BACS-hw6.R

2022-04-22

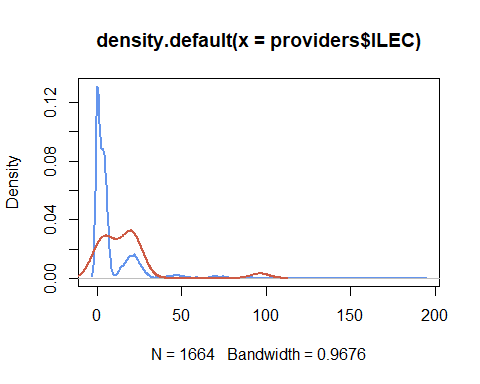
#Question 1  
#a, b  
setwd("C:/Users/eason/Desktop/清大 BACS/資料/")  
verizon\_wide <- read.csv("verizon\_wide.csv")  
library(tidyr)  
verizon\_long <- gather(verizon\_wide, na.rm=TRUE, key = "provider", value = "time")  
providers <- split(x = verizon\_long$time, f = verizon\_long$provider)  
  
#c  
head(verizon\_long)

## provider 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)

## provider 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  
plot(density(providers$ILEC), col="cornflowerblue", lwd=2)  
lines(density(providers$CLEC), col="coral3", lwd=2)



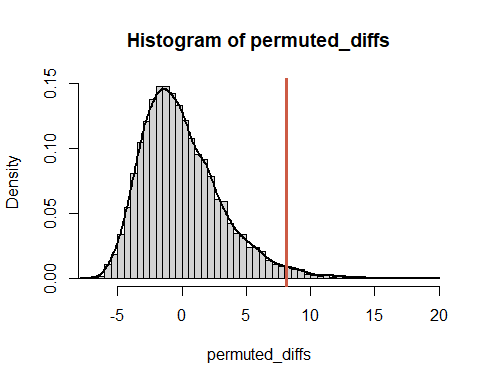
#Question 2  
#a  
  
#b1,2 不太確定significance0.01是甚麼意思  
t.test(providers$CLEC, providers$ILEC, alt="greater", var.equal=TRUE)

##   
## Two Sample t-test  
##   
## data: providers$CLEC and providers$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

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

##   
## Welch Two Sample t-test  
##   
## data: providers$CLEC and providers$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

#c1 看一下函式怎麼寫  
observed\_diff <- mean(providers$CLEC) - mean(providers$ILEC)  
  
permute\_diff <- function(values, groups){  
 permuted <- sample(values, replace=FALSE)  
 grouped <- split(permuted, groups)  
 mean(grouped$CLEC) - mean(grouped$ILEC)  
}  
  
nperms <- 10000  
set.seed(43)  
permuted\_diffs <- replicate(nperms, permute\_diff(verizon\_long$time, verizon\_long$provider))  
hist(permuted\_diffs, breaks = "fd", probability=TRUE)  
lines(density(permuted\_diffs), lwd=2)  
abline(v=observed\_diff, col="coral3", lwd=3)



#c2  
p\_1tailed <- sum(permuted\_diffs > observed\_diff) / nperms  
p\_2tailed <- sum(abs(permuted\_diffs) > observed\_diff) / nperms  
  
p\_1tailed

## [1] 0.0195

p\_2tailed

## [1] 0.0195

#c3   
#We cannot reject Hnull  
  
#Question 3  
#a1  
gt\_eq <- function(a, b){  
 ifelse(a>b, 1, 0) + ifelse(a==b, 0.5, 0)  
}  
W <- sum(outer(providers$CLEC, providers$ILEC, FUN = gt\_eq))  
  
#b  
n1 = length(providers$CLEC)   
n2 = length(providers$ILEC)   
  
n1

## [1] 23

n2

## [1] 1664

wilcox\_p\_1tail <- 1 - pwilcox(W, n1, n2)  
wilcox\_p\_2tail <- 2 \* wilcox\_p\_1tail  
  
wilcox\_p\_1tail

## [1] 0.0003688341

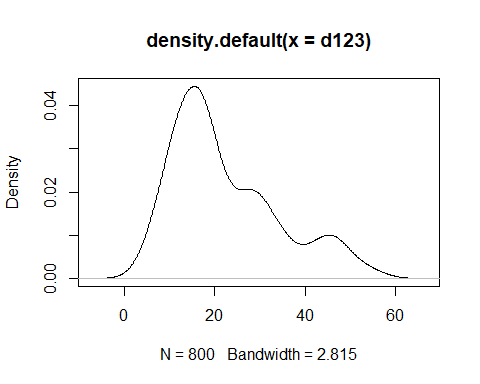
wilcox\_p\_2tail

## [1] 0.0007376683

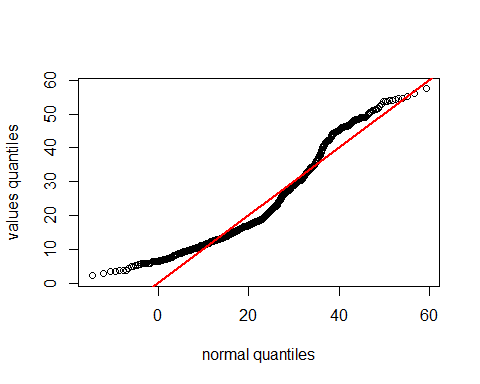
#c paired=false是甚麼意思  
wilcox.test(time ~ provider, data=verizon\_long, alternative="greater", paired=FALSE)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: time by provider  
## W = 26820, p-value = 0.0004565  
## alternative hypothesis: true location shift is greater than 0

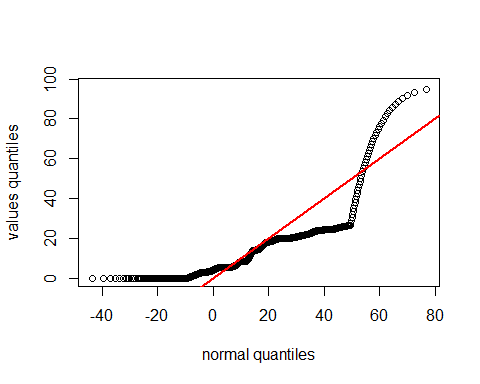
#d自己再想想  
#Values in the two samples are distinct(values seem shifted)  
#Means of samples are not distinguishable, but CLEC vs ILEC customers seem to be treated differently  
  
#Question 4  
#a1~5  
norm\_qq\_plot <- function(values){  
 probs1000 <- seq(0, 1, 0.001)  
 q\_vals <- quantile(values, probs1000)  
 q\_norm <- qnorm(probs1000, mean=mean(values), sd=sd(values))  
 plot(q\_norm, q\_vals, xlab="normal quantiles", ylab="values quantiles")  
 abline(a=0, b=1, col="red", lwd=2)  
}  
  
#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)



#c  
norm\_qq\_plot(providers$CLEC)



norm\_qq\_plot(providers$ILEC)

