BACS - HW6

109006241

Question 1) The Verizon dataset this week is provided as a "wide" data frame. Let's practice reshaping it to a "long" data frame. You may use either shape (wide or long) for your analyses in later questions.

a. Pick a reshaping package (we discussed two in class) – research them online and tell us why you picked it over others (provide any helpful links that supported your decision).

I would pick the reshape2 package, because it has the wider purpose of data reshaping, and it is more adapted for reshaping usages.

Reference:

https://www.r-bloggers.com/2016/06/how-to-reshape-data-in-r-tidyr-vs-reshape2/

b. Show the code to reshape the verizon_wide.csv sample

```
data = melt(data, na.rm=T)
```

```
## No id variables; using all as measure variables
```

```
groups = split(x=data$value, f=data$variable)
```

c. Show us the "head" and "tail" of the data to show that the reshaping worked

```
head(data)
```

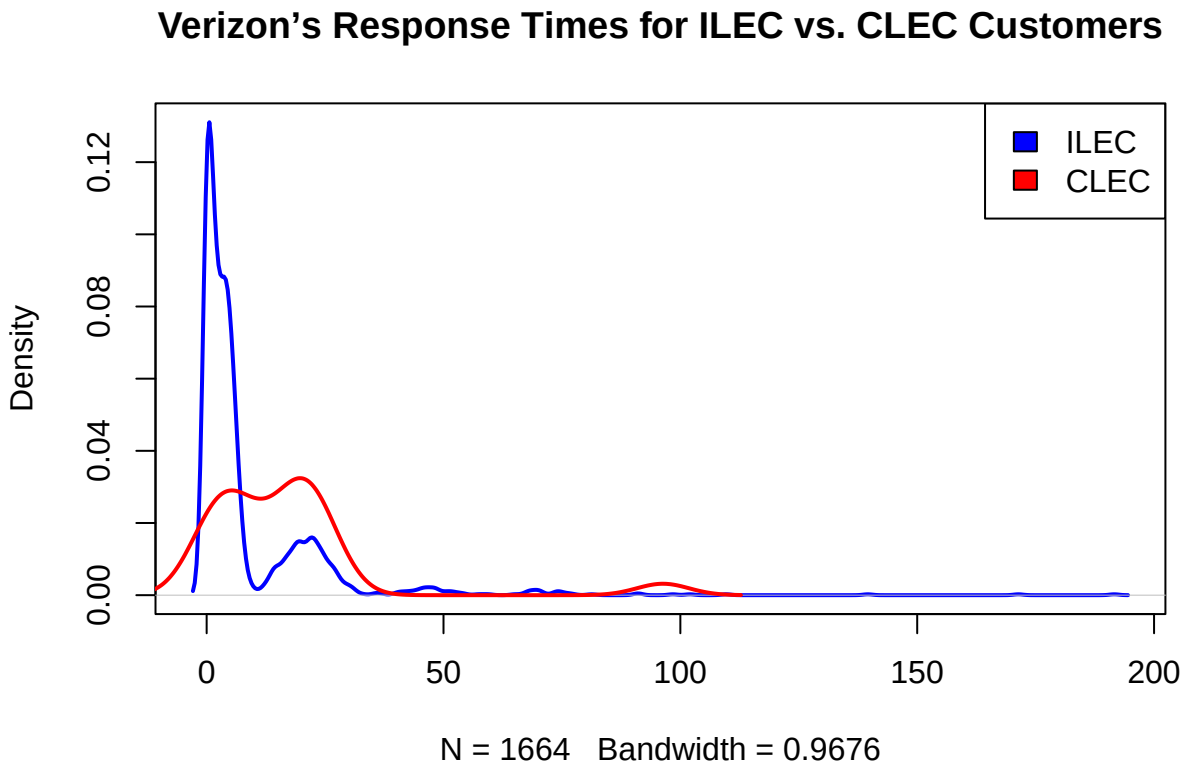
```
##   variable value
## 1     ILEC 17.50
## 2     ILEC  2.40
## 3     ILEC  0.00
## 4     ILEC  0.65
## 5     ILEC 22.23
## 6     ILEC  1.20
```

```
tail(data)
```

```
##      variable value
## 1682     CLEC 24.20
## 1683     CLEC 22.13
## 1684     CLEC 18.57
## 1685     CLEC 20.00
## 1686     CLEC 14.13
## 1687     CLEC  5.80
```

