hw6

111078513

2023-03-26

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.

```
verizon_data <- read.csv("verizon_wide.csv", header = TRUE, sep = ",")
```

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 chose tydir() because it is a newer package compared to reshape() and contains more methods. Additionally, according to the information provided on the website, the computational efficiency of the tydir() package is higher than that of reshape(). Therefore, I decided to use tydir().

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

```
# install.packages("tidyr")
library(tidyr)
loads_long <- gather(verizon_data, na.rm = TRUE,
                     key = "labels", value = "time")
```

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

```
head(loads_long, 5)
```

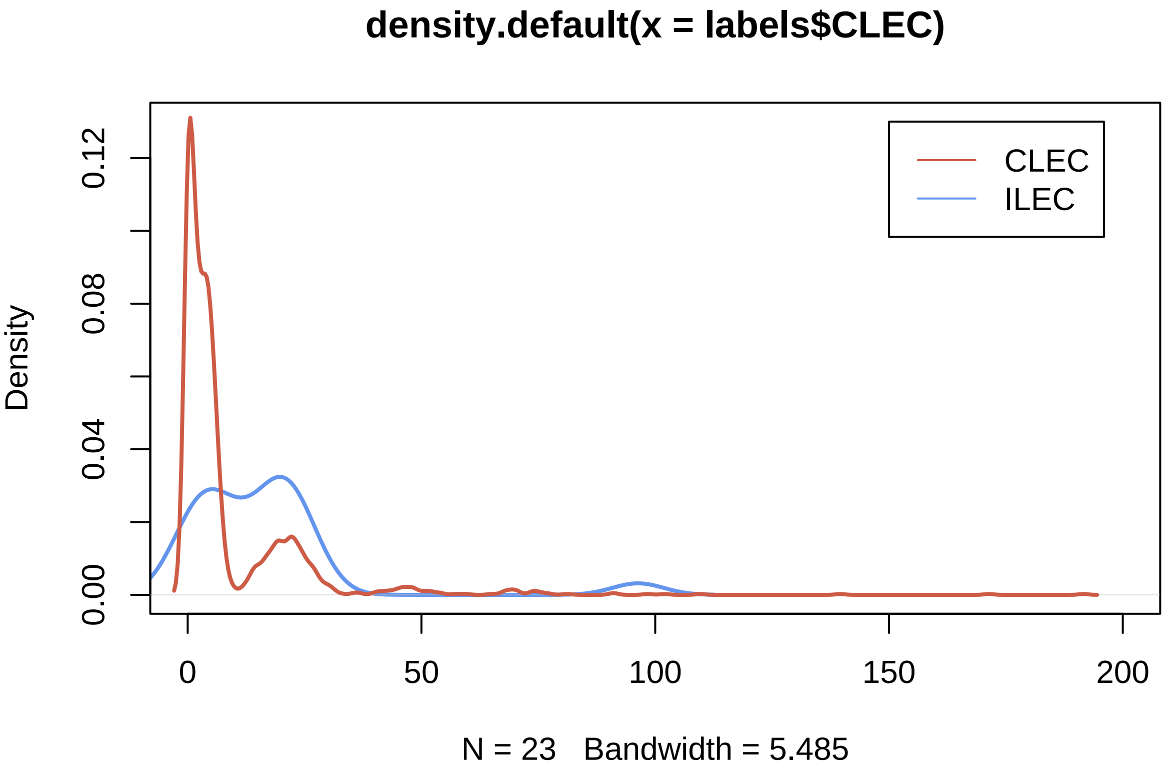
```
##   labels  time
## 1   ILEC 17.50
## 2   ILEC  2.40
## 3   ILEC  0.00
## 4   ILEC  0.65
## 5   ILEC 22.23
```

```
tail(loads_long, 5)
```

```
##      labels  time
## 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

```
labels <- split(x = loads_long$time, f = loads_long$labels)
plot(density(labels$CLEC), col="cornflowerblue", lwd=2, xlim=c(0, 200), ylim = c(0, 0.13))
lines(density(labels$ILEC), col="coral3", lwd=2)
legend(150,0.13 , lty=1, c("CLEC", "ILEC"), col=c("coral3", "cornflowerblue"))
```



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)

- H0: CLEC <= ILEC
- Ha: CLEC > 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.

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

