ST 411/511 Homework 3

E. Alex Soderquist

Summer 2022

Instructions

This assignment is due by 11:59 PM, July 29th on Canvas via Gradescope. **You should submit your assignment as a PDF which you can compile using the provide .Rmd (R Markdown) template.**

> Note: Create a PDF by either compiling a PDF directly (via LaTeX) or from a Word Doc. Do not submit a PDF of the HTML output as they're cumbersome and difficult to read.

Include your code in your solutions and indicate where the solutions for individual problems are located when uploading into Gradescope (Failure to do so will result in point deductions). You should also write complete, grammatically correct sentences for your solutions.

Once you've completed the assignment, and before you submit it to Gradescope, you should read the document to ensure that (1) The computed values show up in the document, (2) the document "looks nice" (i.e. doesn't have extraneous code/outputs and includes *just* the essentials), and (3) ensure the document is "easy" to read.

Goals:

1. Practice two-sample t-tests by hand in order to gain a sense of *intuition* as to how they're conducted.
2. Think critically about study designs and look for instances of "poor design" or underlying issues which may impact analyses.
3. Extend our t-based methods to see how they can be adapted for "paired" samples.
4. See how useful (or useless?) data transformations are to answering questions of interest with statistical methods.

Question 1 (11 points)

122 guinea pigs were randomly assigned to either a control group ($X_1, X_2, \ldots, X_m; m = 64$) or to a treatment group that received a dose of *tubercle bacilli* ($Y_1, Y_2, \ldots, Y_n; n = 58$). The lifetime, in days, for each guinea pig was recorded. The data are available as `ex0211` in the `Sleuth3` library.

Note: Perform these calculations "by hand" (i.e. do not use the `t.test()` function or other built-in equivalents) using code which you write to compute the necessary values. Make sure to output the values requested in your document.

```
# Load the data
data(ex0211)
```

(a) (2 points) Compute the sample mean and sample variance for each group.

```
dataX <- c(ex0211$Lifetime[1:64])
meanX <- mean(dataX)
dataY <- c(ex0211$Lifetime[65:122])
meanY <- mean(dataY)
varX <- var(dataX)
varY <- var(dataY)
meanX
```

```
## [1] 345.2344
```

```
varX
```

```
## [1] 49371.67
```

```
meanY
```

```
## [1] 242.5345
```

```
varY
```

```
## [1] 13907.69
```

(b) (2 points) Compute the pooled variance estimate s_P^2.

```
sp2 <- ((64-1)*varX+(58-1)*varY)/(122-2)
sp2
```

```
## [1] 32526.28
```

(c) (2 points) Compute the t-statistic for testing the null hypothesis that the difference in population mean survival time between these two treatments is zero ($H_0 : \mu_X - \mu_Y = 0$).

```
(meanX-meanY)/sqrt(sp2*((1/64)+(1/58)))
```

```
## [1] 3.141064
```

(d) (2 points) Compute the critical value for a level $\alpha = 0.01$ one-sided test of the null hypothesis vs. the alternative that the difference in population mean survival time is greater than zero ($H_A : \mu_X - \mu_Y > 0$).

```
qt(0.01, df=64+58-2, lower.tail=FALSE)
```

```
## [1] 2.357825
```

(e) (1 point) Compute the p-value for the test using the alternative hypothesis specified in part (d) above.

```
1-pt(3.141064, df=64+58-2)
```

```
## [1] 0.001060091
```

(f) (2 points) Based on your answers to parts (d) and (e), would you reject the null hypothesis $H_0 : \mu_X - \mu_Y = 0$ vs. the alternative ($H_A : \mu_X - \mu_Y > 0$) at level $\alpha = 0.01$? Why? What does this conclusion mean in the context of the problem?

Based on the critical value for an alpha = 0.01 one-sided test, and the probability of getting a t-statistic we actually computed assuming the null hypothesis is true, we reject the null hypothesis because the probability was so low (p=0.001...). In terms of the context of the problem, this means that guinea pigs that receive a dose of tubercle bacili are on average more likely to live a shorter lifespan than guinea pigs that don't.

