por Bartolo Luque

Bartolo Luque es físico y profesor de matemáticas en la Universidad Politécnica de Madrid. Sus investigaciones se centran en la teoría de sistemas complejos.

Correlación no implica causalidad

De las promesas del *Big Data* a los usos y abusos de la estadística

El Gran Colisionador de Hadrones (LHC) del CERN, que hace cuatro años encontró el bosón de Higgs, emplea 150 millones de sensores que generan 40 millones de datos por segundo. El 14 de febrero de 2013, cuando cesó temporalmente su actividad por cuestiones de mantenimiento, acumulaba de sus últimos tres años de experimentos el equivalente a 700 años de películas en alta definición.

Ese ritmo de generación de datos no es exclusivo de las grandes instalaciones científicas: se estima que el tráfico global mensual de datos de telefonía móvil asciende a unos escalofriantes 11 exaoctetos (11.000 millones de gigaoctetos, o *gigabytes*). Si pensamos en Internet y en la cantidad de cachivaches digitales que nos rodean, no extraña que un reciente informe de IBM apuntase que, solo en los últimos dos años, se había generado el 90 por ciento de todos los datos digitales existentes.

Para explorar ese océano de información ha nacido una nueva disciplina: el tratamiento de macrodatos, o *Big Data*. Este hermano forzudo de la estadística se ocupa, en primer lugar, de superar los problemas que plantea buscar, obtener o almacenar esas pantagruélicas masas de información. Y, más delicado e interesante aún, de analizar y visualizar tales datos, en general dispersos y sin clasificar, para extraer información relevante que permita tomar decisiones. Los macrodatos prometen a científicos, Gobiernos y grandes empresas hacer emerger de manera automática relaciones hasta ahora ocultas entre todo tipo de variables.

Sin embargo, dado que vender algo nuevo en un mercado saturado siempre es difícil, algunos seguidores acérrimos del *Big Data* se han venido arriba y han prometido una manera «automática» de hacer nuevos descubrimientos científicos. Algo que recuerda vagamente al programa formalista de Hilbert, pero esta vez en ciencias. Opinan que, con los algoritmos adecuados, podremos encontrar inimaginables correlaciones y regularidades en semejantes cantidades de datos. En palabras de Chris Anderson, que como editor de la célebre revista tecnológica *Wired* levantó polémica en 2008: «No hay necesidad de semántica o de análisis causal. La correlación es suficiente. Podemos introducir los números en el mayor conjunto de ordenadores del mundo y los algoritmos encontrarán patrones donde la ciencia no puede».

Ante semejante optimismo desatado, ha habido respuestas contundentes que han demostrado que, en bases de datos muy grandes, aparecen siempre correlaciones arbitrarias, no debidas necesariamente a la naturaleza de los datos, sino solo a su cantidad. Tales argumentos se basan en la teoría ergódica, la teoría de Ramsey o la teoría algorítmica de la información, entre otras herramientas.

Cum hoc

Con todo, resulta más fácil y gráfica la explicación aportada por Tyler Vigen, estudiante de criminología de Harvard, quien hace poco desarrolló un programa que detecta automáticamente correlaciones entre conjuntos de datos de lo más variopinto. Una rápida búsqueda en su página web (tylervigen.com/spurious-correlations) nos revelará correlaciones estrambóticas, como la existente entre el gasto en I+D de EE.UU. y el número de suicidios por ahorcamiento, estrangulamiento o asfixia a lo largo de una década; o la tasa de divorcios en Maine y el consumo per cápita de margarina, también durante diez años.

¿Cómo cuantifican los estadísticos la bondad de una correlación? El coeficiente de correlación lineal más empleado es el de Pearson, el cual suele denotarse por r y toma un valor comprendido entre +1 y –1. Los extremos indican máxima correlación y anticorrelación, respectivamente, mientras que el valor 0 indica su ausencia. En los dos ejemplos surrealistas de Vigen, $r = 0{,}99$. ¿Significa eso que el incremento en gasto en I+D es responsable del aumento de suicidios, o que cuanta más margarina use una pareja, más probable será que sus miembros se divorcien?

