STA303 - Assignment 1

Last name: Akad
First name: Dogan
Student ID: 1001386083
Course section: STA303H1S-L0101

Feb 5th, 2017

Q1 (a-d) - Data 1: working output

(a) Calculate the means and standard deviations of output for each workman. make a boxplot comparing the part output for the 10 workmen, give a short comment for the boxplot produced.

```
library(ggplot2)
work = read.table("/Users/doganakad/Desktop/uoft/second semester/sta303/Assignments/A1/workmandata.csv"
str(work) # check the type of variables in this data
```

```
## 'data.frame':    200 obs. of  2 variables:
##  $ workman: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ y      : int  318 289 309 317 286 281 284 288 293 264 ...
```

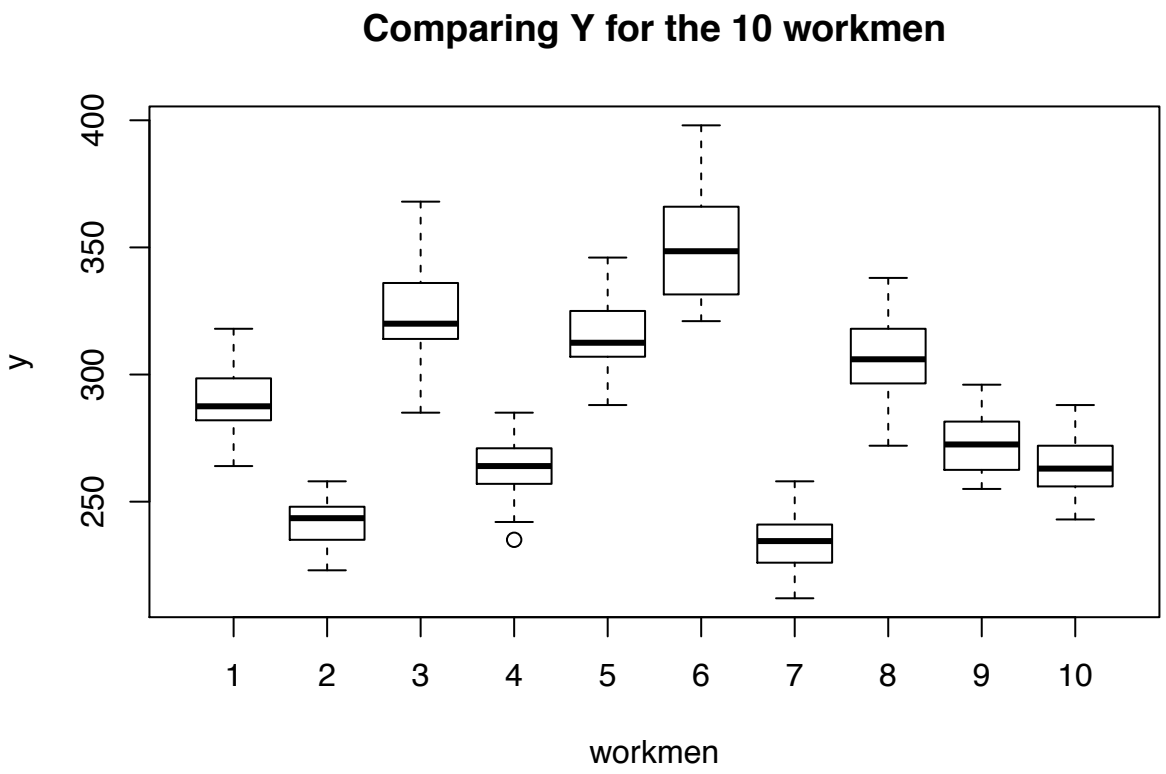
```
work$workman = as.factor( work$workman) # put workman into a factor variable
y = work$y
workman = work$workman

# The means and sd of Y for each workman
with(work ,tapply(y, workman, mean))
```

```
##      1      2      3      4      5      6      7      8      9     10
## 290.25 241.70 324.65 262.85 314.85 350.70 234.80 306.60 273.45 263.90
```

```
with(work, tapply(y, workman, sd))
```

```
##         1         2         3         4         5         6         7
## 16.039015  9.608658 21.086975 11.430684 14.676422 23.517295 11.491874
##         8         9        10
## 16.109494 12.407447 12.130432
```

```
# the boxplot
boxplot(y~workman,data=work, main="Comparing Y for the 10 workmen", xlab="workmen", ylab="y")
```



Comments on boxplot : Looking at the boxplot we can say that each workmen has a different daily part output value. Groups such as 2 and 7 have the lowest daily part output value whereas group 6 has the highest y value with median 348.5. Group 3's median is closer to the lower quartile which can indicate the distribution of y values can be negatively skewed. There's an outlier at group 4.

