Nguyễn Kim đại Nghĩa 5.1 và 5.2

5.1: Bài toán 2 mẫu (gợi ý Kiểm định 2 trung bình)

5.2: Bài toán phân tích phương sai (Gợi ý phân tích phương sai một yếu tố)

Thống kê suy diễn

5.1 Bài toán kiểm định 2 mẫu

Bài toán kiểm định hai mẫu là một bài toán thống kê nhằm kiểm tra xem hai mẫu ngẫu nhiên có đại diện cho hai quần thể có cùng một tham số hay không. Cơ sở lý thuyết của kiểm định hai mẫu cũng dựa trên nguyên lý xác suất nhỏ. Theo nguyên lý này, nếu hai mẫu ngẫu nhiên đại diện cho hai quần thể có cùng một tham số, thì xác suất để thu được một kết quả mẫu nhất định là rất nhỏ.

* Hiệu hai trung bình của hai tổng thể theo phân phối chuẩn đã biết phương sai tổng ~N(0;1):
* Hiệu hai trung bình của hai tổng thể với mẫu nhỏ chưa biết phương sai tổng nhưng có giả thuyết 2 phương sai bằng nhau thì phân phối Student xấp xỉ phân phối chuẩn ~ N(0;1):
* Hiệu hai trung bình của hai tổng thể với mẫu nhỏ chưa biết phương sai tổng nhưng có giả thuyết 2 phương sai khác nhau thì phân phối Student xấp xỉ phân phối chuẩn ~ N(0;1):
* Hiệu hai trung bình của hai tổng thể với mẫu lớn chưa biết phương sai tổng thể thì phân phối Student xấp xỉ quy luật phân phối chuẩn ~N(0;1):
* Hiệu hai trung bình cho hai mẫu phối hợp từng cặp (1 dấu hiệu ở hai kì khác nhau) theo phân phối chuẩn:

Trong đó: =

* Hiệu hai trung bình cho hai mẫu lớn và phối hợp từng cặp (1 dấu hiệu ở hai kì khác nhau) có phân phối tùy ý:

Trong đó: =

Quy tắc kiểm định

Xác định miền giá trị Wα - là miền bác bỏ H0 với mức ý nghĩa α đủ nhỏ (thường lấy giá trị ⩽ 0.05)

• Nếu giá trị của G ∈ Wα thì bác bỏ H0.

• Nếu giá trị của G Wα thì chấp nhận H0.

Các loại sai lầm:

• Sai lầm loại 1: Bác bỏ một điều đúng - Bác bỏ H0 khi H0 đúng. Xác suất mắc sai lầm loại 1 bằng mức ý nghĩa là α

• Sai lầm loại 2: Thừa nhận một điều sai - chấp nhận H0 khi H0 sai. Xác suất mắc sai lầm loại 2 bằng β, 1 - β gọi là lực kiểm định

Các dạng miền tới hạn: Người ra chọn miền tới hạn Wα của tiêu chuẩn G phụ thuộc vào H0,H1 và α như sau:

• Nếu H0: θ =, H1: θ ̸= thì ta thực hiện kiểm định hai phía với:

P(G < G1|H0) = P(G > G2|H0) =

• Nếu H0: θ = θ0, H1: θ < θ0 (đối thuyết lệch trái) thì ta chọn G1 thỏa

P(G < G1|H0) = α

• Nếu H0: θ =, H1: θ > (đối thuyết lệch phải) thì ta chọn G2 thỏa

P(G > G1|H0) = α

Các bước tiến hành kiểm định:

B1: Xác định giả thuyết không và giả thuyết đối.

B2: Chọn mức ý nghĩa.

B3: Chọn tiêu chuẩn kiểm định.

B4: Tính giá trị quan sát của tiêu chuẩn kiểm định.

B5: Quyết định

5.2 **Phân tích phương sai (ANOVA)**

Mục tiêu của phân tích phương sai là so sánh trung bình của nhiều nhóm (tổng thể) dựa trên các số trung bình của các mẫu quan sát từ các nhóm này và thông qua kiểm định giả thuyết để kết luận về sự bằng nhau của các số trung bình này.

