stat359_A3_wducharme

Wesley Ducharme

2024-11-04

Question 1

```r
running_times <- read.csv(file = 'C:/Users/wesch/uvic/stat359 Data
↪  Analysis/Assignments/A3/film_run_times.csv')

attach(running_times)
summary(running_times)
```

```
##     Company1        Company2    
##  Min.   : 86.0   Min.   : 81.0  
##  1st Qu.: 92.0   1st Qu.: 89.5  
##  Median : 98.0   Median : 97.0  
##  Mean   : 97.4   Mean   :110.0  
##  3rd Qu.:102.0   3rd Qu.:124.0  
##  Max.   :109.0   Max.   :165.0  
##  NA's   :2
```

Graphical summaries

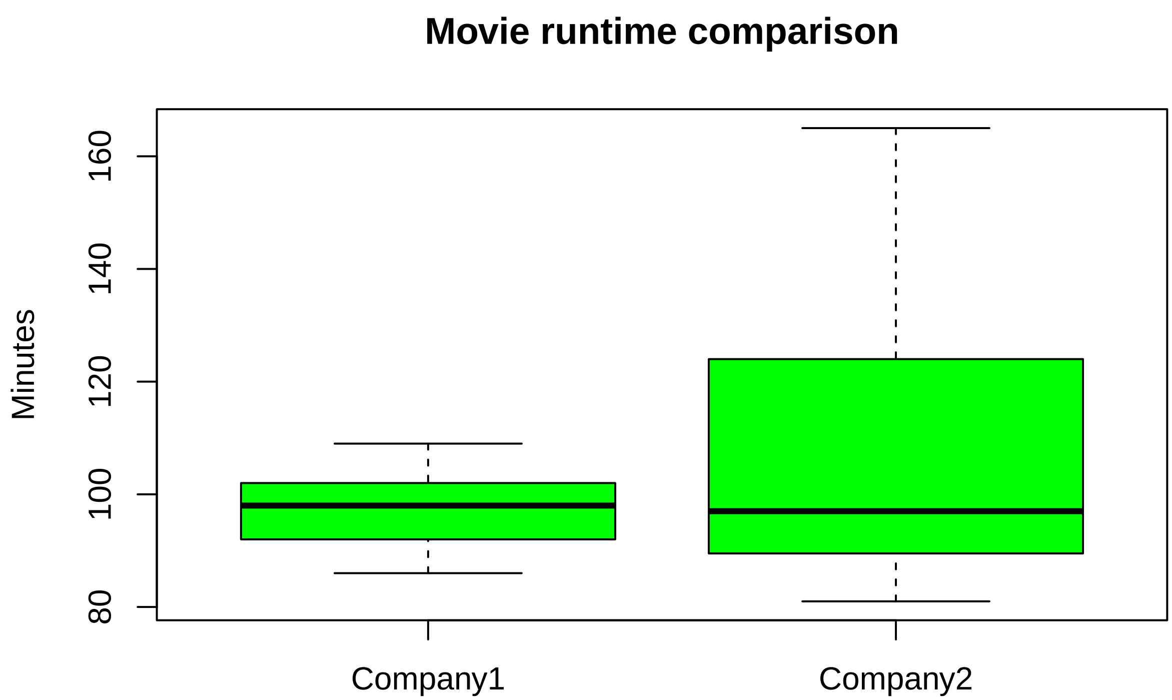
```r
library(sn)
```

```
## Warning: package 'sn' was built under R version 4.3.3

## Loading required package: stats4

## 
## Attaching package: 'sn'

## The following object is masked from 'package:stats':
## 
##     sd
```

