In [4]:
```
install.packages('stats')
install.packages('PMCMR')
```

executed in 5.72s, finished 23:17:11 2019-11-18

. . .

In [7]:
```
library(stats)
library(PMCMR)
```

executed in 36ms, finished 23:17:39 2019-11-18

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PMCMR is superseded by PMCMRplus and will be no longer maintained.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

In [14]:
```
install.packages('PASWR2')
library (PASWR2)
```

executed in 6.46s, finished 14:49:58 2019-11-16

. . .

1 (2 points) Question 1

Let us say you have to decide between two distributions, d_1 and d_2 that a random variable X has. Given below are the probability distributions showing the probabilities that the distributions assign to the values x of X:

X=x	d_1	d_2
0	0.1	0.2
1	0.1	0.1
2	0.1	0.1
3	0.1	0.2
4	0.2	0.2
5	0.1	0.1
6	0.3	0.1

There is a single observation on X available and you want to test

H_0: d_1 is correct

H_1: d_2 is correct (d_1 is not correct)

Consider one possible decision procedure to fail to reject H_0 if X=4 or X=6 and answer the following two parts:

1.1 a. (1 point) Find the probability of a Type I error.

Type I error: H0 is true, but the test reject the H0

P(type I error) = 1-(P(x=4, d_0)+P(x=6, d_0))=1-(0.2+0.3)=0.5

1.2 b. (1 point) Find the probability of a Type II error.

Type II error: H0 is false, but the test did not reject the H0

P(type II error) = P(x=4, d_1)+P(x=6, d_1)=0.3

2 (8 points) Question 2

The current no-smoking regulations in office buildings require workers who smoke to take breaks and leave the building in order to satisfy their habits. A study indicates that such workers average 32 minutes per day taking smoking breaks. The standard deviation is 8 minutes. To help reduce the average break, rooms with powerful exhausts were installed in the buildings. To see whether these rooms serve their designed purpose, a random sample of 110 smokers was taken. The total amount of time away from their desks was measured for 1 day. Test to determine whether there has been a decrease in the mean time away from their desks. Compute the p-value and interpret it relative to the costs of Type I and Type II errors.

A financial analyst has determined that a 2-minute reduction in the average break would increase productivity. As a result, the company would hate to lose this opportunity. In such a case, calculate the probability of erroneously concluding that the renovation would not be successful. If this probability is high, describe how it can be reduced.

Assume this is a normal distribution.

$\varrho_{population}$ is known, so we use z test.

H_0: $\mu \geq 32$

H_1: $\mu < 32$

Assume the test is at 5% significant level.

In [3]:
```
setwd("D:/BAX441/Homeworks/Homeworks 4 and 5 Combined")
Q2<-read.csv('Question2.csv')
Q2
```

executed in 88ms, finished 14:37:23 2019-11-16

. . .

In [18]:
```
mean_pop<-32
sd_pop<-8


z.test(Q2$Minutes,mu=mean_pop,sigma.x=sd_pop,alternative='less')


print('At 5% significance level, there is sufficient evidence to warrant rejection of the claim t
```

executed in 69ms, finished 14:51:43 2019-11-16

```
        One Sample z-test

data:  Q2$Minutes
z = -2.7293, p-value = 0.003174
alternative hypothesis: true mean is less than 32
95 percent confidence interval:
     -Inf 31.17283
sample estimates:
mean of x
 29.91818

[1] "At 5% significance level, there is sufficient evidence to warrant rejection of the claim th
at the renovation would not work."
```

Type I Error: think the renovation would reduce the average smoke time, but actually it does not.

- Cost: spend money on stall the exhausts for no return.

Type II Error: think the renovation would not work, but actually it can reduct average smoke time.

- Cost: the company miss the opportunity to improve the productivity

In [112]:
```
qnorm(0.05,32,8/sqrt(110),lower.tail=TRUE)
```
executed in 29ms, finished 20:25:14 2019-11-16

30.7453548815335

In [114]:
```
beta<-pnorm(30.7453548815335,30,8/sqrt(110),lower.tail=FALSE)

cat('The probability of erroneously concluding that the renovation would not be successful is ',b
print('To reduce this probability, we can use higher significant level and increase the sample si
```
executed in 30ms, finished 20:26:45 2019-11-16

```
The probability of erroneously concluding that the renovation would not be successful is  0.1642
429 , which is high.[1] "To reduce this probability, we can use higher significant level and inc
rease the sample size."
```

3 (5 points) Question 3

This is a case on test marketing. A manufacturer of hardware products needs some help in deciding the price to charge for a new product. The marketing analyst through the use of pricing analytics determined that the new product should sell for 10, but the pricing model made some assumptions that the marketing manager was uncomfortable with. The manager is unsure if the sales volume will differ significantly if it is priced at 9 or 11. To conduct a pricing experiment, she distributes the new product to a sample of 60 stores. These 60 stores are all located in similar neighborhoods. The manager randomly selects 20 stores in which to sell the item at 9, 20 stores to sell it at 10, and the remaining 20 stores to sell it at 11. Sales at the end of the trial period were recorded. Although I do not have the dataset containing the individual sales figures, enough information is available for you to run an appropriate statistical test using principles. The mean sales at stores which sell the item at 9 is 153.60, at 10 is 151.50, and at 11 is 133.25. The standard deviations of sales are 25.57, 30.39, and 25.03 at stores which sell the item at 9, 10, and 11, respectively. Based on the given information, run an appropriate statistical test and describe what should the manager conclude.

In [7]:
```
n9<-20
n10<-20
n11<-20

mean9<-153.6
mean10<-151.5
mean11<-133.25

sd9<-25.57
sd10<-30.39
sd11<-25.03

#One-way ANOVA Requirements:
#Since the total sample size is 60, we assume the distribution of the sample is normal distributi
#Satisfy the requirement of randomly selected and independent
#Although the variances are not equal, with equal sample sizes, violations of this assumption do

#H0: mean9 = mean10 = mean11 (sales of products with different price are equal)
#H1: At least one sales is different
# 5% significant level

mean_grand<-(mean9+mean10+mean11)/3
sst<-(n9*(mean9-mean_grand)^2)+(n10*(mean10-mean_grand)^2)+(n11*(mean11-mean_grand)^2)
sse<-((n9-1)*sd9^2)+((n10-1)*sd10^2)+((n11-1)*sd11^2)

mst<-sst/(3-1)
mse<-sse/(60-3)
f<-mst/mse

cat ('Critical Value: ', qf(0.95, df1=3-1, df2=60-3),'\n')

cat('F score: ',f,'\n')

##########################Conclusion################################################

print('With 5% significant level, we have enough evidence to reject the H0.')
print('Conclusion: there is difference on sales among the prices. ')

###################Tukey's Honestly Significant Difference (HSD)#################
print('\n')
cat('|mean9-mean10|: ',abs(mean9-mean10),'\n')
cat('|mean9-mean11|: ',abs(mean9-mean11),'\n')
cat('|mean10-mean11|: ',abs(mean10-mean11),'\n')

#LSD
cat('Traditional LSD: ',qt(0.975,df=60-3)*sqrt(mse*(1/20+1/20)),'\n')

cat('Tukey LSD: ',qtukey(0.95, 3, 60-3)*sqrt(mse/20),'\n')
```

executed in 73ms, finished 17:11:33 2019-11-17

```
Critical Value:  3.158843
F score:  3.41033
[1] "With 5% significant level, we have enough evidence to reject the H0."
[1] "Conclusion: there is difference on sales among the prices. "
[1] "\n"
|mean9-mean10|:  2.1
|mean9-mean11|:  20.35
|mean10-mean11|:  18.25
Traditional LSD:  17.1632
Tukey LSD:  20.62549
```

4 (20 points) Question 4

If you walk down the aisle of breakfast cereals at a grocery store, you will find that the cereal market is flooded with a bewildering number of breakfast cereals. Each company produces several different kinds of cereal in the belief that there are distinct market segments. For example, there is a market segment comprised primarily of children, another segment for diet-conscious adults, and yet another for health-conscious adults. Each cereal the companies produce has at least one market segment as its target. However, we consumers make our own decisions, which may or may not match the target predicted by the cereal maker.

One such cereal maker in the northeastern part of the United States is attempting to distinguish between consumers. It administered a survey to adults between 25 and 65. Each was asked a few questions, including age, income, and years of education, as well as the brand of cereal each consumed most frequently. The cereal choices are:

1. Sugar Rush, a children's cereal
2. Special K, a cereal aimed at dieters
3. Fiber One, a cereal that is advertised as healthy
4. Cheerios, a cereal that is targeted to a combination of dieters and health conscious consumers

The results of the survey were recorded. The data are available under the CSV file Question 4. The first column contains the cereal choice, second column records the age of respondent, third column records the annual household income, and the fourth column records the years of education.

Run an appropriate statistical test to address the following objectives:

In [51]:
```
setwd("D:/BAX441/Homeworks/Homeworks 4 and 5 Combined")
Q4<-read.csv('Question4.csv')
Cereal<-as.factor(Q4$Cereal)
Age<-Q4$Age
Income<-Q4$Income
Education<-Q4$Education
Q4
```

executed in 57ms, finished 15:42:06 2019-11-17

Cereal	Age	Income	Education
1	33	26	8
1	27	49	11
1	25	35	12
1	31	35	12
1	35	41	15
1	29	46	14
1	31	41	10
1	35	37	12
1	27	45	11
1	25	44	10
1	37	49	10
1	55	36	14

4.1 a. (5 points) Are there differences between the ages of the consumers of the four cereals?

