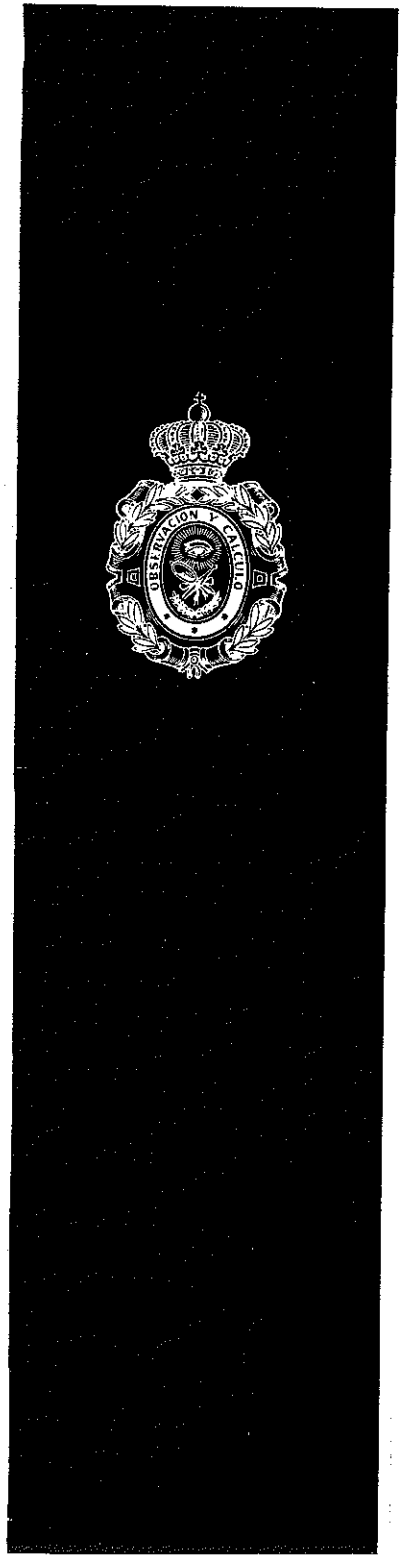

REAL ACADEMIA DE CIENCIAS
EXACTAS, FISICAS Y NATURALES

HISTORIA DE LA CIENCIA

HISTORIA DE LA CIENCIA ESTADISTICA

ANALISIS DE DATOS Y METODOS BAYESIANOS

por

José M. Bernardo

MADRID
1989

ANALISIS DE DATOS Y METODOS BAYESIANOS

JOSE M. BERNARDO *

1. INTRODUCCION

Desde sus comienzos, la práctica profesional de la estadística ha requerido una actitud interdisciplinar, en la que los conocimientos matemáticos del estadístico, la información sobre el significado real de los datos proporcionada por los científicos que los han producido, y los métodos de cálculo accesibles en cada momento histórico, se han combinado para proporcionar soluciones a los problemas de investigación y de decisión que se han ido planteando. Como consecuencia de este proceso, han ido apareciendo un conjunto de técnicas, no siempre consistentes entre sí, que constituyen el núcleo de la formación estadística tradicional.

En las últimas décadas, sin embargo, hemos asistido a una rápida evolución de la metodología utilizada, que se ha debido, fundamentalmente, a dos hechos importantes. Por una parte, el increíble desarrollo de los métodos de cálculo, con la aparición de potentes microordenadores de bajo coste, capaces de realizar de forma interactiva, y con soporte gráfico, cálculos que hace solo unos años eran intratables, ha transformado el análisis de datos; la rigidez de los planteamientos tradicionales está dando paso a una *actitud exploratoria* en la que los datos son estudiados desde múltiples puntos de vista, a la busca de patrones, relaciones o interdependencias, que permitan intuir los posibles modelos subyacentes a las fluctuaciones aleatorias, los errores y la confusión general, que típicamente se observan en los datos reales. Por otra parte, la formalización axiomática, presente en las matemáticas desde Euclides, ha terminado por alcanzar a los métodos estadísticos; los distintos métodos tradicionales están siendo progresivamente sustituidos por una *metodología general unificada*, deducida a partir de los sólidos fundamentos axiomáticos de la teoría de la decisión y caracterizada por una descripción probabilística de la incertidumbre: la metodología Bayesiana.

En este trabajo, se describen con cierto detalle las implicaciones de estos dos factores, y la manera en que su conjunción ha dado lugar a una nueva forma de enfocar y resolver los problemas en los que los métodos estadísticos pueden resultar útiles. El énfasis se sitúa en el plano conceptual, remitiéndose al lector a la literatura especializada para la mayor parte de los detalles técnicos.

---

* Departamento de Estadística, Presidencia de la Generalidad Valenciana.

En la Sección 2 se proporciona una introducción elemental a la moderna teoría de la decisión, cuya estructura axiomática proporciona los fundamentos sobre los que descansan los métodos Bayesianos, y en ese contexto, se describen el análisis de datos y la inferencia estadística como partes del trabajo necesario para incorporar la información de que se dispone al proceso de decisión. En la Sección 3 se subraya el carácter gráfico e interactivo del análisis de datos contemporáneo, se describen algunas de sus técnicas más interesantes, y se discute la forma en que los resultados de tal análisis pueden ser formalizados mediante modelos matemáticos de tipo probabilístico. En la Sección 4 se detalla el proceso formal de inferencia estadística, desde una perspectiva Bayesiana, y se destacan las especiales características del tipo de resultados a que tal proceso da lugar. Finalmente, en la Sección 5 se discuten algunas de las consecuencias de las tesis mantenidas, y se comentan las referencias bibliográficas más relevantes.

2. EL PARADIGMA BAYESIANO

2.1. Teoría de la decisión

Desde sus orígenes, las matemáticas han mostrado una fuerte tendencia a la axiomatización. En efecto, la construcción de un sistema axiomático permite identificar con claridad los conceptos y relaciones que se consideran básicos, y garantiza que las conclusiones que pueden lógicamente derivarse de tal sistema forman una teoría coherente, exenta de contradicciones. Durante muchos años, los métodos estadísticos han permanecido fuera de esta corriente; como consecuencia, han coexistido numerosas "recetas", a veces incompatibles entre sí, para analizar distintos tipos de datos. De tales métodos se sabía que eran frecuentemente útiles, pero su correcto funcionamiento nunca podía ser garantizado a priori. En los años veinte, el genio de Ramsey (1926) advertía que toda la información disponible debe ser analizada en el contexto del problema real que motiva la investigación, argumentaba que los problemas reales pueden ser generalmente descritos como problemas de decisión, y establecía las bases para una teoría axiomática de la decisión, de la que puede deducirse la metodología necesaria para analizar la información relevante.

