MA256 Course Guide

Update: 6 JULY 2020

Task

Develop critical statistical thinkers who can generate precise research questions; identify, collect, and analyze relevant data; and translate this analysis to a complete and correct response that answers the original question.

Purpose

Our daily lives are full of question whose answers should be informed by the growing amount of data surrounding us. Politicians, scientists, sports teams, and military leaders are all looking toward quantitative analysis to provide them with unbiased information with which to make the correct decision. In this course you will gain valuable experience using data in an investigation process to collect, analyze, and report your results to help answer relevant research questions.

You will gain experience in not only completing statistical analysis but also asking the important follow-up questions about why you are getting these results and how you can expect them to change if the experiment was repeated. You will also gain experience in reading, comprehending, and reproducing the results of published scientific papers. All of these tools will help to equip you to be successful in the course project where you are provided the opportunity to do research on a question that interests you. Perhaps you will discover, like some of your fellow cadets that, indeed, more cadets do report to sick call on Mondays.

Grading

Below is the point break-down for this course:

Category	Points
Homework	**350**
Block 1	50
Block 2	125
Block 3	125
Block 4	50
Midterm	**150**
Course Project	**250**
Proposal	25
IPR 1	25
IPR 2	25
Presentation	75
Report	100
TEE	**250**
Total	**1000**

How to use this guide:

The purpose of this course guide is to give you a single document that lists what you need to complete for this course. Given the nature of the learning environment we are in this semester, you can expect that there will be changes to this guide. I will attempt to notify you whenever there are large updates but the version

available on my GitHub (and at the top of the Teams) will always be the latest version.

You will see that there are "Before Class Activities" and "After Class Activities." Hopefully these are pretty self-explanatory. I include "After Class Activities" because I feel that a lot of the "homework" you have will be easier if you complete/think about right after a given class period as opposed to right before the next. I won't be checking though so you do what works best for you.

I have adopted a "Read - Watch - Do" model for laying out your activities. These tasks are the minimum you are expected to complete. There will be some videos assigned that are located under the "Modules" section under our course in WileyPlus. There are a lot more videos available on WileyPlus and please feel free to watch extra videos if it will aid in your understanding. Your textbook also has a lot of exercises if you want some extra practice.

I want to instill a sense of responsibility and ownership of your learning in this course. I have a high expectation of your level of mastery of the concepts, you do what is necessary to get yourself to that level.

Block 1: Question Formulation and Data Exploration

Lesson 1: Course Introduction and RStudio Familiarization

Objectives:

- Introduction to MA256 and course expectations
- Installation and familiarization with RStudio and R
- Tidyverse Tutorial

Before Class Activities:

- Read: *Introduction to Statistical Investigations* - Preliminary 1 (P.1)
- Watch: WileyPlus - Videos P.1.1 - P.1.4 (Optional but recommended)
- Do: Consult *Introduction to RStudio and Tidyverse* (AKA *Tidyverse Tutorial*) and follow the "Download R and RStudio" instructions.
- Do: Take a look at Hadley Wickham's book *R for Data Science* which is available at the link below. This is a great resource for your coding questions as it was written by the guy who wrote a large part of the *Tidyverse*. https://r4ds.had.co.nz/

After Class Activities:

- Ensure that R/RStudio is installed and functional on your computer.
- Carefully go through the *Tidyverse Tutorial*. Attention now will save you a great deal of time later.
- Consider the dataset that we discussed in class today, what sort of research questions could you ask based on the data and variables present in this dataset. No need to answer the questions, but develop three good research questions before next class.

Lesson 2: Research Questions

Objectives:

- Understanding types of and relationships between variables
- Formulation of specific and focused research questions
- Investigation of the "state of the art"
- Presentation of research in a literature review

Before Class Activities:

- Read: *Introduction to Statistical Investigations* - Chapter 4
- Watch: WileyPlus - Videos 4.1.1 - 4.1.5 and Videos 4.2.1 - 4.2.3

- Do: Exploration 4.2 - Problems 1 - 14
- Do: Prepare the three research questions referenced in **Lesson 1 - After Class Activities**.

