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, please see the official course outline for specific details.

There are two summative assessments.

1. Write a book review for The new Statistics with R.

2. Complete a take-home statistical test (with the dataset provided in chapter 6 herein).

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.

Steps

Read a book. The New Statistics with R. (Hector, 2021).

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) and CH9 in textbook
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 and CH7 in textbook
4	Statistics for ecology and evolution II and CH15 in textbook
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

Lortie, CJ (2021): A primer for biostatistics in R. figshare. Book. https://doi.org/10.6084/m9.figshare.15048597.v3

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?

Deeper dive: explore the var.equal or alternative argument. Test nonparametric analog to this test.

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

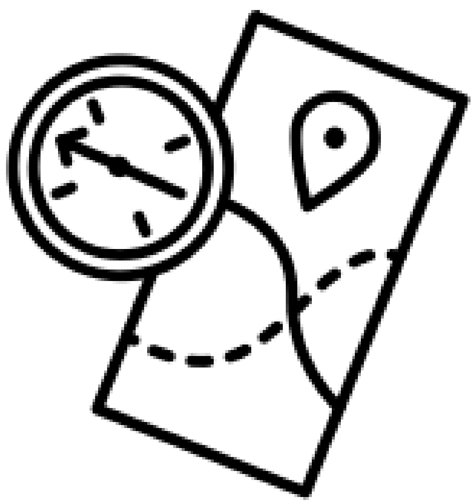
```
## # A tibble: 9 x 7
##   simple_date steps_fitbit sleep_fitbit    hr steps_watch sleep_watch meditati~1
##   <date>             <dbl>        <dbl> <dbl>       <dbl>       <dbl>      <dbl>
## 1 2021-06-02         20913          429    54       25197         314          0
## 2 2021-06-03          6904          447    53       13042         302          0
## 3 2021-06-04         19548          449    56       23285         413         12
## 4 2021-06-05         19311          423    56       25832         355         12
## 5 2021-06-06         26159          435    58       29533         385         12
## 6 2021-06-07         21618          358    56       27796         240          0
## 7 2021-06-08         20890          492    53       24360         434         12
## 8 2021-06-09         12008          541    53       14517         399         12
## 9 2021-06-10         18058          436    57       22392         403         12
## # ... with abbreviated variable name 1: meditation_mins
```

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.

Deeper dive: Check for homoscedasticity or do a post-hoc test.

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

```
## # A tibble: 152 x 6
##     year transect seed_density_per_cm final_plant_density survivorship mean_pl~1
##    <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, and abbreviated variable name 1: mean_plant_size
```

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.

Deeper dive: explore residuals and try the cooks.distance function for outliers.

```
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) (Zuur et al., 2009). 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.

Deeper dive: If you wish to adventure further afield, contrast the two datasets. Explore fitting a different family to the data or explore offset argument versus covariates.

```
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  locat~1 species frequ~2 behav~3 initi~4
##    <dbl> <chr>        <chr>  <dbl> <chr> <chr>   <chr>     <dbl> <chr>   <chr>
##  1  2020 balcony bir~ full       1 10/1~ Holdit~ Agelai~       3 flying  RD
##  2  2020 balcony bir~ full       1 10/1~ Holdit~ Agelai~       4 flying  RD
##  3  2020 balcony bir~ full       1 10/1~ Holdit~ Agelai~       1 perchi~ RD
##  4  2020 balcony bir~ full       1 10/1~ High P~ Aix sp~       4 swimmi~ AB
##  5  2020 balcony bir~ full       1 10/9~ Vaughan Anas p~       4 flying  TA
##  6  2020 balcony bir~ full       1 10/9~ Vaughan Anas p~       6 flying  TA
##  7  2020 balcony bir~ full       1 10/9~ Vaughan Anas p~       9 flying  TA
##  8  2020 balcony bir~ full       1 10/9~ Vaughan Anas p~      10 flying  TA
##  9  2020 balcony bir~ full       1 10/9~ Vaughan Anas p~       2 inacti~ TA
## 10  2020 balcony bir~ full       1 10/9~ Vaughan Anas p~       2 inacti~ TA
## # ... with 816 more rows, 1 more variable: citation_DOI <chr>, and abbreviated
## #   variable names 1: location, 2: frequency, 3: behaviour, 4: inititals
```

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 effectively 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 thinking & creative design mindset to statistical solutions.

2. Practice your workflow and literate coding before a summative test.

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 the critical thinking criteria we have extensively refined and applied to numerous, tidy challenges. Efficiency = sufficiency (provided it is logical, fair, and accurate). Your plots and statistical models should represent a reasonable and likely description of the data at hand. This section is a formative opportunity for you to evaluate your skills and strengths in logic, efficiency, fair adventuring, workflows, and literate coding prior to the final section - a test. You are provided with a general dataset(s). The adventure is solve a very generalized challenge that is embodied in the evidence.

Adventure time

Candy. Candy. Candy. Take a peek at these sweet data. Contrast Canada and USA candy sales at Halloween. Considering including population density in your model for each country for each year so as not to introduce variation and to be more accurate in estimating meaningful differences.

Canadian Candy
USA Candy & Halloween spending
Human populations

Deeper dive: contrast GLMM model performance, examine temporal effects, or explore GAMs.

