Universidad de Cádiz
Departamento de Matemáticas

MATEMÁTICAS
para estudiantes de primer curso
de facultades y escuelas técnicas

Tema 13
Distribuciones bidimensionales. Regresión y correlación lineal

Elaborado por la Profesora Doctora María Teresa González Montesinos

Índice

1. Variables estadísticas bidimensionales

La consideración **simultánea** de dos series estadísticas que provienen de la medida de dos caracteres de una misma población, o de dos poblaciones distintas, recibe el nombre de **distribución bidimensional**.

Las dos series estadísticas consideradas *simultáneamente* constituyen una **serie estadística doble**. Cada elemento de esta serie doble viene medido o representado por dos caracteres (estatura, peso), (nota–física, nota–matemáticas), (estatura–padre, estatura–hijo). La *variación simultánea de ambos caracteres constituye una* **variable bidimensional** *denotada por* (x, y).

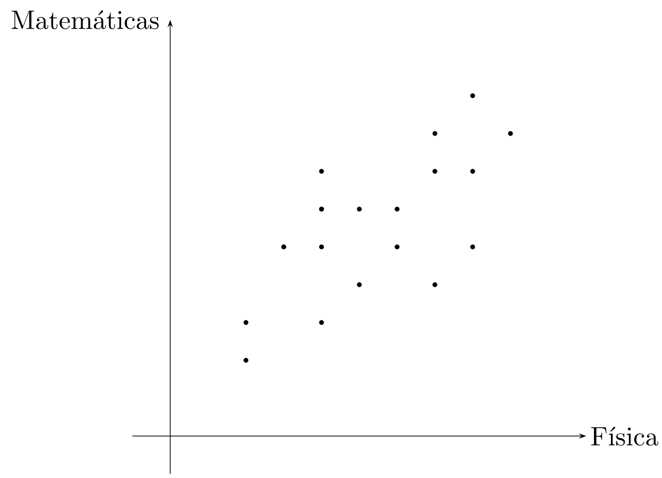
Por ejemplo, supongamos que los alumnos de un curso de tercero de BUP son 20 y que han obtenido las siguientes puntuaciones:

- Física: 5, 4, 3, 7, 2, 9, 8, 6, 4, 6, 5, 4, 7, 3, 8, 2, 4, 7, 4, 8.
- Matemáticas: 4, 6, 5, 7, 2, 8, 5, 5, 7, 6, 6, 5, 4, 5, 7, 3, 6, 8, 3, 9.

Son dos series estadísticas unidimensionales que, considerándolas simultáneamente, constituyen la serie estadística doble de variable bidimensional (x, y) =(Física,Matemáticas):

$$(x, y) = (5,4),\ (4,6),\ (3,5),\ (7,7),\ (2,2),\ (9,8),\ (8,5),\ (6,5),\ (4,7),\ (6,6),\ (5,6),\ (4,5),\ (7,4),$$
$$(3,5),\ (8,7),\ (2,3)\ (4,6)\ (7,8),\ (4,3),\ (8,9).$$

Podemos representar, en el plano, fijando un par de ejes cartesianos, cada pareja de valores correspondientes a las dos series estadísticas, formándose así una **nube de puntos**. Así, la nube de puntos correspondiente al ejemplo anterior sería



2. Regresión lineal. Rectas de regresión

El origen de la palabra *regresión* es debido al biólogo Galton al comparar, en sus investigaciones sobre la herencia genética, la estatura de los padre y de los hijos. Observó este autor que los de padres altos (que sobrepasaban la media de la población) eran también altos, pero menos. Es decir, esos hijos se acercaban (**regresaban**) a la media. Análogamente ocurría con hijos de padres bajos, inferiores a la media.

