**REAL ACADEMIA DE CIENCIAS
EXACTAS, FÍSICAS Y NATURALES
DE ESPAÑA**

OBSERVACIÓN Y CÁLCULO EN ESTADÍSTICA CON DATOS MASIVOS

DISCURSO LEÍDO EN EL ACTO DE SU RECEPCIÓN

COMO ACADÉMICO DE NÚMERO POR EL

EXCMO. SR. D. DANIEL PEÑA SÁNCHEZ DE RIVERA

Y CONTESTACIÓN DEL

EXCMO. SR. D. FRANCISCO JAVIER GIRÓN GONZÁLEZ-TORRE

EL DÍA 4 DE MAYO DE 2022

MADRID
Domicilio de la Academia
Valverde, 22

Índice

1. Agradecimientos

Excmo. Sr. Presidente de la Academia

Excmas. Sras. y Excmos Sres Académicos

Autoridades, familiares, amigas y compañeros, Sras. y Sres.

Mis primeras palabras son de agradecimiento a todas las personas que han hecho posible mi entrada en esta Real Academia.

El académico D. Javier Girón ha sido un constante adalid de mi candidatura desde hace años, primero, como Académico Correspondiente y, después, como numerario. Estoy en deuda con él por su continuo apoyo. Acompañaron al Profesor Girón en su propuesta los académicos D. Enrique Castillo y D. David Ríos, a quienes agradezco mucho su generosa valoración de mis méritos. He tenido en otras muchas ocasiones el respaldo del Profesor Castillo, a quien reitero mi agradecimiento. El reconocido buen hacer del presidente de la Academia, el profesor Jesús Sanz-Serna, ha sido muy importante para llevar a buen término mi candidatura y cuenta con mi gratitud. Agradezco, también, a todos los Académicos de la sección de Ciencias Matemáticas que, con su voto, han hecho posible que el pleno de la Academia, con el suyo, me haya otorgado el privilegio, e inmenso honor, de poder estar hoy aquí. Muchas gracias, a todos, de todo corazón.

Los méritos que puedan haberse encontrado en mi trayectoria investigadora están compartidos con mis coautores, que me han enriquecido con sus ideas, su estímulo y su entusiasmo. Muchos trabajos provienen de tesis dirigidas, que han sido fuente de amistades y del incomparable placer de ver surgir la luz al final de una dura travesía compartida. Mis padres, ya fallecidos, nos inculcaron a todos sus hijos la valoración prioritaria del estudio y del conocimiento y, aunque no puedan disfrutar de este momento, están hoy muy presentes en mi recuerdo. Agradezco, también, todo el apoyo y el cariño que he recibido de mi primera mujer, Mercedes Conde, de mis hijos, Jorge y Alvaro, así como de mi esposa actual, Jette Bohsen, que me ha hecho amar más la vida desde que tuve la suerte de conocerla.

Tengo el honor de ocupar la medalla 32 de esta Academia, que ha tenido 5 muy ilustres predecesores. En el siglo XIX, los Excmos Sres D. Antonio Aguilar, que fué Catedrático de Astronomía de la Universidad Central y Director del Observatorio Astronómico de Madrid,

y D. Alberto Bosch, que, además de sus numerosas contribuciones matemáticas, tuvo una intensa actividad política como Ministro de Fomento, Alcalde de Madrid, Diputado y Senador Vitalicio. En el siglo XX, la llevaron los Excmos Sres, D. Leonardo Torres y Quevedo, uno de nuestros más geniales inventores, precursor de la cibernética. constructor de dirigibles, autómatas y transbordadores, utilizados en todo el mundo, y que fué Presidente de esta Academia; D. Francisco de Asís Navarro, científico versátil que fué químico, arquitecto y matemático y Catedrático de la Facultad de Ciencias de Madrid y, por último, D. Manuel Valdivia Ureña, Catedrático de la Universidad Politécnica de Valencia (UPV) y de la Universidad de Valencia (UV), donde se jubiló en la cátedra de Análisis Matemático. El Profesor Valdivia Ureña ha sido uno de nuestro más destacados investigadores, impulsor del análisis funcional y creador de una escuela de excelencia en nuestro país con gran proyección internacional. Además de estos cinco académicos, D. Francisco Prieto, también electo para esta medalla, no pudo disfrutarla por su prematuro fallecimiento antes de su ingreso.

Situarme el último de la fila formada por tan ilustres académicos es una gran honor que conlleva una gran responsabilidad. Intentaré seguir su ejemplo, y pondré todo mi esfuerzo e ilusión en colaborar con mis compañeros en la academia en difundir la importancia de las Ciencias Matemáticas para el avance y mejora de nuestra sociedad.

El título de mi discurso es *Observación y cálculo en estadística con datos masivos*. Observación y cálculo forman el lema de esta Real Academia y son el fundamento de todas las ciencias. El origen de este lema ha sido investigado en un brillante trabajo por el Académico Ildefonso Díaz, (Díaz, 2009, véase también Sanz-Serna, 2018). Este discurso pretende mostrar la importancia de la observación y del cálculo, en este caso de probabilidades, en el nacimiento y evolución de la estadística y en su adaptación a las nuevas observaciones, los datos masivos, o big data, y a los actuales instrumentos de cálculo informático, para ilustrar como la ciencia estadística esta contribuyendo a resolver los importantes problemas actuales de nuestra sociedad.

2. Introducción

Medir es comparar con un patrón, o clasificar en una escala, la magnitud de una observación. Permite calibrar la importancia de un fenómeno y transmitirlo con exactitud. Las mediciones suelen mostrar variabilidad y su estudio ha sido una fuente de conocimiento. El método científico ha ido evolucionando con nuestra capacidad de medir y analizar datos, y este proceso se ha ido extendiendo a todas las disciplinas actuales. La importancia de la medición explica que uno de los edificios emblemáticos de ciencias sociales de la Universidad de Chicago lleve esculpido el famoso aforismo de Lord Kelvin (1824-1907): When you cannot measure, your knowledge is meager and unsatisfactory".

La medición ha sido generalmente un proceso difícil y costoso. Por ejemplo, en el siglo XVIII la Academia de Ciencias de París organizó dos expediciones para comparar la longitud de un grado del meridiano en el ecuador y el polo norte, y decidir si la tierra era una esfera, o, como propuso Newton, está achatada por los polos. Esta medición supuso un trabajo de tres años. Jorge Juan (1713-1773), un joven oficial integrado en el equipo del ecuador como representante de España, aprendió por experiencia propia la importancia de medir con precisión. A su vuelta, durante el reinado de Carlos III, impulsó la creación de un Observatorio Astronómico y de una Academia de Ciencias y Gabinete de Historia Natural, en los edificios del actual Museo del Prado. Su trabajo a favor de la ciencia y la investigación sembró el germen de esta Real Academia de Ciencias Exactas, Físicas y Naturales de España, nacida un siglo más tarde.

Hoy, por primera vez en nuestra historia, somos capaces de medir automáticamente, con bajo coste y en tiempo real, fenómenos meteorológicos, económicos y sociales cuyo conocimiento requería ingentes cantidades de recursos en el siglo pasado. En la actualidad, muchas de las actividades que realizamos, tanto de ocio como de trabajo, en procesos ambientales, productivos, económicos o sociales, están monotorizadas por sensores. Los aparatos digitales que los contienen, como los teléfonos móviles, recogen datos de forma continua con un coste marginal cada vez más bajo, y crean gigantescas bases de datos, conocidas con el nombre de big data. Estos datos masivos digitales contienen muchas variables, cualitativas y numéricas, pero, también, imágenes, vídeos o audios, con frecuencia asociados a distintos momentos y localizaciones.

Esta abundancia de datos expandirá nuestro conocimiento y un problema crucial es convertirlos en información relevante, para avanzar hacia sociedades más equilibradas y justas. Los nuevos retos están transformando la estadística, y propiciando su convergencia con áreas afines, como la inteligencia artificial y el aprendizaje automático.

En las secciones siguientes abordaremos cómo la observación y el cálculo han ido transformando la estadística. Influyen en su nacimiento, a mediados del siglo XVII, y van dejando su rastro en los cinco periodos en que he dividido su evolución. En el primero, los avances estadísticos fueron impulsados por la astronomía y la física. En el segundo, por la biología. Después por la industria y el comercio, luego por la medicina y, actualmente, por la informática y las tecnologías de la información y la comunicación y la irrupción de los datos masivos. En la sección siguiente, analizamos cinco ideas que han impulsado la transformación de la estadística para aprovechar los datos masivos y las ilustramos con las figuras de sus representantes más destacados. A continuación, analizamos, con cinco ejemplos, cómo los avances en un campo científico se han transmitido a otros a través de los métodos estadísticos, que han facilitado la fertilización cruzada entre disciplinas. Por último, se incluyen algunas conclusiones finales.

3. Los datos como impulsores de la estadística

Numerosos trabajos han estudiado la historia de la estadística y la probabilidad, desde distintos puntos de vista, véase, por ejemplo, David (1962), Gómez-Villegas y Mora (2018), Girón (1994), Fienberg (1992), Hacking (1975, 1990), Hald (1998, 2003), Stigler(1986, 1990, 1996) y Todhunter (1865). También, es muy relevante la colección de artículos que han abierto nuevos caminos, llevada a cabo en tres volúmenes por Kotz y Johnson (1992a, 1992b, 1998). En esta sección vamos a dividir su evolución en cinco periodos, poniendo el énfasis en la relación entre el avance de los métodos estadísticos y los datos disponibles. Este aspecto ha sido estudiado por varios autores, véase Box (1976, 1984) y Stigler (1986).

Desde la antigüedad, los estados han recogido información sobre sus súbditos con motivos fiscales y militares. Una historia de los censos puede verse en Rossi et al. (1983). No obstante, hasta el siglo XVII los datos no se analizan en busca de regularidades, buscando su comprensión o predicción. En el renacimiento comienzan a ser

estudiados acontecimientos futuros tradicionalmente asociados a la voluntad de alguna divinidad, como los juegos con dados o cartas, o los fenómenos meteorológicos, y surge la probabilidad, como medida de la incertidumbre.

Sin embargo, tenemos que esperar hasta mediados del siglo XVII para que se conciba la idea de esperanza matemática y la probabilidad se aplique para resolver un problema de juegos de azar en Francia: la famosa correspondencia entre Pascal (1623 1662) y Fermat (1601 1665) en 1654. Su resolución del problema de repartir las ganancias de un juego de apuestas que se interrumpe antes de finalizarlo, es el inicio del cálculo de probabilidades como disciplina matemática. Pocos años después, Graunt (1662) analiza por primera vez datos demográficos de nacimientos y defunciones en Inglaterra para obtener predicciones demográficas sobre la población de Londres, construyendo tablas de mortalidad. (Véase Glass, 1964, para un buen análisis de sus contribuciones). Podemos concluir que la estadística nace en la segunda mitad del siglo XVII, al disponer de una comprensión del cálculo de probabilidades combinada con las primeras herramientas para estudiar las regularidades observadas en los datos.

3.1. Primer periodo: astronomía y mínimos cuadrados, (1663–1809)

El primer periodo de evolución de la estadística se extiende un siglo y medio, desde mediados del siglo XVII hasta inicios del siglo XIX, con la aparición del método de los mínimos cuadrados.

Veinticinco años después del trabajo pionero de Graunt (1662), se publica en 1687 los *Principia* de Isaac Newton (1642-1727). En este monumental trabajo, Newton propone la primera explicación global de nuestro mundo físico. Su teoría tiene una gran influencia y estimula a los astrónomos a realizar mediciones del movimiento de los cuerpos celestes para contrastar sus predicciones. Al mismo tiempo, se produce un cambio social en la valoración de la experiencia empírica y la importancia de los datos como fuente de conocimiento, gracias al trabajo de los filósofos empiristas británicos, Locke (1630-1704) y Hume (1711-1776), entre otros.

Los instrumentos de medición en esta época eran poco precisos y los errores de medida importantes, lo que planteaba el problema de cómo combinar distintas mediciones de un mismo fenómeno. La solución la proporcionaron Legendre (1752- 1833) y Gauss (1777-

1855), con el descubrimiento del método de mínimos cuadrados, que permite estimar los parámetros para ajustar a los datos una ecuación, minimizando los errores cuadráticos del ajuste (Plackett, 1972). Gauss demostró, además, su optimalidad cuando los errores siguen una distribución normal, que había sido introducida como modelo de la variabilidad de las medidas por Moivre en 1756 (Stigler, 1986).

La teoría de probabilidades avanza mucho en este periodo con los trabajo de J. Bernoulli, C. Huygens, A. De Moivre, J. Arburthnot y T. Bayes (veáse David, 1962, Stigler, 1986 y Girón, 1994, para sus aportaciones).

En resumen, en este periodo, que abarca más de siglo y medio, se mejoran y desarrollan los métodos de medición, especialmente en astronomía, se crean los primeros modelos de distribución de probabilidad para variables continuas, como los errores de medida, y discretas, como los accidentes, y se descubren las ventajas de combinar distintas mediciones de una misma magnitud, por el teorema central del límite. Además, se proponen distintos métodos para calcular ajustes lineales a los datos y se descubre el método de los mínimos cuadrados.

3.2. Segundo período: biología y correlación (1809–1909)

El segundo periodo de evolución de la estadística se extiende exactamente un siglo. Se inicia en 1810, con la disponibilidad de la herramienta de mínimos cuadrados, y finaliza en 1909 con la muerte de Galton y la creación del primer centro de estadística, el Laboratorio Galton en University College en Londres. En este periodo la estadística se desarrolla en las ciencias sociales y se introduce en la administración estatal, se crean los primeros Institutos de Estadística oficial, y se inician sus aplicaciones a la medicina, gracias al trabajo de Florence Nightingale (1820-1910), creadora de la profesión de enfermería, que demostró con datos la importancia de la higiene para prevenir la mortalidad durante la guerra de Crimea (véase Cohen, 1984).

La importancia de los avances ocurridos en astronomía por el análisis de datos lleva a Quetelet (1796–1874), un astrónomo belga, a intentar identificar las leyes de los fenómenos sociales, con la esperanza de construir una teoría general, similar a las leyes del mundo físico de Newton. Hoy sabemos que las leyes sociales son probabilis-

tas y no deterministas: podemos prever el crecimiento económico por término medio, pero no podamos calcular con exactitud su magnitud, aunque eliminemos los errores de medida. A pesar de las limitaciones del trabajo de Quetelet, sus investigaciones estimularon el uso de la estadística en las ciencias sociales y en la administración.

A mediados del siglo XIX los estados europeos comienzan a crear instituciones estables para recoger información sistemática sobre sus territorios. En España, en 1856, se crea la Comisión de Estadística General del Reino, el primer órgano estadístico permanente en nuestro pais (véase Merediz, 2004 para la estadística oficial en España antes de esa fecha). Por otro lado, la necesidad de unificar, para favorecer el comercio, las unidades de medición, impulsó la creación del sistema métrico decimal en 1875, cuando 17 países firmaron un tratado internacional estableciendo un sistema común de pesos y medidas.

Un importante acontecimiento en este periodo es la publicación por Charles Darwin (1809–1882), del *Origen de las Especies*, en 1859. La teoría de la evolución es la primera teoría científica con fundamentos estadísticos, ya que establece que ocurrirá, en promedio, en una población, y no a cada uno de sus elementos. Su base no es la certeza. sino la probabilidad. Galton (1822–1911), un excelente científico, primo de Darwin, comienza a analizar datos biológicos para contrastar esta teoría. Descubre la regresión a la media de los individuos de una población, y dedica todo su esfuerzo, y su fortuna, a desarrollar la estadística, que consideraba fundamental para explicar la evolución de la vida en la tierra. Fundó y financió en 1904 la primera revista de estadística, *Biometrika*, y dejó su herencia a la Universidad de Londres para crear en 1909, el primer centro de estadística, el Galton Laboratory. Su dirección fue encomendada a Karl Pearson (1857–1936), que fue también el primer editor de la revista. Pearson, que había colaborado estrechamente con Galton en sus análisis, había ya introducido el coeficiente de correlación y el primer contraste de ajuste, que lleva su nombre, de los datos a una distribución de probabilidad.

3.3. Tercer período: agricultura, industria e inferencia (1909–1962)

Este tercer periodo se inicia con la creación del Galton Laboratory, en el centenario del nacimiento de Darwin, y finaliza con la aparición del ordenador como herramienta de cálculo en estadística. Está dominado por las figuras de R.A. Fisher (1890-1962), por un

lado, y J. Neyman (1894-1981) y E.S. Pearson (1895 –1980), por el otro. En este periodo se inician las aplicaciones a la agricultura, con el diseño de experimentos, y a los procesos industriales, con el control de calidad, que se convierte en un procedimiento habitual en la fabricación en serie.

Fisher (1925a, 1925b), un licenciado en matemáticas por Cambridge que había rechazado la posición universitaria propuesta por K. Pearson para trabajar como estadístico aplicado en la estación agrícola de Rothamstead, al norte de Londres, introduce el procedimiento de estimación estadística por máxima verosimilitud y desarrolla en su libro, *Statistics for Research Workers*, una metodología rigurosa, basada en su experiencia práctica, para construir modelos estadísticos a partir de los datos: se dispone de una muestra aleatoria de una población, que se supone habitualmente normal, y se desea hacer inferencias sobre sus parámetros. Para ello, se estiman los parámetros del modelo probabilístico escogido y se comprueba el ajuste del modelo seleccionado mediante contrastes de significación. Fisher, desarrolló después los conceptos básicos del diseño experimental (Fisher, 1935) que siguen vigentes hoy (véase Box, 1987, para una buena biografía de Fisher).

