

Stat 120

Deepak Bastola

2023-01-11



# Contents

<b>About</b>	<b>5</b>
<b>1 Class Activity 1</b>	<b>7</b>
1.1 Your Turn 1 . . . . .	7
1.2 Your Turn 2 . . . . .	8
1.3 Quiz . . . . .	9
<b>2 Class Activity 2</b>	<b>11</b>
2.1 Your Turn 1 . . . . .	11
2.2 Your Turn 2 . . . . .	11
2.3 Your Turn 3 . . . . .	12
2.4 Quiz . . . . .	14
<b>3 Class Activity 3</b>	<b>17</b>
3.1 Case Study 1 . . . . .	17
3.2 Case Study 2 . . . . .	18
3.3 Quiz . . . . .	20
<b>4 Class Activity 4</b>	<b>21</b>
4.1 Your Turn 1 . . . . .	21
4.2 Your Turn 2 . . . . .	27
4.3 Quiz . . . . .	36
<b>5 (PART*) Basics R</b>	<b>39</b>

<b>6</b>	<b>What is R?</b>	<b>41</b>
6.1	What is RStudio? . . . . .	41
6.2	R Studio Server . . . . .	41
6.3	R Markdown Basics . . . . .	42
6.4	Installing R/RStudio (not needed if you are using the maize server)	42
6.5	Install LaTeX (for knitting R Markdown documents to PDF): . .	42
6.6	Updating R/RStudio (not needed if you are using the maize server)	43
6.7	Instructions . . . . .	43
6.8	Few More Instructions . . . . .	44
<b>7</b>	<b>R Markdown</b>	<b>45</b>
7.1	Including Plots . . . . .	46
7.2	Read in data files . . . . .	47
7.3	Hide the code . . . . .	48

# 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	56.41977
2	2	Hello	FALSE	48.79070
3	3	Hello	TRUE	55.55897
4	4	Hello	FALSE	25.95434
5	5	Hello	TRUE	26.33548
6	6	Goodbye	FALSE	40.31555
7	7	Goodbye	TRUE	23.83351
8	8	Goodbye	FALSE	43.08431
9	9	Goodbye	TRUE	50.55220
10	10	Goodbye	FALSE	29.80661

*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



### 1.3. QUIZ

9

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] 186 217 162 190 256 58 171 264 116 14
```

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
186	186	8	advanced.

217	217	4	they
162	162	2	is
190	190	3	for
256	256	6	people,
58	58	2	on
171	171	4	here
264	264	3	not
116	116	6	hallow
14	14	1	a

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] 8 4 2 3 6 2 4 3 6 1
```

Then find the mean of these values:

```
mean(mysize)
```

```
[1] 3.9
```

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

# Class Activity 3

### 3.1 Case Study 1

Consider the following case study:

“Swimming with dolphins can certainly be fun, but is it also therapeutic for patients suffering from clinical depression? To investigate this possibility, researchers recruited 30 subjects aged 18-65 with a clinical diagnosis of mild to moderate depression. Subjects were required to discontinue use of any antidepressant drugs or psychotherapy four weeks prior to the experiment, and throughout the experiment. These 30 subjects went to an island off the coast of Honduras, where they were randomly assigned to one of two treatment groups. Both groups engaged in the same amount of swimming and snorkeling each day, but one group (the animal care program) did so in the presence of bottlenose dolphins and the other group (outdoor nature program) did not. At the end of two weeks, each subjects’ level of depression was evaluated, as it had been at the beginning of the study, and it was determined whether they showed substantial improvement (reducing their level of depression) by the end of the study (Antonioli and Reveley, 2005).”

Observed data:

The researchers found that 10 of 15 subjects in the dolphin therapy group showed substantial improvement, compared to 3 of 15 subjects in the control group.

(a). Identify the observational units in this study.

Click for answer

*Answer:* The observational units in this study are the 30 subjects.

(b). Classify each variable as categorical or quantitative.

Click for answer

*Answer:* The variables in this study can be classified as follows: Categorical: Treatment Group (Dolphin and Control) Quantitative: Age, Level of Depression (Beginning and End of Study)

(c). Which variable would you regard as explanatory and which as response?

Click for answer

*Answer:* The explanatory variable would be the Treatment Group and the response variable would be the Level of Depression.

(d). Is this an observational study or an experiment? Justify your answer.

Click for answer

*Answer:* This is an experiment because the researchers randomly assigned the subjects to the two treatment groups, and then observed the effect of the treatment (presence of dolphins) on the response variable (level of depression).

(e). Construct a two-way table based on the results of the experiment.

Click for answer

Two-way table:

Treatment	Improved	Not Improved	Total
Dolphin Group	10	5	15
Control Group	3	12	15
Total	13	17	30

## 3.2 Case Study 2

Consider the following case study:

“Researchers want to find out how a new diet affects weight gain among underweight subjects. This experiment only has two treatment conditions, the new diet and the standard diet. For this study, the researchers recruited 200 subjects which will be grouped into 100 pairs based on shared characteristics such as age, gender, weight, height, lifestyle, and so on. A 20-year-old female within the weight range of 90-110 pounds and the height range of 60-63 inches will be paired with another 20-year-old female that falls into the same weight and height categories. Once all 100 pairs are made, a subject from each pair will be randomly assigned into the treatment group (will be administered the new diet for 2 months) while the other subject from the pair will be assigned to the control group (will be assigned to follow the standard diet for two months).

