S670 Problem set 2

Erik

January 24, 2017

```
library(ggplot2)
library(tidyr)

votes <- read.table("./pennsylvania.txt", header = TRUE)
```

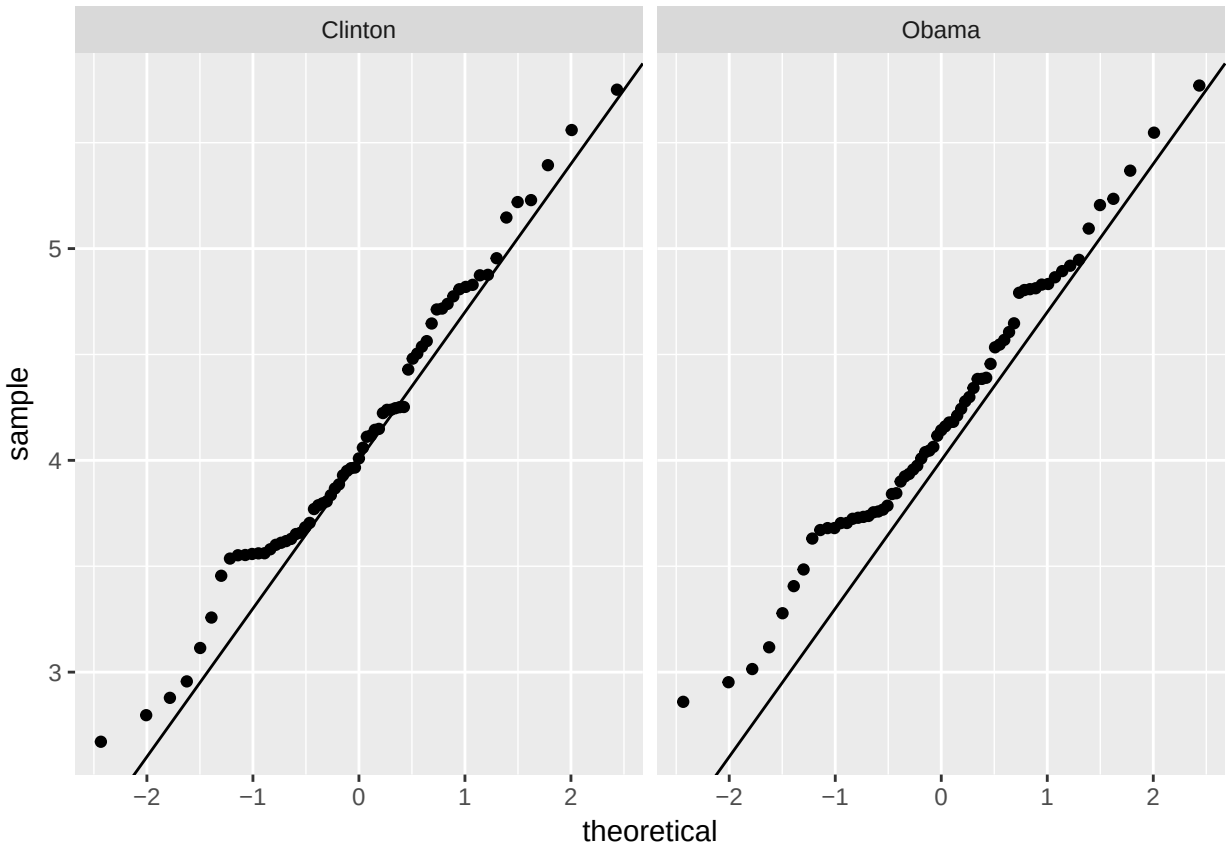
1. Use gather to convert data to "long form"

```
votes.long <- votes %>% gather(Candidate, Votes, Clinton:Obama)
```

2. Reproduce normal QQ plot of log_10 transformed votes data, with clinton and obama as the facet.

```
votes.long.log <- votes.long
votes.long.log[,3] <- log10(votes.long[,3])
# used to log transform just one comumn (the votes). On left, specify which column is changing. Then on

ggplot(votes.long.log, aes(sample = Votes)) + stat_qq() + facet_wrap(~Candidate) + geom_abline(intercep
```



