Metodología de Encuestas
Vol 3, Núm 2, 2001, 227-250

¿EXISTE ALGO MÁS FÁCIL QUE UN ANÁLISIS ESTADÍSTICO?

Antonio Solanas
Lluís de Jover
María Isabel Núñez
Jordi Ocaña
Lluís Salafranca
Begoña Valle
Bienvenido Visauta
Universidad de Barcelona

Joan Manuel Batista-Foguet
Escuela Superior de Administración y Dirección de Empresas

Ana López
Universidad de Sevilla

Rafael San Martín
Universidad Autónoma de Madrid

Jordi Simó
Bertelsmann-Direct

RESUMEN

En este artículo se exponen y argumentan diferentes razones, tanto a favor como en contra, referidas a la utilización de las aplicaciones informáticas desarrolladas para el análisis estadístico. La reflexión no se centra únicamente en torno a una relación de ventajas e inconvenientes. Se abordan otras cuestiones intrínsecamente relacionadas con la aplicación de la tecnología informática en el ámbito de la Estadística, como son la metodología, el razonamiento estadístico, la didáctica y un creciente mercado que demanda este tipo de programas informáticos. Las conclusiones enfatizan la importancia de la validez de los datos, sustentada en una adecuada metodología de investigación, y la necesidad de un amplio conocimiento del razonamiento estadístico. Advertir sobre los posibles problemas no significa en ningún caso obviar las aportaciones de estas aplicaciones informáticas.

Palabras clave: estadística, análisis estadístico, razonamiento estadístico, programas de ordenador.

El contenido de este artículo es una reflexión en torno a las ventajas e inconvenientes del análisis estadístico mediante las aplicaciones informáticas. Los diferentes autores elaboraron separadamente un documento sobre la temática y el coordinador realizó, posteriormente, una síntesis en un único documento que, tras varias revisiones por parte de todos los participantes, tomó su forma final. El texto es resultado de un amplio consenso que intentó cubrir las diferentes visiones. Debe considerarse la existencia de posiciones diversas en el origen, que fueron debidas a las distintas áreas de especialización e interés de los autores; pero, en última instancia, el artículo pretende ser ecléctico y recoger las distintas posiciones, aunque difícilmente puede ser exhaustivo. Al respecto, es importante mencionar que los distintos autores representan los ámbitos de la investigación aplicada y teórica, los estudios aplicados y el ámbito empresarial. También se pretendió proporcionar un enfoque que, en algún sentido, aportara una perspectiva algo diferenciada, motivo por el cual se intentó eludir una reflexión única y exclusivamente centrada en una relación de ventajas e inconvenientes, procurando potenciar nuevas líneas de debate desde posiciones diferentes.

Mercado, Mercaderes y Consumidores

Las aplicaciones informáticas ideadas para el análisis estadístico han sido asociadas desde sus orígenes con la tecnología científica y parece acertado afirmar que están todavía vinculadas con la investigación científica. Desde los inicios de la última década del siglo pasado no es tan evidente que esta tecnología se inscriba únicamente en el campo de la ciencia, pues la difusión y la utilización masiva de esas aplicaciones informáticas llega más allá de ese ámbito. En los últimos años se ha incrementado el número de empresas especializadas en el análisis de datos. Algunas consultoras ofrecen estudios estadísticos para solventar problemas empresariales. Estas empresas no utilizan exclusivamente aplicaciones estadísticas de propósito general, sino que también recurren a programas informáticos más específicos, donde se hallan todo tipo de algoritmos fundamentados en la computación intensiva. Estas nuevas compañías se han añadido al reducido grupo de empresas que existían con anterioridad. La proliferación de estas nuevas empresas se explica por la emergencia de un mercado que requiere soluciones a problemas relacionados con la producción, la comercialización, la gestión, la administración y la formación.

En el ámbito docente, y aproximadamente desde los inicios de la década de los ochenta, se han incorporado las aplicaciones informáticas en la enseñanza universitaria. Las aplicaciones informáticas para el análisis de datos se enseñaron inicialmente en cursos de doctorado, cursos de postgrado y masteres. Actualmente, en mayor o en menor medida, son utilizadas en diversas asignaturas pertenecientes al currículo docente de distintas licenciaturas y diplomaturas. Nuevas titulaciones universitarias han incorporado nuevos perfiles profesionales y, en algunos casos, entre las posibles funciones de estos técnicos está realizar análisis de datos. Con ello, las

empresas, públicas y privadas, disponen de profesionales cada vez más capacitados para realizar análisis estadísticos.

A la existencia de profesionales con mejor formación estadística en las empresas deben sumarse otros cambios. Por un lado, ya sea mito o realidad, se ha extendido la creencia según la cual disponer de grandes cantidades de información está directamente relacionada con el éxito empresarial, la excelencia investigadora y la adecuada gestión de los recursos. Esta idea está reforzada por la forma mediante la cual se denomina este periodo de la historia de la humanidad: sociedad de la información y la comunicación. La ingente cantidad de información disponible debe ser sintetizada, pues de otra forma resulta ininteligible. Por otro lado, las iniciales ideas sobre el control estadístico de los procesos productivos se extendieron a sistemas de gestión de la calidad total, pasando con el tiempo del ámbito de las empresas productivas a aquellas dedicadas a los servicios. Aúnese a esta idea el interés por establecer objetivos y medir su grado de cumplimiento, junto a la cada día más pujante motivación por evaluar cualquier sistema, y se advierte la aparición de nuevos datos que deben ser analizados e interpretados en orden a extraer información relevante.

Puede aceptarse o no, pero existen suficientes indicios a favor de considerar que en torno al análisis de datos existe un mercado, donde las empresas productivas, financieras, de servicios, las administraciones públicas, las instituciones docentes, los centros de investigación y los medios de comunicación social están, por una u otra razón, interesados en disponer de datos, realizar la explotación estadística de los mismos e interpretarlos. El mercado no es reducido y crece continuamente. Al respecto, es importante notar la existencia de otra creencia muy extendida: la idea de que todo aquel razonamiento apoyado en valores numéricos posee un valor añadido de rigurosidad y objetividad. No se trata de negar frontalmente esta concepción, pero sí matizarla. Si las medidas obtenidas no se fundamentan en un diseño metodológicamente correcto, donde se elimine o minimice al máximo posible la incidencia de variables contaminantes, y los instrumentos de medición no son precisos, todo el razonamiento posterior tiene un escaso valor.

Este creciente mercado no ha pasado desapercibido para las compañías dedicadas al desarrollo de aplicaciones estadísticas y las empresas distribuidoras de esta tecnología informática. Como era predecible, aceptado el principio de maximización de beneficios, estas compañías han convertido estas aplicaciones en productos de consumo. La veracidad de esta afirmación puede comprobarse si se observa que cada día es más frecuente que el análisis estadístico se incorpore como una herramienta más en aplicaciones informáticas inicialmente destinadas a otros propósitos, el crecimiento en la inversión dirigida a la publicidad de las aplicaciones estadísticas, la proliferación de cursos de formación destinados a usuarios sin un perfil estadístico y la incorporación de nuevos módulos informáticos orientados a la aplicación en las empresas productivas y al sector de servicios. Es suficiente recuperar alguna versión de las diversas aplicaciones para notar que estaban compuestas por un conjunto de técnicas principalmente orientadas hacia la investigación teórica y la aplicada. Pronto aparecieron más y más módulos que podían ser utilizados en estudios de marketing y en los análisis de cuestionarios obtenidos mediante encuestas, junto a todas aquellas técnicas destinadas al modelado de series cronológicas de datos y

obtención de intervalos de predicción de las mismas. No olvidemos los módulos destinados al análisis estadístico de procesos productivos, centrados especialmente en el control estadístico de procesos. Los últimos desarrollos se centran en procedimientos para la obtención de información mediante la red informática mundial, creación de formularios de respuesta y técnicas para la computación masiva e intensiva, siendo estas últimas generalmente aglutinadas bajo la denominación de *minería de datos* (inducción de reglas, redes neuronales, entre otras).