At the end of the time period of 2 months, researchers will measure the total weight gain for each subject.”

Observed data:

The researchers found that 60 of 100 subjects in the new diet group showed substantial improvement, compared to 43 of 100 subjects in the standard diet group.

(a). Identify the observational units in this study.

Click for answer

*Answer:* The observational units in this study are the 200 subjects.

(b). Classify each variable as categorical or quantitative.

Click for answer

*Answer:* The variables are: age (quantitative), gender (categorical), weight (quantitative), height (quantitative), lifestyle (categorical), and total weight gain (quantitative).

(c). Which variable would you regard as explanatory and which as response?

Click for answer

*Answer:* The explanatory variable is the type of diet (new or standard) and the response variable is the total weight gain.

(d). Is this an observational study or an experiment? Justify your answer.

Click for answer

*Answer:* This is an experiment because the researchers are manipulating the explanatory variables (type of diet) to observe the effects on the response variables (total weight gain).

(e). If it is an experiment, is it randomized comparative experiment or a matched pairs experiment?

Click for answer

*Answer:* This is a matched pairs experiment because each subject is paired with another subject who has similar characteristics and one subject from each pair is randomly assigned to the treatment group and the other to the control group.

(f). Construct a two-way table based on the results of the experiment.

Click for answer

Two-way table:

Outcome	New Diet	Standard Diet	Total
Improvement	60	43	103

Outcome	New Diet	Standard Diet	Total
No Improvement	40	57	97
Total	100	100	200

---

### 3.3 Quiz

**1. A third variable that is associated with both the explanatory variable and the response variable is called a confounding variable.**

A. TRUE

B. FALSE

Click for answer

TRUE

**2. The different levels of an explanatory variable are known as**

A. treatments

B. local groups

C. response

D. cases

Click for answer

treatments

**3. Causality can always be inferred from observational studies.**

A. TRUE

B. FALSE

Click for answer

FALSE

## Chapter 4

# Class Activity 4

### 4.1 Your Turn 1

#### 4.1.1 Flowers v. Mississippi

The data set `APM_DougEvansCases.csv` contains data from 1517 potential black and white jurors for 66 cases that Doug Evans was primary prosecutor for between 1992 and 2017. These jurors were available for Doug Evans to strike using his “peremptory strikes” during the jury selection phase.

(a). Inspect data

Read in the data

```
jurors <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/APM_DougEvansCases.csv")
```

```
# dimension of dataset
dim(jurors)
```

```
[1] 1517    6
```

Look at the first **three rows** of the data set

```
jurors[c(1,2,3), ]
```

	trial__id	race	struck_state	defendant_race
1	4	Black	Not struck by State	White
2	4	Black	Struck by State	White
3	4	White	Not struck by State	White

```

      same_race      struck_by
1 different race Juror chosen to serve on jury
2 different race      Struck by the state
3      same race Juror chosen to serve on jury

```

To get the data from one variable, we use the command `dataset$variable`. For example, `jurors$struck_state` gives us the data values from the `struck_state` variable, which tells us if a juror was struck by the state from the jury pool. Here we can see the first 10 entries in this variable:

```
jurors$struck_state[1:10]
```

```

[1] "Not struck by State" "Struck by State"
[3] "Not struck by State" "Not struck by State"
[5] "Struck by State"     "Not struck by State"
[7] "Struck by State"     "Not struck by State"
[9] "Not struck by State" "Not struck by State"

```

(b). Table of counts and proportions

The `summary` command used with a data frame gives summaries of each variable

```
summary(jurors)
```

```

      trial__id      race      struck_state
Min.   : 4.0   Length:1517   Length:1517
1st Qu.: 52.0   Class :character Class :character
Median : 82.0   Mode  :character Mode  :character
Mean    :112.6
3rd Qu.:170.0
Max.    :301.0

defendant_race      same_race      struck_by
Length:1517   Length:1517   Length:1517
Class :character Class :character Class :character
Mode  :character Mode  :character Mode  :character

```

The `table` command gives the distribution of counts for a single categorical variable. To obtain the count table for `struck_state` you need to

```
counts <- table(jurors$struck_state)
counts
```

Not struck by State	Struck by State
1084	433

We can add the `prop.table` command to turn these counts into proportions:

```
prop.table(counts)
```

Not struck by State	Struck by State
0.7145682	0.2854318

- What proportion of eligible jurors were struck by the state from the jury pool?

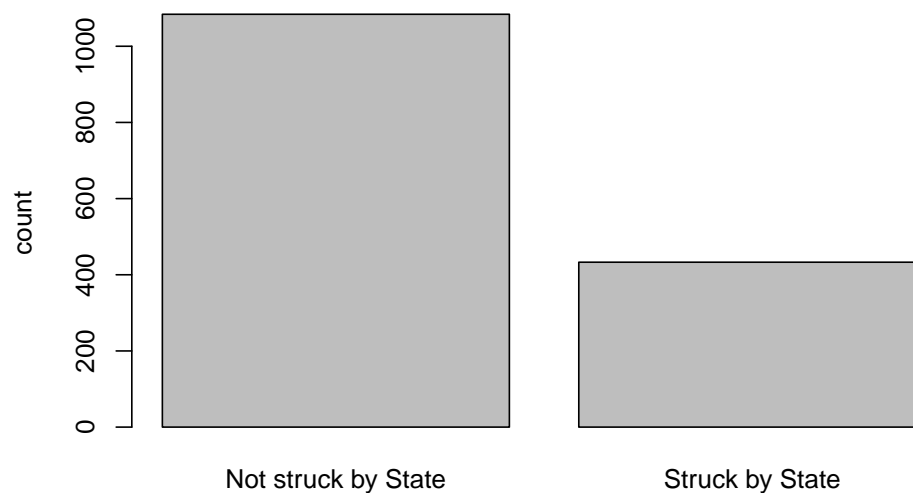
Click for answer

*Answer:* about 28.5% of eligible jurors were struck by the state.

