Interview With Nicholas J. Horton

Allan J. Rossman*, and Nicholas J. Horton?

*Department of Statistics, Cal Poly, San Luis Obispo, CA; ?Department of Mathematics and Statistics, Amherst College, Amherst, MA

Nicholas Horton is Beitzel Professor of Technology and Society and Professor of Statistics and Data Science at Amherst College. He is a Fellow of the American Statistical Association and the American Association for the Advancement of Science. He is also a recipient of ASA’s Founders Award. This interview took place via E-mail on January 11–August 6, 2018.

Beginnings

AR: Thanks very much, Nick, for agreeing to be interviewed for the Journal of Statistics Education. I’m anxious to hear your thoughts on statistics education in the present, but let’s start a few years in the past. Where were you when you were 18 years old, and what were your career plans at that point?

NH: I was a high school student in Albany, NY. I’d been playing around with computers and I was also interested in mathematics and actuarial science. I didn’t have specific career plans at that juncture.

AR: Where did you go to college, and what did you major in?

NH: I went to Harvard and I majored in psychology. In addition to the required courses for the major, I completed both of the advanced statistics courses that were required for PhD students in psychology, at a time when Tukey’s Exploratory Data Analysis was just taking root. It was exciting, as people started to think differently about statistics. It was also a little surreal, since I didn’t have depth in mathematics, but I learned a lot and became more interested in statistics and data analysis.

Early in college I got involved with computer support at the Science Center and started to work as a consultant on the Unix machines (it was a choice between that, VMS, and PRIMOS). By dumb luck that was the right environment to pick, as Mac OS X and Linux are essentially the same interface that I’m still working on some 36 years later.

After my sophomore year I took a year off from college and got a job running the computers for the Harvard Computer Science Department in what was then the Division of Applied Science. (Best job interview story: “you aren’t qualified for the position but if we don’t find anyone else in the next two weeks you’ll get the job.”)

I had my first brush with professional statistical software as a systems administrator when I was asked to install S on our Vax 11/780 (at that time one of the most powerful minicomputers). The installation process took several days of compilation in Fortran. I tested the install by generating a map of the United States using the “usa()” function. It’s really interesting to look back and see how those original ideas for S have influenced modern statistical software.

After I returned to college, I wrote a thesis on sex differences in attitudes toward computing, which focused on the required quantitative literacy exam that first-year students had to complete. It was a longitudinal study of attitudes, with the first questionnaire administered during orientation and the follow-up in February. Not surprisingly, there were missing values! I completed my analyses in SPSS-X, and the thesis and data ended up in the Murray Center archive at the Schlesinger Library.

I taught a discussion section several times for the introductory computer science course at Harvard: this was a wonderful experience that got me excited about teaching. The primary instructor took teaching seriously, and had us videotape one of our lectures and review that with the staff at the Teaching and Learning Center.

I also had the opportunity to interact with Eric Mazur, one of the main founders of peer instruction approaches and active learning techniques. These experiences planted a seed about a future in teaching.

AR: And then …? Did you go to graduate school in statistics immediately, or did you do something else first? When did you go back and fill in your math background?

NH: I took a lot of time to gear up for graduate school. After I graduated from college, I moved to Portland, Oregon to live with my soon-to-be-spouse, who was then a student at Reed
College. For a year or so I did freelance teaching for an outfit that provided week-long training programs in Unix, C, and related technologies for technology companies. This helped me further hone my teaching skills.

Next, I worked for five years as Computer Facilities Manager for the Department of Computer Science and Engineering at a small private graduate school (which was eventually subsumed into the Oregon Health Sciences University).

I enjoyed my day job but also devoted a lot of time to volunteer work with the American Friends Service Committee (a Quaker nonprofit organization) and our local Quaker meeting. I learned so much about how to facilitate consensus decision-making and the importance of process when working collaboratively.

My two wonderful daughters were born while we were in Portland. We were living on campus at Reed, where my spouse, who had founded Reed's community service program, was working in residence life.

My latent interest in statistics was apparently still strong, as I audited a course in "Statistics for Computer Science" that was taught by a visiting Vice President of one of the Bell Operating Companies. One day she asked me what my career plans and aspirations were. I didn't have a good answer. That evening I headed to a woodworking class at the Oregon School of Arts and Crafts, where a colleague of mine who had a similar IT job at a local ad agency was very unhappy. Why was she upset? One of her best employees had given notice and was heading to grad school. Her conclusion: her employees were either great but short-lived, as they went off to bigger or better things, or else unmotivated and less effective in their work. As I drove home, I wondered which group I belonged to. Over the next few weeks, the advice and wisdom of these two women hit home, and I started to think more about my plans for the future.

I'd always been interested in statistics and data analysis, and I had a strong computational background. But to prepare for graduate school, I needed to design a mathematics post-baccalaureate program to retake calculus (it had been a long time), linear algebra, and analysis. Luckily for me, the Biostatistics Department at the Harvard School of Public Health had need for computer support just when I applied. I was accepted into the program and provided an office, and I started to provide computer assistance for faculty and students while attending graduate school.

These years were so rich and formative for me that I can't help but encourage my students to take a few years before heading off to grad school. Seven years is probably an upper bound for how long to wait (my GRE scores from college had long expired), but it was a great chance for me to develop a clear sense of what I wanted to do.

AR: Thank goodness for that woodworking class that eventually led you to statistics! Did your interest in teaching continue to develop in graduate school?

NH: Indeed it did. Starting in my second year, I had the opportunity to teach the introductory statistical computing course for incoming doctoral students in biostatistics. The class was intended to provide a survey of key tools such as Splus, SAS, IMSL, LaTeX, and Unix that the students would need for their future research work. I also had the opportunity to serve as a teaching assistant for the probability, categorical data analysis, and longitudinal regression classes. My thesis was written using Splus, a precursor of R. I implemented the simulation studies for my project by grabbing spare cycles from workstations in the department.

A highlight near the end of my program was the opportunity to teach a two-week short-course on modern statistical methods (resampling, missing data, and semiparametric regression) as part of a collaborative program at the University of Athens, Greece.

All of these experiences helped develop my interest in teaching as some part of my future plans.

AR: Did you foresee a career in academia at that point? What was your next move after earning your degree?

NH: I was definitely planning some form of an academic career, but the plans weren't clearly developed. My next move was an unusual one: I took a postdoc position with my advisor (Nan Laird) to avoid having to move my family (and to get a raise). It was a wonderful opportunity to catch my breath, publish my dissertation, and start new projects. I was working on the development of methods for the analysis of multiple informant reports along with missing data methods with applications in psychiatric epidemiology. It was an amazing environment for a biostatistician.

