HO CHI MINH UNIVERSITY OF TECHNOLOGY
FALCUTY OF COMPUTER SCIENCE AND ENGINERRING

Probability and Statistics

Assignment

Project 1 – Topic 5

Contents

1. Exercise 1

A diary farm raised three breeds of dairy cows A, B, C. The amount of milk of the cows is given in the following table:

Type of cows	The amount of milk		
	Little	Medium	Much
A	92	37	46
B	53	15	19
C	75	19	12

At the significance level of $\alpha = 0.05$, can we conclude that the amount of milk produced by these type of cows is different?

1.1. Solving method

In this exercise, since there are two independent factors and we want to compare the effect of them on a dependent variable (sales), we decided to use ***Two – way ANOVA*** to test this data. Also, we're using ***Tukey's HSD*** test to find any significant different between type of cows

1.2. Theory

1.2.1. Two – way ANOVA

Two – way ANOVA is an extension of the one – way (will be explained in Exercise 4). ANOVA that examines the influence of two different categorical independent variables on one continous dependent variable. The two – way ANOVA not only aims at assessing the main effect of each independent variable but also if there is any interaction between them.

A two – way ANOVA with interaction tests three null hypothesis (H_0) at the same time:

- There is no difference in group means at any level of the first independent variable.
- There is no difference in group means at any level of the second independent variable.
- The effect of one independent variable does not depend on the effect of the other independent variable (no interaction effect)

Assumptions:

To use a two – way ANOVA, the data should meet these certain assumptions:

- Homogentity of variance: homoscedasticity and variances of the populations must be equal
- Independence of observations: The independent variables should not be dependent on one another (i.e. one should not cause the other).
- Normally-distributed dependent variable: The values of the dependent variable should follow a bell curve.
- There should be no significant outliers

1.2.2. Tukey's HSD

Tukey's range test, also known as Tukey's test, Tukey method, Tukey's honest significance test, or Tukey's HSD (honestly significant difference) test is a single-step multiple comparison procedure and statistical test. It can be used to find means that are significantly different from each other.

Tukey's test compares the means of every treatment to the means of every other treatment; that is, it applies simultaneously to the set of all pairwise comparisons $\mu_i - \mu_j$ and identifies any difference between two means that is greater than the expected standard error. The confidence coefficient for the set, when all sample sizes are equal, is exactly $1 - \alpha$ for any $0 \leq \alpha \leq 1$. For unequal sample sizes, the confidence coefficient is greater than $1 - \alpha$. In other words, the Tukey method is conservative when there are unequal sample sizes.

Assumptions

- The observations being tested are independent within and among the groups.
- The groups associated with each mean in the test are normally distributed.
- There is equal within-group variance across the groups associated with each mean in the test (homogeneity of variance).

1.3. Steps to perform

Step 1: Check assumptions and write hypothesis

H_{0a}: There is no difference in group means at any level of the first independent variable.

H_{1a}: There is a difference in group means at any level of the first independent variable

H_{0b}: There is no difference in group means at any level of the second independent variable.

H_{1b}: There is a difference in group means at any level of the second independent variable.

H_{0ab}:The effect of one independent variable does not depend on the effect of the other independent variable (no interaction effect)

H_{1ab}:The effect of one independent variable does not depend on the effect of the other independent variable (no interaction effect)

Step 2: Calculate an approriate test stastistic

We calculate $SSA, SSB, SSAB, SSE, MSA, MSB, MSAB, MSE, F_A, F_B, F_{AB}$ with the formula in the table below

Source	Degree of freedom	Sum of Squares(SS)	Mean Squares(MS)	F-Ratio
A (between groups)	$a - 1$	$SSA = \sum_{i=1}^{a} n_j(\overline{y}_i - \overline{y}..)^2$	$MSA = \frac{SSA}{a-1}$	$F_A = \frac{MSA}{MSE}$

B(between groups)	$b-1$	$SSB = \sum_{i=1}^{b} n_j(\overline{y}_i - \overline{y}..)^2$	$MSB = \frac{SSB}{b-1}$	$F_B = \frac{MSB}{MSE}$
Error(within groups)	$(a-1)(b-1)$	$SSE = SST - SSA - SSB$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	
Total	$N-1$	$SST = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(\overline{y}_{ij} - \overline{y}..\right)^2$		

Notation:

a: number of levels of factor A(first independent factor)

b: number of levels of factor B(second independent factor)

ab: number of treatments(each one a combination of factor A level and a factor B level)

r: number of observarions on each treatment

Step 3: Determine the p-value

The p-value associated with each F-statistic are presented in an ANOVA table.

Step 4: p-value decisions

Based on each p-value for each F-statistic we have obtained:

If $p - value_i < \alpha$: reject H_{0i}

If $p - value_i \geq \alpha$: fail to reject H_{0i}

(where $i \in \{a, b, ab\}$)

Step 5: Post-hoc test

After we point out the differences between levels of the two factors, in order to compare those toghether, we perform a Tukey's Honestly Significant Difference (Tukey's HSD)- post-hoc test for pairwise comparisons.

How to do:

- Choose the calculated data from ANOVA output, especially these field:
  - Means
  - Mean Square Within
  - Number per treatment
  - Degree of freedom Within

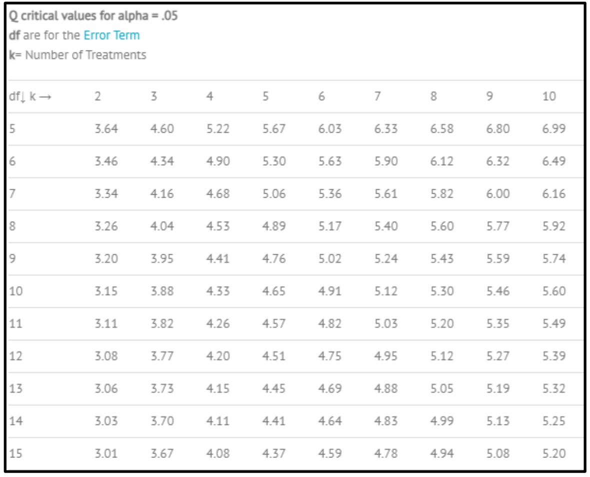
- Calculate the HSD statistic with the following formula

$$HSD = \frac{M_i - M_j}{\sqrt{\frac{MS_w}{n_h}}}$$

Where:

  - $M_i - M_j$ is the difference between the pair of means
  - MS_w is the Mean Square Within
  - n is the number in the group or treatment

- Find the corresponding key in Tukey's critical value table.



Q critical values for alpha = .05

df are for the Error Term

k= Number of Treatments

df↓ k →	2	3	4	5	6	7	8	9	10
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92
9	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20

- Compare the statistic with the key, if the statistic is larger, there are a significant difference between the two means.

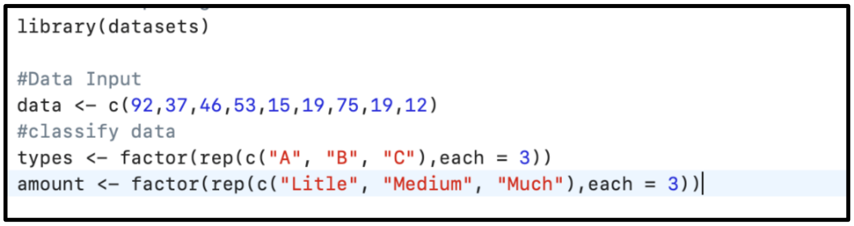
1.4. R/R – Studio Implementation

Step 1: Write hypothesis

H_0: The amount of milk produced by these types of cows is not different

H_1: The amount of milk produced by these types of cows is different

Step 2: Input data

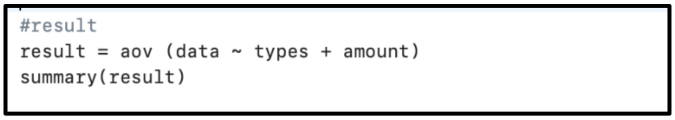


```
library(datasets)

#Data Input
data <- c(92,37,46,53,15,19,75,19,12)
#classify data
types <- factor(rep(c("A", "B", "C"),each = 3))
amount <- factor(rep(c("Litle", "Medium", "Much"),each = 3))
```

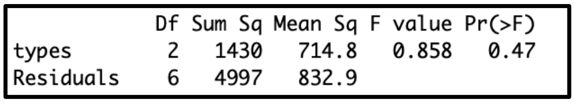
Step 3: Calculate an approriate test stastistic and find the p - value

The R function **aov**() can be used to get the result. The function **summary.aov**() is used to summarize the analysis of variance model.



```
#result
result = aov (data ~ types + amount)
summary(result)
```

After running the code, we got the result:



```
            Df Sum Sq Mean Sq F value Pr(>F)
types        2   1430   714.8   0.858   0.47
Residuals    6   4997   832.9
```

