A THESIS

entitled

The Development of an Expert Systems Approach to the Statistical Analysis of Experimental Data

Submitted in partial fulfillment of the
requirements for the award of the

DEGREE OF DOCTOR OF PHILOSOPHY

of the

COUNCIL FOR NATIONAL ACADEMIC AWARDS

by

EDWINA E BELL

Faculty of Technology
School of Mathematics, Statistics and Computing
Thames Polytechnic
LONDON

October 1988

ABSTRACT

This thesis is concerned with the application of expert systems techniques in the field of statistics. An expert statistician in industry has a twofold role; undertaking the design and analysis of data from complex experiments and providing supervision and help for research workers who analyse data from simpler designs. There is, therefore, a potential role for a statistical expert system which could be used by research workers to enable them to carry out valid analyses. The expert statistician would be freed from the more straightforward analyses and would only need to deal with referrals from the system and to initially 'tune' the system to their own application area. The design and development of such a prototype expert system, THESEUS, is the basis of this work.

The area of application chosen for the prototype system is completely randomised designs with one trial factor. It was initially important to limit the area of study so that knowledge acquisition for the system would be a manageable task. However, once the difficulties in developing an expert system have been tackled, much of the expertise used in analysing this simple type of study could be readily extended to more complex designs.

The knowledge acquisition phase, the most time consuming part of developing any expert system, concentrated on developing a rational prototype rule base by reviewing the available literature, interviewing practising statisticians and undertaking workshops where the analysis of particular data sets was discussed.

The prototype software is a production rule system and is written in Turbo Pascal on an IBM-AT. Pascal was chosen because of the need to access statistical routines during the consultation process. The prototype uses a combination of forward and backward chaining to process the rules. Information required by the system can come from the user, the data or the rules.

The overall system design also includes facilities for entering and editing data, altering and adding knowledge and a report generator. Implementation of these facilities is not incorporated as part of this thesis.

A small number of trial sites were selected for industrial trials in order to validate the system and evaluate the results of the local experts 'tuning' of the rule base to their own particular application area.

Acknowledgements

I would like to express my thanks and appreciation to my supervisors, Peter Watts and John Alexander, whose interest and encouragement throughout the project has been invaluable to me. I am particularly grateful to John Alexander whose enthusiasm for the project initiated my own interest in the development of Statistical Expert Systems.

I would also like to express my appreciation to the undergraduate students who undertook development projects related to this research. In particular, I would like to thank Chris Richardson for his work on the Rulebase Editor and Yiannis Troullides for his work on the graphics and routine interface.

The Statistics Research Group at Thames Polytechnic also made significant contributions towards this research through holding data analysis workshops and testing the software, I am very grateful for their input.

Finally, I would like to thank my family and friends who have supported and encouraged me over the last three years.

CONTENTS

Appendices

Chapter One

Introduction

In this introductory chapter the nature of statistical practice and the problems with existing statistical packages are considered. The concept of expert systems and their potential application to statistics is discussed. The results of a postal survey undertaken in order to obtain some feedback from statisticians in industry on the possible role of expert systems are also presented. Finally, the governing criteria for the research project presented in this thesis are discussed.

1.1 Project Aims

The primary aim of this research was the design and development of a Statistical Expert System that could be used by research workers who are not statisticians but who regularly need to carry out statistical analyses. A further aim of the project was to develop a system in which the expertise contained in the system could be easily modified by a 'local expert statistician'.

These aims required research into a number of different areas; from expert systems technology and knowledge acquisition to the problems of formalising statistical strategy and expertise. The main areas of research pursued in this project are :-

- The development of a knowledge structure and a control mechanism for the system which would be appropriate to statistical analysis.

- The selection and application of knowledge acquisition methods in a targetted area of statistics.

- The development of a prototype system capable of providing help and strategical advice in the analysis of completely randomised designs.

1.2 Statistical Practice and Problems

Statistical consultation is a complex and highly skilled undertaking requiring expertise in communication, analysis and interpretation. In this section we discuss the nature of statistical consultancy and the problems that can occur. The question of statistics being undertaken by non-statisticians is also considered.

It is helpful to consider the work of statisticians in terms of the activities they undertake. A statistician will need to understand and possibly refine the objectives of the research; inspect and possibly modify the data (e.g. by transformation); select and apply appropriate methods and interpret the results (Hand 1986a, Huber 1985, Haux 1985). These activities cannot be expressed as a step-wise progression as statistical practice is an iterative process. For example, it may be necessary to modify the questions or objectives of the research in the light of the statistical methods available or the application of a method may indicate a need to modify the data by transformation. There are potential problems in each of these aspects of statistical consultancy. The most obvious one being lack of understanding between client and consultant. Good communication is essential in data analysis; client and consultant must be able to understand each other's language (Jones 1980).

A skilled statistical consultant is a highly trained and rare resource. A current problem is that with increased access to powerful computers and statistical packages more experimental data is being collected because of the potential for analysis. More seriously, a greater amount of analysis is being undertaken by people who are not statisticians and who have an inadequate grasp of the limitations and suitability of the techniques they are applying (Hand 1986b). There are, quite simply, not enough statisticians to go round.

1.3 Expert systems

There are almost as many definitions of what an expert system is as there are expert systems, for example

> "An expert system is a knowledge-based system that emulates expert thought to solve significant problems in a particular domain of expertise"
> (Sell 1985)

> "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."
> (British Computer Society's Committee of the Specialist Group on Expert Systems, February 1983)

In general terms, an expert system can be viewed as a system which supplies expertise in such a way that a non-expert using the system can arrive at decisions similar to those of an expert.

1.3.1 Historical Overview

The original motivation for the development of computers was to speed up calculation and processing especially for tedious or repetitive tasks. The emphasis was on speed and the most economical use of the machine and the computer was limited to handling numerical tasks or processing 'hard and fast' facts. At the same time as developments to improve the speed and efficiency of computers there has been a growing interest in programming computers to handle more difficult tasks; to process uncertain facts, to make 'reasoned' decisions as opposed to using a brute force approach or where such a brute force approach would not lead to a solution. For example game playing, especially chess, or diagnosis problems where human 'experts' apply their knowledge in terms of heuristics.

The development of the system DENDRAL in the sixties marked the beginning of 'expert systems'. Originally, DENDRAL was designed to enumerate all possible configurations of a set of atoms observing the rules of chemical valence; the aim being to hypothesise on the possible molecular structure of a compound. Extensions to DENDRAL included reducing the set of possible outcomes to a set of likely ones using heuristics or rules based on chemical facts. A description of the development of Dendral is given in a book by Lindsay, Buchanan, Feigenbaum & Lederberg (Lindsey et al 1980).

Other early expert systems included MYCIN and PROSPECTOR. MYCIN was designed to help the physician to diagnose and prescribe for bacterial infections of the blood (Shortliffe 1976). PROSPECTOR was developed to aid the geologist to assess sites for possible deposits. The development of these systems served to illustrate the potential usefulness of expert systems in solving difficult real-world problems (Duda et al 1979). It was the early eighties before any information about research into expert systems for statistics was published.

1.3.2 The Nature and Structure of Expert Systems

An expert system requires a knowledge base, methods of inference and a control mechanism. The knowledge base contains the knowledge about the domain, or area of expertise, usually expressed in terms of facts, heuristics and rules. Methods of inference are necessary to allow the system to make reasoned decisions based on the information available and using the knowledge in the knowledge base. The control mechanism organises the application of the inference methods. Within this context, a reasoned decision is one with which the expert would agree and should have been reached by only considering relevant information and doing so in a logical order.

A major distinction between conventional software and expert systems is that expert systems are process oriented rather than results oriented, the way in which a decision is reached is just as important as the decision itself.

Areas of application

Expert systems are potentially applicable in a wide range of areas, some of which are described in the next section. They are particularly useful where experts are in short supply or where a common form of expertise is required by many. Expert systems can be applied in relatively straightforward areas, where the necessary expertise is not too extensive but is required by many people; for example, a system to give advice on the availability of different loan schemes. Knowledge about an area such as loans is usually 'available' but poorly distributed. The development of an expert system in this area would mean that the information would be drawn together into a single system which can then be made available to many users. Expert Systems may also be applicable in more complex problem areas of expertise where experts exist but are in short supply. For example, process control for an aluminium reduction process where expert knowledge is required to know what information is relevant, what information to request and to reach a decision and act accordingly.

1.3.3 Present Research

ACE is an example of a trouble shooting system designed to aid the manager of a telephone network centre who is responsible for maintenance and trouble shooting (Rauch-Hindin 1988 p293). There is an enormous amount of information available and highly trained specialists are required to identify trouble spots. ACE works through the information available in a data base, using the rules in the knowledge base and presents a report of potential trouble spots and recommended actions for the maintenance engineers.

Expert systems have also begun to appear in the financial sector (Rauch-Hindin 1988 p302). The system ExMarine, developed for Coopers & Lybrand underwriters, collects information about applicants and their insurance brokers, underwrites the risk, and suggests a premium. The system was built using a knowledge acquisition tool, FFAST, and an expert system tool, ART. ExMarine uses both rules and frames to store knowledge.

An example of an expert system in the area of databases is Quist (Rauch-Hindin 1988 p333). The knowledge system generates database access strategies based on knowledge of the database content and general heuristic knowledge about items contained in the database.

Process control is one of the largest growth areas for the development of expert systems. An example of this is the development of a system for automating the control of the kilning stage in the manufacture of cement (Haspel & Taunton 1986). The system uses rules expressed in linguistic terms that can be easily expressed and understood by experienced operators. The system G2 (Rauch-Hindin 1988 p349) has been developed as a tool for building such systems and incorporates a knowledge analysis program and a real-time communications-gateway module. The knowledge-based component receives data from the gateway program, reasons about the data, and offers advice about critical process-control points of interest, multiple alarms, and diagnosis of trouble spots.

Research has continued in the area of medical diagnosis. PUFF (Aikins et al 1984), is a system designed to interpret respiratory tests. Interpretation and diagnosis is based on historic and symptomatic information as well as the test data. GLADYS (Spiegelhalter & Knill-Jones 1984) is a medical diagnosis system for gastroenterology. This system uses information on clinical symptoms, collected by computer interview, to arrive at a probabalistic diagnosis, suitable treatment is then suggested.

1.4 Statistical Expert Systems

The development of interactive statistical software incorporating statistical expertise could help to relieve the professional statistician of the more routine enquiries and also protect the non statistician from inappropriate application of statistical methods. Research undertaken in the area of Statistical Expert Systems is reviewed in Chapter 2. In this section the issues raised by the introduction of Statistical Expert Systems and the requirements of such systems are discussed.

1.4.1 Current Statistical Software

The move towards more 'user-friendly' software and the advent of powerful desk-top micro-computers has meant that general purpose statistical packages are now available to a wide range of users, statistician and non-statistician alike. The current software supplies numerical or algorithmic expertise in a form that is, generally, easy to access and use. It is the responsibility of the user to decide on an appropriate analysis and to interpret the results.

Undertaking a statistical analysis involves determining the questions of interest to the client, selecting an appropriate form of analysis and ensuring that the necessary conditions and assumptions are met. Once an analysis has been carried out, the results need to be interpreted and related back to the original questions. The statistical software currently available can only help with the mechanics of the analysis and not the strategy. The misuse or even abuse of statistical methods is inevitable when such software is readily available to non-statisticians.

Chambers (1981a) in one of the early papers discussing Statistical Expert Systems states :

> 'Statistical software in its present form, made widely available by cheap computing, will precipitate much uninformed, unguided and simply incorrect data analysis. We are obliged to do something to help.'

Hahn (1984) states :

> 'Thus, capabilities for statistical number crunching are no longer limited to a knowledgeable elite, but are readily accessible to those with only limited training in statistics, and, consequently little understanding of the appropriate analyses to perform in a given situation and how to interpret the results.'

1.4.2 The Role of Statistical Expert Systems

The overall aim of Statistical Expert Systems is to incorporate knowledge about statistical strategy into a system, thus supplying users with expertise on both the strategy and the number-crunching aspects of the analysis. There are potential benefits for both the professional statistician and the non statistician.

a) The professional statistician could be relieved of some of the more routine enquiries and thus be able to give greater time to the more difficult tasks.

b) The non-statistician would be protected to a large degree from the inappropriate application of methods and the misinterpretation of results, without needing to have the relevant statistical expertise.

c) The provision of Statistical Expert Systems could also provide an important means of education for non-statisticians. As they follow the systems working and look at the reasons for decisions made, they may, consciously or sub-consciously learn more about statistical analysis. Education need not be limited to non-statisticians. Statisticians themselves may learn by using systems which are expert in areas with which they are not familiar.

d) The development of Statistical Expert Systems will necessitate the thinking through and coding of statistical strategy. Many statisticians employ their own particular strategy and yet are unable to express the reasoning behind the strategy explicitly. There is not necessarily a single correct strategy but by exploring and refining different strategies a clearer understanding of the common aspects of strategy should be gained (Pregibon 1986a).

1.4.3 Requirements of Statistical Expert Systems

Incorporating expertise into statistical software is a complex undertaking which involves the problems associated with developing expert systems in general and problems directly related to applying expert systems methods to statistical analysis. The development of an expert system requires decisions about the form of knowledge representation and the method of inference in addition to the well documented problems of knowledge acquisition. When applying expert systems methods to the area of statistics there are two further important considerations. Information required by the system to make decisions can come from the data as well as the user, thus it is important that the system should be able to access the data during the consultation. The other consideration is related to the problems of knowledge acquisition which is further hampered by the need to formalise statistical strategy in a way that can be expressed within the system.

The issues involved in knowledge acquisition are considered in detail in Chapter 5 and the design requirements for Statistical Expert Systems are discussed in Chapter 3.

A number of authors have agreed that the best way forward for research into Statistical Expert Systems is the development of small-scale systems in specific and well defined areas (Nelder 1984, Tukey 1986, Hahn 1985).

1.5 Prototype system

The aim of this project was to design a Statistical Expert System and develop a prototype system which could be tested in industry. The prototype system, called THESEUS, would provide a rulebase to cover a specific area of statistics and the inference engine necessary to process the rule base. The development of such a system requires the design and implementation of knowledge structures, the inference engine and the user interface. The area of expertise was to be large enough to give a realistic insight into the problems of knowledge acquisition and small enough to allow sufficient consideration to all the aspects of system development. Testing the prototype system in an industrial setting should enable us to assess both the advantages and problems of the different aspects of Statistical Expert Systems development. This assessment process was considered to be very important as it moves the research from being a purely academic exercise to the real world of statistical practice.

1.6 Industrial Review

A document outlining the potential role of intelligent software in statistics (see Appendix I) was sent to a number of statisticians in order to obtain some feedback on the potential for statistical expert systems and to pinpoint suitable application areas.

1.6.1 Format of the Postal Survey

The document was divided into three sections covering the present problems in statistics arising out of the wide availability of powerful statistical packages, the potential role of software which incorporated expertise and finally the general features of such a system.

The document was sent to 57 statisticians who are working in the pharmaceutical industry or research establishments. The list of statisticians was established by a combination of those known by personal contact with members of the Statistics Research Group and by looking through the Royal Statistical Society List of Fellows. Our primary interest was to contact statisticians involved in the analysis of scientific experiments rather than social surveys or official statistics.

1.6.2 Response

Replies were received from 31 of the 57 statisticians and, as anticipated, there was a wide range of opinions. In order to give some impression of the overall response the replies were categorised as follows :

A) Positive	[10 replies]
B) Negative	[5 replies]
C) Cautious or Unsure	[11 replies]
D) Non-committal	[5 replies]

Where quotes have been made from the replies received some indication of the nature of the respondents area of work is given.

The majority of respondents agreed that the misuse and abuse of statistical methods by non-statisticians is a serious problem. For example :

"...strongly endorse your concern about the use of sophisticated statistical software by non-statisticians." (Clinical Research Centre)

"There is a growing demand for skilled statistical analysis throughout industry, commerce and research establishments. Unfortunately there are too many non-statisticians analysing data inappropriately" (Government Research Institute)

However a cautionary note was given by one respondent

"There is as much danger in non-statisticians being over worried by the assumptions of statistical tests as by the misuse of methods, evidenced by letters to the BMJ etc about authors not vigorously testing every variable for non-normality. I fear that 'expert' systems would only encourage this unprofitable approach." (Department of Community Medicine)

Two of the respondents were in the fortunate position of having sufficient statistical resources to deal with all statistical analyses undertaken in their company or department.

Response to the proposal that a statistical expert system could be used both to relieve the statistician of more routine tasks and to protect the non-statisticians from the inappropriate use of statistical techniques was rather more varied. Some respondents were very enthusiastic seeing expert systems as the best way forward. The majority were cautiously optimistic, being aware of some of the possible problems; for example :

"A truly expert system should encapsulate the expert's approach for prescription of the appropriate tools to the end user and when developed and implemented the system should be capable of training the user nearly to the standard of the expert himself. Such a system would require enormous effort; moreover, the size and complexity of the system may not be of much help to strengthen the users motivation...but to begin with a system with simple alternatives should not be unwelcome by most users." (British Telecom)

There was a consensus of opinion that a general statistical expert system would be too complex and ambitious a task at the moment; this agrees with Hahn (1985) who advocates the development of specialised intelligent software.

Several respondents expressed a concern that an expert system could be regarded as a substitute statistician and that this should be avoided at all costs; for example :

"We as pharmaceutical statisticians involved in the analysis of clinical trials, cannot think of many situations where the use of statistics is routine. We have found from our experience and often to our dismay that what originally appears to be a very routine analysis can in fact be much more complicated. ... In situations where there is no access to a statistician, the type of package you are proposing could possibly be of some use, but should not be regarded as a substitute for a statistician. " (Pharmaceutical Company)

1.7 Scope and Application Area for a Prototype System

The main concern of this project is to provide a research worker, who is not a statistician, with the facility to analyse experimental data, offering protection against abuse or misuse of statistical methods.

1.7.1 The End User

The principal end-users of the system have already been defined as the research workers who, though expert in their own particular fields, are not statistically trained. The growing demand for statistical analysis throughout industry and commerce, coupled with increasing sophistication and availability of statistical software leaves statisticians with the ever increasing problem of providing an adequate service and monitoring the use of statistical methods by non-statisticians in their organisation. The possibility of introducing 'intelligent' statistical applications packages is considered as a means of filling the gap and relieving the statistician of some of the more routine work.

1.7.2 Application Area

The other major issue was the choice of application area for the prototype system. As observed above, Hahn (1985) stated that incorporating expertise in a general statistical package is a very large problem and that the best way forward is the development of specialised intelligent software. This was echoed by some of the respondents to the postal survey, for example, British Telecom.

It was important to choose an area that would be of practical use to research workers in industry. At the same time it was also important to select an area small enough for the knowledge acquisition and construction of the system to be a manageable task.

The area chosen was the Analysis of Completely Randomised Experiments with One Trial Factor. Data from experiments of this type are regularly analysed by research workers without statistical help. This area is small and well contained; in addition, much of the expertise used in analysing this simple type of study will readily extend to more complex designs.

1.7.3 Structure of the Thesis

Chapter 2 contains a review of work in the area of Statistical Expert Systems which provided some guidelines on the necessary design criteria. The logical design and structure of the system are described in Chapters 3 and 4. Chapter 5 discusses some of the possible approaches to knowledge acquisition and the methods used in this project.

The next two chapters contain the technical information that was necessary for the development of the prototype knowledge base. Chapter 6 provides an introduction to the concepts involved in hypothesis testing about means and the importance of Normal Theory assumptions; much of the information in this chapter will be relevant in other areas of statistics. Chapter 7 contains more specific information about statistical procedures where there are one, two or several samples to be compared.

Having dealt with the design, structure and knowledge acquisition for the system, Chapter 8 goes on to discuss the development of the system; this chapter also gives examples of the system during a consultation. Chapter 9 deals with the evaluation of the prototype system both within the Statistics Research Group and the evaluation trials in industry; some recommendations for improvements to the prototype system are also given here. In Chapter 10 an assessment of the project is given and areas for future research are identified.

Chapter Two

A Review of Statistical Expert Systems

2.1 Introduction

At the same time that expert systems were being developed in areas outside of statistics in the late sixties and early seventies, the rapidly increasing number and availability of statistical packages gave rise to much concern about the misuse or abuse of statistical procedures.

The concept of statistical expert systems provided a potential solution to these problems. The first statistical expert systems began to appear in the early eighties. This chapter provides a review of some of the research undertaken in statistical expert systems.

2.2 Early Days : 1981 - 1984

One of the first statistical systems to incorporate expert systems techniques was the RX project (Blum 1984). The aim of this project was to design and perform statistical analyses in medicine to establish causal relationships from a large time-oriented clinical data base. The statistical knowledge in RX took the form of a 'robot' statistician which simply applies all the methods it knows in order to try to find evidence of causal relationships.

An initial experiment in building an expert system for data analysis was undertaken at Bell Labs, based on a production rule architecture (Chambers, Pregibon and Zayas 1981) i.e. the knowledge was expressed in terms of

IF condition THEN action

rules. This system interfaced with the package S, providing diagnostic tests to assess the analysis under consideration. Chambers proposed some general design criteria for a statistical expert system, most importantly that the system should aim for a dialogue between client and software and not aim at automatic data analysis. A list of basic requirements was also given and included the need to supply summaries of results, suggestions for action and graphical displays.

Research at Bell Labs continued with the development of REX (Gale and Pregibon 1982). The aim of REX was to assist the novice user in regression analysis by checking for violations of assumptions. The strategy used was to undertake a model independent scrutiny of the data, to assess the model adequacy and to examine the fitting method. REX is written in LISP and interfaces with the package S. The strategy incorporated in the knowledge base was elicited by means of working through examples. Other work undertaken in the early eighties included research by Hájek and Ivánek, Porter and Lai, O'Keefe , Smith, Lee and Hand. The system GUHA 80 ,(Hájek and Ivánek 1982), was aimed at exploratory data analysis, the emphasis being on the formulation of hypotheses. STATPATH is a system which employed a binary tree search to identify appropriate analyses, (Portier and Lai 1983). STATPATH advised on an appropriate analysis and referred the user to the relevant package; as such it did not access the data. ASA, (O'Keefe 1982) was a system which was designed to help a client analyse an experiment which has already been designed. BUMP was constructed as an interface to the package MULTIVARIANCE, (Smith, Lee and Hand 1983). BUMP was not intended as an expert system but nevertheless tackled some of the relevant issues. By means of a dialogue the system helps the user to define the analysis they want, offering help if required. It did not tender advice, nor could it explain why a decision has been made.

Hahn, in his 1985 review paper, suggested that the best opportunities for technical progress seem to be in the development of specialised, rather than general, applications packages. Much of the subsequent research has indeed focussed on specific areas, although some work on building intelligent front ends to general statistical packages has been undertaken.

2.3 More Recent Work : Post 1985

It is interesting to classify the statistical expert systems developed in the mid eighties by the approach used. Some systems have been designed primarily as front ends to existing statistical software while other systems access statistical software to provide the necessary numerical computations for a specific area. A number of systems do not use existing statistical software and a few systems have been written using expert system shells. An expert system shell provides, for a specified form of knowledge representation, an inference engine and some form of explanation and help facilities. The users of expert system shells need only express their knowledge in the form required by the system.

Table I summarises the information available about the development of various expert systems for statistics in 1985 and 1986.

2.3.1 SES Which Use Expert System Shells

The work by Oldford and Peters (1986a, 1986b) was originally undertaken using the expert system shell EMYCIN, although later work has used the expert systems building package LOOPS on a Lisp machine. The system accesses a statistical analysis packages called DINDE which resides on the Lisp machine.

EXPLORA is a system written in LISP, which utilises the expert system shell BABYLON, (Klösgen 1986). The SAS package is used to provide the necessary numerical computations. EXPLORA runs on a Symbolics Lisp machine and is used for exploratory data analysis. Both Klösgen and Oldford and Peters used an object oriented approach where the primary emphasis is placed on the objects within

Reference		Name	Expert System Shell	Language(s)	Statistical Package(s)	Machine	Area of Expertise
Oldford & Peters	1986a 1986b	LOOPS	Mini Mycin Lisp		Dinde	) Xerox) Interlisp	Collinearity
Klösgen	1986	EXPLORA	Babylon	Lisp	SAS	Symbolic Lisp	Exploratory Data Analysis
Berzuini et al	1986		EXPERT	Fortran	MLP		Front End
Nelder	1986	GLIMPSE	APES	Prolog	GLIM	Sun	Front End
Jida et al	1986		-	Prolog	CHADOC	IBM-PC	Front End
Hilhorst et al	1986	ROCHEFORT	-	?	Oracle (dbms) SPSS,SASS BMDP	VAX	Interface data base systems with statistical packages
Gale & Pregibon	1986	STUDENT	-	Lisp	S	?	Strategy Acquisition
Darius	1986		-	SAS	SAS	IBM-AT	Expert system shell
Carlsen & Heuch	1986	EXPRESS	-	Fortran	BMDP	Sperry 1100	Test selection (2 - sample)
Froeschl & Grossmann	1986		-	Fortran Prolog	SPASP	?	Analysis of longitudinal data
Prat et al	1985	STATXPS	-	?	SCA	IBM-PC	Time series
Galmacci	1986	TSX	-	Lisp	SAS	2xIBM-PC	Time Series
Pregibon	1986	TESS	-	Lisp	-	Lisp machine	Shell for encoding statistical strategy
Dambroise & Massotte	1986	MUSE	-	APL, Prolog	-	VAX	Multivariate analysis
Hietala	1986	ESTES	-	Pascal	-	IBM-PC	Time Series
Hakong & Hickman	1985	SASS	-	Nested Interactive array language	-	?	Social Science Statistics
Esposito et al	1986	EXPER	-	Prolog Pascal	-	IBM 4341	Experimental Design

Table I : Overview of Statistical Expert Systems - Post 1984

the system rather than operations or procedures to be undertaken.

Other work in this area includes a front end to the package MLP using the shell EXPERT, (Berzuini et al 1986).

2.3.2 Systems Designed as Front Ends to Existing Statistical Software

GLIMPSE, designed as a rational front end to GLIM (Nelder 1986), is the most well known work in this area. GLIMPSE is written using the Prolog shell APES and runs on a SUN workstation. GLIMPSE offers advice and help on different activities such as data input, data validation, model selection and model prediction.

Rochefort is an ambitious project designed to link data base management systems and statistical software (Hilhorst et al 1987). It is also anticipated by the authors that statistical expertise for selection of appropriate analysis methods would be included.

Other work in this area includes that described by Berzuini et al (1986), mentioned in the previous section, and Jida & Lemaire (1986). The work described by Jida is a front end, written in Prolog, to the statistical package CHADOC. The front end enables the user to generate the necessary command file for CHADOC and also provides a semantic analysis of those commands in order to avoid invalid analyses.

2.3.3 SES Which Access Statistical Packages

There several systems which fall into this category, the best known of which is the system Student, (Gale and Pregibon 1984, Gale 1986). Student is written in LISP and accesses the statistical package S. Student offers an automated learning strategy and is designed to allow a professional statistician to construct a knowledge base by selecting and working examples and by answering questions.

STATXPS is an expert system for time-series analysis which accesses a statistical package called SCA, (Prat et al 1985). Darius (1986) developed an expert system shell written in the SAS language. Other work in this area includes Carlsen and Heuch (1986), Froeschl & Grossmann (1986), Galmacci (1986).

2.3.4 Systems Developed Without an Expert System Shell or Statistical Package

Some Statistical Expert Systems have been developed using an Artificial Intelligence Language, a Procedural language or a combination of both. ESTES is a system for Time Series Analysis written in Pascal on a Macintosh, (Hietala 1986). ESTES makes full use of the windowing facilities available on the Macintosh and is very user-friendly providing both textual and graphical explanations for statistical terms. The SASS system, (Hakong & Hickman 1985), is interesting because it is based on intersecting sets of properties of statistical techniques. SASS has been developed using a Nested Interactive Array Language.

TESS is a system which uses a tree based strategy and is written entirely in LISP, (Pregibon 1986b). In order to assist the statistician in the task of coding numerical routines TESS provides a mini language for statistical computations and enables an expert statistician to encode their strategy for analysing a particular type of data set. Once the knowledge has been encoded the system can be used by non statisticians to analyse their data sets.

Other work in this area is described by Esposito et al (1986) and Dambroise & Massotte (1986).

Chapter Three

Design of a Statistical Expert System

3.1 Introduction

Any expert system should be able to explain and justify its reasoning as well as to offer help and guidance throughout a consultation and the design of the system should take these as basic requirements. There are additional considerations necessary in designing statistical expert systems, including the need to access data during the consultation; these requirements are considered in this chapter. The pattern of consultation to be followed by a system and the choice of knowledge representation are also discussed and finally a logical design for a statistical expert system is proposed.

3.2 Design Considerations for Statistical Expert Systems

3.2.1 Primary Considerations

When developing an expert system it is important to establish both the scope of the system and the prospective users of the system before more specific design work can be undertaken.

The scope of the system will affect both the choice of knowledge representation and the general design of the system. An expert system may be focussed on a narrow and highly specific domain area or may have a wide domain. There is no clear distinction to be made between these two possibilities and it is likely that the scope of a statistical expert system falls somewhere between them. The aim of this project was to develop a software framework suitable for expert systems in small and well defined areas of statistics. The 'end-user' also needs to be considered carefully. There is a wide range of possibilities from the expert statistical consultant to the statistical novice and it would be difficult to cater for all of them in a single system. The statistically naive researcher would need extensive help and guidance to ensure the appropriate analysis is carried out and to interpret the results, whereas experts may want to move through the system quickly, looking only at the results they are interested in. The aim in this project was to develop a system for use by research workers in industry who are regular users of statistical techniques.

3.2.2 Design Features

An expert system should be capable of justifying its conclusions and telling the user why a particular question is being asked. In order to do this it is necessary to keep some form of trace of the consultation process that can be accessed and understood by the user. In addition a statistical expert system should be able to explain statistical terms as well as providing help throughout the consultation.

As with any software, an expert system needs to be structured so that it is easily modifiable, both to allow for ease of maintenance of the system and to cope with developments in the knowledge base. The concept of a dynamic knowledge base is very important in the area of statistics for two reasons; to enable new developments in the domain area to be included and to allow an expert statistician to alter the strategy expressed in the system. There is seldom a single correct strategy in any given area of statistics and different statisticians often use different strategies; thus it is important to have a knowledge base which can be altered easily by an expert statistician.

Statistical expert systems have two main sources of information; the user and the data. Thus in developing a statistical expert system it is essential to access statistical routines or packages during the consultation process as well as providing a flexible and easy to understand user interface. This precludes the use of existing expert system shells which cannot interface with other software.

A statistical expert system also needs to be able to allow for the possibility of multiple objectives; in the domain of statistics a researcher often requires the answer to more than one question.

The system should be able to recommend the most appropriate and most powerful techniques, at the same time allowing the user an element of choice between valid techniques.

A number of people have considered these features; in particular Hand(1985) and Hahn(1985) discuss them more fully. Some of these features need to be considered at the logical design stage, for example, the need to access data during the consultation. The majority of features can be incorporated at the software design stage; this is discussed in more detail in Chapter 8.

3.3 Pattern of Consultation

In order for expert systems to be able to explain and justify their reasoning it is necessary that they use a pattern of consultation that is comprehensible to the user. This does not mean that the expert system must mimic the experts actions, rather that it should operate in a way that can be explained to, and understood by, the user, i.e. it should fit in the 'human window', (Michie and Johnston 1984 p70). A Statistical Expert System can also offer more facilities than a practising statistician because of the speed of processing, for example, running several diagnostic tests takes little time for the computer but would be rather time consuming for a human expert (Hand 1984, Buja 1984).

A great deal of research has been undertaken to try and establish how human consultants interact with their clients (Hand 1984, Clayden - personal communication). Hand suggested that a statistical consultant operates in a similar manner to a medical consultant, initially generating a set of plausible hypotheses and then trying to verify these hypotheses. This has a 'funnelling' effect with the consultant trying to reduce the number of possibilities and thus limit the search space.

One of the major reasons for the development of expert systems stems from the realisation that it is not, in general, practical to foresee and check all possible eventualities. Many techniques used in expert systems concentrate on reducing the number of possibilities to be considered as much as possible. Thus it would seem appropriate to adopt the broad pattern of consultation where the first stage is to establish a subset of appropriate techniques and then to consider each of the techniques in more detail.

When a technique is being considered for use on a particular data set then it is first tested for use on the original data. However, if a parametric technique cannot be verified for use on the original data then the user may wish to try transforming the data. The use of transformations can, therefore, affect the flow of control within the system. Thus the consultation may be cyclic in nature, moving from verification to transformation back to verification where parametric techniques are concerned. This needs to be incorporated in the system design.

3.4 Knowledge Representation

Having established the scope of the expert system and the pattern of consultation the next stage is to decide on an appropriate way to represent the knowledge. There are three main forms of knowledge representation, rules, frames and semantic nets.

Rules are the predominant from of representation used in expert systems and take the form

IF condition THEN action or assertion

These rules may be processed sequentially, forward chaining, or by trying rules that would help to establish a goal the system is interested in, this is known as backward chaining.

Semantic nets are used to represent relationships between objects in the domain as links between nodes, they are particularly useful where inheritance is important.

Frames are generalised record structures which describe a class of objects or events. Slots in the frame may contain default values, procedures, actions or even pointers to other frames. Like semantic nets, it is easy to include inheritance properties when using frames.

It is important to use a knowledge representation that is comprehensible to a statistician who wants to modify the knowledge base. The choice of representation also depends on the scope of the domain. For example, where the domain covers a large area, frames may be most appropriate as they provide a way of describing families of objects.

For this project, the size of domain was intentionally limited to small, well defined areas and production rules were chosen as the most appropriate knowledge representation. The primary reasons for this choice were ease of understanding and flexibility in the ways in which production rules can be processed. The different types of rule and the methods of inference adopted in this project are discussed in Chapter 4.

3.5 Logical Design

The construction of software systems is facilitated by using a structured design methodology which separates the development process into a number of well-defined stages. The motivation behind these methodologies is the emphasis on the problem definition part and the clear separation between the logical and physical design. The advantages of a logical design are that it is independent of hardware and software considerations and that it allows greater interaction between the user and the designer, often via easy to understand graphical methods.

Entity analysis was originally proposed as a methodology for developing database systems (Chen 1977) but it was soon found to be a useful tool in many areas of software engineering (Knight et al 1987). Entity analysis provides a clear diagrammatic view of the logical design of the system and has been used in the design of THESEUS.

3.6 Entity Analysis for THESEUS

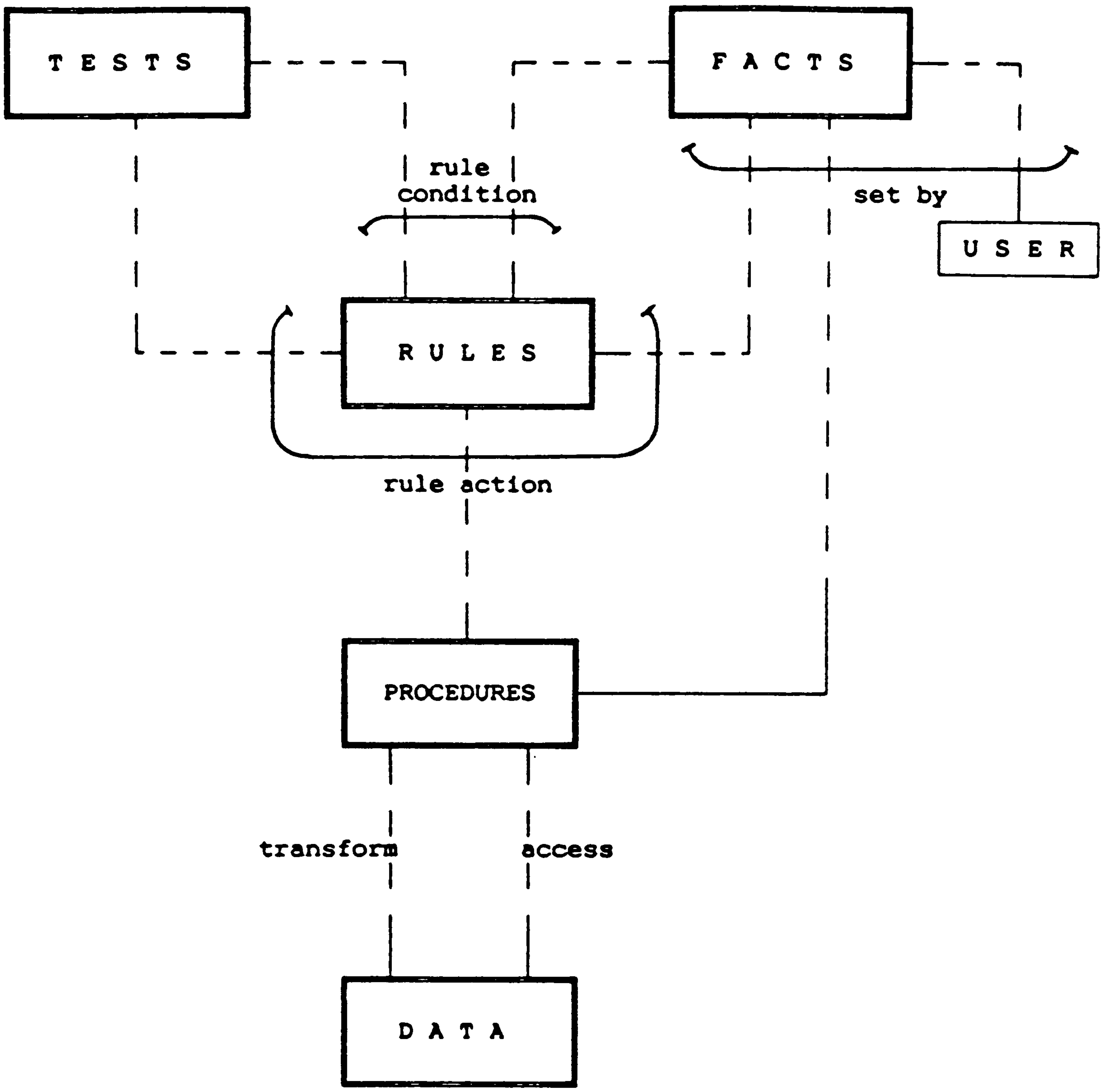
Chen's design representation contains three classes of things : entities, relationships and attribute. There are three different stages in Entity Analysis :

1. Identifying the Entities and the relationships between them in diagrammatic form
2. Identifying attributes for each entity
3. .Constructing Life-Cycle Diagrams for the status of each entity.

When the logical design is translated to software code, each entity is declared as an array of records where the records are defined by the list of attributes for the entity. The Life-Cycle diagrams show how the status of each entity can change within the system, thus indicating the flow of control. The Entity-Relationship diagram shows the relationships between the entities and thus indicates which other entities must be considered when a member of one entity type is being processed .

Figure 3.1 shows the entity relationship model for THESEUS. Entities are objects that can be uniquely identified, and classified into separate types. The entities identified in THESEUS are rules, facts, tests, procedures and experimental data; the lines between the entity types show the relationships. For example, facts

Figure 3.1 : Entity Relationship Diagram



can be set either by the action of a rule or by a procedure or by asking the user. This optionality is shown by the use of dashed lines; that a fact can only be set in one of these ways is shown by the line drawn across the three optional relationships, labelled 'set by'. There are two relationship lines between tests and rules, a test can be part of the condition of a rule or can be set as part of the action of a rule.

After the construction of the graphical model, the attributes of each entity type are determined, these attributes are the properties of the objects which we need to record. The attributes for the entities in THESEUS are given below :

Entity : FACTS	Possible Values
Attributes :	
- Name	character string
- Setby rule	TRUE or FALSE
- Setby procedure	TRUE or FALSE
- Setby user	TRUE or FALSE
- Dataset	character string
- Status	UNTRIED, STRUE, SFALSE CURRENT, UNKNOWN

Entity : TESTS	Possible Values
Attributes :	
- Name	character string
- Parametric	TRUE or FALSE
- Dataset	character string
- Chosen-by-user	TRUE or FALSE
- Status	UNTRIED, LOOK_AT, CURRENT RECOMMENDED, NOT_VALID VALID, UNKNOWN

Entity : PROCS	Possible Values
Attributes :	
- Name	character string
- Called-by	RULES, FINDFACT
- Status	NOT_CALLED, CALLED

Entity : DATA INFO	Possible Values
Attributes :	
- Name	character string
- Form (Algebraic expression)	character string
- Mean [1..number of groups]	array of real numbers
- Var [1..number of groups]	array of real numbers
- Status	UNTRIED, CURRENT ACCEPTED, REJECTED

Entity : RULES	Possible Values
Attributes :	
- Identifier	character string
- Condition	
Any number of	
- operator	' ' or 'NOT'
- fact or test name	character string
pairs	
- Action	
Any number of	
- fact, test or procedure name	character string
- name_is	FACT, TEST, PROC
- action	depends on name_is, see Note 1
triplets	
- Status	UNTRIED, FIRED, FAILED SKIPPED, UNKNOWN

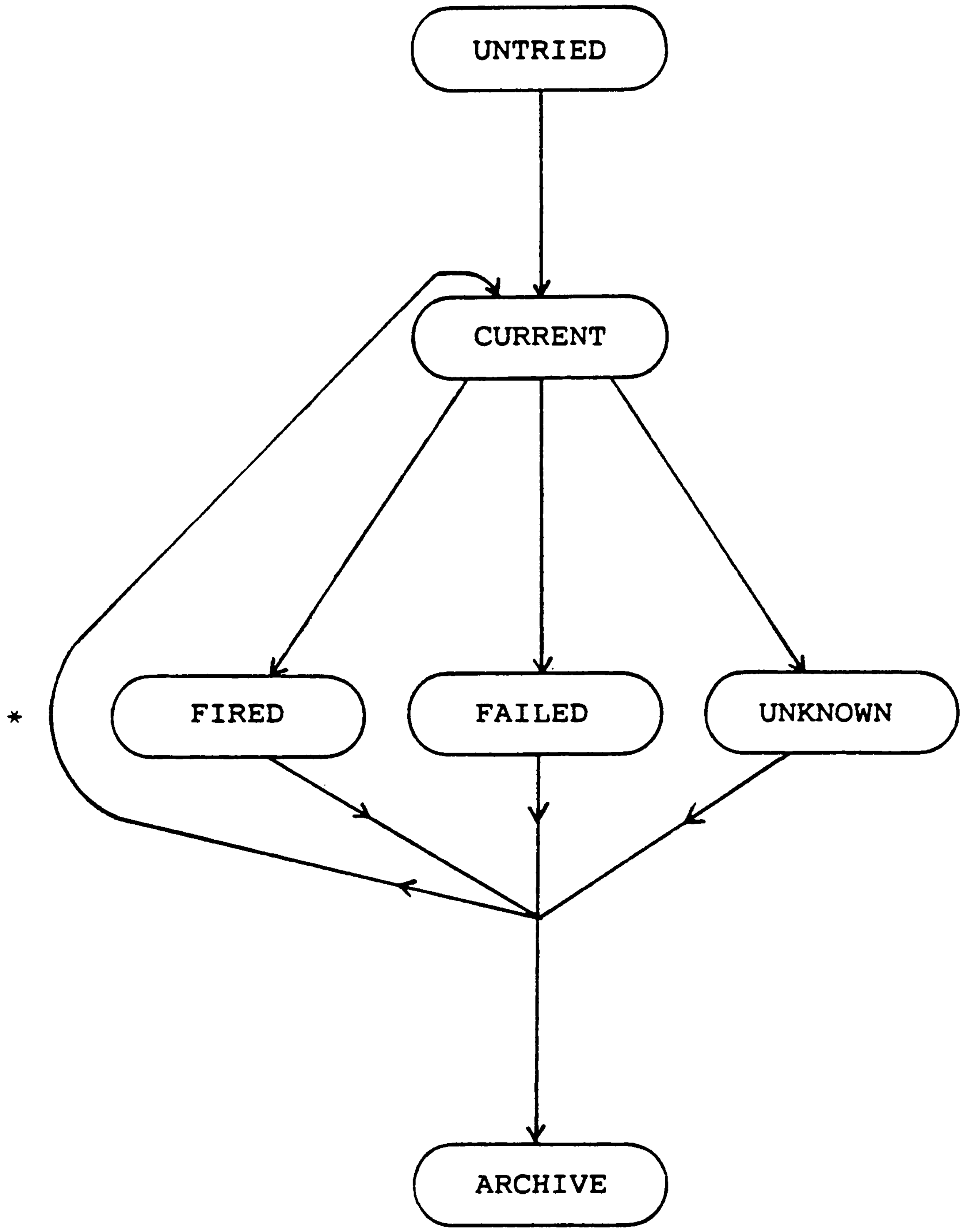
Note 1

name_is	possible values for action
FACT	STRUE, SFALSE
TEST	LOOK_AT, RECOMMENDED, NOT_VALID,VALID
PROC	CALL

Once the attributes have been established the Life-Cycle diagrams for the status of each entity are constructed, showing how the status of each entity may change within the system, see Figures 3.2 to 3.6. For example, in the life-cycle for Test status, the first change of status is from UNTRIED to LOOK_AT, this reflects the first part of the consultation process (establishing a list of potential tests). A test can only be considered further if its status is already LOOK_AT; if this is the case then the test status will, at some stage, become CURRENT when the test will be considered more closely. The possible outcomes are RECOMMENDED, VALID, NOT_VALID or UNKNOWN. RECOMMENDED means that the system considers this technique to be the best of the list under investigation.

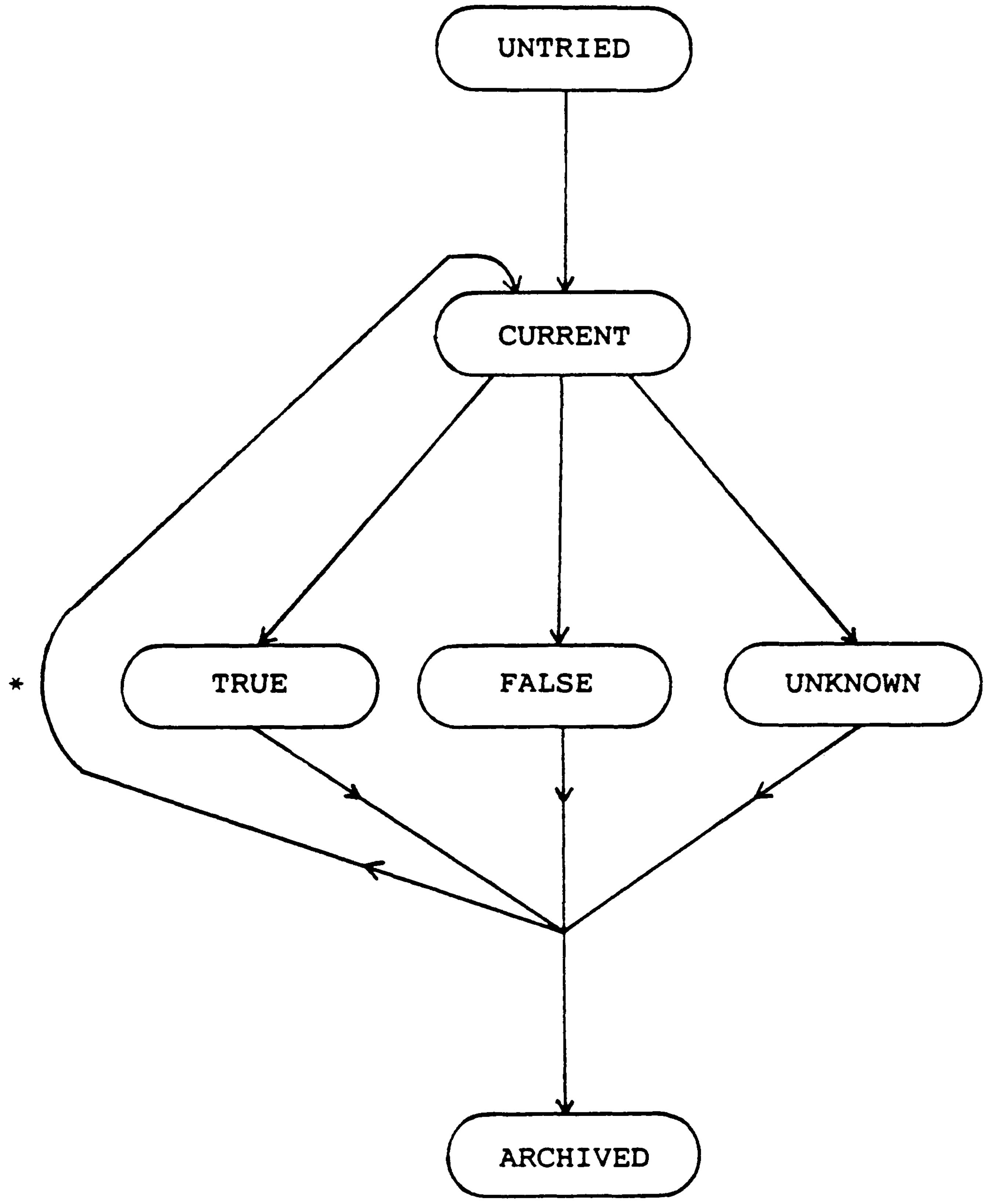
If a parametric test becomes VALID, NOT_VALID or UNKNOWN the status may return to current if the data is transformed. Each Life-Cycle diagram has a node labelled ARCHIVED, which indicates that the status does not change any further and remains at the value given in the previous status node.

Figure 3.2 : Life-Cycle Diagram - Rule Status



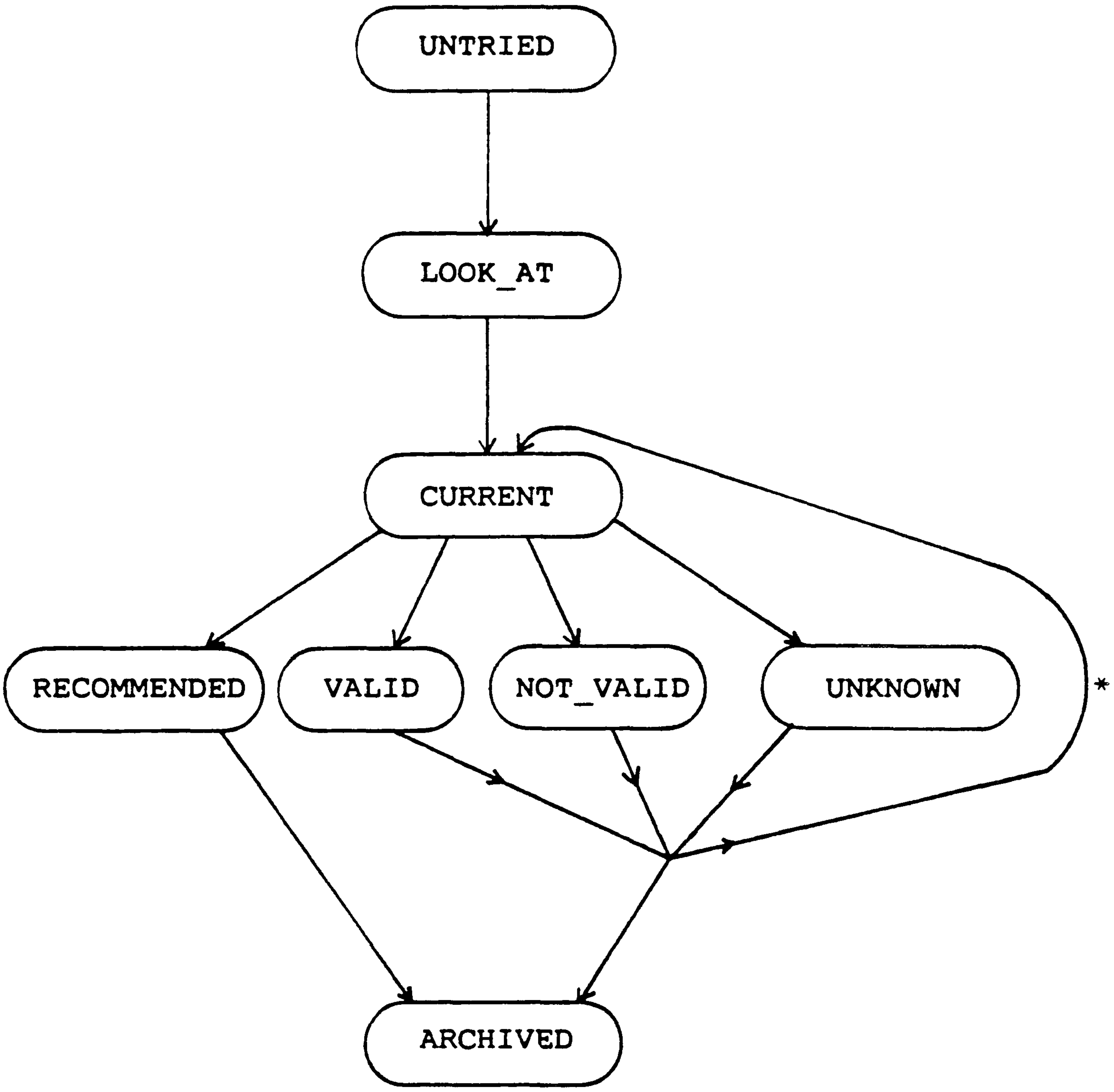
* This can only occur in backward chaining rules when the data is transformed and some facts need to be re-established on the new data set.

Figure 3.3 : Life-Cycle Diagram - Fact Status



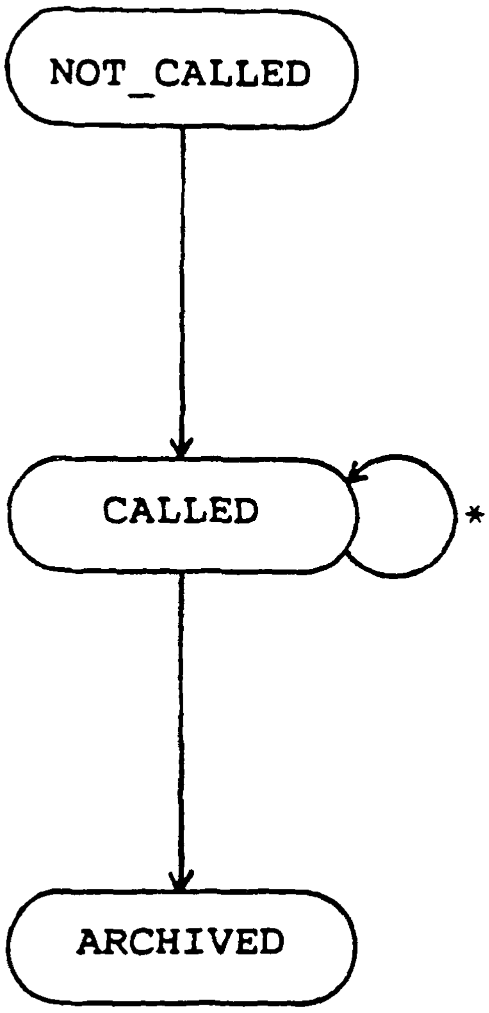
* This will occur when the data is transformed and the fact is a 'dynamic' fact that needs to be re-established on the new data set

Figure 3.4 : Life-Cycle Diagram - Test Status



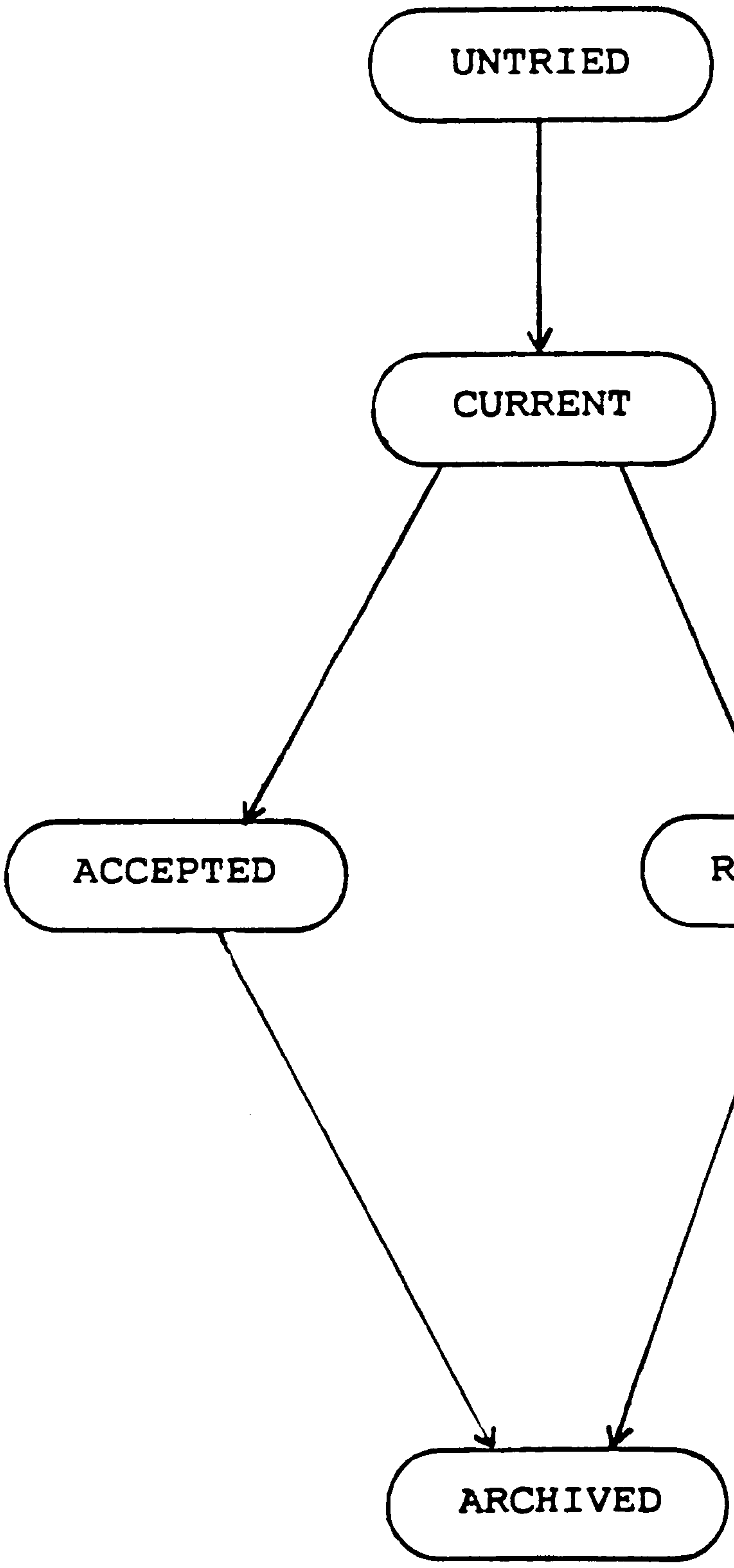
* This only occurs if the data is transformed and the test is a parametric test

Figure 3.5 : Life-Cycle Diagram - Procedure Status



* This only occurs if the data is transformed

Figure 3.6 : Life-Cycle Diagram - Data Status



Chapter Four

Decision Making and Control in a Statistical Expert System

4.1 Introduction

Once the choice of knowledge representation has been made and the form of consultation decided, the next stage, after the logical design, is to consider in more detail the methods of inference and the control structure to be used. Rules can be processed using either forward or backward chaining or using some combination of both. In general terms the prototype system described here uses forward chaining when trying to establish a list of possible methods and backward chaining when trying to check the validity of methods. Forward and backward chaining and the protocol for applying a specific rule are described in the next two sections.

During the development of the prototype system the general structure described above remained the same, however, the actual implementation altered considerably. The reasons for such alterations were to decrease the amount of time the system had to spend looking through the rules and, more importantly, to make progress through the system clearer to the user. The development of the inference process and control structure is discussed in this chapter, and the final method of inference and the control structure are described in detail.

4.2 Applying a Rule

Before going any further it is be useful to establish the way in which an individual rule of any type is processed. Once the system has decided to try to apply a particular rule, it considers each part of the condition in turn. Each part of the condition must be satisfied before the system moves on to consider the next part of the condition. As soon as one part fails then the rule is failed.

In considering each part of the condition, the system will first check whether the status of this fact has already been established as true or false. If the status has not been established then the system looks at the attributes to find out how to establish the fact. As already stated in section 3.6, a fact can be set by asking the user, calling a procedure or by trying other rules.

4.3 Forward and Backward Chaining

Forward chaining involves considering each of the appropriate rules in turn, working through them sequentially and carrying out the actions of those rules whose conditions are satisfied.

Backward chaining is carried out by supplying the system with a goal to backward chain on. The system looks through the rules until it finds one with an action that would establish that goal. The system then tries to apply that rule. If that rule fails then the system continues looking for the next rule which has the goal on the action side of the rule. This process continues until the goal is established or no more relevant rules can be found.

In the course of backward chaining on a particular goal the system may encounter a fact that is not yet known and which is set by other rules. When this occurs the system suspends backward chaining on the original goal and backward chains with this fact as a goal. When the new goal has been established the system resumes backward chaining on the original goal. Figures 4.1 and 4.2 show a simple rule base and an example of backward chaining using that rule base.

Figure 4.1 : Simple Rulebase to Demonstrate Backward Chaining

```
R1  IF   outliers
    THEN not_valid test parametric
         recommend test nonparametric

R2  IF   not outliers and normal_data and variances_equal
    THEN recommend test parametric
         valid test nonparametric

R3  IF   not outliers and not normal_data
    THEN not_valid test parametric
         recommend test nonparametric

R4  IF   not outliers and not variances_equal
    THEN not_valid test parametric
         recommend test nonparametric

R5  IF   shapiro_wilk_sig5 and not user_says_data_normal
    THEN false fact normal_data

R6  IF   shapiro_wilk_sig5 and user_says_data_normal
    THEN true fact normal_data

R7  IF   not shapiro_wilk_sig5
    THEN true fact normal_data

R8  IF   levene_sig5
    THEN false fact variances_equal

R9  IF   not levene_sig5
    THEN true fact variances_equal

outliers              - set by the user
normal_data           - set by other rules
variances_equal       - set by other rules
shapiro_wilk_sig5     - set by a procedure
user_says_data_normal - set by the user
levene_sig5           - set by a procedure
```

Figure 4.2 : Example - backward chaining on 'parametric'

```
Goal : parametric
Trying rule : R1 ask user about outliers (false)
                 [rule fails]
Trying rule : R2 not outliers is true
                 set up normal_data as a goal
                 [R2 remains current]

      Goal : normal_data
      Trying rule : R5 call procedure to set
                       shapiro_wilk_sig5 (false)
                       [rule fails]
      Trying rule : R6 shapiro_wilk_sig5 is false
                       [rule fails]
      Trying rule : R7 not shapiro_wilk_sig5 is true
                       [rule fires]
      Action of   : R7 set normal_data to true

Goal : parametric
Trying rule : R2 normal_data is true
                 set up variances_equal as a goal
                 [R2 remains current]

Goal : variances_equal
Trying rule : R8 call procedure to set
                 levene_sig5 (false)
                 [rule fails]
Trying rule : R9 not levene_sig5 is true
                 [rule fires]
Action of   : R9 set variances_equal to true

Goal : parametric
Trying rule : R2 variances_equal is true
                 [rule fires]
Action of   : R2 recommend parametric test and
                 valid nonparametric test
```

4.4 The Development of an Inference Mechanism

Initially the system was structured so that all the rules were stored in one array. The consultation process used at first can be summarised as follows :

1. Establish a list of possible methods by forward chaining through the rules, only considering those rules which contained an action to LOOK_AT a test or tests.

2. Verify the methods - set up each test as a goal for the system to backward chain on.

3. Return to step 1 - finishing when an empty list is returned from the forward chainer.

It soon became apparent that the system was wasting time looking through the rule array in order to identify the forward chaining rules. Thus the first, and simplest, alteration was to separate the forward and backward chaining rules. This is carried out when the rule-base is picked up by the system, any rule which has an action to LOOK_AT a particular test is stored in a separate array. This makes no noticeable difference to the user but does mean that the system is not wasting time searching to find the appropriate rules to forward chain on.

Once the knowledge acquisition was underway and a realistic rule base was being tried in the system it soon became apparent that dealing with the possibility of transformations within the backward chaining rules was rather complicated. Rules had to be developed for assessing the validity of methods on the original data and other rules had to be developed to deal with transformations and the possibility of trying more than one transformation. Although this was possible it did mean that the condition part of some rules became rather complex and understanding the path the system was following became quite difficult.

This difficulty was overcome by using a two level strategy whereby the backward chaining rules apply to the current data set only. A higher level of rules was introduced which, after a goal has been verified using the backward chaining rules, decide whether to move on to the next test in the list or whether to transform the data. If the data is transformed then the backward chaining rules are applied again to verify the status of the test under consideration on the transformed data. Thus the backward chaining rules may be applied several times in the course of verifying a particular technique.

Three types of rule can now be identified :

I : Forward chaining rules - used to establish a list of possible techniques

II : Backward chaining rules - used to verify the validity of methods on the current data set

III : Meta rules - used to decide whether to move on to the next test in the list or to transform the data

4.5 Control Structure

Flow between the different types of rule is effected by a control module. The structure is described using pseudo code given below.

```
REPEAT

     forward chain to supply a list of possible tests

     WITH each test in the list

          REPEAT

               IF test is not RECOMMENDED yet
               THEN backward chain to establish test

               search meta rules to set NEXT_TEST to true
               or to false (and transform data)

          UNTIL the meta rules have set NEXT_TEST to true
                or current test has been RECOMMENDED

     END of WITH each test in the list

     ask user whether they wish to consider any
     FURTHER_ANALYSIS

UNTIL FURTHER_ANALYSIS is false or
      forward chaining rules supply an empty list
```

4.6 Forward Chaining Rules

Rules which the system uses to establish a list of possible techniques are the most straightforward type. The condition part of these rules is usually composed of facts relating the basic nature of the data, such as the number of groups or the hypotheses of interest to the user. These are the only rules which may also have tests as part of the condition. This may happen where a particular test is used before other techniques are considered; for example the ANOVA may be used before considering multiple comparisons.

These rules are processed by forward chaining as described in section 4.3 . If the condition part of a rule contains a test that has not yet been established then the status of that rule is set to SKIPPED. Each time these rules are considered the system starts at the top of the list and works through considering only those rules whose status is UNTRIED or SKIPPED. The forward chainer stops as soon as one rule has fired. The action part of the rule will be to set the status of a number of tests to LOOK_AT; thus a list of possible techniques has been established.

Examples

```
R3  IF    SEVERAL_GROUPS and
          OVERALL_TEST
    THEN  LOOK_AT TEST ONE_WAY_ANOVA
          LOOK_AT TEST KRUSKAL_WALLIS
```

SEVERAL_GROUPS is set by calling a procedure which counts the number of groups in the data set
OVERALL_TEST is set by asking the user if they wish to consider an overall test of significance

```
R7    IF    MULTIPLE COMPARISONS and
            PAIRWISE and
            ALL_COMPARISONS
      THEN  LOOK_AT TEST NEWMAN_KEULS
            LOOK_AT TEST DUNCANS
            LOOK_AT TEST K_SAMPLE_RANK
            LOOK_AT TEST KRUSKAL_WALLIS_PAIRS
```

PAIRWISE is set by asking the user whether they wish to consider pairwise comparisons
ALL_PAIRWISE is set by asking the user if they wish to look at all possible pairwise comparisons
MULTIPLE COMPARISONS is set by other rules, thus the system would have to backward chain to establish this fact.

4.7 Backward Chaining Rules

These rules are used by the system in order to establish the validity of a technique by checking the appropriate constraints and assumptions. In the logical design a distinction was made between two types of fact, static facts and dynamic facts. Static facts are independent of any transformations of the data set; for example, facts relating to the number of groups or to outliers. These facts once established cannot be changed.

Dynamic facts are those whose status may change if the data is transformed; for example, facts relating to normality.

The rules under discussion here may contain a combination of both types of fact. Thus these rules establish a technique on the current data set, original or transformed. These rules may be processed several times in trying to establish a particular technique, each time with a different transformed version of the data; in this case the status of dynamic facts is re-established for each transformation of the data. A side effect of these rules is that they also set facts used by the Meta rules to decide whether a transformation is necessary.

These rules are processed by backward chaining as described in section 4.3.

