CHALLENGES AND OPPORTUNITIES FOR STATISTICS AND DATA SCIENCE UNDERGRADUATE MAJOR AND MINOR DEGREE PROGRAMS

Nicholas J. Horton[1] and Johanna S. Hardin[2]
[1]Amherst College, United States
[2]Pomona College, United States
jo.hardin@pomona.edu

Recent curricular working groups (e.g., the ASA Undergraduate Guidelines for Statistics Programs and the Park City Math Institute Data Science Guidelines) have provided useful guidance for undergraduate programs in statistics and data science. We review these guidelines, compare and contrast them, and explore successful implementation strategies and problematic hurdles for the teaching of statistics and data science at the post-secondary level. What key emphases from data science (e.g., the increased role of computation and communication) need to be further infused in statistics programs? How can these topics be integrated to ensure that students emerge with the capacity to "think with data"? What are some of the other issues that we need to address to ensure that statistics is interwoven into our data science programs?

INTRODUCTION

This is an exciting time to be a statistician. The flood of data available to address important societal questions and the advent of "data science" have made the study of statistics more relevant than ever. Improved software tools (e.g., R and the tidyverse, https://www.tidyverse.org, Wickham and Grolemund, 2017; McNamara and Horton, in press; Perez and Granger, 2007) and increasingly sophisticated methods (e.g., statistical and machine learning) can easily be applied to rich datasets and databases (Horton, Baumer, and Wickham, 2015; Baumer, 2018a; Baumer, 2018b) using principled and reproducible workflows (Baumer et al, 2014; Bryan, 2018).

The National Academies Committee on Applied and Theoretical Statistics (CATS) opened a report (CATS, 1994) on modern statistics education (circa early 1990's) with a provocative quote:

> Competent statisticians will be front line troops in our war for survival — but how do we get them? I think there is now a wide readiness to agree that what we want are neither mere theorem provers nor mere users of a cookbook. A proper balance of theory and practice is needed and, more important, statisticians must learn to be good scientists, a talent which has to be acquired by experience and example. —George E. P. Box, "Science and Statistics" (CATS, 1994, p. vi)

They also noted (our emphases in *italics*):

> At its August 1992 meeting in Boston, the Committee on Applied and Theoretical Statistics (CATS) noted widespread sentiment in the statistical community that upper-level *undergraduate* and graduate *curricula for statistics majors* and postdoctoral training for statisticians are *currently structured in ways that do not provide sufficient exposure to modern statistical analysis, computational and graphical tools, communication skills*, and the ever growing interdisciplinary uses of statistics. Approaches and materials once considered standard are being rethought. The *growth that statistics has undergone is often not reflected in the education that future statisticians receive*. There is a need to *incorporate more meaningfully* into the curriculum the *computational and graphical tools* that are today so important to many professional statisticians. (CATS, 1994, p. vii).

As these excerpts imply, aspects of data management, data wrangling, exploratory data analysis, visualization, modeling, computation, and communication have long been important components of statistical practice and education. However, the recent dramatic growth of data science has challenged statisticians to ensure that students develop the capacity to make sense of the data they encounter. We remain concerned that while there has been progress in the intervening decades, many of the issues noted in the early 90's remain and that considerable additional work is needed to prepare undergraduate statistics students to enter the world and generate insights from data.

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

We consider the issues in the context of a review of recent statistics and data science curricular guidelines and initiatives at the undergraduate (post-secondary) level, identify key topics for curricular development, and highlight important challenges and issues that require further work. We believe that emphasizing computation and communication tools will help students understand and work with ideas of variability, inference, and design and ensure that key concepts of statistics are embedded in data science education. We focus on ideas that data science brings to statistics.

WHAT IS DATA SCIENCE?