Parece razonable aceptar la premisa que establece que las aplicaciones estadísticas son productos y, por tanto, están sujetas a las leyes del mercado. ¿Y qué requiere el consumidor?:

- En primer lugar, resolución de problemas. O sea, una aplicación informática destinada al análisis de los datos debe solventar un problema empresarial, de investigación o educativo.
- En segundo lugar, facilidad de operación. La mayoría de los operarios y técnicos de las empresas, así como los investigadores, no son expertos estadísticos, pero precisan justificar sus informes o conclusiones ante un órgano de decisión, sea éste un consejo de dirección o unos revisores científicos. Las aplicaciones informáticas aseguran esta facilidad mediante un sistema de menús y simples operaciones consistentes en la selección de variables, proceso en el cual sólo intervienen algunas pulsaciones de botones y desplazamiento de objetos en un monitor. Si se asiste a una sesión de presentación de una aplicación estadística, se advierte cómo los comerciales se afanan en destacar la facilidad y automatización mediante la cual se manejan estas aplicaciones. Las predicciones de ventas se obtienen en cinco minutos y el usuario ni siquiera ha tenido que tomar elección alguna sobre el modelo empírico o teórico.
- En tercer lugar, rapidez de cálculo. No es preciso que quienes se dedican al desarrollo de estas aplicaciones inviertan mucho tiempo en mejorar sus algoritmos, pues el incremento continuo de la velocidad de los procesadores garantiza el progresivo incremento de la rapidez de cómputo.
- En cuarto lugar, un adecuado servicio de ayuda al cliente o consumidor tras la realización de la venta. No nos referimos en este punto ni a la atención al cliente ante la eventual aparición de diversos problemas informáticos relacionados con la aplicación, pues se da por supuesto, ni a los cursos de formación, que también se considera básico en la actual concepción de la atención al cliente. Estamos refiriéndonos a los asesores estadísticos que proliferan como complementos de algunas aplicaciones informáticas destinadas al análisis estadístico. Preguntas relacionadas sobre la naturaleza de las variables, el diseño de investigación y los objetivos son las entradas de un sistema experto que libra al usuario de decisiones relacionadas con la metodología y el razonamiento estadístico. Existe algún asesor estadístico que proporciona interpretaciones de los resultados obtenidos. ¡Nada más fácil que realizar un análisis estadístico!

- En quinto lugar, la comunicación. El razonamiento numérico y estadístico no es la estrategia habitual mediante la cual se aborda la solución de un problema. La estética es una forma más comprensiva para un amplio sector de consumidores, aunque se han publicado varios libros de divulgación donde se remarcan las ventajas derivadas de abordar numéricamente y razonar estadísticamente muchos de los problemas a los cuales nos enfrentamos (Rao, 1989; Paulos, 1990; Paulos, 1995; Dawkins, 1998; Paulos, 1998). Además, el dicho establece que una imagen vale más que mil palabras. En consecuencia, si el consumidor tiene esta creencia, no debe extrañar la importancia que las compañías otorgan en la actualidad a los gráficos. Mercaderes y comerciales destacan frecuentemente las posibilidades que dispone el consumidor para elegir entre varias tramas, colores y posibles disposiciones en el espacio tridimensional mediante rotaciones de los gráficos; sin olvidar que, además, el foco de luz puede llegar desde distintas direcciones y crear sombras a conveniencia. Tal belleza parece confirmar que una imagen vale más que mil palabras, salvo que se olvida el ínfimo detalle de que este tipo de manipulaciones altera la percepción que se alcanza del objeto y, en consecuencia, la interpretación de los resultados. Es por esta razón, y no otra, que los editores de las revistas científicas sugieren contumazmente que las representaciones gráficas sean austeras.

Actualmente, el desarrollo de los programas estadísticos debería entenderse única y exclusivamente como la resultante de los ajustes y mejoras que se realizan a un producto de consumo. Por tanto, la satisfacción del cliente y la innovación y desarrollo del producto son polos que se unen mediante un conexión que permite la alimentación de forma retrógrada y continuada. Resulta difícil mantener una visión romántica al calor del espíritu científico y técnico, que no se duda subsista, pero no pueden contemplarse las aplicaciones estadísticas sólo desde esta perspectiva.

La Cuarta Regla y el Conocimiento Estadístico

Una vez abordada la omnipresencia de las aplicaciones informáticas en el análisis de datos, es conveniente tratar qué se requiere y debe subyacer en el proceso relacionado con tal actividad. No se precisa gran perspicacia para advertir que en este punto tiene capital importancia la formación en Estadística. Al respecto, no parece suficiente el conocimiento de las tres reglas (leer, escribir y contar), sino que es necesario dominar una cuarta regla, que se corresponde con el razonamiento estadístico (Rao, 1989). El propósito de C. Radhakrishna Rao mediante este símil consistió en destacar la importancia de la incertidumbre en la vida cotidiana de las personas. En un solo día la inmensa mayoría de nosotros tomamos numerosas decisiones, bajo mayor o menor grado de incertidumbre, y, por tanto, parece razonable que adquiramos conocimientos que nos ayuden en la toma de decisiones. Como no existen evidencias a favor de que los seres humanos dispongan de mecanismos para manejar de una forma intuitiva la incertidumbre, sino todo lo contrario, es conveniente que el razonamiento estadístico sea el eje sobre el que debe rotar la formación

estadística, ya sea a nivel primario, secundario o universitario. Es muy necesario que la enseñanza del razonamiento estadístico no esté limitada a la formación superior, sino que se incorpore en los niveles fundamentales del sistema educativo.

Pero, ¿cuáles son los elementos básicos donde se fundamenta el razonamiento estadístico? Primero, la cuantificación de la incertidumbre. Segundo, la obtención de datos claros, relevantes y honestos. Tercero, la utilización de modelos apropiados para la extracción de la información. Y en cuarto lugar, el razonamiento científico. La formación debe realizar especial énfasis en el razonamiento estadístico, aunque la tecnología informática pueda ser una ayuda y, por eso, sea utilizada en la formación.

Se acepta que todos los profesionales de la estadística, por el hecho de serlo, son expertos en esa disciplina, utilizan adecuadamente las técnicas e interpretan con suficiente corrección los resultados estadísticos. No entraremos aquí sobre si la anterior creencia es cierta o no, pero sí nos ubicaremos en una hipotética situación. Supongamos que existiera un mundo donde algunos de los profesionales de la estadística no fueran tan expertos en esa disciplina como pudiera parecer. Aceptemos que hubiesen adquirido conocimientos estadísticos en libros de divulgación estadística y que aplicasen técnicas estadísticas sin excesivo rigor a datos correspondientes a diversas investigaciones. Quizás en ese mundo hipotético el análisis estadístico se fundamentaría en una aplicación biyectiva entre los problemas de investigación y las técnicas estadísticas al uso. ¿Sería posible en esa hipotética realidad transmitir unitariamente la importancia del razonamiento estadístico? Posiblemente no. ¿Cómo sería comunicado por algunos el conocimiento estadístico? Es razonable suponer que se trataría de argumentar y transmitir un conjunto de recetas mágicas y crípticas a los clientes o consumidores. No sería aventurado aceptar que se argumentaría y aludiría a la escasez de conocimientos matemáticos y estadísticos disponibles por los destinatarios y al interés por la divulgación para justificar esa aplicación biyectiva entre problema y técnica estadística. Este tratamiento superficial sería lógico que estuviera acompañado por un conjunto de herramientas informáticas o aplicaciones estadísticas que se limitarían a requerir que un usuario buscara, o incluso le ayudara un asesor estadístico, el elemento del conjunto imagen de la mencionada aplicación biyectiva que estuviera en correspondencia con el elemento del conjunto origen. Quizás en un mundo así es donde las aplicaciones estadísticas se desarrollarían como módulos cerrados, donde el usuario sólo podría elegir entre un conjunto reducido de opciones ya predefinidas, que corresponderían a una visión del análisis estadístico que hubieran fijado un grupo de expertos. Añadamos a este mundo imaginario que los clientes, en general, no tuvieran especial interés por conocer el proceso inherente al razonamiento estadístico y las técnicas estadísticas. ¿No sería maravilloso que sólo presionando un botón apareciera un único dato que debe ser interpretado? ¿Y no sería fantástico que un valor superior o inferior a 0.05 dividiera escolásticamente el mundo en verdad y falsedad?

Imaginemos ahora un mundo en el cual la totalidad de los profesionales de la estadística conocieran los fundamentos matemáticos inherentes a las técnicas estadísticas, el cálculo de probabilidades y el razonamiento estadístico. Supongamos también que los consumidores quisieran conocer no tanto la operativa o el saber

cómo hacer las cosas sino el por qué de las mismas. Ni unos ni otros se conformarían con una difusión de la estadística fundamentada en la transmisión y recepción de simples aplicaciones biyectivas simplificadoras de la realidad. De hecho, ¿alguien ha encontrado esa aplicación biyectiva en el análisis de datos? Conocemos hasta un cierto nivel el problema u objetivo principal de una investigación o estudio, pero no es evidente, y seguramente en absoluto cierto, que exista una única forma de analizar estadísticamente un problema concreto. Los estadísticos tienen una tarea mucho más apasionante que buscar entre el listado de recetas aquella correspondencia que relacione el problema con una o varias técnicas estadísticas. El análisis estadístico es más complejo y más laborioso, debiéndose explorar los datos e intentar hallar las diversas formas mediante las cuales destejer la enmarañada información. Para tal fin es útil la experiencia obtenida en anteriores análisis realizados, pero no suficiente. En cada estudio o investigación el estadístico descubre alguna información que guía el siguiente paso del análisis. Como consecuencia, en muchos casos se toman caminos que conducen a ninguna parte y, las menos, llevan a un diamante de información. Convendría que los estadísticos explicáramos que el análisis estadístico es un largo proceso de búsqueda realimentada a partir de los análisis realizados, que nos conducen a nuevas conjeturas, que, a su vez, llevan a nuevos análisis estadísticos, y así sucesivamente. Sería deseable abandonar esa práctica que consiste en atender a un problema de investigación que se nos plantea y responder a la solicitud del interlocutor para que le digamos qué técnica estadística debe aplicar. ¿Por qué? Simplemente porque es una actitud errónea ante el análisis de datos.

