DOCUMENT RESUME

ED 324 371 TM 015 647

AUTHOR Cox, Lawrence H.; Eddy, William F.
TITLE Some Remarks on Guidelines for Evaluating Statistical Software.
SPONS AGENCY National Science Foundation, Washington, D.C.; Office of Naval Research, Arlington, Va.
PUB DATE 89
CONTRACT DMS-88-05676; DMS-89-07274; N00014-87-K-0013
NOTE 11p.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computer Software Evaluation; *Evaluation Criteria; *Guidelines; Statistical Analysis
IDENTIFIERS Statistical Packages

ABSTRACT

The advisability of drafting guidelines for evaluating statistical software is considered. The Committee on Applied and Theoretical Statistics of the National Research Council has decided to initiate a project to articulate issues relating to guidelines and to determine their priorities. Because there has been a proliferation in statistical software, more statistical work is being done by people with little or no training in statistics, a fact that makes guidelines increasingly important. Benefits of the proposed guidelines for users, consumers, producers, and the scientific community are considered. The aspects of statistical packages that require guidelines include: (1) coverage; (2) numerical accuracy; (3) graphics; (4) data retrieval and data manipulation; (5) data transfer; (6) documentation; (7) user interface; (8) device interfaces; (9) speed; and (10) extensibility. Guidelines have at least four important purposes: to provide a commonality to enable consumers to share knowledge; to help code existing knowledge; to raise consumer expectations; and to provide some long-term stability in a rapidly changing environment. (SLD)

SOME REMARKS ON GUIDELINES FOR EVALUATING STATISTICAL SOFTWARE

Lawrence H. Cox, National Academy of Sciences,
William F. Eddy, Carnegie Mellon University

SOME REMARKS ON GUIDELINES FOR EVALUATING STATISTICAL SOFTWARE

Lawrence H. Cox, National Academy of Sciences,
William F. Eddy, Carnegie Mellon University
William F. Eddy, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213

1. Introduction

The Committee on Applied and Theoretical Statistics (CATS), National Research Council (NRC), is a standing committee of the NRC Board on Mathematical Sciences (BMS) designed to monitor research trends and issues affecting the statistical sciences. Dr. Eddy is the Chair of CATS; Dr. Cox directs the BMS. The issue of guidelines for statistical software is broad: on the one hand, it affects all applications of statistics and the perceptions about statistics of users and others outside the field; on the other hand, it is a problem that currently lacks a locus of responsibility and action for its solution. CATS has decided to initiate a project to articulate and prioritize issues in the guidelines area, in an attempt to bring the problems closer to the attention and action agenda of other groups.

2. Summary of the Issues

2.1. The Past

Since 1970, and particularly since 1980 with the advent of inexpensive, powerful personal computing, the use of statistical software packages has proliferated. The outputs of these packages are used to justify and evaluate important decisions affecting segments of national and world populations and economies. As these packages have become more familiar, they are used by a wider and increasingly more statistically naive class of user, with the result that evaluation of outputs and appropriateness of use are less frequently questioned.

Serious quality and reliability issues in statistical software have been documented, including: inconsistent outputs (even by major packages) to identical simple queries, inappropriate algorithms, incomplete or inconsistent repetitions of procedures, poor documentation, numerical instability, and lack of adherence to statistical standards. The effects of these defects are multiplied as statistical software becomes more widely used for routine purposes in business, science, and engineering. The potential repercussions are especially alarming when one realizes that unquestioned outputs are being routinely accepted as reliable for purposes affecting decisions of national importance. As the software increasingly becomes a surrogate for the expert statistician, human abilities to monitor performance decrease as reliance on outputs increases.

In the middle 1970's there were at most 20 commercial statistical software packages. Commonly called statistical packages, they are computer software, each of which consists of a collection of computer programs that performs statistical computations and, in some instances, presents analyses of the computations. These packages were, almost without exception, originally developed at leading research universities. The total number of users was small and the users were generally quite skilled, both in statistics and in computing. Most users submitted batch programs on punched cards to IBM mainframe computers and received support from the local computer center and often directly from the package developers. Even at that time there was great concern in the statistical community about the quality and transportability of statistical software (Francis and Heiberger, 1975).