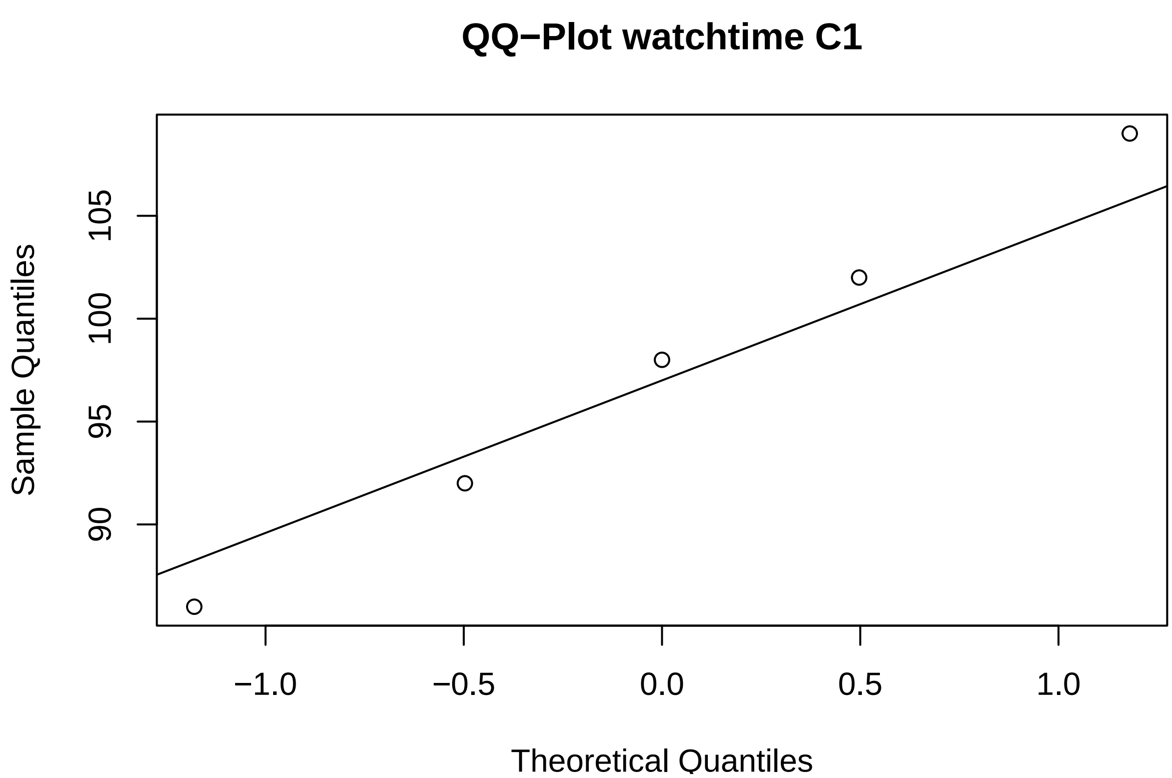
```r
class(running_times)
```

```
## [1] "data.frame"
```

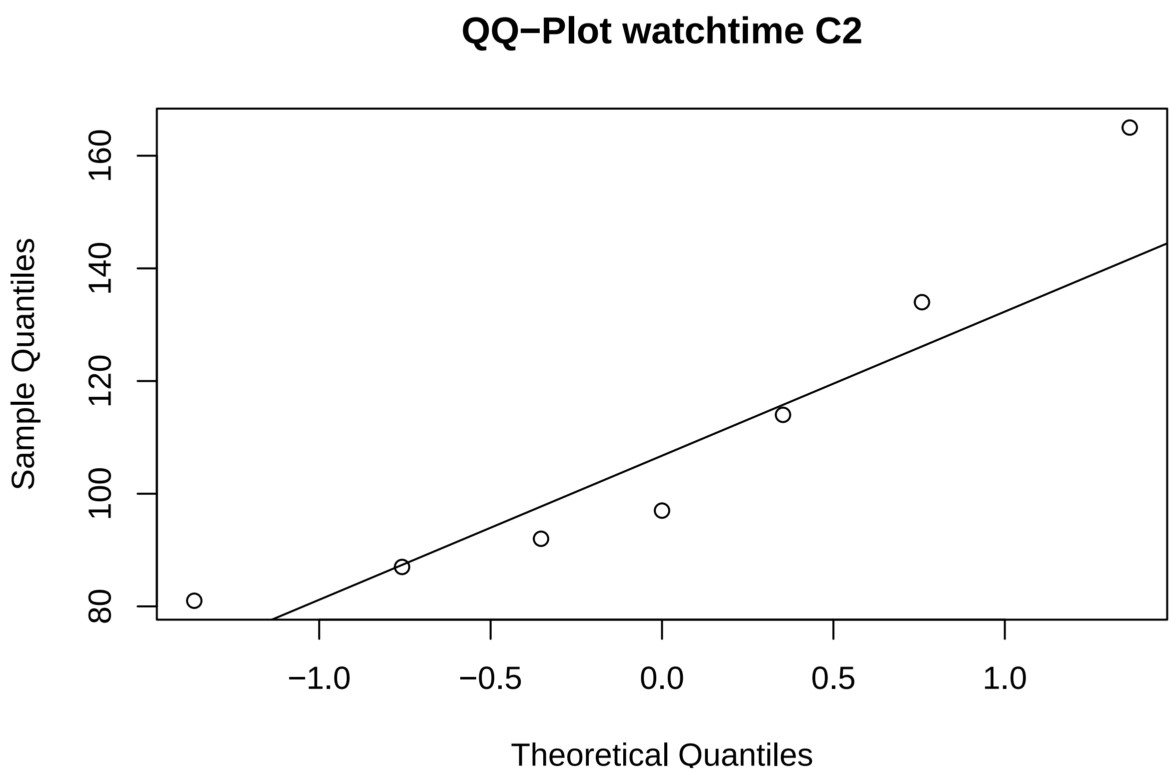
```r
boxplot(running_times, col = "green", ylab = "Minutes")
title("Movie runtime comparison")
```




```r
qqnorm(running_times$Company1, main="QQ-Plot watchtime C1")
qqline(running_times$Company1)
```



```
qqnorm(running_times$Company2, main="QQ-Plot watchtime C2")
qqline(running_times$Company2)
```



