

Introducción

En los temas anteriores nos hemos preocupado de representar gráficamente y de procesar numéricamente los datos relativos a una variable de un colectivo. Nos interesaba, por ejemplo, conocer la distribución de las alturas de los jugadores de los diferentes equipos de baloncesto de la ciudad. Se trataba de estudiar distribuciones unidimensionales. Pero cada miembro de un colectivo presenta diversos aspectos por los que nos podemos interesar. Esto nos lleva a consideraciones muy importantes para establecer con cierta confianza la presencia o ausencia de relaciones significativas entre estos aspectos diferentes.

Lo que ahora vamos a estudiar es la relación o dependencia que existe entre dos variables, y no un estudio aislado de cada una de ellas como ya hicimos en el tema anterior. Se trata pues de cuantificar la relación entre dos variables. En un colectivo bien amplio podemos establecer una medida de la *gordura* y del *grado de buen humor*, cosas ambas bastante subjetivas. Cada miembro del colectivo tiene dos números asignados y así, en un plano con dos ejes rectangulares, el eje de la *gordura* y el eje del *buen humor*, podemos asignar a cada miembro un punto. Obtenemos una *nube de puntos*. Es la representación de una distribución bidimensional.

Si en el espacio tridimensional pudiésemos pintar y manejarnos tan bien como el plano, sería posible estudiar gráficamente las relaciones conjuntas de tres variables distintas, por ejemplo *gordura*, *buen humor* e *ingresos anuales*. Nos llevaría a una distribución tridimensional. Esto que es difícil a mano se hace con la ayuda del ordenador. La *nube de puntos* que se origina se puede manipular y observar desde diferentes ángulos y adquirir así una buena idea de las relaciones conjuntas entre las tres variables.

Relación estadística y relación funcional. Correlación

El importe de la factura de teléfono depende del número de llamadas realizadas y del tiempo empleado en cada una de ellas (del número de pasos); aparte lleva un suplemento por el alquiler de los equipos telefónicos. La expresión que da este importe es de la forma $I = a + np$, donde a es el precio de alquiler de los equipos, n el número de pasos realizados y p el precio de cada paso. La relación que existe entre a y n es funcional y se ajusta a una recta.

Si lanzamos una piedra hacia arriba, llegará tanto más alta cuanto más fuerte la lancemos. Y existe una fórmula que nos permite calcular, exactamente, la altura que alcanzará en función de la velocidad con que se lanza. Es una relación funcional

Las relaciones funcionales se verifican siempre: globalmente y para cada valor particular.

Las personas, en general, pesan más cuanto más altas son, pero no se podría dar una fórmula que permitiera obtener el peso de cualquier persona conociendo su estatura. La relación que existe entre el peso y la estatura de una persona no es funcional. Una relación similar cabría esperar entre el rendimiento de una cosecha y los kilogramos de fertilizantes empleados. Pero en ninguno de estos casos la relación es precisa, exacta.

Dichas relaciones, de algún modo, están sujetas al azar y son efecto de otras muchas causas: además de la altura, la edad o la dieta influyen en el peso de una persona. El agua y la temperatura también afectan al rendimiento de las cosechas.

En los factores asociados al comportamiento humano, la interrelación de efectos y causas es todavía más complejas pues, además de las circunstancias, muchas de ellas azarosas, en las personas juega un papel decisivo la libertad, y este factor sí que es impredecible.

A pesar de las dificultades apuntadas, la Estadística es capaz de dar respuesta a muchos de los problemas fenomenológicos o sociales que pueden presentarse. Y es capaz de dar respuestas porque no pretende que éstas sean unívocas, funcionales, de causa efecto.

Las relaciones estadísticas sólo se verifican de una manera global; para cada valor particular la respuesta puede ser múltiple.

La Estadística da soluciones en clave probabilística; por eso siempre hay que hablar de posible error, y hay que hacer el contraste de hipótesis pertinente (aunque a este nivel, para simplificar las cosas, no lo hagamos). Por lo mismo matizaremos las palabras, y más que de relación entre variables preferimos hablar de **correlación** entre ellas.

Las ciencias exactas tratan, casi exclusivamente, con relaciones funcionales. Sin embargo, para las ciencias sociales y humanas el instrumento de la correlación es muy importante. Por ejemplo:

- En un conjunto de países, ¿qué relación hay entre la renta per cápita y la expectativa de vida al nacer? ¿Y entre la renta per cápita y el índice de natalidad?
- En la promoción publicitaria, ¿qué relación hay entre el número de minutos que cada producto se anuncia en T.V. y el incremento de ventas que se produce?

En estos ejemplos se aprecia una relación estadística más o menos estrecha. Nuestro objetivo es sistematizar este estudio de modo que lo que comienza siendo una idea intuitiva acabe suficientemente estructurado y pueda ser medido mediante una expresión matemática.

Tampoco tendremos en cuenta todas las causas: fertilizantes, lluvia, temperatura, tipos de suelo, destreza y esmero del agricultor... Nos limitaremos a considerar dos variables solamente, de ahí el título de este tema: **distribuciones bidimensionales**.

Diagramas de dispersión. Nubes de puntos

A las variables estadísticas resultantes de la observación de un fenómeno respecto de dos modalidades se las llama **variables estadísticas bidimensionales**.

Las variables estadísticas bidimensionales las representaremos por el par (x, y) , donde x es una variable unidimensional que toma los valores $x_1, x_2, x_3, \dots, x_n$, e y es otra variable unidimensional que toma los valores $y_1, y_2, y_3, \dots, y_n$. Por tanto, la variable estadística bidimensional (x, y) toma estos valores:

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n) \quad \text{o también} \quad (x_i, y_i), (1 \leq i \leq n)$$

Supongamos un ejemplo, en el que la variable x representa la edad e y el grado de psicomotricidad de un grupo de 44 minusválidos mentales.

Edad (años) x	Grado de psic. y	Nº de personas f_i
[5, 7)	[25, 30)	4
[5, 7)	[30, 35)	2
[5, 7)	[35, 40)	1
[7, 9)	[25, 30)	3
[7, 9)	[30, 35)	7
[7, 9)	[35, 40)	1
[7, 9)	[40, 45)	2
[9, 11)	[30, 35)	2
[9, 11)	[35, 40)	11
[11, 13)	[25, 30)	1
[11, 13)	[35, 40)	1
[11, 13)	[40, 45)	6
[11, 13)	[45, 50)	3
		44

Las tablas como las de la izquierda se llaman **tablas simples**. Reflejan el comportamiento de la variable estadística bidimensional (x, y) a partir de los valores individuales que toman cada una de las variables estadísticas unidimensionales x e y .

Variable x_i	Variable y_i	Frec. absoluta f_i
x_1	y_1	f_1
x_2	y_2	f_2
x_3	y_3	f_3
...	...	...
...	...	...
...	...	...
x_n	y_n	f_n
		$\sum_{i=1}^n f_i$

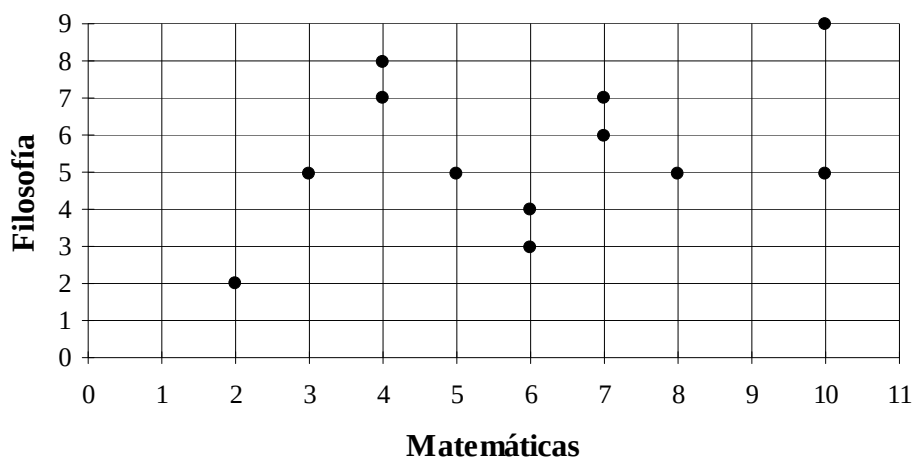
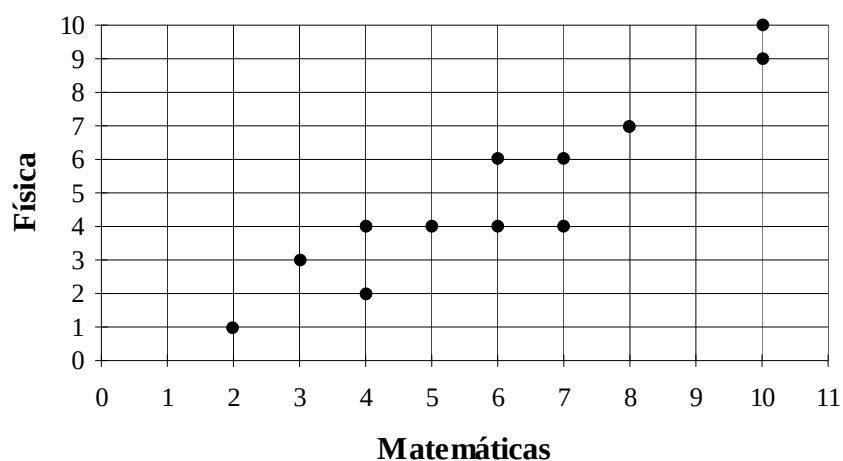
En general, toda tabla bidimensional de frecuencias se puede expresar como muestra la tabla al margen, donde x e y designan las variables estadísticas unidimensionales, x_i e y_i los valores que toman respectivamente, y f_i son las frecuencias absolutas.

Dado que los valores de una variable estadística bidimensional son pares de números reales de la forma (x_i, y_i) , si los representamos sobre un sistema de ejes cartesianos se obtiene un conjunto de punto sobre el plano. A este conjunto de puntos se lo denomina **diagrama de dispersión** o **nube de puntos**.

Ejemplo: A los 12 primeros alumnos (por orden de lista) de un colegio se les toma las notas de los últimos exámenes en Matemáticas, Física y Filosofía.

Representar gráficamente las notas de Matemáticas frente a las de Física en un diagrama de puntos, y las notas Matemáticas frente a las de Filosofía en otro.

Alumno	a	b	c	d	e	f	g	h	i	j	k	l
Mat.	2	3	4	4	5	6	6	7	7	8	10	10
Fís.	1	3	2	4	4	4	6	4	6	7	9	10
Filos.	2	5	7	8	5	3	4	6	7	5	5	9

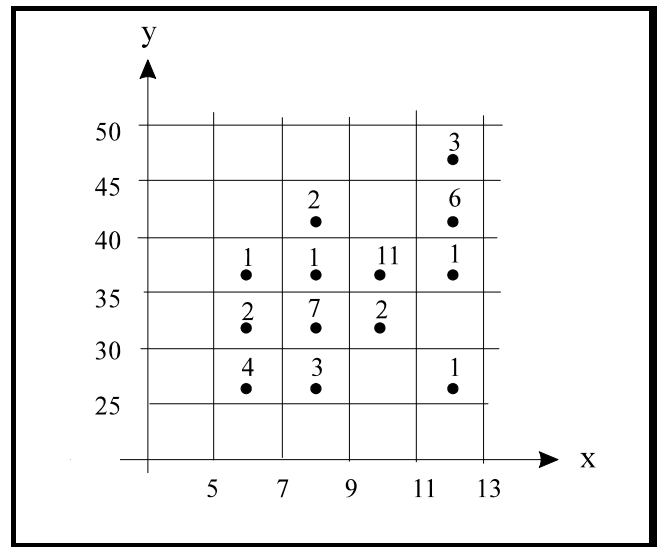


Cuando se tienen muchos datos, y por tanto los pares de valores se repiten muchas veces o bien los valores se encuentran agrupados en intervalos, se utilizan las **tablas de doble entrada** como la representada a la derecha. En ellas los datos se disponen de tal manera que en la primera fila se colocan los valores de la variable x y en la primera columna por la izquierda se colocan los valores de variable y .

