Statistical Programming to Principles of Data Science: Rethinking the traditional statistical programming curricula

Andrew Hoegh*

Abstract

Since the arrival of computers and statistical programming languages in the early 1970s, there has been tension between mathematical statistics and computing. John Tukey, perhaps optimistically, foresaw this as a 'peaceful collision'; however, in hindsight this might have been more of a missed opportunity as the reluctance to 'moderate the romance with mathematics' and fully embrace computing has, arguably, lead to the development of many fields that run parallel or partially include statistics. This article will take a historical look at the interaction between statistics and computing while highlighting select moments where a different path would have lead to a different field of modern statistics. Furthermore, the article will examine the current state of computing in statistics and make recommendations for the future.

Key Words: Computing, Data Science,statistical programming, curriculum, statistics education

1. Introduction

While *Data Scientist* is the popular choice for the best job in America (Glassdoor 2018), there is still open debate about what exactly a data scientist is. This article is not concerned with defining data science; rather, it highlights the historical arc of computing and statistical programming focusing on missed opportunities that arguably lead to the creation of alternate computing-oriented data analysis fields (viz. data mining, machine learning, data science, artificial intelligence) and makes recommendation about what statistical programming courses should focus on in this modern, data-rich era.

The roots for much of the current statistical curriculum lie in a pre-computer era with heavy emphasis on mathematics. It was not until the mid 1960's that discussions about statistical computing were orchestrated (Chambers 1999). In the 1960s John Tukey stated, perhaps optimistically, that the "peaceful collision of statistics and computing" was eminent. In hindsight, a peaceful collision was not the best description of the complicated relationship between statistics and computing. Rather computing as a statistical tool took a backseat to mathematical theory, as Friedman later stated in 2001 we "Must moderate our romance with mathematics."

Moving forward from the 1960's to 2000, computing evolved (and continues to evolve) more rapidly than the corresponding statistics curriculum. As Peng states "a textbook for a typical Ph.D. program in statistics, I believe it would look much like the textbooks used by Cleveland, Chambers, and Tukey in their own studies. In fact, it might even be the same textbook!" (Peng 2017). In 2001, Friedman (2001) had some prescient ideas about the role of computing in statistics.

> We have to make peace with computing. It is here to stay; that's where the data is. Computing has been one of the most glaring omissions in the set of tools that have so far defined statistics. Had we incorporated computing methodology from its inception as a fundamental statistical tool many of the other data related fields would not have needed to exist.

*Department of Mathematical Sciences, Montana State University, Bozeman, MT

While Friedman talks about data mining in his 2001 work, as prophesied, data science programs developed, many which do not have roots in a statistics department. Furthermore, Friedman makes recommendations about education and curriculum.

> If computing is to become one of our fundamental research tools we will have to teach, or be sure that our students learn, the relevant computer science topics. If we are to compete with other data related field in the academic and commercial marketplace, some of our basic paradigms will have to be modified. We may have to moderate our romance with mathematics. Mathematics (like computing) is a tool, a very powerful one to be sure, but not the only one that can be used to validate statistical methodology.

Around the same time, Cleveland (2001) wrote an article arguing that university curriculum needed to evolve to place more emphasis and resources to statistical computing.

In the last fifteen to twenty years, there have been some changes in statistical curriculum (Franklin et al. 2005, Cobb 2007, 2015, Carver et al. 2016). Despite these advances, the need for computing continues to accelerate, along with the growth of data science - as foreseen by Friedman. In the last few years, calls have been made for a renewed focus on programming in the statistics curriculum (Nolan & Temple Lang 2010, McNamara & Horton 2017, Stander & Dalla Valle 2017, Hardin et al. 2015, Cetinkaya-Rundel & Rundel 2017).

The best way to incorporate programming principles into statistical curriculum and specifically what skills are essential for a modern statistician are still open questions. For the most effective learning, we believe programming should be interwoven in each class in the statistics curriculum. While this article focuses on the implementation of a statistical programming course, the basic ideas in Section 4 also apply to the curriculum as a whole. Historically many statistics programs included a course that covered statistically programming with respect to SAS, SPSS, and more recently R. Traditionally in this class examples and assignments from homework assume that the data is contained in a tidy rectangle and focus on implementing statistical methods without addressing topics like data cleaning and wrangling that are essential in most realistic data analysis scenarios.

