R de Pearson

para dos variables cuantitativas

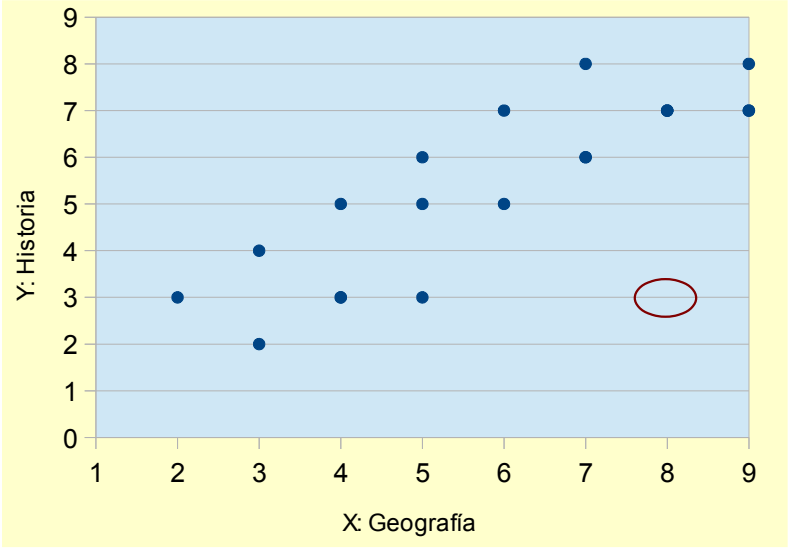
Vicente Manzano Arrondo – 2014

X	Y
5	6
6	5
6	7
7	6
2	3
8	7
9	7
4	3
8	7
9	8
9	7
8	7
7	6
5	5
3	2
4	3
3	4
5	3
4	5
7	8

Tenemos una muestra aleatoria de expedientes académicos de estudiantes de enseñanzas medias. Tomamos sus resultados en las materias de geografía (variable X) y de historia (variable Y). Nos interesa conocer si existe relación entre ambas. Tal vez no sea el objetivo de tu vida, pero se trata de dos variables claramente cuantitativas, medidas ambas en el intervalo habitual de 0 a 10 y, por tanto, muy apropiadas para estudiar cómo se representa gráficamente y cómo se cuantifica una relación entre dos variables cuantitativas.

Estudio de la relación. Diagrama de dispersión

Una representación gráfica idónea es tomar unos ejes cartesianos, ubicar una variable en el eje horizontal y otra en el vertical. Cada estudiante (su par de calificaciones en geografía es historia) será un punto en la gráfica. Ese punto expresa una coordenada, posición o valor en X (puntuación en geografía) y otra en Y (puntuación en historia). Los datos se encuentran a la izquierda de este párrafo.



Observa, por ejemplo, el punto que está rodeado por una elipse roja. Representa a un estudiante que ha obtenido un 5 en geografía y un 3 en historia. Aunque hay 20 estudiantes, podrás contar solo 16 puntos. Esto ocurre porque hay superposiciones: en 4 ocasiones hay coordenadas repetidas. Ocurre, por ejemplo con el par X:9, Y:7.

Esta representación gráfica tiene por nombre *diagrama* y por apellido *de dispersión*, puesto que expresa el grado en que los datos están dispersos a lo largo y alto de la superficie del gráfico. Para que resulte útil, es necesario que exista realmente cierta dispersión. Si existiera mucha coincidencia entre pares de datos, el gráfico no mostraría lo que ocurre en realidad. Nos faltaría representar una tercera dimensión, algo así como la profundidad, indicando el grado en que un mismo punto es visitado por más o menos pares de datos. Por otro lado, también se necesita cierta dispersión o variabilidad en cada uno de los dos ejes. Si contamos, por ejemplo, con una variable con pocos valores (imagina que 4 o 5), aunque se tratara de una variable cuantitativa, no nos serviría para un diagrama de dispersión porque no habría realmente dispersión en la superficie sino básicamente coincidencias, especialmente si el tamaño de la muestra es grande. Así que no nos enteraríamos bien de lo que esté ocurriendo. En tales casos, es preferible acudir a una tabla de contingencia si ambas variables

cuantitativas tienen pocos valores. Si una de ellas tiene pocos valores y la otra muchos, podemos acudir a una representación gráfica propia de una relación entre una nominal y una cuantitativa, como veremos en un apartado posterior. El objetivo es que la representación sea útil. Para ello, según vemos, no solo es importante que se adapte al tipo de escala, sino que las variables representadas cuenten con suficientes valores como sea necesario según el tipo de gráfica que estemos utilizando.