Partway through the second year of my postdoc, I saw an ad for a faculty position at the Department of Biostatistics at Boston University (BU) and went ahead and applied. This was another great opportunity, which allowed me to again avoid moving my family, though I had to leave my postdoc early. Favorite supportive quote from Nan when describing having to give the rest of the money to fund my position back to the NIH: "Nick, it's only money." I'm enormously thankful and appreciative of the time, guidance, and financial support that she provided during my time at Harvard and to this day.

The Biostatistics Department at BU was growing, with active collaborations with clinical researchers at the BU School of Medicine and Boston Medical Center. Half of my salary was supported by projects at the Section of General Internal Medicine, focused on substance abuse and infectious disease epidemiology. I also worked as the main statistician for the Black Women's Health Study at the Slone Epidemiology Center, and also on a project to continue my work on multiple informants and missing data methods. Thirty percent of our salary supported our teaching two courses per year. I taught a longitudinal regression course and regularly taught EB723 (Intro to Stats with SAS) during the summer.

During my four years at BU I became active in the Boston Chapter of the American Statistical Association, which despite the name proudly serves members in Massachusetts, Vermont, New Hampshire, Maine, and Rhode Island. I have served on the planning committee since 2000 and was elected as the BCASA representative to the Council of Chapters Governing Board in 2002.

AR: One of my silly fascinations is with course numbers, so I have to ask: What does EB stand for in the course EB723? I assume
that 723 signifies a graduate course, but why not 501 or something like that for an "Intro to Stats" course?

When I was first at BU, we were an Epidemiology and Biostatistics department, hence the EB prefix. Later we split off and became the Department of Biostatistics, and I started teaching "BS"! Our largest graduate program was the MPH (master of public health) program. Before taking 723 students completed an intro to statistics (BS700/BS701/BS703) course with minimal use of technology.

Liberal Arts Colleges

AR: A less silly question: You eventually moved your family, I think, without moving out of Massachusetts. What led you to Smith College, a liberal arts college for women students?

NH: I'd been interested in liberal arts colleges ever since my father (an oncologist) attended a summer workshop at Colby College in Maine. Later we'd lived on campus at Reed for several years, so I also experienced the environment firsthand. In 2001, while helping my sister-in-law move her family from England to Washington, DC, we stopped by Swarthmore College to talk with Phil Everson about the position they were advertising. On the way back to Boston we spent the night in Northampton, MA, visiting Leanne Robertson (my spouse's college roommate) and seeing an excellent concert by Allison Krauss. Leanne was teaching at Smith at the time and told me that they were hiring. Eighteen (!) months later I received the job offer and we started to make our way the hundred or so miles westward.

AR: What was your biggest adjustment as you moved from BU to Smith? What courses did you teach when you started there? What was your teaching style at that point?

NH: Smith has a "teacher-scholar" model, with excellence expected in both scholarship and teaching. The biggest difference was the emphasis on teaching, with incredibly motivated and engaged students and a talented faculty willing to roll up their sleeves to innovate. Infusing high-impact practices into the curriculum was a major focus. That was really the biggest adjustment I had to make, as there wasn't the same emphasis on teaching at BU at that time.

Active learning became really important in my teaching. I always tried to get my students talking and interacting with each other within the first ten minutes of class as a way to get them to "reset" and engage with the material. I adopted Mosteller's "minute papers" (also known as "one-minute essays") to elicit feedback from students at the end of class. (I continue to use this technique, but now have the results submitted electronically through our course management system by midnight of the class meeting; I find that the resulting feedback is far more thoughtful than what was hurriedly scribbled on the small pieces of paper that I used to provide.)

Smith (along with Amherst College, where I now teach) is one of a small number of institutions with an open curriculum (no general education requirements beyond a first-year seminar). The faculty's goal is to ensure that students are lured into their courses as a way to develop their capacities. As a result students take classes because they are interested (as opposed to required).

To make such an unstructured approach work in practice, the faculty need to offer a variety of accessible and interesting courses that support a heterogeneous group of students while simultaneously challenging them. I taught a range of such classes along with my senior colleague Katherine Halvorsen. I'm particularly proud of the "Statistical Thinking" (SDS107) course that she created and that both of us taught, which was intended as a qualitative literacy course for our Ada Comstock Scholars (a wonderful cohort of nontraditional age students). As with almost all of our courses, the class was project-based and allowed students to pick topics that related to their interests.

We created a special introductory course for our engineering students with the goal of addressing their need for statistical methods that dovetailed with their mathematics and engineering courses. This was the first time that I started to introduce multiple regression early in the course as a purely descriptive approach. This strategy was suggested by a Modern Engineering Statistics text by Lapin (1990) that was published well before its time.

I taught linear algebra three times before our statistics enrollments shot through the roof and we had to hand that course back to the mathematicians.

AR: You mentioned that you have now moved on to Amherst College, continuing your previous tradition of professional moves involving little geographical distance. What motivated this move?

NH: I reserve the right to quibble about the "little geographical distance" part. I did move from the west coast to start grad school, and spent a most productive semester at the University of Auckland (many time zones away).

I moved to Amherst to help build their statistics program. When I first arrived in the Pioneer Valley the only statistician at Amherst was James Denton (he was Amherst's first statistician and first African-American faculty member). After his retirement, the College hired Amy Wagaman and Shu-Min Liao, who helped to build a program that grew so much that they needed more staffing. We have an active Five College Statistics Program (comprising faculty from Amherst, Hampshire, Mount Holyoke, Smith Colleges and the University of Massachusetts/Amherst), so I had interacted with them while I was at Smith.

I left Smith in 2013 and moved across the Connecticut River to begin my position at Amherst, where I now hold the Beitzel Professorship. We now have five tenure track faculty (out of a total of about 200 faculty at Amherst) in statistics, and our program is growing (with 21 graduates out of a class of 450 this year). We've started discussions with our colleagues across the college about a possible data science program.

Undergraduate Guidelines

AR: Rather than ask you about Amherst's program at this point, I'd like to ask about your work chairing the ASA committee that revised guidelines for undergraduate programs in statistical
sciences a few years ago. But before I ask about the guidelines themselves, let me ask about the process: What motivated the formation of the committee, how did you become involved, and how did the committee conduct its work?

NH: Let me take a step back. My career, research, and scholarship have had multiple strands to them. I’ve maintained an active research program in missing data methods with applications in biostatistics. I’ve served on a number of committees and panels for NIH and other organizations (including the Community Influences on Health Behaviors committee for the Center for Scientific Review and multiple data safety monitoring committees). I’ve also been active in statistics and data science education research. But perhaps the most rewarding part of my career has been my service as a volunteer in many capacities for the American Statistical Association and other professional associations. This involvement (that I heartily recommend to others looking to make an impact in our profession) is what led to my selection as the chair of the Undergraduate Statistics Guidelines Working Group that ASA President Nat Schenker formed in 2012.

