Applied Data Analysis: Exam 1

Jiahuan Li

Exam 1 is an open book exam, which means that you can use your notes and R to take the exam. Discussing or sharing any materials or information in regards to the exam with other people and/or students is prohibited. You have 26 hours to take the exam - it is due on Thurs., Feb. 23 before 10 am EST. Upload your completed exam (both the knitted document and your .Rmd file) to Exam 1 under Assignments on Sakai. Save both files with your last and first names separated by a hyphen, such as Smith-Joe.

Use the "exam-template.Rmd" file to show your analyses and write your answers. Even if you have code in an R chunk, you must type out answers to the questions.

Throughout the exam, assume a significance level, α, of 5% unless otherwise stated.

Part 1

1. Calculate the probability of finding 3 acorns in a quadrat and 18 acorns in a quadrat (write your R code and answers)? Which number of acorns (3 or 18) is more likely? (Write your R code and answer.)

```r
# 3 acorns
dpois(3, 12.3)
```

```
## [1] 0.001411699
```

```r
# 18 acorns
dpois(18, 12.3)
```

```
## [1] 0.02952085
```

The probabilities of finding 3 and 18 acorns are 0.0014 and 0.0295, respectively. According to the probability value, finding 18 acorns in a quadrat is more likely.

2. What is the combined probability of counting fewer than 6 and more than 10 acorns in a quadrat (write your R code and answer here)?

```r
ppois(5, 12.3) + ppois(10, 12.3, lower.tail = F)
```

```
## [1] 0.7002536
```

The combined probability is 0.7003.

3. Answer these questions about the previous two questions (Part-1, Questions-1&2).

   a. What probability distribution did you use for the previous two questions?
   b. Why did you choose to use that probability distribution?

   I use Poisson distribution for the previous two questions. That is because Poisson distribution is useful for modeling count/discrete data that occur randomly in an interval of space (quadrat).

Part 2

1. What is the probability that in a random sample of 70 trees, 25 trees do not have seeds? (Write your R code and answer here.)

```r
dbinom(25, 70, 0.38)
```

```
## [1] 0.09165654
```

The probability for this question is 0.0917.

2. What is the probability of finding more than 50 trees with acorns based on the sample of 70 trees? (Write your R code and answer here.)

```r
pbinom(50, 70, 0.62, lower.tail = F)
```

```
## [1] 0.03797778
```

The probability for this question is 0.0380.

3. Answer these questions about the previous two questions (Part-2, Questions-1&2).

(a) What probability distribution did you use for the previous two questions?

(b) Why did you choose to use that probability distribution?

The binomial probability distribution is used for the two questions. It is chosen to describe the outcome of a series of Bernouilli trials. Each tree only has two situations: with or without acorns. Thus, the situation of each tree can be regarded as a Bernouilli trial.

Part 3

1. Write the null and alternative hypotheses for the problem about acorn weights from heated and unheated trees.

   The null hypothesis is that there is no statistically significant difference between the weights of acorns from the experimentally heated oak trees and the unheated trees. The alternative hypothesis is that heated trees will have smaller acorns with lighter weight compared to that of the unheated trees.

2. To test the alternative hypothesis, would you use a one-sided or two-sided statistical test?

   One-sided t-test will be applied since the alternative hypothesis is only interested in whether the acorn will be smaller in the heated group or not.

3. Calculate the 97% confidence intervals for the mean weight of acorns from heated trees and acorns from unheated trees.

   a. What are the 97% confidence intervals? (Clearly label *heated* and *unheated* CI's.)

```r
# heated CIs
c(4.4 - abs(qt(0.015, 19)) * sqrt(3.4) / sqrt(20),
  4.4 + abs(qt(0.015, 19)) * sqrt(3.4) / sqrt(20))
```

```
## [1] 3.432865 5.367135
```

```
# unheated CIs
c(6.7 - abs(qt(0.015, 19)) * sqrt(5.1) / sqrt(20),
  6.7 + abs(qt(0.015, 19)) * sqrt(5.1) / sqrt(20))
```

```
## [1] 5.515506 7.884494
```

The 97% confidence intervals for heated group is (3.43, 5.37). While that for unheated one is (5.52, 7.88).

b. In your own words, what does the confidence interval for one of these types of acorns signify?

