BACS HW 6

109090046

2023-03-23

```
library(readr)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyve
rse 2.0.0 ──
## ✔ dplyr     1.1.0     ✔ purrr     1.0.1
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.1     ✔ tibble    3.1.8
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ── Conflicts ────────────────────────────────────────── tidyverse_co
nflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to
force all conflicts to become errors
```

```
verizon_wide <- read_csv("D:/下載/verizon_wide.csv")
```

```
## Rows: 1664 Columns: 2
## ── Column specification ────────────────────────────────────────────
────────────
## Delimiter: ","
## dbl (2): ILEC, CLEC
##
## ℹ Use `spec()` to retrieve the full column specification for this da
ta.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet
this message.
```

```
head(verizon_wide)
```

```
## # A tibble: 6 × 2
##    ILEC  CLEC
##   <dbl> <dbl>
## 1 17.5  26.6
## 2  2.4   8.6
## 3  0     0
## 4  0.65 21.2
## 5 22.2   8.33
## 6  1.2  20.3
```

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).

ANS:

I picked `tidyr` over other reshaping packages for a few reasons:

- *Consistency with other tidyverse packages:* `tidyr` is part of the `tidyverse`, a collection of R packages designed to work together seamlessly.
- *Community support:* `tidyr` is a widely-used package in the R community, which means that there are plenty of resources available online and also actively maintained.

Here is some links to support my decision:

- [Introducing tidyr]https://posit.co/blog/introducing-tidyr/
- [Tidyr on Stack Overflow]https://stackoverflow.com/questions/tagged/tidyr

b.

Show the code to reshape the versizon_wide.csv sample

```
verizon_long <- gather(verizon_wide, key="Group", value="Time")
```

c.

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

```
head(verizon_long)
```

```
## # A tibble: 6 × 2
##   Group  Time
##   <chr> <dbl>
## 1 ILEC  17.5
## 2 ILEC   2.4
## 3 ILEC   0
## 4 ILEC   0.65
## 5 ILEC  22.2
## 6 ILEC   1.2
```

```
tail(verizon_long)
```

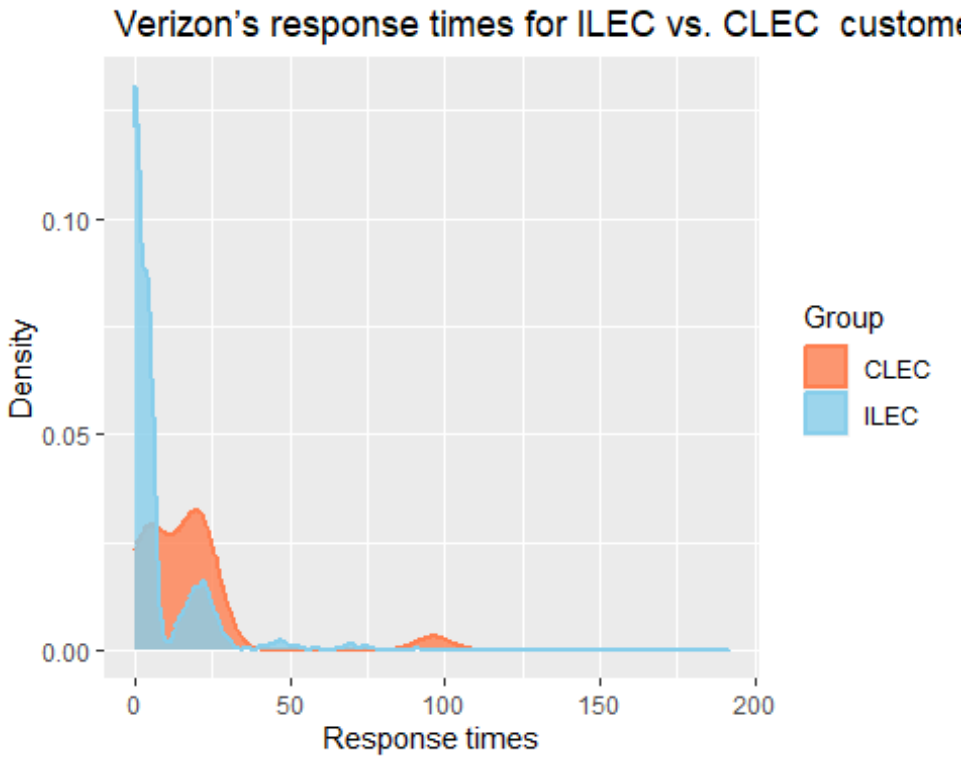
```
## # A tibble: 6 × 2
##   Group  Time
##   <chr> <dbl>
## 1 CLEC     NA
## 2 CLEC     NA
## 3 CLEC     NA
## 4 CLEC     NA
## 5 CLEC     NA
## 6 CLEC     NA
```

