HW1_hwenjun

Wenjun Han

2020-08-29

For each assignment, turn in by the due date/time. Late assignments must be arranged prior to submission. In every case, assignments are to be typed neatly using proper English in Markdown.

This week, we spoke about Reproducible Research, R, Rstudio, Rmarkdown, and LaTeX. To summarize the ideas behind Reproducible Research, we are focusing on Reproducible Analysis. For us, Reproducible Analysis is accomplished by mixing code, figures and text into a cohesive document that fully describes both the process we took to go from data to results and the rational behind our data driven conclusions. Our goal should be to enable a moderately informed reader to follow our document and reproduce the steps we took to reach the results and hopefully conclusions we obtained.

Problem 1

R is an open source, community built, programming platform. Not only is there a plethra of useful web based resources, there also exist in-R tutorials. Please do both of the Primers labeled as The Basics on Rstudio.cloud. (finished)

Problem 2

Now that we have the R environment setup and have a basic understanding of R, let's add Markdown (choose File, New File, R Markdown, pdf).

Let's go ahead and save the file as is. Save the file to the directory containing the *README.md* file you created and committed to your git repo in Homework 0. The filename should be: HW1_pid, i.e. for me it would be HW1_rsettlag.(finished)

You will use this new R Markdown file for the remainder of this homework.

Part A

In this new Rmarkdown file, please type a paragraph about what you are hoping to get out of this class. Include at least 3 specific desired learning objectives in list format.

- Learn several commonly used packages, algorithms
- Learn how to plot data and describe analysis process in a neat format
- Be familiar with R programming language and be able to solve basic problems

Part B

To this, add 3 density functions (Appendix Cassella & Berger) in centered format with equation number, i.e. format this as you would find in a journal.

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty \tag{1}$$

$$f(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1, \alpha > 0, \beta > 0 \tag{2}$$

$$f(x|\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}, -\infty < x < \infty, -\infty < \theta < \infty \tag{3}$$

See (1),(2) and (3) for Normal Distribution, Beta Distribution, Cauchy Distribution, respectively.

Problem 3

A quote from Donoho (1995): "an article about computational results is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result." To the document created in Problem 4, add a summary of the steps in performing Reproducible Research in numbered list format as detailed in:
http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003285.

Next to each item, comment on any challenges you see in performing the step. If you are interested in learning more, a good summary of why this is important can be found in
- https://www.informs.org/ORMS-Today/Public-Articles/October-Volume-38-Number-5/Reproducible-Operations-Research
- https://doi.org/10.1093/biostatistics/kxq028
- http://statweb.stanford.edu/~wavelab/Wavelab_850/wavelab.pdf

- For Every Result, Keep Track of How it was Produced
  - Comment: For novices, they may forget to comment on the analytics process.
- Avoid Manual Data Manipulation Steps
  - Comment: Manual operations between multiple documents can be essential for some cases.
- Archive the Exact Versions of All External Programs Used
  - Comment: You may not have access to previous versions of programs even you archive the files.
- Version Control All Custom Scripts
  - Comment: Some manual operations might not be recorded through version control systems.
- Record All Intermediate Results, When Possible in Standardized Formats
  - Comment: It may require large storage space.
- For Analyses That Include Randomness, Note Underlying Random Seeds.
  - Comment: You may forget the random seeds if they are not noted.
- Always Store Raw Data behind Plots
  - Comment: Sometimes data lost due to inappropriate storage or other mistakes.
- Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected.
  - Comment: The storage context may not allow to do so.
- Connect Textual Statements to Underlying Results
  - Comment: A good report rely on reporter's communication skills.
- Provide Public Access to Scripts, Runs, and Results
  - Comment: The materials can be confidential and they are not suitable for a public access.

Problem 4

Please create and include a basic scatter plot and histogram of an internal R dataset. To get a list of the datasets available use `library(help="datasets")`.

