# *BACS HW (Week6)*

*106070038*

*2021-04-11*

Special thanks to 106070020 for discussing with me.

## *Question 1*

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

```
library(readr)
verizon <- read_csv("verizon.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Time = col_double(),
##   Group = col_character()
## )
```

```
ILEC = verizon[which(verizon$Group == "ILEC"), ]
CLEC = verizon[which(verizon$Group == "CLEC"), ]

plot(density(ILEC$Time),col="red", main = "Density plot of ILEC and CLEC time",)
lines(density(CLEC$Time),col="blue", lty="dashed")

legend("topright", legend=c("ILEC", "CLEC"),col=c("red", "blue"), lty=1:2, cex=1.8, box.lty=0)

#var(ILEC$Time)
#var(CLEC$Time)
```

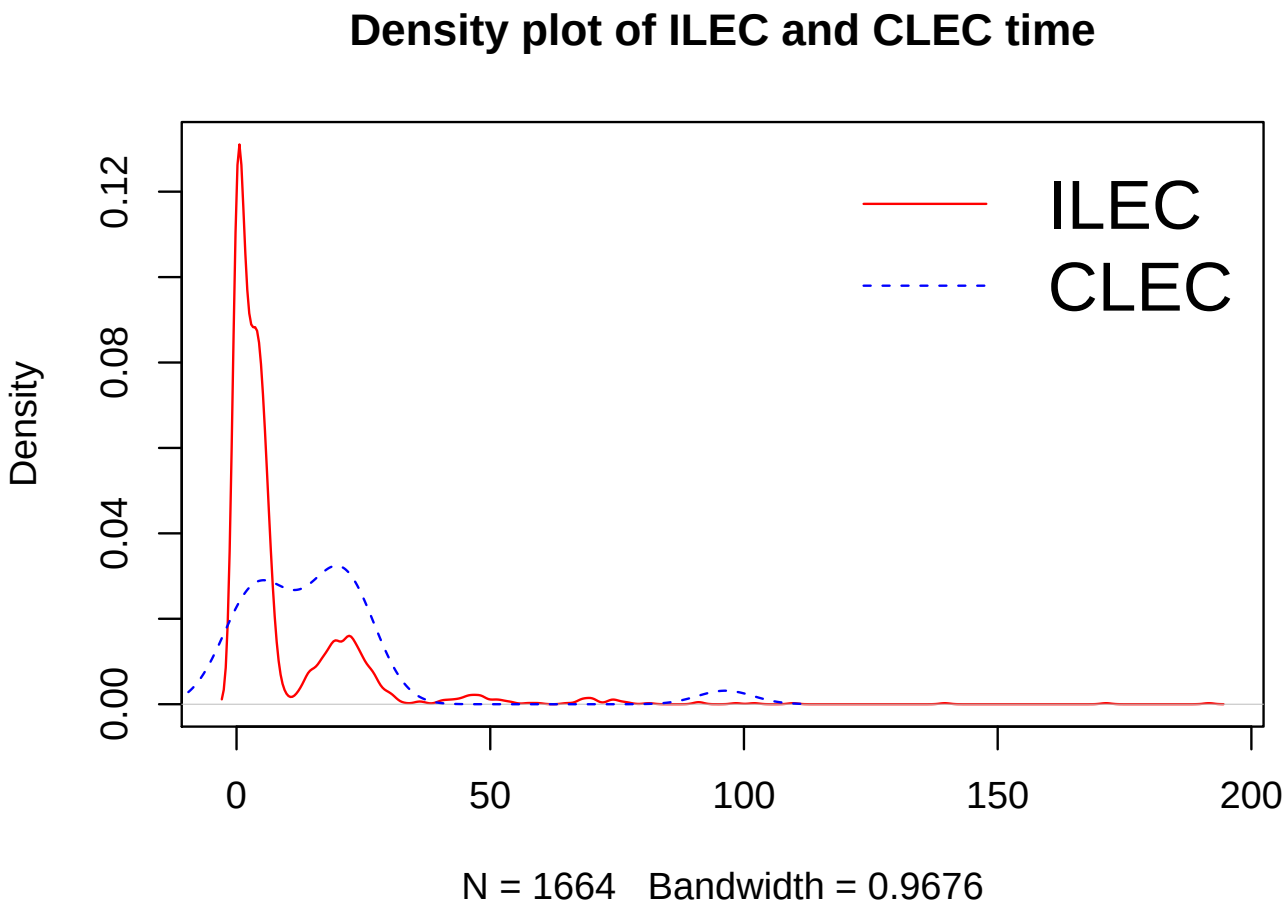
*b. Use the appropriate form of the t.test() function to test the difference between the mean of ILEC sample response times versus the mean of CLEC sample response times. From the output of t.test():*

