HW_Week6_108020033

Che-Wei, Chang

2023-03-22 helped by 108020024

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).

```
# import library
library(tidyr)
```

I will use the "tidyr" package for reshaping the data frame because it provides a simple and easy-to-use set of functions for reshaping data.

Link to documentation: https://tidyr.tidyverse.org/

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

```
# load the data
df <- read.csv("verizon_wide.csv", header = TRUE)

# reshape data
verizon_long <- gather(df, na.rm = TRUE, key = "customer_type", value = "response_time")
```

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

```
# Show the head of the data
head(verizon_long)
```

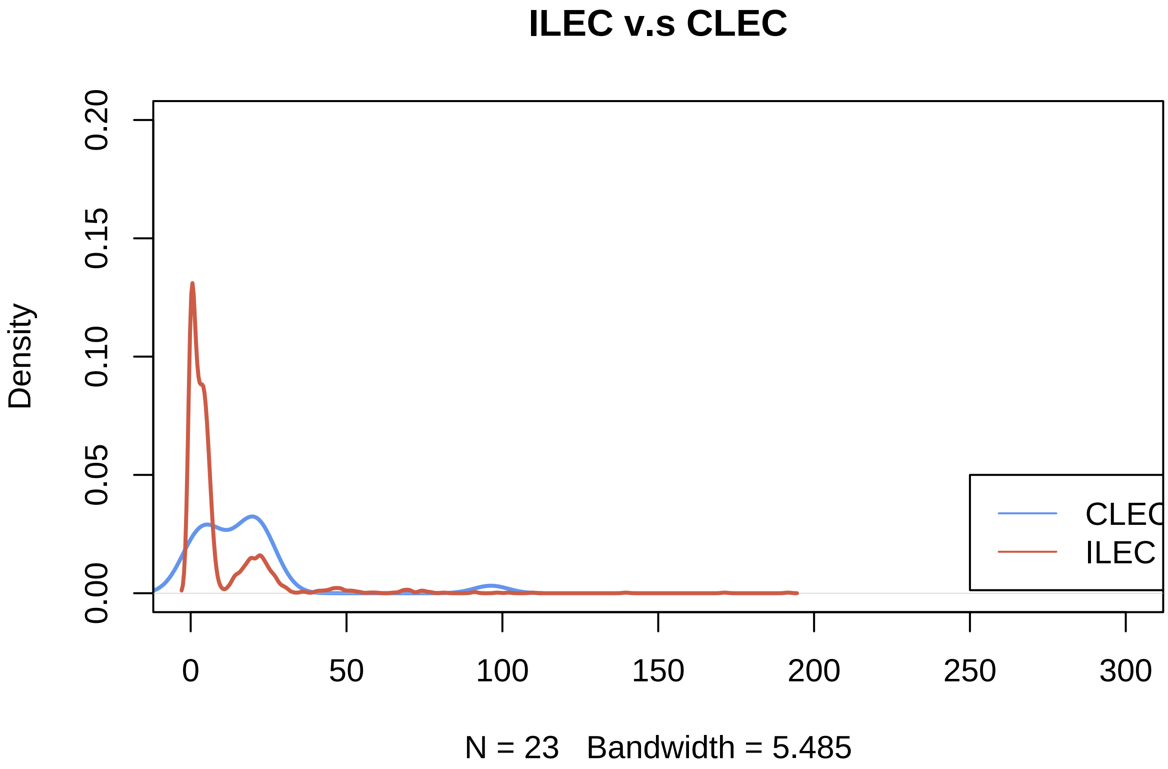
```
##   customer_type response_time
## 1          ILEC         17.50
## 2          ILEC          2.40
## 3          ILEC          0.00
## 4          ILEC          0.65
## 5          ILEC         22.23
## 6          ILEC          1.20
```

```
# Show the tail of the data
tail(verizon_long)
```

```
##      customer_type response_time
## 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

```
# Split the data into ILEC and CLEC
Time <- split(x = verizon_long$response_time, f = verizon_long$customer_type)

# Show the density plot
plot(density(Time$CLEC), col = "cornflowerblue", lwd = 2, xlim = c(0, 300), ylim = c(0, 0.2), main = "I
lines(density(Time$ILEC), col = "coral3", lwd = 2)

