Question:

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 85% 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.

Report

The important parameters are shown below:

> Take level of confidence to be 85% 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.

So that there will be 5 graphs that n equals 5, 10, 30, 50, 100 separately. For each situation, let p = 0.05, 0.1, 0.25, 0.5, 0.9, 0.95.

First I should make a function which describe the distribution and computing of CI. The function is given in professor's class and I do a little changes because in this question, the parameter is p instead of "mu" and "sigma".

```
conf.int <- function(p,n,alpha) {
   x <- rbinom(1, n, p)
   phat<-x/n
   ci <- phat + c(-1, 1) * qnorm(1 - (alpha/2))*sqrt((phat*(1-phat))/n)
   return(ci)
}
```

In this question, 1- alpha = 85%, so alpha = 0.15. We do Monte Carlo simulation for 10000 times. The parameters set up below:

```
p <- 0.05
n <- 5
alpha <- 0.15
nsim <- 10000
ci.mat <- replicate(nsim, conf.int(p,n, alpha))
```

And then we get the proportion of times the interval is correct:

```
mean( (p >= ci.mat[1,])*(p <= ci.mat[2,]) )
[1] 0.2086
>
```

And then we change the value of n and p and finish the form.

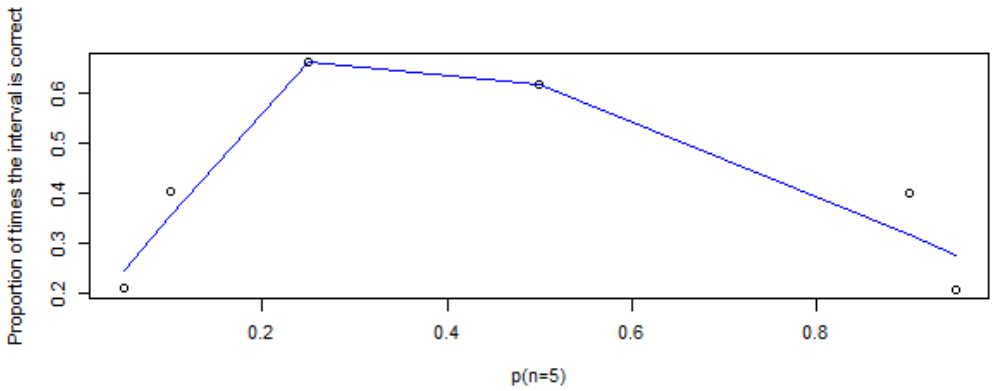
	p = 0.05	p = 0.1	p = 0.25	p = 0.5	p = 0.9	p = 0.95
n = 5	0.2086	0.4052	0.6625	0.6176	0.3995	0.206
n = 10	0.3933	0.6313	0.6814	0.8922	0.6383	0.3917
n = 30	0.7811	0.7868	0.8506	0.8023	0.7894	0.7616
n = 50	0.6819	0.8291	0.8538	0.7988	0.8293	0.6879
n = 100	0.8453	0.8507	0.8389	0.8551	0.844	0.8511

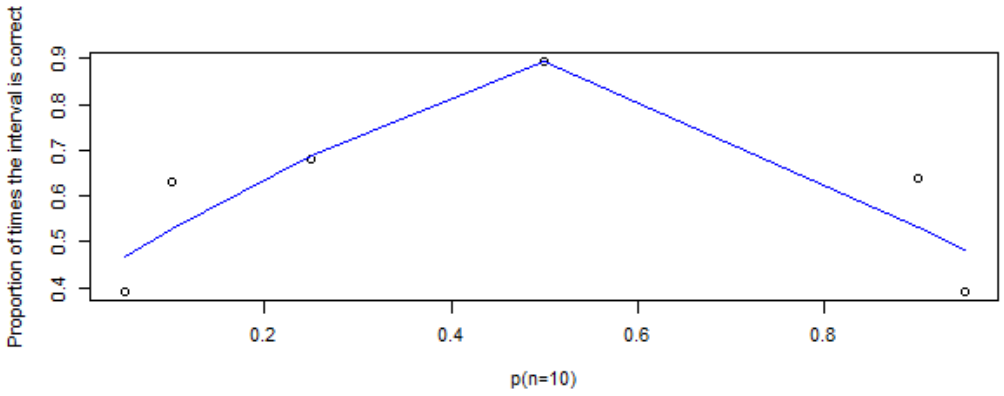
```
p5<-c(0.2086,0.4052,0.6625,0.6176,0.3995,0.206)
p10<-c(0.3933,0.6313,0.6814,0.8922,0.6383,0.3917)
p30<-c(0.7811,0.7868,0.8506,0.8023,0.7894,0.7616)
p50<-c(0.6819,0.8291,0.8538,0.7988,0.8293,0.6879)
p100<-c(0.8453, 0.8507,0.8389,0.8551,0.844,0.8511)
```

