Problem Set 11

Keith Hickman

November 19, 2017

R Markdown

Problem 1

Trosset 12.6 A.

```
library(data.table)
p1.data <- read.csv("C:\\Users\\khickman\\Desktop\\Personal\\IUMSDS\\StatsS520\\Module13\\mydata.csv", 
p1.data
```

```
##       V1 V2
## 1  37.54  A
## 2  37.01  A
## 3  36.71  A
## 4  37.03  A
## 5  37.32  A
## 6  37.01  A
## 7  37.03  A
## 8  37.70  A
## 9  37.36  A
## 10 36.75  A
## 11 37.45  A
## 12 38.85  A
## 13 40.17  B
## 14 40.80  B
## 15 39.76  B
## 16 39.70  B
## 17 40.79  B
## 18 40.44  B
## 19 39.79  B
## 20 39.38  B
## 21 39.04  C
## 22 39.21  C
## 23 39.05  C
## 24 38.24  C
## 25 38.53  C
## 26 38.71  C
## 27 38.89  C
## 28 38.66  C
## 29 38.51  C
## 30 40.08  C
```

1.1.

Use side-by-side boxplots and normal probability plots to investigate the ANOVA assumptions of normality and homoscedasticity. Do these assumptions seem plausible? Why or why not?

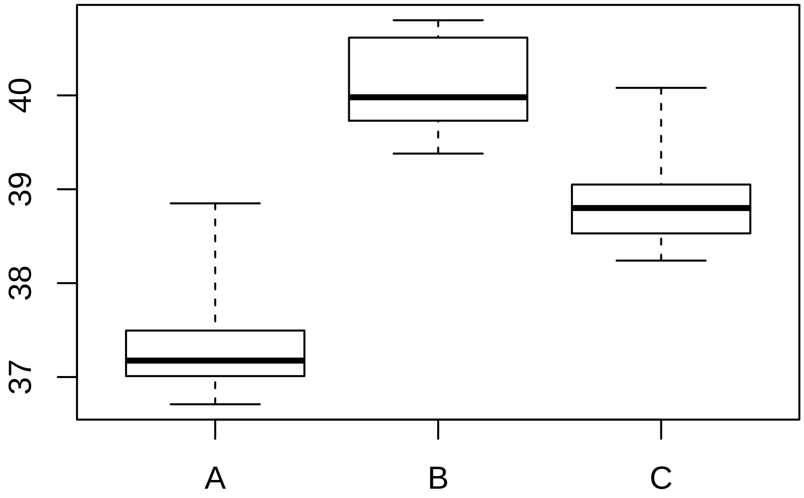
Boxplots:

```
## Split the dataset into three different groups, A, B, and C:
p1.a <- subset(p1.data, V2=="A")
p1.b <- subset(p1.data, V2=="B")
p1.c <- subset(p1.data, V2=="C")

#Checking variables to make sure they worked:
#p1.a
#p1.b
#p1.c
```

Conditioned boxplot for our three variables:

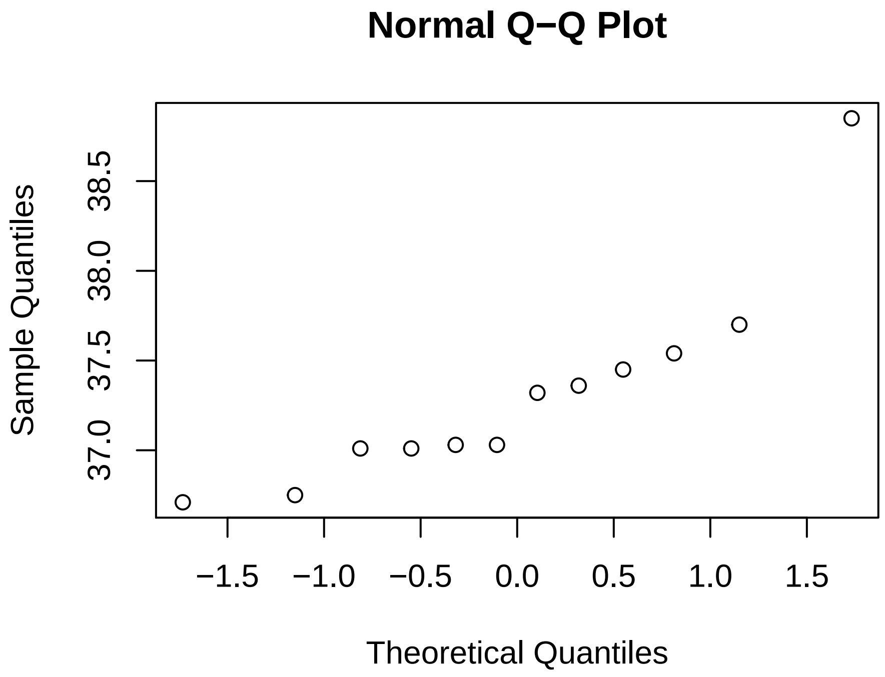
```
boxplot(p1.a$V1, p1.b$V1, p1.c$V1, range=0, names=c("A", "B", "C"))
```



At first glance, there appears to be quite a bit of difference in the median and quantiles between these three variables. The boxplots are a bit misleading I believe, because the scale is relatively small. If we were to zoom out, the values would be much closer together.

