STA 248 - Assignment #2

Due: April 3, 2020 @ 11:59 PM

Submit through Crowdmark (emailed link)

The goal of the assignments is to assess whether you can **apply the concepts discussed in lecture**. While there may be different paths to the solution, you must demonstrate in your solutions that you have, without a doubt, learned the course material. This is an **individual** assignment. Answer the problems to the best of your abilities. There should be no surprises on this assignment if you have adequately reviewed your notes and engaged regularly with the suggested textbook problems!

Problems that require written responses should be answered in the space provided on the assignment document.

Problems labeled with **[R script]** indicate problems that require the use of R script and you will be required to submit your (well organized/labeled) R-script. If you are using R-Studio, you can easily export your script as a pdf file.

Problem 1 [20 points]. A study of report writing by engineers is conducted. A scale that measures the intelligibility of engineers' English is devised. This scale, called an "index of confusion," is devised so that low scores indicate high readability. These data are obtained on articles randomly selected from engineering journals and from unpublished reports written in 1979.

Journals				Unpublished Reports			
1.79	1.75	1.67	1.65	2.39	2.51	2.86	2.14
1.87	1.74	1.94		2.56	2.29	2.49	
1.62	2.06	1.33		2.36	2.58	2.33	
1.96	1.69	1.70		2.62	2.41	1.94	

a) (7 points) Conduct any appropriate graphical checks on assumptions (and include it as part of your submission) before testing for equal variances in "index of confusion" between published and unpublished reports. Use $\alpha = 0.2$. Your solution should be complete with stating hypotheses, assumptions, sample statistics, test statistics, distribution, p-value, and conclusions.

b) (7 points) Does there appear to be a difference in intelligibility of engineers' English in published journals versus unpublished reports? Use $\alpha = 0.05$. Interpret your results. Your solution should be complete from beginning to end.

c) (2 points) If you were to compute a 95% confidence interval for the difference in average "index of confusion" scores between published and unpublished reports, would you expect a difference of 0 to be included in the interval? Explain.

d) (4 points) Construct a 95% confidence interval for the difference in means and interpret the values of your interval.

Problem 2 [26 points]. [R script] In a multi-centre research study, multiple bits of information were collected on lung cancer patients. The data is available on Quercus, in the file `hdp.csv`. In this question, please use a significance level of $\alpha = 0.05$ for all appropriate tests.

a) (11 points) The variable `CRP` represents the concentration of C-reactive protein in the patients' blood. The researchers would like to assume that the distribution of the C-reactive protein concentration follows a Weibull(2, 6) distribution. *Note: $Weibull(\lambda, k)$ distributions have probability density function:*

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, & x \geq 0 \\ 0, & otherwise \end{cases}$$

(i) (1 point) Specify the null hypothesis.

(ii) (4 points) [R script] Provide 10 equally likely intervals for CRP concentration under the null hypothesis. Show how you would find the endpoints of TWO of these intervals by hand, but if your calculus is a little rusty, you may use `qweibull( )` in R to compute your interval endpoints.

(iii) (2 points) [R script] Obtain the number of patients whose CRP concentration levels fall within each of the 10 intervals. Represent the data in a table.

(iv) (4 points) **By hand**, test whether or not the researchers' assumption is correct, and interpret your results.

b) (15 points) Now, the researchers would like to study the relationship between patients' Body Mass Index (BMI, units of kg/m^2) and their stage of cancer (I, II, III, IV). In the dataset, these are represented by the variables `BMI` and `CancerStage`, respectively.

(i) (2 points) **[R script]** Create four vectors containing the BMI's at each stage of cancer.

(ii) (3 points) **[R script]** BMI is usually categorized into four groups: underweight (below 18.5 kg/m^2), normal (18.5 to 25 kg/m^2), overweight (25 to 30 kg/m^2), and obese (over 30 kg/m^2). For each stage of cancer, determine the number of patients that full under each BMI category. Present your results as a table or data frame, which should have 4 rows and 4 columns. Include your table (hand-written or printed) below.

