Lab 6: Normal, Binomial, and Poisson Distribution

Felicia Liu

10/07/2020

- Due date: Friday, October 9th at 11:59 PM.
- Submission process: Follow the submission instructions on the final page. Make sure you do not remove any `\newpage` tags or rename this file, as this will break the submission.

Helpful hints:

- Every function you need to use was taught during lecture! So you may need to revisit the lecture code to help you along by opening the relevant files on Datahub. Alternatively, you may wish to view the code in the condensed PDFs posted on the course website. Good luck!
- Knit your file early and often to minimize knitting errors! If you copy and paste code for the slides, you are bound to get an error that is hard to diagnose. Typing out the code is the way to smooth knitting! We recommend knitting your file each time after you write a few sentences/add a new code chunk, so you can detect the source of knitting errors more easily. This will save you and the GSIs from frustration! **You must knit correctly before submitting.**
- If your code runs off the page of the knitted PDF then you will LOSE POINTS! To avoid this, have a look at your knitted PDF and ensure all the code fits in the file (you can easily view it on Gradescope via the provided link after submitting). If it doesn't look right, go back to your .Rmd file and add spaces (new lines) using the return or enter key so that the code runs onto the next line.

Introduction:

This lab will wrap up some concepts about the Normal distribution. We will then look a binomial distribution and the normal approximation of the binomial. Finally, we will look at calculations under the Poisson distribution.

To prepare for this lab, take a minute to answer the following review questions.

1. What are the two parameters that define the Normal distribution? 2 parameters that define the Normal distribution are mu and sigma, where mu is the mean of the distribution and sigma is the standard deviation.
2. What type(s) of plots can we use to determine if data is approximately normally distributed? We can use a Q-Q plot to determine if data is approximately normally distributed.
3. How is a random variable determined to follow a binomial distribution? There is a fixed number of observations, each observation is independent, and the probability of success is the same for each observation.
4. How is a random variable determined to follow a Poisson distribution? The occurrences are all independent and the probability of an occurrence is the same over all possible intervals of the same size. An interval can be a finite interval of time or space.

Section 1: More practice with the Normal Distribution

```
library(dplyr)
library(ggplot2)
```

Eating disorders affect at least 9% of the population worldwide. One such eating disorder is anorexia which affects approximately 1 in 200 American women. One study was interested in the effects of different therapies in the treatment of eating disorders. 72 young women were recruited and assigned to 3 different groups: control, cognitive behavioural treatment (CBT), and family therapy. Their weights (in pounds) were recorded pre-treatment and post treatment.

```
# The data comes from an R package MASS. We will load it first
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
#save the dataset here
anorexia <- MASS::anorexia

# MASS has functions with the same names as common dplyr functions
# We will detach it now so that we can continue to use our dplyr functions
detach("package:MASS", unload = T)

head(anorexia, 10) # here are the first 10 rows of data
```

```
##    Treat Prewt Postwt
## 1   Cont  80.7   80.2
## 2   Cont  89.4   80.1
## 3   Cont  91.8   86.4
## 4   Cont  74.0   86.3
## 5   Cont  78.1   76.1
## 6   Cont  88.3   78.1
## 7   Cont  87.3   75.1
## 8   Cont  75.1   86.7
## 9   Cont  80.6   73.5
## 10  Cont  78.4   84.6
```

