BACS HW6 106070020

The classmate that help me 106000199

2021年4月8日

```
library(dplyr)
```

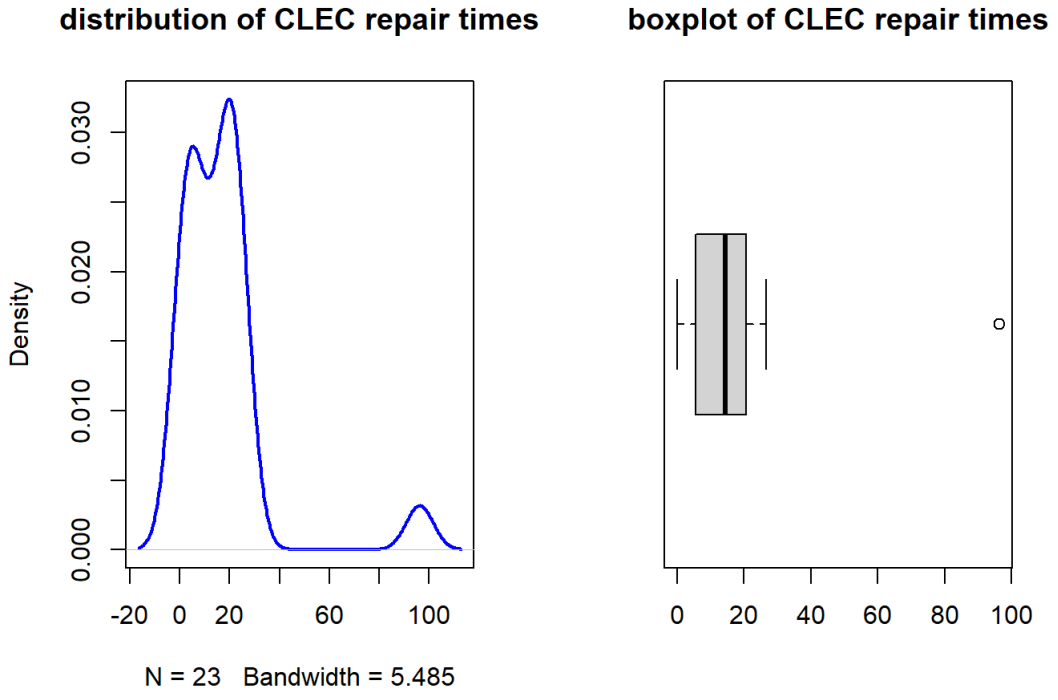
```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
ver <- read.csv("C:/Users/eva/Downloads/verizon.csv", header = T)
clec<-filter(.data=ver, ver$Group == "CLEC")
ct<-clec$Time
ilec<-filter(.data=ver, ver$Group == "ILEC")
it<-ilec$Time
```