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

```
plot(density(groups$ILEC), col="blue", lwd=2,
     main="Verizon's Response Times for ILEC vs. CLEC Customers")
lines(density(groups$CLEC), col="red", lwd=2)
legend("topright", legend=c("ILEC", "CLEC"), fill=c("blue", "red"), )
```



Question 2) Let's test if the mean of response times for CLEC customers is greater than for ILEC customers

a. State the appropriate null and alternative hypotheses (one-tailed)

H0: The mean of response times for CLEC customers is less than or equal to for ILEC customers

H1: The mean of response times for CLEC customers is greater than for ILEC customers

b. Use the appropriate form of the t.test() function to test the difference between the mean of ILEC versus CLEC response times at 1% significance. For each of the following tests, show us the results and tell us whether you would reject the null hypothesis.

i. Conduct the test assuming variances of the two populations are equal

```
t.test(groups$CLEC, groups$ILEC, alternative="greater", conf.level=0.99, var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  groups$CLEC and groups$ILEC
## t = 2.6125, df = 1685, p-value = 0.004534
## alternative hypothesis: true difference in means is greater than 0
## 99 percent confidence interval:
##  0.8801387       Inf
## sample estimates:
## mean of x mean of y
## 16.509130  8.411611
```

ii. Conduct the test assuming variances of the two populations are not equal

```
t.test(groups$CLEC, groups$ILEC, alternative="greater", conf.level=0.99, var.equal=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  groups$CLEC and groups$ILEC
## t = 1.9834, df = 22.346, p-value = 0.02987
## alternative hypothesis: true difference in means is greater than 0
## 99 percent confidence interval:
##  -2.130858       Inf
## sample estimates:
## mean of x mean of y
## 16.509130  8.411611
```

> **When we are assuming that the variances of the two populations are equal, we get a p-value of 0.004534, so we reject the null hypothesis, but if we are assuming that the variances of the two populations are not equal, we get a p-value of 0.02987, meaning that we cannot reject the null hypothesis. We can see that they are bringing us conflicting results.**