Cuando K. Pearson se retira de University College en 1933 su cátedra se divide en dos, ocupadas por su hijo Egon, y Fisher. E.S. Pearson desarrolla con J. Neyman la teoría de contrastes de hipótesis, que tiene muchas aplicaciones industriales, especialmente en el control de calidad. En este periodo se expande la teoría de inferencia estadística, con la creación de los métodos multivariantes, la teoría de los procesos estocásticos, y la de decisión bajo incertidumbre. Con el trabajo de la escuela probabilística rusa, encabezada por Kolmogorov, la estadística se consolida como una rama de las matemáticas y comienza a ser estudiada en los centros universitarios del mundo desarrollado. Las limitaciones de cálculo hacen que los avances se produzcan más en la teoría que en las aplicaciones, pero la estadística se convierte en la tecnología básica del método científico.

La emergente disponibilidad de datos de varias variables estimula el nacimiento del análisis multivariante: Fisher introduce el análisis discriminante (Anderson, 1996), H. Hotelling (1895 - 1973) las correlaciones canónicas y las componentes principales, que habían sido descubiertas por K. Pearson es sus estudios de regresión ortogonal, así como el estadístico para contrastes multivariantes que lleva su

nombre. C. R. Rao (1920-) contribuye a la teoría de la estimación y a los contrastes de hipótesis para variables vectoriales. Un importante resultado en este periodo es el descubrimiento, debido a James y Stein (1961), de la pérdida de optimalidad al aumentar al dimensión de la media muestral para estimar la media poblacional con variables normales multivariantes. Este resultado abrirá la puerta a los estimadores contraídos (shrinkage estimates) y a la regularización. La teoría de procesos estocásticos y series temporales fue desarrollada por M. Barlett (1910 –2002) y H. Cramer (1893 –1985). Finalmente A. Wald (1902-1950) estudia la construcción de reglas de decisión y la teoría de decisiones secuenciales (Wald, 1950).

Hemos seleccionado el año 1962 para finalizar este periodo porque en ese año fallece Fisher y comienza la transformación de la estadística por los ordenadores, una revolución anunciada por Yates (1966). En los años 60 los ordenadores comienzan a introducirse en las universidades y la administración y en 1965 se crea en EE.UU.

el primer centro electrónico de datos que analiza información fiscal y huellas digitales.

3.4. Cuarto período: servicios, computación y modelos complejos y dinámicos (1962–2001)

La aparición del ordenador inicia un periodo de rápido desarrollo de la estadística durante la segunda mitad del siglo XX. Las nuevas facilidades de cálculo electrónico permiten flexibilizar y generalizar las hipótesis para construir y estimar modelos estadísticos, y sus aplicaciones se extienden a la medicina y ciencias de la salud y los servicios. Véase Tanur et al. (1989).

Los modelos de superviencia con variables explicativas para datos biológicos fueron introducidos por Cox (1972) y Nelder and Wedderburn (1972) propusieron la clase de los modelos lineales generalizados. La hipótesis de que todos los datos vienen de la misma población fue modificada por Tukey (1960), Huber (1964) y Box and Tiao (1968) para tener en cuenta en la estimación la aparición de datos atípicos o *outliers*. Heterogeneidad supone, también, la posible presencia de grupos en los datos o clusters, y se proponen nuevos métodos de construcción de clusters basados en mezclas de modelos (Banfield y Raftery, 1993). Tukey (1962, 1977) resalta el componente exploratorio de la estadística, por encima del aspecto inferencial confirmatorio. Se extienden los modelos dinámicos, impulsados por nuevos métodos

para modelizar series temporales y generar predicciones propuestos por Box and Jenkins (1970). Los modelos paramétricos, que habían dominado el periodo anterior, compiten con los enfoques no paramétricos, (véase por ejemplo Härdle, 1990 y Wahba, 1990) y aparecen nuevos métodos de suavizado para relaciones no lineales. Los datos temporales climáticos estimulan la introducción del análisis de datos funcionales para variables dependientes en el tiempo (Ramsay and Silverman, 1997, Ferraty and Vieu, 2006) y el estudio de datos espaciales (Cressie, 1991).

En este periodo surgen nuevas ideas que serán fundamentales para el análisis de datos masivos y que se describen con detalle en la sección 4. Aparece el primer criterio automático de selección de modelos, debidos a Akaike (1973), Efron (1979) propone el bootstrap para calcular la precisión de un estimador en situaciones complejas y la inferencia bayesiana recibe un gran impulso con la aparición de métodos de simulación, o métodos de Monte Carlo, basado en cadenas de Markov o métodos MCMC (Markov Chain Montecarlo Methods), (Gelfand y Smith, 1990). Las dificultades de trabajar con un número relativamente alto de variables correladas entre sí lleva a estudiar estimadores no centrados, pero con menor varianza, y Hoerl y Kennard (1970) proponen estimadores cresta, o *Ridge regression*, y Tibshirani (1996) la estimación lasso.

3.5. Quinto periodo: Big Data, automatización y ciencia de datos, (2001–)

El quinto período de desarrollo de la estadística se inicia en este siglo, con la aparición del big data, y se caracteriza por una convergencia con otros campos en la creación de la disciplina de Ciencia de los Datos. En ella, la estadística tiene un papel central, pero también se incorporan modelos y procedimientos desarrollados en otras disciplinas, como Inteligencia Artificial, Aprendizaje Automático (*Machine Learning*) y Matemática Aplicada e Investigación Operativa. El nombre de *big data* aparece en 1997, utilizado por investigadores de la NASA para ilustrar que el gran aumento del volumen de los datos llevaría a una situación crítica a los sistemas informáticos existentes. En 2001 se propone la caracterización de los bancos de datos masivos con las tres V (Velocidad, Volumen y Variedad).

Con la aparición en 2001 de las redes sociales, Wikipedia y los blogs, y la proliferación de los sensores en los teléfonos móviles y en

otros aparatos de medición, se inicia un crecimiento exponencial de los datos disponibles; véase por ejemplo el informe COTEC, 2017. Muchos trabajos han analizado los cambios en la metodología estadística como consecuencia del big data; véase como ejemplo Bühlmann, P. and van de Geer, S. (2011, 2018), Fan et al. (2014), Efron and Hastie (2016), Donoho (2017), Torrecilla y Romo (2018) y Galeano and Peña (2019). Podemos concluir que en este siglo se produce un cambio de paradigma en el análisis de datos, cuyos orígenes y características desarrollaremos en la sección siguiente.

4. Cinco ideas importantes para la ciencia de datos y sus precursores

La aparición en el siglo XXI de los datos masivos asociados a nuevos problemas produjo inicialmente lo que podríamos resumir como un espejismo del tamaño: si la dimensión de los datos aumenta, con más casos y variables, podremos aplicar los métodos estadísticos ya disponibles, aunque haya que mejorar la velocidad de los procesos de cálculo. Sin embargo, sabemos desde James and Stein (1961), que la dimensión afecta a la optimalidad de un estimador. Esta misma propiedad se ha demostrado para las esperanzas condicionales cuando estimamos una regresión con contracción del estimador de mínimos cuadrados. También aparece el problema de la dimensión en los criterios de comparación de modelos para la predicción: al aumentar la dimensión, un modelo escueto puede ser más eficiente para hacer predicciones fuera de la muestra que otro con más parámetros y mejor ajuste dentro de la muestra, como demostró Akaike (1973). Por tanto, existe desde los años 70 amplia evidencia de que la dimensión de los datos condiciona tanto el estimador que debemos elegir dentro de un modelo, como el modelo que debemos elegir para hacer predicciones en un problema dado.

Por otro lado, la ciencia ofrece muchos ejemplos de que extrapolar loa resultados en una escala a otra más grande, o más pequeña, puede no funcionar. Por ejemplo, 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, al descender a la escala microscópica aparecen las propiedades cuánticas y principios como el de superposición, dónde una partícula puede estar simultáneamente en

dos posiciones distintas con cierta probabilidad. Como ejemplo más cercano a nuestra experiencia, un medicamento tomado en pequeñas dosis nos ayuda a dormir y en dosis elevadas puede producirnos la muerte.

En conclusión, necesitamos nuevos métodos para los nuevos datos masivos. Estos métodos han tenido su germen en ideas desarrolladas en estadística, pero también en inteligencia artificial y en aprendizaje automático. En las secciones siguientes presentamos cinco ideas importantes, surgidas en la segunda mitad del siglo XX en estadística, fundamentales para hacer predicciones con datos masivos. Por razones de espacio hemos seleccionado solamente cinco líneas, aunque otras, como la predicción de la no linealidad con métodos no paramétricos o los métodos para datos funcionales, son también muy relevantes. Véase Gelman and Vehtari (2021) para otra perspectiva de ideas destacadas en estadística en los últimos cincuenta años. Hemos elegido para representar cada una de estas líneas a investigadores reconocidos por sus trabajos pioneros en ellas. Muchas otras personas han realizado contribuciones muy relevantes, pero no pretendemos en este trabajo hacer una revisión exhaustiva de cada línea, sino ilustrar su importancia para el estudio de datos masivos. Estas líneas, que analizaremos en las secciones siguientes son, en orden cronológico del nacimiento de los representantes elegidos:

1. Heterogeneidad en los datos.

   Los datos pueden haber sido generados por distintas distribuciones y no por una sola. Esta idea llevó a Box a diseños experimentales iterativos, métodos robustos de detección de atípicos y modelos integrados de series temporales. El desarrollo de esta idea ha llevado a métodos robustos de estimación y a la estimación de clusters con mezclas de modelos, entre otras extensiones.

2. Estimar muchas variables con penalización.

   Stein demostró que en alta dimensión los estimadores contraídos son mejores que los tradicionales. Estos estimadores añaden una penalización adicional y han llevado a los métodos de regularización para datos masivos.

3. Reglas de predicción automáticas.

Los análisis de datos masivos requiere métodos automáticos. El primer procedimiento de estas características fue descubierto por Akaike, que inició el campo de los criterios automáticos de selección de modelos.

4. Combinar distintas reglas de predicción para aumentar la precisión.

   Breiman demostró que generando muchos conjuntos de datos con pequeñas perturbaciones de los originales, ajustando un modelo a cada conjunto y combinando las predicciones de todos ellos, reducimos el error de predicción. Esta práctica de agregar modelos, o *ensamble models* se ha convertido hoy en habitual.

5. El ordenador para generar nuevos procedimientos.

   Efron propuso sustituir las aproximaciones asintóticas por el cálculo informático con el *bootstrap* o estimador autosuficiente. La aplicación de esta idea en la estimación bayesiana ha llevado a los procedimientos de estimación por simulación de cadenas de Markov o estimación con *Markov Chain Monte Carlo (MCMC)*.

4.1. Box y la heterogeneidad en los datos

El paradigma clásico de la estadística es que disponemos de una muestra aleatoria de una población y deseamos construir un modelo para explicar los valores observados. En los años 60 este paradigma entra en crisis, ya que los datos reales son, con frecuencia, heterogéneos. George Box y John Tukey han sido pioneros en el estudio de la heterogeneidad por datos atípicos y en proponer procedimientos robustos. Box, además, estudió la heterogeneidad en procesos continuos, introduciendo la idea de diseños evolutivos, y en series temporales, con los procesos integrados. Estas ideas facilitan el análisis con datos masivos.

George E. P. Box (1919-2013) nació en Gravesend, cerca de Londres, siendo el hijo menor de una familia con pocos recursos. Compaginó el trabajo en una planta química de depuración de agua con un grado en Química en la universidad de Londres. Al comenzar la segunda guerra mundial es llamado a filas y destinado a un centro de defensa química en el sur de Inglaterra. Su trabajo era estudiar el efecto de gases tóxicos sobre animales y para ello descubre pronto la necesidad de utilizar métodos estadísticos. Encuentra el libro

Figura 1: Professor George E. P. Box, pionero en modelos para datos heterogéneos

de Fisher (1925) y consigue que el ejército solicite su ayuda para la experimentación del centro. Acabada la guerra, aprovecha las becas convocadas por el gobierno para facilitar la integración de los exsoldados británicos para estudiar estadística en el University College, donde se graduó en 1947.

En 1948 Box consiguió un trabajo como estadístico en ICI, una importante empresa química británica, y allí desarrollo una metodología para mejorar de forma continua un proceso mediante la experimentación. Adaptó las ideas de Fisher, desarrolladas para experimentos agrícolas en situaciones estables, a procesos industriales continuos, que cambian en el tiempo. Este trabajo le otorgó un doctorado por la universidad de Londres en 1952, una invitación para pasar un año como profesor invitado en la University of North Carolina, y, después, un contrato con Princeton University como Director of the Statistical Research Group. En Princeton conoce a G. Jenkins y se indicia su colaboración que culmina con la creación de un enfoque unificado para analizar y hacer predicciones de series temporales (Box and Jenkins, 1970). En 1959 Box se traslada a Madison, en Wisconsin, para crear y dirigir el departamento de estadística en la universidad de Wisconsin-Madison. Continuó activo hasta el final de su vida en 2013. Las entrevistas de DeGroot (1987) y Peña (2001) y sus memorias, Box (2013), describen su vida y sus contribuciones

Box fue un pionero en la incorporación de la heterogeneidad en la estadística. En primer lugar, por propia experiencia en ICI, comprobó que habitualmente los procesos productivos evolucionan en el tiempo y la experimentación para mejorarlos requiere tener en cuenta esta di-

námica. La metodología que desarrolló con Wilson, (Box and Wilson, 1951), conduce al proceso a las condiciones óptimas de funcionamiento, y es considerada una aportación fundamental en estadística (Kotz y Johnson, 1992b). En lugar de los métodos de experimentación propuestos por Fisher para la agricultura, diseñó la experimentación continua, *Evolutionary Operation, EVOP*, (Box y Draper 1969), donde cada experimento proporciona información para planear el siguiente. Esta idea, que ha sido muy utilizada en todos los procesos continuos, se aplica actualmente en diseños automáticos de aprendizaje dinámico con datos masivos. Por ejemplo, para maximizar el tiempo que una persona permanece conectada en una red social. Para ello, se envían estímulos al teléfono del usuario, se mide su respuesta y, en función de ella, se ejecuta un nuevo experimento con nuevas variables, para descubrir las que aumentan la probabilidad de continuar activo en la red. También se han utilizado estas ideas para encontrar grupos de usuarios similares en redes sociales (Cui et al., 2018).

En segundo lugar, Box descubrió enseguida la importancia de las desviaciones de los datos de la normalidad e introdujo en estadística el término robustez, (Box, 1953), al analizar el efecto de desviarse de esta hipótesis en los contrastes de igualdad de varianzas. También estudió el efecto de los datos atípicos, y su detección y análisis llevó a nuevas patentes en su trabajo industrial en ICI (Davies et al., 1947). Los primeros modelos para el tratamiento de atípicos son debidos a Tukey (1960) y Huber (1964), que propusieron estimadores para distribuciones normales contaminadas con otra de colas pesadas. El enfoque de Box y Tiao (1968) considera, como Tukey, que los datos provienen de una mezcla de dos distribuciones, una con colas pesadas, pero en lugar de modificar el método de estimación para eliminar, o dar menos peso, a las observaciones extremas, aplican la estimación bayesiana a este modelo para, automáticamente, reducir el peso de los atípicos. El enfoque de Tukey y Huber lleva a cambiar los métodos de estimación con criterios minimax, como los M-estimadores de Huber, o eliminado los datos de las colas, como en el *trimming* de Tukey. El enfoque de Box y Tiao elige un modelo estadístico capaz de generar heterogeneidad y aplica a los datos la estimación del modelo elegido. El inconveniente del primer enfoque es que hay que adaptarlo a cada situación: por ejemplo, los M-estimadores introducidos por Huber(1964) no son robustos en regresión cuando tenemos observaciones muy influyentes, en el sentido de Cook (1977) y se han debido desa-

rrollar nuevos métodos específicos para esta situación (véase Maronna et al., 2019). Sin embargo, el método de Box y Tiao puede aplicarse con los mismos principios, pero, en contrapartida, requiere especificar un modelo que puede ser complejo y, en general, es desconocido.

En tercer lugar, Box y Jenkins (1970) introdujeron en el área de series temporales los modelos integrados, que son modelos no estacionarios pero con derivadas estacionarias, y Box and Tiao (1975) el análisis de intervención, que abrió la puerta al estudio de cambios estructurales en series temporales. El tratamiento hasta ese momento de las series temporales en estadística suponía su estacionaridad, de manera que su variación se producía respecto a un valor medio fijo con una varianza constantes en el tiempo. La genial idea de estos autores es proponer procesos integrados, donde alguna de sus derivadas son estacionarias, y varían respecto a valores fijos. Llegaron a esta idea analizando métodos de predicción utilizados con éxito en la práctica, como el alisado exponencial, y descubriendo su deducción como caso particular de un procedimiento general (véase Box, 1984). Esta idea ha transformado el análisis de series temporales en todas las áreas y muy especialmente en la economía. como veremos en la sección 5.2. Por otro lado, el análisis de intervención permite tratar cambios de nivel generados por factores exógenos en series temporales.

Estos avances han dado lugar a muchas extensiones. El modelo compuesto por dos distribuciones, para generar atípicos, se generaliza a una mezcla de varias distribuciones, que es el problema de cluster. En este caso, se trata de estimar los grupos y sus modelos y clasificar en ellos los datos. Aunque la estimación de una mezcla de modelos tiene una larga historia, que se origina con K. Pearson (véase McLachlan et al., 2019), tenemos que esperar al algoritmo EM de Dempster et al. (1977) para disponer de un procedimiento de estimar los parámetros de la mezcla por máxima verosimilitud. Su aplicación a problemas de cluster es debida a Banfield and Raftery (1993). Casella et al. (2014) han resaltado la importancia de la distribución a priori en el enfoque bayesiano en estos problemas. En series temporales cluster ha sido estudiado, entre otros, por Caiado et al. (2006, 2009), Alonso and Peña (2029) y Alonso et al. (2020).

El estudio de los datos atípicos en regresión llevó a Cook (1977) a definir las observaciones influyentes como aquéllas con un efecto especial en la estimación y que, sin embargo, no se detectan como atípicas. Esto llevó a estudiar el problema de enmascaramiento creado

por grupos de atípicos, (Peña and Yohai, 1995, 1999) y Peña (2005). Atípicos en poblaciones multivariantes han sido estudiados por Maronna (1976), Maronna et al. (2019), Rousseeuw and Van Zomeren (1990) y Peña and Prieto, (2001b, 2007). El estudio de atípicos está muy ligado a los valores extremos, de gran importancia en ingeniería (Castillo, 2012) y otras áreas.