(c). Bar graph for one variable

You can create a simple bar graph for one categorical variable with the `barplot` command. Here we visualize the distribution of struck status for all eligible jurors:

```
barplot(counts, ylab = "count")
```



(d). Two-way tables

First 10 entries of `race` and `struck_state` variable is

```
jurors[(1:10),(2:3)]
```

```

      race      struck_state
1 Black Not struck by State
2 Black      Struck by State
3 White Not struck by State
4 White Not struck by State
5 Black      Struck by State
6 White Not struck by State
7 Black      Struck by State
8 White Not struck by State
9 White Not struck by State
10 White Not struck by State

```

The `table` command also gives two-way tables when two variables are included. Here is the two-way table for juror race and state struck status:

```
mytable <- table(jurors$race, jurors$struck_state)
mytable
```

```

      Not struck by State Struck by State
Black                225                310
White                859                123

```

- How many jurors were white and were not struck by the state?

Click for answer

*answer:* 859

(e). Conditional proportions: state strike status by juror race

The `prop.table` command gives conditional proportions for a two-way table. We plug our two-way table into `prop.table` with a `margin=1` to get proportions grouped by the `row` variable:

```
prop.table(mytable, margin = 1)
```

```

      Not struck by State Struck by State
Black                0.4205607          0.5794393
White                0.8747454          0.1252546

```



Of all eligible black jurors, about 57.9% were struck by the state.

- What proportion of eligible white jurors were struck by the state?  
Click for answer  
*answer:* about 12.5%
- Is there evidence of an association between juror race and state strikes?  
Click for answer  
*answer:* Yes, there is an association because the rate of state strikes varies greatly by juror race with about 60% of black jurors were struck compared to only 13% of white jurors

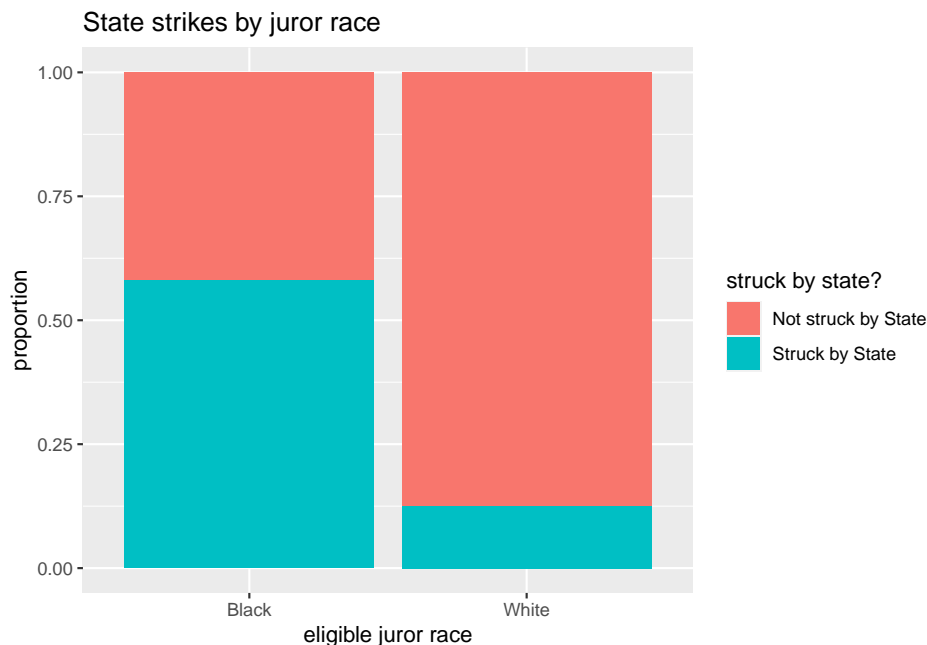
(f). Stacked bar graph for two variables

We can visualize the conditional distribution from part (e) with a stacked bar graph created using the `ggplot2` graphing package. First, load this package's functions with the `library` command:

```
library(ggplot2)
```

Now we can use the `geom_bar` command in this package. Here we get the conditional distribution of `struck_state` given `race`:

```
ggplot(jurors, aes(x = race, fill = struck_state)) +  
  geom_bar(position = "fill") +  
  labs(title = "State strikes by juror race", y = "proportion",  
        x = "eligible juror race", fill = "struck by state?")
```



The basic syntax for this function is to let `ggplot` know your data set name (`jurors`), then specify the grouping or conditional variable on the x-axis (`race`) in the `aes` (aesthetic) argument. The `fill` variable is the response variable (`struck_state`). We add (+) the `geom_bar` geometry to get a bar graph with the `fill` position specified. Adding an informative label and title complete the graph.