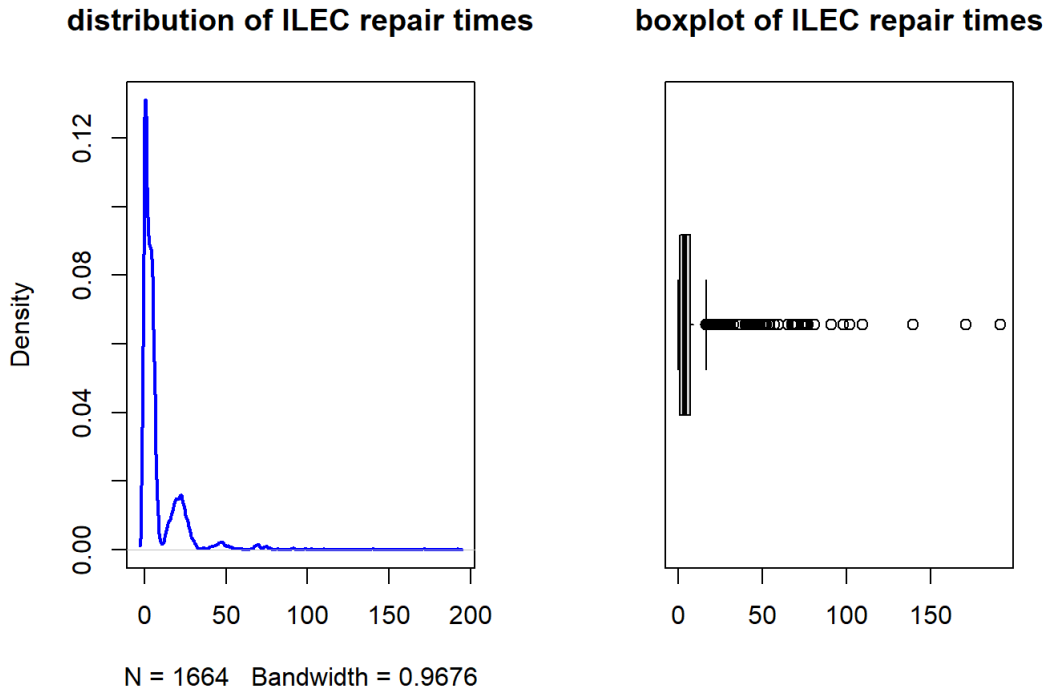
Problem 1

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

```
par(mfrow=c(1,2))
plot(density(ct), lwd=2, col="blue", main="distribution of CLEC repair times")
ctbox<-boxplot(ct, horizontal = T, main="boxplot of CLEC repair times")
```




```
par(mfrow=c(1,2))
plot(density(it), lwd=2, col="blue", main="distribution of ILEC repair times")
itbox<-boxplot(it, horizontal = T, main="boxplot of ILEC repair times")
```



(b) Use the appropriate form of the t.test() function to test the difference between the mean of ILEC sample response times versus the mean of CLEC sample response times. From the output of t.test():

(i) What are the appropriate null and alternative hypotheses in this case?

```
t.test(ct,it,var.equal=FALSE,conf.level = 0.99,alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  ct and it
## 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
```

H0: True difference in means is less than or equal to 0, H1: true difference in means is larger than 0.

(ii) Based on output of the t.test(), would you reject the null hypothesis or not?

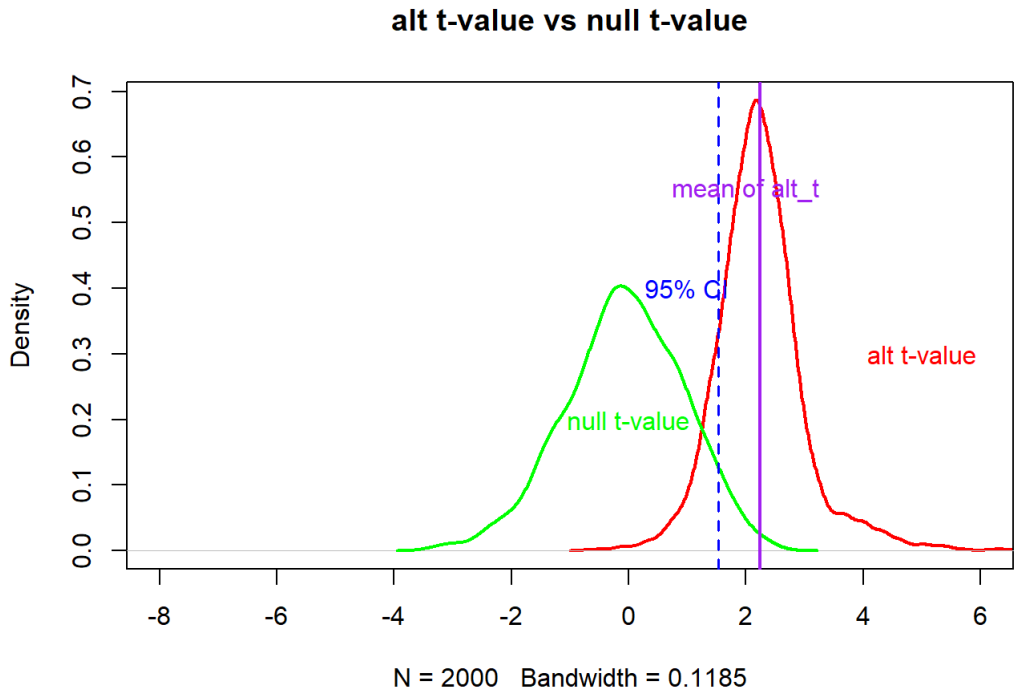
According to the t.test, (mean x) - (mean y) = 8.097519, which is in the 99% confidence level, and the p-value is also larger than 0.01; therefore, we cannot reject H0.

