Histoire et Préhistoire de L'Analyse des Donneés

por *Rubén Hernández Cid*

J. P. Benzécri, *Histoire et Préhistoire de L'Analyse des Donneés*. París: Dunod, 1982.

El análisis de datos fundado en el uso de la computadora es una nueva metodología que la estadística aporta a la ciencia y especialmente a las ciencias humanas. Es con estas palabras (pronunciadas originalmente en 1969 y recogidas en el segundo volumen de *L'Analyse des Données*) que Benzécri inicia la exposición de los "principios fundamentales del análisis de datos" con los cuales sitúa esta nueva disciplina en el contexto general de la estadística. Dichos principios pueden ser resumidos muy brevemente de la siguiente manera:

1º *La estadística no es probabilidad.* En este punto se comenta la ambigüedad implícita en las aplicaciones de la probabilidad debido al hecho que la noción misma de probabilidad no es única. Esto ha dado como resultado la existencia de múltiples corrientes que frecuentemente son opuestas. Se propone en cambio, el uso de un análisis estadístico independiente de interpretaciones particulares de la probabilidad pero que esté basado en el álgebra y en la geometría.

2º El *modelo debe seguir a los datos y no al revés.* Se hace en este caso, una severa crítica a la tendencia, que según el autor, consiste en querer sobre todo *ajustar* modelos a los datos más que tratar de *extraer* estructuras de ellos.

3º *Es conveniente tratar simultáneamente informaciones que conciernen al mayor número posible de dimensiones.* Aquí se comenta la diferencia entre *experimentación* y *observación* en Estadística. Sólo por medio del estudio sistemático de las relaciones

entre variables observadas es posible proponer hipótesis. Este estudio implica el considerar el mayor número de variables posibles.

4º *Para el análisis de hechos complejos y principalmente de hechos sociales, la computadora es indispensable.* El hecho de pretender tratar cada vez un mayor número de informaciones hace necesario implementar nuevos métodos de análisis y de síntesis. La aparición de la nueva tecnología electrónica permite llevar a cabo dichas tareas.

5º *Utilizar una computadora implica abandonar todas las técnicas concebidas antes de la llegada de la computadora. Técnicas, no ciencias.* Se comenta el hecho de que la estadística, habiendo encontrado frecuentemente problemas de cálculo, derivó hacia soluciones donde la "intuición" suplía los argumentos matemáticos. Con el uso de la computadora se impone la revisión de todos los métodos estadísticos, lo que implica el fin de no pocas técnicas.

A casi veinte años de la declaración de estos principios resulta interesante interrogarse no solamente acerca de su origen y su vigencia actual sino también acerca de la evolución e impacto del análisis de datos en el contexto general de la estadística.

En su *Histoire et Préhistoire...* Benzécri y su grupo, creadores y principales practicantes del análisis de datos, intenta mostrar el camino que va desde los más remotos gérmenes de la estadística descriptiva hasta la formulación y desarrollo del análisis de correspondencias, técnica central del análisis de datos.

En cuanto a sus funciones, el análisis de datos puede ser entendido como un conjunto de técnicas fundamentalmente descriptivas basadas en el álgebra y en la geometría que permiten el estudio de grandes volúmenes de datos multidimensionales de todo tipo (numéricos o no) por medio del uso intensivo de programas de cálculo electrónico. Cabe subrayar que en este conjunto de técnicas, al menos en su origen, no tienen carácter inferencial en el sentido clásico de la estadística por lo que debe tenerse presente esta diferencia al tratar de comparar el análisis de datos con otras escuelas estadísticas.[1]

Kendall (1970) fija como punto a partir del cual puede considerarse que la estadística comienza, la publicación de los primeros

[1] Nos referimos exclusivametne al análisis de datos en el sentido de los trabajos de Benzécri. Aun cuando el término es usado también por Tukey su interpretación es diferente a la considerada en el presente artículo.

