



RASC

Real Academia
Sevillana de Ciencias

Observaciones outliers

discurso pronunciado por el

Excmo. Sr. D.

Antonio Pascual Acosta

en el acto de recepción como académico numerario

el día 11 de diciembre de 2003

y

contestación del académico numerario

Excmo. Sr. D.

Rafael Infante Macías

LOS INVESTIGADORES Y LAS OBSERVACIONES OUTLIERS

*Discurso pronunciado por el
Excmo. Sr. D. ANTONIO PASCUAL ACOSTA
en el Acto de su recepción como Académico Numerario
celebrado el día 11 de diciembre de 2003*

*Entregaos sólo a la ciencia.
No esperéis encontrar felicidad en ningún otro sitio.
La fama es un espejismo,
se puede evaporar con la facilidad de un arco iris,
es una ilusión centelleante en los ojos de quienes os rodean,
y si ponéis demasiada confianza en esas cosas,
acabaréis también por creer en ellas.
Confiad sólo en vos y en vuestro trabajo.
La verdad existe en las matemáticas y en ningún sitio más.
Dad caza a las matemáticas, que nunca os desilusionarán.
Carta de D'Alembert a J. C. de la Haye
19 de julio de 1772.*

Excmo. Sr. Presidente del Instituto de Academias de Andalucía,
Excmos. Sres. Rectores y Vicerrectores de las Universidades Andaluzas,
Excmas. e Ilmas. Autoridades,
Excmos. Sres. Directores y Presidentes de Academia,
Excmos. e Ilmos. Sres. Académicos,
Queridos amigos y amigas,
Sras. y Sres.

Mis primeras palabras al iniciar la lectura de mi discurso de ingreso en la Real Academia Sevillana de Ciencias tienen que ser necesariamente de gratitud y de satisfacción.

Gratitud a esta Real Academia por su decisión de incorporarme al conjunto de sus miembros. En cuantas ocasiones he recibido alguna distinción, siempre la he considerado excesiva respecto a mis méritos y no como una cláusula de estilo, tan al uso,

sino como la lógica consecuencia de la valoración que mi labor me merece. Hoy que me veo honrado con esta distinción tan importante para mí, cambiaré el signo de mi habitual inclinación y pensaré que, si la generosidad de tan docto colectivo así lo ha estimado oportuno, no es correcto debatir su pertinencia, ni siquiera conmigo mismo y me limitaré como no podría ser de otra manera a expresar mi sincera gratitud a los miembros de la Real Academia.

Agradecimiento especial al Profesor Infante, a mi maestro, que ha tenido la amabilidad de aceptar el encargo de la Academia de pronunciar el discurso de contestación. Rafael Infante, la persona que me abrió las primeras puertas de la Ciencia, del que tanto he aprendido y al que siempre estaré agradecido por sus atenciones, cariño y magisterio.

Agradecimiento a D. Eduardo Roca Roca, Presidente del Instituto de Academias de Andalucía, que tanto tiene que ver con el momento de auge que viven nuestras Academias en Andalucía.

Agradecimiento también al Secretario General de Universidades de la Junta de Andalucía, D. Francisco Gracia Navarro, al Director General de Universidades, D. Francisco Martos Crespo, a los Rectores de las Universidades de Jaén y Almería y a los Vicerrectores de las Universidades de Sevilla, Granada, Málaga, Córdoba y Cádiz, que han querido acompañarme en este día tan señalado para mí.

Agradecimiento a la Real Academia de Bellas Artes de Santa Isabel de Hungría y a su Presidente, D. Antonio de la Banda y Vargas, por todas sus atenciones hacia mi persona y por permitir que en esta sede tan prestigiosa de la “Casa de los Pinelo” pueda leer mi discurso de ingreso.

Gratitud en suma a todos Uds. por su asistencia y por su amable atención.

Satisfacción no sólo por el alto honor con que se me distingue, sino por la posibilidad que ello me brinda de poder compartir reflexiones con profesores, con investigadores, con maestros de la talla, de la valía, de Francisco González, Rafael Márquez, Manuel Losada, Rafael Infante, Javier Aracil o José Luis Manzanares, entre otros. Satisfacción por la posibilidad de incorporarme a una joven y dinámica Real Academia que gracias al buen hacer de sus sucesivos Presidentes, uno de ellos el profesor Martínez Moreno, desgraciadamente ya no está entre nosotros, ha sabido hacerse un hueco importante, ha sabido ocupar un papel relevante en el conjunto de Reales Academias de Andalucía.

Un tema de investigación que ha centrado mi atención y al que he dedicado tiempo de estudio y de trabajo, es el que hoy en día se conoce como “Análisis Estadístico de Outliers”. Por decirlo de un modo coloquial, con el término anglosajón de outlier se hace referencia a *“aquella observación que por atípica o errónea, se desvía marcadamente del comportamiento general de una masa de datos”*.

Con esto adelanto, aunque sólo sea enunciativamente, el tema sobre el que deseo centrar mi intervención en la parte más marcadamente de producción científica propia y de la que me ocuparé más adelante. Basta por ahora con esta definición para reflejar el factor, simbólico quizás, desencadenante de mi reflexión en torno a la labor del investigador en nuestros días.

Pues bien, la pregunta que recoge esta preocupación se podría formular así: ¿hasta qué punto la persona dedicada a la investigación científica puede considerarse outlier social?, en su doble acepción de atipicidad, por las características de la tarea a la que se dedica y, quizás también, entendiendo como errónea su orientación en la vida, si atendemos, al menos, al común de los valores que imperan en la sociedad de nuestros días, en la que los conceptos de éxito, de triunfo social, de adquisición de cuotas de poder, de preeminencia económica... suelen ser los parámetros por los que se mide y se aprecia la valía de las personas.

Desde luego, si aceptáramos como válido ese modelo de conducta y de actitud que caracteriza al triunfador contemporáneo, difícilmente podríamos escapar de concebir al científico como un auténtico "outlier".

La labor callada, sin brillo exterior, que caracteriza al trabajo investigador; la cantidad de tiempo y de esfuerzo que es preciso invertir para estar en condiciones de abordar, siquiera mínimamente, cualquier planteamiento o cuestión propia del conocimiento científico; la vida ascética, por el rigor y la dedicación exclusiva y obsesiva a un aprendizaje sin fronteras y generalmente arduo y costoso; la obtención feliz de resultados, en la mayoría de los casos modestos, desde el punto de vista del saber, y en absoluto relevantes para los ideales de vida mundana y de reconocimiento social..., todos estos y otros condicionantes que rodean la vida de la ciencia, nos confirmarían en la idea del científico como ser un tanto marginal, cuya nobleza de intenciones no le exime del sambenito de excéntrico, cuando menos.

Sería preciso acudir a la historia de la ciencia, constatar la importancia del saber y del cultivo del conocimiento, el significado y el peso específico de la actividad científica en la vida del ser humano y en la construcción de la realidad social, para contrarrestar, de algún modo, esa visión más bien melancólica, oscura, crepuscular y gris de lo científico, juzgada desde los criterios del éxito social.

Sin duda, es una empresa difícil y excesiva para llevarla a cabo en esta ocasión, pero me van a permitir que les ofrezca unas pinceladas, grosso modo, que, al menos, nos pueden servir para representar más adecuadamente el papel de lo científico en la historia y en la actualidad.

Recurrir a la historia de la ciencia no sólo no es tarea fácil sino que, además no podemos dejar de tener presente la observación en la que insiste un experto de reconocido prestigio. Imre Lakatos (1974) nos advierte "*que la historia de la ciencia es frecuentemente una caricatura de sus reconstrucciones racionales; que las reconstrucciones racionales son frecuentemente caricaturas de la historia real; y que algunas historias de la ciencia son caricaturas de ambas: de la historia y de sus reconstrucciones racionales*".

No obstante esta reserva, si nos adentramos en la génesis del conocimiento científico, cabría distinguir entre antecedentes remotos y lo que podemos considerar ciencia en sentido estricto y moderno del término.

En cuanto a los primeros no podríamos sino remitirnos poco menos que a los orígenes del ser humano. A este respecto dice Friedrichs (1977): "*aunque la mayoría de los estudios contemporáneos rastrear los comienzos de la ciencia en los registros y*

cálculos de los acontecimientos astronómicos y de las relaciones geométricas de los antiguos babilonios y egipcios, podría demostrarse que su origen se remonta por lo menos a los tiempos en los que surgió el lenguaje, durante la evolución del primate. Porque con el surgimiento del primate los progenitores del homo sapiens pudieron por primera vez establecer generalizaciones conceptuales susceptibles de ser compartidas con sus semejantes y transmitidas de generación en generación”.

