HW6_Report

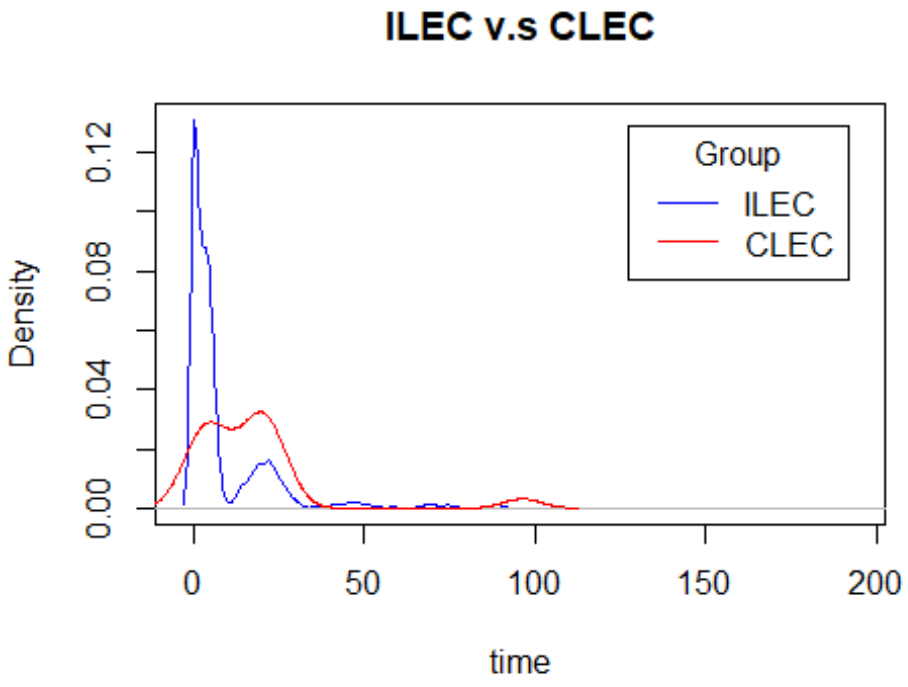
107023058

Question 1:

Verizon claims that mean response time for ILEC and CLEC customers are the same, but the PUC would like to test if CLEC customers were facing greater response times.

(a). Visualize Verizon's response times for ILEC vs. CLEC customers

```
raw_data<-read.csv('D:/Retro/NTHU/課程講義/大三/計算統計於商業分析之應用/HW5/verizon.csv',header = T)
# Loading Data
time_ILEC <- subset(raw_data,Group == "ILEC")$Time
time_CLEC <- subset(raw_data,Group == "CLEC")$Time

plot(density(time_ILEC),col='blue',xlab='time',main = 'ILEC v.s CLEC')
lines(density(time_CLEC),col='red')
legend('topright',inset=.05,title = 'Group',c('ILEC','CLEC'),lty = c(1,1),col=c("blue","red"))

# label for two lines in the graph
```



(b). test the difference between the mean of ILEC sample response times versus the mean of CLEC sample response times.

Null hypothesis (h0): difference of the two mean = 0

Alternative hypothesis (h1): difference of the two mean /= 0

```
t.test(time_ILEC,time_CLEC,var.equal = FALSE,conf.level = 0.99)

##
##  Welch Two Sample t-test
##
## data:  time_ILEC and time_CLEC
## t = -1.9834, df = 22.346, p-value = 0.05975
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -19.588967   3.393927
## sample estimates:
## mean of x mean of y
##  8.411611 16.509130
```

