THE UNIVERSITY OF MANITOBA

PRESENTATION OF A STATISTICAL ON-LINE (SOL) COMPUTER SYSTEM AND AN EVALUATION OF OTHER SYSTEMS

by

ROBERT IAN ROLLWAGEN

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

MASTER OF SCIENCE

DEPARTMENT OF STATISTICS

WINNIPEG, MANITOBA

May, 1974

PRESENTATION OF A STATISTICAL ON-LINE (SOL) COMPUTER SYSTEM AND AN EVALUATION OF OTHER SYSTEMS

by

ROBERT IAN ROLLWAGEN

A dissertation submitted to the Faculty of Graduate Studies of the University of Manitoba in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

ABSTRACT

Statistical computer systems from two major areas of computer processing, batch and on-line, are evaluated. Comprehensive guidelines are developed and used in the evaluation of fifteen statistical systems, namely, ASCOP, OMNITAB, P-STAT, SAS, SPSS, BMD, STAT-PACK (GODDARD), STATISTICAL PACKAGE (MANITOBA), IMPRESS, MIDAS, ISIS, RAX, SSIPP, STATPACK2 - APL, and SOL. Other systems are also evaluated, but in less detail than the above fifteen. The SOL system is evaluated in more detail as well as including its rationale for development and its impact on statistical methods within the author's environment.

ACKNOWLEDGEMENTS

I would like to thank Dr. Bruce Johnston, Dr. Ted Bentley, and Dr. Michael Saunders for their time and help. Their comments and criticisms have enhanced the quality of this thesis.

The typing of this dissertation was done by Mrs. Lynn Wilson. Thank you Lynn.

Thanks are also due to Dr. Glen Atkinson, who was involved in the initial stages of this thesis, and to my colleagues in the Computer Department for their comments.

I would also like to give a special thanks to the developers of the statistical systems for providing the user's manuals or equivalent documentation on which the evaluations are primarily based.

Robert I. Rollwagen

PRESENTATION OF A STATISTICAL ON-LINE (SOL) COMPUTER SYSTEM AND AN EVALUATION OF OTHER SYSTEMS

TABLE OF CONTENTS

Chapter

CHAPTER 1

INTRODUCTION

Most of the technical tools of the future statistician will bear the stamp of computer manufacture, and will be used in a computer. We will be remiss in our duty to our students, if we do not see that they learn to use the computer more easily, flexibly, and thoroughly than we ever have; we will be remiss in our duties to ourselves, if we do not try to improve and broaden our own uses.

John Tukey, 1965

1.1 Computers and Statistics

The success of statistical methods today in providing quantitative measures that assist researchers in explaining 'phenomenological processes' is attributable as much to the tools as to the theory. Without the desk calculator and computer to bridge the gap between theory and practice, statistics would never have made the great strides it has. Nowadays, for many, the process of data analysis using both the computer and statistics provides stimulation, excitement and an impetus to learning more statistics.

Statistics and computers together seem to be another case where the whole is greater than the sum of its parts. Carrying it a little further, a large factor analysis problem would not be attempted without a computer and a simulation study would be near impossible. One tends to agree with Yates (1966), who called computers "the second revolution in statistics" (the first revolution being desk calculators) and Muller (1970), who made the cogent analogy of attributing the qualities of both the microscope and telescope to the computer.

Although the power of computers could not be disputed they were difficult to use because they had to be programmed. Learning to program computers was not the answer for many who had neither the time nor the inclination. Also, the process of writing and debugging programs was a time-consuming task, to say the least, even for those who knew how to program. There existed a need for a set of computer programs that would cover many areas of statistics, as well as manipulating the data (data management) before the actual statistics were performed. Thus, since the early sixties, those involved in data analysis have been trying to combine statistical methodology and computers into statistical computer systems (hereafter called statistical systems or

systems). These systems would enable the non-programmer to use the computer easily for statistical analyses. Designers' goals have been to provide powerful statistical and data management procedures which are efficient and easy to use. Many of these goals are now being reached, but as we will see, however, there is no panacean statistical system.

Statistical systems have their share of disadvantages as they become more powerful and easier to use. The usage of statistical methods in data analysis is growing exponentially, but the production of garbage is not lagging far behind either. It is easy to produce invalid statistical analyses unknowingly or otherwise. Forcing data through a simple factorial analysis when a hierarchical design is appropriate, or performing a factor analysis because it is the 'in thing' is not uncommon of users. Van Reeken (1971) gives a very good example about "someone [who] computed 5000 tests of significance at a 95 percent level and wrote a thesis about the 100 cases that were significant instead of a thesis on why he found only 100 significant." Statistical systems are misused and the best answer to rectify the situation is, of course, more knowledge in statistical concepts imparted through the formal statistical lectures - but it is not the

whole answer. *If the term 'statistical analysis' is to retain its integrity, then it is the statistical community who must get more involved in acquiring and building statistical systems and educating the user on their best use.*

1.2 Early Developments

The number of statistical systems developed as compared to the number that have been documented in statistical journals seems out of proportion. It appears that since the early sixties, when the development of statistical systems essentially began, statistical journals have accepted less than a dozen articles on statistical computer systems. Papers dealing with computers or statistical computing are more plentiful, but, unfortunately, do not deal with specific systems.

Development of statistical computer programs has been chaotic. Wilkinson's comment (Wilkinson, 1969) is appropriate: "it is a sad indictment....[that there are] one thousand or more multiple regression programs in existence, plus a like number for analysis of variance." The development of statistical systems is not quite as gross, but dissemination

of information regarding them is certainly inadequate. Chambers (1967) and Muller (1970) help to alleviate the above situation by listing some of the available systems. The systems are summarized in Table 1.2 to illustrate what some of the earlier systems were. Then, as now, where statistical systems belonged was confusing. This is borne out by Cooper (1969a) in his historical paper, "Statistical Computing - Past, Present, and the Future", where he disposes of statistical systems by saying, "No attempt is made to survey the literature, and packages or systems are not referred to specifically."

There seemed to be a consensus on at least one point in the early sixties in that there existed a need for a statistical computer language or equivalent, so that researchers did not have to waste their time learning 'programming'. Even before this, however, there appeared ahead of its time a system called AUTOSTAT (Douglas and Mitchell, 1960) that recognized the above need. It contained features that could edit, select, tabulate, as well as performing statistical analyses and permitting variables to be labeled.

In 1959 the BMD - Biomedical Computer Programs package appeared. This is a collection of Fortran programs which has become very popular. The BMD concept has not

TABLE 1.2 - LIST OF USER'S GUIDES

REFERENCE	SYSTEM	AUTHOR(S)	DATE	PLACE OR COMPANY
Chambers (1967)	TISER - Fortran routines for time series, linear regression, random numbers	Beaton	1964	Bell Telephone Laboratories, Murray Hill, New Jersey
"	BOMM - time series analysis routines	Bullard et al.	1964	La Jolla
" *	P-STAT			
"	AARDVARK - generalized analysis of variance	Hemmerle	1964	Iowa State University
"	TARSIER - non-linear regression by Hartley's method	Hemmerle	1966	Iowa State University
"	ZORILLA - quadratic programming with linear constraints	Hemmerle	1966	Iowa State University
"	MULTIVARIATE STATISTICAL ANALYZER	Jones	1965	Harvard University
" *	GENSTAT IV - general statistical program	Nelder	1966	Australia: Waite Institute
" *	STORM - statistical oriented matrix program	Pomper	1963	New York: IBM
" *	BMD - Biomedical Computer Programs	Dixon	1964	UCLA
Muller (1970)	ISSS - Information Selection of Sampling Systems	Fan and Shceidemantle	1964	IBM

(* means the system also appears in Muller's paper)

TABLE 1.2 (cont.)

REFERENCE	SYSTEM	AUTHOR(S)	DATE	PLACE OR COMPANY
Muller (1970)	MEDCOMP - Medical Statistical Computer Programs	Sterling and Pollack	1964	Cincinnati University
"	MSP - Multivariate Statistical Programs	Clyde et al.	1966	Miami University
"	OMNITAB - Computer Program for Statistical and Numerical Analysis	Hilsenwrath et al.	1966	National Bureau of Standards
"	SSP - Scientific Subroutine Package		1967	IBM
"	SSUPAC - Computer Programs for Statistical Analysis	Dodson	1964	University of Illinois
"	STATPAC - A Bio-statistical Programming Package	Shannon and Henschke	1967	Goodard Computer Science Institute
"	UNIVAC - Univac Statistical Programs		1965	Sperry Rand Corporation
"	STATJOB	Hutchins et al.	1966	University of Wisconsin
"	SUMX	Champomier	1964	University of California
"	TSAR - Tape Storage Retrieval System	Gabor and Carlitz	1965	Duke University
Wilkinson (1969)	SNAP - Computer Processor for Statistical Analysis	Godfrey	1968	Princeton University
"	STATIST - Conversational Statistics Package for Interactive Consoles	Claringbold	1968	CSIRO, Sidney

changed much and BMD is still a major system (Dixon, 1973). In 1963 R. Buhler of Princeton had a different view when he developed P-STAT. The emphasis was on one comprehensive program and no matter what analyses were performed the same data structure would be used. The P-STAT system worked with data files with the ability to use output from one procedure as input to another. P-STAT is now a major system (Buhler, 1973). Cameron and Hilsenwrath (1965) took an even different approach. They felt a need to provide a language where one would still program, but do so 'easily'. Their language was based on natural English commands. The data was stored in a 'worksheet' (a two-dimensional table) and by using the commands one could 'program' analyses, but in the style of a desk calculator. They called their system OMNITAB.

In 1967 four systems were presented in *Applied Statistics*: METO - Meteorological Office (Craddock and Freeman, 1967); ASCOP - A Statistical Computing Procedure (Cooper, 1967); SEP - Survey and Experiments Program (Gower, Simpson, and Martin, 1967); a conversational system (Colin, 1967). In Colin's system the statistics were limited and the on-line instructions were designed for a completely unskilled user. The instructions, thus, were verbose. The other three systems consist of one large program using procedures and commands that could manipulate the data as well as perform

statistical analyses. METO used numbers instead of words, and SEP was much too 'computer' oriented in its commands. ASCOP was the more user oriented and is discussed in Chapter 3.

Chambers (1967), in his review of statistical systems, discussed the design differences between BMD and P-STAT. It is at this time that another evolutionary stage was set down formally in the move from packages of programs (e.g., BMD) to one comprehensive program (e.g., P-STAT). Besides making references to METO, ASCOP, SEP, and Colin (see above), Chambers makes references to the user's guides in Table 1.2 but does not discuss them.

Cole and Campbell (1969) present a brief discussion of their conversational system STATCHAT which is implemented on the IBM 1620 computer. Although the discussion is too brief to assess the system, nevertheless, this is what is needed - a chance to formally present statistical systems. Chambers (1969) presents his system of 90 subroutines called FIT, which uses either of the following four criteria for fitting the data: linear least squares, non-linear least squares, maximum likelihood, and maximum log - likelihood.

It appeared as if a retrograde step was to be taken in the developemnt of statistical systems when Gower (1969) said, "that the future direction lies in improving the autocodes." This is another way of saying that a special statistical 'language' is needed. He feels the abundance of statistical systems has "demonstrated that existing languages have proved inadequate for statistical purposes," and that we will no longer need statistical packages; "they will be replaced by subroutine packages linked by programming at the autocode level itself." Gower then proceeds to outline a detailed autocode language which appears to be rather complicated. Muller (1969) while acknowledging the power of autocodes as described by Gower in some data analysis problems, makes an important point when he says, "People, given the choice, tend to choose statistical techniques or computing procedures that are easy to use rather than the best to use."

Both Wilkinson (1969) and Muller (1970) list, but do not discuss, statistical systems in their papers (see Table 1.2), except for the STATJOB system where Muller was a participant in its development.

Myers' (1969) survey of social sciences computing systems lists approximately twenty systems of which five could be

classified as general statistical systems. Schucany, Minton, and Shannon (1972) performed the first general survey. They mention thirty-seven systems including some from Table 1.2. Nobody, as yet, appears to have done a full scale evaluation, although Chambers (1967) does compare BMD and P-STAT philosophies on data handling, and Schucany et al. do a brief abstract on each package. There is little discussion of on-line systems (Schucany et al. only list four).

1.3 Sources of Information on Statistical Systems

At a recent conference, 'Computer Science and Statistics: Seventh Annual Symposium on the Interface' held October 18-19, 1973 at Iowa State University, the author asked participants involved in discussing the major statistical systems of today, "Is there a central body where information on statistical systems can be obtained?" The answer was, "No." Some felt, "That is why we have conferences." Many felt there should be a central body. Dissemination of information is poor, as was revealed in a discussion with one individual, who was unaware of a statistical system developed at his university.

The statistical journals have articles on computers and statistics, but as we have seen there is little on the formal presentation of statistical systems and, when mentioned in historical papers, no detail is given as to what the system does, let alone an evaluation of them. *Applied Statistics* has a statistical computer algorithm section. Other types of literature contain information on statistical systems. Some of these are: computer journals and conference proceedings, the *Computer Programs Directory*, computer centre newsletters, journals and conference proceedings from various other fields (e.g., sociology).

The above sources, although not exhaustive, should have captured the major, viable systems and enough minor ones to present the 'state of the art' in statistical systems. From the plethora of statistical systems discovered, it was apparent that many were obsolete, and many that fitted the definition of a statistical system were too restrictive, or when compared to the more major systems did not offer anything extra of value as regards to its present day usefulness or to its promoting ideas of better statistical systems. A user's manual or equivalent documentation was requested from only those institutions whose statistical system appeared to be viable or offered new concepts as regards to future systems.

CHAPTER 2

EVALUATION GUIDELINES

By what criteria can a statistical system be evaluated, and how much weight should each criterion have? In order to answer that question we must first ask, "What do statistical systems provide today?" Most of the guidelines appear by gleaning the pertinent information from the available systems, and the rest are points that one would expect from any system, statistical or otherwise. Thus, the guidelines as drawn up are such that all of the statistical systems contain at least some of the points mentioned in the guidelines, but no one system has them all. What future statistical systems 'might' contain is an important consideration, but is not discussed in this thesis.

Evaluate means 'to examine and judge'. The word 'judge' rings cold - like a teacher marking exams - this one passes, this one fails. In this thesis the word 'judge' should connote the most liberal interpretation possible. Weak areas are mentioned, and some systems are castigated in specific spots.

but no system fails. Each has made a contribution to this rapidly evolving area of statistics. It should be mentioned that the idea did occur to allot points to various criteria and then rank the systems. A few moments of reflection quickly dispelled the idea for a host of reasons. Statistical systems are very complex. There are many subtleties, and procedures are so interwoven that a rigid assigning of ranks could easily be misleading - if not destroying the intangibles. In fact, one of the most important points is, "Is the system successful in its present environment?" But, finding that out for each system is impractical and even if it were true, "How do you rank it?"

The evaluations are based on the point of view of the user. He should not be required to know computer languages (e.g., FORTRAN, ALGOL, APL, etc.) nor for that matter, understand what a computer is. As far as he is concerned, the computer is a black box, and if he has to understand certain aspects of its operaion, then the necessary information should be provided in lay terms in a user's manual or equivalent documentation.

The only problem with evaluating systems from a user's manual is that you do not know if the system works the way

it is supposed to work. Getting and implementing each system is impractical - if not impossible. Some practical experience has been gained by the author with a few of the systems and, with few minor exceptions, all the systems perform as stated in the user's manual.

2.1 Statistical Analyses

The areas of statistical analyses for this evaluation are divided into eight areas: descriptive, frequency analyses, regression and correlation, multivariate analyses, analysis of variance, special two-sample tests, distribution theory, and miscellaneous.

Descriptive statistics should include measures of central tendency, viz., mean, median, and mode; dispersion measures, viz., standard deviation, mean deviation, and range; other measures, viz., skewness, kurtosis, lowest and highest values, and percentiles. Facilities should exist for comprehensive frequency distributions and histograms. N-way tables (e.g., sex by age by...) that include at least the means and standard deviations are commonly needed.

Facilities to construct N-way frequency tables are necessary.

The chi-square statistic should be available for two-way tables including options for displaying expected values. For 2x2 tables one should be able to perform Yate's correction and for small N perform Fisher's exact test. Other tests of association for two-way tables are not uncommon. Options should exist to provide percentages within the cells, as well as for rows and columns.

Simple regression and correlations should be possible, with the facility to generate variance - covariance and correlation matrices. The simple regression should provide, besides the basic estimates and tests of significance, options to display predicted, adjusted, and residual values, as well as confidence limits. Tests for examining residuals are not uncommon nor are means to provide partial correlation coefficients. Non-parametric correlations should be possible, e.g., Spearman and Kendall. Multiple regression with a stepwise mode should be available with good options for the selection and deletion of independent variables. Polynomial regression is commonly needed. Non-linear and weighted regressions are provided by some systems. Bivariate plots (scattergrams) are necessary and other types of graphs are useful.

Factor analysis should be available with options for the type of factoring and types of rotation schemes to be employed. Discriminant analysis in a stepwise mode is necessary. Canonical correlations and multivariate analysis of variance are sometimes needed.

