MATH 3080 Lab Project 3

Your Name Here

### 12/29/2019
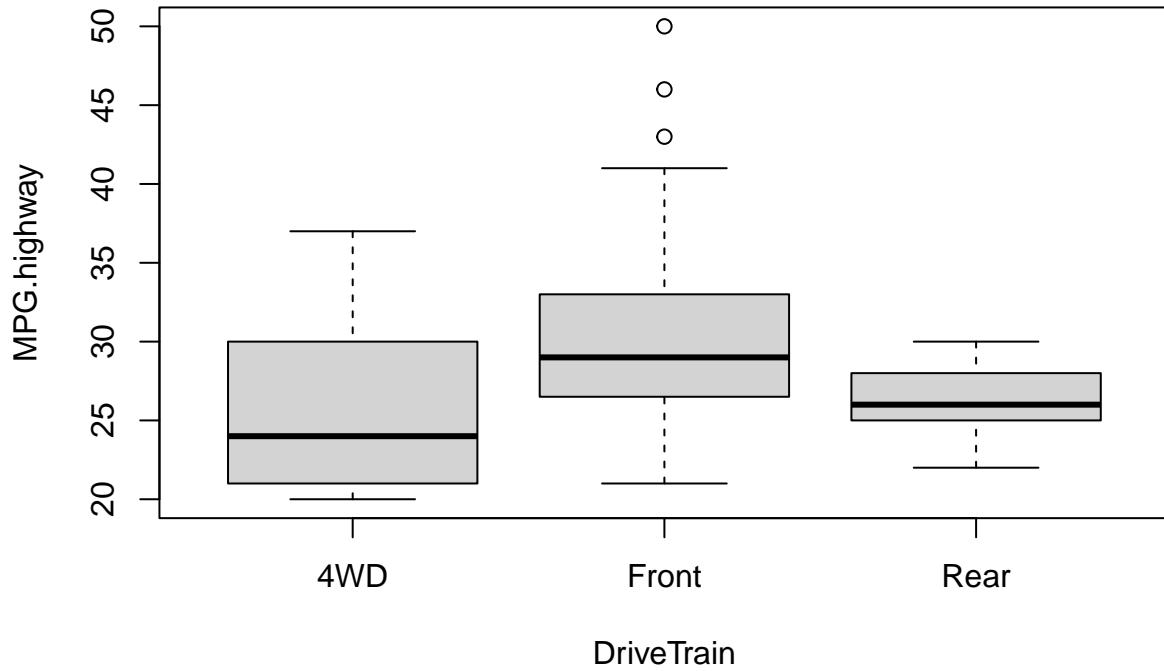
Contents

Problem 1 (Verzani problem 12.2)

*For the data set `Cars93` (**MASS**) perform a one-way analysis of variance of `MPG.highway` for each level of `DriveTrain`. Does the data support the null hypothesis of equal population means? (Use `aov()` for this problem.)*

```
library(MASS)
boxplot(MPG.highway ~ DriveTrain, data = Cars93)
```

```r
test <- aov(MPG.highway ~ DriveTrain, data = Cars93)
oneway.test(MPG.highway ~ DriveTrain, data = Cars93)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  MPG.highway and DriveTrain
## F = 10.725, num df = 2.000, denom df = 22.527, p-value = 0.0005342
```

```r
summary(test)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## DriveTrain   2  320.1   160.1   6.276 0.00281 **
## Residuals   90 2295.2    25.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is strong evidence that the variance is different. Our p-value of .0023 indicates that the populations in these levels are different. If we assume an alpha level of .01 we would reject the null hypothesis. Visually inspecting the boxplot also indicates that there are dramatic differences between these populations.

Problem 2 (Verzani problem 12.4)

*The data set **carsafety (UsingR)** contains car-crash data. For several makes of cars the number of drivers killed per million is recorded in **Drivers.deaths**. The number of drivers of other cars killed in accidents with these cars, per million, is recorded in **Other.deaths**. The variable **type** is a factor indicating the type of car.*

*Perform a one-way analysis of variance of the model **Drivers.deaths ~ type**. Check for equal variances and normality. (Use **oneway.test()** for this problem.)*

```r
library(UsingR)
```

```
## Loading required package: HistData
```

```
## Warning: package 'HistData' was built under R version 4.5.2
```

```
## Loading required package: Hmisc
```

```
## Warning: package 'Hmisc' was built under R version 4.5.2
```
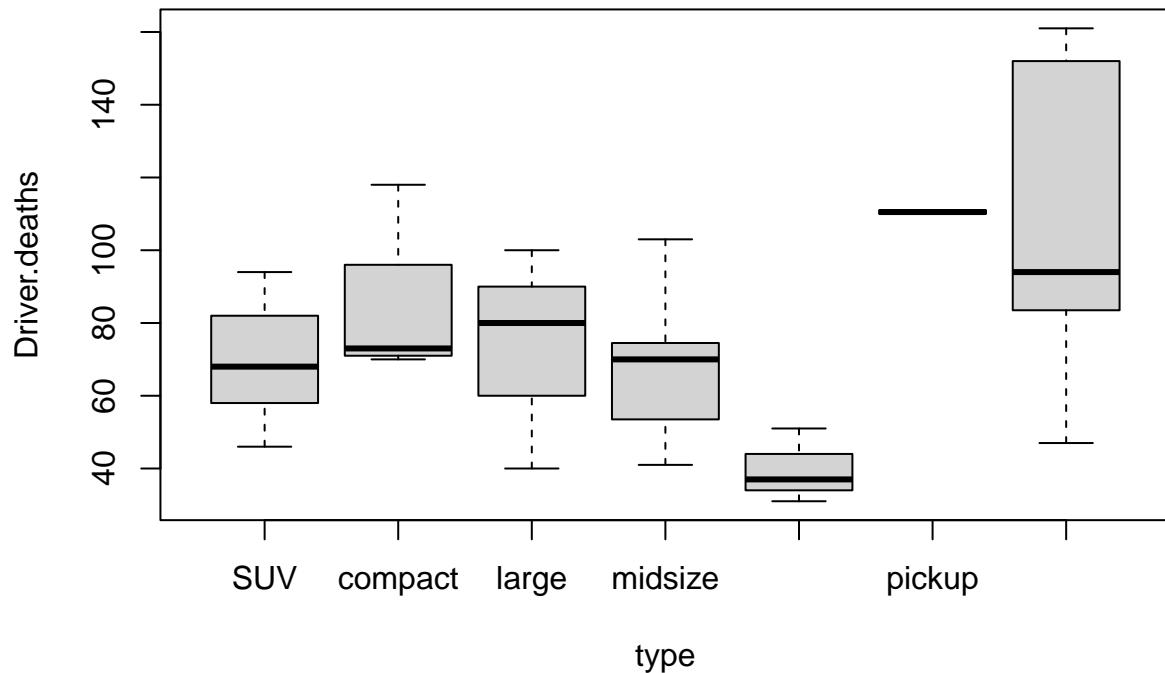
```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##      format.pval, units
```

```r
oneway.test(Driver.deaths ~ type, data = carsafety)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  Driver.deaths and type
## F = 26.465, num df = 6.0000, denom df = 8.7853, p-value = 3.75e-05
```

