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.

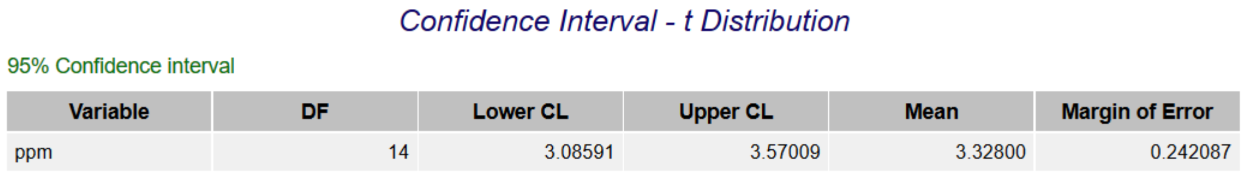
For full credit, include all R code (if using RStudio), graphs, and output. Save your answers as a .docx or .pdf file and upload the file to Titanium.

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.

2 pts for reporting a one sample t confidence interval (3.086, 3.570)

1 pt for showing the output (below)



Confidence Interval - t Distribution

95% Confidence interval

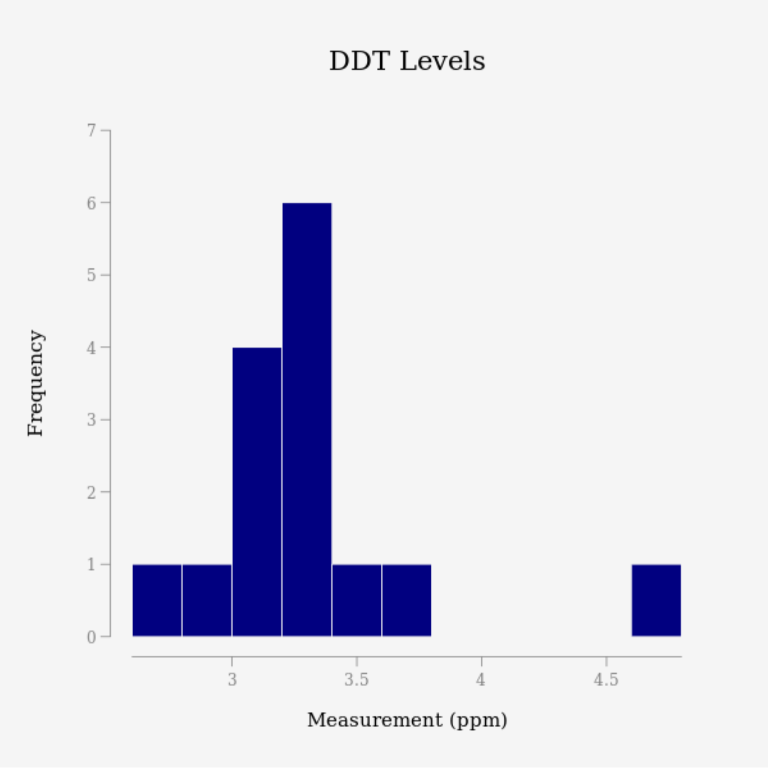
Variable	DF	Lower CL	Upper CL	Mean	Margin of Error
ppm	14	3.08591	3.57009	3.32800	0.242087

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

1 pt No, because the sample size is not high enough

1 pt We have quite an obvious outlier (at 4.64) and only 15 data points

1 pt for showing the histogram below to justify the reason why



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.

0.5 pts for true/false, 0.5 pts for explanation, for each statement

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

FALSE

Explanation: This statement talks about a population, whereas the confidence interval is about a population mean.

