Midterm 2 Solutions

Math 338 Fall 2017

Problem 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.

```
# Read the data set into RStudio
library(readr)
DDT <- read_csv("~/Math 338 Fall 2017/Exams/DDT.csv")
head(DDT)  # view the first few entries in the data set
```

```
## # A tibble: 6 × 1
##     ppm
##   <dbl>
## 1  2.79
## 2  2.93
## 3  3.22
## 4  3.78
## 5  3.22
## 6  3.38
```

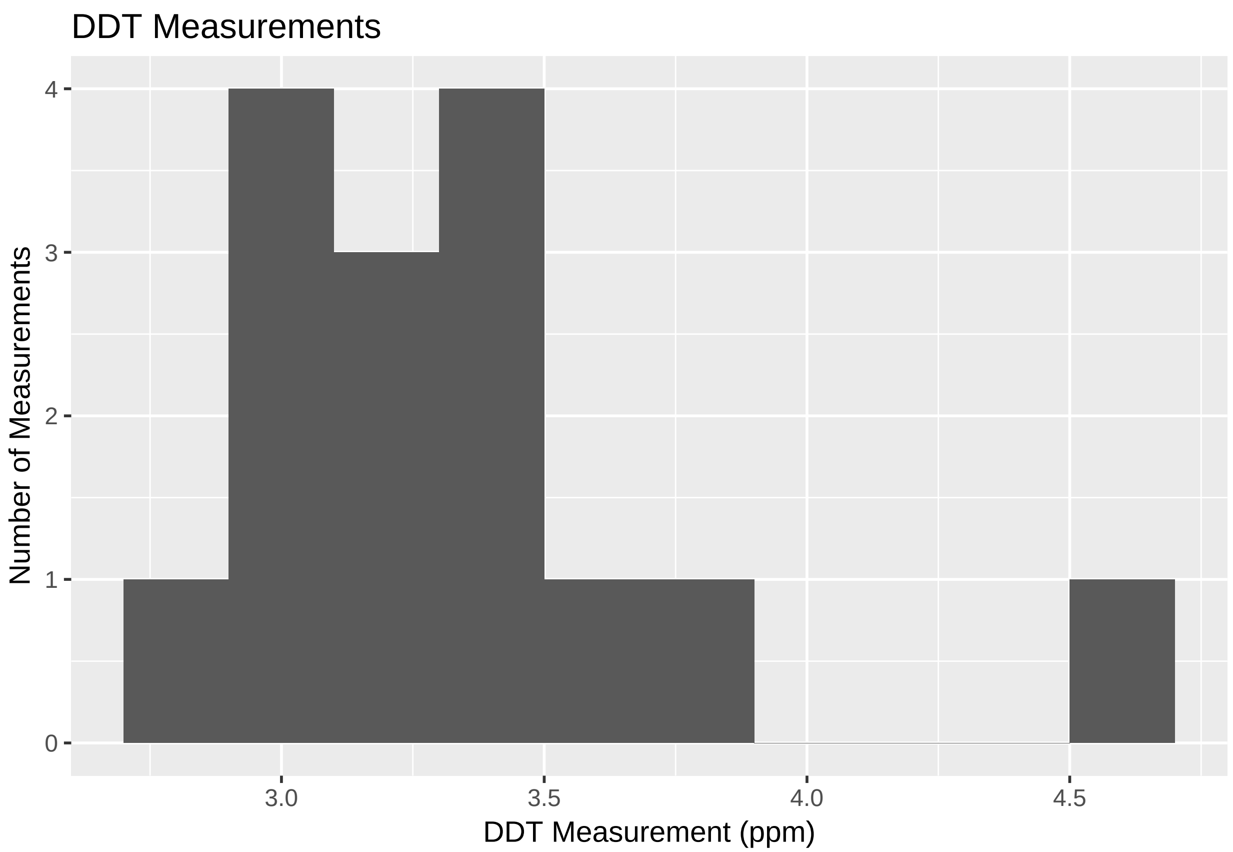
```
# We want a one-sample t CI
t.test(DDT$ppm, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  DDT$ppm
## t = 29.485, df = 14, p-value = 5.299e-14
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  3.085913 3.570087
## sample estimates:
## mean of x
##     3.328
```

The confidence interval is (3.086, 3.570).