What do we mean by data science? Donoho (2017), quoted statistician John Tukey, who in 1962 presaged "an as-yet unrecognized science, whose subject of interest was learning from data, or 'data analysis'". In his recent paper, Donoho updated Tukey's definition and provided a description of data science as a new field that married computational and inferential methods. We present his proposed structure for "Greater Data Science" (GDS), which includes six main divisions for future curricula:

- Data exploration & preparation: the 80% (or more) of data wrangling needed prior to analysis
- Data representation and transformation: including modern databases and special types of data
- Computing with data: multiple environments, high-performance computing, and workflow
- Data visualization and presentation: as a way to explore and present results
- Data modeling: including generative (stochastic model) and predictive (learning)
- Science of data analysis described as one of the most complicated of all sciences

Many efforts are underway to teach statistics and data science in an integrated fashion that allow graduates to extract meaning from data while ensuring their work is on a solid foundation. Some of these efforts are primarily housed in computer science departments, while others are found in statistics departments. Individual structures will vary dramatically by institution; however, we believe that integrated programs will be most successful in developing effective and sustainable data science initiatives. No matter how new programs are developed, there is still need for computational and statistical disciplines to join forces in order to improve learning in both areas.

CURRICULAR GUIDELINES AND WORKING GROUPS

Recently, multiple workshops and ad-hoc committees have been formed to address the issues of updates and modernizations to the statistics and data science curricula. A common theme running through the recommendations is that students should be exposed to full data analysis problems, from data collection and wrangling through reporting. The consistency with which the recommendations are being made gives a clear indication of the direction to take the curriculum, but as we note below, it is not always easy to know *how* to move the curriculum forward.

In 2014, a working group from the ASA updated prior undergraduate guidelines from 2000. Despite having been written 14 years apart, both guidelines stress depth in computational and communication capacities while practicing the statistical problem-solving cycle. The latter document details the expectation as follows:

> All too often undergraduate statistics majors are handed a 'canned' data set and told to analyze it using the methods that are currently being studied. This approach may leave them unable to solve more complex problems out of context, especially those involving large, unstructured data. The statistical analysis process involves formulating good questions, considering whether available data are appropriate for addressing the problem, choosing from a set of different tools, undertaking the analyses in a reproducible manner, assessing the analytic methods, drawing appropriate conclusions, and communicating results. (ASA Undergraduate Statistics Guidelines Group, 2014).

A group met under the auspices of the Park City Mathematics Institute to consider curriculum guidelines for undergraduate data science programs (De Veaux et al, 2017). The effort explored how to construct new courses (and revise others) to ensure that students have the capacity to "think with data" in nimble and supple ways.

> Because data-driven problems are often messy and imprecise, students should be able to impose mathematical [ideas] on [data science] problems by developing structured mathematical problem-solving skills. Students should have enough mathematics to understand the underlying structure of common models used in statistical and machine learning as well as the issues of optimization and convergence of the associated algorithms (De Veaux et al, 2017).

Related curricular initiatives organized by the National Academies include the study entitled "Envisioning the Data Science Discipline: The Undergraduate Perspective" (Committee on Envisioning the Undergraduate Data Science Discipline interim study report, 2017: final report anticipated May 2018) and the Data Science Education Roundtable (http://sites.nationalacademies.org/deps/bmsa/deps_180066). Relevant interim findings from the Envisioning Study include the importance of developing data acumen, and the need to incorporate real data, broad applications, and commonly used methods and approaches.

The aspects of communication and computation in the interim study from the NAS findings reinforce those of the aforementioned statistics and data science guidelines. It is exciting to see that the larger statistics and data science communities have embraced the guidelines and started to implement changes. However, the changes need to be happening at a faster rate and more comprehensively. Below, we provide key information and guidance on how to implement change effectively.

IMPORTANT CHALLENGES AND QUESTIONS

In both statistics and data science guidelines, there is emphasis on computation, communication, and practice implementing the full analysis workflow. As statisticians, we are committed to ensuring that variability, inference, and design remain central to the curriculum, but we see the value in teaching the broader learning outcomes of data science. What needs to be done? We implore educators to make a concerted effort to update every course. To avoid overload, the changes need to be implemented in pieces (and not for all courses at the same time). To ensure this occurs, the updates must happen now and must not be deferred to some future date. (In our own teaching, we have both found it helpful to commit to making at least **one** update to every course each time it is taught.) In this section, we explore important hurdles and suggest ways to overcome barriers.