Meanwhile undergraduate and master's programs in data science have become popular in recent years. While computing is not often relegated to a single course in the program, we will discuss important elements of the computing curriculum across these data science programs: data storage, data manipulation, programming skills. The article by Hardin et al. (2015) provides a summary of courses and curriculum across seven institutions that are including *data science ideas* in statistics curriculum. A general theme in these courses is to teach the necessary programming skills to conduct an entire data analysis. There are also another set of data science programs that have evolved largely out of computer science programs with a slightly different emphases, in particular, more emphasis is placed on algorithms and big data applications like hadoop. We propose a statistical computing course that takes the best of the data science curriculum while also maintaining principles relevant for statistical programming and analysis.

Section 2 provides a review of statistical programming courses taught at the undergraduate and graduate levels. Section 3 discusses important computational concepts contained in data science programs. Section 4 proposes a hybrid approach that contains the essential elements for modern statistical computing and Section 5 concludes with closing thoughts.

2. Review of Statistical Programming Courses

Not a lot of scholarly works are available that document the existence and content of statistical programming courses, but a summary of this material is extracted from university websites and presented for select programs in Table XX. There is more written about computing courses Gentle (2004), Monahan (2004); however, we will focus more on programming which we differentiate from the algorithm-based computing courses.

Table 1: Summary of Stat Programming Material

Dummy for additivity Label	Dummy for dominance Index i	Genotypic value (η)	effect α (x)	effect δ (z)
qq	1	$\mu + 2\alpha$	2	0
Qq	2	$\mu + \alpha + \delta$	1	1
QQ	3	μ	0	0

Some of the key ideas presented in these courses are... XXX. While Lange (2004) summarizes an overview of a course developed for computing and optimization targeting graduate students, a key quote from this article is "There is no doubt that our students need to be more computationally savvy. But the traditional course most departments offer is seriously in need of updating, reformation, and expansion." We whole heartedly agree with this statement and seek to outline principles for the updated, reformed, and expanded statistical programming course.

3. Review of Principles of Data Science Courses

From a statistical perspective, in our mind there are two types of data science programs: those created and oriented with a statistical emphasis and those created by 'others'. Undoubtedly, this is two broad of a generalization, but it allows us to make distinctions between those with input from statistics and the greater statistical curricula from those that arise exclusively elsewhere, such as computer science or business departments. The goal of this paper is to take the best ideas from data science programming courses to meld with a traditional statistical programming to create a modern statistical programming framework for statisticians.

From a more statistics oriented viewpoint, Kross et al. (2017) focused on the democratization of data science education through the lens of Johns Hopkins Data Science programs. The MOOC through JHU was orchestrated by three members of the JHU Biostatistics Department: Brian Caffo, Roger Peng, and Jeff Leek and they focused on three courses on Coursera: Mathematical Biostatistics Boot Camp, R Programming, and Data Analysis. As a follow up a series of 9 data science oriented courses were established: Data scientists toolbox, R Programming, Getting and Cleaning Data, Exploratory Data Analysis, Reproducible Research, Statistical Inference, Regression Models, Practical Machine Learning, and Developing Data Products.

Other data science programs arising from computer science departments have a different emphasis...

Table 2: Summary of Data Science Keywords

Dummy for additivity	Dummy for dominance			
Label	Index i	Genotypic value (η)	effect α (x)	effect δ (z)
qq	1	$\mu + 2\alpha$	2	0
Qq	2	$\mu + \alpha + \delta$	1	1
QQ	3	μ	0	0

4. A Hybrid Approach: Statistical Programming for the Modern Statistician

We don't strongly advocate for the use of a single programming language, but we'd envision this course being taught primarily through `R`. There is value in being bilingual, from a programming perspective, so small modules of `SAS`, `Matlab`, and/or `python` could be taught as well.

Donoho (2017) and references contained within provide a nice overview about what data sciences is, and is not, and also provide an overview of ideas on how to teach data science. In particular, Donoho proposes dividing "greater data science" into six categories: 1.) Data gathering, preparation, and exploration; 2.) Data representation and transformation; 3.) Computing with data; 4.) Data modeling; 5.) Data visualization and presentation; and 6.) Science about data science. Here we are concerned with teaching modern statistics programming, not data science, but there is plenty of overlap between our focus areas and Donoho's categories. We use an approach similar to Hofmann & VanderPlas (2017) and focus on the necessary steps of a data analysis. Specifically, our goal is that students would be well equipped to handle the entire data analysis process of an extremely messy data set: from data acquisition to summarizing the data set, with additional focus on reproduciblity and version control.