```
library(tidyverse)
Canada <- read_csv(url("https://figshare.com/ndownloader/files/30990820"))
Canada
```

```
## # A tibble: 233 x 3
##    month  year  candy
##    <dbl> <dbl>  <dbl>
##  1     1  1997 101014
##  2     2  1997 101938
##  3     3  1997 136057
##  4     4  1997 105601
##  5     5  1997 119123
##  6     6  1997 107689
##  7     7  1997 113399
##  8     8  1997 113934
##  9     9  1997 109441
## 10    10  1997 146876
```

```
## # ... with 223 more rows
```

```r
USA <- read_csv(url("https://figshare.com/ndownloader/files/25190510"))
USA
```

```
## # A tibble: 16 x 6
##     year total costumes candy decorations cards
##    <dbl> <dbl>    <dbl> <dbl>       <dbl> <dbl>
##  1  2005   3.3      1.2   1.2         0.8   0.1
##  2  2006   5        1.8   1.6         1.3   0.3
##  3  2007   5.1      1.8   1.6         1.4   0.3
##  4  2008   5.8      2.1   1.8         1.6   0.3
##  5  2009   4.7      1.7   1.5         1.2   0.3
##  6  2010   5.8      2     1.8         1.6   0.3
##  7  2011   6.9      2.5   2           1.9   0.5
##  8  2012   8        2.9   2.3         2.4   0.6
##  9  2013   7        2.6   2.1         2     0.4
## 10  2014   7.4      2.8   2.2         2     0.4
## 11  2015   6.9      2.5   2.1         1.9   0.3
## 12  2016   8.4      3.1   2.5         2.4   0.4
## 13  2017   9.1      3.3   2.7         2.7   0.4
## 14  2018   9        3.2   2.6         2.7   0.4
## 15  2019   8.8      3.2   2.6         2.6   0.4
## 16  2020   8        2.6   2.4         2.6   0.4
```

```r
humans <- read_csv(url("https://figshare.com/ndownloader/files/30993373"))
humans
```

```
## # A tibble: 249 x 72
##    country `1950` `1951` `1952` `1953` `1954` `1955` `1956` `1957` `1958` `1959`
##    <chr>   <chr>  <chr>  <chr>  <chr>  <chr>  <chr>  <chr>  <chr>  <chr>  <chr>
##  1 Burundi 2 309  2 360  2 406  2 449  2 492  2 537  2 585  2 636  2 689  2 743
##  2 Comoros 159    163    167    170    173    176    179    182    185    188
##  3 Djibou~ 62     63     65     66     68     70     71     74     76     80
##  4 Eritrea 822    835    849    865    882    900    919    939    961    983
##  5 Ethiop~ 18 128 18 467 18 820 19 184 19 560 19 947 20 348 20 764 21 201 21 662
##  6 Kenya   6 077  6 242  6 416  6 598  6 789  6 988  7 195  7 412  7 638  7 874
##  7 Madaga~ 4 084  4 168  4 257  4 349  4 444  4 544  4 647  4 754  4 865  4 980
##  8 Malawi  2 954  3 012  3 072  3 136  3 202  3 271  3 342  3 417  3 495  3 576
##  9 Maurit~ 493    506    521    537    554    571    588    605    623    641
## 10 Mayotte 15     16     16     17     18     19     20     21     22     23
## # ... with 239 more rows, and 61 more variables: `1960` <chr>, `1961` <chr>,
## #   `1962` <chr>, `1963` <chr>, `1964` <chr>, `1965` <chr>, `1966` <chr>,
## #   `1967` <chr>, `1968` <chr>, `1969` <chr>, `1970` <chr>, `1971` <chr>,
## #   `1972` <chr>, `1973` <chr>, `1974` <chr>, `1975` <chr>, `1976` <chr>,
## #   `1977` <chr>, `1978` <chr>, `1979` <chr>, `1980` <chr>, `1981` <chr>,
## #   `1982` <chr>, `1983` <chr>, `1984` <chr>, `1985` <chr>, `1986` <chr>,
```