Capstone Projects / Individual Data Analysis

Possibly the most effective way for students to become proficient at working through the data analysis pipeline is to practice, ideally through a capstone program (Martonosi and Williams, 2016). Some programs have senior theses, others integrate projects into classes, and some institutions have consulting centers with which students can become involved. Outside of the curriculum, we should be encouraging our students to work on data analysis projects that are most interesting to them. There are myriad data challenges ongoing at any time. The ASA regularly organizes a Data Expo (e.g., http://community.amstat.org/stat-computing/data-expo/data-expo-2018), and each spring there are many DataFest competitions internationally (https://ww2.amstat.org/education/datafest).

Even without a grade or faculty feedback, an engagement with data analysis (e.g., made public by the student's GitHub repository) will provide a student with important experience and can demonstrate that the student has the capacity and creativity to solve interesting problems. Whether the practice is within the curriculum (Martonosi & Williams, 2016) or outside of it (Gould, 2014), we highly recommend that faculty encourage their students to be working through the entire analyses from wrangling to final product.

Computation in the Statistical Theory & Related Courses

Much of the energy from the statistics education community is focused on introductory statistics, understandable given the large number of students who take only that course. However, considerable work is needed to restructure other courses to help ensure that they address new learning outcomes based on computation, communication, and the data analysis pipeline. As an

example, consider the augmented focus on computation and communication in the theoretical statistics course (Horton, 2013; Horton, Brown, and Qian, 2004) and other curricular shifts at baccalaureate institutions (McNamara, Baumer, and Horton, 2017). By adding computation to the traditional theory course, students can see the connections between the mathematical ideas they are learning and the results of their data analyses. We do not suggest taking out the mathematical pieces of the course; instead, we hope to emphasize the mathematics using computational tools. A stronger computational component in the statistical theory course will help students know "enough mathematics to understand the underlying structure of common models used in statistical and machine learning" (De Veaux et al, 2017).

Ethics

Examples of data improprieties and ethical lapses highlight the increased importance of ethics in both statistics and data science curricula. From data reproducibility to confounding (see other challenges below) to machine learning algorithms, the analyst must think carefully when communicating statistical results. O'Neil (2016) discusses machine learning bias in a thought provoking book which details the ethics of decision making with AI in areas from public school teacher evaluation, payday lending, and criminal sentencing. Buolamwini and Gebru (2018) discuss racial bias associated with facial recognition software and the problems that ensue. The examples allow students to see how models and analyses are applied in a broader context of decision-making. Additionally, students should be able to think about non-misleading visualizations and aspects of reproducibility putting all of the pieces into context (Baumer et al. 2017).

New or augmented learning outcomes (e.g., confounding and cross-validation)

A challenge to our existing curriculum is that many interesting and important learning outcomes are hidden behind a long list of prerequisite courses. Big and important ideas should be introduced to students early and throughout the curriculum. Consider two vitally important analysis ideas: cross-validation and confounding. Both ideas require time and commitment from the instructor to cover them in detail, but neither concept requires a substantial amount of background knowledge. Pedagogical tools for introducing cross-validation into the classroom can be found in James et al. (2013), and cross-validation can be used as an assessment of whether the inference analysis (e.g., stepwise variable addition in linear models) is appropriate. For confounding, online resources that integrate ideas of confounding and causation into a statistics classroom include Cosma Shalizi's online text (http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV) and Hernan's online EdX course (https://www.edx.org/course/causal-diagrams-draw-assumptions-harvardx-ph559x).