# Add the legend
legend(250, 0.05, lty = 1, c("CLEC", "ILEC"), col = c("cornflowerblue", "coral3"))
```



```
# Show the summary of the data
summary(Time$CLEC)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   5.425  14.330  16.509  20.715  96.320
```

```
summary(Time$ILEC)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.730   3.590   8.412   7.080 191.600
```

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)

null hypothesis: The mean response time for CLEC customers is less than or equal to the mean response time for ILEC customers.

Alternative hypothesis: The mean response time for CLEC customers is greater than the mean response time 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(Time$CLEC, Time$ILEC, alt = "greater", var.equal = TRUE, conf.level = 0.99)
```

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

Since the p-value $= 0.004534 < 0.01$, we can reject the null hypothesis.

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

```
t.test(Time$CLEC, Time$ILEC, alt = "greater", var.equal = FALSE, conf.level = 0.99)
```

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

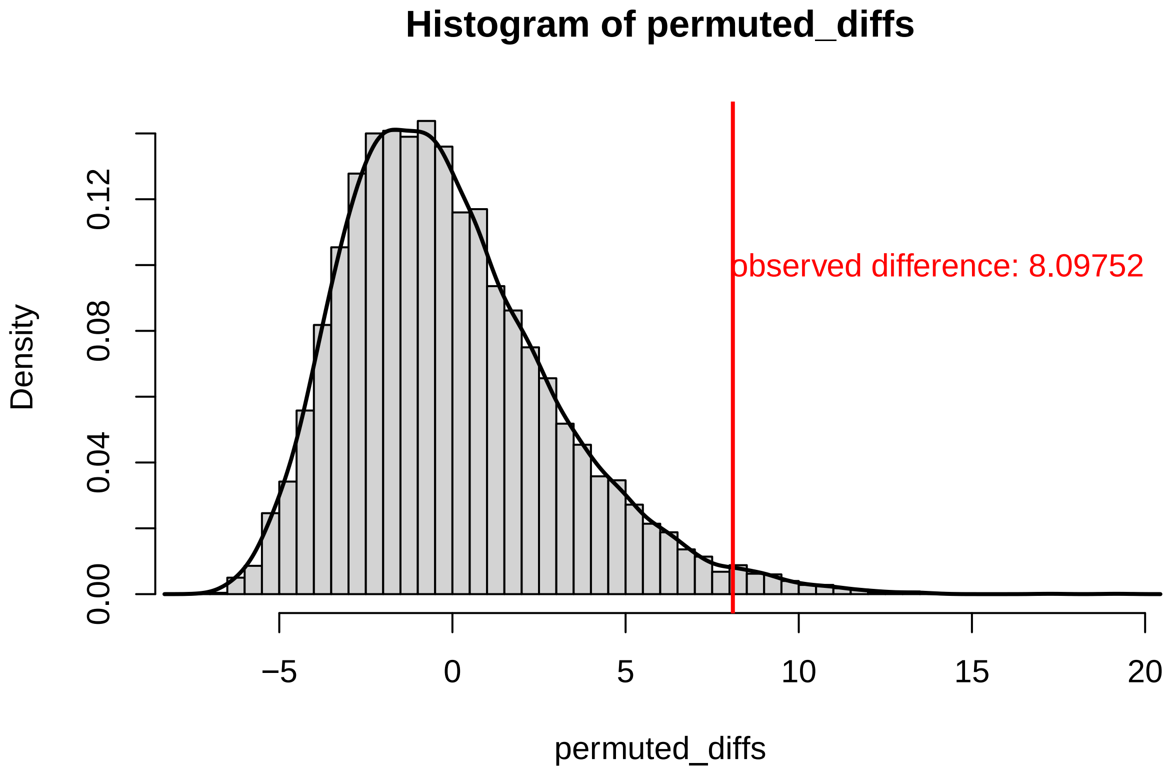
Since p-value $= 0.02987 > 0.01$, we can't reject the null hypothesis.

Therefore, we can't conclude that null hypothesis is false based on these data.