El estudio de atípicos en series temporales univariantes fue iniciado por Fox (1972), y ampliado por Chang et al.(1988), Sánchez y Peña (2003), Justel et al. (2001) y Muler et al. (2009), (Véase Peña et al., 2011 para una comparación de métodos). Las observaciones influyentes en series temporales fueron introducidas por Peña (1990, 1991) y los atípicos fueron definidos para series temporales multivariantes por Tsay et al. (2000). Un método general de detección con series multivariantes ha sido propuesto por Galeano et al. (2006).

4.2. Stein y los estimadores contraídos

Charles M. Stein (1920-2016), nació en Brooklyn, New York. Inició sus estudios de matemáticas en la Universidad de Chicago con 16 años, pero tuvo que interrumpirlos para servir en la fuerza aérea, durante la segunda guerra mundial, haciendo predicciones meteorológicas. Después de la guerra se doctoró en año y medio en el departamento de estadística de la Universidad de Columbia. Inició su carrera académica en 1947 en la Universidad de California en Berkeley, pero fue expulsado tres años más tarde, por no firmar el juramento de lealtad impuesto en la era McCarthy. En 1953 entró en el departamento de Estadística de Stanford University, donde permaneció toda su carrera. Fue activista político contra la guerra de Vietnam y apoyó siempre causas liberales de forma pública.

Stein ha hecho muchas contribuciones fundamentales a la probabilidad, véase DeGroot(1986), Diacones and Holmes (2017) y Efron (2017). Sus demostraciones sobre la convergencia de sumas de variables aleatorias dependientes han influido mucho en la teoría moderna de la probabilidad. Nunca tuvo interés en las aplicaciones y se sintió siempre atraído por la generalidad y la elegancia de las demostraciones matemáticas rigurosas. Sin embargo, sus resultados sobre los estimadores admisibles en alta dimensión con el criterio de error cuadrático medio han revolucionado los métodos de estimación para datos masivos, y han tenido mucha importancia en la práctica. Su trabajo es un ejemplo notorio de la famosa frase, atribuida al psi-

cólogo Kurt Lewin y al físico James C. Maxwell: "no hay nada más práctico que una buena teoría".

Con su estudiante W. James, (James and Stein, 1961) Stein demostró que, en dimensión igual o mayor que tres, el estimador habitual de la media de la población, la media muestra, es inadmisible, es decir, siempre existe un estimador mejor, que es una contracción de la media muestral. Efron (2017) lo ha expresado de forma muy clara: si queremos prever la mortalidad en varios hospitales independientes y con distintos tasas de mortalidad, la mejor predicción para cada hospital combina su media con la de todos ellos. Este resultado, llamado con frecuencia la paradoja de Stein, ha provocado discusiones sobre su interpretación durante 25 años y ha dado lugar a los métodos de regularización para modelos dispersos en muchas dimensiones (*sparse models*). Es interesante que Stein llegó a este resultado tratando de demostrar su opuesto, la optimalidad de la media muestral en cualquier dimensión, y que su trabajo estimuló el uso de los estimadores bayesianos, a pesar de la crítica feroz de Stein a estos métodos. Sus resultados llevaron a Lindley and Smith (1972) a introducir los métodos jerárquicos en la inferencia bayesiana, que han potenciado sus aplicaciones.

Los estimadores contraídos (*shrinkage estimators*) en regresión fueron desarrollados por Hoerl and Kennard (1970) con el nombre de *Ridge regression* or regresión cresta. Estos autores demostraron que una contracción del estimador habitual de mínimos cuadrados en regresión conduce a predicciones más precisas, especialmente cuando las variables dependientes tienen alta correlación, es decir, existe multicolinealidad. Este estimador puede obtenerse imponiendo una penalización cuadrática sobre la suma de los cuadrados de los coeficientes de regresión. Su trabajo estimuló la investigación de otras formas de penalizar el número de parámetros. Un enfoque muy fructífero es debido a Tibshirani (1996), penalizando con la norma L1, es decir la suma de los valores absolutos de los parámetros estimados. Este es el estimador lasso (*least absolute shrinkage and selection operator*).

Una ventaja del estimador lasso con respecto al estimador cresta, o *Ridge*, es que fuerza que los coeficientes con valores pequeños se hagan cero, realizando por tanto una selección de variables que, por ejemplo en regresión, es más eficaz que la obtenida con los métodos tradicionales de regresión paso a paso. Establecer una penalización en

Figura 2: Professor Charles Stein, descubridor de los estimadores contraídos en alta dimensión

la estimación se denomina regularización y tiene muchas aplicaciones en la estimación de los efectos de muchas variables cuando esperamos que la mayoría sean inertes, o sin efectos relevantes. Además, para obtenerlo, tenemos que resolver un problema de optimización convexo, donde es fácil encontrar la solución (Hastie et al, 2015). El grado de penalización que debemos aplicar depende de parámetros que se estiman habitualmente por validación cruzada. El estimador lasso ha tenido muchas generalizaciones, como *elastic net, group lasso, y fused lasso* (Hastie et al., 2015), que han sido aplicados a la estimación de grandes matrices de covarianzas, Bickel and Levina (2008), componentes principales en alta dimensión, (Candès et al., 2011), o correlación canónica, (Witten et al., 2009).

Los estimadores contraídos han sido muy útiles también con series temporales. García-Ferrer et al. (1987) mostraron su utilidad para la predicción de datos macroeconómicos y Peña and Poncela (2004) que los estimadores contraídos surgen de manera natural al hacer predicciones con un modelo factorial dinámico. Las propiedades de la estimación con lasso en series temporales han sido estudiadas por Basu and Michailidis (2015) y Peña et al. (2021). En teoría de la señal Donoho (2006a, 2006b) ha demostrado la optimalidad de minimizar la norma L1 para encontrar combinaciones lineales de variables que conservan toda la información relevante.

4.3. Akaike y los métodos automáticos de selección de modelos

El enfoque tradicional de construcción de modelos estadísticos es artesanal: requiere mucha experiencia y formación. Además, cada etapa depende de los conocimientos adquiridos en las etapas previas y el procedimiento es iterativo, refinando el modelo con el análisis de los residuos, o parte no explicada, de los datos. Este método es muy adecuado con pocos datos y variables pero impracticable con miles de variables y modelos. H. Akaike, trabajando en la predicción de series temporales para la industria japonesa, fue el primero en diseñar un procedimiento riguroso y efectivo para la selección automática del modelo estadístico.

Hirotugu Akaike (1927-2009) nació en Japón, en una familia de granjeros que cultivaba gusanos de seda, y fue el menor de cuatro hermanos. Estudió en la Universidad de Tokio, donde obtuvo su doctorado en matemáticas en 1961. Comenzó trabajando en problemas aplicados en ingeniería en un centro creado por el Gobierno para transferir conocimiento a las empresas japonesas, el Instituto de Estadística Matemática. Posteriormente, el Instituto se transformó en un centro universitario de postgrado que Akaike dirigió desde 1986 hasta su retiro en 1994. Fue profesor visitantes en muchas universidades y recibió muchos honores, como el primer premio Japan Statistical Society Prize (1996), y el más importante de Japón, the Kyoto Prize (2006), por su "Major contribution to statistical science and modeling with the development of the Akaike Information Criterion (AIC)". Su vida y sus aportaciones están descritas en Findley and Parzen (1998) y Tong (2010).

La contribución fundamental de Akaike fue encontrar un criterio general para la selección de un modelo estadístico, el AIC, (Akaike, 1973), que escoge entre un conjunto definido de modelos el que minimiza el error de predicción esperado fuera de la muestra. Akaike demostró que este criterio combina una medida de ajuste del modelo a los datos, el logaritmo de la máxima verosimilitud (sustituyendo los parámetros por estimadores de máxima verosimilitud), con una medida de penalización por el error esperado de estimación, que es dos veces el número de parámetros estimados. El modelo que minimiza el AIC es aquel donde la suma del error de ajuste más la penalización es mínima.

El AIC tiene tres componentes que anticipaban el futuro. En pri-

T 

Figura 3: Professor Hirotugu Akaike, creador del primer criterio de selección automática de reglas de predicción

mer lugar, se centra en seleccionar el mejor modelo en un amplio conjunto, cuando la metodología estadística tradicional se basaba en elegir un modelo y contrastar su ajuste. En segundo lugar, pone el énfasis en la predicción fuera de la muestra, a diferencia del ajuste a los datos, cuando el enfoque tradicional dominante en ese momento suponía, erróneamente, que ambos enfoques llevaban al mismo resultado. Akaike se dio cuenta en su trabajo aplicado con series temporales que modelos autoregresivos de distinto orden podían ser adecuados con los contraste tradicionales de ajuste y que el modelo de mejor ajuste puede no ser el de mejor predicción. Para ello, es imprescindible tener en cuenta el error de estimación de los parámetros, ya que aumentar su número nunca puede empeorar el ajuste y lleva a seleccionar modelos con más parámetros de los necesarios para una buena predicción. En tercer lugar, el AIC une la estimación del modelo y la selección con un criterio único basado en la verosimilitud, expandiendo el enfoque de Fisher sobre la estimación máximo verosímil y relacionándolo con la entropía y la teoría de la información.

El trabajo de Akaike abrió la puerta a la comparación automática de mucho modelos y a la selección del mejor por su comportamiento predictivo, que es el enfoque utilizado, años más tarde, en aprendizaje automático e inteligencia artificial y la base de construcción de reglas de predicción en la emergente ciencia de datos. Poco después, Stone (1974) introduce la validación cruzada, es decir, la predicción fuera de la muestra, como criterio general no parámetrico de selección de

reglas de predicción y establece su relación asintótica con AIC, (Stone, 1977), y Schwarz (1978), propone el criterio BIC para seleccionar modelos desde un punto de vista bayesiano. Los criterios de selección de modelos han servido para unificar el tratamiento de muchos problemas en estadística. Por ejemplo Peña y Galeano (2008) han mostrado que los problemas de discriminación, bondad de ajuste y detección de atípicos en series temporales pueden unificarse con este criterio.

Una gran ventaja de la validación cruzada es su carácter no paramétrico, al no requerir hipótesis sobre la generación de los datos a diferencia de los criterios de selección de modelos, que comparan modelos paramétricos. En su aplicación es imprescindible que la muestra de validación no se utilice en absoluto para ninguna decisión relacionada con la construcción del modelo, y se reserve para su validación. Hay distintas formas de realizar la validación cruzada para datos independientes, por ejemplo dividir la muestra de tamaño n en k partes al azar, dejar una parte fuera para validarlo y estimar el modelo con las restantes $k-1$. El proceso se repite para cada una de las k partes y se hace el promedio de los resultados obtenidos. Este método se llama k-validación cruzada. En el caso particular de $k = n$ el modelo se estima con $n-1$ datos y se valida con n.

Estos métodos no son adecuados para datos dependientes, como series temporales, porque si muestreamos la serie al azar destruimos la dependencia secuencial de las observaciones. Para datos temporales se utiliza la primera parte de la muestra para estimar el modelo y la segunda para validarla, aunque otros procedimientos para dividir la muestra son posibles (véase Peña and Sánchez, 2005). Este campo, sin embargo, debe desarrollarse mucho más para encontrar procedimientos más robustos y eficaces de comparar modelos dinámicos.

Evaluar al modelo por su capacidad predictiva está sustituyendo al enfoque de seleccionar las variables con los contrastes clásicos de significación sobre los coeficientes de un modelo estimado, que puede lleva a modelos con poca capacidad predictiva. En particular, el procedimiento habitual de incluir variables en un modelo cuando su p-valor es menor que 0.05, o su estadístico t es mayor en valor absoluto que dos, no es adecuado para datos masivos. Su análisis, ha puesto de manifiesto que los contrastes de significación tienden a rechazar cualquier hipótesis si el tamaño de la muestra es suficientemente grande. Por estas razones, la utilización de p-valores y de

contrastes de significación en el análisis de datos ha sido formalmente desaconsejada por *The American Statistical Association* (ASA) en un comunicado oficial (Wasserstein and Lazar, 2016). Esta asociación ha publicado un número extraordinario de *The American Statistician* en 2019 con más de 40 trabajos que, desde distintos puntos de vista, recomiendan que el estudio de las relaciones entre variables se haga mediante estimación y no con contrastes de significación.

4.4. Breiman y la combinación de predicciones

El paradigma clásico de la estadística es encontrar el mejor modelo para la muestra observada. El modelo óptimo está bien definido en entornos simples, bajo fuertes hipótesis sobre el proceso generador de los datos, pero empieza a desdibujarse cuando admitimos incertidumbre sobre este proceso, o prescindimos de un modelo generador único. Entonces, resulta razonable considerar todos aquellos modelos compatibles con la muestra observada y combinarlos después para construir la predicción. Por otro lado, el enfoque de ciencia de datos pone más el énfasis en la predicción que en el modelado. (Breiman 2001a). Un enfoque general, propuesto por Breiman, una de las figuras principales en la creación de la disciplina de ciencia de los datos, es introducir modificaciones en la muestra, cambiando algunas observaciones o variables, generar predicciones con cada modificación y combinarlas.

Leo Breiman (1928-2005) nació en Nueva York hijo de emigrantes judíos del este de Europa. Tras estudiar físicas en CALTEC obtuvo su doctorado en matemáticas en la Universidad de California en Berkeley en 1954, con una tesis sobre convergencia en probabilidad dirigida por Michel Lóeve. Su carrera ha combinado la vida académica con la consultoría y ha evolucionado de la dedicación a la probabilidad, al análisis de los datos masivos. Fue contratado como probabilista en el departamento de matemáticas de la Universidad de California en Los Angeles (UCLA) y, siendo ya Professor, lo abandonó para dedicarse a la consultoría durante trece años. Volvió a Berkeley en 1980 y en los siguientes 20 años, desde 1993 como profesor emérito, revolucionó la estadística con la invención de los árboles de clasificación (CART), la alternancia entre esperanzas condicionadas *ACE* y la combinación de modelos de predicción *bagging*, *boosting*, y los bosques aleatorios, *random forest*. Sus contribuciones más importantes las realizó ya jubilado, permaneciendo muy activo en investigación

Figura 4: Professor Leo Breiman, inventor de nuevos métodos para combinar modelos y predicciones

hasta su fallecimiento. Ese año había recibido el *SIGKDD Data Mining and Knowledge Discovery Innovation Award*, el más importante galardón en estas áreas. Fue un hombre de muchos intereses y un reconocido escultor, véase Ohlssen (2001).

Combinar modelos para mejorar las predicciones tiene una larga historia en estadística que se remonta a Bates y Granger (1969). Desde entonces, se ha desarrollado una amplia literatura en combinación de modelos, tanto desde el punto de vista clásico, combinando las predicciones generadas por cada modelo por su precisión relativa, como bayesiano, ponderando las predicciones por las probabilidades a posteriori de los modelos que las generan. Este último método se denomina el promedio bayesiano de modelos (*Bayesian Model Averaging*) y tiene muchas aplicaciones. Véase Draper (1995) y Hoeting et al. (1999). Además de estos métodos, desde 1990 se han desarrollado otros para combinación de predicciones, que se conocen como *ensamble methods*, o combinación de métodos, en la literatura de *machine learning*, y en los que Breiman ha tenido un papel muy destacado.

Breiman fue uno de los artífices de los árboles de clasificación, CART, Breiman et al. (1984), que son reglas de clasificación basadas en decisiones dicotómicas que minimizan el error de clasificación. Breiman observó que, con muchas variables, pequeños cambios de sus valores podían tener grandes efectos en el resultado de la clasificación (Breiman 1996b). Para hacer el resultado más robusto a pequeñas perturbaciones, introdujo para los modelos CART lo que

llamó *bagging* (bootstrap aggregation). El método consiste en tomar muestras con reemplazamiento de los datos, construir el predictor, repetir el procedimiento y, después, promediar todas las predicciones (o modelos) por su error de predicción fuera de la muestra. Este método de perturbar los datos, construir modelos y promediar los resultados, puede aplicarse a todo procedimiento estadístico para mejorar sus predicciones. (Breiman, 1996a)

La idea de perturbar los datos se extendió a la de modificar las variables en los bosques aleatorios,(*random forests*), Breiman(2001b). Para cada muestra de datos se selecciona al azar un conjunto de variables en cada nodo del árbol de clasificación, con lo que se crea un conjunto de árboles cuyas predicciones se combinan por su precisión relativa. *Boosting* es otro enfoque de combinar modelos, propuesto por Freund and Schapire (1996, 1997), donde se crea una regla de predicción compleja combinando modelos muy simples. El modelo va creciendo por la capacidad de los modelos simples de ir explicando los residuos de los modelos anteriores, y, en cada etapa. se da más peso a las observaciones mal clasificadas, o que tienen un residuo más alto. La predicción final es una función compleja obtenida combinando los modelos ajustados en cada de las etapas utilizadas. Breiman (1998) ha propuesto una clase general de modelos con esta filosofía que ha denominado *Arcing (adaptive resampling or adaptive reweighting, and combining).* Su idea es cambiar en cada iteración los pesos de las observaciones más difíciles de clasificar bien, que están cerca de la frontera de clasificación, o de las observaciones más difíciles de prever.

El enfoque de Breiman es distinto del tradicional en estadística de combinar modelos ajustados a la misma muestra. Su idea es aplicar una clase muy general y flexible de modelos a distintas perturbaciones de la muestra, seleccionando los datos al azar o cambiando las ponderaciones de las observaciones, para obtener predicciones más robustas y precisas.