Univariate analysis of variance and covariance procedures should be comprehensive, i.e., be able to print means, standard deviations, residuals, etc. Hierarchical designs should be possible, as well as methods for handling unbalanced designs. Multiple range tests are necessary and non-parametric analyses for one and two-way designs are not uncommon.

Two sample tests are extremely popular and should be provided. These include t-tests for paired and unpaired data, as well as an approximate test when the variances are unequal. A test of homogeneity of variances is necessary. The non-parametric counterparts of the t-tests should be available such as Wilcoxon's paired and unpaired tests, Kolmogorov-Smirnov test, and the sign test, etc.

The ability to test data for goodness of fit to popular distributions is common and useful. Generation of random observations from the common distributions is available in

some systems, as well as the generation of probability density function and cumulative distribution function values.

The above areas seem to be the common ground for general statistical procedures used by the statistical community. As will be seen, some systems cover certain areas comprehensively, while others may not even include certain areas. Some systems have specialized areas, e.g., time series analysis, scalogram analysis, bio-assay analyses, etc., which are noted in the evaluations, but not discussed.

2.2 Data Management

Data management is broken down into five areas: input, output, storage and retrieval of the data, manipulation of the data, and limitations on the data.

The requirements for the structure and input of the data matrix should be the same for all procedures or programs. In general, the best structure to use is the 'variable by case structure'. Thus, each case (synonyms are observation, record, or unit) consists of a number of variables (e.g., identification number, name, age, sex, height, blood pressure, etc.). Each case consists of one or more 80 column data cards

depending on the number of variables used. The variable by case structure is standard and the best way to handle large data bases. Exceptions to this structure would be valid when small sets of data are used, or when procedures are unable to handle missing values or when the design of the system was to facilitate ease of use. Input should be possible from punched cards, tape, disk, or other remote access device. Alphanumeric data, which consists of at least one non-numeric character, as well as numeric data should be possible. Flexible input formats for describing the data is necessary. Free-field, where the data does not have to be in specified columns, is necessary for on-line systems and is a useful option for batch systems. The process of submitting the job for processing should require a minimum of job control language (JCL). IBM 029 keypunches are not standard and means should be provided so that 026 keypunches can be used.

The data matrices should be capable of being named and stored on tape, disk, or other device, as well as being retrieved easily. It should be possible to label variables with an acronym and descriptor (e.g., YRBORN is an acronym and YEARBORN could be its descriptor). Labeling the values a variable takes on should be permitted (e.g., for the variable

SEX it could be: 1 = MALE, 2 = FEMALE). The labeling or numbering of cases is also desirable. The generation, storage, and retrieval of more than one data matrix should be permissible.

File utility procedures such as merging, sorting, concatenating, and forming subsets are necessary. For special files such as correlation matrices, mathematical routines might be desirable (e.g., matrix inversion). Output from statistical procedures should be capable of being stored and used as input to other procedures, e.g., saving and analyzing residuals. Also, files should be readable by other computer and statistical systems.

Program control and data modification commands are needed. Some of these are: relational operators (e.g., greater than, less than, Boolean expressions, etc.); transformations (e.g., square roots, addition of variables, etc.); conditional control (e.g., IF...THEN...); branching (e.g., GO TO...). Very often the same set of modification commands are needed; the only difference being that perhaps a different variable is being used. It would be efficient and convenient if the user did not have to repeat the same set of statements, but yet use the commands already set down. Such facilities

exist and they are called MACROS. A MACRO is analagous to a Fortran subroutine.

Extremely useful are the special data analysis commands. Thus, procedures should exist for recoding, selecting, and deleting either cases or variables. Also, the weighting of cases and the provision to take random samples is desirable. The data analysis and data modification commands should act on a temporary or a permanent basis depending on what the user desires.

For those systems that use the variable by case structure, it is mandatory that provisions should exist for designating and handling missing values. Handling of multiple measurements on a variable within a case is a desirable feature.

Output should be adequately labeled, as well as permitting the user to add his own labeling, if desired. One should have control over procedures that produce voluminous output. For on-line systems this is particulary important, as well as catering to the various teletype widths, e.g., ASR's take 70 columns, IBM 2741's take 120 columns, etc. Also, for on-line systems, teletype printing speeds of up to 30 characters per second should be possible.

Limitations should be realistic. For batch systems and on-line systems that use data files the handling of at least 500 variables and 5000 cases is not unreasonable. The user, however, should not be penalized if the amount of data is small. He should be able to adjust the computer core size to fit his particular problem.

2.3 Ease of Use

Unless the statistical and data management procedures are easy to learn and use, only a handful will benefit. Those few who know the intricacies become the 'experts' and spend large portions of their time instructing users on trivia; a service that a good user's manual should provide. What is lamentable is that it is human nature to take the easy way out and use procedures that are imperfect, often inaccurate, and time-consuming - just so long as it is easy to use. The time and organization required to write an easy to understand user's manual, and easy to use system are reasons why it is often neglected, but these reasons are not a justification.

Ease of use of the system and ease of understanding the user's manual are for all intents and purposes the same thing. The better manuals have most of the following

basics: a) there is explicitness with regards to input, output, and limitations. The output is rigorously defined, i.e., there is no doubt as to what calculations have taken place; b) examples are meaningful, as well as abundant; c) there is as much independence between procedures as possible; d) references are provided; and e) the previous points are presented in a readable style.

The rigorous defining of output is often neglected or done poorly. A recent paper by Francis (1973) shows the confusion that can reign when output is not defined. Four analysis of variance programs (ANOVA procedure from SAS, BMDX64 program from BMD, MANOVA program from OSIRIS, and a program from the University of North Carolina; CAROLINA) were run using data from a standard 2x5 factorial design, but with unequal numbers in the cells. Because the design is non-orthogonal, the sums of squares for the various sources depend on the order of selection; but the answers produced are different because each program enters the source of variation differently, and the ANOVA procedure from SAS is invalid although the user's manual by Barr and Goodnight (1972) does say what is done. BMDX64 is recommended by Francis, but, for many, may be difficult to learn and use. The confusion results because the user's manuals did not

make it clear how the sources of variation were entered, or even, to quote Francis, "that the order might be an important matter."

Ease of use is facilitated by using commands that are in natural English, with ability to abbreviate for experienced users. The editing of commands should provide meaningful error messages and should not stop compiling after only one error. Procedures that edit data (check for bad punches, check sequencing, etc.) are helpful. On-line systems should never lose control and leave the user not knowing what has happened.

2.4 Other Considerations

Details of systems that do not fall into the previous three sections are included as Other Considerations, in addition to the following subsections: evolution, program documentation, transportability, accuracy, and teaching.

2.4.1 Evolution

If a statistical system is to remain viable, it needs to be revised and updated. No system supplies all the answers

and users would like to know the system they are using can be modified and improved, including the facility for the user to add his own routines.

2.4.2 Program Documentation

Well documented user's manuals are few. Well documented internal workings of computer programs are rare. A few reasons for well documented programs are: so those looking at the program at a later date will know the 'who, what, why, when, and where' of the program including the original programmer; so that those involved with implementing, or trying to find bugs, or modifying the system can do so with a minimum of effort; a history of the project is maintained. Good documentation does not require flowcharts or separate systems manuals. Copious comments within the actual program are usually sufficient. There are no excuses for not providing comments. Kreitzberg and Schneiderman (1972) provide some formal standards that should be adhered to.

2.4.3 Transportability

If statistical systems are to be run on other computers, they must be written in the more common languages. Within

the language as much standardization of coding should be used as possible. Some other factors that affect transportability are: the particular computer and software used and other special equipment used; availability; purchasing costs; and ease of implementation. Some other points that are important are the costs of running the system and the reliability of the computer system per se.

2.4.4 Accuracy

Define: $\bar{X} = \Sigma X/N$,

$$S_1 = \Sigma(X-\bar{X})^2 \text{ , and } S_2 = \Sigma X^2 - N\bar{X}^2 \text{ .}$$

The above formulas have been studied to determine their accuracy. Neely (1966) presents four ways of calculating the mean and six ways of calculating the standard deviation, as well as six ways for calculation of the correlation coefficient. He concluded the best algorithm for the mean as $\bar{X} + \Sigma(X-\bar{X})/N$ and S_1 as the best in computing sums of squares. Calculator algorithms like S_2 were the worst. Youngs and Cramer (1971) conclude that the best algorithm for the mean is the conventional $\bar{X}$ but suggest an iterative approach to the sums of squares thus avoiding two passes of the data.

Similarly, Longley (1967) and Wampler (1970) evaluated the accuracy of least squares regression programs, and found that programs that employed elimination algorithms were inaccurate but those using orthogonal Householder transformations and Gram-Schmidt orthogonalization were the most accurate on their ill-conditioned data.

What this suggests is that inaccuracies can occur in many cases (e.g., large numbers, many numbers, ill-conditioned matrices) when certain algorithms are used, and that efforts should be made by developers of statistical systems to use the most accurate numerical analysis methods. Many statistical systems do not use the most accurate algorithms, and users should be constantly on guard and examine output for inaccuracies.

2.4.5 Teaching

The highly specialized area of Computer Assisted Instruction (CAI) in statistics is not germane to this thesis, nor, for that matter, are statistical systems whose sole intention is to assist in the teaching of statistics. How statistical systems can be used in teaching - or even if they should be used - is polemic, and no attempt is made in this thesis to

resolve the issues. The attitude taken is if a system aids in the formal teaching process, 'well and good'; but it is a minor consideration in terms of this thesis framework only. For a survey of teaching statistics at the university level with computers see Evans (1973), and for information on how some specific statistical systems are used in teaching see Kennedy (1973).

CHAPTER 3

BATCH SYSTEMS

The precise definition of a batch computer system is not necessary within the context of this thesis. It is sufficient to say that those systems that use 80 column punched data cards as their source of input and use a standard line printer instead of a teletype may be considered as batch processing systems. Remote job entry by submission of jobs through a teletype is also considered batch, as there is no true 'interaction' between the system and the user. Batch systems, of course, may have the actual raw data entered from tape, disk, or other device, instead of cards.

Within the batch systems there are two main categories: integrated and non-integrated systems. Because the definitions of the above two terms vary from author to author, let us define an integrated system as a system whose input data structure is the same for all statistical and data management procedures and where there exists only one comprehensive program containing all these procedures, instead of an

organized package of independent programs.

The presentation is in alphabetical order except for those discussed in the miscellaneous section where the order does not reflect a judgment of the systems.

3.1 Integrated Systems

3.1.1 ASCOP - A STATISTICAL COMPUTING PROCEDURE

ASCOP was one of the first systems to be documented in a statistical journal (Cooper, 1967). Early success led to more developments (Cooper, 1969b), until ASCOP 3 was written (Cooper, 1972). This evaluation is based on ASCOP 3, but is just referred to as ASCOP.

3.1.1.1 Statistical Analyses

Common measures of central tendency and dispersion are available as well as frequency distributions and histograms, but the user has no control, however, over determining the classes. A normal fit can be applied to the histogram using the chi-square statistic. N-way tables of means are not possible. Multidimensional frequency tables are possible along with percentages, with the chi-square statistic being computed for two-way tables.

Simple product moment correlations are possible. The multiple regression procedure handles replicated values and a quasi-

stepwise regression exists where one can request the system to pick the 'best' variables as determined by the size of the multiple correlation coefficient.

Unfortunately, ASCOP stops selecting variables once the prescribed level is reached and does not give the user a chance to see if other variables might have achieved the same significance level. ASCOP provides facilities for storing the predicted values, residuals, regression coefficients, residual mean square with its degrees of freedom, and the square of the multiple correlation coefficient for use in other procedures. A similar procedure exists for polynomial regression. Other types of regression are not available. A very general bivariate plotting procedure exists, but, unfortunately, it needs a plotter instead of the standard line printer. Non-parametric tests are not available.

Multivariate analyses include factor analysis using either the principal factor solution or Joreskog's method. Varimax is the only rotation method. Factor scores as well as the coefficients from the factor analysis can be stored and used later. A non-stepwise discriminant analysis procedure concludes the multivariate analyses, which permits the storing

of coefficients, scores, and misclassification probabilities.

Analysis of variance procedures are limited to analysis of a balanced complete factorial design. Special procedures for other designs and two-sample parametric and non-parametric tests are not available. Distribution theory is not available. Miscellaneous procedures consist only of Diallel table analysis, serial correlations, and a test of runs.

3.1.1.2 Data Management

Input is the variable by case structure with provisions to identify replicates. If replicates are used then four new variables are automatically created, i.e., the mean, standard deviation, variance, and the number of replicates, and these new variables may be used in other ASCOP procedures. Both alphanumeric and numeric data are permitted with the ability to use flexible formats. Input can be from cards, tape, or disk. Storage and retrieval commands for data files are simple. Labeling of variables is permitted and labeling of cases is done automatically by ASCOP in simple numerical order. Generation of files within one run is permitted and utility procedures consist of sorting and adding new cases. Program control and data modifications procedures are good with the

facility to use many Fortran functions and commands. Use of MACROS is also permitted. Specialized data analysis procedures do not include commands to recode data simply. ASCOP procedures for handling missing values are vague. The manual suggests that blanks are read as missing values. The user, however, cannot specify his own values and has no control as to how the missing values are handled in the statistical analyses.

One can use ASCOP procedures to augment output but the defined procedures in ASCOP do not present enough examples to assess the labeling. Most output is not defined rigorously.

Limitations are stated within a particular section for all procedures and are difficult to understand. The number of cases permitted is not stated, but the maximum number of variables is 230 for the ATLAS computer and 75 for the IBM 360.

3.1.1.3 Ease of Use

The ASCOP manual is well organized but quite verbose and difficult to read. Its 158 pages could easily be cut in half and filled with badly needed examples showing deck

set-ups as well as output. Statistical output is generally not defined and references are lacking. The commands are in natural English, but without the examples they are difficult to understand.

Nothing is mentioned about 026 or 029 keypunches, variable core sizes, error messages, or control when errors are found. It is doubtful whether one could run an ASCOP procedure without assistance from someone with prior experience.

3.1.1.4 Other Considerations

ASCOP is mainly written in Fortran with a few routines written in assember, and has about 18,000 lines of source code. Versions of ASCOP exist for ATLAS, ICL 1800, UNIVAC 1108, and IBM 360 series computers. ASCOP requires 150K and runs under the Fortran 4 H compiler on the IBM 360 system.

3.1.2 OMNITAB II

This evaluation is based on two manuals: *OMNITAB II User's Manual* (Hogben, Peavy, and Varner, 1971) and *Test Problems and Results for OMNITAB II* (Varner and Peavy 1970).

The name OMNITAB comes from 'omnibus' and 'tabular'. There seems to be some ambiguity as to what the OMNITAB II really is. It is a programming language, but is 'highly user oriented' in its design. It is directed toward the individual who needs computing power, but lacks knowledge of high level programming languages such as Fortran, APL, etc. The basic design philosophy consists of a 'workspace' composed of rows and columns. A simple example would be to enter the data into the first column, then using an OMNITAB II instruction, say, SQRT (square root), one could take the square roots of all the numbers and put the resultant answers in column two. Thus, the modus operandi is to move across the worksheet performing calculations much as one would do with a desk calculator.

3.1.2.1 Statistical Analyses

The descriptive statistics of OMNITAB II are comprehensive.

Besides the standard statistical measures, the following statistics are provided for univariate data: 25 percent trimmed mean, mid-range, mean deviation, two-sided confidence intervals for the mean and standard deviation, linear trend statistics, tests for non-randomness, sums of squares, deviations from the mean, and ordering and ranking of the observations. Frequency distributions and histograms are possible with the user allowed to determine exactly what the class boundaries, class widths, etc. should be. Where appropriate, procedures exist to do weighted analysis if desired, as well as storing the answers in the workspace in order that more analyses may be performed. No procedures exist to produce N-way tables containing either frequencies or means.

Simple correlation matrices include product moment and Spearman coefficients. Partial correlation coefficients are also available. Significance levels and confidence intervals are possible as well as a test for linearity. The above coefficients may be stored. Bivariate plots are possible. The regression analyses include simple, multiple, and polynomial procedures with the ability to do a weighted analysis on each. Excellent features are the plots of the standardized residuals versus each of the following: row

numbers, independent variables, and predicted values. A probability plot of the residuals is also possible. Multivariate analyses are not available.

A one-way analysis of variance procedure exists along with two multiple-range tests, viz., Newman-Keuls with Hartley modification and Scheffés. Kruskal-Wallis' non-parametric statistic is also possible. A two-way analysis of variance procedure is available with the ability to utilize weights, if the design is unbalanced. Advanced design procedures and special two-sample procedures are not available.

Distribution theory consists of two procedures: one to find p-values from the F distribution, and another to generate uniform random deviates between 0 and 1.

3.1.2.2 Data Management

Input of data for OMNITAB II is either the variable by case structure (READ command), or by just declaring a variable and assigning the values to it (SET command). For analysis purposes only numeric data is permitted and either free-field or standard Fortran formats are possible. Input can be from cards or tape.