d.

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

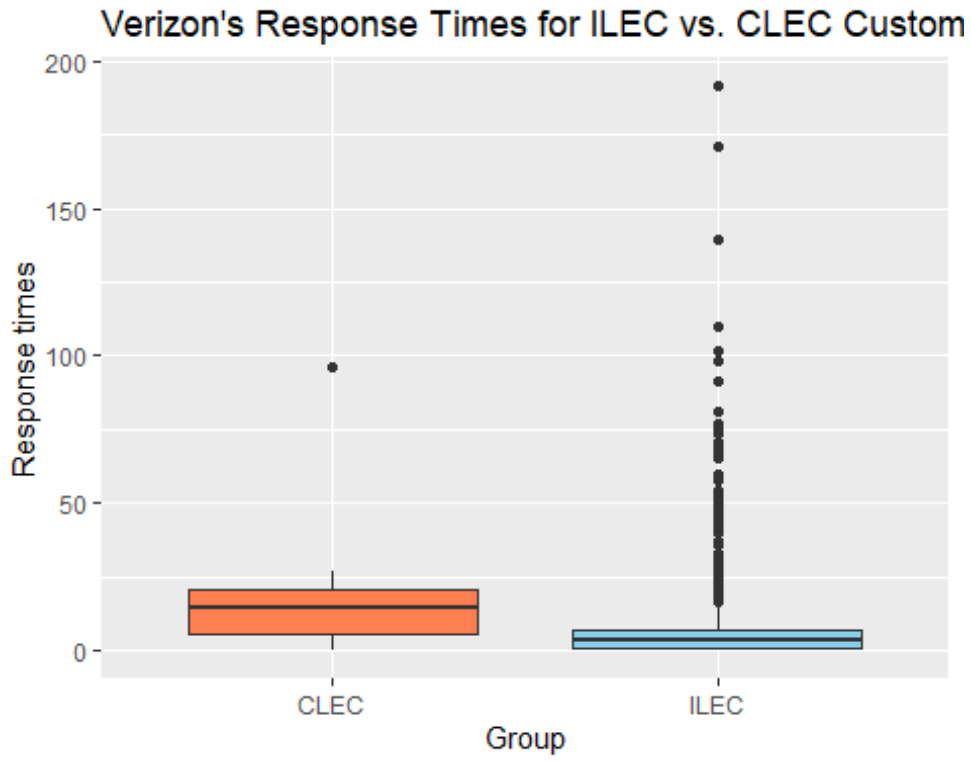
```
ggplot(verizon_long, aes(x = Time, group = Group, color = Group, fill=G
roup)) +
  geom_density(alpha=0.8, linewidth=0.8) +
  scale_color_manual(values = c("coral", "skyblue")) +
  scale_fill_manual(values = c("coral", "skyblue")) +
  labs(x="Response times", y="Density", title=" Verizon's response time
s for ILEC vs. CLEC  customers (Density Plot)")
```

```
## Warning: Removed 1641 rows containing non-finite values (`stat_densi
ty()`).
```



```r
# Boxplot
ggplot(verizon_long, aes(x = Group, y = Time, fill = Group)) +
  geom_boxplot() +
  scale_fill_manual(values = c("ILEC" = "skyblue", "CLEC" = "coral")) +
  theme(legend.position = "none") +
  labs(title = "Verizon's Response Times for ILEC vs. CLEC Customers (Box Plot)",
       x = "Group",
       y = "Response times")
```

```
## Warning: Removed 1641 rows containing non-finite values (`stat_boxplot()`).
```



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)

ANS:

Null hypothesis (H0): The mean response time for CLEC customers is equal to or less than the mean response time for ILEC customers.

