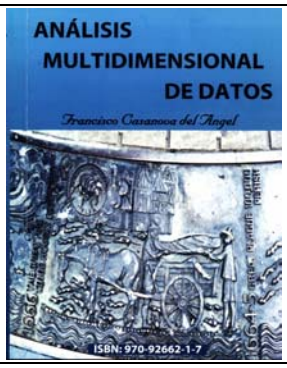
BIBLIOGRAPHY IN STATISTICS TO ENHANCE RESEARCH

Francisco Casanova del Angel
National Polytechnic Institute
Mexico

The document presented here, called "Análisis Multidimensional de Datos" in Spanish ("Multidimensional Data Analysis") is geared toward engineering students in Mexico, Latin America, and Spain. It shows all the theory of data analysis, starting with a biographical sketch of its historical development and how data are organized. It deals with the theory of factorial analysis and scalograms, beginning with establishing information management. A large variety of applications with actual data are presented throughout the text, and a set of programs is furnished that can be implemented easily in a personal computer. The related software is listed at the end.

INTRODUCTION

The academic task is not only concerned with teaching how to gather information. Furthermore, it is essential to teach the student how to study it appropriately so that he/she may learn how to obtain valid results and become skilled in making decisions in building the mathematical model that describes the research or study. Figure 1 shows the cover of the book under discussion.



Copyright: 03-2000-022710031100-01
ISBN Number: 970-92662-1-7
Bar code: 9789709266214

Figure 1. The Cover of the Book.

DISCUSSION

The first chapter presents an attempt to provide a historical description of what data analysis has been since the seventeenth century, when John Graunt made the first formal attempt to statistically interpret one of the social phenomenon of his times, by analyzing and making a clean copy of birth and death figures in London, England. It is precisely this fact that is captured in the photograph that provides the setting for the cover of the book, taken at the monument in July 2001 that gives an account of the history of London, located behind the famous London Tower (Casanova del Angel, F. 2001), see Figure 1.

The beginning of the consolidation of data analysis in the nineteenth century is described from Adolphe Quetelet (1796-1874) to Karl Pearson (1857-1936). In the twentieth century, although the appearance of the famous magazine *Biometrika* at the beginning of the century is not emphasized directly, its importance is brought to light by mentioning that this magazine published some of the works that gave shed light and provided guidance for the consolidation of data analysis, such as the works by W. S. Gossett and Ronald Fisher (1890-1962). Nor is there any direct mention made that this great century ended with the disappearance of one of the most renowned magazines of factorial analysis, namely, the Les Cahiers de l'Analyse des Données founded by Jean Paul Benzécri, as well as the flourishing of the Bernoulli Society for Mathematical Statistics and Probability.

The second chapter presents what should be information organization, the variable concept in statistics, and its codification. The third chapter merits a comprehensive analysis, since

it is the only document to date that graphically and accessibly describes how to build tabular information arrangements, as well as provide step-by-step examples. The 13 main algorithms for building tabular data arrangements assume that they are the entry target in the classical programming of the algorithms of the Multidimensional Data Analysis. Their names and/or denominations are: contingency table, frequency table, intensity notes table, table of types, logical or Boolean description table, Burt or generalized contingency table, Burt one-dimensional or mass diagonal variables, Burt lower triangular table, Burt upper triangular table, split table (Figure 2), range or category table, accumulated table by types (Figure 3), and multiple table. In order to give the reader an idea of its form of presentation, two of those arrangements are exemplified graphically in Figures 2 and 3.

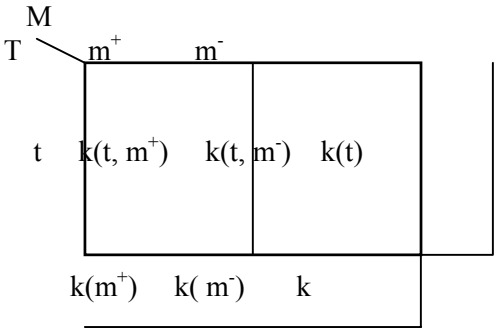


Figure 2. Split Tabular Arrangement.

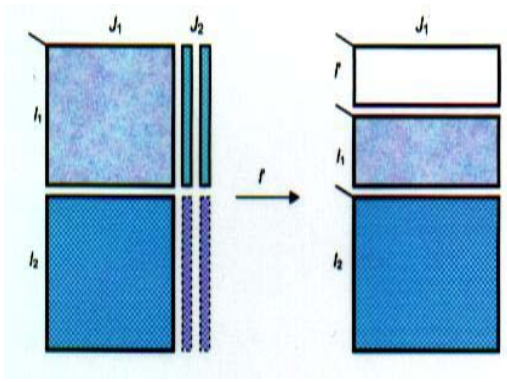


Figure 3. Accumulated Tabular Arrangement by Types.

The fourth chapter is devoted to a review of the vectorial and algebraic concepts necessary for developing the most currently used factorial methods. The fifth chapter presents and exemplifies the theory of the analysis in the main components of the correspondence factor analysis, discriminate analysis and canonical analysis, including a section of the theoretical relationship between all those multivariate analyses that have never before been presented in any multivariate statistical text book. Table 1 shows the objectives and characteristics of these methods. Chapter six is devoted to the main techniques used to describe and interpret a multidimensional data analysis that includes the basic theory of hierarchical classification. Chapter seven contains a step-by-step reading and interpretation of a multidimensional data analysis. The novelty of the second edition is chapter eight, which includes a certain number of explained theoretical exercises of the data factor analysis.