El trabajo de Ramsey, pionero, pero carente de una formalización matemática rigurosa, permaneció ignorado durante casi dos décadas, pero la publicación de la monografía de Savage (1954), *The Foundations of Statistics*, en la que se establecían las mismas tesis de forma rigurosa, ya no pasó inadvertida; la naciente *teoría de la decisión* iba a modificar definitivamente la forma de analizar los problemas estadísticos.

2.2. Conclusiones básicas

A pesar de su notable complejidad matemática, las conclusiones fundamentales que se derivan de adoptar un punto de vista axiomático ante los problemas de decisión son relativamente fáciles de describir.

En efecto, cualquier problema real de decisión puede ser representado como una sucesión de problemas de decisión elementales, en cuya estructura sólo intervienen tres elementos:

(i) el conjunto de decisiones posibles, o *espacio de alternativas*, A,

(ii) el conjunto de sucesos inciertos relevantes, o *espacio paramétrico*, Θ,

(iii) el conjunto de resultados que pueden derivarse de la alternativa que se elija y de los sucesos inciertos que tengan lugar, o *espacio de consecuencias*, C.

Sobre estos elementos básicos, se proponen un conjunto de *axiomas* cuyo objeto es caracterizar lo que se considera un comportamiento "racional", o *coherente*. Por ejemplo, parece razonable exigir de una persona sensata, que si considera que la alternativa a es preferible a la alternativa b, y que la alternativa b es preferible a la alternativa c, entonces *debe* considerar que la alternativa a es preferible a la alternativa c *(transitividad)*. Análogamente, si considera que la alternativa a es preferible a la alternativa b siempre que suceda θ, y también considera que la alternativa a es preferible a la alternativa b siempre que *no* suceda θ, entonces *debe* considerar que la alternativa a es *en todo caso* preferible a la alternativa b *(principio de substitución)*. Naturalmente, existe cierta libertad en la elección del sistema de axiomas; mientras unos sistemas axiomáticos intentan conseguir el mayor grado de generalidad posible, otros tienden a subrayar el contenido intuitivo o el carácter operativo de sus axiomas. Fishburn (1981) proporciona un resumen comparativo de los sistemas axiomáticos anteriores a 1980; Bernardo *et al.* (1985) ofrecen una alternativa más reciente. Sin embargo, las conclusiones básicas que se derivan de *todos* los sistemas axiomáticos propuestos son esencialmente las mismas:

(i) Debe construirse una *medida de probabilidad* que describa la *incertidumbre* del decisor sobre *todos* los aspectos desconocidos del problema *en el momento de tomar la decisión*. Formalmente, debe especificarse en una distribución de probabilidad sobre el espacio paramétrico, que representaremos por su densidad de probabilidad $p(\theta \mid I)$ respecto de la medida dominante apropiada, donde I representa la información disponible en el momento de decidir.

(ii) Debe construirse una *función de utilidad* que describa las *preferencias* del decisor entre los posibles resultados finales. Formalmente, debe especificarse en una función real u que asocia un *valor* $u(a,\theta)$ a cada una de las posibles combinaciones de alternativa a y suceso incierto θ.

(iii) Debe calcularse la *utilidad esperada* de cada decisión, definida por

$$Eu(a \mid I) = \int_{\Theta} u(a,\theta)\, p(\theta \mid I)\, d\theta,$$

lo que proporciona una medida directa de la *deseabilidad* de cada alternativa a, dada la

información I. Consecuentemente, debe tomarse aquella decisión a^* que maximiza la utilidad esperada $Eu\ (a \mid I)$.

Considérese, por ejemplo, la situación de un equipo de gobierno que debe realizar un plan de inversiones públicas a medio plazo. En primer lugar, debería especificarse el conjunto de alternativas posibles A, esto es, el conjunto de *todos* los planes de inversión compatibles con su presupuesto. Seguidamente, debería determinarse el conjunto de *todos* los sucesos inciertos relevantes θ^1, esto es el conjunto de elementos inciertos del problema que pueden afectar a las consecuencias que se deriven de la decisión que eventualmente se tome; así, la demanda mundial de barcos en los próximos años puede ser crucial, si se piensa en un astillero; el crecimiento probable del parque automovilístico, si se contemplan obras de infraestructura viaria; los resultados de las próximas elecciones, si la conclusión del plan requeriría disponer de mayoría absoluta en la próxima legislatura. Finalmente, deberían describirse con detalle las consecuencias que previsiblemente se derivarían de cada posible combinación de plan de inversiones a y suceso posible θ, esto es de cada par (a,θ), lo que completaría la *estructura* del problema de decisión planteado. Para *resolver* el problema, sería entonces necesario

(i) *valorar* cada uno de los resultados posibles, asignando a cada consecuencia c un valor numérico $u\ (c) = u\ (a,\theta\ )$ que debería describir el grado de satisfacción social que se derivaría de tal consecuencia;

(ii) determinar, en función de la información disponible I, una *distribución de probabilidad* $p\ (\theta \mid I\ )$ que describa la plausibilidad de cada uno de los sucesos considerados.

La deseabilidad de un plan de inversiones a vendría entonces medida por su utilidad esperada $Eu\ (a \mid I)$ y, consecuentemente, el plan óptimo a^* sería el que diese lugar a una utilidad esperada máxima. Ni la especificación de la valoración $u\ (a,\theta)$, ni la determinación de la distribución de probabilidad $p\ (\theta \mid I)$ son sencillas, pero ambos elementos son *imprescindibles* para una óptima toma de decisiones.

2.3. Información adicional

Frecuentemente, la información I de que se dispone sobre los elementos inciertos de un problema de decisión no resulta suficiente para decidir con garantías razonables de éxito, pero suele ser posible obtener información adicional que mejore esa situación; formalmente, la utilidad esperada de la decisión óptima, $Eu\ (a^* \mid I)$, puede resultar pequeña, pero generalmente resulta posible conseguir nuevos *datos* D que la mejoren.

En el ejemplo antes comentado, el gobierno puede decidir que necesita más información sobre la evolución del parque móvil, y encargar un estudio basado en los datos de Tráfico, o requerir mayor información sobre la valoración ciudadana de las distintas alternativas y realizar una encuesta entre la población.