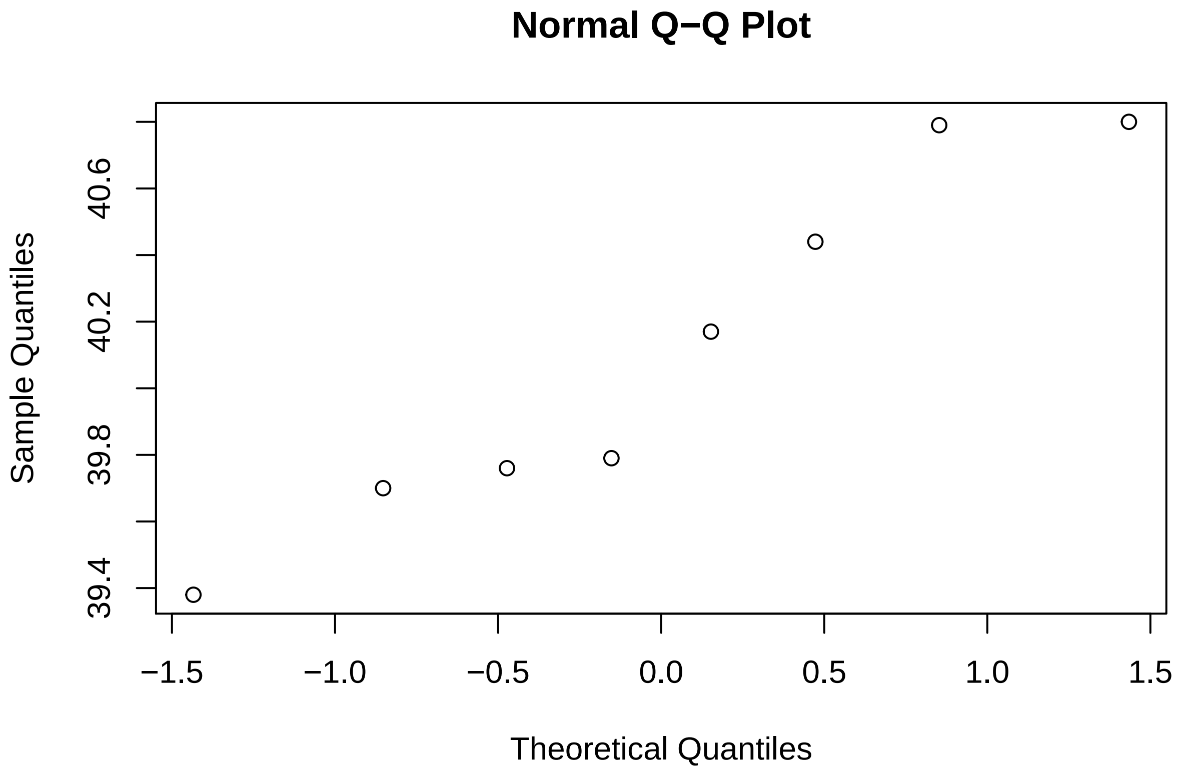
Normality plots for all variables:

```
qqnorm(p1.a$V1)
```



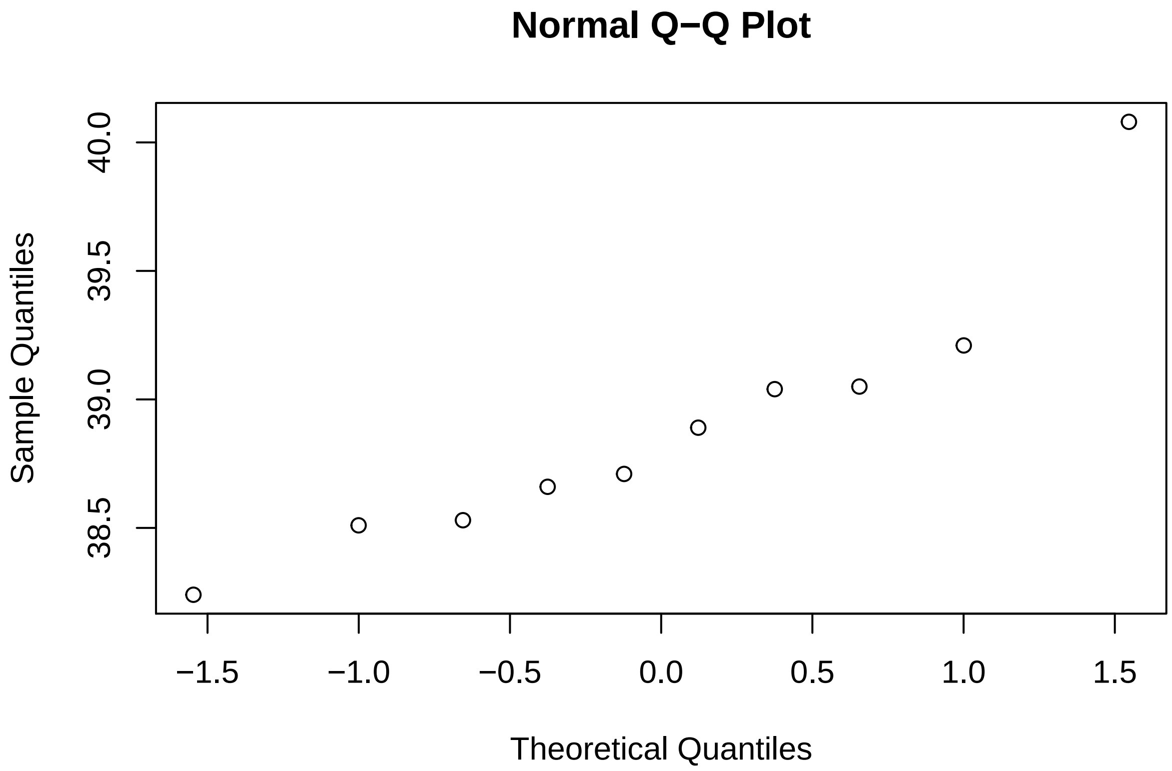
There is definitely a non-normal appearance here, especially at the tails. I'm confident this is a non-normal variable

```
qqnorm(p1.b$V1)
```



No straight line here, either, though this one looks a bit closer to normal. With a small sample size, however, we can't assume normality.

```
qqnorm(p1.c$V1)
```



Most of the variable looks relatively normal, with the exception of one outlier.

Finally, let's compare the standard deviation of all three variables:

```
sd(p1.a$V1)
```

```
## [1] 0.5727975
```

```
sd(p1.b$V1)
```

```
## [1] 0.5313846
```

```
sd(p1.c$V1)
```

```
## [1] 0.510812
```

All three variances are relatively close - within .06 or so. However, relatively speaking, there's supporting evidence that the variables have different standard deviations and thus variances sd^2. The assumption of homoscedasticity is a bit strained here. Variable `A` has a greater variance than B and C by about .04. We could use a log or square root transform, but let's proceed under the assumption of both normality and homoscedasticity.

1.2

Use ANOVA to test the null hypothesis that the three sites have the same mean salinity. Use a significance level of $\alpha = 0.05$ and organize your calculations in an ANOVA table.

Our null hypothesis is $H_0 : \mu_1 = \mu_2 = \mu_3$ where mean of variables A, B, and C in our dataset are μ_1, μ_2, μ_3 respectively.

