Boletín de Estadística e Investigación Operativa
Vol. 30, No. 3, Noviembre 2014, pp. 313-324

Opiniones sobre la profesión

Big Data and Statistics: Trend or Change?

Daniel Peña
Departamento de Estadística
Universidad Carlos III de Madrid
✉ daniel.pena@uc3m.es

Abstract

The analysis of Big Data is not a trend in statistics but a turning point. It is expected to produce deep changes in the teaching of statistics as well as in future research in this field. This article discusses some of these transformations.
Keywords: Multivariate analysis, Time series, Statistical computing.
AMS Subject Classifications: 6262A,6262H

1. Introducción

La aparición de los ordenadores supuso para la Estadística el primer cambio paradigmático, en el sentido de Khun (1962), desde su consolidación como disciplina científica a principios del siglo XX. Los métodos desarrollados por K. Pearson and R. A. Fisher en Inglaterra en el primer cuarto del siglo XX establecieron las bases de la Estadística como la ciencia para analizar los datos y construir modelos matemáticos que expliquen la realidad. En los inicios del siglo pasado empieza a desarrollarse la necesidad de necesidad de disponer de métodos para comprender la información en los los datos existentes, que constituan en general una muestra pequeña del fenómeno de interés. Sólo encontrabamos bancos de datos de tamaño medio con medidas biométricas (altura, peso) o sociales (nacimientos, pago de impuestos), lo que explica la importancia de las aplicaciones en estos campos en el nacimiento de la Estadística. Sin embargo, en el resto de las aplicaciones físicas, técnicas o agronómicas, cada dato era generalmente una medida recogida con esfuerzo, con frecuencia tras un proceso costoso de monitorización, cuya información era necesario aprovechar al máximo. Por ejemplo, cuando Fisher desarrolló el diseño de experimentos en la estación agraria de Rothamstead tuvo que esperar varios meses para disponer del rendimiento en las parcelas sometidas a cambios de condiciones experimentales. Además, las escasas

posibilidades de cálculo en esa época hacían ajustar una regresión lineal, o calcular el valor del estadístico de un contraste de ajuste, trabajos complicados con esfuerzos computacionales apreciables. Esta situación cambió radicalmente con la introducción del ordenador, y en la sección siguiente analizaremos brevemente sus consecuencias.

La generalización de Internet, las nuevas tecnologías de información y telecomunicaciones, el cálculo distribuido y en paralelo y la disminución de los costes de procesamiento y almacenaje de los grandes bancos de datos, los llamados Big Data, pueden dar lugar a un segundo cambio paradigmático en nuestra disciplina. En la actualidad, cualquier medio digital, un termostato, una página web, una red social o una tarjeta de crédito, genera continuamente datos sobre su funcionamiento. Esta información, que crece de día en día de forma exponencial, puede fácilmente compartirse en internet, siendo por tanto accesible a un público cada vez más amplio. Los métodos estadísticos creados por K. Pearson y R. Fisher estaban pensados para muestras pequeñas y tienen limitaciones para el análisis de datos que no son homogéneos, ni en su distribución ni en su formato. El enfoque estadístico tradicional parte de que disponemos de una muestra de una población, con datos medidos con precisión. En los nuevos bancos de datos nos encontramos con medidas heterogéneas: constituyen una mezcla de distintas poblaciones y formatos, al incluir imágenes, gráficos o textos. El primer paso del análisis es comprender esta enorme complejidad.

Un error frecuente en la ciencia ha sido pensar que al crecer el tamaño de un problema, que sabemos resolver a pequeña escala, los métodos establecidos se aplicarán con pequeños ajustes al problema de mayor dimensión. Sin embargo, es bien conocido que al aumentar la velocidad de un objeto y aproximarse a la de la luz, la física clásica deja de ser aplicable y tenemos que utilizar las ecuaciones de la relatividad. En el mismo sentido, si descendemos a escala microscópica es necesario cambiar el enfoque clásico para entender las nuevas fuerzas que actúan a ese nivel. En otro contexto, el efecto de un medicamento no es lineal con la dosis y el compuesto que puede, en pequeñas dosis, ayudarnos a conciliar un sueño reparador, puede producirnos la muerte en cantidades mayores. Este hecho ocurre también en Estadística, donde herramientas muy útiles en pequeñas muestras, como los contrastes de ajuste, son irrelevantes en muestras de millones de datos reales, donde siempre se rechazará un contraste de normalidad.

