Miquel Romero Obón, codirector del curso de Postgrado de Experto en Aplicación Práctica de la Estadística en los Procesos de la Industria Farmacéutica y Afines (2017-2018). Facultad de Farmacia y Ciencias de la Alimentación. Universitat de Barcelona

En Estadística, el tamaño importa hasta que deja de hacerlo: Small & Big Data

Llegamos a esta cuarta edición de fichas estadísticas con un asunto tan simple como poco considerado en todo tratamiento de datos, bien sea con finalidad de estimación, descripción, preparación, filtrado de datos o cualquier otro. Se trata de la determinación del tamaño muestral, del cual toda conclusión posterior va a depender mucho al ser un aspecto de gran influencia en la precisión y significación estadística.

SMALL DATA

Este término no ha existido hasta que no se ha necesitado un concepto antagónico para el Big Data. De hecho, la totalidad de estudios estadísticos realizados hasta hace unos años recae en esta clasificación. No es otro concepto que el de conjuntos de datos relativamente pequeños.

Bien, ¿cómo de pequeños? Cualquier estimación de la popular *n* de todos los estudios estadísticos se basa en cálculos, bien frecuentistas o bayesianos, que ofrecen números bajos. Incluso las *n* de grandes estudios sociológicos o clínicos con decenas de miles de individuos, son cantidades pequeñas al lado del concepto Big Data.

Dicha *n* no crece proporcionalmente a la precisión para una significación estadística determinada, sino que lo hace exponencialmente con acercamiento asintótico al error cercano a cero. Pero, ¿es realmente *n* la única responsable de la credibilidad del estudio? Obviamente no: un tamaño muestral adecuado es una condición necesaria pero no suficiente. El planteamiento y diseño del estudio, así como de la técnica/s estadística/s para procesar los datos, tienen gran influencia en las conclusiones de la investigación. Por lo tanto, no debemos caer en el error de considerar que los estudios con *n* muy grande tienen mayor credibilidad únicamente por haber considerado un tamaño muestral elevado. Errores como emplear métodos paramétricos sobre datos no normales, con heterocedasticidad (variabilidad no estable) o autocorrelacionados, no tener en cuenta la sobreparametrización en los modelos, y, sobre todo, confundir correlación con causalidad, pueden conducir a aseveraciones incorrectas bajo un supuesto fundamento estadístico sólido.

BIG DATA

El crecimiento tecnológico ha favorecido el desmesurado aumento de datos, hasta llegar a tamaños en los que la capacidad de captura y tratamiento superan la capacidad de medios estándar, como ordenadores y software comunes. La imprenta de Gutenberg multiplicó por 500 la capacidad de almacenamiento de información de las tablillas de barro de los sumerios (aprox. 1 dato por pulgada cúbica). Recientemente se ha valorado que los semiconductores nos permiten almacenar $1{,}25 \cdot 10^{11}$ bytes por pulgada cúbica.

Ya, a finales del siglo XIX, necesidades como los censos poblacionales de grandes países llevaron a planteamientos de cómo gestionar en un futuro próximo tales canti-

dades de datos. Se estima que, de mantener los métodos existentes hasta entonces, países como los Estados Unidos habrían necesitado diez años para la gestión de su censo poblacional. Entonces fue la tabuladora de Hollerith (1881) la que permitió capturar y tratar esa gran cantidad de datos; poco después esta misma persona fundó IBM.

Otro de los fenómenos que favoreció la creación de nuevos métodos fue el crecimiento de las bibliotecas ante el rápido incremento de la demanda de nuevas publicaciones e investigaciones. Ya en 1941 se habló de la "explosión de la información" en el periódico Lawton Constitution, con referencia a este fenómeno y posteriormente (1944). Fremont Rider calculaba que el tamaño de las bibliotecas crecía el doble cada 16 años. En 1948, Claude Shannon publicó la Teoría Matemática de la Comunicación, estableciendo un marco de trabajo que posibilitó en los posteriores años que el volumen de los datos fuera muy inferior al que tendríamos sin sus teorías. En 1956 se sumó a todo esto el concepto de memoria virtual, desarrollado por Fritz-Rudolf Güntsch, tratando el almacenamiento finito como infinito. La necesidad de tener soluciones organizativas sólidas volvió a ser objeto de atención al comprobar en la década de 1960 que la velocidad de crecimiento pasó a un factor de 10 cada 50 años, principalmente al descubrir nuevas formas de captura de datos, en especial las procedentes de los sistemas de reconocimiento de voz. En ese momento llegan a la industria los sistemas de computación centralizados para la contabilización de los *stocks*, y, a partir de 1970, las bases de datos relacionales. La generación de datos seguía creciendo, llevándonos al enunciado de la primera ley de Parkinson (1980), simplificada bajo el concepto "los datos se expanden hasta llenar el espacio disponible". En 1989 nace el concepto de inteligencia empresarial, ya mencionado en 1958 por Hans Peter Luhn pero aún no desarrollado. Pero, sin duda, el fenómeno que marca un hito clave es la explosión de la World Wide Web, en la década de los 90, y el crecimiento de la potencia informática con procesadores mucho más veloces.

