STAT 5361: Statistical Computing, Fall 2018

Jun Yan

2018-08-31

Contents

Chapter 1

Prerequisites

We assume that students can use R/Rstudio comfortably. If you are new to R, an extra amount of hard work should be devoted to make it up. In addition to the official documentations from the R-project, the presentations at a [local SIAM workshop in Spring 2018] (https://siam.math.uconn.edu/events/) given by Wenjie Wang, a PhD student in Statistics at the time, can be enlighting:

- https://github.com/wenjie2wang/2018-01-19-siam
- https://github.com/wenjie2wang/2018-04-06-siam

If you have used R, but never paid attention to R programming styles, a style clinic would be a necessary step. A good place to start is Google's R style guide at https://google.github.io/styleguide/Rguide.xml. From my experience, the two most commonly overlooked styles for beginners are spacing and indentation. Appropriate spacing and indentation would immediately make crowdly piled code much more eye-friendly. Such styles can be automatically enforced by R packages such as **formatr** or **lintr**. Tow important styles that cannot be automatically corrected are naming and documentation. As in any programming languate, naming of R objects (variables, functions, files, etc.) shoud be informative, concise, and consistent with certain naming convention. Documentation needs to be sufficient, concise, and kept close to the code; tools like R package **roxygen2** can be very helpful. See Hadley Wichham's online book http://style.tidyverse.org/ for more detailed tips.

For intermediate R users who want a skill lift, The **Advanced R Programming** book by Hadley Wickham is available at https://adv-r.hadley.nz/. The source that generated the book is kindly made available at GitHub: https://github.com/hadley/adv-r. It is a great learning experience to complile the book from the source, during which you may pick up many necessary skills.

The homework, exam, and project will be completed by **R Markdown**. Following the step by step the instructions in Yihui Xie's online book on **bookdown** at https://bookdown.org/yihui/bookdown/, you will be amazed how quickly you can learn to produce cool-looking documents and even book manuscripts. If you are a keener, you may as well following Yihui's **blogdown**, see online book https://bookdown.org/yihui/blogdown/, to build your own website using R Markdown.

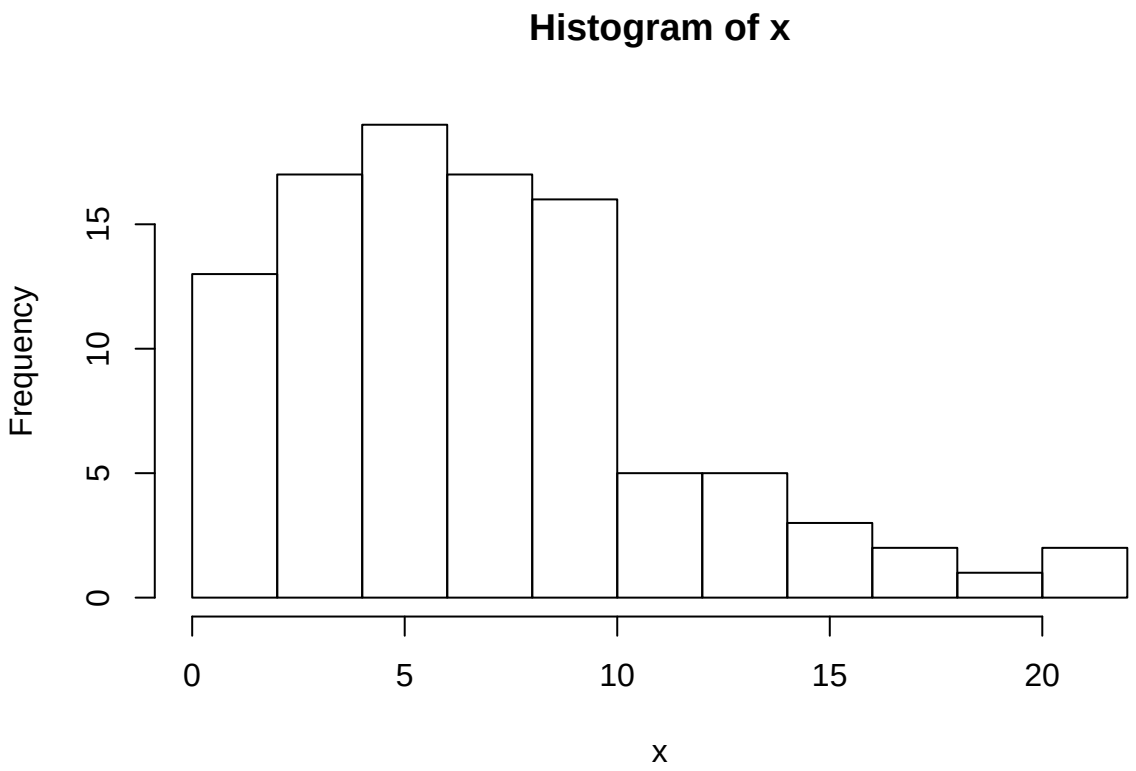
All your source code will be version controlled by **git** and archived on **GitHub**. RStudio has made using git quite straightforward. The online tutorial by Jenny Bryan, Happy Git and GitHub for the useR, is a very useful tool to get started

