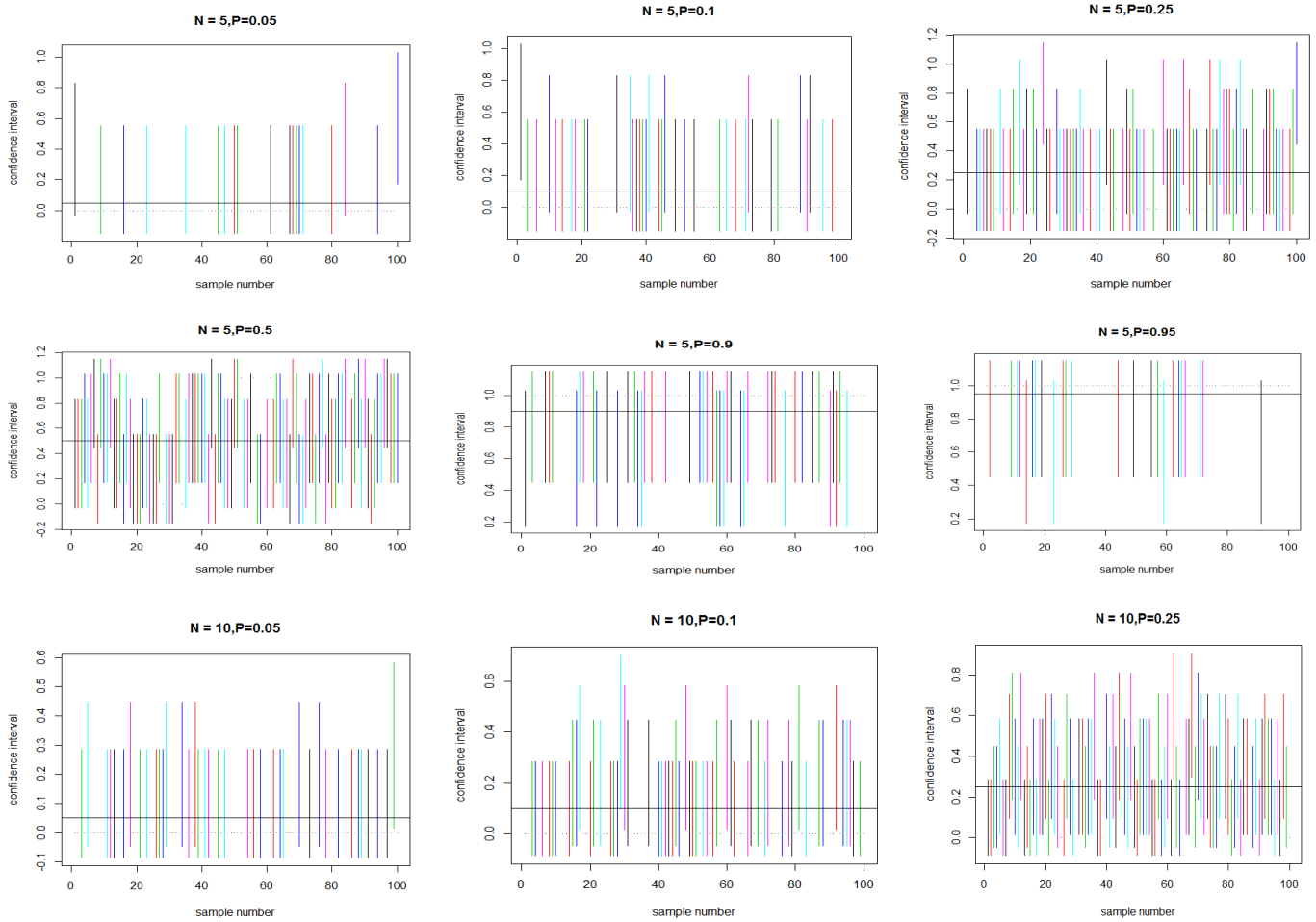
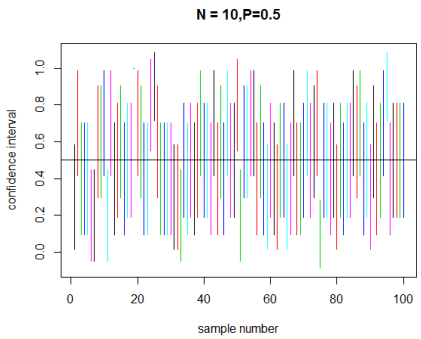
Project 2 **Section: CS 6313.501**

