CS3130/ECE3530: Probability and Statistics for Engineers **Due Tu 4/23**

Name:

uID:

Homework 7: Hypothesis Testing, SImulation, Regression

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. In this problem you will do some hypothesis testing on a built-in R dataset. The dataset is *PlantGrowth*, and it contains samples of weights of harvests from fields that have undergone different treatments ("ctrl", "trt1", "trt2"). The question you will examine is whether each treatment improves plant yields compared to the control. You will assume that the variances of the different groups are the same (but unknown) and you want to set your critical value for type I error rate of 5%.

   (a) State the null and alternative hypotheses for evaluating treatment 1 and treatment 2 (separately) relative to controls.

   (b) Calculate the samples means and variances that are used in this hypothesis test.

(c) Use R to generate the critical value for this experiment with a 95% confidence. Show your code (paste).

(d) Write the equations to generate the test statistics for both tests. Use R to calculate those test statistics and p-values. Show your code.

(e) Explain, based on this analysis, if either treatment 1 and/or treatment 2 can be determined to improve yields.

2. In this problem you will implement and analyze a simulation such as the one described in the class textbook in Section 6.4. The scenario is that users arrive at a well (for water) at random times with intervals between customers following an exponential distribution, with Exp(0.5/minutes). Users each carry a jug to be filled with random size following U(2,5). We are interested in how long people wait to get their jugs filled and the number of people waiting (or the service load) and we will examine the behavior under two different scenarios: when the well pumps 2 liters per minute and 3 liters per minute.

   (a) Write R code that generates a table of: customer number, time since last costumer arrived, absolute arrival times (in minutes, starting at zero), and jug sizes, for each customer. Construct the table with enough entries so that the last customer arrives just within 120 minutes (i.e. only simulate for two hours). Print the head and tail of this table – it should look like the first 4 columns of table 6.5 in the book (different numbers, because it's random).

   (b) Create additional columns in that table with the service and wait times for each customer under the two scenarios. Print the head of this table – it should look like table 6.5 in the book.

(c) Compute the average of the first n waiting times and graph this as a function of n for both pump scenarios.

(d) Create a histogram of waiting times for both scenario, excluding the first 15 minutes.

(e) Perform all of the steps above with a change in the distribution of jug sizes. Instead of U(2,5) use a normal distribution with $\mu = 3.5$ and $\sigma^2 = 0.75$. Do the analysis and show the graphs. Is the result different? Why?

3. This question deals with the built-in R dataset called *mtcars* in order to experiment with linear regression. We will look at gas efficiency as a function of both engine displacement and vehicle weight. Work will be done in R. Show plots and code for the following.

   (a) Do a linear regression of miles per gallon vs engine displacement. Plot the line on a scatter plot of these variables.

   (b) Do a linear regression of miles per gallon vs vehicle weight. Plot the line on a scatter plot of these variables.

(c) Sometimes it makes sense to consider gallons-per-mile (GPM), instead of MPG. GPM is the reciprical of MPG. Do both regressions and plots above with GPM instead of MPG. Are displacement and vehicle weight better linear fits to MPG or GPM? Use correlation coefficients to support your answer.