UNIVERSIDAD DE EXTREMADURA

ESTADÍSTICA, CERTEZA E INCERTIDUMBRE

LECCIÓN INAUGURAL
CURSO ACADÉMICO
2001 - 2002

por

Mª ASUNCIÓN RUBIO DE JUAN

Catedrática de Estadística e Investigación Operativa
Escuela Politécnica de la Universidad de Extremadura

BADAJOZ, 2001

UNIVERSIDAD DE EXTREMADURA

ESTADÍSTICA, CERTEZA E INCERTIDUMBRE

LECCIÓN INAUGURAL
CURSO ACADÉMICO
2001 - 2002

por

Mª ASUNCIÓN RUBIO DE JUAN

Catedrática de Estadística e Investigación Operativa
Escuela Politécnica de la Universidad de Extremadura

BADAJOZ, 2001

Excmas. e Ilmas. Autoridades,

Señoras y Señores:

En este solemne acto de Apertura del Nuevo Curso Académico 2000-2001, me presento ante Vds. con hondo agradecimiento que lleva unido un sincero temor y una esperanza. Agradecimiento, porque se hayan considerado mis limitados méritos para ser la primera mujer que imparta la Lección Inaugural de un Curso Académico en la Universidad de Extremadura. Temor, por la responsabilidad que entraña abordar un tema de carácter tecnológico ante una audiencia pluridisciplinar habituada, por tradición, a atender materias de humanidades. Esperanza, en que mi intervención me posibilite aportar, si no rayos de sol, al menos ráfagas de luz que, unidas, contribuyan a vislumbrar los cambios que el conocimiento de la Estadística conlleva en la sociedad.

Al dar la bienvenida a un nuevo curso, expresemos nuestro deseo de lograr que la Estadística sea fuente de progreso y perfeccionamiento en la evolución de la humanidad. Estas convicciones han configurado el marco de esta Lección Inaugural: Estadística, Certeza e Incertidumbre.

Manifestado este deseo, quiero mostrar mi reconocimiento a dos insignes mujeres y loables estadísticas: Florence Nightingale y Denise Lievesley.

Breve reseña

Florence Nightingale nació en Italia en 1820, enfermera de profesión y estudiosa de métodos y técnicas estadísticas. Estudió simultáneamente Aritmética, Geometría, Álgebra, etc. Además de sus implicaciones como enfermera, dedicó gran parte de su tiempo a enseñar a niños estas disciplinas. Resaltaremos su participación en la guerra de Crimea, donde muestra su honda preocupación por los enfermos, debido al mal estado de los hospitales de campaña. El impacto de los heridos de guerra, la mala situación de los hospitales y sus sentimientos por los enfermos van conformando sus experiencias, motivaciones y conducta.

Sobresale por su dedicación y estímulo para mejorar, mediante la aplicación de sencillas técnicas estadísticas, el sistema hospitalario. Logró el reconocimiento del gobierno inglés por sanear tanto el sistema hospitalario militar como el civil, las maternidades y las escuelas coloniales.

Se la considera una pionera en la aplicación de los métodos epidemiológicos y en la utilización de modelos estadísticos en salud pública. Cabe destacar su visión para saber cómo han de ser manipulados los datos, tratando de obtener información fidedigna, y cómo presentar estos datos mediante simples gráficos. Desarrolló el diagrama conocido como "coxcomb", diagrama de las causas de mortalidad en el ejército durante la guerra de Crimea.

Ha sido la primera mujer elegida, en 1858, miembro de "The Statistical Society". Entre los muchos reconocimientos y premios que alcanzó, mencionaremos el de ser elegida miembro de honor de la Asociación de Estadística Americana. La Reina Victoria le otorgó la Orden de Mérito del gobierno inglés. Murió en su casa, en Inglaterra, en 1910.

Pese a su profesión de enfermera, puede considerársela una verdadera estadística por su amor a esta ciencia, cuestionándose siempre las asunciones y razonando con extremo cuidado en el proceso de toma de decisiones.

Denise Lievesley, centésima Presidenta de "The Royal Society", y segunda mujer en ocupar este cargo, puesto que desempeñará hasta finales del año 2001. Denise pertenece y ha pertenecido a numerosos comités estadísticos en ésta y otras sociedades. En 1986 fue elegida miembro de honor del Instituto Internacional de Estadística y en 1998 fue premiada con la medalla de Henn Willem Methorst por su contribución a la estadística internacional. En 1999 fue nombrada directora del Instituto para Estadística de la UNESCO, con base en París. Es profesora en el departamento de Matemáticas de la Universidad de Essex, directora del ISI en Holanda y desempeña otros altos cargos. Es una entusiasta defensora de las estadísticas oficiales, tratando de promover y establecer nexos entre los usuarios y los realizadores de las estadísticas oficiales. Una de sus metas es la concienciación de la sociedad a favor de los estadísticos, promoviendo a ultranza el uso de la metodología estadística entre los gobiernos tanto centrales como locales.

En su vida privada ha logrado superar un cáncer que le fue diagnosticado con anterioridad a su presidencia y, según manifiesta, se siente feliz de poder probar que los diagnósticos estadísticos pueden equivocarse.

Introducción

¿Quién no ha reflexionado alguna vez sobre las siguientes cuestiones: qué es el conocimiento y cómo adquirirlo, cómo se pueden caracterizar los procesos implícitos?

Éstas y otras preguntas han desconcertado al intelecto humano y han sido durante mucho tiempo objeto de discursos filosóficos. En el desarrollo del conocimiento científico y epistemológico hay épocas épicas, en las que predomina el avance para exploración y conquista, y épocas en las que prevalece la consolidación y profundización.

Tanto en el avance como en la profundización se alcanzan contornos o fronteras. Estas fronteras de lo cierto, lo seguro y lo exacto no siempre son nítidas, sino constituidas por zonas o bandas. La Incertidumbre, la Estadística y la Probabilidad, o quizás la ambigüedad, el

desorden y la estocástica acapararán nuestra atención en estas zonas fronterizas.

Cierto es que nada se sabe de cierto, pero no es menos cierto que hay ansia de saber. De la aspiración por saber nacen esfuerzos para la cuantificación de los fenómenos, o al menos para la clasificación y sistematización, para describir o para inventar clases y órdenes. Corresponden a este campo las mediciones y los comienzos de la tecnología científica, fuentes del reconocimiento de la inevitabilidad de los errores y del nacimiento de la Estadística.

Otra fuente clásica de la Estadística que constituye una zona clara de incertidumbre, entre lo seguro y lo imposible, son los juegos de azar. Una o varias disposiciones aleatorias pueden repetirse convirtiéndose en determinísticas. Por otra parte, varias disposiciones determinísticas pueden a su vez disponerse al azar, como sucede en el caleidoscopio.

En lo que se refiere a un atributo, la banda fronteriza puede ser difusa por el modo de ser; son ejemplos de ello el color y la forma. Por el modo de estar, serían: los cambios de color, por mimetismo; la metamorfosis, el paso de crisálida a mariposa; los cambios en minerales y rocas, etc. También, por el modo de ser percibido, hay una zona de vaguedad relativa al observador e instrumento, natural o artificial, que percibe; por ejemplo, según la luz utilizada (infrarroja, etc.) cambiará la modalidad del color percibido. En un campo muy diferente, el de la realización de encuestas de opinión por muestreo, la interferencia entre el encuestador y el encuestado. Algunos adjetivos, como ligero o pesado, sustantivos, como montón o puñado, adverbios como bastante o insuficiente, dualidades como signo u objeto, serían ejemplos de una zona ambigua por el modo de decir o expresarse.

