STAT 425 Project 01

Daniel Girvitz, Ethan Scott

27/01/2022

Load data

```
eData=read.csv("OrchardSprays.csv")
```

Review data

```
str(eData)
```

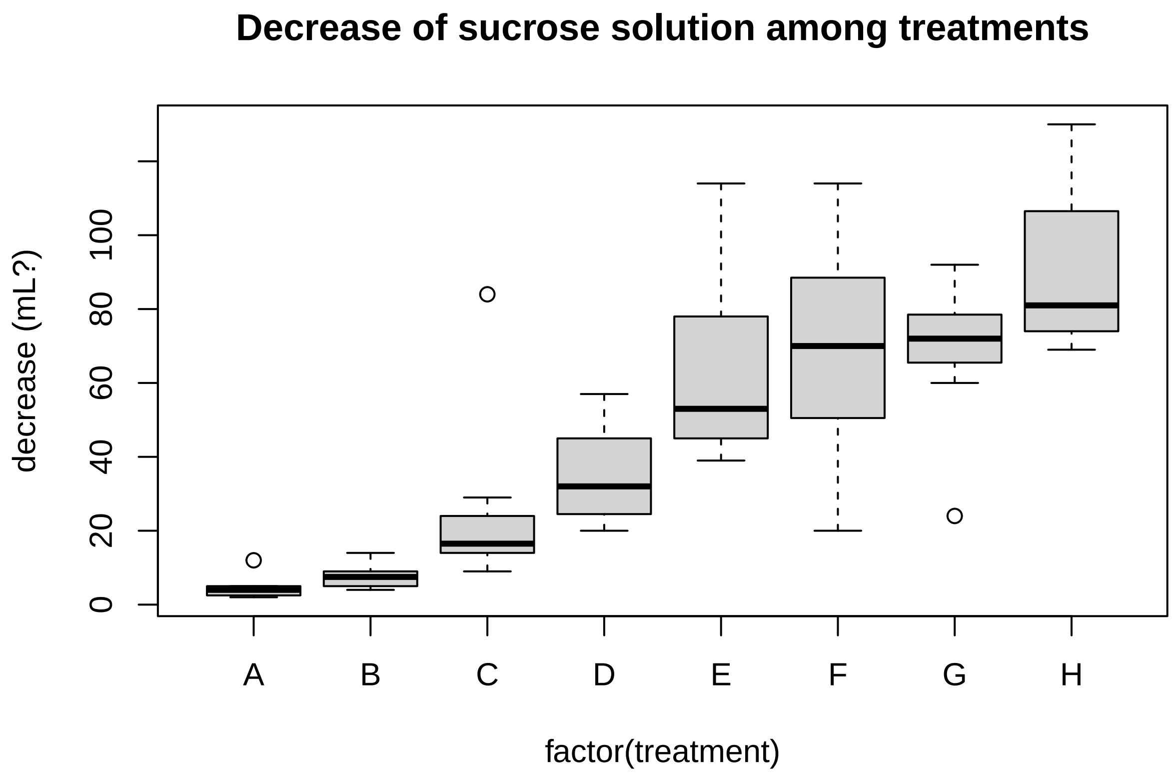
```
## 'data.frame':    64 obs. of  5 variables:
##  $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ decrease : int  57 95 8 69 92 90 15 2 84 6 ...
##  $ rowpos   : int  1 2 3 4 5 6 7 8 1 2 ...
##  $ colpos   : int  1 1 1 1 1 1 1 1 2 2 ...
##  $ treatment: chr  "D" "E" "B" "H" ...
```

```
table(eData$treatment)
```

```
##
## A B C D E F G H
## 8 8 8 8 8 8 8 8
```

Boxplots

```
boxplot(decrease~factor(treatment),data=eData,
        ylab="decrease (mL?)",
        main="Decrease of sucrose solution among treatments")
```

