Stat 120

Deepak Bastola

2023-01-01

Contents

About

This is a *sample* book written in **Markdown**.

Chapter 1

(PART*) Basics R

Chapter 2

What is R?

R is a free and open source statistical programming language that facilitates statistical computation. There are a myriad of application that can be done in R, thanks to a huge online support community and dedicated packages. However, R has no graphical user interface and it has to be run by typing commands into a text interface.

2.1 What is RStudio?

RStudio provides graphical interface to R! You can think of RStudio as a graphical front-end to R that that provides extra functionality. The use of the R programming language with the RStudio interface is an essential component of this course.

2.2 R Studio Server

The quickest way to get started is to go to https://maize.mathcs.carleton.edu, which opens an R Studio window in your web browser. Once logged in, I recommend that you do the following:

- Step 1: Create a folder for this course where you can save all of your work. In the Files window, click on New Folder.
- Step 2: Click on Tools -> Global Options -> R Markdown. Then uncheck the box that says "Show output inline…"

(It is also possible to download RStudio on your own laptop. Instructions may be found at the end of this document.)

2.3 R Markdown Basics

An R Markdown file (.Rmd file) combines R commands and written analyses, which are ‘knit’ together into an HTML, PDF, or Microsoft Word document.

An R Markdown file contains three essential elements:

- Header: The header (top) of the file contains information like the document title, author, date and your preferred output format (pdf_document, word_document, or html_document).
- Written analysis: You write up your analysis after the header and embed R code where needed. The online help below shows ways to add formatting details like bold words, lists, section labels, etc to your final pdf/word/html document. For example, adding ** before and after a word will bold that word in your compiled document.
- R chunks: R chunks contain the R commands that you want evaluated. You embed these chunks within your written analysis and they are evaluated when you compile the document.

2.3.1 R Markdown example:

- Simple R Markdown example
  - compiled pdf

The following handouts, written by Prof Katie St Clair, contain useful information for making the figured and tables in your compiled documents look nice:

- Graph Formatting: Markdown .Rmd file and pdf
- Table Formatting: Markdown .Rmd file and pdf

2.4 Installing R/RStudio (not needed if you are using the maize server)

- Download the latest version of R:
  - Windows: http://cran.r-project.org/bin/windows/base/
  - Mac: http://cran.r-project.org/bin/macosx/
- Download the free Rstudio desktop version (Windows or Mac): https://www.rstudio.com/products/rstudio/download/

Use the default download and install options for each.

2.5 Install LaTeX (for knitting R Markdown documents to PDF):

If you want to compile R Markdown to .pdf files, you also need a LaTeX distribution (Note: this is not necessary if you choose to compile as a Word document.) Click instructions for Windows or instructions for Mac, depending on your operating system to complete the installation.

2.6 Updating R/RStudio (not needed if you are using the maize server)

If you have used a local version of R/RStudio before and it is still installed on your machine, then you should make sure that you have the most recent versions of each program.

- To check your version of R, run the command `getRversion()` and compare your version to the newest version posted on https://cran.r-project.org/. If you need an update, then install the newer version using the installation directions above.
- In RStudio, check for updates with the menu option `Help > Check for updates`. Follow directions if an update is needed.

Chapter 3

R Markdown

This is a R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

You can use asterisk mark to provide emphasis, such as `*italics* or **bold**`.

You can create lists with a dash:

```
- Item 1
- Item 2
- Item 3
  + Subitem 1
* Item 4
```

- Item 1
- Item 2
- Item 3
    - Subitem 1
- Item 4

You can embed Latex equations in-line, $\frac{1}{n}\sum_{i=1}^{n} x_i$ or in a new line as

$$\text{Var}(X) = \frac{1}{n-1}\sum_{i-1}^{n}(x_i - \bar{x})^2$$

## Embed an R code chunk:

Use

~~~
```r
Use back ticks to
create a block of code
```
~~~

to produce:

```
Use back ticks to
create a block of code
```

You can also evaluate and display the results of R code. Each tasks can be accomplished in a suitably labeled chunk like the following:

```r
summary(cars)
```

```
     speed           dist       
 Min.   : 4.0   Min.   :  2.00  
 1st Qu.:12.0   1st Qu.: 26.00  
 Median :15.0   Median : 36.00  
 Mean   :15.4   Mean   : 42.98  
 3rd Qu.:19.0   3rd Qu.: 56.00  
 Max.   :25.0   Max.   :120.00  
```

```r
fit <- lm(dist ~ speed, data = cars)
fit
```

```

Call:
lm(formula = dist ~ speed, data = cars)

Coefficients:
(Intercept)        speed  
    -17.579        3.932  
```

