STAT 355 Project

Part 1

The data I selected analyzes the highway accidents of 39 sections of large highways in Minnesota in the year 1973. I found this data on Github, and I found this data interesting because I wanted to see how driving and safety measures were like back in the 70's. My parents told me that back in the 70's cars were like a luxury commodity, at least for them, so were people a lot more careful back then with their driving, or was it that because there wasn't enough training on how to drive, people would get into accidents more often? Let's dive into our data and figure this out.

```
> highway = read.csv("Highway1.csv")

> head(highway)

  X rate   len adt trks      sigs1 slim shld lane acpt  itg lwid htype

1 1 4.58  4.99  69    8 0.20040080   55   10    8  4.6 1.20   12   FAI

2 2 2.86 16.11  73    8 0.06207325   60   10    4  4.4 1.43   12   FAI

3 3 3.02  9.75  49   10 0.10256410   60   10    4  4.7 1.54   12   FAI

4 4 2.29 10.65  61   13 0.09389671   65   10    6  3.8 0.94   12   FAI

5 5 1.61 20.01  28   12 0.04997501   70   10    4  2.2 0.65   12   FAI

6 6 6.87  5.97  30    6 2.00750419   55   10    4 24.8 0.34   12    PA
```

The data contains the following variables: rate (the rate of accidents in that section per million vehicles), len (the length of that highway section), adt (the average daily traffic count in thousands), trks (the percent of trucks amongst the sample), sigs1 (the number of signals per mile of roadway, this is calculated by: ((number of signalized interchanges per mile * len + 1)/len), slim (the speed limit of that section), shld (the width in feet of the shoulder), lane (the total number of lanes), acpt (the total number of access points), itg (the number of freeway interchanges per mile), lwid (lane width in feet), htype (documents the different kinds of highways classified by FAI, MC, PA, or MA)

Part 2

Question 1:

Cops pull people over for speeding, because it is considered reckless and a danger to others on the road. Essentially the question I want to ask is that is there a difference between the speed limit and the rate at which accidents occur.

To answer this question, I will divide the data into two portions, slower speed limit and faster speed limit. We will define the term slower as 55 mph and below, while faster is anything above that.

```
> avgspeed = mean(slim)

> avgspeed
```

[1] 55

> fasterspeed = highway[which(slim > avgspeed),]

> slowerspeed = highway[which(slim <= avgspeed),]

This separates the data into two population means, and now I can run a t-test on this, to see if there is a difference between the speed limits on accident rates. I hypothesize that greater speed limit would lead to more accidents.

We define the Null hypothesis as:

H_0 = There is no significant difference for accident rates between faster and slower speed limits

H_1 = The faster the speed limit, the greater the accident rate

> t.test(x=fasterspeed$rate, y=slowerspeed$rate, alternative = c("greater"))

Welch Two Sample t-test

data: fasterspeed$rate and slowerspeed$rate

t = -4.2386, df = 34.968, p-value = 0.9999

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

-2.740766 Inf

sample estimates:

mean of x mean of y

2.626923 4.586538

Wow! That's surprising. Because the p-value is greater than the default alpha of 0.05, we fail to reject the null hypothesis, and say there is no significant difference for accident rates between faster and slower speed limits. But hold on, if we look at the mean of x and the mean of y, we are seeing that the slower speed limits are much more prone to accidents.

Let's quickly run a two sided test to make sure that we are right that there is no significant difference between the speed limits on accident rates.

H_0: There is no significant difference between faster and slower speed limits on accident rates

H_1: There is significant difference between faster and slower speed limits on accident rates

> t.test(x=fasterspeed$rate, y=slowerspeed$rate, alternative = c("two.sided"),)

Welch Two Sample t-test

data: fasterspeed$rate and slowerspeed$rate

t = -4.2386, df = 34.968, p-value = 0.0001559

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-2.898216 -1.021015

sample estimates:

mean of x mean of y

2.626923 4.586538

The test gets a p-value that is less than the default alpha, and as such we reject the null hypothesis, saying that there is in fact significant difference between faster and slower speed limits on accident rates.