I’ve already mentioned how I got started through my service to our local chapter. I was involved in several ASA sections and committees. In 2006 I was appointed to the ASA/National Council of Teachers of Mathematics Joint Committee on K-12 Probability, Statistics, and Data Analysis (serving as chair in 2011). This was an exciting time for the committee, as the advent of the Common Core State Standards meant that there was a new openness to statistics in middle and high schools. We set out to publish the previously endorsed GAISE (Guidelines for Assessment and Instruction in Statistics Education) K-12 report, which at that point was just a Microsoft Word file. This was (and is) a worker-bee committee (first chaired by Fred Mosteller), with many moving parts that has had a huge impact on K-12 education. I worked closely with Rebecca Nichols, the energetic and engaged director of ASA’s Education programs. I also served a term on the executive committee of the Section on Statistical Education.

After serving as the Boston Chapter representative to the Council of Chapters Governing Board, I was elected as the Vice-Chair for District 1 (responsible for New York State, New England, and our Canadian chapters). I was nominated for and elected as one of the ASA Board of Directors representatives from the Council of Chapters in 2012. These positions were time-consuming and demanding, but play an important role in building connections between statisticians and supporting the profession.

While I’d known Nat by reputation and by reading his papers, my first real interaction was when he served as a discussant for a talk I gave on my work at the Bureau of Labor Statistics at the Washington Statistical Society on adjusting for nonresponse in the Occupational Employment Statistics survey. Nat was great: he asked all sorts of probing questions that I couldn’t answer, which led to additional work and new findings that ended up being incorporated into the final paper (Horton, Toth, and Phipps 2014, https://projecteuclid.org/euclid.aas/1404229521).

When Nat was elected President of the ASA, I kept on pestering him with possible education initiatives. A few months later he pitched the proposal for the undergraduate guidelines group, and I jumped at the chance. Little did I know what lay ahead of me and the group!

Nat had assembled a dream team of educators from industry, government, and a range of academic institutions. Beth Chance, Steve Cohen, Scott Grimshaw, Jo Hardin, Tim Hesterberg, Roger Hoerl, Chris Malone, Rebecca Nichols, Deb Nolan, and I met once or twice a month for nearly two years. We met primarily via conference call, with an in-person meeting at the ASA office and other gatherings at conferences such as JSM, USCOTS, and ICOTS.

AR: How did you and the committee reach consensus on how to revise the previous undergraduate guidelines? Was there a fairly common agreement at the outset, or did the process require a lot of iteration and compromise? How much did you grapple with the tension between writing guidelines that were aspirational without making them unattainable?

NH: While we had a rough sense for our shared vision, considerable effort was needed to develop the final guidelines for undergraduate statistics programs. The process included a number of webinars on topics such as minors, ethics, observations from large programs, the role of community colleges, the relationship with data science, and preparation for graduate study. We organized open forums, circulated drafts, solicited input, and integrated feedback.

The undergraduate guidelines group spent much of our time finding the balance between being descriptive of what would make an ideal undergraduate program and what might be construed as being prescriptive. We reiterated the importance of following core skill areas: statistical methods and theory, data management, computation, mathematical foundations, and statistical practice. While these were roughly consistent with the prior guidelines from 2000, we were able to clearly delineate places (particularly those that involved computation and communication) where adjustments and improvements in the curriculum were (and still are) needed. The final guidelines were endorsed by the ASA Board in 2014 and can be found at www.amstat.org/asa/education/Curriculum-Guidelines-for-Undergraduate-Programs-in-Statistical-Science.aspx (ASA Undergraduate Guidelines Group 2014).

In parallel with the development of the guidelines, Jo Hardin and I guest-edited a special issue of *The American Statistician* (Horton and Hardin 2015) that is chock-full of great papers that challenge us as a community. I’m proud of the final result and pleased that many institutions have been working to align their curricula with these recommendations. In the U.S. we now have many more undergraduate statistics degree recipients and we’ve seen how successful they’ve been in the workforce or in later graduate study.

**Statistics at Amherst**

AR: Please tell us about Amherst’s undergraduate program in statistics.

NH: Our program is relatively new, with the first four graduates in 2015. We’ve grown considerably since then and have nearly 40 graduates after the first 4 years. We have five
tenure-track faculty (in order of seniority Amy Wagaman, Shu-Min Liao, myself, Brittnie Bailey, and Katherine Correia). The major begins with an introductory statistical modeling course (STAT135) that establishes a foundation in key concepts including variability, inference, modeling, and multivariate thinking. The course leverages a cloud-based version of R/RStudio and RMarkdown where students start their work on day one of the class.

The next course in the sequence is an intermediate course focused on regression and design (STAT230). This can be followed by a series of 200-level elective courses in applied statistics and data analysis. These include STAT231 (Data Science) and STAT240 (Multivariate Analysis). Both courses develop data skills, visualization skills, and communication as part of the process of using data to make insights.

In parallel with these applied courses students develop a foundation in mathematics from two courses in calculus and one in linear algebra and they deepen their understanding of abstraction and algorithmic thinking through a set of two computer science courses.

The theoretical underpinnings of the major are developed in a two-course series of Probability (STAT360) and Theoretical Statistics (STAT370). Both classes are now highly computational. Students learn how to write a function and undertake a simulation in STAT360 and practice empirical and analytical problem solving in groups in STAT370 (Horton et al. 2004). I’ve used STAT370 (and more recently STAT321) as an opportunity to introduce github as a mechanism to support collaborative workflows and to structure their group interactions.

Our capstone course (STAT495) provides an opportunity to integrate their knowledge of statistics, learn some new topics (often in predictive analytics and data science), and work to communicate results. Finally, to demonstrate that they’ve completed the learning outcomes for the major, students complete an independent comprehensive project on their own that features an exposition and analysis.

We’ve structured our program to hew closely to the ASA’s guidelines for undergraduate programs in statistics (ASA Undergraduate Guidelines Group 2014) and the National Academies “Undergraduate Data Science: Opportunities and Options” consensus report (Committee on Envisioning the Data Science Discipline: The Undergraduate Perspective 2018). I’m proud of what we’ve accomplished over a few short years.

AR: Please tell us more about STAT231, the Data Science course. What concepts and skills do students learn in that course? Is it taken by students other than Statistics majors? If so, have those students also taken STAT135 and STAT230 prior to STAT231?

NH: We’ve designed our introductory data science course (STAT231) to require some prior background in statistics and computer science. These prerequisites are flexible but having them allows us to build on those foundations and move further and faster than if the course had no prerequisites. About a third of the students are Statistics majors, with many others in Mathematics and/or Computer Science, and a scattering of other majors across the curriculum (e.g., Economics, Anthropology, Neuroscience, Psychology, Spanish). A handful of students have taken our Intermediate Statistics class (STAT230), since we’ve arranged our curriculum so that they can take either that course or STAT231 (or both at the same time) after completing intro stats.

