BACS HW5

109062710

April 10, 2021

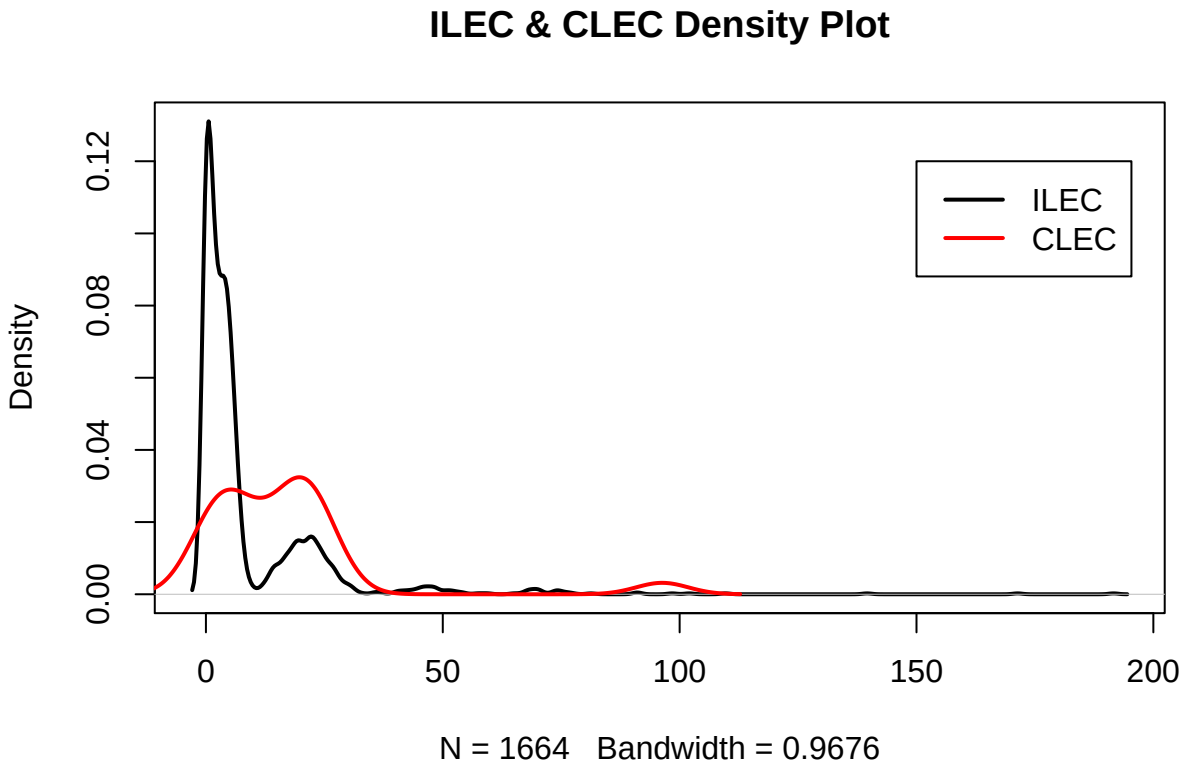
Question 1

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

```
ilec <- as.data.frame(verizon[verizon$Group == "ILEC", ])$Time
ilec_mean <- mean(ilec)

clec <- as.data.frame(verizon[verizon$Group == "CLEC", ])$Time
clec_mean <- mean(clec)

plot(density(ilec), lwd = 2, main = "ILEC & CLEC Density Plot")
lines(density(clec), lwd = 2, col = "red")
legend(
  150,
  0.12,
  c("ILEC", "CLEC"),
  lwd = c(2,2),
  lty = c("solid", "solid"),
  col = c("black", "red")
)
```



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():

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

The null hypothesis is the mean response time for ILEC and CLEC customers are the same. The alternative hypothesis is the mean response time for ILEC and CLEC customers are not the same.

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

```
t.test(ilec, clec, alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  ilec and clec
## t = -1.9834, df = 22.346, p-value = 0.05975
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -16.5568985   0.3618588
## sample estimates:
## mean of x mean of y
##  8.411611 16.509130
```

Since the p-value is greater than equal '0.05', then we say that we accept the null hypothesis.

c. Let's try this using bootstrapping: Estimate bootstrapped null and alternative values of t by using the same t.test() function to compare: bootstrapped samples of ILEC against bootstrapped samples of CLEC (alt t-values); and bootstrapped samples of ILEC against the original ILEC sample (null t-values).

i. Plot a distribution of the bootstrapped null t-values and alternative t-values, adding vertical lines to show the 5% rejection zone of the null distribution (use the same one-vs-two tail logic as 1b).

