MICROPROCESSOR-ORIENTED EXPERT SYSTEMS FOR STATISTICAL ANALYSIS

Turkan K. Gardenier

RAS/OGC, EEOC, 1801 L St. NW
Washington, DC 20507, U.S.A.

ABSTRACT

This paper presents a road map of the decision making process of the simulator relative to the statistical significance of the parameters employed in the mathematical model. It defines strengths and weaknesses of a statistical expert system relative to the cost components involved in its use. It introduces a systems-analytic approach to the definition of parameters before submitting them to statistical scrutiny. A typology of statistical tests is provided. Some available expert systems for statistical applications are reviewed. Examples are given from the *Statistical Navigator*, a statistical expert system available on the market.

1 INTRODUCTION

Computerization of the decision process of a statistician is relatively new. Andrews et al. (1981) develop such a road map which followed the format of a flowchart and which was published in the form of a report. *Statistical Consultant* provides a recent computerization of the same report by Professor Sechrist of Indiana University of Pennsylvania. *Statistical Navigator* is another system offered for the mathematical modeler to assist the decision-making process involved in selecting the types of tests or hypotheses. The objective of the present paper is to (1) provide an overview of the context in which systems of this type can be used, (2) give a typology of commonly used statistical tests, (3) refer to parallel media of instruction, and (4) give examples of user interaction with particular reference to uses within the *Statistical Navigator*.

2 STRENGTHS AND WEAKNESSES OF STATISTICAL EXPERT SYSTEMS

An expert system in statistics provides a "road map" for conducting statistical analyses. The system needs to sensitize the non-statistician to the decisions which a statistician must make in:

- Choosing statistical tools;
- Applying statistical tests to already existing data; and
- Interpreting the results of statistical significance tests.

The strengths of using statistical expert systems may be summarized as follows:

1. Because the systems provide a "road map" of the decision process of a statistician, the number of alternative tests and assumptions can be quickly and easily reviewed. These systems present an overview of the types of statistical tests which are available for the data at hand.
2. If a computerized "expert system" is used as a training tool, it assists in familiarizing the trainee with the basics of computer usage. For example, if the system were menu driven, the knowledge acquired in this training could be transferred to other systems, such as used in word processing and graphics.
3. As the search among alternatives converges upon a best or near-to-best solution to the problem at hand, the user is made aware of the complex decisions that need to be made to analyze data and interpret the results of statistical tests. Knowledge of these decision points aids in future data collection and analysis tasks.
4. In addition to familiarizing users with statistical concepts and terms, a statistical expert system decreases the chance of making mistakes in analysis and interpretation of results. It assists in increasing the probability of using statistical methods properly.

The disadvantages or weaknesses of expert systems in statistics may be summarized as follows:

1. An "expert system" is no better and no worse than the human expert(s) who design it. The expert system may or may not be user-friendly. The expert system may lack in technical sophistication; or it may have too much technical detail and may threaten the novice.

2. Those being trained through the system may get the wrong impression that there is only one "correct" approach, as indicated by going along one path through the "branching tree." The trainees may get frustrated when they find out that (a) many alternatives may lead to the same solution, or (b) there may be more than one course of action for a specified set of circumstances.

3. Those being trained through the use of an "expert system" may feel that the system is merely leading them through an intellectual exercise not providing quick solutions or answers. Many prospective users may feel frustrated as they learn that the "press-a-button" approach to analysis does not apply to statistical expert systems.

3 SYSTEMS APPROACH TO A CASE IN CONTEXT

Figure 1 and Table 1 summarize how inflow and outflow characteristics interact with the definition of the criterion variable being studied. For the case at hand, the relative number of individuals referred or not referred to available jobs constitute the query. The final count may be a function of demographic characteristics, training and experience, and test scores of aptitude and interest. Factors such as job requirements, possible employer bias, and position availability also affect the observations.

We may also have "tree structure" oriented bipartite categorizations such as those described in Figure 2 and Table 2. In this case, the relative number and proportion of individuals in various classes or categories are being analyzed.

Both of these examples display the way in which data from surveys or ongoing records can be compiled or categorized for further statistical analysis and significance testing. An understanding of the context within which "variables" (i.e., inflow and outflow characteristics) interact among themselves and, in turn, affect the criterion measure, is essential in defining an appropriate mathematical model of the process.

4 TYPES OF STATISTICAL TESTS

A general overview of statistical tests which can be applied to the data at hand in order to determine whether or not the observed findings deviate significantly from what would be expected by chance is shown in Figure 3.

