BACS HW6 - 109006234

Credit: 109006278

March 26th 2023

Problem 1

(a) Pick a reshaping package

```r
library(tidyr)
```

The tidyr package is using gather(), which allows you to specify the columns to be reshaped, the names of the new columns, and how to handle missing values, among other options. It is easier to use if you are in the tidyverse ecosystem, such as dplyr, ggplot, etc.

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

```r
file <- read.csv("G:/My Drive/111_2_BACS/HW6/verizon_wide.csv")
verizon <- gather(file, na.rm = TRUE, key = "type", value = "time")
data <- split(x = verizon$time, f = verizon$type)
```

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

```r
head(verizon)
```

```
##   type  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
```

```r
tail(verizon)
```

```
##      type  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

```r
ilec <- data$ILEC
clec <- data$CLEC
plot(density(ilec), lwd=2, main="ILEC and CLEC Density Plots")
lines(density(clec), lwd=2, col="blue")
legend(
    150, 0.12, c("ILEC", "CLEC"),
    lwd = c(2,2), lty = c("solid", "solid"),
    col = c("black", "blue")
)
```

ILEC and CLEC Density Plots

N = 1664   Bandwidth = 0.9676
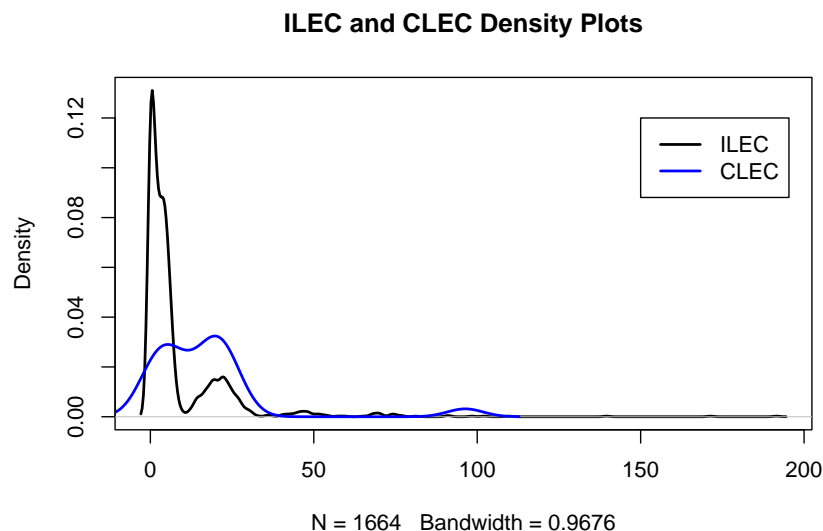
Problem 2

(a) State the appropriate null and alternative hypotheses (one-tailed)
The null hypothesis is the mean of ILEC is less than equal to CLEC. The alternative hypothesis is the mean of ILEC is greater than to the mean of CLEC.

(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.

(i) Conduct the test assuming variances of the two populations are equal

```r
t.test(clec, ilec, alt="greater", var.equal=TRUE, conf.level=0.99)
```

```
##
##  Two Sample t-test
##
## data:  clec and 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
```

2

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

```
t.test(clec, ilec, var.equal=FALSE, conf.level=0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  clec and ilec
## t = 1.9834, df = 22.346, p-value = 0.05975
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -3.393927 19.588967
## sample estimates:
## mean of x mean of y
## 16.509130  8.411611
```

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

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

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

