

A primer for biostatistics in R

cjlortie

Contents

1	Introduction	5
2	Literate coding	11
3	Stats used in eeb I	13
4	Stats used in eeb II	15
5	Hackathon	17
6	Test	19

Chapter 1

Introduction



Welcome to a primer for biostatistics in R.

Mathematical! Adventure time! Well, the mathematical part is up to you, but this is an adventure. This set of learning materials is a guide developed to support you in better developing critical thinking using statistics. Critical

thinking very generally is a mode of thinking that is self-directed and evidence based. Statistical thinking is thus an ideal opportunity and partner in honing literacy adventure skills in this domain. Enhancing clarity, accuracy, precision, relevance, depth, breadth, significance, logic and fairness - all key criteria of critical thinking - with data or evidence both quantitative and qualitative is a profound tool as a scientist and citizen. It should be fundamental to statistics. Hence, the primary goal of this set of materials is to engender statistical thinking that embodies these principles and explores these criteria using data.

The open and free resources associated with learning statistics is nearly infinite online particularly in R. The programming language R is a free, open source programming environment ideal for statistics. There are other similar alternatives, but here R is used to support and scaffold critical thinking and statistical literacy because a significant component of many biologists use R including ecologists (Lai et al., 2019). Importantly, it provides a simple and clear mechanism to document, annotate, tidy up, write down, and literally show your work - like in math class. This benefits you. You see your ideas written down and can explore logic, fairness, and all the criteria listed above. It also enables you to repeat, replicate, and share your work.

Course outline

If you are electing to engage with this learning opportunity formally for BIOL5081 at York University, here is the official course outline.

Learning outcomes

1. Build a tidy, logical data model for a graduate-level dataset.
2. Develop a reproducible data and statistical workflow.
3. Design and complete intermediate-level data visualizations appropriate for a graduate-level tidy dataset.
4. Identify a range of suitable univariate or multivariate statistical approaches that can be applied to any dataset.
5. Interpret statistical output to quantify statistical model performance.
6. Complete fundamental exploratory data analysis on a representative dataset.
7. Appreciate the strengths and limitations of open science, data science, and evidence-based collaboration models.

Structure

Read a book. The new statistics with R. An introduction for biologists (Hector, 2017).

Write a book review. Ten simple rules for writing statistical book reviews (Lortie, 2019) suggests a critical thinking framework to adopt for this process.

Learn-by-doing here.

Do a hackathon.

Do a hackathon as a test and submit for grading & review.

Rationale

Some learn best by reading. Some learn best by doing. We can all benefit from both approaches to refining our critical thinking through statistics.

Two summative (i.e. graded outcomes) include the book review and the test.

Schedule

week	grand challenges
1	[Tidy data in R](https://www.jstatsoft.org/article/view/v059i10)
2	[Literate statistical coding](https://ojs.library.queensu.ca/index.php/IEE/article/view/6559) and [Data sci
3	Statistics for ecology and evolution I [(CH4 in text)](https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780190238281.001.0001/chapter-4)
4	Statistics for ecology and evolution II [(CH10 and 11 in text)](https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780190238281.001.0001/chapter-10)
5	Book review due and hackathon
6	Test

Instructions

Read the text at your own pace. At least hit the key chapters CH4, 10 & 11 to write the review and submit your insights by the fifth week of work (if you choose to do 1-2 tasks per week as suggested in the schedule). If you are taking BIOL5081, please see official course outline and submit all work to turnitin.com as PDF only (even for the R work - knit to pdf).

Each week, read, discuss if you elect to work synchronously, and try the challenge provided.

The final two weeks, that hackathon is a warm up to the test. Grab the dataset, apply your critical thinking skills, code and show your work, and capture code and outputs as PDF. The hackathon is a stepping stone, formative process for to check if you are ready to think on your feet, write code, and apply biostatistical thinking to a challenge. The test is the exact same approach but summative, i.e. you submit for review and grading to a peer or instructor like me.

Citation

License

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Tidy data in R

Tidiness is next to naturalness. We are wired up to see patterns and organize. Put that tendency to good work in data and statistical critical thinking.

Learning outcomes

1. Consider data structures such as long versus wide.
2. Read in a dataset to the R environment.
3. Do a t-test.

Critical thinking

Tidy data thinking was pioneered in the R world (Wickham, 2014). This philosophy to first considering the basic format of your data is transformational and profound. It beautifully connects to logic. Better yet, it sets you up for easier stats and plots in many environments including R. There is an excellent chapter on this topic in the free, open text R for Data Science.

Adventure time

Very simple life data to explore some ideas about meditation, steps, resting heart rate and the importance of instrument variation. Data are here. Explore the t-test in R for this adventure. Is the number of steps or sleep different from 0? Do the means estimated from a watch versus simple Fitbit tracker vary for simple measures? Did 0 versus 12 mins of meditation per day influence a relevant measure?

```
library(tidyverse)
simple_life <- read_csv(url("https://ndownloader.figshare.com/files/28920855"))
simple_life
```

```
## # A tibble: 9 x 7
##   simple_date steps_fitbit sleep_fitbit    hr steps_watch sleep_watch
##   <date>         <dbl>         <dbl> <dbl>         <dbl>         <dbl>
## 1 2021-06-02      20913           429    54      25197          314
## 2 2021-06-03       6904           447    53      13042          302
## 3 2021-06-04      19548           449    56      23285          413
## 4 2021-06-05      19311           423    56      25832          355
```



```
## 5 2021-06-06      26159      435    58      29533      385
## 6 2021-06-07      21618      358    56      27796      240
## 7 2021-06-08      20890      492    53      24360      434
## 8 2021-06-09      12008      541    53      14517      399
## 9 2021-06-10      18058      436    57      22392      403
## # ... with 1 more variable: meditation_mins <dbl>
```

Reflection questions

1. What can a t-test do? Can you imagine other functions for a t-test in the context of your work and life?
2. What are the limitations of a t-test?
3. Is the data structure wide, long, and how can you consider tidying this evidence? Are there variables that represent the same concept?

Chapter 2

Literate coding

Chapter 3

Stats used in eeb I

Chapter 4

Stats used in eeb II

Chapter 5

Hackathon

Chapter 6

Test

Bibliography

- Hector, A. (2017). *The New Statistics with R*. Oxford University Press, Oxford.
- Lai, J., Lortie, C. J., Muenchen, R. A., Yang, J., and Ma, K. (2019). Evaluating the popularity of r in ecology. *Ecosphere*, 10(1):e02567.
- Lortie, C. J. (2019). Ten simple rules for writing statistical book reviews. *PLOS Computational Biology*, 15(1):e1006562.
- Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59:1–23.