So let's run this test again but with the following hypothesis:

H_0 = There is no significant difference for accident rates between faster and slower speed limits

H_1 = The slower the speed limit, the greater the accident rate

> t.test(x=fasterspeed$rate, y=slowerspeed$rate, alternative = c("less"))

Welch Two Sample t-test

data: fasterspeed$rate and slowerspeed$rate

t = -4.2386, df = 34.968, p-value = 7.796e-05

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -1.178465

sample estimates:

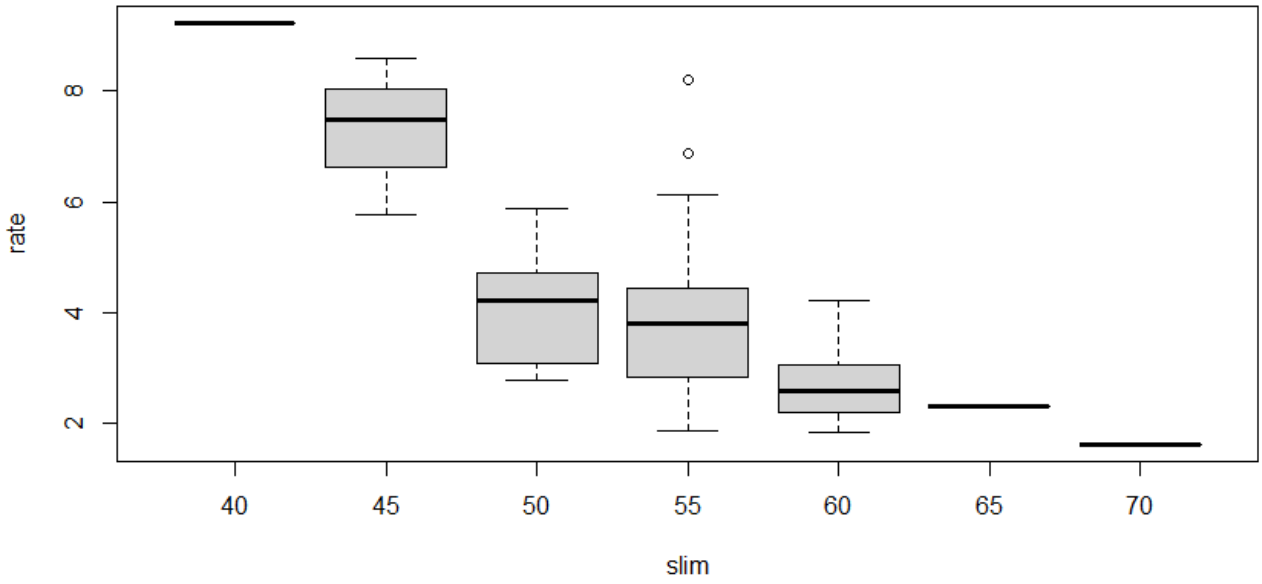
mean of x mean of y

2.626923 4.586538

In this test, we find that the p-value is less than the default alpha, and as such we can successfully reject the null hypothesis and state that slower speed limits actually cause more accidents than faster speed limits.

Let's visualize this data:

> boxplot(rate~slim)



As you can see in this boxplot above, on average, the slower speed limits do have greater accident rates than the faster speed limits.

Question 2:

Do the different types of roadways and funding for these roads could contribute to accident rates? I know back home there's a highway I often tend to go around, as it's maintained by a lower income county and has way too many potholes to my liking.

We find that the sections of highways are classified by the following: FAI, MC, PA, and MA. So let's run an Anova test to see if there's reasonable differences of accident rates among classifications of roadways.

Let the hypothesis be:

H_0: There is no significant difference on accident rates between the different types of roadways

