CHARACTERISTICS OF A STATISTICAL COMPUTING SYSTEM

by

J. H. Goodnight

Institute of Statistics
Mimeograph Series No. 996
Raleigh - April 1975

Characteristics of a Statistical Computing System *

J. H. Goodnight
N. C. State University
Raleigh, North Carolina

1. Introduction

Each generation of computers has had its generation of users. The first generation of computer users was machine language or symbolic machine language oriented. The second generation of users was oriented more toward the higher-level languages of FORTRAN and COBOL. In the third generation, the majority of users tend to rely on problem-oriented languages, specific to their needs, and on other types of user-oriented software.

This past decade has been a period during which the large main frame computer manufacturers have sought to make the new additions to their product lines upwards compatible with their previous computers. Thus, for a good part of this past decade a fairly stable computing environment has existed at many computer locations. This stability has made possible some very large and long-term statistical software development projects.

The similarity of computing equipment from installation to installation and the growth of large-scale computer networks has led to the widespread availability of statistical software and has stimulated its use in almost every field of research.

2. The Community of Users

The community of users of statistical software is perhaps as diverse as the available software itself. At one extreme we see the occasional user with a moderate amount of data for which a specific analysis is needed, while at the other extreme we see users who regularly deal with large amounts of data needing varying types of analyses. Perhaps the availability and accessibility of today's user-oriented statistical software has contributed to the generally larger and more diverse collections of data which are gathered for analysis. Whatever the cause, a portion of the user community does in fact need software which allows for extensive data base management and which allows easy access to a broad range of statistical analysis procedures.

For example, consider a user faced with a large set of data consisting of both continuous and discrete variables, collected in, say, several regions of the country. Suppose that listings of the data, scatter plots for selected variables, regression analyses involving some or all of the continuous variables,

---

* Paper presented at the August, 1974, joint statistical meetings of The American Statistical Association and The Biometric Society ENAR and WNAR, St. Louis, Missouri

and frequency counts or tables are needed. In addition to needing these things on the entire set of data, suppose that some of these things are needed on selected subsets representing one or more regions. Users who have confronted situations of this type and have had only stand-alone programs which perform just one of the specific analyses per run can perhaps best appreciate the type of software which allows for all of these things to be done in a single run with a minimum of user instructions.

3. Statistical Computing System Defined

The August, 1974, Report and Proposal of the Committee on Evaluation of Program Packages to the Section on Statistical Computing of the American Statistical Association suggests some basic guidelines for the review of packages of computer programs for statistical analyses. In the report, a program package is characterized as consisting of a front end for data preparation plus a set of procedures. Stand-alone programs are considered to be degenerate packages with only one procedure. This characterization of packages is of necessity quite liberal and covers a broad range of statistical software. If stand-alone programs are the degenerate packages, then the opposite end of the spectrum is where I feel the statistical computing systems would fall.

I define a statistical computing system as an interdependent collection of both data management and statistical processors which are controlled by a common supervisor and which are accessed through an easily understood control language. The system should be a potentially complete environment for the statistical analysis of data.

Ideally, in this definition of a statistical computing system the words "potentially complete environment" would be replaced by the words "total environment". Perhaps some systems are already a total environment for some users. For this group of users, no knowledge of higher level compilers or knowledge of computer operating systems and their language is necessary. The totality of the data management functions and data analysis procedures they use is incorporated in the one system with which they are familiar.

4. System Characteristics

The definition given for a system points out the basic characteristics of a system, which I will expound.

A. A system must have a control language through which the user describes the tasks he wants performed. The existance of a language is the most cited characteristic of a system (1, 2, 3, 9). In fact, the word "language" is often used interchangeably with the word "system". Elements of the language are used to describe the user's data. Other elements of the language are used in transforming the data and performing additional data management functions. Still other elements of the language are used to indicate which statistical procedures are to be applied to the data.