Alternative hypothesis (Ha): The mean response time for CLEC customers is greater than the mean response time for ILEC customers.

This is a one-tailed hypothesis because we are only interested in the possibility that 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

```
v_equal <- t.test(verizon_wide$ILEC, verizon_wide$CLEC, var.equal = TRU
E, alternative = "greater", conf.level = 0.99)

## p-value (when variences are equal): 0.9954662
```

The p-value for the test is 0.9955, which is greater than the significance level of 0.01. Therefore, we fail to reject the null hypothesis that the mean response time for CLEC customers is equal to or less than the mean response time for ILEC customers. In other words, we do not have enough evidence to conclude that the mean response time for CLEC customers is greater than the mean response time for ILEC customers at 1% significance level.

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

```
v_inequal <- t.test(verizon_wide$ILEC, verizon_wide$CLEC, var.equal = F
ALSE, alternative = "greater", conf.level = 0.99)

## p-value (when variences are not equal): 0.9701269
```

The p-value for the test is 0.9701, which is greater than the significance level of 0.01. Therefore, we cannot reject the null hypothesis that the mean response time for CLEC customers is equal to or less than the mean response time for ILEC customers. In other words, we do not have enough evidence to conclude that the mean response time for CLEC customers is greater than the mean response time for ILEC customers at 1% significance level.

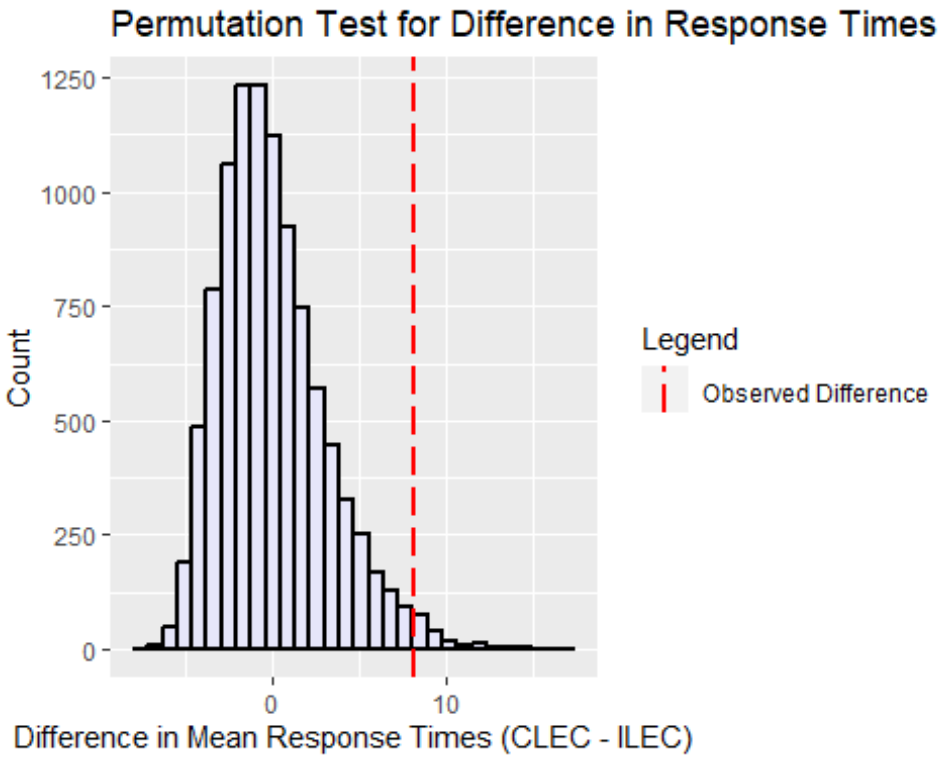
Overall, the p-values of both t-test are greater than significance level of 0.01. We fail to reject the null hypothesis that the mean response time for CLEC customers is equal to or less than the mean response time for ILEC customers at 1% significance level. Therefore, the sample means for the two groups were similar, and we did not have enough evidence to conclude that there was a significant difference in response times between ILEC and CLEC customers at a 1% significance level.