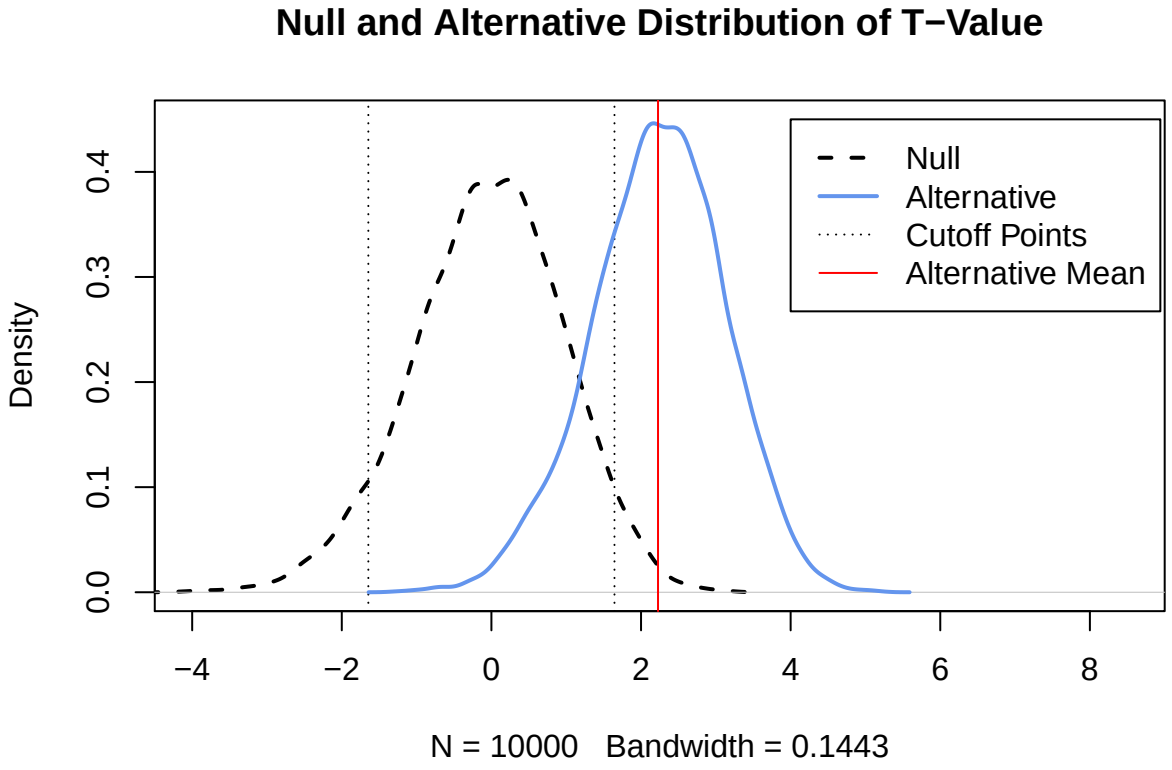
```
t_null_alt <- function(sample0, hyp_mean) {
  resample <- sample(sample0, length(sample0), replace = TRUE)
  resample_se <- sd(resample) / sqrt(length(resample))
  t_null <- (mean(resample) - mean(sample0)) / resample_se
  t_alt <- (mean(resample) - hyp_mean) / resample_se
  c(t_null, t_alt)
}

set.seed(38)
boot <- replicate(10000, t_null_alt(ilec, hypothesized_mean))

# plotting null distribution
t_null <- boot[1, ]
cutoff_value <- qt(c(0.05, 0.95), length(t_null) - 1)
plot(
  density(t_null),
  ylim = c(0, 0.45),
  xlim = c(-4,8.5),
  lwd = 2,
  lty = "dashed",
  main = "Null and Alternative Distribution of T-Value"
)
abline(v = cutoff_value, lty = "dotted")

#plotting alternative distribution
t_alt <- boot[2, ]
lines(density(t_alt), lty = "solid", lwd = 2, col = "cornflowerblue")
abline(v = mean(t_alt), col="red")

legend(
  4,
  0.45,
  c("Null", "Alternative", "Cutoff Points", "Alternative Mean"),
  lty = c("dashed", "solid", "dotted", "solid"),
  lwd = c(2,2,1,1),
  col = c("black", "cornflowerblue", "black", "red"))
```



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

Since the mean of the alternative distribution sits to the far right (outside the cutoff points), we can reject the null hypothesis.