Sigamos en ese mundo tan idílico, siempre desde la perspectiva de aquellos a los cuales les interesa la Estadística, donde se imponen los conocimientos del razonamiento estadístico en profesionales y clientes. Las personas tendrían una visión de la realidad desde la incertidumbre y la información, conocerían adecuadamente las implicaciones de las inferencias estadísticas y entenderían la ciencia como un conjunto de conocimientos sujetos a una cierta aleatoriedad y temporalidad. Pero esa no es la realidad y, por tanto, es preciso reclamar una mayor difusión de los fundamentos estadísticos en los niveles más básicos de la formación, donde el razonamiento estadístico sea algo tan importante como los conocimientos de otras disciplinas. No es conveniente ni admisible una utilización de la tecnología informática sin esa formación tan importante que ha sido reclamada (Rao, 1989). Antes de aceptar un producto que los mercaderes se afanan en vender es preciso que pensemos en la cuarta regla. Si no se conoce, búsquese la formación que se precisa, pues esa tecnología no se utiliza asépticamente, sino que trata con datos. La información que se extrae de esos datos, incorrecta o no, incide sobre decisiones que, de una u otra manera, pueden afectarnos a todos. Pero no se interpreten estas palabras como la negación de que los mercaderes venden un interesante producto que necesitan algunos consumidores. La cuestión es que ese producto sólo interesa a algunas personas, que no son otras que aquellas conocedores del razonamiento estadístico y han sido instruidos en el análisis de datos; o sea, los estadísticos.

La Torre de Babel de la Estadística

A lo largo de la historia la *Torre de Babel* ha representado la expresión del saber. Cantidades ingentes de libros en todas las lenguas a disposición de peRsonas cuyo único objetivo era el conocimiento. Hablamos de la masificación de estudiosos interactuando en pro del descubrimiento de nuevas ideas. Similarmente, las aplicaciones informáticas para el análisis de datos aglutinan a gran cantidad de personas en búsqueda de conocimiento, razón por la cual podemos preguntarnos si estos programas son la Torre de Babel de la Estadística. ¿Podemos esperar que el uso masivo de estas aplicaciones se traduzca en una gran cantidad de nuevos e interesantes conocimientos, ya sean teóricos o aplicados? ¿Representan alguna clave especial en el desarrollo y la innovación del conocimiento?

La respuesta a las anteriores cuestiones es negativa, aunque nadie dudará que la difusión de las aplicaciones informáticas para el análisis estadístico supone mejoras, que trataremos en detalle, pero difícilmente incorpora notables avances en el conocimiento teórico o aplicado. Para proporcionar una respuesta positiva sería necesario que la difusión estuviera acompañada de una adecuada formación en el razonamiento estadístico. La deducción lógica, la inducción a partir de la repetición de los resultados, el control de las variables contaminantes, la validez de las medidas, la precisión o fiabilidad de la medición, la sagaz interpretación de los resultados y, en consecuencia, la adecuada revisión de las hipótesis y teorías son las auténticas herramientas para la adquisición de conocimiento. Una vez garantizadas estas condiciones indispensables, entonces las aplicaciones informáticas suponen una notable ayuda para tratar la información y adquirir más rápidamente conocimiento. Una investigación o estudio metodológicamente adecuado es la clave del desarrollo y no son imputables a las aplicaciones informáticas ni los éxitos ni los fracasos. Por tanto, las aplicaciones informáticas no son la Torre de Babel de la Estadística, sino una cadena de producción continuamente mejorada que incrementa la productividad, sin que exista relación alguna entre el incremento de la producción y la calidad del producto. Es muy importante recordar aquí que la cantidad sólo significa un aumento en la producción y, aunque puedan desarrollarse en paralelo, calidad y cantidad de producción no siempre son inseparables.

Un aprendizaje necesario para toda aquella persona que se inicie en la disciplina estadística consiste en cómo debe organizar y disponer los datos para su posterior análisis estadístico. El procedimiento habitual y estándar es disponer los datos en una estructura tabular o, si se prefiere, en una matriz de datos. En general, las columnas se corresponden con las variables y las filas con los casos. Esta disposición de la información facilita la comprensión de la idea según la cual los datos no son más que coordenadas en un espacio definido por las variables o los individuos, según convenga, para realizar uno u otro análisis estadístico. Los entornos de las aplicaciones informáticas para el análisis de datos suelen utilizar esta útil y óptima forma de disponer los datos para la mayoría de análisis, ya sea dentro de un editor de datos, una hoja de cálculo o un archivo externo para ubicar la información que será posteriormente analizada. La elección de esta forma de disponer la información

también resulta óptima para la programación de la inmensa mayoría de algoritmos informáticos destinados a realizar cómputos estadísticos.

La práctica totalidad de estas aplicaciones informáticas dispone, como mínimo, de un editor de datos, siendo más habitual que posean una hoja de cálculo con, al menos, las funciones más importantes para manipular la información. Este entorno o interfaz de usuario facilita, en general, la fase de introducción de datos al menos en tres sentidos. Primero, el entorno guía al usuario en la forma mediante la cual debe disponer la información. Segundo, agiliza la introducción de datos y mejora la calidad de la información existente en la base de datos, pues es posible en muchas aplicaciones establecer a priori un conjunto admisible de valores para las distintas variables. La detección automática de errores no sólo es una ventaja desde el punto de vista de la reducción del tiempo en la revisión de errores, sino que es una forma de control que también incide directamente en la validez de los datos que serán posteriormente analizados. Tercero, la posible disponibilidad de los datos en distintos formatos permite el intercambio entre diferentes aplicaciones informáticas. Una base de datos suele ser fácilmente exportable a distintos formatos o importada desde diferentes formas de grabar la información, favoreciéndose de esta forma el intercambio de la misma entre distintos usuarios sin apenas dedicar esfuerzo.

Una fase sumamente importante y quizás no siempre realizada del tratamiento de los datos alude a la depuración de los mismos. Cuando se han introducido los datos es frecuente hallar distintos tipos de errores que, por supuesto, nunca serán solventados por análisis estadístico alguno. Más bien todo lo contrario, pues estos errores serán arrastrados a una parte de los resultados obtenidos. ¿Cuáles son las fuentes de estos errores? En ocasiones proceden de una incorrecta trascripción de la respuesta a la hoja de registro por parte de un encuestador o el encuestado, proporcionando un valor no admisible. Otra fuente muy habitual de este tipo de error se produce al pasar la información de las hojas de respuesta o cuestionarios al editor de datos de la aplicación informática utilizada. Con menor frecuencia, aunque también puede suceder, se trata de un error del sistema informático que, aun teniendo una tasa de anomalías reducida, puede provocar cambios en la información grabada. En cuanto a la depuración de los datos, las aplicaciones informáticas proporcionan una ventaja indiscutible. No hace muchos años esta fase del tratamiento de datos resultaba engorrosa y pesada, pues se debían cotejar los datos uno a uno. Las aplicaciones informáticas nos permiten varias vías rápidas y muy seguras para detectar errores. Una consiste en, si la aplicación lo permite, definir los valores admisibles de cada variable antes de iniciar la introducción de datos. De esta forma se detectará una buena parte de los errores. Una segunda vía, si la aplicación no posibilita la detección de errores dado un rango de valores admisibles, es realizar un análisis descriptivo de cada variable para detectar la existencia de valores fuera de rango. Pero mediante estas estrategias sólo se detectarán errores correspondientes a valores no admisibles. Aquéllos debidos a la incorrecta trascripción de un valor por otro durante la introducción de los datos, pero ambos dentro del dominio, pasarán inadvertidos durante la fase de depuración. Existen estrategias para controlar este tipo de errores, pero no está tan claro que en este punto las aplicaciones informáticas permitan aho-

rrar tiempo respecto a otros procedimientos más tradicionales, como el cotejo uno a uno de los datos.