(g). Conditional distribution of race grouped by strike status

We can “flip” our response and grouping variables easily (if we think it makes sense to do so). Here we specify the `margin=2` to get proportions grouped by the `column` variable:

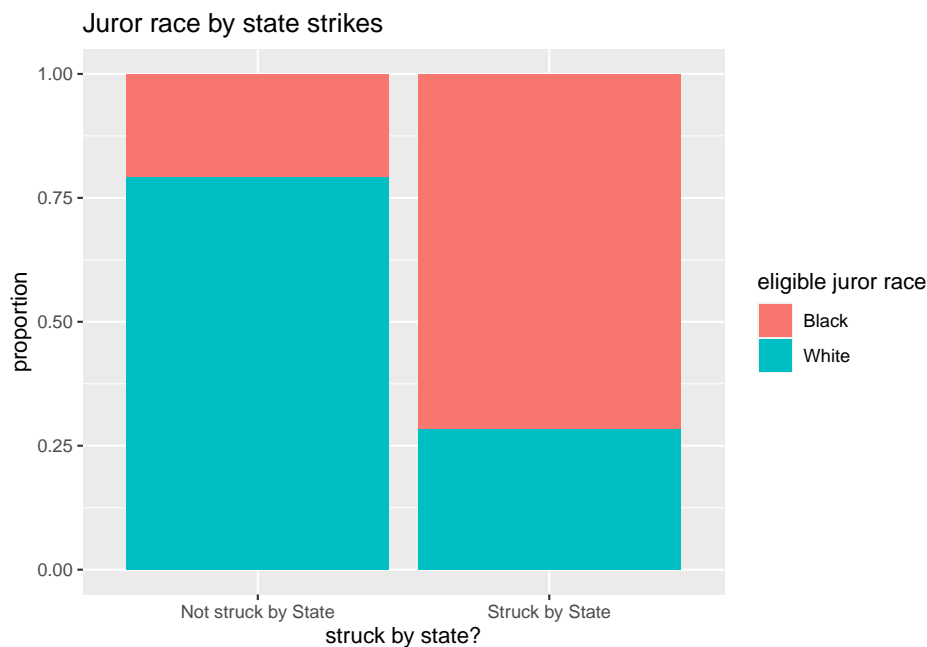
```
prop.table(mytable, margin = 2)
```

	Not struck by State	Struck by State
Black	0.2075646	0.7159353
White	0.7924354	0.2840647

Notice that the proportions add to one **down** each column. Of all eligible jurors struck by the state, about 71.6% were black.

The stacked bar graph for this distribution is

```
ggplot(jurors, aes(x = struck_state, fill = race)) +
  geom_bar(position = "fill") +
  labs(title = "Juror race by state strikes", y = "proportion",
       fill = "eligible juror race", x = "struck by state?")
```



- What proportion of eligible jurors who were not struck by the state were black? were white?

Click for answer

*Answer:* Of all jurors not struck by the state, about 20.8% were black

## 4.2 Your Turn 2

### 4.2.1 Graduate programs acceptance and sex

How are grad school program acceptance rates associated with sex? We will look at a classic data set from Berkeley grad school applications from 1973 (*Science*, 1975). The data cases are applicants to four graduate programs at Berkeley during 1973. The variable **result** tells us if the applicant was accepted to the graduate program, **sex** tells us the sex of the applicant (male or female), and **program** tells us program type (programs 1,2,3 or 4).

```
grad <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/BerkeleyGrad.csv")
```

```
# dimension of the dataset
dim(grad)
```

```
[1] 3014    3
```

```
# first 6 rows
head(grad)
```

```
      program sex result
1 program1 male  accept
2 program1 male  accept
3 program1 male  accept
4 program1 male  accept
5 program1 male  accept
6 program1 male  accept
```

(a). Table of counts and proportions

```
prop.table(table(grad$result))
```

```
      accept    reject
0.4260119 0.5739881
```

- What proportion of applicants were accepted?

Click for answer

*Answer:* About 43% (1284/3014) of applicants were accepted.

(b). Two-way tables

The `table` command also gives two-way tables when two variables are included. Here is the two-way table for result and sex:

```
table(grad$sex, grad$result)
```

```
      accept reject
female    262    587
male      1022   1143
```

- How many applicants involved females who were accepted?

Click for answer

*Answer:* : 262 applicants involved females who were accepted.

(c). Conditional proportions: acceptance given sex

The `prop.table` command gives conditional proportions for a two-way table. First let's save the two-way table in an object named `mytable`:

```
mytable <- table(grad$sex, grad$result)
```

Then use `prop.table` to get the distribution of result conditioned (grouped) on applicant's sex:

```
prop.table(mytable, 1)
```

```
      accept  reject
female 0.3085984 0.6914016
male   0.4720554 0.5279446
```

The value of 1 in this command tell's R that you want *row* proportions (the denominator of the proportion is each row total).

- What proportion of female were accepted?

Click for answer

*Answer:* about 31% ( $262/(262+587)$ )

- What proportion of males were accepted?