tratados de Aritmética Política en el siglo XVII. Agrega que en los tratados más antiguos, tales como las cuantificaciones del pueblo israelita o el inventario de las propiedades de Carlo Magno, todo aspecto numérico es "accidental o aún mera conveniencia" tratándose en realidad "más de registros de una situación que una base para la estimación o la predicción". (Estos documentos son frecuentemente citados como vestigios del pensamiento estadístico contemporáneo). Benzécri, por su parte, al no tomar en cuenta el aspecto inferencial, sitúa el antecedente más remoto del análisis de datos en los registros de la administración egipcia de hace tres mil años. La diferencia no es anecdótica, se trata en realidad de dos concepciones distintas en cuanto a las funciones de la estadística: la *inferencia* y la *descripción*.

Entre estas dos funciones se encuentra el concepto clave: la probabilidad. Benzécri *et al* piensan que el análisis de datos "recibe del cálculo de probabilidades su inspiración pero no sus métodos" (p. 25) consideran además que "aún cuando el abuso del cálculo de probabilidades haya perjudicado a la estadística, el progreso de estas dos disciplinas será ... la historia de una sola ciencia: la del azar" (p. 3). Así, los trabajos fundamentales de Fermat, Pascal, Laplace, Bernoulli en los siglos XVII, así como los más recientes de Kolmogorov, Gnedenko, etcétera, son tomados en cuenta, en la obra estudiada, sólo en tanto puedan coincidir con la óptica del análisis de datos. Así por ejemplo, se ofrece una amplia discusión acerca del descubrimiento de la densidad normal multivariada debido al hecho de que el aspecto multidimensional es fundamental en análisis de datos. El énfasis está puesto en la evolución de la idea de lo multivariado y no en el aspecto del modelo probabilístico propiamente.

En lo que podría llamarse la historia reciente, son estudiadas a fondo tres épocas claves para el pensamiento estadístico: la biometría (de Quetelet a Pearson), la escuela de Fisher y la psicometría (de Spearman a Guttman). Este estudio es hecho partiendo de la siguiente tesis: "la expansión del dominio de la estadística ha determinado también su progreso" (p. 87).

La biometría debe su origen (hacia 1877) a la confluencia de dos corrientes del pensamiento científico que venían desarrollándose hasta entonces de manera independiente: por una parte, la antropometría en donde Quetelet habría incorporado el uso del modelo normal (inspirado en las aplicaciones en la astronomía) y por la

otra, las teorías resumidas en las tesis evolucionistas de Darwin. Es Galton quien movido por el deseo de "probar matemáticamente" las teorías de la evolución descubre empíricamente la noción de correlación entre dos variables biológicas. Posteriormente K. Pearson, basado en el desarrollo matemático de Dickson sobre los descubrimientos de Galton, extiende los conceptos de correlación y correlación parcial en un contexto del modelo normal multivariado. Debido al hecho de que en análisis de datos el estudio de las correlaciones es un tema central, Pearson es considerado como uno de los precursores de esta disciplina estadística.

De la obra de Fisher y sus discípulos (Kendall, Neyman, E. S. Pearson), Benzécri retiene para el análisis de datos, el carácter geométrico que la fundamenta. El diseño de experimentos es analizado con detalle concluyendo que si bien responde en algunas áreas precisas de la investigación, en ciencias sociales no es muy conveniente su aplicación. De hecho, este tema sirve de base para criticar una vez más el empleo del modelo normal así se trate del caso multivariado.

Así como los trabajos de Darwin son el origen de la biometría, las teorías psicológicas vigentes hacia 1935 (y particularmente las basadas en las tesis de Spearman) dan paso a la fundación de la psicometría. Aquélla dejó al análisis de datos el uso de métodos multidimensionales para el estudio de relaciones entre variables físicas de seres vivos, ésta por su parte, permitió el desarrollo de técnicas que tratan de descubrir dimensiones ocultas (tales como la "inteligencia") definidas a partir de combinaciones de medidas observables de manera directa (por ejemplo, notas obtenidas al resolver un examen).