Trong nghiên cứu, phân tích phương sai được dùng như là một công cụ để xem xét ảnh hưởng của một hay một số yếu tố nguyên nhân (định tính) đến một yếu tố kết quả (định lượng)

**Phân tích phương sai một yếu tố**

Phân tích phương sai một yếu tố là phân tích ảnh hưởng của một yếu tố nguyên nhân (dạng biến định tính) đến một yếu tố kết quả (dạng biến định lượng) đang nghiên cứu.

Giả sử cần so sánh số trung bình của k tổng thể độc lập. Ta lấy k mẫu có số quan sát là n1,n2. . . nk; tuân theo phân phối chuẩn. Trung bình của các tổng thể được ký hiệu là µ1, µ2, ....µk thì mô hình phân tích phương sai một yếu tố ảnh hưởng được mô tả dưới dạng kiểm định giả

thuyết như sau:

• H0 : µ1, µ2, ....µk

• H1: Tồn tại ít nhất 1 cặp có µi ̸= µj; i ̸= j

Để kiểm định ta đưa ra 3 giả thiết sau:

• Mỗi mẫu tuân theo phân phối chuẩn N(µ, σ2)

• Các phương sai tổng thể bằng nhau

• Ta lấy k mẫu độc lập từ k tổng thể. Mỗi mẫu được quan sát nj lần.

Các bước tiến hành:

B1: Tính các trung bình mẫu và trung bình chung của k mẫu

Ta lập bảng tính toán như sau

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| TT | k mẫu quan sát | | | | |
| 1 | 2 | 3 | … | k |
| 1 |  |  |  |  |  |
| 2 |  |  |  |  |  |
| 3 |  |  |  |  |  |
| … |  |  |  |  |  |
| … |  |  |  |  |  |
| j |  |  |  |  |  |
| Trung bình mẫu |  |  |  |  |  |

• Trung bình mẫu x1, x2, ...xk được tính theo công thức:

• Trung bình chung của k mẫu được tính theo công thức:

B2: Tính các tổng độ lệch bình phương:

• Tổng các độ lệch bình phương trong nội bộ nhóm (nội bộ từng mẫu - SSW) được tính theo công thức sau:

SSW = + +...+ =

• Tổng các độ lệch bình phương giữa các nhóm (SSB)

SSB =

• Tổng các độ lệch bình phương của toàn bộ tổng thể(SST)

SST = SSW + SSB =

B3: Tính các phương sai (phương sai của nội bộ nhóm và phương sai giữa các nhóm):

B4: Kiểm định giả thuyết

• Tính tiêu chuẩn kiểm định F (F thực nghiệm)

• Nếu F > F ((k-1; n-k); α) ta bác bỏ quả thuyết H0 cho rằng trị trung bình của k tổng thểbằng nhau

• Tìm F lý thuyết (F tiêu chuẩn = F (k-1; n-k; α))

• F lý thuyết là giá trị giới hạn tra từ bảng phân phối F với k-1 bậc tự do của phương sai ở tử số và ; n-k bậc tự do của phương sai ở mẫu số với mức ý nghĩa α.

• Nếu F thực nghiệm > F lý thuyết, bác bỏ H0, nghĩa là các số trung bình của k tổng thể không bằng nhau

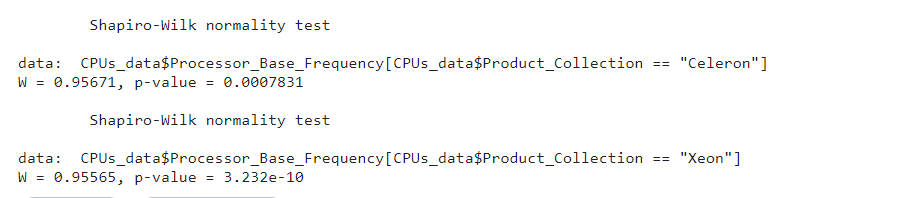
5.1 Kiểm định 2 mẫu

Ta sử dụng số liệu về Processor\_Base\_Frequency của 2 mẫu CPU là Xeon và Celeron.