Lengths:

```
N <- length(p1.a$V1) + length(p1.b$V1) + length(p1.c$V1)
N
```

```
## [1] 30
```

```
a.n <- length(p1.a$V1)
b.n <- length(p1.b$V1)
c.n <- length(p1.c$V1)

mean.a <- mean(p1.a$V1)
mean.b <- mean(p1.b$V1)
mean.c <- mean(p1.c$V1)
```

Total Sum of Squares, Grand Mean and Degrees of Freedom:

```
allvar <- c(p1.a$V1, p1.b$V1, p1.c$V1)
allvar
```

```
##  [1] 37.54 37.01 36.71 37.03 37.32 37.01 37.03 37.70 37.36 36.75 37.45
## [12] 38.85 40.17 40.80 39.76 39.70 40.79 40.44 39.79 39.38 39.04 39.21
## [23] 39.05 38.24 38.53 38.71 38.89 38.66 38.51 40.08
```

```
allvar.df <- length(allvar) - 1
allvar.df
```

```
## [1] 29
```

```
grandmean <- mean(allvar)

SST <- sum((allvar - grandmean)^2)
total.df = N - 1
```

Between Sum of Squares

```
SSB <- a.n * (mean.a - grandmean)^2 + b.n * (mean.b - grandmean)^2 + c.n * (mean.c - grandmean)^2
SSB
```

```
## [1] 38.80088
```

```
between.df <- 2

between.meansquare <- SSB/between.df
between.meansquare
```

```
## [1] 19.40044
```

Within Sum of Squares

```
SSW <- ((a.n-1) * var(p1.a$V1)) + ((b.n - 1) * var(p1.b$V1)) + ((c.n - 1) * var(p1.c$V1))
SSW
```

```
## [1] 7.934014
```

```
within.df <- N-3
within.meansquare <- SSW/within.df
within.meansquare
```

```
## [1] 0.2938524
```

F-statistic: Between MS / Within MS

```r
F <- between.meansquare / within.meansquare
F
```

```
## [1] 66.02105
```

P-value calculation using the F statistic, and two degrees of freedom:

```r
1 - pf(F, between.df, within.df)
```

```
## [1] 4.008649e-11
```

This is an extremely small p-value and well below our significance value α of .05, indicating that there is strong evidence against the null and in support of the alternative hypothesis. This makes sense from an initial analysis of our data using boxplots and `qqnorm` functions. This means that within our dataset, it appears that there is a difference in the average salinity across the different sample spaces.

Checking using the ANOVA function:

```r
anova(lm(allvar ~ p1.data$V2))
```

```
## Analysis of Variance Table
##
## Response: allvar
##            Df Sum Sq Mean Sq F value    Pr(>F)
## p1.data$V2  2 38.801 19.4004  66.021 4.009e-11 ***
## Residuals  27  7.934  0.2939
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Everything looks good!

Problem 2

Trosset 12.6 B. Read in the data:

```r
p2.data <- read.csv("C:\\Users\\khickman\\Desktop\\Personal\\IUMSDS\\StatsS520\\Module13\\sickle.csv", 
p2.data
```

```
##      V1 V2
## 1   7.2 SS
## 2   7.7 SS
## 3   8.0 SS
## 4   8.1 SS
## 5   8.3 SS
## 6   8.4 SS
## 7   8.4 SS
## 8   8.5 SS
## 9   8.6 SS
## 10  8.7 SS
## 11  9.1 SS
## 12  9.1 SS
## 13  9.1 SS
## 14  9.8 SS
## 15 10.1 SS
## 16 10.3 SS
## 17  8.1 ST
## 18  9.2 ST
## 19 10.0 ST
## 20 10.4 ST
## 21 10.6 ST
## 22 10.9 ST
## 23 11.1 ST
## 24 11.9 ST
## 25 12.0 ST
## 26 12.1 ST
## 27 10.7 SC
## 28 11.3 SC
## 29 11.5 SC
## 30 11.6 SC
## 31 11.7 SC
## 32 11.8 SC
## 33 12.0 SC
## 34 12.1 SC
## 35 12.3 SC
## 36 12.6 SC
## 37 12.6 SC
## 38 13.3 SC
## 39 13.3 SC
## 40 13.8 SC
## 41 13.9 SC
```

2.1.

Use side-by-side boxplots and normal probability plots to investigate the ANOVA assumptions of normality and homoscedasticity. Do these assumptions seem plausible? Why or why not?

Here, our null hypothesis is $H_0 : \mu_1 = \mu_2 = \mu_3$ where the expected value of the variables SS, ST, and SC are μ_1, μ_2, and μ_3 respectively.

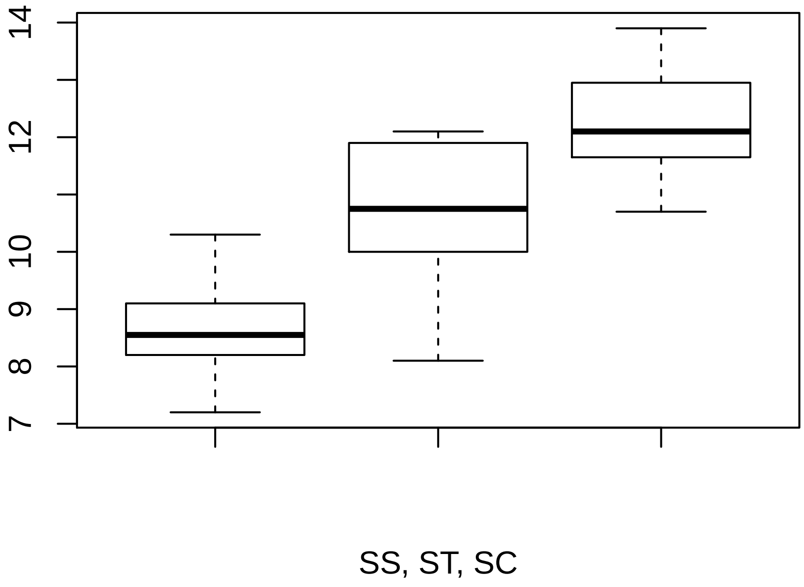
Let's create three variables first:

```
ss <- subset(p2.data, V2=="SS")
st <- subset(p2.data, V2=="ST")
sc <- subset(p2.data, V2=="SC")

##Check the variables:
#ss
#st
#sc
```

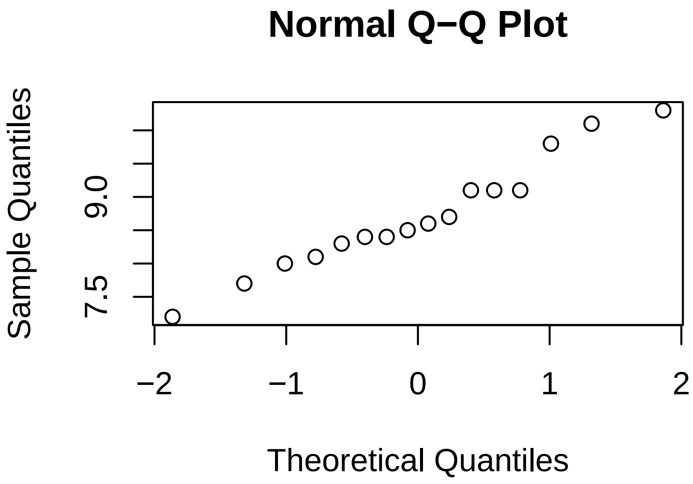
Visually analyze the data with boxplots and qqnorm.

```
boxplot(ss$V1, st$V1, sc$V1, xlab="SS, ST, SC")
```

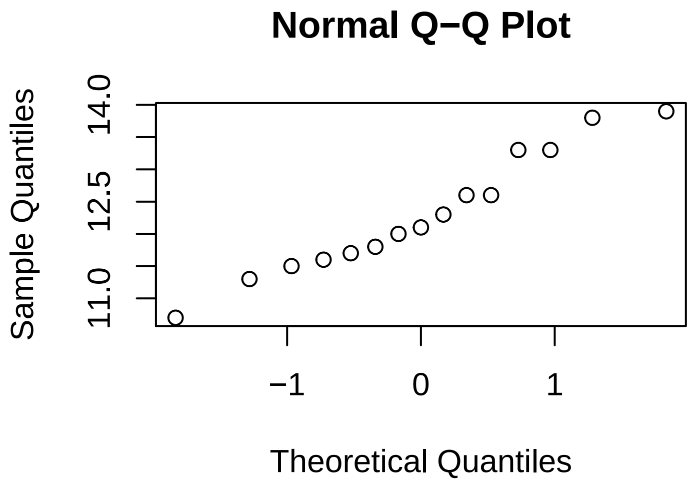


The medians of each variable are clearly different: SS is the lowest, SC is the highest. Let's check normality on the groups:

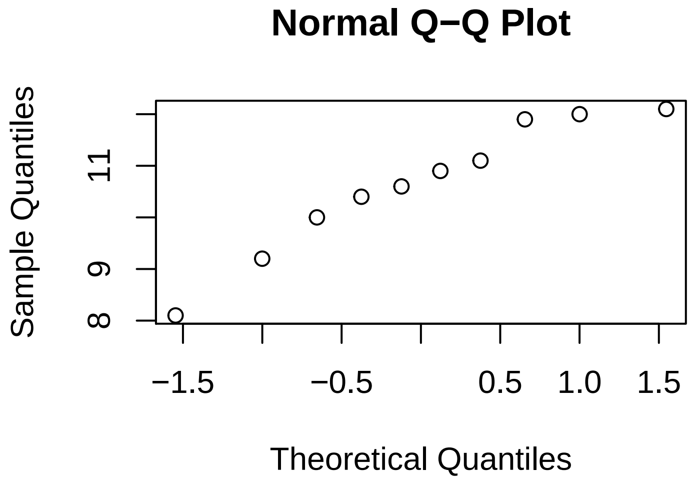
```
qqnorm(ss$V1)
```



```
qqnorm(sc$V1)
```



```r
qqnorm(st$V1)
```



SS: A fairly straight line and no outliers. Normality assumption seems OK here. SC: Similar distribution to the `SS` variable. Some skewness at the right tail, but overall nothing that obviously would violate the normality assumption. ST: Again, no real concerns - we can assume normality on all three variables. Additionally, we have $n = 41$. Since $n \geq 30$, we have a large enough sample to rely on the Central Limit Theorem.

Finally, let's check our assumptions of homoscedasticity with sd():

```r
sd(ss$V1)
```

```
## [1] 0.844492
```

```r
sd(sc$V1)
```

```
## [1] 0.9418826
```

```r
sd(st$V1)
```

```
## [1] 1.284134
```

.84, .94 and 1.28. It's safe to say that the normality assumption is met, but we may have some concerns with an assumption that all three variances are the same. The differences are relatively minor considering the dataset, so we can proceed under the assumption of both normality and homoscedasticity.

2.2

Use ANOVA to test the null hypothesis that the three types of sickle cell disease have the same mean hemoglobin levels. Use a significance level of $\alpha = 0.05$ and organize your calculations in an ANOVA table.

```
mean.ss <- mean(ss$V1)
mean.sc <- mean(sc$V1)
mean.st <- mean(st$V1)

n.ss = length(ss$V1)
n.st = length(st$V1)
n.sc = length(sc$V1)
N = n.ss + n.st + n.sc
```

Total Sum of Squares, Grand mean and Degrees of Freedom:

```
all.s <- p2.data[,1]
grand.mean <- mean(all.s)

SST <- sum((all.s - grand.mean)^2)
total.df = N - 1

SSB <- n.ss * (mean.ss - grand.mean)^2 + n.st * (mean.st - grand.mean)^2 + n.sc * (mean.sc - grand.mean
SSB
```

```
## [1] 99.8893
```

```
#Since we have three variables (SC, ST, and SS) we have 3-1 degrees of freedom.
between.df <- 2

between.meansquare <- SSB/between.df
between.meansquare
```

```
## [1] 49.94465
```

```
SSW <- ((n.ss-1) * var(ss$V1)) + ((n.st - 1) * var(st$V1)) + ((n.sc - 1) * var(sc$V1))
SSW
```

```
## [1] 37.9585
```

```
within.df <- N-3
within.meansquare <- SSW/within.df
within.meansquare
```

```
## [1] 0.9989079
```

Now we can take the ratio of between mean square to within mean square for our F statistic, and use that to calculate our p-value.