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

TRUE

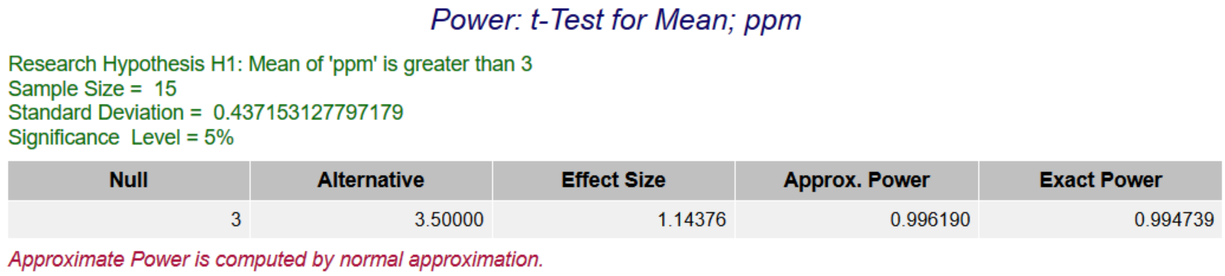
Explanation: The confidence interval is centered at the sample mean, which will likely be different for a different sample.

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 α = 0.05? Assume the population standard deviation is exactly equal to the sample standard deviation of our current set of 15 measurements.

1 pt for noticing we want to do a power analysis using H_0: μ = 3 vs. H_a: μ > 3

1 pt the power is 0.995 (or 0.996 if using normal approximation)

1 pt pasting either the output table or graph (below)

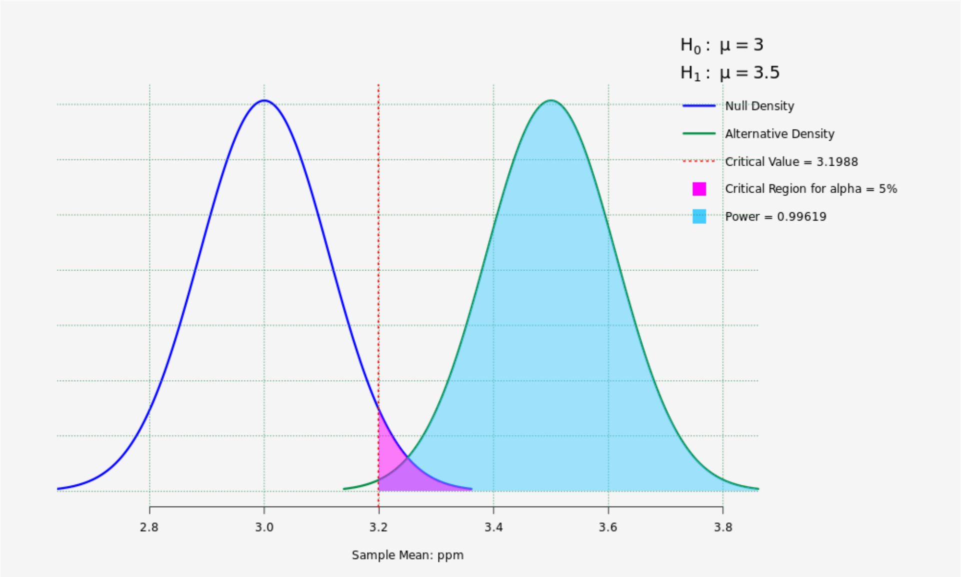


Power: t-Test for Mean; ppm

Research Hypothesis H1: Mean of 'ppm' is greater than 3
Sample Size = 15
Standard Deviation = 0.437153127797179
Significance Level = 5%

Null	Alternative	Effect Size	Approx. Power	Exact Power
3	3.50000	1.14376	0.996190	0.994739

Approximate Power is computed by normal approximation.



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

0.5 pts per answer

Probability of Type I Error: 0.05

Probability of Type II Error: 1 – 0.995 = 0.005

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.

We have a single population of user-sessions. We have a categorical variable with SUCCESS = booking and FAILURE = no booking, so we should use a one-proportion confidence interval.