Resulta difícil mantener ambas cosas, ya que correlación no implica causalidad. Estamos sin duda ante correlaciones espurias. La palabra *espurio* procede del latín *spurius* y posee dos acepciones: «bastardo, degenerado desde su origen» y la que nos interesa aquí, «engañoso o falso». Esta última es la empleada en estadística y fue propuesta por primera vez en 1897 por Karl Pearson para referirse a las correlaciones ilusorias. Aprovecho para advertir que *espúreo* es incorrecto, a pesar de que se encuentra muy extendido incluso entre gente culta, como comentaba Lázaro Carreter en *El dardo en la palabra*, donde confesaba haberlo utilizado alguna vez.

Cuando los estadísticos hablan de la correlación de Pearson entre dos variables se refieren a una buena o mala relación lineal entre ellas. Sin embargo, la causalidad hace referencia a que un suceso constituya el resultado de otro. Causalidad siempre implica correlación, pero la correlación no necesariamente implica causalidad. La cantinela de «correlación no implica causalidad» viene de lejos y se conoce también como falacia *cum hoc ergo propter hoc*, «con esto y, por tanto, a causa de esto».

La versión débil de la correlación espuria puede condensarse en otra famosa expresión latina: *post hoc ergo propter hoc*, «después de esto y, por tanto, a causa de esto». Se trata de una conocida falacia, donde se da por sentado que, si A sucedió antes que B, entonces A debe haber causado B. En este caso, se estira demasiado el hecho de que, efectivamente, las causas preceden a los efectos.

Suena tan absurdo que pensamos que nadie puede caer en semejante trampa mental, pero está al orden del día en esta sociedad tan tecnocientífica como

irracional en la que vivimos. Verbigracia: «¡Pues yo he tomado homeopatía y me he curado!». Es más, cuando encontramos correlación entre dos variables, y aun suponiendo que exista causalidad entre ellas, tampoco estamos capacitados para determinar cuál es la causa y cuál el efecto. Un ejemplo histórico, que hoy nos suena patético, fue la defensa que hicieron las tabacaleras ante la alta correlación entre cáncer y tabaco: los enfermos de cáncer fumaban para aliviar los dolores, argumentaban los muy sagaces.

Propter hoc

De todas maneras, ¿cómo es posible que existan correlaciones tan altas en variables entre las que no hay ningún vínculo causal? Una posibilidad, como ocurre en los casos de Vigen, es el puro azar. Abordemos la cuestión de una manera inocente, sin usar la teoría de Ramsey ni matemáticas elaboradas.

Supongamos una serie temporal $x(t)$ de 10 puntos (t = 0, 1, 2, ..., 9), como las que aparecen en los gráficos. Podemos convertirla en una serie de 9 ascensos y descensos, como:

$$\uparrow\uparrow\uparrow\downarrow\downarrow \cdots \uparrow\downarrow ,$$

donde $\uparrow$ significa:

$$x(t+1) - x(t) > 0$$

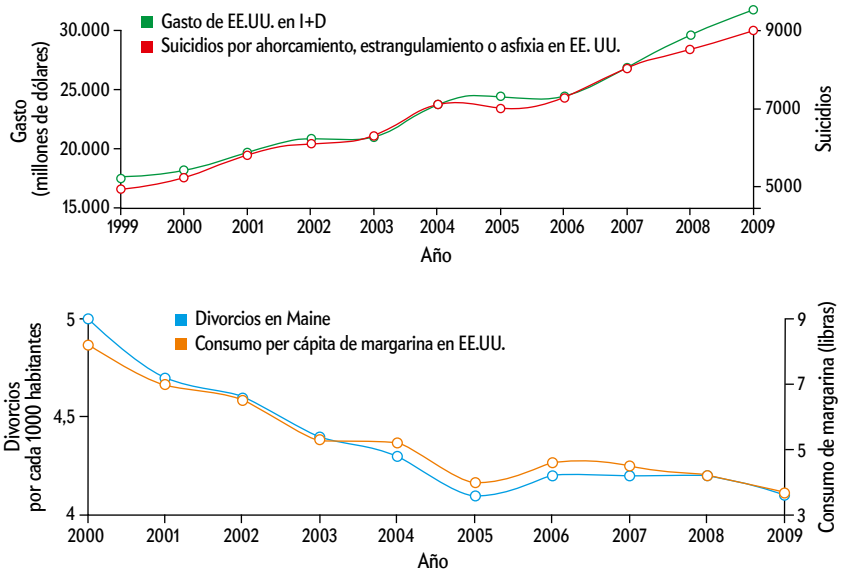
y $\downarrow$ simboliza:

$$x(t+1) - x(t) < 0 .$$

De manera cualitativa, otra serie temporal en otra variable, $y(t)$, se correlacionará positivamente con $x(t)$ si exhibe una secuencia idéntica de subidas y bajadas, y negativamente si exhibe la secuencia opuesta.

Para diez valores tenemos, por tanto, $2^9 = 512$ posibles secuencias. Si tomamos dos de ellas al azar, la probabilidad de que se correlacionen positiva o negativamente será de 2/512 = 1/256. Ahora supongamos que disponemos de 23 series temporales y que escogemos dos de ellas. El número de posibles parejas viene dado por el coeficiente binomial $C(2,23) = 253$. Por lo que, en promedio, siempre podremos esperar encontrar una correlación o anticorrelación por pura suerte.

Otra posibilidad para generar correlaciones espurias es la existencia de una variable oculta. Martin Gardner nos alertaba de ellas hace ya años con ejemplos como la correlación entre el tamaño de los pies y la habilidad para sumar: los niños con pies grandes suman mejor. ¡Claro!



CORRELACIONES ESPURIAS: En conjuntos de datos lo suficientemente amplios siempre es posible encontrar correlaciones casi perfectas entre variables disparatadas. Estas gráficas muestran dos ejemplos recopilados por Tyler Vigen, estudiante de criminología de Harvard. En ambos casos, el coeficiente de correlación es $r > 0{,}99$.

Simplemente tienen más edad, la variable oculta que hace de puente causal.

En el clásico sobre falacias estadísticas *How to lie with statistics* (1954), Darrell Huff pone como ejemplo la correlación entre el salario de los ministros presbiterianos de Massachusetts y el precio del ron en La Habana. ¿Cuál es aquí la causa y cuál el efecto? Sin duda, la cuestión resulta disparatada, como en los ejemplos de Vigen. Y que el salario y los precios crezcan a la par no es más que consecuencia de que, con el paso de los años y a nivel mundial, todo es cada vez más caro.