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

```
library(ggplot2)
DDT_plot <- ggplot(DDT, aes(x = ppm)) + geom_histogram(center = 3,
    binwidth = 0.2) + labs(x = "DDT Measurement (ppm)",
    y = "Number of Measurements", title = "DDT Measurements")
print(DDT_plot)
```



No, the assumptions are not met because the sample size is too small. We have an obvious outlier at 4.64 and only 15 observations.

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).

FALSE: 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: 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 $\alpha = 0.05$? Assume the population standard deviation is exactly equal to the sample standard deviation of our current set of 15 measurements.

```
power.t.test(n = 15, delta = 3.5 - 3, sd = sd(DDT$ppm),
    sig.level = 0.05, type = "one.sample", alternative = "one.sided")
```

```
##
##      One-sample t test power calculation
##
##               n = 15
##           delta = 0.5
##              sd = 0.4371531
##       sig.level = 0.05
##           power = 0.9947395
##     alternative = one.sided
```

The power of our hypothesis test to detect the specific alternative of $\mu = 3.5$ is 0.995.

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

The probability of Type I Error is our significance level, $\alpha = 0.05$.

The probability of Type II Error is 1 - power = 0.005.

Problem 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.

```
binom.test(n = 1000, x = 8, conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  8 and 1000
## number of successes = 8, number of trials = 1000, p-value <
## 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.003459976 0.015702049
## sample estimates:
## probability of success
##                  0.008
```

We are 95% confident that between 0.3% and 1.6% of user-sessions will result in a booking.

Problem 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] 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.

$\mu_d = 0$, where μ_d is the population mean of the difference in heights between cross fertilized and self fertilized plants.

$\mu_d > 0$

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

```
library(readr)
Darwin <- read_csv("~/Math 338 Fall 2017/Exams/Darwin.csv")
head(Darwin)
```

```
## # A tibble: 6 × 4
##    pair   pot  cross   self
##   <int> <int>  <dbl>  <dbl>
## 1     1     1 23.500 17.375
## 2     2     1 12.000 20.375
## 3     3     1 21.000 20.000
## 4     4     2 22.000 20.000
## 5     5     2 19.125 18.375
## 6     6     2 21.500 18.625
```

```
mean(Darwin$self)
```

```
## [1] 17.575
```

```
sd(Darwin$self)
```