1.1 A Teaser Example: Likelihood Estimation

In mathematical statistics, we have learned that, under certain regularity conditions, the maximum likelihood estimator (MLE) is consistent, asymptotically normal, and most efficient. The asymptotic variance of the

estimator if the inverse of the Fisher information matrix. Specifically, let $X_1, \ldots, X_n$ be a random sample from a distribution with density $f(x; \theta)$, where θ is a parameter. How do we obtain the MLE?

```
set.seed(123)
n <- 100
x <- rgamma(n, shape = 2, scale = 4)
hist(x)
```



Package **MASS** provides a function `fitdistr()` to obtain the MLE for univariate distributions with a random sample. We can learn two things from this function. First, an objective function representing the negative loglikelihood is formed, depending on the input of the density function, and fed to the optimizer function `optim`. Second, the variance estimator of the MLE is obtained by inverting the Hessian matrix of the objective function, which is an estimator of the Fisher information matrix. For commonly used distributions, starting values are not necessary. The function computes moment estimator and use them as starting values.

```
MASS::fitdistr(x, densfun = "gamma")
```

```
## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced
```

```
##      shape         rate
##   2.19321713   0.31850407
##  (0.28968120) (0.04724651)
```

For distributions not in the provided list, the `densfun` needs to be a function returning a density evaluated at its first arguments, in which case, the `start` argument needs to be a named list to feed to the optimizer. For example, pretend that someone has a fancy distribution which is just a rename of the gamma distribution. Its density function is defined based on the gamma density `dgamma`.

```
dfancy <- function(x, shape, scale, log = FALSE) {
  dgamma(x, shape = shape, scale = scale, log = log)
}

suppressWarnings(MASS::fitdistr(x, densfun = dfancy, start = list(shape = 10, scale = 20)))
```

```
##      shape       scale
##   2.1931194   3.1400457
##  (0.2897513) (0.4659625)
```

The **stats4** package provides MLE using S4 classes.

```
nll <- function(shape, scale) -sum(dfancy(x, shape, scale, TRUE))
suppressWarnings(fit <- stats4::mle(nll, start = list(shape = 10, scale = 10)))
stats4::summary(fit)
```

```
## Maximum likelihood estimation
##
## Call:
## stats4::mle(minuslogl = nll, start = list(shape = 10, scale = 10))
##
## Coefficients:
##       Estimate Std. Error
## shape 2.193213  0.2896847
## scale 3.139686  0.4657464
##
## -2 log L: 557.1586
```

1.2 Exercises

1. Use git to clone the source of Hadley Wickham's **Advanced R Programming** from his GitHub repository to a local space on your own computer. Build the book using RStudio. During the building process, you may see error messages due to missing tools on your computer. Read the error messages carefully and fix them, until you get the book built. You may need to install some R packages, some fonts, some latex packages, and some building tools for R packages. On Windows, some codes for parallel computing may not work and need to be commented out.

