Homework 17

Youhui Ye

8/24/2020

Problem 1

Part A

The things I wanted to learn from this course including:

- How to collaborate with my teammates using Github;
- How to create my own R package;
- How to conduct statistical learning via R.

Part B

Beta(α, β)

$$f(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \le x \le 1, \quad \alpha > 0, \quad \beta > 0 \tag{1}$$

Exponential(β)

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

Gamma(α, β)

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

Problem 3: Summary

Ten Simple Rules for Reproducible Computational Research

- **For each result, keep track of how it was produced.** When we are very devoted to producing nice results for our projects, we sometimes forget to record how we get them.
- **Avoid manual data manipulation steps.** Sometimes manual modifications can be more convenient than writing a command script or programming. And avoiding manual steps might cause a lot more workload.
- **Archive the exact versions of all external programs used.** Previous code files may not be executable in newest versions. Even though we recorded what we used, it might still be hard to reproduce the results.
- **Version control all custom scripts.** R is an open source software, which means its computational rules change by individual.

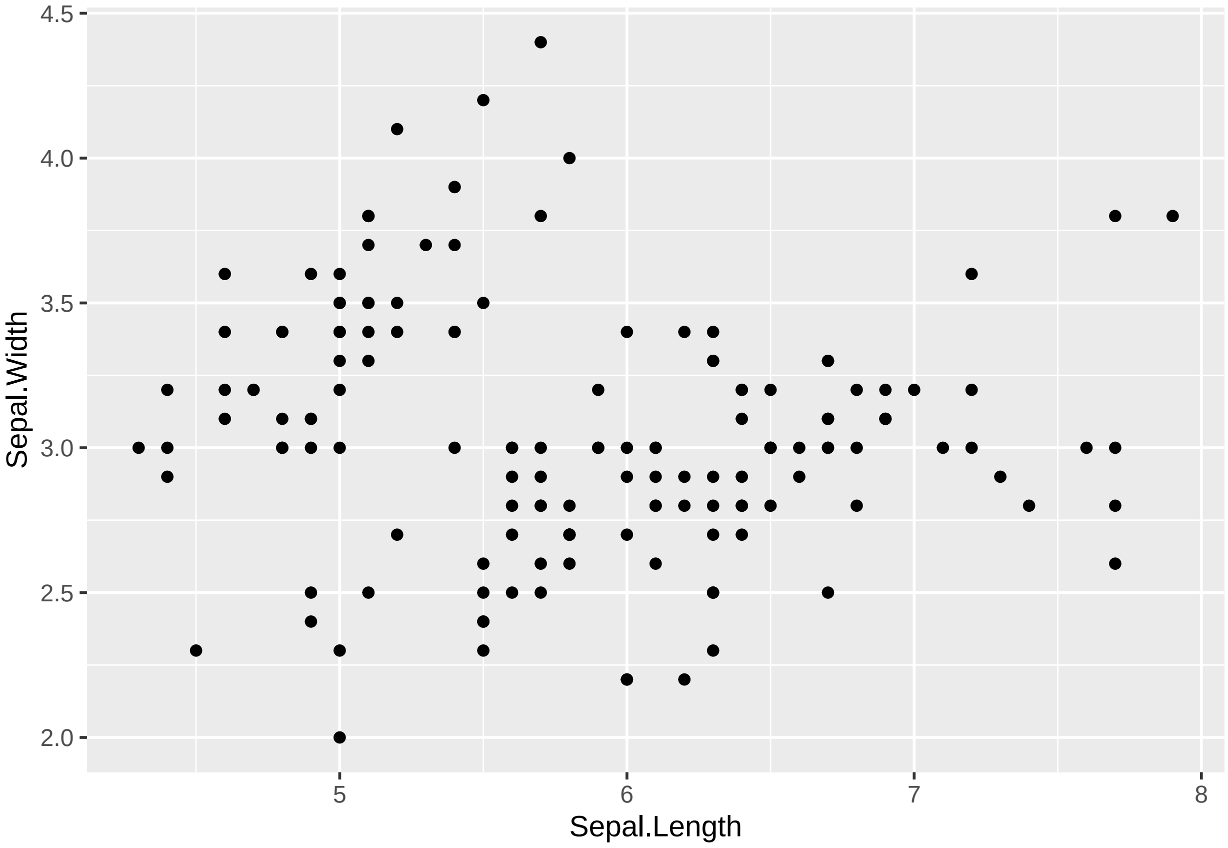
- **Record all intermediate results, when possible in standardized formats.** If the data is massive, it is hard to store intermediate results in each step.
- **For analyses that include randomness, note underlying random seeds.** Sometimes we have good results only when we use a certain random seed.
- **Always store raw data behind plots.** If multiple plots are produced in one project, we may need large space to store them.
- **Generate hierarchical analysis output, allowing layers of increasing detail to be inspected.** In case like machine learning, some intermediate details are not accessible.
- **Connect textual statements to underlying results.** Different statements can be produced on same plot depending on person and version. Thus, choosing the best might be a problem.
- **Provide public access to scripts, runs, and results.** Thanks to Internet, providing scripts and results to public is relatively easy today. However, it is still hard to provide large data set.