```
t.test(labels$CLEC, labels$ILEC, alt="greater", var.equal=TRUE)
```

```
##
##  Two Sample t-test
```

```
##
## data:  labels$CLEC and labels$ILEC
## t = 2.6125, df = 1685, p-value = 0.004534
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.996491      Inf
## sample estimates:
## mean of x mean of y
## 16.509130  8.411611
```

According to the p-value obtained in the test, which is less than 0.01, we suggest rejecting H0. This means that the mean response time of ILEC is not significantly greater than that of CLEC.

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

```
t.test(labels$CLEC, labels$ILEC, alt="greater", var.equal=FALSE)
```

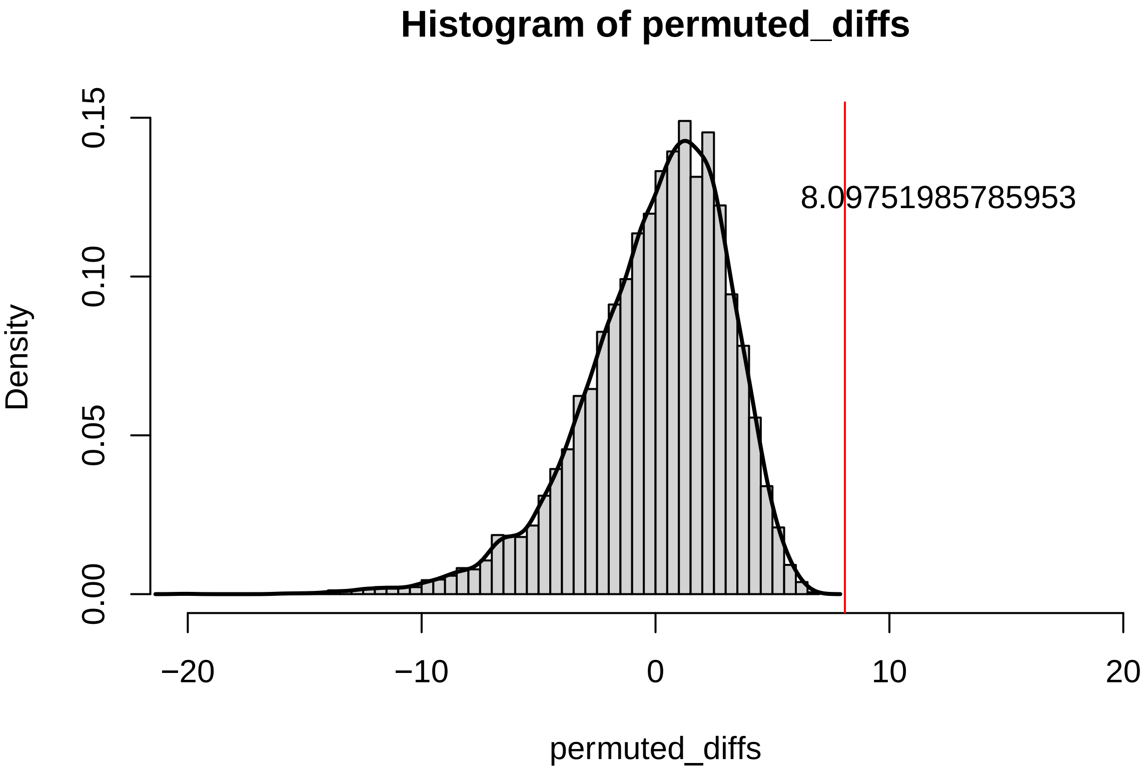
```
##
##  Welch Two Sample t-test
##
## data:  labels$CLEC and labels$ILEC
## t = 1.9834, df = 22.346, p-value = 0.02987
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.091721      Inf
## sample estimates:
## mean of x mean of y
## 16.509130  8.411611
```

According to the p-value obtained in the test, which is greater than 0.01, we suggest not rejecting H0. The evidence here is insufficient for me to demonstrate that the mean response time of ILEC has significantly less than that of CLEC.

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.