En este trabajo vamos a analizar algunas de las implicaciones del estudio de estas grandes masas de datos y como la Estadística tendrá que modificarse para adaptarse a las nuevas necesidades. El artículo se organiza como sigue. En la sección siguiente comenzamos con una nota de prudencia, repasando brevemente la dificultad de prever los avances científicos impulsados por cambios tecnológicos. En efecto, estadísticos muy notables de la segunda mitad del siglo XX no fueron capaces de anticipar la revolución en la metodología estadística que iba a producir el ordenador. En la sección 3 comentaremos las nuevas necesidades

de almacenamiento y cálculo que han aparecido con los grandes bancos de datos actuales y las nuevas herramientas de cálculo paralelo y distribuido, como Hadoop. La sección 4 analiza algunos de los cambios previsibles en la metodología y los métodos estadísticos que van a impulsar la combinación de Big Data con los avances en computación y almacenaje. La sección 5 discute algunas de las implicaciones de estos cambios en la organización de los departamentos universitarios y en la enseñanza de la Estadística. El artículo finaliza con unas breves conclusiones.

2. Los cambios producidos por los ordenadores: la primera revolución en la Estadística

La aparición de los ordenadores supuso cambios fundamentales en la Estadística pero en sus inicios esos cambios fueron difíciles de predecir. Un análisis más completo de las ideas aquí expuestas puede consultarse en Peña (1995). Varios autores han analizado como el desarrollo de los métodos estadísticos ha dependido de las demandas en otras disciplinas científicas y de los métodos de cálculo existentes. Yates (1966) ha ilustrado la importancia de las calculadoras de mesa en el desarrollo inicial de la Estadística y cómo sus limitaciones influyeron en los trabajos desarrollados por Fisher en Rothamsted. La interacción entre la potencia de cálculo y los desarrollos analíticos aparece también claramente expuesta en los libros de Box (1978) y Stigler (1986).

Efron (1979) ha descrito con precisión como la aparición del ordenador ha cambiado la perspectiva de que modelos eran inabordables. Antes, esta distinción dependía del número de cálculos necesarios, que dependía del tamaño de los datos disponibles. Con el ordenador, esta escala cambia, y se desarrollan métodos que requieren cálculos muy intensos con pequeñas muestras y nuevas formas de abordar el análisis de tamaños inimaginables en el pasado. El Bootstrap, los métodos de validación cruzada, los métodos de estimación robusta, el análisis de datos censurados y la creación del algoritmo EM o los métodos MCMC de estimación Bayesiana son algunos ejemplos de esta transformación de la metodología estadística en el último cuarto del siglo XX. Además, la posibilidad de estimar modelos más complejos ha impulsado el desarrollo de áreas como las series temporales, el análisis multivariante, los métodos no paramétricos o la estimación Bayesiana.

Es interesante resaltar que en los inicios de este proceso, en los años sesenta, estos cambios no fueron previstos. En 1965 John W. Tukey, uno de los estadísticos más grandes de su época y una de las figuras líderes en la transformación de la Estadística en la segunda mitad del siglo XX, predecía que los ordenadores cambiarían esta disciplina en cuatro direcciones principales (Tukey, 1965): (1) sustitución de las tablas de las distribuciones por programas informáticos; (2) mayor énfasis en los métodos basados en el orden de las observaciones; (3) méto-

dos de Monte Carlo de experimentación con el ordenador y (4) procedimientos más eficientes de realizar los cálculos algebraicos. Otros famosos estadísticos de la época escribieron sobre el cambio en términos similares.

Cincuenta años después vemos que estas predicciones fueron muy limitadas. El ordenado efectivamente ha permitido olvidarse de las tablas de las distribuciones, pero este efecto ha sido de muy escasa importancia. Los métodos basados en el orden de observaciones han tenido cierta vigencia en la Estadística Robusta, pero su alcance ha sido limitado: mucho más impacto han tenido los métodos iterativos de cálculo, que el ordenado ha hecho posible. Los métodos de Montecarlo sí han tenido un papel central en la evolución de la Estadística en la segunda parte del siglo XX, pero no sólo para comparar un método nuevo con los anteriores, como bien preveía Tukey, sino como base de nuevos procedimientos de estimación, como los métodos MCMC, que han hecho posible el ajuste de modelos mucho más complejos. Se ha seguido avanzando en procedimientos para realizar más eficientemente cálculos algebraicos, pero muchos de estos cálculos han sido sustituidos por métodos de remuestro, como el Bootstrap de Efron.

