REAL ACADEMIA DE CIENCIAS
EXACTAS, FISICAS Y NATURALES

Conceptos y técnicas de la estadística bayesiana: comentarios sobre su estado actual

DISCURSO

LEIDO EN EL ACTO DE SU RECEPCION

POR EL

EXCMO. SR. D. FRANCISCO JAVIER GIRON
GONZALEZ-TORRE

Y

CONTESTACION

DEL

EXCMO. SR. D. SIXTO RIOS GARCIA

EL DIA 13 DE MARZO DE 1991

MADRID
DOMICILIO DE LA ACADEMIA:
VALVERDE, 22 — TELEFONO 521 25 29
1991

Depósito Legal: M. 7.152-1991

REALIGRAF, S. A. - Burgos, 12 - 28039 Madrid

DISCURSO

DEL

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

TEMA:

CONCEPTOS Y TÉCNICAS DE LA ESTADÍSTICA BAYESIANA: COMENTARIOS SOBRE SU ESTADO ACTUAL

Excmo. Sr. Presidente,
Excmos. Sres. Académicos,
Señoras y Señores:

Vaya ante todo mi agradecimiento profundo a todos los miembros de esta Real Academia por haber sido elegido para colaborar en vuestro quehacer científico al que, desde este mismo momento, me incorporo. Constituye para mí un gran honor, y a la vez una gran responsabilidad, ocupar la vacante que se produjo tras el inesperado fallecimiento del Excmo. Sr. D. Francisco Azorín Poch del que espero, a pesar de mis escasos méritos, ser su digno sucesor en las tareas científicas que esta Academia me encomiende.

Aunque no tuve la suerte de ser su discípulo, por encontrarse en Latino-América en la época en que yo era estudiante, tuve ocasión de conocerle, allá por el año 1975, tras su definitivo regreso a nuestro país. Con anterioridad, todos los estudiantes de la especialidad de Estadística le conocíamos por su obra *Curso de Muestreo y Aplicaciones*, que en su época constituyó un hito no sólo en la, entonces escasa, literatura estadística en lengua castellana sino a nivel internacional.

Hay curiosas coincidencias como la de que naciera en Málaga, ciudad en la que resido desde hace más de trece años; y la de haber sido nombrado *Doctor Honoris Causa* por la Universidad Malagueña en el año 1985 a propuesta de la Facultad de Ciencias Económicas. En los días previos a su investidura tuve ocasión de charlar con él sobre todo tipo de temas y darme cuenta de su gran erudición e interés por las más dispares cuestiones.

Otros contactos esporádicos con él tuvieron lugar con ocasión de coincidir en tribunales de tesis doctorales y con motivo de

la celebración de la XVI Reunión Nacional de Estadística, Investigación Operativa e Informática que organizamos en Málaga en noviembre de 1986. En todas estas ocasiones pude disfrutar del beneficio de su amable e inteligente conversación a la que se unía una de sus mayores virtudes: la discreción.

La última vez que tuve oportunidad de verle fue tras la conferencia que pronuncié en esta Real Academia con motivo de mi elección como Académico Correspondiente, dos meses antes de su fallecimiento.

Sus muchos y reconocidos méritos a nivel nacional e internacional, que no ha lugar de recordar ya que fueron magistralmente expuestos por el Excmo. Sr. D. Sixto Ríos en la contestación a su discurso de ingreso en la Academia, solamente se vieron superados por su reconocida modestia.

INTRODUCCIÓN

La elección del tema de este discurso, tarea harto difícil como ya han comentado en más de una ocasión algunos de mis predecesores, no hace sino reflejar algunas de mis preferencias y puntos de vista sobre una manera particular de ver ciertos problemas que caen dentro del ámbito de la ciencia estadística.

Mis comienzos en la actividad investigadora, bajo la tutela del profesor Ríos, estuvieron orientados hacia los fundamentos axiomáticos de la *Teoría de la decisión y la inferencia bayesiana,* en particular a la justificación del uso de la probabilidad subjetiva y del principio de la maximización de la utilidad esperada como modelo o paradigma del comportamiento racional en situaciones que entrañan incertidumbre como son, entre otros, los problemas estadísticos.

Las consecuencias prácticas que se derivan de las diversas axiomáticas que se han propuesto para formular los problemas de decisión en ambiente de incertidumbre, no sólo afectan a lo que Savage denomina *pequeños mundos,* es decir el contexto estricto que rodea a un problema de decisión concreto considerado como problema aislado, sino que algunos van más allá e incluso afirman que la vida misma es un proceso estocástico a lo largo del cual nos vemos obligados, en muchos momentos y bajo circunstancias muy dispares, a tomar decisiones. Este proceso de toma de decisiones se ve generalmente complicado, además, por la existencia de elementos que no controlamos directamente: generalmente intervienen, por una parte, el azar y, por otro lado, las preferencias u opiniones ajenas en el caso de problemas donde hay conflicto de intereses, por todo lo cual sería bueno disponer de

una teoría normativa que, al menos, sirviera de guía de comportamiento. De los axiomas de coherencia se deduce, necesariamente, que la aplicación del *principio de maximización de la utilidad esperada*, en cualquiera de sus formas, sirve a tal fin al menos en el sentido restringido de su aplicación a los pequeños mundos, en el contexto de los problemas de decisión unipersonales en ambiente de incertidumbre. Nuestro propósito es, sin embargo, más modesto y por ello nos concentraremos en los aspectos puramente estadísticos de estos problemas.

En toda exposición sobre técnicas y procedimientos bayesianos, como lo es ésta, es imprescindible comentar, aunque sea brevemente, algo sobre los fundamentos de la inferencia bayesiana.

Desde un punto de vista formal la solución bayesiana a los problemas de decisión en ambiente de incertidumbre ofrecía un marco coherente bien fundamentado en el *Cálculo de probabilidades* y en el *Principio de maximización de la utilidad esperada* y que además, examinada desde el punto de vista clásico de la *Teoría de la Decisión* de Wald, proporcionaba, bajo condiciones muy generales, soluciones admisibles o no dominadas; incluso las propiedades frecuentistas de los procedimientos bayesianos superaban a algunos de los procedimientos clásicos. En esa etapa inicial de mi labor investigadora, me consideraba como un bayesiano teórico —nunca había tenido ocasión de contrastar aquellas teorías en la práctica. Curiosamente, ha sido la práctica de la Estadística la que me ha llevado a adoptar el punto de vista bayesiano como el más adecuado, en general, pero sin exclusivismos, para tratar problemas estadísticos e incluso problemas menos estructurados que, en principio, se pueden considerar como de análisis de datos. Aunque el *Análisis de datos exploratorio* parece en principio un compendio de técnicas informales, sobre todo de tipo interactivo y gráficas, Good nos ha ofrecido argumentos que demuestran que también posee ciertos aspectos bayesianos.

Parece un tanto sorprendente que la estructura formal, tan del gusto de los matemáticos puros, que se deduce de los axiomas de coherencia o principios de racionalidad —debidos fundamentalmente a Ramsey en los años veinte, aunque su trabajo pasó desapercibido en su época, a Bruno de Finetti en los treinta y a L. J. Savage, quién realizó una síntesis del trabajo de los dos precedentes y fue el impulsor de la actitud subjetivista o bayesiana

en la estadística a mediados de los años cincuenta— no haya sido aceptada por la mayoría de los estadísticos tanto teóricos como aplicados.

En esta relación de personalidades relevantes al mundo de la inferencia bayesiana, no podemos dejar de mencionar, aunque su enfoque entronca más con el concepto de probabilidad lógica que subjetiva, el trabajo eminentemente aplicado de Harold Jeffreys —al fin y al cabo era un astrónomo— que se recoge en su libro de 1939 *Theory of Probability*, y que sigue siendo hoy día de obligada consulta, pues es siempre fuente de nuevas ideas. Como comentó Arnold Zellner con ocasión del primer Congreso mundial bayesiano, celebrado en Valencia en 1979, «**todo** está en el Jeffreys», parafraseando el conocido aforismo de *todo está en los libros*.

Una de las consecuencias importantes de los axiomas de racionalidad es la utilización del cálculo de probabilidades como la herramienta básica o motor de la inferencia, como gustan de llamar los que cultivan la *inteligencia artificial*; por lo que la inferencia bayesiana podría en principio atraer a los probabilistas. Toda la potencia y las herramientas del cálculo de probabilidades, en particular la utilización de los conceptos de probabilidad y esperanza condicionada de los que se deriva el teorema de Bayes, se encuentran a disposición del estadístico.

No hay necesidad de recurrir en la inferencia bayesiana, a conceptos exógenos al cálculo de probabilidades como son, por ejemplo, los estimadores puntuales, las medidas de precisión como el error cuadrático medio, el error de tipo I, los niveles de significación, la potencia de un contraste, etc. Desde una perspectiva estrictamente bayesiana, si se adopta un enfoque paramétrico, toda la información relevante está contenida en la distribución a posteriori del parámetro de interés del modelo. Los parámetros ancilares, útiles en la formulación del problema pero generalmente irrelevantes al problema de inferencia o decisión que se esté considerando, se eliminan simplemente calculando la distribución marginal a posteriori del parámetro de interés. No hay necesidad de recurrir a otros procedimientos no probabilísticos como verosimilitudes marginales, métodos pivotales, etc. para eliminar estos parámetros.

El importante problema de hacer predicciones, que muchos

estadísticos consideran como el problema central de la estadística, se reduce sencillamente a calcular la distribución de una o varias observaciones futuras (o estadísticos de éllas) condicionada a los datos ya observados. A esta distribución se le conoce generalmente, en inferencia bayesiana, como *distribución predictiva*, la cual se obtiene fácilmente combinando el modelo probabilístico para las observaciones futuras con la distribución a posteriori, de acuerdo con las reglas del cálculo de probabilidades.

El enfoque predictivista a la inferencia, a diferencia del enfoque paramétrico o del basado en la *Teoría de la Decisión*, tiene su punto de partida en una idea básica debida a de Finetti a finales de los años veinte, que es la de sucesos y variables aleatorias intercambiables.

El concepto de *intercambiabilidad* intenta recoger la idea de lo que debe entenderse por una muestra aleatoria sin hacer referencia o mención explícita a la noción de parámetro ni al concepto de independencia estocástica, que a veces resulta un tanto difícil de justificar en la práctica, como ha señalado Kolmogorov.

La idea de variables aleatorias intercambiables es simple y viene asociada a otra igualmente simple como es la de simetría, con la que se intenta plasmar, en términos matemáticos precisos, el que la estructura probabilística de los observables no depende del orden en que estos aparecen en la muestra; lo que se expresa diciendo que cualquier permutación de las variables observables tiene la misma distribución. Esta definición, como ya hemos comentado, capta perfectamente la idea de muestra aleatoria simple, sin necesidad de recurrir a elementos extraños o ajenos a las observaciones como pudieran ser los parámetros de un modelo estadístico ni a la noción de independencia condicional de las observaciones respecto del parámetro. Sólo hace referencia a variables o magnitudes observables.

Desde un punto de vista teórico es conveniente extender la idea de intercambiabilidad a sucesiones de variables aleatorias, lo que además parece razonable desde el punto de vista práctico. ¿Por qué se va a limitar el concepto de simetría a un número finito prefijado de observaciones, cuando podríamos considerar muestras potencialmente observables de tamaño arbitrario que presenten la misma estructura de simetría?

El teorema de representación de de Finetti, y sus generaliza-

ciones posteriores debidas a Hewitt y Savage, suele tomarse como justificación alternativa de la inferencia bayesiana —la otra justificación es la dada por los axiomas de coherencia. Un resumen excelente del estado actual de la investigación sobre la intercambiabilidad, que incluye extensiones de estos conceptos y resultados tan importantes como las versiones finitas del teorema de de Finetti, intercambiabilidad parcial y su extensión a procesos estocásticos, es el de Diaconis de 1988.