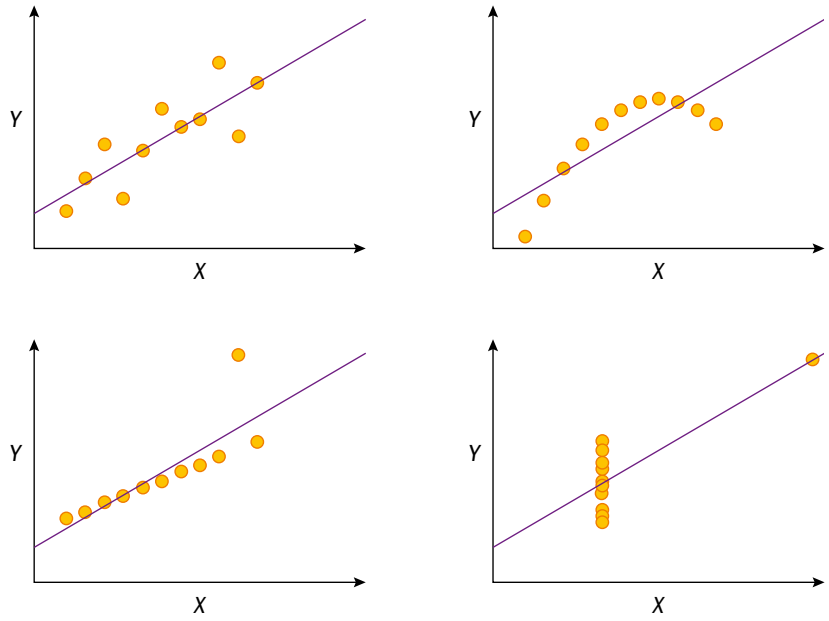
Ciencia espuria

A pesar de que en todas las clases de estadística del planeta se repite una y otra vez la cuestión, siguen apareciendo estudios científicos que caen de una forma u otra en esta vieja falacia. Por ejemplo, cuando se generalizó la terapia de sustitución hormonal (TSH) para amortiguar efectos no deseados de la menopausia, los investigadores hicieron notar que las mujeres que la adoptaban parecían sufrir menos cardiopatías. Algunos estudios adelantaron una relación causal: la TSH reduce el riesgo de enfermedades cardiovasculares. Sin embargo, investigaciones posteriores descubrieron la variable oculta: las mujeres que estaban tomando TSH pertenecían en su mayor parte a grupos socioeconómicos altos, con dietas más sanas y hábito de ejercicio. Cuando se realizaron pruebas a doble ciego con grupos homogéneos para evitar las variables ocultas, se descubrió que, de hecho, el TSH aumentaba ligeramente el riesgo.

Veamos algunos ejemplos más que en su momento tuvieron gran repercusión mediática. En 2000 se publicó un discutido artículo en *Nature* que apuntaba a una fuerte asociación entre miopía y exposición nocturna en niños menores de dos años: los niños que dormían con la luz encendida mostraban una probabilidad cinco veces mayor de desarrollar miopía. Un año después, un estudio publicado también en *Nature* refutaba el resultado señalando que la verdadera causa de la miopía infantil era genética, no ambiental. El trabajo encontraba una fuerte conexión ente la miopía parental y la miopía en desarrollo en niños, haciendo notar, además, que los padres miopes suelen dejar encendida una luz en la habitación de sus vástagos.

En 2012, la revista *New England Journal of Medicine* publicó un artículo cuya conclusión era que el consumo de chocolate mejoraba la función cognitiva. ¿En qué basaban su afirmación los investigadores? Pues en una fuerte correlación entre el número de premios nóbel de una nacionalidad y el consumo per cápita de chocolate en su país. Aquí nos encontramos frente a lo que se conoce como «falacia ecológica»: se alcanza una conclusión sobre individuos a partir de datos agregados de grupos. Se encuentra una correlación para la población de un país

TYLER VIGEN, ADAPTADO DE TYLERVIGEN.COM/SPURIOUS-CORRELATIONS [CC BY SA 4.0]



CUARTETO DE ANSCOMBE: La falta de correlación no indica independencia. Estas cuatro gráficas, conocidas como el «cuarteto de Anscombe», fueron concebidas en 1973 por el estadístico inglés Frank Anscombe para enfatizar la importancia de visualizar los datos antes de elegir el tipo de análisis. Todos los casos muestran dos conjuntos de datos, *X* e *Y* (*naranja*), con el mismo coeficiente de correlación lineal, $r = 0{,}816$ (*ajuste violeta*). Sin embargo, solo el primer gráfico muestra una relación lineal. La segunda y la cuarta gráfica revelan relaciones no lineales a las que no puede aplicarse el coeficiente *r*. La tercera refleja una relación lineal perfecta excepto por un dato fuera de lugar, el cual baja el coeficiente de correlación a 0,816.

y se extrapolan conclusiones para algunos de sus habitantes, los premios nóbel. Pero el consumo real de los laureados les era totalmente desconocido a los investigadores. A pesar de este error elemental, que fue muy criticado en su momento por la comunidad, la prensa se hizo buen eco del resultad. De hecho, el artículo aún no ha sido retirado y goza de 42 citas en el momento de escribir estas líneas.

Niños y cigüeñas

Correlación no implica causalidad. Pero, contrariamente a lo que muchos piensan, una correlación nula tampoco implica independencia. Por ejemplo, una relación funcional en forma de U entre dos variables puede dar una correlación lineal nula. El coeficiente de correlación de Pearson fue creado para determinar la correlación lineal entre variables, por lo que, si hay correlación pero esta es no lineal, podremos encontrar cualquier valor.