Por último, nos referiremos a otra zona no nítida, dada su gran actualidad e interés práctico, a la necesidad de sustituir o inventar datos faltantes por no respuesta, por pérdida, por tergiversación o deterioro.

> "Los lejos del paisaje se esfuman borrosos en la niebla...,
> todo se ve confuso y borroso."
>
> ("Azorín", *El paisaje de España*, cap. VI).

Éstas y otras fuentes de ambigüedad, de borrosidad y de incertidumbre, consideradas en el estudio estadístico, constituyen algunas de las fronteras imprecisas en la evolución del pensamiento humano.

Necesidad e importancia de la Estadística

La ciencia trata de conocer los fenómenos naturales y mejorarlos. Tal entendimiento puede explicarse mediante un proceso de abstracción y con frecuencia se expresa en términos de leyes, axiomas y teorías que permiten predecir eventos futuros sin precisar ciertos límites de exactitud.

Las leyes de movimiento de Newton, la teoría de la relatividad de Einstein, el modelo atómico de Bohr, el efecto Roman, las leyes de herencia de Mendel, la teoría de la evolución debida a Darwin... han sido pilares básicos en la evolución de la tecnología moderna. Sin embargo, nunca sabremos en qué medida estas leyes son ciertas. En el desarrollo del conocimiento científico actual y en la evolución de nuestra tecnología, un hecho irrefutable ha sido la búsqueda de hipótesis que apoyen hechos observables y que, a lo largo del tiempo, puedan reemplazarse por hipótesis que evidencien con mayor nitidez los conjuntos de datos.

El método científico de investigación, debido a Box y Hunter (1988), como muestra la siguiente figura 1, implica un proceso lógico de razonamiento deductivo-inductivo.

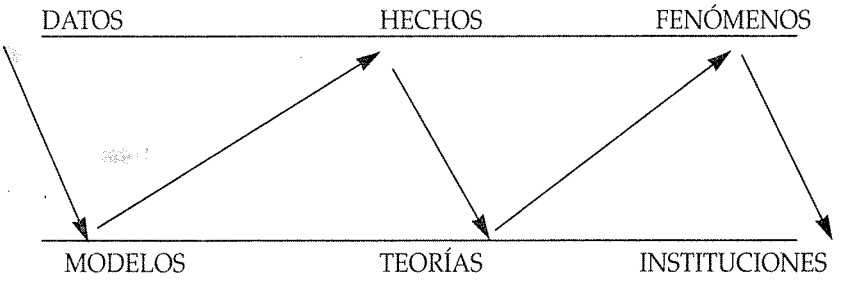

Figura 1. Método científico

El aprendizaje avanza con la iteración representada en la figura anterior y puede también describirse como un bucle de retroalimentación en el que la discrepancia entre los datos y las consecuencias de las hipótesis de partida conduce a las hipótesis modificadas. En este ciclo se deducen las consecuencias de las hipótesis modificadas y se comparan éstas de nuevo con los datos, lo que a su vez puede llevar a nuevas modificaciones y ganancias de conocimiento.

Este modelo de aprendizaje iterativo es la base de toda investigación científica. El objeto de los métodos estadísticos es hacer que ese proceso sea lo más eficiente posible.

Ej.: En una fiesta infantil, una madre servía chocolate en las tazas de los niños, que se encontraban sentados alrededor de la mesa de cumpleaños. La madre preguntó qué sucedería si, en vez de parar, siguiese echando chocolate en las tazas. Uno de los niños de más corta edad dijo que el chocolate ascendería hacia el cielo, y el resto de los niños, salvo los de mayor edad, se mostraron de acuerdo. Los que manifestaron su desacuerdo (probablemente basados en sus propias experiencias) dijeron que el chocolate rebosaría y se vertería sobre la mesa. La madre aprovechó esta oportunidad para realizar el experimento y enseñar a los niños lo que realmente sucede.

En nuestra vida cotidiana cualquier toma de decisiones lleva implícita la incertidumbre. La naturaleza de la incertidumbre dependerá del problema a tratar.

Las siguientes preguntas tipo conducen a la toma de decisiones:

¿Es el acusado culpable del delito que se le imputa?; ¿es el último hijo más inteligente que el primero?; ¿cuál será el precio del petróleo dentro de dos meses?; ¿cuántos grados aumentará la temperatura del planeta en el próximo lustro?; ¿puede una aspirina, tomada a diario, reducir el riesgo de ataques de corazón?

No sirve dar respuestas filosóficas a las situaciones formuladas en los interrogantes anteriores; tampoco podemos encontrar respuestas definitivas con la información disponible, ni existen reglas precisas para seleccionar una respuesta adecuada sin estar sujeta a error.

Abstenerse de tomar decisiones puede considerarse como la alternativa para evitar errores. No obstante, no podríamos hablar de progreso si aceptásemos esa forma de proceder. Por ello, la mejor forma de actuar a la hora de tomar decisiones es optimizar, es decir, minimizar el riesgo que cualquier alternativa lleva implícita. El razonamiento inductivo y la cuantificación de la incertidumbre proporcionan una respuesta a este problema.

Los conceptos de incertidumbre y aleatoriedad han desorientado a la humanidad durante largo tiempo. Cada día, nos enfrentamos a la incertidumbre en el entorno físico y social en el que vivimos. Soportamos incertidumbres de la naturaleza y sufrimos sus catástrofes.

Algunos teólogos argumentan que nada es aleatorio excepto Dios, ya que es la causa de todo lo que sucede; otros afirman que incluso Dios puede estar a merced de algunos eventos aleatorios. En el libro *El jardín de Epicuro,* Anatole France escribe: "El azar es acaso un pseudónimo de Dios cuando no quiere estampar su firma."

Estos teólogos y varios filósofos contemporáneos de Aristóteles no contemplaban la posibilidad de estudiar el azar o medir su incertidumbre.

> "Las leyes científicas no avanzan mediante un principio dictatorial o pueden justificarse mediante fe o filosofía medieval; la Estadística es el único tribunal hacia el nuevo conocimiento."
>
> (P.C. Mahalanobis, 1954)

R. Rao, en su libro *Statistics and Truth,* enfatiza y discute sobre la necesidad de buscar estrategias en el análisis de datos para extraer información relevante de lo observado y ocuparse de la incertidumbre.

La Estadística es un instrumento ineludible para tratar de estudiar y cuantificar la incertidumbre y su uso es inevitable en la búsqueda del conocimiento. En nuestra era las actividades humanas están basadas en predicciones: elección de un trabajo, en qué valores bursátiles habrá que invertir, etc.

Estamos, asiduamente, tratando de predecir bajo incertidumbre y de tomar una decisión que nos posibilite el mejor resultado. La incertidumbre está presente en nuestra realidad. Incertidumbre debida a la falta de información, al desconocimiento en el uso de la información disponible, a los errores producidos por mediciones defectuosas incluso con instrumentación sofisticada, a catástrofes naturales o a los caprichos del comportamiento humano (el más impredecible de todos los fenómenos).

Es importante reflexionar sobre cómo cuantificar la vaguedad formulando estrategias para reducir, controlar y modelar la incertidumbre. En esta búsqueda podemos considerar la Estadística como una tabla de salvación.

> "Estadística es el estudio de cómo debe ser utilizada la información para mostrar los hechos y evidenciarlos mediante la acción en situaciones prácticas donde la incertidumbre esté implícita."
>
> (V. Barnett, 1973).

La palabra azar, término usado en Estadística para describir un fenómeno aleatorio, está vinculada a la incertidumbre y al concepto de probabilidad. El azar, "la suerte", sirve para describir fenómenos aleatorios como la extracción de los números de una lotería. Una secuencia de números, así generada, exhibe un cierto orden a lo largo de la vida del proceso, orden que puede explicarse mediante el Cálculo de Probabilidades. De otro modo, se observa que los números generados mediante un procedimiento determinístico pueden mostrar localmente un comportamiento aleatorio dentro de su regularidad global.