Judging from the qq-plots we can assume that the observations for company 1 and company 2. Company 2 does not follow an approximately normal distribution. Company 2's data looks to be right skewed. Looking at the box-plots we can see that may not be safe to assume that their variances are equal. So we will run a test for this.

```
c1<-running_times$Company1[!is.na(running_times$Company1)]
c2<-running_times$Company2[!is.na(running_times$Company2)]
var.test(c1,c2)
```

```
##
##  F test to compare two variances
##
## data:  c1 and c2
## F = 0.086277, num df = 4, denom df = 6, p-value = 0.03298
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.01385501 0.79351983
## sample estimates:
## ratio of variances
##         0.08627737
```

P-value = 0.03298 Therefore we have strong evidence against the Null hypothesis supporting the assumption that they have equal variances. Therefore it is safe to assume that their variances are not equal.

Based on the above analysis a Wilcox rank-sum test will be performed to test the hypothesis that Company 2's average run time is 10 minutes or longer than company 1's average run time.

I will adjust company 2's data by subtracting 10 minutes from each of company 2's observations. Before performing the test. This will then test whether heir means are equal when Company 2's average is 10 less it's actual value.

```
c2_adjusted <- c2 - 10

#test
wilcox.test(c1, c2_adjusted, alternative="less")
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  c1 and c2_adjusted
## W = 20, p-value = 0.6806
## alternative hypothesis: true location shift is less than 0
```

P-value is 0.68 which means we have little or no evidence against the Null hypothesis for a significance level of 0.1. So we can conclude that Company 2's average movie running time is 10 minutes greater than company 1's.

Question 2

```
tensile_strength <- read.csv(file = 'C:/Users/wesch/uvic/stat359 Data
↪  Analysis/Assignments/A3/machine_seal_strength.csv')

attach(tensile_strength)
summary(tensile_strength)
```

```
##  TensileStrength   Machines
##  Min.   :14.60   Length:24
##  1st Qu.:16.48   Class :character
##  Median :17.60   Mode  :character
##  Mean   :17.79
##  3rd Qu.:18.98
##  Max.   :20.80
```

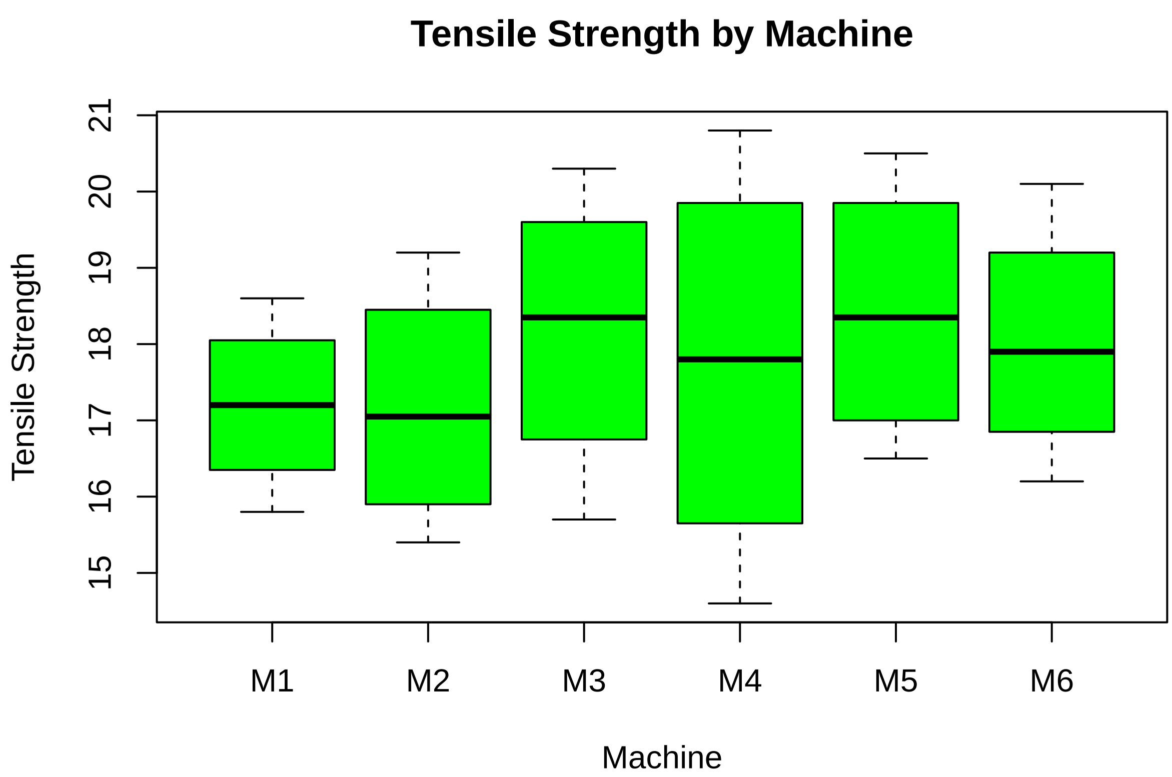
```
kable(tensile_strength, caption = 'Tensile Strength',align='l')
```