*b-i. What are the appropriate null and alternative hypotheses in this case?*

- Answer:
  - $H_{null}$: $\mu_{ILEC} = \mu_{CLEC}$
  - $H_{alt}$: $\mu_{ILEC} \neq \mu_{CLEC}$

*b-ii. Based on output of the t.test(), would you reject the null hypothesis or not?*

```
t.test(ILEC$Time, CLEC$Time, conf.level = 0.99)
```


Density plot of ILEC and CLEC time
Density
0.00
0.04
0.08
0.12
0
50
100
150
200
ILEC
CLEC
N = 1664  Bandwidth = 0.9676

```
##
##  Welch Two Sample t-test
##
## data:  ILEC$Time and CLEC$Time
## 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
```

- Answer: p = 0.05975 > 0.01, Not reject $H_{null}$

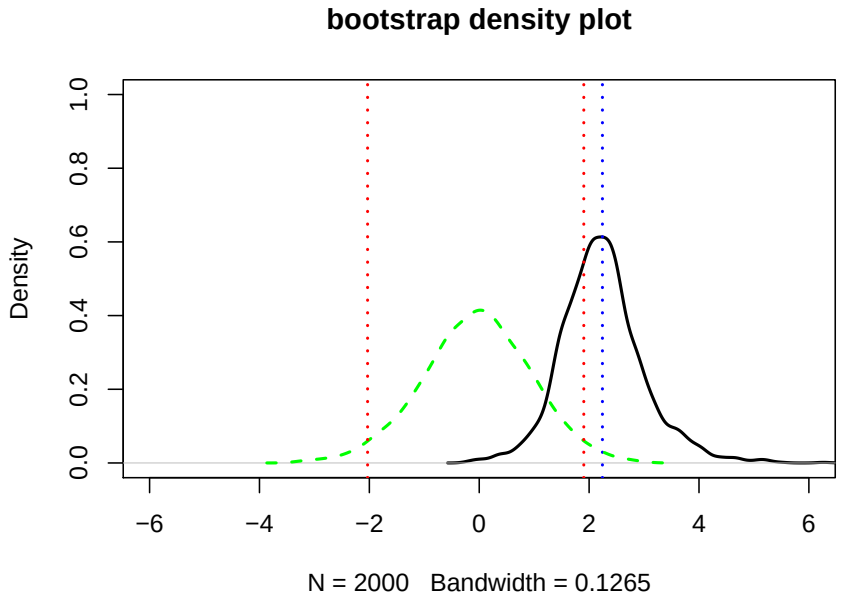
*c. Let's try this using bootstrapping*

```
set.seed(321)
bootstrap_null_alt <- function(sample0, sample1) {
  resample <- sample(sample0, length(sample0), replace=TRUE)
  resample_se <- sd(resample) / sqrt(length(resample))
  resample_1 <- sample(sample1, length(sample1), replace=TRUE)
  resample_se_1 <- sd(resample_1) / sqrt(length(resample_1))

  t_stat_alt <- (mean(resample_1) - mean(resample)) / sqrt(resample_se^2+resample_se_1^2)
  t_stat_null <- (mean(resample) - mean(sample0)) / resample_se
  c(t_stat_alt, t_stat_null)
}
boot_t_stats <- replicate(2000, bootstrap_null_alt(ILEC$Time, CLEC$Time))
```

*c-i. Plot a distribution of the bootstrapped null t-values and alternative t-values.*

- Ans:
  - green line is null t-values
  - red lines are 95% CI
  - black line is alternative t-values

```
t_alt <- boot_t_stats[1,]
plot(density(t_alt), main = "bootstrap density plot", lwd=2, ylim=c(0,1), xlim=c(-6,6))
abline(v=mean(t_alt), lty="dotted", col="blue", lwd = 2)
t_null <- boot_t_stats[2,]
lines(density(t_null), lty="dashed",col='green', lwd = 2)

ci_95 <- quantile(t_null, probs=c(0.025, 0.975))
abline(v=ci_95, lty="dotted",col='red', lwd = 2)
```




*c-ii. Based on these bootstrapped results, should we reject the null hypothesis?*

- Answer: Yes, we should reject $H_{null}$.