Parece poco cuestionable que uno de los puntos a favor para el uso de las aplicaciones informáticas sea la sencillez de ejecución. A lo sumo, y en algunos casos, es preciso introducir algún valor numérico para especificar un parámetro, un nivel de confianza determinado, una probabilidad de error tipo I, el número de factores para una solución dada, realizar una transformación, definir una función objetivo, entre otros. Incluso es posible con una sencillez pasmosa definir los ejes, las divisiones de los mismos, la representación geométrica de los datos en un gráfico y si estarán estos últimos unidos o no por líneas rectas. No es preciso extenderse en demasía en este punto, pues parece evidente que la sencillez está garantizada en estas aplicaciones informáticas.

La rapidez de cómputo es otro de los argumentos a favor de estos programas que tampoco parece cuestionable. Es suficiente considerar el coste temporal que suponía no hace tantos años obtener un valor del coeficiente de correlación lineal producto-momento o, si se prefiere, elaborar una representación gráfica. Pero la ventaja de la rapidez en el cómputo no debe limitarse a un simple y directo análisis comparativo sobre la reducción en el coste de tiempo invertido en el cálculo. Es preciso considerar que permite al analista de los datos disponer de un tiempo adicional para profundizar en el análisis de los mismos. Sólo la relación de análisis imprescindibles antes de alcanzar la solución última sería interminable, resultando de incuestionable valor en este punto una tecnología que permite ahorrar tiempo en el proceso de cálculo y, en consecuencia, disponer de un tiempo precioso para llevar a cabo estudios con mayor corrección estadística. En otros términos, el tiempo que nos permite ahorrar una aplicación informática para el análisis de datos no debe ser utilizado para incrementar la cantidad de bases de datos distintas que se analizan, sino invertido en mejorar el análisis estadístico de una base de datos concreta. No tenemos una total certeza de que eso sea así, pues sí se observa cada vez más un incremento en la cantidad de bases de datos analizadas, pero la dedicación a esos análisis puede que no sea todo lo rigurosa que cabría esperar.

En general, tampoco parece razonable rebatir que estas aplicaciones han mejorado la precisión de los cálculos obtenidos. Aquí nos referimos a la precisión de la representación numérica interna mediante la cual pueden operar estas aplicaciones informáticas, que permite obviar parte de los errores de cálculo derivados del conocido efecto producido por el redondeo. También se consigue alcanzar la práctica inexistencia de errores de cálculo mediante los algoritmos informáticos, lo cual resulta evidente si se compara con los errores de cómputo que se producían mediante cálculos con lápiz y papel. El error derivado del redondeo es muy conocido y, los que utilizaron calculadoras electrónicas o realizaron cálculos con lápiz y papel, han experimentado la desagradable sensación que produce ser consciente del efecto del redondeo y las limitaciones en el sistema de representación numérica. Con las aplicaciones informáticas los problemas no ha desaparecido totalmente, pero para la inmensa mayoría de las investigaciones y estudios no tienen relevancia alguna, pues es posible especificar a estas aplicaciones que operen con niveles de representación

numérica suficientes para evitar ese problema. En cuanto a la falta de precisión derivada durante el proceso correspondiente al algoritmo de cálculo, la comparación entre los cómputos realizados con lápiz y papel, o incluso con aquellos que se alcanzan con las calculadoras electrónicas, se decanta totalmente hacia el haber de las aplicaciones informáticas. Por supuesto, si se piensa en cómputos simples y escasamente voluminosos, la ventaja a favor de estas aplicaciones es casi despreciable, pero, cuando se consideran complejos algoritmos de cálculo y grandes bases de datos, es totalmente evidente. En cualquier caso, es conveniente recordar que no todos los procedimientos incluidos en las aplicaciones informáticas destinadas al análisis de datos son igualmente precisos, al menos si se consideran las distintas aplicaciones, pues su precisión y fiabilidad no siempre se ajusta a nuestras expectativas (McCullough, 1999).

Parece poco cuestionable que determinados estudios o investigaciones sobre voluminosas bases de datos no se hubieran realizado en caso de no disponer de programas para el análisis de datos. Es en este punto donde han encontrado un amplio campo para ofrecer sus productos algunas compañías privadas. Para analizar las ingentes cantidades de datos disponibles en los archivos de las organizaciones, ya sean aquellos adquiridos mediante los sitios informáticos desarrollados por las compañías o los obtenidos mediante tarjetas de crédito, algunas empresas de análisis de datos ofrecen servicios para extraer información. En general, esta información corresponde a perfiles de consumo y caminos o secuencias de acciones. ¿Por qué se realizan y qué finalidad subyace en estos análisis masivos? Básicamente se llevan a cabo estos estudios intensivos para conocer qué productos se pueden ofrecer a los actuales y los nuevos clientes y para predecir la pérdida de clientes e intentar evitar esa acción. De esta forma el *Data Warehouse* de las empresas y el *Data Mining*, básicamente sustentado en algoritmos de optimización matemática no lineales, se unen para extraer la información en orden a facilitar la toma de decisiones. Pero, ¿por qué estas empresas de análisis están logrando un alto grado de penetración en el mercado? Simplemente porque proporcionan soluciones a las empresas, tratando una cantidad elevada de variables y expresando los resultados en el lenguaje del negocio. En un mercado cada vez más competitivo, aunque el resultado de sus análisis diste de ser óptimo, la identificación de pequeños nichos de negocio, utilizando la jerga al uso, supone una ventaja comparativa respecto a la competencia, además de un mayor o menor beneficio. Difícilmente podrían obtenerse resultados si no se dispusiera de aplicaciones informáticas capaces de analizar grandes bases de datos con una gran cantidad de variables, estando estas vinculadas mediante complejas relaciones no lineales.

La rápida reproducibilidad de los análisis estadísticos y sus resultados es otra de las ventajas que aportan las aplicaciones informáticas. Actualmente no es en absoluto necesario almacenar ingentes cantidades de papel en el cual se contengan extensas listas de resultados estadísticos. Es suficiente crear un archivo informático con la secuencia de comandos que han sido ejecutados y, cuando se necesite repetir un análisis estadístico, se dispone casi de inmediato del mismo. No sólo se trata de un aspecto positivo desde una perspectiva ecológica, sino también de un ahorro de

tiempo y espacio, pues el usuario no necesita mantener extensos ficheros físicos, que ocupan espacio y, a veces, dificultan hallar la información.

La mayoría de las expresiones que se requieren para un conocimiento básico de la Estadística incluyen las básicas operaciones aritméticas, comunes funciones matemáticas y un modesto dominio del cálculo de probabilidades. No se precisa un notable conocimiento matemático para realizar los cálculos necesarios. La única salvedad quizás se refiera a los cómputos con algunas funciones de distribuciones de probabilidad correspondientes a variables aleatorias continuas, donde se necesita conocer cálculo matemático. Por tanto, y en general, no implica una excesiva dificultad obtener la mayoría de cálculos para describir una variable o estudiar la dependencia entre dos variables, entre otros análisis de interés. Ahora bien, esta situación ya no se mantiene cuando en el estudio o investigación se necesita analizar conjuntamente un amplio grupo de variables. Muchas de las técnicas estadísticas ideadas para el análisis multivariante requieren invertir matrices, determinar los componentes de varios vectores propios, realizar rotaciones de un eje de coordenadas, obtener gradientes de máxima pendiente, extensos procesos iterativos, entre una larga lista. No se sostiene aquí que sean imposibles estas operaciones con lápiz y papel o con la inestimable herramienta que es una pequeña calculadora electrónica, pero es incuestionable que las aplicaciones informáticas reducen notablemente el tiempo de cómputo. En este punto reside el apreciable crecimiento de estudios en los cuales se realizan análisis conjuntos de un amplio grupo de variables aleatorias, que, de otra forma, es más que probable que jamás se hubieran llevado a cabo. Puede decirse sin temor a equivocarnos que la Estadística Multivariante no hubiera alcanzado jamás las actuales cotas de utilización de no ser por las aplicaciones informáticas y, quizá se hubieran quedado como una curiosidad matemática y estadística (Lagarde, 1983).

Aunque pudiera parecer que la Estadística está compuesta por una serie de técnicas en las cuales se obtiene una solución analítica única, una definición que incluyera esta característica no sería una adecuada descripción de esta disciplina. Es cierto que en la mayoría de cursos de introducción a la Estadística o en la práctica totalidad de los libros de divulgación de la misma suelen aparecer técnicas para los cuales todos los cálculos poseen una única solución, generalmente alcanzada analíticamente, pero eso no es así para todas las técnicas. Algunas estrategias estadísticas se fundamentan en la progresiva minimización o maximización de una función objetivo o función de error, lográndose mediante un proceso iterativo que parte de unas condiciones iniciales que pueden ser variables. No sólo no se alcanza una solución única cada vez que se realiza el análisis de los mismos datos al modificar las condiciones iniciales, que, dicho sea de paso, suelen especificarse con una buena dosis de arbitrariedad, sino que, y más acorde a la idea que pretendemos exponer, se fundamentan en extensos procesos iterativos con cómputos más o menos complejos.