Los teoremas de representación de sucesiones de variables aleatorias intercambiables revelan la estructura subyacente de éstas en forma de:

(i) un modelo estadístico usual, paramétrico o no, en el que las observaciones, condicionadas al parámetro, se comportan como una sucesión de variables aleatorias independientes e idénticamente distribuídas, que equivale a la idea clásica de modelos estadísticos basados en el muestreo aleatorio simple.
(ii) un espacio paramétrico finito o infinito dimensional que, sin más hipótesis adicionales, es el formado por todas las distribuciones de probabilidad sobre el espacio muestral común de los observables.
(iii) una distribución a priori sobre el espacio paramétrico que representa la opinión subjetiva sobre la función de distribución empírica de la sucesión.
(iv) la representación de la distribución marginal del proceso como una mixtura del modelo estadístico de variables aleatorias condicionalmente i.i.d. respecto de la distribución a priori.

Si a la hipótesis de intercambiabilidad se le añaden otras específicas, se pueden obtener, y de este modo justificar, la mayoría de los modelos paramétricos estadísticos usuales. De esta forma se puede justificar el modelo normal, el más frecuentemente utilizado en estadística, si a la sucesión de variables observables se le exige, además de la hipótesis de intercambiabilidad, la de simetría esférica centrada. El modelo usual para datos binarios o dicotómicos —es decir, el basado en pruebas de Bernoulli (condicionalmente) independientes— no es sino una consecuencia obligada, y por tanto inexcusable, del teorema original de de Finetti.

La interpretación de estos resultados clarifica y, en algunas circunstancias, permite la elección de un modelo estadístico paramétrico convencional basado, no en la conveniencia o simplicidad de su tratamiento, como ocurre con harta frecuencia en la estadística aplicada, sino en la estructura que presentan los datos u observables. El teorema anterior establece incluso un nexo de unión, a través de la ley de los grandes números para variables aleatorias intercambiables, con la inferencia clásica y además aclara la interpretación de la distribución a priori.

Con frecuencia se ha criticado la postura bayesiana mediante argumentos basados en la arbitrariedad y subjetividad de los modelos estadísticos y de la distribución a priori utilizados, particularmente esta última. Desde el punto de vista bayesiano ambos ingredientes —la función de verosimilitud, dada por el modelo estadístico y por la muestra observada y la distribución a priori— entran a formar parte de la fórmula de Bayes en igualdad de condiciones.

Se olvida con frecuencia que toda la información relevante al problema de la inferencia bayesiana está contenida en la distribución conjunta de las observaciones y del parámetro, independientemente de cómo ésta se haya obtenido. Generalmente, aunque no necesariamente, esta distribución conjunta se especifica a partir de la distribución de las observaciones condicionada por el parámetro y de la marginal de éste, es decir, la a priori, de las que resulta la distribución conjunta aplicando el teorema de Fubini, en el caso más general, o simplemente multiplicando ambas densidades, en el caso absolutamente continuo. La obligada consulta del clásico, y en su época influyente, libro de Raiffa y Schlaifer de 1961 así nos lo ha recordado.

No debemos olvidar que en la estadística bayesiana, y esto también ocurre en la estadística clásica aunque casi nunca se dice explícitamente, todas las inferencias son condicionales: al modelo y a la información a priori. El por qué considerar uno de los dos ingredientes más subjetivo que otro, centro de las críticas a los bayesianos, no está para mi claro. ¿Será quizás porque el modelo estadístico es elemento común, mientras que la probabilidad a priori no lo es, y no les gusta tirar piedras contra su propio tejado?

No sólo la inferencia clásica está condicionada por la elección del modelo sino también por los procedimientos *ad hoc* de la infe-

rencia. Por procedimientos ad hoc me refiero a la mayoría de las técnicas de optimación de la inferencia, como son los procedimientos insesgados, de varianza mínima, la suficiencia, la eficiencia y un largo etcétera.

No suponga todo lo anterior que la especificación de las opiniones subjetivas en términos de una medida de probabilidad es tarea fácil, sobre todo en problemas multiparamétricos. En algunas circunstancias la aplicación del concepto de intercambiabilidad puede ayudar a reducir la dimensión del espacio paramétrico, aunque algunas veces suele ser a costa de la introducción de hiperparámetros de difícil interpretación. En otros casos ésto no es posible y se recurre a la utilización de las llamadas *distribuciones de referencia* o no informativas. Incluso la determinación de éstas en modelos estadísticos complejos puede ser complicada, e incluso a veces, imposible de obtener; además en muchas ocasiones suelen ser distribuciones impropias. Ni tampoco que la inferencia bayesiana está totalmente libre de paradojas o inconsistencias, cuando se violan algunos de los principios de coherencia básicos. La existencia de paradojas en inferencia bayesiana se debe fundamentalmente a la utilización de distribuciones impropias lo que, en principio, invalida el teorema de Bayes. Aunque su aplicación formal en términos de proporcionalidad a la función de verosimilitud de la derivada de Radon–Nikodym de la distribución a posteriori respecto de la a priori se emplee a veces, sobre todo cuando se utilizan distribuciones de referencia que suelen ser impropias, este uso indebido del teorema de Bayes puede conducir, como demostraron en 1973 Dawid, Stone y Zidek, a las denominadas paradojas de marginalización.

Resulta curioso e irónico a la vez, que dos bayesianos, Diaconis y Freedman, que se autoproclaman, el primero de ellos como subjetivista y el segundo como clásico, hayan demostrado recientemente, en 1986, que puede haber problemas con los estimadores de Bayes, vistos desde el punto de vista de sus propiedades frecuentistas, concretamente con la inconsistencia de los mismos en ciertas situaciones; mientras que, por otro lado, L. Le Cam, un firme y convencido antibayesiano ha dedicado muchos esfuerzos a demostrar las buenas propiedades de los estimadores de Bayes.

Estos resultados teóricos hay que tomarlos en consideración: no todas las combinaciones posibles de la función de verosimili-

tud con la distribución a priori producen distribuciones a posteriori asintóticamente consistentes. Aunque generalmente la inconsistencia de los estimadores de Bayes se produce en situaciones no paramétricas o infinito dimensionales y en conjuntos de medida nula, estos autores sugieren que incluso en el caso muy frecuente de la estimación del parámetro de localización de una familia de distribuciones, el estimador de Pitman puede ser inconsistente para familias paramétricas finito dimensionales.

Otros sistemas de inferencia se han desarrollado posteriormente para tratar algunos problemas que, al parecer, no encuadran dentro del marco de los problemas de decisión como son, por ejemplo, los basados en la teoría de los conjuntos difusos o borrosos, como los denominaba Azorín y en las funciones de credibilidad de Shafer. Contra estos enfoques alternativos a los problemas inferenciales se han alzado en varias ocasiones las voces de Lindley, French y otros muchos, quienes vigorosamente argumentan sobre la inevitabilidad de la probabilidad como la única medida de la incertidumbre.

MÉTODOS CLÁSICOS CONTRA MÉTODOS BAYESIANOS

A pesar del provocativo título de esta sección no voy a entrar en la polémica de bayesianos contra no bayesianos. Simplemente me limitaré a exponer algunos aspectos de ambos enfoques haciendo especial hincapié en su forma de tratar algunas aplicaciones tomadas, fundamentalmente, de la bioestadística.

Debemos recordar aquí que el desarrollo histórico de la estadística clásica ha tenido influencia decisiva en el tipo de enfoque que se ha dado a las aplicaciones, sobre todo a las de tipo biométrico. En estas ciencias donde tradicionalmente se ha utilizado la estadística, e incluso en algunas más recientes como la econometría y la psicometría, la influencia de la escuela clásica basada sobre todo en las ideas de Fisher, por un lado, y de Neyman y Pearson por otro, ha sido determinante.

Los paquetes de programas de ordenador desarrollados para estas disciplinas, como el BMDP, el SPSS y el SAS así lo demuestran; están basados en procedimientos clásicos que hacen un uso intensivo, y casi siempre abusivo, de los contrastes de hipótesis.

Parece aberrante la utilización todavía de los conceptos de tests significativos y muy significativos, que muchas veces se incluyen como valores *por defecto* en estos programas estadísticos, sin tener en cuenta ni la naturaleza ni el contexto del problema que se está tratando.

A este respecto, de una entrevista realizada por M. H. De Groot en el otoño de 1984 a Erich Lehmann, el afamado autor del texto *definitivo* sobre contrastes estadísticos de hipótesis, extraigo los siguientes comentarios a la pregunta de por qué la gente usa el .05 como nivel de significación en los contrastes de hipótesis:

> "En mis clases siempre trato de hablar de ésto porque creo que es una cuestión interesante. Obviamente [el usar .05 como nivel de significación](*) es una tontería, pues hay que considerar también la función de potencia. Y sin embargo hay estudios muy interesantes que demuestran que la gente usa el contraste de hipótesis en situaciones donde la función de potencia es tan pequeña que debería olvidarse de realizar el experimento porque no hay prácticamente posibilidad alguna de descubrir el efecto en el que están interesados. Lo que necesitan es una muestra mayor. ... Pero, a parte de que a la gente le gusta que le digan como hacer las cosas de un modo determinado, y a los editores les gusta aplicar una regla fija de modo que no aceptan un artículo a no ser que el resultado sea significativo al nivel del .05, existe también la ventaja de que si se usa un procedimiento de un modo estándar, uno se acostumbra a él. ... [De este modo] la gente puede comunicarse mucho mejor que si no tuviesen un estándar. Así, aunque básicamente me opongo a ésto [al uso del nivel de significación .05 ó .01], veo que hay algo de positivo."

Incluso el simple y conocido hecho para un estadístico profesional de que el nivel de significación debe variar con el tamaño muestral, parece no tenerse en cuenta por parte de la comunidad científica, que utiliza cada vez con más frecuencia técnicas estadísticas en sus investigaciones. Últimamente parece ser que hay un cambio de actitud al considerar otras técnicas alternativas que utilizan medidas de precisión o evaluación más informativas, como son los intervalos de confianza, en substitución de los clásicos contrastes de hipótesis o valores P.

(*) Los comentarios entre corchetes se han añadido al texto original.

Recientemente se han obtenido resultados que ponen en evidencia la interpretación y utilización de los valores P que habitualmente se emplean en este tipo de problemas, concretamente en el contraste de hipótesis nulas simples frente a alternativas compuestas, que invitan a una reconsideración del empleo en la práctica de estos contrastes. Una de las conclusiones de estos resultados es que muchos usuarios interpretan, erróneamente, la P obtenida de los datos experimentales como la probabilidad de que la hipótesis nula sea cierta; la otra conclusión importante es que los valores P, obtenidos por los procedimientos clásicos en muchos de los problemas comunes de la inferencia, son mucho menores que los factores de Bayes para el problema de contraste, independientemente de la distribución a priori que se elija, siempre que ésta sea razonable. La consecuencia práctica de estos resultados es que para rechazar una hipótesis nula precisa hacen falta valores notablemente más pequeños de la P que los que normalmente se utilizan.

No obstante, y a pesar de todo, tradicionalmente ha habido ciertas parcelas de las aplicaciones estadísticas que han tenido una componente bayesiana, como la evaluación de las primas de riesgo por parte de las compañías aseguradoras que combinan la experiencia pasada con la actual, los tests educacionales que combinan la experiencia acumulada sobre un tipo de test con las respuestas al mismo por parte de un candidato y ciertas aplicaciones a problemas legales, en particular las que se refieren al tratamiento de la evidencia en pruebas periciales. Este tipo de aplicaciones del cálculo de probabiliades al derecho son bastante frecuentes, sobre todo en los Estados Unidos.

Otras disciplinas más recientes como los *sistemas expertos y la inteligencia artificial*, que tratan de problemas en los que hay incertidumbre presente, son, por su propia naturaleza, totalmente ajenas a la estadística clásica y a la noción de probabilidad frecuentista implícita en élla. Sin embargo, la probabilidad subjetiva como descripción de la incertidumbre, junto con otros nuevos enfoques a la misma, como el razonamiento difuso o borroso o la teoría de las funciones de credibilidad de Dempster-Shafer, han encontrado lugar en estas nuevas áreas de conocimiento, especialmente en los sistemas expertos.

Otras parcelas de las aplicaciones de la estadística, como por

ejemplo la bioestadística, han estado totalmente dominadas, hasta hace poco, por la metodología clásica. Sin embargo, la actitud de los bioestadísticos hacia la metodología bayesiana parece haber cambiado en la última década debido sobre todo al desarrollo creciente de herramientas de cálculo para tratar problemas reales. Así, por ejemplo, el diagnóstico médico es un campo donde la experiencia acumulada por los médicos sobre los síntomas causados por diversas enfermedades, expresada en términos probabilísticos, combinada, mediante el teorema de Bayes, con los resultados de análisis clínicos y otros datos objetivos da como resultado una distribución de probabilidad diagnóstica sobre las posibles enfermedades que pueda padecer el paciente.

