Xác suất Thống kê ứng dụng trong kinh tế - xã hội

Nguyễn Thị Nhung

Bộ môn Toán - Đại học THĂNG LONG

Ngày 21 tháng 8 năm 2013

Chương IX

Phân tích phương sai

Chương IX

Chương IX

Chương IX

1. Phân tích phương sai một yếu tố
   - So sánh trung bình của nhiều tổng thể
   - So sánh phương sai của nhiều tổng thể
2. Phân tích sâu One-way ANOVA
3. Phân tích phương sai hai yếu tố
   - Trường hợp có một quan sát trong một ô
   - Trường hợp nhiều quan sát trong một ô
4. Phân tích sâu Two-way ANOVA

Chương IX

1. Phân tích phương sai một yếu tố
   - So sánh trung bình của nhiều tổng thể
   - So sánh phương sai của nhiều tổng thể
2. Phân tích sâu One-way ANOVA
3. Phân tích phương sai hai yếu tố
   - Trường hợp có một quan sát trong một ô
   - Trường hợp nhiều quan sát trong một ô
4. Phân tích sâu Two-way ANOVA

Nội dung chính được giới thiệu trong chương

- Giới thiệu bài toán so sánh nhiều trung bình bằng phương pháp phân tích phương sai (ANOVA);
- Giới thiệu phân tích phương sai mô hình một nhân tố (One-way ANOVA);
- Giới thiệu bài toán phân tích sâu One-way ANOVA;
- Giới thiệu bài toán phân tích phương sai mô hình hai nhân tố: một quan sát trong một ô;
- Giới thiệu bài toán phân tích phương sai mô hình hai nhân tố: nhiều quan sát trong một ô;
- Giới thiệu bài toán phân tích sâu mô hình hai nhân tố (Two-way ANOVA).

Yêu cầu đối với sinh viên

- Nắm được bài toán thực hiện trong chương: Bài toán so sánh nhiều trung bình;
- Hiểu được logic của bài toán so sánh nhiều trung bình bằng phương pháp phân tích phương sai (ANOVA);
- Hiểu được cách thực hiện các loại bài toán phân tích phương sai:
  - Phân tích phương sai mô hình một nhân tố (One-way ANOVA);
  - Phân tích sâu One-way ANOVA;
  - Phân tích phương sai mô hình hai nhân tố: một quan sát trong một ô;
  - Phân tích phương sai mô hình hai nhân tố: nhiều quan sát trong một ô;
  - Phân tích sâu mô hình hai nhân tố (Two-way ANOVA).

Nội dung trình bày

Bài toán tình huống

Câu hỏi tình huống

Trong thời gian gần đây rất nhiều sinh viên trong trường bạn cho rằng thời gian học ở nhà không ảnh hưởng đến kết quả học tập, nghĩa là học nhiều hay ít điểm cũng thế thôi. Để kiểm định xem những suy nghĩ của những bạn sinh viên này có chính xác không, thầy Hiệu trưởng yêu cầu Văn phòng Đoàn điều tra để báo cáo lại kết quả. Sau một thời gian điều tra điểm của những sinh viên trong trường theo ba nhóm sinh viên đánh giá mình là thời gian tự học ít, thời gian tự học bình thường và thời gian tự học nhiều được bảng dữ liệu sau. Bạn làm thế nào để khẳng định được những phát biểu của những sinh viên trên là có cơ sở hay không.

Bài toán tình huống

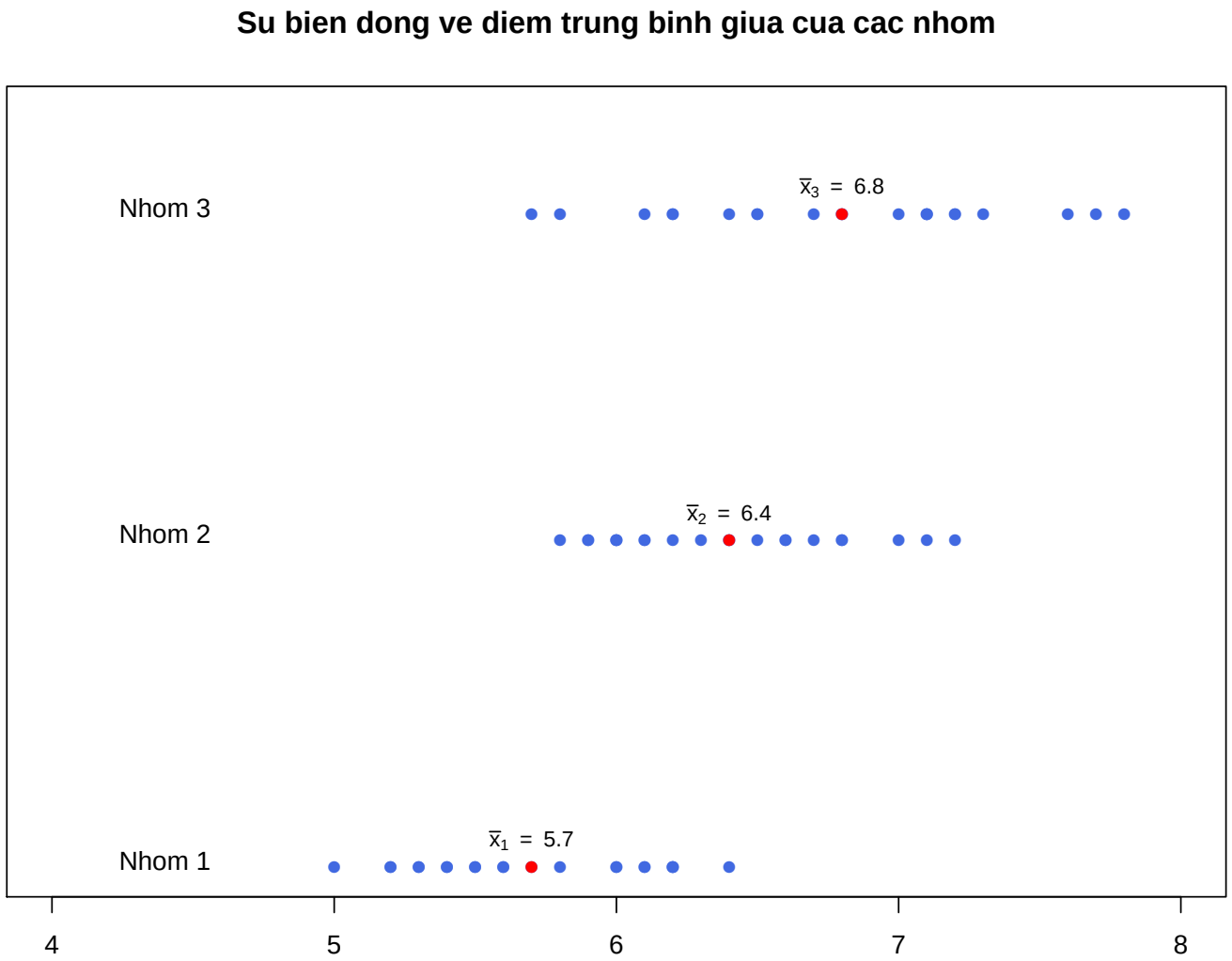
Nhóm I (TG tự học ít)	Nhóm II (TG tự học TB)	Nhóm III (TG tự học nhiều)
5.8	6.0	6.2
6.2	6.6	5.8
5.4	6.1	6.5
6.0	5.8	6.2
5.2	5.9	6.4
5.3	6.0	5.7
5.4	5.9	6.1
5.6	6.0	6.8
6.2	6.7	7.1
5.7	6.5	6.5
5.5	6.3	7.1
6.1	6.1	7.2
6.0	6.8	6.7
5.2	6.4	7.0
6.4	6.8	7.6
5.5	6.6	7.7
5.0	6.4	7.8
5.6	6.2	6.8
6.2	7.1	7.3
6.1	7.0	7.1
5.3	7.2	7.2

Bài toán tình huống

Câu hỏi tình huống

- *Để xét xem thời gian tự học khác nhau có ảnh hưởng đến kết quả học tập không ta đi kiểm định bài toán nào?*
- *Làm thế nào để thực hiện được kiểm định bài toán trên?*

Minh họa sự biến động điểm trung bình của ba nhóm



Bài toán so sánh nhiều trung bình

Bài toán

Giả sử có k tổng thể tuân theo phân phối chuẩn, phương sai bằng nhau với trung bình lần lượt là $\mu_1, \mu_2, \ldots, \mu_k$. Ta cần so sánh trung bình của k tổng thể này dựa trên những mẫu ngẫu nhiên độc lập chọn ra từ k tổng thể này bằng cách kiểm định cặp giả thuyết

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

$$H_1 : \exists i \neq j : \mu_i \neq \mu_j, i, j = \overline{1, k}.$$

Bài toán so sánh nhiều trung bình

Để thực hiện bài toán so sánh nhiều trung bình ta có thể sử dụng phương pháp phân tích phương sai. Dựa trên cách thức điều tra chọn mẫu, ta có thể so sánh nhiều trung bình bằng cách thực hiện:

- Phân tích phương sai mô hình một nhân tố;
- Phân tích phương sai mô hình hai nhân tố một quan sát trong một ô;
- Phân tích phương sai mô hình hai nhân tố nhiều quan sát trong một ô.

So sánh nhiều trung bình bằng phân tích phương sai một nhân tố

- Phân tích phương sai một nhân tố (One-way ANOVA) là phân tích ảnh hưởng của một yếu tố nguyên nhân (dạng biến định tính) ảnh hưởng đến một yếu tố kết quả (dạng biến định lượng) đang nghiên cứu.
- Để so sánh trung bình của k tổng thể ta dùng phương pháp phân tích phương sai một nhân tố.
- Những giả định khi tiến hành phân tích One-way ANOVA:
  - Các tổng thể tuân theo phân phối chuẩn;
  - Các phương sai tổng thể bằng nhau;
  - Các mẫu chọn ra độc lập với nhau.

So sánh nhiều trung bình bằng phân tích phương sai một nhân tố

- Phân tích phương sai một nhân tố (One-way ANOVA) là phân tích ảnh hưởng của một yếu tố nguyên nhân (dạng biến định tính) ảnh hưởng đến một yếu tố kết quả (dạng biến định lượng) đang nghiên cứu.
- Để so sánh trung bình của k tổng thể ta dùng phương pháp phân tích phương sai một nhân tố.
- Những giả định khi tiến hành phân tích One-way ANOVA:
  - Các tổng thể tuân theo phân phối chuẩn;
  - Các phương sai tổng thể bằng nhau;
  - Các mẫu chọn ra độc lập với nhau.

So sánh nhiều trung bình bằng phân tích phương sai một nhân tố

- Phân tích phương sai một nhân tố (One-way ANOVA) là phân tích ảnh hưởng của một yếu tố nguyên nhân (dạng biến định tính) ảnh hưởng đến một yếu tố kết quả (dạng biến định lượng) đang nghiên cứu.
- Để so sánh trung bình của k tổng thể ta dùng phương pháp phân tích phương sai một nhân tố.
- Những giả định khi tiến hành phân tích One-way ANOVA:
  - Các tổng thể tuân theo phân phối chuẩn;
  - Các phương sai tổng thể bằng nhau;
  - Các mẫu chọn ra độc lập với nhau.

Qui trình thực hiện bài toán phân tích phương sai một yếu tố

- Bước 1: Tính các trung bình mẫu.
  - Giả sử ta có k mẫu với số phần tử lần lượt là $n_1, n_2, \ldots, n_k$ chọn từ k tổng thể được cho ở bảng dưới đây:

1	2	3	k
x_{11}	x_{21}	$\ldots$	x_{k1}
x_{12}	x_{22}	$\ldots$	x_{k2}
$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_{1n_1}	x_{2n_2}	$\ldots$	x_{kn_k}

Qui trình thực hiện

- Trung bình mẫu của từng nhóm $x_1, x_2, \ldots, x_n$ theo công thức:

$$\overline{x}_i = \frac{x_{i1} + x_{i2} + \ldots + x_{in_i}}{n_i}.$$

- Trung bình của k mẫu (mẫu gộp) $\overline{x}$ theo công thức:

$$\overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + \ldots + n_k\overline{x}_k}{n_1 + n_2 + \ldots + n_k}.$$

Qui trình thực hiện

- Trung bình mẫu của từng nhóm $x_1, x_2, \ldots, x_n$ theo công thức:

$$\overline{x}_i = \frac{x_{i1} + x_{i2} + \ldots + x_{in_i}}{n_i}.$$

- Trung bình của k mẫu (mẫu gộp) $\overline{x}$ theo công thức:

$$\overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + \ldots + n_k\overline{x}_k}{n_1 + n_2 + \ldots + n_k}.$$

Qui trình thực hiện

- Bước 2: Tính tổng các chênh lệch bình phương
  - Tổng bình phương trong nội bộ nhóm SSW được tính bởi:

$$SSW = SS_1 + SS_2 + \ldots + SS_k,$$

trong đó, SS_i là tổng bình phương của từng nhóm được tính bởi công thức:

$$SS_i = (x_{i1} - \overline{x}_i)^2 + (x_{i2} - \overline{x}_i)^2 + \ldots + (x_{in_i} - \overline{x}_i)^2.$$

  - Tổng bình phương giữa các nhóm SSG được tính bởi công thức

$$SSG = n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 + \ldots + n_k(\overline{x}_k - \overline{x})^2.$$

  - Tổng bình phương toàn bộ SST được tính bởi công thức

$$SST = (x_{11} - \overline{x})^2 + \ldots + (x_{1n_1} - \overline{x})^2 + \ldots + (x_{k1} - \overline{x})^2 + \ldots + (x_{kn_1} - \overline{x})^2.$$

Ta có $SST = SSW + SSG$.