Facilities do not exist within OMNITAB II to store data files. Data manipulation features exist for moving, sorting, and selecting data about the worksheet. For matrices common array operations are possible (e.g., add, transpose, invert, etc.). Other mathematical procedures are numerical analyses and Bessel functions. Assignment statements and arithmetic expressions are possible as well as branching and iteration. OMNITAB II is, however, a programming language and the aforementioned points are to be expected.

Procedures to handle missing values are not mentioned and the system will input zeros, if blanks are punched. Missing values can be handled, of course, just as in any programming language by writing a program, but as said before, programming should not be a prerequisite to using a statistical system. Output is well labeled with the user being able to insert his own labels if wanted. Output is defined. As mentioned before, many forms of output can be further used as input.

The limitations in OMNITAB II are clear but they will vary depending on the procedure used and also with the amount of data. The product of the number of rows and the number of columns must not exceed 12,500. If no intermediary calculations are made by the user on his data, this restriction would permit

about 1000 observations and eight independent variables in the multiple regression procedure. This is restrictive for practical data analysis applications.

3.1.2.3 Ease of Use

The user's manual is not specifically geared to statistical analysis. OMNITAB II is for all intents and purposes a programming language and it attempts to satisfy the needs of many disciplines including statistics. Nevertheless, the section on 'Beginner's Omnitab' initiates the novice so that a user with a relatively easy statistical problem can use the statistical procedures described later in the manual without too much trouble. Although the manual is explicit, and in certain areas very comprehensive, it is not that readable. One has to jump around completely outside of the statistical section very often to answer many of the input, data management, and output problems and it can prove to be frustrating. Two notable features are an abundance of references for the statistical procedures and an attempt to discuss the precision and accuracy of some of the algorithms employed. Use of the data management procedures, although conceptually easy, are cumbersome and difficult to use as one has to remember column numbers and row numbers where

data is stored as well as construct minor 'programs' to manipulate the data as one procedure is not usually powerful enough (e.g., recoding data, handling missing values, etc.).

Jowett, Chamberlain, and Mexas (1972) give an easy to understand account of OMNITAB II including good statistical examples in their paper.

3.1.2.4 Other Considerations

OMNITAB development began in 1963 with Joseph Hilsenwrath setting down the basic philosophy and it has been under development ever since, culminating with Version 5.0, called OMNITAB II. A National Bureau of Standards publication announcement in January 1972 says, "OMNITAB II is a virtually machine-independent computer system." It does, however, require a large computer, i.e., IBM 360/50 and up, GE 265, CDC 3800 and up, Burroughs 5500, and UNIVAC 1108. OMNITAB II is written in as standard a Fortran as is possible. The ability for the user to add his own routines to OMNITAB II is possible but it is difficult. Besides the two manuals referenced in this evaluation, there are two others which were not obtained but are available: *Source Listing of OMNITAB II Program* (371 pp.) and *A System Programmer's Guide*

for Implementing OMNITAB II (43 pp.). The OMNITAB II package is available for $250.

3.1.3 P-STAT

The P-STAT system was developed by Buhler (1973a). Initially, development began in 1963 by Buhler at Princeton University and has evolved to its present state, Version 3.05, 1973. This evaluation is based primarily on the user's manual which is computer generated, as is all documentation described in the reference. Also available, but separate from the manual, was a primer consisting of 33 examples of P-STAT job set-ups with explanations, and nine examples showing deck set-ups and resultant P-STAT output.

3.1.3.1 Statistical Analyses

Using the FREQ and DES procedures univariate descriptive statistics can be obtained as well as frequency distributions. N-way tables of means are implied, but it is not clear how to obtain them from the user's manual or examples. Two-way tables of means are, however, definitely possible. Statistics like the median, mode, skewness, and kurtosis are not available. A histogram is not available from the previously obtained frequency distributions.

Up to six dimensional frequency tables are available with the chi-square and other tests of association being available for two-way tables. Percentages are optional with the frequency tables.

Pearson product moment correlations as well as biserial and tetrachoric correlations are possible. A stepwise regression procedure exists that is very flexible and has many options for the entering and deletion of variables including good output control and storing of residuals. A non-stepwise regression program is also available. A plotting and graphing procedure exists but is mentioned in a different section of the manual. Polynomial, non-linear, and weighted regression procedures are not available.

Multivariate analyses includes factor analysis which does a principal components or an iterative analysis with facilities for orthogonal or oblique rotations. Discriminant analysis is available but not in a stepwise mode.

One and two-way analyses of variance procedures exist while hierarchical designs of analysis of variance and covariance, both balanced and unbalanced, employ the MANOVA program (discussed in the Miscellaneous Section under Non-integrated systems). Two-sample t-tests are possible but non-parametrics

are not. There are no procedures for distribution theory and no miscellaneous statistical procedures.

3.1.3.2 Data Management

Comprehensive input procedures exist. Input is the variable by case structure and three procedures exist for reading data from cards depending on how 'clean' the data is. The most powerful procedure treats blanks automatically as missing values, sets mispunches to missing values, and checks if cards within cases are out of order. Free-field formats are not permitted nor are alphanumeric fields, except when used as case labels. Input formats are flexible, with input being possible from cards, tape, or disk.

The ability to generate, store, retrieve, and purge data files is powerful and relatively simple to use. Facilities exist to label the cases and the variables. Value labels, however, are not available, except for the crosstabulation procedure. There exist procedures to merge data files; the most powerful being MATCH which is a very general procedure that produces output files matched according to the case labels. Thus, the addition of new variables and cases is possible. Sorting procedures as well as miscellaneous

procedures are plentiful, e.g., scaling of variables, ranking. For matrices, procedures are available to add, multiply, invert, etc. P-STAT allows access to many files during one computer run and many procedures allow output from a previous procedure to serve as input to the next or later procedures.

P-STAT has editing facilities that check for bad data, improper sequencing, etc., and informs the user of these problems. Transformation of data is limited to cumbersome codes of which the first 24 transformations are based on the BMD package of transformations. Facilities exist to use logical operations (greater than, equal, etc.); create new variables; modify, select, delete, etc. DO-LOOPS are permissible and the powerful MACRO feature is available in P-STAT. In running P-STAT the user may select only certain cases if desired - a useful feature but not very common. Only one missing value may be designated for each variable and almost all programs can handle missing values.

Output labeling is good with the user being able to supply his own. Value labels as already mentioned are absent - except for crosstabs, but are being implemented in a future version. Output from P-STAT is not defined. Limitations are stated for each procedure.

3.1.3.3 Ease of Use

P-STAT's weakness is its user's manual and command language. Many scans of the whole manual were needed before the P-STAT evaluation could even begin. The deck set-ups and P-STAT printouts for each procedure should be provided within the user's manual. Examples for many statistical procedures are missing. The statistical output is not explicitly defined and references are minimal. Except for the defining of the statistical procedures, the manual is fairly explicit, but difficult to understand. Error checking and retaining control are strong points and help in ease of usage.

3.1.3.4 Other Considerations

P-STAT is written in 'clean' Fortran according to Buhler, except for 100 assembly statements, comprising of 480 subroutines and 45,000 lines of source code. About 40 other computer centres besides Princeton have P-STAT. It has been converted to the following computers: IBM 360/370, PDP-10, SIGMA 7 and UNIVAC 1108. P-STAT comes in 7 sizes: Wee, Tiny, Small, Medium, Large, XLarge, and Jumbo. The Wee size takes 150K on a 360/370 and can handle 150 variables.

The Jumbo takes 680K and can handle 3000 variables. Thus, it becomes easy to adapt to other computers. For a detailed discussion of making P-STAT portable see Buhler (1973b).

If it is installed at a university, P-STAT is available for $540 per year for IBM 360/65 computers and up. It should be noted that Buhler sends free of charge a mini tape of the Wee P-STAT system to prospective users as well as the user's manual. Included with the tape are source cards that theoretically need to be changed only in a couple of places for the user's particular installation. The P-STAT system provides facilities so that a user can add his own routines and is discussed in the user's manual.

Future plans include the general use of value labels, as well as making P-STAT run on CDC 6400/6600 and Burroughs 6700 computers. For an overview of the P-STAT system other than the user's manual see Buhler (1973c).

3.1.4 SAS - STATISTICAL ANALYSIS SYSTEM

This evaluation is based on the user's manual by Barr and Goodnight (1972).

3.1.4.1 Statistical Analyses

A MEANS procedure is used to calculate descriptive statistics. However, statistics like the median, range, standard error of the mean, skewness, and kurtosis are not available. Histograms are also missing. Useful statistics are available that are not usually included in other systems. They are sums and corrected sums of squares. N-way tables of means are possible by using the BY statement. The variables used in the BY statement must, however, be sorted previously to using the MEANS procedure. A SORT procedure is available to do this, but this is inconvenient.

Up to five-way frequency tables are available using the FREQ procedure. Cumulative percentages are missing from the tables. The output of the five-way table is neat and compact, in contrast to a series of two-way tables. The compact method, however, has its disadvantages in readability and in interpretation. For two-way tables the only statistic of

association available is the chi-square. No column and row totals or percentages are available for any tables. Bivariate plots are available using the PLOT procedure. The PLOT procedure when coupled with the SAS data management facility makes it powerful, lending itself to such things as contour plots.

Simple correlations (Pearson, Spearman, and Kendall Tau-b) are available using the CORR and SPEARMAN procedures. P-values are given and for large correlation matrices an option is available to print only the most significant p-values.

The multiple regression and correlation procedures are excellent. The RSQUARE procedure performs all possible regressions of the dependent on the independent variables. Various options are available with this procedure so that some degree of control over the selection of variables and output is available. The STEPWISE procedure enables one to do five types of stepwise regression. Again, various options are available for controlling the process. Tests of significance are available with the STEPWISE procedure, but predicted values, residuals, and confidence limits are not.

Factor, discriminant, and canonical analyses can be performed. The FACTOR procedure allows only the varimax rotation. Useful options are principal component scores and factor loadings. The DISCRIM procedure is good. A test of homogeneity of the within covariance matrices is available, and the classifications scheme allows unequal prior probabilities of group membership. The CANCORR procedure permits the plotting of canonical variates.

The analysis of variance procedure ANOVA is comprehensive. One-way analysis, factorial, and hierarchical designs are possible. For unbalanced designs the sums of squares may be invalid (See Section 2.4.4 on accuracy). The type of model is easily specified in the MODEL statement. Pooling and tests of significance are also possible. Before the ANOVA procedure can be used the data has to be sorted with regard to the factors using the SORT procedure. Although nested analyses are possible using the ANOVA procedure, a more comprehensive procedure is available, called NESTED, its advantage being to estimate variance components and do covariance if applicable. Also, a special LATTICE PROCEDURE is available for analysis of variance and covariance for lattice designs. Analysis of covariance is not available with the ANOVA procedure. A DUNCAN procedure can be used for doing Duncan's multiple range tests.

The most powerful procedure of SAS is REGR. The user's guide says, "The REGR procedure will apply the principle of least squares in fitting a linear model of virtually any type to data." Thus, alternative methods of regression are possible. Polynomial and non-linear regressions can be performed by this procedure. The analysis of covariance lacking in the ANOVA procedure can be done with REGR. Analysis of variance is possible through the creation of dummy variables.

SAS lacks specific procedures which can perform two-sample t-tests and their non-parametric counterparts. Miscellaneous procedures included are: RANK which ranks variates and could enable one to write his own non-parametrics; and RQUE procedure which produces restricted quadratic unbiased estimates of variance components; GUTTMAN to do Guttman scaling; and PLAN which can generate randomized plans for experiments.

3.1.4.2 Data Management

The variable by case structure is used and input can be from cards, tape, or disk in flexible formats. Provisions exist for alphanumeric data. Instead of using a standard Fortran format statement to declare what type the variables

are, i.e., alphanumeric or numeric, and what columns they are in; each variable must be stated, followed by a dollar sign if alphanumeric and then its column numbers of location. For a novice this seems simpler, but for large numbers of variables and many cards per case it is cumbersome as the card number must also follow the variable name. Excellent procedures exist for the generation, sorting, concatenating, merging, and forming subsets of data files. Ouptut from the statistical procedures cannot be stored and later used as input. If invalid characters appear where numeric data should be then SAS assigns missing values to the variables, and then carries on. Blanks found where numeric data should be are also assigned missing values.

SAS contains a strong set of program and data analysis statements with the ability to use MACROS. Functions that generate normal, uniform, and other deviates are possible. There is, however, an absence of a simple way of recoding variables. The cumbersome 'IF....THEN' statements must be used. The ability to designate missing values exists. All the statistical procedures state explicitly what happens when missing values are encountered, but unfortunately the

user has no say whether they should or should not be used.

The output labeling in SAS is good. One can use a TITLE statement to put out one's own headings. The title does not have to remain static and can change within a program run. The options available in the statistical procedures give one control over the amount of output. The output in general is not rigorously defined. The maximum number of variables is 255 and must be contained on 12 or fewer data cards, or a maximum of 1000 characters if read off tape, disk, or drum. Both limitations could be restrictive. The number of cases appears to be solely a function of the storage space available with the computer centre. Limitations on the number of variables, cases, etc. in statistical procedures is usually not mentioned.

3.1.4.3 Ease of Use

With the exception of limitations and the rigorous defining of output, the user's manual is explicit. There could be more examples. Independence between procedures is good, and there is an abundance of references for each procedure to assist interpretation. The command language is simple,

as well as the JCL. SAS does not provide a comment saying that 026 keypunches are possible. Error messages are not listed in the user's manual. When errors are encountered though, SAS implies that they are adequately flagged and SAS carries on, if it can.

As a whole, the user's manual is quite readable. The introduction starts with a simple example which conveys the basic principles clearly.

3.1.4.4 Other Considerations

SAS has 35,000 source statements and is written in Fortran, PL-1, and assembly languages. It runs under an IBM 360 or 370 with at least 120K of core storage. The user's manual lists approximately 60 other SAS installations. Transportability to IBM 360 and 370's is easy but not to other computer systems. Users can add procedures of their own to SAS without great difficulty. SAS development and maintenance has been guaranteed through the next five years and is available for $600 per year for degree granting institutions (Goodnight, 1973).

3.1.5 SPSS - STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES

This evaluation is based on the user's manual by Nie, Bent, and Hull (1970), as well as the three updates, Nie et al. (1971, 1972, and 1973).

3.1.5.1 Statistical Analyses

SPSS is excellent in descriptive statistics and frequency tables - both one and multidimensional. Six independent procedures cover these areas.

Simple correlation (Pearson, Spearman, Kendall) and partial correlation procedures are available. The multiple regression procedure provides plots of residuals and is capable of storing them. A stepwise mode is possible but allows only the method of forward selection of the variables. Bivariate plots are possible, but polynomial and non-linear regression are missing.

An excellent factor analysis procedure is available providing five types of factoring and four types of rotation. Factor scores are available. Many options are available to give one control over the process. Procedures also exist for

canonical correlations and stepwise discriminant analysis.

A one-way analysis of variance procedure with multiple range tests, along with a very useful t-test procedure exist. That, however, is SPSS's scope in analysis of variance and covariance. There are no procedures for distribution theory, and the only miscellaneous procedure is Guttman scaling.

3.1.5.2 Data Management

SPSS uses the variable by case structure for input. Formats can be fixed or free-field and data can be either numeric or alphanumeric with the facility to use 026 keypunches. The data matrix is capable of being easily stored and retrieved. Excellent facilities exist for labeling variables and their values. Case labeling is not possible. SPSS allows only one file to be generated and saved per run. Subsets of files are possible using the SUBFILE LIST procedure, but unfortunately these must be generated at the time of input. Merging and concatenation of files is not possible - though with the use of subfiles, cases or variables can be added, but the procedure is rigid, e.g., when adding cases

all the variables present in the original file must be present with the new cases. A SORT procedure has recently been added. A special WRITE CASES procedure must be invoked, if other than the SPSS system is to read the file.

Most procedures do not permit output to be saved and later used as input. Editing of data is left to the user, but a new feature is the EDIT card which edits the SPSS control card deck prior to a live run. A good feature when fixed format is used is that SPSS will list each variable, its format, and in what record and columns it is located. SPSS program control and data modification procedures and statements are adequate. One, however, cannot define MACROS, and branching with the use of a GO TO is not possible. Specialized data analysis procedures are good. The procedures to recode and select data either temporarily or permanently are easy to understand and use. The ability to designate and handle missing values is exceptionally good. All statistical procedures give options to the user on how he wants to handle the missing values.

SPSS output can be voluminous but it is very readable. Most statistical output is well defined, but some of the updates are not, e.g., the discriminant procedure. SPSS is working towards saving output and then using it for input.

One has fair control over output. Many procedure have options that either present the output in a condensed and compact fashion or if one wants, say, for various reasons more labeling, different format, etc., that too is possible.