```r
boxplot(Driver.deaths ~ type, data = carsafety)
```

```
suvData <- carsafety[carsafety$type == "SUV", ]
compactData <- carsafety[carsafety$type == "compact", ]
largeData <- carsafety[carsafety$type == "large", ]
midsizeData <- carsafety[carsafety$type == "midsize", ]
minivanData <- carsafety[carsafety$type == "minivan", ]
pickupData <- carsafety[carsafety$type == "pickup", ]


qqnorm(suvData$Driver.deaths)
qqline(suvData$Driver.deaths)
```
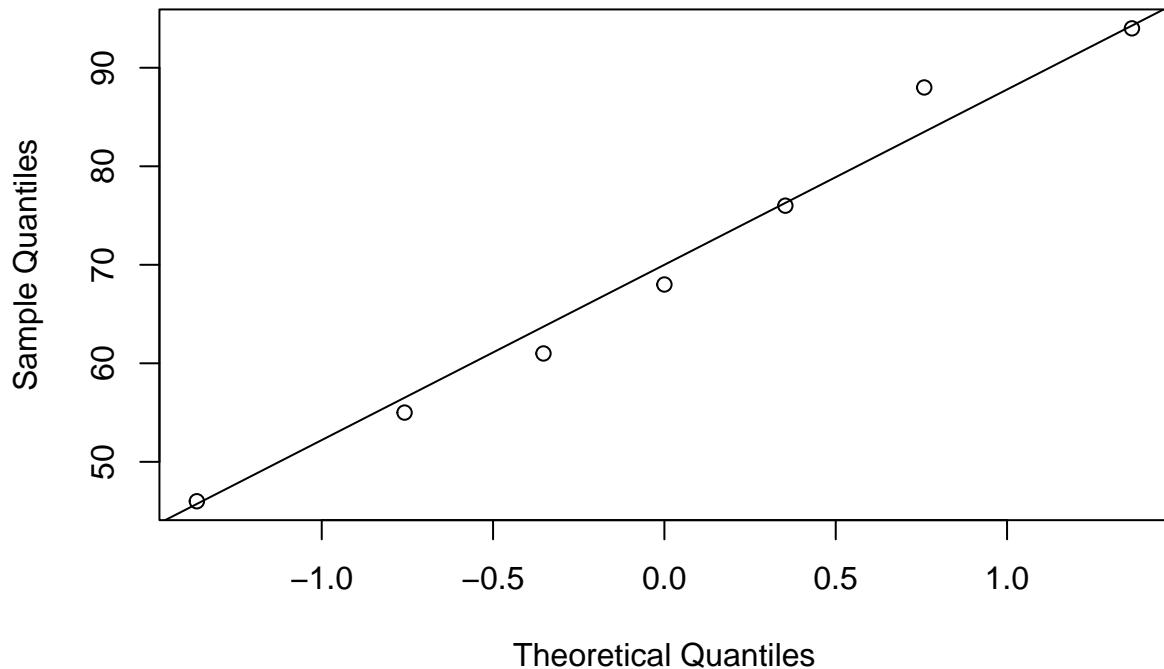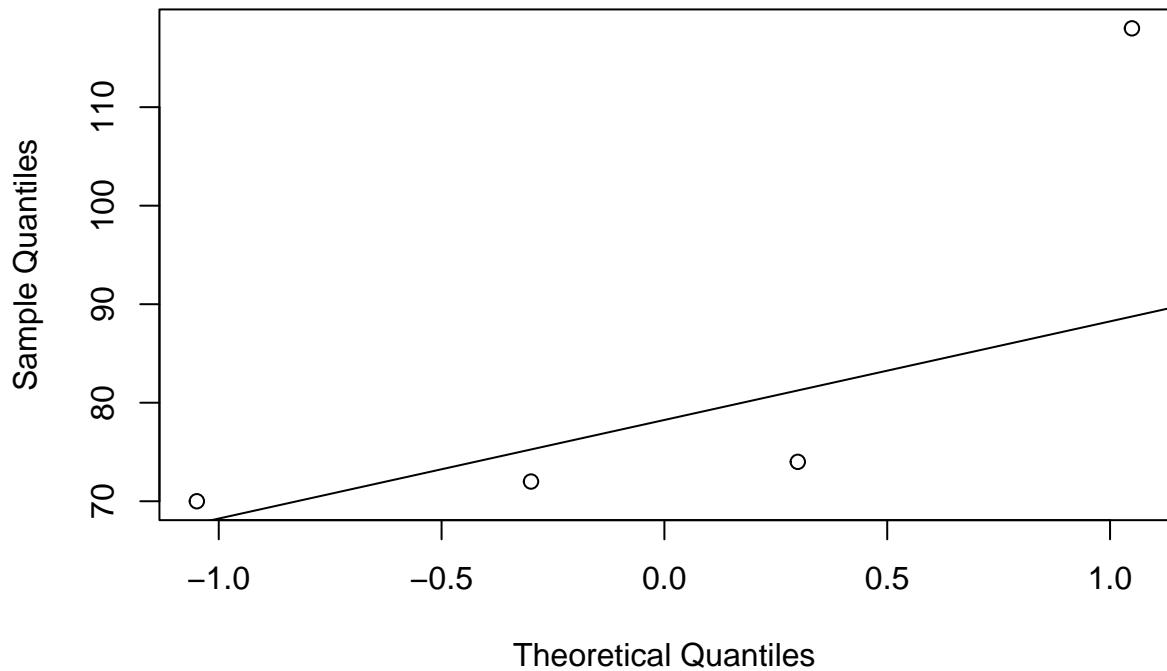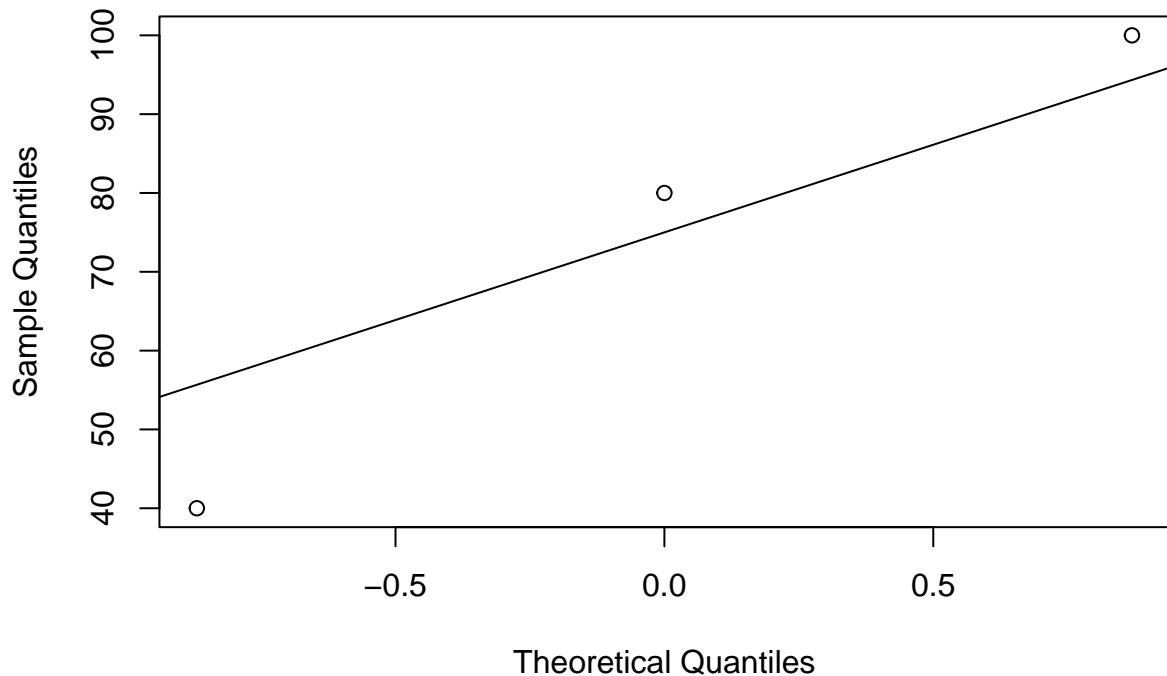
Normal Q–Q Plot


