HW 6

110077443

3/21/2022

Credit given to 109077424 for reminding me to use the smallest W statistic in question 3, affecting the entire question's conclusion

Please note that all code in this document is presented in a grey box and the output reflected below each box

- The below code allows lengthy lines of code to display neatly within the grey box (wrapping it)

```
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
```

1 Importing and reshaping Verizon dataset

```
verizon_wide <- read.csv("verizon_wide.csv", header = TRUE)  # Import
str(verizon_wide)  # Checking structure of data
```

```
## 'data.frame':    1664 obs. of  2 variables:
##  $ ILEC: num  17.5 2.4 0 0.65 22.23 ...
##  $ CLEC: num  26.62 8.6 0 21.15 8.33 ...
```

a) Choosing a reshaping package for data:

Source: https://jtr13.github.io/spring19/hx2259__qz2351.html

- I chose **reshape2** because it provides the melt function which works with arrays, matrices and dataframes, while the gather function as part of the dplyr package only works with dataframes.
- Thus, i picked reshape2 because it allows flexibility with different data types.

b) Reshaping the verizon data to long format

```
require(reshape2)  # Getting package for melting data to long format
```

```
## Loading required package: reshape2
```

```r
verizon_long <- melt(verizon_wide, na.rm = TRUE, variable.name = "customers",
    value.name = "response_times")  # Converting to long format
```

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

```r
str(verizon_long)
```

```
## 'data.frame':    1687 obs. of  2 variables:
##  $ customers     : Factor w/ 2 levels "ILEC","CLEC": 1 1 1 1 1 1 1 1 1 1 ...
##  $ response_times: num  17.5 2.4 0 0.65 22.23 ...
```

```r
defaultW <- getOption("warn")
options(warn = -1)
options(warn = defaultW)
```

c) “head” and “tail” of the data to show that the reshaping worked

```r
head(verizon_long, 5)  # Checking top 5 observations of dataframe
```

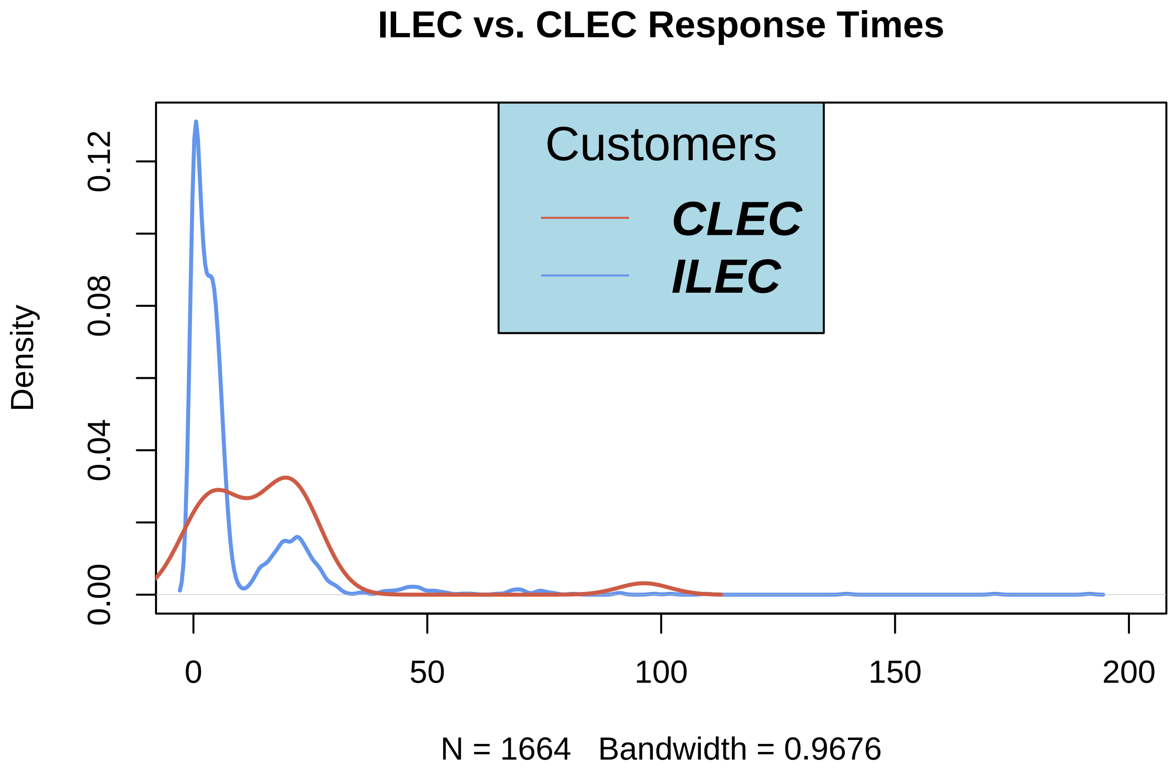
```
##   customers response_times
## 1      ILEC          17.50
## 2      ILEC           2.40
## 3      ILEC           0.00
## 4      ILEC           0.65
## 5      ILEC          22.23
```

```r
tail(verizon_long, 5)  # Checking bottom 5 observations of dataframe
```

```
##      customers response_times
## 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
# Split data in groupings before visualizing
v_customers <- split(x = verizon_long, f = verizon_long$customers)  # Split data

# Visualizing with custom plot function
verizon_plot <- function() {
    plot(density(v_customers$ILEC$response_times), col = "cornflowerblue",
        lwd = 2, xlim = c(0, 200), main = "ILEC vs. CLEC Response Times")  #ILEC
    lines(density(v_customers$CLEC$response_times), col = "coral3",
        lwd = 2)
}

# Adding legend
verizon_plot()
legend(x = "top", legend = c("CLEC", "ILEC"), col = c("coral3",
    "cornflowerblue"), lty = 1, cex = 1.5, title = "Customers",
    text.font = 4, bg = "lightblue")
```



