**VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY**

**UNIVERSITY OF INFORMATION TECHNOLOGY**

**INFORMATION SYSTEM FACULTY**



**LAB 1 REPORT**

**BUSINESS ANALYSIS**

Lecturer: [Nguyen Dinh Thuan](https://courses.uit.edu.vn/user/view.php?id=490&course=1)

Instructor: Nguyen Thi Viet Huong

Student performance: Team 15

Nguyen Cam Tu 20520837

Nguyen Thi Thu Thuy 20520797

Nguyen Thanh Son 20521847

**HO CHI MINH CITY, MARCH 2023**

# TEACHER’S COMMENT

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

# TABLE OF CONTENT

[TEACHER’S COMMENT 2](#_Toc131043109)

[TABLE OF CONTENT 3](#_Toc131043110)

[JOB DISTRIBUTION 4](#_Toc131043111)

[A. 6](#_Toc131043112)

[1. What is Levene Test for Equality of Variances? Explanation and example. 6](#_Toc131043113)

[2. What are post hoc comparison tests used for in ANOVA? Explanation and example 7](#_Toc131043114)

[B. Energy Drink Survey 10](#_Toc131043115)

[1. Using MS Excel 10](#_Toc131043116)

[2. Using R language 13](#_Toc131043117)

[3. Python language 15](#_Toc131043118)

[C. Insurance survey 17](#_Toc131043119)

[1. Using MS Excel 17](#_Toc131043120)

[2. Using R language 21](#_Toc131043121)

[3. Using Python language 23](#_Toc131043122)

[D. VIETNAM NATIONAL HIGHSCHOOL EXAM SCORE 2018 (with optional hypotheses) 26](#_Toc131043123)

[1. Using MS Excel 26](#_Toc131043124)

[2. Using R language 32](#_Toc131043125)

[3. Using Python language 40](#_Toc131043126)

[E. REFERENCES 42](#_Toc131043127)

# JOB DISTRIBUTION

|  |  |  |  |
| --- | --- | --- | --- |
| **Member**  **Work** | **Nguyen Cam Tu (Leader)** | **Nguyen Thi Thu Thuy** | **Nguyen Thanh Son** |
| Problem statement | ✓ | ✓ | ✓ |
| Build a report template | ✓ | ✓ | ✓ |
| Summarize and edit reports | ✓ |  |  |
| Do Exercise 1 | ✓ |  |  |
| Do Exercise 2 *Energy Drink Survey* (Excel, R, Python) | ✓ |  |  |
| Do Exercise 3 *Insurance survey* (Excel, R, Python) |  |  | ✓ |
| Do Exercise 4 VIETNAM NATIONAL HIGHSCHOOL EXAM SCORE 2018(Excel, R, Python) |  | ✓ |  |
| Summarize and edit reports | ✓ | ✓ | ✓ |
| Completion (%) | 90% | 90% | 90% |

|  |
| --- |
| **Lab 2**  **1.** a) What is [Levene Test for Equality of Variances? Explanation and example.](https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm)  b) What are post hoc comparison tests used for in ANOVA? Explanation and example  **2.** Using MS Excel, R language and Python language to perform Chi Square test on the independence of two categorical variables with the data file: *Energy Drink Survey*  **3.** Using MS Excel, R language and Python language to perform ANOVA with data file (including Levene, ANOVA, Tukey Test): *Insurance survey*  **4.** Using R language and Python language to perform ANOVA with data file about VIETNAM NATIONAL HIGHSCHOOL EXAM SCORE 2018 (with optional hypotheses)  <https://www.kaggle.com/ngvietlg/vietnam-national-highschool-exam-score-2018/version/1> |

# A.

## What is [Levene Test for Equality of Variances? Explanation and example.](https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm)

Levene's test is therefore used to test the null hypothesis that the samples to be compared come from a population with the same variance. In this case, possible variance differences occur only by chance, since there are small differences in each sampling.

If the p-value for the Levene test is greater than .05, then the variances are not significantly different from each other (i.e., the homogeneity assumption of the variance is met). If the p-value for the Levene's test is less than .05, then there is a Significant difference between the variances.

H0: Groups have equal variances

H1: Groups have different variances

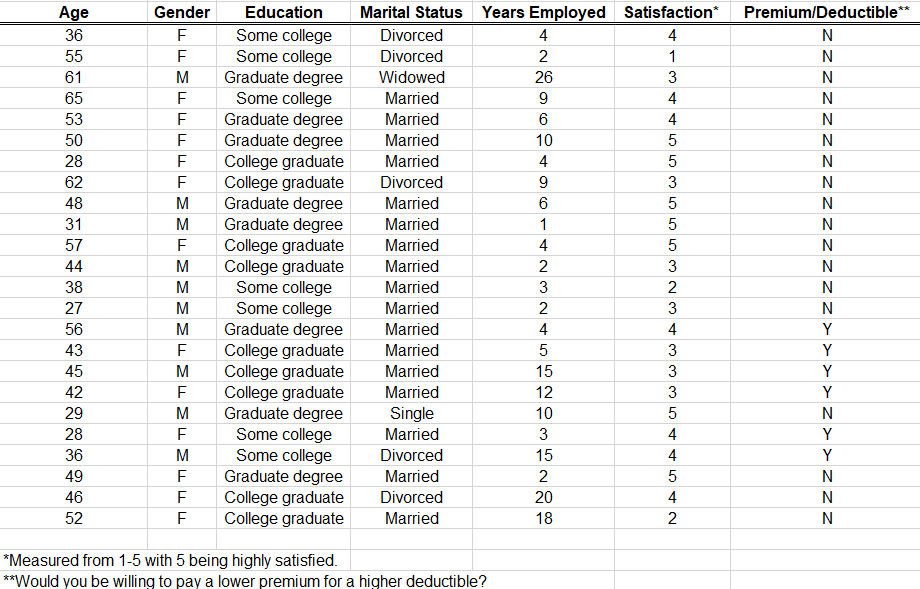
It is important to note that the mean values of the individual groups have no influence on the result, they may differ. A big advantage of Levene's test is that it is very stable against violations of the normal distribution. Therefore, Levene's test is used in many statistics programs.