Esta manera probabilística de diagnosticar puede parecer extraña en principio; no estamos acostumbrados a que el doctor nos diga «tiene Vd. tal y tal otra enfermedad con tales probabilidades»; sin embargo, este tipo de diagnóstico es mucho más informativo —aunque quizás no lo sea para el paciente— que decir, por ejemplo, que uno tiene como enfermedad más probable la moda de la distribución diagnóstica cuando esta probabilidad puede ser pequeña. Además, la decisión de aplicar tal o cual tratamiento o de operar o no operar, que sí afecta al paciente y que a veces puede entrañar graves riesgos e incluso problemas éticos, depende de *toda la distribución diagnóstica no sólo de alguna de sus características.*

Otros conceptos como el de bioequivalencia de compuestos farmaceúticos son perfectamente naturales para los estadísticos bayesianos debido a la naturaleza que presentan como problemas de decisión. Para estos problemas en concreto los modelos jerárquicos bayesianos ofrecen un marco natural.

Concretamente en el campo de las pruebas clínicas que tradicionalmente se han enfocado desde un punto de vista frecuentista como una aplicación del clásico test secuencial de la razón de verosimilitudes de Wald, parece que hay una tendencia a considerar el punto de vista bayesiano o empírico-bayesiano como más satisfactorio y libre de los impedimentos que normalmente afectan a aquellas desde el punto de vista tradicional, como es la dependencia de las conclusiones del experimento de la regla de parada. En casos reales que se dan con frecuencia, como la interrupción de las pruebas por falta de fondos o por cambios de última hora en las prioridades de la investigación, es imposible, en un sentido

estricto desde la óptica frecuentista, hacer inferencias sobre los resultados de las pruebas, ya que los objetivos previstos al principio de las mismas, normalmente el control de los dos tipos de error, no se han cumplido antes de la interrupción. Esto no parece razonable y sin embargo se ha gastado mucho esfuerzo y dinero en análisis de pruebas clínicas poco apropiadas.

El difícil problema de estimar el riesgo de cáncer en los seres humanos por exposición a multitud de agentes ambientales, donde se dispone de pocos datos, a partir de datos experimentales obtenidos con animales, de los que hay abundancia relativa de datos, ha sido y es objeto de mucho debate. En 1988, un comité de la Academia Nacional de Ciencias de los Estados Unidos, encargado de estudiar los efectos cancerígenos de los radioisótopos del plutonio en los seres humanos, a través de los datos obtenidos de estudios realizados con perros y ratas, decidió adoptar la metodología bayesiana desarrollada por DuMouchel y Harris.

EL FILTRO DE KALMAN

El filtro de Kalman, que ha venido siendo utilizado como una herramienta básica en muchas de las técnicas y aplicaciones a las ingenierías eléctrica y de telecomunicaciones y que ha jugado un papel fundamental en el desarrollo de la teoría del control óptimo desde comienzos de los años sesenta, ha estado prácticamente ausente de la literatura estadística hasta mediados de la década de los setenta, si exceptuamos su utilización en el modelo de regresión lineal por Plackett en el año 1950, que desafortunadamente no tuvo la transcendencia que merecía en la práctica estadística de su tiempo, debido casi con seguridad a la inexistencia en esa época de programas de ordenador que implementasen el carácter secuencial del algoritmo.

La contribución principal de Kalman fue la de extender el filtro de Wiener a sistemas multivariantes con coeficientes variables en el tiempo y con errores no estacionarios y obtener una forma secuencial o recursiva de la solución óptima, mediante la reformulación del problema en términos de lo que ahora se conoce como representación en el *espacio de estados*.

El filtro de Kalman puede justificarse desde diversos enfoques, análogamente a como se consideran los modelos lineales desde varios puntos de vista, a saber:

a) Estimación lineal de mínima varianza, según la teoría clásica de la estimación, o minimización del error cuadrático medio si se adopta el punto de vista de la *Teoría de la Decisión.*
b) Procedimientos de tipo geométrico basados en proyecciones ortogonales sobre subespacios.
c) Estimación bayesiana secuencial.

Es interesante destacar que aunque la derivación original del filtro de Kalman se basaba en la idea de estimar un estado futuro del sistema mediante la técnica de mínimos cuadrados, la herramienta básica para la predicción es, como el mismo Kalman señaló en 1978, la *esperanza condicionada.*

Para el caso de errores normales o gaussianos, los tres procedimientos conducen a las mismas ecuaciones recursivas para la estimación secuencial de los parámetros del modelo, lo que se conoce habitualmente con el nombre de *filtrado.* Si se adopta el enfoque de la teoría de la decisión, se puede demostrar además para este caso, que el filtro de Kalman es invariante o robusto, en el sentido de minimizar el riesgo de Bayes, respecto de una amplia clase de funciones de pérdida, no solamente la cuadrática, como generalmente se supone.

Es, sin embargo, analizando el filtro de Kalman desde una perspectiva bayesiana cuando se obtiene una visión más general y más clara del filtro, que además permite su generalización en varios sentidos. Así, el filtro no es sino una consecuencia directa del teorema de Bayes en su forma secuencial y de las propiedades de la distribuciones normales multivariantes. Además, en este caso el enfoque bayesiano permite estimar la distribución a posteriori exacta de los parámetros del modelo, no sólo algunas de sus características, como en los otros métodos que generalmente estiman los momentos de 1º y 2º orden, aunque en el caso gaussiano, éstos determinan unívocamente la distribución, de ahí la equivalencia de los tres métodos.

Su generalización permite incluso la estimación recursiva de la varianza de los errores, para el caso particularmente importante de problemas estadísticos en los que la fuente de variabilidad en

las ecuaciones de observación y del sistema se comporte como un proceso gaussiano i.i.d., extendiendo así el análisis del caso normal mediante la utilización de técnicas bayesianas estándar como son las *familias conjugadas* y las propiedades de ciertas familias de distribuciones. Es interesante señalar cómo el concepto de familia conjugada, inicialmente reservado al caso de procesos independientes tal como lo concibieron Raiffa y Schlaifer en 1961, se puede extender a procesos dependientes como los generados por los modelos lineales dinámicos, en el sentido de que la distribución a posteriori sigue perteneciendo a la misma clase que la a priori inicial en cualquier instante de tiempo t. Esto abre el camino a una posible generalización del concepto de familia conjugada al caso de procesos dependientes. Su existencia parece también depender, al igual que en el caso de independencia, de la existencia de estadísticos suficientes de dimensión fija y posiblemente del carácter markoviano del proceso. La caracterización de estos procesos es un problema abierto interesante.

Recientemente, en 1989, Meinhold y Singpurwalla han demostrado que el filtro de Kalman es débilmente robusto para observaciones y errores distribuídos según una t de Student multivariante, entendiendo por ésto el que las ecuaciones básicas del filtro son idénticas a las del caso normal. La distribución a posteriori de los parámetros es también una t multivariante cuyos parámetros se obtienen de las ecuaciones básicas del filtro y de una nueva ecuación también de carácter recursivo. Este tipo de resultados se asemeja a los ya conocidos en regresión estática, como los obtenidos por Zellner considerando errores que se distribuyen también según una t de Student multivariante y los más generales de Dawid, Jensen y Jensen y Good, sobre la generalización del teorema de Gauss-Markov a modelos lineales con errores no normales que ni siquiera poseen momentos. La simetría esférica de los errores es el único requisito para su validez.

Este resultado puede ser generalizado a una clase más amplia de distribuciones multivariantes; a saber, la clase generada por las mixturas de distribuciones normales multivariantes esféricas respecto de un parámetro de escala, e incluso conjeturamos que el resultado sigue siendo válido, bajo ciertas condiciones, para la clase de todas las distribuciones que presentan simetría esférica. Ésto, junto con el resultado considerado anteriormente de robustez

del filtro respecto de una amplia clase de funciones de pérdida, constituiría un resultado muy general sobre la robustez global del filtro.

Pero la gran ventaja del enfoque bayesiano al filtro de Kalman, sobre los otros dos, reside en su relativa facilidad de ser generalizado a estructuras de error no necesariamente normales. Resultados exactos, en el sentido de que la distribución a posteriori pueda ser calculada en cada instante de tiempo para los parámetros del modelo, solamente se dan bajo condiciones restrictivas, pero que permiten generalizar las ecuaciones del filtro conservando la estructura básica del mismo; es decir, el carácter secuencial o recursivo del proceso de aprendizaje sobre los parámetros del modelo.

Estamos de acuerdo con Lindley cuando en su discurso memorial de Wald de 1988 sobre *El estado actual de la estadística bayesiana* comentaba que el filtro es uno de los avances bayesianos más importantes de los últimos años. Tanto si se adopta una posición bayesiana como si no, el filtro es una herramienta básica que debería incorporarse al bagaje de todo estadístico. De hecho, en algunas de las últimas versiones de los llamados por los usuarios *paquetes estadísticos*, que no están basados en el enfoque bayesiano a la inferencia, ya aparecen versiones y aplicaciones del filtro de Kalman.

De todos modos, como es bien sabido, el filtro de Kalman bajo la hipótesis de errores gaussianos no es fuertemente robusto, en el sentido que hemos descrito anteriormente de ser invariante respecto de variaciones en la distribución de los errores del modelo normal cuando éstos se consideran independientes, por lo que a partir de mediados de los años setenta ha habido muchas aportaciones a la literatura encaminadas a robustecer el filtro de Kalman, de lo que trataremos más adelante.

MODELOS JERÁRQUICOS, MIXTURAS Y LA PARADOJA DE STEIN

Íntimamente relacionados con el filtro de Kalman están los *modelos lineales jerárquicos* desarrollados por Lindley y Smith en los años 1972-73, que desde un punto de vista puramente formal pueden considerarse como un caso particular del filtro. Su génesis,

sin embargo, es totalmente distinta a la del filtro, pues no son modelos dinámicos, es decir, no evolucionan en el tiempo. Son, desde la perspectiva bayesiana, un modelo lineal ordinario en el que los parámetros son variables aleatorias que satisfacen a su vez otro modelo lineal y así sucesivamente a través de una jerarquía finita.

La justificación teórica última de estos modelos y de otros más complejos, que no necesariamente exhiben una estructura lineal, se basa en uno de los conceptos claves del cálculo de probabilidades, que además está en la raíz de los fundamentos de la inferencia bayesiana. Es el concepto de *intercambiabilidad*, que ya comentamos brevemente en la introducción.

La idea de introducir hiperparámetros en los modelos bayesianos cumple una finalidad doble. Por una parte, permite dar una mayor flexibilidad y generalidad a los modelos paramétricos clásicos, aunque a veces a costa de complicar su análisis al no existir, por ejemplo, familias conjugadas y, por otra parte, permite simplificar los modelos que contienen un elevado número de parámetros, precisamente mediante la aplicación del concepto de intercambiabilidad y la utilización del teorema de de Finetti para expresar la distribución de estos parámetros como una mixtura de otros, llamados hiperparámetros, de dimensión menor.

Muchos de los modelos estadísticos paramétricos complejos que no pueden analizarse de un modo simple desde un punto de vista clásico o frecuentista —por ejemplo, aquellos para los que no existen estadísticos suficientes de dimensión fija— tienen oculta, sin embargo, una estructura subyacente que se puede a veces expresar en términos de mixturas, como es el caso de considerar, por ejemplo, muestras de una distribución t de Student k-dimensional.

Es bien sabido que la distribución t de Student es una mixtura de distribuciones normales k-variantes respecto del parámetro de escala, con distribución de mezcla una gamma invertida. Este hecho —que desde el punto de vista de la estadística clásica no tiene relevancia alguna, salvo, quizás, el de proporcionar un procedimiento para simular muestras de una Student multivariante— es realmente importante desde un punto de vista bayesiano, puesto que convierte el modelo original en un modelo jerárquico mediante la introducción de un modelo normal k-variante subyacente y un hiperparámetro *unidimensional*, lo que permite, en principio

e independientemente del costo computacional, un análisis bayesiano estándar del problema, con la ventaja añadida de poder calcular la distribución a posteriori del parámetro latente o hiperparámetro que, por ejemplo, puede utilizarse para contrastar la adecuación del modelo a los datos.

