Mini Project # 3

Name: Niveditha Varadha Chandrasekaran

Exercise 1

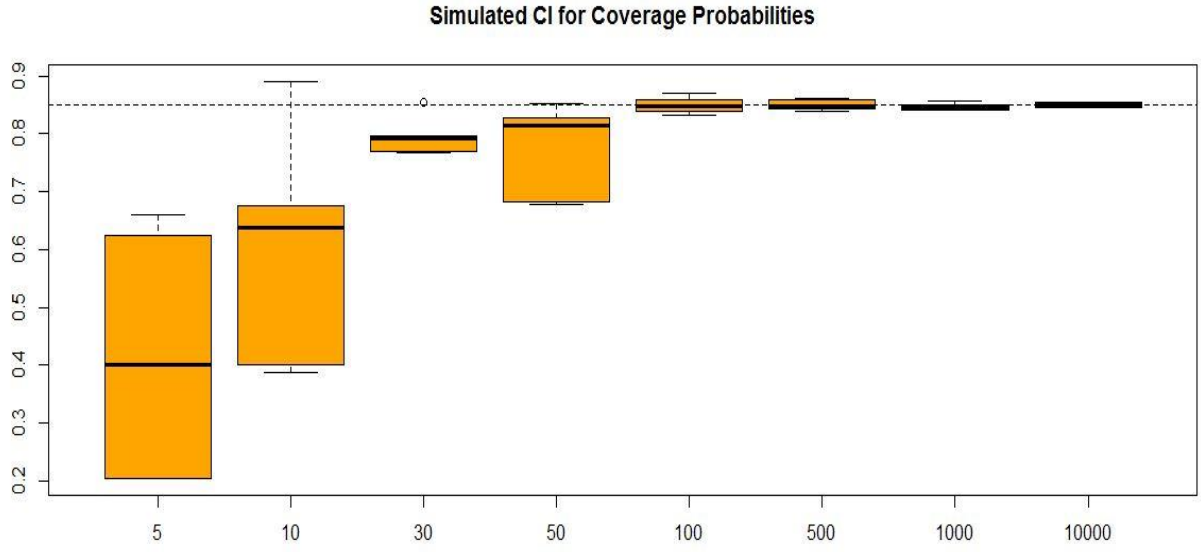
Problem Statement:

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.

Solution:

Coverage Probability for various n and p value is:

p / n	0.05	0.1	0.25	0.5	0.9	0.95
5	0.2037	0.3977	0.6608	0.6237	0.4035	0.2048
10	0.3867	0.6392	0.6767	0.8912	0.6365	0.4001
30	0.7663	0.7952	0.8553	0.7963	0.7897	0.7685
50	0.6784	0.8271	0.8531	0.8004	0.8271	0.6821
100	0.8589	0.8398	0.8320	0.8702	0.8477	0.8463
500	0.8482	0.8389	0.8584	0.8607	0.8438	0.8481
1000	0.8461	0.8399	0.8501	0.8466	0.8420	0.8577
10000	0.8543	0.8536	0.8446	0.8461	0.8515	0.8482



From the above box plot, the recommended value for **n is 10000**. The larger the n the more accurate the coverage probability is. Also from the box plot it is clear that the answer does not depend on p.

Exercise 2

Problem Statement:

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

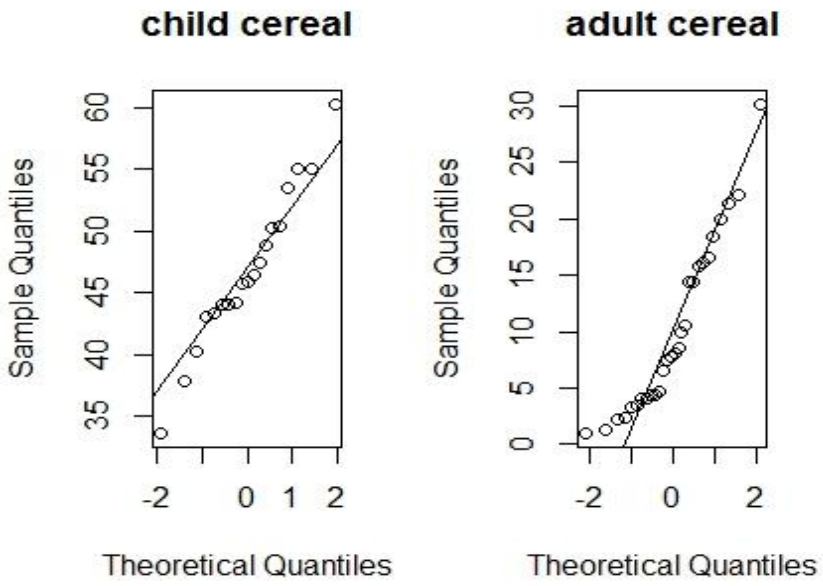
Children's cereals: 40.3, 55, 45.7, 43.3, 50.3, 45.9, 53.5, 43, 44.2, 44, 47.4, 44, 33.6, 55.1, 48.8, 50.4, 37.8, 60.3, 46.5

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.
(b) Can the variances of the two distributions be assumed to be equal? Justify your answer.
(c) Compute an appropriate 90% confidence interval for difference in mean sugar contents of the two cereal types. What assumptions did your make, if any, to construct the interval?
(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.

Solution:

(a) Q-Q Plot:



From the above plot, we see that the most of the points are closer to the line; hence it seems reasonable to assume that each sample comes from a normal distribution.

(b) By calculating CI for ratio of these two variances we get CI as (0.3551741, 1.5116458). The value 1 is included in the interval; hence we cannot conclusively say that the variances of the two distributions can be assumed to be equal.

(c) The 90% confidence interval for difference in mean sugar contents of the two cereal types (child cereal – adult cereal) is (33.18008, 40.10225).
The assumptions made are:
- The variances of population are not known, hence we use sample variances to estimate the whole population.
- Each sample comes from a normal distribution.

(d) On basis of Answer in (C) we can say that children`s cereals have more sugar on average than adult cereals. The mean value of Children cereals is 46.79474 and CI interval is (33.25503, 40.02730) and the amount of sugar is 6 points more than the average value in CI.

Exercise 3

Problem Statement:

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 99% confidence interval.
(b) What assumptions, if any, did you make to compute the interval in (a)? Do the assumptions seem reasonable?

Solution:

(a) The 99% confidence interval is (-0.06102961, 0.06030642). The value 0 is included in the interval, hence we cannot comment on if there is a difference in single-parent and two-parent households when it comes to reporting abuse.

(b) The assumptions made are:
- We are interested in finding Confidence Intervals for the difference between two proportions.
- We consider n1 (number of adults who grew up in a single-parent household) and n2 (number of adults who grew up in a two-parent household) as large.

Yes, these assumptions seem to be reasonable.

R Code:

Exercise 1:

```
> nsim=10000
> # Different n values:
> n.values=c(5,10,30,50,100,500,1000,10000)
> # Different p values
> p.values=c(0.05,0.1,0.25,0.5,0.9,0.95)
> coverage.probability=matrix(NA,nrow=length(n.values),ncol=length(p.values),byrow=TRUE)
> par(mfrow=c(1,1))
> # obtain the lower and upper limit of the 85 %CI and calculate the coverage probability
> for (i in 1:length(n.values)){
+   n=n.values[i]
+   probability=rep(NA, length(p.values))
+   for (j in 1:length(p.values)){
+     p=p.values[j]
+     samples=rbinom(nsim,size=n,p=p)
+     phat=samples/n
+     lower=phat-qnorm(0.925)*sqrt(phat*(1-phat)/n)
+     upper=phat+qnorm(0.925)*sqrt(phat*(1-phat)/n)
+     probability[j]=mean(lower<=p & upper>=p) #proportion of times the interval is correct
+   }
+   coverage.probability[i,]=probability
+ }
>
> coverage.probability
      [,1]   [,2]   [,3]   [,4]   [,5]   [,6]
[1,] 0.2037 0.3977 0.6608 0.6237 0.4035 0.2048
[2,] 0.3867 0.6392 0.6767 0.8912 0.6365 0.4001
[3,] 0.7663 0.7952 0.8553 0.7963 0.7897 0.7685
[4,] 0.6784 0.8271 0.8531 0.8004 0.8271 0.6821
[5,] 0.8589 0.8398 0.8320 0.8702 0.8477 0.8463
[6,] 0.8482 0.8389 0.8584 0.8607 0.8438 0.8481
[7,] 0.8461 0.8399 0.8501 0.8466 0.8420 0.8577
[8,] 0.8543 0.8536 0.8446 0.8461 0.8515 0.8482
> # Boxplot of the coverage probabilities for various n and p values
> boxplot(coverage.probability[1,],coverage.probability[2,],coverage.probability[3,],
+         coverage.probability[4,],coverage.probability[5,],coverage.probability[6,],
+         coverage.probability[7,],coverage.probability[8,],
+         names=c(5,10,30,50,100,500,1000,10000),main="Simulated CI for Coverage Probabilities",
+         cex=1,cex.axis=1,col="orange")
> abline(a=0.85,b=0,lty=2)
```

Exercise 2:

```
> childCereal<-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)
> adultCereal<-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)
>
> par(mfrow=c(1,2))
>
> #contruct Q-Q plot
> qqnorm(childCereal,main="child cereal")
> qqline(childCereal)
> qqnorm(adultCereal,main="adult cereal")
> qqline(adultCereal)
>
> n_cc = length(childCereal)
> n_ac = length(adultCereal)
> sample_mean_cc = mean(childCereal) # sample mean of child cereal
> sample_mean_ac = mean(adultCereal) # sample mean of adult cereal
> sample_variance_cc = var(childCereal) # sample variance of child cereal
> sample_variance_ac = var(adultCereal) # sample variance of adult cereal
> sample_sd_cc = sqrt(sample_variance_cc) # sample standard deviation of child cereal
> sample_sd_ac = sqrt(sample_variance_ac) # sample standard deviation of adult cereal
> alpha = 1 - 0.90
> #calculating CI for ratio of the two variances
> l_critcal_point <- qf(alpha/2, n_cc - 1, n_ac - 1)
> u_critcal_point <- qf(1 - (alpha/2), n_cc - 1, n_ac - 1)
> CI_ratio_variances=((sample_sd_cc/sample_sd_ac)^2) * c(1/u_critcal_point, 1/l_critcal_point)
> CI_ratio_variances #CI for ratio of the two variances
[1] 0.3551741 1.5116458
>
> #calculating satterthwaite's approximation
> v= ((sample_variance_cc/n_cc + sample_variance_ac/n_ac)^2) / ((sample_variance_cc^2/((n_
cc^2)*(n_cc-1))) + (sample_variance_ac^2/((n_ac^2)*(n_ac-1))))
>
> critical_pt = qt(1-alpha/2,v)
> sample_error = sqrt(sample_variance_cc/n_cc + sample_variance_ac/n_ac)
> CI=(sample_mean_cc - sample_mean_ac)+c(-1,1)*(critical_pt*sample_error)
> #90% confidence interval for difference in mean sugar contents of the two cereal types
> CI
[1] 33.18008 40.10225
```

Exercise 3:

```
> n1 = 414 #number of adults who grew up in a single-parent household
> p1_bar = 61/414
> n2 = 501 #number of adults who grew up in a two-parent household
> p2_bar = 74/501
> alpha = 1 - 0.99
> critical_pt = qnorm(1-alpha/2)
> sample_error = sqrt(((p1_bar*(1-p1_bar))/n1) + ((p2_bar*(1-p2_bar))/n2))
> #99% Confidence Interval for the difference between two proportions
> CI=(p1_bar - p2_bar)+c(-1,1)*(critical_pt*sample_error)
> CI
[1] -0.06102961  0.06030642
```