ST 411/511 Homework 3

Pavan Sai Nallagoni

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)
names(ex0211)
```

```
## [1] "Lifetime" "Group"
```

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

```
Lifetimes <- ex0211$Lifetime
Groups <- ex0211$Group

Xmean <- mean(Lifetimes[Groups == "Control"])
Xmean
```

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

```
Ymean <- mean(Lifetimes[Groups == "Bacilli"])
Ymean
```

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

```
Xvar <- var(Lifetimes[Groups == "Control"])
Xvar
```

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

```
Yvar <- var(Lifetimes[Groups == "Bacilli"])
Yvar
```

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

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

```
#Given

m = 64
n = 58

PoolVar <- (((m-1) * Xvar) + ((n-1) * Yvar)) / (m + n - 2)
PoolVar
```

```
## [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 ($H_0 : \mu_X - \mu_Y = 0$).

```
SE <- sqrt(PoolVar * ( 1/m + 1/n))
delta = 0 #Given Ho

t_statistic <- ((Xmean - Ymean) - delta) / SE
t_statistic
```

```
## [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 ($H_A : \mu_X - \mu_Y > 0$).

```
alpha = 0.01
df = m + n - 2

qt(1 - alpha, df)
```

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

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

```
pt(t_statistic, df)
```

```
## [1] 0.9989399
```

(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?

We fail to reject the null hypothesis $H_0 : \mu_X - \mu_Y = 0$ at the $\alpha = 0.01$ significance level and not in favor of the alternative $H_A : \mu_X - \mu_Y > 0$. We have evidence which indicates that the mean difference between the two groups is negligible and equal to 0 where both the groups have an equal lifetime, on average.

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

No. This is not a legitimate way to conduct a statistical analysis. The researchers divided the cultures into two halves, so they're dependent samples since they're related and have same traits. Therefore, we cannot perform an independent two-sample t analysis. Instead, we can perform a paired t-analysis to compare the treatment and control groups.

Regarding the p-value, the paired t-analysis has a greater p-value than the independent two samples t-analysis. Because of the degrees of freedom, the paired t-analysis will have a lesser n than the independent two-sample t-analysis such that the paired t-analysis will have a higher standard error. Furthermore, which decreases the critical value t.

Therefore, it results in the greater p-value for the paired t analysis and lesser for the independent sample t-analysis when compared.

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

Yes. There is a violation of independence within and/or between the $55mph$ and $65mph$ groups. The following are the possible effects that can be violations in this case.

It's likely that a few roads (more dangerous) with $65mph$ or $55mph$ will have more similar number of accidents than a few random roads. Hence, It has a cluster effect since the observations belong to clusters (can be the same set of 100 roads), and observations within a cluster are not independent.

It's likely that samples drawn from 100 roads that are closer to each other will have a similar number of accidents (roads nearer to each other have a similar geographic disadvantage). Hence, It has a spatial effect since the observations from closer locations (closer roads) tend to be more similar and are therefore not independent.

Finally, there's a possibility of a serial effect since the study involves 10 (plus 5) years where a couple of consecutive years can have more accidents due to population increase, traffic, or improper maintenance of roads.

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)
#names(case0202)
#case0202
```

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

Generally, Before/After samples or Twins/Sibling pairs or dependent samples are considered as paired samples. The given data is paired data since the study involves 15 pairs of identical twins that are not independent. There's a high possibility that the twins can share the same genes and traits. Hence, these samples are dependent on each other, and we can use paired t-test.

(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.

```
t.test(case0202$Unaffected, case0202$Affected, alternative = "two.sided", paired = "true")
```

```
##
##  Paired t-test
##
## data:  case0202$Unaffected and case0202$Affected
## 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
```

From the above t-test, the p-value is equal to 0.006062 and thus, we reject the null hypothesis $H_0 : \mu_U - \mu_A = 0$ at the $\alpha = 0.01$ significance level in favor of the alternative $H_A : \mu_U - \mu_A \neq 0$. We have evidence which indicates that the difference in mean left hippocampus volume (Unaffected - Affected) is not equal to zero, on average.

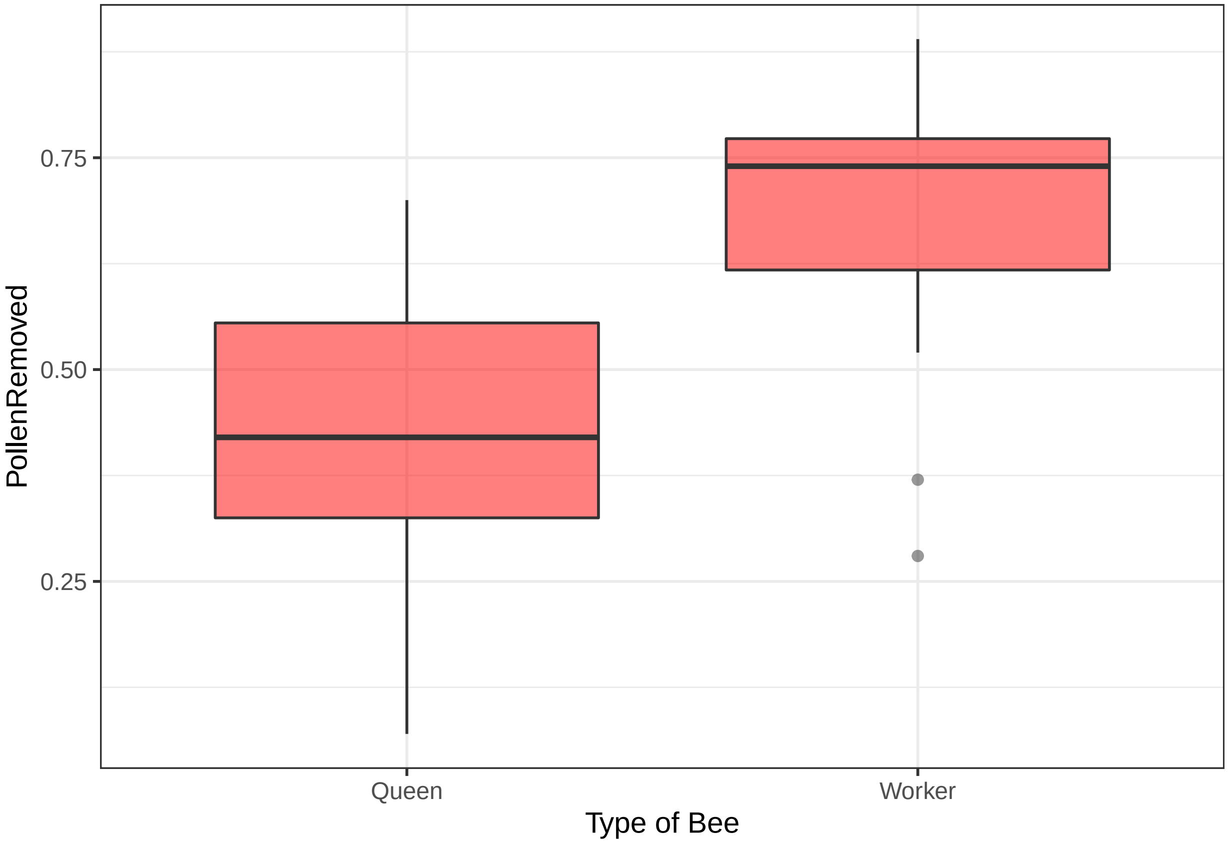
Question 5 (11 points) - Modified from *Sleuth* 3.27(a)

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)
#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?

```
ggplot(data = ex0327, aes(x = as.factor(BeeType), y = PollenRemoved)) +
  geom_boxplot(fill = "red", alpha = 0.5) +
  labs(y = "PollenRemoved", x = "Type of Bee") +
  theme_bw()
```

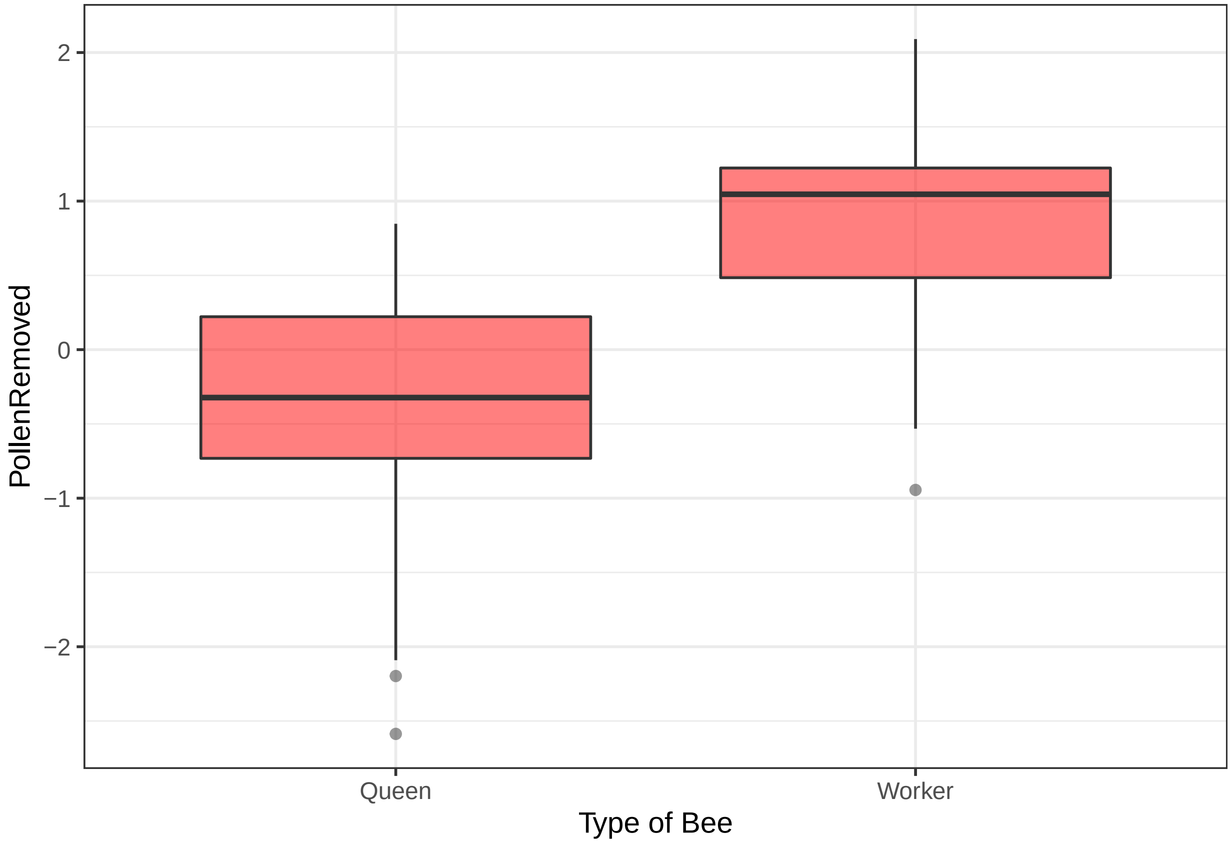


From the above box plot, we have a clear evidence that the distribution of workers in removing the pollen is a skewed distribution whereas the distribution of queen bees is nearly symmetric (normal distribution) with no outliers in it. Therefore, we can consider the transformation of the workers data to make it more normal.

(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.)

```
ex0327$PollenRemoved = log( ex0327$PollenRemoved / (1 - ex0327$PollenRemoved) )

ggplot(data = ex0327, aes(x = as.factor(BeeType), y = PollenRemoved)) +
  geom_boxplot(fill = "red", alpha = 0.5) +
  labs(y = "PollenRemoved", x = "Type of Bee") +
  theme_bw()
```



The transformation seems to have helped us meet the t-test assumptions. The outliers for the workers' distribution are reduced compared to the earlier. Also, with the transformation, the variability of both the samples seems to be reduced (converged) compared to previous such that it is more closer to being symmetric (normally distributed).

(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).

A hypothesis test to examine whether the difference of average of the logit transformed proportion of pollen removed for the two groups ($Queens - Workers$) is equal to $zero$, versus the two-sided alternative.

Null Hypothesis: difference of average of the logit transformed proportion of pollen removed for the two groups ($Queens - Workers$) is equal to $zero$. $H_0 : \mu_Q - \mu_W = 0$ *Alternate Hypothesis*: difference of average of the logit transformed proportion of pollen removed for the two groups ($Queens - Workers$) is not equal to $zero$ $H_A : \mu_Q - \mu_W \neq 0$

where μ_Q is the average of the logit transformed proportion of pollen removed for Queens and μ_W is the average of the logit transformed proportion of pollen removed for Workers.

```r
t.test(PollenRemoved ~ BeeType, data = ex0327, alternative = "two.sided",
       var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  PollenRemoved by BeeType
## t = -3.8493, df = 45, p-value = 0.0003715
## alternative hypothesis: true difference in means between group Queen and group Worker is not equal t
## 95 percent confidence interval:
##  -1.7490870 -0.5474536
## sample estimates:
##  mean in group Queen mean in group Worker
##           -0.3812734            0.7669968
```

The value of t-statistic is -3.8493 and the p-value is equal to 0.0003715 which is less than the given significance level of $\alpha = 0.05$. Therefore, we reject the null hypothesis $H_0 : \mu_Q - \mu_W = 0$ in favor of alternate hypothesis. Thus, we have sufficient evidence that the difference of average of the logit transformed proportion of pollen removed for the two groups ($Queens - Workers$) is not equal to $zero$.

Based on the 95 confidence interval, we are 95% confident that the true mean difference can be in the interval $(-1.7490870, -0.5474536)$ where the average of the logit transformed proportion of pollen removed by the Queens' group is greater than that of Workers' group.

(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?

```r
t.test(PollenRemoved ~ BeeType, data = ex0327, alternative = "two.sided",
       var.equal = TRUE, conf.level = 0.90)
```

```
##
##  Two Sample t-test
##
## data:  PollenRemoved by BeeType
## t = -3.8493, df = 45, p-value = 0.0003715
## alternative hypothesis: true difference in means between group Queen and group Worker is not equal t
## 90 percent confidence interval:
##  -1.649252 -0.647289
## sample estimates:
##  mean in group Queen mean in group Worker
##           -0.3812734            0.7669968
```

```r
#REF: conf.level: https://rc2e.com/generalstatistics#recipe-id123
```

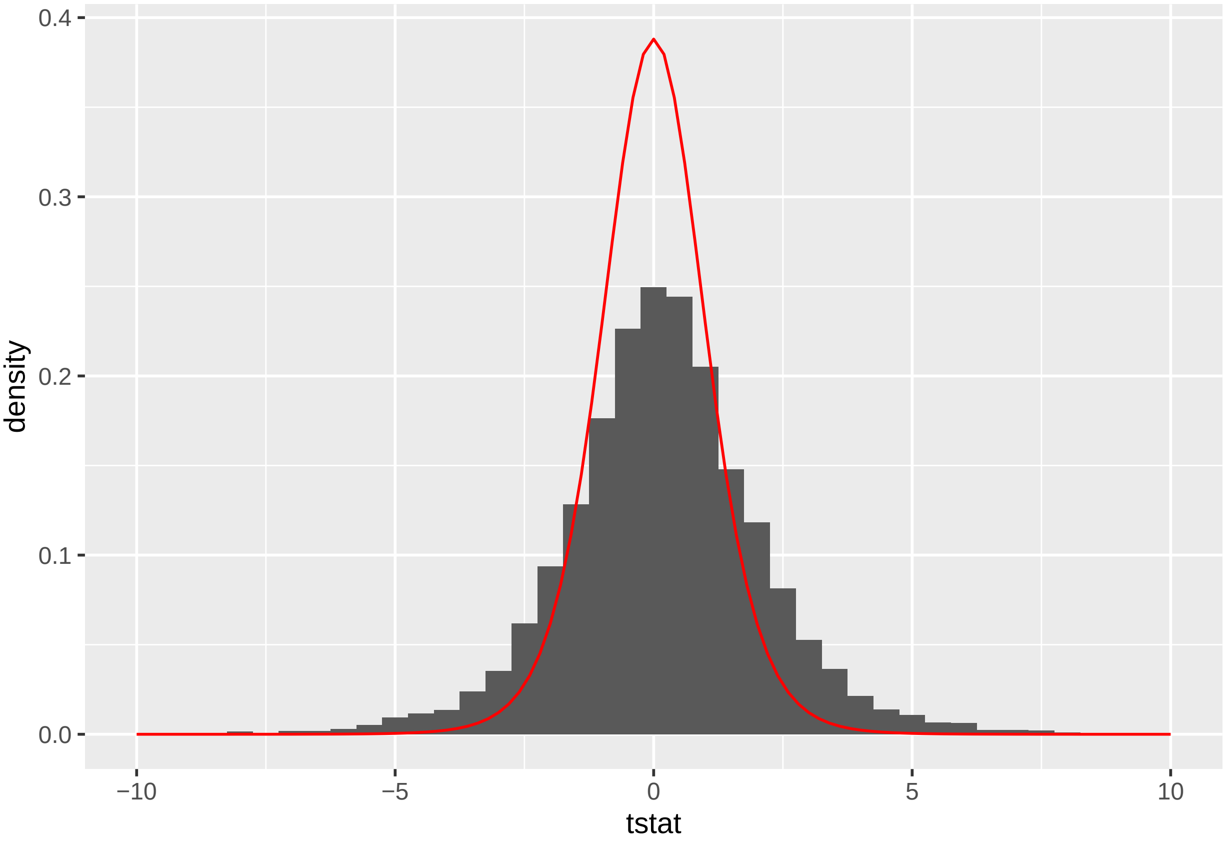
With a 90% confidence interval, the interval gets reduced (narrow), and there's a higher chance of erroneous results than that of a 95% confidence interval since the significance level would be $\alpha = 0.10$ (10% chance of being wrong). Hence, it is more difficult (reduced chance to 90%) to find the true mean difference of the logit proportion of pollen removed with a 90% confidence interval $(-1.649252 - 0.647289)$.

In addition, the 90% confidence interval for the population difference in the mean of the logit proportion of pollen removed may not be the correct representative for the actual population since there's a 10% chance of being wrong, meaning more number of possible values cannot fit in the interval.

Therefore, it is difficult to explain this confidence interval in finding the population difference in the mean.

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").

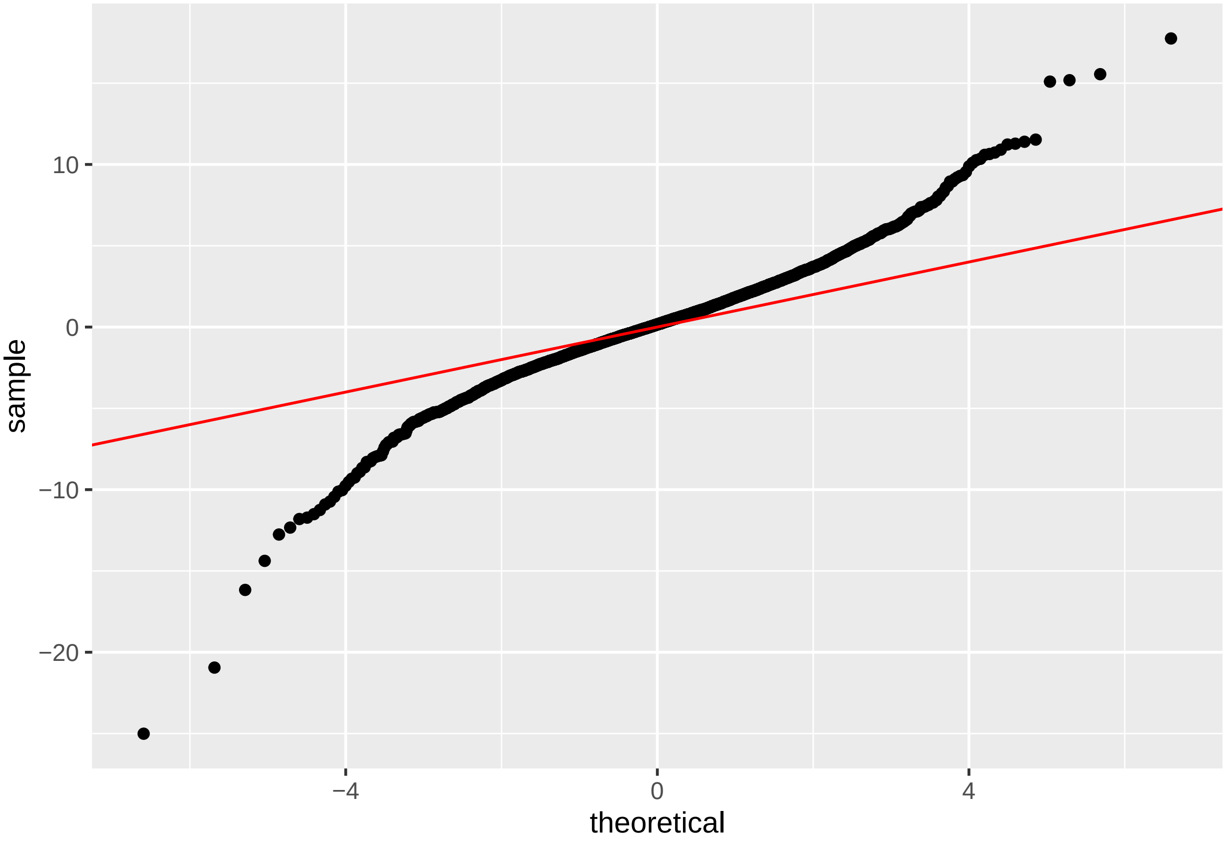


(a) (2 points) Based on what you know about the assumptions of the t-test, the observed distribution of the test statistics (the vertical bars), and the theoretical distribution of the test statistic (the red curve) answer the following questions: (1) Does having duplicated values in our sample violate 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.
```