STATREE: AN EXPERT SYSTEM FOR CHOOSING SUITABLE STATISTICAL DATA PROCESSING TECHNIQUES

Claudio Capiluppi and Luigi Fabbris
Dipartimento di Scienze Statistiche
Via S. Francesco 33, 35121 Padova, Italy

1. Introduction

The choice of a statistical method suitable for the analysis of a statistical dataset depends on several parameters: (i) the survey purposes, to be specified in terms of the research hypotheses and the population, (ii) the data formation environment, *i.e.* the survey and sample designs, the problems faced at the data collection and manipulation stages, (iii) the data measurement scale and the information on and/or the hypotheses about the distribution of variables.

Hierarchical decision trees showing the logical path which the informed researcher should go through have been drawn by many Authors (Andrews *et al.*, 1974; Harshbarger, 1971; Hays, 1973). The computer program STATREE is a personal computer guide for the selection of methods suitable for a directed analysis of data obtained in a given setting.

The system defines the statistical problem by iteratively acquiring from the user his/her aims and the information about the data at hand. The user is presented with a sequence of questions and a set of possible answers between which he/she is asked to choose. The system-user dialogue is assisted by an on-line statistical help accessible at any stage of the search.

The program can perform some analyses, provided the SAS statistical package is present on the user's computer. The user does not need to know how SAS works, nor its specific parameters. He need only to identify the dataset and the variables to process.

STATREE may be defined as an "expert system" because it fits many necessary attributes of a statistical expert system (Hand, 1985b), and in particular it allows the researcher to add new knowledge to the existing database.

2. The software program

The software program is composed of four components: the database containing the available statistical expertise, the structured query module to the database, the on-line help for statistical topics and technical terms, and the interface to the SAS data processing system.

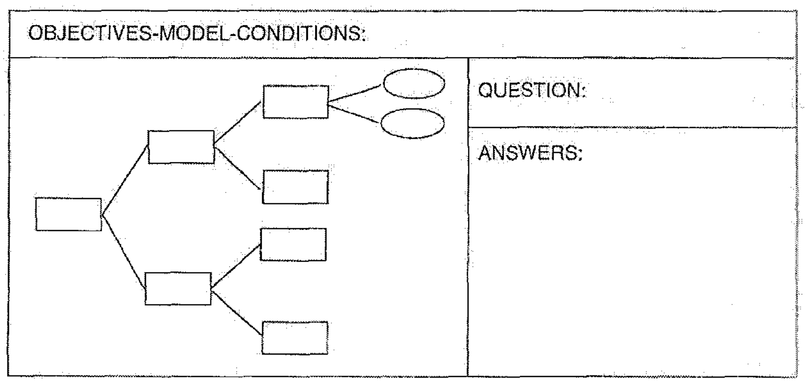
The knowledge of the system is based on a hierarchical network database, that is, a decision tree-like model (Fox, 1983). The database architecture shows some innovations with regards to previous implementations (Hand, 1985a) and allows for consistent updating of the meta-data base. Thus, the system can easily "learn" new techniques and new solutions to problems. The database creation and updating are enabled through an interactive fullscreen oriented managing module (Capiluppi, 1993).

STATREE is a program running on personal computers with Intel 80x86 processors and a DOS 5.0 operative system (Fabbris *et al.*, 1993). An 80286 personal computer with VGA graphic adapter and 1 MB of EMS is the minimal system configuration necessary to run STATREE. The program can also run in the Windows 3.1 environment. STATREE is written in C++ and SAS/SCL languages. The installation on hard disk needs about 1 MB of free space.

3. The query module

The interactive module for the selection of the suitable techniques shows a sequence of questions, in Italian, about the nature of the data and the scope of the analysis, and the user answers the questions by choosing one of the suggested items. The answer is selected using the arrow keys, <UP> and <DOWN>; the answer is then confirmed by hitting the <ENTER>, and the system goes to the next node of the decision tree. Hitting <ESC> the system goes back to the previous node.

The user interface of this module is made up of the following four windows.



The <F1> key shows the program help, that is the functions associated with the active function keys:

F1	*program help*
F2	context statistical help
F3	*index of the statistical help*
F4	references
F10	exit
ESC	back to previous node
↑ ↓	answer selection

4. The statistical help

The interactive selection process is provided with an on-line statistical help, supplying information about the technical terms and concepts in the questions. The help is organized into topics, from general to specific according to the tree structure, and implemented in help frames of variable length. Each question of the decision tree is associated with one such help frame by a keyword, isolating the argument pertinent to the context.