```
## #   `1987` <chr>, `1988` <chr>, `1989` <chr>, `1990` <chr>, `1991` <chr>, ...
```

Reflection questions

1. How does veracity of data from different resources potentially influence your critical thinking?

2. Can joining data introduce errors?
3. How does the available data bias the inference and interpretation of relative variables on key outcomes?

Book review

Instructions

- Read the key chapters that best support your learning from the text 'The New Statistics with R' (Hector, 2017).

- Please use the ten simple rules for reviews (Lortie, 2019) as your instructions how to do a review.

- Write and submit a short, less than 2000 word review of this text and submit to turnitin.com.

Examples

1. Doing Meta-Analysis with R - A Hands-On Guide

2. Python and R for the Modern Data Scientist

3. R for Data Science

4. Applied Time Series Analysis with R (2nd Edition)

5. Open Sesame: R for Data Science is Open Science

Rubric

item	concept	description
1	rule 1 the topic	introduce topic, explain necessity, explain scope
2	rule 2 audience	explain audience-level of book and to what extent blend of expertise is needed
3	rule 3 editions	mention different editions or versions and what is changed
4	rule 4 pedagogy	describe pedagogy and structure of chapters
5	rule 5 content	provide a clear overview of what the text covers
6	rule 6 readability	critique the style and clarity of writing
7	rule 7 links	list and explain linkages to concepts and packages
8	rule 8 compare	briefly list what other resources are out there and compare
9	rule 9 commitment	comment on the commitment and effort need to master text
10	rule 10 benefits	list the main benefits of using this text to learn or solve
11	your writing	your writing and coherence are graded for clarity, balance, directness, and convincing
12	total	sum of above concepts

Chapter 6

Test

Put your practice to the test. Here are some excellent cheatsheets to consider for biostats in R, and this is a useful read on good enough practices in scientific computing (Wilson et al., 2017). The goal here was not to become data scientists nor biostatisticians but to encourage you to consider developing and refining your critical thinking skills in the context of evidence, data, and statistical reasoning.

Learning outcomes

1. Complete fundamental exploratory data analysis on a representative dataset culminating with a fair and reasonable statistical model.

2. Interpret a statistical analyses that you completed with a focus on relevance, significance, and logic.

3. Communicate biostatistical work clearly and effectively to others.

Critical thinking

At times in many disciplines of biological research, we need to be open to experimentation that is fair, transparent, and replicable but that is implemented based on available data. This experimentatation can also happen after we have data. It can be an exercise in fitting the most appropriate or parsimonous models (Cottingham et al., 2005), applying experimental design principles (Ruxton and Colgrave, 2018), and of course invoking critical thinking. This is not to say we are going on fishing expeditions, but that that at times, we have only certain data to describe a system and are tasked or obligated to use the best possible evidence we have to infer relevant processes. For instance, we might compile field data, data from online resources or data products for climate or landscapes, or reuse data on traits on genetics and link these different evidence streams together to explore a question. Critical thinking in statistics can be an important framework that we leverage to not only do the statistics and fit models but also ensure that we are able to ask the questions we need to. In summary, we have data and need an answer but have to use open and transparent thinking with statistics to find the best question.

Workflow for hackathons

A hackathon in data science and the computational science is a fixed-duration, collaborative endeavor to develop a solution for a focussed challenge. The goal is to have a reasonably functional first-approximation that is viable and/or describes the key processes for a system or dataset. It is a blend of hacking and marathon to race or sprint towards a clear endpoint in development. In the data and statistical sciences, we intensively work to deepen our understanding of evidence ideally with key data visualizations and a model that predicts or describes key outcomes. The advantage of setting a reasonably short but fair duration is that it reduces the likelihood that tangents are unduly developed. It also hones your coding, research skills, and statistical reasoning through practiced mental model application of statistics to new data to tell a balanced and reproducible, transparent story.

