Project

Group 1

November 30, 2019

Introduction

The data used in this report was collected from the Harvard Dataverse in an experiment about smokers and non-smokers, and how their social networks differ. The experiement looks into the differences in the participants' family, friend, group, and photo networks within Facebook. Among the metrics used in this experiement is "vertices", a measure of the number of people in a participant's network. Our experiement will test wether smoking has a significant influence over the number of vertices in a participant's Facebook network. And will focus solely on the data collected from underage smokers within the US that are 18 years and below. We will refer to the participants in our experiment as "teenagers".

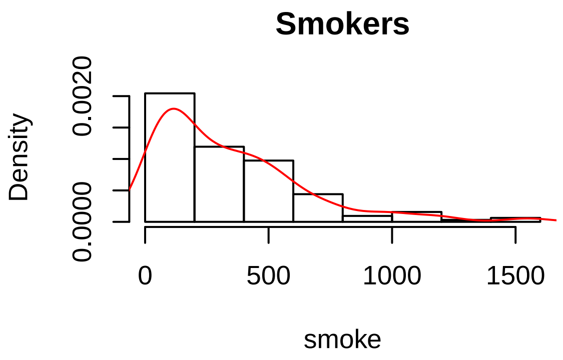
Problem

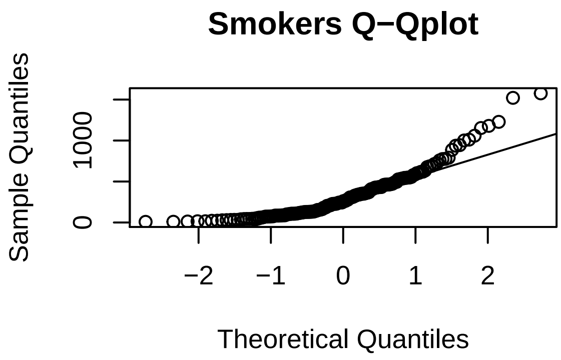
H0: The average number of vertices for a teenage smokers within the US is not significantly different than a non-smoker's. HA: HA !=H0 Alpha = .05

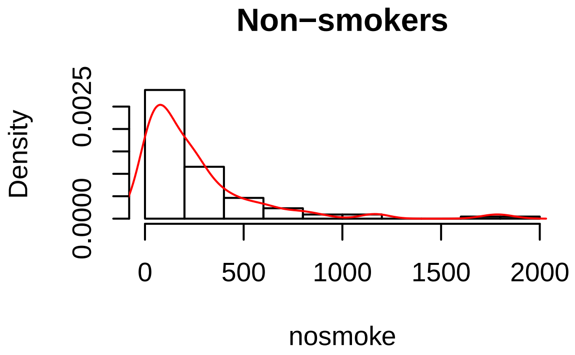
Purpose

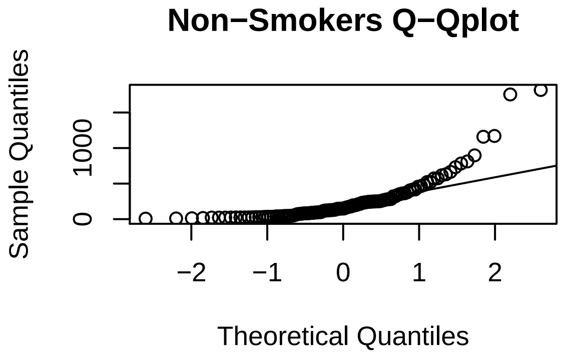
The purpose of this experiment is to test the significance of tobacco's influence, if any, on teenagers and their network of friends. We do so in support of, and to better inform, youth tobacco prevention.