Sin duda, la mayoría de autores coinciden en afirmar que existe una cierta línea de continuidad en el saber desde los orígenes prehistóricos del ser humano hasta el presente, e incluso, Jean Piaget y su escuela han llegado a establecer ciertos paralelismos entre las etapas de formación del espíritu científico y los procesos de desarrollo mental del mismo; lo que implica una similitud o un recorrido paralelo entre las pautas de crecimiento entre individuo y sociedad, ideas que aparecen desarrolladas a lo largo de la obra de Piaget (1977,1974) y con motivo de diversas investigaciones.

Pero, con todo, la ciencia, a diferencia de otras formas de enfrentarse con el reto del conocimiento, más o menos rigurosas, más o menos sistemáticas, –como pueden ser la magia, la religión o el pensamiento reflexivo–, supone una nueva etapa en la evolución del conocimiento, cuya característica más destacada es el rigor, aún extremo en el procedimiento, el rigor del método.

Estamos hablando de un acontecimiento más bien reciente. Dice Bertrand Russell (1949) en su obra “La perspectiva científica”: *“La ciencia como factor en la vida humana es muy reciente. El arte estaba ya bien desarrollado antes de la última época glacial, como sabemos por las admirables pinturas rupestres. No podemos hablar con igual seguridad de la antigüedad de la religión; pero es muy probable que sea contemporánea del arte. Aproximadamente se puede suponer que ambos existen desde hace ochenta mil años. La ciencia, como fuerza importante, comienza con Galileo y, por consiguiente existe desde hace unos trescientos años...”*.

En contra de lo que mucha gente ajena al quehacer científico supone, la ciencia no es un sistema de verdades monolíticas, férreamente asentadas y acumuladas a lo largo del tiempo. Por el contrario, la ciencia es una especie de consagración perpetua del estado de duda. Y ni tan siquiera ansía, como colofón final del proceso de búsqueda, un estadio final de reposo en la verdad indubitable. A este respecto, dice Popper (1980) que la ciencia no es *“un sistema de enunciados seguros y bien asentado, ni uno que avanzase firmemente hacia un estado final (...). Nuestra ciencia nunca puede pretender que ha alcanzado la verdad, ni siquiera el sustituto de ésta que es la probabilidad... El antiguo ideal científico de la “episteme”, de un conocimiento absolutamente seguro y demostrable, ha mostrado ser un ídolo. La petición de la objetividad científica hace inevitable que todo enunciado científico sea provisional para siempre: sin duda cabe corroborarlo, pero toda corroboración es relativa a otros enunciados que son, a su vez, provisionales...”*.

La afirmación de Popper no puede ser más rotunda ni más clara: *“La ciencia nunca persigue la ilusoria meta de que sus respuestas sean definitivas, ni siquiera probables; antes bien, su avance se encamina hacia una finalidad infinita, y sin embargo, alcanzable: la de descubrir incesantemente problemas nuevos, más profundos y más generales,*

y de sujetar nuestras respuestas, siempre provisionales, a contrastaciones renovadas y cada vez más rigurosas”.

Las consideraciones de Popper podrían hacer pensar a alguno que la ciencia es un producto de alcance limitado, de gran provisionalidad y sometida a pruebas y contrapuebas. Sin embargo, es todo lo contrario; incluso programáticamente, la ciencia se presenta como un proyecto ambicioso. Como dice Bunge (1969): *“Lo que afirma la ciencia es: que es más verdadera que cualquier modelo no científico del mundo; que es capaz de probar, sometiéndola a contrastación empírica, esa pretensión de verdad; que es capaz de descubrir sus propias deficiencias, y que es capaz de corregir sus propias deficiencias, o sea, de reconstruir representaciones parciales de las estructuras del mundo que sean cada vez más adecuadas”*.

Ahora bien, la realidad científica no se agota en sus relaciones con la verdad, ni en su peculiar estatuto epistemológico. Muchos autores han señalado convenientemente el carácter ambiguo y amplio del concepto de ciencia, que, bajo un único término, denomina una gran cantidad de cosas distintas, aunque relacionadas entre sí.

Para Merton (1980), los significados de la ciencia pueden ser los siguientes: *“un conjunto de métodos por medio de los cuales se certifica el conocimiento (la ciencia como proceso); un depósito de conocimientos acumulados procedentes de la aplicación de esos métodos (la ciencia como resultado); un conjunto de valores y costumbres culturales que gobiernan las actividades llamadas científicas (la ciencia como ethos) y cualquier combinación de lo anterior”*.

Además de contenidos y objetivos, la ciencia puede cifrarse en un cierto talante mediante el cual se enfrenta al conocimiento y al afán de transformación de la realidad. Ahora bien, ese talante surge en un determinado contexto histórico, social y cultural y, paulatinamente, se va configurando como modo de conocimiento a base de desarrollar una metodología precisa y rigurosa.

Estas condiciones propicias aparecen justamente en la Europa del Renacimiento, por lo tanto podemos pensar, como hace Cordon (1976), que *“la clave de la ciencia está, pues, en la historia de esta época y región”*, o como dice Needham (1980), *“la pregunta es ¿qué ocurrió en la Europa del Renacimiento cuando las matemáticas y el conocimiento de la naturaleza se unieron en una combinación cualitativamente nueva y destinada a transformar el mundo?”*.

A diferencia de otras sociedades y otras etapas históricas, en esa época se desarrolla un espíritu mercantil y práctico que va desplazando al viejo mundo de interpretaciones simbólico-cualitativas. Ahora, la cantidad, la medida, la precisión, cobran carta de naturaleza y configuran una mentalidad nueva y moderna. Pienso que la concurrencia de un conjunto de factores varios como el Renacimiento, la Reforma y el surgimiento del capitalismo mercantil, desencadena el triunfo de esta mentalidad en Europa.

Como dice Farrington (1974): *“lo que hoy llamamos ciencia no pudo comenzar a surgir hasta que se rompieron las barreras que separaban las distintas esferas del saber; hasta que las sugerencias emanadas de los procesos técnicos pudieron ser aplicados con audacia a todos los fenómenos de la naturaleza; hasta que pudo destruirse el hechizo que envolvía el modo de explicación mitológica, y hasta que los conoci-*

mientos matemáticos dejaron de ser parte del equipo administrativo de los burócratas y pasaron a ser parte de la cultura de los ciudadanos libres”.

Esta cultura ciudadana es un fenómeno histórico que se entrelaza, se acopla, se amalgama, con el desarrollo del pensamiento racional y de las matemáticas “*como el lenguaje en el que está escrito el libro de la Naturaleza*”, como decía Galileo. Esa cultura responde a la mentalidad que configura el orden burgués, y el espíritu burgués, impulsado por la Revolución Francesa, va a ser un elemento capital en el desarrollo del pensamiento científico.

El conjunto de valores procedentes de esa sociedad burguesa, que van a sostener e impulsar la auténtica mentalidad científica, está recogida con nitidez por muchos tratadistas. El profesor Salustiano del Campo (1969) los resume del siguiente modo: “*el racionalismo, con el desarrollo de la lógica y de la matemática, el empirismo con su hincapié en la observación y la experimentación de los elementos anteriores, la creencia en la legalidad, tanto en la naturaleza como en la sociedad, que no es sino una parte de aquélla, el pragmatismo, la utilización del conocimiento para la transformación del mundo más que para la obtención de la sabiduría per se; el ascetismo característico de la vocación científica, el escepticismo frente a la autoridad y a la tradición, y finalmente, el individualismo*”.

Pienso, Sras. y Sres., que la renovación es, precisamente, una señal definitoria del conocimiento científico, por su arraigo en el escepticismo metódico, por el sometimiento continuo de hipótesis y teorías a duros procesos de verificación, por la actitud crítica que acompaña a toda incursión científica. Justamente, a partir de la constatación del cambio científico surge la pregunta ¿cómo y por qué evoluciona la ciencia? Apasionante cuestión que ha ocupado a expertos de talla relevante, como Popper, Kuhn, Lakatos, Shapin o Feyebarend, entre otros, durante los últimos tiempos del siglo XX. Pero, esa es ya otra historia que nos llevaría demasiado lejos.

Retomemos, pues, la cuestión inicial: ¿es el científico, el investigador, una especie de “outlier”, de ser marginal, ocupado en oscuras averiguaciones sin trascendencia en la realidad social?. El esbozo que hemos dibujado de la formación del espíritu científico nos legitima para responder negativamente: la actividad científica se sitúa en la médula misma de la configuración del mundo moderno. Son los ideales de racionalidad, de libertad, de emancipación y de progreso, los que están presentes, tanto en el camino de la ciencia, como en el de la sociedad desde los tiempos de la “nueva ciencia”. Hemos llegado a ser lo que somos en gran medida gracias a la aportación de los científicos de los últimos siglos.