Examples

```
R26  IF    NOT OUTLIERS  and
           VARIANCES_EQUAL and
           NORMAL_DATA
     THEN  TRUE  FACT ACCEPT_PARAMETRIC
           FALSE FACT TRANS_FOR_NORMALITY
           FALSE FACT TRANS_FOR_VARIANCES
           FALSE FACT ADJUST_FOR_UNEQ_VAR
```

OUTLIERS, NORMAL_DATA and VARIANCES_EQUAL are all set by other rules

```
R93  IF    ACCEPT_PARAMETRIC and
           BALANCED
     THEN  RECOMMEND TEST NEWMAN_KEULS
           VALID TEST DUNCAN
           VALID TEST K_SAMPLE_RANK
           VALID TEST KRUSKAL_WALLIS_PAIRS
```

ACCEPT_PARAMETRIC is set by other rules
BALANCED is set by calling a procedure which checks that the sample sizes are equal

```
R54  IF    MORE_THAN_20_OVERALL and
           NOT SHAPIRO_WILK_SIG5
     THEN  TRUE FACT NORMAL_DATA
```

MORE_THAN_20_OVERALL is set by calling a procedure which counts the total number of observations

SHAPIRO_WILK_SIG5 is set by other rules, this is because the form of the Shapiro Wilk test may be to consider each group individually or to treat the data as a whole.

4.8 Meta Level Rules

These rules are used to enable the system to decide whether to move on to the next test in the list of possible tests or to call the procedure which transforms the data. They are denoted 'Meta' rules because they govern, to some extent, the flow of control within the system. Meta rules are processed by forward chaining as described in section 4.3 . The status of all Meta rules is returned to UNTRIED before they are processed again.

Examples

```
M1    IF NOT PARAMETRIC
      THEN TRUE FACT NEXT_TEST
```

i.e. IF the test that is being considered is nonparametric then one pass through the backward chaining rules using the original data is sufficient and the system can move on to the next test in the list.
The fact PARAMETRIC is set by looking at the attribute field for the current test

```
M4    IF    PARAMETRIC and
            NOT OUTLIERS and
            TRANS_FOR_VARIANCES and
            MORE_TRANS_TO_TRY
      THEN CALL PROC TRANSFORM
```

OUTLIERS and TRANS_FOR_VARIANCE are set by the backward chaining rules
MORE_TRANS_TO_TRY is set by the procedure TRANSFORM; the initial value is TRUE

Chapter Five

Approaches to Knowledge Acquisition

5.1 Introduction

The logical design and the methods of inference to be used have been established, the next major consideration is knowledge acquisition.

It is widely acknowledged that knowledge acquisition is a major part in the development of an expert system; it is probably true to say that it is the most time consuming and labour intensive part of the development program. (Duda and Shortliffe 1983, Wittkowski 1986, Gale 1987) Duda and Shortliffe in their paper on Expert Systems Research summarised the main problems of knowledge acquisition as follows :

> " The identification and encoding of knowledge is one of the most complex and arduous tasks encountered in the construction of an expert system. The very attempt to build a knowledge base often discloses gaps in our understanding of the subject domain and weaknesses in available representation techniques. Even when an adequate knowledge representation formalism has been developed, experts often have difficulties expressing their knowledge in that form. Thus the process of building a knowledge base has usually required a time-consuming collaboration between a domain expert and an AI researcher."

The usual approach of dialogue sessions between a domain expert and a knowledge engineer is not always appropriate and research into the problems of knowledge acquisition has, to date, concentrated on two different approaches. The first approach has been the development of specific knowledge acquisition techniques for specific types of knowledge (Gammack and Young 1985, Wittkowski 1986).

The other approach to knowledge acquisition is that of rule induction where a system is programmed to acquire the knowledge. Gale (1987) termed this knowledge based knowledge acquisition and it is being used in the development of a system called Student which is designed to learn strategy from examples. Methods of rule induction require a conceptual framework for the domain within which knowledge can be structured; the development of an appropriate framework can be time-consuming in itself. Even when the conceptual framework has been chosen the development of rule induction methods is technically complex and is outside the scope of this project.

In this chapter the different types of knowledge involved in statistical expertise are considered and different methods of knowledge elicitation that are available are discussed. The approach used in building the prototype knowledge base for THESEUS is described in detail.

5.2 Statistical Expertise

Thisted (1986) gives a useful description of the different areas of expertise in statistics :

> "The complete expertise of an expert data analyst encompasses such areas as mathematical statistics; techniques of graphical display and analysis; rules of thumb for judging the importance of apparent indications; copious examples of bad or misleading analyses (coupled with a catalog of common errors made by novices, the avoidance of which is essential to respectability); methods, both ad hoc and those thoroughly grounded in theory, for basic operations such as smoothing, assessment of variability, and model building; and - perhaps most important - knowledge of how and when to elicit specific subject matter information from a scientific collaborator"

It can be seen that there are many different aspects to statistical expertise some of which overlap with other disciplines and some which are unique to statistics. For example, in the area of clinical trials the statistician needs not only expertise relevant to the analysis of the data but should also have a thorough understanding of the problems of data collection and validation. Such data handling problems have much in common with expertise in database management systems which are used in many non-statistical applications.

In considering the application of expert systems techniques to the area of statistics it is helpful to try to classify the different types of statistical expertise. The aim of this classification is to enable a system developer to select both appropriate knowledge acquisition techniques and knowledge representation schemes.

Wittkowski (1986), proposed a way of structuring statistical knowledge in order to establish appropriate knowledge representations. Gammack and Young (1985) proposed a general classification of knowledge so that appropriate knowledge acquisition techniques could be pinpointed; the domain of statistics was used as an example. There are some similarities between the two classifications. For example, Wittowski's knowledge on conceptual problem types seems to correspond with Gammack and Young's knowledge of concepts and relations. The difference between the classifications stem from the reasons for making such classification in the first place, Wittkowski's primary interest was to identify appropriate knowledge representation methods whereas Gammack and Young's main concern was to pinpoint specific knowledge acquisition techniques.

The classification proposed below is based on Gammack and Young's generalised structure but has been expanded to deal with the specific domain of statistics.

Framework : A statistician will have some form of conceptual structure in the domain which will define different types of analysis. This knowledge will be used to select areas of statistics appropriate to the data being considered. For example ANOVA and multivariate analysis could be two such areas.

Concepts : Knowledge about general concepts such as hypothesis tests, distributions, confidence intervals and degrees of freedom. Such concepts are a necessary foundation to understanding and undertaking any analysis.

Procedural Knowledge : Knowledge about the availability and requirements of specific statistical methods for analysis and assumption checking as well as knowledge about graphical representations. For example, knowing what methods are available for testing Normality and how they are implemented.

Heuristics : Rules of thumb used for judging the importance of effects such as violation of assumptions and how to handle them. For example, knowing when to let non-normality affect subsequent decisions.

Methodological Expertise : This enables the statistician to choose the most appropriate method from a range of those that could be used. For example, in selecting a multiple comparisons procedure when there is a control group present and the experimenter is interested in pairwise comparisons then Dunnett's test will be chosen in preference to Tukey's test.

Communication : Surrounding these different types or areas of knowledge is the expertise used in communicating effectively with the user. This involves not just establishing what the experimenter is interested in finding out, but also extracting information about the nature of the data that the statistician needs to make decisions about the most appropriate analysis. This may not be regarded as knowledge in the usual Expert Systems sense but is nevertheless included here because of the influence it should have in developing the knowledge base as well as in the design of the expert system.

Each of these areas of knowledge involves both 'technical' and 'professional' knowledge. 'Technical' knowledge is hard, factual knowledge obtainable from text books and the literature. 'Professional' knowledge is judgmental, experience related and considerably more difficult to elicit and represent, covering decisions such as when to allow unequal variances to affect subsequent decisions. An example of this is deciding to try transforming the data if Levene's test for unequal variances is significant at the 5% level.

5.3 Problems Encountered in Knowledge Acquisition

Knowledge acquisition for expert systems has, in the past, relied heavily on informal interviews between a knowledge engineer and a domain expert. The aim of such a process is to translate the information supplied by the domain expert into some predetermined format and so develop a prototype knowledge base. This knowledge base is then refined by a cyclic process of evaluation and modification. This approach demands a very high level of commitment and enthusiasm from the domain expert. The problem with this is that domain experts, because they are experts, often have little time to spare. Thus it is important to try and develop methods of knowledge acquisition which optimise the time spent with the domain expert.

The knowledge engineer, who has the problem of transferring the knowledge from the domain expert to the knowledge base, also has to ensure that an appropriate and powerful enough form of knowledge representation is used. A great deal of time can be wasted trying to manipulate knowledge in order to make it fit a particular representation; this is a well known disadvantage of expert system shells (Bell 1985). Domain Experts often find it difficult to articulate their decision making processes and face further problems of recognition and interpretation when trying to understand and evaluate the performance of the knowledge base.

Expertise in any domain will contain different types of knowledge (section 5.2 discussed the different types of knowledge in statistics). The development of a knowledge base should be a process of identifying these different types, choosing an appropriate knowledge representation scheme and then employing knowledge elicitation procedures appropriate to the application.

5.4 Knowledge Elicitation Techniques

There are a number of methods available for aiding knowledge elicitation many of which have been borrowed from other fields such as questionnaire design and industrial psychology. An overview of the main methods is given in this section.

5.4.1 Interviews

Interviewing methods are most helpful in the initial stages of knowledge acquisition for establishing the main concepts and components of the domain as well as defining the terminology used. In any area of knowledge acquisition structured interviews can be helpful in ensuring that the domain of interest is covered as completely as possible. However in order to cover the domain in a structured interview it is essential to have a clearly defined model of that domain. Such a model will probably be derived by initial interviews or some other method. It is interesting to note that domain experts often forget to state relevant knowledge and only remember it when the expert system behaves wrongly, Welbank (1983). The limitations of interviewing become more apparent when the domain expert is trying to evaluate the prototype knowledge base and trying to establish what distinguishes the performance of the expert from the inferior performance of the system.

5.4.2 Protocol Analysis

Protocol analysis involves observing and recording the action of the domain experts as they work through scenarios. This method has the advantage that the task situation is completely natural and the task can be done exactly as it normally is. The merit of this approach is that it gives the knowledge engineer a process to model. As the prototype knowledge base begins to take form then more specific scenarios or examples can be used to find out how the expert deals with special situations.

There are disadvantages in protocol analysis which are summed up in the report by Welbank (1983) p23 :

> "The subject cannot verbalise as fast as he reasons, which makes for important deficiencies in the type of material collected. He may not report what is obvious to him. He may leave out steps in his reasoning. Most importantly he does not naturally give 'if x, then y' type rules, or explain his reasons for deciding to do one thing rather than another. He may not have time to explain even if he is asked to."

Protocol analysis is very time-consuming and is a skilled and difficult task. A good understanding of the domain is essential for analysing the protocols accurately. Protocol analysis has most often been used as a way of comparing what experts say they do with what they actually do. (Nii 1984)

The knowledge acquisition in REX (Gale 1987) was undertaken using a form of protocol analysis where the expert (Pregibon) kept records of his own analyses and then studied the records to abstract a description of what he was doing. In this situation, where the knowledge engineer is the domain expert, the most effective use of protocol analysis can be made.

5.4.3 Multi-Dimensional Scaling Methods

The basis of scaling methods is to identify similarities among objects so that they can be grouped conceptually. The repertory grid, Easterby-Smith (1981), which has its roots in personal construct psychology, is probably the most well known of the scaling methods. The repertory grid method works by collecting a set of objects in the domain and presenting them to the expert in groups of three. The expert is asked to identify in what way two of the three are alike and different from the third. This process is continued until all possible groups of three have been considered. An example is given in the paper by Burton and Shadbolt (1987) :

> "As an example, if we were trying to analyse a domain of motor cars, we might choose a Porsche and a BMW as the two similar elements, and a Skoda as the dissimilar. We could then label our construct 'price'. Next time round we might choose a Rolls Royce and an Austin as similar elements, as opposed to a Porsche. This construct could be labelled 'country of origin'. By asking for many constructs we gradually build a map of the domain"

The grid developed through this process is analysed by cluster analysis. There are many variations on the repertory grid method, however all repertory grid methods take a long time to administer, analyse and interpret, even when there are only a small number of objects.

Other multi-dimensional scaling methods exist where elements or objects are rated on a series of dimensions. The analysis then reveals similarities, differences and clusters of objects. These other methods are complex and have not found wide acceptance as knowledge acquisition techniques.

Repertory grid methods are particularly useful where there are small number of closely related concepts and expertise is required to discriminate between them. Gammack and Young (1985) applied this method to elicit knowledge about different types of probability distribution and the extract below summarises their findings in this area:

> "The method first produced the 'objective' distinctions one might expect to find in textbooks, with such dimensions as 'continuous v discrete'. However it also gave more subjective, experientially-based criteria such as the dimension 'useful-in-modelling v common-test-statistic'. An hierarchical cluster analysis applied to the data yielded known families of distributions, such as the closely related F, gamma and log gamma distributions which were highly matched."

5.4.4 Concept Sorting

Concept sorting is applicable when there are a large number of concepts within the domain and some form of structure is required for them to become manageable. In basic terms, concept sorting works by initially establishing a list of the concepts required to cover the domain and then asking the expert to sort the concepts into different groups, describing what each group has in common. The result of this exercise is to enable the concepts to be structured in some hierarchical fashion.

The main difference between concept sorting and scaling methods is that concept sorting results in a structure or framework (meta knowledge) and scaling methods provide a way of discriminating between objects at a lower level.

5.5 Knowledge Acquisition in statistics

Gammack and Young (1985) suggested some appropriate elicitation methods for the different types of knowledge, using the domain of statistics as an example, but these assumed the knowledge engineer to be unfamiliar with the field of statistics. Much of the existing work in statistical expert systems has been undertaken either by statisticians or by people with at least a basic grounding in statistics. The consequence of this was that knowledge engineers were, to some extent, their own experts; and formulating a reasonable set of rules to incorporate technical expertise could be undertaken by a review process of their own knowledge and literature reviews. This is contrary to Nii's (1984) heuristic that the knowledge engineers cannot be their own experts. However, this has been possible, to some extent, in the area of statistics :

> "Expert data analysts have not sat down with trained knowledge engineers so that the latter could encode their expertise. Yet we seem to have made some progress, perhaps even considerable progress. Why? Part of the answer is that statisticians, or at least data analysts, are already in part knowledge engineers; what they do on a daily basis is to elicit and to apply private expertise from experts in a ground domain, using a collection of techniques, strategies, heuristics, and tools for doing so." (Thisted 1986)

Depending on the level of expertise of the knowledge engineer, a certain amount of professional expertise can also be incorporated in the knowledge base. The acquisition of the professional knowledge may be further facilitated by the use of more specific knowledge acquisition techniques and the possible methods are summarised in table II.

The balance between the use of review processes and the use of specific knowledge acquisition techniques depends on the knowledge engineer's level of expertise in the domain area. An academic base provides a good starting point for developing a reasonable prototype knowledge base containing technical expertise and some professional expertise. This knowledge base can then be

Table II : Types of Knowledge and Acquisition Techniques		
Type of Knowledge	Knowledge Elicitation Techniques	
Framework	Concept sorting Interviewing	
Concepts	Repertory Grid Interviewing	*
Procedural	Protocol Analysis	*
Heuristics	Protocol Analysis Structured Interviews	
Methodological	Sorting tasks Scaling methods	
Communication	Interviewing Protocol Analysis	

* Knowledge about concepts and procedural knowledge are primarily technical in nature and can thus be elicited through literature reviews. The acquisition of professional knowledge .in these areas is generally a case of verifying the correctness and completeness of the knowledge established in the literature reviews.

evaluated and modified by 'local experts'. The advantage of this approach is that while it still requires a certain level of commitment from local experts, it is far less time consuming than the conventional dialogue sessions. It also takes into account the variation both within and between application areas.

5.6 Knowledge Acquisition in THESEUS

The selected area of application for THESEUS was the analysis of data from experiments based on the completely randomised design; this incorporates One-Way Analysis of Variance and Multiple Comparisons. The reasons for this choice have been discussed in Chapter 2.

As the application area chosen is a small, well defined one the knowledge acquisition does not need to involve the 'framework' knowledge described above to any great extent but does involve all the other types. Each of the different types of knowledge involves both technical and professional expertise. Some types of knowledge such as procedural knowledge can be regarded as primarily technical in nature whereas knowledge about heuristics is mostly professional.

The knowledge acquisition for the prototype knowledge base of THESEUS was approached by using a combination of literature reviews, semi-structured interviews and workshops (a form of protocol analysis).

Once the prototype knowledge base had been built a process of evaluation and refinement was undertaken involving practicing statisticians. The first stage of the evaluation process was to evaluate the default knowledge base with respect to technical correctness, any problems encountered meant altering the default rulebase. The second stage of the evaluation was modification of the rulebase by practicing statisticians to include their own professional expertise.

5.6.1 Reviews

Literature reviews and small scale investigations were undertaken in order to establish a core of technical knowledge and to form a consistent and rational default rulebase. The review areas included the following :

- Hypotheses of interest to the client
- Choice of multiple comparison procedures
- Handling outliers
- Use of transformations
- Criteria used for checking assumptions

Members of the Statistics Research Group at Thames undertook to review different areas; the results of the review into selection of multiple comparisons procedures is given in Chapter 7. The selection of appropriate multiple comparisons procedures is predominantly professional expertise. However there are a large number of review papers which use simulation techniques to compare different methods in order to increase the technical knowledge in these areas. These review papers can be considered a formalised sorting method where the researchers have ideas about which methods are appropriate under which circumstances and are using simulation techniques to extend their knowledge in the area.

5.6.2 Interviews

A series of interviews with practicing statisticians was undertaken with the purpose of gaining a general insight into the thinking that guides the statistician and the heuristics used, rather than the precise elicitation of rules. Recognising that there is a considerable chance of leading experts into pre-conceived knowledge structures, the interview format was structured with the aim of allowing the expertise to flow unhindered. A loosely structured interview protocol was prepared to ensure that coverage of the relevant knowledge areas was complete while allowing the contributors to describe fully, in their own ways, their approaches to data analysis. The interview schedule covered such areas as attitudes to outliers, rigidity/flexibility on normality assumptions and homoscedasticity, use of transformations and the selection of test procedures.

Selecting statisticians from those who responded favourably in our initial postal survey of 57 statisticians, predominantly in the pharmaceutical and chemical industries and in research institutions, seven such interviews were undertaken. The information gathered demonstrates more than anything else the large variability between statisticians handling similar types of study. For example two statisticians, from different institutions, who present results to the same regulatory authority, have completely different approaches to the use of transformations. The one never uses transformations while the other regularly uses square root or logarithm transformations.

There was a distinct vagueness about multiple comparisons, with each statistician quoting his own favourite test, but being unclear about its use in relation to his client's hypothesis. None of the statisticians used any tests for normality; some justified this on the basis of sample sizes. At least one used the same argument for not investigating the problem of unequal variances. A feature which came through very markedly was the decision to keep everything a simple as possible in the interests of their clients' understanding.

5.6.3 Workshops

A series of statistical workshops was organised in which different approaches to the analysis of data sets, provided in advance, were presented and discussed. The participants in the workshops were members of the Statistics Research Group at Thames Polytechnic. All the data sets presented required a comparison between treatment groups; for example, comparing the weekly food consumption of rats in different treatment groups in a toxicology study.

The idea behind these workshops was to encourage the participants not just to analyse the data but to try and explain the way in which their decisions were made. It was also hoped that discussion between participants would help to identify reasons for any differences in approach.

Some of the approaches to analysis presented were chosen primarily on the basis of theoretical considerations; other approaches were chosen bearing in mind the clients' need to understand the analysis.

The discussions in the workshops highlighted several interesting aspects of the analysis of completely randomised designs. The effect of using the ANOVA as a preliminary screening test was discussed at some length; although this seems a reasonable approach, where it is not actually required it can cause unnecessary conservatism. The use of multiple range techniques is always a source of debate and there was no consensus of opinion about their validity. Decisions about normality and homoscedasticity usually relied on visual methods, with formal tests being occasionally employed where visual inspection was inconclusive. Any outliers were usually detected on Normal or Residual plots; where they were sufficiently extreme to cause concern, the data was often analysed both with and without the offending values.

The workshops were successful in initiating dialogue about different approaches to the analyses although participants rarely found time to write down their thoughts and conclusions after the discussions. Some notes were taken during the workshops but these were of necessity rather brief, conclusions were jotted down but it proved very difficult to keep a written note of the dialogues.

In retrospect, this form of introspective protocol analysis probably has greatest value in two areas. Firstly in understanding the different strategies used and where similarities exist between them. Secondly in dealing with unusual,specific situations it could be beneficial to use such workshops to identify appropriate ways of dealing with these situations. In order to gain the maximum information and benefit from the workshop sessions, it would probably be necessary to record them as well as taking notes.

5.6.4 Prototype evaluation and modification

The interviews, described in section 5.6.2, clearly showed that there are many possible approaches to any given analysis. The consequence of this is that the local experts need to understand sufficient about the knowledge representation and inference methods used to enable them to modify the knowledge base to their own specification.

The expert system was sent to a number of test sites where the collaborating statistician was asked to evaluate the prototype knowledge base and then to try modifying the knowledge base. These industrial trials are described in more detail in Chapter 9. Listings of the knowledge base used by the prototype systems are given in Appendix II.

This evaluation process is regarded as an important part of the development of the knowledge base, both in checking the technical core of knowledge and in incorporating professional expertise.

The next two chapters describe the core of technical knowledge that was established for the prototype knowledge base and are the results of some of the knowledge acquisition described in this chapter.

Chapter Six

Statistical Knowledge - I

Hypothesis Testing About Means

6.1 Introduction

In this chapter the nature of hypothesis testing for inferences about means and the criteria by which these tests can be assessed is discussed. The effects on different test statistic distributions of departures from Normal Theory assumptions is covered; some of the methods for detecting and correcting for such departures are given. Finally the approach chosen for the prototype system is discussed.

The theory covered in this chapter is relevant to many areas of statistics providing a technical core of knowledge and some pointers to the particular situations where professional knowledge plays an important part.

6.2 Hypothesis Testing

6.2.1 Introduction

Hypothesis testing is the process of inferring the truth of a hypothesis when data is obtained from a survey or randomised experiment. The actual data or sample, $\underline{x}$, that we have is regarded as being one of many possible samples that may have been obtained. The set of all possible samples that may have been obtained is the sample space, S. The data will be assumed to have been generated by a probability distribution of a specified form, but unknown exactly. The form of the distribution will be written $f(\underline{x},\theta)$, let the parameters, θ, considered belong to a parameter space, Ω. A statistical hypothesis will say that the data is actually generated by parameters within some subset w. The null and alternative hypotheses will be

$$H_0\colon \theta \in w \quad \underline{v} \quad H_1\colon \theta \in \Omega - w$$

where θ is the true parameter value generating the data. For example, if we want to test whether our data is from a Normal distribution with mean 17 and variance 1 against the alternative that it is from some other Normal distribution of variance 1 our question revolves around the single parameter μ. In this case Ω is the set of real numbers and w={17}; but we would usually write

$$H_0: \mu=17 \quad \underline{v} \quad H_1: \mu<>17$$

The classical problem of hypothesis testing is to test H_0 given the data and we must decide to accept or reject H_0 after examining the data. The set of all samples, S, is divided into two subsets

A : Those samples where we decide not to reject H_0

R : Those samples where we decide to reject H_0

Any particular test of H_0 amounts to a choice of the rejection region, **R**. There are many ways of choosing **R**, the first priority is usually to choose **R** so that the sample only has a small chance of occurring in **R** when H_0 is true, this is restricting the probability of a Type I error. A Type I error occurs if H_0 is rejected when it is true, a Type II error occurs if H_0 is not rejected when it is false. We try to choose **R** so that the probability of a Type I error, $p(\mathbf{R}/H_0)$, is at some small specified level, called the significance level, denoted by α.

Results of hypothesis tests are often expressed in terms of a P-value rather than a stated significance level. The P-value is the probability under the null hypothesis of obtaining a result equal to or more extreme than the test statistic calculated. The smaller the P-value is then the less likely it is that the null hypothesis is true.

There are many regions, **R**, with a given significance level, α, the problem is to decide on the 'best'. The concept of a 'good' or 'best' test is usually defined in terms of reducing the Type II error, or, equivalently, increasing the power. The power of a test is the probability of rejecting H_0 when it is false, (1 - p(Type II error)). Thus power in a test corresponds to sensitivity to a false H_0. The power depends on the actual parameter θ in H_1 model and a power function can be defined as follows

$P(\theta)$ = p(rejecting H_0 when the parameter is θ)

= $p(\mathbf{R}/\theta)$

For $\theta \in w$ then $P(\theta) = \alpha$

The Neyman-Pearson Lemma (Neyman & Pearson 1933), for testing simple hypotheses where the parameter space consists of only two values, tells us that the most powerful test, with significance level α, should be based on the likelihood-ratio. The likelihood function is the likelihood of the data observed given certain values of the parameters for the distribution of the data. The likelihood-ratio is the ratio of the likelihood functions for the observed data given the parameters specified by the alternative hypothesis and the null hypothesis. This gives some confidence in using likelihood ratio tests in more realistic problems.

To summarise, in hypothesis testing the first stage is the selection of appropriate hypotheses. It is sometimes possible to restrict the size of the parameter space Ω by imposing some restriction on the data from prior information. As an example, consider the one sample situation where the hypotheses are

$$H_0 : \mu = \mu_0 \quad \underline{v} \quad H_1 : \mu <> \mu_0$$

then Ω is the set of real numbers and $w = \{\mu_0\}$. However if it is know a-priori that the mean will be equal to or greater than the theoretical value then the alternative hypothesis becomes $H_1 : \mu > \mu_0$ and Ω is the set of real numbers greater than μ_0. Restricting the parameter space in this way can result in tests that are more sensitive for finding these more specific effects. However, a cautionary note, there is always the risk that the restriction made on the parameter space may not be valid. Thus, in an expert system it would be essential to ensure that any restriction required by a statistical method does actually hold.

Once the hypotheses have been selected then the statistician, or expert system, needs to decide on an appropriate test statistic. The choice of Normal Theory, Nonparametric or Robust procedures should be dependent on the nature of the data.

6.2.2 Properties of Hypothesis Tests

In applied statistics there are additional considerations to power (discussed in the previous section) when comparing different test statistics. Many tests use approximations to the distributions of the test statistic for simplicity, this means that the stated significance level, α, is also approximate. A test is said to be **conservative** if the true level of significance is less than that stated, in practice this means that a test is less likely to identify a true alternative hypothesis. Similarly a test is said to be **liberal** if the true level of significance is greater than that stated. This is a can be a more dangerous situation as it increases the chance of falsely accepting the alternative hypothesis i.e. detecting 'differences' that do not exist. The danger, or otherwise, of using a liberal test is dependent on the area of application. For example, in toxicology it is very important to detect differences that are present. The possibility of declaring some differences as significant when they are not is not so important. It is better to declare a compound toxic with an increased chance of being wrong than declare a compound safe when it may be toxic.

The sizes of samples can also have an important effect on the behaviour of a test-statistic. **Efficiency** is a relative term and is used to compare the sample size of one test with another under similar conditions. If the two tests have the same significance level and the same power when testing the same hypothesis then the relative efficiency is the ratio of the larger to the smaller sample size. It is also the case that as sample size increases then the power of a test, its ability to detect real differences, will also increase. The degree of improvement for a given increase in sample size also varies between test statistics. Thus it is possible to have two test-statistics, one of which performs better when the sample sizes are small and the other which performs better for larger samples. The power of both increase with increased sample size but the relative improvement for the latter test-statistic is greater than for the former.

The possibility of two kinds of error has already been discussed (Type I & II), however, Kimball (1957) proposes the concept of a Type III error. This type of error occurs when a false null hypothesis is rejected in favour of the wrong alternative and usually results from inadequate communication between the statistician and the client. This may be of particular concern in Statistical Expert Systems and so developers need to be aware of the dangers of providing the 'right' answers to the wrong questions. This situation could arise for two reasons. The system may not have sufficient understanding of the clients particular problem (i.e. selecting incorrect hypotheses of interest). The system may not be 'smart' enough to realise that the problem is not within the its scope and so tries to push the data into an analysis it does know about.

6.2.3 Different Types of Hypothesis Tests

Hypothesis tests can be divided into three main types, Normal Theory tests, Nonparametric tests and Robust tests. Normal Theory methods, which are usually based on maximum likelihood, likelihood ratio or some approximation

to one of these, are the most powerful methods provided certain assumptions hold. Thus Normal Theory methods are preferable to other methods when they can be used.
Two of the most important, and certainly the most studied, distributions associated with Normal Theory procedures are the t-distribution and the F-distribution.

The t-distribution is associated with tests related to sample means when the variances are not known, the standardized deviate is calculated using the estimated variance and this test statistic follows the t distribution. As degrees of freedom increase the t distribution tends towards the standard normal distribution. The t distribution is important where there are small samples because it adjusts the estimated variance by taking into account the sample size.

The F distribution is associated with inferences about variances, for example in Analysis of Variance. The F test statistic is a ratio of variances estimates which follows the F distribution and depends on the degrees of freedom for each estimate of the variance.

Difficulties arise when one or more of the assumptions are not true and it is in this situation that Nonparametric or Robust techniques may be preferred.

Nonparametric methods are usually based on either ranks or signs of the observations in the sample and have simple assumptions, more easily satisfied than those for Normal Theory methods. The majority of Nonparametric techniques require only that the observations actually have an underlying distribution. Some methods, notably those that depend on the signs of the observations also require that the underlying distribution be symmetrical. Hypothesis tests about means become tests of location in Nonparametric methods. There is a subtle difference here as hypothesis tests about means based on Normal Theory assume that the populations are Normally distributed; in the case of Nonparametric methods the only assumptions about the population distributions is that they exist. Thus it is possible in testing for location, using Nonparametric methods, to have a true null hypothesis where the populations come from completely different distributions but have the same location parameter.

Nonparametric methods are more widely applicable than Normal Theory methods and are most useful when some of the assumptions of those methods do not hold. Nonparametric methods can be applied when the data is Non-Normal or heteroscedastistic. They are also useful if there are outliers present and the experimenter does not want to exclude them from the analysis.

There is also a group of procedures based on 'robust' estimators. Robustness can be defined as signifying insensitivity to small deviations from the assumptions, where primary concern is concentrated on distributional robustness (Huber 1981). Robust estimators are much closer to the classical Parametric ideas than to the Nonparametric concepts, these robust procedures are often assessed in terms of their efficiency relative to the classical Parametric procedures. The median is an example of a robust estimator but its relative efficiency where the data is Normal is quite low in comparison with the mean. There are a number of different types of robust estimators denoted as M, L and R estimates. M estimates are maximum likelihood estimates; L estimates are based on a linear combination of order statistics; R estimates are derived from Rank tests.

In this project attention has focussed on the use of Normal Theory procedures for quantitative data. In certain circumstances nonparametric procedures may be more powerful, especially when some of the assumptions of the Normal Theory procedures do not hold and so they have been included as 'safety nets'.

6.3 Standard Normal Theory Assumptions

Most of the statistical procedures in common usage are based on statistical models which rarely hold true exactly. The standard assumptions for parametric or Normal Theory procedures can be summarised as follows :

1. The observations are a random sample from a Normally distributed population
2. Observations are independently distributed within samples
3. Where samples from two or more populations are being considered then it is necessary to assume that the population variances are equal

Chapter 10 of Scheffé (1959) considers in some detail the effects of departures from these assumptions. Subsequent simulation studies have sought to establish the degree of sensitivity to these assumptions. This is discussed in the following sections.

6.4 Non-Normality

There are two parameters that are usually used to describe the Non-Normality in distributions encountered in practice, namely skewness and kurtosis, for the Normal distribution these are both zero. For a distribution that is heavier in one tail than the other the coefficient of skewness is non zero, for example, the exponential distribution is positively skewed. Non zero kurtosis occurs when the tails of the distribution contain either more (positive kurtosis) or less (negative kurtosis) than the tails of the Normal distribution, the t distribution exhibits positive kurtosis.

For large samples, Non-Normality does not cause major problems because of the effect of the Central Limit Theorem which has the result that if X_i is a random variable with almost any mean μ_i and variance σ_i^2. the distribution of the sample mean is approximately Normal for large enough sample size. However, the size of sample required for the Central Limit Theorem to have sufficient effect will depend on the degree of Non-Normality (Miller 1986 p5-6).

6.4.1 Effect of Non-Normality on the t-test

The distribution of a sample mean, $\bar{x}$, tends rapidly with increasing n to $N(\mu,\sigma^2/n)$ where $E(\bar{x})=\mu$ and $V(\bar{x})=\sigma^2$, even for extreme Non-Normality. Skewness and kurtosis have no effect on the expected value of the sample variance, $E(s^2)$, but do have some effect on $V(s^2)$. However computer simulation has shown that the distribution of t statistic is only affected by extreme values of skewness and kurtosis (Pearson and Please 1975).

In the one-sample t-test the effect of Non-Normality on the P-value varies: for positive kurtosis the t-test becomes conservative and for negative kurtosis the t-test becomes liberal. The one-sided test is much more sensitive to the effects of Non-Normality than the two-sided test. Where sample sizes are small, the effect of Non-Normality is much more marked. Of course, defining what is meant by small is not that straightforward. It is context related and depends on the nature of the data, if data has more inherent variation then larger samples will be necessary. The decision about what constitutes a small sample is dependent on the domain and the statisticians own experience and judgement. Any expert system needs to be able to cater for this, preferably by allowing the local statistician to 'tune' the knowledge base accordingly.

In the two-sample situation, assuming equal variances and equal skewness and kurtosis between samples, Non-Normality has little effect, especially when the sample sizes are equal. In general the two-sample test is less sensitive than the one-sample test to Non-Normality. Where the sample sizes are not equal the effects are much the same as in the one-sample case. More serious distortion of the P-values can occur when the skewness of both samples is not the same; fortunately this does not seem occur too often in practice (Miller 1986 p43).

6.4.2 Effect of Non-Normality on the F-test

Lack of Normality has very little effect on the F statistic, even less than the two-sample case using the t statistic, again this has been verified by computer simulation (Pearson and Please 1975) who showed that the P-values are only distorted where there is extreme Non-Normality occurring in small samples. However if an experiment design is badly unbalanced having samples of very different sizes then skewness can affect the P-values.

6.4.3 Detecting Non-Normality

One of the simplest ways of detecting Non-Normality is by the use of Normal probability plots; data from a Normal distribution will give a straight line plot. If the distribution is skewed then the plot will show marked curvature at one end. If there is non zero kurtosis then the curvature will occur at both ends of the plot. Normal probability plots are very useful for giving the experimenter an idea of the nature of the data but obviously a decision about Normality based on these plots is subjective.

The Shapiro-Wilk test for Non-Normality has been shown to be one of the most effective tests available even for relatively small samples (Shapiro, Wilk and Chen 1968, Dyer 1974, D'Agostino & Stephens 1986 p 405)

6.4.4 Correcting Non-Normality

Although tests based on Normal Theory are robust for validity, they may not be the most powerful for non-Normal distributions and they are not necessarily the most efficient (Miller 1986 p81).

It may be helpful to try transforming the data to convert it to a sample that is approximately Normal. However some statisticians prefer not to transform the data as it is not always easy to interpret what the results on the transformed data actually mean. Normal probability plots are very useful as they can give an indication of a suitable transformation; for example, positively skewed positive data will often come closer to an underlying Normal distribution if a logarithmic or square root transformation is applied.

An alternative approach for handling Non-Normality is to use Nonparametric or Robust procedures, these have already been discussed briefly in section 6.2.3 .

6.5 Unequal Variances

As with Non-Normality, unequal variances have little effect on the t or F test statistics where the sample sizes are equal. In the case of the F test unequal variances may result in a slightly increased P-value. However where the sample sizes are unequal the effect is far more serious for both distributions.

If the largest variance is associated with the smallest sample then the P-values are reduced making the tests more conservative. However if the largest variance is associated with the largest sample the F test will become liberal, this is often more dangerous as it can result in increasing probability of a Type I error, i.e. claiming that there is a difference when the null hypothesis is true.

6.5.1 Detecting Unequal Variances

It is very difficult to decide whether or not the variances are equal, primarily because standard Normal Theory tests such as Bartlett's or Cochran's, are extremely sensitive to Non-Normality. However there are robust tests available, the most well known being Levene's test (Levene 1960). If there are several groups then plotting the standard deviations against the means should show up any relationship such as the variances increasing with the means.

6.5.2 Correcting for Unequal Variances

Transformations are very useful for correcting unequal variances, provided that there is some relationship between the means and the variances. The nature of the relationship between the means and variances can give a good indication of an appropriate transformation. For example, where variances are increasing linearly with the means then a logarithmic transformations may be most helpful; if the relationship is more curved then the square root transformation is a possibility.

However, where there is no discernable relationship between the means and variances then the application of a transformation is not likely to improve the variance heterogeneity. Nonparametric methods are useful when a transformation cannot be found or the experimenter does not want to use transformations, see section 6.8.

If the data is interval scale data or where there is no discernable relationship between the variance and the mean then nonparametric techniques are more appropriate.

6.6 Outliers

The possible presence of outliers needs to be considered carefully as there are several ways in which, if they are true outliers, they can violate the Normal Theory assumptions and affect the analysis of the data. Outliers can be defined as :

> 'An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data' (Barnett and Lewis 1984)

It is possible, and quite common, that human error or ignorance can result in incorrect recording of data, such mistakes can sometimes be traced and corrected. However where this is not the case an outlier may be an extreme value from the population that the sample has been drawn from or a contaminant value from another distribution. Deciding the origin of an outlier, however, is frequently impossible, there are many possibilities but no clear ways of discriminating between them.

6.6.2 Effect of Outlying Values

An outlier that is due to mis-recording and is not detected will distort both the mean and the variance of the sample, the variance is usually more severely affected, the extent of the effect depends on the sample size. This can disguise any treatment effects that may be present as well as causing some of the problems associated with unequal variances, see section 6.5.

Outliers that are extreme values or contaminants will cause similar problems but may also violate some of the Normal Theory assumptions. If the outlier or outliers are extreme values then it is possible that the assumption of Normality does not hold and the data actually comes from a different distribution. Where outlying values are contaminants then the assumptions that the observations are identically distributed is violated and this will seriously affect any inferences made because of the distortion of the mean and variance.

The presence of outliers, from whatever source, can obviously have a serious effect on the analysis of data and it is advisable to detect and deal with such values at the beginning of the analysis.

6.6.3 Detecting and Handling Outlying Values

There are two approaches to dealing with outliers, the use of procedures which can accommodate such values or the detection and possible removal of the outlying value.

Procedures which accommodate outliers are designed to draw valid inferences without being seriously affected by the presence of outliers. 'Robust' statistics, where robustness signifies insensitivity to small deviations from the assumptions (Huber 1981), can be very useful for handling data that may contain outliers; however they may not be particularly robust when the outliers are contaminants. Barnett and Lewis discuss in some detail both general robust methods and more specific accommodation procedures.

The second approach, of testing and possibly rejecting an outlier or outliers requires some criteria of relative discrepancy for deciding when an observation is an outlier. Visual methods, although relying on the observers judgement, can be very useful. Outliers will often show up clearly on a Normal plot, the presence of several apparent outliers on such a plot may indicate Non-Normality or a mixture of distributions. Plots of fitted against observed values are also useful in showing possible outliers. There is a multitude of tests available for testing extreme values, see Chapter 6 of Barnett and Lewis, the more well known methods include Dixon's and Grubb's methods.

If some observation has been classified as an outlier, either by visual inspection or the application of some test procedure (or a combination of both), the experimenter has to decide what to do with the value.

Erroneous measurement or miscalculation is the easiest to handle as it can sometimes be traced and either remeasured or the observation scrapped. Where this is not possible, or where the outlier is an extreme value or contaminant, then the experimenter has a range of options open which include treating the outlier as a missing value or using robust or nonparametric methods.

6.7 Dependence

There are two main types of dependence which can arise in the applications considered here. The first type of dependence is that caused by blocking effects. This can occur when the data has been collected in sub-groups; for example, the data may have been collected on different days. Such factors are referred to as nuisance factors and may have no effect at all but this cannot be assumed. If the blocks are unbalanced, for example if more observations are collected one particular day, then the error variance will be distorted. The easiest and most effective way of detecting and dealing with such block effects is to remodel the design into a higher way classification.

The other main type of dependence can come from a sequence effect either in time or space. If observations are taken serially in time then observations close together in time may be stochastically dependant. Similarly, observations that are taken from physically adjacent or close sites may be dependant because of some local effect or even interaction between sites.

The presence of serial correlation in data has a substantial and serious effect on both Normal Theory and nonparametric procedures, greatly distorting the P-values. It is possible to test for serial dependence by calculating the serial correlation and plotting pairs of observations; for example, plotting the pairs (y_i, y_{i+1}) to check for sequence effect of lag 1. Little is known about correcting for serial dependence. It is possible, where there are only one or two groups, to substitute the correlation coefficient in the expressions for the variances, provided the samples are large enough (Miller 1986 p36,63).

In the context of expert systems the facility to detect a need for using a higher way classification should be considered in the design. For example, if the observations have been collected in blocks such as days or by location, then it may be worth remodelling the experimental design to take these block into account.

6.8 Assumption Checking in THESEUS

In the prototype version of THESEUS attention was concentrated on checking for outliers, Non-Normality and heteroscedasticity. Checking and correcting for dependence beyond remodelling the design if it is suspected, is difficult and was not incorporated in the prototype. There is a facility to view the data, which includes Normal plots, and is available at any stage of the consultation. This facility is provided to assist the user in making decisions about the nature of the data such as checking for Non-Normality or looking for possible outliers.

6.8.1 Outliers

The procedure for detecting and handling outliers or extreme values is fairly simple in the first stage of THESEUS. If each treatment group has more than 25 observations then the decision about outliers is left entirely to the user. For smaller sample sizes, Dixon's test is run and the user is then asked to make a decision based on the outcome of the test and the users own knowledge of the data.

6.8.2 Normality

Although Non-Normality is not regarded as a particularly important problem it is checked anyway, the user being given the option of overriding any decision the system might make. For very small samples (less than 10 observations overall) the decision is left to the user. For large samples (more than 25 observation overall) the Shapiro-Wilk test is run on each group separately; for smaller samples the observations are treated as a single group. In all cases the Shapiro-Wilk test is run on standardised values of the form

$$\frac{\text{observed value - group mean}}{\text{variance}}$$

If the observations are treated as a single group then the standardisation uses the estimate of variance from the ANOVA if the variances are equal and the individual group variances otherwise.

6.8.3 Homoscedasticity

It is usually easier to correct for suspected heteroscedasticity than it is to test for it (Miller 1986 p92). However, two tests have been incorporated to help the user make a decision. The variances are only declared equal by the system if Bartlett's test at 1% and Levene's test at 5% do not show evidence of unequal variances. If either test does show some evidence then the user is asked whether they wish to override this evidence or not.

6.9.4 Transformations

If the data has been found to be Non-Normal or to have unequal variances then the user is asked whether they are prepared to try a transformation. If the user is opposed to the use of transformations then one of the nonparametric methods will be recommended.

If a transformation is to be undertaken then the user is offered a list of possibilities to choose from. The system transforms the data and then repeats the decision process described above to see if the transformed data is satisfies the Normal Theory assumptions. If a transformation has not been successful then other transformations can be tried if the user so desires. If no suitable transformation can be found then the Normal Theory methods will be rejected in favour of nonparametric methods.

Chapter Seven

Statistical Knowledge - II

Analysis of One-Dimensional Data

7.1 Introduction

Chapter 6 provided an overview of the concepts relevant to hypothesis testing about means; this chapter reviews specific statistical methods appropriate to the analysis of data where there is one, two or several treatment groups. The discussion is limited to quantitative data from studies where the interest is in comparisons between the means of the treatment groups. Attention has been concentrated on the Normal Theory methods; nonparametric methods have not been considered in detail, but have been included as they can often be used where the Normal theory methods cannot. The aim of this review is to supply sufficient information for the development of a rational prototype knowledge base for a statistical expert system. Where Normal Theory methods are discussed, only assumptions which are additional to those specified in the previous chapter are stated. Assumptions relevant to Nonparametric methods are stated as each method is discussed.

Notation

x_{ij} is the jth observation from group i
$i = 1(1)t$ where t is the number of treatment groups
$i = 0(1)t-1$ if there is a control group present
$j = 1(1)n_i$ where n_i is the number of observations in group i

$\bar{x}_i$ is the mean for group i
μ_i is the population mean for group i
s_i is the standard deviation for group i
σ_i is the population standard deviation for group i
N total number of observations (Σn_i)
s_p pooled estimate of the common standard deviation

In the single sample case the subscripts for treatment groups are dropped.

7.2 Analysis for a Single Sample

In this situation the researcher is interested in finding out whether or not the mean of the data differs from some hypothesised value. The likelihood ratio test of $H_0 : \mu = \mu_0$ vs $H_1 : \mu <> \mu_0$ leads to Student's t statistic

$$\frac{(\bar{x} - \mu)}{s\surd(1/n)}$$

which has Student's t distribution with n-1 degrees of freedom. With increased sample size the t statistic tends towards a Normal distribution due to the effect of the Central Limit Theorem.

Where some of the assumptions have been violated it may be possible to use a nonparametric test. The Wilcoxon signed rank test, where the differences (observations - hypothesised value) are ranked according to their absolute magnitude, can be used in the one-sample situation. The test statistic is

$$SR_+ = \sum_i r_i I\{z_i > 0\}$$

where
r_i = rank of absolute value of the ith observation
$z_i = y_i - \mu_0$

$$I\{z_i > 0\} = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0 \end{cases}$$

The probabilities $p\{SR_+ = r\}$ can be generated through recursive schemes, tables are readily available for samples of up to 20 observations. For larger samples a normal approximation can be used

$$\frac{SR_+ - [n(n+1)/4]}{\surd[\ n(n+1)(2n+1)/24\ ]}$$

The only assumptions required for this test are that the data is a random sample from a continuous, symmetric distribution and that the observations are independently distributed.

7.3 Analysis for Two Samples

Where there are two treatment groups the experimenter usually wants to compare the two groups in order to detect whether there is any significant difference between them. Under the condition of equal variances the likelihood ratio test of $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 <> \mu_2$ leads to the t statistic

$$t = \frac{(\bar{x}_1-\bar{x}_2)\surd(n_1n_2)}{s_p\surd(n_1+n_2)}$$

where s_p is the pooled variance calculated using

$$s_p = \frac{(n_1-1)s_1^2+ (n_2-1)s_2^2}{(n_1 + n_2 - 2)}$$

The assumption of equal variances is a rather severe one and where this is the only assumption violated one possible approach is to use the approximate Aspin-Welch statistic

$$t' = \frac{\bar{x}_1 - \bar{x}_2}{\surd(s_1^2/n_1 + s_2^2/n_2)}$$

with the degrees of freedom calculated using the approximation

$$\frac{s_1^2/n_1 + s_2^2/n_2}{s_1^2/n_1(n_1-1) + s_2^2/n_2(n_2-1)}$$

For both the t statistic and the Aspin-Welch approximation, the t distribution tends towards the standard Normal distribution as sample size increases.

The nonparametric Wilcoxon rank test can be used where some of the Normal Theory assumptions do not hold. The only assumptions required are that the data are random samples from a continuous distribution and that the observations are independently distributed. The Wilcoxon statistic can be calculated in more than one way, the Mann-Whitney form is given here

$$U = \sum_{i} \sum_{j} I\{y_{1i} > y_{2j}\}$$

$$i = 1,2,..n_1 \quad j = 1,2,..n_2$$

where

$$I\{y_{1i} > y_{2j}\} = \begin{cases} 1 & \text{if } y_{1i} > y_{2j} \\ 0 & \text{if } y_{1i} < y_{2j} \end{cases}$$

Tables are available which give the probabilities associated with values as small as U. For large samples a normal approximation can be used

$$U^* = \frac{U - (n_1 n_2/2)}{\surd[\ n_1 n_2(n_1 + n_2 + 1)/12\ ]}$$

which has an approximately Standard Normal distribution.

7.4 Analysis for Several Groups - Overall Test

Where there are several treatment groups the simplest type of experimental design or layout is the completely randomised designed where treatments are randomly allocated to experimental units. This one-way design is very flexible, allowing any number of treatments and any number of replicates, although the number of replicates should only be varied with good reason as this can affect subsequent analysis. Analysis of the one-way design is straightforward, even with unequal replication or missing data. The loss of information due to missing data is smaller than with any other design because of the relatively large degrees of freedom associated with the error term in the ANOVA.

The major disadvantage of the completely randomised design is that any variation between experimental units is not considered separately from the experimental error. The error can be reduced by using a different design if the experimental units can be handled in groups. For example, in field testing of new varieties of crop there may be a great deal of variation between plots in a field because of different drainage characteristics. In this sort of situation the randomised block design where the plots in the field are divided into blocks and treatments are randomly allocated to plots within each block is useful.

The completely randomised design is most useful in laboratory experiments where the material or units to be tested are homogeneous and so a higher way design is unnecessary. It is also very useful where an appreciable number of missing values may occur because of the easy extension to unequal sample sizes and the large degrees of freedom associated with the error term in the ANOVA. Small scale investigations, where using a more complex design would reduce the error degrees of freedom and so reduce the sensitivity of the experiment, can be analysed using the completely randomised design.

The model for the completely randomised design can be expressed as :

$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$

μ = overall mean

$\mu_i = \mu + \alpha_i$ denotes the mean of the ith population

ϵ_{ij} is the random or unexplained variation

The parameters are constrained by $\Sigma(n_i\alpha_i) = 0$

Where there are several treatment groups, the experimenter is usually interested in constructing point and interval estimates for the group means or in testing hypotheses about these means.

The likelihood ratio approach leads to the standard one way analysis of variance

Source of Variation	Degrees of Freedom	Sum of Squares
Treatments	t-1	$\Sigma n_i(\bar{x}_i - \bar{x})^2$
Error	N-t	$\Sigma\Sigma(x_{ij} - \bar{x}_i)^2$
Total	N-1	$\Sigma\Sigma(x_{ij} - \bar{x})^2$

The significance test of the hypothesis that all the α_i are equal is undertaken by referring the ratio of the mean treatment sum of squares and the mean error sum of squares to F tables on (t-1,N-t) degrees of freedom.

There is often misplaced interest in this significance test, as it is often known a-priori that the treatment effects cannot all be equal. What is more important is to see where the differences between the treatment effects lie; the issue of multiple comparisons is dealt with in some detail in the following sections. The ANOVA table is useful because it gives a summary of the data, showing the amount of variation attributable to the treatment effects. The ANOVA table also supplies an estimate of the pooled variance which is usually required in procedures used to assess possible structures between treatments.

The Kruskal-Wallis one-way analysis of variance nonparametric method can be used in situations where the Normal Theory assumptions do not hold. The Kruskal-Wallis test requires only that the data are random samples from a continuous distribution. The test is carried out by ranking all the samples from the smallest to the largest in a single series. The rank sums for each treatment groups are calculated (R_i), each R_i has a limiting Normal distribution and so the statistic

$$H = 12/(N(N+1)) \sum_j [R_j^2/n_j \quad - 3(N+1)]$$

is approximately distributed as a Chi-square with t-1 degrees of freedom provided none of the treatment groups are too small. Special tables are required for small samples.

7.5 Introduction to Multiple Comparisons

As already mentioned, the experimenter will rarely be satisfied with a simple statement about whether some difference between treatment means exists but is interested in finding out where such differences arise. Thus the experimenter may wish to make a number of statements about the treatment groups, hence the term multiple comparisons. O'Neil and Wetherill (1971) state that there is still much confusion as to what the basic problems of multiple comparisons are, what the various procedures achieve and what properties should be considered!

As an introduction to the issues involved in multiple comparisons, consider a situation where there are two means. If the experimenter constructs 95% confidence intervals for each mean then each interval has a probability of 0.95 of including the corresponding true population mean. However the joint probability that both intervals simultaneously contain their respective population means is (0.95x0.95) iff the two means are totally independent. If there is some dependence between the means then the joint probability is greater than or equal to 1 - (1-0.95) - (1-0.95) (this follows from Boole's inequality $P(A \cup B) <= P(A)+P(B)$). In multiple comparisons these two confidence intervals could be considered a family of statements and the aim of multiple comparisons methods is to control the joint probabilities, under the null hypothesis, for such families.

There is considerable debate between statisticians about whether or not this is an appropriate approach, especially when the null hypothesis is almost certainly false (Nelder commenting on O'Neil & Wetherill 1971); this is discussed further in section 7.7. Even where statisticians consider the general principle in multiple comparisons of controlling the joint probabilities to be acceptable, there is still much debate about exactly what constitutes a family.

7.6 Error Rates and Families

Consider a family of statements $F = \{ S_f \}$ where $N(F)$ is the number of statements in the family and let $N_w(F)$ be the number of incorrect statements in the family. The error rate for the family is

$$Er\{F\} = \frac{N_w(F)}{N(F)} \qquad \text{(assume } N(F) \text{ is finite)}$$

The error rate is a random variable whose distribution depends on the multiple comparisons procedure used and its underlying probability structure. Thus to assess the overall merit of a multiple comparisons procedure some global, non-random parameter of the distribution of the error rate must be selected. The two criteria most commonly used are the probability of a non zero family error rate and the expected family error rate.

7.6.1 Probability of a non zero family error rate

Many of the multiple comparison methods available control this error rate, it is often called the experiment-wise error rate in the literature. It is denoted by

$$\begin{aligned} P(F) &= P(\ N_w(F)/N(F) > 0\ ) \\ &= P(\ N_w(F) > 0\ ) \end{aligned}$$

There is no distinction here between families with only one incorrect statement and families with one or more incorrect statements. As the family size increases then the greater the probability becomes that one of the statements will be wrong; thus the probability associated with each statement will have to be smaller in order to maintain the required overall level for the family. As this probability error rate creates an all or nothing situation for families, great care should be given to just what constitutes a family, see section 7.6.3.

The Bonferroni inequality gives a bound on P(F) related to the individual statement probabilities

Let $\alpha_f = P(I(S_f) = 1), f = 1,2,\ldots,N(F)$

and $I(S_f) = \begin{cases} 1 & \text{if } S_f \text{ is incorrect} \\ 0 & \text{if } S_f \text{ is correct} \end{cases}$

then $1 - P(F) >= 1 - \alpha_1 - \alpha_2 - \ldots - \alpha_{N(F)}$

That is $P(\bigcap_f [I(S_f) = 0]) >= 1 - \sum_f P(I(S_f) = 1)$

This expression becomes an equality when the statements are independent.

7.6.2 Expected Family Error Rate

The expected family error rate or comparison-wise error rate is denoted by

$E\{F\} = E\{ N_w(F)/N(F) \}$ assuming finite N(F)

This error rate is directly related to the marginal performances of each of the statements in the family.

Let $\alpha_f = P(I(S_f) = 1) = E\{ I(S_f) \}, f=1,2,\ldots,N(F)$

then $$E\{F\} = \frac{\alpha_1+\alpha_2+\ldots+\alpha_{N(F)}}{N(F)}$$

Statements can be grouped together in a family where their dependence is difficult to assess, and the family's expected error rate will be known exactly from the behaviour of the individual statements. Thus if an overall significance level, p, is required for the family the procedure to be used must be constructed so that each statement has a probability 1-p. In fact any combination for which $\alpha_1+\alpha_2+...+\alpha_{N(F)} = pN(F)$ will result in the appropriate error rate.

Where the number of statements in the family is 1 then the expected error rate and the probability error rate are equal. Without any knowledge about the structure of the dependence between the statements, the only relation between the two error rates is given by

$E\{F\} <= P(F) <= N(F) . E\{F\}$

The expected family error rate gives exact results for combining dependent statements into one family, however, the probability error rate has a reasonable bound, shown in the Bonferroni inequality, where the number of statements is small. The advantage of using the probability error rate is that it provides a known degree of protection for the entire family and an upper bound on the expected proportion of mistakes.

7.6.3 Families

The concept of what constitutes a family is very subjective. The two extremes are to consider each statement a family or to consider all statements made over a lifetime a single family.

The basic premise of simultaneous statistical inference is to give increased protection to the null hypothesis. Yet it is not always the null hypothesis which is true, and attention must also be given to the error rates under the alternative hypothesis by considering the power function of the test. However it is an inescapable fact that as the error rates are forced down in one direction they must increase in the other. i.e increased protection of the null hypothesis results in decreased power and vice versa.

The introduction of families further complicates this issue. As family size increases then confidence intervals widen and the power is reduced. To increase power, either the size of family must be reduced or sample size or error rate increased.

Miller (1981) states that he usually considers a family to be the individual experiment of the researcher, which could include, for example, a two-way classification analysis of variance, comparison of a half-dozen mean values and perhaps a regression analysis. Included in this is the requirement of reasonable power against reasonable alternatives with reasonable protection for the available sample size. An individual experiment means a related group of observations collected through an autonomous experiment and whose analysis will fall into a single mathematical framework. There are no hard and fast rules for where the family lines should be drawn, and the statistician must rely on his own judgement for the problem at hand.

7.7 Controversy Over the use of Multiple Comparisons

There is considerable debate among statisticians about whether multiple comparison methods should be used at all and the paper by O'Neil and Wetherill(1971) with the subsequent discussion is a good example of the controversy over their use. O'Neil and Wetherill recommend the use of multiple comparisons where the problem is one of fundamental exploration with the aim of discovering the underlying mechanism affecting the results. Some statisticians maintain that such fundamental exploration, where there is no prior pattern, is best approached using other methods (see Placketts and Nelders response to the O'Neil and Wetherill paper). In addition there is concern over the use of methods which are designed to protect against incorrect rejection of the Null hypothesis when it almost certainly the case that the Null hypothesis is false anyway and Type II errors are far more likely to occur.

Chew (1976), although advocating the use of multiple comparison methods, begins by clearly stating some of the abuses of these methods. In the case of the completely randomised design the main abuse is to apply multiple comparison techniques where the treatments are different levels of the same treatment. In this case regression analysis or curve-fitting would seem to be more appropriate. Even this is open to some debate as the experimenter may be more interested in finding out the lowest level at which there is a response (Williams 1971).

7.8 Classification of Multiple Comparison Methods

The majority of multiple comparison methods are based on the following basic techniques or inequalities.

Repeated Normal Statistics : For σ^2 unknown these are separate t tests

Maximum Modulus Method : This method involves finding the constant c such that

$$P(\max[|Y_1|, |Y_2|] <= c) = 0.95$$

Y_i are independent and Normally distributed with means μ_i and variance = 1

The condition of independence means that the constant c is given by the $1-(1-0.95)^2$ percentage point of the Normal distribution. When σ^2 is unknown the t-distribution is used.

Scheffès Chi-squared Projections : these are based on the Chi-squared statistic $Y_1^2 + Y_2^2$. Intervals are obtained by projections of the bivariate Chi-squared region

Multiple Modulus Method : This is an extension of the maximum modulus method and are performed by testing in successive stages. The effect of multiple modulus tests is to enlarge the decision regions for means different from zero at the expense of the situations where one of the means is zero.

Bonferroni Inequality : This has already been stated in section 7.6.1

Sidàks Multiplicative Inequality : (Sidàk 1967) Let Y = $(Y_1, Y_2, \ldots, Y_k)$ be the vector of random variables having the k-dimensional normal distribution with zero means, arbitrary variances σ_1^2, σ_2^2, ..., σ_k^2, and an arbitrary correlation matrix R $=\{p_{ij}\}$. Then for any positive numbers $c_1, c_2, \ldots, c_k$

$$P(|X_1| <= c_1, \ldots , |X_k| <= c_k) >= \prod_i P(|X_i| <= c_i)$$

Sidàks Uncorrelated-t Inequality : (Sidàk 1967) This is related to the multiplicative inequality above but with independent X (i.e. zero correlation). Suppose that s is a positive random variable, independent of X_i then

$$P(|X_1|/s <= c_1, \ldots, |X_k|/s <= c_k) >= \prod_i P(|X_i|/s <= c_i)$$

There is a wide range of methods available and these can be classified according to the hypothesis of interest to the experimenter. The different hypotheses are classified below and the main multiple comparison methods available are briefly introduced. Nonparametric methods have not been considered in detail in this review but are included as they can sometimes be applied in situation where Normal Theory methods cannot.

Pairwise (No control) : The experimenter may be interested in a small number of pairwise comparisons between treatment groups or all possible pairwise comparisons. Much of the work undertaken in multiple comparisons has concentrated on pairwise comparisons. The Least Significant Difference (LSD) and Protected LSD (Fisher

1935) are based on the repeated Normal statistics, the latter requires a preliminary significant F test before any comparisons are made. The LSD and PLSD control the comparison-wise error rate.

The Tukey test (Tukey 1952), which can be extended for contrasts, is based on the Maximum Modulus Method and uses the Studentised Range tables, however it requires equal replication and equal variances and thus many adaptations have been proposed to deal with these problems. The Tukey-Kramer test (Kramer 1956)is a straightforward adaptation of the Tukey test for unequal replication. Other extensions of the Tukey test for handling unequal sample sizes were proposed by Spjøtvoll & Stoline (1973), Genizi & Hochberg(1978). Hochberg (1974) and Gabriel(1978) also proposed similar methods based on Sidàks multiplicative inequality.

A number of methods have been proposed for the case of unequal variances. Games and Howell (1976) suggested a method which uses the Studentised range and an approximation for the degrees of freedom, Welch (1938). Tamhane's (1979) method uses Students-t distribution and is based on Sidàk's multiplicative inequality. A further test, T3 was proposed by Dunnett (1980b) as an adaptation of Tamhane's procedure based on Sidàk's uncorrelated-t inequality. Dunnett (1980b) also proposed a method, C, which is based on the weighted average of Students-t suggested by Cochran(1964).

In addition to the methods already described there are two multiple range methods, Duncans (Duncan 1955) and Newman-Keuls (Newman 1939, Keuls 1952), which are based on the multiple modulus method. The error rate for these tests is rather difficult to define as it is neither comparison-wise nor experiment-wise. The error rates are controlled for each subset of means being considered.

There are two Nonparametric methods which can be used for testing pairwise comparisons, the Steel-Dwass test (Steel 1960, Dwass 1960) and the Kruskal-Wallis test (Nemenyi 1963). The Steel-Dwass test is based on pairwise rankings and requires equal replication and special tables. The Kruskal-Wallis test is based on ranking across all treatment groups, it does not require equal replication and uses the Studentised Range tables. The Kruskal-Wallis test is very versatile as it can also be used as a Nonparametric analog to the One-Way ANOVA, using the Chi-squared tables when the samples are large enough, as well as comparisons with a control, using Dunnett's tables.

Contrasts (No control)

The most commonly used contrasts are linear contrasts of the general form $\Sigma c_i \bar{y}_i$ where $\Sigma c_i = 0$, however it is possible to test non-linear contrasts such as quadratic or polynomial contrasts. In this review only linear contrasts are considered. Scheffé's (1953) method based on his F projections uses the F tables and can be adapted for unequal sample sizes. Brown & Forsythe (1974) proposed a further adaptation for the case of unequal variances. A method based on the Bonferroni inequality (Miller 1981 p67) which uses Student's-t distribution can also be used for testing linear contrasts. The t values are required at significance levels not usually available in standard tables and Dunn(1959) computed necessary values.

Note : If the experimenter wishes to test designed contrasts, that is, contrasts decided on before the experiment, then orthogonal F tests will be the most powerful and should be used where possible.

Comparison with control : When a control group is present there are two different situations

i) Treatment groups are different levels of a single factor, for example, different dose levels of a drug. The experimenter may wish to test for monotonic ordering or to find the lowest dose for which there is a response or possibly to fit a response curve. The latter requires regression techniques but certain specialised multiple comparison methods are available for the other two possibilities.

Bartholomew (1961) proposed a method based on Maximum Likelihood estimates used to test for monotonic alternatives. Williams (1971, 1972) suggested a more specific technique, also based on Maximum Likelihood estimates, for finding the lowest dose at which there is a response. Shirley (1977) proposed a Nonparametric analog to Williams' test which uses the tables developed by Williams.

ii) Treatment groups are different factors; for example, different varieties of a crop. In this situation the experimenter is usually interested in comparing each treatment group with the control.

Dunnett (1955) proposed a test for pairwise comparisons with a control group which requires equal replication. The statistic is a multivariate analog of the t distribution and special tables are required. The Many-One Rank method (Steel 1959) provides a Nonparametric version of Dunnett's test. The Kruskal-Wallis Nonparametric method can also be adapted for comparisons with a control group.

7.9 Simulation Studies

In order to compare the different multiple comparisons procedures properties of power and robustness as well as the conservativeness of the procedure should be considered, Stoline(1981). Practical issues such as ease of use and availability of tables are also important. A great deal of research has been undertaken in studying the robustness and power of the F-distribution and the t-distribution (see Chapter 6) to departures from Normal Theory assumptions. However, little is known about the robustness and power of the Studentised Range, the Studentised Maximum Modulus or the Many-one t statistics. Practically no work on the robustness and power of these statistics has appeared in the literature (Miller 1981 p 102,108). Simulation studies, which are empirical investigations into the behaviour of the different methods under different conditions, provide a very useful way of comparing techniques.

Due to the large numbers of papers on the subject of multiple comparisons, attention has been focussed on the review and simulation papers. Original methodology papers are only referred to where methods have not been included in simulation studies. This section provides an overview of some of the simulation papers.

7.9.1 Carmer & Swanson 1973

Carmer and Swanson compared ten multiple comparison methods for pairwise comparisons. The Type I, Type II and Type III error rates and the correct decision rates were compared. Type III error rates were defined by Carmer and Swanson as the probability of declaring one treatment superior to another when the reverse is true. The methods compared by Carmer and Swanson were the Least Significant Difference (LSD), protected LSD (Using preliminary F at 0.01, 0.05, 0.10 significance levels), Tukey, Newman-Keuls, Duncan, Scheffé and two Bayesian approximations attributed to Waller and Duncan (1969). Data for 1000 Completely Randomised Block experiments were generated for each of 88 combinations of 22 means and four different numbers of replications.

Conclusions

In the conclusions Carmer and Swanson state that although Scheffé, Tukey , Newman-Keuls and the PLSD with a preliminary F test at 0.01%, all provide excellent protection against the Type I errors they are rather conservative, and the ability to detect real differences should have a high priority. The LSD and PLSD with preliminary F test at 0.1% do not give sufficient protection against Type I error. The choice between the remaining procedures is not easy, Duncan's method gives better protection against Type I errors but is less sensitive in detecting real differences than the two Bayesian approximations or the PLSD with a preliminary F test at 0.05%.

Comments

Referring to Carmer and Swanson's Table 3 of observed comparison-wise and experiment-wise Type I error rates, it can be seen that the Bayesian approximations and Duncan's methods control the comparison-wise error rates adequately but not the experiment-wise error rates. If the statistician wishes to control the experiment-wise error rates then in fact the Tukey methods seems, from Table 3, to give the best protection against Type I errors without becoming liberal.

7.9.2 Thomas D.A.H. 1973

The simulation study reported by Thomas compared several methods for pairwise multiple comparisons as well as four methods for constructing confidence intervals about a single mean. The pairwise comparison methods compared were the Protected Least Significant Difference (PLSD), Tukey, Scheffe, Dunn (Bonferroni), Newman-Keuls and Duncan. A non significant F value precluded further testing except for Dunn's method and Duncan's method. The methods were carried out on sets of 5, 10 and 20 means each of four results.

Conclusions

Thomas concluded that the PLSD gives insufficient protection to the null hypothesis. Duncan's test was preferred because it gave adequate protection against Type I errors but was less conservative than the other methods.

Comment

The undue conservatism noted by Thomas for some methods could be related to the use of a preliminary F-test as a filter. Performing a preliminary F test may miss important single effects that get diluted (averaged out) with other effects (Dunnett and Goldsmith 1981).

7.9.3 Tamhane 1979

In this study, Tamhane compares procedures for multiple comparisons in the equal and unequal variance case. The methods reviewed included procedures proposed by Spjøtvoll, Hochberg, Ury and Wiggins, Games and Howell, three proposed by Tamhane, Brown and Forsythe and finally Spjøtvoll and Stoline. The sampling experiments were conducted for all pairwise differences of the means for sets of 4 and 8 means. Selected contrasts for the set of 8 means were also considered. The sample sizes ranged from 7 to 13. For each set of treatment means eight (σ^2,n) configurations were studied, 1000 experiments were run for each configuration.

