A primer for biostatistics in R

cjlortie

Contents

1 Introduction **5**

2 Literate coding **11**

3 Stats used in eeb I **15**

4 Stats used in eeb II **19**

5 Hackathon **23**

6 Test **25**

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.co
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

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
```

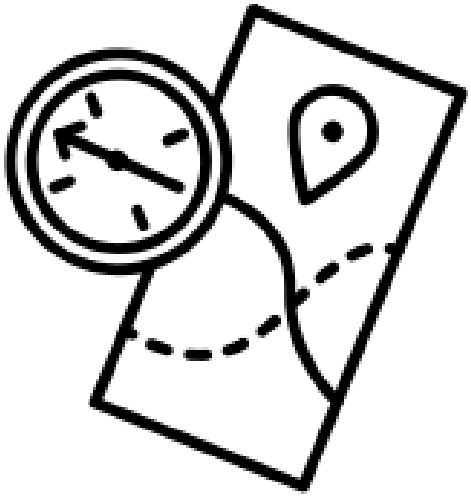
```
## 2 2021-06-03          6904          447    53       13042          302
## 3 2021-06-04         19548          449    56       23285          413
## 4 2021-06-05         19311          423    56       25832          355
## 5 2021-06-06         26159          435    58       29533          385
## 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. For your adventure, test whether a set of groups differ from one another. For instance, test whether transects, or years, or even the density of seeds planted differs in an outcome measure such as mean plant size.

```
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

1. What is the difference between a t-test and an ANOVA?

2. What is the difference between an ANOVA and GLM?

3. What are some of the ways that these simple data can be further analyzed?

4. When you explored annotation and describing your decisions and workflow for these data adventure, was it logical and clear to you if you ignored the R code?

Chapter 3

Stats used in eeb I

Many approaches and critical thinking heuristics in ecology & evolutionary biology (eeb) are relevant to other disciplines.

Learning outcomes

1. Develop your data viz skills.

2. Hone your critical thinking statistically by iterative plotting-modeling a dataset.

3. Do a regression analysis.

Critical thinking

Clean simple graphics are powerful tools in statistics (and in scientific communication). Tufte (Tufte, 2006) and others have shaped data scientists and statisticians in developing more libraries, new standards, and assumptions associated with graphical representations of data. Data viz must highlight the differences, show underlying data structures, and provide insights into the specific research project. R is infinitely customizable in all these respects. There are at least two major current paradigms (there are more these are the two dominant idea sets). Base R plots are simple, relatively flexible, and very easy. However, their grammar, i.e their rules of coding are not modern. Ggplot and related libraries invoke a new, formal grammar of graphics (Leland, 2005) that is more logical, more flexible, but divergent from base R code. It is worth the time to understand the differences and know when to use each.

Evolution of plotting in statistics using R in particular went from base-R then onto lattice then to the ggvis universe with the most recent library being ggplot (Wickham, 2016). Base-R is certainly useful in some contexts as is the lattice and lattice extra library. However, ggplot now encompasses all these capacities with a much simpler set of grammar (i.e. rules and order). Nonetheless, you should be able to read base-R code for plots and be able to do some as well. The philosophy or grammar of modern graphics is well articulated and includes the following key principles. The grammar of graphics layers primacy of ideas (simple first, then more complex) i.e. you build up your plots data are mapped to aesthetic attributes and geometric objects data first then statistics even in plots (Wickham, 2010). This directly supports critical thinking statistically because it promotes depth (literally), precision, and also accuracy in the decisions you make to show your evidence.

Adventure time

Here are a deeper set of quantified life data. Explore whether movement predicts total sleep or its efficiency. Plot out some patterns first, then, do a regression.

```
library(tidyverse)
life <- read_csv(url("https://ndownloader.figshare.com/files/28920729"))
life
```

```
## # A tibble: 4,561 x 7
##    simple_date  year steps mins_asleep efficiency lagged_sleep lagged_efficiency
##    <date>      <dbl> <dbl>       <dbl>      <dbl>        <dbl>              <dbl>
##  1 2011-01-25   2011 13900         481         96          504                 99
##  2 2011-01-26   2011 19229         478         96          481                 96
##  3 2011-01-27   2011 13103         474         96          478                 96
##  4 2011-01-28   2011  7374         491         96          474                 96
##  5 2011-01-29   2011 19132         436         96          491                 96
##  6 2011-01-30   2011 17157         447         98          436                 96
##  7 2011-01-31   2011 19759         456         99          447                 98
```