By the late 1970's a major proliferation of statistical software had begun. Minicomputers and superminicomputers had become available to smaller research groups. The number of packages was increasing. Time-shared interactive computing was becoming commonplace in research environments.

2.2. Today

Today it is possible for anyone with a personal computer and a language processor to create a piece of software and call it a statistical package. There is great demand for statistical software because decision makers increasingly require quantitative information on which to make and defend key decisions. Often, decisions based on a statistical analysis (however flawed) are more readily accepted. There is a shortage of expert statisticians for these purposes and such expertise is expensive to develop and maintain. A statistical package is often an attractive alternative, becoming a pseudo-statistician. This has led to burgeoning growth and the proliferation of both good and bad statistical software.

Today numerous commercial and non-commercial statistical software packages are available. A conservative estimate is that there are 300 commercial packages available and many more packages under development. It is difficult to determine the number of packages that are distributed free-of-charge to the user but the count is similar. Statisticians remain deeply concerned about the quality of the statistical software they and others use (Schervish, 1988).

Because the supply of statistical software packages and computing power is growing quickly and will continue to be less expensive than expert statistical advice, more and more statistical work will be done by people with little or no training in statistics using statistical software packages of increasing "sophistication." While this has the great advantage of satisfying manpower needs with fewer highly trained statisticians, it has the great disadvantage that these untrained users will be unable to evaluate the quality of the program's performance and output. Key national, business, and social decisions will be made in this environment. What is needed to reverse this declining trend are guidelines based upon established statistical principles for benchmarking and assessing the performance and output of statistical software packages. The increased utilization of supercomputers and other high performance computers has emphasized the need for such guidelines. The issue here is not whether these guidelines will eventually lead to standards for statistical software or even whether such standards are desirable. (The difference between guidelines and standards is discussed in Section 6.) Rather, it is necessary to provide information and a framework to ensure minimal acceptable statistical functionality and ease of interpretation and use in statistical software and basic tools to verify that the guidelines have been met. We anticipate that guidelines and benchmarks as proposed will be invaluable to users and useful to developers of statistical software. This will serve to eliminate a well-defined set of problems in many software packages at the development or enhancement stage, thereby reducing greatly at least one kind of misuse of statistics.

3. Need

Wilkinson and Dallal (1977) reported the results of calculating the means, standard deviations, and correlation of three artificial observations on three variables using four statistical packages. Only one of the four packages calculated all of the sample statistics correctly. This test case demonstrates clearly the need for improvements in software quality, even in the most trivial calculations, as well as guidelines and tools for evaluating statistical software.

Teitel (1981) constructed two data sets that were nonrectangular but still had simple relationships between the data records. He then invited the distributors of several widely used statistical packages to perform simple

counts from the data. The results, reported in companion papers, are difficult to summarize simply, but it appears that no two packages got the same counts. These results are most distressing.

A complicating aspect of statistical software packages is that they are used by the most sophisticated researchers (Eddy et. al., 1986) and by the most statistically naive laypersons. Therefore, these packages must have a broader range of capabilities than other kinds of scientific software, and must be integrated vertically so that analyses which should be conforming—such as a detailed analysis by the expert statistician and a broader confirming analysis by management—are not divergent. Guidelines and baseline test data for users and producers will help span this range smoothly.

4. Benefits

4.1. Benefits to Users

The most obvious beneficiaries of the planned set of guidelines will be the users (i.e., purchasers and end-users) of the statistical software. They will have available an objective set of guidelines for evaluation of statistical software. This will allow prospective purchasers to make objective comparisons based on criteria that have been selected by experts (as opposed to the haphazard evaluation criteria that are often used now). This will allow end-users to ascertain whether aspects of the software which they do not fully understand are as trustworthy (based on criteria determined by experts) as the aspects that they do fully understand. The statistical software industry output is estimated at $250 million per year. Nearly all of these purchasing decisions have heretofore been made without the benefit of objective evaluation.

