Homework6_BACS

109090035

3/23/2023

Q1 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.

1. 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 recommend using the tidyverse package, specifically the tidyr package, for reshaping the Verizon dataset in R. The main reason is that the tidyverse package is a collection of R packages designed to work together seamlessly for data manipulation, exploration, and visualization. tidyr is a part of the tidyverse, and it provides an easy-to-understand and consistent syntax for reshaping data.

There are two main functions in the tidyr package for reshaping data: pivot_longer() and pivot_wider(). In this case, we want to convert the wide data frame to a long format, so we’ll use the pivot_longer() function.

```
library(tidyverse)
```

2. Show the code to reshape the versizon_wide.csv sample

```
# Original data
verizon_data <- read.csv("/Users/user/Downloads/verizon_wide.csv")
```

```
# Reshape the data from wide to long format using pivot_longer()
long_verizon_data <- verizon_data %>%
  pivot_longer(cols = c("ILEC", "CLEC"), names_to = "Carrier", values_to = "Response_Time")

# Remove NA
long_verizon_data <- na.omit(long_verizon_data)
```

3. Show us the “head” and “tail” of the data to show that the reshaping worked

```
head(long_verizon_data) %>% ptable()
```

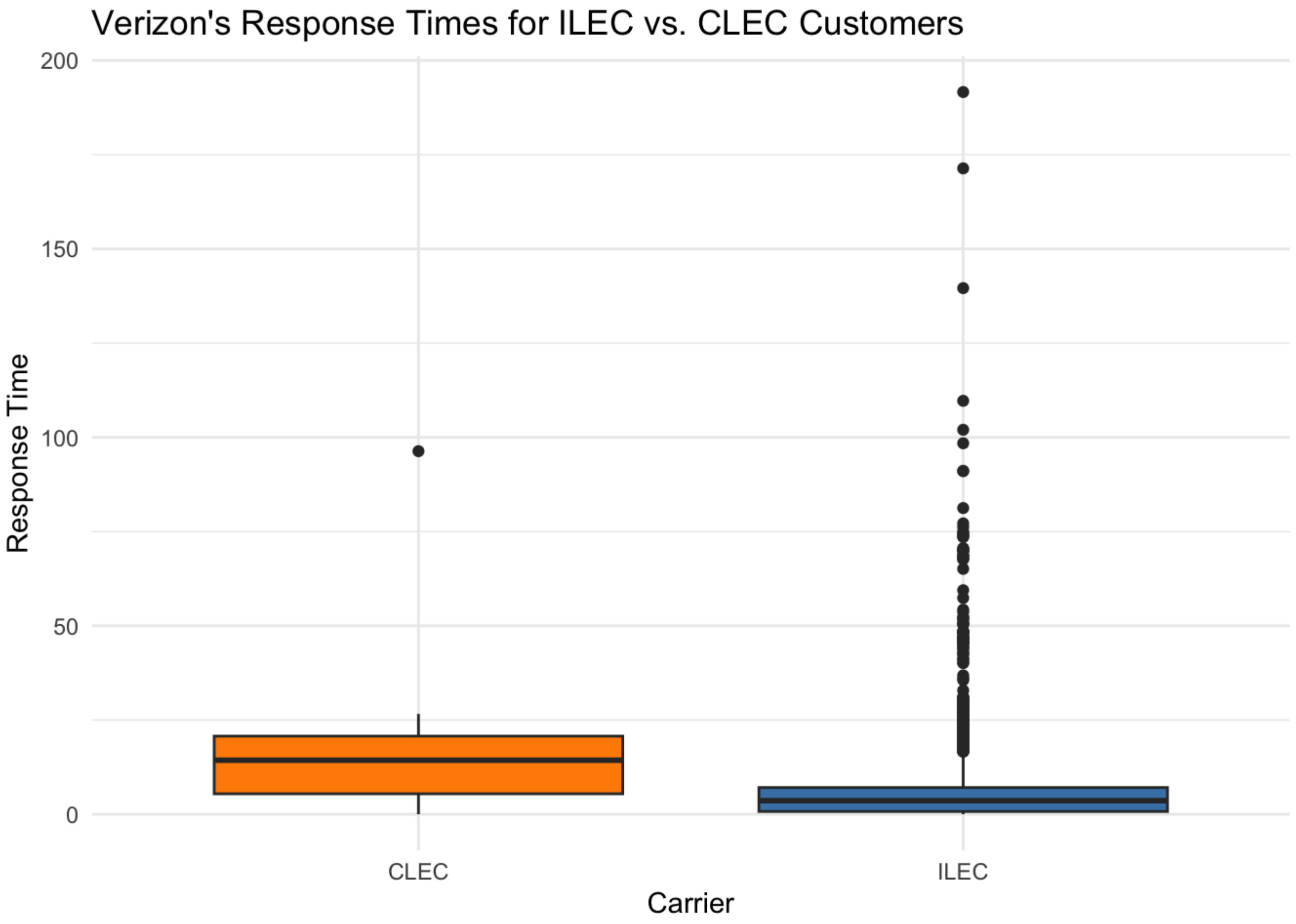
Carrier	Response_Time
ILEC	17.50
CLEC	26.62
ILEC	2.40
CLEC	8.60
ILEC	0.00
CLEC	0.00

```
tail(long_verizon_data) %>% ptable()
```

