Chapter 0

A Quick Introduction to R

This chapter will introduce the datasets, programs and software used in this course. In particular, R will be discussed in more details. It is important for students not familiar with R. First let us starts with the filenames of datasets and programs.

Datasets and Programs

The datasets used in this course have filenames like ***<file>.csv***. The ***csv*** is called the file extension and is the abbreviation of **Common Separated Value**. It is a very common type of data format and can be read by almost all software, including R, EXCEL and SAS. The R programs have extension ***<program>.r***. It is in ***ASCII*** (American Standard Code for Information Interchange) format. Again it is a very common file format and can be read by all text editors, including the built-in editor Notepad. Please download these files and save them in a folder, say ***stat5104/data/***, so that all these files can be read later on.

Packages

In this course, SAS, EXCEL and R will be used for demonstration. There are separate notes on using SAS. However, R will be used mainly in this course and explained in details. The computations are illustrated using R programs. The same calculations will be done using EXCEL whenever possible. These EXCEL files are of filenames ***<excel>.xls***. It is important to get familiar with R as soon as possible. The following section will provide a quick introduction to R.

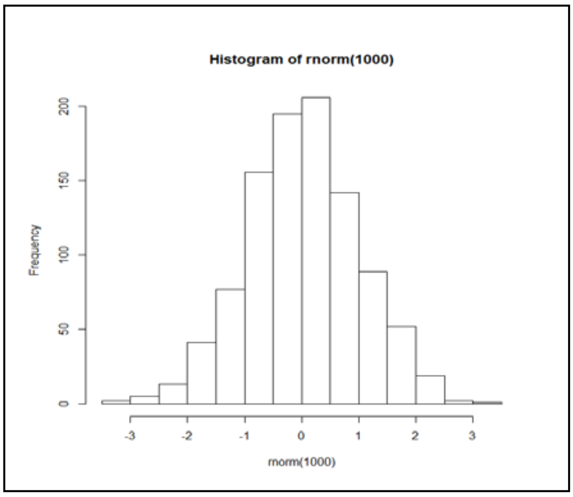
R package

R is a statistical package and programming language which is widely used in the Data Science community. It is an open source freeware and can be downloaded from the official website: www.r-project.org. R is very flexible, powerful and easy to use. Best of all, there are many libraries have been developed by professional statistician all over the world and can be download as well. First, let us download and install the R package. If everything goes smoothly, you should see an R icon on the desktop. Double click this icon to start the R package. You will see a **session window** with the prompt sign >, waiting for the R commands to execute.

Once you begin your R session, it is highly recommended to change the default working directory to your own folder, say ***stat5104/data/*** as soon as possible so that you can read in data and save your output in that folder. On the menu bar, click ***File*** and choose ***change dir…***, and you will see a pop-up window. Choose the folder, say ***stat5104/data/.*** To ensure that you are in the right folder, enter ***dir()*** and you will see the all the files in this folder.

```
> dir()
 [1] "ann.r"       "bank.csv"     "bank.r"       "ch1-dm.r"
 [5] "ch2-ctree.r" "ch3-knn-nb.r" "ch4-lreg.r"   "ch5-ann.r"
 [9] "ch6-clus.r"  "ch7-assoc.r"  "hclus-ex.csv" "hmeq.csv"
[13] "hmeq1.csv"   "iris.csv"     "k_nn.r"       "km.r"
[17] "mdist.r"     "scale.r"      "stand.r"      "titanic.csv"
```

R can be used as a scientific calculator, for example, try to enter: *>1+2*3/4, a<-exp(1)* and *log(a)* and look at the results. There are many built-in functions as well, for example, try to enter: *>hist(rnorm(1000)*. You will see a histogram of 1000 Normal(0,1) random numbers in the **graphic window** as follow:



Now you can right click the graphic window and select ***copy as bitmap*** to save this graph. Then start a word document to paste this graph. Actually this is exactly how I produce this note. There is **online help** for all built-in functions. For example, try ***help(log), help(exp), help(hist)*** and ***help(rnorm)*** to see if you can understand the explanation.

You can save your all your R commands and R outputs, say ***ch0.r*** and ***ch0.txt***, respectively. In the main menu, click ***File-> Save History…*** and enter the filename ***ch0.r***. The file ***ch0.r*** is created, containing all the R commands in the R session window. This is useful as you can save all your R commands to re-run them later on. Click ***File-> Save to file…*** and enter the filename ***ch0.txt***, the file ***ch0.txt*** contains all the R commands and the outputs. This is useful when you want to save your output and include them in your report. You can execute R commands in a file, say ***ch0.r***, by entering ***source('ch0.r', echo=T)***. All the R commands in the following chapters are saved in files and can be executed in a similar way.

Finally you can quit the R session by entering ***quit()*** or clicking the close sign on the upper right hand corner of the window. You will be asked whether you want to save the workspace image each time when you quit. It is recommended **not to save** since we do not want the variable in the current session to be carried forward when we start the R package next time.

That is a quick summary and should be enough for you to get started with R. For those who want to learn more about R, click ***Help-> Manuals (in PDF)*** in the main menu. There is a concise and well-written document: *An introduction to R*.