Trường Đại học Khoa học Tự nhiên TP.HCM

Khoa Toán - Tin học

THỐNG KÊ NHIỀU CHIỀU

Chương 6: So sánh các vectơ trung bình nhiều chiều

Đinh Anh Huy - 18110103

Nguyễn Đức Vũ Duy - 18110004

Ngày 6 tháng 4 năm 2021

Giới thiệu

Những ý tưởng ở chương 5 có thể được mở rộng để giải quyết các bài toán so sánh nhiều véctơ trung bình. Về mặt lý thuyết thì các bài toán này phức tạp hơn và chủ yếu nằm ở giả định phân phối chuẩn nhiều chiều hoặc cỡ mẫu lớn.

So sánh ghép cặp (Paired Comparisons)

Bài toán

Xét bài toán kiểm tra tính hiệu quả của một loại thuốc mới, chúng ta sẽ so sánh các phép đo trước khi "treatment" và sau khi "treatment" ở cùng một đối tượng. Trong nhiều trường hợp khác, có thể có 2 hoặc nhiều treatments tham gia vào cùng một đối tượng thực nghiệm.

Paired Comparisons

Thiết kế bài toán trong trường hợp một chiều

Gọi X_{j1} là phản hồi từ *treatment 1* và X_{j2} là phản hồi từ *treatment 2* với lần thực nghiệm thứ j. Khi đó, n differences

$$D_j = X_{j1} - X_{j2}, j = 1, 2, ..., n \tag{6.1}$$

phản ánh sự chênh lệch ảnh hưởng giữa các treatments.

Paired Comparisons

Cho $D_1, D_2, ..., D_n$ là mẫu ngẫu nhiên được lấy từ tổng thể có phân phối chuẩn $\mathcal{N}(\delta, \sigma_d^2)$ với $D_j, j = 1, 2, ..., n$ là các differences được định nghĩa ở (6.1). Khi đó

$$t = \frac{\overline{D} - \delta}{s_d / \sqrt{n}} \sim t_{n-1} \tag{6.2}$$

trong đó

$$\overline{D} = \frac{1}{n} \sum_{j=1}^{n} D_j \text{ và } s_d^2 = \frac{1}{n-1} \sum_{j=1}^{n} (D_j - \overline{D})^2. \tag{6.3}$$

Paired Comparisons

Xét kiểm định:

$$H_0 : \delta = 0 \quad \text{và} \quad H_1 : \delta \neq 0$$

Với mức ý nghĩa α, ta sẽ kiểm định giả thuyết trên bằng cách so sánh $|t|$ với $t_{n-1}(\alpha/2)$ - phân vị trên thứ $100(\alpha/2)$ của phân phối Student với bậc tự do $n-1$.

Paired Comparisons

Xét kiểm định:

$$H_0 : \delta = 0 \quad \text{và} \quad H_1 : \delta \neq 0$$

Với mức ý nghĩa α, ta sẽ kiểm định giả thuyết trên bằng cách so sánh $|t|$ với $t_{n-1}(\alpha/2)$ - phân vị trên thứ $100(\alpha/2)$ của phân phối Student với bậc tự do $n-1$.

Ta cũng có khoảng tin cậy $100(1-\alpha)\%$ cho $\delta = E(X_{j1} - X_{j2})$ là:

$$\overline{d} - t_{n-1}(\alpha/2)\frac{s_d}{\sqrt{n}} \leq \delta \leq \overline{d} + t_{n-1}(\alpha/2)\frac{s_d}{\sqrt{n}} \tag{6.4}$$

Paired Comparisons

Mô hình bài toán trong trường hợp nhiều chiều

Với p responses, 2 treatments và n unit thực nghiệm. Ta đánh nhãn p responses ở unit thứ j là:

$$
\begin{aligned}
X_{1j1} &= \text{biến 1 trong treatment 1} \\
X_{1j2} &= \text{biến 2 trong treatment 1} \\
&\vdots \\
X_{1jp} &= \text{biến p trong treatment 1} \\
X_{2j1} &= \text{biến 1 trong treatment 2} \\
X_{2j2} &= \text{biến 2 trong treatment 2} \\
&\vdots \\
X_{2jp} &= \text{biến } p \text{ trong treatment 2}
\end{aligned}
$$

Paired Comparisons

Khi đó p biến ngẫu nhiên là các cặp difference trở thành:

$$\begin{aligned} D_{j1} &= X_{1j1} - X_{2j1} \\ D_{j2} &= X_{1j2} - X_{2j2} \\ &\vdots \\ D_{jp} &= X_{1jp} - X_{2jp} \end{aligned} \tag{6.5}$$

Đặt $\mathbf{D}_j = [D_{j1}, D_{j2}, ..., D_{jp}]^T$ với $j = 1, 2, ..., n$. Khi đó

$$E(\mathbf{D}_j) = \boldsymbol{\delta} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_p \end{bmatrix} \quad \text{và} \quad Cov(\mathbf{D}_j) = \mathbf{\Sigma_d} \tag{6.6}$$

Paired Comparisons

Ngoài ra, nếu $\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_n$ là các vectơ ngẫu nhiên độc lập và có cùng phân phối $\mathcal{N}_p(\boldsymbol{\delta}, \boldsymbol{\Sigma}_\mathbf{d})$ thì vectơ của các differences trung bình $\boldsymbol{\delta}$ có thể được "ước lượng" dựa trên thống kê T^2.

$$T^2 = n(\overline{\mathbf{D}} - \delta)^T \mathbf{S}_\mathbf{d}^{-1}(\overline{\mathbf{D}} - \delta) \qquad (6.7)$$

trong đó

$$\overline{\mathbf{D}} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{D}_j \quad \text{và} \quad \mathbf{S}_\mathbf{d} = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{D}_j - \overline{\mathbf{D}})(\mathbf{D}_j - \overline{\mathbf{D}})^T \qquad (6.8)$$

Paired Comparisons

Kết quả 6.1

Cho $\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_n$ là mẫu ngẫu nhiên được lấy từ tổng thể có phân phối chuẩn p chiều $\mathcal{N}_p(\boldsymbol{\delta}, \boldsymbol{\Sigma}_\mathbf{d})$. Khi đó

$$T^2 = n(\overline{\mathbf{D}} - \boldsymbol{\delta})^T \mathbf{S}_\mathbf{d}^{-1}(\overline{\mathbf{D}} - \boldsymbol{\delta}) \sim \frac{(n-1)p}{n-p}\mathcal{F}_{p,n-p}$$

Nếu n và $n-p$ đều lớn thì T^2 xấp xỉ về phân phối χ^2_p.