Furthermore, the variance equality can also be checked graphically, this is usually done with a grouped box-plot or with a Scatterplot.

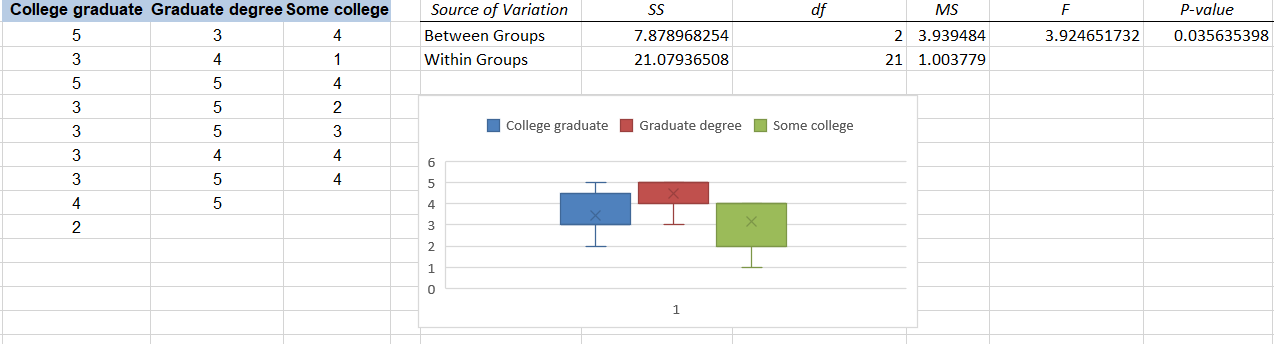
*Explanation:*

The assumption of equality of variances is important in statistical analysis and often shapes the researcher’s procedures when working on measuring the outcomes of experiments and data analysis.

*Example:*

The researcher wanted to know if three different types of education affected their satisfaction with education. They randomly selected 27 different students and divided them into 3 groups of educational attainment: Some College, College Graduate, College Graduate and gave satisfaction levels 1-5.

The data table shows the level of satisfaction measured from 1-5 with 5 being highly satisfied for each different education.



According the figure, p-value is under 0.05, so there is different between the variances of examples.

## What are post hoc comparison tests used for in ANOVA? Explanation and example

In order to find out exactly which groups are different from each other, we must conduct a post hoc test (also known as a multiple comparison test), which will allow us to explore the difference between multiple group means while also controlling for the family-wise error rate.

*One-Way ANOVA:*

An ANOVA is a statistical test that is used to determine whether or not there is a statistically significant difference between the means of three or more independent groups.

The hypotheses used in an ANOVA are as follows:

The null hypothesis (H0): µ1 = µ2 = µ3 = … = µk (the means are equal for each group)

The alternative hypothesis: (Ha): at least one of the means is different from the others

If the p-value from the ANOVA is less than the significance level, we can reject the null hypothesis and conclude that we have sufficient evidence to say that at least one of the means of the groups is different from the others.

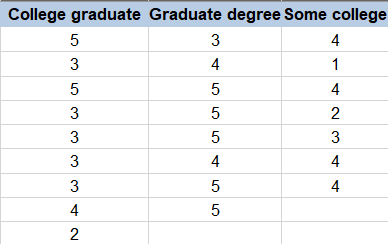
*The Family-Wise Error Rate*

As mentioned before, post hoc tests allow us to test for difference between multiple group means while also controlling for the family-wise error rate.

In a hypothesis test, there is always a type I error rate, which is defined by our significance level (alpha) and tells us the probability of rejecting a null hypothesis that is actually true. In other words, it’s the probability of getting a “false positive”, i.e. when we claim there is a statistically significant difference among groups, but there actually isn’t.

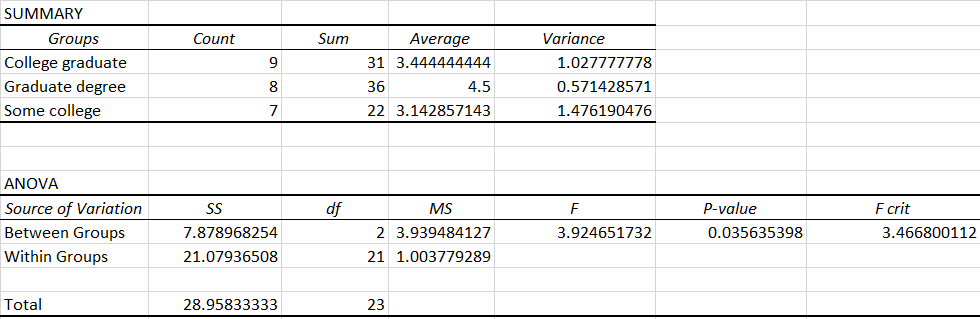
*Multiple Comparisons in ANOVA*

When we conduct an ANOVA, there are often three or more groups that we are comparing to one another. Thus, when we conduct a post hoc test to explore the difference between the group means, there are several pairwise comparisons we want to explore.