Qui trình thực hiện

- Bước 2: Tính tổng các chênh lệch bình phương
  - Tổng bình phương trong nội bộ nhóm SSW được tính bởi:

$$SSW = SS_1 + SS_2 + \ldots + SS_k,$$

trong đó, SS_i là tổng bình phương của từng nhóm được tính bởi công thức:

$$SS_i = (x_{i1} - \overline{x}_i)^2 + (x_{i2} - \overline{x}_i)^2 + \ldots + (x_{in_i} - \overline{x}_i)^2.$$

  - Tổng bình phương giữa các nhóm SSG được tính bởi công thức

$$SSG = n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 + \ldots + n_k(\overline{x}_k - \overline{x})^2.$$

  - Tổng bình phương toàn bộ SST được tính bởi công thức

$$SST = (x_{11} - \overline{x})^2 + \ldots + (x_{1n_1} - \overline{x})^2 + \ldots + (x_{k1} - \overline{x})^2 + \ldots + (x_{kn_1} - \overline{x})^2.$$

Ta có $SST = SSW + SSG$.

Qui trình thực hiện

- Bước 2: Tính tổng các chênh lệch bình phương
  - Tổng bình phương trong nội bộ nhóm SSW được tính bởi:

$$SSW = SS_1 + SS_2 + \ldots + SS_k,$$

trong đó, SS_i là tổng bình phương của từng nhóm được tính bởi công thức:

$$SS_i = (x_{i1} - \overline{x}_i)^2 + (x_{i2} - \overline{x}_i)^2 + \ldots + (x_{in_i} - \overline{x}_i)^2.$$

  - Tổng bình phương giữa các nhóm SSG được tính bởi công thức

$$SSG = n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 + \ldots + n_k(\overline{x}_k - \overline{x})^2.$$

  - Tổng bình phương toàn bộ SST được tính bởi công thức

$$SST = (x_{11} - \overline{x})^2 + \ldots + (x_{1n_1} - \overline{x})^2 + \ldots + (x_{k1} - \overline{x})^2 + \ldots + (x_{kn_1} - \overline{x})^2.$$

Ta có $SST = SSW + SSG$.

Qui trình thực hiện

- Bước 3: Tính các phương sai.
  - Phương sai trong nội bộ nhóm MSW được tính bởi công thức

$$MSW = \frac{SSW}{n-k}.$$

  - Phương sai giữa các nhóm MSG được tính bởi công thức

$$MSG = \frac{SSG}{k-1}.$$

- Bước 4: Kiểm định giả thuyết
  - Đặt $F = \frac{MSG}{MSW}$, khi đó F tuân theo phân phối Fisher với k-1 bậc tự do ở tử và n-k bậc ở mẫu.
  - Bác bỏ giả thuyết H_0 tại mức ý nghĩa α nếu $F > F_{k-1,n-k,\alpha}$.

Qui trình thực hiện

- Bước 3: Tính các phương sai.
  - Phương sai trong nội bộ nhóm MSW được tính bởi công thức

$$MSW = \frac{SSW}{n-k}.$$

  - Phương sai giữa các nhóm MSG được tính bởi công thức

$$MSG = \frac{SSG}{k-1}.$$

- Bước 4: Kiểm định giả thuyết
  - Đặt $F = \frac{MSG}{MSW}$, khi đó F tuân theo phân phối Fisher với k-1 bậc tự do ở tử và n-k bậc ở mẫu.
  - Bác bỏ giả thuyết H_0 tại mức ý nghĩa α nếu $F > F_{k-1,n-k,\alpha}$.

Qui trình thực hiện

- Bước 3: Tính các phương sai.
  - Phương sai trong nội bộ nhóm MSW được tính bởi công thức

$$MSW = \frac{SSW}{n-k}.$$

  - Phương sai giữa các nhóm MSG được tính bởi công thức

$$MSG = \frac{SSG}{k-1}.$$

- Bước 4: Kiểm định giả thuyết
  - Đặt $F = \frac{MSG}{MSW}$, khi đó F tuân theo phân phối Fisher với k-1 bậc tự do ở tử và n-k bậc ở mẫu.
  - Bác bỏ giả thuyết H_0 tại mức ý nghĩa α nếu $F > F_{k-1,n-k,\alpha}$.

Bảng phân tích phương sai một nhân tố

Nguồn biến thiên	Tổng bình phương	Bậc tự do (df)	Phương sai (MS)	Tỉ số F	p-giá trị
Giữa các nhóm	SSG	k-1	$MSG = \frac{SSG}{k-1}$	$F = \frac{MSG}{MSW}$	$P(F_{k-1,n-k} > F)$
Nội bộ các nhóm	SSW	n-k	$MSW = \frac{SSW}{n-k}$		
Toàn bộ	SST	n-1			

Bài toán so sánh nhiều trung bình

Bài toán

Bảng sau đây cho ta dữ liệu về điểm trung bình của các sinh viên theo các nhóm có thời gian tự học khác nhau. Giả sử điểm thi của mỗi nhóm tuân theo phân phối chuẩn với phương sai bằng nhau, hãy so sánh xem điểm trung bình của các nhóm có sự khác biệt không tại mức ý nghĩa $\alpha = 5\%$.

Ví dụ

Nhóm I (TG tự học ít)	Nhóm II (TG tự học TB)	Nhóm III (TG tự học nhiều)
5.8	6.0	6.2
6.2	6.6	5.8
5.4	6.1	6.5
6.0	5.8	6.2
5.2	5.9	6.4
5.3	6.0	5.7
5.4	5.9	6.1
5.6	6.0	6.8
6.2	6.7	7.1
5.7	6.5	6.5
5.5	6.3	7.1
6.1	6.1	7.2
6.0	6.8	6.7
5.2	6.4	7.0
6.4	6.8	7.6
5.5	6.6	7.7
5.0	6.4	7.8
5.6	6.2	6.8
6.2	7.1	7.3
6.1	7.0	7.1
5.3	7.2	7.2

Ví dụ

Gọi μ_1, μ_2, μ_3 lần lượt là điểm trung bình của nhóm có thời gian học ít, trung bình, nhiều.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : \exists i \neq j : \mu_i \neq \mu_j, i, j = 1, 2, 3$$

- Bước 1: Tính trung bình từng nhóm và trung bình chung ba nhóm
  - Trung bình mẫu của từng nhóm $\overline{x}_1 = 5.7, \overline{x}_2 = 6.4, \overline{x}_3 = 6.8$.
  - Trung bình của 3 mẫu $\overline{x} = 6.3$.

Ví dụ

- Bước 2: Tính tổng các chênh lệch bình phương
  - $SSW = SS_1 + SS_2 + SS_3 = (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + (n_3 - 1)S_3^2 = 3.34 + 3.56 + 7.1 = 14.$
  - $SSG = n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 + n_3(\overline{x}_3 - \overline{x})^2 = 21(5.7 - 6.3)^2 + 21(6.4 - 6.3)^2 + 21(6.8 - 6.3)^2 = 13.02.$
- Bước 3: Tính các phương sai
  - Phương sai trong nội bộ nhóm $MSW = \frac{SSW}{n-k} = \frac{14}{63-3} = 0.233.$
  - Phương sai giữa các nhóm $MSG = \frac{SSG}{k-1} = \frac{13.02}{3-1} = 6.51.$
- Bước 4: Tính tỉ số $F = \frac{MSG}{MSW} = \frac{6.51}{0.223} = 27.94,$ $F_{k-1,n-k;\alpha} = F_{3-1,63-3;0.05} = 3.15.$

Ví dụ

- Bước 2: Tính tổng các chênh lệch bình phương
  - $SSW = SS_1 + SS_2 + SS_3 = (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + (n_3 - 1)S_3^2 = 3.34 + 3.56 + 7.1 = 14.$
  - $SSG = n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 + n_3(\overline{x}_3 - \overline{x})^2 = 21(5.7 - 6.3)^2 + 21(6.4 - 6.3)^2 + 21(6.8 - 6.3)^2 = 13.02.$
- Bước 3: Tính các phương sai
  - Phương sai trong nội bộ nhóm $MSW = \dfrac{SSW}{n-k} = \dfrac{14}{63-3} = 0.233.$
  - Phương sai giữa các nhóm $MSG = \dfrac{SSG}{k-1} = \dfrac{13.02}{3-1} = 6.51.$
- Bước 4: Tính tỉ số $F = \dfrac{MSG}{MSW} = \dfrac{6.51}{0.223} = 27.94,$ $F_{k-1,n-k;\alpha} = F_{3-1,63-3;0.05} = 3.15.$

Ví dụ

Hãy lập bảng phân tích phương sai và đưa ra kết luận cho bài toán.

Nguồn biến thiên	Tổng bình phương	Bậc tự do (df)	Phương sai (MS)	Tỉ số F	p-giá trị
Giữa các nhóm					
Nội bộ các nhóm					
Toàn bộ					

Thực hiện phân tích ANOVA trong R

Ta sẽ thực hiện ví dụ trên bằng các bước phân tích ANOVA trong R:

- Nhập dữ liệu theo từng nhóm để phân tích:

```
> DiemTB = c(5.8, 6.2, 5.4, 6.0, 5.2, 5.3, 5.4, 5.6, 6.2,
5.7, 5.5, 6.1, 6.0, 5.2, 6.4, 5.5, 5.0, 5.6, 6.2, 6.1, 5.3,
6.0, 6.6, 6.1, 5.8, 5.9, 6.0, 5.9, 6.0, 6.7, 6.5,
6.3, 6.1, 6.8, 6.4, 6.8, 6.6, 6.4, 6.2, 7.1, 7.0, 7.2,
6.2, 5.8, 6.5, 6.2, 6.4, 5.7, 6.1, 6.8, 7.1, 6.5,
7.1, 7.2, 6.7, 7.0, 7.6, 7.7, 7.8, 6.8, 7.3, 7.1, 7.2)
```

- Tạo ra nhóm thứ bậc để phân loại các phần tử trong mẫu dữ liệu chung:

```
> PhanNhom = rep(1:3,each=21)
> PhanNhom = factor(PhanNhom)
```

- Tiến hành phân tích phương sai bằng hàm `anova()`:

```
> anova(lm(DiemTB~PhanNhom))
```

Ví dụ trong R

```
Analysis of Variance Table

Response: DiemTB
           Df Sum Sq Mean Sq F value    Pr(>F)
PhanNhom    2  13.02  6.5100    27.9 2.712e-09 ***
Residuals  60  14.00  0.2333
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Nội dung trình bày

So sánh phương sai của nhiều tổng thể

Bài toán

Giả sử có k tổng thể tuân theo phân phối chuẩn, ta cần so sánh sự bằng nhau của các phương sai của k tổng thể này dựa trên việc kiểm định cặp giả thuyết sau:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2,$$

$$H_1 : \exists i \neq j : \sigma_i^2 \neq \sigma_j^2, i, j = \overline{1, k}.$$

Thực hiện so sánh phương sai của nhiều tổng thể trong R

Hàm trong R để so sánh phương sai của nhiều tổng thể

```
> bartlett.test(x, g)
```

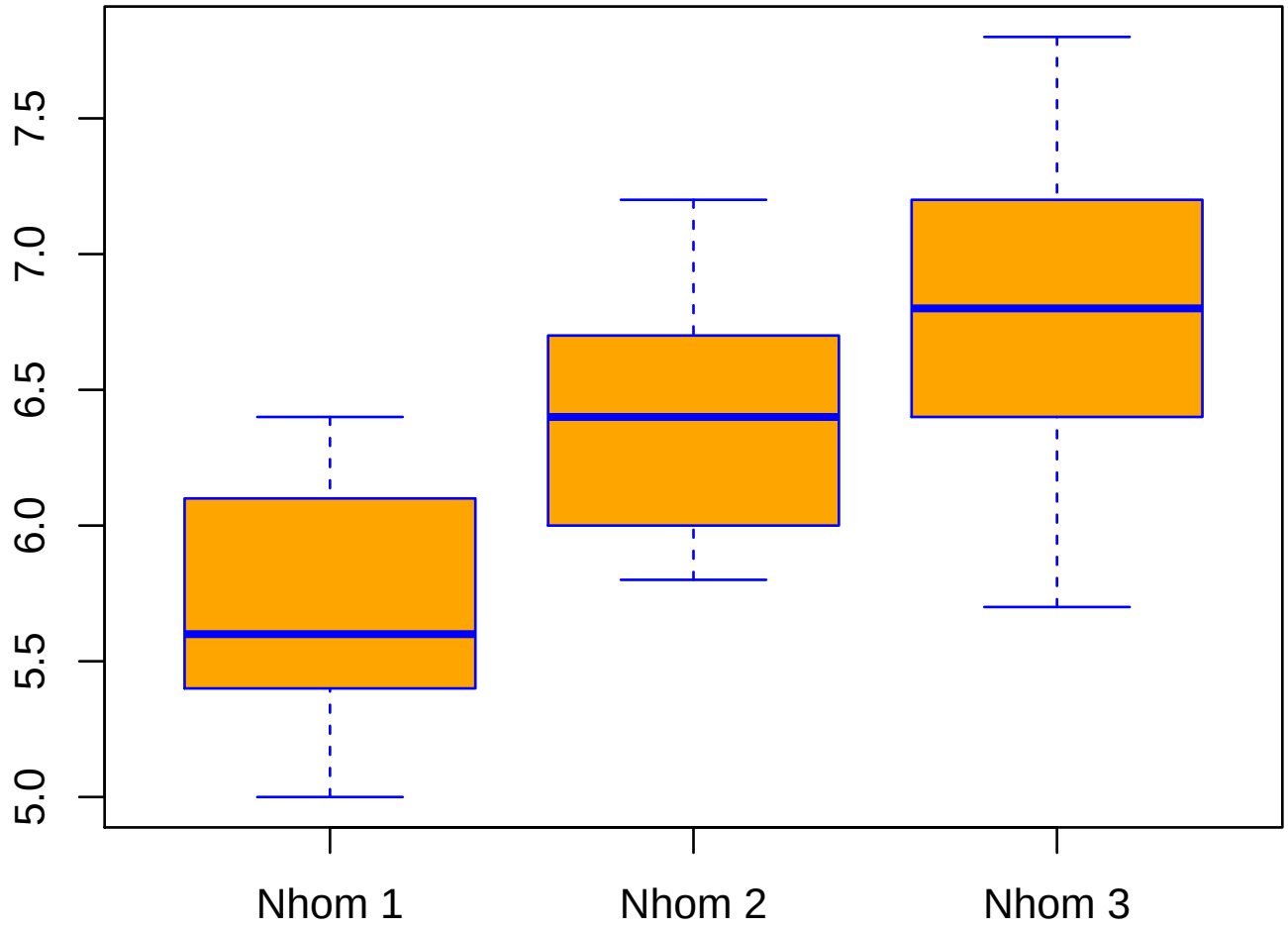
trong đó,

- `x` là véc tơ của dãy giá trị dữ liệu theo từng mẫu;
- `g` là véc tơ chỉ thứ bậc của các giá trị trong mẫu gộp.

Thực hiện so sánh phương sai của nhiều tổng thể trong R

