HW1_sliang

Liang Shi

8/29/2020

Problem 2

Part A

- I used to be a computer engineer, and I find that sometimes the coding does not follow the norms at work. So I just wanna know more about the code specification from this class.
- I used python for data analysis before, but sometimes I find that the python packages for statistics (such as statmodels and sklearn) are not as powerful as R is. Many of the api is not available in python. Thus, I think it will give great help if I'm adept in R.
- I have learned data structure and algorithm before. I think it's a good chance to review them in this class.

Part B

$$f(x \mid \mu, \sigma) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0 \tag{1}$$

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

$$f(x \mid \mu, \beta) = \frac{1}{\beta} \frac{e^{-(x-\mu)/\beta}}{\left[1 + e^{-(x-\mu)/\beta}\right]^2}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \beta > 0 \tag{3}$$

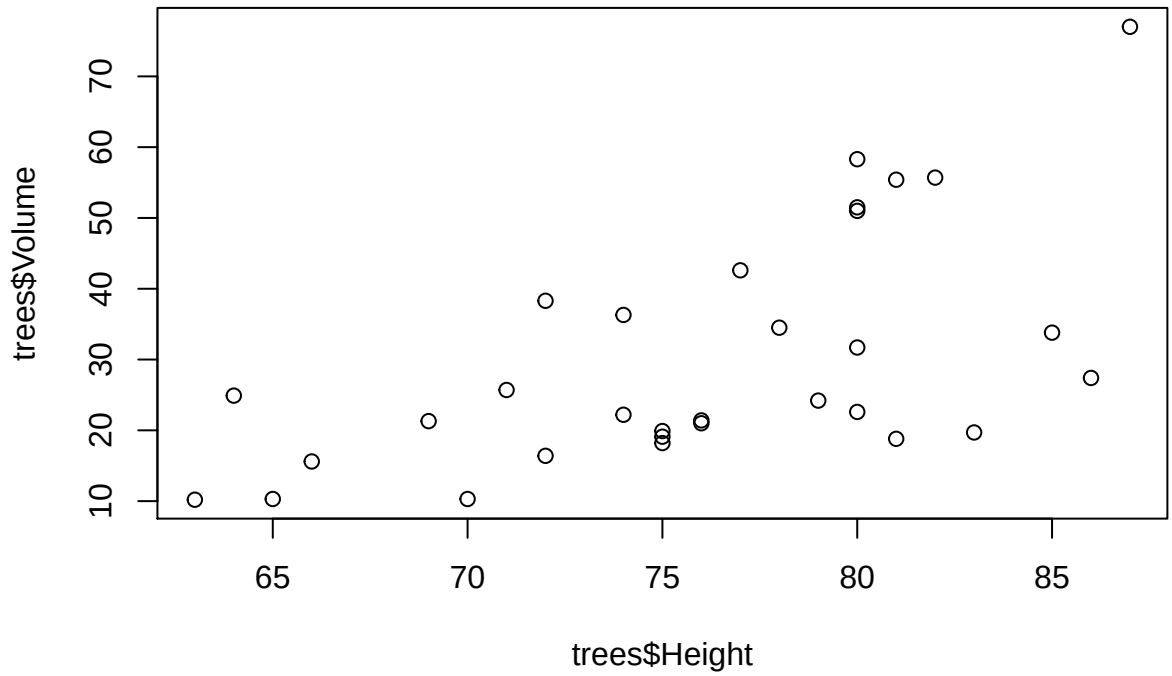
Problem 3

- For every result, keep track of how it was produced
  - I think understanding the detailed operation process of the program is one of the basic qualities of a programmer
- Avoid manual data manipulation steps
  - This is very useful in data analysis, because sometimes we have lots of data. If we do this manually, it will be ineffective and error-prone
- Archive the exact versions of all external programs used
  - I have this experience. My colleague and me use two branches to work together. But we use the similar name about the version. So that when we merge the code, it occurs many bugs.
- Version control all custom scripts
  - Version control is very important because our code is composed of lots of modules. If some version changes, it is very likely to occur bugs.
- Record all intermediate results, when possible in standardized formats
  - This a very useful in debugs, which can help us focus the bug quickly.
- For analyses that include randomness, note underlying random seeds
  - This happend when I build a deep neural network. After I tune the hyperparameters many times. I find that the randomness make me diffcult to know whether the current hyperparameters are better or not. It is essential to fix the randomness by random seeds.