2) Testing if the mean of response times for CLEC customers is greater than for ILEC customers

```
# Computing means and variance before running tests
ILEC_mean <- mean(v_customers$ILEC$response_times)  # 8.411611
ILEC_mean  # Print
```

```
## [1] 8.411611
```

```
ILEC_var <- var(v_customers$ILEC$response_times)  # 215.7973
ILEC_var  # Print
```

```
## [1] 215.7973
```

```
CLEC_mean <- mean(v_customers$CLEC$response_times)  # 6.50913
CLEC_mean  # Print
```

```
## [1] 16.50913
```

```
CLEC_var <- var(v_customers$CLEC$response_times)  # 380.3895
CLEC_var  # Print
```

```
## [1] 380.3895
```

a) Null and Alternative hypotheses (one-tailed)

- H0: mu{CLEC} <= mu{ILEC}
- Ha: mu{CLEC} > mu{ILEC}

b) Using 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

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

```
##
##  Two Sample t-test
##
## data:  v_customers$CLEC$response_times and v_customers$ILEC$response_times
## 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
```

- We do **not reject** the null hypothesis because the p-value < 0.01

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

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

```
##
##  Welch Two Sample t-test
##
## data:  v_customers$CLEC$response_times and v_customers$ILEC$response_times
## 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
```

- We do **reject** the null hypothesis because the p-value > 0.01
- Thus, CLEC's customer response times seems to be greater based on the more robust Welch test

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

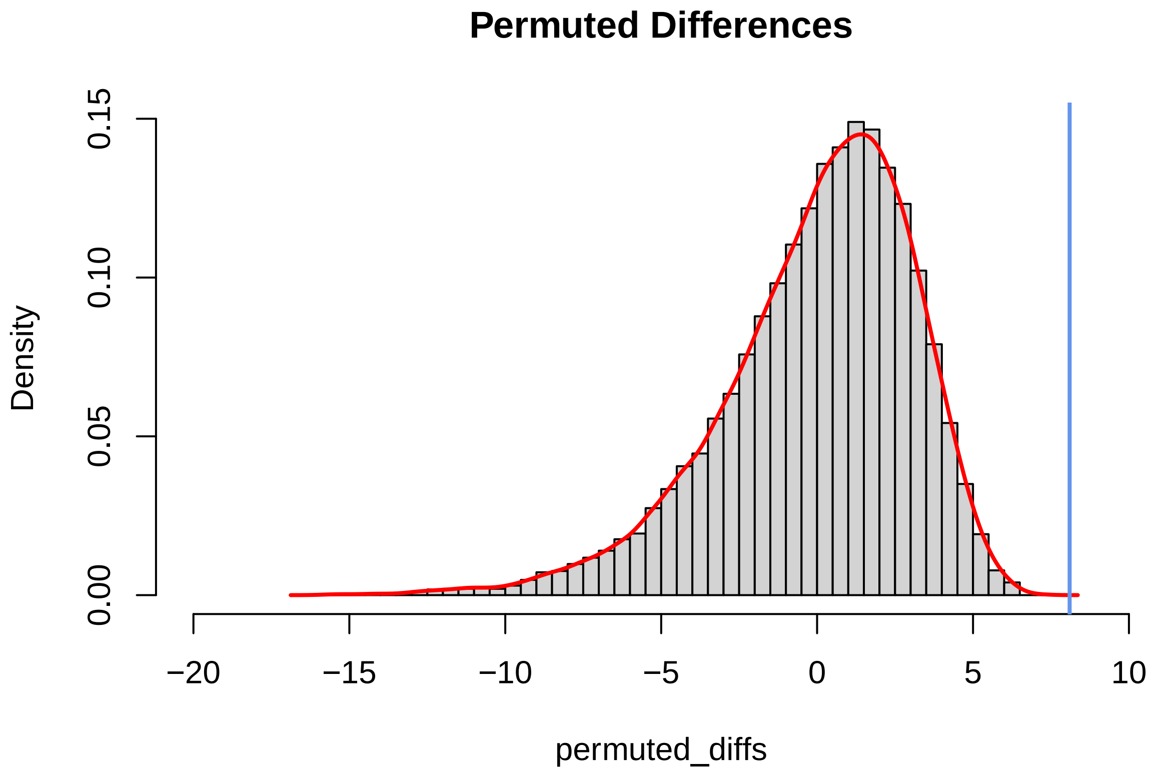
```
# Set seed
set.seed(1990)

# Permute differences function
permute_diff <- function(values, groups) {
    permuted <- sample(values, replace = FALSE)
    grouped <- split(permuted, groups)
    permuted_diff <- mean(grouped[[1]]) - mean(grouped[[2]])
}

nperms <- 10000  # Number of permutations
permuted_diffs <- replicate(nperms, permute_diff(verizon_long$response_times,
    verizon_long$customers))
```