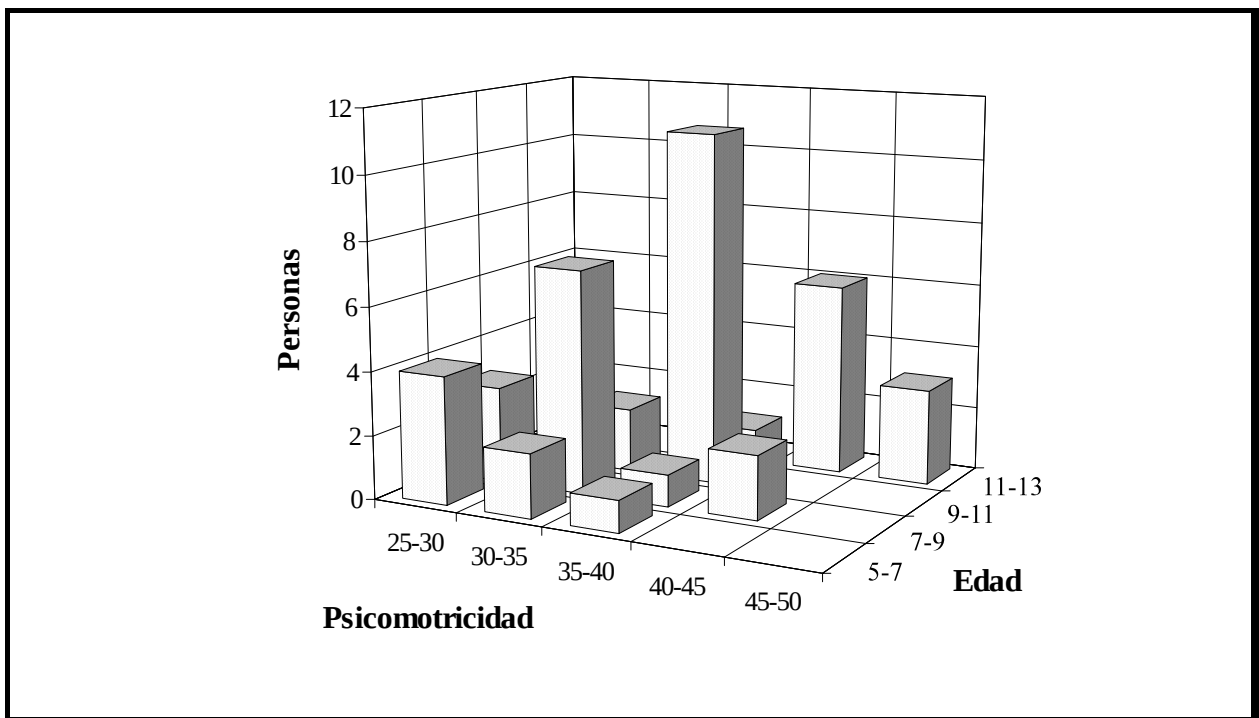
y \ x	[5, 7)	[7, 9)	[9, 11)	[11, 13)
[25, 30)	4	3	—	1
[30, 35)	2	7	2	—
[35, 40)	1	1	11	1
[40, 45)	—	2	—	6
[45, 50)	—	—	—	3
				44

Este tipo de datos también admite un diagrama de dispersión. En este caso, se reticulará el plano cartesiano según estos intervalos y en cada rectángulo inscribiremos tantos puntos como nos indique la tabla de frecuencias absolutas.

Como a veces el número de puntos es excesivamente grande, se suele colocar un sólo punto y encima de él un número equivalente a la frecuencia



Otras veces, la representación gráfica se hace, bien *hinchando* los puntos proporcionalmente a su frecuencia o bien levantando barras de altura proporcional a la frecuencia de cada casilla.



Tratamiento estadístico de datos bidimensionales

Aunque el objetivo es tratar de determinar el posible efecto de una variable sobre otra, empezaremos estudiando cada una de las variables por separado.

Si el conjunto de datos es de la forma $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$, podemos calcular los parámetros estadísticos de cada una de las componentes.

Recordemos las definiciones de media y varianza para distribuciones de variables estadística unidimensional:

Medias

$$\text{Media de la variable } x \quad \bar{x} = \frac{\sum_{i=1}^n x_i f_i}{\sum f_i}$$

$$\text{Media de la variable } y \quad \bar{y} = \frac{\sum_{i=1}^n y_i f_i}{\sum f_i}$$

Varianzas

$$\text{Varianza de la variable } x \quad s_x^2 = \frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{\sum f_i} = \frac{\sum_{i=1}^n x_i^2 f_i}{\sum f_i} - \bar{x}^2$$

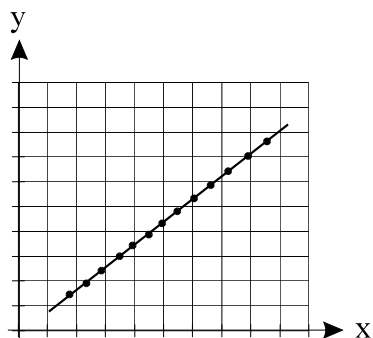
$$\text{Varianza de la variable } y \quad s_y^2 = \frac{\sum_{i=1}^n f_i (y_i - \bar{y})^2}{\sum f_i} = \frac{\sum_{i=1}^n y_i^2 f_i}{\sum f_i} - \bar{y}^2$$

Distintos diagramas de dispersión. Correlación

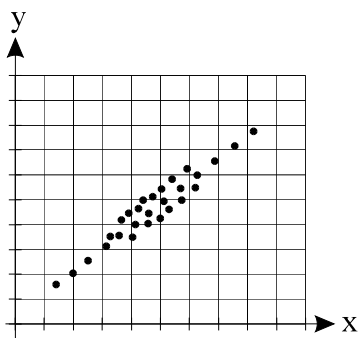
Lo que nos interesa observar en un diagrama de dispersión es si la relación entre las dos variables es lineal o curvilínea, si al crecer una la otra también crece o por el contrario disminuye, si la relación es funcional, etc. Será funcional cuando ambas variables están relacionadas por una función. En caso contrario, la relación será tanto más fuerte o más débil dependiendo de la mayor o menor tendencia de los puntos del diagrama a acercarse a la representación de una función.

Los siguientes diagramas de dispersión reflejan la relación existente entre dos variables, x e y .

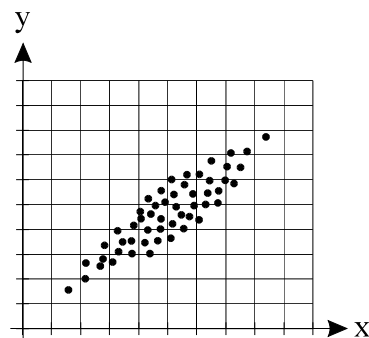
Relación lineal positiva funcional



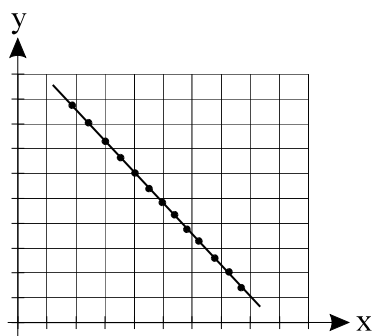
Relación lineal positiva fuerte



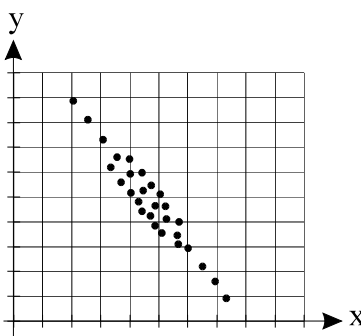
Relación lineal positiva débil



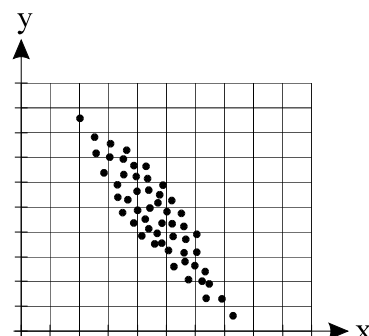
Relación lineal negativa funcional



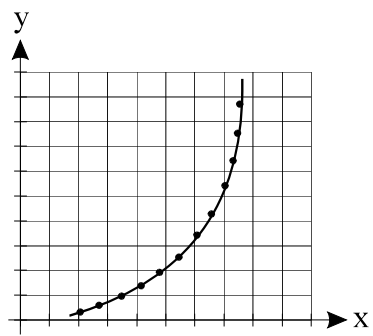
Relación lineal negativa fuerte



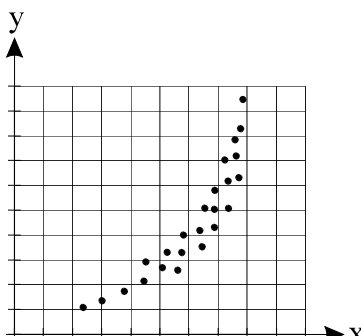
Relación lineal negativa débil



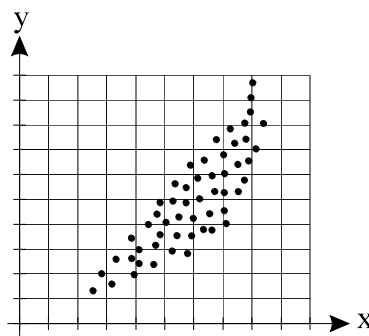
Relación curvilínea positiva funcional

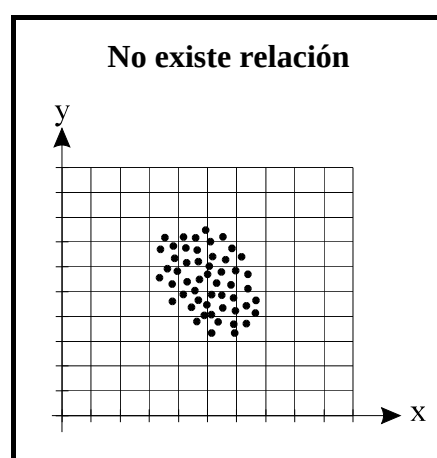
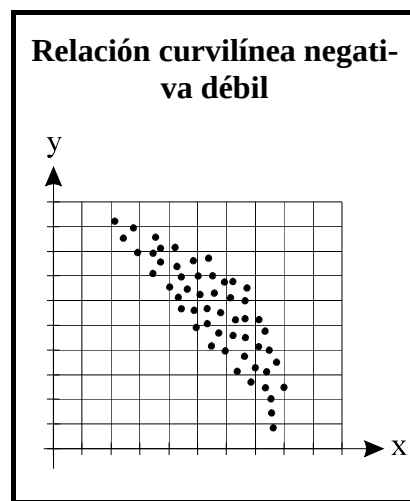
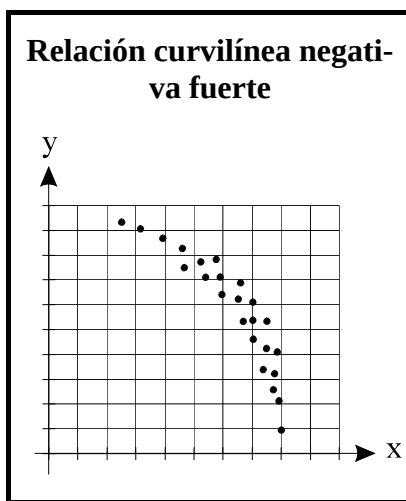
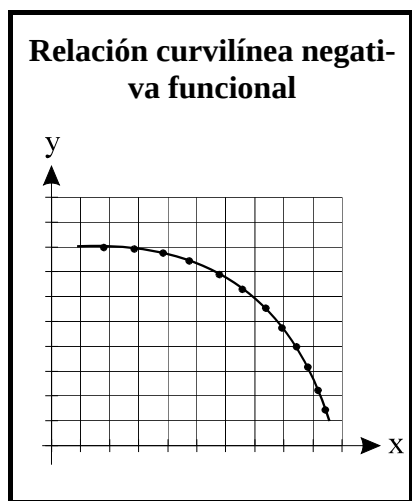


Relación curvilínea positiva fuerte



Relación curvilínea positiva débil





En los ejemplos anteriores hemos analizado lo que significa *relación entre variables*. **Llamaremos correlación a la teoría que trata de estudiar la relación o dependencia que existe entre las dos variables que intervienen en una distribución bidimensional.**

- **La correlación es lineal o curvilínea** según que el diagrama de puntos se condense en torno a una línea recta o a una curva.
- **La correlación es positiva o directa** cuando a medida que crece una variable también crece la otra.

