En defensa de los 'malos' datos

JOSU MEZO

La pandemia de la COVID-19 nos ha traído una explosión del periodismo de datos, que ha saltado a las portadas de los periódicos y a las cabeceras de las webs. Día tras día, hemos seguido la evolución del número de casos, fallecidos, hospitalizados y curados, comparando continentes, países y regiones. Hemos aprendido a distinguir las tasas de letalidad (fallecidos entre contagiados) y las de mortalidad (fallecidos entre habitantes). Mapas y gráficas de todo tipo –con datos diarios o acumulados, con escalas lineales y ¡logarítmicas!– ocupan enormes espacios en las webs de los medios y en las pantallas de televisión. El seguimiento de la actualización diaria de los datos ofrecidos por el Ministerio de Sanidad se ha convertido en un ritual cotidiano para los medios y para muchos ciudadanos, ávidos por tratar de entender lo que está pasando, hasta dónde va a llegar la pesadilla y cuándo va a terminar o va a hacerse, al menos, manejable.

Pero este frenesí estadístico no convence a todo el mundo. Mucha gente, en público y en privado, se ha quejado de la mala calidad de los datos que estamos manejando, ya que la mayoría de ellos son imprecisos o incompletos. Los casos registrados en cada lugar dependen en parte del número de test que se hagan, y este número, que no siempre se conoce, es enormemente distinto de unos países a otros. Además, unos publican cifras de test realizados y otros de personas testadas (y, a veces, no sabemos exactamente cuál de las dos cosas), y algunos solo cuentan los test que miden la presencia del virus (los PCR) y otros incluyen los test que reconocen un contacto previo con él (los de anticuerpos).

La contabilidad de los fallecidos creíamos que era algo más segura y comparable; si bien, luego nos dimos cuenta de que algunos países estaban contando los fallecidos en hospitales, pero no a los que morían en sus domicilios y en residencias de ancianos. Algunos solo cuentan casos confirmados (con test positivos de COVID-19 antes de fallecer) y otros también casos probables, con

Josu Mezo es profesor de la Universidad de Castilla-La Mancha y editor del blog Malaprensa

síntomas semejantes, pero no testados. Más tarde, nos percatamos de que las estadísticas de exceso de mortalidad que tienen muchos países (diferencia entre el número de fallecidos registrados, por todas las causas, y el número esperable, basado en los promedios de años anteriores) diferían considerablemente, al alza, del número de fallecidos registrados atribuidos a la COVID-19.

Esos y otros problemas, como la desconfianza ante los datos facilitados por países como China o Rusia o menos desarrollados, con Estados menos capacitados para una recogida extensa de datos, pueden llevar a algunas personas a una especie de nihilismo estadístico, concluyendo que, en realidad, todo el esfuerzo dedicado a los datos es improductivo, o incluso contraproducente, al generar una falsa sensación de conocimiento y certeza, en donde solo hay ignorancia y confusión. Como podrán ustedes sospechar, por el título de este artículo, y por la línea habitual de esta sección, no soy muy partidario de esta lectura de los datos relativos a la pandemia.

Para empezar, creo que esa visión crítica surge en parte de la creencia errónea de que existe una línea divisoria nítida y clara entre datos *buenos* y datos *malos*, o incluso entre datos *ciertos* y datos *falsos*, de manera que los únicos datos que deberíamos manejar en estudios y publicaciones sociales, económicos o científicos, y en los medios de comunicación que los reflejan, son los *buenos*, es decir, los ciertos y precisos. Pero eso es una ilusión. Verdaderamente, en todas las áreas del conocimiento se trabaja con una amplia gama de datos con diferentes grados de seguridad y certeza, unos más nítidos y otros más desenfocados.