In summary, the methodological question may be one of determining the characteristics of

- One group of individuals on a single variable or on two or more variables;

- More than one group on a single variable.

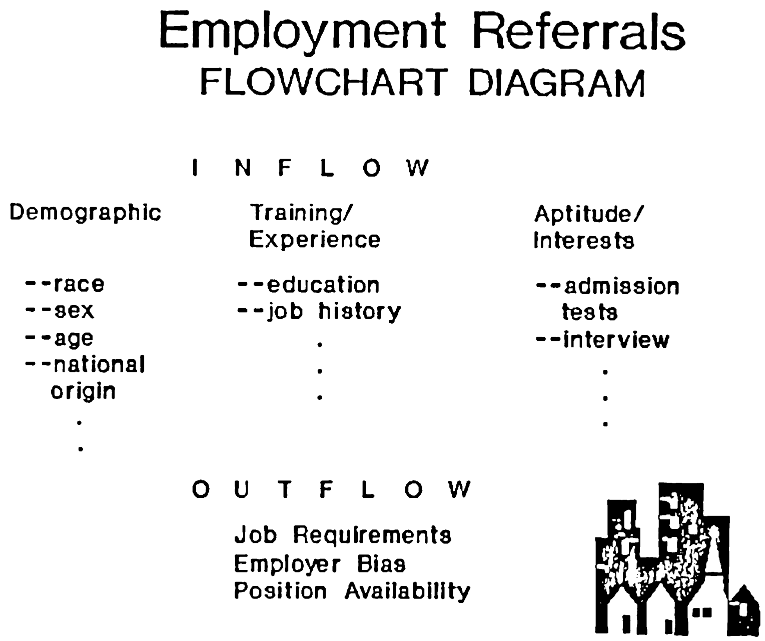
Exploratory data analysis and calculation of indices such as mean and standard deviation are appropriate for describing a specific group when the data are "continuous," i.e., they have not been categorized. For comparing different groups on a single criterion measure, t-tests and analysis of variance (ANOVA) are often used. t-tests are appropriate for the comparison of two groups, ANOVA for more than two groups. If the same group of individuals is being compared on a number of variables, then a "co-relational," or correlational approach is appropriate. Multivariate regression methods are used when more than two variables are being submitted to statistical analysis.

When the data being dealt with are categorized into counts or percentages, then tests such as the Binomial, Chi-square, or Fisher's exact probability test are the techniques used for statistical inference.

The above classification of statistical tests is appropriate for simulated data or data collected from field trials or surveys.

5 SOME GUIDES TO STATISTICAL EXPERT DECISION

Andrews, et al. (1981) in *A Guide for Selecting Statistical Techniques for Analyzing Social Science Data* have presented statistical techniques as a decision tree. Analysis methods are classified as to whether they are appropriate for one variable or two or more variables, the nature of scaling (for example, ranked versus continuous), and definitions of "dependent" or "independent" variables. Statistical terms such as "standard deviation" and "regression coefficient" are indexed by page number and by bibliographic reference. The names of statistical routines which can perform each of the types of tests are indexed by page



Employment Referrals
FLOWCHART DIAGRAM

INFLOW

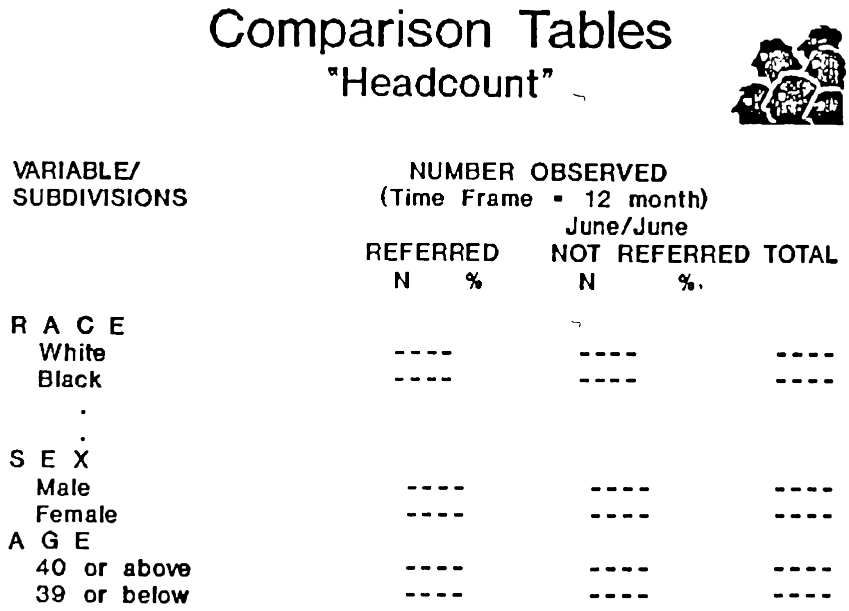
Demographic	Training/ Experience	Aptitude/ Interests
--race	--education	--admission tests
--sex	--job history	--interview
--age	.	.
--national origin	.	.
.	.	.
.		