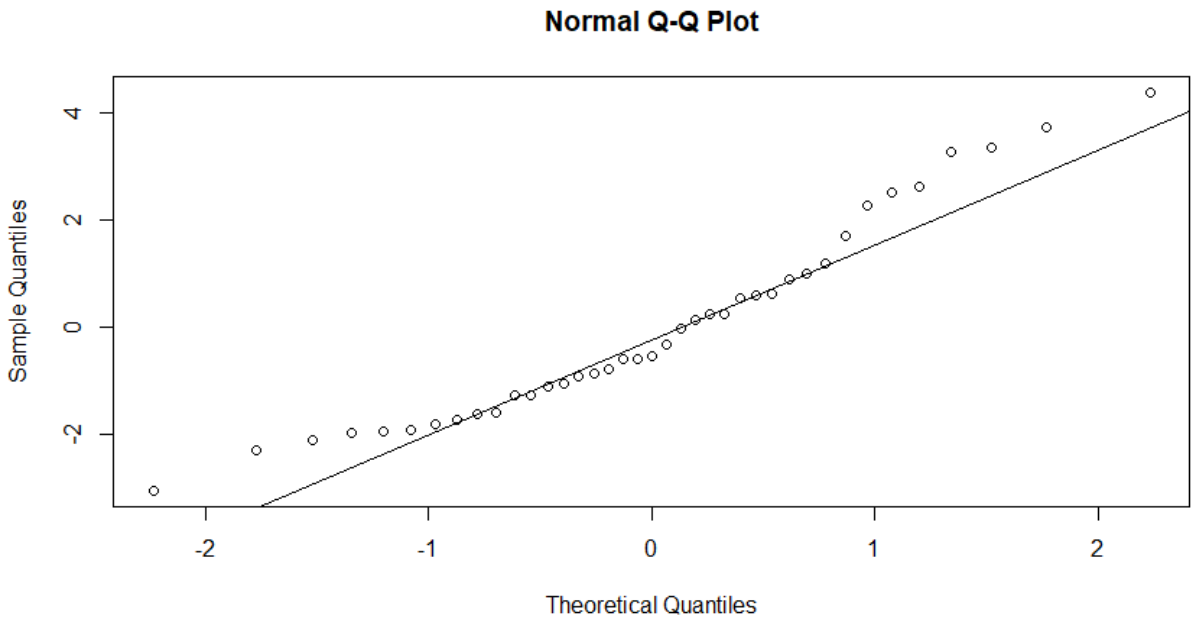
H_1: There is significant difference on accident rates between the different types of roadways

> summary(aov(rate~htype))

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
htype	3	19.29	6.429	1.723	0.18
Residuals	35	130.60	3.731		

From our test, we find the p value to be greater than the default alpha of 0.05, and as such we fail to reject the null hypothesis, saying that there is no significant difference on accident rates between different types of roadways

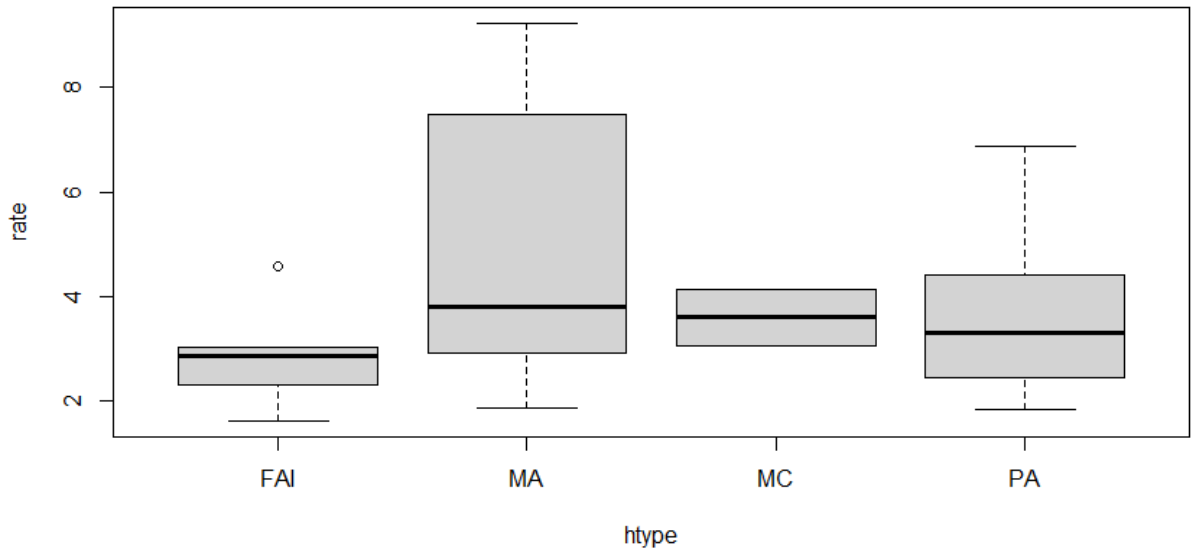
```
> roadways = aov(rate~htype)
> roadways
Call:
   aov(formula = rate ~ htype)

Terms:
                    htype Residuals
Sum of Squares   19.28608 130.59998
Deg. of Freedom         3        35

Residual standard error: 1.93169
Estimated effects may be unbalanced
> qqnorm(roadways$residuals)
> qqline(roadways$residuals)
```



