Análisis estadístico de textos

Ludovic Lebart, André Salem, Mónica Bécue Bertaut

Editorial Milenio, Lleida, Spain, 2000
215 pàgines

The statistical analysis of texts is a quite new discipline that is becoming more and more important, both for the social scientists who want to study and compare any kinds of texts using quantitative methods and for the statisticians who deal with open questionnaires. In this interdisciplinary field the book «*Análisis Estadístico de Textos*» is a remarkable text, the first of its characteristics published in Spanish, which is written with an accesible language for a large range of human science researchers, and which presents, in a systematic way, the main subjects, concepts, methods and techniques of this subject. For these qualities, it constitutes a practical tool for research and also a basic contribution for the consolidation and legitimation of this new speciality.

T. Kuhn, in the papers collected in «The Road Since 'Structure'» (2000), claimed that he had always continued to think on the concept of paradigm and that he would finally relate it with the creation of new specialities of a discipline. We can emphasize that statistics, when applied to an specific matter such as economics take into account at the one hand theories of probability and algebra, that are integrated in the mathematical statistics and, on the other hand, the specific theories of the subject on which it works, Moreover it requires computer facilities and a large set of practical knowledge that is adquired in a long professional experience. For these reasons, modem statistics have been and continue to be very productive in the creation of new specialities or paradigms such as econometry, biometry, psychometry, and more recently, chemiometry. I think that we can look at the statistical analysis of texts from that point of view and consider that it is in a stage of social acceptance and consolidation.

The histoy of this kind of analysis, as professor Daniel Peña points out in his Foreworld of the book, has two periods. In the first period it only used univariate technics, focused on the analysis of frequences of words, and achieved some brilliant results, as those reported by an study of Mosteller and Wallace on the anonimous papers published in 1787-1788 in order to induce the citizens of New York to ratify the Constitution. In the second period, the statistical analysis of texts introduces multivariative technics. Now it is known that all present texts are elaborated and stored in computer supports, that our personal computers have a huge capacity for multivariative analysis and that this method can answer relevant questions about text. These facts justify that we expect a brilliant future for the statistical analysis of texts.

The book presents the following subjects: a short introduction to the notions of text and to statistical analysis, which focus on the exploratory analysis of textual data. The codification of open questionnaires. Lexical units and segmentation of texts. Lexicographical documents. Correspondence analysis of texts. Authomatic classification of tables and texts. Visualization of textual data. Searching for characteristic elements. Analysis of chronological corpus.

The quality of the book reflects the high scientific value and experience of its authors. Ludovic Lebart is research director at the CNRS (Centre National de la Recherche Scientifique) and professor of the ENST (Ecole Nationale Supérieur de Télécommunications) in Paris. He is one of the father founders of the French Correspondence Analysis School, has worked on the analysis of open questionnaires and has created the software pack «SPAD». André Salem is full professor on Language Sciences at the University Paris-Sorbonne. Mónica Bécue is professor of Statistics at the Universitat Politècnica de Catalunya and has collaborated in the development of «SPAD-T». All of them have relevant contributions to the statistical analysis of texts.

Eduard Bonet
Esade (Universitat Ramon Llull)