After Class Activities:

- Pick one of your group's research questions and prepare a literature review for at least two scholarly articles. One submission per group. This will be due at the start of **Lesson 4**.

Lesson 3: Data Exploration - 1

Objectives:

- Understand terminology used in describing distributions
- Understand the information conveyed in a boxplot and five-number summary
- Create appropriate data visualizations in *R*
- Understand reasons for and conclusions drawn from data exploration

Before Class Activities:

- Read: *Introduction to Statistical Investigations* - Preliminary 2 (P.2)
- Read: *Introduction to Statistical Investigations* - Section 6.1
- Watch: WileyPlus - Videos P.2.1 - P.2.5 (Optional but recommended)
- Do: Work on the literature review referenced in **Lesson 2 - After Class Activities**.
- Do: Review and replicate the code examples shown below.

Example Code

There is a lot of example code and important points made in the *Tidyverse Tutorial* so please review the **ggplot2** section again if you have any questions. For this example code I'm going to use the geyser data referenced in the P.2 section of your book. This data is available from the book website at the link included in the code.

```
library(tidyverse)

geyser = read_table2("http://www.isi-stats.com/isi/data/prelim/OldFaithful2.txt")

head(geyser)
```

```
## # A tibble: 6 x 2
##   EruptionType TimeBetween
##   <chr>              <dbl>
## 1 short                 55
## 2 short                 58
## 3 short                 56
## 4 short                 50
## 5 short                 51
## 6 short                 60
```

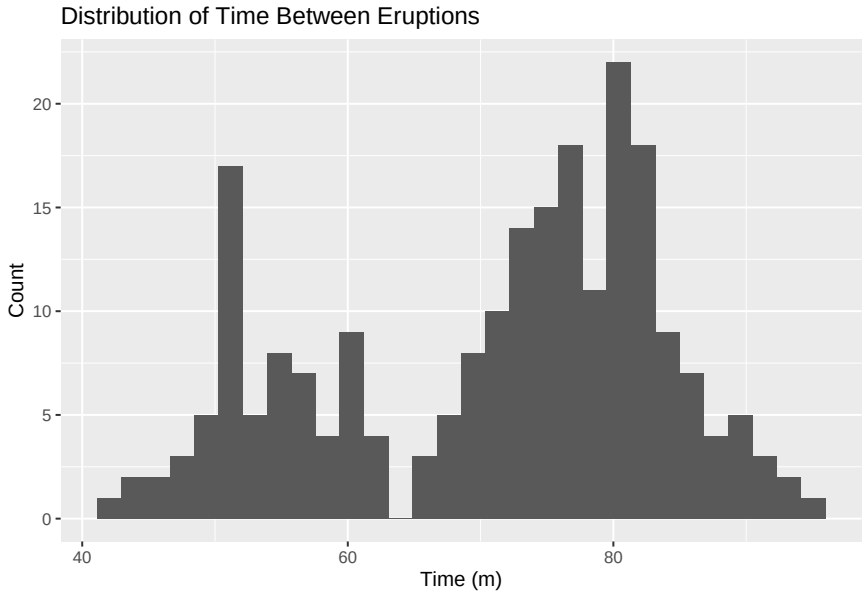
The code above turns on the awesome (loads the tidyverse library) and then reads in our data from the book's website. Notice we use the **read_table2** command here (as opposed to the **read_csv**) covered in the *Tidyverse Tutorial*. I'm doing this because we aren't dealing with a comma-seperated values file but rather a file in which the values for each variable are seperated by the tab character. I find that **read_table2** is a good way to deal with these sort of files.

The last bit of code offers a quick glimpse into the first six rows of the dataframe. It also allows us to ensure that the variables have been "read in" correctly. The *EruptionType* variable was read in as a character (chr) and the *TimeBetween* variable was read in as a double (dbl) so we are good. Occasionally your categorical

variables will be read in as some type of number (integer, double, float, long, etc.) which can cause problems later on.