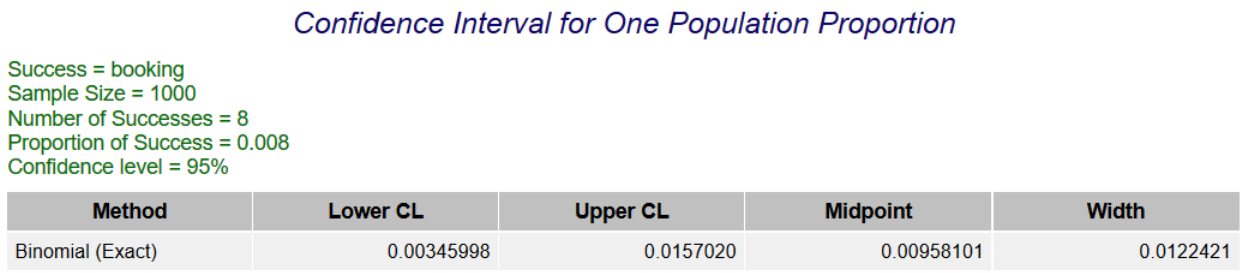
Since we have only 8 successes, a one-proportion z confidence interval is not justified, and we should use a binomial exact confidence interval.

2 pts for justifying choice of methods

1 pt for reporting the binomial exact confidence interval (0.0035, 0.0157) – an incorrect method would give a CI of (0.0024, 0.0135) or (0.0037, 0.0164)

1 pt for showing the output (below)

1 pt for interpretation: We are 95% confident that between 0.3% and 1.6% of user-sessions will result in a booking



Confidence Interval for One Population Proportion

Success = booking
Sample Size = 1000
Number of Successes = 8
Proportion of Success = 0.008
Confidence level = 95%

Method	Lower CL	Upper CL	Midpoint	Width
Binomial (Exact)	0.00345998	0.0157020	0.00958101	0.0122421

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: $\mu_d = 0$

H_a: $\mu_d > 0$ ($\mu_d < 0$ if defined as self – cross)

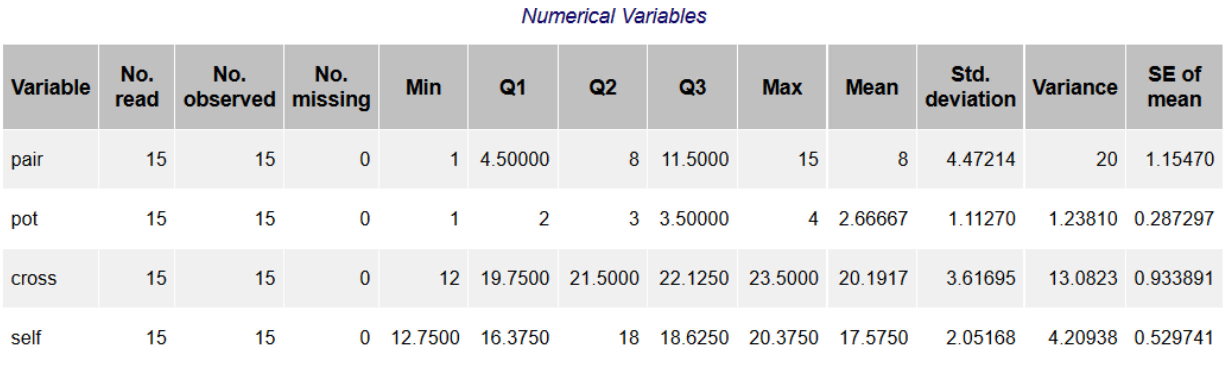
Define what the parameter(s) in your null hypothesis represent: μ_d represents the population mean of (cross – self) heights.