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.

```r
# Calculate the difference of the observations
obs_diff <- mean(Time$CLEC) - mean(Time$ILEC)

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

# Set the number of test
nperms <- 10000

set.seed(42)
# Do the test
permuted_diffs <- replicate(nperms, permute_diff(verizon_long$response_time, verizon_long$customer_type

# Show the result with the plot
hist(permuted_diffs, breaks = "fd", probability = TRUE)
lines(density(permuted_diffs), lwd = 2)

# Indicate the observed difference
abline(v = obs_diff, col = "red", lwd = 2)
text(14, 0.1, labels = "observed difference: 8.09752", col = "red")
```



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

```
# Calculate the p-values for the permutation test
p_1tailed <- sum(permuted_diffs > obs_diff) / nperms
p_2tailed <- sum(abs(permuted_diffs) > obs_diff) / nperms

# Print the one-tailed and two-tailed p-values
cat("one-tailed p-values: ", p_1tailed, "\n")
```

```
## one-tailed p-values:  0.0165
```

```
cat("two-tailed p-values: ", p_2tailed, "\n")
```

```
## two-tailed p-values:  0.0165
```

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

Since the p-value = 0.0165 > 0.01, we won't reject the null hypothesis.

Therefore, we do not have sufficient evidence to conclude that the mean response time for CLEC customers is less than for ILEC customers.

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
# Use rank sum approach
# Rank the time
time_ranks <- rank(verizon_long$response_time)

# Gather and sum the ranks of each group
ranked_groups <- split(time_ranks, verizon_long$customer_type)
U1 <- sum(ranked_groups$CLEC)
# Adjust the rank sum proportionally
n1 <- length(Time$CLEC)
W <- U1 - (n1 * (n1 + 1)) / 2

# Show the result of W statistic
cat("W statistic: ", W, "\n")
```

```
## W statistic:  26820
```

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

```r
n1 <- length(Time$ILEC)
n2 <- length(Time$CLEC)

wilcox_p_1tail <- 1 - pwilcox(W, n1, n2)
cat("one-tailed p-value: ", wilcox_p_1tail, "\n")
```

```
## one-tailed p-value:  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(Time$CLEC, Time$ILEC, alternative = "greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Time$CLEC and Time$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?

Since the p-value $= 0.0004565 < 0.01$, we will reject the null hypothesis that the values of CLEC an 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. Follow the following steps to create a function to see how a distribution of values compares to a perfectly normal distribution. The ellipses (...) in the steps below indicate where you should write your own code.

Make a function called norm_qq_plot() that takes a set of values):

norm_qq_plot <- function(values) { … }

Within the function body, create the five lines of code as follows.

i. Create a sequence of probability numbers from 0 to 1, with ~1000 probabilities in between probs1000 <- seq(0, 1, 0.001)

ii. Calculate ~1000 quantiles of our values (you can use probs=probs1000), and name it q_vals q_vals <- quantile(…)

iii. Calculate ~1000 quantiles of a perfectly normal distribution with the same mean and standard deviation as our values; name this vector of normal quantiles q_norm q_norm <- qnorm(…)

iv. Create a scatterplot comparing the quantiles of a normal distribution versus quantiles of values plot(q_norm, q_vals, xlab="normal quantiles", ylab="values quantiles")

v. Finally, draw a red line with intercept of 0 and slope of 1, comparing these two sets of quantiles abline(… , col="red", lwd=2)

You have now created a function that draws a "normal quantile-quantile plot" or Normal Q-Q plot