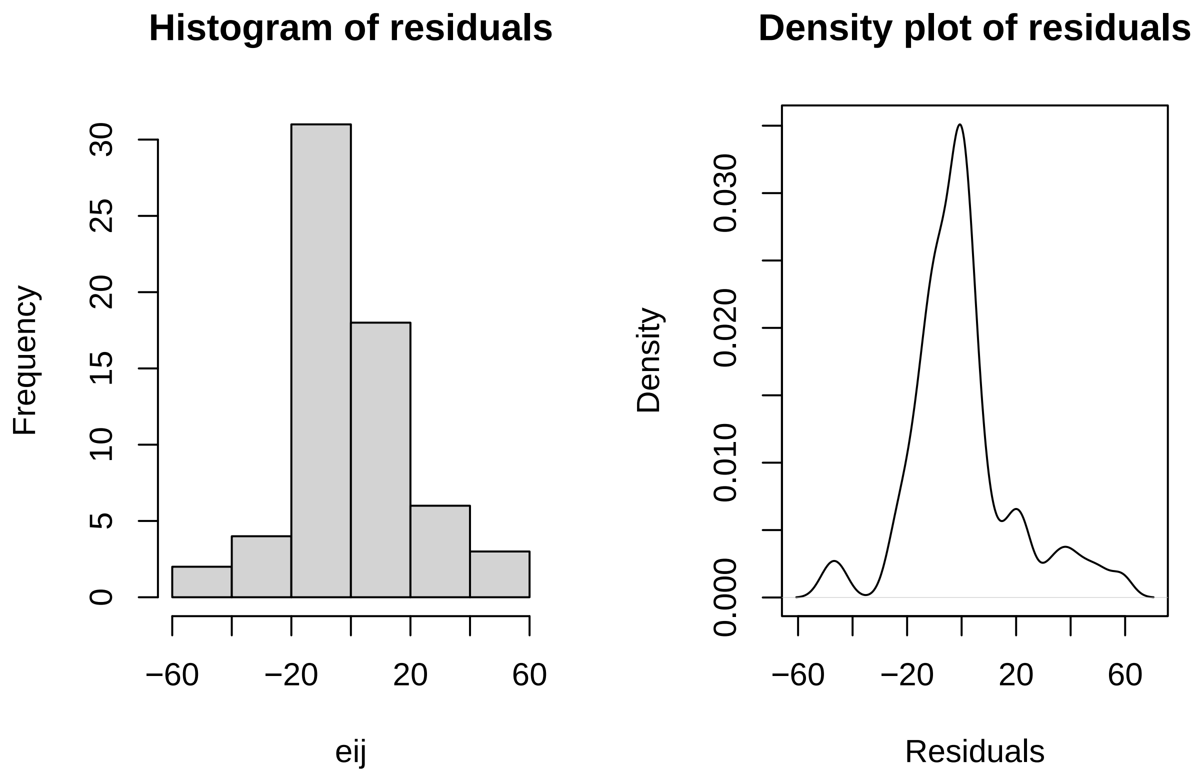



```
one.way=aov(decrease~factor(treatment),data=eData)
summary(one.way)
```

```
##                   Df Sum Sq Mean Sq F value  Pr(>F)
## factor(treatment)  7  56160    8023   19.06 9.5e-13 ***
## Residuals         56  23570     421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

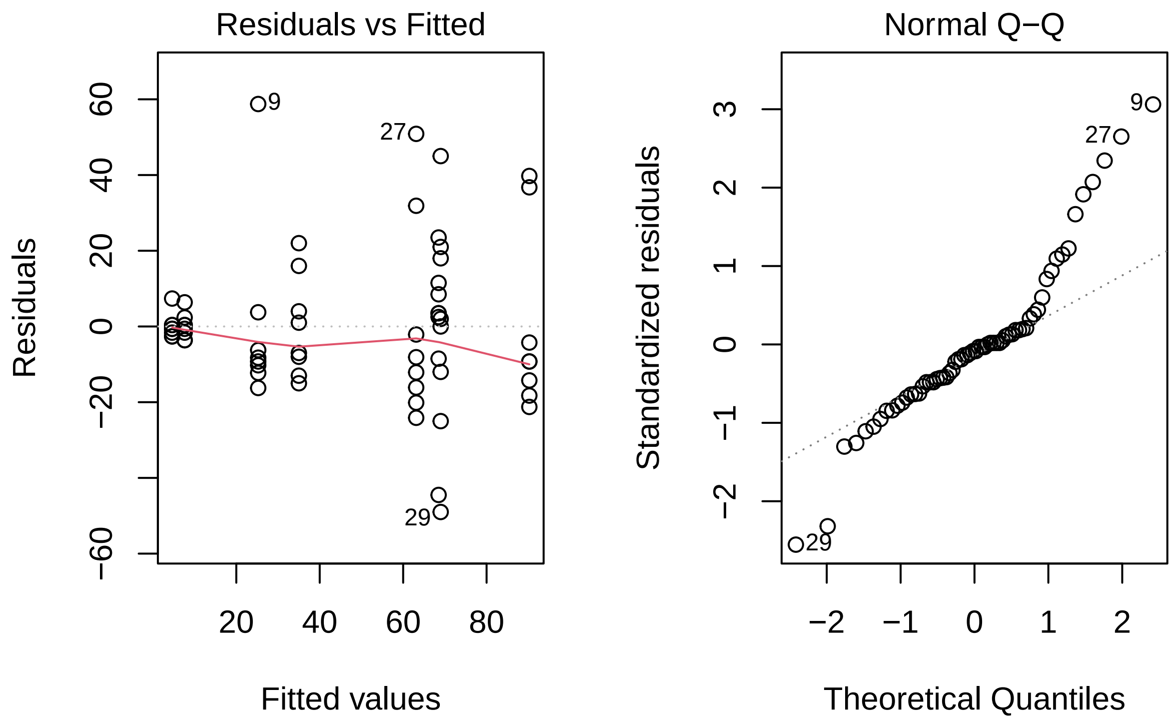
Plots of residuals