```
qqnorm(compactData$Driver.deaths)
qqline(compactData$Driver.deaths)
```
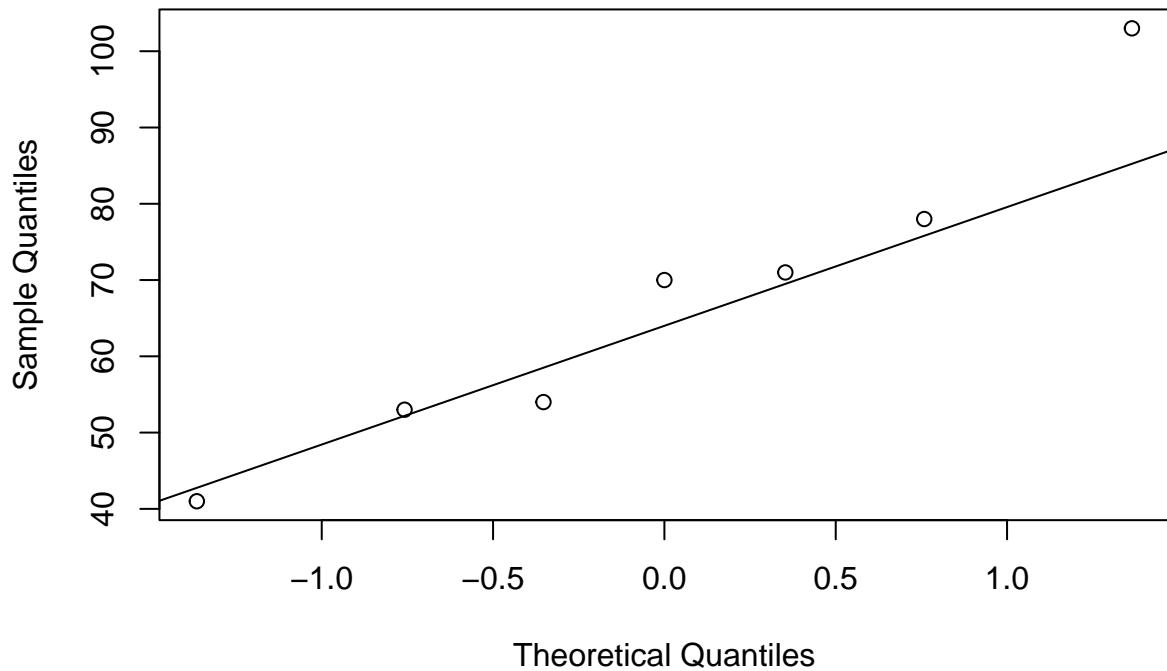
Normal Q–Q Plot


```
qqnorm(largeData$Driver.deaths)
qqline(largeData$Driver.deaths)
```
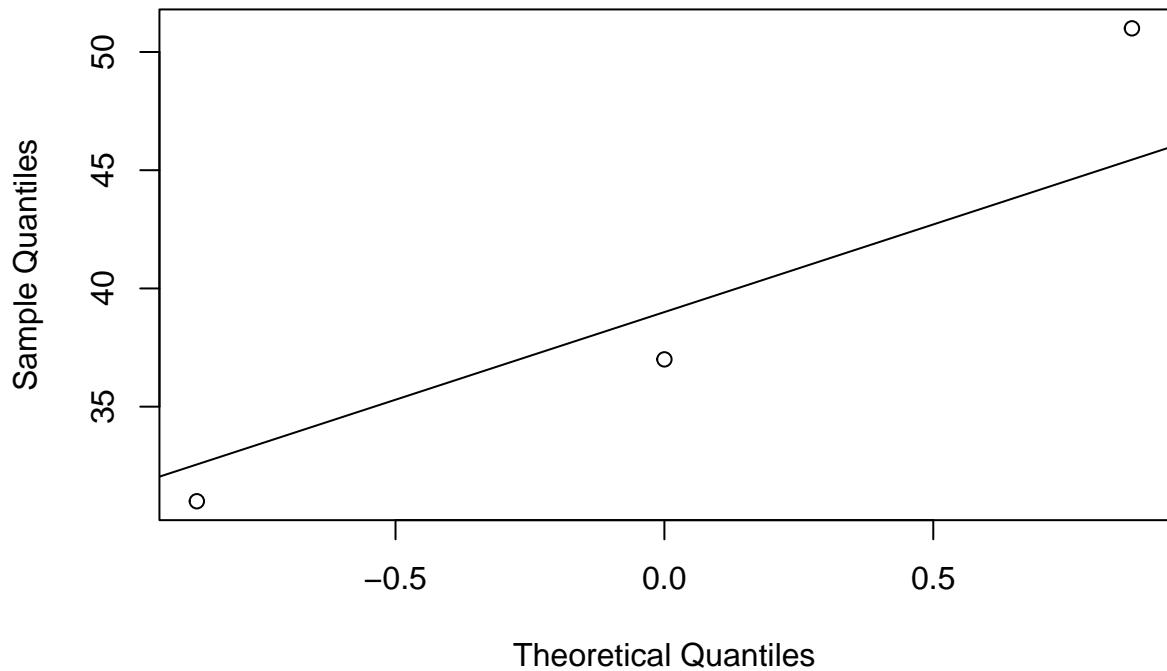
Normal Q−Q Plot


```
qqnorm(midsizeData$Driver.deaths)
qqline(midsizeData$Driver.deaths)
```
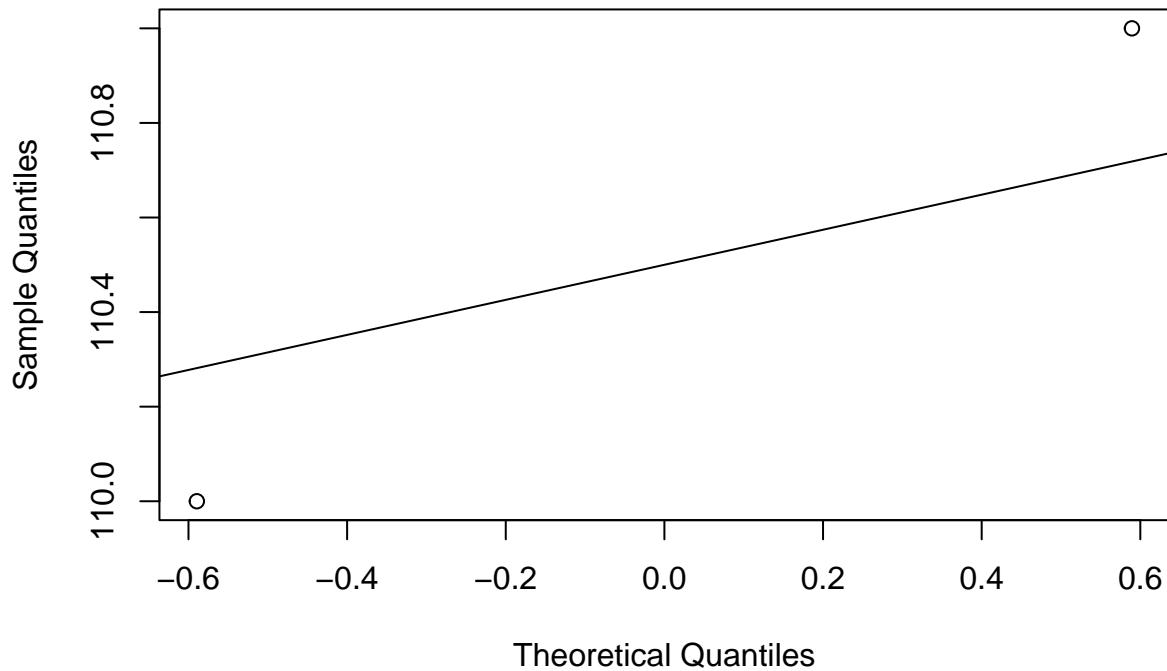
Normal Q–Q Plot