The course provides a practical foundation to think with data by participating in the entire data analysis cycle. Students are provided multiple opportunities to generate statistical and data science questions and then address them through data acquisition, cleaning, transforming, modeling, and interpretation. We introduce modern tools for data management and wrangling (e.g., R/RStudio and the tidyverse) that are common in data science, build on prior work with graphics from intro stat to develop dynamic visualizations, and apply those tools to real-world applications, including text analytics, networks, and maps. We also include a module on ethics early in the course. Students undertake practical analyses of large, complex, and messy data sets leveraging these modern computing tools in a group project at the end of the class where they focus on communicating their insights and the process through a presentation, handout, and Shiny app.

As we reviewed our curriculum, we felt that it was important to develop these data-related skills that are so important for statistical practice earlier than in our capstone course (STAT495). We hope that it provides an accessible platform to introduce or build additional capacities in the following concepts/key skills that the National Academies “Data Science for Undergraduates: Opportunities and Options” report (2018) outlined in terms of what they called “Data Acumen”. I’ve marked topics from that report that are (I)ntroduced or (R)einforced in the class in Table 1.

I’m a firm believer in the importance of introducing topics once, reinforcing them later (perhaps in a different course), and helping students to master skills in a third or later iteration.

AR: Very interesting, thanks. That’s a lot to teach (and to learn!) in one course. You mentioned earlier that the statistics major at Amherst is relatively new and growing. This is a vague question, but what’s your sense for why students are attracted to the program, and what kinds of students are you attracting? Do you think most of them are students who would have majored in mathematics before the statistics major came into existence, or would most of them have been studying computer science, or neither of the above?

NH: There’s a growing interest in statistics and data science amongst our students. I suspect that students are attracted by the ability to dive into complex problems, to use modern tools, and then to communicate their results. Many of our students double major, most commonly with mathematics, computer science, and economics, but also with psychology, music, political science, sociology, and history. Alas, we don’t observe the counterfactual outcome of what students would have majored in if there wasn’t a statistics major, but we know many mathematics majors who are now interested in statistics and data science. That being said, the number of majors in mathematics as well as computer science at Amherst keeps on increasing, so I don’t know if we are taking away from those departments.
Table 1. Topics from Amherst College STAT 231 (Data Science) Course.

Mathematics
- Multivariate thinking via functions and graphical displays (R)
- Networks and graph theory (I)

Algorithmic thinking
- Basic abstractions (R)
- Algorithmic thinking (R)
- Programming concepts (R)

Statistical foundations
- Multivariate thinking (R)
- Exploratory data analysis (R)
- Statistical modeling and model assessment (R)

Data management and curation
- Data provenance (I)
- Data preparation, especially data cleansing and data transformation (I)
- Data management (of a variety of data types) (I)
- Text analytics and regular expressions (I)
- Modern databases (I)

Data description and visualization
- Data consistency checking (I)
- Exploratory data analysis (R)
- Grammar of graphics (I)
- Spatial data (I)
- Attractive and sound static and dynamic visualizations (I)
- Dashboards (I)

Data modeling
- Dimension reduction techniques (clustering) (I)

Workflow and reproducibility
- Workflows and workflow systems (I)
- Reproducible analysis (R)
- Documentation and code standards (I)
- Source code (version) control systems (github) (I)
- Collaboration (R)

Communication and teamwork
- Clear and comprehensive reporting (R)
- Well-structured technical writing without jargon (R)
- Effective presentation skills (R)

Ethics
- Ethical precepts for data science and codes of conduct (I)
- Privacy and confidentiality (I)
- Algorithmic bias (I)

NOTE: I indicates topics that are introduced, while R indicates topics that are reinforced.

Statistical Computing

AK: Statistical computing is an interest of yours, one manifestation of that being your work on Project MOSAIC. What were/are the goals of that project, and what resources have been made available for statistics teachers?

NH: Project MOSAIC was an NSF-funded initiative led by Principal Investigator Danny Kaplan (Macalester College). My fellow co-PIs/co-conspirators Randy Pruim (Calvin College) and Eric Marland (Appalachian State) worked to establish a community of educators working to tie together aspects of quantitative skills that students in science, technology, engineering, and mathematics will need in their professional lives, but which are usually taught in isolation, if at all.

How can we as educators bring new ideas (e.g., confounding and multiple regression) into our courses? The revised GAISE College report (ASA GAISE College Group 2016) enunciated the need for a broadening of our introductory courses beyond bivariate comparisons. One important part of what we wanted to support was the use of modern technology to bring multivariate thinking into the classroom. We’ve encouraged instructors to update their courses by adding multivariate thinking early on in a course to describe relationships using multidimensional graphics and multiple regression models.

A big part of our work was the development of the mosaic package in R, which augmented the existing capabilities of R to facilitate the use of a “Less Volume, More Creativity” approach to descriptive statistics, visualizations, and modeling (Horton et al. 2015; Pruim, Kaplan, and Horton 2017). This builds on a simple but powerful modeling framework. As an example, to calculate summary statistics for an outcome variable Y as a function of a grouping variable X in a dataframe called ds, the commands are:

```r
library(mosaic)
mean(Y ~ X, data = ds)
favstats(Y ~ X, data = ds) # more
    statistics than just the mean
```

Side by side boxplots can be generated using the same framework:

```r
gf_boxplot(Y ~ X, data = ds)
```

An ANOVA can be fit in a similar manner:

```r
lm(Y ~ X, data = ds)
```

Figure 1 displays the results of an analysis of depressive symptoms (CESD score) as a function of substance use groups from the HELP (Health Evaluation and Linkage to Primary Care) study (Samet et al. 2003) available from the mosaicData package in R.

The modeling language can be extended to more than two variables or simplified down to one, though as Chris Wild at the University of Auckland has noted: why not start with a bivariate comparison since it’s more interesting than univariate analyses?

What’s attractive about the mosaic approach is that students can learn about a new kind of object for modeling called a “formula” (e.g., Y ~ X) and then use these formulas as a basis for a modeling language to answer interesting multivariate questions (ASA GAISE College Group 2016; Wang, Rush, and Horton 2017) by adding additional variables to the formula (e.g., Y ~ X1 + X2).

Even though the grant funding has ended, this work continues. To provide a cleaner interface for creating data visualizations, Randy and Danny wrote the ggformula package, which replaces the prior dependency on lattice graphics. The ggformula system utilizes ggplot2 (an elegant but more complicated graphics environment) and extends the existing modeling language in mosaic.
Figure 1. Example analyses using the mosaic modeling language. Here depressive symptoms (CESD score) are modeled as a function of substance use groups.

Many examples and resources for the mosaic package are available as vignettes on the CRAN (Comprehensive R Archive Network) site. These include a minimal R handout (all functions needed to teach intro stat on one side of a piece of paper with examples on the other side), the “Less Volume, More Creativity” sermon that Randy created, and annotated examples from a number of popular textbooks.

