ST 411/511 Homework 4

Due on February 5

Aashish Adhikari

Winter 2020

Instructions

This assignment is due by 11:59 PM, February 5, 2020 on Canvas via Gradescope. **You should submit your assignment as a PDF which you can compile (should you choose – recommended) using the provide .Rmd (R Markdown) template.** If you opt to not use R Markdown, please format your solutions in a similar manner as provided in this document. Include your code in your solutions and indicate where the solutions for individual problems are located when uploading into Gradescope. You should also use complete, gramatically correct sentences for your solutions.

Problems (25 points total)

Question 1 (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 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.

(4 points) 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.

No, this is not a legitimate way to conduct a statistical analysis. What type of t-test is to be done should be determined by whether the two samples are independent from each other or not unlike the researcher's approach where he/ she has chosen one method over the other after seeing results from both the methods.

We can say for sure that the p value from the independent sample analysis will not be too small. It will be too big since the standard error is too small.

Question 2

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.

(4 points) 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.

Within the roads with the speed limit of 55mph, there is a violation of independence because of the cluster effect. It is possible that the roads with a certain angle of elevation or a history of high accidents or bad asphalt conditions could have been assigned a 55mph limit in the first place. i.e., there are clustered in some group according to some feature they have. To add, the road segments closer to each other have a higher probability of being assigned the same 55mph. Hence, there is also a spatial correlation.

Between the 55mph and 65mph group, there is still a violation of independence. All of the roads under observation for 10 years were given a new speed limit of 65mph meaning that there is some relationship between the two. In other words, knowing that a road had 55mph in the initial population tells us that it will have a 65mph in the other population. On a side note, as James mentioned in class and in his slides, serial observations are the ones that are "close" in time, I would not strictly classify this as a serial correlation case.

Within the roads with 65 mph, since they come from the original population of 55mph roads, they are still not independent within themselves as they have the cluster effect as explained for the roads with 55mph. Similarly, they also have a spatial correlation as with the roads in the 55mph group.

Question 3 (Modified from *Sleuth* 3.27(a))

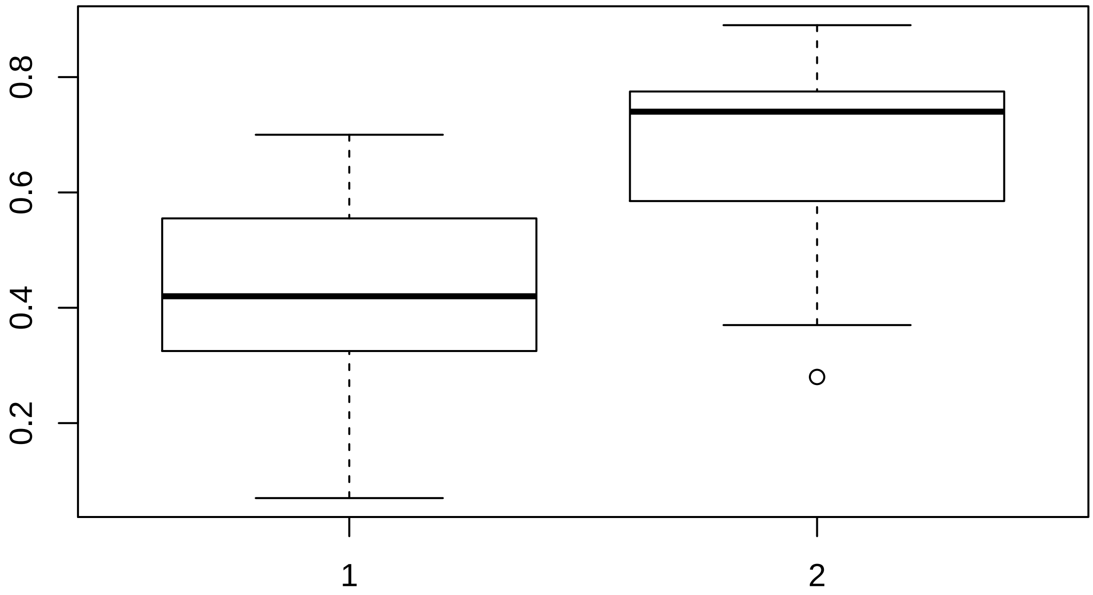
As part of a study to investigate reproductive strategies in plants, biologists recorded the time spent at sources of pollen and the proportions of pollen removed by bumble-bee queens and honeybee workers pollinating a species of lily. These data appear in `ex0327` in the `Sleuth3` package.