Conclusions

Tamhane concluded that the Tukey procedure and Hochberg's procedure are robust and conservative for pairwise comparisons in the equal variance case. In the unequal variance case the Games and Howell procedure gives the shortest intervals but can be liberal. One of the Tamhane procedures gives slightly wider intervals than the Games and Howell but does not suffer from liberality. Where contrasts are required the Brown and Forsythe procedure was recommended. The Brown and Forsythe method is based on the Scheffe projections adapted for unequal variances.

7.9.4 Dodge and Thomas D.R. 1980

This simulation study is particularly interesting because it included nonparametric procedures in the comparison. The Normal Theory methods considered were the LSD, PLSD, Tukey, Duncan, Newman-Keuls, Scheffé and the Bonferroni method. The nonparametric methods were k-sample ranking or pairwise sample ranking analogues of the Normal Theory methods. The simulation considered five different scale-location parameter families (Uniform, Normal, Logistic, 4th power and Extreme value); it did not include the unequal variance situation. Independent sets of 1000 trials were generated for each of 32 different combinations of numbers of treatment groups and numbers of equal pairs between treatment groups.

<u>Conclusions</u>

The Normal Theory procedures were found to be robust with regard to Type I error rates. The k-sample ranking procedures were considered to be extremely conservative, hence methods based on pairwise rankings were preferred. If strict control of experiment-wise error is regarded as essential then the LSD, PLSD and multiple range methods should be rejected. The Scheffé method was found to be more conservative than the Bonferroni or Tukey methods.

7.9.5 Dunnett 1980a

This is the first of a pair of papers on pairwise comparisons and considers the equal variances, unequal sample size situations. Methods proposed by Spjøtvoll and Stoline, Hochberg, Gabriel, Genizi and Hochberg and Tukey-Kramer were compared. Millers suggestion of using the harmonic mean was also included. The simulation was in two parts, the first of which was to calculate the error rates for the Tukey-Kramer intervals for varying sample sizes. Sets of 4, 6 and 10 means were considered and 10,000 simulations undertaken for each configuration. The second stage of the simulation was to consider a set of 6 treatment means and varying sample sizes. 25,000 simulations were undertaken for each combination and the different procedures were compared.

Conclusions

The results of Dunnett's simulation clearly show that the use of the harmonic mean resulted in inflated α values as soon as the ratio of sample sizes moves out of the range 0.25 to 1.25. Gabriel's procedure was also found to be liberal although it performed better than the harmonic mean, only becoming liberal if the sample size ratio was more than about 8. All the other methods were conservative with the Tukey-Kramer method giving the levels closest to $\alpha = 0.05$ and so providing the shortest intervals. This simulation study put to rest fears about the approximate nature of the Tukey-Kramer methods showing that adequate protection is given to the null hypothesis.

7.9.6 Dunnett 1980b

This simulation study dealt with the case of unequal variances. The Games and Howell method and the Tamhane procedure which came out the best in Tamhane's 1979 study were compared along with two newer methods denoted C and T3. For the simulation, sets of 4 and 8 treatment means were chosen with equal replication. For the set of 4 means, some unequal sample sizes were also included. Each configuration of different variances was simulated 10000 times.

Conclusions

The results of Dunnett's study showed that the Games and Howell procedure can be liberal and that the T3 intervals are always shorter than the T2. For large degrees of freedom, the C method has shorter intervals than the T3. From Dunnett's Table 3, the C procedure seems to be better for sample sizes in excess of 25.

7.10 Selection of Multiple Comparison Method

The selection of an appropriate multiple comparison procedure depends upon information on the hypothesis of interest and the nature of the data. Ideally the experimenter requires the most powerful possible method that also provides sufficient protection against wrong decisions. It is apparent from section 7.8 that there are a multitude of methods to choose from. In this section we discuss the different techniques. This discussion is based on the simulation papers summarised in section 7.9 and some of the many review papers available.

The discussion has been divided into sub-sections according to the hypothesis of interest. Some of the methods have been extended for testing other hypotheses; for example, Tukey's test can be extended to test linear contrasts. However it is clear that methods are generally most sensitive when applied to the hypothesis they were originally designed for. For example, Scheffé's test is more sensitive for testing contrasts and Tukey's test is more sensitive for testing pairwise comparisons (Miller 1981 p63, Dodge and Thomas 1980, Scheffé 1959).

Many of the Normal Theory methods have been found to be robust for Non Normality (e.g. Scheffé 1959, Dodge and Thomas 1980, Brown 1974), but these methods may not be the most powerful for Non Normal distributions. Miller (1986) suggests that the use of transformations to improve Normality or the use of other methods may lead to more efficient procedures for Non Normal distributions.

7.10.1 Pairwise Comparisons

The Protected Least Significant Difference method (PLSD), which requires a significant F test before it is used, is probably the most familiar of multiple comparison methods. It is applicable to unbalanced designs, is very easy to use and has sensitivity as good or better than other methods. The preliminary F test guards against falsely rejecting the null hypothesis when it is true. However, when the null hypothesis is false, the PLSD gives no increased protection to that part of the null hypothesis which remains true (Miller 1981). Thus the PLSD has low Type II errors but high Type I errors, the simulation studies which include the PLSD bear this out (Carmer and Swanson 1973, Thomas 1973, Dodge and Thomas 1980); the reviews papers reiterate this problem (e.g. Cornell 1971, Gill 1973). Where this method is not protected by a preliminary F test then the experiment-wise error rate increases still further.

Where an experiment has equal replication and equal variances then the Tukey method has been shown to provide the shortest intervals whilst protecting the experiment-wise error rate (e.g. Carmer & Swanson 1973, Miller 1986). Of the methods capable of handling unequal replication in the equal variance case, the Tukey-Kramer produces the shortest intervals (e.g. Dunnett 1980a, Stoline 1981).

If the condition of equal variances does not hold then there are a number of possible methods available. Dunnetts (1980b) simulation study, which picks up from Tamhanes (1979) study, shows that the T3 and C methods provide the shortest intervals. The C method provides shorter intervals than the T3 method where the number of degrees of freedom is large.

Multiple Range Methods

The multiple range methods are used for comparing all pairs of means but cannot be used for constructing confidence intervals. There is much discussion about the use of multiple range methods and the principle objections are usually to the definition of the error rates which are neither experiment-wise nor comparison-wise. This choice of error rate also makes comparisons between multiple range and other methods rather difficult.

A further disadvantage of the multiple range tests is that the power of testing all pairs of means is subject to the magnitude of the other means. O'Neil and Wetherill (1971) note that techniques based on ranges can be constructed to have precise error rate properties but if standard significance levels are used the techniques are too conservative and so lack power. Such methods are also rather sensitive to deviations from distributional assumptions.

In Duncan's test the probability of a Type I error increases with the number of means being compared, raising the question of· whether sufficient protection is being given to the Null hypothesis or not. The increasing levels for α do make the procedure more powerful. Newman-Keuls test is less powerful and more conservative.

It is difficult to find a consensus of opinion about the use of multiple range methods. For example Gill(1973) considers that the evidence against Duncan's method is so incriminating that use of the test should be discontinued and yet considers the Newman-Keuls to offer sufficient protection to the experiment-wise error rate and greater sensitivity then Tukey's method. Thomas(1974) says that he would undoubtedly choose Duncan's method for pairwise comparisons because of its power.

Spjøtvoll and Stoline, Hochberg, Kramer and Duncan have all suggested ways in which the methods could be extended to allow for unequal variances or sample sizes but in doing so all distributional properties are lost (O'Neil and Wetherill 1971)

Nonparametric Methods

The two best known Nonparametric methods for testing pairwise comparisons are the adapted Kruskal-Wallis test, where the observations from all groups are ranked and the Steel-Dwass methods which rank only the two groups being compared. The Kruskal-Wallis method is very versatile and requires less ranking than the Steel-Dwass. However, the major drawback of the Kruskal-Wallis method is that the outcome of a comparison between two groups depends on the ranking of the observations in the other groups. In addition it is very difficult to construct confidence intervals in the Kruskal-Wallis method. In general the Steel-Dwass method is preferred (Miller 1981, Dodge and Thomas 1980).

7.10.2 Contrasts

If at all possible, designed comparisons should be used rather than comparisons selected post-data, primarily because more powerful methods can be used (Gill 1973, Chew 1976). Linear contrasts which are orthogonal can be tested by partitioning the degrees of freedom for treatments in the ANOVA table. If non-orthogonal contrasts are required then the Bonferroni method can be used.

Where the experimenter wishes to test linear contrasts that were not designed before the experiment, Scheffès method can be used. Scheffé's method controls the experiment-wise error rate for all possible contrasts; as an experimenter is usually interested in a few selected contrasts, the Scheffé method is rather conservative. The Bonferroni-t method can also be used to test linear contrasts and may yield shorter intervals where there are a small number of comparisons (Miller 1981 p69, Gill 1973). The method proposed by Brown & Forsythe for the unequal variance case is recommended by Tamhane(1979).

7.10.3 Techniques for Specific Purposes

A number of methods have been developed specifically for dealing with particular situations, usually where one of the treatment groups is a control group. Two different situations were considered in section 7.8, where treatment groups are different levels of a single factor and where treatment groups are different factors. Where a specialised technique can be applied it tends to perform better than one of the more general techniques already considered.

Different Factors

Where the treatment groups are different factors and the experimenter is interested in comparing each group with a control then Dunnett's test is the most sensitive (Miller 1981 p62, Cornell 1971, Gill 1973) although it does require equal replication and equal variances. The nonparametric analog to Dunnett's test is the Many-One rank test; the Kruskal-Wallis test can also be adapted for comparing groups with control. The comparison between nonparametric methods based on pairwise ranking and those based on ranking over all treatment groups has already been made in section 7.10.1.

Different Levels of a Single Factor

If the treatment groups are different levels of a single factor, for example, different dose levels of a single compound, then procedures proposed by Bartholomew or Williams may be appropriate. However, if interest is centred on estimating the dose level at which the response attains a given magnitude, it may be more appropriate to use regression methods (Chew 1976, Williams 1971).

Bartholomew's method is a test of the null hypothesis against the alternative hypothesis of monotonic ordering. Williams(1971) states that Bartholomews test is superior to those tests which have no order assumptions but that it is not designed to perform best against the most important alternatives in the dose response situation. Williams(1971) method is designed to find the lowest dose at which there is evidence of a response.

In his 1971 paper Williams used simulation methods to compare his method with other methods including Bartholomew's. The results suggested that, on the whole, Bartholomew's test is the most powerful. William's test performs better when the number of observations in the control group is increased and is also more robust against departures from the assumption of monotonic ordering. Shirley(1977) proposed a nonparametric version of William's method which was modified slightly by Wlliams(1986).

7.11 Approach used in THESEUS

The prototype rulebase in THESEUS is not intended to provide knowledge on all the possible methods available. for analysing data in a given situation. Rather, it is intended to supply a rational rulebase which covers the domain adequately. In other words, to be able to suggest or recommend methods which are appropriate in the different situations which come within the scope of the domain.

7.11.1 One Sample

In the single sample case the preferred test is the t-test provided there are no outliers and the data is Normal. For samples with more than 25 observations the Normal approximation is used provided there are no outliers. The Wilcoxon Signed Rank test is used if there are outliers present or, for samples of size less than 25, the data is Non-Normal and no suitable transformation can be used.

7.11.2 Two Samples

Where there are two samples the preferred test is the two-sample t test provided there are no outliers, the data is Normal and the variances are equal. For samples of size greater than 25 the Normal approximation is used, provided there are no outliers, with a pooled estimate of variance if the variances are equal or separate variances if not.

For samples with less than 25 observations, Normal data but with unequal variances the Aspin-Welch method may be used. In this situation the users are asked whether they wish to transform the data. If the answer is no then the Aspin-Welch method will be recommended.

The Wilcoxon Rank test is used where there are outliers present or where a suitable transformation cannot be used when the data is Non Normal or the variances are unequal.

7.11.3 Several Groups

The user is offered the opportunity of carrying out an overall test for a difference between treatment groups but the overall test is not regarded as a precondition to further testing except where a method specifically requires it. The usual overall test is the ANOVA which is recommended provided there are no outliers and the data, or some transformed set of the data, is normal with equal variances. The Kruskal-Wallis test is recommended if the ANOVA cannot be used.

Within THESEUS, multiple comparison methods are initially considered according to the hypothesis of interest and the nature of the treatment groups.

Table III : Initial Choice of Multiple Comparison Technique

Hypothesis of interest	Normal Theory methods	Nonparametric methods
Some pairwise	Tukey Tukey-Kramer T3 C	K-sample-rank Kruskal-Wallis
All pairwise	Newman-Keuls *1 Duncan *1	K-sample-rank Kruskal-Wallis
Contrasts (post-data)	Scheffe Bonferroni	
Designed Contrasts	Linear Contrasts Bonferroni	
Many-one comparisons *2	Dunnett Bonferroni	Many-one rank
Lowest Dose response *3	Williams	Shirley

Notes :

*1 If all pairwise comparisons are required then the user is asked whether they wish to use multiple range methods; if not the methods for some pairwise comparisons are considered.

*2 The many-one comparisons are only considered if there is a control or standard treatment group present and the user wishes to compare each treatment group with the control.

*3 The lowest dose response hypothesis is only considered if the treatment groups are different levels of a single facor and the user is interested in finding out the lowest level at which there is evidence of a response.

Table III summarises the initial choice of methods to be considered. Once a list of possible methods has been established then THESEUS works by establishing whether the Normal Theory assumptions hold (see section 6.9). If this is the case then the Normal Theory method can be applied with appropriate methods for unequal sample sizes being employed where possible. The specialised T3 and C methods are used when the only Normal Theory assumption violated is that of equal variances; the T3 method is used when the sample sizes are less than 25 and the C method is recommended otherwise. The value of 25 is based on the results given in Table 2 of Dunnett(1980b).

The Bonferroni method appears in several sections of Table I above because of its great versatility.

When the choice is between the Sheffè and Bonferroni method the Bonferroni method will be recommended if there are only a few comparisons to be made. When the user is considering designed contrasts the Bonferroni method is recommended if the contrasts are not orthogonal.

When the user wants to tests treatment groups with the control and the sample sizes are not equal then the Bonferroni test may be recommended.

Chapter Eight

Development of the Prototype System

Once a logical design had been developed for the system, see Chapters 3 and 4, and the knowledge acquisition was underway, see Chapter 5, the next stage is to design and implement the software. As already mentioned in section 3.5, each entity in the system is declared as an array of records where the records are defined by the attribute lists for each entity. The Life-Cycle diagrams proposed for each entity define the flow of control within the software code.

In this chapter the choice of implementation language and the software structure are discussed. The expert system user requires other facilities to be available during the consultation process and the design and incorporation of these is covered. Finally the way in which the system interacts with the user is specified and an example consultation is given. The consultation process has already been described in some detail in Chapter 4.

8.1 Choice of Language

Once the system had been designed and the knowledge acquisition was underway it was necessary to decide on the implementation language, for this prototype there were two major constraints. The system was to be developed on an IBM-AT compatible machine, this was chosen because if the system is to used by research workers in industry it is necessary to use a machine that they will have access to. The other major constraint is the need to access the data during a consultation in order to carry out statistical tests. When the software development for this project began none of the Artificial Intelligence languages such as Prolog or Lisp, that were available on the IBM-AT, could access other languages or packages. Such languages are rather hostile for writing statistical routines and thus it was necessary to use a procedural language. By using a procedural language the library routines could be picked up where available and others coded as required. Pascal was chosen as the implementation langauge because of the ease with which user-defined records can be utilised, thus enabling the easy definition of entities within the system. An additional benefit of Pascal was its recursive capability, this meant that developing the code for Backward Chaining was not too difficult.

8.2 Overall Structure

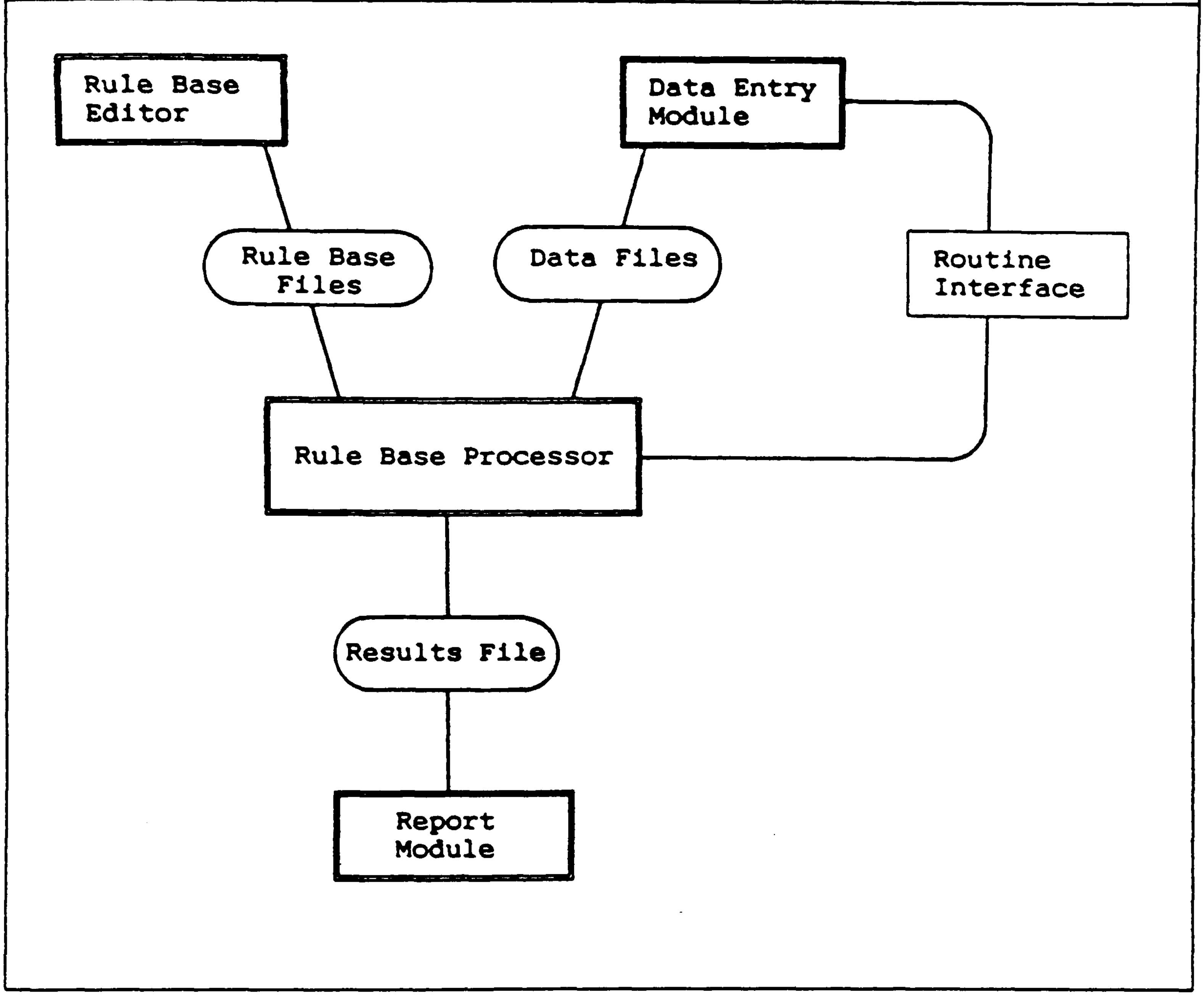
The system overall comprises a number of modules, each of which has a unique function. Communication between modules is effected by means of standard format text files created by each module. Figure 8.1 shows the modules within the system.

Central to the system is the rule base processor or the expert system part, the structure of this has already been described in Chapter 4. Surrounding this rule base processor are the rule base editor, a data entry section and a report module. There is also a routine interface to provide access to statistical routines.

The rule base editor supplies a file of rules which can be picked up by the expert system. The editor enables an expert user to enter, delete and modify rules.

The data entry module allows the system user to enter and edit data, performs basic descriptive analyses and conducts a dialogue with the user to ensure that both the user and the system are satisfied with the representation of the data. This dialogue also serves to ensure that the data under consideration comes within the scope of the system. The rule-base processor works through the rules using a combination of forward and backward chaining, accessing the routine interface and reporting intermediate results as appropriate. This module has only been implemented in part and does not yet contain the dialogue

Figure 8.1 : Component Modules of the System



section. The routine interface allows the system to perform statistical tests on the data both during a consultation and once a particular analysis has been selected.

The report module provides the results of analysis for the user and allows them to structure output in an appropriate way, accessing intermediate results as required, this module has not yet been implemented.

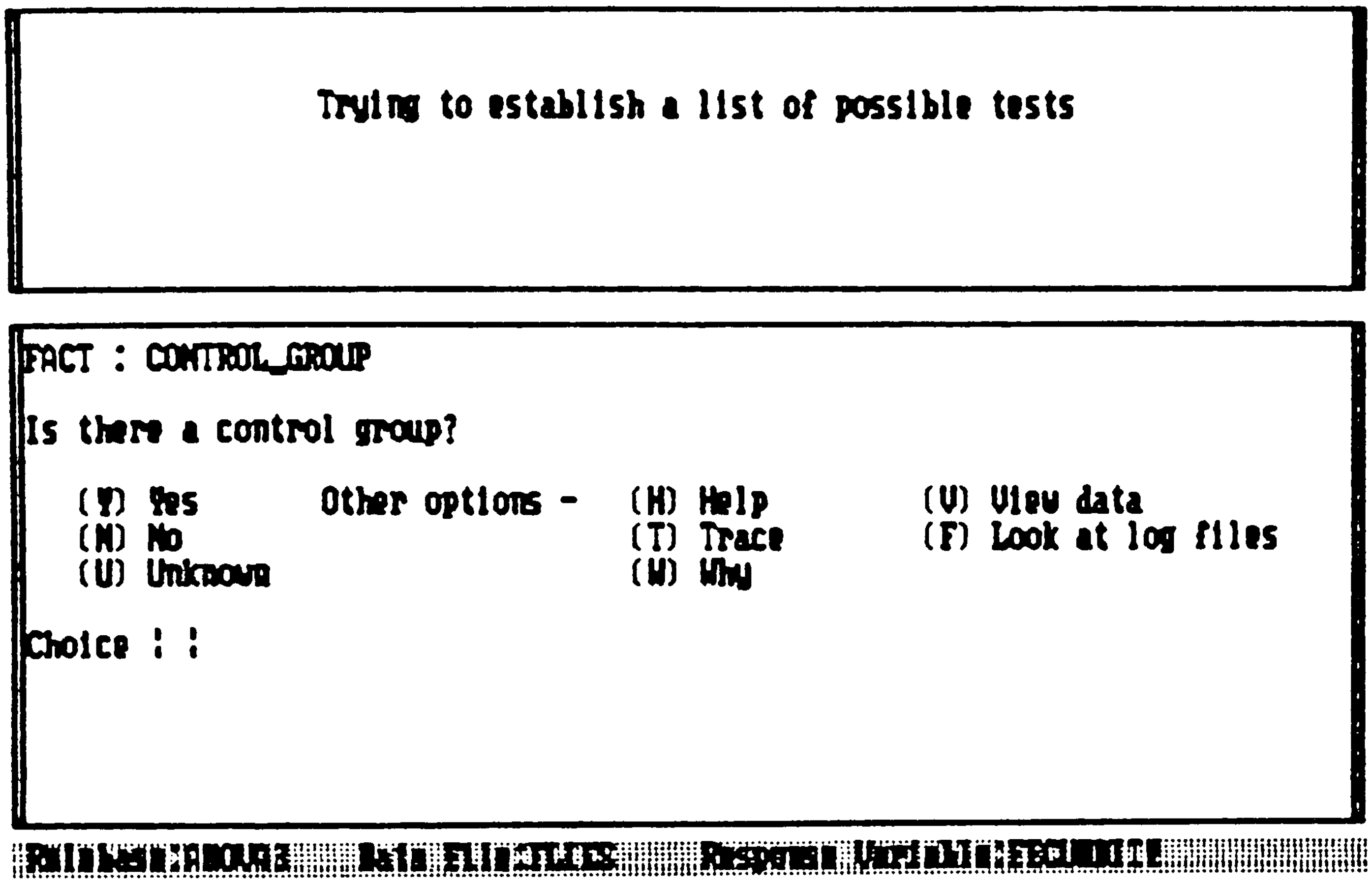
8.3 User Interface

The prototype system presents the user with a split screen consisting of two windows. The top window keeps the user informed of the state of the consultation process. The bottom window is used for interacting with the user and will display menus or questions or requested information during the consultation. The split screen format can be seen in Figure 8.2. There is also a status bar at the bottom of the screen which displays information on the rule-base and data set in use. When the system is being run in test mode the information about the data set is replaced by information on the rules being tried.

User control of the system is effected by means of menus. The main menu allows the user to pick up a rule-base and data set, to look at the data and also permits access to the trace and log facilities. Each facility that can be called during a consultation provides a simple menu of options for the user to choose from. The main menu also provides the point of access to a consultation.

During a consultation, when the system wishes to ask the user for information, a question will be shown in the bottom screen. The user is offered a number of possible responses, Figure 8.2 shows an example of a question screen.

Figure 8.2 : Example of a Question Screen



```
                 Trying to establish a list of possible tests

FACT : CONTROL_GROUP

Is there a control group?

   (Y) Yes       Other options -    (H) Help          (V) View data
   (N) No                           (T) Trace         (F) Look at log files
   (U) Unknown                      (W) Why

Choice : :

Rulebase:ANOVA3   Data File:FLIES   Response Variable:FECUNDITY
```

8.4 Facilities

In order to assist the user a number of facilities are provided. A user can request help or to ask why a particular question is being asked at any stage during the consultation. Within the prototype system described here facilities are also provided for the user to look at the trace arrays or log files as well as to look at the current data set. The system can be run in test-mode so that modifications to the rule-base can be tested.

8.4.1 Help Facility

The help facility is available whenever the system is asking the user for information. Help is provided on a key-word basis, the user can specify any text string and the system will try and find help text on that string. A list of available help can be provided on request. Unless the user specifies otherwise then help is supplied for the question the system is currently asking. Figure 8.3 and 8.4 show the initial help screen and an example of help text.

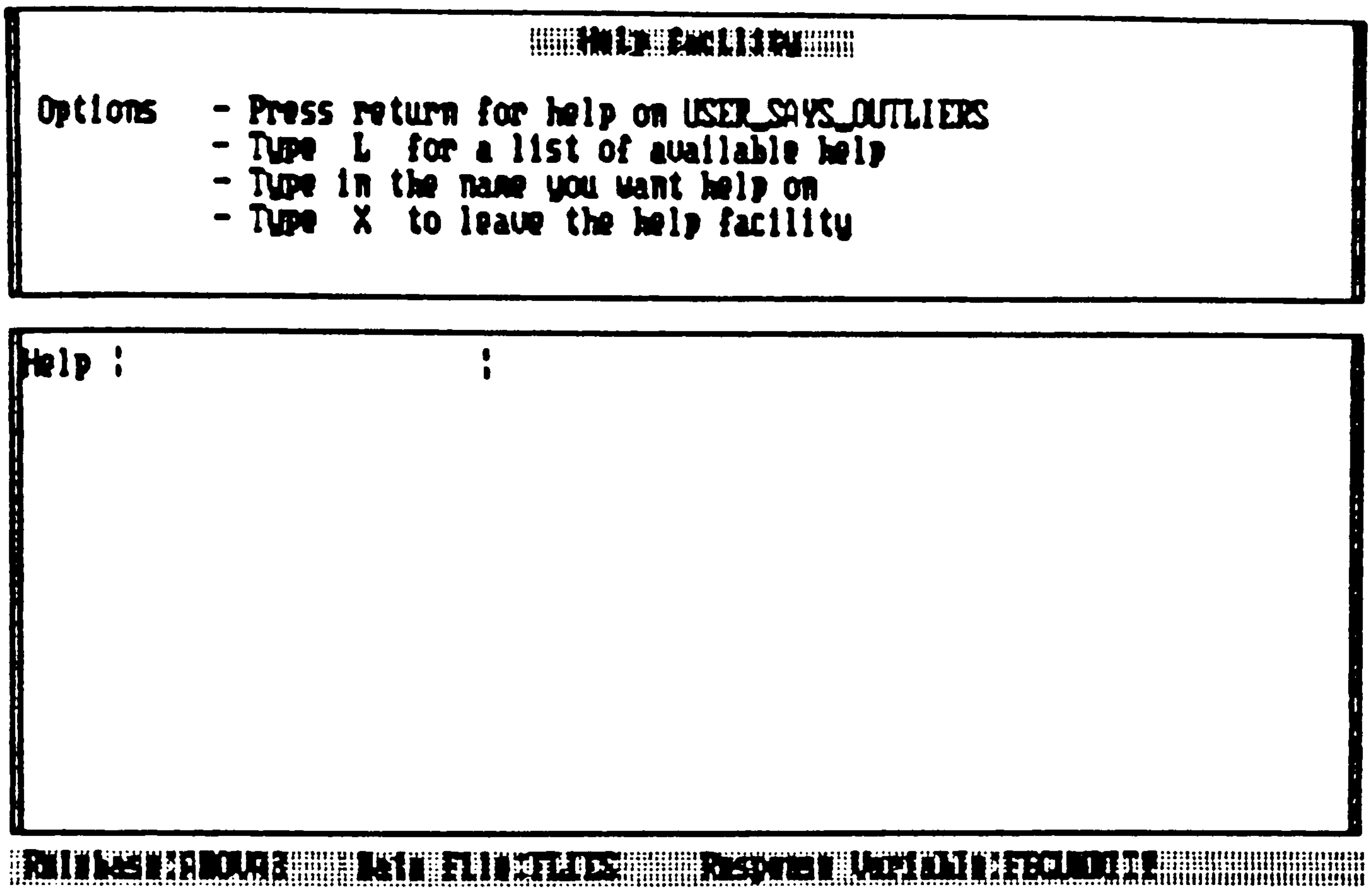
The help text is stored in random access tables; the location for help on a particular text string is generated by calculating a 'hash' function from the text string. The use of random access files means that little time is wasted searching for help text.

8.4.2 Trace Arrays

The trace arrays hold information about the status changes for the entities in the system. There are three trace arrays that can be accessed by the user, the goal trace, the rule trace and the action trace.

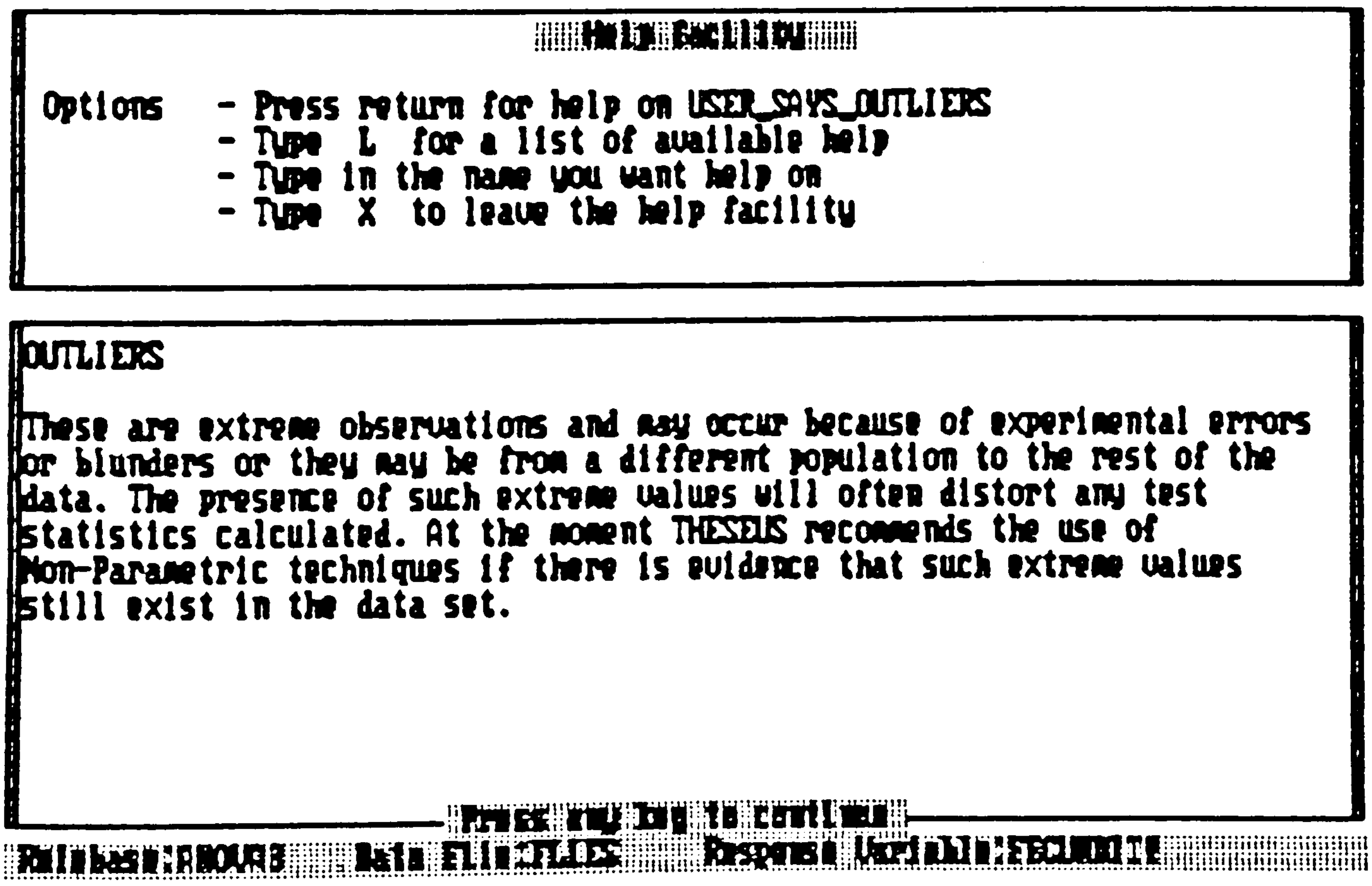
The action trace is the easiest to understand and contains a list of the actions of rules that have been carried out when a rule has fired. An example of information in the action trace is shown in Fig 8.5

Figure 8.3 : Example of the Help Facility - 1



```
                          Help Facility

Options   - Press return for help on USER_SAYS_OUTLIERS
          - Type  L  for a list of available help
          - Type in the name you want help on
          - Type  X  to leave the help facility

Help :                 :

Rulebase:[illegible]   Data File:FLIES   Response Variable:FECUNDITY
```

Figure 8.4 : Example of the Help Facility - 2



```
                          Help Facility

Options   - Press return for help on USER_SAYS_OUTLIERS
          - Type  L  for a list of available help
          - Type in the name you want help on
          - Type  X  to leave the help facility

OUTLIERS

These are extreme observations and may occur because of experimental errors
or blunders or they may be from a different population to the rest of the
data. The presence of such extreme values will often distort any test
statistics calculated. At the moment THESEUS recommends the use of
Non-Parametric techniques if there is evidence that such extreme values
still exist in the data set.

                   Press any key to continue
Rulebase:[illegible]   Data File:FLIES   Response Variable:FECUNDITY
```

Figure 8.5 : Example of the Action Trace

Name	Test or Fact	Action	From Rule	Rule Type
ONE_WAY_ANOVA	TEST	LOOK_AT	R3	F
KRUSKAL_WALLIS_ANOVA	TEST	LOOK_AT	R3	F
OUTLIERS	FACT	FALSE	R35	B
SHAPIRO_WILK_SIG5	FACT	FALSE	R47	B
NORMAL_DATA	FACT	TRUE	R54	B
VARIANCES_EQUAL	FACT	TRUE	R61	B
ACCEPT_PARAMETRIC	FACT	TRUE	R26	B
TRANS_FOR_VARIANCES	FACT	FALSE	R26	B
TRANS_FOR_NORMALITY	FACT	FALSE	R26	B
ADJUST_FOR_UNEQ_VAR	FACT	FALSE	R26	B
ONE_WAY_ANOVA	TEST	RECOMMEND	R76	B
KRUSKAL_WALLIS_ANOVA	TEST	VALID	R76	B
NEXT_TEST	FACT	TRUE	M3	M

Figure 8.6 : Example of Part of a Rule Trace

Rule	Type	Part of condition	Set by	Already Set	Part Satisfied	Rule Status	Data Set
R1	F	ONE_GROUP	FPROC	NO	NO	FAILED	ORIGINAL
R2	F	TWO_GROUPS	FPROC	YES	NO	FAILED	ORIGINAL
R3	F	SEVERAL_GROUPS	FPROC	YES	YES	-	ORIGINAL
R3	F	OVERALL_TEST	USER	NO	YES	FIRED	ORIGINAL
R76	B	ACCEPT_PARAMETRIC	RULE	NO	-	-	ORIGINAL
R22	B	NOT OUTLIERS	RULE	NO	-	-	ORIGINAL
R30	B	MAX_GROUPSIZE_GT_25	FPROC	NO	NO	FAILED	ORIGINAL
R31	B	MAX_GROUPSIZE_GT_25	FPROC	YES	NO	FAILED	ORIGINAL
R32	B	NOT MAX_GROUPSIZE_GT_25	FPROC	YES	YES	-	ORIGINAL
R32	B	DIXONS_SIG_5	FPROC	NO	NO	FAILED	ORIGINAL
R33	B	NOT MAX_GROUPSIZE_GT_25	FPROC	YES	YES	-	ORIGINAL
R33	B	DIXONS_SIG_5	FPROC	YES	NO	FAILED	ORIGINAL
R34	B	NOT MAX_GROUPSIZE_GT_25	FPROC	YES	YES	-	ORIGINAL
R34	B	NOT DIXONS_SIG_5	FPROC	YES	YES	-	ORIGINAL
R34	B	USER_SAYS_OUTLIERS	USER	NO	NO	FAILED	ORIGINAL
R35	B	NOT MAX_GROUPSIZE_GT_25	FPROC	YES	YES	-	ORIGINAL
R35	B	NOT DIXONS_SIG_5	FPROC	YES	YES	-	ORIGINAL
R35	B	NOT USER_SAYS_OUTLIERS	USER	YES	YES	FIRED	ORIGINAL
R22	B	NOT OUTLIERS	RULE	YES	YES		ORIGINAL
R22	B	NORMAL_DATA	RULE	NO	-	-	ORIGINAL
R36	B	NOT MORE_THAN_10_OVERALL	FPROC	NO	NO	FAILED	ORIGINAL
R37	B	NOT MORE_THAN_10_OVERALL	FPROC	YES	NO	FAILED	ORIGINAL
R48	B	MORE_THAN_10_OVERALL	FPROC	YES	YES	-	ORIGINAL
R48	B	NOT MORE_THAN_20_OVERALL	FPROC	YES	NO	FAILED	ORIGINAL

The rule trace keeps track of the rules that the system tries to apply. Fig 8.6 gives an example of part of a rule trace. For each part of the condition of a rule tried by the system a new line is entered into the rule trace. The trace stores information on the rule and its status as well as on the part of the condition and where the system needs to look to establish that part. Information on the current data set is also stored. If the data is transformed then the system will retry some rules on the transformed data.

The goal trace keeps track of the goals that the system tries to backward chain on. In the first instance these goals are the tests that the system has decided it wants to consider. Other goals will be facts the system needs to establish the status of a test. Figure 8.7 gives an example of part of a goal trace. The goal trace stores information on the rules tried and the status of the goal. The data set that the goal is being established on is also recorded. The hyphens used to the left of the goal name specify the depth of recursion. A single hyphen denotes that the backward chainer has been called to try and establish the status of a test. Further hyphens denote recursive calls to the backward chainer while it is still trying to establish a test.

Trace arrays are only stored for the current run; if the user starts a new consultation within the system the trace arrays are all re-initialised. When the user leaves the system the current trace arrays are written to a text file.

8.4.3 Log Files

Three text files are created during a consultation to provide information on the progress of the consultation. These files can be accessed during the consultation and can also be printed out after the consultation has

Figure 8.7 : Example of the Goal Trace

```
                                            Rule  Goal       Data
 Goal                                       Tried Status     Set
-----------------------------------------------------------------------
 -ONE_WAY_ANOVA.......................      R76      -       ORIGINAL
 --ACCEPT_PARAMETRIC..................      R22      -       ORIGINAL
 ---OUTLIERS..........................      R30      -       ORIGINAL
 ---OUTLIERS..........................      R31      -       ORIGINAL
 ---OUTLIERS..........................      R32      -       ORIGINAL
 ---OUTLIERS..........................      R33      -       ORIGINAL
 ---OUTLIERS..........................      R34      -       ORIGINAL
 ---OUTLIERS..........................      R35    FALSE     ORIGINAL
 ---NORMAL_DATA.......................      R36      -       ORIGINAL
 ---NORMAL_DATA.......................      R37      -       ORIGINAL
 ---NORMAL_DATA.......................      R48      -       ORIGINAL
 ---NORMAL_DATA.......................      R49      -       ORIGINAL
 ---NORMAL_DATA.......................      R50      -       ORIGINAL
 ---NORMAL_DATA.......................      R51      -       ORIGINAL
 ---NORMAL_DATA.......................      R52      -       ORIGINAL
 ----SHAPIRO_WILK_SIG5................      R38      -       ORIGINAL
 ----SHAPIRO_WILK_SIG5................      R39      -       ORIGINAL
 ----SHAPIRO_WILK_SIG5................      R40      -       ORIGINAL
 ----SHAPIRO_WILK_SIG5................      R41      -       ORIGINAL
 ----SHAPIRO_WILK_SIG5................      R42      -       ORIGINAL
 ----SHAPIRO_WILK_SIG5................      R43      -       ORIGINAL
 ----SHAPIRO_WILK_SIG5................      R44      -       ORIGINAL
 ----SHAPIRO_WILK_SIG5................      R45      -       ORIGINAL
 ----SHAPIRO_WILK_SIG5................      R46      -       ORIGINAL
 ----SHAPIRO_WILK_SIG5................      R47    FALSE     ORIGINAL
 ---NORMAL_DATA.......................      R53      -       ORIGINAL
 ---NORMAL_DATA.......................      R54     TRUE     ORIGINAL
 ---VARIANCES_EQUAL...................      R55      -       ORIGINAL
 ---VARIANCES_EQUAL...................      R56      -       ORIGINAL
 ---VARIANCES_EQUAL...................      R57      -       ORIGINAL
 ---VARIANCES_EQUAL...................      R58      -       ORIGINAL
 ---VARIANCES_EQUAL...................      R59      -       ORIGINAL
 ---VARIANCES_EQUAL...................      R60      -       ORIGINAL
 ---VARIANCES_EQUAL...................      R61     TRUE     ORIGINAL
 --ACCEPT_PARAMETRIC..................      R23      -       ORIGINAL
 --ACCEPT_PARAMETRIC..................      R24      -       ORIGINAL
 --ACCEPT_PARAMETRIC..................      R25      -       ORIGINAL
 --ACCEPT_PARAMETRIC..................      R26     TRUE     ORIGINAL
 -ONE_WAY_ANOVA.......................      R76 RECOMMEND    ORIGINAL
 KRUSKAL_WALLIS_ANOVA.................      R76    VALID     ORIGINAL
-----------------------------------------------------------------------
```

finished.

The results file provides information on any statistical tests that are undertaken, see Figure 8.8. The log file supplies more detailed information on the progress of the consultation recording the rules tried, procedures called and goals verified, Figure 8.9 gives an example of part of a log file. The message file stores any messages or warnings issued to the user.

When running the expert system it is possible to run more than one consultation, the log files all record the current run number.

8.4.4 Looking at the Data

A facility to look at the data is also provided; this allows the user to look at descriptive statistics or graphical plots of the data. Figure 8.10 shows an example of the descriptive statistics, Figure 8.11 shows an example of a Normal plot.

8.4.5 Why Facility

The Why facility was not included in the prototype system but a simple why facility was included at a later stage. If the system is trying to identify a number of potential methods the Why facility provides information on the rule and the part of its condition that the system is trying, see Figure 8.12. If the Why facility is called when the system is backward chaining the same information is supplied as well as a list of the current goals in the system, see Figure 8.13.

8.4.6 Running the System in Test-Mode

It is possible to run the system in test-mode. This facility is provided to enable a local statistician to test any changes that may have been made to the rulebase without needing to access a data file. The user is asked to supply all the information needed by the system. The status bar shows which rules or goals the system is working with. Figures 8.14 - 8.17 show example screens from running the system in test mode.

Figure 8.8 : Example of the Results File

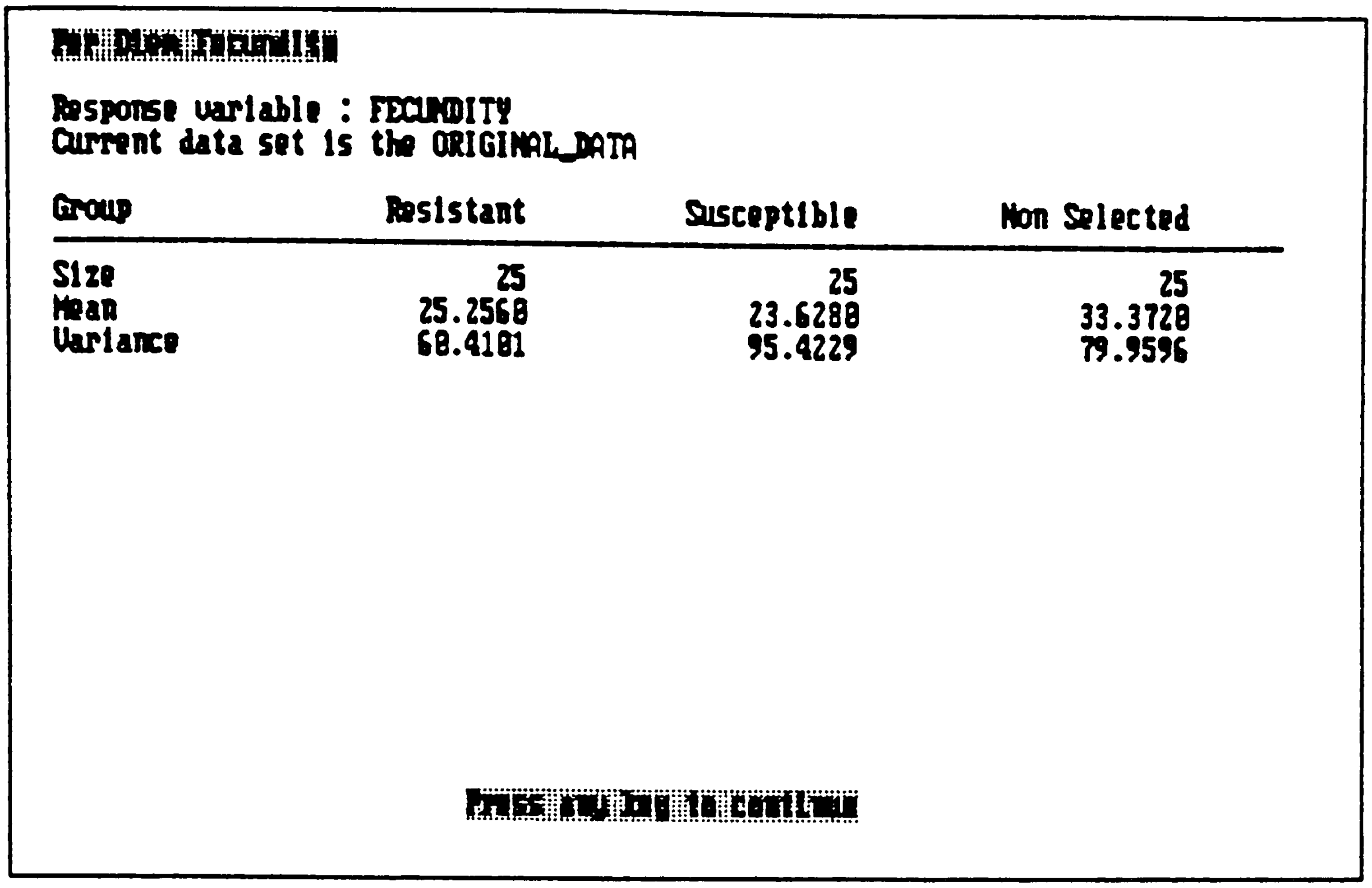
```
RESULTS.LOG - created   Time  13.41 Date 15/9/1988
-----------------------------------------------------
=================================
#  Run number        1          #
=================================
DIXONS test for outliers
Group 1 s_outlier        44.4000 calc   0.1966 crit   0.4060 Not significant
Group 2 s_outlier        48.5000 calc   0.2782 crit   0.4060 Not significant
Group 3 s_outlier        51.8000 calc   0.2984 crit   0.4060 Not significant
-----------------------------------------------------------------------------
Checking for normality using Shapiro wilk test
SHAPWILK_BY_GROUP
Tests each group seperately using the observed values
Calculated values 0.9489
Crit value (5%)   0.9180
Group  1 not significant at 5%
Calculated values 0.9389
Crit value (5%)   0.9180
Group  2 not significant at 5%
Calculated values 0.9831
Crit value (5%)   0.9180
Group  3 not significant at 5%
Shapiro-Wilk test not significant (for any group) at 5% level
-----------------------------------------------------------------------------
LEVENES TEST for unequal variances
TEST CRITERION =    0.62617121
F[5%],( 2,72) = 3.10
F[1%],( 2,72) = 4.92
Accept Ho (5% level)
NO Significant Difference between the   3 Variances
Accept Ho (1% level)
-----------------------------------------------------------------------------
BARTLETTS test for unequal variances
Test Statistic (chi-sq) = 1.225900
Chi-sqrd Table value at 1%[ 2] =     9.210
Cannot Reject Ho at 1% ... implies variances homogenous
-----------------------------------------------------------------------------
ANALYSIS OF VARIANCE TABLE - ONE WAY
-----------------------------------------------------------------------------
SOURCE                     SS              DF              MS               F
-----------------------------------------------------------------------------
TREATMENTS     1362.21147               2      681.10573         8.66574
RESIDUAL       5659.02240              72       78.59753
-----------------------------------------------------------------------------
TOTAL          7021.23387              74
-----------------------------------------------------------------------------
Residual Mean Square (RMS) =    78.59753
One Way Anova : Significant at 5%
Differences Between the Treatments
-----------------------------------------------------------------------------
RESULTS.LOG  closed Time  13.42 Date 15/9/1988
```

RESULTS.LOG closed Time 13.42 Date 15/9/1988

Figure 8.9 : Example of Part of a Log File

```
LOG.LOG  - created   Time  13.41 Date 15/9/1988
----------------------------------------------------
Rule Base Loaded         : \theseus\rulebase\ANOVA3
Data File Loaded         : FLIES
Response Variable Loaded : FECUNDITY
===================================
#  Run number          1          #
===================================
FACT MORE_TRANS_TO_TRY    reset to default value      TRUE
FACT NEXT_TEST            reset to default value     FALSE
========================================================================
Establishing a list of possible tests (Forward chaining)
PROC TEST_NUM_GROUPS      called to set the following fact
FACT ONE_GROUP             is      FALSE on ORIGINAL
RULE   R1     FAILED failed on     ONE_GROUP
RULE   R2     FAILED failed on     TWO_GROUPS
FACT OVERALL_TEST         set to       TRUE on  by USER
RULE   R3      FIRED
Rule   R3 TEST ONE_WAY_ANOVA        is     LOOK_AT on ORIGINAL
Rule   R3 TEST KRUSKAL_WALLIS_ANOVA is     LOOK_AT on ORIGINAL
Possible tests  - ONE_WAY_ANOVA
                - KRUSKAL_WALLIS_ANOVA
========================================================================
Trying to verify the TEST ONE_WAY_ANOVA       on ORIGINAL
* Trying to establish the goal ACCEPT_PARAMETRIC    on ORIGINAL
** Trying to establish the goal OUTLIERS             on ORIGINAL
PROC TEST_GROUP_SIZE      called to set the following fact
FACT MAX_GROUPSIZE_GT_25  is      FALSE on ORIGINAL
RULE  R30     FAILED failed on     MAX_GROUPSIZE_GT_25
RULE  R31     FAILED failed on     MAX_GROUPSIZE_GT_25
PROC DIXONS_TEST          called to set the following fact
FACT DIXONS_SIG_5         is      FALSE on ORIGINAL
RULE  R32     FAILED failed on     DIXONS_SIG_5
RULE  R33     FAILED failed on     DIXONS_SIG_5
FACT USER_SAYS_OUTLIERS   set to      FALSE on  by USER
RULE  R34     FAILED failed on     USER_SAYS_OUTLIERS
RULE  R35      FIRED
Rule  R35 FACT OUTLIERS             is      FALSE on ORIGINAL
** Goal OUTLIERS             set to      FALSE on ORIGINAL
** Trying to establish the goal NORMAL_DATA          on ORIGINAL
PROC TEST_TOTAL_OBS       called to set the following fact
FACT MORE_THAN_10_OVERALL is        TRUE on ORIGINAL
.
.
.
.
```

Figure 8.10 : Example of the View Data Facility - 1



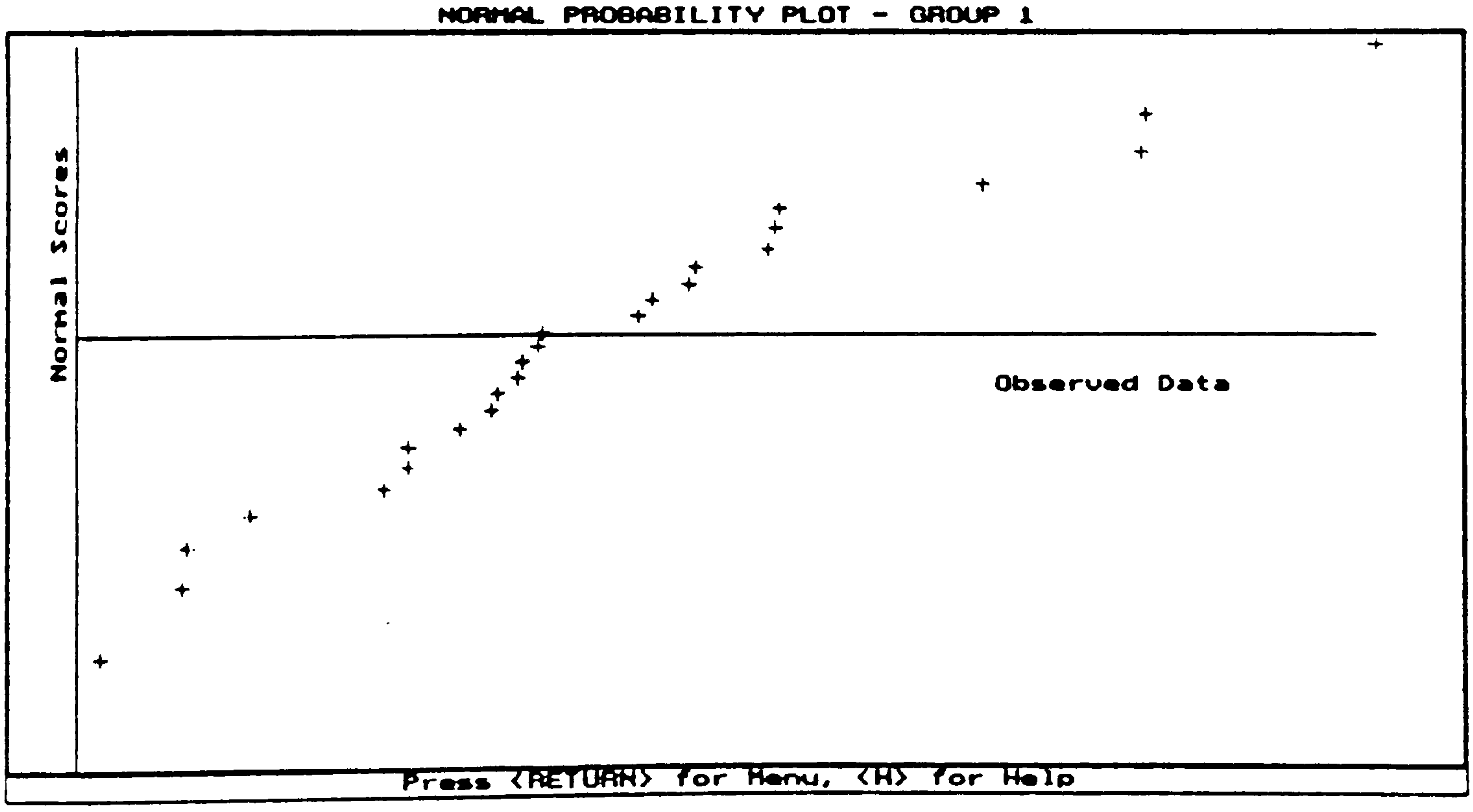
Per Diem Fecundity

Response variable : FECUNDITY
Current data set is the ORIGINAL_DATA

Group	Resistant	Susceptible	Non Selected
Size	25	25	25
Mean	25.2568	23.6288	33.3728
Variance	68.4181	95.4229	79.9596

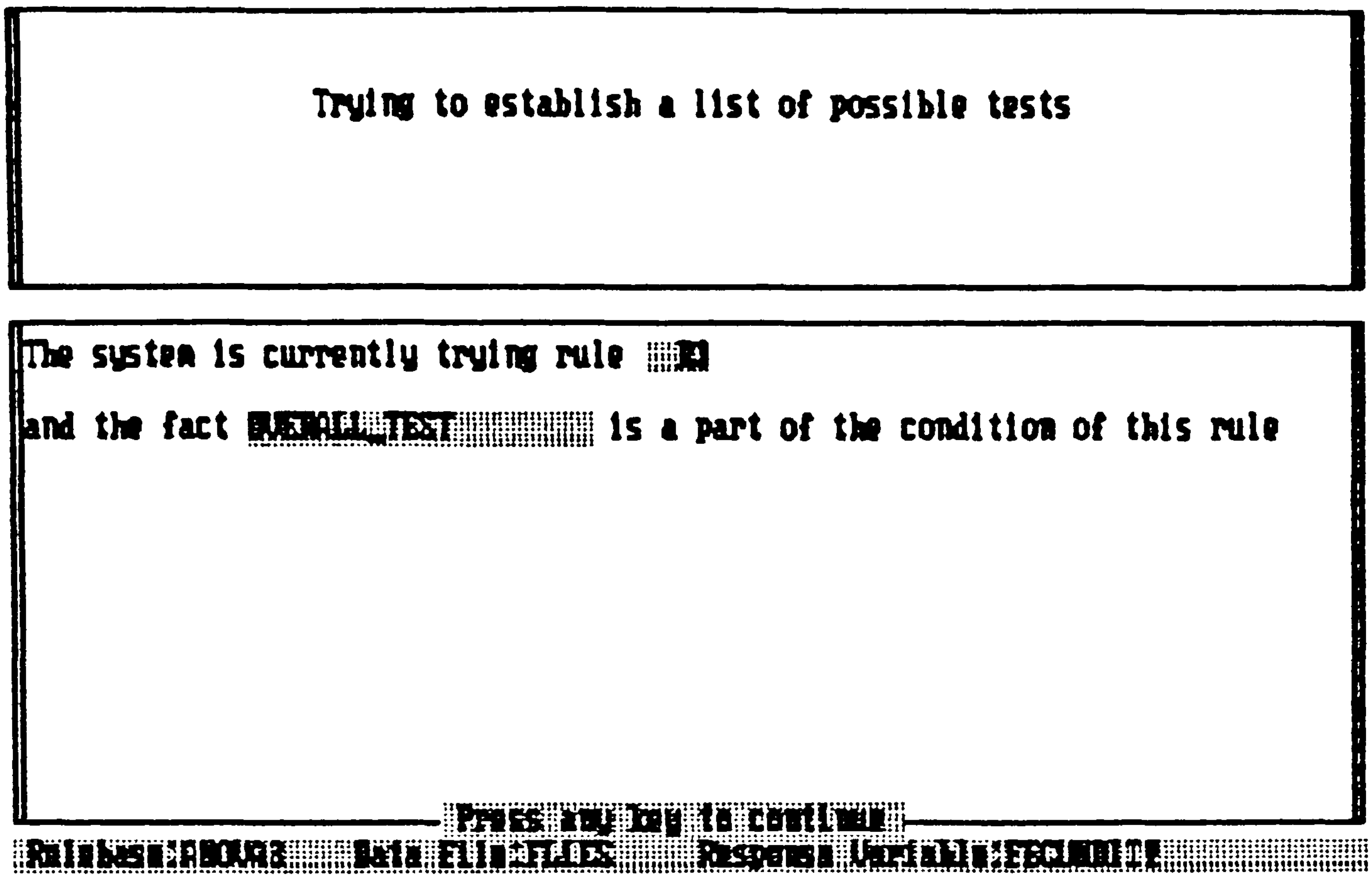
Press any key to continue

Figure 8.11 : Example of the View Data Facility - 2



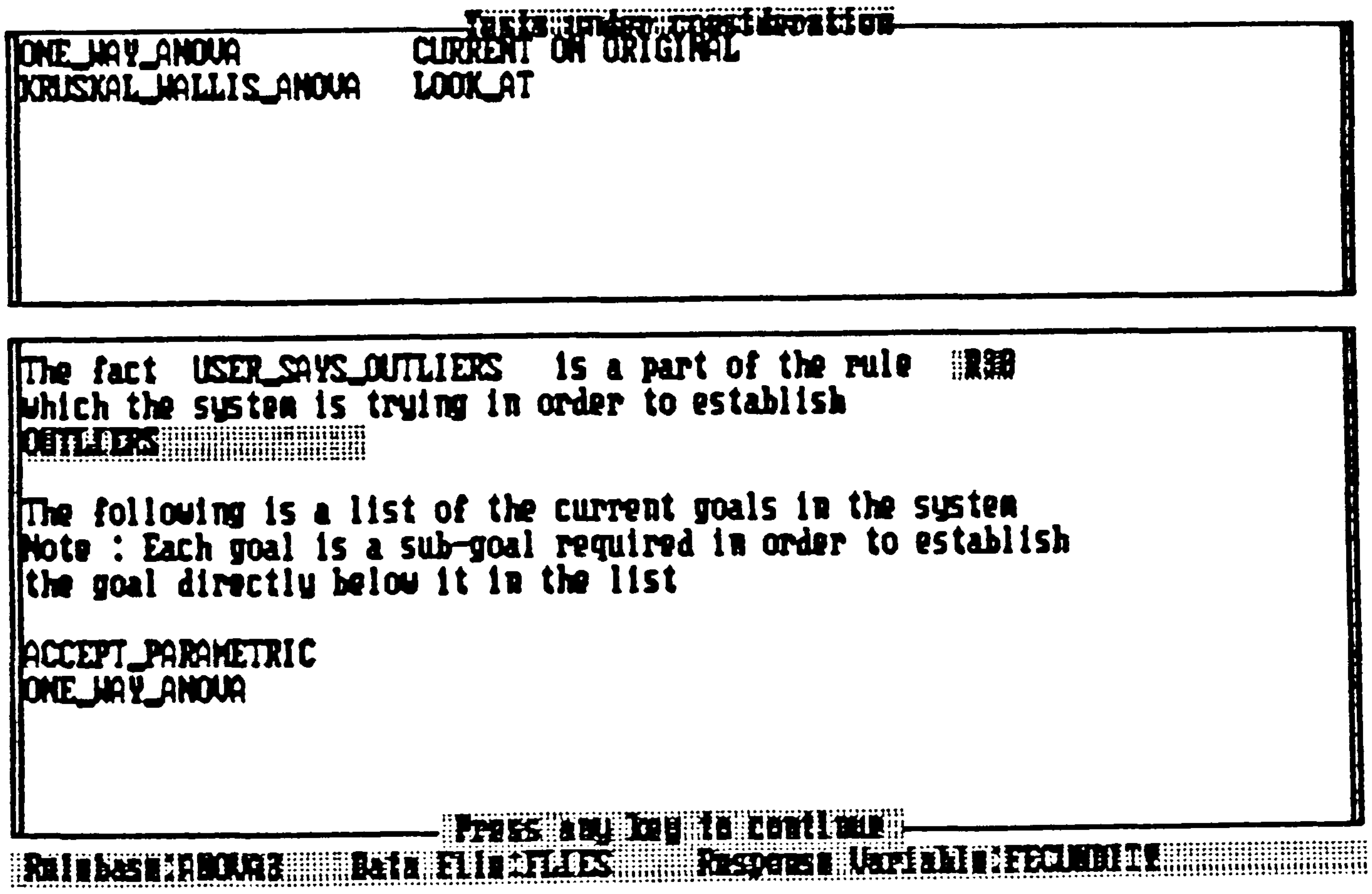
Press <RETURN> for Menu, <H> for Help

Figure 8.12 : Example of the Why Facility - 1



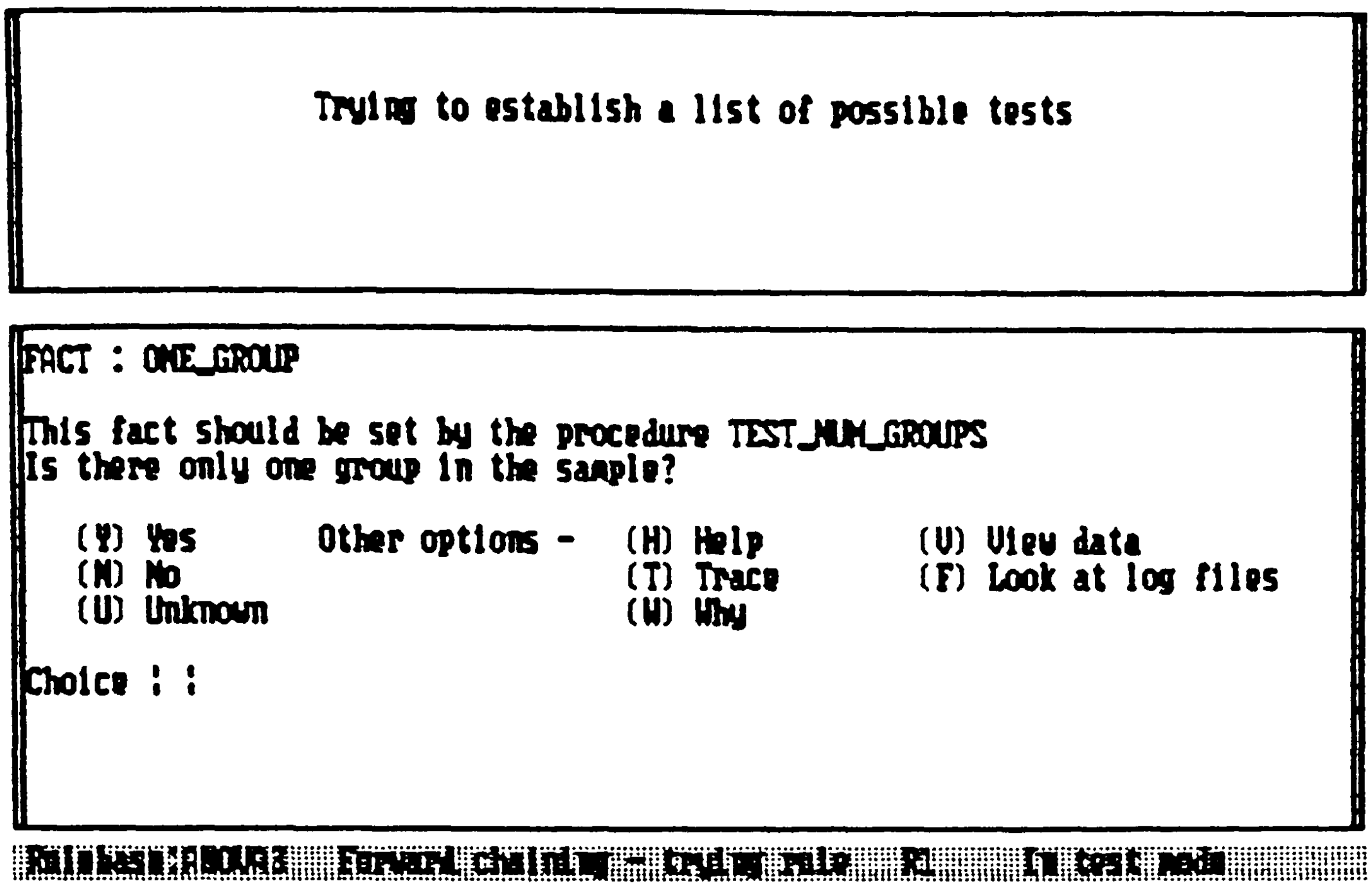
```
                 Trying to establish a list of possible tests

The system is currently trying rule [illegible]

and the fact OVERALL_TEST is a part of the condition of this rule

                            Press any key to continue
Rulebase:ANOVA3   Data File:FLIES   Response Variable:FECUNDITY
```

