Sam Tenney

Homework 4

Section 2

1. **Assumptions.**
   i. Error is independent identically distributed – Use an index plot to check residuals to see if there is a pattern or not. If there is a pattern, then sample is not independent.
   ii. Normally distributed – Look at a normal QQ Plot to compare the observed quantiles to the quantiles in theoretical normal distribution. If the QQ plot follows a straight line, it is normal.
   iii. Does the error have a mean of zero? To find out, we can check to see if the residuals have a mean of zero.
   iv. Check to see if the errors have a constant variance – see if the ratio of the sample standard deviations is less than 2. If so, then the variances are likely constant.

2. **Glyburide versus Insulin**
   a) The plot shown in the homework is checking to see if the error is independent. Since no real pattern is shown in the plot, we can assume the sample data is independent.
   b) Because the mean of the residuals is so small, it's practically zero. We can assume that the error has a mean of zero.
   c) The normal QQ plot is checking to see if the sample is normally distributed. Because the sample plots closely follow a straight line, we can assume that the sample is normally distributed.
   d) Glyburide percent fat mass standard deviation = 5.7
   Insulin percent fat mass standard deviation = 4.2
   5.7 / 4.2 = 1.357
   1.357 < 2 so we can assume that the error has a constant variance.
   e) We want to check four main assumptions to perform our two-sample t-test on our data concerning the effect on neonate percent fat mass when their mother receives glyburide versus insulin. After performing all our checks of our assumptions above, we can move forward with our two-sample t-test since all of our conditions are met.

3. **The Art of Oreo Dunking**

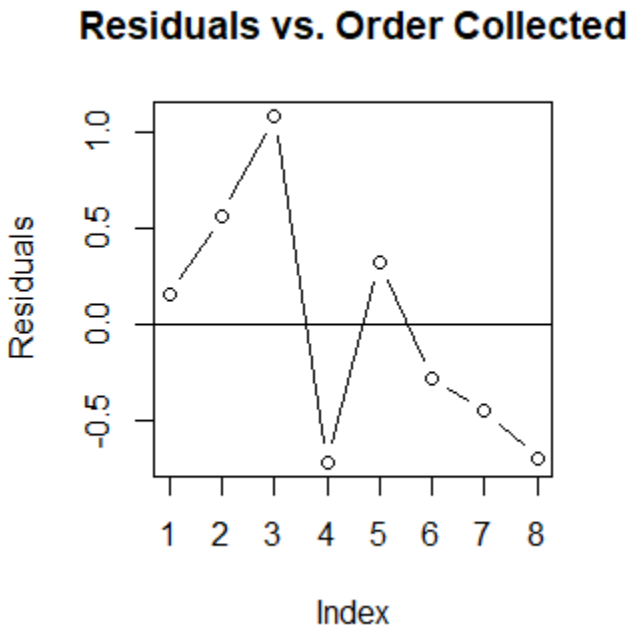
```
a) # Homework 4

b) oreosSUB$resids <- resid(aov(milkabsgram~treatment, data=oreosSUB)
   )
   mean(oreosSUB$resids)

   ## [1] -3.469447e-17
```

The assumption being checked is if the error mean is equal to zero. Since our calculated residual mean of -3.469447e-17 is so small, we can assume that the

error has a mean of zero. The standard deviation of the residuals is calculated below.