```
> bartlett.test(DiemTB,PhanNhom)

        Bartlett test of homogeneity of variances

data:  DiemTB and PhanNhom
Bartlett's K-squared = 3.6295, df = 2, p-value = 0.1629
```

Minh họa sự bằng nhau phương sai điểm của ba nhóm



Bài toán phân tích sâu One-way ANOVA

Bài toán

Trong bài toán so sánh nhiều trung bình, khi giả thuyết H_0 bị bác bỏ có nghĩa là kết luận trung bình của các tổng thể không bằng nhau. Ta cần phân tích sâu hơn (phân tích sâu ANOVA) để xác định trung bình của tổng thể nào khác tổng thể nào, trung bình của tổng thể nào lớn hơn hay nhỏ hơn.

Ta sẽ tiến hành phân tích sâu ANOVA bằng phương pháp Tukey, phương pháp này còn được gọi là kiểm định HSD. Nội dung của phương pháp này là so sánh từng cặp các trung bình nhóm ở mức cùng ý nghĩa α nào đó cho tất cả các cặp kiểm định có thể để tìm ra những nhóm có trung bình khác nhau.

Bài toán phân tích sâu One-way ANOVA

Bài toán

Trong bài toán so sánh nhiều trung bình, khi giả thuyết H_0 bị bác bỏ có nghĩa là kết luận trung bình của các tổng thể không bằng nhau. Ta cần phân tích sâu hơn (phân tích sâu ANOVA) để xác định trung bình của tổng thể nào khác tổng thể nào, trung bình của tổng thể nào lớn hơn hay nhỏ hơn.

Ta sẽ tiến hành phân tích sâu ANOVA bằng phương pháp Tukey, phương pháp này còn được gọi là kiểm định HSD. Nội dung của phương pháp này là so sánh từng cặp các trung bình nhóm ở mức cùng ý nghĩa α nào đó cho tất cả các cặp kiểm định có thể để tìm ra những nhóm có trung bình khác nhau.

Qui trình thực hiện

- Giả sử cần so sánh trung bình của k tổng thể, khi đó ta cần so sánh trung bình của C_k^2 cặp tổng thể:

$$H_0 : \mu_i = \mu_j; H_1 : \mu_i \neq \mu_j, \forall i \neq j, i, j = \overline{1, k}.$$

- Giá trị tới hạn Tukey được tính theo công thức: $T = q_{\alpha,k,n-k}\sqrt{\frac{MSW}{n_{min}}}$,
  - n_{min} là số quan sát nhỏ nhất trong các mẫu chọn ra quan sát;
  - MSW là phương sai trong nội bộ nhóm;
  - $q_{\alpha,k,n-k}$ là giá trị của phân phối kiểm định Tukey tại mức ý nghĩa α, với bậc tự do k và n-k, n là tổng số quan sát $n = \sum n_i$.
- Tiêu chuẩn quyết định là bác bỏ giả thuyết H_0 khi độ lệch tuyệt đối giữa các cặp trung bình mẫu lớn hơn hay bằng T giới hạn.

Qui trình thực hiện

- Giả sử cần so sánh trung bình của k tổng thể, khi đó ta cần so sánh trung bình của C_k^2 cặp tổng thể:

$$H_0 : \mu_i = \mu_j; H_1 : \mu_i \neq \mu_j, \forall i \neq j, i, j = \overline{1, k}.$$

- Giá trị tới hạn Tukey được tính theo công thức: $T = q_{\alpha,k,n-k}\sqrt{\frac{MSW}{n_{min}}}$,
  - n_{min} là số quan sát nhỏ nhất trong các mẫu chọn ra quan sát;
  - MSW là phương sai trong nội bộ nhóm;
  - $q_{\alpha,k,n-k}$ là giá trị của phân phối kiểm định Tukey tại mức ý nghĩa α, với bậc tự do k và n-k, n là tổng số quan sát $n = \sum n_i$.
- Tiêu chuẩn quyết định là bác bỏ giả thuyết H_0 khi độ lệch tuyệt đối giữa các cặp trung bình mẫu lớn hơn hay bằng T giới hạn.

Qui trình thực hiện

- Giả sử cần so sánh trung bình của k tổng thể, khi đó ta cần so sánh trung bình của C_k^2 cặp tổng thể:

$$H_0 : \mu_i = \mu_j; H_1 : \mu_i \neq \mu_j, \forall i \neq j, i, j = \overline{1, k}.$$

- Giá trị tới hạn Tukey được tính theo công thức: $T = q_{\alpha,k,n-k}\sqrt{\frac{MSW}{n_{min}}}$,
  - n_{min} là số quan sát nhỏ nhất trong các mẫu chọn ra quan sát;
  - MSW là phương sai trong nội bộ nhóm;
  - $q_{\alpha,k,n-k}$ là giá trị của phân phối kiểm định Tukey tại mức ý nghĩa α, với bậc tự do k và n-k, n là tổng số quan sát $n = \sum n_i$.
- Tiêu chuẩn quyết định là bác bỏ giả thuyết H_0 khi độ lệch tuyệt đối giữa các cặp trung bình mẫu lớn hơn hay bằng T giới hạn.

Ví dụ

- Trong tính toán ở ví dụ trước, ta có $k = 3, \alpha = 5\%, n = 63$ và $MSW = 0.233$.
- Giá trị q của phân phối Tukey: $q_{0.05,3,60} = 3.4$.
- Giá trị tới hạn: $T = 3.04\sqrt{\frac{0.233}{21}} = 0.36$.
- Độ lệch tuyệt đối của các cặp trung bình mẫu tính được lần lượt như sau:
  - $|\overline{x}_1 - \overline{x}_2| = |5.7 - 6.4| = 0.7$;
  - $|\overline{x}_1 - \overline{x}_3| = |5.7 - 6.8| = 1.1$;
  - $|\overline{x}_2 - \overline{x}_3| = |6.4 - 6.8| = 0.4$.
- Với T = 0.36, qui tắc bác bỏ H_0 cho ta các quyết định sau:
  - Trung bình tổng thể μ_1 và μ_2 khác nhau vì $|\overline{x}_1 - \overline{x}_2| = 0.7 > T$;
  - Trung bình tổng thể μ_1 và μ_3 khác nhau vì $|\overline{x}_1 - \overline{x}_3| = 1.1 > T$
  - Trung bình tổng thể μ_2 và μ_3 khác nhau vì $|\overline{x}_2 - \overline{x}_3| = 0.4 > T$.

  Do $\overline{x}_1 < \overline{x}_2 < \overline{x}_3$ nên ta suy ra $\mu_1 < \mu_2 < \mu_3$.

Ví dụ

- Trong tính toán ở ví dụ trước, ta có $k = 3, \alpha = 5\%, n = 63$ và $MSW = 0.233$.
- Giá trị q của phân phối Tukey: $q_{0.05,3,60} = 3.4$.
- Giá trị tới hạn: $T = 3.04\sqrt{\frac{0.233}{21}} = 0.36$.
- Độ lệch tuyệt đối của các cặp trung bình mẫu tính được lần lượt như sau:
  - $|\overline{x}_1 - \overline{x}_2| = |5.7 - 6.4| = 0.7$;
  - $|\overline{x}_1 - \overline{x}_3| = |5.7 - 6.8| = 1.1$;
  - $|\overline{x}_2 - \overline{x}_3| = |6.4 - 6.8| = 0.4$.
- Với T = 0.36, qui tắc bác bỏ H_0 cho ta các quyết định sau:
  - Trung bình tổng thể μ_1 và μ_2 khác nhau vì $|\overline{x}_1 - \overline{x}_2| = 0.7 > T$;
  - Trung bình tổng thể μ_1 và μ_3 khác nhau vì $|\overline{x}_1 - \overline{x}_3| = 1.1 > T$
  - Trung bình tổng thể μ_2 và μ_3 khác nhau vì $|\overline{x}_2 - \overline{x}_3| = 0.4 > T$.

  Do $\overline{x}_1 < \overline{x}_2 < \overline{x}_3$ nên ta suy ra $\mu_1 < \mu_2 < \mu_3$.

Ví dụ

- Trong tính toán ở ví dụ trước, ta có $k = 3, \alpha = 5\%, n = 63$ và $MSW = 0.233$.
- Giá trị q của phân phối Tukey: $q_{0.05,3,60} = 3.4$.
- Giá trị tới hạn: $T = 3.04\sqrt{\frac{0.233}{21}} = 0.36$.
- Độ lệch tuyệt đối của các cặp trung bình mẫu tính được lần lượt như sau:
  - $|\overline{x}_1 - \overline{x}_2| = |5.7 - 6.4| = 0.7$;
  - $|\overline{x}_1 - \overline{x}_3| = |5.7 - 6.8| = 1.1$;
  - $|\overline{x}_2 - \overline{x}_3| = |6.4 - 6.8| = 0.4$.
- Với T = 0.36, qui tắc bác bỏ H_0 cho ta các quyết định sau:
  - Trung bình tổng thể μ_1 và μ_2 khác nhau vì $|\overline{x}_1 - \overline{x}_2| = 0.7 > T$;
  - Trung bình tổng thể μ_1 và μ_3 khác nhau vì $|\overline{x}_1 - \overline{x}_3| = 1.1 > T$
  - Trung bình tổng thể μ_2 và μ_3 khác nhau vì $|\overline{x}_2 - \overline{x}_3| = 0.4 > T$.

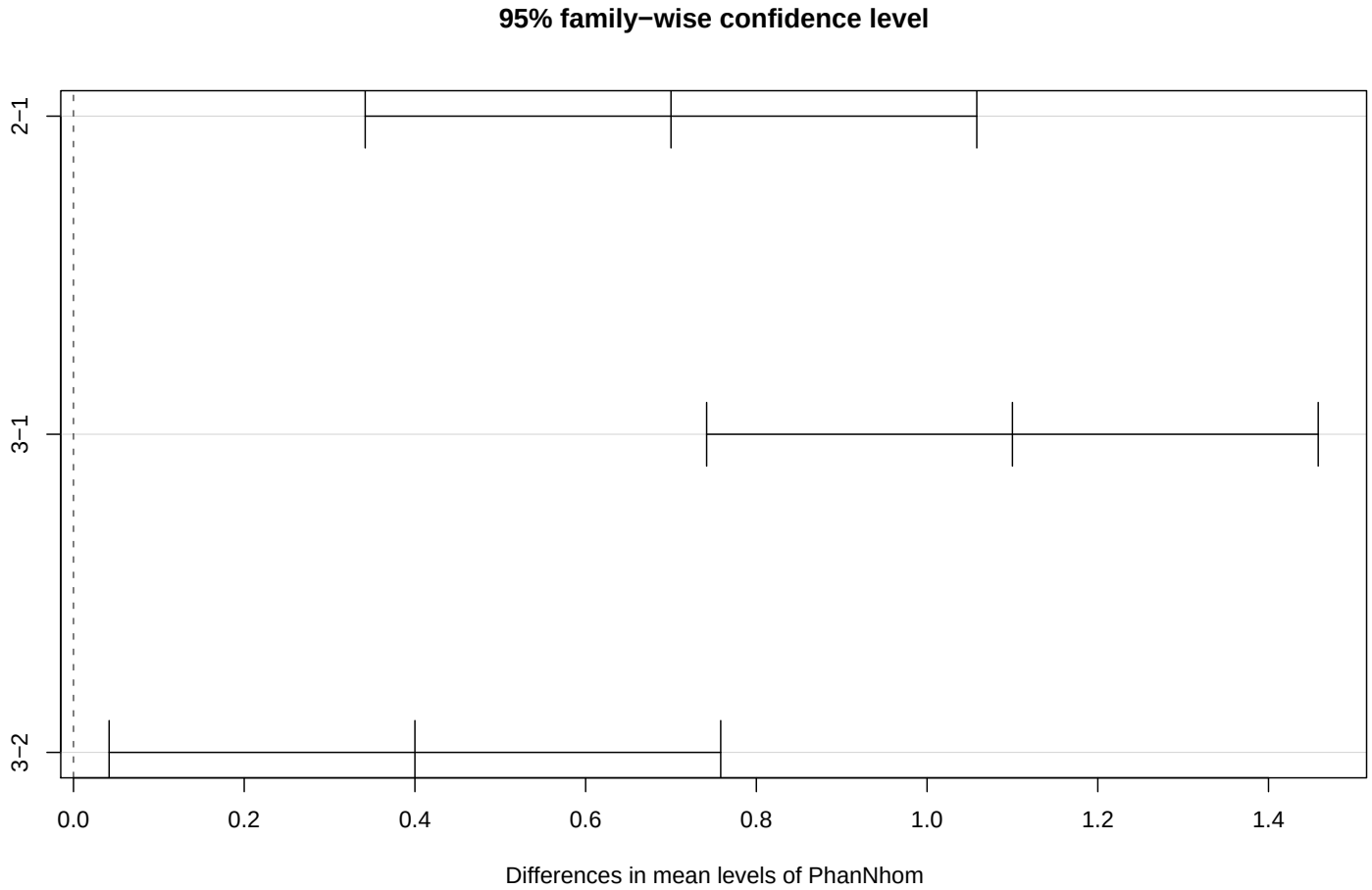
  Do $\overline{x}_1 < \overline{x}_2 < \overline{x}_3$ nên ta suy ra $\mu_1 < \mu_2 < \mu_3$.

Thực hiện phân tích sâu ANOVA trong R bằng hàm TukeyHSD()