Question 2

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

```
print(var(clec))
```

```
## [1] 380.3895
```

```
print(var(ilec))
```

```
## [1] 215.7973
```

Since the variance of CLEC, 380.3894628, is higher than the variance of ILEC, 215.7972572, the null hypotheses is that the variance of CLEC will be the same like ILEC's if the major outliers in CLEC are removed. While the alternative hypotheses is that the variance of CLEC will still be higher than ILEC's even if the major outliers in CLEC are removed.

b. Let's try traditional statistical methods first:

i. What is the F-statistics of the ratio of variances?

```
var(clec) / var(ilec)
```

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

ii. What is the cut-off value of F, such that we want to reject the 5% most extreme F-values? Use the qf() function in R to determine the cutoff.

The cut-off value of F-statistics is 1.5484762.

iii. Can we reject the null hypothesis?

c. Let's try bootstrapping this time:

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

```
f_null_alt <- function(sample_small_var, sample_large_var) {
  resample_small_var <-
    sample(sample_small_var, length(sample_small_var), replace = TRUE)

  resample_large_var <-
    sample(sample_large_var, length(sample_large_var), replace = TRUE)

  f_null <- var(resample_large_var) / var(sample_large_var)
  f_alt <- var(resample_large_var) / var(resample_small_var)

  c(f_null, f_alt)
}

bootstrap_f_null_alt <- replicate(10000, f_null_alt(ilec, clec))

f_null <- bootstrap_f_null_alt[1,]
f_alt <- bootstrap_f_null_alt[2, ]
f_stats <- var(clec) / var(ilec)
```

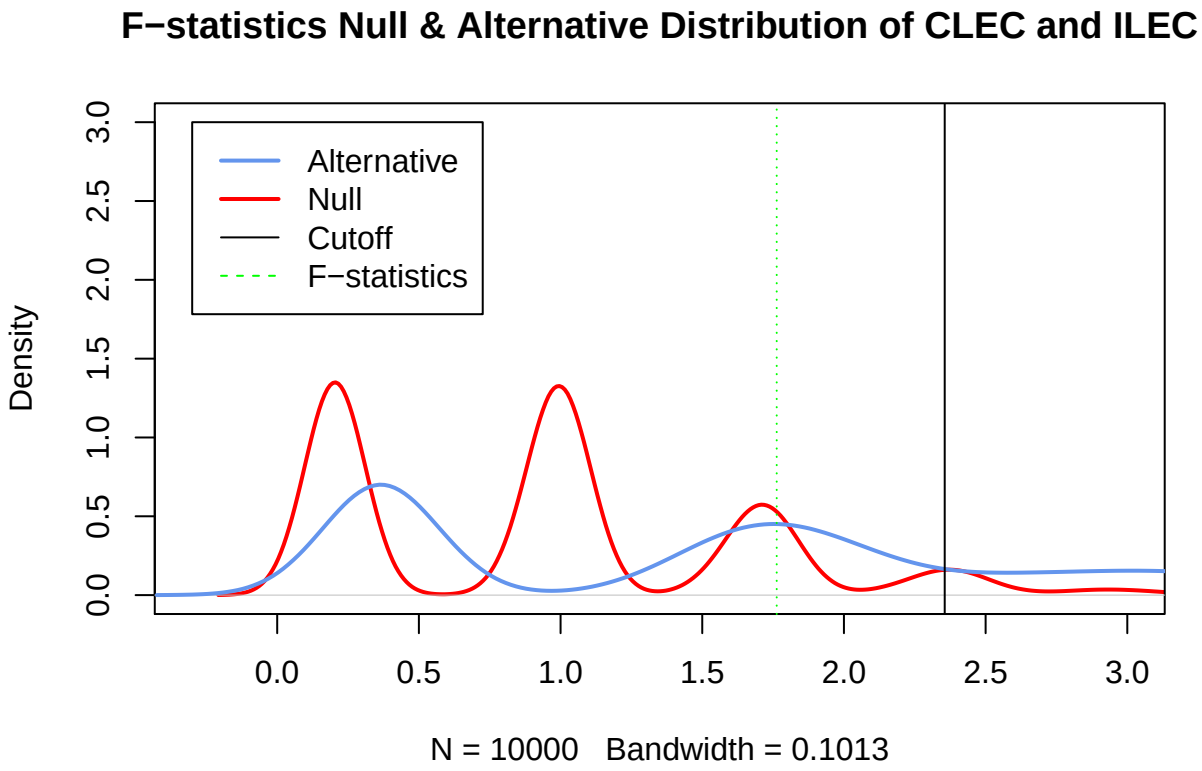
ii. What is the 95% cutoff value according to the bootstrapped null values of F?

```
f_cutoff <- quantile(f_null, probs = 0.95)
f_cutoff
```

```
##     95%
## 2.35566
```

