THE FUNDAMENTAL ROLE OF COMPUTATION IN TEACHING STATISTICAL THEORY

Alison L. Gibbs and Alex Stringer
Department of Statistical Sciences, University of Toronto
Department of Statistics and Actuarial Sciences, University of Waterloo
alison.gibbs@utoronto.ca

What skills, knowledge and habits of mind does a statistician require in order to contribute effectively as an inhabitant of the data science ecosystem? We describe a new course in statistical theory that was developed as part of our consideration of this question. The course is a core requirement in a new curriculum for undergraduate students enrolled in statistics programs of study. Problem solving and critical thinking are developed through both mathematical and computational thinking and all ideas are motivated through questions to be answered from large, open and messy data. Central to the development of the course is the tenet that the use of computation is as fundamental to statistical thinking as the use of mathematics. We describe the course, including its connection to the learning outcomes of our new statistics program of study, and the multiple ways we use and integrate computation.

INTRODUCTION

The emergence of "data science" has attracted enormous attention, resulting in new job titles and career paths, new courses and programs of study, and new scientific journals. But what is data science? While many have offered definitions, we prefer the conception of data science as an ecosystem, as described by Meng (2019), and agree that we need the collaborative and collective efforts of "at least computer science, statistics, engineering and operation research, information and library sciences, law and philosophy, (applied) mathematics, social and behavioral sciences, history of science, and data visualization, not to mention countless areas of application from astronomy to zoology and back to agriculture, and the vital participation of industry, government, NGOs, and beyond" to ensure its health and safety.

At the University of Toronto, the emergence of data science has led to extraordinary demand for undergraduate programs of study in statistics and the number of undergraduate students enrolled in these programs now numbers well over 4,000. Considering that they are drawn to these programs by their interest in data science, how do we best prepare our graduates to most effectively contribute their expertise in statistical thinking to the data science ecosystem? This question motivated a recent major curriculum renewal project.

A key tenet of our new curriculum is summarized by Nolan & Temple Lang (2010): "computational literacy and programming are as fundamental to statistical practice and research as mathematics." Following this tenet, a statistics curriculum must give equal importance to computation and mathematics in developing statistical thinking, and this has become a key principle in all of our courses, including new courses in statistical theory. We regard computation as fundamental to statistical thinking and hence teach statistical theory using computation as an "engine" (Cobb, 2015). Mathematics is present but with equal intellectual weight to computation, avoiding the need for relegation of the latter to a "unit within an existing course or a new course in an existing curriculum" (Cobb, 2015).

In this paper, we turn our focus to a newly developed course in statistical theory that is a core required course in our new curriculum. We begin by outlining the program learning outcomes of this new curriculum and we then describe the course and its connections to these learning outcomes. Among these connections is an emphasis on computation to develop statistical thinking and we next expand further on the three ways that computation is used in this regard.

A STATISTICS PROGRAM OF STUDY FOR THE DATA SCIENCE ECOSYSTEM

What skills, knowledge and habits of mind does a statistician require to contribute effectively as an inhabitant of the data science ecosystem? This question motivated our recent curriculum renewal project for undergraduate programs of study in statistics. The goal of this curriculum is the training of statisticians, who have all the strengths that statisticians have always brought to the scientific process. However, we also considered how our training should be reconsidered to develop modern and adaptable statisticians who are prepared to make significant contributions in the ecosystem of data science.

In: R Helenius, E Falck (Eds.), *Statistics Education in the Era of Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
Aug-Sept 2021, Online conference. ©2021 ISI/IASE iase-web.org/Conference_Proceedings.php

Our new curriculum has five themes. These are:

1. *Theory.* This theme includes probability as a mathematical framework to represent uncertainty, paradigms for statistical inference, and rationales for statistical methods.
2. *Methods.* Knowledge and correct application of a variety of methods for a variety of purposes, including inference, prediction, and explanatory modelling, are included here, as well as data visualization and design of data collection.
3. *Computational thinking.* In addition to the acquisition of programming skills, skills in data wrangling and proficiency in the use of statistical software for data analysis, this theme also includes the use of simulation for multiple purposes: to evaluate methods, supplement mathematical results, and as an approach to inference.
4. *Professional Practice.* All graduating students should develop the ability to present accurate, clear, concise descriptions of statistical methods and the results of analyses to both statisticians and non-statisticians, orally, in writing, and through appropriate data visualizations. They should also be able to contribute effectively to interdisciplinary terms and practice their work in an ethical manner.
5. *Problem Solving Skills.* As our students develop an understanding of statistical analysis as a unified framework, from study design through communication of results, they develop the ability to apply and adapt their knowledge to new areas, to learn new methods independently, to bring multiple approaches to problems and reason critically about their relative merits. All of this must take place in the context of real-world problems.

These themes are fully articulated in more than 20 program learning outcomes. In the mapping of these program learning outcomes to courses, we had two central considerations:

1. Each learning outcome needed multiple exposures, in order to give students the opportunity to be introduced to the outcome, reinforce their learning, and master the knowledge, skill, or habit of mind (Maki, 2004).
2. In order to develop students' understanding of statistics as a unified framework, every course should touch on as many of the themes as possible.

Like many traditional statistics programs of study, our pre-renewal programs for students majoring in statistics required a course in probability followed by a course in mathematical statistics early in the program. In this paper, we describe the results of a major rethink of this traditional mathematical statistics course "from the ground up" (Cobb, 2015) in the context of our new curriculum, with particular consideration to the fundamental role of computation. For descriptions of some other aspects of the new program and its courses see Gibbs (2018), Gibbs & Damouras (2019), and Gibbs & Taback (2021).

RETHINKING A COURSE IN STATISTICAL THEORY FOR THE DATA SCIENCE ECOSYSTEM

For many years, undergraduate programs of study in statistics have only been accessible to students with a strong background in mathematics, reflecting the mid-twentieth century disciplinary emphasis on the mathematical justification of statistical methods (Efron & Hastie, 2016, p. 448). However, we have discovered that the lure of careers in data science has attracted students with a more diverse array of interests and skills than traditional mathematical logic and proof. We agree with Brown & Kass (2009) who argued that such students should be accommodated and in fact actively recruited, and that having stringent mathematical requirements for even introductory courses in statistical theory directly opposes this effort. While we have also retained our traditional courses in theory that rely on mathematical rigour, which might more appropriately be described as mathematical statistics, we describe here a new course in statistical theory designed for students who are specializing in statistics and who will bring this specialization to their contributions to the data science ecosystem. This course was designed to be taken early in students' undergraduate program of study, to cover the essential ideas of statistical theory and to prepare them for all of our advanced statistics courses, but for which mathematics presents much less of a barrier to entry.

Inspired by broader conceptions of statistical theory of others such as Weldon (2010), our course is centred on the foundational ideas that connect all methods for extracting information from data. The course considers three overarching threads in statistical theory and data analysis:

1. Understanding Data: methods for describing data numerically and graphically, error and statistical models,
2. Making Inferences from Data: how data can be used to explain phenomena, and

3. Using Data to Make Predictions.

The course considers various perspectives on these threads, including Bayesian, frequentist, and likelihood approaches. In sharp contrast to our more traditional course in mathematical statistics, the course includes both methods that rely on mathematical thinking and methods that rely on computational thinking, with particular emphasis on computational approaches to analyzing data and understanding statistical methods. Throughout the course, motivation is provided by understanding a dataset collected from a public source, with context and purpose, in the spirit of Nolan & Speed (2000). While others have considered modern approaches to similar courses (for example, Green & Blankenship, 2015), our considerations extended well beyond the integration of effective pedagogical approaches, to a reconsideration of what type of thinking is essential in a modern course.

Our course comes after a first course in data science, a course in probability, and a course in calculus. Problem solving and critical thinking are developed through both mathematical and computational thinking which are integrated throughout the course. Although our approach is non-traditional, the content is standard for an introductory course in statistical theory. The course has five main units:

1. Exploratory data analysis,
2. Limit theorems,
3. Statistical models and estimators,
4. Statistical inference, and
5. Prediction.

These units are typical content for many courses at this level, with the possible exception of the amount of time we devote to exploratory data analysis and prediction. Each unit builds on and uses content from those before it. For example, histograms taught in unit 1 are used to illustrate the central limit theorem in unit 2; limit theorems are then used to visualize and contextualize the basic properties of estimation taught in unit 3, and this makes use of the histograms taught in unit 1; and so on. As this sequence illustrates, computation is a fundamental part of all units. For example, simulation experiments of the type taught in unit 2 are employed in unit 4 to assess the behaviour of statistical estimators and inferences in tandem with mathematical arguments. Ordering the units in this manner, with the computational work integrated throughout and used as a connecting thread, was done in an attempt to minimize the extraneous cognitive burden placed on students by having them learn both statistical concepts and computing tools in a disconnected manner (Woodard & Lee, 2020).

Throughout, we motivate all ideas through large, open and messy data and students are encouraged to take an exploratory approach to problem solving. Given the size, disorder, and multivariate nature of the data, and the limitations that come with exploring it in any other way, our students must rely on computer-aided exploration (as suggested by Cobb, 2015) to prepare and analyze the data. However, the role of computation in the course is much more than the use of statistical software as a tool to obtain a graphical display, estimate or p-value. We contend that, for the statistician in the data science ecosystem, computational thinking (in the broad sense described by Wing, 2006) must be as essential and innate as mathematical thinking. In the next section we describe the various roles of computation in the course.

THE FUNDAMENTAL ROLE OF COMPUTATION IN A MODERN COURSE ON STATISTICAL THEORY

Weldon (2010) discusses strategies for integrating theory and application in a statistics curriculum, and we apply a similar attitude towards integrating math and computation with theory in this single course. As we have noted, a key tenet of our new curriculum is summarized by Nolan & Temple Lang (2010): “computational literacy and programming are as fundamental to statistical practice and research as mathematics.” Similarly, for our new course in statistical theory, computation plays as fundamental a role as mathematics. Putting this principle into practice in the delivery of the course requires careful thought about the role of computation in all of the traditional content to be covered. Our approach is to assume whatever prerequisite knowledge we can, and then just use it; just as we don't draw attention to background knowledge every time we take a derivative, we don’t draw attention to background knowledge every time we wrangle a dataset. In this way, computation is presented as a fundamental and an expected part of statistics, both statistical theory and statistical practice.

We integrate computation into the traditional mathematical statistics course in three ways: computation as a standalone method for making inferences; computation as a means for understanding and appreciating theory; and computation as a means of applying theory. In order to integrate this thinking into our course, we developed supplementary course materials to accompany a more traditional textbook; these materials can be found at https://awstringer1.github.io/sta238-book/. We now briefly describe each of the three roles of computation with examples from the course.

Computation as a standalone method for making inferences

A number of instructors (for example, Chance, Wong & Tintle, 2016) have discovered learning benefits by introducing students to simulation-based inference (bootstrap confidence intervals, randomization tests, etc.). We see these approaches as essential parts of a statistician's repertoire and in our course these concepts are taught by using computation as a means of both motivation and execution, with mathematical details presented in follow-up. For example, the idea of an hypothesis test, including the basic concept that observed effects may be due to chance, is introduced through simulations, followed by the simultaneous teaching of the development and mathematical properties of randomization tests. The classical normal theory tests are then presented as a special case, one where mathematical reasoning can replace intensive computations in practice. When teaching the bootstrap, a similar approach is taken whereby students learn the algorithm, implement it, and explore the behaviour of its output concurrently to learning about its mathematical foundation. This strategy reinforces the notion that computation is as fundamental to statistical practice as mathematics and that statistical ideas and methods depend on both approaches.

Computation as a means for understanding and appreciating theory

We use computation fluidly in the teaching of concepts in theoretical statistics which appear at first glance to be entirely mathematical. Simulations are used to demonstrate probabilistic ideas; interactive apps (developed with the R Shiny package) are used to give students intuition and appreciation for big, difficult concepts; and a computation-focused case study is used to teach the statistical reasoning behind hypothesis testing. For example, the (weak) Law of Large Numbers and the Central Limit Theorem are two of the most important mathematical theorems in an introductory course in statistical theory, and we do not advocate replacing the mathematics with pure computation for these concepts. However, computation is a valuable part of developing students' understanding of these theorems. We use simulations to demonstrate empirically the large sample limiting behaviour of independent sums, with a focus on the difference between what the two theorems are "saying". We found that this use of computation led to a clear and unambiguous delivery of this fundamentally important material.

Computational as a means for applying theory

Finally, we use computation to give context and purpose to theoretical concepts by applying the concepts within larger, ongoing data analyses throughout the course. For example, we began the course with an introduction to data analysis in which the students cleaned and explored a large, messy dataset containing publicly available rental housing quality scores from the city of Toronto. The data did not require substantial manipulation but were perhaps more intimidating than would ordinarily be used in this setting. The size was 3,446 rows and 32 columns, so the dataset could not be meaningfully understood simply by looking at it in a spreadsheet; the data contained a mixture of continuous and categorical variables as well as many missing and questionable values; most variables had unclear scales of measurement, hampering immediate understanding; and there were very few clear, substantial patterns or straightforward conclusions that could be drawn upon initial examination. Having students (with the appropriate support) read in these data and compute basic summary statistics and visualizations represented a significant learning opportunity on its own. Then later in the course, when we taught the concept of maximum likelihood (as an example), we revisited these data and asked the students to formulate some of their earlier summary statistics as maximum likelihood estimators from a suitable statistical model. Connecting the theory with the practice in this manner enhanced the delivery of both concepts, and computation was the engine to make this connection.

Our goal is that the use of computation in statistical work, both theory and practice, becomes innate. To support this, instructional practices we have adopted include the use of participatory live coding and computational work that is literate and reproducible. Participatory live coding (Nederbragt et al., 2020) involves the instructor writing and executing code in real-time during class, narrating their work to make their thinking explicit, including the process of responding to errors. And we are conscious of the importance of setting an example through good practices. Just as we have always been careful to demonstrate rigour and correct form in our mathematical notation, in this new course we are careful to demonstrate literate programming and computational reproducibility. We do this through appropriate use of R Markdown documents (Baumer et al., 2014). Indeed, even the course supplementary materials (available at https://awstringer1.github.io/sta238-book/) are provided as an R Markdown document that students can execute, modify and explore.

CONCLUSION

Many have argued that a modern statistics curriculum needs more computation (e.g., American Statistical Association, 2014; Nolan & Temple Lang, 2010), and more use of real data (GAISE, 2016; Nolan & Speed, 2000), and some have argued for a new conception of statistical theory (Weldon, 2010; Nolan & Speed, 2000; Cobb, 2015). This thinking has also been reflected in recent curriculum guidelines for programs in data science (De Veaux et al., 2017). Much of the impetus for this thinking has been in response to the recent fervour and demand for data scientists, and the role of experts in statistical thinking in this context can be unclear. But if we consider data science to be a broad ecosystem, the role of the statistician becomes clearer. In this ecosystem, statisticians can contribute a perspective on learning from data that is rigorous and scientific, grounded in core ideas such as the power of randomization, the strengths and limitations of statistical models, and the potential pitfalls of confounding and overfitting, while using both mathematical and computational approaches in their work.

Over the past 70 years, statistics has moved from its mathematical and logical centre to a more computational focus (Efron & Hastie, 2016, p. xvii) and our undergraduate curricula should reflect this shift. However, care must be taken to minimize the extraneous cognitive burden (Woodard & Lee, 2020) that is placed on students by the addition of new computational content to our courses. Our belief is that adding in new computational content to courses without fundamentally changing our attitude and approach to the role of computation results in additional extraneous burden that is too much for students and instructors alike. Rather than relegating computation to a boot camp, or restricting our students' exposure to the use of computation as a tool for data manipulation and analysis, we believe that all of our courses, including courses in statistical theory, should treat computational thinking as a natural part of our approach, like we have historically treated mathematical thinking. The resulting modern statisticians will be well-equipped to serve their role as a keystone species in the data science ecosystem.

REFERENCES

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

Baumer, B., 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://dx.doi.org/10.5070/T581020118

Brown, E.N. & Kass, R.E. (2009). What Is Statistics? *The American Statistician, 63*(2), 105-110. DOI: 10.1198/tast.2009.0019

Chance, B., Wong, J. & Tintle, N. (2016). Student Performance in Curricula Centered on Simulation-Based Inference: A Preliminary Report. *Journal of Statistics Education*, *24*(3), 114-126. DOI: 10.1080/10691898.2016.1223529

Cobb, G. (2015). Mere Renovation is Too Little Too Late: We Need to Rethink our Undergraduate Curriculum from the Ground Up. *The American Statistician*, *69*(4), 266-282. DOI: 10.1080/00031305.2015.1093029

De Veaux, R.D. et al. (2017). Curriculum Guidelines for Undergraduate Programs in Data Science. *Annual Review of Statistics and Its Application, 4,* 15-30. https://doi.org/10.1146/annurev-statistics-060116-053930

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

GAISE College Report ASA Revision Committee (2016). "Guidelines for Assessment and Instruction in Statistics Education College Report 2016." *http://www.amstat.org/education/gaise*

Gibbs, A.L. (2018). Building a Foundation in Statistics in the Era of Data Science. *Proceedings of the 10th International Conference on Teaching Statistics, Kyoto, Japan*. http://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_3A2.pdf

Gibbs, A.L. & Damouras, S. (2019). Evolving Statistics Education for a Data Science World. *Proceedings of the 62nd ISI World Statistics Congress 2019, Kuala Lumpur, Special Topic Session Volume 3*. http://isi2019.org/proceeding/2.STS/STS%20VOL%203/index.html#p=48

Gibbs, A. L. & Taback, N. (2021). The Building Blocks of Statistical Education in the Data Science Ecosystem. *Harvard Data Science Review 3*(2). https://doi.org/10.1162/99608f92.8bb28793

Maki, P.L. (2004). Maps and Inventories: Anchoring Efforts to Track Student Learning. *About Campus 9*(4), 2-9. https://doi.org/10.1002/abc.99

Meng, X.-L. (2019). Data science: An artificial ecosystem. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.ba20f892

Nederbragt, A., Harris, R.M., Hill, A.P. & Wilson, G. (2020). Ten quick tips for teaching with participatory live coding. *PLoS Computational Biology*, *16*(9), e1008090. https://doi.org/10.1371/journal.pcbi.1008090

Nolan, D. & Speed, T. (2000). *Stat Labs: Mathematical Statistics Through Applications*. New York: Springer-Verlag.

Nolan, D. and Temple Lang, D. (2010). Computing in the Statistics Curricula. *The American Statistician, 64*(2), 97-107. https://doi.org/10.1198/tast.2010.09132

Weldon, L. (2010). Banishing the Theory-Applications Dichotomy from Statistics Education. *Proceedings of the 8th International Conference on Teaching Statistics*, *Ljubljana*, *Slovenia*. https://iase-web.org/documents/papers/icots8/ICOTS8_4A1_WELDON.pdf?1402524970

Wing, J.M. (2006). Computational Thinking. *Communications of the ACM*, *49*(3), 33-35. https://doi.org/10.1145/1118178.1118215

Woodard, V. & Lee, H. (2020). How Students Use Statistical Computing in Problem Solving. *Journal of Statistics and Data Science Education*, *29*, 145–156. https://doi.org/10.1080/10691898.2020.1847007