ST 411/511 Homework 4

Due on February 6

Winter 2019

Instructions

You should submit your assignment as either a PDF or Word document, which you can compile (should you choose – recommended) from the provided .Rmd (R Markdown) template. Please include your code.

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 legitimate? Discuss whether the p-value from the independent sample t-analysis will be too big, too small, or about right.

It is not legitimate. Since we know that we have to use the paired analysis,we should also know that the inferences may not appear to be as precise. The unpaired analysis is inappropriate.Also, the P-value from the unpaired test/ independent will be smaller in size than the p-value from the paired t-test because the standard error will be smaller and degrees of freedom is higher. Hence we notice that the t- statistic will be far from 0 and it may reject a hypothesis even if though is correct.

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.

Here we see that in each sample the road taken is same and it is considered more than one times. It violates the independence assumption within groups. Clustering means dividing the series into homogeneous groups. Similar elements will appear in similar in groups. So here the independence assumption is violated in relation to a cluster effect. Since we took consecutive years, There may also be serial correlation as the consecutive years may be having same conditions(with no much variations).

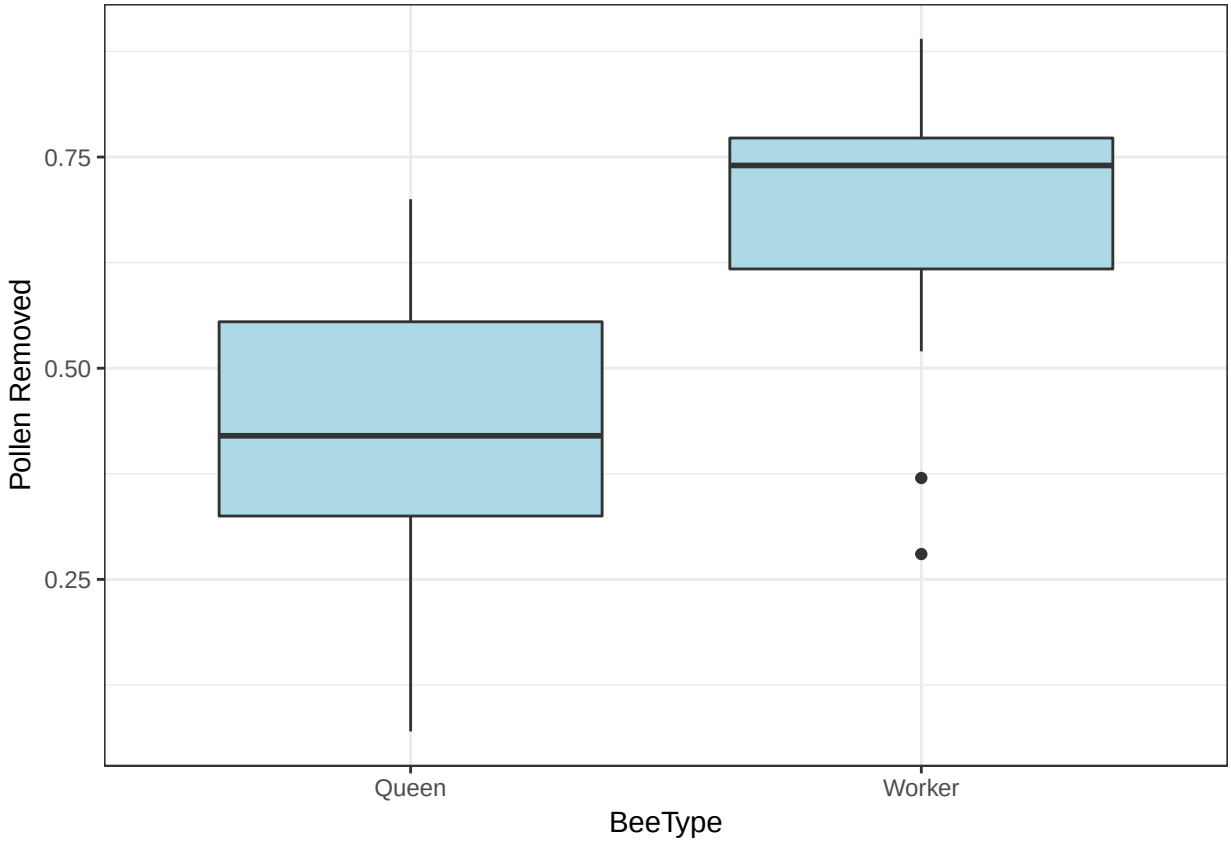
Question 3 (Modified from *Sleuth* 3.27(a))

