Data Analytics
Mini Project 3

Exercise 1 (10 points):

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.

Exercise 2 (10 points):

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.

Exercise 3 (5 points)
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?

Instructions:

- **Due date: Monday, 10th October, 2016.**
- **Total points = 25**
- Submit a typed report and include relevant plots
- You must use the following template for your report:

  Mini Project #
  Name

  Provide the R codes in an appendix. Your code must be annotated. No points may be given if a brief look at the code does not tell us what it is doing.