Tales malentendidos fueron los que, en 1973, llevaron al estadístico inglés Frank Anscombe a divulgar el hoy llamado «cuarteto de Anscombe». Sin embargo, seguimos viendo trabajos científicos que caen en el mismo error.

Hace unos años, por ejemplo, estudios con ratas de laboratorio sobre la ingesta de DEHP, un componente que añadido al plástico lo hace más flexible, apuntaban a que dicha sustancia aumentaba la actividad de la aromatasa, una enzima que induce masculinización cerebral. El problema en estos casos reside en que, a menudo, los toxicólogos dan por sentado que los tests donde se administran dosis altas revelan los efectos más rápidamente y con menor ambigüedad que aquellos en los que se usan dosis bajas durante periodos prolongados. Y esos ensayos solo habían utilizado dosis elevadas. Más tarde, Anderson Andrade, del Hospital Universitario Charité de Berlín, y sus colaboradores mostraron que, a bajas dosis, el DEHP suprimía la aromatasa: un efecto no lineal totalmente inesperado.

Para terminar, y como ejemplo de que no debemos pedir a la estadística más de lo que puede darnos, el siempre sorprendente Robert Matthews retomó hace unos años el conocido ejemplo de correlación espuria entre tasa de nacimientos y población de cigüeñas que en 1952 propuso el matemático polaco Jerzy Neyman. En un artículo titulado «Las cigüeñas traen los bebés ($p = 0{,}008$)», Matthews aborda la cuestión del mismo modo en que lo haría cualquier investigación donde se sospechase la existencia de una correlación, entre dos variables (como, por ejemplo, dieta y cáncer).

Matthews usa un contraste de hipótesis, donde la hipótesis nula es la ausencia de correlación entre las tasas de nacimiento anuales y el número de parejas de cigüeñas blancas (*Ciconia ciconia*) en 17 países europeos. Una regresión lineal de los 17 pares de puntos arroja un coeficiente de correlación lineal de $r = 0{,}62$, no especialmente alto. Sin embargo, un test *t* estándar (una técnica estadística habitual en estos casos) revela que el resultado es significativamente alto, puesto que arroja un valor *p* de 1/125.

En general, el valor *p* indica la probabilidad de obtener un resultado como el observado si asumimos que la hipótesis nula es cierta. En nuestro caso, eso quiere decir que, si no existe una correlación entre nacimientos y cigüeñas, la probabilidad de toparnos con una correlación positiva como la obtenida es de 1 entre 125. Pero, atención, contrariamente a lo que muchos investigadores piensan, eso no implica que la probabilidad de que todo se deba a una mera coincidencia sea de 1/125. Ni, menos aún, que la probabilidad de que las cigüeñas traigan a los bebés sea de 124/125.

La explicación más plausible, como apunta el propio Matthews, es la existencia de una variable oculta, como la extensión de los países. Este caso nos muestra, más allá de que correlación no implica causalidad, que es necesario entender el significado preciso del tan querido para muchos investigadores valor *p*, y que rechazar una hipótesis nula no implica que la hipótesis alternativa sea correcta.

PARA SABER MÁS

Myopia and ambient night-time lighting. Karla Zadnik et al. en *Nature*, vol. 404, págs. 143-144, marzo de 2000.
Storks delivers babies ($p = 0.008$). Robert Matthews en *Teaching Statistics*, vol. 22, págs. 36-28, verano de 2000.
Chocolate consumption, cognitive function, and Nobel laureates. Franz H. Messerli en *The New England Journal of Medicine*, vol. 367, págs. 1562-1564, octubre de 2012.
What everyone should know about statistical correlation. Vladica M. Velickovic en *American Scientist*, vol. 103, enero-febrero de 2015.

EN NUESTRO ARCHIVO

Una sociedad dirigida por datos. Alex P. Pentland en *IyC*, enero de 2014.
El valor resbaladizo de *p*. Regina Nuzzo en *IyC*, diciembre de 2014.

CORTESÍA DEL AUTOR