

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**



**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**

****

**BÀI TẬP THỰC HÀNH**

**MÔN HỌC: PHÂN TÍCH DỮ LIỆU KINH DOANH**

**LAB 02**

*Giảng viên hướng dẫn:*

***PGS.TS Nguyễn Đình Thuân***

*Sinh viên thực hiện:*

|  |  |
| --- | --- |
| ***Nguyễn Hoàng Long*** | ***- 19521788*** |
| ***Phan Phạm Quỳnh Hoa***  ***Huỳnh Minh Thư*** | ***- 19521520***  ***- 19522304*** |

*TP HCM, Ngày 20 tháng 03 năm 2022*

LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành và sự tri ân sâu sắc đối với các thầy cô của trường Trường Đại học Công nghệ thông tin – Đại học Quốc gia TP.HCM, đặc biệt là quý thầy cô khoa Hệ thống thông tin của trường đã giúp cho chúng em trang bị các kiến thức cơ bản, các kỹ năng thực tế và tạo điều kiện để chúng em có thể hoàn thành đồ án môn học của mình.

Đặc biệt chúng em xin chân thành cảm ơn thầy Nguyễn Đình Thuân đã nhiệt tình hướng dẫn hướng dẫn, quan tâm truyền đạt những kiến thức và kinh nghiệm, trực tiếp hướng dẫn tận tình, sửa chữa và đóng góp ý kiến quý báu cho chúng em trong suốt thời gian học tập để chúng em có thể hoàn thành tốt môn học này.

Trong thời gian thực hiện, nhóm tác giả đã vận dụng những kiến thức nền tảng đã tích lũy. Từ đó, nhóm tác giả vận dụng tối đa những gì đã thu thập được để hoàn thành bài tập thực hành một cách tốt nhất. Tuy nhiên, trong quá trình thực hiện, nhóm tác giả không tránh khỏi những thiếu sót. Chính vì vậy, nhóm tác giả rất mong nhận được những sự góp ý từ phía Thầy/Cô nhằm hoàn thiện những kiến thức mà nhóm tác giả đã học tập và là hành trang để nhóm tác giả thực hiện tiếp các đề tài khác trong tương lai.

Nhóm chúng em xin chân thành cảm ơn!

***Nhóm thực hiện***

**NHẬN XÉT CỦA GIẢNG VIÊN**

………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

**MỤC LỤC**

[I. Thực hiện bài tập 1(a, b) 6](#_Toc99052337)

[1. Kiểm định Levene và Turkey 6](#_Toc99052338)

[2. Insurance Survey 7](#_Toc99052339)

[Phát biểu bài toán 7](#_Toc99052340)

[2.1 Sử dụng Excel 7](#_Toc99052341)

[2.2 Sử dụng R 14](#_Toc99052342)

[2.3 Sử dụng Python 16](#_Toc99052343)

[3. Energy Drink Survey 17](#_Toc99052344)

[Phát biểu bài toán 17](#_Toc99052345)

[3.1 Sử dụng Excel 17](#_Toc99052346)

[3.2 Sử dụng R 20](#_Toc99052347)

[3.3 Sử dụng Python 21](#_Toc99052348)

[4. VIETNAM NATIONAL HIGHSCHOOL EXAM SCORE 2018 21](#_Toc99052349)

[Phát biểu bài toán 21](#_Toc99052350)

[4.1 Sử dụng Excel 21](#_Toc99052351)

[4.2 Sử dụng R 24](#_Toc99052352)

[4.3 Sử dụng Python 26](#_Toc99052353)

[5. TITANIC 26](#_Toc99052354)

[Phát biểu bài toán 26](#_Toc99052355)

[5.1 Sử dụng Excel 27](#_Toc99052356)

[5.2 Sử dụng R 29](#_Toc99052357)

[5.3 Sử dụng Python 30](#_Toc99052358)

[6. VN\_housing\_dataset 30](#_Toc99052359)

[Phát biểu bài toán 30](#_Toc99052360)

[6.1 Sử dụng Excel 31](#_Toc99052361)

[6.2 Sử dụng R 32](#_Toc99052362)

[6.3 Sử dụng Python 34](#_Toc99052363)

[II. Bảng phân công công việc 35](#_Toc99052364)

[III. Tài liệu tham khảo 35](#_Toc99052365)

1. **Thực hiện bài tập 1(a, b)**
2. **Kiểm định Levene và Turkey**

***Hàm kiểm định Levene*** (Tiền xử lý kiểm định ANOVA): dùng kiểm định phương sai bằng nhau hay không giữa các nhóm

Nếu chấp nhận giả thuyết H0 -> đủ điều kiện để phân tích tiếp ANOVA.

***Hàm kiểm định ANOVA*** (hàm phân tích phương sai) ***:***Là kỹ thuật thống kê tham số được sử dụng để phân tích sự khác nhau giữa giá trị trung bình của các biến phụ thuộc với nhau.

