Greater or Lesser Statistics: A Choice for Future Research

John M. Chambers
AT&T Bell Laboratories, Murray Hill, New Jersey

Abstract

The statistics profession faces a choice in its future research between continuing concentration on traditional topics, based largely on data analysis supported by mathematical statistics, and a broader viewpoint, based on an inclusive concept of learning from data. The latter course presents severe challenges as well as exciting opportunities. The former risks seeing statistics become increasingly marginal in activities to which it can make important contributions.

This paper is one of a set of short position papers on future research directions in statistics invited by the editor of *Statistics and Computation*.

1 Views of Statistics.

This paper contrasts two views of statistics, *greater* and *lesser* statistics. Greater statistics can be defined simply, if loosely, as everything related to *learning from data*, from the first planning or collection to the last presentation or report. Lesser statistics is the body of specifically statistical methodology that has evolved within the profession — roughly, statistics as defined by texts, journals, and doctoral dissertations. Greater statistics tends to be inclusive, eclectic with respect to methodology, closely associated with other disciplines, and practiced by many outside of academia and often outside professional statistics. Lesser statistics tends to be exclusive, oriented to mathematical techniques, less frequently collaborative with other disciplines, and primarily practiced by members of university departments of statistics.

Future directions for statistics research will reflect the tendency of statisticians to define their work consistently with one or the other of these views.

A general drift towards lesser statistics has been underway for several generations, reflecting a natural desire to give statistics its own theoretical basis, but limiting both the influence of statistics and the benefits the field has provided to society, particularly in recent years.

Statistical computing had the potential to widen our view. To be valuable, statistical software must be integrated into the whole process of learning from data. Observation of current activity suggests, however, that most research in statistical computing remains work in service of lesser statistics, such as support for classical probability-based theory. This may, in its context, be valuable and should not be denigrated. Certainly, it is relatively safe and likely to be appreciated by one's professional colleagues.

The need to learn from difficult but important sources of data provides a strong impetus to research in greater statistics outside of traditional topics. The impetus is strong enough that much of the research advances needed *will* be carried out, by someone. If statisticians remain aloof, others will act. Statistics will lose; in addition, I believe science and society will lose also, because the statistician's mental attitude at its best provides qualities likely to be missing otherwise.

2 Greater Statistics: Learning from Data.

Three broad categories characterize work in greater statistics:

1. *preparing* data, including planning, collection, organization, and validation;

2. *analyzing* data, by models or other summaries;

3. *presenting* data in written, graphical or other form.

Greater statistics suffers from the feeling that statistics' proper domain is the second item, indeed the portion of it that involves probabilistic inference. The rest is viewed as of less intellectual interest and little research potential. This is a fundamental misconception: each aspect of learning from data is rich with intellectual challenges as well as practical importance. Statistics must embrace them *all* if its potential as a discipline is to be realized.

The next section will concentrate on examples from data preparation, the aspect of greater statistics most shamefully underrated and most desperately in need of new insights. However, each aspect presents challenges that will lead to research of both intellectual and practical value, *provided* we can broaden our viewpoint to take them up.

3 One Aspect: Data Preparation.

The argument for greater emphasis on research in data preparation rests on three assertions: the subject is of great practical importance, challenges our best research abilities, and benefits from the unique viewpoint of statisticians.

The most important part of data preparation is often conceptual: what meaningful data can be collected in a particular context, how can the data be organized, and what in broad terms can we expect to learn from them? Data collection needs to mesh both statistics and subject matter insights, just as does data analysis. The statistician brings insights from an understanding of science that is the essence of our professional craft, but which is only partly expressible in terms of traditional mathematical statistics. It is our version of the scientific method, with an emphasis on empirical understanding and a healthy skepticism about theory for its own sake: in short, a methodology for learning from data. It is worth repeating that these insights are as important to data collection and preparation as they are to analysis.

To appreciate some of the research challenges in data preparation, consider a few examples.

1. Many mundane commercial and social activities generate large quantities of potentially valuable data. Examples from business include retail sales, billing, and inventory management. The data were not generated for the purpose of learning; however, the potential for learning is great, if we can cope with some major challenges. The data usually pass through a computer system nowadays, but aside from the enormous quantity, the data are typically not centralized, are inconsistent in organization, and are usually thrown away fairly quickly. The computational challenge of collecting and organizing such data is huge. A more clearly statistical challenge is that the data may represent only a portion of the conceptually relevant data; if so, the sample is often biased in crucial ways.

2. Information networks, such as long-distance telephone networks and local- or wide-area intercomputer networks, are increasingly central to many activities. Understanding how such networks work and how they can be managed effectively is important, difficult and fascinating. The challenge is to learn from the data describing such networks. The data can again be voluminous (several megabytes per hour for some examples), but even more central is that they must be organized

around very different concepts from traditional statistical models. The spatial geometry, the network flows, and the evolution through time must all be considered, while aiming at the underlying activities (e.g., messages) that generate the data but that often cannot be observed directly.

3. Manufacturing of sophisticated, rapidly evolving devices, such as computer components, requires extensive testing. In contrast to the previous examples, the data from these tests *are* collected specifically to learn something; essentially, how well the new device performs. Once again, however, how to think about and prepare the data is a major challenge. Traditional engineering approaches collected large amounts of data, produced a few 'quality' summaries, and threw the rest away. Statistical involvement in a real interdisciplinary approach is leading to a richer concept of what the data really are. The results contribute to research advances in engineering as well as in (greater) statistics.

