Applied Biostats

Yaniv Brandvain

2024-12-24

Table of contents

Preface **3**
Learning in this era . . . 3
I love teaching this course . . . 4
Course philosophy / goals . . . 4
"By the end of this course… . . . 5
`R`, `RStudio`, and the `tidyverse` . . . 6
Installing `R` . . . 7
0.0.1 Tab 1: PC installation . . . 7
0.0.2 Tab 2: Mac installation . . . 7
0.0.3 Tab 3: Linux installation . . . 8
Installing `RStudio` . . . 8
What is this 'book' and how will we use it? . . . 9
How will this term work / look? . . . 10
The Use of Large Language Models . . . 10
Acknowledgements . . . 12
Students . . . 12
Teaching Assistants (TAs) . . . 12
Collaborators . . . 13
Teaching Colleagues . . . 13
Unknowing contributors . . . 13

Preface

In the summer of 2020, the world was on fire – COVID was raging, we – especially in Minnesota – were processing the murder of George Floyd and the subsequent uprising etc, the future was unclear. At that point teaching was likely to be entirely online, and I decided to write a digital book for my course (see the 1st edition and 2nd edition). I didn't really know what I was doing or what my vision was (and to some extent I still do not). There were hiccups: some strangeness in rendering etc, typos, last minute updates, writing at 2am etc etc, but on the whole there were numerous advantages compared to a traditional textbook. I lay these out here:

- My class presentation and the textbook presentation almost always agreed.
- As I was writing and updating as I went the book could be rapidly updated / changed to reflect student needs / interests / timelines / current events etc.
- I could integrate practice problems / youtube links / and even additional readings pretty easily.
- It was free for students.

I think all of these benefits were great, and helped a lot, so I am doing it again. This time I'm bringing in my research collaboration with Dave Moller on speciaiton in *Clarkia xantiana* as a them through this book.

Learning in this era

I know you're dealing with a lot. Every year students are dealing with a lot – from jobs, to supporting family, to the everyday

of being in college and living life etc…. Yet, we are all trying to make the most of life in this era. We want to teach, learn, and grow.

What is more, I believe this content is more important now than it has ever been, statistics is obsessed with the critical evaluation of claims in the face of data, and is therefore particularly useful in uncertain times. Given this focus, and given that you all have different energies, motivations and backgrounds, I am restructuring this course slightly from previous years. The biggest change is a continued de-emphasis on math and programming – that doesn't mean I'm eliminating these features, but rather that I am streamlining the required math and programming to what I believe are the essentials. For those who want more mathematical and/or computational details (either because you want to push yourself or you need this to make sense of things), I am including a bunch of optional content and support. I am also wrestling with the impact of LLMs in our education (more below).

I love teaching this course

The content is very important to me. I also care deeply about you. I want to make sure you get all you can / all you need from this course, while recognizing the many challenges we are all facing. One tangible thing I leave you with is this book, which I hope you find useful as you go on in your life. Another thing I leave you with is my concern for your well-being and understanding – please contact me with any suggestions about the pace / content / you of this course and/or any life updates which may change how and when you can complete the work.

Course philosophy / goals

My motivating goal for this course is to empower you to produce, present, and critically evaluate statistical evidence — especially as applied to biological topics. You should know that stats models are only models and that models are imperfect abstractions of reality. You should be able to think about how a

biological question could be formulated as a statistical question, present graphs which show how data speak to this question, be aware of any shortcomings of that model, and how statistical analysis of a data set can be brought back into our biological discussion.

"By the end of this course...

Students should be statistical thinkers.

Students will recognize that data are comprised of observations that partially reflect chance sampling, & that a major goal of statistics is to incorporate this idea of chance into our interpretation of observations. Thinking this way can be challenging because it is a fundamentally new way to think about the world. Once this is mastered, much of the material follows naturally. Until then, it's more confusing.

Students should think about probability quantitatively.

That chance influences observations is CRITICAL to statistics (see above). Quantitatively translating these probabilities into distributions and associated statistical tests allows for mastery of the topic.

Students should recognize how bias can influence our results.

Not only are results influenced by chance, but factors outside of our focus can also drive results. Identifying subtle biases and non-independence is key to conducting and interpreting statistics.

Students should become familiar with standard statistical tools / approaches and when to use them.

Recognize how bias can influence our results. What is the difference between Bayesian and frequentist thinking? How can

data be visualized effectively? What is the difference between statistical and real-world significance? How do we responsibly present/ interpret statistical results? We will grapple with & answer these questions over the term.

Students should have familiarity with foundational statistical values and concepts.

Students will gain an intuitive feel for the meaning of stats words like *variance, standard error, p-value, t-statistic, and F-statistic*, and will be able to read and interpret graphs, and how to translate linear models into sentences.

Students should be able to conduct the entire process of data analysis in R.

Students will be able to utilize the statistical language, R, to summarize, analyze, and combine data to make appropriate visualizations and to conduct appropriate statistical tests.