c. Use a permutation test to compare the means of ILEC vs. CLEC response times

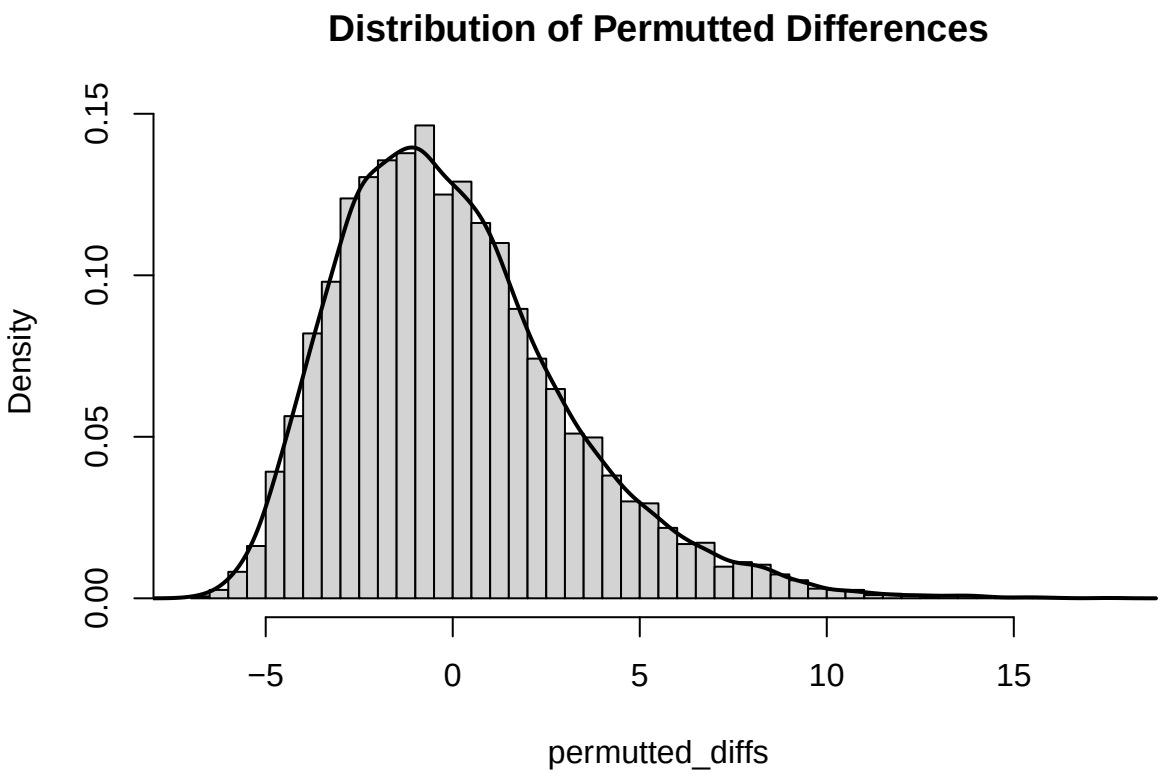
i. Visualize the distribution of permuted differences, and indicate the observed difference as well.

```
obs_diff = mean(groups$CLEC) - mean(groups$ILEC)

permute_diff <- function(values, groups) {
  permuted = sample(values, replace=F)
  grouped = split(permuted, groups)
  permuted_diff = mean(grouped$CLEC) - mean(grouped$ILEC)
}

nperms = 10000
permutted_diffs = replicate(nperms, permute_diff(data$value, data$variable))

hist(permutted_diffs, breaks="fd", probability=T, main="Distribution of Permutted Differences")
lines(density(permutted_diffs), lwd=2)
```




```
cat(" Observed Difference =", obs_diff, "\n",
     "Permutted Difference =", mean(permutted_diffs))
```

```
##  Observed Difference = 8.09752
##  Permutted Difference = 0.0334853
```

ii. What are the one-tailed and two-tailed p-values of the permutation test?

```
p_1tailed = sum(permutted_diffs > obs_diff) / nperms
p_2tailed = sum(abs(permutted_diffs) > obs_diff) / nperms
cat(" One-tailed p-value =", p_1tailed, "\n",
     "Two-tailed p-value =", p_2tailed)
```

```
##  One-tailed p-value = 0.0184
##  Two-tailed p-value = 0.0184
```

iii. Would you reject the null hypothesis at 1% significance in a one-tailed test?

No. Because p-value $>$ 0.01.

Question 3) Let's use the Wilcoxon test to see if the response times for CLEC are different than ILEC.

a. Compute the W statistic comparing the values. You may use either the permutation approach (try the functional form) or the rank sum approach.

```r
gt_eq <- function(a, b) {
  ifelse(a > b, 1, 0) + ifelse(a == b, 0.5, 0)
}
W = sum(outer(groups$CLEC, groups$ILEC, FUN=gt_eq))
cat("W-statistic =", W)
```

```
## W-statistic = 26820
```

b. Compute the one-tailed p-value for W.

```r
n1 <- length(groups$CLEC)
n2 <- length(groups$ILEC)
wilcox_p_1tail = 1 - pwilcox(W, n1, n2)
wilcox_p_1tail
```

```
## [1] 0.0003688341
```

c. Run the Wilcoxon Test again using the wilcox.test() function in R – make sure you get the same W as part [a]. Show the results.

```r
wilcox.test(groups$CLEC, groups$ILEC, alternative="greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  groups$CLEC and groups$ILEC
## W = 26820, p-value = 0.0004565
## alternative hypothesis: true location shift is greater than 0
```

d. At 1% significance, and one-tailed, would you reject the null hypothesis that the values of CLEC and ILEC are similar?

