BACS HW (Week 6)

108020024

due on 03/26 (Sun) Helped by 108020033

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).

Both reshape2 and **tidyr** did similar jobs, however **tidyr** is a package specifically designed for data tidying, and most importantly **tidyr** is a core package to Tidyverse series, a collection of R packages that share common principles and are designed to work together seamlessly, this makes **tidyr** more competitive against reshape2, so I choose **tidyr** as the reshaping package in this homework.

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

```
verizon_wide <- read.csv("verizon_wide.csv")
#install.packages("tidyr")
library(tidyr)
```

```
## Warning:   'tidyr'   R   4.2.3
```

```
verizon_long <- gather(verizon_wide, na.rm = TRUE,key = "host",value = "load_time")
```

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

```
head(verizon_long)
```

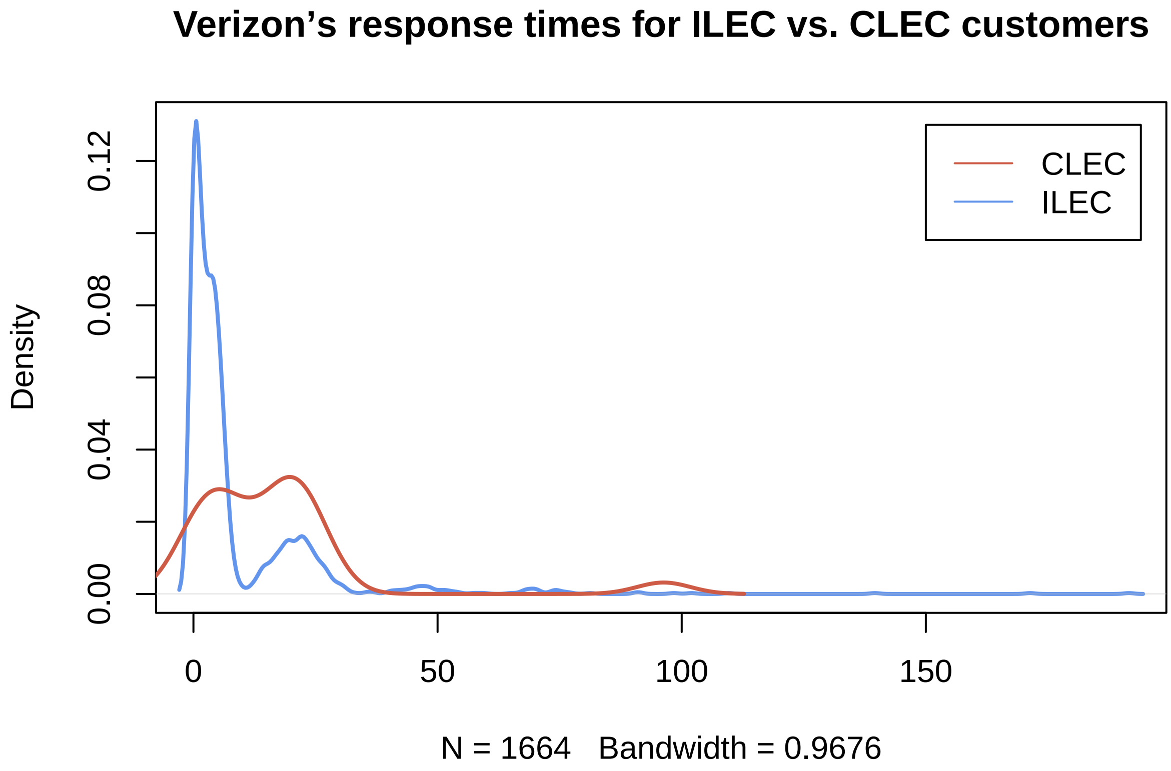
```
##   host load_time
## 1 ILEC     17.50
## 2 ILEC      2.40
## 3 ILEC      0.00
## 4 ILEC      0.65
## 5 ILEC     22.23
## 6 ILEC      1.20
```

```
tail(verizon_long)
```

```
##      host load_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

```
hosts <- split(x = verizon_long$load_time, f = verizon_long$host)
plot(density(hosts$ILEC),
     col="cornflowerblue",
     lwd=2, xlim=c(0,max(verizon_long$load_time)),
     main = "Verizon's response times for ILEC vs. CLEC customers")
lines(density(hosts$CLEC), 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)