La correlación es negativa o inversa cuando a medida que crece una variable la otra decrece.

La correlación es nula cuando no existe ninguna relación entre ambas variables.. En este caso los puntos del diagrama están esparcidos al azar, sin formar ninguna línea, y se dice que las variables están *incorrelacionadas*.

- **La correlación es de tipo funcional** si existe una función tal que todos los valores de la distribución la satisfacen. En caso contrario, será tanto más fuerte o más débil dependiendo de la mayor o menor tendencia de los valores de la distribución a satisfacer una determinada función. Nosotros nos centraremos en el estudio de la correlación lineal.

Medida de la correlación

Supongamos que a los 12 primeros alumnos (por orden de lista) de un colegio se les toma las notas de los últimos exámenes en Matemáticas, Física, Filosofía y educación Física (ver tabla).

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
Filosofía	2	5	7	8	5	3	4	6	7	5	5	9
Ed. Física	5	8	6	8	9	5	7	6	8	5	6	9

$$\rightarrow \bar{x} = 6$$

$$\rightarrow \bar{y} = 5$$

$$\rightarrow \bar{y} = 5'5$$

$$\rightarrow \bar{y} = 6'83$$

Vamos a estudiar la relación que hay entre las notas que han obtenido en Matemáticas y las que han obtenido en Física, Filosofía y Educación Física.

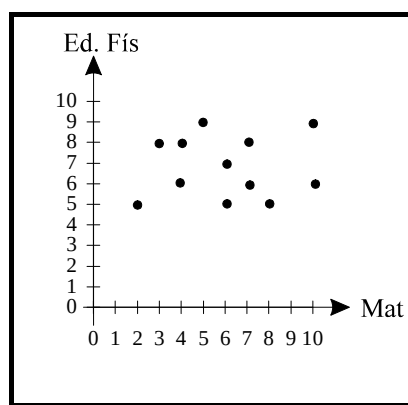
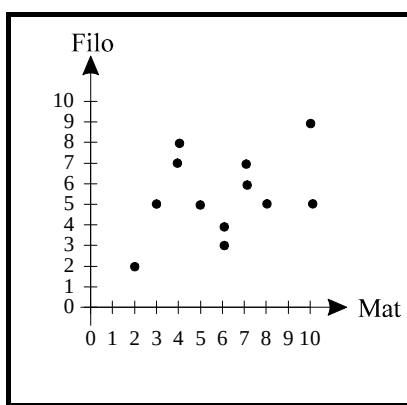
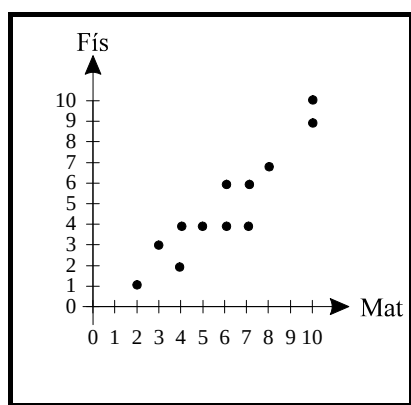
Las distribuciones bidimensionales:

Notas de Matemáticas - Notas de Física

Notas de Matemáticas - Notas de Filosofía

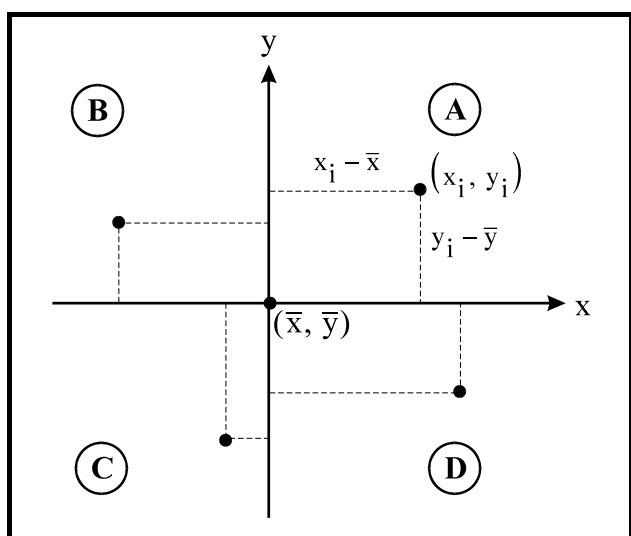
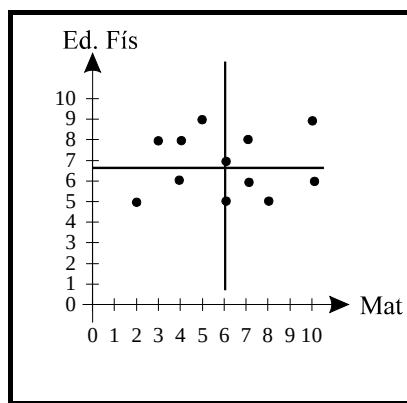
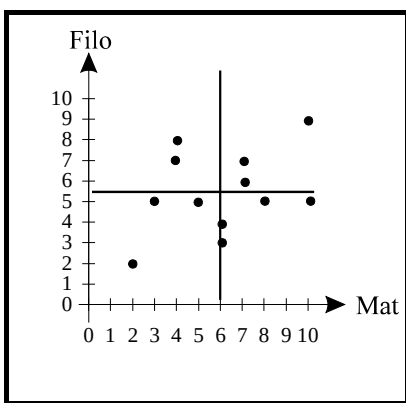
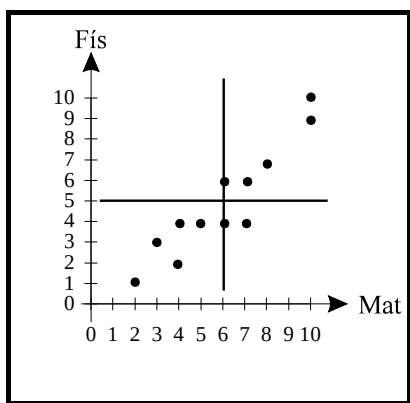
Notas de Matemáticas - Notas de Educación Física

pueden representarse a través de los siguientes diagramas de puntos:



Algunas nubes de puntos sugieren rectas. Cuanto más apretados están los puntos a su correspondiente recta, mayor es la correlación.

Para cuantificar ese efecto, trazamos unos nuevos ejes de coordenadas con origen en el punto de coordenadas $(\bar{x}, \bar{y})$, es decir, por un punto cuya abscisa es la media de la primera variable y cuya ordenada es la media de la segunda variable. Este punto se llama **centro de gravedad de la distribución**.



Cuanto más fuerte es la correlación, cuanto más apretados están los puntos en torno a la recta, más puntos hay en los cuadrantes A y C y menos en los cuadrantes B y D.

Si tenemos en cuenta que en los cuadrantes A y C el producto $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ es positivo, mientras que en los otros dos cuadrantes es negativo, deducimos que, cuanto más fuerte es la correlación, más puntos hay para los que el producto $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ es positivo menos puntos para los que el producto es negativo.

	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
en A	+	+	+
en B	-	+	-
en C	-	-	+
en D	+	-	-

Cuanto mayor es la correlación mayor es la suma:

$$\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

En el razonamiento anterior, hemos supuesto que todos los puntos tienen la misma importancia. Esto no es así. Los puntos, cuanto más alejados están del centro de gravedad más influyen en la correlación, tanto visual como cuantitativamente.

Vamos a calcular los productos $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ correspondientes a la distribución que relaciona las notas de Matemáticas con las de Física.

$$(10 - 6) \cdot (10 - 5) = 20 \quad (10 - 6) \cdot (9 - 5) = 16 \quad (8 - 6) \cdot (7 - 5) = 4 \quad (7 - 6) \cdot (6 - 5) = 1$$

$$(7 - 6) \cdot (4 - 5) = -1 \quad (6 - 6) \cdot (6 - 5) = 0 \quad (6 - 6) \cdot (4 - 5) = 0 \quad (5 - 6) \cdot (4 - 5) = 1$$

$$(4 - 6) \cdot (4 - 5) = 2 \quad (4 - 6) \cdot (2 - 5) = 6 \quad (3 - 6) \cdot (3 - 5) = 6 \quad (2 - 6) \cdot (1 - 5) = 16$$

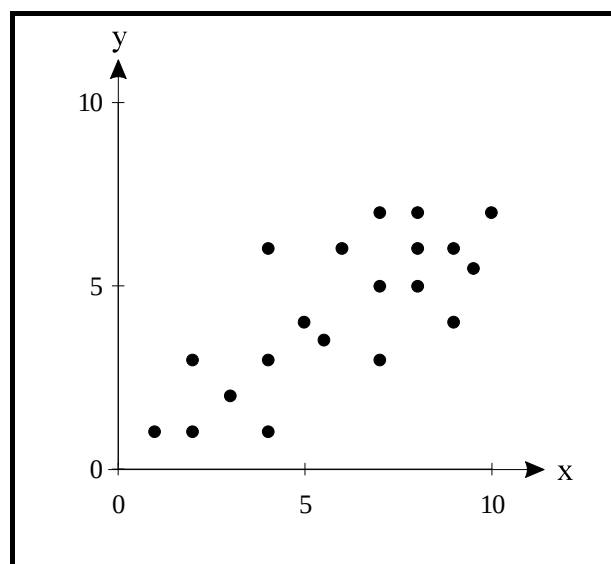
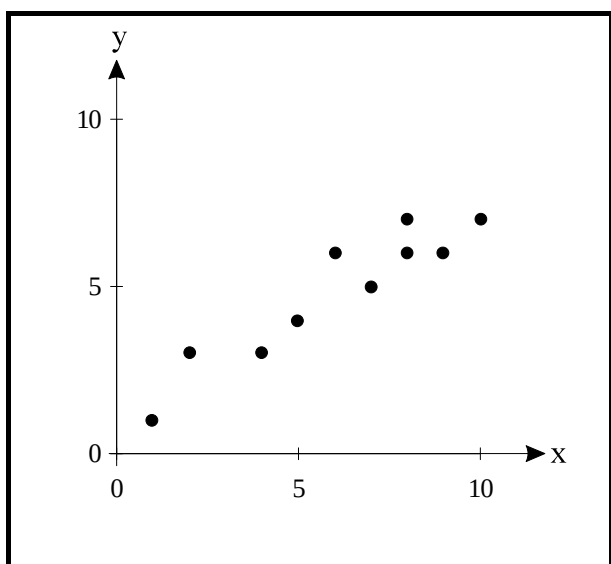
Observamos que los puntos más alejados como el (10,10) y el (2, 1) dan los productos mayores de todos.

Hasta ahora hemos razonado suponiendo la correlación positiva. Si hacemos un razonamiento similar para la correlación negativa llegaremos a la conclusión de que en estos casos la suma

$$\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

es negativa y tanto más grande en valor absoluto cuanto más estrecha sea la relación entre ambas variables.