`R`, `RStudio`, and the `tidyverse`

We will be using R (**version 4.4.0 or above.**) in this course, in the RStudio environment. My goal is to have you empowered to make figures, run analyses, and be well positioned for future work in R, with as much fun and as little pain as possible. RStudio is an environment and the tidyverse is a set of R packages that makes R's powers more accessible without the need to learn a bunch of computer programming.

Some of you might have experience with R and some may not. Some of this experience might be in tidyverse or not. There will be ups and downs — the frustration of not understanding and/or it not working and the joy of small successes. Remember to be patient, forgiving and kind to yourself, your peers, and me. Ask for help from the internet, your friends, TAs, and Yaniv.

Figure 1: This image comes with permissio
Horst, who makes tremendous aRt. I
her work, she would appreciate your s
for Black Lives

Installing R

You can download these onto your computer (Make sure the R is version 4.4.0 or above).

1. Download/update R from here.

Before you can use R you must download and install it*. So, to get started, **download R from CRAN**, and follow the associated installation instructions (see below for detailed instructions for your system).

* **This is not strictly true.** You can use R online via posit cloud. This is a "freemium" service and the free plan is unlikely to meet your needs.

0.0.1 Tab 1: PC installation

If you want a walk through, see Roger Peng's tutorial on installing R on a PC].

"To install R on Windows, click the *Download R for Windows link*. Then click the *base* link. Next, click the first link at the top of the new page. This link should say something like *Download R 4.4.2 for Windows* except the 4.4.2 will be replaced by the most current version of R. The link downloads an installer program, which installs the most up-to-date version of R for Windows. Run this program and step through the installation wizard that appears. The wizard will install R into your program files folders and place a shortcut in your Start menu. Note that you'll need to have all of the appropriate administration privileges to install new software on your machine."

- From Appendix A of *Hands-On Programming wWith R* – Grolemund (2014).

0.0.2 Tab 2: Mac installation

If you want a walk through, see Roger Peng's tutorial on installing R on a mac].

"To install R on a Mac, click the *Download R for macOS* link. Next, click on the [newest package link compatible with your

computer]. An installer will download to guide you through the installation process, which is very easy. The installer lets you customize your installation, but the defaults will be suitable for most users. I've never found a reason to change them. If your computer requires a password before installing new progams, you'll need it here."

- From Appendix A of *Hands-On Programming With R* – Grolemund (2014).

0.0.3 Tab 3: Linux installation

R comes preinstalled on many Linux systems, but you'll want the newest version of R if yours is out of date. The CRAN website provides files to build R from source on [Debian], Redhat, SUSE, and Ubuntu systems under the link "Download R for Linux." Click the link and then follow the directory trail to the version of Linux you wish to install on. The exact installation procedure will vary depending on the Linux system you use. CRAN guides the process by grouping each set of source files with documentation or README files that explain how to install on your system.

- From Appendix A of *Hands-On Programming With R* – Grolemund (2014).

Installing `RStudio`

2. Next download/update RStudio from here.

Alternatively you can simply join the course via RStudioCloud. This could be desirable if you do not want to or have trouble doing this.

What is this 'book' and how will we use it?

A fanstastic feature of this book is that does not stand alone.

I hope that this book provides clear and useful background for the course, and I advise you to regularly go through each book 'chapter' for the relevant week. **Be sure you get familiar with the content BEFORE class.**

Note that this is not the entirety of the course content, and is not an original piece of my own effort – in addition from lifting from a few other course online (with attribution), I also make heavy use of these texts:

- *The Analysis of Biological Data Third Edition (Whitlock and Schluter 2020):* I taught with this book for years. It fantastic and shaped how I think about teaching Biostats. It has many useful resources available online. The writing is great, as are the examples. Most of my material originates here (although I occasionally do things a bit differently). Buy the latest edition.
- *Calling Bullshit (Bergstrom and West 2020):* This book is not technical, but points to the big picture concerns of statisticians. It is very practical and well written. I will occasionally assign readings from this book, and/or point you to videos on their website. All readings will be made available for you, but you might want to buy a physical copy.
- *Fundamentals of Data Visualization (Wilke 2019):* This book is free online, and is very helpful for thinking about graphing data. In my view, graphing is among the most important skills in statistical reasoning, so I reference it regularly.
- *R for Data Science (Grolemund and Wickham 2018):* This book is free online, and is very helpful for doing the sorts of things we do in R regularly. This is a great resource.
- *The storytelling with data podcast* is a fantastic data viz podcast. Be sure to check out Cole Nussbaumer Knaflic's books too!

I will introduce other resources as we go.

How will this term work / look?

- Prep for 'class'. This class is flipped with asynchronous content delivery and synchronous meetings.
- Be sure to look over the assigned readings and/or videos, and complete the short low-stakes homework BEFORE each course.
- During class time, I will address questions make announcements, and get you started on in-class work. The TA & I will bounce around your breakout rooms to provide help and check-in. If you cannot make the class, you could do this on your own time without help, but we do not recommend this as a class strategy.
- The help of your classmates and the environment they create is one of the best parts of this class. Help each other.
- In addition to low stakes work before and in class, there will be a few more intense assignments, some collaborative projects, some in class exams, and a summative project as the term ends.