3.1 Including Plots

You can also embed plots. See Figure 3.1 for example:

```r
par(mar = c(0, 1, 0, 1))
pie(
  c(280, 60, 20),
  c('Sky', 'Sunny side of pyramid', 'Shady side of pyramid'),
```

```
  col = c('#0292D8', '#F7EA39', '#C4B632'),
  init.angle = -50, border = NA
)
```

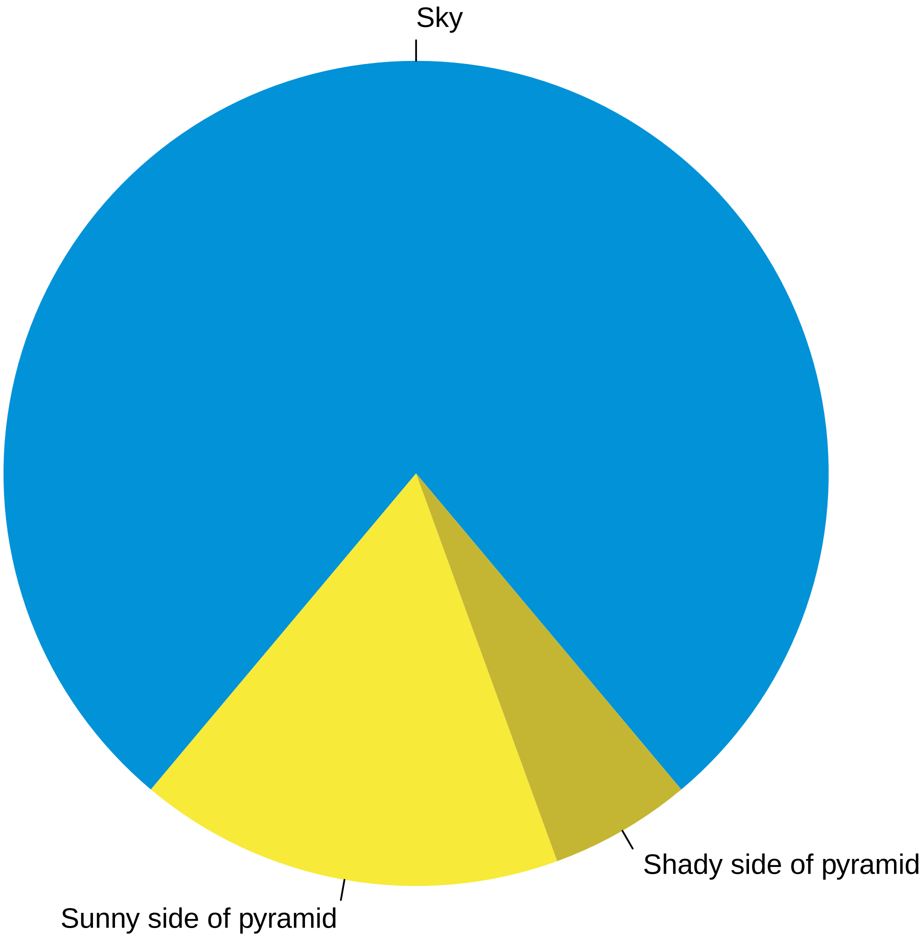

Figure 3.1: A fancy pie chart.