```
par(mfrow=c(1,2))
eij=residuals(one.way)
hist(eij,main="Histogram of residuals")
plot(density(eij),main="Density plot of residuals",ylab="Density",xlab="Residuals")
```



Distribution of residuals appears to satisfy normality assumption, although there is a minor right skew.

```
par(mfrow=c(1,2))
plot(one.way,1)
plot(one.way,2)
```



Analysis of diagnostic plots:
1. Residuals vs Fitted: Overall, the linear is pretty straight, but some treatment levels have impactful outliers.
2. Normal Q-Q: The diagonal line does not describe the points very well, which is a possible violation of the normality assumption.

In light of this, we would like to perform some formal testing, so we perform the "Shapiro-Wilk Test" & Modified Levene's Test".

Shapiro-Wilk's test for normality

```
shapiro.test((eData$decrease))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  (eData$decrease)
## W = 0.91892, p-value = 0.0004483
```

Yikes, REJECT normality.

Modified Levene's test

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.1.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.1.2
```

```
# Levene's test with one independent variable
leveneTest(decrease ~ factor(treatment), data = eData)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  7  1.5132 0.1817
##       56
```

Variance appears to be fine, but Normal Q-Q plot shows heavy skew in the upper tail. To minimize the influence of those residuals, let's try a square-root transformation. (In convex analysis terms, the square root transformation would apply a harsher penality on larger outliers.)