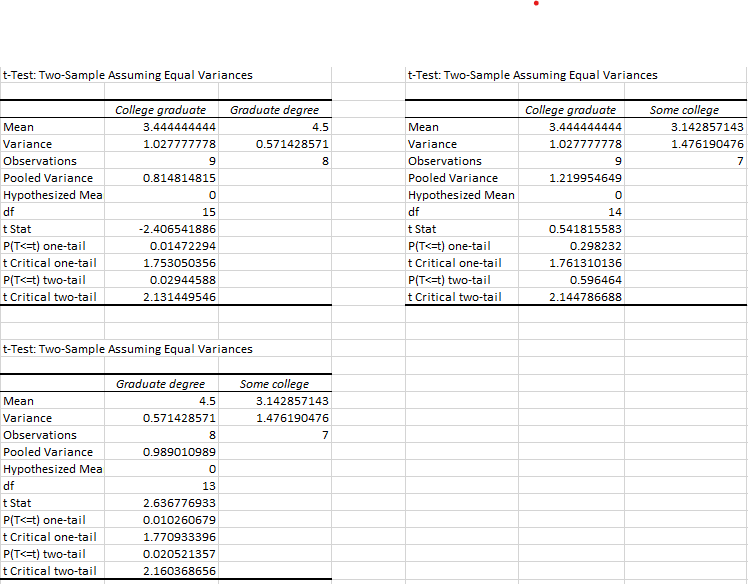
*Example One-Way ANOVA to Use with Post Hoc Tests:*

H0: Satisfaction of 3 groups is not different

H1: Satisfaction of 3 groups is different



We have p-value = 0.035635 < α = 0.05 so we can aject H0. Therefore, Satisfaction of 3 groups is different.

**Use Turkey to find the different number of 3 groups.

# B. Energy Drink Survey

1. Using MS Excel, R language and Python language to perform Chi Square test on the independence of two categorical variables with the data file: *Energy Drink Survey.*

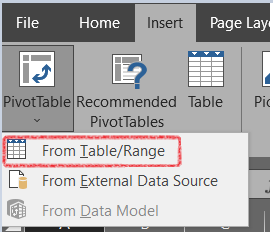
Test relationship between variable Gender and Brand Preference.

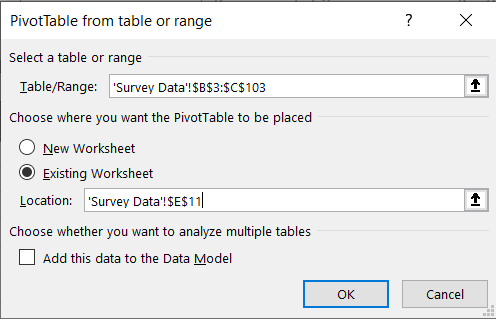
*Define theory:*

H0: Variable Gender and Brand Preference are independent variables.

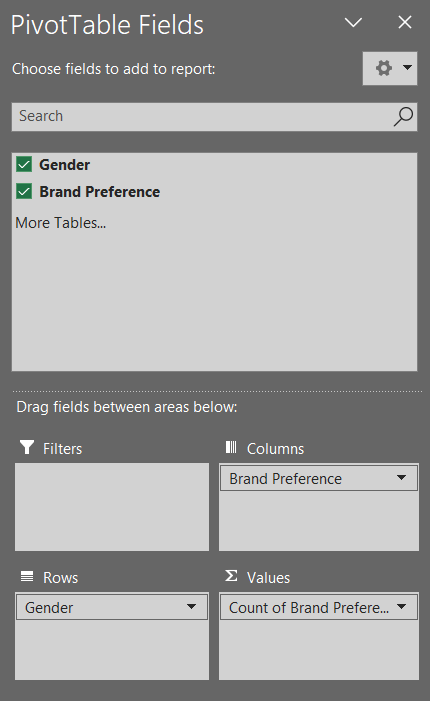
H1: Variable Gender and Brand Preference are not independent variables.

## Using MS Excel

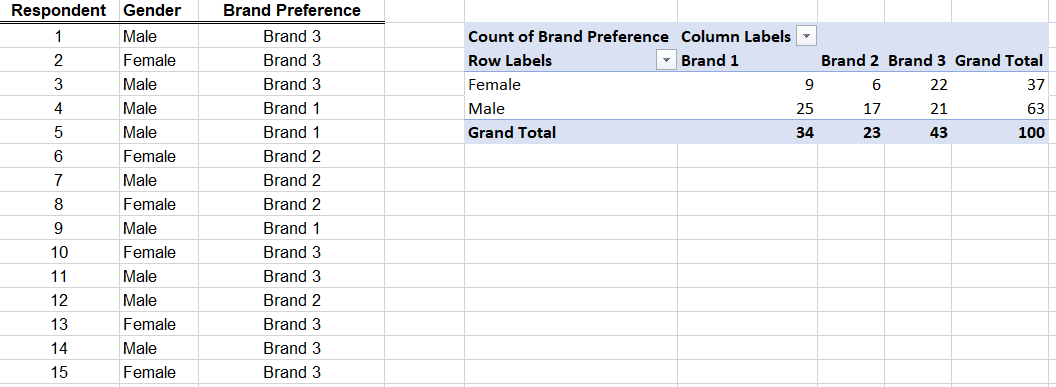
Using Pivot Table to do statistics: Insert PivotTable From Table/Range

Select the range of observed values Choose Existing Worksheet Select location of the analyze table OK

The new sheet exists, with 3 fields that we need to contact: columns, rows and values. Columns and rows form a table. Values are numbers that we want to calculate (normally, that’s the value in columns field).

In two variables, we can choose all of them in columns or rows; or separate them into 2 fields: columns and rows. Finally, select Gender into Values to complete an analyze table.

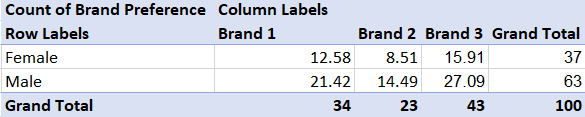
After select all fields, we will have an analyze table.