La abundancia de datos de distintas periodicidad y precisión ha propiciado métodos para combinar datos de distintas clases. La combinación de distintos tipos de datos para la predicción con series temporales fue estudiada por Guerrero, and Peña (2003). Recientemente, se han desarrollado enfoques para la predicción a muy corto plazo llamados de Nowcasting, (nombre tomado de la predicción meteorológica combinando Now and forecasting) que utilizan los datos mensuales

o trimestrales habituales y otros de alta frecuencia, como datos diarios o semanales. Por ejemplo, el método MIDAS (mixed-data sampling), combina datos temporales de distinta frecuencia; véase Kuzin et al. (2021) y Peña and Tsay (2020) para una descripción del mismo y ejemplos de su aplicación. En otros casos, mezclamos datos de series temporales y de sección cruzada, (Galeano y Peña, 2019) o textos, imágenes de video y audios con variables tradicionales. Un ejemplo reciente del uso de vídeos para la predicción puede encontrarse en Sun et al. (2019).

4.5. Efron y el ordenador para generar nuevo métodos

Bradley Efron (1938 -) nació en St. Paul, Minnesota hijo de emigrantes judíos rusos. Sus padres tuvieron cuatro hijos y escasos recursos económicos provenientes del trabajo de su padre como conductor de camiones. Efron pudo ir a la universidad gracias a una beca para estudiar matemáticas en The California Institute of Technology (CALTEC). Después, se doctoró en estadística en Stanford University en 1964. Durante sus estudios de postgrado dirigió una revista satírica estudiantil y fue expulsado de la universidad durante 6 meses por publicar una parodia de la revista Play Boy que se consideró irreverente desde el punto de vista religioso. Ha trabajado siempre como Professor of Statistics and Biostatistics at Stanford con contratos como visitante en Harvard University; Imperial College,London, y the University of California, Berkeley. Es de los estadísticos más premiados: 2014 the Guy Medal in Gold by the Royal Statistical Society, 2005 The National US Medal of Science, 2018 The International Prize in Statistics y 2020 The BBVA Frontiers of Knowledge Price. Holmes et al. (2003), Champking (2010) y Cochran (2015) contienen entrevistas y artículos sobre su vida y sus múltiples aportaciones a la estadística.

La contribución más influyente de Efron es el estimador *Bootstrap* o estimador autosuficiente (Efron, 1979), que es una generalización del jackknife introducido por Quenouille (1949) y Tukey (1958). Dada una muestra aleatoria de tamaño n queremos estimar una característica, como la media o la varianza de una población desconocida. Una vez elegido el estimador, por ejemplo, la media o varianza muestral, el bootstrap nos proporciona automáticamente una estimación del error esperado para el estimador elegido. El procedimiento es tomar B muestras, donde B es grande, con reemplazamiento, de tamaño n

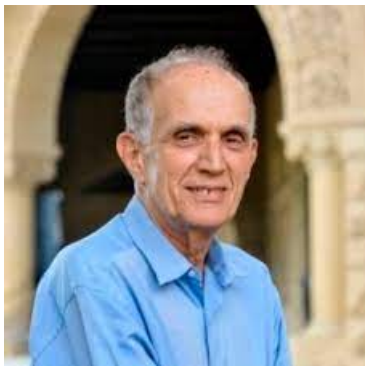

Figura 5: Professor Bradley Efron, creador de métodos de análisis con ordenador

y calcular el estimador en cada una, obteniendo B valores posibles para el estimador. La desviación típica de esos B valores nos estima la desviación típica del estimador utilizado. Es intuitivo que este método debe funcionar en muestras muy grandes cuando B sea también grande, pero inesperado que lo haga bien con valores pequeños de n y B.

Muestrear los datos para obtener errores de estimación ya había sido utilizado antes en estadística. Stigler (1991) cita tres casos de simulación de problemas estadísticos en la última parte del siglo XIX, pero el uso de simulación para obtener el error esperado de un estimador fue llevado a cabo, por primera vez, por Gosset (1876-1937), un matemático inglés que trabajaba en la fábrica de cerveza Guinness investigando cambios en el proceso de fabricación para mejorar su calidad. Supuso que las medidas efectuadas en el proceso seguirían una distribución normal y trató de calcular el error de la media muestral como estimador de la media poblacional en una muestra de pequeño tamaño, n. La influencia de las teoría de Cesare Lombroso (1835-1909) un médico italiano que suponía una relación entre la criminalidad y las características físicas y biológicas, habían llevado a Scotland Yard a recoger datos físicos de delincuentes. Gosset utilizó una base de datos con la altura de 3000 criminales que escribió en fichas y obtuvo muestras al azar sin reemplazamiento de tamaño n de estas fichas, calculando en cada una la media de las estaturas. Por ejemplo, con tamaño muestral 4 obtuvo 750 valores (3000/4), y calculó la desviación promedio entre la estatura media de los 3000

criminales y las 740 estimaciones obtenidas con las muestras. Con este procedimiento, pudo construir tablas de las desviaciones para distintos valores de n.

Su trabajo, Student (1908), fue publicado con seudónimo, ya que Guinness no permitía a sus empleados publicar sus investigaciones. Posteriormente, Fisher (1925c) encontró la forma matemática de la distribución, que denominó t de Student en honor a Gosset. El enfoque de Efron obtiene, gracias a los ordenadores, automáticamente el error muestral de cualquier estimador en situaciones mucho más complejas que la estudiada por Student. Efron demostró que, en condiciones generales, el método autosuficiente es mejor que la aproximación asintótica usando el teorema central del límite y la distribución normal.

El bootstrap, como su antecesor el jacknife, permite la generación de muestras de los datos con pocas perturbaciones y ha sido aplicado en muchos campos. Por ejemplo, para combinar distintas reglas de predicción y obtener un mejor predictor con una red neuronal profunda o un árbol de clasificación (CART), (Hastie et al., 2011), como hemos comentado al explicar el *Bagging (bootstrap aggregation)*, en la sección anterior. En el campo de series temporales el muestro aleatorio no es adecuado, como hemos comentado para la validación cruzada. Kunsch (1988) propuso dividir la serie en bloques y muestrear los bloques (*block bootstrap*) y Bühlmann (1997) ajustar un proceso autorregresivo y muestrear los residuos y llamó a este método el estimador autosuficiente tamizado (*Sieve bootstrap*). Alonso et al. (2003) han propuesto unir los bloques de forma suave suponiendo valores ausentes entre ellos y Alonso et al. (2004) métodos para mejorar la medición de la incertidumbre en la predicción.

La utilización de métodos de Monte Carlo se ha aplicado con éxito a la estimación bayesiana con la simulación de cadenas de Markov o estimación con *Markov Chain Monte Carlo (MCMC)*. Cuando la distribución posterior es compleja y requiere integración en alta dimensión, podemos obtener muestras de ella muestreando iterativamente en distribuciones condicionadas de baja dimensión, con frecuencia univariantes. Este aspecto fue utilizado por Geman and Geman (1984) que introdujeron el nombre de Gibbs sampling para este procedimiento. Posteriormente, Gelfand and Smith (1990) demostraron que esta idea puede aplicarse con gran generalidad en problemas bayesianos de estimación. Justel y Peña (1996) estudiaron las pro-

piedades de Gibbs Sampling en presencia de atípicos. Véase Martin, Frazier, and Roberts (2020) para una revisión de estos métodos, enmarcada en la historia de la computación bayesiana.

5. Cinco ejemplos de áreas generadoras de nuevos métodos estadísticos

Muchos de los modelos utilizados en el análisis de *big data*, o datos masivos, han nacido con las aplicaciones de la estadística en distintos campos científicos. Además, con frecuencia, métodos surgidos en un campo han estimulado nuevos avances en otros, y la estadística ha favorecido la fertilización cruzada o hibridación en la ciencia. En esta sección vamos a ilustrar estos aspectos con cinco ejemplos.

5.1. Psicología, el modelo factorial

Charles Spearman (1863-1945) estudió psicología en Leipzig graduándose con 44 años, tras ejercer durante tres lustros como oficial del ejército británico en la India. Siendo estudiante, publicó su famoso trabajo, Spearman (1904), argumentando que los resultados de los tests de aptitudes, que se estaban desarrollando en esa época, podían explicarse como el efecto lineal de un factor de inteligencia general, que llamó factor g, que era innato al individuo y provenía de su herencia genética. De acuerdo con esta teoría, la inteligencia de los individuos podía ordenarse a lo largo de una sola dimensión. Posteriormente, como la teoría del factor general no explicaba las diferencias en habilidades espaciales o numéricas de las personas, introdujo un segundo factor específico, que dependía del tipo de aptitud.

Unos años más tarde, Louis Thurstone (1887-1955) formuló el modelo factorial general con muchos factores dependientes, como una herramienta básica de análisis de los datos sociales. Las técnicas iniciales utilizadas para la estimación de los factores se basaron en reducir la matriz de covarianzas, o de correlaciones, a una estructura más simple (véase Carroll and Schweiker, 1951) por las dificultades de utilizar los componentes principales, introducidos por Hotelling (1933), (y que tienen antecedentes en el trabajo de K. Pearson, véase Burt, 1949), por las limitaciones de cálculo existentes en esos años. Lawley (1940) estudió su estimación máximo verosímil pero el modelo factorial no se introdujo en la metodología estadística hasta el

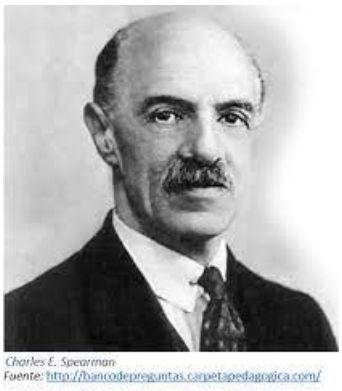

Figura 6: Professor Charles Spearman, creador del análisis factorial

trabajo de Lawley and Maxwell (1963), cuando la aparición del ordenador hizo posible su estimación con varias variables. En España, debemos al psicólogo Mariano Yela (1921-1984), que estudió en la Universidad de Chicago con Thurstone, la introducción del análisis factorial. Yela ha sido un exponente destacado de la psicomatemática en nuestro país, desarrollando en 1969 la especialidad de Psicología en la Universidad Complutense de Madrid.

El modelo factorial ha sido utilizado extensamente en las ciencias sociales, pero, también, en la ingeniería y las técnicas de computación. Una extensión muy utilizada en las ciencias sociales son los modelos estructurales con variables latentes, o LISREL, acrónimo de *linear structural relations*, véase Hayduk (1987) y Jöreskog et al. (2016), para una revisión de este campo y Satorra and Bentler (2001) para test de ajuste en estos modelos.

El análisis factorial está muy presente en las aplicaciones actuales con datos masivos. En primer lugar, el análisis de datos dependientes de alta dimensión requiere un ingente número de parámetros, ya que la dependencia entre dos variables puede manifestarse con distintos coeficientes en h periodos distintos, añadiendo h parámetros para representar cada relación bivariante. Lo mismo ocurre con la dependencia espacial. El análisis factorial de estos datos es una herramienta muy poderosa para modelar datos dependientes y está teniendo un gran auge actualmente, tanto en economía como en las ciencias del medio ambiente, con datos temporales y espaciales. Estos aspectos serán desarrollados en las secciones 5.2 y 5.5.

En segundo lugar, la idea de variables latentes que son combina-

ciones lineales de las variables observadas se generaliza en modelos con distintas etapas o capas introduciendo variables latentes que son combinaciones de las variables anteriores. Por ejemplo, en una red neuronal profunda, las variables de entrada en cada etapa, después de la primera, son combinaciones lineales de las variables latentes precedentes. Las variables latentes están en la base de los métodos actuales de aprendizaje profundo aplicados en ingeniería y computación.

En tercer lugar, las variables latentes se han mostrado muy útiles para el estudio de la heterogeneidad en alta dimensión, buscando proyecciones que revelen una estructura subyacente de grupos. Esta idea fue propuesta por Friedman and Tukey (1974), con el nombre de *projection pursuit*. Peña and Prieto (2001a) han desarrollado un método de cluster con este enfoque y, también, procedimientos robustos para detectar observaciones atípicas multivariantes en alta dimensión (Peña and Prieto, 2001b, 2007).

5.2. Economía, series temporales

La importancia de la estadística en economía queda de manifiesto en la siguiente reflexión de un manual clásico (Lipsey, 1971, p.53): "Adentrarse en el estudio de la economía o en cualquier investigación social sin conocimiento alguno sobre análisis estadístico constituye sin duda alguna un riesgo muy grande que, en definitiva, puede ocasionar que el trabajo emprendido sea completamente inútil o incluso abiertamente erróneo". La estadística aporta la estructura conceptual imprescindible para la comprensión de muchas variables macroeconómicas y de las relaciones entre ellas. El primer premio Nobel de Economía fue concedido a Tinbergen, formado en física y en estadística en los Países Bajos, por incorporar la modelización estadística de variables económicas con un sistemas de ecuaciones estructurales. En contrapartida, también la estadística es deudora de los análisis económicos, ya que conceptos como multicolinealidad o endogeneidad (Engle et al., 1983) han surgido en la econometría, es decir, del análisis estadístico de las relaciones económicas.

Esta interacción entre ambas disciplinas ha sido especialmente fuerte en las series temporales, que son clave para la predicción económica. Los procesos estocásticos estacionarios fueron estudiados por Yule, Bartlett y Cramer, entre otros y la teoría de predicción de series estacionarias es debida a Winner y Kolmogorov. Muchas de las series

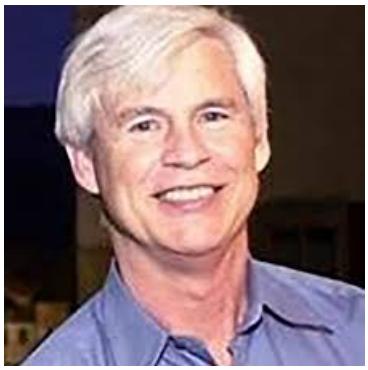

Figura 7: Professor Engle, inventor de nuevos métodos estadísticos para el analisis de datos económicos

que observamos en procesos técnicos o ambientales son aproximadamente estacionarias pero no en economía. En enfoque econométrico tradicional era suponer las series generadas como superposición de una tendencia determinista, una estacionalidad cíclica y un componente estacionario. La introducción de las series temporales integradas, debida a Box y Jenkins (1970) generalizó esta representación determinista a una estocástica, proporcionando un tratamiento riguroso de la predicción y modelado de estas series.

Una de las personas más destacadas en la generación de nuevas ideas de alcance general para las series de datos económicos es Robert F. Engle (1942 -). Se doctoró en 1969 en Cornell University y fue profesor en el MIT y en la University of California, San Diego (UCSD). Actualmente, trabaja en the Stern School of Business, New York University, Engel (1982) propuso un modelo para la varianza condicional, la volatilidad, de las series financieras: el modelo ARCH. Este modelo ha tenido muchas aplicaciones y extensiones y es básico para el análisis financiero. También ha encontrado aplicaciones en otros campos para modelar procesos con valores positivos. Por otro lado, Box and Tiao (1977) descubrieron que dos variables no estacionarias pueden estar relacionadas de forma estacionaria, es decir, que la combinación lineal de ambas que define su relación es una serie estacionaria. Engle and Granger (1987) llamaron a estas variable cointegradas y desarrollaron sus propiedades para la teoría del equilibrio económico. Por estos descubrimientos les concedieron a ambos el Premio Nobel de Economía en 2003.

Cointegración esta muy relacionada con modelos factoriales dinámicos, véase Escribano and Peña (1994). Estos modelos, fundamentales para series temporales masivas, fueron introducido para variables económicas por Geweke (1977), Chamberlain (1983) y Chamberlain, G. and Rothschild, M. (1983). En estadística fue iniciado de forma independiente en los trabajos de Engle and Watson (1981) y Peña and Box (1984). En estos modelos la dependencia entre las series es consecuencia de un componente común, que recoge el efecto de los factores o variables latentes no observadas. Además, cada serie es afectada por un término específico o idiosincrático, que resume la dinámica propia de esa serie. Estos modelos han tenido un desarrollo espectacular en los últimos 20 años, gracias a la disponibilidad de datos masivos. Para variables integradas fueron introducidos por Peña and Poncela (2006). Véase Peña and Tsay (2020) para una revisión de su situación actual. Muy recientemente los modelos factoriales dinámicos han sido generalizados, primero para matrices de series temporales (Wang et al.,2019) y luego para tensores de series temporales de cualquier dimensión (Chen et al., 2022; Peña, 2022; Wang et al., 2022). Estas herramientas surgidas en economía son hoy muy útiles para el análisis de redes de transporte o sociales (Chen et al., 2022).

La estimación de los modelos factoriales dinámicos con retardos en las variable (Forni et al., 2000) se ha realizado con los componentes principales dinámicos introducidos por Brillinger en el dominio de la frecuencia. Una generalización de esta idea es definir estas componentes directamente en el dominio del tiempo y estimarlas por un algoritmo iterativo. Estos componentes han sido propuestos, con aplicaciones a la predicción económica, por Peña and Yohai (2016), y Peña et al., (2019a) pero están ahora siendo utilizado en medicina para el análisis de señales cerebrales (Wang et al., 2019).

5.3. Ingeniería, redes neuronales y deep learning

Muchas ramas de la ingeniería y la computación han estimulado la creación de modelos estadísticos para datos masivos que han sido bienvenidos en otras disciplinas. Los métodos de estimación recursiva, como el filtro de Kalman (1960), aplicados inicialmente a la ingeniría aeronáutica, son ahora una herramienta básica en el diseño de coches autodirigidos y robótica, y se han utilizado para la predicción de variables macroeconómicas. Los métodos de inteligencia artificial para el reconocimiento de imágenes, de texto escrito, y de

voz y sonido digitalizados, han estimulado nuevos métodos que luego han sido aplicados en medicina y ciencias sociales. Por ejemplo, el análisis de imágenes comenzó a introducirse en el control de calidad de ciertos procesos, véase por ejemplo Benito and Peña (2007), pero ahora tiene una importancia creciente en el reconocimiento de imágenes medicas (véase Alfaro-Almagro et al.,2018). Los métodos de reconocimiento de voz se encuentran implantados en muchos medios de comunicación. Además, los nuevos datos como imágenes o sonidos suelen tener estructuras no lineales y han estimulado métodos no paramétricos, como los vecinos más próximos, los métodos de suavizado y el análisis funcional de datos.