La suma $\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$ nos ha parecido un buen indicador para saber cómo es la correlación. Sin embargo, tiene un serio inconveniente. Para apreciarlo observemos estas dos distribuciones:



Es evidente que la correlación de la primera es mayor que la de la segunda. Sin embargo, en la primera la suma $\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$ da 50, mientras que en la segunda da 80'7. Es lógico: la segunda tiene los mismos diez puntos que la primera y otros diez más. Para evitar el efecto producido por el número de puntos, $\sum f_i$, dividimos por él, obteniendo así un nuevo parámetro estadístico llamado **covarianza**.

Un nuevo parámetro estadístico. La covarianza

*Se llama covarianza de una variable bidimensional (x, y) a la media aritmética de los productos de las desviaciones de cada una de las variables respecto a sus medias respectivas. La covarianza se representa por s_{xy} . A la covarianza se le llama también **varianza conjunta** o **sincronizada** de las variables x e y.*

$$s_{xy} = \frac{\sum_{i=1}^n f_i (x_i - \bar{x})(y_i - \bar{y})}{\sum f_i} = \frac{\sum_{i=1}^n f_i x_i y_i}{\sum f_i} - \bar{x} \cdot \bar{y}$$

La deducción de la última fórmula es la siguiente:

$$s_{xy} = \frac{\sum_{i=1}^n f_i (x_i - \bar{x})(y_i - \bar{y})}{\sum f_i} = \frac{\sum_{i=1}^n f_i x_i y_i}{\sum f_i} - \bar{x} \cdot \frac{\sum_{i=1}^n f_i y_i}{\sum f_i} - \bar{y} \cdot \frac{\sum_{i=1}^n f_i x_i}{\sum f_i} + \bar{x} \cdot \bar{y} \cdot \frac{\sum f_i}{\sum f_i} =$$

$$\frac{\sum_{i=1}^n f_i x_i y_i}{\sum f_i} - \bar{x} \cdot \bar{y} - \bar{y} \cdot \bar{x} + \bar{x} \cdot \bar{y} = \frac{\sum_{i=1}^n f_i x_i y_i}{\sum f_i} - \bar{x} \cdot \bar{y}$$

Una covarianza positiva y grande significa que ambas variables se mueven al unísono creciendo o decreciendo de modo bastante acorde. Una covarianza próxima a cero significa que ambas variables funcionan independientemente. Una covarianza negativa y de valor absoluto grande significa que al crecer una variable la otra decrece.

Cálculo de parámetros estadísticos

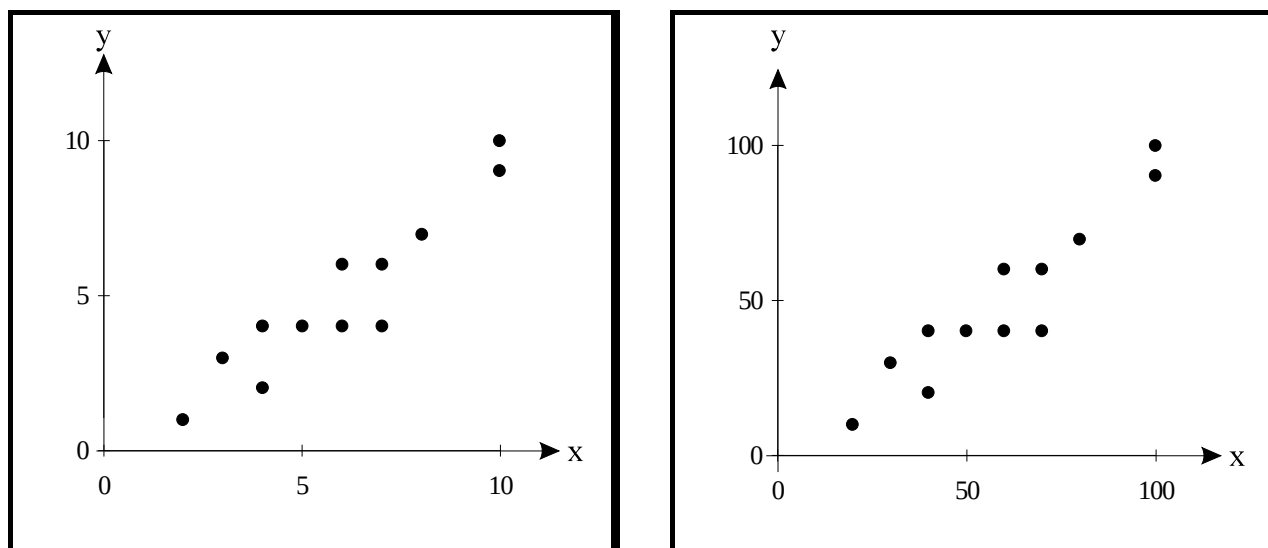
Para realizar los cálculos de los parámetros estadísticos que acabamos de estudiar, dispondremos los datos en una tabla del siguiente modo:

x_i	y_i	f_i	$x_i f_i$	$x_i^2 f_i$	$y_i f_i$	$y_i^2 f_i$	$x_i y_i f_i$
x_1	y_1	f_1	$x_1 f_1$	$x_1^2 f_1$	$y_1 f_1$	$y_1^2 f_1$	$x_1 y_1 f_1$
x_2	y_2	f_2	$x_2 f_2$	$x_2^2 f_2$	$y_2 f_2$	$y_2^2 f_2$	$x_2 y_2 f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	y_n	f_n	$x_n f_n$	$x_n^2 f_n$	$y_n f_n$	$y_n^2 f_n$	$x_n y_n f_n$
		$\sum f_i$	$\sum x_i f_i$	$\sum x_i^2 f_i$	$\sum y_i f_i$	$\sum y_i^2 f_i$	$\sum x_i y_i f_i$

Coeficiente de correlación lineal

La covarianza, que parecía un óptimo parámetro para medir la correlación presenta también algún inconveniente.

Las siguientes distribuciones son idénticas, salvo en la unidad utilizada en los ejes.



La covarianza de la primera es 5'92 y, la de la segunda, 592. El cambio de escala influye en la covarianza de tal manera, que multiplica o divide el resultado en el cuadrado de la equivalencia de la escala (si consideramos cm. en vez de m la covarianza queda multiplicada por 10.000). En nuestro ejemplo, la escala está multiplicada por 10 y sin embargo la covarianza queda multiplicada por el cuadrado, es decir, por 100.

Por eso, ese parámetro todavía no es suficientemente bueno para medir la correlación.

Para medir la correlación, definimos el **coeficiente de correlación lineal de Pearson** como:

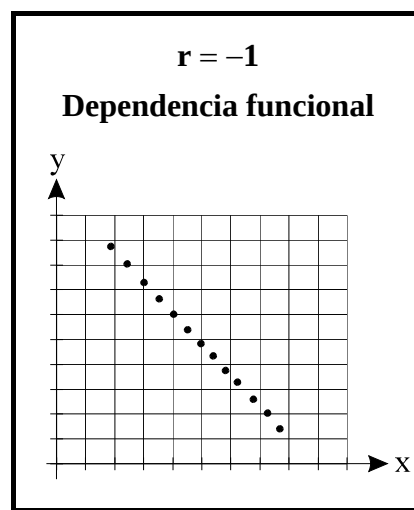
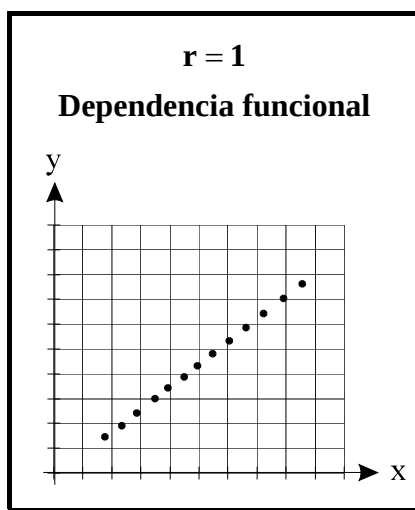
$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

que no cambia de magnitud al hacerlo la escala, pues la covarianza y las desviaciones típicas lo hacen a la vez anulándose mutuamente (se puede expresar la estatura en metros o en centímetros y el dinero en pesetas, dólares o libras y no por eso se modifica el coeficiente de correlación en el que intervenga esa variable, lo cual no pasa con la covarianza).

Propiedades fundamentales del coeficiente de correlación

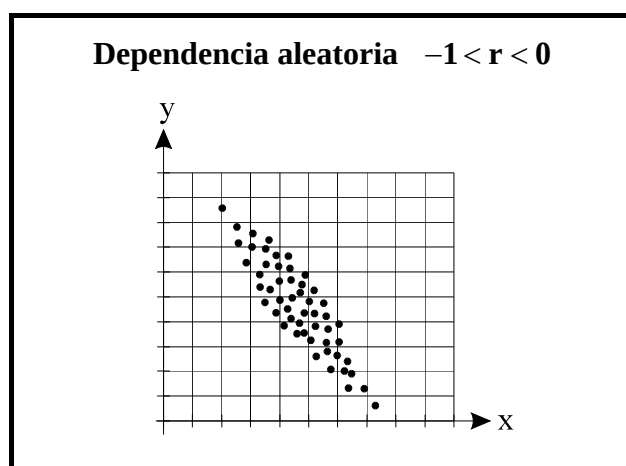
- El coeficiente es una medida objetiva de la correlación lineal entre dos variables y su valor oscila entre -1 y 1 .

- Si $r = -1$ o $r = 1$, se puede demostrar que todos los valores de la variable bidimensional (x, y) se encuentran situados sobre una recta y en consecuencia satisfacen su ecuación. La correlación entre las variables es perfecta y se dice que entre ellas existe una **dependencia funcional**.

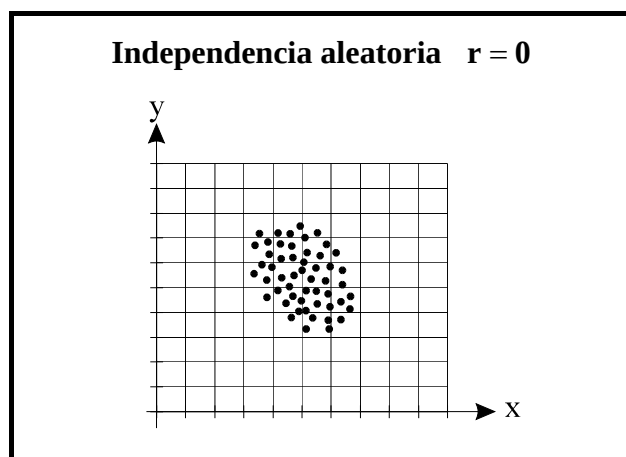


En los dos casos, una variable explica de manera absoluta a la otra (por ejemplo: importe de la factura eléctrica y kilovatios consumidos).

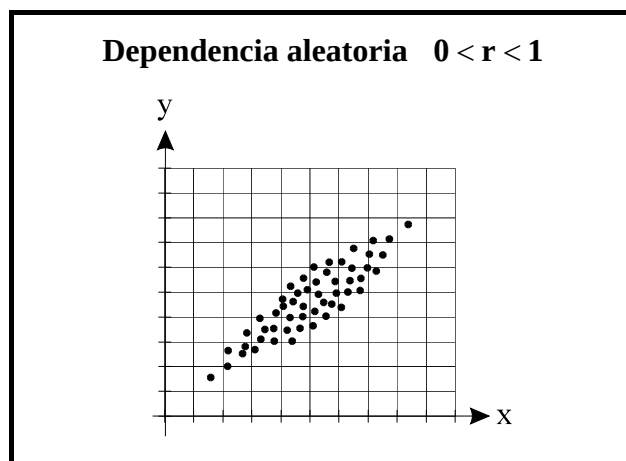
- Si $-1 < r < 0$ la correlación es negativa y será tanto más fuerte a medida que r se aproxima más a -1 y tanto más débil a medida que se aproxima a 0 . La nube de puntos es alargada y el riesgo de hacer estimaciones de una variable a partir de la otra es escaso. Se dice que las variables x e y están en **dependencia aleatoria**.