(Credit: Yihui Xie)

3.2 Read in data files

```
simple_data <- read.csv("https://deepbas.io/data/simple-1.dat", )
summary(simple_data)
```

```
   initials            state                age       
 Length:3           Length:3           Min.   :45.0  
 Class :character   Class :character   1st Qu.:47.5  
 Mode  :character   Mode  :character   Median :50.0  
                                       Mean   :52.0  
                                       3rd Qu.:55.5  
                                       Max.   :61.0  
     time          
 Length:3          
 Class :character  
 Mode  :character  
```

```r
knitr::kable(simple_data, format = "html")
```

initials

state

age

time

vib

MA

61

6:01

adc

TX

45

5:45

kme

CT

50

4:19

3.3 Hide the code

If we enter the `echo = FALSE` option in the R chunk (see the .Rmd file). This prevents the R code from being printed to your document; you just see the results.

initials

state

age

time

vib

MA

61

6:01

adc

TX

45

5:45

kme

CT

50

4:19

Chapter 4

(PART*) Class Activity

Chapter 5

Conclusion

Click for answer

The correct answer is a. If there is a difference, we expect the between group variability to be higher than within group variability. RIGHT TAIL test!

```
Temperature  =  37.7 + 0.231 Chirps
Predictor        Coef    SE Coef      T    Pr(>|t|)
Constant    37.67858    1.97817   19.05  7.35e-06 ***
Chirps       0.23067    0.01423   16.21  1.63e-05 ***
```

```
survey <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/StudentSurvey.csv
mean(survey$Pulse) # the command `mean` computes an average
```

```
[1] 69.57459
```

ROCK	PAPER	SCISSORS	TOTAL
36	12	37	85

First year at Carleton

- Originally from Nepal
- PhD in Applied Statistics from

UC-Riverside

- Diverse education background
- Avid learner and traveler

Chapter 6

Class Activity 1

- Try to knit the file at the present stage and see if it compiles.
- You can add `\vspace*{1in}` in the body of this file to produce a vertical space of 1 inches.

6.1 Your Turn 1

---

a. Run the following chunk. Comment on the output.

```
example_data = data.frame(ID = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10),
                          Greeting = c(rep("Hello", 5), rep("Goodbye",5)),
                          Male = rep(c(TRUE, FALSE), 5),
                          age = runif(n=10, 20,60))
```

Click for answer

```
example_data
```

```
   ID Greeting  Male      age
1   1    Hello  TRUE 40.24020
2   2    Hello FALSE 27.66788
3   3    Hello  TRUE 26.70954
4   4    Hello FALSE 24.22673
5   5    Hello  TRUE 26.88372
6   6  Goodbye FALSE 40.21593
```

```
7   7  Goodbye  TRUE 25.05656
8   8  Goodbye FALSE 36.43088
9   9  Goodbye  TRUE 44.38057
10 10  Goodbye FALSE 59.29558
```

Answer: We see a data frame with four columns, where the first column is an `identifier` for the cases. We have information on the greeting types, gender, and age on these cases in the remaining columns.

b. What is the dimension of the dataset called 'example_data'?

Click for answer

```
dim(example_data)
```

```
[1] 10  4
```

```
nrow(example_data)
```

```
[1] 10
```

```
ncol(example_data)
```

```
[1] 4
```

Answer: There are 10 rows and 4 columns.

6.2 Your Turn 2

a. Read the dataset `EducationLiteracy` from the Lock5 second edition book.

Click for answer

```
# read in the data
education_lock5 <- read.csv("https://www.lock5stat.com/datasets2e/EducationLiteracy.csv")
```

b. Print the header (i.e. first 6 cases by default) of the dataset in part a.

Click for answer

```
head(education_lock5)
```

```
              Country EducationExpenditure Literacy
1         Afghanistan                  3.1     31.7
2             Albania                  3.2     96.8
3             Algeria                  4.3       NA
4             Andorra                  3.2       NA
5              Angola                  3.5     70.6
6 Antigua and Barbuda                  2.6     99.0
```

c. What is the dimension of the dataset in a?

Click for answer

```
dim(education_lock5)
```

```
[1] 188   3
```

Answer: There are 188 rows and 3 columns.

d. What type of variables are `Country`, `EducationExpenditure`, and `Literacy`?