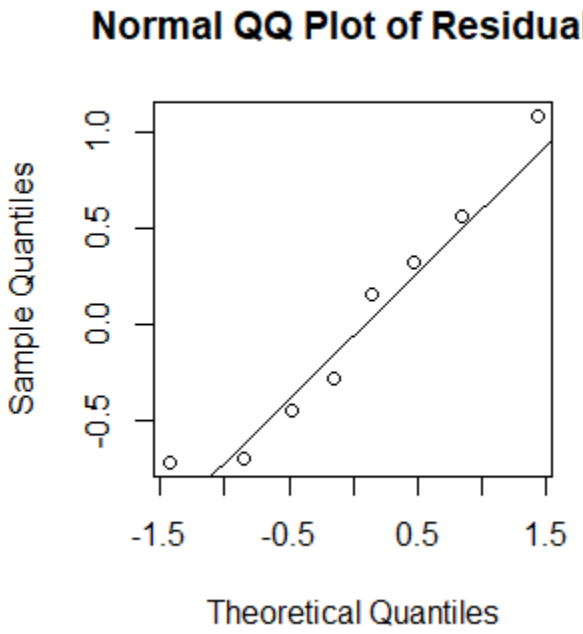
```
sd(oreosSUB$resids)

## [1] 0.6398549
```

```
c) plot(oreosSUB$resids, type = "b", main = "Residuals vs. Order Coll
   ected", ylab = "Residuals")
   abline(h = 0)
```



The plot "Residuals vs. Order Collected" is checking for whether the errors are independent or not. Since there is no distinct pattern in the data, we can assume that the errors are independent.

d)
```
qqnorm(oreosSUB$resids, main = "Normal QQ Plot of Residuals")
qqline(oreosSUB$resids)
```



The plot "Normal QQ Plot of Residuals" is checking whether our errors are normally distributed or not. Since the data is in a relatively straight line, we can assume that our errors are normally distributed.

e)
```
# Standard deviation for Regular  Oreos
Regsd <- sd(oreos$milkabsgram[oreos$treatment == "Reg"])
Regsd
## [1] 0.8658329

# Standard deviation for Trader Joe's Oreos
TJsd <- sd(oreos$milkabsgram[oreos$treatment == "TJ"])
TJsd
## [1] 0.4534681

# Ratio to check if variance is constant
Regsd / TJsd

## [1] 1.909358
```

The ratio of the standard deviations of the two kinds of cookies helps us to know if the errors have a constant variance. 1.909358 is less than 2 so we can assume our errors have a constant variance.

f) We want to know if our errors are independent, are normally distributed, have a mean of zero, and have a constant variance. After checking all our assumptions above, we can verify that all the conditions have been met. We may proceed with our two-sample t-test.

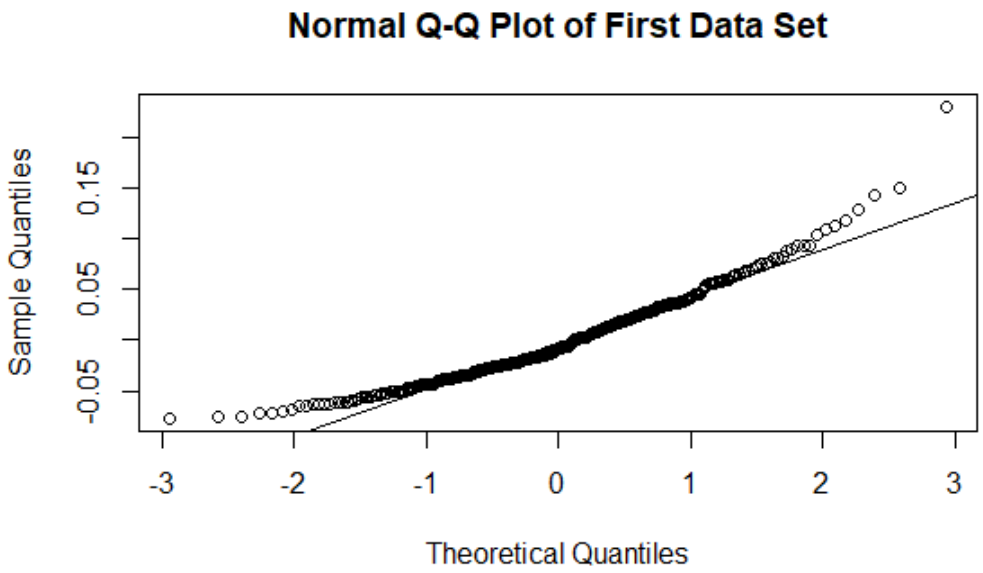
4. **More QQ plot practice**

a)

b) Below are the histogram and normal QQ plots of first set of data.