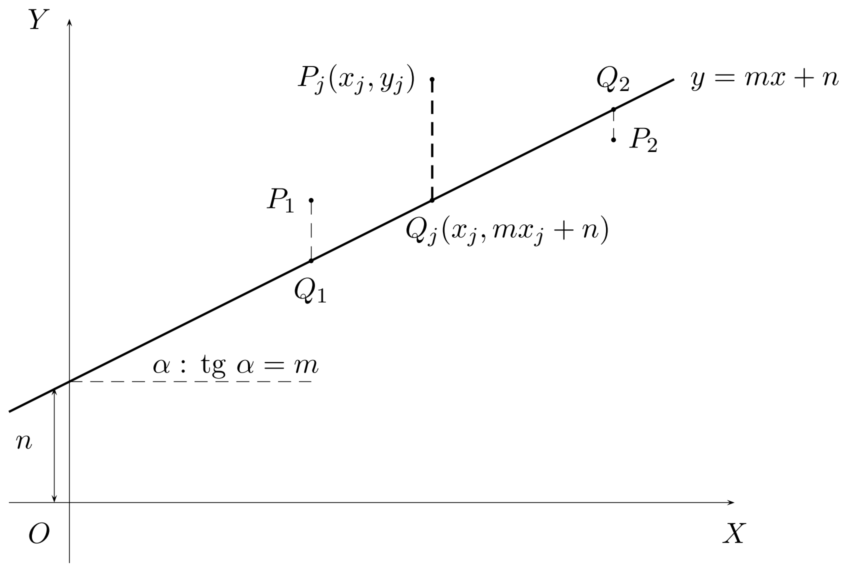


Figura 1

2.1. Recta de regresión de Y sobre X

Supongamos una nube de puntos de una serie bidimensinal (figura 1). Queremos buscar una recta $y = mx + n$ sobre la cual situar los puntos de la nube: $P_i \longrightarrow Q_i$. Cometeremos un error de estimación o cálculo que vamos a precisar. Proyectemos, paralelamente al eje de ordenadas OY, todos los puntos $P_i(x_i, y_i)$ de la nube sobre la recta $y = mx + n$ (figura 1). Obtendremos los nuevos puntos $Q_i(x_i, mx_i + n)$.

Tenemos así dos series de puntos:

- Las coordenadas antiguas, $P_i(x_i, y_i)$, correspondientes a los valores experimentales (reales) de las dos series estadísticas, cuyo estudio simultáneo constituye la serie estadística bidimensional (x_i, y_i).
- Las nuevas coordenadas $Q_i(x_i, mx_i + n)$ que son tales que:
  - Las abscisas x_i son los mismos valores reales de la primera serie.
  - Las ordenadas $mx_i + n$ son los valores estimados ("regresados"). Se calculan por la ecuación $y = mx + n$.

    Los valores $mx_i + n$ de esta serie son distintos, en general, de los de la serie real y_i.

Las diferencias $y_i - (mx_i + n)$ entre los valores reales y los estimados mediante la recta de regresión de y sobre x, $y = mx + n$, se conocen con el nombre de **desviaciones** respecto de esa recta. Son *errores* de cálculo o estimación, denominados también *residuos*.

Parece natural que al elegir la recta de regresión de y sobre x se proceda de tal forma que el error cometido, al sustituir los valores y_i por $mx_i + n$, sea lo más pequeño posible. Es decir, que las desviaciones o residuos sean mínimos. Tratar de encontrar la recta anterior con las condiciones impuestas se conoce por *ajustar* una recta de distribución bidimensional dada.

2.2. Determinación de la recta de regresión: método de los mínimos cuadrados

La recta que mejor se ajusta a los datos es aquella para la cual la suma de los cuadrados de los errores o residuos (distancias de P_i a Q_i) es mínima.

Conoceremos dicha recta cuando determinemos los parámetros m y n cuyo cálculo constituye el *ajuste de una recta a una distribución dada.*

Como ya se vio anteriormente, los errores cometidos vienen dados por $\mathrm{d}(P_i, Q_i) = y_i - mx_i - n$, de manera que $\mathrm{d}(P_i, Q_i)^2 = (y_i - mx_i - n)^2$. El método enunciado de los mínimos cuadrados dice

$$\sum_{i=1}^{N} \mathrm{d}(P_i, Q_i)^2 = \sum_{i=1}^{n} (y_i - mx_i - n)^2 \text{ ha de ser } \textbf{mínima}. \qquad (1)$$

Para que se verifique (1) ha de ser

$$\left.\begin{aligned} \sum_{i=1}^{N} x_i y_i &= m \sum_{i=1}^{N} x_i^2 + n \sum_{i=1}^{N} x_i, \\ \sum_{i=1}^{N} y_i &= m \sum_{i=1}^{N} x_i + nN, \end{aligned}\right\} \qquad (2)$$

denominadas **ecuaciones normales** de la recta de regresión de y sobre x. A partir de las ecuaciones (2) se deducen rápidamente los valores de m y n que determinan la recta.