Table 1: Tensile Strength

TensileStrength	Machines
17.5	M1
16.9	M1
15.8	M1
18.6	M1
16.4	M2
19.2	M2
17.7	M2

TensileStrength	Machines
15.4	M2
20.3	M3
15.7	M3
17.8	M3
18.9	M3
14.6	M4
16.7	M4
20.8	M4
18.9	M4
17.5	M5
19.2	M5
16.5	M5
20.5	M5
18.3	M6
16.2	M6
17.5	M6
20.1	M6

```r
boxplot(TensileStrength ~ Machines, data = tensile_strength,
        xlab = "Machine", ylab = "Tensile Strength",
        main = "Tensile Strength by Machine", col = "green")
```




```r
heights<-tapply(TensileStrength,Machines,mean)
heights
```

```
##       M1       M2       M3       M4       M5       M6
```

```
## 17.200 17.175 18.175 17.750 18.425 18.025
```

```
barplot(heights, col="green", ylim=c(0,25),ylab="mean tensile strength", xlab = "Machine
↪  Number")

## Error bars function
error.bars<-function(y,z){
  x<-barplot(y, plot=F)
  n<-length(y)
  for (i in 1:n)
  {
    arrows(x[i],y[i]-z[i],x[i],y[i]+z[i],code=3,angle=90,length=0.15)
  }
}

## adding the error bars to the plot with a 95% CI
sigma.hat<-summary.lm(aov(TensileStrength~Machines))$sigma
sigma.hat
```

```
## [1] 1.865476
```

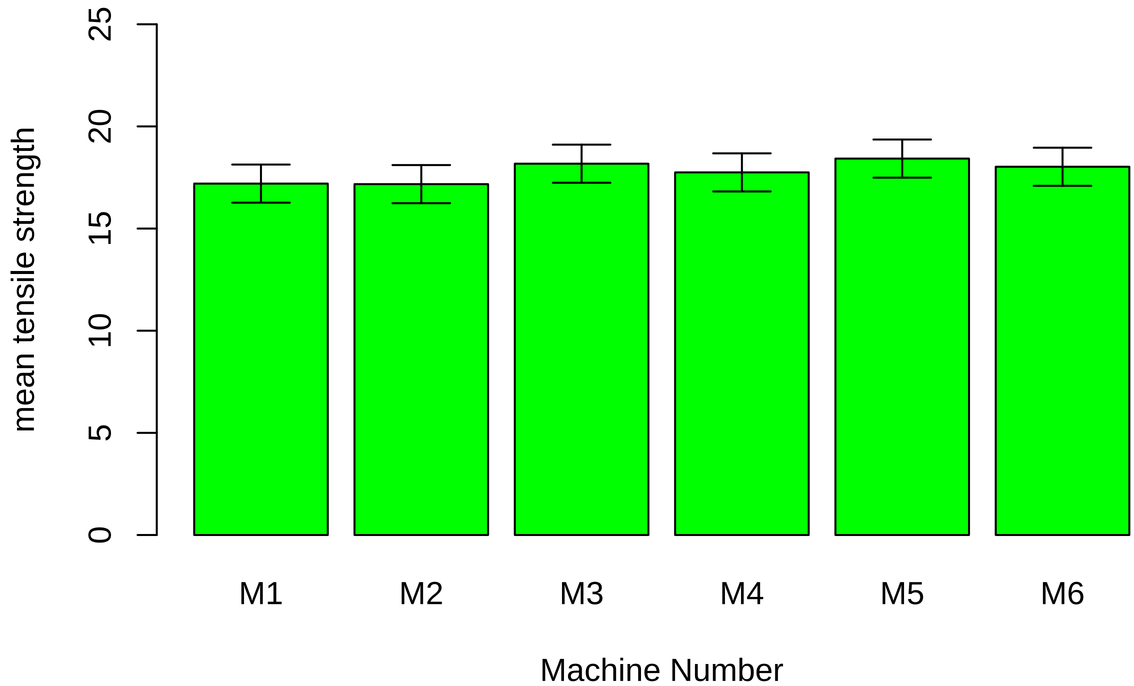
```
table(Machines)
```

```
## Machines
## M1 M2 M3 M4 M5 M6
##  4  4  4  4  4  4
```

```
se.mean<-sigma.hat/sqrt(4)
se.mean
```

```
## [1] 0.9327379
```

```
bar.half.width<-rep(se.mean,6)
error.bars(heights,bar.half.width)
```