```
qqnorm(minivanData$Driver.deaths)
qqline(minivanData$Driver.deaths)
```

Normal Q–Q Plot


```
qqnorm(pickupData$Driver.deaths)
qqline(pickupData$Driver.deaths)
```

Normal Q–Q Plot

Is there a difference in population means? Can you assume equal variances? What about normally distributed populations? There should be a plot associated with each of these.

Response:
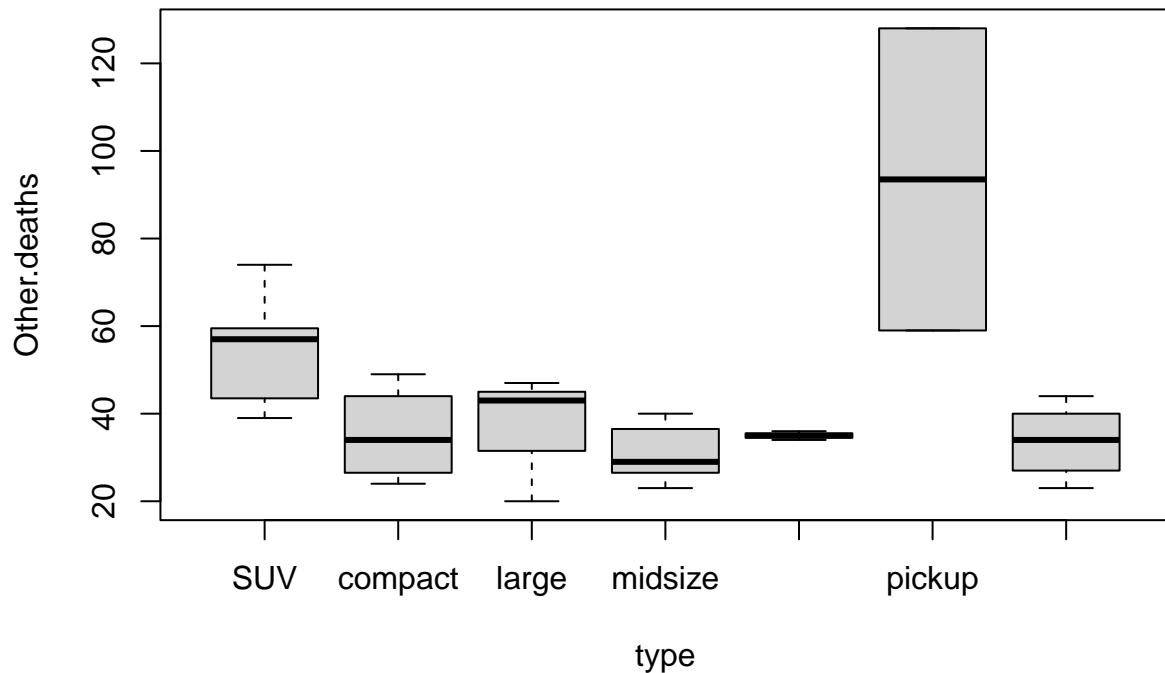
Yes there is a difference in the mean. our oneway test had a p value of 3.75e-05 which is very small. No, we can not assume equal variances based on visual inspection of the boxplot. There are dramatic differences between some of the types, with minivans having a much lower mean and much tighter distribution than any of the other types. I think it is generally safe to assume normality for these populations, however the sample size in some of these groups is too small to be certain.

*Repeat with an analysis of variance of the model `Other.deaths ~ type`.