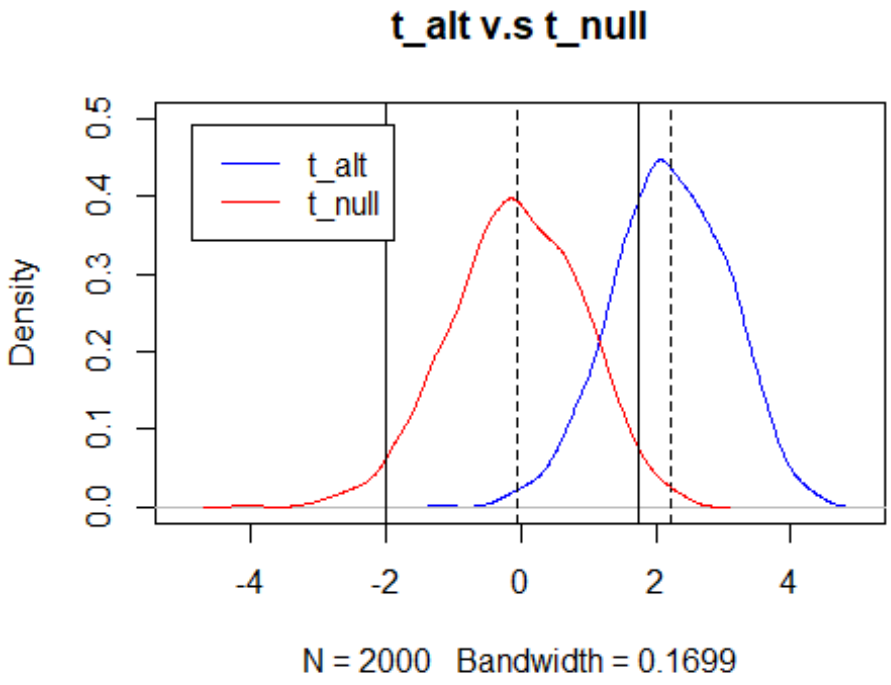
By the result of t.test(), we can see that the p-value is 0.0597,which is larger than alpha = 0.01, besides, the confidence interval also contain 0.

As the result, we can say that we do not reject null hypothesis.

(c). Using bootstrapping to estimate bootstrapped null and alternative values of t .

```
sample_bootstrap <- function(sample0) {
  resample <- sample(sample0, length(sample0), replace=TRUE)
  return(resample)
}

boot_ILEC<-replicate(2000,sample_bootstrap(time_ILEC))
boot_CLEC<-replicate(2000,sample_bootstrap(time_CLEC))

#bootstrapped samples of ILEC against bootstrapped samples of CLEC (alt
 t-values)
#bootstrapped samples of ILEC against the original ILEC sample (null t-
values)

bootstrap_null_alt <- function(sample0,hyp_mean){
  resample<-sample(sample0,length(sample0),replace = TRUE)
  resample_se<-sd(resample)/sqrt(length(resample))

  t_alt <- (mean(resample)-hyp_mean)/resample_se
  t_null <- (mean(resample)-mean(sample0))/resample_se

  return(c(t_alt,t_null))
}
```

```r
set.seed(42)
boot_t_stat <- replicate(2000,bootstrap_null_alt(time_ILEC,7.6))
boot_t_alt <- boot_t_stat[1,]
boot_t_null <- boot_t_stat[2,]

plot(density(boot_t_alt),main='t_alt v.s t_null',xlim = c(-5,5),ylim=c
(0,0.5),col='blue')
abline(v=mean(boot_t_alt),lty="dashed")
lines(density(boot_t_null),col = 'red')
abline(v=mean(boot_t_null),lty='dashed')
abline(v=quantile(boot_t_null,c(0.025,0.975)))
legend('topleft',inset=.05,c('t_alt','t_null'),lty = c(1,1),col=c("blue
","red"))
```




```r
t.test(boot_t_null,boot_t_alt,var.equal = FALSE)

##
##  Welch Two Sample t-test
##
## data:  boot_t_null and boot_t_alt
## t = -77.715, df = 3932.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.330920 -2.216206
## sample estimates:
##   mean of x   mean of y
## -0.06294922  2.21061373
```

Based on the result above, p-value < 2.2e^-16 < 0.01, reject the null hypothesis.

Question 2:

Test whether the variance of ILEC response times is different than the variance of CLEC response times.

(a). Setting null and alternative hypotheses.

```
var(time_ILEC)

## [1] 215.7973

var(time_CLEC)

## [1] 380.3895
```

Because var(time_CLEC) > var(time_ILEC), we set the null hypothesis h0 as $\frac{var(timeCLEC)}{var(timeILEC)} = 1$, and h1: $\frac{var(timeCLEC)}{var(timeILEC)} /= 1$.