From the box-plot above it appears that it is safe to assume that the variances are equal. From the the bar plot above it is observed that there doesn't seem to be a significant difference between their means based on the overlap of the error bars. This is not conclusive evidence however and so the next step is to perform and Analysis of Variance test. The Null hypothesis for this test with be; Each machines mean tensile strength is the same. This will be tested at a 0.05 significance level.

ANOVA test

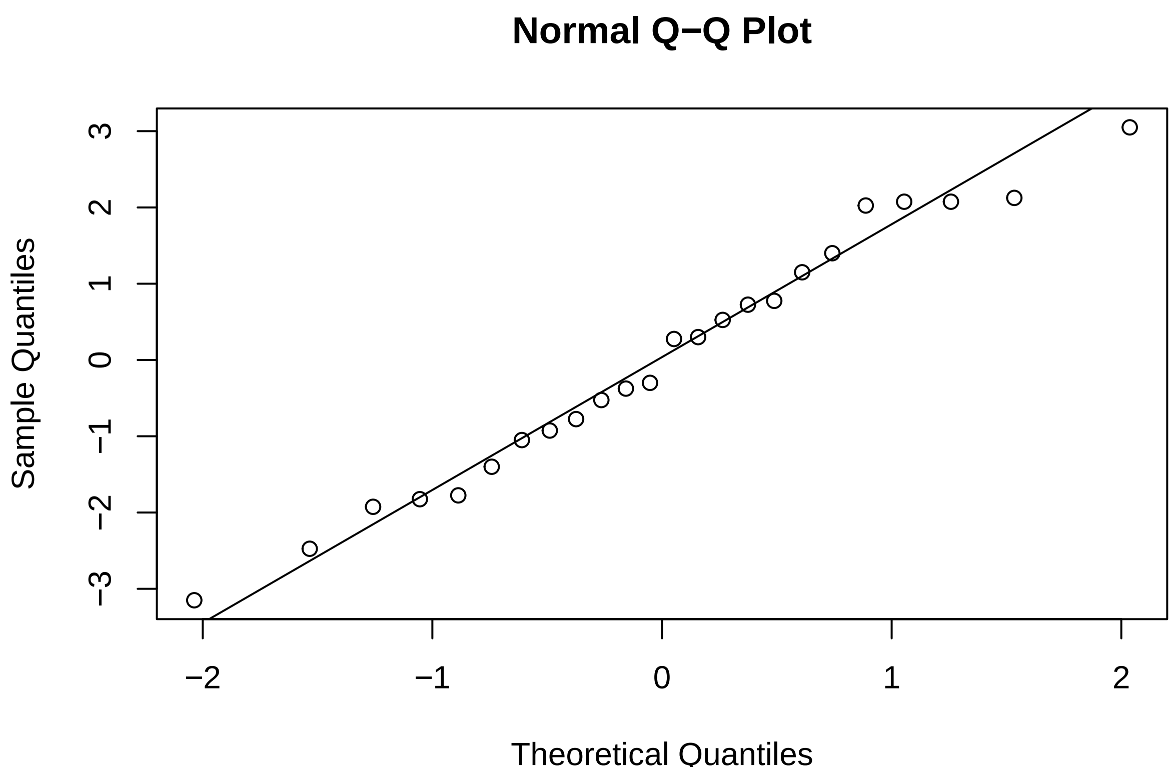
```
anova_result <- aov(TensileStrength ~ Machines)
summary(anova_result)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Machines     5   5.34   1.068   0.307  0.902
## Residuals   18  62.64   3.480
```

With a p-value of 0.902 we can say there is no evidence against the Null hypothesis, so we can't reject it. Thus we can conclude that the mean of tensile strength for each machine is not significantly different.

Double checking residuals

```
qqnorm(residuals(anova_result))
qqline(residuals(anova_result))
```



Residuals look good as they follow an approximately normal distribution.