   The CI means that we are 97% confident that the intervals (3.43, 5.37) and (5.52, 7.88) capture the true mean weight of the acorns of the heated and unheated populations. In other words, if we were to take 100 different samples and compute a 97% confidence interval for each sample, then approximately 97 of the 100 intervals would contain the true population mean value.

4. Based on your calculated confidence intervals, are the weights of the heated and unheated acorns significantly different at a significance level, α, of 3% (explain your answer)?

   Yes, they are significantly different. That is because there is no overlap between the above two intervals. And since they are the 97% confidence intervals of the two groups, the p-value, indicating the intersection of their probability distributions, will be smaller than 0.015 and fall in the rejection zone defined at a 3% significance level.

5. Is the mean weight of acorns from the sample of 20 unheated trees (e.g., 6.7 with a variance of 5.1) significantly different than the population mean weight at a significance level, α, of 0.05? (Write your R code below and write your answer in a complete sentence, correctly reporting the statistics.)

```
# t statistics
t <- (6.7 - 6.3)/ (sqrt(5.1)/sqrt(20))
t
```

```
## [1] 0.792118
```

```
# t > 0, thus
pvalue <- 1 - pt(t, 19)
pvalue
```

```
## [1] 0.2190353
```

```
# Since it is a two-sided test, the above p-value should compare to alpha equal to 0.025.
# Thus, to simplify understanding and still use the significance level
# of 0.05 to evaluate the result, I multiply the above p-value by 2.
# That is also what the t.test() function does.
pvalue2 <- pvalue * 2
pvalue2
```

```
## [1] 0.4380706
```

The mean weight of acorns from the sample of unheated trees is not significantly different from the population mean weight at a significance level of 0.05 ($t = 0.7921$, $df = 19$, $p = 0.4381$).

Note that it is a two-side test and the original calculated p-value is multiplied by 2 to make it comparable to the significance level 0.05 rather than 0.025.

Part 4

1. Answer the following questions using the two samples of data provided for acorns from heated and unheated trees.

    a. Name the *specific* type of statistical test that should be used to determine whether acorns from heated trees are shorter than acorns from unheated trees.

       It is a two-sample one-sided t-test.

    b. In addition to the assumption that the data are independent observations, name two assumptions of your statistical test evaluating whether acorns from heated trees are shorter than acorns from unheated trees.

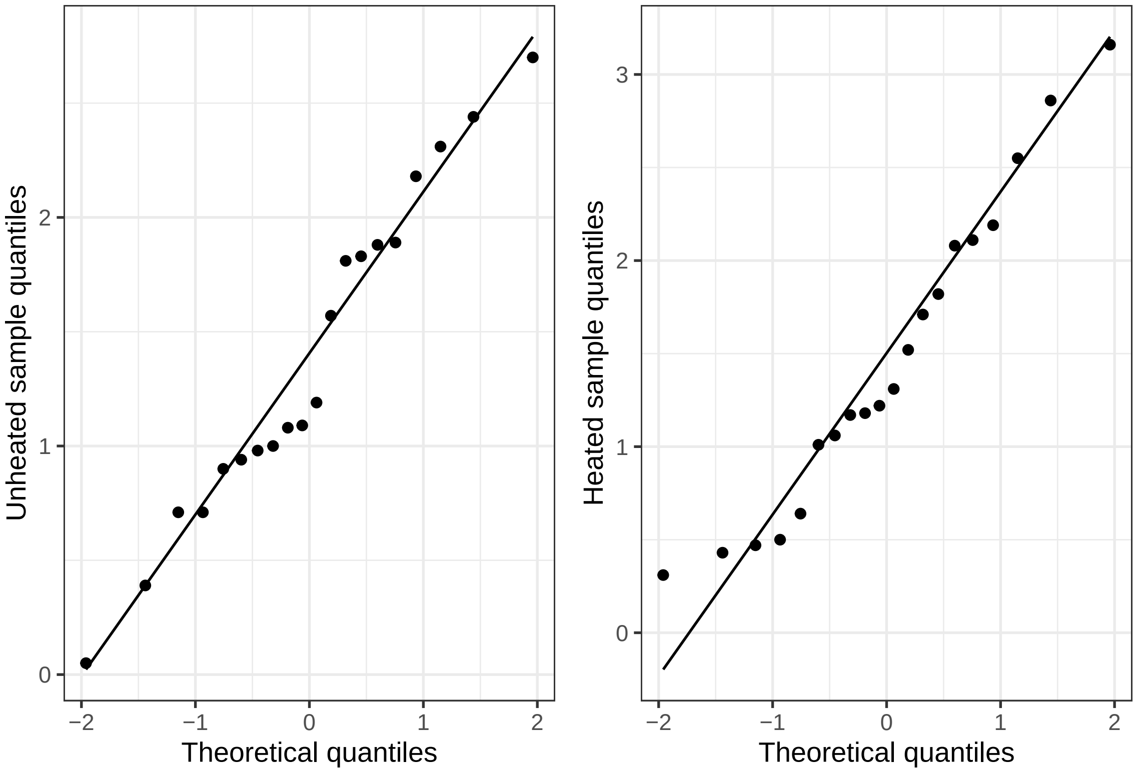
       Assumption 1: The variances of the two samples are assumed to be equal.