2. Use **bookdown** or **rmarkdown** to produce a report for the following task. Consider approximation of the distribution function of $N(0,1)$,

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy, \tag{1.1}$$

   by the Monte Carlo methods:

$$\hat{\Phi}(t) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \le t), \tag{1.2}$$

   where X_i's are iid $N(0,1)$ variables. Experiment with the approximation at $n \in \{10^2, 10^3, 10^4\}$ at $t \in \{0.0, 0.67, 0.84, 1.28, 1.65, 2.32, 2.58, 3.09, 3.72\}$ to form a table. The table should include the true value for comparison. Further, repeat the experiment 100 times. Draw box plots of the bias at all t. The report should look like a manuscript, with a title, an abstract, and multiple sections. It should contain at least one math equation, one table, one figure, and one chunk of R code. The template of our Data Science Lab can be helpful: https://statds.org/template/, the source of which is at https://github.com/statds/dslab-templates.

Chapter 2

Introduction

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 2. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter 3.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 2.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 2.1.

```
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (Xie 2018) in this sample book, which was built on top of R Markdown and **knitr** (Xie 2015).

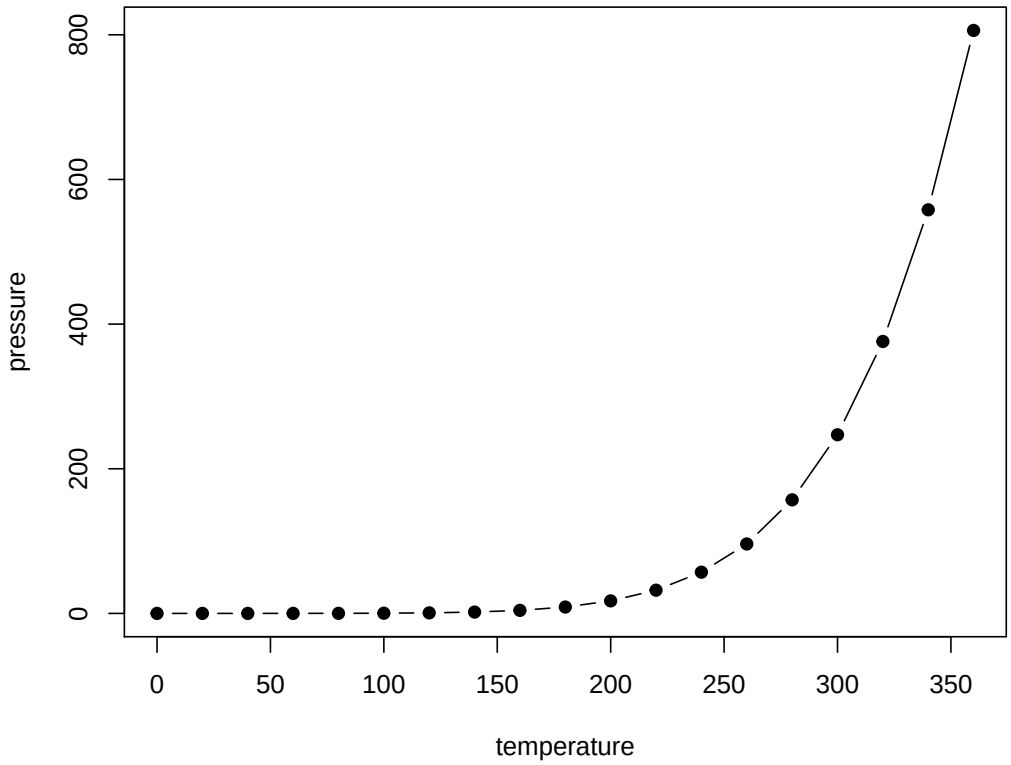


Figure 2.1: Here is a nice figure!

Table 2.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

Chapter 3

Optimization

This chapter is on optimization.

Chapter 4

EM Algorithm

EM has its own chapter for its importance in statistics.

Chapter 5

Random Number Generation

Simulation basics.

Chapter 6

Markov Chain Monte Carlo

Someone may need this for the course project.

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. http://yihui.name/knitr/.

———. 2018. *Bookdown: Authoring Books and Technical Documents with R Markdown*. https://CRAN.R-project.org/package=bookdown.