IST 772 Midterm Exam

Name: Sharat Sripada (vssripad)

Instructions: Compose brief answers to each of the following questions, typing your response in *italics* below each question. Please also include a .R file with IF RELEVANT.

1. Describe the conceptual connection between 𝞵 (“mu,” the population mean), (“x-bar,” a sample mean) and the sampling distribution. How are they connected to each other?

***Answer:*** *The population mean (mu) is defined as: sigma(x) / N*

*where x - All data-points in the population*

*N - Number of data-points in the population*

*However, it is not feasible to obtain all data-points/observations of a population as it is large and can be time-consuming and expensive. Therefore, we work with samples.*

*Sample mean (x-bar) is calculated from samples of data-points/observations taken randomly from a population. Since it is obtained from the population,* ***sample mean is a good estimate of the population mean*** *and can be thought of as directly connected or related to it.*

*As the sample size(n) increases sample mean (x-bar) converges to the population mean (mu).*

*During the process of sampling, we choose a sample size (n) and repeatedly draw samples from the population. For every sample we draw, we measure an aspect of the data (like mean) and record the measurement to create what is known as a* ***sampling distribution****.*

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.

***Answer:*** *We can formulate the NULL and alternate hypothesis as follows:*

* *NULL hypothesis: Average clicks per hour for two headlines are equal*
* *Alternate hypothesis: Average clicks per hour for two headlines are not equal*

*When we run the hypotheses test, we can say with 95% confidence that the difference in average clicks per hour of ZERO between the two headlines is outside the confidence interval (100, 300).*

b. Would you reject or not reject the null hypothesis. Why or why not?

***Answer:*** *After running the t-test, we look to extract the t-value, p-value, confidence intervals etc. from it.*

*p-value helps with determining if we REJECT or FAIL TO REJECT the NULL hypothesis:*

* *if p-value is less than alpha/significance value we will REJECT the null hypothesis*
* *if p-value is greater than alpha/significance value we FAIL TO REJECT the null hypothesis*

*Since we are working with confidence-intervals in this problem and have previously stated that difference in average clicks of ZERO is outside the confidence interval, we can* ***REJECT the NULL Hypothesis****.*

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?

***Answer:*** *We* ***prefer A*** *since we are 95% confident that the true population mean difference would lie in the confidence interval (100, 300).*

*Further, since the interval (100, 300) does not comprise zero, we can say that there would not be a possibility that the average clicks on the two headlines be equal.*

*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?

***Answer:***  *Consider the following two statements:*

* *We are 95% confident that the true population mean difference lies in the confidence interval (100, 300)*
* *More importantly, the possibility for the average means to be equal OR the difference in average means to be equal to ZERO is low*

*Given this, we concluded to REJECT the NULL hypothesis. That means also, that p-value ought to be less than alpha (0.05). Since my friend derived a value of 0.25 > alpha (0.05) this* ***does not sound*** *like a plausible value.*

*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?

***Answer:*** *We can say that 95% of all those tests are likely to contain the true population mean difference. That is, in 950 out of 1000 tests will contain the true population mean difference.*

*f.* Which command in R would you use to produce the confidence interval for each of the 1000 that you constructed?

***Answer:*** *t.test()*

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%.

***Answer:*** *The Bayes’ theorem or formula for this is:*

*P(HIV | +ve test) = P(HIV) \* P(+ve test | HIV) / P(+ve test)*

*From the data we have:*

* *P(+ve test | HIV) = 0.99*
* *P(HIV) = 0.03*
* *P(+ve test) = 0.04*

*Substituting this in the equation above we get:*

*P(HIV | +ve test) = 0.03 \* 0.99 / 0.04 = 0.7425*

*The probability of someone having HIV after receiving a +ve test is 0.7425*

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.

***Answer:*** *Following is the order of the steps for NHST:*

*Assert a null hypothesis*

*Choose an alpha level*

*Collect data*

*Calculate the test statistic*

*Find the p-value associate with the test statistic*

*Evaluate the p-value with respect to alpha*

*Reject the null hypothesis*

*OR*

*Fail to reject the null hypothesis*