En 1997 el término Big Data es empleado por primera vez por la NASA, entendido entonces como un problema, al desconocerse cómo guardar todo lo que se generaba y de qué forma podría procesarse, ya que las estimaciones mostraban que un 80 % de los datos guardados jamás serían consultados por nadie. El crecimiento siguió adelante. En 1999 se cuantificó en 1,5 exabytes (1 EB = 10^{18} bytes) la información mundial. La aparición de Internet of Things en 1999, con el uso de dispositivos RFID para trazabilidad en la cadena de suministro, vuelve a explosionar un nuevo crecimiento de los datos y no se tarda en crear el concepto de Software como Servicio (SaaS), que llegará a duplicarse en menos de diez años. En 2006 Big Data pasa de ser explotado por empresas a serlo, además, por usuarios individuales con herramientas abiertas y gratuitas como Hadoop, lo cual vuelve a impulsar el crecimiento de

los datos, de forma que cada 18 meses se duplica y llegamos al zettabyte (10^{21} bytes) como unidad de medida para referirnos a toda esa gran cantidad de almacenamiento.

¿ES TODA ESTA "MAGIA" LA PANACEA DEL CONOCIMIENTO?

Llegado a este punto, se considera que el Big Data Computing es una nueva revolución que desbanca el método científico clásico y universaliza su utilización. La cantidad de datos es tan grande que la significación estadística debe de ser enorme, así como la capacidad de extraer y deducir información. Entendamos todo ello con cautela y sin precipitación.

Dato no es equivalente a información, sino que requiere depuración, clasificación, estructuración, contextualización e interpretación para que pueda transformarse en información. Por otro lado, información no equivale a conocimiento: requiere también procesamiento y, de nuevo, contextualización e interpretación por especialistas. El uso indiscriminado de los datos con disciplina descuidada presenta un elevado riesgo de capturar falso conocimiento. Los motores de búsqueda y minería de datos tratan de encontrar relaciones entre millones de variables y los encuentran. Dichas relaciones no tienen por qué ser causales; de hecho, mayoritariamente no lo serán: habremos encontrado correlaciones que muestran relaciones espurias. Cuando dichas correlaciones se refieren a variables de difícil interpretación o responden a algo que deseamos encontrar (aunque no necesariamente cierto), existe el riesgo de emplearlas junto a su *p-valor* como pase VIP para la publicación (pseudo)científica y contaminar la comunidad con conocimiento débil, o incluso falso.

¿SON FIABLES LAS CONCLUSIONES BASADAS EN LA "PESCA AUTOMÁTICA" DE CORRELACIONES?

El entorno observacional está afectado de efectos como la confusión, interacción y regresión a la media que, de pasar inadvertidos, nos conducen a conclusiones que pueden no ser ciertas. Adicionalmente, la interpretación de causalidad en una correlación, por alta que sea, no tiene ningún valor si no hay un entorno científico que plantea previamente una hipótesis y no es otra cosa que confirmación de la misma lo que buscamos entre esa ingente cantidad de datos. Incluso el editor de la célebre revista tecnológica Wired publicaba en 2008 "no hay necesidad 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". No tengo palabras publicables para calificar tan errónea y grave afirmación.

Demos un poco de explicación técnica. Si entendemos un conjunto de datos como serie temporal ordenada, podemos reinterpretarla, también, como una serie de ascensos y descensos respecto a cada valor. Esta conversión semicualitativa de una sucesión de datos ayuda a comprender cómo pueden darse correlaciones espurias: solo se requiere que otra serie de datos coincida con el mismo orden de ascensos y descensos. Así, para un estudio con datos de diez años tenemos 2^9=512 posibles secuencias. Tomando al azar dos secuencias, podrán correlacionarse positiva o negativamente con una probabilidad 2/512. En cuanto dispongamos de un número alto de variables, el número de parejas será suficientemente alto (por ejemplo, para 23 variables C(2,23)=253 parejas) para que alguna de ellas cumpla o sobrepase la frecuencia de 2/512. En ese momento aparecerán sorprendentes relaciones de alta correlación, unas descartables a simple vista, otras aceptadas por soportar el deseo que originó el estudio y/o por dificultad de interpretación de las variables emparejadas.

Para ilustrar este hecho, es recomendable echar una ojeada a la divertida web sobre correlaciones espurias calculadas automáticamente a partir de grandes bases de datos públicas (www.tylervigen.com). Por citar algunas, el número de suicidios y el gasto en ciencia espacial en US tienen muy elevada correlación durante un periodo de diez años de estudio, así como también el número de muertos por ahogo en piscinas y cuántas películas rodó Nicolas Cage entre 1999 y 2009, o el consumo de margarina y el ratio de divorcios en Maine.

Aunque de menor diversión, resultan de gran interés y apoyo de lo mencionado en las líneas anteriores las siguientes publicaciones:

- John P.A. Ioannidis. Why most published research findings are false. PloS Medicine, 2005 (www.ncbi.nlm.nih.gov)
- Steven Novella. 0,05 or 0,005? P-value wars continue. Science-Based Medicine, 2017 (www.sciencebasedmedicine.org)
- Ben Locwin. Big Data vs Small Data: what's the proper prescription for you? Bioprocess Online, 2017 (www.bioprocessonline.com)

CONCLUSIÓN

El tamaño muestral es relevante en el ámbito de Small Data, donde mayor *n* conlleva menor error y buen nivel de significación estadística. Una vez superadas las fronteras del Small Data, *n* deja de tener relevancia: por más que se incremente, no aporta mayor solidez a las conclusiones que puedan derivarse. En ningún caso, el tamaño muestral es garantía de solidez de las conclusiones, sino condición necesaria para que subsiguientes estudios puedan apoyarse firmemente.

Próximo artículo:

Simulación Estadística para la definición de planes óptimos de mantenimiento preventivo.