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.

```
data(ex0327)
```

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

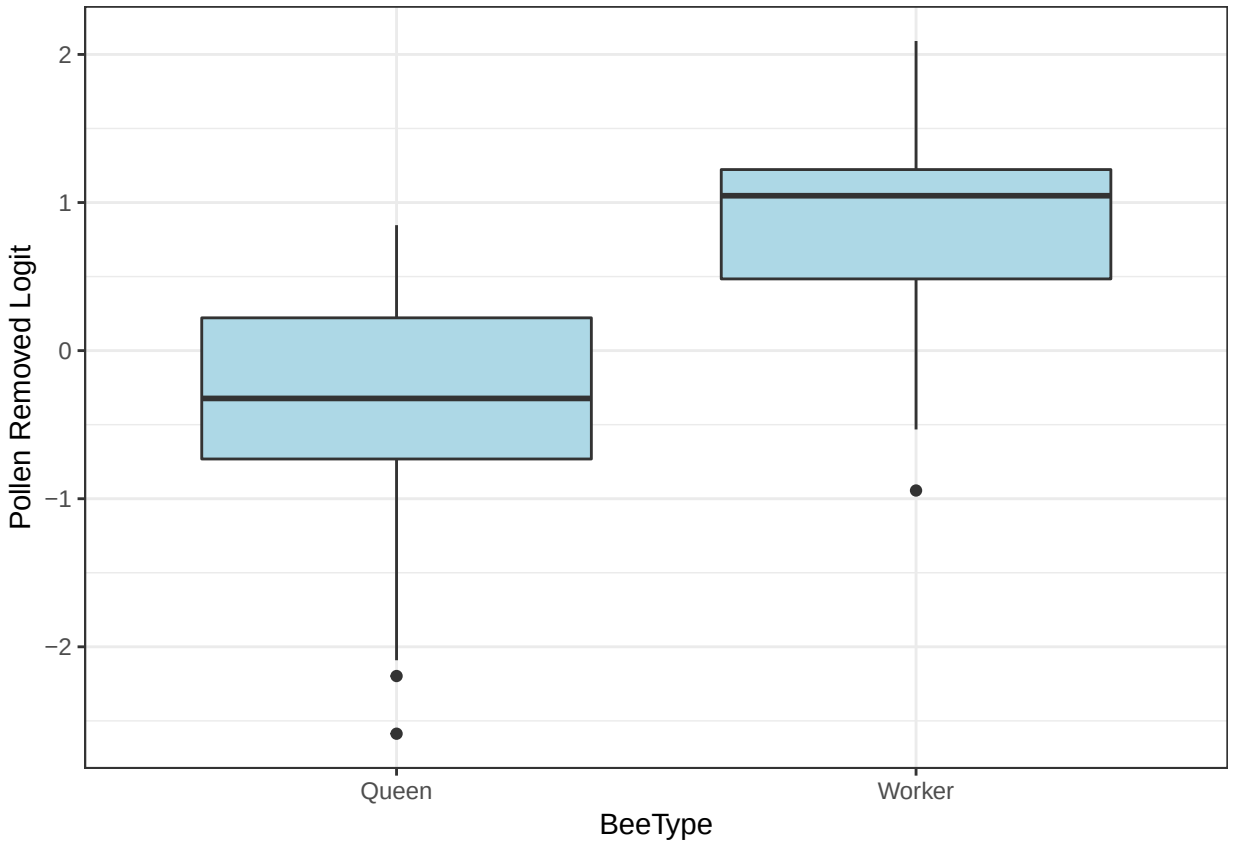
```
ggplot(data = ex0327, aes(x = BeeType, y = PollenRemoved)) + geom_boxplot (fill = "lightblue") + ylab("
theme_bw()
```



In the above plot we notice that the distribution for the worker bees looks assymetric (skewed) with the median not very much at the center. So we can do a transformation here to make it symmetric. We can also see that the distribution of queen bee looks nearly symmetric with median almost at center and this does not require 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)`.

```
ex0327$PollenRemovedLogit=log(ex0327$PollenRemoved/ (1-ex0327$PollenRemoved) ) #View(ex0327)
ggplot(data = ex0327, aes(x = BeeType, y = PollenRemovedLogit)) + geom_boxplot (fill = "lightblue") + yl
theme_bw()
```



We notice that the worker distribution after transformation is little more symmetric than first version but still assymetric. Here the number of outliers before transformation is 2 and after is 1.Since the worker bee distribution is assymetric and mdian is towards the top, we can say that the log transformation wouldnt have helped.

