Statistical Methods for Data Science

MINI PROJECT 3

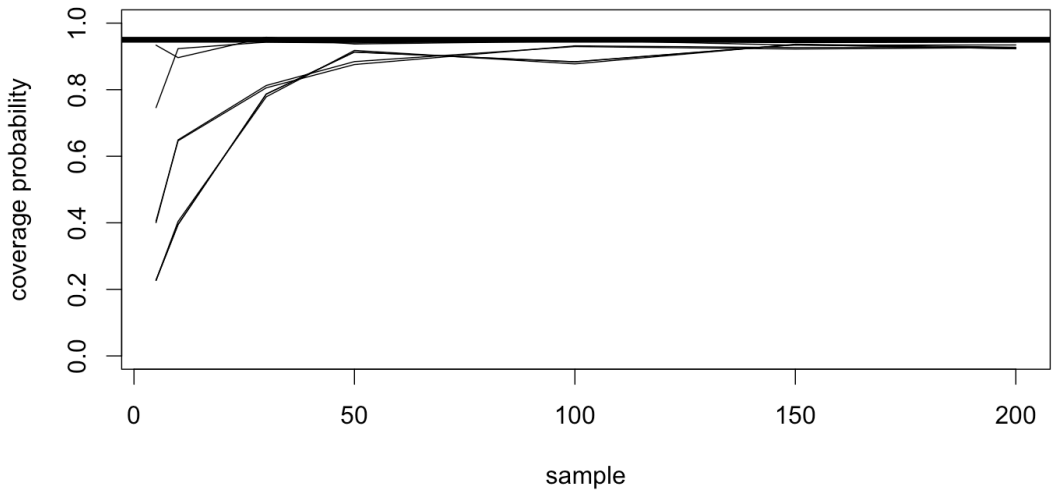
Charu Arora
cxa150730

EXERCISE 1

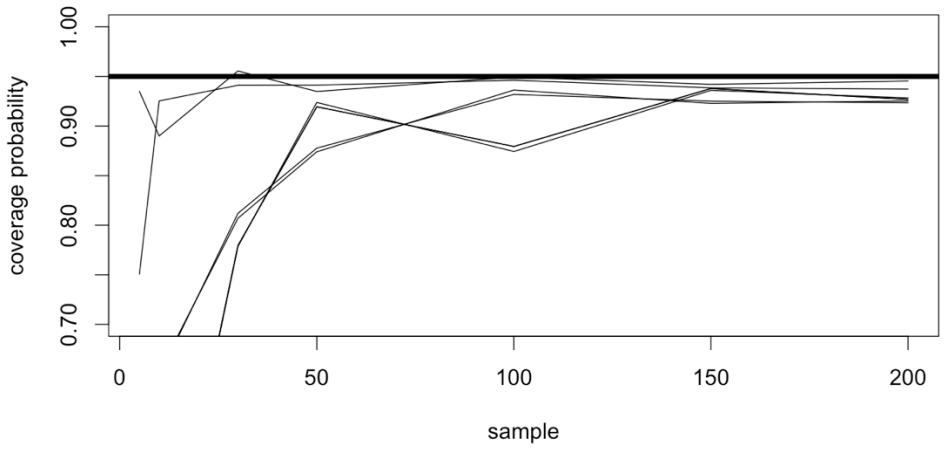
Values used to determine how large the value of n should be:

sample=c(5,10,30,50,100,150,200)
prob=c(0.05,0.1,0.25,0.5,0.9,0.95)

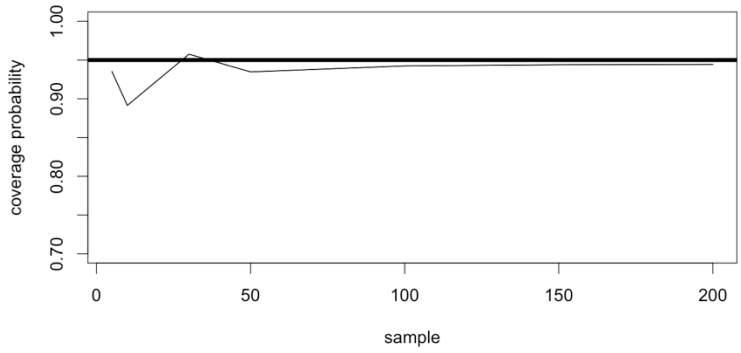
From the graph, it is clear that as the value of n increases, the accuracy increases.
For a smaller value of n (n<30), the values are mostly inaccurate.