(b). Try traditional statistical methods first

```
# (i). find the f-value
f_value = var(time_CLEC)/var(time_ILEC)
f_value

## [1] 1.762717

# (ii). cut-off value
qf(p=0.95,df1=length(time_CLEC)-1,df2=length(time_ILEC)-1)

## [1] 1.548476

# f_value > critical value => reject h0
var.test(time_CLEC,time_ILEC,alternative = 'greater')

##
##  F test to compare two variances
##
## data:  time_CLEC and time_ILEC
## F = 1.7627, num df = 22, denom df = 1663, p-value = 0.01582
## alternative hypothesis: true ratio of variances is greater than 1
## 95 percent confidence interval:
##  1.138356      Inf
## sample estimates:
## ratio of variances
##           1.762717
```

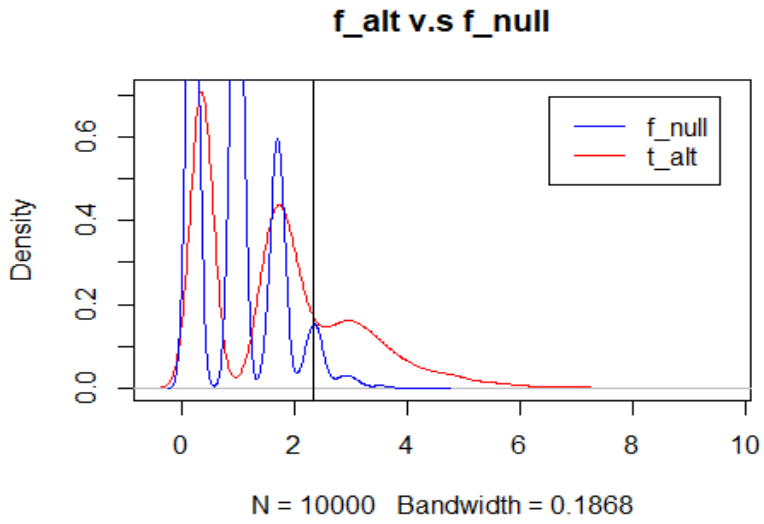
As the result above, f_value is larger than cut-off value, so we will reject the null hypothesis.

(c). Bootstrap

```
# (i). Create bootstrapped values of the F-statistic, for both null and
 alternative hypotheses
set.seed(43)
sd_test<-function(larger_sd,smaller_sd){
  resample_larger_sd <- sample(larger_sd,length(larger_sd),replace=TRU
E)
  resample_smaller_sd <- sample(smaller_sd,length(smaller_sd),replace=T
RUE)
  f_alt<-var(resample_larger_sd)/var(resample_smaller_sd)
  f_null<-var(resample_larger_sd)/var(larger_sd)
  return(c(f_alt,f_null))
}

f_stats <- replicate(10000,sd_test(time_CLEC,time_ILEC))
f_alt<-f_stats[1,]
f_null<-f_stats[2,]


# (ii). the 95% cutoff value according to the bootstrapped null values
of F
quantile(f_null,0.95)
```

```
##      95%
## 2.331735
```

```
# (iii). Visualization
plot(density(f_alt),col='red',main = 'f_alt v.s f_null')
lines(density(f_null),col='blue')
abline(v=quantile(f_null,0.95))
legend('topright',inset=.05,c('f_null','t_alt'),lty = c(1,1),col=c("blu
e","red"))
```



By the result above, rejecting the null hypothesis.

Also, I notice that the graph is a little weird. If my code isn't wrong, then I guess the reason why it looks weird is that the two data set (time_ILEC , time_CLEC) is not normally distributed.

Question 3.

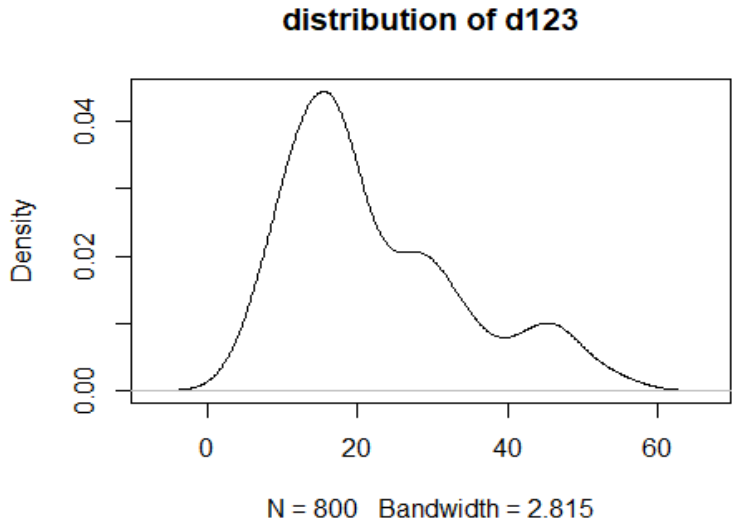
Try to see when we should use the non-parametric bootstrap and when we might be better off with traditional statistical approaches.