El diagrama de dispersión de nuestro ejemplo es muy ilustrativo. Se observa con claridad que conforme las notas en una asignatura tienen un valor mayor, también son mayores las calificaciones en la otra asignatura. Por este motivo, los puntos se distribuyen en torno a una línea recta imaginaria que es ascendente. Este tipo de relación (lineal ascendente) se denomina “relación positiva”. Imagina que ocurriera lo contrario: conforme mayores son las notas en una asignatura, menores son las calificaciones en la otra. En tal caso, los puntos se mostrarían en torno a una línea imaginaria descendente (más alta a la izquierda y más baja a la derecha). En tales casos se habla de “relación negativa”. Así pues, nuestro diagrama de dispersión está hablando y dice “he aquí una relación positiva entre las notas en geografía e historia”.

Cuantificación. Coeficiente de correlación lineal simple de Pearson

Observa este par de conjuntos de datos:

Conjunto A: 3, 12, 9, 1 Conjunto B: 7, 4, 9, 2

Nos planteamos un juego. Consiste en formar parejas de números: uno del A con uno del B. Tendremos al final 4 parejas. Pues bien. Ensaya con ello. Juega generando diferentes parejas. Hay seis combinaciones posibles. Intenta encontrar la combinación que consigue la máxima *suma de productos cruzados*. Un producto cruzado es la multiplicación de los dos números de un par. Si has juntado A:3 con B:9, el producto cruzado es 27. Cuando hayas generado un conjunto de cuatro pares, calcula los cuatro productos cruzados y suma el resultado. Juega con ello y después sigue leyendo.

No sé si has jugado realmente. Supongamos que sí. Si has encontrado las seis agrupaciones posibles, has calculado los productos cruzados y los has sumado, encontrarás que el valor máximo para esa suma ocurre cuando juntas los valores grandes de ambos conjuntos y, por tanto, también los pequeños entre sí. El valor mínimo ocurre cuando reúnes los grandes con los pequeños. La suma mínima es:

	par 1	par 2	par 3	par 4	
A	1	3	9	12	
B	9	7	4	2	Suma:
AB	9	21	36	24	90

Y la máxima:

	par 1	par 2	par 3	par 4	
A	1	3	9	12	
B	2	4	7	9	Suma:
AB	2	12	63	108	185

Las cuatro combinaciones restantes suministran valores mayores que 90 y menores a 185. Aquí está el corazón del procedimiento que vamos a utilizar para cuantificar la relación entre dos variables cuantitativas. Si observas la suma mínima de productos cruzados, ocurre cuando la relación entre las dos variables es negativa, mientras que la suma máxima tiene lugar en el caso contrario: relación positiva. Parece, pues que un buen índice de relación (ir) sería:

$$ir = \sum AB$$

Parece un buen índice, pero no lo es. Tiene un par de inconvenientes importantes. El primero es que es sensible al número de datos. Si en lugar de 4 pares contáramos con 8, el procedimiento suministraría un

valor más grande y eso no estaría señalando una relación mayor, ni una mayor relación positiva, sino registrando únicamente que estamos considerando un conjunto más grande de datos. La solución parece inmediata: en lugar de la suma, la media. No es mala cosa. El nuevo índice (ir2) sería:

$$ir2 = \frac{\sum AB}{n}$$

No obstante nos queda el otro inconveniente: la escala de medida. Imagina que estamos registrando distancias en metros. Si se nos ocurre registrarlas en centímetros, todo queda multiplicado por 100. Es más, los productos cruzados generarán el efecto de que el resultado final quede multiplicado por 100^2. Esto es una barbaridad. Hay que corregirlo. Y sabemos muy bien cómo hacerlo. Lo sabemos porque conocemos una forma de expresar valores o puntuaciones que es independiente de la escala: las puntuaciones típicas, distancias estandarizadas o Zs. Si operamos con ellas en lugar de con las puntuaciones originales, tendremos nuestro índice, que aunque se nos ha ocurrido aquí, ya se le ocurrió también a otra persona hace un siglo, nuestro compañero de viaje Carl Pearson (el mismo que el coeficiente de variación y la chi cuadrado). Se le llama “coeficiente de correlación lineal simple de Pearson” y se simboliza con la letra r.