- Si $r = 0$ no existe ningún tipo de correlación entre las variables (linealmente). Este caso se asocia a nubes de puntos redondeadas. Se dice que las variables x e y son **aleatoriamente independientes**.



- Si $0 < r < 1$ la correlación es positiva y será tanto más fuerte a medida que r se aproxima más a 1 y tanto más débil a medida que se aproxima a 0. La nube de puntos es alargada y el riesgo de hacer estimaciones de una variable a partir de la otra es escaso. Se dice que las variables x e y están en **dependencia aleatoria**.



- *Si r es positivo, la correlación es directa: el aumento de una variable es seguido por el aumento de la otra. Si r es negativo, la correlación es inversa: al aumentar una variable disminuye la otra.*
- El signo de la correlación viene marcado por el de la covarianza, pues s_x y s_y son siempre positivas.

Correlación y causalidad

Muchas veces se confunde el hecho de que dos variables estén correlacionadas con el hecho de que una sea la causa de la otra.

Por ejemplo, podemos leer en un diario: "*Según un estudio hecho en los Estados Unidos, la utilización de microondas para cocinar produce anemia*". Suponiendo que el estudio esté bien hecho, y la muestra bien escogida, lo único que podemos deducir, de hecho, es que se observa una correlación positiva entre las variables "*cocinar con microondas*" y "*padecer anemia*"; pero de aquí no se puede concluir taxativamente que la anemia es la consecuencia de cocinar con microondas. Podría ser que la falta de glóbulos rojos fuese debida al tipo de alimento que el hombre acostumbra a preparar con este aparato y no al uso del aparato por sí mismo; o también, que fuesen personas más bien mayores o delicadas de salud la que mayoritariamente utilizan el microondas para cocinar.

Cualquier experimento que quiera establecer relaciones de causa efecto debe ser diseñado con mucho cuidado. Debe repetirse muchas veces en condiciones variables y debe tener en cuenta la posible influencia común a las dos variables de una tercera que no está controlada por el experimento.

Ejemplo: Calcular la correlación entre las variables *nota en Matemáticas*-*nota en Física* de la distribución dada anteriormente.

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

$$\bar{x} = \frac{72}{12} = 6 \quad \bar{y} = \frac{60}{12} = 5$$

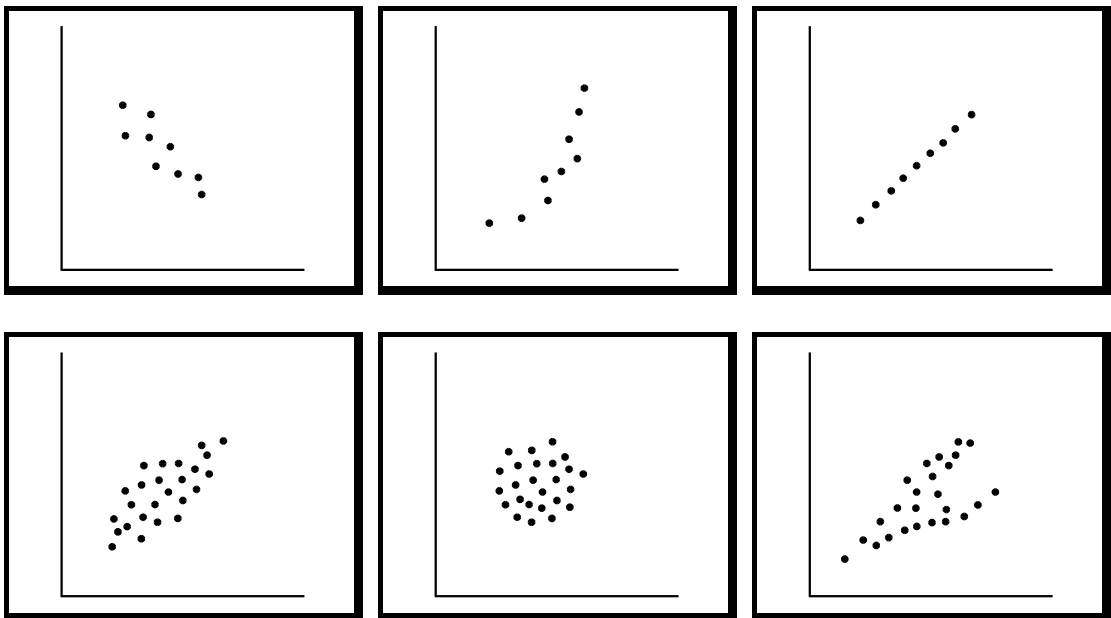
$$s_x = \sqrt{\frac{504}{12} - 6^2} = \sqrt{6} = 2'45$$

$$s_y = \sqrt{\frac{380}{12} - 5^2} = \sqrt{6'66} = 2'58$$

$$s_{xy} = \frac{431}{12} - 6 \cdot 5 = 5'92$$

$$r = \frac{5'92}{2'45 \cdot 2'58} = 0'94$$

Ejemplo: Define fenómenos que puedan adaptarse a los diagramas siguientes dados por este orden a), b), c), d), e), f):



- a) La latitud y la temperatura. A medida que la latitud de un punto de la Tierra va siendo mayor, la temperatura media de ese lugar es más baja.
- b) El consumo de tabaco y el riesgo de contraer cáncer de pulmón. La posibilidad de contraer cáncer de pulmón es pequeña para fumadores moderados y aumenta rápidamente en fumadores empedernidos.

- c) El número de kilovatios gastados y la factura de la compañía eléctrica.
- d) Los ingresos y gastos familiares. El aumento de ingresos permite que los gastos sean mayores.
- e) La simpatía personal y el nivel cultural. Hasta ahora nadie ha demostrado que exista correlación entre ambos aspectos.
- f) La altura y el peso de las personas cuando no se distingue el sexo. Los hombres suelen pesar más que las mujeres, para alturas iguales. En este caso habría que tratar a los varones y a las hembras por separado.

Ejemplo: En la siguiente tabla se dan marcas femeninas de cierta reunión de atletismo.

Distancia (m)	Tiempo
100	0'11"
200	0'22"
400	0'48"
800	1'54"
1500	3'53"

a) Calcula el coeficiente de correlación lineal entre las distancias medidas en metros y los tiempos medidos en seg.

b) Sin efectuar cálculos, razona si debemos esperar correlación positiva o negativa entre las variables velocidad y tiempo.

a)

x_i (distancias)	y_i (tiempos)	x_i^2	y_i^2	$x_i y_i$
100	11	10.000	121	1.100
200	22	40.000	484	4.400
400	48	160.000	2.304	19.200
800	114	640.000	12.996	91.200
1500	233	2.250.000	54.289	349.500
3000	428	3.100.000	70.194	465.400

$$\bar{x} = \frac{3000}{5} = 600 \quad \bar{y} = \frac{428}{5} = 85'6 \quad s_x = \sqrt{\frac{3100000}{5} - 600^2} = 509'90$$

$$s_y = \sqrt{\frac{70194}{5} - 85'6^2} = 81'92 \quad s_{xy} = \frac{465400}{5} - 600 \cdot 85'6 = 41720$$

$$r = \frac{41720}{509'901 \cdot 81'923} = 0'998$$

- b) La velocidad y el tiempo están inversamente relacionadas, luego debemos esperar correlación negativa.

Regresión

En la mayor parte de las cuestiones, el estudio de la relación entre dos variables tiene como finalidad el hacer predicciones acerca de los valores de una de las variables. El coeficiente de correlación permite estudiar la relación entre las variables, pero no resuelve el problema de la predicción.

El objetivo de la regresión es hallar una relación funcional $y = f(x)$ entre las dos variables, que nos de la posibilidad de hacer predicciones. Evidentemente no todos los puntos disponibles verificarán exactamente dicha ecuación; es más, en la mayor parte de las ocasiones ninguno de ellos la verificará. No obstante, el conjunto de ellos se ajustará de forma aproximada a dicha función.

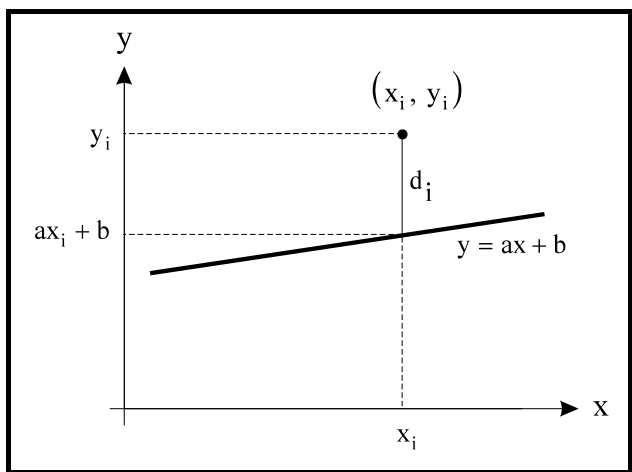
Supongamos que tenemos una distribución bidimensional y que representamos la nube de puntos correspondiente. Observamos que los puntos adoptan una forma marcadamente lineal y queremos trazar la recta que mejor se ajuste a la nube de puntos. Esta recta se llama **recta de regresión**.

Hasta ahora ese "mejor ajuste" lo hemos hecho a *ojo* poniendo, por ejemplo, una regla transparente sobre la nube de puntos, para no perder de vista ninguno, y moviéndola hasta que su posición nos parece satisfactoria. Es fácil comprender que este método es poco riguroso y muy subjetivo, dependiendo en gran medida del acierto y la pericia del sujeto que la realice.

Ahora tratamos de obtener un método más preciso, que nos permita obtener, sin ambigüedad, una única recta que cumpla la condición requerida. Hay varios métodos para buscar un buen ajuste de una nube de puntos mediante una recta, pero, **matemáticamente se demuestra que el mejor ajuste es aquél que minimiza la suma de los cuadrados de las diferencias de ordenadas entre puntos y recta.**

Supongamos que una vez estudiada la correlación existente entre las dos variables x e y que componen una variable bidimensional (x, y) se observa que dichas variables están fuertemente correlacionadas y que el diagrama de puntos se puede ajustar mediante una recta.

Consideremos x como la variable independiente e y como la variable dependiente de x . Entonces el problema consiste en encontrar la ecuación de una recta de la forma $y = ax + b$ que sea la que mejor se ajuste a la nube de puntos.

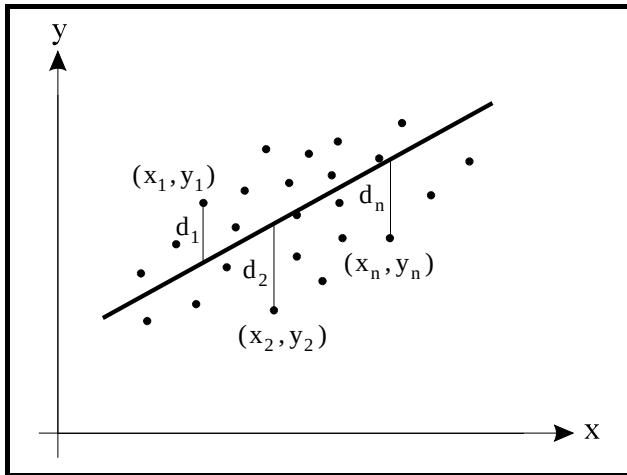


Para cada valor x_i , la correspondiente ordenada en la recta será $ax_i + b$. La diferencia

$$d_i = y_i - (ax_i + b)$$

se denomina **desviación o error**, que puede ser positivo, negativo o nulo.