```
summary(aov(Other.deaths ~ type, data = carsafety))
```
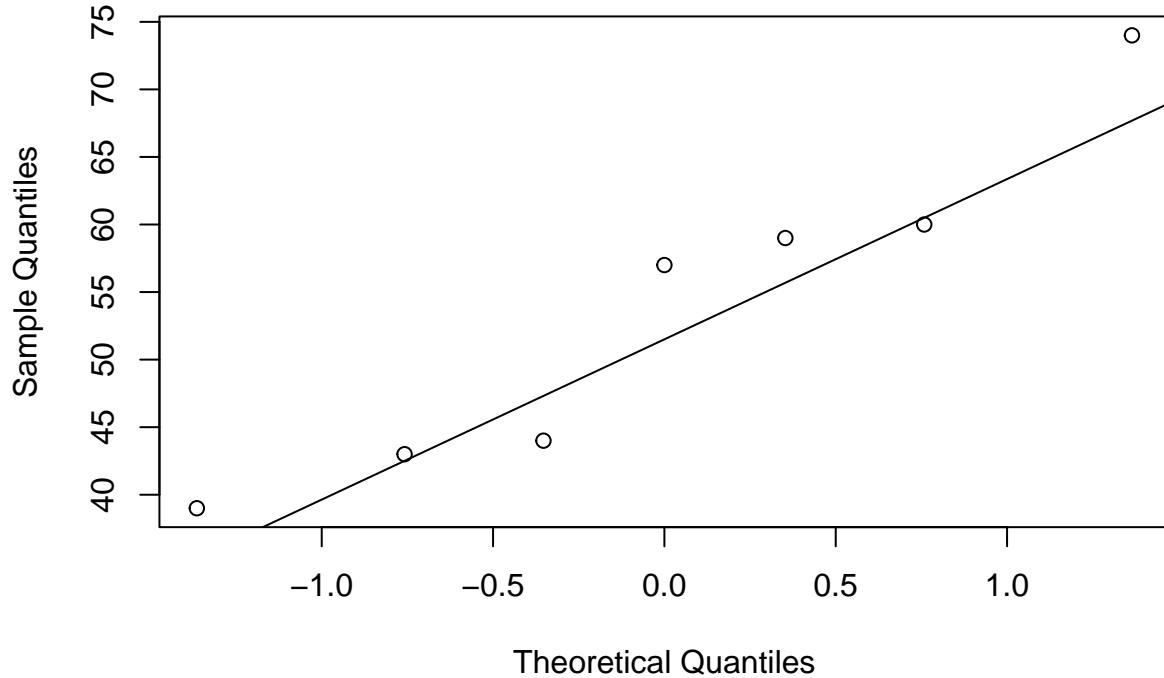
```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## type         6   7996  1332.7   7.289 0.000121 ***
## Residuals   26   4754   182.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
boxplot(Other.deaths ~ type, data = carsafety)
```
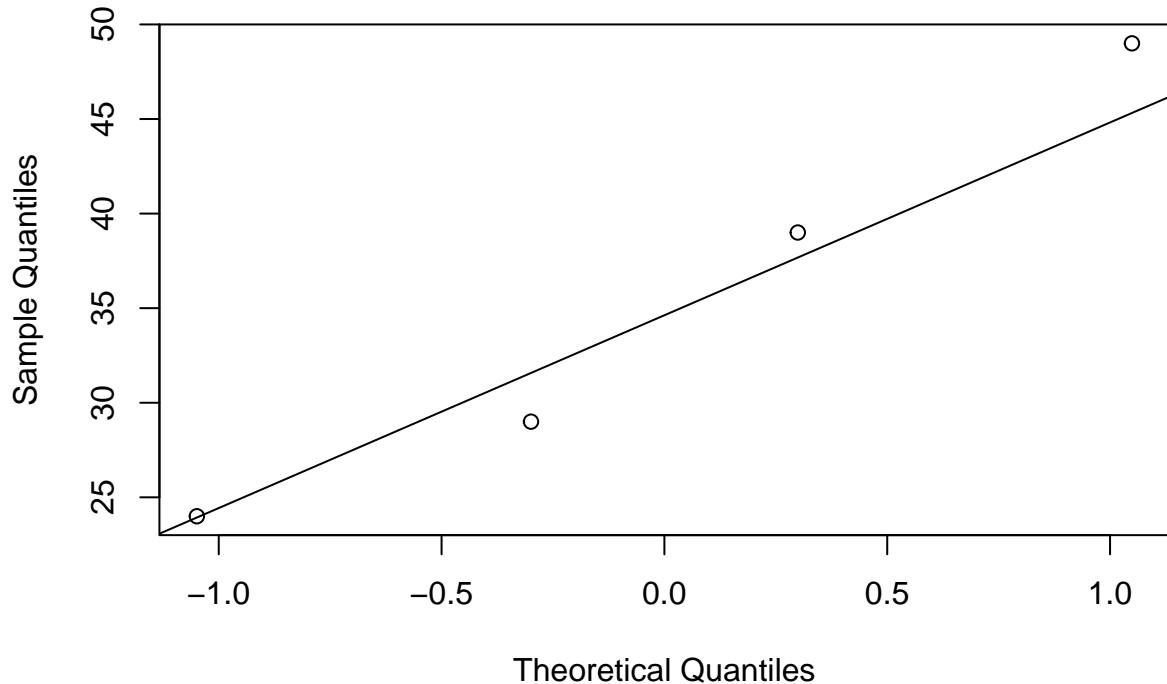
```r
qqnorm(suvData$Other.deaths)
qqline(suvData$Other.deaths)
```
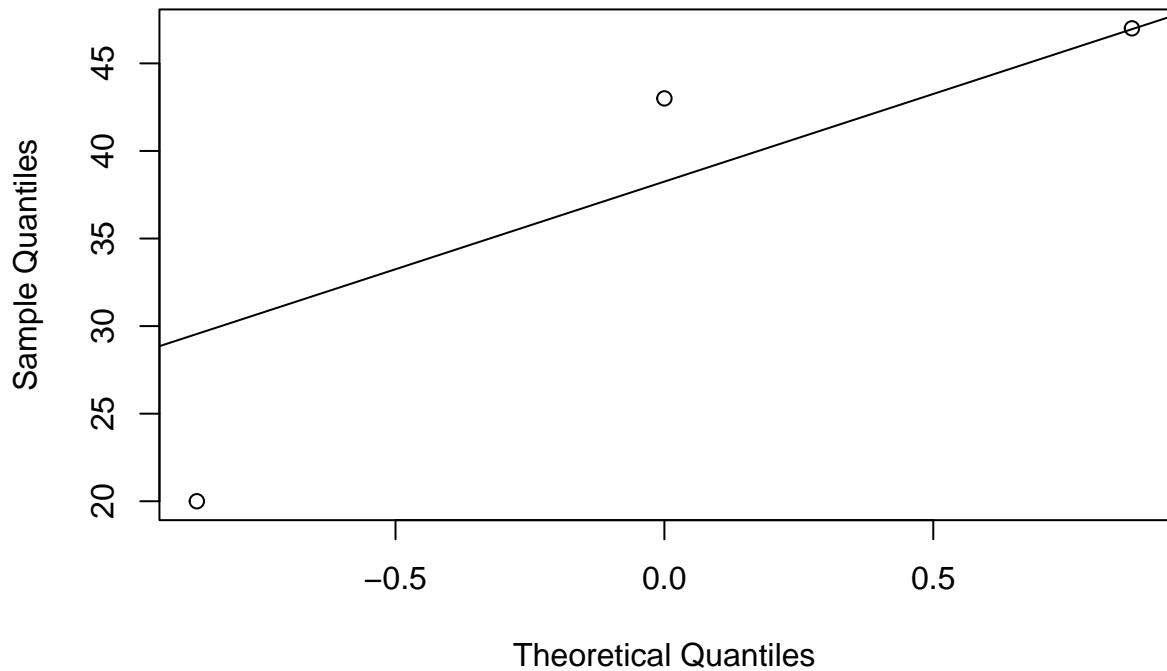
Normal Q-Q Plot


```r
qqnorm(compactData$Other.deaths)
qqline(compactData$Other.deaths)
```
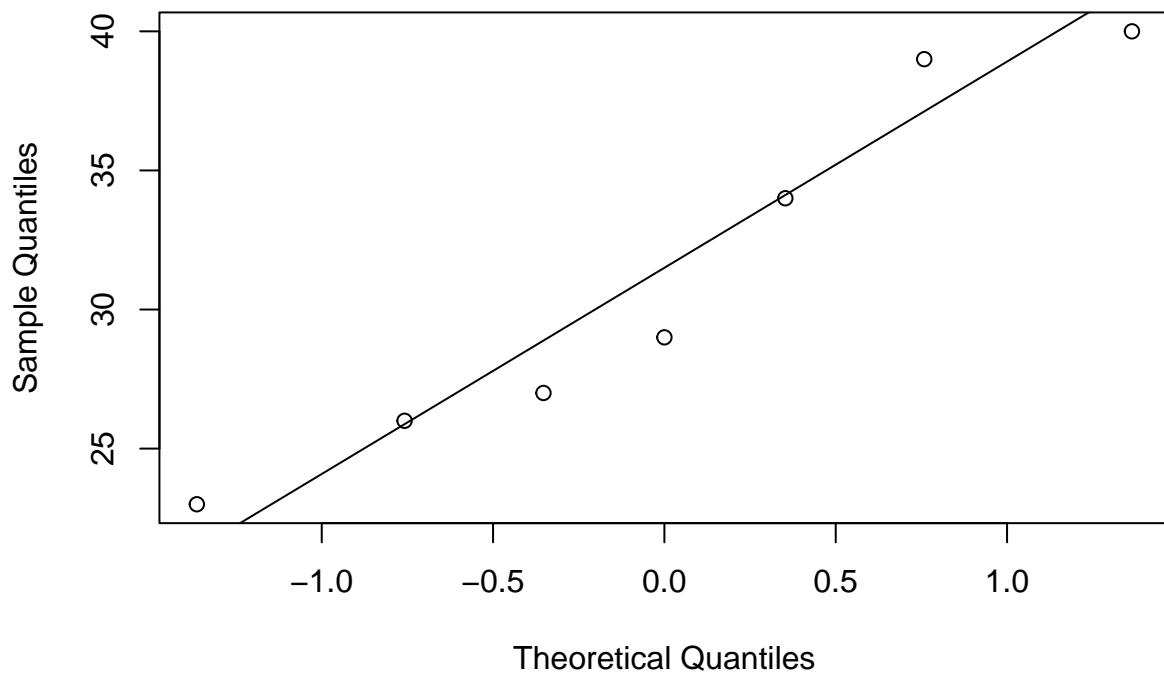
Normal Q–Q Plot