A maximum of five hundred variables can be processed by SPSS. A recently added feature called ARCHIVE, however, enables one to store up to 5000 variables, but only 500 may be processed. A maximum of one hundred subfiles is possible. There appears to be no restriction on the number of cards per case, and the number of cases is limited only by the amount of direct access at any computer centre. Limitations are stated within each procedure and are clear, consistent, and liberal. Also, formulas are given whereby the user can adjust the amount of core size needed to run SPSS depending on the size of his job.

3.1.5.3 Ease of Use

There is no doubt that much of the popularity of SPSS is due to its professional and readable user's manual. The original manual, but not the updates, has all the desirable features mentioned in Section 2.3. The original manual knows 'how' to say it and, hopefully, will set a precedent. With three independent updates, however, it is

becoming monstrous. The command language is lucid. There are a host of minor commands, e.g., DUMP (dump the file), NUMBERED (ability to number control cards), which make data analysis that much easier. Error messages are all listed in the manual as well as being written on the computer printout.

3.1.5.4 Other Considerations

SPSS is currently in use at well over 100 IBM installations with versions also running on other large computer systems or else being adapted. The original version of SPSS was available for $400 with later versions costing around $200.

Future plans call for a truly conversational mode. A notable feature to be included in the next version (6.0) is an N-way analysis of variance procedure.

3.1.6 Other Integrated Systems

3.1.6.1 DATA-TEXT

This evaluation is based on the user's manual (Armor and Couch, 1972) and a paper by Armor (1972). DATA-TEXT was one of the forerunners of integrated systems with the first version being implemented on an IBM 7090/94 system, becoming operational in 1963-64. It has been re-written and now is available on IBM 360/370 computer systems, but requires, however, a minimum of 200K core storage.

The DATA-TEXT system was designed for social scientists and in many respects is like SPSS, but DATA-TEXT lacks procedures that perform discriminant functions; canonical and partial correlations; and histograms. DATA-TEXT provides procedures, however, for an N-way analysis of variance for factorial designs as well as repeated measures designs. Unweighted means using the harmonic means solution is employed for unbalanced complete designs.

Data management is similar to SPSS with both systems being easy to understand and use. DATA-TEXT's ability to define

the variables and their values in one cohesive unit would be preferred by many as compared to other systems. Some other notable features are: data modification commands can be applied to lists of variables; DATA-TEXT uses dynamic core allocation permitting analyses to be run in small amounts of computer core (Armor (1972) says, "All of the statistical routines are almost literally without limitations."); the ability to do analyses by various categories easily using the GROUP BY option.

3.1.6.2 EASYSTAT

The EASYSTAT system (Tucker, 1973) is written in PL-1 and runs on an IBM 360/40 in 108K of core, at Georgetown University. Statistical analyses appear comprehensive including procedures for Friedman's two-way non-parametric analysis of variance and the Cochran Q test. Nothing is mentioned about data management. The user's manual is designed to be especially user oriented. For example, once the user knows what statistical procedure he wants, he then goes through a tree structure answering questions in order that he does not get lost in the maze of rules for data definition.

3.1.6.3 SOUPAC - STATISTICAL ORIENTED USERS, PROGRAMMERS, AND CONSULTANTS

The SOUPAC system has been under development since 1962 and is implemented at the University of Illinois and is presently running using an IBM 360/75. Information on a user's guide has been requested but has not been received to date and this evaluation is based on the paper by Chouinard (1973).

SOUPAC has about 70 statistical procedures. Its statistical strengths lie in the multivariate analysis procedures which includes a factor analysis procedure that contains the latest factoring and rotation schemes, viz., the rotations are binormamin, varimax, varisim, oblimax, oblique, orthogonal procrustes, and personal probability function rotation. Analysis of variance and regression procedures seem comprehensive. Its weakness lies in the inability to perform multiway tables of means and frequencies along with the associated tests of association.

The data management section has facilities to generate data files and manipulate them. The labeling of variables and their values are weak areas. From the commands presented it appears as if SOUPAC is easy to use.

There is some doubt as to the transportability of SOUPAC when Chouinard mentions in his paper that some of the package is written in non-standard Fortran. Its development, however, seems certain. Chouinard says, "The statistical procedures in the remainder of the package are kept up-to-date and represent, we believe, one of the largest and most well rounded parametric statistics library." Also, an interactive version is being planned for the future.

3.1.6.4 STATPAC

STATPAC is evaluated with the aid of the following literature: a simplified user's manual (Nelson, Hendrickson, Phillips, and Thumhart, 1973) and a paper describing STATPAC (Nelson, Phillips, and Thumhart, 1973).

This package, developed at the General Electric Corporation, Schenectady, New York, is written in Fortran for GE and Honeywell 600 and 6000 series computers. The authors stress the need for an easy to understand manual, as well as a system that does housekeeping chores. They say, STATPAC is "a simple and powerful statistical package for all who analyze data." That statement, however, is open to question. STATPAC can only read and store a maximum of thirty variables

and handle only 500 cases. In the two-way crosstabs tables (multidimensional tables are not possible), if a cell has more than nine counts then the following are printed: the letters of the alphabet for numbers between 10 and 35, the dollar sign ($) for numbers between 36 and 50, and so on. Also, the selection and transformation commands are not that simple to use. For example, to say Y=A+B in STATPAC would require the following: USE SUM (A B) TO SET (Y).

STATPAC's strengths lie, however, in a serious attempt to recognize the need of helping the novice but not stifling the sophisticated user and such aids are apparent in the format of the user's manual. STATPAC contains statistical procedures not common in other packages, e.g., fitting distributions to data and analyzing censored life data.

3.1.6.5 TSAR - TELE-STORAGE and RETRIEVAL

The evaluation of TSAR is based on the user's manual (see reference under TSAR). TSAR is implemented at Duke University and runs under IBM 360 series equipment.

The statistical analyses section does not compare with the major systems. That there are no multivariate analyses, and

easy-to-use facilities for multiway tables are big weaknesses. TSAR lacks powerful data management procedures. TSAR, however, contains a procedure that is useful, but is difficult, or absent, on many other systems; namely, the ease of updating files and correcting data. Also, there is a function that calculates the number of years elapsed, if the two dates are given - a minor procedure but very commonly needed in data analysis. Facilities exist to add user subroutines. The user's manual is somewhat disorganized, jumping from statistical to data management procedures and vice-versa, but it is readable and provides good references.

3.1.7 Miscellaneous Integrated Systems

The MINITAB system (Ryan and Joiner, 1973) is the 'little brother' of the OMNITAB system, and is being used successfully for course work and research at Penn State. It is free of charge, except handling charges, etc., and is written up in a 50 page user's manual, which is available from the authors.

The STATLAB system (Atkinson, 1971) although intended for statistics labs is, nevertheless, capable of being used to some extent for small research problems. Commands exist to manipulate scalars, vectors, or matrices as well as perform special statistical operations, e.g., sums, sums of squares, etc. STATLAB is a programming language but is, indeed, "easy-to-learn [and] easy-to-use," as indicated by Atkinson. STATLAB is written in assembler and Fortran IV for an IBM 360/370 computer. The system has changed little from 1971 and the latest user's manual may be obtained from Atkinson.

The STATJOB system (Muller, 1970) appears to be a large integrated system. Muller outlined in detail many of the goals the STATJOB system should adhere to, but unfortunately transportability was not one of them. Personal communication

with the University of Wisconsin indicates that STATJOB runs only on UNIVAC 1100 series equipment and they estimate it would take two man-years to convert to an IBM 360.

See Myers (1969) and Schucany et al. (1972) for some other miscellaneous systems.

3.2 Non-integrated Systems

It might be helpful to restate what a non-integrated system is. A non-integrated system consists either of a series of independent programs, or one large program where the basic input data structure has to be changed for different procedures.

3.2.1 BMD - BIOMEDICAL COMPUTER PROGRAMS

This evaluation is based on the user's manual (Dixon, 1973).

3.2.1.1 Statistical Analyses

Six independent programs are avilable for descriptive statistics, but not one produces the median. N-way tables containing means, etc. are not possible, although two-way is possible using BMD06D. Frequency distributions and histograms, however, are abundant. Multidimensional frequency tables are possible using BMD08D with its 'variable stacking' feature. This program can only handle integers and does not accept missing values, but BMD09D, which is another crosstabulation program, can handle missing values but does not have 'variable stacking'. Also, for two-way tables the chi-square statistic is given for BMD08D but not for BMD09D. Another

program in a different section BMD02S performs a contingency table analysis and gives percentages. A maximum likelihood statistic is produced with this program, along with the chi-square, of course, but other tests of association are missing. As one can see, organization and consistency leave much to be desired.

Simple, multiple (stepwise), and polynomial regression are possible as well as simple correlations. There is more than one simple correlations program - there are three to add confusion. No non-parametric correlations are available. Non-standard regression programs, however, exist. They are: periodic regression and harmonic analyses; asymptotic regression; and non-linear regression where the user specifies the function.

Multivariate analyses include principal components, stepwise discriminant analysis, factor analysis, canonical correlations, and a program to detect outliers in multivariate data. For the initiated a multivariate analysis of variance and covariance as well as a multivariate general linear hypothesis program are available.

The analysis of variance and covariance programs are also very comprehensive. Factorial analysis with multiple covariates,

and analysis of hierarchical designs are possible, but the designs must be balanced and complete. Unbalanced designs can be handled by the general linear hypothesis program. Duncan's multiple range test is available as a separate program. Again, there are ten programs covering the ANOVA and ANCOVA area. Three of these are the general linear hypothesis programs. There is a good t-test program, but there is a complete absence of any non-parametrics throughout the package.

Miscellaneous programs include life tables and survival rates, probit analysis, and four time series programs.

3.2.1.2 Data Management

BMD uses the variable by case structure. Free-field is not permitted. Whether alphanumeric data, integer, or real data is allowed depends solely on the particular program. It is not consistent, e.g., crosstabulation program BMD08D only accepts positive or negative integers. JCL is minimized where applicable and input can come from cards, tape, or disk. Input and output from tapes, etc., must be done with JCL. Variables can have up to a six character label, but only in a small number of programs. BMD does not have its

own file system, so manipulation of data files is non-existent. However, a sort program BMD14S exists. Program control and data modification are possible. Through the use of 'trans-generation codes', one can perform a test of approximately thirty transformations, e.g., square roots, reciprocals, etc. These are difficult to use and very limited. To overcome this, two programs BMD12S (openended transgeneration) and BMD13S (multipass transgeneration) are available. One can supply his own FORTRAN statements to modify data. The process, however, involves putting the transformed data onto an output tape or disk and then using the tape or disk as input to the required program. This is not practical. Specialized data analysis statements that are applicable to all programs are non-existent.

The statistical output is labeled adequately, but, because the input is not labeled, most output refers only to the 'variable number', instead of a concrete name. The 'Computational Procedures Section' does a good job on rigorously defining output, but at perhaps too high a mathematical level and with too much computer jargon, e.g., see BMD03S (probit analysis).

Limitations are clearly stated for each program. Some programs

are liberal, some are not, e.g., BMD08D (crosstabulation) allows a maximum of one hundred variables and 1500 cases. Most programs give the user an estimate of how long the job will take and how many pages of output to expect.

3.2.1.3 Ease of Use

The user's manual is a formidable 773 pages. It is explicit and programs are relatively independent. The first 66 pages are basically about 'how to collect data and use BMD'. Even though it is intended for novices it is still difficult to read. Each program in the manual is consistent in its approach, but unfortunately the programs are difficult to understand and use. The manual has a habit of using words that have no meaning. For example, instead of saying multidimensional, crosstabulation, or N-way tables of frequency counts, they call it crosstabulation with 'variable stacking' (BMD08D). Another is 'multipass transgeneration' (BMD13S). The write-up saying what BMD13S does is just as confusing.

Some programs handle missing values; some do not. Some handle labels; some do not, etc. Error messages are not listed in the manual.

3.2.1.4 Other Considerations

BMD was one of the first major statistical packages. Because BMD consists of independent programs and runs on IBM 360/370 equipment, it is quite transportable. BMD is easily implemented and costs $45.

BMD is under constant development. At the time of this writing the BMDP series is currently under development. Frane (1973) discusses the new BMDP series and says, "...[It is] the start of a complete third generation package." From Frane's paper it appears that the programs are more integrated, i.e., all programs can handle missing values, all programs can have variable labels, etc. More statistics are added, e.g., non-parametrics. The examples have annotations written on them to help in interpretation, etc. All in all it looks like a much superior product.

3.2.2 STAT-PACK - A BIOSTATISTICAL PROGRAMMING PACKAGE

A request to Goddard Computer Science Institute concerning STAT-PACK resulted in the following user's manual (Shannon, 1967), upon which this evaluation is based.

3.2.2.1 Statistical Analyses

Comprehensive descriptive statistics are possible, if one has only one set of data, but if one wants to analyze multiple sets then a different program must be used which is not as comprehensive. A frequency distribution and histogram can be obtained. N-way tables of means are not possible. Cross-tabulations are limited to a primitive two-way program which gives no tests of association.

Simple linear regressions are possible as well as a weighted polynomial regression which can give confidence intervals. Weighted multiple regressions, including stepwise, are possible. A non-linear regression program is available, with the user supplying his own Fortran subroutine of the model. A useful, but not so common, program is available for simple linear regressions where for various dependent values the independent is derived and then confidence limits obtained

about the independent values. A scattergram program exists. A Pearson correlation matrix program is possible. Kendall's tau correlation is possible but not as a matrix. Multivariate analyses are limited to an optionless two group discriminant function.

Analysis of variance of one-way designs with subsampling, and up to three-way balanced complete factorial designs are possible. Unequal cell sizes, however, can be handled in the two-factor model, assuming there is no interaction term. Multiple range tests are limited to Duncan's procedure.

Special two-sample tests include paired and unpaired t-tests, but no non-parametrics. Bartlett's test of homogeneity of variances is available. Distribution theory includes the use of the chi-square and Kolmogorov-Smirnov statistics to fit six common distributions. Random numbers can be generated from the uniform, normal, and exponential distributions, but unfortunately the user must call a Fortran subroutine. A similar feature exists to find probability values for given functions.

Some miscellaneous programs are Kendall's coefficient of concordance, estimation of the parameters of the log and

translated log normal distribution, time series, life tables and survival rates, and a few mathematical routines such as transformations and matrix inversion.

3.2.2.2 Data Management

Only numeric data is permitted and only from cards. The user must supply the data in fixed format, as requested by the read subroutine which can be altered. Input is not the same for all procedures. Data files cannot be generated and thus, there exists no manipulation of them. Program control and data modification features are non-existent except for 25 predefined transformations, which are similar to the BMD transformations and which exist in a separate program and are, for all intents and purposes, valueless as they cannot be used within the other programs. There are no specialized data analysis statements(recode, select, etc.) and there is no way of designating and handling missing values.

Output is adequately labeled, but not rigorously defined statistically. Limitations are stated but very often are not liberal considering this is a batch system, e.g., only nine independent variables in the multiple regression program.

3.2.2.3 Ease of Use

The user's manual lacks clarity on input and output. For example, by reading the manual it is unclear whether the user has to supply his own Fortran subroutine to read the data and if so, where it fits in with the rest of the cards. In actual fact, the user does need his own read subroutine and information on where it fits in the deck set-up is supplied in one of the appendices. The appendices, however, are not listed in the table of contents. Unless the user was familiar with Fortran, he would be unable to use this package due to its heavy reliance on computer jargon in the data set-ups. For some analyses Fortran knowledge is essential. Aside from the previous points the manual needs more examples. References, however, are quite adequate.

Like most non-integrated batch systems the programs are difficult to use due to their different input data structures and other parameters for each program and their rigidity in having to use numeric codes in exactly the right columns to steer the analyses.

3.2.2.4 Other Considerations

STAT-PACK is written entirely in Fortran with non-standard features avoided according to the Preface. Each program runs in 8K words of core which is an advantage for transportability to smaller systems. There are 53 independent programs.

3.2.3 STATISTICAL PACKAGE - UNIVERSITY OF MANITOBA

This package is an organized collection of programs with the evaluation based on the user's manual (Chebib, 1972).

3.2.3.1 Statistical Analyses

A program is available for descriptive statistics including percentiles, frequency distributions, and histograms. N-way tables of means are not possible. Another program exists for multidimensional frequency tables with tests of association being available for the two-way tables. Percentages are not available in the above program. For data that has been previously tabulated, there also exists a separate program to do a chi-square analysis. Fisher's exact test is not available for small 2x2 tables.

Product moment correlation matrices are possible as well as simple, multiple (including stepwise), polynomial and exponential regression. Spearman's rank correlation is available, but only for two variables, i.e., no matrix of correlations is possible. Partial correlations, scattergrams, and facilities for producing non-parametric correlation matrices are unavailable. Multivariate analyses include

factor analysis (principal factoring and varimax rotation) and a canonical correlations program.

Analysis of variance includes factorial analyses using Snedecor's harmonic means solution when cell sizes have unequal numbers. Unfortunately, if the design is one-way this approximation is also used. Analysis of covariance allows up to three factors and nine covariates. There is also a separate, but comprehensive, one-way analysis of covariance. Other programs include latin squares, split plots, experiments replicated in time and space, and partially balanced lattices.