Esto es así incluso en muchas de las estadísticas que manejamos de manera cotidiana, con mucha comodidad, y sin plantearnos casi nunca su precisión. Hablamos con naturalidad de la temperatura o la lluvia en una ciudad, como si los datos de un observatorio colocado en un punto concreto pudieran darnos una información correcta para toda ella. También citamos datos sobre la población de las ciudades, obtenidos del padrón, cuando sabemos que por múltiples razones hay personas que no viven realmente en el municipio en el que están empadronados. Todos los países del mundo calculan cuidadosamente la variación trimestral de su producto interior bruto (PIB), que se utiliza para orientar múltiples decisiones económicas y políticas, pero la estimación inicial suele ser revisada varias veces durante los siguientes meses o hasta años más tarde. Por no hablar, ya que mencionamos el PIB, de la economía sumergida, ese gran fantasma que todo el mundo sabe que existe y que nadie está muy seguro de cuánto mide, aunque sospechamos que difiere bastante de unos países a otros, de manera

que su inclusión en el PIB alteraría notablemente muchas otras estadísticas derivadas (como la renta per cápita o el tamaño del gasto y la deuda públicas en relación con el PIB). También aceptamos los datos publicados por las empresas sobre sus beneficios, y los utilizamos para hacer cálculos sobre su futuro rendimiento, y el valor que deberían tener sus acciones, al tiempo que somos conscientes de que pueden hacer aparecer y desaparecer costes y beneficios en diferentes subsidiarias localizadas en países con mayor o menor fiscalidad, y de que en ciertos sectores financieros la apreciación y depreciación de activos y pasivos y, por tanto, de pérdidas y ganancias puede ser enormemente creativa.

También las áreas de ciencias naturales, de la salud o las ciencias *duras*, frente a lo que tal vez piensan quienes las asocian con la exactitud y precisión absolutas, manejan con total normalidad datos rodeados de incertidumbre, que se expresan típicamente en los artículos científicos a través de los intervalos de confianza. Por ejemplo, se estima que la población de cierta especie de aves en una región es de 12.000 ejemplares, con un intervalo de confianza, al 95% de probabilidad, entre 9.000 y 15.000. Esto quiere decir que, aunque los investigadores creen que el tamaño de la población está en torno a 12.000 individuos, para tener una seguridad del 95% de que están acertando, han de ampliar el rango de valores posibles desde 9.000 hasta 15.000.

Todos esos ejemplos representan datos imprecisos con los que trabajamos a diario y que nos sirven para tomar decisiones de todo tipo. Sabemos de su inexactitud, pero podemos vivir con ella, porque aunque no sean perfectos son *suficientemente buenos* para el uso que les queremos dar. Lo cual no quiere decir que en cada una de las disciplinas del conocimiento tanto los investigadores como los gestores no estén siempre tratando de conseguir mejores datos para tomar decisiones mejor informadas. Pero ni pueden ni necesitan esperar a tener todos los datos, con la máxima precisión, para continuar haciendo avanzar el conocimiento o para adoptar sus resoluciones. Porque sí, hay una diferencia importante entre que la población de un ave sea de 9.000 o de 15.000 ejemplares (un 67% más)..., pero saber que está en ese rango, que no son 600 ni 70.000, puede ser muchas veces suficiente para el científico que quiere entender su impacto sobre el ecosistema correspondiente o para la autoridad gubernamental que tiene que tomar decisiones sobre el periodo de veda adecuado.

Esto no quiere decir tampoco, claro, que todo valga. Ciertamente, en algunos casos, nos encontramos con *datos basura*, absolutamente inútiles, ni siquiera con una gran tolerancia para

la incertidumbre. Es el caso, en el área de las ciencias sociales, de las seudoencuestas, que se basan en muestras autoseleccionadas de lectores de un cierto medio de comunicación o reclutadas por el sistema de bola de nieve entre usuarios de redes sociales, lo que les lleva a carecer de toda representatividad y no servir para prácticamente nada más que entretener al público, llamar su atención y vender esos ojos u oídos a los anunciantes.