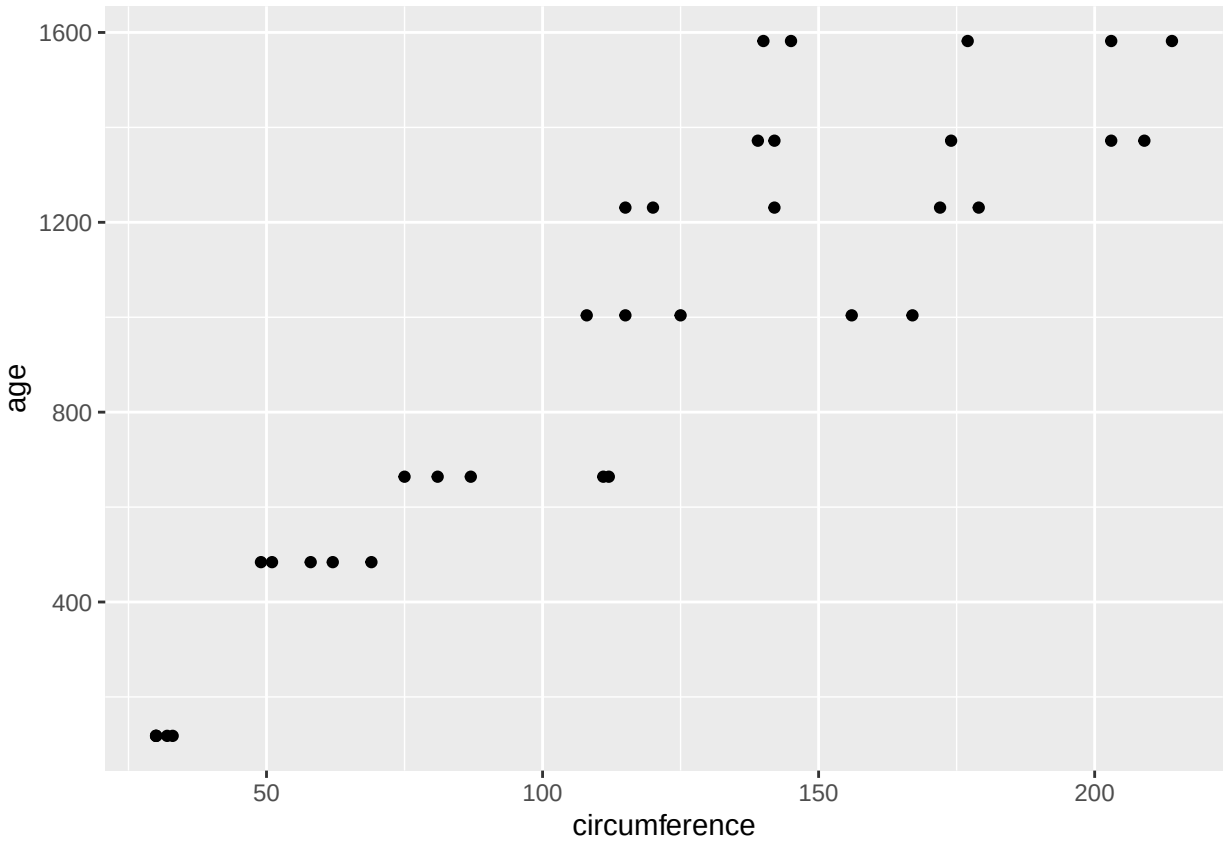
```
Orange
```

```
##    Tree  age circumference
## 1     1  118            30
## 2     1  484            58
## 3     1  664            87
## 4     1 1004           115
## 5     1 1231           120
## 6     1 1372           142
## 7     1 1582           145
## 8     2  118            33
## 9     2  484            69
## 10    2  664           111
## 11    2 1004           156
## 12    2 1231           172
## 13    2 1372           203
## 14    2 1582           203
## 15    3  118            30
## 16    3  484            51
## 17    3  664            75
## 18    3 1004           108
## 19    3 1231           115
## 20    3 1372           139
## 21    3 1582           140
## 22    4  118            32
## 23    4  484            62
## 24    4  664           112
## 25    4 1004           167
## 26    4 1231           179
## 27    4 1372           209
## 28    4 1582           214
## 29    5  118            30
## 30    5  484            49
## 31    5  664            81
## 32    5 1004           125
## 33    5 1231           142
## 34    5 1372           174
## 35    5 1582           177
```