As part of the data exploration process for this dataset, I want to create a visualization that will allow me to get an idea of the distribution of times between eruptions. There are two way to consider this: with regard to the *EruptionType* and without. Let's start with without considering it.

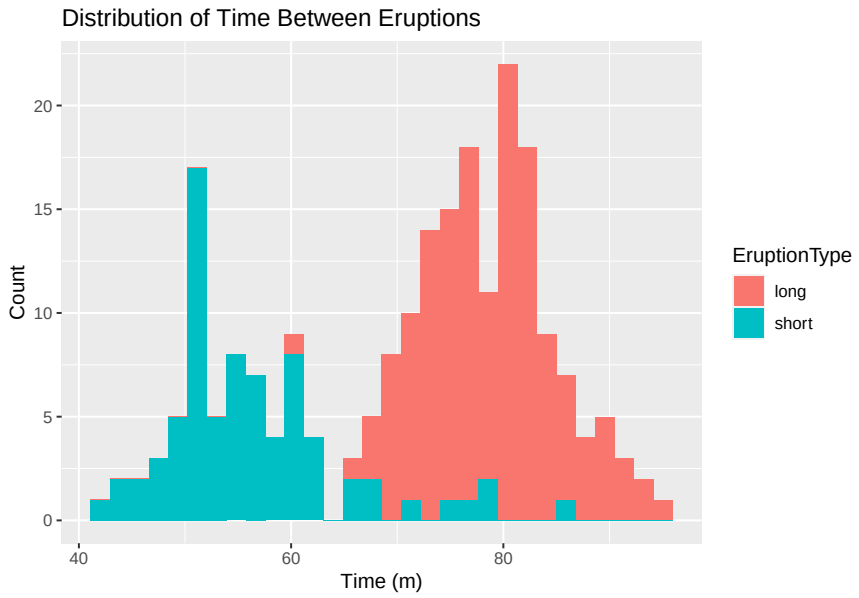
```
geyser %>%
  ggplot(aes(x = TimeBetween)) +
  geom_histogram() +
  labs(x = "Time (m)", y = "Count", title = "Distribution of Time Between Eruptions")
```



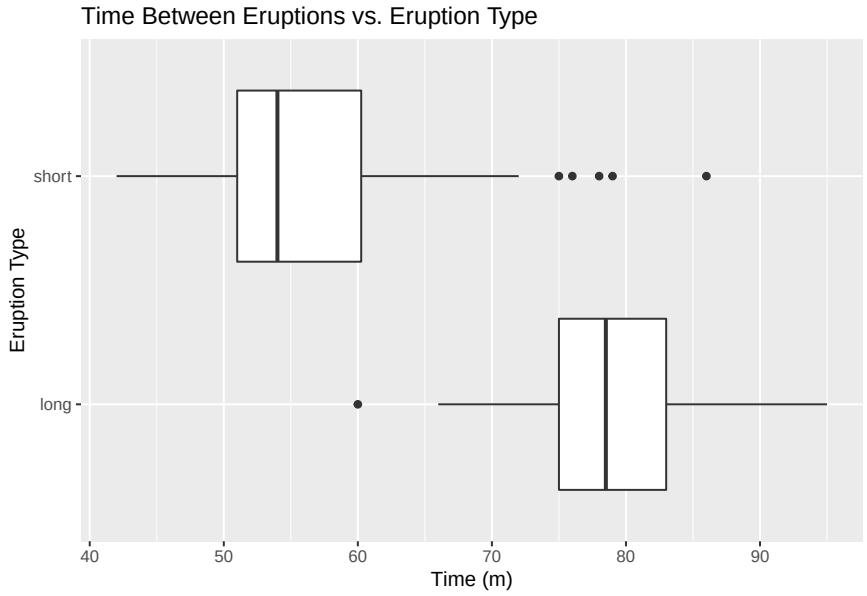
We can get an idea of the shape, center, variability, and unusual observations from this histogram. We can see that we are dealing with a *bi-modal* dataset, a conclusion that is reinforced by the obvious split in the data. We know that we have this other variable *EruptionType* but imagine that we didn't know about that: seeing this split should lead you to question whether there is some additional data you should be collected to try to explore why there is a bifurcation of our data. Furthermore, remember that our second variable is *EruptionType* and this is a split in the times **between** eruptions so, even though we have this second variable to explore, there may still be more information our there we would like to gather.

