36-401 HW2

Sylvia Shuyuan Ding (Shuyuand)

9/9/2019

Problem #1

Load data:

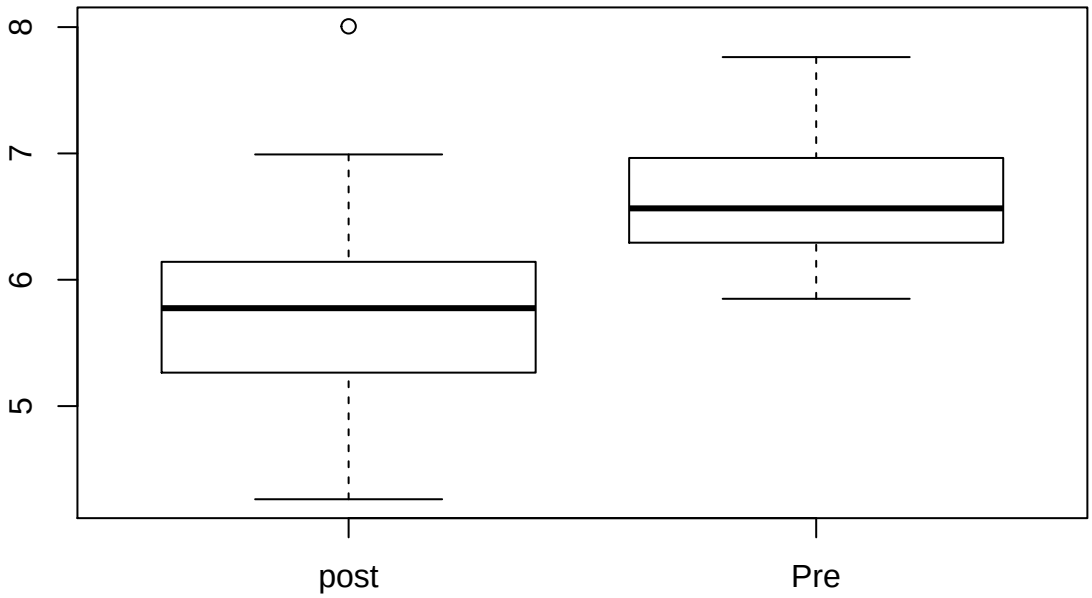
```
setwd("/Users/silviading/Desktop/CMU\U0001f33c/Fall 2019/36401/HW/")
hc.levels = read.table("HC-Levels.txt",header = TRUE)
```

```
hc.levels$logPollution = log(hc.levels$Pollution)
head(hc.levels)
```

```
##   Pollution Period logPollution
## 1      2351    Pre     7.762596
## 2      1293    Pre     7.164720
## 3       541    Pre     6.293419
## 4      1058    Pre     6.964136
## 5       411    Pre     6.018593
## 6       570    Pre     6.345636
```