OUTFLOW

Job Requirements
Employer Bias
Position Availability

Figure 1: Subcomponents of Em>loyee Referral Model

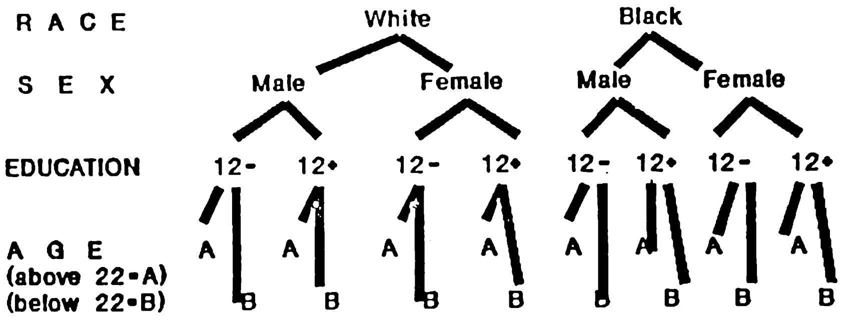
Table 1: Cross-Tabulation Translation of Subcomponents of Employee Referral Model Shown in Figure 1



Comparison Tables
"Headcount"

VARIABLE/ SUBDIVISIONS	NUMBER OBSERVED (Time Frame • 12 month) June/June				
	REFERRED		NOT REFERRED		TOTAL
	N	%	N	%.	
R A C E					
White	----		----		----
Black	----		----		----
.					
.					
S E X					
Male	----		----		----
Female	----		----		----
A G E					
40 or above	----		----		----
39 or below	----		----		----

TREE STRUCTURE
4 Variables



TOTAL = 2 X 2 X 2 X 2 = 16

NEED TO CONSTRUCT TABLE

Nodes=Cells

Figure 2: Tree-Diagram and Nodal Representation of a Four-Category Bipartite System

Table 2: Nested Cell Representation of Tree Structure Shown in Figure 2

CROSS-TABULATION
4 Variables

	RACE			
	White		Black	
EDUC	12-	12+	12-	12+
MALE				
AGE				
19-22	30	300	20	50
23-24	15	150	10	25
FEMALE				
19-22	30	300	30	75
23-24	15	150	15	50

SINGLE VARIABLE

DESCRIPTION OF GROUPS

(Confidence Intervals/ Inference

Continuous

- Mean
- Standard Deviation

"LATERAL SHIFT" TESTS

Subgroups: Mantel-Haenszel

Discrete
(Count or % of Group as Index)

- Single Side: 1: % - Binomial
  2: Compare 2 %
  large Samples: Normal/
  Z-test
- Both Sides of Coin:
  Chi-Square
  Fisher's Exact

  Large Samples:
  Approaches Normal

- Single: Confidence Intervals
  N over 30: Normal
  N under 30: t

- Two groups: t-tests
  with large Samples:
  Normal or Z-tests
  Query: Which is larger?
  "lateral shift"