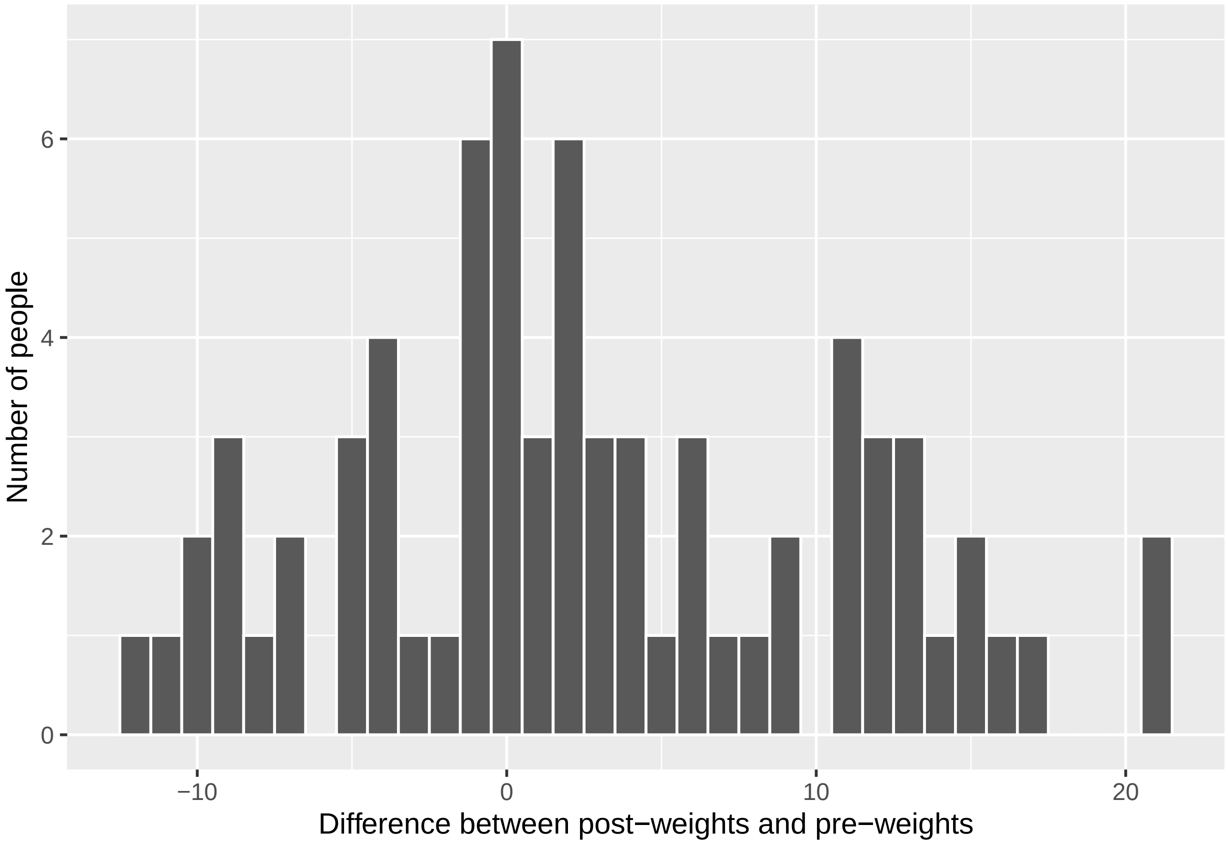
1. First, create a new dataset called ***anorexia_diff*** *that is the same as anorexia with a new column* ***diff*** *that is the difference between the weight after treatment and the weight before treatment for these women.*

```r
anorexia_diff <- anorexia %>% mutate(diff = Postwt - Prewt)

#dim(anorexia_diff) #uncomment this line to check the dimensions of your new dataset


check_problem1()
```

```
## [1] "Checkpoint 1 Passed: anorexia_diff has 72 rows and 4 columns."
## [1] "Checkpoint 2 Passed: You  did the correct mutation!"
##
## Problem 1
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

2. Visualize the distribution of the variable ***diff****. Choose an appropriate binwidth.*

```r
p2 <- ggplot(data = anorexia_diff, aes(x = diff)) +
  geom_histogram(col = "white", binwidth = 1) +
  labs(x = "Difference between post-weights and pre-weights", y = "Number of people")