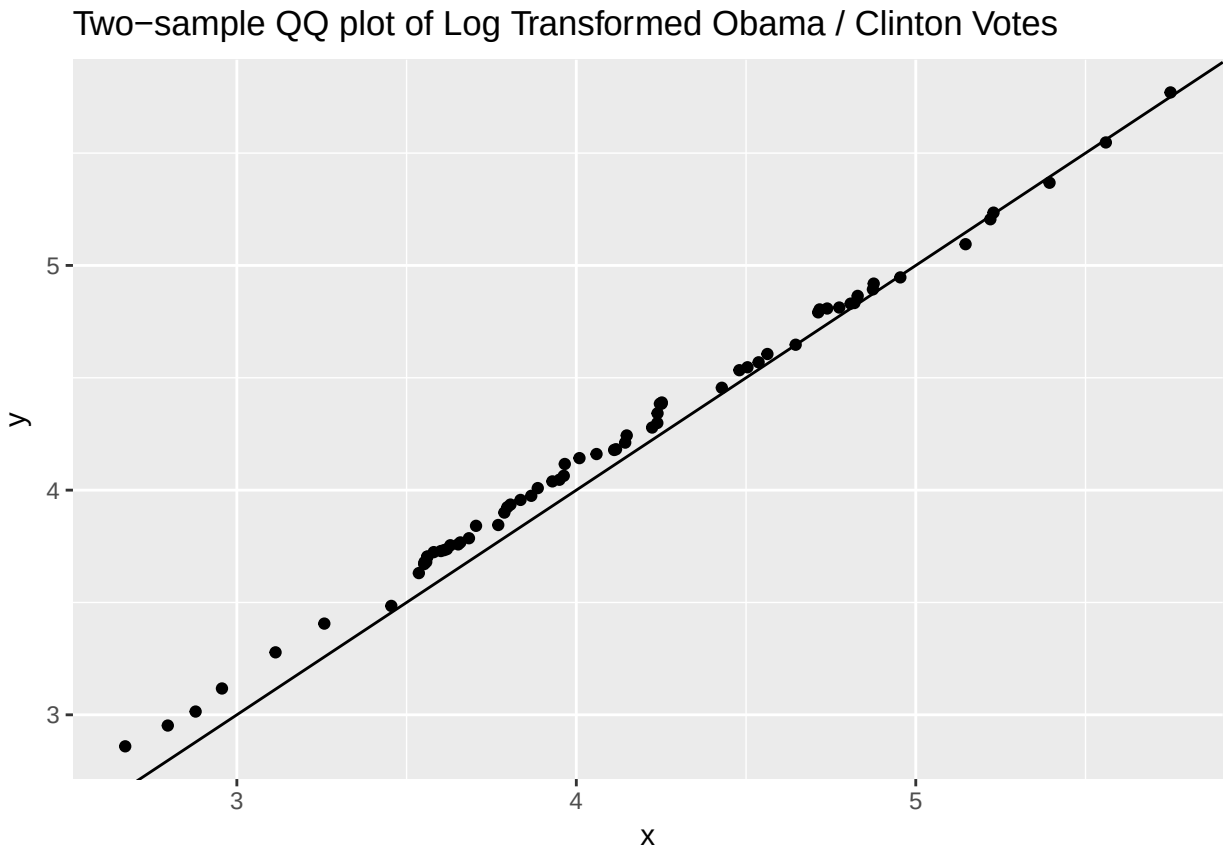
From this plot, we can conclude that the log_10 transformed data (while certainly better than the untransformed data!) still isn't really normal.

Despite the transformation there is a hump in the left side of the plot for both candidates, indicating that the distribution seems to be left skewed slightly for both.

