Comment

Lynne Billard

The Workshop on the Use of Computers in Statistical Research on which Eddy's report was based was a timely and important forum for statistical workers today. The report essentially covers two issues—research with computers and the acquisition of computers. Computers in research, or statistical computing as a subject, is a relatively nebulous concept and is probably described in as many different ways as there are readers of the report, although of course there will be a lot in common, too. The subject is clearly in a state of evolution and it is difficult to forsee future mutations which surely will emerge. For most of us, it is the acquisition of computers with which we are currently grapling. It is here I think that the report becomes of great value to us. It addresses many of the issues that might have eluded most of us simply because we feel we are in uncharted waters. The workshop members, and Eddy in particular, are to be commended for their work. Before commenting on that part of the report, let us look briefly at research with computers.

RESEARCH WITH COMPUTERS

In the past 5 years or so, a question that often seems to be raised relates to what exactly constitutes statistical computing. In the early part of this decade, a commonly heard answer suggested that statistical computing was the development and generation of statistical packages either on a large scale such as IMSL, MINITAB, SAS, SPSS, etc., or on a more individualized small scale. My own feeling is that it is much much more than this, so much so that I see statistical packages per se as only a minor part of statistical computing, and in the future it may not be a direct part of it at all. It will always be true that statisticians will require routine software of some description. However, it is my guess that this aspect of computing will become absorbed as a natural part of the task at hand in much the same way that the use of hand calculators are used, in contrast to the days of the 1960s when students had laboratory sessions for use of the now obsolete desk calculators such as the Monroe machines.

While it is relatively easy to say what statistical computing might not be in the future, it is a much more formidable task to adequately describe what it is or might be. Here in mid-1986, it would seem that statistical computing will encompass two broad categories. To these two categories should be added a third under the heading "Other" which constitutes activities and endeavors presently inconceivable, not yet visualized, and definitiely unable to be clearly defined or described today. Yet, the very nature of the vast and rapid changes associated with computers and computing alone tell us that possibly as soon as next year some of these unforseen avenues will have emerged as clearly defined entities. Certainly, in 5 years they will be more clearly in evidence.

One of the two categories of statistical computing pertains to what might be labeled "redevelopment of 'known' results" necessitated by virtue of the size of the problem at hand. Let me illustrate with three different types of examples. The first example deals with the inversion of a matrix. Prior to approximately 1970 (the exact reference could not be located), those of us who took matrix algebra courses learned to invert matrices in a certain way (using minors—the exact details are not important). While in theory all matrices regardless of the size of the dimension of the matrix could be inverted, in practice, matrices greater than about 30×30 in size brought even the most sophisticated computer hardware of the day to its knees. (Of course, the modern computer of today can fare much better.) However, the practical limitations encountered spurred on the development of new mathematical techniques for inverting matrices, techniques which were not limited by such hardware and practical considerations.

A related but different example arises out of the fact that with the ready availability of computers to research workers in all disciplines, many researchers now generate massive sets of data. These, by their very size, generate in turn difficulties of their own, although again in theory "solutions" may already exist. Some of these involve large matrices with the inherent problems alluded to in the previous example. On the other hand, it is often the case that while the totality of the data is itself large, relevant statistical entities, such as the matrix X of independent variables in a multiple regression model or cells in a $c \times r$ contingency table (where c and r are large), may be quite sparse. Problems here are generated by the sparseness itself.

The third example arises in the area of population modeling. While the derivation of the relevant

Lynne Billard is Professor and Head, Department of Statistics, University of Georgia, Athens, Georgia 30602.

differential-difference or Chapman-Kolmogorov equations usually presents little difficulty, the solving of these equations is often very intractable mathematically except in the simplest cases. In recent years some new techniques have been developed so that, for example, Billard (1981) has provided solutions (for the underlying state probabilities) for a closed bivariate birth and death process. For small population size(s), these solutions can be used easily. However, even for modest population size(s), practical difficulties are encountered although in theory none such exist. Algorithmic (re)development especially for use on a supercomputer should ensure ready calculation of the solutions for any population size. In particular, if the vector of underlying random variables has m dimensions ($m = 4$ in the bivariate birth and death process example), then m of the supercomputer's parallel processors would be engaged simultaneously in the appropriately derived algorithm.

In the three examples presented above, solutions to the situations discussed already exist theoretically. The "new" contribution relates to the way in which one arrived at that solution. In contrast, a second category covered by statistical computing consists of results, solutions, developments, techniques, etc., which are in and of themselves new. Such results would never, could never, have arisen without the computational power and resources of the computer itself. Two examples come immediately to mind. The first is the bootstrap and related techniques developed by Efron (see, for example, Efron, 1982) and several other workers. Another example is projection pursuit and its derivatives introduced by Friedman and Tukey (1974). Eddy's paper provides further illustrations.