p2
```



```
check_problem2()
```

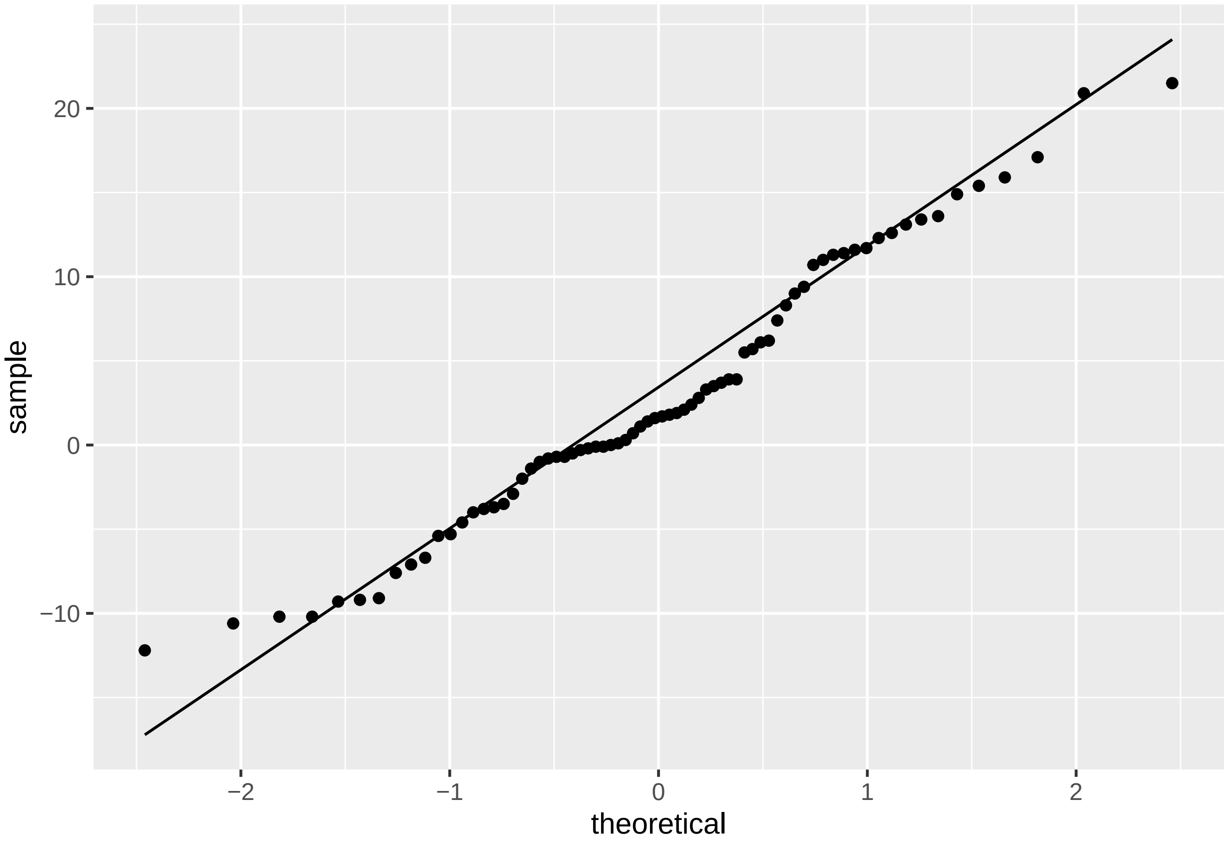
```
## [1] "Checkpoint 1 Passed: You used anorexia_diff"
## [1] "Checkpoint 2 Passed: diff is on the x axis"
## [1] "Checkpoint 3 Passed: A histogram has been defined in ggplot"
##
## Problem 2
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

3. Describe the distribution of the diff variable.

The distribution is unimodal and slightly skewed to the right.

4. Compare this data to a Normal distribution using a different ggplot2 function. Determine if a Normal distribution is a good fit for this data.

```
p4 <- ggplot(anorexia_diff, aes(sample = diff)) +
  stat_qq() +
  stat_qq_line()

p4
```



```
check_problem4()
```

```
## [1] "Checkpoint 1 Passed: You used anorexia_diff"
## [1] "Checkpoint 2 Passed: You are looking at the distribution of diff"
## [1] "Checkpoint 3 Passed: A QQplot has been defined in ggplot"
##
## Problem 4
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

Normal distribution is not a good fit for the data because the points to not match up closely with the line. The closer to a straight line, the closer to normal distribution.

5. Assume this data is sampled from a population distribution that is approximately Normal with mean 2 pounds and standard deviation 7 pounds. Find the probability that a randomly chosen women suffering from anorexia gains 5 pounds or more over the course of the treatment. You may leave this as a number between 0 and 1 and do not need to round.