```
> TukeyHSD(aov(DiemTB ~ PhanNhom))

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = DiemTB ~ PhanNhom)

$PhanNhom
    diff        lwr       upr     p adj
2-1  0.7 0.34174965 1.0582504 0.0000468
3-1  1.1 0.74174965 1.4582504 0.0000000
3-2  0.4 0.04174965 0.7582504 0.0251354
```

Minh họa sự khác biệt giữa các trung bình

```
> plot(TukeyHSD(aov(DiemTB ~ PhanNhom)))
```



Nội dung trình bày

Bài toán phân tích phương sai hai yếu tố (một quan sát)

- Phân tích phương sai hai yếu tố (Two-way ANOVA) xem xét cùng một lúc hai yếu tố nguyên nhân (dưới dạng dữ liệu định tính) ảnh hưởng đến yếu tố kết quả đang nghiên cứu (dưới dạng dữ liệu định lượng).
- Giả sử ta đang nghiên cứu ảnh hưởng của hai yếu tố nguyên nhân định tính đến một yếu tố kết quả định lượng nào đó. Theo yếu tố nguyên nhân thứ nhất ta sắp xếp các đơn vị mẫu thành K nhóm. Theo yếu tố nguyên nhân thứ hai ta sắp xếp các đơn vị mẫu nghiên cứu thành H khối.
- Sắp xếp đồng thời các đơn vị mẫu theo hai yếu tố nguyên nhân ta sẽ có bảng kết hợp gồm K cột và H dòng với $K \times H$ ô dữ liệu.

Bài toán phân tích phương sai hai yếu tố (một quan sát)

- Phân tích phương sai hai yếu tố (Two-way ANOVA) xem xét cùng một lúc hai yếu tố nguyên nhân (dưới dạng dữ liệu định tính) ảnh hưởng đến yếu tố kết quả đang nghiên cứu (dưới dạng dữ liệu định lượng).
- Giả sử ta đang nghiên cứu ảnh hưởng của hai yếu tố nguyên nhân định tính đến một yếu tố kết quả định lượng nào đó. Theo yếu tố nguyên nhân thứ nhất ta sắp xếp các đơn vị mẫu thành K nhóm. Theo yếu tố nguyên nhân thứ hai ta sắp xếp các đơn vị mẫu nghiên cứu thành H khối.
- Sắp xếp đồng thời các đơn vị mẫu theo hai yếu tố nguyên nhân ta sẽ có bảng kết hợp gồm K cột và H dòng với $K \times H$ ô dữ liệu.

Bài toán phân tích phương sai hai yếu tố (một quan sát)

- Phân tích phương sai hai yếu tố (Two-way ANOVA) xem xét cùng một lúc hai yếu tố nguyên nhân (dưới dạng dữ liệu định tính) ảnh hưởng đến yếu tố kết quả đang nghiên cứu (dưới dạng dữ liệu định lượng).
- Giả sử ta đang nghiên cứu ảnh hưởng của hai yếu tố nguyên nhân định tính đến một yếu tố kết quả định lượng nào đó. Theo yếu tố nguyên nhân thứ nhất ta sắp xếp các đơn vị mẫu thành K nhóm. Theo yếu tố nguyên nhân thứ hai ta sắp xếp các đơn vị mẫu nghiên cứu thành H khối.
- Sắp xếp đồng thời các đơn vị mẫu theo hai yếu tố nguyên nhân ta sẽ có bảng kết hợp gồm K cột và H dòng với $K \times H$ ô dữ liệu.

Bảng dữ liệu

Khối (dòng)	Nhóm (cột)			
	1	2	...	K
1	x_{11}	x_{21}	...	x_{K1}
2	x_{12}	x_{22}	...	x_{K2}
⋮	...	...	...	...
H	x_{1H}	x_{2H}	...	x_{KH}

Ví dụ

Ta cần kiểm định cặp giả thuyết H_0, H_1 theo hai trường hợp:

- H_0 : Trung bình của K tổng thể tương ứng với K nhóm mẫu là bằng nhau; H_1 : Tồn tại ít nhất hai tổng thể có trung bình không bằng nhau.
- H_0 : Trung bình của H tổng thể tương ứng với H khối mẫu là bằng nhau; H_1 : Tồn tại ít nhất hai tổng thể có trung bình không bằng nhau.

Để thực hiện các trường hợp kiểm định trên ta thực hiện theo các bước sau:

Qui trình thực hiện

- Bước 1: Tính các trung bình:
  - Trung bình mẫu của từng nhóm (cột) được tính theo công thức:

$$\overline{x}_{i*} = \frac{x_{i1} + x_{i2} + \ldots + x_{iH}}{H}, \quad (i = 1, 2, \ldots, K).$$

  - Trung bình mẫu của từng khối (dòng) được tính theo công thức:

$$\overline{x}_{*j} = \frac{x_{1j} + x_{2j} + \ldots + x_{Kj}}{K}, \quad (j = 1, 2, \ldots, H).$$

  - Trung bình mẫu của toàn bộ mẫu quan sát:

$$\overline{x} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{H} x_{ij}}{n} = \frac{\sum_{i=1}^{K} \overline{x}_i}{K} = \frac{\sum_{j=1}^{H} \overline{x}_j}{H}.$$

Qui trình thực hiện

- Bước 1: Tính các trung bình:
  - Trung bình mẫu của từng nhóm (cột) được tính theo công thức:

$$\overline{x}_{i*} = \frac{x_{i1} + x_{i2} + \ldots + x_{iH}}{H}, \quad (i = 1, 2, \ldots, K).$$

  - Trung bình mẫu của từng khối (dòng) được tính theo công thức:

$$\overline{x}_{*j} = \frac{x_{1j} + x_{2j} + \ldots + x_{Kj}}{K}, \quad (j = 1, 2, \ldots, H).$$

  - Trung bình mẫu của toàn bộ mẫu quan sát:

$$\overline{x} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{H} x_{ij}}{n} = \frac{\sum_{i=1}^{K} \overline{x}_i}{K} = \frac{\sum_{j=1}^{H} \overline{x}_j}{H}.$$

Qui trình thực hiện

- Bước 1: Tính các trung bình:
  - Trung bình mẫu của từng nhóm (cột) được tính theo công thức:

$$\overline{x}_{i*} = \frac{x_{i1} + x_{i2} + \ldots + x_{iH}}{H}, \quad (i = 1, 2, \ldots, K).$$

  - Trung bình mẫu của từng khối (dòng) được tính theo công thức:

$$\overline{x}_{*j} = \frac{x_{1j} + x_{2j} + \ldots + x_{Kj}}{K}, \quad (j = 1, 2, \ldots, H).$$

  - Trung bình mẫu của toàn bộ mẫu quan sát:

$$\overline{x} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{H} x_{ij}}{n} = \frac{\sum_{i=1}^{K} \overline{x}_i}{K} = \frac{\sum_{j=1}^{H} \overline{x}_j}{H}.$$

Qui trình thực hiện

- Bước 2: Tính tổng các chênh lệch bình phương
  - Tổng bình phương giữa các nhóm được tính bởi:

$$SSG = H \sum_{i=1}^{K} (\overline{x}_{i*} - \overline{x})^2.$$

  - Tổng bình phương giữa các khối được tính bởi:

$$SSB = K \sum_{j=1}^{H} (\overline{x}_{*j} - \overline{x})^2.$$

  - Tổng bình phương phần dư được tính bởi:

$$SSE = \sum_{i=1}^{K} \sum_{j=1}^{H} (x_{ij} - \overline{x}_{i*} - \overline{x}_{*j} + \overline{x})^2.$$

  - Tổng bình phương toàn phần được tính theo công thức

$$SST = \sum_{i=1}^{K} \sum_{j=1}^{H} (x_{ij} - \overline{x})^2.$$

Qui trình thực hiện

- Bước 2: Tính tổng các chênh lệch bình phương
  - Tổng bình phương giữa các nhóm được tính bởi:

$$SSG = H \sum_{i=1}^{K} (\overline{x}_{i*} - \overline{x})^2.$$

  - Tổng bình phương giữa các khối được tính bởi:

$$SSB = K \sum_{j=1}^{H} (\overline{x}_{*j} - \overline{x})^2.$$

  - Tổng bình phương phần dư được tính bởi:

$$SSE = \sum_{i=1}^{K} \sum_{j=1}^{H} (x_{ij} - \overline{x}_{i*} - \overline{x}_{*j} + \overline{x})^2.$$

  - Tổng bình phương toàn phần được tính theo công thức

$$SST = \sum_{i=1}^{K} \sum_{j=1}^{H} (x_{ij} - \overline{x})^2.$$

Qui trình thực hiện

- Bước 2: Tính tổng các chênh lệch bình phương
  - Tổng bình phương giữa các nhóm được tính bởi:

$$SSG = H \sum_{i=1}^{K} (\overline{x}_{i*} - \overline{x})^2.$$

  - Tổng bình phương giữa các khối được tính bởi:

$$SSB = K \sum_{j=1}^{H} (\overline{x}_{*j} - \overline{x})^2.$$

  - Tổng bình phương phần dư được tính bởi:

$$SSE = \sum_{i=1}^{K} \sum_{j=1}^{H} (x_{ij} - \overline{x}_{i*} - \overline{x}_{*j} + \overline{x})^2.$$

  - Tổng bình phương toàn phần được tính theo công thức

$$SST = \sum_{i=1}^{K} \sum_{j=1}^{H} (x_{ij} - \overline{x})^2.$$

Qui trình thực hiện

- Bước 2: Tính tổng các chênh lệch bình phương
  - Tổng bình phương giữa các nhóm được tính bởi:

$$SSG = H \sum_{i=1}^{K} (\overline{x}_{i*} - \overline{x})^2.$$

  - Tổng bình phương giữa các khối được tính bởi:

$$SSB = K \sum_{j=1}^{H} (\overline{x}_{*j} - \overline{x})^2.$$

  - Tổng bình phương phần dư được tính bởi:

$$SSE = \sum_{i=1}^{K} \sum_{j=1}^{H} (x_{ij} - \overline{x}_{i*} - \overline{x}_{*j} + \overline{x})^2.$$

  - Tổng bình phương toàn phần được tính theo công thức

$$SST = \sum_{i=1}^{K} \sum_{j=1}^{H} (x_{ij} - \overline{x})^2.$$

Qui trình thực hiện

- Bước 3: Tính các phương sai.
  - Phương sai giữa các nhóm: $MSG = \dfrac{SSG}{K-1}$.
  - Phương sai giữa các khối: $MSB = \dfrac{SSB}{H-1}$.
  - Phương sai phần dư: $MSE = \dfrac{SSE}{(K-1)(H-1)}$.
- Bước 4: Kiểm định giả thuyết về ảnh hưởng của yếu tố nguyên nhân thứ nhất và thứ hai đến yếu tố kết quả tương ứng bằng các tỉ số: $F_1 = \dfrac{MSG}{MSE}$ và $F_2 = \dfrac{MSB}{MSE}$

Qui trình thực hiện

- Bước 3: Tính các phương sai.
  - Phương sai giữa các nhóm: $MSG = \dfrac{SSG}{K-1}$.
  - Phương sai giữa các khối: $MSB = \dfrac{SSB}{H-1}$.
  - Phương sai phần dư: $MSE = \dfrac{SSE}{(K-1)(H-1)}$.
- Bước 4: Kiểm định giả thuyết về ảnh hưởng của yếu tố nguyên nhân thứ nhất và thứ hai đến yếu tố kết quả tương ứng bằng các tỉ số: $F_1 = \dfrac{MSG}{MSE}$ và $F_2 = \dfrac{MSB}{MSE}$

Qui trình thực hiện

- Bước 3: Tính các phương sai.
  - Phương sai giữa các nhóm: $MSG = \dfrac{SSG}{K-1}$.
  - Phương sai giữa các khối: $MSB = \dfrac{SSB}{H-1}$.
  - Phương sai phần dư: $MSE = \dfrac{SSE}{(K-1)(H-1)}$.
- Bước 4: Kiểm định giả thuyết về ảnh hưởng của yếu tố nguyên nhân thứ nhất và thứ hai đến yếu tố kết quả tương ứng bằng các tỉ số: $F_1 = \dfrac{MSG}{MSE}$ và $F_2 = \dfrac{MSB}{MSE}$

Qui trình thực hiện

- Bước 3: Tính các phương sai.
  - Phương sai giữa các nhóm: $MSG = \dfrac{SSG}{K-1}$.
  - Phương sai giữa các khối: $MSB = \dfrac{SSB}{H-1}$.
  - Phương sai phần dư: $MSE = \dfrac{SSE}{(K-1)(H-1)}$.
- Bước 4: Kiểm định giả thuyết về ảnh hưởng của yếu tố nguyên nhân thứ nhất và thứ hai đến yếu tố kết quả tương ứng bằng các tỉ số: $F_1 = \dfrac{MSG}{MSE}$ và $F_2 = \dfrac{MSB}{MSE}$

Qui trình thực hiện

- Bước 5: Qui tắc bác bỏ giả thuyết H_0 theo mỗi trường hợp như sau:
  - Ở mức ý nghĩa α, giả thuyết H_0 cho rằng trung bình của K tổng thể theo yếu tố nguyên nhân thứ nhất (cột) bằng nhau bị bác bỏ khi: $F_1 > F_{K-1,(K-1)(H-1),\alpha}$.
  - Ở mức ý nghĩa α, giả thuyết H_0 cho rằng trung bình của H tổng thể theo yếu tố nguyên nhân thứ hai (dòng) bằng nhau bị bác bỏ khi: $F_2 > F_{H-1,(K-1)(H-1),\alpha}$.

Qui trình thực hiện

- Bước 5: Qui tắc bác bỏ giả thuyết H_0 theo mỗi trường hợp như sau:
  - Ở mức ý nghĩa α, giả thuyết H_0 cho rằng trung bình của K tổng thể theo yếu tố nguyên nhân thứ nhất (cột) bằng nhau bị bác bỏ khi: $F_1 > F_{K-1,(K-1)(H-1),\alpha}$.
  - Ở mức ý nghĩa α, giả thuyết H_0 cho rằng trung bình của H tổng thể theo yếu tố nguyên nhân thứ hai (dòng) bằng nhau bị bác bỏ khi: $F_2 > F_{H-1,(K-1)(H-1),\alpha}$.

Bảng phân tích phương sai hai nhân tố (một quan sát)

Nguồn biến thiên	Tổng bình phương	Bậc tự do (df)	Phương sai (MS)	Tỉ số F
Giữa các nhóm	SSG	K-1	$MSG = \frac{SSG}{K-1}$	$F_1 = \frac{MSG}{MSE}$
Giữa các khối	SSB	H-1	$MSB = \frac{SSB}{H-1}$	$F_2 = \frac{MSB}{MSE}$
Phần dư	SSE	(K-1)(H-1)	$MSE = \frac{SSE}{(K-1)(H-1)}$	
Tổng cộng	SST	n-1		

Ví dụ

Bài toán

Để xét xem quãng đường mà ô tô đi được có ảnh hưởng bởi loại xe và tài xế hay không người ta thử nghiệm ba loại xe A, B, C và tài xế phân theo một số độ tuổi. Bảng sau cho số km đi được trên mỗi lít xăng:

	Hiệu xe		
Loại tài xế	Xe loại A	Xe loại B	Xe loại C
$\leqslant$ 25 tuổi	25.1	23.9	26.0
26-35 tuổi	24.7	23.7	25.4
36-45 tuổi	26.0	24.4	25.8
46-55 tuổi	24.3	23.3	24.4
56-65 tuổi	23.9	23.6	24.2

Tại mức ý nghĩa $\alpha = 5\%$, hãy kiểm định xem quãng đường mà ô tô đi được có bị ảnh hưởng bởi loại xe và quãng đường đi được có ảnh hưởng bởi loại tài xế hay không?

Ví dụ

Lời giải: Cặp giả thuyết cần kiểm định:

- Cặp giả thuyết thứ nhất:

- Cặp giả thuyết thứ hai:

Kết quả kiểm định