Modernizing Statistics Toolbox / Faculty development

A major challenge with the incorporation of computation into the curriculum is that many (most?) faculty who teach introductory statistics don't have substantive experiences with data (ASA/MAA Guidelines, 2014). Instructors need to have an appropriate background in applied statistics and the statistical problem-solving cycle to be able to effectively teach statistics. To teach the tools necessary for data science, instructors need to be comfortable with an even broader set of computational and communication skills. We encourage professional development toward working with modern versions of R (as a start, consider the following tidyverse packages: dplyr, ggplot2, broom, infer, skim, see https://tidyverse.org). As an encouraging note, the recent development of flexible online courses (e.g., Coursera Data Science specialization and a number of Python and R based courses on DataCamp) have some potential to help address the faculty development challenge.

Reproducibility tools

There is no question that the crisis in reproducibility and the tools designed to ameliorate the reproducibility problem have fundamentally changed the way students engage with software like R and Python. By having R Markdown template files to work with on the very first day of class (Baumer et al., 2014), students are able to "fall into the pit of success"

(https://blog.codinghorror.com/falling-into-the-pit-of-success/). Becoming fluent in R and R Markdown takes some learning by an instructor, but with template files and on-line tools, the learning curve for students is extremely short. An additional layer of reproducibility includes tools for collaborative workflow, e.g., Git and GitHub (Bryan, 2018). And while there is still work being done integrating Git more seamlessly in the classroom, GitHub Classroom and Happy Git with R (http://happygitwithr.com) go a long way toward making the resources more accessible. Indeed, Git and GitHub are being used in first year courses in Duke University's "Introduction to Statistics" and "Introduction to Data Science" (Çetinkaya-Rundel & Rundel, in press).

Visualization

There seems to be agreement in the statistics community that good visualization is an important aspect of data analysis. However, a challenge of having visualization as a course topic is that learning about visualization does not often conform to a typical classroom. Teachers might think: How would I structure a lecture? What resources exist? What does a homework assignment look like? Fortunately, textbooks do exist (e.g., Wickham and Grolemund, 2017, Baumer et al, 2017, and http://serialmentor.com/dataviz) with ideas for lectures, assignments, and assessment. Additionally, we encourage bringing creative strategies for getting students thinking about graphs and figures into the classroom (Nolan and Perrett, 2016).

CONCLUSION

While not new, the relationship between statistics and data science is an important question facing the wider statistics community. In terms of educational programs, the statistical education community has been challenged to find ways to teach new learning outcomes that are necessary to make sense of these data, while maintaining core statistical concepts that are fundamental to good data science. In addition to key ideas and concepts, such as variability and inference (broadly defined to include aspects of confounding), that have traditionally been at the center of statistics curricula, new emphasis is needed on computational and communication capacities.

We have not addressed the other side of our ideas: what can statistics bring to data science? We believe that issues of variability and design should play a central role in data science curricula, but we do not have room in the current work to expand on such implementation. If students can recognize the variability associated with predicted values in a machine learning model, they will gain a deeper understanding of not only the models used but also the appropriateness of any conclusions being reported. Additionally, ideas of confounding and experimental design are of utmost importance given reliance on large observational datasets.

Many of our suggestions and observations may seem deliberately provocative. This is intentional. We believe that the statistics community needs to act decisively and to act now if substantive changes in what and how we teach are to be made. We have noted previously (Hardin and Horton, 2017) that the mathematics community needs to engage in curricular discussions and changes to ensure that mathematics is not left behind. The same existential threat exists for the statistical education community. To be part of the future requires working with computational scientists and domain experts in a "team science" approach to help develop new skills and approaches that complement our existing strengths. Without such interdisciplinary collaborations in combination with substantive changes to our curriculum and new approaches to teaching, we run the risk of being left out of the growing field of data science.

REFERENCES

ASA Undergraduate Statistics Guidelines Group (2014). ASA Curriculum Guidelines for Undergraduate Statistics Programs, www.amstat.org/asa/education/Curriculum-Guidelinesfor-Undergraduate-Programs-in-StatisticalScience.aspx.