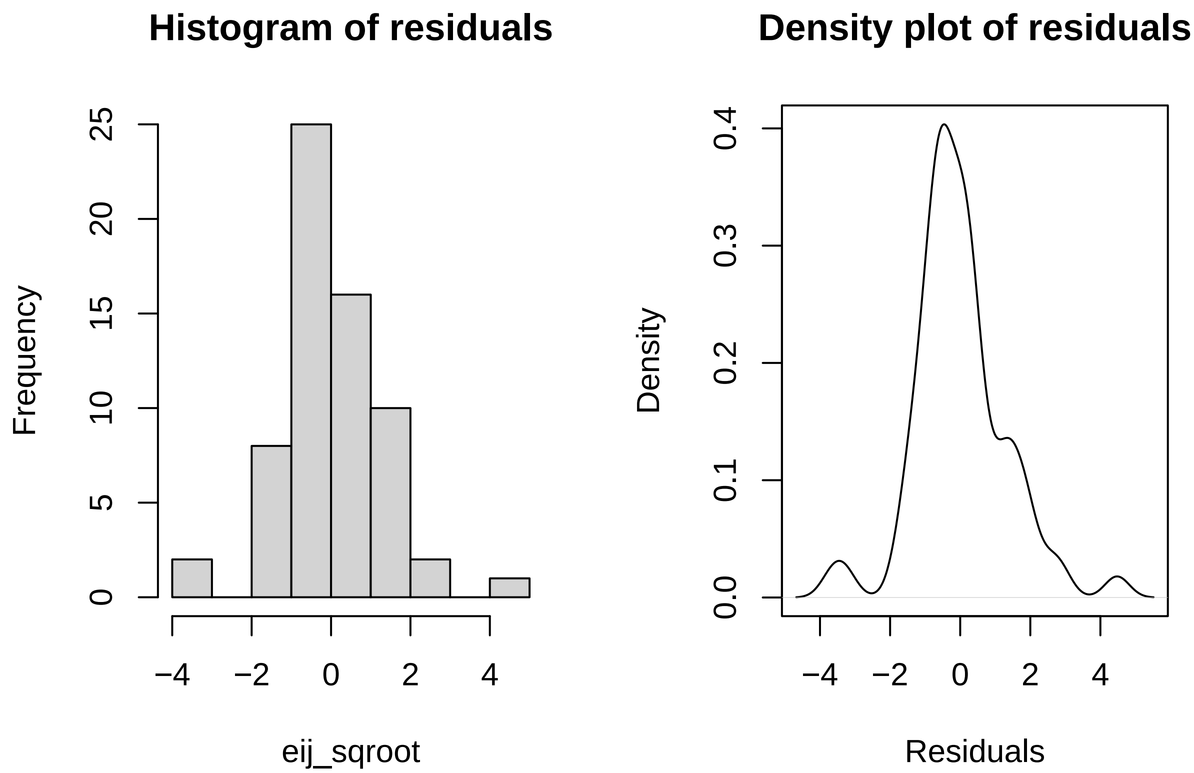
Square root transformation applied to ANOVA

```
one.way.sqroot=aov(sqrt(decrease)~factor(treatment),data=eData)
summary(one.way.sqroot)
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## factor(treatment)  7  417.7   59.68   30.95 <2e-16 ***
## Residuals         56  108.0    1.93
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Plots of residuals

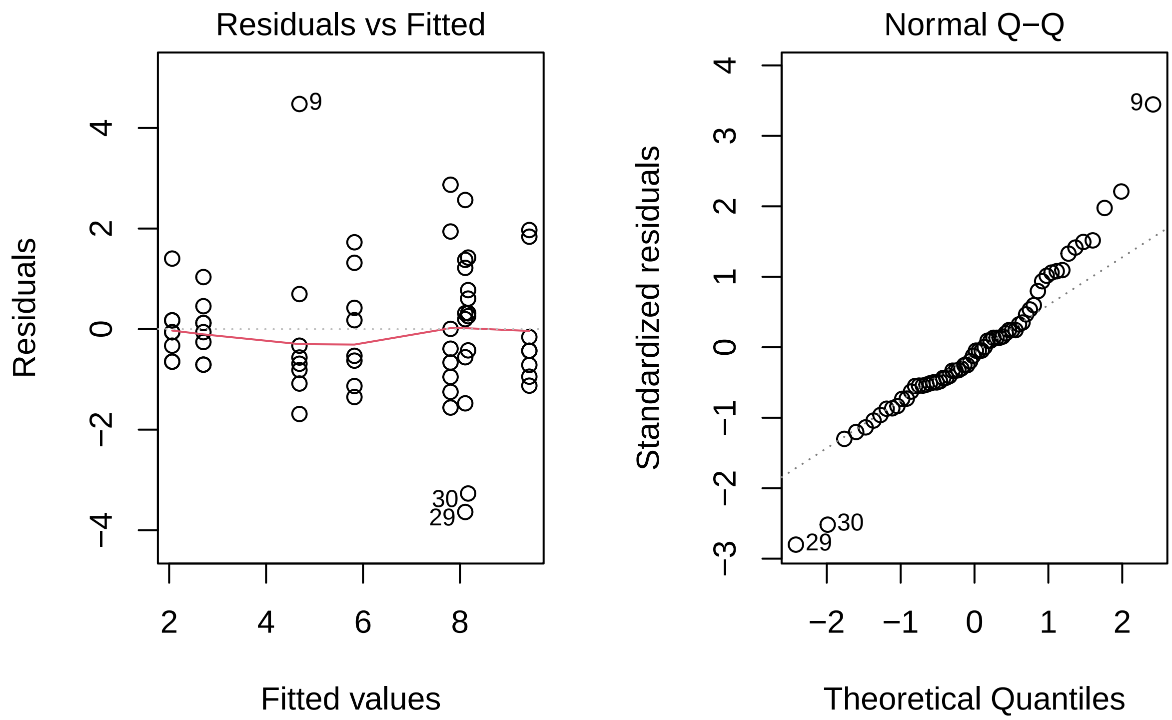
```
par(mfrow=c(1,2))
eij_sqroot=residuals(one.way.sqroot)
hist(eij_sqroot,main="Histogram of residuals")
plot(density(eij_sqroot),main="Density plot of residuals",ylab="Density",xlab="Residuals")
```



Looks a bit better.

Diagnostic plots