```
p5 <- 1 - pnorm(q = 5, mean = 2, sd = 7)

p5
```

```
## [1] 0.3341176
```

```
check_problem5()
```

```
## [1] "Checkpoint 1 Passed: You calculated the correct probability!"
```

```
##
## Problem 5
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

6. Using the information above, find the number of pounds a randomly chosen woman would need to gain in order to be in the 90th percentile according to this probability distribution.

```r
p6 <- qnorm(p = 0.9, mean = 2, sd = 7)

p6
```

```
## [1] 10.97086
```

```r
check_problem6()
```

```
## [1] "Checkpoint 1 Passed: You calculated the correct number of pounds!"
##
## Problem 6
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

Section 2: Binomial Distribution and Normal Approximation

Example from Baldi and Moore

Antibiotic resistance occurs when disease-causing microbes no longer respond to antibiotic drug therapy. Because such resistance is typically genetic and transferred to the next generations of microbes, it is a very serious public health problem. Gonorrhea is the second most commonly reported notifiable disease in the US. According to the CDC, 27% of Gonorrhea cases tested in 2010 were resistant to at least one of the three major antibiotics commonly used to treat sexually transmitted diseases. A physician treated 20 cases of Gonorrhea at some point in 2010.

7. Let X represent the number of patients with antiobiotic resistance seen by this physician. Use notation you learned in lecture to show the distribution that X follows.

X ~ Binom(n = 20, p = 0.27)

8. Calculate (by hand) the probability that exactly 5 people have antibiotic resistance. You can use the `choose(n = , k = )` function in R to help. Confirm your results using an R function.

20 choose 5 = 15504

P(X = 5) = 15504 * (0.27)^5 * (0.73)^(15) = 0.1982

```
p8 <- dbinom(x = 5, size = 20, prob = 0.27)



p8
```

```
## [1] 0.1982008
```

```
check_problem8()
```

```
## [1] "Checkpoint 1 Passed: You calculated the correct probability!"
##
## Problem 8
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

9. Calculate (by hand) the probability that more than 1 person has antibiotic resistance. Confirm your answer using R. Hint: work smarter not harder.

20 choose 1 = 20

P(X > 1) = 1 - (20 * (0.27)^1 * (0.73)^19)= 0.986

```
p9 <- 1 - pbinom(q = 1, size = 20, prob = 0.27)



p9
```

```
## [1] 0.9844906
```

```
check_problem9()
```

```
## [1] "Checkpoint 1 Passed: You calculated the correct probability!"
##
## Problem 9
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

Suppose in one city in the Western United States there were 812 cases of gonorrhea in a population of 100,000. The probability of antibiotic resistance to at lease one major antiobtic remains the same at approximately 27 percent.

10. Calculate the expected number of antibiotic resistant cases of gonorrhea in this population. Make sure to round to the nearest whole number. Also calculate the standard deviation. Round this number to two decimal places.

```
expected_value <- round(812 * 0.27)
standard_deviation <- round(sqrt(812 * 0.27 * 0.73),  2)

expected_value
```

```
## [1] 219
```

```
standard_deviation
```

```
## [1] 12.65
```

```
check_problem10()
```

```
## [1] "Checkpoint 1 Passed: You calculated the correct expected value!"
## [1] "Checkpoint 2 Passed: You calculated the correct standard deviation!"
##
## Problem 10
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

11. We learned in class that the binomial distribution can be approximated by the Normal distribution under some conditions. List the conditions below and determine if this problem satisfies them.

You can use the Normal distribution to perform calculations when data is binomially distributed with a large sample size n. This problem satisfies this condition because we were able to use binomial distribution and the sample size is relatively large (n = 812).