Accordingly, if the value of n is larger than 100, then we get a more accurate value. It is also noticed that to have an acceptable accuracy, value of p does not matter. As the value of n increases, for any probability, i.e., 0.1 to 0.95, we obtain the desired accuracy. Therefore, the accuracy only depends on the value of n and not on p.
I would recommend the value of n to be 150.



Also, for probability = 0.5, its noticed that for all values of n, it generates a pretty accurate value. The graph for p=0.5 and all the values of n is ---->

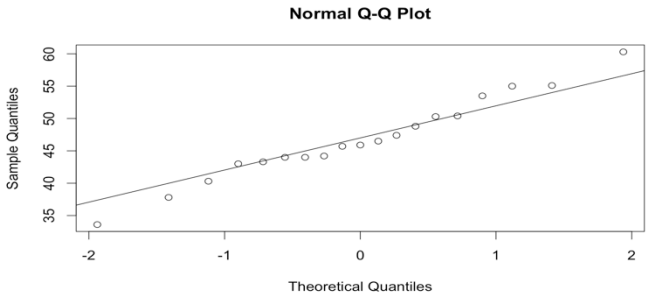


R code

```
Nsim<-10000;                    #simulating 10,000 draws using Monte Carlo Simulations
Sample<-c(5,10,30,50,100,150,200)
Prob<-c(0.05,0.1,0.25,0.5,0.9,0.95)
K<-1
J<-1
for(p in prob)
{
 value<-c()

 obt<-1
 for(n in sample)
 {

  p.hat<-rbinom(nsim,n,p)/n
  se<-sqrt(p.hat*(1-p.hat)/n)
  alpha<-0.05;
  lower<-p.hat - qnorm(1-alpha/2)*se        #finding the lower and upper bound values
  upper<-p.hat + qnorm(1-alpha/2)*se
  sum<-0
  for (i in 1:nsim)                  #check how many times the probability lies within the range
  {
   x<- ((p >= lower[i])&(p <= upper[i]) )
   sum<-sum+x
  }
  avg<-sum/nsim
  value[obt]<-avg
  obt<-obt+1

 }
 if(j){
  plot(sample,value,ylim=c(0,1),ylab="coverage probability",type="l") #plotting all the values
  j<-0
 }
 lines(sample,value,lty=1)
}
abline(h=0.95,lwd=5)          #plot line of y=0.95 , to check which all values are close to the value
                              0.95
```

EXERCISE 2

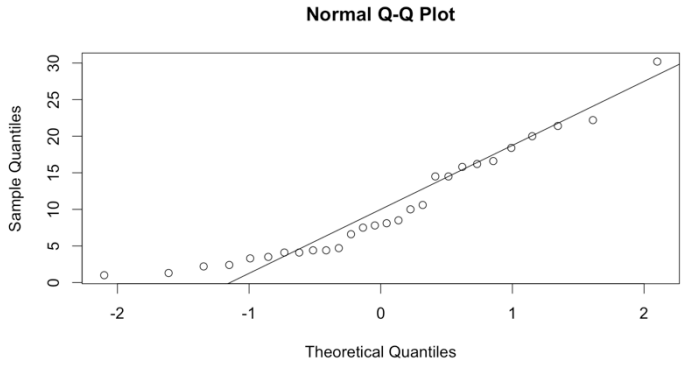
a)



Even though we see a slight curve in the plot, the circles lie very close to the line. At the left and right ends of the plot, the circles are somewhat farther away from the line. This just shows the distribution has a long tail. Therefore, we can say, the distribution is approximately normal.

Here also, since the circles lie very close to the normal line, we can say that it has a normal distribution.

