**ĐẠI HỌC HUẾ**

# KHOA KỸ THUẬT VÀ CÔNG NGHỆ



🙠🙟🕮🙝🙢

**Icon

Description automatically generated**

**BÁO CÁO**

**ĐỒ ÁN**

**NĂM HỌC 2020-2021**

**Giảng viên hướng dẫn: Nguyễn Đình Hoa Cương**

**Lớp: K1- Khoa học dữ liệu và Trí tuệ nhân tạo**

|  |
| --- |
| Số phách  *(Do hội đồng chấm thi ghi)* |

**Thừa Thiên Huế, ngày 26 tháng 6 năm 2021**

**ĐẠI HỌC HUẾ**

# KHOA KỸ THUẬT VÀ CÔNG NGHỆ



🙠🙟🕮🙝🙢

**Icon

Description automatically generated**

(MẪU BÌA PHỤ)

**BÁO CÁO**

**ĐỒ ÁN**

**NĂM HỌC 2020-2021**

**Giảng viên hướng dẫn: Nguyễn Đình Hoa Cương**

**Lớp: K1- Khoa học dữ liệu và Trí tuệ nhân tạo**

**Sinh viên thực hiện: Phan Văn Duy Phúc**

*(ký tên và ghi rõ họ tên)*

|  |
| --- |
| Số phách  *(Do hội đồng chấm thi ghi)* |

**Thừa Thiên Huế, ngày 26 tháng 6 năm 2021**

MỤC LỤC

[KHOA KỸ THUẬT VÀ CÔNG NGHỆ 1](#_Toc75702640)

[KHOA KỸ THUẬT VÀ CÔNG NGHỆ 2](#_Toc75702641)

[**I.** **KHÁI QUÁT VỀ THỐNG KÊ MÔ TẢ** 5](#_Toc75702642)

[**II.** **PHẦN PHÂN TÍCH THỐNG KÊ MÔ TẢ VIẾT CÁC MÔ TẢ VỀ TẬP DỮ LIỆU INC\_EXP\_DATA** 5](#_Toc75702643)

[**III.** **KHÁI QUÁT VỀ THỐNG KÊ SUY DIỄN** 8](#_Toc75702644)

[**IV.** **PHẦN PHÂN TÍCH THỐNG KÊ SUY DIỄN, PHÂN TÍCH ANOVA MỘT CHIỀU CỦA TẬP TIN DIET\_R.CSV** 9](#_Toc75702645)

[**1.** **Khái quát về file Diet\_R.csv:** 9](#_Toc75702646)

[**2.** **Bài toán 1: CÓ SỰ KHÁC NHAU VỀ HIỆU QUẢ CỦA CÁC PHƯƠNG PHÁP GIẢM CÂN?** 10](#_Toc75702647)

[**Tiến hành tìm bảng Anova:** 11](#_Toc75702648)

[**Kiểm tra hậu nghiệm:** 11](#_Toc75702649)

[**Kiểm tra các điều kiện cho kiểm định Anova** 11](#_Toc75702650)

[**3.** **Bài toán 2:CÓ SỰ KHÁC NHAU VỀ HIỆU QUẢ GIẢM CÂN VỚI CÁC GIỚI TÍNH KHÁC NHAU?** 13](#_Toc75702651)

[**Phát biểu giải thuyết thống kê** 13](#_Toc75702652)

[**Tiến hành tìm bảng Anova:** 13](#_Toc75702653)

[**Kiểm tra các điều kiện cho kiểm định Anova:** 14](#_Toc75702654)

[Tài liệu trích dẫn: 15](#_Toc75702655)

[Kết quả kiểm tra đạo văn: 15](#_Toc75702656)

1. **KHÁI QUÁT VỀ THỐNG KÊ MÔ TẢ**

Thống kê mô tả được dùng để mô tả những đặc tính cơ bản của dữ liệu thu thập được từ các nghiên cứu qua các phương pháp khác nhau.