Desde el punto de vista de cambios metodológicos Tukey tuvo más éxito: predijo avances en utilizar modelos más flexibles y un crecimiento de los procedimientos gráficos y exploratorios, que efectivamente han ocurrido. Sin embargo, no intuyó el cambio de paradigma de pasar de trabajar con un modelo único a hacerlo con un conjunto de modelos, como en Bayesian Model Averaging o en predicción adaptativa. Tampoco, el auge de las técnicas de selección de modelos, que han ido sustituyendo a los contrastes de ajuste, y a otros tests, por ejemplo de atípicos, en el trabajo con modelos estadísticos.

Este breve análisis nos indica la dificultad de prever los cambios de una disciplina cuando vienen impulsados, no tanto por su lógica interna, sino por los avances tecnológicos que abren posibilidades antes insospechadas. Es importante tener en cuenta esta experiencia para poner en contexto los comentarios que vienen a continuación.

3. Los Bancos de datos actuales y su tratamiento: Hadoop

La utilización masiva del ordenador como instrumento de recogida de datos en procesos industriales y comerciales permite crear grandes masas de datos dinámicos y multivariantes. Estos datos pueden ser numéricos, pero también imágenes, textos o funciones. La recogida digital de información por lectores ópticos y sensores permite obtener, con un coste marginal despreciable, tantas medidas como se desee. Por otro lado, tanto las redes sociales, como internet en general, proporcionan información masiva sobre los comportamientos, acciones y decisiones de los usuarios. Además, estos bancos de datos crecen continuamente ya que reciben constantemente información adicional. El avance en las telecomunicaciones hace posible acceder a estos datos de forma inmediata y sin coste

a un número creciente de usuarios potenciales.

Analicemos brevemente el crecimiento de las masas de datos. Hace muy pocos años el tamaño de almacenamiento de un ordenador personal era menor de un gigabyte (GB=109 bytes), y hoy hemos pasado ya a terabytes (TB =1012b) , mientras que los servidores se mueven en petabytes (PB=1015 b). Por ejemplo, la colección impresa de la biblioteca del congreso de los EE.UU ocupa actualmente del orden de 10 terabytes, pero el World Data Centre for Climate El WDCC (Centro Mundial de datos para el clima), una de las bases de datos más grande del mundo, almacena unos 400 terabytes de información sobre el clima en todo el mundo. El National Energy Research Scientific Computing Center, NERSC tiene una base de datos de 2.8 petabytes, y Google, que recibe más de 100 millones de consultas al día, se supone que es capaz de almacenar cientos de terabytes de información. Lesk (1977) estimó que toda la información existente a finales del siglo XX podría guardarse en unos poco miles de petabytes. Sin embargo, algunos estudios (Hilbert y López, 2011) evalúan la necesidad actual de almacenamiento en exabytes (EB=1018 b) y, dentro de poco, en zettabyte (ZB= 1021 b).

La evolución de los sistemas operativos y programas de ordenador (software) ha sido paralela a los cambios en los equipos (hardware). En particular, los programas para el cálculo estadístico han pasado de los trabajos por lotes en los 70 y 80, donde se pedía al ordenador una operación concreta cada vez, (como en las versiones antiguas de BMDP o SPSS) a los programas interactivos actuales, concebidos para aplicar distintos tipos de análisis a un mismo conjunto de datos y que permiten acceso directo a los resultados intermedios y capacidad de programación dentro del paquete (como en SAS, S+, SCA, MATLAB, GAUSS etc) y a lenguajes orientados a objetos, que permiten manejar indistintamente funciones, variables o gráficos. La aparición de R, en los años 90, como lenguaje de libre disposición a partir del lenguaje S+, ha creado un estándar donde cientos de investigadores de todo el mundo incorporan nuevas rutinas ampliando cada día las capacidades de análisis. La mayor parte de las funciones de R están escritas en el mismo R, aunque para algoritmos computacionalmente exigentes es posible desarrollar bibliotecas en el lenguaje C. Además, R puede integrarse con distintas bases de datos, aunque todavía debe evolucionar mucho para poder realizar los cálculos requeridos con bases de millones de datos como las actuales.