El problema técnico de *incorporar* tal información adicional en el proceso de decisión no es trivial, y constituye, de hecho, el objetivo último de cualquier análisis estadístico. Desde la perspectiva axiomática que se ha descrito, se trata de obtener la distribución de probabilidad *final* $p\ (\theta \mid I, D)$ que describe la información de que se dispone sobre los aspectos inciertos del problema, dada tanto la información inicial I como la información adicional D, utilizando para ello la distribución *inicial* $p\ (\theta \mid I)$, y la presumible relación existente entre los nuevos datos D y los sucesos inciertos relevantes θ.

Típicamente, la relación entre D y θ puede expresarse mediante un *modelo probabilístico* $p(D \mid \theta)$ que describe la *verosimilitud* de los datos D como función de los posibles valores de θ. Por ejemplo, la relación entre los datos de una encuesta y el porcentaje de ciudadanos que favorece una determinada medida suele ser adecuadamente descrita mediante un modelo multinomial jerárquico. Una vez especificado el modelo $p\ (D \mid \theta)$, la distribución final deseada puede ser obtenida mediante el *Teorema de Bayes*, de forma que

$$p\ (\theta \mid I, D) \propto p\ (D \mid \theta)\, p\ (\theta \mid I),$$

o más explícitamente,

$$p\ (\theta \mid I, D) = k p\ (D \mid \theta)\, p\ (\theta \mid I),$$

donde la constante de proporcionalidad k puede obtenerse como

$$k = \left\{ \int_{\Theta} p\ (D \mid \theta)\, p\ (\theta \mid I)\, d\theta \right\}^{-1}$$

puesto que, por tratarse de una densidad de probabilidad, $\int_{\Theta} p\ (\theta \mid I, D)\, d\theta = 1$.

La especificación de un modelo probabilístico $p\ (D \mid \theta)$, que describa la relación existente entre los datos D y los elementos inciertos del problema θ, no es sencilla; generalmente, es necesario un exhaustivo *análisis de datos*, de tipo exploratorio, que proporcione los elementos de base para una modelización rigurosa. A partir de aquí, la construcción de la distribución final $p\ (\theta \mid I,D)$, basada en el modelo provisionalmente aceptado, constituye el núcleo del proceso de *inferencia* necesario para incorporar al análisis del problema de decisión la información proporcionada por los datos. El uso sistemático del Teorema de Bayes en tal proceso de inferencia es lo que justifica el adjetivo *Bayesiano*, con el que generalmente se describe el paradigma metodológico que nos ocupa.

En el resto de este trabajo, analizaremos con cierto detalle cada uno de estos elementos.

3. ANALISIS DE DATOS

En la descripción del análisis de datos moderno hay dos adjetivos fundamenta-

les: *interactivo* y *gráfico*. La importancia de tales características puede apreciarse mejor recordando, aunque sea brevemente, la historia de los sistemas de cálculo disponibles para el análisis estadístico.

Aunque la mayor parte de los métodos estadísticos clásicos ya se conocían a finales de los años treinta, sólo pudieron ser empleados de forma efectiva con la aparición de los sistemas electrónicos de cálculo. En sus comienzos, en los años 60, los programas de análisis estadístico residían en grandes ordenadores, y eran accedidos por los usuarios mediante tarjetas perforadas que había que rehacer si se cometía un error; el paquete de tarjetas, que contenía las instrucciones de control y los datos, era físicamente entregado en el centro de cálculo, y los resultados eran recogidos en el mismo lugar, impresos en papel contínuo, unas horas más tarde.

A principios de los 70 se trabajaba ya sobre terminales que evitaban los desplazamientos, se habían introducido lenguajes de control con comandos nemotécnicos, como "ajuste polinómico" o "regresión", y las variables eran ya identificadas por sus nombres, en lugar de por su código numérico. Sin embargo, se seguía trabajando en lotes (modo "batch"), de forma que el usuario debía ejecutar todo el programa de golpe y esperar, a veces varias horas, para obtener los resultados, ... o la indicación de que había algún error.

Los sistemas interactivos, en los que se teclea un comando en la terminal y el programa lo ejecuta de inmediato, aparecieron a finales de los años 70. Los programas interactivos permiten analizar los datos de forma secuencial, de forma que los resultados de una etapa pueden utilizarse para decidir lo que se hace en la etapa siguiente, sin que ello exija largas esperas intermedias. La mayor parte de los programas actuales pueden ser utilizados de forma interactiva.

A finales de los años 70, algunos estadísticos, estimulados por el trabajo pionero de Tukey (1977), *Exploratory Data Analysis*, defendieron el uso de programas interactivos para la exploración de datos sin modelos preconcebidos, adaptando para ello algunos de los métodos conocidos y desarrollando otros nuevos, en un intento de permitir que los datos "hablen por sí mismos" y sugieran relaciones, dependencias o interacciones que resulte interesante analizar con más detalle. En una etapa posterior, una vez intuídos los rasgos fundamentales de los datos, se podría proceder a formalizar modelos y obtener conclusiones cuantitativas precisas. Esta forma de trabajar, requería una versatilidad interactiva y una capacidad gráfica de la que los paquetes de programas tradicionales claramente carecían.

La década de los 80 es la del ordenador personal. Por un precio moderado, cualquier profesional puede tener ahora sobre su mesa mayor capacidad de cálculo de la que tenía toda la universidad en los años 60. Al principio, esto se tradujo simplemente en la adaptación de los paquetes de programas tradicionales para su uso en ordenadores personales. Sin embargo, la capacidad gráfica de las estaciones de trabajo personales pronto desbordó tales planteamientos, dando lugar a una nueva generación de programas para análisis de

datos, específicamente diseñados para trabajar de forma gráfica e interactiva.

En el resto de esta Sección, comentaremos brevemente algunas de las técnicas más importantes en el análisis exploratorio de datos.

3.1. Detección de observaciones atípicas

Uno de los elementos cruciales en cualquier análisis es la detección de posibles observaciones atípicas en el conjunto de los datos. Para empezar, los datos reales suelen tener errores; tales errores suelen ser debidos a fallos en el sistema de medida que los originó, a equivocaciones humanas cometidas en su transcripción o, en menor medida, a transmisiones electrónicas defectuosas entre los distintos soportes utilizados; naturalmente, los errores identificados son corregidos o, si ésto no resulta posible, los registros correspondientes son eliminados. Sin embargo, observaciones correctamente realizadas y transcritas pueden resultar tan atípicas, tan extraordinarias, que deban ser eliminadas del análisis general, y estudiadas después por separado, para poder identificar las relaciones subyacentes al conjunto de las observaciones que presentan un comportamiento estándar.

