Data science and statistics

Data science is having a radical effect on the teaching of statistics. Many prominent teachers of statistics are arguing that courses in statistics should follow the data science model (see (Cobb 2015, 2007; Horton 2015; Hardin et al. 2015).

The argument is twofold.

First, most statistics courses are mis-named, because our primary aim is to teach students how to draw valid conclusions from real data, for which inferential statistics is a final and possibly optional step in a long journey of analysis. Teaching statistics rather than data analysis has the effect of pushing the analyst toward fitting the data to the analysis rather than the analysis to the data. Quoting Tukey (1962):

> Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.

Second, current statistics courses need far too much mathematics, and this is a major barrier to understanding. As Cobb (2007, 2015) argues, the traditional emphasis on mathematics and the normal distribution is an artifact of the lack of computing power available to pioneers of statistics at the beginning of the 20th century. We now have readable programming languages and vast computing power. Simple code allows us to use algorithms such as simulation and resampling to explain the ideas of statistical inference in a deeper and more natural way (Cobb 2007, 2015). See (Simon and Holmes 1969; Simon, Atkinson, and Shevokas 1976) for some convincing early examples of using resampling for teaching probability and statistics to high-school and university students. An added benefit of emphasizing algorithms is that students can see the logic of machine learning algorithms in the same framework as classical statistics, giving them more versatile approaches to data analysis.

Algorithms allow us to teach at greater depth in a shorter time, than standard teaching of correlation, t-tests and ANOVA. As Cobb points out (2007), once the students understand the fundamental ideas of inference through algorithms, we can extend the same principles to normal distribution approximations such as those used in the t-test. Once students have seen how to apply the ideas in their own code, and implementations in programming languages such as R or Python, they more quickly learn the interface for GUI tools such as SPSS and Minitab.

The benefits from teaching with data science methods are so dramatic that failure to use them will be a crisis for the future of statistics teaching. Cobb's 2015 paper (2015) has the title "Mere renovation is too little, too late; we need to rethink our undergraduate curriculum from the ground up". An eminent educator in statistics said that "Statistics education remains mired in the 20th (some would say the 19th) century." (Madigan et al. 2014).

References

Cobb, George. 2015. “Mere Renovation Is Too Little Too Late: We Need to Rethink Our Undergraduate Curriculum from the Ground up.” *The American Statistician* 69 (4). Taylor & Francis:266–82.

Cobb, George W. 2007. “The Introductory Statistics Course: A Ptolemaic Curriculum?” *Technology Innovations in Statistics Education* 1 (1).

Hardin, Johanna, Roger Hoerl, Nicholas J Horton, Deborah Nolan, Ben Baumer, Olaf Hall-Holt, Paul Murrell, et al. 2015. “Data Science in Statistics Curricula: Preparing Students to ‘Think with Data’.” *The American Statistician* 69 (4). Taylor & Francis:343–53.

Horton, Nicholas, ed. 2015. “Responses to: Mere Renovation Is Too Little Too Late: We Need to Rethink Our Undergraduate Curriculum from the Ground up.” https://nhorton.people.amherst.edu/mererenovation.

Madigan, David, Peter Bartlett, Peter Bühlmann, Raymond Carroll, Susan Murphy, Gareth Roberts, Marian Scott, et al. 2014. “Statistics and Science: A Report of the London Workshop on the Future of the Statistical Sciences.” http://www.worldofstatistics.org/wos/pdfs/Statistics&Science-TheLondonWorkshopReport.pdf.

Simon, Julian L, David T Atkinson, and Carolyn Shevokas. 1976. “Probability and Statistics: Experimental Results of a Radically Different Teaching Method.” *The American Mathematical Monthly* 83 (9). JSTOR:733–39.

Simon, Julian L, and Allen Holmes. 1969. “A New Way to Teach Probability Statistics.” *The Mathematics Teacher* 62 (4). JSTOR:283–88.

Tukey, John W. 1962. “The Future of Data Analysis.” *The Annals of Mathematical Statistics* 33 (1). JSTOR:1–67. http://projecteuclid.org/euclid.aoms/1177704711.