H_0 : The mean of response times for CLEC customers $\leq$ the mean of response times for ILEC customers

H_1 : The mean of response times for CLEC customers $>$ the mean of response times for ILEC customers

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
```

```r
t.test(hosts$ILEC, hosts$CLEC ,alternative="less",conf.level = 0.99, var.equal = TRUE )
```

```
##
##  Two Sample t-test
##
## data:  hosts$ILEC and hosts$CLEC
## t = -2.6125, df = 1685, p-value = 0.004534
## alternative hypothesis: true difference in means is less than 0
## 99 percent confidence interval:
##        -Inf -0.8801387
## sample estimates:
## mean of x mean of y
##  8.411611 16.509130
```

The overall p-value $= 0.004534 < 0.01$, reject H_0, accept H_1 at 99% confidence.

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

```r
t.test(hosts$ILEC, hosts$CLEC ,alternative="less",conf.level = 0.99, var.equal = FALSE )
```

```
##
##  Welch Two Sample t-test
##
## data:  hosts$ILEC and hosts$CLEC
## t = -1.9834, df = 22.346, p-value = 0.02987
## alternative hypothesis: true difference in means is less than 0
## 99 percent confidence interval:
##      -Inf 2.130858
## sample estimates:
## mean of x mean of y
##  8.411611 16.509130
```

The overall p-value $= 0.02987 > 0.01$, cannot reject H_0 at 99% confidence.

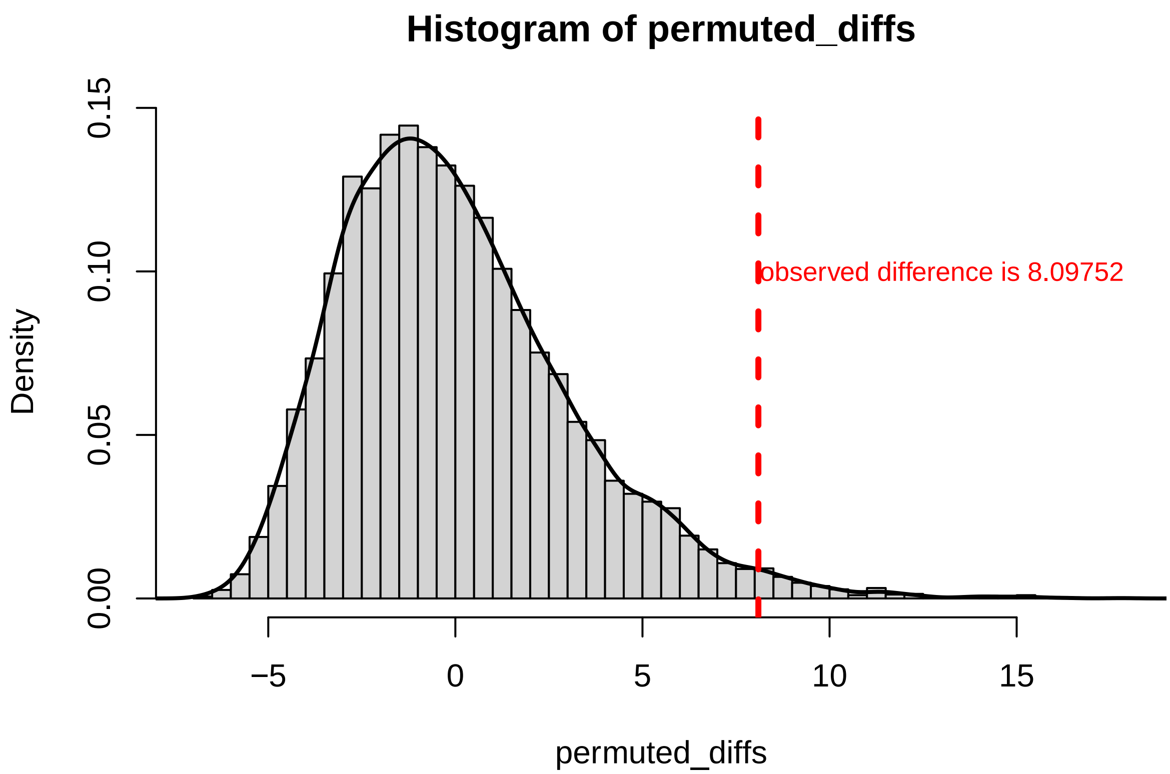
c) Use a permutation test to compare the means of ILEC vs. CLEC response times

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