Click for answer

Answer: `Country` is a categorical variable. `EducationExpenditure` and `Literacy` are both quantitative variables.

e. If we would like to use education expenditure to predict the literacy rate of each countries, which variable is the explanatory variable and which one is the response?

Click for answer

Answer: The education expenditure is the explanatory variable, and the literacy rate is the response.

Chapter 7

Class Activity 1

7.1 Your Turn 2

7.2 Summary of article on It depends on how you ask!

Answer:

This study aimed to measure the effects of psychological biases on estimates of compliance with public health guidance regarding COVID-19. Results showed that compliance estimates were reduced when questions were framed negatively and anonymity was increased. Effect sizes were large, with compliance estimates diminishing by up to 17% points and 10% points, respectively. These findings suggest that standard tracking surveys pose questions in ways that lead to higher compliance estimates than alternative approaches.

7.3 Your Turn 3

7.4 Gettysburg random sample

Let's take a simple random sample (SRS) of Gettysburg words. The "population" is contained in the spreadsheet `GettysburgPopulationCounts.csv`. Carefully load this data into R:

```
pop <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/Gettysberg
head(pop)
```

```
  position size  word
1        1    4  Four
2        2    5 score
3        3    3   and
4        4    5 seven
5        5    5 years
6        6    3  ago,
```

The `position` variable enumerates the list of words in the population (address).

a. Sample

Run the following command to obtain a SRS of 10 words from the 268 that are in the population:

```
samp <- sample(1:268, size=10)
samp
```

```
 [1] 174 237 125 199 222 213 230  14 172 263
```

This tells you the position (row number) of your sampled words. What are your sampled positions? Why are your sampled positions different from other folks in class?

b. Get words and lengths

We will *subset* the data set `pop` to obtain only the sampled rows listed in `samp`. We do this using **square bracket notation** ‘dataset[row number, column number/name]. Run the following command to find your sampled words and sizes:

```
pop[samp,]
```

```
    position size       word
174      174   10 unfinished
237      237    2         in
125      125    3        who
199      199    4       task
222      222    7    measure
213      213    4       that
230      230    4       that
14        14    1          a
172      172    2         to
263      263    5      shall
```

What are your sampled words?

c. Compute your sample mean

The word lengths in part (b) are the data for your sample. You can compute your sample mean using a calculator, or using R. Let's try R (you will find it faster!). First save the quantitative variable `size` in a new variable called `mysize`:

```
mysize <- pop[samp, "size"]
mysize
```

```
 [1] 10  2  3  4  7  4  4  1  2  5
```

Then find the mean of these values:

```
mean(mysize)
```

```
[1] 4.2
```

How does this sample mean (from a truly random sample) compare to your sample mean from the non-random sample?

Answer: The true mean is 4.29. Your two means will likely vary. Since the many non-random samples generally overestimated the population mean length, it is possible (but not guaranteed) that *your* one non-random sample gave a mean length that is greater than the random sample's mean length.

7.5 Driving with a Pet on your Lap

Over 30,000 people participated in an online poll on `cnn.com` conducted in April 2012 asking: "Have you ever driven with a pet on your lap"? We see that 34% of the participants answered yes and 66% answered no.

a. Can you conclude that a random sample was used from the description given? Explain.

Answer: No you can't make this conclusion from the info given. In fact, an online poll at a website like `cnn.com` is almost always reporting results from a non-random sample. The people who respond are individuals who visit `cnn.com`, then see the online poll and decide to respond.

b. Explain why it is not appropriate to generalize these results to all drivers, or even to all drivers who visit cnn.com.

Answer: This is a volunteer sample, and volunteer samples are often biased and can't be generalized to *all drivers* (the population). It is likely that people who have driven with a pet on their lap are more likely to respond to the poll.

c. How might we select a sample of people that would give us results that we can generalize to a broader population?

Answer: A random sample of individuals from all U.S. drivers would need to be selected and given the poll question. (There are many ways to do this, the most common being a variation of random digit dialing where phone numbers are randomly selected from known area codes.)

d. Is the variable measured in this study quantitative or categorical?

Answer: Categorical (yes or no answer to the question).