*So sánh F > F (k-1;n-k) =>* ***Bác bỏ giả thuyết H0***

Trong đó:

MSW : phương sai do các yếu tố khác tạo ra

MSG : phương sai do yếu tố nghiên cứu tạo ra

**Hàm kiểm định Turkey** (Hàm phân tích ANOVA sâu): dùng để phân tích về sự hơn kém giữa các giá trị trung bình trong trường hợp bác bỏ giả thuyết H0.

Các bước kiểm định Turkey:

***Bước 1:*** Tính khoảng biến thiên trung bình giữa hai nhóm:

= | - |

***Bước 2:*** Tính chỉ số Turkey

T = (k, n-k)

***Bước 3:*** Bác hỏ H0 nếu >T

1. **Insurance Survey**

**Phát biểu bài toán**

Kiểm định ANOVA với tập tin dữ liệu (gồm Levene, ANOVA, Tukey Test): Insurance Survey.

Có thể cho rằng có sự khác nhau giữa từng nhóm tình trạng hôn nhân đối với mức độ hài lòng bảo hiểm là đúng hay không?

Với mức ý nghĩa α = 0.05.

* Giả thiết H0: µ1=µ2= µ3 (tức là trình độ học vấn không ảnh hưởng đến mức độ hài lòng bảo hiểm).
* Đối thiết H1: trình độ học vấn ảnh hưởng đến mức độ hài lòng bảo hiểm.
  1. **Sử dụng Excel**

Từ tập tin dữ liệu Insurance Survey, ta lọc lấy hai cột dữ liệu là Education và Satisfaction.

A picture containing table

Description automatically generated

Sau đó lọc dữ liệu thành 3 nhóm dựa trên Education như hình

Table

Description automatically generated

* Kiểm định Levene
* Dùng hàm AVERAGE tính giá trị trung bình của từng nhóm con

Table

Description automatically generated

* Tính giá trị Zij = | Yij - i |

Với i: là giá trị trung bình của nhóm con thứ i.

Thực hiện trong Excel:

Graphical user interface, application, table, Excel

Description automatically generated

Ta được kết quả như hình:

Table

Description automatically generated

* Vào Data -> Data Analysis -> chọn chức năng Anova: Single Factor

Graphical user interface, text, application

Description automatically generated

* Tại hộp thoại Anova: Single Factor:

**Input Range:** Vùng chứa các giá trị đầu vào

**Output Range:** Điểm đầu của kết quả cần hiển thị

Graphical user interface

Description automatically generated

Sau khi chọn các giá trị tương ứng ta thu được bảng kết quả:

Table

Description automatically generated

Giá trị P-value = 0.405206 > 0.05 chấp nhận giả thiết H0

Đủ điều kiện để phân tích tiếp ANOVA.

* Kiểm định ANOVA

Thực hiện kiểm định ANOVA trên dữ liệu ban đầu đã tách:

Table

Description automatically generated

* Vào Data -> Data Analysis -> chọn chức năng Anova: Single Factor

Tương tự phần kiểm định Levene, ta thu được bảng kết quả:

Table, Excel

Description automatically generatedGiá trị P-value = 0.0356353975 < 0.05 nên ta bác bỏ giả thuyết H0.

* Kiểm định Tukey
* Khoảng biến thiên trung bình giữa mỗi nhóm:

Dij = | i - j |

Table

Description automatically generated

* Tính chỉ số Tukey áp dụng công thức:

T = qα (k, n-k)

Với q phân phối Tukey tra từ bảng sau:

Table

Description automatically generated

* Ta tính được giá trị T = 1.355667305

Table

Description automatically generated

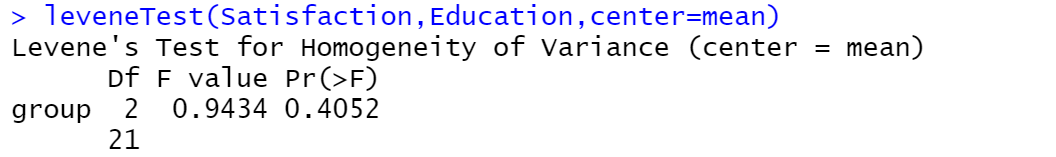
Từ đây ta thấy được:

* College graduate - Graduate degree D12 < T nên phương sai giữa hai nhóm bằng nhau.
* College graduate - Some college D13 < T nên phương sai giữa hai nhóm bằng nhau.
* Graduate degree - Some college D23 > T nên phương sai giữa hai nhóm không bằng nhau.
  1. **Sử dụng R**

Dùng ngôn ngữ R về ANOVA với tập tin dữ liệu (gồm Levene, ANOVA, Tukey Test)