$$r = \frac{\sum Z_A Z_B}{n}$$

Es un índice, estadístico o *coeficiente*. Se denomina de *correlación* porque mide relación entre variables sin que podamos establecer estadísticamente un sentido (una es causa de la otra), por lo que hablamos de co-relación. *Lineal* porque se refiere a una relación que puede ser expresada mediante una línea recta. Digamos que en lugar de hablar de *línearectal* o sencillamente *rectal*, decimos *lineal*, algo que es de agradecer en castellano. Cuando la relación entre dos variables no sigue una línea recta sino curva, hablamos de relación *curvilínea*. Es *simple* porque la lógica de este procedimiento puede aplicarse no solo al caso de relación entre dos variables, sino también de relación entre más de dos variables, en cuya situación ya no hablamos de coeficiente de correlación simple, sino múltiple. Y, por último, de *Pearson*, porque fue el inventor de la cosa. Cuidado con Carl Pearson. En efecto tenía una mente brillante para estos asuntos, pero también es cierto que el hombre se preocupó de divulgar sus hallazgos, hallazgos que en muchas ocasiones eran compartidos con personas que no hicieron lo mismo, por lo que no cuentan con un pedestal a lo pearson.

Interpretación de r

Observa lo que ocurre con las dos combinaciones anteriores (mínima y máxima) cuando calculamos r con ellas.

Mínima:

	par 1	par 2	par 3	par 4	
Z_A	-1,183	-0,732	0,620	1,296	
Z_B	1,300	0,557	-0,557	-1,300	Suma:
$Z_A Z_B$	-1,538	-0,408	-0,345	-1,685	-0,994

Máxima:

	par 1	par 2	par 3	par 4	
Z_A	-1,183	-0,732	0,620	1,296	
Z_B	-1,300	-0,557	0,557	1,300	Suma:
$Z_A Z_B$	1,538	0,408	0,345	1,685	0,994

Los resultados se acercan mucho a -1 para la mínima media de productos cruzados de puntuaciones estandarizadas, y +1 para la máxima. Es más, los resultados obtenidos no llegan a 1 o -1 porque hemos arrastrado errores en los redondeos de Z. Obtener 1 o -1 No es una casualidad. El coeficiente r se mueve en ese intervalo (-1, 1), con este significado:

- Cuando r = -1, la relación entre las dos variables es máxima y negativa. Conforme una aumenta,

la otra disminuye y lo hace de tal modo que bastaría con conocer una de las variables para deducir la otra (eso es lo que significa relación máxima).

- Cuando r = 1, la relación es máxima y positiva. Conforme una aumenta, la otra también.
- Cuando r = 0, la relación es nula. Ocurra lo que ocurra con una de las dos variables, no sabemos nada de la otra. Son independientes.

En cualquier ocasión, lo que expresa r es *grado* y *sentido* de relación. Conforme mayor sea su valor absoluto (más cercano se encuentre a -1 o a 1), mayor es la relación entre ambas variables. El sentido de la relación (positiva o negativa) depende del signo de r.

Para ayudarnos a interpretar su cuantía, tomándolo como una medida de tamaño de efecto, nos valen las sugerencias de Cohen, que hemos visto para el caso de la chi cuadrado, utilizando los mismos puntos de corte:

- $r < 0{,}10$: efecto nulo.
- $0{,}10 \leq r < 0{,}30$: efecto pequeño.
- $0{,}30 \leq r < 0{,}50$: efecto moderado.
- $r \geq 0{,}50$: efecto grande.

A partir de r = 0,10, diremos que existe relación, si bien de efecto pequeño, moderado o grande. Visto

más despacio, personalmente yo pediría al menos r ≥ 0,30 para asumir algo de relación, puesto que al contrario de lo que ocurre con la V de Cramer, r es muy sensible. Podrías ver diagramas de dispersión donde no hay forma de apreciar relación, con valores de r claramente superiores a 0,10. He dicho “personalmente yo pediría”. No es lo que suele hacerse en la comunidad académica y, por tanto, no es exigible en esta asignatura. Pero llamo la atención sobre ello: un poco más de exigencia con r no vendría nada mal.

Una última aclaración: r mide relación lineal, es decir, relación que puede ser representada mediante una línea recta. Si hay relación entre dos variables pero no cumple esa condición de linealidad, entonces r puede suministrar incluso el valor 0. Por eso, entre otras razones ya expuestas, es importante que antes de abordar la cuantificación llevemos a cabo una representación gráfica de la relación entre ambas variables.