Because p-value is less than the significance level 0.01, we can reject the null hypothesis which says that the values of CLEC and ILEC are similar

---

Question 4) One of the assumptions of some classical statistical tests is that our population data should be roughly normal. Let's explore one way of visualizing whether a sample of data is normally distributed.

a. Make a function called norm_qq_plot() to create a function to see how a distribution of values compares to a perfectly normal distribution.

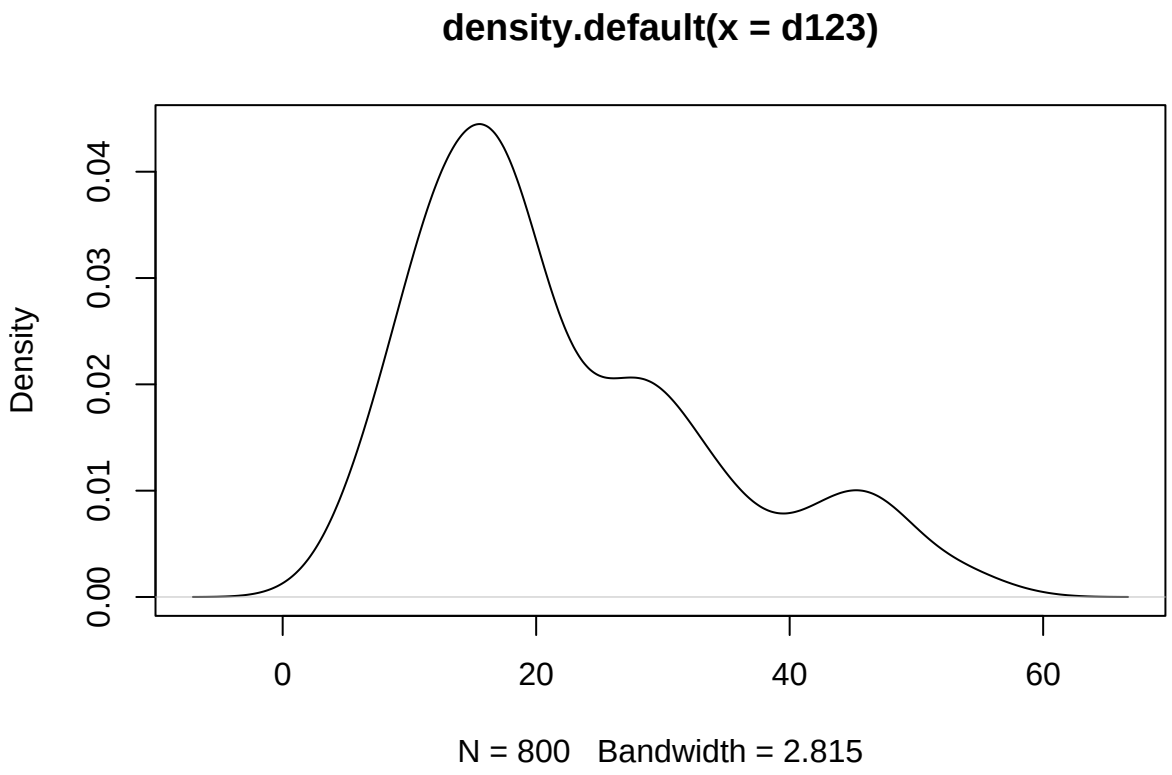
```r
norm_qq_plot <- function(values) {
  probs1000 <- seq(0, 1, 0.001)
  q_vals <- quantile(values, probs=probs1000)
  q_norm <- qnorm(probs1000, mean(values), sd(values))
  plot(q_norm, q_vals, xlab="normal quantiles", ylab="values quantiles")
  abline(a=0, b=1, col="red", lwd=2)
}
```

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:

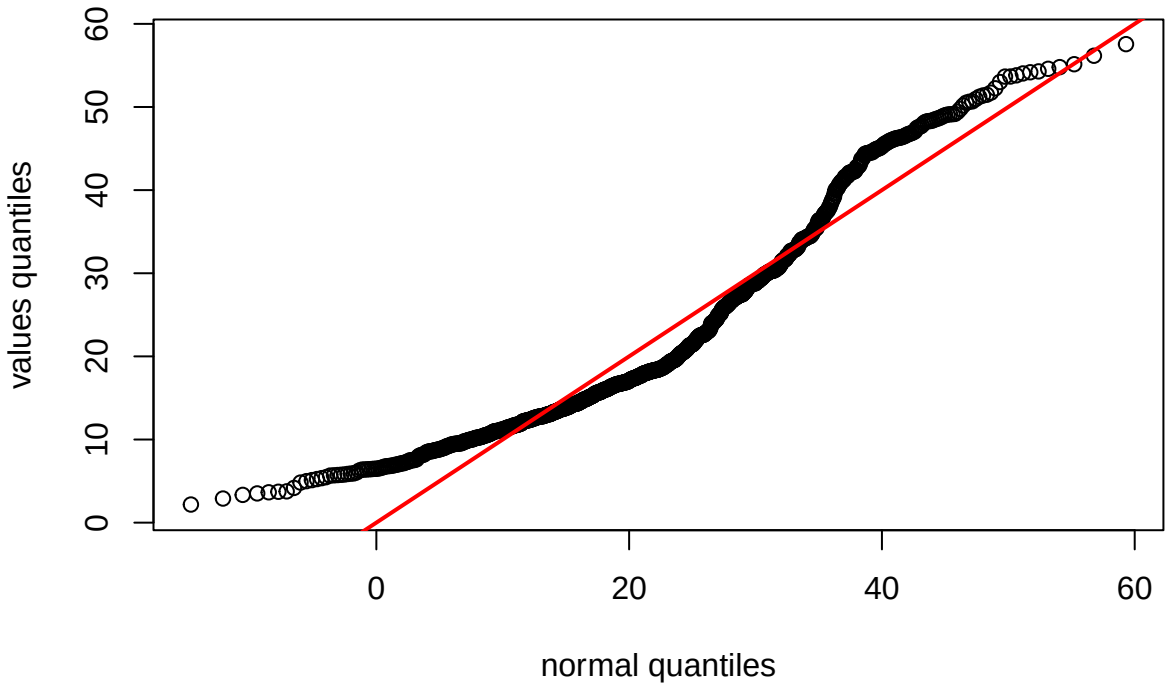
   Interpret the plot you produced (see this article on how to interpret normal Q-Q plots) and tell us if it suggests whether d123 is normally distributed or not.

```
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)
```

