BUILDING A FOUNDATION IN STATISTICS IN THE ERA OF DATA SCIENCE

Alison L. Gibbs
Department of Statistical Sciences
University of Toronto, Toronto, Ontario, Canada M5S 3G3
alison.gibbs@utoronto.ca

In the era of data science, our undergraduate statistics programs need to cover more territory, including data analysis, statistical theory, computational skills for data wrangling, visualization and analysis, and oral and written communication. Ideally, a first course will introduce students to all of this territory, while immersing them in a framework for solving problems with data, illustrating descriptive, explanatory, and predictive purposes. We'll describe an introductory course in statistical reasoning and data science that attempts to do this while placing a particularly strong emphasis on communication. The course is a microcosm of our newly renovated statistics program, introducing students to all the knowledge, skills and attitudes that we want them to acquire as they pursue their degrees.

INTRODUCTION

It has been almost 10 years since Google's chief economist Hal Varian declared "the sexy job in the next 10 years will be statisticians" (Lohr, 2009) and over 5 years since Davenport & Patil (2012) substituted "data scientist" for "statistician" as the "sexiest job for the 21st century." In that time, the number of undergraduate degrees awarded in statistics and biostatistics has risen rapidly (Pierson, 2017). At the University of Toronto, a large publicly-supported, research-intensive university, undergraduate enrolment in statistics programs of study has outpaced this growth, with an almost 7-fold increase in this time period. There are now 3600 undergraduates (in 2nd, 3rd and 4th years of study) enrolled in statistics programs. Over 80% of these students are combining their statistics studies with a major or minor in another discipline. Why are so many students flocking to study statistics? Undoubtedly, they are attracted by the good press, the desire to acquire skills that are in demand in the job market, and the excitement of being part of the new field of data science which is now well-established enough to be listed in 2018 as a job category in the U.S. Bureau of Labor Statistics' Occupational Outlook Handbook (Little, 2016).

In the context of this recognition and demand, how do we effectively design a statistics program that both prepares our students for the needs of the current labour market and prepares them for future challenges that we have not yet imagined, as data sources, technology, and methods continue to evolve? A particular challenge in designing a relevant program of study in the era of data science is that a consensus opinion on the definition of "data science" is still elusive. There is common agreement that data science is focused on *practical* problems, whose solutions require both "multidisciplinary and interdisciplinary approaches for extracting knowledge or insights from data," relying on "processes and systems", and creating "novel techniques to address the 'cracks' between [the foundational] disciplines" (National Academies of Sciences, Engineering, and Medicine, 2017, 1-2). Some have argued that "data science is statistics" (for example, Greenhouse, 2013), but most contend that data science requires skills not traditionally considered part of statistical training, such as the ability to curate data effectively and to build data products. It is also differentiated from traditional statistics by its emphasis on "algorithmic thinking rather than its inferential justification" (Efron & Hastie, 2016, p. xvi). Most agree that statistics does hold a central role (Wickham, 2014, van Dyck et al., 2015). And while some have felt that statistics is at risk of becoming irrelevant "because we have focused too much on doing the right thing, rather than doing something that works" (Wickham, 2015), we hold the view that it is possible to envision a statistics curriculum that is broader than our traditional curriculum which was narrowly focused on inferential methods and their mathematical foundations, that continues to value and nurture a "doing the right thing" mindset, that has aspects of the agility of data science, and that prepares our students to adapt their expertise to new situations.

In light of this, we describe some considerations for the design of a program for undergraduate students majoring in statistics, and, for these students, how one might design the first course. As an illustration, we outline recent developments in statistics programs at the

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

University of Toronto and how these developments are reflected in the first course for students entering statistics programs.

A STATISTICS PROGRAM IN THE ERA OF DATA SCIENCE

Traditionally, undergraduate programs in statistics have been designed around methods and tools for extracting knowledge from data and for understanding the uncertainty associated with this knowledge. With the rise of data science, there is recognition of the need for statistics programs to also include data visualization, experience with more complex data, data wrangling and other computational skills, and more balanced exposure to various purposes of data analysis, including prediction, description and explanation (American Statistical Association, 2014).