0.5 pts for using a matched-pairs H0/Ha, 0.5 points for appropriate definition of parameters

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

Sample Mean: 17.575 inches

Sample Standard Deviation: 2.05168 inches



Numerical Variables

Variable	No. read	No. observed	No. missing	Min	Q1	Q2	Q3	Max	Mean	Std. deviation	Variance	SE of mean
pair	15	15	0	1	4.50000	8	11.5000	15	8	4.47214	20	1.15470
pot	15	15	0	1	2	3	3.50000	4	2.66667	1.11270	1.23810	0.287297
cross	15	15	0	12	19.7500	21.5000	22.1250	23.5000	20.1917	3.61695	13.0823	0.933891
self	15	15	0	12.7500	16.3750	18	18.6250	20.3750	17.5750	2.05168	4.20938	0.529741

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

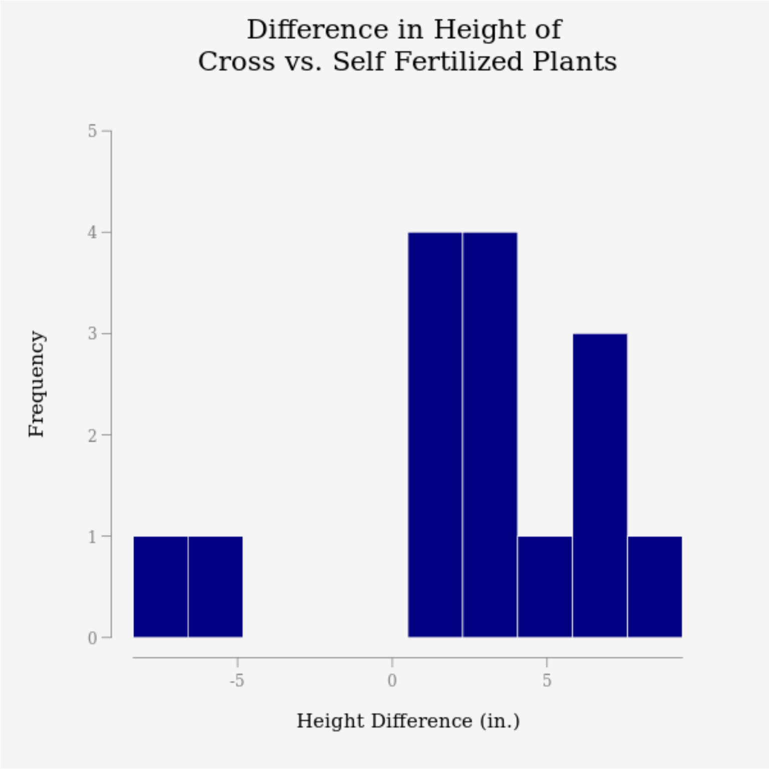
The histogram indicates that we have a couple of major outliers on the low side. If you indicate that we shouldn't do a t-test because of this, that's fine. However, we did not cover an alternative to the matched pairs t-test, so we are sort of stuck here.

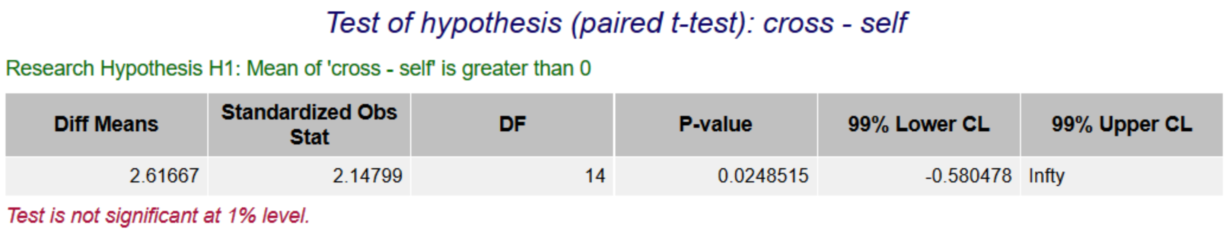
1 pt for obtaining a test statistic of 2.15 and a p-value of 0.025

1 pt for pasting the output table below or the p-value/critical region graph. Note: critical region graphs must have the appropriate significance level indicated.

1 pt since p > α, we fail to reject the null hypothesis.

1 pt At the 1% significance level, we do not have sufficient evidence to prove Darwin's claim that cross-fertilized plants grow larger, on average, than self-fertilized plants. Or, we conclude that the two types of plants are not significantly different.