```
observed_diff <- mean(labels$CLEC) - mean(labels$ILEC)
permute_diff <- function(values, groups) {
  permuted <- sample(values, replace = FALSE)
  grouped <- split(permuted, groups)
  permuted_diff <- mean(grouped$ILEC) - mean(grouped$CLEC)
}
nperms <- 10000
permuted_diffs <- replicate(nperms, permute_diff(loads_long$time, loads_long$labels))
hist(permuted_diffs, breaks = "fd", probability = TRUE, xlim = c(-20, 20))
lines(density(permuted_diffs), lwd=2)
abline(v = observed_diff, col = "red")
text(observed_diff+4, 0.125, observed_diff)
```



(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
```

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

```
## [1] 0.0184
```

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

The probability of permuted_diffs being greater than observed_diff is 0, which is much smaller than the set p-value of 0.01. Therefore, we reject H0, which means that there is no significant difference between the mean values of ILEC and CLEC in the Permutation 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.

```
gt_eq <- function(a, b) {
  ifelse(a > b, 1, 0) + ifelse(a == b, 0.5, 0)
}
W <- sum(outer(labels$CLEC, labels$ILEC, FUN = gt_eq)); W
```

```
## [1] 26820
```

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

```
n1 <- length(labels$CLEC)
n2 <- length(labels$ILEC)
wilcox_p_1tail <- 1 - pwilcox(W, n1, n2) ;wilcox_p_1tail
```

```
## [1] 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.

```
wilcox.test(labels$CLEC, labels$ILEC, alternative = "greater")
```

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

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.

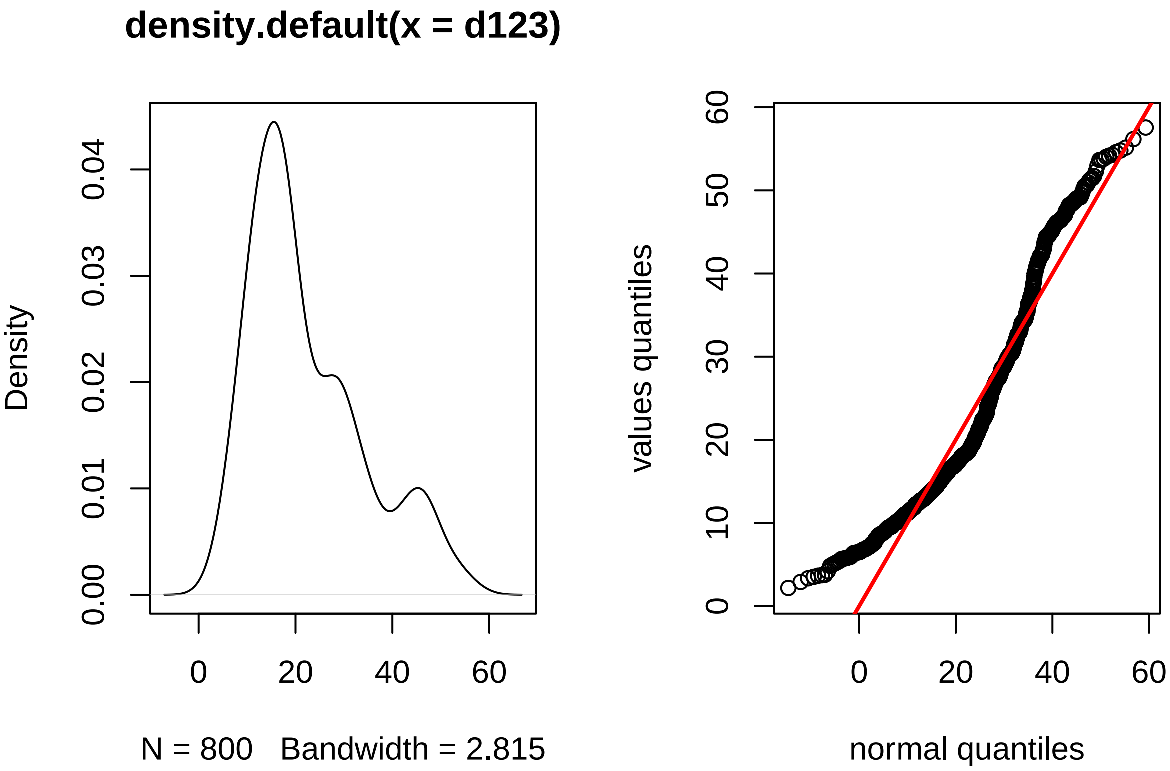
Based on the hypothesis test above, I would reject the null hypothesis (H0) and therefore not reject that CLEC is greater than ILEC.

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.