AR: I take it that you are a proponent of using R in “Stat 101” courses aimed at students in various disciplines including social sciences and life sciences. Am I right about that? If so, please make the case, and respond to those who think that R involves too much of a learning curve for students in those courses.

NH: I am a proponent of introducing students to real tools that they can use during and after “Stat 101” courses. I believe that this is consistent with the recommendations of the revised GAISE college report and makes sense (particularly in the United States) given the increased role that statistics play in 9th and 11th grades in the Common Core State Standards for Mathematics. Most students now have increased exposure to descriptive statistics, informal inference, and formal inference (for univariate and bivariate settings). Many students may need a developmental statistics course that doesn’t use much technology. But most would be well served by a course that lets them practice the entire statistical analysis cycle (data input, data wrangling, exploration, modeling, and interpretation).

Many high schools and colleges introduce R/Studio and RMarkdown on day one of their introductory statistics courses (Baumer et al. 2014; Çetinkaya-Rundel and Rundel 2018; Wang, Rush, and Horton 2017). Incorporating R requires some time in the course, but the tools and interface have become considerably simplified, particularly when using cloud-based implementations where students just “bring a browser” (Çetinkaya-Rundel and Rundel 2018). At Amherst we’ve also hired a set of Statistics and Data Science Fellows who provide support with data wrangling and software, peer tutoring, and assistance with the material in our introductory courses.

And let me be clear: I don’t think that R/Studio is the only environment that might be suitable. I’ve been favorably impressed with the data science library in Python from the data8 folks that provides a simplified interface suitable to new users. The data science library shares much of the “Less Volume, More Creativity” philosophy of the mosaic package (Pruim, Kaplan, and Horton 2017). The new learning outcomes related to computation are kept to a minimum with students being able to practice use of a powerful and elegant modeling language.

This is an exciting time for statistics and data science. New courses such as the Advanced Placement Computer Science Principles course bring “Data as Information” into the fore. I’ve been impressed by how much of our traditional curriculum is included in this fast-growing high school course. At times I worry that this explicitly computer science course is a better option for exploring data than the AP statistics class.

I want to ensure that students taking our introductory courses also see the beauty and power of statistics. They need to be using technology such as R/Studio to allow them to explore multivariate datasets and extract meaning.

AR: You’ve also written some textbooks on statistical computing. Could you provide an overview of them—what’s similar and what’s different, and what are the intended student audiences?

NH: Ken Kleinman and I wrote SAS and R: Data Management, Statistical Analysis, and Graphics, now in its second edition (Kleinman and Horton 2014), as a way to help analysts move back and forth between these two powerful (but not always easy to use nor well-documented) environments. We intended the book to be like a French-English dictionary where the code to undertake a particular task (e.g., fitting a generalized estimating equation or calculating a singular value decomposition of a matrix) was provided in parallel sections with illustrated examples taken from the HELP (Health Evaluation and Linkage to Primary Care) clinical trial (Samet et al. 2003). The book has been really useful to me: I often use it to remind myself of particular syntax or specific idioms. I know of many courses in statistical computing that have used this as a recommended text. We spun off two books from this base: one on R (Horton and Kleinman 2015; also now in its second edition) and one on SAS that were intended for analysts focusing on one package or the other.

My coauthors Ben Baumer and Danny Kaplan wrote Modern Data Science with R (Baumer, Kaplan, and Horton 2017) as a textbook for an introductory one-semester course in data science. The goal was to help students solve problems (or as
Diane Lambert of Google says “to think with data”) using a variety of data types and technologies.

The book begins with an introduction to data visualization, dives into data wrangling, reviews key concepts in data ethics, then transitions to a set of chapters on methods (statistical foundations, supervised machine learning, and unsupervised learning), databases (SQL) then different data types (including text, maps, and graphs). We include appendices on R, how to write a function, reproducible analysis tools, and database administration. I’ve used the book for the STAT231 course that I discussed earlier, while Ben has been teaching a version of the class with no prerequisites.

K-12 Statistics Education

AR: Let’s shift gears a bit to discuss your interests and contributions to K-12 statistics education. How did your interest in this develop?

NH: Historically in the United States statistics hasn’t played a major role in the K-12 curriculum. That changed with a number of mathematics reform initiatives in the 1990’s, including the National Council of Teachers of Mathematics curricular standards in 1989 and the development of the Advanced Placement Statistics course in 1996. Since that time there have been many efforts to identify ways to develop statistical thinking in K-12 (well enunciated in the ASA GAISE K-12 report) and to build connections with science education along with the mathematics curriculum.

I joined the ASA/NCTM Joint Committee to help make connections between post-secondary statistics education and efforts at the K-12 level, and also to support initiatives such as the ASA Poster and Project Competitions, a variety of publications, and efforts such as the Beyond AP Statistics and Meeting within a Meeting.

The challenges in terms of teacher training and faculty development in K-12 are enormous, but the payoff in terms of a quantitatively literate citizenry is huge. There’s still a great amount of work to be done but efforts of the joint committee and its many partners have led to dramatic changes in the curriculum. Statistics now has a larger role in the high school curriculum in the Common Core State Standards and is an important part of the high stakes assessments that, for worse or for better, are integrated into our school systems.

How will these changes affect those of us teaching in colleges and universities? We’re already seeing the impact of a large number of students now completing the AP Statistics and more recently the AP Computer Science Principles course before they come to college.

Potentially more important is the extended exposure that many students are now seeing in their mainstream mathematics courses. Widely adopted curricula such as Eureka Math now devote five weeks of the ninth grade algebra I course to descriptive statistics and informal inference with eight weeks of the eleventh grade algebra II course building on this framework with basic probability and formal inference for two sample comparisons (through resampling).

The SAT Math test includes sample questions about residuals from a linear regression model, interpretation of slopes and intercepts from a model, sample size and its relationship to margin of error, design and possible conclusions from a study, and conditional probabilities calculated from cross-classification tables.

It’s encouraging that many high school students are now developing a background in statistical thinking. These curricular changes will allow us to incorporate changes to our introductory courses to include more multivariate thinking and modeling, among other important topics.

AR: Am I right that you also spent a leave in New Zealand, in part to work on K-12 education initiatives? What did you work on there, and what did you learn about how the role of statistics in K-12 education compares between NZ and US?

NH: I spent eight months in New Zealand during the 2007–2008 academic year. The University of Auckland has the largest statistics department in the southern hemisphere with strengths in many areas. In addition to teaching a course on mixed models at the University of Auckland and writing the SAS and R book, I began to work with Chris Wild, Maxine Pfannkuch, and Matt Regan on the role of informal inference in the K-12 curriculum. It was an amazing experience.

