Homework 1

Due Wednesday Sep 2

Tsering Dolkar

8/26/2020

Problem 1

R is an open source, community built, programming platform. Not only is there a plethora of useful web based resources, there also exists in-R tutorials. As per the instructions, I did both of the Primers labeled as The Basics on Rstudio.cloud.

Problem 2

After having the R environment setup, with the help of a basic understanding of R, I created this Markdown, saved the file to the directory containing the *README.md* file.

Part A

Although I had a brief experience with R before, I have not studied it from the basics properly and therefore, had problems understanding codes during certain instances in the past. With this class, I hope to have a more innate understanding of the language and to have a certain level of comfort using it in the future. Desired learning objectives:

1. Better command and understanding of R
2. Better understanding of statistical simulations in R
3. Hopefully, learn a bit about SAS

Part B

(1) *The density of the sum of two normals*:

$$If[Re[z] < 0, \frac{e^{-\frac{z^2}{4}}}{2\pi}, \int_{-\infty}^{\infty} \frac{E^{-\frac{y^2}{2}-\frac{1}{2}(-y+z)^2}}{2\pi} dy]$$

(2) *The density of the sum of two Cauchys*:

$$\frac{2}{\pi(-2I+z)(2I+z)}$$

(3) *The density of the sum of two t's with 5 degrees of freedom*:

$$\frac{400\sqrt{5}(8400+120z^2+z^4)}{3\pi(20+z^2)^5}$$

Problem 3

Ten Simple Rules for Reproducible Computational Research:

1. **For every result, keep track of how it was produced:**
   It is important to keep track of steps from raw data to final result when the result might be of future interest. While doing so, we will be recording a sequence of steps that are interrelated. This sequence of steps is called an analyses work flow.

   - **Challenge:**
     It is important to record every detail influencing the execution of the steps, including the name and version of the program involved. As diligent as one might be, it is always possible to overlook minute details or be unable to access the said version of the program used when trying to reproduce the research.

2. **Avoid Manual Data Manipulation Steps:**
   Due to inefficiency and possible inaccuracy, rely on execution of programs instead of manual procedures to modify data.

   - **Challenge:**
     While working with integrated frameworks, i.e. a collection of components for data manipulation and format converters is helpful, execution programs don't always run perfectly on repeated implementation.

3. **Archive the Exact Versions of All External Programs Used:**
   To reproduce a concerned result exactly, it is important archive the exact version of program used. There is always the possibility that both input and output formats may change between versions. A newer version of a program may not even run without modifying its inputs.

   - **Challenge:**
     We need to be certain the version of the program used in the original research, specific requirements to other installed packages and possible operating system influence are carefully recreated to fulfill this rule.

4. **Version Control All Custom Scripts:**
   It is important to systematically archive computer codes along its evolution. When a continually developed piece of code (typically a small script) has been used to generate a certain result, only that exact state of the script may be able to produce that exact output, even given the same input data and parameters.

   - **Challenge:**
     As often as you might try to archive copies of your script in Git, there can only be a rough record of the evolution. It is possible to makes changes to your code and to forget to archive all the changes to Git.

5. **Record All Intermediate Results, When Possible in Standardized Formats:**
   Quickly browsing through intermediate results can reveal discrepancies toward what is assumed, and can in this way uncover bugs or faulty interpretations that are not apparent in the final results.

   - **Challenge:**
     Might run into prohibitive storage space when one tries to archive all result files produced while running an analysis.

6. **For Analyses That Include Randomness, Note Underlying Random Seeds:**
   Many analyses and predictions include some element of randomness.For analyses that involve random numbers, this means that the random seed should be recorded.

   - **Challenge:**
     While achieving equal results is a strong indication that a procedure has been reproduced exactly, it is often hard to conclude anything when achieving only approximately equal results.

7. **Always Store Raw Data behind Plots:**
   If raw data behind figures are stored in a systematic manner, so as to allow raw data for a given figure to be easily retrieved, one can simply modify the plotting procedure, instead of having to redo the whole analysis.

   - **Challenge:**
     It is important to pay attention to how your raw data is stored so that it can be easily retrieved.

8. **Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected:**
   For instance, each value along a curve may in turn represent averages from an underlying distribution. When the storage context allows, it is better to simply incorporate permanent output of all underlying data when a main result is generated, using a systematic naming convention to allow the full data underlying a given summarized value to be easily found.

   - **Challenge:**
     The method mentioned could be impractical and inefficient usage of storage space.

9. **Connect Textual Statements to Underlying Results:**
   The results of analyses and their corresponding textual interpretations often involve viewing the results in light of other theories and results.Therefore, it is important to connect a given textual statement (interpretation, claim, conclusion) to the precise results underlying the statement.You can use tools such as Sweave, the GenePattern Word add-in, and Galaxy Pages to achieve that.

   - **Challenge:**
     Making this connection when it is needed may be difficult and error-prone, as it may be hard to locate the exact result underlying and supporting the statement from a large pool of different analyses with various versions.

