Exploring data using R

Kamarul Imran Musa, Wan Nor Arifin

2017-05-22

Contents

Chapter 1

Introduction to R

1.1 Installing R and RStudio

Install R base package: http://www.r-project.org/

Install RStudio: http://www.rstudio.com/

1.2 Getting familiar with the interface

Consists of 4 tabs: 1. Source 2. Console 3. Environment & History 4. Misc. Most important Plots, Packages & Help

1.3 Basic tasks in R

1.3.1 R Script

Text here.

1.3.2 Setting working directory

Text here.

1.3.3 Packages

Text here.

1.3.3.1 Installation

```
install.packages("package.name")
```

1.3.3.2 Loading

```r
library("package.name")
```

1.3.4 Data management

Text here.

1.3.4.1 Loading data

```r
read.csv("file.name")
```

For SPSS file, need `foreign` package

```r
library("foreign")
read.spss("file.name")
```

1.3.4.2 Data dimension

```r
dim(data)
```

1.3.4.3 Entering data

text here

1.3.4.4 Editing data

text here

Chapter 2

Textual

In this chapter, we will go through a number of R functions for basic statistics. We will mostly use the builtin functions (from R standard library). Extra packages will be introduced whenever necessary.

2.1 Descriptive statistics

We are going to use builtin datasets in R. You can view the available datasets by

```
data()
```

```
## Data sets in package 'datasets':

## AirPassengers                        Monthly Airline Passenger Numbers 1949-1960
## BJsales                              Sales Data with Leading Indicator
## BJsales.lead (BJsales)               Sales Data with Leading Indicator
## BOD                                  Biochemical Oxygen Demand
## CO2                                  Carbon Dioxide Uptake in Grass Plants
## ...
```

View the data, for example

```
women
```

```
##    height weight
## 1      58    115
## 2      59    117
## 3      60    120
## 4      61    123
## 5      62    126
## 6      63    129
## 7      64    132
## 8      65    135
## 9      66    139
## 10     67    142
## 11     68    146
## 12     69    150
## 13     70    154
## 14     71    159
## 15     72    164
```

View the dimension, i.e. number of subjects and variables

```
dim(women)
```

```
## [1] 15  2
```

Obtaining mean

```
mean(women$weight)
```

```
## [1] 136.7333
```

and median

```
median(women$weight)
```

```
## [1] 135
```

and sd

```
sd(women$weight)
```

```
## [1] 15.49869
```

and IQR

```
IQR(women$weight)
```

```
## [1] 23.5
```

There 9 types of IQR in R, the default one is type 7. You may change this to type 6 (Minitab and SPSS),

```
IQR(women$weight, type = 6)
```

```
## [1] 27
```

and minimum, maximum and range

```
min(women$weight)
```

```
## [1] 115
```

```
max(women$weight)
```

```
## [1] 164
```

```
range(women$weight)
```

```
## [1] 115 164
```

However, it is actually simpler to obtain most these in one single command for both weight and height

```
summary(women)
```

```
##      height          weight     
##  Min.   :58.0   Min.   :115.0  
##  1st Qu.:61.5   1st Qu.:124.5  
##  Median :65.0   Median :135.0  
##  Mean   :65.0   Mean   :136.7  
##  3rd Qu.:68.5   3rd Qu.:148.0  
##  Max.   :72.0   Max.   :164.0  
```

even simpler, all of the statistics using *psych* package

```
install.packages("psych")
```

```
library(psych)
describe(women)
```

```
##        vars  n   mean    sd median trimmed   mad min max range skew
## height    1 15  65.00  4.47     65   65.00  5.93  58  72    14 0.00
## weight    2 15 136.73 15.50    135  136.31 17.79 115 164    49 0.23
##        kurtosis   se
## height    -1.44 1.15
## weight    -1.34 4.00
```

2.2 Tables

2.2.1 Count, proportion, percentage and cross-tabulation

Use *birthwt* dataset from MASS package.

```
install.packages("MASS")
```

```
library(MASS)
head(birthwt)  # First six subjects
```

```
##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
## 91   0  21 124    3     0   0  0  0   0 2622
```

Count and proportion,

```
table(birthwt$smoke)
```

```
## 
##   0   1 
## 115  74
```

```
prop.table(table(birthwt$smoke))
```

```
## 
##         0         1 
## 0.6084656 0.3915344
```

Cross-tabulation of smoking vs low birth weight baby,

```
table(birthwt$smoke, birthwt$low)  # without row/column labels
```

```
##    
##      0  1
##   0 86 29
##   1 44 30
```

```
table("Smoking status" = birthwt$smoke, "Low birth weight" = birthwt$low)  # with row/column labels
```

```
##               Low birth weight
## Smoking status  0  1
##              0 86 29
##               1 44 30
```

To add value labels to the data for a nicer table, we use *factor*

```
birthwt$smoking = factor(birthwt$smoke, levels = 0:1, labels = c("Non-smoker", "Smoker"))
birthwt$low.weight = factor(birthwt$low, levels = 0:1, labels = c("Low <2.5kg", "Normal >2.5kg"))
head(birthwt)  # we added two new variables with factors
```

```
##    low age lwt race smoke ptl ht ui ftv  bwt    smoking low.weight
## 85   0  19 182    2     0   0  0  1   0 2523 Non-smoker Low <2.5kg
## 86   0  33 155    3     0   0  0  0   3 2551 Non-smoker Low <2.5kg
## 87   0  20 105    1     1   0  0  0   1 2557     Smoker Low <2.5kg
## 88   0  21 108    1     1   0  0  1   2 2594     Smoker Low <2.5kg
## 89   0  18 107    1     1   0  0  1   0 2600     Smoker Low <2.5kg
## 91   0  21 124    3     0   0  0  0   0 2622 Non-smoker Low <2.5kg
```

```
table(birthwt$smoking)
```

```
## 
## Non-smoker     Smoker 
##        115         74
```

```
prop.table(table(birthwt$smoking))*100  # in percent
```

```
## 
## Non-smoker     Smoker 
##   60.84656   39.15344
```

```
cbind(n = table(birthwt$smoking), "%" = 100*prop.table(table(birthwt$smoking)))  # using cbind
```

```
##              n        %
## Non-smoker 115 60.84656
## Smoker      74 39.15344
```

```
table(birthwt$smoking, birthwt$low.weight)
```

```
## 
##              Low <2.5kg Normal >2.5kg
##   Non-smoker         86            29
##   Smoker             44            30
```

Save table for later view and analysis,

```
smoke.x.weight = table(birthwt$smoking, birthwt$low.weight)
smoke.x.weight
```

```
## 
##              Low <2.5kg Normal >2.5kg
##   Non-smoker         86            29
##   Smoker             44            30
```

2.2.2 Entering table data

```
smoking = as.table(rbind(c(15, 5), c(7, 13)))
smoking
```

```
##    A  B
## A 15  5
## B  7 13
```

```
str(smoking)
```

```
##  table [1:2, 1:2] 15 7 5 13
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:2] "A" "B"
##   ..$ : chr [1:2] "A" "B"
```

```
dimnames(smoking) = list(
  Smoking = c("Yes", "No"),
  Lung.CA = c("Yes", "No")
)
smoking
```

```
##        Lung.CA
## Smoking Yes No
##     Yes  15  5
##     No    7 13
```

Chapter 3

Graphical

Test GIT Test GIT 2 - commit

Chapter 4

Reporting results