The normal graph does show linearity, so we can infer the mean is close to the meidan

Additionally to further prove our answer:

> boxplot(rate~htype)



As we can see in the box plot, the means of the different roadway types have similar accident rates, thus showing there is not significant difference on accident rates between roadways

Question 3:

Well looks like the different kinds of roads didn't really cause accidents back then. But what about shoulder width. I personally don't mind driving on a tiny shoulder, but my parents advise me against driving on the left most lane when they have those construction walls near the shoulder. Is there significant difference on accident rates between different shoulder widths?

Lets state the hypothesis:

H_0: There is no significant difference on accident rates between different shoulder widths

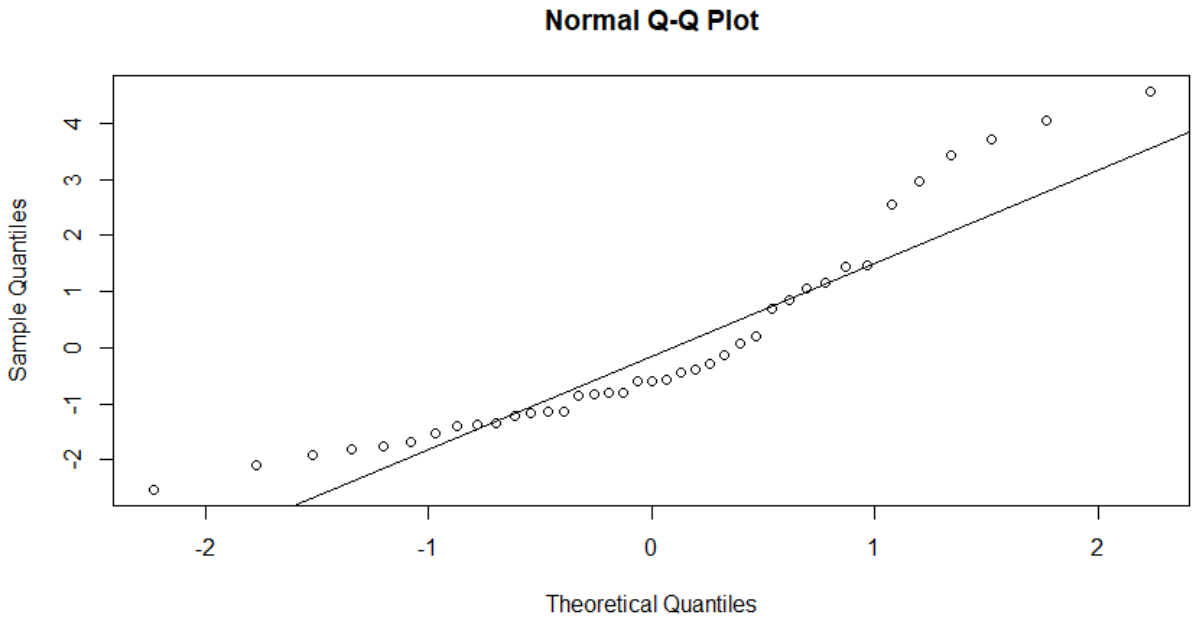
H_1: There is significant difference on accident rates between different shoulder widths

```
> summary(aov(rate~shld))
            Df Sum Sq Mean Sq F value Pr(>F)
shld         1  22.44  22.438   6.514  0.015 *
Residuals   37 127.45   3.445
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Through running this anova test, we find that the p value is less than the default alpha of 0.05, so we can successfully reject the null hypothesis and state that there is significant difference on accident rates between different shoulder widths

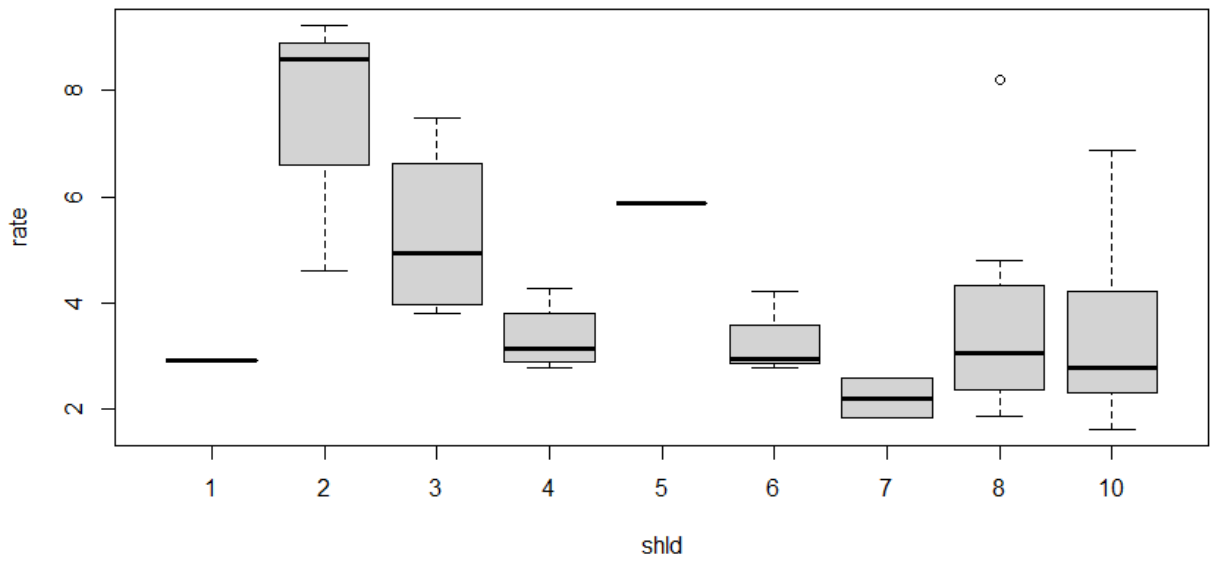
```
> shoulders = aov(rate~shld)
> shoulders
Call:
   aov(formula = rate ~ shld)