In [53]:
```
#########Requirement Test#######################
#1. Randomly selected sample - Satisfy
#2. Normality - Satisfy, because the p-value of the normality test is larger than 0.05
model_ex41 <- lm(Age ~ Cereal)
resids_ex41 <- residuals(model_ex41)
preds_ex41 <- predict(model_ex41)

nortest::ad.test(resids_ex41)

#hist(resids_ex41)

#qqnorm(resids_ex41)
#qqline(resids_ex41)

#3. Variances are constant - Unsatisfy, because the p-value of the homogeneity f variances test i
fligner.test(Age ~ Cereal)
car::leveneTest(Age ~ Cereal)
########Welch's test############################
#H0: no difference between the ages of the consumers of the four cereals
#H1: At least one cereal's customers' age is different from others
oneway.test(Age ~ Cereal, var.equal = FALSE)

#install.packages("userfriendlyscience")
library(userfriendlyscience)

posthocTGH(Age, Cereal, method = "games-howell")

########Conclusion################################
print('Yes, at 5% significant level, there are differences between ages of consumers of the four
```

executed in 91ms, finished 15:42:20 2019-11-17

```
        Anderson-Darling normality test

data:  resids_ex41
A = 0.74779, p-value = 0.051


        Fligner-Killeen test of homogeneity of variances

data:  Age by Cereal
Fligner-Killeen:med chi-squared = 34.012, df = 3, p-value = 1.97e-07
```

	Df	F value	Pr(>F)
group	3	12.08694	1.770382e-07
	291	NA	NA

```
        One-way analysis of means (not assuming equal variances)

data:  Age and Cereal
F = 25.154, num df = 3.00, denom df = 132.24, p-value = 6.11e-13

    n means variances
1  63    31        28
2  81    34        23
3  40    37        31
4 111    40        72

    diff ci.lo ci.hi   t  df    p
2-1  3.1  0.88   5.4 3.6 126 <.01
3-1  6.1  3.16   9.0 5.5  80 <.01
4-1  8.6  5.91  11.3 8.2 170 <.01
3-2  3.0  0.24   5.7 2.9  68  .03
```

```
4-2  5.5  3.00   8.0 5.7 180 <.01
4-3  2.6 -0.57   5.7 2.1 105  .15

[1] "Yes, at 5% significant level, there are differences between ages of consumers of the four c
ereal."
```

4.2 b. (5 points) Are there differences between the incomes of the consumers of the four cereals?

In [54]:
```
#########Requirement Test#######################
#1. Randomly selected sample - Satisfy
#2. Normality - Satisfy, because the p-value of the normality test is larger than 0.05
model_ex42 <- lm( Income~  Cereal)
resids_ex42 <- residuals(model_ex42)
preds_ex42 <- predict(model_ex42)

nortest::ad.test(resids_ex42)

#hist(resids_ex41)

#qqnorm(resids_ex41)
#qqline(resids_ex41)

#3. Variances are constant - Satisfy, because the p-value of the normality test is larger than 0.(
fligner.test( Income~  Cereal)
car::leveneTest(Income~  Cereal)
########One-way ANOVA############################
#H0: no difference between the incomes of the consumers of the four cereals
#H1: At least one cereal's customers' income is different from others
ANOVA_Ex42 <- aov( Income~  Cereal)
summary(ANOVA_Ex42)

########Conclusion################################
print('Yes, at 5% significant level, there are differences between incomes of consumers of the fo
```

executed in 50ms, finished 15:42:42 2019-11-17

```
        Anderson-Darling normality test

data:  resids_ex42
A = 0.38897, p-value = 0.3826


        Fligner-Killeen test of homogeneity of variances

data:  Income by Cereal
Fligner-Killeen:med chi-squared = 4.0702, df = 3, p-value = 0.254
```

	Df	F value	Pr(>F)
group	3	1.301972	0.2739744
	291	NA	NA

```
             Df Sum Sq Mean Sq F value  Pr(>F)
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

[1] "Yes, at 5% significant level, there are differences between incomes of consumers of the fou
r cereal."
```

4.3 c. (5 points) Are there differences between the educational levels of the consumers of the four cereals?

In [62]:
```
#########Requirement Test#######################
#1. Randomly selected sample - Satisfy
#2. Normality - Unsatisfy, because the p-value of the normality test is less than 0.05
model_ex43 <- lm(Education ~  Cereal)
resids_ex43 <- residuals(model_ex43)
preds_ex43 <- predict(model_ex43)

nortest::ad.test(resids_ex43)

#hist(resids_ex41)

#qqnorm(resids_ex41)
#qqline(resids_ex41)

#3. Variances are constant - Satisfy. Although the variances are not equal, with equal sample siz
#violations of this assumption do not seriously affect inferences
fligner.test(Education ~  Cereal)
car::leveneTest(Education ~  Cereal)
########Kruskal-Wallis Test############################
#H0: no difference between the education levels of the consumers of the four cereals
#H1: At least one cereal's customers' education level is different from others
kruskal.test(Education ~  Cereal)


########Conclusion###############################
print('No, there is no difference between education levels of consumers of the four cereal.')
```

executed in 55ms, finished 20:31:27 2019-11-13

```
        Anderson-Darling normality test

data:  resids_ex43
A = 0.87762, p-value = 0.02435


        Fligner-Killeen test of homogeneity of variances

data:  Education by Cereal
Fligner-Killeen:med chi-squared = 1.4531, df = 3, p-value = 0.6931
```

	Df	F value	Pr(>F)
group	3	0.5156102	0.671837
	291	NA	NA

```
        Kruskal-Wallis rank sum test

data:  Education by Cereal
Kruskal-Wallis chi-squared = 5.666, df = 3, p-value = 0.129

[1] "No, there is no difference between education levels of consumers of the four cereal."
```

4.4 d. (5 points) Prepare a summary for the cereal maker describing the differences between the four groups of cereal consumers.

In [64]:
```
###############Summary#########################
#These 4 groups of cereal consumers do have differences of ages and income levels, but not educat

posthocTGH(Age, Cereal, method = "games-howell")
TukeyHSD(ANOVA_Ex42)
```

executed in 58ms, finished 20:32:56 2019-11-13

```
    n means variances
1  63    31        28
2  81    34        23
3  40    37        31
4 111    40        72

    diff ci.lo ci.hi   t  df    p
2-1  3.1  0.88   5.4 3.6 126 <.01
3-1  6.1  3.16   9.0 5.5  80 <.01
4-1  8.6  5.91  11.3 8.2 170 <.01
3-2  3.0  0.24   5.7 2.9  68  .03
4-2  5.5  3.00   8.0 5.7 180 <.01
4-3  2.6 -0.57   5.7 2.1 105  .15


  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Income ~ Cereal)

$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
```

There are age difference between cereal 1 and 2, 1 and 3, 1 and 4, 2 and 3, 2and 4. So the age of consumers of 1 is younger than other brands. Brand 2 is the second younger group. There is no significant difference of ages between brand 3 and brand 4.

There are income difference between cereal 1 and 3, 1 and 4, 2 and 4. Brand 1 customers have lower income level than brand 3 and brand 4. Brand 2 customers havve lower income level than brand 4.

1. Sugar Rush, a children's cereal
2. Special K, a cereal aimed at dieters
3. Fiber One, a cereal that is advertised as healthy
4. Cheerios, a cereal that is targeted to a combination of dieters and health conscious consumers

5 (5 points) Question 5

The first thing that comes to mind when we want to lose weight is to eat less. Of course, we all know that exercise also plays a key role, but many find it hard to fit exercise in their daily routine. Hence, going on a diet (controlling what we eat) seems to be low-hanging. How well do diets work? In a study, 20 people who were more than 20 pounds overweight, were recruited to compare four diets. The people were matched by age. The oldest four became group 1, the next oldest four became group 2, and so on There were five groups in all. The number of pounds that each person lost by following one of the four diets were recorded. Although I do not have access to the data but have enough information for you to run the appropriate statistical test. The mean weight lost by following diet 1, diet 2, diet 3, and diet 4 was 6.2 pounds, 8.0 pounds, 10.8 pounds, and 8.2 pounds, respectively. The mean weight lost by groups 1, 2, 3, 4, and 5 was 5.25 pounds, 7.25 pounds, 7.25 pounds, 10.25 pounds, and 11.5 pounds, respectively. Use this information

and run an appropriate statistical test from principles to infer at 1% significance level whether there are differences among the four diets. What experimental design was used? From your analysis, state if the experimental design was sound.