As a small country with a single educational authority, New Zealand has a remarkably nimble process for curricular changes. Chris and his colleagues were deeply embedded the process of completely overhauling the national mathematics curriculum with the goal of infusing it early and often with statistics. Much of the New Zealand work has been picked up in terms of the statistics strands of the Common Core State Standards.

Our paper (Wild et al. 2011) detailed a progression of rules on “how to make a call” to determine differences between two groups using a series of successively more sophisticated approaches that eventually leads to formal inference. Chris presented this as a read paper at the Royal Statistical Society in October 2010 (on World Statistics Day, no less) and the oral and written comments were thought provoking and stimulating.

AR: Back to your comment about Advanced Placement courses: I’ve heard you deliver mild criticisms of AP Statistics for not changing much over its now 20+ year tenure, and I know that you are quite enthusiastic about AP Computer Science Principles for including many modern ideas of data science. Please expand on these views.

NH: The AP Statistics course represented a modern and forward-facing curriculum when it was unveiled more than two decades ago. But unfortunately, the intro stats course and the discipline as a whole have changed considerably since then. The changes in the Common Core State Standards haven’t been incorporated, so much of the material in the AP Statistics course is somewhat redundant.

I’ve mentioned the importance that the revised GAISE College report places on multivariate thinking. This key topic is underrepresented in the AP Statistics curriculum, which has a decided focus on bivariate comparisons (where students need to assume that other factors that might confound a relationship do not exist). Another area where the course has lagged is the use of technology: graphing calculators are not the tools to use to extract meaning from data. Finally, in my opinion, there’s too much emphasis on inference.
The AP Computer Science Principles course is designed to help students "develop computational thinking skills vital for success across all disciplines, such as using computational tools to analyze and study data and working with large data sets to analyze, visualize, and draw conclusions from trends". The "Data as Information" unit of the course includes a number of learning outcomes that are familiar to us:

LO 3.1.1: Find patterns and test hypotheses about digitally processed information to gain insight and knowledge.
LO 3.1.3: Explain the insight and knowledge gained from digitally processed data by using appropriate visualizations, notations, and precise language.
LO 3.2.1: Extract information from data to discover and explain connections or trends.

The CS Principles course has students submit two "computational artifacts" (think of them as projects) to complement an end-of-year exam. There's a lot there that I wish was reflected in the AP Statistics curriculum.

Data Science

AR: We've talked a bit about data science, but I'd like to return to that topic and take a wide view. You've worked on several initiatives at the National Academies about the enterprise of data science, particularly from an undergraduate perspective. Before I ask about implications at the undergraduate level, let me ask how you see the relationship between data science and statistics.

NH: My definitions of data science and statistics are quite similar: both involve extracting meaning and insights from data. Principled statistical analysis requires domain knowledge, a firm grasp on data issues and data technologies, the ability to select and apply an appropriate statistical method, the experience to be able to assess and interpret results, and facility with a computational workflow that allows transparency and reproducibility. I would argue that statisticians who deal with all of these aspects are indistinguishable from data scientists.

Historically, some statisticians worked in environments where back-office staff handled the vast majority of data issues. Analysts, particularly those hired after an undergraduate degree, now need to have much stronger data skills (what the NAS report called "data acumen"). I liken this shift to the changing role of travel agents, who now primarily exist for complicated itineraries while most of us book our own flights.

AR: Since you describe statistics and data science as very similar, I can't helping wondering (and asking): What is it about "data science" that has led to such an explosion of interest, well beyond what "statistics" has been able to generate? Is there more to it than "data science" being a much better marketing term?

NH: That's a great question. And I don't think that I know the answer. I do know that many people had a bad experience with their prior statistics courses and that they don't hesitate to let me know that when I'm introduced as a statistician. One of my worries is that statistics will be relegated to only what one needs when the sample size is small and a \( p \)-value is needed to test a hypothesis that two groups have the same population mean.

For whatever reason, data science is a far more compelling descriptor that captures the excitement and power of statistics.

AR: Two-year colleges are beginning to develop data science programs (another example where "data science" is making progress that "statistics" has not). Have you been involved with this? What can you tell us about this?

NH: The American Statistical Association recently organized the NSF-funded “Two-Year College Data Science Summit” in Arlington, VA. Rob Gould and Roxy Peck chaired the steering committee which helped to bring together more than 60 participants to discuss data science programs at two-year colleges to identify learning outcomes and resources, and to elicit best practices for associate programs, transfer programs, and certificates. I participated as a member of the writing team for the NSF-funded Broadening Data Science Education workshop (see the report at http://bit.ly/KeepingDataScienceBroad_Report), which noted the key role that community colleges play in higher education in this country (accounting for nearly 40% of undergraduate students).

For the past year I've been working as a Mathematics Pathways Fellow (along with Roxy Peck) at the Dana Center at the University of Texas at Austin. The Dana Center, led by Uri Treisman, works on ways to improve K-16 mathematics and science education with an emphasis on strategies to improve engagement and achievement. Roxy and I have been collaborating with Rebecca Hartzler to identify pathways for data science education at two-year colleges that help to ensure that students have an appropriate foundation in computational, mathematical, and statistical thinking. There's lots of work to be done in determining appropriate learning outcomes, crafting a variety of courses that introduce key topics, reinforcing them in subsequent courses, and providing repeated opportunities for students to master these capacities. We need new materials, curricular frameworks, and above all, faculty development. But as you note, there's an openness to data science at the two-year colleges that was never evident for statistics.

AR: Do you think the interest is more in developing associate degree programs in data science, or in preparing students to transfer to where they can pursue a bachelor's degree?

NH: Another great question. I think that it's too early to tell. Both pathways are important, for different reasons.

Additional work is needed to establish criteria for associates programs leading directly to the workforce. Models for such degree programs include information technology and nursing. I suspect that well-trained associates graduates in data science will find jobs as data technicians if they can develop an appropriate background in the necessary areas (see NAS 2018), depth in data acumen, and repeated practice with the entire data science cycle.

Many students will decide to continue their coursework and complete a bachelor's degree after transferring to a four-year school. Developing programs here is tricky as there's not yet consensus on what a four-year degree in data science should include. Some programs might require more depth in mathematics (e.g., bachelors programs in engineering schools). Others might specify more coursework in other areas. It's important
that two-year colleges prepare students to seamlessly transfer with clear and transparent articulation agreements.

AR: We talked earlier about ASAs guidelines for undergraduate programs in statistical science. Another group developed curriculum guidelines for undergraduate programs in data science. What do you think of those guidelines and how they relate to ASAs guidelines for statistics programs?