Khi tạo các trị thống kê mô tả, người ta có thể nhắm đến 2 mục tiêu:

* Chọn một giá trị thống kê để chỉ ra những đơn vị có vẻ giống nhau thực ra có thể khác nhau thế nào. Các giáo trình thống kê gọi một giải pháp đáp ứng mục tiêu này là thước đo khuynh hướng trung tâm.
* Chọn một trị thống kê khác cho thấy các đơn vị khác nhau thế nào. Loại trị thống kê này thường được gọi là một thước đo phân tán thống kê.

Khi tóm tắt một lượng như độ dài, cân nặng hay tuổi tác, người ta thường sử dụng các trị thống kê như số trung bình cộng, trung vị; hay trong trường hợp một phân bố đơn mode (số trung phương), người ta thường dùng mode.

Các thước đo chung nhất về mức độ phân tán của dữ liệu lượng là phương sai, giá trị căn bậc 2, tức là độ lệch chuẩn; khoảng; khoảng cách giữa các tứ phân vị; và độ lệch bình quân tuyệt đối.(1)

1. **PHẦN PHÂN TÍCH THỐNG KÊ MÔ TẢ VIẾT CÁC MÔ TẢ VỀ TẬP DỮ LIỆU INC\_EXP\_DATA**

Tập dữ liệu Inc\_Exp\_Data có 7 loại dữ liệu, gồn có 4 loại thuộc kiểu dữ liệu định lượng và 3 loại dữ liệu thuộc kiểu dữ liệu đinh tính.

Về dữ liệu định lượng gồm: ***Mthly\_HH\_Income, Mthly\_HH\_Expense, Emi\_or\_Rent\_Amt, Annual\_HH\_Income.***

Để thực hiện thống kê mô tả cho các loại dữ liệu định lượng ta dùng hàm describe() của thư viện pandas. Sau khi áp dụng lệnh ta thu được:

|  |  |
| --- | --- |
| **count** | Số lượng mẫu |
| **mean** | Trung bình mẫu |
| **std** | Độ lệch chuẩn |
| **min** | Giá trị nhỏ nhất |
| **25%**, **50%**, **75%** | Tứ phân vị |
| **max** | Giá trị lớn nhất |

|  |  |
| --- | --- |
| **Mthly\_HH\_Income**  count 50.000000  mean 41558.000000  std 26097.908979  min 5000.000000  25% 23550.000000  50% 35000.000000  75% 50375.000000  max 100000.000000  Name: Mthly\_HH\_Income,  dtype: float64 |  |
| **Mthly\_HH\_Expense**  count 50.000000  mean 18818.000000  std 12090.216824  min 2000.000000  25% 10000.000000  50% 15500.000000  75% 25000.000000  max 50000.000000  Name: Mthly\_HH\_Expense,  dtype: float64 |  |
| **Emi\_or\_Rent\_Amt**  count 50.000000  mean 3060.000000  std 6241.434948  min 0.000000  25% 0.000000  50% 0.000000  75% 3500.000000  max 35000.000000  Name: Emi\_or\_Rent\_Amt, dtype:  float64 |  |
| **Annual\_HH\_Income**  count 5.000000e+01  mean 4.900190e+05  std 3.201358e+05  min 6.420000e+04  25% 2.587500e+05  50% 4.474200e+05  75% 5.947200e+05  max 1.404000e+06  Name: Annual\_HH\_Income, dtype: float64 |  |

Bằng mắt thường ta có thể nhận thấy các dữ liệu trên bị lệch trái, và rất có thể nó không tuân theo phân phối chuẩn.

