Homework 2

Due Wednesday Sep 5

2019-09-01

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. To speed our learning, we will use one such tutorial *swirl*. Please install the *swirl* package, install the "R_Programming_E" lesson set, and complete the following lessons: 1-3 and 15. Each lesson takes about 10 min.

From the R command prompt:

```
install.packages("swirl")
library(swirl)
install_course("R_Programming_E")
swirl()
```

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.

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.

I have already gone through a lot of trouble getting to this point, and in the process, already learned a lot about GitHub and hot it operates. The followings are the three desired learning objectives I have:

- Learn more about GitHub and how it can be useful to me later on.
- Learn different R functions and packages that I had not known before.
- Familiarize myself with other programming languages, if that is part of the course objective.

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.

Exponential distribution

$$f(x|\beta) = \frac{1}{\beta} e^{-\frac{x}{\beta}}, 0 \leq x < \infty, \beta > 0 \tag{1}$$

Lognormal distribution

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \frac{e^{-(logx-\mu)^2/(2\sigma^2)}}{x}, 0 \leq x < \infty, -\infty < \mu < \infty \tag{2}$$

Pareto distribution

$$f(x|\alpha, \beta) = \frac{\beta \alpha^{\beta}}{x^{\beta+1}}, a < x < \infty, \alpha > 0, \beta > 0 \tag{3}$$

I hope this is good enough.

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

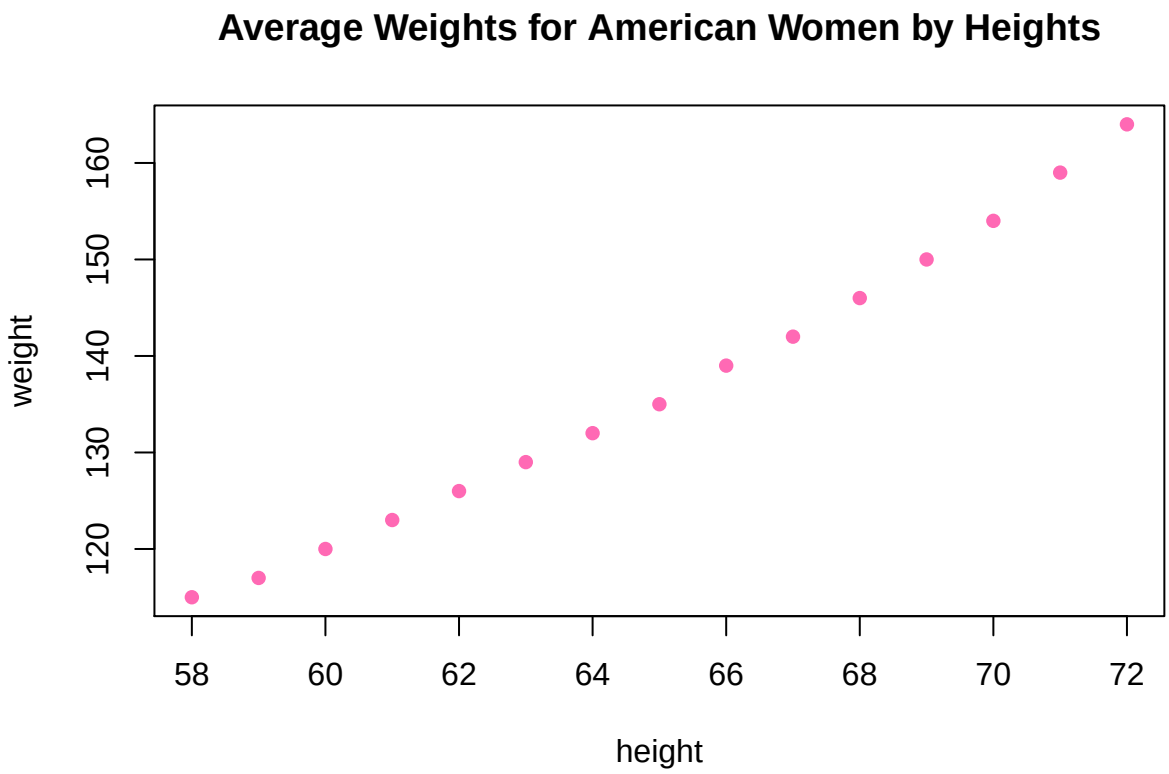
The following is the list of the steps in performing reproducible research, per Sandve et. al.

1. For Every Result, Keep Track of How It Was Produced. I do not see a huge challenge in this unless the methodology that was applied to obtain the result was convoluted and difficult to keep track of, which should not happen in the first place.
2. Avoid Manual Data Manipulation Steps As Much As Possible. Again, I do not see much controversy here, as we should maintain manual data manipulation to minimum. If there is any challenges I see, it would be what to do in case where all other options have failed to return satisfactory results, and one is tempted to lean to manual data manipulation.
3. Archive the Exact Versions of All External Programs Used. One possible obstacle I can see is if there is some user agreement clause or for other technical reasons archiving is not permitted or possible, but I do not know how often this is even an issue. I have faced issues with this; scripts I have produced using outdated version of R does not run in newer version.
4. Version Control All Custom Scripts. I suspect too much of this could compromise efficiency if numerous changes to the programs are required.
5. Record All Intermediate Results, When Possible in Standardized Formats. I suspect this may become increasingly difficult if obtaining each result requires significant amount of time and computing power, or the size of the output is too huge. This is something I have dealt with as a Master's student.
6. For Analyses That Include Randomness, Note Underlying Random Seeds. I think this is a smart move, something I should have done while I was working on research projects. There may be challenges of retaining random seeds when multiple different seeds have been applied.
7. Always Store Raw Data behind Plots. A challenge I can see is in case where contract regarding permissible use of the data has expired, thus would require all raw data to be expunged from the record. This would not be possible anymore.
8. Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected.
9. Connect Textual Statements to Underlying Results. I believe this is important as it allows for better understanding of the results not only from the readers' perspective, but also from my own. However, sometimes I found too much textual statements could lead to greater confusion.
10. Provide Public Access to Scripts, Runs, and Results. A challenge is if the user agreement of data denies public access to those.

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")`.

```
library(help="datasets")
# I used the dataset "women"
plot(women, pch = 16, col = "hotpink",
     main = "Average Weights for American Women by Heights")
```



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 HW2 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