David Valls

ALAN E. GELFAND

Profesor de Estadística en la Universidad de Duke (Durham, EE UU)

«A LA GENTE LE RESULTA MUY DIFÍCIL ENTENDER LAS ESTADÍSTICAS»

David Conesa y Felip Pineda

«Elegí la estadística simplemente porque pensaba que era más interesante y tenía más vida que las matemáticas puras, aunque tengo formación matemática.» Alan Gelfand explica de esta forma su pasión por la estadística, a la que ha dedicado su vida. En los noventa popularizó, junto con Adrian Smith, el método Markov Chain Monte Carlo (MCMC), contribución que ha mejorado considerablemente el subconjunto de la estadística bayesiana mediante el muestreo de distribuciones de probabilidad. «Es una de esas cosas con las que se puede uno sentir muy afortunado porque hay un montón de gente muy inteligente que trabaja mucho para sacar adelante su carrera y resulta que no encuentran nada. Y uno descubre algo y resulta ser un gran avance. Lo único que puedo decir en mi favor es que por lo menos lo aproveché», recuerda.

Durante los últimos años, la investigación de Alan Gelfand ha girado en torno a la estadística espaciotemporal, un campo en auge con una gran cantidad de posibilidades, como explica humildemente, pero con pasión: «Hay una expresión que usamos en inglés, *low-hanging fruit* ("fruta al alcance de la mano"), la que se puede coger sin demasiado trabajo, y en el análisis espaciotemporal había tanta fruta al alcance de la mano que se podía jugar y disfrutar de todas las posibilidades que ofrecía. Muchas otras áreas se han desarrollado tanto, se ha trabajado tanto en ellas, que tienes que alcanzar mucho más alto para poder encontrar algo de fruta. He tenido mucha suerte.» De hecho, ha publicado cuatro libros y más de 250 artículos científicos relacionados con estas cuestiones, y ha recibido varios premios. El último de ellos, la medalla a la distinción en la investigación de la sección de estadística y medio ambiente de la ASA (American Statistical Association).

Actualmente Alan Gelfand es profesor de Estadística en la Universidad de Duke (Durham, EE UU) y miembro de la ASA, el Instituto de Estadística Matemática y el Instituto Internacional de Estadística. Se pueden encontrar aplicaciones relacionadas con su investigación en áreas como la exposición ambiental, los procesos ecológicos espaciotemporales y el desarrollo de modelos climáticos.

«NO ES LO MISMO EL TRABAJO CON 'BIG DATA' QUE LA ESTADÍSTICA. NOSOTROS INTENTAMOS ENTENDER PROCESOS COMPLEJOS, EXPLICAR, PREDECIR, CAPTAR LA INCERTIDUMBRE»

¿Cuál es su visión sobre el uso aplicado de la estadística?

El estadístico del siglo XXI es un conjunto de cosas. Debe tener un buen historial metodológico, ser bueno con la modelización, con la computación, con el análisis de datos, debe poder hacer visualización si es necesario, y tiene todo un conjunto de herramientas que le permiten hacer las cosas que se necesitan hoy en día. Puede darse el caso de que diferentes personas sean más fuertes en las diferentes áreas, pero sí que es necesario algo así como el paquete completo si se quiere ser realmente un estadístico moderno. Y el resultado de todo eso es una revitalización de la estadística. La estadística ha recibido cierta energía, cierta vitalidad, gracias a todo este desarrollo interdisciplinario, a toda esta colaboración. Y no importa mucho si se trata de colaborar con las ciencias medioambientales, con la genómica, la neurociencia, las ciencias sociales o la economía. En el siglo XXI el estadístico es un jugador integrado desde el principio y científicamente esto es mucho más emocionante, porque eres verdaderamente parte del equipo. Así que es un momento muy emocionante para la estadística, en términos de aplicaciones, quiero decir.

¿Qué es exactamente el pensamiento bayesiano?

