STAT5034 Homework6

Zhengzhi Lin
Department of Statistics, Virginia Tech

November 16, 2019

1 Problem 1

1.1 (a)

Step1: Select 16 samples from observations with replace.
Step2: Calculate sample variance from the sample selected.
Step3: Repeat 1 and 2 100 times to get 100 sample variances.
Step4: Calculate 10% quantile from 100 sample variances and it would be the lower bound.

1.2 (b)

The CI for s is: $[1.39, \infty)$ of 90% confidence. That is, there is 90% chance that our CI contains the true standard deviation of population.
The CI calculated from bootstrap is very close to analytic CI calculated from distribution that is presented in notes.
Conclusion: Bootstrap CV is 14.6, conclude that under best situation, the field would not be acceptable.

1.3 (c)

My bootstrap procedure has different assumptions from the procedure of our note. The one from our note assumes the observations are iid normal distributed. Bootstrap only assume the data is collected independently.

2 Problem 2

1. We know that $\hat{p} = \overline{Y}$, and by law of large numbers, we have

$$\hat{p} \xrightarrow{d} N\left(p, \frac{p(1-p)}{n}\right)$$

2. Then by delta method, we have

$$g(\hat{p}) \xrightarrow{d} N\left(g(p), \frac{p(1-p)}{n} g'(p)^2\right)$$

3. Thus we get,

$$log\left(\frac{\hat{p}}{1-\hat{p}}\right) \xrightarrow{d} N\left(log\left(\frac{p}{1-p}\right), \frac{1}{np(1-p)}\right)$$

pluging $\hat{p}$, we get our CI:

$$log\left(\frac{\hat{p}}{1-\hat{p}}\right) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n\hat{p}(1-\hat{p})}}$$

Let $\hat{p} = 0.67, n = 67, \alpha = 1\%$, we calculate our CI is: $[0.0389, 1.377]$

3 Problem 3

3.1 (a)

1. By Fisher transformation, we know,

$$\sqrt{n}\left(\frac{1}{2}log\left(\frac{1+r}{1-r}\right) - \frac{1}{2}log\left(\frac{1+\rho}{1-\rho}\right)\right) \xrightarrow{d} N(0,1)$$

2. Then we apply delta method,

$$\text{Let } T(r) = \frac{1}{2}log\left(\frac{1+r}{1-r}\right) \Rightarrow r = \frac{e^{2T}-1}{e^{2T}+1} \Rightarrow \frac{dr}{dT} = \frac{4e^{2T}}{(e^{2T}+1)^2}$$

3. Now we get asymptotic distribution of r:

$$\sqrt{n}(r-\rho) \xrightarrow{d} N\left(0, \left(\frac{4e^{2T}}{(e^{2T}+1)^2}\right)^2\right)$$

4. Therefore the CI of ρ will be:

$$\left[r - z_{\alpha/2}\frac{4e^{2T}}{\sqrt{n}(e^{2T}+1)^2}, r + z_{\alpha/2}\frac{4e^{2T}}{\sqrt{n}(e^{2T}+1)^2}\right]$$

5. Plugging the data, we get CI: [0.776,0.890]

3.2 (b)

It is fine to use bootstrapping to get CI of ρ, because it only need independence assumption and it is asymptotically consistent.

3.3 (c)

I choose to use bootstrapping to get the point estimation and confidence interval of ρ, because of its simplicity and asymptotic consistency.

How bootstrapping works in this case:
Step1: Select 100 samples from observations with replace.
Step2: Calculate sample correlation from the sample selected.
Step3: Repeat 1 and 2 1000 times to get 1000 sample correlations.
Step4: Calculate 10% and 90% quantile from 1000 sample variances and it would be the CI. The point estimation will be the mean.

We get the point estimation of ρ is 0.83, CI: [0.79,0.87]

4 Problem 4

4.1 (a)

By looking at the histgram of Y, we know that group 0 is more centered than group 1. By looking at the summary result of data we can roughly assume that group 0 has lower mean than group 1.

