**Lab 3**

***Thành viên nhóm:***

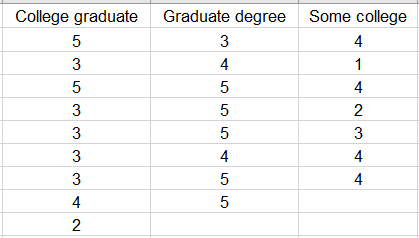
|  |  |  |
| --- | --- | --- |
| Số lượng thành viên | MSSV | *Họ và tên* |
| 1 | 18521320 | *Đoàn Thục Quyên* |
| 2 | 18521554 | *Nguyễn Thành Trung* |
| 3 | 18520454 | *Nguyễn Đức Anh* |

***Phân công công việc:***

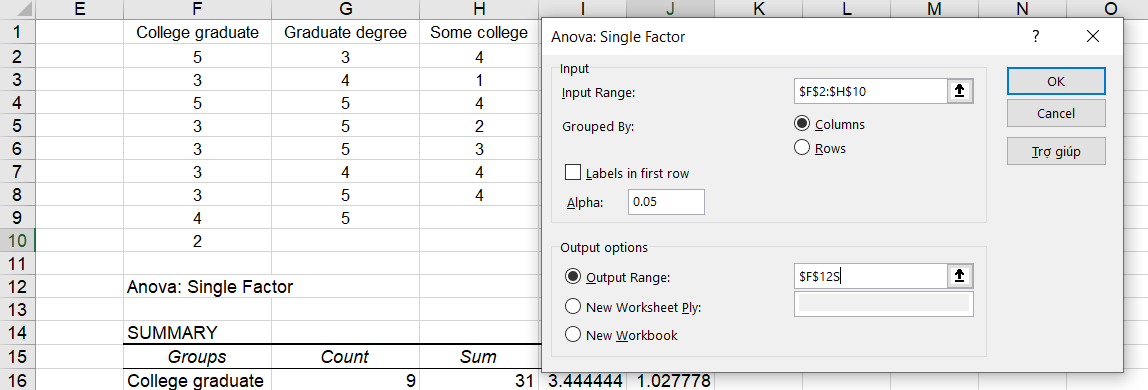
|  |  |  |  |
| --- | --- | --- | --- |
|  | Đoàn Thục Quyên  18521320 | Nguyễn Thành Trung  18521554 | Nguyễn Đức Anh  18520454 |
| Tập dữ liệu:  *Insurance Survey* | | | |
| Câu 3 a: ANOVA | Trên excel | Trên R | Tukey |
| Cau 3a : Chi-square | Trên excel | Trên R |  |
| Kiểm tra |  |  | x |
| Tập dữ liệu:  *Energy Drink Survey* | | | |
| Cau 3a: ANOVA | Trên R, Tukey | Trên Excel |  |
| Câu 3a: Chi-square | Trên R | Trên Excel |  |
| Kiểm tra | x |  | x |
| Trên tập dữ liệu Nhiệt độ của Việt Nam 2002 - 2017 | | | |
| Cau 3a: ANOVA | Trên R, Tukey | Trên Excel |  |
| Câu 3a: Chi-square | Trên R | Trên Excel |  |
| Kiểm tra | x |  | x |
| Trên tập dữ liệu Khảo sát chất lượng trồng sắn | | | |
| Câu 3 a: ANOVA | Trên excel | Trên R | Tukey |
| Cau 3a : Chi-square | Trên excel | Trên R |  |
| Kiểm tra |  |  | x |
| Câu c | x | x | x |

**Câu 3a**

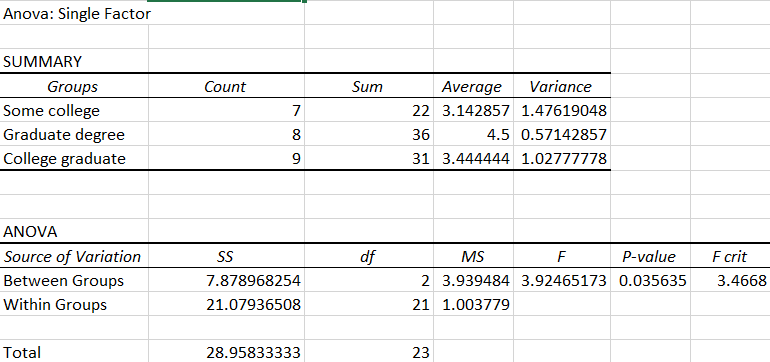
1. Kiểm định Anova và chi-square cho file: Insuance Suvey
2. Kiểm định Anova – xác định mức ảnh hướng của các nhân tố đối với kết quả thực nghiệm
   1. Dùng Excel
      1. Nhập dữ liệu theo cột



* + 1. Chọn tab data >> Chọn Data Analysis >> Chọn Anova Single Factor >> Input range: chọn dữ liệu từ file excel đã nhập ở trên



