# HW5

113078506

2025-03-22

- Gen AI Usage: I use Gen AI to refine my grammar and verify my reasoning.
- Students who helped: 113078505 (Helped with checking the results of Q2__c__ii), 113078502 (Help with checking the plot of Q1__c__iii)

## Question 1

### 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 will choose *reshape*2 over *tidyr* because *reshape*2 can be applied not only to data frames but also to matrices and arrays, whereas *tidyr* works exclusively with data frames. Additionally, *reshape*2 supports data aggregation, which *tidyr* does not. Although *tidyr* is more user-friendly for beginners, it lacks the flexibility that *reshape*2 offers. Therefore, I prefer using *reshape*2 for this task.

- Links that supported my decision:

- R mini camp: Reshape2 and Tidyr

- How to reshape data in R: tidyr vs reshape2

### b. Show the code to reshape the verizon__wide.csv sample

```
data <- read.csv("verizon_wide.csv")
#install.packages("reshape2")
library(reshape2)

data_long <- melt(data, na.rm = TRUE,
variable.name = "Carrier",
value.name = "response_time")
```

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

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

```
# Show the reshaped data
head(data_long)
```

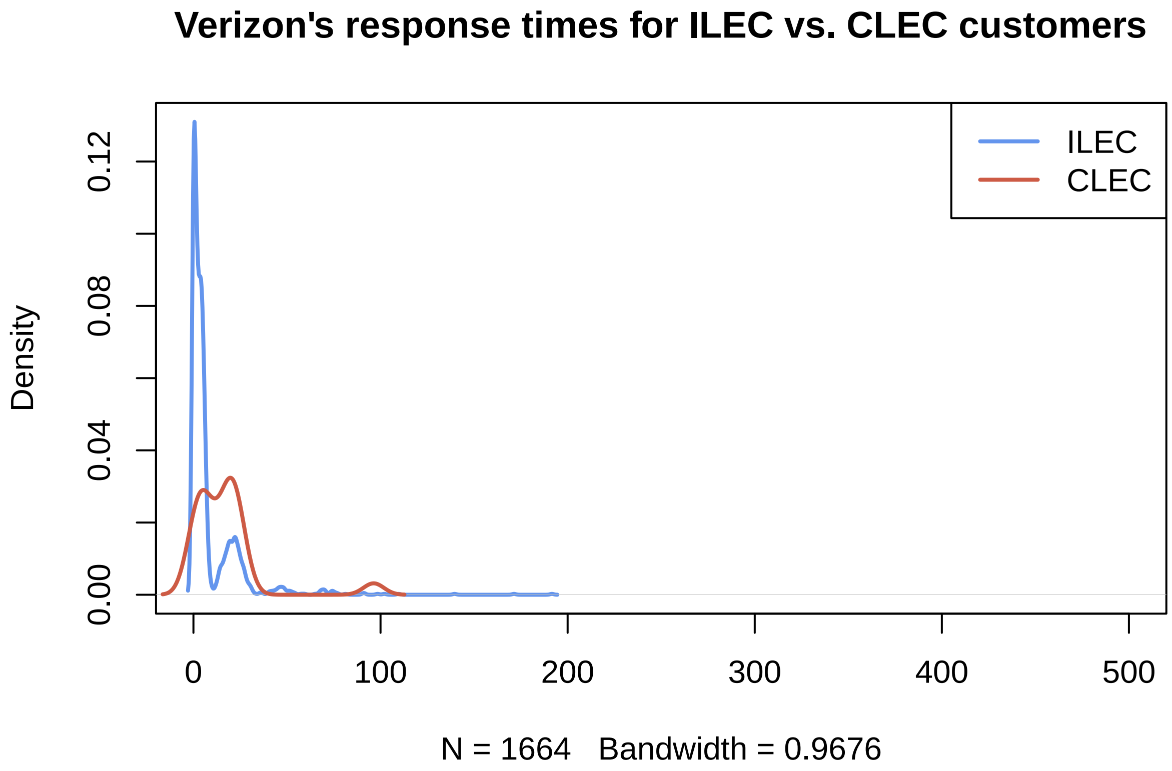
```
##   Carrier response_time
## 1    ILEC         17.50
## 2    ILEC          2.40
## 3    ILEC          0.00
## 4    ILEC          0.65
## 5    ILEC         22.23
## 6    ILEC          1.20
```

```
tail(data_long)
```

```
##      Carrier 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

