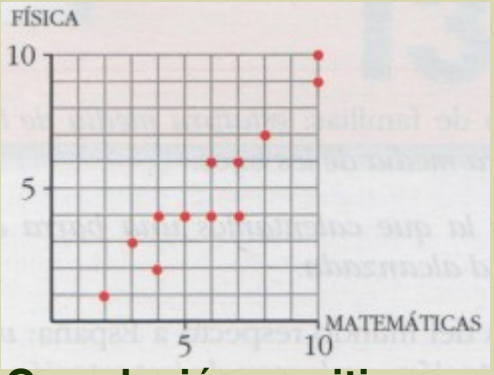
Tema 2. Estadística Bidimensional

1. Nubes de puntos. Correlación
2. Cálculo de la Correlación. Covarianza y Coeficiente de Correlación
3. Datos agrupados en tablas de doble entrada
4. Rectas de regresión. Estimaciones
5. Distribuciones marginales

1. Nubes de puntos. Correlación

Estas son las notas de 12 estudiantes en Matemáticas y en Física:

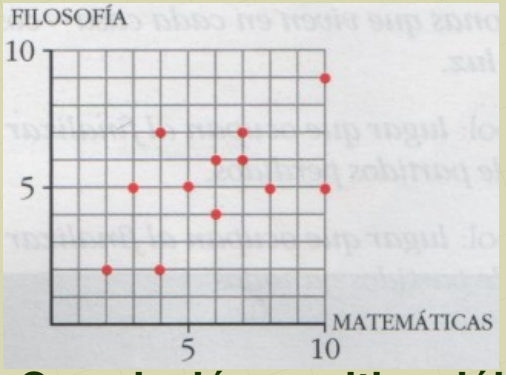
ALUMNO	a	b	c	d	e	f	g	h	i	j	k	l
MATEMÁTICAS	2	3	4	4	5	6	6	7	7	8	10	10
FÍSICA	1	3	2	4	4	4	6	4	6	7	9	10



Correlación positiva

Relacionemos ahora las notas de *Matemáticas* de los mismos alumnos con las de otra asignatura, *Filosofía*.

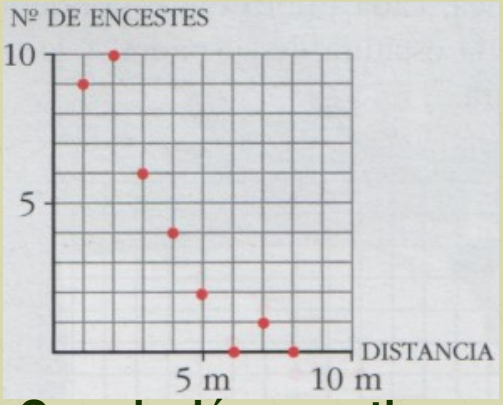
ALUMNO	a	b	c	d	e	f	g	h	i	j	k	l
MATEMÁTICAS	2	3	4	4	5	6	6	7	7	8	10	10
FILOSOFÍA	2	5	2	7	5	4	6	6	7	5	5	9



Correlación positiva débil

Una jugadora de baloncesto lanza a canasta, desde distintas distancias, 10 balones cada vez. Lógicamente, encesta más cuanto más cerca está.

DISTANCIA (m)	1	2	3	4	5	6	7	8
ENCESTES	9	10	6	4	2	0	1	0

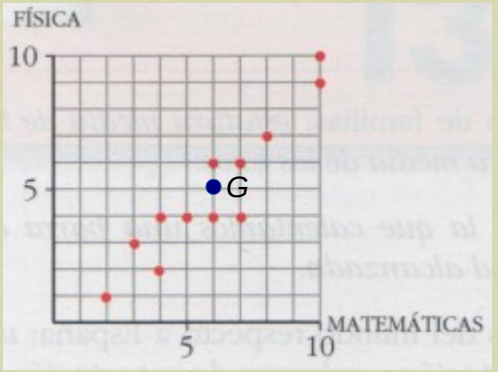


Correlación negativa

2. Cálculo de la correlación

Estas son las notas de 12 estudiantes en Matemáticas y en Física:

ALUMNO	a	b	c	d	e	f	g	h	i	j	k	l
MATEMÁTICAS	2	3	4	4	5	6	6	7	7	8	10	10
FÍSICA	1	3	2	4	4	4	6	4	6	7	9	10



2.1. Centro de gravedad:

- Media de Matemáticas: $\bar{x}=\frac{2+3+4+...+10}{12}=6$
- Media de Física: $\bar{y}=\frac{1+3+2+...+10}{12}=5$

