Homework 1

Adeline Guthrie

September 4, 2019

Problem 1

Swirl tutorials completed.

Problem 2

Part A

By the end of this class, I hope to be more organized and efficient in my research when using programming. I hope produce more comprehendable code using structured formatting and clear comments. Although I have a decent background in R, I would not consider myself an expert and I am looking forward to a more structured overview of some of the useful features of R.

Three specific desired learning objectives I have for this course are:

1. to become proficient in Rmarkdown and reinforce my knowledge of R
2. to effectively use version control to keep track of changes in code
3. to be able to conduct proper reproducible research

Part B

The following are three probability density functions (Appendix Cassella & Berger):

Exponential Distribution pdf:

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

Gamma Distribution pdf:

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

Uniform Distribution pdf:

$$f(x|a, b) = \frac{1}{a-b},\ a \leq x \leq b \tag{3}$$

Problem 3

1. **Maintain an analysis workflow detailing each step performed and all factors that may effect the result of each step.**

   A challenge that may arise from this step is being careful to intentionally document each step as it is performed. It may be easier and more productive to do many steps and document after results are collected but this allows more room for error and details may be forgotten.

2. **When possible, avoid manipulating data manually.**

   It may be challenging to avoid manual data manipulation when it is habit to do simple steps manually (ie: move file, copy and paste, etc) and it may occur without any thought.

3. **Keep track of the names and exact versions of any programs used.**

   If many different programs are used, it may be difficult to keep track of all of them, especially if any programs are updated during the course of research.

4. **Use a version control system to keep track of any edits made to your code.**

   Version control is a very useful tool however, it still requires one to keep track of which version produced which results, which may be challenging.

5. **Keep track of all intermediate results in an organized and standardized fashion.**

   This step may add extra time to research that may feel tedious and uneccesary even though it is important for reproducing results.

6. **Note where randomness is used in the analysis and record the seed used.**

   The biggest challenge with this step may simply be remembering to use a seed and then keeping track of where and what seeds are used.

7. **When a plot is generated, store the raw data behind it (and the code used if possible).**

   A challenge that may arise in this step is storing a data set that is very large or involves sensitive data. It may not be possible to store it locally because of storage size or privacy risks. It may also be a challenge to store the data in a format that could be universally used in the future to reproduce the results.

8. **Detailed underlying data from all main results should be inspected to verify results and then made available for future inspection.**

   This may result in similar challanges to those in step 7 involving the format, size, or sensitivity of the data.

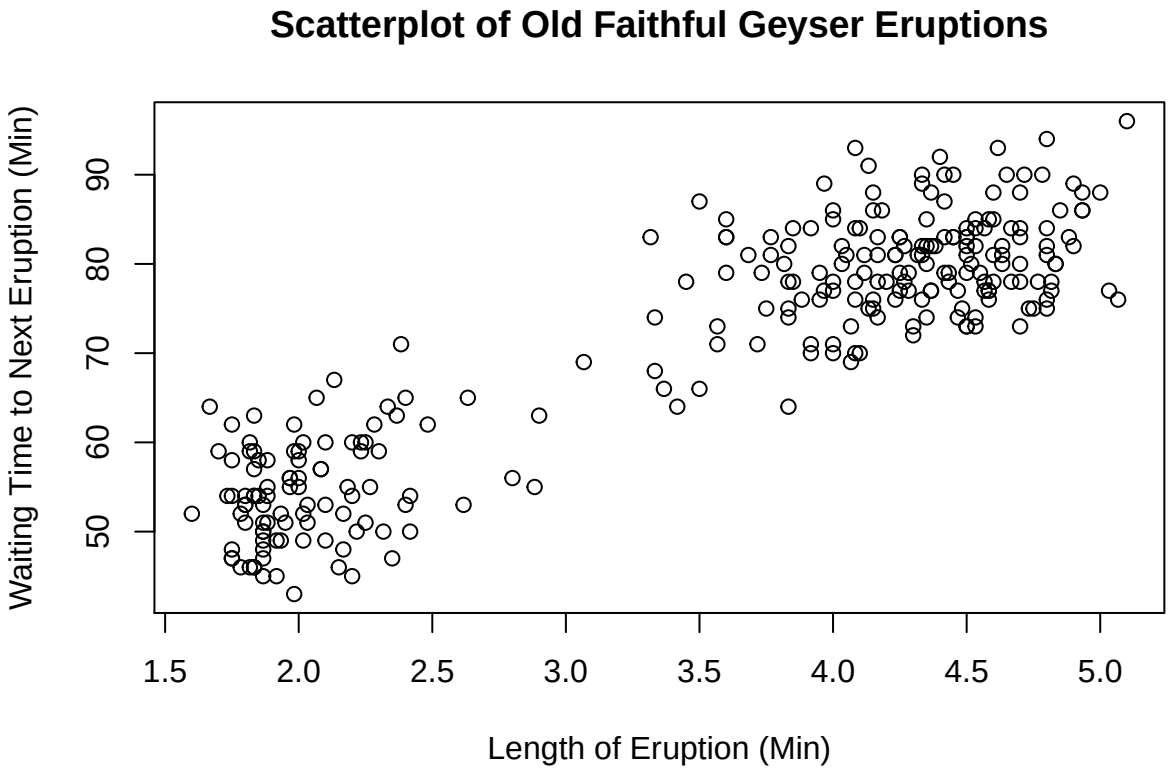
9. **Make direct connections between textual interpretations and specific underlying results along the way instead of at the end of research.**

