bacs\_hw5

110071010

2024-03-18

# Backstory

Verizon was an Incumbent Local Exchange Carrier (ILEC), responsible for maintaining land-line phone service in certain areas. Other competing providers, termed Competitive Local Exchange Carriers (CLEC), could also sell long-distance phone services in Verizon’s areas. When something went wrong, Verizon would be responsible to respond and repair services as quickly for CLEC long-distance customers as for its own ILEC customers. The New York Public Utilities Commission (PUC) monitored fairness by comparing Verizon’s response times for its ILEC customers versus CLEC customers. In each case, a hypothesis test was performed at the 1% significance level, to determine whether response times for CLEC customers were significantly slower than for Verizon’s customers. If Verizon failed to provide fair treatment for CLEC customers, it would have to pay large penalties.

Verizon claims that mean response time for ILEC and CLEC customers are the same, but the PUC would like to test if CLEC customers were facing greater response times.

# Question 1

The Verizon dataset this week is provided as a “wide” data frame. Let’s practice reshaping it to a “long” data frame

## Getting data ready

# Reading in the file  
response\_wide <- read.csv("verizon\_wide.csv")  
View(response\_wide)

## 1a

***Instruction***

Pick a reshaping package – research them online and tell us why you picked it over others.

***Answer***

I chose tidyr over reshape2 for several reasons. For one thing, tidyr is more intuitive and integrated with tidyverse. For another, while reshape2 is powerful for more complex reshaping tasks and works with a wider range of data types including matrices and arrays, this homework does not require such heavy data manipulation.

## 1b

***Instruction***

Show the code to reshape the verizon\_wide.csv sample

# Reshaping the response\_wide sample into response\_long  
library(tidyr)  
response\_long <- gather(response\_wide, na.rm = TRUE, key = "host", value = "response\_time")

## 1c

***Instruction***

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

head(response\_long)

## host response\_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

tail(response\_long)

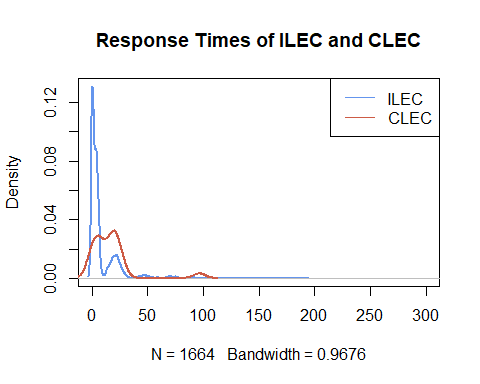
## host response\_time  
## 1682 CLEC 24.20  
## 1683 CLEC 22.13  
## 1684 CLEC 18.57  
## 1685 CLEC 20.00  
## 1686 CLEC 14.13  
## 1687 CLEC 5.80

## 1d

***Instruction***

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

hosts <- split(x = response\_long$response\_time, f = response\_long$host)  
  
plot(density(hosts$ILEC), col = "cornflowerblue", lwd = 2, xlim= c(0,300), main = "Response Times of ILEC and CLEC")  
lines(density(hosts$CLEC), col = "coral3", lwd = 2)  
legend("topright", lty = 1, c("ILEC","CLEC"), col = c("cornflowerblue","coral3"))



# Question 2

Let’s test if the mean of response times for CLEC customers is greater than for ILEC customers

## 2a

***Instruction***

State the appropriate null and alternative hypotheses (one-tailed)

***Answer***

**Null Hypothesis H0:** Mean of CLEC customers’ response times is equal or less compared to that of ILEC customers’ response times

**Alternative Hypothesis H1:** Mean of CLEC customers’ response times is greater than that of ILEC customers’ response times

## 2b

***Instruction***

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(response\_wide$CLEC, response\_wide$ILEC, alt="greater", var.equal = TRUE, conf.level = 0.99)

## Two Sample t-test  
##   
## data: response\_wide$CLEC and response\_wide$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

### (ii)

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

t.test(response\_wide$CLEC, response\_wide$ILEC, alt="greater", var.equal = FALSE, conf.level = 0.99)