```
carrier <- split(x = data_long$response_time, f = data_long$Carrier)

plot(density(carrier$ILEC), col="cornflowerblue", lwd=2, xlim=c(0, 500), main = "Verizon’s response time
lines(density(carrier$CLEC), col="coral3", lwd=2)
legend("topright", legend = c("ILEC", "CLEC"), col = c("cornflowerblue", "coral3"), lty = 1, lwd = 2)
```



## Question 2

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

- $H_0 : \mu_{CLEC} \leq \mu_{ILEC}$
- $H_1 : \mu_{CLEC} > \mu_{ILEC}$

### 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.

```r
t.test(carrier$CLEC, carrier$ILEC,
alt="greater", var.equal=TRUE, conf.level = 0.99)
```

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

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

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

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

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

### c. Implement a permutation test, as we saw in class, to compare the means of ILEC vs. CLEC times

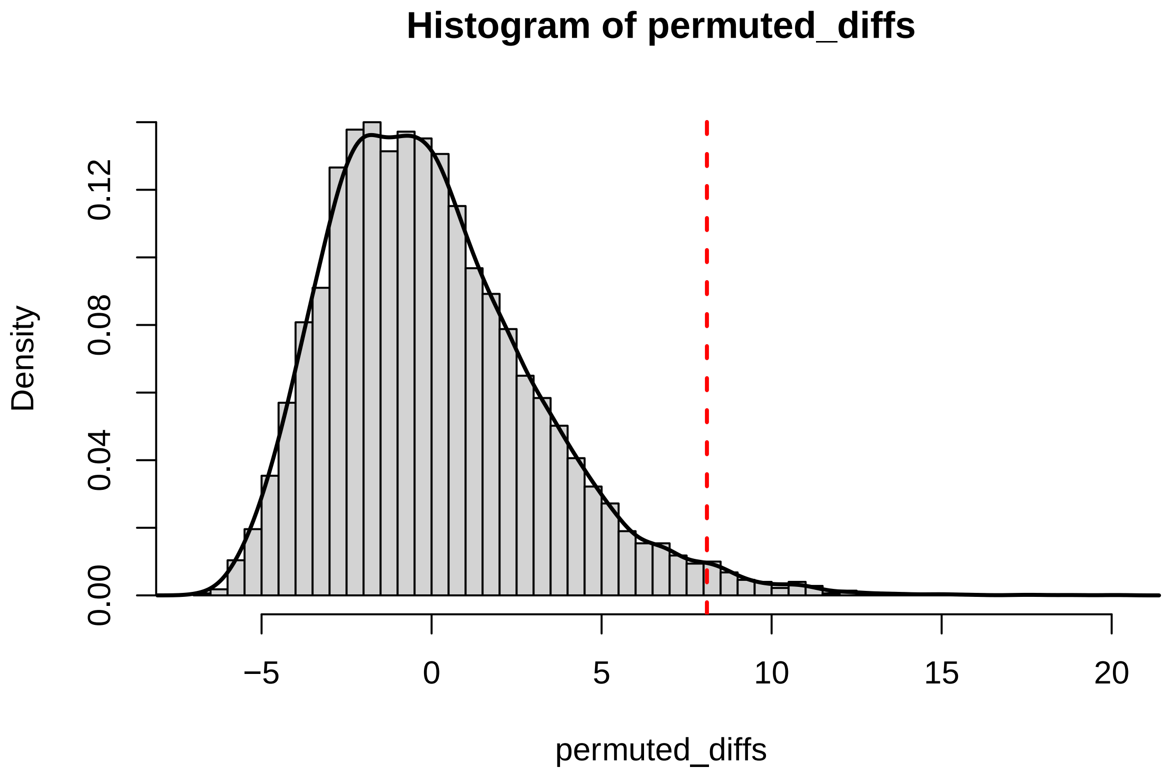
```
# Observed Difference
observed_diff <- mean(carrier$CLEC) - mean(carrier$ILEC)