Statistics programs should also prepare students to apply their knowledge to the solution of real-world problems, involving data of various types collected under a variety of circumstances. To be able to do this effectively, statisticians must be skilled at communication and collaboration, in order to acquire a "nontrivial understanding" of the problem, make "judgments about the relevance and representativeness of the data," and be able "to interpret and communicate the results … so non-statisticians can understand the findings" (Greenhouse, 2013). And to contribute to the solution of real-world problems, statisticians have always needed a combination of flexibility, creativity, and pragmatism. In the era of the "data revolution" (Ridgway, 2015), when much data is collected without design and without questions in mind, the ability to take a flexible approach to learning from data has never been more essential.

Yet the core ideas of statistics remain as important as ever. These core ideas include finding patterns in data and assessing the associated uncertainty, an understanding of when inference is and is not appropriate, necessary, or helpful, and ways to approach situations that are associated with multiple testing, confounders, or overfitting. Through the creation of statistics programs that develop expert understanding of these ideas, we can prepare our students to adequately respond to suggestions that because we now often have huge amounts of data, learning from data can happen without statistical thinking. The development of "post-selection inference," for example, is a demonstration of the contribution of statistical thinking to the reproducibility crisis (Kuffner & Young, 2017). Meng (2009) characterized the difference between someone with statistical training and someone without as the difference between an amateur and a professional, with "[a] key sign distinguishing a professional from an amateur [being] the person's ability to assess what can be done, what cannot be done, and what should not be done even is s/he has all sorts of incentives to do so." This "policing" role is essential for scientific progress and we do not want to cede our ability to lead here and our obligation to train students who will carry on in this way. Moreover, as argued by Horton (2015), a key role for statisticians will be to contribute to creating a field of data science that is more rigorous and scientific. Indeed, continued attention to the core ideas of statistics can differentiate statistics programs from practically oriented data science programs and will ensure their relevance.

In this context of both a broadening discipline and our astounding enrolment growth, the University of Toronto has been reimagining our statistics programs to reflect how we envision training in statistics for the modern era. In our view, training in data science overlaps, but also has a strong emphasis on computer science, and we have new programs (at the undergraduate and graduate levels) in data science that are offered jointly with computer science.

A set of 20 program learning outcomes articulate what students in our undergraduate statistics programs should know and be able to do by the time they graduate. These program learning outcomes encompass "knowledge and understanding, abilities, habits of mind, modes of inquiry, dispositions, [and] values" (Maki, 2002). The learning outcomes of our new programs are categorized into five themes:

1. *Statistical Theory, including Probability:* These outcomes include understanding and application of the role of probability as a mathematical framework to represent uncertainty, paradigms for statistical inference, and theoretical results that support the rationale of statistical methods. Cobb (2015) has exhorted us to rethink the role of mathematics in our curriculum. We have chosen not to dismantle the mathematical structure that has historically been at the core of our program, but instead maintain it in a parallel theme that each student experiences early in their program, and some will choose to dive deeply into.

2. *Methods and Applications:* Program learning outcomes in this category include appropriate application of a variety of statistical methods for purposes of description, prediction and explanation (Shmueli, 2010), study design and data collection strategies, and data visualization. Student development of many of the components of statistical thinking are articulated in these outcomes. There are greater emphases on prediction and methods which rely on algorithms to solve problems than in our previous curriculum, reflecting how statistics has evolved in the past 60 years (Efron & Hastie, 2016, p. xvii).
3. *Computational Thinking:* Our program learning outcomes in this category include the use of simulation for a variety of purposes, data manipulation, and computing for data analysis in a literate format that supports reproducibility. Our new program explicitly includes more data processing than in our previous curriculum, and a recognition of the need for students to be able to produce efficient algorithms that can accommodate large datasets and computationally intensive methods. While our programs did and continue to include a required introductory course in computer science, most of the development of computational methods and thinking is integrated throughout statistics courses, taught in an organic way to support the statistical work being done.
4. *Professional Practice:* Learning outcomes under the professional practice category include the ability to speak and write clearly for both technical and non-technical audiences, apply ethical principles, and collaborate with both statisticians and non-statisticians. It has long been recognized that changes in the nature of statistical work brought about by advances in information technology have increased the importance of communication and the importance of team work (Nicholls, 2001). Our new curriculum calls for more opportunities for our students to put these skills into practice in authentic situations.
5. *Problem Solving:* The learning outcomes in this category are broad, higher order, and applicable to problems in both statistical theory and application. They include applying and adapting existing knowledge to learn new methodologies and applying knowledge in new areas of application, evaluating the relative merit of competing approaches to problems, and critical thinking such as the recognition of strengths and limitations of data analysis. These outcomes reflect our goal to produce graduates who are not just experts in solving standard problems, but are "adaptive experts" who have the disposition to continually learn, to question the applicability of standard methods in new situations, and develop new approaches to solve new problems (National Research Council, 2000, p. 48).