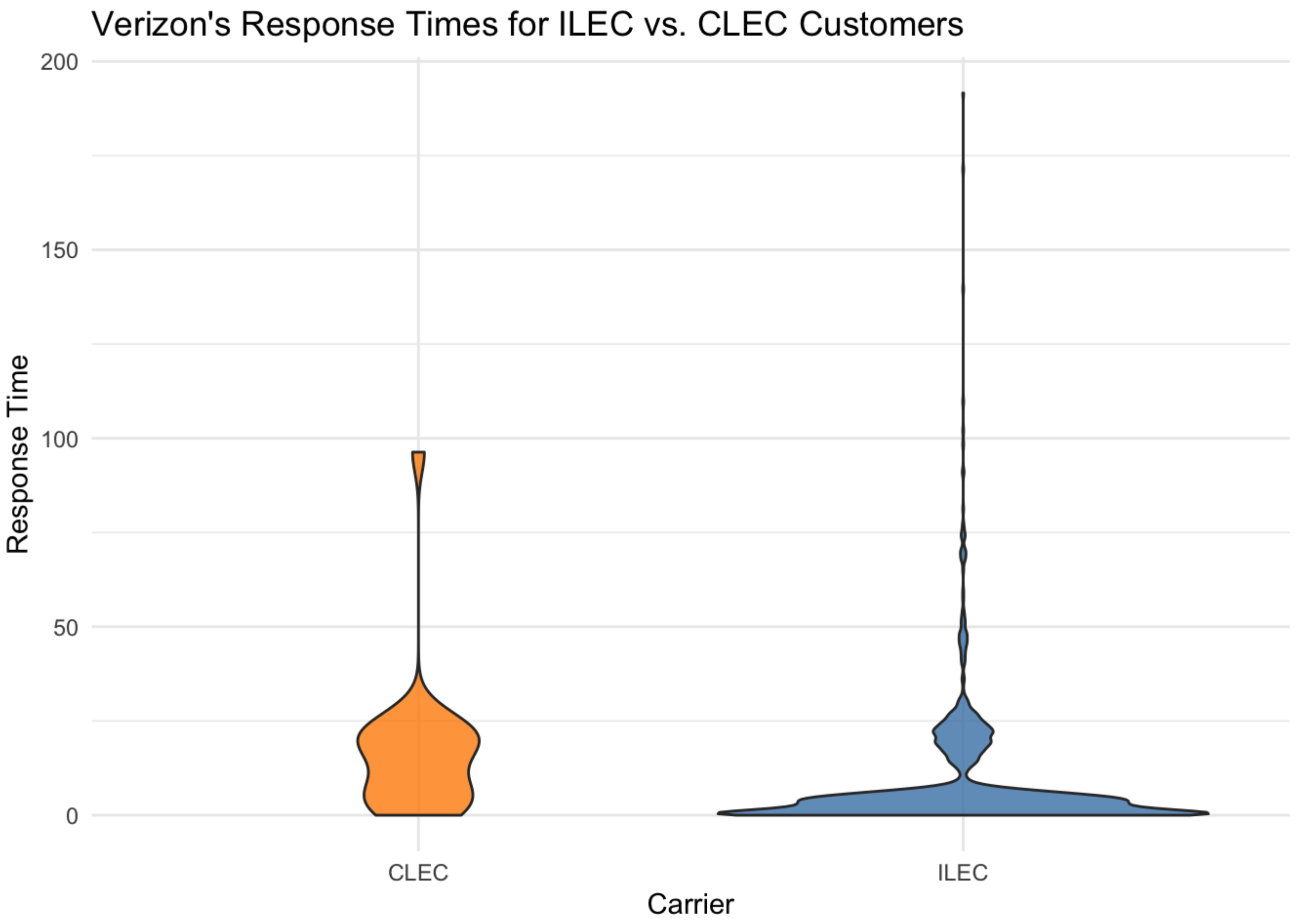
Carrier	Response_Time
ILEC	0.00
ILEC	23.70
ILEC	25.42
ILEC	4.83
ILEC	3.60
ILEC	18.13