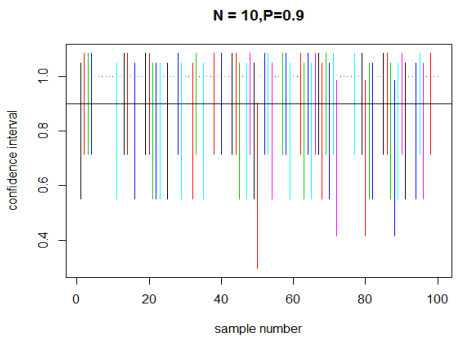
Name : Panchami G. Rudrakshi

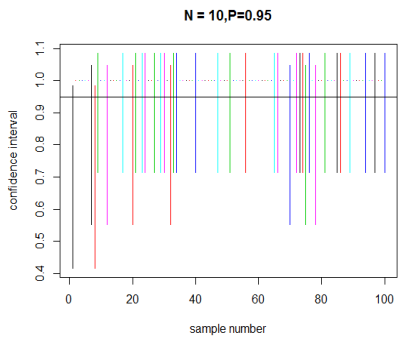
1) We know how to construct a large sample confidence interval for a population proportion p. How large n should be for this interval to have acceptable accuracy? Answer this question by computing the coverage probability of this interval using Monte Carlo simulation, and examining how close the probability is to the nominal confidence level. Take level of confidence to be 95% but use a variety of values for n and p, e.g., n = 5, 10, 30, 50, 100, and p = 0.05, 0.1, 0.25, 0.5, 0.9, 0.95. Summarize your results graphically. Comment on any patterns you see in the results. Based on your findings, what n would you recommend for the use of this confidence interval? Would your answer depend on p? Explain.

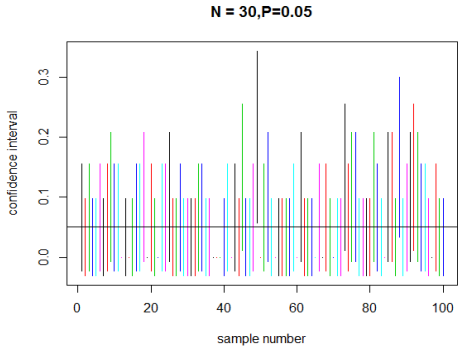
Given: CI=0.95%
Coverage probability is computed using Monte Carlo simulation to examine how close the probability is to the nominal confidence level.
Taking: n = 5, 10, 30, 50, 100, and p = 0.05, 0.1, 0.25, 0.5, 0.9, 0.95
Observations:
As n increases with constant high probability, mean lies in the nominal confidence level.
As p decreases, large n (80-100) is required to get the mean in the nominal confidence level.
With large p, we can get many mean values in the required confidence interval with small n(30-40).
Thus, based on the observations, we can conclude that optimum values of n & p would give most of the values in nominal confidence level.
Also, it depends on p. As p increases, the value of N can decrease below 80 satisfying the interval.

