[A Report on the Future of Statistics]: Comment
Author(s): David Madigan and Werner Stuetzle
Source: *Statistical Science*, Vol. 19, No. 3 (Aug., 2004), pp. 408-409
Published by: Institute of Mathematical Statistics
Stable URL: http://www.jstor.org/stable/4144388
Accessed: 13/05/2014 17:28

acknowledge the fact that these ideas are circulating everywhere and being tackled by all the mathematical sciences—mathematics, operations research, computer science—not to mention subject matter areas such as physics, which have a tradition of stochastic models as old if not older than ours. Everyone is interested in learning from data. Let us not try so hard to distinguish ourselves from other fields, but just do!

Comment

David Madigan and Werner Stuetzle

GRADUATE STATISTICS EDUCATION

The inexorable rise of computing and large-scale data storage has impacted most academic disciplines, sometimes in profound ways. Biology, for example, has become an information science where the tools of data analysis are as commonplace as the microscope. In astronomy, the study and analysis of vast stellar databases takes center stage. In the business arena, financial markets generate rivers of intensely scrutinized data, and all major global-scale retailers store and analyze vast quantities of customer and transaction data. The trend is universal and unstoppable.

Extraordinary opportunities for statistical ideas and for statisticians now present themselves. However, to take advantage of the opportunities, statistics has to change the way in which it recruits and trains students. Statistics has primarily focused on squeezing the maximum amount of information out of limited data. This paradigm is rapidly diminishing in importance and statistics education finds itself out of step with reality. The problems begin at the high school and undergraduate levels, where the standard course includes a narrow set of pre-computing-era topics. At the graduate level, the typical statistics program suffers from the same problem. Most programs focus primarily on problems of estimation and testing, where mathematics brilliantly finesses a paucity of computing power. The demand for graduates of such programs is real and possibly growing. However, students emerging from our programs are ill-prepared to engage in cutting-edge research and collaboration in the burgeoning information-rich arenas.

Statistics as a discipline exists to develop tools for analyzing data. As such, statistics is an engineering discipline and methodology is its core. We should prepare graduate students for methodological research, and note that in methodological research computer science plays a role that is comparable to the role of mathematics. Hence we should try to attract students who are not only mathematically adept, but who also have a background and interest in computing and an inclination toward collaborative research. Unfortunately, however, the current situation is that the computing skills of our incoming and outgoing graduate students are often woeful, and their experience with meaningful collaborative research is nonexistent. "Computing skills" here does not refer to the ability to write a SAS or C program. Rather it refers to the ability to design and evaluate algorithms for computationally challenging statistical methodology.

We examined the statistics Ph.D. programs at 12 major U.S. universities. Almost all of these programs included core courses in statistical estimation and testing, generalized linear models, probability theory and applied statistics. Most research statisticians would probably recognize these as the courses they took in graduate school. However, we contend that a statistician trained only in this manner lacks the skills needed to tackle the kinds of challenges that now present themselves. In particular, most standard programs eschew exposure to and practical experience with the following topics:

- Predictive modeling beyond the classical linear model.
- High-dimension, low sample size statistical analysis.
- Analysis of data that are not in spreadsheet form, such as text data, relational databases and streaming data.
- Bayesian data analysis.

David Madigan is Professor, Department of Statistics, Rutgers University, Piscataway, NJ 08854-8019, USA (e-mail: madigan@stat.rutgers.edu). Werner Stuetzle is Professor, Department of Statistics and Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195-4322, USA (e-mail: wxs@stat.washington.edu).

- Hierarchical and multilevel modeling.
- Causal analysis.
- Design and analysis of algorithms and data structures.

The notion of a common set of core courses for all graduate students is no longer tenable. A strong program will include courses or course sequences in some or all of the topics we list above, in addition to the usual sequences in probability, theoretical statistics and applied statistics. Students would choose among these sequences according to their research interests and talents.

The issues we raise above have nothing to do with the old distinction between applied statistics and theoretical statistics. The traditional viewpoint equates statistical theory with mathematics and thence with intellectual depth and rigor, but this misrepresents the notion of theory. We agree with the viewpoint that David Cox expressed at the 2002 NSF Workshop on the Future of Statistics that "theory is primarily conceptual," rather than mathematical. For example, recent outstanding texts such as *The Elements of Statistical Learning* (Hastie, Tibshirani and Friedman, 2001) or *Learning with Kernels* (Schölkopf and Smola, 2002) are not mathematics texts per se, yet they present primarily theoretical content.

REFERENCES

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. Springer, New York.

SCHÖLKOPF, B. and SMOLA, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond.* MIT Press, Cambridge, MA.

Comment

Marianthi Markatou and Bruce Levin

We would like to congratulate the editors of this report for coherently and succinctly summarizing the challenges and opportunities for the field of statistics as we enter the twenty-first century. The report is the culmination of discussions that took place during the workshop held in May 2002 at the National Science Foundation (NSF). The purpose of this workshop was to assess the current status of the field of statistics, to identify the challenges and opportunities that statistics faces and to develop a strategy for how to position the field to meet its current and future demands. This report also clarifies the often misunderstood role of the statistical sciences and illustrates its position in, and impact on, the scientific enterprise.

Three main themes are addressed in the report: (1) a wealth of interesting and difficult research problems generated by the interaction of statistics with other subject-matter areas, (2) education and (3) resource requirements to meet research needs and educational demands. We will briefly comment on each one of these aspects.

The report makes it clear that this is an exciting time for statistical research. Many important and interesting areas of interdisciplinary work are described in the report. Under the heading of information technology we would here like to add biomedical informatics, with its subareas of clinical informatics and public health informatics. Clinical informatics is the science of effectively extracting and using information in patient care, clinical research and medical education with the ultimate goal of improving quality of care and reducing costs. Public health informatics is the application of information and technology to public health research and practice (Friede et al., 1995).

What connects biomedical informatics with the field of statistics is, among other things, a set of challenges posed by the analysis of data collected both on healthy people and on patients. The problem of contamination (robustness) is very important in this context. If 5% of a large data set amounts to another large data set, the behavior of these different points needs to be explained and addressed. In other words, the outliers may be of medical significance and their behavior needs to be understood. Modeling, especially the creation of predic-

---

Marianthi Markatou is Associate Professor, Department of Biostatistics, Columbia University, New York, NY 10032, USA (e-mail: mm168@columbia.edu). Bruce Levin is Professor and Chair, Department of Biostatistics, Columbia University, New York, NY 10032, USA (e-mail: bruce.levin@biostat.columbia.edu).