       Assumption 2: The two independent samples are normally distributed.

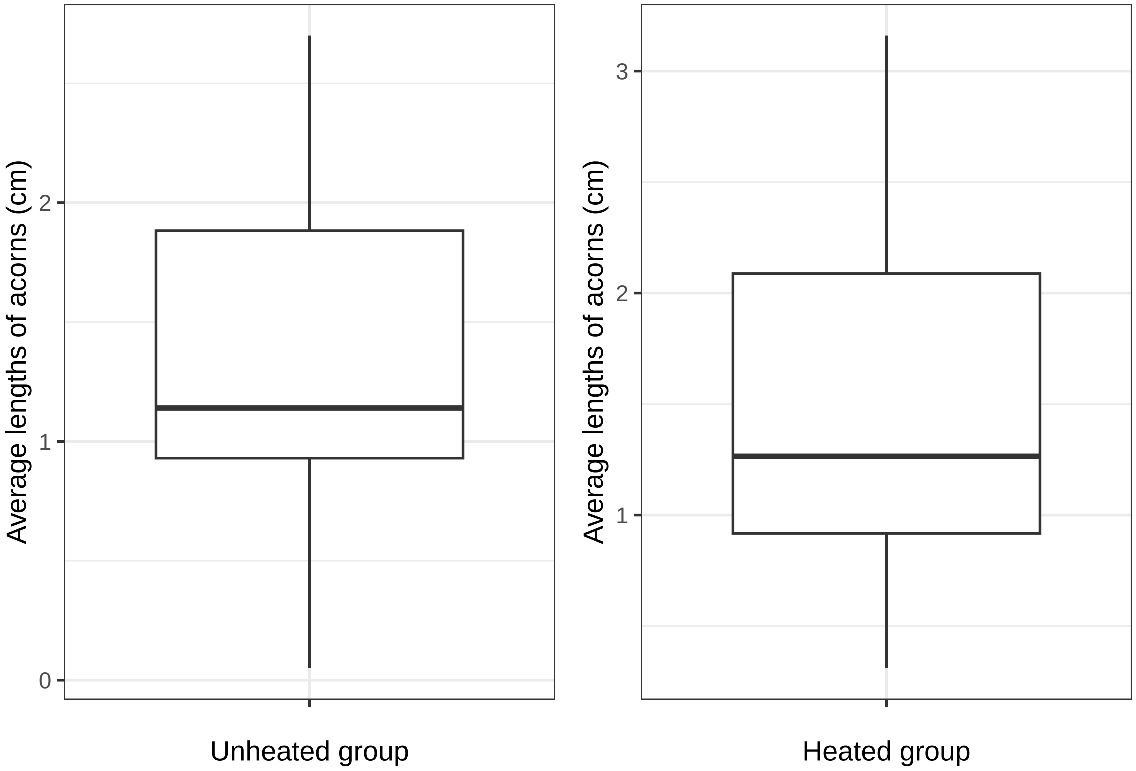
2. Test the assumptions from Part-4, Question-1(b) both (a) graphically and (b) using statistical tests.