## Welch Two Sample t-test  
##   
## data: response\_wide$CLEC and response\_wide$ILEC  
## 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

## 2c

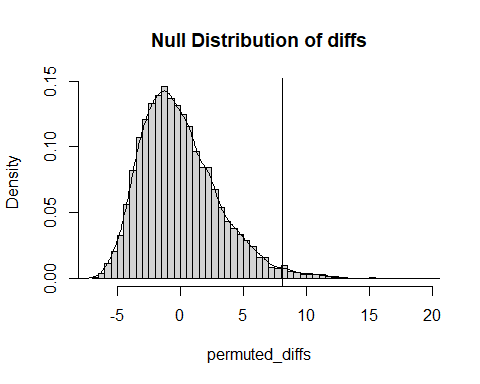
***Instruction***

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.

# Setting seed  
set.seed(123123)  
  
# Observed difference  
observed\_diff <- mean(hosts$CLEC) - mean(hosts$ILEC)  
  
  
# Getting the function ready for permutation  
permute\_diff <- function(values, groups){  
 permuted <- sample(values, replace = FALSE)   
 grouped <- split(permuted, groups)  
 permute\_diff <- mean(grouped$CLEC) - mean(grouped$ILEC)  
}  
  
# Visualize after 10000 permutations   
nperms = 10000  
permuted\_diffs <- replicate(nperms, permute\_diff(response\_long$response\_time, response\_long$host))  
hist(permuted\_diffs, breaks = "fd", probability = TRUE, main = "Null Distribution of diffs")  
lines(density(permuted\_diffs))  
  
# Displaying observed difference   
abline(v = observed\_diff)



### (ii)

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

# one-tailed p-value  
p\_1tailed <- sum(permuted\_diffs > observed\_diff)/ nperms  
p\_1tailed

## [1] 0.0172

# two-tailed p-value  
p\_2tailed <- sum(abs(permuted\_diffs) > observed\_diff)/ nperms  
p\_2tailed

## [1] 0.0172

### (iii)

***Question***

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

***Answer***

No, we would not reject H0 since p-value > 0.01

# Question 3

Let’s use the Wilcoxon test to see if the response times for CLEC are different than ILEC.

## 3a

***Instruction***

Compute the W statistic comparing the values. You may use either the permutation approach (try the functional form) or the rank sum approach

# Define gt\_eq() for comparing values from two groups  
gt\_eq <- function(a,b){  
 ifelse(a > b, 1, 0) + ifelse(a ==b, 0.5, 0)  
}  
  
# W statistic  
W <- sum(outer(hosts$CLEC , hosts$ILEC, FUN = gt\_eq))  
W

## [1] 26820

## 3b

***Instruction***

Compute the one-tailed p-value for W

# Lengths of the CLEC and ILEC groups to be used in the p-value calculation  
n1 <- length(hosts$CLEC)  
n2 <- length(hosts$ILEC)  
  
wilcox\_p\_1tail<- 1 - pwilcox(W,n1,n2)  
wilcox\_p\_2tail <- 2 \* wilcox\_p\_1tail  
wilcox\_p\_1tail