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

```r
str(ex0327) #gives out the heirarchy of how the data is organized.
```

```
## 'data.frame':    47 obs. of  3 variables:
##  $ PollenRemoved  : num  0.07 0.1 0.11 0.12 0.15 0.19 0.28 0.31 0.3 0.34 ...
##  $ DurationOfVisit: int  2 5 7 11 12 11 9 9 16 17 ...
##  $ BeeType        : Factor w/ 2 levels "Queen","Worker": 1 1 1 1 1 1 1 1 1 1 ...
```

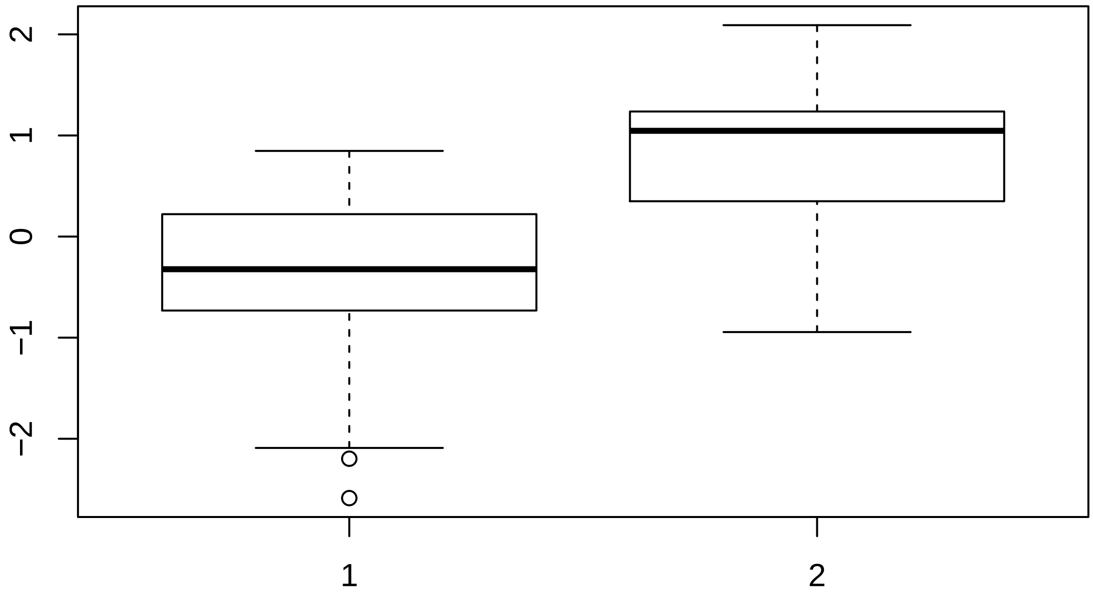
```r
queen_data <-ex0327[ex0327$BeeType=="Queen",]
worker_data <-ex0327[ex0327$BeeType=="Worker",]

boxplot(queen_data$PollenRemoved, worker_data$PollenRemoved)
```



One of the assumptions that needs to be satisfied is that the population should be normal. However, we can see that the worker bees' data is highly-skewed. There are less instances above the median. Also, we can see an outlier. This can could potentially be addressed with a transformation.

(b) (3 points) When the measurement is the proportion P of some amount, one useful transformation is the logit: $\log[P/(1-P)]$. This is the log of the ratio of the proportion removed to the proportion not removed. Draw side-by-side box plots of this transformed data. Does this transformation seem to have helped us meet the t-test assumptions? Note that 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.

```r
queen_data <-ex0327[ex0327$BeeType=="Queen",]
worker_data <-ex0327[ex0327$BeeType=="Worker",]

log_transform <- function(p) { log(p/(1-p)) }


q <- queen_data$PollenRemoved
log_values_q <- log_transform(q)

w <- worker_data$PollenRemoved
log_values_w <- log_transform(w)

boxplot(log_values_q,log_values_w)
```



```
#worker_transformed <- log(worker_data$PollenRemoved)

#boxplot(queen_transformed, worker_transformed)
```

No, this transformation still does not help us meet the requirements of the t-test assumptions. Only the magnitude of values have changed as can be seen above.