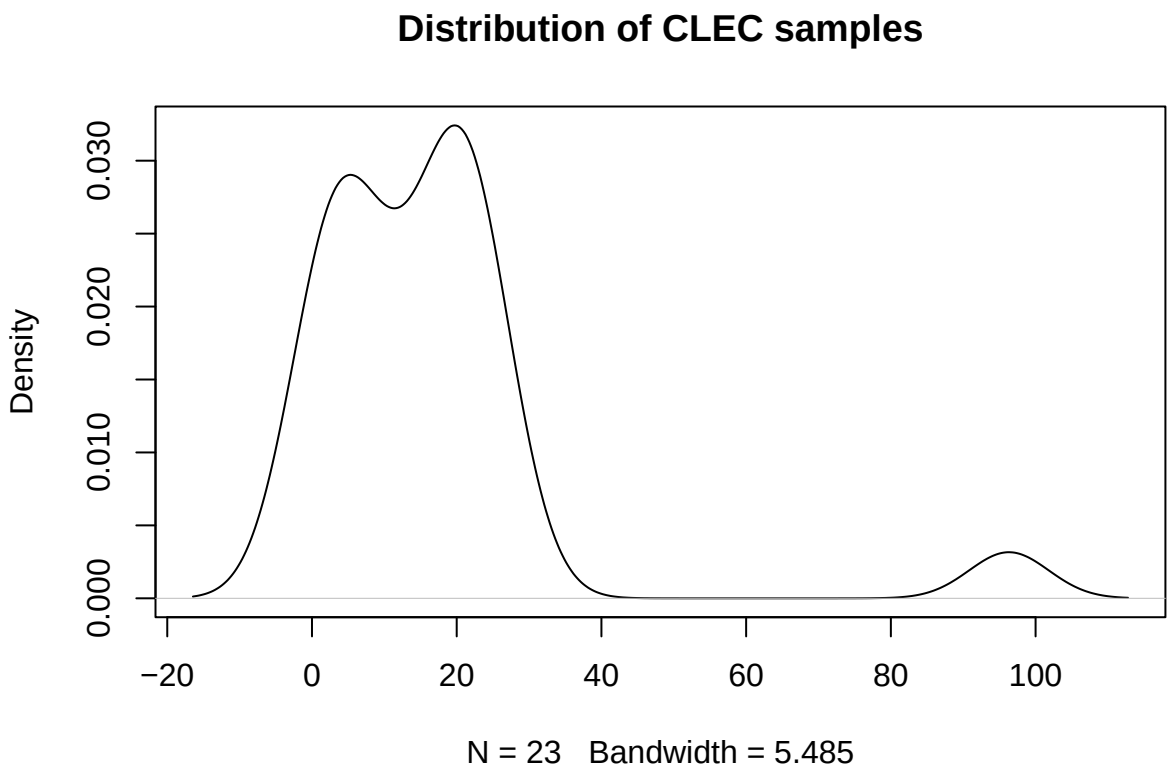


In the Q-Q plot, the straight line indicates the theoretical quantile values of a normal distribution. In other words, it shows where the points would fall if the dataset were normally distributed.

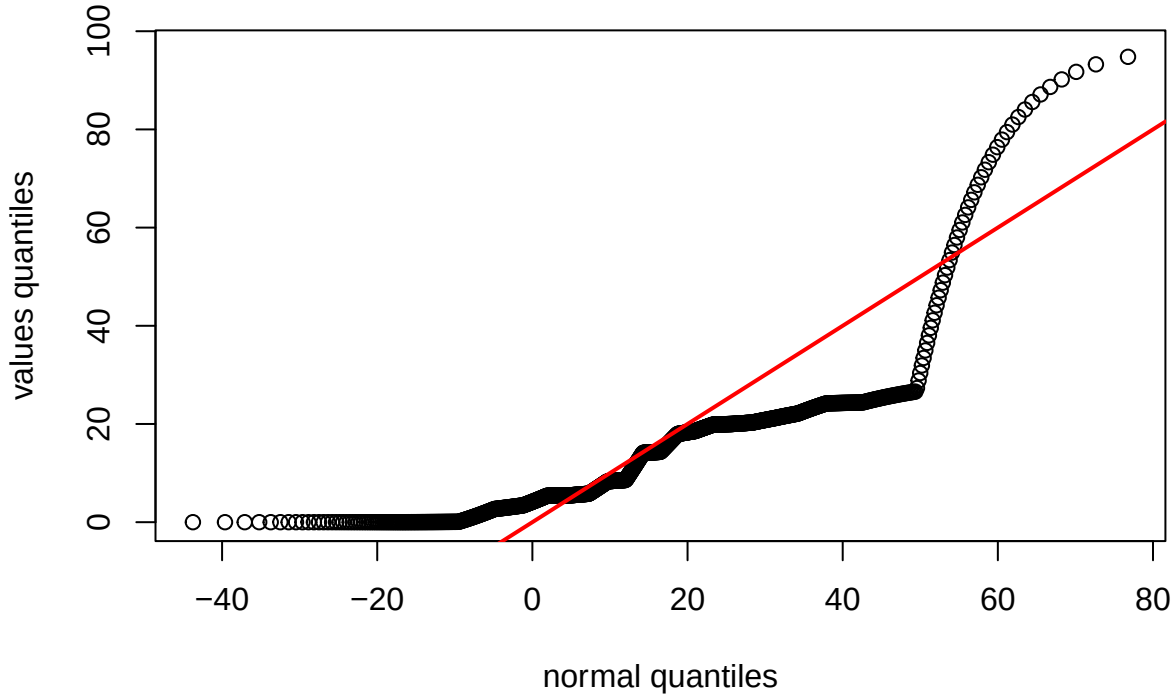
We can see that some of the points in the Q-Q plot departs from the straight line, specifically in the left side and right side of the distribution. So, we can conclude that the distribution is not normally distributed, since some of the actual quantiles do not follow the theoretical quantiles of the dataset if it was normally distributed.

