STATISTICAL EXPERT SYSTEMS IN SAS® SOFTWARE

Pan-Yu Lai, The Upjohn Company

INTRODUCTION

Statisticians and non-statisticians currently have a fairly large number of reliable, powerful, and well-documented statistical computation packages to choose from and SAS system is one of them. It not only provides very extensive statistical analysis procedures with explicit outputs but also offers complete data management tools. Owing to its powerfulness and user-friendliness, SAS is becoming one of the most popular packages in both industry and academia. Anyone with some experiences in SAS and a set of manuals can use the system to analyze data and produce the final statistical report without knowing much about the statistical concepts in the procedures. But without thorough understanding of these procedures, there is great risk on inappropriate analysis and consequently wrong inferences. One of the alternatives would be seeking help from statistical consultants but it is costly and they are not always available. So it would be the best if we can incorporate statisticians' expertise into programs and develop a statistical expert system so that the users can be guided to an appropriate method for analysis and go to statistical consultants for advice only when it is necessary.

EXPERT SYSTEM OVERVIEW

Since World War II, computer scientists have tried to develop techniques that would allow computers to act more like human being. In particular, a collection of artificial intelligent(AI) techniques that enable computers to mimic an expert's thoughts to assist people in solving problems and making decisions, generally called knowledge-based expert systems, have been developed extensively and successfully in many areas such as medicine, chemistry, geology, and mining.

The British Computer Society's Specialist Group has proposed a formal definition as:

> An expert system is regarded as the embodiment within a computer of a knowledge-based component, from an expert skill, in such a form that the system can offer intelligent advice or take an intelligent decision about a processing function. A desirable additional characteristic, which many would consider fundamental, is the capability of the system, on demand, to justify its own line of reasoning in a manner directly intelligible to the enquirer. The style adopted to attain these characteristics is rule-based programming.

It says all of it but I like the shorter definition proposed by Max Bramer:

> An expert system is a computing system which embodies organized human knowledge concerning some specific area of expertise, sufficient to perform as a skillful and cost-effective consultant.

An expert system should consist of at least two major components, knowledge base and inference engine (also called control strategy or knowledge application system). The former is a collection of facts and rules from the problem domain and the latter is a mechanism which draws inferences and controls the reasoning processes over the knowledge base and the external inputs. These two should be kept separately for the reason of maintenance.

Therefore, when the knowledge base changes and grows over the time, we don't have to rewrite the system every time this happens. Besides, the system can be maintained by the domain experts or knowledge engineers instead of programmers. Sometimes people also consider the following two parts as primary components. The first one is the explanatory interface which enables the system to explain why the questions are asked and how the conclusions are drawn. The second one is the development engine (also known as knowledge acquisition subsystem). There are many ways to acquire knowledge such as interview analyses, text analyses, behavior analyses, machine induction, and simulated consultation. Although the problem-solving power of an expert system comes also from the knowledge it possesses, not just from the formalism and inference schemes it employs, we are not going into too much depth in knowledge acquisition subsystem because it is not the main issue in this article. The reader if interested can find more details in [1], [2], and [3].

STATISTICAL EXPERT SYSTEMS

Two-way communication is an important and necessary feature for an expert system. This means that the flow of information is of equal size in both directions. Therefore, a statistical expert system should also be able to provide users elaborated explanations on its reasoning and questioning strategies in addition to guiding users to the appropriate procedures for problem solving. This way users will have better senses about why those questions are asked and how the statistical procedures are selected and learn some concepts on statistics.

Since statistical consultation always involves some degree of data analysis at the end in most cases, it would be necessary that a statistical expert system has easy access to data and is able to extract as much information as possible directly from data to avoid unnecessary user interface. Meanwhile, in order to meet the needs of researchers in various fields, a statistical expert system should incorporate all kinds of scientific disciplines in various fields. But this is quite impossible at this early stage. So before it is done , the system may know more about statistics than the user but may not know as much as the user in a particular field. Therefore, users should be allowed to look up into the data anytime before answering a question prompted by the system. Furthermore, the structures of data sets can be various. It is quite impossible to educate the system to know them all. This further confirms the necessity of direct interfaces among users, data and the statistical expert system.