Two-sample tests include the t-test and the Mann-Whitney test for unpaired data and just the t-test for paired data. There are no programs on distribution theory.

3.2.3.2 Data Management

Only numeric data is permitted, but a variable format is allowed. The input structure is not the same for all the programs and data input is only allowed from cards except for the multidimensional frequency table (MFT) program where input can be from tape.

Data files cannot be generated, thus, storage, retrieval, and manipulation of them does not exist. Editing of data is not provided. Program control and data modifications procedures are absent. Some programs permit up to nine predefined transformation codes. The MFT program uses the variable by case structure and permits up to 24 transformation codes which are a bit more advanced (e.g., can subtract constants, add two variables, etc.). Specialized data analysis procedures do not exist except for the MFT program which permits a cumbersome selection procedure. No provisions exist to handle missing values except for the MFT program where the minimum and maximum can be stated and all values outside this range are excluded, but this could only be called quasi-missing.

Output is adequately labeled. The user, however, is only allowed one heading of 40 characters. Labeling of variables and cases is not permitted except in the MFT program where eight character labels are allowed for variables. Limitations are stated in each program and are liberal.

3.2.3.3 Ease of Use

The user's manual provides no examples of output. Output is not

rigorously defined and references are minimal. No introduction is given as to how to use the programs.

This package, not being an integrated system, does not have a command language and uses codes, numbers, etc. on the cards to specify what to do and is subject to difficulty of use. Besides this, the fact that examples are missing often leaves one at a loss as to what is really meant.

3.2.3.4 Other Considerations

The programs are written in Fortran IV and are written for an IBM 360/370 computer. There are 28 independent programs covering the previously discussed areas. In addition to the statistical analyses discussed, there exists a comprehensive mixed designs (repeated measures) analysis of variance, Kendall's rank correlation and coefficient of concordance, probit analysis, cosine fitting, and discriminant analysis program. These programs, however, are not listed in the user's manual.

3.2.4 Other Non-integrated Systems

3.2.4.1 F4-STAT - FORTRAN IV STATISTICAL SYSTEM

F4-STAT (Beaton, 1973) is a system of Fortran subroutines, but with some differences. Although the user must supply the main program there are 'utility' subroutines that permit, for example, the printing out of headings and error messages, or the ability to read data cards and make assignment statements such as X(1) = COL(17). Thus, the user is not burdened by clumsy 'format' statements.

Statistical analyses are comprehensive including exact analyses for unbalanced designs and regression routines that allow for error in the independent variables. F4-STAT is written in Fortran II and 360 assembler and is implemented on an IBM 360/65, containing approximately 14,000 source statements. According to Beaton (personal communication) F4-STAT is transportable but the "documentation is not really good enough for wide distribution." Also, he feels F4-STAT is designed for use by competent programmers.

Nevertheless, notable points are that F4-STAT lies somewhere between strictly 'subroutine sets' and 'organized collections of programs' as is shown above. Also, something that should be mentioned is that F4-STAT uses the 'special matrix operators'

(Beaton, 1964) in the subroutines. The philosophy behind this is because high speed computers are available, statistical analyses (regression, analysis of variance, etc.) are capable of being stated in the mathematical model form (for analysis of variance dummy variables are created) and then the exact solutions are obtained. Computational formulas that originated for desk calculators are inadequate for unbalanced designs. Beaton feels the emphasis should be placed on specifying the mathematical models rather than on special techniques developed for calculators.

3.2.4.2 IMSL - INTERNATIONAL MATHEMATICAL AND STATISTICAL LIBRARIES

IMSL is an organized collection of Fortran subroutines covering both mathematical and statistical applications. There are three libraries, one for each of the following computers: IBM 370/360, UNIVAC 1100 series, and CDC 6000 series. This evaluation incorporates all updates since 1971 when IMSL first became available.

There are approximately 150 subroutines dealing with statistics covering most areas fairly well. Multiway tables

are not possible for means and frequencies; and the general linear model analysis still requires the user to create his own dummy variables, which is impractical. The sales literature indicates that canonical analysis, cluster analysis, discriminant analysis, and factor analysis are available in the 1973 version, but this author was unable to find them.

Besides what one should definitely expect a good subroutine to offer, i.e., double or single precision, liberal comments, specific instructions, etc., IMSL does provide algorithms that are 'up to date' and gives some theory for each algorithm along with an example. An actual listing of the subroutines is not provided so evaluation of the programming cannot be assessed. IMSL is being constantly updated (last update was approximately 400 pages) and is available for $960 per year.

3.2.5 Miscellaneous Non-integrated Systems

The Scientific Subroutine Package (SSP) has been a popular system of Fortran subroutines (also available in PL-1). It has been indicated that this package may no longer be supported by IBM. Information regarding SSP has been requested from IBM but has not been acknowledged to date.

The MANOVA (univariate and multivariate analysis of variance and covariance) program (Cramer, 1973) although not a system, is of such comprehensiveness that it needs mentioning. To illustrate its comprehensiveness let us quote Cramer, "....it is possible to analyze every design in *Experimental Designs* by Cochran and Cox....". To illustrate its ease of use for, say, a two-way nonorthogonal design all one has to do is put S, D, SD on the control card to signify that the order to be employed is: S ignoring D, D eliminating S, and SD eliminating both S and D. The MANOVA program is implemented in the P-STAT system (see Section 3.1.3). MANOVA, however, can be obtained separately for $50 along with the user's manual (Cramer, 1967).

Lee (1971) provides an organized collection of Fortran programs, but without data management; for multivariate analyses. It is

mentioned because the user's manual also acts as a text for multivariate analysis.

OSIRIS is a large collection of programs developed at the Institute for Social Research, University of Michigan. A user's manual has been requested but has not been received to date.

See Myers (1969) and Schucany et al. (1972) for some other miscellaneous systems.

CHAPTER 4

ON-LINE SYSTEMS

The definition of an on-line statistical system like the definition of a batch system need not be defined in computer terminology within the context of this thesis. It is sufficient to say that on-line systems (synonyms are: conversational time-sharing, interactive computing, or some combination of all three terms) have the following points: 1) the user communicates with the computer using a teletype connected to the computer, usually via telephone lines, and punched cards are not used; 2) there is true line by line interaction between the user and the computer. In the process of communicating the user should not be aware of any delays in computer responses, i.e., printed responses should appear almost instantaneously to the user. Some systems permit the user to access the raw data from a previously defined tape or disk, or to direct voluminous output to a high speed line printer.

Like the batch systems, the on-line systems are similarly

divided into integrated and non-integrated sections, and they are presented in alphabetical order, except SOL, which is presented in Chapter 5.

4.1 Integrated Systems

4.1.1 IMPRESS - INTERDISCIPLINARY MACHINE PROCESSING FOR RESEARCH AND EDUCATION IN THE SOCIAL SCIENCES

The IMPRESS system is "a computer system for selective retrieval and analysis of large data files in the social sciences" in use at Dartmouth College. This evaluation utilizes three user's manuals: *The IMPRESS Manual, Analyzing Tabulations with IMPRESS,* and *The IMPRESS Primer* (see reference under IMPRESS).

4.1.1.1 Statistical Analyses

The 1VAR procedure gives descriptive statistics including higher moments, quartiles, and histograms, but N-way tables of means are not available. The MARG procedure makes one-way frequency distributions. The XTAB procedure can produce up to eight-way frequency tables including percentages. The measures of association that can be optionally obtained with XTAB are comprehensive and exhaustive: Somer's coefficients, Goodman and Kruskal's gamma measure, tau and lambda; Yule's Q and Y measures; the phi coefficient; Pearson's C

statistic; Cramer's V; Tschuprow's T; and Kendall's tau family; and, of course, the chi-square.

The CORREL procedure prints out zero-order or Nth-order correlation matrices, and the PLOT procedure produces bivariate scattergrams. The REGRES and STEP procedures perform multiple regressions and stepwise multiple regressions. Multivariate analyses are limited to a factor analysis procedure.

Analysis of variance, two-sample parametric and non-parametric tests are absent. Some miscellaneous procedures are: path analysis; time-series analysis; Guttman scaling; item analysis; as well as three other special procedures - EFFECT, IDEA, and STD - to study relationships between variables.

4.1.1.2 Data Management

Only numeric data is permitted in a fixed format. The data, unfortunately, must first be entered as a user data file on disk before IMPRESS can access it. The structure is variable by case for all procedures. Storage and retrieval of files is adequate but cumbersome. Labeling

of variables and their values is possible.

By using commands similar to those in the BASIC computer language one can create new variables or modify old ones. Facilities also exist to recode, select, delete, and add cases. Designation and handling of missing values is possible. The facility to take random samples, with or without replacement, from the user's data base is possible. The facility to generate data files from the uniform and normal distributions also exists.

Good labeling is provided, but the output is not rigorously defined. Output from IMPRESS is not capable of being used as input. Limitations on the number of observations are not mentioned and a severely restricting limit of eight variables is allowed for all the statistical procedures as indicated in all the user's manuals. It does state, however, (Chapin and Myers, 1972) that this limit has been increased to twenty, but is left at eight in the user's manuals to prevent indiscriminate use.

4.1.1.3 Ease of Use

The IMPRESS Primer is a successful attempt to present a

sophisticated data analysis system to an inexperienced user. It is, however, oriented toward sociological jargon and examples, as are the rest of the manuals. The *Analyzing Tabulations with IMPRESS* is equally effective as a primer and besides just saying 'how' to run procedures, it explains in simple terms the statistical concepts and how to apply them. It should be noted that it is elementary, being aimed at the novice in statistics and computers. *The IMPRESS Manual* primarily contains information on data management procedures and, unfortunately, does not come up to the readability or simplicity of the other two manuals.

Besides the manuals there are instructional messages which are available on-line to those using IMPRESS. The manuals could provide more examples. IMPRESS uses many esoteric statistical procedures (e.g., see those stated for the crosstabulations), but provides no references. Also, as stated before, the output is not rigorously defined statistically in the user's manuals.

Commands are in a natural language. Effective controls appear to exist in handling errors and in the actual running of IMPRESS with regard to helping the user, controlling output, etc.

4.1.1.4 Other Considerations

IMPRESS is written in the time-sharing language BASIC. Unfortunately, IMPRESS, because of the language and sophisticated file structures, is not transportable to the more common computers. IMPRESS is implemented on a Honeywell G-635 computer system. Chapin and Myers (1972) state, "the trade-off was between a powerful system written in BASIC or a trivial one written in FORTRAN."

Information concerning IMPRESS comes under the heading 'Project IMPRESS'. *The IMPRESS Manual* lists a staff of 33 involved in the project. *The IMPRESS Primer* lists 27 publications including the three aforementioned manuals. Needless to say, development of IMPRESS is not wanting. Perhaps future goals will be to make IMPRESS transportable and to add more statistical procedures that would benefit non-sociologists as well. This author, however, might be doing the IMPRESS project a disservice by neglecting some facts which serve to enhance IMPRESS. The IMPRESS system provides a library of surveys that students or faculty can access on-line. The 'codebooks' (a description of what and where the variables are and what values they can have) are available. Thus, users are encouraged to analyze other people's data and draw their own conclusions. At present

there are 51 surveys with an average of approximately 2000 observations and 240 variables per survey. All these surveys have been edited and analyzed, but in the process the variables have been extensively labeled; values within variables have been grouped into meaningful categories, etc. Thus, all the user has to do is run the statistical procedures without having to process, edit, manipulate, etc. the data. Also, the user does not have to abide by the groupings set up and can change it if he so desires. The accessibility of these surveys to assist in teaching sociology is, indeed, an innovative step and the IMPRESS manuals are geared around this set-up.

4.1.2 MIDAS - MICHIGAN INTERACTIVE DATA ANALYSIS SYSTEM

This evaluation is based on the user's manual (Fox and Guire, 1973) and a paper by Fox (1973).

4.1.2.1 Statistical Analyses

Facilities exist for providing one to N-way tables of means, standard deviations, standard errors, minimum and maximum values. Other descriptive measures are possible but the user must employ cumbersome methods to obtain them. Histograms are available. N-way frequency tables are possible as well as percentages. For the two-way tables various tests of association can be obtained.

Simple and partial product moment correlation matrices are capable of being produced. Simple and multiple regression procedures, including stepwise, are available as well as polynomial regression. The predicted and residual values from the above may be saved and used in other procedures if desired. Scatterplots are also available. Non-parametric correlation matrices include Spearman's and Kendall's statistics. Weighted and non-linear regressions are not available.

Multivariate analyses contain procedures to do principal components analysis, factor analysis, discriminant analyses, including stepwise, and canonical correlations. The principal factor solution is employed in the factor analysis. Orthogonal and oblique rotation schemes are available. The non-stepwise discriminant analysis can also produce quadratic discriminant functions. For the above procedures various types of output can be saved, e.g., factor scores for factor analysis, posterior probabilities for the discriminant analysis, etc.

Analysis of variance is limited to one-way designs but includes two non-parametric tests, viz., Kruskal-Wallis and the median test. Special two-sample tests are available. The unpaired tests are: t-test, Mann-Whitney, median, and Kolmogorov-Smirnov. The paired tests are: t-test, Wilcoxon's, and the median test.

Procedures exist to generate random deviates from the common distributions. Some of the distributions are: normal, F, exponential, gamma, beta, chi-square, student, uniform, and Poisson.

Miscellaneous procedures are time-series analysis, profile

analysis, and two-stage least squares.

4.1.2.2 Data Management

MIDAS uses the variable by case structure although input may be entered 'variablewise' if desired. Alphanumeric data can be read but is translated to 'numbers' internally so it essentially loses its identity. Free-field formats are permitted, with commas being necessary to separate numbers instead of blanks. Input may be from cards, tape, disk, or directly from the teletype.

Variables can have up to eight character labels, but no value labels are possible. Case labels are not possible, but MIDAS automatically numbers the cases consecutively (i.e., 1 to N). The user can create and use multiple datasets. Although the ability to merge and match datasets is mentioned, this author is unable to understand the jargon that MIDAS employs. Many forms of output can be stored as was seen in the statistical analyses section.

MIDAS defines two types of variables and the user must indicate what type his variable is on input, i.e., categorical

(discrete) or analytical (continuous). If one wants to transform data then, unfortunately, either of two procedures must be used depending on the type of variable. Arithmetic expressions are possible, but only one operator is available per command, i.e., it would take two commands to produce Y = A + B + C. Use of IF....THEN, GO TO...., and MACROS are unavailable. However, MIDAS provides functions other than the usual square roots, logs, etc., such as minima, maxima, ranges, sums, percentiles, etc. Good procedures exist for selecting and deleting variables or cases for analyses. Provisions exist to designate missing values, but the user has no say on how they are to be used in the statistical procedures.

Output appears adequately labeled with facilities provided so that the user can add his own labeling. Most of the statistical procedures provide options for good control of output and facilities exist so that output can be either 70, 105, or 132 columns, depending on the make of teletype, or if output is directed to a line printer. The limitations are not stated within each procedure, but it appears as if all procedures are liberal (for the multivariate procedures the product of the cases and variables cannot exceed 131,072).

4.1.2.3 Ease of Use

The output is not rigorously defined, but references are given for all the statistical procedures. The number of examples are minimal. Otherwise, the manual is explicit and procedures are relatively independent. The terminology used throughout the manual, e.g., terms like strata identifiers, does not help the novice. Some procedures are very difficult to understand, e.g., the COMPUTE procedure.

Some strong points are MIDAS's ability to do many analyses in parallel with the use of the STRATA and BYSTRATA commands. Also, in the on-line mode good error control exists as well as facilities for prompting and explaining commands to the user.

4.1.2.4 Other Considerations

MIDAS is written in both Fortran and assembler languages (90% Fortran) and runs on an IBM 360/67 virtual memory computer. MIDAS also operates under the Michigan Terminal System (MTS). The above points limit MIDAS's transportability. MIDAS is also usable in a batch mode.

4.1.3 P-STAT

The P-STAT system (see Section 3.1.3) is also available on-line. Facilities exist to have input from the teletype, cards, tape or disk and to direct output to the line printer; or direct error messages to the teletype but other output to the line printer. For a full description of P-STAT see Section 3.1.3. (Note: on IBM 360/370 computers P-STAT would normally operate under TSO (Time Sharing Operating System) and unless the computer centre has large amounts of core for time-sharing it may be impossible to run P-STAT.)

4.1.4 Miscellaneous Integrated Systems

Joyce (1972) discusses his on-line system, SSTAT, that is "built around an ASCOP-like command language" (ASCOP is evaluated in Section 3.1.1), and runs on a PDP-10 computer.

TROLL - Time-shared Reactive On-Line Laboratory is a statistical system developed at MIT mainly for economic research. Most procedures are 'econometric' but standard regressions and graphs are available along with simple univariate statistics, e.g., means and standard deviations. TROLL runs on an IBM 360/67.

Myers (1969) and Anderson (1971) list some on-line statistical systems. Some that looked general enough were DATANAL, TROLL, ADMINS, and TRACE. Documentation was requested for each of them, but the only documentation received to data was the TROLL system.