Although the actual syntax of the various languages differs from system to system, they generally maintain a fairly uniform syntax within a given system. This feature of a uniform syntactical expression of statements within a given language provides for a great deal of transference of knowledge as a user needs additional procedures within the system. Having once mastered the language of data definition, transformation, and of one analytical procedure, he can easily master its use for other analytical procedures.

The statistical computing languages of today are perhaps the forerunners of a fairly standard statistical computing language of the future. Of today's systems, those which are still being modified and expanded are influenced heavily by their users' requests. Evaluations of the form proposed by the Committee on Evaluation of Program Packages will also tend to influence developers toward uniformity of their languages.

B. A statistical computing system must have statistical analysis procedures sufficient for performing the types of analyses its users most often need. Some systems may specialize in a particular area of analysis, while others may offer a wider spectrum of analytical capabilities without a great deal of depth in a particular area. I will not attempt to list specific analytical procedures which should be in a system. However, regardless of the number of different analytical procedures that are available, the user must have the ability to access any number of these procedures through use of the statistical computing language in a given computer run.

C. A statistical computing system must have data management facilities. The language must provide mechanisms for describing the data which is to be accessed, for transforming the data, and for the generation of new variables. The language must provide a mechanism for specifying which variables and which observations are to be analyzed by a given procedure. The system must be capable of using the same set of data or any part of it in different analytical procedures in the same run.

The data management features of a statistical computing system are perhaps its most prevalent reason for existance. The ability to get a set of data ready for analysis within the same framework in which it is to be analyzed allows the "non-computer expert" access to the analytical capabilities of the system without having to learn other higher level languages. Data management facilities of a system should not, however, be limited to its front end. They should be accessible at any time within a run. This allows the user to access data sets which are themselves created by analytical procedures. The data management facilities in a system should be viewed as being more at the center of the system rather than at the front end.

D. A statistical computing system must have the capability of intermixing data management functions and requests for analytical procedures. The user must be able to structure his sequence of instructions to achieve his particular goals. The system, in order to achieve this end, must have some mechanism to provide for the control of the system. Although this aspect of a system may not be apparent to the user, it is what makes a system a system. This portion of the system is usually referred to as the system supervisor. The prime function of the supervisor is to pass on the user's instructions to the varying parts of the system for action. Once action has been taken by a part of the system, it returns control to the supervisor. The supervisor then examines the next request and additional parts of the system come into play.

5. Accuracy, Flexibility, and Speed

At the last Interface symposium Hemmerle (7) cited accuracy, flexibility, and speed, in that order, as the proper ranking of the criterion to use in considering statistical computations. Although his thoughts centered primarily on hardware considerations, the ranking applies equally well to software.

The statistician must have and indeed expects accuracy for the computations he requests from a particular piece of software.

Flexibility is perhaps the greatest advantage of the systems approach to statistical computation. Using a system allows the user to construct in one job the totality of the analyses he is likely to need. Within the framework of the system he can read his data, transform it, and analyze parts or all of it with different procedures. The flexibility of data handling and movement from one procedure to another is something that cannot be achieved with stand-alone programs without knowledge of the computer operating system conventions and language.

The speed at which statistical computations are performed is too often compared to the actual run time computer costs. Whereas this is perhaps a valid comparison among different systems, it is usually misleading in comparing system performance to the performance of stand-alone programs. Any cost comparisons of this nature should compare total job costs for all analyses envisioned; defined as the cost of data preparation, set-up costs, and actual computer cost.

6. System Modularity

One of the criticisms of systems has been their lack of modularity. At the last Interface conference, Frane (4) stated that

> "The ease with which the source can be changed is one of the advantages of working with a series of stand-alone programs rather than with a system in which all statistical procedures have been combined into a single program."

Hemmerle (7) at the same conference stated that

> "Systems which attempt to incorporate a myriad of algorithms into one master monitor which interprets statements and does whatever the algorithms permit inherently suffer from a lack of modularity."