i) Visualize the distribution of permuted differences and indicate the observed difference

```
# Observed Differences
observed_diff <- CLEC_mean - ILEC_mean  # 8.09752
observed_diff  # Print
```

```
## [1] 8.09752
```

```
# Visualize
hist(permuted_diffs, breaks = "fd", probability = TRUE, xlim = c(-20,
    10), main = "Permuted Differences")
lines(density(permuted_diffs), col = "red", lwd = 2)
abline(v = observed_diff, col = "cornflowerblue", lwd = 2)
```



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

```
p_1tailed <- sum(permuted_diffs > observed_diff)/nperms  # 0.9996305
p_1tailed  # Print
```

```
## [1] 0
```

```
p_2tailed <- sum(abs(permuted_diffs) > observed_diff)/nperms  # 0.0175
p_2tailed  # Print
```

```
## [1] 0.0175
```

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

- p-value is 0 –> mean is never larger than the observed difference in 10000 permutations.
- I would therefore **reject** the null hypothesis at 1% significance in one-tailed test based on the fact that the observed mean difference is too extreme with a p-value of 0%.

3) Using the Wilcoxon test to see if the response times for CLEC are different than ILEC

a) Computing the W statistic comparing the values using the vectorized approach.

```
# Creating function
gt_eq <- function(a, b) {
    ifelse(a > b, 1, 0) + ifelse(a == b, 0.5, 0)
}

# Checking similarity
n1 <- length(v_customers$CLEC$response_times)  #23
n1  # Print
```

```
## [1] 23
```

```
n2 <- length(v_customers$ILEC$response_times)  # 1664
n2  # Print
```

```
## [1] 1664
```

```
# If all the values are equal
(23 * 1664)/2  # 19136
```

```
## [1] 19136
```

```
# W statistic
w_stat <- sum(outer(v_customers$ILEC$response_times, v_customers$CLEC$response_times,
    FUN = gt_eq))  #11452
w_stat  # Print
```

```
## [1] 11452
```

- W (w_stat) is **11452 < 19136**, indicating that ILES response times are smaller than CLEC's, and that their times are not similar overall.

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

```
wilcox_p_1tail <- 1 - pwilcox(w_stat, n1, n2)  # 0.9996305
wilcox_p_1tail  # Print
```

```
## [1] 0.9996305
```

c) Running the Wilcoxon Test again using the wilcox.test() function

```r
# Run test and exchange variables to report the smaller of
# the two W Stat.
wilcox.test(v_customers$CLEC$response_times, v_customers$ILEC$response_times,
    alternative = "greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  v_customers$CLEC$response_times and v_customers$ILEC$response_times
## W = 26820, p-value = 0.0004565
## alternative hypothesis: true location shift is greater than 0
```

```r
wilcox.test(v_customers$ILEC$response_times, v_customers$CLEC$response_times,
    alternative = "greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  v_customers$ILEC$response_times and v_customers$CLEC$response_times
## W = 11452, p-value = 0.9995
## alternative hypothesis: true location shift is greater than 0
```

```r
# Double checking
wilcox.test(response_times ~ customers, data = verizon_long,
    alternative = "greater")  #11452
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  response_times by customers
## W = 11452, p-value = 0.9995
## alternative hypothesis: true location shift is greater than 0
```

- The results match (a) in that W (w_stat) is 11452

d) At 1% significance, and one-tailed, would you reject the null hypothesis that the values of CLEC and ILEC are different from one another?