Our attention thus far has been focused on "statistical computing" as distinct from "computers in statistical research," the topic of this report. There is a subtle but important distinction between the two concepts, with the former being but one aspect of the latter. The report demonstrates quite eloquently the necessity for computers in statistical research in general. My only quarrel with its conclusions on this aspect is its implication that computers are an essential part of *all* statistical research. It is my feeling that it is an integral and essential part of some and of peripheral need to others, while some research areas do not, as yet, require the computer at all. On the other hand, it is apparent, to me at least, that computers are an absolutely essential part of all statistical *education* and *training*. It is hard to conceive of an adequately trained student graduating from a statistics program who is not familiar with computers and computing. The implication here is that the statistics graduate is facile with using a computer. However, there is another side of computers in education, whereby the instructor can change his input, be that data, variables, models, etc., to demonstrate various principles to a class. For example, anyone who in years past has spent endless hours calculating sums of squares for data for an experimental design methods course, often to then discover that that particular choice of data is inadequate to illustrate a certain point, will have no hesitation in appreciating the ready availability of an at hand computer to assist in the class preparation.

ACQUISITION OF COMPUTERS

It would be difficult to add substantially to the information and advice already provided in the report. Let it suffice for me to briefly comment upon and to add emphasis to aspects of the report.

The report suggests that when budgeting for computational facilities, a rough estimate is for $10,000 per researcher per year. At first this sounds enormous and certainly it is hard to conceive of such figures actually becoming available. Yet, it is perhaps disturbingly more accurate than we might want to believe. One of the big reasons for this is because of the rapid changes in computer technology. Indeed by the time computers are delivered, the state of the art machine ordered may have been superceded by a new state of the art machine. The description of our own departmental equipment at the University of Georgia in this report symbollically illustrates this point. In Section 2.2 of the report, our 27 DEC PRO 380s are on order. In Appendix II.4, they, and a VAX 11/750 and other peripherals have arrived and been installed. By this discussion, they have all been networked together. Meantime Digital has phased out its 11/700 machines! An obvious question pertains to how do departments obtain the magical "$10,000 per" budget. Most universities have not seen a growth in equipment budgets in years; indeed many have never had one, having traditionally relied on grants for equipment. Such requests from statistics departments are new. Analogy with the laboratory sciences and their need for maintenance budgets is no doubt the best way to go, but with tight or shrinking budgets, success is not guaranteed.

Personnel considerations encompass both faculty and support staff. Faculty must be in an environment in which they feel comfortable and where they do not feel intimidated by the student who seems so much at home in front of a terminal. In my opinion, the only such place is in the individual faculty office itself. Very rarely is a faculty person seen "trying out" the computer in a computer room "down the hall" regardless of how many machines or terminals may be available for use. Once there is available a computing facility in the office itself, all of us, novice or expert, must be patient. My feeling is that it will take at least

2 years for the faculty member to become a sophisticated user. This learning process is greatly enhanced if there is adequate support staff. Our departmental computer network facility at the University of Georgia has one full time systems specialist (a computer science major) and three to five half time student assistants. All are kept extremely busy. It is important that these staff be responsible for all the support work necessary. In my opinion, it is inefficient for faculty to be utilized in this manner.

CONCLUSION

The report is to be highly commended, most especially as it pertains to the acquisition of equipment and all that that entails (including support personnel). There should be many people presently struggling with bits, bytes, memory, CPUs, and the like, who will be much indebted to those responsible for this report.

ACKNOWLEDGMENT

This work was supported by Grant 2R01-GM30325 from the National Institutes of Health.

ADDITIONAL REFERENCES

BILLARD, L. (1981). Generalized two-dimensional bounded birth and death processes and some applications. *J. Appl. Probab.* **18** 335–347.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* SIAM, Philadelphia.

FRIEDMAN, J. H. and TUKEY, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **C-23** 881–890.

Comment

Douglas M. Bates

I am very pleased that this workshop was held to exchange opinions on the role and funding of computing facilities in statistical research and I am happy to see this report being published here. The members of the workshop are to be commended for their thoughtful and incisive comments on an issue which is important to many of us and which will become even more important in the future.

As mentioned in the report, departments such as ours which have been fortunate to receive computer equipment grants through programs such as Scientific Computing Research Equipment in the Mathematical Sciences from the National Science Foundation and the University Research Instrumentation Program from the Department of Defense have undergone dramatic changes in the way that research is conducted and reported. These changes have not always been painless. This report is particularly helpful in describing the monetary and time costs to a department which is going to start building its own computing resources. A lot of frustration will be avoided if everyone has a realistic expectation of how much time, effort, and money is going to have to be expended to build the facilities.

This is not to indicate at all that I think building departmental computing resources is not worth all this time, effort, and money. Once you have had the opportunity to use such facilities for research, communications, and text preparation, you never want to turn back. The ability to quickly and interactively follow possible avenues of solution to problems in data analysis and presentation then collaborate on the report with colleagues at distant places via electronic mail and finally prepare the report yourself in a typeset form is addicting because it helps you to be more productive.

It is noteworthy that this report mentions the importance of the communications and text preparation aspects of having departmental computers. We tend to visualize computers as being primarily for number crunching. This is an important use because it cannot be done without the computer, but, if we look at what we do as researchers, we spend much more time writing and rewriting our reports about the results than we do actually computing the results. Facilitating our writing and communications is not a trivial use of computers: it is a very important use.

My own experience is that I don't think that I get the writing done any faster with the computer but I do think that the end product is better. Computer text processing can also help in avoiding proofreading for errors created in transcriptions of the text. The Society for Industrial and Applied Mathematics is currently experimenting with allowing authors to submit the final version of their manuscript in *troff* form thereby avoiding a potential source of transcription errors and a proofreading stage for the authors.

Douglas M. Bates is Associate Professor, Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706.