Assignment 6

Your name and student ID

Today's date

- Due date: Tuesday, March 10 10:00pm.
- Late penalty: 50% late penalty if submitted within 24 hours of due date, no marks for assignments submitted thereafter.
- This assignment is marked out of 29. Marks are indicated for each question.
- Remember: autograder is meant as sanity check ONLY. It will not tell you if you have the correct answer. It will tell you if you are in the ball park of the answer so *CHECK YOUR WORK*
- 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 right 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. When it doesn't, 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.

Oklahoma is not historically known for experiencing earthquakes. Up until 2008, Oklahoma experienced a constant rate of about 1.5 perceptible earthquakes per year on average.

1. [1 point] Assuming that earthquakes are random and independent, with a constant rate of 1.5 per year, the count of perceptible earthquakes per year in Oklahoma should have a Poisson distribution with mean 1.5. What is the standard deviation of the number of earthquakes per year? Round to the nearest 3 decimal places.

```
sd_earthquake <- "YOUR ANSWER HERE"
sd_earthquake
```

```
## [1] "YOUR ANSWER HERE"
```

```
check_problem1()
```

```
## [1] "Checkpoint 1 Error: sd_earthquake should be numeric"
## [1] "Checkpoint 2 Error: Round your answer to 3 decimals"
## [1] "Checkpoint 3 Error: In Wrong Range"
##
## Problem 1
## Checkpoints Passed: 0
## Checkpoints Errored: 3
## 0% passed
## --------
## Test: FAILED
```

2. [1 point] Making the same assumptions as in part (a), use one or two R functions to compute the probability of seeing less than two earthquakes per year. Round your answer to three decimal places.

```
probability <- "YOUR ANSWER HERE"
probability
```

```
## [1] "YOUR ANSWER HERE"
```

```
check_problem2()
```

```
## [1] "Checkpoint 1 Error: probability should be numeric"
## [1] "Checkpoint 2 Error: Round your answer to 3 decimals"
## [1] "Checkpoint 3 Error: In Wrong Range"
##
## Problem 2
## Checkpoints Passed: 0
## Checkpoints Errored: 3
## 0% passed
## --------
## Test: FAILED
```

3. [2 points] Do the same calculation as above, this time using only a hand calculator. Show your work and round your final percentage to two decimal places.

<TODO: YOUR ANSWER HERE>

4. [1 point] In 2013, Oklahoma experienced 109 perceptible earthquakes (an average of two per week). Assuming the same model as above, write an equation to show how the chance of experience 109 earthquakes or more can be written as a function of the probability at or below some k.

(Note: You can write these equations using pen and paper and upload the image if you'd like. You can also write the equations using plain text (i.e., P(X >= k)). If you would like to use math equations that render when you knit the pdf (i.e., $P(X >= k)$) you need to be **very careful** with your symbols. For example, to get the symbol for "greater than or equal to" you cannot copy and paste it into R from the slides or another document. This will cause errors! Instead you need to write $P(X \geq k)$. Again, use any of these three methods (hand-written, plain text in R, or "math equations between dollar signs", and you will get points so long as it is human-readable.)

<Note: If you are uploading an image (this is optional), use the following code, or delete if not using.BE SURE TO REMOVE THE OPTION "eval = F" if using this code OR IT WON'T RUN when you knit the file!:>

<TODO: YOUR ANSWER HERE>

5. [1 point] Using R, calculate the probability of observing 109 perceptible earthquakes or more. Round your answer to the nearest whole number.

```
probability_109_or_more <- "YOUR ANSWER HERE"
probability_109_or_more
```

```
## [1] "YOUR ANSWER HERE"
```

```
check_problem5()
```

```
## [1] "Checkpoint 1 Error: Answer should be numeric"
## [1] "Checkpoint 2 Error: In Wrong Range"
##
## Problem 5
## Checkpoints Passed: 0
## Checkpoints Errored: 2
## 0% passed
## --------
## Test: FAILED
```

6. [1 point] Based on your answer to Problem 5, write a sentence describing the chance of seeing such an event assuming the specified Poisson distribution (i.e., is it rare or common?)

<TODO: YOUR ANSWER HERE>

7. [2 points] Based on your answer in question (e), would you conclude that the mean number of perceptible earthquakes has increased? Why or why not? Would knowing that the number of perceptible earthquakes was 585 in 2014 support your conclusion?