Vladimir Vapnik(1936-) nació en la URSS y obtuvo su doctorado en estadística en el Instituto de Ciencias de Control en Moscú en 1964, donde trabajó hasta su emigración a EE.UU. en 1990, contratado por los laboratorios de ATT Bell en New Jersey. Actualmente es profesor de Informática (Computer Science) en la Universidad de Columbia en NY. Ha recibido mucho honores, que pueden consultarse en su página web y en las entrevistas sobre sus trabajo en youtube (véase por ejemplo https://www.youtube.com/watch?v=2M1JCxXkNKE).

Vapnik (1998, 1999) ha elaborado un enfoque para el aprendizaje estadístico que ha tenido mucha repercusión en ingeniería y en *Machine Learning*, o aprendizaje automático. En su trabajo en Bell descubrió que en problemas de clasificación complejos, con muchas variables, en lugar de considerar la distribución conjunta de todas las variables, como se hace en la función discriminante de Fisher, es más efectivo buscar la mejor separación local alrededor de la frontera entre las clases a separar. La propuesta de Vapnik (2013) con las máquinas de vector soporte ha encontrado muchas aplicaciones en ingeniería y ciencias de la computación (Drucker et al., 1996). Además, en problemas no linealmente separables, es posible utilizar un núcleo (kernel) para transformar las variables y llevarlas a un espacio de dimensión mayor donde la separación sea posible.

Un método desarrollado en el aprendizaje automático *(Machine learning)* para representar una relación cualquiera entre una variable respuesta, que puede ser continua o discreta, como ocurre en los problemas de clasificación, y un conjunto amplio de variables potencialmente explicativas son las redes neuronales, o *Neural networks*. Esta representación considera las variables explicativas como variables de entrada y con ellas se forman combinaciones lineales, o factores, que

Figura 8: Professor Vladimir Vapnik, descubridor de máquinas de vector soporte para la clasificación

producen una respuesta no lineal. Las respuestas se combinan entre sí para formar nuevos factores que, de nuevo, actúan no linealmente. El número de capas y de factores necesarios en cada capa, y su composición se determinan de forma empírica, de manera que la respuesta o predicción sea lo mejor posible. La red se entrena, o se estiman sus parámetros, minimizando el error cuadrático de predicción, con un algoritmo no lineal basado en el gradiente. Se denominan redes profundas o métodos de aprendizaje profundo (*Deep learning)* las que utilizan muchas capas para la predicción.

Las redes neuronales tradicionales no están pensadas para variables dinámicas y procesan todas las observaciones sin tener en cuenta su orden temporal. En los últimos años se han desarrollado redes que procesan secuencialmente las observaciones, como las Recurrent Neural Networks (RNN), o redes neuronales recurrentes. Estas redes procesan primero el vector de variables de entrada para obtener la respuesta, pero, en el periodo siguiente, una parte de esa respuesta se introduce como variable de entrada y se combina con el resto de las variables para generar la respuesta, o predicción, y así sucesivamente. De esta forma, la predicción en el instante t depende de los valores de las variables explicativas en el instante $t-1$, pero, también, de las predicciones anteriores, con memoria decreciente. Estas redes tratan de reproducir los procesos autorregresivos de series temporales (véase Peña and Tsay, 2020).

5.4. Medicina, contrastes múltiples

Los problemas médicos y biológicos han sido un estímulo constante en el desarrollo de la estadística durante todo el siglo XX. Han dado lugar a departamentos separados de Bioestadística e impulsado nuevos métodos de contrastes mútiples, el estudio de datos censurados o faltantes, la introducción de variables explicativas para la clasificación y métodos para combinar los resultados de varios estudios con meta análisis, entre otros aspectos. Los artículos estadísticos más citados de todos los tiempos son los aplicados a las ciencias de la salud (Van Noorden, et al., 2014, y Ryan and Woodall, 2005): nueve de los diez artículos más citados van dirigidos a este campo. En particular, el modelo logístico y la función de supervivencia propuesta por Cox (1972) ha tenido una gran repercusión en el último cuarto del siglo XX.

Nan Laird (1943-) ha estado en el centro de los avances en ciencias de la salud como consecuencia de la disponibilidad de datos masivos. Doctorada en Harvard y profesora en Harvard School of Public Health hasta su jubilación en 2015. Su aportacion a los estudios longitudinales le han proporcionado en 2021 el International Prize in Statistics, que solo habían recibido antes David Cox, en 2017, y Bradley Efron, en 2019. Véase Ryan, (2015) para una descripción de su vida y contribuciones. Laird es una de las creadoras del algoritmo EM, Dempster er al (1977), que surgió durante su trabajo de tesis por la necesidad de disponer de un procedimiento general para la estimación de valores ausentes, o missing values, y es uno de los dos artículos de estadística metodológicos más citados (Ryan and Woodall, 2005).

La gran contribución de Laird es al análisis de datos longitudinales, es decir, los provenientes de estudios que siguen grupos de personas a lo largo del tiempo tomando medidas regulares de su evolución. Estos estudios pueden involucrar grandes masas de datos y son mucho más fiables que los tradicionales de corte transversal, donde se comparan los resultados de dos grupos de personas en distintos momentos del tiempo, ya que pueden separar los efectos debidos a la persona de los de la situación de partida del grupo de personas. A cambio, son más difíciles de analizar por la común presencia de datos faltantes y censurados y la necesidad de mezclar la dimensión temporal y la transversal. Laird y sus colaboradores establecieron la metodología estadística para este tipo de estudios (Laird and Ware, 1982, Fitzmaurice et al., 2012). Laird ha hecho también contribuciones im-

Figura 9: Profesora Nan Laird pionera en el análisis longitudinal en medicina

portantes al meta-análisis, desarrollado en medicina para combinar medidas resumen de distintas publicaciones y obtener una valoración global del diagnóstico y pronóstico de una enfermedad (DerSimonian et al., 1986). En la actualidad, se está iniciando la posibilidad de compartir directamente los datos de los pacientes en distintos hospitales, anónimos y convenientemente protegidos, lo que hará decrecer la importancia del meta-análisis que ha ayudado de forma importante al avance de la medicina (véase Hedges and Olkin, 2014).

El estudio del genoma humano ha requerido realizar millones de contrastes de hipótesis para comparar la función de los aproximadamente 24.000 genes que forman el ADN humano. Estos contrastes se desarrollaron en estadística para una hipótesis única. Posteriormente, ante la evidencia de que hipótesis correctas son rechazadas con frecuencia al contrastarlas repetidamente, se establecieron correcciones, de manera que el nivel de significación del contraste dependa del número a realizar, como el método de Bonferroni y otros. Estos procedimientos resuelven el problema de contrastes múltiples cuando el número de contrastes a realizar es de una decenas, pero son claramente inadecuados para realizar millones de comparaciones.

Benjamini and Hochberg (1995) propusieron un procedimiento, generalizado en Benjamin (2010), para realizar un contraste controlando la frecuencia con lo que se rechazan hipótesis que son ciertas al realizar el contraste muchas veces, lo que llamaron *False Discovery Rate (FDR)*, que es el valor esperado del cociente entre las hipóte-

sis rechazadas por error y todas las hipótesis rechazadas. Una tasa de FDR=.05 en un contraste nos dice que es esperable que de todas las hipótesis rechazadas solo el 5 % se ha hecho erróneamente. El procedimiento propuesto considera conjuntamente todos los p-valores correspondientes a los contraste y rechaza todos los que superan una cota, que depende de la FDR establecida.

Este enfoque se ha utilizado mucho con datos genéticos con millones de variables. Por ejemplo, Tzeng et al. (2003) han propuesto un estadístico para descubrir los genes responsables de ciertas enfermedades genéticas que se calcula en muchas regiones del genoma y utiliza el método FDR para controlar la falsa asociación de genes y enfermedades.

Recientemente, la pandemia del COVID-16 está estimulando nuevos métodos para aprender de los numerosos datos que están siendo generados en todo el mundo. Véase Liu et al. (2021) para un ejemplo de estos análisis y López-Cheda, et al. (2021) y Vallejo et al. (2022), para ejemplos de aportaciones españolas.

5.5. Medio ambiente, modelos espaciales y funcionales

Los datos actuales llevan, cada vez con más frecuencia, su localización temporal y espacial. Esto hace posible la construcción de mapas geográficos por ordenador, o geographical information systems (GIS), para describir la distribución de una variable climática o medio ambiental en una zona concreta. También se han desarrollado procedimientos para representar grandes conjuntos de series temporales (Peña et al.,2019b). Por otro lado, la importancia del cambio climático has impulsado la recogida y análisis de datos medio ambientales, por ejemplo masivamente por satélites, propiciado nuevos métodos de alisado de curvas y de análisis de datos funcionales.

Greta Wahba (1934-) ha sido una gran impulsora de los métodos de suavizado para datos climáticos. Tras graduarse en matemáticas en Cornell en 1956 trabajó diez años en la industria y, siendo madre separada con una hija pequeña, consiguió matricularse en Stanford University para realizar un doctorado a tiempo parcial. Simultaneó su trabajo en IBM con su doctorado en estadística, que obtuvo en 1966. Después, fue contratada por la Universidad de Wisconsin-Madison y allí continúa hoy, muy activa como profesora emérita. Ha recibido muchos honores por su trabajo, como el Inaugural Senior Breiman Award, 2017 o el COPSS Fisher Award, 2014. (Véase su página web,

Figura 10: Professora Grace Wahba, pionera en los métodos de suavizado de superficies

https://pages.stat.wisc.edu/ wahba/ y la entrevista de Nychka et al., 2020).

Wahba comenzó trabajando en splines y propuso, con sus coautores, el método más utilizado para estimar el parámetro de regularización, la validación cruzada generalizada o *generalized cross validation, GCV*, Golub et al., (1979). En segundo lugar, encontró una formulación general de muchos problemas de estimación estadística, como splines o máquinas de vector soporte, en el marco de espacios de Hilbert con núcleo reproductor, o *reproducing kernel Hilbert spaces. RKHS*, Wahba, G. (1999). En tercer lugar, generalizó el método de máquinas de vector soporte, introducido por Vapnik para dos clases, para cualquier número de grupos, Lee et al. (2004). Sus métodos se han aplicado extensamente en el análisis de datos masivos meteorológicos y ambientales.

Los datos espaciales aparecen en estadística con los estudios geomineros de Matheron (1963), el creador de la geoestadística (vease Cressi, 1991, y Diggle, 2013), que aglutina los trabajos previos en ingeniería forestal (Matern, 1960) y minera (Krige, 1951), para encuadrarlos dentro de una metodología global de datos espaciales. Besag (1974) estudió los datos espaciales como realizaciones de procesos estocásticos y su trabajo fue importante por atraer el interés de los estadísticos a este campo. Un procedimiento bayesiano para eliminar el ruido o la degradación de una imagen aparece en el trabajo pionero de Grenander (1983) y su método de limpieza de imágenes aplicando Gibbs sampling ha sido extendido y popularizado con el trabajo de Geman and Geman (1984), y con los métodos MCMC, comentados

en la sección 4.1.

Otros ejemplos de cómo el análisis de datos de medio ambiente ha hecho surgir modelos útiles en otras áreas son los siguientes. Para detectar el efecto de los gases CFC (clorofluorocarburos) en la capa de ozono se han desarrollado métodos para detectar tendencias muy pequeñas en series temporales (Reinsel and Tiao, 1987) basados en efectos aleatorios. Estos procedimientos han sido después eficaces en el análisis de datos longitudinales en medicina. El análisis no paramétrico de datos meteorológicos mediante datos funcionales (Wang et al., 2016) ha desarrollado procedimientos utilizados para la comprensión del lenguaje hablado en Inteligencia artificial. Modelos espaciales desarrollados para el estudio del clima han contribuido al estudio epidemiológico de la difusión del COVID-16, véase Amdaoud et al. (2021). Finalmente, modelos en espacio-tiempo para datos masivos de medio ambiente, (véase Peña and Tsay, 2021), se están aplicando con éxito en la econometría.

6. Conclusiones

La estadística ha tenido un papel creciente en el avance de la ciencia, proporcionando las herramientas para construir modelos matemáticos no deterministas para contrastar teorías científicas, comprobar la verosimilitud de las hipótesis, y generar predicciones. Sin embargo, es esperable que, con la abundancia de datos masivos, asuma un nuevo protagonismo en la generación de posibles relaciones entre variables que puedan dar lugar a la creación de teorías científicas. Las nuevas formas de recoger de forma automática datos en muchos entornos incorporarán a los dispositivos de recogida procedimientos automáticos de análisis, encaminados a detectar relaciones no previstas entre variables que puedan mejorar las predicciones. Corresponderá a los científicos en cada campo analizar si las asociaciones detectadas indican o no relaciones causales, y proponer experimentos para verificarlo. De esta manera, la estadística añadirá a su tradicional papel de método científico para la contrastación de teorías, el de herramienta ‘para sugerir otras nuevas.

El análisis automático de datos para construir reglas de decisión en muchos campos sociales se extenderá en el futuro y es importante controlar los sesgos asociados a algoritmos que reproducen el pasado y que funcionan como cajas negras cuyos detalles no son aparentes.

El análisis de las propiedades de los algoritmos mediante diseños de experimentos y otras herramientas para contrastar la causalidad y revelar su estructura interna y los posibles sesgos deberá ser un campo importante de trabajo futuro.

Un impulso importante para nuevos métodos estadísticos va a venir de la abundancia de datos espaciales y temporales muy desagregados. La mayor parte de la estadística desarrollada hasta la segunda mitad del siglo XX ha sido para datos independientes, cuando los nuevos datos masivos van a tener, en general, dependencias temporales y espaciales. Métodos para trabajar de forma más efectiva con este tipo de datos, que en forma desagregada tienen estructura de tensores, va a concentrar muchos avances importantes en el futuro.

Referencias

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L., Griffanti, L., Douaud, G., ... and Smith, S. M. (2018). Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage*, 166, 400-424.

Alonso, A. M., and Peña, D. (2019). Clustering time series by linear dependency. *Statistics and Computing*, 29, 655-676.

Alonso, A. M., Galeano, P., and Peña, D. (2020). A robust procedure to build dynamic factor models with cluster structure. *Journal of Econometrics*, 216, 35-52.

Alonso, A. M., Peña, D., and Romo, J. (2003). Resampling time series using missing values techniques. *Annals of the Institute of Statistical Mathematics*, 55(4), 765-796.

Alonso, A. M., Peña, D., and Romo, J. (2004). Introducing model uncertainty in time series bootstrap. *Statistica Sinica*, 155-174.

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.

Amdaoud, M., Arcuri, G., and Levratto, N. (2021). Are regions equal in adversity? A spatial analysis of spread and dynamics of COVID-19 in Europe. *The European Journal of Health Economics*, 22, 629-642.

Anderson, T. W. (1996). R. A. Fisher and Multivariate Analysis, *Statistical Science*, 11, 20-34.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and Non-Gaussian clustering. *Biometrics*, 49, 803-821.

Basu, S., and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Annals of Statistics*, 43, 1535-1567.

Bates, J. M., and Granger, C. W. (1969). The combination of forecasts. *Journal of the Operational Research Society*, 20, 451-468.

Benito, M. and Peña, D. (2007). Detecting Defects with Image Data *Computational Statistics and Data Analysis*, 51, 6395-6403.

Benjamini, Y. (2010). Discovering the false discovery rate. *Journal of the Royal Statistical Society B*, 72, 405-416.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57, 289-300.

Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B*, 36, 192225.

Bickel, P. J., and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, 36, 2577-2604.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, 40, 318335.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71, 791-799.

Box, G. E. (1984). The importance of practice in the development of statistics. *Technometrics*, 26, 1-8.

Box, G. E. (2013). *An Accidental Statistician: The Life and Memories of George EP Box*. John Wiley & Sons.

Box, G. E., and Draper, N. R. (1969). *Evolutionary operation: A statistical method for process improvement*. Wiley, New York.

Box, G. E. P. and Jenkins, G. (1970). *Times series analysis. Forecasting and control*. Wiley, New York.

Box, G. E. P. and Tiao, G. C. (1968). A Bayesian approach to some outlier problems. *Biometrika*, 55, 119-129.

Box, G. E., and Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, 70, 70-79.

Box, G. E., and Tiao, G. C. (1977). A canonical analysis of multiple time series. *Biometrika*, 64, 355-365.

Box, G. E., and Wilson, K. B. (1951). On the experimental attain-

ment of optimum conditions. *Journal of the Royal Statistical Society B,* 13, 1-38.

Box, J. F. (1978). *R. A. Fisher: The life of a Scientist.* Wiley, New York.

Breiman, L. (1996a). Bagging predictors. *Machine Learning*, 26, 123140.

Breiman, L. (1996b). The heuristics of instability in model selection. *Annals of Statistics*, 24, 23502383.

Breiman, L. (1998). Arcing classifiers (with discussion). *Annals of Statistics*, 26, 801849.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16, 199-231.

Breiman L. (2001b). Random Forests. *Machine Learning*, 45, 532

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, CA.

Bühlmann, P. (1997). Sieve bootstrap for time series. *Bernoulli*, 123-148.

Bühlmann, P. and van de Geer, S. (2018). Statistics for big data: A perspective. *Statistics and Probability Letters*, 136, 37-41.

Bühlmann, P and van de Geer, S (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Berlin: Springer.

Burt, C. (1949). Alternative Methods of Factor Analysis and their Relations to Pearsons Method of Principle Axes. *British Journal of Psychology*, 2, 98121.

Caiado, J., Crato, N., and Peña, D. (2006). A periodogram-based metric for time series classification. *Computational Statistics and Data Analysis*, 50, 2668-2684.