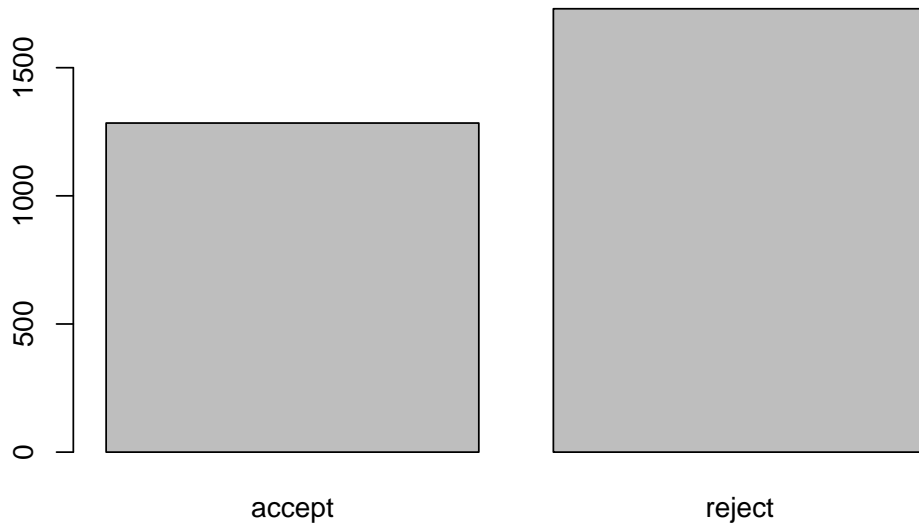
Click for answer

*Answer:* about 47% ( $1022/(1022+1143)$ )

(d). Bar graph for one variable

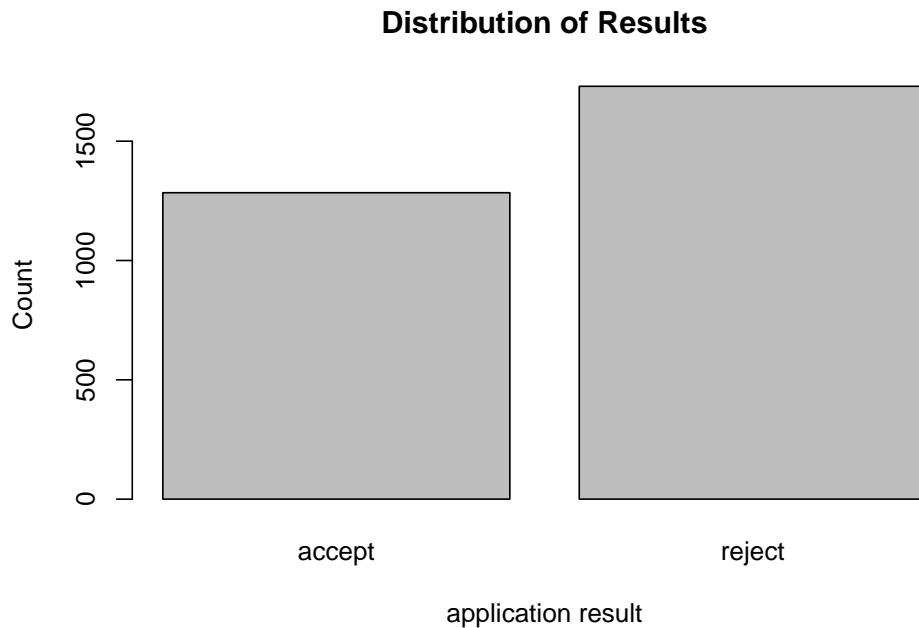
You can create a simple bar graph for one categorical variable with the `barplot` command. Here we visualize the distribution of result:

```
barplot(table(grad$result))
```



We can add in a title and x and y axis labels too:

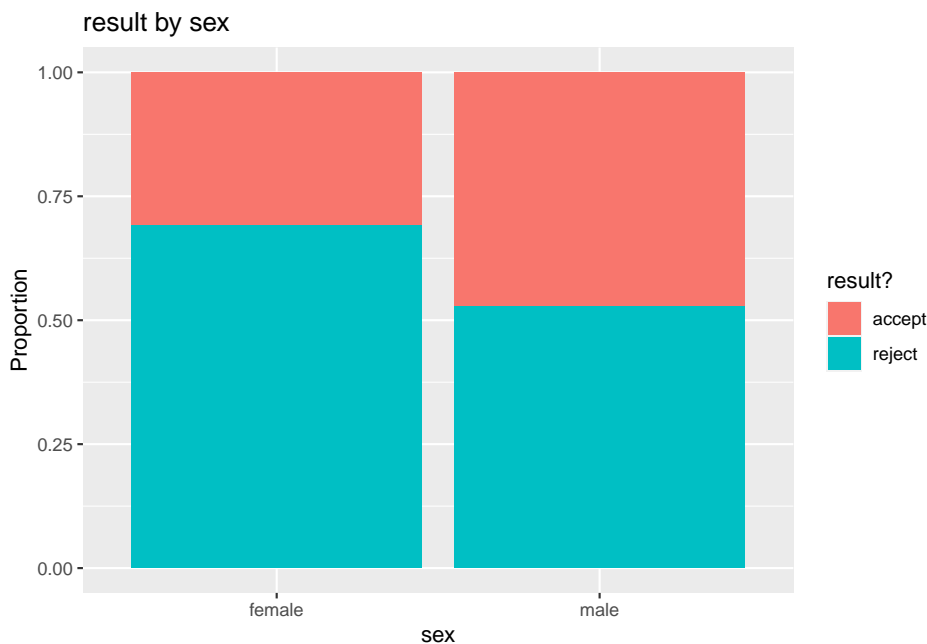
```
barplot(table(grad$result), xlab="application result",
        ylab="Count", main = "Distribution of Results")
```



(e). Stacked bar graph for two variables

Now we can use the `geom_bar` command in this package. Here we get the conditional distribution of `result` given `sex`:

```
library(ggplot2) # don't need if you already entered it for example 1
ggplot(grad, aes(x = sex, fill = result)) +
  geom_bar(position = "fill") +
  labs(y="Proportion", title = "result by sex", fill = "result?", x = "sex")
```



The basic syntax for this function is to let `ggplot` know your data set name (`grad`), then specify the grouping or conditional variable on the x-axis (`sex`) in the `aes` (aesthetic) argument. The `fill` variable is the response variable (`result`). We add (+) the `geom_bar` geometry to get a bar graph with the `fill` position specified. Adding an informative label and title complete the graph.