¿Cómo explicar, entonces, la deficiente valoración del trabajo de las personas dedicadas a la investigación? Quizás es un efecto retórico muy característico del universo actual de la comunicación regido por patrones mediáticos: se ensalza a la ciencia, como entidad abstracta pero se escamotea y se oculta a quienes hacen la ciencia. Se reconoce el valor del conocimiento y su papel en el presente y en el futuro de la sociedad y, a la vez, se le regatea a los equipos y grupos de investigación la financiación necesaria o el material e infraestructuras imprescindibles para su labor. Se trata de una cierta actitud esquizoide y en la que hay quien ve un gusto idealista: se menciona con reverencia

un término abstracto “la ciencia”, pero como si fuera una entidad supraceleste, platónica, que obrase de suyo sin contacto con el operar humano, fuera del tiempo y de toda localización. Se afirma “la ciencia avanza, la ciencia dice”.

Quizás es éste el campo de reivindicación y de protesta en el que todos cuantos tomamos parte en las labores de investigación científica deberíamos alzar la voz y, en nombre de la científicidad, que siempre es realista, denunciar esa inflexión idealista que, en los medios de comunicación, en palabras de no pocas personas ilustres, ensalzan una entidad abstracta tanto como niegan incluso la existencia de los protagonistas reales del hecho científico, salvo que estos, unos pocos, hayan pasado al panteón de la historia y sean, por tanto, gloriosos cadáveres.

Hemos de reclamar, pues, la atención debida a la comunidad científica, cuyos miembros son autores, coautores, de la ciencia; la cual no es un ser autónomo, ni una omnipresente deidad dotada con atributos antropomórficos, sino que ciencia son teorías, propuestas, experimentos, hallazgos, estudios, debates, polémicas, etc. y todo eso es obra de seres de carne y hueso.

En fin, hora es ya de que apoyado y, en gran medida, confortado por el aval que proporciona la trayectoria histórica y cultural de la investigación científica como proceso de construcción de la realidad, proceda a comunicar el trabajo de investigación en el campo de la estadística que ha ocupado una buena parte de mi tarea en los últimos años.

Una de las definiciones clásicas de Estadística es la dada por Hampel (1986) que la define como “*aquella ciencia que permite extraer información relevante y útil del conjunto de datos resultantes de una experimentación*”. Definición ésta que conjuga términos característicos de cualquier proceso de investigación científica: experimentación, datos, conclusiones...

Para extraer la información relevante y útil de los datos es imprescindible que los mismos tengan la mayor precisión y fiabilidad posible, por esta razón, uno de los apartados más importantes dentro del análisis estadístico es la depuración de los datos resultantes de una experimentación.

Como ilustración consideremos el conjunto clásico de datos analizados por Pierce (1852) y Chauvenet (1863), relativos a las quince observaciones hechas por L. Herndon en 1846 del semidiámetro vertical de Venus. Los residuos respecto a un modelo simple se reproducen a continuación:

-0.30	+0.48	+0.6	-0.22	+0.18
-0.44	-0.24	-0.13	-0.05	+0.39
+1.01	+0.06	-1.40	+0.20	+0.10

Obsérvese que existen dos valores -1.40 y +1.01 sobre los que cabría hacerse las siguientes preguntas: ¿proceden dichos valores de la misma fuente que el resto de los datos?, ¿están afectados por algún error?, ¿qué tipo de error?

Partiendo de una forma muy simple de representar una experimentación, comenzaremos realizando una descripción pormenorizada de los errores que pueden presentarse en los datos experimentales, lo que nos permitirá dar respuesta a las preguntas formuladas e indicar, de forma sencilla, la importancia y problemática que presenta actualmente el análisis de este tipo de datos.

La realización de un experimento lleva consigo una tripleta de elementos (Ω , X , E), donde Ω es la fuente o naturaleza que provee los elementos ω_i , objeto de estudio y X es el medio por el cual el experimentador E cuantifica o transforma las características de interés de los elementos de la naturaleza en las observaciones o datos x_i , de los que dispondrá el estadístico para analizar el experimento. Esta estructura permite distinguir diferentes fuentes de variación o error que pueden afectar a las observaciones experimentales: Variabilidad de la fuente o naturaleza, error del medio y error del experimentador.

La *variabilidad de la fuente o naturaleza*, es una manifestación consustancial del carácter estocástico de los problemas que aborda la Estadística. Tal variabilidad ha de considerarse como un comportamiento natural de los datos.

El *error del medio* es el que se produce cuando no se dispone de la técnica adecuada o cuando no existe un procedimiento para realizar la transformación de una forma exacta.

Se incluyen en este error los casos de redondeo forzoso que se ha de realizar cuando se trabaja con variables continuas.

Por último, el *error del experimentador*, es el atribuible al propio experimentador y se puede presentar, a su vez, de tres formas diferentes:

- El que aparece al imponer un modelo matemático no preciso o adecuado a la población, o bien al considerar información o hipótesis iniciales incorrectas, y que denominamos *error de información*.
- El que se presenta cuando el investigador no llega a delimitar de una forma correcta el espacio Ω y el experimento lo realiza sobre un espacio distinto. La selección de muestras sesgadas o la inclusión de individuos de una población diferente a la que debe ser objeto de la experimentación, se contemplan en este error denominado *error de planificación*.
- En tercer lugar, el que se comete al llevar a cabo una valoración errónea de los elementos de la población es el llamado *error de realización*. Aquí se incluyen, entre otras, las transcripciones erróneas de datos o las falsas lecturas realizadas sobre el instrumento utilizado para obtener los resultados experimentales.

Según los errores o variabilidades consideradas anteriormente, Anscombe (1960) distingue dos tipos de observaciones:

- *Observación atípica*, que es aquel valor experimental que presenta una gran variabilidad de tipo inherente de la fuente o naturaleza.

- *Observación errónea*, que es aquella observación que presenta un gran error del medio y/o un gran error del experimentador, que a su vez puede manifestarse a través de uno o varios de los errores definidos anteriormente.

La posible presencia de observaciones atípicas y/o erróneas en una experimentación, hace que algunos autores, entre ellos Chatfield (1985), destaquen, entre las principales etapas que se han de seguir en el análisis estadístico de datos, la de investigar la calidad de las observaciones, así como la de realizar un examen inicial de las mismas, ya que si no se detectan los errores presentes en las observaciones y se actúa suponiendo la ausencia de ellos, los procedimientos estadísticos que se apliquen pueden conducir a conclusiones inexactas.

Al análisis que se realice sobre la calidad de los datos habría que exigirle que distinguiese entre observaciones atípicas y observaciones erróneas, ya que las primeras serían verdaderas observaciones de la población sobre la que se está realizando la experimentación. Sin embargo, desde un punto de vista práctico, tal distinción resulta casi imposible, por lo que se adoptará el término anglosajón de *outlier* para referirse a este tipo de observaciones.

Son muchas las definiciones que sobre el término de *observación outlier* aparecen en la literatura estadística en los últimos años. Generalmente en ellas se caracteriza el outlier, como hace Miller (1981), mediante su desviación del resto de las observaciones muestrales, considerándose además algunas de las siguientes peculiaridades:

- Intervienen los diversos tipos de error que pueden afectar a las observaciones. En este sentido se pueden citar las definiciones dadas por Grubbs (1950), Kendall y Buckland (1957), Anscombe (1960), Hawkins (1980), Beckman y Cook (1983).
- Admiten la subjetividad del experimentador para identificarlas. Entre estas se encuentran las de Grubbs (1950), Kendall y Buckland (1957), Ferguson (1961), Grubbs (1969), Hawkins (1980), Beckman y Cook (1983), Barnett y Lewis (1984).

La idea de caracterizar al outlier mediante su desviación del resto de los elementos que componen la muestra parece evidente, ya que, como indican Beckman y Cook (1983): “...la fiabilidad probable de una observación se refleja mediante su relación con otras observaciones que se obtienen bajo condiciones similares...”.

Con respecto a la primera de las peculiaridades, parece necesario que alguna referencia a los errores aparezcan en la definición, de forma implícita o explícita. Sin embargo, una definición formal de este concepto no debería contener referencia alguna a la influencia subjetiva del experimentador, ya que la presencia de términos como “despertar sospechas”, “parecer inconsistente”..., se ha de considerar como una reminiscencia del tratamiento inicial del problema a finales del siglo XVIII. Basta citar, a modo de ejemplo, a Boscovich (1755), quién para determinar la elipticidad de la tierra rechazaba las dos observaciones extremas resultantes de la experimentación, sin criterio alguno, únicamente por ser la menor y la mayor de las observaciones.

