Homework 1

Han Chen

2019-09-03

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.

- Using tools inlcuding RStudio, Git and GitHub, Overleaf, etc. to create reproducible research work.
- Using packages like ggplots to produce high-quality publishable plots.
- Getting familiar with classic ML packages in Python

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.

- Normal distribution

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

- Beta distribution

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

- Cauchy distribution

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

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.

1. For Every Result, Keep Track of How It Was Produced
- Comment:
2. Avoid Manual Data Manipulation Steps
- Comment: Sometimes data manipulation through code would be tideous.
3. Archive the Exact Versions of All External Programs Used
- Comment: A good archive of all programs is not easy to build
4. Version Control All Custom Scripts
- How to centralize all scripts
5. Record All Intermediate Results, When Possible in Standardized Formats
- Comment: No issue
6. For Analyses That Include Randomness, Note Underlying Random Seeds
- Comment: No issue
7. Always Store Raw Data behind Plots
- Comment: No issue
8. Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
- Comment: No issue

9. Connect Textual Statements to Underlying Results

- Comment: No issue

10. Provide Public Access to Scripts, Runs, and Results

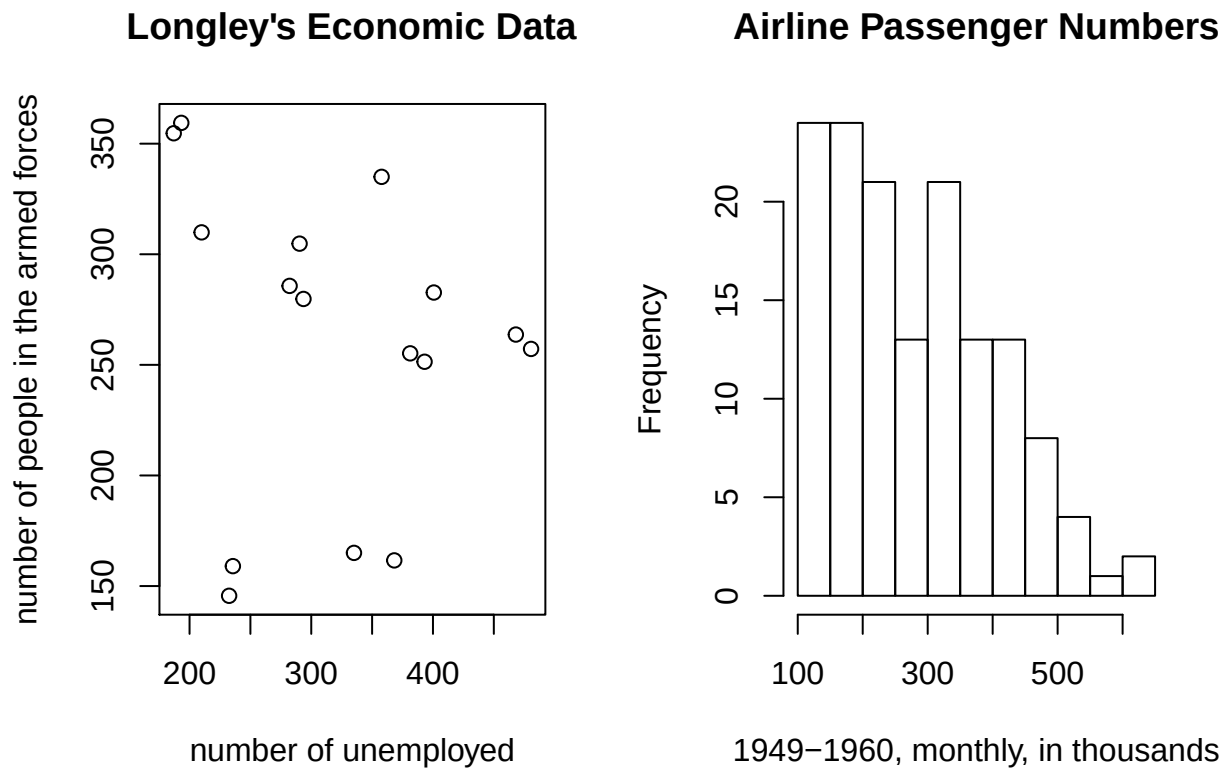
- Comment: Privacy may be a problem

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

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")`.

```
par(mfrow = c(1,2))
plot(longley$Unemployed, longley$Armed.Forces, main = "Longley's Economic Data",
     xlab = "number of unemployed", ylab = "number of people in the armed forces")
hist(AirPassengers, main = "Airline Passenger Numbers",
     xlab = "1949-1960, monthly, in thousands")
```



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