The Use of Large Language Models

We are in the early days of a truly disruptive technology. Large Language Models (LLMs) like ChatGPT and Claude are transforming how we work and learn. While the impact of these tools on future employment, expertise, and citizenry is yet to be settled, it seems clear that no one will hire you to copy and paste AI-generated output. At the same time, no one will hire you to ignore this technology. Success lies in learning how to critically evaluate and work with LLMs—to validate their output,

improve your own understanding, and create high-quality results. Subject-level expertise, in conjunction with strong skills in working with AI, will be essential for the foreseeable future.

I want you to think about this class like learning to play an instrument. You're here to practice, make mistakes, and build mastery over time. Sure, you could ask an LLM to "play the piece" for you by doing your homework. But if you do that, you're not the one learning. Instead, use the LLM as a tutor and "practice partner"—a tool to get feedback, refine your technique, and expand what you're capable of doing. Doing so will allow you to get truly good not just at stats but at using the best tools available to do even better stats.

In fact, over the past year, it has been quite easy for me to differentiate between:

- Truly poor work from people who did not care and did not use AI,
- Minimal ChatGPT output,
- Strong students who used nearly no AI assistance, and
- Strong students who made their work and understanding better by working with AI.

I hope more of you join the final category!

To aid you in achieving this goal, there will be plenty of opportunities for in-class, computer-free efforts to show your mastery of the subject. I will also provide guidance on individual assignments about the appropriate use of AI to help maximize the impact of the assignment on your learning.

Acknowledgements

Students

First and foremost, I would like to thank the more than 500 students who have taken my Applied Biostatistics course. Students provide the most important feedback on whether a particular pedagogical approach is effective. While not every experiment succeeds, I am incredibly grateful to each student who has helped me learn what works and what doesn't as they engaged with the material.

Teaching Assistants (TAs)

I have been fortunate to work with outstanding graduate teaching assistants over the past ten years:

- **Derek Nedveck**: Derek played a key role in helping me establish the course during its early years.
- **German Vargas Gutierrez**: A highly skilled statistician, German's assistance was invaluable in refining the course a few years into its development.
- **Chaochih Liu**: A brilliant programmer, Chaochih contributed greatly to the course's organization and structure.
- **Husain Agha**: Husain has remarkable insights into statistics, genetics, and teaching. My work has greatly benefited from bouncing ideas off him.
- **Brooke Kern**: Brooke was not only an exceptional TA but also a valuable collaborator. Much of the data in this book is drawn from her dissertation research.

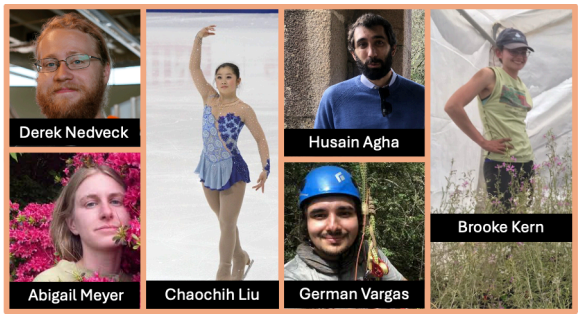

Figure 2: My incredible TAs who have all helped shape this material.

Collaborators

Brooke Kern, Dave Moeller and Shelley Sianta have generated much of the data in this book and have been patient with my delays in turning around our research during teaching times.

Teaching Colleagues

I have learned a lot about statistics and how to teach it from John Fieberg. His book, Statistics for Ecologists is fantastic!

Unknowing contributors

The online community of statistics and R teaching is an amazing place. I have borrowed heavily from the many amazing free resources. Here are the most critical:

- Allison Horst has fantastic illustrations for statistics that she makes freely available.
- Peter D.R. Higgins has created a truly marvelous book – Reproducible Medical Research With R (Higgins (2024)). I have learned a lot and stolen some teaching tricks from this work.
- Jenny Bryan has helped me think about getting students able to do things in R well and quickly. Here book, STAT

545: Data wrangling, exploration, and analysis with R (Bryan (2020)), is a classic.

Bergstrom, Carl T, and Jevin D West. 2020. *Calling Bullshit: The Art of Skepticism in a Data-Driven World*. Random House.

Bryan, Jennifer J. 2020. *STAT 545: Data Wrangling, Exploration, and Analysis with r*. Bookdown. https://stat545.com.

Grolemund, Garrett. 2014. *Hands-on Programming with r: Write Your Own Functions and Simulations*. " O'Reilly Media, Inc.".

Grolemund, Garrett, and Hadley Wickham. 2018. "R for Data Science."

Higgins, Peter D. R. 2024. *Reproducible Medical Research with r*. Bookdown. https://bookdown.org/pdr_higgins/rmrwr/.

Whitlock, Michael C., and Dolph Schluter. 2020. *The Analysis of Biological Data*. Third. Macmillan.

Wilke, Claus O. 2019. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. O'Reilly Media.