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.

```
set.seed(1234)
ILEC <- verizon_wide$ILEC
CLEC <- verizon_wide$CLEC

# removing NA in CLEC
CLEC <- CLEC[!is.na(CLEC)]

# calculate the observed difference in means
obs_diff <- mean(CLEC) - mean(ILEC)

# removing NA in verizon_long
verizon_long <- verizon_long[complete.cases(verizon_long),]

# permutation function
perm_fun <- function(x) {
  perm_time <- sample(x$Time)
  mean_diff <- mean(perm_time[x$Group == "CLEC"]) -
               mean(perm_time[x$Group == "ILEC"])
  return(mean_diff)
}

n_perms <- 10000
perm_diffs <- replicate(n_perms, perm_fun(verizon_long))

perm_diffs <- data.frame(perm_diffs)

# Create a histogram of the permuted differences
ggplot(perm_diffs, aes(x = perm_diffs)) +
  geom_histogram(color = "black", fill = "lavender", linewidth = 0.8) +
  geom_vline(aes(xintercept = obs_diff, color = "Observed Difference"),
 linewidth = 1, linetype="longdash") +
  labs(title = "Permutation Test for Difference in Response Times",
       x = "Difference in Mean Response Times (CLEC - ILEC)",
       y = "Count") +
  scale_color_manual(name="Legend", values=c("Observed Difference"="red
"))

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
one_tailed_p <- mean(perm_diffs >= obs_diff)
two_tailed_p <- mean(abs(perm_diffs) >= abs(obs_diff)) * 2

## One-tailed p-value: 0.0178

## Two-tailed p-value: 0.0356
```

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

In the one-tailed test, the significance level is typically divided by 2, as the null hypothesis is only rejected if the observed difference is in a specific direction. Therefore, for a 1% significance level in a one-tailed test, the critical p-value would be 0.01/2 = 0.005.

The one-tailed p-value is 0.0178, which is greater than the critical p-value of 0.005. Therefore, you would not reject the null hypothesis at a 1% significance level in a one-tailed test.

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.

**Using Rank Sum Approach

```
verizon_long$Rank <- rank(verizon_long$Time)

W <- sum(verizon_long$Rank[verizon_long$Group == "CLEC"])

## Rank sum (W statistic): 27096
```

b.

Compute the one-tailed p-value for W.

```
# Calculate the normal approximation p-value
n_clec <- sum(verizon_long$Group == "CLEC")
n_ilec <- sum(verizon_long$Group == "ILEC")

# Calculate mu(mean)
mu <- n_clec * (n_clec + n_ilec + 1) / 2

# Calculate sigma(standard deviation)
sigma <- sqrt(n_clec * n_ilec * (n_clec + n_ilec + 1) / 12)

# Calculate z(z-score)
z <- (W - mu) / sigma

p_value <- 1 - pnorm(z)

## One-tailed p-value: 0.0004636522
```

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.

```
wilcoxon_test <- wilcox.test(CLEC, ILEC, alternative = "greater")
wilcoxon_test$statistic

##     W
## 26820

wilcoxon_test$p.value

## [1] 0.0004565138

## W statistic: 26820

## One-tailed p-value: 0.0004565138
```

The resulting W statistic and p-value are almost the same, but I think the differences are from the different computing way and algorithm.

d.

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

ANS:

In the one-tailed test, the significance level is typically divided by 2, as the null hypothesis is only rejected if the observed difference is in a specific direction. Therefore, for a 1% significance level in a one-tailed test, the critical p-value would be 0.01/2 = 0.005

The one-tailed p-values we calculated are around 0.00045, which is smaller than critical value of 0.005. Hence, we reject the null hypothesis. In other words, we conclude that the response times for CLEC are different than ILEC.