```
##  8 2011-02-01   2011 18157         455         98         456         99
##  9 2011-02-02   2011  8768         465         97         455         98
## 10 2011-02-03   2011  9150         411         98         465         97
## # ... with 4,551 more rows
```

Reflection questions

1. When do you use regression versus correlation?

2. How could you incorporate time into your plots or statistical models?

3. Did the visualization highlight some of the criteria associated with critical thinking statistically more than others?

Chapter 4

Stats used in eeb II

There is much counting in ecology & evolutionary biology (eeb). We count individuals, species, populations, interactions, and then map out diversity and distributions to infer process. Many disciplines use similar logic in the structure of their evidence and experimental design with statistics.

Learning outcomes

1. Practice your critical workflow for data and statistics that is replicable and literate.

2. Appreciate the value of generalized statistical models that connect to one another conceptually.

3. Do a GLM.

Critical thinking

Exploratory data analyses is everything we have done. This is a primary approach to better understanding your evidence without introducing bias. Transparency is key.

Workflow we have developed but that you nuance based on your cognitive and critical thinking style and strengths.

a. Tidy data.

b. Inspect data structure.

c. Data viz.

d. Basic exploratory data analyses.

However, now that we are ready to apply models, we add in one more tiny step. Continue to visualize the data to better understand its typology and underlying distribution. Then, you are ready to fit your models. Exploratory data analyses is an intermediate step to this end. EDA includes testing assumptions in the data, fitting basic models that ignore covariates, fitting relevant and logical models to explore the data, training your data, and exploring sensitivity (Ellison, 2001). This process builds a viable path for further inquiry, and it is a model builder that is predicated upon critital thinking to ensure you inference (deduction, induction) is aligned with your evidence (Yu, 1994).

A statistical model is an elegant, representative simplification of the patterns you have identified through data viz and EDA (Mengersen et al., 2013). It is a formal mathematical relationship between factors of interest. It should capture data/experimental structure including the key variables, appropriate levels, and relevant covariation or contexts that mediate outcomes. It should support the data viz. It should provide an estimate of the statistical likelihood or probability of differences. Ideally, the underlying coefficients should also be mined to convey an estimate of effect sizes. A t.test, chi.square test, regression/linear model, general linear model, or generalized linear mixed model are all examples of models that describe and summarize patterns and each have associated assumptions about the data they embody. Hence, the final step pre-model fit, is explore distributions.

Conceptually, there are two kind of models. Those that look back and those that look forward. Think tardis or time machine. A model is always a snapshot using your time machine. It can be a grab of what happened or a future snap

of what you predict. In R, there is simple code to time travel in either direction. Actually, there is no time - Arrow of time - only an observer potential perception of it. Statistical models are our observers here. These observers use 'probability distributions' as we described in the first week sensu statistical thinking to calibrate what the think critically when observed or will observe given the evidence at hand. Here are two super resources to further support this in a proximate sense that align with critical thinking. Choosing the correct statistical test made easy (Gunawardana, 2004), and a flowchart for selecting commonly used statistics developed by Bates College.

Adventure time

Here is an impressive dataset describing bird counts in Toronto. These data were collected by York University undergraduates in an experimental design course. Explore whether there is a bias in detection by behaviour and identify the most common species by location in Toronto - at least as estimated using these data. For your curiosity, here are data collected in another larger citizen science endeavour - The Christmas Bird Count for Southern Ontario region centered around the Greater Toronto Area. If you wish to adventure further afield, contrast the two datasets.

```
library(tidyverse)
birds <- read_csv(url("https://knb.ecoinformatics.org/knb/d1/mn/v2/object/urn%3Auuid%3Aa84a9673-8
birds
```

```
## # A tibble: 826 x 11
##     year experiment   source   rep date   location  species  frequency behaviour
##    <dbl> <chr>        <chr>  <dbl> <chr>  <chr>     <chr>        <dbl> <chr>
##  1  2020 balcony bir~ full       1 10/13~ Holditch~ Agelaiu~         3 flying
##  2  2020 balcony bir~ full       1 10/13~ Holditch~ Agelaiu~         4 flying
##  3  2020 balcony bir~ full       1 10/13~ Holditch~ Agelaiu~         1 perching
##  4  2020 balcony bir~ full       1 10/16~ High Park Aix spi~         4 swimming
##  5  2020 balcony bir~ full       1 10/9/~ Vaughan   Anas pl~         4 flying
##  6  2020 balcony bir~ full       1 10/9/~ Vaughan   Anas pl~         6 flying
##  7  2020 balcony bir~ full       1 10/9/~ Vaughan   Anas pl~         9 flying
##  8  2020 balcony bir~ full       1 10/9/~ Vaughan   Anas pl~        10 flying
##  9  2020 balcony bir~ full       1 10/9/~ Vaughan   Anas pl~         2 inactive
## 10  2020 balcony bir~ full       1 10/9/~ Vaughan   Anas pl~         2 inactive
## # ... with 816 more rows, and 2 more variables: inititals <chr>,
## #   citation_DOI <chr>
```

Reflection questions

1. When do you move from EDA to model fitting?

2. Are there ways to mitigate bias and p-hacking through formal workflows?

3. Did building a model such as GLM align with critical thinking and intuition, i.e from critical thinking was it accurate and fair? Did the EDA-to-model process legitimately represent the patterns in the observations recorded.

Chapter 5

Hackathon

All models are wrong but some are useful (Stouffer, 2019; Box, 1976). Critical thinking with statistics is thus critical to ensure that we best support evidence informed decision making in society (Lortie and Owen, 2020; Neelen and Kirschner, 2020).

Learning outcomes

1. Appreciate the challenge of working with data to apply a critical design mindset to statistical solutions.

2. Practice your workflow and literate coding.

3. Refine your thinking and coding for efficiency.

Critical thinking

Efficiency is a fascinating topic in statistics (Craycraft, 1999; Kenett et al., 2003; Norman, 2003). Here, we can simplify this using critical thinking criteria. Efficiency = sufficiency. Provided it is logical, fair, and accurate. Your plots and statistical model should representative a reasonable and likely description of the data at hand.

Adventure time