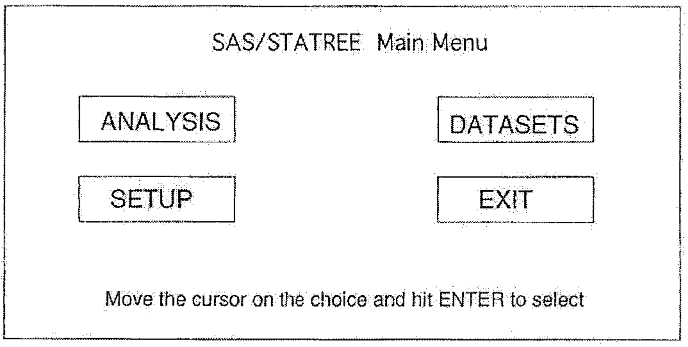
This context help is accessed by hitting the <F2> key, which opens a window on the screen displaying the help frame text. It is always possible to access any other topic through the general help index, by hitting the <F3> key, and then selecting the item of interest in the list.

The help text contains several references, which are available on-line in the system and accessible to the user through the <F4> key.

5. The interface to the SAS package

When a technique has been selected, the analysis can be performed by hitting the <ENTER> key. The analysis is carried out by the interfacing module to the SAS package. The user interface of this module is that of the SAS environment, and is based on dialogue windows and fields which allow the user an easy way to supply the necessary information to the system. Either the arrows or the function keys, or <TAB>, as well as a mouse, can be used to move through the windows fields, while <Fl> always opens the window help. The main menu displays the following four choices:

□ ANALYSIS	carry out the selected analysis;
□ DATASETS	build or modify the content of a dataset;
□ SETUP	set up the SAS interface module;
□ EXIT	return to the decision tree.



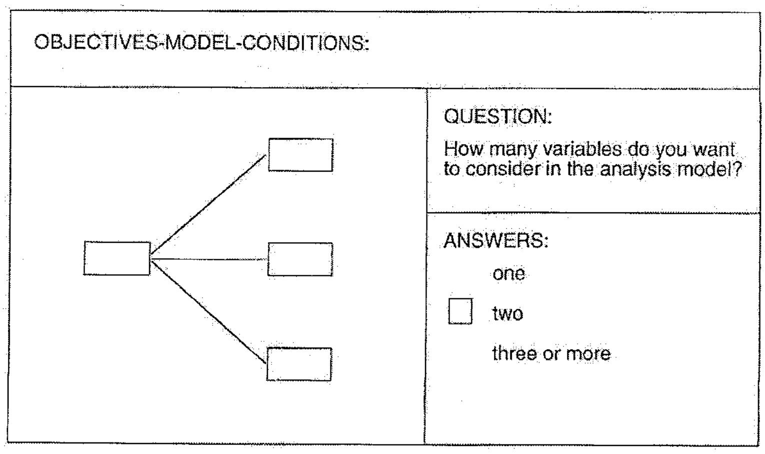
Move the cursor on the choice and hit ENTER to select

The module carries out the analysis by writing a suitable program (job) in the SAS language that is submitted to the SAS package. In this way, all computations are delegated to the SAS routines and procedures. The output is displayed in the OUTPUT window. The SAS program built by the system can be accessed by the user in the PREVIEW window.

6. A sample work session

Suppose we want to measure the relation between income and education of a set of subjects. The system acquires the problem description by asking the following questions:

Q1: How many variables do you want to consider in the analysis model? [A: two];



OBJECTIVES-MODEL-CONDITIONS:

QUESTION:

How many variables do you want to consider in the analysis model?

ANSWERS:

one

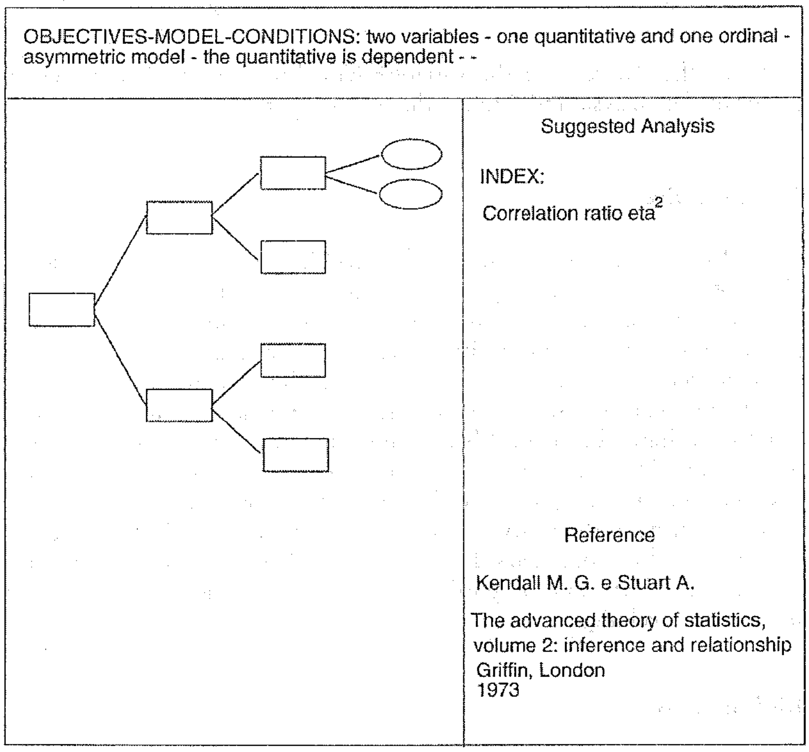
□ two

three or more

Q2: Which is the measurement scale of the variables? [A: one quantitative and one ordinal];
Q3: Is the model symmetric or asymmetric? [A: asymmetric];
Q4: Which one is the dependent variable? [A: the quantitative variable].

At this stage the system knows that the analysis is a bivariate asymmetric analysis with a quantitative dependent variable and an ordinal independent variable. Now it is possible to choose between two indices, the correlation ratio η^2 or the serial correlation ratio:

Q5: Do you want to consider the ordinal variable as a discrete class reduction of a normal variable? [A: no]
The suitable solution is then the correlation ratio η^2:



OBJECTIVES-MODEL-CONDITIONS: two variables - one quantitative and one ordinal - asymmetric model - the quantitative is dependent - -

Suggested Analysis

INDEX:

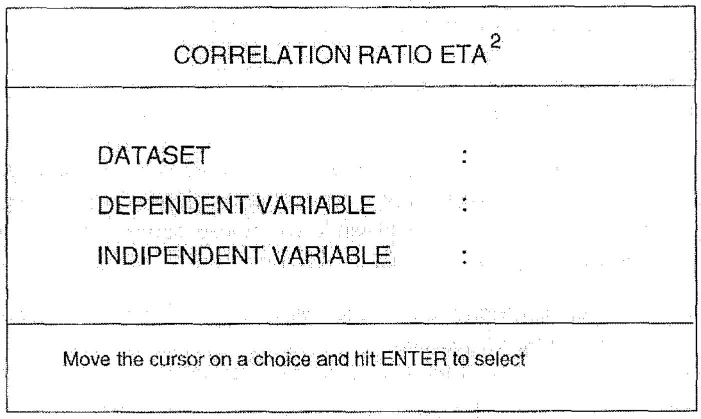
Correlation ratio eta^2

Reference

Kendall M. G. e Stuart A.

The advanced theory of statistics,
volume 2: inference and relationship
Griffin, London
1973

Hitting <ENTER> activates the SAS interface module and displays its main menu. Choosing ANALYSIS the system gets the necessary inputs to process the data:



CORRELATION RATIO ETA^2

DATASET :

DEPENDENT VARIABLE :

INDIPENDENT VARIABLE :

Move the cursor on a choice and hit ENTER to select

Once the dataset and the variables have been specified, the system performs the analysis and displays the final output.

7. Perspectives

STATREE has been designed for teaching purposes. The iterative *dialogue* between user and system to identify the method suitable for the research problem at hand develops the user's ability to screen the properties of the methods available. So the system may be used as a self-teaching tool, with an occasional consultation of the manual

The program may be used also by researchers who need to use statistics but have a limited statistical background. It represents an operative tool for performing data analysis without experience of mathematical packages, and avoiding misuse of inappropriate analysis techniques.

Up until now, the statistical analyses supported in STATREE are the univariate and bivariate ones. Later versions will embed in the current tree the analysis of rates and proportions in longitudinal studies, and the main multivariate techniques.

Bibliography

Andrews F.M., Klem L., Davidson T.M., O'Malley P.M. and Rodgers W. (1974), *A guide for selecting statistical techniques for analyzing social science data*, ISR The University of Michigan.

Capiluppi C. (1993), *Implementation of the knowledge database of STA-TREE, an expert system for the selection of the statistical method for data analysis,* Statistic '93, Wollongong, NSW, Australia.

Fabbris L., Capiluppi C., Giancotti G. and Meneghello A. (1993), *STA-TREE 1.0 manuale per l'uso,* Summa, Padova.

Fox J. (1983), *The theory and practice of expert systems,* British Computer Society Specialist Group on Expert Systems, Newsletter No. 8 (May), pp. 14-16.

Hand D.J. (1985a), Choice of statistical technique, in *Proceedings of the 45th Session of the International Statistical Institute,* Amsterdam.

Hand D.J. (1985b), Statistical expert systems: necessary attributes, *Journal of Applied Statistics,* 12(1).

Harshbarger T.R. (1971), *Introductory statistics: a decision map,* The Macmillan Company, New York.

Hays W.L. (1973), *Statistics for the Social Sciences,* Second Edition, Holt, Rinehart and Winston, New York.