La técnica fundamental en psicometría es el análisis en factores cuya solución numérica requiere de la búsqueda de los valores propios de una matriz de varianzas y covarianzas. Esta tarea (que en la actualidad no representa mayores problemas) hizo que en los años cuarenta se desarrollaran una serie de técnicas alternativas para salvar el problema numérico. Como consecuencia de estas búsquedas se desarrollan una gran cantidad de técnicas dentro de las cuales caben mencionar, por su relación con el análisis de correspondencias, aquellas que se refieren al tratamiento de datos no métricos (por ejemplo, respuestas en una escala nominal a una serie de preguntas). De esta época datan gran parte de las medidas de asociación entre variables nominales (*cf.* Goodman y Kruskall

(1954)), la teoría de las estructuras latentes y la construcción de escalas (*cf.* Reynolds (1980)).

La última etapa considerada en esta *Histoire et*... es la relativa al análisis de correspondencias. Es a partir de un problema en lingüística (la búsqueda de un método para la traducción automática) que las primeras formulaciones van tomando lugar. En 1983, Benzécri presenta en el Colegio de Francia la primera versión coherente del análisis de correspondencias, mostrando sus propiedades algebraicas y geométricas fundamentales. En realidad algunas de las fórmulas simplificadas por este análisis estaban ya contempladas en trabajos de Fisher, Maung y Kendall y Stuar pero de manera tangencial únicamente. En la versión presentada por Benzécri, el análisis de correspondencias permite tratar tablas de contingencias de dos criterios (variables cualitativas) de manera tal que es posible describir gráficamente las "proximidades" o asociaciones entre los niveles de los criterios estudiadas. Con la llegada de las primeras computadoras, el análisis de correspondencias multiplica sus aplicaciones (véanse los dos volúmenes de *L'Analyse des Données*) y desde el punto de vista teórico los avances se hacen cada vez más notorios. Así, el análisis de datos fue formulado desde un contexto general incluyendo dos grandes grupos de métodos: las técnicas de clasificación automática (*cf.* Roux (1985) y los análisis factoriales, en donde se sitúa el análisis de correspondencias. La historia narrada en la obra de Benzécri se detiene en 1975 con la generalización del análisis de correspondencias para cualquier número de criterios nominales y con la existencia de un buen acervo de programas de cómputo (para grandes configuraciones) realizando todos los análisis de datos, de manera eficiente.

Desde entonces, podemos observar tres características importantes que se desarrollan en la evolución del análisis de datos:

1. El fin del aislamiento que había mantenido al análisis de datos como una "escuela francesa de estadística". Los trabajos de Hill (1974), de Nishisato (1980), del grupo holandés bajo el nombre de A. Gifi (1981) y sobre todo la aparición de los libros de Greenacre (1984) y Lebart, Morineau y Warwick (1977) han dado una difusión más amplia de las técnicas en cuestión. Desde hace solamente 10 años, el *Index of Statistics* (de la ASA), incluye en sus páginas los trabajos relativos al análisis de correspondencias.

2. La aparición de las microcomputadoras vino a dar un nuevo impulso al análisis de datos. Los paquetes interactivos (SPAD,

SICLA, STATIS, etcétera) han permitido que el análisis de datos se utilice en nuevas áreas. Es de señalarse que en este campo, los paquetes tradicionales (SPSS, BMOP, SYSTAT, etcétera) aún no incluyen al análisis de correspondencias. Creemos que esta situación cambiará en corto tiempo.

3. El análisis de correspondencias, en su práctica actual, ha traído una preocupación acerca de la estabilidad en sus resultados. Esto ha desembocado, con el uso de técnicas de remuestreo como el método Bootstrap, en un planteamiento original respecto a la inferencia en análisis de datos (*cf.* Lebart, Morineau (1977), Volle (1981)). Por otra parte se empiezan a combinar de manera explícita técnicas del análisis de datos y de la estadística clásica inferencial, por ejemplo los trabajos de Escoufier (1986) (selección de variables en análisis de correspondencias) o de Goodman (1986) (análisis de correspondencias y modelos log-lineales).