Ejemplo 2.1 Las notas de diez alumnos de una lista de tercero de BUP en Matemáticas y Física han sido las siguientes:

- Matemáticas: $x_i = 7, 4, 3, 5, 8, 6, 9, 4, 6, 8$.
- Física: $y_i = 6, 3, 5, 4, 6, 7, 7, 5, 6, 7$.

Calculemos la recta de regresión de y sobre x. Para llegar a las ecuaciones (2), es conveniente disponer los cálculos en forma de tabla:

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
7	6	42	49	36
4	3	12	16	9
3	5	15	9	25
5	4	20	25	16
8	6	48	64	36
6	7	42	36	49
9	7	63	81	49
4	5	20	16	25
6	6	36	36	36
8	7	56	64	49
$\sum = 60$	$\sum = 56$	$\sum = 354$	$\sum = 396$	$\sum = 330$

Tabla 1

Sustituyendo estos valores en las ecuaciones (2) se tiene que

$$\left.\begin{aligned} 354 &= 396m + 60n, \\ 56 &= \ \ 60m + 10n. \end{aligned}\right\}$$

Despejando n en la segunda ecuación, $n = 5'6 - 6m$, y sustituyendo en la primera, $354 = 396m + 60(5'6 - 6m) = 36m + 336$, de donde

$$36m = 354 - 336 = 18 \Longrightarrow m = \frac{18}{36} = \frac{1}{2}.$$

Por lo tanto,

$$n = 5'6 - 6\frac{1}{2} = 2'6 = \frac{13}{5}.$$

Luego la recta de regresión viene dada por $y = \frac{1}{2}x + \frac{13}{5}$.

2.3. Cálculo abreviado de la recta de regresión de Y sobre X

Consideremos la nube de puntos de la serie estadística bidimensional (x, y) referida al sistema de referencia $(O; x, y)$. Efectuemos una traslación de ejes al nuevo origen $(O'; \bar{x}, \bar{y})$, siendo las coordenadas de O' las medias aritméticas de las respectivas series:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

Llamaremos X' e Y' a los nuevos ejes. Para todo punto P_i se tiene la siguiente relación vectorial:

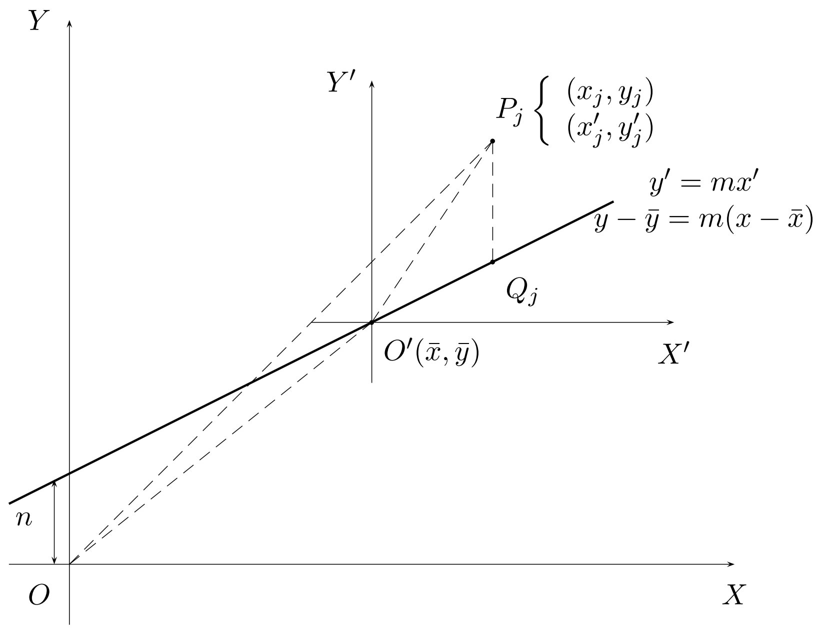


Figura 2

$$\overrightarrow{OP_i} = \overrightarrow{OO'} + \overrightarrow{O'P_i} \Longrightarrow \overrightarrow{O'P_i} = \overrightarrow{OP_i} - \overrightarrow{OO'} \Longrightarrow \left\{ \begin{array}{l} x'_i = x_i - \bar{x}, \\ y'_i = y_i - \bar{y}, \end{array} \right.$$

que constituyen dos nuevas series cuyos valores son las *desviaciones* de las dos series x_i e y_i respecto de sus medias respectivas:

$$x' = d_x = x - \bar{x}, \quad y' = d_y = y - \bar{y}. \tag{3}$$

El punto $O'(\bar{x}, \bar{y})$ es el **centro de gravedad** de la distribución, y las rectas de regresión (véase la figura 2) van a pasar por dicho punto.