Hardin et al. (2015) summarize attempts to guarantee students are able to "think with data", which focus on the statistical and computational issues for non-textbook data and present examples curricula and courses at six institutions. One emphasis in Nolan & Temple Lang (2010) is for students to *Compute with data in the practice of statistics*, in other words, "statistical computing should be taught in the context of statistical practice, just like statistical methods." Zheng (2017) focuses on the idea of *statistical thinking* through conducting data analysis on data from the "real jungle" that are messy, with vague scientific problems, and lack ground truth. We agree with all of these approaches and to summarize, the central theme is that programming, in addition to statistical pedagogy as a whole, should be taught from a data-focused paradigm.

The next four categories are designed to map the data analysis process, and the remaining two: reproducibility and version control are important elements of the data analysis process as well.

4.1 Data Gathering and Data Storage / Access

In a traditional, experimental design based setting, data may come in a nice, tidy rectangle. However, in many modern data analyses and data visualization settings, data must be acquired from sources online, many of which are not as easy as downloading a spreadsheet file. The ability to scrape data from various online data structures. In other settings, data may be contained in a database structure and require SQL to extract data from the database. Students need to be able to extract information from both of these data structures, and often

store data in alternative data structure forms too.

4.2 Data Cleaning / Wrangling

Once data has been obtained, there will almost always been the need to clean and transform the data into the desired, usable form. (McNamara & Horton 2017) details the importance of data wrangling and provides examples in the categorical data case. Many of the R packages in the `tidyverse` world Wickham (2017), are designed to make data wrangling easier.

4.3 Data Visualization

With data cleaned and usable, the next step would another of Tukey's contributions to the field of statistics, exploratory data analysis. In particular, `ggplot2` (Wickham 2016) and ggplot2, R Shiny should be tools for data exploration and also for creating interactive graphics and publication quality figures. Ultimately students should also be data storytellers that can distill datasets and present them in a fashion that are informative, but also maintain fidelity to the data itself. (Stander & Dalla Valle 2017)

4.4 Statistical Modeling

There still should be an element of running the actual statistical methods in R, but in this context we are focused more on the "button pushing" programming and the theory is covered elsewhere in courses. As part of the "statistical thinking" or data-focused paradigm, emphasis should also be placed on the qualitative elements of the dataset, both prior to and after applying statistical methods. Leman et al. (2015) uses the moniker, QQQ, to detail this qualitative-quantitative-qualitative process. Like statistical thinking, in the best case, the QQQ idea is embedded in the statistical curriculum as a whole.

4.5 Reproducibility

R Markdown is an absolute must for creating and sharing reproducible documents. (Stander & Dalla Valle 2017) highlights a course that focuses on using the whole R studio interface, including R Markdown. We believe that R Markdown should be incorporated in conjunction with R from the first time students program as in Baumer et al. (2014). Perhaps, it should be the default way that students learn to program in R itself.

4.6 Version Control

Just like research and business projects after graduation, many classes are incorporating group work into data analysis projects. With this, version control becomes essential and `github` provides nice interfaces with R and other statistical languages.

5. Concluding Thoughts

Ultimately had that "peaceful collision," that Tukey mentioned, actually occured, the field of statistics the associated data analysis fields would have a very different look. What we now consider, data mining, data science, and/or artificial intelligence (feel free to add other fields here too...) might be an accepted part of statitics, rather than those separate fields that Friedman saw coming back in 2001. Nevertheless, as we can't go back in time, the question is, *How should we proceed today*? One challenge with the modern take on

statistical programming, particular when thinking about research and development, is the antiquated value placed on mathematical theory over computing tools as detailed in Bryan & Wickham (2017). In 1962, Tukey wrote a voluminous overview on the future of data analysis (Tukey 1962) in which he professed his "central issue" was in data analysis, which he defined as,

> among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

By this definition, Tukey would support equal value for the computing tools, with mathematical theory, for academic researchers. We feel that it is essential to modernize statistics and equally value computational and mathematical contributions and, furthermore, we must make sure to teach students, both undergraduate and graduate students, to be competent programmers. More specifically, our students must be able to carry out all of the aspects of the data analysis cycle and not just fit a model and report a p-value.

References

Baumer, B., Cetinkaya-Rundel, M., Bray, A., Loi, L. & Horton, N. J. (2014), 'R markdown: Integrating a reproducible analysis tool into introductory statistics', *arXiv preprint arXiv:1402.1894* .

Bryan, J. & Wickham, H. (2017), 'Data science: A three ring circus or a big tent?', *arXiv preprint arXiv:1712.07349* .

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., Rossman, A., Rowell, G. H., Velleman, P. & Witmer, J. (2016), 'Guidelines for assessment and instruction in statistics education (GAISE) college report 2016', *Alexandria, VA: American Statistical Association.[Online: www. amstat. org/education/gaise]* .

Cetinkaya-Rundel, M. & Rundel, C. W. (2017), 'Infrastructure and tools for teaching computing throughout the statistical curriculum', (e3181v1).
URL: *https://peerj.com/preprints/3181*

Chambers, J. (1999), 'Computing with Data: Concepts and Challenges', *The American Statistician* **53**(1), 73–84.
URL: *http://www.tandfonline.com/doi/abs/10.1080/00031305.1999.10474434*

Cleveland, W. S. (2001), 'Data science: an action plan for expanding the technical areas of the field of statistics', *International statistical review* **69**(1), 21–26.

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.
URL: *http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2015.1093029*

Cobb, G. W. (2007), 'The Introductory Statistics Course: A Ptolemaic Curriculum? - eScholarship', *Technology Innovations in Statistics Education* **1**(1).
URL: *https://escholarship.org/uc/item/6hb3k0nz*

Donoho, D. (2017), '50 years of data science', *Journal of Computational and Graphical Statistics* **26**(4), 745–766.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M. & Scheaffer, R. (2005), 'Guidelines for assessment and instruction in statistics education (GAISE) report', *Alexandria, VA: American Statistical Association* .

Friedman, J. H. (2001), 'The role of statistics in the data revolution?', *International Statistical Review* **69**(1), 5–10.

Gentle, J. E. (2004), 'Courses in statistical computing and computational statistics', *The American Statistician* **58**(1), 2–5.

Glassdoor (2018), 'Best jobs in america'.
URL: *https://www.glassdoor.com/List/Best-Jobs-in-America.htm*

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**(4), 343–353.
URL: *http://amstat.tandfonline.com/doi/full/10.1080/00031305.2015.1077729*

Hofmann, H. & VanderPlas, S. (2017), 'All of this has happened before. all of this will happen again: Data science', *Journal of Computational and Graphical Statistics* **26**(4), 775–778.

Kross, S., Peng, R. D., Caffo, B. S., Gooding, I. & Leek, J. T. (2017), The democratization of data science education, Technical Report e3195v1, PeerJ Inc. DOI: 10.7287/peerj.preprints.3195v1.
URL: *https://peerj.com/preprints/3195*

Lange, K. (2004), 'Computational Statistics and Optimization Theory at UCLA', *The American Statistician* **58**(1), 9–11.
URL: *http://dx.doi.org/10.1198/0003130042890*

Leman, S., House, L. & Hoegh, A. (2015), 'Developing a new interdisciplinary computational analytics undergraduate program: A qualitative-quantitative-qualitative approach', *The American Statistician* **69**(4), 397–408.

McNamara, A. & Horton, N. J. (2017), Wrangling categorical data in R, Technical Report e3163v2, PeerJ Inc. DOI: 10.7287/peerj.preprints.3163v2.
URL: *https://peerj.com/preprints/3163*

Monahan, J. (2004), 'Teaching Statistical Computing at North Carolina State University', *The American Statistician* **58**(1), 6–8.
URL: *http://dx.doi.org/10.1198/0003130042881*

Nolan, D. & Temple Lang, D. (2010), 'Computing in the statistics curricula', *The American Statistician* **64**(2), 97–107.

Peng, R. D. (2017), 'Comment on 50 years of data science', *Journal of Computational and Graphical Statistics* **26**(4), 767–767.

Stander, J. & Dalla Valle, L. (2017), 'On Enthusing Students About Big Data and Social Media Visualization and Analysis Using R, RStudio, and RMarkdown', *Journal of Statistics Education* **25**(2), 60–67.

Tukey, J. W. (1962), 'The Future of Data Analysis', *The Annals of Mathematical Statistics* **33**(1), 1–67.
URL: *http://www.jstor.org/stable/2237638*

Wickham, H. (2016), *ggplot2: elegant graphics for data analysis*, Springer.

Wickham, H. (2017), 'Tidyverse', *R package version* **1**(1).

Zheng, T. (2017), 'Teaching data science in a statistical curriculum: Can we teach more by teaching less?', *Journal of Computational and Graphical Statistics* **26**(4), 772–774.