```
F <- between.meansquare/within.meansquare
F
```

```
## [1] 49.99926
```

```
1-pf(F, between.df, within.df)
```

```
## [1] 2.281786e-11
```

Again, an extremely small p-value suggests evidence against the null. Let's check the calculations are correct with ANOVA:

```
anova(lm(all.s ~ p2.data$V2))
```

```
## Analysis of Variance Table
##
## Response: all.s
##            Df Sum Sq Mean Sq F value    Pr(>F)    
## p2.data$V2  2 99.889  49.945  49.999 2.282e-11 ***
## Residuals  38 37.959   0.999                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The values look the same using Anova as with our calculations. Additionally, the p-value here is indicated as significantly small enough to reject the null with this data. With our n = 41, this seems like compelling evidence that there is a difference between the three means.

Problem 3

Trosset 12.6 G

Read in the data:

```
p3.data <- read.csv("C:\\Users\\khickman\\Desktop\\Personal\\IUMSDS\\StatsS520\\Module13\\mice.csv")
p3.data
```

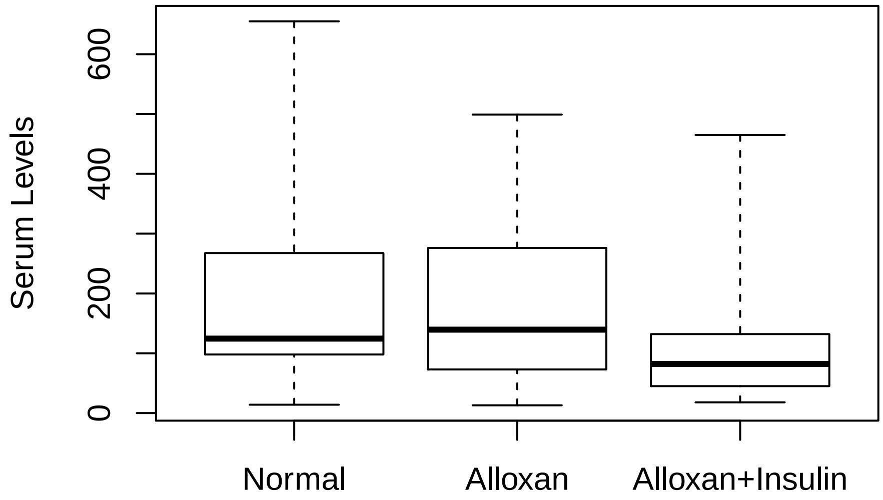
```
##    Value    Class
## 1    156   Normal
## 2    282   Normal
## 3    197   Normal
## 4    297   Normal
## 5    116   Normal
## 6    127   Normal
## 7    119   Normal
## 8     29   Normal
## 9    253   Normal
## 10   122   Normal
## 11   349   Normal
## 12   110   Normal
## 13   143   Normal
## 14    64   Normal
## 15    26   Normal
## 16    86   Normal
## 17   122   Normal
## 18   455   Normal
## 19   655   Normal
## 20    14   Normal
## 21   391  Alloxan
## 22    46  Alloxan
## 23   469  Alloxan
## 24    86  Alloxan
## 25   174  Alloxan
## 26   133  Alloxan
## 27    13  Alloxan
## 28   499  Alloxan
## 29   168  Alloxan
## 30    62  Alloxan
## 31   127  Alloxan
## 32   276  Alloxan
## 33   176  Alloxan
## 34   146  Alloxan
## 35   108  Alloxan
## 36   276  Alloxan
## 37    50  Alloxan
## 38    73  Alloxan
## 39    82 AlloxIns
## 40   100 AlloxIns
## 41    98 AlloxIns
## 42   150 AlloxIns
## 43   243 AlloxIns
## 44    68 AlloxIns
## 45   228 AlloxIns
## 46   131 AlloxIns
## 47    73 AlloxIns
## 48    18 AlloxIns
## 49    20 AlloxIns
## 50   100 AlloxIns
## 51    72 AlloxIns
## 52   133 AlloxIns
## 53   465 AlloxIns
## 54    40 AlloxIns
## 55    46 AlloxIns
## 56    34 AlloxIns
## 57    44 AlloxIns
```

3.1

Using the above data, investigate the ANOVA assumptions of normality and homoscedasticity. Do these assumptions seem plausible for these data? Why or why not?

```
normal <- subset(p3.data, Class=="Normal")
allox <- subset(p3.data, Class=="Alloxan")
alloxins <- subset(p3.data, Class=="AlloxIns")
## check that it worked:
##allox
##alloxins
```

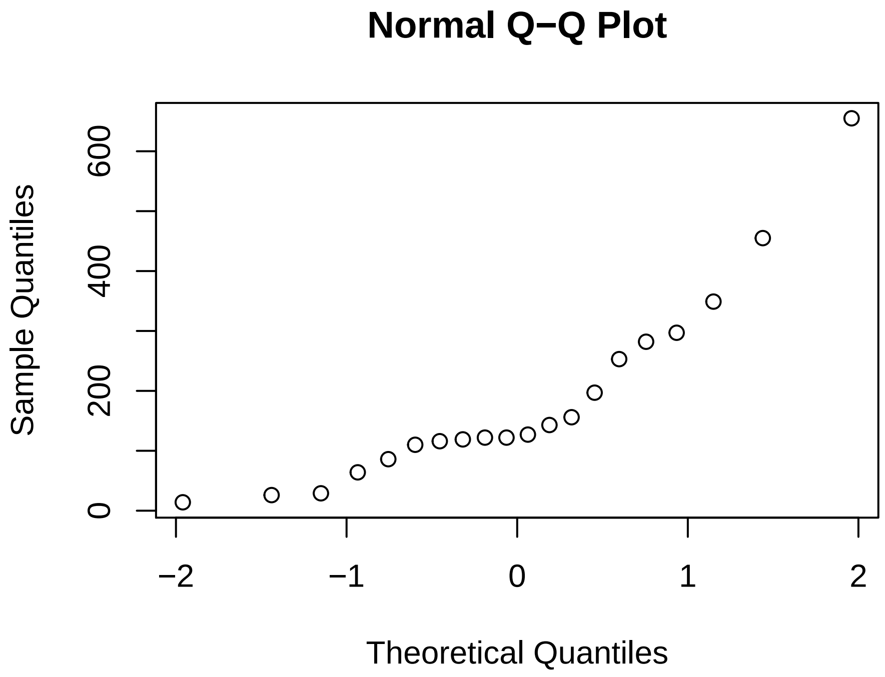
```
boxplot(normal$Value, allox$Value, alloxins$Value, range=0, names = c("Normal", "Alloxan", "Alloxan+Insu
```