```
library(tidyverse)
birds <- read_csv(url("https://knb.ecoinformatics.org/knb/d1/mn/v2/object/urn%3Auuid%3A
birds
```

```
## # A tibble: 826 x 11
##     year experiment   source   rep date   location  species  frequency behaviour
##    <dbl> <chr>        <chr>  <dbl> <chr>  <chr>     <chr>        <dbl> <chr>
##  1  2020 balcony bir~ full       1 10/13~ Holditch~ Agelaiu~         3 flying
##  2  2020 balcony bir~ full       1 10/13~ Holditch~ Agelaiu~         4 flying
##  3  2020 balcony bir~ full       1 10/13~ Holditch~ Agelaiu~         1 perching
##  4  2020 balcony bir~ full       1 10/16~ High Park Aix spi~         4 swimming
##  5  2020 balcony bir~ full       1 10/9/~ Vaughan   Anas pl~         4 flying
##  6  2020 balcony bir~ full       1 10/9/~ Vaughan   Anas pl~         6 flying
##  7  2020 balcony bir~ full       1 10/9/~ Vaughan   Anas pl~         9 flying
##  8  2020 balcony bir~ full       1 10/9/~ Vaughan   Anas pl~        10 flying
##  9  2020 balcony bir~ full       1 10/9/~ Vaughan   Anas pl~         2 inactive
## 10  2020 balcony bir~ full       1 10/9/~ Vaughan   Anas pl~         2 inactive
## # ... with 816 more rows, and 2 more variables: inititals <chr>,
## #   citation_DOI <chr>
```

Reflection questions

1. When do you move from EDA to model fitting?

2. Are there ways to mitigate bias and p-hacking through formal workflows?

3. Did building a model such as GLM align with critical thinking and intuition, i.e from critical thinking was it accurate and fair? Did the EDA-to-model process legitimately represent the patterns in the observations recorded.

Chapter 6

Test

Bibliography

Box, G. (1976). Science and statistics. *Journal of American Statistical Association*, 71:791–799.

Craycraft, C. (1999). A review of statistical techniques in measuring efficiency. *Journal of Public Budgeting, Accounting and Financial Management*, 11(1):19–27.

Ellison, A. (2001). *Exploratory data analysis and graphic display.*, pages 37–62. Oxford University Press, Oxford, second edition.

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

Gunawardana, N. (2004). Choosing the correct statistical test made easy. *SMJ*, 7:33–37.

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

Kenett, R. S., Coleman, S., and Stewardson, D. (2003). Statistical efficiency: The practical perspective. *Quality and Reliability Engineering International*, 19(4):265–272.

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.

Leland, W. (2005). *The Grammar of Graphics.* Springer, Chicago, second edition.

Lortie, C. J. (2019). Ten simple rules for writing statistical book reviews. *PLOS Computational Biology*, 15(1):e1006562.

Lortie, C. J. and Owen, M. (2020). Ten simple rules to facilitate evidence implementation in the environmental sciences. *FACETS*, 5(1):642–650.

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.

Mengersen, K., Schmid, B., Jennions, M. D., and Gurevitch, J. (2013). *Statistical models and approaches to inference*, book section 8, pages 89–107. Princeton University Press, Princeton and Oxford.

Neelen, M. and Kirschner, P. (2020). *Evidence-informed learning design: creating training to improve performance*. Kogan Page, London.

Norman, P. (2003). Statistical discrimination and efficiency. *The Review of Economic Studies*, 70(3):615–627.

Stouffer, D. B. (2019). All ecological models are wrong, but some are useful. *Journal of Animal Ecology*, 88(2):192–195.

Tufte, E. (2006). *Beautiful Evidence.* Graphics Press, Chesire, Connecticut.

Wickham, H. (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, 19:3–28.

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

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York, second edition.

Yu, C. H. (1994). Abduction? deduction? induction? is there a logic of exploratory data analysis? *ERIC*, ED376173:1–28.