<u>*Where as:*</u>

- ***Df***: the degree of freedom
- ***Sum sq***: sum square
- ***Mean Sq***: mean square
- ***F Value***: test static from the F test.
- ***Pr(>F)**:* the p-value of the F-statistic

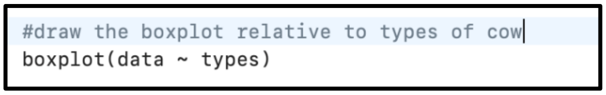
Step 4: Compare the p – value

Because $p - value > \alpha$ $(0.47 > 0.05)$, we will not reject H_0

Step 5: Post – hoc

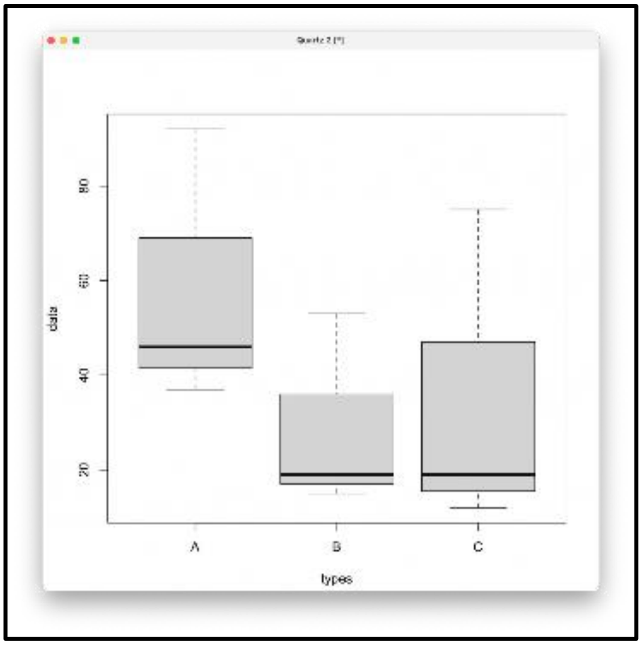
To confirm the interaction, we use boxplot tool to do the task.

Below is the code:



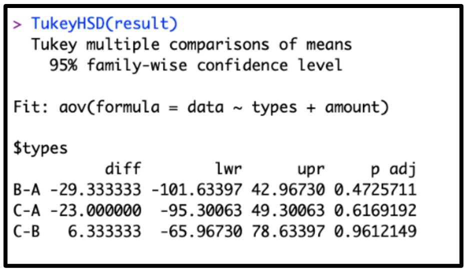
```
#draw the boxplot relative to types of cow
boxplot(data ~ types)
```

We got the following boxplot



Clearly, we see that the interaction of type A are significantly higher than other types.

After that, we perform post-hoc analysis by using TukeyHSD() in R to complete this analysis and here is the result:



```
> TukeyHSD(result)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = data ~ types + amount)

$types
          diff        lwr      upr     p adj
B-A -29.333333 -101.63397 42.96730 0.4725711
C-A -23.000000  -95.30063 49.30063 0.6169192
C-B   6.333333  -65.96730 78.63397 0.9612149
```

The p adj (adj means "adjusted" and is adjusted to avoid false positive rate) can be interpreted as p-value. To reject the null hypothesis, p adj must be less than α.

By observation, there are no p adj value less than 0.05, which tells none of the comparisons above are statistically significant.

=> H_0 cannot be rejected

Step 6: Conclusion

With α = 0.05, we do not have enough evidence to confirm that the amount of milk produced by these types of cows is different

Below is the full code implementation of the exercise:

```
#Exercise 1

#H0 : The amount of milk produced by these types of cows is not different

#H1 : The amount of milk produced by these types of cows is different

#Load base packages:

library(datasets)

#Data Input

data <- c(92,37,46,53,15,19,75,19,12)

#classify data

types <- factor(rep(c("A", "B", "C"),each = 3))

amount <- factor(rep(c("Litle", "Medium", "Much"),each = 3))

#result

result = aov (data ~ types + amount)

summary(result)

#draw the boxplot relative to types of cow

boxplot(data ~ types)

TukeyHSD(result)
```

2. Exercise 2

At the significance of α = 5%, compare the business performance of some industries in the four urban districts on the basis of the sales of some stores given in the following table.

Industries	Districts			
	1	2	3	4
Refrigeration	2.5, 2.7, 2.0, 3.0	13.1, 3.5, 2.7	2.0, 2.4	5.0, 5.4
Construction Materials	0.6, 10.4	15.0	9.5, 9.3, 9.1	19.5, 17.5
Computer services	1.2, 1.0, 9.8, 1.8	2.0, 2.2, 1.8	1.2, 1.3, 1.2	5.0, 4.8, 5.2

2.1. Solving method

In this exercise, since there are two independent factors (Districts and Industries) and we want to compare the effect of them on a dependent variable (sales), we decided to use ***Two – way ANOVA*** to test this data (assuming the data meet some certain conditions:the Homogeneity of variance – homoscedasticity and variances of the populations must be equal). Also, we're using ***Tukey's HSD*** test to compare the business performance of the industries in the districts.

2.2. Theory

The theory of the Two – way ANOVA and Tukey's HSD can be seen in section 1.2

2.3. Steps to perform

Steps to perform this exercise can be seen in section 1.3

2.4. R/R – Studio Implementation

Step 1: Check assumptions and write hypothesis

H_{0a}: There is no difference in business performance in different industries.

$H_{1a:}$ There is a difference in business performance in different industries

$H_{0b:}$ There is no difference in business performance in different districts.

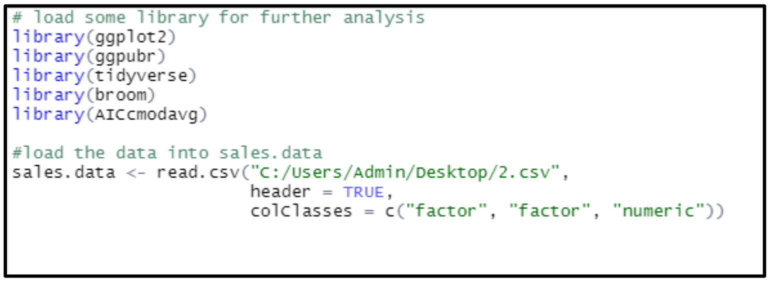
$H_{1b:}$ There is a difference in business performance in different districts.

$H_{0ab:}$ There business performance in different industries does not depend on the districts.

$H_{1ab:}$There business performance in different industries depend on the districts.

Step 2: Load data

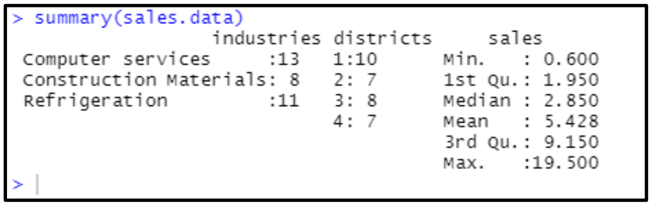
In order to use the data from the question, We transfer it into a .csv file and import the data to Rstudio with the help of read.csv() and specify whether each of the variables should be quantitative ("numeric") or categorical ("factor")



```
# load some library for further analysis
library(ggplot2)
library(ggpubr)
library(tidyverse)
library(broom)
library(AICcmodavg)

#load the data into sales.data
sales.data <- read.csv("C:/Users/Admin/Desktop/2.csv",
                       header = TRUE,
                       colClasses = c("factor", "factor", "numeric"))
```

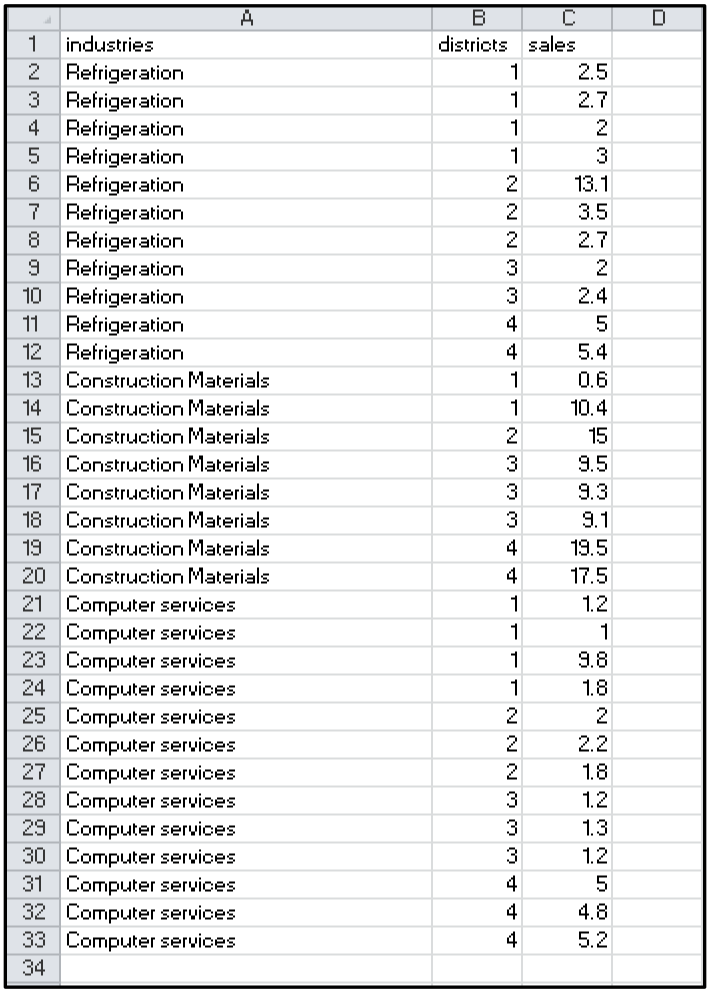
Check the data we have just loaded:



```
> summary(sales.data)
                 industries districts     sales
 Computer services     :13   1:10      Min.   : 0.600
 Construction Materials: 8   2: 7      1st Qu.: 1.950
 Refrigeration         :11   3: 8      Median : 2.850
                             4: 7      Mean   : 5.428
                                       3rd Qu.: 9.150
                                       Max.   :19.500
> 
```