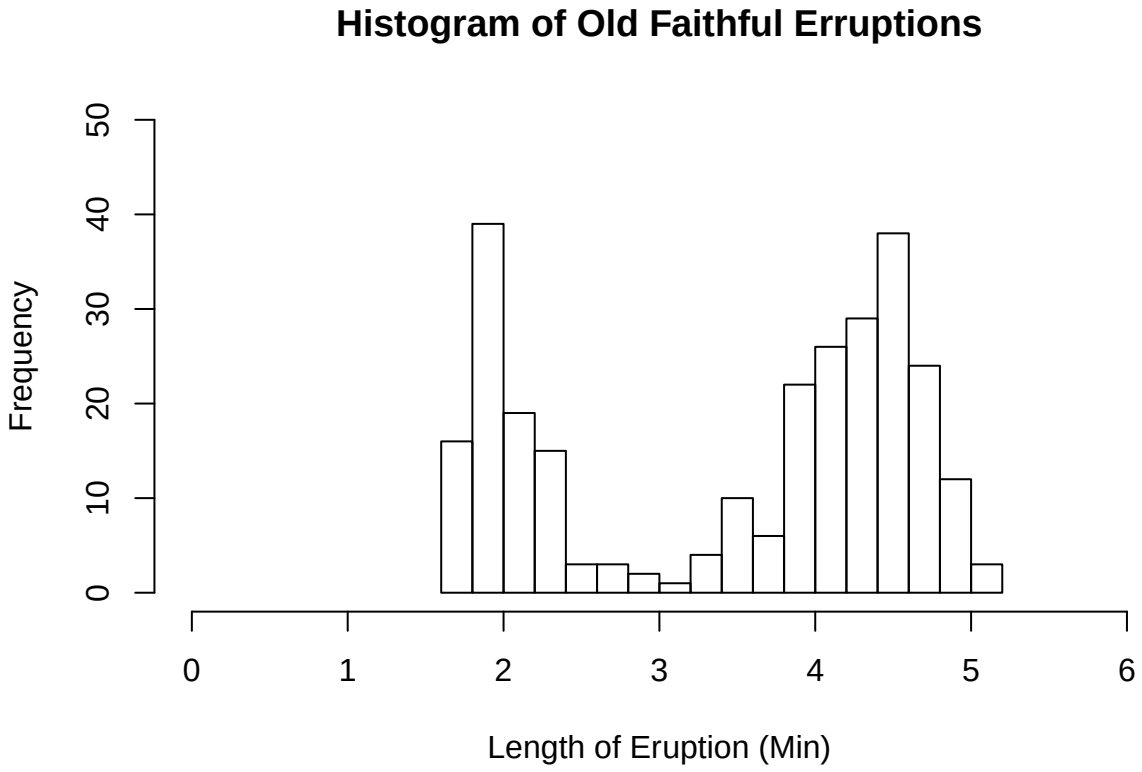
   It may be challenging to make connections to results immediately as it may not be possible to make interpretations until more results are collected.

10. **Make all data, source code, intermediate results, and any other important information on what you did to produce your results publically available along with the main results and analysis.**

    This step requires the researcher to maintain a record of this information in an organized and comprehendable format for it to be easily accessible, which may be challenging. Also, sensitive data may prevet it from being published.

Problem 4


Scatterplot of Old Faithful Geyser Eruptions
Waiting Time to Next Eruption (Min)
50
60
70
80
90
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Length of Eruption (Min)


Histogram of Old Faithful Erruptions
Frequency
0
10
20
30
40
50
0
1
2
3
4
5
6
Length of Eruption (Min)