El problema de la subjetividad se pone de manifiesto en trabajos, como el Collet y Lewis (1976), donde se afirma: “...*existen comportamientos diferentes entre los individuos en su reacción a valores sorprendentes e incluso entre los juicios realizados por un mismo individuo en distintas ocasiones...*”, o el de Beckman y Cook (1983) donde muestran su preocupación hacia definiciones que dependan de una noción tan subjetiva como “parecer sorprendente”.

Es más, en las “situaciones más estructuradas”, según la clasificación realizada por Barnett y Lewis (1984), o cuando se observan datos multivariantes, resulta en general imposible la influencia subjetiva del experimentador sobre la apreciación de observaciones outliers. Esto hace que la definición que adoptaremos evite cualquier grado de subjetividad. De hecho, existen autores que ponen en entredicho la influencia del experimentador.

Asimismo, la definición también debería reflejar el hecho de que una observación será outlier para un determinado modelo o método, no siéndolo necesariamente para otro. Baste citar para ello lo apuntado por Gnanadesikan y Kettinger (1972): “...*la complejidad del caso multivariante sugiere que sería conveniente investigar procedimientos que ofrezcan protección contra cualquier tipo de outlier, aunque una aproximación más razonable sería la de construir procedimientos que permitan proteger los datos ante situaciones específicas...*”.

Teniendo en cuenta las consideraciones realizadas, Muñoz, Moreno y Pascual (1990) en la Revista International Statistical Review proponen la siguiente definición para el término outlier: “*Es aquella observación que siendo atípica y/o errónea, se desvía marcadamente del comportamiento general de la masa de datos, respecto del criterio (modelo, método o situación) que se desea analizar en los mismos*”.

De lo expuesto hasta aquí, podría deducirse que el análisis y caracterización de los outliers tiene como único objetivo validar la calidad de los datos resultantes de una experimentación. Sin embargo, esto no siempre es así, ya que existen bastantes situaciones en las que el estudio e identificación de outliers es precisamente el objetivo más importante de la experimentación. Consideremos para ello los dos ejemplos siguientes.

El primero es recogido por Barnett (1978) y en él se describe el caso legal de Hadlum contra Hadlum en 1949, en el que Mr. Hadlum alega en su petición de divorcio adulterio de Mrs. Hadlum. La evidencia que presentaba para ello era que Mrs. Hadlum alumbró un niño el 12 de Agosto de 1945, 349 días después que Mr. Hadlum abandonara Inglaterra para cumplir el servicio militar en el extranjero. En realidad la evidencia que aportaba era que el período de 349 días era muy diferente al período de gestación promedio de 280 días. En resumen, el valor que se había observado era un outlier. Para esta situación el tribunal juzgó que 349 días era un período de gestación improbable pero científicamente posible según testimonios médicos, y por ello no se concedió el divorcio.

De este caso legal hay que destacar que el interés fundamental del mismo se centraba en la observación outlier y que la sentencia del tribunal se reduce a afirmar que el outlier era producido por la variabilidad inherente.

El segundo es recogido en la prestigiosa revista de medicina *Lance* (1984) y corresponde a un trabajo realizado por profesores de los departamentos de Medicina Interna y de Estadística de la Universidad de Sevilla. En 1983, el Departamento de Medicina Interna del Hospital General Virgen del Rocío de Sevilla, diagnosticó correctamente desde un punto de vista clínico los dos primeros casos de S.I.D.A., sin embargo restaba por comprobar que dichos enfermos tenían un cociente $T4/T8$, donde $T4$ representa el número de linfocitos cooperadores y $T8$ el número de linfocitos supresores, significativamente inferior al de las personas sanas.

Para realizar esta comprobación se utilizaron técnicas de identificación de outliers, para lo cual se tomó una muestra control constituida por personal sanitario a la que se añadió los dos enfermos, midiéndoles a todos ellos el cociente $T4/T8$. El análisis de los datos concluyó que existían dos observaciones outliers correspondientes a los dos enfermos de S.I.D.A. y que los valores de sus cocientes estaban deslizados significativamente a la izquierda del resto de los elementos de la muestra. En este caso, los outliers presentaban el denominado error de planificación.

Con estos ejemplos se pone de manifiesto que el análisis de outliers puede utilizarse para detectar la presencia de distintos fenómenos o peculiaridades en una muestra, y que en muchas experimentaciones suele ser el objetivo fundamental.

Analizada pues la importancia que para el estadístico tiene las observaciones outliers en un conjunto de datos procedentes de una experimentación, parece lógico detenerse en analizar las diferentes formas de abordar la posible presencia de este tipo de observaciones.

Los primeros procedimientos para tratar el problema pueden datarse hacia la segunda mitad del siglo XVIII, época en la que se proponían métodos o criterios de tipo subjetivo. Huber (1.972) recoge el siguiente texto correspondiente a un trabajo anónimo de 1821: “...*Por ejemplo, existen ciertas provincias de Francia, donde para determinar la producción media de las fincas, es costumbre observar su producción durante veinte años, quitar la mayor y la menor producción y calcular la media de los dieciocho elementos restantes*”.

Surge aquí el problema de la subjetividad, al que antes nos referíamos y que se arrastra hasta nuestros días, aunque desde luego de una forma menos determinante.

En las últimas décadas del pasado siglo XX, aparecen los primeros métodos o procedimientos científicos para abordar la presencia de outliers en un conjunto de datos. Beckman y Cook (1983) distinguen entre *métodos de acomodación* y *métodos de identificación*, siendo el objetivo o la información que se desee obtener de la muestra resultante de la experimentación determinante para que se utilicen un tipo de técnicas u otro.

Dentro de los *métodos de acomodación* se engloban dos formas distintas de abordar el problema:

- Dotar a la fuente o naturaleza de un modelo matemático que explique todos los valores muestrales. Conviene señalar que en caso de que las observaciones

outliers fuesen manifestaciones de la variabilidad inherente, este tipo de procedimiento puede dar lugar a un modelo matemático erróneo.

- Utilizar procedimientos estadísticos que no sean influenciados por la presencia de outliers (Métodos robustos). Desde esta perspectiva se presupone que todos los outliers son observaciones erróneas, sin admitir la posibilidad de que existan observaciones con cierto grado de variabilidad inherente, las cuales desde un punto de vista teórico deben ser consideradas también.

Así pues, puede afirmarse que la mayor dificultad que plantean estos métodos está contenida en la conjetura de Anscombe (1960), presentando además el inconveniente de que requieren de una gran información para poder cumplir con sus objetivos, información que además generalmente no está disponible.

Dentro de los *métodos de identificación*, se engloban todos aquellos procedimientos que se caracterizan por las dos etapas siguientes:

- ETAPA I. Determinación de la observación u observaciones outliers.
- ETAPA II. Análisis pormenorizado de cada outlier, y en conjunto de todos ellos, con el objetivo de determinar las variabilidades o errores presentes. Ello permitiría llegar a una o varias de las siguientes conclusiones: Mantener o rechazar las observaciones detectadas como elementos del conjunto inicial de datos; confirmar el modelo matemático caso que este hubiese sido supuesto inicialmente o determinar las desviaciones que se presenten sobre las hipótesis realizadas sobre la población; identificar rasgos de interés práctico que en un principio son desconocidos por el investigador y/o determinar comportamientos singulares en la experimentación.

Estas dos últimas situaciones originan en muchos casos que los outliers tengan más importancia que la muestra en sí.

No obstante, se ha de reconocer, que a veces se tienen experimentos en los que no se llega a conclusiones como las expuestas, bien por la conjetura de Anscombe (1960), o bien porque no es posible el análisis a posteriori de las observaciones identificadas como outliers, lo cual puede ocurrir por muy diversas razones, como es el caso de aquellos experimentos donde los elementos del espacio empírico no vuelven a su estado inicial tras la experimentación. En estos casos, se tiende a rechazar las observaciones determinadas en la Etapa I, basándose para ello en las afirmaciones realizadas al respecto por diversos autores, entre ellos Hampel (1986): “...*el peligro causado por el outlier es mucho mayor que el peligro de pérdida de eficiencia que motiva su rechazo en caso de que no lo sea ...*”.

De las dos etapas que acabamos de distinguir en los métodos de identificación, es la primera de ellas la que es propia del estadístico, ya que la segunda es más apropiada para el experimentador.