The 'industry' and 'districts' are listed as 2 factor as well as their number of observation at each level (for example, there are 13 observations in Computer Service, 8 observations Construction Materials and 11 observations in Refrigeration industry).

At the right most column 'sales' is our dependent variable which is treated as numeric data with its summary (min, max, mean, median,...)

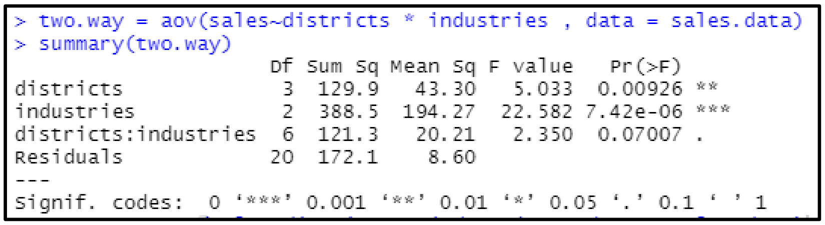


	A	B	C	D
1	industries	districts	sales	
2	Refrigeration	1	2.5	
3	Refrigeration	1	2.7	
4	Refrigeration	1	2	
5	Refrigeration	1	3	
6	Refrigeration	2	13.1	
7	Refrigeration	2	3.5	
8	Refrigeration	2	2.7	
9	Refrigeration	3	2	
10	Refrigeration	3	2.4	
11	Refrigeration	4	5	
12	Refrigeration	4	5.4	
13	Construction Materials	1	0.6	
14	Construction Materials	1	10.4	
15	Construction Materials	2	15	
16	Construction Materials	3	9.5	
17	Construction Materials	3	9.3	
18	Construction Materials	3	9.1	
19	Construction Materials	4	19.5	
20	Construction Materials	4	17.5	
21	Computer services	1	1.2	
22	Computer services	1	1	
23	Computer services	1	9.8	
24	Computer services	1	1.8	
25	Computer services	2	2	
26	Computer services	2	2.2	
27	Computer services	2	1.8	
28	Computer services	3	1.2	
29	Computer services	3	1.3	
30	Computer services	3	1.2	
31	Computer services	4	5	
32	Computer services	4	4.8	
33	Computer services	4	5.2	
34				

The .csv file contain the data.

Step 3: Perform ANOVA test and point the difference

We can perform an ANOVA in R using the **aov()** function. This will calculate the test statistic for ANOVA and determine whether there is significant variation among the groups formed by the levels of the independent variable.

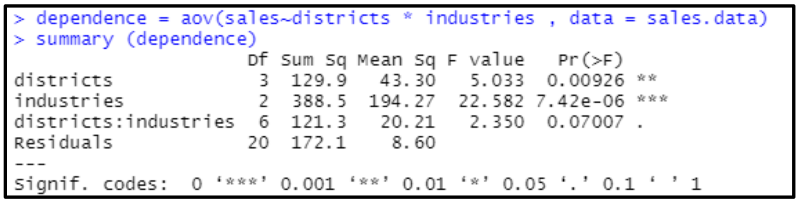


```
> two.way = aov(sales~districts * industries , data = sales.data)
> summary(two.way)
                     Df Sum Sq Mean Sq F value   Pr(>F)    
districts             3  129.9   43.30   5.033  0.00926 ** 
industries            2  388.5  194.27  22.582 7.42e-06 ***
districts:industries  6  121.3   20.21   2.350  0.07007 .  
Residuals            20  172.1    8.60                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- *Df*: the degree of freedom
- *Sum sq*: sum square
- *Mean Sq*: mean square
- *F Value*: test statistic from the F test.
- *Pr(>F)*: the p-value of the F-statistic

For the ‘districts’ and ‘industries’ factor, p-value of them are less than α = 0.05. So with a significant level of α = 0.05, we may have evidence that the type of industry and location of stores have an impact on the sales. In other words, there are differences based on the sales when we consider 2 types of industry as well as different districts

Test the dependence of 2 factors:

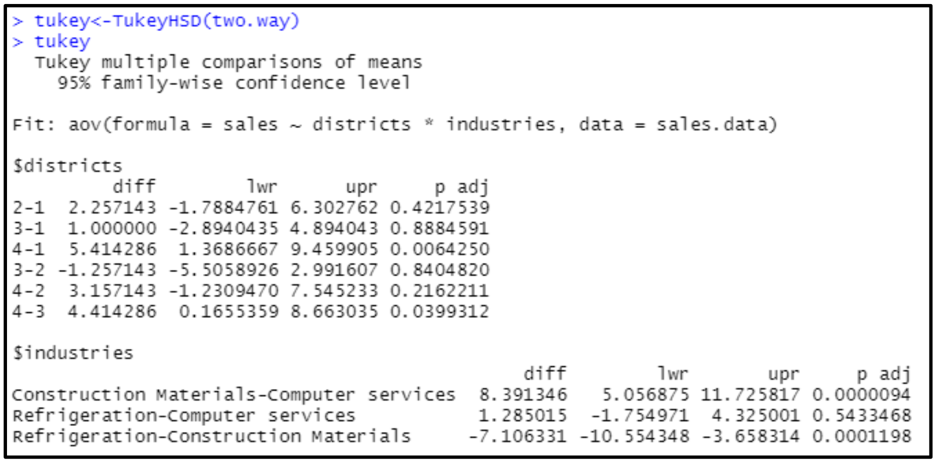


```
> dependence = aov(sales~districts * industries , data = sales.data)
> summary (dependence)
                     Df Sum Sq Mean Sq F value   Pr(>F)    
districts             3  129.9   43.30   5.033  0.00926 ** 
industries            2  388.5  194.27  22.582 7.42e-06 ***
districts:industries  6  121.3   20.21   2.350  0.07007 .  
Residuals            20  172.1    8.60                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As the p-value is larger than α = 0.05, the interaction between the 2 factors seem to be not significant.

Step 4: Post - hoc test

Using Tukey’s HSD test described above, we have the output below:



```
> tukey<-TukeyHSD(two.way)
> tukey
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = sales ~ districts * industries, data = sales.data)

$districts
         diff        lwr      upr     p adj
2-1  2.257143 -1.7884761 6.302762 0.4217539
3-1  1.000000 -2.8940435 4.894043 0.8884591
4-1  5.414286  1.3686667 9.459905 0.0064250
3-2 -1.257143 -5.5058926 2.991607 0.8404820
4-2  3.157143 -1.2309470 7.545233 0.2162211
4-3  4.414286  0.1655359 8.663035 0.0399312

$industries
                                           diff        lwr       upr     p adj
Construction Materials-Computer services  8.391346   5.056875 11.725817 0.0000094
Refrigeration-Computer services           1.285015  -1.754971  4.325001 0.5433468
Refrigeration-Construction Materials     -7.106331 -10.554348 -3.658314 0.0001198
```

Take a look at what the test give us. We can see that there are some significant differences between (as the p-value is small):

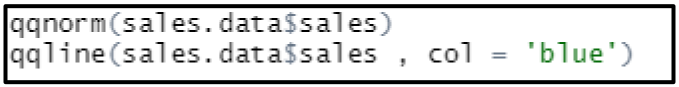
District 4 – District 1
District 4 – District 3
Construction Materials – Computer Service