- Verify that this graph is plotting the conditional proportions from part (c)

(f). Subsetting by program type

Finally, we will repeat the previous analysis of result and sex, but this time we will divide (or subset) the data set by program type. To do this we need to know how the values of `program` are coded:

```
table(grad$program)
```

```
program1 program2 program3 program4
    933      585      782      714
```

Here we use the `filter` command available from the `dplyr` package to get only the applicants to program 1:

```
library(dplyr)
grad.p1 <- filter(grad, program == "program1") # gets rows where program equal program1
head(grad.p1)
```

```
  program sex result
1 program1 male accept
2 program1 male accept
3 program1 male accept
4 program1 male accept
5 program1 male accept
6 program1 male accept
```

```
dim(grad.p1)
```

```
[1] 933  3
```

Verify that the number of rows in the subsetting program 1 data set matches the number of program 1 applicants shown in the `table` of counts above.

- Repeat the `filter` command to get a data set for program 2 and call the new data set `grad.p2`. Verify that the number of rows in this dataset matches the number of program 2 applicants in the original data set.

```
# enter R code for (f) here
grad.p2 <- filter(grad, program == "program2") # gets rows where program equal program2
head(grad.p2)
```

```
  program sex result
1 program2 male accept
2 program2 male accept
3 program2 male accept
4 program2 male accept
5 program2 male accept
6 program2 male accept
```

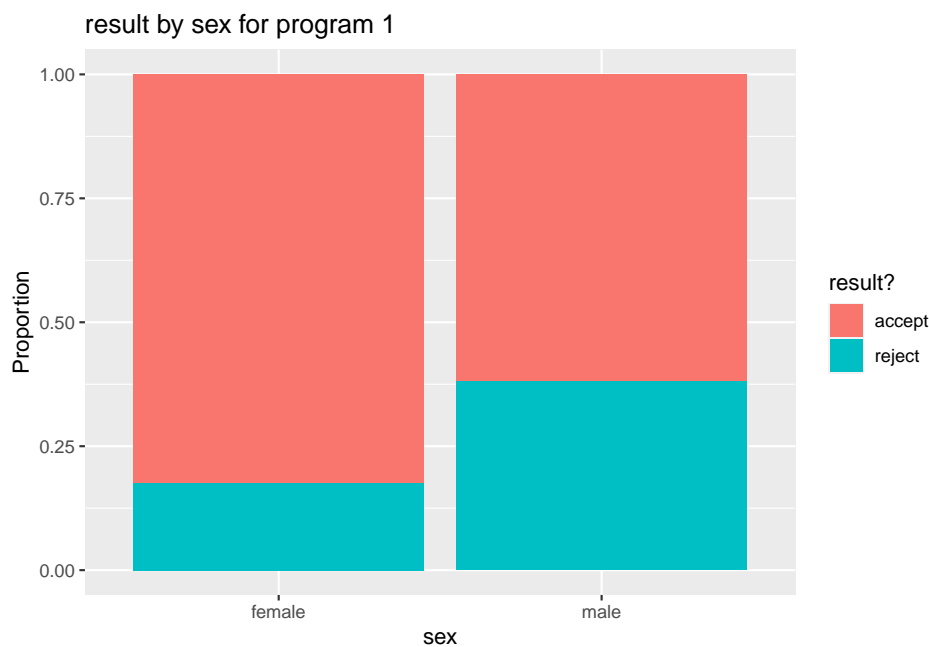
(g). Result by sex for program 1.

- Show the distribution of result conditioned on applicant's sex for the program 1 data set. Get both a table of conditional proportions (or percentages) and a stacked bar graph.



Click for answer

```
# enter R code for (g) here
ggplot(grad.p1, aes(x = sex, fill = result)) +
  geom_bar(position = "fill") +
  labs(y="Proportion", title = "result by sex for program 1",
       fill = "result?", x = "sex")
```



```
prop.table(table(grad.p1$sex, grad.p1$result),1)
```