* + 1. Ta được bảng Anova như sau



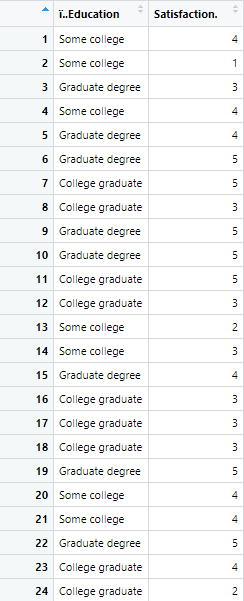
Nhìn bảng ta thấy F > Fcrit (3.9246 > 3.4668) nên ta bác bỏ giả thuyết H0, chấp nhận giả thuyết có trình độ học vấn có ảnh hưởng đến mức độ hài lòng khi mua bảo hiểm nhân thọ.

* 1. Dùng R:

Các bước kiểm định ANOVA như sau:

* + 1. Import file

IS2 = read.csv("IS2.csv", header = TRUE)



* + 1. Tạo 2 biến

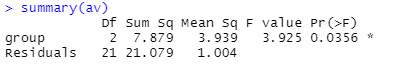
group = IS2[,1]

satisfaction = IS[,2]

* + 1. Tiến hành kiểm định phương sai ANOVA

av = aov(Satisfaction ~ group)

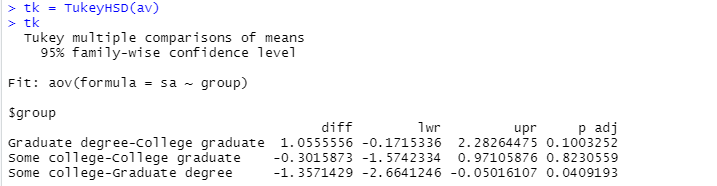
summary(av)



Bảng ANOVA giống với bảng Excel đã làm ở trên

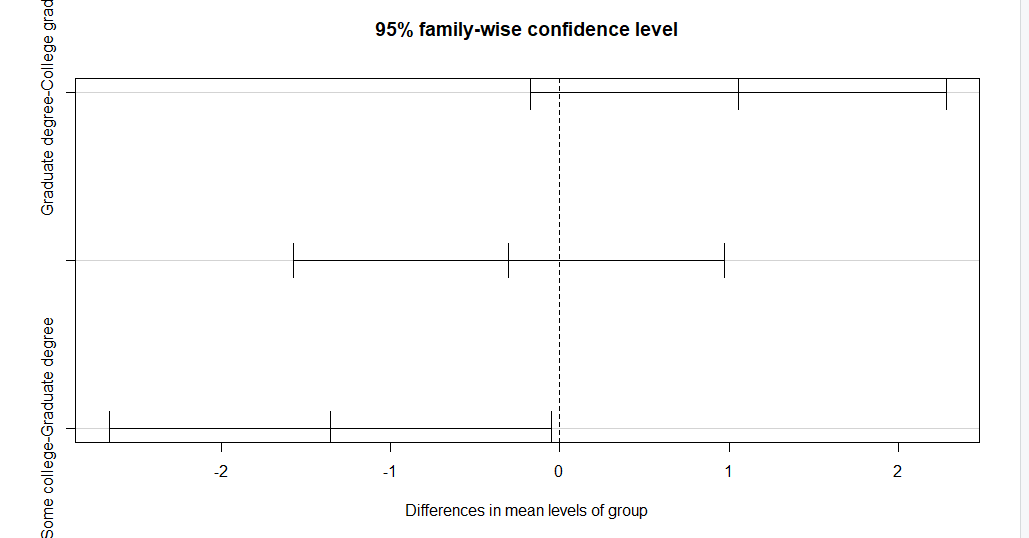
Ta cũng đưa ra kết luận F > Fcrit (3.9246 > 3.4668) nên ta bác bỏ giả thuyết H0, chấp nhận giả thuyết có trình độ học vấn có ảnh hưởng đến mức độ hài lòng khi mua bảo hiểm nhân thọ.

* + 1. Để kiếm tra sự khác nhau giữa các nhóm

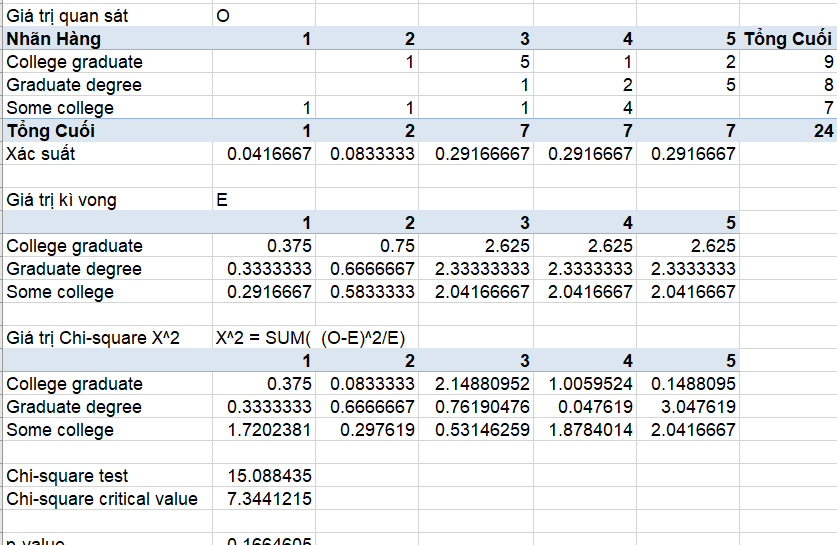


* + 1. Trực quan hơn thì bằng hình vẽ

Plot(tk)