Question 2 (4 points) - Modified from *Sleuth* 3.16

A researcher has taken tissue cultures from 25 subjects. Each culture is divided in half, and a treatment is applied to one of the halves chosen at random. The other half is used as a control. After determining the percent change in the sizes of all culture sections, the researcher calculates an independent two sample t-analysis and a paired t-analysis to compare the treatment and control groups. Finding that the paired t-analysis gives a slightly larger standard error (and gives only half the degrees of freedom), the researcher decides to use the results from the unpaired analysis.

Is this a legitimate way to conduct a statistical analysis? Discuss whether the p-value from the independent sample t-analysis will be too big, too small, or about right. Write your answer as a short paragraph and be sure to explain your answer/reasoning

This is an invalid result, as we can only use a two-sample t-test when the samples are independent from each other. Because the samples are completely matched pairs, they are very much dependent. Because of this, and because the paired t-test gives only half the degrees of freedom, the p-value from the independent sample will be a little smaller than the paired samples based on the t-distribution.

Question 3 (4 points)

Researchers are interested in studying the effect of speed limits on traffic accidents. For a set of 100 roads with a speed limit of 55 miles-per-hour (mph), they record the number of accidents per year on each road for 10 consecutive years. The posted speed limit on each of these roads is then increased to 65 mph, and the number of accidents per year is recorded for each of the next 5 years.

Is there a violation of independence within and/or between the 55 mph and 65 mph groups? If so, discuss why the independence assumption is violated in relation to a cluster effect, serial correlation, and/or spatial correlation. Write your answer as a short paragraph and be sure to explain your answer/reasoning

I believe there is a violation of independence, assuming we are drawing conclusions about the effect of ALL speed limits on traffic limits. This would be a cluster effect because there is a certain subset of the population which would be more likely to drive on 55-65 mph roads such as interstates and highways: Truck drivers, for instance. Because we are only taking 55 mph zones into account, we may end up deriving extraneous information or encountering confounding variables while trying to make conclusions about ALL traffic accidents. (They happen anywhere!)

Question 4 (4 points)

Researchers studied 15 pairs of identical twins where only one twin was schizophrenic ('Affected'). They measured the volume of the left hippocampus region of each twin's brain. This data is available as `case0202` in the `Sleuth3` library.

```
data(case0202)
```

(a) (1 point) Is this paired data or two independent samples? Explain.

This is paired data, because each twin has a large amount of dependence on the other. The two observations together form a pair of data points, or a single difference depending on the test question.

(b) (3 points) Consider a hypothesis test to examine whether the difference in mean left hippocampus volume (Unaffected - Affected) is equal to zero, versus the two-sided alternative. Use the `t.test()` function in R to perform the appropriate t-test at significance level $\alpha = 0.01$. Report the p-value and what you conclude from the test.

```
unaff <- case0202$Unaffected
aff <- case0202$Affected
t.test(unaff, aff, alternative = "two.sided", paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  unaff and aff
## t = 3.2289, df = 14, p-value = 0.006062
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.0667041 0.3306292
## sample estimates:
## mean difference
##       0.1986667
```

The p-value of 0.006 at the 0.01 significance level means that we reject the null hypothesis in favor of the alternative hypothesis, i.e., that there is sufficient evidence to believe that the size of the left hippocampus in the affected twin is very likely to be different than the size of the left hippocampus in the unaffected twin.