*Question 2 We also wish to test whether the variance of ILEC response times is different than the variance of CLEC response times.*

*a. What is the null and alternative hypotheses in this case?*

- Answer:
  - $H_{null}$: $\sigma_{ILEC} <= \sigma_{CLEC}$
  - $H_{alt}$: $\sigma_{CLEC} > \sigma_{ILEC}$

*b. Let's try traditional statistical methods first:*

*b-i. What is the F-statistic of the ratio of variances?*

```
var.test(CLEC$Time, ILEC$Time, alternative="greater")
```

```
##
##  F test to compare two variances
##
## data:  CLEC$Time and ILEC$Time
## 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
```

*b-ii. What is the cut-off value of F, such that we want to reject the 5% most extreme F-values?*

```
f_value = var(CLEC$Time)/var(ILEC$Time)
f_value
```

```
## [1] 1.762717
```

*b-iii. Use the qf() function in R to determine the cutoff. Can we reject the null hypothesis?*

```
qf(p=0.95, df1=length(CLEC$Time)-1, df2=length(ILEC$Time)-1)
```

```
## [1] 1.548476
```

- Answer: Yes, we should reject $H_{null}$.

*c. Let's try bootstrapping this time:*

*c-i. Create bootstrapped values of the F-statistic, for both null and alternative hypotheses.*

var(CLEC) > var(ILEC), so CLEC is larger and ILEC is smaller.

```
set.seed(4321)
sd_providers_test <- function(larger_sd_sample, smaller_sd_sample) {
  resample_larger_sd <- sample(larger_sd_sample, length(larger_sd_sample), replace=TRUE)
  resample_smaller_sd <- sample(smaller_sd_sample, length(smaller_sd_sample), replace=TRUE)
  f_alt <- var(resample_larger_sd) / var(resample_smaller_sd)
  f_null <- var(resample_larger_sd) / var(larger_sd_sample)
  c(f_alt, f_null)
}

f_stats <- replicate(10000, sd_providers_test(CLEC$Time, ILEC$Time))
f_alts <- f_stats[1,]
f_nulls <- f_stats[2,]
```

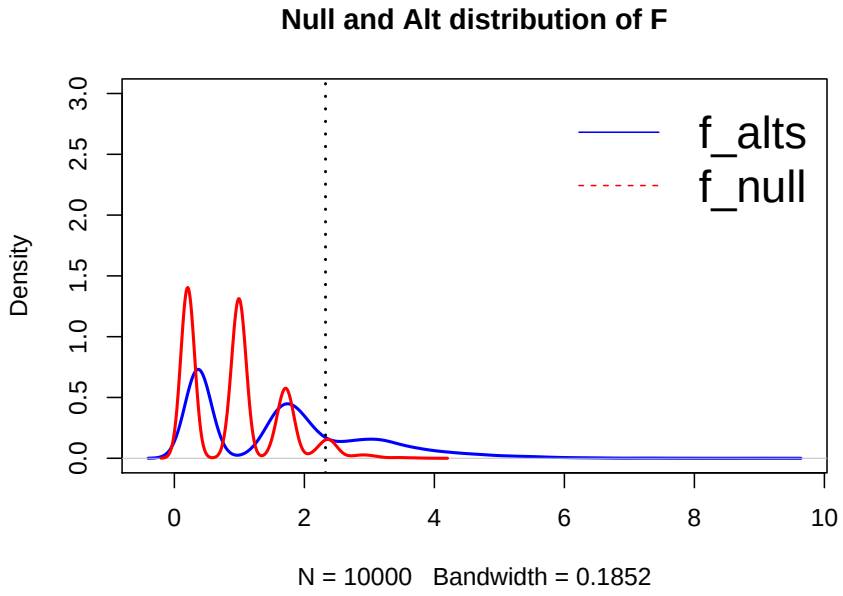
*c-ii. What is the 95% cutoff value according to the bootstrapped null values of F?*

```
quantile(f_nulls, probs=0.95)
```

```
##      95%
## 2.325051
```