Evidentemente $d_i = 0 \Leftrightarrow y_i = ax_i + b$



Con los puntos:

$$(x_1, y_1), \dots, (x_n, y_n)$$

se obtienen n-desviaciones.

$$d_1 = y_1 - (ax_1 + b); d_2 = y_2 - (ax_2 + b) \dots\dots\dots$$

$$d_n = y_n - (ax_n + b)$$

Intuitivamente se ve que la recta estará mejor ajustada cuanto más próximos estén de ella los puntos, es decir, cuanto más pequeñas sean las desviaciones en valor absoluto.

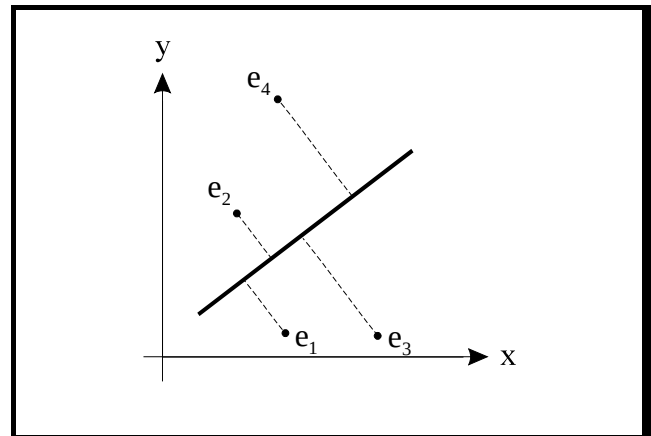
Consideremos la expresión: $D = d_1^2 + d_2^2 + \dots + d_n^2$.

Se observa que $D = 0 \Leftrightarrow d_i^2 = 0 \quad \forall i \Leftrightarrow$ los puntos (x_i, y_i) están sobre la recta.

Cuanto mayor sea D tanto peor será el ajuste, luego nos interesa minimizar D (problema de máximos y mínimos condicionados), ya que para conseguir una aproximación óptima entre la recta y el conjunto de puntos, interesa que D sea lo menor posible.

Se llama recta de regresión de mínimos cuadrados de los puntos $(x_1, y_1), \dots, (x_n, y_n)$ a aquella para la cual la expresión $D = d_1^2 + d_2^2 + \dots + d_n^2$ sea mínima.

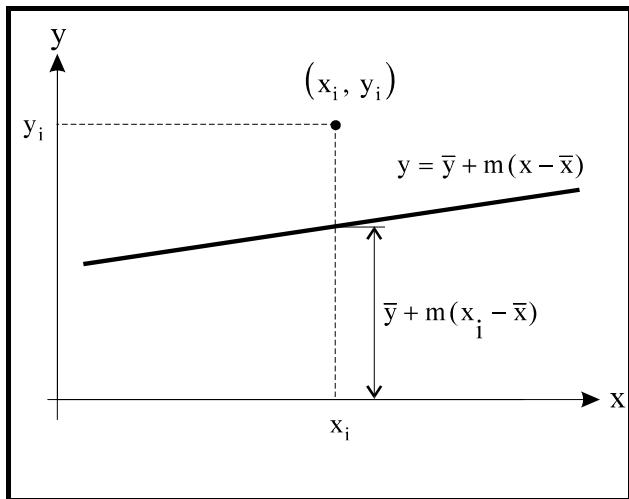
La razón por la que se toman las desviaciones o errores al cuadrado es para evitar las posibles compensaciones entre errores positivos y negativos al considerar la totalidad de los puntos, pues podemos encontrar situaciones como la de la figura. La suma $\sum_{i=1}^n e_i$ es prácticamente nula, con lo que obtendríamos conclusiones erróneas en la mayor parte de los casos.



Vamos a buscar la recta que mejor se ajuste a la nube de puntos siguiendo el siguiente criterio:

a) La recta debe pasar por el punto $(\bar{x}, \bar{y})$.

b) Los cuadrados de las diferencias de ordenadas entre puntos y recta deben sumar lo menos posible.



La condición a) impone que la recta tenga la ecuación $y - \bar{y} = m(x - \bar{x})$. La pendiente se calcula teniendo en cuenta la condición b).

Si consideramos un punto (x_i, y_i) implica que la ordenada de la recta para la abscisa x_i será $\bar{y} + m(x_i - \bar{x})$

La diferencia de ordenadas entre el punto y la recta para la misma abscisa será:

$$y_i - [\bar{y} + m(x_i - \bar{x})] = (y_i - \bar{y}) - m(x_i - \bar{x})$$

La condición b) dice que: $\sum_{i=1}^n [(y_i - \bar{y}) - m(x_i - \bar{x})]^2$ debe ser mínimo.

En esta expresión $x_i, y_i, \bar{x}$ e $\bar{y}$ son números, y sólo es variable la m . Hemos de encontrar el valor de m para el cual la función:

$$D(m) = \sum_{i=1}^n [(y_i - \bar{y}) - m(x_i - \bar{x})]^2$$

llamada variación total presenta un mínimo.

$D(m)$ será mínimo $\Leftrightarrow D'(m) = 0$.

$$D'(m) = 2 \cdot \sum_{i=1}^n [(y_i - \bar{y}) - m(x_i - \bar{x})] \cdot (x_i - \bar{x})(-1) = 0 \Rightarrow \sum_{i=1}^n [(y_i - \bar{y}) \cdot (x_i - \bar{x}) - m(x_i - \bar{x})^2] = 0$$

$$\sum_{i=1}^n (y_i - \bar{y}) \cdot (x_i - \bar{x}) - m \sum_{i=1}^n (x_i - \bar{x})^2 = 0 \Leftrightarrow m = \frac{\sum_{i=1}^n (y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Dividiendo por $\sum f_i$ arriba y abajo tenemos: $m = \frac{s_{xy}}{s_x^2}$

luego la recta para la cual $D(m)$ es mínimo es aquella para la cual la pendiente es $m = \frac{s_{xy}}{s_x^2}$, es decir:

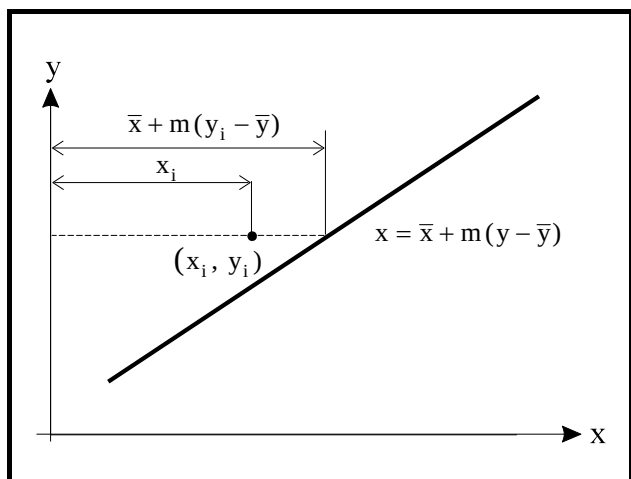
$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

y se llama **recta de regresión** de y sobre x .

A la pendiente $\frac{s_{xy}}{s_x^2}$ se le llama **coeficiente de regresión de y sobre x** y se representa por m_{yx} .

A esta recta de regresión se le llama **recta de regresión de y sobre x** , ya que **hemos considerado la variable x independiente y la variable y dependiente**. Al calcular la recta de regresión hay que tener absolutamente clara la noción de cuál es la variable independiente (x) y cuál es la variable dependiente (y). No son intercambiables; una cosecha puede depender de la lluvia caída, pero la lluvia no depende de la cosecha.

Análogamente se puede obtener la **recta de regresión de x sobre y** . En este caso la variable independiente es y , siendo x la variable dependiente de y .



La ecuación de la recta será de la forma:

$$x = \bar{x} + m(y - \bar{y})$$

Diferencia de abscisas:

$$x_i - [\bar{x} + m(y_i - \bar{y})]$$

Variación total:

$$D(m) = \sum_{i=1}^n [(x_i - \bar{x}) - m(y_i - \bar{y})]^2$$

La ecuación de la recta de regresión de x sobre y es de la forma:

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

Observaciones

1. Cuando la correlación es fuerte (es decir, cuando $|r|$ es grande), las dos rectas están muy próximas y coinciden, a su vez, con la recta que trazamos a ojo. Sin embargo, si la correlación es floja, las dos rectas forman un ángulo grande.
2. La recta que trazamos a ojo pretendiendo seguir la tendencia de la nube, está entre medias de ambas. Es aproximadamente su bisectriz.
3. Existe un procedimiento (**TEST**) encaminado a valorar hasta qué punto es aceptable la hipótesis de linealidad. Cuando la disposición de los puntos es tal que no resulta plausible la idea de que se amolde a una recta, el test lo pone de manifiesto y se rechaza la hipótesis de linealidad.

Uso de la calculadora

Los cálculos de la recta de regresión y del coeficiente de correlación lineal se automatizan con una calculadora que tenga análisis de regresión. Hay que colocarla en el **modo de regresión lineal** (en muchas **LR**, iniciales del inglés "linear regression") y **borrar el contenido de la memoria (MR) así como el número de datos (n)**.

Vamos a hacer los cálculos ayudándonos de un ejemplo:

Un país ha registrado las rentas y gastos anuales siguientes en billones de pesetas:

Año	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973
Renta = x_i	0'9	1'1	1'2	1'3	1'5	1'6	1'9	2	2'5	2'7
Gasto = y_i	0'6	0'7	0'7	0'8	0'9	1	1'1	1'3	1'5	1'7

Introduciremos los datos así:

0'9 x_D, y_D 0'6 DATA 1'1 x_D, y_D 0'7 DATA 1'2 x_D, y_D 0'7 DATA 1'3
 x_D, y_D 0'8 DATA 1'5 x_D, y_D 0'9 DATA 1'6 x_D, y_D 1 DATA 1'9 x_D, y_D
1'1 DATA 2 x_D, y_D 1'3 DATA 2'5 x_D, y_D 1'5 DATA 2'7 x_D, y_D 1'7 DATA

La tecla x_D, y_D enlaza los pares de datos. La tecla DATA los introduce en la memoria.

- Si la frecuencia no es unitaria y queremos introducir por ejemplo 15 veces el dato $x = 2$ e $y = 7$ operaremos así:

2 x_D, y_D 7·15 DATA

Si ahora queremos borrar 3 de esos datos escribiremos:

$$2 \boxed{x_D, y_D} 7 \cdot 3 \boxed{\text{SHIFT}} \boxed{\text{DEL}}$$

El registro $\boxed{\text{DEL}}$ está en la misma tecla que $\boxed{\text{DATA}}$

- Si queremos introducir $x_1 = 1$ e $y_1 = 3$ y por error tecleamos $2 \boxed{x_D, y_D}$ entonces con la tecla $\boxed{\text{C}}$ volvemos al 0 y podemos volver a teclear.

Si advertimos el error cuando ya hemos introducido en memoria el par de datos, tendremos que borrar ese dato de la memoria de la siguiente manera:

$$2 \boxed{x_D, y_D} 3 \boxed{\text{SHIFT}} \boxed{\text{DEL}}$$

❑ Cálculo de la recta de regresión de y sobre x

Para obtener la constante **a** de la recta de regresión lineal $y = ax + b$ basta con pulsar la tecla $\boxed{\text{B}}$ precedida de la tecla $\boxed{\text{SHIFT}}$, para obtener la constante **b** basta con pulsar la tecla $\boxed{\text{A}}$ precedida de la tecla $\boxed{\text{SHIFT}}$ y para obtener el coeficiente de correlación lineal **r** basta con pulsar la tecla $\boxed{\text{r}}$ precedida de la tecla $\boxed{\text{SHIFT}}$.

