

Stat 120

Deepak Bastola

2023-01-08

Contents

About	5
1 Class Activity 1	7
1.1 Your Turn 1	7
1.2 Your Turn 2	8
1.3 Quiz	9
2 Class Activity 2	11
2.1 Your Turn 1	11
2.2 Your Turn 2	11
2.3 Your Turn 3	12
2.4 Quiz	14
3 (PART*) Basics R	17
4 What is R?	19
4.1 What is RStudio?	19
4.2 R Studio Server	19
4.3 R Markdown Basics	20
4.4 Installing R/RStudio (not needed if you are using the maize server)	20
4.5 Install LaTeX (for knitting R Markdown documents to PDF): . .	20
4.6 Updating R/RStudio (not needed if you are using the maize server)	21
4.7 Instructions	21
4.8 Few More Instructions	22

5	R Markdown	23
5.1	Including Plots	24
5.2	Read in data files	25
5.3	Hide the code	26

About

This is a *sample* book written in **Markdown** to guide STAT 120 students interactively explore various class activities and projects in R.

Chapter 1

Class Activity 1

1.1 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	21.39270
2	2	Hello	FALSE	55.18330
3	3	Hello	TRUE	54.35298
4	4	Hello	FALSE	38.84701
5	5	Hello	TRUE	46.65463
6	6	Goodbye	FALSE	57.27510
7	7	Goodbye	TRUE	59.32366
8	8	Goodbye	FALSE	56.82235
9	9	Goodbye	TRUE	29.72396
10	10	Goodbye	FALSE	39.89474

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, whether male or not, 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.

1.2 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.

1.3 Quiz

1. Cases are a set of individual units where the measurements are taken.

- A. TRUE
- B. FALSE

Click for answer

TRUE

2. The characteristic that is recorded for each case is called a

- A. ledger

- B. caseholder
- C. placeholder
- D. variable

Click for answer

variable

3. Variables can be either categorical or quantitative.

- A. TRUE
- B. FALSE

Click for answer

TRUE

Chapter 2

Class Activity 2

2.1 Your Turn 1

This exercise is about finding the average word length in Lincoln's Gettysburg's address.

2.2 Your Turn 2

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

Click for answer

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.

2.3 Your Turn 3

2.3.1 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/Gettysburg")
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] 22 73 40 245 62 21 67 107 135 121
```

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
22	22	2	to

73	73	4	that
40	40	7	testing
245	245	4	have
62	62	2	of
21	21	9	dedicated
67	67	4	come
107	107	5	sense,
135	135	5	power
121	121	3	men,

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] 2 4 7 4 2 9 4 5 5 3
```

Then find the mean of these values:

```
mean(mysize)
```

```
[1] 4.5
```

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

Click for answer

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.

2.3.2 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.

Click for answer

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`.

Click for answer

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?

Click for answer

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?

Click for answer

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

2.4 Quiz

1. A group of researchers investigated the effect of media usage (whether or not subjects watch television or use the Internet) in the bedroom on "Tiredness" during the day (measured on a 50 point scale). The explanatory and response variables are

- A. Explanatory is media usage in the bedroom and response is "tiredness"

B. Explanatory is “tiredness” and response is media usage in the bedroom

Click for answer

The correct answer is A.

2. An October 2016 Gallup poll estimates that 60% of US adults support legalizing the use of marijuana. Their results were based on a “random sample of 1,017 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia”. The population for this study is

A. all adults (18 and older) living in the U.S. (including D.C)

B. the 1,017 adults (18 and older) living in the U.S. (including D.C) who were sampled

C. the 1,017 adults (18 and older) living in the U.S. (including D.C) who were sampled and support legalizing marijuana

D. all adults (18 and older) living in the U.S. (including D.C) who support legalizing marijuana

Click for answer

The correct answer is A.

3. An October 2016 Gallup poll estimates that 60% of US adults support legalizing the use of marijuana. Their results were based on a “random sample of 1,017 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia”. Which statement below regarding bias is true?

A. The results are biased because Gallup only contacted a small fraction of people in the population.

B. The results may be biased because people may not have answered a survey question about marijuana truthfully

Click for answer

The correct answer is B.

Chapter 3

(PART*) Basics R

Chapter 4

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.

4.1 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.

4.2 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.)

4.3 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.

4.4 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.

4.5 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.

4.6 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.

4.7 Instructions

If using Rstudio on your computer, using the **File>Open File** menu to find and open this .Rmd file.

If using Maize Rstudio from your browser:

- In the Files tab, select **Upload** and **Choose File** to find the .Rmd that you downloaded. Click *OK* to upload to your course folder/location in the maize server account.
- Click on the .Rmd file in the appropriate folder to open the file.