i.
```
# Histogram of qqdata
hist(qqdata$x, xlab = "Data Values", main = "Histogram of Firs
t Data Set")
```

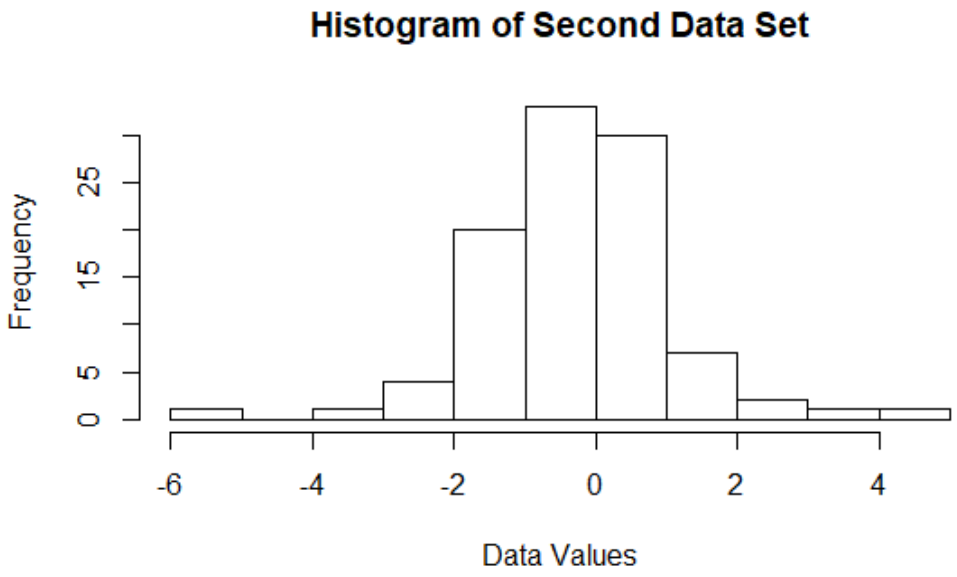

```
# Normal QQplot of qqdata

qqdata$resids <- qqdata$x - mean(qqdata$x)
qqnorm(qqdata$resids, main = "Normal Q-Q Plot of First Data Se
t")
qqline(qqdata$resids)
```



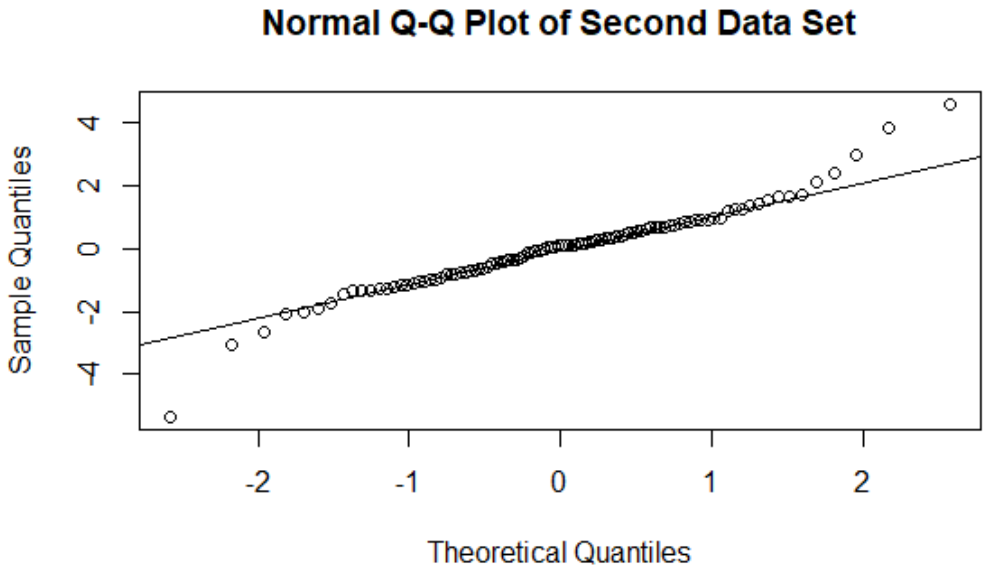
ii. The histogram's shape is slightly skewed right. From the Normal Q-Q Plot, we can see that the data points are flat on the left half of the data, indicating a large part of the data is on the left of the distribution. The data then begins to trend upwards showing that the data gets more spread out as it goes right, indicating it is right skewed.

