IST 772 Midterm Exam

Name: Zane Alderfer

Instructions: Compose brief answers to each of the following questions, typing your response in *italics* below each question.

1. Describe the conceptual connection between $\boldsymbol{\mu}$ ("mu," the population mean) , {$\bar{?}$}("x-bar," a sample mean) and the sampling distribution. How are they connected to each other?

The mean of the entire population is the population mean and the sample mean is used as an estimate for the population mean when you use a sample of the population as the sampling distribution. As you use a larger sample size, the distribution will look more and more like a bell-shaped curve and the sample mean will get closer and closer to the population mean.

2. Your boss at the New York Times asks you to conduct an A/B test on two different headlines about the same story. Each headline is displayed on n=140 high traffic social media pages:

Headline 1 gets an average of 2400 clicks per hour.
Headline 2 gets an average of 2200 clicks per hour.

The 95% confidence interval is as follows:
100 < (mean difference between Headline 1 and 2) < 300.

Answer the following questions about that confidence interval:

a. On the basis of this confidence interval, conduct a hypothesis test at the 0.05 level under the alternative hypothesis that average clicks per hour are not equal. State the null and alternative hypotheses.
The Null hypothesis would be: The mean difference between the headlines equal to 0 and the alternative Null hypothesis would be: The mean difference between the headlines is not equal to 0.

b. Would you reject or not reject the null hypothesis. Why or why not?
I would not reject the Null hypothesis because the confidence interval contains 0 meaning the expected difference between the actual difference of the population means and theorized difference of the population means is anywhere between –100 and 100 which would suggest the difference could be 0.

c. Based on your answer in b. What is your conclusion about the difference between the headlines? Is headline 1 or 2 better and why?
Although I could assume headline 1 is a better option because it has a higher population mean of clicks, the actual difference could be 0 according to the confidence interval so NO conclusion toward which is better is made.

d. Your friend calculates the p-value for the hypothesis mentioned in questions a-c and finds that p = 0.25. Does this sound plausible to you? Why or why not?
The p-value is a calculation of the percent of how possible the NULL hypothesis which seems to be 25% here. This would confirm that we FAIL to reject the NULL hypothesis and also supports my answers from a-c that it is possible that the difference of these population means can be 0 so yes this seems plausible.

e. Your boss tells you to run the same experiment 999 more times, calculating a new confidence interval each time. Now you have a collection of 1000 confidence intervals, each of which was constructed in the same way, but from new data samples: What can you say about this collection of confidence intervals?
I could use these samples to calculate a true population mean difference between the two populations and this can help to either accept or reject our initial Null hypothesis with the initial sample populations of data.

f. Which command in R would you use to produce the confidence interval for each of the 1000 that you constructed?
I would store a variable called something like clicks_confidence and use the "replicate" function which would look something like:
click_confidence <- replicate(1000, mean(sample(clicks, size = 140, replace = TRUE)), simplify = TRUE)
The variable "clicks" was created just as a placeholder for the initial confidence interval population but then would use the quantile function to get the 95% confidence internal like:
quantile(click_confidence, .025) and quantile(click_confidence, .975) and this should give me the two values I'm looking for

3. Tests for detecting diseases such as HIV are not 100% accurate and one can use Bayes' theorem to assess the probability that someone is actually infected HIV given a positive test. Please use the following facts to calculate the probability that someone has HIV after having received a positive test:

- For someone with HIV, the probability of a positive test is 99%
- The probability that someone has HIV is 3%.
- The probability of getting a positive test is 4%.

Bayes' theorem can help us find out the probability of having HIV after receiving a positive test (P(Positive|HIV)) while using the variables probability of having HIV (P(HIV)), probability of having positive test (P(Positive)) and probability of someone with HIV producing a positive test (P(HIV|Positive)) with the equation looking something like:
*P(Positive|HIV) = (P(HIV) * P(HIV|Positive))/P(Positive)*
*= (.03 * .99)/.04*
= 74.25%
This makes sense as 1% of the population test positive when they're negative for HIV and 3% are accurate positive and 3% is 75% of 4% so this seems accurate.

4. : The Null Hypothesis Significance test (NHST) is the classic inferential test used throughout the 20th century. The NHST comprises a set of logical steps that lead to a consideration of the viability of a stated null hypothesis. Following the material presented on page 77 of *Reasoning with Data*, here is an unordered list of the steps:

Calculate the test statistic
Assert a null hypothesis
Collect data
Find the p-value associate with the test statistic
Choose an alpha level
Reject the null hypothesis
Fail to reject the null hypothesis
Evaluate the p-value with respect to alpha

Place these steps in the correct order and add a brief one or two sentence explanation that describes the purpose and importance of each step.

- *Assert a Null Hypothesis*
    - *This allows us to create an assumption or statement that there isn't significance between the two groups in question. The following tests or steps use this as a foundation to decide whether the variables involved are worth exploring*
- *Choose an alpha level*
    - *This allows to decide with later tests whether the Null hypothesis can be rejected or not. A normal alpha level is .05 meaning a p-value below the alpha level can lead to you rejecting the Null hypothesis and a p-value above the alpha level fails to reject the Null hypothesis*
- *Collect Data*

  - *The collection of data is obviously vital. If there's no data to develop statisitics for then it wouldn't be possible to reject or fail to reject the assigned Null hypothesis*
- *Calculate the test statisitic*
  - *This allows us to see the difference between the actual data and what we would expect according to the Null hypothesis. This can help us develop other values for proper evalution*
- *Find the p-value associate with the test statistic*
  - *The p-value allows us to see the probability that the mean difference of the two populations is contained within the 95% confidence interval.*
- *Evalute the p-value with respect to alpha*
  - *This will help us determine whether we reject the Null hypothesis or fail to reject the Null hypothesis. If the p-value is below the alpha level then we reject the Null hypothesis and look at the alternative Null hypothesis as more plausible and if the p-value is greater than we fail to reject the Null hypothesis.*
- *Reject the Null hypothesis*
  - *This is the assumed outcome but when we reject the Null hypothesis, we look to alternative Null hypotheses as a better explanation for the two sets of data*
- *Fail to reject the Null hypothesis*
  - *If this happens, then there appears to be significance between the two populations and should be explored further.*