Cabría intentar la contraposición de lo estocástico, cósmico o azar sujeto a leyes, con lo errático o caótico. El azar se ocupa del orden en desorden mientras que el caos trata el desorden en orden. Ambos términos son de enorme relevancia en el modelado de los fenómenos observables. El aprendizaje de los fenómenos "aleatorios" y sus relaciones con el estudio del azar y la incertidumbre ha llevado, en los últimos treinta años, al estudio de la *teoría del caos*.

Se han propuesto nuevas aproximaciones para modelar complejas formas y sombras tales como la formación de una nube, una turbulencia, la línea costera de un país, e incluso para tratar de explicar variaciones en los precios de valores mediante el uso de ecuaciones matemáticas relativamente sencillas. Esta forma de pensar, en algún sentido diferente, lleva implícito un mecanismo debido al azar para tratar de describir las salidas de un sistema.

La introducción de la Probabilidad como medida de la incertidumbre fue iniciada por Niccolo Fontana, llamado "Tartaglia" (1500-1522), y Gicolano Cardano (1501-1576), y recibió decisivo impulso con Blais Pascal (1628-1662). La probabilidad de un evento se indica con una medida normalizada, comprendida entre 0 y 1. Por otro lado, se entiende por posibilidad de un evento o un suceso que éste pueda ocurrir, acaecer, tener lugar. La posibilidad de un suceso está situada entre dos extremos: seguridad o certeza, e imposibilidad.

La investigación científica es un proceso de aprendizaje dirigido. El objeto de los métodos estadísticos es hacer que ese proceso sea lo más eficiente posible.

> "La vida es el arte de extraer conclusiones suficientes de insuficientes premisas."
>
> Samuel Butler (véase, por ej. C. R. Rao, 1989).

> "Adivinar es barato, adivinar erróneamente es caro."
>
> (Antiguo proverbio chino).

La investigación científica no posibilita una vía única de resolución del problema. Dos o más investigadores, enfrentados a un mismo problema, partirán en general de premisas distintas, continuarán su investigación por sendas diferentes y aun así podrán llegar a la misma conclusión. Lo que se trata de lograr no es la uniformidad, sino la convergencia.

Aun bajo la suposición de que los datos científicos no contengan ruido, la inducción de la realidad inherente a sistemas complejos es muy difícil. La presencia de errores experimentales hace la tarea toda-

vía más costosa. La convergencia hacia el resultado se producirá, en mayor grado y con más certeza, si el investigador dispone de métodos eficientes de Diseño de Experimentos. Estos métodos le permiten obtener respuestas menos ambiguas y afectadas en menor medida por los errores experimentales. Puede ser importante un análisis de sensibilidad de los datos. Un diseño experimental mal elegido conduciría a una carencia de información relevante en los datos, por lo que poco se podría extraer de ellos por muy sofisticado que sea su análisis. En contraposición, un adecuado diseño conllevaría información relevante y podría no ser necesario un complejo análisis estadístico.

La Estadística no es ningún procedimiento mágico que permita obtener cualquier conclusión que se desee con independencia de la información disponible, sino que es un método riguroso de análisis de las observaciones basado en hipótesis fundamentales que no pueden ser violadas. No sólo es importante conocer los métodos estadísticos, sino también sus limitaciones. Un determinado procedimiento puede ser de utilidad para responder a una o varias cuestiones, bajo ciertas premisas; pero el incumplimiento de alguna de estas condiciones invalidaría el método.

El usuario no informado suele cometer graves errores en el análisis de sus datos. Es importante planificar de forma adecuada una experiencia a fin de que los resultados garanticen una respuesta apropiada a las suposiciones originarias de la investigación y ello conduzca a un mínimo coste. Un planteamiento equívoco o defectuoso puede llevar a la obtención de datos inservibles que habrá que desechar, o a valores incorrectos que, siendo también desechables, son muy difíciles de localizar, o simplemente, a datos insuficientes o excesivos con un alto coste asociado.

Método

El método científico se basa en dos tipos de razonamientos: el deductivo y el inductivo, como ya hemos mencionado con anterioridad. Una investigación empírica usa ambos modos de razonamiento siguiendo un ciclo deductivo-inductivo. El método estadístico es el

procedimiento por medio del cual se sistematiza y organiza este proceso de aprendizaje iterativo. Las etapas básicas en este proceso pueden ser: planteamiento del problema; construcción de un modelo; recogida de los valores muestrales; depuración de los datos; estimación de los parámetros estadísticos; contrastes de simplificación; crítica y diagnosis del modelo propuesto.

Los métodos estadísticos no son la panacea para eliminar en su totalidad las principales fuentes de dificultad a las que se enfrenta el investigador. Sin embargo estas técnicas sí pueden aliviar, en gran medida, las dificultades más generalizadas. Así, el error experimental considerado como la variación producida por factores distorsionantes, tanto conocidos como desconocidos, puede reducirse significativamente. La confusión entre correlación y causalidad puede evitarse y puede clarificar la complejidad de los efectos estudiados.

Consideremos, a modo de sencillo ejemplo, un estudio experimental de los efectos del alcohol y del café en los tiempos de respuesta de conductores de automóviles sentados en un simulador. Supongamos que se ha observado que, sin haber ingerido café, un trago de alcohol incrementa el tiempo de respuesta en una media de 0,45 segundos, y, sin haber tomado alcohol, una taza de café reduce el tiempo de respuesta en una media de 0,25 segundos. Al evaluar los efectos de varias copas de licor y varias tazas de café y su efecto combinado, se podría simplificar mucho el estudio si ambos efectos fueran lineales y aditivos. Si el efecto del alcohol no es lineal y existe una iteración entre el alcohol y el café, habría que optar por un diseño experimental que generen los datos.

Por desgracia, están bastante extendidas determinadas afirmaciones y usos estadísticos erróneos y ello abona el campo a otras afirmaciones despectivas que sólo reflejan desconocimiento del método estadístico o una crítica al uso inadecuado del mismo. En ocasiones, la célebre frase de Disraeli: "Hay tres tipos de mentiras: mentiras, condenadas mentiras y mentiras estadísticas", puede reflejar un principio negativo de la sociedad ante esta disciplina.

Un singular modo de descubrir los diferentes errores, las manipulaciones, voluntarias o no, y los tratamientos y conclusiones erróneos

consiste en estar avezado en el lenguaje y en el método estadístico. Debemos renovar nuestra apreciación de lo que la Estadística supone como unidad dentro de la diversidad. El único antídoto frente a una posible manipulación es un conocimiento básico de la metodología estadística, consiguiendo participar, de forma efectiva, en la argumentación pública (basada en cifras y datos) consustancial a la vida democrática.

Una tarea que los estadísticos hemos descuidado es la de explicar a la sociedad las posibilidades y el valor de la experimentación. La sociedad necesita las explicaciones convenientes para poder tener una base sólida a la hora de decidir si quiere que la experimentación tenga lugar. Por otro lado, la Estadística se ha convertido en una herramienta fundamental e indispensable para la investigación, tanto en las ciencias experimentales como en las basadas en observaciones. Su importancia ha sido reconocida hasta tal punto que la revista *Science*, en 1984, calificaba el desarrollo y la difusión de los métodos estadísticos para interpretar datos en condiciones de incertidumbre como uno de los 20 desarrollos científicos "más significativos" entre los ocurridos en el siglo XX, por su impacto sobre nuestra forma de vida y sobre nuestra forma de conocernos a nosotros mismos y al mundo que nos rodea.