(c) (4 points) Test whether the distribution of proportions removed is the same or different for the two groups by using the t-test on the transformed data. You may use the `t.test` function. State your null and alternative hypotheses, the t-statistic and p-value of your test. What do you conclude at significance level $\alpha = 0.05$?

```
d<-ex0327
d$lprop <-log(d$PollenRemoved/(1 - d$PollenRemoved))
d
```

```
##    PollenRemoved DurationOfVisit BeeType       lprop
## 1           0.07               2   Queen -2.58668934
## 2           0.10               5   Queen -2.19722458
## 3           0.11               7   Queen -2.09074110
## 4           0.12              11   Queen -1.99243016
## 5           0.15              12   Queen -1.73460106
## 6           0.19              11   Queen -1.45001018
## 7           0.28               9   Queen -0.94446161
## 8           0.31               9   Queen -0.80011930
## 9           0.30              16   Queen -0.84729786
## 10          0.34              17   Queen -0.66329422
## 11          0.35              12   Queen -0.61903921
## 12          0.39              14   Queen -0.44731222
## 13          0.38              23   Queen -0.48954823
## 14          0.40              35   Queen -0.40546511
## 15          0.42              21   Queen -0.32277339
## 16          0.40              10   Queen -0.40546511
## 17          0.41               9   Queen -0.36396538
## 18          0.42                 7   Queen -0.32277339
## 19          0.48                11   Queen -0.08004271
## 20          0.48                13   Queen -0.08004271
## 21          0.47                14   Queen -0.12014431
## 22          0.49                16   Queen -0.04000533
## 23          0.50                14   Queen  0.00000000
## 24          0.51                17   Queen  0.04000533
## 25          0.53                22   Queen  0.12014431
## 26          0.58                13   Queen  0.32277339
## 27          0.59                13   Queen  0.36396538
## 28          0.65                12   Queen  0.61903921
## 29          0.60                19   Queen  0.40546511
## 30          0.60                23   Queen  0.40546511
## 31          0.69                21   Queen  0.80011930
## 32          0.70                27   Queen  0.84729786
## 33          0.70                28   Queen  0.84729786
## 34          0.51                58   Queen  0.04000533
## 35          0.70                15   Queen  0.84729786
## 36          0.28                 3  Worker -0.94446161
## 37          0.37                12  Worker -0.53221681
## 38          0.52                10  Worker  0.08004271
## 39          0.65                17  Worker  0.61903921
## 40          0.76                24  Worker  1.15267951
## 41          0.89                33  Worker  2.09074110
## 42          0.74                44  Worker  1.04596856
## 43          0.70                46  Worker  0.84729786
## 44          0.79                48  Worker  1.32492541
## 45          0.78                51  Worker  1.26566637
## 46          0.74                64  Worker  1.04596856
## 47          0.77                78  Worker  1.20831121
```

```
Queens<-d[d$BeeType=="Queen",]
Workers<-d[d$BeeType=="Worker",]

t.test(Queens$lprop,Workers$lprop,alternative = "two.sided", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Queens$lprop and Workers$lprop
## 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
```

Null Hypothesis: The true difference in means of the logit transformation of the proportions of the pollens removed is equal to 0. Alternative Hypothesis: The true difference in means of the logit transformation of the proportions of the pollens removed is not equal to 0. The t-statistic is -3.849. The p-value is 0.0003715 The p value obtained is less than the significance level hinting that the value lies in the rejection region, we reject the null hypothesis.

(d) (2 points) Construct a 90% confidence interval for the population difference in the mean of the logit pollen removed between the two bee groups. What is one possible issue with presenting this confidence interval?

```
t.test(Queens$lprop,Workers$lprop, conf.level=0.90, alternative = "two.sided", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Queens$lprop and Workers$lprop
## 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
```

One possible issue with it is that it does not encompass all the plausible values.