*Calculate according to the formula:*

**Step 1:** Calculate Expected Values

Expected value = (row sum \* column sum) / table sum

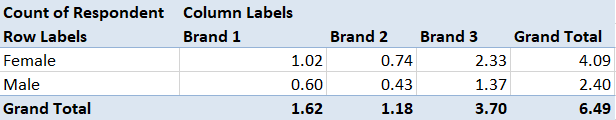
*For example: at (Brand 1, Female), 37 \* 34 / 100 12.58*

**Step 2:** (O-E)^2/E

O (observed): observed value that in the first analyze table.

E (Expected): expected value that in the step 1 table.

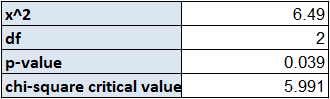
*For example: at (Brand 1, Female), we have O = 9, E = 12.58*

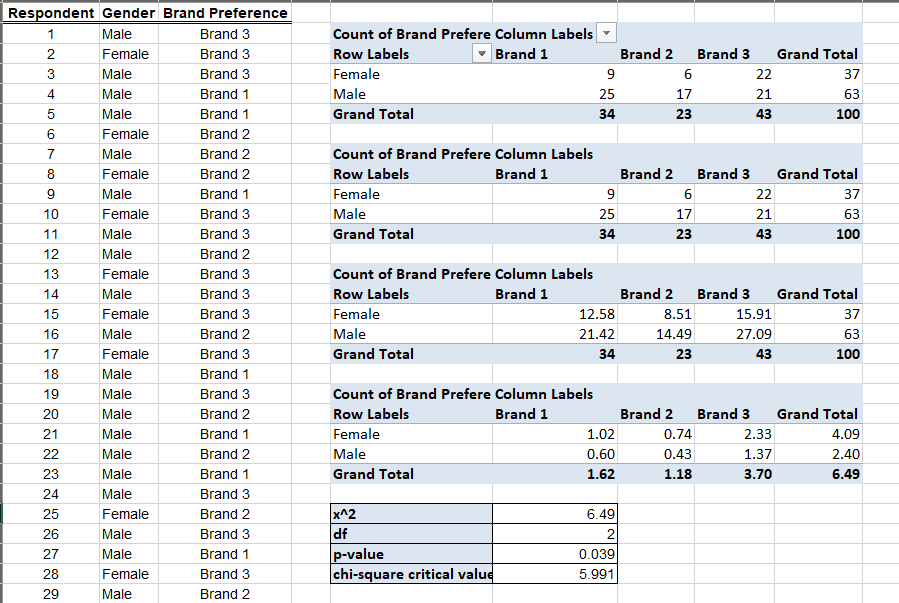
 *(9 – 12.58) ^ 2 / 12.58 1.02*

**Step 3:** Calculate test statistics x2, p-value and df

X2 = SUM(F21:H22)

df = (row – 1) \* (column – 1)

p-value = CHISQ.DIST.RT(X2, df)  
 chi-square critical value = CHISQ.INV.RT(0.05, df)

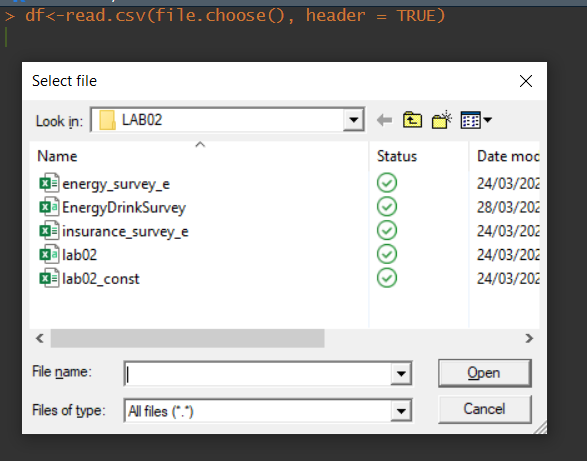
All calculations in this problem:

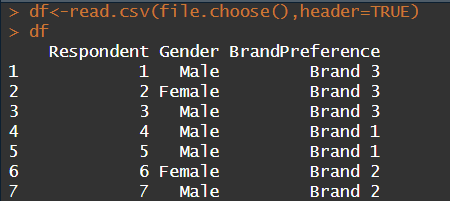
According the result:

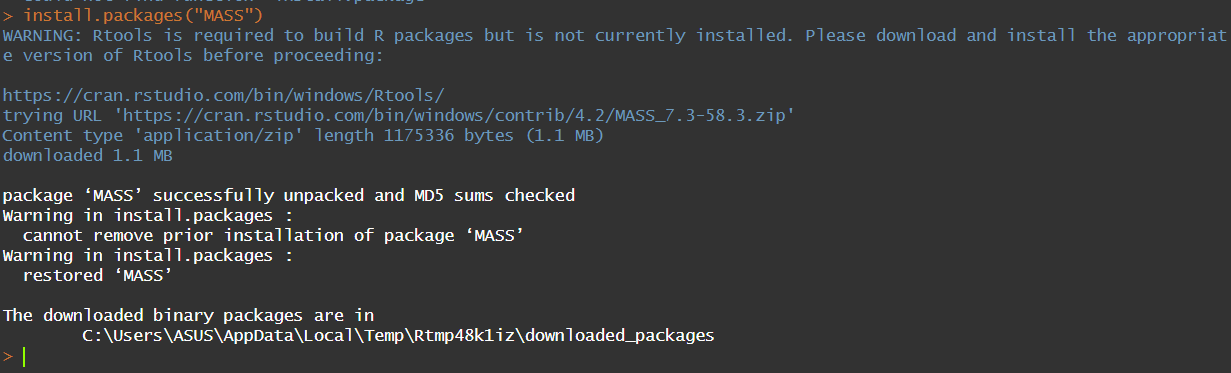
p-value 0.039 < α = 0.05 so wereject H0

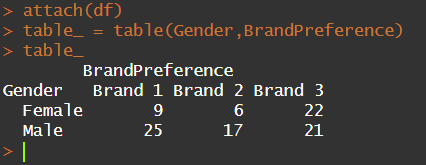
Therefore, variable Grand and Brand Preference are not independent variables.