Vivimos, pues, en la era de la Estadística y cada aspecto de la actividad humana se mide e interpreta en términos estadísticos. Es una falacia considerar que la importancia real de las técnicas estadísticas es proporcional a su complejidad aritmética. De hecho, algunos métodos gráficos y otras herramientas de análisis descriptivo son extremadamente útiles para la interpretación de datos reales pese a la sencillez de su aparato matemático.

Concepción actual y logros

La finalidad de este epígrafe será la de tratar de difundir una concepción actual de la metodología estadística como un proceso iterativo de aprendizaje.

Los objetivos de la Estadística no son los de la Matemática. Tradicionalmente, la enseñanza de la Estadística ha estado vinculada a la

enseñanza de los modelos matemáticos de la Estadística. El matemático, en muchas ocasiones, por su tendencia hacia la "ciencia exacta", ha desdeñado el aprendizaje y la enseñanza de importantes fases del método estadístico, tratando de convertir la Estadística en una ciencia abstracta carente de nexo con la realidad y totalmente alejada de sus orígenes como ciencia eminentemente aplicada.

Lindley (1988) señala: "La Estadística no sólo es matemática sino que también se refiere al mundo real."

Sir Clauss Moser, insigne estadístico inglés, destacaba, en una conferencia pronunciada en el Instituto Nacional de Estadística en la década de los noventa, que al asistir a reuniones y congresos internacionales de Estadística con frecuencia se sentía anonadado porque, después de toda una vida dedicada a la profesión estadística, apenas podía entender un pequeño porcentaje de las ponencias y comunicaciones allí presentadas debido a su elevado grado de abstracción.

La Estadística trabaja con datos, trata de dar solución a problemas reales, de tomar decisiones y hacer inferencias en ambiente de incertidumbre, de diseñar y analizar experiencias que permitan extrapolar información, de deducir conclusiones acerca de la población basadas en las observaciones de una muestra, etc. La práctica profesional de la Estadística ha requerido una actitud interdisciplinar, en la que los conocimientos del estadístico, la información sobre el significado real de los datos proporcionada por los profesionales de otras áreas que los han producido y los métodos de cálculo accesibles en cada momento histórico, se han combinado para posibilitar soluciones a los problemas de investigación y de decisión que se han ido planteando a lo largo de la historia.

Se producen opiniones encontradas sobre el papel que juega la Estadística en la sociedad. Para unos, es la ciencia en la que las diferencias individuales quedan ocultas a través de las medias; para otros, es la ciencia mediante la cual se manipula la opinión desde la publicidad, la tecnología o la economía; para una minoría, facilita una metodología para evaluar y juzgar las discrepancias entre la realidad y los modelos.

En las últimas décadas hemos asistido a una rápida evolución de la metodología estadística. Dos hechos lo han motivado: el increíble desarrollo de los métodos de cálculo y la Informática. La evolución 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 sólo unos años eran intratables. La Informática, a través de nuevos paquetes y programas más potentes, flexibles y de fácil manejo, ha revolucionado también el análisis y tratamiento de la información.

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, en la búsqueda 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. El vertiginoso auge de Internet hace posible, en tiempo real, acceder y lograr información relevante sobre el problema a tratar, así como consultar opiniones o establecer debates sobre las conclusiones obtenidas entre distintos profesionales.

Los programas interactivos permiten analizar los datos de manera secuencial, de forma que los resultados de una etapa puedan utilizarse para decidir lo que se hace en la siguiente, sin que ello exija largas esperas intermedias. Con este *software*, de uso sorprendentemente sencillo, un profesional no estadístico puede explorar sus datos experimentales o dinámicos, económicos, sociales, médicos, etc., por sí mismo. La posibilidad de que el propio experto en el tema pueda participar en el análisis de datos es esencial. Un conocimiento profundo del contexto en el que se encuadran los datos hace mucho más probable que se descubran relaciones importantes en el proceso de analizar datos que en la simple observación de una tabla de resultados estadísticos o de una colección de representaciones gráficas.

Elementos esenciales de todo análisis exploratorio y del estudio de fiabilidad son la detección e interpretación de observaciones atípicas (*outliers*) y el tratamiento de observaciones incompletas (*missing values*).

El uso adecuado de la información conseguida a partir de un análisis de datos sofisticado exige unos conocimientos mínimos sobre Estadística. Es necesario difundir la importancia de los procedimientos interactivos y gráficos de última generación, y acostumbrar a los investigadores y profesionales de otras áreas a trabajar desde el principio con datos reales, no depurados, cuyo contexto estén en condiciones de entender. Es interesante prevenir al usuario no profesional de la Estadística contra un uso indiscriminado de algunos paquetes estadísticos. En palabras de Hartley (1980), "los resultados de un uso indiscriminado de paquetes informáticos pueden ser deplorables. De hecho, en ciertos casos, somos testigos de una producción altamente eficiente de resultados completamente irrelevantes para el problema estadístico que se analiza." Un buen modo de actuación, para evitar este abuso, consiste en recalcar la importancia de las opciones que presentan los citados paquetes para la diagnosis y crítica del modelo utilizado.

Las nuevas metodologías: técnicas de remuestreo como el "jacknife" o como el "bootstrap", la idea de robustez y el estudio de procedimientos de validación y diagnóstico, junto con el fácil acceso a las autopistas de información, pueden potenciar la metodología estadística y permitirán tratar problemas dinámicos y multivariantes de creciente importancia en nuestra sociedad.

Aplicaciones

Para utilizar las técnicas estadísticas hay que tener presente que todos los problemas manifiestan particularidades que deben analizarse antes de que se adopten los procedimientos más efectivos para su resolución. Además de esta consideración, habrá que tratar de averiguar cuanto se pueda del problema: ¿Cuál es el objetivo de la investigación? ¿Qué tipos de datos se tienen? ¿Cómo han sido recogidos? ¿Quién o quienes los han recogido? ¿En qué orden? ¿Cuál es la relación entre ...?

Los métodos estadísticos son más efectivos cuando se combinan de forma adecuada con el conocimiento del tema al que se aplican. No deberíamos olvidar, pues, el conocimiento no estadístico para lograr una definición precisa de los objetivos. Es sorprendente cómo, muy a menudo, se

ignora este conocimiento o no se le presta suficiente atención, circunstancia que conlleva dificultades y a veces, puede conducir al desastre.

Es loable aprender unos de otros, buscar la interconexión de la teoría con la práctica. Un claro ejemplo lo tenemos en Sir Ronald Fisher, eminente estadístico e investigador laudable, que trabajó íntimamente unido a otros científicos. Para él no había mayor placer que discutir problemas con otros colegas frente a una jarra de cerveza. Su amigo y colaborador William S. Gosset, mejor conocido por "Student", era considerado por muchos como asesor estadístico de las fábricas de cerveza Guinness; para otros, era un cervecero que dedicaba su tiempo libre a la Estadística. En ambos científicos había una enorme preocupación y conexión entre las investigaciones estadísticas que realizaban y los problemas prácticos que tenían que resolver. Sus éxitos como investigadores y sus habilidades para desarrollar técnicas estadísticas útiles estaban muy relacionados con sus intensas participaciones en el trabajo experimental.

¡Yo no creo en modelos!, ¡los modelos no sirven para nada! ¿Cuántas veces en la vida profesional nos hemos encontrado con expresiones de este tipo? En base a éstos y otros comentarios similares podemos formularnos la siguiente pregunta: ¿cómo puede seguir existiendo esa opinión en determinados sectores sociales? La respuesta a este interrogante no es sencilla. Una posible explicación puede ser la falta de conocimiento. Desconocimiento tanto de la ciencia en sí como de sus potencialidades de aplicación.

