DATA 1 DATA ... - 5 DATA

Unha vez introducidos os datos, coas teclas SHIFT \bar{x} e as teclas SHIFT σ_n

obtemos $\bar{x} = 3,288$ e $s_x = 6,266$ que elevando ao cadrado resulta $s_x^2 = 39,267$
2º. Despois de borrar a memoria, introducimos os valores de Y

34.3 DATA 38.4 DATA 39.7 DATA ... 42.7 DATA

E co as teclas \bar{y} SHIFT atopamos $\bar{y} = 36,577$

3º. Por último, despois de borrar a memoria, introducimos $\sum x_i y_i$
8.3 × 34.3 DATA 5.5 × 38.4 DATA 1 × 39.7 DATA ... 5 × 42.7 DATA

e coas teclas SHIFT Σ x obtemos que $\sum x_i y_i = 819,7$

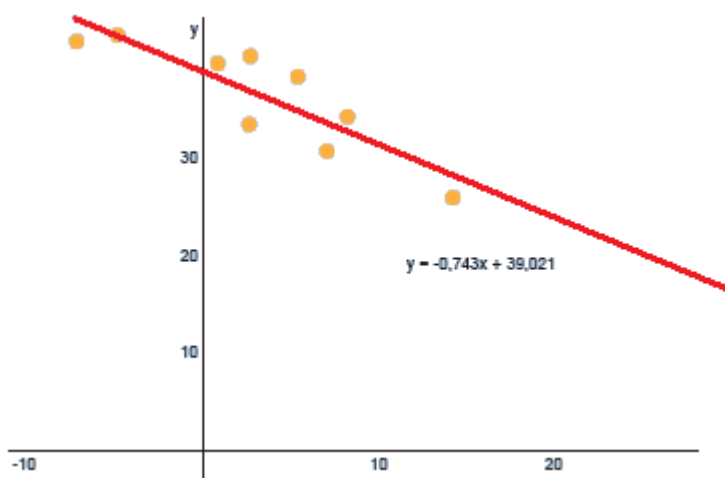
4º. Escribimos a recta de regresión

$$y - 36,577 = \frac{\frac{819,7}{9} - 3,288 \cdot 36,577}{39,267} \cdot (x - 3,288)$$

Facendo operacións, resulta a recta de regresión

$$y = -0,743x + 39,021$$

Graficamente sería a recta da figura:



A principal utilidade da recta de regresión é facer predicións. Se quixésemos saber cal é a latitude dunha cidade de Estados Unidos cuxa media das temperaturas mínimas no mes de xaneiro é 4,5º C, substituímos x por 4,5 e obtemos unha estimación da latitude:

$$y = -0,743 \cdot 4,5 + 39,021 = 35,677$$

A cidade tería 35,677 graos de latitude norte. Queda un problema por resolver: que fiabilidade proporciona a recta de regresión para facer estimacións? Iso saberémolo coñecendo o coeficiente de correlación lineal das dúas variables que estudamos no próximo apartado.

4. Concepto de correlación

O grao de dependencia lineal entre dúas variables mídese co **coeficiente de correlación** lineal, e cando a dependencia lineal é débil a recta de regresión carece de interese.

4.1. Covarianza

En primeiro lugar queremos descubrir se a relación entre dúas variables é directa, é dicir, cando ao aumentar a variable independente aumenta tamén a variable dependente, ou se é inversa, que acontece cando ao aumentar a variable X diminúe a variable Y.

A **covarianza** é un parámetro que mide este tipo de relación e está definida como a media aritmética dos produtos da desviacións de cada un dos valores das variables respecto ás súas medias, simbolízase por S_{xy} e vén dada por:

$$S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

A covarianza ten unha formulación máis coñecida se realizamos as operacións indicadas e quedaría:

$$S_{xy} = \frac{\sum x_i y_i}{n} - \bar{x} \bar{y}$$

A covarianza resulta ser o numerador da pendente da recta de regresión.

4.2. Coeficiente de correlación

A medida precisa da relación de dúas variables estatísticas proporciona o coeficiente de correlación lineal, representado pola letra r , e que está definido pola expresión seguinte:

$$r = \frac{S_{xy}}{S_x S_y}$$

É dicir, é o cociente entre a covarianza e o produto das desviacións típicas de X e Y. Como a desviación típica dunha variable estatística é sempre positiva, o signo do coeficiente de correlación depende do signo da covarianza, e podemos afirmar:

Covarianza positiva indica correlación directa.

Covarianza negativa indica correlación inversa.

Covarianza nula indica que non hai correlación entre as variables.

Pódese demostrar que o coeficiente de correlación é un número comprendido entre -1 e 1 , e, en consecuencia, pódense dar as seguintes situacións:

Que $r = 1$, entón a relación entre as variables é funcional directa e a nube de puntos está sobre unha recta de pendente positiva.