1. Chi-square test
   1. Dùng excel



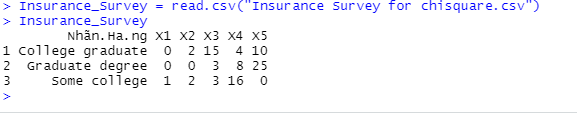
Bước 1: Ta tính xác suất theo hàng và theo cột

Bước 2: Ta tính giá trị kì vọng: lấy xác xuất của hàng \* xác xuất cột \* 89

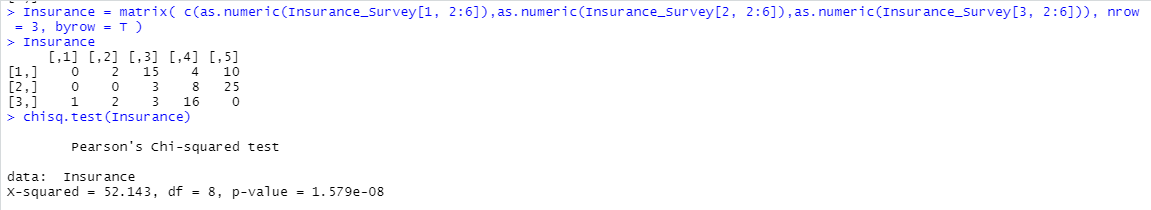
Bước 3: Ta tính giá trị chi-square theo công thức X^2 = Sum( (O-E)^2/E)

Bước 4:

* Cách 1: Ta tính giá trị chi-square critical bằng câu lệnh: = CHISQ.INV.RT(0.5,8)
* Cách 2: Ta tính p-value bằng câu lệnh: =CHISQ.TEST(vùng dữ liệu thực tế,vùng dữ liệu mong đợi)
  + Theo hướng p-value: vì p-value rất nhỏ( <0.05) nên ta bác bỏ giả thuyết H0. Điều đó có nghĩa là ta chấp nhận giả thuyết trình độ học vấn có liên quan đến mức độ hài lòng đối với chế độ bảo hiểm
  1. Dùng R
     1. Nhập dữ liệu theo file csv như hình



* + 1. Tính giá trị chi-square bằng câu lệnh chisq.test(dữ liệu)



Kết luận: Vì p-value quá nhỏ (<0.05) nên ta có thể bác bỏ giả thuyết H0. Đồng nghĩa với việc chấp nhận giả thuyết trình độ học vấn có liên quan đến độ hài lòng bảo hiểm.

II.Kiểm định Anova và chi-square test cho file Energy Drink Survey.xlxs

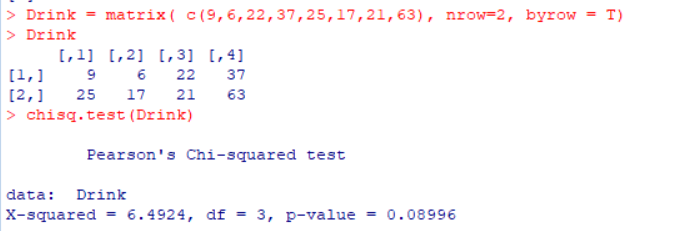
1. ANOVA

Sử dụng ANOVA một chiều khi bạn đã thu thập dữ liệu về một biến độc lập phân loại và một biến phụ thuộc định lượng . Biến độc lập phải có ít nhất ba mức (tức là ít nhất ba nhóm hoặc danh mục khác nhau).

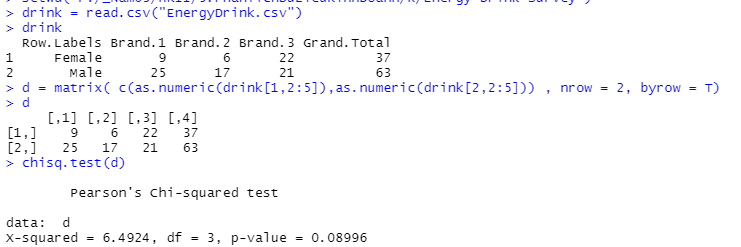
Ở đây ta có 2 biến (Genre và Brand) không thỏa các điều kiện trên nên ta không thể sử dụng phương pháp kiểm định phương sai ANOVA.