4. Visualize Verizon’s response times for ILEC vs. CLEC customers

```
# Colorful boxplot
ggplot(long_verizon_data, aes(x = Carrier, y = Response_Time, fill = Carrier)) +
  geom_boxplot() +
  scale_fill_manual(values = c("ILEC" = "steelblue", "CLEC" = "darkorange")) +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(title = "Verizon's Response Times for ILEC vs. CLEC Customers",
       x = "Carrier",
       y = "Response Time")
```



```
# Violin plot
ggplot(long_verizon_data, aes(x = Carrier, y = Response_Time, fill = Carrier)) +
  geom_violin(alpha = 0.8) +
  scale_fill_manual(values = c("ILEC" = "steelblue", "CLEC" = "darkorange")) +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(title = "Verizon's Response Times for ILEC vs. CLEC Customers",
       x = "Carrier",
       y = "Response Time")
```



Q2 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 and alternative hypotheses (one-tailed):

Null hypothesis (H0): The mean response time for ILEC customers (μ1) is less than or equal to the mean response time for CLEC customers (μ2). Mathematically, H0: μ1 <= μ2.

Alternative hypothesis (H1): The mean response time for CLEC customers (μ2) is greater than the mean response time for ILEC customers (μ1). Mathematically, H1: μ2 > μ1.

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.

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

```
# Perform the t-test assuming equal variances
equal_var_ttest <- t.test(Response_Time ~ Carrier, data = long_verizon_data,
                          alternative = "greater", var.equal = TRUE, conf.level = 0.99)

# Display the results
print(equal_var_ttest)
```

```

    Two Sample t-test

data:  Response_Time by Carrier
t = 2.6125, df = 1685, p-value = 0.004534
alternative hypothesis: true difference in means between group CLEC and group ILEC is greater than 0
99 percent confidence interval:
 0.8801387       Inf
sample estimates:
mean in group CLEC mean in group ILEC
         16.509130           8.411611
```

```
# Check if the null hypothesis should be rejected
if (equal_var_ttest$p.value < 0.01) {
  cat("Reject the null hypothesis: There is evidence to support that CLEC response times are significantly greater than ILEC response times.\n")
} else {
  cat("Do not reject the null hypothesis: There is not enough evidence to support that CLEC response times are significantly greater than ILEC response times.\n")
}
```

```
Reject the null hypothesis: There is evidence to support that CLEC response times are significantly greater than ILEC response times.
```

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

```
# Perform the t-test assuming unequal variances (Welch's t-test)
unequal_var_ttest <- t.test(Response_Time ~ Carrier, data = long_verizon_data,
                            alternative = "greater", var.equal = FALSE, conf.level = 0.99)

# Display the results
print(unequal_var_ttest)
```

```
    Welch Two Sample t-test

data:  Response_Time by Carrier
t = 1.9834, df = 22.346, p-value = 0.02987
alternative hypothesis: true difference in means between group CLEC and group ILEC is greater than 0
99 percent confidence interval:
 -2.130858       Inf
sample estimates:
mean in group CLEC mean in group ILEC 
         16.509130           8.411611 
```

```
# Check if the null hypothesis should be rejected
if (unequal_var_ttest$p.value < 0.01) {
  cat("Reject the null hypothesis: There is evidence to support that CLEC response times are significantly greater than ILEC response times.\n")
} else {
  cat("Do not reject the null hypothesis: There is not enough evidence to support that CLEC response times are significantly greater than ILEC response times.\n")
}
```

```
Do not reject the null hypothesis: There is not enough evidence to support that CLEC response times are significantly greater than ILEC response times.
```

