Homework 4&5 solutions

Shalini Mishra

11/16/2019

##Q1 $H_0 : d_1$ is correct $H_1 : d_2$ is correct Condidering one possible decision procedure to fail to reject H_0 if X=4 or X=6

a the probability of a Type I error H_0 is true,Reject H_0

```
#Probability of when null hypothesis is true and we don't reject null hypothesis
#Prob(X=4|X=6) for d1=P(X=4)+P(X=6)
prob_correct <- 0.2 +0.3
alpha <- 1-prob_correct
cat("alpha:probability of type 1 error",alpha)
```

```
## alpha:probability of type 1 error 0.5
```

b the probability of a Type II error

```
#Probability of when null hypothesis is true and we reject null hypothesis
#Prob(X=4|X=6) for d2=P(X=4)+P(X=6)
beta <- 0.2 +0.1
cat("beta:probability of type 2 error",beta)
```

```
## beta:probability of type 2 error 0.3
```

Q2

ATQ, Hypotheses for our analysis are: - $H_0 : \mu \geq 32$ - H_a : $< 32 $

Please look into attached doc for Type 1, Type 2 error analysis

```
library(PASWR2)
```

```
## Warning: package 'PASWR2' was built under R version 3.5.3
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
q2_data <- read.csv('Question2.csv')
#sample size
n <- 110
x_bar <- mean(q2_data$Minutes)
z_score <- (x_bar - 32)/8/sqrt(110)

prob <- pnorm(0.05,z_score,lower.tail=FALSE)
prob
```

```
## [1] 0.4701823
```

```
#Type 1 error would be a more dangerous in this scenario than type 2
#Hence, we will go for a significance of 5%

#Our sample size >30 and population standard deviation is known.
#Calculate the critical value for rejection region
X_c <- qnorm(0.05,32,8/sqrt(110),lower.tail=TRUE)
X_c
```

```
## [1] 30.74535
```

```
#A financial analyst has determined that a 2-minute reduction in the average break would increas
e productivity
#So considering our mu as 30(=32-2)
beta <- pnorm(X_c,30,8/sqrt(110),lower.tail=FALSE)
cat('beta:',beta) #~16.4%
```

```
## beta: 0.1642429
```

probability of erroneously concluding that the renovation would not be successful= Probability of conducting a type 2 error = β The value of β can be reduced by: 1. increasing the sample size, n 2. increasing the value of α

Q3

It is an example of one-way ANOVA We have price as a factor Levels: Different prices provided i.e. 9, 10 and 11$ Since, it's a one way ANOVA, our treatments are same as levels (20 stores under each) - H_0 : $9 = {10} =_{11} $ - H_a : Atleast one of $\mu_9, \mu_{10}, \mu_{11}$ is different

The manager is unsure if the sale would differ significantly among these 3 prices (levels) stated above

```
#sample means
x_9 <- 153.60
x_10 <- 151.50
x_11 <- 133.25

#sample sizes
n_9 <- 20
n_10 <- 20
n_11 <- 20

#standard deviations
s9 <- 25.57
s10 <- 30.39
s11 <- 25.03

#grand_mean
x_bar <- (x_9*n_9 +x_10*n_10 +x_11*n_11)/(n_9+n_10+n_11)
x_bar
```

```
## [1] 146.1167
```

```
#compute SST -Sample Square for Treatments (between treatment variation)
sst <- n_9*(x_9-x_bar)^2 + n_10*(x_10-x_bar)^2 + n_11*(x_11-x_bar)^2
sst
```

```
## [1] 5010.633
```

```
#compute SSE -Sample Square for Errors (within treatment variation)
sse <- (n_9-1)*(s9^2)+(n_10-1)*(s10^2)+(n_11-1)*(s11^2)
sse
```

```
## [1] 41873.68
```

```
#number of treatments,k(9$,10$,11$)
k <- 3

#compute MSE and MST
mst <- sst/(k-1)
cat("mst=",mst)
```

```
## mst= 2505.317
```

```
mse <- sse/(60-k)
cat("mse=",mst)
```

```
## mse= 2505.317
```

```
#compute F statistic
f <- mst/mse
f
```

```
## [1] 3.41033
```

```
#We reject null hypothesis if our f-statistic,f > f-statistic(0.05,k-1=2,n-k=60-3=57)
qf(0.05, df1=2, df2=60-k, lower.tail=FALSE)
```

```
## [1] 3.158843
```

```
pf(f,2,57,lower.tail=FALSE)
```

```
## [1] 0.03990587
```

As the observed F statistic is greater than 3.16,f-statistic(0.05,2,57) & p-value<0.05 -> Reject H0 At 5% significance level, there is sufficient evidence to support the claim that the sales vary significantly across these three different prices Manager was unsure if the sale would vary significantly. But the results from ANOVA test tells otherwise.

Q4

a. It is a case of one-way ANOVA We have cereals as a factor Levels: Different cereals provided i.e. 1,2,3 & 4 Since, it's a one way ANOVA, our treatments are same as levels For mean age, - H_0 : $*{1}* = {2}= *3* ={4} $ - H_a : Atleast one of μ_1,μ_2,μ_3,μ_4 is different

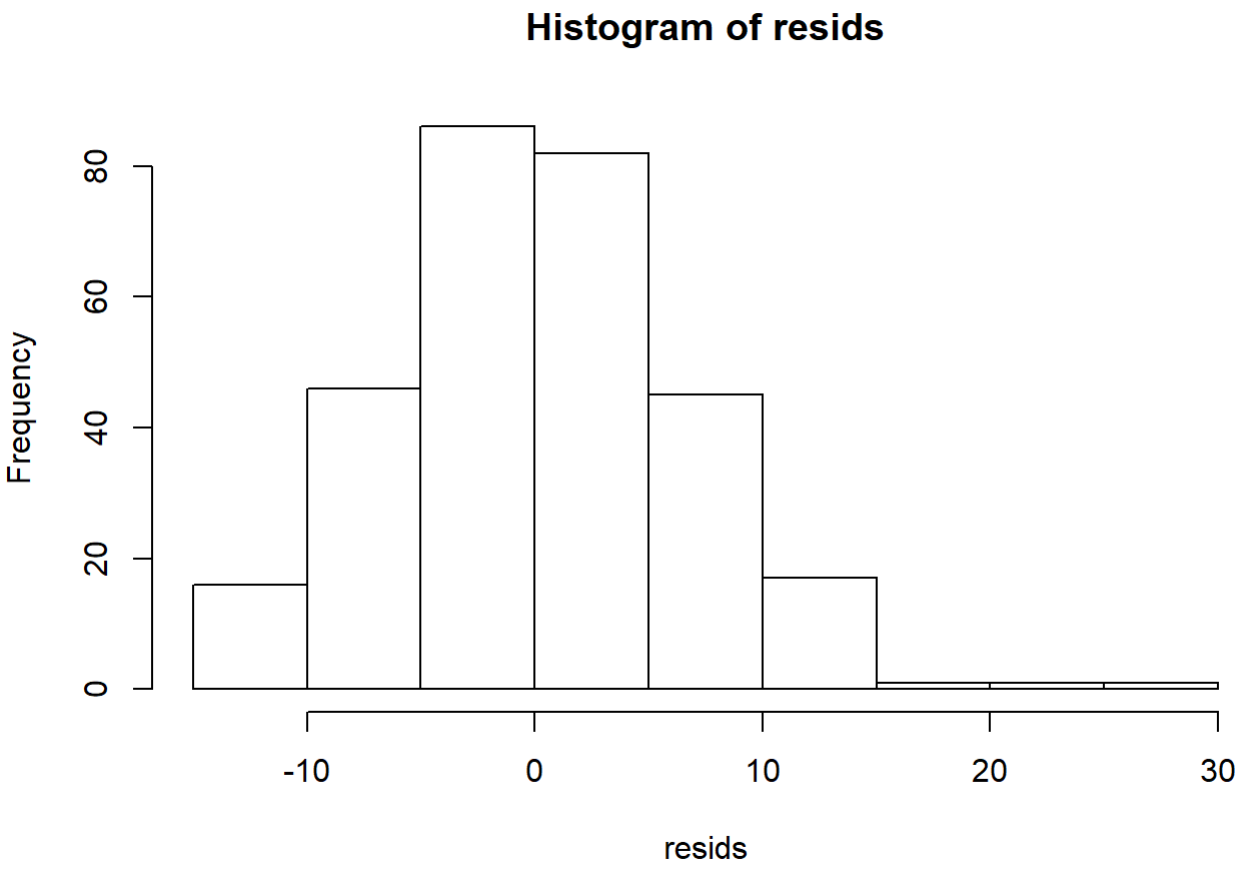
```
q4 <- read.csv('Question4.csv')
#Changing data type for cereal type to factors
q4$Cereal <- factor(q4$Cereal)
#Run one-way ANOVA
ANOVA_4 <- aov(q4$Age ~ q4$Cereal)
summary(ANOVA_4)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## q4$Cereal     3   3366  1122.0    25.6 9.96e-15 ***
## Residuals   291  12752    43.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Check for normality
#Creating a residual plot
model <- lm(q4$Age ~ q4$Cereal)
resids <- residuals(model)

nortest::ad.test(resids)
```

```
##
##  Anderson-Darling normality test
##
## data:  resids
## A = 0.74779, p-value = 0.051
```

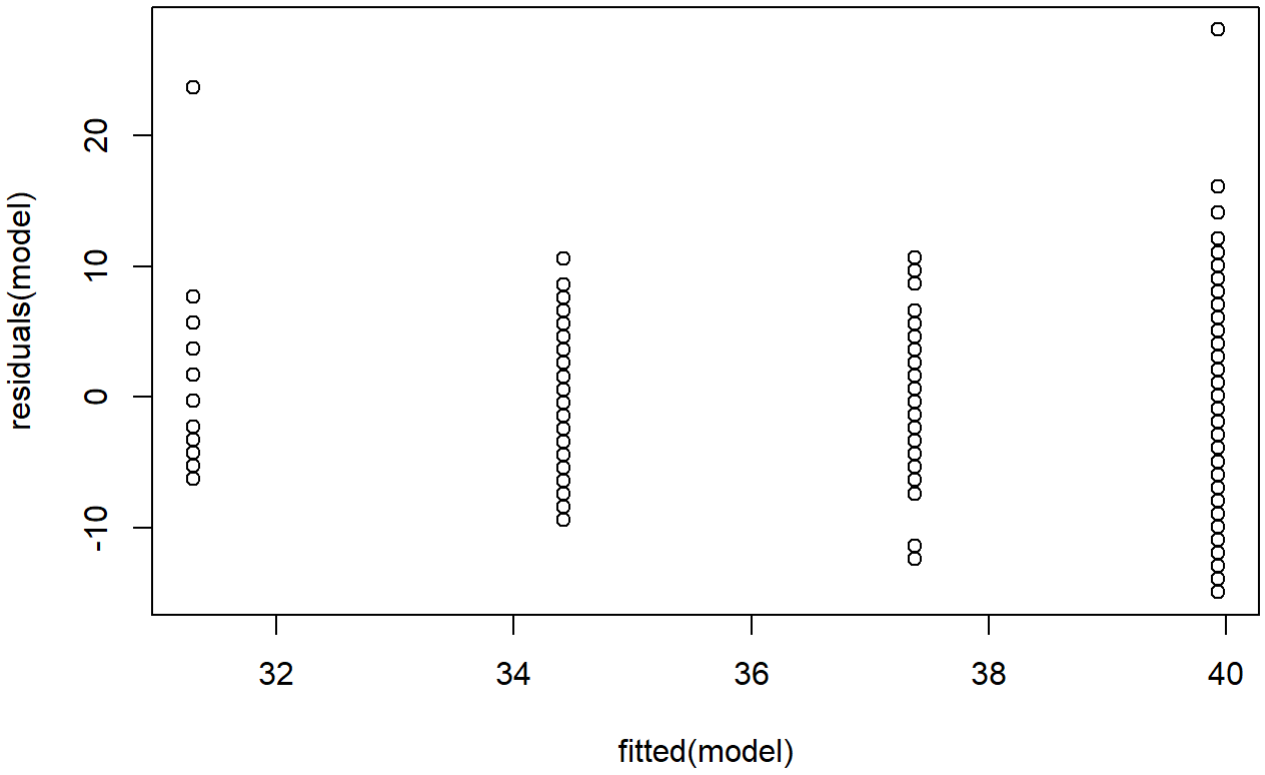
```
hist(resids)
```




```
#p-value>0.05 Failed to reject
#Histogram also resembles (right-skewed) distribution


#Check for variance
car::leveneTest(q4$Age ~ q4$Cereal)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)
## group   3  12.087 1.77e-07 ***
##       291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Plot for variance test
plot(fitted(model), residuals(model))
```




```
#p-value<0.05 --rejecting the null hypothesis of constant variance
#The plot for variance also shows that the variance is increasing not constant

#As normality test passed but variance test failed, we will go for Welch test
oneway.test(q4$Age ~ q4$Cereal, var.equal = FALSE)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  q4$Age and q4$Cereal
## F = 25.154, num df = 3.00, denom df = 132.24, p-value = 6.11e-13
```

As p-value<<0.05, We reject null hypothesis stating that there is no difference in age of the consumers of the four cereals. At 5% significance level, there is sufficient evidence to support the claim that the age vary significantly across the consumers of the 4 different cereals

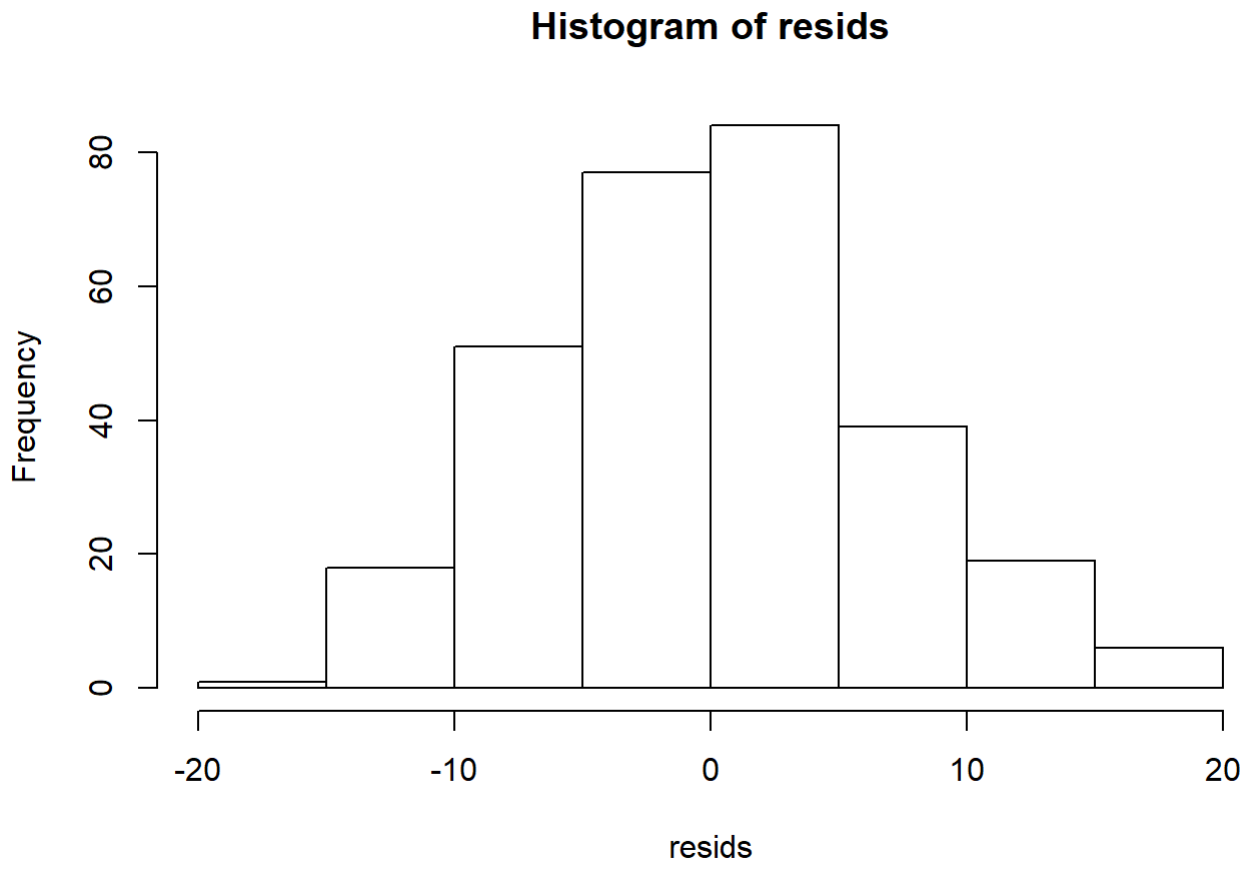
b

It is a case of one-way ANOVA We have cereals as a factor Levels: Different cereals provided i.e. 1,2,3 & 4 Since, it's a one way ANOVA, our treatments are same as levels For mean income, - H_0 : $*{1}* = {2}= *3* ={4} $ - H_a : Atleast one of μ_1,μ_2,μ_3,μ_4 is different

```
#Check for normality
#Creating a residual plot
model <- lm(q4$Income ~ q4$Cereal)
resids <- residuals(model)

nortest::ad.test(resids)
```

```
##
##  Anderson-Darling normality test
##
## data:  resids
## A = 0.38897, p-value = 0.3826
```

