Introduction

Steven Hobbs

2022-08-06

Purpose and Scope

Data Essentials With R is intended to support new college students who may soon be working with data through current or future coursework or academic research. I assume no prior knowledge in any software language, statistics, calculus, or even mathematics beyond basic algebra. The primary goal of this resource is to build a versatile skill set for working with data, one that is motivated by biological research in humans (medical research and physiology), but which is readily transferable to any field

Why learn about data before learning statistics?

Learning about data, a computer language and statistics simultaneously is hard, very hard. In the biological sciences, and most likely other disciplines, many sharp and highly motivated students struggle in their first college statistics course, in part because learning the nuances of data and the technology used to tackle statistics are heavy cognitive burdens *in addition* to learning statistics. Many students are so intimidated by college statistics that they put off taking a course until their 2nd, 3rd or even 4th year, by which time they have already been struggling with data and data analysis in other courses or in research labs.

So, why do we pack so much into college statistics courses? Unfortunately, it is virtually impossible to teach a modern statistics course without using data and a computer language like R or Python. Furthermore, statistics courses that seek to give their students a usable skill set must also contend with the "Thankless 80%" that is importing data, recoding variables, dealing with missing values, restructuring data, creating new variables using logical statements and control loops and visualizing data.

Data Essentials with R seeks to tame the "Thankless 80%" of data analysis, essentially everything that happens before "statistics." Students that finish all lessons will have a solid understanding of R, R Studio and R Markdown as well as strong data literacy and data wrangling skills. Even modest gains on these fronts will set students up for success in their first, next or current statistics course and any scholarly pursuit that involves data.

Why R

R is a computer language built for statistics. R is powerful, free and open source. R is arguably the dominant language in academia for data analysis and statistics. In the professional statistics-leaning data science world, Python and R are the two most common languages and many professionals use both.

Why R Markdown

R Markdown is a versatile file format that allows authoring and analyzing data in the same document. From an educators perspective, R Markdown is a tool we can use to provide explanations and interpretations intermixed with executable code. Similarly, students can use R markdown to write and annotate code and provide interpretations and explanations.

Why R Studio

RStudio is a multi-language Integrative Development Environment (IDE) for data analysis. RStudio is loaded with features for writing and executing code, visualizing graphs and output, and managing data. RStudio is accessible to beginners, but is also the industry standard for data analysis with R.

Reading Data Essentials with R

I recommend that readers work with two formats on their computer:

1) The rendered pdf or html
2) the individual .Rmd files contained within an R project.

The pdf and html file types present the reader-friendly finished product, while the .Rmd files allow the reader to run code on their own computer while learning the RStudio application and the R Markdown file type.