(c)

(i)Plot a distribution of the bootstrapped null t-values and alternative t-values, adding vertical lines to show the 5% rejection zone of the null distribution

```
bootstrap_null_alt<-function(sample0,sample1) {
  resample0 <-sample(sample0, length(sample0), replace=TRUE)
  resample1 <-sample(sample1, length(sample1), replace=TRUE)
  resample_se0<-sd(resample0)/sqrt(length(resample0))
  resample_se1<-sd(resample1)/sqrt(length(resample1))
  t_stat_alt<-(mean(resample1)-mean(sample0)) /sqrt(resample_se1^2+resample_se0^2)
  t_stat_null<-(mean(resample0)-mean(sample0))/resample_se0
  c(t_stat_alt,t_stat_null)
}
set.seed(42)

boot_stats<-replicate(2000,bootstrap_null_alt(it,ct))
alt_t<-boot_stats[1,]
null_t<-boot_stats[2,]
{plot(density(alt_t),lty=1,col="red",lwd=2, xlim=c(-8,6), main="alt t-value vs null t-value")
  lines(density(null_t),lty=1,col="green",lwd=2)
  abline(v=quantile(null_t,probs=0.95),lty=2,col="blue",lwd=1.5)
  abline(v=mean(alt_t),lty=1,col="purple",lwd=2)
  text(x=0, y=0.2, "null t-value", col="green")
  text(x=5, y=0.3, "alt t-value", col="red")
  text(x=2, y=0.55, "mean of alt_t", col="purple")
  text(x=1, y=0.4, "95% CI", col="blue")}
```



(ii) Based on these bootstrapped results, should we reject the null hypothesis?

> According to the bootstrap results, the mean of alt t-value is out of the 95% CI of the null hypothesis, so we should reject H0.

Problem 2

(a) What is the null and alternative hypotheses in this case?

```
var(ct)
```

```
## [1] 380.3895
```

```
var(it)
```

```
## [1] 215.7973
```

```
var.test(ct,it,alternative="greater")
```

```
##
##  F test to compare two variances
##
## data:  ct and it
## F = 1.7627, num df = 22, denom df = 1663, p-value = 0.01582
## alternative hypothesis: true ratio of variances is greater than 1
## 95 percent confidence interval:
##  1.138356      Inf
## sample estimates:
## ratio of variances
##           1.762717
```

> H0: True ratio of variances is less than or equal to 1, H1: True ratio of variances is larger than 1.

(b)

(i) What is the F-statistic of the ratio of variances?

```
f_value=var(ct)/var(it)
f_value
```

```
## [1] 1.762717
```

(ii) What is the cut-off value of F, such that we want to reject the 5% most extreme F-values?

```
cut_off<-qf(p=0.95, df1=length(ct)-1,df2=length(it)-1)
cut_off
```

```
## [1] 1.548476
```

(iii) Can we reject the null hypothesis?

> The f_value is 1.762717, and the cutoff in this example is 1.548476. The f_value is grater than the cutoff, so we can reject H0.

(c)

(i)Create bootstrapped values of the F-statistic, for both null and alternative hypotheses.

```
set.seed(43)
sd_providers_test<-function(larger_sd_sample,smaller_sd_sample)
   {resample_larger_sd<-sample(larger_sd_sample, length(larger_sd_sample), replace=TRUE)
    resample_smaller_sd<-sample(smaller_sd_sample, length(smaller_sd_sample), replace=TRUE)
    f_alt<-var(resample_larger_sd) /var(resample_smaller_sd)
    f_null<-var(resample_larger_sd) /var(larger_sd_sample)
    c(f_alt,f_null)}
f_stats<-replicate(10000,sd_providers_test(ct,it))
f_alts<-f_stats[1,]
f_nulls<-f_stats[2,]
```