```
> QuangDuong = scan()
25.1 24.7 26.0 24.3 23.9 23.9 23.7 24.4 23.3 23.6 26.0 25.4 25.8 24.4 24.2
> PLXe = gl(3,5)
> PLTuoi = gl(5,1,length=15)
> anova(lm(QuangDuong~PLXe+PLTuoi))

Analysis of Variance Table

Response: QuangDuong
          Df Sum Sq Mean Sq F value    Pr(>F)
PLXe       2  5.124  2.5620 20.5783 0.0007015 ***
PLTuoi     4  4.944  1.2360  9.9277 0.0034222 **
Residuals  8  0.996  0.1245
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Bảng phân tích phương sai

Nguồn biến thiên	Tổng bình phương	Bậc tự do (df)	Phương sai (MS)	Tỉ số F
Giữa các nhóm	SSG =	K-1 =	$MSG =$	$F_1 =$
Giữa các khối	SSB =	H-1 =	$MSB =$	$F_2 =$
Phần dư	SSE =	(K-1)(H-1) =	$MSE =$	
Tổng cộng	SST =	n-1 =		

- p-giá trị cho cặp giả thuyết thứ nhất:

- p-giá trị cho cặp giả thuyết thứ hai:

Ví dụ

- Kết luận cho cặp giả thuyết thứ nhất:

- Kết luận cho cặp giả thuyết thứ hai:

Nội dung trình bày

Phân tích phương sai hai nhân tố (nhiều quan sát trong một ô)

Để tăng tính chính xác khi kết luận về ảnh hưởng của hai yếu tố nguyên nhân đến yếu tố kết quả của mẫu cho tổng thể, ta tăng cỡ mẫu trong điều kiện cho phép. Gọi L là số quan sát trong một ô, ta có dạng tổng quát của L quan sát trong một ô như sau:

Khối (dòng)	Nhóm (cột)			
	1	2	...	K
1	$x_{111}x_{112}\ldots x_{11L}$	$x_{211}x_{212}\ldots x_{21L}$	...	$x_{K11}x_{K12}\ldots x_{K1L}$
2	$x_{121}x_{122}\ldots x_{12L}$	$x_{221}x_{222}\ldots x_{22L}$	...	$x_{K21}x_{K22}x_{K2L}$
⋮	...	...	...	...
H	$x_{1H1}x_{1H2}\ldots x_{1HL}$	$x_{2H1}x_{2H2}\ldots x_{2HL}$	...	$x_{KH1}x_{KH2}\ldots x_{KHL}$

Tương tác trong bài toán phân tích phương sai

Definition

Tương tác xảy ra giữa các yếu tố nguyên nhân khi ảnh hưởng của một yếu tố nguyên nhân này thay đổi theo các mức của yếu tố nguyên nhân còn lại.

Nhận xét: Khi có sự tương tác giữa dòng và cột, thì hiệu trung bình giữa các mức của yếu tố nguyên nhân cột (dòng) thay đổi hay phụ thuộc theo các mức của yếu tố nguyên nhân dòng (cột).

Các bài toán kiểm định

Ta cần kiểm định cặp giả thuyết H_0, H_1 theo ba trường hợp:

- H_0 : Trung bình của K tổng thể theo yếu nguyên nhân thứ nhất bằng nhau; H_1 : Tồn tại ít nhất hai tổng thể có trung bình không bằng nhau.
- H_0 : Trung bình của H tổng thể theo yếu tố nguyên nhân thứ hai bằng nhau; H_1 : Tồn tại ít nhất hai tổng thể có trung bình không bằng nhau.
- H_0 : Không có sự tương tác giữa yếu tố nguyên nhân thứ nhất và yếu tố thứ hai; H_1 : Có sự tương tác giữa yếu tố nguyên nhân thứ nhất và yếu tố thứ hai.

Để thực hiện các trường hợp kiểm định trên ta thực hiện theo các bước sau:

Qui trình thực hiện

- Bước 1: Tính các trung bình:
  - Trung bình của từng nhóm (cột) được tính theo công thức:

$$\overline{x}_{i*} = \frac{\sum_{j=1}^{H}\sum_{s=1}^{L} x_{ijs}}{H \times L}, \quad (i = 1, 2, \ldots, K).$$

  - Trung bình mẫu của từng khối (dòng) được tính theo công thức:

$$\overline{x}_{*j} = \frac{\sum_{i=1}^{K}\sum_{s=1}^{L} x_{ijs}}{K \times L}, \quad (j = 1, 2, \ldots, H).$$

  - Trung bình mẫu của từng ô được tính bởi:

$$\overline{x}_{ij} = \frac{\sum_{s=1}^{L} x_{ijs}}{L},$$

  - Trung bình mẫu của toàn bộ mẫu quan sát:

$$\overline{x} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{H}\sum_{s=1}^{L} x_{ijs}}{K \times H \times L}.$$

Qui trình thực hiện

- Bước 1: Tính các trung bình:
  - Trung bình của từng nhóm (cột) được tính theo công thức:

$$\overline{x}_{i*} = \frac{\sum_{j=1}^{H}\sum_{s=1}^{L} x_{ijs}}{H \times L}, \quad (i = 1, 2, \ldots, K).$$

  - Trung bình mẫu của từng khối (dòng) được tính theo công thức:

$$\overline{x}_{*j} = \frac{\sum_{i=1}^{K}\sum_{s=1}^{L} x_{ijs}}{K \times L}, \quad (j = 1, 2, \ldots, H).$$

  - Trung bình mẫu của từng ô được tính bởi:

$$\overline{x}_{ij} = \frac{\sum_{s=1}^{L} x_{ijs}}{L},$$

  - Trung bình mẫu của toàn bộ mẫu quan sát:

$$\overline{x} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{H}\sum_{s=1}^{L} x_{ijs}}{K \times H \times L}.$$

Qui trình thực hiện

- Bước 1: Tính các trung bình:
  - Trung bình của từng nhóm (cột) được tính theo công thức:

$$\overline{x}_{i*} = \frac{\sum_{j=1}^{H}\sum_{s=1}^{L} x_{ijs}}{H \times L}, \quad (i = 1, 2, \ldots, K).$$

  - Trung bình mẫu của từng khối (dòng) được tính theo công thức:

$$\overline{x}_{*j} = \frac{\sum_{i=1}^{K}\sum_{s=1}^{L} x_{ijs}}{K \times L}, \quad (j = 1, 2, \ldots, H).$$

  - Trung bình mẫu của từng ô được tính bởi:

$$\overline{x}_{ij} = \frac{\sum_{s=1}^{L} x_{ijs}}{L},$$

  - Trung bình mẫu của toàn bộ mẫu quan sát:

$$\overline{x} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{H}\sum_{s=1}^{L} x_{ijs}}{K \times H \times L}.$$

Qui trình thực hiện

- Bước 1: Tính các trung bình:
  - Trung bình của từng nhóm (cột) được tính theo công thức:

$$\overline{x}_{i*} = \frac{\sum_{j=1}^{H}\sum_{s=1}^{L} x_{ijs}}{H \times L}, \quad (i = 1, 2, \ldots, K).$$

  - Trung bình mẫu của từng khối (dòng) được tính theo công thức:

$$\overline{x}_{*j} = \frac{\sum_{i=1}^{K}\sum_{s=1}^{L} x_{ijs}}{K \times L}, \quad (j = 1, 2, \ldots, H).$$

  - Trung bình mẫu của từng ô được tính bởi:

$$\overline{x}_{ij} = \frac{\sum_{s=1}^{L} x_{ijs}}{L},$$

  - Trung bình mẫu của toàn bộ mẫu quan sát:

$$\overline{x} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{H}\sum_{s=1}^{L} x_{ijs}}{K \times H \times L}.$$

Qui trình thực hiện

- Bước 2: Tính tổng các chênh lệch bình phương
  - Tổng bình phương giữa các nhóm được tính bởi:

$$SSG = HL \sum_{i=1}^{K} (\overline{x}_{i*} - \overline{x})^2.$$

  - Tổng bình phương giữa các khối được tính bởi:

$$SSB = KL \sum_{j=1}^{H} (\overline{x}_{*j} - \overline{x})^2.$$

  - Tổng bình phương giữa các ô được tính bởi:

$$SSI = L \sum_{i=1}^{K} \sum_{j=1}^{H} (\overline{x}_{ij} - \overline{x}_{i*} - \overline{x}_{*j} + \overline{x})^2.$$

Qui trình thực hiện

- Bước 2: Tính tổng các chênh lệch bình phương
  - Tổng bình phương giữa các nhóm được tính bởi:

$$SSG = HL \sum_{i=1}^{K} (\overline{x}_{i*} - \overline{x})^2.$$

  - Tổng bình phương giữa các khối được tính bởi:

$$SSB = KL \sum_{j=1}^{H} (\overline{x}_{*j} - \overline{x})^2.$$

  - Tổng bình phương giữa các ô được tính bởi:

$$SSI = L \sum_{i=1}^{K} \sum_{j=1}^{H} (\overline{x}_{ij} - \overline{x}_{i*} - \overline{x}_{*j} + \overline{x})^2.$$

Qui trình thực hiện

- Bước 2: Tính tổng các chênh lệch bình phương
  - Tổng bình phương giữa các nhóm được tính bởi:

$$SSG = HL \sum_{i=1}^{K} (\overline{x}_{i*} - \overline{x})^2.$$

  - Tổng bình phương giữa các khối được tính bởi:

$$SSB = KL \sum_{j=1}^{H} (\overline{x}_{*j} - \overline{x})^2.$$

  - Tổng bình phương giữa các ô được tính bởi:

$$SSI = L \sum_{i=1}^{K} \sum_{j=1}^{H} (\overline{x}_{ij} - \overline{x}_{i*} - \overline{x}_{*j} + \overline{x})^2.$$

Qui trình thực hiện

- Bước 2: Tính tổng các chênh lệch bình phương
  - Tổng bình phương phần dư được tính bởi:

$$SSE = \sum_{i=1}^{K}\sum_{j=1}^{H}\sum_{s=1}^{L}(x_{ijs} - \overline{x}_{ij})^2.$$

  - Tổng bình phương toàn phần được tính theo công thức

$$SST = \sum_{i=1}^{K}\sum_{j=1}^{H}\sum_{s=1}^{L}(x_{ijs} - \overline{x})^2.$$

Ta có thể chứng minh được rằng: $SST = SSG + SSB + SSI + SSE$.