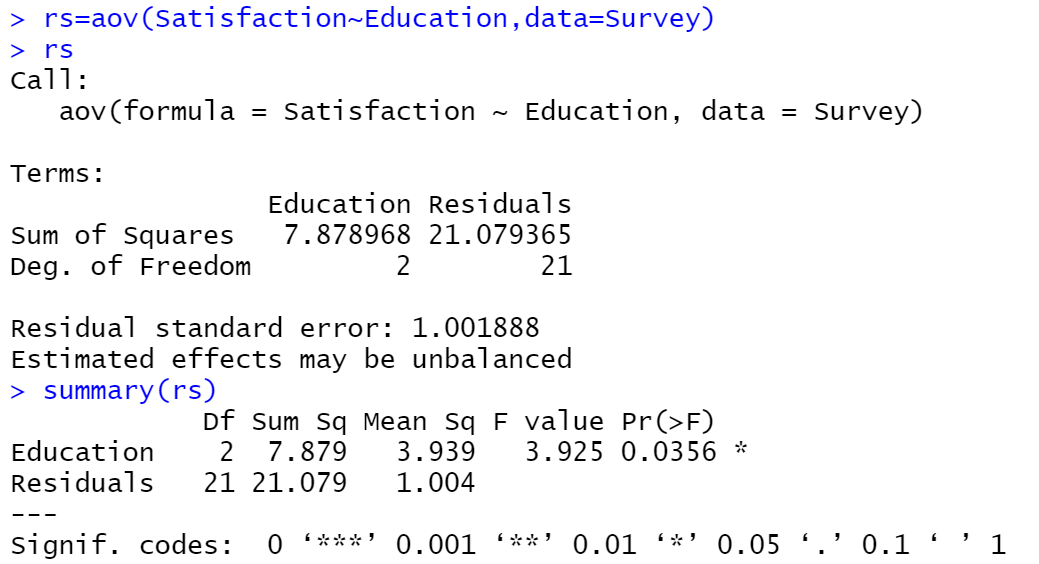
* **Hàm kiểm định Levene:** Sử dụng hàm ***leveneTest(value, group, center=mean),*** cần cài đặt package car và require để sử dụng hàm.

***Kết quả:***

****

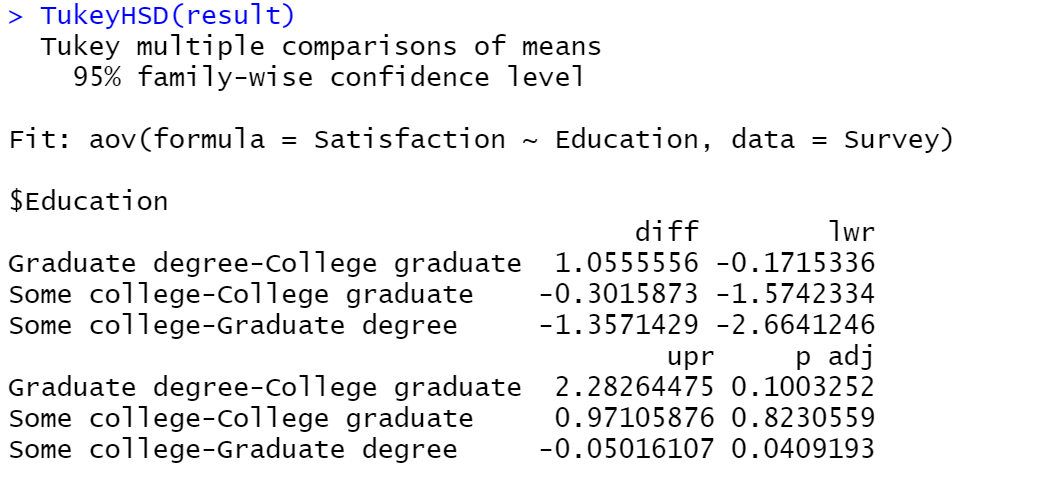
* Giá trị Pr = 0.4052 > 0.05, vì vậy ta chấp nhận giả thuyết H0 và kiểm định ANOVA.
* ***Hàm kiểm định ANOVA:*** Sử dụng hàm ***aov (value ~ group, data = data)***, cần cài đặt package car và require để sử dụng hàm.

***Kết quả:***



* Kết luận: Vì Pr = 0.0356 < 0.05 nên bác bỏ giả thuyết H0. Vậy có ít nhất một cặp có giá trị khác nhau.
* **Hàm kiểm định Turkey:** Sử dụng hàm ***TukeyHSD (result of ANOVA)***, cần cài đặt package car và require để sử dụng hàm.