In summary, a statistical expert system should contain at least six components, the inference engine, the knowledge base, the help file, the interface, the data processing, and the working memory. Their connections are shown in Figure 1. The working memory is used to store the instances or dynamics of the knowledge of the target problem provided by the user and/or extracted from the data. It is invoked only when the system is in working. A knowledge base can be constructed in many ways such as decision tree, rules, frame, and semantic networks. They are all qualified for constructing a statistical knowledge base. The decision tree approach is used for an example later. Forward chaining and backward chaining are two of the most common ways for the inference engine. Forward chaining, also known as data driven or antecedent reasoning, starts with the initial conditions and searches forward through the knowledge base toward a solution. It works well for decision tree data base. Backward chaining, also known as goal driven or consequent reasoning, starts with a possible solution and works backward to search evidences to prove or disprove it. It only works fine when there are not many

solutions in the knowledge base. In most cases, these two reasoning processes are combined to reach a better performance. References [4] and [5] give more details about various types of knowledge base and inference engine.

APPROACHES AND TOOLS IN SAS SOFTWARE

Knowledge Base

Portier and Lai [6] had developed a prototype statistical expert system which performs a binary tree search to aid users identifying an appropriate analysis for the data in hand. Their programs were written in FORTRAN 77 and implemented on a DEC/PDP 11/23 microcomputer. A part of their decision tree given in Figure 2 will be used here as an example. The questions need to be answered at each node are given in Table 1. There are many ways to convert this decision tree into a knowledge base using SAS software. The basic idea here is to convert it into a SAS data set as shown in Table 2. Q_NO indicates the node number. Q_TEXT stores the action to be taken. TYP indicates the type of this action which is either to run a subprogram or to ask a question. Y_GO and D_Y tell the system the next node and decision tree,respectively, to go when the user answers 'yes' at this node. N_GO and D_N, U_GO and D_U, and H_GO and D_H are to tell the system where to go when the user answers 'NO', 'UNKNOWN', or 'HELP',respectively. For instance, at node one Q_TEXT tells us that the action is to run a subprogram named NORMAL.PRG to check the normality of a particular variable. If NORMAL.PRG returns 'Y' to the system, then it will execute node #2 of the decision tree specified by D_Y. If the return code is 'H' (for help), then the system will go to node #1 of the help file indicated by D_H to display explanations for the user. Since the structure of the decision tree may be changed from time to time and so is the length of the text for a node, the observation number of a node may not be consistent all the time. It is crucial to store the node numbers instead of the observation numbers in the pointers (Y_GO, N_GO, ...). Also, in order for the system to locate the target node quickly, it is necessary to create an index data set for the decision tree data set as shown in Figure 3. For example, if the return code of node #2 is 'NO' (see Table 2), it implies that the next move is go to node #5 of the decision tree specified by D_N because the value of N_GO is 5. The system will directly access to the fifth observation of the index file where the LINKER says 6. This tells the system that node #5 start with the 6th observation of the decision tree data set and this observation is then accessed directly. This kind of structure allows the whole knowledge base to be partitioned into several parts and each part can be updated and modified by different groups of domain experts and the knowledge engineers easily and separately.

Help File

All the explanations and reasoning can be stored in SAS data sets with structures similar to the knowledge base but simpler. In fact, only Q_NO and Q_TEXT are necessary. Again, an index file is helpful for the maintenance and may speed up the process. Figure 4 gives a glimpse of such help file.

Inference Engine

With the well-structured knowledge base and help file mentioned above, the major mechanism of the inference engine is solely to retrieve a particular observation from a SAS data set and display the contents of this observation to the user and then decide on the next move according to the user's response. This process is repeated over and over again until the goal is reached or the user quits. This task can be accomplished by a SAS/AF PROGRAM entry. A PROGRAM entry contains two major components, the display panel and the source panel. The display panel is what the user sees on the screen when the application is run in AF execution mode. It is used to display questions and collect the answers from the user. The source panel contains three sections. They are INIT, MAIN, and TERM.