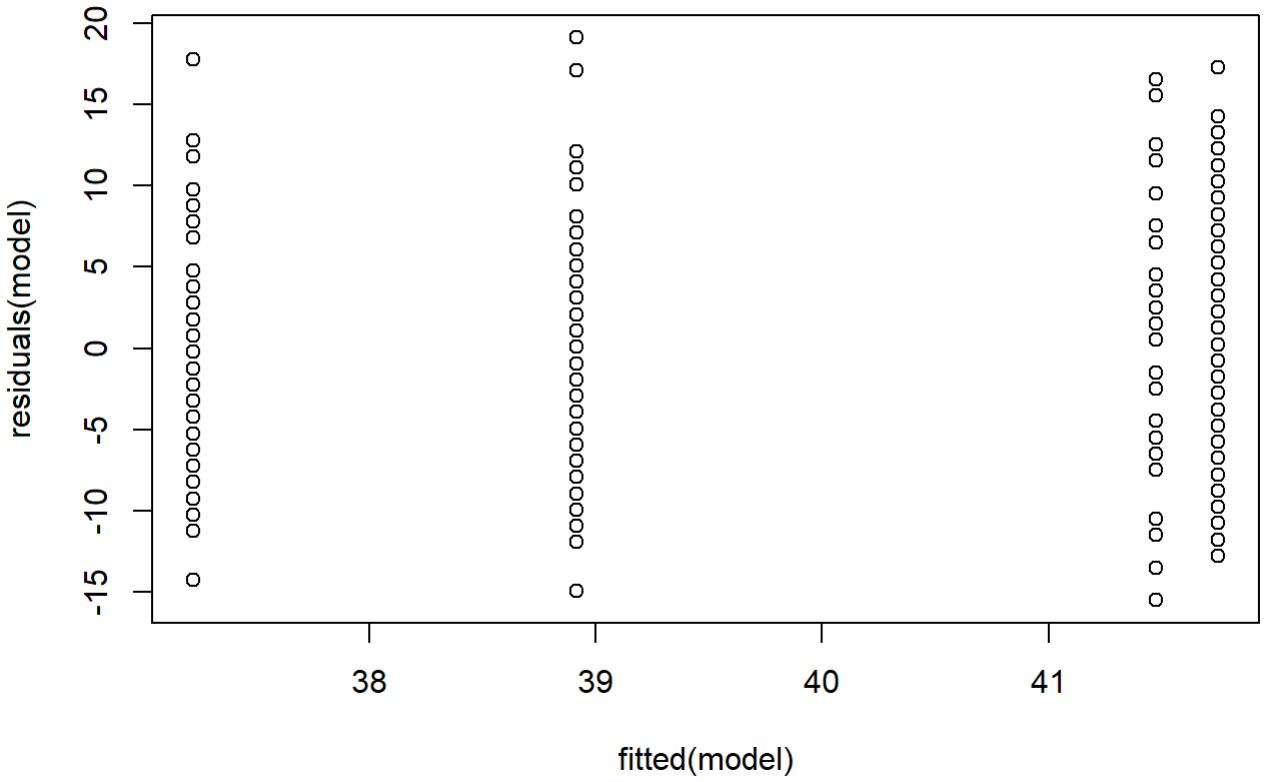
```
hist(resids)
```




```
#p-value>>0.05-satisfies normality
#Histogram also resembles normal distribution like plot

#Check for variance
car::leveneTest(q4$Income ~ q4$Cereal)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   3   1.302  0.274
##       291
```

```
#Plot for variance test
plot(fitted(model), residuals(model))
```



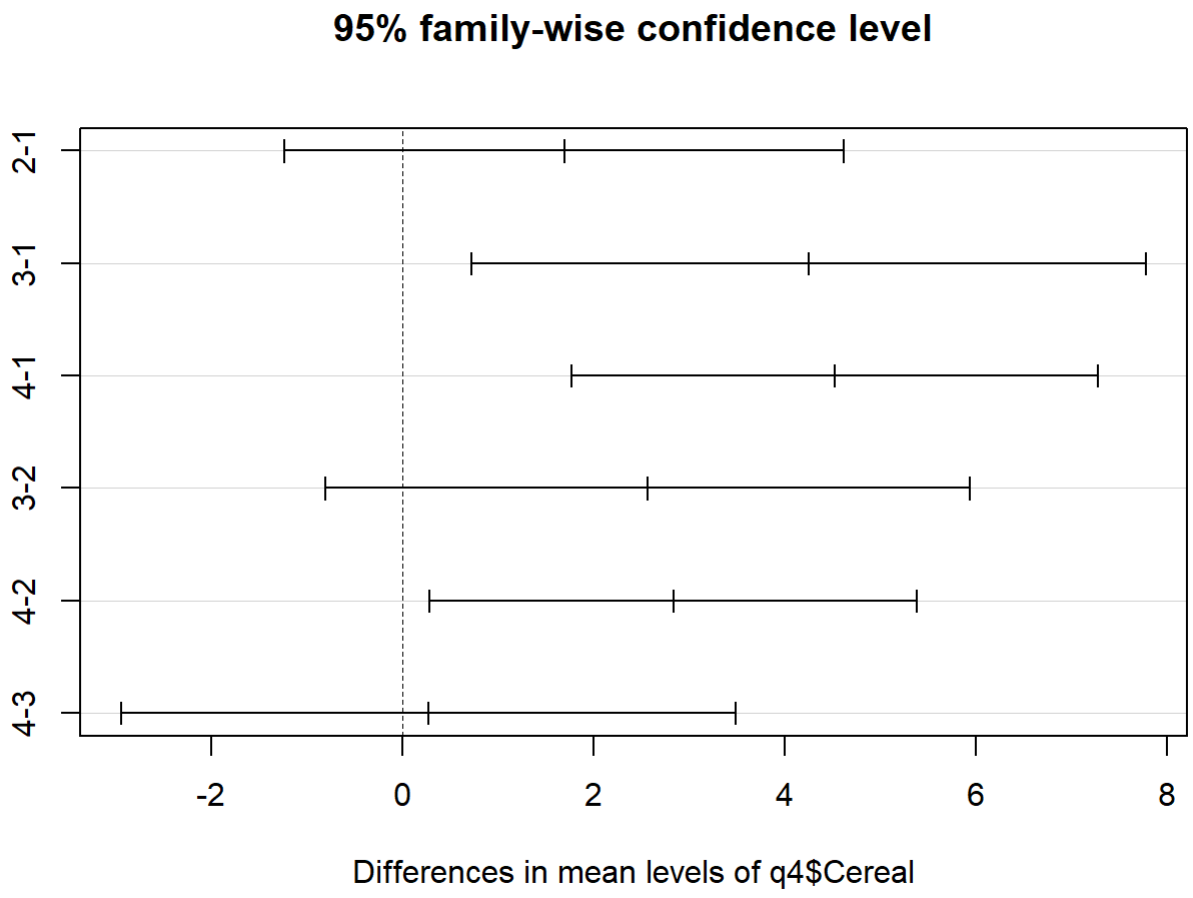

```
#p-value>0.05 --failing to reject null hypothesis:constant variance
#The plot for variance also shows that the variance is constant

#As both normality test & variance test were passed
#Run one-way ANOVA
ANOVA_4 <- aov(q4$Income ~ q4$Cereal)
summary(ANOVA_4)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## q4$Cereal     3   1007   335.8   7.372 8.9e-05 ***
## Residuals   291  13256    45.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#As we reject null hypothesis, we proceed with tukey
TukeyHSD(ANOVA_4)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = q4$Income ~ q4$Cereal)
##
## $`q4$Cereal`
##          diff        lwr      upr     p adj
## 2-1 1.6913580 -1.2382761 4.620992 0.4437664
## 3-1 4.2527778  0.7269324 7.778623 0.0107801
## 4-1 4.5255255  1.7745433 7.276508 0.0001679
## 3-2 2.5614198 -0.8088534 5.931693 0.2042006
## 4-2 2.8341675  0.2856270 5.382708 0.0225024
## 4-3 0.2727477 -2.9434438 3.488939 0.9962792
```

```
plot(TukeyHSD(ANOVA_4))
```



As p-value<<0.05, We reject null hypothesis stating that there is no difference in Income of the consumers of the four cereals. At 5% significance level, there is sufficient evidence to support the claim that the income vary significantly across the consumers of the 4 different cereals

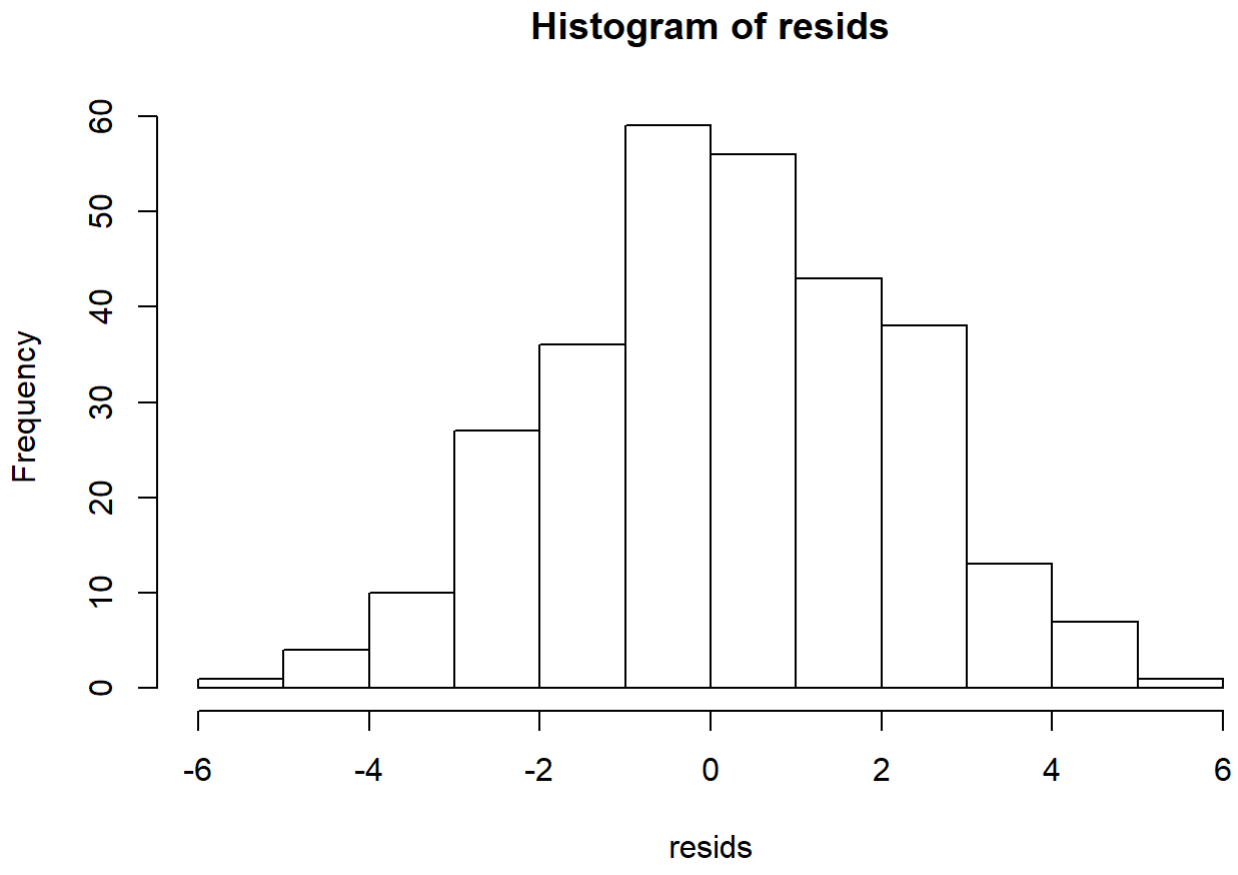
After which we conduct tukey test, to look which pairs have significant difference between each other Pair of Cereal groups 1&3, 1&4,2&4 have significant difference in their ratings

c. It is a case of one-way ANOVA We have cereals as a factor Levels: Different cereals provided i.e. 1,2,3 & 4 Since, it's a one way ANOVA, our treatments are same as levels For mean education_level, - H_0 : $*{1}* = {2}= *3* = {4} $ - H_a : Atleast one of $\mu_1, \mu_2, \mu_3, \mu_4$ is different

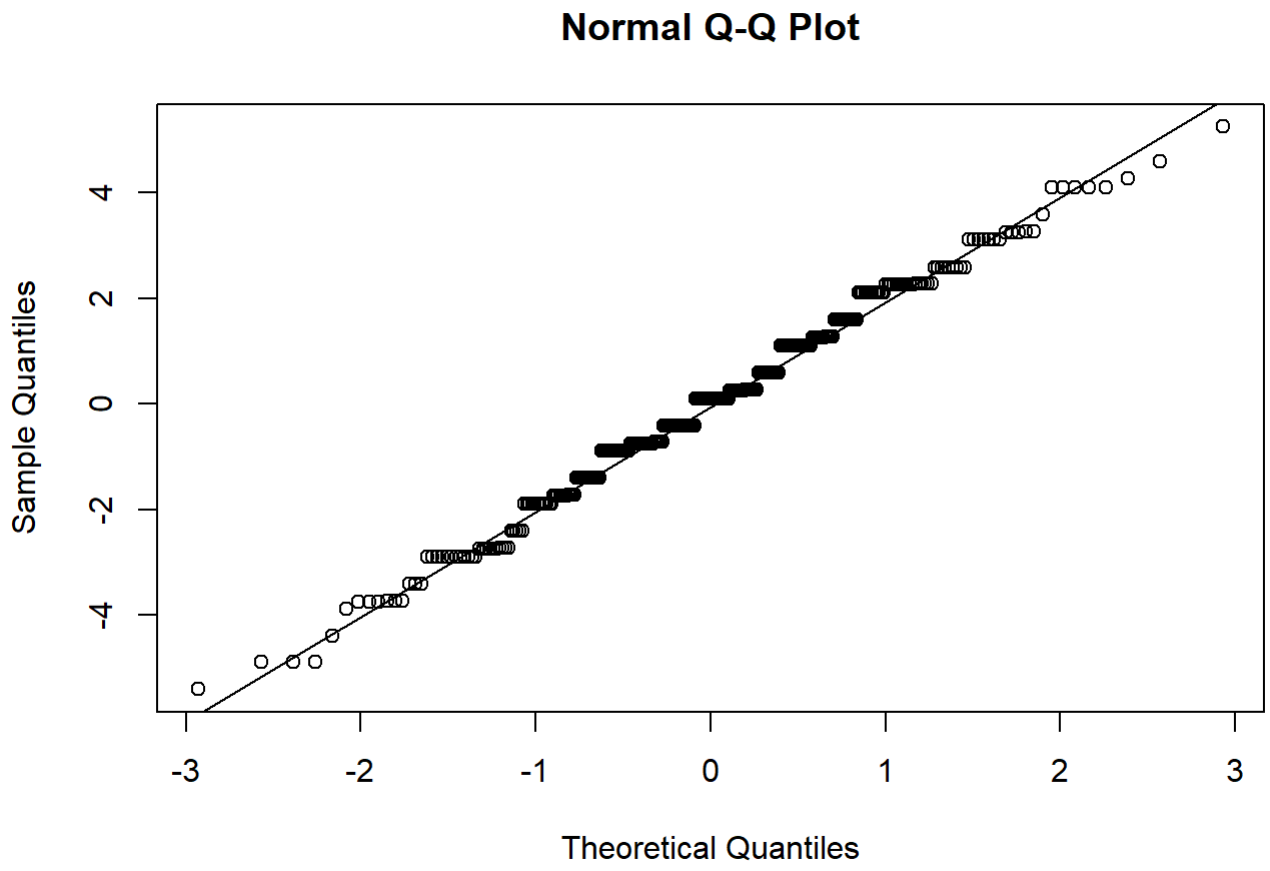
```
#Check for normality
#Creating a residual plot
model <- lm(q4$Education ~ q4$Cereal)
resids <- residuals(model)

nortest::ad.test(resids)
```

```
##
##  Anderson-Darling normality test
##
## data:  resids
## A = 0.87762, p-value = 0.02435
```

```
hist(resids)
```