c) Below are the histogram and normal QQ plots of second data set

i.
```r
# Histogram of qqdata2
hist(qqdata2$x, xlab = "Data Values", main = "Histogram of Sec
ond Data Set")
```



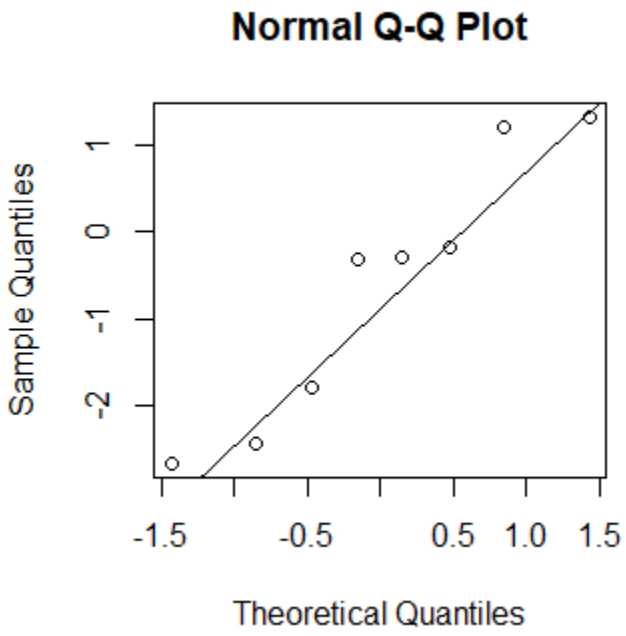

```r
# Normal QQplot of qqdata2
qqdata2$resids <- qqdata2$x - mean(qqdata2$x)
qqnorm(qqdata2$resids, main = "Normal Q-Q Plot of Second Data Set")
qqline(qqdata2$resids)
```

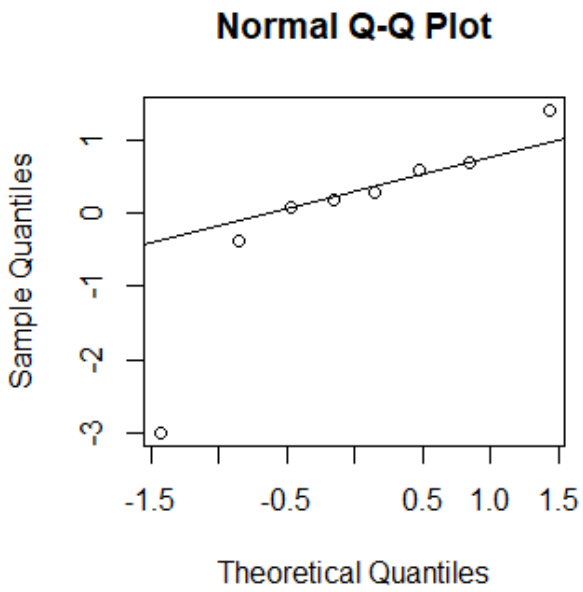


ii. The histogram has a normal distribution with skinny tails. The normal QQ plot has a large part of the data going in a straight line along the trend line. This shows that it is a normal distribution. The two tails on the ends of the data are indications that the tails are skinny of the distribution.

d) All the plots have a relatively straight line of data, even though it doesn't perfectly fit the trend line made by qqline(). We can assume that the data is normally distributed. Most of the plots are like the Oreo data my group collected, in that they are all normally distributed and have a straight line in their QQ plots. The first plot below is the most similar. The second plot was the one that would've given me the most trouble.

