Exam #2 STA-209 Sections 03, 05

Name:_____

Directions

- Several questions have a *suggested* number of sentences for your answer. This is to help indicate the scope of solution I am looking for (i.e., you do not always need every single detail) and to discourage you from "information dumping"
- Information that is included with your answer that is not relevant to the problem will not help you but *will still be graded for correctness*. In other words, including more information than is asked for can generally only hurt you
- You **do not** need to write in complete sentences: bullet points are completely acceptable and even preferred

Formulas

$$\overline{X} \sim N(\mu, \sigma/\sqrt{n})$$

$$\hat{p} \sim N\left(p,\ \sqrt{\frac{p(1-p)}{n}}\right)$$

Question 1

Part A: Explain the relationship between a *population* and a *sample* and how each of them are used within the statistical framework.

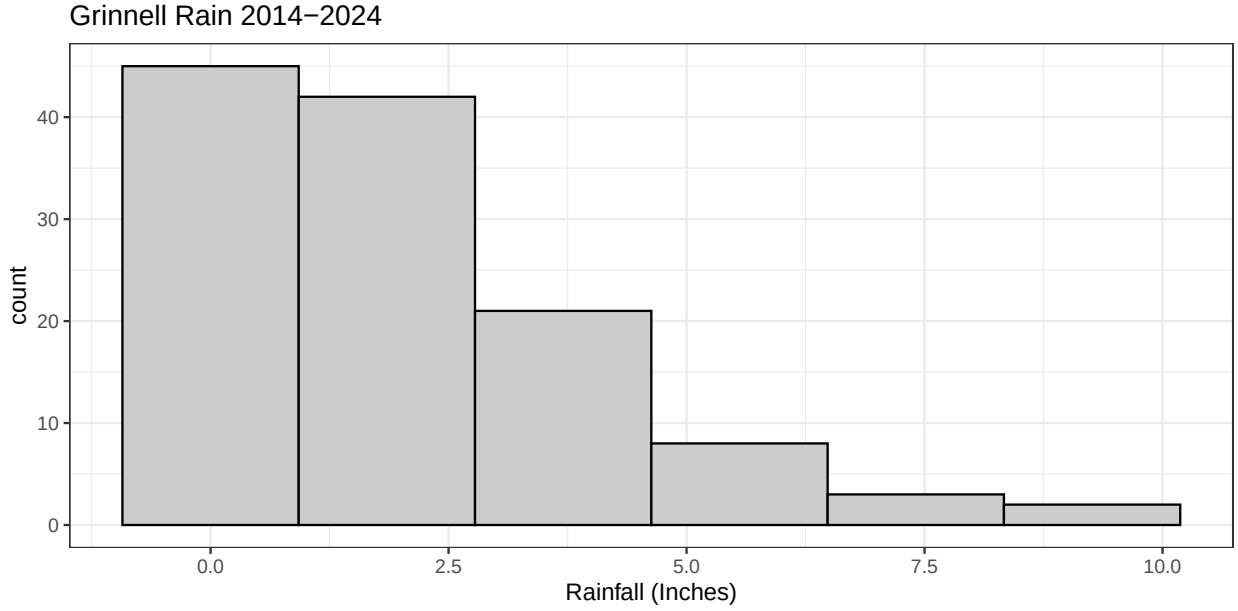
Part B: Write out the formula for a t-statistic. Explain how each term in the formula changes the value in the t-statistic (you may consider the difference $\overline{x} - \mu$ to be a single term).

Part C: Write out the formula for finding a confidence interval using the point estimate/margin of error method. Assuming that our test statistic follows a *t-distribution*, explain the relationship between the calibration value, C, and the degrees of freedom of the t-distribution i.e., for constructing a 95% confidence interval, how does changing the degrees of freedom change the value of C?