4.2 Non-integrated Systems

4.2.1 ISIS - INTERACTIVE STATISTICS INSTRUCTIONAL SYSTEM

This evaluation is based on the user's manual (Brown, Goodin, Meeter, and Soller, 1972).

4.2.1.1 Statistical Analyses

Comprehensive univariate descriptive statistics are available, but N-way tables are missing. Four-way frequency tables are possible with the ability to analyze two-way tables with the chi-square test.

The simple, multiple, and polynomial regression programs allow plotting and storing of residuals. Scattergrams, product moment correlation, and rank order correlations are available as well as partial correlations.

A factorial analysis program exists using unweighted cell means in unbalanced designs. Analysis of covariance, hierarchical designs, and multiple range tests are absent. Two sample parametric and non-parametric tests are not available, except for the Wilcoxon signed-rank test.

Program GENDAT can generate random observations from any one of fifteen common distributions. A test of normality also exists. There are no multivariate statistics or miscellaneous statistical programs.

4.2.1.2 Data Management

Alphanumeric data as well as free-field is permissible. Input is the variable by case structure and can be stored and retrieved from external data files with relative ease. Labeling of variables is not possible.

Separate programs exist for manipulating data files including sorting, listing, editing, dividing, and selecting data, but they are limited. For example, to divide one data file up into smaller subsets (maximum of 20) the division can be made on the basis of only one variable. Programs for transforming data and merging and weaving files are mentioned but no details are given for these procedures. ISIS was designed so that output from one program may serve as input to another program. Other than the above no more data manipulation procedures are possible. Missing values are input as question marks (?), but not all programs can handle missing values. For those programs that can handle missing values, no mention is made

of what is done when they are encountered.

Output labeling is adequate with the user being able to add his own for some programs (e.g., scattergram). Procedures exist for controlling voluminous output. The output, however, is not defined.

For an on-line, instructional system the limitations are varied but liberal. For the simple descriptive statistics program the number of observations allowed is stated as "practically unlimited", but only fifty pairs, however, are allowed for the Wilcoxon matched-pair test.

4.2.1.3 Ease of Use

The user's manual for ISIS is stored on disk and printed out on computer paper. Instructions for communicating with the computer and requesting the desired program are clear, as well as the logistics of handling typing errors, making data files, etc.

Each program is designed to be self-instructional when used on-line and consequently only a few programs discuss detailed

usage in the user's manual (where all that is done is to print out what would occur on-line with the user, if he was unsure what to type in). The programs where the self-instruction messages are printed out are fairly easy to understand for simple programs but would be very difficult to follow for, say, the factorial analysis program. Again, no mention is made as to what statistics are actually computed. For production purposes the lengthy instructional messages would prove to be frustrating and time-consuming. Complicating features are that most programs require the input data to come from a previously defined external file and that not all programs can handle the missing data feature previously described.

4.2.1.4 Other Considerations

ISIS was originally developed at Florida State University and was designed primarily for instructional use.

The programs were written in GE time-sharing Fortran and re-written in Fortran IV by the University of Calgary to run on the CDC 6400 KRONOS time-sharing system. Additional non-parametric tests are under development.

4.2.2 RAX - REMOTE ACCESS STATISTICAL SYSTEM

RAX is strictly an IBM system. The version presently being evaluated was developed by Tauchi (1968). Information on the latest status of RAX has been requested from IBM but has not been received to date. As late as May, 1971, however, the RAX system had not changed appreciably (Kelly, 1970).

4.2.2.1 Statistical Analyses

Some descriptive statistics are available but medians, modes, skewness, kurtosis are not available; neither are N-way tables for means. One is able to generate two-way frequency tables and then perform a chi-square analysis, but N-way frequency tables are not possible.

Simple product moment and Kendall correlation coefficients are available as well as a simple linear regression, with the ability to test for extreme residuals, and bivariate plots. Multiple, stepwise, and polynomial regression complete the regression programs. Multivariate analyses programs are available. They are: canonical correlations, factor analysis, and discriminant analyses.

Standard factorial designs are available using Hartley's algorithm (1962). Unfortunately, the write-up also uses Hartley's 'operator' terminology which would prove incomprehensible to a user and for practical purposes is unnecessary to explain in a user's manual. Hierarchical, unbalanced designs, and multiple-range tests are absent, although paired and unpaired t-tests are available.

Some miscellaneous statistical programs are triple exponential smoothing and probit analysis. It should be noted that RAX is built around the Scientific Subroutine Package (SSP).

4.2.2.2 Data Management

Only numeric data is permitted in RAX but it can be accessed from both disk or tape in either free-field or fixed (but variable) format. All input is done with the variable by case structure, and facilities exist within the statistical program to edit (replace, add, delete, etc.) data. Manipulation of data files is limited to fourteen pre-defined transformations, e.g., square roots, X + Y, etc.

A very serious limitation of RAX is its inability to designate

and handle missing values. If a system uses the variable by case structure then missing values are bound to occur for real data in such a matrix, and unless these can be designated and handled accordingly, the system is very limited.

Output is adequately labeled and defined. Tauchi's version has unreasonable limitations (i.e., regardless of the program, the input matrix can have a maximum of twelve variables and one hundred observations); Kelly's version has thirty variables by two hundred observations - better, but still restrictive.

4.2.2.3 Ease of Use

The user's manual gives a lucid and easy to understand introduction along with an example. Each program is clear and easy to use. Examples exist for each program as well as references where necessary. The facility exists to use either standard ASR-33 teletypes or the IBM 2741's.

A good feature is the 'SOS' command. When RAX asks the user to reply, and if the user does not understand, then if 'SOS' is typed by the user, instructional help is printed back.

4.2.2.4 Other Considerations

RAX is written almost entirely in Fortran IV. To run RAX (Tauchi's version) requires an IBM 360/40 or 50 and requires the remote access computing system with 128K. Kelly's version needs an IBM 360/67 model with 256K.

4.2.3 SSIPP - SOCIAL SCIENCES INTERACTIVE PROGRAMMING PACKAGE

SSIPP was developed at Beloit College, Beloit, Wisconsin as an instructional system for statistics labs, with instructions for use being obtained when actually running the program. This evaluation is based on the user's manual (see reference under SSIPP).

4.2.3.1 Statistical Analyses

Three programs are available for calculating moments and frequency distributions. N-way tables of descriptive statistics are not possible. Multidimensional frequency tables are not possible, but a chi-square program is available to analyze two-way tables.

Regression analyses of all types are absent as well as multivariate statistics. A program for product moment correlations is available. There are no analysis of variance programs except for the one-way design. For unpaired data for two-samples the t-test, Kolmogorov-Smirnov, and Mann-Whitney tests are available.

Distribution theory is SSIPP's forté. For the common

distributions, programs are available to find critical values given the level of significance, and vice-versa, for one or two-tailed tests. Computations of the power of the tests is also possible as well as computations of minimum sample sizes needed for given powers. Graphs can be generated for the binomial distribution increasing the sample sizes to illustrate the central limit theorem.

Forty-four individual programs cover the previous areas with over eighty percent applied to distribution theory.

4.2.3.2 Data Management

Free-field numeric data is required. Input is not variable by case and depends on the requirements of each individual program. Use of data files is not possible and no features exist for manipulation and transformation of data. The output is not rigourously defined. Limitations when mentioned are small and no provisions exist for missing values.

4.2.3.3 Ease of Use

The user's manual is self-instructional on-line. For the novice using a program for the first time the on-line instructions would be useful. Although each program asks if the instructions should be typed out, the mandatory commands needed for actual use are too long and unnecessary. For example, the one-way analysis of variance program types "TYPE NUMBER OF SAMPLE VALUES FOR RANDOM SAMPLE X", where X is the number of the sample. If one has, say, five samples, this statement would be repeated five times even before the data values are typed in. Even for instructional use the amount of instructional output is verbose. There is no discussion of error messages or how to correct errors.

4.2.3.4 Other Considerations

The user's manual was obtained from the University of Calgary. The author, upon discovering where SSIPP was developed (Beloit College, Wisconsin), requested information directly from Beloit College. Beloit College did not send a user's manual, but sent a brief list of available programs. Besides having computer assisted teaching programs and

miscellaneous programs in other disciplines (e.g., psychology) they mention the existence of a multiple regression program and a simple two-way tabulation program which the University of Calgary did not have in their user's manual.

Beloit College says, "Almost all SSIPP programs and subprograms are written in Fortran IV....". The programs average less than 8K words of core which is good for transportability to small computers.

The SSIPP project is no longer funded, but the programs are still obtainable from the Beloit College Computing Centre, with costs they say, "[only for] labor, material, and shipping."

4.2.4 STATPACK2 - AN APL STATISTICAL PACKAGE

This evaluation is based on the user's manual (Smillie, 1969). In the author's environment, and in others, STATPACK2 is implemented but no updates have taken place since 1969. Like other systems, except subroutine sets, the user should not be expected to know the computer language involved, whether it be Fortran, Cobol, PL-1 or APL. Thus, the approach in evaluating STATPACK2 is from the user's manual presented and not from what APL may be capable of doing if a user were fluent in the APL language.

4.2.4.1 Statistical Analyses

Programs are available for descriptive statistics. N-way tables for means are not possible. Multidimensional frequency tables are not available although one and two-way tables are. The chi-square statistic for two-way tables is available.

Simple and partial correlations are possible along with a stepwise multiple linear regression program. An analysis of residuals is possible. Non-parametric correlations and other types of regression programs are not available. Multivariate statistics are absent.

A complete factorial, cross-nested design, and a one-way design are all the analysis of variance programs available. There are miscellaneous programs involved in matrix operations and operations research.

4.2.4.2 Data Management

Input is from files or directly from the teletype. Format is free-field and is not variable by case, except for correlations and multiple regression programs. Although data can come from files there are no provisions for their manipulation. Some programs can handle missing values. For those programs that handle missing values the designation of them is impractical. For example, the simple correlations program, SCORR, requests that all missing values be input as any negative number, and the one-way analysis of variance program, ANOVA2, requests that missing values be input as zero.

Labeling of output is almost totally absent and the output is not rigorously defined. There is no discussion of limitations.

4.2.4.3 Ease of Use

A user without an acquaintance with APL would be unable to use the simple descriptive statistics program if he relied

on the user's manual. The first few pages describe the organizational structure of STATPACK2 using APL computer jargon (e.g., workspace, function, etc.) which does not promote understanding. Only one example is done and it is not helpful. To find the mean and standard deviation of a set of numbers one has to prepare the data separately using the APL assignment statements and then submit this to the particular program. For all the programs the user must assign values to 'vectors', 'arrays', etc., but nowhere does it say how to do it. In short, the user's manual has no examples (except one), is not explicit, and is difficult to understand.

4.2.4.4 Other Considerations

The computer statements are listed for each program in the user's manual. The fact that APL is a powerful, yet compact language, cannot be argued, as can be seen by program MVSD where the mean, variance, and standard deviation are calculated for M variates and N observations in two program statements. For this algorithm, however, it does not say whether or not each variate has to have the same number of observations. Similar feats of compactness are achieved for other programs. It may be efficient for running the programs, but in terms of

understanding what is being done, it is difficult. Not one comment is given for any of the programs. The programs are well below the minimum programming standards (see Section 4.2 on Documentation).

To run STATPACK2 requires that the computer installation has APL. To run APL requires IBM teletypes (usually 2741's) that need special APL typeballs for its set of characters.

Smillie in his introduction says, "STATPACK2 is a package of APL programs...". However, it is the author's feeling that from a typical user's point of view the programs just described are not programs, but are really APL subroutines and belong under the category of 'subroutine sets'.

4.2.5 Miscellaneous Non-integrated Systems

Kilpatrick (1972) presents two systems in his paper: CCSS (Conversational Computer Statistical System) written in Fortran IV and APL-SS (APL Statistical System). The APL system appears more user-oriented than the APL system in Section 4.2.4. Many universities have some type of conversational statistics programs. Carleton in Canada

and Oregon State in the U.S.A. have sent some literature on their developments (this is not referenced). Anderson (1971) lists four commercial systems.

CHAPTER 5

THE SOL SYSTEM

5.1 History and Rationale for Development

An increasing dissatisfaction with batch packages of programs by both the author and users led to the initial development of SOL. The process of keypunching and using job control cards; the process of coding the control cards within the canned program; the process of coding the options; and finally assembling the data cards amongst the above was rigid, difficult, and time-consuming. If the control cards for both the computer system and statistical program were not specified to the letter and placed in the correct card column, the run ended prematurely producing wrong, incomplete, or no results. The error messages were usually in computer jargon and provided little assistance in informing the average user of the exact nature of the problem. One of the resident statisticians usually was involved in correcting the problem, and spent, for routine statistical problems, more time on computer preparation and problems than necessary.

The above process has taken up to a week for a routine problem of finding descriptive statistics and producing a frequency distribution on a sample of data containing about one hundred observations.

The coding of data into rigid formats and then keypunching the data or submitting the data to keypunching services and waiting was too time-consuming and restrictive for most users who had small amounts of data - even apart from the process of submitting and running the job.

It should be noted that prior to the implementation of SOL in 1970 none of the major statistical systems emphasizing the role of data management and catering to the novice in statistical computing was available in the author's environment. Those packages or programs available were not designed to initiate the novice into analyzing his own data. For example, a typical program would say, "Keypunch the number of observations in columns 1-5." The user has no idea that 'right justified' is implied; or even if 'right justified' is stated, what does it mean? Thus, aside from using a program, understanding how to use it was just as much a problem, and, unfortunately, the statistician was involved in explaining non-statistical problems more often than necessary.

It was recognized that most of the problems involved small or moderate amounts of data and that many types of statistical analyses requested were of a re-occurring nature. For this type of problem the following points of need emerged: speed up the process; provide a statistical system that is easy to understand and use; minimize the role of the statistician as much as possible from becoming involved in the 'computer aspect'. These points led to the design of an on-line conversational statistical system.

5.2 SOL - STATISTICS ON-LINE

The SOL system (Rollwagen, 1973) was implemented in April, 1970 and can be classified as an organized collection of programs. Like the previous systems it is evaluated along the same guidelines, however, some of the objectives of the system are mentioned throughout the evaluation.

5.2.1 Statistical Analyses

Descriptive statistics are available as well as chi-square tests for frequency data. Both areas are comprehensive in what they give (e.g., expected values are printed out along

with chi-square contribution of each cell and Fisher's test for 2x2 tables for frequency data), but N-way tables are not possible for means or frequency data. Reasons for not including N-way tables were lack of computer core, and it would have made the package harder to use. If N-way tables were used then input would have had to be the variable by case structure; which would have required the user with small amounts of data to organize it in a rigid manner not necessarily conducive to ease of use, or to his collection technique.

A comprehensive simple linear regression is available including confidence limits and bivariate plots. Multiple regression is possible and because the program is interactive, decisions can be made after each fit as to what variables to include in the equation. A polynomial regression is also written along the same lines. Tests of significance and examination of residuals are possible for each of the above.

A one-way analysis of variance program exists and has options to do the Kruskal-Wallis non-parametric test and five multiple range tests (i.e., LSD, Dunnett's, Duncan's, Student-Newman-Keul's, and Tukey's). Factorial analysis programs are available. There are three separate programs: one observation

per cell with the ability to estimate missing values; equal observations per cell; and unequal observations per cell using Snedecor's harmonic means solution. All programs have options to estimate the variance components and residuals. Analysis of covariance and hierarchical designs are absent.

SOL provides two-sample parametric and non-parametric tests. Equal and unequal variance (using Welch's procedure) t-test programs are possible as well as Bartlett's test of homogeneity of variances. The non-parametric tests are included as options within the t-test programs. This facilitates and encourages their use. Examples of the non-parametric tests are: Wilcoxon-Mann-Whitney, Kolmogorov-Smirnov, sign, and Wilcoxon matched-pair test.

The chi-square and Kolmogorov statistics provide goodness of fit tests for five common distributions: normal, exponential, uniform, binomial, and Poisson. A program is available to generate random numbers from 1 to N. Generation of distribution functions is not possible.

Some miscellaneous programs are: one-sample t-test, rankit transformations and test for normality, least-squares fit to a hyperbolic equation, a comprehensive probit analysis program,

a program that simulates the operations of a desk calculator, and a simple transformation program.

A notable missing area is multivariate analyses. Because multivariate analyses usually involve large quantities of data, and usually require extensive editing before statistics can be computed, the data is best punched on cards and stored on disk or tape. SOL presently cannot access data files, that being the reason for no multivariate statistics. Also, the original design philosophy was to include common procedures and for most people multivariate statistics are still esoteric. The underlying philosophy was to make the common statistical procedures as comprehensive as possible and to educate the user on their best use.

5.2.2 Data Management

The variable by case structure is not used, except in the regression programs, for reasons previously stated. For the amount of data usually presented to SOL and the types of analyses required the input structure adopted (see Fig. 5.2.1) has proved to be the best to date in terms of understanding and ease of use for the majority of the users of SOL. There is no doubt, however, that the variable by case structure is

the best system for batch processing or for large amounts of data containing many variables. Input is a flexible free-field format with data being entered directly from a teletype. Use of data files as input is not available and consequently storage and retrieval and manipulation of files is absent. No provisions exist to alter data except for a few basic transformations available in most programs such as square roots, logs, etc.