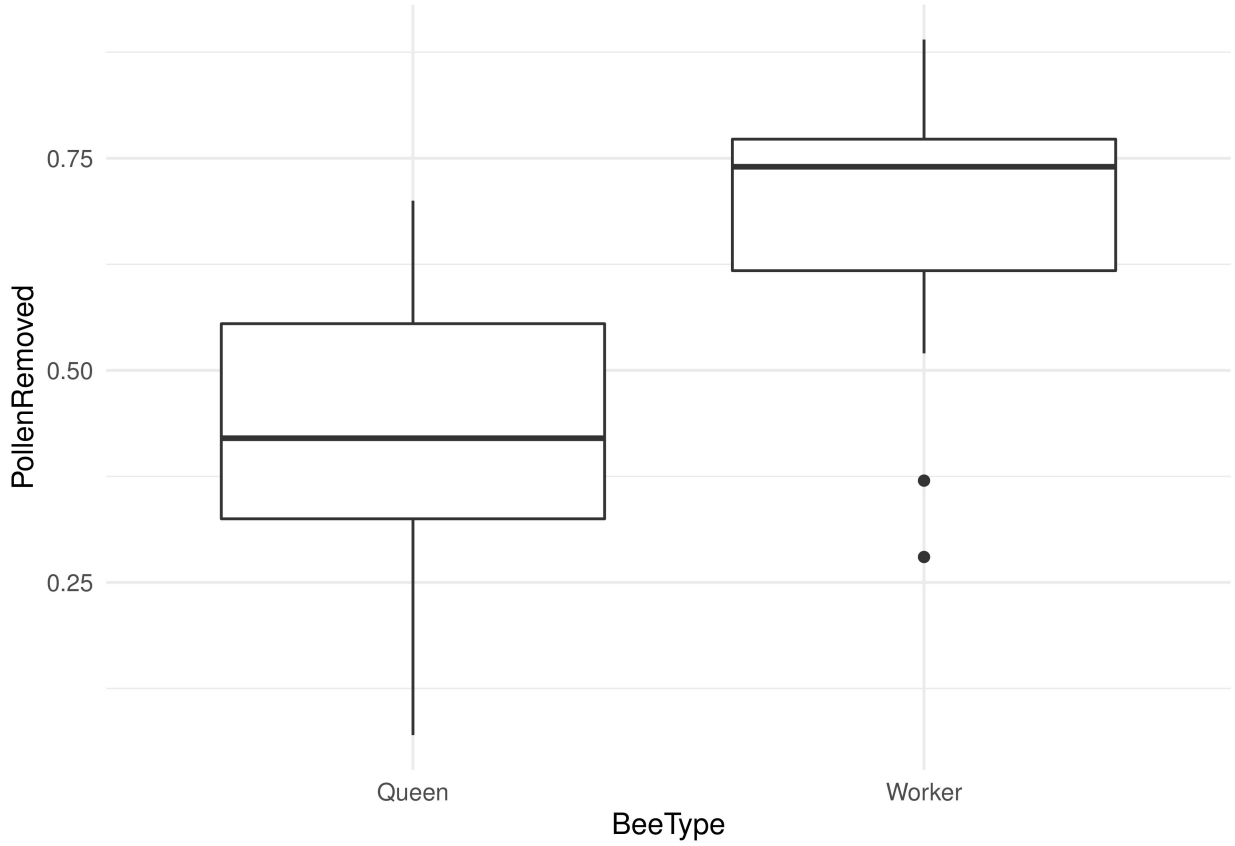
Question 5 (11 points) - Modified from *Sleuth* 3.27(a)

As part of a study to investigate reproductive strategies in plants, biologists recorded the time spent at sources of pollen and the proportions of pollen removed by bumble-bee queens and honeybee workers pollinating a species of lily. These data appear in `ex0327` in the `Sleuth3` package.

```
data(ex0327)
```

(a) (2 points) Create a side-by-side box plot for the proportion of pollen removed by queens and workers. What evidence do you see for doing a transformation?