(a). Create a function to see if key statistics/assumptions of normality are met in our distributions.

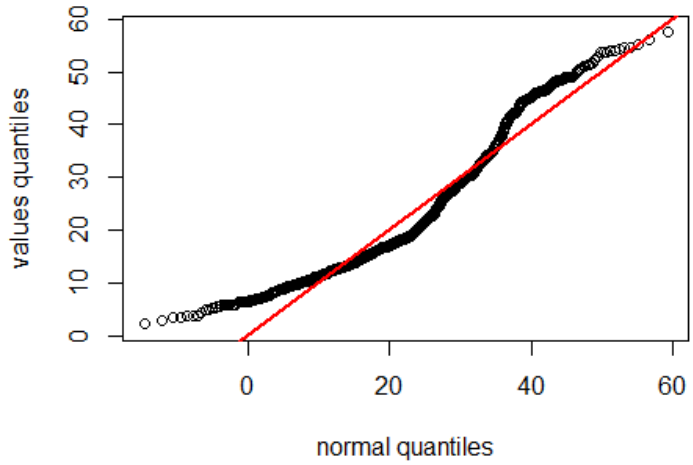
```
norm_qq_plot <- function(values){
  probs1000 <- seq(0, 1, 0.001)
  # Create a sequence of probability numbers from 0 to 1, with ~1000 pr
obabilities in between
  q_vals <- quantile(values,prob=probs1000)
  # Calculate ~1000 quantiles of our values
  q_norm <- qnorm(probs1000,mean(values),sd(values))
  # Calculate ~1000 quantiles of a perfectly normal distribution with t
he same mean and standard deviation as our values
  plot(q_norm, q_vals, xlab="normal quantiles", ylab="values quantiles
")
  # Create a scatterplot comparing the quantiles of a normal distributi
on versus quantiles of values
  abline(a=0,b=1, col="red", lwd=2)
  # draw a red line with intercept of 0 and slope of 1
}
```

(b).

```
set.seed(978234)
d1 <- rnorm(n=500, mean=15, sd=5)
d2 <- rnorm(n=200, mean=30, sd=5)
d3 <- rnorm(n=100, mean=45, sd=5)
d123 <- c(d1, d2, d3)

plot(density(d123),main = 'distribution of d123')
```



