A primer for biostatistics in R

cjlortie

Contents

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 (Facionie, 2017). 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

Slide decks are optional. The decks simply highlight some of the connections between the criteria for critical thinking and statistical heuristics.

week	adventure
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/1(
4	Statistics for ecology and evolution II [(CH10 and 11 in text)](https://oxford.universitypressscholarship.cor
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>
```
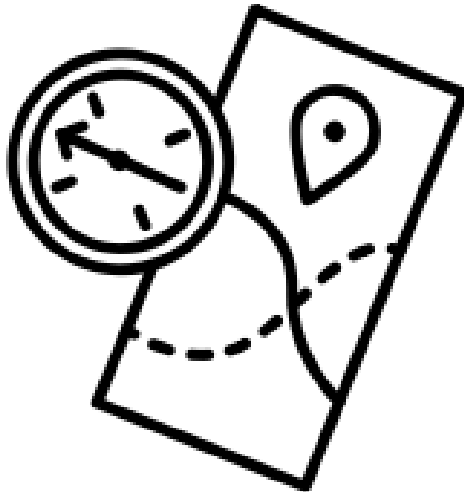
Reflection questions

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

Your code is a story too. Use your code and annotation of decisions (en)coded in your data manipulations, calculations, models, and plots to communicate clarity, logic, relevance, and depth. This story is not just for your collaborators - it is for you. Writing down your ideas and work down makes it more clear. It also reminds you later, even a week later, why you elected to make a particular decision in your workflow. Tidy data and tidy thinking make for better science.

Learning outcomes

1. Practice writing code and using annotation.

2. Consolidate your understanding of tidy data and critical thinking statistically.

3. Do an ANOVA.

Critical thinking

Tidy data make your life easier. Data structures should match intuition and common sense. Data should have logical structure. Rows are are observations, columns are variables. Tidy data also increase the viability that others can use your data, do better science, reuse science, and help you and your ideas survive and thrive.

Literate coding (Knuth, 1992) should capture a workflow that includes the wrangling you did to get your data ready. Literate code should be able to read by a human AND a machine. If data are already very clean in a spreadsheet, they can easily become a literate, logical dataframe. Nonetheless, you should still use annotation within the introductory code to explain the meta-data of your data to some extent and what you did pre-R to get it tidy. The philosophy here is very similar to the data viz lesson forthcoming that promotes critical thinking statistically through documented and described steps that are replicable and clear.

Adventure time

Many years ago in a galaxy far, far away, a student sowed seeds in the desert at different densities for their PhD research. Here are the data, and here is the publication too (Lortie and Turkington, 2002). This student was not strong in the force, but it was a good adventure in beginning to understand the relative importance of significance biologically and statistically by exploring critical thinking.

```
library(tidyverse)
density <- read_csv(url("https://ndownloader.figshare.com/files/28934310"))
density
```

```
## # A tibble: 152 x 6
##     year transect seed_density_pe~ final_plant_den~ survivorship mean_plant_size
##    <dbl>    <dbl>            <dbl>            <dbl>        <dbl>           <dbl>
## 1   1998        1           0.0625               41        0.461           0.554
## 2   1998        1           0.0625               47        0.712           0.356
## 3   1998        1           0.0625               60        0.698           0.301
## 4   1998        1           0.25                 31        0.525           0.808
## 5   1998        1           0.25                 50        0.505           0.212
## 6   1998        1           0.25                 58        0.563           0.148
## 7   1998        1           1                    30        0.273           0.578
## 8   1998        1           1                    42        0.243           1.28
```

```
##  9  1998          1            1                     73          0.619              0.719
## 10  1998          1            2                     46          0.263              0.652
## # ... with 142 more rows
```

Reflection questions

Chapter 3

Stats used in eeb I

Chapter 4

Stats used in eeb II

Chapter 5

Hackathon

Chapter 6

Test

Bibliography

Facionie, P. (2017). Critical thinking: what it is and why it counts. *Insight Assessment*, California Academic Press.

Hector, A. (2017). *The New Statistics with R*. Oxford University Press, Oxford.

Knuth, D. (1992). *Literate Programming*. University of Chicago Press, Illinois.

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.

Lortie, C. J. and Turkington, R. (2002). The effect of initial seed density on the structure of a desert annual plant community. *Journal of Ecology*, 90(3):435–445.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59:1–23.