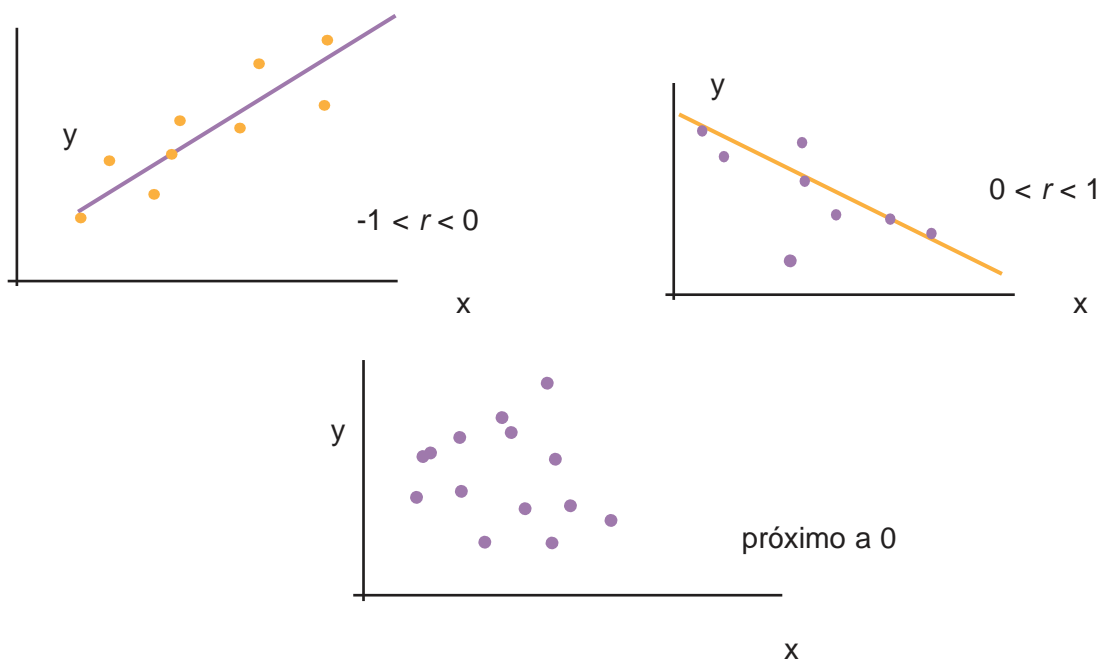
Que $0 < r < 1$, entón hai unha correlación directa entre as variables. Correlación forte cando r está próximo a 1 e débil cando r se aproxima a 0 .

Que $r = 0$, entón non existe ningún tipo de relación ou dependencia entre as variables.

Que $-1 < r < 0$, entón hai correlación inversa entre as variables. Correlación forte cando r está próximo a -1 e débil cando r está próxima a 0 .

Que $r = -1$, entón a relación entre as variables é funcional inversa e a nube de puntos está sobre unha recta de pendente negativa.

Nas figuras ilustramos algunhas destas situacións:



Resumindo: a recta de regresión permite facer previsións ou estimacións, pero non debemos esquecer que estas estimacións só son fiables cando r toma valores próximos a -1 ou a 1 .

Exemplos

1. Atopar o coeficiente de correlación lineal correspondente á táboa das idades e alturas de 12 nenos rexistrados por un pediatra

	18	19	20	21	22	23	24	25	26	27	28	29
	76,1	77	78,1	78,2	78,8	78,2	79,5	81	81,2	81,8	82,8	83,5

Solución. O coeficiente de correlación lineal vén dado pola fórmula:

$$r = \frac{s_{xy}}{s_x s_y} \quad \text{Donde} \quad s_{xy} = \frac{\sum x_i y_i}{n} - \bar{x} \bar{y} \quad \text{e } s_x \text{ e } s_y \text{ son as desviacións}$$

típicas de X e Y. Ya sabemos, calculámolo no exemplo 1, que $S_{xy} = 7,6191$

Tamen conocemos \bar{x} que $= 23,5$ e $S_x = 3,45$

Introducimos de novo os valores de Y

76.1 DATA 77 DATA 78.1 DATA... 83.5 DATA

e coas teclas \bar{x} SHIFT e as teclas SHIFT σ_n atopamos $\bar{y} = 79,68$ $S_y = 2,24$.

Logo:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{7,61}{3,45 \cdot 2,24} = 0,98$$

O que indica un alto grao de correlación e as previsións que se fagan coa recta de regresión son altamente fiables.

2. Calcular o coeficiente de correlación lineal das cualificacións de 12 alumnos en Matemáticas e Lingua:

Matemáticas	2	4	4	6	4	6	3	6	5	7	3	7
Lingua	2	7	4	2	5	5	6	4	8	1	7	6

Solución:

1º Despois de borrar a memoria, introducimos os datos da variable Matemáticas, que chamaremos X,

2 DATA 4 DATA 4 DATA... 7 DATA

Coas teclas \bar{x} coñecidas obtemos $\bar{x} = 4,75$ e $S_x = 1,68$.

2º Borramos a memoria e introducimos os datos de Lingua, variable Y

2 DATA 7 DATA 4 DATA... 6 DATA

E atopamos $\bar{y} = 4,75$ e $S_y = 2,12$

3º Por último, introducimos

2 DATA 4 × 7 DATA 4 × 4 DATA... 7 × 6 DATA

coas teclas SHIFT Σ x obtemos que $\Sigma xy = 262$

Agora,

$$s_{xy} = \frac{262}{12} - 4,75 \cdot 4,75 = -0,7291$$

entón,

$$r = \frac{s_{xy}}{s_x s_y} = \frac{-0,7291}{1,68 \cdot 2,12} = -0,20$$

O que indica unha correlación negativa pero moi débil.