El siguiente salto es la computación en paralelo y distribuida y el almacenamiento en la nube. El cálculo en paralelo consiste en ejecutar muchas instrucciones simultáneamente. Esto exige una programación donde en lugar de resolver un problema secuencialmente se descompone en partes que pueden realizarse en paralelo en hardware con procesadores con varios núcleos, o con varios procesadores, que realizan los cálculos en paralelo y se comunican entre sí. El cálculo en paralelo muestra toda su potencia cuando se conectan varios ordenadores para que trabajen conjuntamente. Puede hacerse de forma remota, donde los ordenadores no están físicamente cerca y se conectan por la web, o formando un cluster

o grupo de ordenadores de potencia media, pero conectados entre sí mediante un sistema de red de alta velocidad (gigabit de fibra óptica por lo general). Además, debe existir un programa que controle la distribución de la carga de trabajo entre los equipos. Por lo general, este tipo de sistemas cuentan con un centro de almacenamiento de datos único.

Una infraestructura digital en código abierto, dentro de la licencia de la Fundación APACHE, es Hadoop, creado por Doug Cutting. Hadoop combina la computación en paralelo y distribuida permitiendo desarrollar tareas muy intensivas de computación dividiéndolas en pequeñas parte y distribuyéndolas en un conjunto tan grande como se quiera de máquinas. Al ser de licencia libre este software está siendo adoptado no solo por usuarios particulares sino también por grandes sistemas (Oracle, Dell, etc.), lo que está llevando a una aceleración tanto de su difusión como de sus prestaciones. A diferencia de las soluciones anteriores para datos estructurados, la tecnología Hadoop introduce técnicas de programación nuevas y más accesibles para trabajar en almacenamientos de datos masivos con datos tanto estructurados como no estructurados.

Un problema de la computación distribuida es que aumenta la probabilidad de fallo. Cuando el trabajo se hace en una sola máquina esta probabilidad es pequeña, pero si trabajamos con miles de máquinas, que tienen además que comunicarse entre sí, la probabilidad de un fallo parcial aumenta mucho: puede haber congestiones que retrasen las comunicaciones, un conmutador o un router pueden estropearse, una máquina puede quedarse sin memoria RAM disponible o sin espacio en disco. Entonces, los cálculos que deben estar finalizados para que el sistema funcione no estarán disponibles, causando interrupciones en todo el sistema. Hadoop se ha diseñado para manejar de forma muy robusta situaciones de fallos en los equipos y en las redes de transmisión. En un cluster Hadoop de ordenadores los datos se distribuyen entre todos ellos cuando se cargan, y una herramienta del software, The Hadoop Distributed File System (HDFS), divide grandes ficheros de datos en trozos que se manejan en distintos nodos del cluster. Además, cada trozo se replica en varias máquinas, de manera que un fallo en un ordenador tenga poco efecto sobre el conjunto. Cada máquina del cluster realiza sus cálculos independientemente, de acuerdo a un modelo de programación llamado “MapReduce”. En este sistema los datos se procesan independientemente con tareas llamadas Mappers y los resultados se ponen en común con un segundo conjunto de tareas llamado Reducers. El sistema maneja toda la información que se transmite entre los nodos. Una ventaja adicional de Hadoop es que si disponemos de más máquinas no es necesario reprogramar los cálculos y el sistema utiliza siempre de forma eficiente la capacidad de cálculo total disponible.

4. Big Data y los cambios en la Estadística

La introducción de los ordenadores ha hecho patente que la forma de cálculo condiciona el modo en que establecemos, estimamos y testamos los modelos estadísticos. Por ejemplo, el cálculo de los ordenadores tradicionales ha estimulado modelos estadísticos secuenciales. Con cálculos algebraicos, como la famosa formula de la inversa de la suma de dos matrices, en los modelos lineales podemos saber el efecto de introducir o eliminar una variable o una observación sin repetir todos los cálculos. Esto ha hecho posible métodos eficientes de diagnóstico y estudio de sensibilidad, incorporación y supresión de variables, etc.