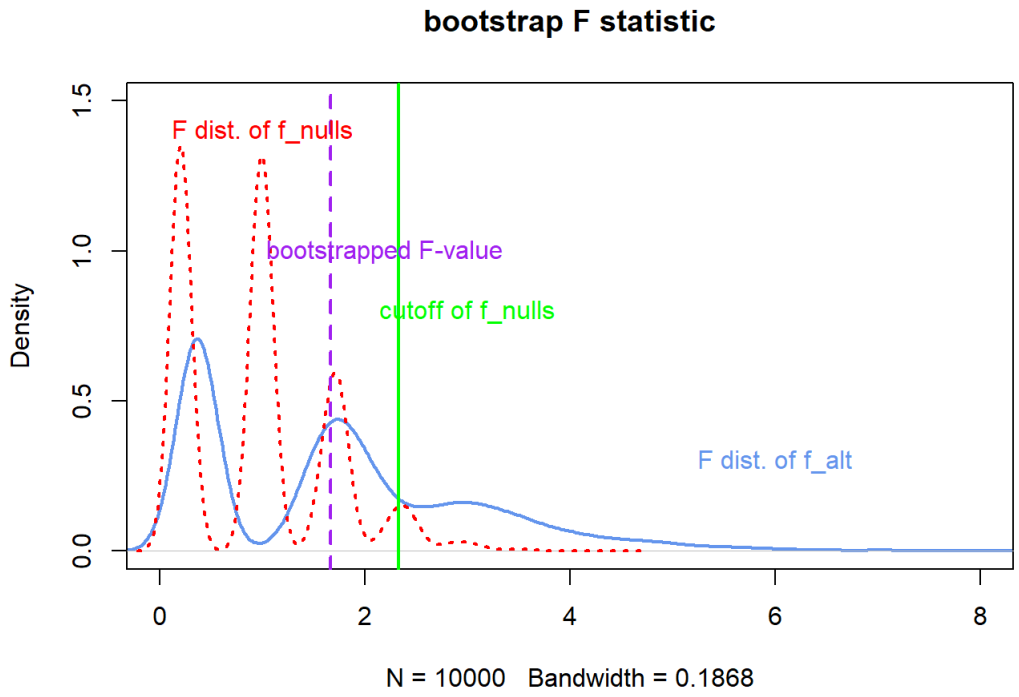
(ii) What is the 95% cutoff value according to the bootstrapped null values of F?

```
quantile(f_nulls,probs=0.95)
```

```
##      95%
## 2.331735
```

(iii) Plot a visualization of the null and alternative distributions of the bootstrapped F-statistic, with vertical lines at the cutoff value of F nulls.

```
{plot(density(f_alts),ylim=c(0,1.5),xlim=c(0,8),col="cornflowerblue",lwd=2, main="bootstrap F statistic")
lines(density(f_nulls),lty="dotted",col="Red",lwd=2)
abline(v=quantile(f_nulls,probs=0.95),lty=1,col="green",lwd=2)
abline(v=median(f_alts),lty=2,col="purple",lwd=2)
text(x=6, y=0.3,"F dist. of f_alt", col='cornflowerblue')
text(x=1, y=1.4,"F dist. of f_nulls",col='red')
text(x=3, y=0.8,"cutoff of f_nulls",col='green')
text(x=2.2, y=1,"bootstrapped F-value",col='purple')}
```



(iv) What do the bootstrap results suggest about the null hypothesis?

> According to the graph, the bootstrap F value is less than the cutoff of f_nulls, so we cannot reject the null hypothesis(True ratio of variances is equal to 1).