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.

```
norm_qq_plot <- function(values) {

  # 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=prob
s1000), and name it q_vals
  q_vals <- quantile(values, probs=probs1000, na.rm=TRUE)

  # iii. Calculate ~1000 quantiles of a perfectly normal distribution w
ith the same mean and standard deviation as our values; name this vecto
r of normal quantiles q_norm
```

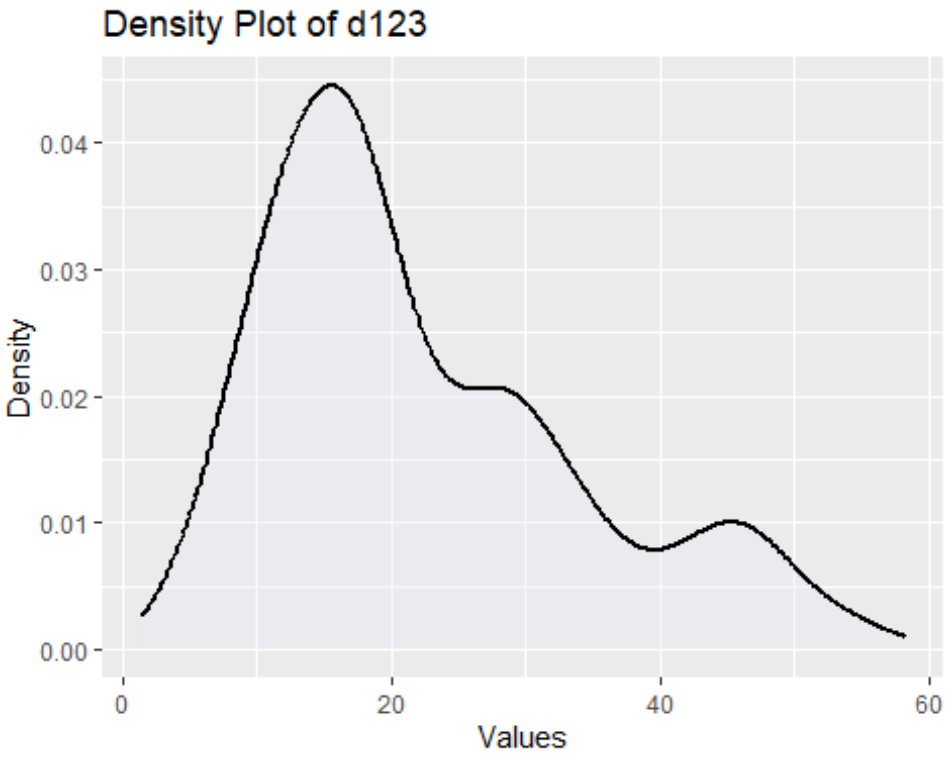
```
  q_norm <- qnorm(probs1000, mean=mean(values), sd=sd(values))

  # iv. Create a scatterplot comparing the quantiles of a normal distri
bution versus quantiles of values
  ggplot(data = data.frame(x = q_norm, y = q_vals), aes(x = x, y = y))
+
    geom_point() +
    labs(x="quantiles of a normal distribution", y="quantiles of values
") +

  # v. Draw a red line with intercept of 0 and slope of 1, comparing th
ese two sets of quantiles
    geom_abline(aes(intercept = 0, slope = 1, color = "abline"), linewi
dth = 0.8) +
    scale_color_manual(name="Legend", values=c("abline"="red"))
}
```

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.