Figure 8.13 : Example of the Why Facility - 2



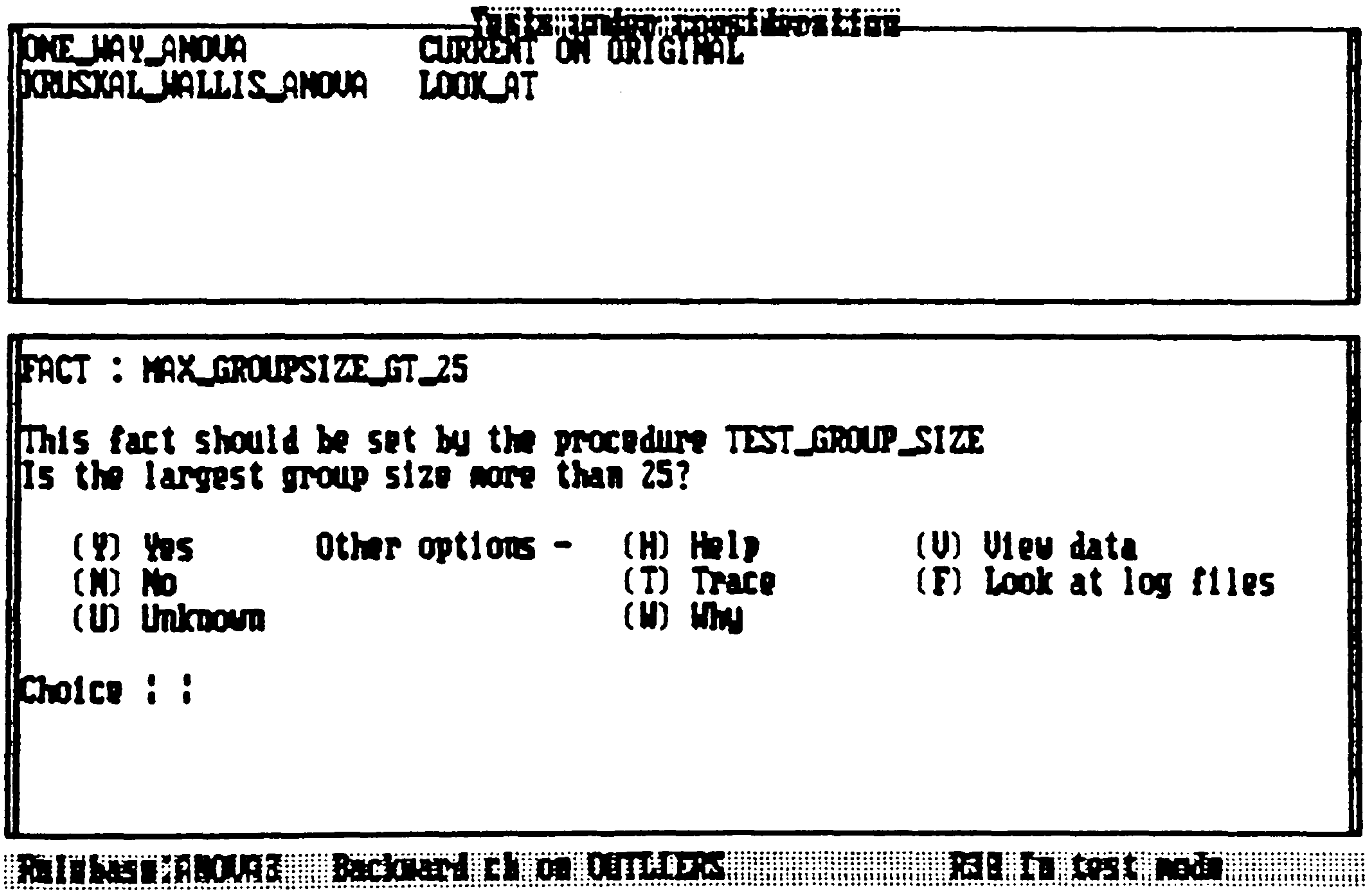
```
                            Tasks under consideration
ONE_WAY_ANOVA          CURRENT ON ORIGINAL
KRUSKAL_WALLIS_ANOVA   LOOK_AT

The fact  USER_SAYS_OUTLIERS  is a part of the rule  [illegible]
which the system is trying in order to establish
OUTLIERS

The following is a list of the current goals in the system
Note : Each goal is a sub-goal required in order to establish
the goal directly below it in the list

ACCEPT_PARAMETRIC
ONE_WAY_ANOVA

                            Press any key to continue
Rulebase:ANOVA3   Data File:FLIES   Response Variable:FECUNDITY
```

Figure 8.14 : Running the System in Test Mode - 1



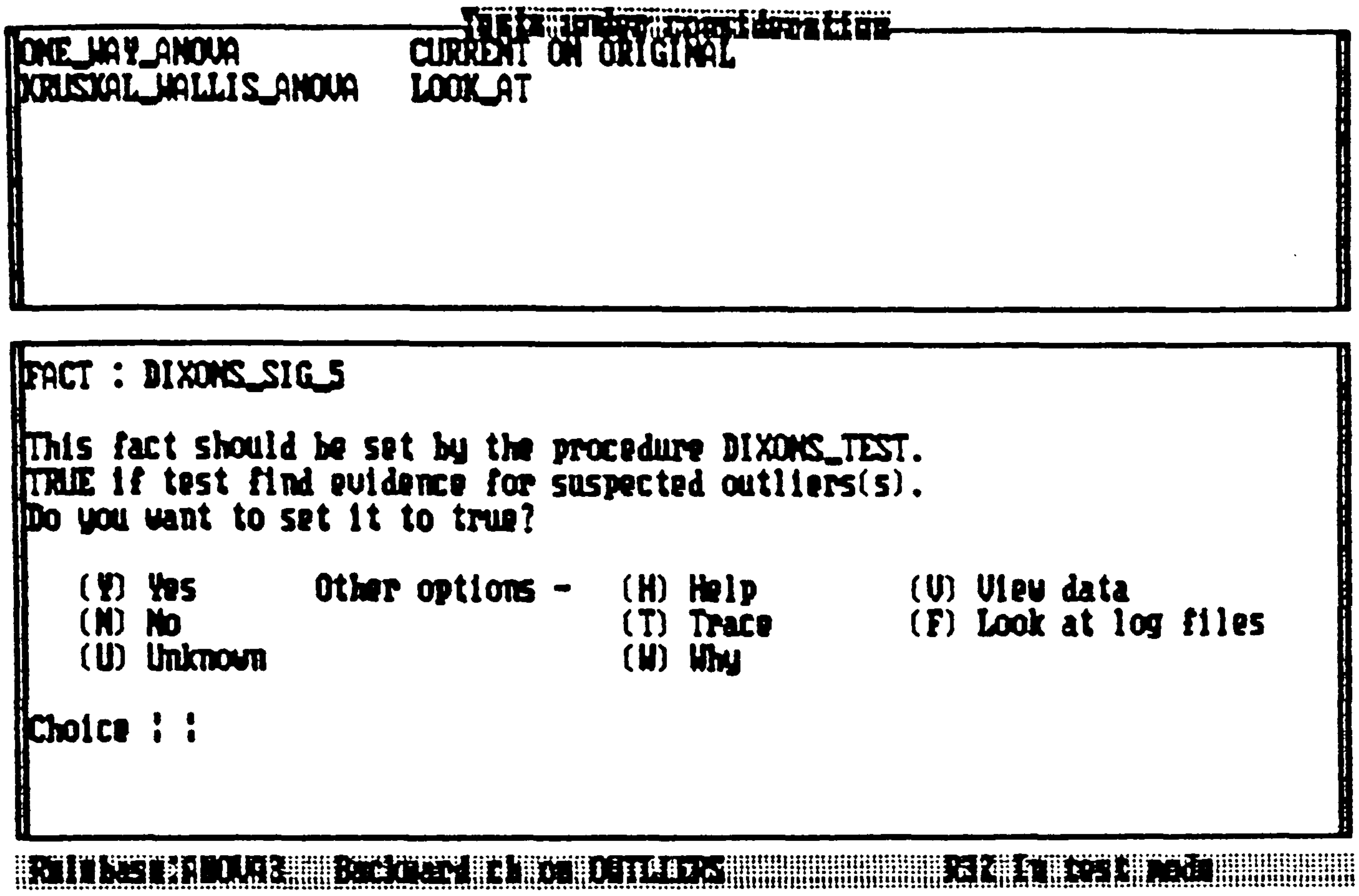
```
                Trying to establish a list of possible tests

FACT : ONE_GROUP

This fact should be set by the procedure TEST_NUM_GROUPS
Is there only one group in the sample?

   (Y) Yes       Other options -  (H) Help        (V) View data
   (N) No                         (T) Trace       (F) Look at log files
   (U) Unknown                    (W) Why

Choice : :

Rulebase:ANOVA3  Forward chaining - trying rule  R1    In test mode
```

Figure 8.15 : Running the System in Test Mode - 2



```
                           Tests under consideration
ONE_WAY_ANOVA          CURRENT ON ORIGINAL
XKRUSKAL_WALLIS_ANOVA  LOOK_AT

FACT : MAX_GROUPSIZE_GT_25

This fact should be set by the procedure TEST_GROUP_SIZE
Is the largest group size more than 25?

   (Y) Yes       Other options -  (H) Help        (V) View data
   (N) No                         (T) Trace       (F) Look at log files
   (U) Unknown                    (W) Why

Choice : :

Rulebase:ANOVA3  Backward ch on OUTLIERS            R38 In test mode
```

Figure 8.16 : Running the System in Test Mode - 3

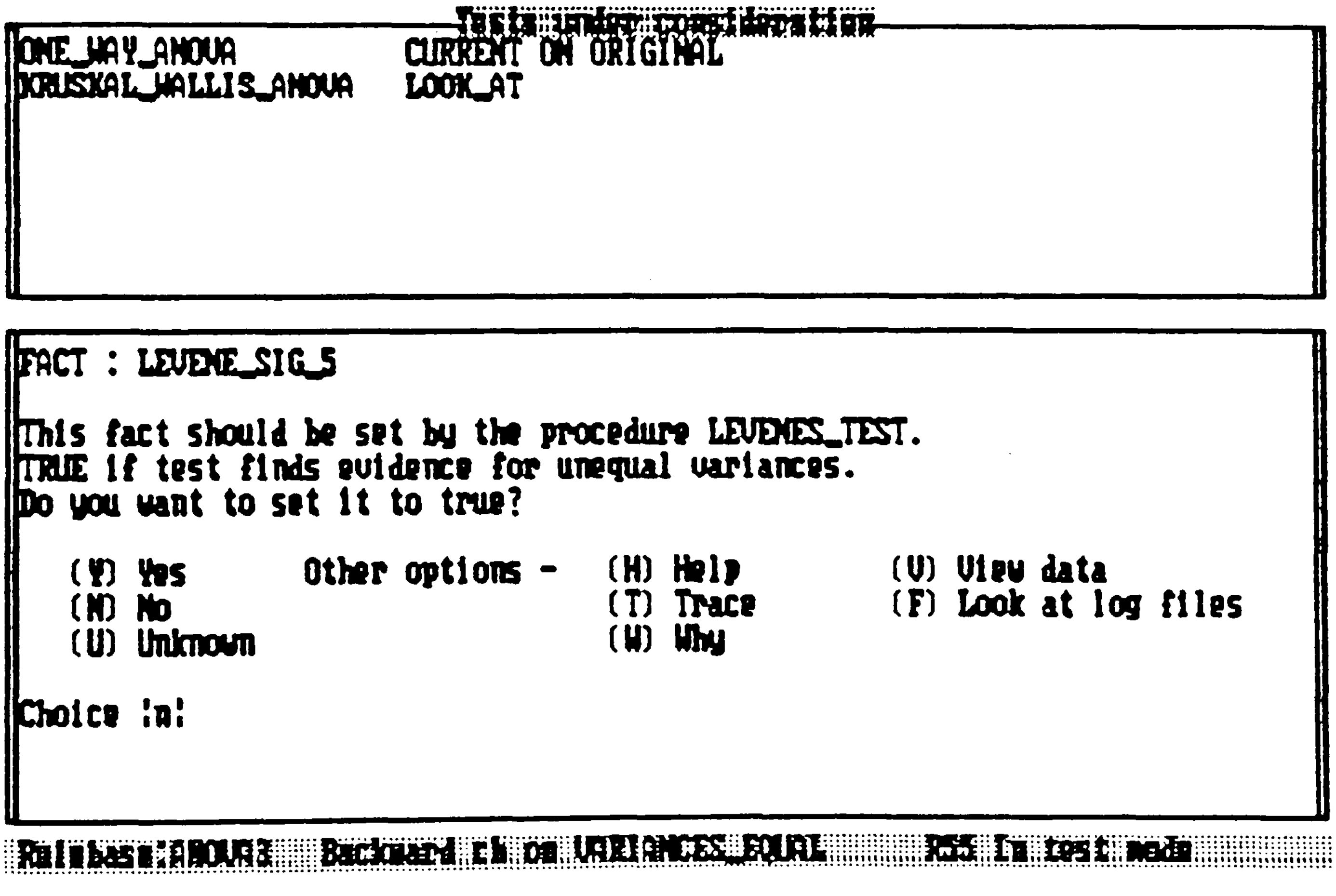


```
                       Tests under consideration
ONE_WAY_ANOVA            CURRENT ON ORIGINAL
KRUSKAL_WALLIS_ANOVA     LOOK_AT
```

```
FACT : DIXONS_SIG_5

This fact should be set by the procedure DIXONS_TEST.
TRUE if test find evidence for suspected outliers(s).
Do you want to set it to true?

   (Y) Yes        Other options -   (H) Help           (V) View data
   (N) No                           (T) Trace          (F) Look at log files
   (U) Unknown                      (W) Why

Choice : :
```

```
Rulebase:ANOVA3  Backward ch on OUTLIERS             R32 In test mode
```

Figure 8.17 : Running the System in Test Mode - 4



```
                       Tests under consideration
ONE_WAY_ANOVA            CURRENT ON ORIGINAL
KRUSKAL_WALLIS_ANOVA     LOOK_AT
```

```
FACT : LEVENE_SIG_5

This fact should be set by the procedure LEVENES_TEST.
TRUE if test finds evidence for unequal variances.
Do you want to set it to true?

   (Y) Yes        Other options -   (H) Help           (V) View data
   (N) No                           (T) Trace          (F) Look at log files
   (U) Unknown                      (W) Why

Choice :n:
```

```
Rulebase:ANOVA3  Backward ch on VARIANCES_EQUAL      R55 In test mode
```

8.5 Examples of Consultations

Two examples of consultations are given. The first goes as far as the One Way Analysis of Variance and shows the use of transformations, the data is simulated data from a negative exponential distribution. Figures 8.18 to 8.46 show this consultation :-

8.18 - 8.23 Using the main menu to select a rule-base and data set.

8.24 - 8.25 Trying to establish a list of possible techniques using forward chaining.

8.26 - 8.27 The system is now considering the One Way ANOVA more closely. The first 'goal' the system is trying to establish is whether there are any outliers present in the data.

8.28 - 8.30 Showing the use of the Viewdata option by the user to assist in answering the question put by the system.

8.31 The system now returns to the question about outliers.

8.32 - 8.35 The system is now trying to establish whether the data is Normal and the variances are equal.

8.36 - 8.37 The Shapiro-Wilk test had found some evidence of Non-Normality and so the system tries transforming the data.

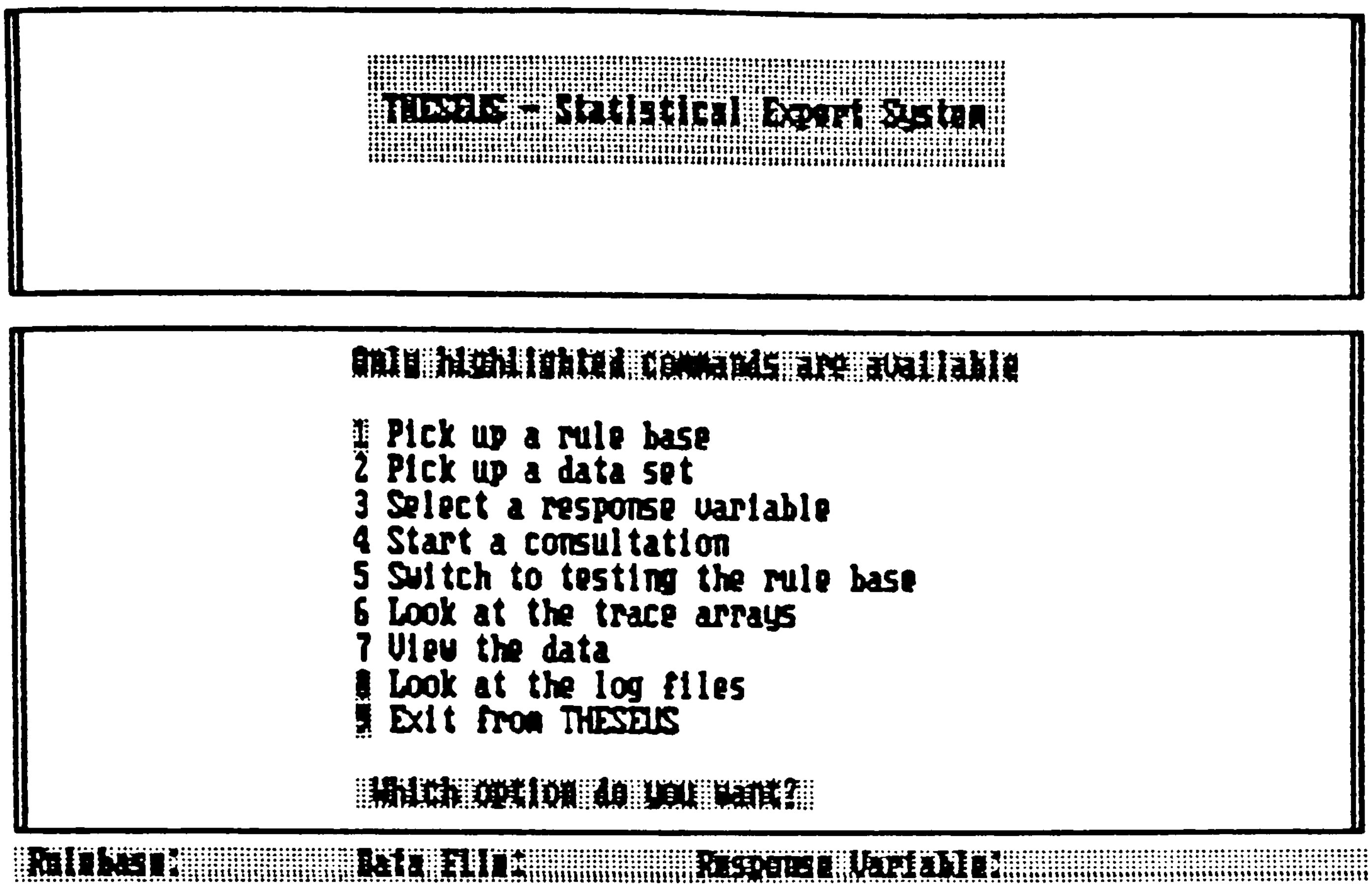
8.38 - 8.40 The system now rechecks for Normality and equality of variances.

8.41 - 8.42 The system has finished considering the possible tests and asks the user to select a test.

8.43 The results of the analysis of variance

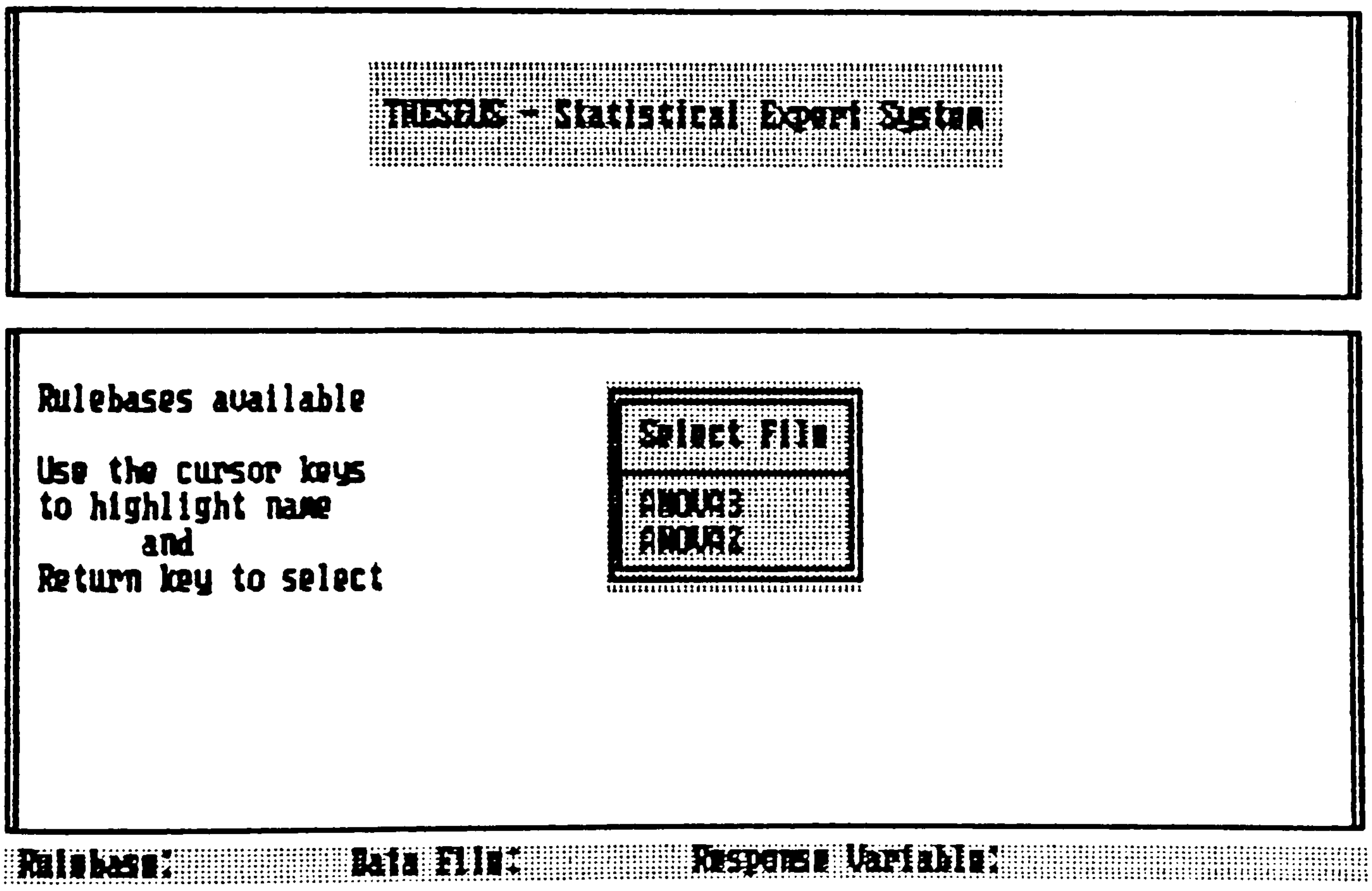
8.44 - 8.46 The system asks the user whether they wish to consider any further analysis, if not it informs them of all techniques selected and then returns to the main menu.

Figure 8.18 : Consultation A - 1



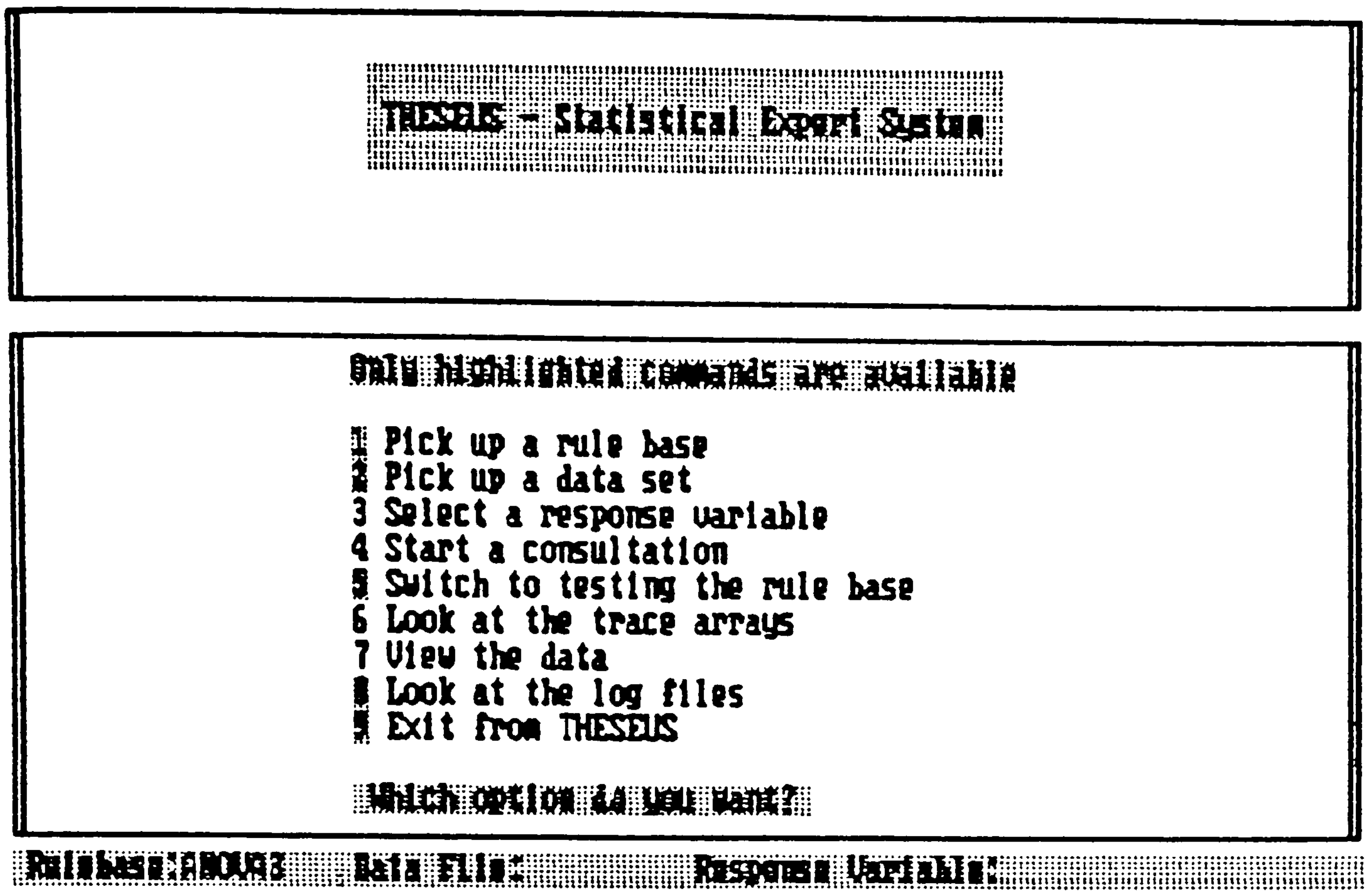
```
THESEUS - Statistical Expert System

Only highlighted commands are available

1 Pick up a rule base
2 Pick up a data set
3 Select a response variable
4 Start a consultation
5 Switch to testing the rule base
6 Look at the trace arrays
7 View the data
8 Look at the log files
9 Exit from THESEUS

Which option do you want?

Rulebase:          Data File:          Response Variable:
```

Figure 8.19 : Consultation A - 2



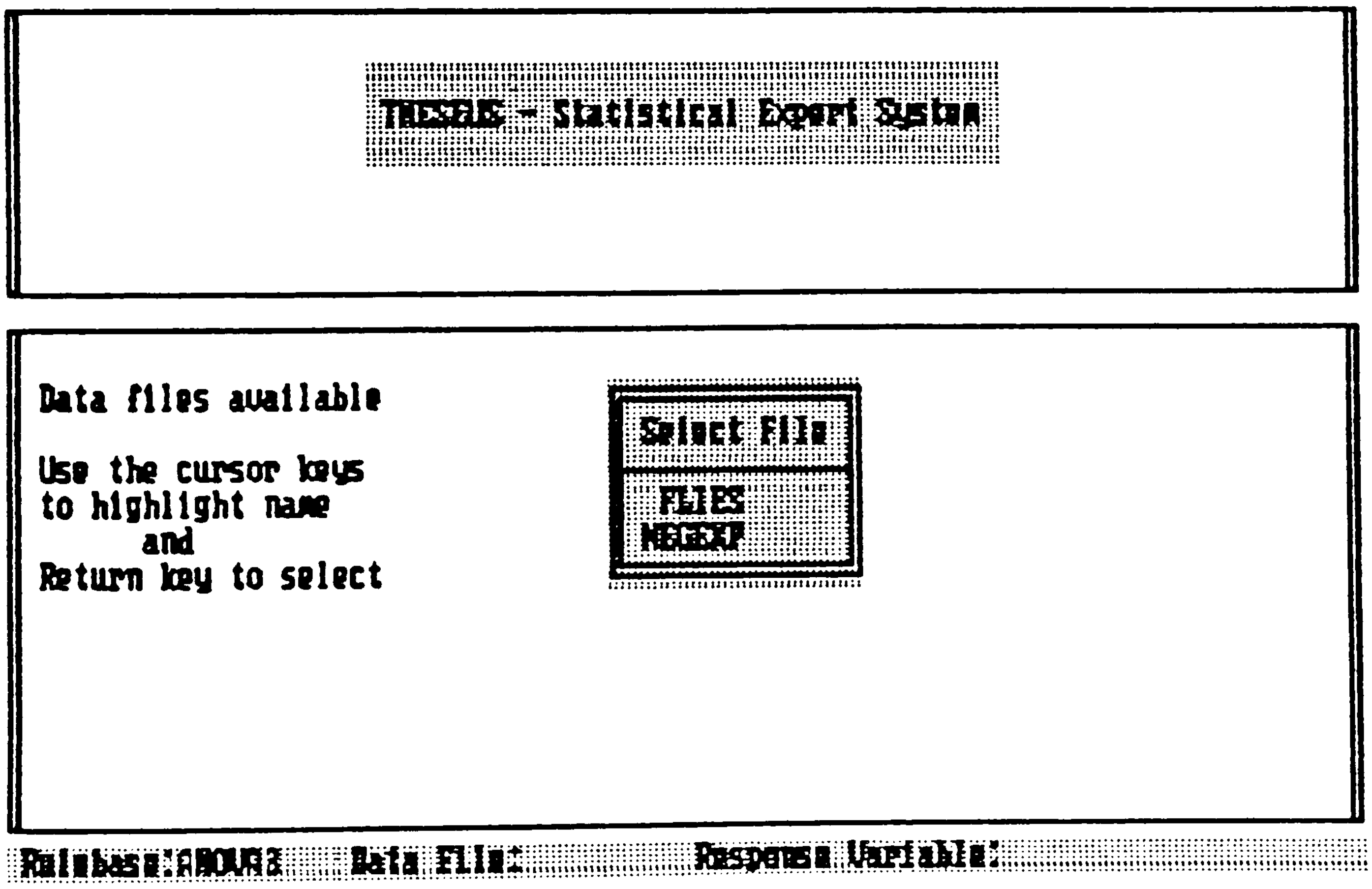
```
THESEUS - Statistical Expert System

Rulebases available                Select File
                                   ANOVA3
Use the cursor keys                ANOVA2
to highlight name
      and
Return key to select

Rulebase:          Data File:          Response Variable:
```

Figure 8.20 : Consultation A - 3



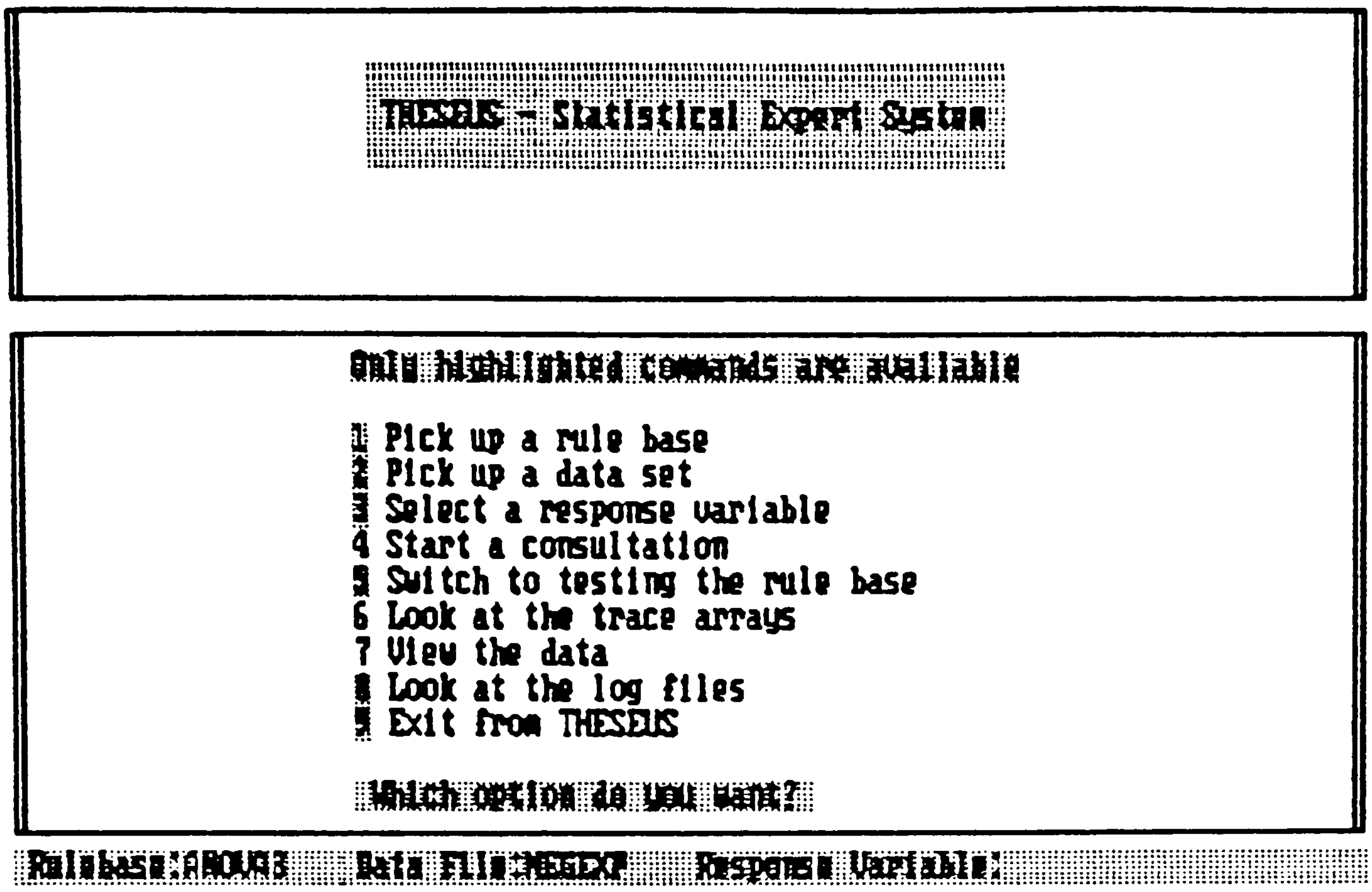
```
THESEUS - Statistical Expert System

Only highlighted commands are available

1 Pick up a rule base
2 Pick up a data set
3 Select a response variable
4 Start a consultation
5 Switch to testing the rule base
6 Look at the trace arrays
7 View the data
8 Look at the log files
9 Exit from THESEUS

Which option do you want?
Rulebase:ANOVA3   Data File:         Response Variable:
```

Figure 8.21 : Consultation A - 4



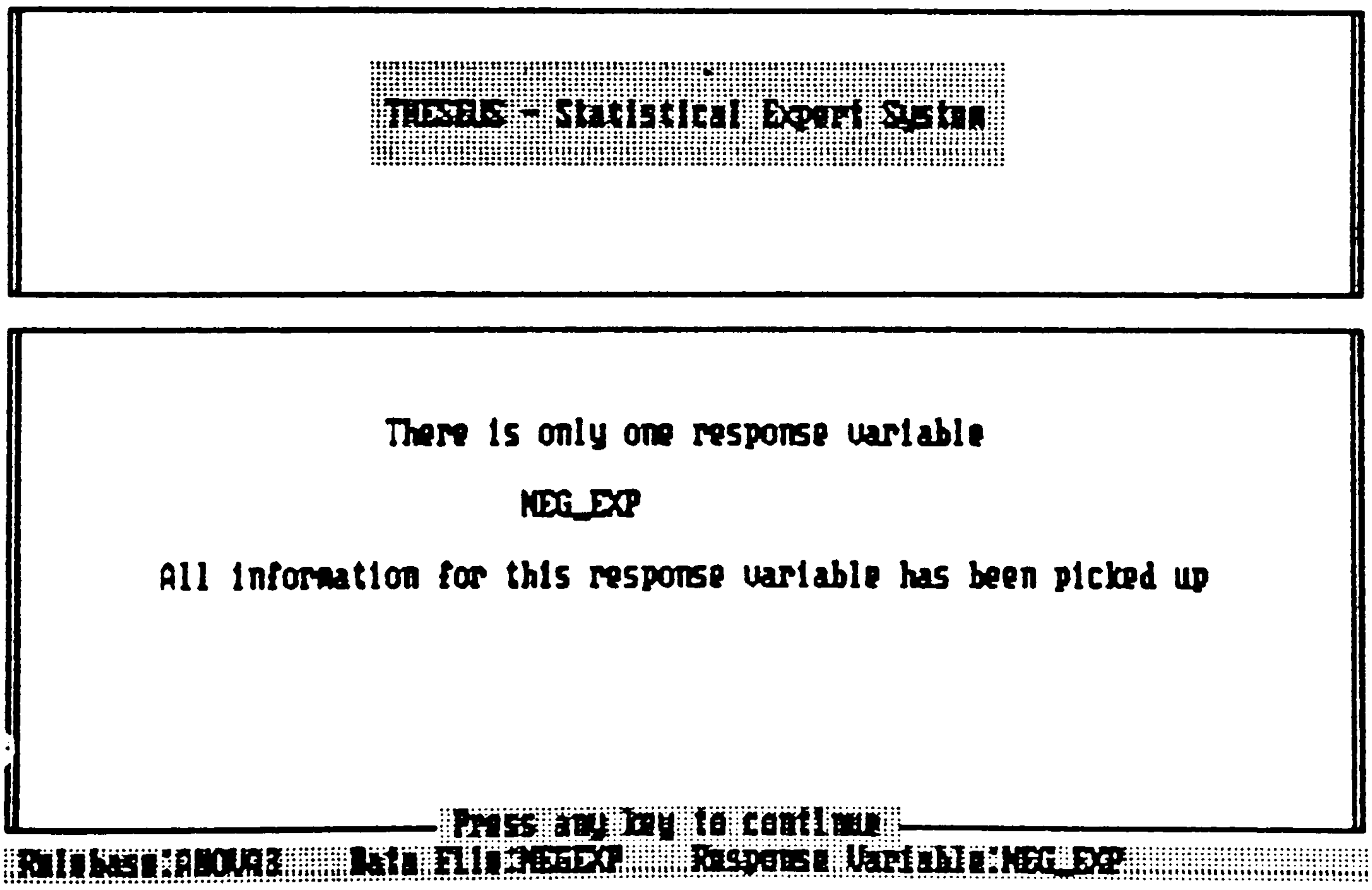
```
THESEUS - Statistical Expert System

Data files available            Select File
                                 FLIES
Use the cursor keys              NEGEXP
to highlight name
      and
Return key to select

Rulebase:ANOVA3   Data File:         Response Variable:
```

Figure 8.22 : Consultation A - 5



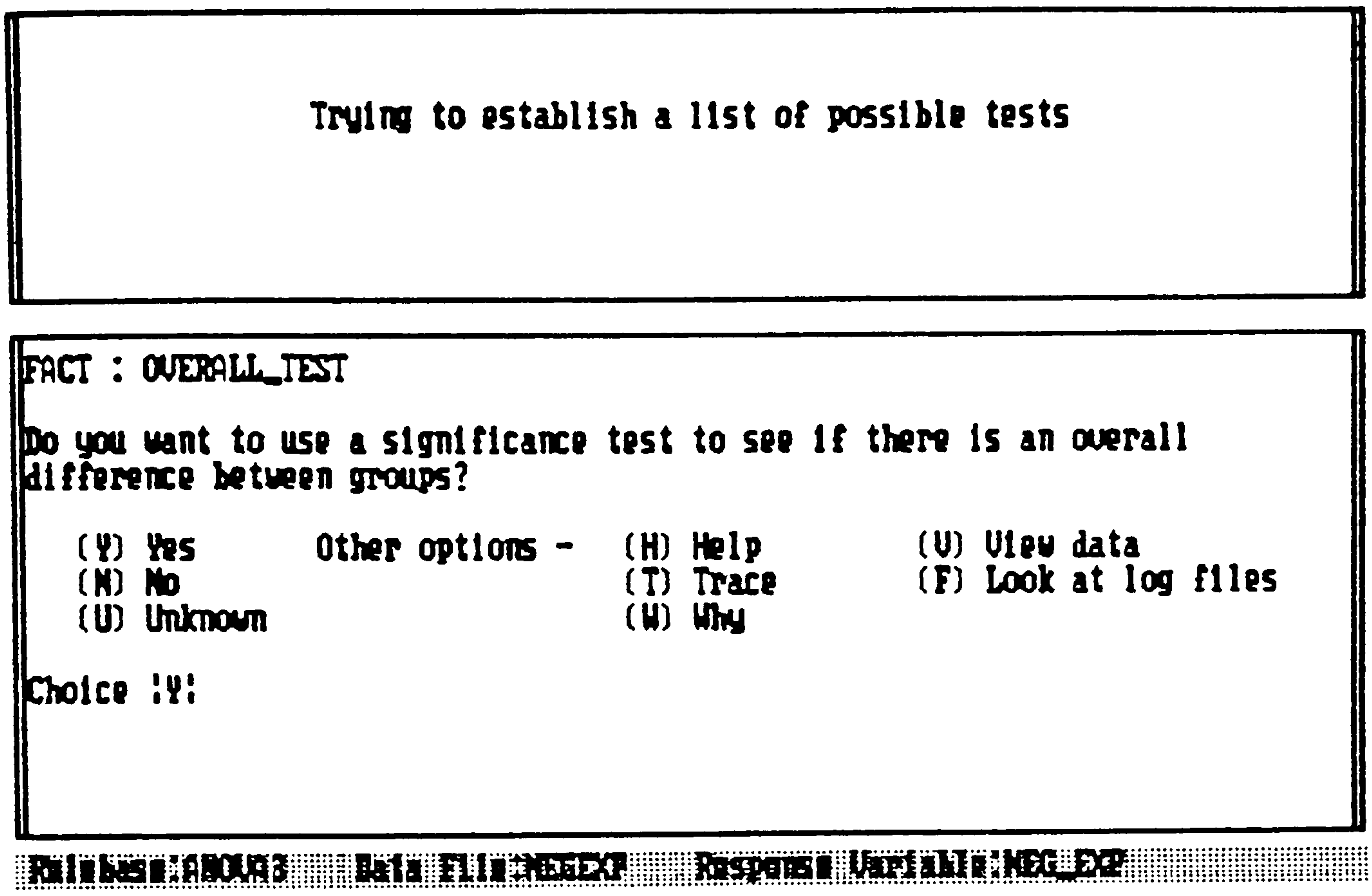
```
THESEUS - Statistical Expert System

Only highlighted commands are available

1 Pick up a rule base
2 Pick up a data set
3 Select a response variable
4 Start a consultation
5 Switch to testing the rule base
6 Look at the trace arrays
7 View the data
8 Look at the log files
9 Exit from THESEUS

Which option do you want?
Rulebase:ANOVA3   Data File:NEGEXP   Response Variable:
```

Figure 8.23 : Consultation A - 6



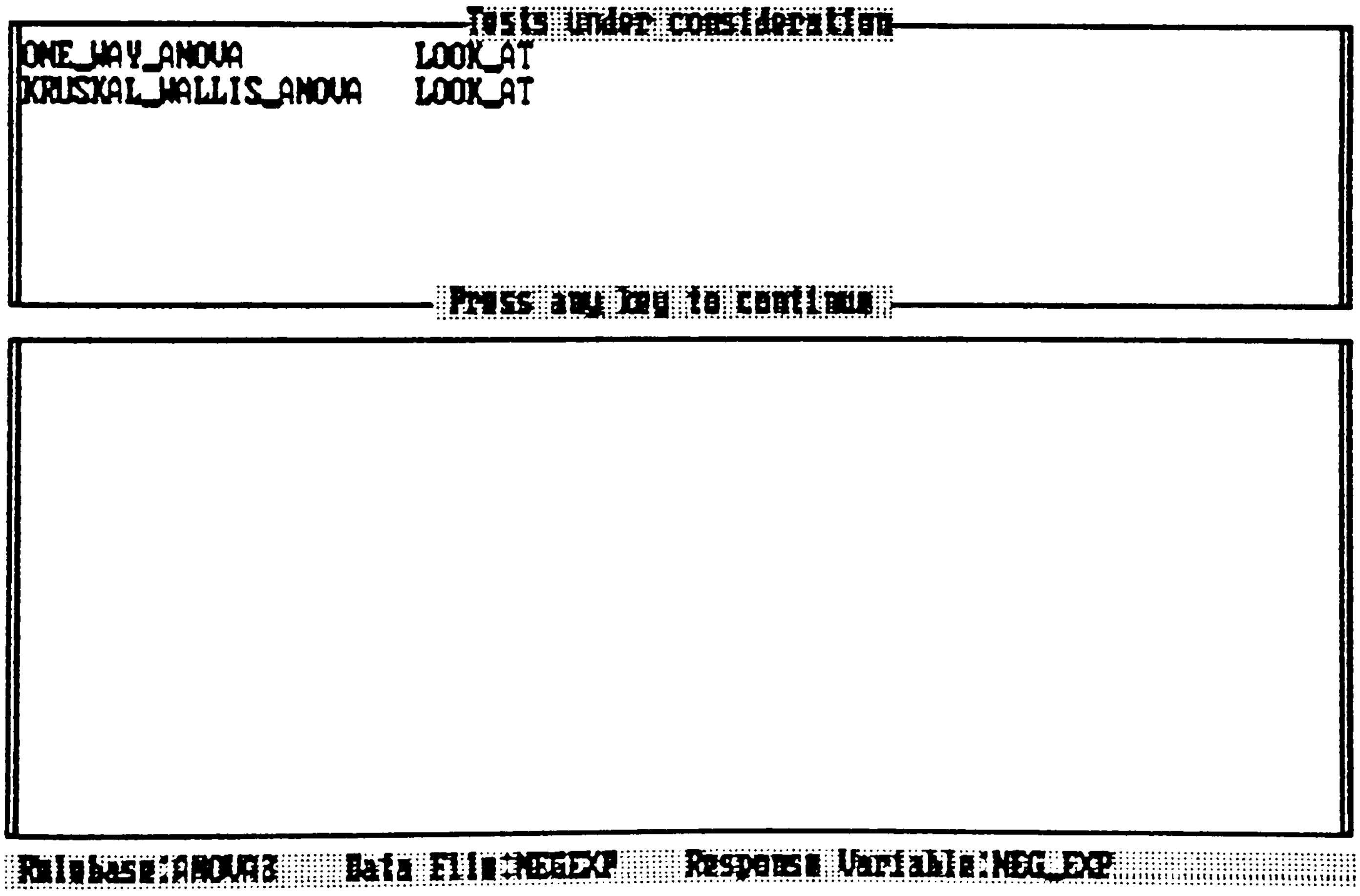
```
THESEUS - Statistical Expert System

There is only one response variable
NEG_EXP
All information for this response variable has been picked up

Press any key to continue
Rulebase:ANOVA3   Data File:NEGEXP   Response Variable:NEG_EXP
```

Figure 8.24 : Consultation A - 7



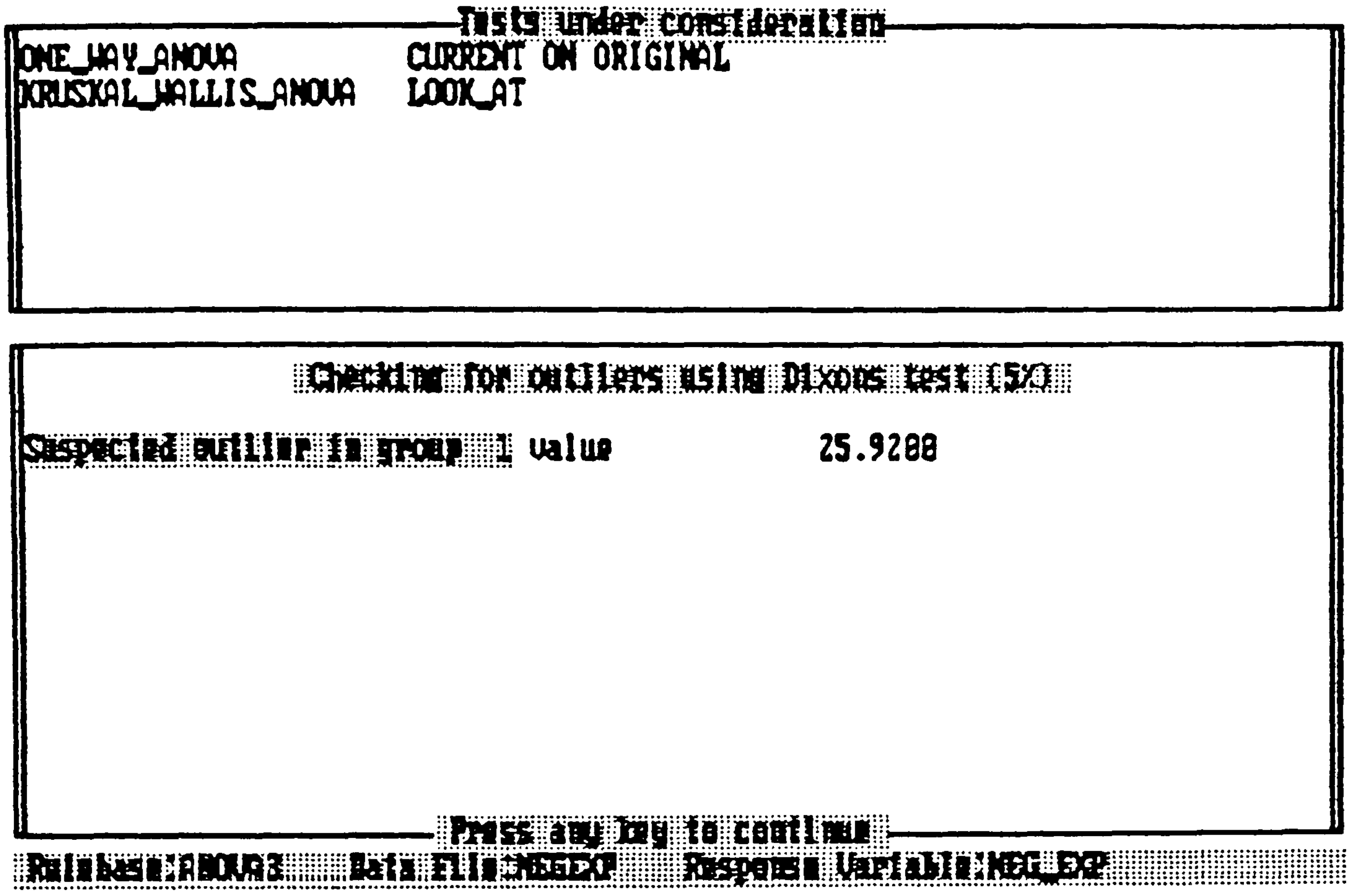
```
                Trying to establish a list of possible tests

FACT : OVERALL_TEST

Do you want to use a significance test to see if there is an overall
difference between groups?

   (Y) Yes       Other options -  (H) Help        (V) View data
   (N) No                         (T) Trace       (F) Look at log files
   (U) Unknown                    (W) Why

Choice :Y:

Rulebase:ANOVA3   Data File:NEGEXP   Response Variable:NEG_EXP
```

Figure 8.25 : Consultation A - 8



```
                       Tests under consideration
ONE_WAY_ANOVA          LOOK_AT
KRUSKAL_WALLIS_ANOVA   LOOK_AT

                       Press any key to continue

Rulebase:ANOVA3   Data File:NEGEXP   Response Variable:NEG_EXP
```

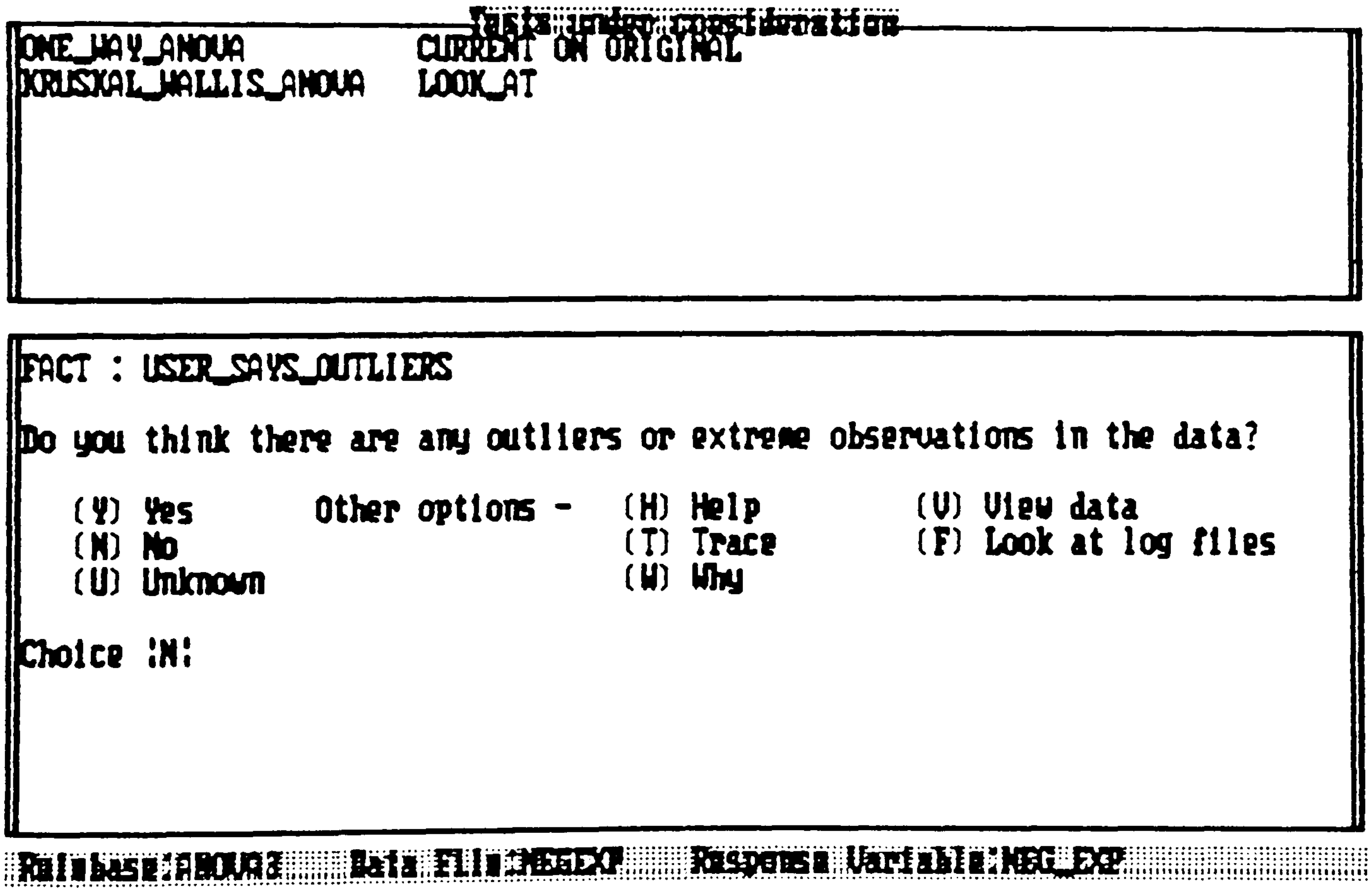
Figure 8.26 : Consultation A - 9



```
                         Tests under consideration
ONE_WAY_ANOVA           CURRENT ON ORIGINAL
KRUSKAL_WALLIS_ANOVA    LOOK_AT
```

```
              Checking for outliers using Dixons test (5%)

Suspected outlier in group 1 value           25.9288



                        Press any key to continue
Rulebase:ANOVA3   Data File:NEGEXP   Response Variable:NEG_EXP
```

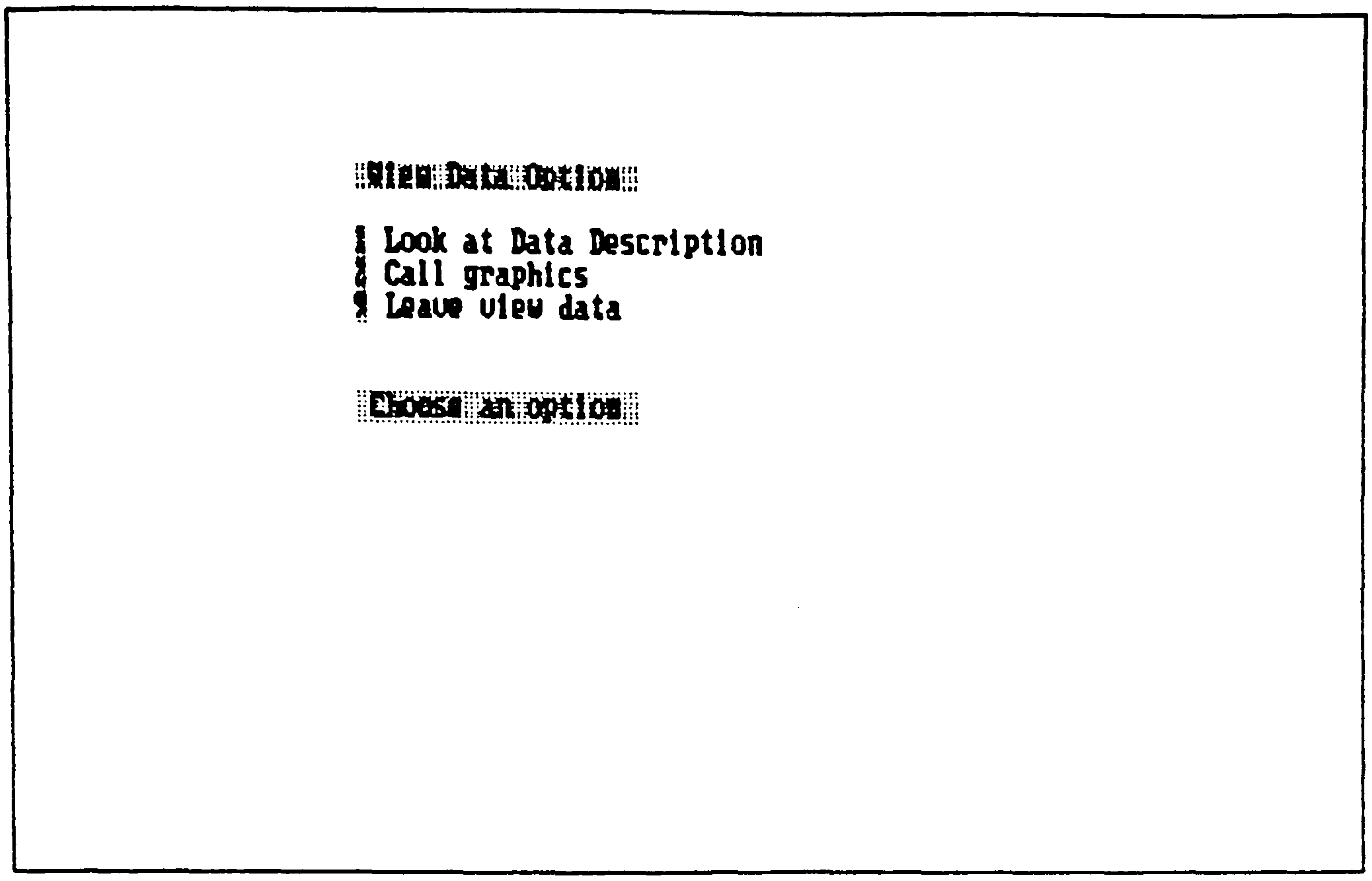
Figure 8.27 : Consultation A - 10



```
                         Tests under consideration
ONE_WAY_ANOVA           CURRENT ON ORIGINAL
KRUSKAL_WALLIS_ANOVA    LOOK_AT
```

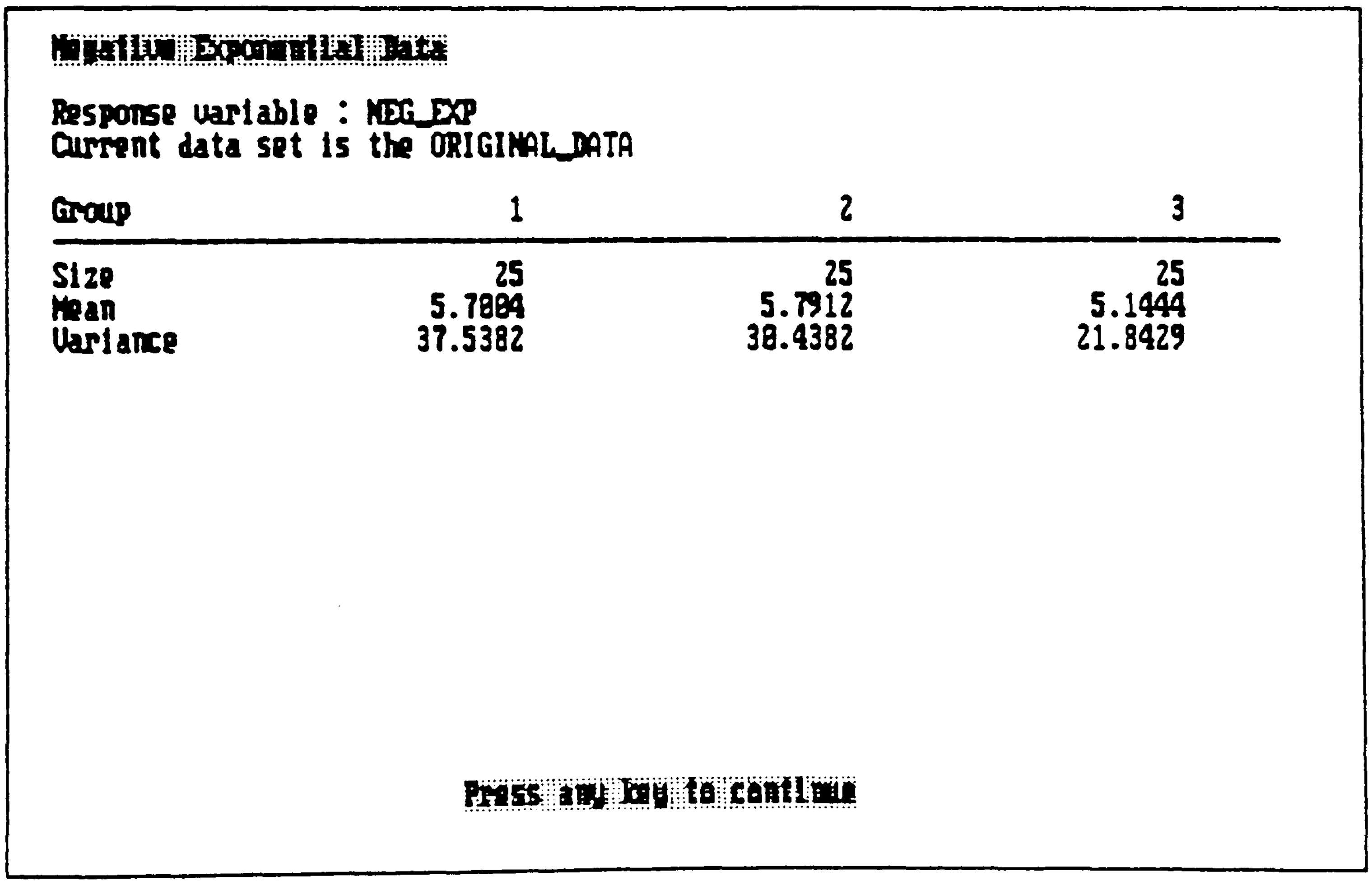
```
FACT : USER_SAYS_OUTLIERS

Do you think there are any outliers or extreme observations in the data?

   (Y) Yes      Other options -   (H) Help        (V) View data
   (N) No                         (T) Trace       (F) Look at log files
   (U) Unknown                    (W) Why

Choice :N:

Rulebase:ANOVA3   Data File:NEGEXP   Response Variable:NEG_EXP
```

Figure 8.28 : Consultation A - 11



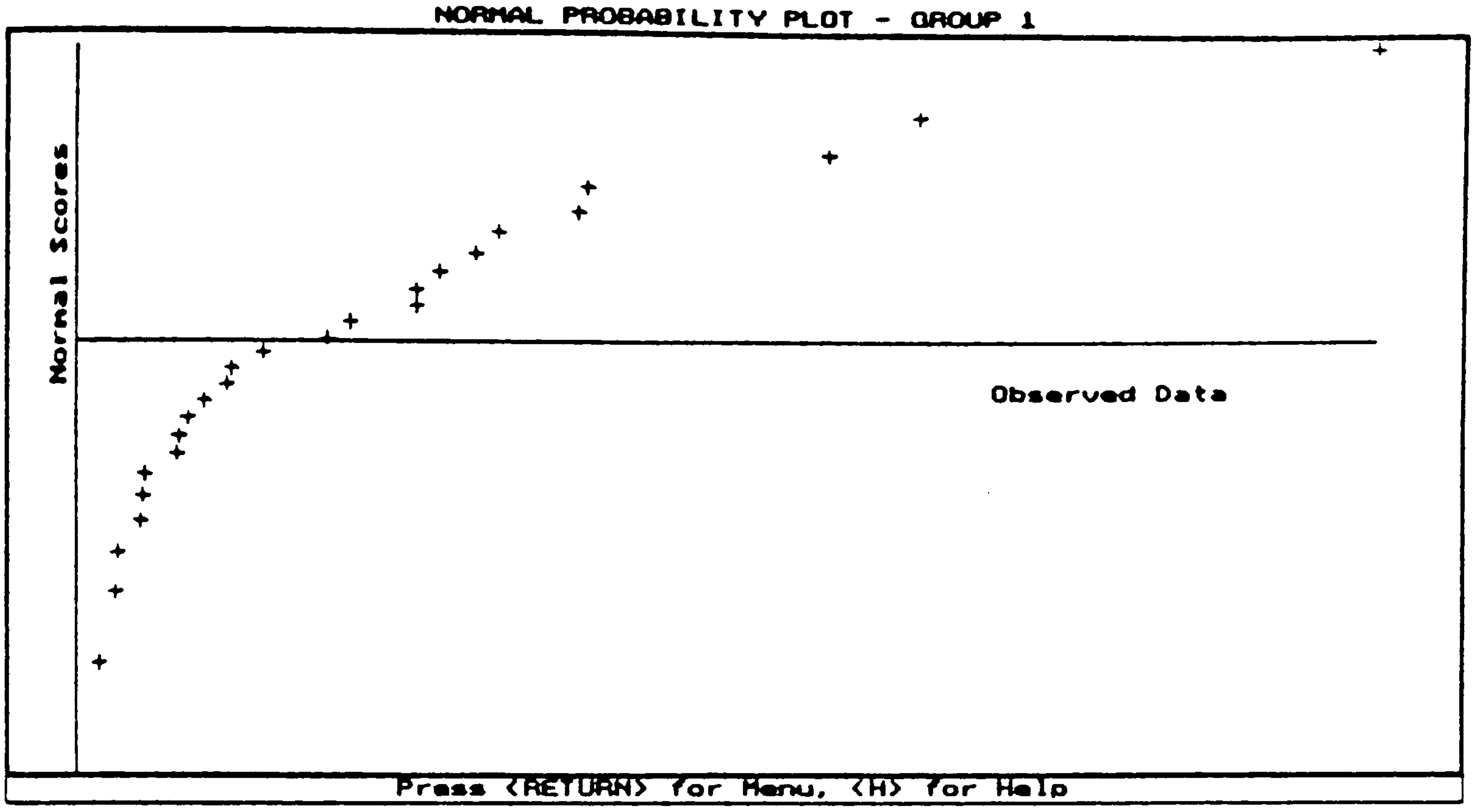
```
View Data Options

1 Look at Data Description
2 Call graphics
9 Leave view data


Choose an option
```