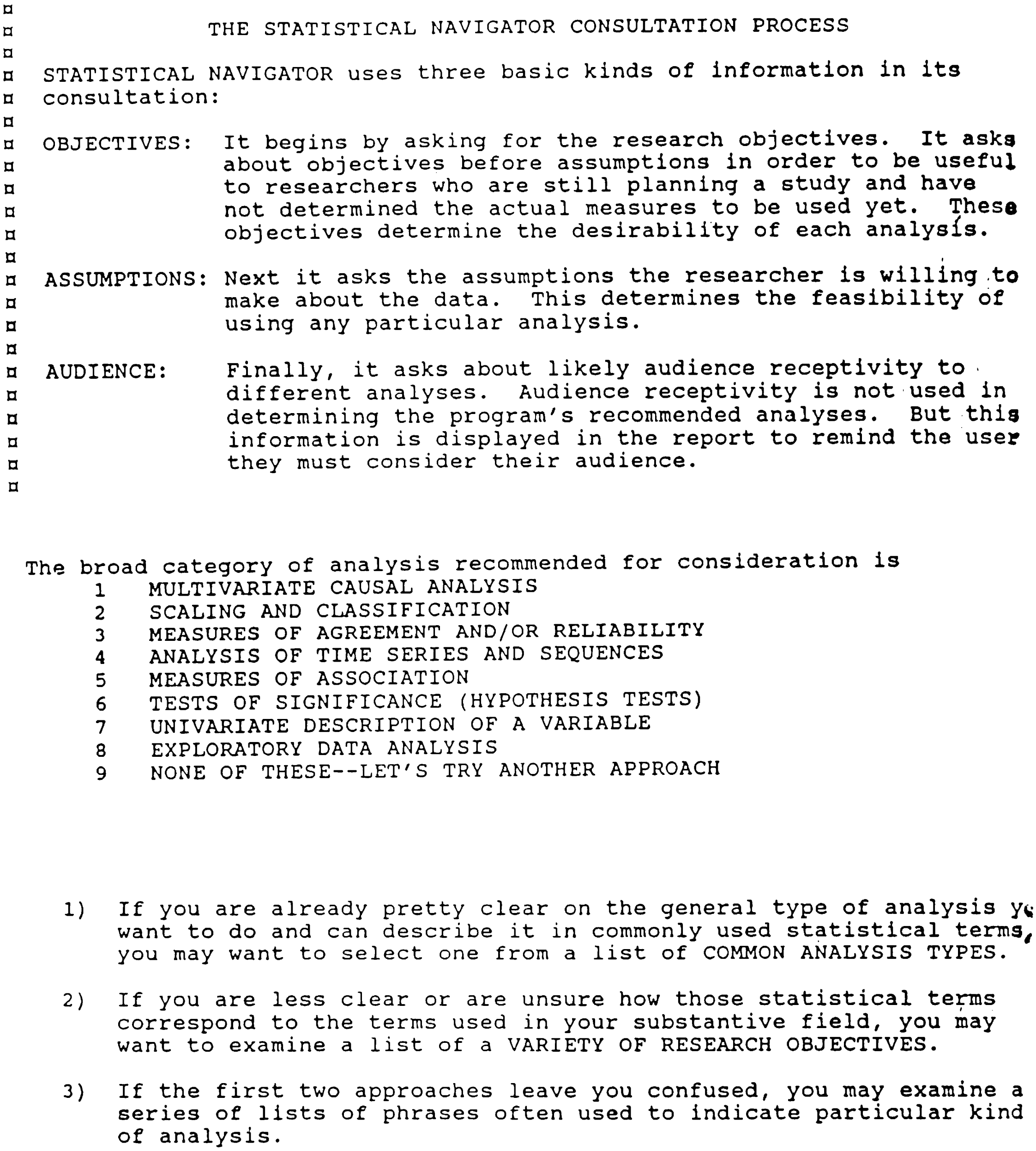
- More than 2 groups:
  Analysis of Variance
  (ANOVA)

TWO OR MORE VARIABLES
"CO-RELATION" (Correlation)

Two Variables	More than Two Variables
• Simple correlation	• Multiple correlation
• 2-variable regression	• Multiple regression

r = correlation coefficient
b = regression coefficient
	R = multiple correlation coefficient
	b1. . . . . bn

Figure 3: Typological Representation of Statistical Tests by Types of Variables Being Studied



```
¤
¤              THE STATISTICAL NAVIGATOR CONSULTATION PROCESS
¤
¤  STATISTICAL NAVIGATOR uses three basic kinds of information in its
¤  consultation:
¤
¤  OBJECTIVES:   It begins by asking for the research objectives.  It asks
¤                about objectives before assumptions in order to be useful
¤                to researchers who are still planning a study and have
¤                not determined the actual measures to be used yet.  These
¤                objectives determine the desirability of each analysis.
¤
¤  ASSUMPTIONS: Next it asks the assumptions the researcher is willing to
¤                make about the data.  This determines the feasibility of
¤                using any particular analysis.
¤
¤  AUDIENCE:     Finally, it asks about likely audience receptivity to
¤                different analyses.  Audience receptivity is not used in
¤                determining the program's recommended analyses.  But this
¤                information is displayed in the report to remind the user
¤                they must consider their audience.
¤
```

```
The broad category of analysis recommended for consideration is
     1    MULTIVARIATE CAUSAL ANALYSIS
     2    SCALING AND CLASSIFICATION
     3    MEASURES OF AGREEMENT AND/OR RELIABILITY
     4    ANALYSIS OF TIME SERIES AND SEQUENCES
     5    MEASURES OF ASSOCIATION
     6    TESTS OF SIGNIFICANCE (HYPOTHESIS TESTS)
     7    UNIVARIATE DESCRIPTION OF A VARIABLE
     8    EXPLORATORY DATA ANALYSIS
     9    NONE OF THESE--LET'S TRY ANOTHER APPROACH
```

```
1)  If you are already pretty clear on the general type of analysis yo
    want to do and can describe it in commonly used statistical terms,
    you may want to select one from a list of COMMON ANALYSIS TYPES.

2)  If you are less clear or are unsure how those statistical terms
    correspond to the terms used in your substantive field, you may
    want to examine a list of a VARIETY OF RESEARCH OBJECTIVES.

3)  If the first two approaches leave you confused, you may examine a
    series of lists of phrases often used to indicate particular kind
    of analysis.
```