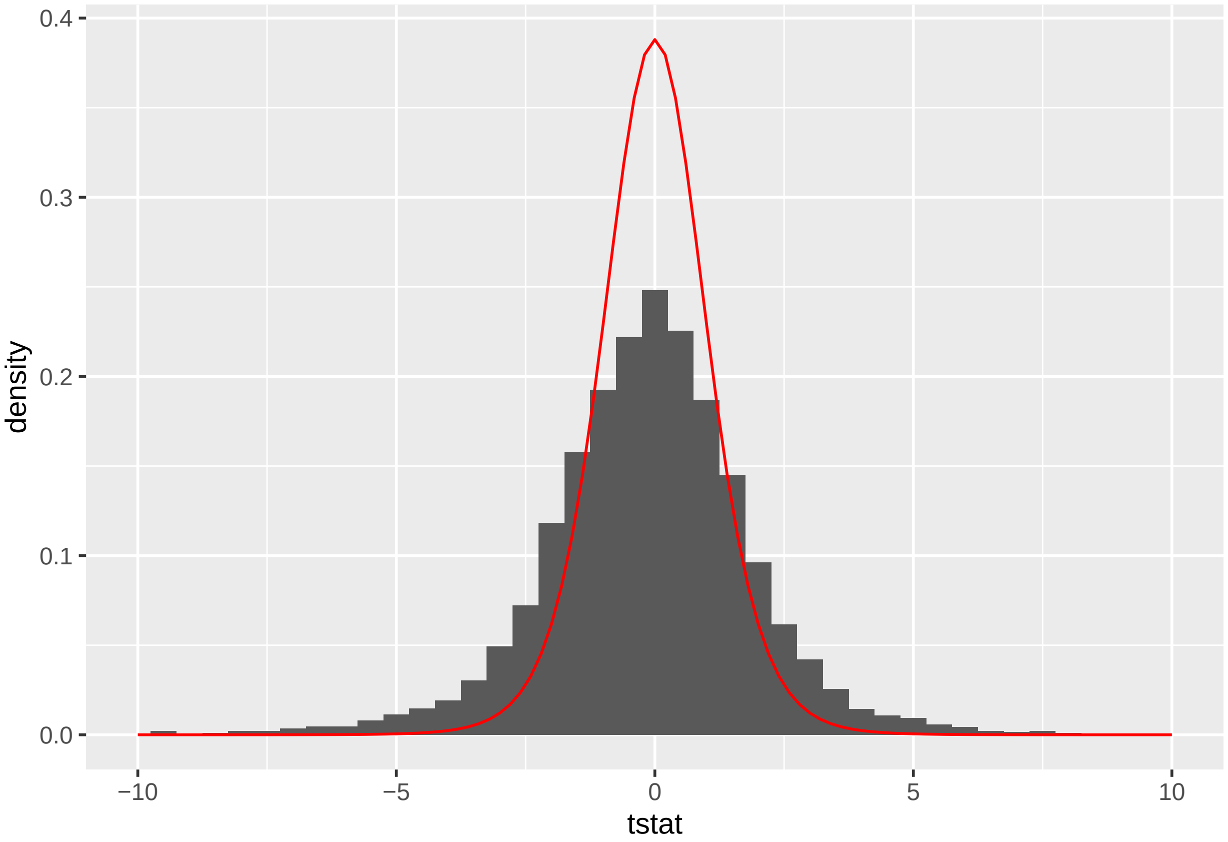
Question 4

Suppose you have a normally distributed population with mean 60 and variance 25. Consider drawing samples of size 5 that accidentally get duplicated such that there are 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 $t_{(9)}$ distribution. (Note: You can ignore warnings about "removed rows containing non-finite values/missing values").

```
pop <- rnorm(1000, mean=60, sd=5)

tstat <- vector(length = 10)
set.seed(411511)
for (i in 1:10000) {
  samp <- rep(sample(pop, size=5), each = 2)
  tstat[i] <- (mean(samp) - 60) / sqrt(var(samp) / 10)
}

df <- data.frame(tstat)

ggplot(df, aes(x = tstat)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.5) +
  stat_function(fun = dt, args = list(df = 9), color = "red") +
  scale_x_continuous(limits=c(-10, 10))
```

```