Figure 8.29 : Consultation A - 12



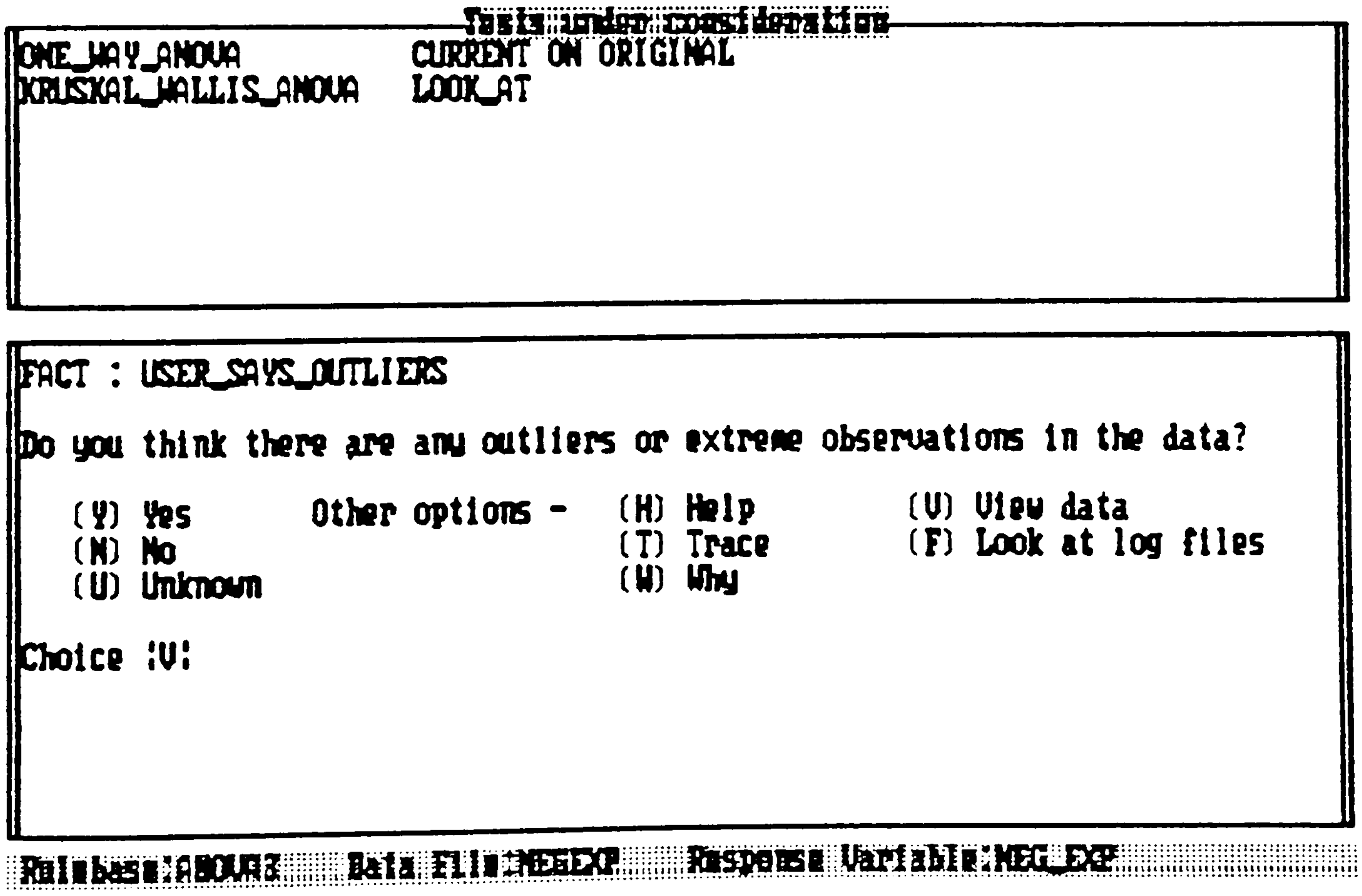
```
Negative Exponential Data

Response variable : NEG_EXP
Current data set is the ORIGINAL_DATA

Group                          1                  2                  3
-----------------------------------------------------------------------
Size                          25                 25                 25
Mean                      5.7884             5.7912             5.1444
Variance                 37.5382            38.4382            21.8429



                 Press any key to continue
```

Figure 8.30 : Consultation A - 13



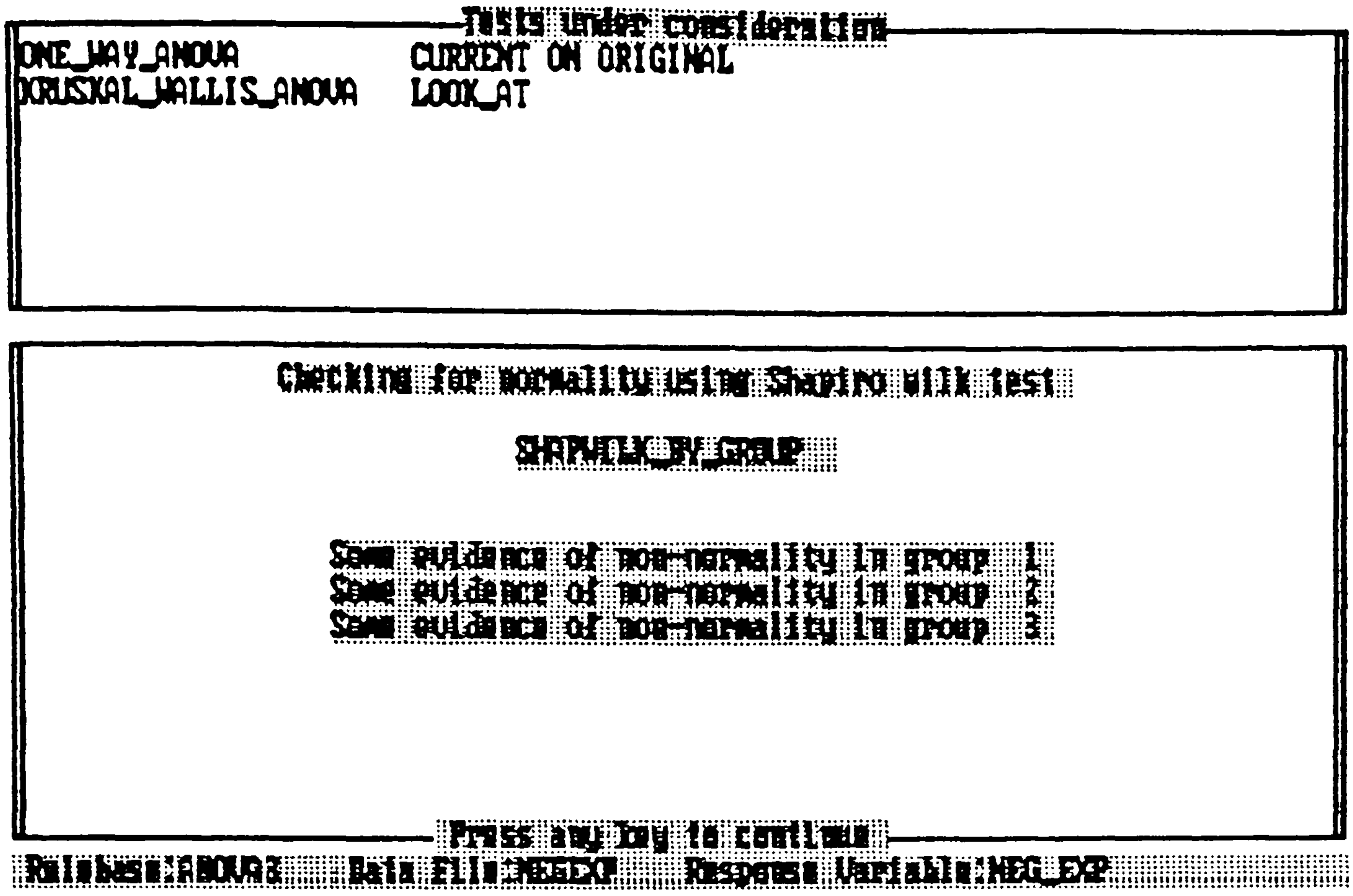
```
Press <RETURN> for Menu, <H> for Help
```

Figure 8.31 : Consultation A - 14



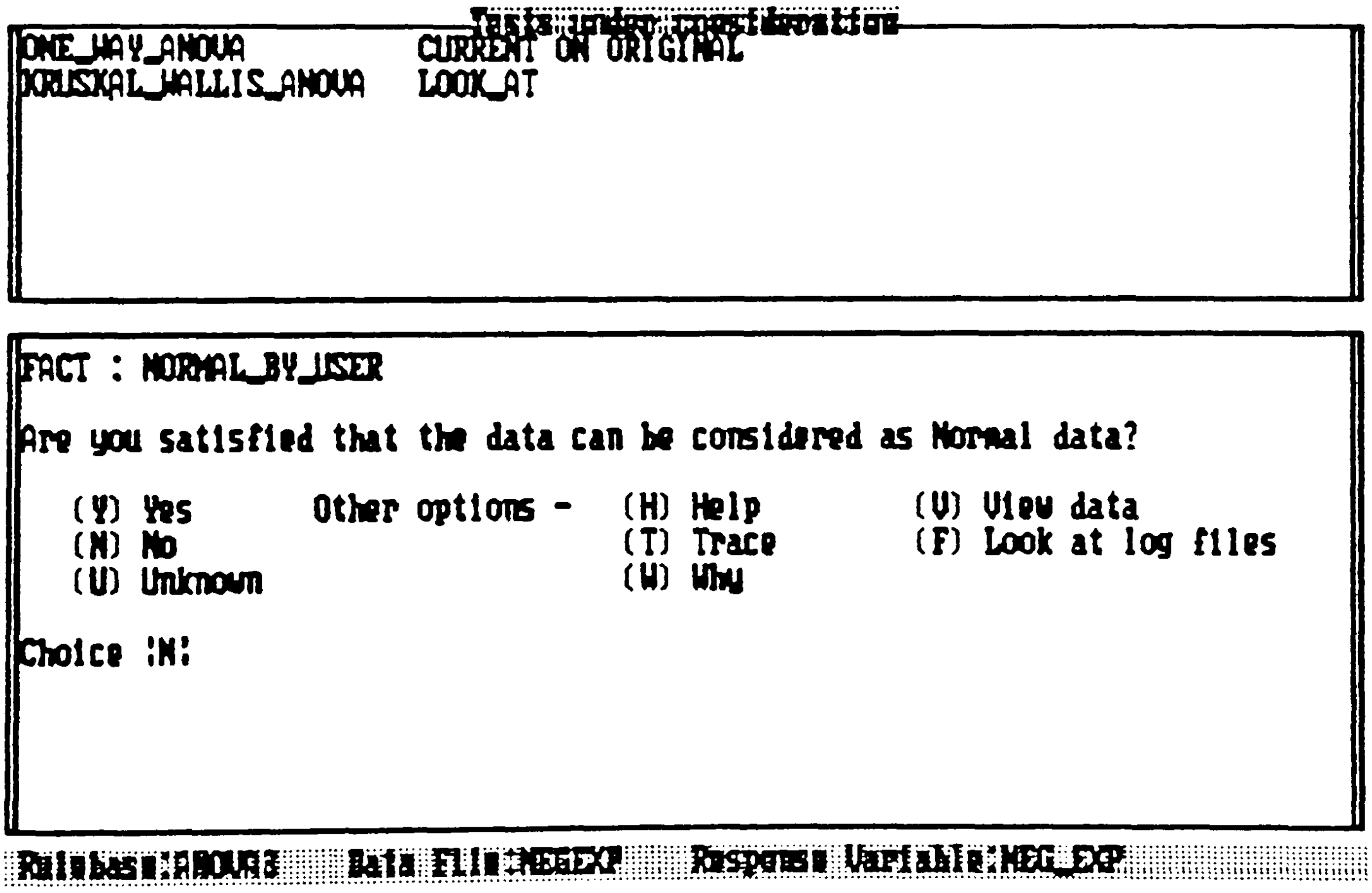
```
Tests under consideration
ONE_WAY_ANOVA           CURRENT ON ORIGINAL
KRUSKAL_WALLIS_ANOVA    LOOK_AT

FACT : USER_SAYS_OUTLIERS

Do you think there are any outliers or extreme observations in the data?

   (Y) Yes       Other options -   (H) Help          (V) View data
   (N) No                          (T) Trace         (F) Look at log files
   (U) Unknown                     (W) Why

Choice :V:

Rulebase:ANOVA3   Data File:MEGEXP   Response Variable:MEG_EXP
```

Figure 8.32 : Consultation A - 15



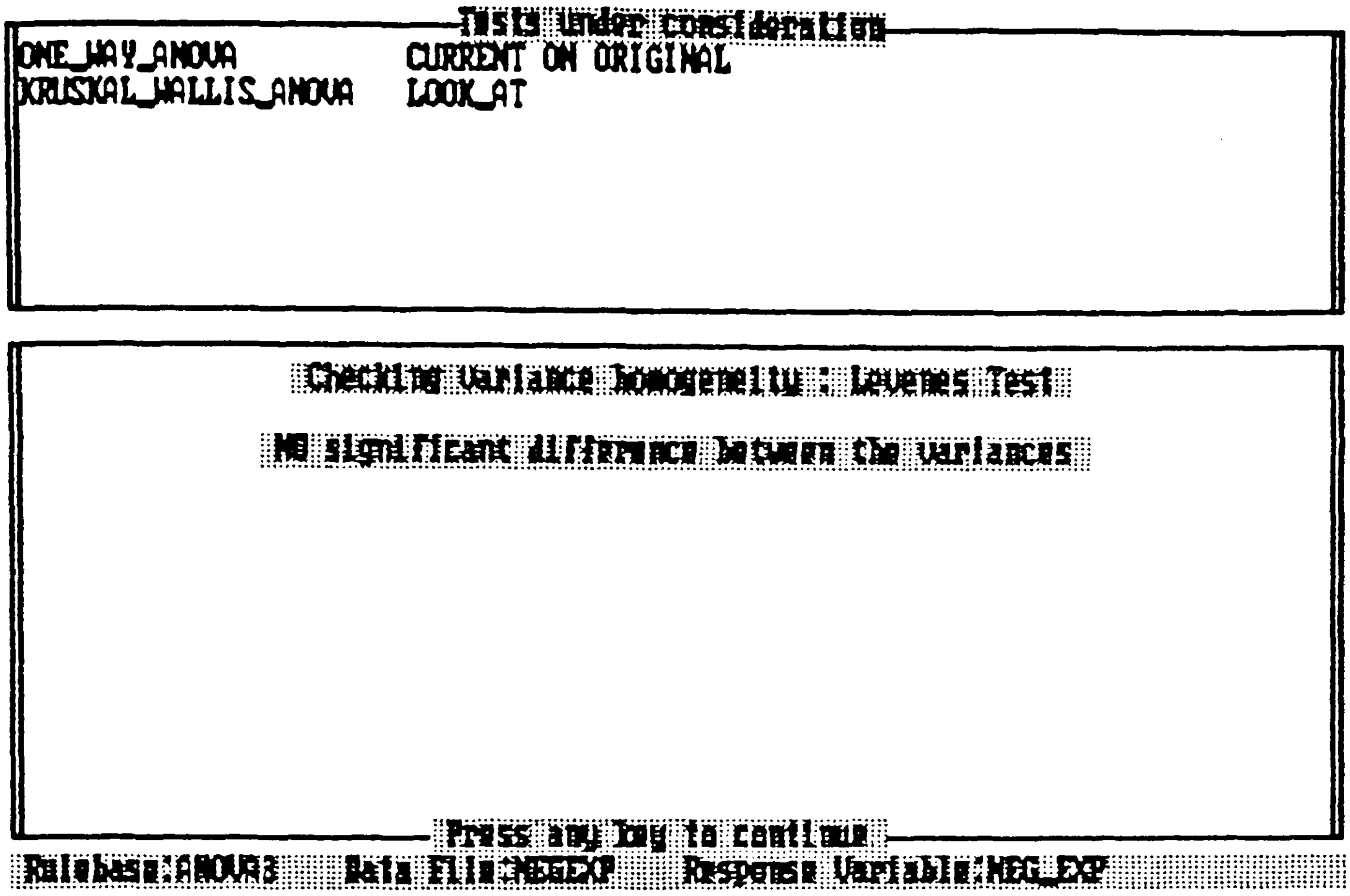
```
                         Tests under consideration
ONE_WAY_ANOVA          CURRENT ON ORIGINAL
KRUSKAL_WALLIS_ANOVA   LOOK_AT

              Checking for normality using Shapiro Wilk test

                         SHAPWILK_BY_GROUP

              Some evidence of non-normality in group  1
              Some evidence of non-normality in group  2
              Some evidence of non-normality in group  3

                          Press any key to continue
Rulebase:ANOVA3    Data File:NEGEXP    Response Variable:NEG_EXP
```

Figure 8.33 : Consultation A - 16



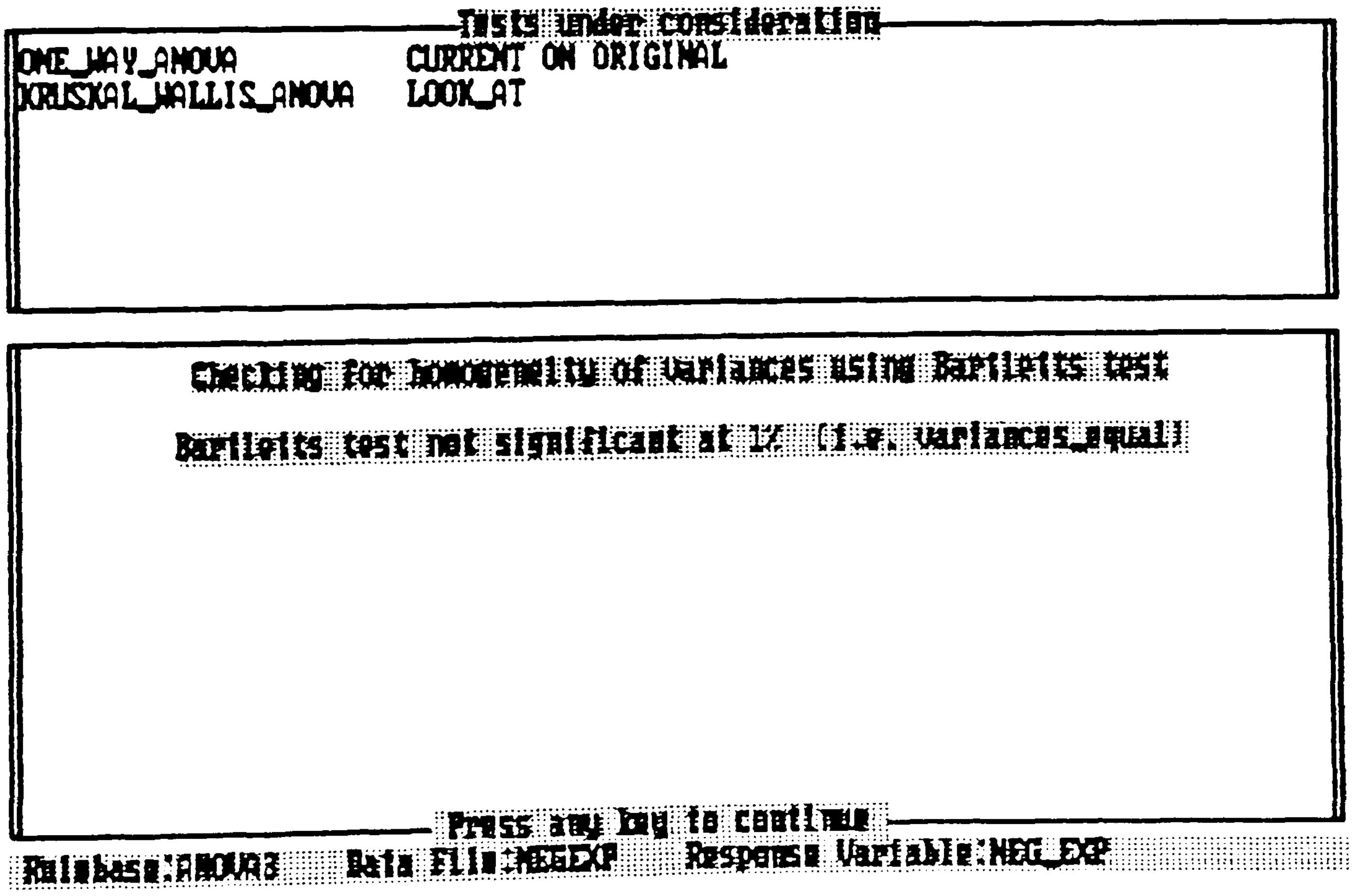
```
                         Tests under consideration
ONE_WAY_ANOVA          CURRENT ON ORIGINAL
KRUSKAL_WALLIS_ANOVA   LOOK_AT

FACT : NORMAL_BY_USER

Are you satisfied that the data can be considered as Normal data?

   (Y) Yes      Other options -   (H) Help       (V) View data
   (N) No                         (T) Trace      (F) Look at log files
   (U) Unknown                    (W) Why

Choice :N:

Rulebase:ANOVA3    Data File:NEGEXP    Response Variable:NEG_EXP
```

Figure 8.34 : Consultation A - 17



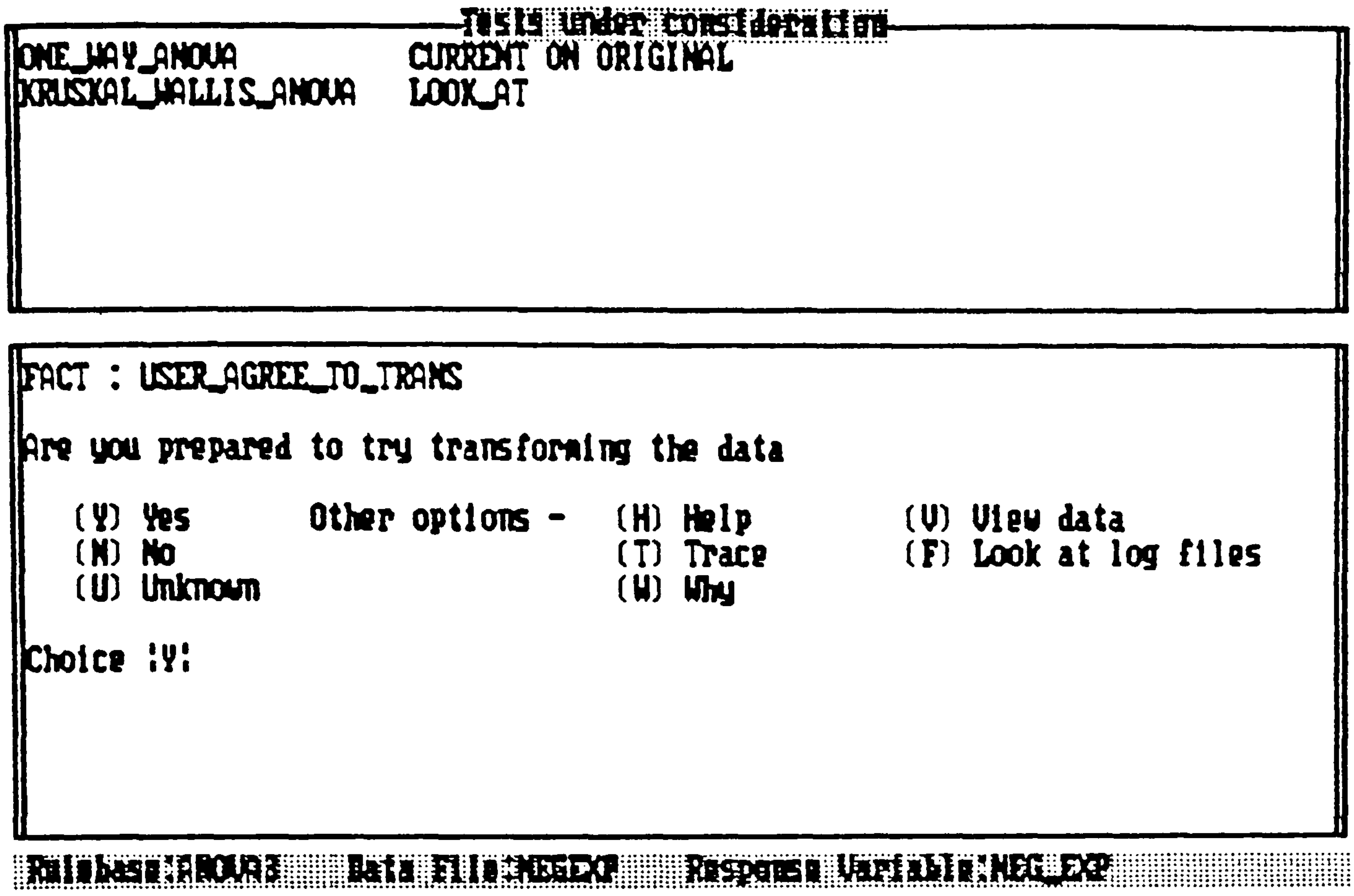
```
                       Tests under consideration
ONE_WAY_ANOVA          CURRENT ON ORIGINAL
KRUSKAL_WALLIS_ANOVA   LOOK_AT

           Checking variance homogeneity : Levenes Test

          NO significant difference between the variances

                       Press any key to continue
Rulebase:ANOVA3   Data File:NEGEXP   Response Variable:NEG_EXP
```

Figure 8.35 : Consultation A - 18



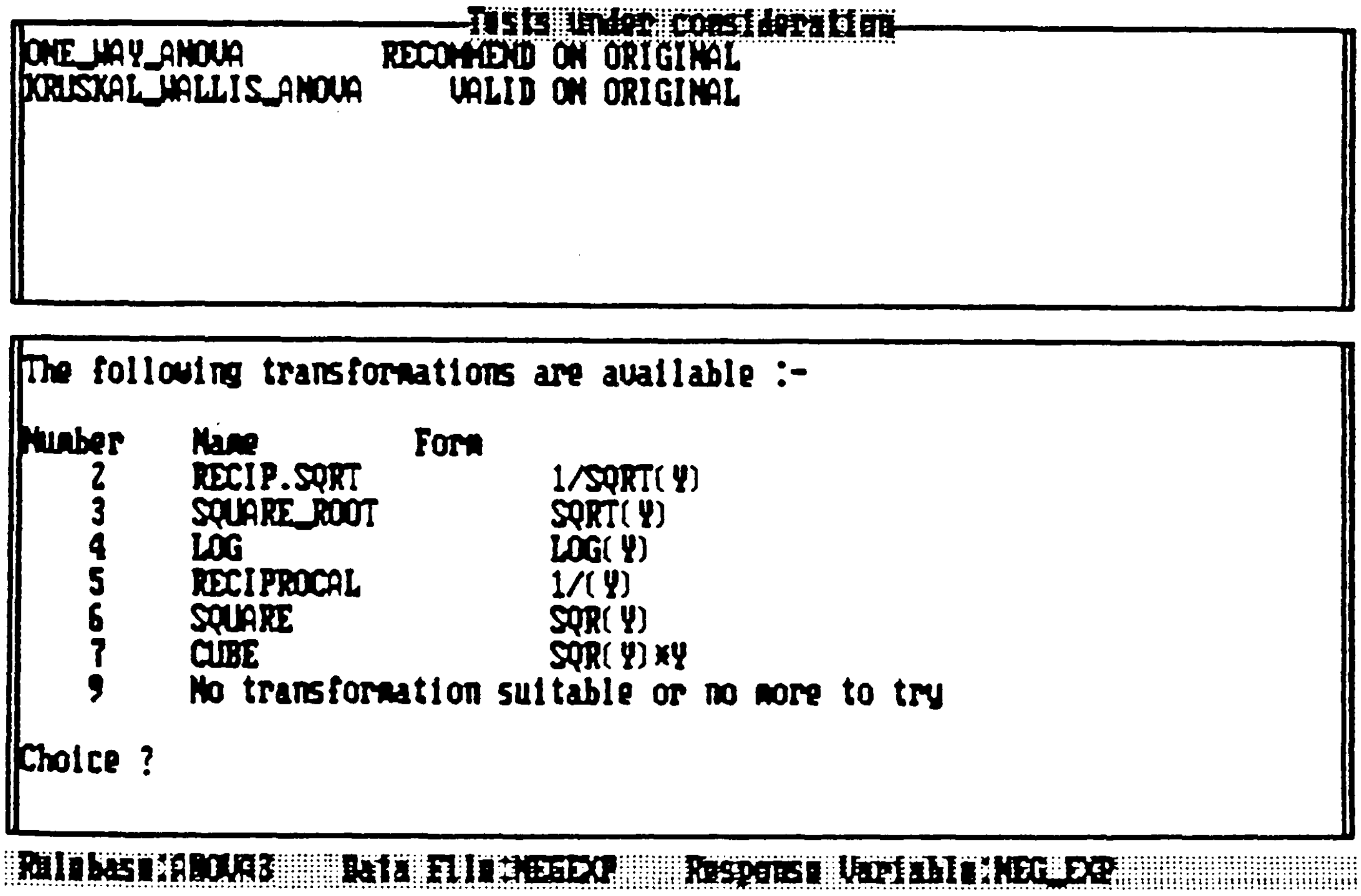
```
                       Tests under consideration
ONE_WAY_ANOVA          CURRENT ON ORIGINAL
KRUSKAL_WALLIS_ANOVA   LOOK_AT

      Checking for homogeneity of variances using Bartletts test

     Bartletts test not significant at 1% (i.e. variances_equal)

                       Press any key to continue
Rulebase:ANOVA3   Data File:NEGEXP   Response Variable:NEG_EXP
```

Figure 8.36 : Consultation A - 19



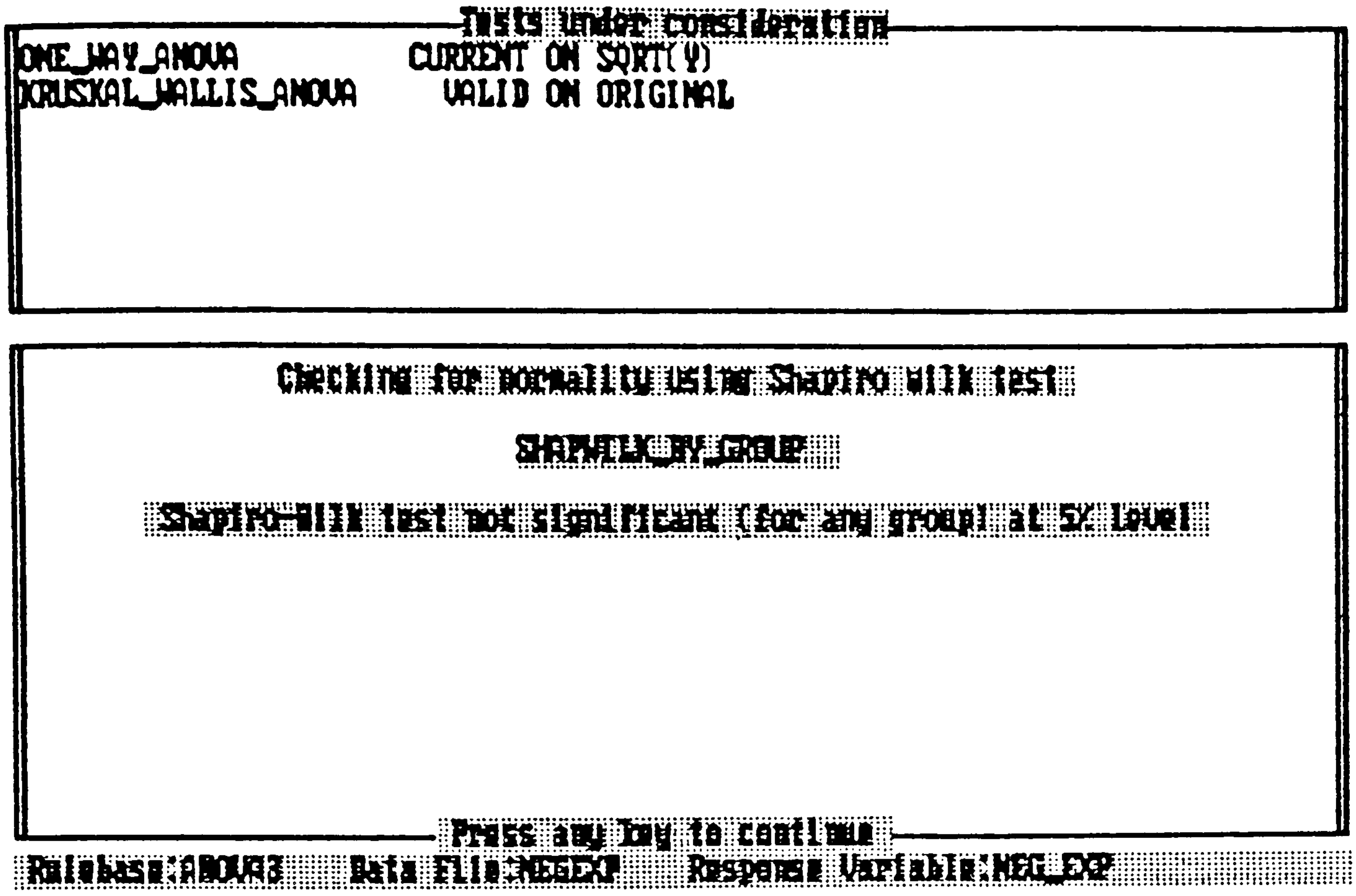
```
                              Tests under consideration
ONE_WAY_ANOVA              CURRENT ON ORIGINAL
KRUSKAL_WALLIS_ANOVA       LOOK_AT

FACT : USER_AGREE_TO_TRANS

Are you prepared to try transforming the data

   (Y) Yes        Other options -   (H) Help          (V) View data
   (N) No                           (T) Trace         (F) Look at log files
   (U) Unknown                      (W) Why

Choice :Y:

Rulebase:ANOVA3    Data File:NEGEXP    Response Variable:NEG_EXP
```

Figure 8.37 : Consultation A - 20



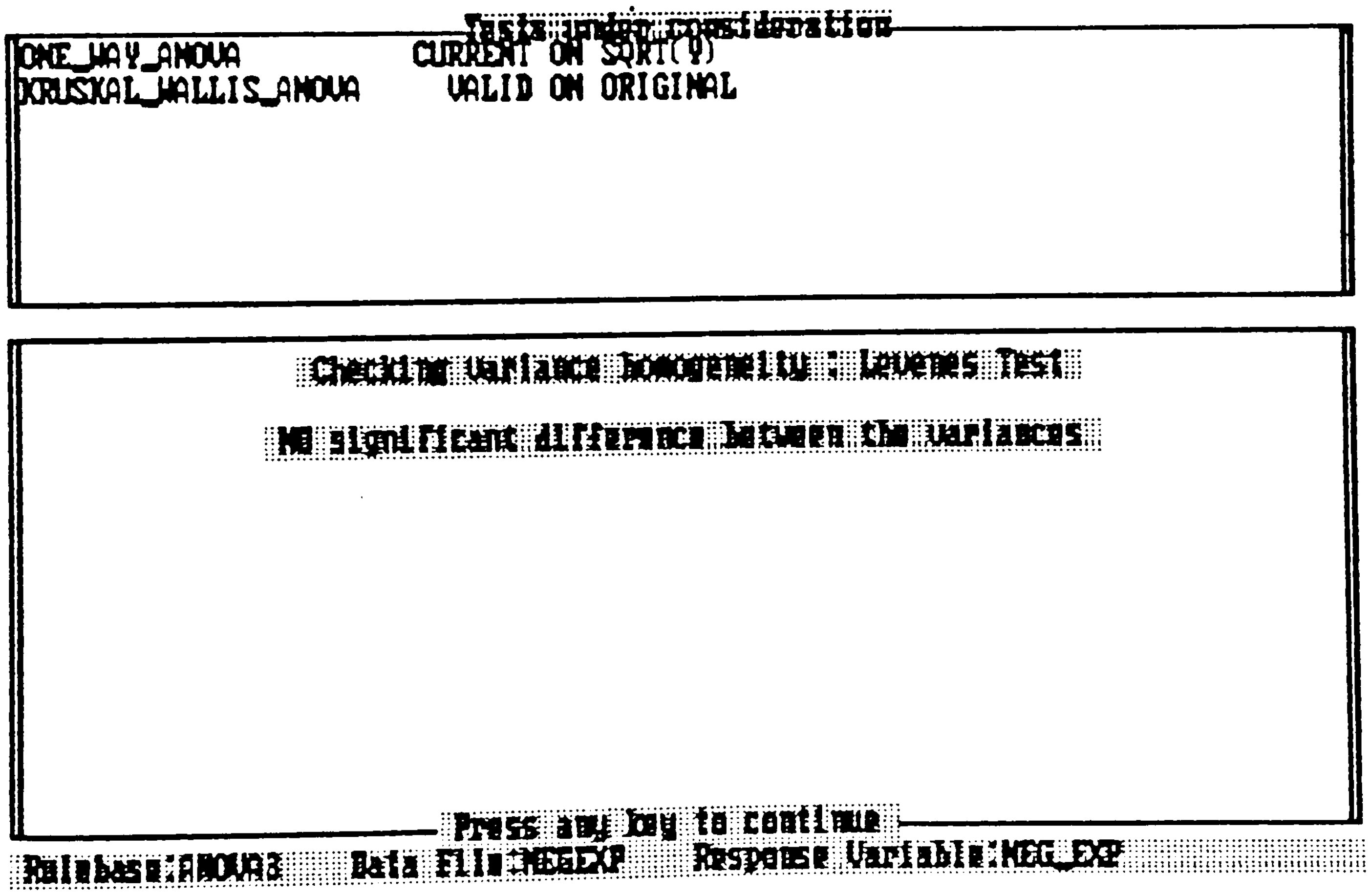
```
                              Tests under consideration
ONE_WAY_ANOVA            RECOMMEND ON ORIGINAL
KRUSKAL_WALLIS_ANOVA         VALID ON ORIGINAL

The following transformations are available :-

Number     Name           Form
   2       RECIP.SQRT             1/SQRT(Y)
   3       SQUARE_ROOT            SQRT(Y)
   4       LOG                    LOG(Y)
   5       RECIPROCAL             1/(Y)
   6       SQUARE                 SQR(Y)
   7       CUBE                   SQR(Y)*Y
   9       No transformation suitable or no more to try

Choice ?

Rulebase:ANOVA3    Data File:NEGEXP    Response Variable:NEG_EXP
```

Figure 8.38 : Consultation A - 21



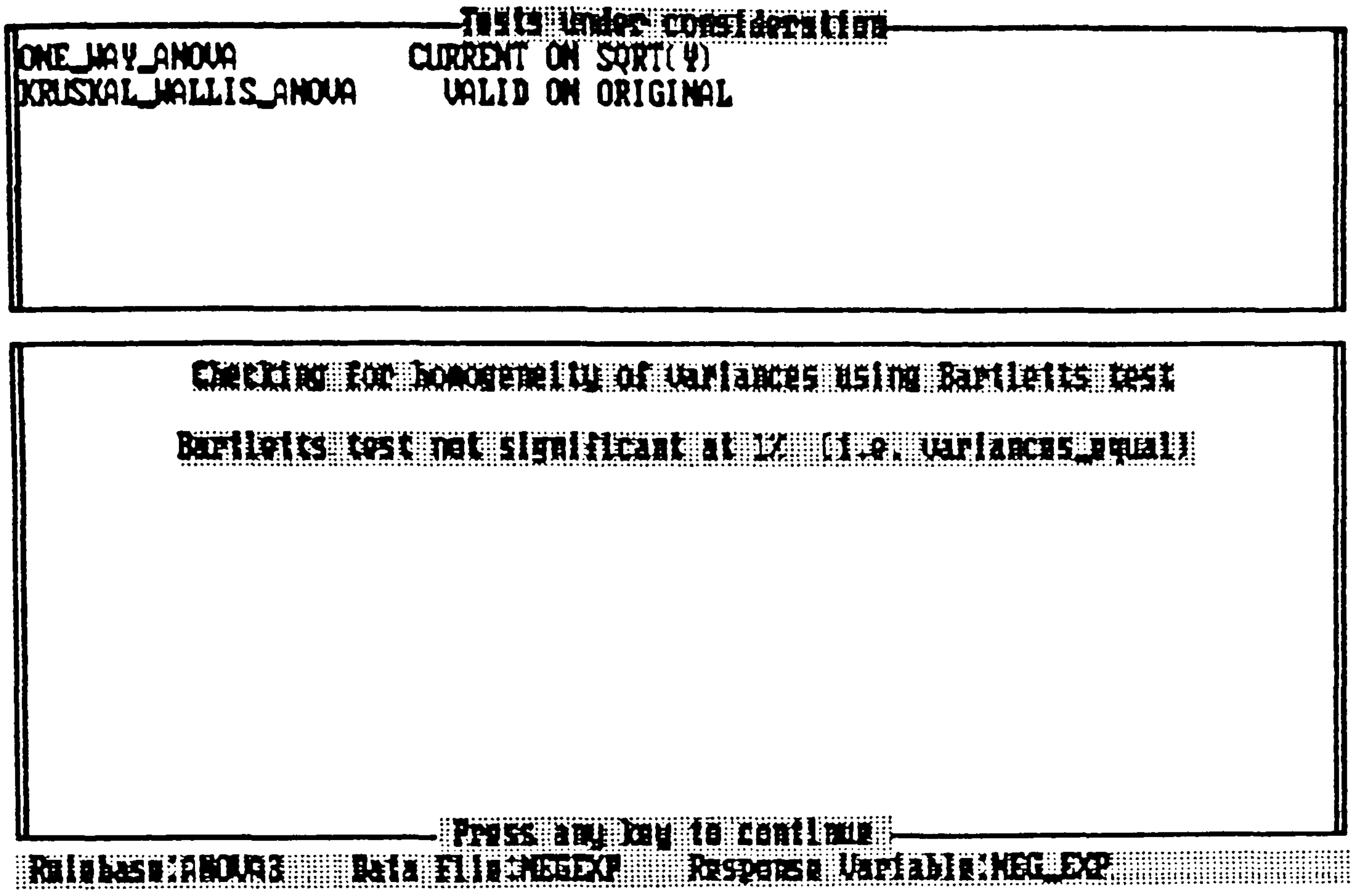
```
                    Tests under consideration
ONE_WAY_ANOVA           CURRENT ON SQRT(Y)
KRUSKAL_WALLIS_ANOVA      VALID ON ORIGINAL

          Checking for normality using Shapiro Wilk test

                        SHAPWILK_BY_GROUP

    Shapiro-Wilk test not significant (for any group) at 5% level

                      Press any key to continue
Rulebase:ANOVA3    Data File:NEGEXP    Response Variable:NEG_EXP
```

Figure 8.39 : Consultation A - 22



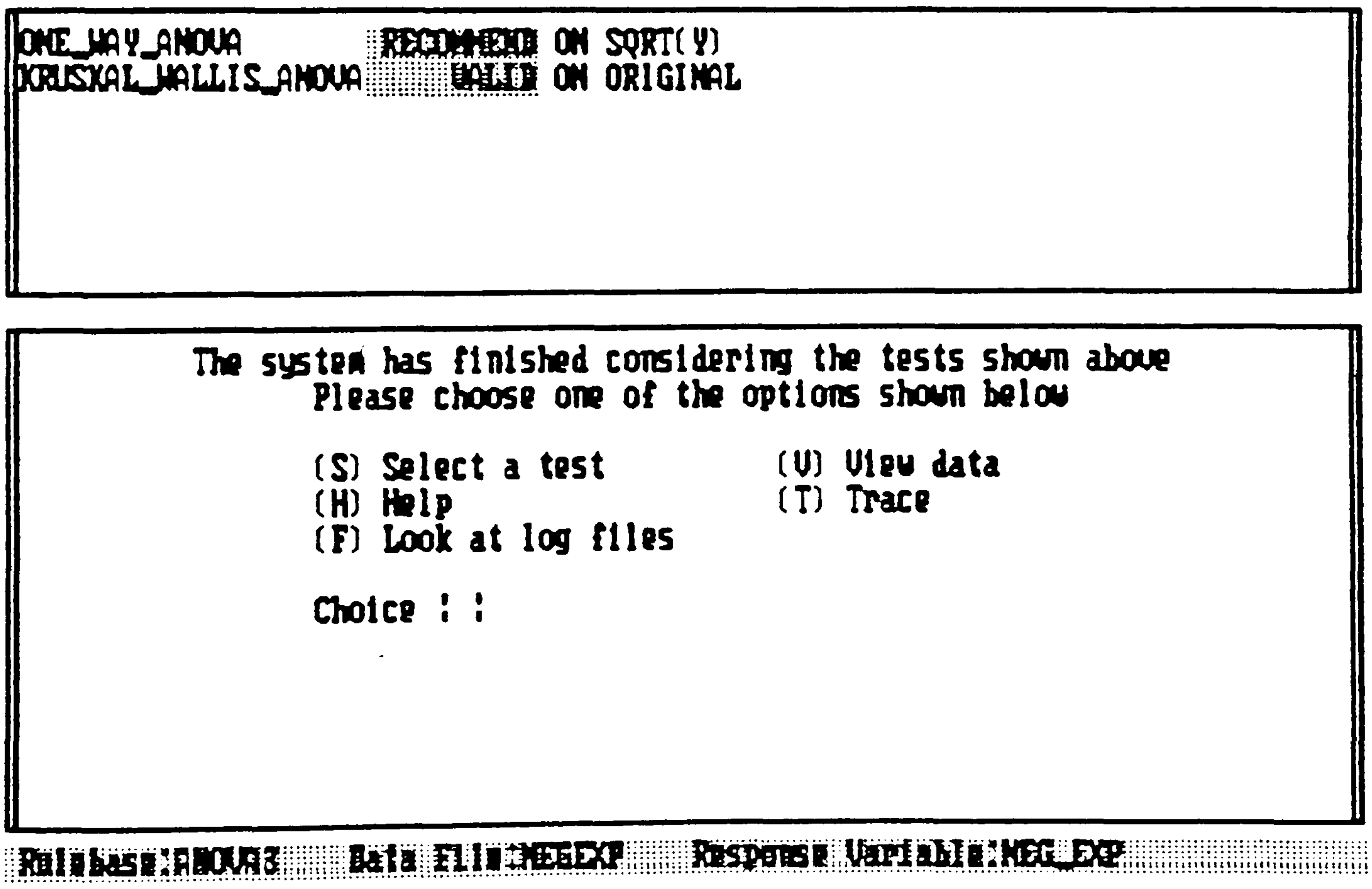
```
                    Tests under consideration
ONE_WAY_ANOVA           CURRENT ON SQRT(Y)
KRUSKAL_WALLIS_ANOVA      VALID ON ORIGINAL

            Checking variance homogeneity : Levenes Test

           NO significant difference between the variances

                      Press any key to continue
Rulebase:ANOVA3    Data File:NEGEXP    Response Variable:NEG_EXP
```

Figure 8.40 : Consultation A - 23



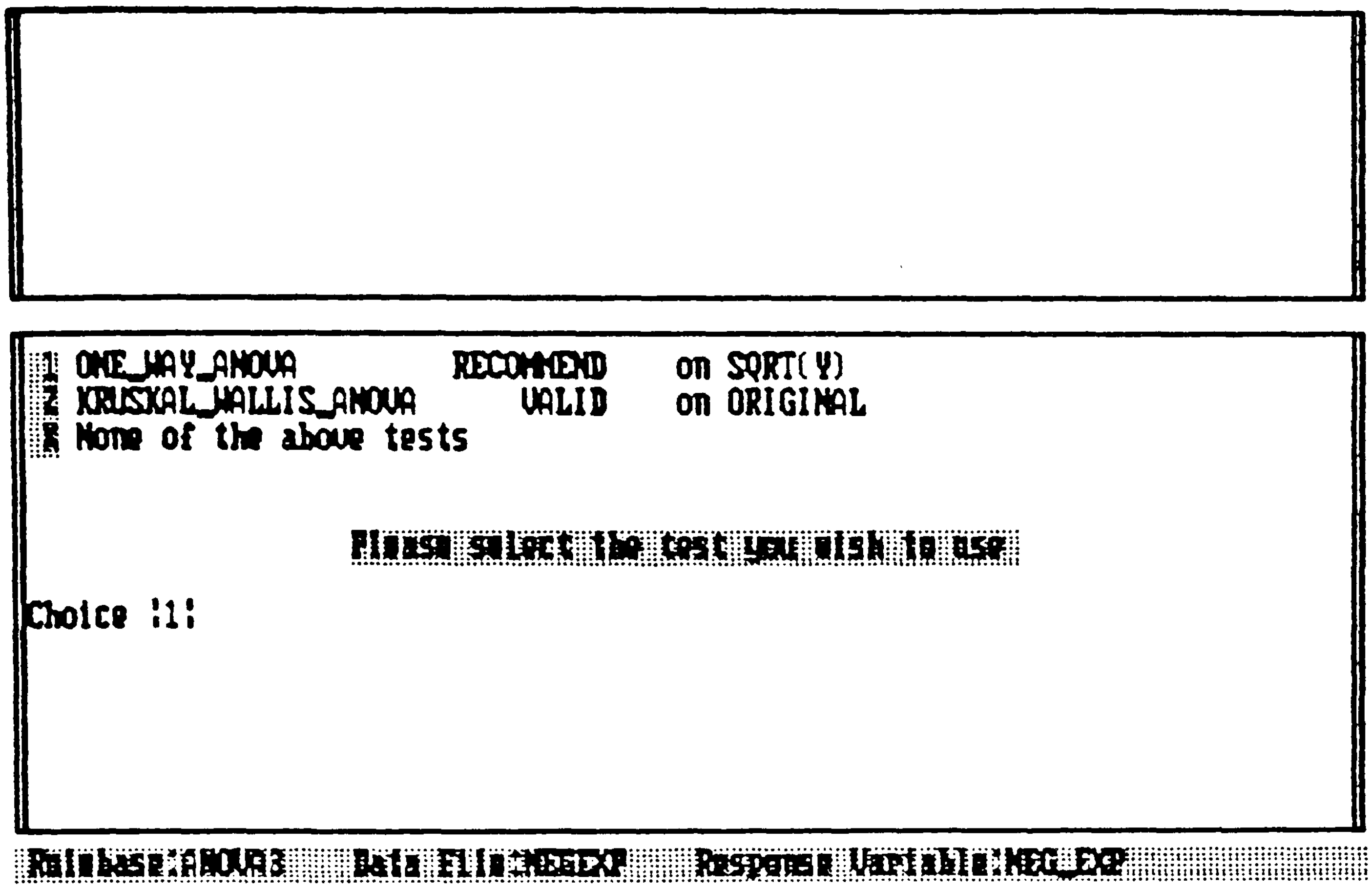
```
                              Tests under consideration
ONE_WAY_ANOVA          CURRENT ON SQRT(Y)
KRUSKAL_WALLIS_ANOVA     VALID ON ORIGINAL

        Checking for homogeneity of variances using Bartletts test

        Bartletts test not significant at 1%  (i.e. variances equal)

                              Press any key to continue
Rulebase:ANOVA3   Data File:NEGEXP   Response Variable:NEG_EXP
```

Figure 8.41 : Consultation A - 24



```
ONE_WAY_ANOVA          RECOMMEND ON SQRT(Y)
KRUSKAL_WALLIS_ANOVA       VALID ON ORIGINAL

        The system has finished considering the tests shown above
              Please choose one of the options shown below

              (S) Select a test           (V) View data
              (H) Help                    (T) Trace
              (F) Look at log files

              Choice : :

Rulebase:ANOVA3   Data File:NEGEXP   Response Variable:NEG_EXP
```

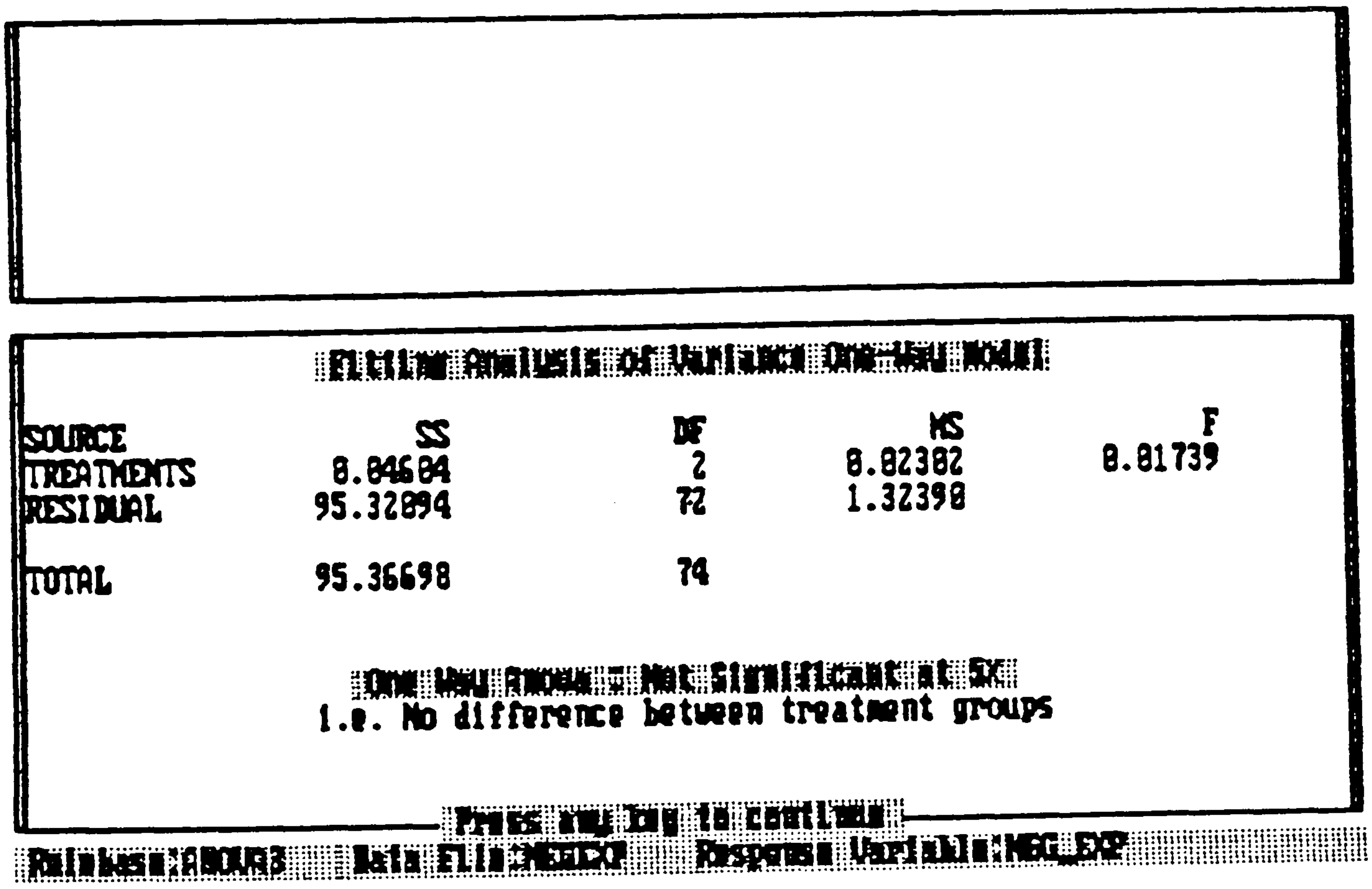
Figure 8.42 : Consultation A - 25



```
1 ONE_WAY_ANOVA          RECOMMEND    on SQRT(Y)
2 KRUSKAL_WALLIS_ANOVA       VALID    on ORIGINAL
X None of the above tests


              Please select the test you wish to use

Choice :1:
```

```
Rulebase:ANOVA3   Data File:NEGEXP   Response Variable:NEG_EXP
```

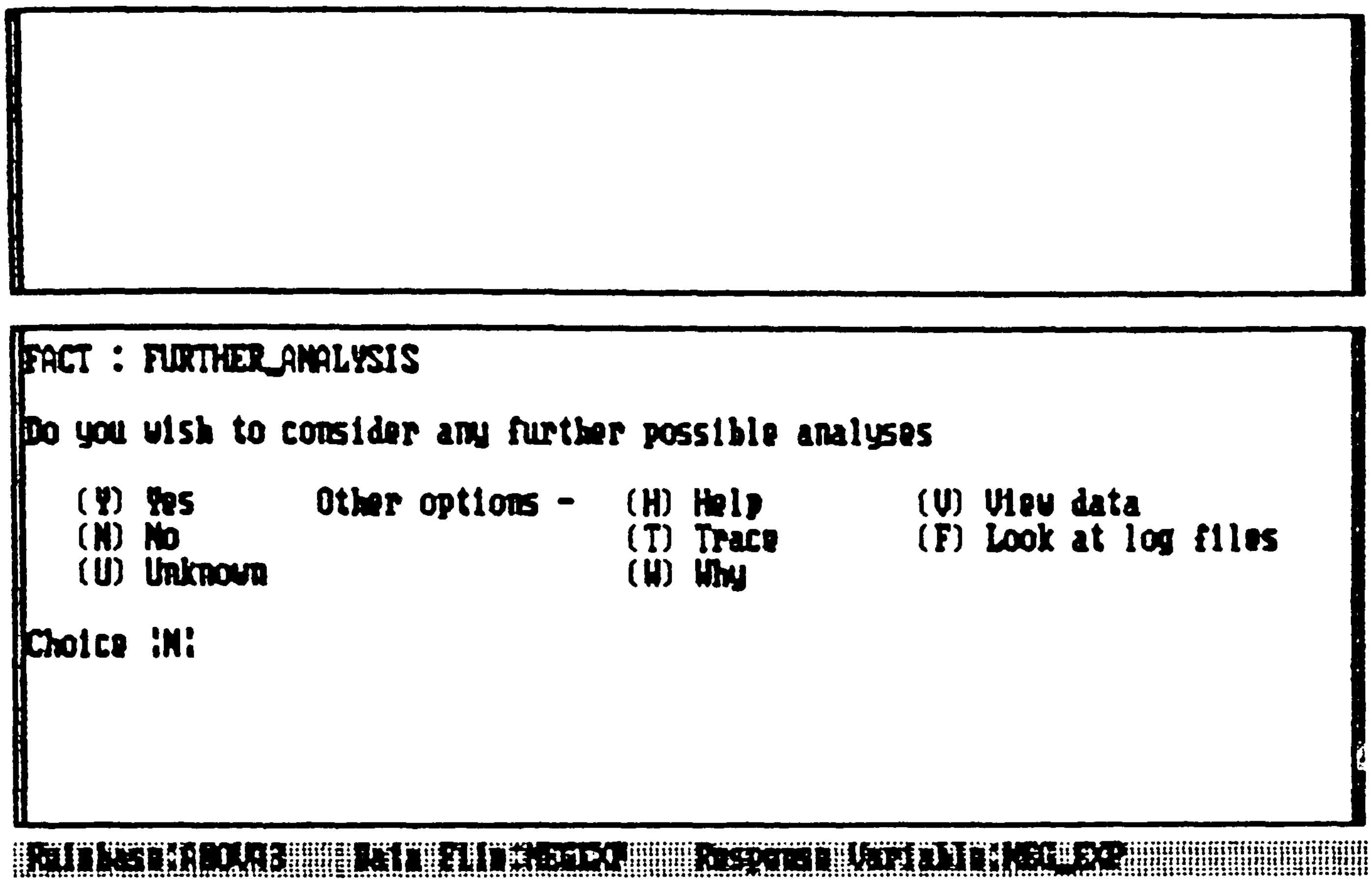
Figure 8.43 : Consultation A - 26



```
              Fitting Analysis of Variance One-Way Model

SOURCE              SS             DF          MS              F
TREATMENTS     0.04604              2     0.02302        0.01739
RESIDUAL      95.32094             72     1.32390

TOTAL         95.36698             74


              One Way Anova : Not Significant at 5%
              i.e. No difference between treatment groups

                      Press any key to continue
```

```
Rulebase:ANOVA3   Data File:NEGEXP   Response Variable:NEG_EXP
```

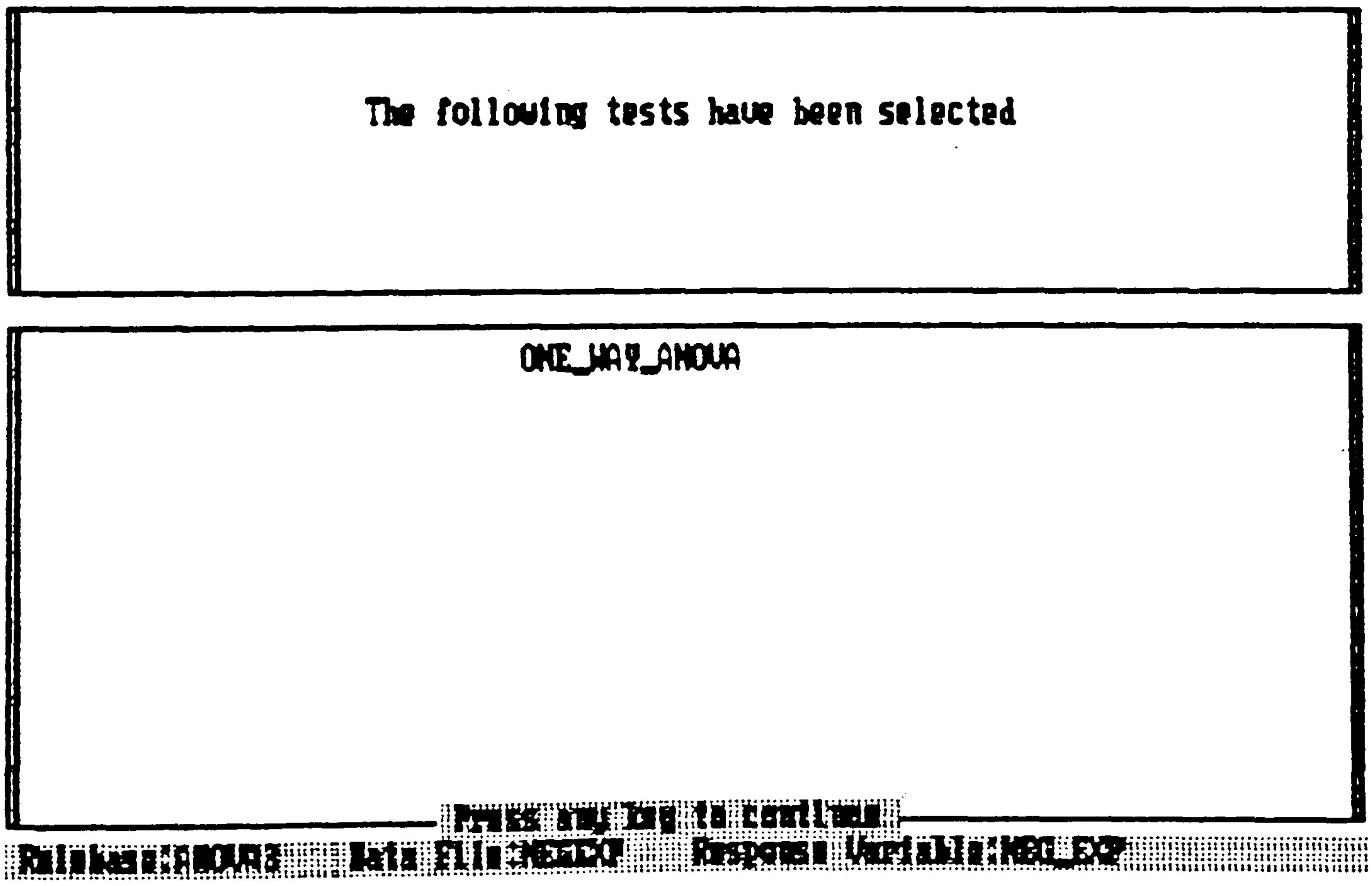
Figure 8.44 : Consultation A - 27



```
FACT : FURTHER_ANALYSIS

Do you wish to consider any further possible analyses

   (Y) Yes       Other options -  (H) Help        (V) View data
   (N) No                         (T) Trace       (F) Look at log files
   (U) Unknown                    (W) Why

Choice :N:
```

```
Rulebase:ANOVA3   Data File:MEGEXP   Response Variable:MEG_EXP
```

Figure 8.45 : Consultation A - 28

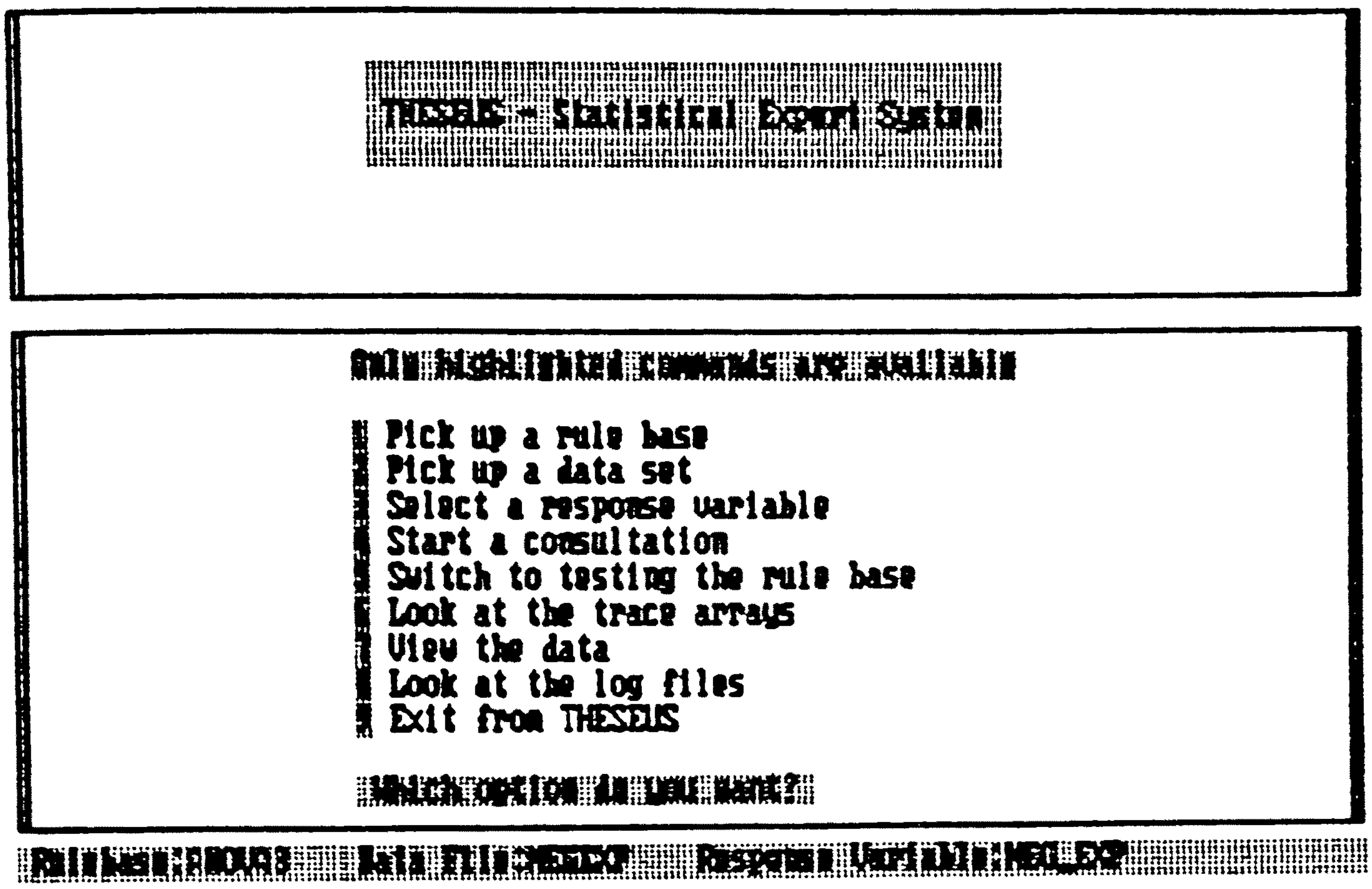


```
                    The following tests have been selected
```

```
                          ONE_WAY_ANOVA
```

```
                        Press any key to continue
Rulebase:ANOVA3   Data File:MEGEXP   Response Variable:MEG_EXP
```

Figure 8.46 : Consultation A - 29



```
THESEUS - Statistical Expert System

Only highlighted commands are available

 Pick up a rule base
 Pick up a data set
 Select a response variable
 Start a consultation
 Switch to testing the rule base
 Look at the trace arrays
 View the data
 Look at the log files
 Exit from THESEUS

Which option do you want?
Rule base:GROUP8   Data File:MEDEXP   Response Variable:MED_EXP
```

The second consultation is picked up after the Analysis of Variance has been carried out and shows the selection of appropriate multiple comparison methods. The data here is taken from page 239 of 'Biometry' (Sokal R.R. and Rohlf F.J., 1981, Freeman). Figures 8.47 to 8.59 show this consultation.