These three examples represent areas in which my colleagues at AT&T Bell Labs have participated. In each case, statistics has contributed substantially to an interdisciplinary collaboration and statisticians have found interesting new research. In our local context, the work and its research content has been appreciated. Relatively little of it, however, fits into the tradition of lesser statistics. To encourage and reward such research is the challenge for our profession, if statistics is to have a wider role.

The practical importance of data preparation is easy to defend: data collection, organization, and validation often takes the majority of the human time and effort expended on substantial interdisciplinary projects, even though this effort is typically not emphasized in reporting the work. No such project can proceed until the analyst is satisfied that the data are what they are claimed to be, and are adequately free from mistakes of specification or recording. Described by terms such as data acquisition, data cleaning, or data verification, these activities are often the limiting constraint on the project.

Data preparation is, as argued above, broader than these activities. However, the data acquisition and cleaning aspects themselves present interesting challenges. A prevailing negative view holds that these activities are tedious, difficult, specialized to the particular project, and unlikely to yield interesting (that is, publishable) research results. This view itself suggests some of the research potential. The tediousness and difficulty of the task as currently handled emphasize the value of even imperfect approaches to data preparation software and other tools. Progress in this area will reduce a substantial burden and cost. The specialization of knowledge about data cleaning to individual projects means that we need to express such knowledge in suitable computational and statistical language (some work on constraint systems points to approaches) so that it can begin to be reused and communicated. Research on data cleaning and verification will be rewarding and exciting. Again, can we accommodate it to our view of statistics?

4 Do Statisticians Matter?

The examples above, and many others, represent practical and intellectual challenges. Computer software and other tools will be developed in response; in fact, considerable existing software to some extent addresses these needs. Is it then important whether statistics as a profession takes a central part in this work? By and large, statisticians have a record of participating only marginally in many recent new research areas where our expertise and interests are relevant, such as expert software, scientific visualization, chaos theory, and neural networks. Whether our relatively small involvement in these fields represents timidity or wise avoidance of temporary fads can be argued in each individual case. However, the overall pattern is that of a profession neglecting opportunities to broaden its role.

If statisticians do not take up the challenges of the greater statistics viewpoint, both the profession and the larger community will be losers. The practical importance of these needs is large and growing. Good work will be well rewarded, intellectually and practically. A more serious point, perhaps, is that statistics *does* have a unique contribution to make. Without it, an important component is likely to be missing from solutions to these problems.

Statistics at its best provides methodology for dealing empirically with complicated and uncertain information, in a way that is both useful and scientifically valid. The problems arising in greater statistics very much need such methodology. One example will have to suffice. Data cleaning and verification is to some extent handled by current database management systems and by software explicitly designed for data acquisition. The approach taken can be summarized as constraint-oriented; that is, the software allows the user some mechanism for constraining the data in various ways (ranges for the values of individual variables, equality or inequality relations among variables, etc.). This approach is obvious to a computer scientist. However, a statistician thinking about the problems at all seriously would, I believe, point out a fundamental inadequacy of the constraint approach. Knowledge that can be specified so simply and precisely represents only part, perhaps a small part, of the relevant knowledge about data being acquired. A serious attack on the problem involves more difficult kinds of knowledge, such as relations whose violation is surprising but not logically impossible, or relations whose form might be assumed but which depend on unknown parameters. An approach to data validation that incorporates good statistical thinking in both the testing and the display of possibly anomalous data would be much more valuable (and also much more difficult) than current techniques based on computational ideas alone.

5 What Needs to be Done.

While pointing out desirable new directions is all very well, the exercise has little value if no practical path leads there from here. The considerations I have grouped under the term greater statistics relate to concerns expressed by many statisticians, particularly those involved with applications. An effective response, however, can only come with the efforts of university statistics departments. Whatever is done elsewhere, it is the approach of academic statistics to research and to training that will determine where statistics goes in the next decade. Many statistics departments have attempted to improve the training offered in data analysis, statistical consulting, or statistical computing in recent years. Other steps must be taken if anything like the greater statistics viewpoint is to spread.

The key conceptual hurdle is simple, but difficult. The view of acceptable research by faculty or graduate students must be broadened to the inclusive, multi-disciplinary perspective required for greater statistics rather than the traditional theoretical basis of mathematical statistics. A number of specific implications follow, among them:

- Much of this research will be done by designing and implementing computer software. We must train students in these activities; this training is *not* provided by teaching the use of statistics packages to support standard analyses.

- Assuming the research can be done, it must be evaluated, good work must be 'published', and its authors must be given professional rewards. Dealing with software-as-research in each of these ways will require new approaches in the profession.

- Interdisciplinary projects involving all aspects of learning from data must be an accepted part of research activity for statistics faculty. Involving students along with faculty in such projects is essential to foster the view of greater statistics.

- Research must lead to teaching. Teaching requires good source material, in the form of papers and books that present important techniques for greater statistics. Here the non-academic community must contribute. We need a few good books on under-documented aspects of greater statistics.

Any efforts that university statistics departments can take in these directions need to be supported by enlightened funding agencies. Co-operation between academic and non-academic statistics groups, which has improved substantially over the last generation, needs to contribute through support, teaching tools, and individual involvement. The evolution will be difficult, but only by such an evolution will statistics ensure itself the important future role it deserves.

Acknowledgements

For comments on the current paper, thanks to colleagues Bill Cleveland, Mike Luvalle, and Daryl Pregibon. The underlying ideas owe too much to too many to acknowledge, but two names deserve mention, Sir David Cox and John W. Tukey, both of whom have offered many related insights over the years.