Paired Comparisons

Cho các quan trắc differences $\mathbf{d}_j = [d_{j1}, d_{j2}, ..., d_{jp}]^T, j = 1, 2, ..., n$ là mẫu ngẫu nhiên với các biến tương ứng ở (6.5) được lấy từ tổng thể có phân phối chuẩn $\mathcal{N}_p(\boldsymbol{\delta}, \boldsymbol{\Sigma}_\mathbf{d})$. Với mức ý nghĩa α, xét kiểm định

$$H_0 : \boldsymbol{\delta} = \mathbf{0} \quad \text{và} \quad H_1 : \boldsymbol{\delta} \neq \mathbf{0}.$$

Ta bác bỏ giả thuyết H_0 với mức ý nghĩa α nếu

$$T^2 = n\overline{\mathbf{d}}^T \mathbf{S}_\mathbf{d}^{-1} \overline{\mathbf{d}} > \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$$

trong đó $F_{n,n-p}(\alpha)$ là phân vị trên thứ 100α của phân phối *Fisher* với hệ số tự do là p và n-p; $\overline{\mathbf{d}}$ và $\mathbf{S}_\mathbf{d}$ được cho bởi (6.8).

Paired Comparisons

Khoảng tin cậy $100(1-\alpha)\%$ đồng thời cho từng differnece trung bình riêng lẻ là:

$$\delta_i : \overline{d_i} \pm \sqrt{\frac{(n-1)p}{n(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{d_i}^2}{n}} \qquad (6.10)$$

trong đó $\overline{d_i}$ là phần tử thứ i của $\overline{\mathbf{d}}$ và $s_{d_i}^2$ là phần tử thứ i trên đường chéo chính của $\mathbf{S_d}$.

Trường hợp $n-p$ lớn, $[(n-1)p/(n-p)]F_{p,n-p}(\alpha) = \chi_p^2(\alpha)$.

Paired Comparisons

Khoảng tin cậy $100(1-\alpha)\%$ Bonferroni đồng thời cho từng difference trung bình riêng lẻ là:

$$\delta_i : \overline{d}_i \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{d_i}^2}{n}} \qquad (6.10a)$$

Trong đó $t_{n-1}(\frac{\alpha}{2p})$ là phân vị trên thứ $100(\frac{\alpha}{2p})$ của phân phối Student với bậc tự do $n-1$.

Paired Comparisons

Ví dụ 6.1 (Kiểm tra độ chênh lệch trung bình với các cặp quan sát)

Table 6.1 *Effluent Data*

Sample j	Commercial lab x_{1j1} (BOD)	x_{1j2} (SS)	State lab of hygiene x_{2j1} (BOD)	x_{2j2} (SS)
1	6	27	25	15
2	6	23	28	13
3	18	64	36	22
4	8	44	35	29
5	11	30	15	31
6	34	75	44	64
7	28	26	42	30
8	71	124	54	64
9	43	54	34	56
10	33	30	29	20
11	20	14	39	21

Source: Data courtesy of S. Weber.

Paired Comparisons

Xét thống kê T^2 cho kiểm định giả thuyết $H_0 : \boldsymbol{\delta} = [\delta_1, \delta_2]^T = [0, 0]^T$ được xây dựng từ độ chênh lệch giữa các cặp quan trắc:

d_{j1}	-19	-22	-18	-27	-4	-10	-14	17	9	4	-19
d_{j2}	12	10	42	15	-1	11	-4	60	-2	10	-7

trong đó $\mathbf{d}_{ji} = x_{1ji} - x_{2ji}, i = 1, 2.$

Paired Comparisons

Giả sử giả thuyết H_0 đúng, ta có:

$$\overline{\mathbf{d}} = \begin{bmatrix} \overline{d}_1 \\ \overline{d}_2 \end{bmatrix} = \begin{bmatrix} -9.36 \\ 13.27 \end{bmatrix}, \quad \mathbf{S}_d = \begin{bmatrix} 199.26 & 88.38 \\ 88.38 & 418.61 \end{bmatrix}$$

Với mẫu thực nghiệm, ta có giá trị thống kê

$$\begin{aligned} t^2 &= n\overline{\mathbf{d}}^T \mathbf{S}_d^{-1} \overline{\mathbf{d}} \\ &= 11 \begin{bmatrix} -9.36 & 13.27 \end{bmatrix} \begin{bmatrix} 0.0055 & 0.0012 \\ -0.0012 & 0.0026 \end{bmatrix} \begin{bmatrix} -9.36 \\ 13.27 \end{bmatrix} \approx 13.6 \end{aligned}$$

Paired Comparisons

Với mức ý nghĩa $\alpha = 0.05$ ta có

$$\frac{p(n-1)}{n-p} F_{0.05}(p, n-p) = \frac{2 \times 10}{9} F_{2,9}(0.05) \approx 9.47$$

Vì $t^2 > \dfrac{p(n-1)}{n-p} F_{n,n-p}(0.05)$ nên ta bác bỏ H_0 với mức ý nghĩa $\alpha = 0.05$.

Như vậy, có sự chênh lệch giữa các phép đo ở hai phòng thí nghiệm. Từ việc kiểm tra dữ liệu, phòng thí nghiệm thương mại có xu hướng tạo ra các chỉ số BOD thấp hơn và SS cao hơn phòng thí nghiệm công lập.