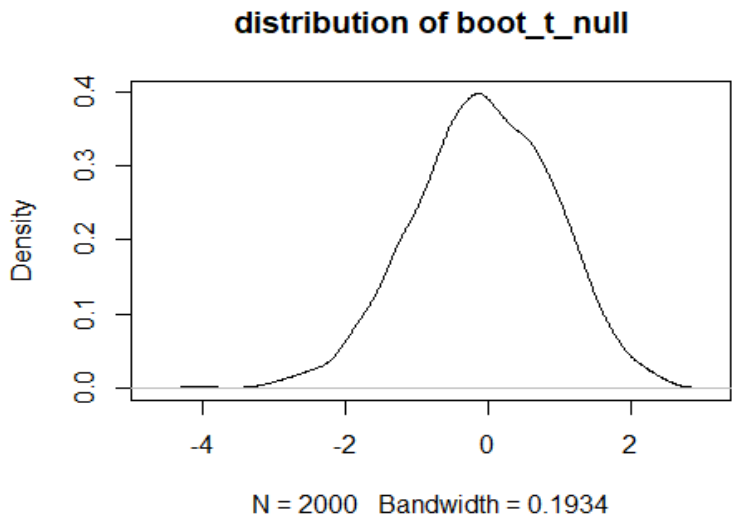

```
norm_qq_plot(d123)
```



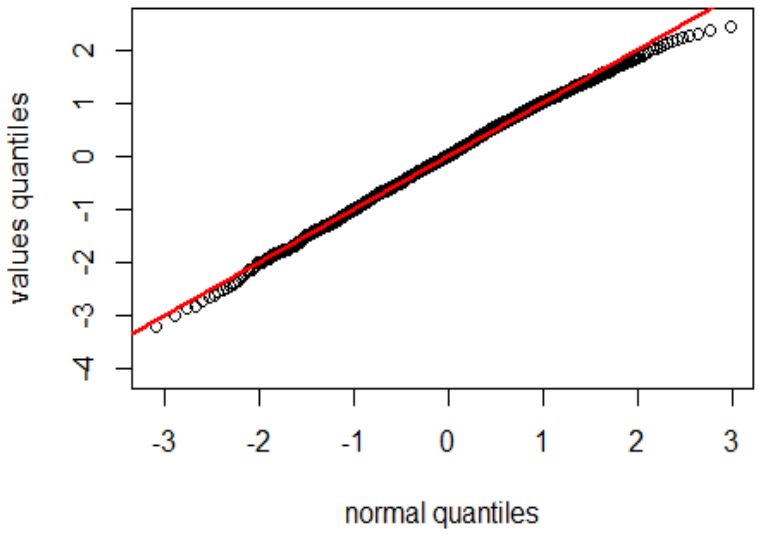
the graph in Q-Q plot is distributed around $y = x$, so it's normally distributed.

(c). Check if the bootstrapped distribution of null t-values in question 1c was normally distributed.

```
plot(density(boot_t_null),main = 'distribution of boot_t_null')
```



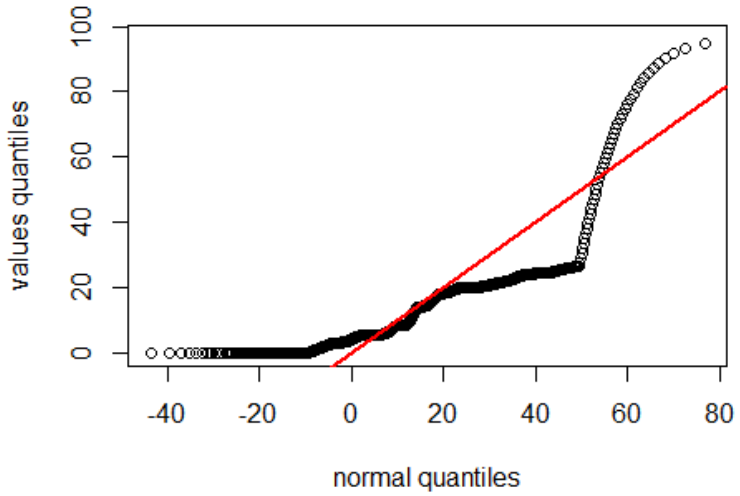

```
norm_qq_plot(boot_t_null)
```



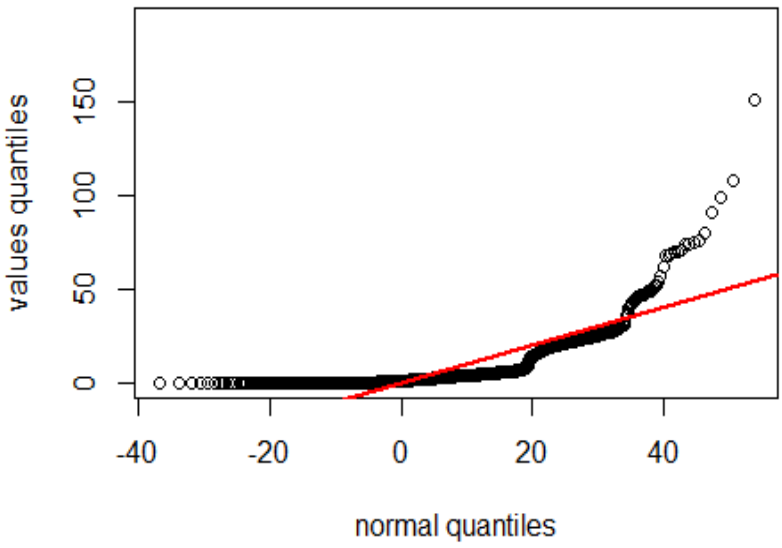
From the graph of Q-Q plot, I think that it's not normally distributed.

(d). Check if the two samples we compared in question 2 could have been normally distributed.

```
norm_qq_plot(time_CLEC)
```




```
norm_qq_plot(time_ILEC)
```



From the graph of Q-Q plot, I think that they are neither normally distributed.