12. Let's generate some data from this distribution to check the normal approximation!

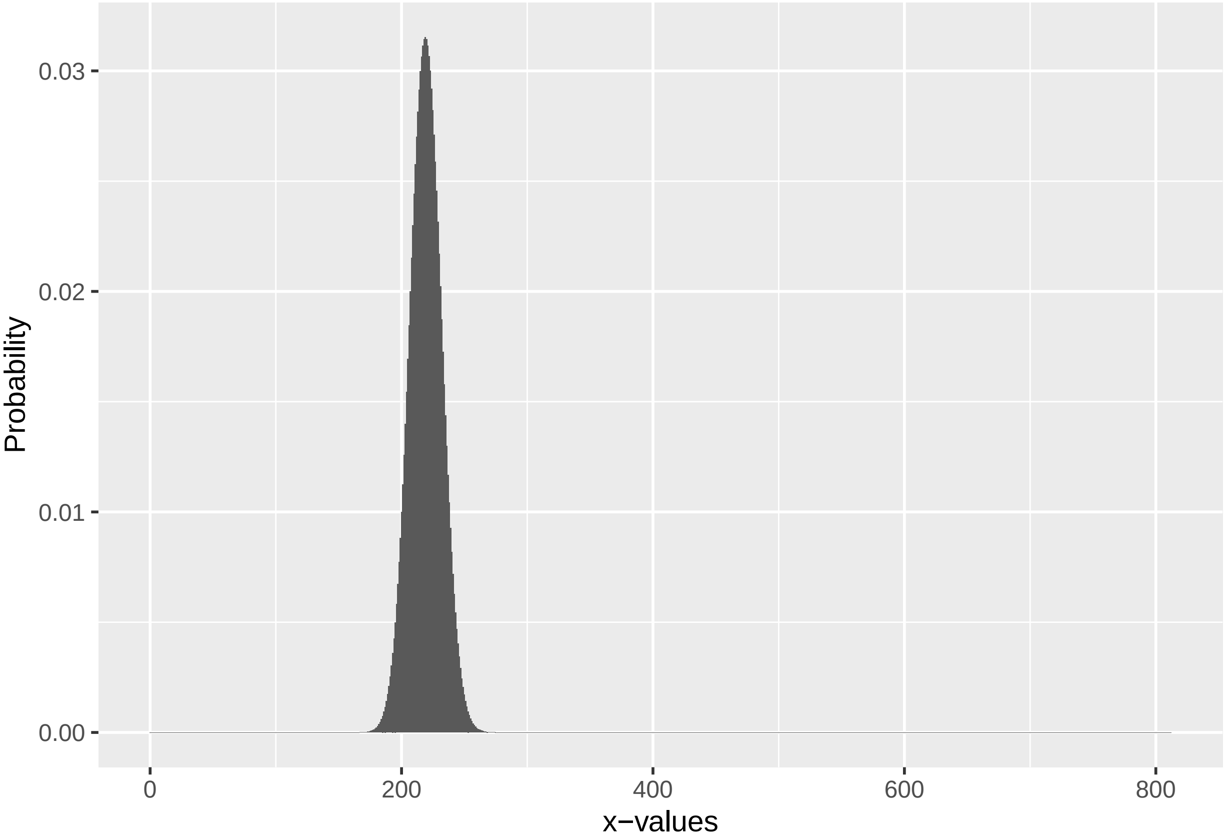
The first step is to generate the probabilities of observing each of the possible values of X ~ Binom(n = 812, p = 0.27). We will use the familiar dbinom() function to do this. However, instead of just plugging in one value, we will plug in the entire range of values (0 through 812) and save it as a vector called `obs_data`.

Note: You will not be tested on this use of code but you should understand what's happening at every step. It is useful to print out the object in your console to get an idea of what's happening at each stage.

```
#this is just the range of values x can take
x_vals <- 0:812

# this generates the probabilities
probs <- dbinom(0:812, size = 812, prob = 0.27)

#this combines them together as a dataframe with 2 columns: x_vals and probs.
#View(obs_data) in your console to see what this dataframe looks like
obs_data <- as.data.frame(cbind(x_vals, probs))
```

Now use ggplot2 to plot a histogram of `obs_data` with `x_vals` on the x axis and the respective probabilities on the y axis.

```
p12 <- ggplot(data = obs_data, aes(x = x_vals, y = probs)) +
  geom_bar(stat = "identity") +
  labs(x = "x-values", y = "Probability")


p12
```



```
check_problem12()
```

```
## [1] "Checkpoint 1 Passed: You used obs_data"
## [1] "Checkpoint 2 Passed: x_vals is on the x axis"
## [1] "Checkpoint 3 Passed: probs is on the y axis"
## [1] "Checkpoint 4 Passed: A histogram has been defined in ggplot"
##
## Problem 12
## Checkpoints Passed: 4
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

Section 3: Poisson Distribution

Example from Baldi and Moore

The state of New York reported 1484 live births in which the infants had Down syndome between 2006 and 2010, which averages to about 5.7 cases per week. While the causes of Down syndrome are not fully understood, it is reasonable to assume that live births are independent and the weekly rate is constant. Let X be the count of babies born with Down syndrome in New York in a given week.

13. What distribution does X approximately follow? Write it using notation learned in lecture. What are the possible values X can take?

Distribution: X ~ P(5.7). The possible values that X can take are: 0, 1, 2, and so on. If k is any one of these values, then P(X = k) = (e^(-mu)*mu^k)/k!.

14. What are the mean and standard deviation of X?