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

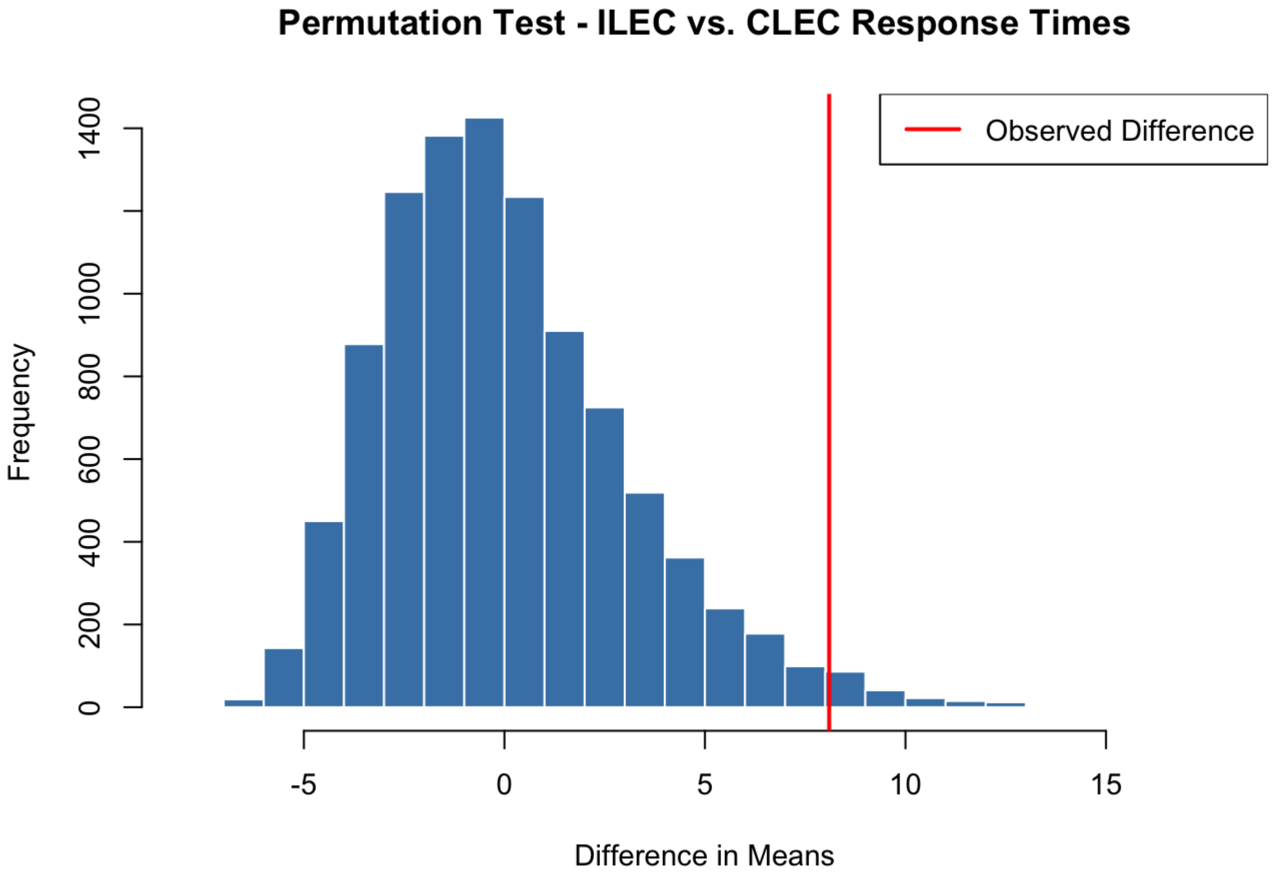
1. Visualize the distribution of permuted differences, and indicate the observed

difference as well.

```
# Function to perform a single permutation
permute_means <- function(data) {
  permuted_response_time <- sample(data$Response_Time)
  mean_diff <- mean(permuted_response_time[data$Carrier == "CLEC"]) -
               mean(permuted_response_time[data$Carrier == "ILEC"])
  return(mean_diff)
}



# Perform the permutation test
n_permutations <- 10000
permuted_diffs <- replicate(n_permutations, permute_means(long_verizon_data))

# Check the first few values of permuted_diffs
head(permuted_diffs)
```

```
[1]  6.400029  2.147706  5.742808 -3.392611  1.967863  1.116693
```

```
observed_diff <- mean(long_verizon_data$Response_Time[long_verizon_data$Carrier == "CLEC"]) -
                 mean(long_verizon_data$Response_Time[long_verizon_data$Carrier == "ILEC"])

# Plot the distribution of permuted differences
hist(permuted_diffs, breaks = 30, col = "steelblue", border = "white",
     main = "Permutation Test - ILEC vs. CLEC Response Times",
     xlab = "Difference in Means",
     ylab = "Frequency")
abline(v = observed_diff, col = "red", lwd = 2)
legend("topright", legend = c("Observed Difference"), col = c("red"), lty = 1, lwd = 2)
```



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

```
# Calculate the one-tailed and two-tailed p-values
one_tailed_p_value <- mean(permuted_diffs >= observed_diff)
two_tailed_p_value <- mean(abs(permuted_diffs) >= abs(observed_diff)) * 2

cat("One-tailed p-value:", one_tailed_p_value, "\n")
```

```
One-tailed p-value: 0.0177
```

```
cat("Two-tailed p-value:", two_tailed_p_value, "\n")
```

```
Two-tailed p-value: 0.0354
```

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

To determine whether to reject the null hypothesis at 1% significance in a one-tailed test, compare the one-tailed p-value with the significance level (0.01).