The INIT section is executed when the user invokes the PROGRAM entry. It is used for initiation such as opening data sets and assigning the starting point. The MAIN section is executed whenever the user presses the ENTER key or a LINK MAIN or GOTO MAIN statement is encountered. It is used to go through the knowledge base and link data processing procedures and the help file. The TERM section is executed only once after the control is switched from the MAIN section. It is used to close all data sets opened previously and summarize the consultation session for the user. Figure 5 gives a brief flow chart of an inference engine using SAS/AF.

Data Processing

Before performing the final analyses, the statistical expert system should be able to gather information from the target data sets for the purpose of decision making. By invoking the abundant data management tools, statistical procedures, and graphic procedures provided by SAS system through SAS/AF, data processing has become a straight forward job in building an statistical expert system.

Interface

From previous explanations a picture of interface among users, data processing procedures, help file, and the inference engine has been outlined. In summary, the display panel of the PROGRAM entry in SAS/AF is used to present questions and explanations to the user, to pass the user's response to the inference engine, and to receive the user's SAS statements which will be submitted to the SAS system for execution later by SUBMIT and ENDSUBMIT statements. The information obtained from data processing is passed to the inference engine and the user through the output data sets generated by the data processing procedures. If any particular procedure doesn't produce output data set, PROC PRINTTO can be used to generate a SAS data set from the listing file.

Working Memory

A working memory is a temporary space to store all the information gathered in the whole session such as goals specified by the user, variable names, their characteristics, their relationships, and the path of the whole consultation period in the forms of SAS data sets or SAS/AF macro variables.

CONCLUSION

The quality and the speed are equally important in judging an expert system. Although I do believe that a team of outstanding statisticians, well-trained knowledge engineers, and comprehensive SAS programmers can build a high quality statistical expert system, the slow speed of SAS software in PC is making it less valuable. For example, producing graphs is a very important feature in a statistical expert system to help users to examine and understand the data and the analysis results but it's low speed in SAS PC is making it difficult to build a very comprehensive system because the most accurate diagnosis is worthless if the patient dies.
Therefore, the SAS system is only recommended to build a small scale or less-data-processing statistical expert system before the speed is improved.

REFERENCES

[1]. Graham,I. and Jones, P.L. (1988). "Expert Systems : Knowledge, Uncertainty, and Decision". Chapman and Hall.

[2]. Cleal, D.M. and Heaton, N.O. (1988). "Knowledge-Based Systems: Implications for Human-Computer Interfaces". Halsted Press.

[3]. Marcus, S.M. (1988). "Automating Knowledge Acquisition for Expert Systems". Kluwer Academic Publishers.

[4]. Wolfgram, D.D., Dear, T.J., and Galbraith, G.S. (1987). "Expert System for the Technical Professional". John Wiley & Sons, Inc.

[5]. Harmon,P. and King, D. (1985). "Expert Systems : Artificial Intelligence in Business". John Wiley & Sons, Inc.

[6]. Portier, K.M. and Lai, P.Y. (1983). "A Statistical Expert System for Analysis Determination". Proceeding of the American Statistical Association, Statistical Computing Section.