Por tanto, al enfrentarnos a nuevos datos, nuestra tarea no es tan simple como identificar y desechar los malos y quedarnos con los buenos. Es, más bien, averiguar lo más posible sobre cómo se han generado esos datos haciéndonos preguntas como las que han ocupado otros artículos de esta sección: ¿cómo se han definido los conceptos que se trata de estudiar? ¿Cómo se han medido? ¿Son datos de toda la población o de una muestra y, en ese caso, cómo se ha seleccionado? ¿Existen sesgos importantes en el proceso de selección de la muestra que limiten su representatividad? ¿Son coherentes los nuevos datos entre sí y con datos de otras fuentes? ¿Encajan con lo que sabemos de otros lugares u otros tiempos? En algunos casos, esas preguntas nos llevarán a reconocer y descartar algunos casos catastróficos, de datos insalvables para hacer con ellos ningún aprendizaje. La mayor parte de las veces, sin embargo, no será así, y encontraremos datos en algún punto de una amplia gama de posibilidades entre la perfección ultraprecisa (las acciones de cierta empresa se vendían al cierre de la bolsa a 3,43 euros) y la estimación aproximada, pero suficientemente buena para nuestros fines, como cuando decimos que el 21 de abril cayeron 2,5 litros de agua de lluvia por metro cuadrado en Sevilla. En realidad, lo hicieron solo en un lugar muy concreto de su aeropuerto, donde está colocado el pluviómetro del observatorio; no obstante, sabemos, por la gran variabilidad espacial del fenómeno de la lluvia, que seguramente ese día se recogieron cantidades bastante diferentes en otros puntos de la ciudad.

En esta línea, los datos con los que contamos sobre el coronavirus son imperfectos, como todos, pero *suficientemente buenos* para muchas cosas. A pesar de todas las dificultades e inconsistencias entre unos lugares y otros, y a través del tiempo, en la forma de contar los casos conocidos, los test, los fallecidos, los hospitalizados…, esas series de datos, procedentes de diversas fuentes, se comportan con considerable sintonía entre ellas. Ello es un indicio muy importante de que están midiendo razonablemente bien la realidad, y nos permiten saber con bastante certeza muchas cosas. Sabemos, con pocas dudas, qué países se encuentran entre los más afectados y cuáles lo están mucho menos (al menos, entre los países demo-

cráticos y desarrollados). Sabemos, dentro de cada país, que hay unas regiones mucho más afectadas que otras, porque todos sus datos (casos, hospitalizaciones, fallecidos medidos de diferentes maneras) nos lanzan un mensaje parecido. Sabemos, por la evolución de esas diferentes medidas interrelacionadas, en qué países y regiones la epidemia está aún expandiéndose, en cuáles se ha frenado y en cuáles está en retroceso. Sabemos, por la comparación entre los datos de fallecidos registrados con COVID-19 y los datos de excesos de mortalidad, que hay en casi todas partes un número importante de fallecidos causados por la enfermedad y no contabilizados; y que ese número no es, salvo excepciones en algunas regiones, dos, tres o cuatro veces más que el número oficialmente reconocido, sino más bien en torno a un 40%-60% adicional a la cifra oficialmente contabilizada. Sabemos que los confinamientos funcionan, puesto que en casi todas partes, en países y regiones con muy diferente prevalencia de la enfermedad, los casos detectados y los fallecimientos dejan de crecer en una fecha muy similar, en torno a unos 14-16 días después de la entrada en vigor de las medidas de confinamiento.

Naturalmente, querríamos que los datos fueran mucho mejores y que pudiéramos saber muchas más cosas, con mucha más precisión. Pero es inevitable que los datos obtenidos en tiempo real, en mitad de una epidemia, en la que la prioridad es salvar vidas, no tengan la calidad que a todos nos gustaría. Porque, en general, el proceso de datos es mucho más laborioso de lo que tendemos a pensar, y por eso, incluso en circunstancias normales, datos aparentemente sencillos como la población de los municipios o las estadísticas de mortalidad se publican muchos meses después de su fecha de referencia.

En definitiva, casi todos los datos que manejamos sobre casi todos los asuntos de interés público son, en alguna medida, imprecisos. Eso no los hace inútiles. Lo que es necesario, en cada caso, es hacer un ejercicio cuidadoso de comprensión de lo que miden y de cómo lo hacen, de sus posibles lagunas, sus sesgos, sus incoherencias con otras fuentes. Además, hay que tener presente que la información extraída puede ser borrosa, y que no nos permitirá estimar diferencias muy pequeñas entre lugares, cambios mínimos en las tendencias. Sin embargo, eso no significa que no sea valiosa, porque el conocimiento aproximado es *suficientemente bueno*, en muchas circunstancias y, desde luego, mucho mejor que la ignorancia sin evidencia alguna.