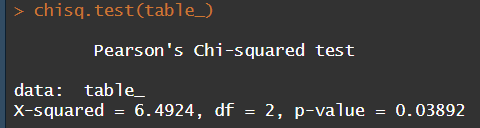
## Using R language

Choose data file

Read and check data

 Install package MASS to define Chi-Square

 Do statistics to define Chi-Square

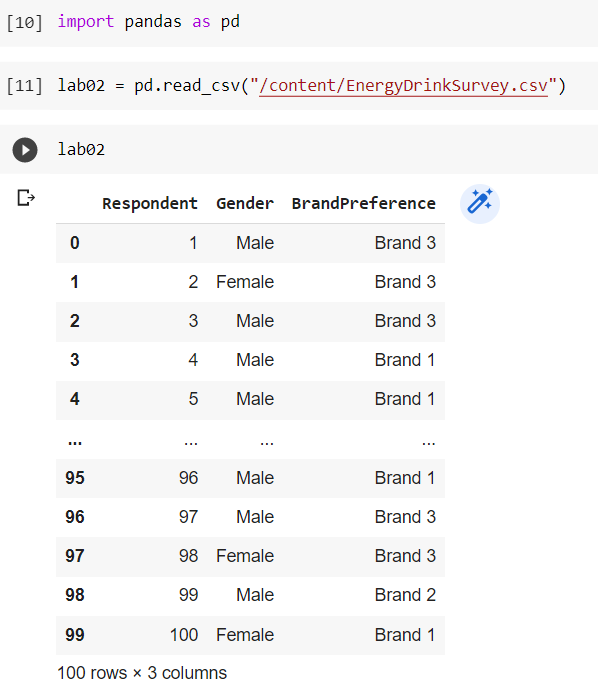
 Use command chiq.test() to define

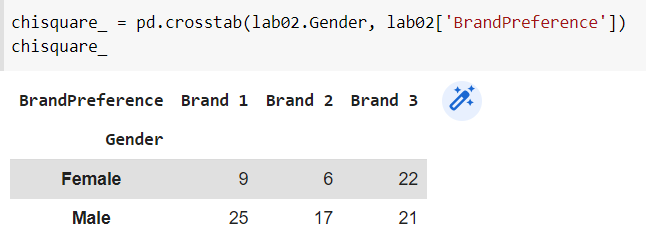
According the result:

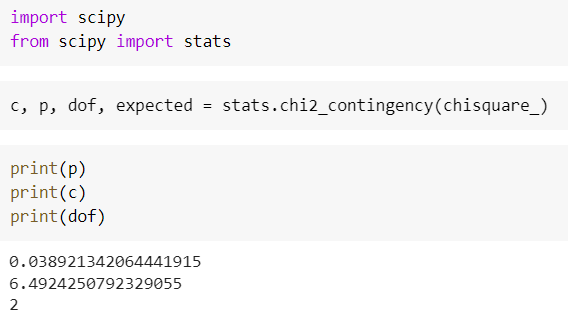
p-value 0.039 < α = 0.05 so wereject H0.

Therefore, variable Grand and Brand Preference are not independent variables.

## Python language

Import libraries, read and check data

Do statistics to define Chi-Square

Result:

Note:

chi2: test statistics

p: p-value

dof: free degree

expected: values that we hope it achieves.

According the result:

p-value 0.039 < α = 0.05 so wereject H0.

Therefore, variable Grand and Brand Preference are not independent variables.

# C. Insurance survey

to perform ANOVA with data file (including Levene, ANOVA, Tukey Test)

## Using MS Excel

* 1. ANOVA

Determine your hypothesis:

H0: x1 = x2 = x3

H1: At least one mean is different from the rest of the mean

In there:

x1: College graduate mean

x2: Average of Graduate degree

x3: Mean value of Some college

**Calculation method**: Select Data -> Data Analysis -> Anova: Single Factor -> Select observed values -> OK

A screenshot of a computer

Description automatically generated with medium confidence

Graphical user interface, application, table, Excel

Description automatically generated

**In there:**

Count: Number of observed values

Sum: Sum of observed values

Average: Average of observed values

Variance: a measure of the difference between the numbers in a data set relative to the mean of the data set.

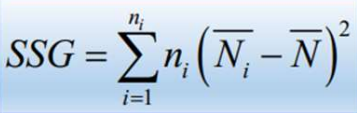
Table

Description automatically generated

In there:

Between groups:

+ SSG (between-groups sum of squares): sum of squares of deviations between groups.



+ df (Degree of Freedom): degrees of freedom. Calculated by the formula

where k is the number of groups

**df=k-1**

k is the number of groups

+ MSG (Between-groups mean square): mean squared between groups.

Text

Description automatically generated

Within groups:

+ SSW: Sum of intra-group variations.

A picture containing text, clock, gauge

Description automatically generated

+ df (Degree of Freedom): degrees of freedom, calculated by the formula:

**𝑑𝑓 = (the value group1 – 1) + (the value group2 – 1) + (the value group3 – 1)**

+ MSW (Within-groups mean square): the internal mean square of the

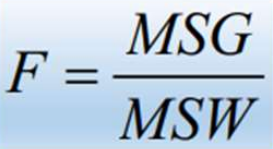
group.

SST (Total-sum of squares): sum of squares of deviations between each observation and average of all observations

Text

Description automatically generated with medium confidence

F: test value F



If Fstat > fcrit : reject H0

If Fstat < fcrit : accept H0

**Note:**

F = 3.939

F crit = 3.467 < 3.939

Cause F > F crit -> rejects H0. So there is at least one mean is different from the others.

* 1. Tukey

Since we reject H0, we will now perform a ANOVA’s Testing to determine specifically which group's mean is different from which group, larger or smaller.