	Diet			
Group	1	2	3	4
1	5	2	6	8
2	4	7	8	10
3	6	12	9	2
4	7	11	16	7
5	9	8	15	14

In [1]:
```
x <- data.frame("pounds" = c(5,2,6,8,4,7,8,10,6,12,9,2,7,11,16,7,9,8,15,14),
                "group" = c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4,5,5,5,5), "diet" = c(1,2,3,4,1,2,3,4,
pounds<-x$pounds
group<-as.factor(x$group)
diet<-as.factor(x$diet)
```

executed in 49ms, finished 16:56:49 2019-11-17

In [2]:
```
#########Requirement Test#######################
#1. Randomly selected sample - Satisfy
#2. Normality - Satisfy, because the p-value of the normality test is larger than 0.05
model_ex5 <- lm(pounds ~  group+diet)
resids_ex5 <- residuals(model_ex5)
preds_ex5 <- predict(model_ex5)

nortest::ad.test(resids_ex5)

#hist(resids_ex41)

#qqnorm(resids_ex41)
#qqline(resids_ex41)

#3. Variances are constant - Satisfy. Although the variances are not equal, with equal sample siz
#violations of this assumption do not seriously affect inferences
fligner.test(pounds ~  diet)
car::leveneTest(pounds ~  diet)

fligner.test(pounds ~  group)
car::leveneTest(pounds ~  group)
############Random Block Design#############################
#H0: There are not differences among the four diets
#H1: At least one diet is different from others
ANOVA5 <- aov(pounds ~  group+diet)
summary(ANOVA5)

########################Conclusion#############################################
print('At 1% significant level, there is no difference among these 4 diet.')
```

executed in 453ms, finished 16:56:51 2019-11-17

```
        Anderson-Darling normality test

data:  resids_ex5
A = 0.17884, p-value = 0.905


        Fligner-Killeen test of homogeneity of variances

data:  pounds by diet
Fligner-Killeen:med chi-squared = 1.9144, df = 3, p-value = 0.5904
```

	Df	F value	Pr(>F)
group	3	0.6227709	0.610514
	16	NA	NA

```
        Fligner-Killeen test of homogeneity of variances

data:  pounds by group
Fligner-Killeen:med chi-squared = 2.4353, df = 4, p-value = 0.6563
```

	Df	F value	Pr(>F)
group	4	0.798913	0.5443667
	15	NA	NA

```
            Df Sum Sq Mean Sq F value Pr(>F)
group        4  102.2   25.55   2.355  0.113
diet         3   53.8   17.93   1.653  0.230
Residuals   12  130.2   10.85

[1] "At 1% significant level, there is no difference among these 4 diet."
```

6 (5 points) Question 6

A recruiter for a computer company would like to determine whether there are differences in sales ability between business, arts, and science graduates. She takes a random sample of 20 business graduates who have been working for the company for the past 2 years. Each is then matched with an arts graduate and a science graduate with similar educational and working experience. The commission earned by each (in $1,000s) in the last year was recorded. The dataset is available on the CSV file Question6. Use 5% significance level.

In [44]:
```
setwd("D:/BAX441/Homeworks/Homeworks 4 and 5 Combined")
Q6<-read.csv('Question6.csv')
Q6$Group<-as.factor(Q6$Group)
Q6<-stack(Q6)
names(Q6) <- c("Commission", "Major")
list1<-rep(1:20, times=3)
Group<-data.frame(list1)
Q6<-cbind(Q6,Group)
names(Q6) <- c("Commission", "Major","Group")
Q6
```

executed in 59ms, finished 15:39:48 2019-11-17

. . .

6.1 a. Is there sufficient evidence to allow the recruiter to conclude that there are differences in sales ability between the holders of the three types of degrees?

In [47]:
```
########One-Way ANOVA#######################
#H0: There is no differences in sales ability between the holders of the three types of degrees
#H1: At least one type of degree's holders have different sales ability

Group <- factor(Q6$Group)
Commission<-Q6$Commission

ANOVA61 <- aov(Commission ~  Major)
summary(ANOVA61)
print('At 5% significant level, there is no difference bwteen the holders of the three types of d
```

executed in 40ms, finished 15:40:15 2019-11-17

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

[1] "At 5% significant level, there is no difference bwteen the holders of the three types of de
grees."
```

6.2 b. Conduct a test to assess if any other experimental design would have been a better choice.

In [48]:
```
###########Random Block Design#############################
#H0: There is no differences in sales ability between the holders of the three types of degrees
#H1: At least one type of degree's holders have different sales ability

Group <- factor(Q6$Group)
Commission<-Q6$Commission
Major<-factor(Q6$Major)

ANOVA6 <- aov(Commission ~ Group + Major)
summary(ANOVA6)

print('At 5% significant level, there is no difference bwteen the holders of the three types of d
```

executed in 33ms, finished 15:40:39 2019-11-17

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

[1] "At 5% significant level, there is no difference bwteen the holders of the three types of de
grees."
```

6.3 c. What are the required conditions for the test in Part a? Are the required conditions satisfied?

1. Normality - Satisfy
2. Independence - Satisfy
3. Homogeneity of Variances - Satisfy

In [36]:
```
#Test normality
model_ex6 <- lm(Commission ~ Group + Major)
resids_ex6 <- residuals(model_ex6)
preds_ex6 <- predict(model_ex6)

nortest::ad.test(resids_ex6)

#hist(resids_ex6)

#qqnorm(resids_ex6)
#qqline(resids_ex6)

print("This is a normal distribution")
```

executed in 34ms, finished 20:03:33 2019-11-13

```
        Anderson-Darling normality test

data:  resids_ex6
A = 0.23463, p-value = 0.7838

[1] "This is a normal distribution"
```

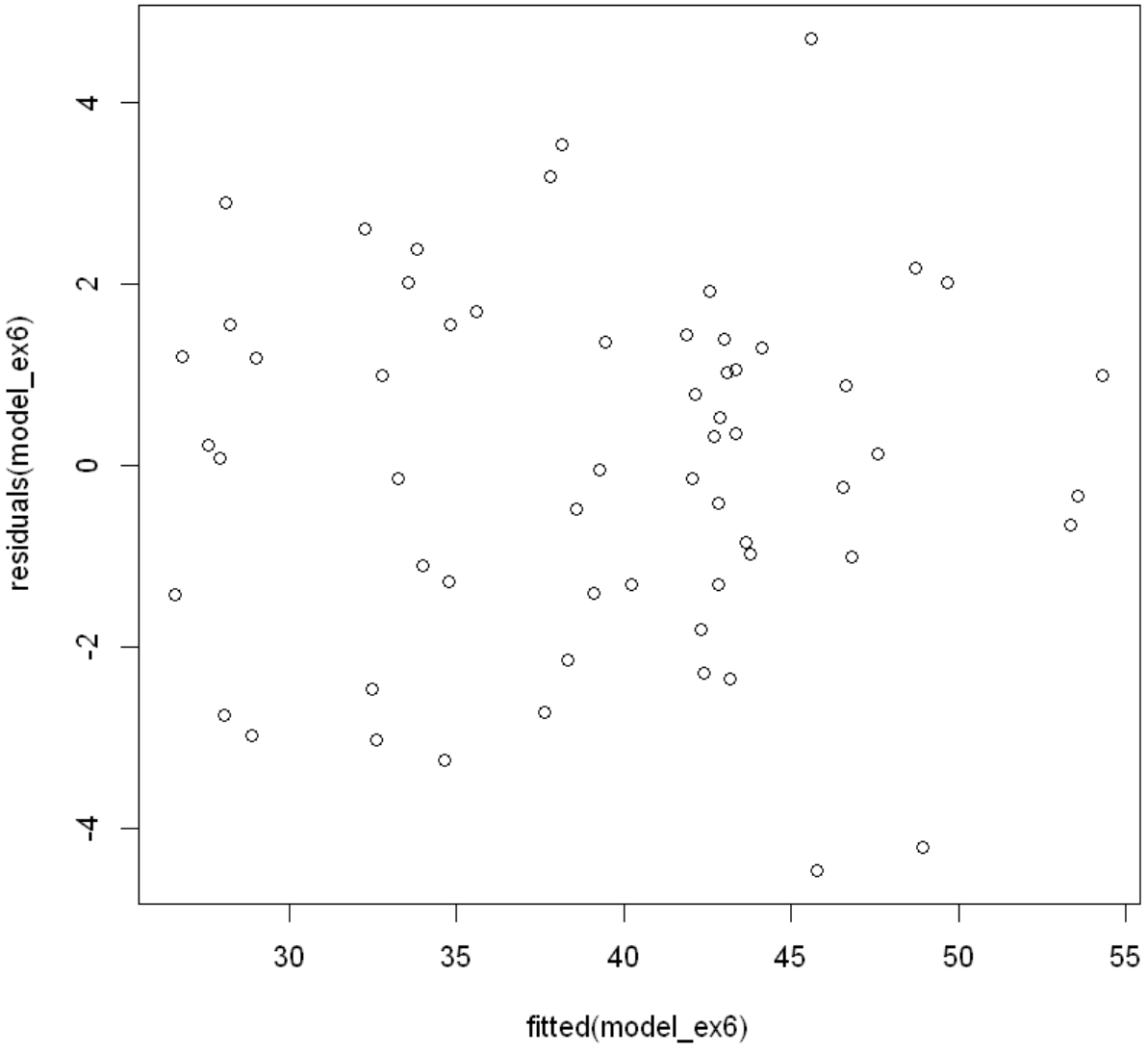
In [37]:
```
#Test Homogeneity of Variances
car::leveneTest(Commission ~  Major)
car::leveneTest(Commission ~  Group)

plot(fitted(model_ex6), residuals(model_ex6))
print('Satisfy the homogeneity of variances')
```

executed in 79ms, finished 20:03:36 2019-11-13

	Df	F value	Pr(>F)
group	2	0.4145875	0.6625886
	57	NA	NA

	Df	F value	Pr(>F)
group	19	0.4560112	0.9658094
	40	NA	NA

```
[1] "Satisfy the homogeneity of variances"
```



7 (5 points) Question 7

Statistical techniques are used to determine auto insurance premiums. The premiums are proportional to the risks and costs of accidents. AutoState, a leading insurance agency in the United States, hired an analyst to conduct a study that looked at miles driven in the previous year, ages of the drivers, and their gender. The age categories are 16-19, 20-34, 35-54, 55-64, and 65+. The dataset is provided on the CSV file Question7. Conduct an appropriate statistical test at 5% significance level to determine if we have enough evidence to conclude that males and female drivers differ in the number of miles they drive. Can we infer that there are differences between the age categories in the number of miles they drive? What other analysis can you perform to generate additional insights for statistical inference?

