Comment

Andreas Buja, E. B. Fowlkes, and J. R. Kettenring

INTRODUCTION

We are delighted that *Statistical Science* is providing a forum in which broader aspects of statistics, like the impact of computing, can be discussed, and we are grateful to the editors for recognizing this issue, as well as to the authors of the present paper for providing an unusually informative account which will help many academic departments in the initiation or expansion of computing activity. Computing is having fundamental implications for directions of statistics research, the types of people we train and hire, our value systems, and our financial priorities. We are in basic agreement with the present paper, and we would hope that the recommendations put forward by the authors will be taken seriously. The present discussion will make a few points which are based on our own personal involvement in computing, both in academic and industrial settings, and we will add a few more specific recommendations to the authors' list of suggestions for the profession.

ISSUES OF STANDARDIZATION

The authors have important things to say about standardization, an issue with many ramifications. On the level of a single computer system, standardization may mean an effort to cut down on the number of nonstandard components both in hardware and software for the purpose of getting the system to run with the least expense of human and financial resources. Very often standardization can be interpreted as buying a complete "off the shelf" system from a single vendor. This approach has its advantages, even if the initial purchase costs are higher. A multitude of suppliers can cause problems in a number of ways: special interfaces and driver software may be needed, the sources of malfunctions may be harder to pinpoint, software updates on the main system may have disconcerting effects on nonstandard peripherals, etc.

On the other hand, there are global aspects to standardization which might make deviations from standards at a lower level necessary. A first example is the following. If a university establishes a campus wide network, a statistics department should set up the required interfaces and software, irrespective of availability from the usual vendor. A second example is provided by the UNIX-type operating systems (UNIX is a trademark of AT&T Bell Laboratories). Such systems are nonstandard alternatives to vendor-supplied operating systems on most mini-sized time-shared computers, but they have grown into a standard, transcending hardware brands. Standardization is meant to grant certain compatibilities which can be in conflict with each other, and statistics departments will have to face questions of compatibility with a user community (committed to an operating system standard), with a vendor (to facilitate maintenance), or across a campus (to access networks or tap expertise in computer science departments).

Yet another complicating aspect of standardization enters if computing is made a research topic rather than kept as an activity to achieve old goals by new means. We will then want to keep up with developments in computer science which may imply radical departures from most of what we are used to. Some may decide that we cannot afford to miss trying out certain innovations for their potential relevance, for example, to data analysis. Only history can prove them right or wrong, but one should not discard the possibility that new standards can be made to happen if research shows their superiority.

BARRIERS IN THE WAY OF COMPUTING

Barriers which obstruct access to computers are many. They range from psychological to institutional and financial. At the level of an individual who would like to get started, there arises the question of how to find expertise and documentation. A more fundamental problem is how such an individual can learn the mental models needed to develop perspective, organize a plan of action, and achieve goals. Failing at this stage imposes a feeling of backwardness on otherwise willing beginners, while a more appropriate response would be to fault the documentation and backwardness of the software. The only way to deal with the current situation is by hiring competent systems personnel from the outset and making instruction a major focus when newcomers like first year students or new faculty arrive in a department.

This brings us to another topic well worked out by the authors: the need of competent staff and the costs

Andreas Buja is affiliated with Bell Communications Research and the University of Washington. E. B. Fowlkes and J. R. Kettenring are affiliated with Bell Communications Research. Their mailing address at Bellcore is 435 South Street, Morristown, New Jersey 07960.

related to it. A typical yearly salary for such people may compare with or even rise above senior faculty salaries. This may hurt, but we have to face up to market realities. It would be a serious mistake to dump systems work on graduate students for two reasons: progress on their thesis work may be adversely affected and also they cannot provide the necessary continuity.

We would also like to reiterate the importance of the costs of maintenance of hardware and software: a typical estimate is 1% of the purchase price per month(!). Fortunately, some funding agencies have grown more responsive to this ongoing problem and allow the inclusion of maintenance as an item in budgets for research proposals.

Another barrier concerns attitudes within the statistical community: the arrival of computers may affect our value systems and upset some long standing priorities. One of the major questions is whether Ph.D. theses of a computer science flavor will be accepted in statistics as a valid alternative to the traditional theses which have grown out of the mathematical orientation of our discipline in the 1950s. Statistics departments are facing the problem of establishing standards of intellectual acceptance for new types of work which do not fit the old mold. If this doesn't happen, statistics may become irrelevant to data analysis and lose some of its raison d'être. Moreover, students from departments which refuse to embrace modern computing developments may suffer a disadvantage in the job market.

OTHER EXPERIENCES

While the focus in this paper is on universities, its title, "Computers in Statistical Research," suggests that experiences from government and industry may be relevant too. As an illustration, we can mention a few aspects of the computing environment in the Statistics Research Group at Bell Communications Research, a 2-year-old company, where many of the start up issues addressed in this study were faced.