- We **do not reject** the null hypothesis since the p-value at **0.9995** > significance level of 0.01

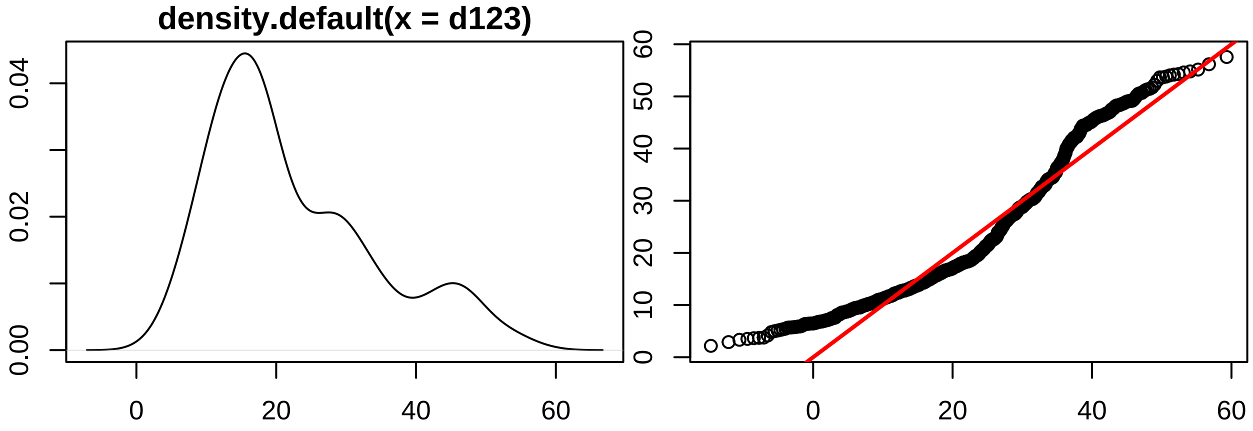
4) Visualizing whether a sample of data is normally distributed

a) Following steps (i) to (v) to create a function to see how a distribution of values compares to a perfectly normal distribution

```r
# Creating function with commented steps mentioned
norm_qq_plot <- function(values) {
    probs1000 <- seq(0, 1, 0.001)  # Step (i)
    q_vals <- quantile(values, probs1000)  ## Step (ii)
    q_norm <- qnorm(probs1000, mean = mean(values), sd = sd(values))  # Step (iii)
    plot(q_norm, q_vals, xlab = "normal quantiles", ylab = "value
       quantiles")  # Step (iV)
    abline(a = 0, b = 1, col = "red", lwd = 2)  # Step (v)
}
```

b) Confirming the function works by running it against the values of our d123 distribution from week 3

```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)

par(mfrow = c(2, 2), mar = c(2, 2, 1.5, 0.1))
plot(density(d123))  # Plot
norm_qq_plot(d123)  # Plot
```