8.47 Shows the analysis of variance table for the data.

8.48 Asking the user whether they wish to consider any further analysis

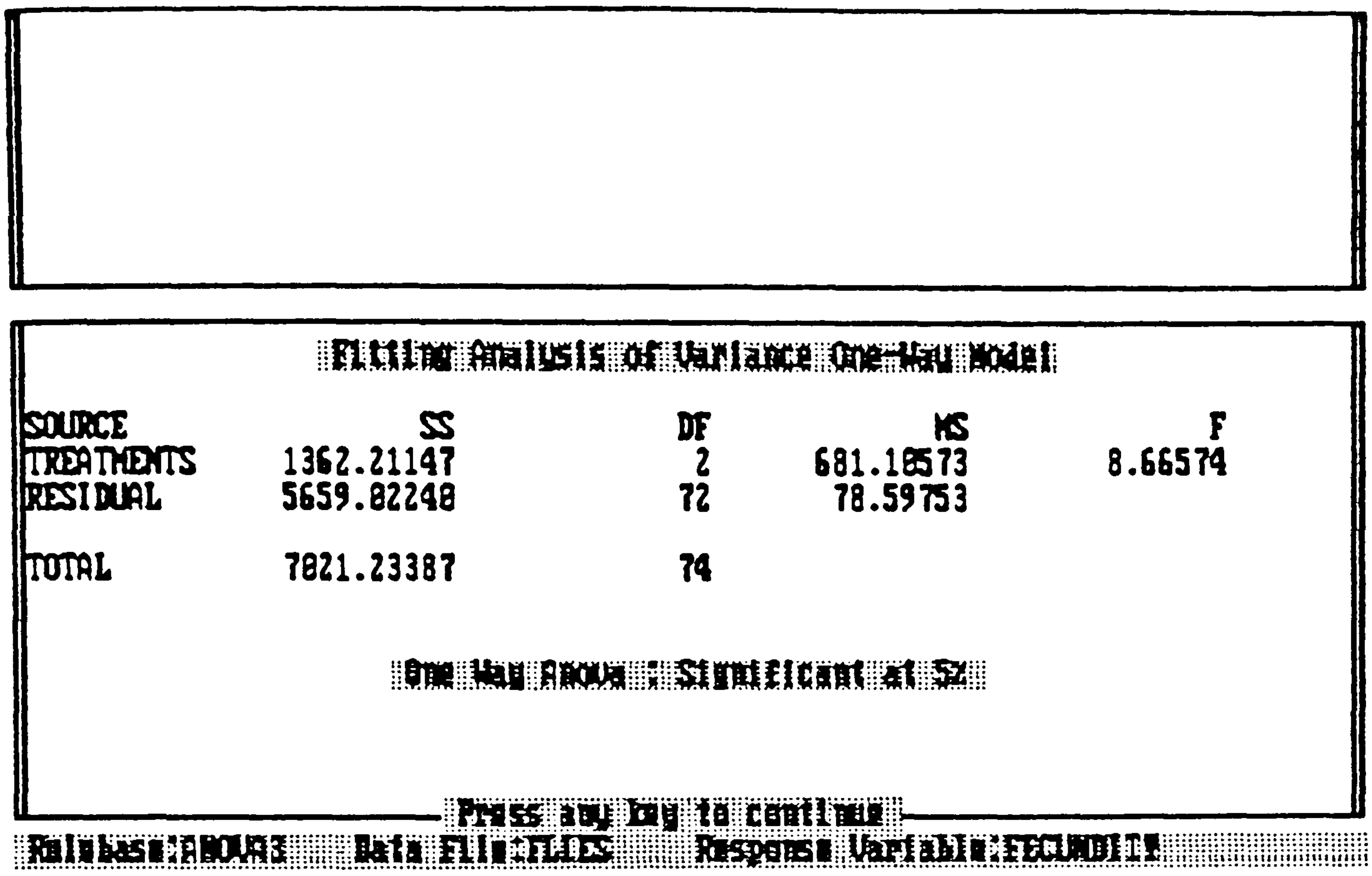
8.49 - 8.54 The system is trying to establish a list of possible methods. 8.50 shows an example of the Help facility accessed during this stage.

8.55 Having established a list of possible methods the system now tries to verify these techniques. This consultation was picked up after the Analysis of Variance had been undertaken so the system has already established that there were no outliers and the data was Normal with equal variances. Consequently the system is able to make its recommendations on the basis of information already known.

8.56 The user is asked to select the methods they wish to be used.

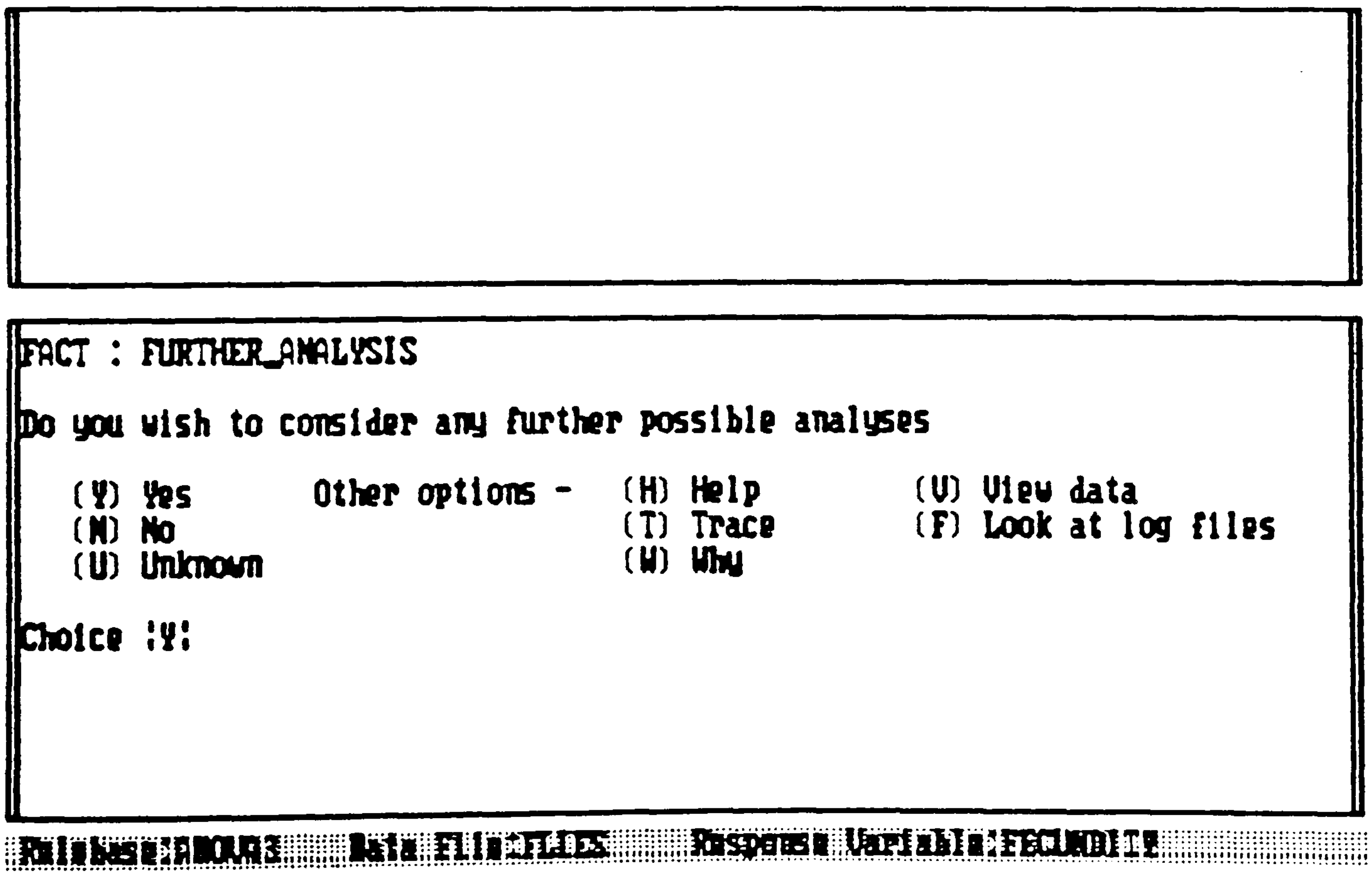
8.57 - 8.59 The system establishes that the user does not wish to consider any further analyses and informs the user which tests were selected.

Figure 8.47 : Consultation B - 1



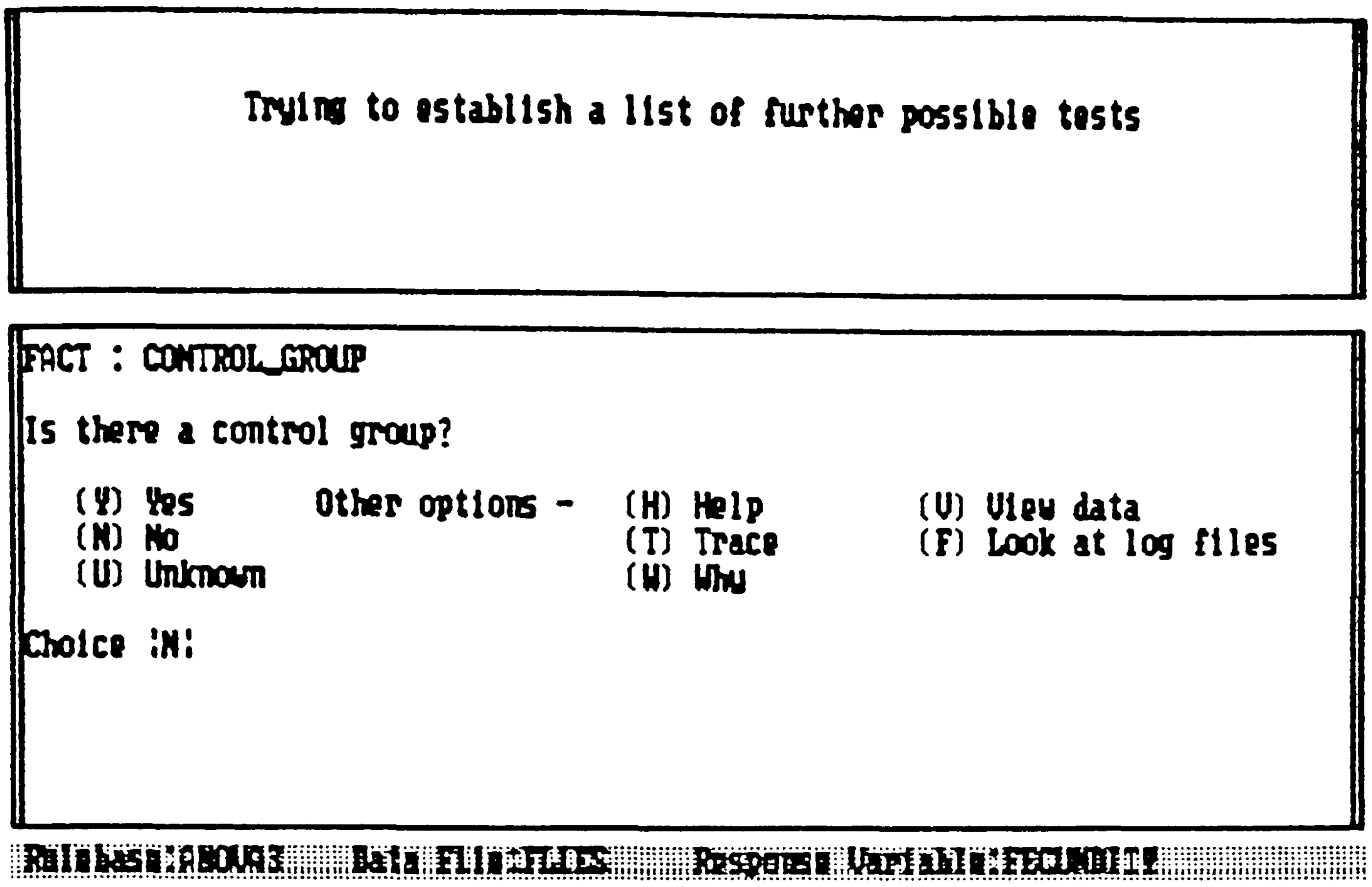
```
                Fitting Analysis of Variance One-Way Model

SOURCE              SS            DF            MS             F
TREATMENTS     1362.21147          2      681.18573        8.66574
RESIDUAL       5659.82248         72       78.59753

TOTAL          7821.23387         74


                    One Way Anova : Significant at 5%

                          Press any key to continue
Rulebase:ANOVA3    Data File:FLIES    Response Variable:FECUNDITY
```

Figure 8.48 : Consultation B - 2



```
FACT : FURTHER_ANALYSIS

Do you wish to consider any further possible analyses

   (Y) Yes        Other options -   (H) Help         (V) View data
   (N) No                           (T) Trace        (F) Look at log files
   (U) Unknown                      (W) Why

Choice :Y:

Rulebase:ANOVA3    Data File:FLIES    Response Variable:FECUNDITY
```

Figure 8.49 : Consultation B - 3

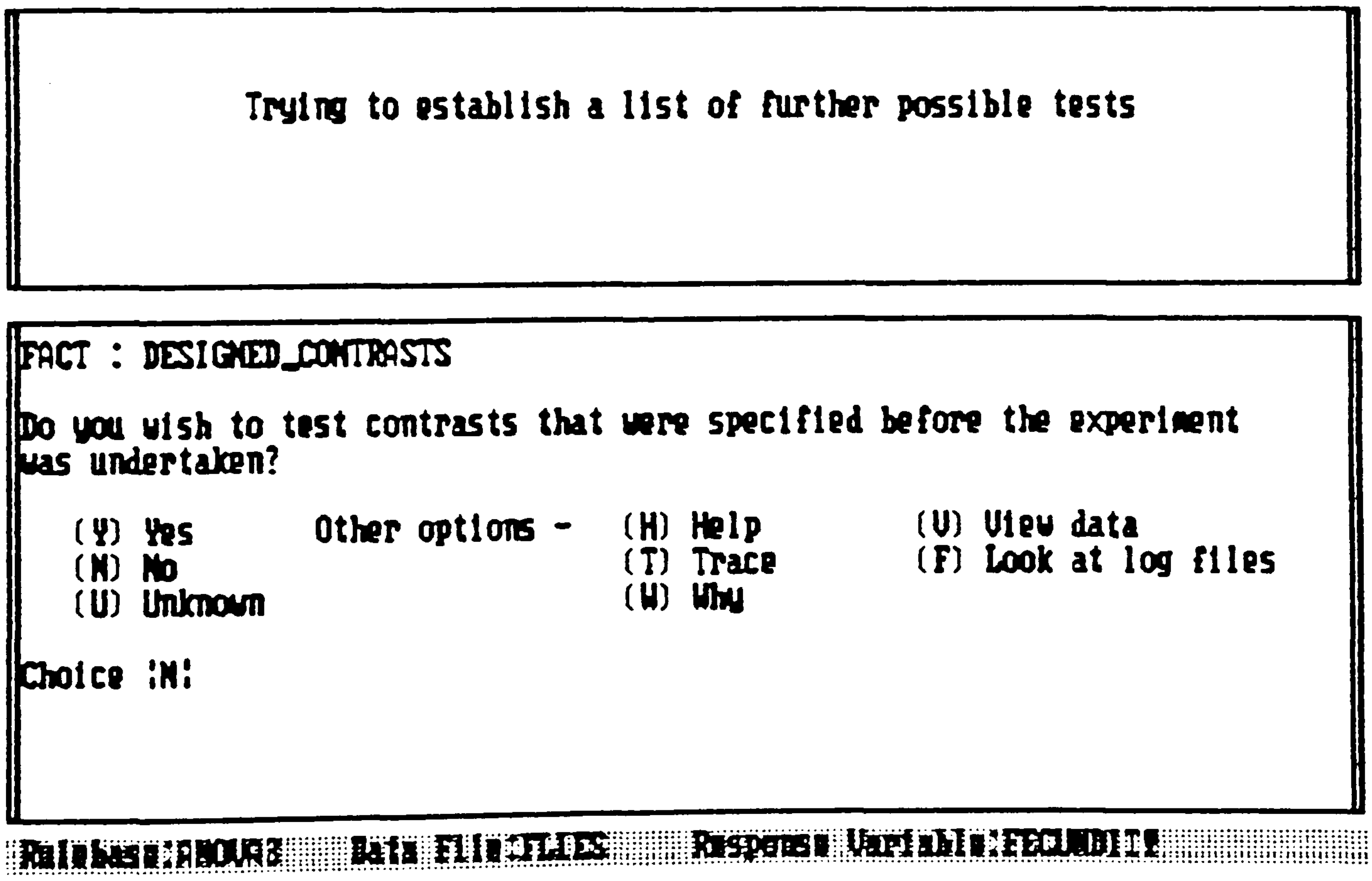


Trying to establish a list of further possible tests

```
FACT : CONTROL_GROUP

Is there a control group?

   (Y) Yes      Other options -   (H) Help       (U) View data
   (N) No                         (T) Trace      (F) Look at log files
   (U) Unknown                    (W) Why

Choice :N:
```

Rulebase:ANOVA3   Data File:FLIES   Response Variable:FECUNDITY

Figure 8.50 : Consultation B - 4

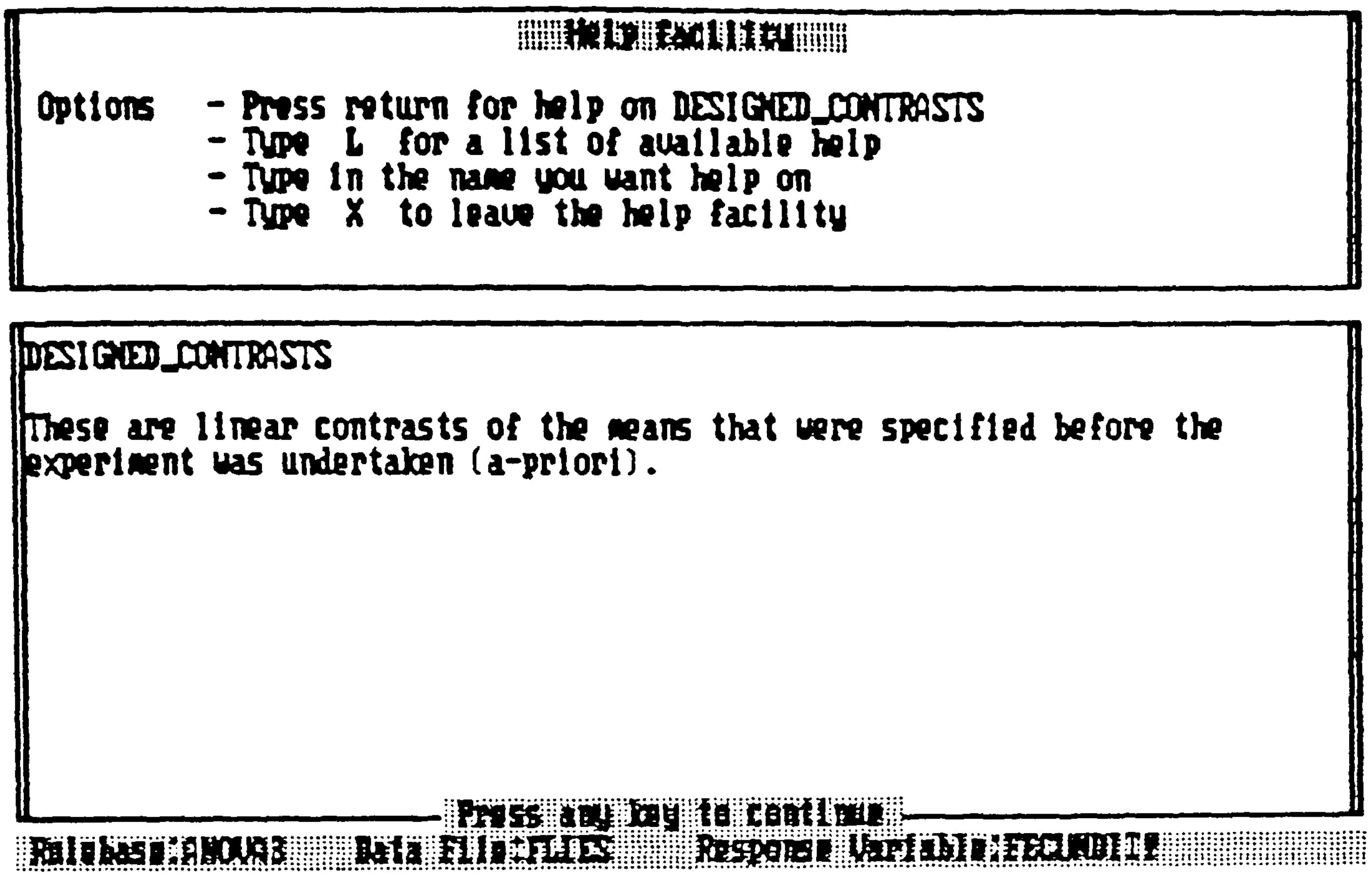


Trying to establish a list of further possible tests

```
FACT : DESIGNED_CONTRASTS

Do you wish to test contrasts that were specified before the experiment
was undertaken?

   (Y) Yes      Other options -   (H) Help       (U) View data
   (N) No                         (T) Trace      (F) Look at log files
   (U) Unknown                    (W) Why

Choice :N:
```

Rulebase:ANOVA3   Data File:FLIES   Response Variable:FECUNDITY

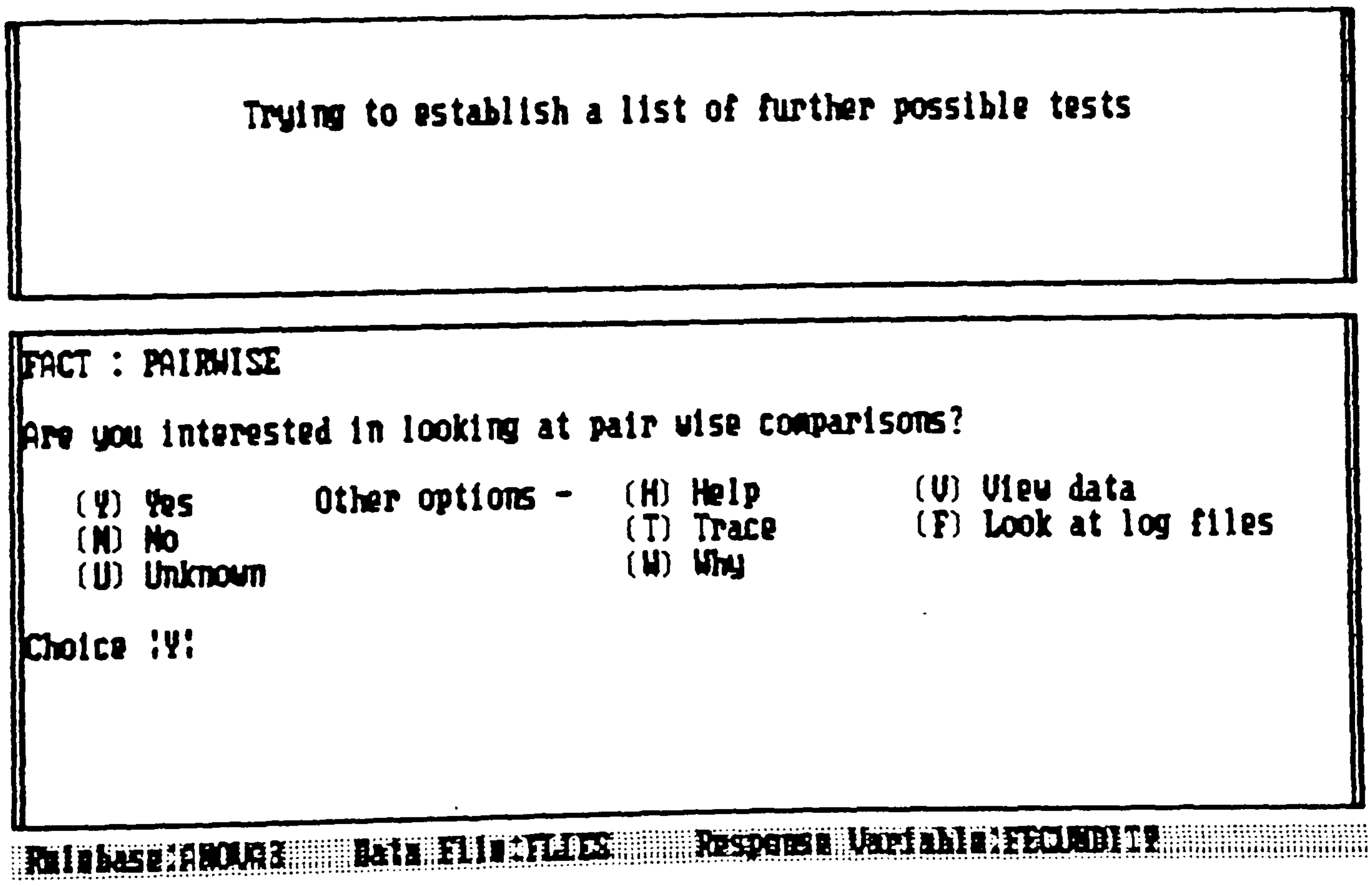
Figure 8.51 : Consultation B - 5



```
                          Help facility

Options   - Press return for help on DESIGNED_CONTRASTS
          - Type  L  for a list of available help
          - Type in the name you want help on
          - Type  X  to leave the help facility
```

```
DESIGNED_CONTRASTS

These are linear contrasts of the means that were specified before the
experiment was undertaken (a-priori).

                       Press any key to continue
Rulebase:ANOVA3   Data File:FLIES   Response Variable:FECUNDITY
```

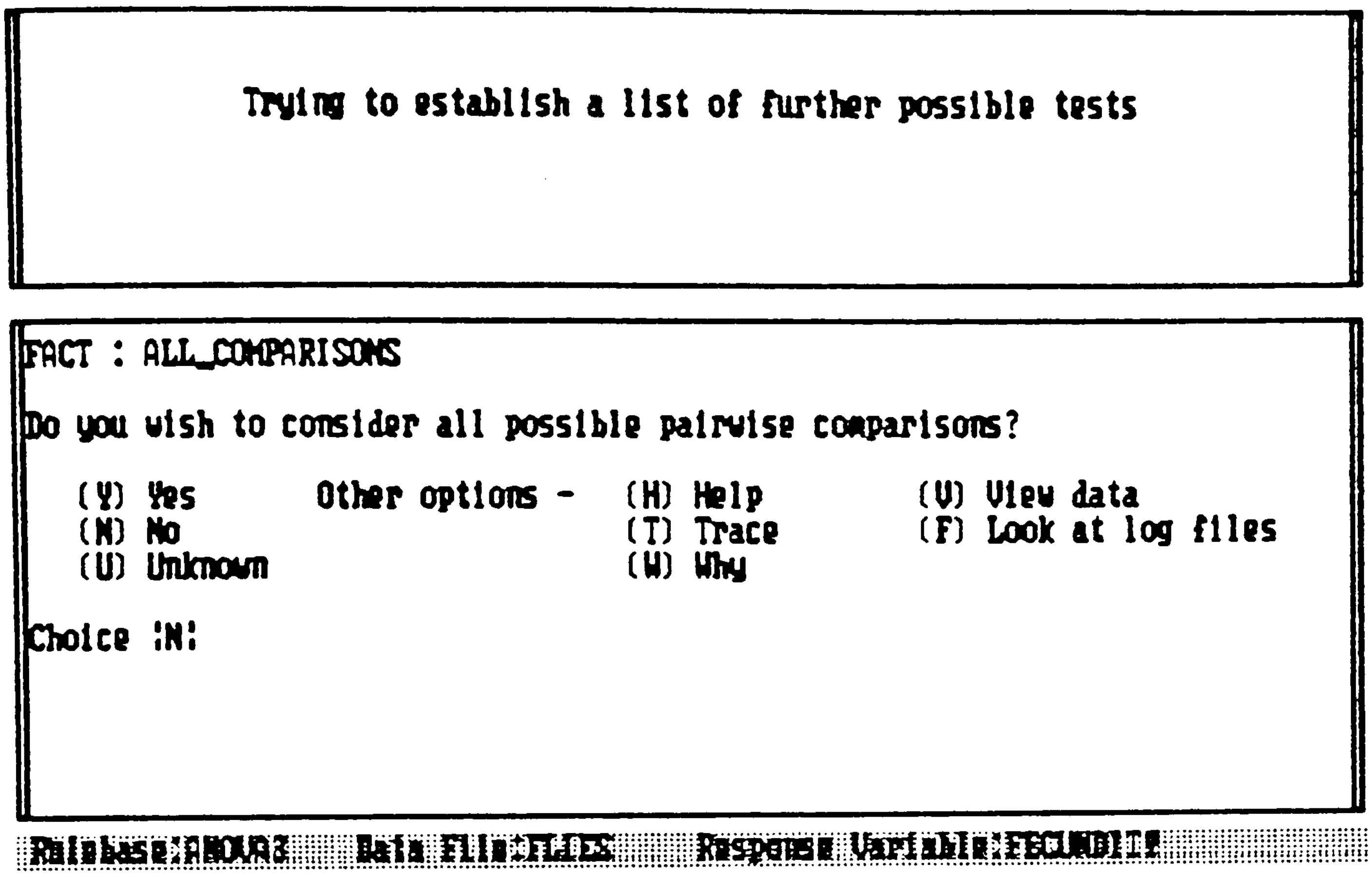
Figure 8.52 : Consultation B - 6



```
          Trying to establish a list of further possible tests
```

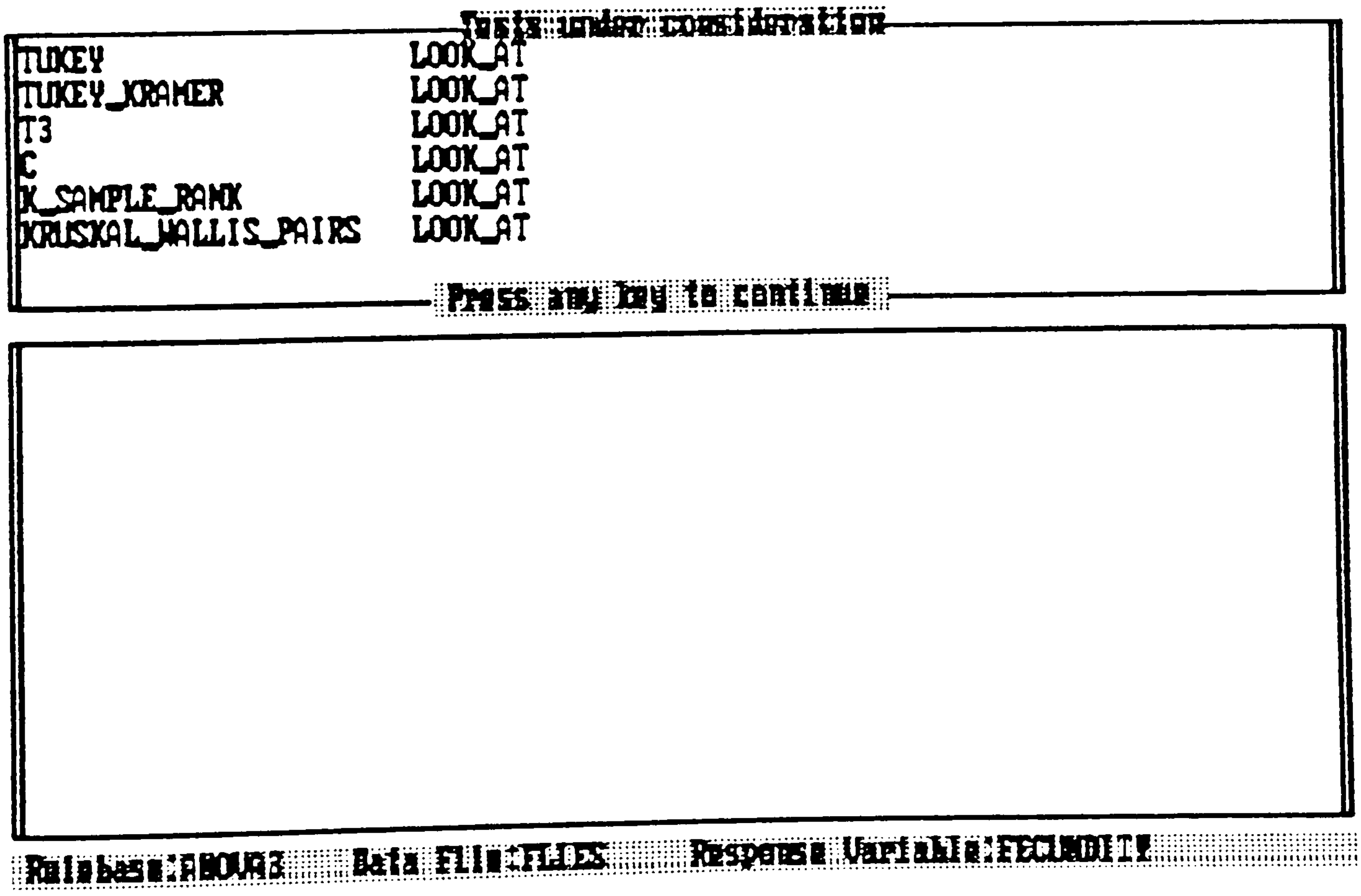
```
FACT : PAIRWISE

Are you interested in looking at pair wise comparisons?

   (Y) Yes       Other options -  (H) Help        (V) View data
   (N) No                         (T) Trace       (F) Look at log files
   (U) Unknown                    (W) Why

Choice :Y:

Rulebase:ANOVA3   Data File:FLIES   Response Variable:FECUNDITY
```

Figure 8.53 : Consultation B - 7



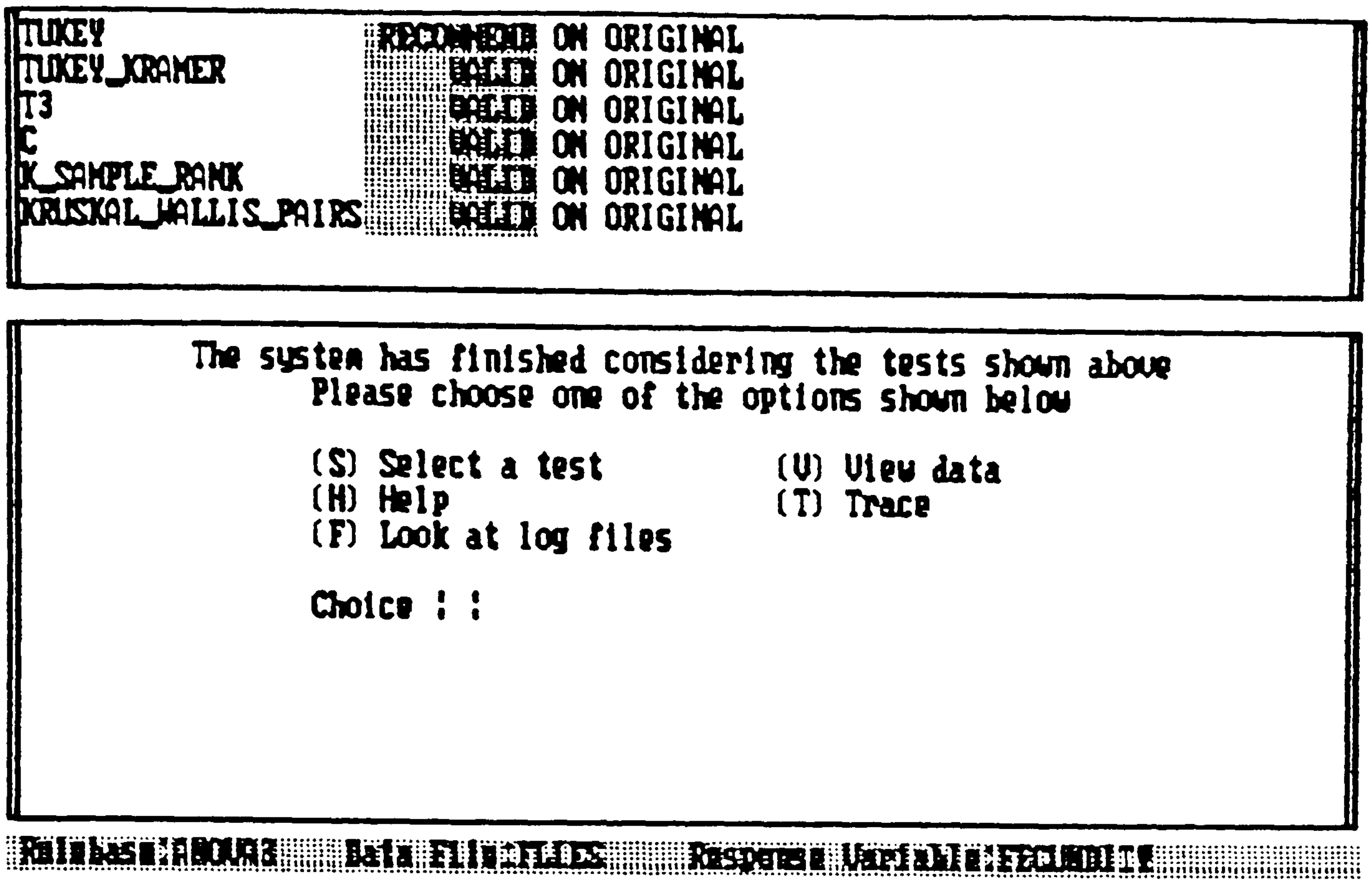
```
Trying to establish a list of further possible tests

FACT : ALL_COMPARISONS

Do you wish to consider all possible pairwise comparisons?

  (Y) Yes       Other options -   (H) Help        (V) View data
  (N) No                          (T) Trace       (F) Look at log files
  (U) Unknown                     (W) Why

Choice :N:

Rulebase:ANOVA3   Data File:FLIES   Response Variable:FECUNDITY
```

Figure 8.54 : Consultation B - 8



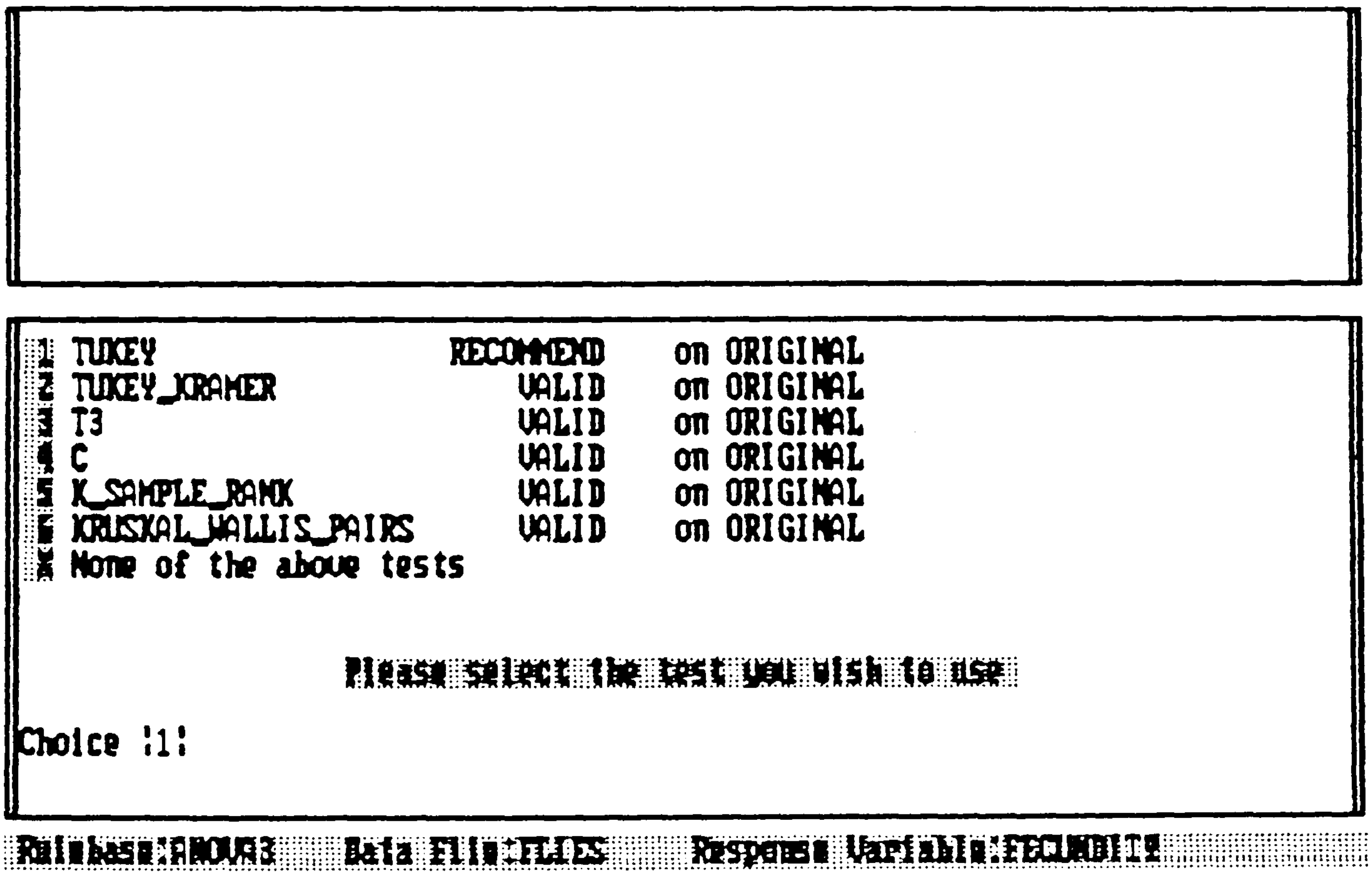
```
                                   Tests under consideration
TUKEY                    LOOK_AT
TUKEY_KRAMER             LOOK_AT
T3                       LOOK_AT
C                        LOOK_AT
K_SAMPLE_RANK            LOOK_AT
KRUSKAL_WALLIS_PAIRS     LOOK_AT
                                   Press any key to continue

Rulebase:ANOVA3   Data File:FLIES   Response Variable:FECUNDITY
```

Figure 8.55 : Consultation B - 9



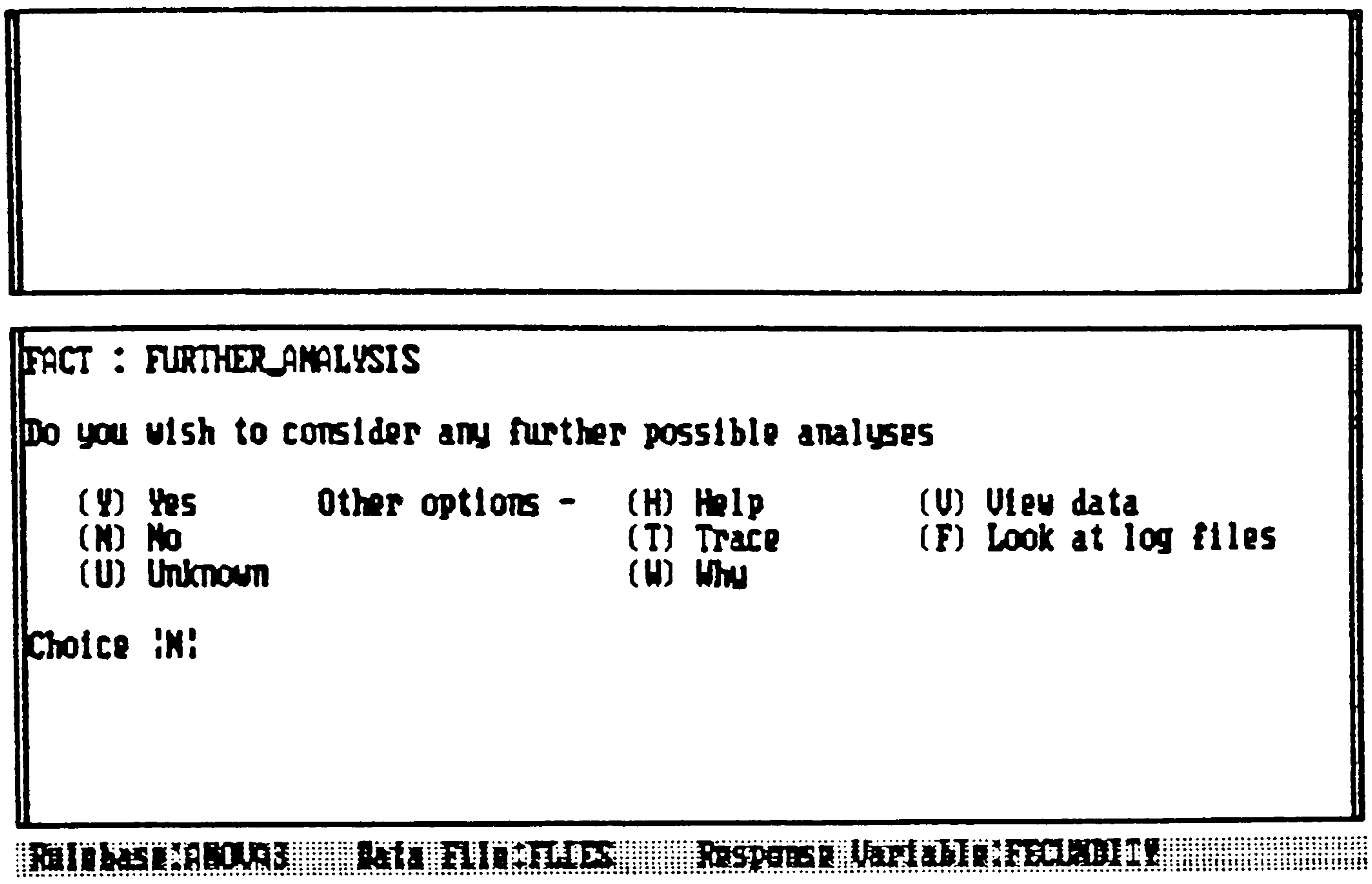
```
TUKEY                     RECOMMEND ON ORIGINAL
TUKEY_KRAMER                  VALID ON ORIGINAL
T3                            VALID ON ORIGINAL
C                             VALID ON ORIGINAL
K_SAMPLE_RANK                 VALID ON ORIGINAL
KRUSKAL_WALLIS_PAIRS          VALID ON ORIGINAL

          The system has finished considering the tests shown above
               Please choose one of the options shown below

               (S) Select a test            (V) View data
               (H) Help                     (T) Trace
               (F) Look at log files

               Choice : :

Rulebase:ANOVA3   Data File:FLIES   Response Variable:FECUNDITY
```

Figure 8.56 : Consultation B - 10



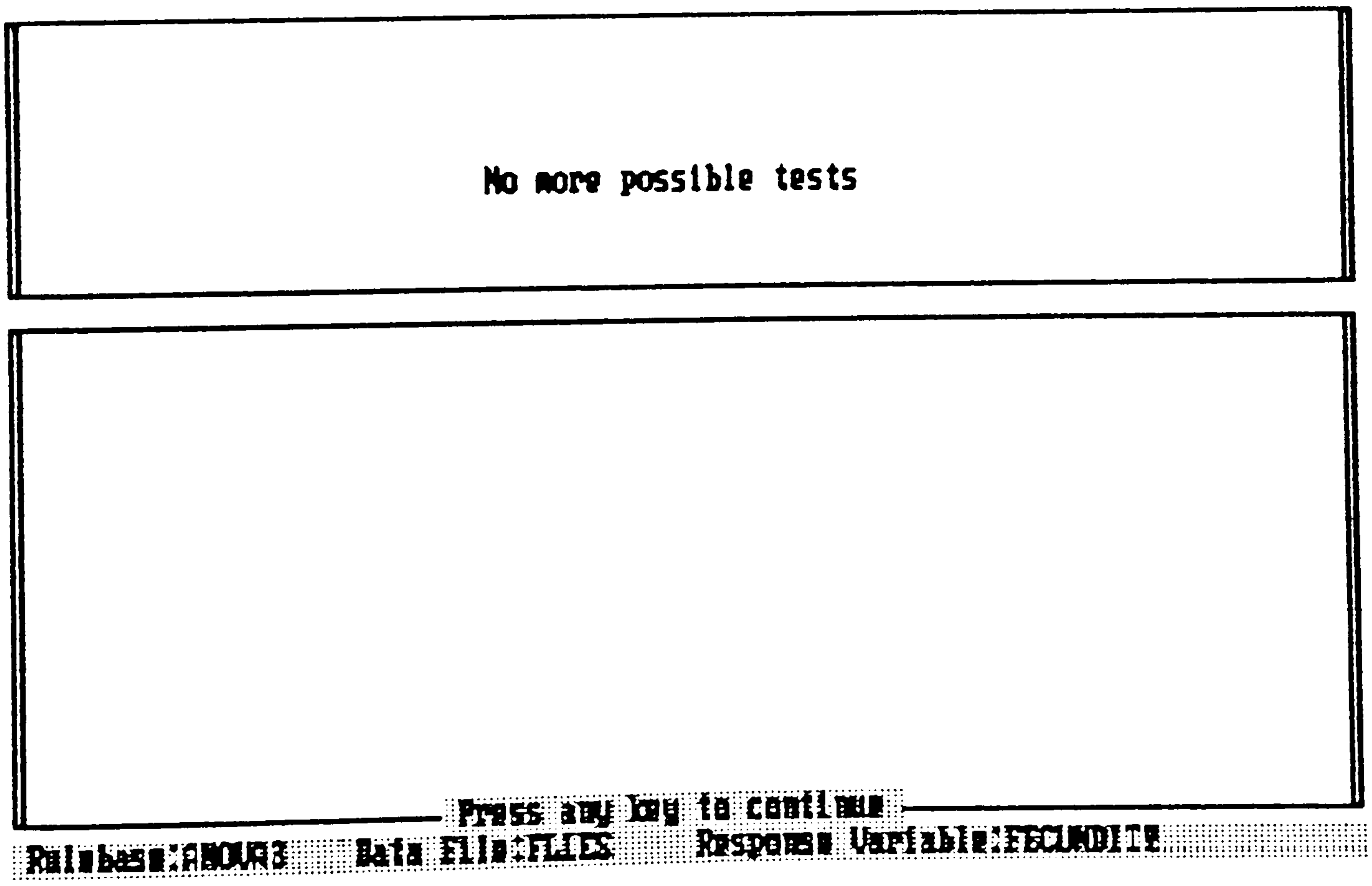
```
1 TUKEY                   RECOMMEND    on ORIGINAL
2 TUKEY_KRAMER                VALID    on ORIGINAL
3 T3                          VALID    on ORIGINAL
4 C                           VALID    on ORIGINAL
5 K_SAMPLE_RANK               VALID    on ORIGINAL
6 KRUSKAL_WALLIS_PAIRS        VALID    on ORIGINAL
X None of the above tests

              Please select the test you wish to use

Choice :1:

Rulebase:ANOVA3   Data File:FLIES   Response Variable:FECUNDITY
```

Figure 8.57 : Consultation B - 11



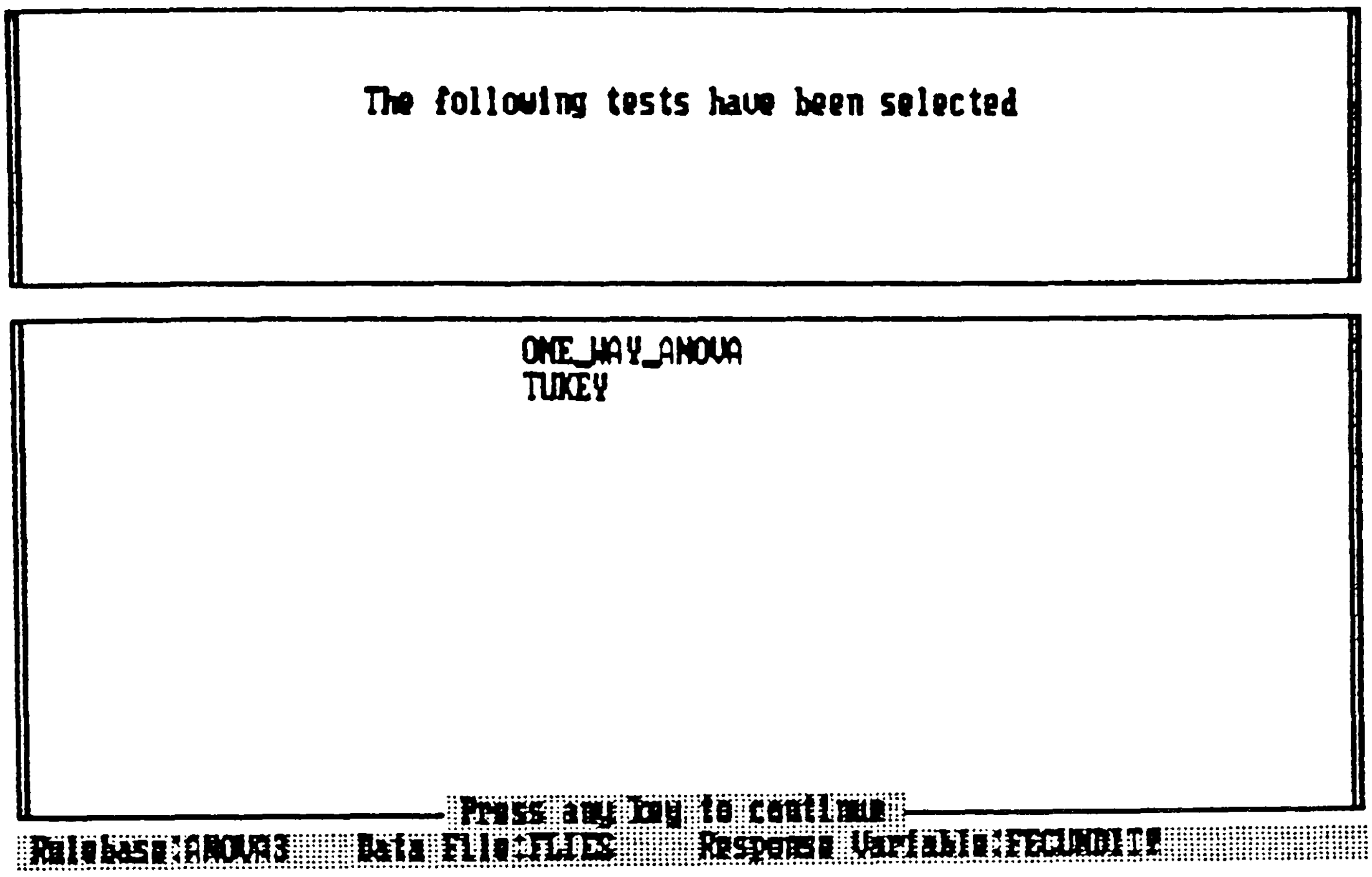
```
FACT : FURTHER_ANALYSIS

Do you wish to consider any further possible analyses

   (Y) Yes       Other options -   (H) Help         (V) View data
   (N) No                          (T) Trace        (F) Look at log files
   (U) Unknown                     (W) Why

Choice :N:

Rulebase:ANOVA3   Data File:FLIES   Response Variable:FECUNDITY
```

Figure 8.58 : Consultation B - 12



```
                        No more possible tests

                        Press any key to continue
Rulebase:ANOVA3   Data File:FLIES   Response Variable:FECUNDITY
```

Figure 8.59 : Consultation B - 13



The following tests have been selected

ONE_WAY_ANOVA
TUKEY

Press any key to continue

Rulebase:ANOVA3 Data File:FLIES Response Variable:FECUNDITY

Chapter Nine

Evaluation of the Prototype

9.1 Introduction

The prototype system was demonstrated to members of the Statistics Research Group while the knowledge acquisition was in progress and the software was being developed. This enabled frequent feedback and modifications where appropriate, thus aiding the development of the system in terms of the user-interface as well as the knowledge acquisition. An advantage of undertaking the knowledge acquisition and software development in 'parallel' was that inadequacies in the rule-base soon showed up inadequacies in the software.

Once the prototype system had been developed to the extent that a rational rule-base was available and the software was sufficiently complete to demonstrate the potential for a Statistical Expert System then small scale evaluation trials were instigated.

The overall aim of these evaluation trials was to assess the advantages and problems of implementing a Statistical Expert System. By undertaking preliminary trials at this stage it was envisaged that any major difficulties with the software or structure of the system could be identified and corrected. When the results of the trials have been assessed the system can be modified and extended accordingly before being sent out for evaluation on a larger scale.

9.2 Format of Trial

The prototype software was initially sent to several industrial sites, based mostly in the pharmaceutical industry, and two University departments. Users were asked to assess the system with respect to

1. The general structure and pattern of the consultation
2. The default rule-base
3. Modification of the rule-base

4. The potential for use by research workers in routine data analysis tasks

The trial was divided into three sections :-

i) initial system assessment
ii) assessment of the default rule-base
iii) modifying the rule-base

Some modifications to prototype software were made, after some of the questionnaires had been returned, as a result of comments made by the respondents. The updated system has been assessed by two statisticians, one each from academia and industry. The alterations were to allow the user to answer 'unknown' in response to a question, to ask the system 'Why?' and to provide access to the facilities when selecting a statistical method.

9.3 Response to Questionnaire

The responses to the questionnaires are summarised in Table IV, the additional comments are detailed in Table V.

9.3.1 User Interface

Respondents were generally satisfied with the main menu, the only additional suggestion given was to include an option to allow the user to edit the data.

There was a consensus of opinion that the ask text, used for eliciting a 'Yes' or 'No' response from the user, were readable but would be better if it supplied more information. Some respondents would have liked to have more facilities available at this stage, or at least improvement of the existing facilities. Only one respondent thought it unnecessary to permit the user to answer 'Unknown' to a question maintaining that the Help text available should supply sufficient information to allow the user to answer 'Yes' or 'No'.

Table IV : Response to Questionnairre

Respondent	1	2	3	4	5	6
(A)cademic (I)ndustrial Statistician	I	I	I	A	I	A
<u>Section A : User Interface</u>						
Main Menu						
A1 Is it easy to understand	5	4	4	4	5	4
A2 Do you find it easy to use	5	4	4	4	5	5
A3 Is it flexible enough	5	3	2	3	5	4
Question Screens						
A4 Enough information	4	4	3	3	3	4
A5 Is the text readable	3	4	3	4	5	4
A6 Are the facilities useful	4	3	4	4	2	3
A7 Are there enough facilities	3	4	4	5	N	5
A8 Answer unknown	Y	Y	Y	Y	Y	Y
Selecting a Technique						
A9 Is the information clear enough	3	4	4	3	N	3
A10 Facilities available here	Y	Y	Y	Y	Y	Y
General						
A11 Is it easy to follow the system	3	3	4	3	2	4
<u>Section B : Facilities</u>						
Help						
B1 Is it easy to use	5	2/3	2	4	5	5
B2 Is the text understandable	4	4	3	3	5	5
B3 Is this facility useful	5	Y	4	3	5	5
B4 Is enough information given	4	3	3	3	N	2
B5 Is it versatile enough	3	2	1	3	-	2
Trace						
B6 Is it easy to use	5	4	3	3	5	5
B7 Is it understandable	3	2	3	2	1	3
B8 Is it useful	5	3	3	3	depends	3
Viewdata						
B9 Is it easy to use	5	-	4	4	5	5
B10 Is it understandable	5	4	4	3	5	5
B11 Is it useful	5	-	3	3	5	4
B12 Is it powerful enough	2	4	2	2	1	3
Log Files						
B13 Is it easy to use	5	3	3	4	3	5
B14 Is it understandable	2	3	3	2	3	3
B15 Is it useful	5	4	4	3	3	3
<u>Section C : Documentation</u>						
C1 Documentation presented helpfully	5	4	3	4	4	4
C2 Information presented helpfully	4	3	3	3	4	4
C3 Is enough information given	4	4	4	3	toomuch	4
C4 Is it easy to read	4	4	3	5	3	4

Table V : Additional Comments from Questionnairres

Respondent 2

- creating the data file is difficult
- appears potentially useful, some omissions in the 'help' and 'view data' sections. We assume these are still to be completed

Respondent 3

- Non-statisticians may require more detailed questions, perhaps an option of brief or detailed questions would be useful
- Help facility should be enough for the 'dont knows' to make a yes or no decision
- Typing in the help required takes too long
- The help facility is confusing with so many similar sounding items on the list. Maybe ID numbers should be entered or a way of moving the cursor to the help item required

Respondent 4

- Include the median in the descriptive statistics of the view data facility
- Normal Plots are difficult to follow, more labelling ?

Respondent 6

- Viewdata : Plots fine but data description could be fuller e.g. maximum, minimum, ranges. Raw data.
- Trace : Should be able to jump out of a long one (currently the user must view the whole trace). Trace only really useful to someone altering the rule-base
- Help : Could refer to well known texts. Some ok but some do not give enough information
- Why : Only useful from expert systems point of view, would be better if it provided a statistical explanation of what is happening
- Transformations : needs tightening up, no facility for going through all transformations and then selecting the one you want. Would also be helpful to be able to look for outliers again after a transformation

When the system gets to the stage of asking the user to select a technique, all the respondents indicated that they would like the facilities to be available at this stage. The way the information is presented here was not regarded as sufficiently clear by some of the respondents.

The majority of the respondents did not find it particularly easy to follow the pattern of the consultation, the most common response was 'sufficient'. As the users during this trial were all statisticians this pin-points a weakness in the system, if a statistician finds it difficult to follow the systems pattern of consultation a non-statistician will probably have even greater difficulty.

9.3.2 Facilities

Opinion on the ease of use of the Help facility was divided with some respondents quite happy with the facility as it is. Other respondents were concerned because of missing help or difficulties in using the facility; typing in the name and the occurrence of many similar sounding names increased the difficulty of using the system. One of the respondents suggested using identification numbers or allowing the user to move the cursor to the name on which help is required; this would improve the versatility of the help facility. The help text was generally considered to be readable but could be extended or improved in some cases.

The trace arrays, although easy to use and potentially beneficial, were not considered easy to understand. It may be that too much information at a detailed level is supplied or that the information is not presented clearly enough. Whatever the case this facility needs careful thinking through. The facility for looking at the log files seems to have elicited a similar response.

The view data facility received the most enthusiastic response, the only disappointment being that it was not extensive enough.

One respondent thought that the way in which transformations were handled was too rigid and should be extended to allow the user greater flexibility in trying a number of transformations and then being able to select the most appropriate.

9.4 Local Tuning of the Rule-base

Only one statistician was able to try tuning the rule-base and the following comments are based on an interview with this statistician.

When the prototype system was being developed it was difficult to assess how much a local expert would need to know about the structure of the knowledge and the method of inference used in order to successfully modify the rule-base. For the evaluation trials the local expert was supplied with a brief description of the knowledge representation used and the method of inference. The description about the inference was limited to an explanation of the flow of control, as given in section 4.5 and a brief description of the different types of rule.

The local expert found the rule-base editor relatively easy to use and helpful in checking the syntax of the rules, leaving him free to concentrate on the problem of encoding the knowledge that he wished to add to the rule-base. The local expert was trying to include a further two multiple comparison methods for testing for trend where the data groups are levels of treatment. This involved adding new rules, tests and facts and also involved altering the attributes of an existing fact from being one set by rules to one set by the user.

There seemed to two main difficulties encountered by the local expert. The first was that he hadn't included a forward chaining rule which would enable the system to consider the two new techniques. The second difficulty was one of being able to see easily the consequences of altering the attributes of a given fact. The fact was handled correctly in the rules added by the local expert but it was not immediately apparent what other rules were affected.

The local expert did find it difficult to 'debug' the rules that were put in, primarily because the trace arrays were not that easy to follow. This was remedied in part by talking through the relevant traces. The local expert was shown the proposed graphical representation, see section 9.5, of the goal trace and found that far easier to understand.

9.5 Recommended Improvements to the Prototype System

As a result of the response to the evaluation trials a number of potential improvements are considered here, the majority related to making the facilities available more versatile and understandable.

The WHY facility currently provides explanation in expert system terms, explaining which rule is being tried and what current goals the system is trying to establish. This facility would be more useful to the non-statistician if it provided a statistical explanation about what the system is trying to do.

The HELP facility could be extended to include different levels of Help ranging from text which would serve as a reminder for the user to more extensive Help for the statistically naive user, perhaps with reference to well known texts. The means of accessing the required help could be improved by allowing selection by moving the cursor. If there was plenty of memory available, a hypertext system would probably provide the most versatile and comprehensible help facility. A hypertext system would provide screens of appropriate help which are linked by keywords highlighted on the screen, a user can access further help by selecting these keywords.

The facility to look at the data is currently quite limited and there are a wide range of potential improvements. For example the user may wish to be able to view the raw data or other summary statistics such as the median or quartiles. The graphics capabilities are currently limited by the amount of memory available but could be extended to incorporate other plots of the data such as histograms and plots of the means and variances.

A certain amount of data handling should be available within the rule-base processor to enable the user to exclude some observations from subsequent analyses or to undertake local edits; this will help in the handling of outliers. The users should also be able to aggregate groups within the data set if they so wish.

The data entry module needs to be fully developed so that it provides interactive data editing as well as an initial dialogue with the user. This initial dialogue is very important as it should be able to assess whether the users data comes within the scope of the expert system.

The trace facilities are currently not particularly easy to follow, a graphical method of representing the progress through the system would probably be more helpful. One possible graphical method is shown in Figures 9.1 to 9.9. Figure 9.1 shows a goal trace from the current system, Figures 9.2 to 9.9 show graphical representation of the trace at each of the 'snap-shot' positions marked on Figure 9.1.