Now let's look at a couple visualizations that take into account our second variable of *EruptionType*.

```
geyser %>%
  ggplot(aes(x = TimeBetween, fill = EruptionType)) +
  geom_histogram() +
  labs(x = "Time (m)", y = "Count", title = "Distribution of Time Between Eruptions")
```




```
geyser %>%
  ggplot(aes(x = TimeBetween, y = EruptionType)) +
  geom_boxplot() +
  labs(x = "Time (m)", y = "Eruption Type", title = "Time Between Eruptions vs. Eruption Type")
```



These two visualizations allow us to gain some information about the relationship between this two variables in our sample. We can now describe the shape, center, variability, and unusual observations for each group in the histogram. The side-by-side boxplot can provide very similar information but in a different, and perhaps easier to understand form. Please reference Section 6.1 for more information about the specific parts of the boxplot.

After Class Activities:

- Finalize and submit the literature review referenced in **Lesson 2 - After Class Activities**.

Lesson 4: Data Exploration - 2

Objectives:

- Demonstate proficiency of data exploration skills

Before Class Activities:

- Read: *Introduction to Statistical Investigations* - Preliminary 3 (P.3)
- Watch: No videos for this lesson.
- Do: Ensure you have the *NYPD Arrest* data loaded for this lesson.
- Do: Literature review from **Lesson 2** is due before the start of class.

After Class Activities:

- Finalize your work from class. Submission is due before the start of **Lesson 5**. One submission per group.

Block 2: Questions of Single Variables

Lesson 5: Single Categorical Variable - 1

Objectives:

- Understand chance models
- Calculate and utilize various "strength of evidence" measures
- Understand what affects the strength of evidence
- Apply the simulation-based and theory-based approach to solving these problems

Before Class Activities:

- Read: *Introduction to Statistical Investigations* - Chapter 1
- Watch: WileyPlus - Videos 1.5.1 - 1.5.3
- Do: *Introduction to Statistical Investigations* - Exploration 1.4
- Do: Submit your work from *Lesson 4* before the start of class.

After Class Activities:

- Ensure you understand the content discussed in class today... you'll get a chance to prove it next class.

Example Code:

Simulating a Null Distribution:

While the applets make it very easy to simulate a null distribution, I want to provide you with some code to demonstrate how to do a similar thing in *R*. I find that more exposure to code make students better coders (radical thought... I know) so don't just skip over this section. I will use the rock-paper-scissors example in your book for this code.

```
library(tidyverse)

replications_dataframe = NULL

num_reps = 1000
sample_size = 12
null_prob = 1/3
sample_stat = 0.167
```

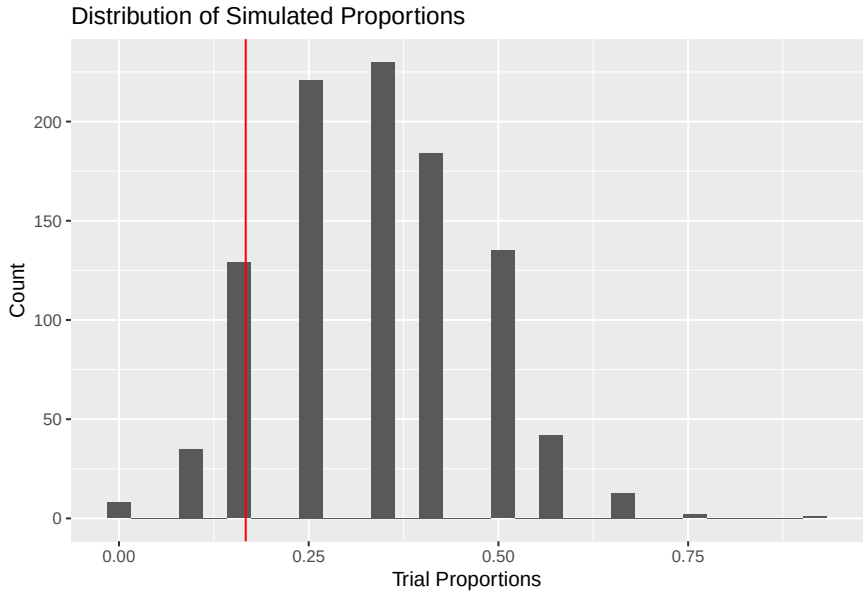
```
for (i in 1:num_reps){

  trial = sample(x = c(1,0),
                 size = sample_size,
                 prob = c(null_prob, 1-null_prob),
                 replace = TRUE)

  trial_proportion = sum(trial)/sample_size

  replications_dataframe = rbind(replications_dataframe, data.frame(trial_proportion))

}
```

