BACS HW - Week 6

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. You may use either shape (wide or long) for your analyses in later questions.

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).

Ans: I’ll pick the

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

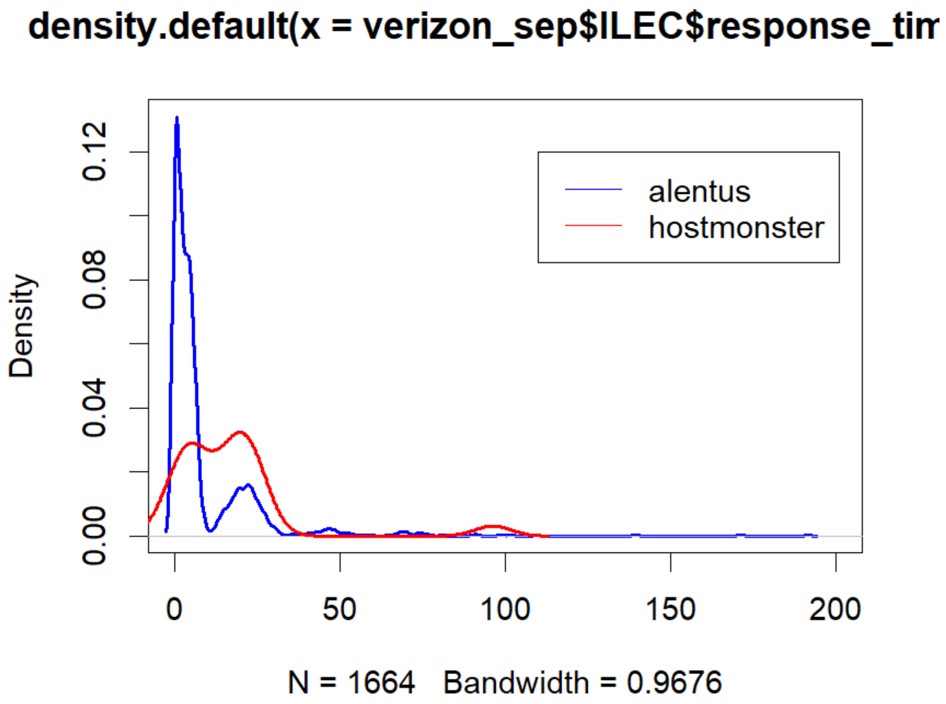
```
> #1-(b)
> #install.packages("reshape2")
> library(reshape2)
> verizon<-read.csv("verizon_wide.csv")
> verizon_long<-melt(verizon,na.rm=TRUE,variable.name =
"company",value.name ="response_time")
```

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

```
> head(verizon_long)
  company 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(verizon_long)
     company 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
```

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

```
> verizon_sep<-split(verizon_long,f=verizon_long$company)
>
plot(density(verizon_sep$ILEC$response_time),col="blue",lwd=2,xlim=c(0,200)
)
> lines(density(verizon_sep$CLEC$response_time),col="red",lwd=2)
>
legend(x=110,y=0.12,legend=c("alentus","hostmonster"),col=c("blue","red"),l
ty=1)
```



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)

null-hypothesis: the mean response time for ILEC and CLEC customers are the same

alternative hypothesis: the mean response time for ILEC and CLEC customers are not the same

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

```
> #2-(b)
>
t.test(verizon_sep$ILEC$response_time,verizon_sep$CLEC$response_time,alt="g
reater",var.equal=TRUE)

        Two Sample t-test

data:  verizon_sep$ILEC$response_time and verizon_sep$CLEC$response_time
t = -2.6125, df = 1685, p-value = 0.9955
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -13.19855       Inf
sample estimates:
mean of x mean of y
 8.411611 16.509130
```

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

```
>
t.test(verizon_sep$ILEC$response_time,verizon_sep$CLEC$response_time,alt="g
reater",var.equal=FALSE)

        Welch Two Sample t-test

data:  verizon_sep$ILEC$response_time and verizon_sep$CLEC$response_time
t = -1.9834, df = 22.346, p-value = 0.9701
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -15.10332       Inf
sample estimates:
mean of x mean of y
 8.411611 16.509130
```