**Mthly\_HH\_Income: Chủ yếu tập trung từ 20000 – 50000**

**Mthly\_HH\_Expense: Chủ yếu tập trung từ 10000 – 25000**

**Emi\_or\_Rent\_Amt: Chủ yếu tập trung từ 0 - 4000**

**Annual\_HH\_Income: Chủ yếu tập trung từ 0.2 - 0.6**

Về dữ liệu đinh tính gồm: ***Highest\_Qualified\_Member, No\_of\_Fly\_Members, No\_of\_Earning\_Members.***

Để thực hiện thống kê mô tả cho các loại dữ liệu định lượng ta dùng hàm describe() của thư viện pandas. Sau khi áp dụng lệnh ta thu được các chỉ số của dữ liệu :

|  |  |
| --- | --- |
| **count** | Số lượng mẫu |
| **mean** | Trung bình mẫu |
| **std** | Độ lệch chuẩn |
| **min** | Giá trị nhỏ nhất |
| **25%** | Tứ phân vị |
| **50%** |
| **75%** |
| **max** | Giá trị lớn nhất |

Cho chạy hàm describe() ta được kết quả của 4 loại dữ liệu định lượng:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Mthly\_HH\_Income** | **Mthly\_HH\_Expense** | **Emi\_or\_Rent\_Amt** | **Annual\_HH\_Income** |
| **count** | 50.000000 | 50.000000 | 50.000000 | 5.000000e+01 |
| **mean** | 41558.000000 | 18818.000000 | 3060.000000 | 4.900190e+05 |
| **std** | 26097.908979 | 12090.216824 | 6241.434948 | 3.201358e+05 |
| **min** | 5000.000000 | 2000.000000 | 0.000000 | 6.420000e+04 |
| **25%** | 23550.000000 | 10000.000000 | 0.000000 | 2.587500e+05 |
| **50%** | 35000.000000 | 15500.000000 | 0.000000 | 4.474200e+05 |
| **75%** | 50375.000000 | 25000.000000 | 3500.000000 | 5.947200e+05 |
| **max** | 100000.000000 | 50000.000000 | 35000.000000 | 1.404000e+06 |

Với kiểu dữ liệu là dữ liệu định lượng, ta sử dụng boxplot để mô tả dữ liệu một cách trực quan:

|  |  |
| --- | --- |
| **Highest\_Qualified\_Member**  count 50  unique 5  top Graduate  freq 19  Name: Highest\_Qualified\_Member,  dtype: object |  |
| **No\_of\_Fly\_Members**  count 50  unique 7  top 4  freq 15  Name: No\_of\_Fly\_Members,  dtype: object |  |
| **No\_of\_Earning\_Members**  count 50  unique 4  top 1  freq 33  Name: No\_of\_Earning\_Members,  dtype: object |  |

**Highest\_Qualified\_Member: Level chiếm đa số: ‘Graduate’ (19)**

**No\_of\_Fly\_Members: Level chiếm đa số: ‘4’ (15)**

**No\_of\_Earning\_Members: Level chiếm đa số: ‘1’ (33)**

1. **KHÁI QUÁT VỀ THỐNG KÊ SUY DIỄN**

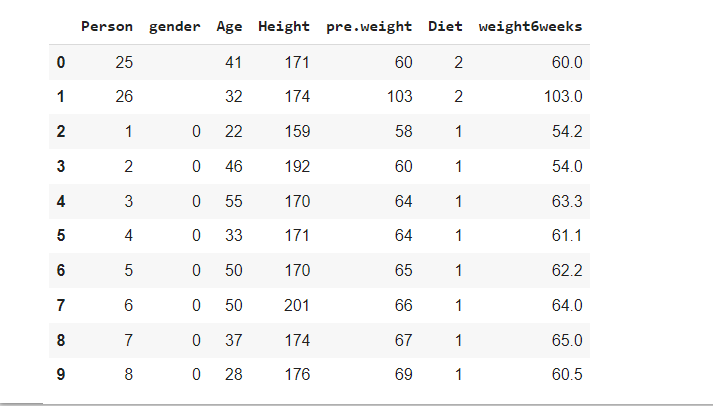
Thống kê suy diễn là quá trình suy luận suy ra các đặc điểm của một phân phối cơ bản bằng việc phân tích dữ liệu. Thống kê suy diễn ra tính chất của tổng thể: điều này bao gồm các giả thuyết thử nghiệm và các ước tính phát sinh. Tổng thể được giả định là lớn hơn so với tạo ra các dữ liệu quan sát, nói cách khác, các dữ liệu quan sát được giả định là lấy mẫu từ một tổng thể lớn hơn.(2)