After turning on the awesome, this code sets up several "parameters" for our simulation. This is "parameter" in the general sense, not to be confused with statistical parameters. We create a dataframe, or storage spot, to put our simulated proportions. We then define the number of repetitions (1000), the size of each simulated repetitions (12... based on the problem in the book), the null probability (probability of getting scissors under the null hypothesis), and the sample statistic (what we observed in our experiment).

What follows is a **for** loop that serves as the "simulation" in our experiment. The code here is a bit more advanced but nothing insurmountable. The variable *trial* is basically a list of size 12 (sample size) of ones and zeros. This list is populated using the **sample** function which chooses from the vector we give it in x with the probability of each elements being chosen as defined by the *prob* parameter in the function. Said another way: the function will choose *1* with a probability of *null_prob* and *0* with the probability of *1-null_prob*. Finally, we need to sample *with replacement* otherwise you can't get twelve numbers out of a list of two.

We calculate the proportion of scissors (ones) in our vector by summing the elements of the vector (again, a bunch of ones and zeros) and dividing by the sample size. Finally, we take this *trial_proportion* and "bind" it to the end of our big list of simulated proportions. What we are left with is 1000 proportions that are simulated under the assumption that the probability of picking scissors is 1/3... the null hypothesis. We can visualize the distribution of these numbers using a histogram.

```
replications_dataframe %>%
  ggplot(aes(x = trial_proportion)) +
  geom_histogram() +
  labs(x = "Trial Proportions", y = "Count", title = "Distribution of Simulated Proportions") +
  geom_vline(xintercept = sample_stat, color = "red")
```



One-proportion z-test

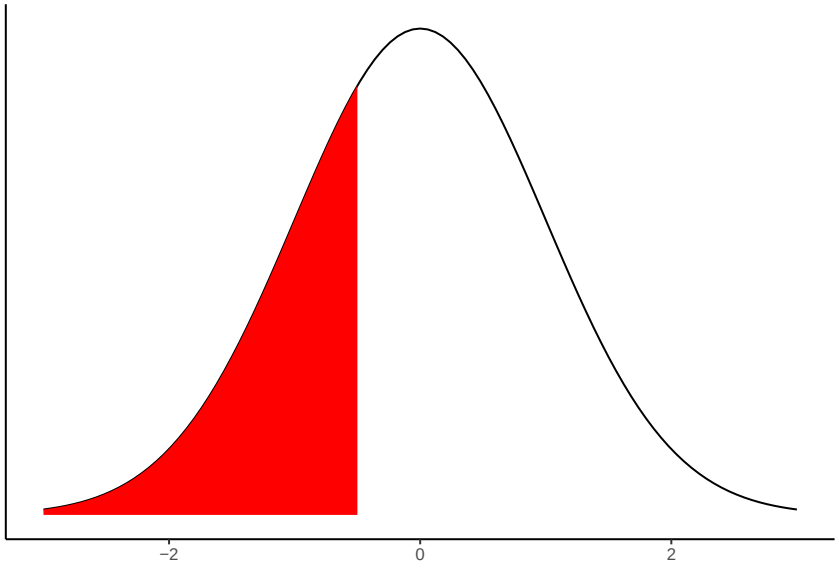
Here is some example code for conducting a one-proportion z-test in *R*. I will use the "Halloween Treats" example from your textbook.

```
sample_proportion = 148/(135+148)

std_null = sqrt(0.5 * (1-0.5) / 283)

z = (sample_proportion - 0.5) / std_null

p = 2 * (1 - pnorm(abs(z)))

p
```

```
## [1] 0.4396586
```