Esta importante idea de revelar la estructura oculta o subyacente de algunos modelos estadísticos no estándar de modo que se pueden considerar como distribuciones marginales o baricentros de una multidimensional obtenida combinando la distribución condicionada de un modelo respecto de un parámetro ancilar con la marginal de éste a través de la operación de mixtura, puede ser una de las más interesantes y que ya está empezando a explotarse en campos tan recientes como el denominado *muestreo bayesiano*. Lo único que debe exigirse a la anterior representación de la distribución marginal como mixtura de otras es que éstas sean distribuciones fácilmente tratables desde un punto de vista analítico o bién sean fáciles de simular.

Curiosamente, la consideración de los modelos lineales jerárquicos permitió dar una justificación bayesiana y también empírico-bayes al denominado efecto o paradoja de Stein y a toda la amplia literatura sobre la estimación contraída, que surgió a partir del sorprendente descubrimiento de Stein en 1956 de que, en dimensiones estrictamente mayores que 2, el estimador usual del vector de medias de una distribución normal multivariante, es decir, la media muestral que es el estimador de máxima verosimilitud, es inadmisible para funciones de pérdida cuadráticas. Posteriormente James y Stein, en 1961, demostraron la existencia y dieron fórmulas explícitas de una amplia clase de estimadores (denominados contraídos) que dominaban al estimador usual. El nombre de estimadores contraídos se debe a que se pueden considerar como una media ponderada del origen y de la observación correspondiente, lo que equivale a una contracción del estimador ordinario hacia el origen. Efron y Morris han generalizado estos estimadores a otros que contraen la observación no hacia el origen sino hacia el vector de medias y han demostrado que estos estimadores también dominan a la media muestral para dimensiones mayores que tres.

Stigler, en el discurso memorial de Neyman de 1988, comenta con su ingenio habitual sobre estos estimadores:

"Cuando uno se encuentra por vez primera con este fenómeno puede parecer ridículo —¿cómo puede utilizarse la información del precio de las manzanas de Washington y de las naranjas de Florida para mejorar una estimación del precio del vino francés, cuando se supone que no están relacionados? La mejor explicación heurística que puede ofrecerse es la bayesiana."

Stigler se refiere evidentemente a la aplicación de la hipótesis de intercambiabilidad en los precios. Desde el punto de vista práctico la utilización de la hipótesis de intercambiabilidad debe realizarse con cautela y siempre tras un análisis cuidadoso del problema en cuestión. Hay situaciones reales donde la utilización del concepto de intercambiabilidad no es adecuada y hay que recurrir al más débil de intercambiabilidad parcial. Estas ideas encuentran una gran aplicación en cierto tipo de modelos de diseño experimental.

Como ilustración de estas ideas consideremos un ejemplo clásico tomado del *Análisis de la varianza*. Supongamos el caso más sencillo del modelo de clasificación simple con un sólo factor o *modelo completamente aleatorizado* donde las variables observables y_{ij} tienen la estructura dada por

$$y_{ij} = \theta_i + u_{ij} \qquad i = 1, \ldots, m; \quad j = 1, \ldots, n_i;$$

y son independientes condicionalmente a $(\theta_1, \ldots, \theta_i, \ldots, \theta_m)$ y a la varianza común de los errores σ_w^2.

Este modelo es formalmente análogo al modelo II de componentes de la varianza. La diferencia entre ambos estriba en dónde radica el interés del estadístico, si en estimar las medias θ_i o las componentes de la varianza global, σ_b^2 y σ_w^2. Desde el punto de vista bayesiano los dos modelos son idénticos, aunque el análisis bayesiano se suele centrar principalmente en el estudio de las medias θ_i. La diferencia entre ambos modelos se suele reflejar en la elección de la varianza σ_b^2; si ésta es muy grande se tiene el modelo I.

Si suponemos que los efectos θ_i son intercambiables y se consideran como una muestra de una distribución normal $N(\mu, \sigma_b^2)$, tenemos un modelo jerárquico bietápico, dado por

$$y_{ij} \sim N(\theta_i, \sigma_w^2),$$
$$\theta_i \sim N(\mu, \sigma_b^2).$$

Si completamos la descripción del problema suponiendo, por sencillez, que las varianzas σ_w^2 y σ_b^2 son ambas conocidas y que μ sigue una distribución localmente uniforme, se obtiene que los estimadores de Bayes de los θ_i son de la forma

$$\hat{\theta_i} = \omega_i y_{i.} + (1 - \omega_i)\bar{y};$$

donde los pesos $\omega_i = n_i\sigma_b^2/(n_i\sigma_b^2 + \sigma_w^2)$. Es decir, son los estimadores usuales $y_{i.}$ contraídos hacia la media global $\bar{y}$.

Otra de las aplicaciones interesantes de los modelos lineales jerárquicos es la de proporcionar una justificación a la denominada regresión cresta debida a Hoerl y Kennard en 1970, que Azorín gustaba de denominar regresión riscal *(ridge regression)*. La idea de la regresión riscal, que está también relacionada con el efecto Stein, se basa en generalizar la regresión mínimo-cuadrática, minimizando en su lugar el error cuadrático medio, lo que implica prescindir de la hipótesis de que los estimadores de regresión sean insesgados. Los estimadores así obtenidos presentan una gran similitud con los estimadores contraídos de Stein y pueden obtenerse fácilmente utilizando un modelo lineal jerárquico trietápico en el que al último hiperparámetro de la jerarquía se le asigna una distribución localmente uniforme.

Una alternativa bayesiana a la estimación no paramétrica es la propuesta por Ferguson en los años 1973-74, que ha dado lugar a una gran cantidad de trabajos, sobre todo orientados hacia problemas de supervivencia y datos censurados. Otro enfoque alternativo a los problemas no paramétricos se puede realizar a través de los modelos paramétricos basados en mixturas finitas de distribuciones. La justificación teórica de este enfoque se basa en un resultado obtenido independientemente por Diaconis e Ylvisaker y Dalal y Hall en 1983, que substancialmente afirma que cualquier distribución puede aproximarse, en el sentido de la topología débil, por una mixtura finita de familias conjugadas, lo que permite su utilización como modelo para una aproximación bayesiana a los problemas no paramétricos y semi-paramétricos. Una consecuencia inmediata de estos resultados es la extensión de la noción de familia conjugada al caso de mixturas finitas: éstas son también cerradas por el muestreo aunque no satisfacen la definición más restrictiva de famila conjugada dada en términos de linealidad de la esperanza a posteriori.

Los modelos basados en mixturas de familias conjugadas, particularmente de distribuciones normales, se han utilizado también como una alternativa paramétrica a la técnica del análisis multivariante conocida como *análisis de conglomerados.* Sin embargo, la estimación de los parámetros de la mixtura presenta problemas que no suelen aparecer en modelos estadísticos más convencionales. Así, por ejemplo, los estimadores basados en el método de los momentos, aunque son asintóticamente consistentes, son muy ineficientes para muestras de tamaño pequeño o moderado y también cuando hay un elevado número de componentes en la mixtura; además, la función de verosimilitud no está acotada; pueden existir varios máximos locales e incluso puntos de silla. Ni siquiera la aplicación del método de estimación basado en el algoritmo iterativo de esperanza y maximización conocido como *EM*, debido a Dempster, Laird y Rubin, que considera el modelo de mixtura como un problema de datos incompletos (no se sabe a que conglomerado pertenece cada elemento de la muestra) produce, en muchas ocasiones, resultados aceptables. El algoritmo es muy sensible a la solución inicial y aunque suele convergir a un máximo local, no hay garantía de que sea éste el que corresponda a la solución fuertemente consistente.

La solución bayesiana al problema es, en principio, inmediata; sin embargo, la solución exacta es una mixtura de k^n términos, tantos como particiones pueden hacerse de la muestra de tamaño n en k subconjuntos, donde k es el número de conglomerados; lo que plantea problemas computacionales incluso con muestras de tamaño moderado, por lo que hay que recurrir, bien a métodos aproximados de tipo secuencial, que se conocen en la literatura especializada como *métodos de aprendizaje secuencial*, o bien a procedimientos basados en las nuevas ideas del muestreo bayesiano.

Para complicar aún más la situación, la determinación de la distribución de referencia en problemas de mixturas no es analíticamente tratable: no se puede calcular explícitamente la distribución de referencia dada por la regla de Jeffreys. Este problema está obviamente emparentado con el comportamiento anómalo de la función de verosimilitud que hemos comentado en el parágrafo precedente.

La utilización de las distribuciones de referencia en proble-

mas estadísticos estándar, que en general suelen ser impropias, da lugar a resultados satisfactorios. Por el contrario, el modelo de mixtura **no** permite la utilización de distribuciones impropias, problema que está relacionado con la falta de identificabilidad de los componentes de la mixtura. Baste observar que hay una probabilidad estrictamente positiva de que ninguna observación provenga de una de las poblaciones de la mixtura lo que, obviamente, no permitiría estimar los parámetros de esa componente.

Todos estos comentarios señalan algunos de los problemas que se presentan en el análisis de modelos estadísticos basados en mixturas finitas de distribuciones, que aún están por resolver satisfactoriamente. Otras soluciones alternativas a éste y otros problemas de la inferencia las comentamos en una sección posterior, que trata de nuevas técnicas, posibles gracias a la utilización de métodos de cálculo intensivo —podríamos decir una especie de versión o réplica bayesiana a las técnicas del *bootstrap*— y que, al parecer, prometen convertirse en una de las áreas de investigación aplicada más activas de la inferencia bayesiana de la futura década.

Observaciones anómalas y robustez bayesiana

El disponer de procedimientos de cálculo rápidos y eficientes ha permitido a la ciencia y a la praxis estadística, en general, liberarse de las restricciones que imponían los modelos clásicos basados en la existencia de estadísticos suficientes y procedimientos de inferencia óptimos. Por otra parte, el análisis estadístico de grandes bases de datos presenta serios problemas acerca de la calidad de tales datos. Un porcentaje no despreciable de la información contiene errores de transcripción, algunos datos están incompletos, otros censurados, etc. Un análisis directo de estos datos —que se suele hacer con más frecuencia de lo que sería deseable, incluso con un modelo estadístico adecuado a los mismos— produciría en general resultados y, por consiguiente, conclusiones erróneas. Es lo que los angloparlantes denominan *garbage in-garbage out*: si en una terminal de ordenador se introducen datos erróneos, las salidas, con independencia del programa que se utilice, también lo son.

De los problemas que afectan a toda base de datos solamente vamos a ocuparnos del primero de ellos: el de los posibles erro-

res. Una primera solución sería realizar un análisis exploratorio de la base de datos que permitiera eliminar, al menos, parte de la información espuria y, seguidamente, proceder con el análisis estadístico. Incluso así quedarían sin detectar muchos errores: unos porque su existencia solamente se revela al modelar los datos y otros porque no se pueden considerar como observaciones anómalas ni son realmente errores; son observaciones que proceden de un modelo distinto, generalmente más complejo que el inicialmente contemplado para los datos.

El distinguir entre estas dos situaciones no es taréa fácil, pues existe una clara interdependencia; lo que es una observación anómala para un modelo relativamente sencillo puede no serlo para otro modelo distinto generalmente más complejo, pero podría serlo para ambos si se tratase efectivamente de un error. Si el número de observaciones anómalas es pequeño será preferible explicar los datos mediante un modelo simple, apelando al principio de economía o parsimonia del escolástico medieval Guillermo de Ockham «*non sunt multiplicanda entia praeter necessitatem*» (no hay que multiplicar los entes más allá de lo que sea necesario). En caso contrario hay que recurrir a procedimientos de clasificación que permitan separar las observaciones anómalas del resto o considerar un modelo más complejo.

En cualquier caso, queda de manifiesto la naturaleza subjetiva de las observaciones anómalas, y su condición de tales está condicionada al modelo que se vaya a utilizar para explicar los datos y, a posteriori, a las técnicas o criterios que para su detección se empleen.