c. Use your normal Q-Q plot function to check if the values from each of the CLEC and ILEC samples we compared in question 2 could be normally distributed. What's your conclusion?

```
plot(density(groups$CLEC), main="Distribution of CLEC samples")
```




```
norm_qq_plot(groups$CLEC)
```



As we can see from the density plot, we can see that the distribution of CLEC samples is positively skewed / skewed to the right. This means that most of the data lies on the left side, with a long "tail" extending to the right side.

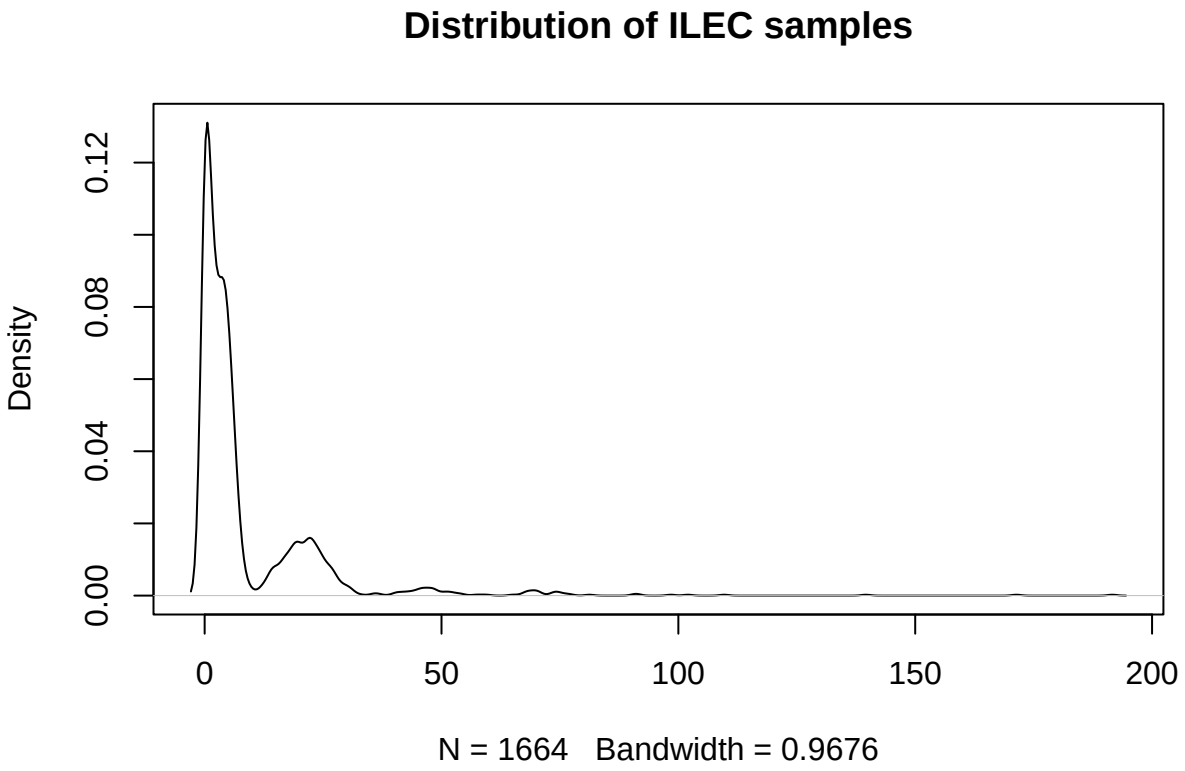
From the Q-Q plot, we can see that the points start to depart downwards and then upwards from the straight line right after, as we follow the quantiles from left to right.