Test of hypothesis (paired t-test): cross - self

Research Hypothesis H1: Mean of 'cross - self' is greater than 0

Diff Means	Standardized Obs Stat	DF	P-value	99% Lower CL	99% Upper CL
2.61667	2.14799	14	0.0248515	-0.580478	Infty

Test is not significant at 1% level.

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.

0.5 pts for true/false, 0.5 pts for explanation, for each statement

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

TRUE

Explanation: If the test statistic is in the critical region, we reject H_0. If we reject H_0 when it is true, that is a Type I Error.

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 $\sigma = 5$.

FALSE

Explanation: The theoretical standard deviation of the sampling distribution is 5/sqrt(15) = 1.291

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.**

We have two populations – stroke researchers and cancer researchers. We have a categorical response variable with SUCCESS = could not determine sample size and FAILURE = could determine sample size.

Let Population 1 = stroke researchers. We have 203 successes and 316 – 203 = 113 failures in population 1. Let Population 2 = cancer researchers. We have 148 successes and 206 – 148 = 58 failures in population 2. As far as we know the BINS assumptions are met in each population. Therefore, we can do two-proportion z procedures.

2 pts for justification

1 pt for one of the following:

CI: For a 95% CI, (-0.157, 0.005) or (-0.161, 0.009) if using the continuity correction

HT of H_0: $p_1 = p_2$ against H_a: $p_1 \neq p_2$: $z = -1.809$, p-value = 0.0704

Fisher's Exact Test: p-value of 0.0858

1 pt for showing the output (below)

1 pt for interpretation: CI: We are 95% confident that between 15.7% (16.1%) fewer and 0.5% (0.9%) more stroke researchers fail to report sample size, compared to cancer researchers. Therefore, we cannot determine at the 5% significance level whether one group is worse with respect to failing to report sample size.

HT: Since $p > \alpha$, we fail to reject the null hypothesis. It is reasonable to assume that the two groups of researchers are equivalent with respect to failing to report sample size.

Confidence Interval for Difference of Two Population Proportions

Success = No
Population 1 = Stroke,   Population 2 = Cancer
Sample Size: Stroke = 316,   Cancer = 206
Number of Successes: Stroke = 203,   Cancer = 148
Proportion of Success: Stroke = 0.6424,   Cancer = 0.7184
Confidence level = 95%

Method	Lower CL	Upper CL	Midpoint	Width
Large Sample z	-0.157064	0.00498138	-0.0760415	0.162046
Large Sample z with cc	-0.161074	0.00899084	-0.0760415	0.170065

cc: Continuity correction is used in computing the interval.

Method: Large Sample z Test (Pooled Standard Error)

Success = No
Population 1 = Stroke,   Population 2 = Cancer
Sample Size: Stroke = 316,   Cancer = 206
Number of Successes: Stroke = 203,   Cancer = 148
Proportion of Success: Stroke = 0.6424,   Cancer = 0.7184
Significance level = 5%
Research Hypothesis H1: Proportion of 'Stroke - Cancer' is not equal to 0

Proportion Stroke	Proportion Cancer	Difference	Standardized Obs Stat	P-value	95% Lower CL	95% Upper CL
0.642405	0.718447	-0.0760415	-1.80930	0.0704037	-0.158415	0.00633192

Test is not significant at 5% level.

Method: Fisher Exact Test

Success = No
Population 1 = Stroke,   Population 2 = Cancer
Sample Size: Stroke = 316,   Cancer = 206
Number of Successes: Stroke = 203,   Cancer = 148
Proportion of Success: Stroke = 0.6424,   Cancer = 0.7184
Significance level = 5%
Research Hypothesis H1: Proportion of 'Stroke - Cancer' is not equal to 0

Proportion Stroke	Proportion Cancer	Difference	P-value
0.642405	0.718447	-0.0760415	0.0857701

Test is not significant at 5% level.