In designing our programs around these themes and mapping courses to our program learning outcomes, we found it helpful to keep in mind Wiggins & McTighe's (2005) advice that an effective program is not a "march through content" but is about "learning to perform with content" (p. 292). We also found it helpful to think in terms of "essential questions," such as "*How do we effectively learn from data?",* that recur throughout the curriculum and that may help students make sense of the knowledge and skills accumulated throughout their program in a holistic and integrated way. Developing an understanding of the essential questions of a discipline "requires rethinking and constant (re-)application" (Wiggins & McTighe, 2005, p. 291). With this in mind, we were tasked with designing a new first course for statistics students.

THE FIRST COURSE

Traditionally, introductory courses in statistics programs have taken one of two general approaches: (1) a probability course followed by a course in statistical theory with an emphasis on inference, giving a mathematical introduction to the discipline; this approach can delay meaningful engagement with data to later in the program and requires students to develop their understanding of how to learn from data by discovering practical applications of mathematical abstractions, or (2) a non-mathematical introduction, with varying degrees of emphasis on statistical literacy and statistical practice, typically focused on students' development of conceptual understanding of estimation and inference for purposes of explaining causes or relationships. Our previous curriculum allowed students to choose one of these two starting points.

There have been many innovative ideas for alternative first courses. Gould (2010) gives a history of major developments in approaches to teaching introductory statistics, including the movement toward using real data to develop conceptual understanding and a call to give students

exposure to more and a greater variety of data. Other important ideas include Weldon's (2014) call for immersion in practical problems with data, with the development of the problem-solving process taking precedence over the need to cover a long list of content. Some more recent innovative first courses integrate elements of data science. Examples of this are a joint course with computer science (Jordan, 2016), an introductory course which moves quickly to topics more typically seen in more advanced multivariate analysis courses (Wagaman, 2016), a course strongly focused on modeling (Kaplan, 2017), and first and second courses designed to supplement statistical methods and concepts with computation (Hardin et al., 2015).

More broadly, first university courses in a discipline can serve many purposes. They can be the public face of the discipline, introducing many students to the ways of knowing that the discipline uses. They can give an overview of the discipline, introducing its vocabulary, methodologies and the core ways of thinking and types of arguments used. Or, taking a more focused approach, they can establish prerequisite foundational knowledge and conceptual understanding needed for future courses in a program of study.

Recognizing that learning is recursive, and that rethinking and re-application are required to connect ideas and develop a broad understanding of both the range and the common core of statistical methods, and that this will happen for different students at different times, we decided to use the model of a survey course for our first course. Our goal is to have students leave the course with the understanding that statistics is not a collection of methods, but a way of thinking and doing. The course includes elements of inferential thinking, mathematical modeling, computing with data, and solving problems using computationally intensive methods, with examples chosen to provide illustrations of cases where learning from data may require prediction, inference, or description. It was intentionally designed to introduce students to all themes in our program learning outcomes, and, in contrast with other introductory statistics courses, it is solely intended for students who will enroll in our statistics major program and will be engaging more deeply with the curriculum in later courses. The course gives students a glimpse of the statistics program so they can later better choose among program options that are appropriate to their interests and strengths. Our current enrolment in statistics programs at the University of Toronto means we do not have a pressing need to recruit more students, but we do want our students to be able to make an informed decision about whether a statistics program is right for them.

The course structure includes:

- Equal time in large lecture sections (2 hours per week) and communication focused tutorials (20 students, 2 hours per week).
- Tutorials which include short writing assignments and brief oral presentations, designed to develop students' ability to write clearly and speak fluently in the context of a statistical problem or a real-world scenario.
- A final project in which groups of students present their analyses of a large multivariate data set and scientific findings in a poster session. The project integrates the use of computational skills, statistical reasoning, and the oral and written communication skills developed in lectures, practice problems, and tutorials. It is both a learning and assessment task, designed to encourage deeper and integrated understanding of the course material and, by extension, the discipline.
- Teamwork, weekly in tutorials and through the final project. The final project also gives students experience managing a small project.

