ANALES DE PSICOLOGÍA N.º 4, 1987. PÁGS. 97-105

Análisis de datos exploratorio y formulación de hipótesis

ANTONIO P. VELANDRINO NICOLÁS
JULIO SÁNCHEZ MECA
JOSÉ A. LÓPEZ PINA
Departamento de Psicología. Universidad de Murcia

RESUMEN

De acuerdo con las nuevas técnicas estadísticas de análisis de el enfoque exploratorio pretende ampliar la finalidad de la etapa clásica de la descripción. Es dentro de esta visión más globalizadora donde recientemente ha surgido una polémica respecto a si los orígenes empiricistas de Análisis de Datos Exploratorio (ADE) permiten entenderlo como una metodología generadora de hipótesis. Tras considerar los dos polos de tal polémica se propone que, dentro del actual realismo científico, la relación datos-hipótesis debe suponer un proceso interactivo.

Palabras clave: Ralismo Científico, Empiricismo, Análisis de Datos Exploratorio, Estadística Descriptiva.

ABSTRACT

Among the new statistical techniques of data analysis, the EDA approach tries to wide up the goal of the descriptive clasical stage. This new look has

raised a controversy concerning to empiricist origins of EDA. We proposed, within the current scientific realism, that the relationship data-hypotheses must carry over an interactive process.

Key words: Scientific Realism, Empiricism, Exploratory Data Analysis, Descriptive Statisfics.

INTRODUCCIÓN

Las técnicas estadísticas clásicas, tanto descriptivas como inferenciales [1], han sido establecidas para resultar útiles en la medida en que sean ciertos una serie de supuestos de variada índole que en algunos casos llegan a ser altamente restrictivos. Lamentablemente, como es bien conocido, los datos en un gran número de ocasiones no se ajustan a tales supuestos. Esta situación se ha intentado suavizar en parte efectuando numerosos estudios para determinar bajo qué condiciones de trabajo son admisibles determinadas violaciones de dichos supuestos. Pero los resultados obtenidos son, en la mayoría de las ocasiones, imprecisos y las conclusiones obtenidas ambiguas [2]. Dentro de un amplio marco de renovación estadística actual uno de los avances más prometedores, en un intento de solventar la anterior dificultad y, en consecuencia, de ampliar los objetos fundamentalmente de carácter aplicado de la estadística, se encuentra formalizado en una serie de técnicas conocidas con el nombre de *Análisis de Datos Exploratorio*, ADE (Exploratory Data Analysis, EDA), y *Análisis de Datos Confirmatorio*, ADC (Conformatory Data Analysis, CDA).

Ambos grupos de técnicas son, en primera instancia, las dos grandes fases que podemos considerar en el proceso de análisis de datos: una primera de exploración u observación y una posterior de confirmación o decisión (véase Hoaglin, Mosteller y Tukey, 1983).

Ahora bien, recientemente ha surgido una enconada disputa entre meto-

1 No consideramos aquí la actual revisión de lo oportuno que haya resultado la tradicional distinción entre estadística descriptiva e inferencial. Puede verse a este respecto la aportación de Shvyrkov - (1984).

2 Otra alternativa (si bien con fines inferenciales) bastante aceptada (sobre todo en Ciencias Sociales) ha consistido en reducir o eliminar buena parte de las restricciones. Esta alternativa es la conocida como estadística no paramétrica. Entre las desventajas que presenta puede contabilizarse (a) una pérdida de precisión (aunque una ganancia en generalización), (b) una disminución en la potencia-eficiencia, sólo subsanable a expensas de aumentar los costos de términos de tamaño muestral. Una amplia discusión de este tema puede encontrarse entre otros, en Siegel (1979) y Jenkins et al. (1984).

dólogos al establecer la naturaleza teórica de la primera de las etapas mencionadas del análisis de datos. Disputa que se extiende tanto a sus orígenes como a los objetivos que persigue. Esta discusión es la que pretendemos recoger en el presente informe.

ESTADÍSTICA DESCRIPTIVA

De acuerdo con I. J. Good (1983) la teoría estadística clásica en su aspecto descriptivo se encuentra estructurada principalmente por pretensiones. La primera de ellas tiene que ver con lo que podríamos llamar «composición visual de los datos». La simple disposición de una tabla de contingencia nos ayudará a comprenderla. Cuando tenemos dos variables, A y B (categóricas o numéricas), cada una de ellas con A_1, A_2, ..., A_r y B_1, B_2, ..., B_s niveles, respectivamente, y con n_{ij} observaciones o frecuencias para cada una de las combinaciones de niveles de A y B (con i = 1, 2, ..., r y j = 1, 2, ..., s), la presentación de esta masa de datos la efectuamos según:

	B_1	B_2	...	B_s
A_1	n_{11}	n_{12}	...	n_{1s}
A_2	n_{21}	n_{22}	...	n_{2s}
.	.	.	.	.
.	.	.	.	.
A_r	n_{r1}	n_{r2}	...	n_{rs}

Si en cambio, estos datos los introdujéramos en un ordenador, la disposición necesaria para cada dato debería ser: (A_1, B_1, n_{11}), (A_1, B_2, n_{12}), ..., (A_r, B_s, n_{rs}), o cualquiera similar (dependiendo de las exigencias concretas de la entrada de datos de cada máquina). Resulta claro que la presentación tabular le permite al ojo humano ver con relativa claridad buena parte de lo que está sucediendo en tal cuerpo de datos. Lo cual resultaría muy difícil con cualquier otra disposición.

La segunda de las intenciones descriptivas se encuentra en la reducción de datos. Con esta estrategia se pretende suprimir todas aquellas características de escaso o nulo valor informativo en favor de aquellas otras que resulten más importantes para comprobar la conducta manifestada por los datos. Un ejemplo tópico para ilustrar esta meta de la estadística descriptiva lo suministra el hecho de que aún en el caso de que conociéramos la posición en un instante

determinado de todas las moléculas de un gas, sería preciso desechar la mayor parte de tal información si pretendemos observar el gas como una totalidad.

Podemos resumir, por tanto, afirmando que la estadística descriptiva pretende fundamentalmente organizar y destacar la información relevante de una masa de datos en función de las capacidades cognitivas humanas (Good, op. cit.).

ANÁLISIS EXPLORATORIO DE DATOS

¿Qué supone de nuevo el ADE con respecto a la estadística descriptiva clásica? En primer lugar podemos afirmar que el ADE no supone una ruptura con la estadística tradicional, sino su desarrollo y complemento (véase Leinhardt y Leinhardt, 1980, p. 87). Así lo entiende también Good (op. cit.) quien reserva para la estadística descriptiva el objetivo de buscar patrones (entendiendo por tales todas aquellas conductas puestas de manifiesto por los datos que presentan una elevada probabilidad, lógica o subjetiva, de ser, al menos en parte, potencialmente explicables), mientras caracteriza el ADE como una metodología de formulación de hipótesis.

Para este autor el origen del ADE hay que buscarlo en la tradición iniciada con Francis Bacón (1561-1626), quien afirma que «el método básico de la ciencia es la colección sistemática y tabulación de observaciones, lo que permitirá casi siempre y automáticamente el descubrimiento de importantes verdades científicas, dando lugar a que ciertas formas de razonamiento incorrecto sean evitadas» (Good, op. cit., p. 287). De esta manera el ADE iría un paso más allá de la estadística descriptiva en la dirección de la auténtica ciencia en el sentido de que sugiere hipótesis. Esta forma de proceder se concretaría en la regla básica y principal del ADE: mirar los datos, para que, tras obtener de ellos toda la información que pueden suministrar, plantear las hipótesis pertinentes que han de ser evaluadas a través del análisis confirmatorio. Esta misma idea la expresa también P. D. Finch (1981) cuando afirma que «la estadística inferencial presupone que las hipótesis relevantes han sido formuladas previamente, y su papel estriba primariamente en someterlas a prueba y confirmarlas. Pero las hipótesis han de ser formuladas antes de que puedan ser probadas, confirmadas o rechazadas. La estadística clásica a menudo falla al explicar de dónde provienen las hipótesis probadas y por ellas misma proporciona poca o ninguna base teórica para determinar qué hipótesis deben ser probadas. Sus procedimientos están basados en la suposición de

que estas cuestiones han sido ya establecidas. La descripción de datos juega un papel preliminar en la formación de hipótesis (...)» (p. 138).

Este papel de los datos en la formación de hipótesis es cuestionado por S. A. Mulaik (1985). Para él la estadística en cuanto estadística trata de los datos, su descripción e inferencias dependiendo de la distribución de probabilidad que los datos originan. Según este autor «ninguna hipótesis sobre los acontecimientos del mundo puede ser generada a partir de los datos sin establecer algunos supuestos acerca de la conexión entre los datos y los fenómenos del mundo» (Mulaik, op. cit., p. 425). De acuerdo con esta forma de entender los datos y su análisis, la generación de hipótesis es tarea exclusiva del científico apoyado en los instrumentos que le suministra la estadística. Mulaik (op. cit.) entiende que es el científico-especialista quien está en condiciones de establecer las hipótesis relativas a los eventos que suceden en el mundo y evaluar la importancia de los patrones que aparecen en los datos.