La recta de regresión de y' sobre x' será $y' = mx' + n$. Para determinar los parámetros m y n procedemos por el método de los mínimos cuadrados:

$$\left.\begin{array}{r}\displaystyle\sum_{i=1}^{N} x_i'y_i' = m\sum_{i=1}^{N}(x_i')^2 + n\sum_{i=1}^{N} x_i', \\ \displaystyle\sum_{i=1}^{N} y_i' = m\sum_{i=1}^{N} x_i' + nN.\end{array}\right\} \tag{4}$$

Pero sabemos que la suma de las desviaciones de cualquier serie respecto de su media es cero:

$$\sum x_i' = \sum(x_i - \bar{x}) = \sum x_i - N\bar{x} = \sum x_i - N\frac{\sum x_i}{N} = 0.$$

Análogamente $\sum y_i' = 0$. De esta forma, las ecuaciones (4) quedan reducidas a

$$\left.\begin{array}{c}\sum x_i'y_i' = m\sum(x_i')^2 \\ 0 = nN,\end{array}\right\}$$

o bien,

$$n = 0, \quad m = \frac{\sum x_i'y_i'}{\sum(x_i')^2} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \tag{5}$$

donde m recibe el nombre de **coeficiente de regresión** de y respecto de x.

Sustituyendo m y n de (5) en la recta $y' = mx' + n$ queda

$$y' = \frac{\sum x_i'y_i'}{\sum(x_i')^2}x',$$

recta que pasa por $O'(\bar{x}, \bar{y})$. Volviendo a los ejes (X, Y), teniendo en cuenta (3), resulta

$$y - \bar{y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}(x - \bar{x}). \tag{6}$$

Ejemplo 2.2 Consideremos de nuevo el ejemplo 2.1. Para calcular la recta de regresión de y sobre x por el método abreviado, disponemos los cálculos de la siguiente manera:

x_i	y_i	$x_i' = x_i - \bar{x}$	$y_i' = y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
7	6	1	0’4	1	0’16	0’4
4	3	-2	-2’6	4	6’76	5’2
3	5	-3	-0’6	9	0’36	1’8
5	4	-1	-1’6	1	2’56	1’6
8	6	2	0’4	4	0’16	0’8
6	7	0	1’4	0	1’96	0
9	7	3	1’4	9	1’96	4’2
4	5	-2	-0’6	4	0’36	1’2
6	6	0	0’4	0	0’16	0
8	7	2	1’4	4	1’96	2’8
$\bar{x} = 6$	$\bar{y} = 5'6$	$\sum = 0$	$\sum = 0$	$\sum = 36$	$\sum = 16'4$	$\sum = 18$

Tabla 2

Sustituyendo los valores hallados en (6), tenemos

$$y - 5'6 = \frac{18}{36}(x - 6) \iff y - 5'6 = \frac{1}{2}(x - 6),$$

que es la misma recta que se obtuvo por el procedimiento anterior.

2.4. Recta de regresión de X sobre Y

La recta tendrá por ecuación $x = m'y + n'$. Para su obtención se procede de forma totalmente análoga a la anterior: proyectamos, paralelamente al eje de abscisas OX, todos los puntos $P(x_i, y_i)$ de la nube sobre la recta $x = m'y + n'$, obteniendo los nuevos puntos $Q_i(m'y_i + n', y_i)$.

Las **ecuaciones normales** serán en este caso

$$\left.\begin{array}{l} \displaystyle\sum_{i=1}^{N} x_i y_i = m' \sum_{i=1}^{N} y_i^2 + n' \sum_{i=1}^{N} y_i, \\ \displaystyle\sum_{i=1}^{N} x_i = m' \sum_{i=1}^{N} y_i + n' N. \end{array}\right\}$$

A partir de las ecuaciones normales reducidas se llega a que

$$n' = 0, \quad m' = \frac{\sum x_i' y_i'}{\sum (y_i')^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2},$$

donde m' se denomina **coeficiente de regresión** de x sobre y. La recta de regresión es ahora

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

Ejemplo 2.3 Calcular la recta de regresión de x sobre y para la serie bidimensional considerada en el ejemplo anterior.

