Teaching Safe-Stats, Not Statistical Abstinence

Hadley WICKHAM

I thoroughly agree with George Cobb that it is time to rethink the statistics curriculum. We are at grave risk of becoming irrelevant, not because we are useless, but because we have focussed too much on doing the right thing, rather than doing something that works. As a field, statistics tends to lean towards abstinence based outreach: you should only do statistics if you're in a committed long-term relationship with a professional statistician. If you experiment on your own (or with friends), you will hurt yourself and others. Abstinence based approaches don't work because people will take the risk anyway, and most of the time they will safely enjoy themselves. Abstinence based statistics is particularly problematic as there are simply not enough professional statisticians to go around.

Rather than stigmatizing amateur statistics, we should be doing out best to provide tools to make it safer. We can't force people to use our tools (as much as we'd like to!), so those tools must be more fun and more empowering than the alternatives. The undergraduate statistics curriculum is a vital place to develop and promulgate these tools. As statistics grows ever more popular, we must be able to provide a curriculum that give students the skills they need to practice safe-stats for the rest of their lives.

To me, one of the keys to teaching safe-stats is to develop grammars of data analysis. A grammar is a framework that lays out the minimal set of independent components and a means of composing them to solve a wide range of problems within a domain. Much of my own work has been in this area: How can we provide an accessible grammar of graphics (Wilkinson 2005) that makes it easy to create graphics tailored to the problem at hand (Wickham 2009)? What are the verbs of data manipulation that allow us to solve 95% of problems (Horton, Baumer, and Wickham 2015)? What are the key components of tidy data, and how can we tidy messy data (Wickham 2014)?

The best grammars are both flexible and constraining. Once you've seen the grammar used to solve two problems, you should be able to recombine the pieces to solve a third, new, problem. But a grammar also needs to be constraining: unmitigated freedom is both overwhelming and potentially dangerous. Constraints can guide users towards better outcomes, but can not guarantee them. A system that prevents the user from doing the wrong thing must necessarily prevent many right things. (A personal example is the use of rather elegant ggplot2 graphics in the paper discussed by Singal 2015.)

To have that needed flexibility, a grammar must be embedded in a programming language. This offers an escape clause: each grammar need solve only the 90% of most common problems, leaving the long-tailed 10% to other parts of the language. This implies that statistics students must be taught to program.

Teaching programming even in the first statistics course is eminently achievable, as Baumer et al. (2014) and others have shown. The key is to focus on immediate pay-offs. You don't need to study the foundations of programming before you can start to program data analyses. Students can start with recipes, code snippets that students can use and adapt, to show immediately useful (and interesting) tools. Over the length of the course, students can grow from simple duplication to creative rearrangement of entire components.

We must give statistics students the skills to dive into the data ocean. Yes, there are sharks and jellyfish and rip tides, but we can not be paralyzed by all the potential dangers. Students will go swimming with or without us, and all we can do is prepare them as best we are able.

References

Baumer, B., Cetinkaya-Rundel, M., Bray, A., Loi, L., and Horton, N. J. (2014), "R Markdown: Integrating a Reproducible Analysis Tool into Introductory Statistics," *Technology Innovations in Statistics Education*, *https://escholarship.org/uc/item/90b2f5xh*.

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.

Horton, N. J., Baumer, B. S., and Wickham, H. (2015), "Setting the Stage for Data Science: Integration of Data Management Skills in Introductory and Second Courses in Statistics," *ArXiv E-Prints*, February, *http://chance.amstat.org/2015/04/setting-the-stage/*.

Singal, J. (2015), "The Case of the Amazing Gay-Marriage Data: How a Graduate Student Reluctantly Uncovered a Huge Scientific Fraud." *NyMag*, May, *http://nymag.com/scienceofus/2015/05/how-a-grad-student-uncovered-a-huge-fraud.html*.

Wickham, H. (2009), *ggplot2: Elegant Graphics for Data Analysis*, Berlin: Springer.

——— (2014), "Tidy Data," *The Journal of Statistical Software*, 59, *http://www.jstatsoft.org/v59/i10*.

Wilkinson, L. (2005), *The Grammar of Graphics* (2nd ed.), Berlin: Springer.

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. Hadley Wickham, RStudio 1719 Drew Houston Texas 77004 (Email: hadley@rstudio.com).