1. Get the data.

2. Read the metadata (and if you get stuck, look up from online resources or related/similar datasets the potential meaning of opaque variable names). Nomenclature and annotation shorthand terminology in a field can be highly specific at times.

3. Consider and ensure that you understand the individual vectors or variables (inspect the dataframe).
4. Develop an informal or formal data map - picture a Sankey diagram (conceptual semantic visualization of relationships between variables).
5. Dig into online resources or literature to ideate on important questions, novel gaps, key theories, or even basic fundamental science that supports these data.
6. Decide on focus and key purpose and begin to plan out an analytical workflow.
7. Determine if you have sufficient data, i.e., consider if you need to augment these data. Augmenting data can be from novel data sources or from reclassification of existing data.
8. Begin your exploration of the dimensions and scope of the variables you were interested in using (skimr, min, max, fitdistrplus, or str-like functions or tools in R).
9. Now, adopt the r4ds workflow such as Fig 1.1, and use plots such as histograms or boxplots to understand depth and range of data, use basic tests as the t.test to explore differences, and prepare for your final statistical model and keystone plot to show the differences you tested.
10. Code and test your main model to address the overarching goal. Decide and revise the best/most representative instance of data viz that illuminates the salient process or patterns examined.

If you favor this method of collaborative work in your lab or team, here are ten simple rules to run a successful BioHackathon.

Test adventure time

York University, Keele Campus is a small urban forest mixed with grasslands and open space. The master gardeners measured nearly 7000 trees over the course of two years. These data were recently compiled and published. There are many fascinating and compelling questions to explore that can support evidence-informed decisions and valuation estimates for this place ecologically, environmentally, and from a trait or species-level perspective. This challenge as a summative test is thus relatively more open ended. Given these data, collected and now published, what can we do to enhance our biological and social understanding and appreciation for a university campus that support people, other animals, and plants. Explore the data, define a relevant challenge or set of questions that would benefit the stakeholders or local community or inform our understanding of a biological theory, and demonstrate your mastery of critical thinking in statistics. Submit your work to turnitin.com as PDF including the code, annotation, rationale, interpretation, and outputs from the viz, EDA, and model(s) that supported your thinking.

Metadata for test data

attribute	description
FID	FID refer to an unique identifier of an object within a table in ArcGIS data
OBJECTID	unique instance of measurement counting rows
Date	month, day, year format
Block	block that York uses in some maps to organize campus into grid
Street_or_	road names
Building_C	building code
Tree_Tag_N	number on the metal tag affixed to each tree on campus
Species_Co	species code acronyms used to abbreviate species names
Common_Nam	not the Latin binomial name for a species, the common name used
Genus	genus is a taxonomic unit that may contain one species (monotypic)
Species	most basic category in the system of taxonomy
DBH	diameter at breast height measured at approximately 1.3 m (4.3 ft)
Number_of_	number of main branches
Percentage	percentage of canopy cover
Crown_Widt	width of the crown
Total_Heig	the total height of tree to the top of canopy, actual top of tree
Latitude	degrees decimals, a notation for expressing latitude and longitude geographic coordi
Longitude	degrees decimals, a notation for expressing latitude and longitude geographic coordi
Height_to_	height to first branch of the main trunk of the tree
Unbalanced	the number of times a tree splits or branches out from main trunk
Reduced_Cr	a measure of reduced crown treatment by the foresters on campus, number of branc
Weak_Yello	an indirect of tree health, Likert Score from 0 to 3 with 0 being no yellow and three
Defoliatio	an indirect of tree health, Likert Score from 0 to 3 with 0 being no evidence of leaf l
Dead_Broke	number of dead or broken branches
Poor_Branc	an indirect of tree health, Likert Score from 0 to 3 with 0 being no evidence of poor
Lean	A tree that leans because it has grown towards the sun often has a curving trunk, s
Trunk_Scar	number of tree scars on main trunk of tree

Test data

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