Más fácil de calcular

Para despedirnos de *r* es bueno tener en cuenta que existen otras formas de calcularlo. Obviamente, todas llevan al mismo resultado. Ocurre como pasó con la varianza. Recuerda sus dos versiones:

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n} = \frac{\sum X_i^2}{n} - \bar{X}^2$$

Ambas expresiones permiten calcular el valor de la varianza. La primera es preferible para comprender mejor en qué consiste el cálculo. La segunda lo hace más sencillo y rápido. Lo mismo ocurre con el coeficiente de correlación lineal simple de Pearson. Hemos visto la expresión de cálculo que permite comprender mejor en qué consiste. Pero para calcular el índice es preferible otra, que se deduce de la primera cuando se sustituyen las distancias estandarizadas por sus expresiones de cálculo:

$$r = \frac{\sum Z_A Z_B}{n} = \frac{\overline{X_A X_B} - \overline{X}_A \overline{X}_B}{S_A S_B}$$

La segunda expresión es más sencilla de calcular que la primera, aunque la primera expresa con más claridad el significado de lo que estamos haciendo. La nueva expresión puede ser recordada como *media de productos menos producto de medias entre las desviaciones*.

Vamos realizar los cálculos para nuestro ejemplo, de las dos formas. Para ello, partimos de una tabla con resultados intermedios. Esta tabla va a contar con una fila para cada caso (en nuestro ejemplo, un

caso se refiere a una persona), con la siguiente información: puntuación directa en las variables A y B (lo necesitamos para el cálculo de la media de cada variable y para el producto $X_A X_B$ en la segunda versión del cálculo de *r*), puntuaciones cuadráticas de A y B (lo necesitamos para el cálculo de las desviaciones tipo), producto cruzado de las puntuaciones directas (para el cálculo de *r* en la segunda versión) y puntuaciones tipo con sus productos cruzados (para el cálculo de *r* en su primera versión). Con todo ello:

	X	X^2	Y	Y^2	XY	Zx	Zy	ZxZy
	5	25	6	36	30	-0,445	0,301	-0,134
	6	36	5	25	30	0,023	-0,246	-0,006
	6	36	7	49	42	0,023	0,847	0,020
	7	49	6	36	42	0,492	0,301	0,148
	2	4	3	9	6	-1,852	-1,339	2,480
	8	64	7	49	56	0,961	0,847	0,814
	9	81	7	49	63	1,430	0,847	1,212
	4	16	3	9	12	-0,914	-1,339	1,224
	8	64	7	49	56	0,961	0,847	0,814
	9	81	8	64	72	1,430	1,394	1,993
	9	81	7	49	63	1,430	0,847	1,212
	8	64	7	49	56	0,961	0,847	0,814
	7	49	6	36	42	0,492	0,301	0,148
	5	25	5	25	25	-0,445	-0,246	0,110
	3	9	2	4	6	-1,383	-1,886	2,609
	4	16	3	9	12	-0,914	-1,339	1,224
	3	9	4	16	12	-1,383	-0,793	1,096
	5	25	3	9	15	-0,445	-1,339	0,597
	4	16	5	25	20	-0,914	-0,246	0,225
	7	49	8	64	56	0,492	1,394	0,686
Suma	119	799	109	661	716	0,000	0,000	17,288
Media	5,95	39,95	5,45	33,05	35,8	0,000	0,000	0,864

$$\overline{X}_A = \frac{\sum X_A}{n} = \frac{119}{20} = 5{,}95$$

$$\overline{X}_B = \frac{\sum X_B}{n} = \frac{109}{20} = 5{,}45$$

$$S_X = \sqrt{\frac{\sum X_i^2}{n} - \bar{X}^2} = \sqrt{\frac{799}{20} - 5{,}95^2} = 2{,}1325$$

$$S_Y = \sqrt{\frac{661}{20} - 5{,}45^2} = 1{,}8296$$

$$r = \frac{\sum Z_A Z_B}{n} = \frac{17{,}288}{20} = 0{,}864$$

$$r = \frac{\overline{X_A X_B} - \overline{X}_A \overline{X}_B}{S_A S_B} = \frac{35{,}8 - 5{,}95 \cdot 5{,}45}{2{,}1325 \cdot 1{,}8296} = 0{,}864$$

En nuestro ejemplo, con la relación entre calificaciones de geografía y de historia, el valor que obtenemos es r = 0,86, lo que indica claramente un efecto grande.