Attracting Undergraduates to Statistics Through Data Science

Jim ALBERT and Mark GLICKMAN

We agree with George Cobb that statisticians need to rebuild their undergraduate curricula in statistics in the wake of big data and the many opportunities for employment in Data Science. As Cobb notes, our statistics curricula are currently facing several threats such as big data in computer science and analytics in business, and we agree that it is high time for statisticians to seriously rethink our undergraduate curriculum.

In particular, we believe the one-year sequence in probability and mathematical statistics, the standard introduction to statisticians for the past 50 years, is no longer a suitable foundation for training for a modern applied statistician. At Bowling Green, one of us has been fortunate to participate in the creation of a new major in Data Science within a department of mathematics and statistics. At Harvard, where one of us has been visiting faculty for 10 years, a new Data Science track for the statistics concentration is actively under development. Here we focus on what we believe are the important components of a data science program/track that can attract majors and provide a good foundation for employment as a data scientist.

Introduce Statistics Through Exploratory and Visualization Methods

For students with minimal statistics prerequisites, a good foundation course in a data science major focuses on computation with data. A version of this course is already in existence at Harvard, and Baumer (2015) and Hardin et al. (2014) describe similar data science courses. The student learns basic methods for importing, manipulating, and exploring data using a scripting language such as R or python. Particular data wrangling tools such as the use of regular expressions for textual data play an important role since text data is representative of modern data with a different structure from the traditional "data frame" rectangular grid. This course provides a good opportunity to learn about categorical and quantitative variables, as well as data management that aids in accessing elements of large data sets. Many of the themes of Tukey's exploratory data analysis can be introduced in this computing with data course. Additionally, such a course can emphasize communication of results through visualization and interpretable summaries, including generation of hypotheses by simple data explorations.

Statistical Programming

Much of the work of a data scientist consists of the process of data collection and data wrangling and it seems clear that the student needs sufficient training in a statistics programming language. One required course in the new Bowling Green data science program is "Statistical Programming". This course is an in-depth look at data types and containers, and the students get experience writing scripts and functions for different data science tasks. The tools for collecting, managing, and visualizing data are changing quite rapidly and the student with a solid foundation in a statistics language such as R will likely be able to adapt to these new data science tools.

The Importance of Context

The computing with data course uses several interesting big data applications to demonstrate the value of statistical thinking. Nolan and Lang (2015) describe a number of real-life data science projects, and ideas from these projects can help the instructor in building interesting homework assignments and projects. Ideally, a curriculum in data science can prepare the student to work on an extended data science project in collaboration with a faculty member from an applied discipline. Each undergraduate program needs to develop a network of internships, advisers and summer programs that can help in the development of these capstone projects.

From Statistical Inference to a Broad View of Statistical Algorithms

We agree with Cobb that the statistics curriculum needs to move away from traditional statistical inference courses with their emphasis on testing hypotheses and normal error and independence assumptions. But what should a modern statistical inference course look like? One possibility is to offer a statistical learning course. One of the required courses in the Bowling Green data science program is a course based on James et al (2013) that gives an applied overview of various statistical learning algorithms together with lab exercises on interesting datasets. Another alternative would be one based on generalized linear models (GLM), but this would require the students to have knowledge of a variety of probability distributions. In the Harvard GLM course that one of us teaches, which assumes students have had exposure to basic probability but not mathematical statistics, we have incorporated a data prediction competition on Kaggle as a course project. This type of exercise provides students an opportunity not only to engage in model criticism and refinement, but also to explore machine learning prediction algorithms. Students are exposed through this project both to stochastic and algorithmic cultures that Breiman (2001) identified. Given the increasing popularity of Bayesian methods, we think the time is right for the development of an applied Bayesian course. Link and Barker (2009) is one example of an applied Bayesian text that illustrates basic concepts within the

Online discussion of "Mere Renovation is Too Little Too Late: We Need to Rethink Our Undergraduate Curriculum From the Ground Up," by George Cobb, *The American Statistician*, 69. Jim Albert is Professor of Statistics, Bowling Green State University, Bowling Green, OH (Email: albert@bgnet.bgsu.edu). Mark Glickman is Research Professor of Health Law, Policy and Management, Boston University, Boston, MA.

context of interesting examples from a particular discipline.

The Intro Statistics Course?

We believe the most challenging task is to redesign the introductory statistics class. Too many classes focus on learning recipes and the student leaves the course with a distaste for the subject, and more troublesome a lack of appreciation of the discipline of statistics. One of us has incorporated magic tricks (Lesser and Glickman 2009) and class-participatory demonstrations (Gelman and Glickman 2000) as ways to enhance interest in introductory statistics concepts. One of us has had fun experimenting (Albert 2003) with a baseball version of our introductory class. The course arguably succeeds in part since the students are genuinely interested in the sports application and the statistics concepts make more sense when discussed in this context. Generally, any type of project in which the students get to implement all of the steps of a statistics investigation is one of the best ways of making the discipline real for the students.

References

Albert, J. (2003), *Teaching Statistics Using Baseball*, Mathematical Association of America.

Baumer, B. (2015), "A Data Science Course for Undergraduates: Thinking with Data," available online at arXiv:1503.05570v1.

Breiman, L. (2001). "Statistical Modeling: The Two Cultures" (with comments and a rejoinder by the author), *Statistical Science*, 16, no. 3, 199–231.

Cobb, G. W. (2015), "Mere Renovation is Too Little Too Late: We Need to Rethink Our Undergraduate Curriculum from the Ground Up," *The American Statistician*, 69(4), doi:10.1080/00031305.2015.1093029.

Gelman, A., and Glickman, M.E. (2000), "Some Class-Participation Demonstrations for Introductory Probability and Statistics," *The Journal of Educational and Behavioral Statistics*, 25, 84–100.

Hardin, J., Hoerl, R., Horton, N., and Nolan, D. (2014), "Data Science in the Statistics Curricula: Preparing Students to 'Think with Data,"' available online at arXiv:1410.3127v2.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning: With Applications in R*, Springer.

Lesser, L.M., and Glickman, M.E. (2009), "Using Magic in the Teaching of Probability and Statistics," *Model Assisted Statistics and Applications*, 4 (4), 265–274.

Link, W., and Barker, R. (2009), *Bayesian Inference: with Ecological Applications*, Academic Press.