*c-iii. Plot a visualization of the null and alternative distributions of the bootstrapped F-statistic, with vertical lines at the cutoff value of F nulls.*

```
plot(density(f_alts), lwd = 2, col = "blue", ylim = c(0,3), main = "Null and Alt distribution of F")
lines(density(f_nulls), lwd = 2, col = "red")
abline(v=quantile(f_nulls, probs=0.95), lwd = 2, lty="dotted")
legend("topright", legend=c("f_alts", "f_null"),col=c("blue", "red"), lty=1:2, cex=1.8, box.lty=0)
```




*c-iv. What do the bootstrap results suggest about the null hypothesis?*

- Answer: We should reject $H_{null}$.

*Question 3 Let's try to see when we should use the non-parametric bootstrap and when we might be better off with traditional statistical approaches.*

*a.*

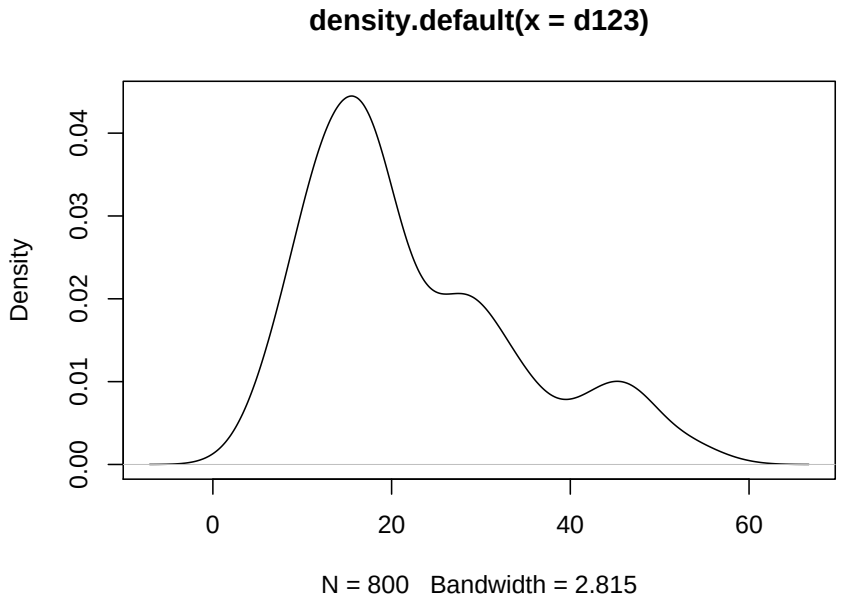
```
norm_qq_plot <- function(values) {
  probs1000 <- seq(0, 1, 0.001)
  q_vals <- quantile(values, probs=probs1000)
  ?qnorm
  q_norm <- qnorm(probs1000, mean = mean(values), sd = sd(values))
   plot(q_norm, q_vals, xlab="normal quantiles", ylab="values quantiles")
   # a: y intercept, b: slope
   abline( a = 0, b = 1 , col="red", lwd=2)
```

```
}
```

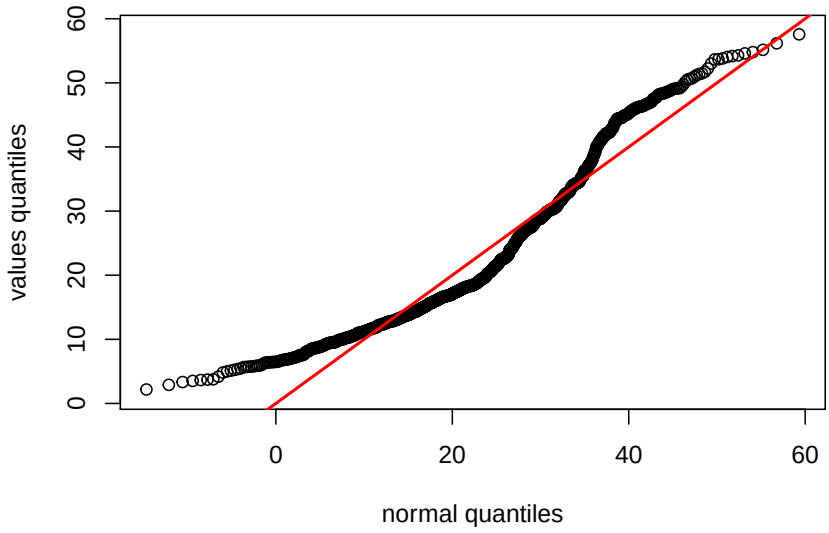
*b. Confirm that your function works by running it against the values of our d123 distribution from week 3 and checking that it looks like the plot on the right:*