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_null),
  col = "red",
  lwd = 2,
  xlim = c(-0.3, 3),
  ylim = c(0,3),
  main = "F-statistics Null & Alternative Distribution of CLEC and ILEC"
)
lines(density(f_alt), col = "cornflowerblue", lwd = 2)
abline(v = f_cutoff, lty = "solid", col = "black")
abline(v = f_stats, lty = "dotted", col = "green")
legend(
  -0.3,
  3,
  c("Alternative", "Null", "Cutoff", "F-statistics"),
  lty = c("solid", "solid", "solid", "dashed"),
  lwd = c(2,2,1,1),
  col = c("cornflowerblue", "red", "black", "green"))
```



iv. Can we reject the null hypothesis?

Since the F-statistics value doesn't cross over the cutoff point, thus we can accept the null hypothesis.

Question 3

Part a

Create a `qq_plot` function.

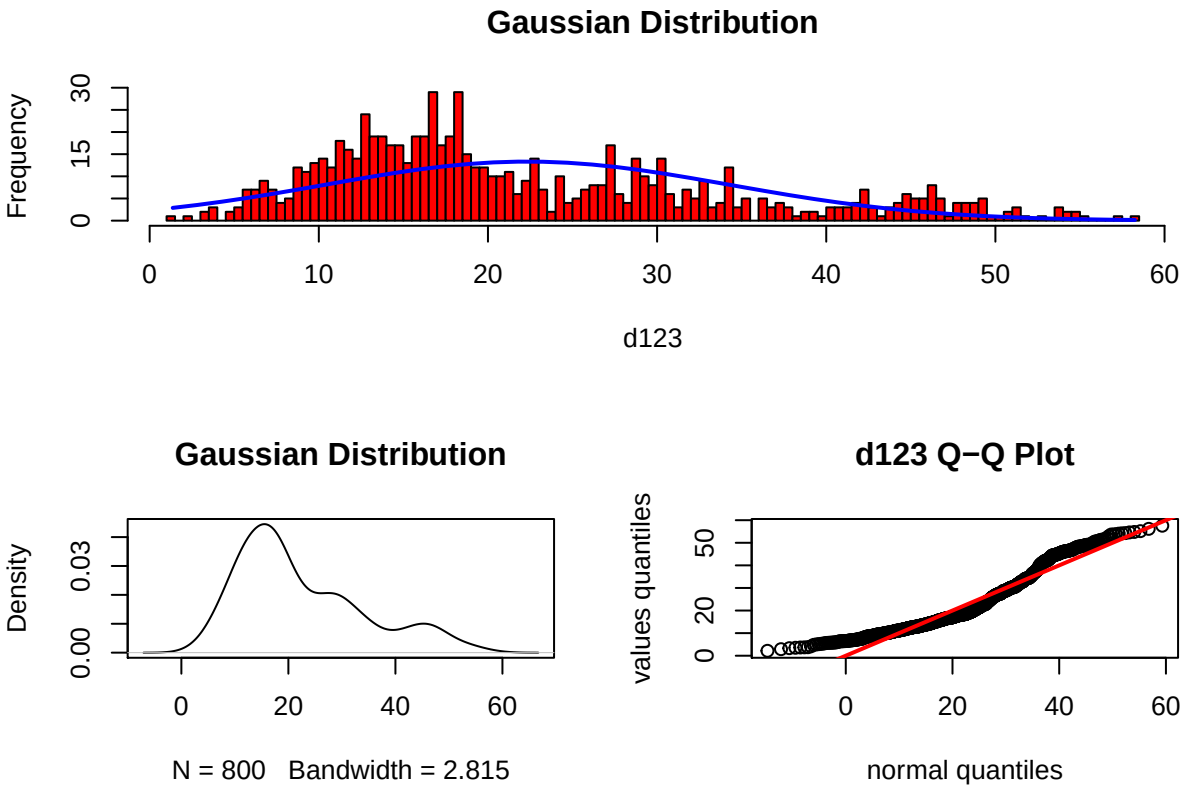
```
norm_qq_plot <-function(values, main) {
  probs1000 <- seq(0, 1, 0.001)
  q_vals <- quantile(values, probs=probs1000)
  q_norm <- qnorm(probs1000,mean=mean(values), sd = sd(values))
  plot(q_norm, q_vals, xlab = "normal quantiles", ylab = "values quantiles", main = main)
  abline(a = 0, b = 1, col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