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.

```
> #2-(c)-(i)
> observed_diff<-mean(verizon_sep$ILEC$response_time)-
mean(verizon_sep$CLEC$response_time)
> observed_diff
[1] -8.09752
> permute_diff<-function(value,group){
+   permuted<-sample(value,replace = FALSE)
+   grouped<-split(permuted,group)
+   permute_diff<-mean(grouped$ILEC)-mean(grouped$CLEC)
+ }
> permuted_diffs<-
replicate(10000,permute_diff(verizon_long$response_time,verizon_long$compan
y))
> hist(permuted_diffs,probability = TRUE,breaks = "fd",xlim=c(-20,10))
> lines(density(permuted_diffs),lwd=1.5)
> abline(v=observed_diff,col="blue",lwd=1.5)
```

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

```
> #2-(c)-(ii)
> p_onetail<-sum(permuted_diffs>observed_diff)/10000
> p_twotail<-sum(abs(permuted_diffs)>observed_diff)/10000
> p_onetail
[1] 0.9808
> p_twotail
[1] 1
```

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

Ans: No, both p-value of one-tail and two-tail are very close to 1, so we don’t have the power to reject H0.

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 (with either for-loops or the vectorized form) or the rank sum approach.

```
> #3-a
> time_ranks<-rank(verizon_long$response_time)
> ranked_groups<-split(time_ranks,verizon_long$company)
> U1<-sum(ranked_groups$CLEC)
> n1<-length(verizon_sep$CLEC)
> w<-U1-(n1*(n1+1))/2
> w
[1] 27093
```

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

```
> #3-b
> n2<-length(verizon_sep$ILEC)
> wilcox_p_1tail<-1-pwilcox(w,n1,n2)
> wilcox_p_2tail<-2*wilcox_p_1tail
> wilcox_p_1tail
[1] 0
> wilcox_p_2tail
[1] 0
```

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.

```
> #3-c
> wilcox.test(verizon$CLEC,verizon$ILEC,alternative = "greater")

        Wilcoxon rank sum test with continuity correction

data:  verizon$CLEC and verizon$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? ~~different from one another~~

Ans: Yes,I will reject the null hypothesis.(the p value is 0, which isvery small)

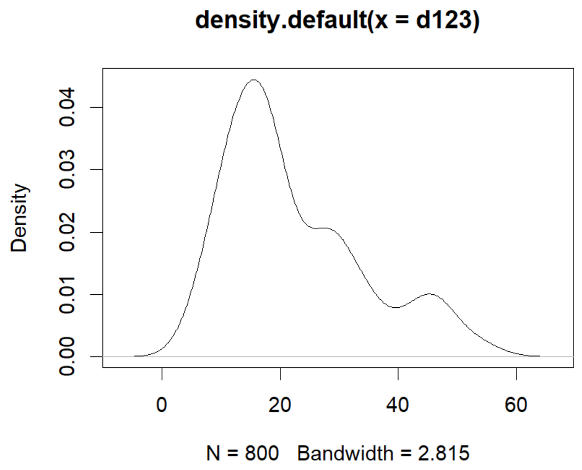
Question 4) One of the assumptions of some classical statistical tests is that our population data should be roughly normal. Let’s explore one way of visualizing whether a sample of data is normally distributed.

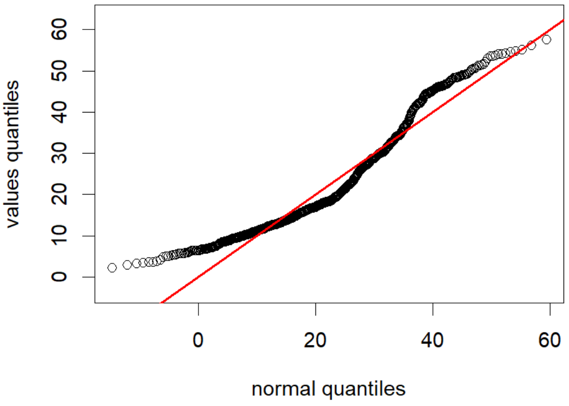
a. Follow the following steps to create a function to see how a distribution of values compares to a perfectly normal distribution. *The ellipses (...) in the steps below indicate where you should write your own code*.