***Kết quả:***



* Kết luận cặp Some college – Graduate degree có giá trị khác nhau, vì p = 0.0409193 < 0.05.
  1. **Sử dụng Python**

**Kiểm định Leven**

**Bước 1:** Chia nhóm



**Bước 2:** Sử dụng hàm Levene để kiểm định Levene

Ảnh có chứa văn bản

Mô tả được tạo tự động

Ta thấy, p-value >= 0.05 -> tiến hành kiểm định ANOVA

**Kiểm định ANOVA**

**Bước 1:** Ta sử dụng hàm f\_oneway để kiểm định ANOVA

Ảnh có chứa văn bản

Mô tả được tạo tự động

Vì Pr = 0.0356 < 0.05 nên bác bỏ giả thuyết H0. Vậy có ít nhất một cặp có giá trị khác nhau

**Kiểm định Turkey**

**Bước 1:** Ta sử dụng hàm pairwise\_turkey để kiểm định ANOVA

Ảnh có chứa bàn

Mô tả được tạo tự động

Kết luận cặp Some college – Graduate degree có giá trị khác nhau, vì p = 0.04 < 0.05

1. **Energy Drink Survey**

**Phát biểu bài toán**

Hãy cho biết có sự liên quan giữa giới tính và thương hiệu ưa thích hay không?

Với Mức ý nghĩa α = 0.05

* Giả thiết H0: giới tính và thương hiệu ưa thích độc lập với nhau.
* Đối thiết H1: giới tính và thương hiệu ưa thích phụ thuộc nhau.
  1. **Sử dụng Excel**

Kiểm định Chi Square về tính độc lập của 2 biến định danh với tập tin dữ liệu Energy Drink Survey

Vào Insert -> PivotTable để tạo bảng Count of Respondent từ cột Gender và Brand Preference

Table

Description automatically generated

Tính toán giá trị cho bảng expected frequency theo công thức:

fe của dòng i cột j =

Graphical user interface, application, table, Excel

Description automatically generated

Ta được bảng expected frequency như hình:

Table

Description automatically generated

Tính toán giá trị cho bảng Chi Square Statistic theo công thức:

**Table

Description automatically generated**

Ta được bảng Chi Square Statistic như hình:

**Table

Description automatically generated**

* **χ2 = 6.492**

Sử dụng hàm CHISQ.INV.RT(α, df) để tính giá trị Critical Value

Với α: mức ý nghĩa bằng 0.05

df: bậc tự do

Ta có: df = (r - 1)(c-1) = (2 - 1)(3 - 1) = 2

A picture containing table

Description automatically generated

* Giá trị **Critical Value = 5.9915**
* **χ2 > Critical Value.** Vậy ta từ chối H0
  1. **Sử dụng R**

Dùng ngôn ngữ R kiểm định Chi Square về tính độc lập của 2 biến định danh với tập tin dữ liệu Energy Drink Survey

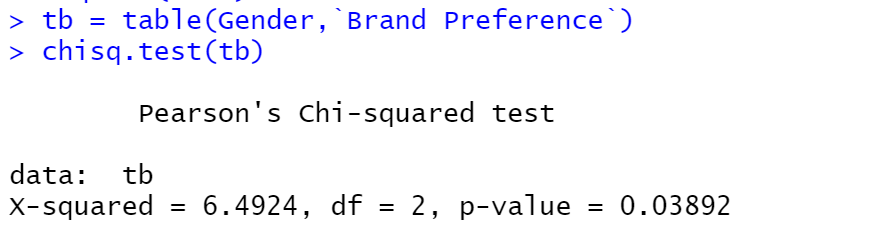
* **Kiểm định Chi Square:** kiểm định sự độc lập/phụ thuộc của hai biến dạng phân loại.
* Phát biểu bài toán

***H0:*** Hai biến phân loại là *độc lập*

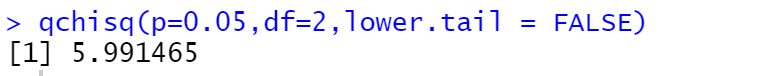
***H1:*** Hai biến phân loại là *phụ thuộc*

Sử dụng hàm **table (***group1, group2) và* **chisq.test(** *tb* **),** cần cài đặt package *MASS* và require để sử dụng hàm.

***Kết quả:***

****

* Sử dụng hàm ***qchisq*** để tính Critical-value cho bảng dữ liệu:

****

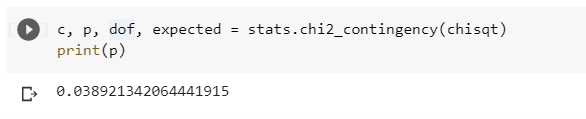
* + Bác bỏ giả thuyết H0 vì P = 0.03892 < 0.05
  1. **Sử dụng Python**

**Bước 1:** Sử dụng hàm crosstab để thống kê bảng nhóm

Ảnh có chứa bàn

Mô tả được tạo tự động

**Bước 2:** Sử dụng hàm chi2\_contingency để kiểm định chi-square



Ta từ chối Ho do p-value < 0.05, vậy 2 biến phân loại là độc lập

1. **VIETNAM NATIONAL HIGHSCHOOL EXAM SCORE 2018**

**Phát biểu bài toán**

* 1. **Sử dụng Excel**

Từ tập tin dữ liệu VIETNAM NATIONAL HIGHSCHOOL EXAM SCORE 2018, ta lọc lấy ba cột dữ liệu là Math, Viet và English

**Chart

Description automatically generated with low confidence**

Kiểm định Levene

* Dùng hàm AVERAGE tính giá trị trung bình của từng nhóm con

**Table

Description automatically generated**

Tương tự phần 2.1, lần lượt ta tính được các giá trị Zij

A screenshot of a computer

Description automatically generated with low confidence

Sử dụng chức năng Anova: Single Factor ta được kết quả như hình:

Application, table, Excel

Description automatically generated

Giá trị P-value = 0 < 0.05 bác bỏ giả thiết H0.