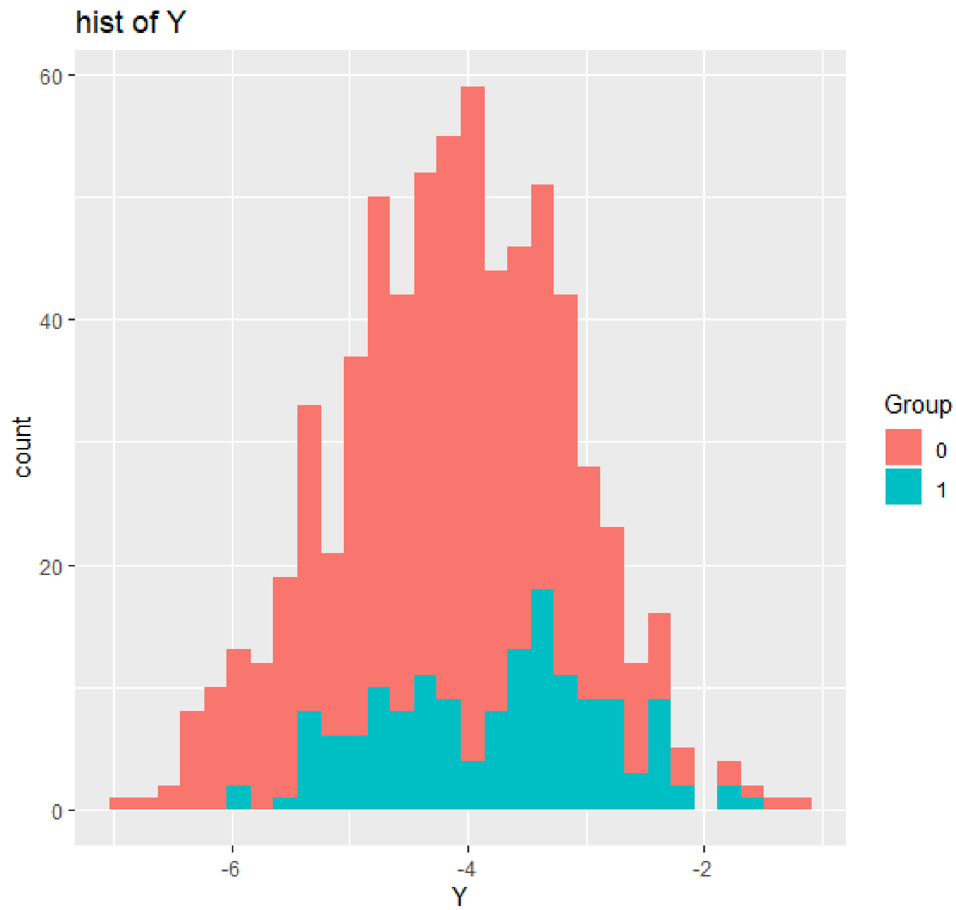

Figure 1: histgram Plot

	Y	Group
	Min. :-6.970	0:540
	1st Qu.:-4.832	1: 0
	Median :-4.150	NA
	Mean :-4.222	NA
	3rd Qu.:-3.560	NA
	Max. :-1.230	NA

Figure 2: summary of Y(group=0) Plot

```
|   |      Y         |Group  |
|:--|:---------------|:------|
|   |Min.   :-5.920  |0:   0 |
|   |1st Qu.:-4.522  |1:150  |
|   |Median :-3.645  |NA     |
|   |Mean   :-3.784  |NA     |
|   |3rd Qu.:-3.103  |NA     |
|   |Max.   :-1.510  |NA     |
```

Figure 3: summary of Y(group=1) Plot

4.2 (b)

$$d = \frac{\overline{Y_1} - \overline{Y_0}}{S_p} = \frac{-3.78 + 4.22}{0.96} = 0.454$$

4.3 (c)

Function is in Appendix

4.4 (d)

My function ch(p4dd,"Group",0.05) gives us estimation of cohen's d : -0.45 with CI [-0.98,0.003]. We have 95% chance that our CI contains true cohen's d value.

Appendix A Code

A.1 Code for problem 1

```
dat <- c(10.54,8.58,11.28,12.43,10.34,11.31,
9.03,9.79,10.49,11.26,7.37,6.08,9.89,8.28,7.28,8.00)
boost <- matrix(0,17,100)
for(i in 1:100){

  boost[1:16,i] <- sample(dat,16,replace = T)
  boost[17,i] <- sqrt(var(boost[1:16,i]))

}
s <- quantile(boost[17,],0.1)
```

```
s/mean(dat) * 100
s^2
```

A.2 Code for problem 2

```
qnorm(0.005,0,1,lower.tail = F)
log(0.67/(1-0.67)) -
    qnorm(0.005,0,1,lower.tail = F)*sqrt(1/(67*0.67*(1-0.67)))
```

A.3 Code for problem 3

```
#a
treedat <- read.csv("treedat.csv")
r <- cor(treedat$DBH,treedat$height)
T <- 1/2*(log((1+r)/(1-r)))
r+4*exp(2*T)/(exp(2*T)+1)^2/sqrt(111) * 1.96
r-4*exp(2*T)/(exp(2*T)+1)^2/sqrt(111) * 1.96
#b

#c
boost <- matrix(0,1,1000)
for(i in 1:1000){

  index <- sample(seq(1:111),100,replace = T)
  t <- treedat[index,c(1,3)]
  boost[i] <- cor(t[,1],t[,2])
}
mean(boost)
quantile(boost,0.1)
quantile(boost,0.9)
```

A.4 Code for problem 4

```
library(ggplot2)
library(knitr)
p4dd <- read.csv("P4DD.csv")
p4dd$Group <- as.factor(p4dd$Group)
ggplot(data = p4dd) + geom_histogram(aes(Y, fill=Group)) +
    ggtitle("hist of Y")
```

```
kable(summary(p4dd[which(p4dd$Group==1),]))
kable(summary(p4dd[which(p4dd$Group==0),]))

n1 <- nrow(p4dd[which(p4dd$Group==1),])
n0 <- nrow(p4dd[which(p4dd$Group==0),])
s1 <- var(p4dd[which(p4dd$Group==1),1])
s0 <- var(p4dd[which(p4dd$Group==0),1])
sp <- sqrt(((n1-1)*s1+(n0-1)*s0)/(n1+n0-2))

(mean(p4dd$Y[which(p4dd$Group==1)]) -
    mean(p4dd$Y[which(p4dd$Group==0)]))/sp

#function of 4(c)
ch <- function(data,group,alpha){
  t <- paste(group)
  l <- levels(data[,t])
  boost <- matrix(0,1,1000)
  for(i in 1:1000){
    index <- sample(seq(1:nrow(data)),100,replace = T)
    bdata <- data[index,]
    g1 <- bdata[which(bdata[,t]==l[1]),!names(data) %in% t]
    g2 <- bdata[which(bdata[,t]==l[2]),!names(data) %in% t]
    n1 <- length(g1)
    n2 <- length(g2)
    s1 <- var(g1)
    s2 <- var(g2)
    sp <- sqrt(((n1-1)*s1+(n2-1)*s2)/(n1+n2-2))
    boost[i] <- (mean(g1) - mean(g2))/sp
  }
  lb <- quantile(boost,alpha/2)
  ub <- quantile(boost,1-alpha/2)
  return(c(mean(boost),lb,ub))

}

ch(p4dd,"Group",0.05)
```