Qui trình thực hiện

- Bước 2: Tính tổng các chênh lệch bình phương
  - Tổng bình phương phần dư được tính bởi:

$$SSE = \sum_{i=1}^{K}\sum_{j=1}^{H}\sum_{s=1}^{L}(x_{ijs} - \overline{x}_{ij})^2.$$

  - Tổng bình phương toàn phần được tính theo công thức

$$SST = \sum_{i=1}^{K}\sum_{j=1}^{H}\sum_{s=1}^{L}(x_{ijs} - \overline{x})^2.$$

Ta có thể chứng minh được rằng: $SST = SSG + SSB + SSI + SSE$.

Qui trình thực hiện

- Bước 3: Tính các phương sai.
  - Phương sai giữa các nhóm: $MSG = \dfrac{SSG}{K-1}$.
  - Phương sai giữa các khối: $MSB = \dfrac{SSB}{H-1}$.
  - Phương sai giữa các ô: $MSI = \dfrac{SSI}{(K-1)(H-1)}$.
  - Phương sai phần dư: $MSE = \dfrac{SSE}{K \times H \times (L-1)}$.
- Bước 4: Kiểm định giả thuyết về ảnh hưởng của yếu tố nguyên nhân thứ nhất (cột) và nguyên nhân thứ hai (dòng), tương tác hai yếu tố đến yếu tố kết quả tương ứng bằng các tỉ số:

$$F_1 = \frac{MSG}{MSE}, \quad F_2 = \frac{MSB}{MSE}, \quad F_3 = \frac{MSI}{MSE}.$$

Qui trình thực hiện

- Bước 3: Tính các phương sai.
  - Phương sai giữa các nhóm: $MSG = \dfrac{SSG}{K-1}$.
  - Phương sai giữa các khối: $MSB = \dfrac{SSB}{H-1}$.
  - Phương sai giữa các ô: $MSI = \dfrac{SSI}{(K-1)(H-1)}$.
  - Phương sai phần dư: $MSE = \dfrac{SSE}{K \times H \times (L-1)}$.
- Bước 4: Kiểm định giả thuyết về ảnh hưởng của yếu tố nguyên nhân thứ nhất (cột) và nguyên nhân thứ hai (dòng), tương tác hai yếu tố đến yếu tố kết quả tương ứng bằng các tỉ số:

$$F_1 = \frac{MSG}{MSE}, \quad F_2 = \frac{MSB}{MSE}, \quad F_3 = \frac{MSI}{MSE}.$$

Qui trình thực hiện

- Bước 5: Qui tắc bác bỏ giả thuyết H_0 theo mỗi trường hợp như sau:
  - Đối với F_1, ở mức ý nghĩa α, giả thuyết H_0 cho rằng trung bình của K tổng thể theo yếu tố nguyên nhân thứ nhất (cột) bằng nhau bị bác bỏ khi: $F_1 > F_{K-1,KH(L-1),\alpha}$.
  - Đối với F_2, ở mức ý nghĩa α, giả thuyết H_0 cho rằng trung bình của H tổng thể theo yếu tố nguyên nhân thứ hai (dòng) bằng nhau bị bác bỏ khi: $F_2 > F_{H-1,KH(L-1),\alpha}$.
  - Đối với F_3, ở mức ý nghĩa α, giả thuyết H_0 cho rằng không có tác động qua lại giữa yếu tố thứ nhất (cột) và yếu tố thứ hai (dòng) bị bác bỏ khi: $F_3 > F_{(K-1)(H-1)),KH(L-1),\alpha}$.

Qui trình thực hiện

- Bước 5: Qui tắc bác bỏ giả thuyết H_0 theo mỗi trường hợp như sau:
  - Đối với F_1, ở mức ý nghĩa α, giả thuyết H_0 cho rằng trung bình của K tổng thể theo yếu tố nguyên nhân thứ nhất (cột) bằng nhau bị bác bỏ khi: $F_1 > F_{K-1,KH(L-1),\alpha}$.
  - Đối với F_2, ở mức ý nghĩa α, giả thuyết H_0 cho rằng trung bình của H tổng thể theo yếu tố nguyên nhân thứ hai (dòng) bằng nhau bị bác bỏ khi: $F_2 > F_{H-1,KH(L-1),\alpha}$.
  - Đối với F_3, ở mức ý nghĩa α, giả thuyết H_0 cho rằng không có tác động qua lại giữa yếu tố thứ nhất (cột) và yếu tố thứ hai (dòng) bị bác bỏ khi: $F_3 > F_{(K-1)(H-1)),KH(L-1),\alpha}$.

Qui trình thực hiện

- Bước 5: Qui tắc bác bỏ giả thuyết H_0 theo mỗi trường hợp như sau:
  - Đối với F_1, ở mức ý nghĩa α, giả thuyết H_0 cho rằng trung bình của K tổng thể theo yếu tố nguyên nhân thứ nhất (cột) bằng nhau bị bác bỏ khi: $F_1 > F_{K-1,KH(L-1),\alpha}$.
  - Đối với F_2, ở mức ý nghĩa α, giả thuyết H_0 cho rằng trung bình của H tổng thể theo yếu tố nguyên nhân thứ hai (dòng) bằng nhau bị bác bỏ khi: $F_2 > F_{H-1,KH(L-1),\alpha}$.
  - Đối với F_3, ở mức ý nghĩa α, giả thuyết H_0 cho rằng không có tác động qua lại giữa yếu tố thứ nhất (cột) và yếu tố thứ hai (dòng) bị bác bỏ khi: $F_3 > F_{(K-1)(H-1)),KH(L-1),\alpha}$.

Bảng phân tích phương sai hai nhân tố (nhiều quan sát)

Nguồn biến thiên	Tổng bình phương	Bậc tự do (df)	Phương sai (MS)	Tỉ số F
Giữa các khối	SSB	K-1	MSB	F_2
Giữa các nhóm	SSG	H-1	MSG	F_1
Tương tác giữa hai yếu tố	SSI	(K-1)(H-1)	MSI	F_3
Phần dư	SSE	KH(L-1)	MSE	
Tổng cộng	SST	KHL-1		

Ví dụ

Xét thời gian tự học và mức độ yêu thích ngành học đối với điểm trung bình của sinh viên.
Các cặp giả thuyết H_0 được đặt ra:

1. Điểm trung bình học tập của sinh viên có thời gian học tập khác nhau đều bằng nhau;
2. Điểm trung bình học tập của sinh viên có mức độ yêu thích ngành học khác nhau đều bằng nhau;
3. Không có ảnh hưởng tương tác giữa thời gian tự học và mức độ yêu thích ngành học của sinh viên.

Ví dụ

Xét thời gian tự học và mức độ yêu thích ngành học đối với điểm trung bình của sinh viên.
Các cặp giả thuyết H_0 được đặt ra:

1. Điểm trung bình học tập của sinh viên có thời gian học tập khác nhau đều bằng nhau;
2. Điểm trung bình học tập của sinh viên có mức độ yêu thích ngành học khác nhau đều bằng nhau;
3. Không có ảnh hưởng tương tác giữa thời gian tự học và mức độ yêu thích ngành học của sinh viên.

Ví dụ

Xét thời gian tự học và mức độ yêu thích ngành học đối với điểm trung bình của sinh viên.
Các cặp giả thuyết H_0 được đặt ra:

1. Điểm trung bình học tập của sinh viên có thời gian học tập khác nhau đều bằng nhau;
2. Điểm trung bình học tập của sinh viên có mức độ yêu thích ngành học khác nhau đều bằng nhau;
3. Không có ảnh hưởng tương tác giữa thời gian tự học và mức độ yêu thích ngành học của sinh viên.

Ví dụ

Mức độ yêu thích		Thời gian tự học	
	Ít giờ	Trung bình	Nhiều giờ
Không yêu thích	5.8	6.0	6.2
	6.2	6.6	5.8
	5.4	6.1	6.5
	6.0	6.8	6.2
	5.2	6.9	6.4
	5.3	6.0	5.7
	5.4	5.9	6.1
Thích	5.6	7.0	6.8
	6.2	7.7	7.1
	5.7	6.5	6.5
	6.5	7.3	7.1
	7.1	6.1	7.2
	6.0	6.8	6.7
	5.2	6.4	7.0
Rất thích	6.4	6.8	7.6
	6.5	6.6	7.7
	5.0	6.4	7.8
	6.6	6.2	6.8
	6.2	7.1	7.3
	6.1	7.0	7.1
	5.3	7.2	7.2

Ví dụ

- Bước 1: Tính trung bình từng nhóm và trung bình chung ba nhóm
  - Trung bình mẫu của từng nhóm: $\overline{x}_{1*} = 5.89, \overline{x}_{2*} = 6.64, \overline{x}_{3*} = 6.8.$
  - Trung bình mẫu của từng khối: $\overline{x}_{*1} = 6.02, \overline{x}_{*2} = 6.6, \overline{x}_{*3} = 6.71.$
  - Trung bình một ô: $x_{11} = 5.61, x_{21} = 6.33, x_{31} = 6.13, x_{22} = 6.04, x_{12} = 5.76, x_{32} = 6.91, x_{13} = 6.01, x_{23} = 6.76, x_{23} = 7.36.$

Ví dụ

- Bước 1: Tính trung bình từng nhóm và trung bình chung ba nhóm
  - Trung bình mẫu của từng nhóm: $\overline{x}_{1*} = 5.89, \overline{x}_{2*} = 6.64, \overline{x}_{3*} = 6.8.$
  - Trung bình mẫu của từng khối: $\overline{x}_{*1} = 6.02, \overline{x}_{*2} = 6.6, \overline{x}_{*3} = 6.71.$
  - Trung bình một ô: $x_{11} = 5.61, x_{21} = 6.33, x_{31} = 6.13, x_{22} = 6.04, x_{12} = 5.76, x_{32} = 6.91, x_{13} = 6.01, x_{23} = 6.76, x_{23} = 7.36.$

Ví dụ

- Bước 1: Tính trung bình từng nhóm và trung bình chung ba nhóm
  - Trung bình mẫu của từng nhóm: $\overline{x}_{1*} = 5.89, \overline{x}_{2*} = 6.64, \overline{x}_{3*} = 6.8.$
  - Trung bình mẫu của từng khối: $\overline{x}_{*1} = 6.02, \overline{x}_{*2} = 6.6, \overline{x}_{*3} = 6.71.$
  - Trung bình một ô: $x_{11} = 5.61, x_{21} = 6.33, x_{31} = 6.13, x_{22} = 6.04, x_{12} = 5.76, x_{32} = 6.91, x_{13} = 6.01, x_{23} = 6.76, x_{23} = 7.36.$

Ví dụ

- Bước 2: Tính tổng các chênh lệch bình phương
  - $SSG = 9.89$, $SSB = 5.67$, $SSI = 1.58$, $SSE = 11.0$, $SST = 28.14$.
- Bước 3: Tính các phương sai
  - Phương sai giữa các khối: $MSB = \frac{SSB}{H-1} = \frac{5.67}{3-1} = 2.84$.
  - Phương sai giữa các nhóm $MSG = \frac{SSG}{K-1} = \frac{9.89}{3-1} = 4.95$.
  - Phương sai giữa các ô
    $MSI = \frac{SSI}{(K-1)(H-1)} = \frac{1.58}{(3-1)(3-1)} = 0.4$.
  - Phương sai giữa phần dư $MSE = \frac{SSE}{KH(L-1)} = \frac{11.0}{3 \times 3 \times (7-1)} = 0.2$.