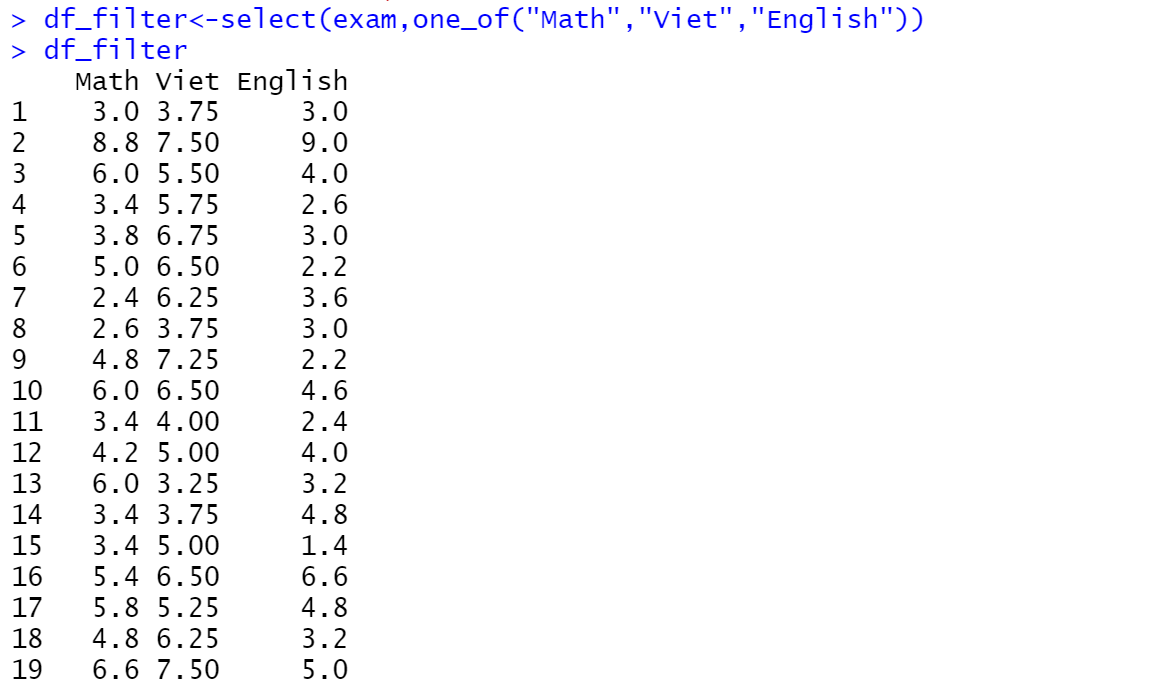
Không đủ điều kiện để phân tích tiếp ANOVA.

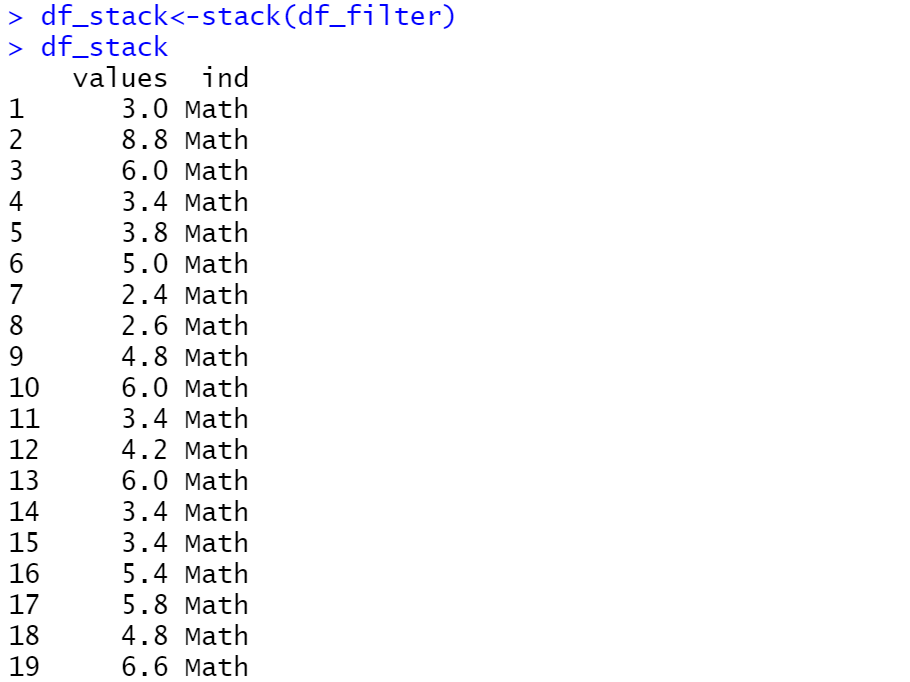
* 1. **Sử dụng R**

Dùng ngôn ngữ R về ANOVA với tập tin dữ liệu (gồm Levene, ANOVA, Tukey Test)