**Calculation method:**

**Step 1:** Calculate Q-statistics

We have k = 3,

df = 21,

α = 0.05

Looking up the Turkey distribution table, we get Q statistic = 3,565

**Step 2**: Comparative Criterion T

Text, whiteboard

Description automatically generated

In there:

Q: q-statistics

K: Group number

n - k: k is the number of groups, n is the number of observations

α = 0.05

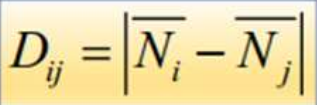
MSW: mean squared within groups

Nmin: Minimum number of observed values

Graphical user interface, application, table, Excel

Description automatically generated

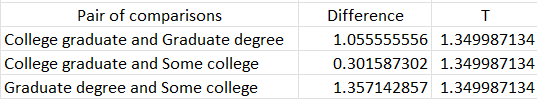
**Step 3**: Calculate the difference between 2 groups



In there:

xi: mean value of group i

xj: Mean value of group j



**Step 4**: Compare the difference between 2 pairs with T and make a conclusion

College graduate vs Graduate degree < T => x1 = x2

College graduate vs Some college < T => x1 = x3

Graduate degree vs Some college > T => x2 ≠ x3

**Conclusion:** So the average value of insurance satisfaction of Graduate degree

significantly different from Some .'s insurance satisfaction mean

college

## Using R language

Read and check data



Test data

Text

Description automatically generated with medium confidence

Check the relationship between Years Employed and Satisfaction

The hypothesis is as follows:

H0: The group variances are not significantly different.̃

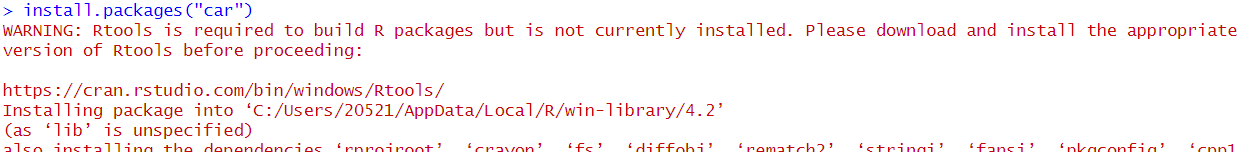
H1: The group variances are significantly different.

Prepare data:

Graphical user interface, text, application

Description automatically generated with medium confidence

Install the library R car:



Use the library:

Text

Description automatically generated

Levene’s Testing

Text

Description automatically generated

Note: Equal variance -> Anova’s Testing

Anova’s Testing

Text

Description automatically generated

Note: p-value > 0.05 accept the hypothesis H0

## Using Python language

Import libraries, read and check data

A picture containing graphical user interface

Description automatically generated



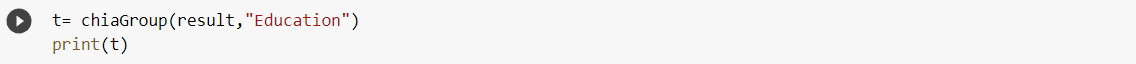
Table

Description automatically generated

Write group division function

Chart

Description automatically generated with medium confidence



Return column data to list

Graphical user interface, text, application, email

Description automatically generated

Leneve’s Testing

Graphical user interface, application

Description automatically generated

Anova’s Testing

Graphical user interface, application

Description automatically generated

Tukey’s Testing

Text

Description automatically generated with medium confidence

# D. VIETNAM NATIONAL HIGHSCHOOL EXAM SCORE 2018 (with optional hypotheses)

to perform ANOVA

Here the group will randomly select 99 rows of data about the point columns of blocks A, B, C, D, A1

## Using MS Excel

**Determine the Levene test hypothesis:**

**H0:** There is no difference in variance of 5 overall

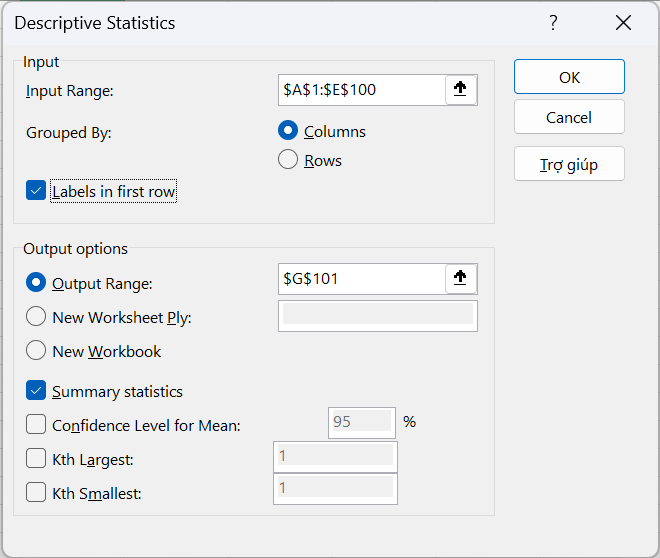
**H1:** There is a difference in the variance of the 5 populations

**Step 1:** Calculate statistics. On the toolbar, select "Data" 🡪 Data Analysis🡪Appears dialog... Click "Decription Statistics" 🡪OK

Ảnh có chứa bàn

Mô tả được tạo tự động

Select "Input Range", Grouped By: " by Column, check "Labels in first row", select "Output Range" and check "Summary statistics". Click OK