La perspectiva bayesiana es científicamente intuitiva y el único desafío al que se ha enfrentado el pensamiento bayesiano vino de dentro de la comunidad estadística, no tanto de la comunidad científica general. La idea que hay detrás del pensamiento bayesiano es lo más simple que se pueda imaginar. Se infiere lo que no se sabe partiendo de lo que se observa. ¿Hay algo más na-

tural que eso? Sin embargo la visión más clásica está en cierto modo invertida. La visión clásica dice: ¿qué podría usted imaginar partiendo de lo que no sabe? Es una forma extraña de ver las cosas; un punto de vista peculiar. Yo creo que por eso el paradigma bayesiano para la inferencia es tan atractivo, y la razón por la que nos costó tanto tiempo pasar al frente de la estadística fue el problema computacional. Por eso la comunidad bayesiana solo podía resolver problemas muy simples. Los bayesianos pasamos mucho tiempo jugando con la teoría axiomática formal y la perspectiva matemática, que fue importante en su fundación. Pero desde el punto de vista aplicado, desde el punto de vista práctico, realmente no ayudaba a la gente.

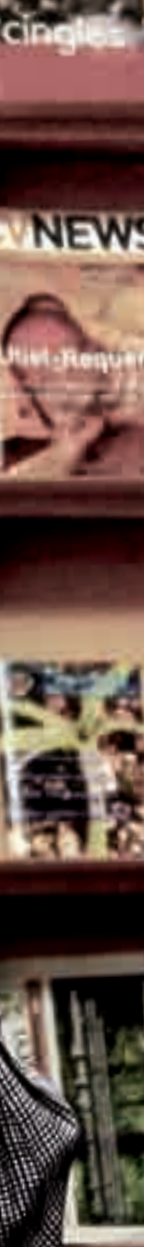
¿Qué significó el método Markov Chain Monte Carlo (MCMC) para la estadística bayesiana en su momento?

Para un público profano, el problema computacional bayesiano tiene que ver con la integración de datos de alta dimensionalidad y otras integraciones complicadas que no se podían hacer de forma explícita. Necesitábamos un mecanismo para romper lo que se conocía en aquel momento como la maldición de la dimensionalidad, porque los problemas de alta dimensionalidad no se podían resolver analíticamente, y necesitábamos una estrategia. Y la estrategia resulta ser la idea más básica en toda la estadística. Si se quiere aprender sobre una población, se hace un muestreo a partir de esa misma población. Eso es lo que hacemos, y resulta que el muestreo de Gibbs y el método Markov Chain Monte Carlo son mecanismos que permiten la toma de muestras en poblaciones complicadas, en distribuciones multidimensionales complicadas. La verdad es que es una idea muy elegante. Pero ¿por qué funciona como funciona? Porque era imposible muestrear una distribución de muy alta dimensionalidad, digamos por ejemplo una distribución de cien o de mil dimensiones, pero sí se puede romper en pedazos más pequeños y muestrear distribuciones de baja dimensionalidad. Si lo hiciéramos de la manera correcta, con la toma de muestras de las distribuciones de baja dimensionalidad obtendríamos muestras de la distribución de alta dimensionalidad, que es lo que realmente nos interesa. Ése era su poder y la primera vez que lo vimos nos sentimos como un niño con zapatos nuevos.

También ha estado trabajando en la estadística espacial. ¿Cuál fue su experiencia en este campo?

Me di cuenta de lo adecuada que era la estadística espacial para la inferencia bayesiana. La idea es que, para gran parte del trabajo estadístico, la gente usa la probabilidad y la inferencia probabilística, y cuando se juntan la incertidumbre y la inferencia probabilística, la incertidumbre proviene generalmente de las ideas asintóticas. Y lo que pasa con la estadística espacial es que el análisis asintótico no funciona y el único análisis asintótico que se podía usar en la estadística espacial era el de series temporales, en las que se dejaba que el tiempo llegara hasta el infinito. Pero con el espacio no se puede dejar que llegue hasta el infinito. Si queremos estudiar una región, Valencia o incluso España, no vamos a ocuparnos de una región del tamaño de la Tierra. Lo que en realidad queremos es pensar en una inferencia en la que las muestras aumentan de tamaño si observamos más localizaciones dentro de la región, no si hacemos aumentar de tamaño la región. Y ese tipo de comportamiento asintótico no funciona con las series temporales clásicas ni con otro tipo de teoría asintótica, así que resulta que Bayes viene muy bien para esto porque nos da la inferencia exacta y no requiere ningún argumento asintótico. El espacio es interesante porque cuando se trabaja con series temporales en una dimensión existe el orden, y se puede distinguir qué va antes y qué va