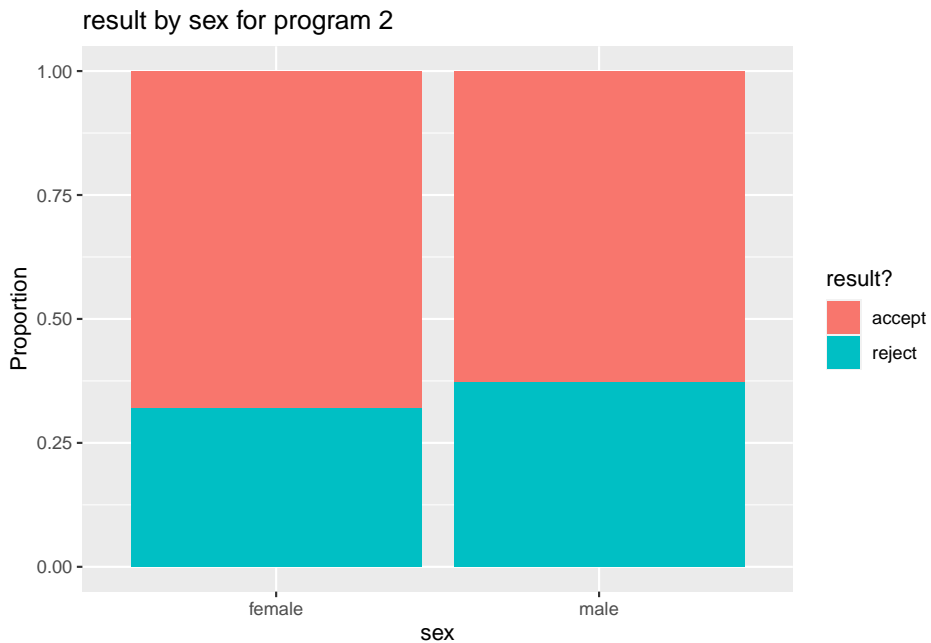
	accept	reject
female	0.8240741	0.1759259
male	0.6193939	0.3806061

(h). Result by sex for program 2.

- Repeat part (g) but this time use the program 2 data set. Compare the two bar graphs for (g) and (h) and explain how they show that females have a higher acceptance rate after accounting for program type (1 or 2).

Click for answer

```
# enter R code for (h) here
ggplot(grad.p2, aes(x = sex, fill = result)) +
  geom_bar(position = "fill") +
  labs(y="Proportion", title = "result by sex for program 2",
  fill = "result?", x = "sex")
```



```
prop.table(table(grad.p2$sex, grad.p2$result),1)
```

	accept	reject
female	0.6800000	0.3200000
male	0.6285714	0.3714286

*Answer:* For both programs 1 and 2, we see that female applicants have a slightly higher rate of acceptance than male applicants. After accounting for program type, we now see that black defendants have a higher rate of death penalty than white defendants. Without accounting for program type, the opposite was true (see parts (c) and (e)).

Why? the confounding affect of program type which is associated with both result and sex:

Click for answer

- females prefer to apply to programs 3 and 4 while males prefer programs 1 and 2 (more than 3 and 4).
  - 44% of females applied to program 3 and 40% to program 4
  - 38% of males applied to program 1 and 26% to program 2

```
prop.table(table(grad$sex, grad$program), 1)
```

	program1	program2	program3	program4
female	0.12720848	0.02944641	0.44169611	0.40164900
male	0.38106236	0.25866051	0.18799076	0.17228637

-Programs 3 and 4 were much harder to get into than programs 1 and 2 - 64% of applicants to program 1 were accepted and 63% of applicants to program 2 were accepted - 6% of applicants to program 4 were accepted and 34% of applicants to program 3 were accepted

```
prop.table(table(grad$program, grad$result), 1)
```

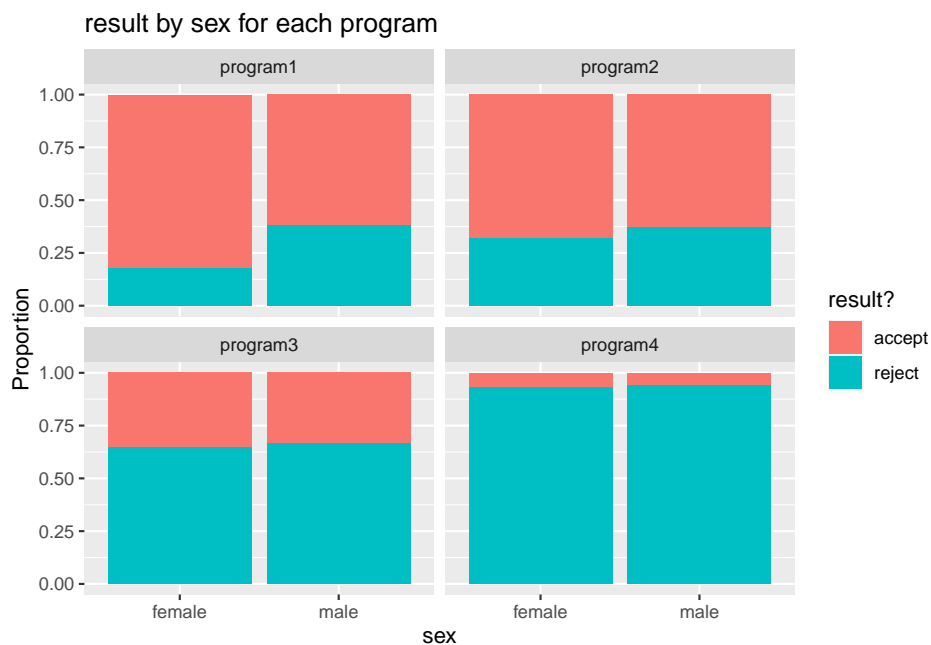
	accept	reject
program1	0.64308682	0.35691318
program2	0.63076923	0.36923077
program3	0.34398977	0.65601023
program4	0.06442577	0.93557423

So since the majority of females applied to the toughest programs (as measured by acceptance rates), there overall rate of acceptance was lower for females compared to males. But when we break down these rates by program type, we see that females have higher acceptance rates than males (see the visual in part (i)).

