HW1

Tanner Glass

8/22/2020

Problem 1

No submission required, Primers are complete.

Problem 2

Part A

I am looking forward to my return to Virginia Tech as well as my return to Virginia Tech's computing and coding programs. As an undergraduate in Statistics, I took the time to get an intermediate use of R under my belt. I would like to focus more this year on making my work that I produce more professional as well as making it a publishable standard. As far as content, I am happy to jump into high level practices such as:

- Deep learning tactics
- Bayesian Computing methods vs. frequentist approach
  - Time and efficiency between the two
  - Choosing which approach is most optimal
- Compiling files into a tidy data set and compiling from internet data sources through code. (Webscraping)

Part B

I have included three density functions below. Note the following use the Gamma function which is defined as follows

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$$

I present the Chi squared, F, and Gamma, PDFs respectively as such.

$$f(x|p) = \frac{1}{\Gamma(p/2)2^{(p/2)}}x^{(p/2)-1}e^{-x/2};\ 0 \leq x < \infty;\ p = 1, 2, ... \tag{1}$$

$$f(x|\nu_1, \nu_2) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}(\frac{\nu_2}{\nu_1})^{\nu_1/2}\frac{x^{(\nu_1-2)/2}}{(1+(\frac{\nu_1}{\nu_2})x)^{(\nu_1+\nu_2)/2}};\ 0 \leq x < \infty;\ \nu_1, \nu_2 = 1, 2, ... \tag{2}$$

$$f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta};\ 0 \leq x < \infty;\ \alpha, \beta > 0 \tag{3}$$

Problem 3

From this article regarding reproducible research, taking action to ensure that your data is compatible with other researchers is taking the steps to ensure that a researcher can take your data and produce the exact same results. The article consolidates this ideal down into 10 rules. They are as follows:

1. For Every Result, Keep Track of How It Was Produced
2. Avoid Manual Data Manipulation Steps
3. Archive the Exact Versions of All External Programs Used
4. Version Control All Custom Scripts
5. Record All Intermediate Results, When Possible in Standardized Formats
6. For Analyses That Include Randomness, Note Underlying Random Seeds
7. Always Store Raw Data behind Plots
8. Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
9. Connect Textual Statements to Underlying Results
10. Provide Public Access to Scripts, Runs, and Results

With these rules and why they are suggested is because there can be challenges with the data sets that must be considered. I shall correspond a comment for each of the rules.

1. It can be hard to discover how the data was produced if it isn't outlined with the data set.
2. This could be hard to do if the data frame has a lack of structure or missing data points.
3. Programs can be from different libraries non-native to the base programming. So ensuring that those extra packages are archived can be a challenge.
4. Custom scripts may be hard to decipher by other users
5. We want to record them, but not display EVERY result made through the computation, noting it for those he wish to find it, but not bombarding the researchers with possibly non-important intermediate results.
6. Setting seeds can be a challenge to ensure EVERY random process in your code is set. Its easy to forgot one line that may have a variation, even if you set a seed for another line.
7. If it isn't stored, the research will have to pull in themselves, and it may be edited from when it was originally pulled. Making sure the data set is available for the researcher is a challenge.
8. The challenge is making this hierarchical model and then explaining it for the researcher to understand what is being produced.
9. If the researcher cant decipher your code from what is written, its hard to declare your research as reproducible. The reference should allow and point to the code on what the research is exactly doing, line by line.
10. The challenge here is making your scripts and such public. Sometimes the research has restrictions where it may not be allowed to be made public by a company or firm. The challenge is to making it publicly accessible.

Problem 4

For this problem, I have decided to choose the data set "Trees". The data set contains data regarding black cherry trees. The data set is in a data frame format with 31 observations and 3 variables. They are Girth,

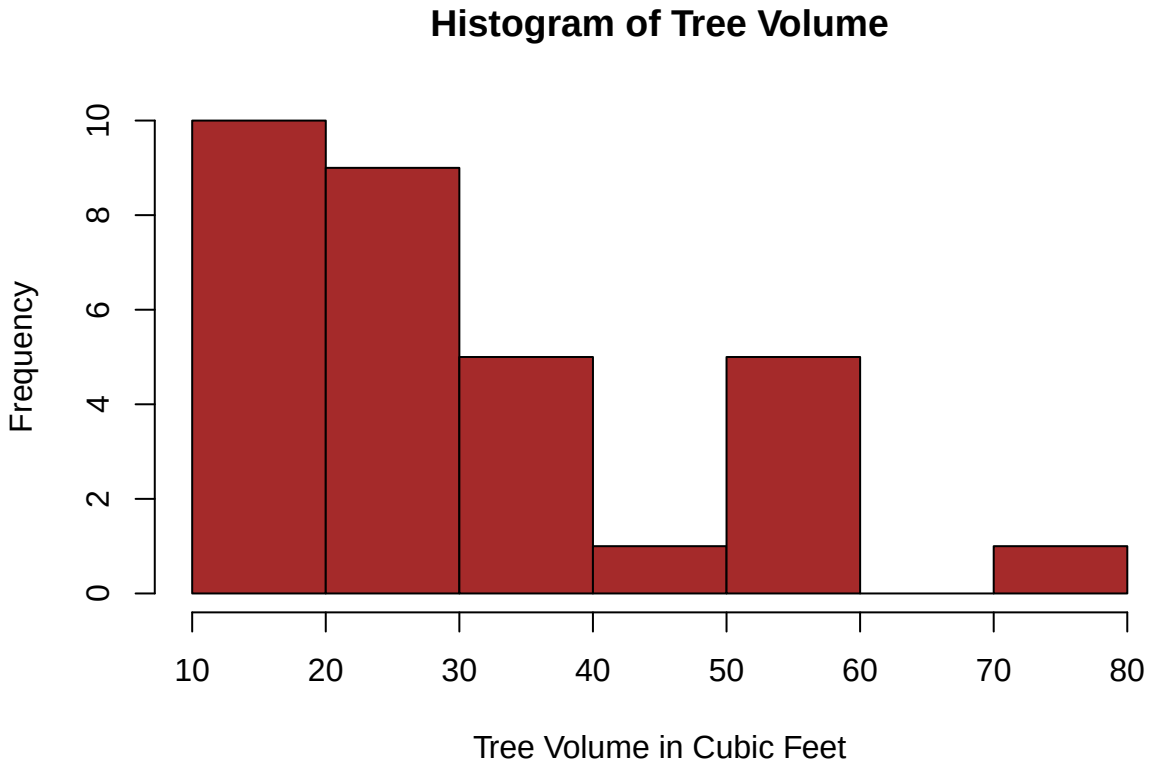
Height, and Volume respectively. I use this data set to create a simple scatter plot and histogram. Let us view the first 10 observations. Then proceed to create our visuals.

Table 1: Cherry Trees

Girth	Height	Volume
8.3	70	10.3
8.6	65	10.3
8.8	63	10.2
10.5	72	16.4
10.7	81	18.8
10.8	83	19.7
11.0	66	15.6
11.0	75	18.2
11.1	80	22.6
11.2	75	19.9

I made this plot to do a quick observation regarding the correlation of height and diameter. There seems to

be a weak positive correlation between the two variables and thus analysis via simple linear regression would be a next step.



As can be seen for the graph, There seems to be a gathering of most trees staying relatively small. Some trees were recorded to be larger which causes a right skew trend.