And the result is as shown

Ảnh có chứa bàn

Mô tả được tạo tự động

**Step 2:** Calculate the deviation in each group = ABS(xij – Mean(xi))

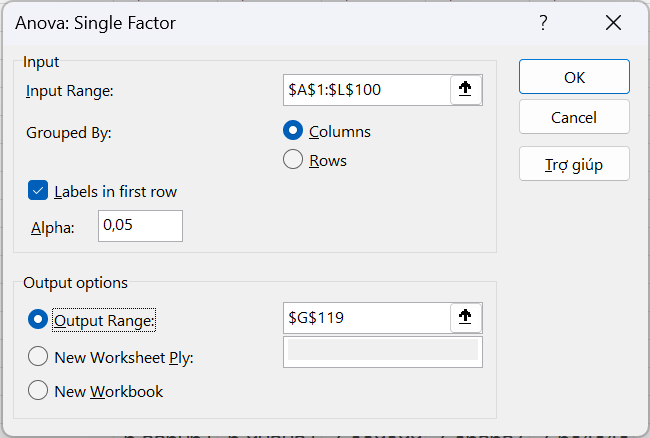
Ảnh có chứa bàn

Mô tả được tạo tự động

**Step 3:** On the toolbar select "Data" 🡪Data Analysis🡪A dialog box🡪appears Select the function "Anova: Single Factor" Select 🡪the area to observe🡪Click OK

Ảnh có chứa bàn

Mô tả được tạo tự động



And the result is as shown below

Ảnh có chứa bàn

Mô tả được tạo tự động

We can see that: Since p-value = 0 < a = 0.05, we do not accept the H0 hypothesis

🡪So there is a difference in the variance of the 5 masters

**ANOVA Accreditation:**

**H0:** There is no difference in variance of 3 overall(KhoiA, KhoiB, KhoiC)

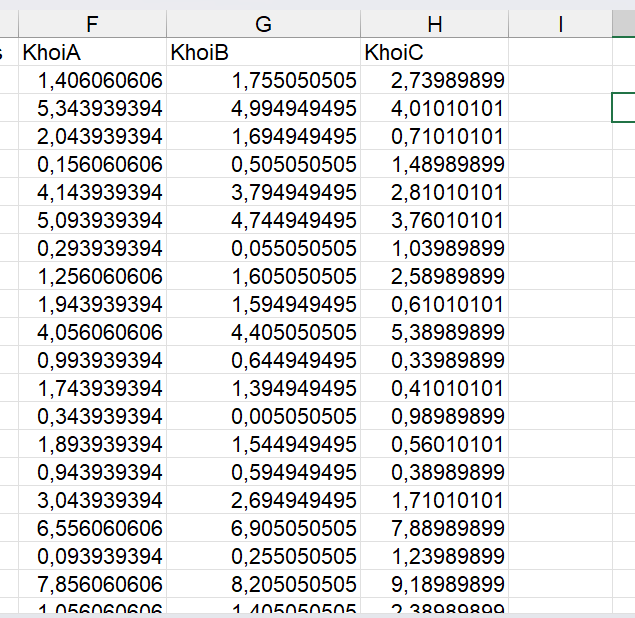
**H1:** There is a difference in the variance of the 3 populations

Where:

X1: mean value KhoiA

X2: average value KhoiB

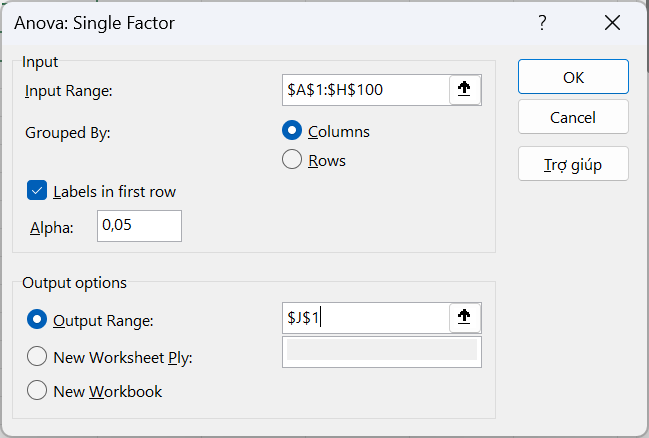
X3: mean KhoiC



On the toolbar select "Data" 🡪Data Analysis🡪A dialog box🡪appears Select the function "Anova: Single Factor" Select 🡪the area to observe🡪Click OK

Ảnh có chứa bàn

Mô tả được tạo tự động



And the result:

Ảnh có chứa bàn

Mô tả được tạo tự động

Meaning of the values in the result table:

**Count:** observed values

**Sum:** sum of observed values

**Average:** average of observed values

**Variance:** variance

**Between Groups:**

**SS**(between-groups sum of squares): The sum of the squares of the deviations between groups. Calculated by the formula:

Ảnh có chứa văn bản

Mô tả được tạo tự động

**Df**(Degree of freedom): degrees of freedom. Calculated by the formula:

Df = k – 1

Với: k là số nhóm

**MSG**(Between-groups mean square): Mean squared between groups. Calculated by the formula:

Ảnh có chứa văn bản

Mô tả được tạo tự động

**Within Groups:**

**SSW**(Sum of squares within): Total variability of groups. Calculated by the formula:

Ảnh có chứa văn bản

Mô tả được tạo tự động

**Df**(Degree of Freedom): degrees of freedom, calculated using the formula:

Df = (number of group values 1 – 1) + (number of group values 2 – 1) + (number of group values 3 – 1)