Un gerente o un presidente de empresa no tiene por qué saber Estadística; por lo tanto, este desconocimiento no le es imputable, o por lo menos no en su totalidad, al máximo gestor. Si a la falta de información sobre potencialidades de la Estadística se añade la carencia de enfoque práctico en estudios y propuestas de otros escalones del organigrama no es de extrañar los resultados en este campo de ciertas empresas españolas frente a las de otros países más competitivos.

Los profesionales de la Estadística tampoco estamos exentos de culpa; se trata de lograr un acercamiento a las empresas, asesorándolas en problemas específicos y estableciendo eslabones de colaboración entre las universidades y las empresas.

Tratar de enumerar los posibles campos de aplicación de la Estadística sería tan prolijo como enumerar todas las ciencias del saber. No obstante, creemos interesante presentar en este apartado una breve sistematización de diversas aplicaciones estadísticas.

- Cronológicamente, el primer objetivo que, como aplicación, ha abordado la Estadística es la descripción de datos, la búsqueda de procedimientos adecuados para resumir la información en ellos contenida.

- Otra meta es la elección de muestras representativas para resolver problemas como:

  - La orientación de la estrategia electoral de un partido político.
  - La interpretación de un test de inteligencia.
  - La previsión de la demanda potencial de un producto mediante un estudio de mercado.

- Otro objetivo frecuente suele ser dar una respuesta adecuada a preguntas como:

  - ¿Son efectivos el cinturón de seguridad o la limitación de velocidad para reducir la gravedad de los accidentes de tráfico?
  - ¿Cómo se relaciona el rendimiento escolar con variables familiares y sociológicas?
  - ¿Cuál es la relación entre paro o inflación?

A cuestiones como éstas debemos facilitar una respuesta en términos estadísticos que determine y mida las relaciones entre variables físicas, sociales o técnicas.

- La previsión de la evolución futura de muchas variables económicas y físicas, que presentan cierta inercia, es otra meta de la Estadística aplicada.

En un sinfín de situaciones, formularios, etc., los números han sustituido desde hace tiempo a los términos descriptivos, a menudo de forma muy beneficiosa. El empleo de metros, gramos y grados centígra-

dos tiene claras ventajas sobre expresiones como bastante corto, muy ligero y poco caliente. Hace algún tiempo, febril tenía un significado tan vago como simpático o antipático. En áreas más sutiles, como tonos musicales, timbres de voz y grados de inteligencia, los números tienen cuantiosa utilidad. El color es un ejemplo sorprendente en este sentido. Los aparatos físicos pueden medir longitud, peso y temperatura; los colores se manipulan por medio de aparatos sofisticados que mezclan y difuminan el color, y la inteligencia se evalúa con un test estándar.

¿Es posible hacer uso de la Estadística para explicar algo tan difícil de precisar como es el significado? Admitida esta suposición, ¿no se perdería todo el romanticismo de la poesía y el encanto de la elocuencia?

En el trabajo estadístico que referiremos a continuación, la persona actúa como un instrumento de medición en vez de hacerlo como un objeto de estudio. Su impresión subjetiva, obtenida y analizada de modo muy cuidadoso, conforma la medición.

Imaginemos, por un momento, que un amigo nos describe como reservado o introvertido, o bien como espontáneo o extrovertido, y pensemos en las diferentes sensaciones que nos producirían estas descripciones. Cuando éstas y otras palabras se utilizan para redactar informes, la elección de unas u otras puede tener repercusiones muy diferentes.

No sólo el empleo cotidiano de estas palabras es muy distinto, sino que en otros campos se presenta la misma ambigüedad de significado; esto sucede tanto con palabras comunes como con términos técnicos, tales como esquizofrénico o autista.

Podemos clarificar el significado, ¿pero cómo? La Estadística puede aportar, en principio, una aproximación a la solución del problema, y un método de análisis:

- ¿Qué es lo que queremos? Una descripción de términos relativos a rasgos de personalidad, según su significado.
- ¿De qué medios disponemos? Un método estadístico como es el escalograma multidimensional debido a Rosenberg.

- ¿Qué datos trata? Similitudes o disimilitudes entre cada par de palabras de entre un número selecto más usado.

- ¿Qué resultados aporta? El resultado del escalograma multidimensional es un mapa o esquema. Este mapa carece de escala y no tiene ninguna orientación espacial. La cercanía de los ítems en el mapa debe corresponder a pequeñas disimilitudes, y el alejamiento a grandes disimilitudes; o viceversa para similitudes entre significados de palabras.

El catorce de noviembre de 1985, Gay Taylor, un estudioso de Shakespeare, localizó un poema de nueve estrofas, oculto desde 1775, en un volumen encuadernado de una colección de la Biblioteca "The Bodleian".

Nuevo poema atribuido a Shakespeare: Una Oda a la Estadística:

> *Nor marble, nor the guilded monument of princess,*
> *Shall outlive this powerful rhyme.*
>
> Shakespeare.

El poema analizado consta de 429 palabras. ¿Podía atribuirse a Shakespeare? Dos estadísticos, Thisted y Efron (1987), realizaron un estudio del problema de anonimato y concluyeron que el poema podía encajar, fácilmente, en el uso de palabras al estilo de Shakespeare.

El número total de palabras distintas usadas en las obras conocidas de Shakespeare es de 884. Aunque la complejidad de un poema pueda medirse por el número de palabras utilizadas, dicha medida no es lo suficientemente sensible como para distinguir, fidedignamente, entre autores que escriben con estilos similares. Las variables usadas en diversos estudios de autoría dudosa son las *razones de ocurrencia* de palabras individuales específicas. La elección de determinadas palabras en una obra, lo que se denomina registro, constituyen detalles cuya presencia proporciona una fuerte indicación de la autoría de uno de los posibles autores. Los registros, cuando se detectan, pueden contribuir a la discriminación, pero también pueden presentar dificultades a la hora de atribuir la autoría de un poema.

Las cuestiones de autoría son frecuentes y, a veces, importantes. La mayoría de las personas han oído hablar de la controversia Shakespeare-Bacon-Marlowe sobre quién escribió las grandes obras generalmente atribuidas a Shakespeare. Una cuestión menos conocida, pero ampliamente estudiada, se refiere a la autoría de ciertos escritos religiosos cristianos llamados paulinos, algunos de los cuales forman parte de los libros del Nuevo Testamento. Un análisis estadístico puede contribuir a la resolución de muchos problemas de autoría.

Una parte esencial de la Estadística, como ya se ha dicho, consiste en inferir en situaciones difusas. En vez de pensar en una palabra como en un registro, podemos tomar la razón o frecuencia relativa del uso de cada palabra en los escritos del autor como medida para distinguir su obra.

El estudio de las distribuciones de frecuencias de palabras distintas en poemas escritos por Shakespeare y su comparación con los cánones seguidos por Samuel Johnson, Christopher Marlowe y John Donne, autores contemporáneos de Shakespeare, en escritos de similar longitud, muestran que el poema puede atribuirse a Shakespeare.

Otro modelo de autoría lo conforma el análisis estadístico llevado a cabo para estudiar el complejo problema sobre la autoría de los doce artículos federalistas de la Historia Americana. Una discusión extensa de este problema, que incluye detalles históricos, discusión de técnicas actuales y una variedad de análisis lo proporcionan F. Mosteller y D. L. Wallace (1964).

Estas referencias son muestras de cómo la Estadística puede facilitar soluciones válidas para un país, para la ciencia, o para las personas que inicialmente plantearon el problema.

A continuación, citaremos ejemplos de aplicaciones de la estadística en diversos campos:

- La preocupante situación de las ballenas. D.G. Chapman.