Results and Discussion









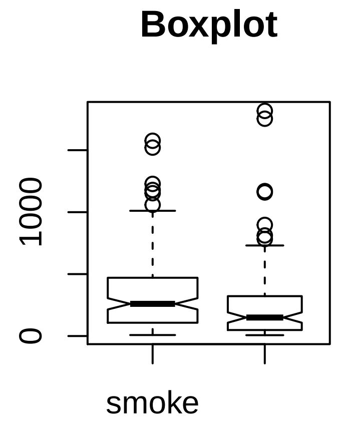
```
## [[1]]
## [1] "Decision: Fail to Reject the Null"
##
## $Alpha
## [1] 0.05
##
## $NumberOfBootSamples
## [1] 10000
```

```
##
##  Shapiro-Wilk normality test
##
## data:  smoke
## W = 0.8614, p-value = 6.112e-11
```

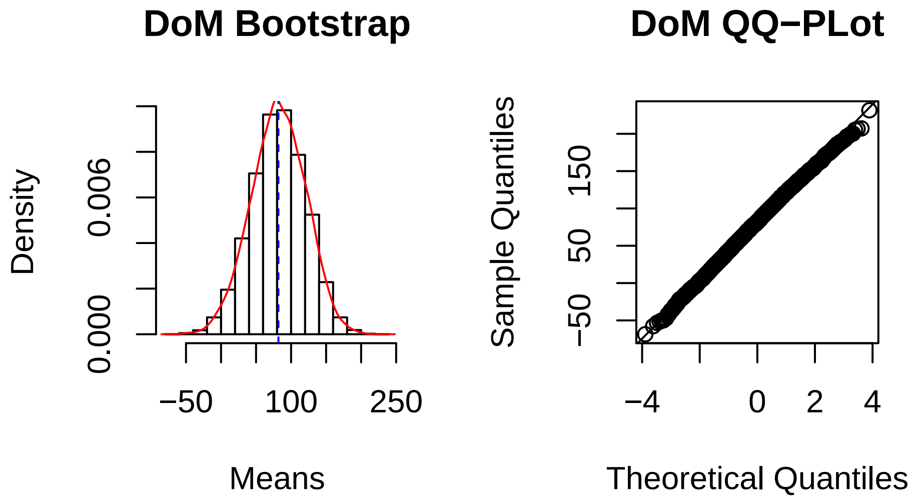
```
##
##  Shapiro-Wilk normality test
##
## data:  nosmoke
## W = 0.70014, p-value = 1.521e-13
```

Both histograms are not at all symmetric and the Q-Q plots are greatly skewed, suggesting the data is not normally distributed. Thus, in order to test whether there is a significant difference between the two, we need to bootstrap. The W statistic of both ‘smoke’ and ‘nosmoke’ from the Shapiro-Wilk test is significant at a 5% significance level since we have .86 and .70 with p values of 6.112e-11 and 1.521e-13 respectively. Thus with basically no evidence for normality, we must reject the hypothesis that their distributions are normal.

Equarvartest shows both groups have equal variances. This makes a comparison of their variances an option since both plots are skewed. But since, they are both skewed to the right, we choose instead to analyze their difference in means. We will refer to our chosen statistic as DoM from here on.



Since we have decided to focus on the DoM, we take a quick look at a visul representation of both means. The boxplot suggests means are different since there looks to be a significant drop from the smoke to nosmoke mean. We can see also that the medians are different because the notches of both boxplots do not overlap. Finally the boxplot is a reminder of the previously unaddressed but important fact that the smoke and nosmoke sets have unequal sizes.



```
## [1] "Smokers have this many vertices on average:"

## [1] 339.3396

## [1] "Non-Smokers have on average, this many vertices:"

## [1] 257.3889

## [1] "The observed difference:"
## [1] 81.95073
```

```
## [1] "The Bootstrap Mean difference:"
```

```
## [1] 81.97261
```

```
## [1] "Bootstrap STD"
```

```
## [1] 38.88226
```

```
## [1] "Bootstrap bias"
```

```
## [1] 0.0218768
```

```
##       2.5%      97.5%
##   4.264212 154.965894
```

```
## [1] 0.4988501
```