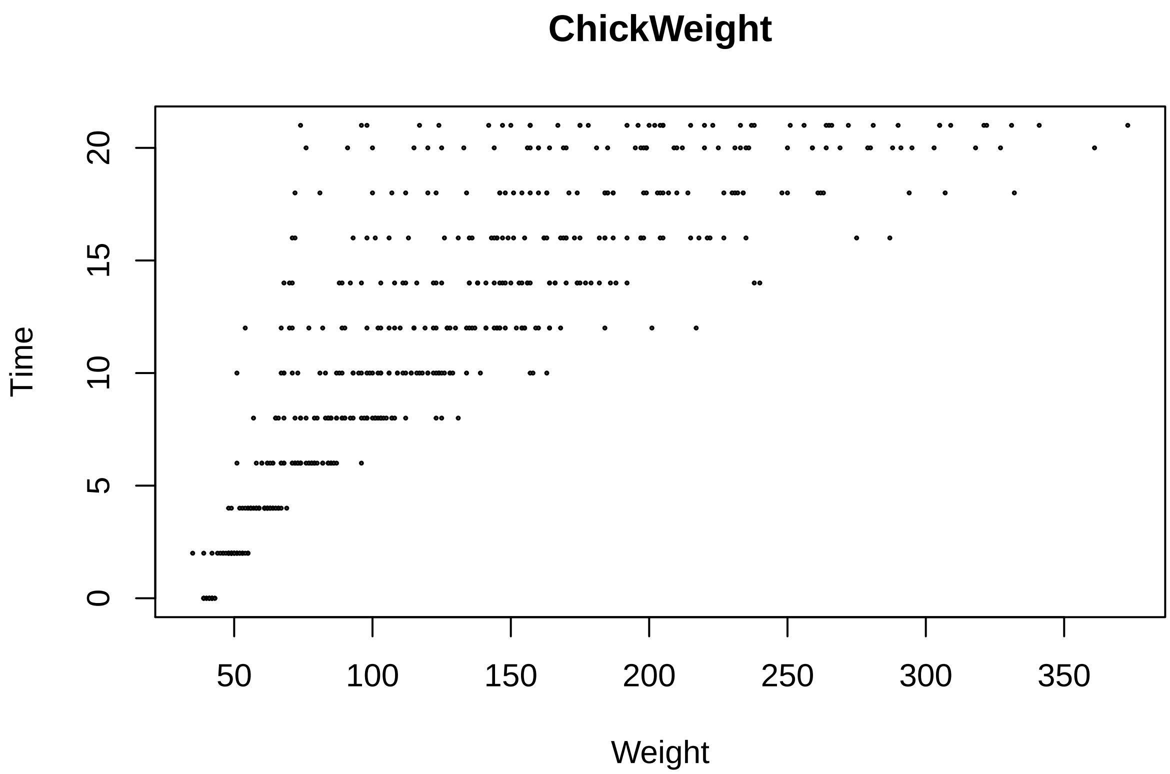
10. **Provide Public Access to Scripts, Runs, and Results:**
    All input data, scripts, versions, parameters, and intermediate results should be made publicly and easily accessible.

    - **Challenge:**
      Such transparency requires excellent quality, and confidence in one's research.

Problem 4

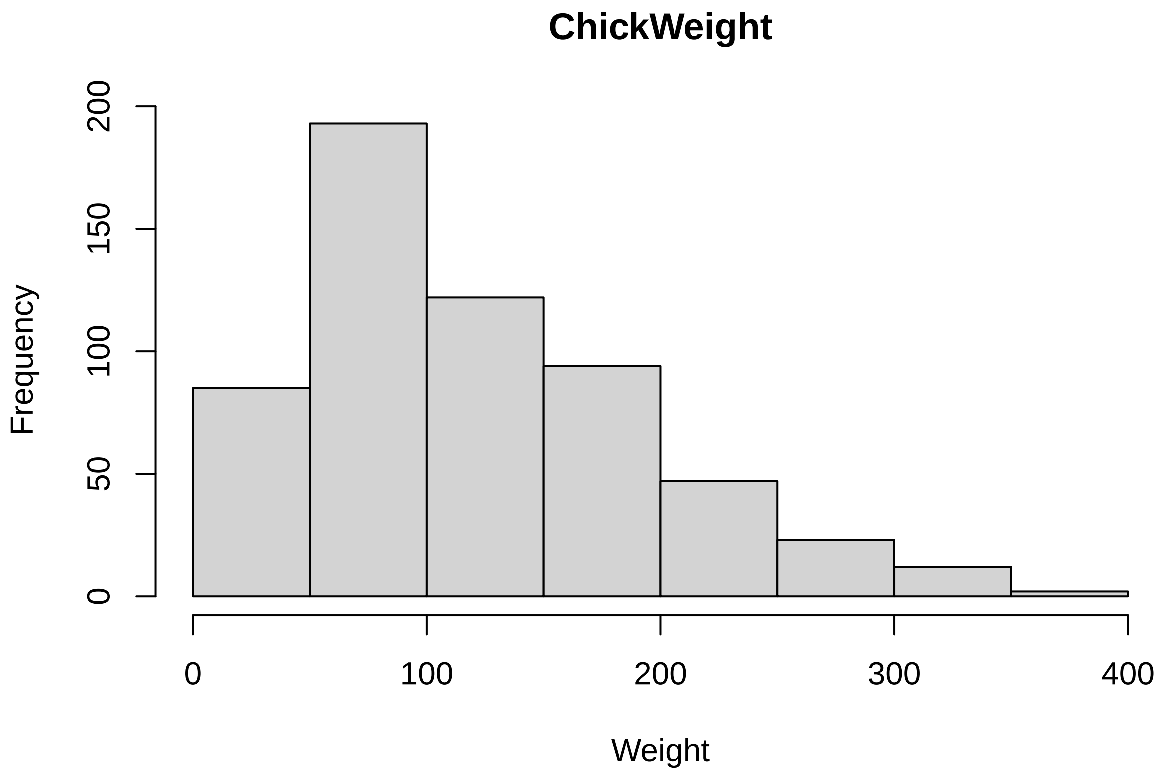
Following is a basic scatter plot:

```r
plot(ChickWeight$weight,ChickWeight$Time, cex = 0.2, xlab = "Weight", ylab = "Time",
     main = "ChickWeight")
```



Following is a histogram:

```r
hist(ChickWeight$weight, xlab = "Weight", ylab = "Frequency",
     main = "ChickWeight")
```


ChickWeight
Frequency
Weight