# Assignment 3

Alexander Williams

2024-11-07

## 1.

First, create the dataframe

```
films <- data.frame(time=c(102,86,98,109,92,  81,165,97,134,92,87,114),
                    company=c(1, 1, 1, 1,  1,   2, 2,  2, 2,  2, 2, 2))
attach(films)
```

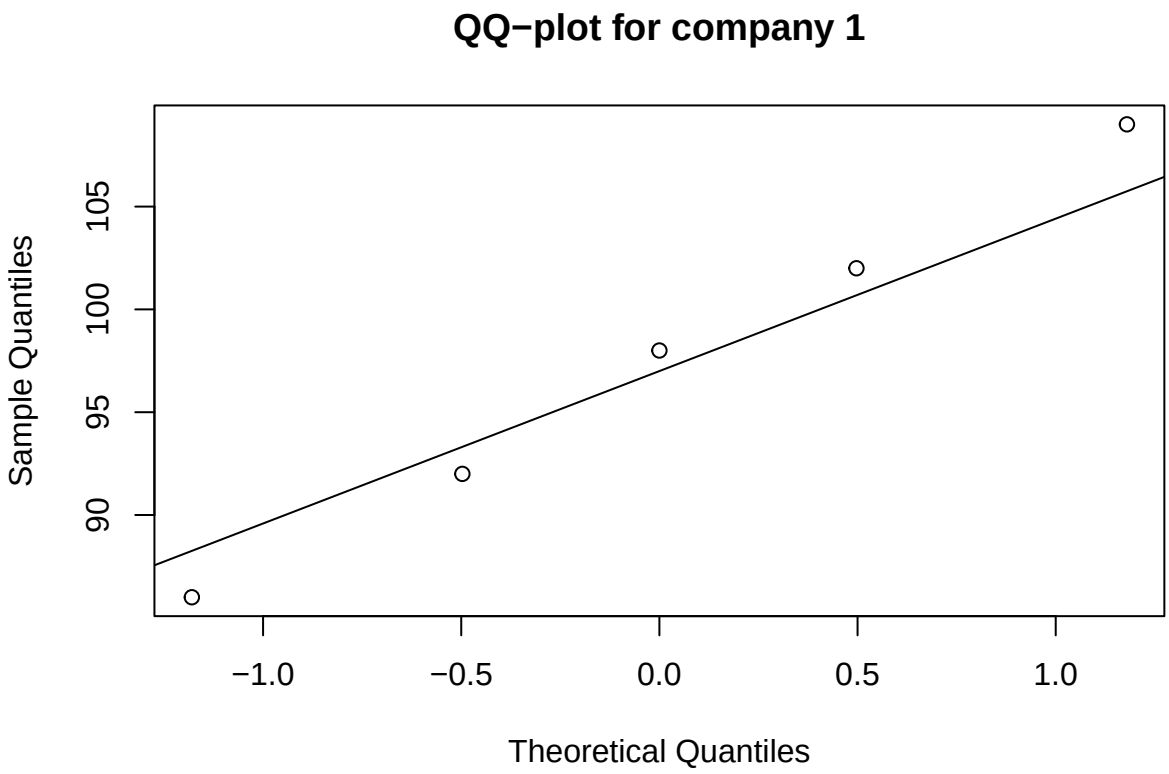
first, check if they have the same variance

```
var.test(time[company == 1], time[company == 2])
```