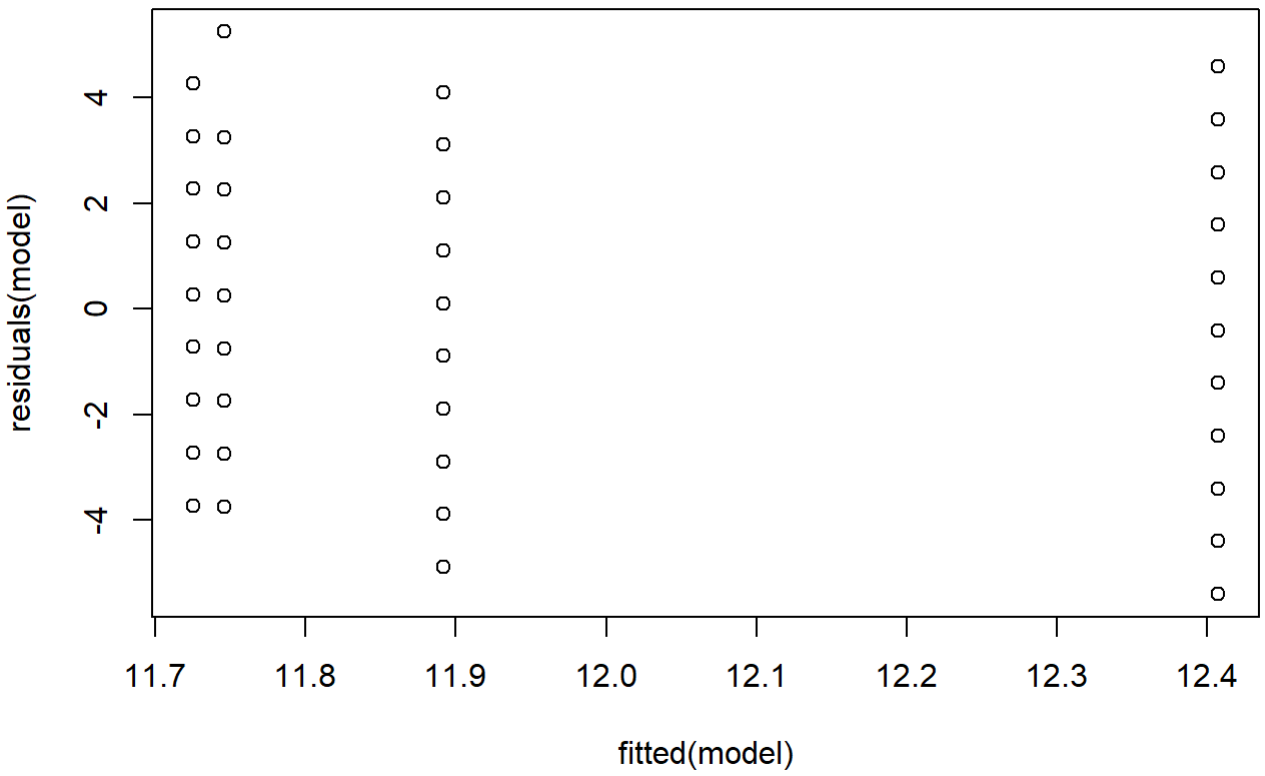

```
qqnorm(resids)
qqline(resids)
```




```
#Based on formal test p-value<0.05-does not satisfy normality
#Hence, considering that normality is maintained

#Check for variance
car::leveneTest(q4$Education ~ q4$Cereal)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   3  0.5156 0.6718
##       291
```

```
#Plot for variance test
plot(fitted(model), residuals(model))
```




```
#p-value>0.05 --failing to reject null hypothesis:constant variance
#The plot for variance also shows that the variance is approximately constant

#As normality test failed and equal variance test passed
#We will use a non-parametric test
kruskal.test(q4$Education ~ q4$Cereal)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  q4$Education by q4$Cereal
## Kruskal-Wallis chi-squared = 5.666, df = 3, p-value = 0.129
```

As p-value>0.129, We failed to reject the null hypothesis stating that there is no difference in education levels of the consumers of the four cereals. At 5% significance level, we lack evidence to support the claim that the education level vary significantly across the consumers of the 4 different cereals *d*

```
##On average, Cereal 2 is better than 1 by 3.11 years
##On average, Cereal 3 is better than 1 by 6.07 years
##On average, Cereal 4 is better than 1 by 8.6 years
##On average, Cereal 4 is better than 2 by 5.5 years
##### Income
###On average, Cereal 3 is better than 1 by $4.2k
###On average, Cereal 4 is better than 1 by $4.5k
###On average, Cereal 4 is better than 2 by $2.8k
##### Education
###There is no difference in education of consumers
```

Q5

There are five groups in total. The number of pounds that each person lost by following one of the four diets were recorded. Each of these 5 groups are blocks. This is a RBD (Randomized Block Design) Diet types 1,2,3 & 4 are treatments

```
#number of treatments
k <- 4
#number of blocks
b <- 5
#treatment-means
#treatment mean for diet 1, and so on
x_trt1 <- 6.2
x_trt2 <- 8.0
x_trt3 <- 10.8
x_trt4 <- 8.2

#block means
#block mean for first block, second block and so on
x_b1 <- 5.25
x_b2 <- 7.25
x_b3 <- 7.25
x_b4 <- 10.25
x_b5 <- 11.50

#grand-mean
x_bar <- ((x_b1+x_b2+x_b3+x_b4+x_b5)*4)/20
# OR x_bar <- ((x_trt1+x_trt2+x_trt3+x_trt4)*5)/20
print('grand mean:')
```

```
## [1] "grand mean:"
```

```
x_bar
```

```
## [1] 8.3
```

```
#Sum of Squares for the given RBD
SST <- 5*((x_trt1-x_bar)^2+(x_trt2-x_bar)^2 +(x_trt3-x_bar)^2 +
            (x_trt4-x_bar)^2)
SSB <- 4*((x_b1-x_bar)^2+(x_b2-x_bar)^2+(x_b3-x_bar)^3+
            (x_b4-x_bar)^2+(x_b5-x_bar)^2)
SSE <- (5 - x_trt1 - x_b1 + x_bar)^2 +
       (4 - x_trt1 - x_b2 + x_bar)^2 +
       (6 - x_trt1 - x_b3 + x_bar)^2 +
       (7 - x_trt1 - x_b4 + x_bar)^2 +
       (9 - x_trt1 - x_b5 + x_bar)^2 +
       (2 - x_trt2 - x_b1 + x_bar)^2 +
       (7 - x_trt2 - x_b2 + x_bar)^2 +
       (12 - x_trt2 - x_b3 + x_bar)^2 +
       (11 - x_trt2 - x_b4 + x_bar)^2 +
       (8 - x_trt2 - x_b5 + x_bar)^2 +
       (6 - x_trt3 - x_b1 + x_bar)^2 +
       (8 - x_trt3 - x_b2 + x_bar)^2 +
       (9 - x_trt3 - x_b3 + x_bar)^2 +
       (16 - x_trt3 - x_b4 + x_bar)^2 +
       (15 - x_trt3 - x_b5 + x_bar)^2 +
       (8 - x_trt4 - x_b1 + x_bar)^2 +
       (10 - x_trt4 - x_b2 + x_bar)^2 +
       (2 - x_trt4 - x_b3 + x_bar)^2 +
       (7 - x_trt4 - x_b4 + x_bar)^2 +
       (14 - x_trt4 - x_b5 + x_bar)^2

# Mean Squares for RBD
MST <- SST/(k-1)
MST
```

```
## [1] 17.93333
```

```
MSB <- SSB/(b-1)
MSB
```

```
## [1] 23.28987
```

```
MSE <- SSE/((b*k)-b-k+1)
MSE
```

```
## [1] 10.85
```

For treatment Effect: - H_0 : ${1} = {2}= 3 ={4} $ - H_a : Atleast one of $\mu_1, \mu_2, \mu_3, \mu_4$ is different For Block Effect: - H_0 : ${1}={2}=3 ={4}=_5 $ - H_a : Atleast one of $\delta_1, \delta_2, \delta_3, \delta_4, \delta_5$ is different

```
#Test Treatment Effect
F_trt <-MST/MSE
F_trt
```

```
## [1] 1.652842
```

```
#Compute corresponding p-value
pf(q=F_trt, df1=3, df2=12, lower.tail=FALSE)
```

```
## [1] 0.2296292
```

```
#Or compute the F statistic for alpha=0.01 and degrees of freedom using df()
qf(0.01,3,12,lower.tail=FALSE)
```

```
## [1] 5.952545
```

```
#Test Block Effect
F_Block <- MSB/MSE
F_Block
```

```
## [1] 2.146532
```

```
#Compute corresponding p-value
pf(q=F_Block, df1=4, df2=12, lower.tail=FALSE)
```

```
## [1] 0.1374097
```

```
#Or compute the F statistic for alpha=0.01 and degrees of freedom using df()
qf(0.01,4,12,lower.tail=FALSE)
```

```
## [1] 5.411951
```

Since, F_Block<f-statistic(alpha=0.01,4,12) & p-value >0.01 =>failed to reject H_0 for *block effect* Similarly, F_trt<f-statistic(alpha=0.01,3,12) & p-value>0.01=>failed to reject H_0 for *treatment effect* At 1% significance level, we lack evidence to support our claim that there is a difference among the four diets both treatment effect and block effect. From my analysis, I think this design was not sound enough because there is no block effect which means there was no additional value added to our analysis by creating blocks -> blocks were not meaningful

Q6

The experimental design used is a one-factor randomized block design Factor is the type of graduate Levels are business, art and science graduate Block is group of the 3 graduates with similar educational and working experience

For treatment Effect (graduate type):
- H_0 : ${bu} = {art}= _{sci} $
- H_a : Atleast one of μ_{bu},μ_{art},μ_{sci} is different

For Block Effect:
- H_0 : ${1}={2}=3 =….=20} $ - H_a : Atleast one of $\delta_1, \delta_2, \delta_3, . . \delta_{20}$ is different

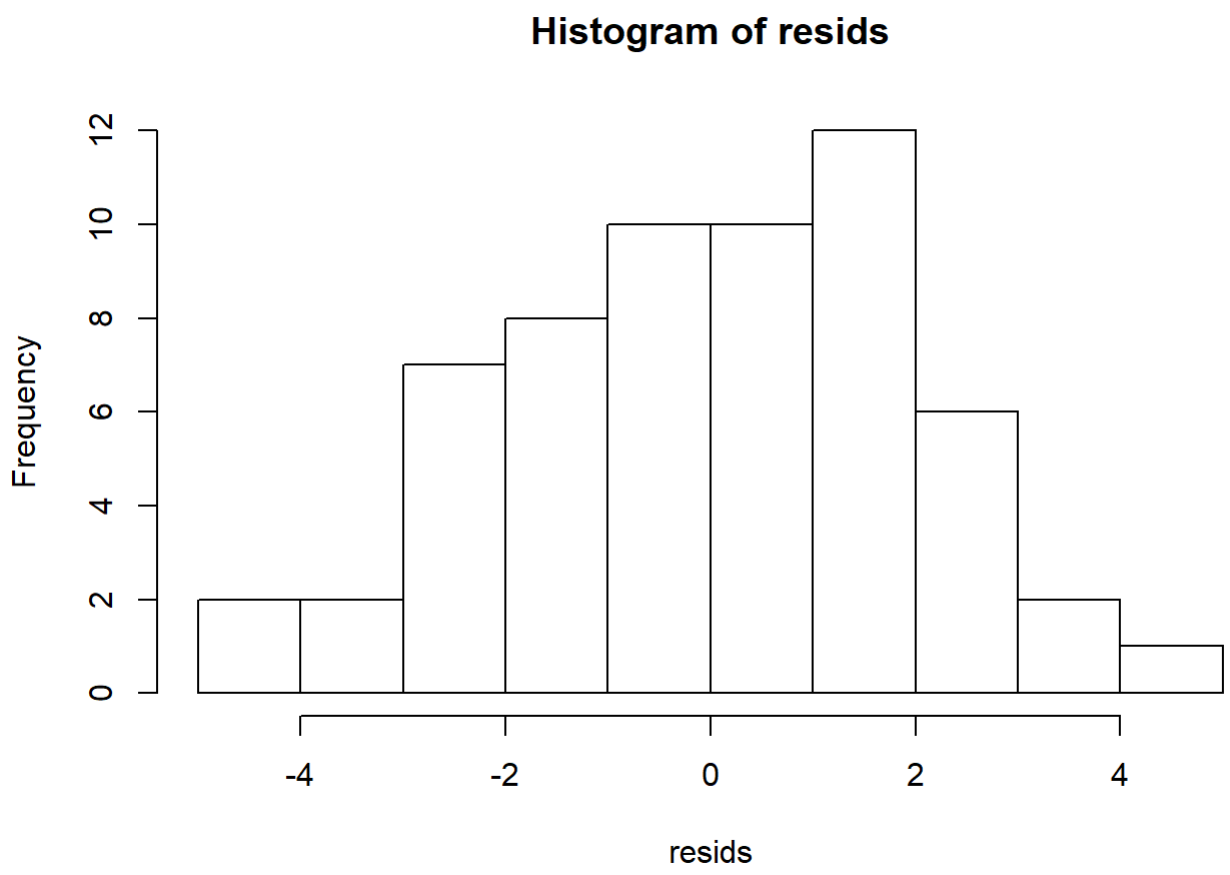
```
#The dataset provided is not stacked format
q6 <- read.csv('Question6.csv')

#Reloading the same data in stacked format
q6 <- read.csv('Question6_stacked.csv')
#View(q6)

Sales <- q6$Sales
Grad <- factor(q6$Graduate)
Group <- factor(q6$Group)

#Check for normality
#Creating a residual plot
model <- lm(Sales ~ Grad + Group)
resids <- residuals(model)

nortest::ad.test(resids)
```

```
##
##  Anderson-Darling normality test
##
## data:  resids
## A = 0.23463, p-value = 0.7838
```

```
shapiro.test(resids)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resids
## W = 0.99054, p-value = 0.9242
```

```
hist(resids)
```



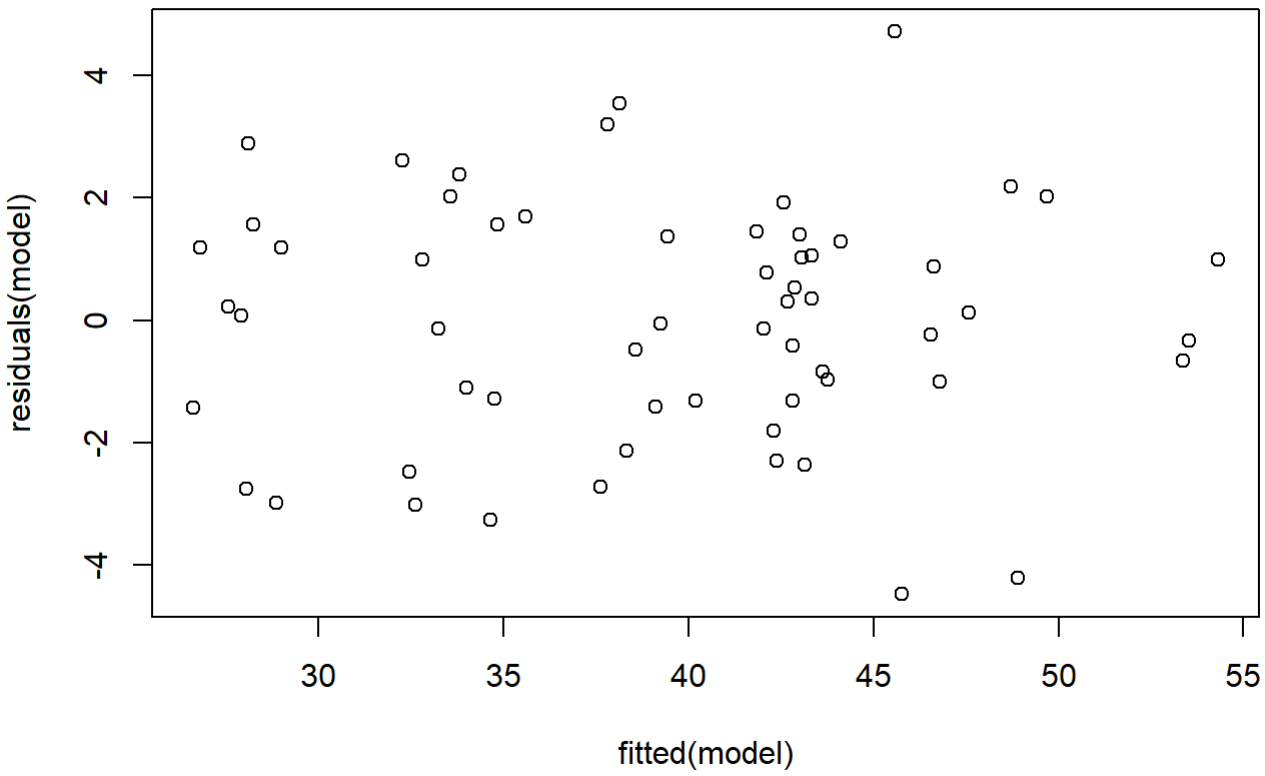

```
#p-value>>0.05-satisfies normality
#Histogram also resembles close to normal distribution like plot

#Check for variance
car::leveneTest(Sales ~ Grad)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.4146 0.6626
##       57
```

```
car::leveneTest(Sales ~ Group)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group 19   0.456 0.9658
##       40
```

```
#Plot for variance test
plot(fitted(model), residuals(model))
```