    a. graph

```
unheated <- c(0.90, 1.88, 1.89, 0.94, 2.31, 1.57, 1.19, 2.70, 1.83, 0.71,
              2.18, 0.98, 1.81, 1.08, 1.09, 1.00, 2.44, 0.39, 0.05, 0.71)

heated <- c(1.06, 2.11, 0.64, 1.71, 1.82, 2.86, 2.19, 1.31, 2.55, 1.18,
            2.08, 0.50, 3.16, 0.31, 1.17, 1.22, 0.43, 1.52, 0.47, 1.01)

# normality
treated <- ggplot(data.frame(heated), aes(sample=heated)) +
  stat_qq() + stat_qq_line() +
  ylab("Heated sample quantiles") +
  xlab("Theoretical quantiles") + theme_bw()

control <- ggplot(data.frame(unheated), aes(sample=unheated)) +
  stat_qq() + stat_qq_line() +
  ylab("Unheated sample quantiles") +
  xlab("Theoretical quantiles") + theme_bw()

library(patchwork)
control + treated
```



```
# equal variance
tr <- ggplot(as.data.frame(heated), aes(x="", y = heated)) +
  geom_boxplot() +
  ylab("Average lengths of acorns (cm)") + xlab("Heated group") +
  theme_bw()

con <- ggplot(as.data.frame(unheated), aes(x="", y = unheated)) +
  geom_boxplot() +
  ylab("Average lengths of acorns (cm)") + xlab("Unheated group") +
  theme_bw()

con + tr
```



b. formal statistical tests

```
# normality
shapiro.test(unheated)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  unheated
## W = 0.96415, p-value = 0.6297
```

```
shapiro.test(heated)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  heated
## W = 0.95231, p-value = 0.4035
```

```
# equal variance
var.test(unheated, heated)
```

```
##
##  F test to compare two variances
##
## data:  unheated and heated
## F = 0.75289, num df = 19, denom df = 19, p-value = 0.5422
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2980042 1.9021469
## sample estimates:
## ratio of variances
##          0.7528929
```

c. Do the data meet the assumptions for conducting a statistical test? Explain using the results from both your graphical and formal test of the assumptions.

   Yes, the data meet the assumptions.

   For normality, points on the q-q plot fall approximately on a straight one-to-one line, which implies the match between sample quantiles and the theoretical standard normal distribution quantiles. Also, Shapiro-Wilk normality test results of the two samples are both larger than the 0.05 significance level. Thus, the null hypothesis that they are normally distributed cannot be rejected.

   For equal variances, we can find the locations of the two samples' boxplots are close and the overlap is huge. The median, box quartiles, and whiskers are all similar shown in the two graphs. Besides, the F test to compare two variances reports a p-value much higher than the signicance level. Thus, we cannot reject the null hypothesis that the true ratio of variances is equal to 1. That is, their variances are much likely to be equal.

3. Now conduct a statistical test to examine whether heated trees have significantly shorter acorns than unheated trees. Answer the following questions.

```
t.test(heated, unheated, alternative = "less", var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  heated and unheated
## t = 0.33515, df = 38, p-value = 0.6303
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 0.4975098
## sample estimates:
## mean of x mean of y
##    1.4650    1.3825
```

a. What are your null and alternative hypotheses?

   Null hypothesis: the length of acorns of heated trees is not significantly different from that of unheated trees.

   Alternative hypothesis: acorns from heated trees are significantly shorter (cm) than acorns from unheated trees.

b. Do you reject or fail to reject your null hypothesis?

   I fail to reject the null hypothesis since the p-value is larger than 0.05.

c. Write a sentence articulating your conclusion from the test, making sure to correctly cite the statistics.