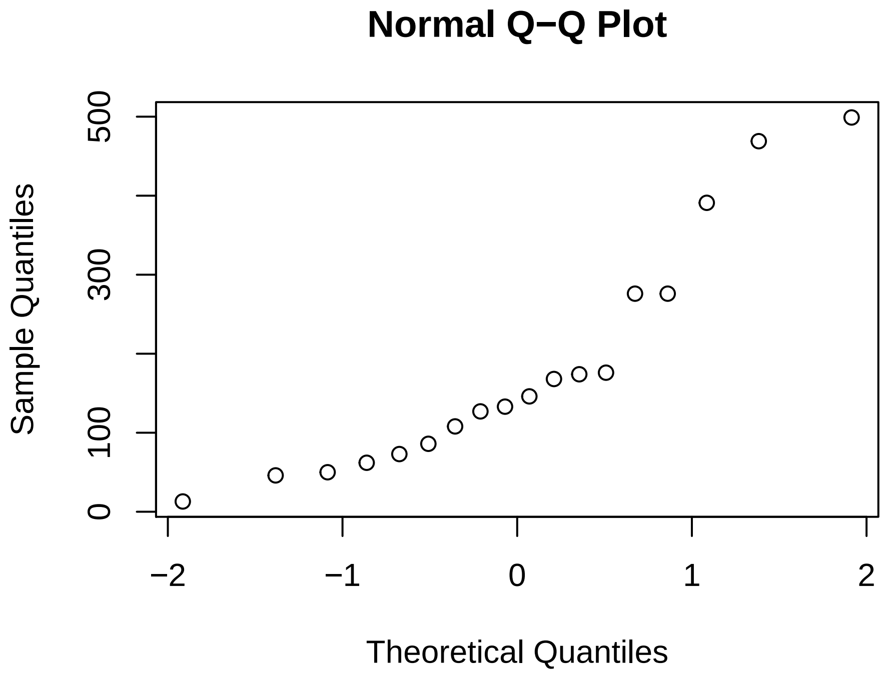
The variables have roughly the same median, and possibly the same variance, though the alloxan-diabetic mice treated with insulin had lower serum levels on average, and at the tails.

Let's check the qqnorm plots:

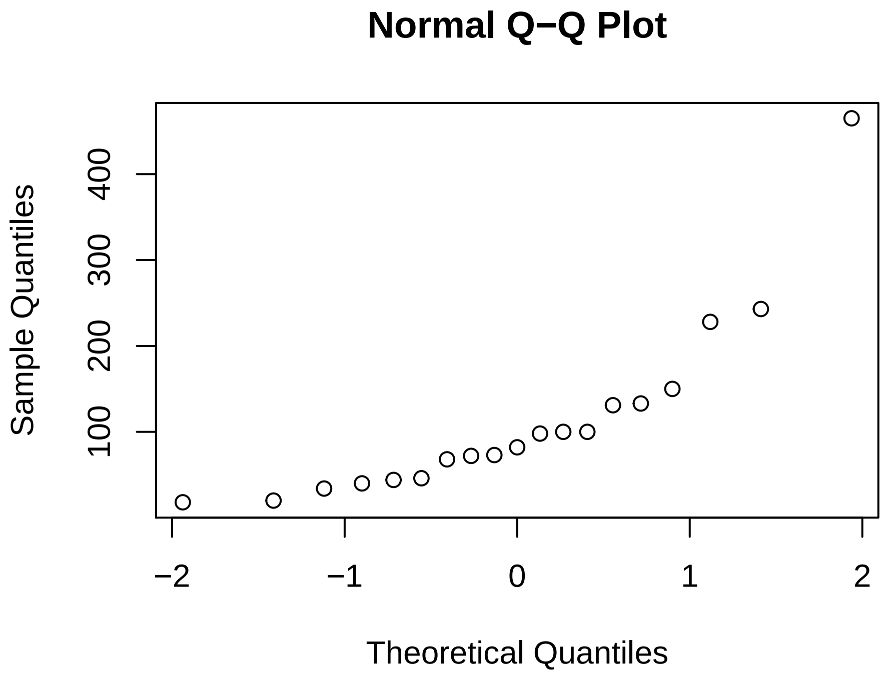
```r
qqnorm(normal$Value)
```



```r
qqnorm(allox$Value)
```



```r
qqnorm(alloxins$Value)
```



Some definitely non-normal characteristics in all three variables.

Let's check homoscedasticity:

```
sd(normal$Value)
```

```
## [1] 158.8349
```

```
sd(allox$Value)
```

```
## [1] 144.8493
```

```
sd(alloxins$Value)
```

```
## [1] 105.7896
```

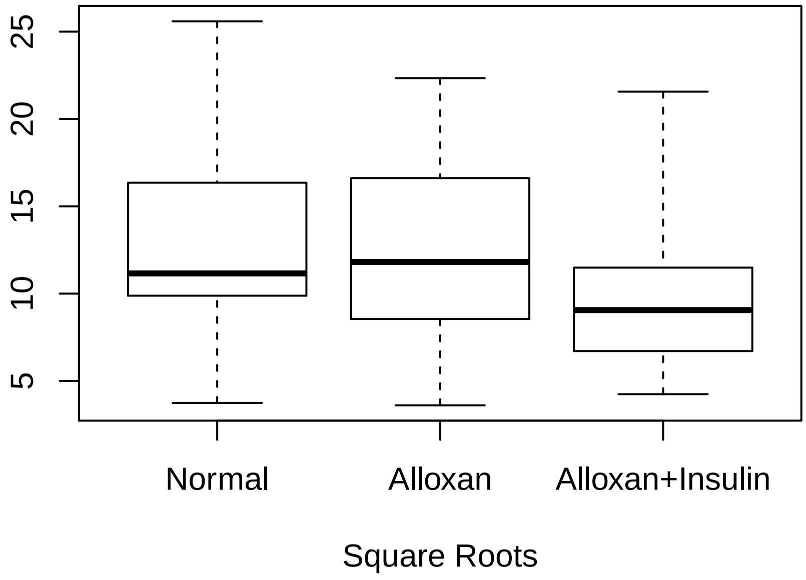
The serum levels in the first two (Normal, Alloxan) are very close, but the third variable, Alloxan+Insulin has noticeably less variance. We can assume normality, but homoscedasticity might be a stretch.

