STATISTICS RESEARCH:
THE NEXT TEN YEARS

by

David S. Moore
Purdue University

Technical Report #95-9

Department of Statistics
Purdue University

April 1995

Statistics Research:
The Next Ten Years

David S. Moore
Purdue University
West Lafayette, Indiana, USA

Invited contribution to a forum on "The Future of Statistical Research," *Statistics and Computing*, **3** (1993), pp. 200–201

I once spent an undergraduate summer with my left hand chained to a desk calculator, calculating double star orbits for a distinguished astronomer. Each lengthy calculation produced a point, which I plotted on graph paper. Theory required that these points, which contained a good deal of observational error, form a parabola. After some weeks of calculation, I therefore sharpened a pencil and drew a parabola through the scattered points by hand. The distinguished astronomer was pleased; he said it was clear that I was a mathematics student, because the parabola I drew "looked like a parabola." When I arrived at Purdue in 1967, I repeated this story to an older colleague. He replied that he had once visited some engineering faculty who also had a theory that demanded a parabola. They plotted their data on a large piece of paper, put it on the wall, and fit a parabola by hanging a chain over the paper.

Where we are. The point of these anecdotes from what now seem like the dark ages is that things have changed. The computing revolution has made data-analytic procedures much more sophisticated than fitting a parabola easily available to engineers and scientists everywhere. General journals such as *Science* carry brightly-colored advertisements for software packages that promise to carry out complex statistical analyses as well as to prepare elaborate presentation graphics in several colors. Astronomers now fit parabolas instantly, and they no longer hire undergraduates to calculate double star orbits by hand. Expectations rise with prosperity, of course. Present-day astronomers are more likely to search large data bases of the Doppler redshifts of galaxies for voids and filaments and to ask if these features are "significant" in the sense of requiring systematic explanation.

Statisticians have not been slow to take advantage of fast and cheap computing to extend their range. Old methods such as regression now come equipped with a bewildering variety of diagnostic tools. More general classes of models (generalized linear models, generalized additive models) describe a wider variety of processes. Bootstrapping produces error estimates and confidence intervals in previously intractable settings. Each year seems to bring new ways of smoothing data by fitting very general classes of functions, so that kernels, splines, and wavelets compete for use, not to mention the irresistibly-named "supersmoother." The most important statistical research of the past twenty years has brought us this cornucopia.

Where we need to go. What then will be the most important statistical research of the next ten years?

A look around us reveals that most students and many users of statistics, and not a few statisticians as well, are a bit overwhelmed. Beginning students often fail to grasp the reasoning of tests of significance. More advanced students struggle to decide which multiple comparisons method to use, and to understand why the size and significance of the regression coefficient for an explanatory variable depend so strongly on what other variables are in the model. Many users don't yet have a firm hold on these issues, but their needs push them on to questions of study design and to more complex (but always automated) analyses. The standards of many fields (medicine comes to mind) now require very much more sophisticated statistical analyses than was once the case. Having mastered linear and logistic regression, the researcher now finds that she is expected to make sense of published studies that use Kalbfleish's extension of Cox's semiparametric regression model for censored data. Moreover, she is given to understand that this is standard and rather old methodology.

I propose that *the most important research of the next ten years will aim to provide, through technology, students with tools for learning and users with tools for choosing methods and understanding their results.* This proposition recognizes the increasing emphasis (by society, if not by academics) on technology transfer rather than basic research. Statisticians have a great deal of impressive and useful technology to transfer, and we ought to pay more attention to better means of doing so. I do not at all suggest that we have "enough" statistical technology, or that additional important statistical techniques will not emerge in the next decade. I only propose that technology-based help for students and users is the single most important advance we are likely to see.

Is this statistical research? The same question could be asked of work on detecting voids and filaments in the positions of galaxies, or on image-processing, or on statistical software. Statistics is a methodological discipline, and our most important advances often occur on the many boundaries we share with other disciplines. In particular, most statisticians now recognize that work on the interface between statistics and computer science has had great influence on statistical practice. Developing better tools for aiding learning and assisting judgment requires expertise in both computer science and statistics, but also in learning theory and human factors. Software designers, who understand the great importance of the "user interface" to the effectiveness of software, are ahead of statisticians in recognizing that a mix of "soft" disciplines with "hard" technical knowledge is increasingly necessary.

How will we get there? It is of course not clear just what form the new technology will take. But some paths are becoming apparent. First, *learning and receiving guidance will be highly interactive.* Research on learning emphasizes that students learn by their own activity, not by passively receiving information. Working statisticians and users of statistics make much the same point when they emphasize the importance of actual experience with data. Another finding is that "multiple linked representations" of a phenomenon (e.g., by several dynamically linked graphs as well as by a model) are very helpful for understanding. Some statistical software already offers means of interacting with data, linking several representations, and manipulating objects such as graphs and models directly. No doubt these trends in software will be extended and unified.

Second, *users will interact with a multimedia system.* As video becomes digitized, the boundary between computing and video is breaking down. The user at the end of the next decade will face a screen that offers text, dynamic graphics, full-motion video and sound as well as computational capability. The system will respond to keyboard, mouse and voice. (This is a conservative vision; virtual reality technology will eventually go much further in integrating media and responding to the user.) Students, for example, will see a video of the process to be analyzed, and perhaps even slow the video in order to take measurements from the screen. Sound and dynamic graphics will combine with text for exposition, problem posing and response to student questions and efforts at analysis. An analysis window will be continuously available for carrying out any numerical or graphical analysis suggested by the student.

Third, *an intelligent interface will guide the user.* The pace of development in artificial intelligence and expert systems has been slower than was first hoped. It has proved easy to automate what humans are bad at, such as immense calculations, and difficult to automate what we are good at, common sense and integrating background information. As the old joke says, you know it's *artificial* intelligence if it continues to make brilliant chess moves when the room is on fire. Yet substantial intelligence is essential if users and students are to be offered real help. A user who wants to describe a response surface needs guidance on both design of data production and model fitting, guidance that is adapted to the users's individual research setting (are theoretical models available?), goals (qualitative description, prediction, search for maxima?) and level of knowledge.

Advocating development of interactive systems to help students learn and guide users' judgments does not mean that I expect them to replace human teachers and consultants. The rising level of statistical methodology in medical research, for example, has made participation of a statistician more important despite much more capable software. Technology can, however, make learning more efficient by encouraging individual student activity and enable statistical consultants to focus on unusual or advanced aspects of the researchers' substantive problems.

It is unlikely that developing technological tools for assisting students and users will win as much prestige in academic statistics departments as inventing yet another method for nonparametric regression. But the test of an innovation in statistics is whether it is used. Tools for guiding and accelerating the progress of students and users will not only be used themselves but will greatly increase the use of other statistical innovations. That leverage marks their singular importance.

Some references

1. Biehler, R. (1993) Software tools and mathematics education: The case of statistics. To appear in W. Dörfler, C. Keitel and K. Ruthven, eds., *Learning form Computers: Mathematics Education and Technology.* Springer, Berlin.

2. Chambers, J. M. and Hastie, T. J., eds. (1992) *Statistical Models in S.* Wadsworth & Brooks/Cole, Pacific Grove, California.

3. Feigelson, E. D. and Babu, G. J., eds. (1992) *Statistical Challenges in Modern Astronomy.* Springer, New York.

4. Garfield, J. (1993) How students learn statistics. *International Statistical Review*, to appear.

5. Grabinger, R. S., Wilson, B. W., and Jonassen, D. H. (1990) *Building Expert Systems in Training and Education.* Praeger, New York.