The content covered incorporates:

- A broad view of statistical practice, with a variety of applications for a wide range of purposes. Students engage in activities designed to illustrate how domain questions can be developed into questions that can be addressed with statistical methods, how to manipulate and explore data, and the types of arguments, founded in mathematics or algorithms, used by statisticians.
- Computation emphasizing literate programming in a reproducible format, through examples that can be emulated and starter code that can be adapted and combined, encouraging the acquisition of good habits in style and reproducibility.
- Discussion of some ethical issues encountered in statistical practice and reflection on how they are relevant to the variety of situations encountered in the course, including the connection

between carrying out reproducible work and ethical practice (Thompson & Burnett, 2012). Through this reflection and the development of technical and communication competencies, students begin to develop their professional identity.

We're aware that there is room for improvement in our course design. Because of the breadth of our goals, the depth of coverage of any of the content is very limited. But we hope that students leave the course with a window into our program and curiosity to learn more. We are optimistic that a broad introduction in their first course will allow them to make connections among the ideas in the more specialized courses they will encounter later in their programs. And we hope that the students are developing habits in their approaches with respect to computation and writing, in particular, that they will carry forward.

CONCLUDING REMARKS

Many have recently commented that this is an "exciting time to be a statistician" (for example, Horton & Hardin, 2015, Ridgway, 2015). Even more so, it is an exciting time to be a statistics educator. The rapid growth in prominence and reach of data science and the rapid increase in demand for statistics programs has forced us to think carefully about how we are training our students. Despite some early hand-wringing about the continued relevance of statistics, we believe that there is an essential role for statistical thinking in data science and that we have the obligation and opportunity to prepare students to be leaders in an increasingly data-driven world.

While we have made many changes and we have expanded our vision of what we teach, we must continue to adapt. The rise of data science has shown us that we need to be prepared to respond to the continuing evolution of applications, technology, and computational algorithms in a much more agile way than we have needed to before. Similarly, we must train our student to think beyond what they have already mastered, and to learn how to learn, so that they graduate from our programs with the agility to adjust to changing technologies and data, and the evolving nature of statistics and data science roles.

REFERENCES

American Statistical Association. (2014). "2014 Curriculum Guidelines for Undergraduate Programs in Statistical Science." Retrieved from http://www.amstat.org/asa/files/pdfs/EDU-guidelines2014-11-15.pdf.

Cobb, G.W. (2015). Rejoinder to Discussion of "Mere Renovation is Too Little Too Late: We Need to Rethink Our Undergraduate Curriculum from the Ground Up." *The American Statistician, 69*(4), online supplement. Retrieved from http://www.tandfonline.com/doi/suppl/10.1080/00031305.2015.1093029.

Davenport, T.H., & Patil, D.J. (2012). "Data Scientist: The Sexiest Job of the 21st Century." *Harvard Business Review*. Retrieved from https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.

Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence and Data Science.* New York: Cambridge University Press.

Gould, R. (2010). Statistics and the Modern Student. *International Statistical Review, 78*(2), 297-315.

Greenhouse, J. (2013, July 26). "Statistical Thinking: The Bedrock of Data Science." *Huffington Post Blog*. Retrieved from http://www.huffingtonpost.com/american-statistical-association/statistical-thinking-the-bedrock-of-data- science_b_3651121.html.

Hardin, J., Hoerl, R., Horton, N.J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Temple Lang, D., & Ward, M.D. (2015). Data Science in Statistics Curricula: Preparing Students to "Think with Data." *The American Statistician 69*(5), 343-353.

Horton, N.J. (2015). Challenges and Opportunities for Statistics and Statistical Education: Looking Back, Looking Forward. *The American Statistician 69*(5), 138-145.

Horton, N.J., & Hardin, J.S. (2015). Teaching the Next Generation of Statistics Students to "Think With Data": Special Issue on Statistics and the Undergraduate Curriculum. *The American Statistician, 69*(4), 259-265.