```
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=probs1000),
  #      and name it q_vals
  q_vals <- quantile(values, prob = probs1000)
  # (iii) 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))
  # (iv) 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")
  # (v) 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)
}
```

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:

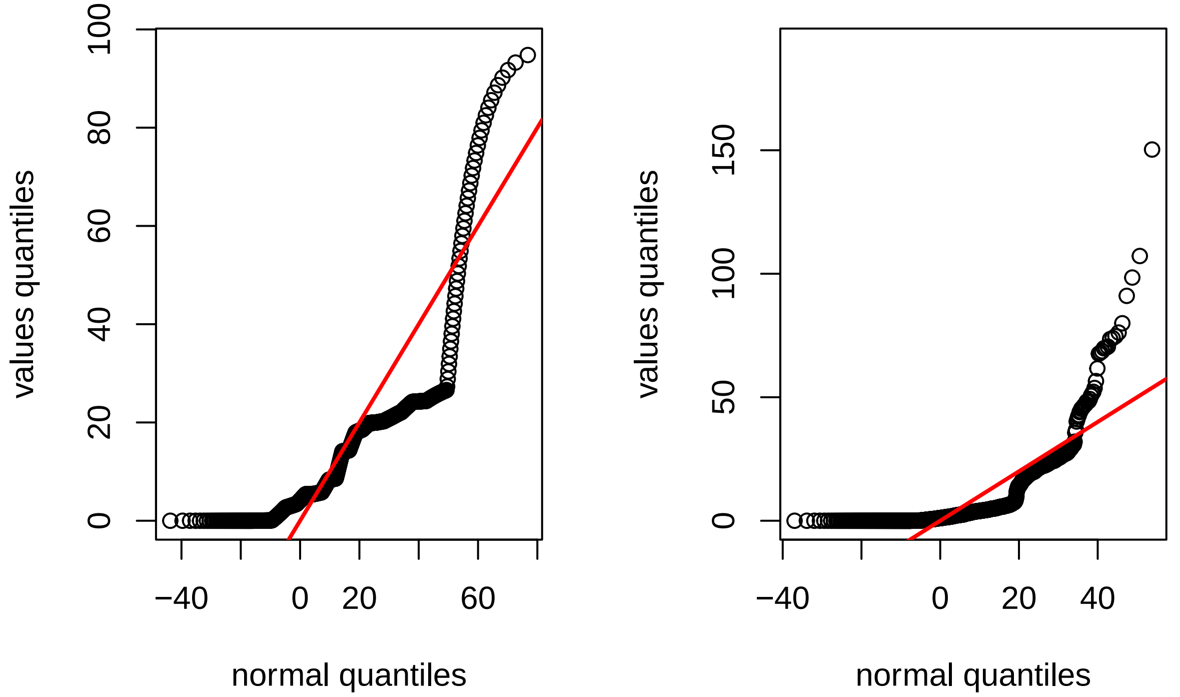
```
set.seed(978234)
d1 <- rnorm(n=500, mean=15, sd=5)
d2 <- rnorm(n=200, mean=30, sd=5)
d3 <- rnorm(n=100, mean=45, sd=5)
d123 <- c(d1, d2, d3)

par(mfrow = c(1, 2))
plot(density(d123))
norm_qq_plot(d123)
```



If a qq-plot conforms to a normal distribution, the result should be close to the diagonal line with an intercept of 0 and a slope of 1. Even if the distribution is normal but with differences in kurtosis and skewness, the plotted graph may not fit the diagonal line perfectly, but should still show symmetry at the left and right ends of the graph.
However, the qqplot of the d123 dataset winds around the diagonal line with an intercept of 0 and a slope of 1 like a caterpillar, indicating that the dataset does not follow a normal distribution.

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?

```
par(mfrow = c(1, 2))
norm_qq_plot(labels$CLEC)
norm_qq_plot(labels$ILEC)
```



By comparing the two plots, it can be observed that the data in both datasets are concentrated below 50. Moreover, according to the explanation of the qqplot, the plots for CLEC and ILEC do not conform to the expected state of a Normal Distribution.