In [1]:
```
library(dplyr)
library(tidyverse)
```

executed in 1.14s, finished 13:10:02 2019-11-14

. . .

In [24]:
```
setwd("D:/BAX441/Homeworks/Homeworks 4 and 5 Combined")
Q7<-read.csv('Question7.csv')
Q7<-as.data.frame(Q7)
names(Q7) <- c("Gender", "age1",'age2','age3','age4','age5')
df1<-Q7[Q7[, "Gender"] == 'Males',]
colnames(df1)<-c('Gender','1','2','3','4','5')
df1<-stack(df1)
df1$gender<-'male'

df2<-Q7[Q7[, "Gender"] == 'Females',]
colnames(df2)<-c('Gender','1','2','3','4','5')
df2<-stack(df2)
df2$gender<-'female'
df<-rbind(df1, df2)
colnames(df)<-c('miles','age','gender')
df
```

executed in 62ms, finished 13:53:15 2019-11-14

. . .

In [26]:

```
#GENDER
############Requirement Test##############################
df_test<-df
df_test$agegender <- paste(df$age, df$gender, sep="_")

gender<-as.factor(df$gender)
age<-as.factor(df$age)
miles<-as.numeric(df$miles)
agegender<-factor(df_test$agegender)
#1. Normality - Unisatisfy, because the p-value is smaller than 0.05
model_ex7 <- lm(miles ~ gender)
resids_ex7 <- residuals(model_ex7)
preds_ex7 <- predict(model_ex7)
nortest::ad.test(resids_ex7)
shapiro.test(resids_ex7)
#2. Constant Variance - Unsatisfy, because the p-value is smaller than 0.05
car::leveneTest(miles ~  gender)
#3. Independent - Satisfy

########Kruskal-Wallis Test#############################
#H0: There is not difference between males and female in the number of miles they drive
#H1: There is difference between males and female in the number of miles they drive

kruskal.test(miles ~  gender)

############Conclusion##############################
#At 5% significance level, there is enough evidence to conclude that
#males and female drivers differ in the number of miles they drive.
```

executed in 65ms, finished 13:58:42 2019-11-14

```
        Anderson-Darling normality test

data:  resids_ex7
A = 1.3937, p-value = 0.001285


        Shapiro-Wilk normality test

data:  resids_ex7
W = 0.97747, p-value = 0.00264
```

	Df	F value	Pr(>F)
group	1	22.58946	3.850781e-06
	198	NA	NA

```
        Kruskal-Wallis rank sum test

data:  miles by gender
Kruskal-Wallis chi-squared = 60.259, df = 1, p-value = 8.318e-15
```

In [30]:

```
#AGE
############Requirement Test#############################
df_test<-df
df_test$agegender <- paste(df$age, df$gender, sep="_")

gender<-as.factor(df$gender)
age<-as.factor(df$age)
miles<-as.numeric(df$miles)
agegender<-factor(df_test$agegender)
#1. Normality - Satisfy, because the p-value is smaller than 0.05
model_ex7 <- lm(miles ~ age)
resids_ex7 <- residuals(model_ex7)
preds_ex7 <- predict(model_ex7)
nortest::ad.test(resids_ex7)
shapiro.test(resids_ex7)
#2. Constant Variance - Unsatisfy, because the p-value is smaller than 0.05
car::leveneTest(miles ~  age)
#3. Independent - Satisfy

########Welch's test#########################
#H0: There is not difference among ages in the number of miles they drive
#H1: There is difference among ages in the number of miles they drive

oneway.test(miles ~ age, var.equal = FALSE)

############Conclusion#############################
#At 5% significance level, there is enough evidence to say that there are differences
#between the age categories in the number of miles they drive.
```

executed in 49ms, finished 14:03:17 2019-11-14

```
        Anderson-Darling normality test

data:  resids_ex7
A = 0.54848, p-value = 0.1562


        Shapiro-Wilk normality test

data:  resids_ex7
W = 0.99264, p-value = 0.4143
```

	Df	F value	Pr(>F)
group	4	13.07112	1.889869e-09
	195	NA	NA

```
        One-way analysis of means (not assuming equal variances)

data:  miles and age
F = 56.85, num df = 4.000, denom df = 90.827, p-value < 2.2e-16
```

In [34]:
```
######More Insight###############################
#install.packages("userfriendlyscience")
library(userfriendlyscience)

posthocTGH(miles, age, method = "games-howell")
posthocTGH(miles, gender, method = "games-howell")
```

executed in 76ms, finished 14:04:49 2019-11-14

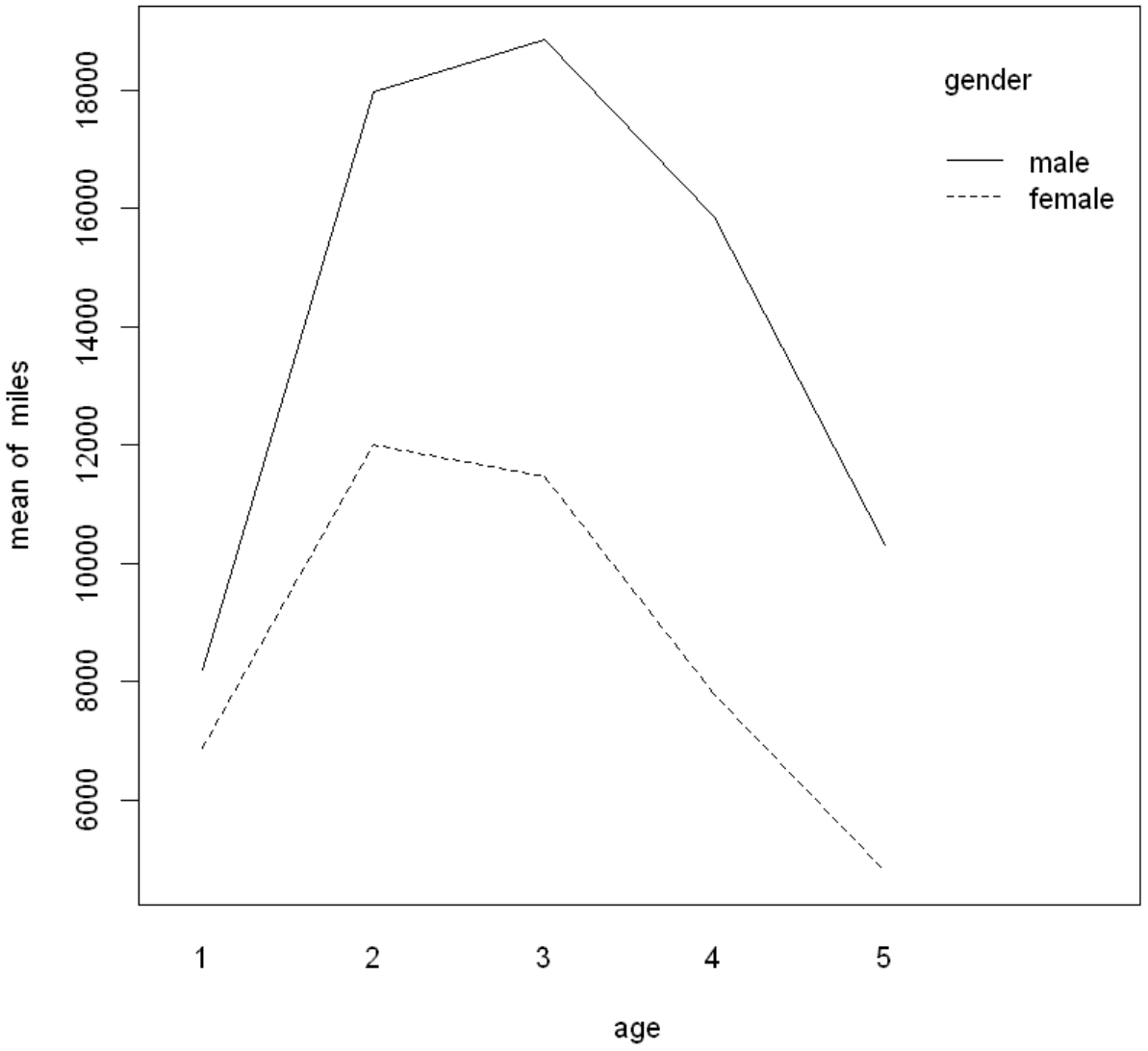
```
   n means variances
1 40  7539   2.2e+06
2 40 14989   1.5e+07
3 40 15160   2.1e+07
4 40 11819   2.2e+07
5 40  7544   9.9e+06

       diff  ci.lo ci.hi       t df    p
2-1  7450.4   5620  9281 11.5123 51 <.01
3-1  7621.4   5463  9779 10.0151 47 <.01
4-1  4280.3   2066  6495  5.4828 47 <.01
5-1     5.2  -1548  1559  0.0094 56    1
3-2   171.0  -2460  2802  0.1816 76    1
4-2 -3170.1  -5846  -494  3.3114 75  .01
5-2 -7445.2  -9630 -5261  9.5264 75 <.01
4-3 -3341.1  -6239  -443  3.2198 78  .02
5-3 -7616.2 -10076 -5157  8.6734 69 <.01
5-4 -4275.1  -6784 -1766  4.7755 68 <.01

         n means variances
female 100  8581   1.1e+07
male   100 14240   2.3e+07

            diff ci.lo ci.hi   t  df    p
male-female 5659  4503  6815 9.7 177 <.01
```