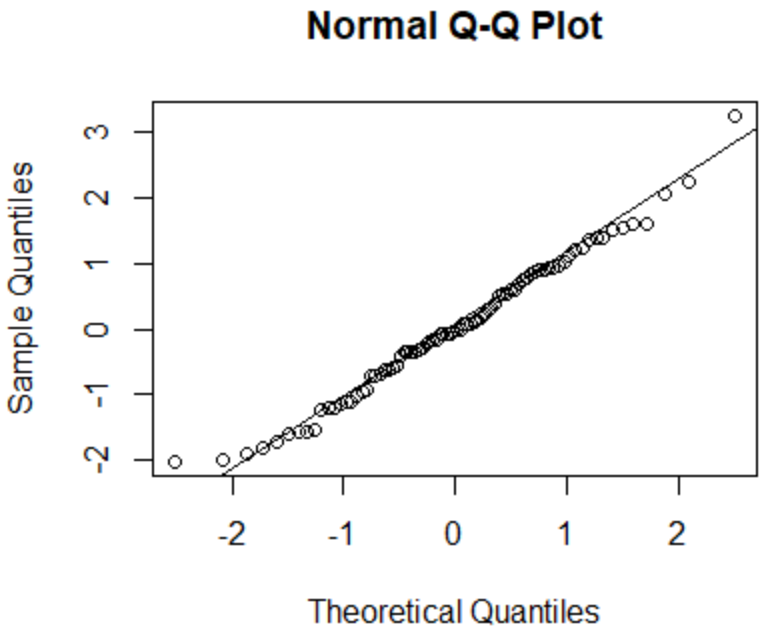
```
# part d
set.seed(42)
replicate(10, {
  y <- rnorm(8)
  qqnorm(y)
  qqline(y)
})
```