Extra notes:

- You can run a line of code by placing your cursor in the line of code and clicking **Run Selected Line(s)**
- You can run an entire chunk by clicking the green triangle on the right side of the code chunk.
- After each small edit or code addition, **Knit** your Markdown. If you wait until the end to Knit, it will be harder to find errors in your work.
- Format output type: You can use any of pdf_document, html_document type, or word_document type.
- **Maize users:** You may also need to allow for “pop-up” in your web browser when knitting documents.

4.8 Few More Instructions

The default setting in Rstudio when you are running chunks is that the “output” (numbers, graphs) are shown **inline** within the Markdown Rmd. If you prefer to have your plots appear on the right of the console and not below the chunk, then change the settings as follows:

1. Select Tools > Global Options.
2. Click the R Markdown section and uncheck (if needed) the option Show output inline for all R Markdown documents.
3. Click OK.

Now try running R chunks in the .Rmd file to see the difference. You can recheck this box if you prefer the default setting.

Chapter 5

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 task can be accomplished in a suitably labeled chunk like the following:

```
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 |

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

Call:

```
lm(formula = dist ~ speed, data = cars)
```

Coefficients:

| | |
|-------------|-------|
| (Intercept) | speed |
| -17.579 | 3.932 |

5.1 Including Plots

You can also embed plots. See Figure 5.1 for example:

```
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  
)
```

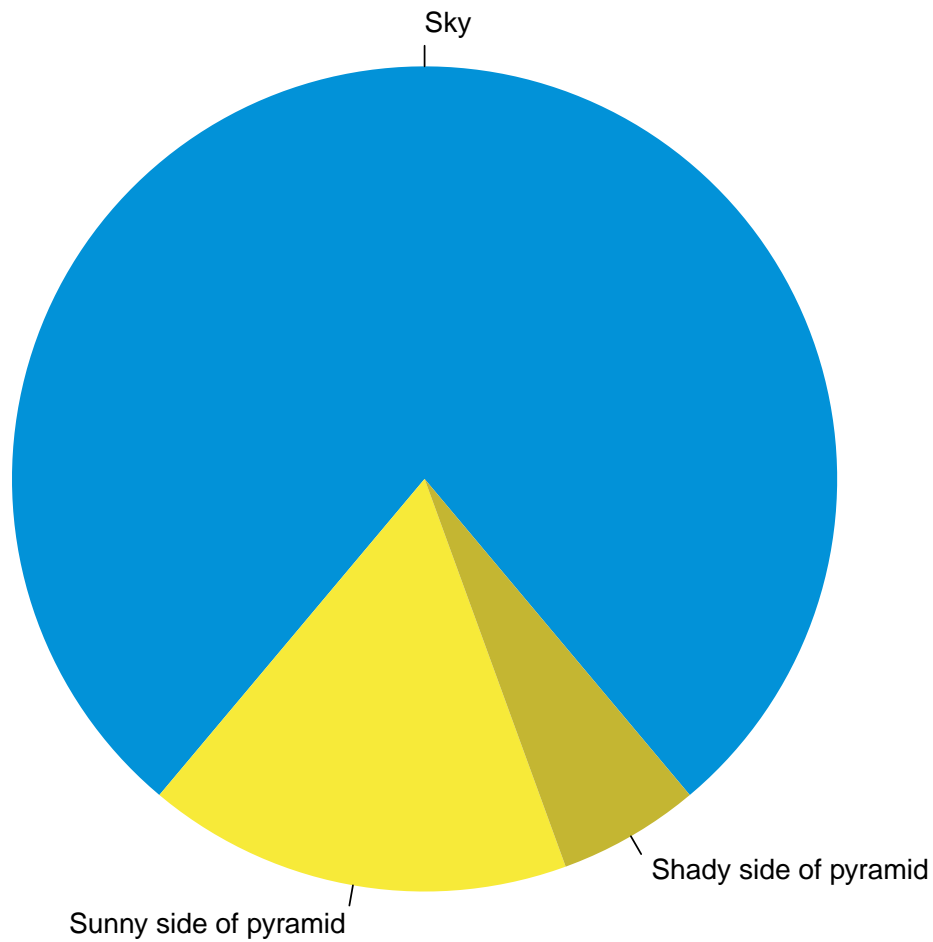


Figure 5.1: A fancy pie chart.

(Credit: Yihui Xie)

5.2 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

```

```
knitr::kable(simple_data)
```

| initials | state | age | time |
|----------|-------|-----|------|
| vib | MA | 61 | 6:01 |
| adc | TX | 45 | 5:45 |
| kme | CT | 50 | 4:19 |

5.3 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 |