ST 411/511 Homework 4

Chimdi Chikezie

Winter 2022

Instructions

This assignment is due by 11:59 PM, February 4th 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. Think critically about study designs and look for instances of "poor design" or underlying issues which may impact analyses.
2. Extend our t-based methods to see how they can be adapted for "paired" samples.
3. See how useful (or useless?) data transformations are to answering questions of interest with statistical methods.

Question 1 (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 isn't a legitimate way to conduct this particular statistical analysis. From the question, we are working with cultures from cultures gotten from one subject which means that the divided samples have a relationship with one another.

The p-value will be small because the independent two sample t-analysis has more degrees of freedom.

Question 2 (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 55 mph and 65 mph groups. These roads have common relationship. These roads probably are found in the same area that have same weather that affects the roads. This leads to violation of independence as we can know the beahavior of one road from another.

Question 3 (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 a paired data because the twins are from the same family and have traits that affect them.

(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", mu = 0, paired = TRUE, conf.l
```

```
##
##  Paired t-test
##
## data:  case0202$Unaffected and case0202$Affected
## t = 3.2289, df = 14, p-value = 0.006062
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  0.01551009 0.38182325
## sample estimates:
## mean of the differences
##               0.1986667
```

We reject the null hypothesis because the p-value is less than alpha. This means that the difference in mean left hippocampus volume (Unaffected - Affected) is not equal to zero.

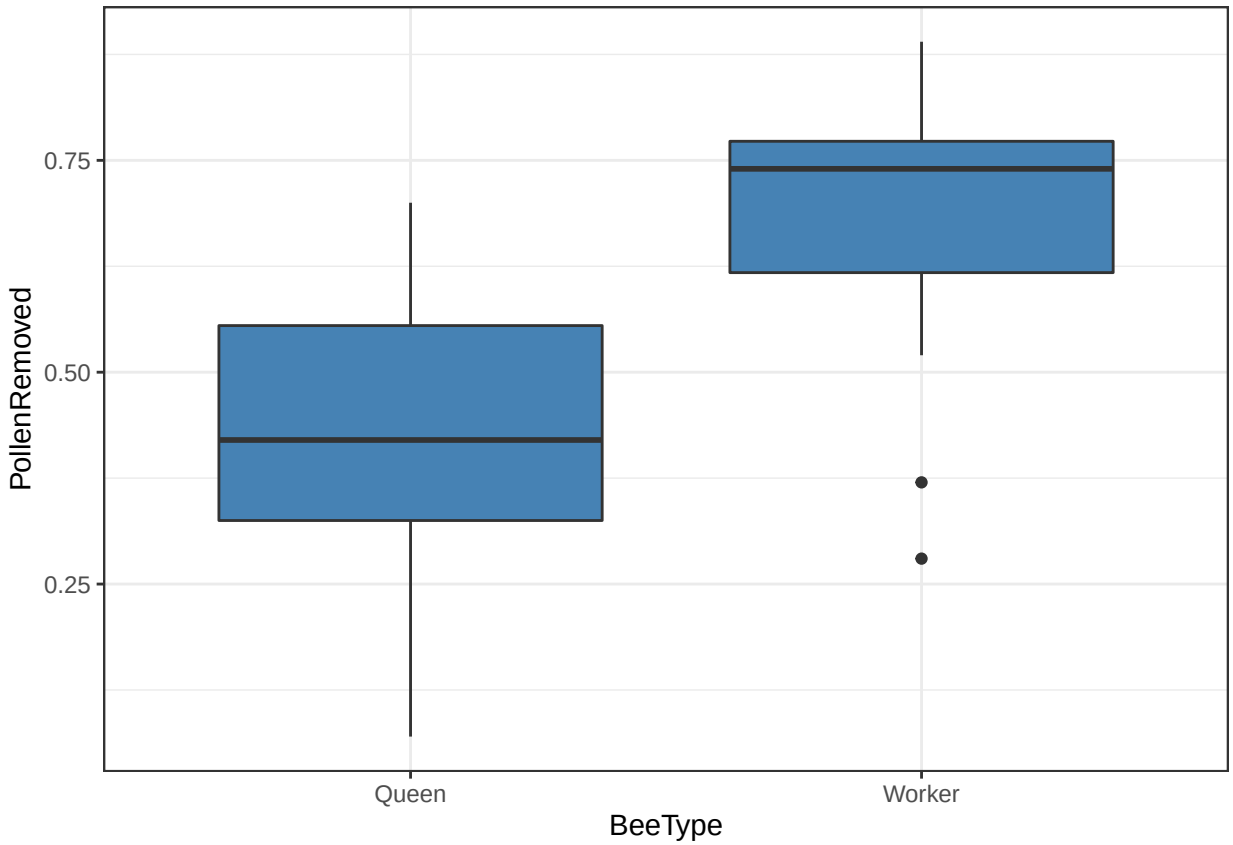
Question 4 (11 points) - Modified from *Sleuth* 3.27(a)

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?

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

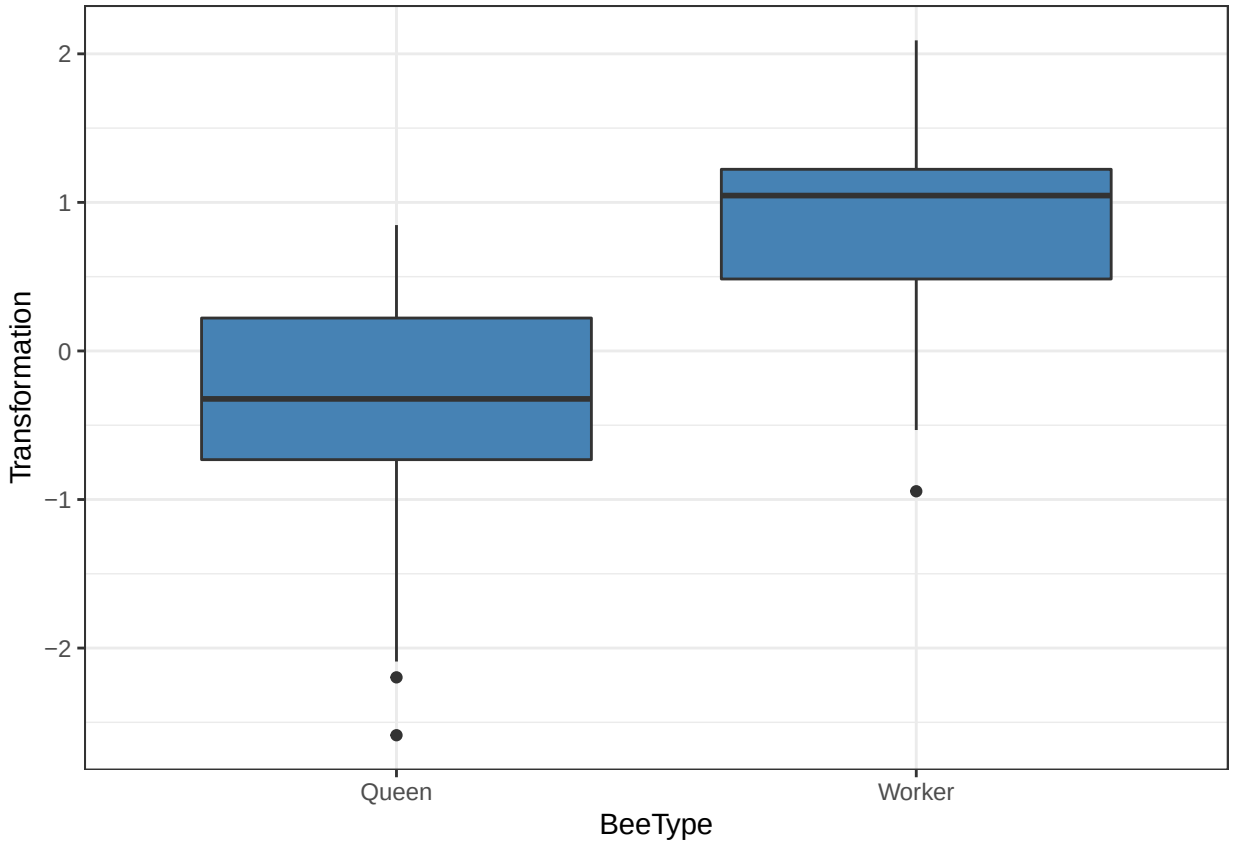


There are outliers from the worker boxplot which will call for transformation.

(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$Transformation <- log(ex0327$PollenRemoved/(1-ex0327$PollenRemoved))

ggplot(data = ex0327, aes(x = BeeType, y = Transformation)) +
  geom_boxplot(fill = "steelblue") +
  theme_bw()
```



No, this did not help us inteprete the values. It two samples now have large spread just like in the initial graph worker value in the initial graph.