Por ejemplo, en el análisis de una encuesta orientada a determinar las preferencias mostradas por los ciudadanos entre distintos planes de inversión, hay que eliminar los errores físicos, posiblemente debidos a un grabado incorrecto de las respuestas obtenidas, pero también hay que identificar y tratar por separado, observaciones atípicas que oscurecerían la tendencia general; por ejemplo, cuando se analizan los datos de una encuesta sobre posibles inversiones en la red viaria, tal vez deban tratarse por separado los datos correspondientes a la zona que debería ser expropiada para construir una nueva autopista.

Los métodos gráficos interactivos son especialmente efectivos en la detección de observaciones atípicas. Por ejemplo, las gráficas móviles tridimensionales correspondientes a la representación simultánea de las tres primeras componentes principales de cada elemento del conjunto de datos, permiten una inmediata identificación visual de las observaciones atípicas. Menos poderosos, pero a veces muy efectivos, son los métodos gráficos en una o dos dimensiones; por ejemplo, los diagramas de caja ("boxplots") proporcionan un interesante resumen gráfico de la distribución de una magnitud univariante.

3.2. Transformaciones

La elección de una métrica adecuada para cada una de las magnitudes presentes en los datos originales es frecuentemente crucial para su análisis efectivo.

En primer lugar, es preferible trabajar siempre con magnitudes tipificadas, puesto que permiten comparaciones intuitivas de forma mucho más sencilla. Por otra parte, una proporción importante de los métodos estadísticos están diseñados para trabajar con datos *normales;* consecuentemente, suele resultar deseable transformar las magnitudes utilizadas de forma que sus distribuciones marginales se ajusten cuanto sea posible a una distribución

normal. En algunos casos, la experiencia acumulada sugiere transformaciones obvias, como la transformación logística cuando se trabaja con proporciones; en otros casos, sin embargo, es necesaria una búsqueda secuencial, enormemente facilitada por aquellos sistemas que permiten la observación simultánea de los histogramas y representaciones en escala probabilística normal de las distribuciones correspondientes a las distintas transformaciones consideradas.

Naturalmente, el análisis de transformaciones, como cualquier otra manipulación de los datos, debe realizarse una vez excluidas las observaciones atípicas, puesto que en otro caso tales observaciones podrían sesgar el análisis.

3.3. Detección de subpoblaciones homogéneas

Para establecer un modelo probabilístico que describa fielmente la relación entre los datos observados y los sucesos inciertos que se investigan, es frecuentemente necesario *separar* el conjunto de datos en subpoblaciones razonablemente homogéneas entre sí.

Las representaciones tridimensionales en componentes principales antes mencionadas pueden utilizarse con este propósito, lo que a veces permite la identificación visual de conglomerados. Sin embargo, es frecuentemente más efectivo procesar los datos con algunos de los numerosos algoritmos específicamente diseñados para la determinación de conglomerados.

La diferencia más importante entre los distintos métodos de construcción de conglomerados ("clusters") reside en el tipo de *distancia* utilizada para determinar el grado de semejanza entre cada par de registros; cada definición de distancia da lugar a un resultado diferente, gráficamente descrito por un *dendograma*, en el que resultan aparentes los subconjuntos de datos relativamente homogéneos con respecto a *esa* distancia. El uso de diferentes distancias (o alternativamente, el uso repetido de la misma definición de distancia con distintas transformaciones de las magnitudes incluidas en cada registro) proporciona distintos conglomerados que, en el contexto de los datos analizados, suelen tener una interpretación interesante.

Por ejemplo, en el análisis de datos electorales, la distancia logarítmica entre las distribuciones de voto en cada comarca, origina conglomerados de comportamiento político homogéneo, mientras que la distancia euclídea permite la identificación inmediata de comarcas que resultan atípicas como consecuencia del extraordinario nivel de implantación en ellas de alguno de los partidos.

3.4. Tratamiento de las observaciones incompletas

Uno de los aspectos más polémicos del análisis de datos es el tratamiento que se da a las observaciones incompletas presentes en los datos. Esencialmente, hay dos posibilidades: el tratamiento convencional consiste en ignorar en el análisis aquellos elementos

en los que falte una o más de las magnitudes requeridas; alternativamente, pueden *imputarse* los valores que faltan, utilizando la información proporcionada por los datos completos para *predecir* el valor más probable de las magnitudes que faltan en un registro, en función de las magnitudes presentes en ese mismo registro.

Si puede suponerse que la distribución de observaciones incompletas es aleatoria, la segunda opción es más arriesgada: si se hace correctamente, se dispone de mucha más información, obteniéndose por lo tanto conclusiones más precisas, pero un modelo de imputación incorrecto puede introducir importantes sesgos. Sin embargo, la distribución en la población de las observaciones incompletas es rara vez independiente del objeto de la investigación y, en este caso, ignorar las observaciones incompletas puede producir serios sesgos en el análisis.

Por ejemplo, en las elecciones generales de 1982, la imputación probabilística, a partir de un modelo logístico basado en su perfil sociológico, de las contestaciones que presumiblemente hubieran dado las personas encuestadas que no quisieron revelar su intención de voto, permitió unas predicciones extremadamente precisas (Bernardo, 1984); ignorar las contestaciones incompletas hubiera desviado las predicciones, debido a la correlación existente entre la intención de voto y la disposición a darlo a conocer.

El análisis de datos moderno permite estudiar interactivamente, por simulación, las consecuencias de distintas formas de imputación, reduciendo así notablemente el peligro de introducir inadvertidamente algún tipo de sesgo.

3.5. Identificación de modelos

Una vez eliminadas las observaciones atípicas, realizadas las transformaciones oportunas, separados los datos en subpoblaciones homogéneas, y decidido en cada una de ellas un tratamiento apropiado para las observaciones incompletas, pueden empezar a estudiarse, con garantías, los posibles modelos que servirían para explicar el comportamiento observado.

Histogramas de las distribuciones marginales de las variables, y de sus transformaciones, en las distintas subpoblaciones; representaciones gráficas *relacionales*, de forma que los elementos seleccionados en una de ellas son inmediatamente identificados en las demás; regresión de algunas variables sobre distintas transformaciones de otras, y análisis de los correspondientes residuos; análisis de las correlaciones existentes entre pares de variables; comparación múltiple de las características correspondientes a las distintas subpoblaciones, ... son algunos de los elementos que pueden contribuir a sugerir un modelo capaz de explicar el comportamiento de los datos.

