**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 2 yếu tố**

Phân tích phương sai 2 yếu tố là phân tích ảnh hưởng của 2 yếu tố nguyên nhân (dạng biến định tính) đến một yếu tố kết quả (dạng biến định lượng) đang nghiên cứu

Giả sử ta nghiên cứu ảnh hưởng của 2 yếu tố nguyên nhân định tính đến một yếu tố kết quả định lượng nào đó.

Ta lấy mẫu không lặp lại, sau đó các đơn vị mẫu của yếu tố nguyên nhân thứ nhất sắp xếp thành K nhóm (cột), các đơn vị mẫu của yếu tố nguyên nhân thứ hai sắp xếp thành H khối (hàng). Như vậy, ta có bảng kết hợp 2 yếu tố nguyên nhân gồm K cột và H hàng và (K x H) ô dữ liệu. Tổng số mẫu quan sát là n = (K x H).

Để kiểm định ta đưa ra 2 giả thiết sau:

* Mỗi mẫu tuân theo phân phối chuẩn N(μ, σ­2)
* Ta lấy K mẫu độc lập từ K tổng thể, H mẫu độc lập từ H tổng thể. Mỗi mẫu được quan sát 1 lần không lặp.

Các bước tiến hành:

Bước 1: Tính các trung bình riêng của từng nhóm/khối và trung bình trung của toàn bộ mẫu quan sát

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **1** | **2** |  | **c** |  |  |
| **1** |  |  | … |  |  |  |
| **2** |  |  | … |  |  |  |
| … | … | … | … | … | … | … |
| **h** |  |  | … |  |  |  |
|  |  |  | … |  |  |  |
|  |  |  | … |  |  |  |

Trung bình riêng của từng nhóm (c cột) và từng khối (h hàng) được tính theo công thức:

Trung bình chung của toàn bộ mẫu quan sát được tính theo công thức:

Bước 2: Tính các tổng độ lệch bình phương:

* Tổng các độ lệch bình phương chung (SST)
* Tổng các độ lệch bình phương giữa các cột (SSK)
* Tổng các độ lệch bình phương giữa các hàng (SSH)
* Tổng các độ lệch bình phương phần dư (ERROR)

Bước 3: Tính các phương sai

* Phương sai giữa các cột (MSK)
* Phương sai giữa các hàng (MSH)
* Phương sai phần dư

Bước 4: Kiểm định giả thuyết

Tính tiêu chuẩn kiểm định F (F thực nghiệm)

Trong đó:

- MSH là phương sai giữa các nhóm (hàng)

- MSK là phương sai giữa các khối (cột)

- MSE là phương sai phần dư

Tìm F lý thuyết cho 2 yếu tố nguyên nhân

- Yếu tố nguyên nhân thứ nhất:

F tiêu chuẩn = F (k-1; (k-1)(h-1), α) là giá trị giới hạn tra từ bảng phân phối F với k-1 bậc tự do của phương sai ở tử số và (k-1)(h-1) bậc tự do của phương sai ở mẫu số với mức ý nghĩa α.

- Yếu tố nguyên nhân thứ hai:

F tiêu chuẩn = F (h-1; (k-1)(h-1), α) là giá trị giới hạn tra từ bảng phân phối F với h-1 bậc tự do của phương sai ở tử số và (k-1)(h-1) bậc tự do của phương sai ở mẫu số với mức ý nghĩa α.

Nếu F1 thực nghiệm > F1 lý thuyết, bác bỏ Ho, nghĩa là các số trung bình của k tổng thể nhóm (cột) không bằng nhau.

Nếu F2 thực nghiệm > F2 lý thuyết, bác bỏ Ho, nghĩa là các số trung bình của k tổng thể khối (hàng) không bằng nhau.

Dưới đây là bảng tổng kết các công thức:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nguồn** | **SS** | **df** | **MS** | **F** |
| **Yếu tố A** |  | h-1 |  |  |
| **Yếu tố B** |  | c-1 |  |  |
| **Sai số** |  | (h-1)(c-1) |  |  |
| **Tổng** |  | h.c-1 |  |  |

2/ Chuyển đổi số liệu (có thể cho phần này vô chung với xử lý dữ liệu cũng được á)

Xử lý chuyển đổi số liệu ở cột Processor\_Base\_Frequency về đơn vị GHz

A screenshot of a computer code

Description automatically generated

Hợp nhất các dữ liệu ở cột nb\_of\_Cores thành 3 nhóm: CPU có số nhân dưới 4, số nhân từ 4 đến 10 và số nhân trên 10

A screen shot of a computer code

Description automatically generated

3/ ANOVA trong R

Đầu tiên ta kiểm tra điều kiện để thực hiện ANOVA, với mục đích là so sánh hiệu suất CPU giữa các nhóm CPU khác nhau với nhóm nhân khác nhau.