(a).

```
boxplot(logPollution~Period, data = hc.levels)
```

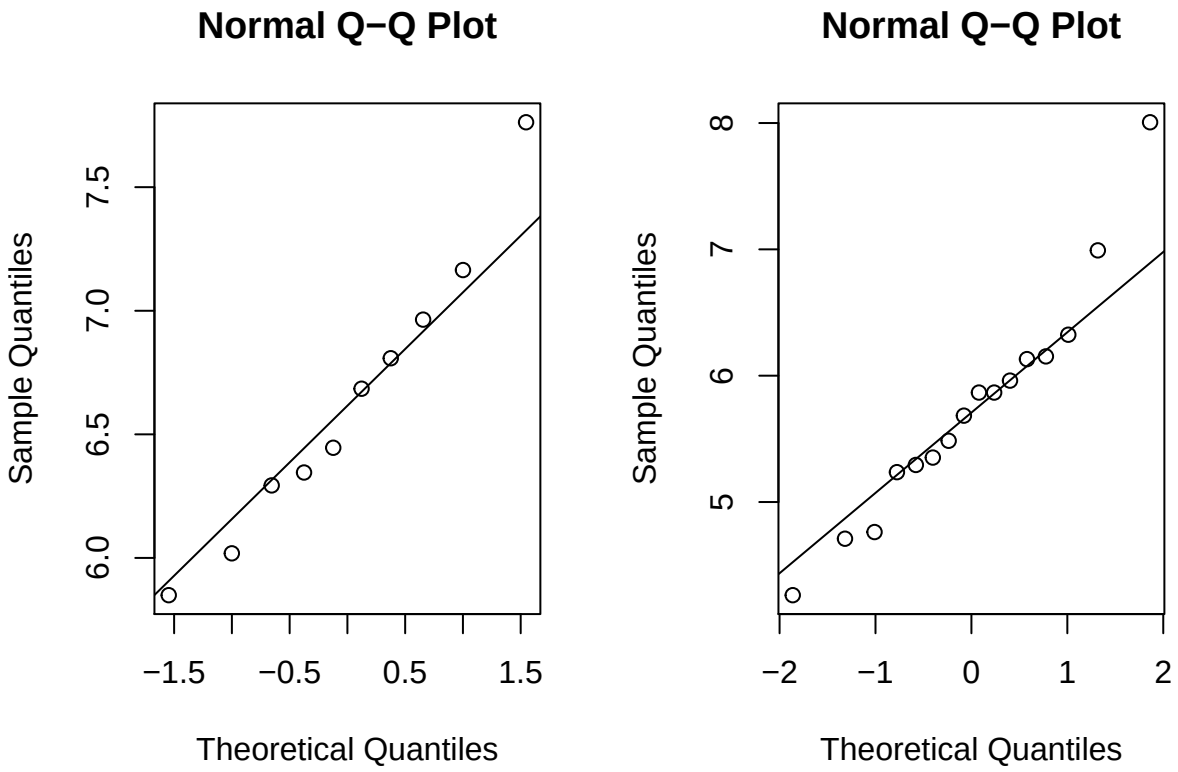


From the distribution of each infividual graph, we can see that the range of logPollution for post period is wider than that of pre period, which indicates the median of pre period may be more informative than that of post period. Also, The post period has one really big outlier so that the distribution is right skewed. IQR for post period is also wider than IQR for pre period. It is also interesting that the difference between min and Q1 is distinguishably bigger than the difference between Q3 and max for pre period.

Then compare the between the two graphs, we see that overall, min, max, and median for pre are all higher than the according values of post period, which indicates to us that it is highly possible that the emission control is working and log(Pollution) effectively decreased as the control took place.

(b).

```
par(mfrow = c(1,2))
qqnorm(hc.levels$logPollution[hc.levels$Period == "Pre"])
qqline(hc.levels$logPollution[hc.levels$Period == "Pre"])
qqnorm(hc.levels$logPollution[hc.levels$Period == "post"])
qqline(hc.levels$logPollution[hc.levels$Period == "post"])
```



The distribution seems like violates the nomality assunption, particularly due to the high outliers.

(c)

```
t.test(data = hc.levels, logPollution~Period, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  logPollution by Period
## t = -2.7341, df = 24, p-value = 0.01156
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.5414286 -0.2153153
## sample estimates:
## mean in group post  mean in group Pre
##           5.755297           6.633669
```

Interpretation:The 95% interval is [-1.5414286, -0.2153153], and from this result, we can conclude that we are 95% confident that the difference of mean between logPollution for pre and post period falls in the range between -1.5414286 and -0.2153153. And since 0 does not fall in the range, we conclude there is a difference of mean in pollution level in pre and post period, which corresponds to our observation from the boxplots.

```r
t.test(data = hc.levels, Pollution~Period, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Pollution by Period
## t = -1.4301, df = 24, p-value = 0.1656
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -938.8829  170.3079
## sample estimates:
## mean in group post  mean in group Pre
##           506.3125           890.6000
```

The t-confidence interval for the difference in means of the untransformed pollution is really big and it is [-938.8829, 170.3079] and it also includes 0. Since the confidence interval is too big, it is not informative.

(d) Abstract

Investigators have been interested in whether the standards established by federal government for automobile emission control in 1967 have effectively controlled pollution from Hydrocarbon (HC, in unit of ppm). Specifically, they are interested in whether there is any difference in automobile pollution levels during the period prior (PRE) to the establishment of the standard (1967) compared to the period after (POST). To address this problem, pollution level was measured for 26 times randomly, either prior to the standard (PRE) or post the standard (POST). Since the data is positively skewed and not noramlly distributed, we will analyze by making log transformation on it and also conduct two-sample t-test to find 95% confidence interval for the difference in means of the log-transformed values between two time periods.