- Ans: The points in the Q-Q plot form a relatively straight line since the quantiles of the dataset not match what the quantiles of the dataset would theoretically be.=, so the dataset was not normally distributed.

```
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))
```



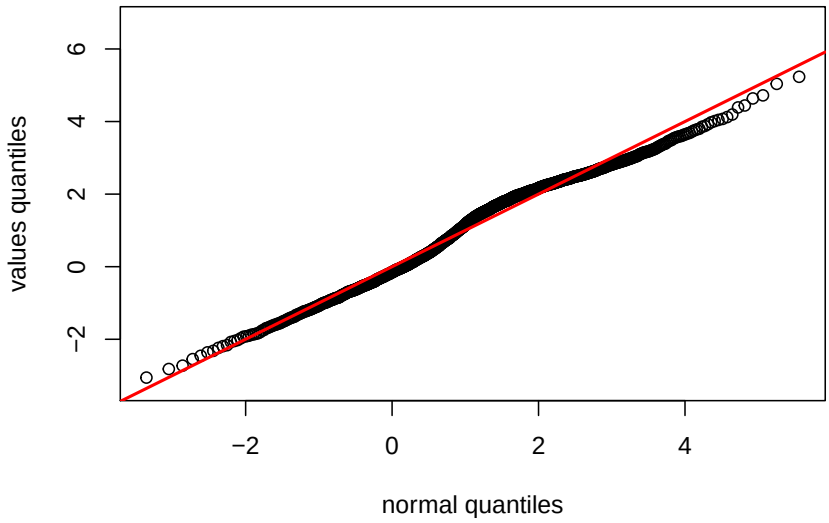

```
norm_qq_plot(d123)
```




*c. Based on 1-c*

- Answer: The bootstrapped distribution of null t-values in **question 1c was normally distributed**. Because the points in the Q-Q plot form a relatively straight line since the quantiles of the dataset nearly match what the quantiles of the dataset would theoretically be if the dataset was normally distributed.

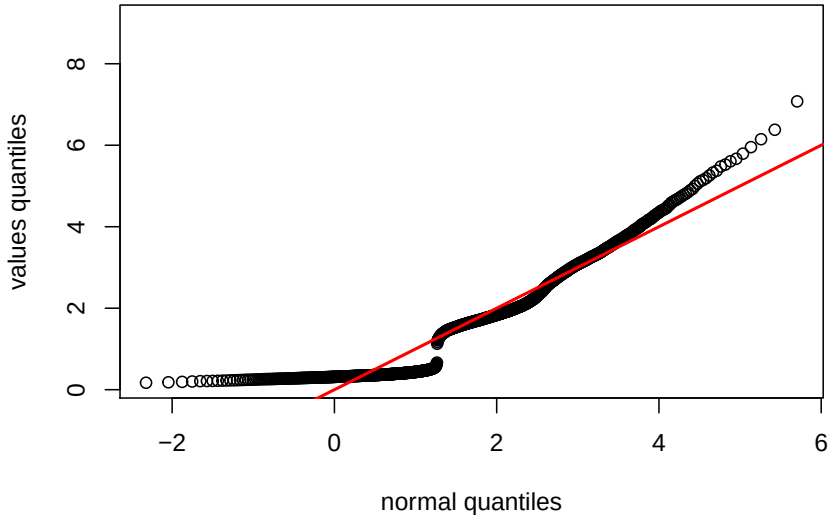
```
norm_qq_plot(boot_t_stats)
```



*d. Hypothesis tests of variances (f-tests) assume the two samples we are comparing come from normally distributed populations. Use your normal Q-Q plot function to check if the two samples we compared in question 2 could have been normally distributed. What's your conclusion?*

- Answer: The two samples we compared in question 2 **are not normally distributed**. Because the points in the Q-Q plot are not close to the straight red line.

```
norm_qq_plot(f_alts)
```




```
norm_qq_plot(f_nulls)
```