Fig. 1

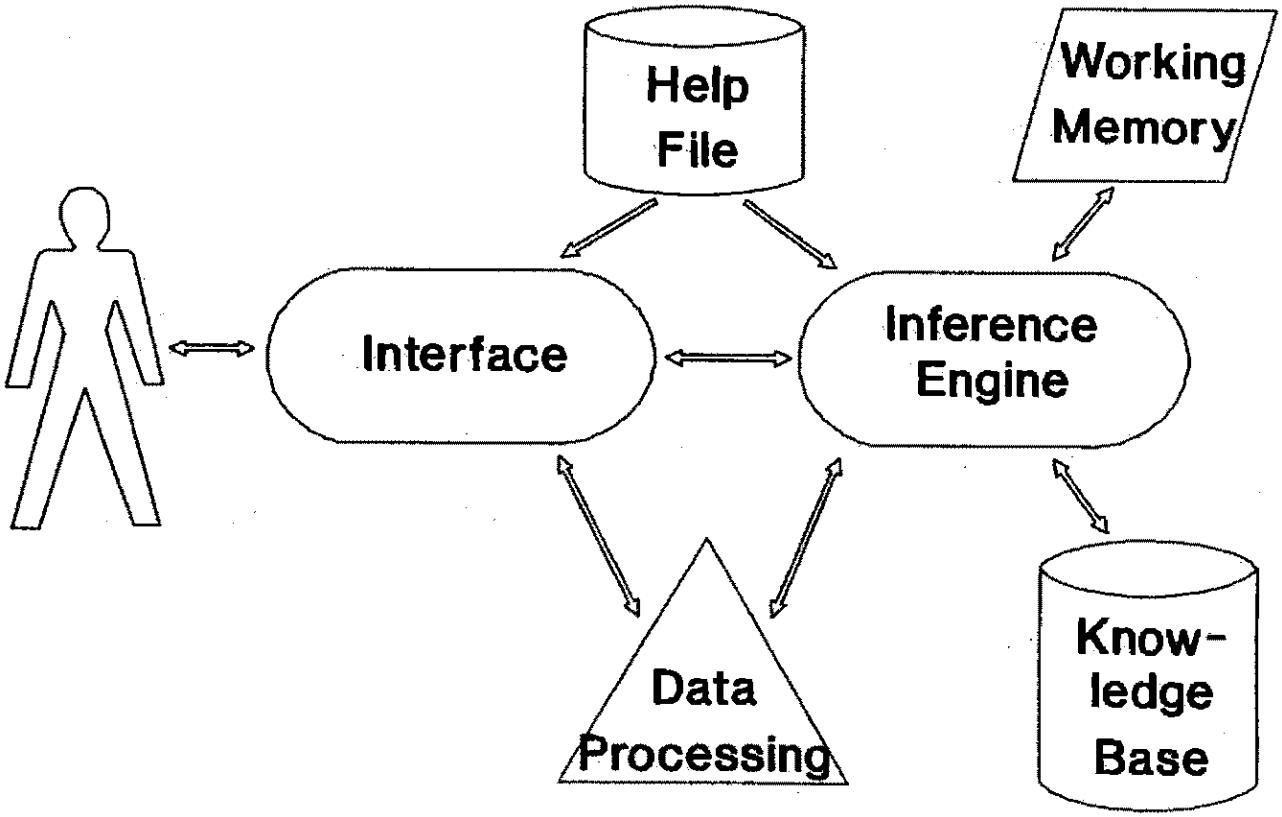


Fig. 2

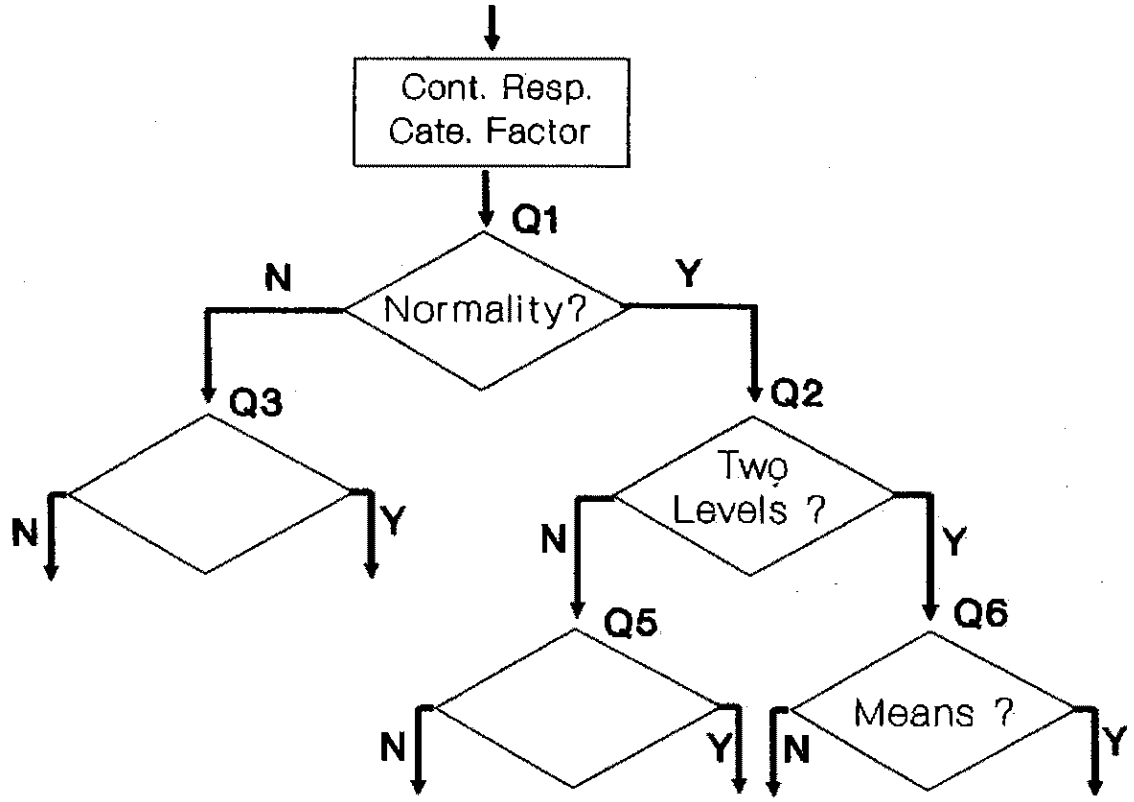


Table 1

Q1. According to your responses, we know that you have the continuous response variable and the categorical factor variable.
Do you think it is fair to assume that response variable is normally distributed ?

Q2. Are there more than two levels (categories) in the factor variable ?

Q3. Since the normality assumption does not hold, nonparametric procedures will be considered. Are there more than two categories in the factor variable ?

Q6. Are you interested in testing a hypothesis that the means of the response variables for the two levels of the factor variable are different from each other ?