Los procedimientos existentes para identificar las observaciones outliers se pueden clasificar en dos grupos, Métodos no Formales o Gráficos y Métodos Formales.

Los *métodos no formales* tienen como objetivo la representación de los datos en un sistema de ejes cartesianos bi ó tridimensional, utilizando procedimientos muy diversos. El más usual consiste en transformar cada observación mediante una función con valores en la recta real.

La función o estadístico se elige de acuerdo con el rasgo o propiedad que se desee analizar en los datos experimentales, considerándose outliers aquellas observaciones cuya representación gráfica se desvía de manera relevante de la masa de datos.

Estas técnicas presentan el inconveniente de que no existe una estructura teórica común a todas ellas, ya que lo único que se exige a las funciones que se utilizan en estos procedimientos es que tengan un fundamento relacionado con el fenómeno que se desea analizar.

El problema fundamental, en casi todas estas técnicas, es que no cuantifican la magnitud ni indican la dirección de la desviación del outlier respecto de la masa de datos.

Los *métodos formales* consisten básicamente en la realización de un contraste de hipótesis, donde la hipótesis nula o hipótesis de trabajo establece que la muestra obtenida de la experimentación no contiene outlier alguno. Para aceptar o rechazar esta hipótesis se construye una función test, que permite, decidir sobre la presencia o no de outliers en los datos.

No obstante, para una misma hipótesis de trabajo, suelen presentarse situaciones en las que es necesario elegir entre diversas funciones tests, por lo que, al igual que en la teoría clásica de contrastes de hipótesis, es necesario introducir algunas medidas que permitan discriminar entre dichas funciones. Así, se define la potencia del test, de forma similar a como se hace en la teoría clásica de Neymann-Pearson, y por tanto a cada hipótesis de trabajo se le ha de asociar una hipótesis alternativa que fija el modelo que provoca la presencia de outliers en la muestra.

Barnett y Lewis (1984), estudian diferentes tipos de hipótesis alternativas en caso de que la hipótesis nula fije la pertenencia de las observaciones a un determinado modelo de distribución.

En estos contrastes de hipótesis, además de la potencia, aparecen otras medidas para la comparación de las funciones tests. David y Paulson (1965) definen diversos criterios para el caso de una observación outlier, que han sido analizados por diversos autores, entre ellos Hawkins (1978,1980), quién indica la importancia y utilidad de algunos de ellos.

Al igual que ocurre en la teoría clásica de los tests, no siempre es posible encontrar funciones con propiedades óptimas globales sobre los parámetros definidos, por ello se suelen buscar tests con condiciones de optimalidad local, insesgadez o que satisfagan ciertas condiciones de invarianza, o bien se utilizan procedimientos que tengan buenas propiedades, como ocurre con el principio de razón de verosimilitud.

Los *Métodos Formales* así expuestos, presentan el inconveniente de que el rechazo de la hipótesis nula no determina qué observación u observaciones son outliers, ya que la hipótesis alternativa únicamente formula el modelo que seguirían estas observaciones, caso de existir. Esta circunstancia hace que se introduzcan las denominadas *Hipótesis o Modelos Etiquetados* y las *Hipótesis o Modelos no Etiquetados*.

En los Modelos Etiquetados la hipótesis alternativa especifica las observaciones que se considerarán outliers en caso de rechazar la hipótesis nula.

Los Modelos No Etiquetados no explicitan las observaciones que se declararán como outliers en caso de rechazar la hipótesis de trabajo. En éstos se suele indicar el número de outliers a detectar, o una cota superior del mismo. Para la determinación de esta cota se utilizan, entre otros, los denominados procedimientos de saltos.

En los Métodos Formales se representan muy diversos problemas, que analizaremos a continuación atendiendo a los elementos característicos de los mismos.

- a) *Respecto de la hipótesis alternativa.* Como ya se ha indicado, esta hipótesis fija el modelo que explica los outliers en caso de que estos existan, lo que resulta muy difícil por las razones que se exponen a continuación:
 - i) Porque no todos los outliers han de proceder de un mismo modelo, ya que esto sería equivalente a afirmar que todos ellos presentan el mismo tipo de error.
En el caso de modelos de poblaciones multivariantes, o de situaciones más estructuradas, la problemática que se plantea es aún mayor, ya que en el caso de los primeros se han estudiado pocos modelos alternativos y en situaciones más estructuradas a la diversidad de modelos existentes, se les une el conjunto de hipótesis iniciales que lleva consigo cada uno de ellos.
 - ii) Aun en el caso en que todos los outliers pudiesen explicarse por un único modelo, cabría preguntarse por la unicidad de éste.
 - iii) La incertidumbre existente sobre la elección de la hipótesis alternativa reduce la utilidad de cualquier propiedad de optimalidad para los tests, ya que éstas dependen de la forma que se adopte para dicha hipótesis.
Estos problemas hacen que la mayor parte de las técnicas formales estén referidas únicamente a la hipótesis nula.
- b) *Respecto a la hipótesis alternativa etiquetada.* Los modelos etiquetados obligan a determinar, a priori, las posibles observaciones outliers. Esto puede considerarse como una reminiscencia del carácter subjetivo con que se trataba esta teoría en sus inicios.
- c) *Respecto a la hipótesis nula.* El problema fundamental es, al igual que en el caso de la hipótesis alternativa, fijarla ya que generalmente no se tiene un conocimiento suficientemente preciso sobre el modelo que siguen los datos.
En este sentido se puede citar la afirmación de Barnett (1984): “...un outlier para una distribución normal puede no serlo para una distribución Cauchy...”.
Por tanto es necesario tener gran certeza sobre el modelo que se establece en esta hipótesis ya que se podría producir el denominado error de información.
- d) *Respecto al estadístico o test.* La problemática que se plantea en este punto está muy relacionada, e incluso podría incluirse en el apartado anterior, ya que la distribución del estadístico en el que se basa la función test, se calcula bajo el modelo que se propone en la hipótesis nula.

El mayor inconveniente radica en el hecho de que generalmente no es posible calcular la distribución exacta del estadístico, por lo que debe aproximarse, utilizándose para ello diversos métodos, siendo los más comunes: los Métodos Conservativos y los Métodos Asintóticos.

- *Método Conservativo*: la aproximación de la distribución del estadístico da lugar a una región crítica conservativa, lo que presenta el inconveniente de que observaciones outliers pueden no ser identificadas como tales.
- *Método Asintótico*: se considera la distribución asintótica del estadístico. En esta situación los outliers pierden parte de su importancia ya que estas observaciones son más relevantes en el caso de muestras de tamaño reducido, que es donde ejercen su mayor efecto perturbador.

Como decíamos antes, también pueden presentarse situaciones en las que se desee analizar la presencia de más de un outlier en la masa de datos, a lo que algunos autores denominan problema de múltiples outliers.

Este problema es abordado de formas muy diversas desde el punto de vista de las técnicas formales y presenta inconvenientes adicionales a los ya destacados anteriormente para las mismas.

La primera forma de abordar el problema de múltiples outliers consiste en la aplicación sucesiva de procedimientos de identificación de un outlier, eliminando de la masa de datos la observación declarada como tal y aplicando la misma técnica de forma sucesiva a la muestra restringida hasta aceptar la hipótesis de trabajo.

Este procedimiento plantea como inconveniente lo que se denomina *Efecto de Enmascaramiento de tipo A*: considerar no outlier a ciertas observaciones que lo son.

El efecto de enmascaramiento de tipo A fue citado por primera vez por Pearson y Chandrasekar (1936), quienes afirman sobre su criterio: “...este criterio parece eficiente únicamente en caso de que exista una única observación outliers...”.

Este efecto causa la disminución de la potencia del método considerado, y presenta, además, un inconveniente de tipo metodológico, ya que el nivel de significación referido a la hipótesis de trabajo (no existen outliers en los datos muestrales), no será el mismo que se le asigne al conjunto de observaciones declaradas como outliers.

Estos inconvenientes, hacen que algunos autores propongan otros tipos de procedimientos para abordar el problema de la presencia de múltiples outliers.

Así, se sugiere un procedimiento iterativo, que en una etapa preeliminar selecciona las posibles observaciones outliers, de donde se extrae la observación que menos sospecha desperta. A esta observación se le aplica una técnica de identificación de outlier en la muestra restringida. En caso de aceptarse la hipótesis nula, se incluiría esta observación en la muestra restringida y se procedería a la selección de una nueva observación de entre los restantes posibles outliers. En el momento en que se rechace la hipótesis nula, se concluye que la observación que se estaba considerando, junto a las que aún no lo habían sido, son outliers.