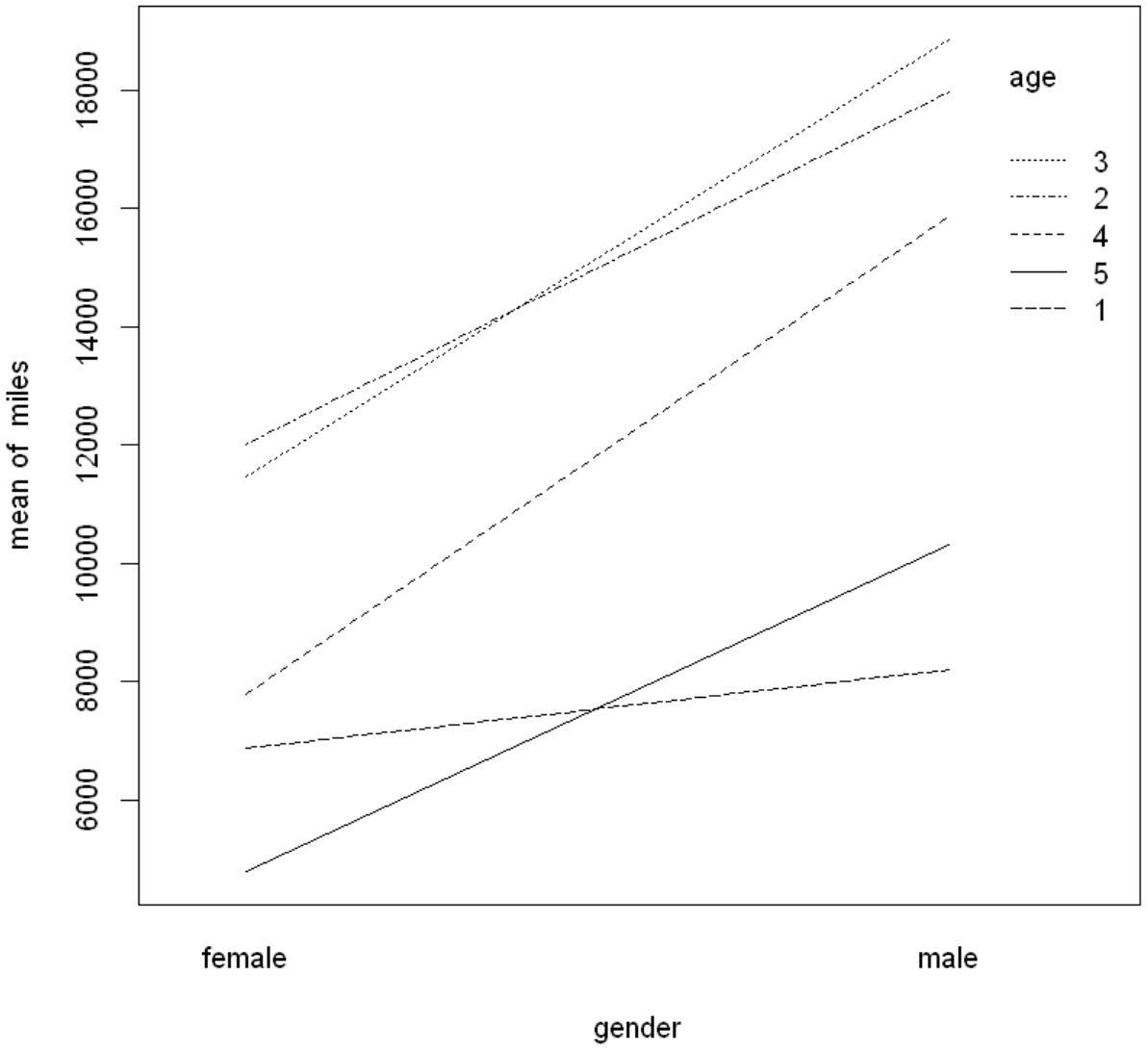
In [127]:

```
# Also use 2 factor ANOVA. The interaction of gender and age plays a significant role.
ANOVA7 <- aov(miles ~ gender + age+gender*age)
summary(ANOVA7)

model_ex7 <- lm(miles ~ gender + age+gender*age)
resids_ex7 <- residuals(model_ex7)
preds_ex7 <- predict(model_ex7)

residuals_Ex7 <- resid(ANOVA7)

#plot(fitted(model_ex7), residuals(model_ex7))

nortest::ad.test(residuals_Ex7)

#qqnorm(residuals_Ex7)
#qqline(residuals_Ex7)

shapiro.test(residuals_Ex7)

interaction.plot(age, gender, miles)
interaction.plot(gender, age, miles)

#There is a significant interaction effect on the miles.
```

executed in 115ms, finished 19:20:10 2019-11-14

```
             Df    Sum Sq   Mean Sq F value   Pr(>F)
gender        1 1.601e+09 1.601e+09  361.57  < 2e-16 ***
age           4 2.279e+09 5.697e+08  128.64  < 2e-16 ***
gender:age    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


        Anderson-Darling normality test

data:  residuals_Ex7
A = 0.49791, p-value = 0.2089


        Shapiro-Wilk normality test

data:  residuals_Ex7
W = 0.98936, p-value = 0.1438
```


mean of miles
6000
8000
10000
12000
14000
16000
18000
1
2
3
4
5
age
gender
male
female



8 (5 points) Question 8

On Saturday morning, Barbara Kenworth received a call at her home from the production supervisor at Kenworth Electronics Plant 1. The supervisor indicated that she and the supervisors from Plants 2, 3, and 4 had agreed that something must be done to improve company morale and thereby increase the production output of their plants. Barbara Kenworth, president of Kenworth Electronics, agreed to set up a Monday morning meeting with the supervisors to see if they could arrive at a plan for accomplishing these objectives.

By Monday, each supervisor had compiled a list of several ideas, including a four-day work week and interplant competitions of various kinds. A second meeting was set for Wednesday to discuss the issue further.

Following the Wednesday afternoon meeting, Barbara Kenworth and her plant supervisors agreed to implement a weekly contest called the BEST Game of the Week. The plant producing the most each week would be considered the BEST Game of the Week winner and would receive 10 points. The second-place plant would receive 7 points, and the third- and fourth-place plants would receive 3 points and 1 point, respectively. The contest would last 26 weeks. At the end of that period, a $200,000 bonus would be divided among the employees in the four plants proportional to the total points accumulated by each plant.

The announcement of the contest created a lot of excitement and enthusiasm at the four plants. No one complained about the rules because the four plants were designed and staffed to produce equally.

At the close of the contest, Barbara Kenworth called the supervisors into a meeting, at which time she asked for data to determine whether the contest had significantly improved productivity. She indicated that she had to know this before she could authorize a second contest. The supervisors, expecting this request, had put together the data available. The data, shown below, contains the proportion of days each category of units was produced before the contest and the number of days each category of units was produced during the contest.

Units Produced (4 plant total)	Proportion of Days (Before Contest)	During-Contest No.of Days
0-2500	0.105	0
2501-8000	0.219	20
8001-15000	0.533	83
15001-20000	0.143	52

Barbara examined the data and said, “There must be some way to statistically test the worthiness of this contest. I have to see the results before I will authorize the second contest.”

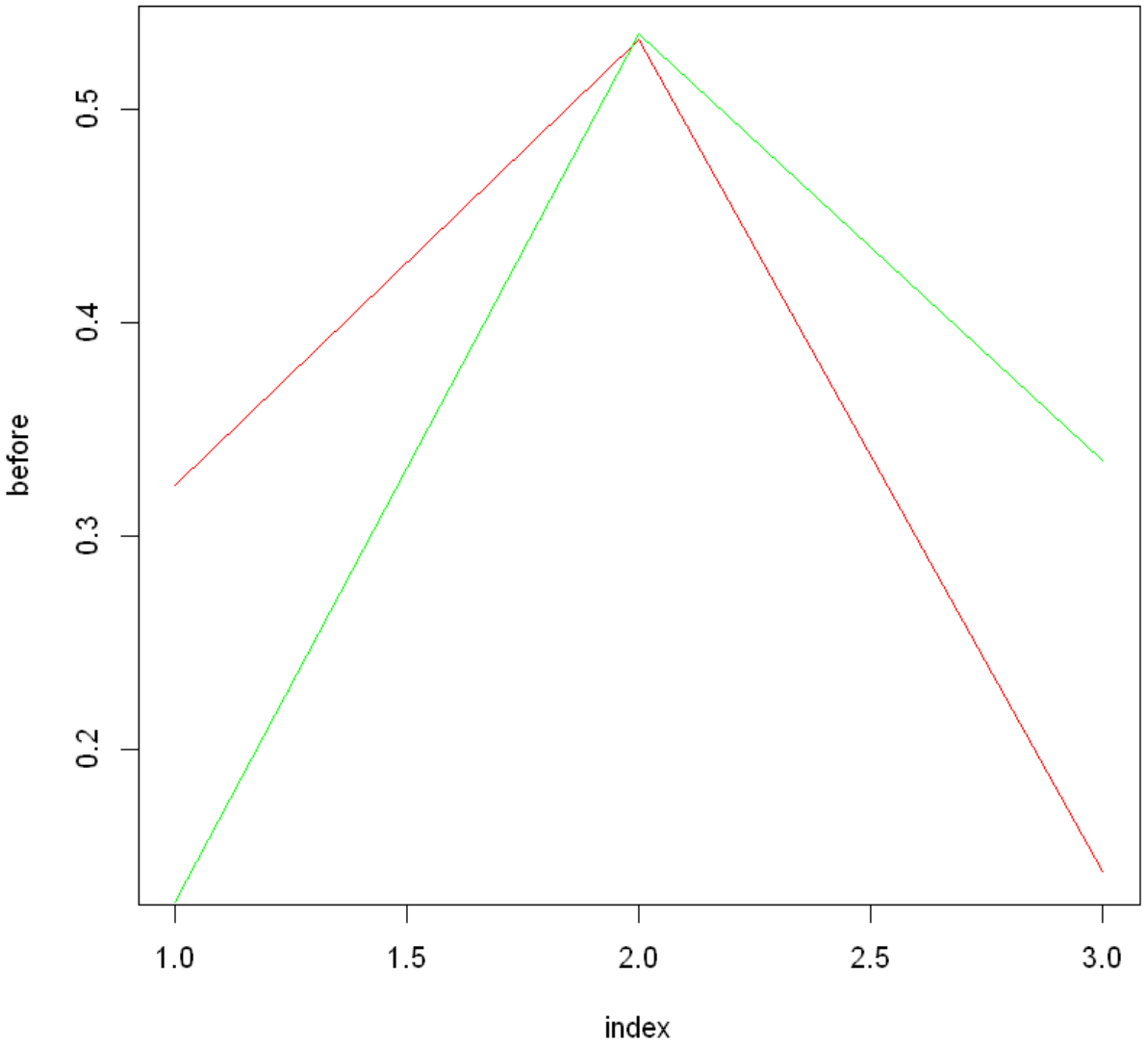
Your job as an analyst is to help Barbara assess the statistical significance of the results of the contest by running an appropriate statistical test.