Therefore, from the qqplot it is reasonable to assume that each sample comes from a normal distribution.



b) Yes, the variances of the two distributions can be assumed to be equal. We construct the confidence interval for the ratio of population variances.
RESULT: [0.3102977 , 1.7564758]
As

$$\frac{\text{Var(child)}}{\text{Var(adult)}}$$

could be equal to 1, we conclude that it is in fact possible for the two variances to be equal.

c) The 95% confidence interval for the difference in mean sugar contents of the two cereal types is –
RESULT: [32.35553 , 40.92680]
Here, the values of n.child and n.adult is 19 and 28 respectively, both of which lie below 30. Based on the results obtained in the qqplot, we can conclude that both our distributions are normally distributed. When we construct the confidence interval for the ratio of population variances (result in previous part). Since 1 lies in the confidence interval, we assume that the two variances are equal.

d) Since are confidence interval for the difference is completely above 0, we can conclude that the sugar content of child's cereal is more that the sugar content of adult cereals.
Yes, the child's cereal has more sugar content on average than adult cereals by 32.35 to 40.92 percentage of weight.

R code

```
child<-c(40.3,55,45.7,43.3,50.3,45.9,53.5,43,44.2,44,47.4,44,33.6,55.1,48.8,50.4,37.8,60.3,46.5)
qqnorm(child)
qqline(child)          #qqplot for the sugar content in child's cereal
```

```
adult<-c(20,30.2,2.2,7.5,4.4,22.2,16.6,14.5,21.4,3.3,6.6,7.8,10.6,16.2,14.5,4.1,15.8,4.1,2.4,3.5,8.5,10,1,
4.4,1.3,8.1,4.7,18.4)
qqnorm(adult)
qqline(adult)          #qqplot for the sugar content in adult's cereal
qqplot(adult)

alpha<-0.05

n.c<-length(child)
n.a<-length(adult)

x.c<-mean(child)
x.a<-mean(adult)

sd.c<-sd(child)
sd.a<-sd(adult)

var.c<-var(child)
var.a<-var(adult)

f.1<-qf(alpha/2,n.c-1,n.a-1)
f.2<-qf(alpha/2,n.a-1,n.c-1)
ratio<-((sd.c/sd.a)^2)*c(f.2,1/f.1)      #to check if they have equal variances or not using f
                                          distribution

pooled.var<-(((n.c-1)*var.c) + ((n.a-1)*var.a))/(n.c+n.a-2)      #using pooled variance as var(child) can
                                                                  be assumed to be equal to var(adult)

se<-sqrt(pooled.var)*sqrt((1/n.c)+(1/n.a))
ci<-(x.c - x.a) + c(-1,1) * (qt(1-(alpha/2),n.c+n.a-2) * se)      #formula to calculate the confidence
                                                                  interval for the difference in mean
```

EXERCISE 3

a) We have two proportions, 61/414 adults who grew up in a single-parent household reporting to suffer at least one incident of abuse and 74/501 adults who grew up in a two-parent household reporting abuse. We construct the confidence interval for the difference between two proportions.
RESULT : [-0.04652425 , 0.04580106]
is a 95% confidence interval for the difference in reporting abuse.

The estimator $\hat{p}_1 - \hat{p}_2 = -0.0003615956$, suggests that larger number of adults who grew up in a two parent household reported abuse, but since our confidence interval includes the value 0, we conclude that the difference between adults growing up in single and two parent household reporting abuse is not statistically significant, i.e., the two proportions are the same.

b) Here, we assume that our distribution is normal as the values of n1<-414 and n2<-501 is large. Central limit theorem states that the sampling distribution will be normal or nearly normal, if the sample size is large enough and a value of n<-30 is considered large. This makes our assumption reasonable.

R code

```
p1<-61/414      #proportion 1
p2<-74/501      #proportion 2
n1<-414
n2<-501
alpha<-0.05
z<-qnorm(1-(alpha/2))
se<-sqrt(((p1*(1-p1)/n1)+((p2*(1-p2)/n2))))
ci<-p1 - p2 +  c(-1,1) * z * se          #formula to construct 95% confidence interval of difference in proportions
```