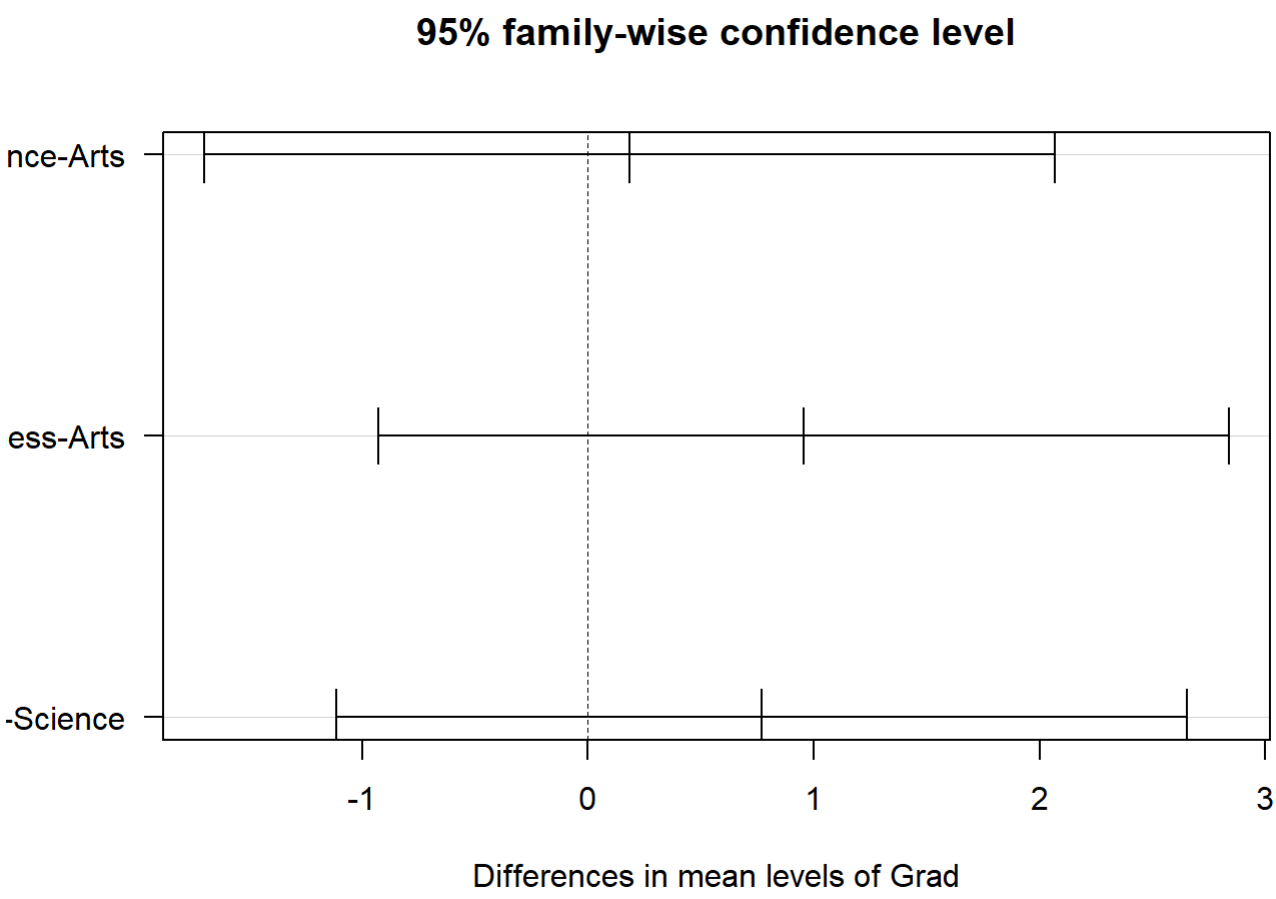

```
#p-value >>0.05 for both formal tests-> constant variance is maintained
#Visual anlaysis also shows approx. constant variance throughout

#Both normality and variance tests have been cleared
#Now we proceed with ANOVA for RBD

ANOVA6 <- aov(Sales ~ Grad + Group)
summary(ANOVA6)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Grad         2   10.3    5.13    0.86   0.431
## Group       19 3020.3  158.96   26.64 2.4e-16 ***
## Residuals   38  226.7    5.97
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

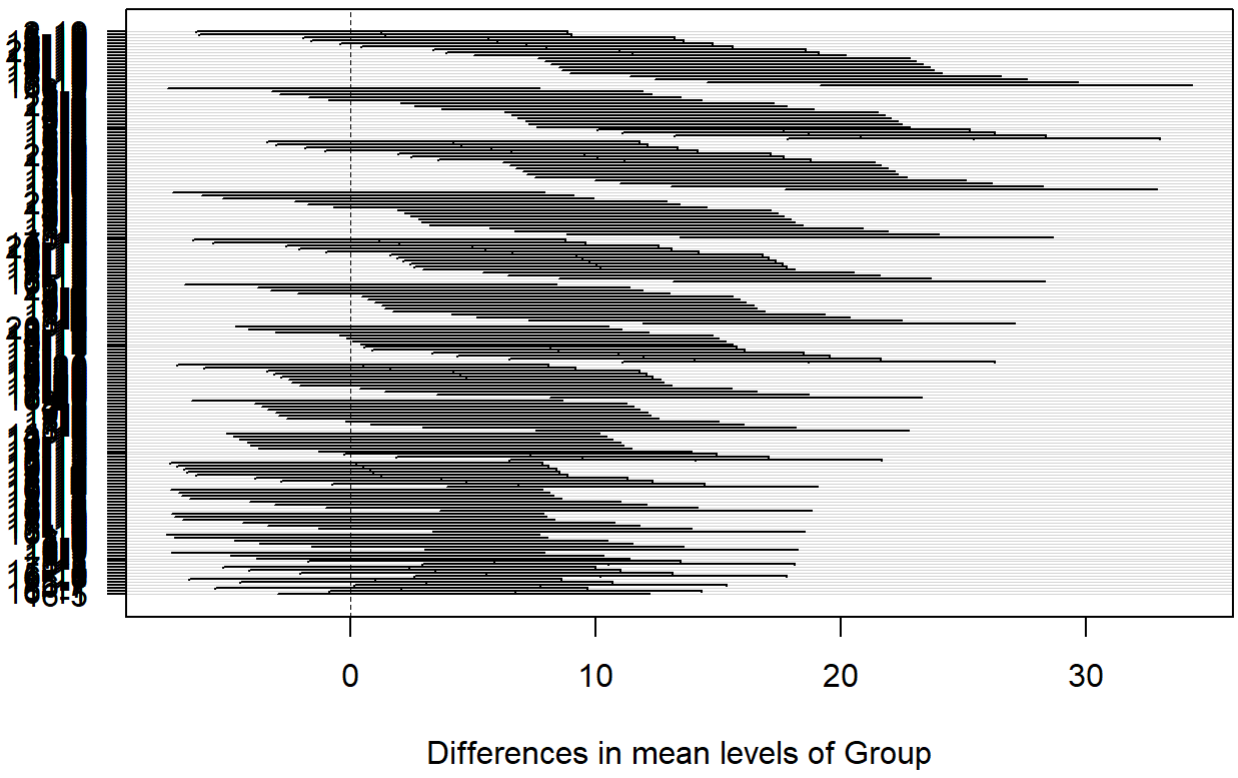
```
TukeyHSD(ANOVA6, which = 'Grad', ordered = TRUE)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##     factor levels have been ordered
##
## Fit: aov(formula = Sales ~ Grad + Group)
##
## $Grad
##                   diff        lwr      upr     p adj
## Science-Arts     0.185 -1.6987645 2.068764 0.9688969
## Business-Arts    0.955 -0.9287645 2.838764 0.4395209
## Business-Science 0.770 -1.1137645 2.653764 0.5833203
```

```
plot(TukeyHSD(ANOVA6, which = 'Grad', ordered = TRUE), las = 1)
```




```
#TukeyHSD(ANOVA6, which = 'Group', ordered = TRUE) --too long a table

plot(TukeyHSD(ANOVA6, which = 'Group', ordered = TRUE), las = 1) #to give a visual idea
```



a. As, p-value for treatment effect > 0.05 => Failed to reject H_0 At 5% significance level, we don't have sufficient evidence to conclude that there are differences in sales ability between the holders of different graduation types As, p-value for block effect << 0.05 => Reject H_0 There is a significant difference between the average of sales ability among blocks (groups 1-20) At 5% significance level, we have sufficient evidence to conclude that there are differences in sales ability between employees with different working and educational experience

b As Block effect turned out to be significant as mentioned above p-value<<0.05 Thus, blocks were meaningful RBD was a good experiment design selected for this analysis

```
#One factor ANOVA with graduate types
ANOVA6 <- aov(Sales ~ Grad)
summary(ANOVA6)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Grad         2     10    5.13    0.09  0.914
## Residuals   57   3247   56.97
```

c. Required consitions to conduct test in part a. As seen in the analysis done above, For Normality: a.Plotted a Histogram b.Conducted following formal tests-> Anderson-Darling test, Shapiro-Wilk test For Constant Variance: a.Levene'sTest of Equality of Variances
b.Plot residuals and predicted values

All these tests were passed, As a result of which we conducted one-factor RBD Anova test

Q7

The experimental design used is a two-factor ANOVA design Factors are the age groups and gender

For Effect between the levels of Age Groups:
- H_0 : ${age1} = {age2}=… =_{age5} $
- H_a : Atleast one of $\mu_{age1}, \mu_{age2}, \mu_{age3}$ is different

For Effect between the levels of Gender:
- H_0 : ${M}={F} $ - H_a : μ_M, μ_F are different

For Effect of interaction between age and gender: - H_0 : ${M:Age1}={M:Age2}=..={M:Age5}={F:Age1}=… =_{F:Age5} $ - H_a : Atleast one of them is different

```
q7 <- read.csv("Question7.csv")
head(q7)
```

```
##       X Age.1  Age.2  Age.3  Age.4  Age.5
## 1 Males 7,707 18,361 22,577 18,541 13,193
## 2       7,916 17,179 17,251 18,071 10,547
## 3       7,859 20,785 16,207 14,020 12,090
## 4       6,908 19,566 21,694 13,545 11,307
## 5       8,733 20,044 18,667 14,023 10,520
## 6       8,601 15,740 25,967 14,097 11,311
```

```
#Provided data is not stacked, we upload in stacked format
q7 <- read.csv("Question7.csv",na.strings=c('',"  ","NA"))

q7_stacked <- reshape2::melt(q7,id.vars=1)#,measure.vars=c('Age.1','Age.2','Age.3','Age.4','Age.
5'))
```

```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```

```
colnames(q7_stacked) <- c('Gender','Age','Miles')

library(stringr)
```

```
## Warning: package 'stringr' was built under R version 3.5.3
```

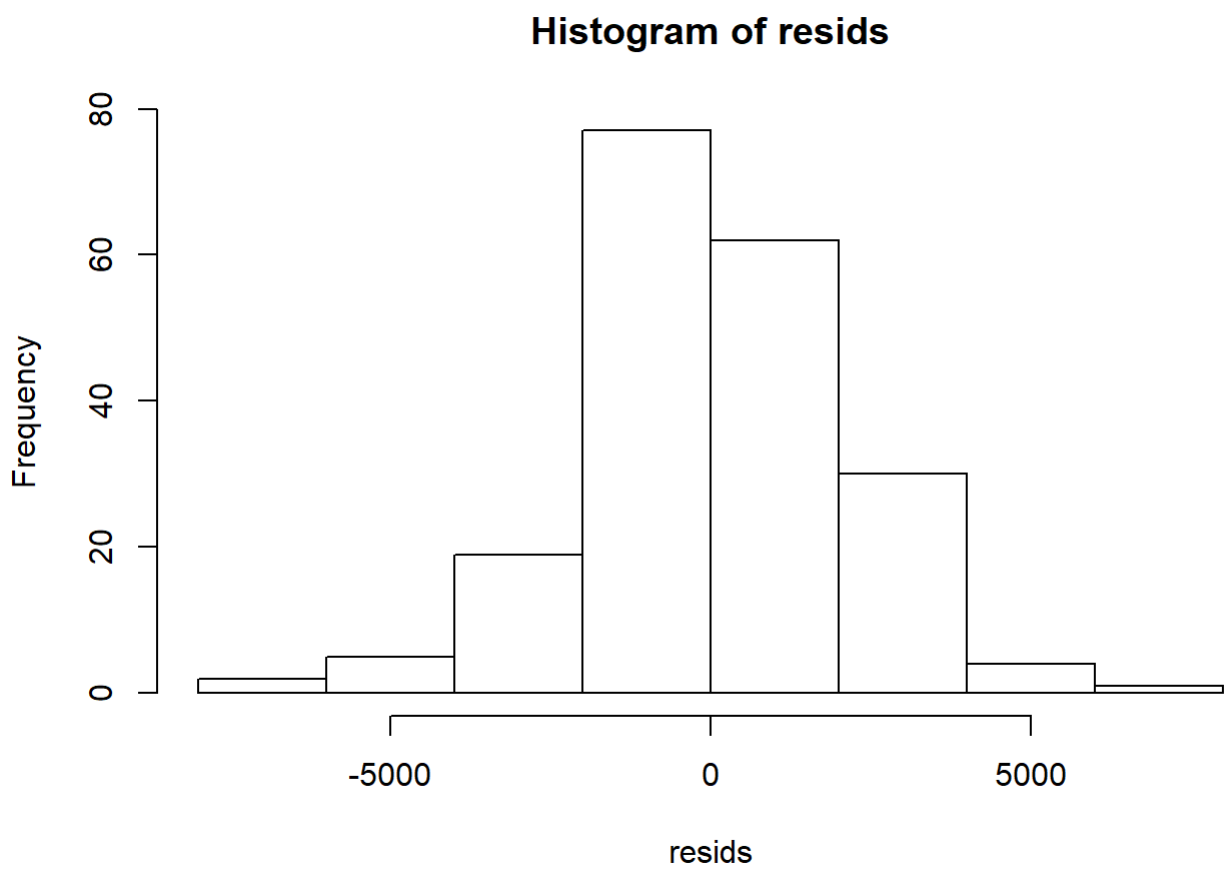
```
q7_stacked$Miles <- as.numeric(gsub(",","",q7_stacked$Miles))

q7_stacked <- tidyr::fill(q7_stacked,Gender)
#View(q7_stacked)
#Casting data types
Gender <- q7_stacked$Gender
Age <- q7_stacked$Age
Miles <- q7_stacked$Miles

#Examining the condition of normality
#Creating a residual plot
model <- lm(Miles ~ Age + Gender + Gender*Age)
resids <- residuals(model)

nortest::ad.test(resids)
```

```
##
##  Anderson-Darling normality test
##
## data:  resids
## A = 0.49791, p-value = 0.2089
```

```
shapiro.test(resids)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resids
## W = 0.98936, p-value = 0.1438
```