```
observed_diff <- mean(hosts$CLEC) - mean(hosts$ILEC)

permute_diff <- function(values, groups) {
permuted <- sample(values, replace = FALSE)
grouped <- split(permuted, groups)
permuted_diff <- mean(grouped$CLEC) - mean(grouped$ILEC)
}
nperms <- 10000

permuted_diffs <- replicate(nperms,
                            permute_diff(verizon_long$load_time, verizon_long$host))
hist(permuted_diffs, breaks = "fd", probability = TRUE)
lines(density(permuted_diffs), lwd=2)
abline(v = observed_diff, col="red", lwd=3, lty=2)
text(13,0.1,labels="observed difference is 8.09752", cex=0.8, col="red")
```



The observed difference is 8.09752.

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

```r
p_1tailed <- sum(permuted_diffs > observed_diff) / nperms
p_2tailed <- sum(abs(permuted_diffs) > observed_diff) / nperms
```

The one-tailed p-value is 0.018.

The two-tailed p-value is 0.018.

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

The overall p-value = 0.018 > 0.01, cannot reject H_0 at 1% significance.

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.

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

The W statistic is 26820.

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

```r
n1 <- length(hosts$CLEC)
n2 <- length(hosts$ILEC)

wilcox_p_1tail <- 1 - pwilcox(W, n1, n2)
```

The one-tailed p-value for W is 3.6883415×10^{-4}.

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.

```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
```

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

The overall p-value = 0.0004565 < 0.01, reject the null hypothesis that the values of CLEC and ILEC are similar at 1% significance.

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.

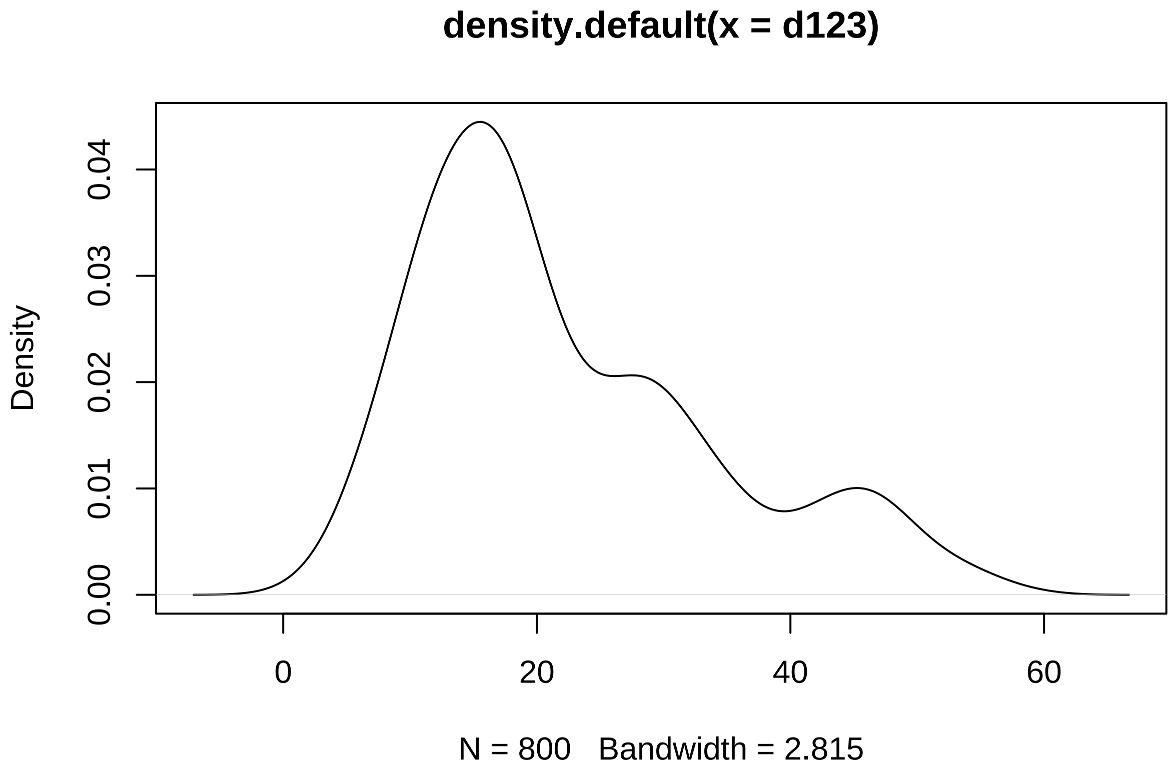
```
i) Create a sequence of probability numbers from 0 to 1, with ~1000 probabilities in
between
ii) Calculate ~1000 quantiles of our values (you can use probs=probs1000), and name it
q_vals
iii) calculate ~1000 quantiles of a perfectly normal distribution with the same mean
and standard deviation as our values; name this vector of normal quantiles q_norm
iv) Create a scatterplot comparing the quantiles of a normal distribution versus
quantiles of values
v) Finally, draw a red line with intercept of 0 and slope of 1, comparing these two
sets of quantiles
```