3.2.

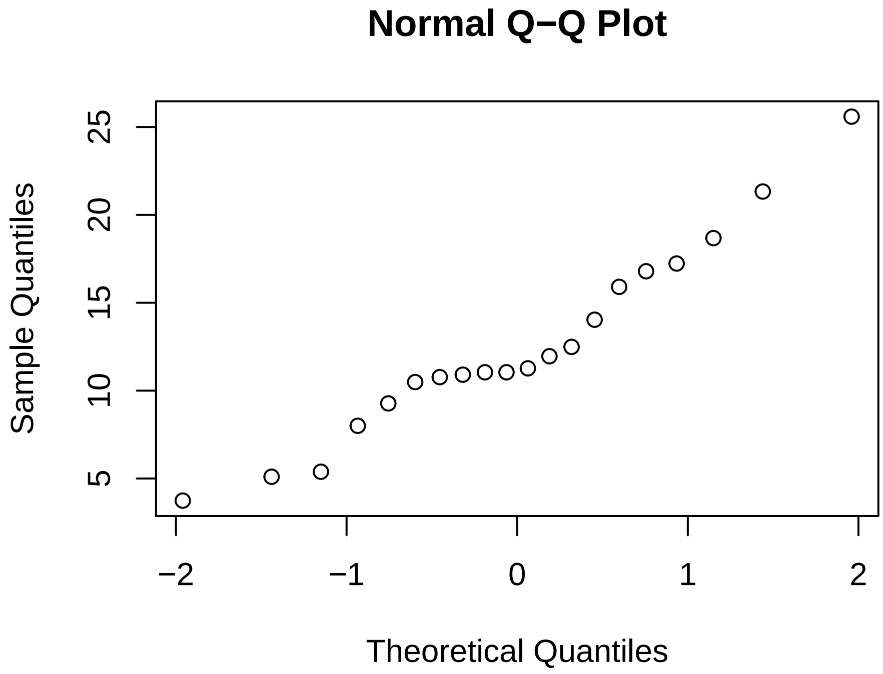
Now transform the data by taking the square root of each measurement. Using the transformed data, investigate the ANOVA assumptions of normality and homoscedasticity. Do these assumptions seem plausible for the transformed data? Why or why not?

```
norm.sq <- sqrt(normal$Value)
allox.sq <- sqrt(allox$Value)
alloxins.sq <- sqrt(alloxins$Value)
```

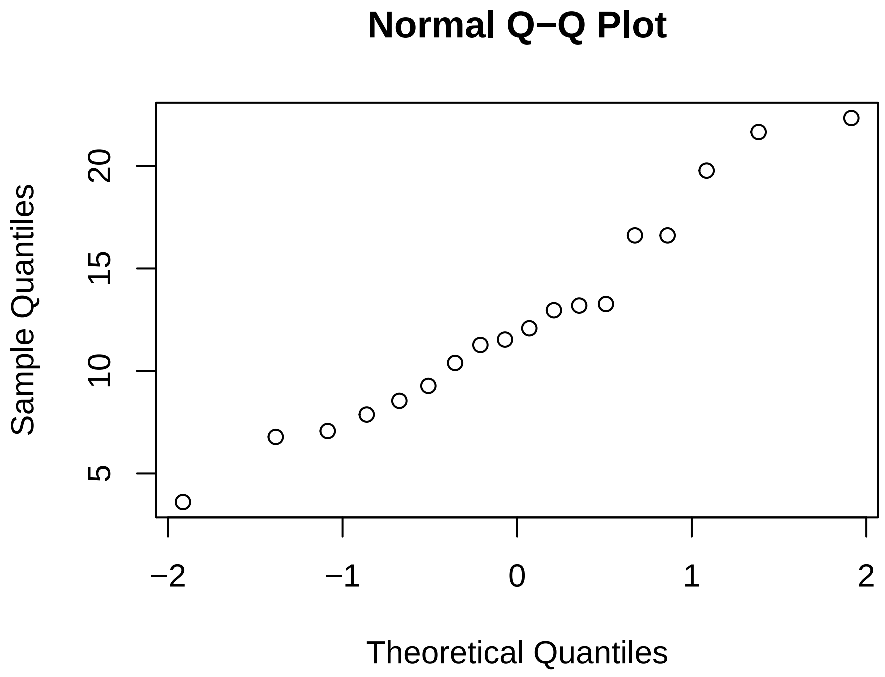
```
boxplot(norm.sq, allox.sq, alloxins.sq, range=0,
        names=c("Normal", "Alloxan", "Alloxan+Insulin"),
        xlab="Square Roots")
```



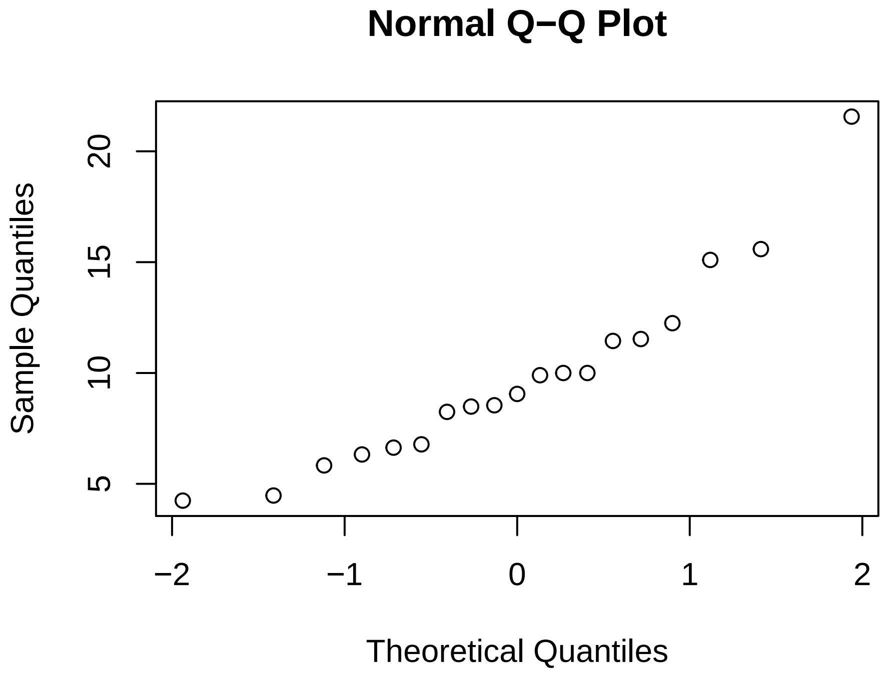
```
qqnorm(norm.sq)
```



```
qqnorm(allox.sq)
```



```r
qqnorm(alloxins.sq)
```