Uno de los métodos más poderosos en esta fase final del análisis de datos es la búsqueda de elementos invariantes, de simetrías. Se puede comprobar, por ejemplo, que una reordenación aleatoria de los datos en una subpoblación no modifica las caracte-

rísticas observadas, lo que permitiría suponer que los elementos que la integran son *parcialmente intercambiables* esto es, en un lenguaje más convencional, que constituyen una muestra aleatoria de un cierto modelo; se puede verificar que los resíduos correspondientes a una regresión lineal tienen simetría esférica, lo que sugeriría un modelo lineal normal; puede comprobarse que una determinada función vectorial de los datos proporciona el mismo tipo de información que los datos completos, lo que sugeriría que se trata de un estadístico suficiente y, por tanto, que el modelo probabilístico adecuado pertenece a la correspondiente familia exponencial generalizada. Las posibilidades son esencialmente innumerables; sólo están limitadas por la competencia y la imaginación del estadístico.

En definitiva, el resultado del análisis exploratorio de los datos debe proporcionar información suficiente sobre su estructura como para permitir la especificación de un modelo matemático con el que puedan deducirse conclusiones cuantitativas precisas. La deducción de tales conclusiones constituye el proceso formal de inferencia estadística.

4, INFERENCIA ESTADISTICA

4.1. Modelización

De acuerdo con las conclusiones que se derivan de la teoría de la decisión, para determinar la alternativa óptima, dada la información inicial I y los datos adicionales D, correspondiente a un problema de decisión cuyo espacio de sucesos inciertos es θ, es necesario determinar la *distribución final* $p(\theta \mid I,D)$, que describe la información de que se dispone sobre θ en el momento de tomar la decisión.

A partir de un análisis de datos intensivo, en la línea que acaba de ser descrita, es a veces posible construir un modelo matemático, $p(D \mid \theta)$ que describa *directamente* la relación probabilística entre los datos D y el parámetro de interés θ; más frecuentemente, sin embargo, el modelo encontrado es de la forma $p(D \mid \theta, \omega)$, donde aparece un nuevo parámetro desconocido ω, al que llamaremos *parámetro marginal* que, como θ, es frecuentemente multidimensional.

Por ejemplo, el análisis de datos puede sugerir, para cada una de las k subpoblaciones descubiertas, una distribución normal multivariante alrededor de un vector media, en el que estamos interesados, pero su especificación también depende de una matriz de covarianzas desconocida. En este caso, el modelo completo sería de la forma

$$p(D \mid \theta, \omega) = \sum_{i=1}^{k} \alpha_i N(D_i \mid \mu_i V_i)$$

donde el parámetro de interés $\theta = \{\mu_1, \ldots, \mu_k\}$ contiene a las medias, y el parámetro marginal $\omega = \{V_1, \ldots, V_k, \alpha_1, \ldots, \alpha_k\}$ está constituido por las varianzas y por los pesos relativos de cada una de las subpoblaciones.

4.2. Distribuciones finales

De acuerdo con la axiomática de comportamiento racional, es necesario especificar mediante probabilidades la información de que se dispone sobre cualquier elemento incierto del problema. Consecuentemente, será necesario especificar una distribución inicial conjunta $p(\theta,\omega \mid I)$ que describa la información de que inicialmente se dispone tanto sobre θ como sobre ω.

Una vez determinada la distribución inicial, es fácil obtener, mediante el Teorema de Bayes, la distribución final conjunta

$$p(\theta, \omega \mid I,D) \propto p(D \mid \theta, \omega)\, p(\theta, \omega \mid I)$$

que describe la información de que se dispone sobre θ y ω, una vez incorporada la información proporcionada por los datos D. A partir de aquí, puede marginalizarse para obtener la distribución final buscada

$$p(\theta \mid I,D) = \int_{\Omega} p(\theta, \omega \mid I,D)\, d\omega,$$

que es el elemento necesario para determinar la utilidad esperada de cada alternativa

$$Eu(a \mid I,D) = \int_{\Theta} u(a,\theta)\, p(\theta \mid I,D)\, d\theta.$$

Esto proporciona una medida de la deseabilidad de cada una de las alternativas consideradas en el momento de decidir y constituye, por lo tanto, la *solución* al problema de decisión planteado.

En algunas ocasiones, el modelo probabilístico es de la forma $p(D \mid \omega)$, de manera que el parámetro de interés θ no aparece explícitamente en el modelo, pero es en cambio una *función* matemática conocida $\theta = \theta(\omega)$ del parámetro marginal utilizado. Por ejemplo, en el análisis de datos electorales, el modelo probabilístico multinomial es expresado en función de las proporciones $\omega = \{\omega_1, \ldots, \omega_k\}$ de votos conseguidos por cada partido, pero el parámetro de interés θ, la distribución de escaños en el Parlamento, es una función suya, $\theta = \theta(\omega)$, definida por la Ley d'Hont.

En estos casos, el Teorema de Bayes permite obtener directamente la distribución final de ω,

$$p(\omega \mid I,D) \propto p(D \mid \omega)\, p(\omega \mid I),$$

de donde la distribución final buscada,

$$p(\theta \mid I,D) = T\left\{p(\omega \mid I,D)\right\},$$

puede obtenerse, analítica o numéricamente, mediante la transformación probabilística apropiada, T. Obsérvese que este tipo de transformación *no* es posible en la metodología clásica.

El proceso de cálculo de las distribuciones finales puede resultar técnicamente difícil; por ejemplo, pueden aparecer serios problemas de integración multidimensional, o puede resultar que no existan soluciones analíticas y deba recurrirse al cálculo numérico. Sin embargo, debe subrayarse que se trata de dificultades *técnicas*, no conceptuales. El proceso a seguir es único y bien definido; se dispone de una teoría *general* de inferencia.

4.3. Predicción

Frecuentemente, puede suponerse que los datos constituyen una muestra aleatoria, constituida por un determinado número n, de observaciones independientes de un determinado modelo; por ejemplo, en control de calidad, los datos relevantes consisten frecuentemente en los resultados de las medidas x realizadas a n productos elegidos aleatoriamente dentro de una partida. En este caso, $D = \{x_1, ..., x_n\}$, con

$$p(D \mid \theta, \omega) = \prod_{i=1}^{n} p(x_i \mid \theta, \omega).$$

En este tipo de situaciones es frecuente que la función de utilidad venga expresada en términos de los valores de un nuevo registro x, y sea por tanto de la forma $u\,(a,x)$, en lugar de ser de la forma $u\,(a,\theta)$. Por ejemplo, en control de calidad, las consecuencias de la decisión de aceptar o rechazar una partida generalmente dependen de que existan o no productos cuyas magnitudes x excedan las especificaciones permitidas, no del valor medio de tales magnitudes en la partida analizada.