Figure 9.1 Goal Trace Showing Where Snapshots for Proposed Trace have been Taken

```
                                       Rule  Goal       Data      Snap-
Goal                                   Tried Status     Set       Shot
--------------------------------------------------------------------------
-ONE_WAY_ANOVA...................      R76        -     ORIGINAL
--ACCEPT_PARAMETRIC..............      R22        -     ORIGINAL
---OUTLIERS......................      R30        -     ORIGINAL  - 1
---OUTLIERS......................      R31        -     ORIGINAL
---OUTLIERS......................      R32        -     ORIGINAL
---OUTLIERS......................      R33        -     ORIGINAL
---OUTLIERS......................      R34        -     ORIGINAL
---OUTLIERS......................      R35    FALSE     ORIGINAL
---NORMAL_DATA...................      R36        -     ORIGINAL  - 2
---NORMAL_DATA...................      R37        -     ORIGINAL
---NORMAL_DATA...................      R48        -     ORIGINAL
---NORMAL_DATA...................      R49        -     ORIGINAL
---NORMAL_DATA...................      R50        -     ORIGINAL
---NORMAL_DATA...................      R51        -     ORIGINAL
---NORMAL_DATA...................      R52        -     ORIGINAL
----SHAPIRO_WILK_SIG5............      R38        -     ORIGINAL  - 3
----SHAPIRO_WILK_SIG5............      R39        -     ORIGINAL
----SHAPIRO_WILK_SIG5............      R40        -     ORIGINAL
----SHAPIRO_WILK_SIG5............      R41        -     ORIGINAL
----SHAPIRO_WILK_SIG5............      R42        -     ORIGINAL
----SHAPIRO_WILK_SIG5............      R43        -     ORIGINAL
----SHAPIRO_WILK_SIG5............      R44        -     ORIGINAL
----SHAPIRO_WILK_SIG5............      R45        -     ORIGINAL
----SHAPIRO_WILK_SIG5............      R46        -     ORIGINAL
----SHAPIRO_WILK_SIG5............      R47    FALSE     ORIGINAL  - 4
---NORMAL_DATA...................      R53        -     ORIGINAL
---NORMAL_DATA...................      R54     TRUE     ORIGINAL
---VARIANCES_EQUAL...............      R55        -     ORIGINAL  - 5
---VARIANCES_EQUAL...............      R56        -     ORIGINAL
---VARIANCES_EQUAL...............      R57        -     ORIGINAL
---VARIANCES_EQUAL...............      R58        -     ORIGINAL
---VARIANCES_EQUAL...............      R59        -     ORIGINAL
---VARIANCES_EQUAL...............      R60        -     ORIGINAL
---VARIANCES_EQUAL...............      R61     TRUE     ORIGINAL  - 6
--ACCEPT_PARAMETRIC..............      R23        -     ORIGINAL
--ACCEPT_PARAMETRIC..............      R24        -     ORIGINAL
--ACCEPT_PARAMETRIC..............      R25        -     ORIGINAL
--ACCEPT_PARAMETRIC..............      R26     TRUE     ORIGINAL  - 7
-ONE_WAY_ANOVA...................      R76 RECOMMEND    ORIGINAL  - 8
KRUSKAL_WALLIS_ANOVA.............      R76    VALID     ORIGINAL
--------------------------------------------------------------------------
```

Figure 9.2 : Snapshot 1

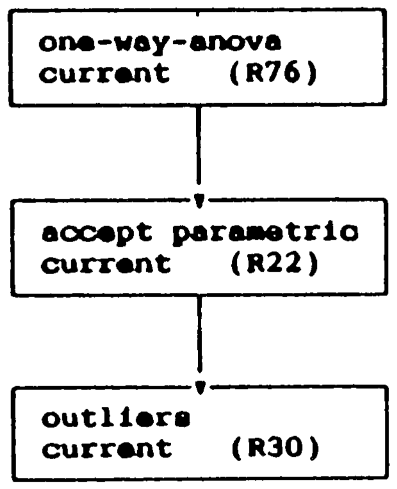


Figure 9.3 : Snapshot 2

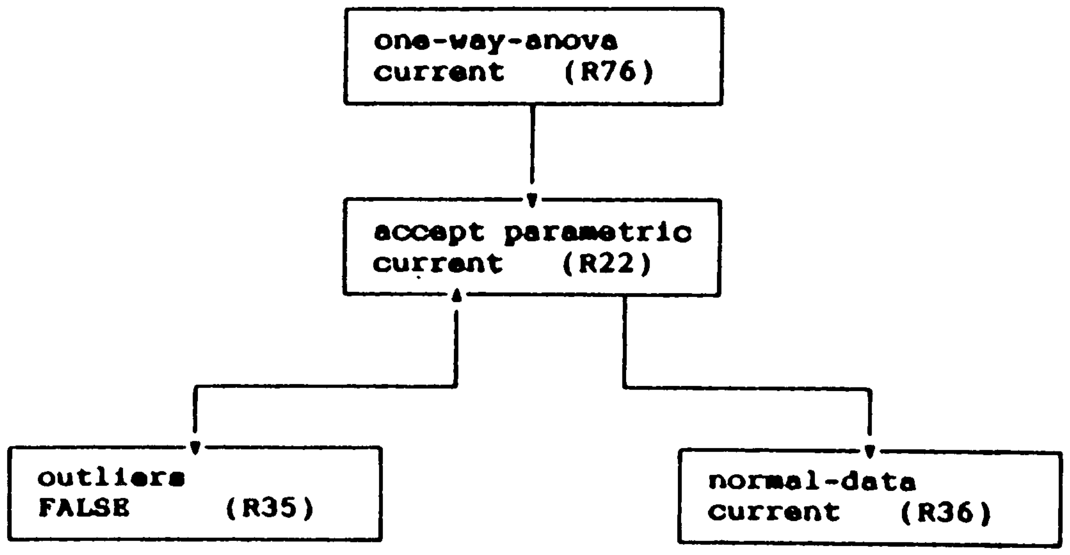


Figure 9.4 : Snapshot 3

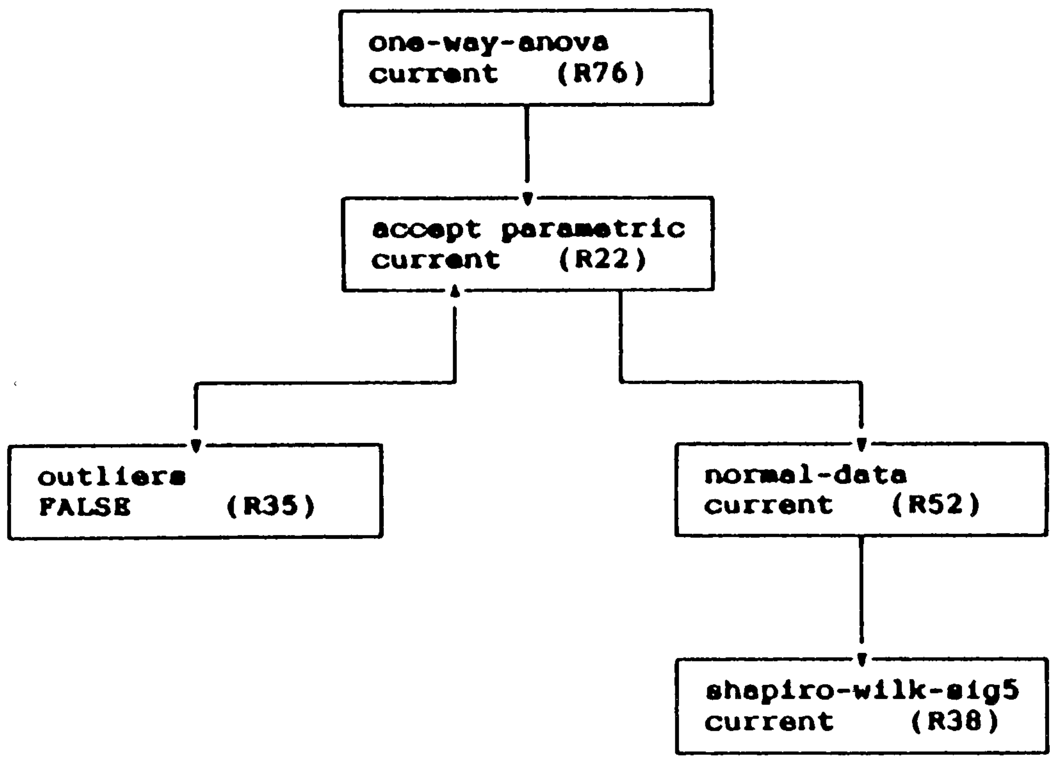


Figure 9.5 : Snapshot 4

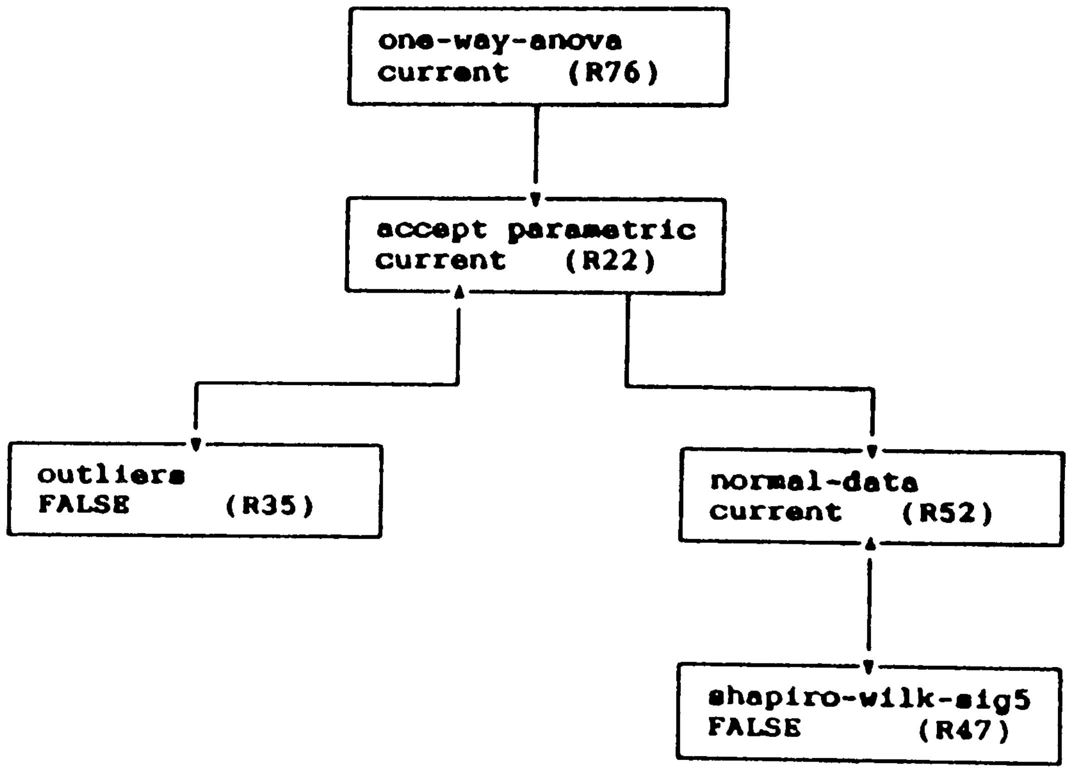


Figure 9.6 : Snapshot 5

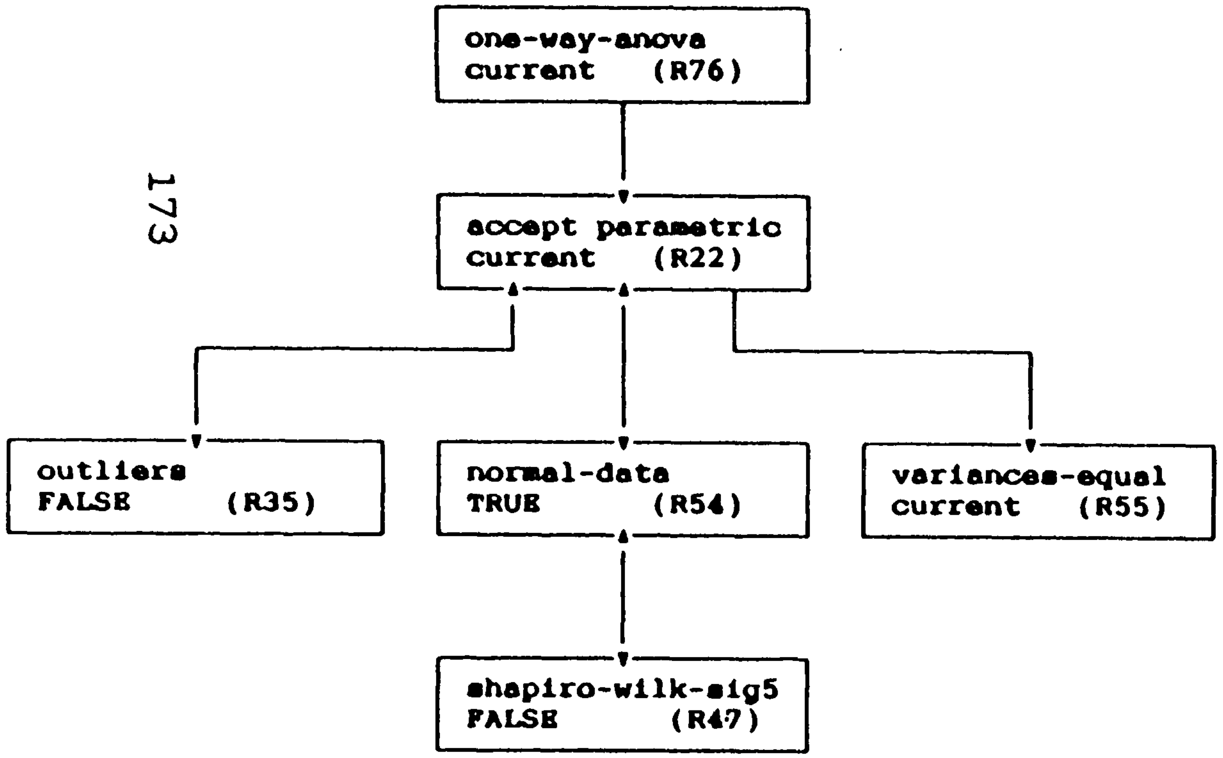


Figure 9.7 : Snapshot 6

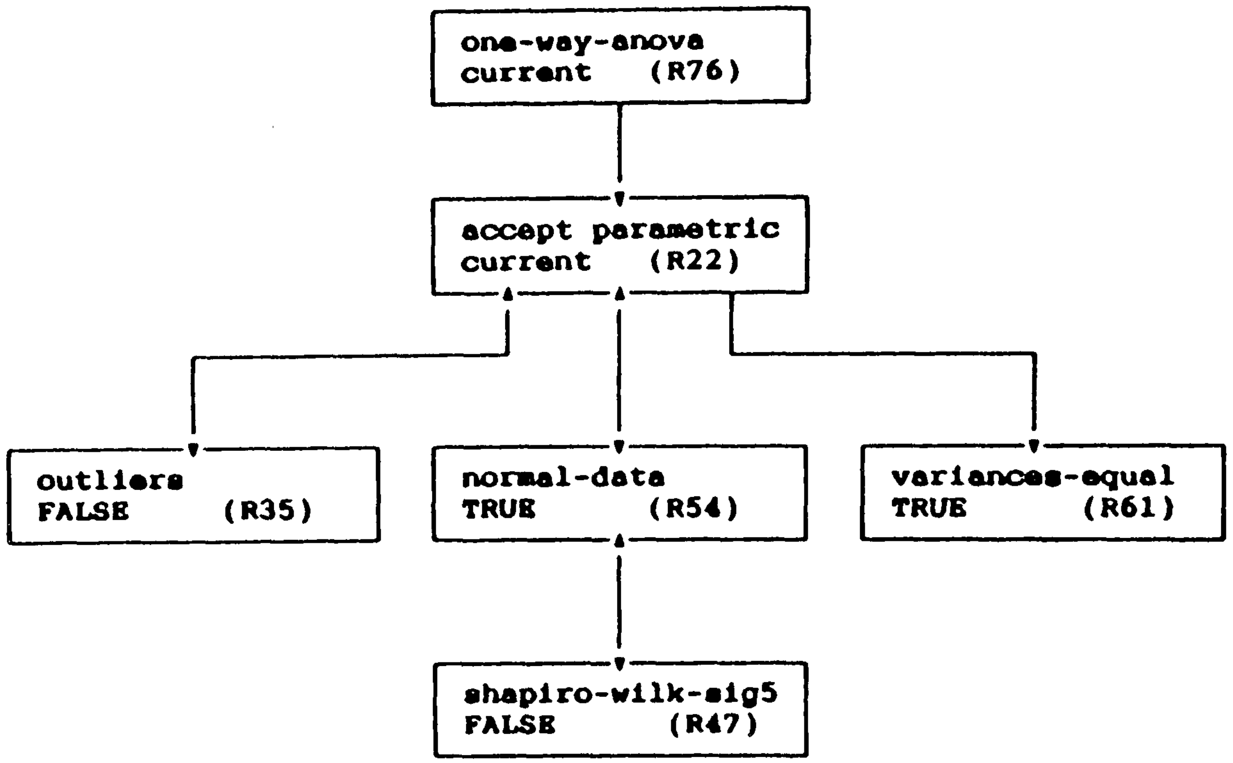


Figure 9.8 : Snapshot 7

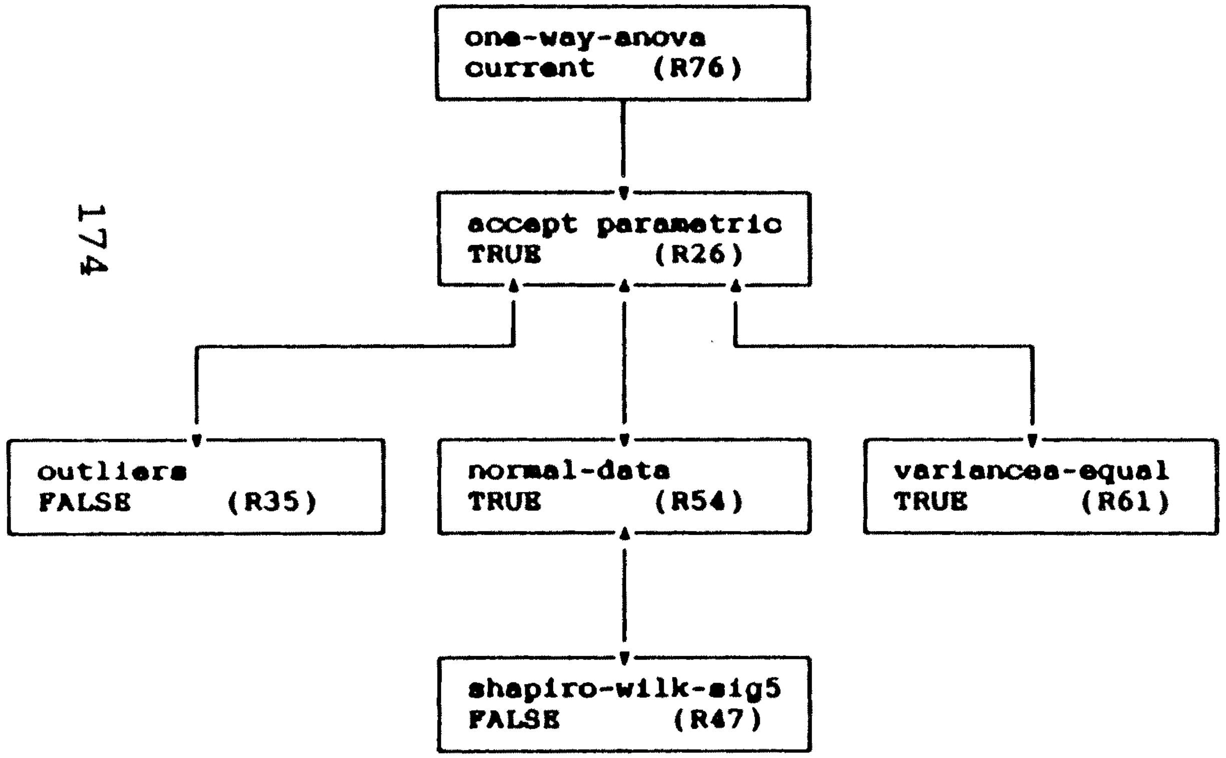
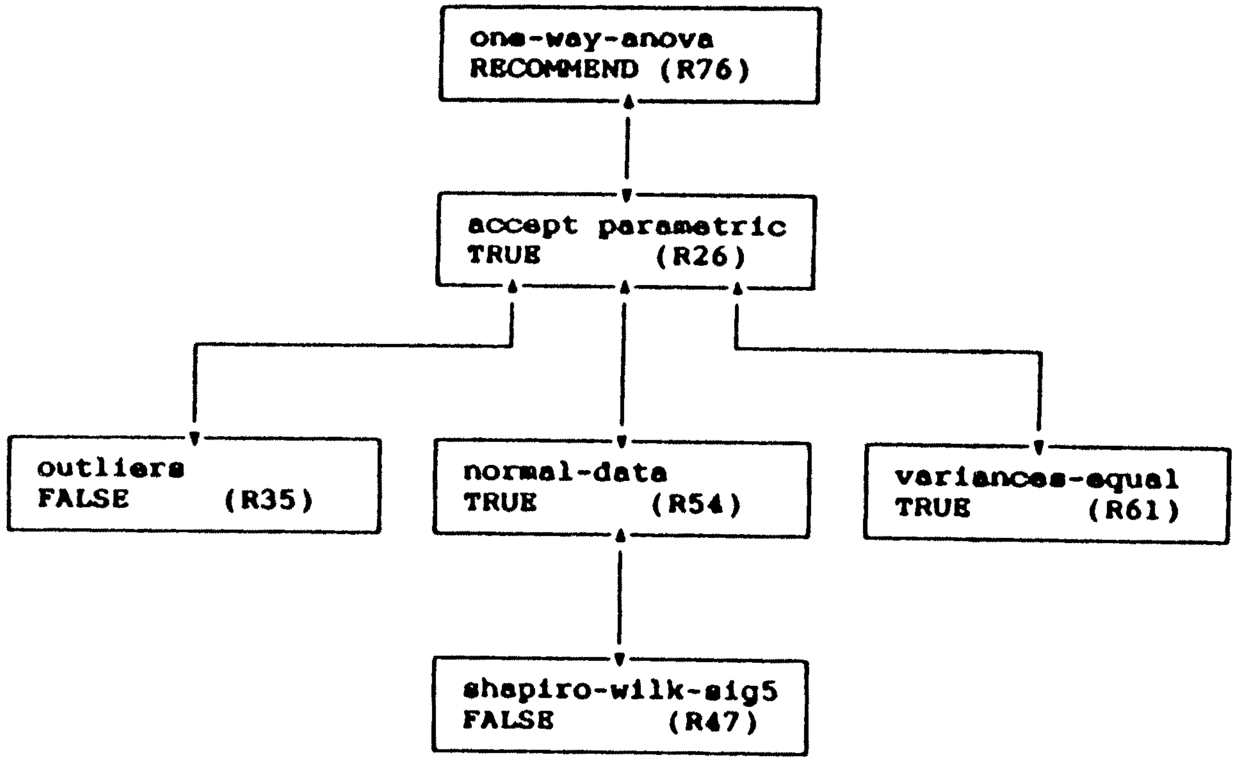


Figure 9.9 : Snapshot 8



More versatility is required for handling transformations of the data. The user may wish to consider a number of possible transformations and then select the most appropriate, this may be possible by extending the meta rules.

At any stage at the consultation it would be helpful to be able to go back a stage, the user may wish to change their response to a previous question in the light of subsequent knowledge.

In order to assist the local expert in modifying the rule-base it may be helpful group rule according to their context; for example, rules used to establish normality. In addition the rule-base editor could be extended to allow the user to identify all the rules that would be affected by a given change, so helping to ensure consistency within the rule-base.

Chapter Ten

Conclusions

10.1 Objectives of the Project

The main objective of this research was to investigate the practical aspects of designing and developing a Statistical Expert System that could be used by research workers who are not statisticians but who regularly need to carry out statistical analyses. A further aim of this project was to develop a system where the rule-base could be easily modified by a 'local expert statistician'.

This entailed research into a number of different areas from expert systems technology and knowledge acquisition to the problems of formalising statistical strategy and expertise.

10.2 Work Undertaken

A review of work already carried out in the area of Statistical Expert Systems was undertaken in order to establish some of the design criteria for such systems and to identify potential problems likely to be encountered in the development of Statistical Expert Systems. A postal şurvey of statisticians in industry supplied information on the potential role and problems of Expert Systems for Statistics and possible areas of application. The area of application chosen was the analysis of Completely Randomised Designs and multiple comparisons.

The logical design of the system was undertaken using Entity Analysis which provided a clear definition of the system and its requirements that was independent of software and hardware considerations. Entity Analysis also enables the designer to specify the flow of control within the system using graphical methods which can be easily understood and interpreted. Once the logical design was complete the methods of inference were considered in more dètail and different types of rule established for different parts of the consultation process. The software development used the logical design as a basis and further facilities such as options to access help text and to view the data were incorporated.

Knowledge acquisition was undertaken in parallel with the system design and development. A basic understanding of the knowledge domain is necessary to ensure that the knowledge representation is appropriate and sufficient. The process of knowledge acquisition involved interviews with practising statisticians, a review of knowledge available in papers and texts and a number of workshops with academic statisticians to consider specific case studies.

Evaluation trials were undertaken to assess the prototype system in terms of the ease of use, the pattern of consultation, the facilities available and the default rule-base supplied with the prototype system. A rule-base editor was also supplied with the prototype system to allow some assessment of the problems of allowing a local statistician to alter or extend the rule-base.

10.3 Results Achieved

10.3.1 Design and Development

The logical design provided a clear diagrammatic representation of the system and ensured careful consideration of the way in which knowledge was to be stored and processed within the system. This logical design was independent of hardware or software considerations and yet, when completed, provided a clear and comprehensive specification of the system that greatly facilitated the software development.

In this project attention was concentrated on the development of the rule-base processor. The rule-base editor was the subject of an undergraduate project. Some work has been done on the data entry module, sufficient to allow a user to create the necessary data files for the expert system. There were two main aspects to the rule-base processor, the coding of the inference engine to process the knowledge and the development of the facilities and user interface. The development of the backward chaining part of the inference engine was the most complex part of the coding requiring the use of recursion.

10.3.2 Knowledge Acquisition

Despite the limited area of expertise the knowledge acquisition proved to be a complex and time consuming undertaking. Areas of expertise such as the effect of Non Normality, heteroscedasticity or outliers are common to many areas of statistical analysis. Practising statisticians could be expected to have clear ideas about which statistical tests are useful for assessing the validity of assumptions related to these areas. They should also be aware of the relative importance of these assumptions with respect to their own application area. It was envisaged that 'local' statisticians would want to incorporate their own expertise for these areas. Consequently, some knowledge acquisition was undertaken in these areas but attention was focussed on the area of multiple comparisons where local statisticians would be less likely to have extensive expertise.

By dealing with the knowledge acquisition in terms of 'technical' and 'professional' expertise it was possible to develop a rational default rule-base which could then be modified by local experts. The Statistics Research Group provided a useful forum for assessing the default rule-base during its development. Some form of feedback is necessary during the development of a rule-base to prevent problems which may arise from lack of understanding on the part of the knowledge engineer. The workshops, as a form of protocol analysis, gave useful insight into the different possible approaches. The interviews with statisticians, described in section 5.6.2, were useful in giving some insight into their general strategy. One interesting point that came out of these interviews was the limited experience in the area of multiple comparisons; most of the statisticians interviewed used only one or two different techniques. This re-enforced our decision to concentrate on the area of multiple comparisons. We expect the statistician to be more likely to want to alter the rule-base with respect to such areas as Non Normality.

10.3.3 Evaluation Trials

The evaluation trials enabled us to assess the efficacy of the system particularly with respect to the general pattern of consultation and the facilities available to the user. The overall impression of the system gained from these trials was that it was easy to use. Opinions on the facilities varied widely, some helpful suggestions for improvements were made. The consensus of opinion seemed to be that there were enough facilities but that some of them needed developing or extending further. The inclusion of graphical procedures for looking at the data were very popular but not extensive enough. These facilities to view the data do not use any of the information in the knowledge base but are simply regarded as decision support facilities, nevertheless they seem very popular with users. A number of recommendations for improvements to the prototype system have already been made in the previous chapter.

10.3.4 Drawbacks to This Approach

The prototype system was written in Turbo Pascal which meant that the development of the code for the inference engine was more difficult than it would have been if an Artificial Intelligence language such as Prolog had been used. By using Pascal it was possible to incorporate statistical routines into the system fairly easily. However it would be more helpful to be able to interface to an existing statistical package, current technology for microcomputers makes this difficult. The provision for incorporating new procedures is particularly important as it forms an integral part of allowing a statistician to extend the rulebase.

The system was developed on an IBM-AT compatible and this also put some limitations on the development of the system. The large amount of memory required to hold the knowledge base during a consultation means that making the system user-friendly has been more difficult than it would have been if windowing systems or more extensive graphics, both of which are memory intensive, could have been used.

10.4 Associated Areas of Research

10.4.1 Developments in Computing and Expert Systems Research

The speed, memory and power available in microcomputers is changing so rapidly that it is becoming easier to interface with other packages or languages. For example, it is now possible to incorporate routines written in the language C into a program developed in Prolog using Borland's Turbo Prolog. This could be a useful development tool as the inference part of an expert system could be written much more easily and quickly at the same time as allowing statistical routines written in C to be incorporated. Improvements in operating systems technology also means that interfacing different software will become easier. The great advances being made in microcomputers also mean that it will become easier to incorporate more extensive graphics and windowing systems, both of which will help to make systems more user-friendly.

Expert Systems technology is developing rapidly, for example, methodologies for knowledge acquisition are becoming more powerful. This is a particularly important area of research because knowledge acquisition is so time consuming. Research is also being undertaken in the area of rule induction systems where the system 'learns' rules from examples, however this is a particularly complex task and it could be some time before such systems become widely used.

Expert system shells are rapidly becoming more versatile in their forms of knowledge representation and methods of inference and so will become more widely applicable

10.4.2 Study of Statistical Methods and Strategy

Statistics is itself a dynamic science and is continually developing and changing. It is essential that Statistical Expert Systems be flexible enough to be able to incorporate new techniques and methodologies as they are developed. In order to do this, the manner in which knowledge is expressed within a system needs to be comprehensible to the expert statistician who wishes to either extend or assess the knowledge base.

The application of expert systems techniques in the domain of statistics has resulted in an increased interest into the formalization of statistical strategy. In order to construct an expert system some conceptual model of the decision making process or strategy is required. For example, in the system GLIMPSE (Nelder 1986) the analytic process in perceived in terms of nine activities including data definition, model selection and model checking, each of which also contain a specific strategy. The system TESS (Pregibon 1986b) provides a way in which expert statisticians can encode their own strategy in a tree based structure which can then be used by non statisticians. As the understanding of statistical strategy is improved it will become easier to develop knowledge based systems. However it is also the case that as knowledge based systems are developed for statistics, understanding of strategy will be extended.

10.5 Recommendations for Further Research

10.5.1 Different Languages and Packages

As computing and expert systems research is developing so rapidly it would now be feasible to build a Statistical Expert System using a combination of an Artificial Intelligence language, such as Prolog, and a procedural language such as Pascal or C. By using more than one language it is possible to use the language most appropriate and powerful for each of the different aspects of the system. This would make maintaining and developing the system easier. It would also be worthwhile investigating the possibility of interfacing with existing statistical software available on microcomputers such as SAS.

10.5.2 Other Areas of Statistics

The prototype system was developed by considering a specific, well defined area of expertise. It would be useful to undertake the knowledge acquisition for a different area of expertise to enable the researcher to assess more closely the problems of knowledge acquisition and the structure of the system.

10.6 The Future Role of Expert Systems in Statistics

The advent of widely available and powerful computing facilities has had a marked effect on statistics in two ways. New, computationally intensive, methods for statistical analysis became possible and so statisticians were able to tackle larger and more complex analyses than previously. The other effect is the increasing availability of powerful statistical packages, available to statistician and non-statistician alike. This has meant that much more analysis is being undertaken by non-statisticians on a regular basis and some protection of these users is necessary.

The introduction of Expert Systems technology into Statistics is also likely to have a radical effect on statistical analysis in the future. The immediate advantage is that of providing protection against misuse of statistical methods for the non-statistician. Possible long term effects include changes in emphasis for both consulting and research statisticians. Consulting statisticians will see less of the routine analyses and concentrate on more complex analyses and advising about the availability of different consulting systems; the consulting statistician will also be involved in 'tuning' the knowledge base to their own particular field of application. The research statistician will not only be researching new methodologies and processes but also be involved in helping to formalize statistical strategy so that it can be incorporated into consulting systems. The research statisticians will also be involved in the development of new knowledge based systems.

The use of expert systems will also enable the statisticians to consider multiple answers in two different ways. Firstly, an expert system can consider several alternatives at a given stage in the analysis, whereas a statistician using conventional software will usually only pursue one possible route. Secondly an expert system can be structured so that the user can consider multiple answers in terms of a sequential analysis; for example, using a combination of statistical techniques in sequence rather than relying on the result of a single procedure.

Statistical Expert Systems should have a great impact on education, particularly of statistically naive users. Tukey (1986) suggested that in the future non-statisticians will probably benefit most from learning how to use a number of Statistical Expert systems for specific areas rather than only being given a basic course in statistics. However the educational aspect is currently seen as a potentially beneficial 'side-effect' of expert systems, primarily because of their potential to explain a course of reasoning on request. The naive user can follow the system as it works through a problem, requesting explanations as necessary. The difficulty with this is that the explanations currently offered by most systems are not particularly helpful, often consisting of a list of rules that have fired. In order to produce knowledge based systems that are useful for education it will be necessary to improve the current methods of providing explanations and to have a better understanding of the needs of the student. The ideal would be a system that could tailor itself to the student by learning from its own interaction with the student.

10.7 Conclusion

Research into statistical expert systems is still in the early stages and much remains to be done. This project provided an insight into the issues that need to be tackled in building such a system. As such it has shown that the development of systems that are expert in small, well-defined areas is a realistic proposition.

The software developed can also be considered as an expert system shell; knowledge bases relevant to other areas of statistics could be developed using the system. However it would be necessary to extend the rule-base editor and routine interface, as discussed in previous sections, before statisticians could begin to develop their own rule-bases.

By developing the system as a 'shell' and encouraging the development of different rule-bases a closer understanding would be gained of the nature of statistical knowledge and strategy. This understanding would then lead to further developments and improvements in future statistical expert systems.

<u>REFERENCES</u>

Chapter 1

Aikins J.S., Kunz J.C., Shortliffe E.H. and Fallat R.J. (1984) PUFF: An Expert System for Interpretation of Pulmonary Function Data
Readings in Medical Artificial Intelligence : The First Decade pp444-455 Ed. Clancey W.J. & Shortliffe E.H. (Addison Wesley)

Chambers J.M. (1981a) Some Thoughts on Expert Software
Proc. Interface of Computer Science and Statistics 13th Symposium pp36-40

Duda R., Gashnig J. and Hart P. (1979) Model Design in the PROSPECTOR Consultant System for Mineral Exploitation
Expert Systems in the Micro Electronic Age pp153-167
Ed. Michie D. (Edinburgh University Press)

Hahn G.J. (1984) Statistical Expert Systems and Intelligent Statistical Software
General Electric Report 84CRD173

Hahn G.J. (1985) More Intelligent Statistical Software and Statistical Expert Systems : Future Directions
The American Statistician Vol.39 No.1 pp1-16

Hand D.J. (1986a) Patterns in Statistical Strategy
Artificial Intelligence and Statistics pp355-388
Ed. Gale W. (Addison Wesley)

Hand D.J. (1986b) Expert Systems in Statistics
The Knowledge Engineering Review Vol.1 No.3 pp2-10

Haspel D. and Taunton C. (1986) Application of Rule-base Control in the Cement Industry
Expert Systems and Optimisation in Process Control pp53-61
Ed. Mamdani A. and Efstathiou J. (Technical Press)

Haux R. (1985) Expert Systems in Statistics : Some Problems and Some New Views
Proc. 9th German Workshop on Artificial Intelligence pp313-322

Huber P.J. (1985) Environments for Supporting Statistical Strategy
Artificial Intelligence and Statistics pp285-294
Ed. Gale W. (Addison Wesley)

Jones B. (1980) The Computer as a Statistical Consultant
Bulletin in Applied Statistics Vol.7 No.2 pp168-195

Lindsay R.K., Buchanan B.G., Feigenbaum E.A. and Lederberg J. (1980) Applications of Artificial Intelligence for Organic Chemistry (McGraw Hill)

Nelder J. (1984) Present Position and Potential Developments : Some Personal Views : Statistical Computing
J.R.Statist.Soc. A Vol.147 Part 2 pp151-160

Pregibon D. (1986a) A D.I.Y. Guide to Statistical Strategy
Artificial Intelligence and Statistics pp389-400
Ed. Gale W. (Addison Wesley)

Rauch-Hindin W. (1988) A Guide to Commercial Artificial Intelligence : Fundamentals and Real-World Applications (Prentice Hall)

Sell P.S. (1985) Expert Systems - A Practical Introduction (Macmillan)

Shortliffe E.H. (1976) Computer Based Medical Consultation : MYCIN (Elsevier)

Spiegelhalter D.J. and Knill-Jones R.P. (1984)
Statistical and Knowledge-based Approaches to Clinical Decision-support Systems with an Application in Gastroenterology
J.R.Statist.Soc. A Vol.147 Part 1 pp35-77

Tukey J. (1986) An Alphabet for Statisticians Expert Systems
Artificial Intelligence and Statistics pp401-409
Ed. Gale W. (Addison Wesley)

Chapter 2

Berzuini C., Ross G. and Larizza C. (1986) Developing Intelligent Software for Non-Linear Model Fitting as an Expert System
Proceedings of Compstat86 pp259-264 (Physica-Verlag)

Blum R.L. (1984) Discovery, Confirmation, and Incorporation of Causal Relationships from a Large Time-Oriented Clinical Database : The RX Project
Readings in Medical Artificial Intelligence pp399-425
Ed. Clancey W.J. & Shortliffe E.H. (Addison Wesley)

Carlsen F. and Heuch I. (1986) Express - An Expert System Utilizing Standard Statistical Packages
Proceedings of Compstat86 pp265-270 (Physica-Verlag)

Chambers J.M., Pregibon D. and Zayas E.R. (1981) Expert Software for Data Analysis: An Initial Experiment
Proc. 43rd Session of the International Statistical Institute, Beunos Aires pp294-309

Dambroise E. and Massotte P. (1986) Muse : An Expert System in Statistics
Proceedings of Compstat86 pp271-276 (Physica-Verlag)

Darius P.L. (1986) Building Expert Systems with the Help of Existing Statistical Software : An Example
Proceedings of Compstat86 pp277-282 (Physica-Verlag)

Esposito F., Capozza F. and Altini F. (1986) Exper: An Expert System in the Experimental Design
Short Communication - Compstat86 p83

Froeschl K.A. and Grossmann W. (1986) Knowledge Base Supported Analysis of Longitudinal Data
Proceedings of Compstat86 pp289-294 (Physica-Verlag)

Gale W.A. and Pregibon D. (1982) An Expert System for Regression Analysis
Comp. Sc. & Statistics - 14th Symposium on the Interface pp110-117

Gale W.A. and Pregibon D. (1984) AI Research in Statistics
AI Magazine Vol.5 Part 4 pp72-75

Gale W.A. (1986) Student Phase 1 - A Report on work in progress
Artificial Intelligence and Statistics pp239-266
Ed. Gale W. (Addison Wesley)

Galmacci G. (1986) A Knowledge Based System for the Time Series Analysis
Short Communication - Compstat86 p93

Hahn G.J. (1985) More Intelligent Statistical Software and Statistical Expert Systems: Future Directions
The American Statistician Vol.39 No.1 pp1-16

Hájek P., Ivánek J. (1982) Artificial Intelligence and Data Analysis
Proceedings of Compstat82 pp54-60 (Physica-Verlag)

Hakong L. and Hickman F.R. (1985) Expert Systems Techniques: An Application in Statistics
Expert Systems 85 pp43-63 Ed. Merry M. (Cambridge University Press)

Hietala P. (1986) How to Assist an Inexperienced User in the Preliminary Analysis of a Time Series : First Version of the ESTES Expert System
Proceedings of Compstat86 pp295-300 (Physica-Verlag)

Hilhorst R.A., Van Romunde L.K.J., Troquay T.P.H. and Van Den Berg J.V. (1987) Rochefort: Research on Creating a Human Environment for On-Line Research Tools
Statistical Software Newsletter Vol.13 pp47-56

Jida J. and Lemaire J. (1986) Expert Systems and Data Analysis Package Management
Proceedings of Compstat86 pp251-258 (Physica-Verlag)

Klösgen W. (1986) EXPLORA : An Example of Knowledge Based Data Analysis
Expert Systems in Statistics pp45-60 Ed. Haux R. (Gustav Fischer)

Nelder J. (1986) AI and Generalized Linear Modelling : An Expert System for GLIM
AI Methods in Statistics pp34-43 (Unicom Seminar)

O'Keefe R. (1982) An Expert System for Statistics
Proc. of Technical Conference on the Theory and Practice of Knowledge Based Systems, Brunel University

Oldford R.W. and Peters S.C. (1986a) Object-Oriented Data Representations for Statistical Data Analysis
Proceedings of Compstat86 pp301-308 (Physica-Verlag)

Oldford R.W. and Peters S.C. (1986b) Implementation and Study of Statistical Strategy
Artificial Intelligence and Statistics pp335-354
Ed Gale W. (Addison Wesley)

Portier K.M. and Lai P. (1983) A Statistical Expert System for Analysis Determination
Proc. Statistical Computing Section of American Statistical Assoc. pp309-311

Prat A., Marti M. and Catot J.M. (1985) Incorporating Expertise in Time Series Modelling : The STATXPS System
Statistical Software Newsletter Vol.11 No.2 pp55-62

Pregibon D. (1986b) Data Analysis as Search
AI Methods in Statistics pp1-17 (Unicom Seminar)

Smith A.M.R., Lee L.S. and Hand D.J. (1983) Interactive User-Friendly Interfaces to Statistical Packages
The Computer Journal Vol.26 No.3 pp199-204

Chapter 3

Buja A. (1984) Why Mimicking Data Analysts by Expert Systems
Appendix to Pregibon (1986)

Chen P.P. (1977) The Entity-Relationship Approach to Logical Database Design (Wellesley, Mass)

Hahn G.J. (1985) More Intelligent Statistical Software and Statistical Expert Systems: Future Directions
The American Statistician Vol.39 No.1 pp1-16

Hand D.J. (1984) Statistical Expert Systems: Design
The Statistician No.33 pp351-369

Hand D.J. (1985) Statistical Expert Systems: Necessary Attributes
Journal of Applied Statistics Vol.12 No.1 pp19-27

Knight B., Cross M. and Edwards D. (1987) Software Design Strategies for Numerical Software
Reliability and Robustness of Engineering Software pp121-136 Edited papers of 1st International Conference, Como, Italy (Elsevier)

Michie D. and Johnston R. (1984) The Creative Computer : Machine Intelligence and Human Knowledge (Pelican Books)

Pregibon D. (1986b) Data Analysis as Search
AI Methods in Statistics pp1-17 (Unicom Seminar)

Chapter 5

Bell M.Z. (1985) Why Expert Systems Fail
Jnl.Op.Res.Soc. Vol.36 No.7 pp613-619

Burton, M, and Shadbolt, N. (1987) Knowledge Engineering Technical Report 87-2-1 University of Nottingham, Dept. of Psychology, Artificial Intelligence Group

Duda R.O. and Shortliffe, E.H. (1983) Expert Systems Research
Science Vol.220 No.4594 pp261-268

Easterby-Smith M. (1981) The Design, Analysis and Interpretation of Repertory Grids
Recent Advances in Personal Construct Technology pp9-30 (Academic Press)

Feigenbaum E.A. and McCorduck P. (1983) The Fifth Generation
(Addison Wesley)

Gale W.A. (1987) Knowledge-based Knowledge Acquisition for a Statistical Consulting System
Int.J.Man-Machine Studies Vol.26 pp55-64

Gammack J.G and Young R.M. (1985) Psychological Techniques for Eliciting Expert Knowledge
Research and Development in Expert Systems pp105-112
Ed. M Bramer (Cambridge University Press)

Nii H.P. (1984) The Knowledge Engineer at Work
pp80-84 in Feigenbaum and McCorduck (1983)

Thisted R.A. (1986) Representing Knowledge for Expert Data Analysis Systems
Artificial Intelligence and Statistics pp267-284
Ed. Gale W. (Addison Wesley)

Welbank M. (1983) A Review of Knowledge Acquisition Techniques for Expert Systems
British Telecommunications, Ipswich

Wittkowski K.M. (1986) Generating and Testing Statistical Hypotheses : Strategies for Knowledge Engineering
Expert Systems in Statistics pp139-154
Ed. Haux R. (Gustav Fischer)

Chapter 6

Barnett V. and Lewis T. (1984) Outliers in Statistical Data
(Wiley)

Dyer A.R. (1974) Comparisons of tests for Normality with a Cautionary Note
Biometrika Vol.61 pp185-189

D'Agostino R.B. & Stephens M.A. (1986) Goodness-of-Fit Techniques
(Dekker)

Huber P.J. (1981) Robust Statistics (Wiley)

Kimball A.W. (1957) Errors of the Third Kind in Statistical Consulting
Jnl.Am.Stat.Assoc. Vol.52 pp133-142

Levene H. (1960) Robust Tests for Equality of Variances
Contributions to Probability and Statistics
Ed. Olkin I. et al (Stanford University Press)

Miller R.G. (1981) Simultaneous Statistical Inference
(Springer-Verlag)

Miller R.G. (1986) Beyond ANOVA : Basics of Applied Statistics
(Wiley)

Neyman J. and Pearson E.S. (1933) On the Problem of the most Efficient Tests of Statistical Hypotheses
Phil.Trans.A Vol.231 pp289-337

Pearson E.S. and Please N. W. (1975) Relation Between the Shape of Population Distribution and the Robustness of Four Simple Test Statistics
Biometrika Vol.62 No.2 pp223-241

Shapiro S.S., Wilk M.B. and Chen H.J. (1968) A Comparative Study of Various Tests for Normality
Jnl.Am.Stat.Assoc. Vol.63 pp1343-1372

Scheffè H. (1959) The Analysis of Variance (Wiley)

Chapter 7

Bartholomew D.J. (1961) Ordered Tests in Analysis of Variance
Biometrika Vol.48 pp325-332

Brown R.A. (1974) Robustness of the Studentized Range Statistics
Biometrika Vol.61 No.1 pp171-175

Brown M.B. and Forsythe A.B. (1974) The Analysis of Variance and Multiple Comparisons for Data With Heterogeneous Variances
Biometrics Vol.30 pp719-724

Carmer S.G. and Swanson M.R. (1973) An Evaluation of Ten Pairwise Multiple Comparison Procedures by Monte Carlo Methods
Jnl.Am.Stat.Assoc. Vol.66 pp66-74

Chew V. (1976) Comparing Treatment Means : A Compendium
Hortscience Vol.11 No.4 pp348-357

Cochran W.G. (1964) Approximate Significance Levels of the Behrens-Fisher Test
Biometrics Vol.20 pp191-195

Cornell J.A. (1971) A Review of Multiple Comparison Procedures for Comparing a Set of K Population Means
Proc.Soil and Crop Science Soc.,Florida Vol.31 pp92-97

Dodge Y. & Thomas D.R. (1980) On the Performance of Nonparametric and Normal Theory Multiple Comparison Procedures
Sankyha (B) 42 Part 1&2 pp11-27

Duncan D.B. (1955) Multiple Range and Multiple F Tests
Biometrics Vol.11 pp1-42

Dunnett C.W. (1955) A Multiple Comparisons Procedure for Comparing Several Treatments With a Control
Jnl.Am.Stat.Assoc. Vol.50 pp1096-1121

Dunnett C.W. (1980a) Pairwise Multiple Comparisons in the Homogeneous Variance, Unequal Sample Size Case
Jnl.Am.Stat.Assoc. Vol.75 pp789-795

Dunnett C.W. (1980b) Pairwise Multiple Comparisons in the Unequal Variance Case
Jnl.Am.Stat.Assoc. Vol.75 pp796-800

Dunnett C. and Goldsmith C. (1981) When and How to do Multiple Comparisons
Statistics in the Pharmaceutical Industry pp397-433
(Dekker)

Dwass M. (1955) A Note on Simultaneous Confidence Intervals
Ann. of Math. Vol.26 pp146-147

Fisher R.A. (1935) Statistical Methods for Research Workers
(Oliver and Boyd)

Gabriel K.R. (1978) A Simple Method of Multiple Comparisons of Means
Jnl.Am.Stat.Assoc. Vol.73 pp724-729

Games P.A. and Howell J.F. (1976) Pairwise Multiple Comparison Procedures With Unequal n's and/or Variances : A Monte Carlo Study
Journal of Educational Statistics Vol.1 pp113-125

Genizi A. and Hochberg Y. (1978) On Improved Extension of the T-Method of Multiple comparisons for Unbalanced Designs
Jnl.Am.Stat.Assoc. Vol.73 pp879-884

Gill J.L. (1973) Current Status of Multiple Comparison of Means in Designed Experiments
Jnl. of Dairy Science Vol.56 pp973-977

Hochberg Y. (1974) Some Generalizations of the T-Method in Simultaneous Inference
Journal of Multivariate Analysis Vol.4 pp224-234

Keuls M. (1952) The Use of the "Studentized Range" in Connection With an Analysis of Variance
Euphytica Vol.1 pp112-122

Kramer C.Y. (1956) Extension of Multiple Range Tests to Group Means With Unequal Number of Replications
Biometrics Vol.12 pp307-310

Miller R.G. (1981) Simultaneous Statistical Inference
(Springer-Verlag)

Miller R.G. (1986) Beyond ANOVA : Basics of Applied Statistics
(Wiley)

Nemenyi P. (1963) Distribution-free Multiple Comparisons
Unpublished doctoral thesis, Princeton University

Newman D. (1939) The Distribution of the Range in Samples From a Normal Population, Expressed in Terms of an Independent Estimate of Standard Deviation
Biometrika Vol.31 pp20-30

O'Neil R. & Wetherill G.B. (1971) The Present State of Multiple Comparison Methods
J.R.Stat.Soc Series B No.33 pp218-250

Scheffé H. (1953) A Method For Judging All Contrasts in the Analysis of Variance
Biometrika Vol.40 pp87-104

Scheffè H. (1959) The Analysis of Variance (Wiley)

Shirley E. (1977) A nonparametric Equivalent of Williams Test for Contrasting Increasing Dose Levels of a Treatment
Biometrics Vol.33 pp386-389

Sidàk Z. (1967) Rectangular Confidence Regions for the Means of Multivariate Normal Distributions
Jnl.Am.Stat.Assoc. Vol.62 pp626-633

Spjøtvoll E. and Stoline M.R. (1973) An Extension of the T-Method of Multiple Comparison to Include the Cases With Unequal Sample Sizes
Jnl.Am.Stat.Assoc. Vol.68 pp975-978

Steel R.G.D. (1959) A Multiple Comparison Rank Sum Test : Treatments Versus Control
Biometrics Vol.15 pp560-572

Steel R.G.D. (1960) A Rank Sum Test for Comparing All Pairs of Treatments
Technometrics Vol.2 pp197-207

Stoline M.R. (1981) The Status of Multiple Comparisons : Simultaneous Estimation of all Pairwise Comparisons in One-Way ANOVA Designs
The American Statistician Vol.35 No.3 pp134-141

Tamhane A.C. (1979) A Comparison of Procedures for Multiple Comparison of Means with Unequal Variances
Jnl.Am.Stat.Assoc. Vol.74 pp471-480

Thomas D.A.H. (1973) Multiple Comparisons Among Means - A Review
The Statistician Vol.22 No.1 pp16-42

Thomas D.A.H. (1974) Error Rates in Multiple Comparisons Among Means : Results of a Simulation Exercise
J.R.Stat.Soc. Series C No.23 pp284-294

Tukey J.W. (1952) Allowances for Various Types of Error Rates
Unpublished IMS address, Virginia Polytechnic Institute, Blacksburg

Waller R.A. and Duncan D.B. (1969) A Bayes Rule for the Symmetric Multiple Comparisons Problem
Jnl.Am.Stat.Assoc. Vol.64 pp1484-1503

Welch B.L. (1938) The Significance of the Difference Between Two Means When the Population Variances are Unequal
Biometrika Vol.29 pp350-362

Williams D.A. (1971) A Test for Differences Between Treatment Means When Several Dose Levels are Compared with a Zero Dose Control
Biometrics Vol.27 pp103-117

Williams D.A. (1972) The Comparison of Several Dose Levels
With a Zero Dose Control
Biometrics Vol.28 pp519-531

Williams D.A. (1986) A Note on Shirleys Nonparametric
Procedure for Comparing Several Dose Levels With a Zero
Dose Response
Biometrics Vol.42 pp183-186

Appendices

Appendix I

Document Sent to Statisticians

Dear

The growing demand for statistical analysis throughout industry and commerce, coupled with increasing sophistication and availability of statistical software leaves the statistician with the ever increasing problem of providing an adequate service and monitoring the use of statistical methods by non-statisticians in his organisation.

The enclosed document outlines an approach which the Statistics Research Group at Thames Polytechnic is pursuing as a means of dealing with this problem. The possibility of introducing 'intelligent' statistical applications packages is considered as a means of filling the gap and relieving the statistician of some of the more routine work.

We would be most interested to hear of your views on future developments in computing in this area, together with any steps that you have already taken along this road. We would also be extremely interested to hear of any routine statistical analysis problems which statistically untrained members of your organisation handle, together with the amount of statistical protection that you, or the software currently in use, provides. We would appreciate your opinion on whether expert systems such as we are proposing could be of value in these areas.

Thank you in anticipation of your reply.

Yours sincerely,

J.R. Alexander MSc M.I.S.
E.E. Bell BSc

INTELLIGENT STATISTICAL APPLICATIONS PACKAGES

INTRODUCTION

In industry and research there is a permanent shortage of professional statisticians, with the inevitable consequence that statistical analysis is often undertaken by non-statisticians who have limited access (if any) to the expertise of the statistician. With the advent of powerful general purpose statistical packages, methods can easily be inappropriately applied which can lead to potentially misleading results. There is a clear need to make the knowledge of the expert statistician available to these users without needing to take up too much of the time of the expert in answering routine enquiries.

INTELLIGENT SOFTWARE

The development of interactive statistical software incorporating statistical expertise could help to meet the needs outlined above. We perceive that there will be two major advantages:-

a) The professional statistician could be relieved of some of the more routine enquiries and thus be able to give greater time to the more difficult tasks.

b) The non-statistician would be protected to a large degree from the inappropriate application of methods and the misinterpretation of results, without needing to have the relevant statistical expertise.

Such systems would need to incorporate many of the features of expert-systems, the more important features are mentioned below.

APPLICATIONS PACKAGES

1. To be of manageable size to implement on small computers and to relate closely to the needs of users, these systems would need to be designed for specific application areas for which there is a frequent demand. For example:-

   a) Design of acceptance sampling schemes.

   b) Analysis of animal carcinogenicity studies.

   c) Sales forecasting.

2. The package would provide advice and would include a dialogue with the user to ensure that the selected method(s) for analysis is appropriate to the data. The user would be referred to a professional statistician when the problem does not fall within the class for which the system was designed.

3. The system would be capable of explaining its reasoning, on request, so that the statistical rationale for the methods employed is made clear to the user.

4. Such packages would need to cater for a variety of inferential and computational procedures, so that the statistician installing and maintaining the service could set the advice and judgements provided by the system according to his own usual practice.

5. For computation, the system would either interface with a statistical package or provide its own 'number crunching' facility.

Appendix II

Listing of Prototype Rulebase

The rules in the rulebase ANOVA3 have been divided into sections which are dependent on their context as follows ; R1 - R10 are Forward Chaining rules, the remainder are processed using backward chaining;