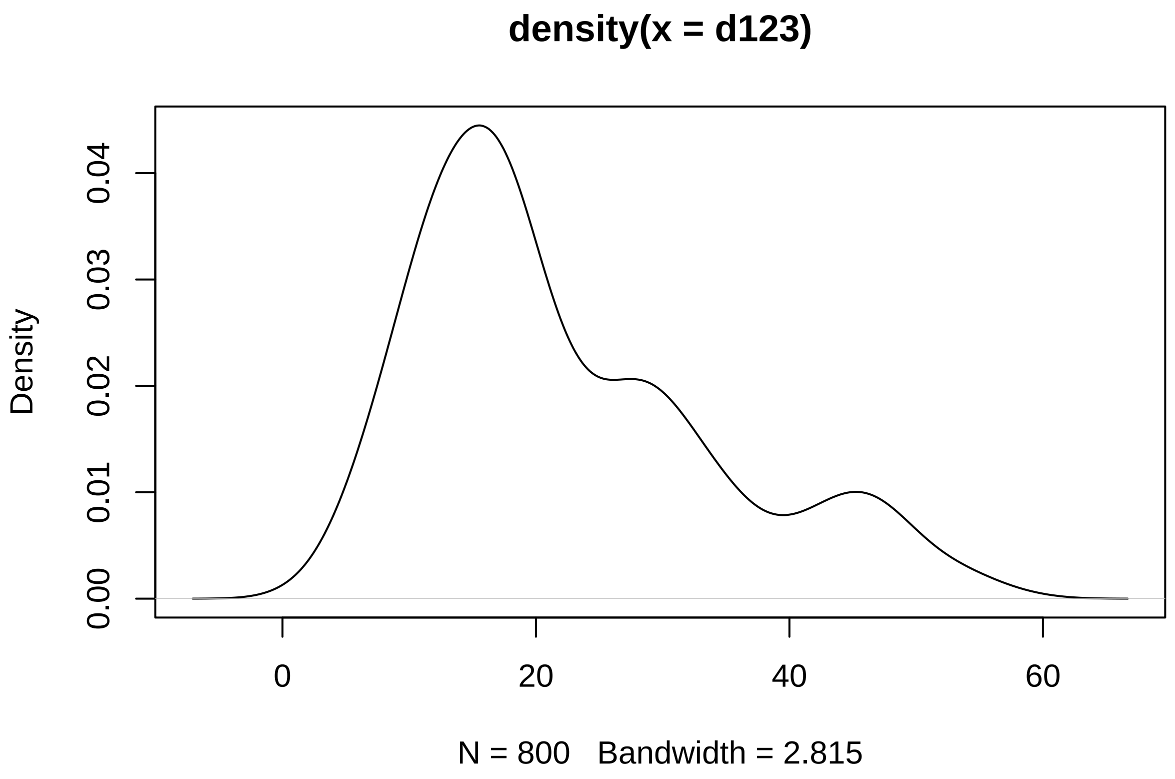
(please show code for the whole function in your HW report)

```r
norm_qq_plot <- function(values) {
  # Create sequence of probabilities
  probs1000 <- seq(0, 1, 0.001)

  # Calculate quantiles of values
  q_vals <- quantile(values, probs = probs1000)

  # Calculate quantiles of normal distribution
  q_norm <- qnorm(probs1000, mean = mean(values), sd = sd(values))

  # Create scatterplot
  plot(q_norm, q_vals, xlab = "normal quantiles", ylab = "values quantiles")

  # Add red line
  abline(0, 1, col = "red", lwd = 2)
}
```

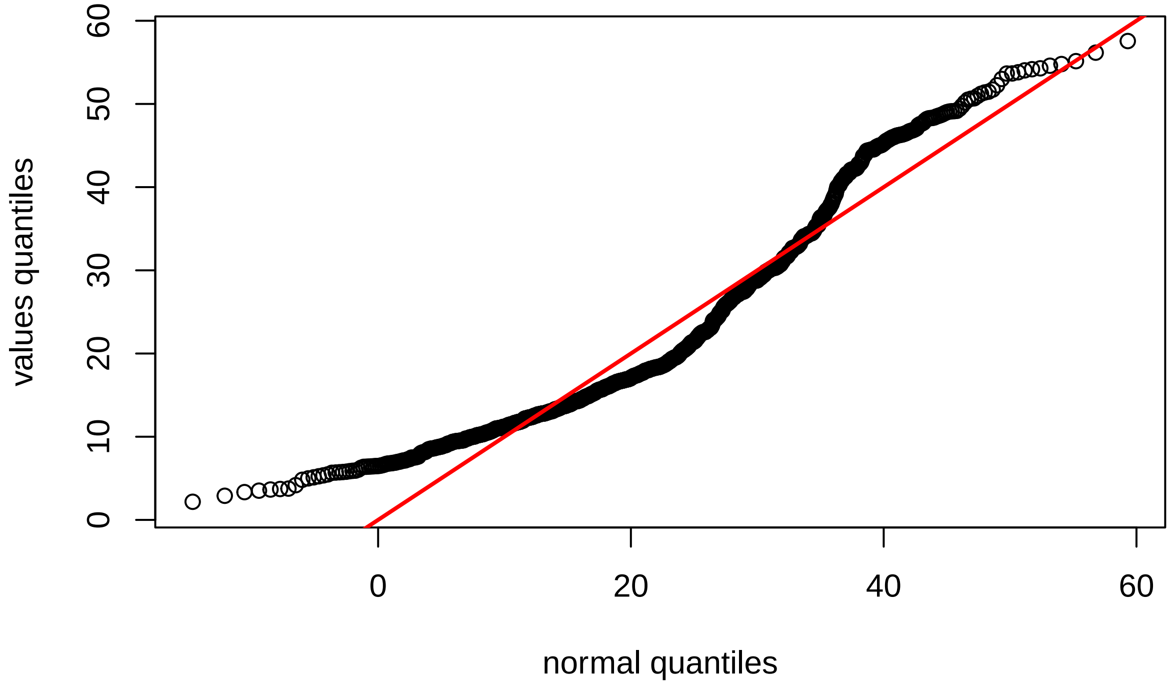
b. Confirm that your function works by running it against the values of our d123 distribution from week 3 and checking that it looks like the plot on the right:

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.

```r
# Set the seed and declare the data
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)

# Show the result by plot function and norm_qq_plot function
plot(density(d123))
```



```
norm_qq_plot(d123)
```

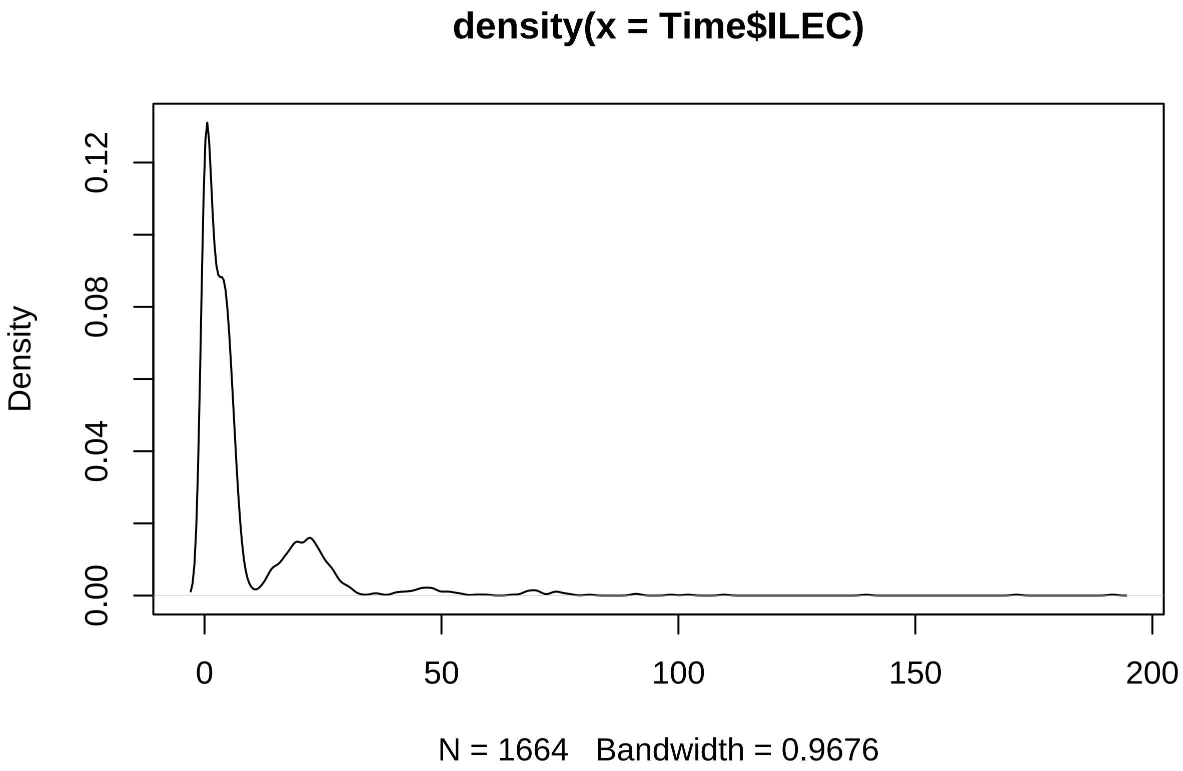


By the plot, we can see that the data doesn't tightly fitted to the line.

