

Relaciones entre las variables

Al concluir un estudio, como cualquiera de los que se han puesto como ejemplo en los temas desarrollados, se obtiene una gran cantidad de datos agrupados en sus variables. Así, nos podemos encontrar con una gran tabla que contiene la edad de los pacientes, su valor de colesterol total, su presión arterial, el número de veces que ha acudido a la farmacia, etc.; es decir, una diversidad de variables cuantitativas que derivan de cada individuo que ha participado en nuestro estudio. El primer paso para su análisis es investigar si hay o no alguna relación entre ellas; es decir, examinar si la variación de alguna de ellas implica inexorablemente la variación de otra.

La relación

La existencia de una relación entre dos variables conlleva que si la primera cambia, la segunda lo hará también, sea en sentido positivo o negativo. La importancia de este efecto es muy grande. Piense en la conclusión que se podría derivar si colegimos que el resultado se ha obtenido como consecuencia de la intervención aplicada, si en realidad lo hubiera sido en función de otra variable que no estaba controlada. O a la inversa, el resultado puede haberse conseguido por una variable con la que no contábamos, con lo que se anula la influencia de la intervención, por lo que se podría inferir que esa intervención tiene incluso un efecto negativo (fig. 1). No obstante, tiene que quedar claro que la correlación entre dos variables no implica causalidad entre ellas, sino simplemente una interrelación.

Veámoslo con un ejemplo. Supongamos que se implementa una intervención para mejorar el control de los pacientes diabéticos de nuestra farmacia. Al cabo de un tiempo, al analizar el comportamiento de su glucemia,

se observa que en los pacientes a los que hemos dedicado mayor tiempo en la intervención se ha producido una proporción de hipoglucemias superior a la del grupo control. A la vista de estas dos únicas variables, valor de glucemia y tiempo invertido, se podría concluir que la intervención no favorece el control del paciente, dado que se producen hipoglucemias en una proporción significativa. Imaginemos ahora que, después de hablar con los pacientes, descubrimos que, por algún motivo, ellos han entendido que había que incrementar notablemente el nivel de ejercicio físico sin que fuera necesario compensarlo con ninguna acción. En este caso, si ahora se analiza el valor de glucemia y el tiempo de ejercicio físico se observaría que los pacientes que más ejercicio han realizado se corresponden con los que han presentado episodios de hipoglucemia.

La relación comentada, que se ha deducido de una forma empírica, puede analizarse mediante el análisis de datos con una prueba llamada correlación, que evalúa la existencia o no de relaciones entre variables, con lo que colabora en la explicación de los hechos y facilita la elaboración de conclusiones verdaderas.

Correlación entre variables

La primera acción para investigar si hay o no correlación entre todas las variables que hemos estimado en el estudio consiste en la representación gráfica de cada par de variables. Ello producirá una nube de puntos, cuya forma nos ofrece ya una idea de lo que está sucediendo: cuanto más agrupados estén los puntos (más estrecha es la nube), mayor relación habrá entre las variables. Además, si la nube de puntos presenta una posición ascendente en el gráfico, es decir, cuando aumenta una variable aumenta también la otra, la correlación es positiva; en caso contrario, será negativa.

Posteriormente, se estima el coeficiente de correlación, conocido como de *Pearson* y representado como «r». Éste toma el valor entre +1 y -1. Cuanto más se aproxime al valor unidad, bien sea positivo o negativo, mayor correlación habrá. La ausencia total de relación entre las variables ofrece un coeficiente 0, donde la nube presenta una gran dispersión en sus puntos. Finalmente, un coeficiente igual o superior a $\pm 0,7$ indica una buena correlación. Como en otros temas, Excel ofrece una solución muy sencilla para analizar la correlación.

Fig. 1. La intervención parece ser la causante del resultado. Sin embargo, puede haber alguna variable que tenga una relación positiva o negativa, causante de la variación del resultado.