## [1] 0.0003688341

## 3c

***Instruction***

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

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

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

## 3d

***Question***

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

***Answer***

Yes, we would reject H0 since p-value < 0.01

# 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.

## 4a

***Instruction***

Make a function called norm\_qq\_plot() that takes a set of values

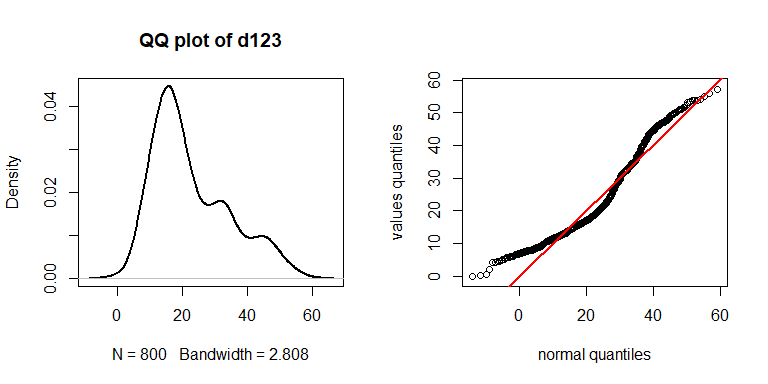
norm\_qq\_plot <- function(values) {  
 # (i) Create a sequence of probability numbers from 0 to 1, with ~1000 probabilities in between  
 probs1000 <- seq(0, 1, 0.001)  
 # (ii) Calculate ~1000 quantiles of our values  
 q\_vals <- quantile(values, probs = probs1000)  
 # (iii) Calculate ~1000 quantiles of a perfectly normal distribution with the same mean and standard deviation as our values  
 q\_norm <- qnorm(p = 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) 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) # a:intercept, b:slope  
}

## 4b

***Instruction***

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 given plot

set.seed(123123)  
  
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), main = "QQ plot of d123", lwd = 2)  
norm\_qq\_plot(d123)



***Comment***

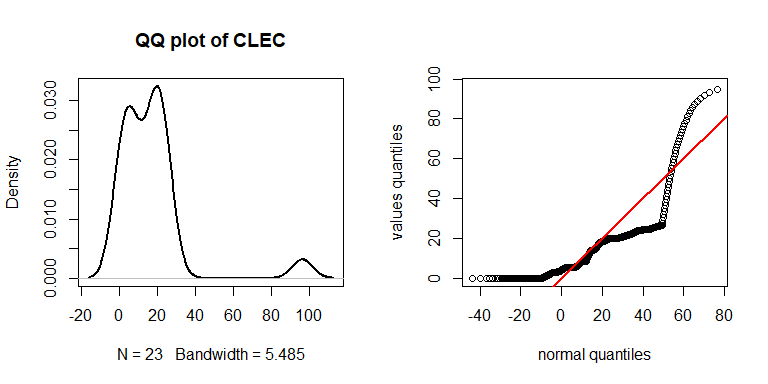
The QQplot of d123 suggests that it is not normally distributed. Rather, d123 appears to exhibit patterns of a bimodal distribution.

## 4c

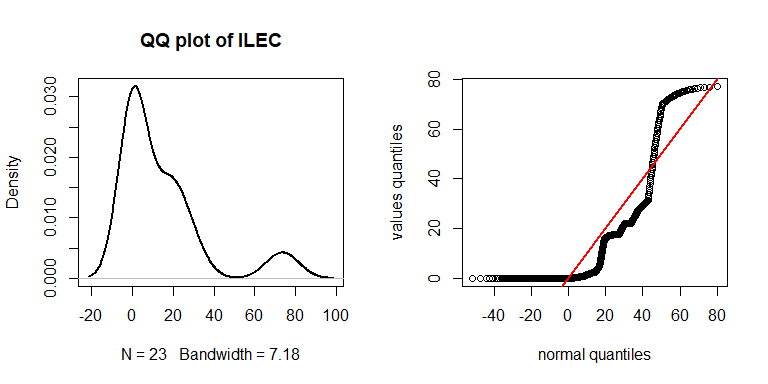
***Instruction***

Traditional statistical t-tests to compare the means of two populations require that the two populations are normally distributed. Use your normal Q-Q plot function, norm\_qq\_plot, 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?

# QQ-plotting for CLEC  
par(mfrow = c(1,2))  
response\_wide <- response\_wide[!is.na(response\_wide$CLEC), ]  
plot(density(response\_wide$CLEC), main = "QQ plot of CLEC", lwd = 2)  
norm\_qq\_plot(response\_wide$CLEC)



# QQ-plotting for ILEC  
par(mfrow = c(1,2))  
plot(density(response\_wide$ILEC), main = "QQ plot of ILEC", lwd = 2)  
norm\_qq\_plot(response\_wide$ILEC)



***Conclusion***

As observed in the graphs above, there are many unaligned data points in both CLEC and ILEC’s QQplots, so one could easily conclude that both are not normally distributed.