```
hist(permuted_diffs, breaks = "fd", probability = TRUE,
    labels=seq(-10, 20, 5))
lines(density(permuted_diffs), lwd=2, col="darkgreen",
    main="permuted_diffs Plot")
abline(v=observed_diff, lty="dashed", col="red")
```


Histogram of permuted_diffs
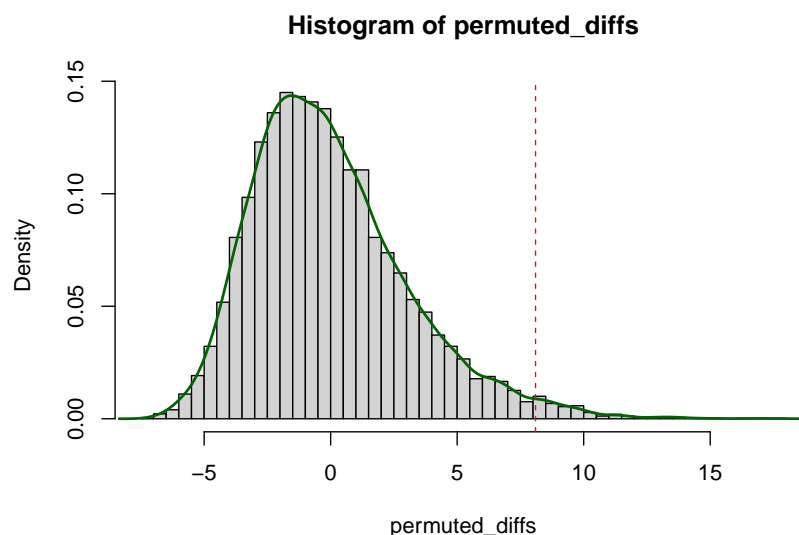
(ii) What are the one-tailed and two-tailed p-values of the permutation test?

```
p_1tailed <- sum(permuted_diffs > observed_diff) / nperms
p_1tailed
```

[1] 0.0185

```
p_2tailed <- sum(abs(permuted_diffs) > observed_diff) / nperms
p_2tailed
```

[1] 0.0185

(iii) Would you reject the null hypothesis at 1% significance in a one-tailed test?
Based on the calculation above, we can conclude that we fail to reject the null hypothesis.

Problem 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.

```
#Functional form
gt_eq <- function(a, b) {
    ifelse(a > b, 1, 0) + ifelse(a == b, 0.5, 0)
}
wilcox <- sum(outer(clec, ilec, FUN = gt_eq))
wilcox
```

[1] 26820

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

```
n1 <- length(ilec)
n2 <- length(clec)
wilcox_p_1tail <- 1 - pwilcox(wilcox, n1, n2)
wilcox_p_1tail
```

[1] 0.0003688341

(c) Run the Wilcoxon Test again using the wilcox.test() function in R

```
wilcox.test(clec, ilec, alternative = "greater", conf.level=0.01)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  clec and 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 (c), we observe that the Wilcoxon test yield a small p-value. Hence, we can say that we reject the null hypothesis.