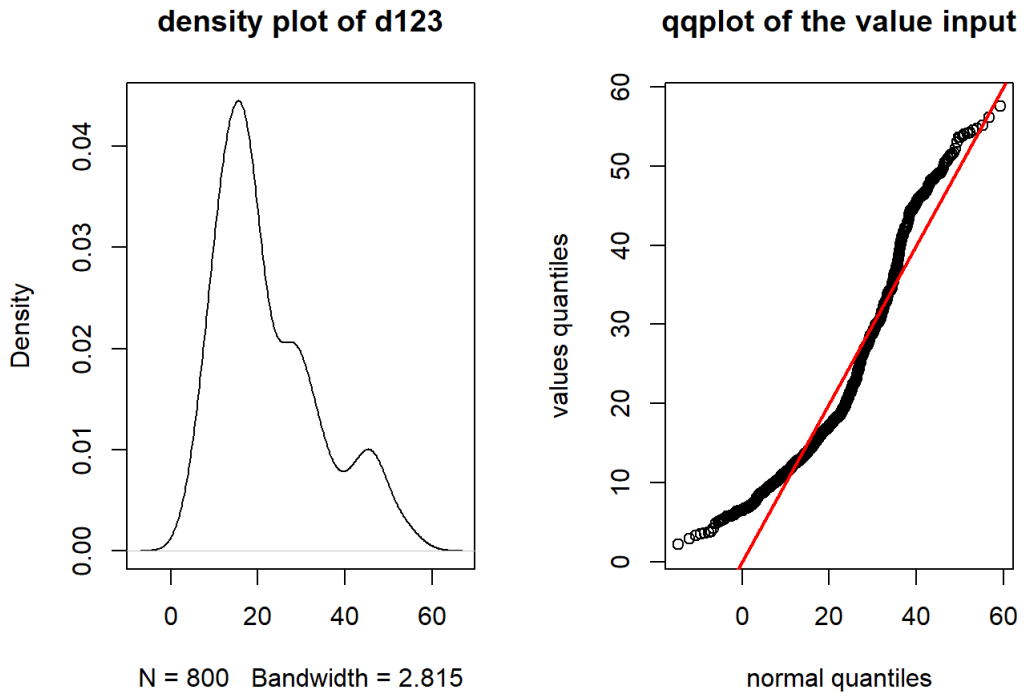
Problem 3

(a) Make a function called norm_qq_plot() that takes a set of values)

```
norm_qq_plot <- function(values) {
  probs1000 <- seq(0, 1, 0.001)
  q_vals <- quantile(values, probs1000)
  q_norm<-qnorm(probs1000,mean(values),sd(values))
  plot(q_norm, q_vals, xlab="normal quantiles", ylab="values quantiles", main="qqplot of the value input")
  abline(0, 1, col="red", lwd=2)
}
```

(b) Confirm that your function works by running it against the values of our d123

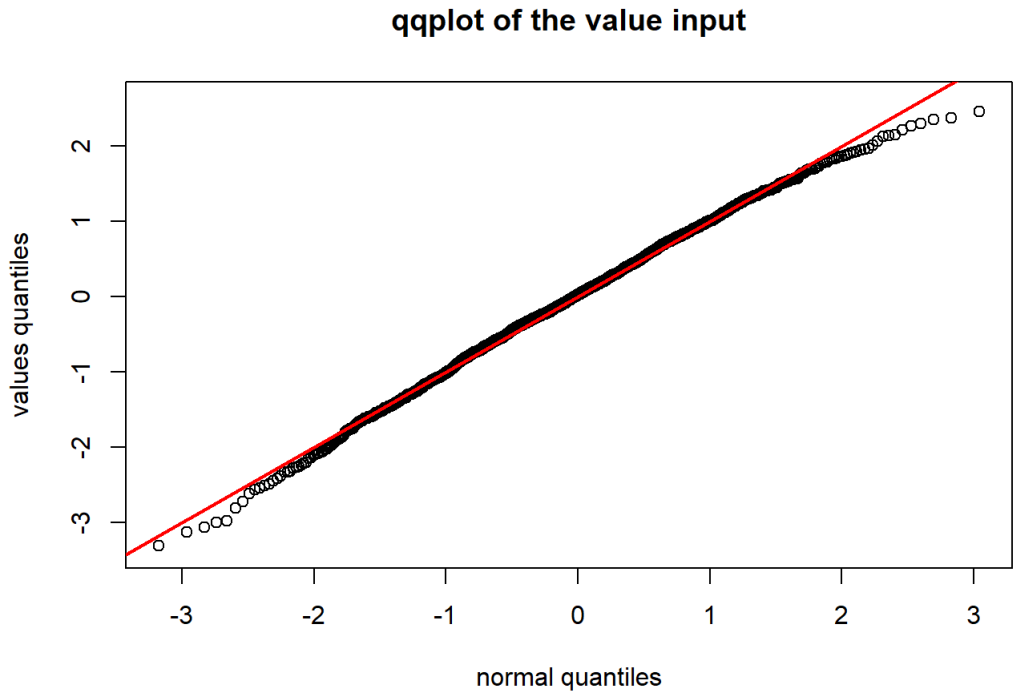
```
set.seed(978234)
d1 <- rnorm(n=500, mean=15, sd=5)
d2 <- rnorm(n=200, mean=30, sd=5)
d3 <- rnorm(n=100, mean=45, sd=5)
d123 <- c(d1, d2, d3)
par(mfrow=c(1,2))
plot(density(d123), main="density plot of d123")
norm_qq_plot(d123)
```