Ahí reside la dificultad de definir lo que se entiende por una observación anómala, *outlier* en inglés. Ninguna de las definiciones propuestas parece ser plenamente satisfactoria ni elude el carácter subjetivo del concepto. Barnett y Lewis, en la segunda edición de su libro de 1984, dan como definición de observación o subconjunto de observaciones anómalas de un conjunto de datos aquellas que parecen ser inconsistentes con el resto de los datos. Esta definición no hace sino traspasar la vaguedad del concepto a la frase igualmente imprecisa de «parecer inconsistentes».

Los métodos que se han propuesto para la detección de observaciones anómalas no distinguen, en principio, entre éstas y lo que realmente constituyen errores, pues, como ya hemos comen-

tado, la distinción a priori entre ambos depende no sólo del criterio sino del modelo estadístico empleado. Una vez detectadas las observaciones sospechosas de ser anómalas, se puede contrastar si efectivamente son errores de transcripción de los datos originales o, realmente, se trata de datos verdaderos, lo que exigiría, en este caso, la reconsideración del modelo estadístico empleado.

En muchas ocasiones la decisión de rechazar una observación anómala ha de hacerse en el mismo momento de su detección. Esto ocurre con frecuencia en sistemas *on line* donde los datos van llegando uno a uno o en lotes pequeños, como ocurre, por ejemplo, tras el cierre de los colegios electorales en unas elecciones, donde los datos van llegando secuencialmente y pueden contener errores de diversa índole: de recuento, de transcripción, etc. La decisión de incluir o no estos datos puede sesgar las inferencias, en particular las estimaciones o predicciones y ha de realizarse teniendo en cuenta los datos recogidos hasta el momento y también la experiencia acumulada en el pasado o la información obtenida mediante encuestas realizadas a la salida de los colegios electorales.

Es éste un ejemplo particularmente importante donde los métodos bayesianos ofrecen soluciones satisfactorias al difícil problema de amalgamar diversos tipos de información provenientes de fuentes heterogéneas.

Mi opinión personal sobre este tipo de problemas de carácter secuencial, en los que hay que realizar inferencias y tomar decisiones sobre la marcha contando únicamente con la información pretérita, es la de desarrollar métodos robustos que acomoden y, a la vez, detecten las observaciones anómalas. De un modo genérico no muy preciso, por métodos o procedimientos de acomodación de observaciones anómalas se entienden aquellos que no eliminan las observaciones bajo sospecha sino que las incorporan al modelo pero a cambio de darles menor peso o importancia que al resto de las observaciones con el fin de que el procedimiento sea robusto.

El estudio de los problemas teóricos que plantean estas cuestiones desde el punto de vista bayesiano se realiza a través de las propiedades del operador de Bayes, entendiendo por tal el operador que transforma el espacio producto del espacio de las distribuciones a priori y el espacio de los modelos estadísticos en el espacio de las distribuciones a posteriori o de las distribuciones

predictivas si se está interesado en la robustez de las predicciones.

De esta formulación general se pueden obtener, por ejemplo, resultados cualitativos y cuantitativos sobre la robustez respecto de la distribución a priori utilizando los resultados de Diaconis y Freedman, que ya mencionamos en la introducción, sobre la diferenciabilidad del operador de Bayes respecto de la distribución a priori. Seguramente, el estudio de esta derivada, en particular la derivada de Gâteaux, que enlaza directamente con la idea de contaminación a través de modelos de mixtura, será una de las áreas de investigación más prometedoras en el estudio de la sensibilidad global, que además enlaza con las ideas clásicas de robustez y medidas de influencia de la escuela de Huber.

Otros resultados importantes —en otra línea totalmente distinta de la anterior pero complementaria, debida principalmente a J. O. Berger— se refieren a la robustez de la distribución a posteriori o de ciertos procedimientos estadísticos respecto de amplias clases de distribuciones a priori, que además conducen a resultados cuantitativos sobre las probabilidades a posteriori o sobre ciertas características de ellas generalmente expresados en forma de intervalos o acotaciones.

El estudio de la sensitividad o robustez bayesiana respecto de todos los ingredientes de un problema de decisión con información a priori, es decir, la distribución a priori, el modelo y la función de pérdida, que Leamer denomina *análisis de la sensibilidad global*, parece más complicado aunque ya hay ciertos resultados recientes en esta dirección.

Volviendo de nuevo al aspecto práctico del tratamiento de los problemas de robustez y observaciones anómalas, muchos de los modelos clásicos tanto univariantes como multivariantes, sobre todo los que de algún modo presentan una estructura lineal subyacente, pueden reformularse en términos del filtro de Kalman, por lo que vamos a concentrarnos en él.

Como ya hemos visto, el filtro de Kalman con fuentes de error gaussianas, proporciona una técnica muy general para tratar una gran variedad de modelos estadísticos estructurados, como la regresión, los modelos dinámicos y una clase amplia de series temporales. Pero no es robusto salvo para la clase de las distribuciones esféricas, y ésto en un sentido débil, por lo que se han realizado muchos intentos de robustecerlo.

Desde el punto de vista bayesiano el robustecimiento del filtro puede conseguirse mediante la consideración en el modelo de fuentes de error que sean *resistentes* a las observaciones anómalas, como las que se distribuyen según una t de Student, o bien utilizar familias de distribuciones paramétricas más generales que la normal, como, por ejemplo, la famila exponencial de potencias de Box y Tiao, o bien suponer que el mecanismo que genera las observaciones anómalas puede ser representado por una mixtura de dos distribuciones normales, una que correspondería a la fuente de error de la que provendrían la mayoría de las observaciones y otra que generaría las observaciones anómalas debidas bien a un cambio de nivel o desplazamiento del error o a una mayor dispersión.

Todos estos modelos para la estructura de los errores del filtro pueden considerarse como casos particulares de mixturas finitas e infinitas de distribuciones normales. Para esta clase de errores las ecuaciones del filtro no son válidas por lo que se hace necesario recurrir a métodos aproximados. La idea central es intentar conservar la forma recursiva del filtro mediante aproximaciones en cada etapa usando la estructura jerárquica de las mixturas de distribuciones normales, lo que produce filtros aproximados que además de ser robustos y que acomodan de un modo automático a las observaciones anómalas, permiten la identificación de éstas.

Como una ilustración de estas técnicas, consideremos los datos de la Tabla siguiente (Tabla 1) que representan los pesos de los cerebros (medidos en gramos) y de los cuerpos (medidos en kilogramos) de 28 animales. Esta tabla está tomada del libro *Robust Regression & Outlier Detection*, de P. J. Rousseeuw y A. M. Leroy, publicado en el año 1987 (pag. 57). Con estos datos se pretende investigar si existe una correlación positiva entre el tamaño del cerebro y del cuerpo.

La Figura 1 muestra los datos de la tabla expresados en una escala logarítmica, con lo que se consigue que los datos transformados presenten una estructura más lineal y, a la vez, desaparezca también la heterocedasticidad presente en los datos originales. En ella se muestran, además, la recta de regresión mínimo cuadrática (línea discontinua) y la recta obtenida mediante la aplicación de un modelo de regresión robusto (línea de trazo continuo), basado en una distribución t de Student con un grado de libertad para los

Tabla 1. Pesos del cuerpo y del cerebro de 28 animales

Índice (i)	Especies	Peso cuerpo (Kg.) (x_i)	Peso cerebro (g.) (y_i)
1	Castor	1.350	8.100
2	Vaca	465.000	423.000
3	Lobo gris	36.330	119.500
4	Cabra	27.660	115.000
5	Conejo de Indias	1.040	5.500
6	Diplodoco	11700.000	50.000
7	Elefante asiático	2547.000	4603.000
8	Burro	187.100	419.000
9	Caballo	521.000	655.000
10	Mono capuchino	10.000	115.000
11	Gato	3.300	25.600
12	Jirafa	529.000	680.000
13	Gorila	207.000	406.000
14	Hombre	62.000	1320.000
15	Elefante africano	6654.000	5712.000
16	Triceratops	9400.000	70.000
17	Macaco rheso	6.800	179.000
18	Canguro	35.000	56.000
19	Hámster	0.120	1.000
20	Ratón	0.023	0.400
21	Conejo	2.500	12.100
22	Oveja	55.500	175.000
23	Jaguar	100.000	157.000
24	Chimpancé	52.160	440.000
25	Braquiosaurio	87000.000	154.500
26	Rata	0.280	1.900
27	Topo	0.122	3.000
28	Cerdo	192.000	180.000

errores, utilizando los métodos aproximados que hemos comentado al final de esta sección.

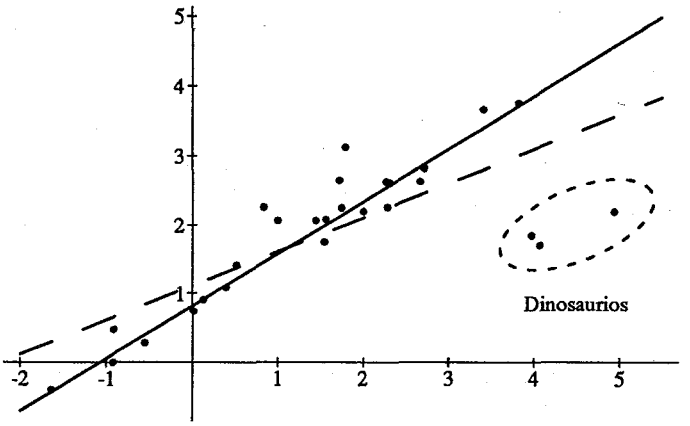

Figura 1. Logaritmos de los datos de la Tabla 1.

La recta de regresión mínimo cuadrática viene dada por la ecuación

$$\log y = 0.495995 \log x + 1.10958,$$

mientras que la obtenida por el método robusto, utilizando el filtro de Kalman aproximado y una distribución inicial no informativa, es

$$\log y = 0.814433 \log x + 0.758945.$$

La ecuación anterior se obtuvo aplicando el filtro aproximado a los datos en el orden en que aparecen en la tabla. Sin embargo, conviene destacar que la solución robusta *exacta*, basada en la moda de la verdadera distribución a posteriori, utilizando también la distribución no informativa usual, viene dada por

$$\log y = 0.82305 \log x + 0.760081,$$

que prácticamente no se diferencia, para estos datos, de la aproximada.

La aplicación rutinaria de un análisis de regresión estándar de estos datos, tal como se deduce del examen del gráfico de los residuos ordinarios (Figura 2a), no revelaría la existencia de observaciones anómalas . Éstas se hallan enmascaradas por la sobreestimación de la varianza del modelo y también por el bajo coeficiente de determinación R^2 del modelo basado en los mínimos cuadrados ordinarios, 0.608. Lo que induciría a pensar que el modelo lineal no es adecuado para explicar la relación existente entre los logaritmos de los pesos del cerebro y del cuerpo en los mamíferos. De todos modos, se observa que las observaciones 6 y 25 podrían considerarase como anómalas por estar en el umbral de la banda de confianza al 95%.

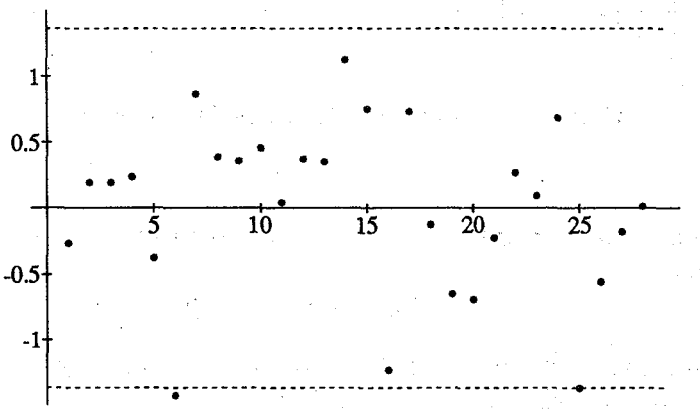

Figura 2a. Residuos ordinarios mínimo cuadráticos.