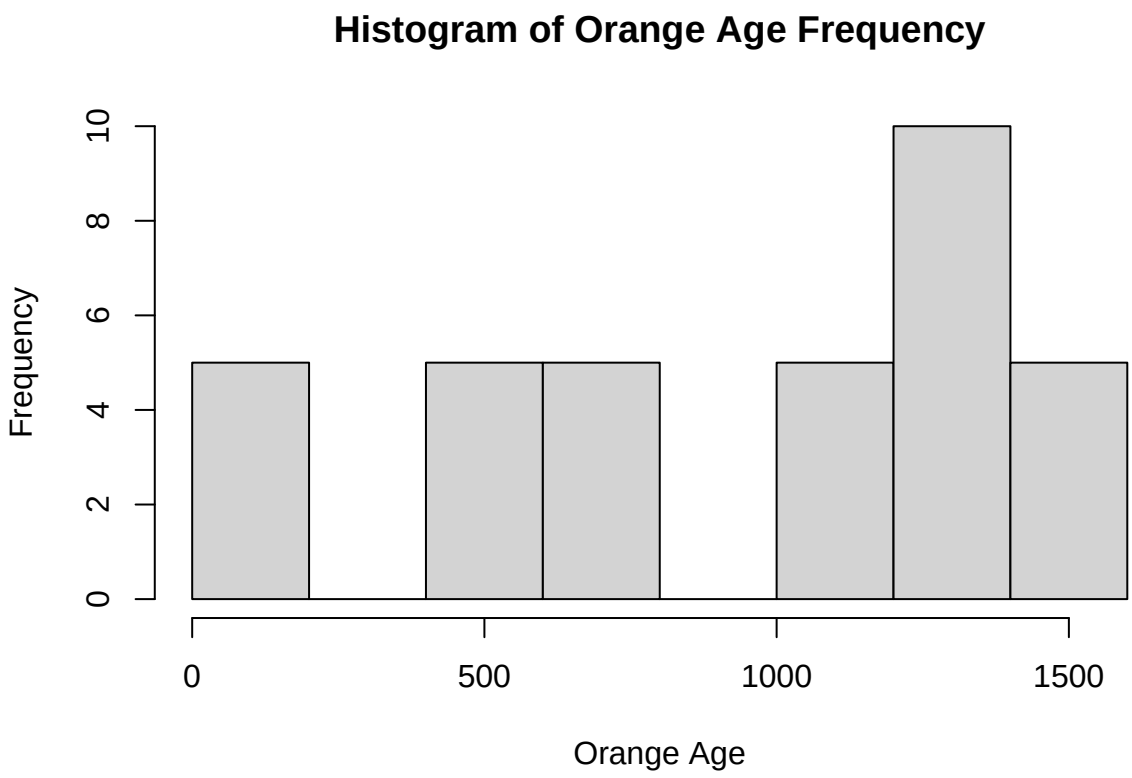
```
install.packages("ggplot2")
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
```

```
library(ggplot2)
ggplot(data = Orange) +
  geom_point(mapping = aes(x = circumference, y = age))
```




```
hist(Orange$age, main = "Histogram of Orange Age Frequency", xlab = "Orange Age")
```



This document containing solutions to Problems 2-4 should be typed in RMarkdown, using

proper English, and knitted to create a pdf document. Do NOT print, we will use git to submit this assignment as detailed below.

Problem 5

Please knit this document to PDF (name should be `HW2_pid`) and push to GitHub:

In the R Terminal, type:

1. git pull
2. git add HW1_pid.[pR]* (NOTE: this should add two files)
3. git commit -m "final HW1 submission"
4. git push

A more detailed description is on the course website under *Submitting Homework*.

Reminder on where to find Git help:

Read through the Git help Chapters 1 and 2. https://git-scm.com/book/en/v2