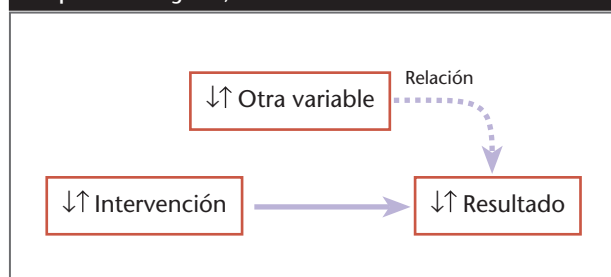


Tabla 1. Datos obtenidos en el ejemplo supuesto citado en el texto*

ID	EDAD (AÑOS)	COLESTEROL (MG/DL)	TIEMPO (MIN/MES)	ID	EDAD (AÑOS)	COLESTEROL (MG/DL)	TIEMPO (MIN/MES)
1	59	155	41	11	75	204	35
2	65	215	24	12	76	226	24
3	78	235	21	13	84	255	24
4	84	266	14	14	66	215	29
5	65	176	41	15	64	185	36
6	69	181	37	16	76	183	31
7	76	241	25	17	81	224	33
8	81	257	22	18	69	188	33
9	58	215	26	19	72	286	19
10	67	188	29	20	62	235	27

*ID es el número identificativo de cada paciente.

Cálculo práctico

Para simplificar al máximo el ejemplo, supongamos una intervención efectuada en 20 pacientes para la prevención primaria de episodios cardiovasculares; al finalizar esa intervención se dispone de los datos de las variables edad (años), colesterol (mg/dl) y tiempo invertido en cada paciente (minutos al mes) (tabla 1).

En un primer paso se representarán los gráficos de cada par de variables cuantitativas (edad frente a colesterol, edad frente a tiempo de intervención y tiempo de intervención frente a colesterol) (fig. 2). La simple visión de esos gráficos parece indicar que en este ejemplo hay una relación muy pobre entre colesterol y edad, si bien se observa una tendencia de incrementar su valor en sangre al aumentar la edad; respecto a la

Fig. 2. Gráficos entre cada par de variables cuantitativas para apreciar la forma de cada «nube» de puntos.

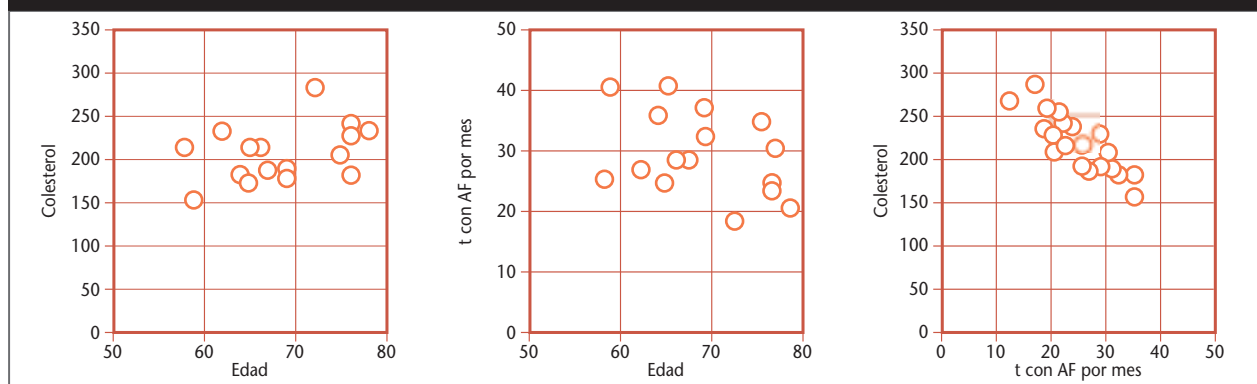


Fig. 3. Rectas de regresión entre cada par de variables.