Good (op. cit.) también reconoce la necesidad del científico para explicar el significado de los patrones en los datos, pero tanto él como Finch (op. cit.) defienden que es el análisis exploratorio de datos la más importante fuente generadora de hipótesis.

Además Mulaik entiende que a pesar de no poder descartarse completamente ciertas influencias baconianas en la base histórica de la exploración de datos (no debemos olvidar que Bacon junto con Hume y Locke son los principales representantes de la filosofía empiricista inglesa), es el empiricismo más ortodoxo el que da origen a dicho análisis de datos. Según Mulaik, Good no tiene demasiadas razones en su aproximación Baconiana puesto que el método de Bacon «no implica más que: (1) la realización de numerosas y cuidadas observaciones, (2) haciéndolas sin prejuicios o hipótesis, y después (3) un examen completo de los datos recogidos para ver las relaciones de cara a efectuar una taxonomía o clasificación de los hechos» (Daniels, 1968, citado por Mulaik, op. cit., p. 412).

La postura de Mulaik se radicaliza cuando efectúa una valoración del ADE. Admite este autor que la actuación conjunta estadístico-científica puede establecerse en los siguientes términos: el estadístico mediante su análisis exploratorio localiza patrones en los datos y sugiere al científico lo que tales patrones pueden representar. Es entonces cuando el científico actúa para determinar si ello es posible. Pero también acepta Mulaik que el científico por sí mismo puede tener de antemano una concepción, en forma de hipótesis de lo que acontece en el mundo y predecir, en consecuencia, que un determinado patrón aparecerá en los datos. El científico según esto, no nece-

sita mirar los datos para generar hipótesis: «Puede haber más en la experiencia e incluso en nuestras nociones de lo que existe en el mundo de lo que es dado por los datos» (Mulaik, op. cit., p. 426). A este respecto Mulaik recoge la interesante aportación del sociólogo L. Guttman. Este autor ha propuesto recientemente un método para desarrollar marcos teóricos conceptuales dentro de los cuales es posible establecer hipótesis; tal método conocido como Análisis de Facetas (Guttman, 1977a, 1977b, citado por Mulaik, op. cit., p. 421). Sin entrar a detallar esta nueva metodología, sí diremos que es consecuencia de la postura crítica de Guttman hacia lo que él entiende excesiva preocupación de sociólogos y psicólogos por los métodos de reducción y exploración de datos a expensas de una debida atención a las relaciones entre los conceptos importantes de un área o dominio substantivo de conocimientos y la estructura empírica de las observaciones (datos) para que pueda tenerse una idea clara de qué es lo que se está tratando en los datos.

DISCUSIÓN

Sin dejar de valorar la crítica que expone Mulaik en el sentido de que probablemente haya sido forzar demasiado el método baconiano para situarlo como el origen de una metodología generadora de hipótesis, no podemos estar del todo de acuerdo con él cuando deja reducido al ADE a una mera técnica descriptiva que «(...) puede a menudo provocar resultados ambiguos» (Mulaik, op. cit., p. 427). Y esto aunque sólo sea tomando como base el quehacer cotidiano de la mayor parte de los analistas de datos que adoptan el punto de vista exploratorio, y las recomendaciones elaboradas por los más prestigiosos pioneros de esta metodología: Tukey (1977), Velleman y Hoaglin (1981), Hoaglin, Mosteller y Tukey (1983).

Sí estamos de acuerdo con Mulaik que, dentro del actual clima de realismo científico, el dato es una entidad relativa. El dato, en efecto, puede ser valorado desde distintas ópticas en función del modelo teórico que el investigador tenga en mente. Pero también es cierto que, de acuerdo con Tukey (op. cit.), los datos pueden entenderse como las «pistas» que sigue un detective en su trabajo de búsqueda de pruebas para presentarlas ante la valoración de un jurado.

No podemos, por último, dejar de valorar los cinco objetivos que propone Good (op. cit.) para establecer el marco teórico del enfoque exploratorio. A saber:

1. Presentación de datos.

2. Aislamiento de patrones.
3. Formulación de hipótesis.
4. Búsqueda de nuevas hipótesis de mayor alcance explicativo, y
5. Racionalidad de tipo II.