(i). A bar graph with three variables

If we simply want to graph the relationship between result and sex for each type of program, we can avoid subsetting the data by using the `facet_wrap` command in `ggplot2`. It is one simple addition to the stacked bar graph in part (e):

```
ggplot(grad, aes(x = sex, fill = result)) +
  geom_bar(position = "fill") +
  labs(y = "Proportion",
       title = "result by sex for each program",
       fill = "result?",
       x = "sex") +
  facet_wrap(~program)
```



- Verify that this command creates side-by-side stacked bar graphs that match your graphs in parts (g) and (h) for programs 1 and 2.

Click for answer

*Answer:* The graphs match.

### 4.3 Quiz

1. A two-way table is shown for two groups, 1 and 2, and two possible outcomes, A and B.

	Outcome A	Outcome B	Total
Group 1	40	10	50
Group 2	30	120	150
Total	70	130	200

What proportion of all cases are in Group 1?

- A. 0.33

B. 0.20

C. 0.25

D. 0.75

Click for answer

C. 0.25

**2. A disruption of a gene called DYXC1 on chromosome 15 for humans may be related to an increased risk of developing dyslexia. Researchers studied the gene in 109 people diagnosed with dyslexia and in a control group of 195 others who had no learning disorder. The DYXC1 break occurred in 10 of those with dyslexia and in 5 of those in the control group. Is this an experiment or an observational study?**

A. Experiment

B. Observational Study

Click for answer

Observational Study

**3. The data from question 3 can be summarized in a two way table as:**

	Gene Break	No Break	Total
Dyslexia Group	10	99	109
Control Group	5	190	195
Total	15	289	304

**What is the proportion of Dyslexia group who have the break on the DYXC1 gene? Round your answer to 3 significant digits after the decimal.**

A. 0.026

B. 0.667

C. 0.127

D. 0.092

Click for answer

D. 0.092



## Chapter 5

### (PART\*) Basics R





## Chapter 6

# 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.

### 6.1 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.

### 6.2 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.)

## 6.3 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.

## 6.4 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.

## 6.5 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.

## 6.6 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.

## 6.7 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.

## 6.8 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 7

# 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 |

## 7.1 Including Plots

You can also embed plots. See Figure 7.1 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  
)
```

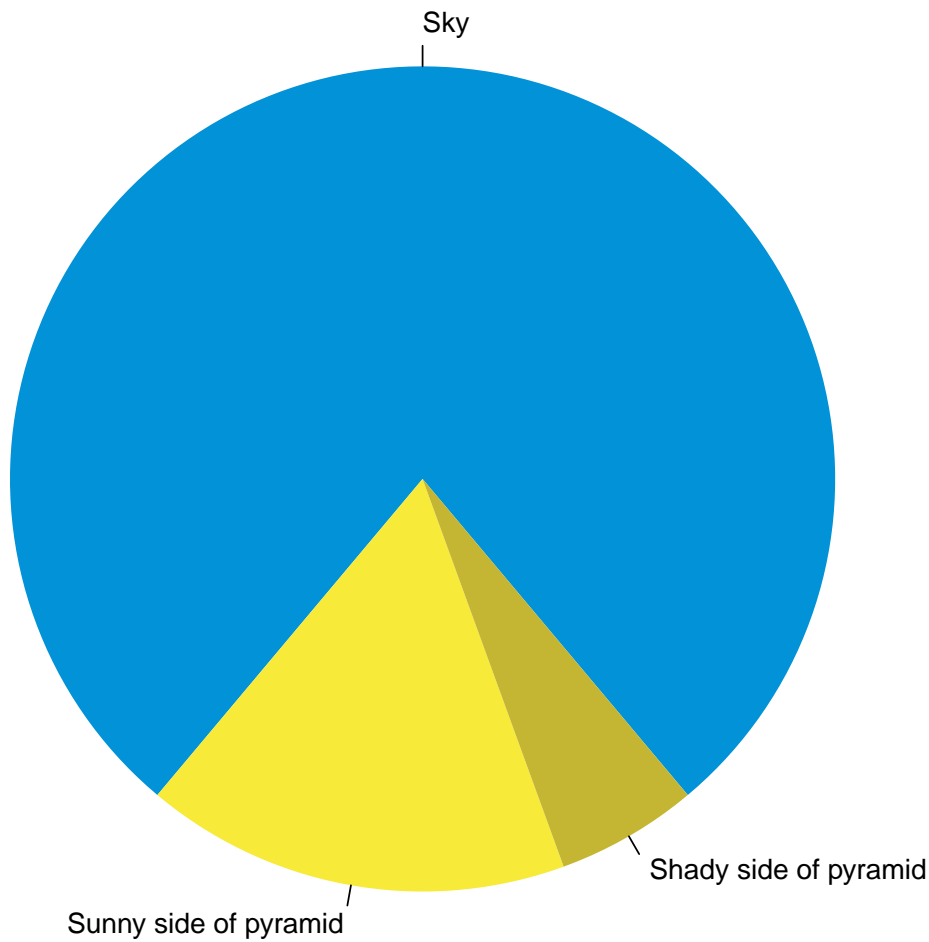


Figure 7.1: A fancy pie chart.

(Credit: Yihui Xie)

## 7.2 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 |

### 7.3 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 |