#Permutation Function
permute_diff <- function(values, groups) {
permuted <- sample(values, replace = FALSE)
grouped <- split(permuted, as.factor(groups))
permuted_diff <- mean(grouped$CLEC) - mean(grouped$ILEC)
return(permuted_diff)
}

permute_diff(data_long$response_time, data_long$Carrier)
```

```
## [1] -3.652678
```

```
set.seed(567)
nperms <- 10000
permuted_diffs <- replicate(nperms, permute_diff(data_long$response_time, data_long$Carrier))
```

```r
hist(permuted_diffs, breaks = "fd", probability = TRUE)
lines(density(permuted_diffs), lwd=2)
abline(v = observed_diff, col = "red", lwd = 2, lty = 2)
```

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



```r
# one-tailed
p_1tailed <- sum(permuted_diffs > observed_diff) / nperms
# two-tailed
p_2tailed <- sum(abs(permuted_diffs) > observed_diff) / nperms

cat("one-tailed:", p_1tailed, "two-tailed:", p_2tailed)
```

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

```
## one-tailed: 0.0188 two-tailed: 0.0188
```

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

- I would not reject $H_0$ since one-tailed p-value is greater than 0.01 (1% significance), which means there's no strong evidence that the response time of $CLEC$ is greater than $ILEC$.

## Question 3

**a. Compute the W statistic comparing the values. You may use either the permutation approach (try the functional form) or the rank sum approach, but you must implement it yourself.**

```r
# permutation approach (try the functional form)
gt_eq <- function(a, b) {
ifelse(a > b, 1, 0) + ifelse(a == b, 0.5, 0)
}
W <- sum(outer(carrier$CLEC, carrier$ILEC, FUN = gt_eq))
cat("W-statistic:", W)
```

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

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