Caiado, J., Crato, N., and Peña, D. (2009). Comparison of times series with unequal length in the frequency domain. *Communications in StatisticsSimulation and Computation*, 38, 527-540.

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis?. *Journal of the ACM*, 58, 1-37.

Carroll, J. B. and Schweiker, R. F. (1951). Factor Analysis in Educational Research. *Review of Educational Research* 21, 5, 368-388.

Casella, G., Moreno, E., and Girón, F. J. (2014). Cluster analysis,

model selection, and prior distributions on models. *Bayesian Analysis*, 9, 613-658.

Castillo, E. (2012). *Extreme value theory in engineering*. Elsevier.

Cochran, J. J. (2015). ASA leaders reminisce: Brad Efron. *AMSTAT news: the membership magazine of the American Statistical Association*, 459, 12-13.

Cohen, I. B. (1984). Florence Nightingale. *Scientific American,* 250, 128-137.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19, 15-18.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society B*, 34, 187-202.

Cressie, N. (1991). *Statistics for Spatial Data*. New York: Wiley.

Cui, L., Hu, H., Yu, S., Yan, Q., Ming, Z., Wen, Z., and Lu, N. (2018). DDSE: A novel evolutionary algorithm based on degree-descending search strategy for influence maximization in social networks. *Journal of Network and Computer Applications*, 103, 119-130.

Chang, I., Tiao, G. C., and Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics*, 30, 193-204.

Champking, J. (2010). Bradley Efron. *Significance*, December 178–181.

Chamberlain, G. (1983). Funds, factors, and diversification in arbitrage pricing models, *Econometrica*, 51, 1305–1323.

Chamberlain, G. and Rothschild, M.(1983). Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets, *Econometrica*, 51, 1281–1304.

Chen, R., Yang, D., and Zhang, C. H. (2022). Factor models for high-dimensional tensor time series. *Journal of the American Statistical Association*, in press.

David, F. N. (1962). *Games, gods and gambling: The origins and history of probability and statistical ideas from the earliest times to the Newtonian era*. Hafner Publishing Company.

Davies, O. L.(Ed.) (1947). *Statistical methods in research and production*. Oliver and Boyd, London.

DeGroot, M. H. (1986). A conversation with Charles Stein. *Statistical Science*, 1, 454-462.

DeGroot, M. H. (1987). A conversation with George Box. *Statistical Science*, 2, 239-258.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39, 1-22.

DerSimonian, R., and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, 7, 177-188.

Diaconis, P. and Holmes, S. (2017). Obituary: Charles M. Stein, 19202016. *Institute of Mathematical Statistics*.

Díaz, J. I. (2009). *Observación y Cálculo: los comienzos de la Real Academia de Ciencias y sus primeros correspondientes extranjeros*. Discurso inaugural del año académico 2009-2010. Real Academia de Ciencias Exactas, Físicas y Naturales de España

Diggle, P. J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns*. CRC press.

Donoho, D. (2006a). For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59, 797-829.

Donoho, D. (2006b). Compressed sensing. *IEEE Transactions on Information Theory*, 52, 1289–1306.

Donoho, D. (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, 26, 745-766.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society B* , 57, 45-70.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., and Vapnik, V. (1996). Support vector regression machines. *Advances in neural information processing systems*, 9.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*. 7, 1-26.

Efron, B. (2017). Charles Stein, 1920-2016. *Journal of the Royal Statistical Society A*, 923-925.

Efron, B., and Hastie, T. (2016). *Computer age statistical inference*. Cambridge University Press.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50, 987-1007.

Engle, R. F. and Granger, C. W. J. (1987), Cointegration and Error Correction: Representation, Estimation and Testing, *Econometrica*, 55, 251276.

Engle, R., and Watson, M. (1981). A one-factor multivariate time

series model of metropolitan wage rates. *Journal of the American Statistical Association*, 76, 774-781.

Engle, R. F., Hendry, D. F., and Richard, J. F. (1983). Exogeneity. *Econometrica*, 277-304.

Escribano, A., and Peña, D. (1994). Cointegration and common factors. *Journal of time series analysis*, 15, 577-586.

Fan, J., Han, F. and Liu, H. (2014). Challenges of big data analysis. *National Science Review*, 1, 293-314.

Ferraty, F., and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. New York: Springer.

Fienberg, S. E, (1992). A brief history of statistics in three and one-half chapters: A review essay. *Statistical Science*, Vol. 7, No. 2 : 208-225.

Findley, D. F., and Parzen, E. (1998). A conversation with Hirotugu Akaike. *Statistical Science,* 10, 104- 117.

Fisher, R. A. (1925a).*Statistical Methods for Research Workers*. Olyver and Boyd.

Fisher, R. A. (1925b). Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society* 22, 5, 700-725. Cambridge University Press.

Fisher, R. A. (1925c). Applications of Students Distribution, *Metron*, 5, 90104

Fisher, R. A. (19235). *The Design of Experiments*. Olyver and Boyd.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). *Applied longitudinal analysis*. John Wiley and Sons.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized dynamic factor model: Identification and estimation, *The Review of Economic and Statistics*, 82, 540–554.

Fox, A. J. (1972). Outliers in time series. *Journal of the Royal Statistical Society B*, 34, 350-363.

Freund, Y. and Schapire, R. (1996). Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference*, 148156.

Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.*, 55 119139.

Friedman, J. H., and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on com-*

puters, 100(9), 881-890.

Galeano, P., Peña, D. (2019). Data science, big data and statistics. *TEST*, 28, 289-329.

Galeano, P., Peña, D., and Tsay, R. S. (2006). Outlier detection in multivariate time series by projection pursuit. *Journal of the American Statistical Association*, 101, 654-669.

Garcia-Ferrer, A., Highfield, R. A., Palm, F., and Zellner, A. (1987). Macroeconomic forecasting using pooled international data. *Journal of Business and Economic Statistics*, 5, 53-67.

Gelfand, A. E., and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Gelman, A., and Vehtari, A. (2021). What are the most important statistical ideas of the past 50 years?. *Journal of the American Statistical Association*, 116, 2087-2097.

Geman, S., and Geman, D. (1984), Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721741.

Geweke, J. (1977). The dynamic factor analysis of economic time series. In *Latent variables in socio-economic models.* D. Aigner and A. Goldberger, eds. North Holland, Amsterdam, NL.

Girón, F. J. (1994). Historia del cálculo de probabilidades: de Pascal a Laplace. *Historia de la Ciencia Estadistica.* Real Acad. Cien. Exac Fis. y Nat. Madrid.

Glass, D. V. (1964). John Graunt and His Natural and Political Observations. *Notes and Records of the Royal Society of London*, 19,63-100.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21, 215-223.

Gómez-Villegas, M. A. and de Mora, M.S. (2018). *Historia de la Probabilidad y la Estadística.* UNED.

Graunt, J. (1622). *Natural and Political Observations Mentioned in a Following Index and Made up upon the Bills of Mortality.* London J. Martyn and J. Allestry.

Grenander, U.(1983). Tutorial in pattern theory. *Tech Report.* Division of Applied Mathematics, Brown University, Providence.

Guerrero, V. M., and Peña, D. (2003). Combining multiple time series predictors: a useful inferential procedure. *Journal of Statistical*

Planning and Inference, 116, 249-276.

Hacking, I. (1975). *The Emergence of Probability*. Cambridge University Press.

Hacking, I.. (1990). *The Taming of Chance*. Cambridge University Press.

Hald, A. (1998). *A History of Mathematical Statistics from 1750 to 1930*. New York: Wiley.

Hald, A. (2003). *A History of Probability and Statistics and Their Applications before 1750*. Hoboken, NJ: Wiley.

Härdle, W. (1990). *Applied nonparametric regression* (No. 19). Cambridge university press.

Hastie, T., Tibshirani, R. and Friedman, J. (2011). *The elements of statistical learning: data mining, inference, and prediction*, 2th edition. Springer.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances*. Jhu Press.

Hedges, L. V., and Olkin, I. (2014). *Statistical methods for meta-analysis*. Academic press.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14, 382-417.

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67.

Holmes, S., Morris, C., Tibshirani, R., and Efron, B. (2003). Bradley Efron: A conversation with good friends. *Statistical Science*, 268-281.

Hotelling, H.(1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, 24: 417441,498520.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, 73-101.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. 4th Berkeley Symposium on Mathematical Statistics and Probability, vol. I*, 361-379. University of California Press.

Jöreskog, K. G., Olsson, U. H., and Wallentin, F. Y. (2016). *Multivariate analysis with LISREL*. Basel, Switzerland: Springer.

Justel, A., and Peña, D. (1996). Gibbs sampling will fail in outlier

problems with strong masking. *Journal of Computational and Graphical Statistics*, 5(2), 176-189.

Justel, A., Peña, D., and Tsay, R. S. (2001). Detection of outlier patches in autoregressive time series. *Statistica Sinica*, 651-673.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 83, 95-108.

Kotz, S., and Johnson, N. L. (Eds.). (1992a). *Breakthroughs in Statistics: Foundations and basic theory.* Springer Science & Business Media.

Kotz, S., and Johnson, N. L. (Eds.). (1992b). *Breakthroughs in Statistics: Methodology and distribution.* Springer Science & Business Media.

Kotz, S., and Johnson, N. L. (Eds.). (1998). *Breakthroughs in Statistics, Volume III.* Springer Science & Business Media.

Krige, D.G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52, 119139.

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 1217-1241.

Kuzin, V., Marcellino, M., and Schumacher, C. (2011). MIDAS vs. mixed-frequency VAR: Nowcasting GDP in the euro area. *International Journal of Forecasting,* 27, 529-542.

Laird, N. M., and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 963-974.

Lawley, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh, Section A,* 60, 6482.

Lawley, D. N. ans Maxwell, A. E.(1963). *Factor analysis as a statistical method*. London: Butterworths.

Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99, 67-81.

Lindley, D. V., and Smith, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society B*, 34, 1-18.

Lipsey, R. G.(1971). *Introducción a la economía positiva*, Vincens-vives.

Liu, X., Ahmad, Z., Gemeay, A. M., Abdulrahman, A. T., Hafez, E. H., and Khalil, N. (2021). Modeling the survival times of the

COVID-19 patients with a new statistical model: A case study from China. *Plos one*, 16, https://doi.org/10.1371/journal.pone.0254999

López-Cheda, A., Jácome, M., Cao,R, and De Salazar, M. (2021). Estimating lengths-of-stay of hospitalized COVID-19 patients using a non-parametric model: a case study in Galicia (Spain), *Epidemiology and Infection*, 149 102, 1-8.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 51-67.

Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2019). *Robust statistics: theory and methods* (with R). John Wiley and Sons.

Martin, G. M., Frazier, D. T., and Robert, C. P. (2020). Computing Bayes: Bayesian computation from 1763 to the 21st century. arXiv preprint arXiv:2004.06425.

Matern, B. (1960). *Spatial variation*. Meddelanden fran Statens Skogsforsknings Institut, Stockholm. Band 49, No. 5.

Matheron, G. (1963). Principles of Geostatistics. *Economic Geology* 58, 124666.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*, 6, 355-378.

Merediz, A. (2004). *Historia de la estadistica oficial como institución pública en España*. Consejería de Educación y Hacienda, Junta de Andalucía.

Muler, N., Peña, D., and Yohai, V. J. (2009). Robust estimation for ARMA models. *The Annals of Statistics*, 37, 816-840.

Nelder, J. A., and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society A* 135, 370-384.

Nychka, D., Ma, P., and Bates, D. (2020). A conversation with Grace Wahba. *Statistical Science*, 35, 308-320.

Olshen, R. (2001). A conversaton with Leo Breiman. *Statistical Science*, 16, 184-198.

Peña, D. (1990). Influential observations in time series. *Journal of Business and Economic Statistics*, 8, 235-241.

Peña, D. (1991). Measuring influence in dynamic regression models. *Technometrics*, 33, 93-101.

Peña, D. (2001). George Box: an interview with the International Journal of Forecasting. *International Journal of Forecasting*, 17, 1-9.

Peña, D. (2005). A new statistic for influence in linear regression.

Technometrics, 47, 1-12.

Peña, D. (2022). Comment on Factor Models for High-Dimensional Tensor Time Series. *Journal of the American Statistical Association*, in press.

Peña, D. and Box, G. E. P. (1987). Identifying a simplifying structure in time series. *Journal of the American Statistical Association*, 82, 836-843.

Peña, D., and Galeano, P. (2008). A unified approach to model selection, discrimination, goodness of fit and outliers in time series. *In Advances in Mathematical and Statistical Modeling, in honor of Enrique Castillo*, 267-278. Birkhäuser Boston.

Peña, D. and Poncela, P. (2004). Forecasting with nonstationary dynamic factor models. *Journal of Econometrics*, 119, 291-321.

Peña, D., amd Poncela, P. (2006). Nonstationary dynamic factor analysis. *Journal of Statistical Planning and Inference*, 136, 1237-1257.

Peña D. and Prieto, F. J. (2001a). Cluster identification using projections. *Journal of the American Statistical Association*, 96, 1433-1445.

Peña, D. and Prieto, F. J. (2001b). Robust covariance matrix estimation and multivariate outlier detection. *Technometrics*, 43, 286-310.

Peña, D., and Prieto, F. J. (2007). Combining random and specific directions for outlier detection and robust estimation in high-dimensional multivariate data. *Journal of Computational and Graphical Statistics*, 16, 228-254.

Peña, D. and Sánchez, I. (2005). Multifold predictive validation in ARMAX time series models. *Journal of the American Statistical Association*, 100, 135-146.

Peña, D. and Tsay, R. S. (2020). *Statistical Learning with Big Dependent Data*. John Wiley & Sons.

Peña, D., and Yohai, V. J. (1995). The detection of influential subsets in linear regression by using an influence matrix. *Journal of the Royal Statistical Society B*, 57, 145-156.

Peña, D., and Yohai, V. (1999). A fast procedure for outlier diagnostics in large regression problems. *Journal of the American Statistical Association*, 94, 434-445.

Peña, D. and Yohai, V. J. (2016). Generalized dynamic principal components. *Journal of the American Statistical Association*, 111,

1121-1131.

Peña, D., Tiao, G. C. and Tsay, R. S. (2011). *A Course in Time Series Analysis*. John Wiley & Sons.

Peña, D., Smucler, E., and Yohai, V. J. (2019a). Forecasting multiple time series with one-sided dynamic principal components. *Journal of the American Statistical Association*, 114, 1683-1694.

Peña, D., Smucler, E., and Yohai, V. J. (2021b). Sparse One-Sided Dynamic Principal Components. *International Journal of Forecasting*, 37, 1498-1508.

Peña, D. Tsay, R. S, and Zamar, R. (2019b). Empirical Dynamic quantiles fortime series. *Technometrics*, 61, 429-444.

Plackett, R. L. (1972). Studies in the History of Probability and Statistics : The discovery of the method of least squares. *Biometrika*, 59, 239-251.

Quenouille,M. H. (1949). Problems in Plane Sampling. *Annals of Mathematical Statistics*, 20, 355375.

Ramsay, J. O., and Silverman, B. W. (2008). *Functional data analysis*. Springer.

Reinsel, G. C., and Tiao, G. C. (1987). Impact of chlorofluoromethanes on stratospheric ozone: A statistical analysis of ozone data for trends. *Journal of the American Statistical Association*, 82, 20-30.

Rossi, P. H., Wright, J. D., and Anderson, A. B. (1983). Sample surveys: History, current practice, and future prospects. in *Handbook of survey research*, 1-20.

Rousseeuw, P. J., and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association*, 85, 633-639.

Ryan, L. (2015). A conversation with Nan Laird. *Statistical Science*, 30, 582-596.

Ryan, T. P., and Woodall, W. H. (2005). The most-cited statistical papers. *Journal of Applied Statistics*, 32, 461-474.

Sánchez, M. J., and Peña, D. (2003). The identification of multiple outliers in ARIMA models. *Communications in Statistics-Theory and Methods*, 32, 1265-1287.

Sanz-Serna, J. M. (2018). *Un pequeño elogio de la ciencia pequeña.* Discurso inaugural del año académico 2018-2019. Real Academia de Ciencias Exactas, Físicas y Naturales de España

Satorra, A., and Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*,

66, 507-514.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics,* 6, 461-464.

Spearman, C. (1904). General Intelligence. *American Journal of Psychology*, 15, 201-292.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.

Stigler, S. M. (1996). The history of statistics in 1933. *Statistical Science*, 11, 244-252.

Stigler, S. M. (1990). *Statistics on the Table*. Harvard University Press.

Stigler, S. M. (1991). Stochastic simulation in the nineteenth century. *Statistical Science*, 89-97.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, 36, 111-147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society B*, 39, 44-47.

Student (1908). The probable error of a mean, *Biometrika*, 6, 1-25.

Sun, C., Shrivastava, A., Vondrick, C., Sukthankar, R., Murphy, K., & Schmid, C. (2019). Relational action forecasting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , 273-283.

Tanur, J. M., Mosteller, F., Kruskal, W. H., Lehmann, E. L., Link, R. F., Pieters, R. S., and Rising, G. R. (1989). *Statistics: A Guide to the Unknown*, Pacific Grove, CA: Wadsworth & Brooks.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B* , 267-288.

Todhunter, I. (1865). *History of the Mathematical Theory of Probability from the time of Pascal to that of Laplace*. Macmillan and Company.

Tong, H. (2010). Professor Hirotugu Akaike, 1927-2009. *Journal of the Royal Statistical Society A*, 173, 451-454.

Torrecilla, J. L. and Romo, J. (2018). Data learning from big data. *Statistics and Probability Letters*, 136, 15-19.

Tsay, R. S., Peña, D., and Pankratz, A. E. (2000). Outliers in multivariate time series. *Biometrika*, 87, 789-804.

Tukey, J. W. (1958). Bias and Confidence in Not Quite Large

Samples, *The Annals of Mathematical Statistics*, 29, 614.

Tukey J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics:Essays in Honor of Harold Hotelling* (I. Olkin et al., eds.) 448485. Stanford Univ. Press

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33, 1-67.

Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley.

Tzeng, J-Y., Byerley, W., Devlin, B., Roeder, K. and Wasserman, L. (2003). Outlier detection and false discovery rates for whole-genome DNA matching. *Journal of the American Statistical Association*, 98, 236-246.

Van Noorden, R., Maher, B., and Nuzzo, R. (2014). The top 100 papers. *Nature News*, 514(7524), 550.

Vallejo, J. A., Trigo, N. , Rumbo-Feal, S., y otros (2022). Highly predictive regression model of active cases of COVID-19 in a population by screening wastewater viral load, *Science of the Total Environment*, in press..

Vapnik, V. N. (1998). *Statistical Learning Theory.* Wiley-Interscience

Vapnik, V. N. (1999). *The Nature of Statistical Learning Theory.* Springer

Wahba, G. (1990). *Spline models for observational data.* Society for industrial and applied mathematics.

Wahba, G. (1999). Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. *Advances in Kernel Methods-Support Vector Learning*, 6, 69-87.

Wald, A. (1950). *Statistical decision functions.* Wiley.

Wang, D., Liu, X., and Chen, R. (2019). Factor models for matrix-valued high-dimensional time series. *Journal of Econometrics*, 208, 231-248.

Wang, Y., Ting, C. M., Gao, X., and Ombao, H. (2019). Exploratory analysis of brain signals through low dimensional embedding. *9th International IEEE/EMBS Conference on Neural Engineering (NER)* 997-1002. IEEE.

Wang, D., Zheng, Y., Lian, H., and Li, G. (2022). High-dimensional vector autoregressive time series modeling via tensor decomposition. *Journal of the American Statistical Association*, in press.

Wang, J. L., Chiou, J. M., and Müller, H. G. (2016). Functional

data analysis. *Annual Review of Statistics and Its Applications*, 3, 257-295.

Wasserstein, R. L., and Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, 70, 129-133.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10, 515-534.

Yates, F. (1966). Computers, the second revolution in Statistics.*Biometrics*, 22, 233-251.

CONTESTACIÓN
DEL
EXCMO. SR. D. FRANCISCO JAVIER
GIRÓN GONZÁLEZ-TORRE

Excmo. Sr. Presidente,
Excmas. Sras. Académicas,
Excmos. Srs. Académicos,
Señoras y Señores

Quisiera agradecer a la Presidencia de la Academias el haberme propuesto para contestar el discurso de ingreso del Prof. Daniel Peña, lo que me produce por una parte enorme satisfacción y por otra una gran responsabilidad a la hora de glosar sus méritos y contribuciones a la Ciencia Estadística a lo largo de su dilatada carrera, cuyo comienzo podemos situar hacia la mitad de los años 70 del pasado siglo y continúa en la actualidad con gran vigor ya que, en estos momentos, tiene entre sus manos la publicación de un libro, varios artículos con sus numerosos discípulos y colaboradores y la dirección de cuatro tesis doctorales. Ni siquiera sus dos períodos como rector de la Universidad Carlos III de Madrid apenas mermaron su dedicación a la Estadística.

Conocí al Prof. Peña en el año 1975, cuando era profesor de la Escuela de Organización Industrial, situada cerca de la Facultad de Ciencias, tras coincidir con él en una conferencia dictada en alguno de los seminarios del Departamento de Estadística e Investigación Operativa de la Universidad Complutense de Madrid, entonces dirigido por el Prof. Sixto Ríos a quien quiero recordar esta tarde ya que fue él quien contestó a mi discurso de ingreso en esta Academia. Desde ese momento hemos tenido una relación personal que ha durado hasta nuestros días: hemos coincidido en muchos Congresos de Estadística; en particular, los *Valencia Meeting on Bayesian Statistics* y los de la *Sociedad de Estadística e Investigación Operativa*. También compartimos cargos de responsabilidad en la dirección de la SEIO durante un mandato; él como Presidente y yo como Vicepresidente. Curiosamente, desde el punto de vista profesional solamente hemos tenido una ocasión de trabajar juntos para un ambicioso proyecto originado en nuestra Academia, del que más adelante daré debida cuenta.

Quiero comentar, muy brevemente, como ahora nos sugieren y recomiendan los nuevos estatutos, el discurso de ingreso del Prof. Peña quién comienza con un breve repaso histórico de la ciencia estadística hasta nuestros días, señalando algunas de las ideas más importantes aparecidas en esta disciplina y que supusieron notables cambios de paradigma así como la introducción de nuevas técnicas estadísticas.

Entre ellas hay que señalar las debidas a la irrupción de los ordenadores en la estadística, que han permitido el uso de métodos de simulación o MonteCarlo como son el *bootstrap* de Efron en la estadística frecuentista, y los métodos de MonteCarlo basados en cadenas de Markov en la estadística bayesiana.

Finaliza su discurso comentando cómo muchas de las ciencias sociales —la Economía, la Psicología, y otras como la Medicina, la Ingeniería o el estudio de los problemas Medio Ambientales— han impulsado la aparición o el auge de nuevas técnicas estadísticas. Esta interacción o simbiosis entre teoría y aplicaciones se da con mayor incidencia en la estadística al tratarse de un ciencia que se nutre sobre todo de las aplicaciones, como también ocurre en algunas ramas de la matemática aplicada. Muchas veces, la necesidad de resolver nuevos problemas, que anteriormente no se habían planteado, ha dado origen a nuevos métodos estadísticos que hubiesen sido inimaginables en su momento, debido a que en la actualidad los datos ya no son simplemente escasos y numéricos sino que pueden ser masivos como las imágenes, vídeos, audios, sonidos, etc.

Paso, a continuación, a detallar su trayectoria académica y científica.

LAUDATIO

Comienzo con una breve descripción de sus principales méritos que desglosaré a continuación, con especial atención a sus contribuciones científicas.

Daniel Peña es Doctor Ingeniero Industrial por la Universidad Politécnica de Madrid y Diplomado en Sociología y en Estadística por la Universidad Complutense de Madrid .

Actualmente es Director del Instituto UC3M-BS en *Financial Big Data* y Catedrático emérito del Departamento de Estadística de la Universidad Carlos III de Madrid. Ha sido profesor titular y posteriormente catedrático de la Escuela Técnica Superior de Ingenieros Industriales de la Universidad Politécnica de Madrid y director del Departamento de Estadística de la misma. A continuación fue director y fundador del Departamento de Economía de la Universidad Carlos III de Madrid y, poco después, director del Departamento de

Estadística e Investigación Operativa de la misma Universidad y, posteriormente, fue elegido Rector de esta Universidad por dos mandatos consecutivos, desde abril de 2007 hasta abril de 2015.

Anteriormente, fue director del Laboratorio de Estadística de la Universidad Politécnica de Madrid y Visiting Full Professor en las Universidades de Wisconsin–Madison, Chicago, y British Columbia.

Es miembro de las siguientes instituciones nacionales e internacionales: The Institute of Mathematical Statistics, The International Statistical Institute, The American Statistical Association, The Royal Statistical Society, The Bernoulli Society of Mathematical Statistics, la Sociedad Española de Estadística e Investigación Operativa de la que fue presidente, el Instituto Interamericano de Estadística, la Asociación Española para la Calidad, The American Society for Quality Control y The International Society for Bayesian Analysis.

Hasta la fecha ha dirigido 31 tesis doctorales, más otras cuatro en curso, publicado dieciocho libros y 232 artículos de investigación sobre Estadística y sus aplicaciones. Su índice h, según Google Scholar, es 45 y el número de citas de sus artículos se eleva a unas 9 500. Tiene seis sexenios de investigación.

Ha dirigido proyectos del Plan Nacional de Investigación como Investigador principal ininterrumpidamente desde 1982 hasta la actualidad, de la Fundación BBVA, y ha participado en redes y proyectos europeos como el Economic Applications of Time Series (1995-1998), el Factor Models (1998-2001), y ha sido Director del grupo español en la red europea SACD (Statistical Analysis of Complex Data) financiado por la European Science Foundation.

Su trayectoria internacional ha sido reconocida como: profesor invitado en las Universidad de Chicago, British Columbia, Toronto y Madison-Wisconsin. Ha sido miembro de honor de las asociaciones principales de Estadística (Fellow of the American Statistical Association y del Institute of Mathematical Statistics). Como conferenciante invitado ha participado en más de 70 congresos de Estadística, entre ellos los más importantes de la profesión, como los Joint Statistical Meetings (ASA, IMS, etc.) en EEUU, the World Congress of the Bernoulli Society, the European Meeting of Statisticians, the Congress of the International Statistical Institute.

En el ámbito de las series temporales, ha sido invitado frecuente en los congresos más prestigiosos sobre series temporales y predicción como, por ejemplo, fue Keynote Speaker en el International Sympo-

sium on Forecasting celebrado en Lisboa en el año 2000, y seis veces conferenciante invitado de la NBER/NSF Time Series Conference.

Además ha impartido seminarios de investigación en más de 80 universidades, que incluyen varios de los mejores departamentos de Estadística del mundo, como Stanford, Chicago, Wisconsin, Minnesota, Carnegie-Mellon, London School of Economics, Lancaster, Liverpool, Lovaina, Cambridge, Hong-Kong, British–Columbia y Toronto.

Su investigación se ha centrado en los modelos de series temporales, los métodos robustos, el análisis multivariante, los modelos bayesianos, y las aplicaciones de la Estadística a la mejora de la economía y la administración de empresas, la historia, la antropología, la medicina, la ingeniería y el medio ambiente, que más adelante se comentarán con más detalle.

Como resultado de sus investigaciones se han creado nuevos métodos estadísticos que han demostrado ser importantes en sus aplicaciones a distintos ámbitos de la Ciencia. En especial, como indicaba la American Statistical Association en su nombramiento como Fellow, es uno de los investigadores más reconocidos del mundo en el modelado y predicción de conjuntos de series temporales: "for his outstanding and pathbreaking research contributions to time series analysis".

Entre sus muchas contribuciones a las series temporales, el método que propuso para medir la influencia de cada dato en la predicción se cita como el método estándar en muchos libros de texto en varios idiomas. El modelo que desarrolló con George Box, the Peña-Box model or Exact Dynamic Factor Model, ha sido extendido y generalizado por muchos autores y, por su importancia en Economía, se le concedió el Premio Jaime I de investigación en esa área.

Su interés investigador ha cubierto un amplio espectro de temas con contribuciones destacadas en muchas áreas de la Estadística, como los métodos bayesianos, la estimación robusta de modelos, el análisis multivariante, la econometría y los métodos estadísticos para la calidad, con contribuciones que se han publicado en las mejores revistas de estos campos. En concreto, el enfoque que propuso para medir la sensibilidad de los datos en un modelo de regresión fue reconocido con el 2006 Youden Prize al mejor trabajo publicado en Technometrics, concedido por The American Statistical Association y The American Society for Quality.

Además de su intensa labor investigadora ha impulsado la investigación estadística y la colaboración internacional impartiendo cursos

y seminarios en más de 40 universidades españolas y más de 65 en todo el mundo; dirigiendo asociaciones científicas españolas (Presidente de la Sociedad Española de Estadística e Investigación Operativa, SEIO), europeas (Presidente of European Courses on Advanced Statistics, ECAS) e internacionales (Vicepresidente del Instituto Interamericano de Estadística, IIE), creando nuevos centros de investigación estadística como el Departamento de Estadística y el Instituto de Big Data de la Universidad Carlos III de Madrid y desarrollando una intensa labor editorial como Director de Estadística Española y Editor Asociado de numerosas revistas de Estadística de alto impacto. Ha contribuido a transformar la enseñanza de la Estadística en España y Latinoamérica escribiendo libros de texto ampliamente utilizados y dirigiendo la tesis doctoral de 31 profesores que imparten actualmente docencia en 8 universidades españolas, 5 europeas y 8 latinoamericanas. Finalmente, ha trabajado para mejorar los métodos docentes y la administración universitaria mediante herramientas estadísticas de calidad, recibiendo varios premios por estas actuaciones, tanto de organismos públicos (Comunidad de Madrid), como privados (Asociación Española para la Calidad).

A continuación, quisiera señalar la relación que ha tenido el Prof. Peña con nuestra Academia. Ha sido invitado a impartir, al menos que yo recuerde, dos conferencias plenarias en nuestra casa, fue elegido Académico Correspondiente en 2018 y participó como coordinador, junto con el Prof. Toni Espasa y con quien les está hablando, en un proyecto pionero dirigido por nuestro recordado presidente D. Ángel Martín Municio sobre *Econometría de la lengua española* que culminó, poco después de su fallecimiento, con la publicación en el año 2003 de la primera monografía dedicada a estudiar *El valor económico de la lengua española*, patrocinado por el *Instituto Cervantes*, la *Fundación del Banco Santander Central Hispano* y nuestra *Academia*.

La metodología empleada se basó en el examen de las cuentas nacionales, en concreto en las tablas de origen destino, con el fin de extraer de estas tablas la parte del VAB y PIB que se puede asignar a la lengua cuando se tiene información completa, y usar métodos econométricos cuando solamente se tiene información parcial o se hacen predicciones. Como resultado de este trabajo se estimó la aportación de la lengua española al PIB en un 14 % en el período 1995-1997, con tendencia creciente. Una actualización del estudio anterior, que incluye la aportación de la lengua al empleo puede verse en Girón

& Cañada (2009).[1]

Sorprendentemente, en una conferencia organizada por la Sección de Matemáticas con fecha 5 de junio de 2019, el Prof. Tomás Chacón de la Universidad de Sevilla presentó el Estudio del impacto socioeconómico de las Matemáticas en España, elaborado por la empresa AFI, por encargo de la Red Estratégica en Matemáticas. En el informe sobre la aportación de las matemáticas al PIB y al empleo en varios países europeos se utilizó la misma metodología que en el estudio sobre la aportación económica de la lengua española. En concreto se mostraba que alrededor del 11 % del PIB de España se puede atribuir a las matemáticas, algo menos que en otros países europeos pero, en cualquier caso, una cifra notable. Un estudio similar podría llevarse acabo en las otras disciplinas de nuestra Academia.

Tras esta enumeración de su trayectoria científica, me he permitido analizar someramente los diversos temas tratados por el Prof. Peña y su evolución en el tiempo. Una constante a lo largo de este período son las contribuciones, tanto teóricas como aplicadas, como ya se ha apuntado anteriormente, de las series temporales; en segundo lugar aparecen los temas referentes a las observaciones atípicas o anómalas. Poco a poco, van apareciendo temas como el análisis de conglomerados o *cluster analysis*, los modelos factoriales.y más recientemente el análisis de imágenes, el aprendizaje automático y el aprendizaje profundo o *Deep learning*, este último basado en redes neuronales, que revelan su interés por temas más novedosos relacionados con el Big Data, la Analítica o Ciencia de Datos, e incluso, la Inteligencia Artificial.

En relación con los últimos temas quiero destacar dos libros recientes, todavía en prensa, y editados por FUNCAS —un centro de análisis o *think tank* dedicado a la investigación económica y social y a su divulgación—, que incluyen valiosas aportaciones del Prof. Peña:

- *Nuevos métodos de predicción económica con datos masivos.*
- *Análisis econométrico con Big Data.*

Sobre sus publicaciones, quisiera resaltar el elevado número de libros de texto en castellano, que han sido de gran utilidad a más de una generación de matemáticos, estadísticos, ingenieros, y estudiantes

[1] Girón, F J. & Cañada, A. (2009). Las cuentas del español. *Fundación Telefónica.* Ariel: España.

de ciencias sociales. Entre ellos destacaría los dedicados a las *Series Temporales* y al *Análisis de datos multivariantes*. Su última publicación, en la prestigiosa editorial Wiley, en colaboración con Ruey S. Tsaya, que está a punto de aparecer, versa sobre *Statistical Learning with Big Dependent Data.*

Sus primeros artículos, que datan de 1976, versan sobre dos temas importantes que estaban en efervescencia en ese momento como eran la teoría de la decisión y la inferencia bayesiana y la metodología de Box-Jenkins para el modelado de las series temporales publicada en un artículo pionero de 1973, con aplicaciones a problemas reales. A partir de ese momento, el tema de las series temporales y otros relacionados con ellas, como hemos comentado con anterioridad, constituyen una parte importante de sus aportaciones, tanto teóricas como aplicadas.

Otro de los temas en que ha destacado la investigación del Prof. Peña es al estudio de la robustez de los modelos estadísticos, con particular énfasis en los modelos lineales, los modelos lineales generalizados y las series temporales uni y multidimensionales, y a la detección de observaciones anómalas o atípicas, donde sus trabajos han tenido un reconocimiento unánime.

A estos resultados, que han sido una constante en su dilatada carrera, hay que añadir, más recientemente, su interés por nuevas áreas de la estadística como son la selección de modelos y de variables en regresión, los datos funcionales, los modelos factoriales dinámicos, el análisis estadístico de imágenes, el análisis de conglomerados o —como habitualmente se le conoce— *cluster analysis* y, por último, en problemas con datos masivos o *Big Data* como se describen en el título de su discurso.

Dando un paso más, si examinamos cuidadosamente su Curriculum Vitae, un simple análisis estadístico de su producción científica, excluyendo los libros de texto, revela un número elevado de temas de investigación que van desde la estadística bayesiana hasta temas muy actuales como el aprendizaje automático y el Big data. Sin embargo, hay al menos cuatro temas importantes en los que se ha centrado una buena parte de su investigación que son: las series temporales (80 artículos del total de 232); en segundo lugar, el importante tema de la robustez y detección de observaciones atípicas o anómalas y el desarrollo de medidas de influencia en los modelos estadísticos (49 del total); en tercer lugar, los artículos de índole bayesiana (32 del total),

y en cuarto lugar, los modelos lineales generalizados y en particular, los modelos de regresión (24 del total). Cabe señalar que los cuatro temas mencionados no son disjuntos en absoluto; de hecho, muchos artículos tratan de aplicar resultados de algunos temas a otros como, por ejemplo, medir la influencia de las observaciones atípicas en series temporales y en los modelos de regresión, tanto desde la perspectiva frecuentista como desde la bayesiana.