```r
norm_qq_plot <- function(values) {
  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",
       xlim=c(min(values),max(values)),
       ylim=c(min(values),max(values)),
       main = 'Q-Q plot')
  abline( a = 0, b = 1 , col="red", lwd=2)
}
```

b) Interpret the plot you produced (see this article on how to interpret normal Q-Q plots) and tell us if it suggests whether d123 is normally distributed or not.

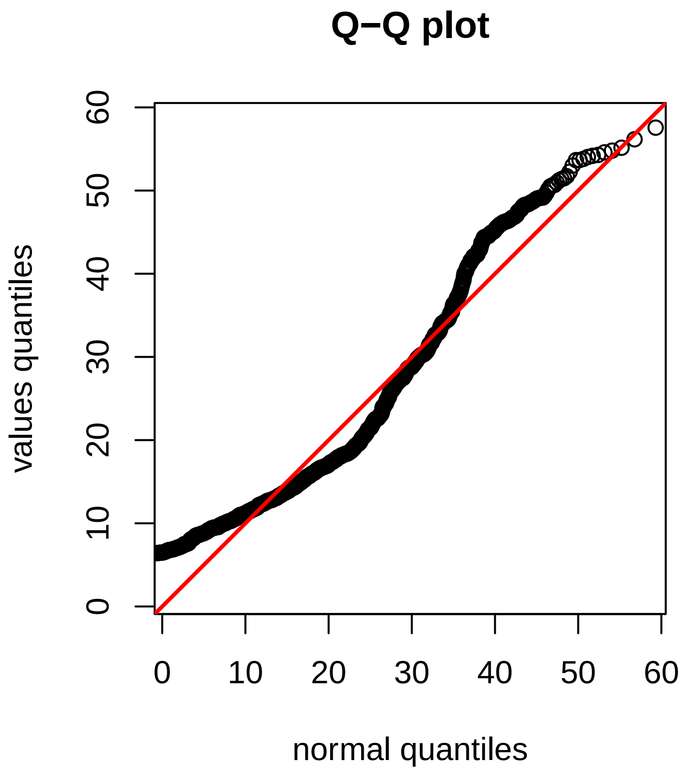
```r
set.seed(978234)
d1 <- rnorm(n=500, mean=15, sd=5)
d2 <- rnorm(n=200, mean=30, sd=5)
d3 <- rnorm(n=100, mean=45, sd=5)
d123 <- c(d1, d2, d3)

plot(density(d123))
```



```
par(pty="s")

norm_qq_plot(d123)
```



From the Q-Q plot we see the points stray from linearity, so d123 is not normally distributed. Check with shapiro normality test.