David Valls



después. Pero en los espacios no hay orden. Hace que la vida sea más interesante, más exigente, más divertida, y nos ha abierto muchas posibilidades. La combinación de espacio y tiempo es realmente hermosa.

¿En qué dirección cree usted que va a progresar la estadística? ¿Cree que será más teórica o más aplicada?

La contribución de la estadística es en realidad la inferencia, por lo que la pregunta es si podemos mantener la inferencia como contribución científica o no. Aquí es donde entran los desafíos del *big data*, y la pregunta de si necesitamos las estadísticas, si necesitamos la inferencia, o si podemos operar solo con algoritmos, si podemos hacer ciencia solo con resúmenes descriptivos y explorar bases de datos sin la necesidad de recurrir a la inferencia probabilística. Antes de ello, y resulta interesante debido a la crisis de identidad a la que la estadística se va a enfrentar, la pregunta es: ¿hacia dónde debemos dirigirnos? Y hay algunas personas que dirán: «Bueno, siempre necesitaremos la parte teórica, porque necesitamos cierto rigor, asegurarnos de que la gente está haciendo las cosas de la manera correcta». Y es cierto, si ofreces herramientas, la gente las utiliza. Pero puede ser peligroso si la gente no ha analizado con rigor los retos fundamentales que parte de este trabajo comporta. Así que creo que siempre habrá un lugar para la parte teórica, pero no hay duda de que el futuro está en la parte aplicada. Y, en particular, lo que realmente está cambiando es el estilo de hacer las cosas. Tiene que cambiar la forma en la que pensamos en las cosas, porque es necesario un pensamiento más integral, mayor síntesis de diferentes fuentes de información, porque no somos capaces de controlarlo todo. Estamos fundamentalmente ante un proceso más complejo, con características que debemos intentar captar mediante todas las maneras posibles. En resumen, el futuro de la estadística reside en trabajar en proyectos interdisciplinarios complicados, en trabajar en proyectos y sistemas exigentes.

«SI LOS ESTADÍSTICOS NO SOMOS LO SUFICIENTEMENTE VISIBLES, LA GENTE SIMPLEMENTE ASUMIRÁ QUE NO TENEMOS MUCHO QUE APORTAR»

Viktor Mayer-Schönberger y Kenneth Cukier comienzan su libro sobre *big data*[1] explicando cómo las búsquedas en Google predijeron la propagación de la epidemia de gripe H1N1 en 2009. Este ejemplo les sirve a los autores para citar el artículo de Chris Anderson. «The End of Theory: The Data Deluge Makes the Scientific Method Obsolete»[2], publicado en 2008 en la revista *Wired*, en el que el autor afirmaba de manera provocadora que «los petabytes nos permiten decir: "La correlación es suficiente". Podemos dejar de buscar modelos». ¿Qué piensa usted sobre ello? ¿El futuro de la estadística pasa por ser un simple análisis descriptivo de datos?