  Los métodos para estimar el tamaño de las poblaciones de ballenas son de gran ayuda para promulgar leyes para su con-

servación. En este estudio se desarrollan métodos estadísticos para tratar de resolver cuestiones como: ¿Cuántas ballenas hay en la reserva que se alimenta en la Atlántida?, ¿cuántas nacen cada año?, ¿cuántas mueren por causas naturales cada año?, ¿cómo se ven afectadas estas tasas de nacimiento y muerte por factores que puede controlar el ser humano?

- Cómo podemos conseguir que las calificaciones de los exámenes sean más justas mediante la Estadística: H. I. Braun y H. Wainer.

  En este experimento, se distribuyen exámenes de alumnos entre diferentes correctores para ajustar las diferencias entre correctores a lo largo del tiempo con el fin de obtener calificaciones más justas.

- La importancia del ser humano: W.W. Howells.

  Se utiliza un procedimiento estadístico para determinar si un esqueleto fosilizado corresponde a un hombre o a un mono; además puede servir para tratar de determinar la edad de la raza humana.

- Evaluación preliminar de un nuevo producto alimenticio: E. Strett y M. B. Carroll.

  Experimentos y encuestas pueden determinar el grado de aceptación del sabor y de las características nutritivas de un nuevo producto alimenticio.

- La memoria infantil de la información gráfica: D. R. Entwiste y W. H. Huygins.

  Un experimento cuidadosamente diseñado demuestra que a una edad temprana los niños son capaces de recordar con gran precisión dibujos que se les había mostrado previamente.

- Optimización y el problema del vendedor ambulante: C. A. Whitney.

  Los métodos aleatorios de búsqueda son de gran ayuda para determinar cómo se debe orientar de modo más eficiente un

telescopio hacia un grupo de estrellas, en qué orden se deben visitar una serie de ciudades, en qué lugar se debe inspeccionar en busca de gas natural y cómo optimizar el recorrido en el problema del viajante o vendedor ambulante.

Éstas y otras aplicaciones lingüísticas, taxonómicas, de comportamiento humano, sociales, físicas, etc., se recogen en la obra *La Estadística: una guía de lo desconocido* (1992).

Aplicaciones a las ciencias de la salud

- Influencia de la herencia en las enfermedades: D. D. Raid

  Comparaciones entre gemelos idénticos y no idénticos proporcionan información sobre las enfermedades en los que la herencia juega un papel importante.

- La seguridad de los anestésicos: L. E. Moses y F. Mosteller.

  Un estudio nacional utiliza el muestreo y la normalización de tasas para comparar la seguridad de diversos anestésicos utilizados en cirugía.

- El precio del tabaco, el hábito de fumar y la política tributaria: K. E. Warner.

  Mediante el análisis de regresión, los economistas estiman la demanda de tabaco y los efectos sobre la salud pública de una disminución de los impuestos sobre el tabaco. De este modo se presiona al Congreso estadounidense para que mantenga el impuesto, y además, posiblemente, se disuade a muchos adolescentes de fumar.

  Los datos que proporcionan las ciencias englobadas bajo el epígrafe Ciencias de la Salud presentan un problema de variabilidad biológica de los sujetos experimentales: dos seres vivos nunca son iguales, ni un mismo ser es igual a sí mismo en diferentes etapas de su vida. Esta variabilidad sólo puede ser descrita en términos estadísticos. En el mundo clínico, destacaremos dos objetivos prioritarios: el diagnóstico y el tratamiento.

A ellos añadiremos un tercer objetivo, el pronóstico, importante en muchos casos para una adecuada elección del tratamiento. La Estadística puede contribuir a alcanzar estas metas.

En palabras de la Organización Mundial de la Salud, "en todos los dominios de las Ciencias de la Salud, en su vertiente clínica, administrativa o de investigación, es indispensable conocer los principios estadísticos para comprender bien los problemas y el profesional de la salud necesita de los datos estadísticos para tomar decisiones válidas."

Aplicaciones económicas y políticas

- Un método para contar mejor: aplicación de la Estadística para mejorar la calidad de un censo: M. H. Hansem y B. A. Bailer.

  Métodos de encuesta muestrales han contribuido a hacer más exacto el censo de Estados Unidos.

- Desarrollo y análisis de los indicadores económicos: G. H. Moore.

  Para predecir y medir la tendencia de la economía, los expertos combinan información procedente de centenares de indicadores económicos.

- Medición de los efectos de las innovaciones sociales con series temporales: D. T. Campbell.

  Datos recogidos antes y después de una reforma social son útiles a la hora de evaluar su efectividad.

- Noche de elecciones en televisión: R. F. Link.

  La predicción correcta de los resultados en unas elecciones requiere procedimientos adecuados de recogida de datos y un modelo estadístico especial para controlar los errores.

Las aplicaciones a las ciencias de la salud, las aplicaciones económicas y las políticas reseñadas en el texto pueden, también, encontrarse en el libro *La Estadística: una guía de lo desconocido* (1992).

Sobresale en estas aplicaciones, como una primera fase del método estadístico, la recogida y depuración de los datos. Uno de los problemas más graves que se plantea el usuario de las estadísticas (económicas y demográficas) es la depuración de los errores y falsedades que contienen las respuestas de los encuestados a las preguntas formuladas en los cuestionarios.

El gran economista Morgenstern (1970) señala al respecto: "Las estadísticas económicas y sociales se basan, con frecuencia, en respuestas evasivas y mentiras deliberadas de varios tipos. Estas falsedades nacen, en principio, de malas interpretaciones, del miedo a las autoridades fiscales, de la incertidumbre o disgusto de los planes o del deseo de confundir a los competidores. Nada de esto ocurre en la naturaleza."

Las contestaciones falsas o sesgadas son corrientes en cualquier campo de investigación estadística; basta considerar, por ejemplo, el fracaso de las encuestas electorales en las que el encuestado no dice la verdad respecto a su intención de voto.

En lo que se refiere a la información económica, se producen valores erróneos al formular preguntas como el volumen de los ingresos de una persona física, el valor de las ventas o el volumen de producción de un empresario. Las cifras de producción obtenidas en un censo industrial suelen ser disparatadas. Sin embargo, los profesionales de la Estadística pueden y deben estudiar y proponer metodologías que permitan elaborar estadísticas económicas básicas suficientemente aceptables o exactas para que puedan emplearse en el análisis económico. Un estudio poco científico y un análisis incompleto de las estadísticas conducen a conclusiones equivocadas.

La Estadística institucional y la Estadística pública tienen como objetivo prioritario el de devolver a los individuos, familias, administraciones y empresas informantes un cuadro coherente de la estructura y la evolución del colectivo social del que forman parte. La contabilidad nacional y las tablas "input-output" constituyen ejemplos de sistemas estadísticos que engloban multitud de estadísticas económicas, industriales y financieras aisladas.

De esta vocación de globalidad se desprende la capacidad de integración de perspectivas en conflicto que constituye la principal aportación de la Estadística a la formación de la conciencia colectiva. Como corolario de este objetivo de globalidad de la Estadística pública se deduce que su desarrollo exige un esfuerzo constante tanto para vencer la resistencia de los nuevos colectivos sociales que pueden beneficiarse de la oscuridad, como para lograr la transparencia que compense la tendencia de los gobernantes a convertir la Estadística en hagiografía política.

Ingeniería

La Estadística como herramienta básica para mejora de procesos en la industria alcanza un auge inusitado como consecuencia de su utilización masiva en Japón y de la adopción por las más importantes multinacionales occidentales de la filosofía de Deming sobre calidad y productividad.

El convencimiento de que los ingenieros no pueden permanecer al margen de innovaciones tan importantes como el diseño robusto de productos y procesos, de los conceptos de calidad de W. E. Deming, J. M. Juran y otros, o de la enorme cantidad de problemas industriales que se pueden resolver con las siete herramientas básicas de K. Ishikawa, ha sido el motor en esta área de aplicación.