## Warning: Removed 25 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

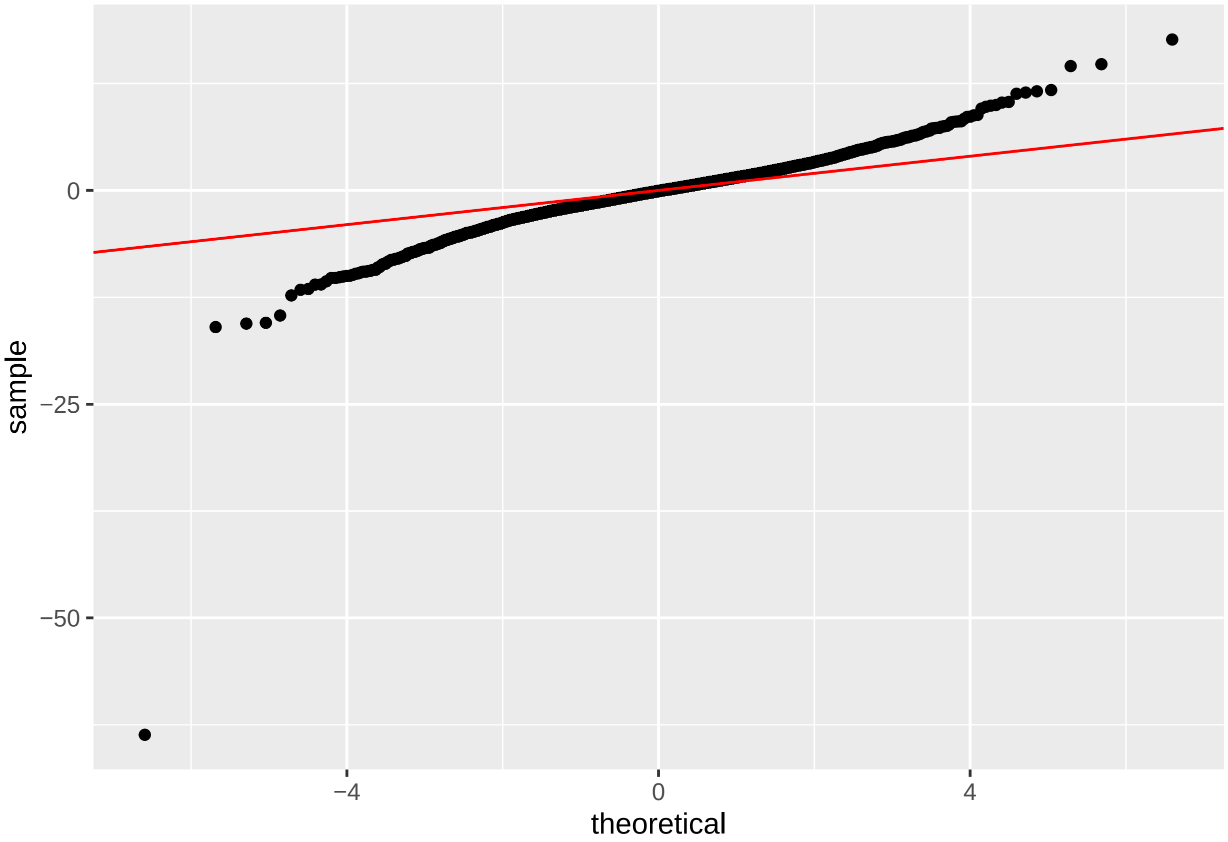


(a) (2 points) Does this plot indicate that duplication violates any of our t-test assumptions? If so, which one(s)? Discuss how the histogram helps you make this conclusion.

Since there are duplicates for each value, it violates the assumption that each observation is independent. This is supported by the histogram because the distribution of the test statistic does not follow the t distribution for a degree of freedom = 9 and has lower tails.

(b) (2 points) Use the following code to produce a quantile-quantile plot for this test statistic. Does this plot indicate that duplication violates any of our t-test assumptions? If so, which one(s)? Discuss how the plot helps you make this conclusion.