Question 3

For clarification “HoR” in the data stands for “Hours of Relief”.

```r
tablet_relief <- read.csv(file = 'C:/Users/wesch/uvic/stat359 Data
↪  Analysis/Assignments/A3/tablet_hours_of_relief.csv')
attach(tablet_relief)
summary(tablet_relief)
```

```
##       HoR            Tablet
##  Min.   :1.000   Length:25
##  1st Qu.:3.400   Class :character
##  Median :5.800   Mode  :character
##  Mean   :5.516
##  3rd Qu.:7.200
##  Max.   :9.300
```

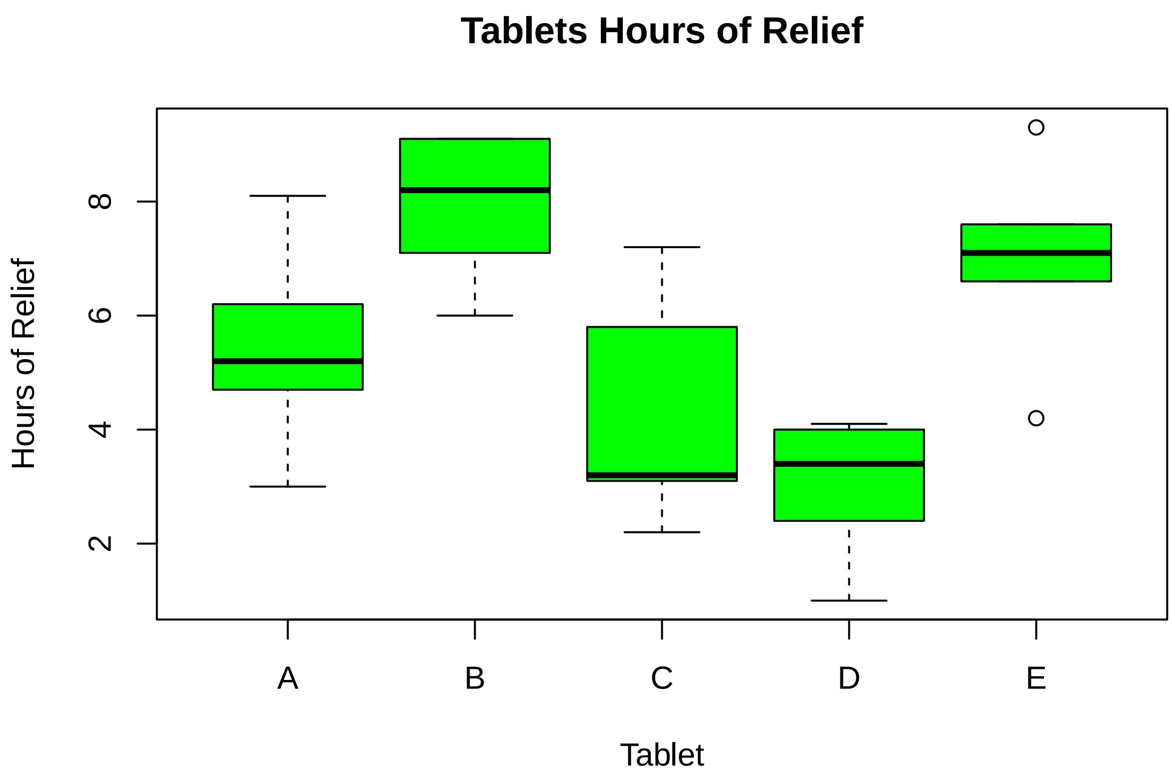
```r
kable(tablet_relief, caption = 'Tablet Hours of Relief',align='l')
```

Table 2: Tablet Hours of Relief

HoR	Tablet
5.2	A
4.7	A

HoR	Tablet
8.1	A
6.2	A
3.0	A
9.1	B
7.1	B
8.2	B
6.0	B
9.1	B
3.2	C
5.8	C
2.2	C
3.1	C
7.2	C
2.4	D
3.4	D
4.1	D
1.0	D
4.0	D
7.1	E
6.6	E
9.3	E
4.2	E
7.6	E