$G(6\ ,\ 5)$

$G(\bar{x}\ ,\ \bar{y})$ Punto que está en el centro de la nube de puntos

2.2. Covarianza:

$$s_{xy}=\frac{2\cdot 1+3\cdot 3+4\cdot 2+...+10\cdot 10}{12}-6\cdot 5=5{,}92$$

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

2.2. Coeficiente de correlación:

$$s_x=\sqrt{\frac{2^2+3^2+4^2+...+10^2}{12}-6^2}=2{,}45$$

$$s_y=\sqrt{\frac{1^2+3^2+2^2+...+10^2}{12}-5^2}=2{,}58$$

$$r=\frac{5{,}92}{2{,}45\cdot 2{,}58}=0{,}94\rightarrow \text{Correlación positiva fuerte}$$

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

El coeficiente de correlación siempre es un número en el intervalo ***[–1 , 1]***.
Cuanto más cerca de -1 o 1, la correlación es más fuerte

3. Datos agrupados en tablas de doble entrada

Si tenemos muchos datos hay que agruparlos en una tabla de frecuencias.

Tenemos 100 alumnos y las notas van de 1 a 5:

Fís y_j \ Mat x_i	1	2	3	4	5	
1	3	2	4	4	3	
2	1	2	3	2	5	
3	1	4	6	4	1	
4	3	6	7	10	9	
5	0	2	5	5	8	
						100

x_1 y_3 f_{13} f_{45}

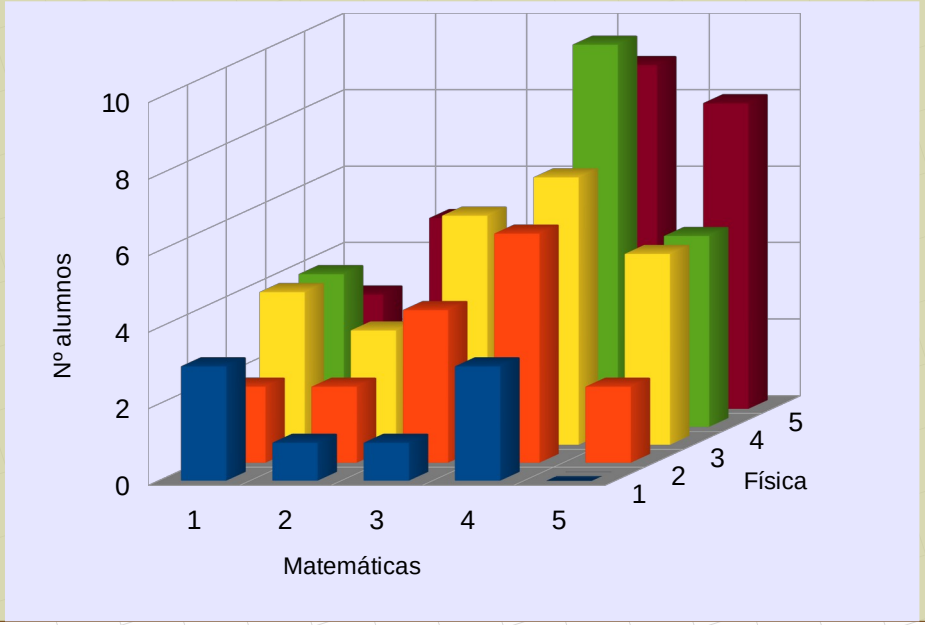
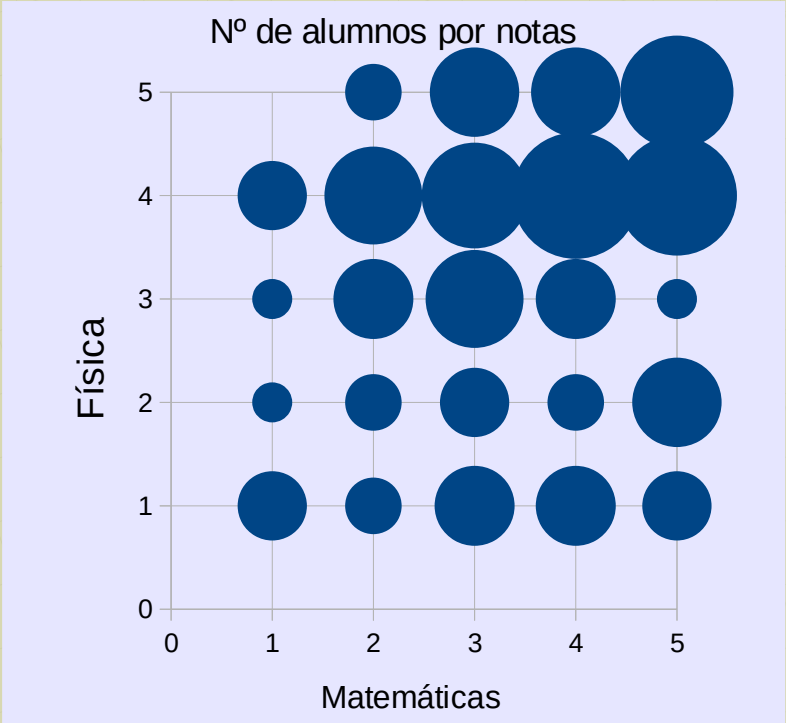