```
ggplot(df, aes(sample = tstat)) +
  stat_qq(distribution = qt, dparams = 9) +
  geom_abline(slope = 1, intercept = 0, color="red")
```



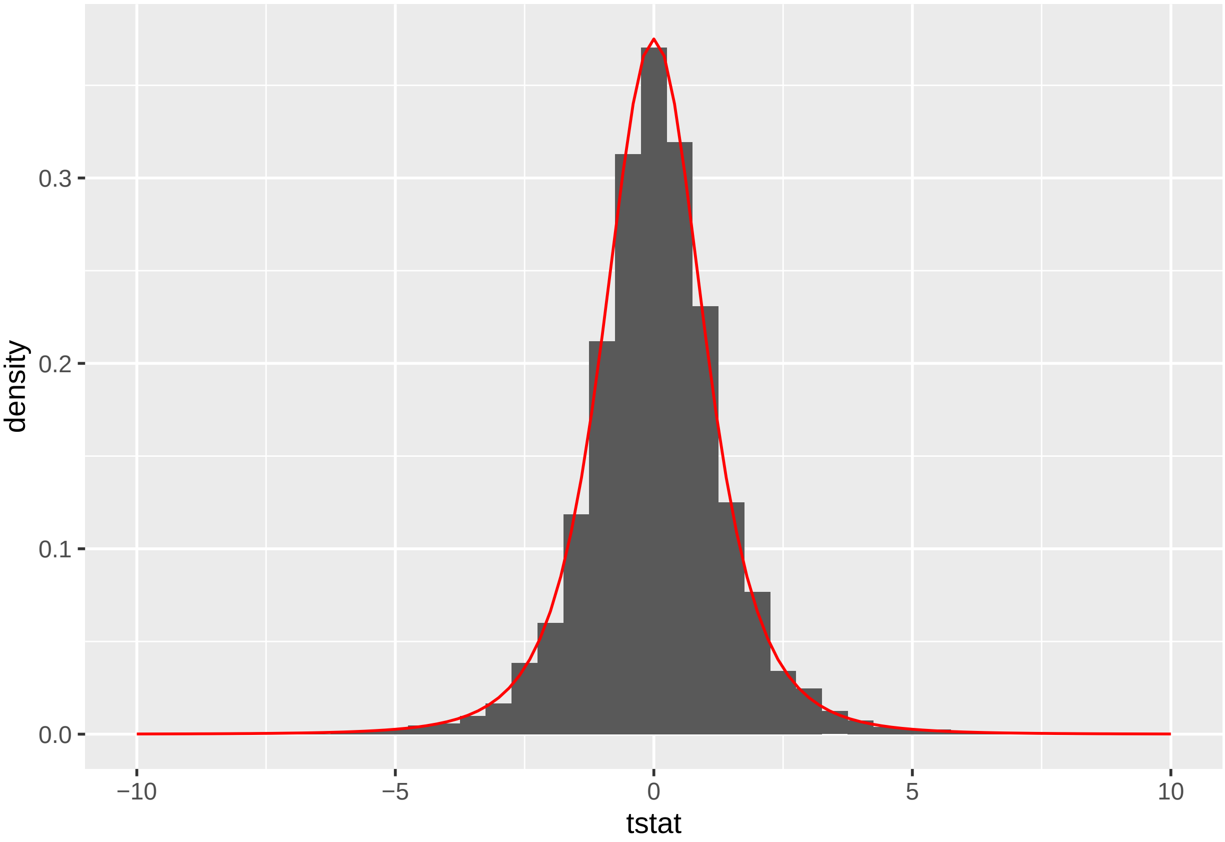
Yes, this quantile-quantile plot indicates that duplication violates our t-test assumption of independence of observation. The values should have fallen on the line if the observed values had come from the t distribution with a degree of freedom = 9. i.e., theoretical quantiles should have matched the observed quantiles.

(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 `sample` to get rid of the duplicated values. Now the samples should be of size 5 with no duplication. Create a histogram of the distribution of the test statistic with the appropriate null distribution superimposed. Do the t-test assumptions appear to be better met in this case? Note: You will also need to modify n and the degrees of freedom in the code.

```
pop <- rnorm(1000, mean=60, sd=5)

tstat <- vector(length = 10)
set.seed(411511)
for (i in 1:10000) {
  samp <- sample(pop, size=5)
  tstat[i] <- (mean(samp) - 60) / sqrt(var(samp) / 5)
}

df <- data.frame(tstat)

ggplot(df, aes(x = tstat)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.5) +
  stat_function(fun = dt, args = list(df = 4), color = "red") +
  scale_x_continuous(limits=c(-10, 10))
```

```
## Warning: Removed 6 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



Yes, it appears that with this modification, the sample distribution aligns with the null hypothesis meaning that the assumptions have been fairly met. The sample distribution can be seen to conform with the corresponding t distribution above.