En cambio, el examen de los residuos del modelo robusto (Figura 2b) revela, inmediatamente, la existencia de un conglomerado de datos claramente diferenciado del resto, correspondiente a las observaciones 6, 16 y 25 (con residuos muy negativos), y que son responsables de la baja estimación mínimo cuadrática de la pendiente de la recta de regresión. Curiosamente, estas observaciones corresponden a los tres dinosaurios, que tienen un cerebro pequeño, comparado con el tamaño de su cuerpo, si se comparan con el resto de los mamíferos. También, aunque de forma no tan

acusada, se detecta que las observaciones 14 y 17 (que tienen residuos positivos), a saber, el macaco rheso y el hombre, son también anómalas. En este caso, el peso del cerebro es superior al predicho por el modelo lineal. Si eliminamos estos casos anómalos, el resto de las observaciones se ajustan muy bien al modelo lineal.

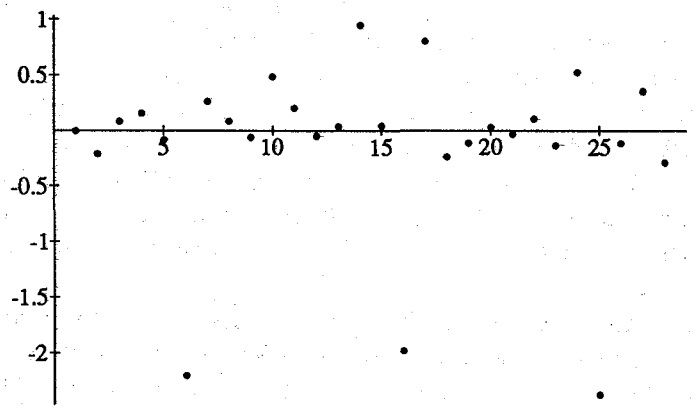

Figura 2b. Residuos ordinarios del modelo robusto.

La conclusión del análisis de estos datos mediante un modelo de regresión robusto, suponiendo que los errores son independientes e idénticamente distribuidos y siguen una distribución de Cauchy (t de Student con 1 grado de libertad) es que los dinosaurios, y en menor medida el hombre y el macaco rheso, no obedecen al mismo modelo que la mayoría de los otros mamíferos. No hay necesidad de eliminar estos datos del análisis de regresión —como suele hacerse en los enfoques clásicos, volviendo a calcular la nueva regresión con los datos restantes. En este ejemplo, el modelo de regresión robusta acomoda y detecta, automáticamente, las observaciones anómalas. Si se compara la nueva ecuación de regresión

$$\log y = 0.750872 \log x + 0.869174,$$

calculada eliminando las observaciones 6, 14, 16, 17 y 25, con la robusta, no se observan diferencias apreciables entre ambas rectas, en el rango de variación de las variables.

MUESTREO BAYESIANO

Ha sido en la última década cuando han aparecido varios métodos orientados a facilitar el cálculo de distribuciones a posteriori en problemas complejos basados todos ellos en variantes del método de Monte Carlo, y que evitan los complejos problemas numéricos que aparecen al analizarlos desde un punto de vista bayesiano puramente analítico. Estos métodos constituyen una alternativa a las técnicas de integración numérica en espacios de gran dimensión, desarrolladas por el grupo de Nottingham, y a las aproximaciones analíticas para el cálculo de distribuciones a posteriori o ciertas características de éstas, obtenidas recientemente por Kass, Tierney y Kadane, utilizando variantes y extensiones del clásico método de Laplace sobre el desarrollo asintótico de integrales.

Dos de ellos, el algoritmo de substitución y el muestreo de Gibbs, son de carácter iterativo mientras que el algoritmo *SIR* de Rubin, basado en el muestreo por importancia, es un procedimiento no iterativo que constituye una alternativa a las técnicas de integración basadas en el método de Monte Carlo. Todos estos procedimientos son de aplicación inmediata en muchos de los problemas que frecuentemente ocurren en la práctica, por lo que estas técnicas merecerían ser conocidas y ensayadas en un número mayor de problemas.

Sobre este tipo de técnicas, Rubin en 1985 aseguraba que «...teniendo en cuenta la capacidad de cálculo de hoy día, es a menudo deseable realizar inferencias en problemas aplicados utilizando técnicas de simulación, debido a la gran flexibilidad que resulta de poder adecuar datos a modelos sin tener que depender de complejos, y a veces indirectos, análisis matemáticos».

El influyente artículo de Tanner y Wong de 1987 parte de una idea que ya se había utilizado previamente en el algoritmo *EM*, en un contexto no bayesiano orientado a calcular estimadores de máxima verosimilitud, que es la de aumentar los datos observados mediante la introducción de *datos latentes*, lo que en muchas ocasiones permite analizar el problema de un modo más sencillo. La importancia del artículo consiste en extender esta idea a la inferencia bayesiana permitiendo el cálculo de la distribución a posteriori de los parámetros de interés.

El procedimiento es simple: los datos observados se aumentan —de ahí el nombre que recibe el método— añadiendo otros llamados latentes, de modo que, si ambos fuesen conocidos, el problema de calcular la distribución a posteriori del parámetro condicionado a los datos aumentados fuese fácil de resolver. Lo que interesa, sin embargo, es la distribución a posteriori del parámetro condicionada a los datos originales, que en general es difícil de calcular. El análisis de las diversas distribuciones condicionadas involucradas revela la dependencia y dualidad mutua entre la distribución a posteriori que se quiere calcular y la distribución predictiva de los datos latentes, lo que a su vez proporciona un algoritmo iterativo para calcular la verdadera distribución a posteriori como solución de una ecuación integral.

Desde un punto de vista analítico el algoritmo es equivalente a los métodos de substitución sucesiva para encontrar los puntos fijos de un operador integral. Aprovechando resultados estándar de la teoría de los operadores integrales se demuestra la unicidad de la solución, la convergencia monótona del proceso iterativo hacia la verdadera densidad en la norma de L_1 y que la tasa de convergencia es geométrica.

Para que el algoritmo sea eficiente es necesario que la obtención de muestras de las distribuciones del parámetro condicionadas por la muestra aumentada y de los datos latentes condicionados al parámetro y los datos originales sea inmediata; es decir, que se disponga de algoritmos de simulación eficientes, lo que suele ocurrir en muchos casos.

El llamado muestreo de Gibbs, introducido por Geman y Geman en 1984, ya había sido empleado en problemas de procesado de imágenes, en redes neuronales y en sistemas expertos, pero su potencialidad para tratar problemas estadísticos convencionales parece haber pasado desapercibida. El muestreo de Gibbs se basa en el resultado bien conocido de que las distribuciones de cada variable condicionadas por el resto caracteriza la distribución conjunta de todas ellas. Es por consiguiente un procedimiento de aprendizaje markoviano en el que en cada ciclo o etapa se generan variables aleatorias de las distribuciones da cada variable condicionadas por las demás, partiendo de unos valores iniciales que, bajo condiciones muy débiles, convergen débilmente hacia la distribución marginal. Utilizando la norma del supremo en vez

de la del espacio L_1, también se demuestra la convergencia del procedimiento iterativo resultante hacia la verdadera densidad, y que la tasa de convergencia es también geométrica. En este caso, además, se obtiene una versión del teorema ergódico.

La principal diferencia entre ambos métodos es que el muestreo de Gibbs no está en principio pensado para datos incompletos, como ocurre con el método de Tanner y Wong, y a diferencia de éste, es completamente simétrico respecto de los parámetros o variables aleatorias que aparecen en el modelo, lo que en algunas circunstancias resulta más práctico, aún a costa de introducir más aleatoriedad en el proceso iterativo y, a veces, decelerar la tasa de convergencia.

Recientemente, Gelfand y Smith, en 1990, han generalizado estos resultados, en particular el método aumentativo de Tanner y Wong que ellos denominan de substitución, estableciendo la relación que existe entre éllos y lo han aplicado al cálculo de distribuciones a posteriori para un gran número de problemas que incluyen, entre otros, modelos con datos incompletos, modelos jerárquicos y modelos de componentes de la varianza.

El algoritmo *SIR* de Rubin o de *muestreo seguido de remuestreo por importancia* es una técnica de tipo no iterativo que se aplica a la simulación de muestras de distribuciones complicadas. Su nombre se debe a que inicialmente se obtiene una muestra grande de una distribución parecida a la original que sea fácil de simular y, a continuación, de esta muestra se extrae una submuestra con probabilidades de extracción aproximadamente proporcionales a los cocientes de las dos densidades evaluados en cada elemento de la primera muestra, cocientes a los que Rubin, siguiendo la terminología usual, denomina *razones o cocientes de importancia*. La ventaja de este método es que puede ser aplicado en muchas circunstancias aunque para algunos casos particulares pueda resultar ineficiente. Esto último depende de la elección que se haga de la aproximación a la distribución original, lo que a su vez también influye en el método de remuestreo, que puede ser más conveniente elegirlo con o sin reemplazamiento.

Desde la perspectiva de su aplicación a la inferencia bayesiana, el algoritmo resulta ser más útil cuando exista una buena aproximación a la verdadera distribución a posteriori de la que sea fácil extraer muestras. Esta situación se presenta habitualmente en

problemas análogos a los que originaron el algoritmo de substitución; es decir, cuando la distribución a posteriori del parámetro condicionada por todos los datos —los observados y los que faltan, o no observados— sea simple. Sin embargo, el algoritmo de Rubin es más general y se puede aplicar a problemas más complicados como, por ejemplo, la estimación de los parámetros de un modelo de regresión logística con muestras de tamaño pequeño.

La gran ventaja de estos métodos, en contraposición con las sofisticadas técnicas numéricas que se han desarrollado para tratar estos problemas, es que son conceptualmente sencillos y además fáciles de aplicar por aquellos usuarios que dispongan de recursos de cálculo eficientes.

COMENTARIOS FINALES

Llegados a este punto, es hora de poner fin a este discurso en el que no hemos pretendido hacer una apología del bayesianismo sino mostrar aquellas contribuciones y aspectos realmente importantes y positivos que la estadística bayesiana ha aportado a la ciencia estadística en general.

Los temas que hemos seleccionado y comentado en este discurso tienen todos conexión con aspectos relevantes de la teoría estadística clásica. No hay exclusivismos: las buenas ideas de cada enfoque suelen tener su contrapartida en el otro, seguramente porque detrás de todo ésto están los teoremas que garantizan la optimalidad de las reglas de Bayes desde un punto de vista clásico o frecuentista, aunque el recíproco no sea generalmente cierto: hay procedimientos clásicos que son inadmisibles, como ocurre, por ejemplo, con el problema estadístico que dió origen a la paradoja de Stein. Otros procedimientos clásicos, como muchos de los que habitualmente se emplean en la práctica, son numéricamente —aunque no conceptualmente— equivalentes a los bayesianos. Creo, sin embargo, que hay que abandonar ciertos prejuicios de la inferencia clásica en favor de la postura coherente o bayesiana, como ya comentamos con ocasión de los contrastes de hipótesis.

Por otro lado, queda claro que, para poner en práctica muchos de los procedimientos bayesianos, hay que desarrollar y me-

jorar las técnicas de asignación de probabilidades subjetivas, cuestión particularmente importante sobre todo en lo que se refiere a aquellos individuos que han de tomar decisiones relevantes: políticos, médicos, abogados, etc.; y en aquellos problemas reales, que resultan complejos de modelar, en los que intervienen muchos parámetros. Como ya hemos comentado, los modelos jerárquicos suelen ofrecer una solución satisfactoria a este tipo de problemas, si se aplican con buen criterio.

Pero ésto en la práctica suele presentar graves problemas. Aunque vivimos en un mundo donde la incertidumbre y el azar están constantemente presentes en la mayoría de las actividades humanas, no estamos acostumbrados a pensar en términos probabilísticos, ni tenemos constancia real de la escala de las mismas, sobre todo cuando se trata de probabilidades muy pequeñas o muy próximas a la unidad. Así como tenemos una idea relativamente precisa de lo que son las distancias, pesos y medidas y disponemos de instrumentos para su medición, en cambio no tenemos la misma capacidad de apreciar o medir probabilidades o riesgos, que parece ser una tarea más difícil. Creo que efectivamente lo es, pero, como en otros campos de la actividad humana, no sólo hay que aprender a desarrollar métodos de medición adecuados sino también enseñarlos.

A todos, muchas gracias por su atención.

DISCURSO DE CONTESTACIÓN

DEL

EXCMO. SR. D. SIXTO RÍOS GARCÍA

Excmo. Sr. Presidente,
Excmos. Sres. Académicos,
Señoras y Señores:

Gran honor ha sido para mi que esta Real Academia me encargara la contestación al Discurso de ingreso de mi querido discípulo Francisco Javier Girón, santanderino de nacimiento, madrileño en su juventud y aposentado hace años en Málaga como catedrático, impulsor eficaz de los estudios estadísticos.

Viene a ocupar la vacante académica de nuestro inolvidable compañero Francisco Azorín, especialista de nivel internacional en muestreo, que a pesar de su breve paso por la Academia, realizó contribuciones notables a nuestro Vocabulario Científico y a nuestra Revista.

Han sido ciertamente desgraciados estos dos últimos años para la Sección de Exactas, pués, además del fallecimiento de Francisco Azorín, hemos tenido otras tres importantes pérdidas de eminentes Académicos: Felipe Lafita, Federico Goded y Enrique Linés. Cumplido el triste deber de recordarlos, deseemos larga vida, plena de éxitos científicos, a los nuevos colegas Montesinos, Liñán, Jiménez Guerra y Girón; elegidos para dar continuidad y brillo a nuestra tareas académicas.

Javier Girón cursó el bachillerato como brillante alumno de los escolapios, siempre muy interesado por los experimentos físico-químicos, y su inclinación por las matemáticas surgió pronto gracias, en gran parte, a las enseñanzas de su profesor, D. Pedro García Varona, que supo inculcarle el gusto por la resolución de problemitas adecuados a su edad. Tras ser un excelente estudiante de la Licenciatura en Ciencias Matemáticas en la Complutense,

inició su especialización en Investigación Operativa en 1967, trabajando con Becas de IBM, de la Fundación March, etc., desarrollando trabajos en la *Teoría de la Decisión*, en la que yo le inicié y siendo especialmente destacable su actividad en el seminario sobre *Comprobación experimental de teorias normativas de la decisión* que realizamos en 1975–76, bajo el patrocinio del Fondo Nacional para el Desarrollo de la Investigación Científica.

Su tesis de Doctor en Ciencias Matemáticas fué publicada en la Revista de esta Academia (que dicho sea de paso, tiene para sus artículos de matemáticas un nivel similar al que marcan Revistas internacionales incluidas en la reciente lista seleccionada para la calificación de Profesores universitarios). La tesis logra, en una época temprana del desarrollo del bayesianismo, una interesante caracterización axiomática de la regla de Bayes y la probabilidad subjetiva, en respuesta al tema que le propuse y que desarrolló en buena parte como Honorary Research Fellow en el University College de Londres. Aquel ambiente magnífico, creado a la sombra de Egon Pearson, por entonces ya jubilado, pero que conservaba y frecuentaba su despacho del College, contribuyó, junto con las enseñanzas de los Lindley, Birnbaum, etc., a conformar de un modo muy efectivo la personalidad futura de Girón.

Algunas de sus aficiones a la lectura, a la música (es actualmente maestro cantor de la Coral *Santa María de la Victoria de Málaga*, que pronto cantará delante del Papa), probablemente se reafirmaron durante su estancia en Londres.

En el curso 1974–75 se incorpora nuevamente a la docencia e investigación, ya como Profesor Adjunto del Departamento de Estadística e Investigación Operativa de la Universidad Complutense y continúa sus trabajos en teoría de la decisión, tema fundamental sobre el que pronto inicia la dirección de algunas tesis de gran calidad, contribuyendo de este modo a la fama y prestigio de la generalmente llamada *Escuela de Madrid*.

En 1977 obtiene la plaza de Profesor Agregado de Estadística Matemática de la Universidad de Málaga, en la que encuentra un ambiente de vida, trabajo y amor que le satisfacen y gratifican y le llevan a no aceptar otras proposiciones de trabajo ofrecidas por algunas Universidades españolas y extranjeras. En Málaga parte casi de cero, con un Departamento recién creado y poco a poco lo va nutriendo de Doctores y colaborando con otros Departamen-

tos españoles, marroquíes, franceses e ingleses, constituyendo el más destacado conductor de una escuela de Estadística digna hija, como algunas más, de la Escuela de Estadística de la Universidad Complutense, en que se iniciaron sus cabezas directivas.

Sus excelentes trabajos y los de sus numerosos discípulos (Criado, Prieto, Domínguez, Imlahi, Ríos, Caro, Martínez, etc.) continúan estudiando nuevas orientaciones, fundamentalmente de tipo axiomático, en la teoría de la decisión y la inferencia, análisis de conglomerados, mixturas de distribuciones, filtro de Kalman, robustez, etc. Pero me permito destacar entre sus trabajos personales el que presentó, en colaboración conmigo, al Coloquio bayesiano Internacional de Valencia (1979). Este trabajo profundiza mi comunicación de 1975 al Congreso Internacional de Varsovia, en la que introduje la idea de edificar una teoría de la inferencia con probabilidades supuestamente imprecisas. Más concretamente, adoptamos el punto de vista de estimar las probabilidades por conjuntos convexos, partiendo de una axiomática de preferencias, generalización natural de la bayesiana. Su interés y éxito se han confirmado recientemente al publicar Peter Walley un libro de más de 600 páginas con el título *Statistical Reasoning with Imprecise Probabilities*, en que desarrolla ampliamente dicha metodología, que ha resultado expecialmente adaptable al progreso de la Inteligencia Artificial y los Sistemas Expertos.

Estos y otros resultados suyos son citados y utilizados en libros, monografías y trabajos de Berger, Kotz, Johnson y French, que dan una idea de su alcance e interés general. Y en gracia a la brevedad, no voy a hacer referencia detallada de sus meritos de pertenencia a Consejos Directores de Revistas Internacionales, de Vicepresidente de la *Sociedad Española de Estadística e Investigación Operativa*, de sus contratos de investigación para trabajos prácticos de su equipo dedicados al estudio de la delincuencia en la Costa del Sol, de su actividad como organizador de Congresos, etc.

Como Girón ha sido un frecuente colaborador de nuestra Revista —antes de ser Académico correspondiente y también después, simultaneando tal labor con su amplia contribución al Vocabulario Científico, ya desde su primera edición— es natural esperar que continúe y redoble su actividad en tales tareas, así como en los trabajos multidisciplinares que proyecta para un futuro próximo nuestro Presidente. Es decir, que no es difícil supo-

ner que el nuevo Académico, a pesar de su distancia geográfica a nuestra sede, dedicará una buena parte de su tiempo a colaborar al brillo de nuestra Academia y también logrará, a través de sus trabajos universitarios en Málaga, contribuciones que sean origen de una efectiva participación regional andaluza en nuestras actividades.

Tras esta breve exposición de elogios académicos pasemos a hacer algunos comentarios sobre el magnífico discurso del nuevo Académico relativo a conceptos, ideas y resultados actualmente vigentes del problema de la inferencia probabilística, tema que es central en la Estadística y la Filosofía de la Ciencia y que juega un papel importante en la Economía, Psicología, Investigación Operativa, Inteligencia Artificial, etc.

Nuestras notas y observaciones van a tener un carácter principalmente histórico-filosófico y crítico, proyectadas a perspectivas futuras de las aplicaciones a la Ingeniería del Conocimiento.

La Filosofía de la Ciencia viene preocupándose hace más de dos mil años del problema de la inducción, como base para llegar a un método científico general que permita realizar inferencias inductivas, es decir, procesos discursivos, que partiendo de una proposición particular permitan llegar a otra más general.

Es sabido que en la obra aristotélica se distingue ya la inferencia deductiva, regulada por leyes lógicas bien establecidas, de la inferencia inductiva, que no resulta con certeza de las premisas y cuya validez puede cambiar al variar la información asociada a las mismas. Es decir, la inferencia inductiva supone incertidumbre, lo que es característico de las conclusiones científicas, así como de las decisiones del hombre educado en su actividad diaria. Como ha dicho el matemático Polya «... estrictamente hablando, todo nuestro conocimiento, aparte de las matemáticas y la lógica demostrativa, consiste en conjeturas...».

«Existe una diferencia clara entre el razonamiento deductivo, que permite el conocimiento matemático y el conocimiento plausible, que comprende la evidencia inductiva del físico, la evidencia circunstancial del abogado, la evidencia documental del historiador, la evidencia estadística del economista, del sociólogo, etc...».

Inferencias de la muestra a la población, de los datos a la hipótesis, de los efectos observados a las causas inciertas, del pa-

sado al futuro, son aspectos de la inducción con que continuamente nos enfrentamos en nuestra actividad de conocer y decidir y que, muy frecuentemente, no va asociada a fenómenos aleatorios de carácter repetitivo.

Dos puntos de vista adoptados para abordar estos problemas son el de los filósofos y el de los estadísticos. El primero representa un intento de identificación de las características universales de las inferencias inductivas para llegar a una *Teoría Magna* como justificación general de los métodos inductivos. Los estadísticos se restringen a problemas inductivos más concretos, logrando, con modelos matemáticos apropiados, establecer reglas de inferencia en condiciones bien especificadas.

Refiriéndonos al enfoque filosófico diremos que con precursores como Avicena (980–1037), Roger Bacon (1214–1294) y otros posteriores, llegamos a Francis Bacon (1561–1626) con su *Novum Organum*, en que inicia la ruptura con los métodos aristotélicos deductivos y continúa con John Stuart Mill (1806–1873), introductor de una metodología inductiva, cuyo propósito es establecer relaciones de causa a efecto en un proceso dialéctico en que, a partir de un conjunto potencialmente infinito de observaciones, se obtendrían unas primeras afirmaciones inductivas, que se someterían a comprobación. Si se confirmaban por las nuevas experiencias, permitirían llegar a teorías científicas sucesivamente perfeccionadas. En él aparecen ya las ideas iniciales de los métodos eliminativos de Popper, más tarde tan en boga, y a los que luego nos referiremos.

Pero antes es obligado señalar la contribución de David Hume, que define por primera vez claramente en su *Treatise of Human Nature (1739-40)*, «el problema de la inducción» en la siguiente forma: «Cuando se pasa de lo observado O a lo inobservado I, O e I son lógicamente distintos, al menos en el sentido de que se puede concebir O como evidente, mientras I no. En consecuencia, no existe necesidad lógica de que I se siga de O. ¿Cuáles podrían entonces ser los fundamentos para afirmar I dado O?». En definitiva, plantea el problema de dar una justificación racional a las inferencias inductivas. Su postura empiricista le lleva a negar que tal problema pudiera tener solución. Pero como dice muchos años después Bertrand Russell «es necesario elegir entre la inducción con su irracionalidad relativa y la irracionalidad absoluta».

Las respuestas filosóficas al problema han sido de dos tipos. Las primeras reconocen que las inferencias inductivas son injustificables y que no deben figurar en ningún libro científico.

En compensación tratan estos filósofos (Popper, Kuhn, Feyerabend, etc.) de auxiliar al científico con sus métodos de falsación o refutación, que se relacionan con los métodos estadísticos de los contrastes de hipótesis (Fisher, Neyman y Pearson), como término de una línea de trabajos que se inicia en Bacon y Mill, y que un análisis detallado permite descubrir en algunas de sus fases el empleo tácito de trampas inductivas.

Se ocupan los filósofos del segundo grupo en justificar racionalmente las inferencias inductivas, bien tratándolas como argumentos deductivos incompletos, bien agregando algún principio de carácter general como el de uniformidad de la Naturaleza.

Hay que reconocer que en esta línea no se avanza mucho en tiempos recientes, a pesar del enfoque pragmático de Reichenbach (1949) y de las aportaciones de Keynes que, en su *A Treatise on Probability (1921)*, introduce la probabilidad lógica o necesaria que se interpreta como una función numérica para medir el grado de confirmación inductiva de una hipótesis, dadas ciertas proposiciones de evidencia. Hoy se considera fallido este ambicioso intento de reconstrucción y formalización racional de la realidad científica, a pesar de los importantes trabajos del filósofo Carnap (1950), cuyo complicado sistema de axiomas de simetría, invariancia, etc., construido para poder manejar simultáneamente probabilidades lógicas y frecuentistas, fue pronto abandonado. Junto con Keynes y Jeffreys contribuyó a resucitar la antigua línea bayesiana de trabajos que pueden considerarse hoy como el primer tratamiento frontal con éxito de la inferencia inductiva.

