ESTADÍSTICA BIDIMENSIONAL

1º Bachillerato CC.SS.

VARIABLES BIDIMENSIONALES.

Supongamos que en nuestra clase de 16 alumnos se dan los siguientes valores de peso y estatura: (60, 167), (65, 170), (65, 170), (65,170), (70, 180), (70,180), (70,180), (70,180), (70,170), (70,170), (65, 170), (65,170), (68, 170), (68,170), (50, 155), (60,160)

Peso	Estatura	f_i
60	167	1
65	170	5
70	170	2
70	180	4
68	170	2
50	155	1
60	160	1
		16

VARIABLES BIDIMENSIONALES.

Supongamos que en nuestra clase de 12 alumnos se dan las siguientes notas en matemáticas y en física:

Matemáticas	Física	f_i
2	1	1
3	3	1
4	2	1
4	4	1
5	4	1
6	4	1
6	6	1
7	4	1
7	6	1
8	7	1
10	9	1
10	10	1
		12

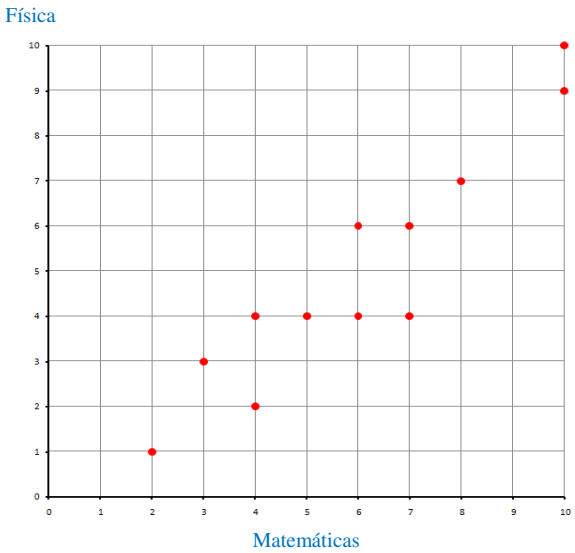
TABLAS DE CONTINGENCIA.

Cuando hay muchos datos, la tabla se suele mostrar de la siguiente manera:

		Matemáticas										
	X \ Y	1	2	3	4	5	6	7	8	9	10	**Tot**
	1	7	8	5	4	9	7	6	4	3	2	**55**
	2	12	3	9	8	5	7	9	8	5	3	**69**
	3	5	4	7	8	6	2	5	6	4	7	**54**
	4	6	3	2	5	8	9	7	9	9	8	**66**
Física	5	4	7	8	5	2	5	4	7	8	9	**59**
	6	6	5	4	1	2	3	5	4	7	9	**46**
	7	6	5	4	7	8	9	5	8	7	9	**68**
	8	3	6	5	4	7	8	5	4	7	9	**58**
	9	6	5	4	7	8	5	6	9	8	7	**65**
	10	9	8	9	8	7	4	5	2	4	5	**60**
	Tot	**64**	**54**	**57**	**57**	**62**	**59**	**57**	**61**	**62**	**67**	**600**

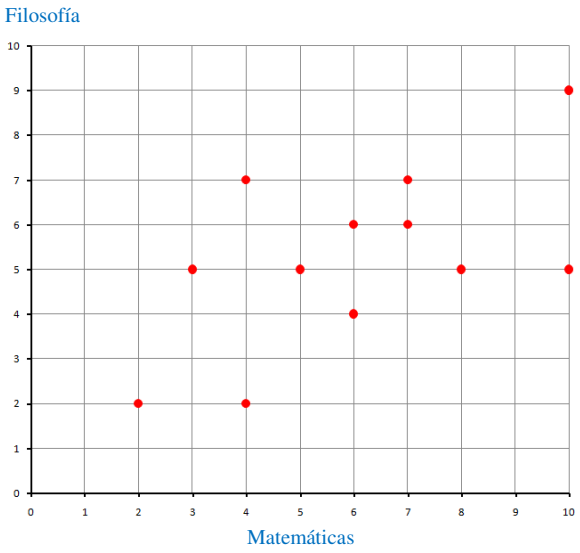
CORRELACIÓN. NUBE DE PUNTOS.