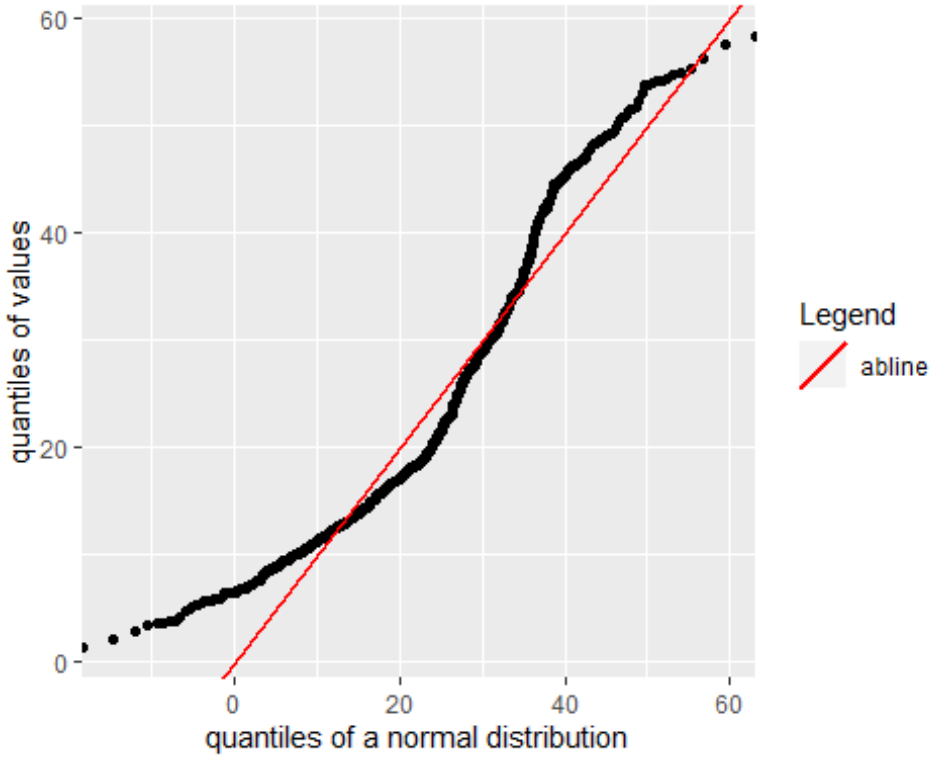
Confirm that your function works by running it against the values of our d123 distribution from week 3 and checking that it looks like the plot on the right:

```
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)

ggplot(data.frame(d123), aes(x=d123)) +
  geom_density(linewidth=0.8, fill="lavender", alpha=0.3) +
  labs(x = "Values", y = "Density", title="Density Plot of d123")
```



Normal Q-Q plot of d123

```
norm_qq_plot(d123)
```



Interpret the plot you produced and tell us if it suggests whether `d123` is normally distributed or not.

**ANS:*

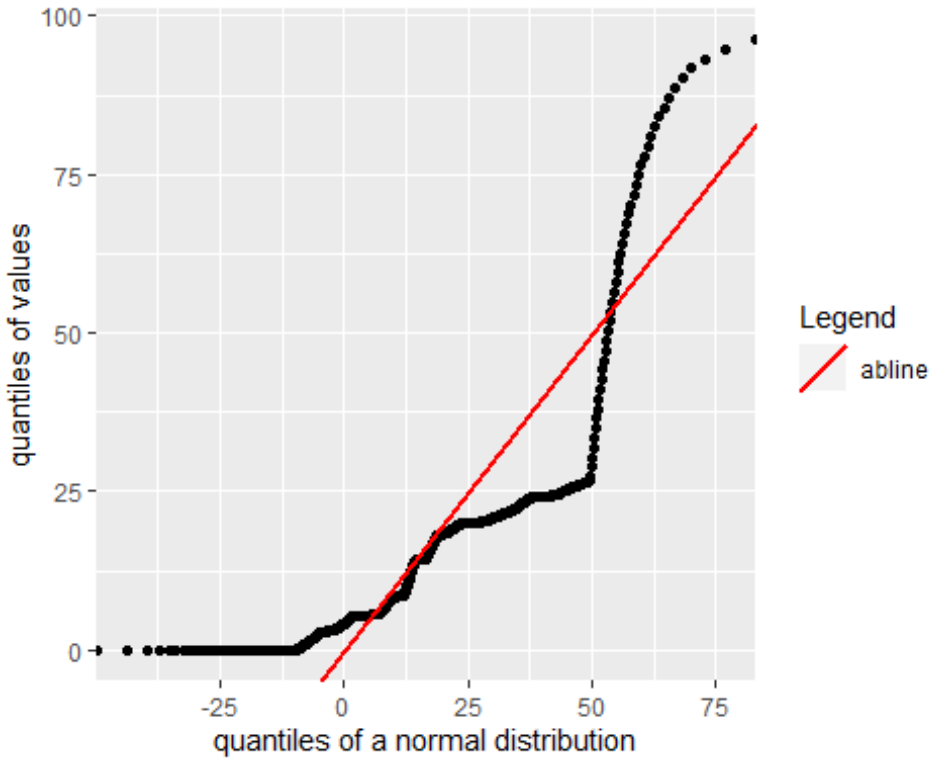
The `norm_qq_plot()` function indicates if the CDF(Cumulative Distribution Function) of input value is similar with the normal distribution, then the black dots will follow the red line. From the plot shown above, we can see that the black dots do follow the red line, though there are some deviations from the red line. Hence, we can preliminary confirm that `d123` is normally distributed by `norm_qq_plot()`.