Paired Comparisons

Khoảng tin cậy đồng thời với độ tin cậy 95% cho độ chênh lệch trung bình δ_1 và δ_2 là:

$$\delta_1 : \overline{d}_1 \pm \sqrt{\frac{(n-1)p}{n-p} F_\alpha(p, n-p)} \sqrt{\frac{s_{d_1}^2}{n}} \approx -9.36 \pm \sqrt{9.47} \sqrt{\frac{199.26}{11}}$$

$$\approx (-22.46, 3.74)$$

và

$$\delta_2 : 13.27 \pm \sqrt{9.47} \sqrt{\frac{418.61}{11}} \approx (-5.71, 32.25)$$

Paired Comparisons

Một cách tổng quát, ta có thể tính T^2 từ các đại lượng mẫu đầy đủ $\bar{\mathbf{x}}$ và $\mathbf{S}$. Trong đó, $\bar{\mathbf{x}}$ là một vectơ $(2p \times 1)$ của trung bình mẫu cho p biến trên hai treatments

$$\bar{\mathbf{x}} = [\bar{x}_{11}.\bar{x}_{12}, ..., \bar{x}_{1p}, \bar{x}_{21}, \bar{x}_{22}, ..., \bar{x}_{2p}] \tag{6.11}$$

và $\mathbf{S}$ là ma trận $(2p \times 2p)$ của các hiệp phương sai mẫu

$$\mathbf{S} = \begin{bmatrix} \underset{(p\times p)}{\mathbf{S}_{11}} & \underset{(p\times p)}{\mathbf{S}_{12}} \\ \underset{(p\times p)}{\mathbf{S}_{21}} & \underset{(p\times p)}{\mathbf{S}_{22}} \end{bmatrix} \tag{6.12}$$

Paired Comparisons

Ta định nghĩa ma trận $\underset{(p\times 2p)}{\mathbf{C}}$ như sau:

$$\underset{(p\times 2p)}{\mathbf{C}} = \left[\begin{array}{cccc|cccc} 1 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & -1 \end{array}\right] \tag{6.13}$$

trong đó mỗi dòng $\mathbf{c}_i$ là một *contrast vectơ*, vì $\sum_{j=1}^{2p} c_{ij} = 0$. Khi đó

$$\begin{aligned} \mathbf{d}_j &= \mathbf{C}\mathbf{x}_j, \quad j = 1, 2, ..., n \\ \overline{\mathbf{d}} &= \mathbf{C}\overline{\mathbf{x}} \quad \text{và} \quad \mathbf{S}_d = \mathbf{C}\mathbf{S}\mathbf{C}^T \end{aligned} \tag{6.14}$$

Do đó

$$T^2 = n\overline{\mathbf{x}}^T\mathbf{C}^T(\mathbf{C}\mathbf{S}\mathbf{C}^T)^{-1}\mathbf{C}\overline{\mathbf{x}} \tag{6.15}$$

Thiết kế phương pháp lặp lại cho so sánh các treatment

Một cách tổng quát khác của các cặp thống kê t đơn biến phát sinh trong trường hợp q treatments được thực hiện trên từng unit. Với quan trắc thứ j ta có

$$\mathbf{X}_j = \begin{bmatrix} X_{j1} \\ X_{j2} \\ \dots \\ X_{jq} \end{bmatrix}, \quad j = 1, 2, ..., n$$

trong đó X_{ji} là phản hồi cho treatment thứ i trên unit thứ j.

Repeated Measures Design

cho bài toán so sánh các treament

Xét các contrast của từng thành phần của $\boldsymbol{\mu} = E(\mathbf{X}_j)$

$$\begin{bmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \vdots \\ \mu_1 - \mu_q \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{bmatrix} = \underset{((q-1)\times 1)}{\mathbf{C}_1 \boldsymbol{\mu}}$$

hoặc

$$\begin{bmatrix} \mu_2 - \mu_1 \\ \mu_3 - \mu_2 \\ \vdots \\ \mu_q - \mu_{q-1} \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{bmatrix} = \underset{((q-1)\times 1)}{\mathbf{C}_2 \boldsymbol{\mu}}$$

trong đó $\mathbf{C}_1$ và $\mathbf{C}_2$ được gọi là *contrast matrix* vì $q-1$ dòng của chúng là các contrast vectơ và chúng độc lập tuyến tính với nhau.

Repeated Measures Design

cho bài toán so sánh các treatment

Khi các treatment mean bằng nhau, $\mathbf{C}_1\boldsymbol{\mu} = \mathbf{C}_2\boldsymbol{\mu} = \mathbf{0}$ thì giả thuyết kiểm định trở thành $H_0 : \mathbf{C}\boldsymbol{\mu} = \mathbf{0}$ với mọi contrast matrix $\mathbf{C}$.

Ta có trung bình $\mathbf{C}\overline{\mathbf{x}}$ và ma trận hiệp phương sai $\mathbf{CSC}^T$, ta kiểm định giả thuyết $\mathbf{C}\boldsymbol{\mu} = \mathbf{0}$ sử dụng thống kê T^2

$$T^2 = n(\mathbf{C}\overline{\mathbf{x}})^T(\mathbf{CSC}^T)^{-1}\mathbf{C}\overline{\mathbf{x}}$$

Repeated Measures Design

Kiểm định tính tương đồng giữa các treatment

Xét tổng thể có phân phối chuẩn nhiều chiều $\mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ và $\mathbf{C}$ là một contrast matrix. Với mức ý nghĩa α, ta kiểm điểm giả thuyết $H_0 : \mathbf{C}\boldsymbol{\mu} = \mathbf{0}$ với đối thuyết $H_1 : \mathbf{C}\boldsymbol{\mu} \neq \mathbf{0}$. Khi đó, ta bác bỏ giả thuyết H_0 nếu

$$T^2 = n(\mathbf{C}\overline{\mathbf{x}})^T(\mathbf{CSC}^T)^{-1}\mathbf{C}\overline{\mathbf{x}} > \frac{(n-1)(q-1)}{n-q+1} F_{q-1,n-q+1}(\alpha) \quad (6.16)$$

Repeated Measures Design

Kiểm định tính tương đồng giữa các treatment

Miền tin cậy cho $\mathbf{C}\boldsymbol{\mu}$ với $\boldsymbol{\mu}$ là trung bình của tổng thể có phân phối chuẩn được xác định như sau

$$n(\mathbf{C}\overline{\mathbf{x}} - \mathbf{C}\boldsymbol{\mu})^T(\mathbf{CSC}^T)^{-1}(\mathbf{C}\overline{\mathbf{x}} - \mathbf{C}\boldsymbol{\mu}) \leq \frac{(n-1)(q-1)}{(n-q+1)} F_{q-1,n-q+1}(\alpha) \tag{6.17}$$