```r
qqnorm(largeData$Other.deaths)
qqline(largeData$Other.deaths)
```
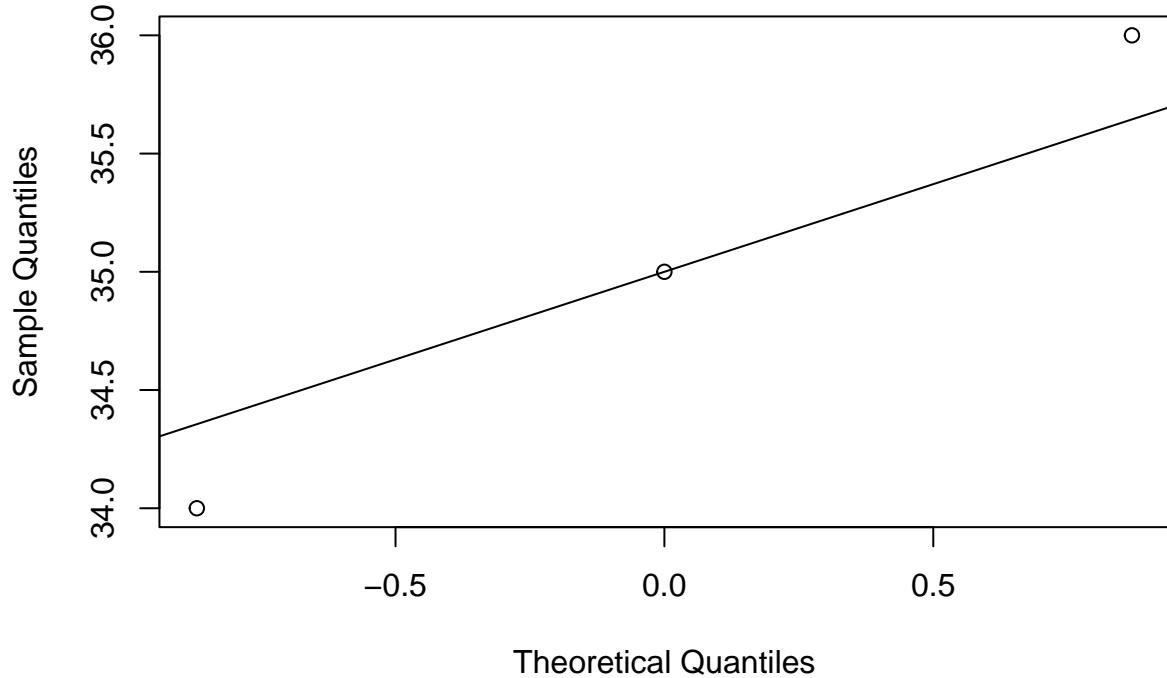
Normal Q–Q Plot


```r
qqnorm(midsizeData$Other.deaths)
qqline(midsizeData$Other.deaths)
```
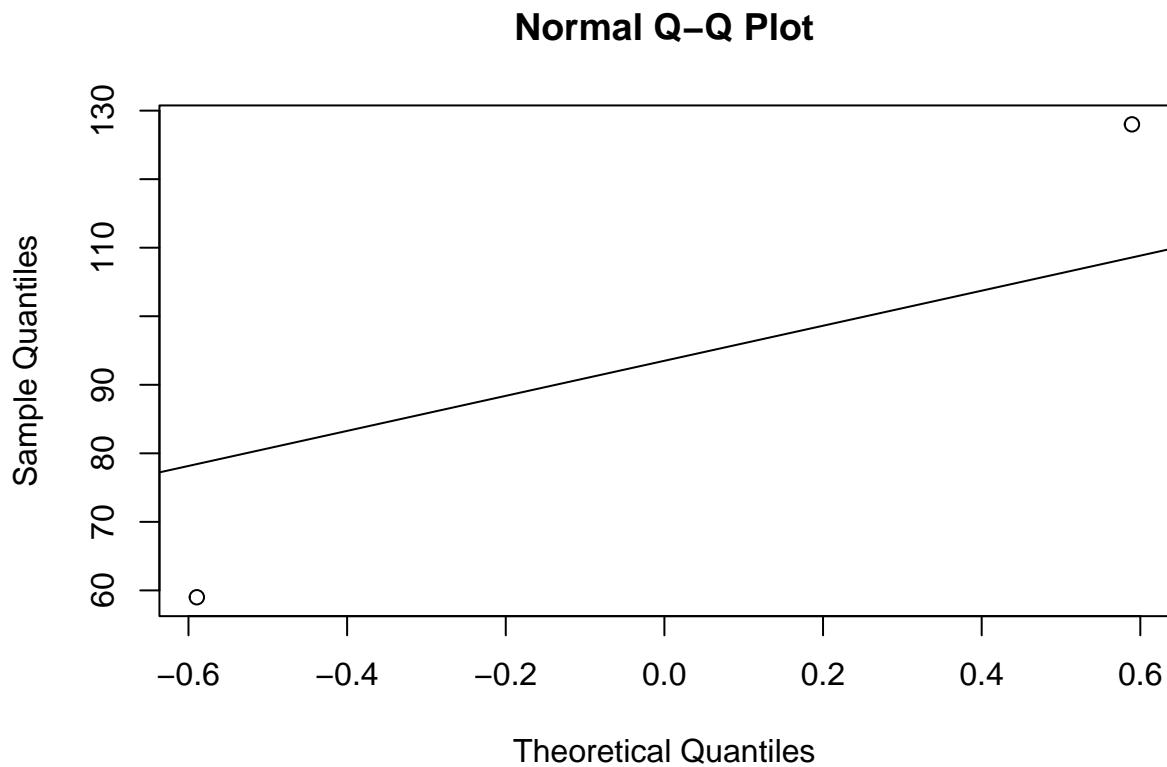
Normal Q–Q Plot


```
qqnorm(minivanData$Other.deaths)
qqline(minivanData$Other.deaths)
```

Normal Q–Q Plot


```
qqnorm(pickupData$Other.deaths)
qqline(pickupData$Other.deaths)
```

Normal Q–Q Plot

Is there a difference in population means? Can you assume equal variances? What about normally distributed populations? There should be a plot associated with each of these.

Response

The ANOVA test strongly suggest we reject the null hypothesis, indicating that there is a difference in means between at least one of the data sets. I don't think we can assume equal variances after visually inspecting the box plot. The minivan test data is very tightly distributed while pickup test data has an extremely large spread. The normality plots appear normal so we can assume norma, populations.

Problem 3 (Verzani problem 12.7)

A manufacturer of point-of-sale merchandise tests three types of ENTER-button markings. They wish to minimize wear, as customers get annoyed when the markings on this button wear off. They construct a test of the three types, and conduct several trials for each. The results, in unspecified units, are recorded in the following table:

Type 1	303	293	296	299	298	
Type 2	322	326	315	318	320	320
Type 3	309	327	317	315		

Is there a difference in wear time among the three types? Answer this using a one-way ANOVA.
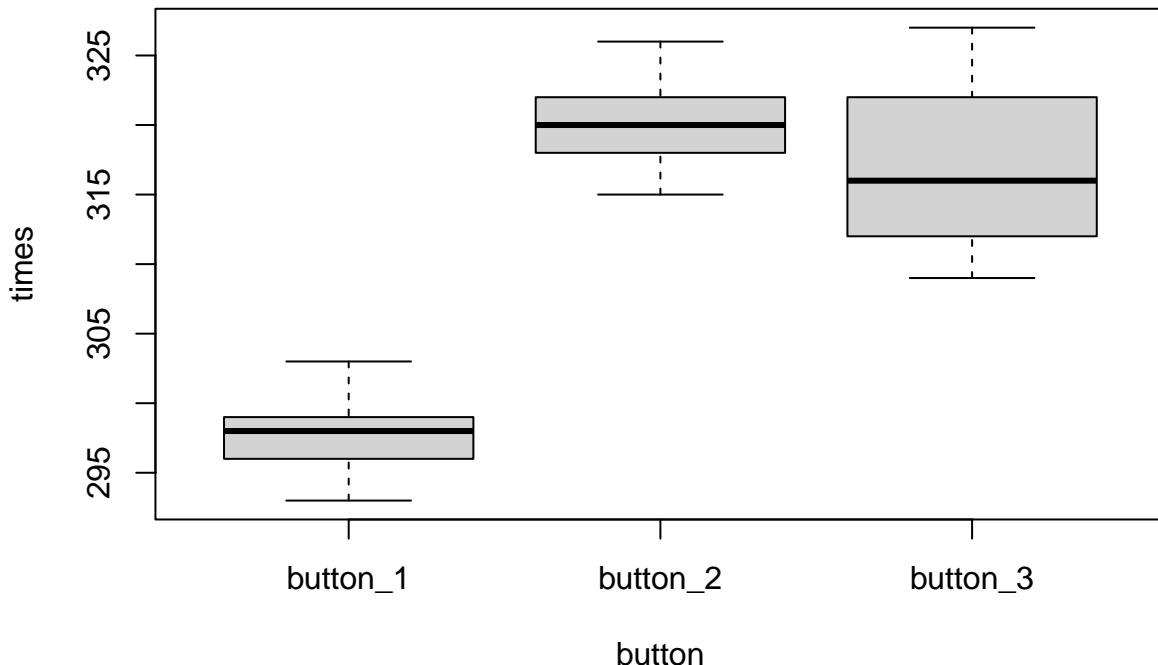
Hint: To create your dataframe, the command `rep(something , times = c(something))` might be helpful. Replace `something` with your own code.