Este procedimiento evita el problema del enmascaramiento de tipo A, pero presenta el inconveniente de que es necesario, en cada etapa, seleccionar la observación a estudiar, lo que quizás sea factible en el caso de datos univariantes pero no así en el caso multivariante o en las denominadas situaciones más estructuradas.

Otros tipos de procedimientos que evitan el efecto de enmascaramiento de tipo A, son las denominadas técnicas de bloques, las que se aplican para la determinación de un número fijo de outliers.

Desde un punto de vista metodológico, estos últimos procedimientos, necesitan fijar inicialmente el número de observaciones a contrastar como outliers, lo que puede realizarse mediante los denominados procedimientos de saltos.

Fieller (1976), indica que estos procedimientos para la identificación de múltiples outliers pueden presentar el denominado *efecto de enmascaramiento de tipo B*: considerar outliers observaciones que no lo son.

Además, tanto la determinación del número de outliers a contrastar, como el efecto de enmascaramiento de tipo B, afectan a las probabilidades de error de tipo I y tipo II.

Un desarrollo completo de todos estos aspectos puede encontrarse en las distintas ediciones de 1978, 1984 y 1994 del libro de Barnett y Lewis "Outliers in Statistical Data" en las que se recogen muchas de las técnicas que se engloban dentro de los métodos anteriormente citados y se clasifican las técnicas de identificación de outliers recogidas en la literatura de acuerdo al campo específico para el que han sido propuestas: métodos para datos univariantes, métodos para datos multivariantes, métodos para datos estructurados (regresión, modelo lineal, diseño de experimentos,...), métodos bayesianos, métodos en series cronológicas, métodos para datos direccionales y métodos en áreas poco exploradas (tablas de contingencia, muestreo,...). Esta clasificación ilustra claramente la importancia del análisis de outliers en cualquier contexto donde se precise de un análisis estadístico y por lo tanto en cualquier problema real cuyo análisis y diagnóstico de soluciones se fundamente en la Estadística.

Quisiera destacar la importancia que tuvo para nuestro grupo la primera edición de este libro, ya que en esos momentos eran escasas las revistas científicas en el área de la Estadística en la Universidad de Sevilla, teniendo que recurrir continuamente al Consejo Superior de Investigaciones Científicas (CSIC) en Madrid, para poder consultar revistas de nuestra especialidad.

Nuestros primeros esfuerzos en el estudio de las observaciones outlier se encaminaron hacia dos objetivos: tratar de mejorar algunas de las técnicas recogidas en la literatura y proponer nuevos criterios de identificación en modelos no estudiados hasta la fecha. Dentro de la segunda de las líneas antes señaladas, en el año 1982 (Muñoz García, J. y Pascual Acosta, A.) proponemos un criterio de identificación de outliers univariantes en modelos tipo Pareto. La larga cola por la derecha, característica de estos modelos, motiva el interés por identificar outliers en esta dirección ya que pueden incidir notablemente en las medidas de concentración, pudiéndose obtener una errónea representación de la realidad si no se corrige su efecto. Nuestro interés por estas familias de distribuciones tiene su origen en el debate sobre el modelo de Impuesto de

la Renta de las Personas Físicas que se produce en nuestro país a finales de los 70 y principios de los 80, debate que se reabre en los últimos años.

Paralelamente la identificación de outliers en datos multivariantes centra también nuestra atención, interés que nace del convencimiento de que la naturaleza es multivariante y de las posibilidades de cálculo que se vislumbran con el desarrollo de los nuevos sistemas informáticos. Sin duda, el estudio de técnicas en el campo multivariante es más complejo y menos intuitivo que en el caso univariante. Baste citar que además de su incidencia perturbadora en los parámetros de localización y escala, surgen problemas específicos, asociados a la dependencia entre las componentes y a la diversidad de objetivos de los métodos estadísticos multivariantes.

Fruto de este trabajo en el campo multivariante es la primera tesis doctoral que dirigió en el año 1980, titulada "Algunas Técnicas sobre Detección de Outliers". En ella se proponen nuevos procedimientos de identificación de outliers en poblaciones normales multivariantes y en el marco del análisis discriminante bajo la hipótesis de normalidad de los datos con matrices de covarianzas distintas, problema éste último que hasta la fecha había recibido poca atención en la literatura.

Permítame, en estos momentos, hacer una breve reflexión para recordar las circunstancias en las que se desarrollaba la investigación en esos años. La aplicación práctica de los criterios propuestos en la tesis antes referida requerían de la tabulación de distintas distribuciones, lo que originaba un problema computacional para cuya resolución era necesario trabajar con el ordenador, un ordenador de los de entonces, en doble precisión. Trabajar en doble precisión significaba poder correr el programa informático una vez al mes y con un permiso especial del Ministerio de Educación y Ciencia. Necesitábamos meses para tabular una simple función de distribución. En dos décadas el avance en el mundo de la informática ha sido espectacular y las circunstancias actuales, afortunadamente, no se asemejan en nada a las del pasado reciente. Esta reflexión debe de servirnos para valorar el trabajo realizado por los investigadores que nos precedieron.

Nuestro trabajo en el campo multivariante originó publicaciones como Muñoz García, J. (1984a), Muñoz García, J., Pascual Acosta, A. y Pérez Díez de la Ríos, J. L. (1984b), entre otras, en las que se aborda el estudio de los outliers en el modelo de correlación intraclases, modelo de gran importancia en los estudios de órganos biológicos, ya que a partir de él se generan los estadísticos tamaño total y forma del órgano y, además, culminó con la lectura de una nueva tesis bajo el título "Identificación de outliers en muestras multivariantes".

Como decíamos anteriormente, la ciencia no avanza si se aceptan como definitivos los resultados obtenidos o los planteamientos considerados, avanza al cuestionarse lo que otros hacen, incluso lo que uno mismo hace. Nos cuestionamos si la perspectiva desde las que se estaba abordando el problema de las observaciones outliers era la más adecuada. Las razones que nos llevaron a este cambio de visión del problema, algunas de ellas expuestas en un artículo divulgativo publicado en el año 1986 en la revista *Thales*, fueron las siguientes:

1. La ausencia de una metodología formal para abordar la construcción de las técnicas de identificación de outliers.
2. El establecimiento de las hipótesis iniciales sobre el modelo de distribución para la población de la que se extrae la muestra.
3. La dificultad de plantear la hipótesis alternativa.
4. El convencimiento de que las técnicas de identificación de observaciones outliers no deberían construirse sobre distribuciones de tipo asintótico. Las perturbaciones ocasionadas por las observaciones outliers son más acentuadas en muestras de reducido tamaño. Y por último,
5. La percepción de que la caracterización de una observación como outlier no puede hacerse al margen del resto de los datos muestrales y del objetivo que persigue el análisis estadístico.

Algunas de estas cuestiones se abordaron en la tesis leída en el año 1987 titulada "Análisis cualitativo de datos estadísticos" y en algunas publicaciones como Muñoz García, J., Moreno Rebollo, J. L. y Pascual Acosta, A. (1990). En este trabajo se propone un método formal a partir del cual se derivan, como casos particulares, muchos de los criterios de identificación, tanto para datos univariantes como multivariantes, recogidos en la literatura hasta ese momento. Entre éstos podemos citar los criterios propuestos por Grubbs (1950), Tietjen y Moore (1972), Dixon (1950, 1951), Lewis y Fieller (1978), Fisher (1929), Cochran (1941) y Wilks (1963). También es de señalar que el procedimiento propuesto no necesitaba de las condiciones iniciales, restrictivas, que cada autor consideraba al desarrollar la correspondiente técnica.

Otra de las conclusiones de estos trabajos es que los criterios de identificación de outliers debían de ser específicos para cada modelo estadístico, tal y como se recogía en la definición a la que antes hicimos referencia. En esta línea podemos encuadrar el estudio de outliers en el marco del análisis de homogeneidad para poblaciones estructuradas como modelos multinomiales. Problema éste motivado por algunas cuestiones que se nos plantearon en el campo de la medicina y la agricultura, y que utilizamos para ilustrar las técnicas propuestas. Nuestros trabajos se vieron recompensados con la lectura de una nueva tesis doctoral "Técnicas de identificación de outliers para el Modelo Multinomial" en 1988, y la publicación de nuevos trabajos, Muñoz García, J., Moreno Rebollo, J. L. y Pascual Acosta, A. (1987) y García de las Heras, J., Muñoz García, J. y Pascual Acosta, A. (1993).

La ampliación del grupos de trabajo debido al crecimiento del Departamento de Estadística e Investigación Operativa nos permitió ampliar nuestra línea inicial de trabajo y dedicar parte del esfuerzo hacia un nuevo campo como es el Análisis de Observaciones Influencia, en el que se trata de identificar aquellas observaciones muestrales cuya supresión produce un cambio sustancial en las conclusiones del análisis, problema con ciertos paralelismos con el de los outliers, pero con objetivos distintos.

Nuestros primeros trabajos en este nuevo campo se desarrollan a partir del concepto de sesgo condicionado y su aplicación en el Modelo Lineal General, resultados que se recogen en la tesis doctoral "Análisis de Influencia en el Modelo Lineal General. Sesgo

Condicionado" (1992) y en diversas publicaciones como Muñoz Pichardo, J. y Muñoz García, J. (1995) o Moreno Rebollo, J. L., Muñoz Reyes, A. y Muñoz Pichardo, J. (1999), entre otros. En estos trabajos se pone de manifiesto la utilidad del sesgo condicionado como herramienta en el análisis de influencia y como a partir de este concepto se pueden obtener de forma unificada distintos diagnósticos de influencia.

Casi todos los diagnósticos de influencia propuestos en la literatura, están orientados hacia el estudio de la influencia sobre las estimaciones de los parámetros de interés del modelo, sin tener en cuenta que en todas las experiencias y estudios científicos se le pide a la Estadística que estas estimaciones vayan acompañadas de una región de confianza o de su error de estimación. Por esta razón orientamos nuestro trabajo a diseñar técnicas que permitieran analizar la influencia sobre la distribución muestral de dichos estimadores. Esto es, no sólo sobre el valor del estimador obtenido en una muestra concreta, sino en toda la distribución de valores que puede tomar el estimador bajo estudio. Ello nos proporcionaría una medida de influencia más global, más completa y, por tanto, menos sensible a posibles fluctuaciones muestrales.

Así, en un trabajo publicado recientemente (Jiménez-Gamero y otros, 2002), proponemos una modelización genérica del problema y, en particular, se utiliza la distancia de Rao para cuantificar la discrepancia en las distribuciones muestrales. A raíz de esta modelización se ha abierto una línea de trabajo concreta que puede aportar una nueva e interesante visión del problema de la influencia, con la aplicación de otras distancias distribucionales.

Aunque se podría seguir mencionando frutos de la investigación desarrollada por el grupo al que pertenezco, me gustaría dar respuesta ahora a una cuestión que la sociedad a veces nos plantea: ¿qué les ha llevado a ustedes estudiar lo que estudian?, ¿qué extraña causa es la que les ha movido a investigar en lo que investigan? Aunque pueda parecer una paradoja, el desarrollo teórico de estas técnicas nació de nuestro contacto con problemas reales de otras áreas científicas. Nuestra colaboración con médicos, economistas, psicólogos, sociólogos, geólogos, químicos, biólogos y con el sector empresarial es esa causa extraña que nos ha movido a investigar lo que investigamos, pero también nos ha movido la duda científica con el objeto de garantizar la máxima fiabilidad de las conclusiones extraídas de los datos resultantes de una experimentación. El hacer de la duda un objeto de estudio ha sido, es y será uno de los grandes motores de la ciencia y la colaboración interdisciplinar ha sido, es y será uno de los grandes vehículos que hacen avanzar la ciencia.

Ese motor denominado duda y ese vehículo llamado colaboración nos ha abierto una nueva línea de investigación entrelazada con las técnicas estadísticas orientadas al análisis de la calidad de los datos y con los estudios medioambientales. La creciente preocupación social por nuestro medio ambiente ha motivado que desde muy diversas áreas de la ciencia se estudien problemas medioambientales, problemas que en muchas ocasiones requieren del análisis de la información proporcionada por una amplísima gama de magnitudes químicas y físicas. En consecuencia, la investigación se fundamenta en grandes y, en ocasiones, dispersos conjuntos de datos, que han de ser analizados y desmenuzados desde la perspectiva de la Estadística. En un primer momento

se ha tendido hacia la aplicación directa de las técnicas estadísticas ya existentes; sin embargo, se hace necesario el desarrollo de otras específicas que den respuestas a cuestiones nuevas que le son comunes y exclusivas a los estudios medioambientales. Al igual que en los estudios econométricos el efecto “tiempo” impide la aplicación directa de algunas técnicas estadísticas y origina que surjan nuevas técnicas que contemplen la presencia de este efecto, en los estudios medioambientales el efecto “espacio” desempeña un papel similar.

En esta consideración encontramos el origen de nuestro interés por el Análisis de Observaciones Outliers Espaciales. El efecto “espacio” en el análisis de este tipo de observaciones resulta ser bastante más complejo dado que puede darse la circunstancia de que una observación sea considerada como outlier en un entorno de ella y sin embargo no tiene por qué ser considerada como tal de una forma global, o viceversa, como afirman Watenberg (1990) y Sekhar y Zhang (2003). Ello hace necesario la tipificación de observación outlier, además de globalmente, de forma local.

Más aún, la realidad medioambiental indica que las magnitudes no sólo cambian en función de las distancias en el espacio, sino que pueden verse afectadas por las direcciones en que se recorren. En consecuencia, además del carácter local debe ser tipificado adecuadamente el carácter direccional en el concepto de observación outlier espacial, y por tanto, se han de desarrollar nuevas técnicas de identificación de estas observaciones que contemplen la anisotropía de estos problemas.

En esta línea de investigación ya se han obtenido algunos resultados (Pino Mejías y otros, 2003), sin embargo queda un largo camino por recorrer para dar adecuadas respuestas a la necesidades planteadas.

Como dice Barry Mazur (Profesor de la Universidad de Harvard), las matemáticas son una larga conversación entre aquellos que resuelven los problemas y aquellos que construyen las teorías, entre las zorras y los erizos que pueblan todas las ciencias, por utilizar una expresión del poeta griego Arquiloco el cual escribía: “muchas cosas sabe la zorra pero el erizo sabe una sola y grande”.

Yo, tengo la suerte de pertenecer a un gran equipo que ha tratado de resolver problemas reales y en ocasiones también ha tratado de diseñar sus propios modelos teóricos. Nuestra aportación, grande o pequeña, al desarrollo de la ciencia ha sido factible gracias al espíritu de colaboración y a la conjugación del esfuerzo con el compañerismo. Son muchos los nombres y apellidos, los hombres y mujeres que han participado, que han colaborado para hacer posible la obtención de los resultados a alguno de los cuales me he referido en la última parte de mi discurso. Casi todas estas personas se encuentran esta noche aquí, y saben de sobra de mi reconocimiento y agradecimiento a su trabajo, a su esfuerzo y su buen hacer, como saben de mi cariño hacia ellos. Y me ilusiona y me emociona comprobar que este afecto es recíproco al verlos aquí esta noche en este acto tan importante para mí.

No quisiera retener por más tiempo vuestra atención pero sí quiero agradecerles de corazón a todos ustedes señoras y señores su presencia en este acto, lo que sin duda entiendo como una manifestación de cariño y amistad.

Termino reiterando que me siento muy honrado en ingresar en la Real Academia Sevillana de Ciencias, lo que significa para mí el estímulo de considerarme unido estrechamente a un conjunto de colegas académicos comprometidos con lo que Unamuno calificó como uno de los mayores servicios que se pueden prestar a una nación, entregarse de una manera callada y persistente a cultivar lo que él denominó el heroísmo del trabajo y el culto a la verdad.

Muchas Gracias

BIBLIOGRAFÍA

- [1] LAKATOS, IMRE. Historia de la Ciencia y sus reconstrucciones racionales. Tecnos. Madrid. 1974.
- [2] FRIEDRICHS, ROBERT. Sociología de la Sociología. Amorrortu. Buenos Aires, 1977.
- [3] BRINGUIER, JEAN CLAUDE. Conversaciones con Piaget. Gedisa. Barcelona 1977.
- [4] GOLDMANN, LUCIEN y OTROS. Jean Piaget y las ciencias sociales. Sígueme. Salamanca, 1974.
- [5] POPPER, KARL R. La lógica de la investigación científica. Tecnos. Madrid, 1980.
- [6] BUNGE, MARIO. La investigación científica Ariel. Barcelona 1969.
- [7] MERTON, ROBERT K. Teoría y estructura social F.C.E. México, 1964.
- [8] CORDON, FAUSTINO. La función de la ciencia en la sociedad. Edicusa. Madrid 1976.
- [9] NEEDHAM, JOSEPH. Las matemáticas y las ciencias en China y en Occidente, en B. Barnes, Estudios sobre sociología de la Ciencia. Alianza. Madrid 1980.
- [10] FARRINGTON, BENJAMÍN. Mano y cerebro en la Grecia antigua. Ayuso. Madrid 1974.
- [11] DEL CAMPO, SALUSTIANO. La sociología científica moderna. IEP Madrid 1969.
- [12] LOSEE, JOHN. Introducción histórica a la filosofía de la Ciencia: Alianza Editorial. Madrid, 1982.
- [13] KUHN, THOMAS. La estructura de las revoluciones científicas, F.C.E. México, 1971.
- [14] MERTON, ROBERT K. Los imperativos institucionales de la ciencia, en B. Barnes, Estudios sobre sociología de la Ciencia. Alianza. Madrid 1980.