En el ejemplo se obtiene:

$$\begin{array}{ccccc} \boxed{\text{SHIFT}} & \boxed{\text{A}} & \boxed{\text{SHIFT}} & \boxed{\text{B}} & \boxed{\text{SHIFT}} & \boxed{\text{r}} \\ \downarrow & & \downarrow & & \downarrow & \\ b = 9'127 \cdot 10^{-3} & a = 0'6113 & r = 0'99287 \end{array}$$

Cálculo de la recta de regresión de x sobre y

El proceso de cálculo es el mismo que para la recta de regresión de y sobre x, pero en este caso hay que introducir los datos al revés, es decir, primero los valores de la variable y y luego los valores de la variable x. En nuestro ejemplo:

$$0'6 \boxed{x_D, y_D} 0'9 \boxed{\text{DATA}} 0'7 \boxed{x_D, y_D} 1'1 \boxed{\text{DATA}} \dots\dots\dots$$

❑ Cálculo de la media y desviación típica de ambas variables.

$$\begin{array}{cccc} \boxed{\text{SHIFT}} & \boxed{\bar{x}} & \boxed{\text{SHIFT}} & \boxed{x\sigma_n} & \boxed{\text{Kout}} & \boxed{\sum x^2} & \boxed{\text{Kout}} & \boxed{\sum x} \\ \downarrow & & \downarrow & & & \downarrow & & \downarrow \\ 1'67 & & 0'56753 & & & 31'11 & & 16'7 \end{array}$$

SHIFT	$\bar{y}$	SHIFT	$y\sigma_n$	Kout	$\sum y^2$	Kout	$\sum y$
	↓		↓		↓		↓
	1'03		0'3494		11'83		10'3

donde los registros $\bar{x}$ e $\bar{y}$ representan la media aritmética de las variables x e y respectivamente, y donde $x\sigma_n$ y $y\sigma_n$ representan las desviaciones típicas, que en los apuntes vienen simbolizadas por s_x y s_y .

❑ Cálculo de los productos $x_i y_i$ y del número de datos $\sum f_i$

Kout	$\sum xy$	Kout	n
	↓		↓
	19'17		10

❑ Cálculo de $f(x_i)$

La calculadora proporciona el valor $f(x_i)$, que es la aproximación a y_i , con las instrucciones:

$$x_i \quad \hat{y}$$

Por ejemplo obtendríamos $f(0'9)$, $f(1'1)$ y $f(1'2)$ así:

0'9	$\hat{y}$	1'1	$\hat{y}$	1'2	$\hat{y}$
	↓		↓		↓
	0'559		0'682		0'743

Con la renta de 3 billones el valor de $f(3)$ será:

$$3 \quad \hat{y} \\ \downarrow \\ 1'843$$

que será una predicción o estimación aproximada del gasto, suponiendo que la función de regresión lineal continúa aproximándose bien a los valores y_i fuera del intervalo donde se tienen los datos.

También da la calculadora el valor de x que hace que $f(x)$ sea igual a cierto número. Para calcular para que valor de x es $f(x) = 1'5$ operaremos de la siguiente manera:

$$1'5 \quad \text{SHIFT} \quad \hat{x} \\ \downarrow \\ 2'438$$

La recta de regresión para hacer predicciones

¿Qué fiabilidad podemos conceder a estos cálculos obtenidos a través de las rectas de regresión? La fiabilidad será tanto mayor cuanto mayor sea el coeficiente de correlación lineal en valor absoluto. Así pues:

- Si r es muy pequeño, *no tiene sentido realizar ningún tipo de estimaciones o predicciones.*
- Si r es próximo a -1 o a 1 , *probablemente los valores serán próximos a nuestras estimaciones.*
- Si $r = -1$ o $r = 1$, *las estimaciones realizadas coincidirán con los valores reales.*

Pero incluso para valores de r próximos a uno, las estimaciones que obtengamos pueden resultar poco fiables; por ejemplo, cuando se pretende **extrapolar** más allá del recinto de los datos observados. Es cierto que muchas veces es necesario realizar extrapolaciones, pero es conveniente tener en cuenta siempre el elevado riesgo que se corre de obtener resultados totalmente erróneos.

Ejemplo: Supongamos que el número de casos de SIDA detectados desde su aparición hasta el momento tiene una fuerte correlación con el número de individuos fallecidos. En el supuesto de que dentro de 15 años se presenten 5000 casos, ¿se puede predecir cuántos individuos fallecerán de esos 5000?

No es posible contestar afirmativamente, ya que es fácil que durante ese período se haya podido obtener algún tipo de tratamiento que haga disminuir la mortalidad.

La recta de regresión sirve para predecir la variable dependiente (o regresor), y por eso suele hablarse de recta de regresión de y sobre x .

A partir de la recta de regresión de y sobre x podemos calcular con cierta aproximación los valores de y conocidos los de x , sin más que sustituir estos últimos en la ecuación. A estos cálculos se les suele llamar estimaciones o predicciones. A partir de la recta de regresión de x sobre y podemos calcular con cierta aproximación los valores de x conocidos los de y , sin más que sustituir estos últimos en la ecuación.

Ejemplo: Dada la tabla adjunta, donde la variable x significa el número de hermanos de una persona y la variable y representa el número de hermanos de su padre, calcular:

- El coeficiente de correlación.
- Las rectas de regresión de Y/X y de X/Y .
- Si Pedro tiene 3 hermanos ¿cuántos tenía su padre? Si Juan tiene 7 hermanos ¿cuántos tenía su padre?
- Si el padre de Jaime tiene 6 hermanos ¿cuántos tendrá Jaime?

x	y
0	1
0	3
2	3
2	4
4	5
6	6
6	7
6	5
8	2

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
0	1	0	1	0
0	3	0	9	0
2	3	4	9	6
2	4	4	16	8
4	5	16	25	20
6	6	36	36	36
6	7	36	49	42
6	5	36	25	30
8	2	64	4	16
34	36	196	174	158

$$\bar{x} = \frac{34}{9} = 3'78 \quad \bar{y} = \frac{36}{9} = 4$$

$$s_{xy} = \frac{158}{9} - 3'78 \cdot 4 = 2'44$$

$$s_x^2 = \frac{196}{9} - 3'78^2 = 7'49$$

$$s_y^2 = \frac{174}{9} - 4^2 = 3'33$$

Coefficiente de correlación

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{2'44}{\sqrt{7'49} \cdot \sqrt{3'33}} = \mathbf{0'4885}$$

Recta de regresión de y sobre x

$$y - \bar{y} = \frac{s_{xy}}{s_x^2} \cdot (x - \bar{x}) \rightarrow y - 4 = \frac{2'44}{7'49} \cdot (x - 3'78) \rightarrow y = \mathbf{0'32x + 2'79}$$

Esta ecuación nos indica cuántos hermanos tenía el padre de una persona que tiene x hermanos.

Recta de regresión de x sobre y

$$x - \bar{x} = \frac{s_{xy}}{s_y^2} \cdot (y - \bar{y}) \rightarrow x - 3'78 = \frac{2'44}{3'33} \cdot (y - 4) \rightarrow x = \mathbf{0'73y + 0'86}$$

Esta ecuación nos indica cuántos hijos es previsible que tenga un hombre que tiene y hermanos.

Estimaciones o predicciones

Si Pedro tiene 3 hermanos su padre tenía 4, como se comprueba a continuación.

$$x = 3 \Rightarrow y = 0'32 \cdot 3 + 2'79 = 3'75 \Rightarrow y \cong 4$$

Si Juan tiene 7 hermanos, su padre tenía 5:

$$x = 7 \Rightarrow y = 0'32 \cdot 7 + 2'79 = 5'03 \Rightarrow y \cong 5$$

Si el padre de Jaime tiene 6 hermanos, Jaime tendrá 5.

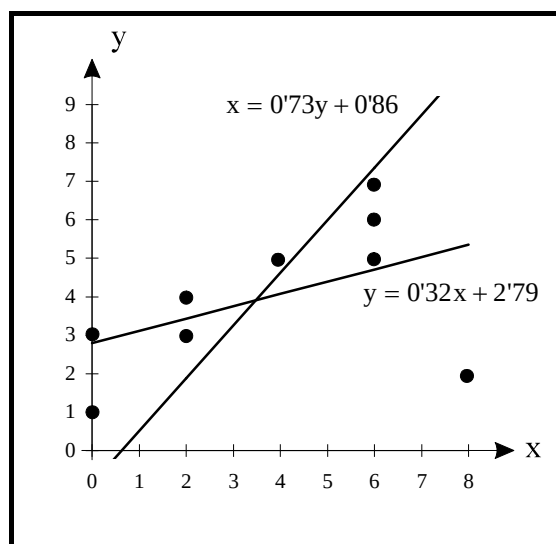
$$y = 6 \Rightarrow x = 0'73 \cdot 6 + 0'86 = 5'24 \Rightarrow x \cong 5$$

¿Qué fiabilidad hay en estas estimaciones? En este caso, en el que el coeficiente de correlación es 0'4885, no hay garantía de que las predicciones sean muy fiables. Sin embargo, en distribuciones cuya correlación sea próxima a 1 o a -1, probablemente los valores coinciden con las predicciones.

Es interesante observar, que en los datos del problema aparece un punto (8,2) bastante alejado de la recta de regresión que hace que el ajuste de esta recta sea inferior al que obtendríamos si dicho punto no estuviera.

Estos puntos especiales, cuando no son errores de medida, suelen tener explicaciones particulares que hacen que se puedan excluir de los datos iniciales, en caso de que convenga, sin pérdida de generalidad.

Por ejemplo, en nuestro caso podría ser que el hombre estuviera casado en segundas nupcias, de donde se deduce que el número de hijos aumentase, a pesar de que tuviera pocos hermanos.



En general, el método de los mínimos cuadrados da lugar a una recta que tiene muy en cuenta los puntos alejados, es decir, el ajuste resultante favorece la predicción de estos puntos, en detrimento de los más cercanos. Es pues importante asegurarse de que no se trata de un error de medida y debe estudiarse si es o no conveniente su inclusión.

Para calcular el valor de Y conocido el de X, hay que hacerlo en la recta de regresión Y/X porque se trata de hacer mínima la suma de los cuadrados de las diferencias entre las **ordenadas** de los puntos y las correspondientes de la recta de regresión, mientras que para calcular el valor de X conocido el de Y, hay que hacerlo en la recta de regresión X/Y porque se trata de hacer mínima la suma de los cuadrados de las diferencias entre las **abscisas** de los puntos y las correspondientes de la recta de regresión.

Analogías y diferencias entre correlación y regresión

Las técnicas de correlación y regresión lineal tienen ciertas analogías y diferencias, aunque estas últimas son más acusadas que las primeras.

Analogías

- Tanto la correlación como la regresión lineal estudian conjuntamente dos variables y tratan de averiguar si existe o no relación entre ellas. Así, por ejemplo, en el caso lineal (el único que hemos estudiado) el signo de r es el mismo que el de la pendiente de la recta de regresión, pero su valor no es igual. Es más, podemos estar ante un caso en el que r tiene un valor grande y la pendiente uno pequeño o viceversa. Sólo coinciden en el caso en que $s_x = s_y$, conclusión fácilmente deducible si se observan las expresiones de r y la pendiente.

$$r = \frac{s_{xy}}{s_x \cdot s_y} \qquad m_{yx} = \frac{s_{xy}}{s_x^2} \qquad m_{xy} = \frac{s_{xy}}{s_y^2}$$

- La recta de regresión se puede obtener siempre, pero sólo es útil en el caso de que $|r|$ sea grande y x sea próximo a $\bar{x}$.

Diferencias

- Dado que las dos técnicas parten de puntos de vista diferentes, nos facilitan distinta información.
- La correlación mide el grado de covariación entre las dos variables sin exigir ningún tipo de relación entre ellas.
- La regresión nos proporciona una relación funcional entre las variables y por tanto una relación causal y lo más importante, y aquí radica la diferencia fundamental, nos permite predecir la variable dependiente en función de la independiente.
- La elección entre una técnica y otra dependerá, por tanto, del tipo de estudio que se desee realizar.