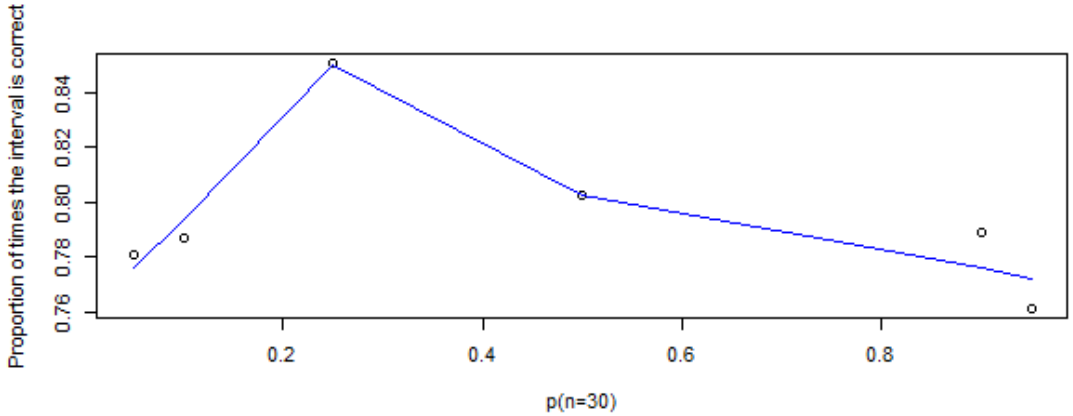
Based on those information, we can get 5 graphs by running the command:

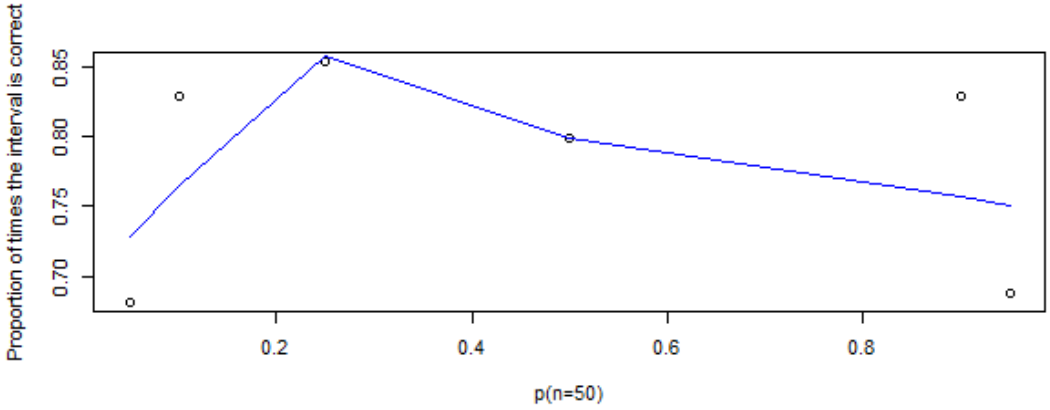
```
p<-c(0.05, 0.1, 0.25, 0.5, 0.9, 0.95)
plot(p5~p, xlab="p(n=5)", ylab="Proportion of times the interval is correct")
lines(lowess(p,p5), col="blue")
```

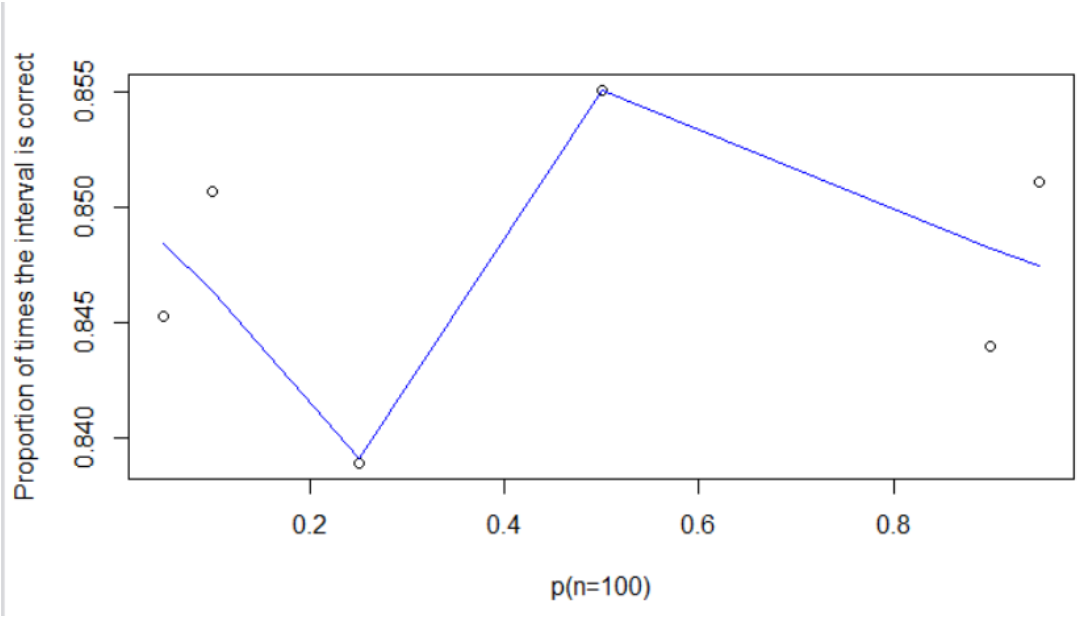
What we need change is p5, p10, p30, p50, p100 and xlab="p(n=5)"
Finally the graphs show below:




Proportion of times the interval is correct
0.4
0.5
0.6
0.7
0.8
0.9
0.2
0.4
0.6
0.8
p(n=10)


Proportion of times the interval is correct
0.76
0.78
0.80
0.82
0.84
0.2
0.4
0.6
0.8
p(n=30)


Proportion of times the interval is correct
0.70
0.75
0.80
0.85
0.2
0.4
0.6
0.8
p(n=50)



From those graphs I can find when n equals 5 or 10, the proportion (probability) may get a low value. When n is larger than 30, the proportion is high.

So I think n should be larger than 30.

Also, when the value of p is near 0.5, proportion will get the highest value. However, this regulation is not shown on the last graph (n=100). I think the reason is because those six values in graph 5 is really close with each other, even a really small difference may be shown as a huge difference in graph.