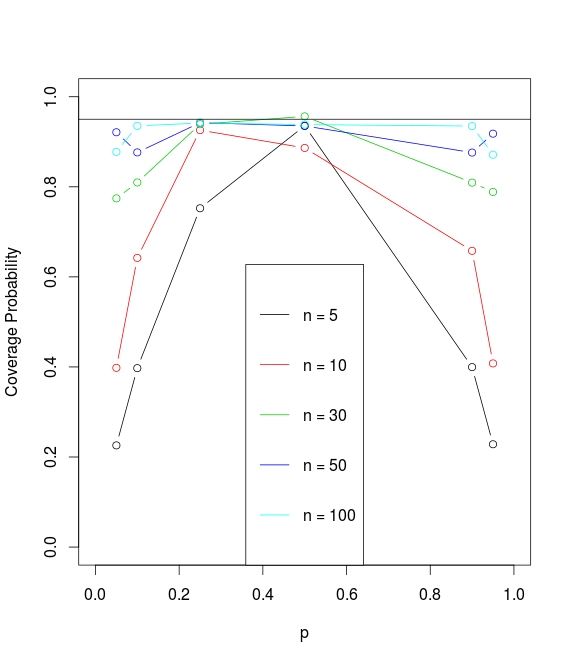
**CS 6313.501**

**Mini Project 2**

**Group members:** Arnav Sharma (axs144130), Divyanshu Paliwal (dxp151630)

**Contribution of each member:** Both members contributed equally to the following tasks:

* Writing the R code
* Generating graphs
* Deriving conclusions

**Exercise 1**

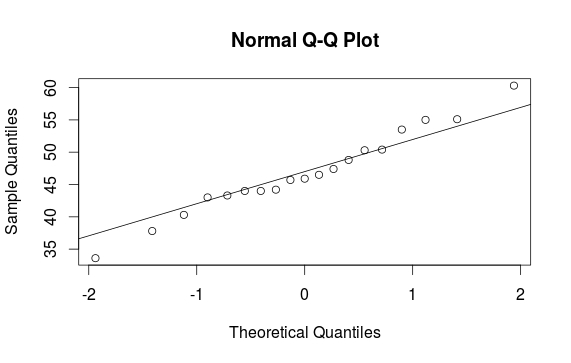
Straight line at coverage probability 0.95 independent of p.

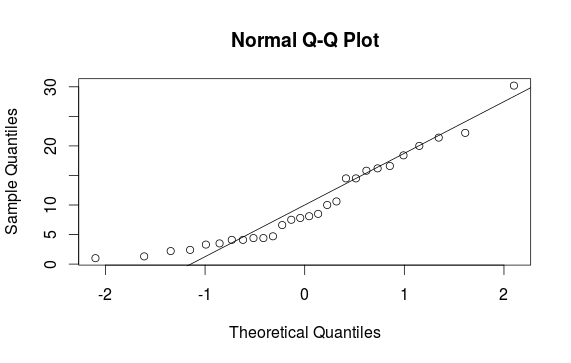
The above graph indicates the following result:

1. For p = 0.5, the coverage probability reaches approximately 95% for each sample size.
2. The coverage probability decreases with value of p tending to either 0 or 1.
3. For n = 100 has the best coverage probability for different values of p, therefore n =100 is the recommended sample size for use of 95% confidence interval.

**Exercise 2**

**2. a)**

Q-Q Plot for Sugar content in Children’s Cereals. The values have equal deviation on both sides of the linear line for the plot, therefore it follows a normal distribution.



Q-Q Plot for Sugar content in Adult's Cereals. The values do not have equal deviation on both sides of the linear line for the plot, therefore it follows some skewed distribution.

**2. b)**

On computing a 95% Confidence Interval for ratio of two variances we get the following values:

> ci

[1] 0.3102977 1.7564758

which indicates that the interval also includes 1. Therefore, we can assume the variance of two distribution to be equal.