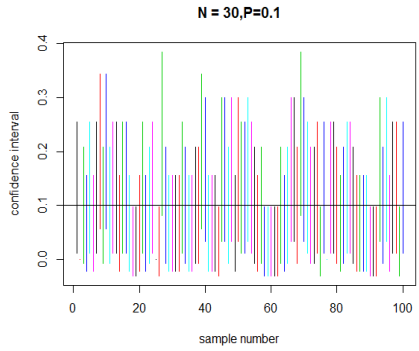


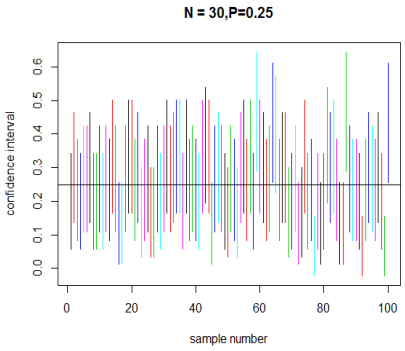

N = 10,P=0.5
confidence interval
sample number

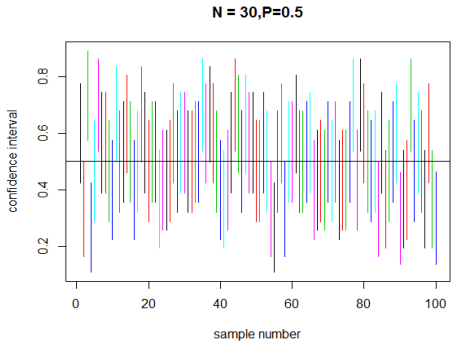

N = 10,P=0.9
confidence interval
sample number

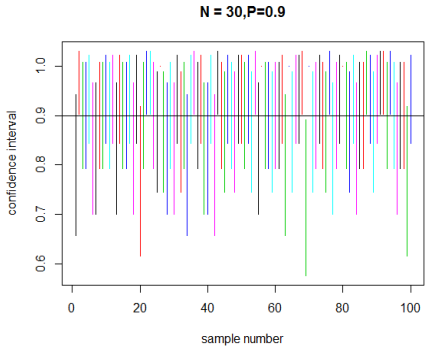

N = 10,P=0.95
confidence interval
sample number

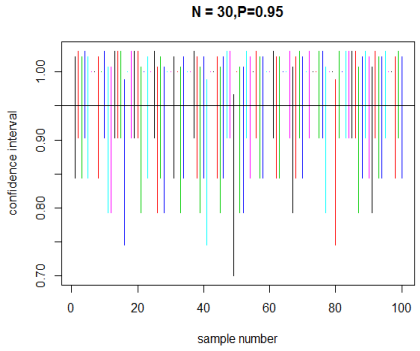

N = 30,P=0.05
confidence interval
sample number

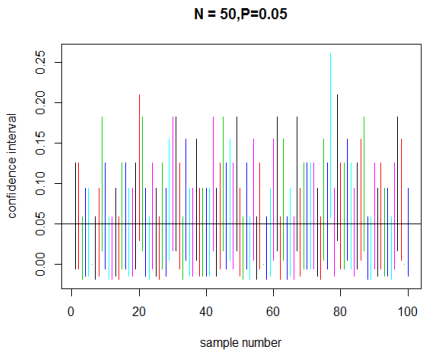

N = 30,P=0.1
confidence interval
sample number

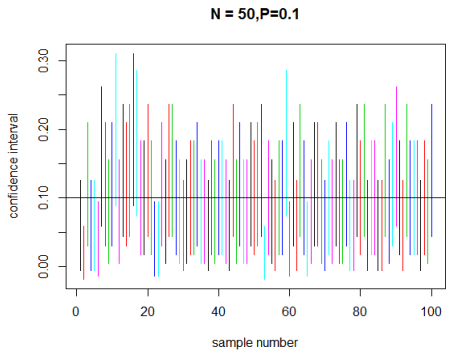

N = 30,P=0.25
confidence interval
sample number

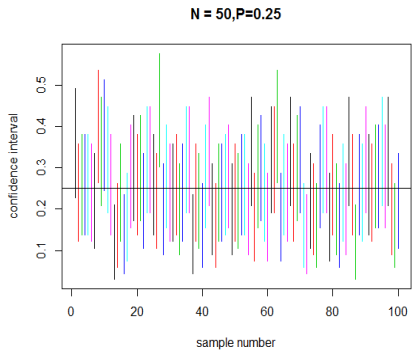

N = 30,P=0.5
confidence interval
sample number

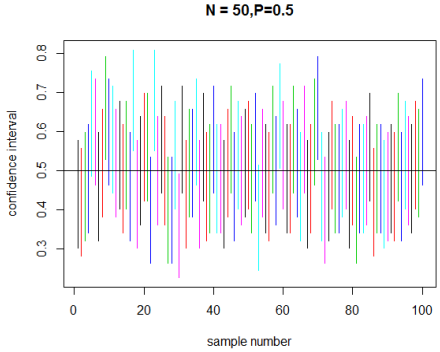

N = 30,P=0.9
confidence interval
sample number

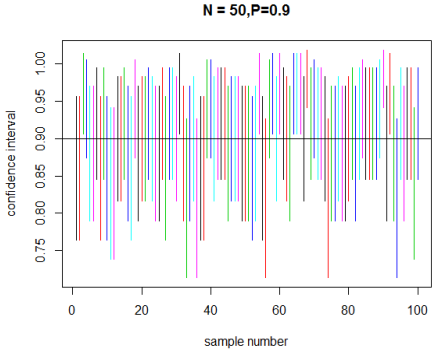

N = 30,P=0.95
confidence interval