CONTRIBUCIONES CIENTÍFICAS

- GARCÍA DE LAS HERAS, J. (1988). Técnicas de identificación de outliers para el Modelo Multinomial. Tesis Doctoral. Universidad de Sevilla.
- GARCÍA DE LAS HERAS, J., MUÑOZ GARCÍA, J. and PASCUAL ACOSTA, A. (1993). Criteria for the detection of outliers in the multinomial model. Journal of Applied Statistics. Vol. 20. pp. 137-142.
- MUÑOZ GARCÍA, J. (1980). Algunas Técnicas sobre Detección de Outliers. Tesis Doctoral. Universidad de Sevilla.
- MUÑOZ GARCÍA, J. y PASCUAL ACOSTA, A. (1982). Criterio de detección de outliers en modelos probabilísticos tipo Pareto. Estadística Española n.º 95, pp. 113-120.
- MUÑOZ GARCÍA, J. (1984 a). Criterio para detectar outliers en poblaciones normales bivariantes. Trabajos de Estadística e Investigación Operativa. Vol. 35. pp. 65-82.
- MUÑOZ GARCÍA, J.; PASCUAL ACOSTA, A. y PÉREZ DIEZ DE LOS RÍOS, J.L. (1984 b). Detección de outliers en el modelo de correlación intraclase. Actas de la XIV Reunión Nacional de Estadística e Investigación Operativa. Granada. pp. 336-340.

- MUÑOZ GARCÍA, J. y PASCUAL ACOSTA, A. (1986). Análisis singular de un conjunto de datos. Las observaciones outliers. *Thales*. Vol. 6. pp. 33-39.
- MUÑOZ GARCÍA, J. MORENO REBOLLO, J.L. and PASCUAL ACOSTA, A. (1987). Detecting outliers in s - mulynomial distributions through adjusted residuals. *J. Appl. Statist.* Vol. 14, pp. 171-176.
- MUÑOZ GARCÍA, J., MORENO REBOLLO, J.L. and PASCUAL ACOSTA, A. (1990). Outliers: A formal approach. *Int. Stat. Rev.* Vol. 58, pp. 215-226.
- MUÑOZ PICHARDO, J. (1992). Análisis de Influencia en el Modelo Lineal General. Sesgo condicionado. Tesis doctoral. Universidad de Sevilla.
- MUÑOZ PICHARDO, J.; MUÑOZ GARCÍA, J. (1995). A new approach to the influence análisis in linear models. *Sankhya, Serie A*. Vol. 57. pp. 393-409.
- JIMÉNEZ GAMERO, M.D.; MUÑOZ PICHARDO, J.; MUÑOZ GARCÍA, J. and PASCUAL ACOSTA, A. (2002). Rao distance as a measure of influence in the multivariate linear model. *Journal of Applied Statistics*. Vol. 29, pp. 841-854.

DISCURSO PRONUNCIADO POR EL EXCMO. SR. D. RAFAEL INFANTE MACÍAS

Académico Numerario

en contestación al leído por el

EXCMO. SR. D. ANTONIO PASCUAL ACOSTA

en el Acto de su recepción como Académico Numerario

celebrado el día 11 de diciembre de 2003

Excmo. Sr. Presidente del Instituto de Academias de Andalucía; Excmo. Sr. Secretario General de Universidades e Investigación; Excmo. Sr. Presidente de la Real Academia de Bellas Artes de Santa Isabel de Hungría de Sevilla, que nos abre hoy sus puertas generosamente para celebrar este acto; Excmo. Sr. Representante del Tribunal Constitucional; Excmo. Sr. Director de la Real Academia de Buenas Letras; Excmo. Sr. Representante de la Academia Iberoamericana de Farmacia; Excmos. Sres. Rectores de las Universidades de Jaén y Almería; Excmas. e Ilmas. Autoridades; Excmos. e Ilmos. Sres. Académicos, Señoras y Señores.

La incorporación de un nuevo académico a nuestra Real Academia es siempre motivo de júbilo. Y en esta ocasión, la del Excmo. Sr. D. Antonio Pascual Acosta lo es por varios motivos.

En primer lugar porque se incorpora a la Academia un prestigioso matemático. Cubre sobradamente el profesor Pascual Acosta el requerimiento que impone el artículo noveno de los estatutos de la Academia, de que para ser elegido académico numerario, el aspirante deberá poseer un reconocido prestigio científico.

En segundo lugar, el profesor Pascual Acosta viene a aumentar la nómina de académicos numerarios de la institución, que aún lejos de alcanzar la cifra de 48 miembros de número que permiten sus estatutos, con su incorporación sobrepasa ampliamente el 60 % de su composición total.

Además, se incorpora un académico con excepcionales dotes personales. Por su manera de ser, siempre afable, alegre, animoso y comunicativo, hace que sea un placer tratar con él aunque sea por unos minutos, lo que es corriente para los que convivimos con el profesor Pascual Acosta en el Campus de Reina Mercedes de nuestra universidad, adonde acude diariamente a cumplir con sus obligaciones universitarias, no importa cuantas tareas le lleven fuera de la universidad. Y es que contagia su entusiasmo, dinamismo y buen hacer.

Pero el acto de hoy es especialmente jubiloso porque por fin se incorpora a la academia el Excmo. Sr. D. Antonio Pascual Acosta. Fue nombrado académico electo el

18 de noviembre de 1991, como ha indicado al comienzo de este acto el Ilmo. Sr. D. Francisco Sánchez Burgos, secretario de la institución. Han pasado por tanto 12 años desde su nombramiento.

Los estatutos de la Academia requieren que los académicos electos entreguen su discurso de entrada en el plazo de un año desde su nombramiento. Pero también indican que si por una causa justificada el académico electo solicita una prórroga de dicho plazo, es potestad de su presidente concederla. El Excmo. Sr. D. Antonio Pascual Acosta la tenía y de peso. Había sido nombrado Consejero de Educación y Ciencia de la Junta de Andalucía y el Presidente de la Academia, por aquellas fechas el Excmo. Sr. D. Juan Martínez Moreno, así lo entendió y accedió a su petición para ampliar el plazo.

Entregó su discurso en 1999 y sus ocupaciones extra-académicas han sido, en buen aparte, la causa de que la lectura del mismo se haya ido retrasando. Tras varias fechas fijadas y sus propuestas, la lectura ha tenido lugar, finalmente, en el día de hoy, a pesar del accidente sufrido por el recipiendario hace varias semanas, lo que sin duda le ha supuesto un doble esfuerzo.

Tengo la seguridad de que el nuevo académico se incorporará a las tareas de la academia en la medida que se lo permitan sus compromisos fuera de los ámbitos académicos y universitarios a que le han llevado el que los responsables políticos de nuestra comunidad, conocedores de las extraordinarias dotes y excepcional capacidad de trabajo que acompañan al profesor Pascual Acosta, le hayan pedido que dedicara parte de sus esfuerzos al bien común de nuestra sociedad.

Excmo. Sr. D. Antonio Pascual Acosta, esta Real Academia Sevillana de Ciencias lo recibe con los brazos abiertos, como simboliza el abrazo que ha recibido de sus miembros y compañeros, reunidos hoy en esta sesión académica para acogerlo con júbilo, con la seguridad de que su incorporación redundará en beneficio de la academia que espera de usted todo, que es mucho, sea bienvenido.

Excmo. Sr. Presidente del Instituto de Academias de Andalucía, Excmo. Sr. Secretario General de Universidades e Investigación, Excmo. e Ilmas. Autoridades, Sres. Académicos, Señoras y Señores, quiero terminar estas palabras para en nombre de la institución, felicitar al Excmo. Sr. D. Antonio Pascual Acosta por haberse incorporado a la Academia; y a la Real Academia Sevillana de Ciencias, por el acierto de esta designación. Y para agradecer a todos ustedes, Excmos. e Ilmos. Señores, Señoras y Señores, el que hayan querido compartir con nosotros este acto.

Muchas gracias.