Para aquellos que no están dentro del mundo, o al corriente, de la estadística les puede parecer extraño el que haya, básicamente, dos enfoques muy diferentes a los llamados métodos estadísticos. Si nos fijamos, a lo largo del discurso de ingreso y de mi contestación ya han aparecido referencias a los enfoques frecuentista y bayesiano a la estadística, a veces en concordancia, como suele ocurrir en la estimación de parámetros, y otras veces en discrepancia, como son los problemas de contraste de hipótesis simples y múltiples, los de selección de variables y los más generales de selección de modelos. Esto diferencia a la estadística del resto de las matemáticas, ya que sus fundamentos teóricos son distintos según el enfoque que se quiera adoptar: las distribuciones en el muestreo para el enfoque frecuentista y la teoría de la probabilidad para el segundo. En este sentido, la estadística no es una ciencia exacta estrictamente hablando, lo cual es de esperar pues trata con un concepto elusivo, y difícil de definir y cuantificar, como es la *incertidumbre*, aunque su desarrollo depende de varias ramas de las matemáticas y, en concreto y en mayor medida, del cálculo de probabilidades.

En relación con lo que acabo de exponer me van a permitir que haga un breve inciso en este momento.

En 1812, Laplace en su *Teoría Analítica de las Probabilidades* expone los principios y las aplicaciones de lo que él llama *geometría del azar*[2]. Esta obra representa la introducción de los recursos del análisis matemático, que él mismo había desarrollado, en el estudio de los fenómenos aleatorios, y recopila además toda una serie de memorias publicadas desde 1771, algunas de las cuales no tenían relación con las probabilidades.

Laplace expresa de forma sencilla el significado del cálculo de probabilidades como sigue: "En el fondo, la teoría de probabilidades

[2]En esa época a las matemáticas se las conocía con el nombre genérico de *geometría* y a los matemáticos como *geómetras*.

es sólo sentido común expresado con números".

La importancia de esta materia la resalta Laplace con las siguientes palabras: "Es notable que una ciencia que comenzó con las consideraciones sobre los juegos de azar había de llegar a ser el objeto más importante del conocimiento humano. Las cuestiones más importantes de la vida constituyen en su mayor parte, en realidad, solamente problemas de probabilidad".

Es sorprendente que después de Laplace, el interés por esta materia fue disminuyendo hasta prácticamente desaparecer como disciplina matemática durante el siglo XIX.

Sin embargo, su comentario se puede considerar profético, ya que hoy día no se concibe el progreso en ninguna ciencia, ni en cualquier actividad humana, sin la presencia de la probabilidad.

De hecho, la fórmula o teorema de Bayes, tal como hoy lo conocemos y se explica en los cursos sobre teoría de la probabilidad, se debe a Laplace. Tanto Laplace como Gauss usaron la inferencia bayesiana para resolver problemas reales. De hecho Gauss y Laplace descubrieron, por caminos muy distintos y casi al mismo tiempo, la distribución *normal* que es conocida universalmente como *distribución gaussiana* o distribución de *Gauss-Laplace* en el mundo francófono. Laplace lo hizo como resultado del teorema central del límite y Gauss a resultas de imponer principios básicos en la distribución que siguen los errores de observación de modo que justificasen como estimador bayesiano el estimador de mínimos cuadrados.

Reanudo mi discurso comentando las aportaciones del Prof. Peña como ingeniero y estadístico a la amplia área que abarca el término *calidad*, que afecta a muchos ámbitos, no solamente a la industria y las fábricas en sus procesos industriales, sino a instituciones como puedan ser las universidades, las titulaciones universitarias, la educación, los servicios informáticos y un largo etcétera. En total, los artículos en esta área suman 22 aportaciones. La Tabla 1 resume sus contribuciones a las diferentes áreas mayoritarias.

De todo lo anterior podríamos definir a Daniel Peña como un estadístico ecléctico. En efecto, en su extensa producción científica aparecen tanto artículos bayesianos como artículos frecuentistas. No está adscrito a ninguna de las dos maneras de ver la estadística sino que utiliza las mejores y más apropiadas herramientas que proporcionan ambos enfoques para cada tipo de problemas estadísticos.

Series temporales	80
Robustez y observaciones atípicas	49
Estadística bayesiana	32
Modelos lineales	24
Control de calidad	22
Miscelánea	25

Tabla 1: Resumen de la producción científica, desglosada según temas de investigación.

En su dilatada carrera, Daniel Peña ha recibido muchos e importantes premios y honores como reconocimiento de su actividad científica y aportaciones relevantes a la Ciencia Estadística. Aunque en mi discurso ya he comentado alguno de los premios o distinciones que ha recibido, paso a continuación a señalar los más relevantes en orden inverso en el tiempo:

- Premio Nacional de Estadística 2020. Premiado en la primera convocatoria y concedido por el Gobierno de España a través del INE.
- Miembro de la Real Academia de Ciencias Exactas, Físicas y Naturales de España como Académico Correspondiente Nacional. 2018.
- Medalla de Honor de la Universidad Carlos III de Madrid, 2015.
- Medalla de la Sociedad Española de Estadística e Investigación Operativa. 2014.
- Premio Rey Jaime I de Economía. 2011.
- Premio Ingeniero del año, del Colegio Oficial de Ingenieros Industriales de Madrid. 2011.
- Miembro de honor de la American Statistical Association.
- Jack Youden Prize 2005 concedido por The American Society for Quality and The American Statistical Association for the best article published in Technometrics this year for the article: *A new statistic for influence in regression.*

- Premio de Investigación de la Universidad Carlos III de Madrid. Bienios 2003–5 y 2005–07.
- Miembro de honor del Institute of Mathematical Statistics.
- Miembro del International Statistical Institute, 1985.
- Premio de la Comunidad de Madrid a la Excelencia y Calidad del Servicio público 2000 por el proyecto: Mejora de la calidad mediante la simplificación de procesos en la Universidad Carlos III de Madrid.

Creo y estoy seguro de que tu incorporación a la que ya es tu casa, va a ser muy útil y beneficiosa dado tu interés y conocimiento de las nuevas tecnologías del análisis de datos masivos que ya están cambiando nuestra sociedad.

Finalizo mi discurso, dándote la bienvenida a nuestra Academia en nombre de todos mis compañeros.

Muchas gracias por su atención.

COMENTARIOS AL HILO DEL DISCURSO

En este apartado de mi discurso de contestación hago unas breves reflexiones sobre algunos de los aspectos más llamativos —y que considero importantes— del discurso de ingreso.

- **Sobre la predicción**

Una de las tareas más importantes de los métodos estadísticos actuales, como se enfatiza en el discurso, es la de predecir el comportamiento futuro de observaciones basado en los datos presentes. En los últimos tiempos se ha cambiado el paradigma de *estimar* los parámetros de un modelo por el de *predecir* el comportamiento futuro de los datos generados por el modelo. La estadística bayesiana, por su propia estructura probabilística basada en la idea importantísima

de la *intercambiabilidad*, está mejor diseñada, a priori, que la clásica para hacer predicciones. En efecto, así como la estadística clásica está más enfocada a las predicciones puntuales, la bayesiana se basa en la llamada *distribución predictiva*, que es una distribución de probabilidad sobre los acontecimientos futuros que permite, entre otras cosas, medir la incertidumbre asociada a la predicción. En los casos más interesantes de datos dependientes, también se puede calcular la distribución predictiva, aunque su cálculo pueda ser más complejo y haya que recurrir a los métodos de Monte Carlo basados en cadenas de Markov. Cómo ejemplo simple de la diferencia entre los dos enfoques tenemos el siguiente: cuando se dice, p. ej., que el IPC del mes próximo va a ser un 0.5 %, lo más probable es que la predicción —en este caso, la estimación puntual de lo que va a ocurrir— no sea exactamente esa, aunque se supone será parecida. Sin embargo, es mucho más sensato e informativo decir que el IPC futuro no es un número sino una cantidad aleatoria que sigue una determinada distribución, llamada predictiva, centrada alrededor de 0.5 % de la que se puede obtener su incertidumbre, muchas veces medida por la anchura de un cierto intervalo de credibilidad, que se deduce de esa distribución.

- **Sobre la robustez de los modelos y las observaciones atípicas**

Uno de los temas en los que más ha trabajado el Prof. Peña es el de *robustez*, un concepto, en principio elusivo, ya que la robustez hay que entenderla como condicionada o referida a un modelo base, generalmente basado, directa o indirectamente, en la distribución normal. Es bien sabido que los modelos normales son muy sensibles a la aparición de datos atípicos o anómalos (siempre entendidos respecto del modelo base), que son datos no esperados por el modelo pero que influyen no solamente en la estimación y en los contrastes de hipótesis de esos parámetros sino también en las predicciones, por lo cual este tema ha generado mucho interés y dentro del ámbito estadístico se han propuesto diferentes enfoques, como los M-estimadores, los estimadores recortados (*trimmed)*, y un largo etcétera.

Sin embargo, otra manera de atacar el problema de la robustez es el considerar modelos *robustos* en lugar de modelos basados en la distribución normal. La idea es simple: la distribución normal tiene colas que decrecen de forma exponencial con lo cual la presencia de

observaciones atípicas es, valga la redundancia, algo anómalo, mientras que los modelos robustos se basan en distribuciones con colas más pesadas, con lo cual los posibles datos atípicos tiene menos peso o influencia en el análisis estadístico. El problema con este enfoque es encontrar el modelo más robusto dentro de alguna familia, paramétrica o no paramétrica, que mejor se ajuste a los datos, lo que implica tener que resolver un problema de selección de modelos. Ambos enfoques son complementarios, aunque el primero permite hacer un reconocimiento de las observaciones atípicas y medir su influencia en el modelo base, cuando se supone que este es cierto para la mayoría de las observaciones.

- **Sobre la selección de modelos y variables**

El tema de selección de modelos y de variables en modelos multiparamétricos es uno de los más importantes de la estadística actual para el que se han propuesto muy diversas soluciones como se comenta en el discurso.

Debemos señalar la diferencia entre la selección de modelos y la selección de variables. El primero es un problema más general que el segundo pues permite comparar modelos de muy distinta índole, mientras que la selección de variables se aplica a modelos más estructurados, como los modelos lineales generalizados, que dependen de diversas covariables que, en principio, se supone sirven para explicar los datos. La selección de variables consiste en buscar el *mejor modelo* explicativo, es decir, en eliminar aquellas covariables que son irrelevantes para explicar los datos y que solo sirven para introducir *ruido*. En general, se trata de encontrar el modelo más económico, en el sentido de ser el más pequeño, tal que su valor predictivo sea alto.

El elevado número de procedimientos frecuentistas y bayesianos que hay en la actualidad para atacar ambos problemas se describe en el discurso. La selección de modelos desde el punto de vista de la inferencia bayesiana se basa en el concepto de factor de Bayes propuesto por Harold Jeffreys en su libro[3], que ha tenido gran influencia en la aceptación y desarrollo de la estadística bayesiana.

Como es un problema en el que he trabajado desde la perspectiva bayesiana, quisiera comentar mi modesta contribución al problema

[3] Jeffreys, H. (2000). *3rd Edition. The Theory of Probability.* Oxford Classical Texts in Physical Sciences: OUP Oxford.

de selección de variables en modelos de regresión normal. En Girón (2021)[4] se presenta un procedimiento sencillo, basado en factores de Bayes y modelos jerárquicos para comparar modelos anidados y una extensión al problema de comparar modelos no anidados como es el de selección de variables. El factor de Bayes que se detalla tiene buenas propiedades asintóticas, como es la consistencia bajo la hipótesis nula y las alternativas, y además tiene un buen comportamiento para muestras pequeñas y medianas, desde el punto de vista de su evaluación frecuentista, lo que resulta en un balance adecuado entre los errores de tipo I y II.

La extensión del factor de Bayes anterior a la comparación de modelos paramétricos arbitrarios, resulta ser una extensión del BIC que permite obtener un nuevo criterio que, al ser una generalización, podemos denominar $GBIC$. A continuación adjunto sus expresiones para facilitar la comparación entre ambos.

Si M_k es un modelo estadístico dependiente de un parámetro o vector $\boldsymbol{\theta}$ de dimensión k, $\mathbf{y} = (y_1, \ldots, y_n)$ es una muestra aleatoria de tamaño n del modelo, y $\tilde{\boldsymbol{\theta}}$ es el estimador máximo verosímil de $\boldsymbol{\theta}$, el BIC se define como

$$BIC(M_k) = k \log n - 2 \log L(\tilde{\boldsymbol{\theta}}\,; \mathbf{y}),$$

y el $GBIC$ como

$$GBIC(M_k) = k \log \frac{n}{k} - 2 \log L(\tilde{\boldsymbol{\theta}}\,; \mathbf{y}),$$

donde $L(\boldsymbol{\theta}\,; \mathbf{y})$ es la función de verosimilitud.

Para comparar modelos con parámetros unidimensionales el $GBIC$ es idéntico al BIC, pero es bien sabido que el BIC no funciona adecuadamente como selector de modelos cuando la dimensión del vector de parámetros es grande. Sin embargo, el nuevo criterio tiene más en cuenta la dimensión de los modelos que se comparan.

Desde el punto de vista bayesiamo, la selección de modelos y la de variables termina con un hiper-modelo o meta-modelo que es simplemente una distribución de probabilidad sobre modelos individuales, donde los pesos de los modelos son sus probabilidades a posteriori. De este modo, la selección de modelos o variables coincide con la idea de *model average* o promedio de modelos.

[4] Girón, F, J. (2021). *Bayesian Testing of Statistical Hypotheses*. Arguval: Málaga.

El promedio de modelos se presenta cuando el modelo que describe al conjunto de datos es complejo y puede haber varios modelos competitivos, por lo que una idea importante es construir un modelo a partir de estos, que se realiza construyendo un meta-modelo mezclando los modelos —por ejemplo, los que se obtienen aplicando técnicas de selección de modelos— y ponderándolos con sus probabilidades a posteriori.

La importancia de esta idea es que la predicción usando un meta-modelo se realiza con una distribución predictiva que es una combinación lineal o mixtura de las distribuciones predictivas de cada modelo individual, no solamente del más probable, lo cual produce predicciones más acuradas o precisas.

Como afirma el Prof. Peña en la Subsección 4.4, el usar p-valores para eliminar variables en el problema de selección de variables no funciona adecuadamente, en general, y menos en el caso de datos masivos, lo que suscribo totalmente.

▪ Sobre el concepto de intercambiabilidad o simetría

Una consecuencia importante de la paradoja de Stein en la estadística bayesiana, como señala el Prof. Peña en su discurso, es el de la extensión del concepto de *intercambiabilidad* o *simetria*, debido originalmente a De Finetti, que es clave en el desarrollo de métodos bayesianos simples para crear modelos más generales y refinados para tratar, entre otros, el problema de la *sobredispersión*, que es la presencia de más variabilidad en los datos que lo que sería de esperar del modelo estadístico elegido, y que, desde el punto de vista frecuentista es complicado de tratar.

Muchos de los tests de homogeneidad de la estadística —como la comparación de la igualdad de los parámetros de dos o más poblaciones independientes— pueden reformularse como tests de homogeneidad utilizando la idea de intercambiabilidad (véase Girón, 2021) y, de ese modo, verlos como modelos jerárquicos que permitxen más flexibilidad a la hora de modelarlos aún a costa de un coste computacional más elevado, lo que en la actualidad no es ningún impedimento, debido al enorme desarrollo del cálculo numérico y de los métodos de Monte Carlo.

El meta-análisis de ensayos clínicos de diversos centros que usan el mismo protocolo pero que pueden diferir entre si debido a otros

factores externos no controlables se modela y se adapta, de modo más natural, a la filosofía subyacente en la estadística bayesiana ya que el concepto de intercambiabilidad está asociado a la idea de homogeneidad de los datos de los diferentes centros.

▪ Sobre las aplicaciones a la medicina

Es interesante saber que, como nos recuerda el Prof. Peña, dentro del ámbito de la estadística aplicada, nueve de cada diez de las publicaciones más referenciadas están relacionadas con la medicina; en particular, el modelo de regresión logística, el meta-análisis y el modelo de supervivencia de Cox son tres de los métodos estadísticos más utilizados.

Si nos fijamos en algo de tanta importancia como son los ensayos clínicos en la medicina, estos siempre utilizan métodos estadísticos generalmente basados en dividir la cohorte bajo estudio en dos o tres partes para inferir la eficacia de un nuevo tratamiento: una a la que se le aplica el tratamiento base, otra a la que se le aplica el nuevo tratamiento y otra a la que se le suministra un placebo. Últimamente también se observa una tendencia a incluir en las aplicaciones médicas métodos bayesianos.

Otro aspecto importante, dentro del ámbito de la medicina que involucra a la economía es el de los estudios de coste-efectividad de tratamientos clínicos. Suele ocurrir que los tratamientos más efectivos son los más caros por lo que se hace necesario diseñar nuevos métodos para combinar ambos aspectos de un modo óptimo. El problema se complica muchas veces por la existencia de subgrupos que responden a los tratamientos de forma distinta. En mis aportaciones a este tema hemos introducido una novedad importante y es el de considerar la distribución predictiva de las combinaciones de los costes y la efectividad de los datos proporcionados por un ensayo clínico como herramienta fundamental y, de paso, considerar el análisis de subgrupos como un problema de selección de variables.

▪ Colofón

En las conclusiones de su discurso de ingreso, el Prof. Peña refleja las enormes posibilidades de la Ciencia de Datos para afrontar los retos asociados a la ingente cantidad y diversidad de los datos —datos

espacio-temporales, imágenes, vídeos, audios y sonidos— que últimamente se transforman en datos computables para ser analizados con algún algoritmo, entre los que destacan los procedimientos predictivos. De modo que la nueva estadística sirva no solo para contrastar teorías científicas existentes sino que dé un paso adelante para *sugerir otras nuevas*.