Puesto que siempre es necesario obtener la distribución de probabilidad que describe la información disponible sobre el suceso incierto *relevante* y en este caso tal suceso es el valor x de un nuevo registro, será necesario determinar la distribución final de x, que denominaremos distribución *predictiva* de x, y que puede obtenerse, utilizando la distribución final de los parámetros en el Teorema de la probabilidad total, como

$$p(x \mid I,D) = \int_{\Theta}\int_{\Omega} p(x \mid \theta, \omega)\, p\,(\theta, \omega \mid I,D)\, d\theta d\omega.$$

Consecuentemente, la utilidad esperada de una alternativa a será ahora de la forma

$$Eu\,(a \mid I,D) = \int_{X} u(a,x)\, p\,(x \mid I,D)\, dx,$$

y la mejor alternativa a^*, la que maximize esa expresión.

Obsérvese la notable diferencia entre la solución Bayesiana al *problema de predicción*, definida por la distribución predictiva $p(x \mid I,D)$, y la solución tradicional $p(x \mid \hat{\theta}, \hat{\omega})$

que se limita a substituir los valores desconocidos de θ y ω por los de sus estimadores $\hat{\theta}$, $\hat{\omega}$, lo que ni siquiera tiene en cuenta el tamaño n de la muestra en que se basa la predicción. Aichison y Dunsmore (1975) proporcionan un detallado análisis de las peligrosas consecuencias de este tipo de error.

4.3. Distribuciones de referencia

La especificación de la distribución $p(\theta, \omega \mid I)$ que describe la información inicial es frecuentemente muy difícil, especialmente en las frecuentes situaciones en las que los parámetros θ y ω son multidimensionales. Por otra parte, la información inicial es frecuentemente pequeña comparada con la información proporcionada por los datos, de manera que puede esperarse que $p(\theta \mid I,D)$ y $p(\theta \mid D)$ sean básicamente idénticas. Finalmente, existen situaciones, por ejemplo en la redacción del informe relativo a los resultados de un experimento científico, en las que no quiere incorporarse más información que la directamente proporcionada por los resultados del experimento realizado. Para todos estos casos, es necesario disponer de una distribución inicial de *referencia* que matemáticamente describa la propiedad deseada, esto es que la información inicial sea despreciable frente a la información proporcionada por los datos.

La teoría de la información proporciona una forma de resolver el problema; en efecto, existen argumentos de tipo axiomático que permiten deducir la expresión adecuada para medir la *cantidad de información desconocida*, $H^{\theta}\{p(\theta)\}$, sobre el parámetro de interés θ, que naturalmente depende de la información $p(\theta)$ de que se dispone: cuanto *mayor* sea la información contenida en la distribución inicial $p(\theta)$, *menor* será la cantidad de información desconocida sobre θ. Recíprocamente, la distribución inicial "no-informativa" o de referencia, $\pi\,(\theta)$, será aquella que maximice la información desconocida $H^{\theta}\{p(\theta)\}$ (Bernardo, 1979). En el caso de que existan parámetros marginales, se puede utilizar el mismo argumento de forma secuencial, para obtener una distribución inicial de referencia de la forma $\pi\,(\theta)\,\pi\,(\omega \mid \theta)$ (Berger y Bernardo, 1989).

Las correspondientes distribuciones finales, $\pi\,(\theta \mid D)$ y $\pi\,(x \mid D)$ son las *distribuciones finales de referencia* y describen la información que el conjunto de los datos D proporcionan sobre el parámetro de interés θ y sobre futuras observaciones x, dado el modelo supuesto. Puesto que la metodología tradicional siempre ignora la información inicial, las distribuciones finales de referencia proporcionan la única forma de comparar los resultados clásicos con los resultados Bayesianos: ha podido observarse que, generalmente, los resultados clásicos pueden ser reproducidos a partir de las distribuciones finales de referencia, por lo que, pragmática, *aunque no conceptualmente*, el paradigma Bayesiano constituye una ***extensión*** de la metodología clásica.

4.4. Descripción de resultados

El resultado final de un proceso de inferencia Bayesiano es siempre una distribución de probabilidad, la distribución final del parámetro de interés, $p(\theta \mid I,D)$. Sin em-

bargo, con objeto de comunicar con facilidad los rasgos más característicos del resultado obtenido, es conveniente presentar, junto con la ecuación formal que define $p(\theta \mid I,D)$, algunas medidas de localización, algunas medidas de dispersión, y una representación gráfica adecuada.

La elección de una medida de localización puede ser descrita en sí misma como un problema de decisión, en el que el espacio de alternativas coincide con el espacio paramétrico, esto es, $A = \theta$. Es la versión Bayesiana de la *estimación puntual*. Para distintas funciones de pérdida $l(\theta,\hat{\theta})$, se obtienen distintas soluciones, esto es distintos *estimadores* θ o medidas de localización de la distribución final; la media, la mediana y la moda son casos particulares (ver, por ejemplo, Berger, 1985, p. 161). Generalmente, sin embargo, no hay razones para limitarse a calcular un sólo estimador, y es frecuente determinar al menos los tres anteriores; en efecto, la media de la distribución final suele tener interesantes propiedades matemáticas y el intervalo determinado por la moda y la mediana de la distribución final contiene a la mayor parte de los estimadores que resultan de funciones de pérdida razonables, por lo que constituye una interesante "medida de localización", en forma de intervalo.

La familia más interesante de medidas de dispersión, desde un punto de vista Bayesiano está constituída por las llamadas *regiones de máxima densidad*. La región de máxima densidad y nivel de probabilidad p, o *región p-creíble*, R_p, correspondiente a la distribución $p(\theta \mid I,D)$, es aquella de menor volumen entre las que contienen θ con probabilidad p, esto es, tales que minimizan $\int_{R_p} d\theta$ bajo la condición

$$\int_{R_p} p(\theta \mid I,D)\, d\theta = p.$$

Es fácil demostrar que toda región creíble tiene la propiedad de que cualquier punto de ella tiene mayor densidad de probabilidad que todos los que no pertenecen a la región; de ahí la expresión "máxima densidad". Las regiones creíbles, que tienen forma de intervalos cuando el parámetro de interés es unidimensional y su distribución final es unimodal, constituyen la versión Bayesiana de los *intervalos de confianza* clásicos, y frecuentemente coinciden numéricamente con ellos, cuando no se utiliza información inicial; obsérvese, sin embargo, que la interpretación *correcta* de una región creíble R_p es que "de acuerdo con la información disponible, el valor del parámetro pertenece a R_p con probabilidad p", mucho más intuitiva que la forzada interpretación frecuencial de los intervalos de confianza. Debe asímismo subrayarse que las regiones creíbles *no* juegan en los métodos Bayesianos el papel privilegiado de los intervalos de confianza en la metodología clásica; constituyen simplemente una descripción parcial de la distribución final.