Hopefully, the first three steps here are pretty clear from your reading in your book. The last equation (to get the p-value) may be a little new because of the **pnorm** function. Please spend some time experimenting with this function becuase it is going to become very useful to you in this class. You can see this function as saying "give me an x and I'll give you an area under the normal curve from negative infinity to that x." Now maybe you think I'm crazy because functions talk to me but that's ok as long as you remember that. Here are a couple examples of using the **pnorm** function. Note: You won't be responsible for knowing all functions below but I'm using a few new ones to illustrate my point.

```
ggplot(NULL, aes(x = c(-3, 3))) +
  stat_function(fun = dnorm,
                args = list(mean = 0, sd = 1)) +
  stat_function(fun = dnorm,
                args = list(mean = 0, sd = 1),
                xlim = c(-3, -0.5), #This is (-3, 0) for the purposes of plotting.
                geom = "area", fill = "red") +
  labs(x = "", y = "") + theme_classic() +
  scale_y_continuous(breaks = NULL)
```

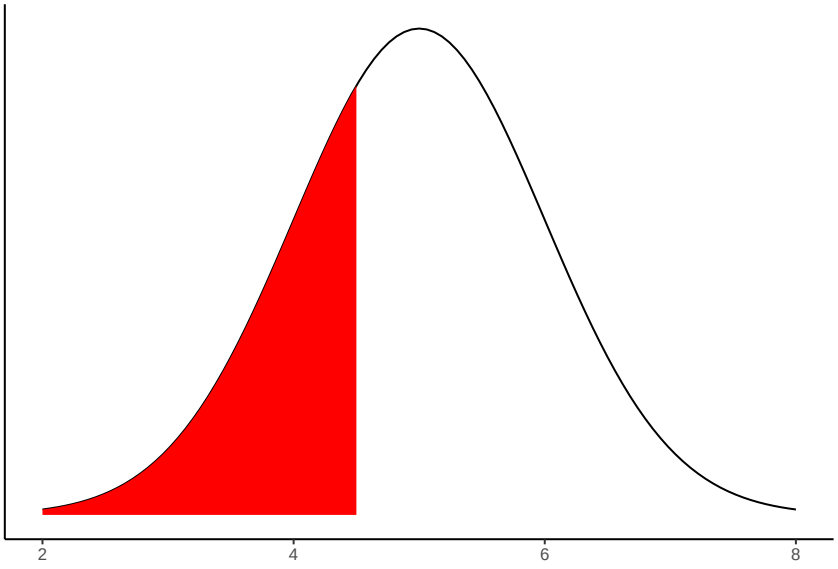



```
pnorm(-0.5)
```

```
## [1] 0.3085375
```

In this case, **pnorm** gives us the area of the region shaded in red above. When we use **pnorm** without any other parameters, it defaults to the *standard normal distribution* which has a mean of 0 and a standard deviation of 1. One of the great things about *R*, however, is that we don't have to "standardize" our sample statistics to calculate our areas. We can also do something like the following two examples if one or both of our parameters are different than the standard normal.

