CS3130/ECE3530: Probability and Statistics for Engineers **Due Tu 4/9**

Name:

uID:

Homework 6: Estimators and Confidence Intervals

Instructions: Write your answers directly on this pdf (via an editor, iPad, or pen/pencil). The answers should be in the specified place. Students will be responsible for loading their assignments to GradeScope, and identifying what page contains each answer.

The assignment should be uploaded by 11:50pm on the date it is due. There is some slack built into this deadline on GradeScope. Assignments will be marked late if GradeScope marks them late.

If the answers are too hard to read you will lose points (entire questions may be given 0).

Please make sure your name appears at the top of the page.

You may discuss the concepts with your classmates, but write up the answers entirely on your own. **Be sure to show all the work involved in deriving your answers! If you just give a final answer without explanation, you may not receive credit for that question.**

1. Say a Utah pollster conducted m=15 polls among people who would vote in the 2020 presidential elections, and reports that 60% of the respondents would vote for Chris Stewart. But the pollster did *not* report how many people n were interviewed in each poll, although each poll they conduct always includes the same number of people n. However, the pollster did report how many people voted for Spencer Cox in the 2020 gubernatorial election. Also, suppose we know that exactly 3/5 of all voters voted for Cox in 2020.

   (a) Each poll conducted can be represented as a random variable X_i for i = 1, 2, ..., 15, representing the number of people who said they would vote for Cox. What is the distribution of each X_i? (Remember to include the parameters of the distribution in your description.)

   (b) Given the random sample $X_1, ..., X_{15}$, suppose we want to estimate the number of people n included in each poll. Give an unbiased statistic that you would use to estimate the parameter n. Please include a verification that the bias is zero. (Hint: Start with a mean statistic).

(c) What is the variance of this statistic?

2. In this problem you are going to analyze the built-in R data set `iris`. First, extract the petal width of the Virginica species and save it to a vector `x` using this command: `x = iris$Petal.Width[iris$Species == "virginica"]`

   Answer the following: **Write down the R code you use to compute parts (a), (b), (c), (d) - it should be very short.**

   (a) Using a Normal approximation, what is the 95% confidence interval for the mean of `x`? (Assume you know the true standard deviation of x is $\sigma = 0.2747$).

   (b) Using a Student t distribution, what is the 95% confidence interval for the mean of `x`?

(c) Now assume that you only have the first 10 measurements. That is, create the vector `y = x[1:10]`. Repeat parts (a) and (b) for the vector `y`.

(d) How did the decrease in sample size affect the results? Were the two different confidence intervals affected differently?

3. For the following scenarios, state the null hypothesis (H_0) and the Alternative hypothesis (H_1) in terms of the appropriate parameter. Remember to provide the distribution for the null hypothesis.

   (a) Kroger regularly receives large shipments of apples. For each shipment, a supervisor takes a random sample of apples to see if they are bruised and performs a significance test. If the sample shows significant evidence that more than 5% of the apples are bruised, they will request a new shipment of apples. (Let p represent the proportion of the apples in the shipment that are bruised.)

   (b) My spouse recently bought an automatic cat food dispenser. It is supposed to dispense 20 mL of cat food per portion, but my he thinks that it is not dispensing the right amount. We took 50 samples of cat food, weighing each, and we would like to test whether it is dispensing the right amount of cat food on average. (Let μ represent the amount of cat food dispensed on average.)

   (c) Lionel Messi has scored on 77% of his penalties when taking penalties for either club or country. He has decided to alter his penalty run up and he would like to test whether his new run up produces worse results. (Let p represent the proportion of penalties made.)

4. Suppose we have a random sample $X_1, ..., X_n$ of n random variables from an unknown distribution D with mean μ_D. We want to design an experiment that will evaluate if we should deviate from the null hypothesis with a critical value at $\alpha = 0.03$.

$$H_0 : \mu = \mu_D, \qquad H_1 : \mu > \mu_D.$$

   (a) We want to define a threshold $t_{0.03}$ (for a critical value at 0.03) for a statistic T so that $\Pr(T \leq t_{0.03}) < 0.97$. What is the statistic T and what is $t_{0.03}$? Be sure to write the mathematical expression for T. You can use R syntax to express $t_{0.03}$. (**Hint:** First ask yourself "which distribution/statistic should we use?" Then start with the $(1 - \alpha)$ quantile of this statistic.)

   (b) Now, suppose we know that the random sample $X_1, ..., X_n$ is from a Bernoulli distribution with known parameter p. Now, what is the statistic T and what is $t_{0.03}$? (**Hint:** You can assume n is sufficiently large so the central limit theorem applies.)

5. Here we are going to test a couple of hypotheses about the Old Faithful data in R. Remember, this is the `faithful` data frame that is built into R. First, split `faithful` into two separate data frames: (1) those entries with eruption times less than 3 minutes (`faithful$eruptions < 3`) and (2) those entries with eruption times greater than or equal to 3 minutes (`faithful$eruptions >= 3`). Answer the following about the entry **wait time** (`faithful$waiting`):
(**Hint:** You might try using the R function `t.test` to double-check the answers you get, but you may not use it as the R commands that are asked for below.)

   (a) For the entries with short eruption times, you want to test the hypothesis that the associated waiting time is on average less than 60 minutes. What is the null hypothesis? What is the alternative hypothesis?

   (b) Give R commands to compute the t statistic and the resulting p-value (one line of code for t and one line for p). What values did you get? Would you reject the null hypothesis at the $\alpha = 0.05$ level?

(c) For the entries with long eruption times, you want to test the hypothesis that the associated waiting time last on average longer than 80 minutes. What is the null hypothesis? What is the alternative hypothesis?

(d) Give R commands to compute the t statistic and the resulting p-value to test the hypothesis you came up with the part (c) (again, one line of code for t and one line for p). What values did you get? Would you reject the null hypothesis at the $\alpha = 0.05$ level?