Aprovechando los cálculos de las tablas 1 y 2 y sustituyéndolos en cualquiera de las expresiones para la recta se tiene que

$$x - 6 = \frac{18}{16'4}(y - 5'6).$$

3. Correlación lineal

El significado de la palabra *correlación* es el de relación mutua. El término estadístico **correlación** nos indica, por tanto, la *mutua dependencia que existe entre dos o más series estadísticas*. Así, la **correlación estadística** es la *interdependencia (la relación de dependencia) existente entre sus respectivas variables aleatorias.*

3.1. coeficiente de correlación lineal

Definición 3.1 *Llamamos* ***coeficiente de correlación lineal*** *entre dos series estadísticas, x_i e y_i, a la media geométrica de los dos coeficientes de regresión, m y m', correspondientes a las dos rectas de regresión; esto es,*

$$r = \sqrt{m \cdot m'}.$$

Partiendo de los valores m y m' hallados en la sección anterior por el método abreviado, tenemos que

$$r = \sqrt{m \cdot m'} = \sqrt{\frac{\left(\sum x_i' y_i'\right)\left(\sum x_i' y_i'\right)}{\left(\sum (x_i')^2\right)\left(\sum (y_i')^2\right)}} = \frac{\sum x_i' y_i'}{\sqrt{\left(\sum (x_i')^2\right)\left(\sum (y_i')^2\right)}} =$$
$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum (x_i - \bar{x})^2\right)\left(\sum (y_i - \bar{y})^2\right)}}.$$

Ejemplo 3.1 Calculemos el coeficiente de correlación del ejemplo dado en la sección anterior:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum (x_i - \bar{x})^2\right)\left(\sum (y_i - \bar{y})^2\right)}} = 0'74.$$

3.2. Variación y estudio de las propiedades del coeficiente de correlación

Dada la seria estadística bidimensional (x_i, y_i), recordemos que la varianza de los valores x e y vienen dadas por

$$\sigma_x^2 = \frac{\sum(x_i - \bar{x})^2}{N} = \frac{\sum(x_i')^2}{N}, \quad \sigma_y^2 = \frac{\sum(y_i - \bar{y})^2}{N} = \frac{\sum(y_i')^2}{N}.$$

Definimos la **covarianza** de la serie como

$$\sigma_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N} = \frac{\sum x_i' y_i'}{N}.$$

Se deduce fácilmente que

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N\sigma_x\sigma_y} = \frac{\sigma_{xy}}{\sigma_x\sigma_y} = \frac{m\sigma_x}{\sigma_y} = \frac{m'\sigma_y}{\sigma_x}.$$

Operando, se obtiene como suma de los cuadrados de los residuos:

$$\sum \overline{P_iQ_i}^2 = N\sigma_y^2(1 - r^2). \tag{7}$$

Como el primer miembro de esta igualdad es mayor o igual que cero, el segundo miembro también lo será; al ser $N, \sigma_y^2 > 0$, tendremos que

$$1 - r^2 \geq 0 \iff -1 \leq r \leq 1,$$

es decir, *el coeficiente correlación lineal es un número comprendido entre* -1 *y* 1.

La variación de r en el intervalo $[-1, 1]$ nos permite un estudio sencillo de la dependencia de las dos series estadísticas de una distribución bidimensional. (El coeficiente de correlación es el valor más empleado, por su sencillez y rápida interpretación en sociología, psicología, etc. Así, por ejemplo, gracias a dicho coeficiente podemos verificar el grado de fiabilidad y veracidad de un test de inteligencia, etc.)

Podemos distinguir varios casos:

(1) Si $|r| = 1$, diremos que se produce una *correlación total, completa o perfecta*. Llevando este valor a (7), nos queda

$$\sum \overline{P_iQ_i}^2 = N\sigma_y^2(1 - 1) = 0 \Longrightarrow P_i = Q_i, \ \forall\, i.$$

Los valores reales P_i y los estimados Q_i sobre las rectas de regresión coinciden. Al ser nulos los residuos o errores, $\bar{P_iQ_i}$, todos los puntos de la nube están en una recta, por lo que la dependencia entre las variables aleatorias es *funcional* y las dos rectas de regresión coinciden.