Khoảng tin cậy động thời với độ tin cậy $100(1-\alpha)\%$ cho từng contrast vectơ $\mathbf{c}^T\boldsymbol{\mu}$ được cho bởi

$$\mathbf{c}^T\boldsymbol{\mu}: \quad \mathbf{c}^T\overline{\mathbf{x}} \pm \sqrt{\frac{(n-1)(q-1)}{(n-q+1)} F_{q-1,n-q+1}(\alpha)} \sqrt{\frac{\mathbf{c}^T\mathbf{S}\mathbf{c}}{n}} \tag{6.18}$$

Repeated Measures Design

Kiểm định tính tương đồng giữa các treatment

Ví dụ 6.2 (Kiểm định sự tương đồng giữa các treatment trong repeated measures design)

Table 6.2 Sleeping-Dog Data

Dog	Treatment 1	2	3	4
1	426	609	556	600
2	253	236	392	395
3	359	433	349	357
4	432	431	522	600
5	405	426	513	513
6	324	438	507	539
7	310	312	410	456
8	326	326	350	504
9	375	447	547	548
10	286	286	403	422
11	349	382	473	497
12	429	410	488	547
13	348	377	447	514
14	412	473	472	446
15	347	326	455	468
16	434	458	637	524
17	364	367	432	469
18	420	395	508	531
19	397	556	645	625

Source: Data courtesy of Dr. J. Atlee.

Repeated Measures Design

Kiểm định tính tương đồng giữa các treatment

Bảng 6.2 chứa 4 treatmeans cho mỗi 19 con chó, trong đó

$$\begin{aligned}
\text{Treatment 1} &= \text{áp suất } CO_2 \text{ cao không thêm } H \\
\text{Treatment 2} &= \text{áp suất } CO_2 \text{ thấp không thêm } H \\
\text{Treatment 3} &= \text{áp suất } CO_2 \text{ cao có thêm } H \\
\text{Treatment 4} &= \text{áp suất } CO_2 \text{ thấp có thêm } H
\end{aligned}$$

Repeated Measures Design

Kiểm định tính tương đồng giữa các treatment

Xét $\mu_1, \mu_2, \mu_3, \mu_4$ lần lượt là thời gian phản hồi trung bình của các treatment 1,2,3,4. Khi đó

$$(\mu_3 + \mu_4) - (\mu_1 + \mu_2) = \left(\begin{array}{c} \text{Độ chênh lệch của halothane} \\ \text{khi có hoặc không có halothane} \end{array} \right)$$

$$(\mu_1 + \mu_3) - (\mu_2 + \mu_4) = \left(\begin{array}{c} \text{Độ chênh lệch của } CO_2 \\ \text{khi áp suất } CO_2 \text{ cao hoặc thấp} \end{array} \right)$$

$$(\mu_1 + \mu_4) - (\mu_2 + \mu_3) = \left(\begin{array}{c} \text{Độ chênh lệch thể hiện} \\ \text{sự ảnh hưởng của halothane} \\ \text{lên sự sai khác áp suất } CO_2 \end{array} \right)$$

Với $\boldsymbol{\mu} = [\mu_1, \mu_2, \mu_3, \mu_4]^T$, ta có ma trận tương phản $\mathbf{C}$:

$$\mathbf{C} = \begin{bmatrix} -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

Xét thống kê $\mathbf{T}^2 = n(\mathbf{C}\overline{\mathbf{x}})^T(\mathbf{CSC}^T)^{-1}(\mathbf{C}\overline{\mathbf{x}})$ kiểm định giả thuyết $H_0 : \mathbf{C}\boldsymbol{\mu} = \mathbf{0}$.

Khi đó

$$T^2 \sim \frac{(n-1)(q-1)}{n-q+1}\mathcal{F}_{q-1,n-q+1}$$

Giả sử giả thuyết H_0 đúng, từ dữ liệu ta có:

$$\overline{\mathbf{x}} = \begin{bmatrix} 368.21 \\ 404.63 \\ 479.26 \\ 502.89 \end{bmatrix} \quad \text{và} \quad \mathbf{S} = \begin{bmatrix} 2819.29 & & & \\ 3568.42 & 7963.14 & & \\ 2943.49 & 5030.98 & 6851.32 & \\ 2295.35 & 4065.44 & 4499.63 & 4878.99 \end{bmatrix}$$

Khi đó

$$\mathbf{C}\overline{\mathbf{x}} = \begin{bmatrix} 209.31 \\ -60.05 \\ -12.79 \end{bmatrix}; \quad \mathbf{CSC}^T = \begin{bmatrix} 9432.32 & 1098.92 & 927.62 \\ 1098.92 & 5195.84 & 914.54 \\ 927.62 & 914.54 & 7557.44 \end{bmatrix}$$

Trên mẫu thực nghiệm, ta có

$$t^2 = n(\mathbf{C}\overline{\mathbf{x}})^T(\mathbf{CSC}^T)^{-1}(\mathbf{C}\overline{\mathbf{x}}) \approx 19 \times 6.11 \approx 116.$$

Với mức ý nghĩa $\alpha = 0.05$, ta có

$$\begin{aligned} \frac{(n-1)(q-1)}{n-q+1} F_{0.05}(q-1, n-q+1) &= \frac{18 \times 3}{16} F_{0.05}(3, 16) \\ &\approx \frac{18 \times 3}{16} 3.24 \approx 10.94 \end{aligned}$$

Vì $t^2 > \frac{(n-1)(q-1)}{n-q+1} F_{0.05}(q-1, n-q+1)$ nên ta bác bỏ H_0 với mức ý nghĩa 0.05.

Repeated Measures Design

Kiểm định tính tương đồng giữa các treatment

Để xem yếu tố nào ảnh hưởng đến việc bác bỏ H_0, ta xây dựng khoảng tin cậy 95% cho các yếu tố trên. Ảnh hưởng của halothane $=$ $(\mu_3 + \mu_4) - (\mu_1 + \mu_2)$ được ước lượng trong khoảng

$$(\overline{x}_3 + \overline{x}_4) - (\overline{x}_1 + \overline{x}_2) \pm \sqrt{\frac{18 \times 3}{16} F_{0.05}(3, 16)} \sqrt{\frac{\mathbf{c}_1^T \mathbf{S} \mathbf{c}_1}{19}}$$

$$\approx 209.31 \pm \sqrt{10.94} \sqrt{\frac{9432.32}{19}}$$

$$\approx 209.31 \pm 73.70$$

Ảnh hưởng của áp suất $CO_2 = (\mu_1 + \mu_3) - (\mu_2 + \mu_4)$:

$$-60.05 \pm 54.70,$$

Sự tác động lẫn nhau giữa halothane và áp suất CO_2:

$$-12.79 \pm 65.97.$$

So sánh vectơ trung bình từ hai tổng thể

Thống kê T^2 cho ta kiểm định về tính bằng nhau giữa các vectơ trung bình từ 2 tổng thể nhiều chiều. Quy trình kiểm định ở trường hợp nhiều chiều có thể được phát triển khá giống với trường hợp một biến. Thống kê T^2 này phù hợp cho việc so sánh giữa tổng thể này với tổng thể khác có cùng số lượng unit. Nhưng nếu như 2 tổng thể có số lượng unit không bằng nhau thì sao?