Antiguamente me consideraba un estadístico, pero, si me preguntas ahora, me consideraría un modelador estocástico, es decir, modelo problemas con incertidumbre. Mi visión del mundo es que cuando nos fijamos en estos procesos complejos no somos capaces de explicarlos de manera perfecta y, por lo tanto, si no podemos explicarlos de manera perfecta, debemos de introducir un cierto grado de incertidumbre, y para mí eso es lo que significa el modelado probabilístico o estocástico. Esto no lo consideraría del todo una idea del *big data*. Hay muchos problemas que no requieren *big data*. Puedes trabajar en problemas científicos interesantes de mucha menor escala, pero lo que es un verdadero problema en este campo es que el aprendizaje automático y el *big data* tratan de imaginar un gran conjunto que comprenda la estadística, la informática, quizás la ingeniería informática y las matemáticas aplicadas. Realmente esa no es la manera en la que quiero pensar sobre la ciencia. Creo que gran parte de estas cosas del aprendizaje automático son una especie de búsqueda de estructura en los *big data*, una búsqueda de patrones, de relaciones. Pero no es lo mismo que lo que hacemos, nosotros intentamos entender procesos complejos, explicar, intentamos predecir, captar la incertidumbre. No es la misma manera de ver las cosas, no puede responder a las mismas preguntas. Necesitamos ciencia con datos, está claro, y la ciencia no es ciencia sin estadística. Realmente creo que el enfoque de la ciencia que estoy describiendo es muy diferente a simplemente explorar una base de datos enorme y tratar de sacar alguna estructura de ella, que es lo más común en el aprendizaje automático.

Nos gustaría hablar de la visibilidad de la estadística. ¿Qué deben hacer los estadísticos para hacerla más visible?

El año pasado fue el Año Internacional de la Estadística, una oportunidad para que pensáramos un poco más en

[1] Mayer-Schönberger, V. y K. Cukier, 2013. *Big Data. La revolución de los datos masivos*. Turner. Madrid.

[2] Anderson, C., 2008. «The End of Theory: The Data Deluge Makes the Scientific Method Obsolete». *Wired*, 16 (7). Disponible en: <http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory>.

cómo vamos a seguir participando en la comunidad científica más grande. Nos vino bien. Históricamente, los estadísticos se han contentado con estar en un segundo plano. Hemos sido gente de baja visibilidad. No necesitábamos estar en *El País* o en *The New York Times*. No necesitábamos salir por televisión. Es decir, el líder científico del proyecto solía ser un estadístico al que no le importaba estar en segundo plano. Creo que eso tiene muchas cosas buenas, en el sentido en que hace que la comunidad estadística sea más amistosa, un poco menos cruel, un poco más solidaria. Porque no buscamos tanto la gloria, no competimos tanto, nos apoyamos más, y esa clase de inocencia es en verdad algo muy bueno. No quiero perder eso, pero reconozco que, si no somos lo suficientemente visibles, la gente simplemente asumirá que no tenemos mucho que aportar. Asumirán que un científico de otra especialidad puede hacer lo que hacemos nosotros y que no tenemos un papel muy importante que desempeñar. Realmente, no sé cómo evitar este problema, porque no quiero sacrificar la inocencia, pero por otra parte tampoco quiero estar ahí fuera en busca de la gloria. No es una motivación. Para mí lo importante es la ciencia. Ejercer buena ciencia, tratar de resolver los problemas importantes, y tratar de hacer una contribución a los problemas importantes. Creo que ese es el reto en realidad.

David Valls

¿Cuál es su opinión sobre el futuro de la financiación a la investigación?

Hace tiempo podías proponer un programa de investigación que fuera estadística teórica pura y dura y, bueno, a lo mejor no te hacías rico, pero podías obtener un poco de dinero para financiarla y quizás te daban dinero para contar con uno o dos estudiantes, para unos pocos viajes, y esa era más o menos la manera de hacer las cosas. Pero ahora, ese tipo de investigación no va a ser fácil de financiar, por lo menos en los Estados Unidos –y sé que en el Reino Unido, y tal vez en España, tampoco–, como la crisis ha dificultado tanto las cosas, conseguir financiación es más complicado, pero sé que el modelo que los organismos financiadores buscan es una idea de proyecto interdisciplinar. Creo que seguramente este será el futuro de la financiación. La dificultad es que la cantidad de fondos se está reduciendo porque la gente ajena a la comunidad científica no valora demasiado la ciencia básica, cree que el gobierno no debería gastar dinero en ciencia cuando hay otras cosas en las que debería gastarlo, y quizás que no debería haber gastado tanto dinero en primer lugar. Y así, lo que está sucediendo es que la cantidad de financiación se está reduciendo pero hay más presión porque, por lo menos en los Estados Unidos, las universidades esperan que su profesorado les reporte dinero. Así que es una partida que no pueden ganar. No sé cómo se va a desarrollar. Algunas personas sugieren que la industria privada ayude. Yo no estoy tan seguro, creo que las industrias privadas tienen sus propias prioridades. No creo que podamos esperar que proporcionen la financiación que queramos para mantener la investigación científica básica. No estoy seguro de que haya una solución.