<TODO: YOUR ANSWER HERE>

To track epidemics, the Center for Disease Control and Prevention requires physicians to report all cases of important transmissible diseases. In 2014, a total of 350,062 cases of gonorrhea were officially reported, 53% of which were individuals in their 20s. Assume this 53% stays the same every year. Researchers plan to take a simple random sample of 400 diagnosed cases of gonorrhea to study the risk factors associated with the disease. Call $\hat{p}$ the proportion of cases in the sample corresponding to individuals in their 20s.

8 [1 point] What is the mean of the sampling distribution of $\hat{p}$ in random samples of size 400?

```
sampling_dist_mean <- "YOUR ANSWER HERE"
sampling_dist_mean
```

```
## [1] "YOUR ANSWER HERE"
```

```
check_problem8()
```

```
## [1] "Checkpoint 1 Error: Answer should be numeric"
## [1] "Checkpoint 2 Error: In Wrong Range"
##
## Problem 8
## Checkpoints Passed: 0
## Checkpoints Errored: 2
## 0% passed
## --------
## Test: FAILED
```

9. [1 point] What is the standard deviation of the sampling distribution of $\hat{p}$ in random samples of size 400? Round your answer to 3 decimal places.

```
sampling_dist_sd <- "YOUR ANSWER HERE"
sampling_dist_sd
```

```
## [1] "YOUR ANSWER HERE"
```

```
check_problem9()
```

```
## [1] "Checkpoint 1 Error: Answer should be numeric"
## [1] "Checkpoint 2 Error: Round your answer to 3 decimals"
## [1] "Checkpoint 3 Error: In Wrong Range"
##
## Problem 9
## Checkpoints Passed: 0
## Checkpoints Errored: 3
## 0% passed
## --------
## Test: FAILED
```

10. [3 points] Describe the conditions required for the sampling distribution of $\hat{p}$ to be Normally distributed. Use the numbers provided in the question to check if the conditions are likely met.

<TODO: YOUR ANSWER HERE>

Read this short article in the New York Times Upshot from 2016. (All Berkeley students should have access to a free NY Times subscription.)

11. [2 points] Explain sampling variation, in the context of this article. Does the 3 percentage point margin of error account for sampling variation?

<TODO: YOUR ANSWER HERE>

12. [1 point] The authors provides several reasons why the true margin of error is larger than three percent. Describe one of the primary reasons provided in 1-2 sentences.

<TODO: YOUR ANSWER HERE>

13. [1 point] Based on the information in article, if we're doing a study in public health, choose the answer that is most correct:

    (a) The confidence interval accounts for random error only. If a study suffers from other sources of bias (i.e., confounding, or mismeasurement) the CI will not account for this limitation.
    (b) Increasing the sample size will reduce the chance of other sources of bias (i.e., confounding, or mismeasurement), which is why a larger sample is better.
    (c) both (a) and (b)
    d) neither (a) or (b)

Assign your letter choice as a string. Example: nytimes_answer <- "c"

```
nytimes_answer <- "REPLACE WITH a, b, c, or d. Keep the quotes"
nytimes_answer
```

```
## [1] "REPLACE WITH a, b, c, or d. Keep the quotes"
```

```
check_problem13()
```

```
## [1] "Checkpoint 1 Error: Wrong Format"
##
## Problem 13
## Checkpoints Passed: 0
## Checkpoints Errored: 1
## 0% passed
## --------
## Test: FAILED
```

Deer mice are small rodents native in North America. Their adult body lengths (excluding tail) are known to vary approximately Normally, with mean $\mu = 86$ mm and standard deviation $\sigma = 8$ mm. It is suspected that depending on their environment, deer mice may adapt and deviate from these usual lengths. A random sample of $n = 14$ deer mice in a rich forest habitat gives an average body length of $\bar{x} = 91.1$ mm. Assume that the standard deviation σ of all deer mice in this area is 8 mm.

14. [1 point] Calculate a 99% confidence interval based on this information (you can use R as a calculator to perform the calculation, or use a hand calculator). Round your final values to three decimal places.

```
lower_tail <- "REPLACE WITH YOUR ANSWER FOR THE LOWER BOUND"
upper_tail <- "REPLACE WITH YOUR ANSWER FOR THE UPPER BOUND"
ci_99 <- c(lower_tail, upper_tail)

check_problem14()
```

```
## [1] "Checkpoint 1 Error: Both answer should be numeric"
## [1] "Checkpoint 2 Error: Round your answer to 3 decimals"
##
## Problem 14
## Checkpoints Passed: 0
## Checkpoints Errored: 2
## 0% passed
## --------
## Test: FAILED
```

15. [1 point] Interpret the confidence interval from Problem 14.