sample number

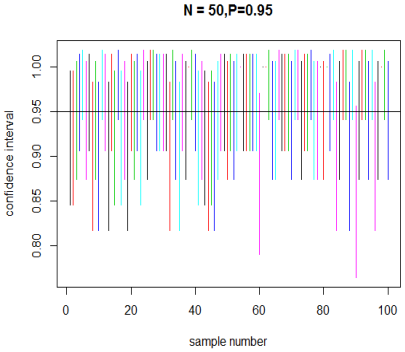

N = 50,P=0.05
confidence interval
sample number

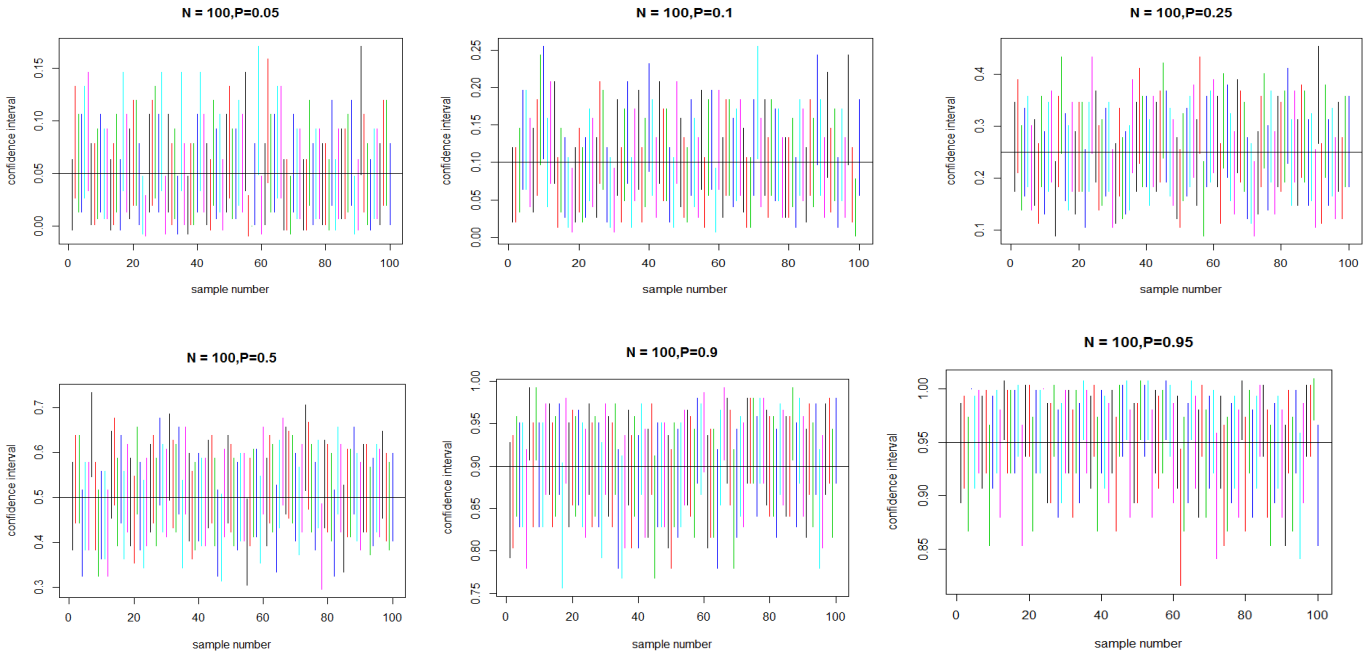

N = 50,P=0.1
confidence interval
sample number


N = 50,P=0.25
confidence interval
sample number


N = 50,P=0.5
confidence interval
sample number


N = 50,P=0.9
confidence interval
sample number


N = 50,P=0.95
confidence interval
sample number



2) The data below show the sugar content (as a percentage of weight) of several national brands of children's and adults' cereals.

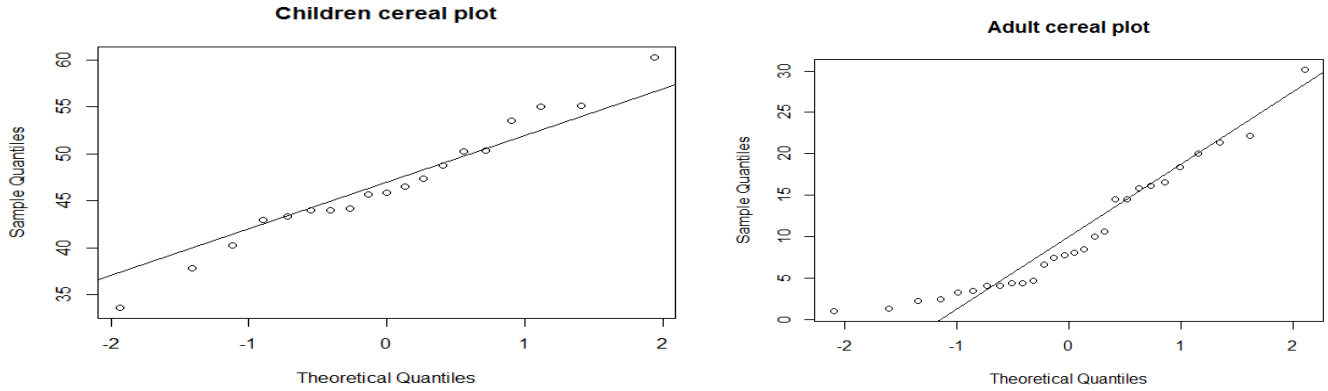
Given:

Children's cereals: 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

Adults' cereals: 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

(a) Does it seem reasonable to assume that each sample comes from a normal distribution? Draw Q-Q plots to answer this question.

Ans: Seeing the Q-Q plots for x & y, we can conclude that each sample comes from an approximately normal distribution as the plots does not exactly lie on the line but are very near to the line in both the cases.



(b) Can the variances of the two distributions be assumed to be equal? Justify your answer.

Ans: Yes. The variances can be assumed equal as 1 lies in the ratio of variances i.e. from 0.3102977 1.7564758.

(c) Compute an appropriate 95% confidence interval for difference in mean sugar contents of the two cereal types. What assumptions did your make, if any, to construct the interval?

Ans: As the variances are assumed to be equal, pooled sample variance is used.

As n is small, T distribution is used.

The pooled variance is 7.158801. The confidence interval obtained is 32.35553 - 40.92680.