Jordan, M. (2016). Computational Thinking and Inferential Thinking: Foundations of Data Science. Electronic Conference on Teaching Statistics 2016 Keynote Presentation. Retrieved from https://www.causeweb.org/cause/ecots/ecots16/keynotes/jordan.

Kaplan, D. (2017). Teaching Stats for Data Science, *The American Statistician*, (to appear). Retrieved from https://doi.org/10.1080/00031305.2017.1398107.

Kuffner, T.A., & Young, G.A. (2017). Principled Statistical Inference in Data Science. Preprint retrieved from https://www.math.wustl.edu//~kuffner/papers/KuffnerYoung2017a.pdf.

Ledbetter, M.L., & Campbell, A.M. (2005). A Survey of Survey Courses: Are They Effective? Argument Favoring a Survey as the First Course for Majors. *Cell Biology Education, 4*, 133-137.

Little, H.L. (2016, November 3). Jobs of the Future [Blog post]. Retrieved from http://knowledgequest.aasl.org/jobs-of-the-future/.

Lohr, S. (2009, August 5). For Today's Graduate, Just One Word: Statistics. *The New York Times*. Retrieved from http://www.nytimes.com/2009/08/06/technology/06stats.html.

Maki, P. (2002). Developing an Assessment Plan to Learn about Student Learning. *The Journal of Academic Librarianship, 28*(1), 8–13.

Meng, X.-L. (2009). Desired and Feared – What Do We Do Now and Over the Next 50 Years? *The American Statistician, 63*(3), 202-210.

National Academies of Sciences, Engineering, and Medicine. (2017). *Envisioning the Data Science Discipline: The Undergraduate Perspective: Interim Report*. Washington, DC: The National Academies Press. Retrieved from https://doi.org/10.17226/24886.

National Research Council. (2000). *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. Washington, DC: The National Academies Press. Retrieved from https://doi.org/10.17226/9853.

Nicholls, D. (2001). Future Directions for the Teaching and Learning of Statistics at the Tertiary Level. *International Statistical Review, 69*(1), 11-15.

Pierson, S. (2017). Bachelor's, Master's Statistics and Biostatistics Degree Growth Strong Through 2016. *AmStat News*. Retrieved from http://magazine.amstat.org/blog/2017/10/01/degrees16/.

Ridgway, J. (2015). Implications of the Data Revolution for Statistics Education. *International Statistical Review, 84*(3), 528-549.

Shmueli, G. (2010). To Explain or Predict? *Statistical Science, 25*(3), 289-310.

Thompson, P.A., & Burnett, A. (2012). Reproducible Research. *CORE Issues in Professional and Research Ethics, 1*, Paper 6. Retrieved from https://nationalethicscenter.org/resources/734/download/Thompson.pdf.

van Dyck, D., Fuentes, M., Jordan, M., Newton, M., Ray, B.K., Temple Lang, D., & Wickham, H. (2015, October 1). ASA Statement on the Role of Statistics in Data Science. *AmStat News*. Retrieved from http://magazine.amstat.org/blog/p2015/10/01/asa-statement-on-the-role-of-statistics-in-data-science/.

Wagaman, A. (2016). Meeting Student Needs for Multivariate Data Analysis: A Case Study in Teaching an Undergraduate Multivariate Data Analysis Course. *The American Statistician, 70*(4), 405-412.

Weldon, K.L. (2014). Experience Early, Logic Later. In H. MacGillivray et al. (eds.), *Topics from Australian Conferences on Teaching Statistics: OZCOTS 2008-2012*. New York: Springer, Springer Proceedings in Mathematics & Statistics, 81, 25-42.

Wickham, H. (2014). "Data Science: how is it different to statistics?" *IMS Bulletin 43*(6), 7. Retrieved from http://bulletin.imstat.org/2014/09/data-science-how-is-it-different-to-statistics/.

Wickham, H. (2015). Teaching Safe-Stats, Not Statistical Abstinence. *The American Statistician, 69*(4), online supplement. Retrieved from http://www.tandfonline.com/doi/suppl/10.1080/00031305.2015.1093029.

Wiggins, G. & McTighe, J. (2005). *Understanding by Design* (expanded 2nd edition). Alexandria, VA: ASCD.