Part D: Suppose that researchers are investigating the difference between the lengths of male and female turtles. Assume that their sample was representative and the study was done correctly. Testing the null hypothesis that the difference between the two is the same at $\alpha = 0.05$, the investigators find a p-value for the difference of $p = 0.0001$. Would this be considered *statistically significant*? Does it also imply that the difference is a meaningful one? Explain in 1-2 sentences what would need to be true for the difference to be *statistically significant* without being *practically significant*.

Question 2

The plot below again shows the monthly distribution of rainfall in Grinnell over a 10 year period from 2014-2024.



From this collection of 121 months, we collect a sample of size $n = 40$, with the following summary statistics:

N	Mean	Median	SD
40	1.419	1.305	1.143

Part A We are interested in using our sample to estimate the population *median*. Explain what steps you would take to find a 95% confidence interval for the true value of the *median*.

Part B From your work in Part A, you find that your bootstrapped median has a 95% confidence interval of (0.745, 1.935). In 1-2 sentences, explain what this means.

Part C Your friend observes that you have computed confidence intervals the hard way. Because the median and the mean are so similar in the summary statistics, a better way to find a 95% confidence interval would be to use the central limit theorem. Because $N > 30$, we can use a normal approximation and find the confidence interval using:

$$\text{Median} \pm 2 \times \frac{\hat{\sigma}}{\sqrt{n}} = (0.943, 1.666)$$

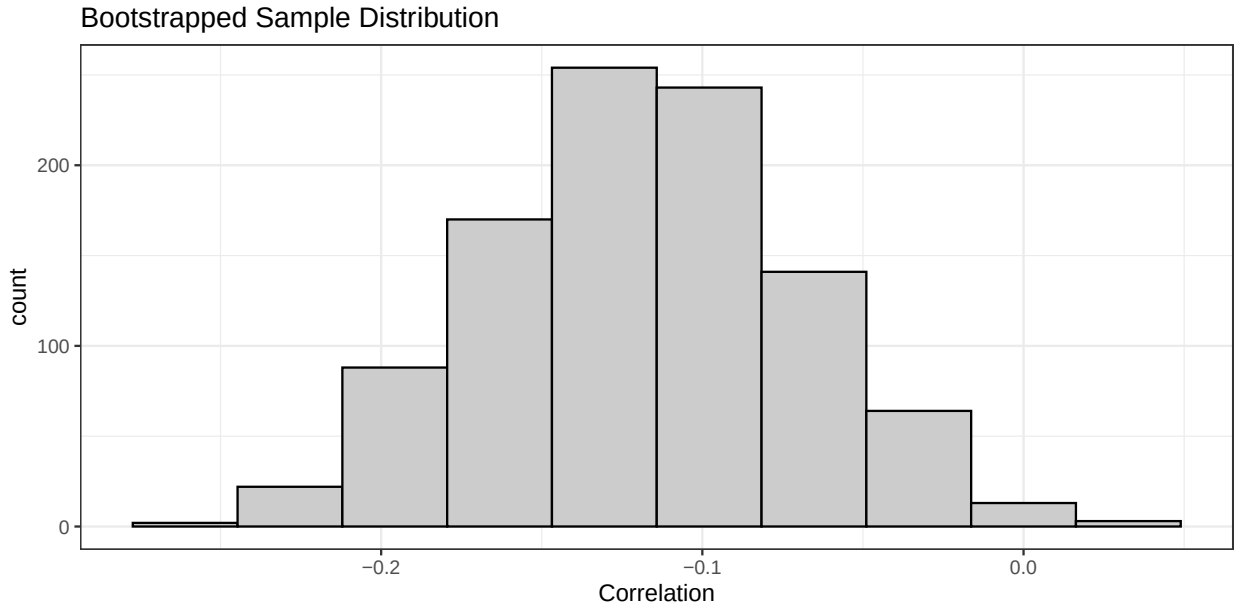
Additionally, because these confidence interval is narrower than the one you derived, it is going to be more accurate. Do you agree with your friend's assessment? Explain why or why not.