Otras regresiones basadas en la regresión lineal

Directamente basadas en la regresión lineal estudiada, podemos intentar el ajuste a los datos de otro tipo de curvas menos simples. Son las regresiones logarítmica, exponencial y potencial.

Regresión logarítmica

Si dados los datos (x_i, y_i) por la tabla, introducimos los valores $(\ln x_i, y_i)$, la recta de regresión obtenida para la nube de puntos es, en consecuencia:

$$y = a + b \cdot \ln x$$

Se utiliza, cuando los datos y_i experimentan un crecimiento (o decrecimiento) **suave** con respecto al crecimiento de los datos x_i . Plantea problemas si los valores x_i están próximos a 0 y no tiene sentido si son negativos.

Regresión exponencial

Si dados los datos (x_i, y_i) , introducimos los valores $(x_i, \ln y_i)$, la recta obtenida es $\ln y = a + bx$.

De aquí, despejando y tenemos: $y = e^{a+bx} = e^a \cdot e^{bx}$, y tomando $A = e^a$ resulta:

$$y = A \cdot e^{bx}$$

regresión que ajusta bien los casos de crecimiento (o decrecimiento) **rápido** de los valores y_i respecto de los valores x_i .

Regresión potencial

Si dados los datos (x_i, y_i) , introducimos los valores $(\ln x_i, \ln y_i)$, la recta de regresión obtenida es $\ln y = a + b \cdot \ln x$. De aquí, despejando y y aplicando propiedades del logaritmo:

$$y = e^{a+b \cdot \ln x} = e^a \cdot e^{b \cdot \ln x} = e^a \cdot (e^{\ln x})^b = e^a \cdot x^b$$

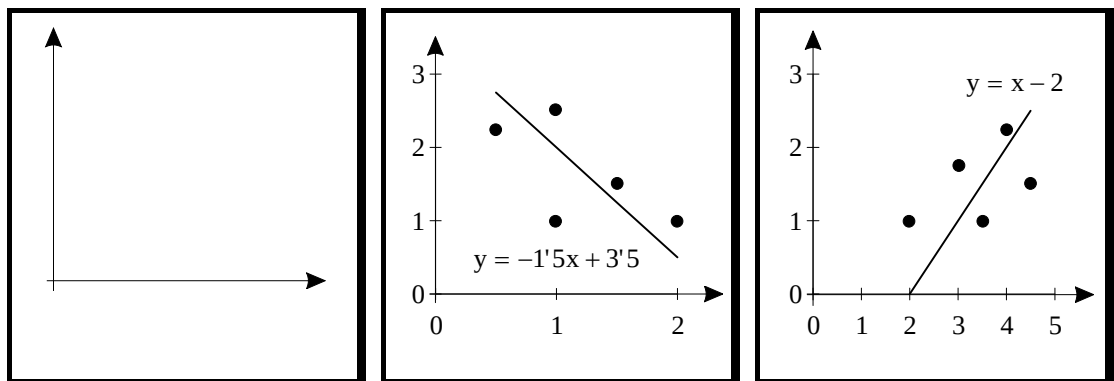
Tomando $A = e^a$, resulta:

$$y = A \cdot x^b$$

Para determinar si es adecuado alguno de estos ajustes se utiliza el papel semilogarítmico (en el que sobre uno de los ejes se toman medidas en logaritmos) y el papel log-log (en el que las medidas en logaritmos son tomadas sobre los ejes).

Ejemplo: Asocia razonadamente las rectas de regresión y las nubes de puntos siguientes:

$$y = x - 2; \quad y = -1'5x + 3'5; \quad y = 0'5x + 1$$



Ejemplo: Un conjunto de datos bidimensionales (x,y) tiene coeficiente de correlación $r = 0'8$ y las medias respectivas son $\bar{x} = 3$, $\bar{y} = 10$. Sin efectuar cálculos, razona por qué las siguientes ecuaciones no pueden corresponder a la recta de regresión de y sobre x .

$$y = -2x + 16; \quad y = 1'5x + 1; \quad y = -3'5x - 1$$

Sabemos que el signo de la pendiente de la recta de regresión es el mismo que el de la correlación. Por otra parte, la recta de regresión debe contener al punto $(\bar{x}, \bar{y})$.

Como la correlación es positiva, la primera y tercera rectas hay que descartarlas. Nos queda como posible sólo la segunda recta.

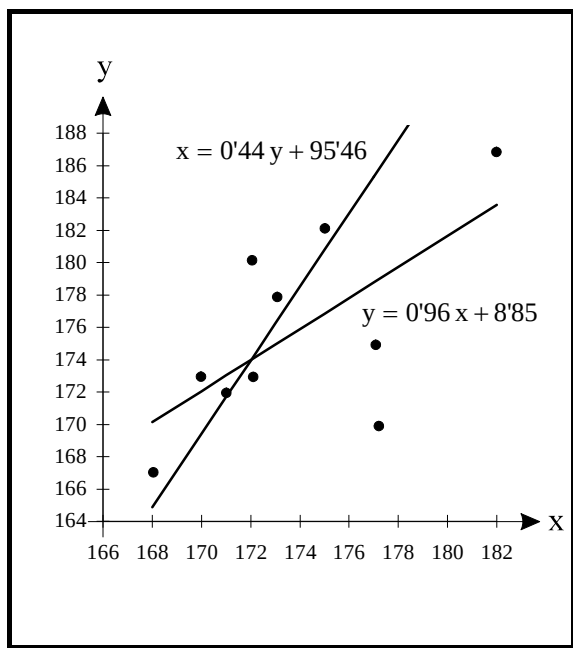
Como el punto $(3,10)$ no pertenece a ella, tampoco puede ser la recta de regresión.

Ejemplo: En la siguiente tabla se dan las alturas de diez padres y del primer hijo varón de cada uno de ellos:

Padre(x)	170	168	172	177	182	177	175	171	173	172
Hijo(y)	173	167	173	170	187	175	182	172	178	180

Dibuja el diagrama de dispersión y las recta de regresión de y sobre x y de x sobre y .

x	y	x^2	y^2	xy
170	173	28900	29929	29410
168	167	28224	27889	28056
172	173	29584	29929	29756
177	170	31329	28900	30090
182	187	33124	34969	34034
177	175	31329	30625	30975
175	182	30625	33124	31850
171	172	29241	29584	29412
173	178	29929	31684	30794
172	180	29584	32400	30960
1737	1757	301869	309033	305337



$$\bar{x} = \frac{1737}{10} = 173'7 \quad \bar{y} = \frac{1757}{10} = 175'7$$

$$s_x^2 = \frac{301869}{10} - 173'7^2 = 15'21$$

$$s_y^2 = \frac{309033}{10} - 175'7^2 = 32'81$$

$$s_x = 3'9 \quad s_y = 5'728$$

$$s_{xy} = \frac{305337}{10} - 173'7 \cdot 175'7 = 14'61$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{14'61}{3'9 \cdot 5'728} = 0'654$$

La recta de regresión de y sobre x es:

$$y - \bar{y} = \frac{s_{xy}}{s_x^2} (x - \bar{x}) \rightarrow y - 175'7 = \frac{14'61}{15'21} (x - 173'7) \rightarrow y = 0'96x + 8'852$$

La recta de regresión de x sobre y es:

$$x - \bar{x} = \frac{s_{xy}}{s_y^2} (y - \bar{y}) \rightarrow x - 173'7 = \frac{14'61}{32'81} (y - 175'7) \rightarrow x = 0'445y + 95'462$$

Ejemplo: Una compañía de seguros considera que el número de vehículos (y) que circulan por la autopista Vasco-Aragonesa puede ponerse en función del número de accidentes (x) que ocurren en ella. Durante cinco días obtuvo los siguientes resultados:

x (en unidades)	5	7	2	1	9
y (en miles)	15	18	10	8	20

a) Calcula el coeficiente de correlación lineal.

b) Si ayer se produjeron seis accidentes, ¿cuántos vehículos podemos suponer que circulaban por la autopista?

c) ¿Es buena esta predicción? Razona la respuesta.

x	y	x ²	y ²	xy
5	15	25	225	75
7	18	49	324	126
2	10	4	100	20
1	8	1	64	8
9	20	81	400	180
24	71	160	1113	409

$$\bar{x} = \frac{24}{5} = 4'8 \quad \bar{y} = \frac{71}{5} = 14'2$$

$$s_x^2 = \frac{160}{5} - 4'8^2 = 8'96 \rightarrow s_x = 2'993$$

$$s_y^2 = \frac{1113}{5} - 14'2^2 = 20'96 \rightarrow s_y = 4'57$$

$$s_{xy} = \frac{409}{5} - 4'8 \cdot 14'2 = 13'64$$

a) El coeficiente de correlación es: $r = \frac{s_{xy}}{s_x s_y} = 0'995$

b) Hallamos la recta de regresión del número de vehículos sobre el de accidentes.

$$y - \bar{y} = \frac{s_{xy}}{s_x^2} (x - \bar{x}) \rightarrow y - 14'2 = \frac{13'64}{8'98} (x - 4'8) \rightarrow y = 1'52x + 6'89$$

Si el número de accidentes de ayer fue de 6, podemos suponer que circularon:

$$y = 1'52 \cdot 6 + 6'89 = 16.000 \text{ vehículos}$$

c) Como la correlación es muy alta, la predicción puede considerarse muy buena.

Ejemplo: Considera el siguiente conjunto de datos bidimensionales:

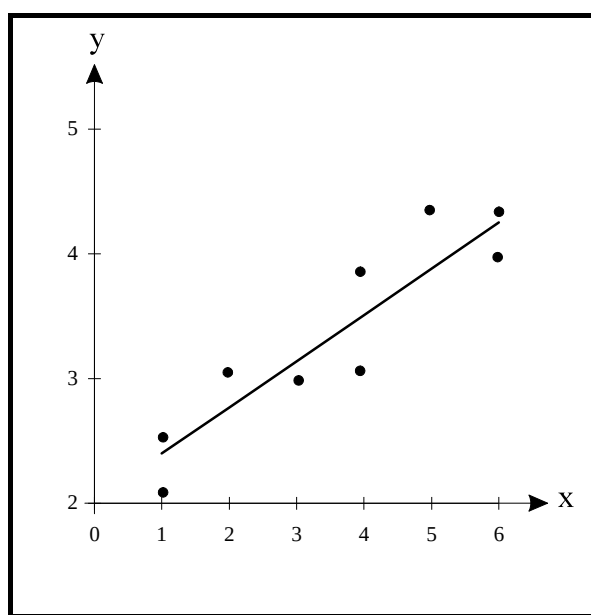
x	1	1	2	3	4	4	5	6	6
y	2'1	2'5	3'1	3	3'8	3'2	4'3	3'9	4'4

a) Sin efectuar ningún cálculo razona cuál de estos valores es su coeficiente de correlación: 0'3, -0'9, -0'1, 0'92, y cuál de las siguientes rectas es la de regresión de y sobre x:

$$y = 2'03 + 0'37x; \quad y = 5'53 + 0'37x; \quad y = 2'3 - 1'37x; \quad y = 2'03 - 0'72x$$

b) Para el valor $x = 3'5$, ¿qué predicción de la variable y es razonable efectuar?

a) Representamos la nube de puntos:



La correlación es directa, luego su coeficiente tiene que ser positivo; además, por el alargamiento de la nube de puntos, debe ser alto. Por tanto tiene que ser 0'92.

Por disposición de la nube de puntos, la recta debe tener pendiente positiva y su ordenada en el origen debe estar en torno a dos. La única posible es:

$$y = 2'03 + 0'37x$$

b) Para $x = 3'5$, y debe tomar un valor próximo a $y = 2'03 + 0'37 \cdot 3'5 = 3'325$