MATH 338

MIDTERM 2
WED/THURS, NOVEMBER 1-2, 2017

Your name: _____

Your scores (to be filled in by Dr. Wynne):

Problem 1: ____/12

Problem 2: ____/5

Problem 3: ____/11

Problem 4: ____/5

Total: ____/33

You have 50 minutes to complete this exam.

You may refer to your textbook, any notes/code you wrote, anything on Titanium, and software help menus. You may ask Dr. Wynne to clarify what a question is asking for, or to help you troubleshoot RGuroo errors and/or debug your R code. You may not ask other people for help or use online resources other than those on Titanium or the software itself.

For full credit, include all R code (if using RStudio), graphs, and output. Save your answers as a .docx or .pdf file and upload the file to Titanium.

1. Finsterwalder (1976) explored a method of determining the amount of pesticide in food. The DDT dataset contains 15 measurements of the amount of the pesticide DDT in kale, in parts per million (ppm). Assume each measurement was conducted by an independent laboratory. Download the DDT.csv dataset from Titanium and import it to your software of choice.

A) [3 pts] Construct, **but do not interpret**, a 95% confidence interval for the mean amount of DDT in this particular batch of kale.

Write your answer in this box. Paste any code and output (tables/graphs/etc.) below the box.

B) [3 pts] Are the assumptions for correct interpretation of a 95% confidence interval met? Support your answer using software output.

Write your answer in this box. Paste any code and output (tables/graphs/etc.) below the box.

C) [2 pts] Identify each of the following statements as either true (T) or false (F). Argue why the statement is true or false using mathematics/logic and/or software output.

95% of all possible measurements of DDT in kale will fall within the interval you computed in Part (A).

TRUE/FALSE? Explanation: Paste any code and output (tables/graphs/etc.) below the box.

If you had a different sample of 15 measurements, you would have a different interval than you computed in Part (A).

TRUE/FALSE? Explanation: Paste any code and output (tables/graphs/etc.) below the box.

D) [3 pts] Suppose we take a different sample of 15 measurements. We are interested in testing whether the population mean measured amount of DDT in a new piece of kale is 3 ppm, or if it is greater. What is the power of the hypothesis test to detect the specific alternative that the true mean amount of DDT is 3.5 ppm, using a significance level of $\alpha = 0.05$? Assume the population standard deviation is exactly equal to the sample standard deviation of our current set of 15 measurements.

Write your answer in this box. Paste any code and output (tables/graphs/etc.) below the box.

E) [1 pt] What are the probabilities of Type I Error and Type II Error for the hypothesis test in Part (D)?

Probability of Type I Error:

Probability of Type II Error:

You do not need to paste any code/output beyond that you pasted in Part (D).

2. [5 pts] At DataFest 2017, students investigated over 10 million user-sessions from Expedia's hotel booking website. Some of the sessions resulted in the user booking a hotel and some did not. Suppose we take a simple random sample of 1000 user-sessions and find that 8 sessions ended in a booking. Construct and interpret a 95% confidence interval for the population proportion of sessions on Expedia's website that result in the user booking a hotel. Paste the appropriate output from software below. **Justify your choice of methods.**

Describe and justify your inferential technique, and write your interpretation/conclusion, in this box. Paste any code and software output (tables/graphs/etc.) below the box.

3. In 1876, Charles Darwin published the results of an experiment in which he recorded the height (to the nearest eighth of an inch) of 15 pairs of corn plants. One plant in each pair was produced by self-fertilization (variable "self") and one plant was produced by cross-fertilization (variable "cross"). Download the Darwin.csv dataset from Titanium and import it to your software of choice.

A) [2 pts] Darwin wanted to show that the cross-fertilized plants grew higher than self-fertilized plants. Given his experiment, convert his claim to an **appropriate** null and alternative hypothesis.

H_0:

H_a:

Define what the parameter(s) in your null hypothesis represent:

B) [1 pt] What are the sample mean and standard deviation of the heights of the 15 self-fertilized plants?

Sample Mean:

Sample Standard Deviation:

Paste any code and software output (tables/graphs/etc.) below the box.

C) [4 pts] At the $\alpha = 0.01$ significance level, does Darwin provide sufficient statistical evidence for his claim? Perform the hypothesis test suggested by your answer to Part (A).

Describe and justify your inferential technique, and write your conclusions, in this box. Paste any code and software output (tables/graphs/etc.) below the box.

D) [2 pts] Which of the following would stay the same if Darwin used a different sample of corn plants? Indicate all correct answers below by using **BOLD AND UNDERLINE**, highlighting, and/or red text.

null hypothesis test statistic p-value significance level

E) [2 pts] Identify each of the following statements as either true (T) or false (F). Argue why the statement is true or false using mathematics/logic and/or software output.

If the test statistic is within the critical region, but H_0 is true, you will commit a Type I Error.

TRUE/FALSE?

Explanation:

Paste any code and output (tables/graphs/etc.) below the box.

If the population distribution is normally distributed with theoretical mean $\mu = 2$ and theoretical standard deviation $\sigma = 5$, then the sampling distribution of the mean of 15 samples is also normally distributed with theoretical mean $\mu = 2$ and theoretical standard deviation $\sigma = 5$.

TRUE/FALSE?

Explanation:

Paste any code and output (tables/graphs/etc.) below the box.

4. [5 pts] In a recent meta-analysis (Holman et al., 2016; doi: 10.1371/journal.pbio.1002331), researchers investigated attrition rates of animals in pre-clinical studies. In 203 out of 316 stroke-related studies, and in 148 out of 206 cancer-related studies, the researchers were unable to determine even the initial sample size of animals in the study. Determine whether stroke researchers and cancer researchers have different standards for publishing sample size in their studies. **Justify all assumptions used to reach your conclusions.**

Describe and justify your inferential technique, and write your interpretation/conclusion, in this box. Paste any code and software output (tables/graphs/etc.) below the box.