Thống kê suy diễn có thể được so sánh với số liệu thống kê mô tả. Thống kê mô tả chỉ quan tâm tới tính chất của dữ liệu quan sát, và không giả sử các dữ liệu đến từ dữ liệu lớn hơn.

Thống kê suy diễn giải thích rõ ràng về một tổng thể, sử dụng dữ liệu rút ra từ tổng thể thông qua hình thức lấy mẫu cho một giả thuyết về tổng thể mà chúng ta muốn suy ra. Ví dụ:

* + - Bác bỏ giả thuyết.
    - Ước lượng khoảng thời gian
    - Phân nhóm hoặc phân tầng các điểm dữ liệu thành các nhóm.

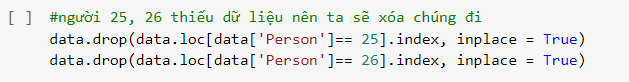
1. **PHẦN PHÂN TÍCH THỐNG KÊ SUY DIỄN, PHÂN TÍCH ANOVA MỘT CHIỀU CỦA TẬP TIN DIET\_R.CSV**
2. **Khái quát về file Diet\_R.csv:**



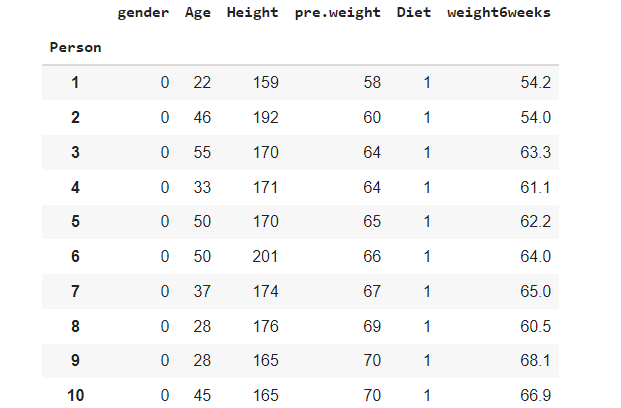
-Dữ liệu gồm các chỉ số cân năng của một số người trước và sau khi thực hiện các phương pháp giảm cân.

-Bởi vì người 25, 26 thiếu dữ liệu nên để giải quyết các bài toán 1 và 2. Tôi sẽ tiến hành xóa 2 cột dữ liệu đó đi.

-Ta sử dụng câu lệnh sau:



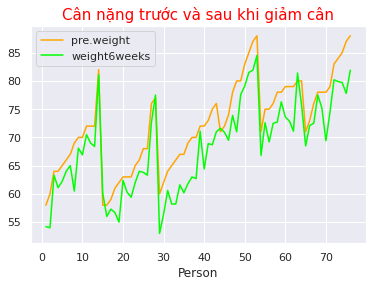
-Kết quả thu được là :

**

1. **Bài toán 1: CÓ SỰ KHÁC NHAU VỀ HIỆU QUẢ CỦA CÁC PHƯƠNG PHÁP GIẢM CÂN?**

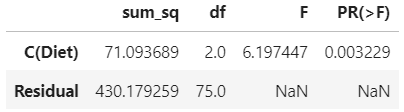
**Phát biểu giải thuyết thống kê**

* Cân nặng trung bình của trước và sau khi giảm cân là bằng nhau: H0:μ0=μ1
* Cân nặng trung bình của trước và sau khi giảm cân là khác nhau: H1:μ0<>μ1

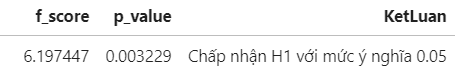


Nhìn vào đồ thị boxplot trên ta có thể nhận thấy có một chút khác việt về cân nặng trước và sau khi giảm cân.