Tremendous benefits from the statisticians' perspective have been realized by working in an interdisciplinary and physically intermingled environment in which dealing with limited resources is not only a necessity but, in many respects, a virtue. The computer scientists have led the way in establishing a local environment in which many computers and workstations are networked together for joint use by roughly 100 scientists. There have been many benefits of such a setup. For example, if a machine is scheduled for maintenance, it is an easy matter to transfer essential files to another one quickly. Computing-related problems can be referred to "staff" allowing a knowledgeable person to pick up on them. This provides a range of consulting experts that extends well beyond the base of statisticians. While much of this assistance transpires electronically, it is a bonus to have the live expert next door or down the hall for personal interaction.

In many respects, starting from scratch is easier than changing an existing environment. However, evolution is to be expected. Even within a 2-year span many changes have occurred as advanced products became available and new research challenges arose. Careful initial planning can certainly make such evolution easier.

RECOMMENDATIONS TO THE PROFESSION

As much as being a new era in computing, the penetration of science by computers is a revolution in communications. The advent of electronic mail facilities and computer typesetting marks the beginning of a process at whose end we might see the virtual disappearance of printed publications and proceedings. Some journals have begun accepting submission of papers as typeset files via electronic mail, and there is no reason in principle why distribution of reprints and even entire journals to readers could not follow the same routes. The leading statistical societies could agree on one or two typesetting systems as standards. Such standards have already been adopted in mathematics and physics.

Drawing further on examples from other disciplines, we would like to mention numerical analysis, where a very lively computer network with a central node at Stanford has been established. It consists basically of a large electronic mail box on the node computer and takes advantage of the many already existing networks. In spite of the apparent overhead in sending electronic mail to a central node rather than the shortest known route, there are several overwhelming advantages to this concept:

- If an individual moves, there is only one item in the central mail box to be updated, and one does not have to notify a large number of colleagues of imminent address changes.
- Users will not have to be bothered with the historically grown idiosyncrasies of electronic mail address syntax. Due to the existence of several networks (like ARPAnet, CSnet, BITnet, USEnet), straightening out addresses can be a nontrivial problem. With a central mail box there is a one-time effort of finding the correct address.
- Individuals can subscribe to special interest groups. For example, this could link together statisticians with interests in generalized linear models and facilitate the sharing of new problems and results on this topic.

- In a similar way, speakers, organizers, and attendees of meetings can be electronically connected to facilitate quick and informal exchanges at all levels, largely replacing slow and costly conventional mail.

Regarding the last point, in the numerical analysis network, small meetings are often announced very early to tailor contents according to the requests from the community. This may be a more democratic way of managing meetings if insider circles can be put under pressure from the outside.

A central computer node run by the leading societies of our profession could provide further services, like maintaining and updating various databases useful to the general audience. One of them could be an on-line equivalent of a professional society directory which could be queried by sending it an electronic mail message. Also, one could establish a database of abstracts of technical reports, collected from statistics groups at universities, industrial laboratories, and journals. Full copies of reports could then be ordered from and sent by the authors or their departments, again preferably via electronic mail. An exciting aspect of this facility would be the capability for keyword search among reports and papers.

Another type of database maintained at a central node could be a collection of data sets, documented according to standards to be established. Access to the database would be provided on request from researchers trying out new methods as well as from teachers in need of data sets for their classes. One could go even farther and keep a record of short accounts of data analyses done in the past, with new accounts being incorporated as data sets are reanalyzed. This should ultimately result in an interesting history of data analysis by way of multiply analyzed data sets. A database of statistical software, or at least pointers to programs, could be tremendously useful, too.

We should point out that the proposals made above, although formulated in terms of a central node computer, could be realized in more decentralized ways as well. Schemes exist whereby copies of some or most of the databases could be kept locally in every computer, with some nodes providing updates periodically. We do not propose a particular implementation but the provision of services and capabilities of interest to the profession.

Last, we recommend that the statistical societies designate a few individuals with suitable expertise as consultants on computing issues for the profession. We hesitate to call this yet another "committee," but we sense that there are a sizeable number of statistics departments which have not yet developed sufficient expertise in computing matters and who need someone to turn to for start up help. It seems to us that some of the panel members, whose report we have at hand, would be especially qualified for this task.

Comment

David W. Scott

1. ABSTRACT

Eddy and his coauthors are to be warmly thanked for bringing together such a complete array of information for creating successful research computing environments at the departmental level. I have added a few observations, comments, and predictions directed to the very difficult task of communicating the "feel" of a good computer environment. I believe the authors' report should have a positive effect accelerating the availability of quality computing for statistical research.

2. EXPANDING COMPUTER HORIZONS

Some 5 years ago in the **R** room overlooking Rice stadium, a data processing manager for an oil company addressed an assortment of academic and industrial computer experts. He remarked that for his large seismic processing operation to be successful, it required, in part, at least a 60-day backlog of jobs. Ten years ago when I graduated I believed that on-line computing was too expensive and wasteful of my time, encouraging unproductive experimentation rather than careful and thoughtful program development possible with batch processing, which of course was the only kind of computing available to me. While I cannot speak for the first gentleman, my own attitude toward computing has undergone some changes. My primary computing resource has seldom lasted longer than 2 years, beginning with an IBM 1620; in graduate

David W. Scott is Professor of Mathematical Sciences, Rice University, Houston, Texas 77251. These comments were written while the author was on sabbatical leave at the Department of Statistics, Stanford University.