Matemáticas	Física
2	1
3	3
4	2
4	4
5	4
6	4
6	6
7	4
7	6
8	7
10	9
10	10



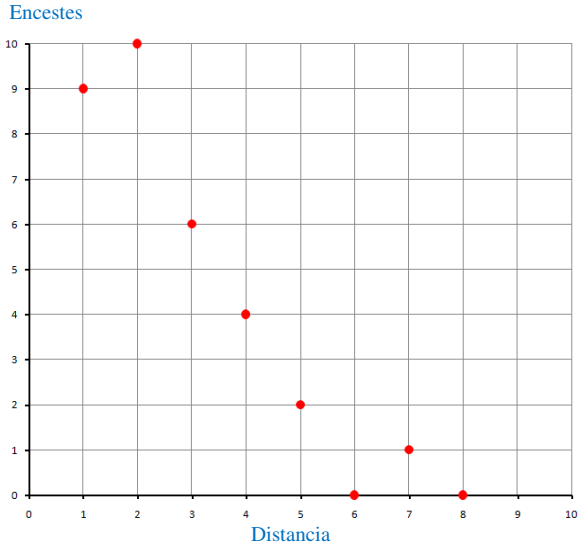
CORRELACIÓN. NUBE DE PUNTOS.

Matemáticas	Filosofía
2	2
3	5
4	2
4	7
5	5
6	4
6	6
7	6
7	7
8	5
10	5
10	9



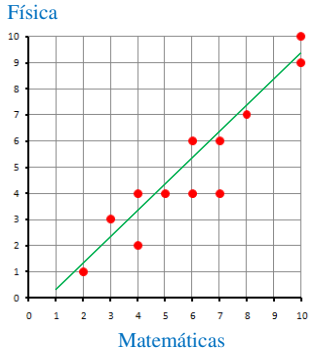
CORRELACIÓN. NUBE DE PUNTOS.

Distancia	Encestes
1	9
2	10
3	6
4	4
5	2
6	0
7	1
8	0

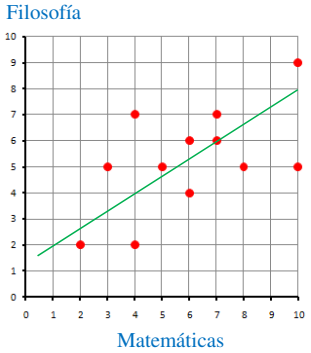


CORRELACIÓN. RECTA DE REGRESIÓN.

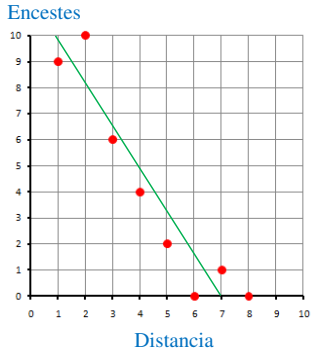
Correlación positiva



Correlación positiva débil



Correlación negativa fuerte



DISTRIBUCIONES CONJUNTA Y MARGINALES.

Matem	Física
2	1
3	3
4	2
4	4
5	4
6	4
6	6
7	4
7	6
8	7
10	9
10	10
72	**60**

Mat: x_i	f_i	$f_i \cdot x_i$	$f_i \cdot x_i^2$
2	1	2	4
3	1	3	9
4	2	8	32
5	1	5	25
6	2	12	72
7	2	14	98
8	1	8	64
10	2	20	200
	12	**72**	**504**

Fis: y_i	f_i	$f_i \cdot y_i$	$f_i \cdot y_i^2$
1	1	1	1
2	1	2	4
3	1	3	9
4	4	16	64
6	2	12	72
7	1	7	49
9	1	9	81
10	1	10	100
	12	**60**	**380**

$$\bar{x} = \frac{\sum f_i \cdot x_i}{n} = \frac{72}{12} = 6$$

$$\sigma_x^2 = \frac{\sum f_i \cdot x_i^2}{n} - \bar{x}^2 = \frac{504}{12} - 6^2 = 6$$

$$\bar{y} = \frac{\sum f_i \cdot y_i}{n} = \frac{60}{12} = 5$$

$$\sigma_y^2 = \frac{\sum f_i \cdot y_i^2}{n} - \bar{y}^2 = \frac{380}{12} - 5^2 = 6'66$$

DISTRIBUCIONES CONJUNTA Y MARGINALES.