(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, and the t-statistic and p-value of your test. What do you conclude at significance level $\alpha = 0.05$?

```
t.test (PollenRemovedLogit ~ BeeType, data = ex0327, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  PollenRemovedLogit by BeeType
```

```
## 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 in group Queen mean in group Worker
##           -0.3812734            0.7669968
```

The Null Hypothesis is : Ho : uqueenlog — pworkerlog = 0 , i.e) the difference between the transformed (logit) means of the pollens removed by the mean Queen bees and Worker bees is 0. The Alternative Hypothesis : HA : uqueenlog — pworkerlog! = 0 i.e) True difference in means is not equal to 0. t statistics t = -3.8493 p-value = 0.0003715 Also the p value is less than alpha. So we reject the Null hypothesis that the difference in means of logs of the pollens removed by the Queen and Worker bees is 0 in favor of the alternative hypothesis at alpha = 0.05 since our p value is less than alpha.

(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 (PollenRemovedLogit ~ BeeType, data = ex0327, conf.level = .9, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  PollenRemovedLogit by BeeType
## 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 in group Queen mean in group Worker
##           -0.3812734            0.7669968
```

From the above we can observe that, 90% confidence interval for the population difference in the mean of the logit pollen removed between the logit Queen bees logit Worker bees is [-1.649252 -0.647289]. We do it in a such that : instead of taking the difference in means , we are taking the log values. Hence, the values in confidence interval are not the actual data collected so it can not be very easy to comprehend.

Question 4

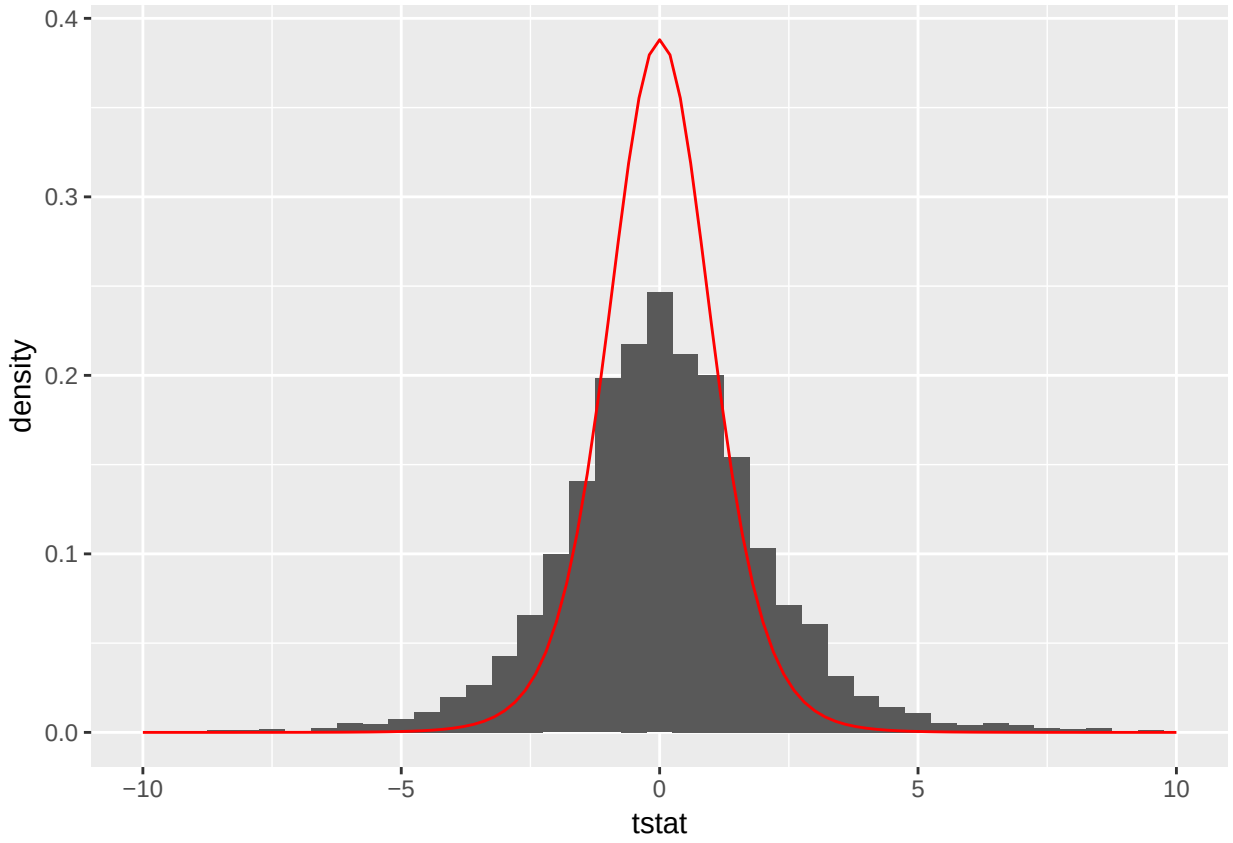
Suppose you have a Normal population with mean 60. Consider drawing samples of size 5 that accidentally get duplicated such that there are 10 observations in the sample, with each unique value occurring twice. Use the following code to produce a histogram of distribution of the test statistic for a one-sample t-test of $H_0 : \mu = 60$. The superimposed red curve is the $t_{(9)}$ distribution.

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

tstat <- c()
set.seed(411511)
for (i in 1:10000) {
  samp <- rep(sample(pop, size=5), 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 23 rows containing non-finite values (stat_bin).
```