Fig. 3

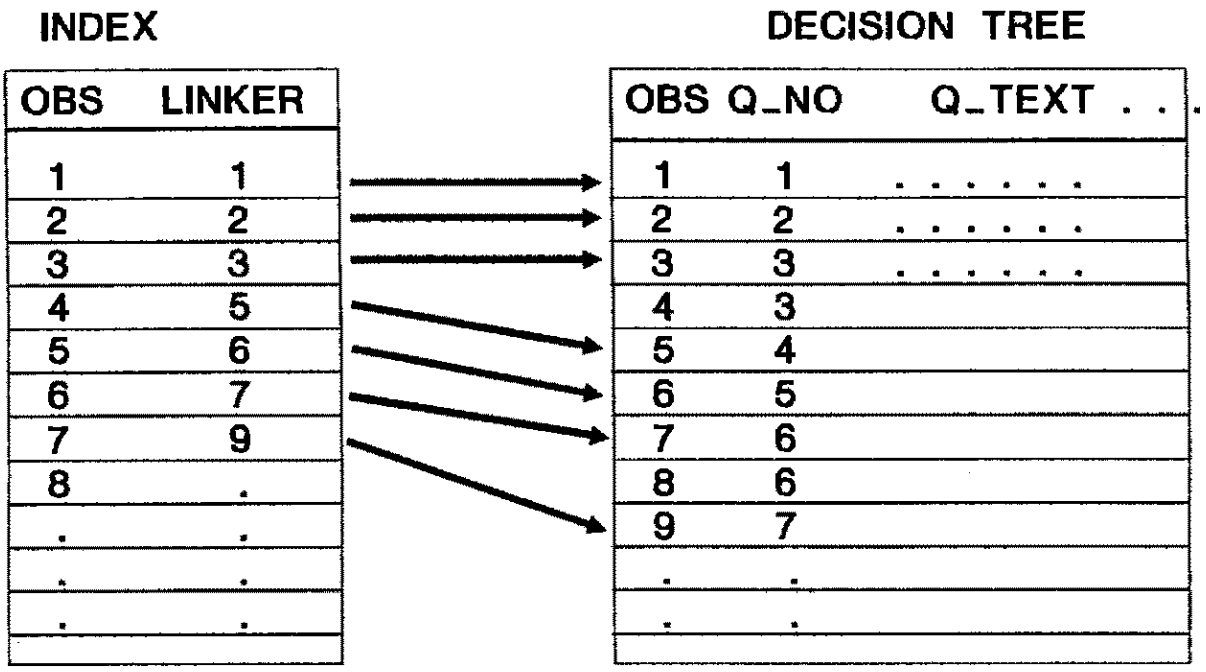


Table 2

OBS	Q_NO	Q_TEXT	TYP	Y_GO	D_Y	N_GO	D_N	U_GO	D_U	H_GO	D_H
1	1	NORMAL.PRG	A	2	1	3	1	1	2	1	3
2	2	LEVEL.PRG	A	6	1	5	1	0	2	0	3
3	3	...	Q	4	1	7	1	2	2	2	3
4	3	...	Q	4	1	7	1	2	2	2	3
5	4	...	Q	9	1	10	1	3	2	3	3
6	5	...	Q	12	1	13	1	4	2	4	3
7	6	Are you interested in ...	Q	10	1	11	1	5	2	5	3
8	6	Are means different ?	Q	10	1	11	1	5	2	5	3
.	.	...	.	.	.	.	.	.	.	.	.
.	.	...	.	.	.	.	.	.	.	.	.
20	10	T_TEST.PRG	A	0	0	0	0	0	0	0	0
.	.	...	.	.	.	.	.	.	.	.	.
.	.	...	.	.	.	.	.	.	.	.	.