Si disponemos de un conjunto muy amplio de datos que procesamos en paralelo cortándolo en partes, un problema central es comprobar si estos datos son homogéneos y pueden recombinarse, o en lugar de un modelo único los datos provienen de un conjunto de modelos distintos con zonas de transición entre ellos. Esta situación se observó primero en series temporales con los modelos no lineales threshold introducidos por Tong (1980), que permiten que según los valores pasados de algunas variables el modelo que generará los datos siguientes cambien en el tiempo. Esta misma situación puede aparecer en modelos estáticos, donde puede existir un amplio conjunto de modelos $M_1, \ldots M_k$ correspondientes a distintas zonas del espacio muestral. El concepto de robusted también requerirá una reformulación, porque interesa que un modelo sea robusto no solo antes observaciones atípicas, sino ante heterogeneidad más general.

Un problema central será como combinar información de fuentes muy diversas: datos numéricos, funciones, gráficos, imágenes, información de textos, y para ello habrá que desarrollar nuevos métodos de Meta Análisis (véase Olkin, 1995) en el marco de una análisis Bayesiano. En general los métodos Bayesianos son más flexibles para manejar distintos tipos de información, por lo que es esperable su crecimiento, aunque como complemento de los métodos clásicos o frecuentistas. Esta complementaridad proviene de que, en primer lugar, los métodos descriptivos sin modelo tendrán cada vez mayor importancia en situaciones en que prácticamente tenemos toda la población relevante a nuestra disposición, y en segundo, porque siempre será necesario suponer en algún momento que todo lo que hemos hecho puede estar equivocado, y encontrar métodos de chequeo de nuestras hipótesis que no dependan del modelo de partida, lo que requiere un enfoque frecuentista.

Los métodos dinámicos y multivariantes, tanto los clásicos como los desarrollados bajo el nombre de Machine learning and Data Mining, ganarán en generalidad. Un campo de especial relevancia serán los métodos factoriales y de reducción de la dimensión, pero orientados a un objetivo específico. Por ejemplo, las redes neuronales pueden verse como modelos factoriales. Sabemos que los métodos útiles para reducir la dimensión preservando la máxima variabilidad en los datos, como componentes principales, no son adecuados si queremos que

esta reducción de dimensión sea óptima para detectar grupos en los datos. Por eso Fisher desarrolló las direcciones óptimas discriminantes, cuya generalización son las de máxima curtosis, introducidas por Peña y Prieto (2001) o, desde otro enfoque, las direcciones de SVM de Cortes y Vapnik (1995). Además, estas direcciones óptimas van a depender del tipo de clusters que existan en los datos, por lo que es de esperar cambios importantes en el área de clasificación, que se convertirá en uno de los ejes de los análisis de grandes masas de datos.

En el campo dinámico algunos de los modelos desarrollados en los últimos años sólo tienen sentido con Big Data. Por ejemplo, los modelos factoriales dinámicos necesitan un número infinito de series para estar identificados y sus propiedades se establecen cuando tanto el número de observaciones en cada serie como el número de series tiende a infinito. Muchas de las ideas de series podrán trasladarse al análisis de imágenes, viendo las filas o las columnas de las imágenes como conjunto de series y unificando el tratamiento de los datos espaciales y las series temporales, que hasta ahora han evolucionado con cierta independencia.

Los métodos automáticos irán teniendo cada vez más peso por las necesidades del análisis. Hasta la introducción del criterio de Akaike (1973) los estadísticos han confiado en el trabajo artesanal de construcción de modelos como la mejor forma de extraer la información de los datos. Sin embargo, las necesidades de análisis de grandes masas de datos han hecho cada vez más populares los métodos automáticos y el éxito de los programas TRAMO y SEATS desarrollados por Gómez y Maravall (1996) para el análisis de series temporales y la desestacionalización es una muestra de la enorme demanda en todo el mundo por buenos métodos automáticos de análisis.

