Honors Thesis Proposal

Audrey Bertin
Statistical and Data Sciences

April 10, 2020

Abstract

For my honors thesis in the department of Statistical and Data Sciences, I would like to focus on the topic of reproducibility in data science. Reproducibility is critical to the advancement of knowledge,yet there are not very many effective tools to address it and it does not appear to be widely taught at US academic institutions. In this honors project, I will discuss several main areas - 1) The need for widespread knowledge about reproducibility. 2) An R package I am working on developing to address this issue, along with new ideas and ways of coding I am learning through the process. 3) How my work is benefitting, and will continue to benefit introductory data science students and other users. 4) Further recommendations on how best to incorporate reproducibility into data science education.

Keywords: reproducibility, statistical software, workflow, collaboration, teaching, curriculum recommendations

1 The Broad Issue: Reproducibility in Data Science

As research is becoming increasingly data-driven, and because knowledge can be shared worldwide so rapidly, reproducibility is critical to the advancement of scientific knowledge.

Data-based research cannot be fully *reproducible* unless the requisite code and data files produce identical results when run by another analyst. When researchers provide the code and data used for their work in a well-organized and reproducible format, readers are more easily able to determine the veracity of any findings by following the steps from raw data to conclusions.

The creators of reproducible research can more easily receive more specific feedback (including bug fixes) on their work. Moreover, others interested in the research topic can use the code to apply the methods and ideas used in one project to their own work with minimal effort.

However, while the necessity of reproducibility is clear, there are significant behavioral and technical challenges that impede its widespread implementation, and no clear consensus on standards of what constitutes reproducibility in published research. Not only are the *components* of reproducible research up for discussion (e.g., need the software be open source?), but the corresponding *recommendations* for ensuring reproducibility also vary.

Much of the discussion around reproducibility is also generalized—it is written to be applicable to users working with a variety of statistical software programs. Since all statistical software programs operate differently, generalized recommendations on reproducibility are often shallow and unspecific. While they provide useful guidelines, they can often be difficult to implement, particularly to new analysts who are unsure how to apply such recommendations within the software programs they are using. Thus, reproducibility recommendations tailored to specific software programs are more likely to be adopted.

One of the most effective ways to work on improving reproducibility in the data science community is to focus on a specific piece of software used by data analysts. In the program in Statistical and Data Sciences at Smith College, we use the program R. Since R is freely available online and very popular in the data science community, it is a good candidate for research focus. As a result, it is the primarily language used to teach data science at many academic institions around the US.

Unfortunately, there are not very many academic papers or software packages discussing reproducibility in R, and much of the work that does exist is not ideal for students just beginning to learn data science. Much of this work is narrowly tailored, with each package effectively addressising a small component of reproducibility—file structure, modularization of code, version control, etc. Many existing software packages available to R users succeed at their area of focus, but at a cost. They are often difficult to learn and operate, providing a barrier to entry for less-experienced data analysts.

Due to the above reasons, reproducibility is often left behind and not prioritized to the degree it should be.

2 My Focus

In order to address this issue, I have been working to help develop an R package that can provide important information about the reproducibility of a data analysis project and share recommendations for how to improve it, all in a few short and simple lines of code. The package is written in such a way as to be accessible to introductory data science students, and can easily be brought into the classroom as a tool for integrating the concept of reproducibility into data science education.

The package is still in the development process, and I will be spending a significant portion of my thesis time improving it and expanding its functionality.

One major focus will be delving into adding `make`-like functionality that can analyze an R project structure and files and use this information to generate a Makefile. This Makefile would have information about target files and their prerequisites and would assist with making sure that re-running an analysis is done as quickly as possible by ensuring that only the necessary code and files that have been updated are run when rebuilding and re-running code.

In my thesis, I would like to focus much of my discussion on this package while covering the following four general topics:

1) The need for widespread knowledge about reproducibility.

In this background section, I would like to delve into evidence that widespread repro-

ducibility is lacking across the sciences. Several reports from different academic fields dive into this idea and share the state of reproducibility in their areas of study.

I will also cover the benefits of reproducible research, many of which I mentioned in brief in the previous section. I will also delve into the importance of teaching and working to integrate reproducibility early on in the education of students studying data science.

2) A discussion of my R package and the ideas I have learned through the package development process.

In this section, I will describe the purpose of the R package and how it works (covering several of the major functions).

The package has a very unique structure compared to many other available packages. Much of the functionality of the package operates under the hood and in the background. Unlike most packages, it does not have functions that are used when coding, but rather functions that analyze users' coding and project building practices. Many of the functions utilize features of R programming not covered in any courses at Smith College, which required a significant amount of outside research and troubleshooting to put into use.

I would like to highlight several of the unique functionalities of the package that taught me new ways of programming in R. These include:

- File path and directory manipulation
- Function shimming/masking
- Environment variables
- Utilizing hidden files
- Leveraging the dots (...) for additional functionality
- Understanding dependencies, keeping track of file update histories, and using Makefiles

As package development continues throughout the course of this honors project, the list above will likely expand as new issues are tackled.

3) How my work has benefitted and will continue to benefit introductory data science students and other users.

In this section, I will discuss my work as a statistics tutor, discussing some of the challenges that seem to have regularly and demonstrating how my package works to address these issues.

I will also include some preliminary reviews and discussion of how functional the software is to users who have never seen it before.

Looking back at the bigger picture, I will then talk about the software's wider applications to reproducibility in general (not just for introductory students) and its unique place in the world of the tools available to data scientists.

4) Further recommendations on how best to incorporate reproducibility into data science curricula, including a look at approaches currently being pioneered at some universities.

In this section, I will work on defining the ideal way to integrate reproducibility into the data science curriculum. I will refer to the guidelines on teaching data science from the 2017 Annual Review of Statistics and Its Applications, as well as to a variety of other sources discussing easy methods for implementing different aspects of reproducibility.

I will also look into some of the new reproducibility curriculum additions that are beginning to appear at colleges across the country, including one course offered at the University of Washington that has reproducibility as its sole focus, as well as recommendations on how to teach reproducibility as shared by several leaders in the field.

I will focus on how my work fits into the discussion of teaching reproducibility and how it might be integrated into data science education in a way that combines effectively with other available tools such as RMarkdown and GitHub.

3 Relevant Preparation

I have spent the last year studying reproducibility as part of research I have been conducting with Prof. Ben Baumer in the Statistical and Data Sciences department. That research

has focused on creating a publicly available piece of software that can be downloaded by people all around the world. In the process of sharing that work, I have already begun the process of researching the need for reproducibility and available resources, so I already have a relatively well developed background of knowledge on reproducibility coming into this honors project. I have become familiar with several reproducibility experts through my research, whose work I can delve into further to gather insights on effective ways to develop the idea data science curriculum.

As a statistics tutor, I also have a good sense of which types of tools and which aspects of the curriculum work best for students of all different levels, and can use this knowledge to help inform the recommendations in this project. I also have made connections with data scientists and students of all different levels, who I can reach out to in order to build a diverse pool of testers for my package.

Finally, I have completed a significant number of data analysis and research projects in my time at Smith. I am very familiar with strategies for finding/summarizing sources and collecting/analyzing data, as well as with sharing my findings. I have had experience with writing lengthy (35+ page) papers before, and have developed excellent time management skills through that process that will allow me to successfully complete this project.

Thank you for taking the time to consider my honors thesis proposal. I have had a wonderful time partipating in my reproducibility research over the last year, and look forward to this opportunity to expand on it further in a meaningful way.

Audrey Bertin '21