Since the one-tailed p-value is greater than or equal to 0.01, you cannot reject the null hypothesis. There is insufficient evidence to conclude that the mean response time for CLEC customers is significantly greater than for ILEC customers.

Q3 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.

Rank sum approach

```
# Function of ranks of CLEC (W statistic)
gt_eq <- function(a,b){
 ifelse(a > b, 1, 0) + ifelse(a == b, 0.5, 0)
}

# Calculate the observed rank sum (W statistic) for CLEC customers
observed_rank_sum <- sum(outer(long_verizon_data$Response_Time[long_verizon_data$Carrier=="CLEC"], long_verizon_d
ata$Response_Time[long_verizon_data$Carrier=="ILEC"], FUN = gt_eq))

cat("Observed rank sum (W statistic):", observed_rank_sum, "\n")
```

```
Observed rank sum (W statistic): 26820
```

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

To compute the one-tailed p-value for W without using any package or function, we Use the normal approximation method.

First, we need to calculate the mean (mu) and standard deviation (sigma) of the W distribution under the null hypothesis, and then we can calculate the Z-score for the observed W statistic. Finally, we can use the standard normal distribution to find the p-value.

```
# Number of ILEC and CLEC observations
n_ILEC <- sum(long_verizon_data$Carrier == "ILEC")
n_CLEC <- sum(long_verizon_data$Carrier == "CLEC")

# Calculate the mean (mu) and variance (var) of the W distribution under the null hypothesis
mu <- n_CLEC * (n_ILEC + n_CLEC + 1) / 2
var <- n_ILEC * n_CLEC * (n_ILEC + n_CLEC + 1) / 12

# Calculate the Z-score for the observed W statistic
Z <- (observed_rank_sum - mu) / sqrt(var)

# Calculate the one-tailed p-value using the standard normal distribution
one_tailed_p_value <- pnorm(Z, lower.tail = FALSE)

cat("One-tailed p-value:", one_tailed_p_value, "\n")
```

```
One-tailed p-value: 0.0007046266
```

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.

```
# Convert the Carrier variable to a factor
long_verizon_data$Carrier <- as.factor(long_verizon_data$Carrier)

# Separate response times for ILEC and CLEC
ILEC_response_times <- long_verizon_data$Response_Time[long_verizon_data$Carrier == "ILEC"]
CLEC_response_times <- long_verizon_data$Response_Time[long_verizon_data$Carrier == "CLEC"]

# Perform the Wilcoxon Rank Sum test
wilcoxon_test <- wilcox.test(CLEC_response_times, ILEC_response_times,
                             alternative = "greater", exact = TRUE, conf.int = TRUE)
```

```
Warning in wilcox.test.default(CLEC_response_times, ILEC_response_times, :
cannot compute exact p-value with ties
```

```
Warning in wilcox.test.default(CLEC_response_times, ILEC_response_times, :
cannot compute exact confidence intervals with ties
```

```
# Get the W statistic and one-tailed p-value
W_statistic <- wilcoxon_test$statistic
one_tailed_p_value <- wilcoxon_test$p.value

cat("W statistic:", W_statistic, "\n")
```

```
W statistic: 26820
```

```
cat("One-tailed p-value:", one_tailed_p_value, "\n")
```

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

We can see the p-value result in (a) and (c) is nearly the same,Without using package and function or use the base r function wilcox.test().

In this case, the W statistic would be 26820. The important part is that the p-value is nearly consistent and indicates a significant difference between the groups.

The W statistic value can be different depending on the way it is calculated, as it can represent either the sum of ranks for the first group or the second group.

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

To determine whether to reject the null hypothesis at a 1% significance level in a one-tailed test, compare the one-tailed p-value with the significance level (0.01).

Since the one-tailed p-value is less than 0.01, reject the null hypothesis, concluding that the values of CLEC and ILEC are significantly different.