The point's trend downward in the middle part of the distribution shows that the actual quantiles are lower than the normal quantiles, meaning that there is a lower concentration of data in the middle parts of the distribution.

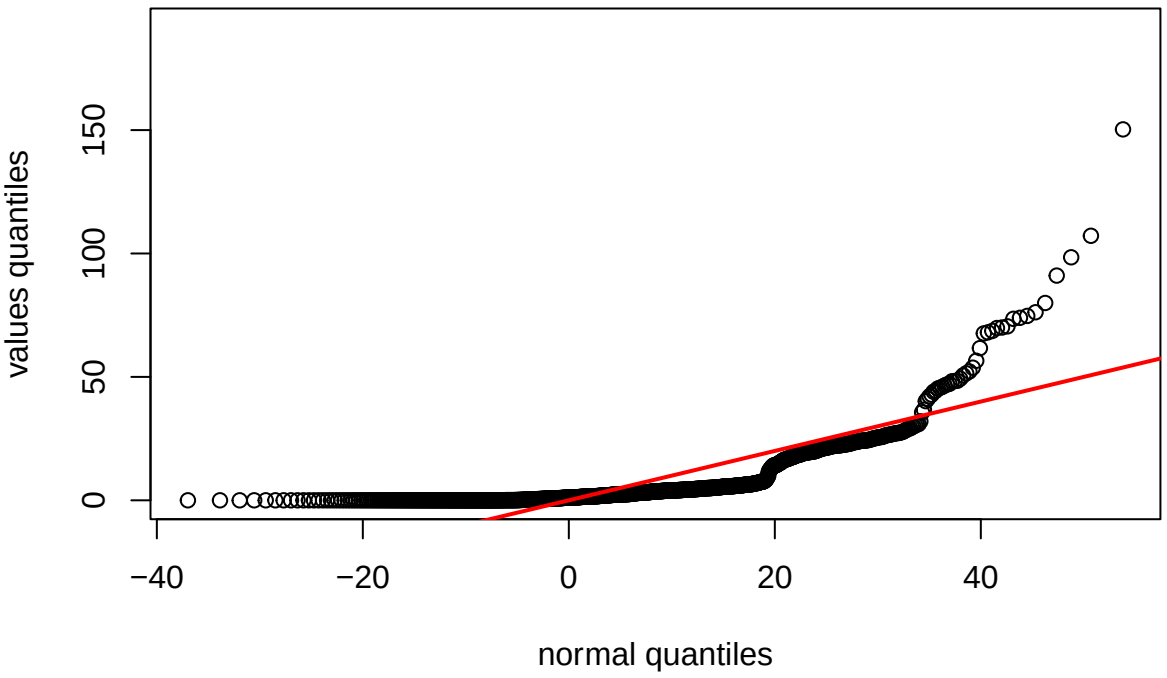
The point's trend upward in the right part of the distribution shows that the actual quantiles are greater than the normal quantiles, meaning that there is a greater concentration of data in the right parts of the distribution.

So, because there are many points that are far away from the straight line (the theoretical quantiles), we can conclude that the CLEC samples does not follow a normal distribution.

```
plot(density(groups$ILEC), main="Distribution of ILEC samples")
```




```
norm_qq_plot(groups$ILEC)
```



As we can see from the density plot, we can see that the distribution of CLEC samples is positively skewed / skewed to the right. This means that most of the data lies on the left side, with a long "tail" extending to the right side.

From the Q-Q plot, we can see that the points start to depart upwards from the straight line, as we follow the quantiles from left to right.

The point's trend upward in the right part of the distribution shows that the actual quantiles are greater than the normal quantiles, meaning that there is a greater concentration of data in the right parts of the distribution.

So, because there are many points that are far away from the straight line (the theoretical quantiles), we can conclude that the ILEC samples does not follow a normal distribution.