The plot shows the dataset d123 has “fat tails,” which means that compared to the normal distribution there is more data located at the extremes of the distribution and less data in the center of the distribution. In terms of quantiles this means that the first quantile is much less than the first theoretical quantile and the last quantile is greater than the last theoretical quantile. This trend is reflected in the corresponding Q-Q plot.

(c) We generally don’t need to use bootstrapping for hypothesis tests of the mean (t-tests) if the null distribution of the t-statistic follows a normal distribution (traditional statistics measures would work fine). Use your normal Q-Q plot function to check if the bootstrapped distribution of null t-values in question 1c was normally distributed. What’s your conclusion?

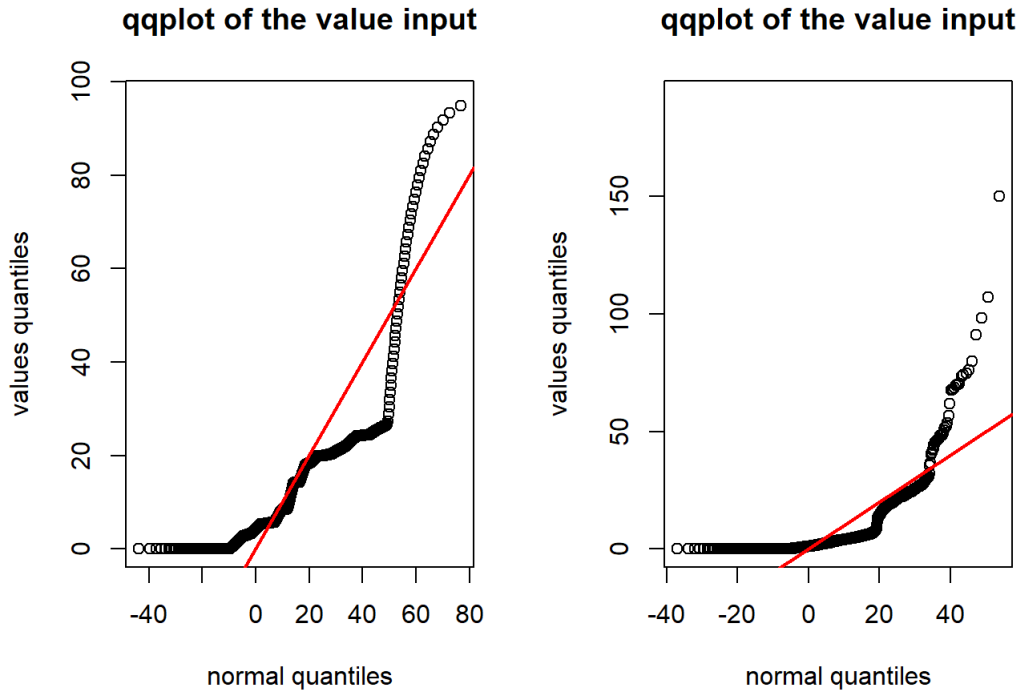
```
norm_qq_plot(null_t)
```



The distribution of null t-values in the question 1c is normally distributed, as the scatter plot constructs a relatively straight line as the red line does.

(d) Hypothesis tests of variances (f-tests) assume the two samples we are comparing come from normally distributed populations. Use your normal Q-Q plot function to check if the two samples we compared in question 2 could have been normally distributed. What's your conclusion?

```
par(mfrow=c(1,2))
norm_qq_plot(ct)
norm_qq_plot(it)
```



Both of the samples (it, ct) in question 2 are not normally distributed, as the scatter plots are not formed as straight lines, and the graphs are seemed to have the fat tails in the Q-Q plot.