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



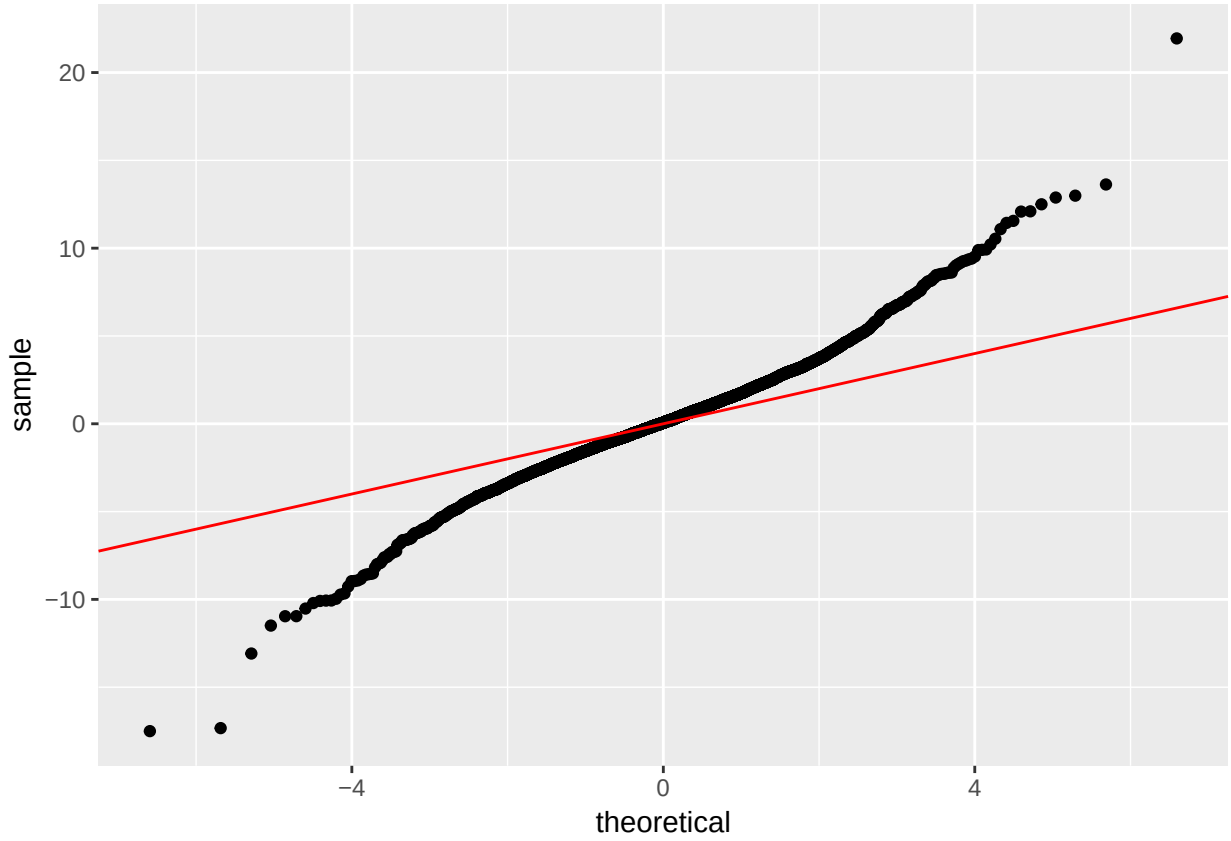
(a) (2 points) Does this plot indicate that duplication violates any of our t-test assumptions? If so, which one(s)? Discuss how the histogram helps you make this conclusion.