In [85]:

```
days<-20+83+52
before<-c(0.105,0.219,0.533,0.143)
after<-c(0/days,20/days,83/days,52/days)

#combine the first 2 rows
before<-c(0.105+0.219,0.533,0.143)
after<-c(20/days,83/days,52/days)
index<-c(1,2,3)
plot(index, before, type='l',col="red")
lines(index, after, col="green")

############Categorical Fit Test###############
#H0: before = after
#H1: at least one p is different

chisq.test(x=after*days,p=before)

#########Conclusion########################### -1
print('At 5% significant level, we have enough evidence to say there are differences between cont

############Confidencial Interval for The Third Row###########
#H0: contest_3 <= non-contest_3
#H1: contest_3 > non-contest_3

z95<-qnorm(0.95,0,1)
#############Conclusion#################### -2
cat('The lower critical value is ',52/days-z95*sqrt((52/days)*(1-(52/days))/days), ', which is hi
print('Yes, the company should do the second contest.')
```

executed in 85ms, finished 19:27:22 2019-11-16

```
        Chi-squared test for given probabilities

data:  after * days
X-squared = 58.346, df = 2, p-value = 2.14e-13

[1] "At 5% significant level, we have enough evidence to say there are differences between conte
st and non-contest."
The lower critical value is  0.2731032 , which is higher than the before (0.143).[1] "Yes, the c
ompany should do the second contest."
```



9 (5 points) Question 9

A multinational investment banking firm on the Wall Street is trying to devise a new investment strategy for its clients. Investment banking firms keep a close watch on how the market performed each day of the week so that it can monitor the soundness of the strategies. They hire “quants” to design and formulate such strategies. Quants usually have PhD in Math or Physics and they use financial engineering to develop techniques that go into formulating such strategies. Say, you are hired as an analyst to work on this project along with the quant. Your task on this project is to assess whether trading on some days of the week is better or worse than any other. The following table summarizes whether the stock market went up or down during each trading day of last year. Run the appropriate statistical test to address the objective. Verify the required conditions necessary for calling your statistical method a valid one. Do the data impose any limitations on the conclusions drawn from the study?

		Day of Week				
		Monday	Tuesday	Wednesday	Thursday	Friday
Market	Down	42	49	46	43	41
Direction	Up	53	55	58	59	58

If the market direction and day of week are independent, then there is not a specific day that trading would be better or worse.

H_0:Market direction and day of week are independent

H_1:Market direction and day of week are dependent

In [2]:
```
down<-c(42,49,46,43,41)
up<-c(53,55,58,59,58)
df<-cbind(down,up)
rownames(df)<-c('mon','tue','wed','thu','fri')
df
#########Requirement Test#######################
#1. randomly select, independent - yes
#2. categorical daya - yes


#########Test###################################
chisq.test(df)
chisq.test(df)$expected

############Conclusion#################################
#At 5% significant level, market direction and day of week are independent. So, there is no such 
#better or worse than any other days.
#Limitation: The days of down and up cannot show how much the stock price change.
```

executed in 165ms, finished 12:18:35 2019-11-17

	down	up
mon	42	53
tue	49	55
wed	46	58
thu	43	59
fri	41	58

```
        Pearson's Chi-squared test

data:  df
X-squared = 0.81886, df = 4, p-value = 0.9359
```

	down	up
mon	41.65675	53.34325
tue	45.60317	58.39683
wed	45.60317	58.39683
thu	44.72619	57.27381
fri	43.41071	55.58929

10 (5 points) Question 10

The Survey of Consumer Finances (SCF) is conducted every 3 years to provide detailed information on the finances of U.S. households. The study is sponsored by the Federal Reserve Board in cooperation with the Department of the Treasury. Since 1992, data have been collected by the National Opinion Research Center (NORC) at the University of Chicago. The full dataset is very large, and the number of variables is extremely large as well. For this exercise, I have created a subset of the dataset. This subset contains data for the middle-class households (40% - 60% of the net worth).

We believe that people who completed a college degree fare better financially than those who started college and never finished. One way to judge financial success is by measuring assets. Is there enough evidence to conclude that heads of households with college degrees have more assets than those who have some college? Run an appropriate statistical technique to infer.

The dataset in the CSV file Question10 contains data for the education level (EDCL) and the total value of assets (ASSET) for middle class households.

The variable definitions are:

ASSET: Total value of assets held by household

EDCL Education category of head of household: 1. No high school diploma, 2. High school diploma, 3. Some college, 4. College degree

In [40]:
```
setwd("D:/BAX441/Homeworks/Homeworks 4 and 5 Combined")
Q10<-read.csv('Question10.csv')
Q10
```

executed in 55ms, finished 15:37:29 2019-11-17

. . .

In [41]:

```
edu<-as.factor(Q10$EDCL)
asset<-as.numeric(Q10$ASSET)

#########Requirement Test#########################
#1. Randomly selected sample - Satisfy
#2. Normality - Unsatisfy, because the p-value of the normality test is less than 0.05
model_ex10 <- lm( asset~  edu)
resids_ex10 <- residuals(model_ex10)
preds_ex10 <- predict(model_ex10)

nortest::ad.test(resids_ex10)

#hist(resids_ex41)

#qqnorm(resids_ex41)
#qqline(resids_ex41)

#3. Variances are constant - Satisfy, because the p-value of the normality test is larger than 0.(
fligner.test( asset~  edu)
car::leveneTest(asset~  edu)
########Kruskal-Wallis Test############################
#H0: There is not difference between different education level and the asset of people
#H1: At least one education level people have different asset level from others
kruskal.test(asset ~  edu)

pairwise.wilcox.test(asset,edu,p.adjust='bonferroni')
########Conclusion###############################
print('Yes, at 5% significant level, there is enough evidence to conclude that heads of household
with college degrees have more assets than those who have some college')
```

executed in 89ms, finished 15:37:33 2019-11-17

```
        Anderson-Darling normality test

data:  resids_ex10
A = 8.0446, p-value < 2.2e-16


        Fligner-Killeen test of homogeneity of variances

data:  asset by edu
Fligner-Killeen:med chi-squared = 5.1307, df = 3, p-value = 0.1625
```

	Df	F value	Pr(>F)
group	3	0.9231836	0.4288763
	1199	NA	NA

```
        Kruskal-Wallis rank sum test

data:  asset by edu
Kruskal-Wallis chi-squared = 43.915, df = 3, p-value = 1.573e-09


        Pairwise comparisons using Wilcoxon rank sum test

data:  asset and edu

  1       2       3
2 1.0000  -       -
3 0.1411  0.6583  -
4 1.1e-05 5.2e-08 0.0089

P value adjustment method: bonferroni

[1] "Yes, at 5% significant level, there is enough evidence to conclude that heads of househo
```

```
lds \nwith college degrees have more assets than those who have some college"
```

11 (5 points) Question 11

A hospital conducted a study of the waiting time in its emergency room. The hospital has a main campus, along with three satellite locations. Management had a business objective of reducing waiting time for emergency room cases that did not require immediate attention. To study this, a random sample of 15 emergency room cases at each location were selected on a particular day, and the waiting time (recorded from check-in to when the patient was called into the clinic area) was measured. The results are stored in the CSV file Question11. Conduct an appropriate statistical test to determine if there is any evidence of a difference in the waiting times.