1. Chi-square test
2. Dùng R

Cách 1: Nhập dữ liệu bằng tay



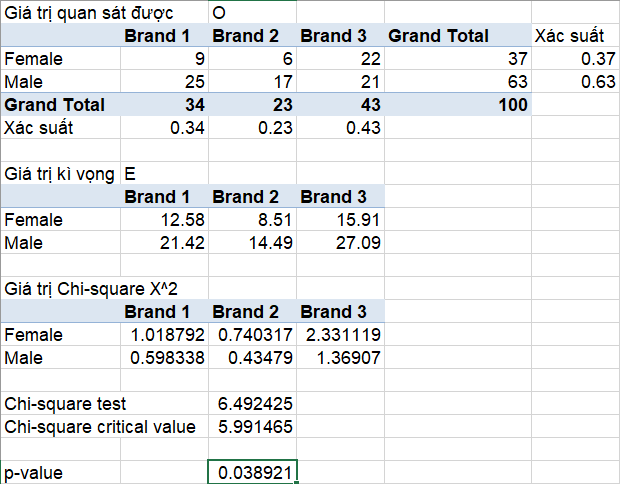
Cách 2: Đọc file csv bằng câu lệnh



Chúng ta đều có kết quả:

Giá trị p-value nhỏ hơn 0.05. Vậy bác bỏ giả thiệt H0, nghĩa là hai giá trị giới tính và hãng nước tang lực có liên quan với nhau

1. Dùng Excel



Bước 1: Ta tính xác suất theo hàng và theo cột

Bước 2: Ta tính giá trị kì vọng: lấy xác xuất của hàng \* xác xuất cột \* 100

Bước 3: Ta tính giá trị chi-square theo công thức X^2 = Sum( (O-E)^2/E)

Bước 4:

* Cách 1: Ta tính giá trị chi-square critical bằng câu lệnh: = CHISQ.INV.RT(0.5,2)
* Cách 2: Ta tính p-value bằng câu lệnh: =CHISQ.TEST(vùng dữ liệu thực tế,vùng dữ liệu mong đợi)
  + Kết luận: Giá trị p-value nhỏ hơn 0.05. Vậy bác bỏ giả thiệt H0, nghĩa là hai giá trị giới tính và hãng nước tang lực có liên quan với nhau

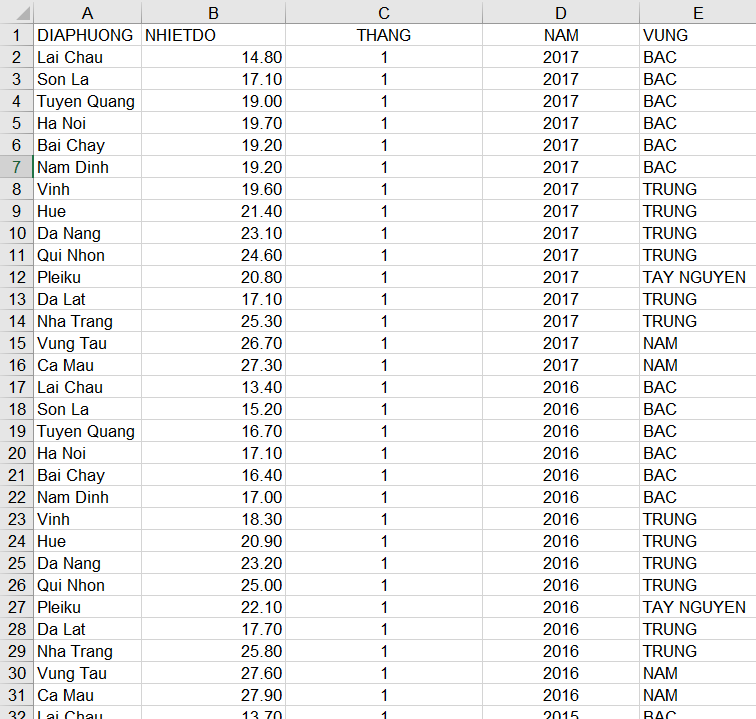
**Câu 3b:**

Tên nguồn dữ liệu: VIETNAM WEATHER TEMPERATURE 2002 2017

Chọn nguồn dữ liệu: <https://www.kaggle.com/ngvietlg/vietnam-weather-temperature-2002-2017>