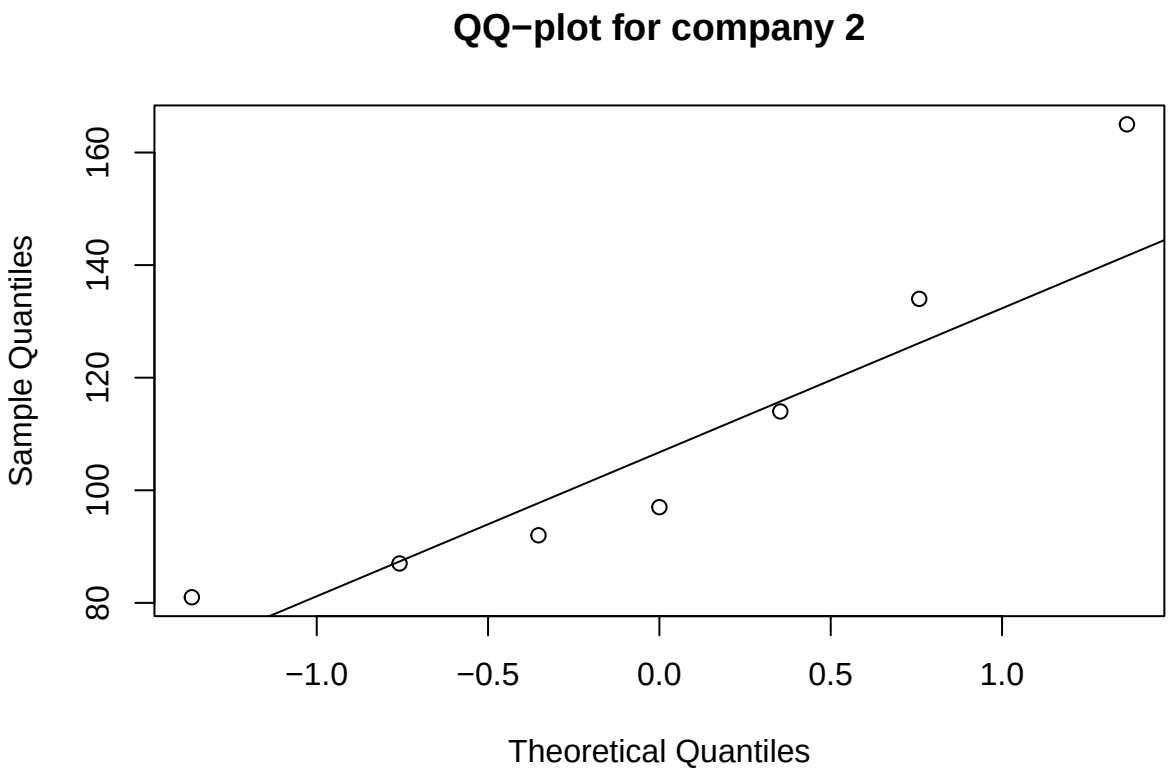
```
##
##  F test to compare two variances
##
## data:  time[company == 1] and time[company == 2]
## F = 0.086277, num df = 4, denom df = 6, p-value = 0.03298
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.01385501 0.79351983
## sample estimates:
## ratio of variances
##          0.08627737
```

since the p-value = 0.03, there's evidence that the two companies have different variances. We'll have to keep that in mind for our tests.

```
qqnorm(time[company == 1], main='QQ-plot for company 1')
qqline(time[company == 1])
```




```
qqnorm(time[company == 2], main='QQ-plot for company 2')
qqline(time[company == 2])
```




While the QQ-plot for company one does appear normal, the plot for company 2 appears to have less weight in the upper values. Let's test the skew value to see if it could be normal. While we're at it, lets check kurtosis as well.

```
skew <- function(x)
{
  if (var(x) == 0)
    return(0)
  m3 <- sum((x-mean(x))^3) / length(x)
  s3 <- sqrt(var(x))^3
  m3 / s3
}
kurtosis <- function(x)
{
  if (var(x) == 0)
    return(0)
  m4 <- sum((x-mean(x))^4) / length(x)
  s4 <- sqrt(var(x))^4
  m4 / s4 - 3
}

B <- 1000
bootstrap <- matrix(data=sample(time[company == 2],
                                size = length(time[company == 2]) * B,
                                replace=TRUE),
                    nrow = length(time[company == 2]),
                    ncol = B)
```

```
boot.skew <- apply(bootstrap,
                   2,
                   skew)
boot.kurt <- apply(bootstrap,
                   2,
                   kurtosis)
quantile(boot.skew, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## -0.338231  1.485299
```

```
quantile(boot.kurt, c(0.025, 0.975))
```

```
##       2.5%      97.5%
## -2.1558339  0.5424375
```

there's no reason to believe that either skew or kurtosis aren't equal to zero, so we can proceed with the assumption that company 2's times are normally distributed. For the sake of completeness, lets do the same for company 1's times

```
bootstrap <- matrix(data=sample(time[company == 1],
                                size = length(time[company == 1]) * B,
                                replace=TRUE),
                    nrow = length(time[company == 1]),
                    ncol = B)

boot.skew <- apply(bootstrap,
                   2,
                   skew)
boot.kurt <- apply(bootstrap,
                   2,
                   kurtosis)
quantile(boot.skew, c(0.025, 0.975))
```

```
##       2.5%      97.5%
## -0.8596392  0.8153360
```

```
quantile(boot.kurt, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## -2.253333 -0.920000
```

For company 1, there's no evidence that skew isn't zero, but the kurtosis implies more weight in the center of the distribution. However, it's only a slight difference, so I'm willing to move forward with the assumption that both populations are normally distributed.