```r
button1times = c(303, 293, 296, 299, 298)
button2times = c(322, 326, 315, 318, 320, 320)
button3times = c(309, 327, 317, 315)

buttontimescombine = c(button1times, button2times, button3times)

button1vec <- rep("button_1", times = 5)
button2vec <- rep("button_2", times = 6)
button3vec <- rep("button_3", times = 4)
buttontype <- c(button1vec, button2vec, button3vec)

button_df <- data.frame(
  "times" = buttontimescombine,
  "button" = buttontype
)
boxplot(times ~ button, data = button_df)
```



```r
oneway.test(times ~ button, data = button_df)
```
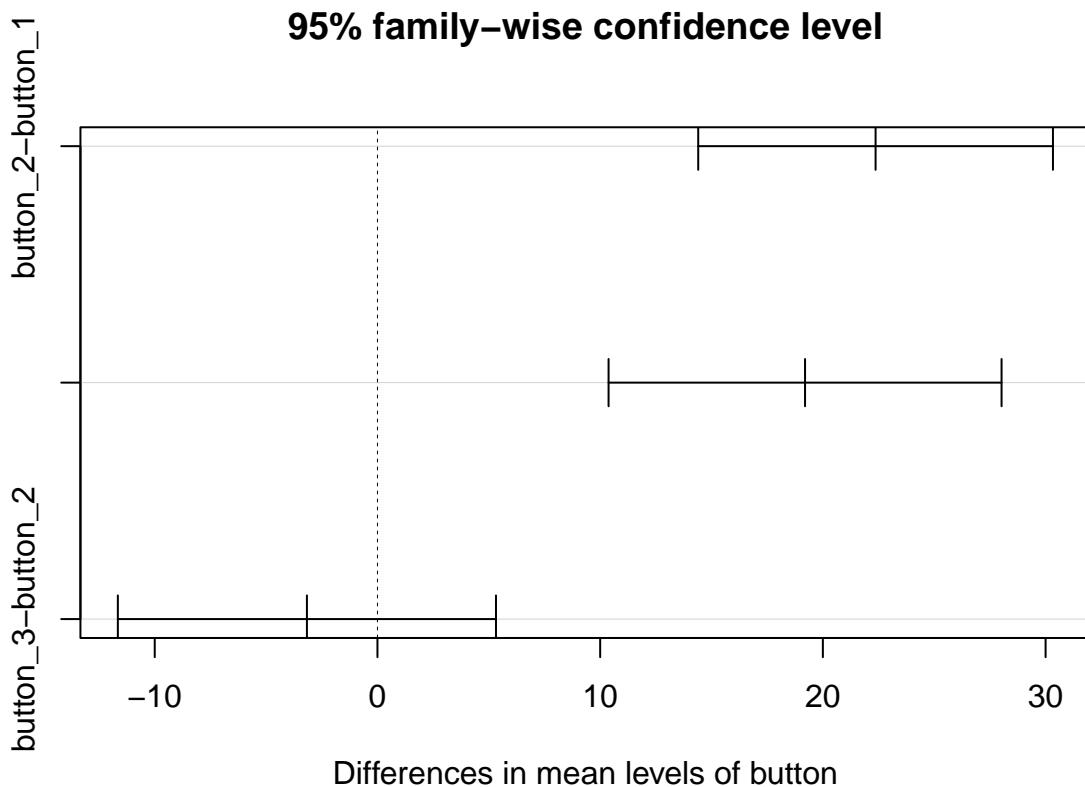
```
## 
##  One-way analysis of means (not assuming equal variances)
## 
```

```
## data:  times and button
## F = 46.475, num df = 2.0000, denom df = 6.4165, p-value = 0.0001522
```

```
tuk <- TukeyHSD(aov(times ~ button, data = button_df))
plot(tuk)
```


95% family–wise confidence level

The ANOVA test supports that there is a difference in wear time between the buttons. visually inspecting the boxplot leads me to believe that button 1 population is significantly different than button 2 or 3. The p-value of the one way analysis is .0001522 which indicates very strong evidence that the means are different

Problem 4 (Verzani problem 12.19)

The `InsectSprays` data set contains a variable `count`, which counts the number of insects and a factor `spray`, which indicates the treatment given.

First perform an analysis of variance to see whether the treatments make a difference (use aov this time). If so, perform the Tukey HSD procedure to see which pairwise treatments differ.

InsectSprays

```
##     count spray
## 1      10     A
## 2       7     A
## 3      20     A
```

```
## 4    14    A
## 5    14    A
## 6    12    A
## 7    10    A
## 8    23    A
## 9    17    A
## 10   20    A
## 11   14    A
## 12   13    A
## 13   11    B
## 14   17    B
## 15   21    B
## 16   11    B
## 17   16    B
## 18   14    B
## 19   17    B
## 20   17    B
## 21   19    B
## 22   21    B
## 23    7    B
## 24   13    B
## 25    0    C
## 26    1    C
## 27    7    C
## 28    2    C
## 29    3    C
## 30    1    C
## 31    2    C
## 32    1    C
## 33    3    C
## 34    0    C
## 35    1    C
## 36    4    C
## 37    3    D
## 38    5    D
## 39   12    D
## 40    6    D
## 41    4    D
## 42    3    D
## 43    5    D
## 44    5    D
## 45    5    D
## 46    5    D
## 47    2    D
## 48    4    D
## 49    3    E
## 50    5    E
## 51    3    E
## 52    5    E
## 53    3    E
## 54    6    E
## 55    1    E
## 56    1    E
## 57    3    E
```