En la época de los ochenta hemos sido testigos de un cambio extraordinario en los enfoques de los problemas de calidad y productividad en todos los sectores, y quizás la consecuencia más destacada de estos nuevos enfoques es el protagonismo que otorgan, a todos los niveles dentro de las grandes empresas, a la utilización sistemática de la Estadística.

Como considera Juran (1983), "ningún recurso es tan escaso en las empresas como el conocimiento estadístico. No hay conocimiento que pueda contribuir tanto a mejorar la calidad, productividad y competitividad de las empresas como el de los métodos estadísticos."

Ishikowa (1976) afirma: "Las herramientas estadísticas básicas deben ser conocidas y utilizadas por todo el mundo en una empresa, desde la alta gerencia a los operarios de las líneas."

Para el ingeniero de este nuevo siglo la Estadística será un arma de trabajo esencial para liderar la mejora continua de la calidad y de la productividad en todos los procesos que de él dependan.

Aplicaciones a las Nuevas Tecnologías

Mencionaremos, en último término, algunas aportaciones relevantes y de expansión futura de la Estadística.

> "La Estadística es la ciencia cuyo objeto es la obtención y el análisis de datos mediante el recurso a modelos matemáticos y a herramientas informáticas."
>
> (R. Gnanadesikan, 1977).

Es la palabra datos la clave en la definición anterior, como afirmó Joiner (1986): "Los datos son el centro de nuestra ciencia, el centro no es la variación aleatoria ni la probabilidad".

Los datos pueden presentar contaminación. Las observaciones anómalas y atípicas deben ser objeto de especial atención. En un censo, la edad de un individuo no debería superar un cierto valor de años admisible. Podría dudarse de las respuestas 100 años de edad o 110 años; sin embargo, si la respuesta fuese 160 años quedaría definitivamente rechazada.

Existen diferentes alternativas en el tratamiento de situaciones anómalas. Una propuesta sería no prescindir de dato alguno, una vez recogidos, si bien se debería poner el máximo esmero y atención en su recogida. La supresión de las observaciones anómalas constituye otro criterio contradictorio con el anterior. La sustitución automática de las observaciones anómalas por imputación o asignación de valores sustitutivos, puede ser otra perspectiva a considerar; por supuesto que el reemplazamiento de valores debería efectuarse en el menor número posible de ocasiones, para evitar, en lo que se pueda, la reconstrucción de datos que siempre puede dar lugar a sesgos, más o menos previsibles. Un procedimiento complementario de actuación es el empleo de estimaciones robustas, que elimina o atenúa el efecto de posibles valores anómalos a cambio de pagar un moderado precio por la pérdida de información.

Cabría formular una similitud entre la situación que se acaba de describir y la restauración de monumentos u obras de arte históricas.

Una forma de proceder sería preferir la ruina a la falsificación; otra restaurar haciendo distinción entre lo auténtico y lo restaurado.

Es evidente que, además de la naturaleza probabilística del método estadístico, existe una ambigüedad no cuantificada en cuanto a los supuestos. Ocurre esto con todas las hipótesis simplificadoras y con la de independencia e idéntica distribución de los elementos de una muestra. De aquí la enorme importancia, en el campo de las aplicaciones, de los procedimientos robustos, poco vulnerables al no cumplimiento de dichas hipótesis simplificadoras.

Desde comienzos de 1996, el Estudio General de Medios (EGM) y los paneles de audiencia como Media Metrix o Net Value, con una muestra procedente de más de 40.000 entrevistas anuales, analizan la evolución del uso de Internet en España. A estos efectos, se recoge a través de encuestas y de un *software* adecuado, tanto información general sobre la población como información sobre el uso del ordenador, acceso, utilidad de la red, etc.; así como datos adicionales asociados al uso del medio: lugar de acceso, perfil sociodemográfico, etc.

Sencillas técnicas estadísticas sirven para resumir la información y contestar cuestiones del tipo: ¿cuántos usuarios tiene la red en España?, ¿cuál ha sido la evolución de los medios digitales? ¿cuáles son las páginas más visitadas?, etc.

Con esta información, es posible determinar el grado de satisfacción de los usuarios respecto a sus proveedores.

Data Mining es una nueva disciplina posicionada como interfaz de la Estadística, Tecnología de Bases de Datos, Reconocimiento de Patrones, Ingeniería del Conocimiento y otras áreas. Esta materia está teniendo una vertiginosa difusión, sobre todo en el mundo financiero y económico, por aportar estructuras de conocimiento que pueden guiar a la toma de decisiones en condiciones de certeza limitada.

Data Mining puede entenderse como un proceso analítico de búsqueda de información valiosa o interesante en enormes conjuntos de datos, para facilitar la toma de decisiones desde un mejor conocimiento del entorno. Aunque se fundamenta en principios esenciales del

análisis exploratorio, presenta nuevos cambios y desafíos en el enfoque de los problemas tradicionales estadísticos. Los gigantescos conjuntos de datos implican que se originen insólitos problemas.

Las bases de datos actuales contienen millones de registros; así, el proyecto del genoma humano, del que todos hemos oído hablar, lleva recopilados miles de millones de bits. Las técnicas estadísticas estándar, por sí solas, no pueden manejar cifras de estas dimensiones.

Tres etapas se consideran importantes en el proceso de *Data Mining*: exploración, construcción de modelos o reconocimiento de patrones, y validación o verificación de los modelos propuestos. Idealmente, si la naturaleza de la información lo permite, se repetiría iterativamente el procedimiento hasta llegar a identificar un modelo "robusto".

La Estadística puede favorecer el desarrollo de futuras posibilidades de aplicación de esta técnica; así, desde el proceso de organización del almacenamiento de grandiosos conjuntos de datos multidimensionales, pasando por la búsqueda de relaciones sistemáticas entre el amplio conjunto de variables consideradas en el estudio, para finalizar con la validación de los modelos o patrones detectados en los nuevos subconjuntos de datos.

Un requerimiento clásico en los análisis estadísticos, del que ya hemos hablado, es la depuración de observaciones anómalas, atípicas y faltantes. Una estrategia ideal sería la de tratar de buscar la fuente que los ha producido. En un análisis mediante *Data Mining*, es casi seguro que un elevado porcentaje de observaciones no será válido, en algún sentido. Este porcentaje será especialmente alto cuando los datos describan interacciones humanas de alguna clase, tales como información de mercado, de transacciones financieras o de recursos humanos. La contaminación es, por supuesto, otro problema importante presente en estudios de *Data Mining*. También habrá que tener presente la asunción de la independencia y la idéntica distribución de los datos; si los datos no satisfacen los requerimientos del análisis estadístico, deberán examinarse las variables para determinar el tratamiento pertinente para adecuarlas al mismo.

No siempre será posible conseguir de forma sencilla patrones interesantes de búsqueda o determinar alguna estructura en los datos.

La situación puede complicarse cuando existan patrones o modelos de comportamiento que puedan, sencillamente, ser producto de fluctuaciones aleatorias y no representar una verdadera estructura subyacente. Habrá ocasiones en que necesitemos procesar los datos en tiempo real. Los resultados de un análisis obtenidos en septiembre, pueden ser de poco valor para predecir el comportamiento de un determinado sector en el mes de julio. Las conclusiones no tienen un largo ciclo de vida, se necesitan soluciones, rápidas, a corto y medio plazo.