Terms:
                    shld Residuals
Sum of Squares   22.43752 127.44855
Deg. of Freedom         1        37

Residual standard error: 1.855951
Estimated effects may be unbalanced
> qqnorm(shoulders$residuals)
```

> qqline(shoulders$residuals)



Uh, this seems relatively linear right? So we can infer that the means are close to the median

> boxplot(rate~shld)



If we look at the boxplot, yes, we can see that smaller width shoulders lead to more accidents than larger width shoulders. This backs up the hypothesis we selected.

Question 4:

I find this assignment pretty interesting, so let's do another question for the sake of it. I would assume that if there are less lanes, then the chances of accidents would be much lower. I mean, if you think about it, there's nowhere to go but straight. So the question I am asking is, does the number of lanes provide significant difference for accident rates.

> avglanes = mean(lane)

> avglanes

[1] 3.128205

> avglanes = round(avglanes, 0)

> avglanes

[1] 3

Let's define smaller highways as 3 lanes and lower, while larger lanes would be greater than 3 lanes.

> smallhghwy = highway[which(lane<=avglanes),]

> largehghwy = highway[which(lane>avglanes),]

This separates the data into two population means, and now I can run a t-test on this, to see if there is a difference between the number of lanes on accident rates. I hypothesize that greater amount of lanes would lead to more accidents.

H_0: There is no significant difference between the number of lanes on accident rates

H_1: The greater the number of lanes, the greater the rate of accidents

> t.test(x=largehghwy$rate, y=smallhghwy$rate, alternative = c("greater"),)

Welch Two Sample t-test

data: largehghwy$rate and smallhghwy$rate

t = -0.21403, df = 36.947, p-value = 0.5842

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

-1.22464 Inf

sample estimates:

mean of x mean of y

3.862632 4.000500

The p-value of this test is greater than the default alpha of 0.05, thus we fail to reject the null hypothesis that there is no significant difference between the number of lanes and the rate of accidents.

Lets verify this by running the following tests:

Two Sided:

H0: There is no significant difference between larger highways and smaller highways on accident rates

H1: There is significant difference between larger highways and smaller highways on accident rates

> t.test(x=largehghwy$rate, y=smallhghwy$rate, alternative = c("two.sided"),)

Welch Two Sample t-test

data: largehghwy$rate and smallhghwy$rate

t = -0.21403, df = 36.947, p-value = 0.8317

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.443092 1.167355

sample estimates:

mean of x mean of y

3.862632 4.000500

Less:

H0: There is no significant difference between larger highways and smaller highways on accident rates

H1: The smaller the number of lanes, the greater the accident rates

> t.test(x=largehghwy$rate, y=smallhghwy$rate, alternative = c("less"),)

Welch Two Sample t-test

data: largehghwy$rate and smallhghwy$rate

t = -0.21403, df = 36.947, p-value = 0.4158

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf 0.9489034

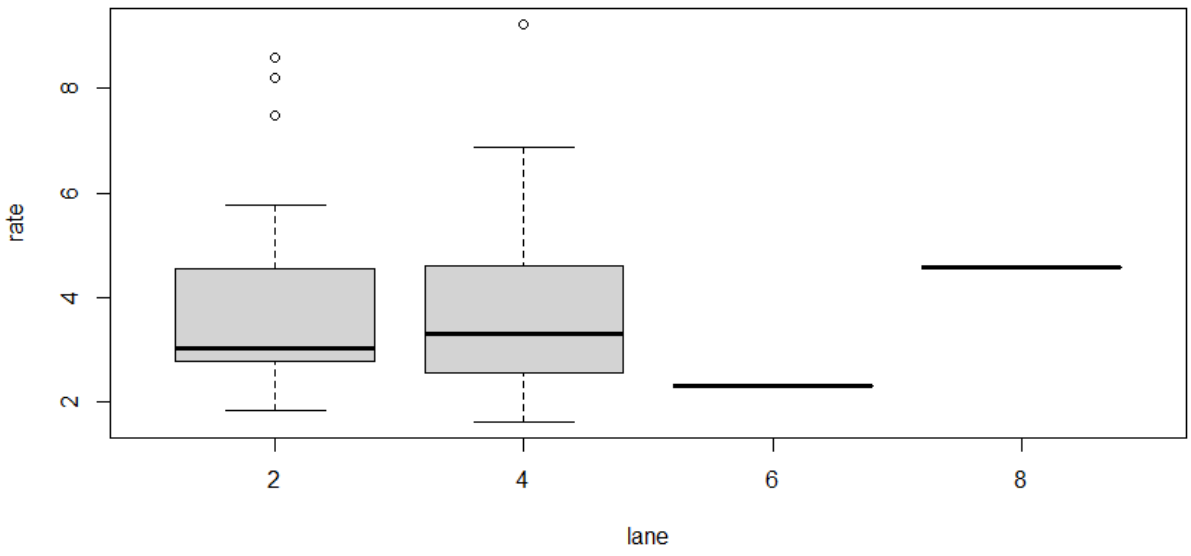
sample estimates:

mean of x mean of y

3.862632 4.000500

We see that the p value in both of these tests proved to be greater than the default alpha of 0.05, thus we fail to reject the null hypothesis certifying that there is no significant difference in the number of lanes on accident rates.

> boxplot(rate~lane)



This boxplot does back our data showing that there is not that significant of a difference in the number of lanes on accident rates. Although it is worth noting that there was only one 8 lane highway section in the sample, so this worth noting outside this dataset.

Conclusion

So in conclusion what can we conclude from the questions posed on this dataset. Well 1. The slower the speed limit, the greater the accident rate. 2. The type of roadways didn't really alter the number of accidents. 3. The smaller the shoulder, the greater the accident rate. 4. The number of lanes didn't really cause any difference in the number of accidents.