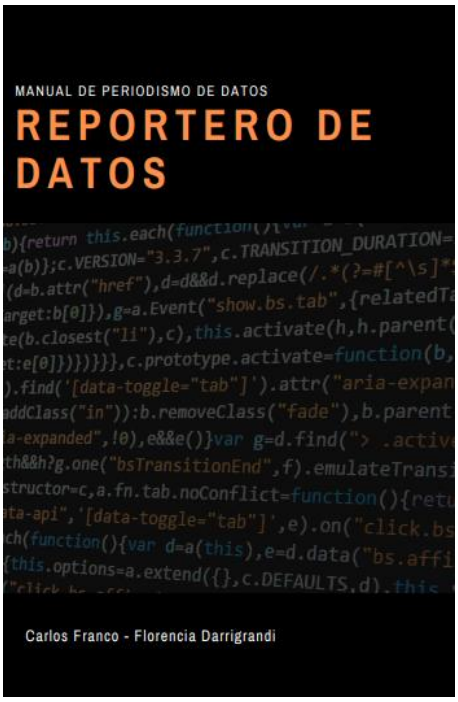


Correlación no es causalidad

El 21 de mayo del 2019 la página CriptoNoticias (www.criptonoticias.com) publicó un reportaje cuyo titular era el siguiente "***Disminuye la correlación*** *entre bitcoin y el resto de las criptomonedas en 2019*"

https://www.criptonoticias.com/mercados/precios-trading/disminuye-correlacion-bitcoin-criptomonedas-2019/

Más adelante, en el cuerpo del reportaje, se menciona lo siguiente:

"En el transcurso del 2018 se observó que ***los niveles de correlación entre las principales criptomonedas y bitcoin fueron especialmente altas****. El 75% de las principales 200 criptomonedas por capitalización de mercado,* ***tuvieron una correlación promedio de 0,89****. Según hodlbot en un informe publicado a comienzo de abril, 150 de las 200 primeras monedas por capitalización de mercado tuvieron una correlación de al menos 0,87 o más en todo el 2018."*

¿A qué se refiere el artículo cuando habla de correlación? ¿Qué significará que esta sea de 0,89?

El coeficiente de correlación lineal ayuda a responder la siguiente pregunta acerca da la relación entre dos variables. Si conozco el valor de la primera variable, ¿con qué precisión puedo predecir el valor de la segunda variable?

En estadística, a la primera variable se le llama variable independiente, pronosticadora o explicativa. Mientras que a la segunda variable se le llama variable dependiente o de respuesta.

Volviendo al ejemplo del inicio, en este caso se tiene que la variable dependiente es el precio del bitcoin, en tanto que la variable independiente sería el precio de otras criptomonedas. En otras

palabras, el artículo de alguna manera apunta a, si es posible predecir el valor del bitcoin en base al precio de otras criptomonedas.

El coeficiente de correlación mide la fuerza y sentido que tiene la relación lineal entre dos variables. Esta puede ser directa o inversamente proporcional. Por ejemplo, directamente proporcional sería que, a mayor cantidad de horas trabajadas, más dinero se recibe a fin de mes, en tanto, inversamente proporcional sería que, a mayor cantidad de horas trabajadas, se tiene un menor nivel de felicidad.

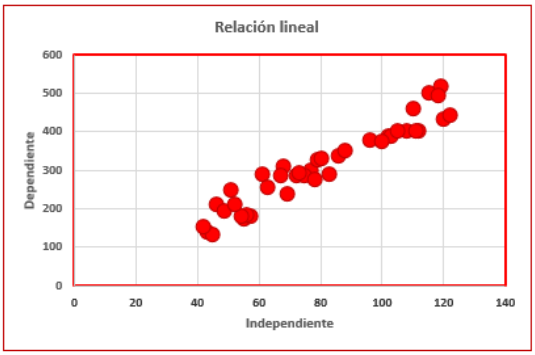
Los clásicos ejemplos de correlaciones son: la altura y peso de una persona, el nivel de escolaridad e ingresos, el número de horas dedicadas al estudio y las calificaciones obtenidas.

