BACS HW (Week 6)

due on 03/26 (Sun)

### 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)

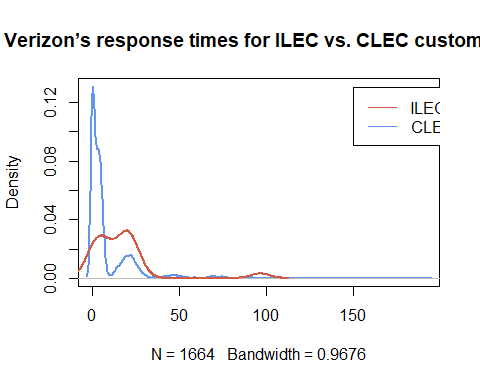
## Warning: 套件 'tidyr' 是用 R 版本 4.2.3 來建造的

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

### 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("ILEC", "CLEC"), 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)

: The mean of response times for CLEC customers the mean of response times for ILEC customers : 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

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 , accept at 99% confidence.

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

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

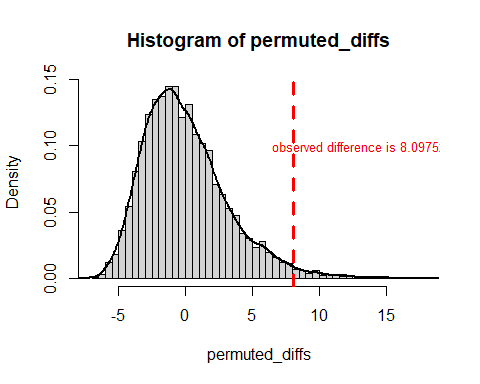
##   
## Welch Two Sample t-test  
##   
## data: hosts$ILEC and hosts$CLEC  
## t = -1.9834, df = 22.346, p-value = 0.02987  
## alternative hypothesis: true difference in means is less than 0  
## 99 percent confidence interval:  
## -Inf 2.130858  
## sample estimates:  
## mean of x mean of y   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

The overall p-value = 0.02987 > 0.01, cannot reject 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?

p\_1tailed <- sum(permuted\_diffs > observed\_diff) / nperms  
p\_2tailed <- sum(abs(permuted\_diffs) > observed\_diff) / nperms

The one-tailed p-value is 0.0175. The two-tailed p-value is 0.0175.

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

The overall p-value = 0.0175 > 0.01, cannot reject 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.

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))

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

### 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.

### 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.

### 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

### b) nterpret 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.

### 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?