NH: The curriculum guidelines for data science that Dick De Veaux and colleagues developed during the summer of 2016 were really helpful in clarifying the differences between data science and statistics (De Veaux et al. 2017). They share much of the general framework as the ASA statistics guidelines, but broadened them in important ways. Most notably, they identified key learning outcomes for computational thinking that go beyond what we had suggested for undergraduate statistics programs. Some examples include increased depth in abstraction, variable scoping, data structures, algorithmic complexity, high-performance computing, computing security, and data technologies. The data science guidelines were referenced in the National Academies “Undergraduate Data Science: Opportunities and Options” consensus report.

Professional Associations

AR: You have served on the ASA Board of Directors, on the National Academies Committee on Applied and Theoretical Statistics (now there’s a good acronym), and as chair of the Committee of Presidents of Statistical Societies (COPSS). How have these experiences informed your outlook on our discipline?

NH: All three of those appointments have served to broaden my exposure to the discipline. The ASA promotes the practice and profession of statistics (and increasingly data science). CATS was founded 40 years ago to raise the visibility and improve the practice of statistics within government agencies not well connected to statistics (see for example, CATS 1994). In recent years there has been increasing attention to statistical issues of big data and data science. COPSS works to improve inter society communication between the ASA, IMS, SSC, ENAR, and WNAR as well as broader conversations with other affiliated statistical associations. All of these organizations have identified education as an increasingly important part of their work. All three have been asking questions about how we can help ensure that the next generation of citizens and statisticians/data scientists have the necessary skills and capacities to analyze the data in their lives. Much of my work has been to build connections between teaching and the broader worlds of statistical practice and statistical research.

AR: In terms of those connections, what do you see as the biggest misconception that statistics practitioners and researchers have about statistics education? Conversely, what's the biggest misconception that statistics educators have about statistical practice and research? Or if you don't like the misconception focus, what message do you most want each group to hear about the other?

NH: Statistics practitioners and researchers sometimes think of themselves as experts in statistics education, since they mastered the discipline in their prior studies. It’s important for those involved with education (e.g., as a dean, department chair, director of undergraduate studies, or curriculum committee member) as well as those teaching to engage with the statistics education community to ensure that they aren’t just doing what has been done in the past out of a misplaced sense of tradition.

Conversely, statistics educators need to ensure that they are current in the practice of statistics, which as we’ve noted in this era of data science has gotten far broader. The ASA/MAA Guidelines for Teaching Statistics talk about the importance of multiple exposures to the entire statistical analysis cycle. We need to make this type of experience a component of professional development programs.

AR: Here’s a devil’s advocate question about professional associations. I believe that ASA is doing quite well (please correct me if I'm wrong). But I’ve heard some statisticians express concern that considering how well our profession is doing in terms of numbers earning degrees and employment opportunities and publicity in general, it’s a shame that ASA has not doubled or tripled its membership in the past decade. How do you respond to this concern? Do you think professional associations need to change in order to appeal to people entering the profession in the 21st century?

NH: I can’t speak for the ASA, though I do understand that they are doing reasonably well financially, amidst changes in how conferences, publications, and membership fees are used to support associations. You are correct that the ASA membership is growing but not as fast as the discipline. Some of this is societal: memberships in many organizations are shrinking or not keeping pace with population growth.

Associations provide important connections, foster synergistic activities, and build networks. Sustained over time it is possible to have a big impact on our profession. I know that I’ve benefited greatly as a statistician due to the efforts of many associations, including the ASA, IMS, CAUSE, AAAS, and the ISI. I believe that it’s important for statisticians to join these associations, contribute their time and expertise, and encourage their colleagues and students to do the same.

AR: ASA is sponsoring StatFest as one effort to promote the profession to undergraduate students from historically underrepresented groups. You’ll be hosting StatFest 2018 at Amherst soon. Please tell us about this effort and your involvement.

NH: StatFest began in 2001 due to the tireless efforts of Professor Ngambal Shah (Spelman College). Since then it has flourished as an annual conference to showcase careers and graduate studies in statistics and data science. Now organized by the ASA Committee on Minorities in Statistics, StatFest is highly interactive, with presentations from established professionals, academic leaders, and current graduate students, opportunities for networking, panel forums, and a poster session.

I agreed to help run the local organizing committee for this year’s gathering because I believe that it’s critically important to ensure that the statistics profession is more inclusive. This commitment to diversifying the profession is at the heart of the mission of Amherst College and the ASA.


Pop Quiz

AR: Let’s turn to the “pop quiz” portion of this interview, where I’ll ask a series of short questions and will encourage you to keep responses brief as well. Please tell us about your family.

NH: I have two adult children that I’m really proud of: one lives in the Twin Cities, MN and the other in Western Massachusetts. I’ve been married to my spouse Julia since 1990. I have a sister in England and a half-brother in Atlanta. My parents, their partners, and my mother-in-law are still alive and thriving.

AR: What are some of your hobbies?

NH: Julia and I helped to found the Friends of Northampton Trails and Greenways (www.fntg.net), a nonprofit that supports the rail trail network in and around our community. It’s been really rewarding to see the trail networks connected and extended. Planning is ongoing to make connections from Northampton to Boston, MA and New Haven, CT. I live and work near one trail and regularly explore it on bike as well as make my way to work or errands.

AR: Speaking of travel (how’s that for a seamless segue?), name a favorite travel spot that you’ve visited for work and one that was purely a pleasure trip.

NH: My absolute favorite place to visit has been New Zealand. We toured the Central Otago Rail Trail (how’s that for a shameless continuation of the previous answer?) and explored Dunedin and the Otago Peninsula a few years back. It was magical. Closer to home, Vermont and Maine are places that I love.

AR: What are some recent (nonstatistics) books that you’ve read or movies that you’ve seen?

NH: I tend to read nonfiction in my spare time. Some books that I’ve enjoyed recently include “Space Odyssey: Stanley Kubrick, Arthur C. Clarke, and the Making of a Masterpiece” by Michael Benson, “A First-Class Catastrophe: The Road to Black Monday, the Worst Day in Wall Street History” by Diana Henriques, and “Dreaming the Beatles” by Rob Sheffield. But I also like a well-spun novel, like “Seven Days of Us” by Francesca Hornax. An all-time favorite of mine is Scarlett Thomas’s novel “Popco”: it’s a lovely story of an ideator navigating a complicated world.

AR: Please tell us something about yourself that is likely to surprise most JSE readers.

NH: I was a child model (surprise, surprise) and a member of the United Food and Commercial Workers Union (not related to the modeling).

AR: Well done, I am indeed surprised by both. And that’s great that you have alternate career paths in place, just in case data science and statistics and education do not work out for you. My next two questions are fanciful ones: First let’s suppose that the JSE editor offers to treat you to dinner with three guests of your choice, at a site of your choosing anywhere in the world. Where would you go, and who would you invite to join you?

NH: I would invite M.F.K. Fisher, Aimee Mann, and Barack Obama to a meal at the Sylvia Beach Hotel in Newport, Oregon. I would look forward to an evening full of discussions of food, music, and hope.