```
> #4-a
> norm_qq_plot <- function(values){
+   probs1000 <- seq(0, 1, 0.001)
+   q_vals <- quantile(values,prob=probs1000)
+   q_norm <- qnorm(probs1000,mean(values),sd(values))
+   plot(q_norm, q_vals, xlab="normal quantiles", ylab="values
quantiles",ylim=c(min(values)-5,max(values)+5))
+   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:

```
> #4-b
> 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)
>
> plot(density(d123))
> norm_qq_plot(d123)
```

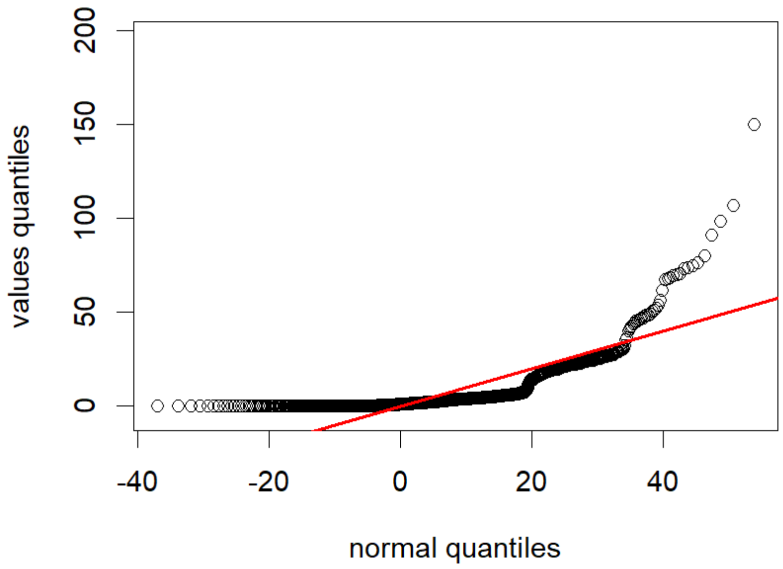




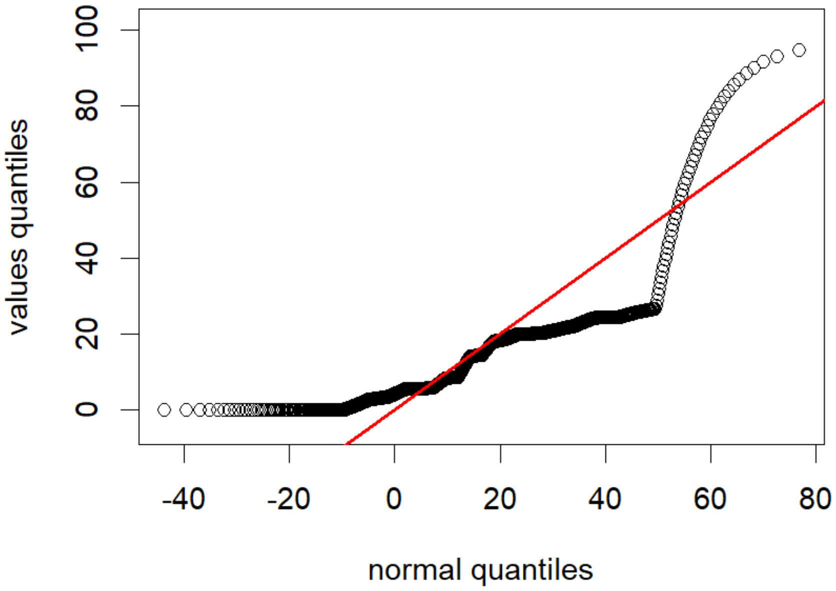
Ans: Yes, the plot above is quite similar to the provided plot.Also,from the right plot above, we can see that most points follows the pattern of normal distribution(quit close to the line), while there are still some points looks a little bit far from the line around zero.

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?

```
> #4-c
> plot(density(verizon$ILEC))
> norm_qq_plot(verizon$ILEC)
> plot(density(verizon$CLEC[!is.na(verizon$CLEC)]))
> norm_qq_plot(verizon$CLEC[!is.na(verizon$CLEC)])
```



Ans: This is the plot of ILEC, we can see that there are some points seriously far from the line around 40 normal quantiles, which indicates that the data might be skewed right.



Ans: This is the plot of CLEC, same as IELC, although the points in the middle looks a little bit different from IELC, we can still see that there are some points seriously far from the line around 40 normal quantiles, which indicates that the data might be skewed right.