Dữ liệu gốc như bên dưới

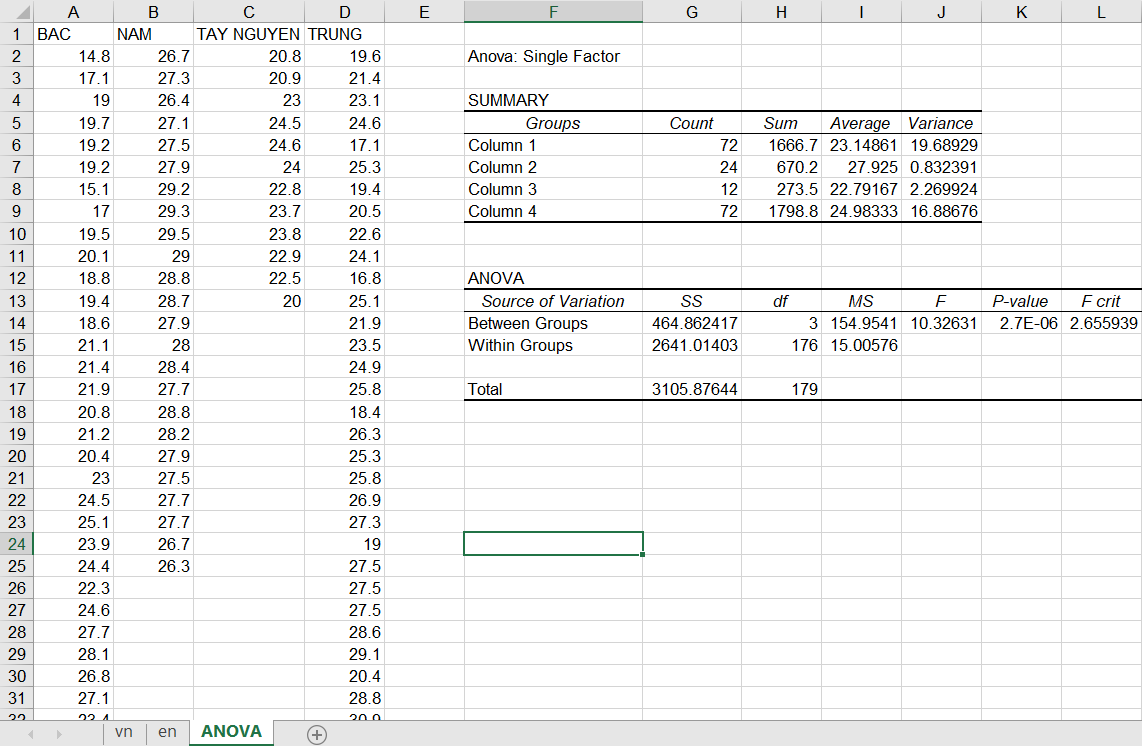
1. ANOVA



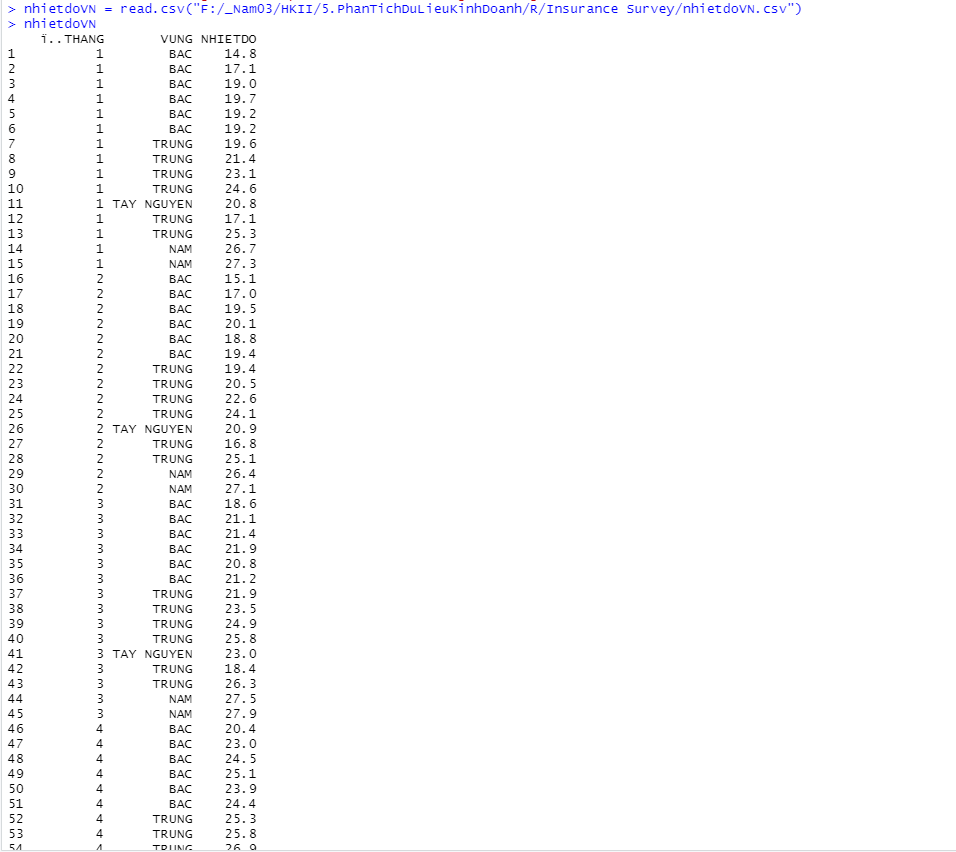
1. ANOVA

Phân tích nhiệt độ có liên quan đến các vùng của Việt Nam hay không?