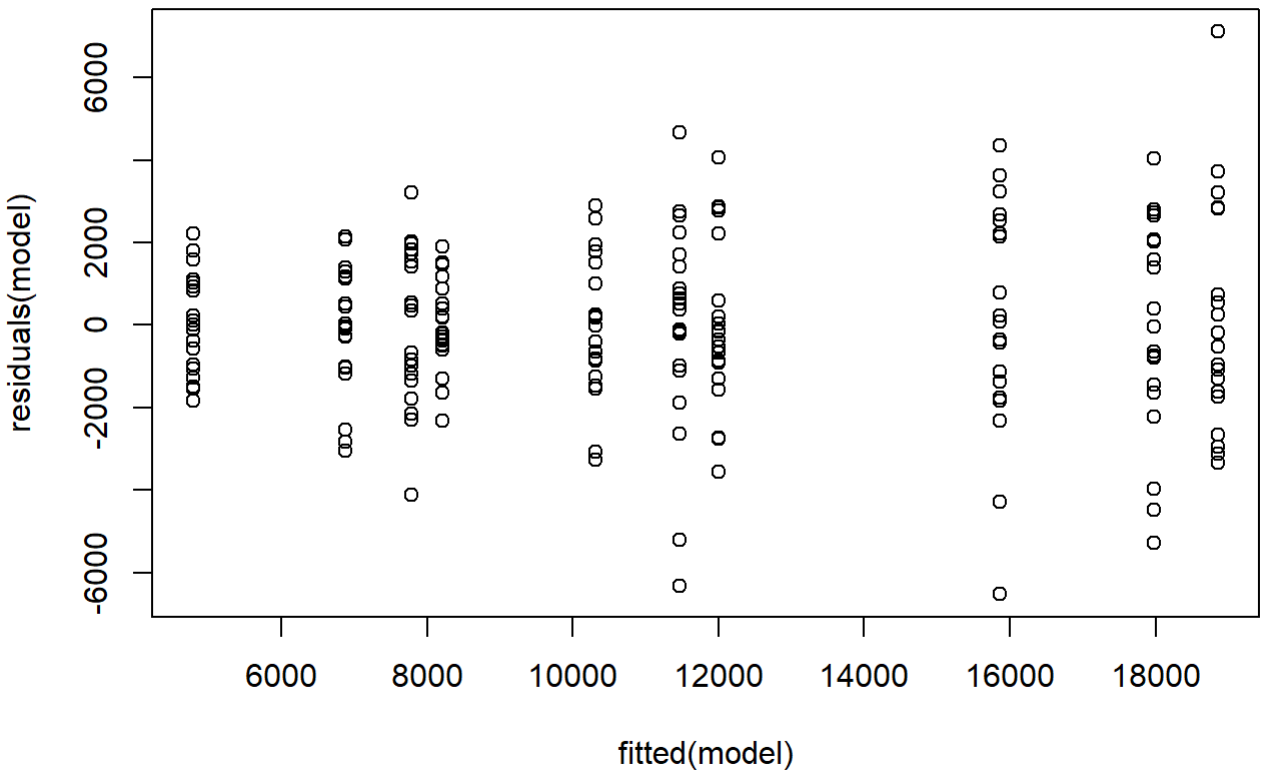
```
hist(resids)
```




```
#Check for variance
car::leveneTest(Miles ~ Gender)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   1   22.59 3.851e-06 ***
##       198
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

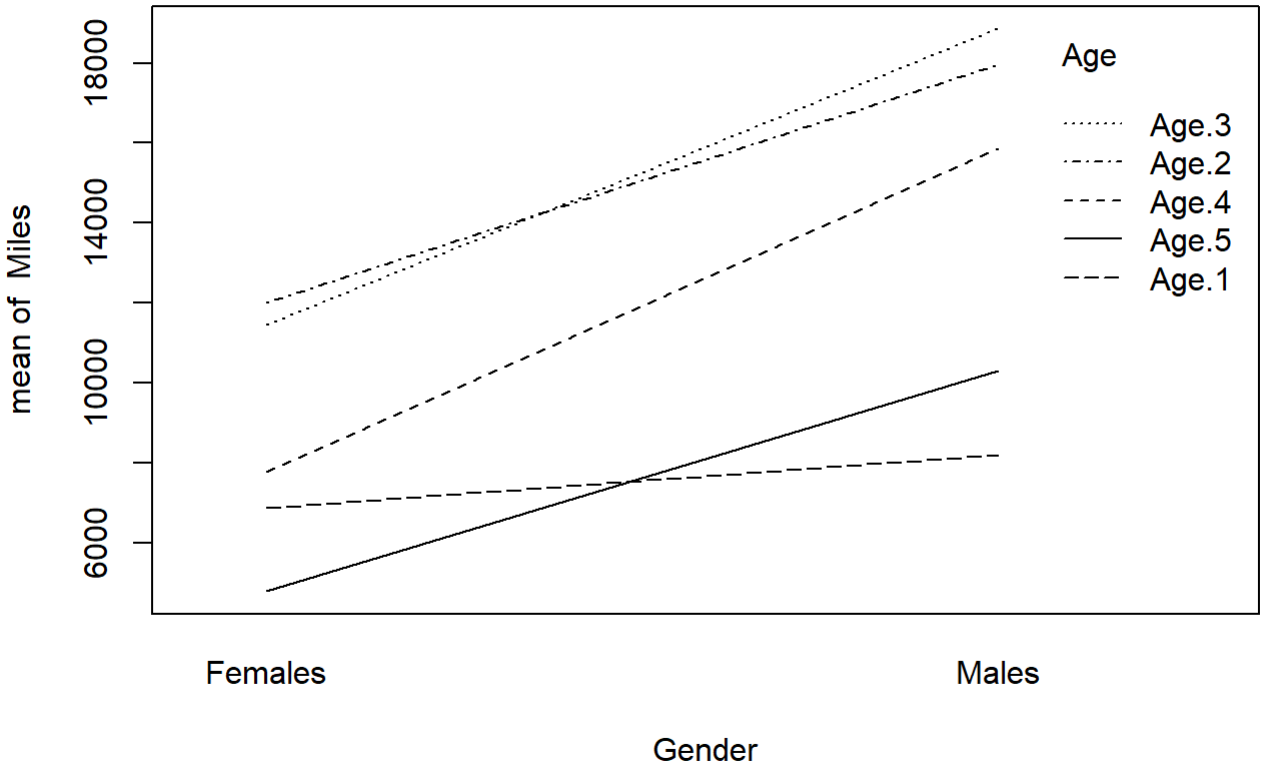
```
#Plot for variance test
plot(fitted(model), residuals(model))
```



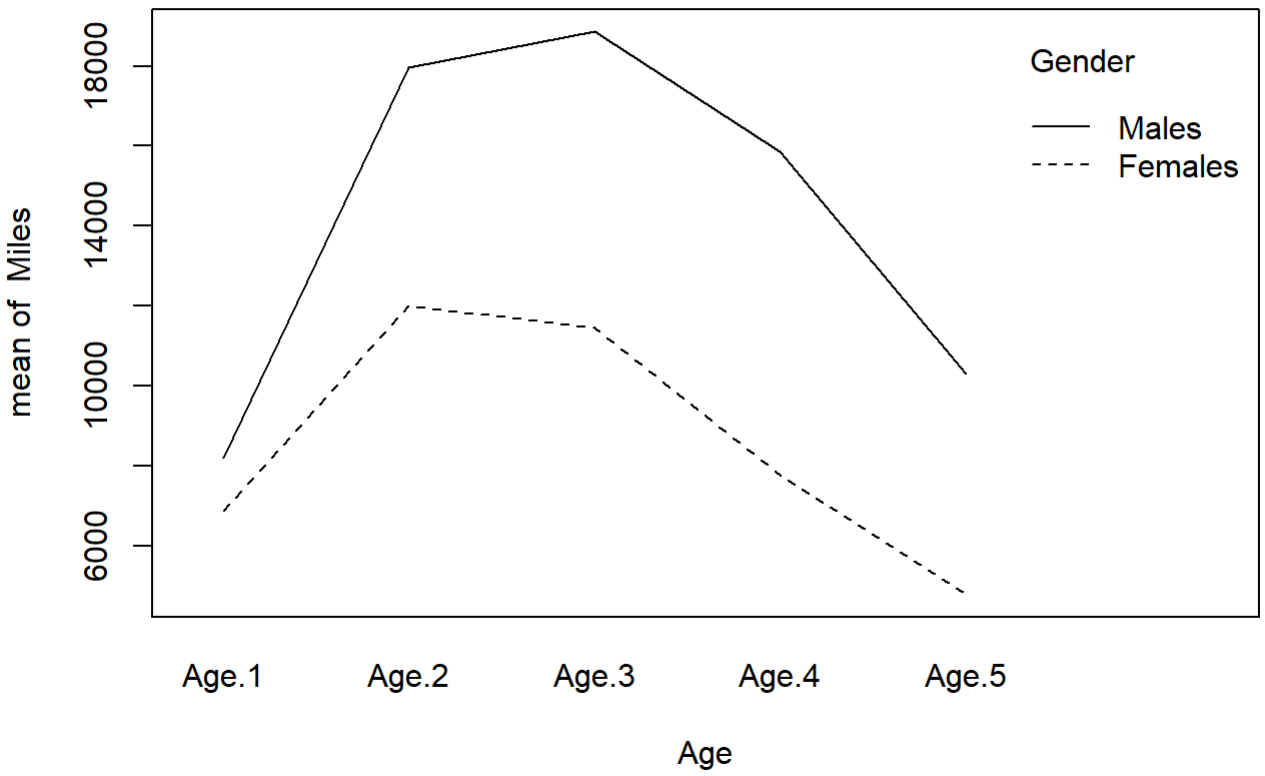

```
#As both normality condition and variance passed,
#Performing two factor ANOVA
Model <- aov(Miles ~ Age + Gender + Gender*Age)
summary(Model)
```

```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)
## Age           4 2.279e+09 5.697e+08  128.64  < 2e-16 ***
## Gender        1 1.601e+09 1.601e+09  361.57  < 2e-16 ***
## Age:Gender    4 2.769e+08 6.923e+07   15.63 4.49e-11 ***
## Residuals   190 8.415e+08 4.429e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Visual plots to see effect of interaction
interaction.plot(Gender,Age, Miles)
```




```
interaction.plot(Age, Gender, Miles)
```



As

the p-value for interaction effect > 0.05(or any other relevant significance level) We reject the null hypothesis Gender and Age groups strongly interact to affect the mean no. of miles they drive As there is significant interaction effect, we don't proceed further to check main effects.

Q8

According to the question, the four plants are designed and staffed to produce equally. So being from different plants won't make much difference on the productivity

Hypotheses for the Chi-squared goodness of fit test:
- H_0 : $ p_{(0-2500)} =p_{(2501-8000)}=p_{(8001-15000)}=p_{(15001-20000)}$ - H_a : At least one proportion differs after the contest got launched

Required Conditions are all met: 1. Contingency table condition-All categories are mutually exclusive - units produced 2. Sample Size Condition- Rule of five was valid 3. Simple Random Sample - group of employees working in the four plants
4. For more than 2 rows, chi-square should notbe used if more than 20% of the expected frequency cells have expected frequencies less than 5.

```
#proportions before the contest got launched
expected <- c(0.105,0.219,0.533,0.143)

#observations during the contest
observed <- c(0,20,83,52)
chisq.test(x=observed,p=expected)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 62.165, df = 3, p-value = 2.026e-13
```

```
#Display expected values
chisq.test(x=observed,p=expected)$expected
```

```
## [1] 16.275 33.945 82.615 22.165
```

```
#95% confidence interval for the proportion of the days
#when  150001-20000 units will be produced during the contest
z95 <- qnorm(0.95,0,1)
p_Obs <- 52/155
p_Obs
```

```
## [1] 0.3354839
```

```
se <- sqrt(p_Obs*(1-p_Obs)/155)
se
```

```
## [1] 0.03792475
```

```
LCL <- p_Obs - z95*se
UCL <- p_Obs + z95*se
LCL
```

```
## [1] 0.2731032
```

```
UCL
```

```
## [1] 0.3978645
```

As p-value<<0.05 => Reject Null Hypothesis At 5% significance level, we have overwhelming evidence that supports our claim that the proportion of days has changed before and during the contest Also, We are 95% confident that the population proportion of days where number of units are between 15001-20000 are between 0.27 and 0.40. Both the limits are greater than the proportion before the contest. The contest has improved the productivity of the plant significantly

Q9

According to the question, our goal is to assess whether trading on some days of the week is better or worse than any other

Hypotheses for the Chi-squared test:
- H_0 : Trading status and weekdays are independent - H_a : Trading status and weekdays are dependent

Conditions required for calling the chisquared test valid: 1.Categories are mutually exclusive When market is up and down is defined by the company and clearly demarcated Weekdays are mutually exclusive to each other 2.Have enough data in each cell, count>10 everywhere 3.Simple Random Sample to correctly represent the underlying population

```
#Step1:Create observed values contigency table
q9_observed <- data.frame(row.names=c('Down','Up'),"Mon"=c(42,53),"Tue"=c(49,55),"Wed"=c(46,58),
"Thu"=c(43,59),
                 "Fri"=c(41,58))

#Step2: Create expected values contigency table
q9_expected <- data.frame(row.names = c('Down','Up'),
                          "Mon"=c(95*221/504,95*283/504),
                          "Tue"=c(104*221/504,104*283/504),
                          "Wed"=c(104*221/504,104*283/504),
                          "Thu"=c(102*221/504,102*283/504),
                          "Fri"=c(99*221/504,99*283/504)
                          )
#q9_observed
q9_expected
```

```
##           Mon      Tue      Wed      Thu      Fri
## Down 41.65675 45.60317 45.60317 44.72619 43.41071
## Up   53.34325 58.39683 58.39683 57.27381 55.58929
```

```
#Step3: To compute Chisquared test statistic
i=0
j=0
chisq=0
for(i in 1:2)
{
  for(j in 1:5)
  {
    #print(i)
    #print(j)
    chisq=chisq+(q9_observed[i,j]-q9_expected[i,j])^2/q9_expected[i,j]
    #print(chisq)
  }
}
chisq
```

```
## [1] 0.8188568
```

```
#Step4: Compute Degrees of freedom=(r-1)(c-1)
df9 <- (nrow(q9_observed)-1)*(ncol(q9_observed)-1)
df9
```

```
## [1] 4
```

```
pchisq(chisq, df=df9,lower.tail=FALSE)
```

```
## [1] 0.9359023
```

As p-value>>0.05(and any reasonable significance level), Failed to reject null hypothesis i.e. they are independent At 5% significance level, there isn't enough evidence to support the claim that weekdays and trading performance are dependent i.e. they don't affect each other.

Limitations imposed by data on the conclusion of the test Condition 1&2 are clearly satisfied by the data But condition 3 isn't completely satisfied:cause data is used for last year to make an assumption on the population The data can be biased by effect of economy specific to last year. Would have been a better option to randomly collect across years

Q10

Provided data is a sample data comprising info about middle class households Our goal is to establish: is there enough evidence to conclude that heads of households with college degrees have more assets than those who have some college?

EDCL levels: 3- Enrolled in some college 4- Obtained college degree

Initial assumption: For mean asset of a household:
- $H_0 : \mu_{EDCL=3} = \mu_{EDCL=4} \geq 0$
- $H_a : \mu_{EDCL=3} - \mu_{EDCL=4} < 0$
i.e. heads of Households with college degree have higher assets

```
q10 <- read.csv('Question10.csv')
#View(q10)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#filtering out households with EDCL of 3&4
q10_filtered <- q10 %>% filter(EDCL==3 | EDCL==4)

q10_filtered$ASSET <- as.numeric(gsub(",","",q10_filtered$ASSET))
#View(q10_filtered)
#computing the number of rows:
q10%>%filter(EDCL==3 | EDCL==4)%>%group_by(EDCL)%>% summarise(n())
```

```
## # A tibble: 2 x 2
##    EDCL `n()`
##   <int> <int>
## 1     3   214
## 2     4   450
```

```
#Checking Normality
#Creating a residual plot
model <- lm(q10_filtered$ASSET ~ q10_filtered$EDCL)
resids <- residuals(model)

nortest::ad.test(resids)
```

```
##
##  Anderson-Darling normality test
##
## data:  resids
## A = 9.4906, p-value < 2.2e-16
```

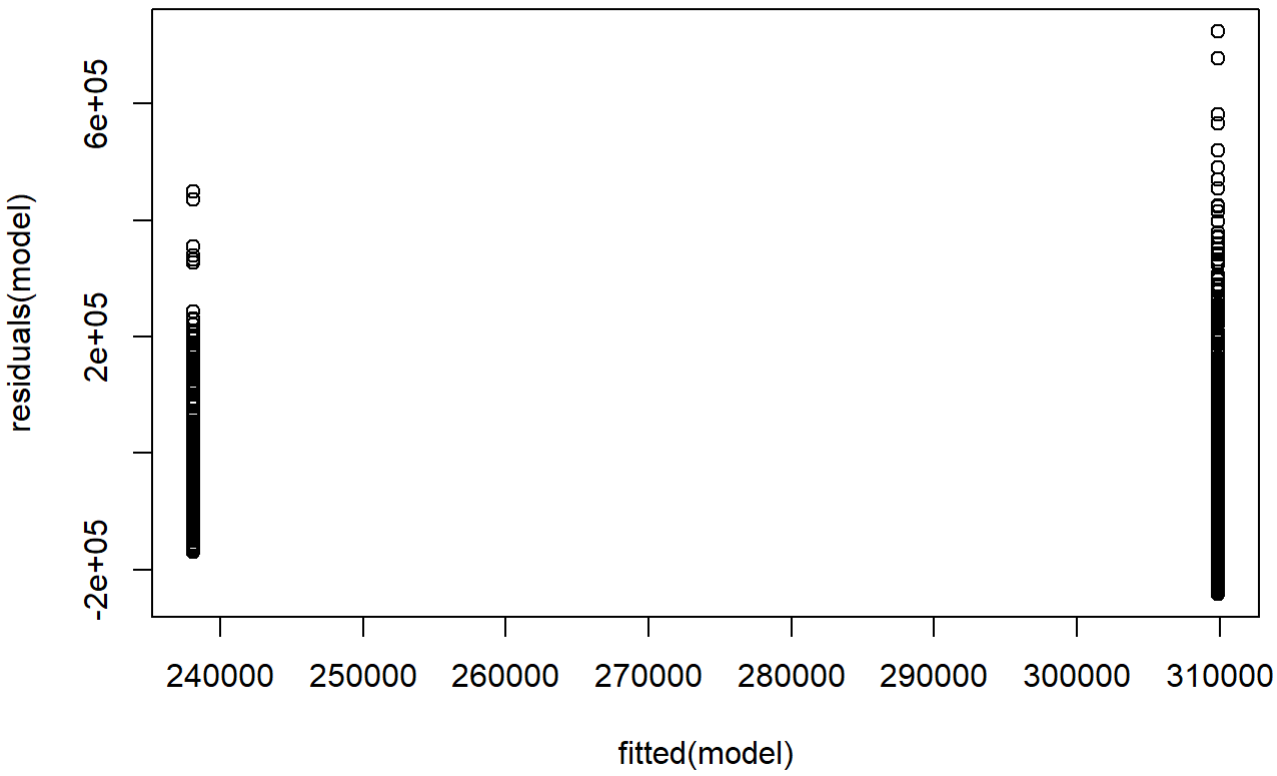
```
shapiro.test(resids)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resids
## W = 0.9349, p-value < 2.2e-16
```

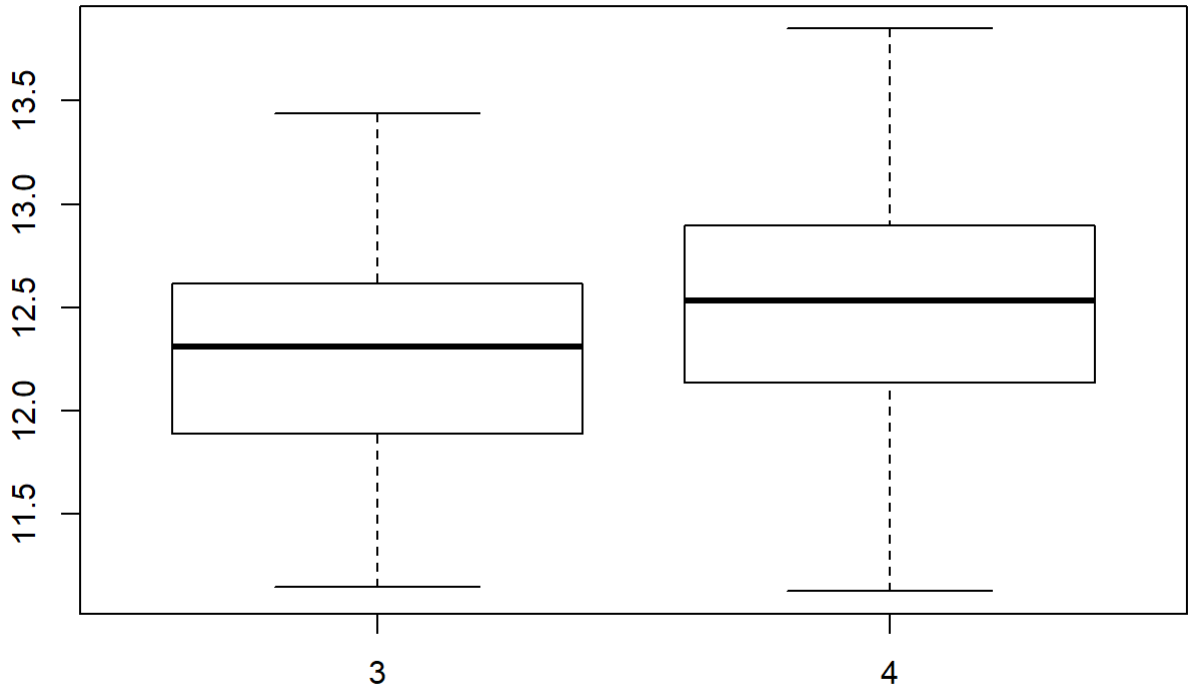
```
#Checking for variance
car::leveneTest(q10_filtered$ASSET ~ factor(q10_filtered$EDCL))
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   1  16.613 5.141e-05 ***
##       662
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