Respecto a los objetivos 1 y 2 parece deseable que, como de hecho sucede, el ADE haga suyos los propósitos de la estadística descriptiva de organizar y presentar los datos de acuerdo a las peculiaridades cognitivas humanas, y el de servir como instrumento para detectar los posibles patrones que revelen la conducta de los datos. En cuanto al tercero de los objetivos, Good parece olvidar que la teoría también juega un importante papel como generadora de problemas e hipótesis. El objetivo 4 recoge el trabajo cotidiano de mejorar la(s) primera(s) hipótesis una vez establecida(s) —si ello es posible— estableciendo otra(s) de mayor potencia explicativa, mediante una de sus técnicas más conocidas: el análisis de residuales [3]. Ahora bien, lo que puede suceder con esta forma de actuar es que el analista acabe haciendo ciencia substantiva, con el evidente riesgo que ello puede suponer. El último de los objetivos es totalmente incuestionable por cuanto se refiere al hecho de que el analista debe tener siempre presente como propósito prioritario el maximizar la utilidad de su análisis teniendo presente los costos necesarios para llegar a una conclusión respecto de sus datos (Good hablaría de hipótesis en lugar de solución). Es claro que si este principio del quehacer científico es siempre deseable, lo es especialmente para el ADE debido a los varios procedimientos para analizar los datos.

CONCLUSIONES

La polémica se encuentra centrada, como vemos, en establecer cuál es el momento de establecer las hipótesis y, como consecuencia, a quién compete hacerlo. Nosotros entendemos que no hay ninguna inconsistencia en admitir que precisamente por tener el ADE sus orígenes en el más ortodoxo empiricismo, el acercamiento a los datos no puede realizarse con ideas preconcebidas, y son ellos los que deben mostrarnos posibles explicaciones sobre lo que sucede en el mundo real. No obstante, no podemos ser tan ingenuos (como el mismo Mulaik establece) como para negar que precisamente los datos son

3 En realidad el análisis de residuales no es una técnica desarrollada estrictamente por el análisis exploratorio de datos. Pero sí la ha incorporado como un instrumento más muy valioso para reanilizar los datos. Véase entre otros Goodall (1983).

recogidos en función de algunas expectativas previas. Expectativas más o menos formalizadas o explícitas que dependerán fundamentalmente (aunque no exclusivamente) de un cuerpo teórico establecido.

Además si consideramos el hecho de que el análisis exploratorio tiene un carácter observacional antes que el de una técnica para analizar resultados procedentes de rigurosos procedimientos experimentales, resultará fácil admitir que sólo tras la atenta mirada de los datos se podrá intentar generar hipótesis sobre la conducta que ellos nos revelan.

Pensamos, para finalizar, que el ADE forma parte de una nueva visión más globalizadora de la estadística, en el sentido de que mientras el enfoque tradicional daba por establecido la formulación de hipótesis (y modelos) sin atender a su orígen, esto es sólo una parte del trabajo de análisis de datos. Es necesario establecer un proceso interactivo en el que se considere un segundo aspecto donde los datos puedan dar lugar a la proposición de ciertas hipótesis. En definitiva, un proceso datos $\rightleftharpoons$ hipótesis en el cual el camino no es de una única dirección y la secuencia temporal no se encuentra prefijada.

REFERENCIAS BIBLIOGRÁFICAS

DANIELS, G. H. (1968). *American Science in the Age of Jackson*. New York: Columbia University Press.

FINCH, P. D. (1981): On the Role of Description of Statistical Inquiry. *Brit. Jour. for the Phil. of Scie., 32*, 127-144.

GOOD, I. J. (1983): The Philosophy of Exploratory Data Analysis. *Philosophy of Science, 50*, 283-95.

GOODALL, C. (1983): Examining Residuals. En D. C. Hoaglin, F. Mosteller and J.W. Tukey (Eds.) (op. cit.).

GUTTMAN, L. (1977a) What Is Not What in Statistics. En I. Borg (Ed.) (1981). *Multidimensional Data Representations: When & Why*. Ann Arbor: Mathesis Press.

— (1977b). What Is Not What in Theory Construction. En I. Borg (Ed.) (1981). *Multidimensional Data Representations: When & Why*. Ann Arbor: Mathesis Press.

D. C. HOAGLIN, F. MOSTELLER and J. W. TUKEY (eds.) (1983): *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley.

JENKINS, S. J. et al. (1984): Evaluating Criteria for Selection of Nonparametrics Statistics. *Perceptual and Motor Skill. 58*, 979-984.

LEINHARDT, G. and LEINHARDT, S. (1980): Exploratory Data Analysis: New Tools for the Analysis of Empirical Data. *Review of Research in Education, 8*. 85-157.

MULAIK, S. A. (1985): Exploratory Statistics and Empiricism. *Philosophy of Science, 52*, 410-430.

SIEGEL, S. (1979): *Estadística no paramétrica*. México: Trillas (original: McGraw-Hill, 1956).

SHYVYRKOV, V. V. (1984): Epistemological Foundations of Statistics. *Quality and Quantity, 18*, 351-366.

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

VELLEMAN, P. F. and HOAGLIN, D. C. (1981): *Applications, Basics and Computing of Exploratory Data Analysis*. Boston: Duxbury Press.