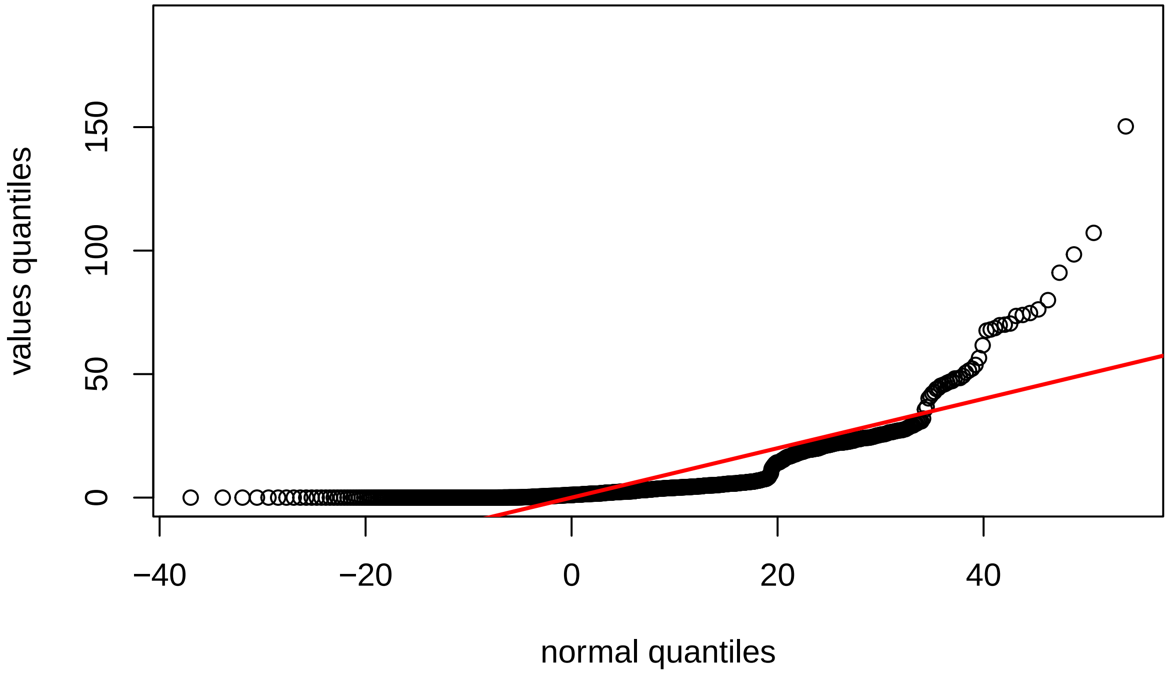
Therefore, d123 is not normally distributed.

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?

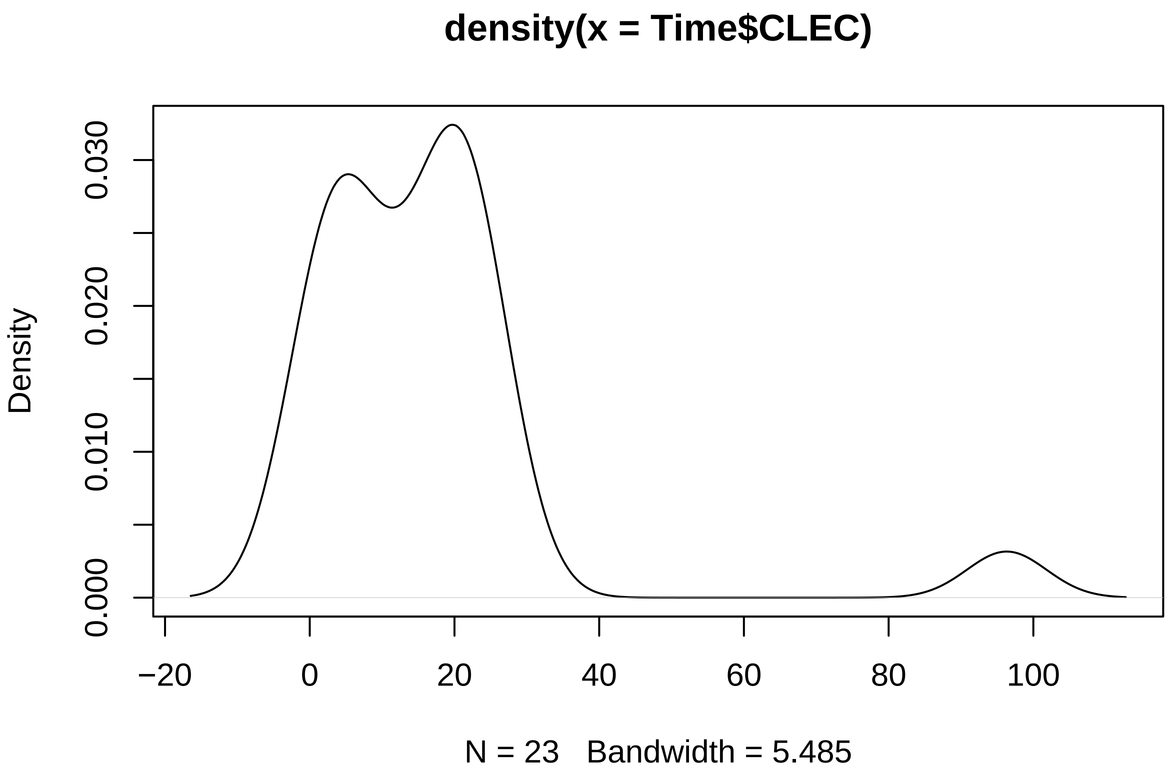
```r
# show the plot of ILEC
plot(density(Time$ILEC))
```