Gráfico de barras. Estereograma



Nube de puntos, diagrama de burbujas

Parece tener una débil correlación positiva

3.1. Cálculo de la correlación

Y \ X	1	2	3	4	5	$f_{\cdot j}$	$y_j \cdot f_{\cdot j}$	$y_j^2 \cdot f_{\cdot j}$	$y_j \cdot (\Sigma x_i \cdot f_{ij})$
1	3	2	4	4	3	16	16	16	50
2	1	2	3	2	5	13	26	52	94
3	1	4	6	4	1	16	48	144	144
4	3	6	7	10	9	35	140	560	484
5	0	2	5	5	8	20	100	500	395
$f_{i\cdot}$	8	16	25	25	26	100	330	1272	1167
$x_i \cdot f_{i\cdot}$	8	32	75	100	130	345			
$x_i^2 \cdot f_{i\cdot}$	8	64	225	400	650	1347			

3.1. Centro de gravedad:

$$\bar{x} = \frac{345}{100} = 3{,}45$$

$$\bar{y} = \frac{330}{100} = 3{,}3$$

$$G(3{,}45\ ,\ 3{,}3)$$

3.2. Desviaciones típicas:

$$s_x = \sqrt{\frac{1347}{100} - 3{,}45^2} = 1{,}25$$

$$s_y = \sqrt{\frac{1272}{100} - 3{,}3^2} = 1{,}35$$

3.3. Covarianza

$$s_{xy} = \frac{1167}{100} - 3{,}45 \cdot 3{,}3 = 0{,}285$$

$$s_{xy} = \frac{\sum x_i \cdot y_j \cdot f_{ij}}{n} - \bar{x} \cdot \bar{y}$$

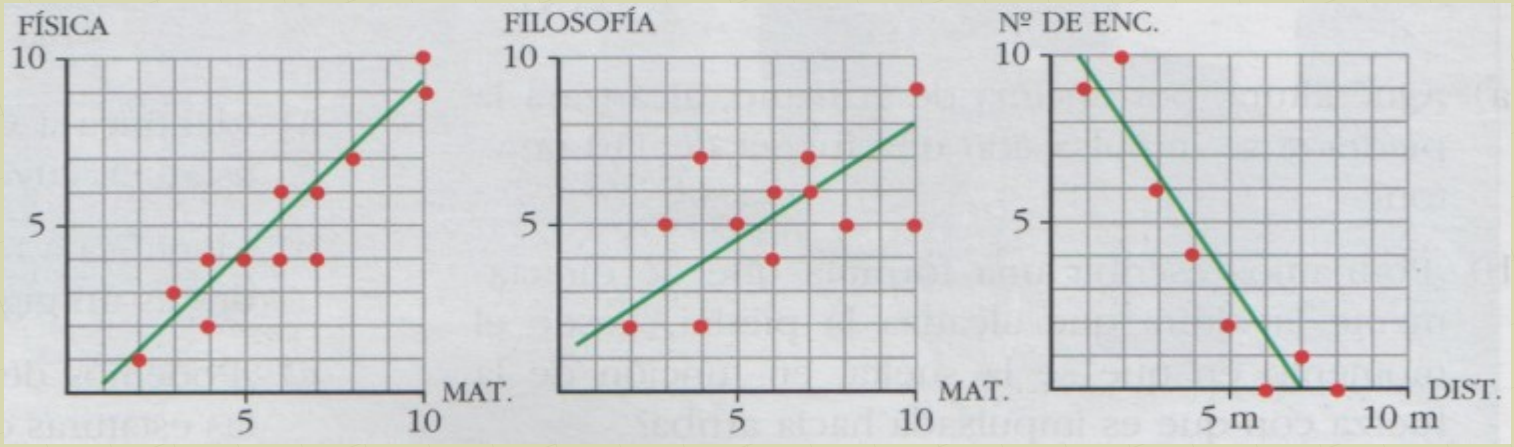
3.4. Coeficiente de correlación

$$r = \frac{0{,}285}{1{,}25 \cdot 1{,}35} = 0{,}17 \rightarrow \text{Correlación positiva muy débil}$$