```r
n1 <- length(carrier$CLEC)
n2 <- length(carrier$ILEC)
wilcox_p_1tail <- 1 - pwilcox(W, n1, n2)
cat("one-tailed p-value for W:", wilcox_p_1tail)
```

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

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

- Based on the result, since the p-value is smaller than 0.01, we reject $H_0$. The Wilcoxon test provides strong evidence that the response time of $CLEC$ is significantly greater than that of $ILEC$.

## Question 4

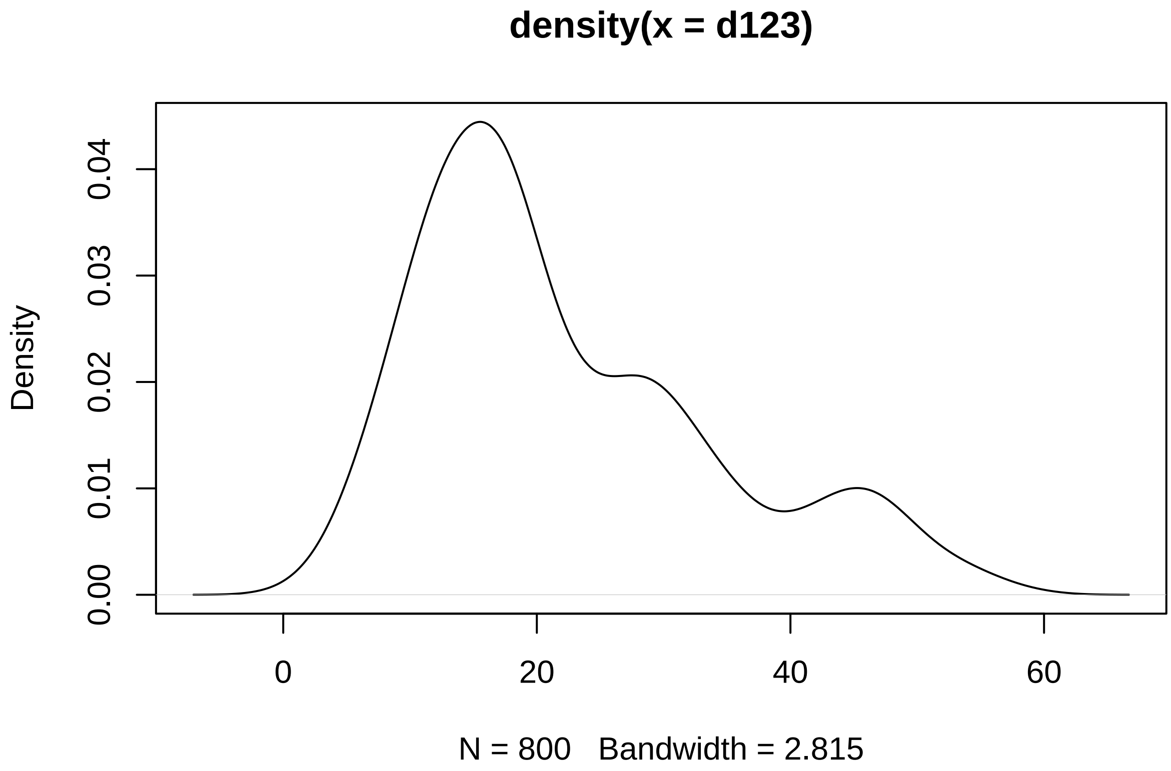
**a. Follow the following steps to create a function to see how a distribution of values compares to a perfectly normal distribution. The ellipses (...) in the steps below indicate where you should write your own code. Make a function called norm_qq_plot() that takes a set of values:**

```r
norm_qq_plot <- function(values) {
