A Response to “Mere Innovation is Too Late”: Data Cowboys and Statistical Indians

Jim Ridgway

“Mere Innovation is Too Late” is an important paper calling for reflection and constructive discussions about the future of statistics education. George Cobb offers a metaphor from California real estate, namely that serviceable properties are often rebuilt by their owners to bring them up to date. However, I fear that he has mapped out the most positive scenario for the future of statistics. A “middle ground” metaphor is that of Indian tribes being moved from their reservations to less desirable and less fertile ground—here, statisticians being displaced by data scientists from their sacred territory; a darker metaphor is the fate of the Indian tribes in California in the Mission and post-Mission eras—condemned to servitude, and random acts of genocide. Data cowboys are unlikely to shoot statisticians—but then they don’t need to, because they seem to take over core territory, with rather little effort. Data science is seen to be sexy; it uses data that everyone actually generates (twitter streams, purchasing data, mobile phone locations), and creates applications that are really useful and part of everyday life in developed countries—fingerprint access, speech recognition, weather forecasts, shopping and vacation advice. So there is an existential crisis for statistics—if you can ride the data revolution, who needs a statistician? Well, some statisticians think you do.

“Most real life statistical problems have one or more nonstandard features. There are no routine statistical questions; only questionable statistical routines” (Cox, quoted in Chatfield, 1991, p. 240). “All models are wrong, but some are useful” (Box and Draper, 1987, p. 424).

Statistics has its origins in solving novel, practical problems. Both the Royal Statistical Society and the American Statistical Association were established by heterogeneous collections of individuals united by a common goal to tackle exciting problems by inventing methods and mathematics (see Pullinger, 2014). A problem for statistics education is that the curriculum devotes too much time to modeling well-understood problems with traditional (1920s) methods, and too little time modeling unfamiliar ones—thereby ignoring the raison d’etre of the discipline. Here, I consider introductory statistics courses.

Many introductory courses focus on one- and two-variable problems, work with small samples, and use made-up data. This runs real risks, pedagogically, namely reinforcing the common notion that every sample is representative of the population from which it is drawn, and that small samples are as representative as large ones (i.e. Tversky and Kahneman’s (1974) “representativeness” heuristic). Starting from one- and two-sample problems makes the leap to understanding multivariate data, and notions of interaction, rather difficult. Similarly, the emphasis on correlation and linear relationships makes the idea of nonlinear relationships hard. The range of applications of this approach is narrow (assuming additivity and linearity even in school science would be a big mistake). Some positive alternative approaches can be found in Ridgway (2015).

I agree with George Cobb that a focus on mathematical technique hides statistical ideas, and with his assertion that accessible ideas (e.g. Bayesian inference) are made obscure by dressing them up in heavy mathematics. Rakow et al. (2015) reported presenting funnel plots on outcomes from child heart surgery to 172 participants (a quarter of whom had no education beyond compulsory schooling) who (predominantly) were related to a child under the care of a specialist cardiac unit. The researchers asked questions which required an understanding of the funnel plot, and questions about which hospital or surgeon to choose. Around 90% of the responses were correct. They concluded that “. . . funnel plots can be readily understood. . . ” (p.327). Our own work supports the assertion that formal mathematics is not necessary for the understanding of important statistical ideas. We used Rasch scaling to show that computer-based problems involving three variables can be easier (in psychometric terms) than one and two variable problems presented on paper (Ridgway, McCusker and Nicholson, 2003). We have also shown that statistically naíve students aged 16 years can express sophisticated ideas such as interaction, nonlinear functions, and piecewise functions when asked to describe patterns in data presented in interactive displays (here, on alcohol use by young people, sexually transmitted diseases, and photosynthesis). Data visualisation makes it possible for students to explore large, authentic data sets and to reason about complex situations using these data. For example, we created the *constituency explorer* in collaboration with the House of Commons Library. This presents data on 150+ variables (demographics, health, voting patterns in two national elections etc.) relevant to every constituency in the UK in a visualization that runs on desktops and mobile devices. The primary target audience was politicians and their aides, but the resource has been used in the teaching of political science, social history, and geography, as well in school statistics classes. Is this approach strong on short term gratification and weak on long-term need? It *is* strong on short-term “wow”, but it can also bring core statistical ideas into introductory courses (data provenance, metadata descriptions, nonlinearity, discontinuities, and effect size, as well as means, medians and spread).

The range of phenomena that need to be modeled has expanded, and students know it. Linear additive models are not the only game in town. For example, analysis of Wikipedia use has been used to predict stock exchange movements (Moat et al. 2013). The key words associated with upward and downward movements are in the public domain. Should you use these

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 Ridgway, School of Education, Durham University, Leazes Road, Durham DH1 1TA, UK (Email: jim.ridgway@durham.ac.uk).

keywords when making investment decisions? Students have created imaginary traffic jams on Satnavs (see Bilton, 2104). These examples illustrate situations where agents can "game" important systems for their own purposes. Modeling systems is hard; modeling systems undergoing change is harder; now students are familiar with systems undergoing change that are "self aware". Should we be teaching students 1920s mathematics, or introducing big statistical ideas relevant to their own lives? Can we cede all algorithmic thinking to data science? A modest set of targets for statistics education is: use data visualisations of big open data sets to teach big statistical ideas; teach about the statistics that underpin the lived experiences of technology-savvy students (notably pattern recognition in its many applications); introduce modeling, early. And make friends with the data cowboys.

References

Bilton, N. (2014), "Friends, and Influence, for Sale Online," available online at *http://bits.blogs.nytimes.com/2014/04/20/friends-and-influence-for-sale-online/?emc=edit_ct_20140424&nl=technology&nlid=66226932*.

Box, G., and Draper, N. (1987), *Empirical Model-Building and Response Surfaces*, New York: Wiley.

Chatfield, C. (1991), "Avoiding Statistical Pitfalls," Statistcal Science, 6(3), 240–268.

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.

The Constituency Explorer, *http://www.constituencyexplorer.org.uk/*. Accessed 7 August 2015.

Moat, H.S., Curme, C., Avakian, A., Kenett, D.Y., Stanley, H.E., and Preis, T. (2013), "Quantifying Wikipedia Usage Patterns Before Stock Market Moves," *Nature Scientific Reports* (3), 1801. Doi:10.1038/srep01801.

Pullinger, J. (2014), "Statistics Making an Impact," *Jounal of the Royal Statistical Society*, Series A, 176 (4), 819–839.

Rakow, T., Wright, R.J., Spiegelhalter, D.J., and Bull, C. (2015), "The Pros and Cons of Funnel Plots as an Aid to Risk Communication and Patient Decision Making," *British Journal of Psychology*, 106, 327-348. DOI 10.1111/bjop.12081.

Ridgway, J. (2015), "Implications of the Data Revolution for Statistics Education," *International Statistical Review*, Doi:10.1111/insr.12110.

Ridgway, J., McCusker, S., and Nicholson, J. (2003), "Reasoning with Evidence—Development of a Scale," IAEA Conference proceedings. Available online at *https://www.dur.ac.uk/resources/smart.centre/Publications/ReasoningwithEvidence-Developmentofascale.pdf*.

Tversky, A., and Kahneman, D. (1972), "Judgement Under Uncertainty: Heuristics and Biases," *Science*, 185, 1124–1131. DOI:10.1126/science.185.4157.1124