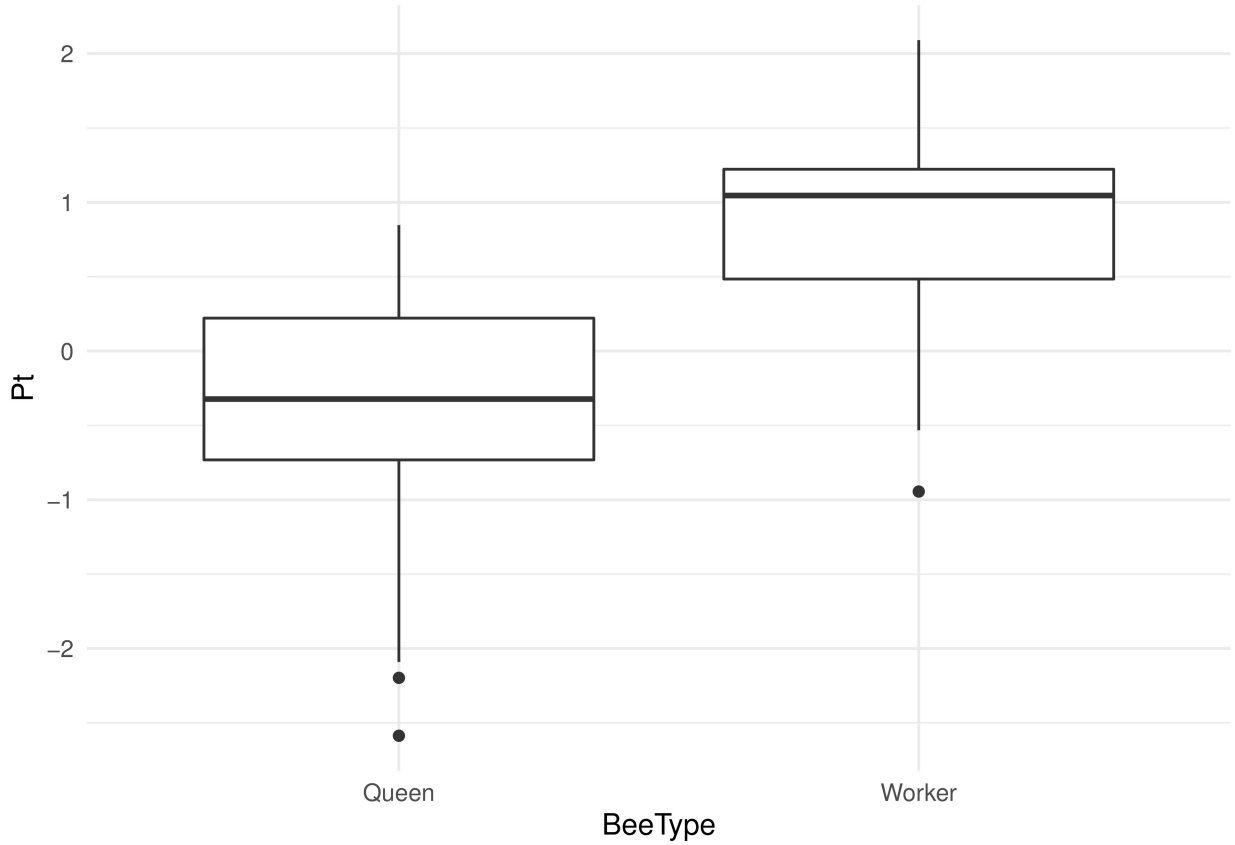
```
df <- data.frame(ex0327)
ggplot(df, aes(x=BeeType,y=PollenRemoved))+geom_boxplot() +theme_minimal()
```



The only evidence I see for doing a transformation resides in the skewed data of the worker bees. In this case, however, only one of the groups is highly skewed.

(b) (3 points) When the measurement is a proportion, P, of some amount, one useful transformation is the logit transformation which is defined as: $\log[P/(1-P)]$ with P being a proportion. This transformation is the log of the ratio of the proportion removed to the proportion not removed. Create a side-by-side box plot using the logit transformation of the pollen removed by queens and workers. Does this transformation seem to have helped us meet the t-test assumptions? Justify your answer. You can take the log of a vector `x` in R using `log(x)` (Note: The `log()` function is base e and not base 10.)

```
P <- ex0327$PollenRemoved
Pt <- log(P/(1-P))
ggplot(df, aes(x=BeeType,y=Pt))+geom_boxplot() +theme_minimal()
```



There is less skew among the worker bee group, and the data is slightly more centered for both groups. However the queen bee group has two outliers now, but overall this does help us satisfy t-test assumptions to perform a log/proportional log test.