Part b

Confirm that if the function works.

```
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)

layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
x <- hist(d123, breaks=100, col = "red", main = "Gaussian Distribution")
xfit <- seq(min(d123), max(d123), length = 40)
yfit <- dnorm(xfit, mean = mean(d123), sd = sd(d123))
yfit <- yfit * diff(x$mids[1:5]) * length(d123)
lines(xfit, yfit, col = "blue", lwd = 2)
plot(density(d123), main = "Gaussian Distribution")
norm_qq_plot(values = d123, main = "d123 Q-Q Plot")
```

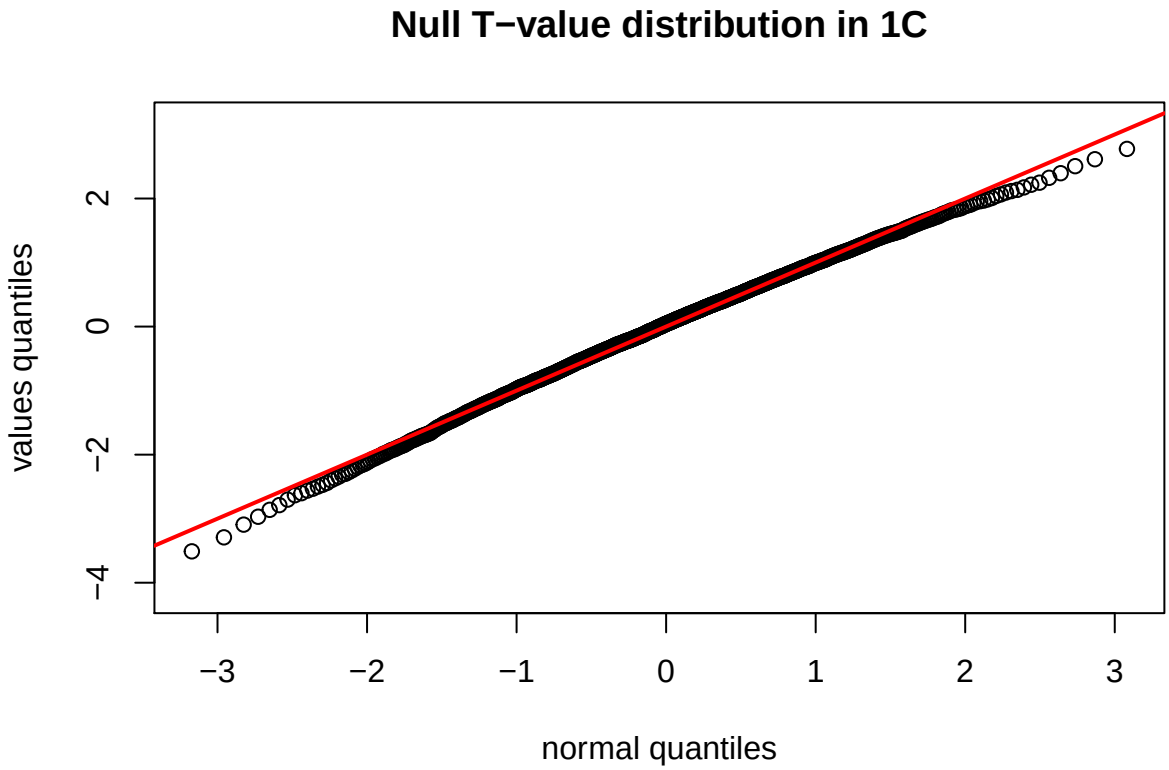


The Q-Q plot suggests that our Gaussian distribution of this data is more skewed to the right. Moreover, some of group in the data set seems to have higher frequencies, thus making them more higher than the density graph in the first graph. In other words, it's not normally distributed.