Since we can assume normality but can't assume they share the same variance, we can use an unpooled t-test to compare the two populations.

$H_0 : \hat{\mu}_2 - \hat{\mu}_1 \geq 10$

$H_1 : \hat{\mu}_2 - \hat{\mu}_1 < 10$

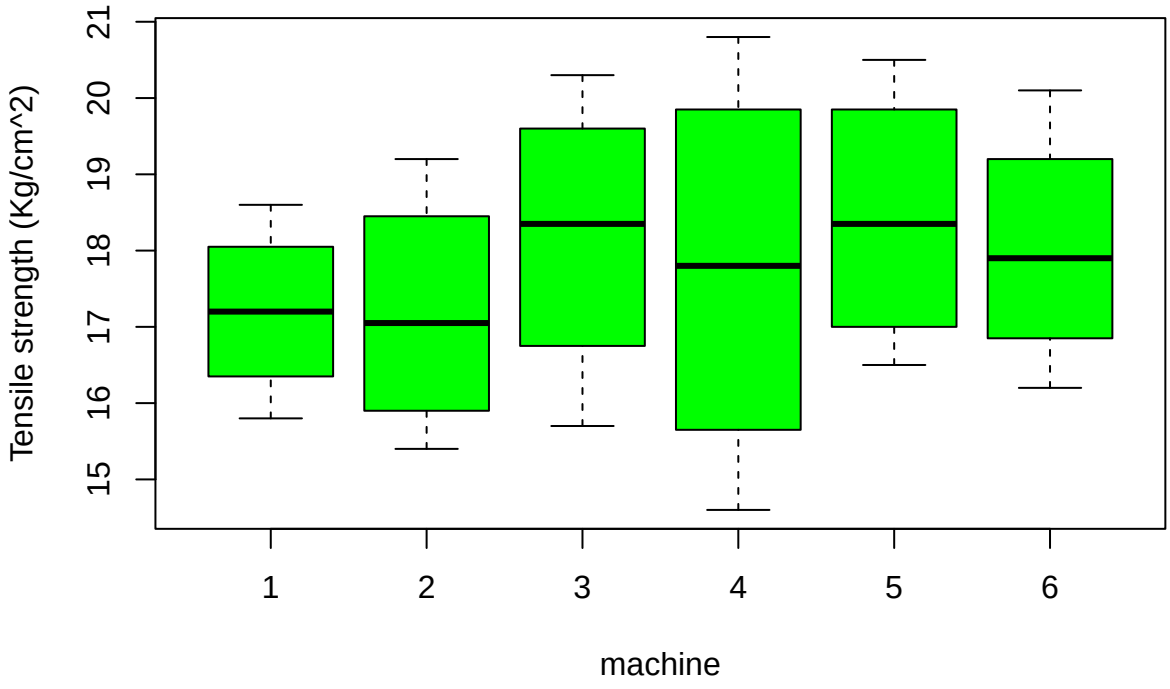
```
t.test(time[company == 2],
       time[company == 1],
       var.equal = FALSE,
       alternative = 'less',
       mu = 10)
```

```
##
##  Welch Two Sample t-test
##
## data:  time[company == 2] and time[company == 1]
## t = 0.215, df = 7.3756, p-value = 0.5822
## alternative hypothesis: true difference in means is less than 10
## 95 percent confidence interval:
##      -Inf 35.33647
## sample estimates:
## mean of x mean of y
##     110.0      97.4
```

The p-value is 0.582, meaning we have no evidence against the null hypothesis, or that we have no reason to believe company 2's movies are less than 10 minutes longer on average than company 1's.