  # Create a sequence of probability numbers from 0 to 1, with ~1000 probabilities in between
  probs1000 <- seq(0, 1, 0.001)

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

  # Calculate ~1000 quantiles of a perfectly normal distribution with the same mean and standard deviat
  q_norm <- qnorm(probs1000, mean = mean(values, na.rm = TRUE),
                  sd = sd(values, na.rm = TRUE))

  # 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")

  # Finally, draw a red line with intercept of 0 and slope of 1, comparing these two sets of quantiles
  abline( a = 0, b=1 , col="red", lwd=2)
  }
```

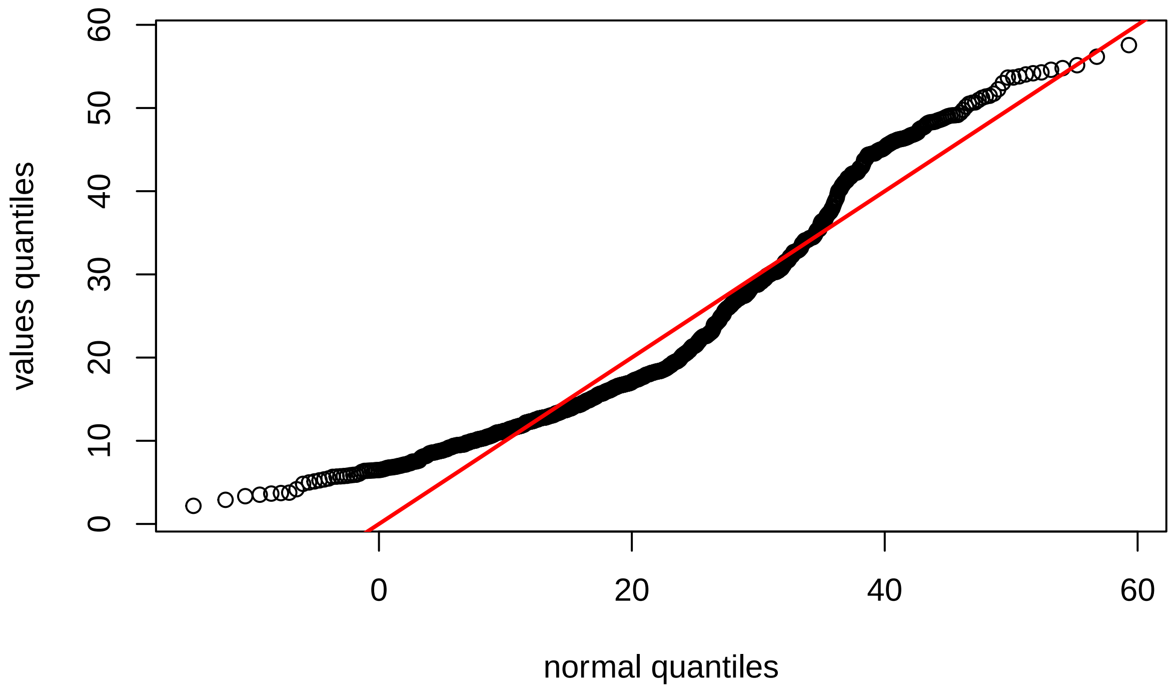
**You have now created a function that draws a "normal quantile-quantile plot" or Normal Q-Q plot (please show code for the whole function in your HW report)**

**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 and tell us if it suggests whether d123 is normally distributed or not.**