   There is no significant evidence to support that heated trees have shorter acorns than unheated trees ($t = 0.3352$, $df = 38$, $p = 0.6303$).

d. What is the 95% confidence interval from your test? In 1-2 sentences, explain what the confidence interval means in this situation.

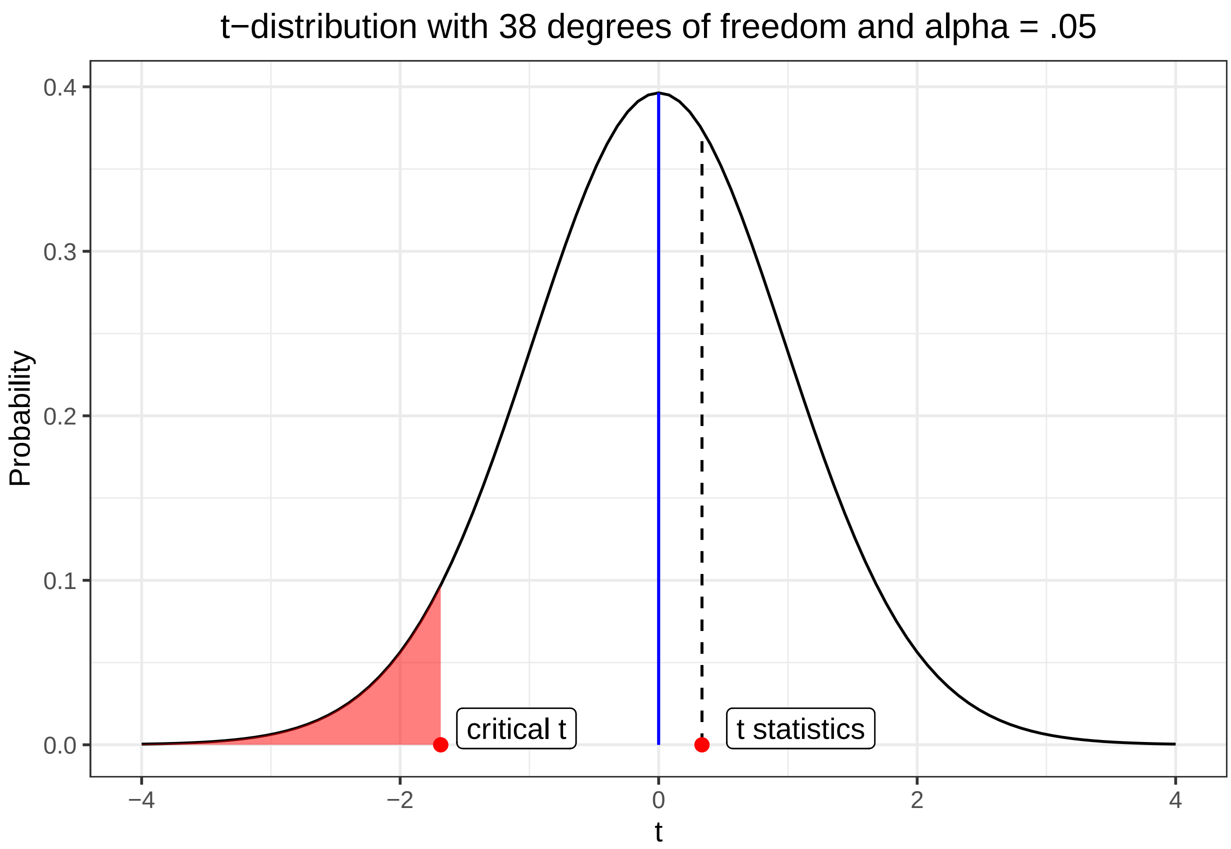
The 95% CI is $(-\infty, 0.4975)$. It means that we have 95% confidence that the acorns' true mean length of the heated tree population will not exceed that of **the unheated ones plus 0.4975**. Note that this conclusion is drawn under the specific one-sided "less than" alternative hypothesis.

4. Draw by hand the probability distribution associated with the above problem (Part 4 - Question 3): "test whether heated trees have significantly shorter acorns than unheated trees".