The output labeling is abbreviated to speed up the printing process (standard ASR-33 teletypes only print at ten characters per second), but output is rigorously defined. The limitations are clearly stated for each program and are liberal considering all data is to be entered by teletype input. The t-test allow 400 observations per sample; the multiple regression allows ten variables and up to four hundred observations; the factorial analysis allows up to six factors and fifty replications per cell.

5.2.3 Ease of Use

Ease of use is one of SOL's stronger points. The user has only to read three pages before he can use SOL. Entry into the particular statistical program is easy, as can be seen in

Figure 5.2.1 - Examples

Following are two annotated examples. All user input is boxed. To establish communications with the computer system, the teletype is turned on and the appropriate phone no. dialed.

```
HEALTH SCIENCES COMPUTER SYSTEM
WED 21 NOV 73, PORT 08

JOBNO.. 7001

?? RUN ST11
 ST11 IN

 TITLE.. AN EXAMPLE OF ST11

 OPS... MSFH
    LCB.. 4.5
    CW... 1
    L OR B.. L

 DATA...
 7.8 9.3 10.2 6.5
 8.4 9.3 10.1 7.4
 8.9 9.8 6.8 8.2
 5.8   5.5 7.9 6.3
   6.9 4.9/

 N    =          18

 MEAN =       7.778
 VAR  =       2.641
 SD   =       1.625
 SDM  =       0.383
 CV   =       0.209

 MED  =       7.850
 RAN  =       5.300

 SKEW =      -0.085
 KURT =       1.693
```

The system acknowledges the phone call and requests the user's job number. Each line of user input must end by pressing the RETURN key.

Anything may be typed in the TITLE, but usually something identifying the data.

By typing MSFH the

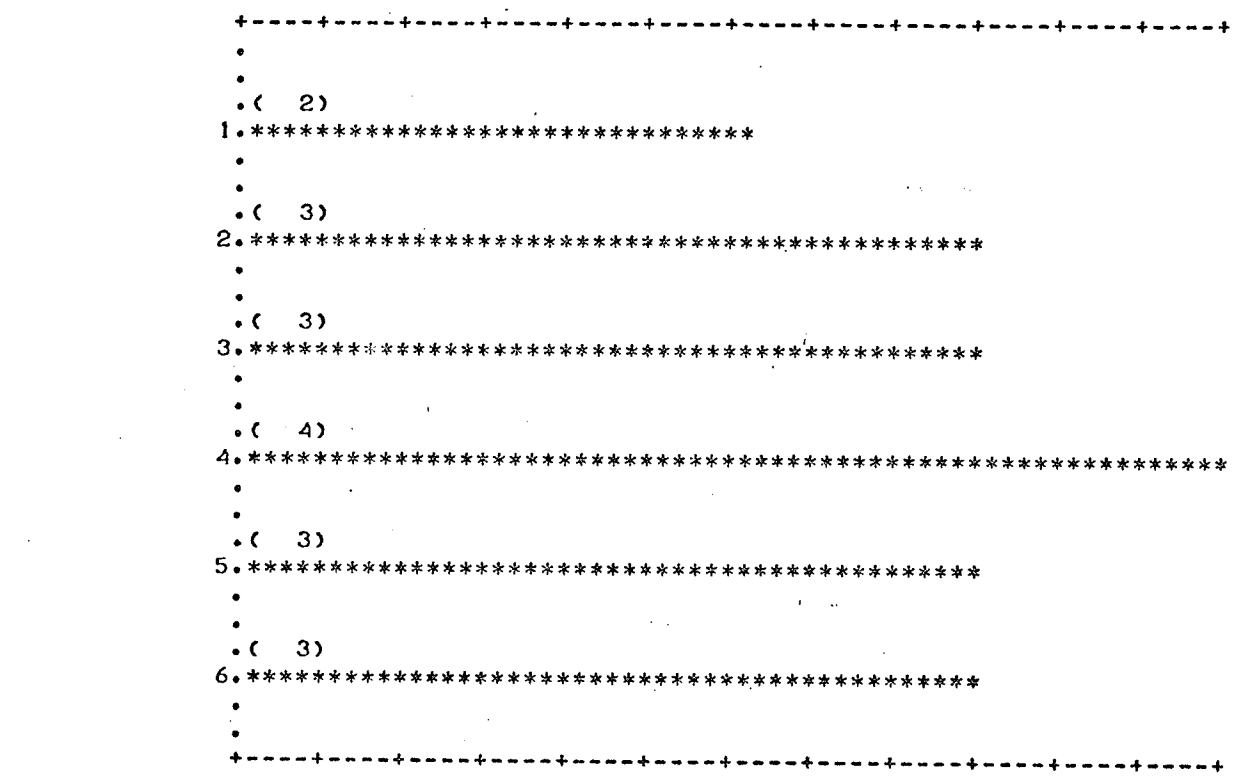
1. *median and range,*
2. *skewness and kurtosis,*
3. *frequency distribution (LCB=lowest class boundary, CW=class width),*
4. *and histogram (L=line graph, B= bar graph) will be output.*

The data values are typed in free format and terminated by typing a slash (/).

```
FREQUENCY DISTRIBUTION

CLASS    CLASS BOUNDARIES      FREQ   REL FREQ   CUM FREQ   CUM REL FREQ

  1       4.500     5.500       2      .111         2         0.111
  2       5.500     6.500       3      .167         5         0.278
  3       6.500     7.500       3      .167         8         0.444
  4       7.500     8.500       4      .222        12         0.667
  5       8.500     9.500       3      .167        15         0.833
  6       9.500    10.500       3      .167        18         1.000
```




```
RU FINISHED..Y
```

The user is finished with program ST11.

```
 ST11 OUT
ST11   USED FOR 00.07 HRS
```

```
?? RUN ST41
 ST41 IN
```

User now wishes to use ST41.

```
TITLE.. AN EXAMPLE OF ST41
```

```
TEST..AK
   MRTEST..Y
   TFD... N
   NO. GROUPS.. 5
```

By typing AK the analysis of variance and the Kruskul-Wallis test are both requested. The multiple-range tests are also wanted. No transformation is required on the data. Are 5 groups.

```
GRP 1..
75 70 67 75 65 71 67 67 76 68/
```

Each group is typed separately, and terminated by a slash (/).

```
GRP 2..
57 58 60 59 62 60 60 57 59 61/
```

```
GRP 3..
58 61 56 58 57 56 61 60 57 58/
```

```
GRP 4..
58 59 58 61 57 56 58 57 57 59/
```

```
GRP 5..
62 66 65 63 64 62 65 65 62 67/
```

GROUP	N	MEAN	SD	SDM
1	10	.7010000E 02	.3984671E 01	.1260063E 01
2	10	.5930000E 02	.1636444E 01	.5174893E 00
3	10	.5819999E 02	.1873842E 01	.5925608E 00
4	10	.5800000E 02	.1414213E 01	.4472136E 00
5	10	.6410000E 02	.1791909E 01	.5666513E 00
TOTAL	50	.6194000E 02	.5195784E 01	.7347949E 00

```
ANALYSIS OF VARIANCE

SOURCE      DF        SS          MS          F

BETWEEN      4     1077.312     269.328     49.368
WITHIN      45      245.500       5.456
TOTAL       49     1322.812


MULTIPLE COMPARISON

                  DUNCAN
          LSD     SNK
          DUNNETT TUKEY     P = NO. OF MEANS FOR RANGE TESTED

 4: 3      0.191    0.271    2
 4: 2      1.245    1.760    3
 4: 5      5.840    8.259    4
 4: 1     11.584   16.382    5

 3: 2      1.053    1.489    2
 3: 5      5.648    7.988    3
 3: 1     11.392   16.111    4

 2: 5      4.595    6.499    2
 2: 1     10.339   14.622    3

 5: 1      5.744    8.123    2


KRU-WAL TEST

 GROUP      RANK SUM

    1        450.00
    2        196.50
    3        138.50
    4        131.50
    5        358.50


 CHISQ =   38.437
 DF    =    4


 RU FINISHED.. [Y]


 ST41 OUT
ST41   USED FOR 00.08 HRS


?? [LOGOUT]
LOGGED OUT
```

These values may be compared directly with the tables for significance.

The Kruskal-Wallis test and the chi-square value to test for significance.

The user is finished.

The user logs out. Teletype shuts off.

the examples. The user's manual is explicit as to input, output, and limitations, and provides many examples. Error messages use codes (again to speed up the process and minimize core storage). Unfortunately, one cannot go back and correct data on previous lines, although corrections can be made on the line of data being entered at the time. The user cannot cause the system to lose control.

5.2.4 Other Considerations

Since SOL's implementation in 1970, the following changes have been made: make the programs easier to use; increase the amounts of data that can be input; add non-parametric tests; prepare the programs to handle data files; and last but not least - improve the readability of the user's manual.

The computer programs are well documented and are written entirely in Fortran IV. The coding is very modular and is as standard as possible to facilitate transportability. Efforts were made to restrict that coding that would have to be modified on other systems to a few specific subroutines placed at the very beginning of the package. Also, the data acquisition routines have the decoding facilities, using standard Fortran coding, built into them, helping to give it

independence of other operating systems. SOL consists of twenty-four independent programs, 250 subroutines and 21,000 lines of source code.

SOL originally was implemented on a Control Data Corporation (CDC) computer with 32,767 (32K) words of core storage of which only 12K was available for the on-line statistics package, as it was called then. The time-sharing system had not been developed at this time for the CDC machine; however, using an overlay processor (Winspur, 1970) each program if modularly designed could fit into approximately 4K and is described by Rollwagen, Protti, and Saunders (1971). Without overlaying the average program is under 10K.

The only problem with the above restriction of 4K was that only a maximum of three programs could be run simultaneously, regardless of whether the program was statistical or otherwise. When the usage of the statistics package increased users were unable to gain access to the system - a serious situation for on-line systems. To rectify the problem an internally developed time-sharing system was developed for SOL (Winspur, 1971), where up to six users may access SOL simultaneously. When the need arises (probably late 1974), the time-sharing system will be modified to handle 12 users

simultaneously. Response times are excellent when six users are on simultaneously and system reliability is equally good. SOL is available free of charge to non-profit organizations and the original version has been sent to over twenty other installations.

5.3 Influence on Statistical Methods

Initially in 1970 only two teletypes could access SOL; only one hour per day was permitted, and this time had to be booked. The acceptance of SOL was good and user needs have reached the point where twenty-five ASR-33 teletypes and four portable teletypes are distributed throughout the Health Sciences complex as well as five remote teletypes at other institutions - all with twenty-four hour accessibility. For the last two years (1973 and 1972) users have spent approximately six hours per day of connect time at the teletype terminals using SOL (i.e., the time the user logs on the system to the time he logs off).

SOL was not designed to be a teaching aid in formal statistics courses, but is, nevertheless, being used as such. In the Medical College the graduate Bio-statistics and Biometrics courses design course work, assignments, and take home exams to a fair extent around SOL. In the Faculty of Education

(University of Manitoba) SOL is also used in undergraduate and graduate statistics courses (Sandals, 1974) with success. Sandals singles out the major advantage being that SOL is easier and faster to use than batch systems. Also, more time is available to do examples and in depth interpretation - especially with those students having difficulty.

In both teaching and for production purposes SOL has helped in decreasing the fear of using both computers and statistics. In a typical consulting situation after decisions have been reached on how to analyze the data, the researcher is often incapable of doing his own statistical computations. Hand calculations are impractical due to: lack of knowledge of what formulas to employ; high source of error in calculations; large quantities of data; large number of experiments; and time considerations. If the researcher has an electronic calculator he may wish to use it, provided the analyses is preprogrammed, but, in general, calculators are within specific departments and inaccessible. Even so, electronic programmable calculators in the past have tended to be limited in statistical analyses and data management and often difficult to understand and use for most researchers within the author's environment. If possible, the researcher uses SOL. He is not aware of the computer side other than the conversational

interaction and he is not frustrated by the batch problems discussed previously. Results are often obtained a short time after a consultation dispelling fears of both computers and statistics. The fact that the user is involved in the process reinforces confidence and encourages a more 'statistical attitude' toward future experiments.

In counteracting fears of statistical methods SOL's major contribution has been availability, instant turn-around time, and ease of use. For simplicity the t-test may be discussed. If one can explain the t-test concept to a researcher in terms of two samples of data along with the assumptions and then provide an easy method of doing it at the same time, then the researcher will perform the test. Of course, the easier something is to use the more likely it will be misused, but, in general, most researchers will not knowingly abuse statistics. For example, in ST13 (two sample t-test, Wilcoxon-Mann-Whitney, and Kolmogorov-Smirnov program) a researcher unfamiliar with non-parametric statistics requested these options without knowing what they meant. The user did not understand the results from the two non-parametric tests and consequently asked, "What are these tests? What makes them different from the t-test and when do I use them?" There are cases, however, where ease of use has permitted indiscriminate use of

SOL, as well as inefficient use, e.g., analyzing large quantities of data.

Summarizing, more time is spent by the consulting statistician on concepts and interpretation; and the user, because he does his own work gains confidence in statistics enabling him to prepare for more advanced statistical techniques, as well as preparing the way for more formal data collection methods needed in large scale batch systems.

5.4 Future

SOL has been programmed to accept input from data files and will go into production when a general file system has been developed for the CDC 1700 computer system (now nearing completion). Implementation of comprehensive data management procedures is unfeasible on the CDC 1700 computer system and future developments in this area may necessitate more computing power.

Addition of CAI is being considered. A user will be able to ask questions at any step of analysis and get appropriate answers. For example, if the user is running the descriptive statistics program and types, '??WHAT IS A STANDARD DEVIATION'

then the program stops what it currently is doing (because it recognizes two question marks (??) as being a signal that the user needs help) and scans the question looking for keywords and finds 'STANDARD DEVIATION'. From the bank of answers the reply is printed out. The user may then continue asking questions or proceed on with running the program. Statistical analyses will continue to be added. With the addition of files, multivariate analyses will probably be the next to be added.

SOL's original design was to complement - not compete with - the larger batch systems and future additions will be along the same philosophy.

CHAPTER 6

CONCLUSION

Excluding the miscellaneous statistical systems, a total of 22 systems has been presented and evaluated, of which 15 have been discussed in detail. Some general observations can be made. It appears that the batch systems are more powerful in providing statistical procedures and data management than the on-line systems. The on-line systems are faster and easier to use, and many are excellent teaching systems. The major on-line systems, however, suffer from a lack of transportability. Looking at the integrated systems versus the non-integrated it can be seen that the integrated systems are the most powerful, and even with these extra features, they are generally easier to learn and use for the batch systems.

Whether the integrated systems are going to phase out the non-integrated systems remains to be seen. The trend indicates they will, although the BMD system containing the new P-series of programs makes a strong case for retention of a package of independent programs. For those involved in programming or

in developing statistical systems, packages of 'subroutines' are an asset and it is probable that such packages will continue to be developed. Two such systems are F4STAT and IMSL. It was stated in the introduction that only those systems that did not require a knowledge of computer programming would be evaluated, but it was felt that an exposure to these two systems would be of benefit, even for non-programmers.

Besides there being four major areas: batch and on-line systems, integrated and non-integrated, there is considerable variety among the systems within each area. The reasons for this are understandable. For example, 1) the particular type of computer available; 2) the type of person using the system (i.e., statistician, non-statistician, student, computer scientist, etc.); 3) the personality of the user, regardless of his professional type; 4) the types of data base and the amounts of data to be expected; 5) ease of implementation, costs, etc.; 6) the speed of obtaining results. In short, there is no one best system. Each of the criteria in the guidelines has different weights depending on the particular application and user. Nevertheless, some very general conclusions can be made about those systems that were evaluated in detail.

The batch integrated systems (ASCOP, OMNITAB, P-STAT, SAS, and SPSS) are the most powerful. All provide fairly comprehensive statistical procedures with P-STAT containing advanced analysis of variance procedures; SPSS providing strong frequency analysis and multivariate procedures; SAS containing comprehensive regression techniques and an easy to use analysis of variance procedure for balanced designs and other procedures for unbalanced designs; OMNITAB provides comprehensive descriptive and regression procedures with facilities for weighted analyses. All have powerful data management procedures with SAS containing the most powerful in terms of allowing programming; and P-STAT and ASCOP strong in the use of creating, storing, and using data files and using output from one procedure as input to others. SPSS and SAS are the easiest to use with the SPSS user's manual (not the updates) being the most readable of all systems. The OMNITAB system has sufficient power to enable the user to program non-standard analyses and would prove useful more for teaching than for production.

The batch non-integrated systems (BMD, STAT-PACK - GODDARD, and STATISTICAL PACKAGE - MANITOBA) suffer from a lack of data management and ease of use. The statistical analyses provided, however, are good with BMD providing an extremely

powerful set of statistical programs.