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):

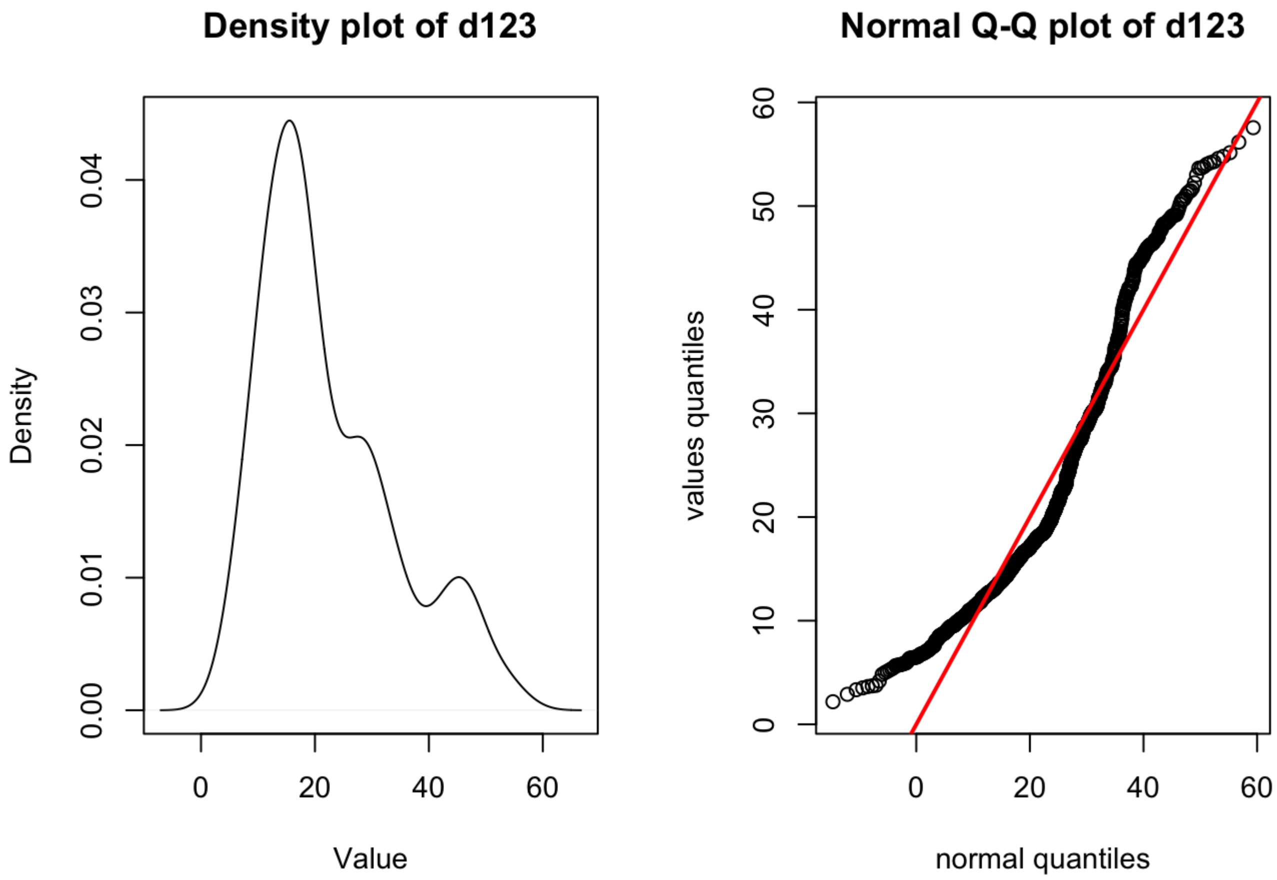
```
norm_qq_plot <- function(values) {
  # Create a sequence of probability numbers from 0 to 1, with ~1000 probabilities in between
  probs1000 <- seq(0, 1, 0.001)

  # Calculate ~1000 quantiles of our values, and name it q_vals
  q_vals <- quantile(values, probs = probs1000)

  # 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(probs1000, mean = mean(values), sd = sd(values))

  # 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")

  # 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:

```
set.seed(978234)
# Three normally distributed data sets
d1 <- rnorm(n=500, mean=15, sd=5)
d2 <- rnorm(n=200, mean=30, sd=5)
d3 <- rnorm(n=100, mean=45, sd=5)