Thomas Bayes, clérigo presbiteriano, que vivió en Inglaterra entre 1702 y 1761, cultivó con profundidad el Cálculo de Probabilidades, legado de Pascal y Fermat, dejando entre sus papeles un sencillo teorema que fué publicado dos años después de su muerte, como trabajo póstumo, en las *Philosophical Transactions* de la *Royal Society (1763)*. Bayes, que ni siquiera pretendió publicar su teorema, probablemente nunca imaginó que 200 años después un eminente estadístico inglés, el Prof. D. V. Lindley, escribiera: «Es difícil encontrar un trabajo que contenga ideas tan importantes y originales como el de Bayes. Su teorema debía figurar al lado

de la fórmula de Einstein, $E = mc^2$, como una de las grandes y sencillas verdades».

Concretamente abordaba Bayes en su trabajo el llamado problema inverso de la probabilidad, que anteriormente J. Bernoulli, en su *Ars Conjectandi (1973)*, planteaba así: «Para ilustrar ésto con un ejemplo supongo que tengo una urna con 3000 bolas blancas y 2000 bolas negras, que alguien ha puesto allí y que yo, sin conocer tales números, trato de averiguar la proporción de bolas blancas y negras, mediante un experimento reiterado que consiste en sacar una bola al azar, observar su color y reponerla a la urna, etc.».

En definitiva el problema consiste en pasar de la información que da la realización de un cierto número de experimentos o muestra de resultados, al conocimiento, en forma probabilística, de la composición de la urna o población. Si llamamos θ a un parámetro que caracteriza el modelo, en este caso la composición desconocida de la urna, y x un conjunto de datos, en este caso la frecuencia observada, el teorema de Bayes permite realizar la inferencia, que consiste en pasar del dato particular x a la proposición general θ, mediante el cálculo de $P(\theta \mid x)$ a través de $P(\theta)$ y $P(x \mid \theta)$.

Este teorema, que Laplace redescubrió y aplicó muchos años después, constituye la base de la inferencia bayesiana, que se desarrolló y utilizó hasta la Primera Guerra Mundial, junto con otras metodologías, que se suelen denominar estimación y contraste de hipótesis.

Después surgieron los importantes trabajos de Fisher, que con su variada gama de técnicas, más fáciles de elaborar, arrinconaron las ideas bayesianas, a lo que contribuyen también los trabajos más formalizados de Neyman-Pearson y Wald, que campean a partir de los años treinta entre los estadísticos teóricos y aplicados.

No vamos a reiterar la relación de las etapas más recientes de la constitución del paradigma bayesiano: de Carnap y Keynes a Jeffreys, y de Ramsey, von Neumann y de Finetti a Savage, admirablemente expuestos en el discurso de Girón.

Y tras este breve comentario histórico voy a referirme especialmente al campo de aplicaciones a la Inteligencia Artificial y Sistemas Expertos, que a su vez han contribuido a encuadrar la inferencia probabilística en el poderoso marco preparado por el Cálculo de Probabilidades que, de acuerdo con la profecía de Bo-

ole, viene así a jugar un papel para la inferencia inductiva análogo al que el Álgebra de Boole representa para la inferencia deductiva.

Si 50 años después de la publicación del *Álgebra de Boole (1854)*, pudo decir Bertrand Russell que: «era el único libro de Matemáticas que conocía», y más tarde se vió su importancia en el diseño de ordenadores, algo análogo está ocurriendo en nuestros días con la inferencia bayesiana en relación con la Inteligencia Artificial y los Sistemas Expertos.

En efecto, los primeros pasos de los Sistemas Expertos, p. ej., el MYCIN (1976), se dieron tratando la incertidumbre como una generalización del valor de verdad, en que la certidumbre de una fórmula resultaba como una función única de las certidumbres de las subfórmulas que la componían. Por ejemplo en los llamados *sistemas intensionales* la certidumbre de la conjunción $A \cap B$ viene dada por una función fijada, concretamente el producto o el mínimo, de las medidas de certidumbre de A y B consideradas individualmente. Contrariamente, en los llamados *sistemas extensionales* se utiliza el Cálculo de Probabilidades, u otros cálculos inspirados en éste, para asignar medidas de certidumbre a conjuntos de universos, y éstos se combinan mediante las reglas de las operaciones de la teoría de conjuntos y las medidas se calculan con las reglas de las probabilidades. Así por ejemplo, $P(A \cap B)$ no depende solo de $P(A)$ y $P(B)$ sino también de la relación de dependencia entre A y B.

Desde el artículo de McCarthy y Hayes (1969) que consideraba las probabilidades epistemológicamente inadecuadas, muchos especialistas de Inteligencia Artificial han huido de las probabilidades justificando su actitud con frases como: «las probabilidades son difíciles de calcular», «el uso de las probabilidades requiere muchos datos», etc. Son fáciles de rebatir estas posiciones: quizá basta considerar con Shafer que en Inteligencia Artificial la contribución más importante de la probabilidad no es el aspecto numérico, sino la penetración en la estructura del razonamiento inductivo.

Cuando un médico afirma que un paciente con la enfermedad A debe desarrollar el síntoma B con la probabilidad p, lo importante de la afirmación no es tanto el valor numérico de p, como la razón específica de la creencia del médico, el contexto o suposiciones bajo las cuales la creencia debería ser firmemente

mantenida y las fuentes de información que podrían causar un cambio en esta creencia.

Pero por qué las creencias se combinarían como las frecuencias. A primera vista no hay razón que empuje a tratar las creencias o disposiciones mentales respecto a sucesos irrepetibles mediante las mismas reglas que se aplican a las frecuencias de los sucesos que se presentan en las pruebas repetitivas de los juegos de azar u otros experimentos aleatorios. Ciertamente la afortunada relación entre la intuición humana y las leyes de operación con las frecuencias no es una pura coincidencia, sino una consecuencia de que nuestras creencias están fuertemente influidas por la acumulación de experiencias que retenemos en nuestra mente en forma de promedios, frecuencias, pesos y relaciones cualitativas, que nos ayudan para futuras situaciones nuevas. Ésta es la explicación de que el cálculo con grados de creencia no puede ser diferente del cálculo con promedios y frecuencias, es decir, del cálculo de probabilidades convencional.

Ésto lleva a la construcción axiomática de un modelo probabilístico subjetivo mediante axiomas que han ido perfeccionándose a través de los trabajos de de Finetti, Kraft, Savage, Aumann, Machina, etc. Con ellos se llega finalmente a las mismas reglas de Kolmogorov, que permiten el cálculo de la probabilidad subjetiva de cada sentencia S, bien construida mediante las reglas de Boole, a partir de un conjunto de proposiciones atómicas A, B, C, ..., dadas en un cierto contexto provisto de una cierta información K. Sobre esta construcción axiomática se edifica el formalismo bayesiano para razonar en condiciones de incertidumbre.

Dos aspectos del Cálculo de Probabilidades que ha sido necesario estudiar con más profundidad a fin de responder a las demandas de la Inteligencia Artificial y la Teoría de la Decisión son las relaciones de dependencia y el cálculo numérico de las probabilidades iniciales de sucesos no necesariamente repetibles.

Como ha dicho Pearl (1989), un argumento para utilizar teoría de la probabilidad, aunque estemos realizando modelización psicológica, es que: «aunque los hombres difícilmente recuerdan las frecuencias exactas de los sucesos complejos, aprenden a pensar en términos de las dependencias e independencias implicadas por dichas frecuencias».

La manera tradicional de presentar la Teoría de Probabilida-

des partiendo de la función de distribución conjunta, no es la más adecuada cuando se trata de reflejar el esquema del razonamiento probabilístico, que sigue justamente el camino inverso: partir de juicios probabilísticos relativos a un pequeño número de proposiciones de significado intuitivo directo. Con el nuevo enfoque se trata de establecer una sucesión de dependencias que permitan tener en cuenta en los sistemas de Inteligencia Artificial las relaciones primitivas que se designan como verosimilitud comparada, condicionamiento, relevancia, causación, que adecuadamente establecidas, permiten utilizar la inferencia bayesiana para llegar a las probabilidades de las complejas proposiciones finales. Este estudio, que ha obligado a profundizar la axiomática de las dependencias, ha originado la introducción de los *grafos markovianos*, después superados por los *grafos bayesianos, diagramas de influencia*, etc., con sus teoremas debidos a Dawid, Lauritzen, Spiegelhalter, etc. Ellos permiten establecer un conjunto coherente de dependencias directas o indirectas que constituyen la esencia cualitativa de los modelos de bases de conocimientos.

Probabilidades por intervalos, lógica de razonamientos defectuosos, funciones de creencia de Dempster-Shafer, conjuntos difusos, medidas inexactas, lógicas no monótonas, etc., son otros muchos caminos que tratan de dar solución a los problemas planteados, en concurrencia con los métodos bayesianos. No es fácil hacer predicciones en este campo ni siquiera empleando métodos bayesianos; pero parece claro que siempre serán criticables los sistemas que no tengan en cuenta los esquemas de condicionamiento que ofrece la teoría de probabilidades moderna asociada a los grafos bayesianos y otros antes citados.

En cuanto a los problemas de asignación de probabilidades numéricas a sucesos iniciales, podemos decir que gracias, especialmente, a los trabajos de los psicólogos y estadísticos Edwards, Winkler, Kahneman, Tversky, Machina, Fishburn, etc., se van superando paradojas y tendencias, y estableciendo técnicas de medida, algo complicadas, pero efectivas, para resolverlos. Y otro tanto podemos decir de las funciones de utilidad para los problemas de decisión.

Pero volvamos al hilo del discurso del Prof. Girón, para referirnos a las amplias discusiones entre bayesianos y no bayesianos, a las que habría que agregar las no menos complicadas

entre filósofos y estadísticos, entre estadísticos e investigadores científicos, etc. Querría, en relación con éstas, comenzar con algunas palabras sobre la validación de modelos.

El problema de la validación de un modelo trata de ver hasta que punto el modelo propuesto es una representación fiel de la realidad, es decir, que las deducciones obtenidas del modelo, al traducirlas a la realidad, sean suficientemente aproximadas al fin a que se dedica el modelo. Si tratamos de aplicar estas ideas a los modelos de inferencia y decisión, parece que debíamos suponer una realidad objetiva que intentamos modelizar. Pero, ¿cuál es esta realidad? ¿Se trata quizá de alguna propiedad de nuestros cerebros que estamos tratando de revelar buscando alguna verdadera estructura de preferencia que queremos encontrar a pesar de nuestras limitaciones como instrumentos de medida?

Los puntos de vista que son naturales al validar modelos de las ciencias o las ingenierías no parecen adecuados en los modelos de inferencia o decisión. No se trata ahora de construir modelos descriptivos que reflejen el comportamiento real o experimental de los individuos, sino de *modelos normativos* que representen una afinada percepción de valores e incertidumbres en una estructura, que satisfaga ciertos principios de coherencia lógica o racionalidad que deseamos adoptar en nuestras decisiones racionales.

No puede, sin embargo, esta postura llevarnos a considerar alguna de las axiomáticas construidas como un dogma intocable y válido para la eternidad. Ni apoyarnos en una para atacar a las otras. Por ésto ha sido y continuará siendo útil el aprendizaje obtenido de los experimentos realizados con personas o sistemas intencionales, empresas, comités, etc., que toman decisiones y que han permitido y continuarán permitiendo perfeccionar las teorías de la inferencia y la decisión. Por ejemplo, los recientes trabajos de Fishburn, Machina, etc., y los del Premio Nobel Prof. Simon, que marcan nuevos caminos de perfeccionamiento o de rectificación del paradigma de la utilidad esperada de von Neumann.

Continúan pues, existiendo en este campo de la inducción y la decisión posibilidades de mejorar las soluciones y resolver nuevos problemas, aun limitándonos al caso de individuos racionales, pues no hemos hablado de conocimiento común, decisiones colectivas, decisiones con criterios múltiples, en concurrencia, etc.

Plácemes merece nuestro nuevo compañero por cultivar con

éxito parcelas de este fecundo y promisorio campo de la inferencia y la decisión, en que se van afincando también sus valiosos discípulos, a los que cariñosamente, me atrevo a considerar como nietos científicos. Sea pues, bienvenido a nuestra Casa el Profesor Girón y hónrela con sus trabajos por muchos años.