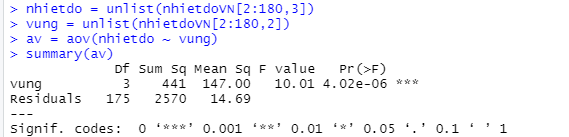
* Đặt giả thiết H0: giá trị nhiệt độ độc lập, không liên quan giữa các vùng
* Ta sắp xếp lại dữ liệu như hình và tính được ANOVA như bên dưới
* Do p-value = 2.7E-06 <0.05 nên ta bác bỏ giả thuyết H0
* Do F > F crit nên ta bác bỏ giả thuyết H0. Nghĩa là có sự liên quan, ảnh hướng của giữa các vùng với nhiệt độ ở Việt Nam



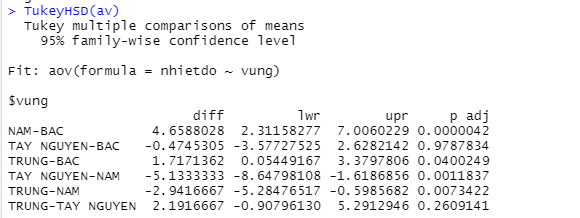
1. Kiểm chứng lại bằng phần mềm R
   1. Đọc dữ liệu trong file csv



* 1. Lấy giá trị nhiệt độ và vùng. Sau đó tính ANOVA. Ta thấy kết quả như tính bằng excel trên



* 1. Phân tích Turkey để thấy rõ sự khác biết giữa các vùng như thế nào.