Si para una buena parte de los estadísticos la representación del conocimiento mediante valores numéricos es la mejor forma de destejer la compleja realidad, librándonos de las confusiones que pueden derivarse mediante otras maneras de codificar la misma, ciertamente no coinciden en ese punto ni todos los estadísticos

ni, por supuesto, la inmensa mayoría de los usuarios de técnicas estadísticas. Una amplia cantidad de personas prefieren las representaciones gráficas de la realidad a la pura descripción numérica. ¿Por qué? Posiblemente porque los gráficos llegan a un más amplio sector de personas, especialmente a aquellas que no están familiarizadas con la Estadística, pues la inspección visual de gráficos parece requerir un nivel menor de conocimientos matemáticos y estadísticos. No debemos descartar el factor evolutivo, si se nos permite denominarlo de esta forma, ya que los seres humanos, como otras especies, somos el resultado de un largo camino de evolución en el cual se han desarrollado mecanismos fisiológicos que nos permiten codificar la realidad de una determinada manera. Reconocer estímulos familiares, como formas y colores, además de reaccionar a estímulos infrecuentes, que captan nuestra atención, son características que poseemos. Las gráficas se fundamentan en estas estrategias, resultando especialmente atractivas cuando consiguen reclamar nuestra atención mediante una sorprendente mezcla de formas y colores. La combinación de ambas razones puede explicar por qué gozan de tanta aceptación las representaciones gráficas. Lo cierto es que, sea o no por esos motivos, aquellos que crean aplicaciones informáticas conocen muy bien las necesidades de sus clientes y por ello invierten grandes esfuerzos en el desarrollo de representaciones gráficas, cada vez con mayores niveles de operación para los usuarios. Tanto es así que existen compañías que están desarrollando sistemas para, mediante visores, conseguir que los usuarios dispongan de un efecto de representación tridimensional de los datos, desplazándose en el espacio mediante movimientos realizados mediante los brazos, las manos y la cabeza. En cualquier caso, una posición que es aceptada por la mayoría de los estadísticos consiste en que la representación numérica y la visual son necesarias. La razón reside en que la segunda ayuda a la interpretación de los resultados. Pero para la mayoría de usuarios se considera una aportación importante de las aplicaciones informáticas la facilidad y flexibilidad con la cual permiten manejar los gráficos. Sólo por captar personas que se interesen en el análisis de datos, y en la medida que las representaciones visuales constituyan una aproximación a la Estadística, los gráficos resultan de máximo interés. En este punto, de nuevo, debe reconocerse el papel de las aplicaciones informáticas, pues una parte de personas se aproximan a la Estadística quizá animados por la facilidad con la cual piensan que pueden interpretarse las gráficas.

El ámbito docente, y nos referimos a la disciplina de la Estadística, se ve beneficiado por estas aplicaciones informáticas. Es fácil realizar simulaciones y ejemplos que permiten ayudar a la comprensión teórica de conceptos clave (Scott y Jackman, 1999; Jackson y Smith, 2001) como la distribución muestral, los teoremas límite, los valores de probabilidad asociados a distintos valores de diferentes estimadores, la magnitud de los errores tipo II, las representaciones en el espacio de las variables, entre una extensa gama de posibilidades. La principal utilidad reside en la ayuda didáctica que supone para el profesor cuando debe comunicar conceptos abstractos. Es cierto que para muchos alumnos no es preciso tal tipo de recurso, ya que logran un notable conocimiento abstracto mediante el análisis. Ahora bien, pueden ayudar a aquellos menos familiarizados con el razonamiento matemático y estadístico. Tampoco sería conveniente otorgar una especial importancia a la aplicación docente de

estas herramientas, pues no olvidemos que no están diseñadas para tal fin, debiéndose considerar una utilización secundaria de estos programas (Biehler, 1997).

La última razón que mencionamos a favor de las aplicaciones informáticas destinadas al análisis de datos se refiere a cómo facilitan la comunicación y difusión de los resultados. Ya sea a los investigadores, estadísticos aplicados o profesionales que realizan informes técnicos, se les requiere la elaboración de artículos o informes donde se expongan los nuevos descubrimientos o resultados. Es evidente que se potencia la difusión a todos sus niveles, pero la comunicación científica y técnica debe hacerse por escrito. En este sentido, las aplicaciones informáticas para el análisis de datos, al explotar las funcionalidades de intercambio de información de los distintos sistemas operativos y sistemas de encriptación de la información, favorecen notablemente la elaboración de informes técnicos y científicos. No es nada desdeñable la comodidad mediante la cual incluimos un gráfico elaborado por medio de una aplicación informática y lo insertamos en un procesador de textos.

En ningún caso podemos pensar que toda la colección de argumentos a favor hacen de las aplicaciones estadísticas una Torre de Babel. Como ya ha sido mencionado, el fundamento principal reside en el razonamiento estadístico que, una vez garantizado su conocimiento y dominio, permite utilizar las aplicaciones informáticas como herramientas que nos proporcionan una inestimable colaboración.

La espada de Damocles de la Estadística

Hemos argumentado algunas de las ventajas derivadas de la utilización de las aplicaciones informáticas, pero ¿existe una espada de Damocles? Tal y como se desprenderá de las distintas razones que expondremos, y siempre que se opte por potenciar el cálculo numérico sin acompañarlo con el conocimiento del razonamiento estadístico, la respuesta debe ser afirmativa. La decisión de optimizar la rapidez de cálculo y el ahorro de tiempo derivado, junto a facilitar una herramienta técnica y favorecer su difusión, conlleva que la espada de Damocles caiga sin que se pueda evitar. En última instancia es una decisión humana aventurarse en el uso de una tecnología sin los conocimientos necesarios. Quizá también aquellos que participamos en la formación estadística no hemos sido muy eficaces a la hora de explicar que una cosa es la instrucción estadística, necesaria para todos, y otra, muy distinta, que de ésta se derive la capacitación para realizar análisis estadísticos. Pero el problema es mucho mayor y con muchos más agentes incidiendo en el futuro de la Estadística. Las revistas científicas requieren análisis estadísticos en los artículos y, por descontado, la mayoría de revistas no son de contenido estadístico ni sus autores tienen como campo de trabajo la Estadística; pero se nota un apreciable incremento en la cantidad de análisis estadísticos que se mencionan en las publicaciones científicas (Smith, 1996). Una parte de docentes universitarios transmiten a los discentes la necesidad de utilizar técnicas estadísticas, aunque la mayoría de los alumnos dedicarán su actividad profesional a cuestiones escasamente relacionadas con esta materia. Su función profesional no será llevar a cabo investigaciones o estudios ni para el desarrollo científico ni para obtener nuevo conocimiento. También se exige con

mayor énfasis a las administraciones públicas que justifiquen sus acciones y eso se traduce en una extensa cantidad de resultados estadísticos. Las empresas privadas destinadas al análisis de datos insisten en la importancia de extraer información y las empresas cuantifican sus procesos para maximizar beneficios y reducir costes. No olvidemos que algunas empresas han formado a sus operarios en control estadístico de procesos, implicando a los mismos en la mejora del sistema productivo. Todas estas fuerzas, entre otras, colaboran en transmitir la idea según la cual la Estadística debe ser conocida por todos. Como parece potenciarse el conocimiento sobre cómo se hacen las cosas más que la comprensión de aquello que se realiza, resulta difícil exigir a las personas que se limiten a un conocimiento teórico e intelectualmente enriquecedor. Por tanto, conocer es aplicar. En consecuencia, y como la cantidad de datos disponibles crece sin cesar, la mayoría de personas que tengan posibilidad y necesidad de utilizar técnicas estadísticas recurrirán a las aplicaciones informáticas. Por supuesto, y de forma irremediable, la espada de Damocles caerá.

La primera razón a la que haremos referencia, entre aquellos problemas derivados de las aplicaciones informáticas, es que favorecen la automatización del análisis estadístico. Podemos establecer un símil entre las actuales aplicaciones informáticas para el análisis de datos y las máquinas. No sería nada complejo instruir a una persona sin conocimientos de Estadística para que se responsabilizara del control estadístico de procesos, indicándole que cuando ocurra uno de entre una serie de hechos avisara a un técnico. No se advierte una diferencia entre el manejo de una máquina, reducido a una serie de rutinas y el análisis estadístico mediante estas aplicaciones. Una serie de acciones, se comprenda o no lo que implican, lleva a un resultado. ¿Cuál es el problema? Simplemente que no se requiere un especialista. Mientras un experto realiza su proceso de formación continuamente, además de los conocimientos que ya posee, no podemos esperar que aquellas personas que operan sobre estas aplicaciones de forma rutinaria tengan un interés especial por mejorar sus conocimientos estadísticos. Por mucho que a quienes estamos en el campo de la Estadística nos apasione esa temática, no podemos olvidar que otras personas contemplan esos análisis como un trabajo rutinario. En consecuencia, la automatización no es una garantía para extraer toda la información relevante de los datos ni la correcta utilización de las técnicas.