2.5. Check the pre – assumed condition of the data

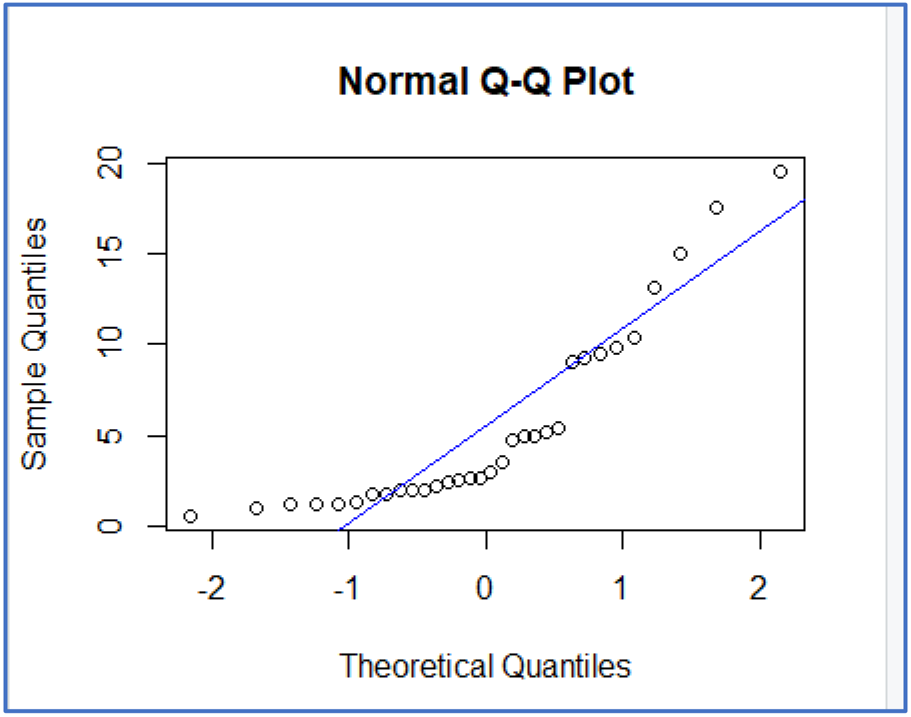
So far we have done analyzing the data and do the comparison with the help of ANOVA method.

This section will check again the assumption of using this method by a Q-Q plot.

We draw the qq plot in Rstudio:



```
qqnorm(sales.data$sales)
qqline(sales.data$sales , col = 'blue')
```



As clearly in the Q-Q Plot, the data does not fit closely to the line and there are lots of outliners. So maybe, two way ANOVA is not the best method for analyzing this data.

In addition, we can use more test to check if the population have common variation by Levene's test

The **null hypothesis** for Levene's test is that **the groups we're comparing all have equal population variances.**

And also, we can use a non parametric alternative that is Kruskal Wallis Test

Since these approaches are very interesting and too complicated, we did not go that far.

Below is the full code implementation of the exercise:

```
# load some library for further analysis
library(ggplot2)
library(ggpubr)
library(tidyverse)
library(broom)
library(AICcmodavg)

#load the data into sales.data
sales.data <- read.csv("C:/Users/Admin/Desktop/2.csv",
                       header = TRUE,
                       colClasses = c("factor", "factor", "numeric"))
summary(sales.data)
two.way = aov(sales~districts * industries , data = sales.data)
summary(two.way)
dependence = aov(sales~districts * industries , data = sales.data)
summary (dependence)
tukey<-TukeyHSD(two.way)
tukey
qqnorm(sales.data$sales)
qqline(sales.data$sales , col = 'blue')
# R program to illustrate
# Levene's test
# Import required package
library(car)
# Using leveneTest()
result = leveneTest(sales ~ interaction(districts, industries),
                        +                     data = sales.data)
# print the result
print(result)
```

3. Exercise 3

A group of US businessmen is categorized by their annual incomes and their ages. The results are as follows.

Age range	Income levels		
	Under $100000	$100000 - $400000	Over $400000
Under 40	6	9	5
40-54	18	19	8
Over 54	11	12	17

At the 1% significance level, are the age and the income level related or not?

3.1. Solving method

We're using ***Chi-Square*** test for data analysis

3.2. Theory

There are **two types of chi-square tests**. Both use the chi-square statistic and distribution for different purposes:

- A **chi-square goodness of fit test** determines if sample data matches a population.
- A **chi-square test for independence** compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of categorical variables differ from each other.

In this problem, we use ***Chi-Square test for independence***.

3.3. Steps to perform

Step 1: Check assumptions and write hypotheses

Hypothesis:

H_0: There is no relationship exists on the the categorical variables in the population (they are independent)

H_1: There is an relationship exists on the the categorical variables in the population (they are dependent)

Step 2: Calculate an appropriate test statistic

The formula for the chi-square statistic used in the chi-square test is:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where:

The subscript “c” is the degrees of freedom.
“O” is your observed value
“E” is your expected value

Step 3: Determine the $\chi^2_{critical}$

The $\chi^2_{critical}$ can be found in the Chi-Square distribution table with the correct degree of freedom. Chi-Square tests are always right-tailed tests.

$$\chi^2_{critical} \sim \chi^2(df)$$

Degree of Freedom of Chi-Square Test of Independence:

$$df = (number\ of\ rows - 1)(number\ of\ columns - 1)$$

Step 4: Make a decision:

If $\chi_c^2 \geq \chi^2_{critical}$: reject H_0

If $\chi_c^2 < \chi^2_{critical}$: fail to reject H_0

Step 5: State a “real world” conclusion

Write a conclusion in terms of the original research question

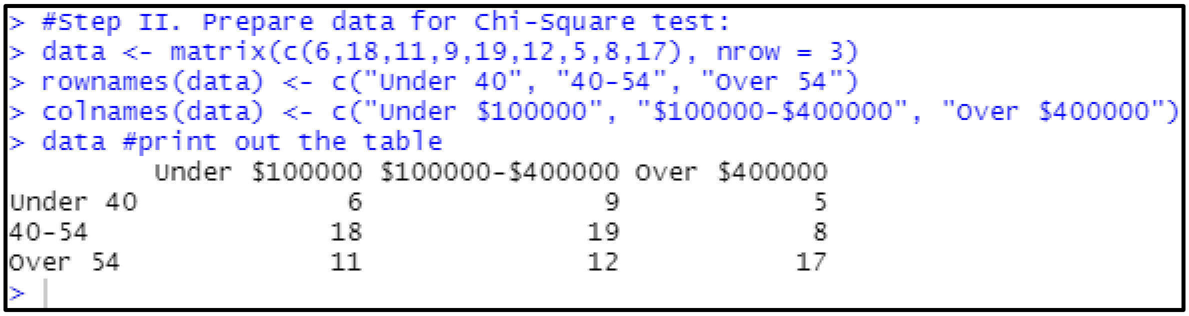
3.4. R/R – Studio Implementation

Step 1: Check assumptions and write hypotheses

H_0: The age and the income level are not related.

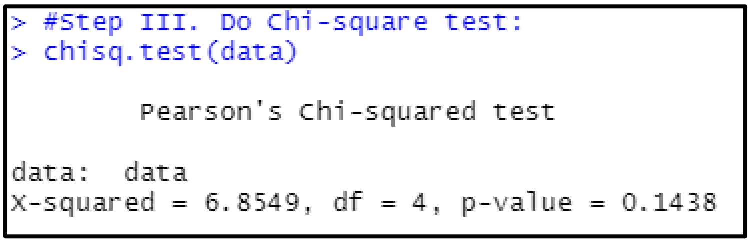
H_1: The age and the income level are related

Step 2: Prepare data by using matrix



```
> #Step II. Prepare data for Chi-Square test:
> data <- matrix(c(6,18,11,9,19,12,5,8,17), nrow = 3)
> rownames(data) <- c("Under 40", "40-54", "Over 54")
> colnames(data) <- c("Under $100000", "$100000-$400000", "Over $400000")
> data #print out the table
         Under $100000 $100000-$400000 Over $400000
Under 40             6               9            5
40-54               18              19            8
Over 54             11              12           17
> 
```

Step 3: Do Chi-Square test

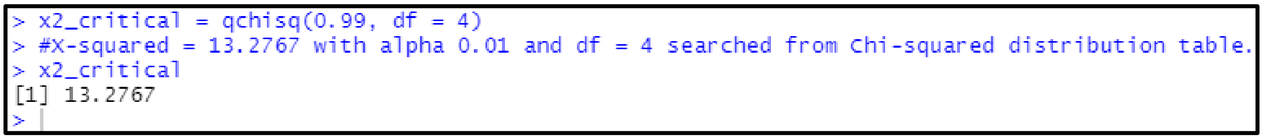


```
> #Step III. Do Chi-square test:
> chisq.test(data)

        Pearson's Chi-squared test

data:  data
X-squared = 6.8549, df = 4, p-value = 0.1438
```