So sánh vectơ trung bình từ hai tổng thể

Giả định về cấu trúc bộ dữ liệu

1. Mẫu $\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1}$ là mẫu ngẫu nhiên kích thước n_1 lấy từ tổng thể p chiều có vectơ trung bình $\boldsymbol{\mu}_1$ và ma trận hiệp phương sai $\boldsymbol{\Sigma}_1$.
2. Mẫu $\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2}$ là mẫu ngẫu nhiên kích thước n_2 lấy từ tổng thể p chiều có vectơ trung bình $\boldsymbol{\mu}_2$ và ma trận hiệp phương sai $\boldsymbol{\Sigma}_2$.
3. $\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1}$ độc lập với $\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2}$.

Nếu cả hai cỡ mẫu n_1, n_2 đều nhỏ thì ta cần thêm giả định:

1. Cả hai tổng thể đều có phân phối chuẩn nhiều chiều.
2. $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$.

So sánh vectơ trung bình từ hai tổng thể

Khi $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$, ta có thể gộp (pool) thông tin từ hai mẫu để ước lượng một ma trận hiệp phương sai chung $\mathbf{\Sigma}$ cho hai tổng thể. Khi đó ta có

$$\begin{aligned} S_{pooled} &= \frac{\sum_{j=1}^{n_1}(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)^T + \sum_{j=1}^{n_2}(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)^T}{n_1 + n_2 - 2} \\ &= \frac{n_1 - 1}{n_1 + n_2 - 2}\mathbf{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2}\mathbf{S}_2 \qquad (6.21) \end{aligned}$$

là một ước lượng của $\mathbf{\Sigma}$.

So sánh vectơ trung bình từ hai tổng thể

Ta có

$$E(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) = E(\overline{\mathbf{X}}_1) - E(\overline{\mathbf{X}}_2) = \boldsymbol{\mu}_1 \boldsymbol{\mu}_2$$

và

$$Cov(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) = Cov(\overline{\mathbf{X}}_1) + Cov(\overline{\mathbf{X}}_2) = \frac{1}{n_1}\boldsymbol{\Sigma} + \frac{1}{n_2}\boldsymbol{\Sigma} = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{\Sigma} \tag{6.22}$$

Vì $\mathbf{S}_{pooled}$ là ước lượng của $\boldsymbol{\Sigma}$ nên $\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{\Sigma}$ là ước lượng của $Cov(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)$.

So sánh vectơ trung bình từ hai tổng thể

Xét kiểm định tỷ lệ hợp lý

$$H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_0 \quad \text{và} \quad H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \boldsymbol{\delta}_0$$

với $\boldsymbol{\delta}_0$ là vectơ tham số được chỉ định. Ta dùng thống kê T^2 để kiểm định giả thuyết trên. Ta bác bỏ H_0 nếu

$$T^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta}_0)^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{S}_{pooled} \right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta}_0) > c^2 \tag{6.23}$$

trong đó c^2 được xác định từ phân phối của thống kê T^2 trên hai mẫu.

So sánh vectơ trung bình từ hai tổng thể

Kết quả 6.2

Nếu $\mathbf{X}_{11}, \mathbf{X}_{12}, ..., \mathbf{X}_{1n_1}$ là mẫu ngẫu nhiên kích thước n_1 lấy từ phân phối $\mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ và $\mathbf{X}_{21}, \mathbf{X}_{22}, ..., \mathbf{X}_{2n_2}$ là mẫu ngẫu nhiên kích thước n_2 lấy từ phân phối $\mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ thì

$$T^2 = [\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \mathbf{S}_{pooled}\right]^{-1} [\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]$$

có phân phối

$$\frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} \mathcal{F}_{p, n_1+n_2-p-1}$$

So sánh vectơ trung bình từ hai tổng thể

Kết quả 6.2

Hơn nữa

$$P\left[(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \mathbf{S}_{pooled}\right]^{-1} (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)) \leq c^2\right] = 1 - \alpha \tag{6.24}$$

trong đó

$$c^2 = \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p, n_1+n_2-p-1}(\alpha)$$

So sánh vectơ trung bình từ hai tổng thể

Ví dụ 6.3 (Xây dựng miền tin cậy cho differece giữa hai vectơ trung bình)

Xét bài toán sản xuất xà phòng. Có 50 thanh xà phòng được sản xuất theo hai cách. Có 2 đặc tính được quan tâm dùng để đánh giá là $X_1 =$ *lượng bọt* và $X_2 =$ *tính dịu nhẹ*. Quá trình sản xuất theo hai phương pháp có các thống kê được tóm tắt như sau:

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 8.3 \\ 4.1 \end{bmatrix}. \qquad \mathbf{S}_1 = \begin{bmatrix} 2 & 1 \\ 1 & 6 \end{bmatrix}$$

$$\bar{\mathbf{x}}_2 = \begin{bmatrix} 10.2 \\ 3.9 \end{bmatrix}. \qquad \mathbf{S}_2 = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$$

Ta cần tìm vùng tin cậy 95% cho $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

So sánh vectơ trung bình từ hai tổng thể

Do $\mathbf{S}_1$ và $\mathbf{S}_2$ xấp xỉ bằng nhau nên ta có thể gộp chúng lại thành

$$\begin{aligned} \mathbf{S}_{pooled} &= \frac{n_1 - 1}{n_1 + n_2 - 2}\mathbf{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2}\mathbf{S}_2 \\ &= \frac{49}{98}\mathbf{S}_1 + \frac{49}{98}\mathbf{S}_2 \\ &= \begin{bmatrix} 2 & 1 \\ 1 & 5 \end{bmatrix} \end{aligned}$$

Hơn nữa

$$\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 = \begin{bmatrix} -1.9 \\ 0.2 \end{bmatrix}$$

Suy ra confidence ellipse có tâm tại $[-1.9, 0.2]^T$.

So sánh vectơ trung bình từ hai tổng thể

Các trị riêng và vectơ riêng của $\mathbf{S}_{pooled}$ thu được từ phương trình

$$0 = |\mathbf{S}_{pooled} - \lambda \mathbf{I}| = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 5-\lambda \end{vmatrix} = \lambda^2 - 7\lambda + 9$$

Suy ra $\lambda = \dfrac{7 \pm \sqrt{49-36}}{2}$ hay $\lambda_1 \approx 5.303$ và $\lambda_2 \approx 1.697$, và các vectơ riêng tương ứng là $\mathbf{e}_1$ và $\mathbf{e}_2$ được xác định từ

$$\mathbf{S}_{pooled}\mathbf{e}_i = \lambda_i \mathbf{e}_i, \quad i = 1, 2$$

là

$$\mathbf{e}_1 = \begin{bmatrix} 0.290 \\ 0.957 \end{bmatrix} \quad \text{và} \quad \mathbf{e}_2 = \begin{bmatrix} 0.597 \\ -0.290 \end{bmatrix}$$

So sánh vectơ trung bình từ hai tổng thể

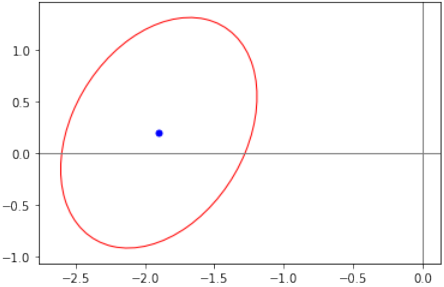
Với mức ý nghĩa $\alpha = 0.05$, theo kết quả 6.2 ta có

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right) c^2 = \left(\frac{1}{50} + \frac{1}{50}\right) \frac{98 \times 2}{97} F_{0.05}(2, 97) \approx 0.08 \times 3.1 \approx 0.25$$

Khi đó confidence ellipse được mở rộng theo

$$\sqrt{\lambda_i}\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) c^2} = \sqrt{\lambda_i}\sqrt{0.25}$$

So sánh vectơ trung bình từ hai tổng thể



Ta có thể thấy $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$ không nằm trong ellipse. Do đó ta có thể kết luận rằng hai phương pháp sản xuất xà phòng có sự khác biệt. Sự sai khác ở đây được tạo ra từ việc quá trình thứ 2 tạo ra sản phẩm có lượng bọt (X_1) cao hơn, trong khi độ dịu nhẹ (X_2) thì gần như tương đồng ở cả hai quá trình.

So sánh vectơ trung bình từ hai tổng thể

Khoảng tin cậy đồng thời

Khoảng tin cậy đồng thời cho từng thành phần của vectơ $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ có thể được tổng quát hoá cho tất cả các tổ hợp tuyến tính có thể có của các difference trong các vectơ trung bình. Giả sử rằng các tổng thể đa chiều ban đầu có phân phối chuẩn với hiệp phương sai là $\boldsymbol{\Sigma}$.

So sánh vectơ trung bình từ hai tổng thể

Khoảng tin cậy đồng thời

Kết quả 6.3

Cho $c^2 = \frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)} F_{p,n_1+n_2-p-1}(\alpha)$. Với xác suất $1-\alpha$ thì

$$\mathbf{a}^T(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) \pm c\sqrt{\mathbf{a}^T\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\mathbf{S}_{pooled}\mathbf{a}}$$

sẽ bao hết $\mathbf{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ với mọi $\mathbf{a}$. Cụ thể hơn là $\mu_{1i} - \mu_{2i}$ sẽ bị bao bởi

$$(\overline{X}_{1i} - \overline{X}_{2i}) \pm c\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S_{ii,pooled}} \quad \text{với } i = 1, 2, ..., p$$

So sánh vectơ trung bình từ hai tổng thể

Khoảng tin cậy đồng thời

Nhận xét

Với kiểm định giả thuyết $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$, tổ hợp tuyến tính $\hat{\mathbf{a}}^T(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$ với vectơ hệ số $\hat{\mathbf{a}} \propto \mathbf{S}_{pooled}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$ ước lượng difference tổng thể lớn nhất. Nghĩa là, nếu thống kê T^2 bác bỏ giả thuyết H_0 thì $\hat{\mathbf{a}}^T(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$ sẽ có trung bình khác không.

Ví dụ 6.4 (Tìm khoảng tin cậy đồng thời cho differences giữa các thành phần của trung bình)

Các mẫu có kích thước $n_1 = 45$ và $n_2 = 55$ được lấy từ các chủ nhà ở Wisconsin tương ứng với việc dùng hay không dùng máy điều hoà. Hai phép đo lượng sử dụng điện (tính bằng kWh): trong giờ cao điểm (X_1) và ngoài giờ cao điểm (X_2) trong suốt tháng 7. Ta có các kết quả tóm tắt như sau

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 204.4 \\ 556.6 \end{bmatrix}, \qquad \mathbf{S}_1 = \begin{bmatrix} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{bmatrix}, \quad n_1 = 45$$

$$\bar{\mathbf{x}}_2 = \begin{bmatrix} 130.0 \\ 355.0 \end{bmatrix}, \qquad \mathbf{S}_1 = \begin{bmatrix} 8632.0 & 19616.7 \\ 19616.7 & 55964.5 \end{bmatrix}, \quad n_1 = 55$$

Ta cần tìm khoảng tin cậy đồng thời 95% cho sự chênh lệch giữa các thành phần của vectơ trung bình.

So sánh vectơ trung bình từ hai tổng thể

Khoảng tin cậy đồng thời

Ta có

$$\begin{aligned}\mathbf{S}_{pooled} &= \frac{n_1 - 1}{n_1 + n_2 - 2}\mathbf{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2}\mathbf{S}_2 \\ &= \begin{bmatrix} 10963.7 & 21505.5 \\ 21505.5 & 63661.3 \end{bmatrix}\end{aligned}$$

và với mức ý nghĩa $\alpha = 0.05$ thì

$$\begin{aligned}c^2 &= \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_\alpha(p, n_1 + n_2 - p - 1) \\ &= \frac{98 \times 2}{97} F_{0.05}(2, 97) \\ &\approx 2.02 \times 3.1 = 6.26\end{aligned}$$

So sánh vectơ trung bình từ hai tổng thể

Khoảng tin cậy đồng thời

Với $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = [\mu_{11} - \mu_{21}, \mu_{12} - \mu_{22}]^T$ thì khoảng tin cậy đồng thời 95% cho sự sai khác của tổng thể là

$$\mu_{11} - \mu_{21} : (204.4 - 130.0) \pm \sqrt{6.26}\sqrt{\left(\frac{1}{45} + \frac{1}{55}\right) 10963.7}$$

hoặc

$$21.7 \leq \mu_{11} - \mu_{21} \leq 127.1 \qquad \text{(cao điểm)}$$

$$\mu_{12} - \mu_{22} : (556.6 - 355.0) \pm \sqrt{6.26}\sqrt{\left(\frac{1}{45} + \frac{1}{55}\right) 63661.3}$$

hoặc

$$74.7 \leq \mu_{12} - \mu_{22} \leq 328.5 \qquad \text{(ngoài cao điểm)}$$

So sánh vectơ trung bình từ hai tổng thể

Khoảng tin cậy đồng thời

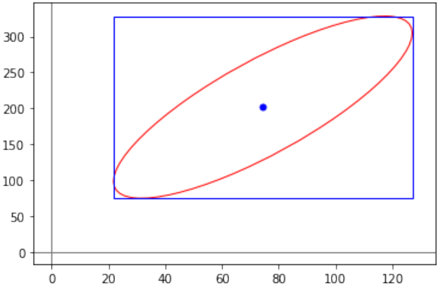
Ta có thể kết luận rằng có sự chênh lệch trong việc tiêu thụ điện giữa những nhà có và không có sử dụng máy điều hoà. Sự chênh lệch này cũng thể hiện rõ ở cả vào giờ cao điểm và ngoài cao điểm.

Hình confidence ellipse 95% cho $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ được xác định từ các cặp trị riêng-vectơ riêng $\lambda_1 = 71323.5, \mathbf{e}_1 = [0.336, 0.942]^T$ và $\lambda_2 = 3301.5, \mathbf{e}_2 = [0.942, -0.336]^T$. Khi đó

$$\sqrt{\lambda_1}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)c^2} \approx 134.4 \quad \text{và} \quad \sqrt{\lambda_2}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)c^2} \approx 28.9$$

So sánh vectơ trung bình từ hai tổng thể

Khoảng tin cậy đồng thời



Vì confidence ellipse cho sự sai khác ở trung bình không chứa $\mathbf{0} = [0, 0]^T$ nên thống kê T^2 bác bỏ giả thuyết $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$ ở mức ý nghĩa 5%.

So sánh vectơ trung bình từ hai tổng thể

Khoảng tin cậy đồng thời

Với độ tin cậy $100(1-\alpha)\%$, khoảng tin cậy Bonferroni đồng thời cho p difference trung bình tổng thể là

$$\mu_{1i} - \mu_{2i} : (\overline{x}_{1i} - \overline{x}_{2i}) \pm t_{n_1+n_2-2}\left(\frac{\alpha}{2p}\right)\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)s_{ii,pooled}}$$

Trong đó, $t_{n_1+n_2-2}(\dfrac{\alpha}{2p})$ là phân vị trên thứ $100(\alpha/2p)$ của phân phối Student với bậc tự do n_1+n_2-2.

So sánh vectơ trung bình từ hai tổng thể

Trường hợp $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$

- Khi $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, ta không thể tìm ra phép đo "khoảng cách" nào như thống kê T^2 mà phân phối của nó không phụ thuộc vào $\mathbf{\Sigma}_1$ và $\mathbf{\Sigma}_2$.
- Kiểm định Bartlett được sử dụng để kiểm tra về tính bằng nhau của $\mathbf{\Sigma}_1$ và $\mathbf{\Sigma}_2$. Tuy nhiên, nếu cả hai tổng thể đều không chuẩn thì kiểm định Bartlett không còn đúng.
- Trong trường hợp cỡ mẫu lớn, ta có thể tránh được vấn đề trên.

So sánh vectơ trung bình từ hai tổng thể

Trường hợp $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$ với cỡ mẫu lớn

Kết quả 6.4

Cho các cỡ mẫu thoả mãn $n_1 - p$ và $n_2 - p$ đều lớn. Khi đó, một xấp xỉ confidence ellipsoid với độ tin cậy $100(1-\alpha)\%$ cho $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ được cho bởi tất cả $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ thoả mãn

$$[\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^T \left[\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right]^{-1} [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \le \chi_p^2(\alpha)$$

Hơn nữa, khoảng tin cậy đồng thời $100(1-\alpha)\%$ cho tất cả các tổ hợp tuyến tính $\mathbf{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ là

$$\mathbf{a}^T(\bar{x}_1 - \bar{x}_2) \pm \sqrt{\chi_p^2(\alpha)}\sqrt{\mathbf{a}^T\left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right)\mathbf{a}}$$

So sánh vectơ trung bình từ hai tổng thể

Trường hợp $\Sigma_1 \neq \Sigma_2$ với cỡ mẫu lớn

Nhận xét

Nếu $n_1 = n_2 = n$ thì $\dfrac{n-1}{n+n-2} = \dfrac{1}{2}$, do đó

$$\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2 = \frac{1}{n}(\mathbf{S}_1 + \mathbf{S}_2) = \frac{(n-1)\mathbf{S}_1 + (n-1)\mathbf{S}_2}{n+n-2}\left(\frac{1}{n} + \frac{1}{n}\right)$$
$$= \mathbf{S}_{pooled}\left(\frac{1}{n} + \frac{1}{n}\right)$$

Với cỡ mẫu bằng nhau, phương pháp mẫu lớn về cơ bản giống với phương pháp dựa trên ma trận hiệp phương sai gộp. Trong một chiều, ảnh hưởng của các phương sai không bằng nhau nhỏ nhất khi $n_1 = n_2$ và lớn nhất khi $n_1 \ll n_2$ hoặc ngược lại.