(b) Applying one-way ANOVA to this data, testing the equality of the output means for the ten workmen. State the null and alternative hypothesis for the p-value in ANOVA output. How significant is the result ?

```
# one way anova
summary(aov(y~workman, data=work))
```

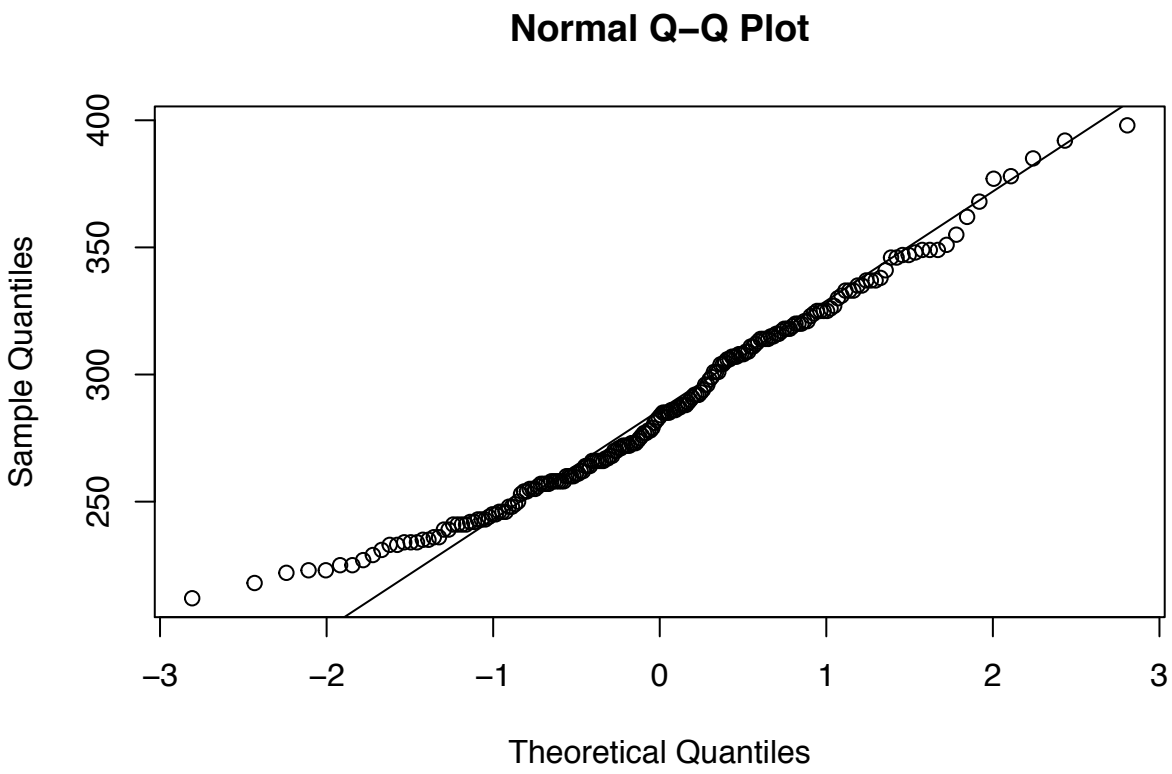
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## workman       9 254380   28264   118.5 <2e-16 ***
## Residuals   190  45335     239
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null and alternative hypothesis for the F test: Null hypothesis: Output means for all the ten workmen are the same. Alternative hypothesis: At least two of the workmen have different output means.

How significant is the result?: Because our p value is very low, we have strong significant evidence to reject our null hypothesis. Therefore, at least two of the workmen have different output means and the result is significant.

(c) ANOVA assumes that the data in each group are distributed normally. This assumption is equivalent to saying that the residuals of the best-fitting model are distributed normally. Check the normality assumption by doing a qqnorm plot in conjunction with qqline based on the residuals from the linear regression model fitting. What conclusion do you have from the plot?

```
# qqnorm plot
qqnorm(y); qqline(y)
```



Comments on normal Q-Q plot: In the lower(-3- -1.5) and upper(2-3) quantiles, residuals do not fall on to the straight line. Other than these quantiles, residuals follow a straight line which indicates that between -1 and 2 the residuals follow a normal distribution. Therefore, the residuals in general in this plot are not perfectly normally distributed.

(d) Examine the output variability for the ten workmen using the Bartlett test. What is your conclusion?

```
# bartlett test
bartlett.test(y~workman, data=work)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  y by workman
## Bartlett's K-squared = 28.792, df = 9, p-value = 0.0007024
```

Conclusion: With 0.05 significance level, p-value is less than the significance level. Therefore, we strongly reject our null hypothesis and can conclude that 10 groups don't have equal variances.