Problem 4

(a) norm_qq_plot() Function

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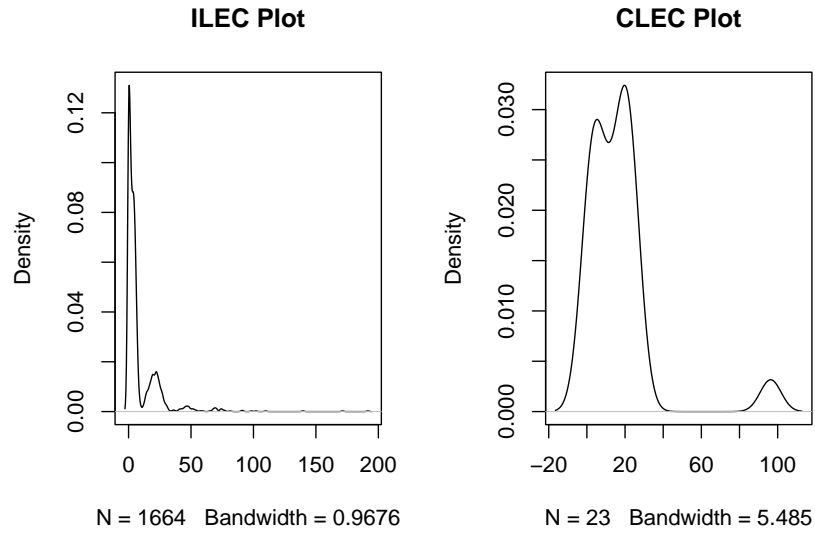
(b) Confirming norm_qq_plot() function

```
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)
layout(matrix(c(1,2), 1, 2, byrow = TRUE))
plot(density(d123), main = "Gaussian Distribution")
norm_qq_plot(values = d123, main = "d123 Q-Q Plot")
```
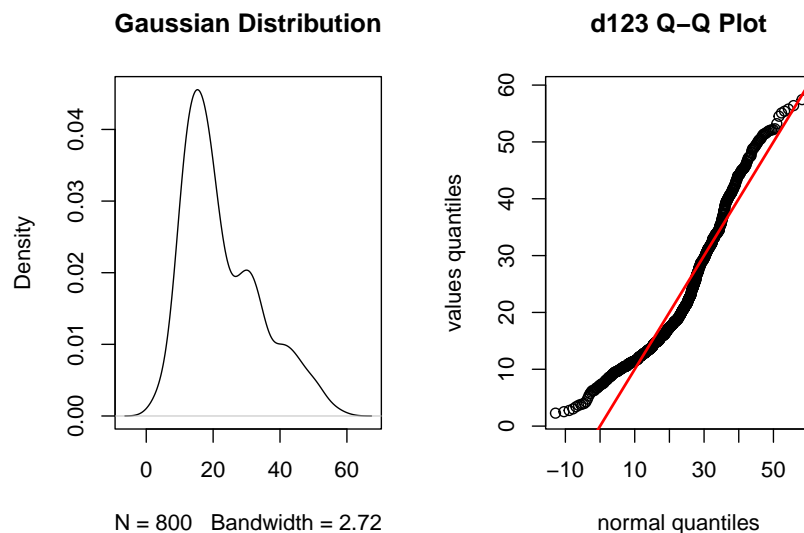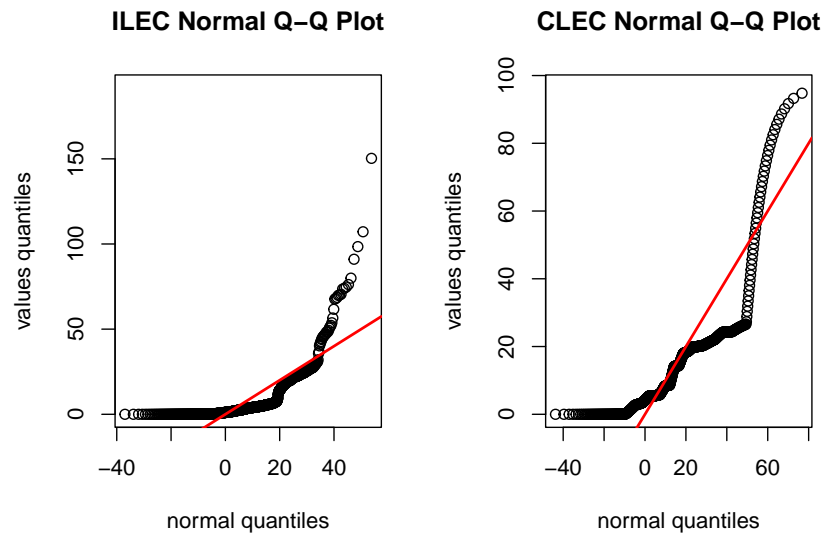

As can be seen in the Q-Q plot displayed, we can determined that the distribution is light tailed, meaning that the data is skewed, in this case it is skewed to the right.

(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?

```
layout(matrix(c(1,2), 1, 2, byrow = TRUE))
plot(density(ilec), main = "ILEC Plot")
plot(density(clec), main = "CLEC Plot")
```

ILEC Plot

CLEC Plot

Density (ILEC): y-axis 0.00, 0.04, 0.08, 0.12; x-axis 0, 50, 100, 150, 200

N = 1664   Bandwidth = 0.9676

Density (CLEC): y-axis 0.000, 0.010, 0.020, 0.030; x-axis −20, 20, 60, 100

N = 23   Bandwidth = 5.485

```r
layout(matrix(c(1,2), 1, 2, byrow = TRUE))
norm_qq_plot(ilec, main = "ILEC Normal Q-Q Plot")
norm_qq_plot(clec, main = "CLEC Normal Q-Q Plot")
```

ILEC Normal Q-Q Plot

CLEC Normal Q-Q Plot

values quantiles (ILEC): 0, 50, 100, 150; normal quantiles: −40, 0, 20, 40

values quantiles (CLEC): 0, 20, 40, 60, 80, 100; normal quantiles: −40, 0, 20, 60

As both plots do not follow the red line, we can conclude that both ILEC and CLEC samples are not normally distributed.