We notice that the duplication violates the t-test assumptions of independence of observations. The distribution does not fit completely in the t(9) distribution shown in red curve(Null distribution). In the tail area the observtion is more than expected and less than expected at peak .t(9) distribution does not have enough area in tails cover observed distribution.

(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, the above plot indicates the independence of observation assumption is violated.The t(9) does not have enough area in the tails so there is variations on distribution from red line(null distribution). Variation is more towards the ends. The observed points are not properly fitted in the Null distribution.

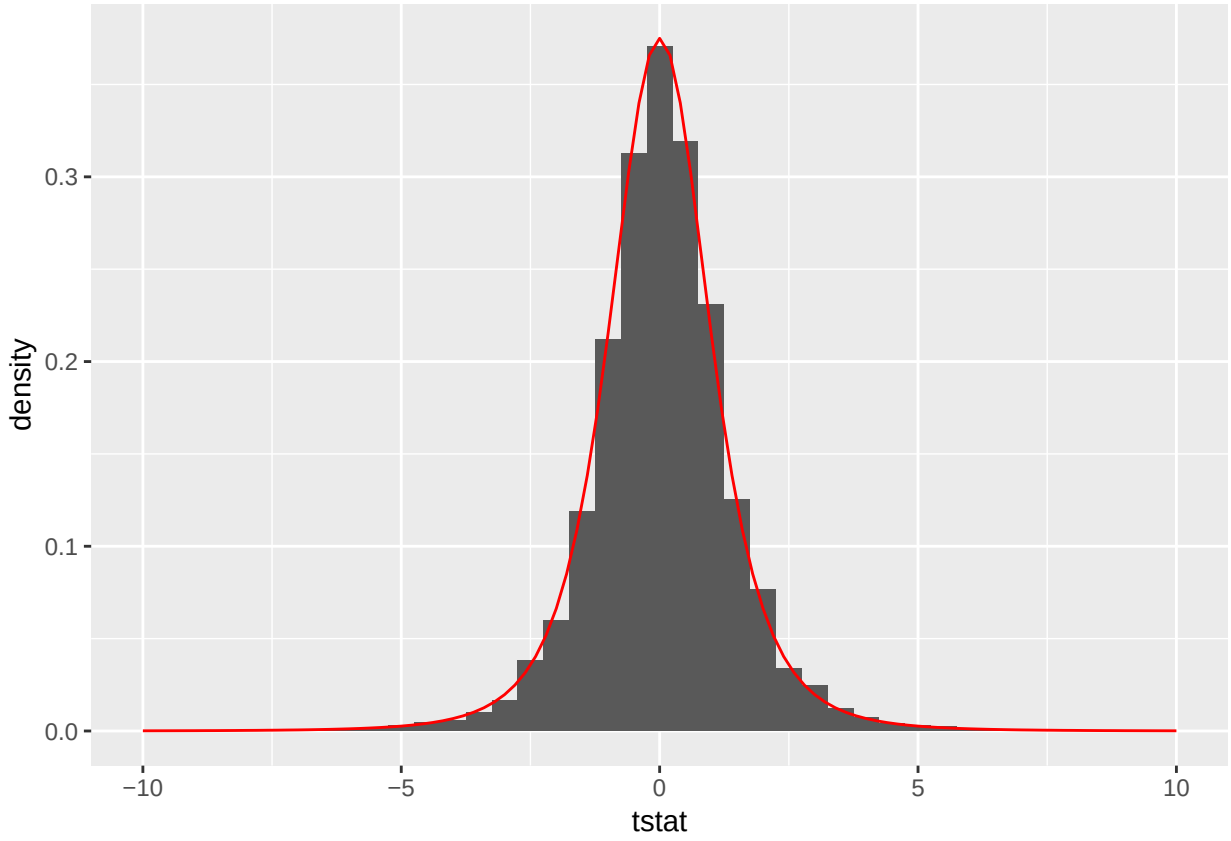
(c) (2 points) Alter the code I provided by removing the `rep()` function wrapped around `sample` to get rid of this duplication. Now the samples should be of size 5 with no dublication. Create a histogram of the distribution of the test statistic with the appropriate null distribution superimposed. 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 <- c()
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 we can now observe that the distribution looks similar to the null distribution and also the t-test assumptions appear to be better met here in this case.