## 2.

```
tensile <- data.frame(strength=c(17.5, 16.9, 15.8, 18.6,
                                 16.4, 19.2, 17.7, 15.4,
                                 20.3, 15.7, 17.8, 18.9,
                                 14.6, 16.7, 20.8, 18.9,
                                 17.5, 19.2, 16.5, 20.5,
                                 18.3, 16.2, 17.5, 20.1),
                      machine =c(1, 1, 1, 1,
                                 2, 2, 2, 2,
                                 3, 3, 3, 3,
                                 4, 4, 4, 4,
                                 5, 5, 5, 5,
                                 6, 6, 6, 6))
attach(tensile)

boxplot(strength~machine,
        data=tensile,
        ylab="Tensile strength (Kg/cm^2)",
        col='green')
```




Visually there does appear to be some variance, though the means all appear similar. Instead of a boxplot, we can use a barplot with error bars.

First, lets make the error bars using a 0.05 confidence interval.

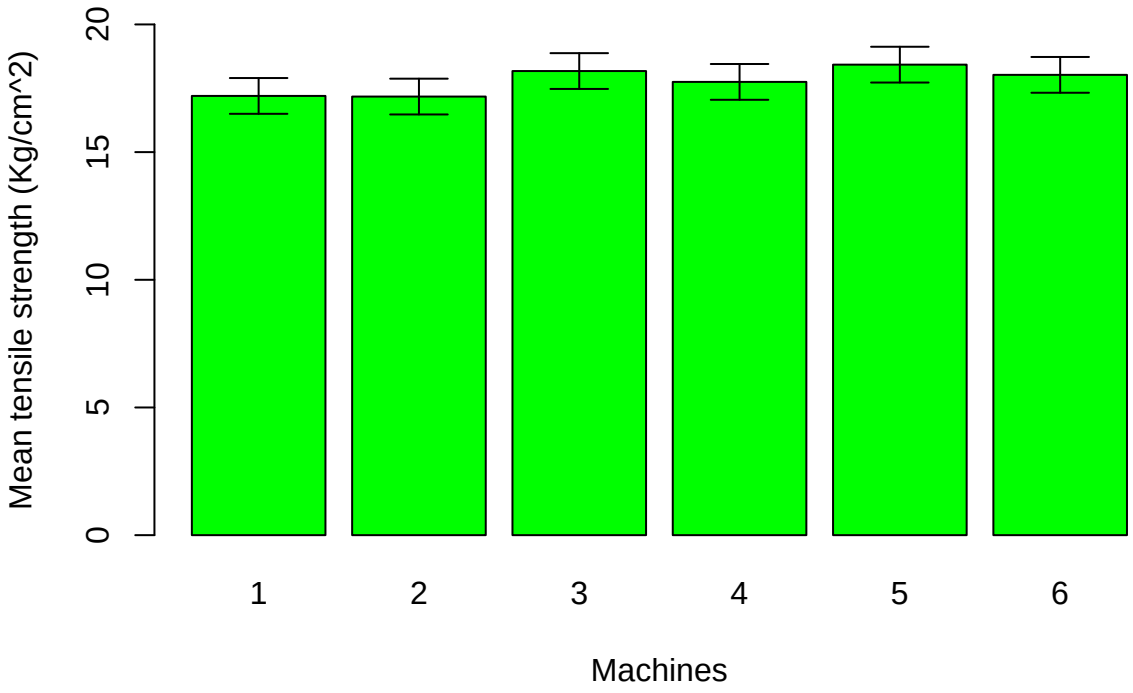
```r
sigma.hat <- summary.lm(aov(strength~machine))$sigma
sigma.hat
```

```
## [1] 1.716322
```

```r
# There's 6 populations, so divide by sqrt(6)
se.mean <- sigma.hat / sqrt(6)
se.mean
```

```
## [1] 0.7006856
```

```r
means <- rep(0, 6)
for ( i in 1:6)
  means[i] <- mean(strength[machine==i])

barplot(means,
        xlab="Machines",
        ylab="Mean tensile strength (Kg/cm^2)",
        names=c(1,2,3,4,5,6),
        ylim=c(0, 20),
        col='green')

x <- barplot(means, plot=F)
for (i in 1:6)
{
  arrows(x[i], means[i]-se.mean,
         x[i], means[i]+se.mean,
         code=3, angle=90, length=0.15)
}
```



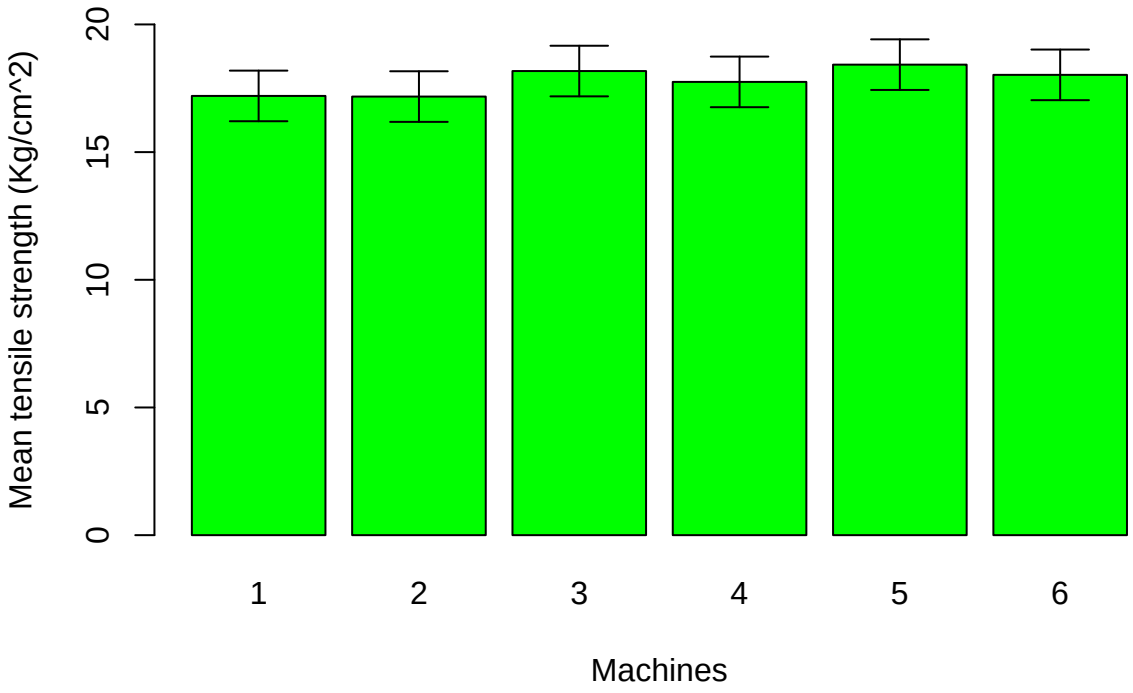

All of the error bars appear to overlap, so using this test we have no evidence that any machines produce rubber with different tensile strength.

This method of testing can give a large chance of a type-1 error, it's better to use the least significant difference to make the error bars. Since we have no evidence against the null hypothesis it won't make a difference here, but for completeness, lets do it anyways.

```
LSD <- 2 * sqrt(2) * se.mean
LSD
```