(c) (4 points) Conduct a test, at the $\alpha = 0.05$ significance level, to decide whether the average of the logit transformed proportion of pollen removed is different for the two groups (Queens and Workers) using an appropriate t-test. You should use the `t.test()` function and answer this question using complete sentences. Be sure to state your null and alternative hypotheses, include the R output from the `t.test()` function, and write a complete conclusion for your test. A complete conclusion should include items such as whether or not you reject the null hypothesis at what significance level, the values of the test statistic and p-value, a confidence interval describing what values the true population parameter might plausibly be, as well as a sentence describing what the result of the test means in the context of the problem (bees in this case).

I hypothesize the average of the transformed proportion of pollen removed is different for the two groups.Judging by the IQR, there is no overlap between the two boxes, and suggests there is a measurably significant difference between the two groups. Null Hypothesis: H0: mu1 = mu2, i.e., the two population means are equal. Alternative Hypothesis: H0: mu1 $\neq$ mu2, i.e., the two population means are not equal (different)

```
t.test(c(Pt[1:35]), c(Pt[36:47]), var.equal=TRUE, alternative="two.sided")
```

```
##
##  Two Sample t-test
##
## data:  c(Pt[1:35]) and c(Pt[36:47])
## t = -3.8493, df = 45, p-value = 0.0003715
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.7490870 -0.5474536
## sample estimates:
##  mean of x  mean of y
## -0.3812734  0.7669968
```

Based on our significance level of 0.05, a given test statistic of -3.8493 translates to a p-value of 0.0003715 which states that there is a 0.037% chance that we would derive samples as or more unusual than the ones we found in sampling. Therefore because our p-value is less than the significance level, we reject the null hypothesis in favor of the idea that there is strong evidence to say the two bee populations have different means for harvesting pollen.

(d) (2 points) Use the `t.test()` function to construct a 90% confidence interval for the population difference in the mean of the logit proportion of pollen removed between the two bee groups. What is one issue with presenting this confidence interval to someone who is perhaps not as well-versed in statistics as yourself? In other words, why might this confidence interval be difficult to explain?

```
t.test(c(Pt[1:35]), c(Pt[36:47]), var.equal=TRUE, alternative="two.sided", conf.level=0.90)
```

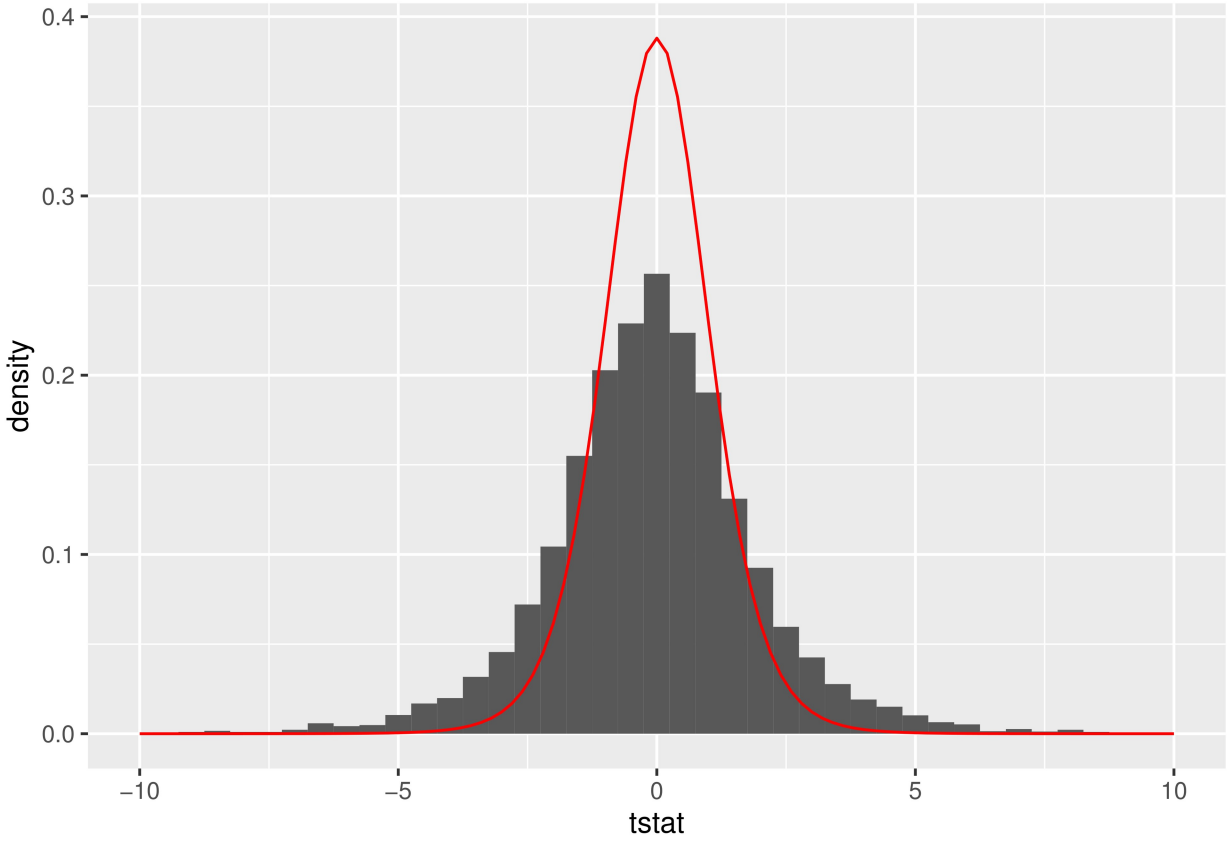
```
##
##  Two Sample t-test
##
## data:  c(Pt[1:35]) and c(Pt[36:47])
## t = -3.8493, df = 45, p-value = 0.0003715
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 90 percent confidence interval:
##  -1.649252 -0.647289
## sample estimates:
##  mean of x  mean of y
## -0.3812734  0.7669968
```

Our 90% confidence interval is (-1.649252, -0.647289), which means we are 90% confident that the mean difference of logit-proportional bee pollen removal is between -1.649252 logit-proportion pollen removal and -0.647289 logit-proportion pollen removal. That mangled sentence in itself should be evidence enough as to why non-statistically trained people might have trouble wrapping their heads around any kind of logarithmic transformation... especially since there is no way to transform such a scale back without making heavy and/or risky assumptions.

OPTIONAL QUESTION - Question 5 (0 points)

Suppose you have a normally distributed population with mean 60 and variance 25. Further, suppose that you draw samples of size 5 from the population but each sampled value accidentally gets duplicated so that you end up with 10 observations in the sample with each unique value occurring twice. Use the following code to produce a histogram of distribution of the test statistic for a one-sample t-test of $H_0 : \mu = 60$. The superimposed red curve is the theoretical $t_{(9)}$ distribution. (Note: You can ignore warnings about "removed rows containing non-finite values/missing values").

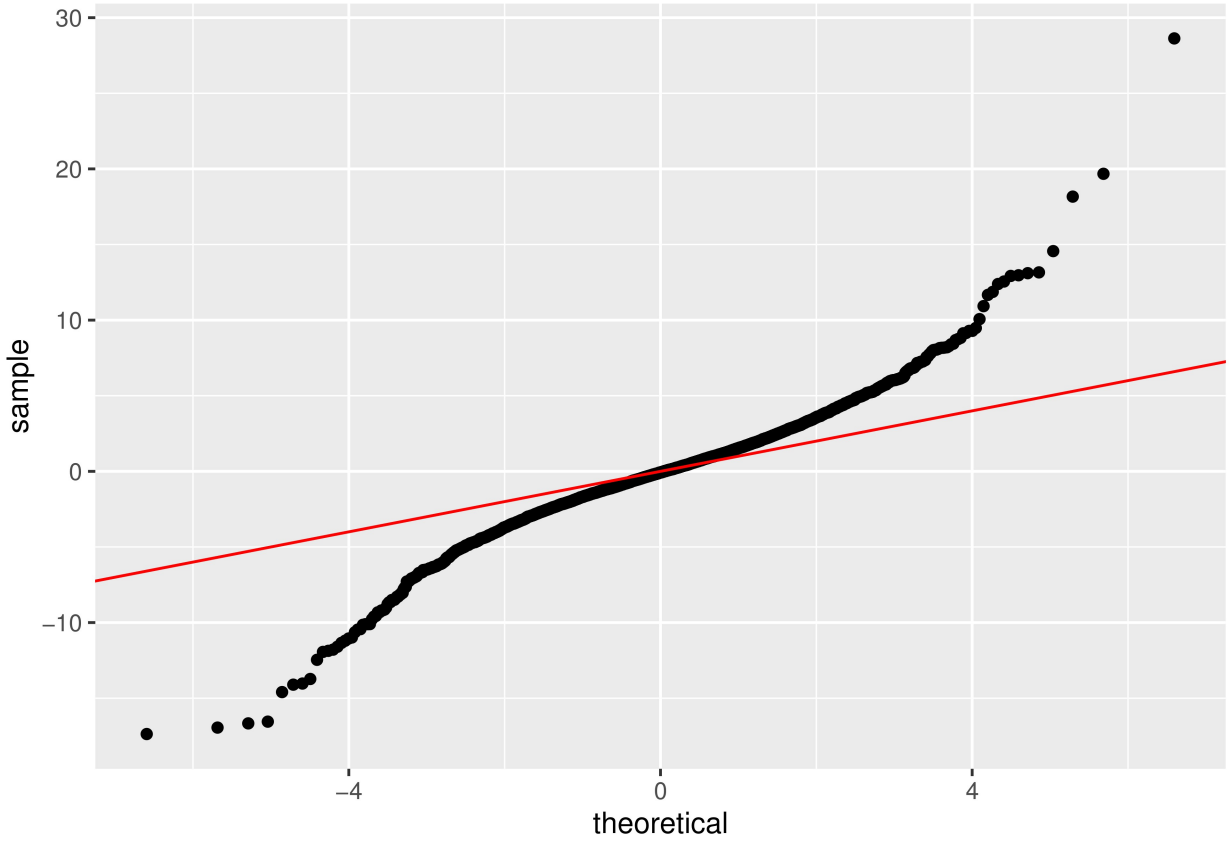


(a) (2 points) Based on what you know about the assumptions of the t-test, the observed distribution of the test statistics (the vertical bars), and the theoretical distribution of the test statistic (the red curve) answer the following questions: (1) Does having duplicated values in our sample violate any of our t-test assumptions? (2) How does the plotted histogram and curve help you see that a violation has occured? That is, what about the plot doesn't look "right" and how *should* the plot look if no assumption violations had occured?

1.

2.

(b) (2 points) Use the following code to produce a quantile-quantile plot for these simulated test statistics. Then answer the following questions: (1) Does this plot indicate that the duplicated values in each sample violates one of our t-test assumptions? If so, which one(s)? (2) Discuss how the plot helps you make this conclusion.



1.

2.

(c) (2 points) Copy and paste the code provided in part (a) of this question and alter the code by removing the `rep()` function wrapped around the `sample()` function to get rid of the duplicated values. Now the samples should be of size 5 with no duplication. Create a histogram of the distribution for the test statistic with the appropriate null distribution superimposed (i.e. How many degrees of freedom do you have now that $n = 5$?). Do the t-test assumptions appear to be met in this case? How can you tell?

```
# Copy, paste, and then alter the code from Question 4 Part (a) here.
```