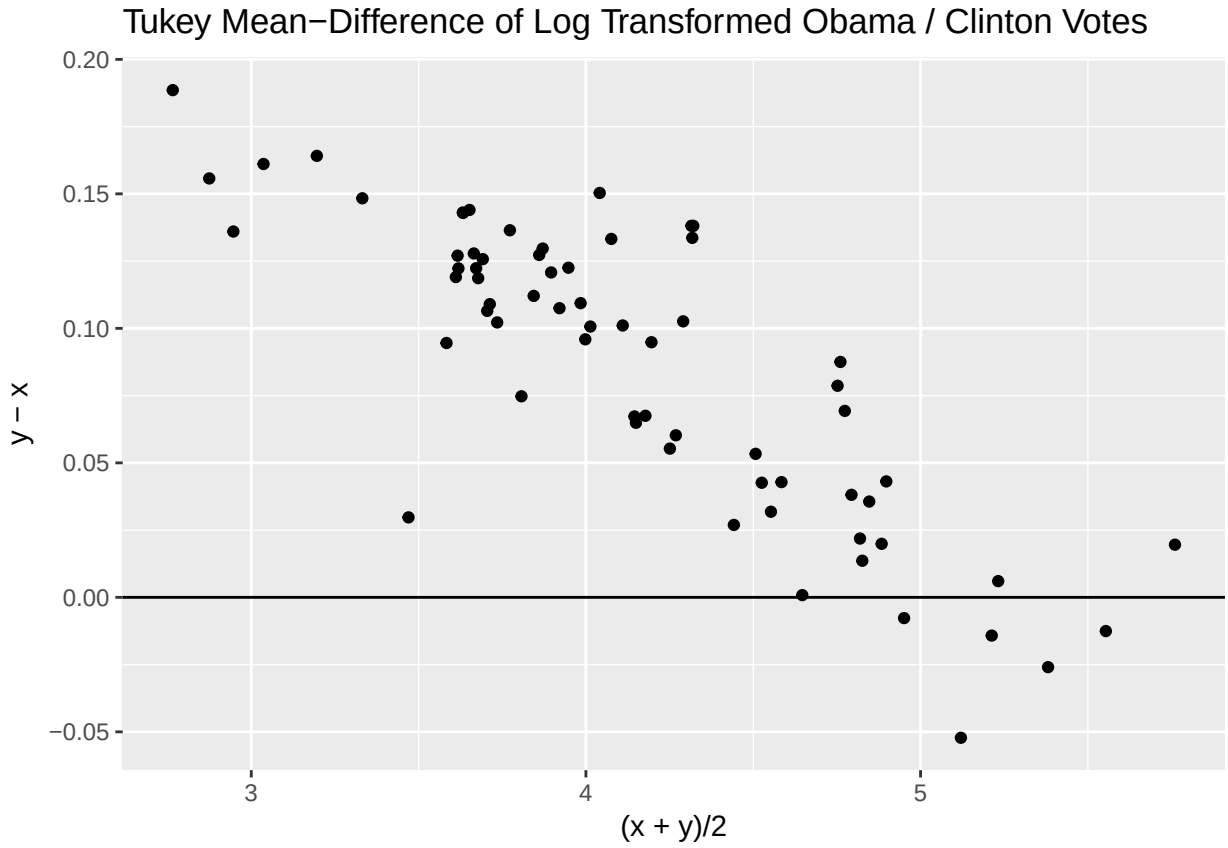
3. Is the relationship between Clinton and Obama's votes additive, multiplicative, or more complicated than that?

```
Clinton.log <- votes.long.log$Votes[votes.long.log$Candidate=="Clinton"]
Obama.log <- votes.long.log$Votes[votes.long.log$Candidate=="Obama"]
qq.df.log <- as.data.frame(qqplot(Clinton.log, Obama.log, plot.it = FALSE))

ggplot(qq.df.log, aes(x=x, y=y)) + geom_point() + geom_abline() + ggtitle("Two-sample QQ plot of Log Tra
```




```
ggplot(qq.df.log, aes(x=(x+y)/2, y=y-x)) + geom_point() + geom_abline(slope=0) + ggtitle("Tukey Mean-Di
```



From the first figure, we see that the log transformed data seems to be well described as a straight line, which is indicative of a multiplicative shift as we are working with log_10 transformed values. However, the story doesn't seem as simple as this being just a standard multiplicative shift as shown best in the second figure. In this Tukey mean-difference plot, it seems that if the average vote total is high (right side of x-axis), the difference between y (Clinton votes) and x (Obama votes) is negative, meaning Clinton earned more votes in high turnout places than did Obama. On the other hand, if the average number of votes (turnout) was low or middling, the difference between y and x is positive, meaning that Obama earned more votes. This second case held true for the majority of datapoints, so for the majority of counties in Pennsylvania Obama got more votes than Clinton did. These two graphs together seem to be saying that the relationship between the votes for Obama and Clinton is more complicated than just additive vs multiplicative.