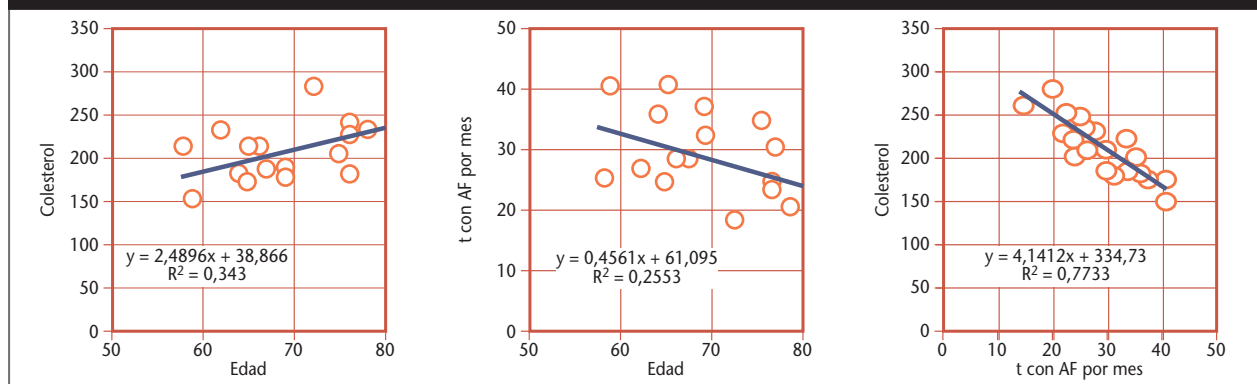


Tabla 2. Tabla de correlaciones entre todas las variables

	EDAD	COLESTEROL	TIEMPO
Edad	1		
Colesterol	0,5856	1	
Tiempo	-0,5053	-0,8794	1

edad y al tiempo invertido en la intervención, la nube es más dispersa, lo que indica aún una menor relación. Finalmente, el gráfico del tiempo invertido respecto al colesterol muestra una nube estilizada descendente, lo que indica que parece existir una relación negativa entre ambas, es decir, cuanto más tiempo se ha dedicado al paciente, menor valor de colesterol ha obtenido finalmente. Recordemos de nuevo que ello no implica necesariamente que la intervención presente una relación de causa a efecto con el valor de colesterol, sino que habrá que demostrarlo más adelante.

Posteriormente, con Excel se determina el coeficiente de regresión de Pearson. Esto se hace con una función específica ubicada en: «Herramientas»/«Análisis de datos»/«Coeficiente de correlación». El resultado que ofrece la aplicación se muestra en la tabla 2, que ofrece el coeficiente de correlación de Pearson para cada par de variables (p. ej., coeficiente entre tiempo y colesterol: -0,8794, lo que significa que cuanto mayor tiempo se ha intervenido, menores valores de colesterol se muestran).

A partir de los gráficos de la figura 2, se pone el ratón sobre cada uno y se va a «Gráfico»/«Agregar línea de tendencia». En la solapa «Opciones» se marca la casilla «Presentar el valor R cuadrado en el gráfico». El resultado queda como la figura 3, donde se observa que se ha dibujado la recta de regresión de la variable en ordenadas (el eje vertical) sobre la ubicada en abscisas (el horizontal), además de ofrecer el cuadrado del coeficiente de Pearson (expresado como R^2). Nuevamente, se ve claro que, cuanto más agrupados estén los puntos sobre su recta de regresión, mayor será la correlación y el coeficiente de correlación se aproximará más a 1, positivo o negativo. Hay que tener en cuenta que el valor del coeficiente buscado se puede deducir de cada gráfico después de hallar la raíz cuadrada de cada valor de R^2 ; p. ej., en la de colesterol (en ordenadas) frente a edad (en abscisas): como R^2 es 0,343, el valor de R será: $\sqrt{0,343} = 0,5856$, como se había estimado previamente en la tabla 2 con la función de Excel. ■