* **Xử lý dữ liệu:** dùng hàm stack để gom nhóm 3 môn học Math, Van, English, sau đó drop các dòng có giá trị NA, cần cài đặt package DataCombine và require để sử dụng hàm.

***Kết quả:***

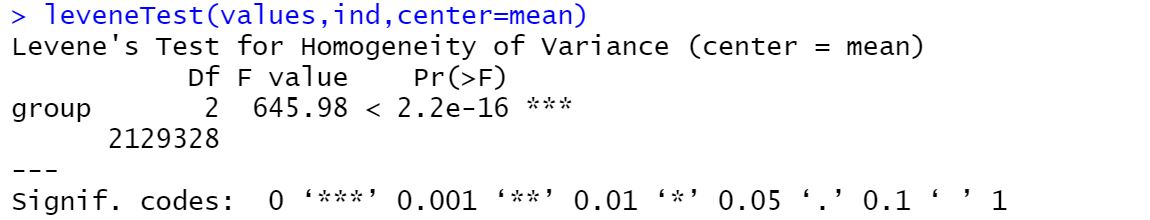
******

******

Text

Description automatically generated

* **Hàm kiểm định Levene:** Sử dụng hàm ***leveneTest(value, group, center=mean),*** cần cài đặt package car và require để sử dụng hàm.

****

* + Vì giá trị Pr < 0.05 nên bác bỏ H0.
  + Không đủ điều kiện để thực hiện kiểm định ANOVA.
  1. **Sử dụng Python**

**Kiểm định Levene**

**Bước 1:** Chọn ra 3 group

Ảnh có chứa văn bản

Mô tả được tạo tự động

**Bước 2:** Thực hiện kiểm định với hàm levene

Ảnh có chứa văn bản

Mô tả được tạo tự động

Ta thấy p-value < 0.05, bác bỏ Ho -> Không đủ điều kiện để kiểm định ANOVA

1. **TITANIC**

**Phát biểu bài toán**

Hãy cho biết có sự liên quan giữa cảng lên tàu và sự sống sót trên tàu hay không?

Với Mức ý nghĩa α = 0.05

* Giả thiết H0: Embarked và Survived độc lập với nhau.
* Đối thiết H1: Embarked và Survived phụ thuộc nhau.
  1. **Sử dụng Excel**
* Kiểm định Chi Square về tính độc lập của 2 biến định danh Embarked – Survived với tập tin dữ liệu Titanic

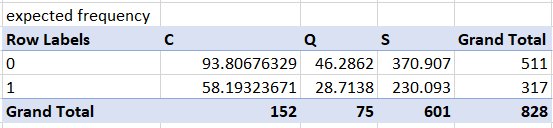
Vào Insert -> PivotTable để tạo bảng Count of Embarked từ cột Embarked và Survived

Table

Description automatically generated

Tương tự mục 3.1 ta được:

Bảng expected frequency:



Bảng Chi Square Statistic:

Table

Description automatically generated

* **χ2 = 20.308**

Sử dụng hàm CHISQ.INV.RT(α, df) để tính giá trị Critical Value

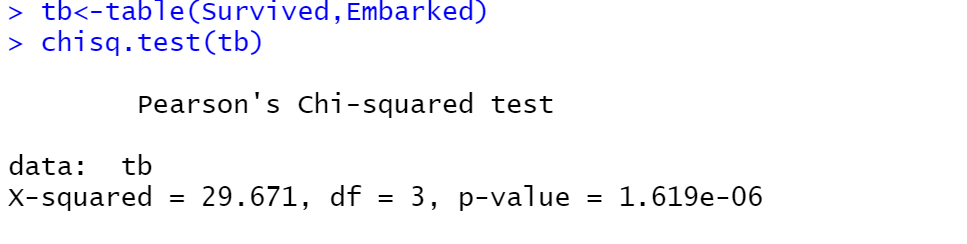
Ta có: df = (r - 1)(c-1) = (2 - 1)(3 - 1) = 2

Text

Description automatically generated with low confidence

* Giá trị **Critical Value = 5.99**
* **χ2 > Critical Value.** Vậy ta từ chối H0
  1. **Sử dụng R**
* **Kiểm định Chi Square:** kiểm định sự độc lập/phụ thuộc của hai biến dạng phân loại.
* Sử dụng hàm **table (***group1, group2) và* **chisq.test(** *tb* **),** cần cài đặt package *MASS* và require để sử dụng hàm.