**2. c)**

We assumed :

* Normal distribution for both the samples.(as n = 28 for 2nd distribution)
* common variance

**2. d)**

We can conclude based on the result that children’s cereal has more sugar than adult's cereal with 95% confidence interval

> ci

[1] 29.37357 43.90876

**Exercise 3**

**3. a)**

n1 = 414, n2 = 501, p1.hat = 61/414 = 0.1473, p2.hat = 74/501 = 0.1477 . p1.hat ~ p2.hat.

Margin = 0.05 . Hence, 95% confidence interval is [-0.05, 0.05].

* Since p1.hat ~ p2.hat, there is no difference between the two proportions. 95% confidence interval is [-0.05, 0.05].
* Since the confidence interval contains value 0 as well therefore, we can conclude that there is no difference in single parent and two parent household when it comes to reporting child abuse.

**3. b)**

We assumed normal distribution for the data. It seems reasonable because we have a large sample size for both the distributions. Therefore, we can assume the sample data to follow approximately normal distribution.

**Appendix**

**#################################################################**

#Exercise-1

#Computing Coverage Probability for different values of p and sample size.

conf.int <- function(n, p, m=10000, alpha=0.05){

x <- rbinom(m, n, p)

p.hat <-x/n

#ci <- p.hat + c(-1,1)\*qnorm(1 - (alpha/2))\*sqrt(p.hat\*(1-p.hat)/n)

lci <- p.hat - qnorm(1 - (alpha/2))\*sqrt(p.hat\*(1-p.hat)/n)

uci <- p.hat + qnorm(1 - (alpha/2))\*sqrt(p.hat\*(1-p.hat)/n)

return(sum(lci < p & uci> p)/m)

}

conf.vec <- Vectorize(conf.int)

n.vec <- c(5, 10, 30, 50, 100)

p.vec <- c(0.05, 0.1, 0.25, 0.5, 0.9, 0.95)

cov.mat <-outer(n.vec, p.vec, conf.vec)

rownames(cov.mat) <- n.vec

colnames(cov.mat) <- p.vec

print(cov.mat)

plot(NA, xlim = c(0,1), ylim = c(0,1), ylab = "Coverage Probability", xlab = "p")

for(i in seq\_along(n.vec)){

lines(p.vec, cov.mat[i,], type = "b", col = i)

}

abline(h = 0.95)

legend("bottom", col = seq(5), lwd = 1, legend = paste0("n = ", n.vec))

###################################################################

#Exercise 2-a

X = 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)

Y = 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(X)

qqline(X)

qqnorm(Y)

qqline(Y)

##########################################

#Exercise 2-b

X = 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)

Y = 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)

x.sigma <- sd(X, na.rm = FALSE)

y.sigma <- sd(Y, na.rm = FALSE)

x.n <- length(X)

y.n <- length(Y)

alpha = 0.05

ci <- ((x.sigma/y.sigma)^2)\*c(1/(qf(1 - (alpha/2),x.n - 1, y.n - 1)), 1/(qf(alpha/2, x.n - 1, y.n - 1)))

#> ci

#[1] 0.3102977 1.7564758

#########################################

#Exercise 2-c

X = 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)

Y = 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)

x.mean <- mean(X)

y.mean <- mean(Y)

d.mean <- mean(X) - mean(Y)

alpha <- 0.05

x.n <- length(X)

y.n <- length(Y)

x.sigma <- sd(X, na.rm = FALSE)

y.sigma <- sd(Y, na.rm = FALSE)

ssp <- ((x.n - 1)\*(x.sigma)^2 + (y.n - 1)\*(y.sigma)^2)/x.n + y.n -2

sp<- sqrt(ssp)

ci <- d.mean + c(-1, 1) \* qt(1 -(alpha/2), df = (x.n + y.n - 2)) \* sp\*sqrt(1/x.n + 1/y.n)

#> ci

#[1] 29.37357 43.90876

#######################################################################

#Exercise 3

p1 = 61/414

p2 = 74/501

n1 = 414

n2 = 501

mean.diff = p2-p1

margin = 1.96\*sqrt((p1\*(1-p1))/n1+(p2\*(1-p2))/n2)

lower = mean.diff - margin

upper = mean.diff + margin