(2) Si $r = 0$ nos encontramos ante una *correlación nula*, y la expresión (7) admite su valor máximo:

$$\sum \overline{P_iQ_i}^2 = N\sigma_y^2(1 - 0) = N\sigma_y^2,$$

lo cual indica que los puntos P_i están separados lo máximo posible de Q_i (dispersión completa). Los errores $\overline{P_iQ_i}$ son máximos: se trata de la llamada *independencia aleatoria*. En este caso los coeficientes de regresión vienen dados por

$$m = r\frac{\sigma_y}{\sigma_x}, \quad m' = r\frac{\sigma_x}{\sigma_y}.$$

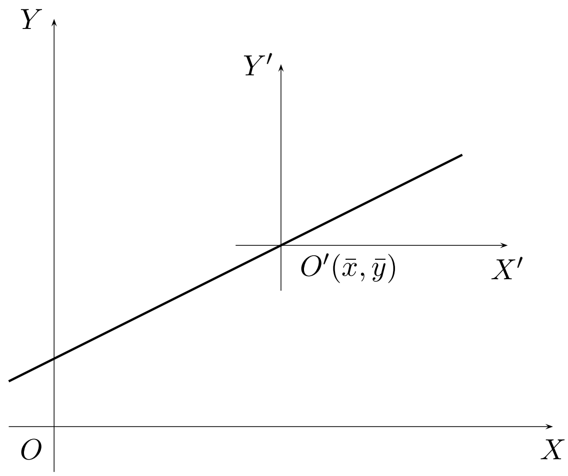


Figura 3: $r = 1$.

Luego las rectas de regresión serán

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}), \quad x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}).$$

Al ser $r = 0$, se deduce que $m = m' = 0$, quedando

$$y = \bar{y}, \quad x = \bar{x}.$$

Las dos rectas regresión son paralelas a los ejes coordenados, formando, por tanto, un ángulo de 90° (véase figura 4):

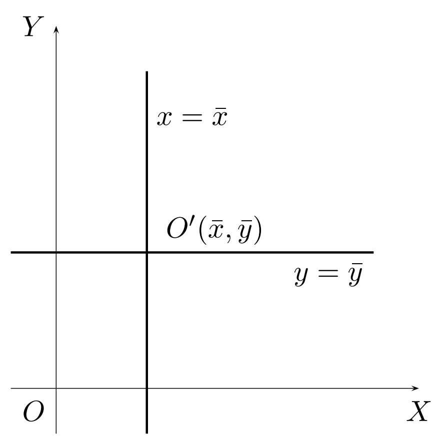


Figura 4: $r = 0$.

(3) Si $0 < |r| < 1$ nos encontramos ante una *dependencia aleatoria*. Este caso comprende todos los numerosos fenómenos de la vida real que se encuentran entre los dos casos límites anteriores.

   (3.1) Si $|r|$ se aproxima a 1, entonces la dependencia aleatoria se aproxima a la funcional; P_i estaría próximo a Q_i y el ángulo que forman las dos rectas tiende a cero.

(3.2) Si $|r|$ tiene un valor próximo a cero, entonces P_i y Q_i están muy separados, con lo que la dependencia aleatoria es pequeña (tiende a la independencia aleatoria) y el ángulo entre las dos rectas de regresión tiende a 90°.

Ejemplo 3.2 Si $r = 0'92$ entonces la correlación o dependencia entre las dos series es alta y las rectas de regresión casi coinciden (figura 5). Si $r = 0'25$, la correlación es baja y el ángulo entre las dos rectas de regresión se aproxima al ángulo recto (figura 6).

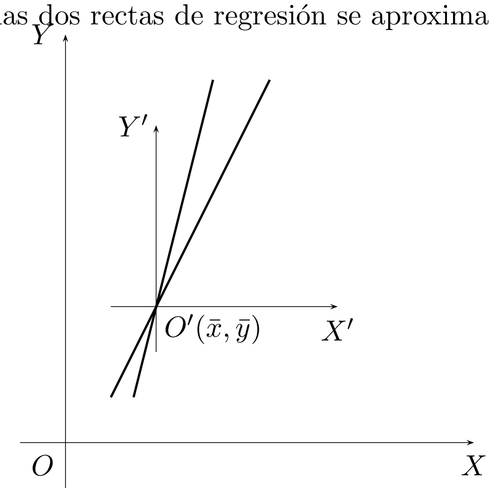


Figura 5: $r = 0'92$.

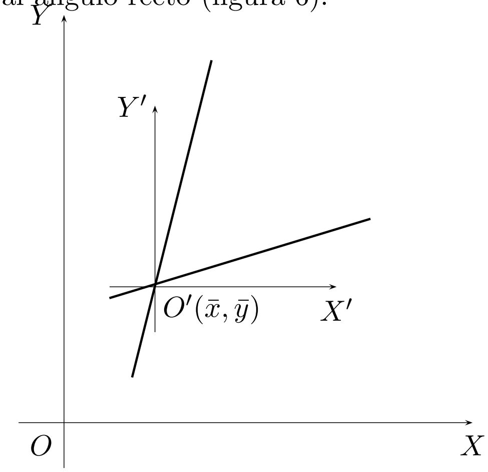


Figura 6: $r = 0'25$.

(4) Si $r > 0$ diremos que se produce una *correlación directa o positiva*. En este caso las pendientes de las rectas de regresión, $m = r\dfrac{\sigma_y}{\sigma_x}$ y $m' = r\dfrac{\sigma_x}{\sigma_y}$, son ambas positivas, pues las desviaciones típicas son siempre positivas. Al crecer x, también crece y.

(5) Si $r < 0$ diremos que se produce una *correlación inversa o negativa*. Razonando de forma análoga al caso anterior, se tiene que $m, m' < 0$, esto es, las pendientes de las rectas de regresión son negativas, y al crecer x disminuye y.

En resumen, el coeficiente de correlación r mide la dependencia aleatoria de dos series estadísticas:

$\lvert r\rvert$	0	$0 < \lvert r\rvert < 1$	1
Dependencia	Correlación nula Indep. aleatoria	Dep. aleatoria	Correlación perfecta Dep. funcional

Vamos a considerar un ejemplo que nos ayude en la resolución de otros ejercicios. Supongamos las dos series estadísticas siguientes:

$$x_i : 7, 4, 8, 6, 9, 5, 2, 4, 9, 8, 6, 7.$$
$$y_i : 6, 5, 7, 6, 8, 7, 1, 3, 8, 7, 8, 6.$$

Se pide:

a) Representación del diagrama de dispersión.

b) Rectas de regresión.

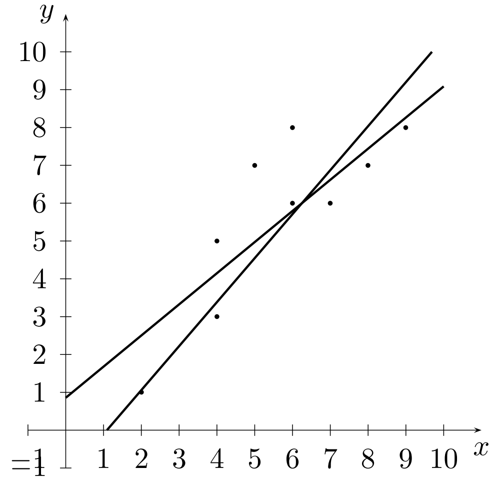
c) Coeficiente de correlación.

d) Valores estimados en la recta de regresión de y sobre x.

e) Errores cometidos en esta estimación.

Nota 3.1 Si nos pidieran la dependencia estadística, el único dato que necesitaríamos sería el coeficiente de correlación. Por ser dos muestras pequeñas no agrupamos los datos repetidos.

a) La siguiente figura representa la nube de puntos:



b) Dispongamos los datos necesarios en la siguiente tabla:

x_i	y_i	x'_i	y'_i	$(x'_i)^2$	$(y'_i)^2$	$x'_i y'_i$
7	6	0'75	0	0'5625	0	0
4	5	-2'25	-1	5'0625	1	2'25
8	7	1'75	1	3'0625	1	1'75
6	6	-0'25	0	0'0625	0	0
9	8	2'75	2	7'5625	4	5'5
5	7	-1'25	1	1'5625	1	-1'25
2	1	-4'25	-5	18'0625	25	21'25
4	3	-2'25	-3	5'0625	9	6'75
9	8	2'75	2	7'5625	4	5'5
8	7	1'75	1	3'0625	1	1'75
6	8	-0'25	2	0'0625	4	-0'5
7	6	0'75	0	0'5625	0	0
$\bar{x} = 6'25$	$\bar{y} = 6$	$\sum = 0$	$\sum = 0$	$\sum (x'_i)^2 = 52'25$	$\sum (y'_i)^2 = 50$	$\sum x'_i y'_i = 43$