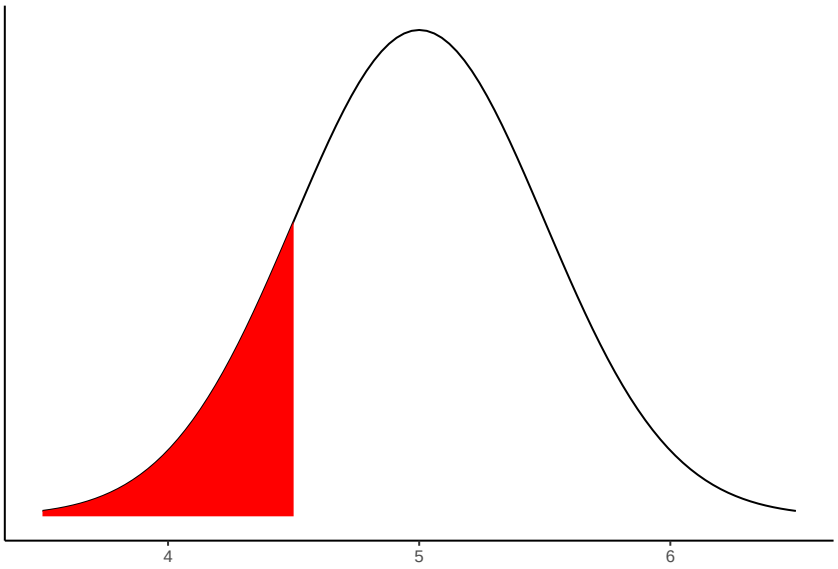
```
ggplot(NULL, aes(x = c(2, 8))) +
  stat_function(fun = dnorm,
                args = list(mean = 5, sd = 1)) +
  stat_function(fun = dnorm,
                args = list(mean = 5, sd = 1),
                xlim = c(2, 4.5),
                geom = "area", fill = "red") +
  labs(x = "", y = "") + theme_classic() +
  scale_y_continuous(breaks = NULL)
```




```
pnorm(4.5, mean = 5, sd = 1)
```

```
## [1] 0.3085375
```

```
ggplot(NULL, aes(x = c(3.5, 6.5))) +
  stat_function(fun = dnorm,
                args = list(mean = 5, sd = 0.5)) +
  stat_function(fun = dnorm,
                args = list(mean = 5, sd = 0.5),
                xlim = c(3.5, 4.5),
                geom = "area", fill = "red") +
  labs(x = "", y = "") + theme_classic() +
  scale_y_continuous(breaks = NULL)
```




```
pnorm(4.5, mean = 5, sd = 0.5)
```

```
## [1] 0.1586553
```

Lesson 6: Single Categorical Variable - 2

Objectives:

- Demonstrate proficiency in answering research questions involving a single categorical variable.

Before Class Activities:

- Read: *Introduction to Statistical Investigations* - Chapter 3.1 - 3.2
- Watch: WileyPlus - Videos 3.1.1 - 3.1.3, 3.2.1 - 3.2.2
- Do: *Introduction to Statistical Investigations* - Exploration 3.2, Problems 1 - 11

After Class Activities:

- Finalize your work from class. Submission is due before the start of **Lesson 7**. One submission per group.

Lesson 7: Single Quantitative Variable - 1

Objectives:

- Understand the concepts of convenience sample, random sample, and sampling variability
- Apply the simulation-based and theory-based approach to solving these problems
- Understand the two types of statistical errors and relate them to the selection of a significance level

Before Class Activities:

- Read: *Introduction to Statistical Investigations* - Chapter 2
- Watch: WileyPlus - Videos 2.2.4 - 2.2.5, Videos 2.3.3/2.3.4 - 2.3.5
- Do: *Introduction to Statistical Investigations* - Exploration 2.1B
- Do: Submit your work from *Lesson 6* before the start of class.

After Class Activities:

- Ensure you understand the content discussed in class today... you'll get a chance to prove it next class.

Lesson 8: Single Quantitative Variable - 2

Objectives:

- Demonstrate proficiency in answering research questions involving a single quantitative variable.

Before Class Activities:

- Read: *Introduction to Statistical Investigations* - Chapter 3.3 - 3.5
- Watch: WileyPlus - Videos 3.3.1 - 3.3.2, 3.4.1.1 - 3.4.1.4
- Do: *Introduction to Statistical Investigations* - Exploration 3.5, Problems 1 - 6

After Class Activities:

- Finalize your work from class. Submission is due before the start of **Lesson 9**. One submission per group.