Las regiones creíbles pueden ser utilizadas para proporcionar un análogo Bayesiano a los contrastes de hipótesis clásicos, "rechazando", con nivel $1-p$, la hipótesis $\theta = \theta_0$ si, y solamente si, θ_0 no pertenece a la región creíble de nivel p. Sin embargo, si realmente se quiere plantear un problema de contraste desde una perspectiva Bayesiana,

debe formularse el problema de decisión correspondiente, definido por un espacio de decisiones con sólo dos alternativas, aceptar o rechazar la hipótesis nula, y por una función de utilidad que describa el valor de aceptar o rechazar tal hipótesis en función del verdadero valor de θ (Bernardo, 1985).

Si, por ejemplo, el resultado de una pequeña encuesta de intención de voto sobre 100 electores aleatoriamente elegidos da lugar, una vez adecuadamente procesadas las no-respuestas, a 10 votantes que piensan abstenerse, la distribución final de referencia correspondiente a la proporción θ de abstención sería una distribución Beta, $Be(\theta \mid 10.5,90.5)$. Puede comprobarse que la media, moda y mediana correspondientes son, respectivamente, *0.104, 0.096*, y *0.101*, mientras las regiones creíbles de niveles *0.90, 0.95* y *0.99* resultan ser, respectivamente, los intervalos *(0.055, 0.152), (0.048, 0.164)*, y *(0.037, 0.189)*. Consecuentemente, en términos más coloquiales, podría entonces afirmarse que la abstención oscilaría en torno al *10%*, situándose probablemente entre el *6%* y el *15%* y casi con seguridad entre el *4%* y el *19%*; en particular, con estos datos, no sería posible garantizar, con probabilidad de error menor del *1%*, que la abstención se situase por debajo del *18%*.

Cuando el parámetro de interés es unidimensional, es natural representar gráficamente su densidad de probabilidad, señalando además algunas de las medidas de localización y dispersión mencionadas. Cuando el parámetro de interés es multivariante, es interesante acompañar las gráficas de las distribuciones marginales de sus componentes, con representaciones de las "curvas de nivel" correspondientes a regiones creíbles bidimensionales de los pares de componentes que se estime oportunos, así como de las gráficas de las densidades de probabilidad correspondientes a funciones del parámetro de interés especialmente significativas.

4.5. Remodelización

Los resultados del proceso de inferencia correspondientes al modelo probabilístico que se juzgó más adecuado, pueden y deben ser utilizados para proceder a un análisis crítico del modelo y estudiar sus posibles mejoras. Por ejemplo, puede determinarse la distribución predictiva de distintas funciones de los datos *(funciones diagnósticas)* y verificar si tales *predicciones* son o no compatibles con los valores que toman tales funciones al actuar sobre los datos de que se dispone. La forma concreta de decidir sobre tal "compatibilidad" es un tema polémico; véase, por ejemplo, Box (1980).

Alternativamente, a partir de la distribución predictiva, pueden generarse observaciones artificiales del modelo por el método de Monte Carlo y analizar si los datos así obtenidos presentan o no elementos que los distingan de los datos originales; las posibles discrepancias proporcionan, cuando existen, importante información sobre los elementos del modelo que deben ser mejorados.

Naturalmente, esas actividades se sitúan de nuevo en el contexto del análisis

exploratorio de datos, iniciándose así un ciclo que sólo debe darse por terminado cuando el modelo final sea capaz de predecir *todas* las características descubiertas en el conjunto de las observaciones que no han sido consideradas atípicas.

5. DISCUSION

Tanto el análisis gráfico interactivo de datos como la metodología Bayesiana han conocido un crecimiento espectacular en los últimos años, pero el esfuerzo se ha centrado mayoritariamente en unos cuantos centros de investigación de primera fila; lamentablemente, su implementación dista mucho de ser mayoritaria en las universidades, y es prácticamente inexistente en la industria o la administración. La distancia temporal que tradicionalmente separa la investigación de punta de su difusión generalizada, no debe servir de consuelo.

La difusión de los métodos modernos de análisis de datos, requiere la disponibilidad por parte de los analistas de microordenadores y de paquetes de programas adecuados. En los centros de investigación de todo el mundo, la gama de SUN, sobre sistema operativo UNIX, se está convirtiendo rápidamente en un estándard; su escaso número y elevado precio reduce sin embargo la oferta de programas comerciales disponibles, de forma que la mayor parte de los centros programan sus propias aplicaciones. Una alternativa más modesta la constituye la gama de los Apple Macintosh, para la que están apareciendo una serie de programas comerciales específicamente diseñados para trabajar de forma gráfica e interactiva; ejemplos impresionantes de esta nueva tendencia son *Data Desk*, *MacSpin* y *Mathematica.*

La difusión de los métodos estadísticos Bayesianos requiere su incorporación sistemática a los programas de las asignaturas de introducción a la estadística que, con distintos nombres, se imparten en las universidades. Lamentablemente la mayor parte de los libros de texto elementales que se utilizan, ignoran completamente la metodología Bayesiana y, lo que es más grave, presentan los métodos tradicionales sin ningún tipo de análisis crítico (una excepción notable es DeGroot, 1988); consecuentemente, sólo algunos alumnos postgraduados, que frecuentemente se orientan a una actividad académica, llegan a poder apreciar las limitaciones de la estadística tradicional y a conocer la alternativa Bayesiana. En esta situación no puede sorprender que la *práctica* de la estadística sea todavía abrumadoramente tradicional.

Desde hace algunos años, los métodos Bayesianos vienen ocupando un espacio notable en las revistas profesionales más importantes, pero los libros de texto Bayesianos son relativamente escasos. Concluiremos este trabajo con una sucinta referencia cronológica a algunos de ellos.