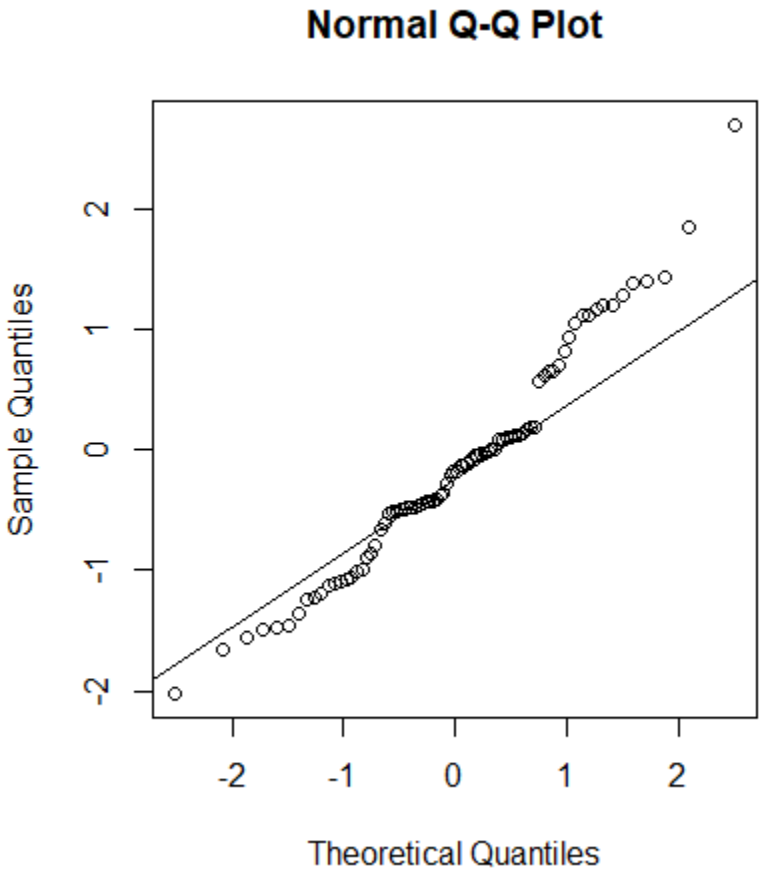




e) All the plots have data points that have a zig-zag shape around the trend line. Although the data isn't perfectly straight, because the data is linear in its path, we can assume the data is normally distributed. The simulated graphs are on a different scale, but they are like the residuals in our glyburide/insulin normal QQ plots. The first graph is our graph that is most similar to the glyburide/insulin graph, and the second graph is the most problematic of the simulated graphs.

```
# part e
set.seed(42)
replicate(10, {
  y <- rnorm(82)
  qqnorm(y)
  qqline(y)
})
```


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2
3


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2

5. **Hand Sanitizer Technique**
   a) A type I error is rejecting the null hypothesis when the null hypothesis is true. In the case of hand sanitizing techniques, this would mean determining that there is a difference in effectiveness of the hand washing techniques when there is no difference in whichever technique you use. The mean bacterial count after using CDC's or WHO's technique is the same, but you think that one is more effective than the other.
   b) A type II error is failing to reject the null hypothesis when the null hypothesis is false. Concerning the hand washing techniques, this would mean using either technique when there exists a more effective technique out of the two. One technique results in a different mean bacterial count than the other, but you assume that both have the same mean bacterial count.
   c) It would be more important to minimize β in this case because a type II error could result in using a less effective technique of hand washing, which could lead to more diseases and uncleanliness. Whereas making a type I error would just mean that you prefer one method over the other, but you could use either one and have the same results.

```
bacteria <- read.table(text = "count
700.67
420.82
278.78
647.36
361.25
164.19
154.06
356.10
274.30
319.89", header = TRUE, sep="")

bacteria_count_sd <- sd(bacteria$count)

d <- mean(bacteria$count) * 0.20

power.t.test(sig.level = .05, delta = d, power = .80, sd = bacteria_
count_sd)
##      Two-sample t test power calculation
##
##               n = 97.07331
##           delta = 73.5484
##              sd = 181.9802
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

The sample size needed to detect a 20% reduction in the mean bacterial count when using WHO's 6-step method with 80% power and an $\alpha = .05$ is about 97.

```
d) power.t.test(sig.level = .05, delta = d, power = .90, sd = bacteria_
   count_sd)##
   ##      Two-sample t test power calculation
   ##
   ##               n = 129.6233
   ##           delta = 73.5484
   ##              sd = 181.9802
   ##       sig.level = 0.05
   ##           power = 0.9
   ##     alternative = two.sided
```

The sample size needed to detect a 20% reduction in the mean bacterial count when using WHO's 6-step method with 90% power and an $\alpha = .05$ is about 130.

6. **The Art of Oreo Dunking**

```
a) power.t.test(n = 4, sig.level = .05, delta = mean_diff, sd = sp2)##
   ##      Two-sample t test power calculation
   ##
   ##               n = 4
   ##           delta = 1.335
   ##              sd = 0.47765
   ##       sig.level = 0.05
   ##           power = 0.905521
   ##     alternative = two.sided
```

The power needed to detect a difference in the mean amount of milk absorbed in 10 Seconds as large as 1.335 is 0.9055 for our Oreos experiment.

```
b) # Part 6b
   power.t.test(sig.level = .05, delta = mean_diff, power = .80, sd = s
   p2)
   ##      Two-sample t test power calculation
   ##
   ##               n = 3.305898
   ##           delta = 1.335
   ##              sd = 0.47765
   ##       sig.level = 0.05
   ##           power = 0.8
   ##     alternative = two.sided
```

The sample size needed to achieve 80% power to detect a difference in the mean amount of milk absorbed in 10 seconds as large as 1.335 is about 3.306.