En conclusión es posible decir que, al menos para el análisis de correspondencias, se esperan aún importantes contribuciones tanto en el terreno descriptivo (*cf.* D'Ambra, Lauro (1982)) como en el inferencial. De esta manera, de los *principios fundamentales* enunciados en 1969, sólo el primero ("la estadística no es probabilidad") está siendo cuestionado por el carácter inferencial *sui generis* surgido en análisis de correspondencias. En cuanto a los demás (¿quién dudaría en esta época acerca del uso de la computadora en la práctica estadística?), su vigencia es incontestable. De esta manera, el análisis de datos aparece cada vez más como un grupo de técnicas intermedias entre lo exclusivamente descriptivo y lo exclusivamente inferencial.

BIBLIOGRAFÍA COMPLEMENTARIA

D'Ambra, L., Lauro, N.: (1982), "Alcune estensioni dell'analisi in componenti principali in rapporto a sottospazi di riferimento". En: R. Leoni (ed.) *Alcuni lavori di analisi statistica multivariata*. SIS, Florencia, p. 41-67.

Benzécri, J. P.: (1973), *L'Analyse des Donnés*. 2 vols. París: Dunod.

Benzécri, J. P.: (1980), *Pratique de l'analyse des données*. 3 vols. París: Dunod.

Escoufier, Y.: (1986), "L'Analyse des Correspondances: ses propriétés er ses extensions". "Document de travait Université de Noutpellier.

Goodman, L. A.; Kruskal: (1954), "Measures of association for cross classifications" *JASA* 49: 736-64.

———: (1959), "Measures of association for cross classifications II", *JASA* 54: 123-63.

———: (1963), "Measures of association for cross classifications III", *JASA* 58: 310-64.

———: (1972), "Measures of association for cross classifications IV", *JASA* 67: 415-21.

Goodman, L. A.: (1986), "Correspondence analysis models, log-linear models and log-bilinear models for the analysis of con-ingency tables".

Gifi, A.: (1981), *Nonlinear Multivariate Analysis*. Leiden: R.V.I./ F.S.W.

Greenacre, M. J.: (1984), *Theory and applications of correspond-ence analysis*. London: Academic Press.

Hill, M. O.: (1974), "Correspondence Analysis: A neglected Mul-tivariate method". *Appl. Statist.* 23 Nº 3, pp. 340-354.

Hill, M. O.: (1982), "Correspondence Analysis" en *Encyclopedia of Statistical Sciences* (Kotz & Johnonsons edts.). 2 pp. 204-210. Wiley.

Kendall, M. G.: (1970), "Where Shall the History of Statistics begin", en *Studies in the History of Statistics an dProbability* (Pearson and Kendall edts.). Londres: Friffin.

Lebart, L.: (1976), "The significance of eigenvalues issued from correspondence analysis", en *Proceedings in computational sta-tistics* (COMSTAT) pp. 38-45. Viena. Physica-Verlag.

Lebart, L.; Morineau, A.; Tabard, N.: (1977), *Techniques de la description Statistique, Méthodes er programmes pour l'nalyse des grands tableaux*. Paris, Dunod.

Lebart, L.; Morineau, A.; Warwick, W.: (1984), *Multivariate Descriptive Statistical Analysis. Correspondence Analysis and Related Techniques*, N. Y.: Wiley.

Nishisato, S.: (1980), *Analysis of categonical data, dual sclaing and its applications*. Toronto: University of Toronto.

Reynold, H. T.: (1980), *The analysis of cross-classifications*. N.Y.: The Free Press.

Roux, M.: (1985), *Algorithmes de Classifications*. Paris: Masson.