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

4. Rectas de Regresión

En una nube de puntos podemos trazar una recta que se ajuste lo más posible a todos los puntos a la vez



- Recta de regresión Y sobre X:
  - Pasa por el centro de gravedad: $G(\bar{x}, \bar{y})$
  - Su pendiente es:
    Coef. de regresión de Y sobre X

$$m_{yx} = \frac{s_{xy}}{s_x^2}$$

Ecuación punto – pendiente:

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

$$y - \bar{y} = m_{yx}(x - \bar{x})$$

- Recta de regresión X sobre Y:
  - Pasa por el centro de gravedad: $G(\bar{x}, \bar{y})$

Coef. de regresión de X sobre Y

$$m_{xy} = \frac{s_{xy}}{s_y^2}$$

- Su pendiente es: $m = \frac{s_y^2}{s_{xy}}$

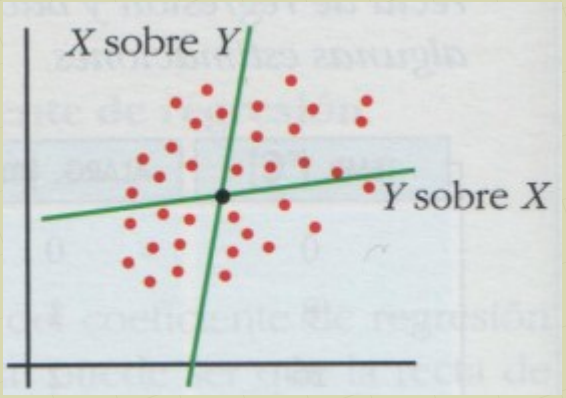
Ecuación punto – pendiente:

$$x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y}) \rightarrow y - \bar{y} = \frac{s_y^2}{s_{xy}}(x - \bar{x})$$

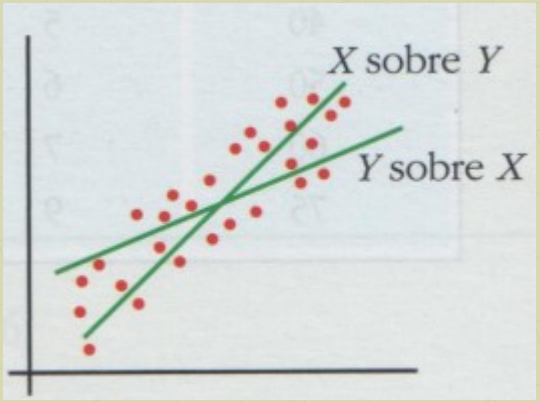
$$x - \bar{x} = m_{xy}(y - \bar{y}) \qquad y - \bar{y} = m(x - \bar{x})$$

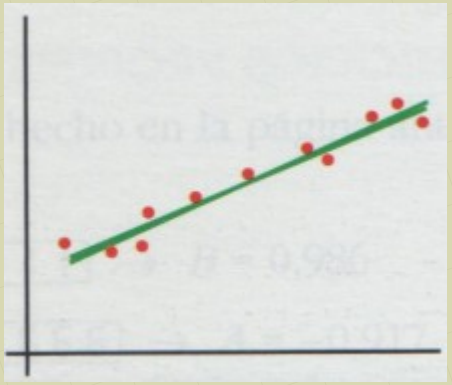
Ejercicio: Calcular la recta de regresión de Y sobre X en el caso de Matemáticas/Física y en el caso Metros/Canastas

- Cuando hay mucha correlación, el ángulo entre las dos rectas es muy pequeño: $\rightarrow 0°$
- Cuando hay poca correlación, el ángulo entre las dos rectas es muy grande: $\rightarrow 90°$