4.2. Benefit to Consumers

The largest group of beneficiaries of the planned set of guidelines are the consumers of the results produced by the software. Their benefit is indirect in two ways. First, they do not actually use the software but rather just the output. Second, the consumers may not even have the advantage of knowing that the software that produced the results has been evaluated using guidelines and benchmarks. In any case, these consumers will have the advantage that their results are based on software that is acceptable to the community of statistical experts (if that is the case).

4.3. Benefits to Producers

The smallest group of beneficiaries of the planned set of guidelines, but certainly the most important, are the producers of the software. They will benefit in two direct ways. First, they will have the opportunity to compare their software to other software based on an objective set of guidelines and benchmarks. Currently, the only comparisons are generated by individual commercial producers and there is an attendant suspicion of bias in the selection of test cases. Second, producers of software, by having test cases and guidelines for evaluation, will have the opportunity to raise the quality of their software. We expect that the user and ultimately the producer community will come to the point where software will be unacceptable unless it can meet the guidelines that are planned.

4.4. Benefits to the Scientific Community

In a recent article, Molenaar (1989) wrote, "The rapid growth of data analysis by non-experts has aggravated the problem that the scientific community lacks a suitable system for assessing and controlling the quality of statistical software." He went on to discuss the costs to the scientific community of bad statistical software and to argue that the community should modify the system of rewards and punishments to draw attention to the problem. Molenaar (1989) and Victor (1985) before him obviously felt that information about the quality of software must be made available to the larger scientific community, not just the

statisticians, in order to effect the desired improvements.

In the final analysis, the most important benefit is, naturally, the confidence that future research and applications will not be dangerously flawed through the naive and unwitting use of poor, inaccurate or incorrect statistical packages.

5. Aspects Needing Guidelines

There are a variety of different aspects or components of statistical software that require guidelines for evaluation and use. The only previous systematic effort aimed at evaluating the quality of statistical software (Francis, Heiberger, and Velleman, 1975) has provided some background on the appropriate aspects. However, as computing environments have changed, so have the important aspects. A list of those aspects of statistical software that require evaluation would include at least the following topics.

5.1. Coverage

What statistical procedures does the package include? The procedures which are available and their implementation limit the quality of the user's results as well as the ease with which they are produced. Commonly used procedures include descriptive statistics (both numerical and graphical), random number generation, confidence intervals and tests, regression analysis, correlation and linear model fitting, and categorical data analysis. Nonlinear model fitting is important in the physical sciences. Guidelines as to what are the best procedures known to date that make possible good statistical practice are needed.

Some software packages are more comprehensive or are specialized for particular kinds of problems. In such cases too, guidelines as to what procedures are necessary for good modern statistical practice are needed by consumers. Included in a list of more advanced/specialized topics would be: time series analysis, loglinear models (and extensive cross-tabulation), survival analysis, nonparametric smoothing, linear algebra, construction and analysis of complex experimental designs, dose-response modeling, and dynamic data display.

5.2. Numerical Accuracy

Both the common mathematical functions and the algorithms used to implement statistical procedures must exhibit acceptable numerical quality. There already exist hardware standards such as ANSI/IEEE 754-1985 for floating point arithmetic and most hardware manufacturers adhere to those standards. These standards provide a foundation on which to build statistical software which will have the highest possible numerical accuracy. There already exist a few test data sets and some unsupported recommendations for numerical accuracy of statistical software. The project described here will result in a systematic body of data and some systematic guidelines for evaluation of this aspect. It is actually becoming more important to have guidelines available with the increased use of supercomputers and other high performance computers, as their added speed allows users to do more computations on larger data sets than in the past.