No abundaremos en la idea, pues se ha tratado ya, pero el uso masivo de las aplicaciones informáticas no permite contar con un mínimo de garantías sobre la calidad de los análisis realizados, una opinión con la cual no creemos que nos diferenciemos en exceso de otros autores (Abraira, Cadarso, Gómez, Martín y Pita, 2001). La tecnología estadística está ahí, y su facilidad de manejo también. En una sociedad fuertemente competitiva la ventaja diferencial es fundamental y, por esa razón, quien pueda demostrar un dominio de la tecnología que otros no poseen tiene un valor añadido en su perfil profesional. La estrategia óptima consiste en disminuir la ventaja del competidor y, si es posible, obtener otra que no posea. En esta dinámica está la Estadística y, por descontado, sus aplicaciones informáticas al ser ambas dos valores que pueden colaborar a conseguir un perfil diferencial.

Consecuencia de la automatización y la masiva difusión y utilización, se detecta una creciente falta de rigurosidad estadística. Se advierte la carencia de conocimien-

tos teóricos y flagrantes errores. De nuevo, la razón es la facilidad de acceso a ese tipo de tecnología en ausencia de firmes conocimientos estadísticos. En parte, la responsabilidad es de las empresas productoras y distribuidoras de estos programas, pues en el diseño del producto se favorece que cualquier persona no experta pueda operar ese sistema, mientras los distribuidores enfatizan la facilidad con que se pueden obtener interesantes resultados estadísticos. Pero también existe un error en el enfoque docente, si no se consigue transmitir que conocer algo de Estadística no es suficiente para realizar análisis estadísticos. Por supuesto, utilizar esas aplicaciones e interpretar los resultados es una decisión libre, pero sería aconsejable una dosis de autocrítica.

Un problema muy importante que se deriva de las aplicaciones informáticas es que el usuario, en general, se ciñe a la oferta de técnicas estadísticas que proporciona un programa. Eso se debe seguramente a dos motivos. Primero, es muy probable que sólo disponga de una única aplicación informática para el análisis de datos. Segundo, muchos usuarios considerarán que un programa de renombre debe contener todas las técnicas estadísticas que se correspondan con los distintos problemas de investigación. La segunda razón es especialmente preocupante porque, otra vez, aparece la pertinaz idea de la existencia de una aplicación biyectiva entre el problema de investigación y la técnica estadística. Ninguna aplicación informática para el análisis de datos es una Torre de Babel en la cual el usuario hallará la totalidad de técnicas que puedan ser necesarias para realizar un análisis estadístico, y eso sin entrar en los diferentes coeficientes e índices que pudieran ser precisos en esa investigación.

Es inhabitual que, en los resultados que proporcionan los programas, se advierta al usuario del incumplimiento de los supuestos requeridos por las técnicas estadísticas. Estas advertencias son más la excepción que la regularidad. Si el usuario posee conocimientos estadísticos suficientes, el problema queda mitigado. No afirmamos que se elimina el problema debido a que ningún estadístico es un profundo conocedor de la totalidad de técnicas estadísticas. Cuando quien realiza el análisis posee escasos conocimientos de Estadística, las consecuencias en la estimación y la decisión estadística pueden resultar catastróficas. Pero el punto crítico que queremos destacar en este momento es que resulta sorprendente esta aparente falta de interés por los supuestos en las aplicaciones informáticas. ¿Por qué estos programas no proporcionan mensajes escritos sobre los supuestos en los listados de resultados? ¿Cuál es la razón por la cual no se advierte al usuario sobre la violación de los supuestos? No es tan complejo ni supone constreñir al analista de datos; al contrario, se agradecerían los mensajes de aviso. La explicación reside en los inicios de estas aplicaciones informáticas. Al principio eran programas cuyo mercado era única y exclusivamente el ámbito universitario, donde la mayoría de investigadores que las utilizaban tenían conocimientos suficientes para realizar los análisis estadísticos. Para acabar de complicar la situación, esas iniciales aplicaciones requerían conocer un complejo conjunto de comandos, siendo necesario elaborar algo parecido a un programa informático para obtener un análisis estadístico, siempre bajo entornos no gráficos, como eran, por ejemplo, los sistemas operativos UNIX y DOS de la época.

Incluso podemos remontarnos unos años antes y recordar que los análisis estadísticos se obtenían mediante una tarjeta perforada. De hecho eran unas pocas personas quienes tenían la tecnología y los conocimientos estadísticos. Eran una pocas personas las que realizaban sus propios análisis estadísticos y los de los demás. En general, salvo algunas técnicas muy simples y de fácil cálculo, el resto de investigadores no utilizan complejas técnicas multivariantes. La mayoría de aplicaciones, que hoy en día se comercializan para el análisis de datos en entornos o interfaces gráficas de usuario, arrancan de aquellos tiempos que parecen remotos. Un notable número de sus actuales procedimientos se remontan a aquellos días. Ciertamente han sido retocadas las salidas de resultados, aprovechando las mejoradas posibilidades que ofrecen los entornos gráficos. Eso implica modificar el código informático, ya sea cambiando valores de variables de salida o mejorando la salida gráfica con nuevo código. Éste es el tipo de modificación que ha primado, junto a crear código nuevo para dotar a las aplicaciones de procedimientos totalmente nuevos. Ahora bien, completar el viejo código o dotar al nuevo código informático de líneas de instrucciones para advertir al usuario sobre la violación de los supuestos no ha sido una prioridad. Es más, en muchos casos los resultados para poder realizar el análisis de los supuestos no se halla en las opciones por defecto de los procedimientos. Mientras el análisis estadístico sea realizado por expertos, el problema será relativamente poco importante. Pero no olvidemos que para estas empresas informáticas los consumidores ya no son sólo expertos estadísticos y, como era de esperar, el escaso énfasis en los supuestos que se arrastra desde los inicios tiene efectos descontrolados sobre la decisión estadística. En general, el programa proporciona el resultado, se cumplan o no las condiciones.

Hemos defendido que deben entenderse las aplicaciones informáticas para el análisis de datos como productos. Desde esta perspectiva, las empresas que desarrollan estos programas están obligadas a mantener procedimientos y algoritmos durante un tiempo para asegurar su rentabilidad, no siendo evidente que estas compañías respondan con rapidez a los avances teóricos e incorporen con prontitud estos desarrollos en sus programas. Pero un problema más importante es que se mantienen algunos procedimientos anticuados o cuestionados, que seguramente se utilizan por la seguridad que transmite a muchos usuarios el aval que supone su inclusión en una aplicación informática de renombre (Bartholomew, 1997).

El Santo Grial de la Estadística

Entre argumentos a favor de las aplicaciones informáticas y aquellos en contra, ¿dónde está el *Santo Grial* que nos conducirá por un camino equilibrado entre el riguroso conocimiento estadístico y las ventajas computacionales? Por descontado no se halla en dar una respuesta más o menos tajante ni a favor ni en contra de las aplicaciones informáticas. Nuestro objetivo, y a modo de conclusión, es proporcionar algunas de las que consideramos pautas fundamentales para compatibilizar el conocimiento estadístico y las ventajas que proporcionan las aplicaciones informáticas.

El más importante de todos los requisitos para analizar datos es una notable formación. No nos referimos a las aplicaciones informáticas, sino al razonamiento estadístico (Rao, 1989; Altman, 1994). El dominio de la metodología y el razonamiento que sigue un estadístico es fundamental y no se puede reducir a un conjunto reducido y simple de reglas de decisión, al menos por el momento. El analista de datos debe conocer los principios del método científico, la teoría de probabilidades, las distintas técnicas de muestreo, el diseño de experimentos, las técnicas estadísticas, la interpretación de los resultados, la inferencia estadística y la revisión de las teorías y modelos. Cuáles son los errores metodológicos y cuál es su trascendencia, son aspectos que ya han sido descritos (Anderson, 1990). Por descontado, esa formación debe verse complementada con el conocimiento de aplicaciones informáticas, pues optimizarán la productividad del análisis de datos. Entonces, ¿qué profesional está capacitado para realizar análisis estadísticos? Parece razonable defender que sólo aquellos formados en una diplomatura o licenciatura de Estadística o los que hayan cursado un doctorado, postgrado o máster especializado en el análisis de datos. Puede ser conveniente que se programe para estos profesionales un periodo durante el cual deban realizar investigaciones y estudios aplicados, siendo supervisados por profesionales experimentados en el análisis de datos. En cualquier caso, parece cada vez menos sostenible que el análisis de datos sea realizado por profesionales cuya formación fundamental dista mucho de ser estadística.