****

* 1. Ta có thể thấy rõ trực quan qua hình vẽ

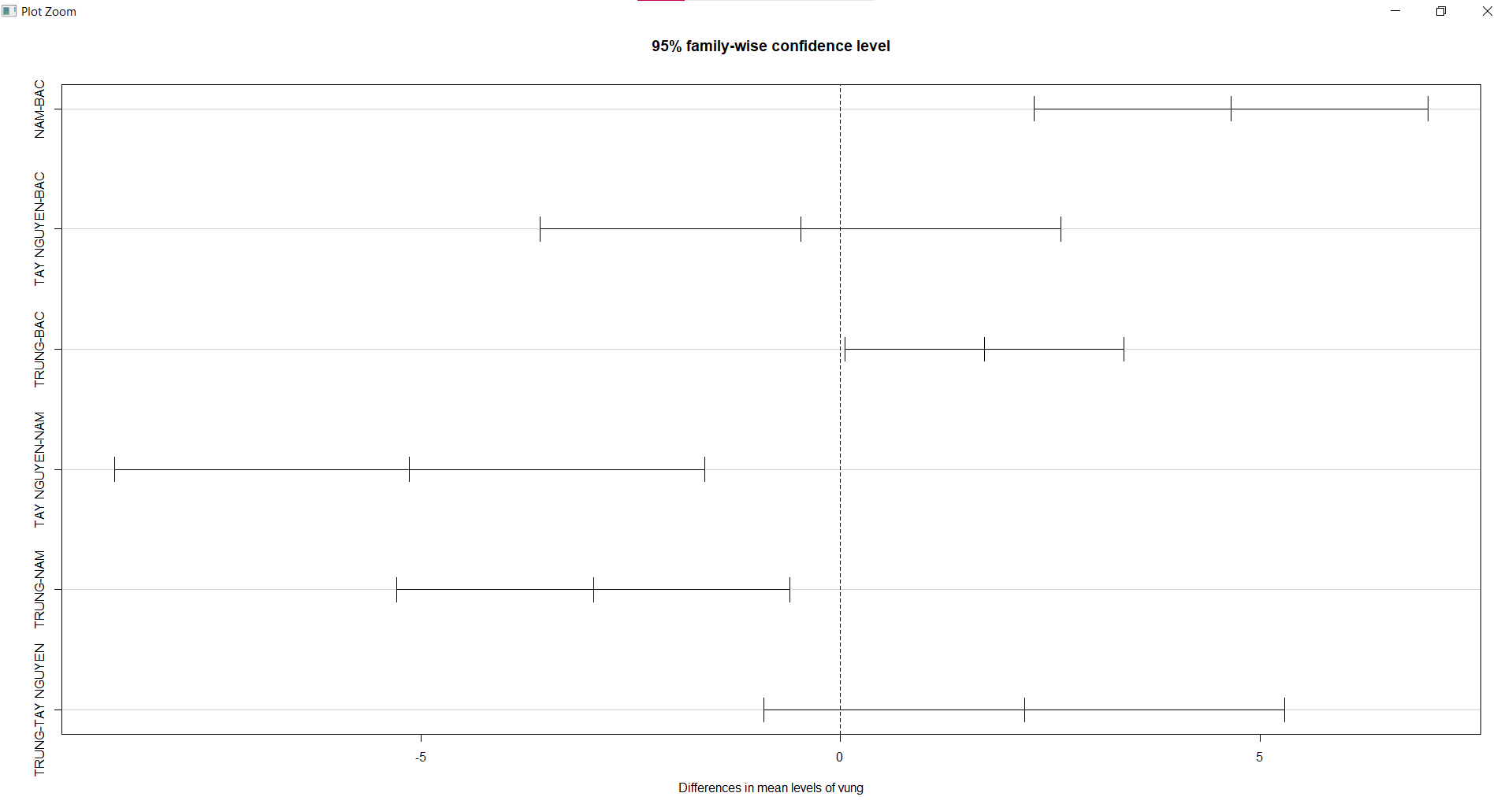


Diff: chỉ số khác biệt giữa các nhóm

Lwr: khoảng tin cậy 95% dưới

Upr: khoảng tin cậy 95% trên

P adj: p-value sau khi điều chỉnh



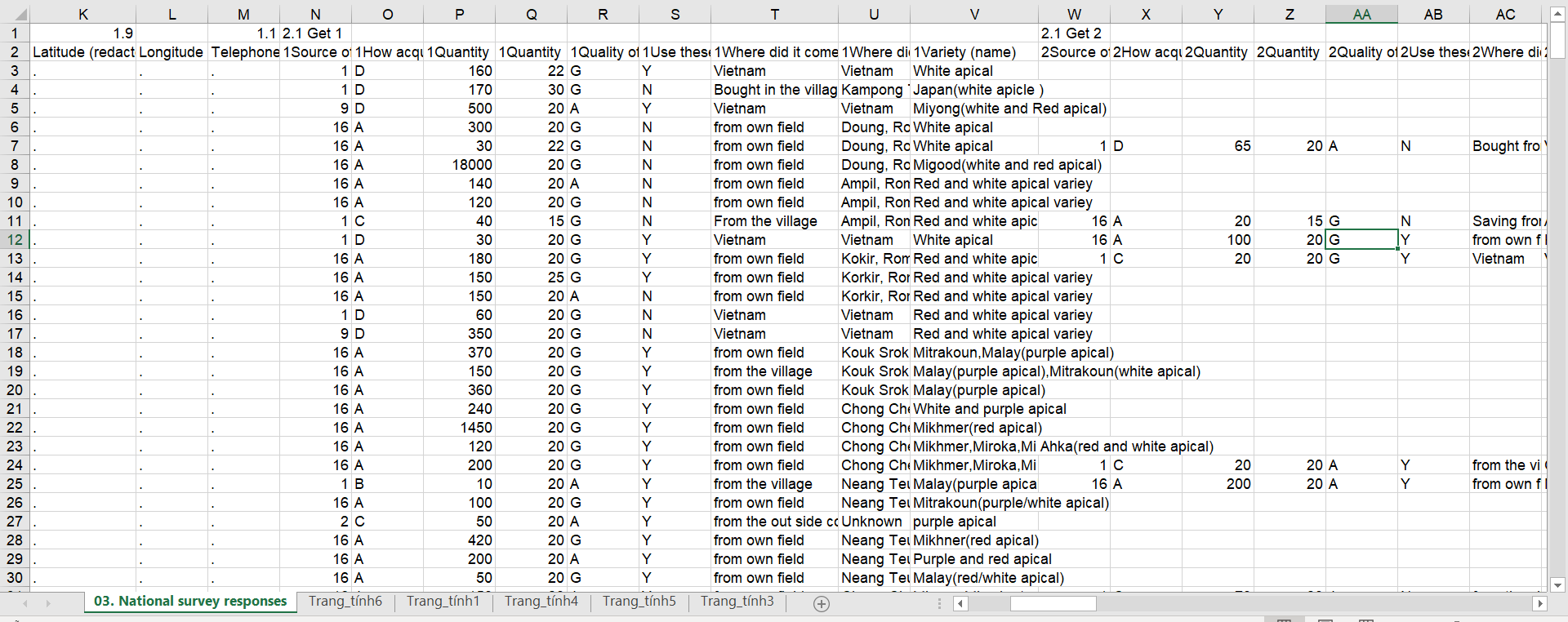
Nhận xét: nhìn biểu đồ ta thấy:

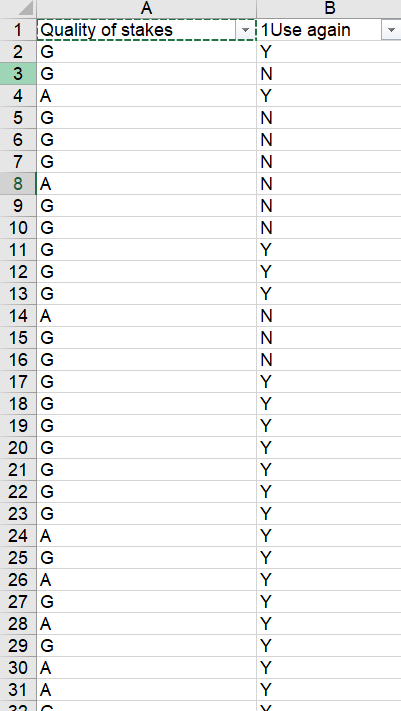
+ Sự khác biệt nam-bắc, trung-bac, tay nguyen-nam, trung-nam có ý nghĩa thống kê và sự sự khác biệt thống kê ở mực độ lần lượt 0.00000042, 0.0400249, 0.0011837, 0.0073422)