Label the following components: (a) mean; (b) rejection region for the test at a significance level of $\alpha = 0.05$; (c) the critical value(s) of the statistic, (d) the statistic calculated from your test of this question, and (e) x- and y-axis titles.

You do not need to upload your drawing. Answer the below questions from your drawing:

```r
ggplot(data.frame(x = c(-4, 4)), aes(x = x)) +
  stat_function(fun = dt, args = list(df = 38)) +
  stat_function(fun = dt, args = list(df = 38),
                xlim = c(-4, qt(0.05, 38)), geom = "area",
                fill = "red", alpha = 0.5) +
  geom_segment(x = 0.33515, y = 0,
               xend = 0.33515, yend = dt(0.33515, 38),
               linetype = "dashed") +
  geom_segment(x = 0, y = 0,
               xend = 0, yend = dt(0, 38), color = 'blue') +
  geom_point(aes(x=0.33515, y=0), colour="red", size = 2) +
  geom_point(aes(x=qt(0.05, 38), y=0), colour="red", size = 2) +
  geom_label(label = "t statistics", x = 1.1, y =0.01) +
  geom_label(label = "critical t", x = -1.1, y =0.01) +
  ylab("Probability") + xlab("t") + theme_bw() +
  ggtitle("t-distribution with 38 degrees of freedom and alpha = .05") +
  theme(plot.title = element_text(hjust = 0.5))
```



```r
qt(0.05, 38)
```

```
## [1] -1.685954
```

a. What is the mean of the distribution?

   As illustrated in the graph, the mean of the t-distribution is 0 indicated by the blue line.

b. describe the rejection region of the distribution.

   The rejection region of the distribution is colored red. It is a one-tail test and the significance level is 0.05. The region is located on the left since the alternative hypothesis states that heated trees have shorter acorns than unheated trees.

c. what is the critical value of the statistic(s)?

   The critical t in the situation for question 3 is - 1.6860.

Part 5

1. Answer the questions from part 5.

   a. Do the data meet the assumptions of your test? Explain.

```
unheated1 <- c(0.90, 1.88, 1.89, 0.94, 2.31, 1.57, 1.19, 2.70, 1.83, 0.71,
               2.18, 0.98, 1.81, 1.08, 1.09, 1.00, 2.44, 0.39, 0.05, 0.71)

heated1 <- c(0.65, 1.43, 1.44, 0.68, 1.78, 1.18, 0.88, 2.09, 1.39, 0.50,
             1.67, 0.71, 1.38, 0.79, 0.80, 0.73, 1.88, 0.24, 0.03, 0.50)

# normality
shapiro.test(unheated1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  unheated1
## W = 0.96415, p-value = 0.6297
```

```
shapiro.test(heated1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  heated1
## W = 0.96068, p-value = 0.5575
```

```
# equal variance
var.test(unheated1, heated1)
```

```
##
##  F test to compare two variances
##
## data:  unheated1 and heated1
## F = 1.5956, num df = 19, denom df = 19, p-value = 0.317
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6315762 4.0313217
## sample estimates:
## ratio of variances
##           1.595646
```

According to the p-values of the tests above, we can find that they are all higher than 0.05, indicating the null hypotheses cannot be rejected. Thus, they are considered to be approximately normally distributed and have equal variances. In fact, as the two samples are paired, their variances are thought to be equal and do not need to be tested.

b. Are acorns from the heated treatments significantly shorter than acorns from the same trees before they were heated? (Write your R code below and write your answer in a complete sentence, correctly reporting the statistics.)

```
t.test(heated1, unheated1, paired = T, alternative = "less")
```

```
##
```

```
##  Paired t-test
##
## data:  heated1 and unheated1
## t = -10.219, df = 19, p-value = 1.854e-09
## alternative hypothesis: true mean difference is less than 0
## 95 percent confidence interval:
##        -Inf -0.2866245
## sample estimates:
## mean difference
##          -0.345
```

Acorns from the heated treatments are significantly shorter than acorns from the same trees before they were heated (t = -10.219, df = 19, p = 1.854e-09).