The on-line integrated systems (IMPRESS, MIDAS, and P-STAT) are similar to the batch systems with MIDAS being the most general. IMPRESS is an excellent teaching system and comprehensive in the statistics it provides, but is not very general. Neither MIDAS nor IMPRESS is that transportable. P-STAT is transportable to IBM 360/370 computers but whether it works as it should may be suspect, as Buhler (1973c) says, "...works fairly well...".

The on-line non-integrated systems (ISIS, RAX, SSIPP, STATPACK2 - APL, and SOL) are all weak in data management. RAX and SOL are the more comprehensive and general systems. RAX is not that transportable and may be rather limited because it uses the variable by case structure but cannot handle missing values. ISIS and SSIPP are primarily instructional systems. STATPACK2 - APL has potential, but it leaves much to be desired in its present format.

Concerning ourselves with the SOL system the evidence indicates that it serves an important role in providing a fast and easy to use tool for aiding students and researchers in learning and applying statistical methods to their data.

It complements the major batch systems and has, within the author's environment, greatly assisted in banishing fears of computers and statistics.

REFERENCES

[1] ANDERSON, RONALD E. (1971). "A survey of application software for social data analysis instruction," *Conference on Computers in the Undergraduate Curricula.* Dartmouth College, Hanover, N.H., 135-141.

[2] ARMOR, DAVID J. (1972). "The DATA-TEXT system - an application language for the social sciences," *AFIPS,* Vol. 40, Montvale, N.J. 333-342.

[3] ARMOR, DAVID J. and COUCH, ARTHUR S. (1972). *DATA-TEXT Primer.* The Free Press, New York. 227 pp.

[4] ATKINSON, GLEN F. (1971). "STATLAB, a simple programming system for the statistics laboratory," *Conference on Computers in the Undergraduate Curricula.* The University Press of New England, Hanover, N.H., 456-461.

[5] BARR, ANTHONY J. and GOODNIGHT, JAMES H. (1972). *SAS - A User's Guide to the Statistical Analysis System* by Jolayne Service. Student Supply Stores, North Carolina State University, Raleigh, North Carolina. 260 pp.

[6] BEATON, ALBERT E. (1964). *The Use of Special Matrix Operators in Statistical Calculus.* Bulletin RB-64-51, Educational Testing Service, Princeton, N.J. 226 pp.

[7] BEATON, ALBERT E. (1973). "F4STAT statistical system," *Computer Science and Statistics: 7th Annual Symposium on the Interface,* ed. William J. Kennedy. 117 Snedecor Hall, Iowa State University, Ames, Iowa, 279-282.

[8] BROWN, DOUGLAS; GOODIN, DONNA; MEETER, DUANE; and SOLLER, RAYMOND. (1972). *ISIS - Interactive Statistics Instructional System.* Florida State University Computing Centre and Department of Statistics, Tallahassee, Florida. 22 pp.

[9] BUHLER, ROALD (1973a). *P-STAT - A Computing System for File Manipulation and Statistical Analysis of Social Science Data.* Princeton University Computer Centre, Princeton, New Jersey. 187 pp.

[10] BUHLER, ROALD (1973b). "Supporting P-STAT on several computers," *Computer Science and Statistics: 7th Annual Symposium on the Interface*, ed. William J. Kennedy. 117 Snedecor Hall, Iowa State University, Ames, Iowa, 181-188.

[11] BUHLER, ROALD (1973c). "The P-STAT system," *Computer Science and Statistics: 7th Annual Symposium on the Interface*, ed. William J. Kennedy. 117 Snedecor Hall, Iowa State University, Ames, Iowa, 283-286.

[12] CAMERON, JOSEPH M. and HILSENWRATH, JOSEPH (1965). "Use of general-purpose coding systems for statistical calculations," *IBM Scientific Computing Symposium - Statistics*. 281-299.

[13] CHAMBERS, JOHN M. (1967). "Some general aspects of statistical computing," *Applied Statistics*, Vol. 16, 124-132.

[14] CHAMBERS, JOHN M. (1969). "A computer system for fitting models to data," *Applied Statistics*, Vol. 18, 249-263.

[15] CHAPIN, DAVID A. and MYERS, EDMUND D. (JR.) (1972). "Project IMPRESS: an interactive social science software package," a preprint of a contribution to a book tentatively titled *Effective applications of Time-Sharing at Dartmouth College*, ed. Robert F. Hargraves, Jr. 104 pp.

[16] CHEBIB, FAROUK (1972). *STATISTICAL PACKAGE*, Part B of the Programmer's Guide. Computer Centre, University of Manitoba, Winnipeg, Manitoba. 96 pp.

[17] CHOUINARD, PAUL (1973). "Soupac system development: past, present, and future," *Computer Science and Statistics: 7th Annual Symposium on the Interface*, ed. William J. Kennedy. 117 Snedecor Hall, Iowa State University, Ames, Iowa, 270-273.

[18] COLE, A. J. and CAMPBELL, R. M. (1969). "Yet another conversational mode program," *Applied Statistics*, Vol. 18, 190-191.

[19] COLIN, A. J. T. (1967). "On-line access systems in statistics," *Applied Statistics*, Vol. 16, 111-119.

[20] COOPER, BRIAN E. (1967). "ASCOP - A statistical computing procedure," *Applied Statistics*, Vol. 16, 100-110.

[21] COOPER, BRIAN E. (1969a). "Statistical computing - past, present, and the future," *The Statistician*, Vol. 19, 125-141.

[22] COOPER, BRIAN E. (1969b). "The continuing development of a statistical system," *Statistical Computation*, eds. R. C. Milton and J. A. Nelder. Academic Press, New York, 295-315.

[23] COOPER, BRIAN E. (1972). *ASCOP User Manual*. National Computing Centre, Quay House, Quay St., Manchester. 158 pp.

[24] CRADDOCK, J. M., and FREEMAN, M. H. (1967). "The METO computer language," *Applied Statistics*, Vol. 16, 120-122.

[25] CRAMER, ELLIOT M. (1967). *Revised MANOVA Program and Simplified Instructions*. Psychometric Laboratory, University of North Carolina, Chapel Hill, N. C. 14 pp.

[26] CRAMER, ELLIOT M. (1973). "MANOVA, A computer program for univariate and multivariate analysis of variance," *Computer Science and Statistics: 7th Annual Symposium on the Interface*, ed. William J. Kennedy. 117 Snedecor Hall, Iowa State University, Ames, Iowa, 245-251.

[27] DIXON, W. J. (1973). *BMD - Biomedical Computer Programs*. University of California Press, Los Angeles, California. 773 pp.

[28] DOUGLAS, A. S., and MITCHELL, A. J. (1960). "Autostat - a language for statistical data processing," *Computer J.*, Vol. 3, 61-66.

[29] EVANS, DENNIS A. (1973). "The influence of computers on the teaching of statistics," *J.R.S.S., Series A*, Vol. 136, 153-190.

[30] FOX, DANIEL J. (1973). "Data manipulation with MIDAS," *Computer Science and Statistics: 7th Annual Symposium on the Interface,* ed. William J. Kennedy. 117 Snedecor Hall, Iowa State University, Ames, Iowa, 254-261.

[31] FOX, DANIEL J. and GUIRE, KENNETH E. (1973). *Documentation for MIDAS - Michigan Interactive Data Analysis System* (second edition). Statistical Research Laboratory. University of Michigan. 173 pp.

[32] FRANCIS, IVOR (1973). "A comparison of several analysis of variance programs," *JASA,* Vol. 68, 860-865.

[33] FRANE, JAMES W. (1973). "Educational aspects of the BMDP and BMD series of statistical computer programs - point of view of a developer," *Computer Science and Statistics: 7th Annual Symposium on the Interface,* ed. William J. Kennedy. 117 Snedecor Hall, Iowa State University, Ames, Iowa, 238-242.

[34] GOODNIGHT, J. H. (1973). "Design philosophy of the Statistical Analysis System," *Computer Science and Statistics: 7th Annual Symposium on the Interface,* ed. William J. Kennedy. 117 Snedecor Hall, Iowa State University, Ames, Iowa, 233-235.

[35] GOWER, JOHN C. (1969). "Autocodes for the statistician," *Statistical Computation,* eds. R. C. Milton and J. A. Nelder. Academic Press, New York, 37-62.

[36] GOWER, JOHN C., SIMPSON, H. R. and MARTIN, A. H. (1967). "A statistical programming language," *Applied Statistics,* Vol. 16, 89-99.

[37] HARTLEY, H. O. (1962). "Analysis of variance," *Mathematical Methods for Digital Computers,* eds. A. Ralston and H. Wilf. John Wiley and Sons, Chapter 20.

[38] HOGBEN, DAVID; PEAVY, SALLY T.; and VARNER, RUTH N. (1971). *OMNITAB II - User's Reference Manual.* Statistical Engineering Laboratory, National Bureau of Standards, Tech. Note 552, Washington, D.C. 264 pp.

[39] IMPRESS:

a) *The IMPRESS Manual* (July, 1972). 197 pp.
b) *Analyzing Tabulations with IMPRESS* (second edition, Aug. 1973). 34 pp.
c) *The IMPRESS Primer* (third edition, Sept. 1973). 84 pp.

Project IMPRESS, Silsby Hall, Dartmouth College, Hanover, N.H.

[40] *IMSL - International Mathematical and Statistical Libraries* (1973). Suite 510, 6200 Hillcroft, Houston, Texas. approx. 800 pp.

[41] JOWETT, D; CHAMBERLAIN, R. L.; and MEXAS, A. G. (1972). "OMNITAB - A simple language for statistical computations," *J. Statis. Comput. Simul.*, Vol. 1, 129-147.

[42] JOYCE, S. M. (1972). "The development of an interactive statistical language," *On-Line 72*, Vol. 1, Brunel University, Uxbridge, England, 477-496.

[43] KELLY, R. F. (1970). *CALLDATA Statistical System.* Grumman Data Systems Corp., Bethpage, New York, 150 pp.

[44] KENNEDY, WILLIAM J. (ed.) (1973). *Computer Science and Statistics: 7th Annual Symposium on the Interface.* 117 Snedecor Hall, Iowa State University, Ames, Iowa. 440 pp.

[45] KILPATRICK, S. J. (JR.) (1972). "Interactive computing - four years experience in medical research and training," *On-Line 72*, Vol. 1, Brunel University, Uxbridge, England, 591-604.

[46] KREITZBERG, CHARLES B. and SCHNEIDERMAN, BEN (1972). *The Elements of FORTRAN Style: Techniques for Effective Programming.* Harcourt, Brace, Jovanovich, Inc., New York. 121 pp.

[47] LEE, P. J. (1971). *Multivariate Analysis for the Fisheries Biology* (Technical Report No. 244). Fisheries Research Board of Canada, Freshwater Institute, Winnipeg, Manitoba. 182 pp.

[48] LONGLEY, JAMES W. (1967). "An appraisal of least squares programs for the electronic computer from the point of view of the user," *JASA*, Vol. 62, 819-841.

[49] MULLER, MERVIN E. (1969). "Statistics and computers in relation to large data bases," *Statistical Computation*, eds. R. C. Milton, and J. A. Nelder. Academic Press, New York.

[50] MULLER, MERVIN E. (1970). "Computers as an instrument for data analysis," *Technometrics*, Vol. 12, 259-293.

[51] MYERS, EDMUND D. (JR.) (1969). *Survey of Social Science Computing Systems.* Project IMPRESS. Dartmouth College, Hanover, N.H. 50 pp.

[52] NEELY, PETER M. (1966). "Comparison of several algorithms for computation of means, standard deviations and correlation coefficients," *Communications of the ACM,* Vol. 9, 496-499.

[53] NELSON, W. B.; HENDRICKSON, R.; PHILLIPS, M. C.; THUMHART L. (1973). *STATPAC Simplified - A Short Introduction to How to Run STATPAC, A General Statistical Package for Data Analysis.* Technical Information Series, Report no. 73CRD046, General Electric Co., Schenectady, N.Y. 29 pp.

[54] NELSON, W. B.; PHILLIPS, M; and THUMHART, L. (1973). "More effective computer packages for applications," *AFIPS,* Vol. 42. AFIPS press, Montvale, N.J., 607-614.

[55] NIE, NORMAN, H; BENT, DALE H.; HULL, C. HADLAI (1970). *SPSS - Statistical Package for the Social Sciences.* McGraw-Hill, New York. 343 pp.

[56] NIE, NORMAN H.; HULL, C. HADLAI; KIM, JAE-ON; STEINBRENNER, KARIN (1971). *SPSS UPDATE MANUAL.* National Opinion Research Centre, University of Chicago. 79 pp.

[57] NIE, NORMAN H.; HULL, C. HADLAI; KIM, JAE-ON; STEINBRENNER, KARIN (1972). *SPSS UPDATE MANUAL.* National Opinion Research Centre, University of Chicago. 104 pp.

[58] NIE, NORMAN H.; HULL C. HADLAI; KIM, JAE-ON; STEINBRENNER, KARIN; JENKINS, JEAN (1973). *SPSS UPDATE MANUAL.* National Opinion Research Centre, University of Chicago. 143 pp.

[59] ROLLWAGEN, ROBERT I. (1973). *SOL - Statistics On-Line.* Computer Department for Health Sciences, University of Manitoba, Winnipeg, Manitoba. 189 pp.

[60] ROLLWAGEN, ROBERT I.; PROTTI, DENIS J.; and SAUNDERS, MICHAEL G. (1971). "A conversational statistics package," *FOCUS-5 Proceedings* (CDC User's Forum), St. Paul, Minnesota, 223-239.

[61] RYAN, T. A. (JR.) and JOINER, BRIAN L. (1973). "MINITAB: A statistical computing system for students and researchers," *The American Statistician*, Vol. 27, 222-225.

[62] SANDALS, LAURAN (1974). Personal communication re: using SOL as teaching aid. Faculty of Education, University of Manitoba.

[63] SCHUCANY, W. R.; MINTON, PAUL D.; SHANNON, B. STANLEY, (JR.) (1972). "A survey of statistical packages," *ACM Computing Surveys*, Vol. 4, 65-79.

[64] SHANNON, S. (ed.) (1967). *STAT-PACK; A Biostatistical Programming Package* (second edition). Goddard Computer Science Institute, 9000 Harry Hines Blvd., Dallas, Texas. 192 pp.

[65] SMILLIE, K. W. (1969). *STATPACK2: AN APL STATISTICAL PACKAGE* (second edition), Publication No. 17. Department of Computing Science, University of Alberta, Edmonton, Alberta. 67 pp.

[66] *SSIPP - Social Sciences Interactive Programming Package, Users' Guide (CDC 6400 Version)* (1972). Computer Services, University of Calgary. 93 pp.

[67] *SSP - System/360 Scientific Subroutine Package.* (360A-CM-03X) Verion III, Programmer's Manual, IBM, Technical Publications Department, 112 East Post Road, White Plains, N.Y. 450 pp.

[68] TAUCHI, H. J. (1968). *RAX - Remote Access Statistical System.* IBM Corporation, 1930 Century Park West, Los Angeles, 144 pp.

[69] *TROLL: An Introduction and Demonstration* (1973). Computer Operations Activity, National Bureau of Economic Research, Inc., 575 Technology Square Cambridge, Massachusetts. 54 pp.

[70] *TSAR - TELE-STORAGE AND RETRIEVAL SYSTEM* (1971). Duke University Computation Center, Durham, North Carolina. 138 pp.

[71] TUCKER, ALLEN B. (1973). "EASYSTAT - An easy to use statistics package," *AFIPS*, Vol. 42, Montvale, N.J., 615-619.

[72] TUKEY, JOHN W. (1965). "The technical tools of statistics," *The American Statistician*, Vol. 19, 23-28.

[73] VAN REEKEN, A. J. (1971). "Report on the work of the Dutch working party on statistical computing," *Applied Statistics*, Vol. 20, 73-79.

[74] VARNER, RUTH N. and PEAVY, SALLY T. (1970). *Test Problems and Results for OMNITAB II.* Statistical Engineering Laboratory, National Bureau of Standards, Tech. Note 551, Washington, D.C. 190 pp.

[75] WAMPLER, ROY H. (1970). "A report on the accuracy of some widely used least squares computer programs," *JASA*, Vol. 65, 549-565.

[76] WILKINSON, GRAHAM N. (1969). "Facilities in a statistical program system for analysis of multiply-indexed data," *Statistical Computation*, eds. R. C. Milton and J. A. Nelder. Academic Press, New York, 201-228.

[77] WINSPUR, WILLIAM (1970). "Design and implementation of an overlay processor for the CDC 1700 MSOS operating system," *FOCUS-3 Proceedings* (CDC User's Forum), St. Paul, Minnesota, 140-171.

[78] WINSPUR, WILLIAM (1971). *Swapping task manager* (Software Maintenance Specification No. 10.). Computer Department for Health Sciences, University of Manitoba, unpublished report.

[79] YATES, F. (1966). "Computers, the second revolution in statistics," *Biometrics*, Vol. 22, 233-251.

[80] YOUNGS, EDWARD A. and CRAMER, ELLIOT M. (1971). "Some results relevant to choice of sum and sum-of-product algorithms," *Technometrics*, Vol. 13, 657-665.