+ Sự khác biệt tay nguyen – bac và trung-tay nguyen không có ý nghĩa thống kê

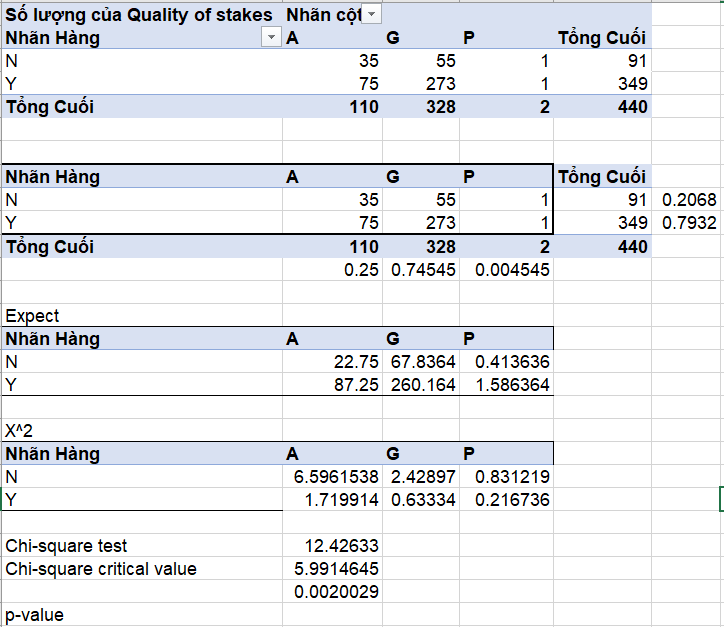
1. Chi-square

Nguồn lấy dữ liệu: Khảo sát chất lượng trồng sắn của việt nam



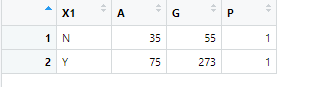
Em lấy ra 2 dữ kiện có giá trị phân loại: “Quality of stakes (G,A,P)” và “Use these again next season?” (Y/N)

Em có được kết quả chi-square như sau:

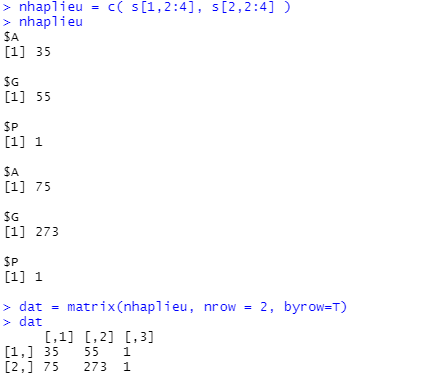


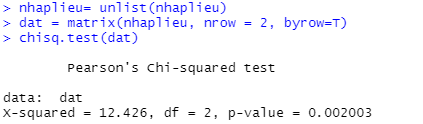
Từ kết quả, ta kết luận do p-value < 0.05 nên ta loại bỏ giả thuyết H0, nghĩa là ta chấp nhận giả thuyết chất lượng loại cọc (Quantity of stakes) ảnh hưởng đến quyết định có dùng lại vào mùa tiếp theo hay không(Use these again next season)

Dùng R ta được kết quả tương tự:









**Câu 3c**

Kết quả kiểm định gồm hai phần:

Phần 1:

Levene test: dùng kiểm định phương sai bằng nhau hay không giữa các nhó

Ho: “Phương sai bằng nhau”

p-value <= 0.05: bác bỏ Ho

p-value >0.05: chấp nhận Ho -> đủ điều kiện để phân tích tiếp anova

Phần 2:

ANOVA test: Kiểm định anova

Ho: “Trung bình bằng nhau”

p-value <=0.05: bác bỏ Ho -> đủ điều kiện để khẳng định có sự khác biệt giữa các nhóm đối với biến phụ thuộc

p-value >0.05: chấp nhận Ho -> chưa đủ điều kiện để khẳng định có sự khác biệt giữa các nhóm đối với biến phụ thuộc

Khi có sự khác biệt thì có thể phân tích sâu hơn để tìm ra sự khác biệt như thế nào giữa các nhóm quan sát bằng các kiểm định Tukey.

Vd: như bài tập file Insurance Survey

Phần 1: Levene Test

* Đặt ra giả thuyết H0: “Phương sai của 3 nhóm Some college, graduate degree và college graduate bằng nhau”
* Kiếm tra p-value = 0.0356 < 0.05
* Suy ra bác bỏ giả thuyết H0

Phần 2: Anova Test

* Do p-value < 0.05: đủ điều kiện khẳng định có sự khác biệt giữa các nhóm đối với biến satisfaction