### **Tiến hành tìm bảng Anova:**



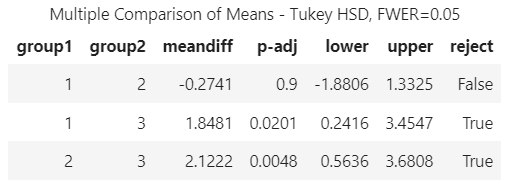
Vì p\_value bé hơn 0.05 nên ta sẽ bác bỏ giải thuyết H0 với mức ý nghĩa 0.05.



Ta sẽ tạm chấp nhân H1: có ít nhất hai kiểu phương pháp khác nhau về hiệu quả giảm cân.

Vì chưa thể biết được sự khác biệt là ở cặp giảm cân nào nên chúng ta sẽ tiến hành kiểm tra hậu nghiệm.

### **Kiểm tra hậu nghiệm:**

****

*Với mức ý nghĩa là 0.05, ta nhận thấy*

Phương pháp 1 và phương pháp 2 không có sự khác biệt về hiệu quả giảm cân

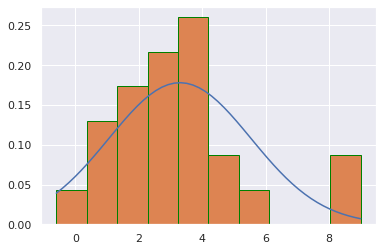
Phương pháp 1 và phương pháp 3 có sự khác biệt về hiệu quả giảm cân

Phương pháp 2 và phương pháp 3 có sự khác biệt về hiệu quả giảm cân

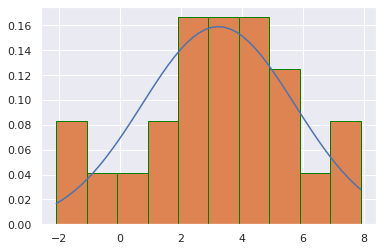
### **Kiểm tra các điều kiện cho kiểm định Anova**

**Vẽ biểu đồ histogram cho diet1; diet2; diet3 :**

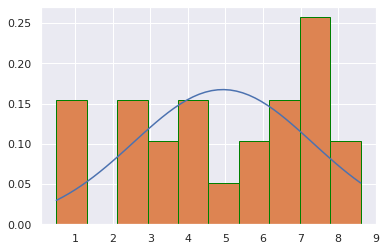
**+Diet 1**

****

**+Diet 2**

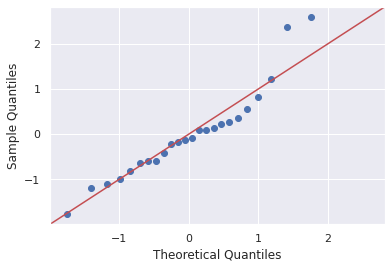
****

**+Diet 3**

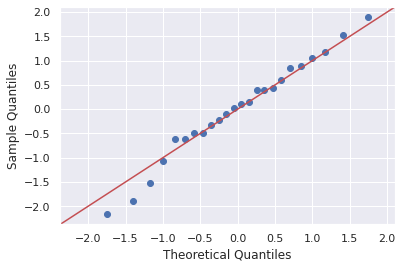
****

**Vẽ biểu đồ qq plot cho diet1; diet2; diet3 :**

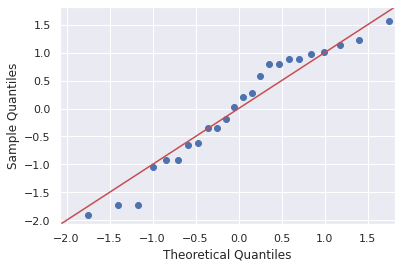
**+Diet 1**

**++**

**+Diet 2**

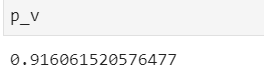
****

**+Diet 3**

****

+Bằng mắt thường ta có thể nhận thấy dữ liệu trên phân phối chuẩn, để kiểm tra chính xác hơn ta tiến hành kiểm định shapiro:

p\_v = stats.shapiro(model.resid)[1]