5.3. Graphics

Computer graphics is fast becoming an essential component of statistical data analysis. Because of the larger size of data sets and the increased complexity of the kinds of analyses which are done using high performance computers, graphics has become a crucial analytical tool. Research on graphical perception demonstrates clearly that the eye-brain system performs certain tasks required to decode statistical data represented graphically better than other tasks (indeed, such tasks can be arranged hierarchically). The common means of representing statistical data graphically give rise systematically to perceptual distortions (e.g., use of areas to

represent linear variables such as in statistical maps and improper uses of color) (Cleveland, 1985). Statistical data are typically multidimensional and efficient, comprehensible means of representing multidimensional data are needed. The need for guidelines for statistical graphics is critical, particularly because non-scientific users of graphics tend to emphasize the ability of graphics to capture the reader's attention with less regard for the accuracy of the message received.

5.4. Data Retrieval and Manipulation

Teitel (1981) uncovered enormous discrepancies between major commercial packages in handling and making counts and inferences from data sets of only moderate complexity, and proposed some fairly simple but quite revealing data sets to test the database management capabilities of statistical software systems. It is advisable to go further and consider searching, splitting, aggregating, transforming, handling of character as well as numeric data, handling of missing values, sorting, array construction and manipulation. Statistical databases differ from traditional transaction databases in that records typically are processed in the attribute domain rather than the record domain. Paradigms leading to models of efficient statistical databases are needed. Performance guidelines are needed to construct and evaluate such paradigms.

5.5. Data Transfer

One critical feature of a program in a statistical package is its ability to read data from and write data to the world outside the program. This can be as simple as reading an ASCII text file (a feature missing from some programs) and as complex as transferring an internal data structure from the format of one program to the format of another. The importance of this feature can be inferred from the fact that there exist commercial programs which provide only this capability.

5.6. Documentation

Muller (1978) (also, Berk and Francis, 1978) has provided a set of objectives for the evaluation of statistical software documentation.

"A user guide should make it possible for a reader to determine:

- what capabilities are provided
- how they are achieved for the user
- how they are presented in the user guide
- what types of data are acceptable
- what types of input are used to create the output
- how the statistical capabilities should be used
- what constraints or limitations must be observed
- how required resources should be estimated."

5.7. User Interface

Issues include:

- clarity and ease of use (Thisted, 1976)
- diaries and histories
- commands or menus (or both)
- output format
- error handling and messages (Eddy 1981).

5.8. Device Interfaces

As hardware environments have become more diverse, it has become more difficult to find statistical program packages which support the variety of output devices available. The possible devices include printers, plotters, and multi-window workstations. There are also a multitude of more exotic devices for

which interfaces would be useful, for example, film recorders. Some statistical packages, especially those which are oriented towards graphics, provide a wide variety of interfaces; other packages provide only line printer output.

5.9. Speed

As the power of computer hardware increases, concern about the speed with which a particular calculation is performed within a statistical package diminishes. Nevertheless, the time a package takes to perform a particular procedure is important to the user and can often provide clues to the overall quality of the package. Test data sets which can serve as a basis for comparing the speed of statistical packages would be helpful.

5.10. Extensibility

A modern statistical package provides its own language for performing statistical calculations. The simplest of these merely allows for the creation of temporary variables for storage of intermediate calculations. The more sophisticated ones have many of the capabilities of more traditional programming languages; for example, loops, branches, subroutines, recursion, complex data structures, etc. Ling (1980) has provided a preliminary sketch of what a user should expect.

6. What Are Guidelines?

In order to appreciate the value of guidelines (for evaluating statistical software) it is important to understand the difference between guidelines and standards. Generally, standards fall into two broad categories:

1. those that are legislated; and
2. those that are accepted.

We begin by considering the latter.

6.1. *De Facto* Standards

There is, of course, no black and white here. Government contractual requirements which force software to be SVID-compliant are much closer to the legislative end of the gray area. Standards making bodies such as ANSI and ISO operate near the opposite end of the gray area. De facto standards are, in many senses, the best possible kind. There are few, if any, significant social or economic dislocations caused by the standards. They provide a generally higher level of functionality.