```
shapiro.test(d123)
```

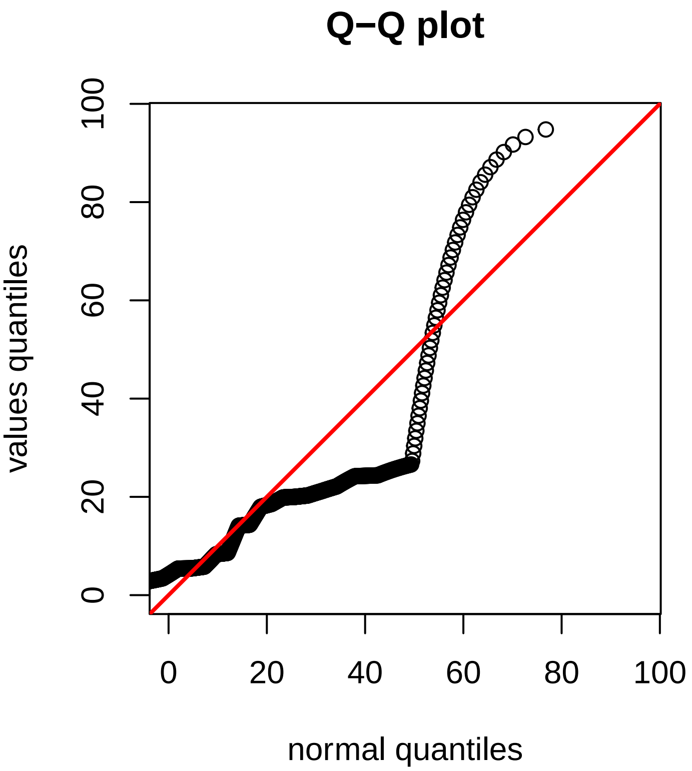
```
##
##  Shapiro-Wilk normality test
##
## data:  d123
## W = 0.92439, p-value < 2.2e-16
```

The result shows d123 is not normally distributed.

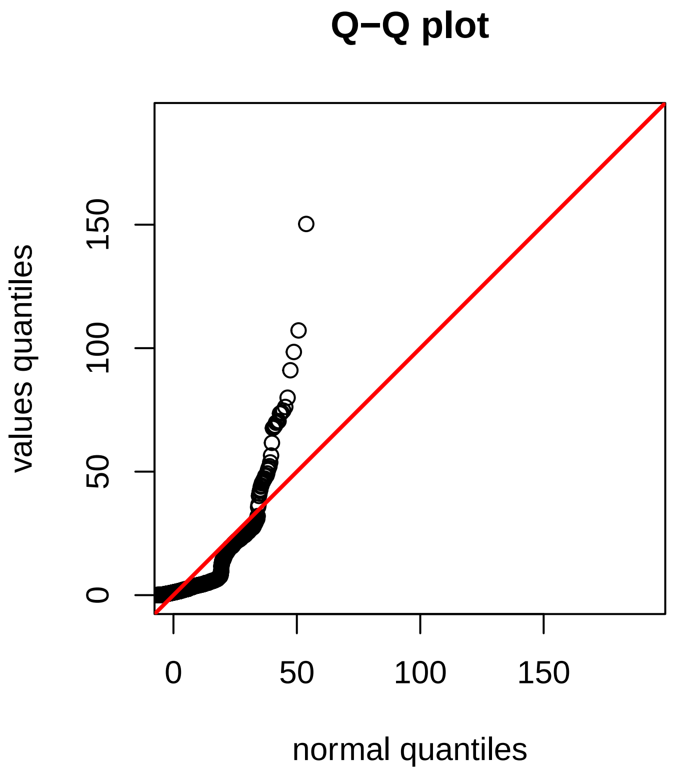
c) Use your normal Q-Q plot function to check if the values from each of the CLEC and ILEC samples we compared in question 2 could be normally distributed. What's your conclusion?

```
par(pty="s")

norm_qq_plot(hosts$CLEC)
```



```
norm_qq_plot(hosts$ILEC)
```



From the Q-Q plot we see the points stray from linearity, so CLEC and ILEC is not normally distributed. Check with shapiro normality test.

```
shapiro.test(hosts$CLEC)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  hosts$CLEC
## W = 0.63665, p-value = 2.339e-06
```

```
shapiro.test(hosts$ILEC)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  hosts$ILEC
## W = 0.56012, p-value < 2.2e-16
```

The result shows CLEC and ILEC is not normally distributed.