Cuando hay muchos datos, la tabla se suele mostrar de la siguiente manera:

Hijos

Mascotas \ Y	1	2	3	4	**Mar X**	$f_i \cdot x_i$	$f_i \cdot x_i^2$
1	7	8	5	4	**24**	24	24
2	12	3	9	8	**32**	64	128
3	5	4	7	8	**24**	72	216
4	6	3	2	5	**16**	64	256
Mar Y	**30**	**18**	**23**	**25**	**96**	**224**	**624**
$f_i \cdot y_i$	30	36	69	100	**235**		
$f_i \cdot y_i^2$	30	72	207	400	**709**		

$$\bar{x} = \frac{\sum f_i \cdot x_i}{n} = \frac{224}{96} = 2'33$$

$$\bar{y} = \frac{\sum f_i \cdot y_i}{n} = \frac{235}{96} = 2'45$$

$$\sigma_x^2 = \frac{\sum f_i \cdot x_i^2}{n} - \bar{x}^2 = \frac{624}{96} - (2'33)^2 = 1'07$$

$$\sigma_y^2 = \frac{\sum f_i \cdot y_i^2}{n} - \bar{y}^2 = \frac{709}{96} - (2'45)^2 = 1'38$$

DISTRIBUCIONES CONDICIONADAS.

Cuando se fija un valor de x se puede determinar la distribución de la otra variable:

Hijos

Mascotas \ Y	1	2	3	4	**Mar X**
1	7	8	5	4	**24**
2	12	3	9	8	**32**
3	5	4	7	8	**24**
4	6	3	2	5	**16**
Mar Y	**30**	**18**	**23**	**25**	**96**

Para x_2 la distribución Y queda:

$Y \mid x_2$	1	2	3	4	
f_i	12	3	9	8	**32**
h_i	12/32	3/32	9/32	8/32	**1**

Para y_3 la distribución X queda:

$X \mid y_3$	f_i	h_i
1	5	5/23
2	9	9/23
3	7	7/23
4	2	2/23
	23	**1**

DEPENDENCIA E INDEPENDENCIA.

Para saber si dos variables son independientes se calcula la tabla con h_{ij}:

f_{ij}

Hijos

Mascotas \ Y	1	2	3	4	f_i
1	7	8	5	4	**24**
2	12	3	9	8	**32**
3	5	4	7	8	**24**
4	6	3	2	5	**16**
f_j	**30**	**18**	**23**	**25**	**96**

h_{ij}

Hijos

Mascotas \ Y	1	2	3	4	h_i
1	7/96	8/96	5/96	4/96	**24/96**
2	12/96	3/96	9/96	8/96	**32/96**
3	5/96	4/96	7/96	8/96	**24/96**
4	6/96	3/96	2/96	5/96	**16/96**
h_j	**30/96**	**18/96**	**23/96**	**25/96**	**1**

Son independientes si se cumple que $h_{ij} = h_i \cdot h_j$

$$\frac{4}{96} \neq \frac{25}{96} \cdot \frac{24}{96} \rightarrow \text{Son dependientes}$$

DEPENDENCIA E INDEPENDENCIA.