Take, as an example, the lids on the metal cans that various processed foods are sold in. We always refer to these as "tin" cans although they probably haven't had much tin in them for years. As far as we know every can opener will open every can. This is a remarkable degree of standardization and a study of how it came to be would probably be very instructive for the standard makers. We should admit parenthetically that there are some specialized can openers which are not so ubiquitous. The kind of opener that is used to punch holes in the lid so that liquids can be poured out (beer drinkers usually call these "church keys") can only be used to remove solid food in an act of desperation well-known to campers. The resulting many-pointed star is one of the most dangerous parts of a camping trip. The kind of opener often delivered with a can of sardines cannot be used for opening cans other than the special type that it was designed for. However, the can can (we apologize but couldn't resist) always be opened by a more traditional can opener; we know this from many experiences in which the key-like opener delivered with the sardine can failed to open it.

Another example that leaps to mind is automobile transmissions. These are classified into two large categories:

1. automatic transmissions; and
2. manual transmissions

and crossed with this classification are two other large categories:

1. steering-column shifter; and

2. floor-mounted shifter.

There do exist "semi-automatic" transmissions, although we haven't seen a recently-manufactured one. And within each of the two categories there are a number of minor variations. But, for example, the gear pattern is generally a straight line for the automatics in the order **P R N D L**. For the manual transmission there are a wider variety of patterns but the basic H-pattern seems extremely wide-spread. The importance of this is that (once one has have learned where **R** is on the manual transmission) you can drive any car in the world. You cannot however drive a 18-wheel tractor trailer combination. Again it would be instructive to understand how the automobile transmission came to be standardized. When was the last time that you saw an automobile with an automatic transmission operated by push buttons on the dashboard?

6.2. *De Jure* Standards

These standards usually have the force of law behind them although one could argue that some ANSI standards are in this category. One of the most important examples in this class is electric receptacles in the United States. The first and most important point is that no matter where you go in the United States the receptacle tells you what sort of electricity will be delivered. And to focus sharply on standard home 15 amp 120V AC receptacles, every plug will go into every receptacle (which meets the standard). Like any good standard the National Electric Code (NEC) is evolving and modern polarized plugs will not always go into the receptacles installed thirty years ago; this is a genuine safety feature. However, thirty-year old plugs will go into modern receptacles. Another dimension of compatibility concerns higher amperage receptacles. A plug designed to deliver between 15 and 20 amps will not fit into a 15 amp receptacle. However, a plug designed to deliver less than 15 amps will fit into a 20 amp receptacle. So the standard preserves a specific kind of compatibility across time and size; the specific compatibility was carefully planned to increase safety for the user.

How is the NEC enforced? Unfortunately, electric codes are deemed the provenance of local governments. So, each locality has (if it chooses, and most so choose because of the additional government jobs that are automatically created) its own inspectors to enforce the code. The reason we wrote "unfortunately" is that not every locality adopts the NEC; many write in their own variations on the NEC. Occasionally these variations are of good intent, reflecting some minor correction to the NEC (and such variations are often adopted into the NEC in its next revision). More often, however, these variations are adopted for local political/economic reasons. We are thinking, for example, of requiring the use of BX cable instead of the more widespread NM. BX is more expensive so, we presume, the cable distributor makes a larger profit. Or the details of some installation procedure may be varied so that it takes more time for the electrical contractor to perform a specific installation, again generating a greater income.

The situation with the NEC should be compared with other countries. A traveler to Europe (including Great Britain) must be prepared to deal with two different voltages (120V and 240V) and about five different sorts of receptacles; often there are several kinds in one country. And we have even seen cases where there were three different-shaped receptacles in one room; pity the poor folks who have only one kind of receptacle and they have an appliance with an incompatible plug.

7. What Good Are Guidelines?

First, we must understand the difference between guidelines and standards. A set of guidelines could be a preliminary standard, a sort of draft standard waiting to be adopted. A set of guidelines could be a standard that

failed to be adopted; a rule with no teeth. A set of guidelines could be advice to consumers.

