Authentic Data Analysis Experience

Duncan TEMPLE LANG

I'd like to applaud and thank George for a very stimulating and entertaining paper, and the challenge to reinvent the statistics undergraduate curriculum. I am very hopeful that it will lead to real discussions, experimentation and, importantly, significant changes.

Do we need such radical changes to our undergraduate curriculum? To answer this, I hope every instructor will seriously reflect on whether their graduates have the capabilities to perform good data analyses? and whether we could be doing substantially better in this regard? My focus here is on data analysis, broader than statistics, as that is what the majority will do with what they have learned, and is increasingly in high demand.

In my opinion, most of our students are not prepared for data analysis after their statistics major. My explanation is simple—they have done very little actual data analysis. Instead, they have learned methods, solved homework problems corresponding to the method taught that week, and perhaps done a project or a single capstone course. They think there is "one correct answer" and that the data analysis process starts by being told to fit a model or perform a hypothesis test, and ends by reporting the parameter estimates or a p-value. For a model, they may have generated the expected diagnostic plots, but not necessarily have really interpreted them.

Many students enjoy the mathematics, and others lose sight of the "why" of the methods due to the mathematical and computational details. Many see only the deterministic aspects of the methods, and not the variability in the data and the approximations and rough precision needed for the insights and qualitatively solving real problems. They get drawn into the details of a test, only to forget to ask if the observations form the population or a sample, or are dependent. The statistical methods are important elements of data analysis, but there is so much more to the data analysis process than these methods and we don't spend much time teaching these other components. Exploring data is something they may feel compelled to do because that is "recommended", but is an obligation before the "real statistics" are done. Indeed, George says we teach EDA in a first course, but typically, this is labeled "Graphical Summaries" or "Descriptive Statistics" in textbooks. Unfortunately, the very important steps of cleaning and exploring of both the data and the problem are not emphasized as being essential parts of the data analysis process.

While a laudable goal is statistical reasoning, students need to develop, *at least*, a sense of intelligent data analysis, the ability to frame a data analysis question and identify the goals, and the skills to express the necessary computations and create graphical displays. They develop these by repeating the process multiple times, not just once in a capstone course. They must learn the process by first watching how data analysts actually work, via guided case studies. The ideas and motivation of common methods can be introduced at this point without the details. This is, as many have written, quite a different learning experience from presenting a long list of methods and their underpinnings that we typically teach in courses without the actual data analysis context and connecting the thinking to the question. The NSF-funded Explorations in Statistical Research (ESR) workshops (Nolan and Temple Lang 2015) that Deborah Nolan (UC Berkeley) and I organized exposed students to the data analysis process. While very short, they illustrate the need and potential for quarter/semester–long immersion with real data analysis.

Students enjoy exploring data when they understand what is being measured (e.g., cost of apartments in different cities, car traffic patterns, airline delays, climate change, social network patterns.) Students can acquire reasonable computational skills by exploring data. These are the skills they will need to process and analyze data. Given these computational skills, they can then simulate data and explore the characteristics of statistical methods. This can augment, or substitute for, the mathematical understanding for different students.

On reading George's paper, I was led to Friedman's 1997 article on Data Mining and Statistics (Friedman 1997), which led me back to Tukey's "The Future of Data Analysis" (Tukey 1962), and of course to Brown and Kass (2009), and other historical calls for changes to the curriculum. We should reflect on these papers and see how much has changed. Indeed, George writes about adoption of Bayesian statistics: "Statisticians read the arguments, followed the proofs, nodded in agreement, and continued in their pursuit of incoherence." All of these calls for change are important, and I believe statistics education is slowly changing. However, it is too slow and there are many other fields that are providing alternatives, and for teaching the actual practice of data analysis, this may be a good thing.

I recently became the director of the Data Sciences Initiative at University of California, Davis. While there are many logistical challenges, I feel liberated and less constrained. We have an opportunity to develop programs from the ground up, as George is encouraging. There are many constituents hungry for Data Sciences education and research skills that span the entire data pipeline, including data identification, acquisition, cleaning, exploration, visualization, analysis and dissemination of insights. This demand for data pipeline knowledge is something we should embrace. We should work with various different fields (both consumers and producers of data sciences methods) to create students with the essential fundamentals and problem solving capabilities needed in data science. Data science requires data analysis, computational reasoning, and actual experience and practice.

Deciding if and how we should change the curriculum in-

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. Duncan Temple Lang, University of California, Davis (Email: dtemplelang@ucdavis.edu)

volves clearly articulating and prioritizing our goals. For me, the ability to be creative, independent, problem solve, work in a team, be able to map ideas into computations and results, and, most importantly, to make sense of data are important for the majority of our students. Whether this involves more or less mathematics, computing, methods, . . . is up for debate and, importantly, experimentation and evaluation. Learning "just-in-time" or "on-demand" is an important skill for problem solving, and can help the students escape the "multiple-choice/one-correct-answer" mindset into which they are led from high-school through college. A mode of teaching that leads students to "discover" traditional statistical methods, rather than just being told them, will be much richer.

George's suggestions of a "Teacher's corner" that present innovations and experiences in teaching is a good idea. Facilitating instructors to develop and share case studies and projects would also be very useful. Having these recognized as scholarly contributions could help. While change in the curriculum is hopefully proceeding, we can act more quickly by having students participate in data analysis challenges. These might be as short as the weekend-long DataFest^{TM} that has emerged through UCLA, the ASA, and others, or 6–7 week long ongoing problem solving activities centered around real problems (e.g., extra-curricular team-based data analysis competitions).

Let's embrace the new opportunities that data science presents and include Statistics in the next generation of data analysis.

References

Brown, E. N., and Kass, R. E. (2009), "What is Statistics?" (with discussion), *The American Statistican*, 63, 105–123.

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.

Friedman, J. H. (1997), "Data Mining and Statistics: What's the Connection," in *Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics*.

Nolan, D., and Temple Lang, D. (2015), "Explorations in Statistics Research: An Approach to Expose Undergraduates to Authentic Data Analysis," *The American Statistician*.

Tukey, J. W. (1962), "The Future of Data Analysis," *Annals of Mathematical Statistics*, 33(1):1–67,