R1 - R10	Establishing the tests to be looked at
R11 - R21	Also used when trying to establish which tests to look at
R22 - R29	Accepting parametric or nonparametric techniques and/or transformations
R30 - R35	Outliers
R36 - R54	Normality
R55 - R61	Equality of variances
R62 - R67	One sample methods
R68 - R75	Methods for two samples
R76 - end	Methods for several samples

```
R1    IF     ONE_GROUP
      THEN   ONE_SAMPLE_NORMAL (TEST) LOOK_AT
             ONE_SAMPLE_T (TEST) LOOK_AT
             ONE_SAMPLE_WILCOXON (TEST) LOOK_AT

R2    IF     TWO_GROUPS
      THEN   NORMAL_POOLED_VAR (TEST) LOOK_AT
             NORMAL_SEPARATE_VAR (TEST) LOOK_AT
             TWO_SAMPLE_T (TEST) LOOK_AT
             ASPIN_WELCH (TEST) LOOK_AT
             TWO_SAMPLE_WILCOXON (TEST) LOOK_AT

R3    IF     SEVERAL_GROUPS AND
             OVERALL_TEST
      THEN   ONE_WAY_ANOVA (TEST) LOOK_AT
             KRUSKAL_WALLIS_ANOVA (TEST) LOOK_AT

R4    IF     MULTIPLE_COMPARISONS AND
             COMP_WITH_CONTROL AND
             USER_COMP_W_CONTROL
      THEN   DUNNETT (TEST) LOOK_AT
             BONFERRONI_T (TEST) LOOK_AT
             MANY_ONE_RANK (TEST) LOOK_AT

R5    IF     MULTIPLE_COMPARISONS AND
             LOWEST_DOSE_RESPONSE AND
             USER_DOSE_RESPONSE
      THEN   WILLIAMS (TEST) LOOK_AT
             SHIRLEYS (TEST) LOOK_AT

R6    IF     MULTIPLE_COMPARISONS AND
             DESIGNED_CONTRASTS
      THEN   LINEAR_CONTRASTS (TEST) LOOK_AT
             BONFERRONI (TEST) LOOK_AT
             SUGGESTED_CONTRASTS (FACT) FALSE

R7    IF     MULTIPLE_COMPARISONS AND
             PAIRWISE AND
             ALL_COMPARISONS
      THEN   NEWMAN_KEULS (TEST) LOOK_AT
             DUNCANS (TEST) LOOK_AT
             K_SAMPLE_RANK (TEST) LOOK_AT
             KRUSKAL_WALLIS_PAIRS (TEST) LOOK_AT

R8    IF     MULTIPLE_COMPARISONS AND
             PAIRWISE AND
             NOT ALL_COMPARISONS
      THEN   TUKEY (TEST) LOOK_AT
             TUKEY_KRAMER (TEST) LOOK_AT
             T3 (TEST) LOOK_AT
             C (TEST) LOOK_AT
             K_SAMPLE_RANK (TEST) LOOK_AT
             KRUSKAL_WALLIS_PAIRS (TEST) LOOK_AT

R9    IF     MULTIPLE_COMPARISONS AND
             SUGGESTED_CONTRASTS
      THEN   SCHEFFE (TEST) LOOK_AT
             BONFERRONI (TEST) LOOK_AT

R10   IF     MULTIPLE_COMPARISONS AND
             DESIGNED_CONTRASTS AND
             NOT ORTHOGONAL
      THEN   BONFERRONI (TEST) LOOK_AT
```

```
R11  IF    CONTROL_GROUP AND
           LEVELS_OF_TREATMENT
     THEN  LOWEST_DOSE_RESPONSE (FACT) TRUE
           COMP_WITH_CONTROL (FACT) FALSE

R12  IF    CONTROL_GROUP AND
           NOT LEVELS_OF_TREATMENT
     THEN  COMP_WITH_CONTROL (FACT) TRUE
           LOWEST_DOSE_RESPONSE (FACT) FALSE

R13  IF    NOT DESIGNED_CONTRASTS
     THEN  SUGGESTED_CONTRASTS (FACT) TRUE

R14  IF    ONE_GROUP
     THEN  MULTIPLE_COMPARISONS (FACT) FALSE

R15  IF    TWO_GROUPS
     THEN  MULTIPLE_COMPARISONS (FACT) FALSE

R16  IF    ONE_WAY_ANOVA AND
           FURTHER_ANALYSIS
     THEN  MULTIPLE_COMPARISONS (FACT) TRUE

R17  IF    KRUSKAL_WALLIS_ANOVA AND
           FURTHER_ANALYSIS
     THEN  MULTIPLE_COMPARISONS (FACT) TRUE

R18  IF    SEVERAL_GROUPS AND
           NOT OVERALL_TEST
     THEN  MULTIPLE_COMPARISONS (FACT) TRUE

R19  IF    ONE_WAY_ANOVA AND
           NOT FURTHER_ANALYSIS
     THEN  MULTIPLE_COMPARISONS (FACT) FALSE

R20  IF    KRUSKAL_WALLIS_ANOVA AND
           NOT FURTHER_ANALYSIS
     THEN  MULTIPLE_COMPARISONS (FACT) FALSE

R21  IF    NOT KRUSKAL_WALLIS_ANOVA AND
           NOT ONE_WAY_ANOVA
     THEN  MULTIPLE_COMPARISONS (FACT) FALSE
```

```
R22  IF     NOT OUTLIERS AND
            NORMAL_DATA AND
            NOT VARIANCES_EQUAL AND
            USER_AGREE_TO_TRANS
     THEN   ACCEPT_PARAMETRIC (FACT) FALSE
            TRANS_FOR_NORMALITY (FACT) TRUE
            TRANS_FOR_VARIANCES (FACT) FALSE
            ADJUST_FOR_UNEQ_VAR (FACT) TRUE

R23  IF     NOT OUTLIERS AND
            NOT VARIANCES_EQUAL AND
            NORMAL_DATA AND
            NOT USER_AGREE_TO_TRANS
     THEN   ACCEPT_PARAMETRIC (FACT) FALSE
            TRANS_FOR_VARIANCES (FACT) FALSE
            TRANS_FOR_NORMALITY (FACT) FALSE
            ADJUST_FOR_UNEQ_VAR (FACT) TRUE

R24  IF     NOT OUTLIERS AND
            NOT VARIANCES_EQUAL AND
            NOT NORMAL_DATA AND
            USER_AGREE_TO_TRANS
     THEN   ACCEPT_PARAMETRIC (FACT) FALSE
            TRANS_FOR_NORMALITY (FACT) TRUE
            TRANS_FOR_VARIANCES (FACT) TRUE
            ADJUST_FOR_UNEQ_VAR (FACT) FALSE

R25  IF     NOT OUTLIERS AND
            VARIANCES_EQUAL AND
            NOT NORMAL_DATA AND
            USER_AGREE_TO_TRANS
     THEN   ACCEPT_PARAMETRIC (FACT) TRUE
            TRANS_FOR_VARIANCES (FACT) FALSE
            TRANS_FOR_NORMALITY (FACT) TRUE
            ADJUST_FOR_UNEQ_VAR (FACT) FALSE

R26  IF     NOT OUTLIERS AND
            VARIANCES_EQUAL AND
            NORMAL_DATA
     THEN   ACCEPT_PARAMETRIC (FACT) TRUE
            TRANS_FOR_VARIANCES (FACT) FALSE
            TRANS_FOR_NORMALITY (FACT) FALSE
            ADJUST_FOR_UNEQ_VAR (FACT) FALSE

R27  IF     NOT OUTLIERS AND
            VARIANCES_EQUAL AND
            NOT NORMAL_DATA AND
            NOT USER_AGREE_TO_TRANS
     THEN   ACCEPT_PARAMETRIC (FACT) FALSE
            TRANS_FOR_NORMALITY (FACT) FALSE
            TRANS_FOR_VARIANCES (FACT) FALSE
            ADJUST_FOR_UNEQ_VAR (FACT) FALSE

R28  IF     NOT OUTLIERS AND
            NOT VARIANCES_EQUAL AND
            NOT NORMAL_DATA AND
            NOT USER_AGREE_TO_TRANS
     THEN   ACCEPT_PARAMETRIC (FACT) FALSE
            TRANS_FOR_VARIANCES (FACT) FALSE
            TRANS_FOR_NORMALITY (FACT) FALSE
            ADJUST_FOR_UNEQ_VAR (FACT) FALSE

R29  IF     OUTLIERS
     THEN   ACCEPT_PARAMETRIC (FACT) FALSE
            TRANS_FOR_VARIANCES (FACT) FALSE
            TRANS_FOR_NORMALITY (FACT) FALSE
            ADJUST_FOR_UNEQ_VAR (FACT) FALSE
```

```
R30  IF     MAX_GROUPSIZE_GT_25 AND
            USER_SAYS_OUTLIERS
     THEN   OUTLIERS (FACT) TRUE

R31  IF     MAX_GROUPSIZE_GT_25 AND
            NOT USER_SAYS_OUTLIERS
     THEN   OUTLIERS (FACT) FALSE

R32  IF     NOT MAX_GROUPSIZE_GT_25 AND
            DIXONS_SIG_5 AND
            USER_SAYS_OUTLIERS
     THEN   EXPLAIN_OUTLIERS (PROC) CALL
            OUTLIERS (FACT) TRUE

R33  IF     NOT MAX_GROUPSIZE_GT_25 AND
            DIXONS_SIG_5 AND
            NOT USER_SAYS_OUTLIERS
     THEN   OUTLIERS (FACT) FALSE

R34  IF     NOT MAX_GROUPSIZE_GT_25 AND
            NOT DIXONS_SIG_5 AND
            USER_SAYS_OUTLIERS
     THEN   OUTLIERS (FACT) TRUE

R35  IF     NOT MAX_GROUPSIZE_GT_25 AND
            NOT DIXONS_SIG_5 AND
            NOT USER_SAYS_OUTLIERS
     THEN   OUTLIERS (FACT) FALSE
```

```
R36  IF    NOT MORE_THAN_10_OVERALL AND
           NORMAL_BY_USER
     THEN  NORMAL_DATA (FACT) TRUE

R37  IF    NOT MORE_THAN_10_OVERALL AND
           NOT NORMAL_BY_USER
     THEN  NORMAL_DATA (FACT) FALSE

R38  IF    NOT ONE_GROUP AND
           NOT MORE_THAN_25_OVERALL AND
           VARIANCES_EQUAL AND
           SHAPWILK_ALL_RMS_5 AND
           SHAPWILK_BY_GROUP_5
     THEN  SHAPIRO_WILK_SIG5 (FACT) TRUE

R39  IF    NOT ONE_GROUP AND
           NOT MORE_THAN_25_OVERALL AND
           VARIANCES_EQUAL AND
           SHAPWILK_ALL_RMS_5 AND
           NOT SHAPWILK_BY_GROUP_5
     THEN  SHAPIRO_WILK_SIG5 (FACT) TRUE

R40  IF    NOT ONE_GROUP AND
           NOT MORE_THAN_25_OVERALL AND
           VARIANCES_EQUAL AND
           NOT SHAPWILK_ALL_RMS_5 AND
           SHAPWILK_BY_GROUP_5
     THEN  SHAPIRO_WILK_SIG5 (FACT) FALSE

R41  IF    NOT ONE_GROUP AND
           NOT MORE_THAN_25_OVERALL AND
           VARIANCES_EQUAL AND
           NOT SHAPWILK_ALL_RMS_5 AND
           NOT SHAPWILK_BY_GROUP_5
     THEN  SHAPIRO_WILK_SIG5 (FACT) FALSE

R42  IF    NOT ONE_GROUP AND
           NOT MORE_THAN_25_OVERALL AND
           NOT VARIANCES_EQUAL AND
           SHAPWILK_ALL_GSD_5 AND
           SHAPWILK_BY_GROUP_5
     THEN  SHAPIRO_WILK_SIG5 (FACT) TRUE

R43  IF    NOT ONE_GROUP AND
           NOT MORE_THAN_25_OVERALL AND
           NOT VARIANCES_EQUAL AND
           SHAPWILK_ALL_GSD_5 AND
           NOT SHAPWILK_BY_GROUP_5
     THEN  SHAPIRO_WILK_SIG5 (FACT) TRUE

R44  IF    NOT ONE_GROUP AND
           NOT MORE_THAN_25_OVERALL AND
           NOT VARIANCES_EQUAL AND
           NOT SHAPWILK_ALL_GSD_5 AND
           SHAPWILK_BY_GROUP_5
     THEN  SHAPIRO_WILK_SIG5 (FACT) FALSE
```

```
R45  IF     NOT ONE_GROUP AND
            NOT MORE_THAN_25_OVERALL AND
            NOT VARIANCES_EQUAL AND
            NOT SHAPWILK_ALL_GSD_5 AND
            NOT SHAPWILK_BY_GROUP_5
     THEN   SHAPIRO_WILK_SIG5 (FACT) FALSE

R46  IF     MORE_THAN_25_OVERALL AND
            SHAPWILK_BY_GROUP_5
     THEN   SHAPIRO_WILK_SIG5 (FACT) TRUE

R47  IF     MORE_THAN_25_OVERALL AND
            NOT SHAPWILK_BY_GROUP_5
     THEN   SHAPIRO_WILK_SIG5 (FACT) FALSE

R48  IF     MORE_THAN_10_OVERALL AND
            NOT MORE_THAN_20_OVERALL AND
            SHAPIRO_WILK_SIG5 AND
            NORMAL_BY_USER
     THEN   NORMAL_DATA (FACT) TRUE

R49  IF     MORE_THAN_10_OVERALL AND
            NOT MORE_THAN_20_OVERALL AND
            SHAPIRO_WILK_SIG5 AND
            NOT NORMAL_BY_USER
     THEN   NORMAL_DATA (FACT) FALSE

R50  IF     MORE_THAN_10_OVERALL AND
            NOT MORE_THAN_20_OVERALL AND
            NOT SHAPIRO_WILK_SIG5 AND
            NORMAL_BY_USER
     THEN   NORMAL_DATA (FACT) TRUE

R51  IF     MORE_THAN_10_OVERALL AND
            NOT MORE_THAN_20_OVERALL AND
            NOT SHAPIRO_WILK_SIG5 AND
            NOT NORMAL_BY_USER
     THEN   NORMAL_DATA (FACT) FALSE

R52  IF     MORE_THAN_20_OVERALL AND
            SHAPIRO_WILK_SIG5 AND
            NOT NORMAL_BY_USER
     THEN   NORMAL_DATA (FACT) FALSE

R53  IF     MORE_THAN_20_OVERALL AND
            SHAPIRO_WILK_SIG5 AND
            NORMAL_BY_USER
     THEN   NORMAL_DATA (FACT) TRUE

R54  IF     MORE_THAN_20_OVERALL AND
            NOT SHAPIRO_WILK_SIG5
     THEN   NORMAL_DATA (FACT) TRUE
```

```
R55  IF    LEVENE_SIG_5 AND
           BARTLETT_SIG1 AND
           OVERIDE_VAR_EQ_TEST
     THEN  VARIANCES_EQUAL (FACT) TRUE

R56  IF    LEVENE_SIG_5 AND
           BARTLETT_SIG1 AND
           NOT OVERIDE_VAR_EQ_TEST
     THEN  VARIANCES_EQUAL (FACT) FALSE

R57  IF    LEVENE_SIG_5 AND
           NOT BARTLETT_SIG1 AND
           OVERIDE_VAR_EQ_TEST
     THEN  VARIANCES_EQUAL (FACT) TRUE

R58  IF    LEVENE_SIG_5 AND
           NOT BARTLETT_SIG1 AND
           NOT OVERIDE_VAR_EQ_TEST
     THEN  VARIANCES_EQUAL (FACT) FALSE

R59  IF    NOT LEVENE_SIG_5 AND
           BARTLETT_SIG1 AND
           OVERIDE_VAR_EQ_TEST
     THEN  VARIANCES_EQUAL (FACT) TRUE

R60  IF    NOT LEVENE_SIG_5 AND
           BARTLETT_SIG1 AND
           NOT OVERIDE_VAR_EQ_TEST
     THEN  VARIANCES_EQUAL (FACT) FALSE

R61  IF    NOT LEVENE_SIG_5 AND
           NOT BARTLETT_SIG1
     THEN  VARIANCES_EQUAL (FACT) TRUE
```

```
R62  IF    MORE_THAN_25_OVERALL AND
           NOT OUTLIERS
     THEN  ONE_SAMPLE_NORMAL (TEST) RECOMMEND
           ONE_SAMPLE_T (TEST) VALID
           ONE_SAMPLE_WILCOXON (TEST) VALID

R63  IF    NOT MORE_THAN_25_OVERALL
     THEN  ONE_SAMPLE_NORMAL (TEST) NOT_VALID

R64  IF    OUTLIERS
     THEN  ONE_SAMPLE_NORMAL (TEST) NOT_VALID
           ONE_SAMPLE_T (TEST) NOT_VALID
           ONE_SAMPLE_WILCOXON (TEST) RECOMMEND

R65  IF    NOT MORE_THAN_25_OVERALL AND
           NOT OUTLIERS AND
           NORMAL_DATA
     THEN  ONE_SAMPLE_T (TEST) RECOMMEND
           ONE_SAMPLE_WILCOXON (TEST) VALID

R66  IF    NOT MORE_THAN_25_OVERALL AND
           NOT OUTLIERS AND
           NOT NORMAL_DATA AND
           USER_AGREE_TO_TRANS
     THEN  ONE_SAMPLE_T (TEST) NOT_VALID
           ONE_SAMPLE_WILCOXON (TEST) RECOMMEND
           TRANS_FOR_NORMALITY (FACT) TRUE

R67  IF    NOT MORE_THAN_25_OVERALL AND
           NOT OUTLIERS AND
           NOT NORMAL_DATA AND
           NOT USER_AGREE_TO_TRANS
     THEN  ONE_SAMPLE_T (TEST) NOT_VALID
           ONE_SAMPLE_WILCOXON (TEST) RECOMMEND
           TRANS_FOR_NORMALITY (FACT) FALSE
```

```
R68  IF    MORE_THAN_25_OVERALL AND
           ACCEPT_PARAMETRIC
     THEN  NORMAL_POOLED_VAR (TEST) RECOMMEND
           NORMAL_SEPARATE_VAR (TEST) VALID
           TWO_SAMPLE_T (TEST) VALID
           ASPIN_WELCH (TEST) VALID
           TWO_SAMPLE_WILCOXON (TEST) VALID

R69  IF    NOT MORE_THAN_25_OVERALL
     THEN  NORMAL_POOLED_VAR (TEST) NOT_VALID
           NORMAL_SEPARATE_VAR (TEST) NOT_VALID

R70  IF    NOT MORE_THAN_25_OVERALL AND
           ACCEPT_PARAMETRIC
     THEN  TWO_SAMPLE_T (TEST) RECOMMEND
           ASPIN_WELCH (TEST) VALID
           TWO_SAMPLE_WILCOXON (TEST) VALID

R71  IF    MORE_THAN_25_OVERALL AND
           NOT ACCEPT_PARAMETRIC AND
           NOT ADJUST_FOR_UNEQ_VAR
     THEN  NORMAL_POOLED_VAR (TEST) NOT_VALID
           NORMAL_SEPARATE_VAR (TEST) NOT_VALID
           TWO_SAMPLE_T (TEST) NOT_VALID
           ASPIN_WELCH (TEST) NOT_VALID
           TWO_SAMPLE_WILCOXON (TEST) RECOMMEND

R72  IF    MORE_THAN_25_OVERALL AND
           NOT ACCEPT_PARAMETRIC AND
           ADJUST_FOR_UNEQ_VAR AND
           NOT USER_AGREE_TO_TRANS
     THEN  NORMAL_SEPARATE_VAR (TEST) RECOMMEND
           ASPIN_WELCH (TEST) VALID
           NORMAL_POOLED_VAR (TEST) NOT_VALID
           TWO_SAMPLE_T (TEST) NOT_VALID
           TWO_SAMPLE_WILCOXON (TEST) VALID

R73  IF    MORE_THAN_25_OVERALL AND
           NOT ACCEPT_PARAMETRIC AND
           ADJUST_FOR_UNEQ_VAR AND
           USER_AGREE_TO_TRANS
     THEN  NORMAL_SEPARATE_VAR (TEST) VALID
           ASPIN_WELCH (TEST) VALID
           NORMAL_POOLED_VAR (TEST) NOT_VALID
           TWO_SAMPLE_T (TEST) NOT_VALID
           TWO_SAMPLE_WILCOXON (TEST) VALID

R74  IF    NOT MORE_THAN_25_OVERALL AND
           NOT ACCEPT_PARAMETRIC AND
           NOT ADJUST_FOR_UNEQ_VAR
     THEN  TWO_SAMPLE_T (TEST) NOT_VALID
           ASPIN_WELCH (TEST) NOT_VALID
           TWO_SAMPLE_WILCOXON (TEST) RECOMMEND

R75  IF    NOT MORE_THAN_25_OVERALL AND
           NOT ACCEPT_PARAMETRIC AND
           ADJUST_FOR_UNEQ_VAR
     THEN  TWO_SAMPLE_T (TEST) NOT_VALID
           ASPIN_WELCH (TEST) RECOMMEND
           TWO_SAMPLE_WILCOXON (TEST) VALID
```

```
R76  IF    ACCEPT_PARAMETRIC
     THEN  ONE_WAY_ANOVA (TEST) RECOMMEND
           KRUSKAL_WALLIS_ANOVA (TEST) VALID

R77  IF    NOT ACCEPT_PARAMETRIC
     THEN  ONE_WAY_ANOVA (TEST) NOT_VALID
           KRUSKAL_WALLIS_ANOVA (TEST) RECOMMEND

R78  IF    ACCEPT_PARAMETRIC AND
           BALANCED
     THEN  DUNNETT (TEST) RECOMMEND
           BONFERRONI_T (TEST) VALID
           MANY_ONE_RANK (TEST) VALID
           KRUSKALWALLIS_MANY_1 (TEST) VALID

R79  IF    ACCEPT_PARAMETRIC AND
           NOT BALANCED
     THEN  DUNNETT (TEST) NOT_VALID
           BONFERRONI_T (TEST) RECOMMEND
           MANY_ONE_RANK (TEST) NOT_VALID
           KRUSKALWALLIS_MANY_1 (TEST) VALID

R80  IF    NOT ACCEPT_PARAMETRIC AND
           BALANCED
     THEN  DUNNETT (TEST) NOT_VALID
           BONFERRONI_T (TEST) NOT_VALID
           MANY_ONE_RANK (TEST) RECOMMEND
           KRUSKALWALLIS_MANY_1 (TEST) VALID

R81  IF    NOT ACCEPT_PARAMETRIC AND
           NOT BALANCED
     THEN  DUNNETT (TEST) NOT_VALID
           BONFERRONI_T (TEST) NOT_VALID
           MANY_ONE_RANK (TEST) NOT_VALID
           KRUSKALWALLIS_MANY_1 (TEST) RECOMMEND

R82  IF    ACCEPT_PARAMETRIC
     THEN  WILLIAMS (TEST) RECOMMEND
           SHIRLEYS (TEST) VALID

R83  IF    NOT ACCEPT_PARAMETRIC
     THEN  WILLIAMS (TEST) NOT_VALID
           SHIRLEYS (TEST) RECOMMEND

R84  IF    ACCEPT_PARAMETRIC AND
           BALANCED
     THEN  TUKEY (TEST) RECOMMEND
           TUKEY_KRAMER (TEST) VALID
           K_SAMPLE_RANK (TEST) VALID
           KRUSKAL_WALLIS_PAIRS (TEST) VALID
```

```
R85  IF     ACCEPT_PARAMETRIC AND
            NOT BALANCED
     THEN   TUKEY (TEST) NOT_VALID
            TUKEY_KRAMER (TEST) RECOMMEND
            K_SAMPLE_RANK (TEST) NOT_VALID
            KRUSKAL_WALLIS_PAIRS (TEST) VALID

R86  IF     NOT ACCEPT_PARAMETRIC AND
            NOT BALANCED
     THEN   TUKEY (TEST) NOT_VALID
            TUKEY_KRAMER (TEST) NOT_VALID
            K_SAMPLE_RANK (TEST) NOT_VALID
            KRUSKAL_WALLIS_PAIRS (TEST) RECOMMEND

R87  IF     ACCEPT_PARAMETRIC AND
            BALANCED
     THEN   NEWMAN_KEULS (TEST) RECOMMEND
            K_SAMPLE_RANK (TEST) VALID
            KRUSKAL_WALLIS_PAIRS (TEST) VALID

R88  IF     ACCEPT_PARAMETRIC AND
            NOT BALANCED
     THEN   NEWMAN_KEULS (TEST) NOT_VALID
            K_SAMPLE_RANK (TEST) NOT_VALID
            KRUSKAL_WALLIS_PAIRS (TEST) RECOMMEND

R89  IF     NOT ACCEPT_PARAMETRIC AND
            BALANCED
     THEN   NEWMAN_KEULS (TEST) NOT_VALID
            K_SAMPLE_RANK (TEST) RECOMMEND
            KRUSKAL_WALLIS_PAIRS (TEST) VALID

R90  IF     NOT ACCEPT_PARAMETRIC AND
            NOT BALANCED
     THEN   NEWMAN_KEULS (TEST) NOT_VALID
            K_SAMPLE_RANK (TEST) NOT_VALID
            KRUSKAL_WALLIS_PAIRS (TEST) RECOMMEND

R91  IF     ONE_GROUP AND
            SHAPWILK_BY_GROUP_5
     THEN   SHAPIRO_WILK_SIG5 (FACT) TRUE

R92  IF     ONE_GROUP AND
            NOT SHAPWILK_BY_GROUP_5
     THEN   SHAPIRO_WILK_SIG5 (FACT) FALSE

R93  IF     NOT CONTROL_GROUP
     THEN   COMP_WITH_CONTROL (FACT) FALSE
            LOWEST_DOSE_RESPONSE (FACT) FALSE

R94  IF     NOT PAIRWISE
     THEN   ALL_COMPARISONS (FACT) FALSE

R95  IF     NOT ACCEPT_PARAMETRIC AND
            BALANCED
     THEN   TUKEY (TEST) NOT_VALID
            TUKEY_KRAMER (TEST) NOT_VALID
            K_SAMPLE_RANK (TEST) RECOMMEND
            KRUSKAL_WALLIS_PAIRS (TEST) VALID
```

```
R96  IF    NOT ACCEPT_PARAMETRIC AND
           BALANCED
     THEN  TUKEY (TEST) NOT_VALID
           TUKEY_KRAMER (TEST) NOT_VALID
           K_SAMPLE_RANK (TEST) RECOMMEND
           KRUSKAL_WALLIS_PAIRS (TEST) VALID

R97  IF    ACCEPT_PARAMETRIC AND
           BALANCED
     THEN  NEWMAN_KEULS (TEST) RECOMMEND
           DUNCANS (TEST) VALID
           K_SAMPLE_RANK (TEST) VALID
           KRUSKAL_WALLIS_PAIRS (TEST) VALID

R98  IF    ACCEPT_PARAMETRIC AND
           NOT BALANCED
     THEN  NEWMAN_KEULS (TEST) NOT_VALID
           DUNCANS (TEST) NOT_VALID
           K_SAMPLE_RANK (TEST) NOT_VALID
           KRUSKAL_WALLIS_PAIRS (TEST) RECOMMEND
           ALL_COMPARISONS (FACT) FALSE

R99  IF    NOT ACCEPT_PARAMETRIC AND
           BALANCED
     THEN  NEWMAN_KEULS (TEST) NOT_VALID
           DUNCANS (TEST) NOT_VALID
           K_SAMPLE_RANK (TEST) RECOMMEND
           KRUSKAL_WALLIS_PAIRS (TEST) VALID

R100 IF    NOT ACCEPT_PARAMETRIC AND
           NOT BALANCED
     THEN  NEWMAN_KEULS (TEST) NOT_VALID
           DUNCANS (TEST) NOT_VALID
           K_SAMPLE_RANK (TEST) NOT_VALID
           KRUSKAL_WALLIS_PAIRS (TEST) RECOMMEND

R101 IF    ACCEPT_PARAMETRIC AND
           MANY_CONTRASTS
     THEN  SCHEFFE (TEST) RECOMMEND
           BONFERRONI (TEST) VALID

R102 IF    ACCEPT_PARAMETRIC AND
           NOT MANY_CONTRASTS
     THEN  BONFERRONI (TEST) RECOMMEND
           SCHEFFE (TEST) VALID

R103 IF    NOT ACCEPT_PARAMETRIC
     THEN  SCHEFFE (TEST) NOT_VALID
           BONFERRONI (TEST) NOT_VALID

R104 IF    ACCEPT_PARAMETRIC AND
           ORTHOGONAL
     THEN  LINEAR_CONTRASTS (TEST) RECOMMEND
           BONFERRONI (TEST) VALID

R105 IF    ACCEPT_PARAMETRIC AND
           NOT ORTHOGONAL
     THEN  BONFERRONI (TEST) RECOMMEND
           LINEAR_CONTRASTS (TEST) NOT_VALID
```

```
R106 IF    NOT ACCEPT_PARAMETRIC
     THEN  BONFERRONI (TEST) NOT_VALID
           SCHEFFE (TEST) NOT_VALID

R107 IF    ACCEPT_PARAMETRIC
     THEN  T3 (TEST) VALID
           C (TEST) VALID

R108 IF    NOT ACCEPT_PARAMETRIC AND
           NOT ADJUST_FOR_UNEQ_VAR
     THEN  T3 (TEST) NOT_VALID
           C (TEST) NOT_VALID

R109 IF    NOT ACCEPT_PARAMETRIC AND
           ADJUST_FOR_UNEQ_VAR AND
           MAX_GROUPSIZE_GT_25
     THEN  C (TEST) RECOMMEND
           T3 (TEST) VALID

R110 IF    NOT ACCEPT_PARAMETRIC AND
           ADJUST_FOR_UNEQ_VAR AND
           NOT MAX_GROUPSIZE_GT_25
     THEN  T3 (TEST) RECOMMEND
           C (TEST) VALID
```

FACTS USED IN RULE BASE ANOVA3

OUTLIERS	Set by -	Rule
VARIANCES_EQUAL	Set by -	Rule
NORMAL_DATA	Set by -	Rule
TRANS_FOR_NORMALITY	Set by -	Rule
TRANS_FOR_VARIANCES	Set by -	Rule
SEVERAL_GROUPS	Set by -	Procedure TEST_NUM_GROUPS
OVERALL_TEST	Set by -	User
FURTHER_ANALYSIS	Set by -	User
PAIRWISE	Set by -	User Rule
DIXONS_SIG_5	Set by -	Procedure DIXONS_TEST
BARTLETT_SIG1	Set by -	Procedure BARTLETTS_TEST
MORE_THAN_10_OVERALL	Set by -	Procedure TEST_TOTAL_OBS
MORE_THAN_20_OVERALL	Set by -	Procedure TEST_TOTAL_OBS
SHAPIRO_WILK_SIG5	Set by -	Rule
OVERIDE_VAR_EQ_TEST	Set by -	User
MULTIPLE_COMPARISONS	Set by -	Rule
ONE_GROUP	Set by -	Procedure TEST_NUM_GROUPS
TWO_GROUPS	Set by -	Procedure TEST_NUM_GROUPS
LEVELS_OF_TREATMENT	Set by -	User
LOWEST_DOSE_RESPONSE	Set by -	User Rule
SUGGESTED_CONTRASTS	Set by -	User Rule
ACCEPT_PARAMETRIC	Set by -	Rule
USER_AGREE_TO_TRANS	Set by -	User
MORE_THAN_25_OVERALL	Set by -	Procedure TEST_TOTAL_OBS
NORMAL_BY_USER	Set by -	User
ADJUST_FOR_UNEQ_VAR	Set by -	Rule
BALANCED	Set by -	Procedure TEST_BALANCED
SHAPWILK_ALL_RMS_5	Set by -	Procedure SHAPWILK_ALL_RMS
SHAPWILK_BY_GROUP_5	Set by -	Procedure SHAPWILK_BY_GROU
SHAPWILK_ALL_GSD_5	Set by -	Procedure SHAPWILK_ALL_GSD
LEVENE_SIG_5	Set by -	Procedure LEVENES_TEST
USER_SAYS_OUTLIERS	Set by -	User
MAX_GROUPSIZE_GT_25	Set by -	Procedure TEST_GROUP_SIZE
COMP_WITH_CONTROL	Set by -	Rule
USER_COMP_W_CONTROL	Set by -	User
USER_DOSE_RESPONSE	Set by -	User
ORTHOGONAL	Set by -	User
ALL_COMPARISONS	Set by -	User Rule
CONTROL_GROUP	Set by -	User
DESIGNED_CONTRASTS	Set by -	User
MANY_CONTRASTS	Set by -	User

```
TESTS USED IN RULE BASE ANOVA3

          TUKEY_KRAMER  ( Parametric    )
               SCHEFFE  ( Parametric    )
         MANY_ONE_RANK  ( Nonparametric )
              WILLIAMS  ( Parametric    )
              SHIRLEYS  ( Nonparametric )
                 TUKEY  ( Parametric    )
     ONE_SAMPLE_NORMAL  ( Parametric    )
     NORMAL_POOLED_VAR  ( Parametric    )
           ASPIN_WELCH  ( Parametric    )
         ONE_WAY_ANOVA  ( Parametric    )
   ONE_SAMPLE_WILCOXON  ( Nonparametric )
   TWO_SAMPLE_WILCOXON  ( Nonparametric )
          ONE_SAMPLE_T  ( Parametric    )
          TWO_SAMPLE_T  ( Parametric    )
               DUNNETT  ( Parametric    )
          NEWMAN_KEULS  ( Parametric    )
  KRUSKAL_WALLIS_ANOVA  ( Nonparametric )
          BONFERRONI_T  ( Parametric    )
  KRUSKALWALLIS_MANY_1  ( Nonparametric )
         K_SAMPLE_RANK  ( Nonparametric )
  KRUSKAL_WALLIS_PAIRS  ( Nonparametric )
      LINEAR_CONTRASTS  ( Parametric    )
                    T3  ( Parametric    )
                     C  ( Parametric    )
               DUNCANS  ( Parametric    )
            BONFERRONI  ( Parametric    )
   NORMAL_SEPARATE_VAR  ( Parametric    )
```

```
PROCEDURES USED IN RULE BASE ANOVA3

Procedure            TRANSFORM    called by RULES
Procedure        TEST_BALANCED    called by FINDFACT
Procedure          DIXONS_TEST    called by FINDFACT
Procedure     EXPLAIN_OUTLIERS    called by RULES
Procedure       BARTLETTS_TEST    called by FINDFACT
Procedure         LEVENES_TEST    called by FINDFACT
Procedure       TEST_TOTAL_OBS    called by FINDFACT
Procedure        SHAP_WILK_RMS    called by FINDFACT
Procedure   SHAP_WILK_GROUP_SD    called by FINDFACT
Procedure      TEST_NUM_GROUPS    called by FINDFACT
Procedure       TEST_TOTAL_OBS    called by FINDFACT
Procedure    SHAPWILK_BY_GROUP    called by FINDFACT
Procedure     SHAPWILK_ALL_RMS    called by FINDFACT
Procedure     SHAPWILK_ALL_GSD    called by FINDFACT
Procedure      TEST_GROUP_SIZE    called by FINDFACT
```

RULES OF RULE BASE META

```
M1   IF     NOT PARAMETRIC
     THEN   NEXT_TEST (FACT) TRUE

M2   IF     PARAMETRIC AND
            OUTLIERS
     THEN   NEXT_TEST (FACT) TRUE

M3   IF     PARAMETRIC AND
            NOT OUTLIERS AND
            NOT TRANS_FOR_NORMALITY AND
            NOT TRANS_FOR_VARIANCES
     THEN   NEXT_TEST (FACT) TRUE

M4   IF     PARAMETRIC AND
            NOT OUTLIERS AND
            TRANS_FOR_VARIANCES AND
            MORE_TRANS_TO_TRY
     THEN   TRANSFORM (PROC) CALL

M5   IF     PARAMETRIC AND
            NOT OUTLIERS AND
            TRANS_FOR_VARIANCES AND
            NOT MORE_TRANS_TO_TRY
     THEN   NEXT_TEST (FACT) TRUE

M6   IF     PARAMETRIC AND
            NOT OUTLIERS AND
            TRANS_FOR_NORMALITY AND
            MORE_TRANS_TO_TRY
     THEN   TRANSFORM (PROC) CALL

M7   IF     PARAMETRIC AND
            NOT OUTLIERS AND
            TRANS_FOR_NORMALITY AND
            NOT MORE_TRANS_TO_TRY
     THEN   NEXT_TEST (FACT) TRUE
```

FACTS USED IN RULE BASE META

PARAMETRIC	Set by -	Procedure TEST_PARAMETRIC
NEXT_TEST	Set by -	Metarule
OUTLIERS	Set by -	Rule
TRANS_FOR_NORMALITY	Set by -	Rule
TRANS_FOR_VARIANCES	Set by -	Rule
MORE_TRANS_TO_TRY	Set by -	Procedure TRANSFORM called by r

PROCEDURES USED IN RULE BASE META

Procedure	TEST_PARAMETRIC	called by FINDFACT
Procedure	TRANSFORM	called by RULES

Appendix III

Papers Presented or Published During the Course of This Research

Copies of the papers listed below, which were presented at conferences during the course of this research project, are included in this appendix.

(1) 'THESEUS : An Expert Statistical Consultant'
Presented at ICOSCO-I (International Conference on Statistical Computing), Izmir, Turkey, March 1987
To be included in the proceedings when they are published.

(2) An updated version of the same paper was presented at the DOSES (Development of Statistical Expert Systems) Seminar, Luxembourg, December 1987

(3) 'Knowledge Acquisition in the Development of THESEUS, a Statistical Expert System' Presented by John Alexander at the Royal Statistical Society Charter Centenary Conference, Cambridge, U.K., April 1987

(4) 'Building a Statistical Knowledge Base : A Discussion of the Approach Used in the Development of THESEUS, a Statistical Expert System' Presented at COMPSTAT 88, Copenhagen, Denmark, August 1988.

THESEUS : An Expert Statistical Consultant

Edwina Bell and Peter Watts:

Thames Polytechnic, London

John Alexander: Hazleton UK, Harrogate, N. Yorks.

1. Introduction

Traditionally, statistical packages have been general purpose computer programs such as SPSS, SAS, GLIM, and GENSTAT. Whilst programs such as these have wide applicability and are often very powerful, there are a number of problems associated with their use: particularly those of checking the appropriateness of an analysis and lack of 'user-friendliness' in the package.

With the advent of Artificial Intelligence techniques, it is becoming possible to tackle these problems. Expert systems are designed to approach a problem in a user-friendly manner and to incorporate 'real-world' expertise in their structure.

It is becoming clear that incorporating expertise in a general statistical package is a very large problem and that the best way forward is the development of specialised intelligent software, Hahn(1985).

In this paper, we describe THESEUS, an Expert System being developed at Thames Polytechnic which concentrates on the area of one-way analysis of variance and related techniques, which our researches have shown to be heavily in demand in industry.

THESEUS is being written in Turbo Pascal and it is envisaged that the final system will be available for use on an IBM XT micro-computer.

2. Problems of Statistical Expert Systems

Developing an expert system has its own difficulties, the major bottleneck being that of knowledge elicitation. The process of acquiring the knowledge appropriate to the area of application is often time-consuming and difficult and is further hampered, especially in statistics, by conflicts in expertise. There is the additional problem that once knowledge has been incorporated into the system an extensive process of testing and verification needs to be carried out to ensure consistency and acceptability of the expertise.

The 'end-user' also needs to be considered carefully. There is a wide range of possibilities from the expert statistical consultant to the statistical novice and it would be difficult to cater for all of them in a single system. The statistically naive researcher would need extensive help and guidance to ensure the appropriate analysis is carried out and to interpret the results, whereas the expert may want to move through the system quickly, looking only at the results he or she is interested in.

THESEUS attempts to cater for this by having a wide range of help and explanation facilities available and a flexible mode of operation. Although the system will be able to cope with a range of users, it has been designed for use primarily by research workers in industry who are regular users of statistical analytical techniques.

3. Design Considerations

There are a number of features that are desirable in a statistical expert system. As with any software, an expert system needs to be structured so that it is easily modifiable, both to allow for ease of maintenance of the system and to cope with developments in the knowledge base. An expert system in particular, has the additional features of explaining its operation to the user, being adaptable to changing circumstances in a consultation and catering for mistakes by the user.

A statistical expert system also needs to be able to explain statistical terms to the user, to allow for the possibility of multiple objectives and to identify appropriate tests and data exploration methods as well as selecting the most powerful technique.

Another consideration for a statistical expert system is the need to access statistical routines or packages in order to extract information from the raw data.

A number of people have considered these features: in particular Hand(1985a) and Hahn(1985) discuss them more fully.

4. Knowledge Acquisition

The problem of knowledge acquisition is particularly difficult in the area of statistics. Statistical expertise can come from a number of areas: academic knowledge to be found in text books and research papers, and expertise arising from the practice of statisticians in industry, government and commerce.

Although in theory there may be a large number of possible alternatives for an analysis, in practice an industrial statistician may not use the most powerful and appropriate technique but choose a less powerful technique because it will be understood more easily or because the statistician's client expects a particular technique to be used.

In developing THESEUS, a survey of the relevant literature was undertaken in order to build a preliminary knowledge base able to deal with one-way analysis of variance and multiple comparisons. In respect of the latter area, it is worth noting that the application of multiple comparison methods is a matter of considerable debate among statisticians (see for example, the paper by O'Neil and Wetherill(1971) and the subsequent discussion). This controversial area holds considerable interest for the development of statistical expert systems, since, as the above mentioned discussion shows, there are some grey areas in the knowledge domain; the extensive and continuing literature on the subject indicates the need for a dynamic knowledge base that can be easily updated to reflect new techniques and research findings. The process of knowledge acquisition in the area of multiple comparisons has, not surprisingly, proved to be both time consuming and open ended.

In addition to the literature search, a number of workshops have been held at the Polytechnic. In these, a member of the research group presents a set of data,in advance, to be analysed and acts as the research scientist. The rest of the research group then report back on how they, either individually or in small groups of two or three, analysed the data. The aim of these workshops is to see if a consensus of opinion can be reached about the nature of an analysis.

In order to find out what is applied in practice and how practice varies from statistician to statistician, a program of structured interviews with industrial statisticians was set up following response to a questionnaire. Preliminary results have indicated that there are marked differences both between individual statisticians in industry and between academic theory and industrial practice. This has had a direct bearing on the design of the expert system as there is obviously a need to allow the statistician in industry to 'tune' the system to their own requirements; this is discussed in more detail in a later section.

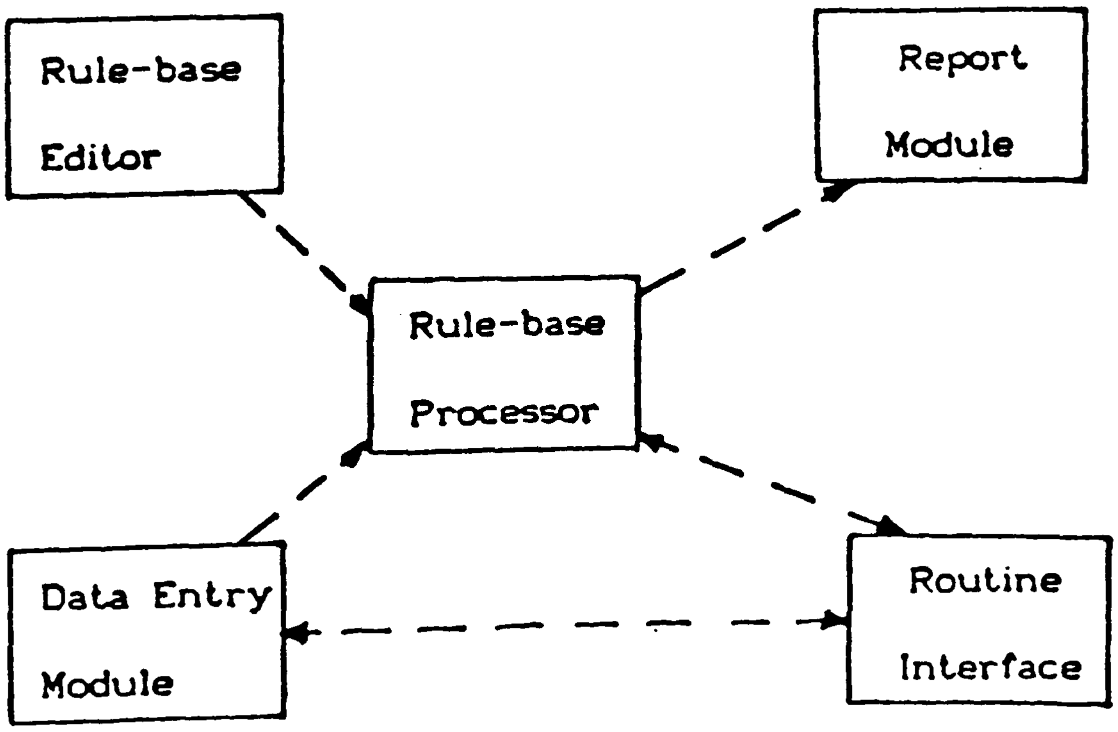
5. Structure of THESEUS

As in most diagnostic areas, a statistical consultant proceeds by initially generating plausible hypotheses about the analysis and then checking each hypothesis against the data to decide whether or not the analysis considered is appropriate, Hand(1985b). This leads to a natural structure for a statistical knowledge base.

THESEUS is basically a production rule system with two types of rule. One type deals with the selection of possible appropriate techniques, and is processed using forward chaining. The other type, processed by backward chaining, is concerned with verifying the applicability of these tests by checking assumptions, selecting transformations where necessary and dealing with possible outliers.

The software is designed to be highly modular to aid development and maintenance as well as facilitating comprehension and flexibility. (See Fig.1)

Figure 1 : Structure of THESEUS



The rule base editor produces a file of rules which can be picked up by the expert system. The editor enables an expert user to enter, delete and modify rules and checks the rule base for consistency and redundancy.

The data entry module allows the system user to enter and edit data, performs basic descriptive analyses and conducts a dialogue with the user to ensure that both the system and the user are satisfied with the representation of the data structure.

The rule-base processor works through the rules using a combination of forward and backward chaining, accessing the routine interface and reporting intermediate results as appropriate. Information required by the system can come from a number of sources:

- asking the user questions
- backward chaining through the rules
- initial data entry section
- intermediate analyses of the data during a consultation

The routine interface allows the system to perform statistical tests on the data by accessing libraries or packages, either during the consultation process or when carrying out recommended analyses .

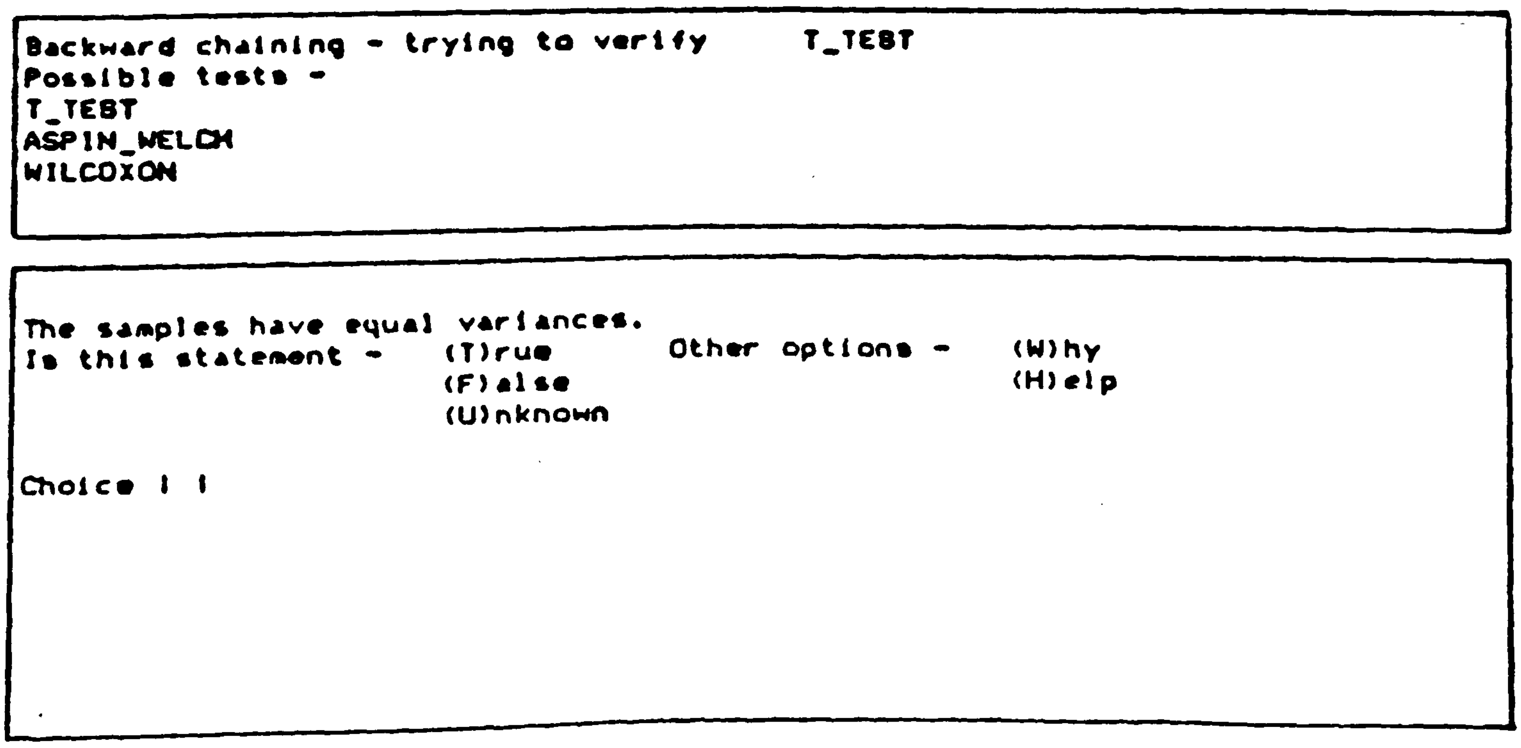
The report module provides the results of analyses for the user and allows the user to structure output in an appropriate way, accessing intermediate results as necessary.

6. Interfacing with the user

THESEUS presents the user with a split screen consisting of two windows (See Fig.2). The top window keeps the user informed of the state of the consultation process. The bottom window normally displays the current question during the running of the consultation. In response to a request for help, this window will display appropriate help screens. Similarly, the response to a 'why' request will be displayed in this window.

User control of the system is effected by means of menus allowing the selection of different processes at different stages in the consultation. Initial menus allow the user to select from a number of options including viewing introductory help pages, entering new data, editing existing data files and consulting the system. Duŗing the consultation the user is presented with menus which allow them to respond to a question in a number of ways including 'why' and 'help' and 'unknown' if they are unable to give a true or false response.

Figure 2 : Example of Split-screen Presentation



```
Backward chaining - trying to verify      T_TEST
Possible tests -
T_TEST
ASPIN_WELCH
WILCOXON
```

```
The samples have equal variances.
Is this statement -   (T)rue      Other options -    (W)hy
                      (F)alse                        (H)elp
                      (U)nknown

Choice : :
```

7. Facilities Offered by THESEUS

The 'help' facility is available to explain statistical terms and is designed primarily for use by the research worker who does not fully understand the question being asked. The 'why' facility is available to inform the user which method the system is trying to verify and the fact it is trying to establish.

One of the important features of THESEUS is that it should allow the statisticians in industry to 'tune' the system to their particular requirements; this is to be achieved at two different levels. The first and simplest level is to allow the statistician to change those values used to enable the system to decide whether a fact is true or not, for example the level of significance used in a test for normality. The second level is to allow the statistician to modify the rule-base by altering, adding or deleting rules.This will work by keeping a 'default' rule-base completely separate and allowing the statistician to edit a'duplicate version and so create his or her own rule-base. The rule-base processor will use the modified rule-base but give the user, at the end of a consultation, the option to see the decisions that would have been reached by the default rule-base without having to go through the consultation process again.

In order to assist the statistician in editing and debugging the rule-base there is a trace facility which, rather than giving the straight forward list of accessed rules which is usually supplied by an expert system, gives more detailed information on the progress of the system. Information is supplied on the methods being considered, intermediate goals used by the system in backward chaining, rules being tried and actions undertaken. This trace facility is available at any time of the consultation so that the effects of changes to the rule-base can be monitored more easily.

8. Conclusion

The system outlined above is currently under development at Thames Polytechnic and involves a number of separate projects. To date the rule-base editor and rule-base processor have been written and the data entry and routine interface are currently being developed. It is expected that the prototype system will be ready by the end of 1987 for evaluation trials in industry. Concurrently with the development of the software, consultations with statisticians in industry will be continued and a first version of the default rule-base written and tested.

One aspect of the system design which is expected to be developed during the coming year is that of 'tunability' whereby expert users are able to tailor the system to suit their own requirements allowing them to impose their preferences, especially in areas where there is much debate as to the most appropriate technique.

References

Hahn, Gerald J. (1985), "More Intelligent Statistical Software and Statistical Expert Systems: Future Directions" The American Statistician, February 1985, Vol.39, No.1, pp 1-16.

Hand, David J. (1985a), "Statistical Expert Systems : Necessary Attributes" Journal of Applied Statistics, Vol. 12, No. 1, pp 19-27

Hand, David J. (1985b), "Choice of Statistical Technique" Bulletin of the International Statistical Institute, Proceedings of the 45th session, Vol. 3, Amsterdam, pp 21.1-1 to 21.1-16

Miller, R.G.,Jr (1981) "Simultaneous Statistical Inference" 2nd ed. Springer, New York

O'Neil, R. & Wetherill, G.B. (1971), "The Present State of Multiple Comparison Methods" Journal Royal Statistical Society Series B, Vol. 33, pp 218-250

Wetherill, G. Barrie (1985), "The Design and Evaluation of Statistical Software for Microcomputers" The Statistician Vol. 34, pp 391-427

THESEUS : An Expert Statistical Consultant

Edwina Bell and Peter Watts;
Thames Polytechnic, London

ABSTRACT

Two of the major problems in developing any expert system are the time and effort involved in knowledge elicitation and the less well documented area of software development. These two problems are particularly apparent in the area of statistical expert systems and a further complication is introduced by the need to use analysis packages or routines which are currently used by statisticians. Another problem is that practising statisticians tend to use those techniques with which they are familiar and to be reluctant to change their methods.

A statistical expert system which attempts to tackle these problems is under development at Thames Polytechnic.

Since the area of applied statistics is so vast, it was decided to concentrate effort on a limited area of statistical expertise; the area chosen was that of one-way analysis of variance and related techniques which our researches have shown to be heavily in demand in industry. Nevertheless it is expected that the principles established in this area will be applicable to larger scale systems.

The system has been written in Turbo Pascal and is designed to run on an IBM AT. To date a prototype system has been implemented and is running with an experimental rule-base.

THESEUS : An Expert Statistical Consultant

Edwina Bell and Peter Watts;
Thames Polytechnic, London

1. Introduction

Traditionally, statistical packages have been general purpose computer programs such as SPSS, SAS, GLIM, and GENSTAT. Whilst programs such as these have wide applicability and are often very powerful, there are a number of problems associated with their use; particularly those of checking the appropriateness of an analysis and lack of 'user-friendliness' in the package.

With the advent of Artificial Intelligence techniques, it is becoming possible to tackle these problems. Expert systems are designed to approach a problem in a user-friendly manner and to incorporate 'real-world' expertise in their structure.

It is becoming clear that incorporating expertise in a general statistical package is a very large problem and that the best way forward is the development of specialised intelligent software, Hahn(1985).

In this paper, we describe THESEUS, an Expert System being developed at Thames Polytechnic which concentrates on the area of one-way analysis of variance and related techniques, which our researches have shown to be heavily in demand in industry. It is anticipated that many of the lessons learnt in this area will be applicable to larger scale systems. Our primary aim is to develop a methodology appropriate to statistical expert systems and secondly to produce a prototype using this methodology.

THESEUS is being written in Turbo Pascal and it is envisaged that the final system will be available for use on an IBM AT micro-computer.

2. Problems of Statistical Expert Systems

Developing an expert system has its own difficulties, the major bottleneck being that of knowledge elicitation. The process of acquiring the knowledge appropriate to the area of application is often time-consuming and difficult and is further hampered, especially in statistics, by conflicts in expertise. Within a statistical expert system it is not only necessary to be able to update or alter any knowledge base relatively easily but also to allow expert statisticians to 'tune' the system to their own requirements. There is the additional problem that once knowledge has been incorporated into the system an extensive process of testing and verification needs to be carried out to ensure consistency and acceptability of the expertise.

The 'end-user' also needs to be considered carefully. There is a wide range of possibilities from the expert statistical consultant to the statistical novice and it would be difficult to cater for all of them in a single system. The statistically naive researcher would need extensive help and guidance to ensure the appropriate analysis is carried out and to interpret the results, whereas the expert may want to move through the system quickly, looking only at the results he or she is interested in.

THESEUS attempts to cater for this by having a wide range of help and explanation facilities available and a flexible mode of operation. Although the system will be able to cope with a range of users, it has been designed for use primarily by research workers in industry who are regular users of statistical analytical techniques.

Another consideration for a statistical expert system is the need to access statistical routines or packages in order to extract information from the raw data.

3. Design of THESEUS

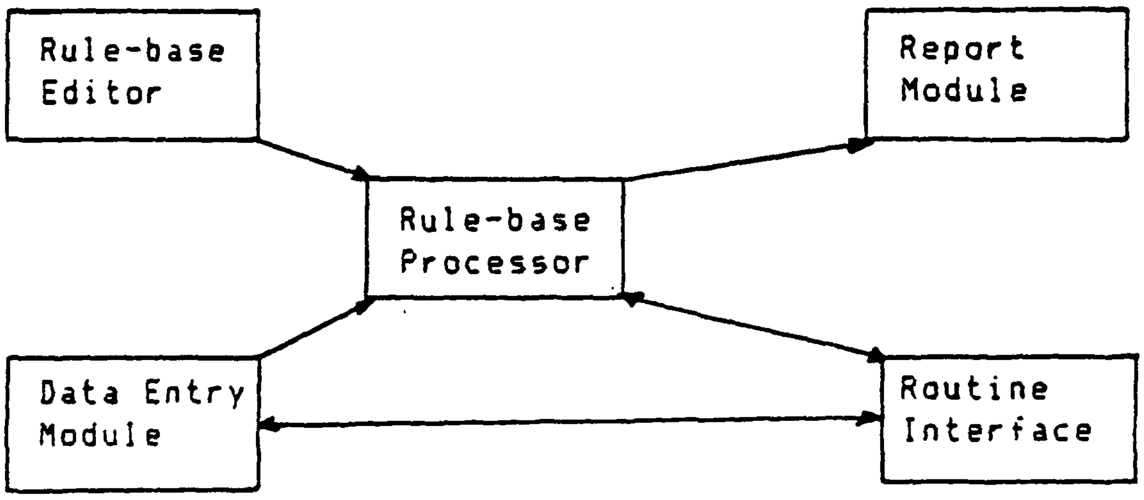
As in most diagnostic areas, a statistical consultant proceeds by initially generating plausible hypotheses about the analysis and then checking each hypothesis against the data to decide whether or not the analysis considered is appropriate, Hand(1985b). This leads to a natural structure for a statistical knowledge base.

THESEUS is basically a production rule system with rules in the form IF <condition> THEN <action> . The rules are processed in such a way as to reflect the decision making process of a consultant where possible. THESEUS works by initially identifying list of potential techniques, this is similar to the plausible hypothesis generation of the consultant. Once this list has been established the system will attempt to verify the validity and appropriateness of the technique in the list. If the knowledge base is sufficiently complete then the system will recommended a particular technique and also inform the user which of the remaining techniques are valid, allowing the user to select the technique they wish to employ.

4. Structure of THESEUS

The software is designed to be highly modular to aid development and maintenance as well as facilitating comprehension and flexibility. (See Fig.1)

Figure 1 : Structure of THESEUS



The rule base editor produces a file of rules which can be picked up by the expert system. The editor enables an expert user to enter, delete and modify rules and checks the rule base for consistency and redundancy.

The data entry module allows the system user to enter and edit data, performs basic descriptive analyses and conducts a dialogue with the user to ensure that both the system and the user are satisfied with the representation of the data structure.

The rule-base processor works through the rules using a combination of forward and backward chaining, accessing the routine interface and reporting intermediate results as appropriate. Information required by the system can come from a number of sources:

- asking the user questions
- backward chaining through the rules
- initial data entry section
- intermediate analyses of the data during a consultation

The structure of the rules and the rule-base processor are described in more detail in the next section.

The routine interface allows the system to perform statistical tests on the data by accessing libraries or packages, either during the consultation process or when carrying out recommended analyses .

The report module provides the results of analyses for the user and allows the user to structure output in an appropriate way, accessing intermediate results as necessary.

5. Rules and Inference

All rules in THESEUS have the same IF...THEN... construction but can be divided into three types :-

- Rules used to generate the initial list of possible techniques, these are processed by forward chaining (i.e. the system works sequentially through the rules)

- Rules which are used to establish the validity of a technique on the current data set, these are processed by backward chaining (i.e. the system sets up each technique as a goal and then trys to fire rules which would lead to establishing the goal)

- Rules which are used to decide, on the basis of information already available, whether to try transforming the data and so change the current data set, or whether to move on to the next technique in the list. These rules are processed by forward chaining and are denoted as META RULES as they govern the flow of control within the system.

Figure 2 : Control Structure of THESEUS

```
PROCESS                                 CONSEQUENCES
==========================================================================
Forward Chain                           Supplies a list of possible
                                        techniques

Set current data set to
Original Data

Point to first technique in list        Sets first technique as the
                                        current one

While more techniques in list Do
Begin

    Backward Chain                      Establishes whether or not the
                                        current technique is valid on the
                                        current data set

    Search Meta Rules                   Decides whether to move onto the
                                        next test (set next to true) or to
                                        try  a  transformation (set next to
                                        false and change current data set)

    If next is true
    Then point to next technique

End
```

6. Facilities Offered by THESEUS

The 'help' facility is available to explain statistical terms and is designed primarily for use by the research worker who does not fully understand the question being asked. The 'why' facility is available to inform the user which method the system is trying to verify and the fact it is trying to establish.

One of the important features of THESEUS is that it should allow the statisticians in industry to 'tune' the system to their particular requirements; this is to be achieved at two different levels. The first and simplest level is to allow the statistician to change those values used to enable the system to decide whether a fact is true or not, for example the level of significance used in a test for normality. The second level is to allow the statistician to modify the rule-base by altering, adding or deleting rules.This will work by keeping a 'default' rule-base completely separate and allowing the statistician to edit a duplicate version and so create his or her own rule-base. The rule-base processor will use the modified rule-base but give the user, at the end of a consultation, the option to see the decisions that would have been reached by the default rule-base without having to go through the consultation process again.

In order to assist the statistician in editing and debugging the rule-base there is a trace facility which, rather than giving the straight forward list of accessed rules which is usually supplied by an expert system, gives more detailed information on the progress of the system. Information is supplied on the methods being considered, intermediate goals used by the system in backward chaining, rules being tried and actions undertaken. This trace facility is available at any time of the consultation so that the effects of changes to the rule-base can be monitored more easily.

7. Conclusions

While THESEUS is still only a prototype under development, several conclusions of interest have already emerged from the research programme. The disparity, both between industrial statisticians and between industrial statisticians and academic theory has been highlighted. Leading directly from that, the importance of allowing expert statisticians to incorporate their own expertise into the system has been emphasised. It is thus essential that the knowledge is represented in a clear manner that can be easily understood by the statistician.

Our experience has shown that knowledge acquistion, even for such a limited area, is time consuming especially when the experts do not agree! Although a great deal of research has been undertaken in the area of knowledge elicitation this remains one of the major difficulties in building any expert system.

Finally we expect that the principles and structure embedded in THESEUS will be applicable to more general statistical expert systems.

References

Hahn, Gerald J. (1985), "More Intelligent Statistical Software and Statistical Expert Systems: Future Directions" The American Statistician, February 1985, Vol.39, No.1, pp 1-16.

Hand, David J. (1985a), "Statistical Expert Systems : Necessary Attributes" Journal of Applied Statistics, Vol. 12, No. 1, pp 19-27

Hand, David J. (1985b), "Choice of Statistical Technique" Bulletin of the International Statistical Institute, Proceedings of the 45th session, Vol. 3, Amsterdam, pp 21.1-1 to 21.1-16

Wetherill, G. Barrie (1985), "The Design and Evaluation of Statistical Software for Microcomputers" The Statistician Vol. 34, pp 391-427

Knowledge acquisition in the development of Theseus, a statistical expert system

E. Bell, P.J. Watts (Thames Polytechnic) & J. Alexander (Hazleton UK)

1. Introduction

It is becoming clear that incorporating expertise in a general statistical package is a very large problem and that the best way forward is the development of specialised intelligent software (Hahn, 1985).

At Thames Polytechnic a research programme has been set up to develop a software framework (or shell) for building specialised statistical expert systems.

This system, Theseus (fig 1) differs from existing expert system shells in a number of ways. The system includes the facility to interrogate the data as well as the user in order to establish facts. Perhaps its most novel feature is the combination of a "tunability" concept which will enable a "local expert" to reset the rules to his own specification, and a development module which will allow the local expert to experiment with the system.

2. The local expert

The concept of the "local expert" is central in the design of Theseus.

It is probably true to say that most research workers have some kind of access to statistical support, but those of us who provide that support know just how limited the resource is. That is why many researchers are in practice obliged to fall back on their own limited knowledge, using (or misusing) a package, get their printout, and after a little star-gazing write their reports, without having their statistics monitored by anyone, or anything, with expertise in statistics. If the local statistician were able to make available to his clients a system which incorporates his or her own expertise in those, often quite straightforward, areas of application where there is insufficient time to monitor the work personally, there would be obvious benefits to both statistician and researcher (Fig 2). Such a system should relate to the types of study design and data encountered locally and should obviously reflect the statistician's own experience. Thus the concept of an expert system accessible to and tunable by the local expert.

3. Selection of an application area

After carrying out a postal survey of practising statisticians, we elected to make the analysis of completely randomised experiments with one trial factor our prototyping application for THESEUS. When the user first interacts with the system, the opening dialogue will establish that the problem in hand does in fact fall into this category, otherwise she or he will be referred on to a human expert.

The reason for selection of this application area are straightforward. Firstly there is a little point in producing a system for which there is no demand; we believe that in this area there is a demand, and that, moreover, researchers regularly analyse experiments of this type without statistical help. We are aware that many statisticians are uneasy about this situation in their own institutions, but have insufficient resources to be able to offer the help needed.

We also felt that the knowledge acquisition and construction of this system would be a manageable task, and that there would not be a heavy development overhead in producing the numerical processing routines. Lastly, most of the expertise used in analysing this simple type of study will readily extend to more complex designs in later developments of the system.

4. The knowledge acquisition process

We firstly set out to define the domain of expertise we are dealing with. After some debate we identified two classifications of the domain which are valuable in structuring the acquisition process. Firstly, the domain involves both technique selection expertise and technique application expertise (Hand, 1984).

Secondly, it involves both "technical" and "professional" knowledge areas. "Technical knowledge" is hard, factual knowledge obtainable from text books and the literature, "Professional knowledge" is judgemental, experience related and immensely more difficult to encompass. For our technical knowledge resource we believe that our academic base provided the best starting point. For the professional knowledge we knew that we must involve practising statisticians. We therefore set about a series of literature reviews and investigations and at the same time set up a series of semi-structured interviews with statisticians in industry and research institutions. A series of workshops in which alternative approaches to data sets are discussed in terms of production rules brings the two aspects together. From these three areas of activity the prototype rule-base is being formulated.

5. "Technical Knowledge" - acquisition

To pursue this subset of the knowledge domain, a number of review areas were defined and members of our group set out to perform extensive literature reviews and to initiate small-scale research projects to ensure that a rational set of rules could be formulated.

The review areas fall into four groups; firstly, those concerned with study objectives - determining the structure of the client's hypotheses, selecting appropriate error rates and the choice of suitable multiple comparison procedures; secondly, those concerned with the data itself - handling outliers, making transformations; thirdly, a review of test procedures - their comparative performance; and fourthly the validity of methods - notably in respect of appropriateness of method to scale of measurement, and the properties of homoscedasticity and normality.

6. "Professional Knowledge" - acquisition

It is in this area that we come up against all of the problems that are making knowledge acquisition a growing discipline in its own right. The problems of encapsulating the knowledge used by a statistician, say, viewing a normal plot and in his subsequent decisions presents difficulties of pattern classification that could well prove intractable were we to attempt to extract it and translate it into production rules.

Two alternatives to direct elicitation present themselves. They are both rule inductive. The first is to produce intelligent software that will learn directly from observing decisions: that is, the rule base would be self-adaptive. We felt that this was not only technically complex, but would be sensitive to the paramaterisation used for the decision process. The second approach (Mingers, 1986) is based on the use of case studies with appropriate statistical classification procedures used to determine an underlying rule structure. We did not believe, however, that we would be able to sustain the motivation of our busy professional collaborators to work through a substantial number of test data sets.

Instead of following either of these approaches we have developed an evolutionary view. We have conducted a series of interviews with practising statisticians. The purpose of these interviews was to give general insight into the kind of thinking that guides them rather than the precise elicitation of rules. Recognising that there is a considerable chance of leading experts into pre-conceived knowledge structures, we tried at all times to allow the expertise to flow unhindered. A loosely-structured interview protocol was prepared to ensure that coverage of the relevant knowledge areas was complete while allowing the contributors to describe fully, in their own ways, their approaches to data analysis. The interview schedule covered such areas as attitudes to outliers, rigidity/flexibility on normality assumptions and homoscedasticity, use of transformations and the selection of test procedures.

Selecting statisticians purposively from those who responded favourably in our initial postal survey of 50 statisticians, predominantly in the pharmaceutical and chemical industries and in research institutions, we carried out a series of seven such interviews. The information gathered demonstrates more than anything else the enormous variability between different statisticians handling similar types of study. One statistician, for example, never transforms data because, he says, the regulatory authority (in this case the FDA) to whom the results of studies are presented, do not approve of the practice. Another colleague, working in the same area with the same regulatory authority to satisfy, regularly uses square root and logarithm transformation, according to his own discretion.

There was distinct vagueness about multiple comparisons, with each statistician quoting his own favourite, but being unclear about its use in relation to his client's hypotheses. Not one used any tests for normality; some justified this on the basis of sample sizes. At least one used the same argument form not investigating varying homogeneity. A feature which came through very markedly was the decision to keep everything as simple as possible in the interests of their clients' understanding.

From these interviews we are however obtaining sufficient data on the basis of decision commonly used by statisticians. This provides us with a starting point for a prototype default rule-base. Rather than trying to provide a faithful representation of the knowledge base at this point, we will provide our panel of collaborators with a prototype utilising our default rule-base, and with the aid of the development module and rule base processor, they will, in an operational setting, be able to modify and tune the rule base to their own taste. This will then form the basis for a further development cycle.

7. Development

The selected strategy for developing the knowledge base can thus be summarised as follows:

1. Members of the research group will contribute to a default rule-base utilising the fruits of their academic investigation and the interview findings.

2. An ongoing series of workshops provides opportunities for the group to evaluate and develop this rule-base.

3. The emergent rules will be entered into the shell to provide a prototype system.

4. The prototype system will be distributed to test sites where the collaborating statistician will be asked to

   (a) use the default rule base with his own and simulated data.

(b) encourage selected clients to use the system with its default rule-base.

(c) use the development module to try out modifications to the rule-base and compare its performance with that of the default set on real and simulated data.

(d) provide feedback in the form of suggestions and sending back his or her modified rule-base.

5. The feedback will be reviewed and the prototype modified as appropriate both structurally, and in terms of its production rules. (Fig 3)

It is recognised that we are expecting a great deal of input from our collaborators, but we believe that in a production situation they will be both motivated and able to contribute. Of course it is likely that in using the system the statistician will gain both in experience and technical awareness. If an expert system designed for the naive user can also find a role in training and extending the skills of the professional statistician then this is an opportunity we can ill afford to miss.

Building a Statistical Knowledge Base: A Discussion of the Approach Used in the Development of THESEUS, a Statistical Expert System

E. Bell and P. Watts, London

Keywords : Expert Systems, Statistical Analysis, ANOVA, Multiple Comparisons.

Introduction

Knowledge acquisition is one of the major problems in developing any expert system. It _s recognised that different methods of knowledge acquisition should be used for eliciting different types of knowledge and that the usual approach of dialogue sessions between a domain expert and a knowledge engineer is not always appropriate.

Statistical expertise involves both technical expertise and professional or experiential expertise; professional expertise is very difficult to encompass and varies both within and between application areas. It is because of this variability in professional expertise that the concept of a local expert who can modify or extend the knowledge base became central to the design of THESEUS, a statistical expert system under development at Thames Polytechnic which provides advice and guidance in the fields of ANOVA and Multiple Comparisons.

An initial knowledge base has been built up through a program of literature reviews and statistics workshops as well as drawing on information gained during interviews with practicing statisticians. This initial knowledge base has been supplied to a number of local experts for them to modify in a manner appropriate to their own operational setting, incorporating not only their own preferred testing procedures but also such constraints as customer preferences.

Compstat 1988
© Physica-Verlag Heidelberg for IASC
(International Association for Statistical
Computing), 1988

In this paper different methods of knowledge acquisition and the nature of statistical expertise are discussed leading to the approach used in the development of THESEUS.

The Nature of Statistical Expertise

Statistical expertise can be divided into several different types, the boundaries may not be particularly clear but such classifications can assist in the choice of knowledge elicitation technique, Gammack and Young (1985).

FRAMEWORK : A statistician will have some form of conceptual structure in the domain which distinguishes different types of analysis. This knowledge will be used to select areas of statistics appropriate to the data being considered. For example ANOVA and regression could be two such areas.

CONCEPTS : Such concepts as hypothesis testing, population distributions, confidence intervals and degrees of freedom are necessary foundations to understanding and undertaking any analysis within the area of ANOVA.

PROCEDURAL KNOWLEDGE : This will include knowledge about the availability and requirements of specific statistical methods for analysis and assumption checking as well as knowledge about graphical representations.

HEURISTICS : These include rules of thumb used for judging the importance of effects such as violation of assumptions and how to handle them, and information about when methods such as transformations are applicable.

METHODOLOGICAL EXPERTISE : This enables the statistician to choose the most appropriate method from a range of those available. For example, selecting Dunnett's method in preference to Tukey's method when there is a control group present.

COMMUNICATION : Surrounding these different types or areas of knowledge is the expertise used in communicating effectively with the client. This involves not just establishing what the experimenter is interested in finding out, but also extracting information about the nature of the data that the statistician needs to assist in making decisions about the most appropriate analysis. This may not be regarded as knowledge but is nevertheless included here because of the influence it should have in developing the knowledge base as well as in the design of the expert system.

Each of these areas of knowledge involves both 'technical' and 'professional' knowledge. 'Technical' knowledge is hard, factual knowledge obtainable from text books and the literature. 'Professional' knowledge is judgmental, experience related and considerably more difficult to elicit and represent, covering decisions such as when to allow unequal variances to affect subsequent decisions.

Much of the existing work in statistical expert systems has been undertaken by statisticians or at least by people with a basic grounding in statistics; à great deal of technical knowledge (which comprises a part of each of the different types of knowledge) can be established by literature reviews. The professional knowledge acquisition may be facilitated by the use of more specific knowledge acquisition techniques. Interviewing methods can be helpful in understanding framework knowledge, concepts and communication expertise.

Problems of Knowledge Acquisition

Knowledge acquisition for expert systems has, in the past, relied heavily on informal interviews between a knowledge engineer and a domain expert. The aim is to encode the information supplied by the domain expert into some predetermined format and so develop a prototype knowledge base. The knowledge base is then refined by a cyclic process of evaluation and modification.

The knowledge engineer, who has the problem of transferring the knowledge from the domain expert to the knowledge base, also has to ensure that an appropriate and powerful enough form of knowledge representation is used. A great deal of time can be wasted trying to manipulate knowledge in order to make it fit a particular representation; this is a well-known disadvantage of expert system shells. Domain experts often find it difficult to articulate their decision making processes and face further problems of recognition and interpretation when trying to understand and evaluate the performance of the knowledge base.

Expertise in any domain will contain different types of knowledge and thus the development of a knowledge base is the process of identifying the different types, choosing an appropriate knowledge representation scheme and then employing knowledge elicitation procedures appropriate to the situation.

Knowledge Elicitation Techniques

There are a number of methods available for aiding knowledge elicitation. INTERVIEWING METHODS are most helpful in the initial stages of knowledge acquisition for establishing the main concepts and components of the domain as well as defining the terminology used. The limitations of interviewing become more apparent when the domain expert is trying to evaluate the prototype knowledge base, trying to establish what distinguishes the performance of the expert from the performance of the system.

PROTOCOL ANALYSIS involves observing and recording the actions of the domain experts as they work through scenarios. The merit of this approach is that it gives the knowledge engineer a process to emulate. As the prototype knowledge base begins to take form, specific examples can be used to find out how the expert deals with special situations.

The basis of SCALING METHODS, which includes the repertory grid method, is to identify similarities among objects so that they can be grouped. Such methods result in values for a number of attributes used to define the objects. Cluster analysis of these attribute values enables discrimination between the objects. These methods are particularly useful where there is a number of closely related concepts, and expertise is required to discriminate between them.

CONCEPT SORTING is applicable when there is a large number of concepts within the domain. Concept sorting works by initially establishing a list of the concepts required to cover the domain and then asking the expert to sort the concepts into different groups, describing what each group has in common. This allows the concepts to be structured in an hierarchical fashion.

Protocol analysis is more appropriate for eliciting procedural knowledge and facts and heuristics, it may also be useful for understanding communication expertise. Concept sorting is really only appropriate for establishing the framework knowledge. Scaling methods could be useful for establishing concepts and understanding methodological expertise.

The Approach Used in THESEUS

As the application area chosen is small and well defined (Completely Randomised Designs and Multiple Comparisons) the knowledge acquisition does not involve the 'framework' knowledge described above but does involve all the other types. Each of the different types of knowledge involves both technical and professional expertise. Some types of knowledge such as procedural knowledge can be regarded as primarily technical in nature whereas knowledge about heuristics is mostly professional.

The knowledge acquisition for the prototype knowledge base of THESEUS was approached by using a combination of literature reviews, semi-structured interviews and workshops (protocol analysis). The aim was then to refine the prototype knowledge base with the help of practicing statisticians.

Literature reviews and small scale investigations were undertaken in order to establish a core of technical knowledge and ensure that a rational set of rules could be formulated. The review areas included hypotheses of interest to the client, choice of multiple comparison procedures, handling of outliers, use of transformations and criteria used for checking assumptions.

A series of interviews with practicing statisticians was undertaken with the purpose of gaining a general insight into the thinking that guides the statistician rather than the precise elicitation of rules. A loosely structured interview protocol was prepared to ensure that coverage of the relevant knowledge areas was complete, while allowing the contributors to describe fully, in their own ways, their approaches to data analysis. The interview schedule covered such areas as attitudes to outliers, flexibility on normality assumptions and homoscedasticity, use of transformations and the selection of test procedures. Seven such interviews were performed, the information gathered demonstrating more than anything else the enormous variability between statisticians handling similar types of study.

A series of statistical workshops was organised in which different approaches to the analysis of data sets, supplied two weeks in advance, were presented. This enabled close examination of the rationale behind decisions about the most appropriate way to undertake the analysis.

The expert system was sent to a number of test sites where the previously interviewed statisticians were asked to evaluate the prototype knowledge base and then tó try modifying the knowledge base.

Conclusions

An academic base provides a good starting point for developing a rational prototype knowledge base containing technical expertise and some professional expertise. This prototype knowledge base can then be evaluated and modified by 'local experts'.

This approach still requires a certain level of commitment from local experts but is far less time consuming than the conventional dialogue sessions. It also takes into account the variation in approaches both within and between application areas.

The industrial trials are still going on; preliminary results are encouraging with the local experts being able to identify areas where they disagree with the prototype knowledge base and suggest alternative approaches. As the trials progress any shortcomings detected in the technical knowledge in the knowledge base will be corrected. Modification of the knowledge base, by local experts, to include alternative approaches will be monitored to assess the ease with which the knowledge representation and inference process can be understood.

References

Bell E.E., Watts P.J. & Alexander J.R. (1987), 'THESEUS : An Expert Statistical Consultant' Proc. ICOSCO-I, Turkey

Brooking A.G. (1986), 'The analysis phase in development of knowledge based systems' A.I. & Statistics , ed W. Gale, Addison-Wesley

Gammack and Young (1985), 'Psychological techniques for eliciting expert knowledge' Research and Development in Expert Systems, ed M. Bramer, Cambridge University Press

Hand, D.J. (1985), 'Choice of statistical technique', Proc. ISI Centennial Meeting, Amsterdam