```
mean <- 5.7
sd <- sqrt(5.7)



mean
```

```
## [1] 5.7
```

```
sd
```

```
## [1] 2.387467
```

```
check_problem14()
```

```
## [1] "Checkpoint 1 Passed: Mean is 5.7!"
## [1] "Checkpoint 2 Passed: SD is the sqrt(5.7)!"
##
## Problem 14
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

15. What is the probability that no child will be born with Down syndrome in a given week in New York? Calculate the probability by hand and confirm your answer in R.

P(X = 0) = (e^(-5.7)*5.7^0)/0! = 0.003346

```
p15 <- dpois(x = 0, lambda = 5.7)

p15
```

```
## [1] 0.003345965
```

```
check_problem15()
```

```
## [1] "Checkpoint 1 Passed: You computed the correct probability!"
##
## Problem 15
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

16. What is the probability that 2 or more children will be born with Down sydrome in a given week in New York? Calculate the probability by hand and confirm your answer in R.

P(X >= 2) = 1 - (P(X = 0) + P(X = 1)) = 1 - ((e^(-5.7) * 5.7^0)/0! + (e^(-5.7) * 5.7^1)/1!) = 0.97758

```r
p16 <- 1 - ppois(q = 1, lambda = 5.7)



p16
```

```
## [1] 0.977582
```

```r
check_problem16()
```

```
## [1] "Checkpoint 1 Passed: You calculated the correct probability!"
##
## Problem 16
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

17. Use R to calculate the probability that more than 12 children are born with Down syndrome?

```r
p17 <- 1 - ppois(q = 12, lambda = 5.7)

p17
```

```
## [1] 0.005921731
```

```r
check_problem17()
```

```
## [1] "Checkpoint 1 Passed: You calculated the correct probability!"
##
## Problem 17
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

Check your score

Click on the middle icon on the top right of this code chunk (with the downwards gray arrow and green bar) to run all your code in order. Then, run this chunk to check your score.

```r
# Just run this chunk.
total_score()
```

```
##                      Test Points_Possible          Type
## Problem 1          PASSED               1    autograded
## Problem 2          PASSED               1    autograded
## Problem 3  NOT YET GRADED               1 free-response
## Problem 4          PASSED               1    autograded
## Problem 5          PASSED               1    autograded
## Problem 6          PASSED               1    autograded
## Problem 7  NOT YET GRADED               2 free-response
## Problem 8          PASSED               1    autograded
## Problem 9          PASSED               1    autograded
## Problem 10         PASSED               1    autograded
## Problem 11 NOT YET GRADED               1 free-response
## Problem 12         PASSED               1    autograded
## Problem 13 NOT YET GRADED               1 free-response
## Problem 14         PASSED               1    autograded
## Problem 15         PASSED               1    autograded
## Problem 16         PASSED               1    autograded
## Problem 17         PASSED               1    autograded
```

Submission

For assignments in this class, you'll be submitting using the **Terminal** tab in the pane below. In order for the submission to work properly, make sure that:

1. Any image files you add that are needed to knit the file are in the `src` folder and file paths are specified accordingly.
2. You **have not changed the file name** of the assignment.
3. The file knits properly.

Once you have checked these items, you can proceed to submit your assignment.

1. Click on the **Terminal** tab in the pane below.
2. Copy-paste the following line of code into the terminal and press enter.

```
cd; cd ph142-fa20/lab/lab06; python3 turn_in.py
```

3. Follow the prompts to enter your Gradescope username and password. When entering your password, you won't see anything come up on the screen–don't worry! This is just for security purposes–just keep typing and hit enter.
4. If the submission is successful, you should see "Submission successful!" appear as output.
5. If the submission fails, try to diagnose the issue using the error messages–if you have problems, post on Piazza.

The late policy will be strictly enforced, **no matter the reason**, including submission issues, so be sure to submit early enough to have time to diagnose issues if problems arise.