Ví dụ

- Bước 2: Tính tổng các chênh lệch bình phương
  - $SSG = 9.89$, $SSB = 5.67$, $SSI = 1.58$, $SSE = 11.0$, $SST = 28.14$.
- Bước 3: Tính các phương sai
  - Phương sai giữa các khối: $MSB = \dfrac{SSB}{H-1} = \dfrac{5.67}{3-1} = 2.84$.
  - Phương sai giữa các nhóm $MSG = \dfrac{SSG}{K-1} = \dfrac{9.89}{3-1} = 4.95$.
  - Phương sai giữa các ô
    $MSI = \dfrac{SSI}{(K-1)(H-1)} = \dfrac{1.58}{(3-1)(3-1)} = 0.4.$
  - Phương sai giữa phần dư $MSE = \dfrac{SSE}{KH(L-1)} = \dfrac{11.0}{3 \times 3 \times (7-1)} = 0.2.$

Ví dụ

- Bước 4: Tính tỉ số F:
  - $F_1 = \frac{MSG}{MSE} = \frac{4.95}{0.2} = 24.75,\ F_{K-1,KH(L-1);\alpha} = F_{2,54;0.05} = 3.17;$
  - $F_2 = \frac{MSB}{MSE} = \frac{2.84}{0.2} = 14.2,\ F_{H-1,KH(L-1);\alpha} = F_{2,54;0.05} = 3.17;$
  - $F_3 = \frac{MSI}{MSE} = \frac{0.4}{0.2} = 2.0,\ F_{(K-1)(H-1),KH(L-1);\alpha} = F_{4,54;0.05} = 2.54.$
- Qui luật bác bỏ
  - Vì $F_1 = 24.75 > F_{2,54,0.05}$ nên ta có đủ bằng chứng thống kê để bác bỏ H_0 trong cặp giả thuyết thứ nhất, tức là điểm trung bình của sinh viên có thời gian tự học khác nhau thì không bằng nhau.
  - Vì $F_2 = 14.2 > F_{2,54,0.05}$ nên ta có đủ bằng chứng thống kê để bác bỏ H_0 trong cặp giả thuyết thứ hai, tức là điểm trung bình của sinh viên có mức độ yêu thích ngành học khác nhau thì không bằng nhau.
  - $F_3 = 2.0 < F_{4,54;0.05}$ nên ta không có đủ bằng chứng thống kê để bác bỏ H_0 trong cặp giả thuyết thứ ba, tức là không có sự tương tác giữa thời gian tự học và mức độ yêu thích ngành học trong việc ảnh hưởng đến điểm trung bình của sinh viên.

Ví dụ

- Bước 4: Tính tỉ số F:
  - $F_1 = \dfrac{MSG}{MSE} = \dfrac{4.95}{0.2} = 24.75$, $F_{K-1,KH(L-1);\alpha} = F_{2,54;0.05} = 3.17$;
  - $F_2 = \dfrac{MSB}{MSE} = \dfrac{2.84}{0.2} = 14.2$, $F_{H-1,KH(L-1);\alpha} = F_{2,54;0.05} = 3.17$;
  - $F_3 = \dfrac{MSI}{MSE} = \dfrac{0.4}{0.2} = 2.0$, $F_{(K-1)(H-1),KH(L-1);\alpha} = F_{4,54;0.05} = 2.54$.
- Qui luật bác bỏ
  - Vì $F_1 = 24.75 > F_{2,54,0.05}$ nên ta có đủ bằng chứng thống kê để bác bỏ H_0 trong cặp giả thuyết thứ nhất, tức là điểm trung bình của sinh viên có thời gian tự học khác nhau thì không bằng nhau.
  - Vì $F_2 = 14.2 > F_{2,54,0.05}$ nên ta có đủ bằng chứng thống kê để bác bỏ H_0 trong cặp giả thuyết thứ hai, tức là điểm trung bình của sinh viên có mức độ yêu thích ngành học khác nhau thì không bằng nhau.
  - $F_3 = 2.0 < F_{4,54;0.05}$ nên ta không có đủ bằng chứng thống kê để bác bỏ H_0 trong cặp giả thuyết thứ ba, tức là không có sự tương tác giữa thời gian tự học và mức độ yêu thích ngành học trong việc ảnh hưởng đến điểm trung bình của sinh viên.

Ví dụ

- Bước 4: Tính tỉ số F:
  - $F_1 = \frac{MSG}{MSE} = \frac{4.95}{0.2} = 24.75$, $F_{K-1,KH(L-1);\alpha} = F_{2,54;0.05} = 3.17$;
  - $F_2 = \frac{MSB}{MSE} = \frac{2.84}{0.2} = 14.2$, $F_{H-1,KH(L-1);\alpha} = F_{2,54;0.05} = 3.17$;
  - $F_3 = \frac{MSI}{MSE} = \frac{0.4}{0.2} = 2.0$, $F_{(K-1)(H-1),KH(L-1);\alpha} = F_{4,54;0.05} = 2.54$.
- Qui luật bác bỏ
  - Vì $F_1 = 24.75 > F_{2,54,0.05}$ nên ta có đủ bằng chứng thống kê để bác bỏ H_0 trong cặp giả thuyết thứ nhất, tức là điểm trung bình của sinh viên có thời gian tự học khác nhau thì không bằng nhau.
  - Vì $F_2 = 14.2 > F_{2,54,0.05}$ nên ta có đủ bằng chứng thống kê để bác bỏ H_0 trong cặp giả thuyết thứ hai, tức là điểm trung bình của sinh viên có mức độ yêu thích ngành học khác nhau thì không bằng nhau.
  - $F_3 = 2.0 < F_{4,54;0.05}$ nên ta không có đủ bằng chứng thống kê để bác bỏ H_0 trong cặp giả thuyết thứ ba, tức là không có sự tương tác giữa thời gian tự học và mức độ yêu thích ngành học trong việc ảnh hưởng đến điểm trung bình của sinh viên.

Ví dụ

- Bước 4: Tính tỉ số F:
  - $F_1 = \dfrac{MSG}{MSE} = \dfrac{4.95}{0.2} = 24.75,\ F_{K-1,KH(L-1);\alpha} = F_{2,54;0.05} = 3.17;$
  - $F_2 = \dfrac{MSB}{MSE} = \dfrac{2.84}{0.2} = 14.2,\ F_{H-1,KH(L-1);\alpha} = F_{2,54;0.05} = 3.17;$
  - $F_3 = \dfrac{MSI}{MSE} = \dfrac{0.4}{0.2} = 2.0,\ F_{(K-1)(H-1),KH(L-1);\alpha} = F_{4,54;0.05} = 2.54.$
- Qui luật bác bỏ
  - Vì $F_1 = 24.75 > F_{2,54,0.05}$ nên ta có đủ bằng chứng thống kê để bác bỏ H_0 trong cặp giả thuyết thứ nhất, tức là điểm trung bình của sinh viên có thời gian tự học khác nhau thì không bằng nhau.
  - Vì $F_2 = 14.2 > F_{2,54,0.05}$ nên ta có đủ bằng chứng thống kê để bác bỏ H_0 trong cặp giả thuyết thứ hai, tức là điểm trung bình của sinh viên có mức độ yêu thích ngành học khác nhau thì không bằng nhau.
  - $F_3 = 2.0 < F_{4,54;0.05}$ nên ta không có đủ bằng chứng thống kê để bác bỏ H_0 trong cặp giả thuyết thứ ba, tức là không có sự tương tác giữa thời gian tự học và mức độ yêu thích ngành học trong việc ảnh hưởng đến điểm trung bình của sinh viên.

Ví dụ trong R

```
> DiemTB = scan()
1: 5.8 6.2 5.4 6.0 5.2 5.3 5.4 5.6 6.2
10: 5.7 6.5 7.1 6.0 5.2 6.4 6.5 5.0 6.6
19: 6.2 6.1 5.3 6.0 6.6 6.1 6.8 6.9 6.0
28: 5.9 7.0 7.7 6.5 7.3 6.1 6.8 6.4 6.8
37: 6.6 6.4 6.2 7.1 7.0 7.2 6.2 5.8 6.5
46: 6.2 6.4 5.7 6.1 6.8 7.1 6.5 7.1 7.2
55: 6.7 7.0 7.6 7.7 7.8 6.8 7.3 7.1 7.2
64:
Read 63 items
> PhanThoiGian = factor(rep(1:3,each=21))
> PhanMucYT = factor(rep(1:3,each=7,length=63))
> anova(lm(DiemTB~PhanThoiGian+PhanMucYT+PhanThoiGian*PhanMucYT))
```