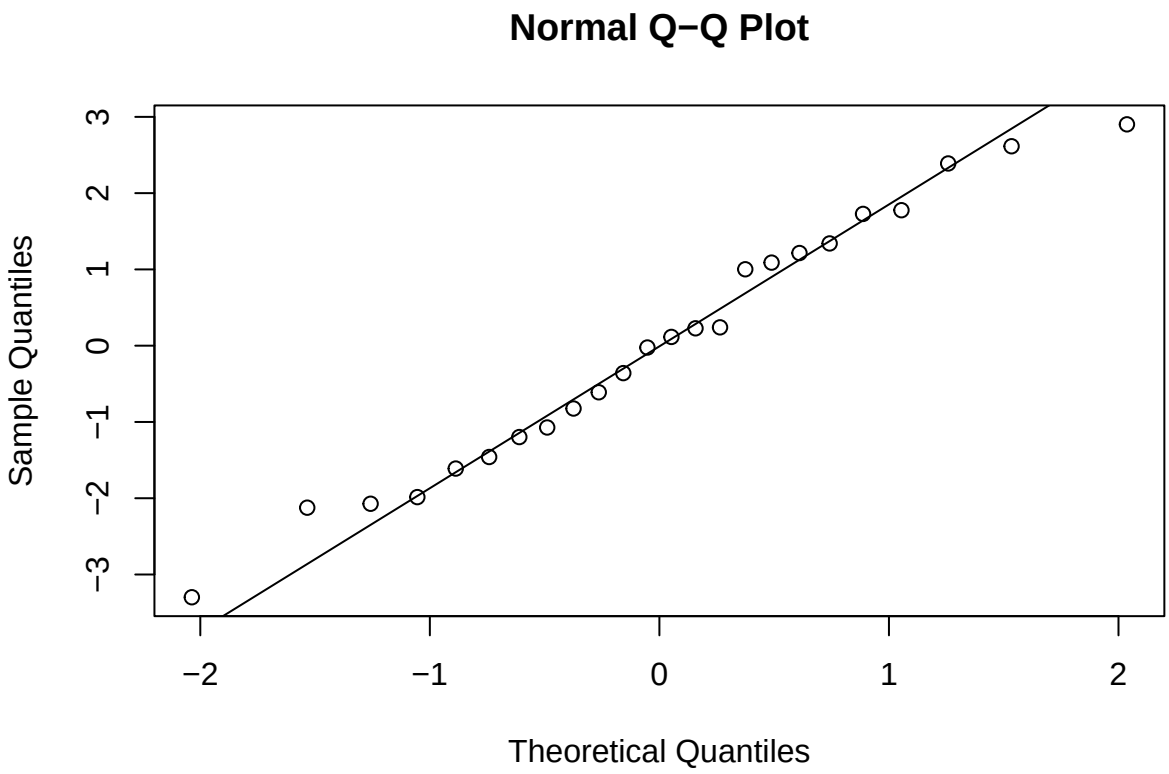
```
## [1] 1.981838
```

```
error = LSD / 2

barplot(means,
        xlab="Machines",
        ylab="Mean tensile strength (Kg/cm^2)",
        names=c(1,2,3,4,5,6),
        ylim=c(0, 20),
        col='green')

x <- barplot(means, plot=F)
for (i in 1:6)
{
  arrows(x[i], means[i]-error,
         x[i], means[i]+error,
         code=3, angle=90, length=0.15)
}
```




Before we can come to any conclusions, we should ensure that our assumption that the residuals are normally distributed holds.

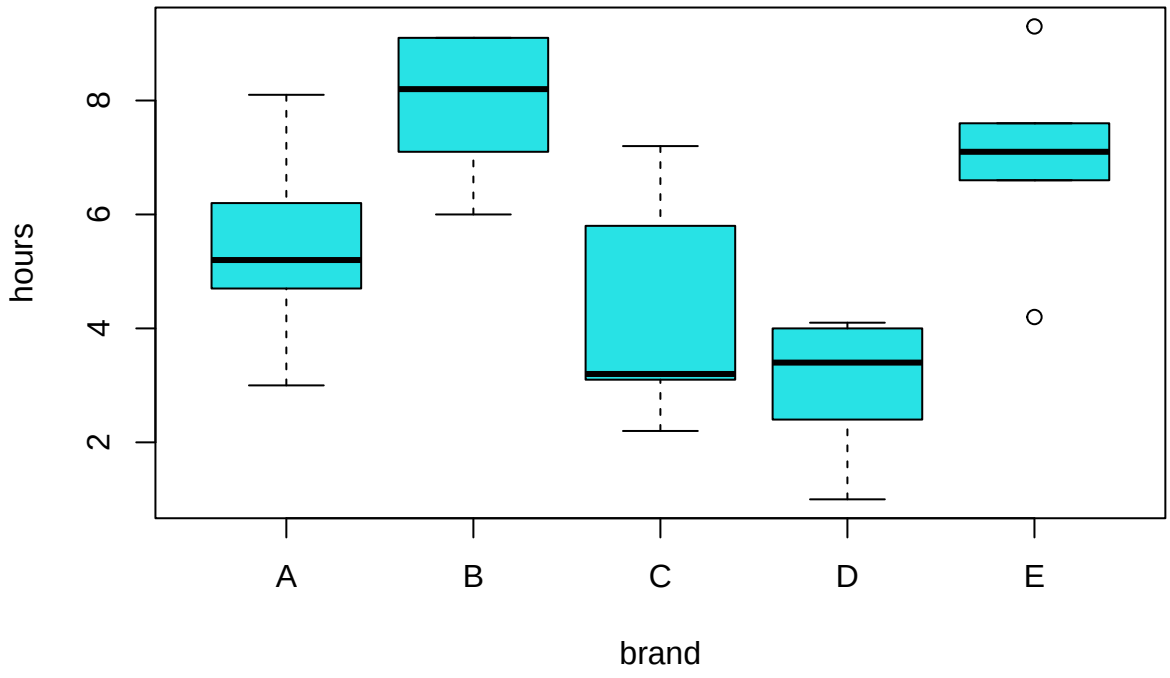
```
resid.strength <- resid(aov(strength~machine))
qqnorm(resid.strength)
qqline(resid.strength)
```




The residual distribution appears normal, so our testing with ANOVA was valid. It would appear that there is no evidence that any machine produces rubber with different tensile strength than any other machine.

**3.**

```
tablet <- data.frame(hours=c(5.2, 4.7, 8.1, 6.2, 3.0,
                             9.1, 7.1, 8.2, 6.0, 9.1,
                             3.2, 5.8, 2.2, 3.1, 7.2,
                             2.4, 3.4, 4.1, 1.0, 4.0,
                             7.1, 6.6, 9.3, 4.2, 7.6),
                     brand=c("A","A","A","A","A",
                             "B","B","B","B","B",
                             "C","C","C","C","C",
                             "D","D","D","D","D",
                             "E","E","E","E","E"))
attach(tablet)
boxplot(hours~brand, tablet, col=5)
```




Visually, Brands B and E seem to give more hours of relief than the others. Lets go straight to LSD error bars on a barplot to better analyse the variance.

```
sigma.hat <- summary.lm(aov(hours~brand))$sigma
sigma.hat
```

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

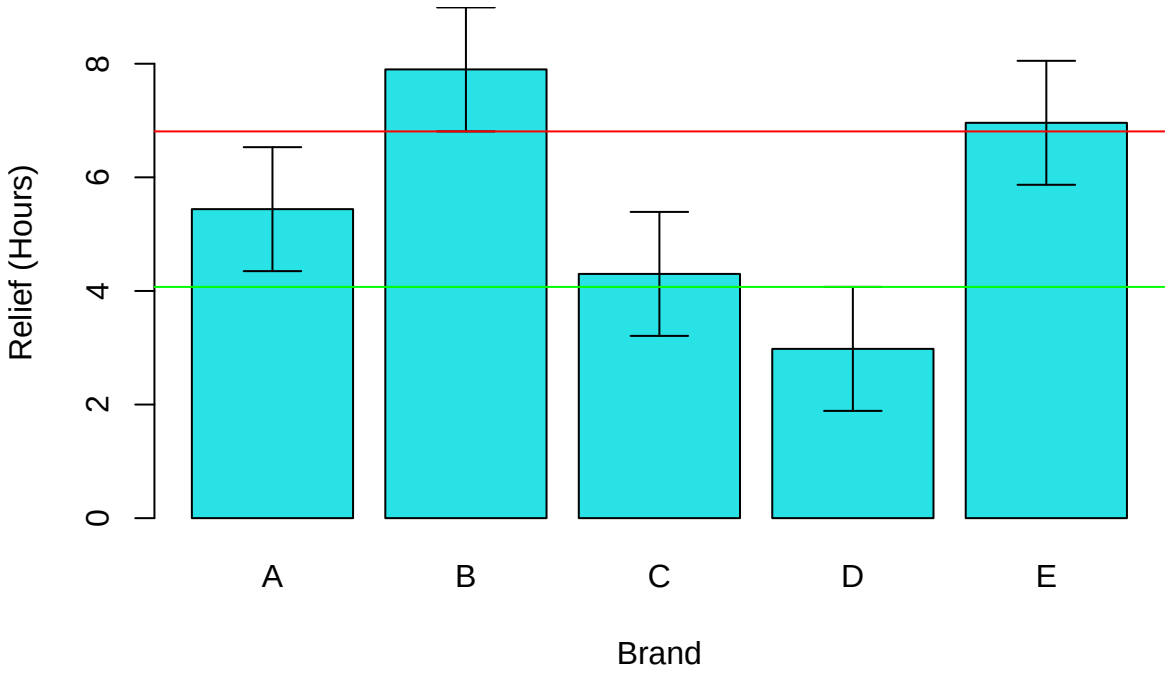
```
# There's 5 populations, so divide by sqrt(5)
se.mean <- sigma.hat / sqrt(5)
se.mean
```

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

```r
LSD <- 2 * sqrt(2) * se.mean
error = LSD / 2
error
```

```
## [1] 1.091165
```

```r
means <- rep(0, 5)
brand.list <- c("A","B","C","D","E")
for ( i in 1:5)
{
  means[i] <- mean(hours[brand == brand.list[i]])
}

x <- barplot(means,
        xlab="Brand",
        ylab="Relief (Hours)",
        names=brand.list,
        ylim=c(0,max(means)+error),
        col=5)

for (i in 1:5)
{
  arrows(x[i], means[i]-error,
         x[i], means[i]+error,
         code=3, angle=90, length=0.15)
}

abline(h=max(means)-error, col='red') # show lower bounds for error on column with the greatest mean
abline(h=min(means)+error, col='green') # show upper bounds for error on column with the least mean
```

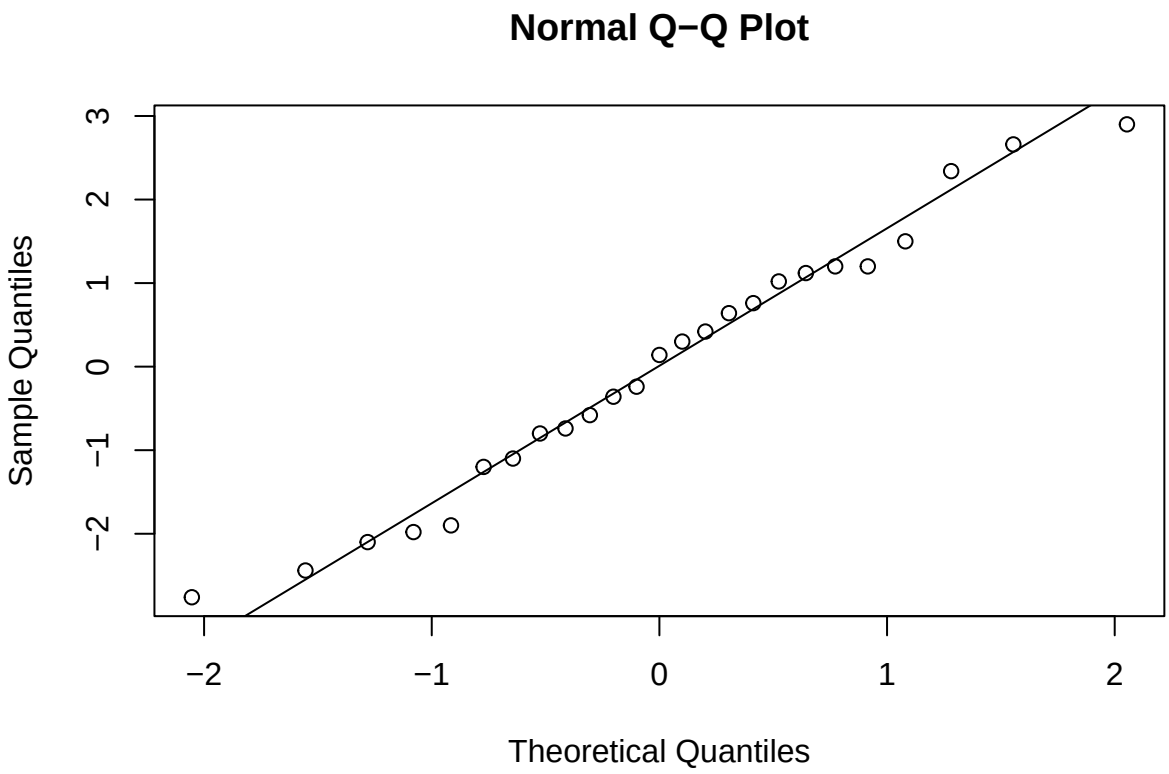



The red and green lines are added for clarity, to better see which error margins overlap.

Looking at the error margins on this barplot, there is evidence that some brands give more hours of relief than others. Brand B gives more relief than A, C, or D, and the brand that gives the most relief is either B or E. Brand D gives less release than A, B, or E, so when looking for headache relief, this brand is one of the less efficient options available.

To ensure this analysis is valid, lets check to see if the residuals are normally distributed.

```
resid.relief <- resid(aov(hours~brand))
qqnorm(resid.relief)
qqline(resid.relief)
```




The residuals appear normal, so our assumptions hold.