«HAY UN VERDADERO RETO EN EL MAL USO DE LOS DATOS Y ESTO CREA ESCEPTICISMO EN LA MENTE DEL PÚBLICO, CREA UNA VERDADERA DUDA SOBRE LA VALIDEZ DEL ANÁLISIS ESTADÍSTICO»

El mal uso de la estadística en los medios de comunicación puede dar lugar, a propósito o sin querer, a la manipulación de las cifras. ¿Cree que esto es un problema común?

La vieja mentira solía ser, al menos en inglés, *liars, damned liars and statisticians* ("mentirosos, malditos mentirosos y estadísticos"). La idea era que los estadísticos podían abusar de los datos. No hay duda de que se ha hecho. Lo que tenemos ahora es una gran

presión para publicar, y esa gran presión para publicar comporta una presión por encontrar cosas. Ese tipo de presión conduce a un mal uso o una tergiversación de los resultados estadísticos. Esto se une al desafío al que llamamos reproducibilidad. Por ejemplo: ¿puedes dar un conjunto de datos sobre el que has escrito algo a otro equipo científico, y pueden ellos reproducir lo que has averiguado? Y la historia de esto es notablemente decepcionante, depende del campo, pero hace poco escuché una maravillosa charla sobre la reproducibilidad de estadísticas médicas, y los números asustan. Se ha podido reproducir menos del 50 % de los conjuntos de datos que han sido publicados. Hay un verdadero reto en el mal uso de los datos y esto crea escepticismo en la mente del público, crea una verdadera duda sobre la validez del análisis estadístico, y la sensación entre el público de que un estadístico puede contarte la historia que quiera porque puede manipular los datos a su antojo. Y en cierto sentido es verdad.

¿Considera usted que la población está preparada para entender correctamente el análisis estadístico en los medios de comunicación?

Estoy profundamente preocupado por las nociones numéricas, como lo llamamos nosotros, de la población. No creo que mucha gente entienda ni siquiera los órdenes de magnitud. Sé, por ejemplo, que si entro en una pequeña tienda de los Estados Unidos, compro una barrita de chocolate y doy veinte dólares, la persona de detrás de la máquina registradora no sabe cuánto dinero me tiene que devolver. No tienen ni idea de aritmética. La única forma que tienen de saberlo es introducir el número y esperar a que la máquina les diga el cambio. Y sé que eso no solo pasa en los Estados Unidos, sino en todas partes. Creo que hay una pérdida completa de la magnitud de los números, por lo que a la gente le resulta muy difícil entender las estadísticas. Además de todo esto, cuando proporcionamos relaciones estadísticas, aunque la mayoría de la gente sea capaz de entender los promedios, quizás las proporciones, como los porcentajes, cuando intentas ir más allá, hasta la incertidumbre o la variabilidad, la gente tiene muchas dificultades con estas cuestiones, incluso los que tienen una buena formación. Todas estas cosas dificultan mucho que el público acepte y comprenda las estadísticas a gran escala. Creo que hemos malacostumbrado a la gente a no apreciar los números porque todos podemos hacer cálculos sencillos en un móvil y todo está automatizado cuando compramos, sin que tengamos que hacer nada.

David Conesa. Profesor del departamento de Estadística e Investigación Operativa. Universitat de València.
Felip Pineda. Periodista. Revista *Mètode*, Universitat de València.