Mục tiêu: Kiểm định được tần số cơ sở trung bình của bộ xử lí của 2 dòng CPU là Xeon và Celeron.

Ta sử dụng hàm shapio.test để kiểm định Processor\_Base\_Frequency của 2 dòng CPU Xeon và Celeron có tuân theo phân phối chuẩn không.

Kết quả chạy code:

****

Hinhx. Kết quả kiểm định shapiro.test

Nhận xét: kết quả sau khi kiểm tra thì ta thấy trị số p-value của 2 dòng CPU Xeon và Celeron đều nhỏ hơn 0,05 rất nhiều, cho nên chúng ta có thể nói rằng CPU Xeon và Celeron không tuân theo luật phân phối chuẩn. Trong trường hợp này, chúng ta có thể dựa vào phương pháp phi tham số có tên là kiểm định Wilcoxon, vì kiểm định này không tùy thuộc vào giả định phân phối chuẩn.Ta chọn ra 1000 biến ngẫu nhiên của 2 mẫu để thực hiện kiểm định.

Kiểm định tần cơ bản trung bình của 2 dòng CPU là Xeon và Celeron.

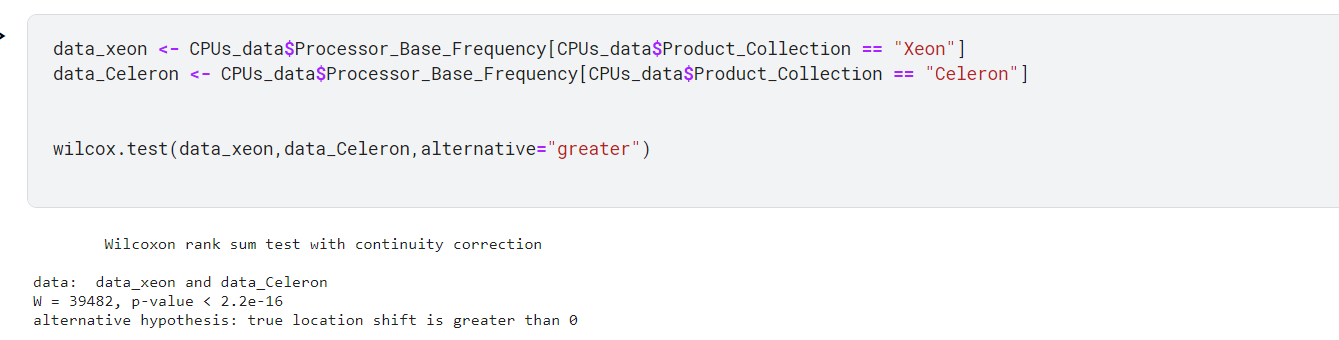
Gọi là tần số cơ sở trung bình của dòng CPU Xeon

Gọi là tân số cơ sở trung bình của dòng CPU Celeron.

Giả thuyết kiểm định :

Giả thuyết đối là :

Kết quả chạy code:



Từ trị số p<0.05, ta kết luận rằng có thể bác bỏ chấp nhận .

Vậy có thể kết luận rằng có sự chênh lệch ý nghĩa giữa 2 dòng CPU Xeon và Celeron với hướng chênh lệch là lớn hơn, kết quả này có ý nghĩa thống kê.

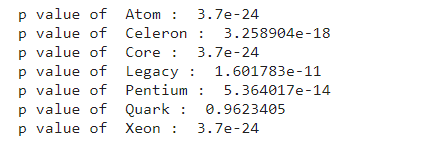
5.2

**Phân tích phương sai một yếu tố**

Để có thể kiểm định Anova, ta cần phải xác định được điều kiện của nó. Trong phần sử dụng Recommemnd\_Customer\_Price của các loại sản phẩm intel.Với mục đích so sánh giá tiền của các dòng sản phẩm của Intel

**Giả định phân phối chuẩn**

Sử dụng kiểm định Aderson\_Darling test để kiểm tra biến Recommend\_Customer\_Price có phân phối theo từng nhóm Product\_Collection hay không. Bằng cách dựa vào giá trị p được tính toán, nếu giá trị nhỏ hơn 0.05 thì ta có thể bác bỏ giả thuyết rằng biến có phân phối chuẩn: Dựa vò kết quả kiểm định, chỉ có loại “Quark” tuân theo phân phối chuẩn với khoảng giá trị là 0.9623405.