Un texto que presenta muchos de estos nuevos problemas y las soluciones que desde la ingeniería y las ciencias de la computación han aportado para resolverlas es el de Hastie, Tibshirani and Friedman (2011), aunque sus métodos están pensados para variables estáticas y no temporales. Estos autores presentan en un lenguaje estadístico unificado los métodos que llaman de aprendizaje supervisado (supervised learning) y aprendizaje no supervisado (unsupervised learning). En los primeros se desea predecir el valor de una variable, conocidos los valores de otras relacionadas con el objeto de la predicción. El caso más simple es regresión lineal, cuando la variable respuesta es continua, o discriminación, cuando esta es un atributo. Los autores presentan métodos de regresión no lineal, incluyendo métodos locales cuando la forma de relación cambia dentro del espacio estudiado, y redes neuronales. Para variables de atributo tenemos regresión logística, que puede aplicarse localmente, con funciones nucleo de suavización, y métodos como SVM (support vector machines) y métodos de árboles de clasificación. Además estos modelos pueden combinarse para la predicción, con Bayesian Model Averaging o mediante otros procedimientos descritos en el libro.

Los métodos de aprendizaje no supervisado pueden ser de clasificación, que corresponde a los de Cluster en la terminología estadística tradicional, y los de reducción de dimensión. Respecto a los primeros los autores presentan los métodos clásicos y algunos de los nuevos métodos propuestos basados en proyecciones sobre espacios de dimensión menor. Véase Peña, Prieto y Viladomat (2010) para un nuevo método que se compara con otros propuestos en los últimos años. Para la reducción de dimensión además de las clásicas componentes principales y análisis factorial los autores introducen los componentes independientes.

No existe, que yo sepa, ningún texto que aborde los problemas dinámicos cuando disponemos de grandes bancos de datos. En Peña and Poncela (2006) se revisan los métodos para reducir la dimensión en series temporales que es uno de los problemas centrales con Dynamic Big Data. Textos recientes que presentan ejemplos interesantes de los análisis con grandes masas de datos son O'Neil and Schutt (2013) and Provost and Fawcett (2013). Mayer-Schönberger and Cukier (2013) discuss the future effect of Big Data in our lives.

5. Las titulaciones de Estadística y los departamentos de Estadística

El fenómeno de los Big Data está estimulando la aparición de grados en Ciencias de los Datos o Ingeniería de Datos. En mi opinión su futuro, como ha ocurrido ya también con los Grados de Estadística, es incierto. Las titulaciones de Estadística han tenido una demanda moderada que ha ido además cayendo en los últimos años, y los grados que han sobrevivido se han orientado hacia una cierta especialización. No es previsible que esta tendencia cambie en el futuro, y una lección que podemos extraer de esta experiencia es que hubiera sido mejor concentrar los recursos y los esfuerzos en programas de Master en Estadística, donde estudiantes formados en Economía, Biología, Ingeniería u otras titulaciones básicas, profundizaran en las técnicas más útiles para su profesión. Existe un alto riesgo de que lo mismo ocurra con estas nuevas titulaciones. Considero que la mejor formación para resolver problemas reales de grandes datos es primero un Grado en Ciencias, Economía o Ingeniería y después especializarse con un Máster en Estadística, Big Data o Data Mining. La formación de Máster debería impartirse en colaboración con otros departamentos, especialmente de informática y TIC. Es importante que el estudiante conozca tanto los métodos de análisis como los procedimientos para manejar el gran volumen de datos y los metodos de computación necesarios para llevarlos a cabo. Además, un enfoque interdiciplinar puede aportar una visión mucho más amplia y realista de estos problemas.

Una situación semejante se presenta respecto a la investigación en este campo. Los métodos analíticos para Big Data deben desarrollarse en paralelo con métodos de computación eficaces para llevarlos a cabo y para tratar adecuada-

mente el flujo continuo de datos que aparece en muchas aplicaciones.

Por estas razones creo que es importante desarrollar lazos de colaboración estrechos entre los departamentos de Estadística y los departamentos de Informática, Matemáticas y Tecnologías de la Comunicación, así como otros departamentos universitarios interesados en el análisis de grandes masas de datos que aparecen en Medicina, Economía y Empresa, Sociología o Periodismo y Comunicación. La creación de institutos de investigación interdiciplinarios sobre estos temas puede ser una buena herramienta para asegurar estas colaboraciones.