Figure 4: Overview of User-Interactive Consultation Process Used in **Statistical Navigator**

number and by bibliographic reference. The names of statistical routines which can perform each of the types of tests are indexed as they appear in commercial packages such as SAS *(Statistical Analysis System)*, BMDP *(Biomedical Computer Programs)*, and SPSS *(Statistical Package for the Social Sciences)*. Professor Robert Sechrist of the Research Laboratory, Indiana University of Pennsylvania, has computerized this 68-page document. It is called *Statistical Consultant*.

The Future of Statistical Software: Proceedings of a Forum, prepared by The Committee on Applied and Theoretical Statistics of the National Research Council (1991), includes the following recommendations for the development of expert-systems-oriented approaches:

1. The desirability of a "branching process" for using and displaying statistical results. For example, all possible analysis options for contingency tables (e.g., Chi-square, Fisher's exact probability test, Cohen's Kappa, and Normal approximation to the Binomial) may not be appropriate to the specific query at hand.

2. Advantages and disadvantages of developing functional models based upon "serendipitous" data, defined as data which are not collected from sample surveys or designed experiments.

Statistical Navigator Professional, developed and distributed by Idea Works, Columbia, MO, is a software package with a user-friendly interface which uses "windows," "hypertext," and/or a "mouse." This package includes over 200 statistical analysis options. In the "consult" mode, the options in analysis can be matched to the problem at hand. The tests include measures of association (e.g., Fisher's exact probability test and Chi-square), and significance tests for continuous data (e.g., *t*-tests and analysis of variance). Definitions of statistical tests can be accessed through the "browse" mode using "hypertext." A report is produced for the user identifying several types of analytic procedures for the problem at hand. Procedures are ranked by the extent to which they are appropriate to the specified scenario. An example of the type of output during the interactive process with the user is shown in Figure 4.

6 CONCLUSIONS AND RECOMMENDATIONS

In summary, statistical expert systems can be very useful in portraying the wide array of techniques available to the statistical researcher and to the simulation analyst. These systems need to be supplemented, however, with other teaching tools. For example, *Against All Odds*, developed and distributed by the Annenberg Foundation, consists of 26 units; each is 20–25 minutes in length, and covers illustrative examples of where each of the types of tests shown in Table 3 would be applicable. Hands-on experience on analysis-oriented statistical software is also essential, preferably employing the user's data during interactive analysis.

ACKNOWLEDGMENTS

The facts and opinions presented in this paper are the sole responsibility of the author and do not represent policies and procedures of the United States Government.

The author wishes to thank Joseph R. Donovan for the guidance and editorial comments during the development of the ideas presented in this report.

REFERENCES

Andrews, F. M., L. Klem, T. N. Davison, P. M. O'Malley, and L. R. Willard. 1981. *A Guide for Selecting Statistical Techniques for Analyzing Social Science Data*, 2d ed. Ann Arbor: University of Michigan.

Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences, National Research Council. 1991. *The Future of Statistical Software*. Washington, DC: National Academy Press.

Statistical Navigator. 1989. Columbia, MO: Idea Works, Inc.

AUTHOR BIOGRAPHY

TURKAN K. GARDENIER is Mathematical Statistician, Research and Analytical Services, Office of General Counsel, U. S. Equal Employment Opportunity Commission, Washington, DC. She received her A.B. at Vassar College and her M.A. and Ph.D. at Columbia University. She has taught at Columbia University, George Washington University, and American University, and was Chairman of the Industrial Engineering Department at Middle East Technical University. She served as Associate Editor for Health Statistics for *Simulation*. Her research interests include optimization techniques for multivariate analysis through the use of statistical experiment design based preprocessors.