```
## # A tibble: 6,951 x 27
##      FID OBJECTID Date   Block Street_or_  Build~1 Tree_~2 Speci~3 Commo~4 Genus
##    <dbl>    <dbl> <chr>  <chr> <chr>         <dbl>   <dbl> <chr>   <chr>   <chr>
##  1     0        1 9/7/12 A     Stedman Le~      22       1 lochon  Honey ~ Gled~
##  2     1        2 9/7/12 A     Stedman Le~      22       2 lochon  Honey ~ Gled~
##  3     2        3 9/7/12 A     Stedman Le~      22       3 lochon  Honey ~ Gled~
```

```
## 4      3         4 9/7/12 A     Stedman Le~      22       4 lochon  Honey ~ Gled~
## 5      4         5 9/7/12 A     Stedman Le~      22       5 lochon  Honey ~ Gled~
## 6      5         6 9/7/12 A     Stedman Le~      22       6 lochon  Honey ~ Gled~
## 7      6         7 9/7/12 A     Stedman Le~      22       7 lochon  Honey ~ Gled~
## 8      7         8 9/7/12 A     Stedman Le~      22       8 lochon  Honey ~ Gled~
## 9      8         9 9/7/12 A     Stedman Le~      22       9 lochon  Honey ~ Gled~
## 10     9        10 9/7/12 A     Stedman Le~      22      10 lochon  Honey ~ Gled~
## # ... with 6,941 more rows, 17 more variables: Species <chr>, DBH <dbl>,
## #   Number_of_ <dbl>, Percentage <dbl>, Crown_Widt <dbl>, Total_Heig <dbl>,
## #   Latitude <dbl>, Longitude <dbl>, Height_to_ <dbl>, Unbalanced <dbl>,
## #   Reduced_Cr <dbl>, Weak_Yello <dbl>, Defoliatio <dbl>, Dead_Broke <dbl>,
## #   Poor_Branc <dbl>, Lean <dbl>, Trunk_Scar <dbl>, and abbreviated variable
## #   names 1: Building_C, 2: Tree_Tag_N, 3: Species_Co, 4: Common_Nam
```

Clean code

Effective coding so that others can read it and understand it - not just machines - is an art and a science. Object and function naming that is intuitive really helps. Functions to streamline repeated operations, and annotation to explain steps with headers are all useful. This approach to literate coding for humans is sometimes entitled ‘clean code’. **Here** is a short paper with some tips and tricks relevant to your work when you need to share it (Filazzola and Lortie, 2022).

Rubric

Remember, we are working together to hone our statistical reasoning skills.

The goal is to tell a story with these data.

It does not need to be super complex, but it does need to showcase your skills in understanding key principles such as a GLM with appropriate data visualizations - but any reasonable test that MATCHES the story you tell is great.

Show your work of exploring the data in plots and basic stats, develop your idea, test it, and then have a final key plot showing the relationship you tested.

item	concept	description
1	effective data viz	are there figures exploring the data and is the final main figure publishable in t
2	effective EDA	is the distribution of and relationship between variables explored
3	final data model(s)	does the final model(s) address the purpose of study, appropriate, and assumpt
4	annotation and reporting	is there annotation in the r-code chunks, reporting in the markdown, and an in
5	total	sum of above

Bibliography

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

Cottingham, K., Lennon, J., and Brown, B. (2005). Knowing when to draw the line: designing more informative ecological experiments. *Frontiers in Ecology and the Environment*, 3:145–152.

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.

Filazzola, A. and Lortie, C. (2022). A call for clean code to effectively communicate science. *Methods in Ecology and Evolution*, 3:1–10.

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.

Hector, A. (2021). *The New Statistics with R.* Oxford University Press, Oxford, second edition.

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.

Ruxton, G. and Colgrave, N. (2018). *Experimental Design for the Life Sciences*. Oxford University Press, Oxford, UK, fourth edition.

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.

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., and Teal, T. K. (2017). Good enough practices in scientific computing. *PLOS Computational Biology*, 13(6):e1005510.

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

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., and Smith, G. M. (2009). *GLM and GAM for Count Data*, pages 209–243. Springer New York, New York, NY.