Ví dụ trong R

```
Analysis of Variance Table

Response: DiemTB
                       Df  Sum Sq Mean Sq F value    Pr(>F)
PhanThoiGian            2  9.8867  4.9433 24.2673 3.028e-08 ***
PhanMucYT               2  5.6686  2.8343 13.9138 1.338e-05 ***
PhanThoiGian:PhanMucYT  4  1.5790  0.3948  1.9379    0.1174
Residuals              54 11.0000  0.2037
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Bài toán phân tích sâu Two-way ANOVA

Trong phân tích phương sai hai yếu tố, khi đã xác định được có sự khác biệt giữa các nhóm so sánh, chúng ta có thể dùng kiểm định Tukey HSD để xác định các cặp trung bình khác nhau xét theo yếu tố thứ nhất (so sánh giữa K nhóm) hay xét theo yếu tố thứ hai (so sánh giữa H khối).

Hai giá trị giới hạn Tukey được tính theo công thức:

- $T = q_{\alpha,K,KH(L-1)}\sqrt{\frac{MSE}{H \times L}}$, đối với việc so sánh theo yếu tố thứ nhất;
- $T = q_{\alpha,H,KH(L-1)}\sqrt{\frac{MSE}{K \times L}}$, đối với việc so sánh theo yếu tố thứ hai.

Tiêu chuẩn quyết định là bác bỏ giả thuyết H_0 khi độ lệch tuyệt đối giữa các cặp trung bình mẫu theo nhóm (dòng) lớn hơn hay bằng T giới hạn tương ứng.

Bài toán phân tích sâu Two-way ANOVA

Trong phân tích phương sai hai yếu tố, khi đã xác định được có sự khác biệt giữa các nhóm so sánh, chúng ta có thể dùng kiểm định Tukey HSD để xác định các cặp trung bình khác nhau xét theo yếu tố thứ nhất (so sánh giữa K nhóm) hay xét theo yếu tố thứ hai (so sánh giữa H khối).
Hai giá trị giới hạn Tukey được tính theo công thức:

- $T = q_{\alpha,K,KH(L-1)}\sqrt{\frac{MSE}{H \times L}}$, đối với việc so sánh theo yếu tố thứ nhất;
- $T = q_{\alpha,H,KH(L-1)}\sqrt{\frac{MSE}{K \times L}}$, đối với việc so sánh theo yếu tố thứ hai.

Tiêu chuẩn quyết định là bác bỏ giả thuyết H_0 khi độ lệch tuyệt đối giữa các cặp trung bình mẫu theo nhóm (dòng) lớn hơn hay bằng T giới hạn tương ứng.

Bài toán phân tích sâu Two-way ANOVA

Trong phân tích phương sai hai yếu tố, khi đã xác định được có sự khác biệt giữa các nhóm so sánh, chúng ta có thể dùng kiểm định Tukey HSD để xác định các cặp trung bình khác nhau xét theo yếu tố thứ nhất (so sánh giữa K nhóm) hay xét theo yếu tố thứ hai (so sánh giữa H khối).
Hai giá trị giới hạn Tukey được tính theo công thức:

- $T = q_{\alpha,K,KH(L-1)}\sqrt{\frac{MSE}{H \times L}}$, đối với việc so sánh theo yếu tố thứ nhất;
- $T = q_{\alpha,H,KH(L-1)}\sqrt{\frac{MSE}{K \times L}}$, đối với việc so sánh theo yếu tố thứ hai.

Tiêu chuẩn quyết định là bác bỏ giả thuyết H_0 khi độ lệch tuyệt đối giữa các cặp trung bình mẫu theo nhóm (dòng) lớn hơn hay bằng T giới hạn tương ứng.

Ví dụ

Trong tính toán ở ví dụ trên, ta có
$K = 3, H = 3, L = 7, \alpha = 5\%, MSE = 0.2$. Giá trị q của phân phối Tukey: $q_{0.05,3,54} = 3.4$.

- So sánh giữa các nhóm theo yếu tố thứ nhất (thời gian tự học), giá trị tới hạn: $T = 3.4\sqrt{\frac{0.2}{21}} = 0.33$.
- Trung bình nhóm lần lượt là: $5.89, 6.64, 6.8$ và chênh lệch giữa các nhóm là
  - $D_{\text{ít, TB}} = |5.89 - 6.64| = 0.75$;
  - $D_{\text{ít, nhiều}} = |5.89 - 6.8| = 0.91$;
  - $D_{\text{TB, nhiều}} = |6.64 - 6.8| = 0.16$.
- Chênh lệch $D_{\text{ít, TB}}, D_{\text{ít, nhiều}}$ đều lớn hơn giá trị giới hạn Tukey T, chênh lệch $D_{\text{TB, nhiều}}$ nhỏ hơn giá trị tới hạn T nên ta có cơ sở thống kê để cho rằng nhóm học ít có điểm trung bình khác biệt so với nhóm học trung bình và nhóm học nhiều. Tuy nhiên, điểm trung bình của nhóm học trung bình và nhóm học nhiều lệch nhau không có ý nghĩa thống kê.

Ví dụ

Trong tính toán ở ví dụ trên, ta có
$K = 3, H = 3, L = 7, \alpha = 5\%, MSE = 0.2$. Giá trị q của phân phối Tukey: $q_{0.05,3,54} = 3.4$.

- So sánh giữa các nhóm theo yếu tố thứ nhất (thời gian tự học), giá trị tới hạn: $T = 3.4\sqrt{\frac{0.2}{21}} = 0.33$.
- Trung bình nhóm lần lượt là: $5.89, 6.64, 6.8$ và chênh lệch giữa các nhóm là
  - $D_{\text{ít, TB}} = |5.89 - 6.64| = 0.75$;
  - $D_{\text{ít, nhiều}} = |5.89 - 6.8| = 0.91$;
  - $D_{\text{TB, nhiều}} = |6.64 - 6.8| = 0.16$.
- Chênh lệch $D_{\text{ít, TB}}, D_{\text{ít, nhiều}}$ đều lớn hơn giá trị giới hạn Tukey T, chênh lệch $D_{\text{TB, nhiều}}$ nhỏ hơn giá trị tới hạn T nên ta có cơ sở thống kê để cho rằng nhóm học ít có điểm trung bình khác biệt so với nhóm học trung bình và nhóm học nhiều. Tuy nhiên, điểm trung bình của nhóm học trung bình và nhóm học nhiều lệch nhau không có ý nghĩa thống kê.

Ví dụ

Trong tính toán ở ví dụ trên, ta có
$K = 3, H = 3, L = 7, \alpha = 5\%, MSE = 0.2$. Giá trị q của phân phối Tukey: $q_{0.05,3,54} = 3.4$.

- So sánh giữa các nhóm theo yếu tố thứ nhất (thời gian tự học), giá trị tới hạn: $T = 3.4\sqrt{\frac{0.2}{21}} = 0.33$.
- Trung bình nhóm lần lượt là: $5.89, 6.64, 6.8$ và chênh lệch giữa các nhóm là
  - $D_{\text{ít, TB}} = |5.89 - 6.64| = 0.75$;
  - $D_{\text{ít, nhiều}} = |5.89 - 6.8| = 0.91$;
  - $D_{\text{TB, nhiều}} = |6.64 - 6.8| = 0.16$.
- Chênh lệch $D_{\text{ít, TB}}, D_{\text{ít, nhiều}}$ đều lớn hơn giá trị giới hạn Tukey T, chênh lệch $D_{\text{TB, nhiều}}$ nhỏ hơn giá trị tới hạn T nên ta có cơ sở thống kê để cho rằng nhóm học ít có điểm trung bình khác biệt so với nhóm học trung bình và nhóm học nhiều. Tuy nhiên, điểm trung bình của nhóm học trung bình và nhóm học nhiều lệch nhau không có ý nghĩa thống kê.

Ví dụ

- So sánh giữa các nhóm theo yếu tố thứ hai (mức độ yêu thích ngành học), giá trị tới hạn: $T = 3.04\sqrt{\frac{0.2}{21}} = 0.33$.
- Trung bình khối lần lượt là: $6.02, 6.6, 6.71$ và chênh lệch giữa các khối là
  - $D_{\text{không thích, thích}} = |6.02 - 6.6| = 0.58$;
  - $D_{\text{không thích, rất thích}} = |6.02 - 6.71| = 0.69$;
  - $D_{\text{thích, rất thích}} = |6.6 - 6.71| = 0.11$.
- Chênh lệch giữa nhóm thích và rất thích $D_{\text{thích, rất thích}}$ nhỏ hơn giá trị T giới hạn, nên ta kết luận rằng nhóm sinh viên có mức độ yêu thích ngành học nhiều hay rất nhiều thì kết quả học tập không khác biệt nhau đáng kể. Nhóm không thích ngành học có kết quả học tập kém hơn hẳn hai nhóm thích và rất thích ngành đang học.

Ví dụ

- So sánh giữa các nhóm theo yếu tố thứ hai (mức độ yêu thích ngành học), giá trị tới hạn: $T = 3.04\sqrt{\frac{0.2}{21}} = 0.33.$
- Trung bình khối lần lượt là: $6.02, 6.6, 6.71$ và chênh lệch giữa các khối là
  - $D_{\text{không thích, thích}} = |6.02 - 6.6| = 0.58;$
  - $D_{\text{không thích, rất thích}} = |6.02 - 6.71| = 0.69;$
  - $D_{\text{thích, rất thích}} = |6.6 - 6.71| = 0.11.$
- Chênh lệch giữa nhóm thích và rất thích $D_{\text{thích, rất thích}}$ nhỏ hơn giá trị T giới hạn, nên ta kết luận rằng nhóm sinh viên có mức độ yêu thích ngành học nhiều hay rất nhiều thì kết quả học tập không khác biệt nhau đáng kể. Nhóm không thích ngành học có kết quả học tập kém hơn hẳn hai nhóm thích và rất thích ngành đang học.

Ví dụ

- So sánh giữa các nhóm theo yếu tố thứ hai (mức độ yêu thích ngành học), giá trị tới hạn: $T = 3.04\sqrt{\frac{0.2}{21}} = 0.33$.
- Trung bình khối lần lượt là: $6.02, 6.6, 6.71$ và chênh lệch giữa các khối là
  - $D_{\text{không thích, thích}} = |6.02 - 6.6| = 0.58$;
  - $D_{\text{không thích, rất thích}} = |6.02 - 6.71| = 0.69$;
  - $D_{\text{thích, rất thích}} = |6.6 - 6.71| = 0.11$.
- Chênh lệch giữa nhóm thích và rất thích $D_{\text{thích, rất thích}}$ nhỏ hơn giá trị T giới hạn, nên ta kết luận rằng nhóm sinh viên có mức độ yêu thích ngành học nhiều hay rất nhiều thì kết quả học tập không khác biệt nhau đáng kể. Nhóm không thích ngành học có kết quả học tập kém hơn hẳn hai nhóm thích và rất thích ngành đang học.

Ví dụ trong R

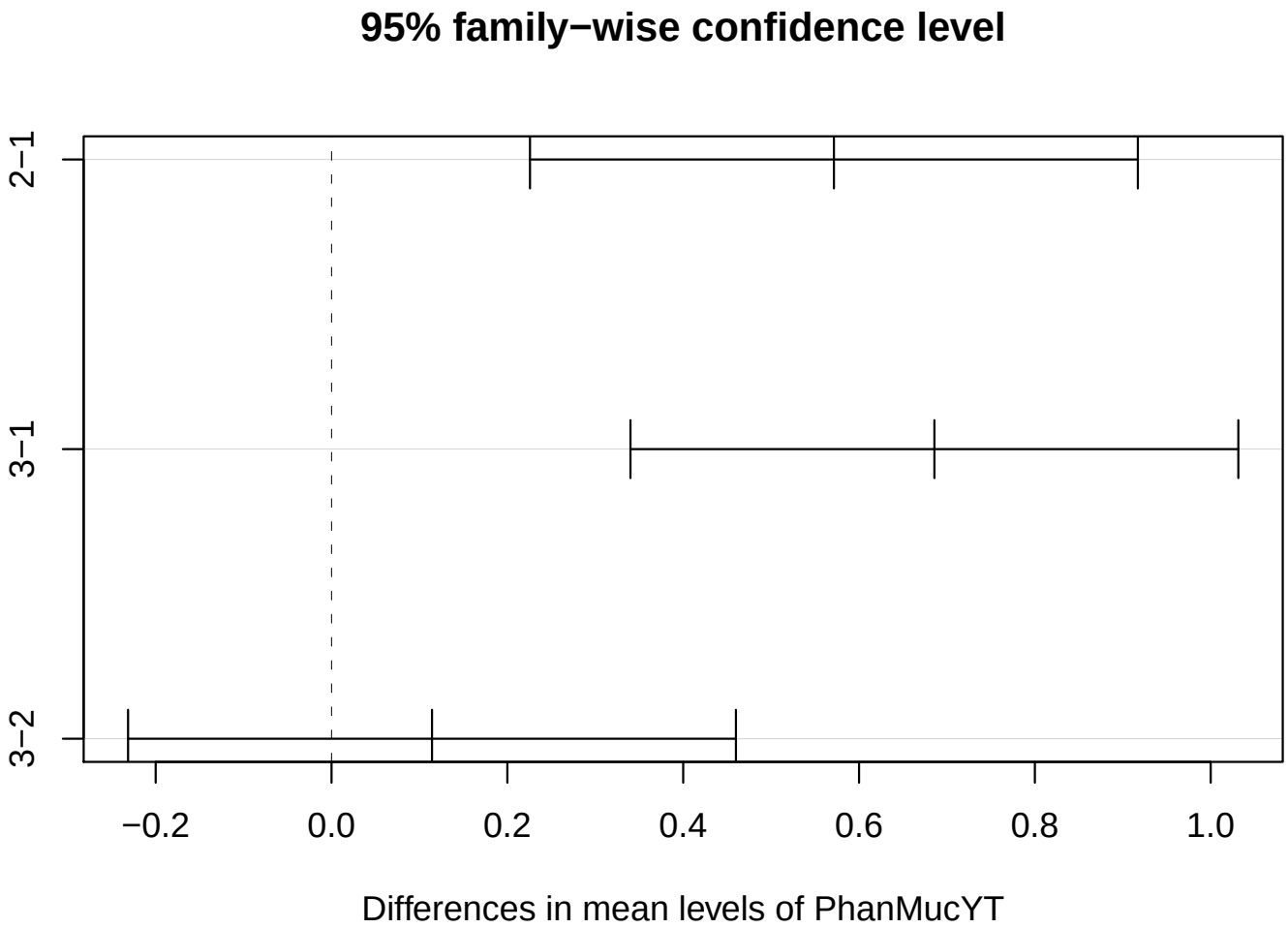
- Hàm `TukeyHSD()` giúp thực hiện phân tích sâu Two-way ANOVA trong R. Trong ví dụ trên ta có thể thực hiện:

```
> TukeyHSD(aov(DiemTB ~ PhanThoiGian + PhanMucYT +
                        PhanThoiGian*PhanMucYT))
```

- Ta có thể minh họa sự khác biệt giữa các trung bình như sau:

```
> plot(TukeyHSD(aov(DiemTB~PhanThoiGian)))
> plot(TukeyHSD(aov(DiemTB~PhanMucYT)))
```

Minh họa phân tích sâu Two-way ANOVA trong R

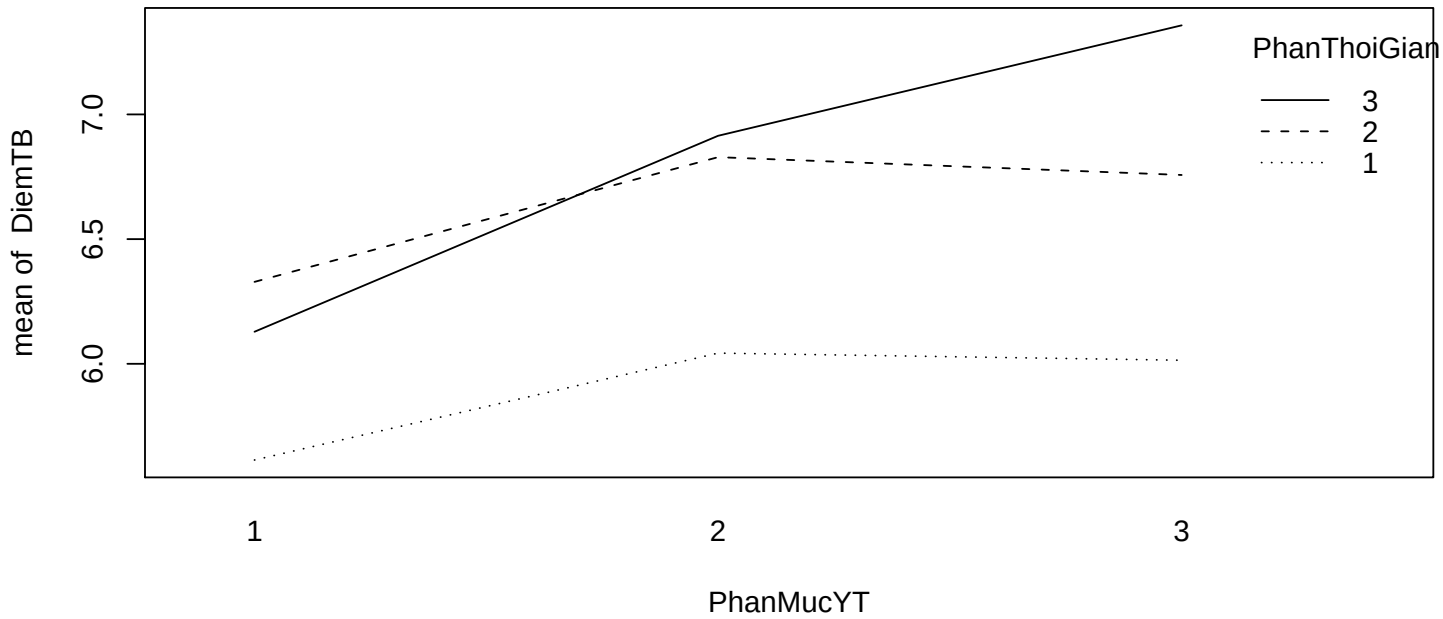


Minh họa tương tác trong R

- Hàm `interaction.plot` dùng để minh họa sự tương tác giữa các yếu tố nguyên nhân đến yếu tố kết quả:

```
> interaction.plot(PhanMucYT,PhanThoiGian,DiemTB)
```

- Kết quả được minh họa qua hình vẽ sau:



Bài tập ôn luyện

Bài toán

Một lớp gồm 23 sinh viên. Vào đầu học kì mỗi kì mỗi sinh viên được chọn ngẫu nhiên để theo một trong 4 phụ giảng A, B, C hay D. Các sinh viên này được khuyến khích gặp người phụ giảng để nhờ hướng dẫn giải đáp các khó khăn trong bài học. Cuối học kì họ thi chung một bài thi và điểm thi được ghi lại (ứng với mỗi phụ giảng) như sau:

A	B	C	D
72	78	80	79
69	93	68	70
84	79	59	61
76	97	75	74
64	88	82	85
	81	68	63

a. Hãy kiểm định ở mức ý nghĩa $\alpha = 5\%$ giả thiết H_0 rằng điểm trung bình tổng thể bằng nhau ứng với 4 phụ giảng.

b. Nếu giả thiết H_0 bị bác bỏ, hãy cho biết ở mức ý nghĩa $\alpha = 5\%$ có thể chỉ ra thêm điểm trung bình tổng thể ứng với phụ giảng nào là khác nhau?