Ejemplo:

f_{ij}

X \ Y	2	4	6	8	f_i
1	3	12	6	15	**36**
3	5	20	10	25	**60**
f_j	**8**	**32**	**16**	**40**	**96**

h_{ij}

X \ Y	2	4	6	8	h_i
1	0,0312	0,125	0,0625	0,1563	**0,375**
3	0,0521	0,2083	0,1042	0,2604	**0,625**
h_j	**0,0833**	**0,3333**	**0,1667**	**0,4167**	**1**

Son independientes si se cumple que $h_{ij} = h_i \cdot h_j$. Comprobamos:

$0{,}0833 \cdot 0{,}375 = 0{,}0312...$ $\quad$ $0{,}3333 \cdot 0{,}625 = 0{,}2083...$

Son independientes

MEDIDA DE LA CORRELACIÓN

Centro de Gravedad:

El centro de gravedad de una distribución es $(\overline{x}\ ,\ \overline{y})$

Covarianza:

$$\sigma_{xy} = \frac{\sum (x_i - \overline{x}) \cdot (y_i - \overline{y})}{n} = \frac{\sum x_i \cdot y_i}{n} - \overline{x} \cdot \overline{y}$$

Coeficiente de determinación:

$$R^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \cdot \sigma_y^2}$$

Coeficiente de correlación:

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

- El valor de r está comprendido entre -1 y 1
- Si la correlación es perfecta r vale 1 o -1
- Si la correlación es fuerte r está cerca de 1 o -1
- Si la correlación es débil r está cerca de 0

x_i	y_i
x_1	y_1
x_2	y_2
x_3	y_3
…	…
x_n	y_n

MEDIDA DE LA CORRELACIÓN

x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
2	1	4	1	2
3	3	9	9	9
4	2	16	4	8
4	4	16	16	16
5	4	25	16	20
6	4	36	16	24
6	6	36	36	36
7	4	49	16	28
7	6	49	36	42
8	7	64	49	56
10	9	100	81	90
10	10	100	100	100
72	60	504	380	431

$$\overline{x} = \frac{\sum x_i}{n} = \frac{72}{12} = 6 \qquad \overline{y} = \frac{\sum y_i}{n} = \frac{60}{12} = 5$$

El centro de gravedad es (6 , 5)

$$\sigma_x = \sqrt{\frac{\sum x_i^2}{n} - \overline{x}^2} = \sqrt{\frac{504}{12} - 6^2} = \sqrt{6} = 2{,}45$$

$$\sigma_y = \sqrt{\frac{\sum y_i^2}{n} - \overline{y}^2} = \sqrt{\frac{380}{12} - 5^2} = \sqrt{6{,}67} = 2{,}58$$

$$\sigma_{xy} = \frac{\sum x_i \cdot y_i}{n} - \overline{x} \cdot \overline{y} = \frac{431}{12} - 6 \cdot 5 = 5{,}92$$

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{5{,}92}{2{,}45 \cdot 2{,}58} = 0{,}94$$

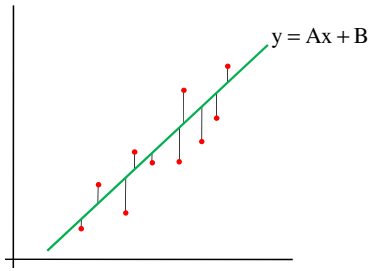
$$R^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \cdot \sigma_y^2} = \frac{5{,}92^2}{2{,}45^2 \cdot 2{,}58^2} = 0{,}8836$$

La correlación es fuerte.

El 88'36% de la variabilidad es explicada

RECTAS DE REGRESIÓN

Recta de regresión de Y sobre X:

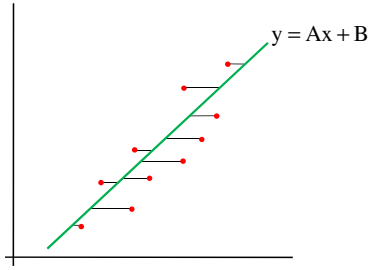


$$y = \overline{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x - \overline{x})$$

$\frac{\sigma_{xy}}{\sigma_x^2}$ Coeficiente de regresión de Y sobre X

RECTAS DE REGRESIÓN

Recta de regresión de X sobre Y:



$$x = \bar{x} + \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})$$

$\frac{\sigma_{xy}}{\sigma_y^2}$ Coeficiente de regresión de X sobre Y

RECTAS DE REGRESIÓN

x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
2	1	4	1	2
3	3	9	9	9
4	2	16	4	8
4	4	16	16	16
5	4	25	16	20
6	4	36	16	24
6	6	36	36	36
7	4	49	16	28
7	6	49	36	42
8	7	64	49	56
10	9	100	81	90
10	10	100	100	100
72	60	504	380	431