The ninth chapter merits special attention, in as much as it contains all the theory of the scalograms built in the shadow of a multivariate data analysis where the theory developed for weighted scalograms is present for the factor type scalograms, as well as for uni-rational and multi-rational scalograms. Applying the mathematical statistic in social sciences has centered on building indices and has developed them in number and variety, due to their extensive use in continuity forecasts and measurements. In the times of Louis Guttman, a contemporary of Ronald A. Fisher who established this theory, a large number of techniques were developed that combined individual scale values and indices. They were highly simple from a theoretical viewpoint and highly used in the area of human behavior. Guttman, a devotee of scalograms, builds a large diversity of order structures from a data analysis perspective, relative to simple or complex scalograms in alphabetical order, by using the components of an analysis in principal components. The applications were made in the study of expressions of fear in combat. Their scales are one-dimensional and generally cumulative.

Table 1
Objectives and Main Characteristics of the Data Analysis Methods

Method	***Objective***	***Characteristics***
Canonical Analysis	Search for relationships, if any, between two groups of variables	• Mathematical formalization of the problem. • Search for canonical variables. • Geometrical type note. • Graphic representation of individuals and variables. • Practical inconvenience.
Multiple Regression	Search for relationships between one variable (variable to be explained) and a group of variables (explanatory variables).	• Mathematical formalization of the problem. • Determination of the regression model parameters, and • As a particular case of canonical analysis.
Discriminant Analysis	Differentiate well defined k-groups of individuals a priori and determine a rule for applying a new individual to one of these groups.	• Mathematical formulation of the problem. • Discriminant factor axis. • Determination of discriminant variables step-by-step. • Obtaining the inherent values of the $T^{-1}E$ matrix. • Graphic representation of individuals and variables, and • As a particular case of canonical analysis.
Analysis of Principal Components.	Opening data reduction technique with appreciation criteria of that reduction.	• Mathematical formalization of the problem. • Principal axes. • Graphic representation of variables, and • As a particular case of canonical analysis.
Factor Analysis of Correspondences	Graphic representation of the lines and columns of a tabular data arrangement on plans created by the (main) axes of maximum length of the line points cloud (individuals) referred to in space $\Re^p$ with p as the number of columns or variables in the tabular arrangement.	• Mathematical formalization of the problem. • Factorial axes • Graphic representation of individuals and variables, and • As a particular case of canonical analysis.

In the last years of World War Two, Paul f. Lazarsfeld performed studies with *quasi* scales by proposing a return to the old concept of a latent ongoing attitude. His basic hypothesis was that there was a set of latent types such as ratios expressed between either two or more questions in a questionnaire that could be had for the existence of the basic types. Lazarsfeld defines an attitude as an inference to latent types. Other advances in this direction have been made by J. P. Benzécri who developed a weighted scale model of Guttman in 1975, starting with the analysis of answers to a questionnaire. Furthermore, two factor models are presented by this

writer, developed and applied for non-psychological variables, but used to measure continuities. Even though this part of Mathematical Statistics has evolved considerably in these last years, with a large number of works appearing at each world congress of the specialty, its bibliographical index has not been updated significantly.

Finally, a certain number of programs are presented in an exhibit, on the multidimensional data analysis in one of the simplest programming languages, in order for them to be implemented and modified in a personal computer. The ANABAS program is built in a Basic language and holds 14 Kb in the central memory, of which 4 Kb apply to the graphics subroutine. This program stores the factor coordinates in a 6.681 Kb. matrix. Once the program is saved on a disk, loaded into the memory of the microcomputer and executed with the respective instruction, this program will question the dimensions of the data table to be analyzed. After introducing the dimensions, the amount of remaining memory appears on the screen that will be available to work the graphics. The keyboard provides for reading the data or by a file previously stored on the disk. If the data are already stored on the disk, the program reads them automatically from the files denominated *vardes.dat* for the identifiers of the variables and individuals, and *desdo.dat* for numerical values. The names of these files can be changed freely, on the condition that it should be done prior to executing those files. One decides whether or not to print the data on paper. The program calculates the inherent values and vectors, and prints their histogram. It calculates factors and their contributions for both individuals and variables. It asks if it is necessary to edit the factors. If the answer is yes, the factors of individuals and their contributions are immediately printed on paper. Immediately afterwards, the same question is asked, but relative to variables. There is an option to store inertia, weight, and the factors of individuals, variables or of individuals and variables. The file denominated factores.dat is stored. The last important question is concerned with the subject of the graphics of the calculated factors. If the answer is *no*, the program stops executing. But if the answer is *yes*, there are two forms of graphics. One is through an E that means that the graph is shown on the screen. The other is through a P, which means that the graph is printed on paper. The program only graphs two factor axes at a time, and it is necessary to give the axis values separated by a comma every time they are required. There is the facility of obtaining separate graphics of individuals only or variables only, or if desired, of individuals and variables at the time of the desired combination of axes.

CONCLUSION

The best way to disseminate knowledge is through books, magazines, notes, and the now famous communication network via computers known as Internet (information superhighway). Professor G. Achenwall (1719-1772) probably never thought that introducing the term “Statistic” in class at the University of Gotinga in 1746 would encompass gathering, treating, and interpreting information, and that his methodology would benefit all types of societies. The academic and didactical task is easy if one is mindful that not only is it important to teach the concept and how to use it, but it is also essential to teach how to gather information, analyze and interpret it to be able to formulate and reformulate the building of the mathematical model that accurately describes the research or study performed.

REFERENCES

Casanova del Angel, F. (2001). *Multidimensional data analysis*. Mexico: Logiciels.