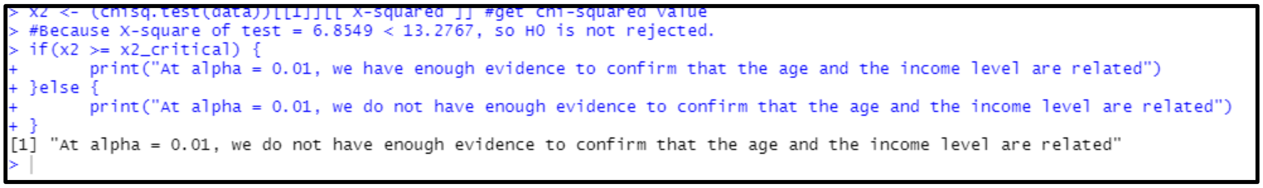
Step 4: Determine the $\chi^2_{critical}$



```
> x2_critical = qchisq(0.99, df = 4)
> #X-squared = 13.2767 with alpha 0.01 and df = 4 searched from Chi-squared distribution table.
> x2_critical
[1] 13.2767
> 
```

We can see that $\chi^2_c < \chi^2_{critical}$ $(6.8549 < 13.2767)$

Step 5: Conclusion



```
> x2 <- (chisq.test(data))[[1]][["X-squared"]] #get chi-squared value
> #Because X-square of test = 6.8549 < 13.2767, so H0 is not rejected.
> if(x2 >= x2_critical) {
+       print("At alpha = 0.01, we have enough evidence to confirm that the age and the income level are related")
+ }else {
+       print("At alpha = 0.01, we do not have enough evidence to confirm that the age and the income level are related")
+ }
[1] "At alpha = 0.01, we do not have enough evidence to confirm that the age and the income level are related"
> 
```

Conclusion: At the 1% significance level, we do not have enough evidence to confirm that the age and the income level are related.

Below is the full code implementation of the exercise:

```
#Exercise 3-Topic 5 (Chi-square)
#Step I: Write assumption and Hypothesis
#H0 : the age and the income level are not related
#H1 : the age and the income level are related
#Load base packages:
library(datasets)
#Step II. Prepare data for Chi-Square test:
data <- matrix(c(6,18,11,9,19,12,5,8,17), nrow = 3)
rownames(data) <- c("Under 40", "40-54", "Over 54")
colnames(data) <- c("Under $100000", "$100000-$400000", "Over $400000")
data #print out the table
#Step III. Do Chi-square test:
chisq.test(data)
#X-squared of test = 6.8549
#df = (num_row-1)*(num_column-1) = (3-1)*(3-1) = 4
#Step IV. Find X2_Critical
x2_critical = qchisq(0.99, df = 4)
#X-squared = 13.2767 with alpha 0.01 and df = 4 searched from Chi-squared distribution table.
x2_critical
#Step V. Conclusion:
x2 <- (chisq.test(data))[[1]][["X-squared"]] #get chi-squared value
#Because X-square of test = 6.8549 < 13.2767, so H0 is not rejected.
if(x2 >= x2_critical) {
    print("At alpha = 0.01, we have enough evidence to confirm that the age and the income level are related")
}else {
    print("At alpha = 0.01, we do not have enough evidence to confirm that the age and the income level are related")
}
```

4. Exercise 4

The following table shows the number of daily newspapers sold in 5 urban districts:

Day	Districts				
	District 1	District 2	District 3	District 4	District 5
Monday	22	18	22	18	18
Tuesday	21	18	22	18	19
Wednesday	25	25	25	19	20
Thursday	24	24	18	20	22
Friday	28	19	15	22	25
Saturday	30	22	28	25	25

Is there any significant difference in the amount of newspapers sold in the 5 districts at α = 1%? Is the amount of newspapers sold affected by the days of week?

4.1. Solving method

We're using ***Analysis Variance 1 factor (One-way ANOVA)*** for data analysis.

4.2. Theory

One-way ANOVA is a hypothesis test in which only one categorical variable or single factor is taken into consideration. With the help of F-distribution, it enables us to compare the means of three or more samples. The null hypothesis (H_0) is the equity in all population means while an alternative hypothesis is a difference in at least one mean.

X_1	X_2	...	X_k
X_{11}	X_{12}	...	X_{1k}
X_{21}	X_{22}	...	X_{2k}
...	...	...	...
$X_{n1\,1}$	$X_{n2\,2}$	....	$X_{nk\,k}$

Hypothesis:

H_0 : $a_1 = a_2 = \ldots = a_k$

H_1 : Exist $j_1 \neq j_2$ so that $a_{j1} \neq a_{j2}$

4.3. Steps to perform

Step 1: Check assumptions and write hypotheses

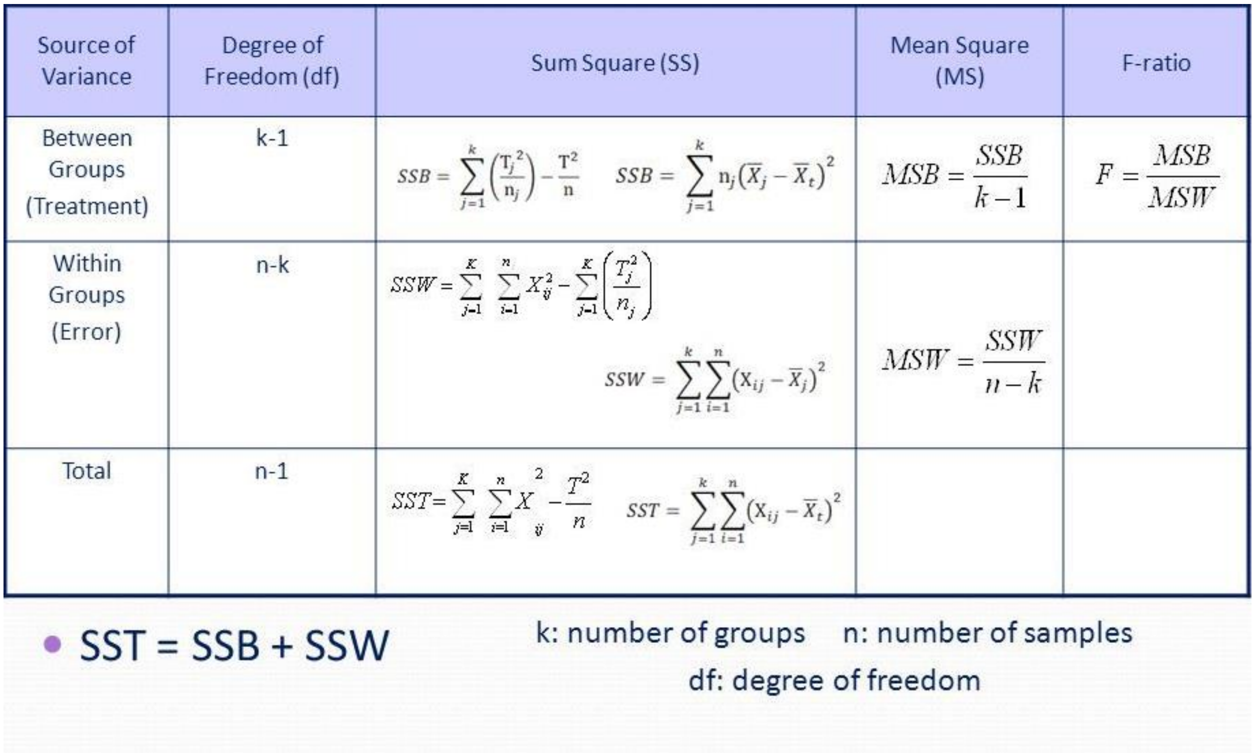
Hypothesis:

H_0 : There is no difference among group means

H_1 : There is at least one group differs significantly from the overall mean of the dependent variable

Step 2: Calculate an appropriate test statistic

We calculate SSB, SSW, SST, MSB, MSW and F-ratio based on the formula in the image below:



Source of Variance	Degree of Freedom (df)	Sum Square (SS)	Mean Square (MS)	F-ratio
Between Groups (Treatment)	k-1	$SSB = \sum_{j=1}^{k}\left(\frac{T_j^2}{n_j}\right) - \frac{T^2}{n}$ $\quad SSB = \sum_{j=1}^{k} n_j(\overline{X}_j - \overline{X}_t)^2$	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within Groups (Error)	n-k	$SSW = \sum_{j=1}^{K}\sum_{i=1}^{n} X_{ij}^2 - \sum_{j=1}^{K}\left(\frac{T_j^2}{n_j}\right)$ $SSW = \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2$	$MSW = \frac{SSW}{n-k}$	
Total	n-1	$SST = \sum_{j=1}^{K}\sum_{i=1}^{n} X_{ij}^{2} - \frac{T^2}{n}$ $\quad SST = \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \overline{X}_t)^2$		

- SST = SSB + SSW

k: number of groups n: number of samples

df: degree of freedom

Step 3: Determine the $F_{critical}$

$F_{critical}$ can be found in the Fisher distribution table with degree of freedom (k-1; n-k)

Step 4: Make a decision:

If $F \geq F_{critical}$: reject H_0

If $F < F_{critical}$: fail to reject H_0

Step 5: State a "real world" conclusion

Write a conclusion in terms of the original research question

4.4. R/R – Studio Implementation

4.4.1. Is there any significant difference in the amount of newspaper sold in the 5 districts at α=1%?

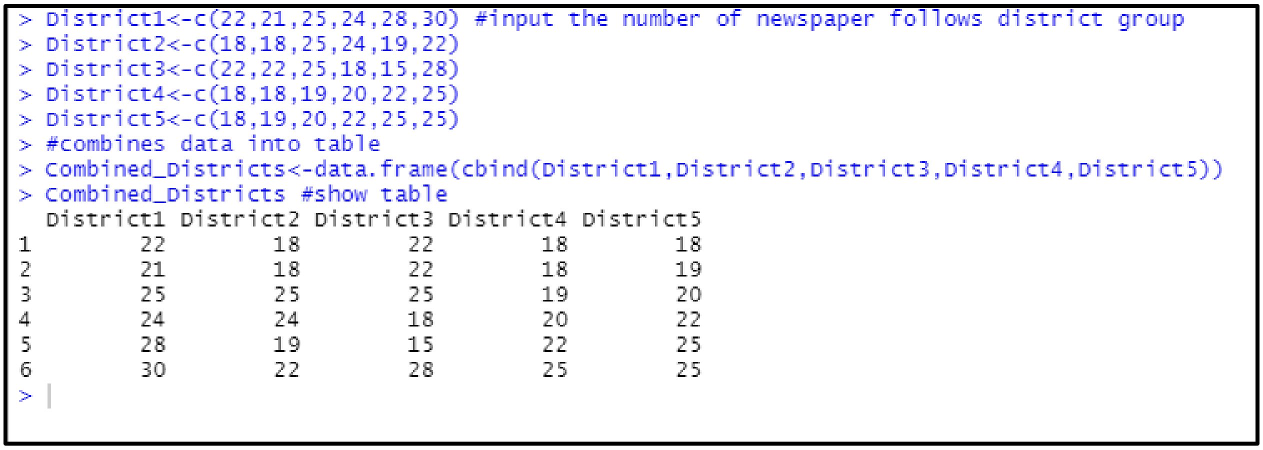
Step 1: Check assumptions and write hypotheses

Hypotheses

H0 : Assume that the amount of newspaper sold is the same in 5 districts.

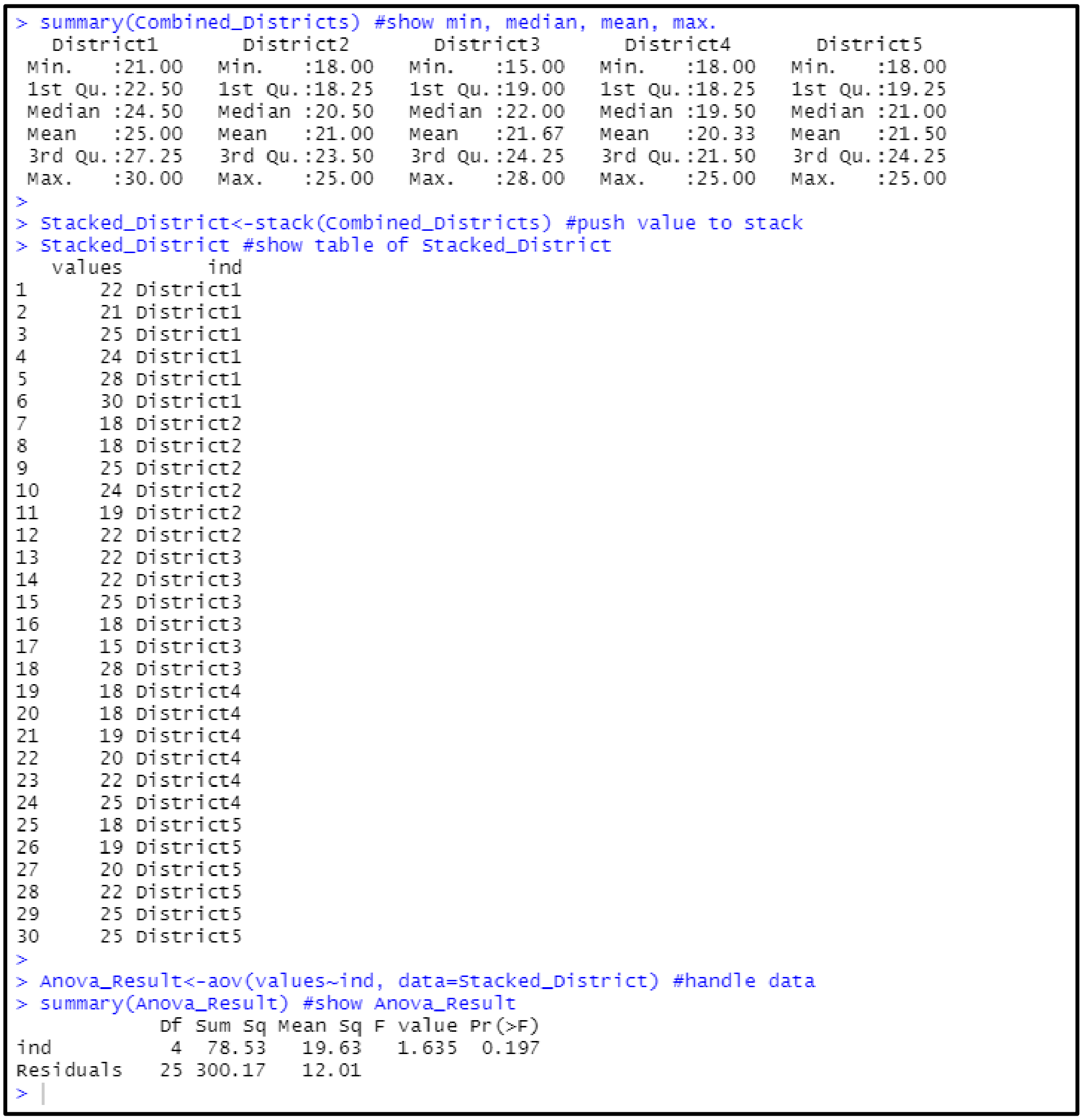
H1 : Assume that the amount of newspaper sold is different in 5 districts.

Step 2: Input the number of newspaper follows district group and combines data into table



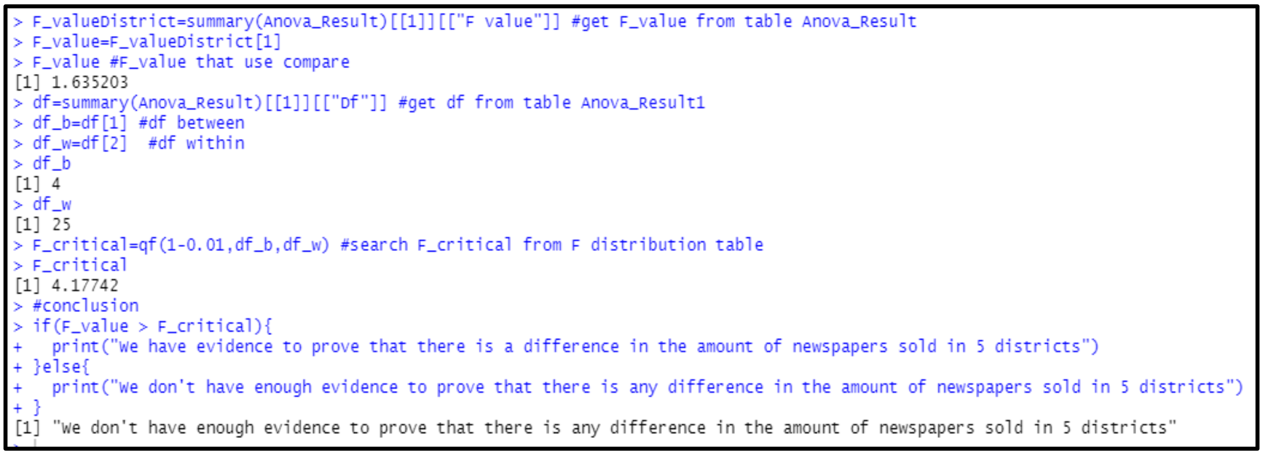
```
> District1<-c(22,21,25,24,28,30) #input the number of newspaper follows district group
> District2<-c(18,18,25,24,19,22)
> District3<-c(22,22,25,18,15,28)
> District4<-c(18,18,19,20,22,25)
> District5<-c(18,19,20,22,25,25)
> #combines data into table
> Combined_Districts<-data.frame(cbind(District1,District2,District3,District4,District5))
> Combined_Districts #show table
  District1 District2 District3 District4 District5
1        22        18        22        18        18
2        21        18        22        18        19
3        25        25        25        19        20
4        24        24        18        20        22
5        28        19        15        22        25
6        30        22        28        25        25
> 
```