```
## 58     2      E
## 59     6      E
## 60     4      E
## 61    11      F
## 62     9      F
## 63    15      F
## 64    22      F
## 65    15      F
## 66    16      F
## 67    13      F
## 68    10      F
## 69    26      F
## 70    26      F
## 71    24      F
## 72    13      F
```
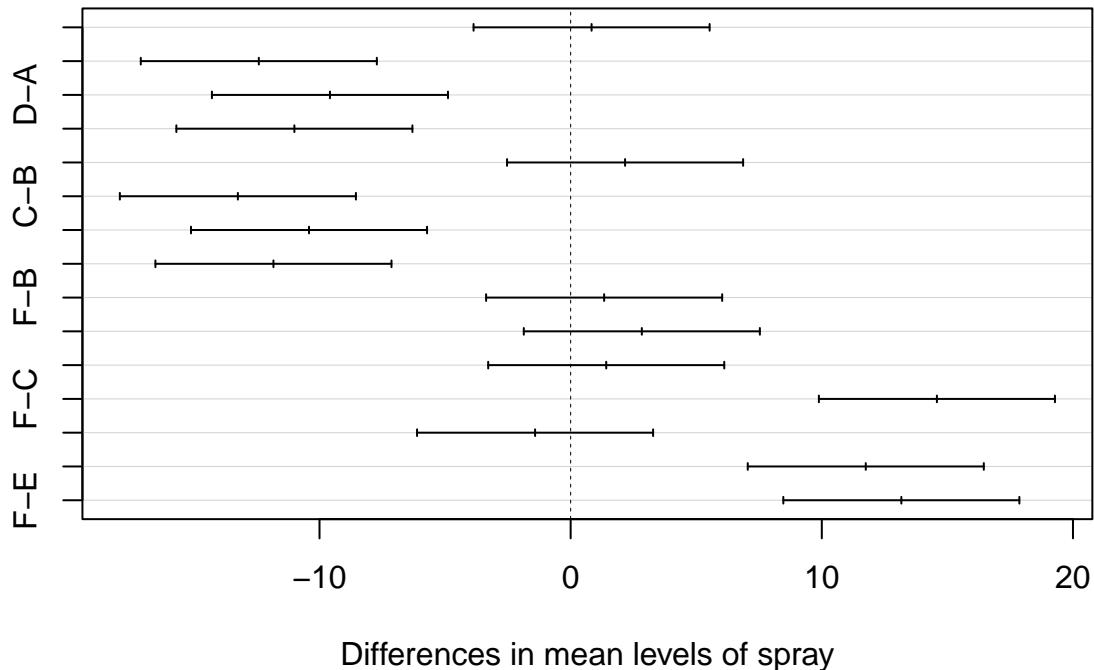
```
summary(aov(count ~ spray, data=InsectSprays))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## spray        5   2669   533.8    34.7 <2e-16 ***
## Residuals   66   1015    15.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
res <- aov(count ~ spray, data=InsectSprays)
plot(TukeyHSD(res))
```

95% family−wise confidence level

Differences in mean levels of spray

```
TukeyHSD(res)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = count ~ spray, data = InsectSprays)
##
## $spray
##              diff        lwr        upr     p adj
## B-A    0.8333333  -3.866075   5.532742 0.9951810
## C-A -12.4166667 -17.116075  -7.717258 0.0000000
## D-A  -9.5833333 -14.282742  -4.883925 0.0000014
## E-A -11.0000000 -15.699409  -6.300591 0.0000000
## F-A   2.1666667  -2.532742   6.866075 0.7542147
## C-B -13.2500000 -17.949409  -8.550591 0.0000000
## D-B -10.4166667 -15.116075  -5.717258 0.0000002
## E-B -11.8333333 -16.532742  -7.133925 0.0000000
## F-B   1.3333333  -3.366075   6.032742 0.9603075
## D-C   2.8333333  -1.866075   7.532742 0.4920707
## E-C   1.4166667  -3.282742   6.116075 0.9488669
## F-C  14.5833333   9.883925  19.282742 0.0000000
## E-D  -1.4166667  -6.116075   3.282742 0.9488669
## F-D  11.7500000   7.050591  16.449409 0.0000000
## F-E  13.1666667   8.467258  17.866075 0.0000000
```

Do the treatments make a difference? How do you know?
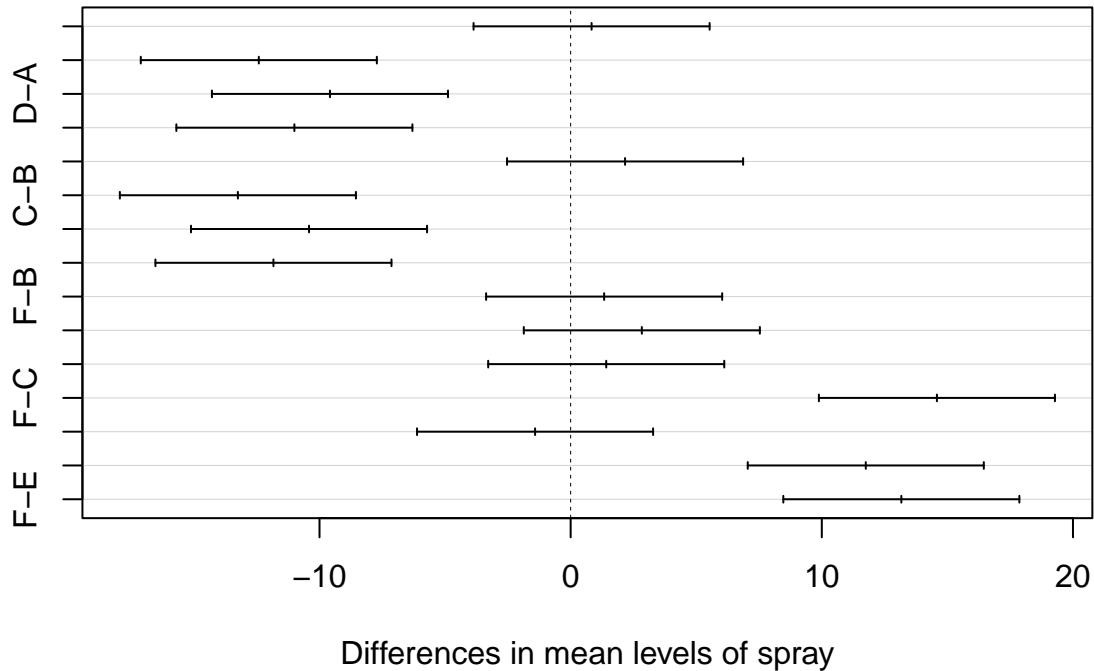
Response:

Yes, the treatments differ. There are several combinations with very low p-values, indicating a difference in means.

If yes, perform the Tukey HSD procedure to see which pairwise treatments differ. Include the plotted version of TukeyHSD. *

```
plot(TukeyHSD(res))
```

95% family–wise confidence level

Differences in mean levels of spray

```
TukeyHSD(res)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = count ~ spray, data = InsectSprays)
##
## $spray
##              diff        lwr       upr     p adj
## B-A    0.8333333  -3.866075   5.532742 0.9951810
## C-A -12.4166667 -17.116075  -7.717258 0.0000000
## D-A  -9.5833333 -14.282742  -4.883925 0.0000014
## E-A -11.0000000 -15.699409  -6.300591 0.0000000
## F-A    2.1666667  -2.532742   6.866075 0.7542147
## C-B -13.2500000 -17.949409  -8.550591 0.0000000
## D-B -10.4166667 -15.116075  -5.717258 0.0000002
## E-B -11.8333333 -16.532742  -7.133925 0.0000000
## F-B    1.3333333  -3.366075   6.032742 0.9603075
## D-C    2.8333333  -1.866075   7.532742 0.4920707
## E-C    1.4166667  -3.282742   6.116075 0.9488669
## F-C  14.5833333   9.883925  19.282742 0.0000000
## E-D  -1.4166667  -6.116075   3.282742 0.9488669
## F-D  11.7500000   7.050591  16.449409 0.0000000
## F-E  13.1666667   8.467258  17.866075 0.0000000
```

What do these results tell us? Is 0 in any of the confidence intervals? What does this mean?

Response:

Several of the pairings have a 95% confidence interval that doesn't capture 0, indicating a statistically significant difference in means. If 0 is captured by the interval in this 95% test then we know that for that pairing the difference in means is not statistically significicant and can be explained by random variations in the samples.