```
par(mfrow=c(1,2))
plot(one.way.sqroot,1)
plot(one.way.sqroot,2)
```



Much better.

Tests for homogenity of variance (homoscedastiticity)

```
bartlett.test(sqrt(decrease) ~ factor(treatment), data = eData)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  sqrt(decrease) by factor(treatment)
## Bartlett's K-squared = 15.529, df = 7, p-value = 0.02979
```

```
leveneTest(sqrt(decrease) ~ factor(treatment), data = eData)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  7   0.763 0.6205
##       56
```

Bartlett test rejects equality of variances, but this is suspect because it is heavily influenced by normality dist. requirement. Levene's test FTR, and is perhaps more accurate because it is more robust.

Shapiro-Wilk's test for normality

```
shapiro.test(sqrt(eData$decrease))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sqrt(eData$decrease)
## W = 0.94271, p-value = 0.005088
```

Yikes, guess it's better to go with the Modified Levene test after all…

#getting means of treatments

```
means = tapply(eData$decrease,eData$treatment, mean)
means[1]
```

```
##     A
## 4.625
```

```
means[2]
```

```
##     B
## 7.625
```

```
means[3]
```

```
##     C
## 25.25
```

```
means[4]
```

```
##  D
## 35
```

```
means[5]
```

```
##      E
## 63.125
```

```
means[6]
```

```
##  F
## 69
```

```
means[7]
```

```
##    G
## 68.5
```

```r
means[8]
```

```
##     H
## 90.25
```

#Fischer LSD Test

```r
N=64
n=8
a=8
MSE=8023
Fisher.LSD=qt(0.05, N-a, lower.tail=F)*sqrt(MSE*2/n)
Fisher.LSD
```

```
## [1] 74.90492
```

```r
comparisons=c(
  means[1]-means[2],
  means[1]-means[3],
  means[1]-means[4],
  means[1]-means[5],
  means[1]-means[6],
  means[1]-means[7],
  means[1]-means[8],#Positive
  means[2]-means[3],
  means[2]-means[4],
  means[2]-means[5],
  means[2]-means[6],
  means[2]-means[7],
  means[2]-means[8],#positive
  means[3]-means[4],
  means[3]-means[5],
  means[3]-means[6],
  means[3]-means[7],
  means[3]-means[8],
  means[4]-means[5],
  means[4]-means[6],
  means[4]-means[7],
  means[4]-means[8],
  means[5]-means[6],
  means[5]-means[7],
  means[5]-means[8],
  means[6]-means[7],
  means[6]-means[8],
  means[7]-means[8])
abs(comparisons)
```

```
##      A      A      A      A      A      A      A      B      B      B      B
##  3.000 20.625 30.375 58.500 64.375 63.875 85.625 17.625 27.375 55.500 61.375
##      B      B      C      C      C      C      C      D      D      D      D
## 60.875 82.625  9.750 37.875 43.750 43.250 65.000 28.125 34.000 33.500 55.250
##      E      E      E      F      F      G
##  5.875  5.375 27.125  0.500 21.250 21.750
```

```r
Fisher.LSD
```

```
## [1] 74.90492
```

```r
abs(comparisons)-Fisher.LSD
```

```
##          A          A          A          A          A          A          A
## -71.904915 -54.279915 -44.529915 -16.404915 -10.529915 -11.029915  10.720085
##          B          B          B          B          B          B          C
## -57.279915 -47.529915 -19.404915 -13.529915 -14.029915   7.720085 -65.154915
##          C          C          C          C          D          D          D
## -37.029915 -31.154915 -31.654915  -9.904915 -46.779915 -40.904915 -41.404915
##          D          E          E          E          F          F          G
## -19.654915 -69.029915 -69.529915 -47.779915 -74.404915 -53.654915 -53.154915
```

mean 1-8 and 2-8 are significantly different, the rest of the means are not

#Tukey's Test