(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).

$H_0 : \mu_D = 0$ The logit transformed proportion of pollen removed is the same for the 2 groups. $H_A : \mu_D \neq 0$ The logit transformed proportion of pollen removed is not the same for the 2 groups.

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

```
##
##  Two Sample t-test
##
## data:  Transformation 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
```

We reject the null hypothesis at 0.05 significance level. test statistics:-3.8493 P-value:0.0003715 At 95% confidence interval, the difference in mean is between -0.3812734 & 0.7669968 We do not have enough evidence that the logit transformed proportion of pollen removed is the same for the 2 groups.

(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(Transformation~BeeType, data = ex0327, var.equal = TRUE, alternative = "two.sided", conf.level =
```

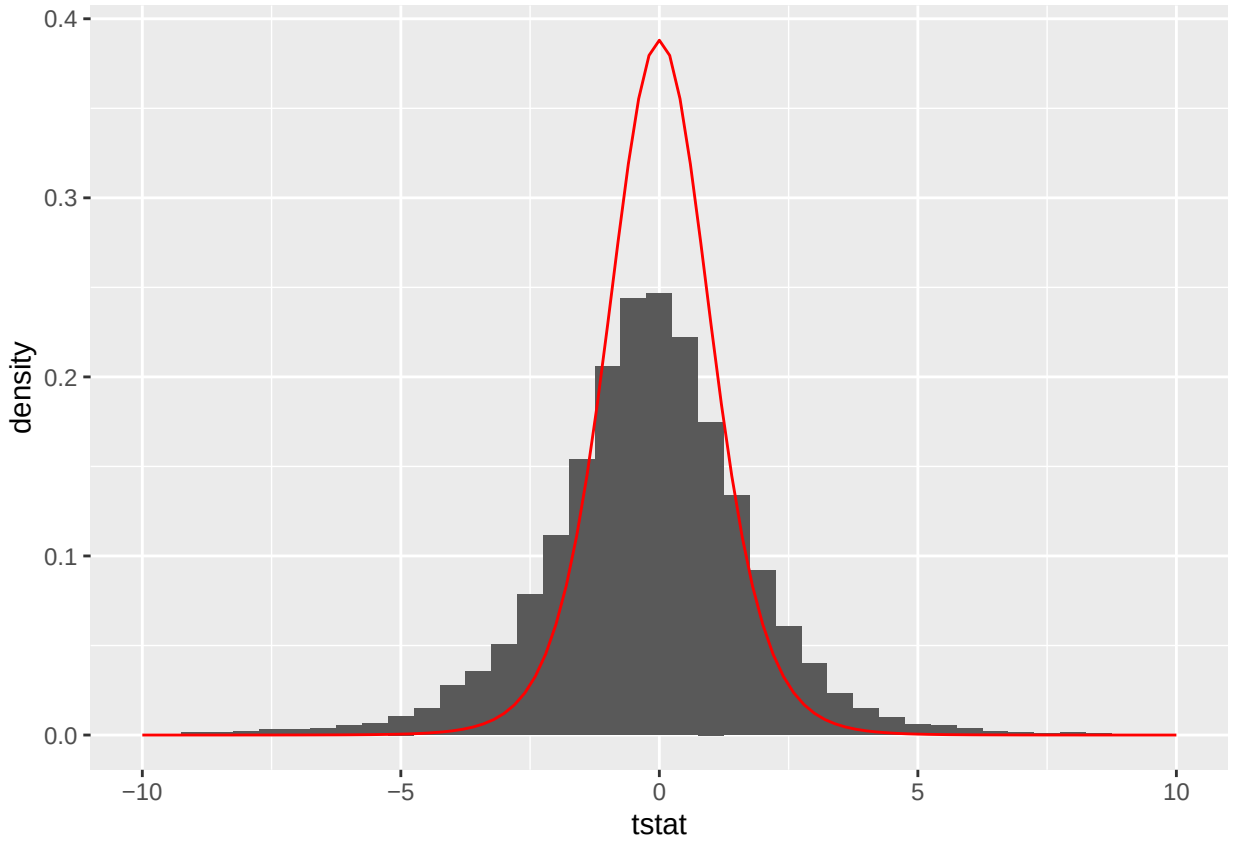
```
##
##  Two Sample t-test
##
## data:  Transformation by BeeType
## t = -3.8493, df = 45, p-value = 0.0003715
## alternative hypothesis: true difference in means between group Queen and group Worker is not equal t
## 90 percent confidence interval:
##  -1.649252 -0.647289
## sample estimates:
## 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

The values calculated are for the Logic Transforming and does not give us information about our original values and how they are changing.