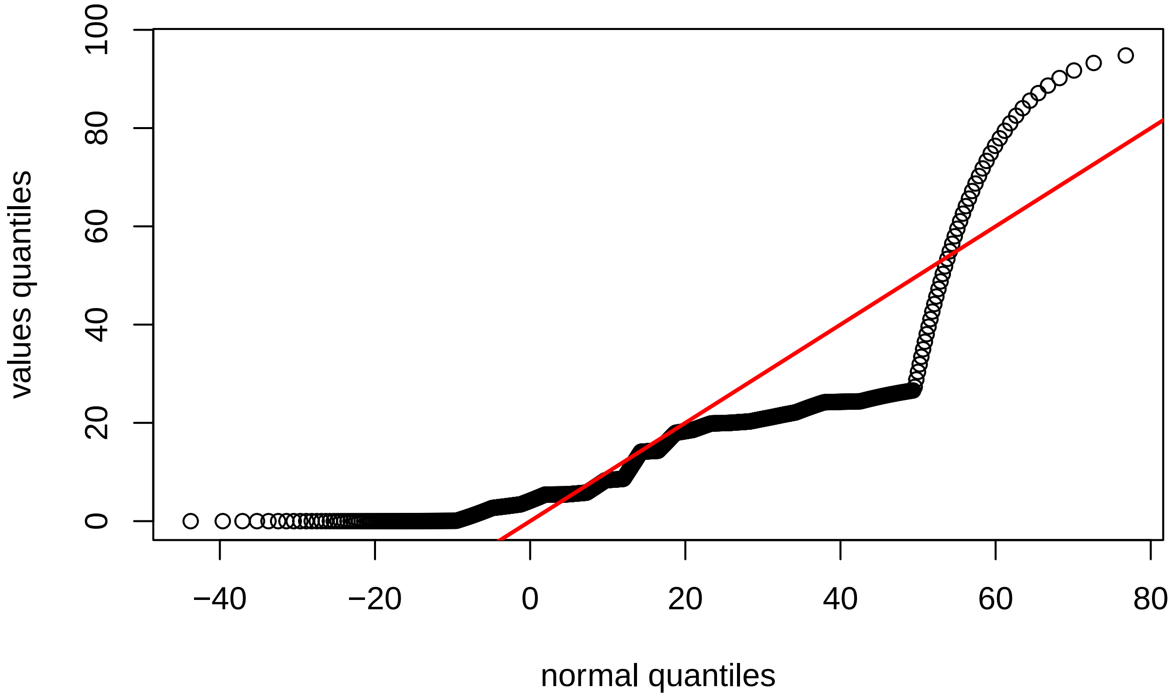
```r
norm_qq_plot(Time$ILEC)
```



```
# Show the plot of CLEC
plot(density(Time$CLEC))
```



```
norm_qq_plot(Time$CLEC)
```



By these plots, we can conclude that ILEC and CLEC samples are not normally distributed.