```r
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)
```

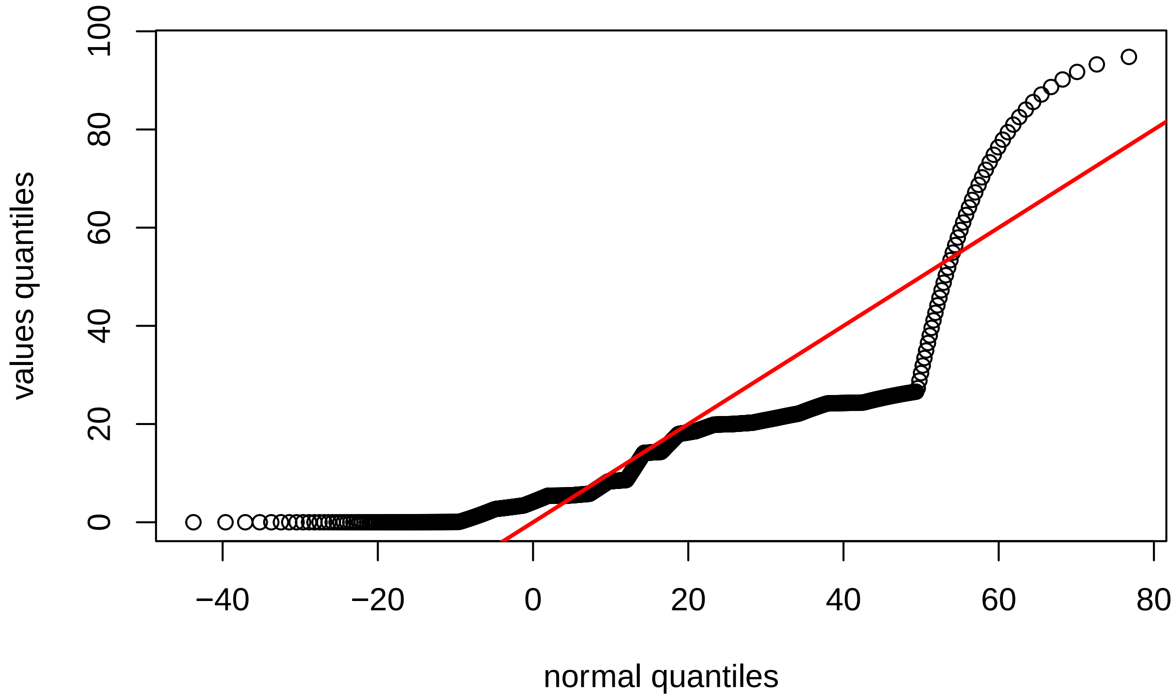


- According to the article, if the quantiles of $d123$ corresponded to those of a normally distributed $d123$, the black thick line would overlap the red reference line. However, in this case, we can clearly see that the black line does not overlap the red line, indicating that $d123$ is not normally distributed. Additionally, the Q-Q plot shows that $d123$ exhibits fat tails and is right-skewed.

**c. Traditional statistical t-tests to compare the means of two populations require that the two populations are normally distributed. Use your normal Q-Q plot function, norm_qq_plot, 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?**

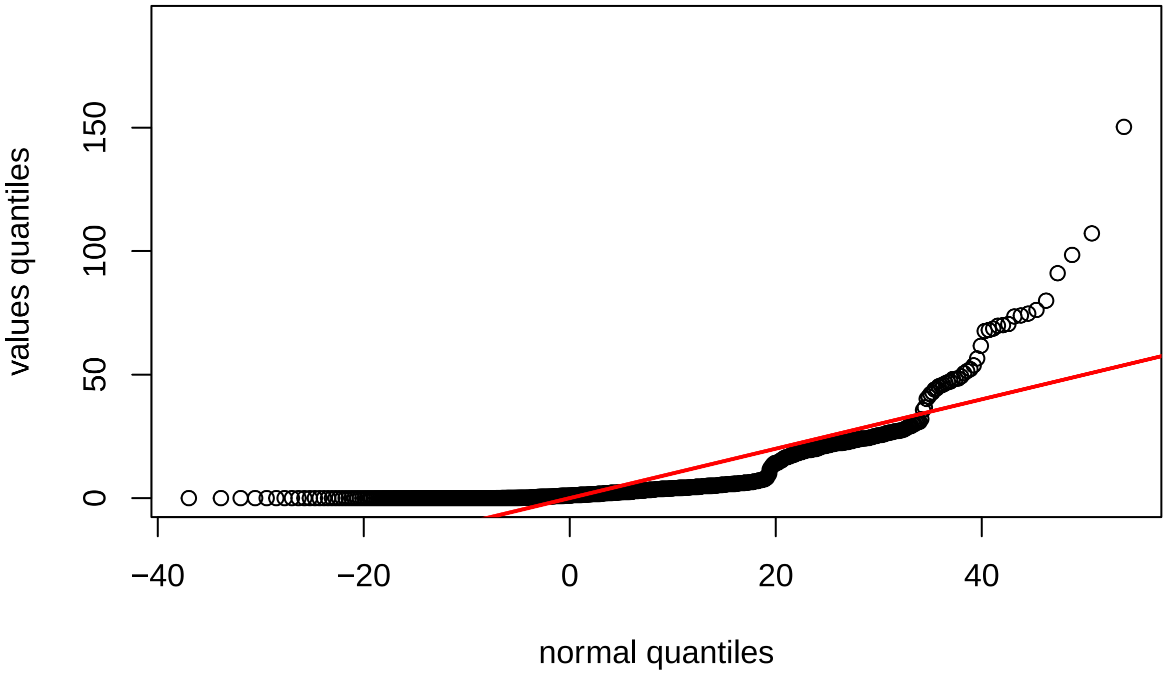
- QQ plot CLEC

```
norm_qq_plot(carrier$CLEC)
```



- The black line isn't corresponding the red line, which shows that *CLEC* sample is not normal distributed, instead it is strongly right skewed and has fat tails on both sides.
- QQ plot ILEC

```
norm_qq_plot(carrier$ILEC)
```



- The sample of $ILEC$ also does not follow normal distribution, the rise on the right side also shows a right skewness, and the kind of s-shape shows the fat tails on this distribution.
- Conclusion: Neither this 2 samples are normal distributed, which is not suitable to use the statistical test requiring normal distribution assumption (such as: Student's Two-Sample t-Test, Welch's Two-Sample t-Test). We could use the statistical test which does not require such assumption instead, such as MWW test.