Step 3: Show max, min, mean median and table of Stacked_District.



```
> summary(Combined_Districts) #show min, median, mean, max.
   District1       District2       District3       District4       District5    
 Min.   :21.00   Min.   :18.00   Min.   :15.00   Min.   :18.00   Min.   :18.00  
 1st Qu.:22.50   1st Qu.:18.25   1st Qu.:19.00   1st Qu.:18.25   1st Qu.:19.25  
 Median :24.50   Median :20.50   Median :22.00   Median :19.50   Median :21.00  
 Mean   :25.00   Mean   :21.00   Mean   :21.67   Mean   :20.33   Mean   :21.50  
 3rd Qu.:27.25   3rd Qu.:23.50   3rd Qu.:24.25   3rd Qu.:21.50   3rd Qu.:24.25  
 Max.   :30.00   Max.   :25.00   Max.   :28.00   Max.   :25.00   Max.   :25.00  
> 
> Stacked_District<-stack(Combined_Districts) #push value to stack
> Stacked_District #show table of Stacked_District
   values       ind
1      22 District1
2      21 District1
3      25 District1
4      24 District1
5      28 District1
6      30 District1
7      18 District2
8      18 District2
9      25 District2
10     24 District2
11     19 District2
12     22 District2
13     22 District3
14     22 District3
15     25 District3
16     18 District3
17     15 District3
18     28 District3
19     18 District4
20     18 District4
21     19 District4
22     20 District4
23     22 District4
24     25 District4
25     18 District5
26     19 District5
27     20 District5
28     22 District5
29     25 District5
30     25 District5
> 
> Anova_Result<-aov(values~ind, data=Stacked_District) #handle data
> summary(Anova_Result) #show Anova_Result
            Df Sum Sq Mean Sq F value Pr(>F)
ind          4  78.53   19.63   1.635  0.197
Residuals   25 300.17   12.01               
> 
```

Step 4: Use qf to find $F_{critical}$ to compare and conclusion



```
> F_valueDistrict=summary(Anova_Result)[[1]][["F value"]] #get F_value from table Anova_Result
> F_value=F_valueDistrict[1]
> F_value #F_value that use compare
[1] 1.635203
> df=summary(Anova_Result)[[1]][["Df"]] #get df from table Anova_Result1
> df_b=df[1] #df between
> df_w=df[2]  #df within
> df_b
[1] 4
> df_w
[1] 25
> F_critical=qf(1-0.01,df_b,df_w) #search F_critical from F distribution table
> F_critical
[1] 4.17742
> #conclusion
> if(F_value > F_critical){
+   print("We have evidence to prove that there is a difference in the amount of newspapers sold in 5 districts")
+ }else{
+   print("We don't have enough evidence to prove that there is any difference in the amount of newspapers sold in 5 districts")
+ }
[1] "We don't have enough evidence to prove that there is any difference in the amount of newspapers sold in 5 districts"
```

We can see that:

$$F_{Value} < F_{Critical} \ (1.635 < 4.177)$$

Step 5: Conclusion

Conclusion: We don’t have enough evidence to prove that there is any differences in the amount of newspapers sold in 5 districts.

Below is the full code implementation of the exercise:

```
#H0 :assume that the amount of newspaper sold is the same in 5 districts
#H1 :the amount of newspaper sold is at least different in 5 districts

#input the number of newspaper follows district group
District1<-c(22,21,25,24,28,30)
District2<-c(18,18,25,24,19,22)
District3<-c(22,22,25,18,15,28)
District4<-c(18,18,19,20,22,25)
District5<-c(18,19,20,22,25,25)
#combines data into table
Combined_Districts<-
data.frame(cbind(District1,District2,District3,District4,District5))
Combined_Districts #show table
summary(Combined_Districts) #show min, median, mean, max.

Stacked_District<-stack(Combined_Districts) #push value to stack
Stacked_District #show table of Stacked_District

Anova_Result<-aov(values~ind, data=Stacked_District) #handle data
summary(Anova_Result) #show Anova_Result

F_valueDistrict=summary(Anova_Result)[[1]][["F value"]]
#get F_value from table Anova_Result
F_value=F_valueDistrict[1]
F_value #F_value that use compare

df=summary(Anova_Result)[[1]][["Df"]] #get df from table Anova_Result1
df_b=df[1] #df between
df_w=df[2]  #df within

df_b
```

```r
df_w
F_critical=qf(1-0.01,df_b,df_w) #search F_critical from F distribution table
F_critical

#conclusion
if(F_value > F_critical){
  print("We have evidence to prove that there is a difference in the amount of newspapers sold in 5 districts")
}else{
  print("We don't have enough evidence to prove that there is any difference in the amount of newspapers sold in 5 districts")
}
```

4.4.2. Is the amount of newspapers sold affected by the day of week?

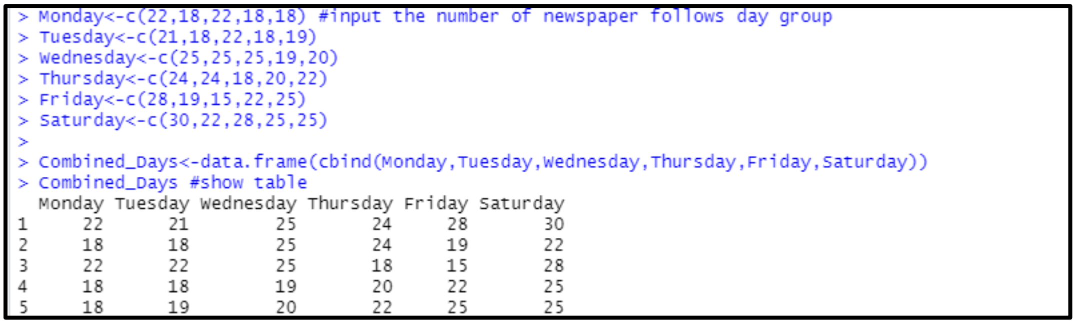
Step 1: Check assumptions and write hypotheses

Hypotheses

H_0 : Assume that the amount of newspaper sold is the same in 5 districts.

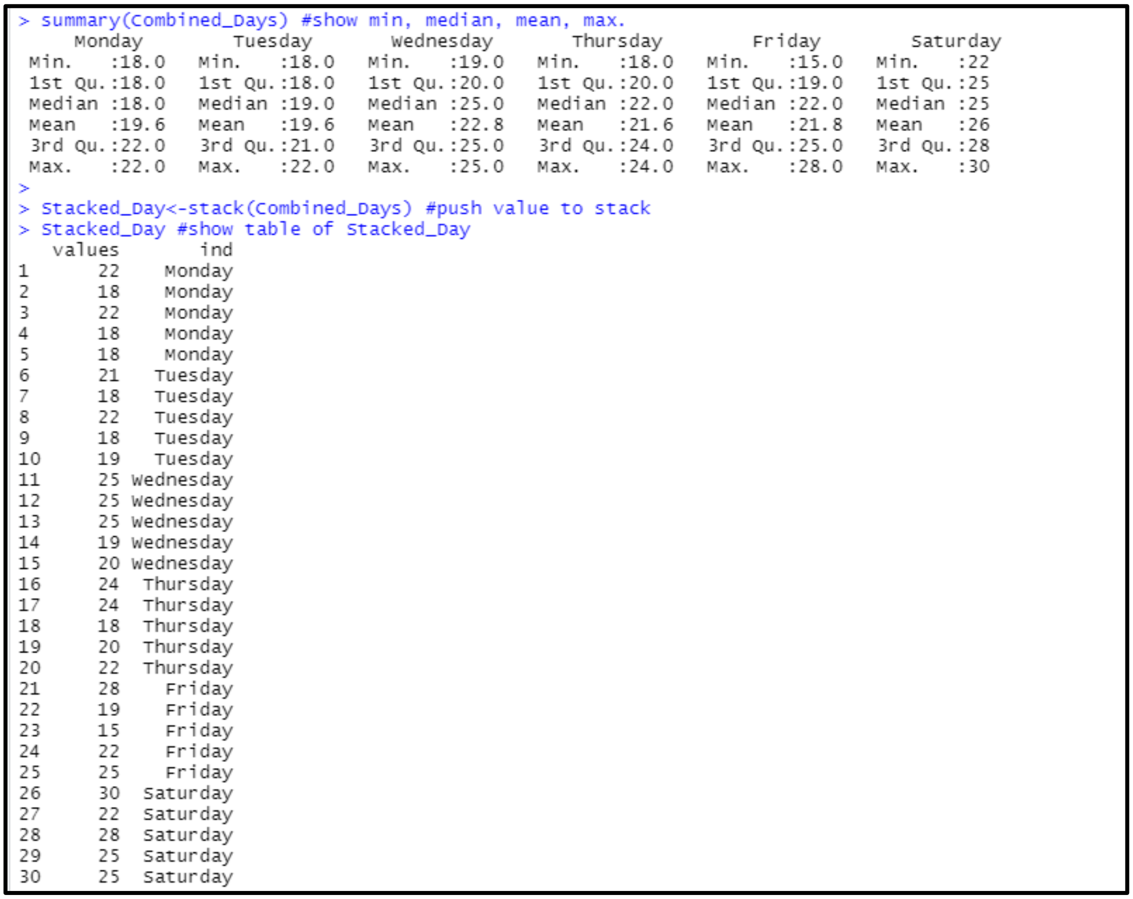
H_1 : Assume that the amount of newspaper sold is different in 5 districts.