Part c

Create a Q-Q Plot for Null T-value in 1C.

```
norm_qq_plot(t_null, main = "Null T-value distribution in 1C")
```

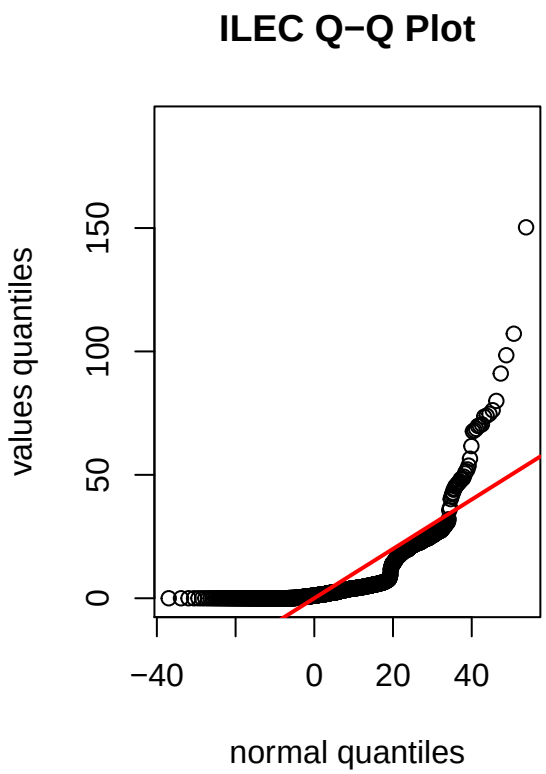


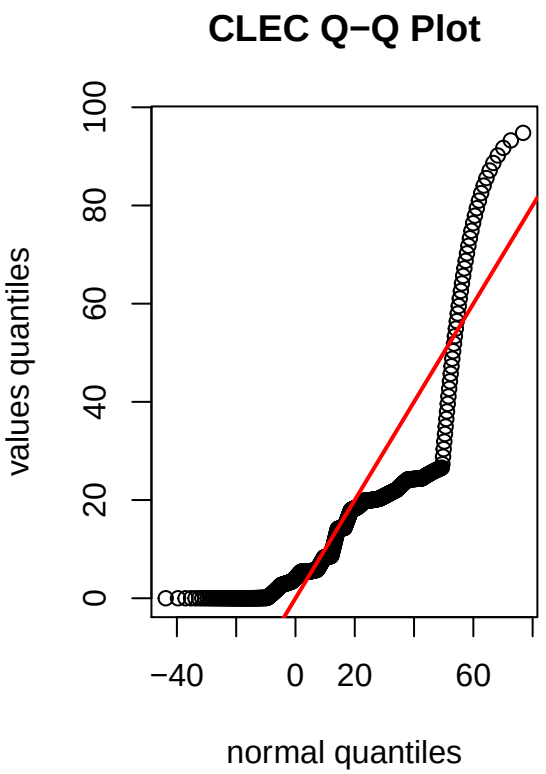
The bootstrapped distribution of null t-values is normally distributed since most of the data points sit on the red line.

Part 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?

```
layout(matrix(c(1,2), 1, 2, byrow = TRUE))
norm_qq_plot(ilec, main = "ILEC Q-Q Plot")
norm_qq_plot(clec, main = "CLEC Q-Q Plot")
```





From the two graphs above, we can conclude that both data sets are not normally distributed because most of the data points in both data sets don't sit around the red lines.