AR: The second odd question is about time travel. If you could travel to any point in the past or future to observe for a day, which would you pick, and why?

NH: I’d head into the near future, to see if we were able to address the challenges of global climate change (and if we weren’t, to come back with even more of an imperative for what’s needed for us to avoid the worst of it).

Conclusions

AR: That concludes the pop quiz. I have a few concluding questions for you, but first let me ask if there’s anything that you’d like me to ask that I haven’t yet.

NH: The mentoring that I’ve received has been fundamental in my career. In recent years I’ve been working to help pay things back.

I helped to establish the Stat Ed Section mentoring program two years ago that paired up seasoned mentors with more junior mentees. It’s been very rewarding to see how a relatively simple pilot program has now grown since Matt Hayat took over as the chair.

I’ve also worked as a biostatistics faculty member for the Advanced Research Institute (ARI) that matches new investigators with a focus on geriatric mental health with mentors to help them make a successful transition to an independent research career (Bruce et al. 2011). I’ve learned so much from this intensive boot-camp about ways to guide junior faculty in terms of their research agendas, grant-preparation, and time- and career-management. I hope someday to be able to organize something similar in the statistics education community.

AR: It’s too easy to ask about what has changed in your teaching over the years, so instead let me ask: What has not changed in your teaching throughout your career?

NH: One thing that has not changed is my commitment to help students develop the capacities that they need to solve complex problems out in the world. The tools and the technologies we use change all the time, but the commitment to problem-solving remains constant.

AR: What is your greatest hope, and what is your most troubling concern, about statistics education in the next decade?

My greatest hope is that statistics education can fully embrace data science and broader learning outcomes. My most troubling concern is that we will double down on our focus on inference and be tagged the “p-value police” and labeled as the discipline to turn to when one wants to compare two groups when the sample size is small.
AR: Among all of your professional accomplishments, can you name one in which you take the most pride?

I’m proudest of the work that I’ve done as part of the National Academies and the American Statistical Association to help improve statistics and data science education at the undergraduate level.

AR: Thanks again for taking the time to answer all of my questions. My final one is: What advice to you have for those new to statistics education?

NH: My advice is to become active in the profession, say yes when editors, government agencies, or colleagues ask for reviews and groups look for volunteers, stay focused on your research and scholarship, and learn how to say “no” at the right times for the right reasons.

References

ASA GAISE College Group (2016), “Guidelines for Assessment and Instruction in Statistics Education revised College Report,” American Statistical Association, available at www.amstat.org/asa/files/pdfs/ GAISE/GaiseCollege_Full.pdf.

ASA Undergraduate Guidelines Group (2014), “Curriculum Guidelines for Undergraduate Programs in Statistical Science,” American Statistical Association, available at www.amstat.org/asa/education/Curriculum-Guidelinesfor-Undergraduate-Programs-in-StatisticalScience.aspx.

Baumer, B., Çetinkaya-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, 8, available at http://escholarship.org/uc/item/90b2f5xh.

Baumer, B., Kaplan, D., and Horton, N. J. (2017), Modern Data Science With R, Boca Raton, FL: CRC Press, available at http://msdr-book.github.io.

Bruce, M., Bartles, S., Lyness, J., Sirey, J., Sheline, Y., and Smith, G. (2011), “Promoting the Transition to Independent Scientist: A National Career Development Program,” Academic Medicine, 86, 1179–1184.

Çetinkaya-Rundel, M., and Rundel, C. (2018), “Infrastructure and Tools for Teaching Computing Throughout the Statistical Curriculum,” The American Statistician, 72, 58–65.

Committee on Applied and Theoretical Statistics (CATS), National Research Council (1994), Modern Interdisciplinary University Statistics Education: Proceedings of a Symposium, available at http://www.nap.edu/catalog/2355.html.

Committee on Envisioning the Data Science Discipline: The Undergraduate Perspective (2018), Undergraduate Data Science: Opportunities and Options, Washington, DC: National Academies Press, available at https://www.nap.edu/catalog/25104/data-science-for-undergraduates-opportunities-and-options.

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, I. Z., Francis, A., Gould, R., and Kim, A. Y. (2017), “Curriculum Guidelines for Undergraduate Programs in Data Science,” Annual Review of Statistics and Its Application, 4, 15–30.

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,” CHANCE, 28, 40–50, available at http://chance.amstat.org/2015/04/setting-the-stage.

Horton, N. J., and Hardin, J. (2015), “Teaching the Next Generation of Statistics Students to ‘Think With Data’: Special Issue on Statistics and the Undergraduate Curriculum,” The American Statistician, 69, 259–265, available at https://amstat.tandfonline.com/doi/full/10.1080/00031305.2015.1094283.

Horton, N. J., and Kleinman, K. (2015), Using R and RStudio for Data Management, Statistical Analysis, and Graphics (2nd ed.), Boca Raton, FL: CRC Press.

Horton, N. J., Brown, E. R., and Qian, L. (2004), “Use of R as a Toolbox for Mathematical Statistics Exploration,” The American Statistician, 58, 343–357.

Horton, N. J., Toth, D., and Phipps, P. (2014), “Adjusting Models of Ordered Multinomial Outcomes for Nonignorable Nonresponse in the Occupational Employment Statistics Survey,” The Annals of Applied Statistics, 8, 956–973, available at https://projecteuclid.org/euclid.aoas/1404229521.

Kleinman, K., and Horton, N. J. (2014), SAS and R: Data Management, Statistical Analysis, and Graphics (2nd ed.), Boca Raton, FL: CRC Press.

Lapin, L. L. (1990), Modern Engineering Statistics, Boston, MA: PWS-Kent Pub. Co.

Pruiim, R., Kaplan, D., and Horton, N. J. (2017), “The Mosaic Package: Helping Students to ‘Think With Data’ Using R,” R Journal, 9, 77–102, available at https://journal.r-project.org/archive/2017/RJ-2017-024/ RJ-2017-024.pdf.

Samet, J., Larson, M., Horton, N. J., Doyle, K., Winter, M., and Saitz, R. (2003), “Linking Alcohol and Drug-Dependent Adults to Primary Medical Care: A Randomized Controlled Trial of a Multi-disciplinary Health Intervention in a Detoxification Unit,” Addiction, 98, 509–516.

Wang, X., Rush, C., and Horton, N. J. (2017), “Data Visualization on Day One: Bringing Big Ideas Into Intro Stats Early and Often,” Technology Innovations in Statistics Education, 10, available at http://escholarship.org/uc/item/84v3774z.

Wild, C. J., Pfannkuch, M., Regan, M., and Horton, N. J. (2011), “Towards More Accessible Conceptions of Statistical Inference,” Journal of the Royal Statistical Society, Series A, 174, 247–295.