Lab 6: Normal, Binomial, and Poisson Distribution

Your name and Student ID

Today’s Date

* 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. What type(s) of plots can we use to determine if data is approximately normally distributed?
3. How is a random variable determined to follow a binomial distribution?
4. How is a random variable determined to follow a Poisson distribution?

### 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.*

anorexia\_diff <- "YOUR CODE HERE"  
  
#dim(anorexia\_diff) #uncomment this line to check the dimensions of your new dataset  
  
  
check\_problem1()

## [1] "Checkpoint 1 Error: anorexia\_diff should have 72 rows and 4 columns."  
## [1] "Checkpoint 2 Error: diff should be a column of Postwt-Prewt?"  
##   
## Problem 1  
## Checkpoints Passed: 0  
## Checkpoints Errored: 2  
## 0% passed  
## --------  
## Test: FAILED

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

p2 <- "YOUR CODE HERE"  
  
p2

## [1] "YOUR CODE HERE"

check\_problem2()

## [1] "Checkpoint 1 Error: Make sure to use anorexia\_diff."  
## [1] "Checkpoint 2 Error: You should be visualizing the diff column."  
## [1] "Checkpoint 3 Error: Did you define a histogram?"  
##   
## Problem 2  
## Checkpoints Passed: 0  
## Checkpoints Errored: 3  
## 0% passed  
## --------  
## Test: FAILED

*3. Describe the distribution of the diff variable.*

[TODO: Your answer here]

*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 <- "Your code here"  
  
p4

## [1] "Your code here"

check\_problem4()

## [1] "Checkpoint 1 Error: Make sure to use anorexia\_diff."  
## [1] "Checkpoint 2 Error: You should be visualizing the diff column."  
## [1] "Checkpoint 3 Error: Did you define a QQplot?"  
##   
## Problem 4  
## Checkpoints Passed: 0  
## Checkpoints Errored: 3  
## 0% passed  
## --------  
## Test: FAILED

[TODO: Your answer here]

*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 <- "Your code here"  
  
p5

## [1] "Your code here"

check\_problem5()

## [1] "Checkpoint 1 Error: Check your calculation."  
##   
## Problem 5  
## Checkpoints Passed: 0  
## Checkpoints Errored: 1  
## 0% passed  
## --------  
## Test: FAILED

*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.*

p6 <- "Your code here"  
  
p6

## [1] "Your code here"

check\_problem6()

## [1] "Checkpoint 1 Error: Check your calculation."  
##   
## Problem 6  
## Checkpoints Passed: 0  
## Checkpoints Errored: 1  
## 0% passed  
## --------  
## Test: FAILED

## 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.*

[TODO: Your answer here]

*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.*

[TODO: Your calculation here]

p8 <- "Your code here"  
  
  
p8

## [1] "Your code here"

check\_problem8()

## [1] "Checkpoint 1 Error: Check your calculation."  
##   
## Problem 8  
## Checkpoints Passed: 0  
## Checkpoints Errored: 1  
## 0% passed  
## --------  
## Test: FAILED

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

[TODO: Your calculation here]

p9 <- "Your code here"  
  
  
p9

## [1] "Your code here"

check\_problem9()

## [1] "Checkpoint 1 Error: Check your calculation."  
##   
## Problem 9  
## Checkpoints Passed: 0  
## Checkpoints Errored: 1  
## 0% passed  
## --------  
## Test: FAILED

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 <- "Your code/answer here"  
standard\_deviation <- "Your code/answer here"  
  
expected\_value

## [1] "Your code/answer here"

standard\_deviation

## [1] "Your code/answer here"

check\_problem10()

## [1] "Checkpoint 1 Error: Remember expected value is n\*p."  
## [1] "Checkpoint 2 Error: Remember sd is sqrt(n\*p\*(1-p))."  
##   
## Problem 10  
## Checkpoints Passed: 0  
## Checkpoints Errored: 2  
## 0% passed  
## --------  
## Test: FAILED

*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.*

[TODO: Your answer here]

*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 <- "Your code here"  
  
  
p12

## [1] "Your code here"

check\_problem12()

## [1] "Checkpoint 1 Error: Make sure to use obs\_data."  
## [1] "Checkpoint 2 Error: x\_vals should be on the x axis."  
## [1] "Checkpoint 3 Error: probs should be on the y axis."  
## [1] "Checkpoint 4 Error: Did you define a histogram?"  
##   
## Problem 12  
## Checkpoints Passed: 0  
## Checkpoints Errored: 4  
## 0% passed  
## --------  
## Test: FAILED

## 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?*

[TODO: Your answer here]

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

mean <- "Your code here"  
sd <- "Your code here"  
  
  
mean

## [1] "Your code here"

sd

## [1] "Your code here"

check\_problem14()

## [1] "Checkpoint 1 Error: Remember lambda is the mean!"  
## [1] "Checkpoint 2 Error: Remember that sd is the sqrt of lambda!"  
##   
## Problem 14  
## Checkpoints Passed: 0  
## Checkpoints Errored: 2  
## 0% passed  
## --------  
## Test: FAILED

*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.*

[TODO: Your answer here]

p15 <- "Your code here"  
  
p15

## [1] "Your code here"

check\_problem15()

## [1] "Checkpoint 1 Error: Check your calculation."  
##   
## Problem 15  
## Checkpoints Passed: 0  
## Checkpoints Errored: 1  
## 0% passed  
## --------  
## Test: FAILED

*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.*

[TODO: Your answer here]

p16 <- "Your code here"  
  
  
p16

## [1] "Your code here"

check\_problem16()

## [1] "Checkpoint 1 Error: Check your calculation"  
##   
## Problem 16  
## Checkpoints Passed: 0  
## Checkpoints Errored: 1  
## 0% passed  
## --------  
## Test: FAILED

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

p17 <- "Your code here"  
  
p17

## [1] "Your code here"

check\_problem17()

## [1] "Checkpoint 1 Error: Check your calculation"  
##   
## Problem 17  
## Checkpoints Passed: 0  
## Checkpoints Errored: 1  
## 0% passed  
## --------  
## Test: FAILED

### 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.

# Just run this chunk.  
total\_score()

## Test Points\_Possible Type  
## Problem 1 FAILED 1 autograded  
## Problem 2 FAILED 1 autograded  
## Problem 3 FAILED 0 autograded  
## Problem 4 FAILED 1 autograded  
## Problem 5 FAILED 1 autograded  
## Problem 6 FAILED 1 autograded  
## Problem 7 FAILED 0 autograded  
## Problem 8 FAILED 1 autograded  
## Problem 9 FAILED 1 autograded  
## Problem 10 FAILED 1 autograded  
## Problem 11 FAILED 0 autograded  
## Problem 12 FAILED 1 autograded  
## Problem 13 FAILED 0 autograded  
## Problem 14 FAILED 1 autograded  
## Problem 15 FAILED 1 autograded  
## Problem 16 FAILED 1 autograded  
## Problem 17 FAILED 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

1. 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.
2. If the submission is successful, you should see “Submission successful!” appear as output.
3. 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.