(d) What do you conclude on the basis of your answer in (c)? Can we say that children's cereals have more sugar on average than adult cereals? If yes, by how much? Justify your answers.
Ans: Yes, children's cereals has a higher value than adult cereals. Also, since CI (32.35553 40.92680) does not include 0, we can conclude children's cereals have more sugar on average than adult cereals.

3) A study shows that 61 of 414 adults who grew up in a single-parent household report that they suffered at least one incident of abuse during childhood. By contrast, 74 of 501 adults who grew up in two-parent households report abuse.

(a) Is there a difference in single-parent and two-parent households when it comes to reporting abuse? Answer this question by computing an appropriate 95% confidence interval.

Ans: Since, the null value i.e. 0 lies in the CI interval of 95%, we cannot find statistically significant difference in single-parent and two-parent households when it comes to reporting abuse. The CI 95% does not provide sufficient evidence for the children with single or two parent about the abusement.

(b) What assumptions, if any, did you make to compute the interval in (a)? Do the assumptions seem reasonable?
Ans: As N>30 & M>30 i.e. sample size is large, Z distribution is used.

R CODE:
1)

```
conf.int <- function(p, alpha, n, se.p )
  {
  y = rbinom(n,1,p)                  # calling binomial to find probability proportion
  p1 = mean(y)                       # mean calculation
  se.p <- sqrt(p1 * (1-p1)/n)        # standard error calculation
  ci <- p1 + c(-1, 1) * qnorm(1 - (alpha/2)) * se.p               # CI calculation
  return(ci)
}

n = 80                                                  # n is changed for every iteration
p = 0.5                                                 # p is changed for every iteration
alpha <- 0.05                                           # CI of 95%
conf.int(p, alpha, n, se.p)                             # function call
nsim <- 10000                                           # no of simulation
ci.mat <- replicate(nsim, conf.int(p, alpha, n, se.p))
matplot(rbind(1:100, 1:100), ci.mat[, 1:100], type = "l", lty = 1, main = "N =
5,P=0.05", xlab = "sample number", ylab = "confidence interval")        # matrix of CI

abline(h = p)                                                  # drawing plot of CI Vs mean

# Proportion of times the interval is correct i.e. convergence probability
mean( (p >= ci.mat[1,])*(p <= ci.mat[2,]) )
```

2)

```
# input Children's cereals:
```

```
a = 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(a, main='Children cereal normal plot');                          #to plot Q-Q plot
qqline(a)
# input Adult's cereals:
b = 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(b, main='Adult cereal normal plot');                             #to plot Q-Q plot
qqline(b)
n.a = length(a)                                              # length of Children's cereals
n.b = length(b)                                              # length of Children's cereals
alpha = 0.05                                                                  # CI of 95%
f.l.crit = qf(alpha/2, n.a - 1, n.b - 1)
f.u.crit = qf(1 - (alpha/2), n.a - 1, n.b - 1)
ratio=((sd(a)/sd(b))^2) * c(1/f.u.crit, 1/f.l.crit)    # to check equality of variances
ratio
#calculating pooled standard deviation
sp = sqrt(((n.a-1)*var(a) + (n.b -1) * var(b))/(n.a+n.b -2))
sp
#calculating CI
ci=mean(a) - mean(b) + c(-1,1)*qt(1-alpha/2,n.a+n.b-2)*sqrt((1/n.a)+(1/n.b))*sp
ci
```

Results:

```
Ratio
[1] 0.3102977 1.7564758
sp
[1] 7.158801
Ci
[1] 32.35553 40.92680p1 <- 61/414
```

3)

```
p1 <- 61/414
p2 <- 74/501
n <- 414
m <- 501
phat= p1-p2
phat
alpha <- 0.05
se=sqrt(((p1*(1-p1))/n)+((p2*(1-p2))/m))
se
ci=(phat) +c(1,-1)*qnorm (1-(alpha/2))*se
ci
```

Result:

```
phat
[1] -0.0003615956
se
[1] 0.02355281
ci
[1]  0.04580106 -0.04652425
```