Correlación débil: r ≈ 0





Correlación muy fuerte: r ≈ 1

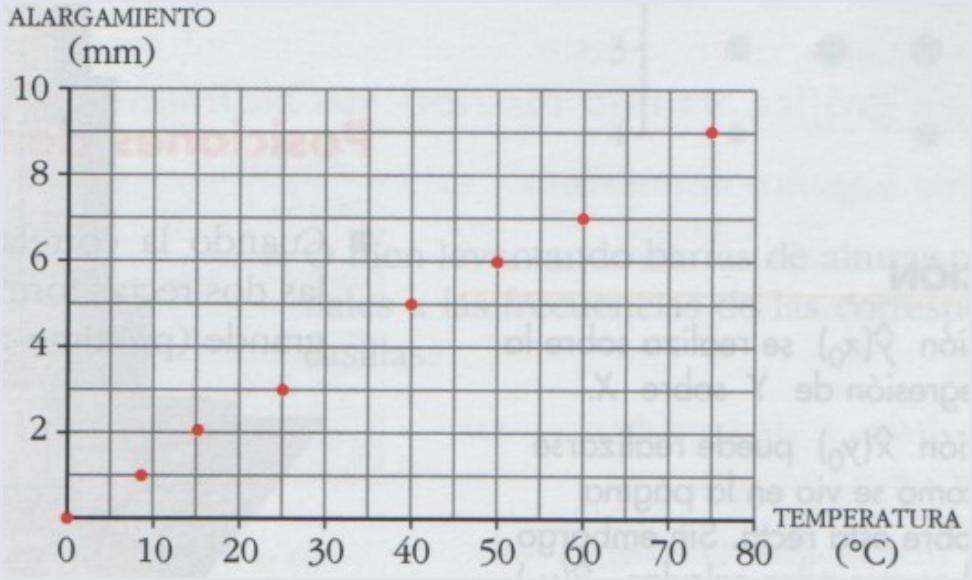
Con las rectas de regresión pueden hacerse estimaciones, obtener un valor para *Y* a partir de un cierto valor para *X*, o viceversa.

Estas estimaciones sólo serán fiables cuando la correlación sea fuerte, y sólo deben hacerse para valores dentro del rango estudiado o muy cerca de él

4.1. Estimaciones

Ejemplo. Se estudia la dilatación de una barra metálica a partir de la temperatura aplicada. Obtenemos la siguiente tabla:

TEMP. (°C)	ALARG. (mm)
0	0
8	1
16	2
25	3
40	5
50	6
60	7
75	9



Es evidente que hay mucha correlación. La calculamos y se obtiene:

$$r = 0,9994$$

La recta de regresión de Y sobre X es: $r_{yx}: y = 0,119x + 0,06$

Estimamos la dilatación para 55°: $\hat{y}(55) = 0,119 \cdot 55 + 0,06 = 6,6$

Estimamos la temperatura para 4 mm: $4 = 0,119 \cdot \hat{x} + 0,06 \rightarrow \hat{x}(4) = 33,1$

- Como la correlación es muy fuerte, y los valores a estimar están dentro del rango de datos, estas estimaciones son muy fiables.
- También sería muy fiable para hacer una estimación de, por ejemplo, 80° u 11 mm.
- No sería fiable si estimamos 100°, y mucho menos 200°
- Como la correlación es muy buena, las dos rectas de regresión son casi coincidentes, por lo que podemos usar solo una de ellas para hacer las estimaciones.
  - Si no hubiera tanta correlación, para estimar *y* a partir de *x* habría que usar la de *y* sobre *x*. Para estimar *x* a partir de *y*, usaríamos la recta de regresión de *x* sobre *y*. (Aunque en este caso las estimaciones no serían muy fiables)

5. Distribuciones marginales

Se trata de estudiar una variable tomando de la otra todos o solo algún valor. Se puede así estudiar la media, moda, cuartiles, etc

Tenemos 100 alumnos y las notas van de 1 a 5:

Mat x_i

Fís y_j \	1	2	3	4	5	
1	3	2	4	4	3	
2	1	2	3	2	5	
3	1	4	6	4	1	
4	3	6	7	10	9	
5	0	2	5	5	8	
						100

- Estudio de las notas de Física:

Primer cuartil:

$$\frac{100}{4}=25$$

y_j	$f_{\cdot j}$	$F_{\cdot j}$
1	16	16
2	13	29
3	16	45
4	35	80
5	20	100
	100	

$$Q_1(X)=2$$

- Estudio de las notas de Física en los alumnos que han sacado '*2*' en Matemáticas:

Y/X_2

y_j	f_{2j}
1	2
2	2
3	4
4	6
5	2

Media:

y_j	f_{2j}	$y_j \cdot f_{2j}$
1	2	2
2	2	4
3	4	12
4	6	24
5	2	10
	16	52

$$\left(\bar{Y}/X_2\right) = \frac{52}{16} = 3{,}25$$

- Estudio de las notas de Matemáticas en los alumnos que han sacado '*5*' en Física:

X/Y_5

x_i	f_{i5}
1	0
2	2
3	5
4	5
5	8

Mediana:

$$\frac{20}{2} = 10$$

x_i	f_{i5}	F_{i5}
1	0	0
2	2	2
3	5	7
4	5	12
5	8	20

$$\left(Me(X)/Y_5\right) = 4$$