# Combining them into a composite dataset
d123 <- c(d1, d2, d3)

# Create a 2x1 plot grid
par(mfrow = c(1, 2))

plot(density(d123), main = "Density plot of d123", xlab = "Value", ylab = "Density")
norm_qq_plot(d123)
title(main = "Normal Q-Q plot of d123")
```



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.

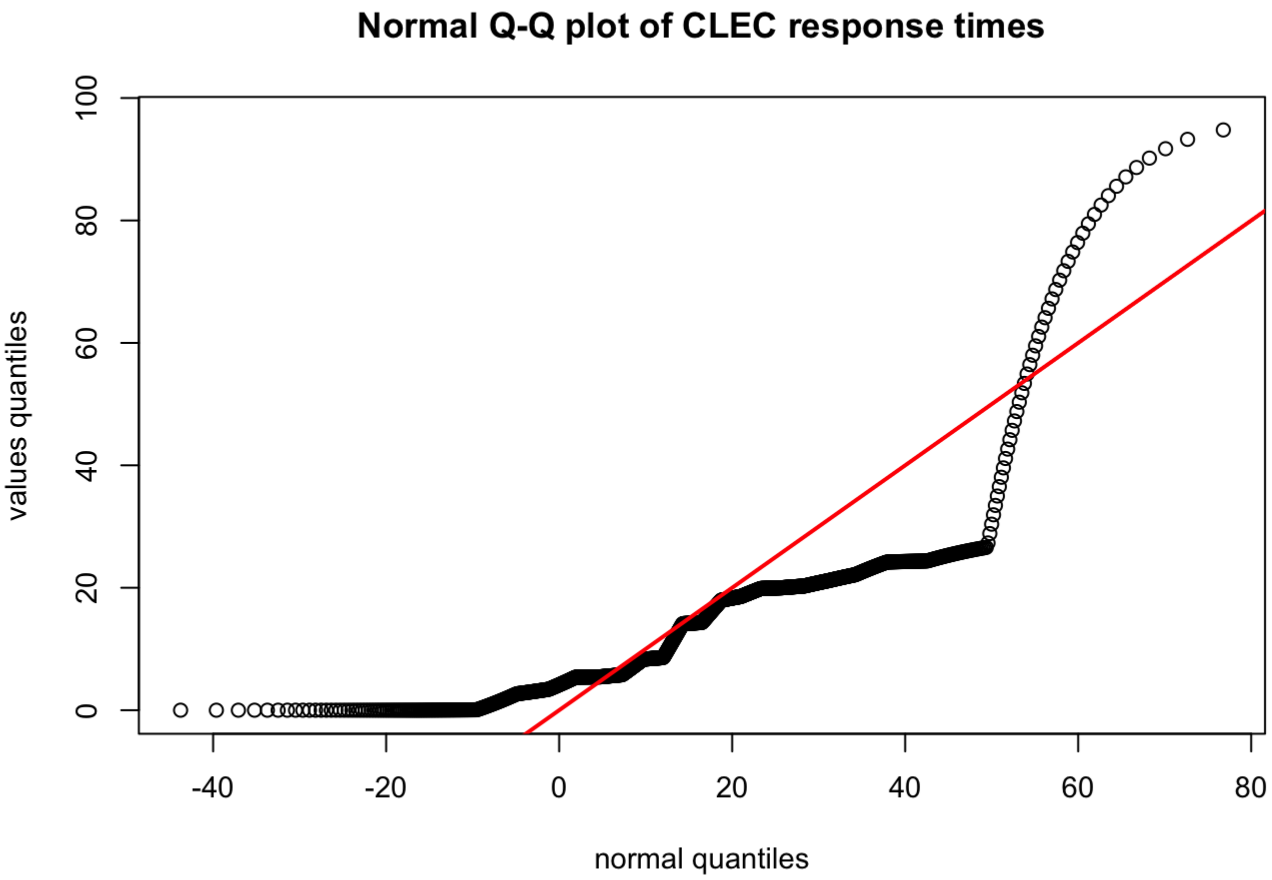
The Normal Q-Q plot produced by the norm_qq_plot() function suggests that d123 is approximately normally distributed.

The blue dots in the plot follow the red line fairly closely, which indicates that the distribution of d123 is not too different from a normal distribution. However, there are some deviations from the red line towards the lower end and upper end of the distribution, which suggests that there may be some slight departures from normality in the tails of the distribution.

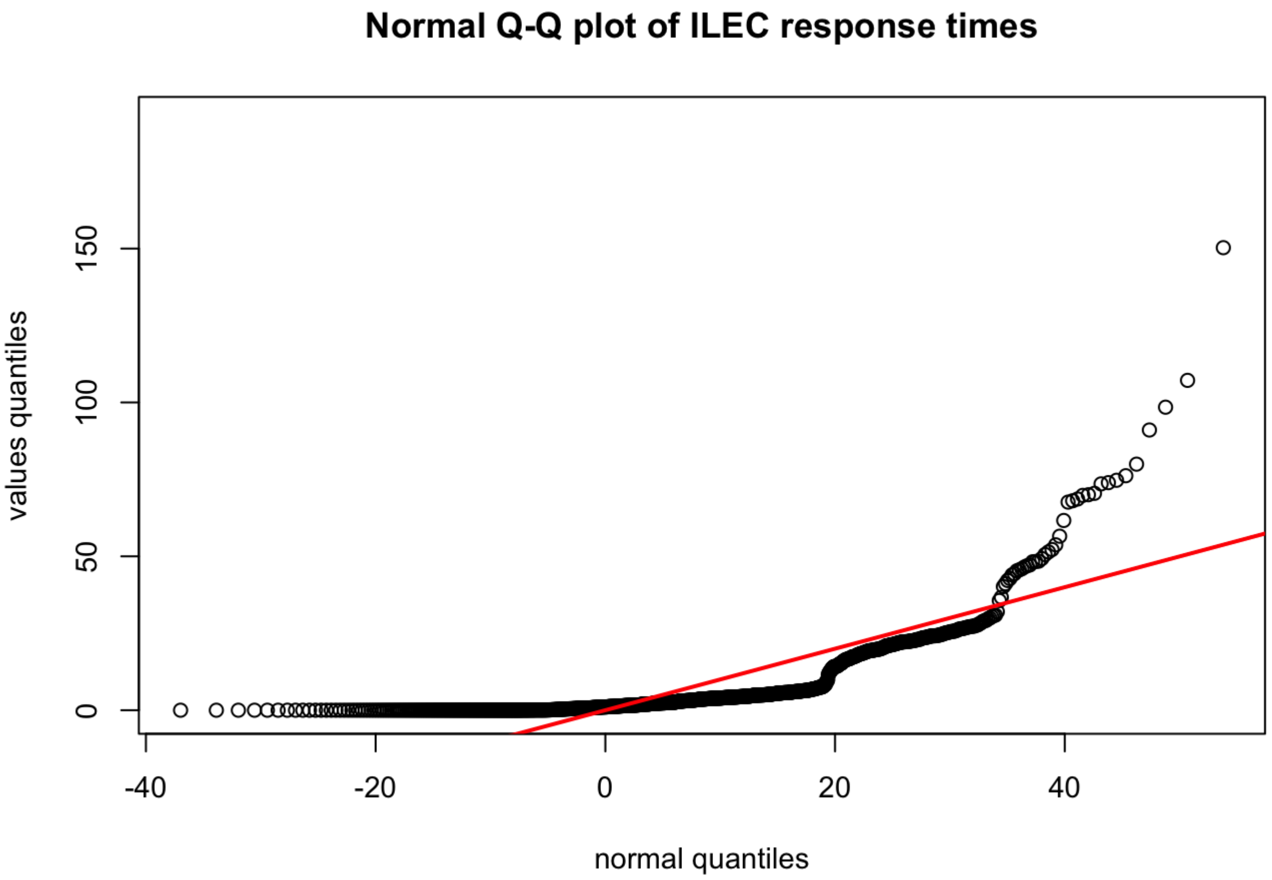
Overall, based on the Normal Q-Q plot, we can tentatively conclude that d123 is approximately normally distributed, but we should complement this visual inspection with formal statistical tests for normality, such as the Shapiro-Wilk or Anderson-Darling tests, to confirm this conclusion.

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
# Extract response times for CLEC and ILEC customers
clec_times <- long_verizon_data$Response_Time[long_verizon_data$Carrier == "CLEC"]
ilec_times <- long_verizon_data$Response_Time[long_verizon_data$Carrier == "ILEC"]

# Create Normal Q-Q plots for CLEC and ILEC response times
norm_qq_plot(clec_times)
title(main = "Normal Q-Q plot of CLEC response times")
```



```
norm_qq_plot(ilec_times)
title(main = "Normal Q-Q plot of ILEC response times")
```



We can see that the CLEC and ILEC response times do not appear to be normally distributed. In both plots, there are significant deviations from the diagonal red line, especially towards the tails of the distribution.

This suggests that the CLEC and ILEC response times may be non-normally distributed, and this conclusion is consistent with the results of the Shapiro-Wilk test and Wilcoxon rank-sum test we performed earlier.