ASA/MAA Guidelines. (2014). *ASA/MAA develop guidelines for teaching introductory statistics course*, https://magazine.amstat.org/blog/2014/04/01/asamaaguidelines.

Baumer, B.S. (2018a). A grammar for reproducible and painless Extract-Transform-Load (ETL) operations on medium data (https://arxiv.org/pdf/1708.07073.pdf).

Baumer B.S. (2018b). Lessons from between the white lines for isolated data scientists. *The American Statistician*, in press.

Baumer B.S., Cetinkaya-Rundel M., Bray A., Loi L., & Horton N.J. (2014). R Markdown: integrating a reproducible analysis tool into introductory statistics, *Technology Innovations in Statistics Education*, *8*(1), http://escholarship.org/uc/item/90b2f5xh.

Baumer, B.S., Kaplan, D. T., and Horton, N.J. (2017). Modern Data Science with R, Boca Raton, FL: CRC Press, http://mdsr-book.github.io.

Buolamwini, J. and Gebru, T. (2018). Gender shades: intersectional accuracy disparities in commercial gender classification, *Proceedings of Machine Learning Research*, *81*, 1-15, http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Bryan, J. (2018). Excuse me, do you have a moment to talk about version control? (https://peerj.com/preprints/3159). *The American Statistician*, in press.

Çetinkaya-Rundel, & Rundel, C. (in press). Infrastructure and tools for teaching computing throughout the statistical curriculum, *The American Statistician*, in press.

Chambers, J.M. (1993). Greater or lesser statistics, *Statistics and Computing*, *3*(4), 182-184, https://link.springer.com/article/10.1007/BF00141776.

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

Committee on Envisioning the Data Science Discipline: The Undergraduate Perspective: Interim Report (2017). Washington, DC: National Academies Press, https://www.nap.edu/catalog/24886/envisioning-the-data-science-discipline-the-undergraduate-perspective-interim-report.

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., ... & Kim, A. Y. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, *4*, 15-30. DOI:10.1146/annurev-statistics-060116-053930.

Donoho, D. (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766. http://www.tandfonline.com/doi/abs/10.1080/10618600.2017.1384734.

Gould, R. (2014). Datafest: Celebrating data in the data deluge. In *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics*. https://icots.info/9/proceedings/pdfs/ICOTS9_4F2_GOULD.pdf.

Horton, N.J., Baumer B.S., & Wickham H. (2015). Setting the stage for data science: Integration of data management skills in introductory and second courses in statistics, *CHANCE*, *28*(2), 40-50, http://chance.amstat.org/2015/04/setting-the-stage.

Horton, N.J., Brown E.R., & Qian L. (2004). Use of R as a toolbox for mathematical statistics exploration. *The American Statistician*, *58*(4), 343-357.

Horton, N.J., & Hardin, J.S. (2017). Ensuring that mathematics is relevant in a world of data science. *Notices of the American Mathematical Society*, *64*(9), 986-990. https://www.ams.org/publications/journals/notices/201709/rnoti-p986.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. New York, NY: Springer.

Martinosi, S.E. & Williams, T.D. (2016). A survey of statistical capstones, Journal of Statistics Education, *24*(3). http://tandfonline.com/doi/full/10.1080/10691898.2016.1257927.

McNamara A., & Horton N.J. (in press). Wrangling categorical data in R. *The American Statistician*, in press..

McNamara A., Horton N.J., & Baumer B.S. (2017). Greater data science at baccalaureate institutions, *JCGS*, *26*(4), 781-783.

Nolan, D. and Perrett, J. (2016). Teaching and learning data visualization: ideas and assignments, *The American Statistician*, *70*(3), 260-269.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality & Threatens Democracy*. New York, NY: Crown Publishing Group.

Pérez F., Granger B.E. (2007) IPython: A system for interactive scientific computing. *Computing in Science & Engineering*, *9* (3), 21–29.

Wickham, H., & Grolemund, G. (2017). *R for Data Science*. Sebastopol, CA: O'Reilly Media.