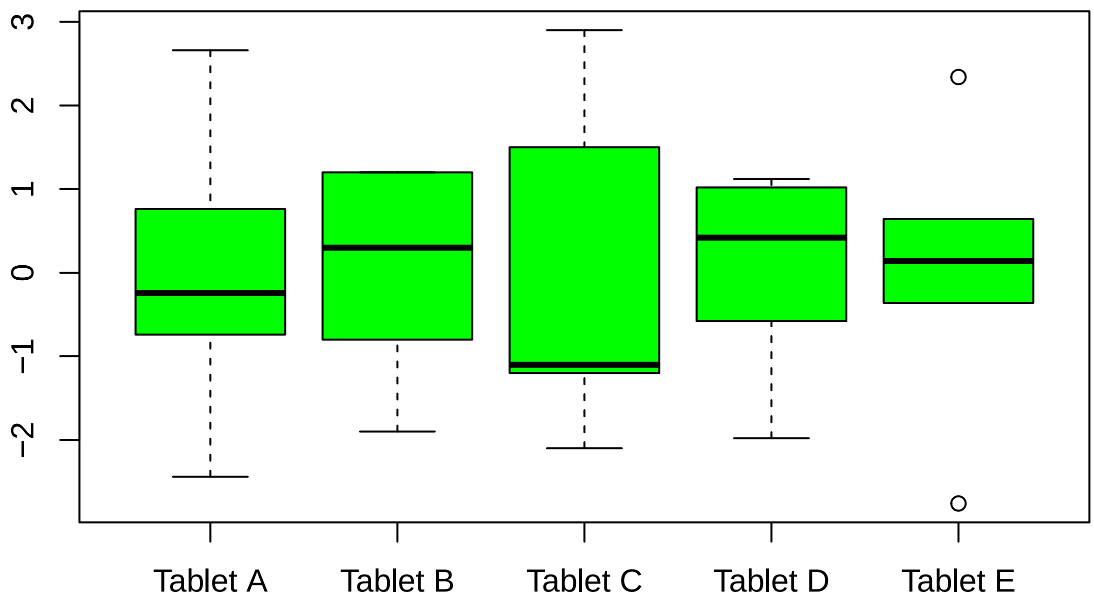
```
# Boxplot to get a first look at the variances
boxplot(HoR ~ Tablet, data = tablet_relief,
        xlab = "Tablet", ylab = "Hours of Relief",
        main = "Tablets Hours of Relief", col = "green")
```



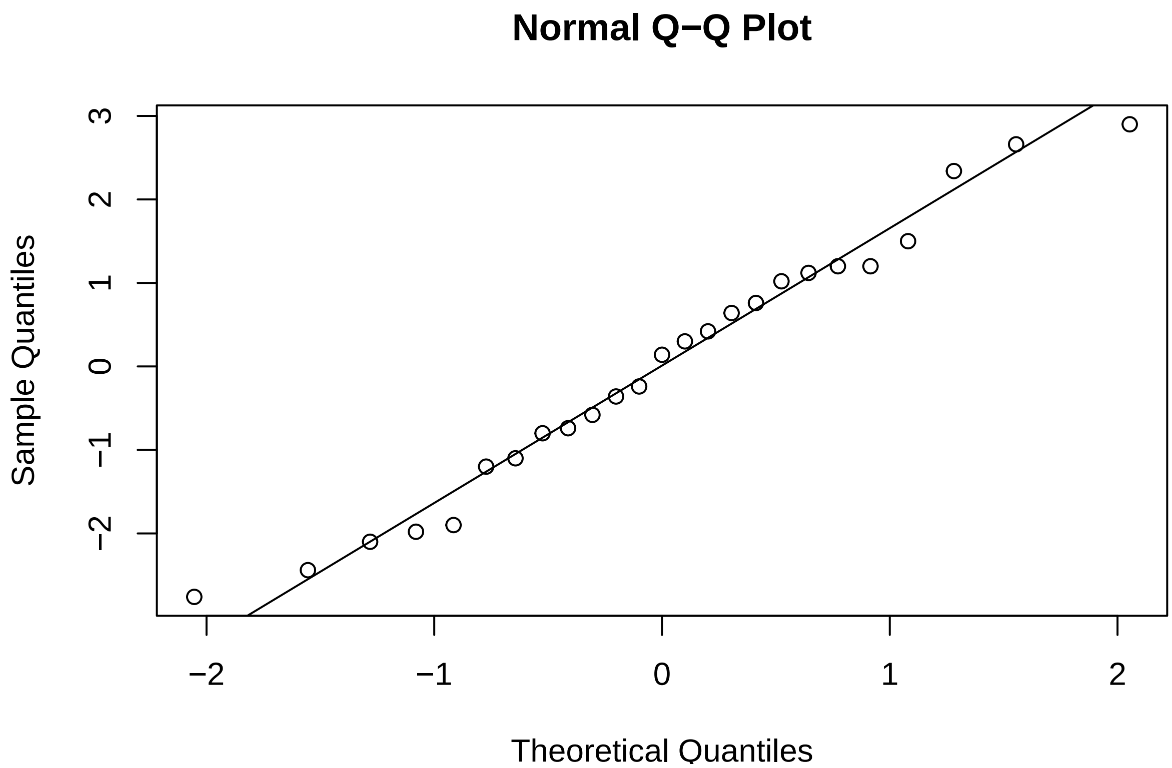
Variances may not be approximately equal based on the box-plot above.

Will check the assumptions of ANOVA by checking for approximatly equal variances across the residuals and normality.

```
resid.Tablet<-resid(aov(HoR~Tablet))
boxplot(resid.Tablet[Tablet=='A'],resid.Tablet[Tablet=='B'],
        resid.Tablet[Tablet=='C'],resid.Tablet[Tablet=='D'],
        resid.Tablet[Tablet=='E'],
        names=c('Tablet A','Tablet B','Tablet C','Tablet D','Tablet E'),
        col="green")
```



```r
qqnorm(resid.Tablet)
qqline(resid.Tablet)
```



Box-plot appears to show approximately equal variance across the data and the qq-plot shows that the data is normal distributed.

Will now perform the ANOVA test on the H0: Mean hours of relief are the equal across all Tablets.

```r
anova_result_HoR <- aov(HoR ~ Tablet)
summary(anova_result_HoR)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Tablet       4  78.42  19.605   6.587 0.0015 **
## Residuals   20  59.53   2.977
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value = 0.0015

There is very strong evidence against H0. So we can conclude that the mean hours of relief is not the same across all the different tablets.

Barplot to visualize the means.

```r
heights<-tapply(HoR,Tablet,mean)
heights
```

```
##    A    B    C    D    E
## 5.44 7.90 4.30 2.98 6.96
```

```r
barplot(heights, col="green", ylim=c(0,25),ylab="Mean Hours of Relief", xlab = "Tablets")

## Error bars function
error.bars<-function(y,z){
  x<-barplot(y, plot=F)
  n<-length(y)
  for (i in 1:n)
  {
    arrows(x[i],y[i]-z[i],x[i],y[i]+z[i],code=3,angle=90,length=0.15)
  }
}

## adding the error bars to the plot with a 95% CI
sigma.hat<-summary.lm(aov(HoR~Tablet))$sigma
sigma.hat
```

```
## [1] 1.725283
```

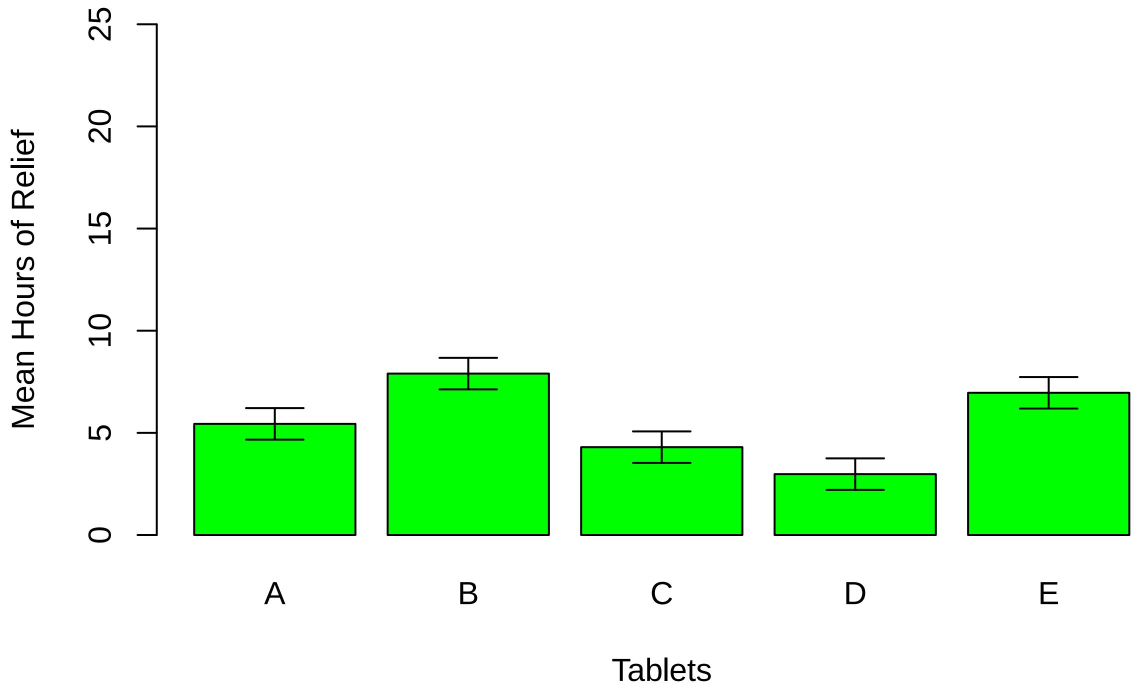
```r
table(Tablet)
```

```
## Tablet
## A B C D E
## 5 5 5 5 5
```

```r
se.mean<-sigma.hat/sqrt(5)
se.mean
```

```
## [1] 0.7715698
```

```r
bar.half.width<-rep(se.mean,5)
error.bars(heights,bar.half.width)
```



Bar-plot shows that our test results make sense and it is clear that there are unequal means.