Guidelines can provide at least four important and useful purposes:

1. Guidelines can provide a commonality that enables consumers to share knowledge. Universality is particularly important in complex areas as it provides a sort of *lingua franca* for communication among those who conform to the implied standards.

2. Guidelines can help serve to codify existing knowledge. Codification is equally important; with the passage of time certain kinds of knowledge erodes from the collective consciousness unless it is preserved in some formal way.

3. Guidelines can help raise consumer expectations. Raising the minimum acceptable standard seems particularly important with respect to statistical software. There is a great deal of very bad software; providing users a sort of guaranteed lower bound on the quality seems highly desirable.

4. Guidelines can provide some long-term stability in a rapidly changing environment. Upward compatibility is invaluable in the NEC; it is less clear what value such compatibility might have in statistical software.

8. References

Berk, K.N. and Francis, I.S. (1978). "A Review of the Manuals for BMDP and SPSS," *Journal of the American Statistical Association*, **73**, 65–71 (with discussion).

Cleveland, W.S. (1985). *The Elements of Graphing Data*, Wadsworth, Monterey.

Eddy, W.F. (1981). "Messages from Systems to Users: A Proposal for Standardization," Applied Systems and Cybernetics, Vol. V, 2331–2335.

Eddy, W.F., Huber, P.J., McClure, D.E., Moore, D.S., Stuetzle, W., and Thisted, R.A. (1986). "Computers in Statistical Research-Report of a Workshop on the Use of Computers in Statistical Research," *Statistical Science*, **1**, 419–453 (with discussion).

Francis, I. and Heiberger, R.M. (1975). "The Evaluation of Statistical Program Packages-The Beginning," *Proceedings of Computer Science and Statistics: 8th Annual Symposium on the Interface*, (J.D.W. Frane, Ed.), 106–109.

Francis, I., Heiberger, R.M., and Velleman, P.F. (1975). "Criteria and Considerations in the Evaluation of Statistical Program Packages," *The American Statistician*, **29**, 52–56.

Ling, R.F. (1980). "General Considerations on the Design of an Interactive System for Data Analysis," *Communications of the ACM*, **23**, 147–154.

Molenaar, I.W. (1989). "Producing, Purchasing and Evaluating Statistical Scientific Software: Intellectual and Commercial Challenges," *Statistical Software Newsletter*, **15**, 45–48.

Muller, M.E. (1978). "A Review of the Manuals for BMDP and SPSS," *Journal of the American Statistical Association*, **73**, 71–80 (with discussion).

Schervish, M.J. (1988). "Quality of Statistical Software," *Chance*, **1**, 1, 48–51.

Teitel, R.F. (1981). "Volume Testing of Statistical/Database Software," *Proceedings of Computer Science and Statistics: 13th Annual Symposium on the Interface*, (W.F. Eddy, Ed.), 113–115.

Thisted, R.A. (1976). "User Documentation and Control Language: Evaluation and Comparison of Statistical Computer Packages," *Proceedings of the Statistical Computing Section*, American Statistical Association, 24–30.

Victor, N. (1985). "Computational Statistics - Tool or Science?" *Statistical Software Newsletter*, **10**, 105–125.

Wilkinson, L. and Dallal, G.E. (1977). "Accuracy of Sample Moments Calculations Among Widely Used Statistical Programs," *The American Statistician*, **31**, 128–131.

Lawrence H. Cox is Director, Board on Mathematical Sciences, National Academy of Sciences, Washington, DC 20418. Dr. Cox's work was supported under National Science Foundation Contract Number DMS-89-07274. William F. Eddy is Professor of Statistics, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890. Dr. Eddy's work was partially supported by the Office of Naval Research under Contract Number N00014-87-K-0013 and by the National Science Foundation under Grant DMS-88-05676. The opinions and observations presented are solely those of the authors, and do not necessarily represent the views, policies, or conclusions of the National Academy of Sciences or the National Research Council.

END

U.S. Dept. of Education

Office of Education
Research and
Improvement (OERI)

ERIC

Date Filmed

March 21,1991