Step 2: Input the number of newspaper follows Dates and combines data into table



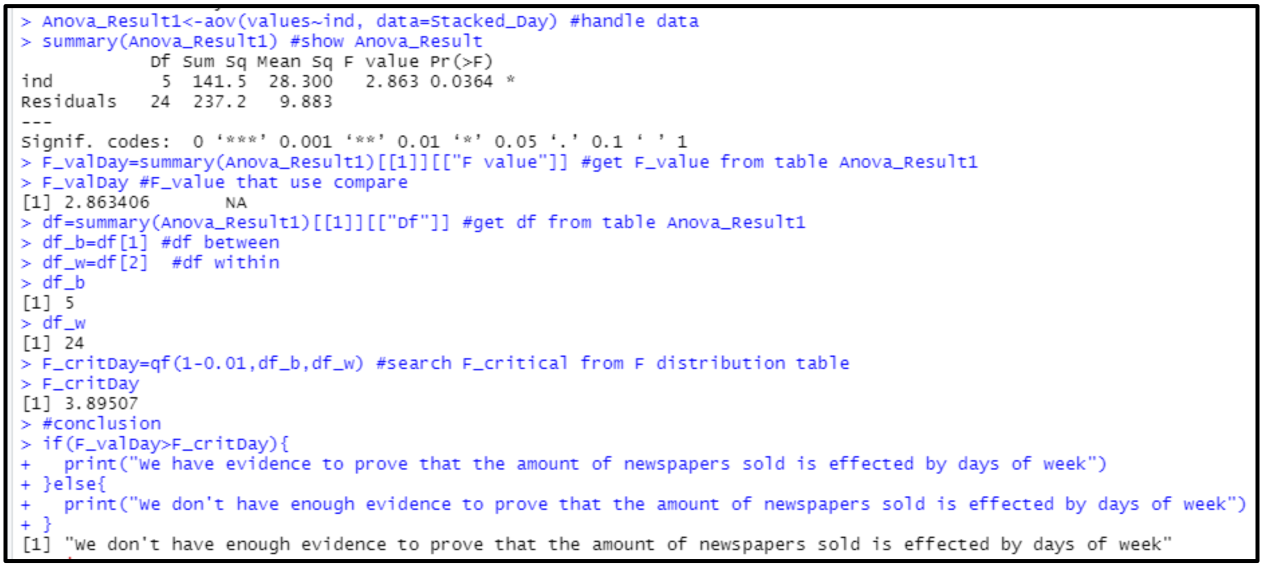
```
> Monday<-c(22,18,22,18,18) #input the number of newspaper follows day group
> Tuesday<-c(21,18,22,18,19)
> Wednesday<-c(25,25,25,19,20)
> Thursday<-c(24,24,18,20,22)
> Friday<-c(28,19,15,22,25)
> Saturday<-c(30,22,28,25,25)
>
> Combined_Days<-data.frame(cbind(Monday,Tuesday,Wednesday,Thursday,Friday,Saturday))
> Combined_Days #show table
  Monday Tuesday Wednesday Thursday Friday Saturday
1     22      21        25       24     28       30
2     18      18        25       24     19       22
3     22      22        25       18     15       28
4     18      18        19       20     22       25
5     18      19        20       22     25       25
```

Step 3: Show max, min, mean median and table of Stacked_District.



```
> summary(Combined_Days) #show min, median, mean, max.
     Monday          Tuesday         Wednesday        Thursday         Friday          Saturday
 Min.   :18.0   Min.   :18.0   Min.   :19.0   Min.   :18.0   Min.   :15.0   Min.   :22
 1st Qu.:18.0   1st Qu.:18.0   1st Qu.:20.0   1st Qu.:20.0   1st Qu.:19.0   1st Qu.:25
 Median :18.0   Median :19.0   Median :25.0   Median :22.0   Median :22.0   Median :25
 Mean   :19.6   Mean   :19.6   Mean   :22.8   Mean   :21.6   Mean   :21.8   Mean   :26
 3rd Qu.:22.0   3rd Qu.:21.0   3rd Qu.:25.0   3rd Qu.:24.0   3rd Qu.:25.0   3rd Qu.:28
 Max.   :22.0   Max.   :22.0   Max.   :25.0   Max.   :24.0   Max.   :28.0   Max.   :30
>
> Stacked_Day<-stack(Combined_Days) #push value to stack
> Stacked_Day #show table of Stacked_Day
   values       ind
1      22    Monday
2      18    Monday
3      22    Monday
4      18    Monday
5      18    Monday
6      21   Tuesday
7      18   Tuesday
8      22   Tuesday
9      18   Tuesday
10     19   Tuesday
11     25 Wednesday
12     25 Wednesday
13     25 Wednesday
14     19 Wednesday
15     20 Wednesday
16     24  Thursday
17     24  Thursday
18     18  Thursday
19     20  Thursday
20     22  Thursday
21     28    Friday
22     19    Friday
23     15    Friday
24     22    Friday
25     25    Friday
26     30  Saturday
27     22  Saturday
28     28  Saturday
29     25  Saturday
30     25  Saturday
```

Step 4: Use qf to find $F_{CritDay}$ to compare and conclusion



```
> Anova_Result1<-aov(values~ind, data=Stacked_Day) #handle data
> summary(Anova_Result1) #show Anova_Result
            Df Sum Sq Mean Sq F value Pr(>F)
ind          5  141.5  28.300   2.863 0.0364 *
Residuals   24  237.2   9.883
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> F_valDay=summary(Anova_Result1)[[1]][["F value"]] #get F_value from table Anova_Result1
> F_valDay #F_value that use compare
[1] 2.863406       NA
> df=summary(Anova_Result1)[[1]][["Df"]] #get df from table Anova_Result1
> df_b=df[1] #df between
> df_w=df[2]  #df within
> df_b
[1] 5
> df_w
[1] 24
> F_critDay=qf(1-0.01,df_b,df_w) #search F_critical from F distribution table
> F_critDay
[1] 3.89507
> #conclusion
> if(F_valDay>F_critDay){
+   print("We have evidence to prove that the amount of newspapers sold is effected by days of week")
+ }else{
+   print("We don't have enough evidence to prove that the amount of newspapers sold is effected by days of week")
+ }
[1] "We don't have enough evidence to prove that the amount of newspapers sold is effected by days of week"
```

We can see that:

$$F_{valDay} < F_{CritDay}(2.863 < 3.895)$$

Step 5: Conclusion

Conclusion: We don’t have enough conclusion to prove that the amount of newspapers is effected by days of the week.

Below is the full code implementation of the exercise:

```
#H0: assume that the amount of newspaper sold is not effected by days of weeks
#H1: the amount of newspaper sold is effected by days of weeks
#input the number of newspaper follows day group
Monday<-c(22,18,22,18,18)
Tuesday<-c(21,18,22,18,19)
Wednesday<-c(25,25,25,19,20)
Thursday<-c(24,24,18,20,22)
Friday<-c(28,19,15,22,25)
Saturday<-c(30,22,28,25,25)

Combined_Days<-data.frame(cbind(Monday,Tuesday,Wednesday,Thursday,Friday,Saturday))
Combined_Days #show table
summary(Combined_Days) #show min, median, mean, max.

Stacked_Day<-stack(Combined_Days) #push value to stack
Stacked_Day #show table of Stacked_Day

Anova_Result1<-aov(values~ind, data=Stacked_Day) #handle data
summary(Anova_Result1) #show Anova_Result


F_valDay=summary(Anova_Result1)[[1]][["F value"]] #get F_value from table
Anova_Result1
F_valDay #F_value that use compare

df=summary(Anova_Result1)[[1]][["Df"]] #get df from table Anova_Result1
df_b=df[1] #df between
df_w=df[2]  #df within

df_b
df_w
```

```
F_critDay=qf(1-0.01,df_b,df_w) #search F_critical from F distribution table
F_critDay
#conclusion
if(F_valDay>F_critDay){
  print("We have evidence to prove that the amount of newspapers sold is effected by days of week")
}else{
  print("We don't have enough evidence to prove that the amount of newspapers sold is effected by days of week")
}
```