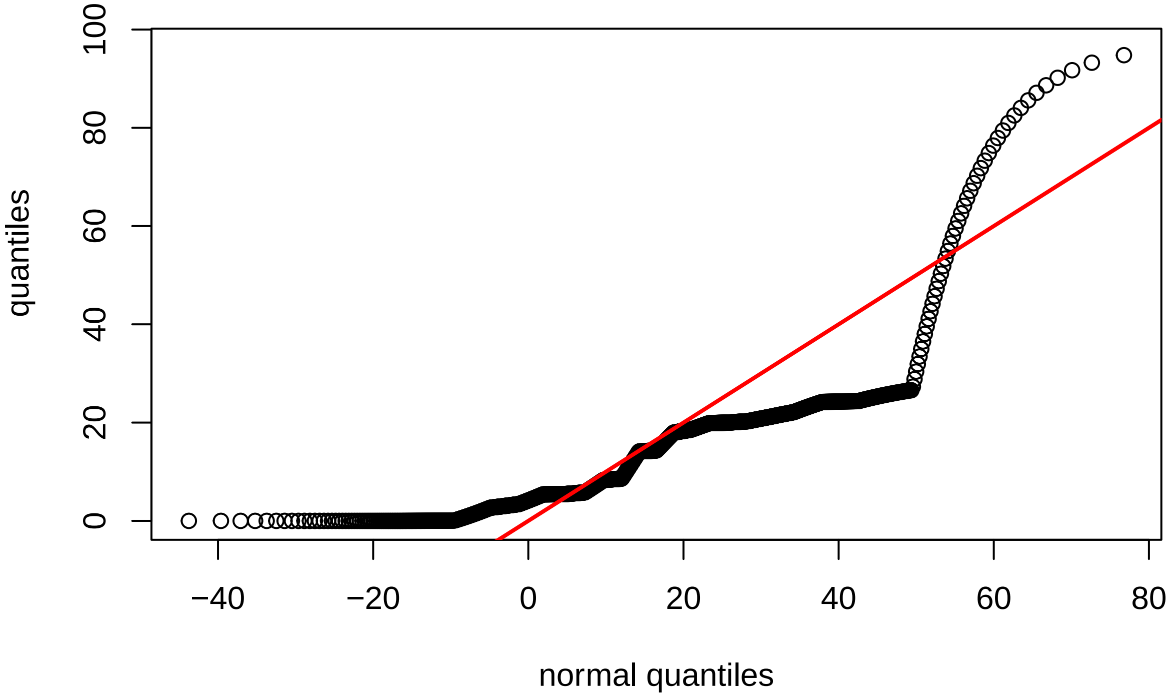


Interpreting the plot:

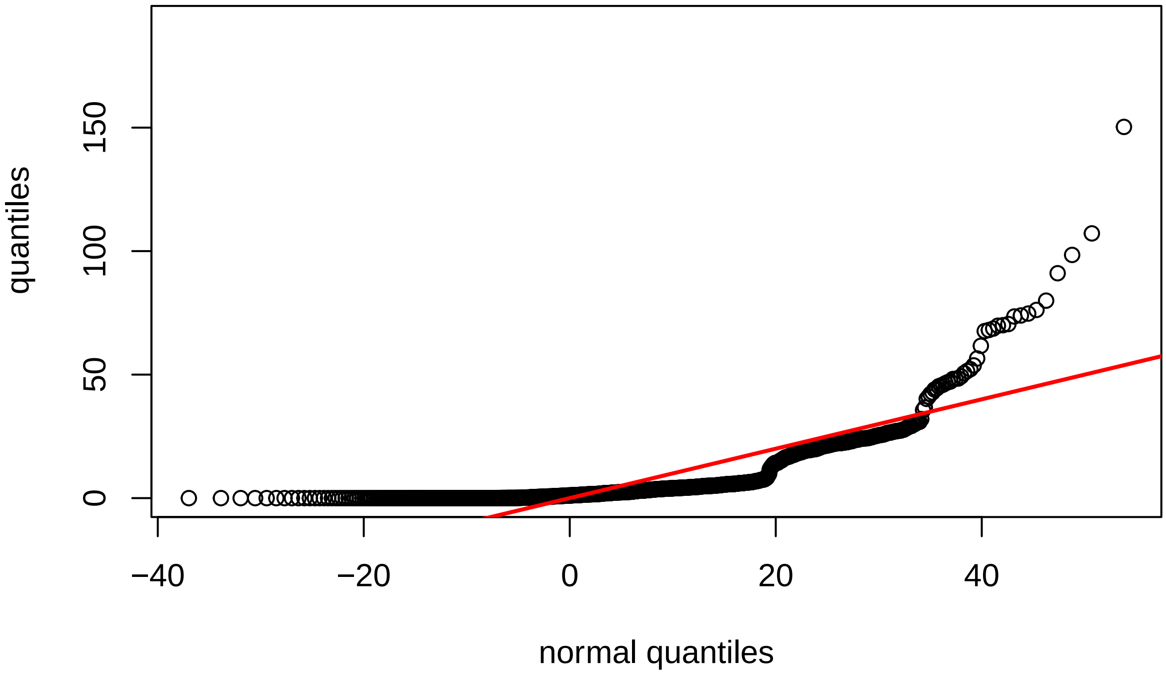
- We can see that the data is skewed right and not normally distributed since the qunatiles deviate from the red line, which is what the quantiles of the dataset would theoretically be if the dataset was normally distributed.

c) Using the normal QQ plot function to check if the values from each of the CLEC and ILEC samples are normally distrubuted

```
norm_qq_plot(v_customers$CLEC$response_times)  # CLEC samples
```



```
norm_qq_plot(v_customers$ILEC$response_times)  # ILEC samples
```



Conclusion:

- We can conclude from the plots that the CLEC and ILEC samples are **not normally distributed** since the distribution points for both samples are not linear.