In [38]:
```
setwd("D:/BAX441/Homeworks/Homeworks 4 and 5 Combined")
Q11<-read.csv('Question11.csv')
Q11<-stack(Q11)
colnames(Q11)<-c('wait_time','location')
wait_time<-Q11$wait_time
location<-as.factor(Q11$location)
```

executed in 31ms, finished 15:36:52 2019-11-17

In [39]:
```
#########Requirement Test#######################
#1. Randomly selected sample - Satisfy
#2. Normality - Satisfy, because the p-value of the normality test is less than 0.05
model_ex11 <- lm( wait_time~  location)
resids_ex11 <- residuals(model_ex11)
preds_ex11 <- predict(model_ex11)

nortest::ad.test(resids_ex11)

#hist(resids_ex11)

#qqnorm(resids_ex11)
#qqline(resids_ex11)

#3. Variances are constant - Satisfy, because the p-value of the normality test is larger than 0.(
fligner.test(wait_time~  location)
car::leveneTest(wait_time~  location)

########One-way ANOVA############################
#H0: There is not difference in waiting time among the locations
#H1: At least one location's waiting time is different from others

ANOVA_Ex11 <- aov(wait_time~  location)
summary(ANOVA_Ex11)

TukeyHSD(ANOVA_Ex11)

plot(TukeyHSD(ANOVA_Ex11), las = 1)

########Conclusion################################
print('Yes, at 5% significant level, there are differences in waiting time among different locati

print('Satellite 1 and 3 have significantly shorter waiting time than the main campus.')
```

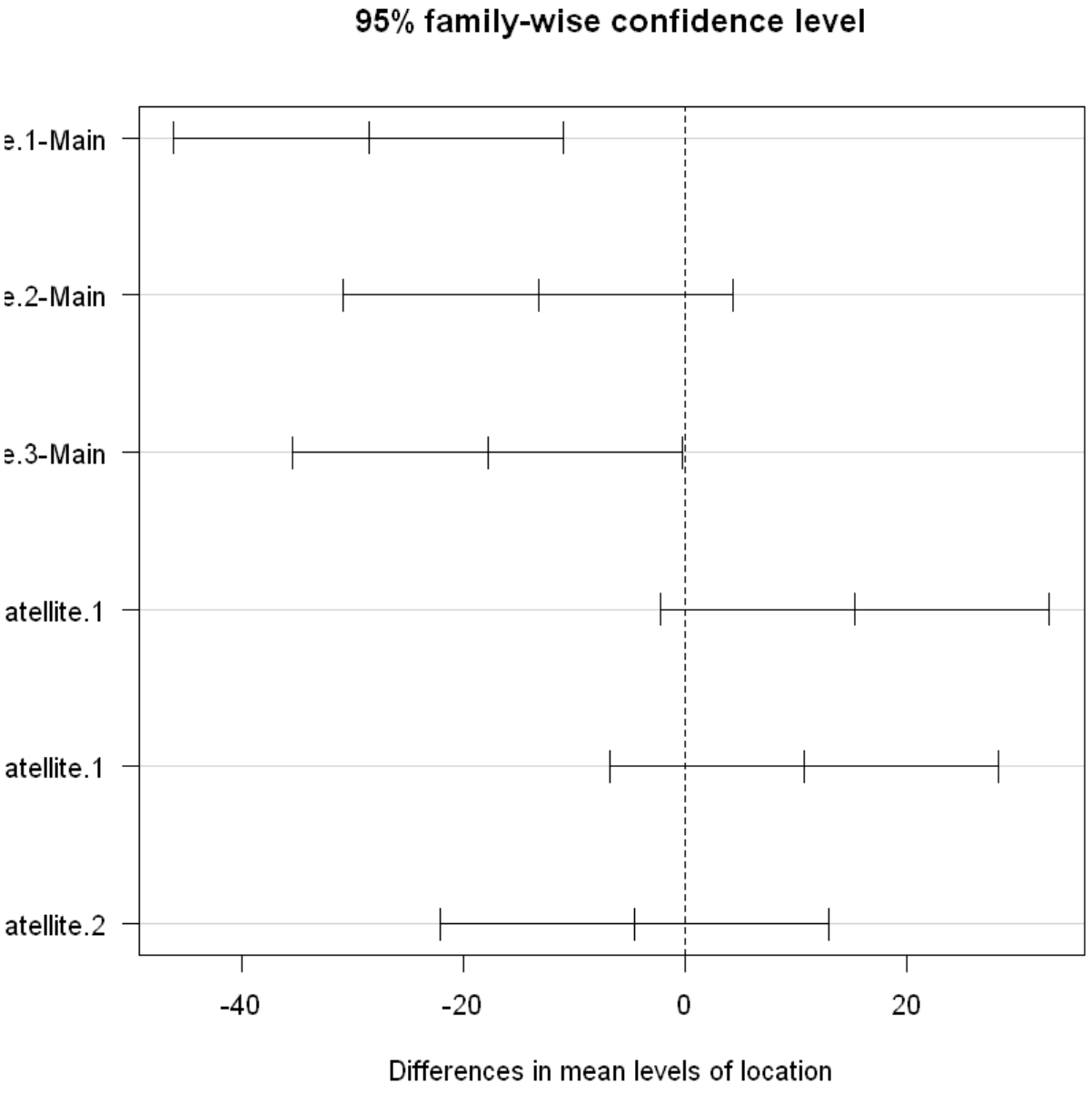
executed in 111ms, finished 15:36:53 2019-11-17

```
        Anderson-Darling normality test

data:  resids_ex11
A = 0.29114, p-value = 0.5973


        Fligner-Killeen test of homogeneity of variances

data:  wait_time by location
Fligner-Killeen:med chi-squared = 3.1187, df = 3, p-value = 0.3737
```

	Df	F value	Pr(>F)
group	3	0.8200883	0.4882562
	56	NA	NA

```
            Df Sum Sq Mean Sq F value   Pr(>F)
location     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

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = wait_time ~ location)

$location
                             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

[1] "Yes, at 5% significant level, there are differences in waiting time among different locatio
ns."
[1] "Satellite 1 and 3 have significantly shorter waiting time than the main campus."
```



12 (5 points) Question 12

A well-known soft-drink manufacturer has used the same secret recipe for its product since its introduction over 100 years ago. In response to a decreasing market share, however, the president of the company is contemplating changing the recipe. He has developed two alternative recipes. In a preliminary study, he asked 20 people to taste the original recipe and the two new recipes. He asked each to evaluate the taste of the product on a 5-point scale, where 1 = Awful, 2=Poor,3=Fair,4=Good, and 5=Wonderful. The dataset is available on the CSV file Question12. The president decides that unless significant differences exist between evaluations of the products, he will not make any changes. At 5% significance level, use an appropriate statistical test to conclude if there are any differences in the ratings of the three recipes.

In [1]:
```
setwd("D:/BAX441/Homeworks/Homeworks 4 and 5 Combined")
Q12<-read.csv('Question12.csv')
df<-Q12[,-1]
df<-stack(df)
colnames(df)<-c('rating','type')
rating<-df$rating
type<-as.factor(df$type)
df2<-read.csv('Question12.csv',header=TRUE,row.names="Person")
df2<-data.matrix(df2)
block<-Q12$Person
block<-rep(block,times=3)
df_block<-cbind(df,block)
Q12
```

executed in 61ms, finished 23:16:34 2019-11-18

. . .

In [9]:
```
#########Requirement Test#######################
#1. Randomly selected sample - Satisfy
#2. Normality - Unsatisfy, because the p-value of the normality test is less than 0.05


rating<-df_block$rating
type<-as.factor(df_block$type)
block<-as.factor(df_block$block)
########Friedman Test (Random Block Design)##############
#H0: There is not differences in the ratings of the three recipes
#H1: At least one recipe is different from others

friedman.test(rating,type,block)
posthoc.friedman.nemenyi.test(rating,type,block)

########Conclusion###############################
print('Yes, at 5% significant level, there is difference between original and new recipe 2.')
```

executed in 48ms, finished 23:18:20 2019-11-18

```
        Friedman rank sum test

data:  rating, type and block
Friedman chi-squared = 7.1525, df = 2, p-value = 0.02798


        Pairwise comparisons using Nemenyi multiple comparison test
             with q approximation for unreplicated blocked data

data:  rating , type and block

             Original New.Recipe.1
New.Recipe.1  0.559    -
New.Recipe..2 0.057    0.415

P value adjustment method: none

[1] "Yes, at 5% significant level, there is difference between original and new recipe 2."
```

13 (5 points) Question 13

Kira Breyan works as a financial advisor at a large investment firm. She meets with an inexperienced investor who has some questions regarding two approaches to mutual fund investing: growth investing versus value investing. The investor has heard that growth funds invest in companies whose stock prices are expected to grow at a faster rate, relative to the overall stock market, and value funds invest in companies whose stock prices are below their true worth. The investor has also heard that the main component of investment return is through capital appreciation in growth funds and through dividend income in value funds. The investor shows Kira the annual return data for Vanguard's Growth Index mutual fund (Growth) and Vanguard's Value Index mutual fund (Value). The CSV file Question13 shows the annual return data for these two mutual funds for the years 2007–2016. Using the appropriate statistical test, determine if the performance of the returns of these two funds were similar over the 2007-2016 timeframe.

In [2]:
```
library(reshape2)
library(PMCMR)
```

executed in 191ms, finished 18:55:09 2019-11-19

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

In [1]:
```
setwd("D:/BAX441/Homeworks/Homeworks 4 and 5 Combined")
Q13<-read.csv('Question13.csv',header=TRUE, row.names="Year")
df<-data.matrix(Q13)

Q13
```

executed in 51ms, finished 18:55:05 2019-11-19

. . .