(iii) (5 points) Among patients with stage IV cancer, is there evidence to suggest an imbalance in the number of patients in each BMI category? **Please do this problem by hand.** Your answer should include the null and alternative hypotheses, calculation of an appropriate test statistic, a p-value, and an interpretation in the context of the problem.

(iv) (5 points) To answer the researchers' question, is there evidence of association between a patient's BMI category and their stage of cancer? **Please do this problem by hand.** Your answer should include the null and alternative hypotheses, calculation of an appropriate test statistic, a p-value, and an interpretation in the context of the problem.

Problem 3 [20 points]. Let x denote the number of lines of executable SAS code, and let Y denote the execution time in seconds. Use the following summary information to answer the following problems:

$$n = 10 \quad \sum_{i=1}^{10} x_i = 16.75 \quad \sum_{i=1}^{10} y_i = 170 \quad \sum_{i=1}^{10} x_i \cdot y_i = 285.625$$

$$\sum_{i=1}^{10} x_i^2 = 28.64 \quad \sum_{i=1}^{10} y_i^2 = 2898$$

a) (3 points) Estimate the regression line as a function to estimate conditional means of Y.

b) (3 points) Estimate $V(Y_i) = \sigma^2$, the standard deviation of B_1, and B_0.

c) (4 points) Test the hypothesis $\beta_1 = 0$ using $\alpha = 0.01$, and state your conclusion in the context of this problem.

d) (4 points) Test the hypothesis $\beta_0 = 25$ using $\alpha = 0.05$, and discuss the conclusion in the context of this problem. (e.g. What is the hypothesis testing? When is this a valid hypothesis to test, when is it not?)

e) (2 points) If significant regression is found, estimate the average time required to run a SAS program with 15 lines of executable code. If regression is not significant, what does this mean mathematically? Can you think of a practical reason from a computing standpoint that regression might not be significant in this case?

Problem 4 [15 points]. In a sleep-deprivation study, reaction times (in milliseconds) to a flashing dot on a computer screen were measured before sleep deprivation, and after 24 hours without any sleep. You may assume the reaction times before and after sleep deprivation are normally distributed. The data for 8 participants is shown below:

Before	21.2	23.3	22.7	21.8	24.6	26.9	25.5	23.9
After	22.4	27.1	22.1	21.9	30.1	28.7	25.0	23.7

a) (3 points) What is the most appropriate statistical test to use in this scenario? Justify your answer.

b) (1 point) Write out the test statistic you will use. Your answer should be consistent with your response in (a).

c) (5 points) Perform a hypothesis test at $\alpha = 0.05$ to determine whether there is a difference between the reaction times before and after sleep deprivation. Be sure to provide a complete solution, and your conclusion in the context of the problem.

d) (3 points) Construct a 99% confidence interval for the true mean difference in reaction times, and provide an interpretation for the confidence interval.

e) (3 points) Between the interpretations for your results in (c) and (d), which is more informative or gives a more complete picture of the effects of sleep deprivation on reaction times? Explain your answer.

Problem 5 [10 points]. In the video lecture, I derived for you the mean, variance, and the distribution for the slope estimator B_1 for the simple linear regression model. In this problem, you will do the same for B_0, the intercept estimator, which we derived to be $B_0 = \overline{Y} - B_1\overline{x}$. Remember to be careful that in SLR, the Y_i are random variables, while $x_i's$ are treated as constants.

a) (2 points) Show that B_0 is an unbiased estimator of β_0.

b) (2 points) Express B_0 as a linear combination of Y_i. i.e. $\sum_{i=1}^{n} a_i \cdot Y_i$.

c) (3 points) Use your answer in (b) to derive the $V(B_0)$. Simplify your expression for variance as much as possible.

d) (3 points) Use your answers above to define the probability distribution of B_0. Justify your distribution choice by referencing the appropriate assumptions of SLR.