6. Conclusión

El fenómeno Big Data es una oportunidad para resituar la Estadística en el centro de la adquisición de conocimiento. Los estadísticos debemos trabajar conjuntamente con los científicos que entienden los datos y con los técnicos que saben cómo transmitirlos y manipularlos eficazmente. Este papel central que pueden jugar los estadísticos debería generar una gran demanda de profesionales con esta formación. Ayudará a ésta tarea atraer a graduados brillantes y aportarles la formación estadística necesaria para avanzar en su campo de especialización mediante el análisis de datos. Necesitamos profesores con amplia experiencia en el análisis de datos masivos en diversos campos científicos para dar formación e impulsar investigación relevante y de impacto en las aplicaciones. El campo está abierto y las posibilidades son enormes, pero debemos recordar que si los estadísticos no estamos a la altura de esta tarea otros ocuparán nuestro lugar.

Referencias

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood method. *Proceeding of the Second Symposium on Information Theory,* N.B. Petrov and F. Caski, eds., Academiai Kiado, Budapest, 267-281.

[2] Box, J (1978). *Fisher: The life of a Scientist.* Wiley, N.Y.

[3] Cortes, C. and Vapnik, V. (1995). "Support-vector networks". *Machine Learning* 20 (3): 273.

[4] Efron, B. (1979). Computers and the Theory of Statistics: Thinking the Unthinkable, *SIAM Review* Vol. 21, No. 4, 463-480.

[5] Gómez, V. and Maravall, A. (1996). Programas TRAMO and SEATS. *Documento de Trabajo, Banco de España.* SGAPE-97001.

[6] Hastie, T., Tibshirani, R. and Friedman, J. (2011). *The Elements of Statistical Learning:Data Mining, Inference, and Prediction*, 2th edition. Springer Series in Statistics.

[7] Hilbert, M. and López, P. (2011). The world's technological capacity to store, communicate, and compute information, *Science*, February 10.

[8] Khun, T. (1962). *The structure of scientific revolutions*, University Chicago press.

[9] Lesk, M. (1977). How Much Information Is There In the World? http://www.lesk.com/mlesk/ksg97/ksg.html

[10] Mayer-Schönberger, V. and Cukier, K.(2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think,* John Murray (Publishers).

[11] Olkin. I. (1995). Meta-Analysis: Reconciling the results of independent studies. *Statistics in Medicine*, 14, 457-472.

[12] O'Neil, C. and Schutt, R. (2013) *Doing Data Science: Straight Talk from the Frontline,* O'Really Media Inc.

[13] Peña, D. (1995). El futuro de los métodos estadísticos, *Jornadas de Estadística Española*, libro por el 50 Aniversario de la Fundación del Instituto Nacional de Estadística, 93-108.

[14] Peña, D. and Poncela. P. (2006) Dimension Reduction in Time Series , in *Advances in Distribution Theory, Order Statistics and Inference*, Balakrishnan, N, Castillo, E. and Sarabia, J. M. (eds), Chapter 27, 437-461, Birkhauser: Boston.

[15] Peña D. and Prieto, J. (2001). Cluster identification using Projections. *Journal of American Statistical Association*, 96, 1433-1445.

[16] Peña D., Prieto, J. and Viladomat, J. (2010). Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure. *Journal of Multivariate Analysis* 101, 9, 1995 -2007.

[17] Provost, F. and Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking,* O'Really Media Inc.

[18] Stigler, S. M. (1986). *The History of Statistics*. Harvard University Press.

[19] Tong H. (1965). Threshold autorregression, limit cycles and cyclical data. *Journal of Royal Statistical Society B*, 42,245-292.

[20] Tukey J. W. (1965). The Technical Tools of Statistics. *The American Statistician*, 19, 2, 23-28.

[21] Yates, F. (1966). Computers, the second revolution in Statistics, Biometrics, 22,3, 233-251.

Acerca de los autores

Daniel Peña es Catedrático de Estadística y Rector de la Universidad Carlos III de Madrid. Titulado por la Politécnica de Madrid (UPM), Complutense, y Harvard University, ha sido Catedrático en la UPM, U. Wisconsin- Madison y U. de Chicago. Ha sido Presidente de la SEIO y de European Courses in Advanced Statistics y Vicepresidente de Interamerican Statistical Institute. Ha dirigido 26 tesis doctorales, publicado 14 libros y más de 200 artículos de investigación. Ha recibido el premio Youden Prize al mejor artículo publicado en Technometrics en 2005, es Ingeniero del año por el Colegio de Ingenieros Industriales de Madrid y Premio Jaime I de investigación en Economía. Es Miembro de honor (Fellow) IMS y ASA.