In [4]:
```
#########Requirement Test#######################
#1. Randomly selected sample - Satisfy
#2. Normality - Unsatisfy


growth<-Q13$Growth
value<-Q13$Value


########Spearman Rank Correlation#######################
#H0: There is no correlation between growth and value
#H1: There is correlation between growth and value

cor.test(x=growth, y=value,method='spearman')

#######Conclusion######################
print('Yes, at 5% significant level, the performance of the returns of these two funds were simil
```

executed in 46ms, finished 18:57:19 2019-11-19

```
        Spearman's rank correlation rho

data:  growth and value
S = 32, p-value = 0.008236
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8060606

[1] "Yes, at 5% significant level, the performance of the returns of these two funds were simila
r over the given time frame."
```

14 (5 points) Question 14

Refer to the dataset from the General Social Survey provided on the CSV file Question14. If one works longer hours (HRS1) does the chances of losing one's job (JOBLOSE: 1 = Very likely, 2 = Fairly likely, 3 = Not too likely, 4 = Not likely) become less likely? Conduct an appropriate statistical test to answer the question.

In [16]:
```
library(tidyverse)
```

executed in 1.89s, finished 15:04:49 2019-11-17

. . .

In [32]:

```
setwd("D:/BAX441/Homeworks/Homeworks 4 and 5 Combined")
Q14<-read.csv('Question14.csv')
df<-Q14 %>% drop_na(c("HRS1",'JOBLOSE'))
df<-df[,2:5]
df_1ANOVA<-df[,3:4]
joblose<-as.factor(df_1ANOVA$JOBLOSE)
hours<-df_1ANOVA$HRS1
df
```

executed in 66ms, finished 15:32:22 2019-11-17

. . .

In [33]:
```
#########Requirement Test#######################
#1. Randomly selected sample - Satisfy
#2. Normality - Unsatisfy, because the p-value of the normality test is less than 0.05
model_ex14 <- lm( hours~  joblose)
resids_ex14 <- residuals(model_ex14)
preds_ex14 <- predict(model_ex14)

nortest::ad.test(resids_ex14)


#3. Variances are constant - Unsatisfy, because the p-value of the normality test is larger than (
fligner.test(hours~  joblose)
car::leveneTest(hours~  joblose)

########Spearman Rank Correlation#####################
#H0: There is no correlation between work hours and joblose
#H1: There is correlation between work hours and joblose

cor.test(x=df$HRS1, y=df$JOBLOSE,method='spearman')

########Conclusion#########################
print('Yes, at 5% significant level,there is correlation between work hours and job lose: more wo
But the correlation is weak.')
```

executed in 57ms, finished 15:32:27 2019-11-17

```
        Anderson-Darling normality test

data:  resids_ex14
A = 19.506, p-value < 2.2e-16


        Fligner-Killeen test of homogeneity of variances

data:  hours by joblose
Fligner-Killeen:med chi-squared = 13.297, df = 3, p-value = 0.004036
```

	Df	F value	Pr(>F)
group	3	4.68726	0.002945543
	997	NA	NA

```
Warning message in cor.test.default(x = df$HRS1, y = df$JOBLOSE, method = "spearman"):
"Cannot compute exact p-value with ties"


        Spearman's rank correlation rho

data:  df$HRS1 and df$JOBLOSE
S = 151103791, p-value = 0.002339
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.09609079

[1] "Yes, at 5% significant level,there is correlation between work hours and job lose: more wor
k hours, less chance to lose job. \nBut the correlation is weak."
```

15 (5 points) Question 15

Does the brand name of an ice cream affect consumers' perceptions of it? The marketing manager of a major dairy pondered this question. She decided to ask 60 randomly selected people to taste the same flavor of ice cream in two different dishes. The dishes contained exactly the same ice cream but were labeled differently. One was given a name that suggested that its maker was European and sophisticated; the other was given a name that implied that the product

was domestic and inexpensive. The tasters were asked to rate each ice cream on a 5-point scale, where 1 = Poor, 2 = Fair, 3 = Good, 4 = Very good, and 5 = Excellent. Do the results allow the manager to conclude at the 10% significance level that the European brand is preferred? Dataset is on the CSV file Question15.

In [5]:
```
setwd("D:/BAX441/Homeworks/Homeworks 4 and 5 Combined")
Q15<-read.csv('Question15.csv')
Q15
```

executed in 58ms, finished 19:32:56 2019-11-19

...

In [8]:
```
#########Requirement Test#######################
#1. Randomly selected sample - Satisfy
#2. Normality - Unsatisfy because is ordinal data
#hist(Q15$European)
#hist(Q15$Domestic)

#########Sign Test#######################
#H0: The European brand is less preferred than or equal to the local brand
#H1: The European brand is more preferred than the local brand

install.packages('BSDA')
library(BSDA)

SIGN.test(x = Q15$European,
          y = Q15$Domestic,
          alternative = "greater",
          conf.level = 0.9)

########Conclusion###############################
print('Yes, at the 10% significant level, the survey result show that European brand is preffered
```

executed in 1.23s, finished 19:35:05 2019-11-19

```
Warning message:
"package 'BSDA' is in use and will not be installed"


        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

[1] "Yes, at the 10% significant level, the survey result show that European brand is preffere
d."
```

16 (5 points) Question 16

The New York Times has been publishing its weekly list of best-selling books in the United States for more than 80 years. For certain book subcategories, the best-selling lists are published monthly. On the New York Times best-selling book website, there is a “buy” button below (or next to) each listed book. The button directs the viewer to online bookstore options. Two such options are: Amazon and Barnes and Noble. The CSV dataset Question16 contains online

prices of the best-selling business books for the month of September a couple of years ago. Older titles are paperback versus paperback comparisons, while newer titles are hardback versus hardback comparisons. Is there a systematic difference in book prices between the two online booksellers? Use an appropriate statistical test at the 5% significance level to infer.

In [34]:
```
setwd("D:/BAX441/Homeworks/Homeworks 4 and 5 Combined")
Q16<-read.csv('Question16.csv')
Q16
```

executed in 32ms, finished 15:35:00 2019-11-17

Title	BN	Amazon
Outliers	10.39	10.19
The Path Between the Seas	13.44	13.18
Thinking, Fast and Slow	9.99	9.90
The Power of Habit	9.97	9.78
#GIRLBOSS	16.25	16.17
The Organized Mind	18.03	17.68
Capital in the Twenty-First Century	24.08	23.97
Think Like a Freak	17.73	17.73
Business Adventures	10.25	10.25
Lean In	16.11	16.03

In [35]:
```
##########Requirement Test#########################
#1. Randomly selected sample - Satisfy
#2. Normality - Unsatisfy

##########Wilcoxon Signed Rank Sum Test (Paired)################
#H0: There is not difference in book price between bn and amazon
#H1: There is difference in book price between bn and amazon
bn<-Q16$BN
amazon<-Q16$Amazon
wilcox.test(bn,amazon,alt='two.sided',paired=TRUE,exacted=FALSE,conf.level=0.95)

##########Conclusion############################
print('Yes, at 5% significant level, there is difference in price between these two bookstore.')
```

executed in 39ms, finished 15:35:04 2019-11-17

```
Warning message in wilcox.test.default(bn, amazon, alt = "two.sided", paired = TRUE, :
"cannot compute exact p-value with ties"Warning message in wilcox.test.default(bn, amazon, alt =
"two.sided", paired = TRUE, :
"cannot compute exact p-value with zeroes"


        Wilcoxon signed rank test with continuity correction

data:  bn and amazon
V = 36, p-value = 0.01415
alternative hypothesis: true location shift is not equal to 0

[1] "Yes, at 5% significant level, there is difference in price between these two bookstore."
```

17 (5 points) Question 17

Suppose that a random sample of 100 observations was drawn from a population. After calculating the mean and standard deviation, each observation was standardized and the number of observations in each of the following intervals was counted. Using the chi-squared test for normality, can we infer at the 5% significance level that the data

were not drawn from a normal population? Note: We have covered using chi-squared test to assess if the observed counts match the Poisson distribution. This exercise is to assess if the data comes from (fits) the normal distribution. This is another way to assess normality.

Interval	Frequency
Z≤1.5	10
-1.5<Z≤-0.5	18
-0.5<Z≤0.5	48
0.5<Z≤1.5	16
Z>1.5	8

In [34]:
```
#########Requirement Test#########################
#1.Simple Random Sample:satisfy
#2.Contingency table condition: The categories are mutually exclusive and the question made sure
#3.Sample size condition: all frequency numbers are >5

########Testing the Fit of a Probability Model############
#H0: the data was drawn from a normal population
#H1: the data was not drawn from a normal population

p1<-pnorm(-1.5,mean=0,sd=1)
p2<-pnorm(-0.5,mean=0,sd=1)-p1
p3<-pnorm(0.5,mean=0,sd=1)-p1-p2
p4<-pnorm(1.5,mean=0,sd=1)-p1-p2-p3
p5<-1-pnorm(1.5,mean=0,sd=1)

ex_p<-c(p1,p2,p3,p4,p5)
ex_f<-ex_p*100
ex_p

obs<-c(10,18,48,16,8)
chisq.test(x=obs,p=ex_p)
```

executed in 93ms, finished 16:08:39 2019-11-16

```
0.0668072012688581  0.241730337457129  0.382924922548026  0.241730337457129  0.0668072012688581


        Chi-squared test for given probabilities

data:  obs
X-squared = 8.7104, df = 4, p-value = 0.06876
```

In [122]:
```
pchisq(8.7104,3,lower.tail=FALSE)
```

executed in 27ms, finished 19:14:16 2019-11-14

```
0.0333996807612279
```

Yes, we can infer at the 5% significance level that the data were not drawn from a normal population