

Statistics Ready for a Revolution

Next Generation of Statisticians Must Build Tools for Massive Data Sets

Mark van der Laan, Jiann-Ping Hsu/Karl E. Peace Professor in Biostatistics and Statistics at UC Berkeley, and Sherri Rose, PhD candidate at UC Berkeley

The statistics profession has reached a tipping point. The need for valid statistical tools is greater than ever; data sets are massive, often measuring hundreds of thousands of measurements for a single subject. The field is ready for a revolution, one driven by clear, objective benchmarks by which tools can be evaluated.

The new generation of statisticians must be ready to take on this challenge. They have to be dynamic and thoroughly trained in statistical concepts. They have to work effectively on an interdisciplinary team and understand the immense importance of objective benchmarks to evaluate statistical tools. They have to produce energetic leaders who stick to a roadmap, but who also break with current practice when necessary.

Why do we need a revolution? Sadly, 99.99% of all data analyses are based on the application of so-called parametric (or other restrictive) statistical models that assume the data-generating distributions have specific forms. Many agree that these models are wrong.

That is, statisticians know linear or logistic regression models and Cox

proportional hazards models are specified incorrectly. But, they still use them to draw conclusions and then hope these conclusions are not too wrong.

The original purpose of a statistics model was to develop a set of realistic assumptions about the probability distribution generating the data set (i.e., incorporating background knowledge). However, restrictive parametric models are almost always used because standard software is available. These models also allow the user to obtain p -values and confidence intervals for the target parameter of the probability distribution, which are desired to make sense out of data.

Unfortunately, these measures of uncertainty about our estimates are even more susceptible to bias than the effect estimates. We know that for large enough sample sizes, every study—including ones in which the null hypothesis of no effect is true—will declare a statistically significant effect.

Some practitioners will tell you they have extensive training, are experts in applying these tools, and should be allowed to choose the models to use in

response to the data. Be alarmed. It is no accident that the chess computer beats the world champion in chess. Humans are not as good at learning from data and easily susceptible to beliefs about those data.

For example, an investigator may be convinced his or her data have a particular functional form, but if you bring in another expert, his or her belief about the functional form may differ. Or, many models may be run, dropping variables that are non-significant in each model. While this is common, it leaves us with faulty inference.

With high-dimensional data, not only is the correct specification of the parametric model an impossible challenge, but the complexity of the parametric model also may increase so that there are more unknown parameters than observations. The true function also might be described by a complex function not easily approximated by main terms.

For these reasons, allowing humans to include only their true, realistic knowledge (e.g., treatment is randomized, such as in a randomized controlled trial, and our data set represents

an independent and identically distributed observations of a random variable) is essential.

What about machine learning, which is concerned with the development of black-box algorithms that map data (and few assumptions) into wished objects? Indeed, this is in contrast to using misspecified parametric models, but the goal is often the whole prediction function, instead of particular effects of interest.

Even in machine learning, however, there is often unsupported devotion to beliefs. In this case, to the belief that certain algorithms are superior. No single algorithm (e.g., random forests, support vector machines, etc.) will always outperform all others in all data types, or even within specific data types (e.g., SNP data from genomewide association studies). One can't know a priori which algorithm to choose. It's like picking the student who gets the top grade in a course on the first day of class.

The concept of a model is also important. We need to be able to incorporate true knowledge in an effective way. In addition, we need such data-adaptive tools for all parameters of the data-generating distribution, including parameters targeting causal effects of interventions on the system underlying the data-generating experiment. The latter typically represents our real interest: We are not only trying to sensibly observe, but also to learn how the world operates.

The tools we develop must be grounded in theory, such as an optimality theory, that shows certain methods are more optimal than others. For example, one can compare methods based on mean squared error with respect to the truth. It is not enough to have tools that use the data to fit the truth well. We also require an assessment of uncertainty, the very backbone of statistical learning. That is, we cannot give up

on reliable assessment of uncertainty in our estimates.

The new generation of statisticians cannot be afraid to go against standard practice. Remaining open to, interested in, and a developer of newer, sounder methodology is perhaps the one key act statistics students can perform. We must all continue learning, questioning, and adapting as new statistical challenges are presented.

The science of learning from data (i.e., statistics) is arguably the most beautiful and inspiring field—one in which we try to understand the very essence of human beings. However, we should stop fooling ourselves and actually design and develop powerful machines and statistical tools that can carry out specific learning tasks.

There is no better time to make a truly meaningful difference. ■

Applet Fun

Below are a few websites with collections of applets and other resources to help teach and learn statistics.

ASA Section on Statistical Education

www.amstat.org/sections/educ/applets.html

www.amstat.org/sections/educ/statedlinks.html

ASA Education Useful Sites for Teachers

www.amstat.org/education/usefulsitesforteachers.cfm

CAUSEweb

www.causeweb.org/cwis/SPT--BrowseResources.php?ParentId=1

PASS


**Power Analysis
and Sample Size**

GESS

Microarray Analysis


NCSS

**Statistical Analysis
and Graphics**



**PASS 2008
Upgrade Now
Available!**

sales@ncss.com
1-800-898-6109



Order Today at:
www.ncss.com