Problem 4

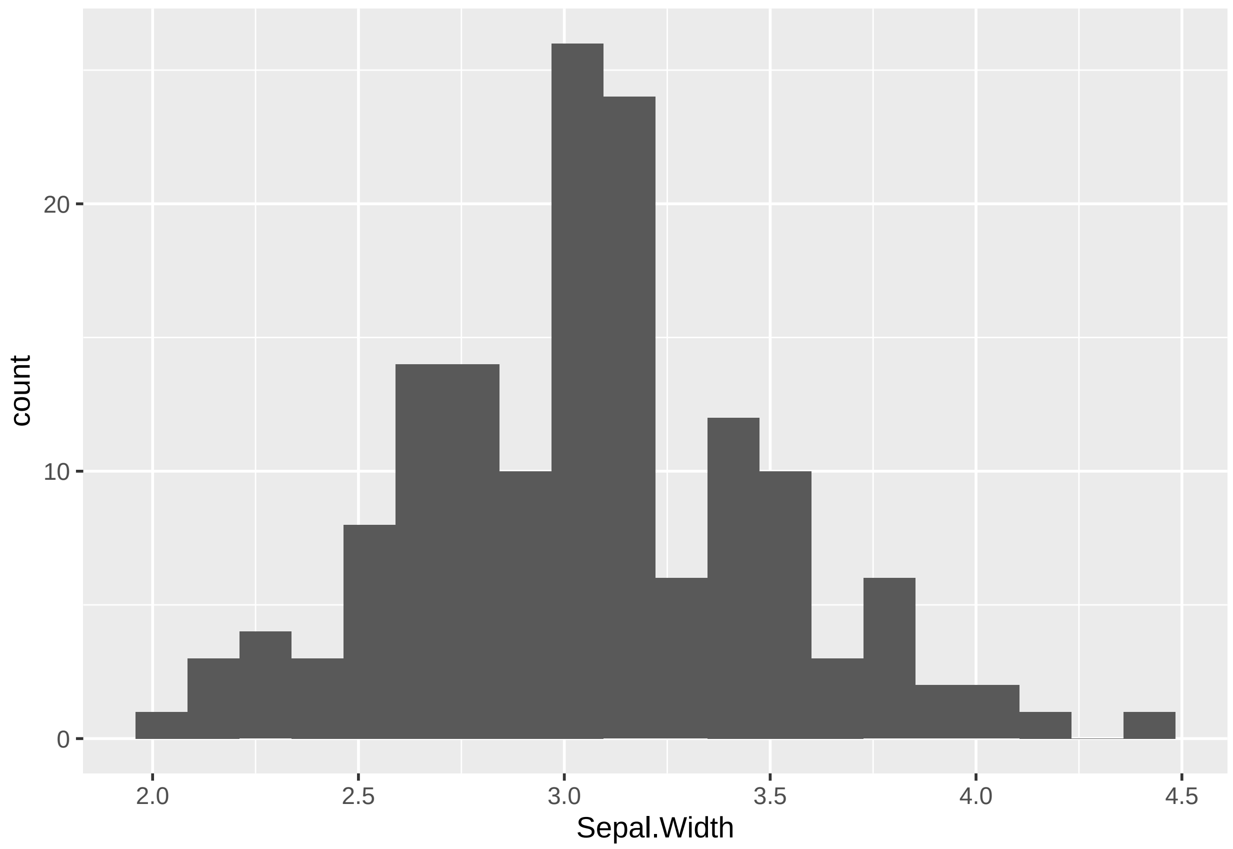
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
df <- iris
ggplot(data = df, mapping = aes(x = Sepal.Length, y = Sepal.Width)) + geom_point()
```



```
ggplot(data = df, mapping = aes(Sepal.Width)) + geom_histogram(bins = 20)
```


count
20
10
0
2.0
2.5
3.0
3.5
4.0
4.5
Sepal.Width