Los métodos para la construcción de modelos en Data Mining incluyen técnicas de barrido en el análisis exploratorio de datos; métodos gráficos dinámicos e interactivos; análisis *cluster* y de correspondencias; análisis de regresión y correlación; funciones de análisis discriminante; correlaciones canónicas; análisis de series de tiempo; métodos de clasificación y validación; métodos robustos; árboles de clasificación, etc. Es frecuente cada vez en mayor medida que las bases de datos contengan tipos de datos no numéricos: imágenes, audibles, en forma de texto, geográficos, etc. Un papel notable en el tratamiento de estos datos lo desempeña la teoría y las redes Bayesianas.

Desde una vertiente comercial se ha implementado una gran cantidad de *software* para el análisis de *Data Mining*. La finalidad de este *software*, integrado por técnicas avanzadas, será lograr los máximos beneficios en el análisis de datos, preferentemente en los sectores económicos y financieros, para la identificación de hábitos de compra y de grupos de compradores, en el análisis de riesgos y para la detección y segmentación de mercados.

En el futuro desarrollo de *Data Mining*, la Estadística será una técnica complementaria que permitirá obtener conocimiento inédito en nuestros almacenes de datos y concretar respuestas que ayuden a la toma de decisiones. Nuevos problemas requerirán nuevas soluciones.

Conclusión

Creemos oportuno repetir algunos de los aspectos positivos en la introducción de la teoría y de las técnicas estadísticas, así como las ventajas que conlleva la aplicación de la Estadística a otros campos:

- Mejor comprensión y tratamiento de la realidad.

  El mundo real es en su globalidad inaprehensible para el ser humano. Hemos de efectuar abstracciones y simplificaciones, para tratar de trazar y establecer regularidades que nos faciliten su comprensión y nos ayuden en nuestras predicciones y en la toma de decisiones.

  Puede ser que el azar sea una manifestación de nuestro desconocimiento, pero es indudable que existe un sustrato permanente y profundo de incertidumbre en los fenómenos.

  La aplicación a fenómenos reales de la teoría estadística y de sus técnicas y procedimientos permite una mejor comprensión y tratamiento de situaciones y procesos de naturaleza aleatoria o estocástica, y difusa o imprecisa.

  Los rápidos cambios que se están produciendo en todas las facetas de la actividad humana, acompañados, por una parte, del desarrollo y la evolución de la tecnología y de las comunicaciones para la transmisión y el procesamiento de la información; y de otra, por la enorme cantidad de datos almacenados, abren un amplio abanico de estudio a la experimentación y a la investigación.

- Estímulo y cooperación en los esfuerzos de síntesis.

  La investigación científica es una aventura singular que descubrimos desde diferentes sendas, pero con un método y una perspectiva común: la búsqueda de la verdad y la comprensión del mundo y de nosotros mismos.

  La evolución científica es un proceso continuo y es tarea de los investigadores preparar el camino a las nuevas generaciones.

  Debemos contemplar la producción estadística más como un sistema estadístico coherente que como un conjunto de estadísticas aisladas. La creación de un sistema estadístico contribuye, por una parte, a la armonización de clasificaciones, registros y nomenclaturas y, por otra, a la detección de errores en la veracidad y consistencia de los datos individuales.

Es conveniente estimular la creación de equipos multidisciplinarios que favorezcan la síntesis e integración de perspectivas.

- Entendimiento y comunicación.

Sería infructuoso que la labor de un profesor universitario quedase relegada a una simple enseñanza del conocimiento recuperado y almacenado en nuestras instituciones universitarias. Otro cometido, no menos importante, consiste en adaptar el lenguaje de la Estadística a los conocimientos y a la sensibilidad de los usuarios finales.

Es evidente que debemos dedicar gran parte de nuestro tiempo a estar con los estudiantes, tratando de solventar, en la práctica, problemas reales no estructurados. Abordar situaciones reales, o, al menos, realistas, relevantes, debe ser una finalidad para conseguir un serio conocimiento por el alumnado de que los métodos estadísticos son valiosísimas herramientas para el análisis y la toma de decisiones a las que tendrá que enfrentarse en el futuro ejercicio profesional de su carrera.

Tan importante como enseñar conocimientos concretos es despertar en el universitario una motivación positiva hacia la Estadística. Un claro ejemplo de este estímulo, a todos los niveles de la sociedad, es el modelo nipón.

A lo largo de lo expuesto hemos tratado de visionar cómo los métodos estadísticos pueden utilizarse para tratar de mejorar, no sólo las organizaciones industriales, sino también las organizaciones sociales, gubernamentales, etc. La Estadística contempla tanto individualidades como globalidades que influirán, de seguro, en el desarrollo de nuestra sociedad.

La Universidad, como institución, debe poseer siempre juventud y vigor para transmitir a la sociedad que el aprendizaje puede ser motivo de divertimento y la comprensión es posible que sea uno de los mayores placeres que la vida nos ofrece.

Bibliografía

Azorín, F. (1979). *Algunas aplicaciones de los conjuntos borrosos a la Estadística.* Instituto Nacional de Estadística. Madrid.

Barnett, V. (1973). *Comparative Statistical Inference.* Wiley.

Box, G. E. P., Hunter, W. G., y Hunter, J. S. (1989). *Estadística para Investigadores. Introducción al Diseño de Experimentos. Análisis de Datos y Construcción de Modelos.* Reverte. Barcelona.

Gnanadesikan, R. (1977). *Statistical Data Analysis of Multivariate Observations.* Wiley, New York.

Hand, D. J. (1998). *Database management; Statistics.* American Statistician, May 98, Vol. 52, Issue 2, pp. 112-117.

Hartley, H. O. (1980). *Statistics as a Science and as a Profession.* JASA, 75, 1-8.

Ishikawa, K. (1976). *Guide to Quality Control.* Asian Productivity Organization.

Joiner, B. L. (1986). *Transformation of the American Style of Teaching Statistics.* Center for Quality and Productivity Improvement (University of Wisconsin). Report 10.

Juran, J.M. et al. (1983). *Manual de Control de Calidad.* Reverté. Barcelona.

Lindley, D. V. (1988) *Statistical Inference concerning Hardy-Weinberg equilibrium.* In bayesian Statistics. (J. M. Bernardo, M. H. De Groot, et al, eds.), pp. 307-326.

Mahalanobis, P. C. (1954). *The foundations of Statistics.* Dialectica 8, 95-11.

Morgenstern, O. (1970). *Sobre la exactitud de las observaciones económicas.* Tecnos. Madrid.

Mosteller, F. And Wallace, D. L. (1964). *Inference and disputed Authorship: The Federalist.* Reading, Mass: Addison-Wesley.

Rao, C. R. (1989). *Statistics and Truth.* Council of Statistics and I Research. New Delhi.

Tanur, J. M., Mosteller, F., et al. Eds. (1992). *La Estadística: Una guía de lo desconocido.* Alianza. Madrid.

Reseña biográfica

Mª Asunción Rubio de Juan nació en Madrid el 16 de agosto de 1956. Ingeniera por la Escuela Técnica Superior de Ingenieros de Montes de Madrid, en 1980. Doctorada en Informática, en 1985, por la Universidad Politécnica de Madrid. En 1981 se incorporó a la Facultad de Informática de la U. P. M., donde ingresa en el Cuerpo de Profesores Titulares de Universidad. Durante 1985-86 realiza el Curso de Postgraduados del Centro de Formación del Banco de España, Especialidad de Economía Cuantitativa. Se traslada en 1988 al equipo de profesores de la Universidad de Extremadura. Desde 1992 es Catedrática de Universidad del área de Estadística e Investigación Operativa, puesto que desarrolla en la actualidad en la Escuela Politécnica de la UEx. Profesora visitante y colaboradora del Departamento de Estadística de la Universidad Carolina de Praga. Especialista en Análisis de Series de Tiempo. Miembro de la Sociedad Española de Estadística e Investigación Operativa.

ÍNDICE