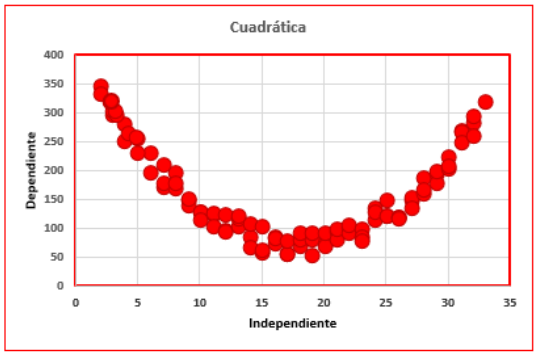
Algunos ejemplos recurrentes en los medios de comunicación son la correlación entre la bolsa chilena y la de Estados Unidos, precio del dólar y precio del cobre, horas de trabajo y productividad.

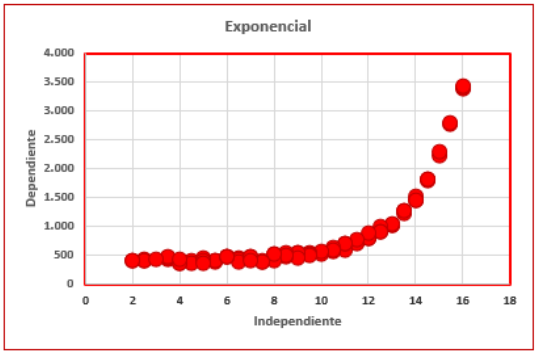
La relación entre dos variables no siempre es lineal. Existen distintos tipos de relaciones, como, por ejemplo, cuadráticas, exponenciales, etc. En este contexto, resulta valioso el uso del diagrama de dispersión. Este nos permite tener una primera aproximación del tipo de relación que hay entre dos variables… si es que la hay.

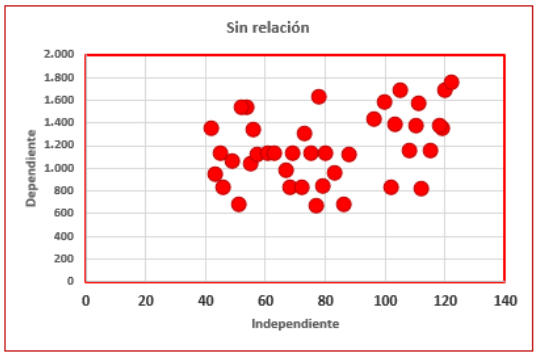
Para construir un diagrama de dispersión, en el eje X se pone la variable independiente y en el eje Y la variable dependiente.

Veamos algunos ejemplos de diagramas de dispersión:









Antes de seguir profundizando acerca del coeficiente de correlación lineal nos detendremos para explicar la diferencia entre lo que son las relaciones determinísticas y no determinísticas.

Supongamos que envío mi sobrino a comprar pan al negocio de la esquina. Yo sé que el kilogramo de marraqueta en ese negocio cuesta $1.400, por lo que, si le pido que traiga un kilogramo, debo darle $1.400, si le pido dos kilogramos, $2.800 y así sucesivamente. Lo relevante aquí es que sé exactamente cuanta plata va a necesitar, por eso es que podemos hablar de una relación determinística.

Si la variable dependiente que llamaremos, Y, es el monto a pagar y la variable dependiente X, son los kilos de pan a comprar, gráficamente tenemos lo siguiente:

Ahora, ¿qué pasa si yo no sé a dónde mi sobrino va a ir a comprar el pan? Teniendo en cuenta que el valor del pan puede variar de un negocio a otro, en este caso no sé exactamente cuánto va a tener que pagar, probablemente tenga una idea aproximada, pero el valor exacto lo desconozco.

Gráficamente tendríamos lo siguiente:

(*) Precios recopilados por alumnos de Ingeniería Civil de la Universidad Adolfo Ibañez, mayo 2019.

Este es un ejemplo de relación no determinística, pues, para un número fijo de kilogramos de pan no sé exactamente lo que tendré que pagar. Si sé, por ejemplo, que si compro dos kilogramos de pan tendré que pagar entre $1.700 y $3.780.

En estadística interesan principalmente las relaciones no determinísticas.

La correlación toma valores entre -1 y 1. Cuando se analiza la correlación se debe analizar, por un lado, su signo, si es positivo o negativo, y por otro lado su magnitud (valor). Si la correlación es positiva hay una relación directamente proporcional. Si la correlación es negativa hay una relación inversamente proporcional. Si la correlación es 1 diremos que existe una relación lineal, inversamente proporcional, perfecta, entre ambas variables. Si la correlación es -1 diremos que existe una relación lineal, inversamente proporcional, perfecta, entre ambas variables. En ambos casos estamos hablando de relaciones determinísticas.