<TODO: YOUR ANSWER HERE>

16. [2 points] Suppose deer mice researchers thought your CI was too wide to be useful. Given that you cannot change the standard deviation, what two things could you do to provide a narrower confidence interval?

<TODO: YOUR ANSWER HERE>

17. [1 point] You decide to create a 95% confidence interval, rather than a 99% confidence interval. Perform this calculation below and round your answer to 3 decimal places.

```
lower_tail95 <- "REPLACE WITH YOUR ANSWER FOR THE LOWER BOUND"
upper_tail95 <- "REPLACE WITH YOUR ANSWER FOR THE UPPER BOUND"
ci_95 <- c(lower_tail95, upper_tail95)

check_problem17()
```

```
## [1] "Checkpoint 1 Error: Both answer should be numeric"
## [1] "Checkpoint 2 Error: Round your answer to 3 decimals"
##
## Problem 17
## Checkpoints Passed: 0
## Checkpoints Errored: 2
## 0% passed
## --------
## Test: FAILED
```

18. [2 points] Based on this 95% CI, is there evidence against the hypothesis H_0 that these mice have a significantly different mean length compared to the population described in the first part of the question? Without performing a calculation, what amounts do you know the p-value to be bounded between for a two-sided hypothesis test of H_0?

Hint: Use information from question 17 and from question 14.

<TODO: YOUR ANSWER HERE>

We want to perform a z-test with the two-sided alternative hypothesis the true mean length is not equal to 86mm. In the next four problems, we will conduct a z-test step by step.

19. [1 point] Write out the null and alternative hypotheses for the above problem using notation.

<TODO: YOUR ANSWER HERE>

20. [1 point] Calculate the z test statistic. Round your answer to 3 decimal places.

```
z_stat <- "REPLACE WITH YOUR ANSWER FOR THE LOWER BOUND"
z_stat
```

```
## [1] "REPLACE WITH YOUR ANSWER FOR THE LOWER BOUND"
```

```
check_problem20()
```

```
## [1] "Checkpoint 1 Error: Answer should be numeric"
## [1] "Checkpoint 2 Error: Round your answer to 3 decimals"
## [1] "Checkpoint 3 Error: In Wrong Range"
##
## Problem 20
## Checkpoints Passed: 0
## Checkpoints Errored: 3
## 0% passed
## --------
## Test: FAILED
```

21. [1 point] Calculate the p-value as a decimal. Round your answer to 3 decimal places.

```
p_val <- "REPLACE WITH YOUR ANSWER FOR THE LOWER BOUND"
p_val
```

```
## [1] "REPLACE WITH YOUR ANSWER FOR THE LOWER BOUND"
```

```
check_problem21()
```

```
## [1] "Checkpoint 1 Error: Answer should be numeric"
## [1] "Checkpoint 2 Error: Round your answer to 3 decimals"
## [1] "Checkpoint 3 Error: In Wrong Range"
##
## Problem 21
## Checkpoints Passed: 0
## Checkpoints Errored: 3
## 0% passed
## --------
## Test: FAILED
```

22. [1 point] Interpret your above p-value.

<TODO: YOUR ANSWER HERE>

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  NOT YET GRADED               2 free-response
## Problem 4  NOT YET GRADED               1 free-response
## Problem 5          FAILED               1    autograded
## Problem 6  NOT YET GRADED               1 free-response
## Problem 7  NOT YET GRADED               2 free-response
## Problem 8          FAILED               1    autograded
## Problem 9          FAILED               1    autograded
## Problem 10 NOT YET GRADED               3 free-response
## Problem 11 NOT YET GRADED               2 free-response
## Problem 12 NOT YET GRADED               1 free-response
## Problem 13         FAILED               1    autograded
## Problem 14         FAILED               1    autograded
## Problem 15 NOT YET GRADED               1 free-response
## Problem 16 NOT YET GRADED               2 free-response
## Problem 17         FAILED               1    autograded
## Problem 18 NOT YET GRADED               2 free-response
## Problem 19 NOT YET GRADED               1 free-response
## Problem 20         FAILED               1    autograded
## Problem 21         FAILED               1    autograded
## Problem 22 NOT YET GRADED               1 free-response
```

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 is saved (the file name in the tab should be **black**, not red with an asterisk).
4. 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-sp20/hw/hw06; 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.