En el puro ámbito docente, las aplicaciones informáticas para el análisis de datos sólo deben presentarse como una forma de reducir el tiempo dedicado al cómputo, enfatizar el aprendizaje en la extracción de la información contenida en los resultados estadísticos, entrenar en la interpretación de las representaciones gráficas y aprovechar las posibilidades de simulación que ofrezcan las aplicaciones para mostrar conceptos estadísticos. Las aplicaciones no son el fin en sí mismas, sino un medio o recurso más para agilizar el aprendizaje estadístico. Esta tecnología sólo debe ser un complemento para ayudar en la formación orientada al razonamiento estadístico. Es importante tener en cuenta que el significado y la información contenida en un valor correspondiente a un coeficiente de correlación jamás se ha entendido por calcularlo, ya sea a lápiz y papel, mediante calculadora electrónica o con un programa informático. Realizar un cálculo sólo implica que se ha aprendido por imitación, generalmente, un algoritmo de cómputo y nada más, siendo el conjunto de operaciones o pasos que lo componen diferentes según se utilice lápiz y papel, calculadora electrónica o un programa informático. Sostener lo contrario significa que entendimos el significado del operador producto por memorizar tablas de multiplicar y realizar, hasta la saciedad, distintas multiplicaciones Sólo mostrábamos ser organismos capaces de imitar y reproducir algoritmos. Sí es clave para el aprendizaje estadístico, no sólo que el aprendiz conozca el razonamiento estadístico, sino razonarle aquello que un índice, por ejemplo, pretende describir, y mostrar distintos valores indicando cómo puede interpretarse esa información y por qué se les dota de esa significación en cada caso concreto. Muy probablemente, al inicio, el aprendizaje sea puramente imitativo, en el sentido que se adquiere el mismo sistema de interpretación de quien enseña. También es posible que la mayoría de aprendices se que-

de a ese nivel puramente imitativo y reproduzcan aquello que se les enseñó; pero otros quizá alcancen un nivel de compresión, por medio del análisis y no por imitación, que les permita establecer su propio entendimiento del concepto y su personal criterio de interpretación, que pueden o no coincidir exactamente con aquellos que le fueron enseñados. En cualquier caso, por el bien del conocimiento científico, es esperable que se parezcan bastante.

Pero una docencia en parte sustentada en estas aplicaciones informáticas también tiene efectos no deseados. Es fácil que el alumno se demore al seguir la secuencia de pasos que se le propone realizar, el ritmo de exposición es más lento, las posibilidades de distracción aumentan y el profesor debe atender a distintos estados distintos del proceso de ejecución de la práctica. Para este conjunto de problemas, parece que una buena solución es incorporar alumnos del curso anterior para colaborar como monitores y ayudar al profesor en estas clases prácticas. En caso de no realizarse así, el profesor lleva a cabo una tarea hercúlea.

Sin abandonar el ámbito de la docencia, todos aquellos docentes responsables de la exposición de temáticas no estadísticas deberían enfatizar durante la formación el carácter no determinista de los conocimientos que disponemos sobre la realidad. Seguramente por simplificar y facilitar la comprensión de los discentes se exponen los conocimientos como un conjunto de relaciones deterministas. Nuestro conocimiento no es así, sino todo lo contrario, pues, ya sea en estudios aplicados o investigaciones científicas, las relaciones entre variables se nos presentan, en algún sentido, de una forma estocástica. Se podrá argumentar que estas relaciones probabilistas se nos muestran así debido a errores de los instrumentos de medición, a la falta de control en los experimentos o a la carencia de conocimiento. ¡Qué más da! En cualquier caso, la naturaleza se nos presenta de una forma aleatoria, en cierto sentido caprichosa, y, en consecuencia, así debería transmitirse. Cuando se enseña que los electrones siguen una trayectoria u órbita en torno a un núcleo, ¿por qué es necesario dibujar un círculo? Muchas personas no conocen qué es un orbital y no parece muy difícil entender ese concepto. La formación sobre la aleatoriedad debe iniciarse en fases iniciales de la educación y, a partir de ese momento, los profesores podrían mostrar la realidad tal como se nos muestra. Los efectos de un intrincado concepto estocástico de la realidad se traducirían en visiones totalmente distintas de la misma para la mayoría de personas y se facilitaría entender el razonamiento estadístico. Si la Estadística está destinada a jugar un papel fundamental en la ciencia (Rao, 1989), también es cierto que la interdisciplinariedad y visión estocástica en la exposición del conocimiento favorecería el aprendizaje, aunque no parece que sea un práctica muy extendida (Cordani, 2001).

Otro principio que debería seguirse es enfatizar menos la obtención de resultados e invertir más esfuerzo en la fase de diseño de los estudios estadísticos. O sea, el estadístico debe participar en las investigaciones desde el inicio, entendiendo al máximo posible el problema de investigación, planificar un adecuado muestro, previendo el control de las variables de confundido en el diseño del experimento, asegurando la validez y fiabilidad de las medidas, garantizando un adecuado nivel de precisión en las inferencias realizadas y colaborando en la interpretación de los resultados en relación a los objetivos teóricos de la investigación. Debe evitarse que el

estadístico inicie su participación cuando los datos ya se han obtenido y se le solicita sólo que explote estadísticamente los mismos. Esta práctica, que ha sido muy común, trivializa la profesión del estadístico, extendiendo la opinión según la cual el profesional de la Estadística únicamente se dedica a obtener resultados, cuando, de hecho, es experto en el razonamiento estadístico; además, favorece la difusión de la idea de que es suficiente ejecutar aplicaciones estadísticas para sustituir a un estadístico. El profesional de la Estadística no sólo debería renunciar a este tipo de análisis *ad hoc*, sino también manifestar públicamente la incorrección de este tipo de proceder.

Para evitar la dependencia de una única aplicación informática es conveniente disponer de varias herramientas de ese tipo. Es obvio que los programas comercializados no disponen de la totalidad de las técnicas estadísticas y, con frecuencia, no se encuentra en la aplicación que habitualmente utiliza el usuario aquella técnica o procedimiento que se necesita. Por tal motivo es importante conocer diversas aplicaciones informáticas (Goldstein, 1996), desde aquellas de propósito general a otras desarrolladas para análisis más específicos. No existen en la actualidad problemas para intercambiar datos entre los diferentes programas. Conociendo distintas aplicaciones, es más factible utilizar la técnica adecuada al problema de investigación y no caer en el error de buscar una aproximación estadística a la solución entre las técnicas incluidas en la aplicación que se utiliza habitualmente. En ocasiones, la técnica que precisamos no está en ninguna aplicación comercializada, en cuyo caso es recomendable buscar en la red informática mundial, en el caso de que quien ideó la técnica desarrollara una aplicación para realizar los cómputos. Esta búsqueda, en general, tiene éxito y, en caso contrario, sería conveniente contactar con quien desarrolló la técnica y explicarle que estamos buscando una aplicación que realice los cálculos del modelo que propuso. Esa acción también suele tener éxito.

Existe otra posibilidad para solventar las limitaciones de las aplicaciones informáticas para el análisis de datos. Algunos programas no ofrecen un sistema cerrado de técnicas estadísticas, permitiendo que el usuario programe la técnica estadística que le interesa. Se dispone de un conjunto de librerías (entre las cuales se encuentran varias librerías estadísticas), operadores (aritméticos, lógicos y relaciones) y funciones matemáticas, de forma que el usuario puede configurar según sus necesidades el análisis estadístico. El problema es que consume tiempo de programación y, por descontado, requiere un periodo de aprendizaje mucho más extenso que otras aplicaciones. La ventaja evidente consiste en que estas aplicaciones son muy dúctiles, de tal forma que el usuario puede programar mediante un conjunto de comandos la técnica estadística que le interese y que no está disponible en otras aplicaciones comercializadas. Sería similar a realizar un programa con un lenguaje de programación informática, pero disponiendo de un conjunto de librerías que agilizan notablemente la programación. Por ejemplo, no sería necesario programar el cálculo de una correlación, pues bastaría llamar a la función de la librería que realiza ese cálculo.