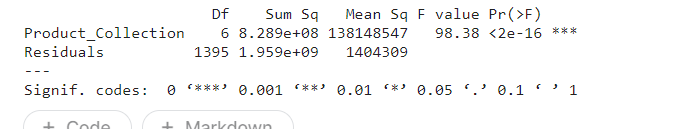


Hình x. Kết quả kiểm định Aderson Darling Test

**Giả định phương sai đồng nhất**

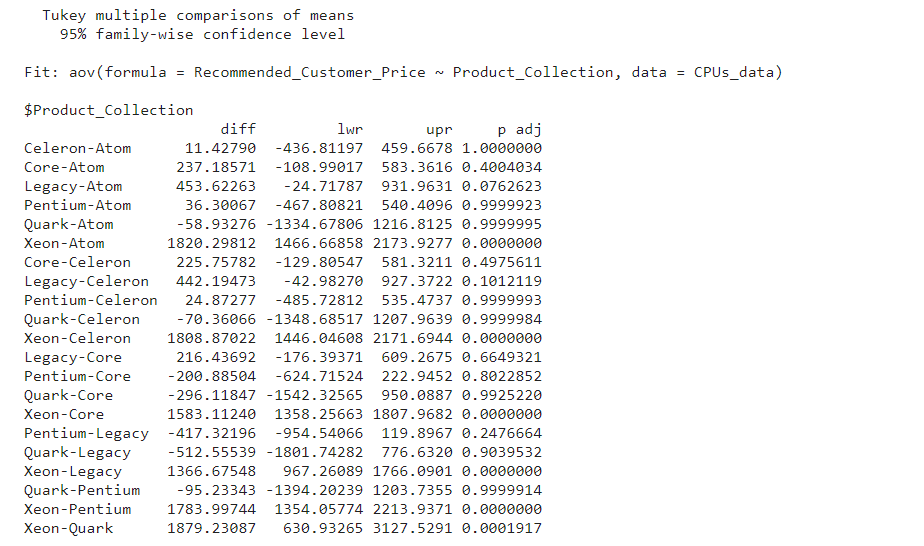
**Kiểm định Anova:**

Sử dụng Anova để kiểm tra: Kết quả thu được giá trị p < 0.05, cho thấy rằng có sự khác biệt đáng kể giữa các nhóm sản phẩm đối với giá tiền đề xuất



Hinhx. Kết quả kiểm định anova

Để làm rõ sự khác biệt trên Tukey HSD test để kiểm tra sự khác biệt cụ thể của các nhóm nhà sản xuất: Kết quả cho thấy có sự khác biệt giữa các nhóm dữ liệu với nhau. Khác biệt giá tiền đề xuất lớn nhất là giữa loại CPU Xeon và Quark với 1879.2309 và khoảng tin cậy từ 630.9327 đến 3127.5291.



d. Kết luận:

Dựa vào kết quả phân tích Anova ta có thể thấy có sự khác biệt lớn về giá tiền giữa các nhóm CPU. Thật vậy, khi ta vẽ biển đồ hộp hay biểu đồ đường sẽ thấy sự khác biệt giá tiền rõ rệt của các nhóm.Tuy nhiên số lượng mẫu nhiều nhưng không có phân phối chuẩn và không đồng nhất về phương sai nên kết quả phân tích Anova có thể không chính xác và đáng tin cậy. Vì vậy mô hình Anova không hoạt động tốt trong trường hợp này và kết quả phân tích trên chỉ nên tham khảo.