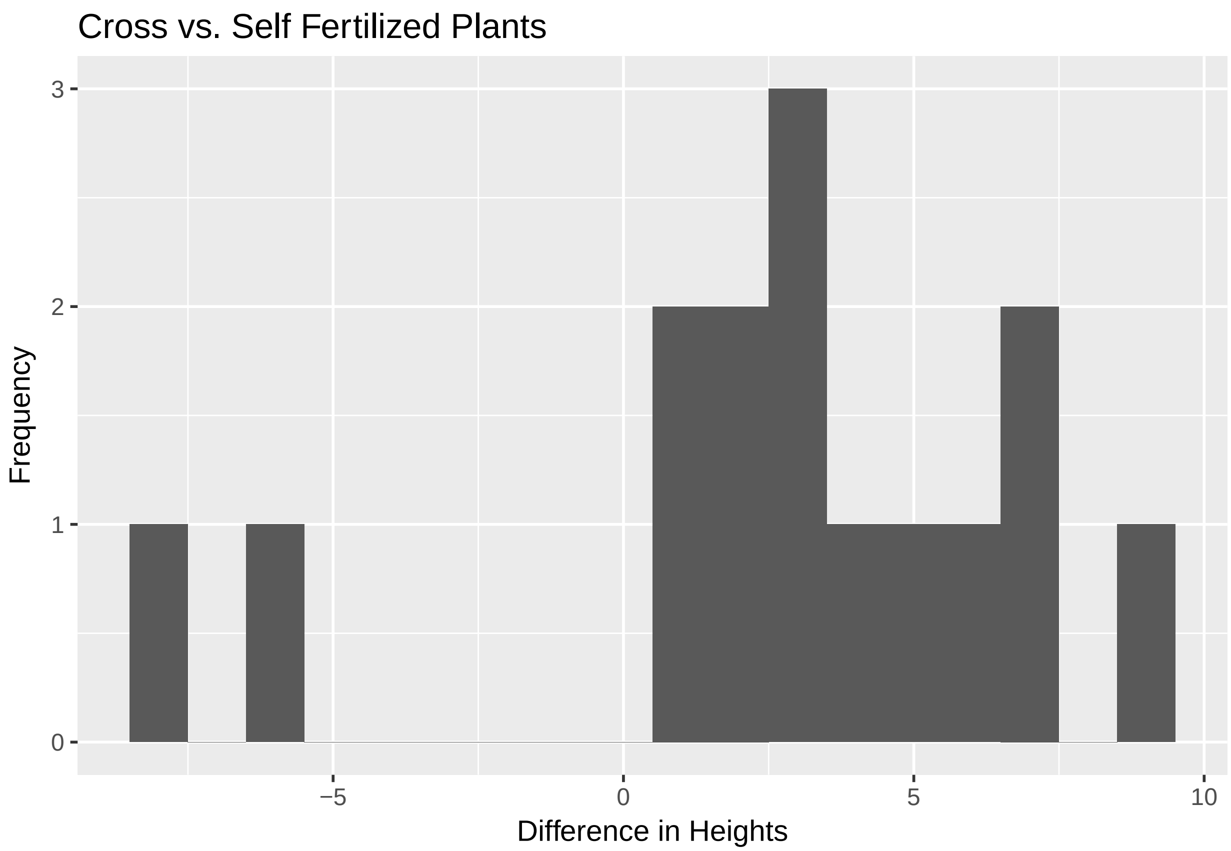
```
## [1] 2.051676
```

The sample mean is 17.575 inches and the sample standard deviation is 2.05168 inches.

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).

```
library(dplyr)
Darwin_diff <- Darwin %>% mutate(diff = cross - self)

ggplot(Darwin_diff, aes(x = diff)) + geom_histogram(center = 0,
    binwidth = 1) + labs(x = "Difference in Heights",
    y = "Frequency", title = "Cross vs. Self Fertilized Plants")
```



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.

```
t.test(Darwin_diff$diff, mu = 0, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  Darwin_diff$diff
## t = 2.148, df = 14, p-value = 0.02485
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  0.4710482       Inf
## sample estimates:
## mean of x
##  2.616667
```

We have a t-statistic of 2.148 and a p-value of 0.025. Since the p-value is greater than $\alpha = 0.01$, we do not have a significant result. We conclude that the cross and self fertilized plants are not significantly different.

D) [2 pts] Which of the following would stay the same if Darwin used a different sample of corn plants?

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: 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 $\sigma = 2$ and theoretical standard deviation $\sigma = 5$.

FALSE: The theoretical standard deviation of the sampling distribution is $5/\sqrt{15} = 1.291$

Problem 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 papers. 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.

```r
prop.matrix <- matrix(c(203, 316 - 203, 148, 206 -
    148), nrow = 2, ncol = 2, byrow = TRUE)
prop.test(prop.matrix, alternative = "two.sided")
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  prop.matrix
## X-squared = 2.9375, df = 1, p-value = 0.08655
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.161073920  0.008990843
## sample estimates:
##    prop 1    prop 2
## 0.6424051 0.7184466
```

If we interpret the confidence interval, we claim that we are 95% confident that between 16.1% fewer and 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.

If we interpret the hypothesis test, at the 5% significance level we do not have significance. It is reasonable to assume that the two groups of researchers are equivalent with respect to failing to report sample size.

Alternatively we can do Fisher's exact test. We will get a slightly different p-value but the conclusion is the same.

```r
prop.matrix <- matrix(c(203, 316 - 203, 148, 206 -
    148), nrow = 2, ncol = 2, byrow = TRUE)
fisher.test(prop.matrix, alternative = "two.sided")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  prop.matrix
## p-value = 0.08577
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.4711812 1.0470524
## sample estimates:
## odds ratio
##  0.7044921
```