Question 3

Suppose we are interested in finding the average song length of all of Taylor Swift's music. In order to collect data, we listen to several radio stations every day for a week, recording the length of each of the Taylor Swift songs we hear. What is my population? Should we expect the sample we collect to be representative? Why or why not?

Question 4

Suppose now that we are interested in determining if there is any association between the length of Taylor Swift's songs and how popular it is. Collecting data now from Spotify instead, we randomly sample 100 songs from her entire catalog, treating this as our sample. The Spotify data includes two variables: *popularity*, a numerical score from 0 (least popular) to 100 (most popular), and *duration*, measured in milliseconds. The correlation between these variables was bootstrapped 1,000 times, resulting in the following sample distribution:



Additionally, the following percentiles of the distribution were found:

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
-0.18166	-0.16068	-0.14368	-0.13124	-0.11836	-0.10513	-0.09379	-0.07724	-0.05503	0.03346

Part A: Explain how we might use the information above to draw conclusions about the association between popularity and song length.

Part B: Using the information above, calculate a 80% confidence interval for the correlation between song length and popularity. Based on this, would you say there is an association between song length and popularity? If so, how would you describe this relationship?

Part C: Suppose in testing the hypothesis that the correlation is not zero (i.e., there *is* an association between the two), the p-value based on our observed data is $p = 0.008$. If before we conducted our test, we specified that we would accept a Type I error rate of 1%, would we reject our null hypothesis based on this data?

Question 6

This question also involves Taylor Swift Spotify data.

"Danceability" is a quantitative variable defined as how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm, stability, beat strength, and overall regularity. For reference (this information i *not* needed to answer this question), the following Taylor Swift songs have the associated danceability score:

- "Blank Space" – 0.753
- "Picture to Burn" – 0.658
- "Fearless" – 0.598
- "Tear Drop on My Guitar (Live)" – 0.344

Based on this, you decide to investigate what proportion of her songs have a danceability score greater than 0.6. Before we begin our study, we hypothesize that the true proportion is 75%.

Part A: To begin, you randomly sample $n = 20$ songs and find that 25 of them have a danceability score greater than 60%. Using this information, create a t-statistic for the observed data.

Part B: Assuming a normal approximation, calculate a 95% confidence interval for the true proportion of songs with a danceability score greater than 0.6. Based on this, what decision would you make regarding your null hypothesis?

Part C: Suddenly, you remember that you don't need to use a sample of Taylor Swift songs to find the true proportion, as we have data on every Taylor Swift song. Evaluating the full dataset, we find the true value of this parameter to be $p = 0.48$. Using this new information, answer the following questions:

- Thinking about the confidence interval you created in Part B, did it contain the true value of the population parameter?
- Do you think the conclusion that you came to in Part B was the correct one?
- What about your experiment could you have changed to have reached a different conclusion?

```
##
##  Exact binomial test
##
## data:  sum(taylor$danceability > 0.6) and 530
## number of successes = 257, number of trials = 530, p-value
## <0.0000000000000002
## alternative hypothesis: true probability of success is not equal to 0.75
## 95 percent confidence interval:
##  0.44160 0.52838
## sample estimates:
## probability of success
##                0.48491
```

```
## [1] 0.40849 0.84151
```

Part D: In the context of Part B and Part C, explain the difference between "Failing to Reject H_0" and "Accepting H_0". Why is this distinction important in this case? Limit your answer to 3-5 sentences.

Part E: Create an illustration the demonstrates what occurred in this problem. That is, draw what the sampling distribution would look like under the null hypothesis (centered at $p_0 = 0$) and what the true sampling distribution would look like knowing that the true population parameter was $p = 0.48$. Your drawing does not need to be exact, but it should demonstrate the following:

- Where did $\hat{p}$ lay relative to p_0 and p?
- How much did the null and true distribution overlap? I.e., what role did variability play in this experiment