```r
test_statistic = (max(means)-min(means))/sqrt(MSE/n)
test_statistic
```

```
## [1] 2.703816
```

```r
q_Alpha = 2.10 #from table
Tukey = q_Alpha*sqrt(MSE/n)
Tukey
```

```
## [1] 66.50322
```

```r
abs(comparisons)-Tukey
```

```
##          A          A          A          A          A          A          A
## -63.503224 -45.878224 -36.128224  -8.003224  -2.128224  -2.628224  19.121776
##          B          B          B          B          B          B          C
## -48.878224 -39.128224 -11.003224  -5.128224  -5.628224  16.121776 -56.753224
##          C          C          C          C          D          D          D
## -28.628224 -22.753224 -23.253224  -1.503224 -38.378224 -32.503224 -33.003224
##          D          E          E          E          F          F          G
## -11.253224 -60.628224 -61.128224 -39.378224 -66.003224 -45.253224 -44.753224
```

```r
# means 1-8 and 2-8
```

Same as Fisher. LSD, mean 1-8 and 2-8 are significantly different, the rest of the means are not

#Confidence Intervals

```r
(means[1]-means[2])+c(-1,1)*Tukey
```

```
## [1] -69.50322  63.50322
```

```r
(means[1]-means[3])+c(-1,1)*Tukey
```

```
## [1] -87.12822  45.87822
```

```r
(means[1]-means[4])+c(-1,1)*Tukey
```

```
## [1] -96.87822  36.12822
```

```r
(means[1]-means[5])+c(-1,1)*Tukey
```

```
## [1] -125.003224    8.003224
```

```r
(means[1]-means[6])+c(-1,1)*Tukey
```

```
## [1] -130.878224    2.128224
```

```r
(means[1]-means[7])+c(-1,1)*Tukey
```

```
## [1] -130.378224    2.628224
```

```r
(means[1]-means[8])+c(-1,1)*Tukey# does not cross 0
```

```
## [1] -152.12822  -19.12178
```

```r
(means[2]-means[3])+c(-1,1)*Tukey
```

```
## [1] -84.12822  48.87822
```

```r
(means[2]-means[4])+c(-1,1)*Tukey
```

```
## [1] -93.87822  39.12822
```

```r
(means[2]-means[5])+c(-1,1)*Tukey
```

```
## [1] -122.00322   11.00322
```

```r
(means[2]-means[6])+c(-1,1)*Tukey
```

```
## [1] -127.878224    5.128224
```

```r
(means[2]-means[7])+c(-1,1)*Tukey
```

```
## [1] -127.378224    5.628224
```

```r
(means[2]-means[8])+c(-1,1)*Tukey#does not cross 0
```

```
## [1] -149.12822  -16.12178
```

```r
(means[3]-means[4])+c(-1,1)*Tukey
```

```
## [1] -76.25322  56.75322
```

```r
(means[3]-means[5])+c(-1,1)*Tukey
```

```
## [1] -104.37822   28.62822
```

```r
(means[3]-means[6])+c(-1,1)*Tukey
```

```
## [1] -110.25322   22.75322
```

```r
(means[3]-means[7])+c(-1,1)*Tukey
```

```
## [1] -109.75322   23.25322
```

```r
(means[3]-means[8])+c(-1,1)*Tukey
```

```
## [1] -131.503224    1.503224
```

```r
(means[4]-means[5])+c(-1,1)*Tukey
```

```
## [1] -94.62822  38.37822
```

```r
(means[4]-means[6])+c(-1,1)*Tukey
```

```
## [1] -100.50322   32.50322
```

```r
(means[4]-means[7])+c(-1,1)*Tukey
```

```
## [1] -100.00322   33.00322
```

```r
(means[4]-means[8])+c(-1,1)*Tukey
```

```
## [1] -121.75322   11.25322
```

```r
(means[5]-means[6])+c(-1,1)*Tukey
```

```
## [1] -72.37822  60.62822
```

```r
(means[5]-means[7])+c(-1,1)*Tukey
```

```
## [1] -71.87822  61.12822
```

```r
(means[5]-means[8])+c(-1,1)*Tukey
```

```
## [1] -93.62822  39.37822
```

```r
(means[6]-means[7])+c(-1,1)*Tukey
```

```
## [1] -66.00322  67.00322
```

```r
(means[6]-means[8])+c(-1,1)*Tukey
```

```
## [1] -87.75322  45.25322
```

```r
(means[7]-means[8])+c(-1,1)*Tukey
```

```
## [1] -88.25322  44.75322
```

These results are the same as well, we can conclude that the means 1-8 and 2-8 are significantly different