Đối với yếu tố loại CPU (tức Vertical\_Segment):

* H0: Không có sự khác biệt đáng kể về xung nhịp của CPU giữa các loại CPU (Embedded, Desktop, Mobile, Server).
* H1: Có sự khác biệt đáng kể về xung nhịp của CPU giữa các loại CPU (Embedded, Desktop, Mobile, Server).

Đối với yếu tố số nhân CPU (Tức CPU\_Group)

* H0: Không có sự khác biệt đáng kể về xung nhịp của CPU giữa các nhóm nhân CPU (dưới 4 nhân, từ 4 đến 10 nhân, trên 10 nhân).
* H1: Có sự khác biệt đáng kể về xung nhịp của CPU giữa các nhóm nhân CPU (dưới 4 nhân, từ 4 đến 10 nhân, trên 10 nhân).

Đối với sự tương tác giữa 2 yếu tố trên:

* H0: Không có sự tương tác đáng kể giữa loại CPU và nhóm nhân CPU đối với xung nhịp CPU.
* H1: Có sự tương tác đáng kể giữa loại CPU và nhóm nhân CPU đối với xung nhịp CPU.

Kiểm tra shapiro-wilk Processor\_Base\_Frequency có phân phối chuẩn hay không?

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

Giá trị p của các loại CPU và nhóm nhân CPU đều nhỏ hơn 0.05, mẫu không tuân theo luật phân phối chuẩn. Ngoài ra ta có thể dùng qqplot để kiểm tra

Kiểm tra các nhóm có phương sai đồng nhất không

A screenshot of a computer

Description automatically generated

Với p < 0.005 cho thấy được các nhóm không có phương sai đồng nhất.

Kiểm định ANOVA

A screenshot of a computer

Description automatically generated

Kết quả Anova cho thấy ta bác bỏ H0 và chấp nhận H1 cho 2 yếu tố đơn lẻ “Vertical\_Segment” và “CPU\_Group” (p < 0.05), chứng tỏ rằng có sự khác biệt giữa 2 yếu tố đơn lẻ trên đối với xung nhịp CPU. Tuy nhiên đối với tương tác giữa 2 yếu tố, vì p = 0.175 > 0.05, vì vậy ta bác bỏ H1, chấp nhận H0, chứng tỏ rằng không có sự tương tác giữa 2 yếu tố đối với xung nhịp của CPU

Để làm rõ sự khác biệt trên, ta sử dụng Tukey HSD test để kiểm tra sự khác biệt cụ thể của các nhóm.

A screenshot of a computer

Description automatically generated

Ở phần Vertical\_Segment

* Embedded và Desktop: xung nhịp của CPU ở phân khúc Embedded thấp hơn ở Desktop khoảng 0.717GHz, với p < 0.05
* Mobile và Desktop: xung nhịp của CPU ở phân khúc Mobile thấp hơn ở Desktop khoảng 0.724GHz, với p < 0.05
* Server và Desktop: xung nhịp của CPU ở phân khúc Server thấp hơn ở Desktop khoảng 0.172 GHz, với p < 0.05
* Mobile và Embedded: không có sự khác biệt đáng kể, với p = 0.9994
* Server và Embedded: xung nhịp của CPU ở phân khúc Server cao hơn ở Embedded khoảng 0.544GHz, với p < 0.05
* Server và Mobile: xung nhịp của CPU ở phân khúc Server cao hơn ở Mobile khoảng 0.552 GHz, với p < 0.05

Ở phần CPU\_Group:

* Từ 4 đến 10 nhân và Dưới 4 nhân: CPU có từ 4 nhân đến 10 nhân có xung nhịp cao hơn CPU dưới 4 nhân khoảng 0.44 GHz, với p < 0.05
* Trên 10 nhân và Dưới 4 nhân: Không có sự khác biệt đáng kể về xung nhịp của CPU có trên 10 nhân so với CPU dưới 4 nhân, với p = 0.8537
* Trên 10 nhân và Từ 4 đến 10 nhân: CPU có trên 10 nhân có xung nhịp thấp hơn CPU có từ 4 đến 10 nhân khoảng 0.47GHz, với p < 0.05

**Kết luận:** Do không thỏa ở cả 2 điều kiện phân phối chuẩn và phương sai đồng nhất, cho nên phân tích phương sai 2 yếu tố chỉ mang tính chất tham khảo