As a proponent of statistical systems, I feel it necessary to offer some clarification to these statements. The very nature of a system requires that it be extremely modular. Systems which offer dozens of different procedures for data analysis cannot afford the luxury of maintaining all of the routines in storage at one time. Indeed, with a systems approach, only the supervisor, along with the procedure being used for the current analysis, is in storage. The supervisor brings in the specific portion of the system needed to handle each request. Although some systems have been written as one large overlaid program, others exist for which each procedure is itself an independent module. Systems of this type require no modification to the supervisor when a new procedure is added, and the language in which the new procedure is written is up to its developer.

7. Programming Advantages

Let us consider the task which faces a developer of a new statistical procedure, say one that is to be made generally available to other users. To incorporate his routine into a statistical system will require him to learn the routines necessary to interface with the system supervisor. In general, the supervisor has built into it the necessary routines for parsing the user's language statements, so he need not write his own parsing routines. Since the system offers extensive data management facilities, he need not build any data manipulation routines into his procedure. Use of his new procedure will require little additional learning by persons already using the system, since they already comprehend data input to the system, data manipulation, and the general language syntax of the system statements. Access to the new procedure is greatly simplified for the user, since no special operating system statements are needed. The task of documenting his new procedure is generally simplified since only the new language statements and the resultant output along with a description of the methods employed need be explained.

8. Use and Misuse

Perhaps one of the greatest concerns to statisticians is the misuse of statistical systems by those who lack sufficient knowledge in the underlying theory and assumptions of the techniques they can so readily use. In the days of the desk calculator, we didn't have to be overly concerned about misuse of our techniques, since the sheer burden of computation protected our more elaborate methods. The power is now at hand to begin to verify that data being analyzed meet the underlying assumptions. Some methods, such as lack of fit tests, tests of homogeneity of variance, and examination of residuals are available. Much more work in these areas needs to be done. Methods which are available should be included in systems. The statistician must, however, take the lead in encouraging this type of user protection.

References

1. Allerbeck, Kalus R. 1971. Data Analysis Systems: A User's Point of View. Social Science Information 10:23-35

2. Anderson, Ronald E., and Edwin R. Coover. 1972. Wrapping Up the Package: Critical Thoughts on Applications Software for Social Data Analysis. Computers and the Humanities 7:81-95

3. Armor, David J. 1970. Developments in Data Analysis Systems for the Social Sciences. Social Science Information 3:145-156.

4. Frane, James W. 1973. Educational Aspects of the BMDP and BMD Series of Statistical Computer Programs. Proceedings of Computer Science and Statistics: 7th Annual Symposium on the Interface 238-242.

5. Goodnight, James H. 1973. Design Philosophy of the Statistical Analysis System. Proceedings of Computer Science and Statistics: 7th Annual Symposium on the Interface 233-235.

6. Heiberger, R. M. 1973. Statistical Computing Through Statistical Packages: An Introductory Course. Proceedings of Computer Science and Statistics: 7th Annual Symposium on the Interface 218-222.

7. Hemmerle, William J. 1973. Computer Science and Statistics - Interface or Intersection. Proceedings of Computer Science and Statistics: 7th Annual Symposium on the Interface 213-217.

8. Meeker, Jeff B. 1973. Special Languages and Systems for Statistics. Proceedings of Computer Science and Statistics: 7th Annual Symposium on the Interface 427-432.

9. Muller, M. E. 1970. Computers as an Instrument for Data Analysis. Technometrics 12:259-

10. Nelder, J. A., and J. C. Gower. 1972. Statistical Systems and General-Purpose Languages. International Statistical Institute Bulletin 44:296-301.

11. Schucany, W. R., B. S. Shannon, and P. D. Minton. 1972. A Survey of Statistical Packages. ACM Computing Surveys 4:65-

12. Slysz, William D. 1974. An Evaluation of Statistical Software in the Social Sciences. Communications of the ACM 17:326-