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?

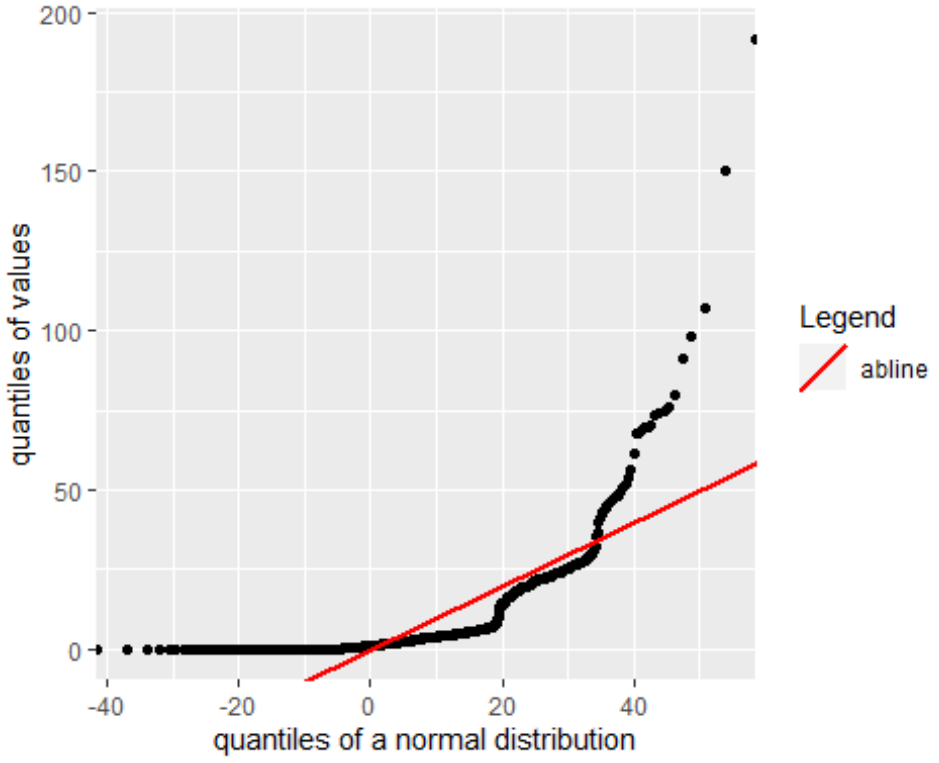
Normal Q-Q plot of CLEC response times

```
norm_qq_plot(CLEC)
```



Normal Q-Q plot of ILEC response times

```
norm_qq_plot(ILEC)
```



The response times of both ILEC and CLEC seems to be not normally distributed because of the large deviation from the normal distribution CDF. Especially, the head and tail of these two distribution are far way from the normal distribution. We can preliminary conclude that they are not normally distributed, but for further discussion, we can use the Kolmogorov–Smirnov test and the Shapiro–Wilk test to test them if they are normally distributed or not.