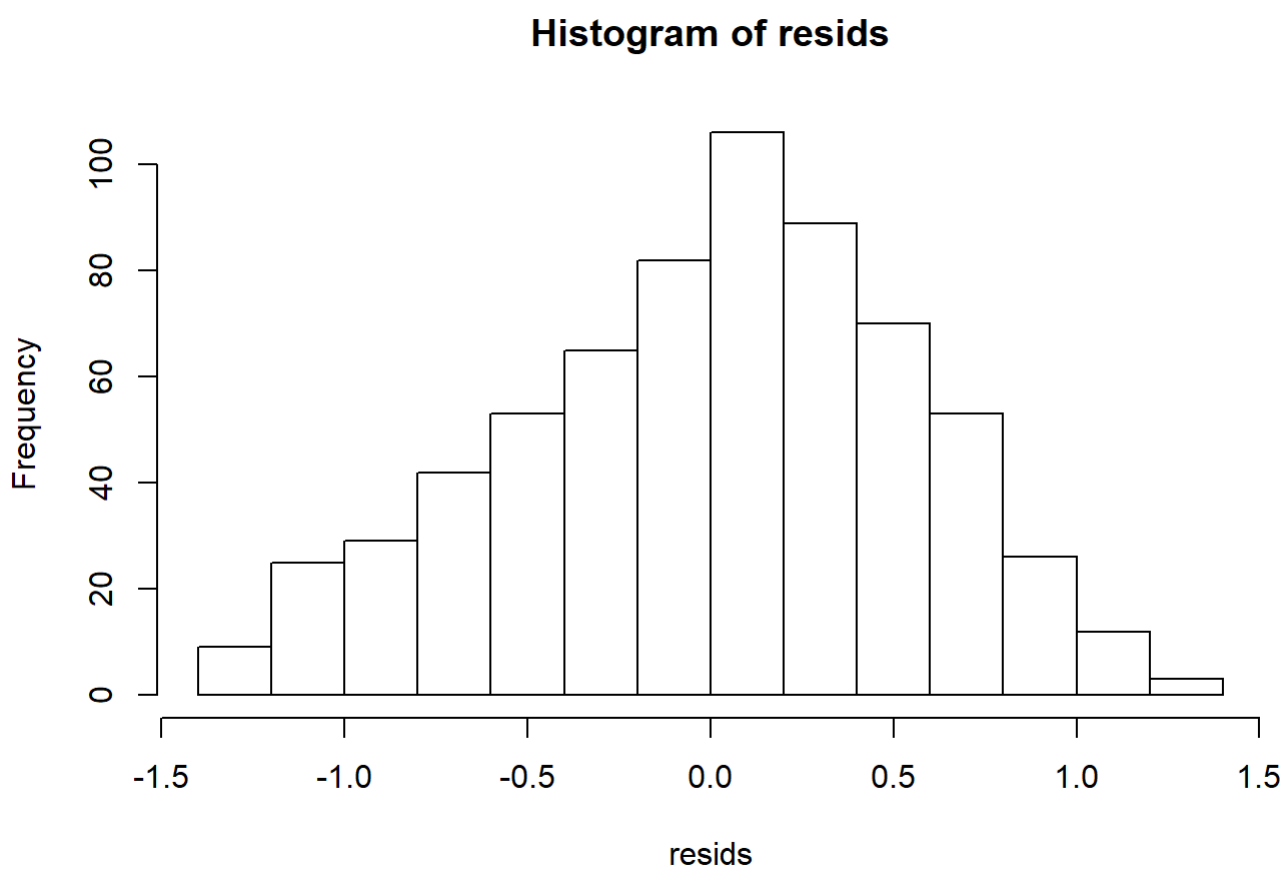
```
#Plot for variance test
plot(fitted(model), residuals(model))
```




```
#Both variance and noramlity test is violated, so we go for transformations

#Applying log transformations on Asset(dependent variable)
q10_filtered$log_asset <- log(q10_filtered$ASSET)
boxplot(q10_filtered$log_asset ~ factor(q10_filtered$EDCL))
```




```
#Testing for normality
model <- lm(q10_filtered$log_asset ~ factor(q10_filtered$EDCL))
resids <- residuals(model)
hist(resids)
```




```
nortest::ad.test(resids)
```

```
##
##  Anderson-Darling normality test
##
## data:  resids
## A = 1.7156, p-value = 0.0002125
```

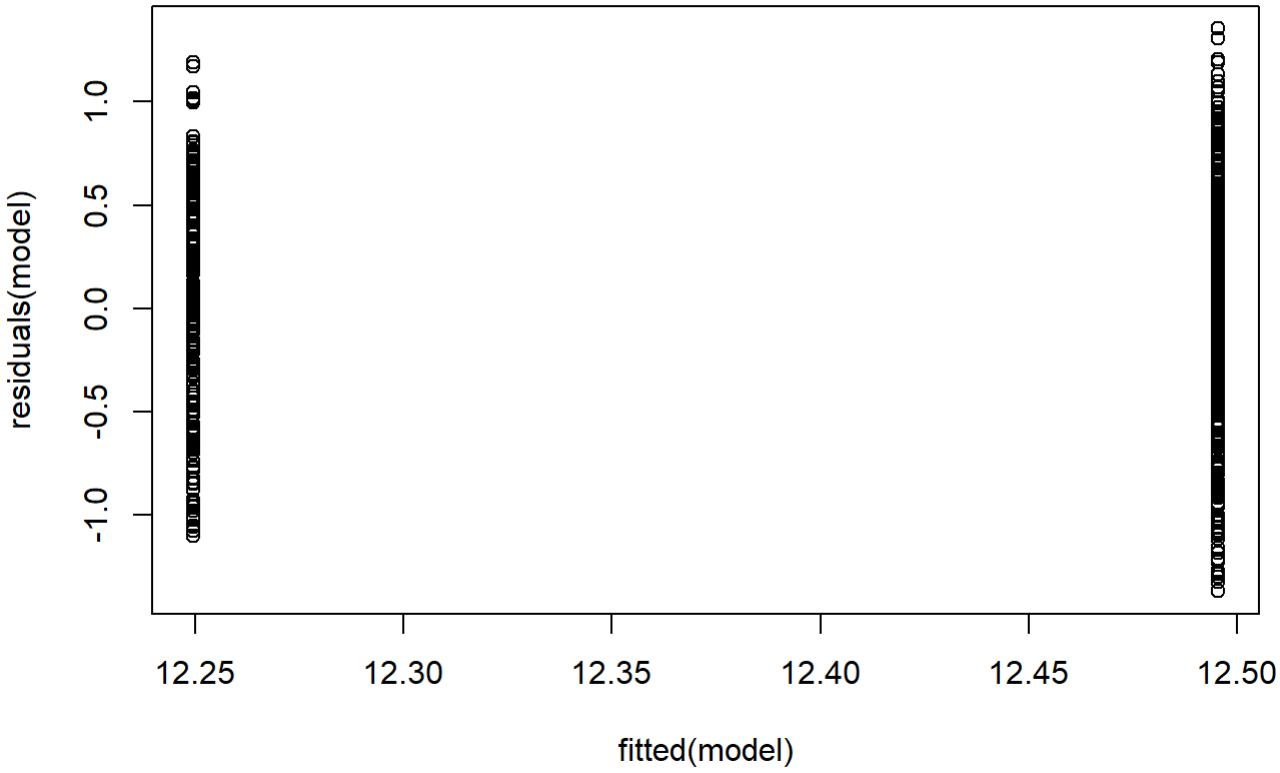
```
shapiro.test(resids)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resids
## W = 0.98929, p-value = 9.076e-05
```

```
#the transformed ASSET fails the normality test as the p-value <<0.05

#Testing for variance
car::leveneTest(q10_filtered$log_asset ~ factor(q10_filtered$EDCL))
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.9384  0.333
##       662
```

```
plot(fitted(model), residuals(model))
```




```
#the variance test has passed for the log transformed as the p-value is well above 0.05
#As equal variance failed and normality passed, using Kruskal-wallis test
kruskal.test(q10_filtered$log_asset ~ factor(q10_filtered$EDCL))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  q10_filtered$log_asset by factor(q10_filtered$EDCL)
## Kruskal-Wallis chi-squared = 27.112, df = 1, p-value = 1.92e-07
```

```
#As p-value is way below 0.05-> Reject the null hypothesis
#So we run posthoc analysis
library(PMCMR)
```

```
## Warning: package 'PMCMR' was built under R version 3.5.3
```

```
## PMCMR is superseded by PMCMRplus and will be no longer maintained. You may wish to install PM
CMRplus instead.
```

```
posthoc.kruskal.nemenyi.test(q10_filtered$log_asset ~ factor(q10_filtered$EDCL),p.adjust.methods
='bonferroni')
```

```
## Warning in posthoc.kruskal.nemenyi.test.default(c(12.2971465924295,
## 12.8174673242179, : Ties are present, p-values are not corrected.
```

```
##
##  Pairwise comparisons using Tukey and Kramer (Nemenyi) test
##                    with Tukey-Dist approximation for independent samples
##
## data:  q10_filtered$log_asset by factor(q10_filtered$EDCL)
##
##   3
## 4 1.9e-07
##
## P value adjustment method: none
```

For mean asset of a household:
- $H_0 : \mu_{EDCL=4} - \mu_{EDCL=3} \leq 0$
- $H_a : \mu_{EDCL=4} - \mu_{EDCL=3} > 0$

```
q10_EDCL3 <- q10_filtered%>%filter(EDCL==3)%>%select(ASSET)
q10_EDCL4 <- q10_filtered%>%filter(EDCL==4)%>%select(ASSET)
t.test(q10_EDCL4$ASSET,q10_EDCL3$ASSET,,alternative='greater',mu=0,paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  q10_EDCL4$ASSET and q10_EDCL3$ASSET
## t = 6.2422, df = 559.28, p-value = 4.267e-10
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  52838.31      Inf
## sample estimates:
## mean of x mean of y
##  309872.8  238087.1
```

As p-value<0.05, we reject the null hypothesis. The data provides sufficient evidence to infer that heads of households with college degrees(EDCL=4) have higher mean value of assets than those who have some college(EDCL=3)

Q11

ATQ, there are four campuses in total 15 emergency cases randomly selected from each campus It is a random assignment design Factor: Campus Levels: Main campus, Satellite.1, Satellite.2, Satellite.3 Levels and treatments are same in one factor ANOVA

For mean waiting time: - H_0 : $\mu_{main} = \mu_{sat1} = \mu_{sat2} = \mu_{sat3}$
- H_a : Atleast one mean is different

```
q11 <- read.csv('Question11.csv')
head(q11)
```

```
##     Main Satellite.1 Satellite.2 Satellite.3
## 1 120.08       30.75       75.86       54.05
## 2  81.90       61.83       37.88       38.82
## 3  78.79       26.40       68.73       36.85
## 4  63.83       53.84       51.08       32.83
## 5  79.77       72.30       50.21       52.94
## 6  47.94       53.09       58.47       34.13
```

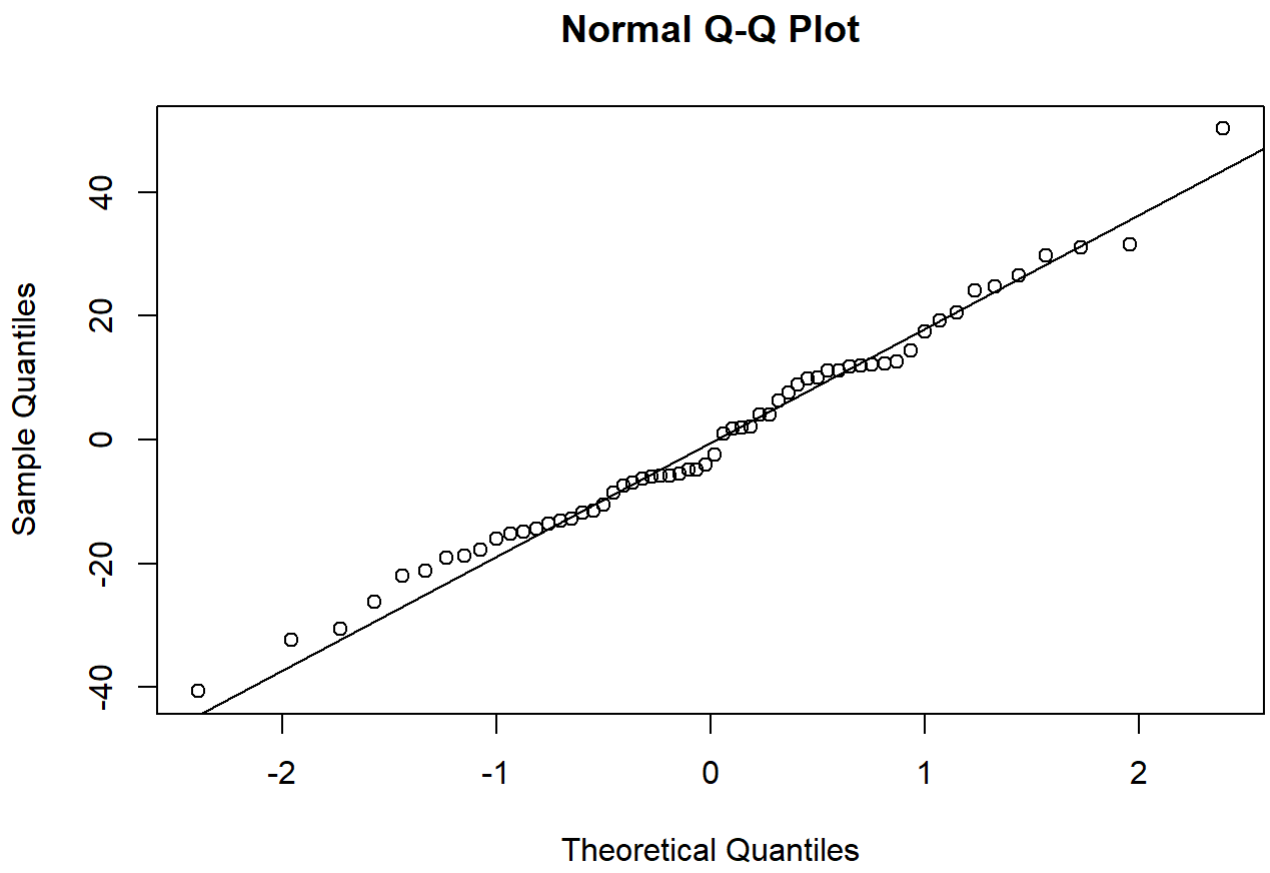
```
q11 <- stack(q11)
colnames(q11) <- c('w_time','campus')
class(q11$campus)
```

```
## [1] "factor"
```

```
#Checking normality
#Creating a residual plot
model<- lm(q11$w_time ~ q11$campus)
resids <- residuals(model)
preds<- predict(model)

nortest::ad.test(resids)
```

```
##
##  Anderson-Darling normality test
##
## data:  resids
## A = 0.29114, p-value = 0.5973
```

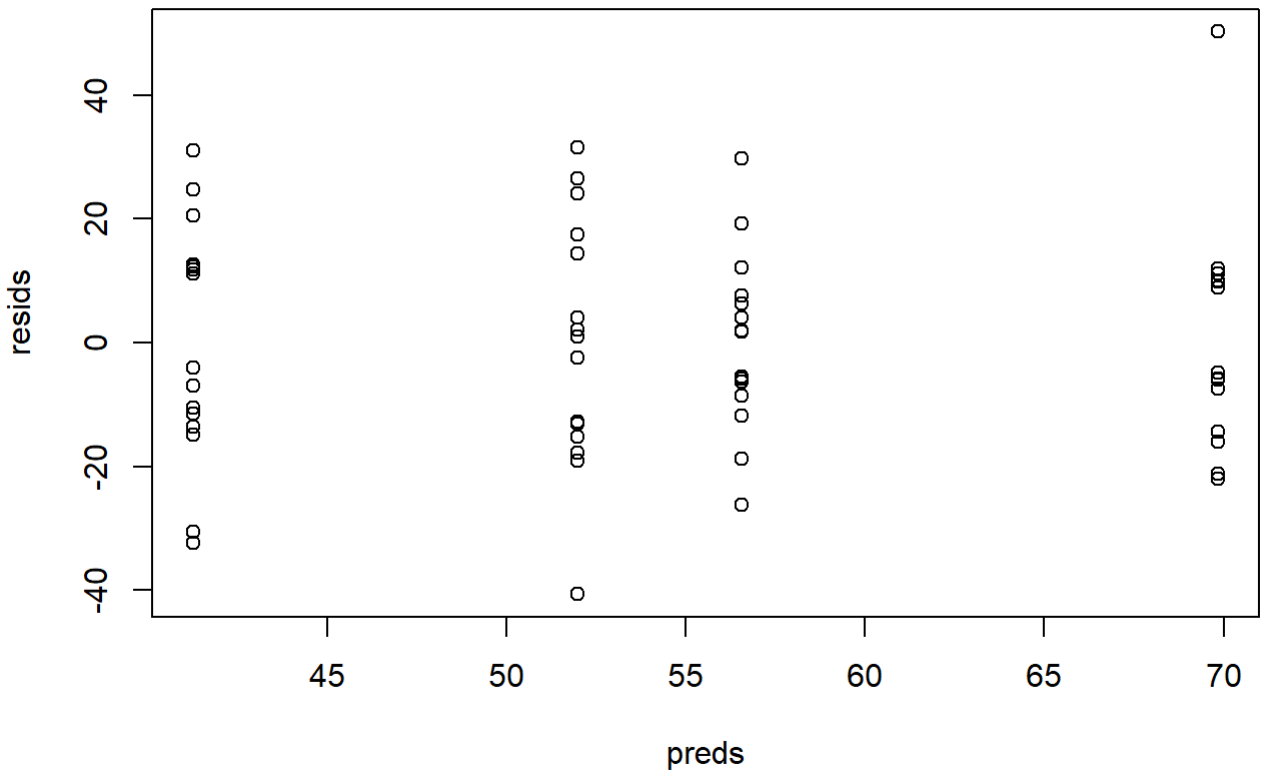
```
qqnorm(resids)
qqline(resids)
```




```
#p-value >0.05--> Normality test passed
#All the QQ plot approx. aligns with normal distribution

#Checking for variance
car::leveneTest(q11$w_time ~ q11$campus)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  0.8201 0.4883
##       56
```

```
#visual for variance test
plot(preds, resids)
```



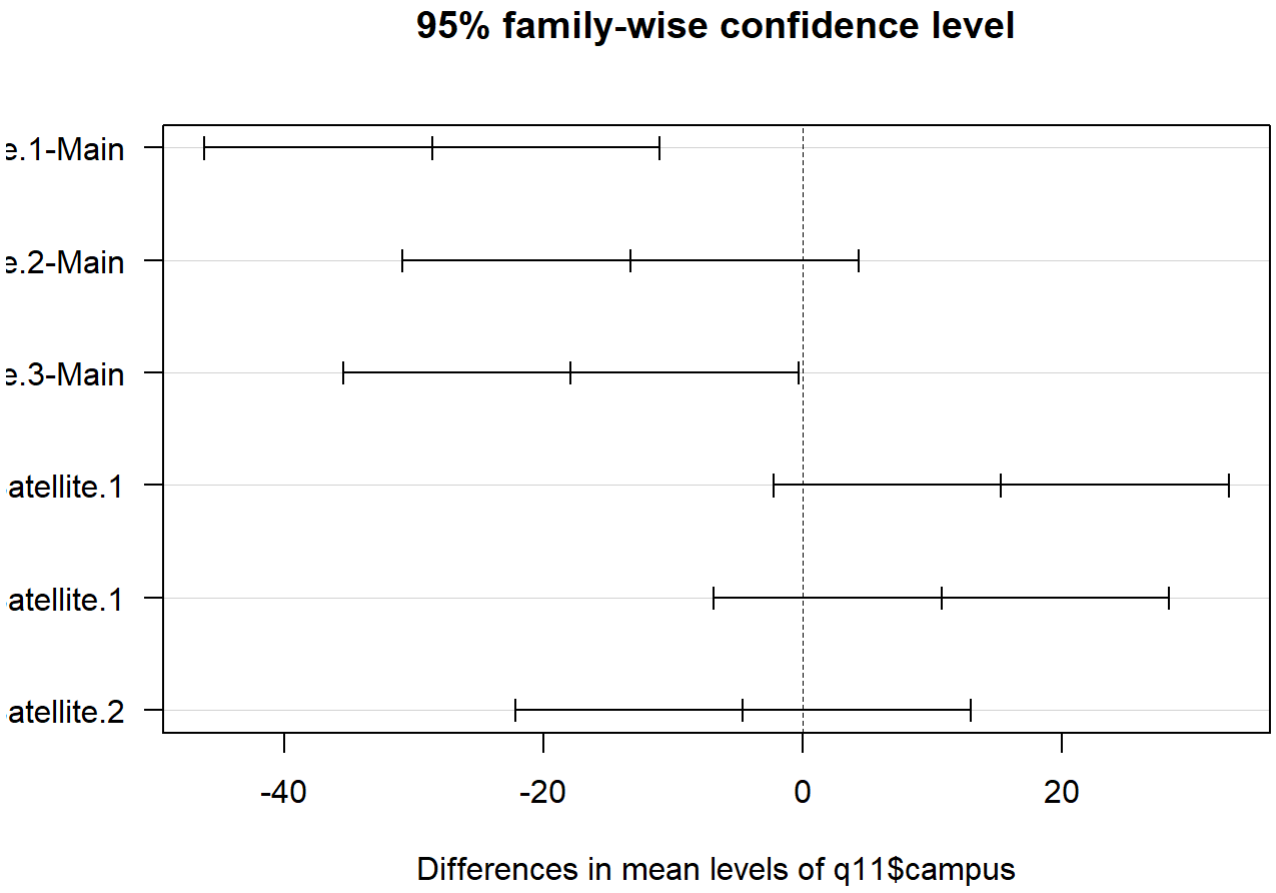

```
#p-value>0.05 -->Constant variance
#and visually it appears somewhat constant variance except few outliers

#Method for one-factor ANOVA
ANOVA_11 <- aov(q11$w_time ~ q11$campus)
summary(ANOVA_11)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## q11$campus   3   6312  2104.1   6.372 0.000859 ***
## Residuals   56  18493   330.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Since, p-value <0.05. We reject the null hypothesis
#Proceed with tukey test
TukeyHSD(ANOVA_11)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = q11$w_time ~ q11$campus)
##
## $`q11$campus`
##                               diff        lwr         upr     p adj
## Satellite.1-Main        -28.588667 -46.158986 -11.0183478 0.0003814
## Satellite.2-Main        -13.281333 -30.851652   4.2889855 0.1997139
## Satellite.3-Main        -17.870667 -35.440986  -0.3003478 0.0447878
## Satellite.2-Satellite.1  15.307333  -2.262986  32.8776522 0.1086697
## Satellite.3-Satellite.1  10.718000  -6.852319  28.2883188 0.3785196
## Satellite.3-Satellite.2  -4.589333 -22.159652  12.9809855 0.8998737
```

```
plot(TukeyHSD(ANOVA_11), las = 1)
```

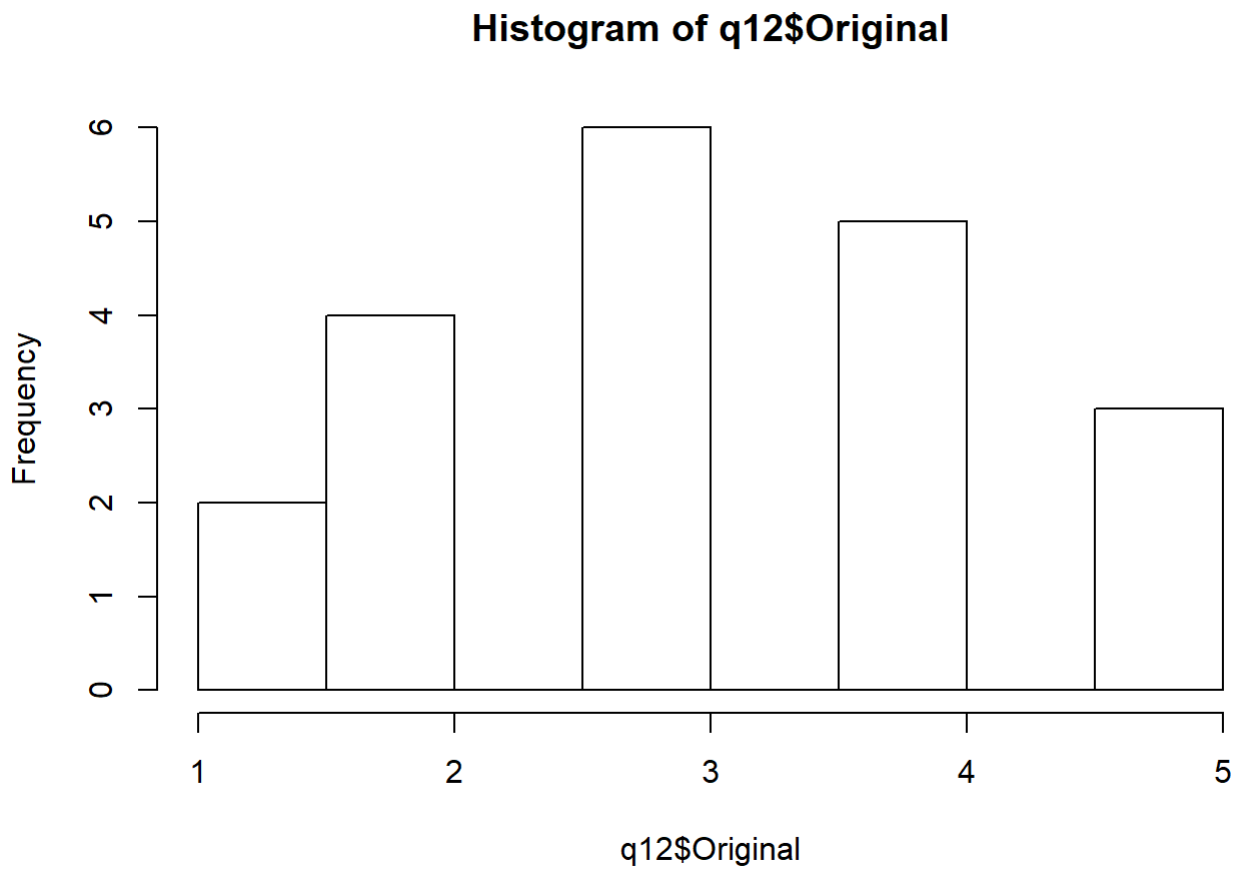


As p-value<0.05 (and any other relevant significance level) => Reject the H_0 At 5% significance level, there is sufficient evidence to support the claim that there is a difference between the mean waiting times of campuses To determine which population means significantly differ, we used Tukey test p-value for two combination of campuses <0.05 -> Null hypothesis for the two pairs rejected At 5% significance level, there is sufficient evidence to support the claim that mean waiting times of Satellite 1 and main campus are different mean waiting times of Satellite 3 and main campus are different

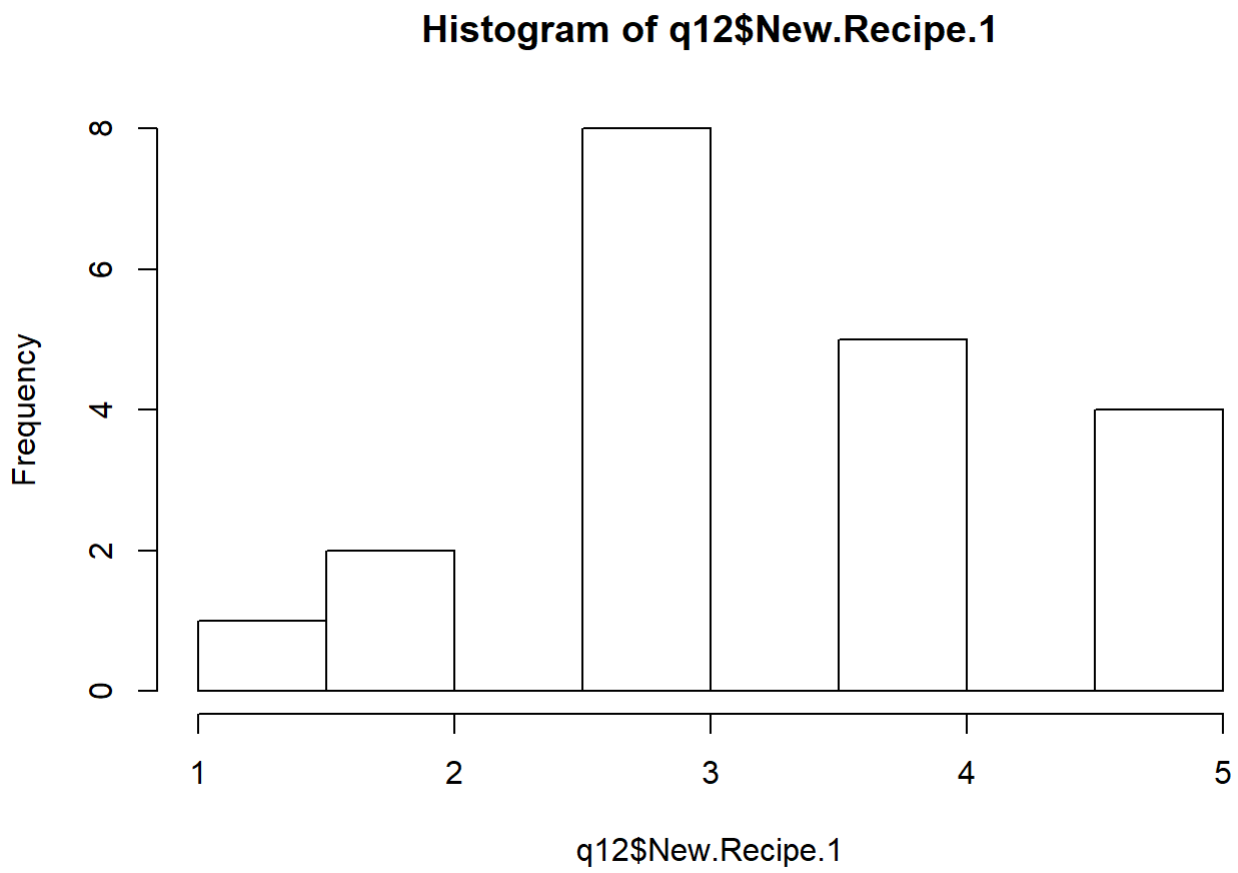
Q12

Objective-Determine if there difference in ratings of the recipes Type of data- Ordinal Type of sample- Independant It is a randomized block design as 20 people are selected Each taste all the three recipes

```
q12 <- read.csv('Question12.csv')
#head(q12)
hist(q12$Original)
```




```
hist(q12$New.Recipe.1)
```




```
hist(q12$New.Recipe..2)
```

We

notice that 2 out of the 3 populations clearly don't align to norma distribution Since, the condition for normality is failed and our dataset has ordinal values We will try non-parametric test to solve Number of samples-3