***Kết quả*** với nhóm Embarked – Survived:

****

* Sử dụng hàm ***qchisq*** để tính Critical-value cho bảng dữ liệu:

****

* Bác bỏ giả thuyết H0 vì p < 0.05.
  1. **Sử dụng Python**

**Bước 1:** Sử dụng hàm crosstab để thống kê bảng nhóm

Ảnh có chứa văn bản

Mô tả được tạo tự động

**Bước 2:** Sử dụng hàm chi2\_contingency để kiểm định chi-square

Ảnh có chứa văn bản

Mô tả được tạo tự động

Ta từ chối Ho do p-value < 0.05, vậy 2 biến phân loại là độc lập

1. **VN\_housing\_dataset**

**Phát biểu bài toán**

* 1. **Kiểm định ANOVA**

Có thể cho rằng có sự khác nhau giữa từng Huyện đối với giá nhà là đúng hay không?

Với mức ý nghĩa α = 0.05.

* Giả thiết H0: µ1=µ2= µ3 (tức là Huyện không ảnh hưởng đến giá nhà).
* Đối thiết H1: Huyện ảnh hưởng đến giá nhà.
  1. **Kiểm định Chi Square**

Hãy cho biết có sự liên quan giữa Loại hình nhà ở và Giấy tờ pháp lý hay không?

Với Mức ý nghĩa α = 0.05

* Giả thiết H0: Loại hình nhà ở và Giấy tờ pháp lý độc lập với nhau.
* Đối thiết H1: Loại hình nhà ở và Giấy tờ pháp lý phụ thuộc nhau.
  1. **Sử dụng Excel**
  2. **Kiểm định ANOVA**

Từ tập tin dữ liệu VN\_housing\_dataset, ta lọc lấy 310 cột dữ liệu là tên Huyện.

Vào Data -> Data Analysis -> chọn chức năng Anova: Single Factor

Tương tự phần 2.1 kiểm định ANOVA, ta thu được kết quả:

Table

Description automatically generated

Giá trị P-value 1 > 0.05 nên ta chấp nhận giả thuyết H0.

* 1. **Kiểm định Chi Square**

Kiểm định Chi Square về tính độc lập của 2 biến định danh Loại hình nhà ở và Giấy tờ pháp lý với tập tin dữ liệu VN\_housing\_dataset

Vào Insert -> PivotTable để tạo bảng Count of Loại hình nhà ở từ cột Loại hình nhà ở và Giấy tờ pháp lý

**Table

Description automatically generated**

Tương tự mục 3.1 ta được:

Bảng expected frequency:

Graphical user interface, table, Excel

Description automatically generated

Bảng Chi Square Statistic:

Table, Excel

Description automatically generated

* **χ2 = 815.059**

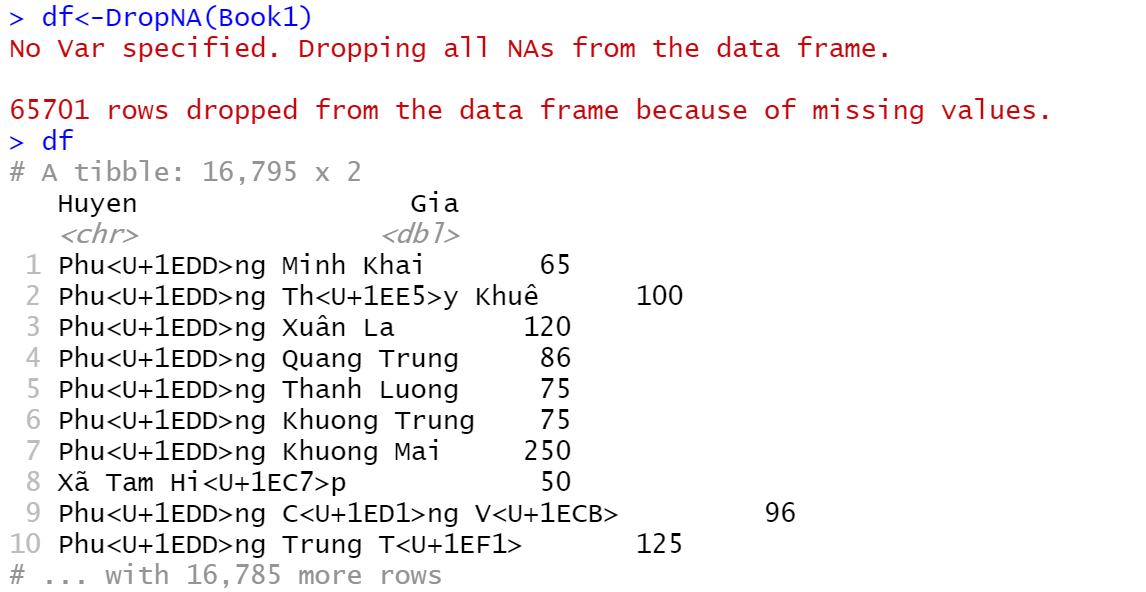
Sử dụng hàm CHISQ.INV.RT(α, df) để tính giá trị Critical Value