La obra de Jeffreys (1961) es un clásico que contiene el germen de muchos de los desarrollos posteriores de los métodos Bayesianos. Lindley (1965) presenta una introducción a la probabilidad y a la estadística Bayesiana que subraya la equivalencia numérica de los resultados tradicionales con los que se derivan de las distribuciones finales de referencia. El libro de DeGroot (1970) es un curso para postgraduados en teoría Bayesiana de la decisión. Zellner (1971) introduce la econometría Bayesiana. Winkler (1972) proporciona una introducción elemental a la inferencia Bayesiana. Box y Tiao (1973) exploran las consecuencias de adoptar una perspectiva Bayesiana ante los problemas estadísticos que aparecen en la investigación científica. De Finetti (1975) proporciona una reconstrucción de la teoría de la probabilidad desde un punto de vista subjetivista. La monografía de Aichison y Dunsmore (1975) está centrada en los problemas de predicción. Hartigan (1983) presenta un análisis riguroso de los modelos matemáticos presentes en la metodología Bayesiana. El libro de Berger (1985) es un curso postgraduado en teoría de la decisión, que incluye tanto los aspectos clásicos como los Bayesianos.

Mención aparte merecen las colecciones de artículos publicadas en honor de estadísticos Bayesianos famosos: Jimmy Savage (Fienberg y Zellner, 1974), Harold Jeffreys (Zellner, 1980) y Bruno de Finetti (Goel y Zellner, 1984), o con ocasión de congresos monográficos Bayesianos: Fontainebleau (Aykac y Brumat, 1977), Innsbruck (Viertl, 1987), Purdue (Gupta y Berger, 1988) y Valencia (Bernardo *et al.*, 1980, 1985, 1988); los congresos de Valencia constituyen un reconocido foro permanente, de carácter mundial, donde cada cuatro años se discuten los nuevos avances logrados en el desarrollo de la metodología Bayesiana.

BIBLIOGRAFIA

AICHISON, J. and DUNSMORE, I.R. (1975). *Statistical Prediction Analysis.* Cambridge: University Press.

AYKAC, A. and BRUMAT, C., eds. (1977). *New Developments in the Applications of Bayesian Methods.* Amsterdam: North-Holland.

BERGER, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.* New York: Springer.

BERGER, J.O. and BERNARDO, J.M. (1989). *Estimating a product of means: Bayesian analysis with reference priors.* J. Amer. Statist. Assoc. **84** (in press).

BERNARDO, J.M. (1979). *Reference posterior distributions for Bayesian inference.* J. Roy. Statist. Soc. B. **41**, 113-147 (with discussion).

BERNARDO, J.M. (1984). *Monitoring the 1982 Spanish Socialist victory: a Bayesian analysis.* J. Amer. Statist. Assoc. **79**, 510-515.

BERNARDO, J.M. (1985). *Análisis Bayesiano de los contrastes de hipótesis paramétricos.* Trab. Estadist. **36,** 45-54.

BERNARDO, J.M., DEGROOT, M.H., LINDLEY, D.V., and SMITH A.F.M., eds. (1980). *Bayesian Statistics* . Valencia: University Press.

BERNARDO, J.M., DEGROOT, M.H., LINDLEY, D.V., and SMITH A.F.M., eds. (1985). *Bayesian Statistics 2.* Amsterdam: North-Holland.

BERNARDO, J.M., DEGROOT, M.H., LINDLEY D.V., and SMITH A.F.M., eds. (1988). *Bayesian Statistics 3.* Oxford: University Press.

BERNARDO, J.M., FERRANDIZ, J.R. and SMITH, A.F.M. (1985). *The foundations of decision theory: an intuitive, operational approach with mathematical extensions.* Theory and Decision **18,** 127-150.

BOX, G.E.P. and TIAO, G.C. (1973). *Bayesian Inference in Statistical Analysis.* Reading, Mass.: Addison-Wesley.

BOX, G.E.P. (1980). *Sampling and Bayes inference in scientific modelling and robustness.* J. Roy. Statist. Soc. A **143,** 383-430 (with discussion).

DE FINETTI, B. (1975). *Theory of Probability.* New York: Wiley.

DE GROOT, M.H. (1970). *Optimal Statistical Decisions.* New York: McGraw-Hill.

DE GROOT, M.H. (1988). *Probabilidad y Estadística.* Madrid: Addison-Wesley Iberoamericana.

FIENBERG, S.E. and ZELLNER, A., eds. (1974). *Studies in Bayesian Econometrics and Statistics.* Amsterdam: North-Holland.

FISHBURN, P.C. (1981). *Subjective expected utility: a review of normative theories.* Theory and Decision **13,** 139-199.

GOEL, P.K. and ZELLNER, A., eds. (1984). *Bayesian Inference and Decision Techniques with Applications.* Amsterdam: North-Holland.

GUPTA, S.S. and BERGER, J.O., eds. (1988). *Statistical Decision Theory and Related Topics IV,* 1. New York: Springer.

HARTIGAN, J.A. (1983). *Bayes Theory.* New York: Springer.

JEFFREYS, H. (1961). *Theory of Probability.* Oxford: University Press.

LINDLEY, D.V. (1965). *An Introduction to Probability and Statistics from a Bayesian Viewpoint.* (2 vol.) Cambridge: University Press.

RAMSEY, F.P. (1926). *Truth and probability.* Reprinted in Studies in Subjetive Probability (H. Kyburg and H. Smoker, eds.), New York: Wiley, 1964, 61-92.

SAVAGE, L.J. (1954). *The Foundations of Statistics.* New York: Wiley.

TUKEY, J.W. (1977). *Exploratory Data Analysis.* Reading, Mass.: Addison-Wesley.

VIERTL, R., ed. (1987). *Probability and Bayesian Statistics.* London: Plenum Press.

WINKLER, R.L. (1972). *Introduction to Bayesian Inference and Decision.* New York: Holt, Rinehart and Wiston.

ZELLNER, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley.

ZELLNER, A., ed. (1980). *Bayesian Analysis in Econometrics and Statistics.* Amsterdam: North-Holland.

SOFTWARE

Data Desk. Odesta Corporation. 4084 Commercial Avenue. Northbrook, Illinois 60062, U.S.A. Requiere Apple Macintosh. (U.S. $ 495).

MacSpin. D^2 Sofware. 3001 North Lamar Blvd., # 110 Austin, Texas 78705, U.S.A. Requiere Apple Macintosh. (U.S.$ 200).

Mathematica. Wolfram Research, Inc. P.O. Box 6059. Champaign, Illinois 61821, U.S.A. Requiere Apple Macintosh. (U.S.$ 795).