**

Chấp nhận kiểu phân bố dữ liệu trên là phân phối chuẩn.

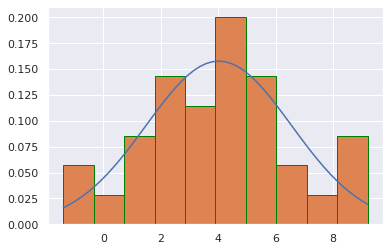
Tiến hành kiểm định về phương sai. Thỏa mãn các điều kiện để kiểm định ANOVA một chiều

1. **Bài toán 2:CÓ SỰ KHÁC NHAU VỀ HIỆU QUẢ GIẢM CÂN VỚI CÁC GIỚI TÍNH KHÁC NHAU?**

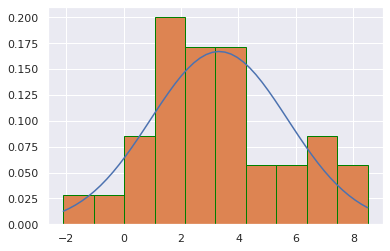
**Phát biểu giải thuyết thống kê**

* *H0*: không có sự khác biệt về hiệu quả giảm cân đối với các giới tính khác nhau
* *H1*: Các giới khác nhau về hiệu quả giảm cân

**+Biểu đồ giảm cân của giới tính nam:**

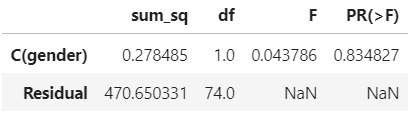


**+Biều đồ giảm cân của giới tính nữ:**

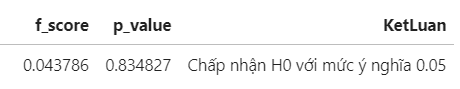


Ta nhận thấy sự tương đồng về sự hiệu quả giảm cân trung bình của 2 giới tính nam và nữ.

### **Tiến hành tìm bảng Anova:**



Vì p\_value lớn hơn 0.05 nên ta sẽ chấp nhận giả thuyết H0 với mức ý nghĩa 0.05.

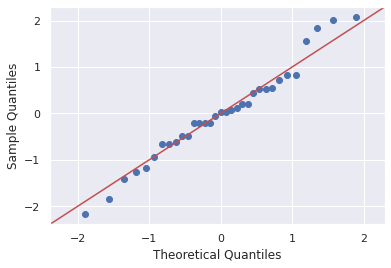


Vì dữ liệu liên quan đến giới tính là một biến nhị phân, chỉ có 0 và 1, nên ta không cần tiến hành kiểm tra hậu nghiệm.

**Kiểm tra các điều kiện cho kiểm định Anova:**

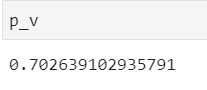
**Vẽ biểu đồ qqplot :**

+Giới tính nam

****

*Bằng mắt thường ta có thể nhận thấy dữ liệu trên phân phối chuẩn, để kiểm tra chính xác hơn ta tiến hành kiểm định shapiro:*

*p\_v = stats.shapiro(model\_2.resid)[1]*

****

Chấp nhận kiểu phân bố dữ liệu trên là phân phối chuẩn.

Tiến hành kiểm định về phương sai. Thỏa mãn các điều kiện để kiểm định ANOVA một chiều

# Kết quả kiểm tra đạo văn:

