Golas, Ekal

UNIVERSITY OF TEXAS AT DALLAS | 800 W CaMPBELL RD RICHARDSON TX 75080

MINI PROJECT 3

Statistical methods for data science



Table of Contents

[Methodology 2](#_Toc433726801)

[Exercise 1 2](#_Toc433726802)

[Exercise 2 2](#_Toc433726803)

[Exercise 3 2](#_Toc433726804)

[Answers 3](#_Toc433726805)

[Exercise 1 3](#_Toc433726806)

[Exercise 2 3](#_Toc433726807)

[Exercise 3 4](#_Toc433726808)

[R Code 6](#_Toc433726809)

[Exercise 1 6](#_Toc433726810)

[Exercise 2 6](#_Toc433726811)

[Exercise 3 7](#_Toc433726812)

[References 8](#_Toc433726813)

# Methodology

As discussed in the problem statement (Choudhary, 2015) , the algorithms for both the exercises are as follows:-

## Exercise 1

The algorithm to implement the solution in R is:-

1. Generate sequences for N and P
2. Define a function to calculate confidence interval, given N and P
   1. Calculate mean and standard deviation from N and P
   2. Simulate a random variable using rnorm function, N times
   3. Get the standard error and mean for this random variable
   4. Using qnorm function, compute the confidence interval for this variable
3. Define a function to calculate coverage probability for a given N and P
   1. Simulate the function in step 2 10000 times to get a confidence interval matrix
   2. Filter the intervals that contain P and return the coverage probability
4. Get all possible combinations of N and P
5. For each combination, compute coverage probability using step 3
6. Plot the result compared to N, P and product of N and P

## Exercise 2

The algorithm to implement the solution in R is:-

1. Define adult`s and children`s cereal data
2. Draw Q-Q plots for both these parameters
3. Do a variance test to check if variances are equal
4. Get the confidence interval using student`s t-distribution

## Exercise 3

The algorithm to implement the solution in R is:-

1. Define probabilities for single and two parent reports
2. Get 95% confidence interval on the difference between two probabilities using the Z distribution

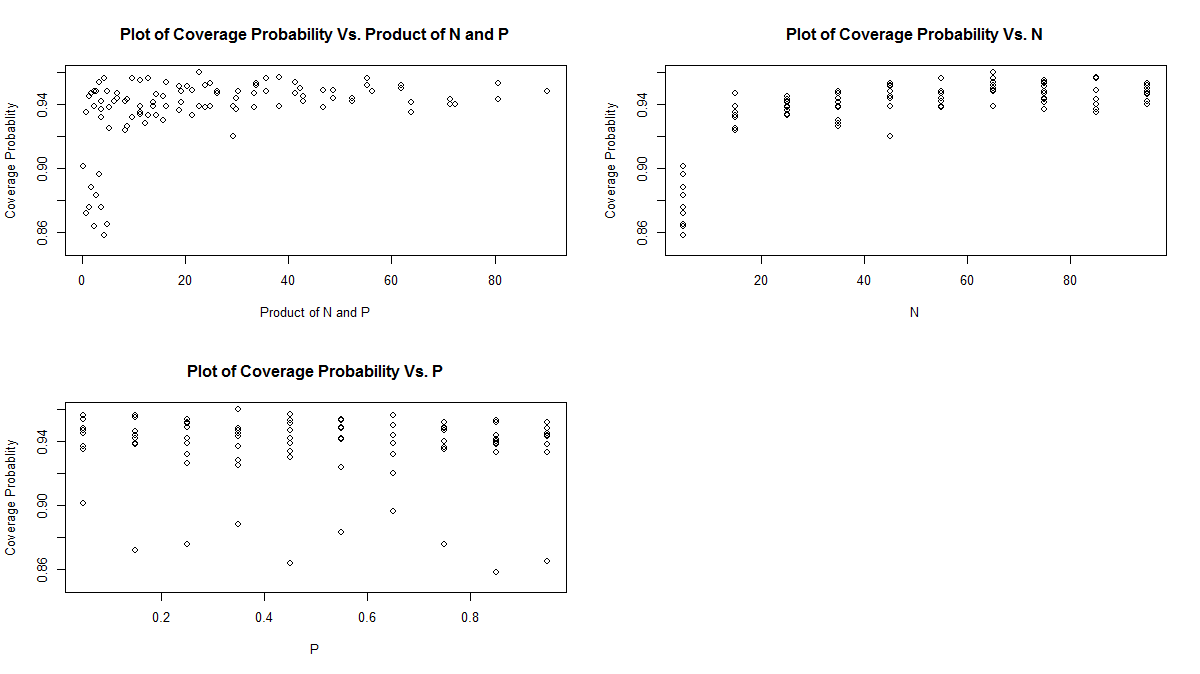
# Answers

## Exercise 1

The answers to the questions stated in (Choudhary, 2015) are as follows:-

1. **Summarize your results graphically.**

The results are displayed as follows:-



1. **Comment on any patterns you see in the results.**

From these results we can observe that:-

* Accuracy increases as N increases
* Accuracy improves greatly when NP > 5
* After a certain value, increase in N has a little/no effect on accuracy
* For large N, the confidence interval is very accurate

1. **Based on your findings, what n would you recommend for the use of this confidence interval?**

From the above results, N > 25 seems a good size as all the data points after this size are in a region of acceptable accuracy

1. **Would your answer depend on p? Explain.**

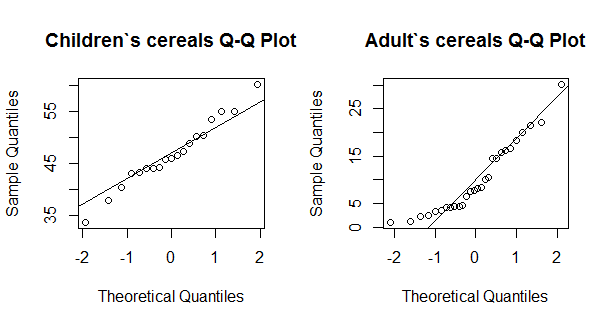
Only when N is small, it is seen to vary with P. If N is large, change in P has little effect. This is because of the equation of margin of error which is dependent on P (1 – P) and N. If N is large, the change in P won’t be noticed as margin of error has become very small.

## Exercise 2

The answers to the questions are stated as follows:-

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

The Q-Q plots are as follows:-



Ignoring the high and the low end, we can assume that each sample comes from normal distribution.

1. **Can the variances of the two distributions be assumed to be equal? Justify your answer.**

Variances of two distributions cannot be assumed as equal. The same is verified by the var.test function which calculates the ratio as 0.71.

1. **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?**

The confidence interval is computed as [32.48894, 40.79339]. To arrive at this answer, the assumptions made were:-

* Data came from a normal distribution
* Since variances were not equal, confidence internal is computed by Student`s t-distribution
* Degrees of freedom are given by Satterthwaite approximation

1. **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.**

We can conclude that children`s cereal on average have more sugar than adult cereals because the confidence interval for difference in mean clearly has a positive lower and upper limit. Also with 95% confidence we can say that children`s cereals will have sugar greater than adult`s cereals by 32.4% to 40.8%.

## Exercise 3

The answers to the questions are stated as follows:-

1. **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.**

The 95% confidence interval for the difference in mean is computed as [-0.04580106, 0.04652425]. As 0 is a part of this interval, we can say that there is no difference is single-parent and two-parent households when it comes to reporting abuse.

1. **What assumptions, if any, did you make to compute the interval in (a)? Do the assumptions seem reasonable?**

The following assumptions were made:-

* Data was assumed to come from a normal distribution. This might not have been valid as there is no such information in the question that suggests this.
* Size of the samples were assumed to be large and hence Z distribution was used to compute the confidence interval for difference in mean. This seems valid if the first assumption holds.

# R Code

## Exercise 1

# Generate N and P

N = seq.int(from = 5, to = 100, by = 10)

P = seq.int(from = 5, to = 95, by = 10) / 100

# Define function to get 95% confidence interval

conf.int = function(N, P) {

# Get interval with this N and P

alpha = 0.05

std.dev = sqrt(P \* (1 - P) / N)

x = rnorm(N, mean = P, sd = std.dev)

ci = mean(x) + c(-1, 1) \* qnorm(1 - alpha / 2) \* sd(x) / sqrt(N)

ci

}

# Define function to get coverage probablity

cov.prob = function(N, P) {

# Simulate 10000 times

ci.mat = replicate(1000, conf.int(N, P))

# Proportion of times interval is correct

mean((P >= ci.mat[1,]) \* (P <= ci.mat[2,]))

}

# Get all possible combinations of N and P

combinations = expand.grid(N, P)

NP = (combinations$Var1 \* combinations$Var2)

# Get coverages for combinations

coverage = apply(combinations, 1, function(x) {cov.prob(x[1], x[2])})

# Plot the coverage against N \* P

par(mfrow = c(2, 2))

plot(NP, coverage, ylim = c(0.85, 0.96), xlab = "Product of N and P", ylab = "Coverage Probablity", main = "Plot of Coverage Probability Vs. Product of N and P")

# Plot coverage against N and P

plot(combinations$Var1, coverage, ylim = c(0.85, 0.96), xlab = "N", ylab = "Coverage Probablity", main = "Plot of Coverage Probability Vs. N")

plot(combinations$Var2, coverage, ylim = c(0.85, 0.96), xlab = "P", ylab = "Coverage Probablity", main = "Plot of Coverage Probability Vs. P")

## Exercise 2

# Define data

children.cereals = 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)

adult.cereals = 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)

# Draw Q-Q plots

par(mfrow = c(1, 2))

qqnorm(children.cereals, main = "Children`s cereals Q-Q Plot")

qqline(children.cereals)

qqnorm(adult.cereals, main = "Adult`s cereals Q-Q Plot")

qqline(adult.cereals)

# Can the variances assumed to be equal?

var.test(children.cereals, adult.cereals)

# Compute an appropriate 95% confidence interval for difference in mean sugar contents of the two cereal types

t.test(children.cereals, adult.cereals)

## Exercise 3

# Define data

single.parent.prob = 61 / 414

two.parent.prob = 74 / 501

alpha = 0.05

# Get standard error

SE = sqrt((single.parent.prob \* (1 - single.parent.prob) / 414) + (two.parent.prob \* (1 - two.parent.prob) / 501))

# Compute 95% confidence interval

c.l.crit = qnorm(alpha / 2) \* SE

c.u.crit = qnorm(1 - alpha / 2) \* SE

c(two.parent.prob - single.parent.prob + c.l.crit, two.parent.prob - single.parent.prob + c.u.crit)