Ta có: df = (r - 1)(c-1) = (5 - 1)(4 - 1) = 12

Table

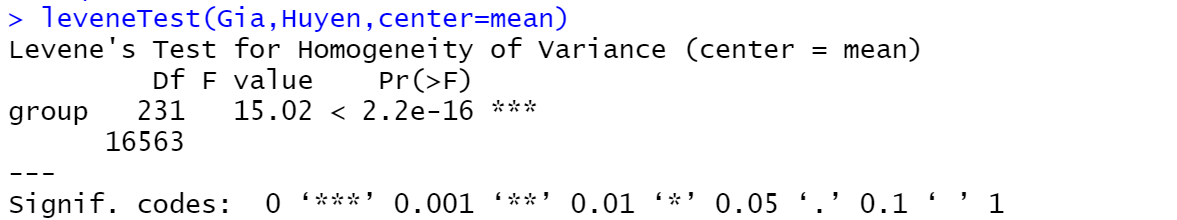
Description automatically generated with low confidence

* Giá trị **Critical Value = 21.026**
* **χ2 > Critical Value.** Vậy ta từ chối H0
  1. **Sử dụng R**
* **Xử lý dữ liệu:** Sử dụng hàm ***DropNA*** để drop các dòng có giá trị NA, cần cài đặt package *DataCombine* và require để sử dụng.

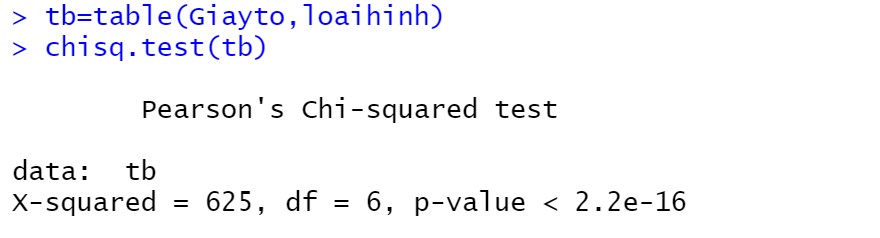


* **Hàm kiểm định Levene:** Sử dụng hàm ***leveneTest(value, group, center=mean),*** cần cài đặt package *car* và require để sử dụng hàm.

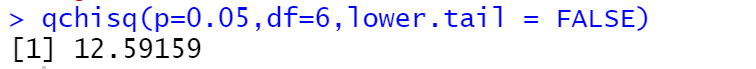
**Kết quả:**



* + Vì giá trị Pr < 0.05 nên bác bỏ H0.
  + Không đủ điều kiện để thực hiện kiểm định ANOVA.
* **Kiểm định Chi Square:** kiểm định sự độc lập/phụ thuộc của hai biến dạng phân loại.
* Sử dụng hàm **table (***group1, group2) và* **chisq.test(** *tb* **),** cần cài đặt package *MASS* và require để sử dụng hàm.



* Sử dụng hàm ***qchisq*** để tính Critical-value cho bảng dữ liệu:



* Bác bỏ H0 vì p < 0.05.
  1. **Sử dụng Python**

**Kiểm định Levene**

Ảnh có chứa văn bản

Mô tả được tạo tự động

Ta thấy p-value < 0.05, bác bỏ Ho -> Không đủ điều kiện để kiểm định ANOVA

**Kiểm định Chi-square**

Ảnh có chứa bàn

Mô tả được tạo tự động

Ta từ chối Ho do p-value < 0.05, vậy 2 biến phân loại là độc lập

1. **Bảng phân công công việc**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Nguyễn Hoàng Long | Phan Phạm Quỳnh Hoa | Huỳnh Minh Thư |
| Xây dựng Template báo cáo | X |  |  |
| Tìm hiểu và sử dụng Excel |  |  | X |
| Tìm hiểu và sử dụng R |  | X |  |
| Tìm hiểu và sử dụng Python | X |  |  |
| Làm bài tập 2 a, b,c,d,e,f,g (Sử dụng Excel) |  |  | X |
| Làm bài tập 2 a, b,c,d,e,f,g (Sử dụng R) |  | X |  |
| Làm bài tập 2 a, b,c,d,e,f,g (Sử dụng Python) | X |  |  |

1. **Tài liệu tham khảo**

[1] <https://nguyenvantien0405.files.wordpress.com/2018/03/hd_excel_xstk_161223.pdf>

[2] <https://corporatefinanceinstitute.com/resources/excel/functions/>

[3] <https://www.statology.org/tukey-kramer-post-hoc-test-excel/>

[4] <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html>

[5] <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html>