La recta de regresión de y sobre x será

$$y - \bar{y} = \frac{\sum x'_i y'_i}{\sum (x'_i)^2}(x - \bar{x}) \iff y - 6 = \frac{43}{52'25}(x - 6'25),$$

y la de x sobre y vendrá dada por

$$x - \bar{x} = \frac{\sum x'_i y'_i}{\sum (y'_i)^2}(y - \bar{y}) \iff x - 6'25 = \frac{43}{50}(y - 6).$$

c) El coeficiente de correlación será

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum (x_i - \bar{x})^2\right)\left(\sum (y_i - \bar{y})^2\right)}} = \frac{43}{\sqrt{52'25 \cdot 50}} = 0'8412.$$

d) Para calcular los valores estimados y los errores debemos tener en cuenta la recta de regresión de y sobre x: $y-6=\dfrac{43}{52'25}(x-6'25)$. En la siguiente tabla se presentan las estimaciones y los errores cometidos:

x_i	7	4	8	6	9	5	2	4	9	8	6	7
y_i	6	5	7	6	8	7	1	3	8	7	8	6
y'_i	6'62	4'15	7'44	5'79	8'26	4'97	2'50	4'15	8'26	7'44	5'76	6'62
Errores	0'38	0'85	-0'44	0'21	-0'26	2'03	-1'50	-1'15	-0'26	-0'44	2'24	-0'62

Para hallar, por ejemplo, el sexto valor de la serie bidimensional, $(5,7)$, hacemos lo siguiente:

$$x_6=5\Longrightarrow y_6-6=\frac{43}{52'25}(5-6'25)\Longrightarrow y_6=\frac{43}{52'25}(5-6'25)+6=4'97.$$

El error de estimación cometido será $7-4'97=2'03$.

4. Ejercicios propuestos

(1) Las notas obtenidas por un grupo de alumnos de tercero de BUP en Matemáticas y Física son las siguientes:

- Matemáticas: $x_i:\ 4,6,8,3,7,5,3,3,7,8.$
- Física: $y_i:\ 5,4,7,5,9,6,4,3,6,6.$

a) Calcúlense las dos rectas de regresión.

b) Estímense los valores de y cuando x vale 4 y 9.

c) Estímense los valores de x cuando y vale 3 y 8.

(2) Se ha aplicado una batería de tests de inteligencia a un grupo de seis estudiantes, y al mismo tiempo, se ha anotado su rendimiento académico según la tabla adjunta:

Inteligencia	102	96	122	110	100	130
Rendimiento	6	7	7	8	4	5

Halla las dos rectas de regresión.

(3) En un hospital se ha aplicado un medicamento, A, a 100 enfermos, y en otro hospital se aplicado otro medicamento, B, a otros 100 enfermos. El número de curados cada día durante los diez primeros días es el siguiente:

X: Medicamento A	7	3	2	8	6	5	4	1	3	1
Y: Medicamento B	4	5	2	4	6	7	2	2	1	2

a) Hállense las rectas de regresión.

b) Estímense los valores de y en la primera recta, y los valores de x en la segunda.

c) Hállense los residuos o errores de estimación.

(4) Calcula el coeficiente de correlación de las siguientes series estadísticas bidimensionales:

a) $\left\{\begin{array}{ll} x_i: & 3,5,7,9 \\ y_i: & 5,7,8,9 \end{array}\right.$,

b) $\left\{\begin{array}{ll} x_i: & 1,9,7,6,8,7 \\ y_i: & 1,7,4,5,7,4 \end{array}\right.$,

c) $\left\{\begin{array}{ll} x_i: & 4,7,3,9,6,5,3,2 \\ y_i: & 8,8,5,3,8,1,9,5 \end{array}\right.$.

Interprétense los resultados.

(5) Al aplicar dos tests de memoria a un grupo de alumnos, se han obtenido los siguientes resultados:

$$\begin{array}{ll} x_i: & 3,5,7,4,9,8,7,6,5,3,9,3. \\ y_i: & 4,6,8,5,7,7,8,7,6,4,8,5. \end{array}$$

Indica la dependencia y la correlación entre los dos tests.

(6) El cambio de la moneda de dos naciones no europeas respecto al euro ha sufrido las siguientes fluctuaciones:

$$\left\{\begin{array}{l} 1'3,\ 2'5,\ 1'2\,,1'1\,,0'9; \\ 1'1,\ 2'3,\ 0'9\,,1'0\,,0'8. \end{array}\right.$$

Indica la dependencia comercial y económica de esas dos naciones mediante el coeficiente de correlación.