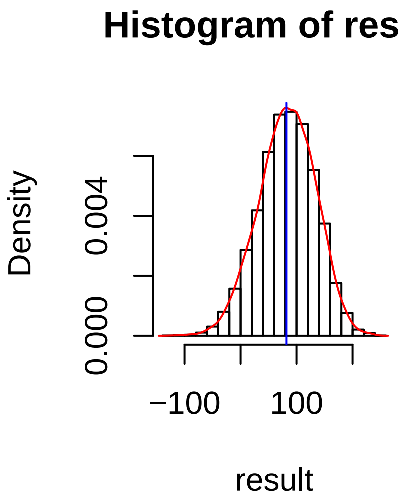
Note that the histogram is almost symmetric and tapers off with thin tails on both sides. The normal qq-plot also suggests normality of the bootstrap statistic. The observed DoM seems to be in the center of the distribution. The 95% Confidence Interval for the difference in means doesn't contain zero. And our interval, (4.560679, 155.463505), suggests that there is a SIGNIFICANT difference in the mean vertices between teenage smoker and non-smoker networks. However, with a p value of .49, we fail to reject the null at a 5% significance level suggesting no significant DoM. And so the bootstrap gives conflicting results.

```
##        5%       95%
##  16.79408 144.80117
```

We try to correct this by modifying alpha, but changing alpha to .1 does not change the discrepancy in the results, it only trims the CI (17.93173, 144.96056).

Next we perform a permutation test to see whether the difference is significant.

```
## [1] 0.4817518
```

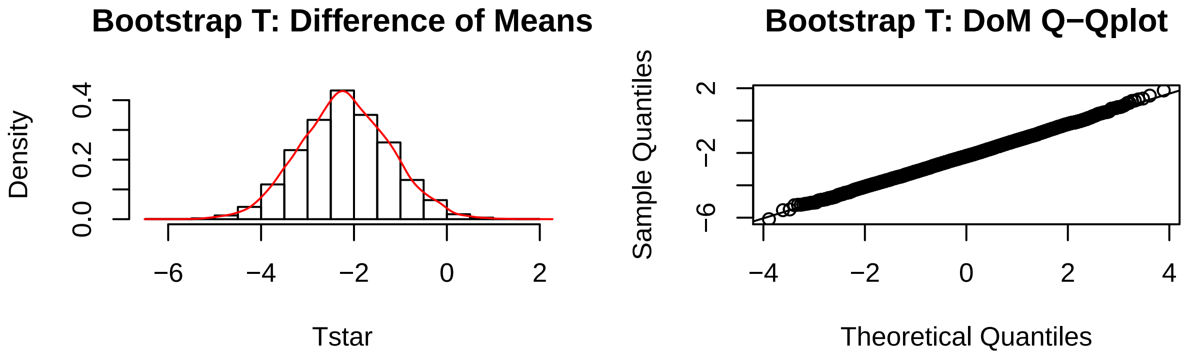


Once again, the histogram is almost symmetric and tapers off with thin tails on both sides. Once again, with a p value of .48, we fail to reject the null at a 5% significance level suggesting no significant DoM.

```
## [1] FALSE
```

At this point of the test, it appears that we should claim no significant DoM. We perform a bootstrap T test to construct another CI.

```
##     97.5%      2.5%
##  91.44448 238.81746
```



This CI (90.79513 238.49807) is even more outrageous compared to our previous analysis suggesting significant DoM. We need to check the power of the test to determine if we are right to reject the null hypothesis. This could provide more suport one way or another.

```
## [1] 1
```

To test the power, we simulated the percentage of times we correctly reject the null hypothesis when the null is false. We ensured this false null by setting the difference in means to distinctly different numbers, 1:6. When we set the difference >= 6, we get a power of 0, thus incorrectly rejecting the null hypothesis 100% of the time. When the difference is <= 5, we get a power of 100, thus correctly rejecting the null hypothesis 100% of the time. The lack of middle ground in the power test does however raise a red flag and lead us to question its validity.

Conclusion

We failed to reject the null hypothesis, which means that tobacco has no significant influence on teenagers and their network of friends. Thus we can tentatively conclude that smoking has no impact on teenage facebook networks. Future work should take into account the unequal size of the samples, consider other metrics, like transitivity and edges, from this dataset and look into a time series model of the data.

References:

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/XMPAUQ