It is difficult to interprete the values to someone who isn't versed in statistics.

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").

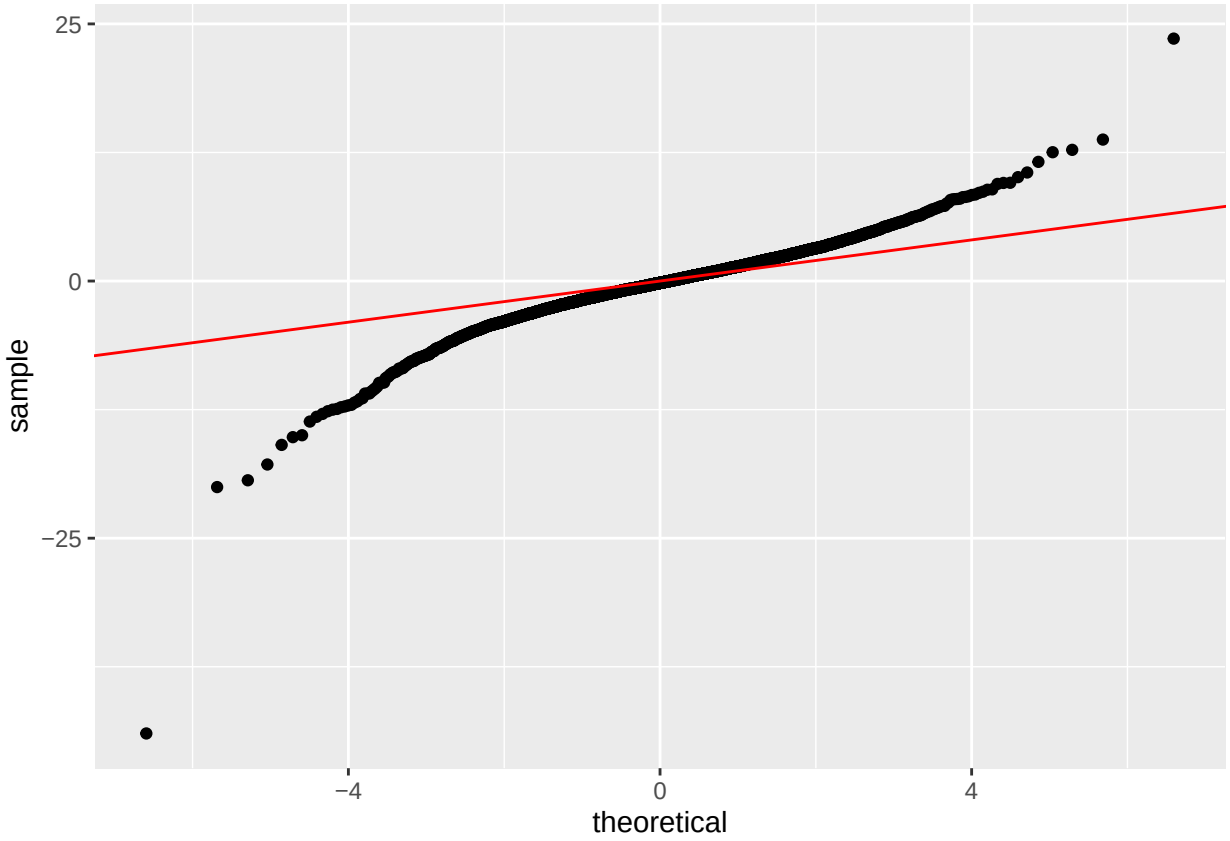


(a) (2 points) Based on what you know about the assumptions of the t-test, the observed distribution of the test statistics (the vertical bars), and the theoretical distribution of the test statistic (the red curve) answer the following questions: (1) Does having duplicated values in our sample violate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