Fig. 4

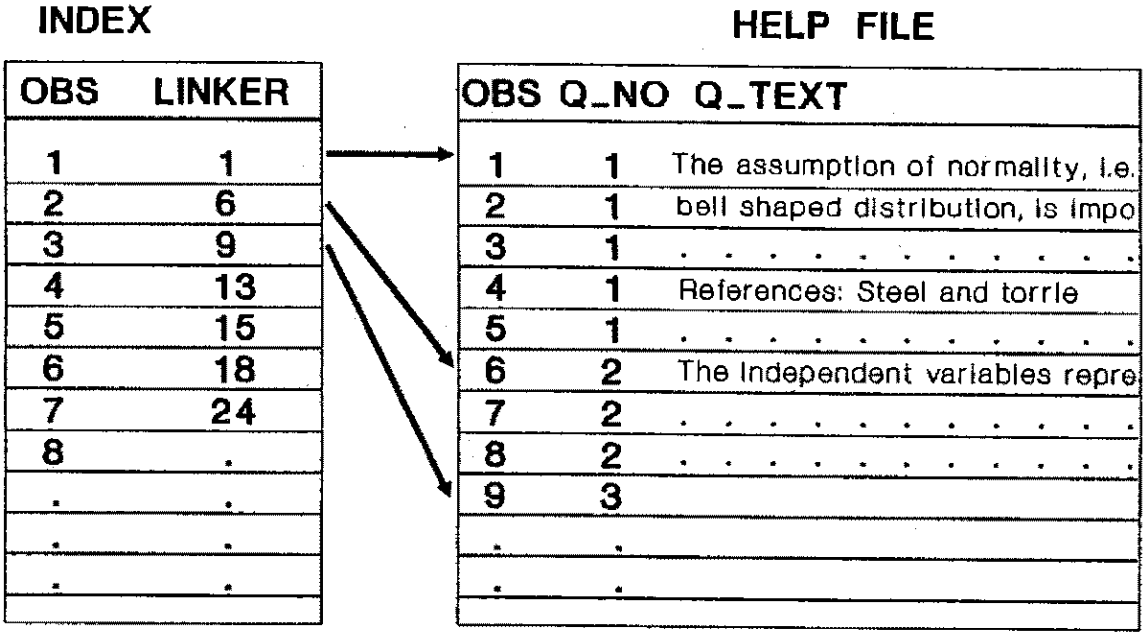
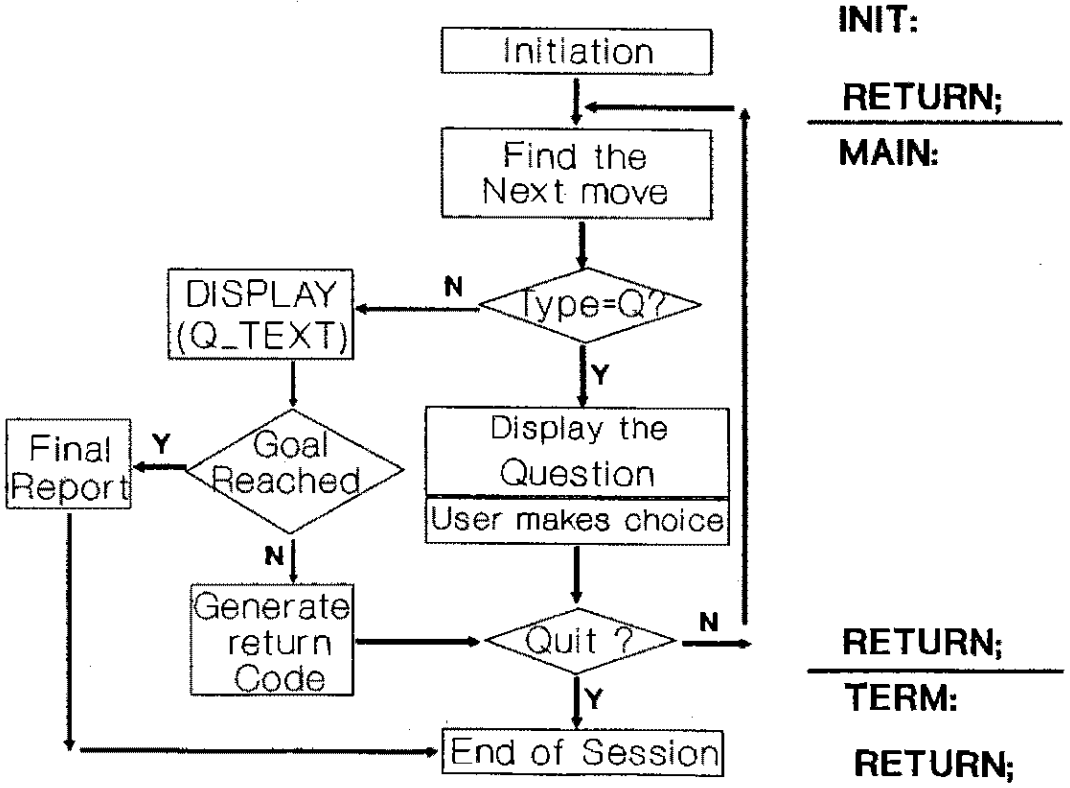

INDEX
HELP FILE
OBS LINKER
1 1
2 6
3 9
4 13
5 15
6 18
7 24
8 .
OBS Q_NO Q_TEXT
1 1 The assumption of normality, i.e.
2 1 bell shaped distribution, is impo
3 1 . . . . . .
4 1 References: Steel and torrie
5 1 . . . . . .
6 2 The independent variables repre
7 2 . . . . . .
8 2 . . . . . .
9 3

Fig. 5


Initiation
Find the Next move
Type=Q?
N
Y
DISPLAY (Q_TEXT)
Display the Question
User makes choice
Goal Reached
Y
N
Final Report
Generate return Code
Quit ?
N
Y
End of Session
INIT:
RETURN;
MAIN:
RETURN;
TERM:
RETURN;