Friedman Test - H_0 : the location of all the three populations is the same - H_a : Atleast two population locations differ

```
#Convert the dataset to stacked format
q12
```

```
##    Person Original New.Recipe.1 New.Recipe..2
## 1       1        5            5             5
## 2       2        3            4             5
## 3       3        4            5             5
## 4       4        2            4             4
## 5       5        3            3             5
## 6       6        2            2             3
## 7       7        3            3             2
## 8       8        4            3             5
## 9       9        1            1             1
## 10     10        1            3             2
## 11     11        2            4             3
## 12     12        3            3             4
## 13     13        5            3             4
## 14     14        3            2             3
## 15     15        4            3             5
## 16     16        4            4             4
## 17     17        2            4             3
## 18     18        5            3             5
## 19     19        4            5             5
## 20     20        3            5             4
```

```
q12_stacked <- cbind(q12[,1:2],'Original')
q12_stacked
```

```
##    Person Original "Original"
## 1       1        5   Original
## 2       2        3   Original
## 3       3        4   Original
## 4       4        2   Original
## 5       5        3   Original
## 6       6        2   Original
## 7       7        3   Original
## 8       8        4   Original
## 9       9        1   Original
## 10     10        1   Original
## 11     11        2   Original
## 12     12        3   Original
## 13     13        5   Original
## 14     14        3   Original
## 15     15        4   Original
## 16     16        4   Original
## 17     17        2   Original
## 18     18        5   Original
## 19     19        4   Original
## 20     20        3   Original
```

```
colnames(q12_stacked) <- c('Person','Rating','Recipe')
q12_stacked_t <- cbind(q12[,c(1,3)],'NewRecipe_1')
colnames(q12_stacked_t) <- c('Person','Rating','Recipe')
q12_stacked <- rbind(q12_stacked,q12_stacked_t)
q12_stacked_t <- cbind(q12[,c(1,4)],'NewRecipe_2')
colnames(q12_stacked_t) <- c('Person','Rating','Recipe')
q12_stacked <- rbind(q12_stacked,q12_stacked_t)
rm(q12_stacked_t)

#Make sure column types are correct
q12_stacked$Person <- factor(q12_stacked$Person)
q12_stacked$Recipe <- factor(q12_stacked$Recipe)

#Run Friedman Test
stats::friedman.test(q12_stacked$Rating,q12_stacked$Recipe,q12_stacked$Person)
```

```
##
##  Friedman rank sum test
##
## data:  q12_stacked$Rating, q12_stacked$Recipe and q12_stacked$Person
## Friedman chi-squared = 7.1525, df = 2, p-value = 0.02798
```

The p-value > 0.05 significance level, so we failed to reject the null hypothesis At 5% significance level, there isn't enough evidence to support the claim that a difference in mean ratings across these recipes exist

Q13

Hypotheses for the analysis:
- $H_0 : \rho_{hrs} = 0$ –no correlation - $H_a : \rho_{hrs} \neq 0$ –correlation exists

```
q13 <- read.csv('Question13.csv')
head(q13)
```

```
##   Year Growth  Value
## 1 2007  12.56   0.09
## 2 2008 -38.32 -35.97
## 3 2009  36.29  19.58
## 4 2010  16.96  14.28
## 5 2011   1.71   1.00
## 6 2012  16.89  15.00
```

```
cor.test(q13$Growth,q13$Value,alternative='greater')
```

```
##
##  Pearson's product-moment correlation
##
## data:  q13$Growth and q13$Value
## t = 7.0698, df = 8, p-value = 5.254e-05
## alternative hypothesis: true correlation is greater than 0
## 95 percent confidence interval:
##  0.7720455 1.0000000
## sample estimates:
##       cor
## 0.9284542
```

Our p-value<<0.05, we have sufficient evidence to reject our null hypothesis Correlation exists between Growth and value mutual funds Our value for ρ ~ 1 => indicating positive correlation between the two types of mutual funds, i.e. they both move in the same direction Henceit's safe to say that the performance of the returns of these two funds were similar over the 2007-2016 timeframe.

Q14

JOBLOSE:1,2,3,4 Lower the value of JOBLOSE more the likelihood of losing the job Hypotheses for the analysis:
- $H_0 : \rho_{hrs} \leq 0$
- $H_a : \rho_{hrs} > 0$ –positive correlation

```
q14 <- read.csv('Question14.csv')
#As JOB LOSE is ordinal in nature
cor.test(q14$HRS1,q14$JOBLOSE, alternative='greater',method='spearman')
```

```
## Warning in cor.test.default(q14$HRS1, q14$JOBLOSE, alternative =
## "greater", : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  q14$HRS1 and q14$JOBLOSE
## S = 151100000, p-value = 0.00117
## alternative hypothesis: true rho is greater than 0
## sample estimates:
##       rho
## 0.09609079
```

As p-value<<0.05, Reject the null hypothesis At 5% significance level, we have enough eveidence to support the claim that there is a positive correlation between workhours and JOBLOSE rating Which basically means higher the workhours; higher the JOBLOSE Rating i.e. *less is the likelihood* of them losing the job

Q15

She decided to ask 60 randomly selected people to taste the same flavor of ice cream in two different dishes. So it's a paried sample because same set of people are made to test the ice creams under two different names Data Type for Ratings - Ordinal -> we will go for non-parametric test As the rating is discrete and bounded (1,2,3,4,5), it

will never have a normal distribution Factor - Name type (Domestic / European) population 1 -European Name and population 2 -Domestic Name

Following are the hypothesis for the paired sample test: - H_0 : The locations of two populations are the same - H_a : The location of population 1 is to the right of location 2

```
q15 <- read.csv('Question15.csv')
#View(q15)
q15$PersonID <- seq.int(nrow(q15))

#Run the sign test for paired samples
library(BSDA)
```

```
## Warning: package 'BSDA' was built under R version 3.5.3
```

```
##
## Attaching package: 'BSDA'
```

```
## The following objects are masked from 'package:PASWR2':
##
##     normarea, nsize, ntester, SIGN.test, tsum.test, z.test,
##     zsum.test
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```
SIGN.test(x = q15$European,
          y = q15$Domestic,
          alternative = "greater",
          conf.level = 0.90)
```

```
##
##  Dependent-samples Sign-Test
##
## data:  q15$European and q15$Domestic
## S = 30, p-value = 0.000236
## alternative hypothesis: true median difference is greater than 0
## 90 percent confidence interval:
##    0 Inf
## sample estimates:
## median of x-y
##           0.5
##
## Achieved and Interpolated Confidence Intervals:
##
##                   Conf.Level L.E.pt U.E.pt
## Lower Achieved CI     0.8775      0    Inf
## Interpolated CI       0.9000      0    Inf
## Upper Achieved CI     0.9225      0    Inf
```

```
#Alternate method
#PASWR2::SIGN.test(x = q15$European,
#          y = q15$Domestic,
#          alternative = "greater",
#          conf.level = 0.90)
```

As p-value is less than any reasonable significance level. Hence we reject Null hypothesis. At 10% significance level, The results indicate that brand name of an ice cream affect consumers' perceptions of it. We have sufficient evidence to support our claim that european brand names are preferred over domestic names.

Q16

The data is collected for the month of September, a couple of years ago -fixed time frame It represents price of best-selling books across two sellers- Amazon and Barnes&Noble So Say a Book A –> it's price in Amazon and B&N …and so on

Since it's a paperback vs paperback and harback vs hardback comparison,and the books are same we will consider it as paired sample.

For the paired t test (two-tailed) - $H_0 : \mu_{AM} = \mu_{BN}$
- $H_a : \mu_{AM} \neq \mu_{BN}$

```
q16 <- read.csv('Question16.csv')
#head(q16)
#Conducting a test to check if population variances are equal using F test
var.test(q16$BN,q16$Amazon, ratio=1,
         alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  q16$BN and q16$Amazon
## F = 0.99968, num df = 9, denom df = 9, p-value = 0.9996
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2483056 4.0246931
## sample estimates:
## ratio of variances
##          0.9996768
```

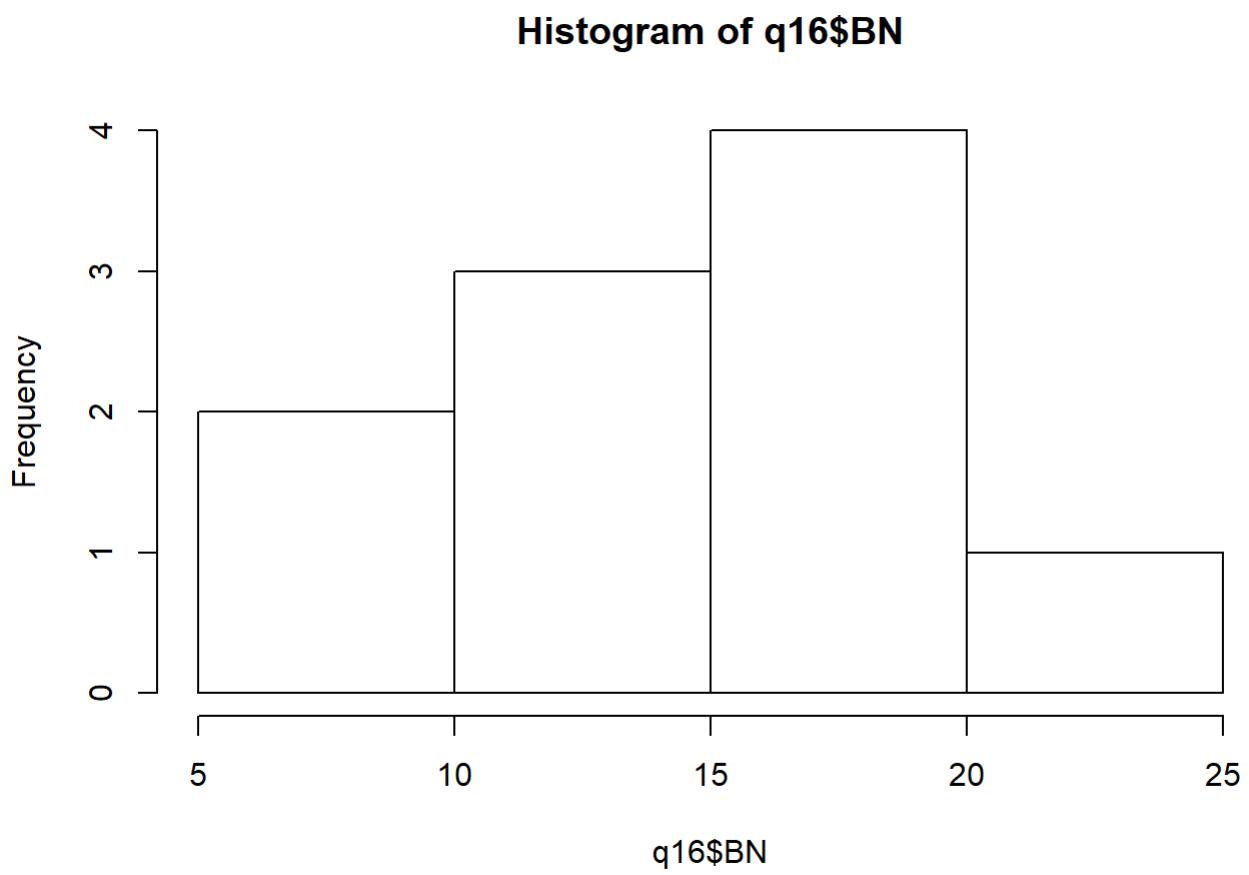
```
#Test for normality
nortest::ad.test(q16$BN)
```

```
##
##  Anderson-Darling normality test
##
## data:  q16$BN
## A = 0.49253, p-value = 0.1655
```

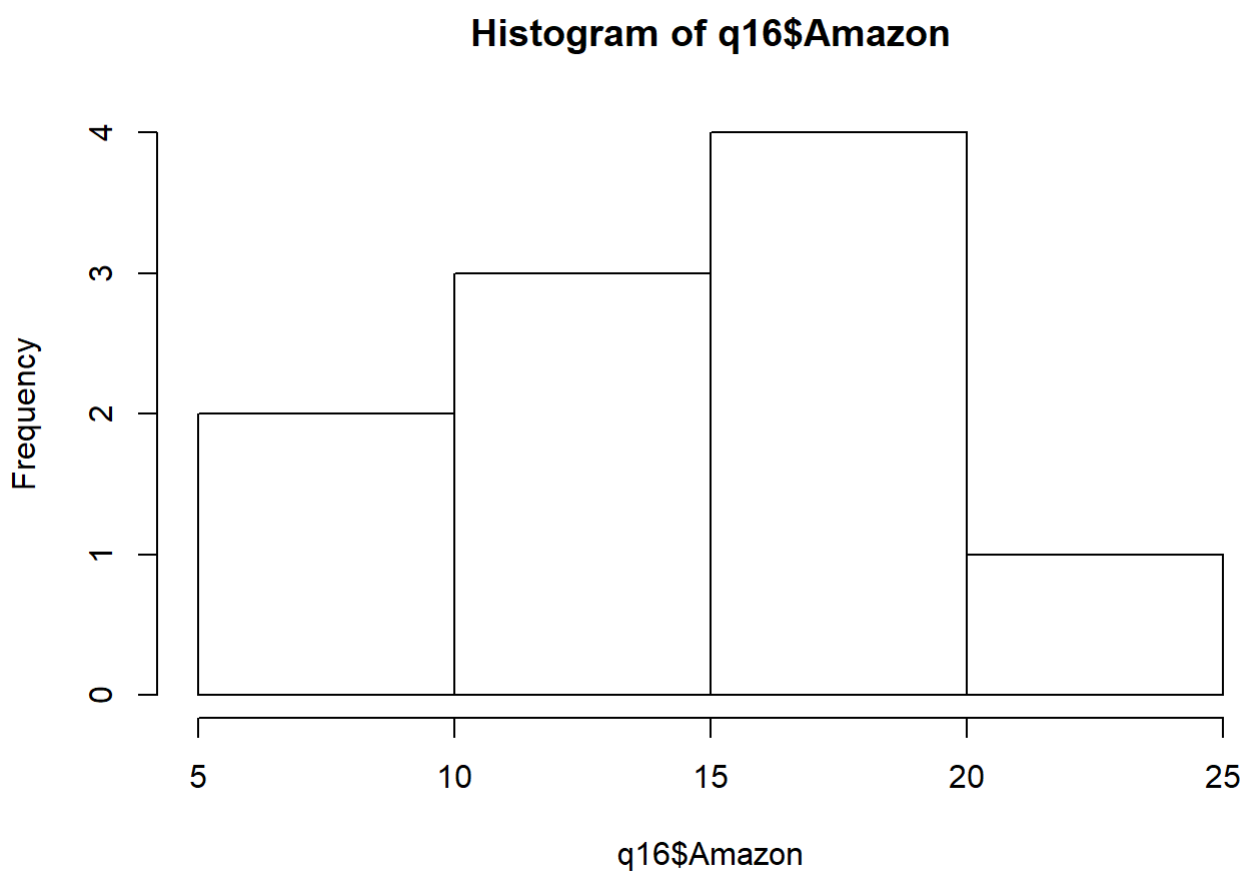
```
nortest::ad.test(q16$Amazon)
```

```
##
##  Anderson-Darling normality test
##
## data:  q16$Amazon
## A = 0.50318, p-value = 0.1547
```

```
#Histogram
hist(q16$BN)
```




```
hist(q16$Amazon)
```




```
#p-value>0.05 -->failed to reject null hypothesis-normality condition met

#As both normality and variance conditions are met,
#proceed with paired t-test
t.test(q16$BN,q16$Amazon,mu=0,paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  q16$BN and q16$Amazon
## t = 3.8173, df = 9, p-value = 0.004107
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.05540453 0.21659547
## sample estimates:
## mean of the differences
##                   0.136
```

```
#As p-value<0.05,Reject null hypothesis
#There is significant difference between prices of both the sellers
```

The p-values<0.05, Reject null hypothesis At 5% significance, so we conclude that there is difference in prices across the two sellers

Q17

All the necessary conditions are met for the goodness of fit test: 1.SRS -random sample of 100 observations-standardized 2.Contigency table consition met-intervals are mutually exclusive 3.Sample Size condition: expected counts>5

Hypotheses for the Chi-squared goodness of fit test:
- H_0 : $ p_{1} =p_{2}=p_{3}=p_{4}=p_{5}$ - H_a : At least one proportion differs from the expected normal distribution

```
observed <- c(10,18,48,16,8)
#Computing Expected Probabililties
# p(z<=-1.5)
p1 <- pnorm(-1.5,0,1)
# p(-1.5< z <= -0.5) = p(z<= -0.5) - p(z<= -1.5)
p2 <- pnorm(-0.5,0,1)-pnorm(-1.5,0,1)
# p(-0.5< z <= 0.5) = p(z<=0.5) - p(z<=-0.5)
p3 <- pnorm(0.5,0,1) - pnorm(-0.5,0,1)
# p(0.5< z <=1.5) = p(z<=1.5) - p(z<=0.5)
p4 <- pnorm(1.5,0,1) - pnorm(0.5,0,1)
# p(z>1.5)= 1- p(z<=1.5)
p5 <- pnorm(1.5,0,1,lower.tail=FALSE)

expected <-c(p1,p2,p3,p4,p5)

#Performing chi-square goodness of fit test
chisq.test(x=observed,p=expected)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 8.7104, df = 4, p-value = 0.06876
```

As the p-value>0.05 => Failed to Reject the null hypothesis As a result of which at 5% significance level, these counts are in agreement with the normal distribution

No,we can’t infer at the 5% significance level that the data were not drawn from a normal population But for significance level >=7%, we can reject the null hypothesis