These qqnorm plots indicate that the transformed variables are much closer to normal. There is one apparent outliers with the alloxan+insulin variable, but the normality assumption is safe here.

Let's investigate homoscedasticity:

```
sd(norm.sq)
```

```
## [1] 5.480753
```

```
sd(allox.sq)
```

```
## [1] 5.226939
```

```
sd(alloxins.sq)
```

```
## [1] 4.244474
```

Again, our third variable is slightly lower, but the transformed variances are fairly close. We can proceed under the assumptions required by ANOVA.

3.3

Using the transformed data, construct an ANOVA table. State the null and alternative hypotheses tested by this method. Should the null hypothesis be rejected at the $\alpha = 0.05$ level?

```
all.v <- c(norm.sq, allox.sq, alloxins.sq)
all.v
```

```
##  [1] 12.489996 16.792856 14.035669 17.233688 10.770330 11.269428 10.908712
##  [8]  5.385165 15.905974 11.045361 18.681542 10.488088 11.958261  8.000000
## [15]  5.099020  9.273618 11.045361 21.330729 25.592968  3.741657 19.773720
## [22]  6.782330 21.656408  9.273618 13.190906 11.532563  3.605551 22.338308
## [29] 12.961481  7.874008 11.269428 16.613248 13.266499 12.083046 10.392305
## [36] 16.613248  7.071068  8.544004  9.055385 10.000000  9.899495 12.247449
## [43] 15.588457  8.246211 15.099669 11.445523  8.544004  4.242641  4.472136
## [50] 10.000000  8.485281 11.532563 21.563859  6.324555  6.782330  5.830952
## [57]  6.633250
```

```
anova(lm(all.v ~ p3.data$Class))
```

```
## Analysis of Variance Table
##
## Response: all.v
##               Df  Sum Sq Mean Sq F value Pr(>F)
## p3.data$Class  2   94.74  47.368  1.8815 0.1622
## Residuals     54 1359.47  25.175
```

The null hypothesis, $H_0 : \mu_1 = \mu_2 = \mu_3$ where Normal, Alloxan, and Alloxan+Insulin are μ_1, μ_2, and μ_3 respectively, cannot be rejected with a p-value of .16, which is well above our significance level of .05, so it appears that there is no difference between the means and the treatment does not have an effect.

3.4.

Using the transformed data, construct suitable contrasts for investigating the research questions framed above. State appropriate null and alternative hypotheses and test them using the method of Bonferroni t-tests. At what significance level should these hypotheses be tested in order to maintain a family rate of Type I error equal to 5%? Which null hypotheses should be rejected? We would like to use Bonferroni tests to compare the following variables:

Normal : Alloxan

Alloxan : alloxan + insulin

Alloxan + insulin : normal

At $\alpha = .05$. To find the appropriate significance level, we take .05/3, and will look for an average significance level of .0167 across three Welch's t-tests.

Normal : Alloxan $H_0 : \mu_1 - \mu_2 = 0$ where the expected value of the antibody response of normal mice is μ_1 and the expected response of the antibody response of mice with alloxan diabetes is μ_2. We can conduct a Welch's t-test. The alternative hypothesis $H_1 : \mu_1 - \mu_2 \neq 0$.

```
t.test(normal$Value, allox$Value)
```

```
##
##  Welch Two Sample t-test
##
## data:  normal$Value and allox$Value
## t = 0.086606, df = 35.991, p-value = 0.9315
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -95.64859 104.18192
## sample estimates:
## mean of x mean of y
##  186.1000  181.8333
```

Our p-value here is .93, which is not small, and we cannot therefore reject the null here.

Moving on to the second contrast, Alloxan : Alloxan-Insulin: $H_0 : \mu_1 - \mu_2 = 0$ where the expected value of the antibody response of alloxan diabetic mice is μ_1 and the expected response of the antibody response of mice with alloxan diabetes treated with insulin is μ_2. We can conduct a Welch's t-test. The alternative hypothesis $H_1 : \mu_1 - \mu_2 \neq 0$.

```
t.test(allox$Value, alloxins$Value)
```

```
##
##  Welch Two Sample t-test
##
## data:  allox$Value and alloxins$Value
## t = 1.6458, df = 31.037, p-value = 0.1099
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -16.48957 154.36676
## sample estimates:
## mean of x mean of y
##  181.8333  112.8947
```

Again, the p-value is .10, which means we do not have enough evidence to reject the null for this contrast.

Finally, comparing the Alloxan-diabetic mice treated with insulin to the Normal mice (which is probably our main contrast of interest). Our null - $H_0 : \mu_1 - \mu_2 = 0$ where the expected value of the antibody response of normal mice is μ_1 and the expected response of the antibody response of mice with alloxan diabetes treated with insulin is μ_2. We can conduct a Welch's t-test. The alternative hypothesis $H_1 : \mu_1 - \mu_2 \neq 0$.

```
t.test(normal$Value, alloxins$Value)
```

```
##
##  Welch Two Sample t-test
##
## data:  normal$Value and alloxins$Value
## t = 1.7018, df = 33.237, p-value = 0.09813
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -14.28944 160.69996
## sample estimates:
## mean of x mean of y
##  186.1000  112.8947
```

The smallest of the p-values yet, though still not small at .09. We have failed to reject null hypotheses in all three contrasts.