**MSW**(Within-groups mean square): the average square in each group. Calculated by the formula:

Ảnh có chứa văn bản

Mô tả được tạo tự động

## Using R language

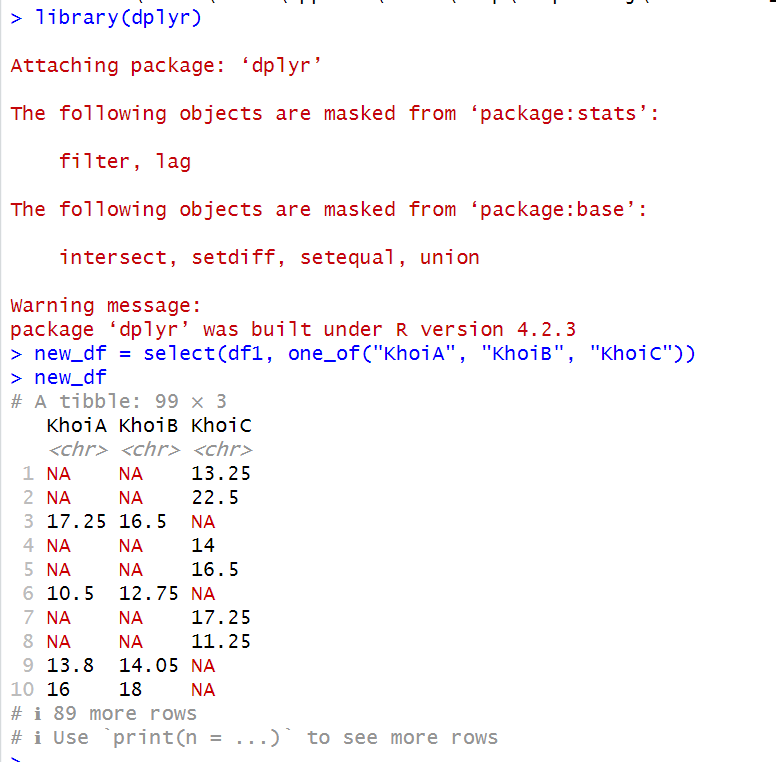
**Step 1:** Import data from the original file

Ảnh có chứa văn bản

Mô tả được tạo tự động

**Step 2:** Pre-process the data. Because the initial data contains many null data and many different populations, it is necessary to perform a mistaken preprocessing step to "clean" the data

Choose 3 overalls to be inspected



Use the stack() function to split groups

Ảnh có chứa bàn

Mô tả được tạo tự động

Use the omit() command to remove empty data fields

Ảnh có chứa bàn

Mô tả được tạo tự động

Randomly select 100 lines for analysis

Ảnh có chứa bàn

Mô tả được tạo tự động

Rename the value and group columns "Diem" and "Khoi" respectively

Ảnh có chứa bàn

Mô tả được tạo tự động

Export to new csv file to conduct inspection



**Step 4:** Calculate the Levene

Ảnh có chứa văn bản

Mô tả được tạo tự động

Since p-value = 0.09483 > a = 0.05, we accept H0. So there is no difference in the variance of the 3 masters

🡪Eligibility to conduct ANOVA accreditation

**ANOVA Accreditation**

Ảnh có chứa văn bản

Mô tả được tạo tự động

Since p-value = 0, 000354 < a = 0.05 🡪 Reject the H0 hypothesis. So there is at least one mean value that is different from the rest

**Tukey Accreditation**

Ảnh có chứa văn bản, bức thư

Mô tả được tạo tự động

Because:

BlockB - BlockA has p-value > a => Accept hypothesis H0(x1 = x2)

BlockC - BlockA has p-value < a => Does not accept hypothesis H0(x1 != x3)

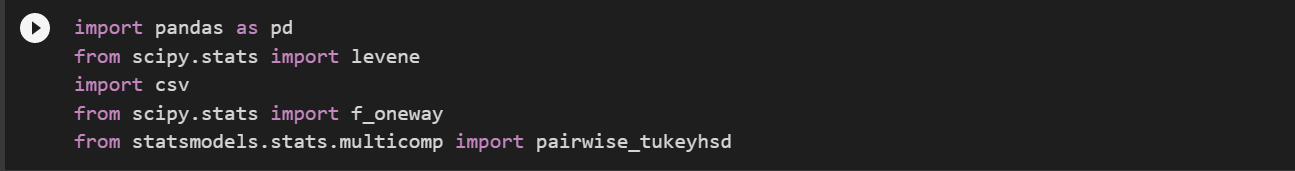
BlockC - Block B has p-value > a => Accept hypothesis H0(x1 != x2)

Conclusion: Two Blocks A and Block C, Block B and Block C have different average results, Block A and Block B have the same average results.

## Using Python language

Import libraries and data files

Use the divisiongroup() function to list groups that are unique in the dataframe



Ảnh có chứa văn bản

Mô tả được tạo tự động

Ảnh có chứa bàn

Mô tả được tạo tự động

Convert Blocks information to lists

Ảnh có chứa văn bản

Mô tả được tạo tự động

Ảnh có chứa văn bản

Mô tả được tạo tự động

Ảnh có chứa văn bản

Mô tả được tạo tự động

Since p-value < a = 0.05, we reject H0. So there is a difference in the variance of the 3 totals

Ảnh có chứa văn bản

Mô tả được tạo tự động

Ảnh có chứa văn bản

Mô tả được tạo tự động

# E. REFERENCES

[1] Nguyen Thi Viet Huong, Huynh Thi Anh Truc, Le Trinh Thanh Thuy, “BÁO CÁO LAB 2 PHÂN TÍCH DỮ LIỆU KINH DOANH”

[2] "Anova", https://www.excel-easy.com/examples/anova.html (accessed Mar. 30,2023)

[3] "Tukey Test Honest Significant Difference", https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/post-hoc/tukey-test-honest-significant-difference/ (accessed Mar. 30,2023)

[4] [Levene Test • Simply explained | DATAtab](https://datatab.net/tutorial/levene-test) (accessed Mar.30, 2023)

[5] Jim Frost, “Using Post Hoc Tests with ANOVA”, [Using Post Hoc Tests with ANOVA - Statistics By Jim](https://statisticsbyjim.com/anova/post-hoc-tests-anova/), (accessed Mar. 30, 2023)

[6] Zack, “A Guide to Using Post Hoc Tests with ANOVA”, April 14, 2019, “[A Guide to Using Post Hoc Tests with ANOVA - Statology](https://www.statology.org/anova-post-hoc-tests/)” (accessed Mar.30, 2023)

[6] Nguyen Minh Nhut, “LAB02\_PTDLKD”