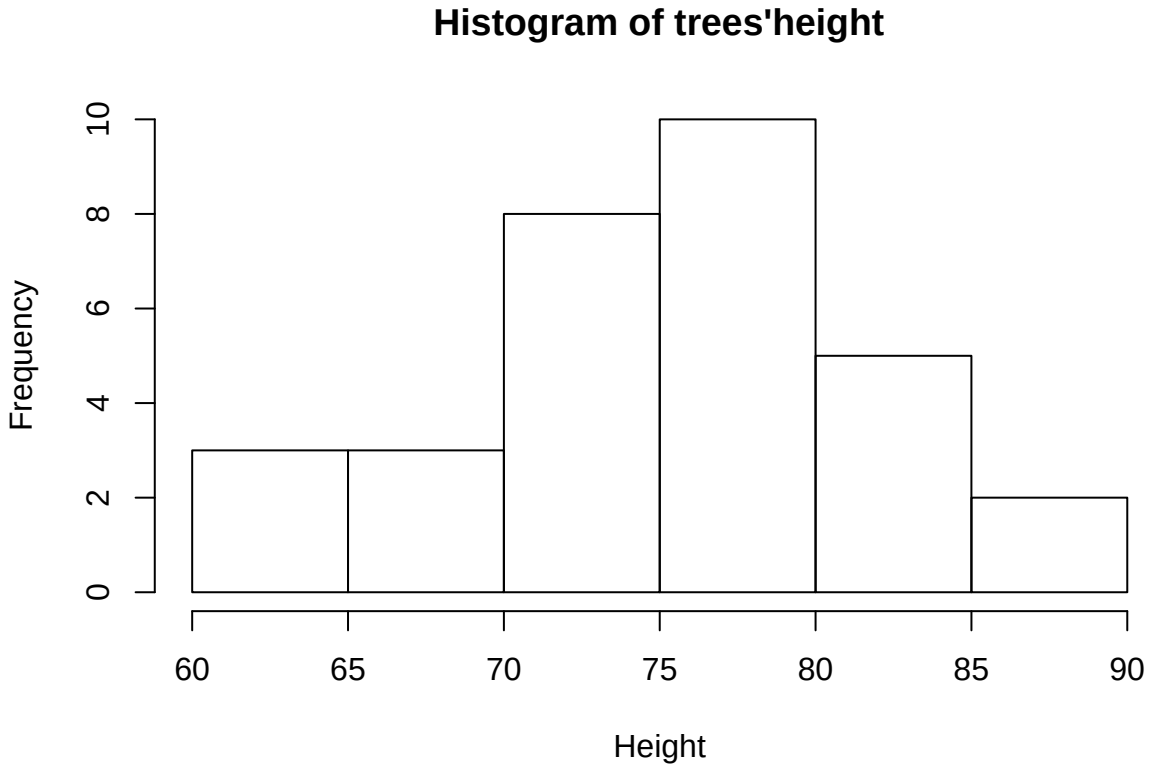
- Always store raw data behind plots
  - I have made a website to show the data for VTTI. I haven't store the raw data of one boxplot. This makes me did a lot of extro work when the changed the data.
- Generate hierarchical analysis output, allowing layers of increasing detail to be inspected
  - It is very common that the paper will be revised, and the data and experiments will be revised, too. Thus, it is essential to know the data underlying the tables and plots.
- Connect textual statements to underlying results
  - This is very important in collaborative development. Such connection can give clearer instructions to our collaborators.
- Provide public access to scripts, runs, and results
  - Many papers will publish their data and code to github. Usually such articles will have more citations than articles with hidden data and code.

Problem 4

```
library(help="datasets")
plot(trees$Height,trees$Volume)
```




```
hist(trees$Height,xlab="Height",main="Histogram of trees'height")
```


Histogram of trees'height
Frequency
0
2
4
6
8
10
60
65
70
75
80
85
90
Height