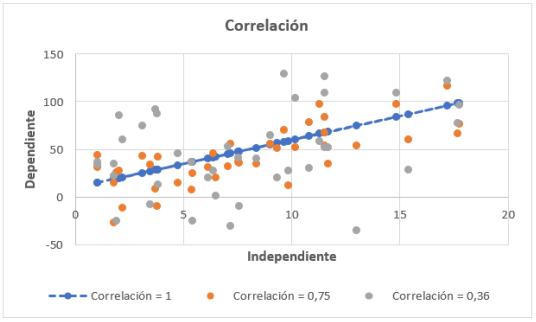
¿Qué pasa cuando la correlación es mayor que -1 y menor que 0 o mayor que 0 y menor que uno? Mientras mayor sea el valor absoluto de la correlación, mayor o más fuerte será la relación lineal entre la variable dependiente e independiente. Visto de otra manera, mientras mayor sea la correlación lineal entre dos variables, la variable independiente podrá predecir con mayor precisión el valor de la variable dependiente.

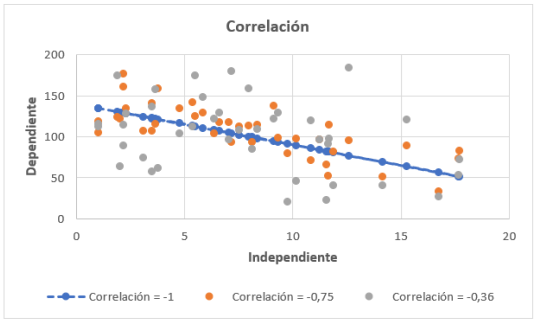
Observemos los siguientes ejemplos.

El primer gráfico muestra relaciones directamente proporcionales. La línea y los puntos azules corresponden a un caso de correlación igual a 1, es decir de una relación perfecta o determinística. A media que disminuye la correlación vemos como la nube de puntos se dispersa entorno a los puntos azules. Los puntos naranjos tienen una correlación de 0,75, mientras que los grises tienen una correlación de 0,36.

El segundo gráfico muestra relaciones inversamente proporcionales. La línea y los puntos azules corresponden a un caso de correlación igual a -1. Al igual que antes, a media que disminuye la

correlación en valor absoluto, se observa como la nube de puntos se dispersa entorno a los puntos azules. Los puntos naranjos tienen una correlación de -0,75, mientras que los grises tienen una correlación de -0,36.





Correlación 0 no implica que las variables sean independentes. Como vimos anteriormente, pueden haber diferentes tipos de relaciones entre dos variables. Hay que tener en cuenta que la correlación mide asociación lineal y en consecuencia, si la correlación entre dos variables es 0, lo único que podemos asegurar es que no hay relación lineal entre ellas. Por el contrario si las variables son independientes podemos asegurar que la correlación es 0 pues entre ellas no existe relación lineal ni de ningún tipo.

Volviendo al ejemplo del inicio, donde se menciona que la correlación promedio entre las principales criptomonedas y bitcoin fue de 0,89, podemos concluir que dado que esta correlación es mayor a 0, existe una relación directamente proporcional entre ambas variables y que esta es bastante alta dado que es relativamente cercana a 1.

Es importante tener en cuenta que la correlación no es causalidad. O sea, aunque exista correlación, esto no quiere decir *per se* que el comportamiento de una variable sea causada por la otra. De hecho, en internet está lleno de correlaciones espurias: en que se tiene un alto grado de correlación, pero no una lógica en que una variable explique la otra.

Hay un sitio llamado “Tyler Vigen” (https://www.tylervigen.com/spurious-correlations) donde se muestran diversos casos de correlaciones espurias, por ejemplo, la tendencia directamente proporcional entre el consumo per cápita de pollo y las importaciones de petróleo crudo de Estados Unidos.

Al interpretar la correlación, es muy importante considerar la opinión de un experto.

Analicemos este caso estudiado por los alumnos de Periodismo de la Universidad Adolfo Ibañez.

La gráfica muestra la evolución de la inmigración venezolana y haitiana a Chile desde los años 2005 a 2016. En este caso hay una alta correlación entre ambas variables ¿Podemos hablar de causalidad en este caso? ¿O será que ambas variables varían simultáneamente debido a una tercera? (Poner la gráfica del ejemplo.)