Para no forzar a los usuarios a realizar búsquedas de distintas aplicaciones y aprender otra vez largas listas de comandos, ¿no sería conveniente incorporar nuevos procedimientos en las herramientas informáticas? Se nota en éstas un más que

evidente descuido de todas aquellas técnicas que se fundamentan en la computación intensiva. Un buen ejemplo, pues hace varios años que fueron desarrollados los métodos de re-muestreo, son las pruebas de permutación y la metodología *bootstrap*. Se ha mencionado la importancia de los supuestos y sabemos que muchas técnicas estadísticas se utilizan sin mayor problema, aunque se violen los mismos. Entonces, ¿por qué estas aplicaciones no incorporan procedimientos de re-muestreo, que sabemos requieren un mínimo de supuestos? No se pueden aducir costes computacionales porque, aunque consumen más tiempo de procesador, ese no suele ser un problema crítico para la mayoría de investigaciones, pues no tienen un número excesivo de datos. Además, omitir esos métodos provoca la siguiente secuencia de retroalimentación: como no se hallan en las herramientas informáticas, no se conocen; a consecuencia de que se desconocen, no se reclama su inclusión en las aplicaciones. Quizá su ausencia no sea más que una consecuencia de un tipo de diseño de las aplicaciones, que fue fundamentada en algoritmos representados mediante procedimientos o funciones. La solución se hallaría en una programación orientada a objetos, donde la ejecución de un aplicación informática requiriese que el usuario especificara un secuencia de acciones primarias (podrían estar representadas mediante iconos), como 'abrir datos', 'especificar procedimiento', 'fijar condiciones', 'seleccionar estadístico', 'determinar prueba de hipótesis', 'establecer distribución' y 'fijar regla de decisión'. Sería algo más complejo para el usuario, pero también favorecería la comprensión del proceso de razonamiento estadístico, sería un sistema más flexible y no se produciría un renacer de los lenguajes de comandos, tal como se ha sugerido (Chambers, 2000). Ahora bien, y sin tocar la programación estructurada, sería totalmente posible desarrollar nuevos procedimientos en los cuales se implementaran las pruebas de permutación y la metodología *bootstrap*, aprovechándose de las funciones ya existentes. El problema para el desarrollo de ese nuevo código no puede ser precisamente que deban generarse nuevos procedimientos. Existe un ejemplo muy evidente, como son las técnicas relacionadas con la minería de datos. Su importancia en el ámbito aplicado es incuestionable, aunque debe quedar claro que no es conocimiento científico aquello que suele obtenerse en estos estudios. Es evidente el esfuerzo realizado por las compañías que desarrollan aplicaciones por incorporar nuevos procedimientos orientados a la minería de datos, comprar los derechos de los procedimientos a otras compañías o, incluso, llegar a alianzas comerciales. Ahora bien, aquí hay negocio. La minería de datos está aportando resultados muy interesantes a las empresas. Aquí está con casi total seguridad la razón por la cual algunos procedimientos se han incorporado y otros no. Pero, insistimos, parece aconsejable, no sólo que se mejoren los módulos para la minería de datos, sino que se desarrollen procedimientos para realizar pruebas de permutación y estimaciones *bootstrap*. Esto resulta especialmente necesario si atendemos a la utilización masiva de estas aplicaciones y al escaso cumplimiento de las condiciones que requieren las técnicas estadísticas correspondientes.

Otra regla conveniente que debe seguirse es no ejecutar los análisis suponiendo adecuadas para nuestros datos las opciones que por defecto fijan los programas. Las cosas no son tan simples, y menos la Estadística. Antes de obtener una solución debemos conocer exactamente por qué utilizamos un determinado procedimiento de

estimación, fijamos un criterio para detener un proceso iterativo y un largo etcétera. Tampoco la máxima rigurosidad nos garantiza unos resultados incuestionables. La razón reside en que las técnicas estadísticas son modelos sujetos a distintos supuestos y, por tanto, la solución no es independiente de los mismos. Un ejemplo aclarará esta cuestión. La taxonomía numérica se ha utilizado para clasificar especies y determinar su proximidad evolutiva. Ahora bien, no todos los antropólogos aceptan las soluciones halladas mediante las técnicas multivariantes, pues se fundamentan en las características fenotípicas y, además, suponen que morfologías similares indican proximidad. Pero, ¿siempre la semejanza morfológica implica esa proximidad? ¿no han podido obtenerse morfologías parecidas entre especies cuyas líneas evolutivas se separaron hace varios millones de años? Este es sólo un ejemplo donde se muestra que las cosas no son en absoluto simples, pues ni siquiera seleccionando con total rigor los procedimientos estadísticos se llega a que los resultados serán incuestionables, a tenor de los supuestos sustantivos dominantes en una determinada disciplina científica.

No es menos importante el estudio de los supuestos que subyacen a todas las técnicas estadísticas, y no sólo nos referimos a los puramente estadísticos, sino también a los requerimientos que atañen a la escala de medida. Se cumplan o no los supuestos o, como algunos prefieren denominar, las condiciones de aplicación, los programas realizan los cómputos y proporcionan diversos resultados. Es preciso siempre comprobar los supuestos y, si su credibilidad es dudosa, no deben interpretarse los resultados obtenidos mediante la técnica estadística que los exige. La magnitud del error que se comente cuando se violan los supuestos es conocida para muchas técnicas estadísticas y, por descontado, el efecto no es despreciable, llevándonos con frecuencia a conclusiones erróneas. Las aplicaciones no realizan esta tarea por nosotros.

No olvidemos que el más experto de los estadísticos no conoce, a buen seguro, con suficiente profundidad la totalidad de las técnicas estadísticas. ¿Se produce acaso esa quimera en alguna profesión? Nos tememos que no. Pero en la mayoría de las profesiones, por el momento, no existe una tecnología informática que ponga a disposición del usuario, experto o no en Estadística, tan amplio conjunto de técnicas propias de la profesión. Como conocer una aplicación informática estadística implica obtener resultados con todas aquellas técnicas incluidas en el programa, aunque no significa que quien la utilice domine todas esas estrategias analíticas, un buen principio sería no usar técnicas que no se conocen con suficiencia. El programa no discrimina la capacitación del usuario y, por tanto, sólo queda que la autocrítica evite problemas mayores.

En síntesis, utilicemos las aplicaciones informáticas para el análisis de datos, siempre que dispongamos de sólidos conocimientos estadísticos, y beneficiémonos de sus ventajas; pero invirtamos el ahorro de tiempo que nos proporcionan estos programas para mejorar los análisis estadísticos. No parece razonable en la actualidad dejar de reconocer que la utilización de las herramientas informáticas en el análisis estadístico de datos tiene un carácter casi inevitable, irreversible y positivo;

pues los problemas no están en las herramientas, sino que derivan de acciones de quien las utiliza.

Referencias

Abraira, V.; Cadarso, C.; Gómez, G.; Martín, A. y Pita, S. (2001). Mesa redonda: La Estadística en la investigación médica. *Qüestió*, 25(1), 121-156.

Anderson, B. (1990). Methodological errors in medical research: an incomplete catalogue. Blackwell, Oxford.

Biehler, R. (1997). Software for learning and for doing statistics. *International Statistical Review*, 65(2), 167-189.

Altman, D.G. (1994). The scandal of poor medical research [editorial]. *British Medical Journal*, 308, 283-284.

Bartholomew, D.J. (1997). Fifty years of multivariate analysis. *British Jouunal of Mathematical and Statistical Psychology*, 50, 205-214.

Chambers, J.C. (2000). Journal of Computational and Graphical Statistics, 9(39, 404-422.

Cordani, L.K. (2001). Comments about the article research in statistical education: Some priority questions. *Statistical Education Research Newsletter*, 2(1), 7-9.

Dawkins, R. (1998). *Unwearing the rainbow*. Richard Dawkins. [Traducción española: *Destejiendo el arco iris. Ciencia, ilusión y el deseo de asombro*, Tusquets (Metatemas 61), Barcelona, 2000].

Goldstein, R. (1996). Software, Biostatistical. Encyclopedia of Biostatistics. New York: Wiley.

Jackson, V. y Smith, J.D. (2001). Using technology to reinforce data display and analysis. *The Statistics Teacher Network*, 56, 6-7.

Lagarde, J. (1983). *Initiation à l'analyse des données*. París : Bordas.

McCullough, B.D. (1999) Assessing the reliability of statistical software – Part II. *American Statistician*, 53(2), 149-159.

Paulos, J.A. (1990). *Innumeracy: Mathematical illiteracy and its consequences*. New York: Vintage. [Traducción española: *El hombre anumérico*, Tusquets (Metatemas 20), Barcelona 1990].

Paulos, J.A. (1995). *A mathematican reads the newspaper*. New York: Basic Books. [Traducción española: *Un matemático lee el periódico*, Tusquets (Metatemas 44), Barcelona, 1996].

Paulos, J.A. (1998). *Once upon a number. The hidden mathematical logic of stories*. John Allen Paulos. [Traducción española: *Érase una vez un número*, Tusquets (Metatemas 60), Barcelona, 1999].

Rao, C.R. (1989). *Statistics and truth*. New Delhi: CSIR. [Traducción española: *Estadística y verdad. Aprovechando el azar*, Promociones y Publicaciones Universitarias, S.A., Barcelona, 1994].

Scott, T. y Jackman, S. (1999). Computer corner: Examples of the use of technology in teaching statistics. *Teaching Statistics*, 21.

Smith, R. (1996). Statistical Review for Medical Journals, *Journal's Perspective. Encyclopedia of Biostatistics*. New York: Wiley.