$\bar{x} = \frac{72}{12} = 6$ $\quad \sigma_x = \sqrt{\frac{504}{12} - 6^2} = 2,45$ $\quad \sigma_{xy} = \frac{431}{12} - 6 \cdot 5 = 5,92$

$\bar{y} = \frac{60}{12} = 5$ $\quad \sigma_y = \sqrt{\frac{380}{12} - 5^2} = 2,58$ $\quad r = \frac{5,92}{2,45 \cdot 2,58} = 0,94$

Recta de regresión de Y sobre X:

$$y = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x}) \rightarrow y = 5 + \frac{5,92}{6}(x - 6) \rightarrow y = 0,99x - 0,92$$

Recta de regresión de X sobre Y:

$$x = \bar{x} + \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y}) \rightarrow x = 6 + \frac{5,92}{6,67}(y - 5) \rightarrow x = 0,89y + 1,56$$

Si $x = 9 \rightarrow y = 0,99 \cdot 9 - 0,92 = 7,99$

Si $y = 5 \rightarrow x = 0,89 \cdot 5 - 1,56 = 2,89$

Las predicciones son buenas.

RECTAS DE REGRESIÓN. PRECAUCIONES.

Relación de causalidad entre las variables

x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
2	1	4	1	2
3	3	9	9	9
4	2	16	4	8
4	4	16	16	16
5	4	25	16	20
6	4	36	16	24
6	6	36	36	36
7	4	49	16	28
7	6	49	36	42
8	7	64	49	56
10	9	100	81	90
10	10	100	100	100
72	60	504	380	431

$\bar{x} = \frac{72}{12} = 6$ $\quad \sigma_x = \sqrt{\frac{504}{12} - 6^2} = 2,45$ $\quad \sigma_{xy} = \frac{431}{12} - 6 \cdot 5 = 5,92$

$\bar{y} = \frac{60}{12} = 5$ $\quad \sigma_y = \sqrt{\frac{380}{12} - 5^2} = 2,58$ $\quad r = \frac{5,92}{2,45 \cdot 2,58} = 0,94$

En este caso parece que las variables X e Y están fuertemente relacionadas, pero si nos dijeran que X es la temperatura media mensual en Alaska e Y el número de puertas en las casas de 12 familias de Alaska...

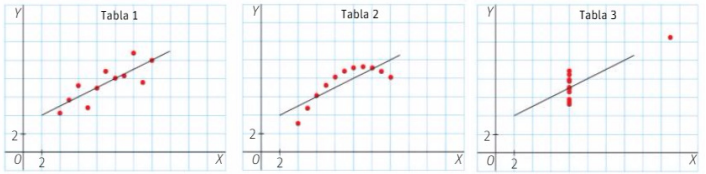
Para ajustar un modelo de regresión no basta con que las dos variables se puedan asociar con un coeficiente de correlación alto. Además, debe existir una **relación lógica de causa-efecto** entre dichas variables.

RECTAS DE REGRESIÓN. PRECAUCIONES.

La importancia de la representación gráfica

1	X	10	8	13	9	11	14	6	4	12	7	5
	Y	8,04	6,95	7,58	8,81	8,33	9,96	7,24	4,26	10,8	4,82	5,68
2	X	10	8	13	9	11	14	6	4	12	7	5
	Y	9,14	8,14	8,74	8,77	9,26	8,1	6,13	3,1	9,13	7,26	4,74
3	X	8	8	8	8	8	8	8	8	8	8	19
	Y	6,58	5,76	7,71	8,84	8,47	7,04	5,25	5,56	7,91	6,89	12,5

Es fácil comprobar que estas tres distribuciones bidimensionales tienen el mismo coeficiente de correlación y la misma recta de regresión de Y sobre X: r = 0'82 y = 3 + 0'5x



Sólo la 1ª se puede representar razonablemente con la recta de regresión.

Antes de llevar a cabo cualquier análisis hay que representar gráficamente los datos.