So sánh vectơ trung bình từ hai tổng thể

Trường hợp $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ với cỡ mẫu lớn

Ví dụ 6.5 (Phương pháp mẫu lớn cho suy luận về differences trong kỳ vọng)

Ta sẽ đi phân tích lượng điện tiêu thụ được thảo luận trong ví dụ 6.4 bằng cách tiếp cận mẫu lớn.

So sánh vectơ trung bình từ hai tổng thể

Trường hợp $\Sigma_1 \neq \Sigma_2$ với cỡ mẫu lớn

Ta có các kết quả sau được lấy từ ví dụ 6.4

$$\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2 = \frac{1}{45}\begin{bmatrix} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{bmatrix} + \frac{1}{55}\begin{bmatrix} 8632.0 & 19616.7 \\ 19616.7 & 55964.5 \end{bmatrix}$$
$$= \begin{bmatrix} 464.17 & 886.08 \\ 886.08 & 2642.15 \end{bmatrix}$$

và khoảng tin cậy đồng thời 95% cho $\mu_{11} - \mu_{21}$ và $\mu_{12} - \mu_{22}$ khi phương pháp gộp (pooling) được áp dụng là

$$\mu_{11} - \mu_{21} : 74.4 \pm \sqrt{5.99}\sqrt{464.17} \quad \text{hay} \quad (21.7, 127.1)$$
$$\mu_{12} - \mu_{22} : 201.6 \pm \sqrt{5.99}\sqrt{2642.15} \quad \text{hay} \quad (74.7, 328.5)$$

So sánh vectơ trung bình từ hai tổng thể

Trường hợp $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ với cỡ mẫu lớn

Xét thống kê T^2 cho kiểm định giả thuyết $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$ là

$$
\begin{aligned}
T^2 &= [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2]^T \left[\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right]^{-1} [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2] \\
&\approx 15.66
\end{aligned}
$$

Với mức ý nghĩa $\alpha = 0.05$, ta có giá trị tới hạn là $\chi_2^2(0.05) \approx 5.99$, vì $T^2 \approx 15.66 > \chi_2^2(0.05) \approx 5.99$ nên ta bác bỏ giả thuyết H_0.

So sánh vectơ trung bình từ hai tổng thể

Trường hợp $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ với cỡ mẫu lớn

Khi đó, tổ hợp tuyến tính dẫn đến việc bác bỏ H_0 có vectơ hệ số

$$\hat{\mathbf{a}} \propto \left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2 \right)^{-1} = \begin{bmatrix} 0.041 \\ 0.063 \end{bmatrix}$$

Ta có thể thấy rằng, sự chệnh lệch lượng điên tiêu thụ ngoài giờ cao điểm giữa những nhà có và không có máy điều hoà đóng góp nhiều hơn với sự chênh lệch tương ứng ở giờ cao điểm cho việc bác bỏ giả thuyết $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$.

So sánh vectơ trung bình từ hai tổng thể

Trường hợp $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$ với cỡ mẫu nhỏ

- Cho hai tổng thể có phân phối chuẩn nhiều chiều có cỡ mẫu n_1 và n_2 đều nhỏ và $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$.
- Trường hợp trên được gọi là bài toán Behrens-Fisher nhiều chiều. Điều kiện cho bài toán trên là cả hai cỡ mẫu n_1 và n_2 đều phải lớn hơn số lượng biến p.
- Phương pháp Behrens-Fisher là một xấp xỉ của thống kê T^2. Thay vì sử dụng xấp xỉ phân phối chi bình phương để tìm ra một giá trị để kiểm định giả thuyết H_0 thì ta tìm một xấp xỉ mới cho cho cỡ mẫu nhỏ hơn.

So sánh vectơ trung bình từ hai tổng thể

Trường hợp $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ với cỡ mẫu nhỏ

Xấp xỉ được đưa ra là

$$T^2 = \frac{vp}{v-p+1}\mathcal{F}_{p,v-p+1} \tag{6.28}$$

trong đó bậc tự do v được ước lượng từ ma trận hiệp phương sai mẫu như sau:

$$v = \frac{p+p^2}{\sum_{i=1}^{2}\frac{1}{n_i}\left\{tr\left[\left(\frac{1}{n_i}\mathbf{S}_i\left(\frac{1}{n_1}\mathbf{S}_1+\frac{1}{n_2}\mathbf{S}_2\right)^{-1}\right)^2\right]+\left(tr\left[\frac{1}{n_i}\mathbf{S}_i\left(\frac{1}{n_1}\mathbf{S}_1+\frac{1}{n_2}\mathbf{S}_2\right)^{-1}\right]\right)^2\right\}} \tag{6.29}$$

trong đó $\min(n_1, n_2) \leq v \leq n_1 + n_2$.

So sánh vectơ trung bình từ hai tổng thể

Trường hợp $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ với cỡ mẫu nhỏ

Với cỡ mẫu trung bình từ 2 tổng thể có phân phối chuẩn, xét kiểm định với mức ý nghĩa α của giả thuyết: $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = 0$. Ta bác bỏ H_0 nếu:

$$[\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^T \left[\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right]^{-1} [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] > \frac{vp}{v-p+1} F_{p,v-p+1}(\alpha)$$

Tương tự, miền tin cậy xấp xỉ với độ tin cậy $100(1-\alpha)\%$ cho bởi tất cả $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ thoả mãn

$$[\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^T \left[\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right]^{-1} [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \leq \frac{vp}{v-p+1} F_{p,v-p+1}(\alpha) \tag{6.30}$$

So sánh vectơ trung bình từ hai tổng thể

Trường hợp $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ với cỡ mẫu nhỏ

Ví dụ 6.6 (Xấp xỉ phân phối T^2 khi $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$)

Mặc dù dữ liệu mức tiêu thụ điện năng ở ví dụ 6.4 có cỡ mẫu khá lớn nhưng ta vẫn sử dụng lại dữ liệu này kết hợp với các tính toán ở ví dụ 6.5 để minh hoạ các phép tính dẫn đến xấp xỉ phân phối T^2 khi các ma trận hiệp phương sai tổng thể không bằng nhau.

So sánh vectơ trung bình từ hai tổng thể

Trường hợp $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$ với cỡ mẫu nhỏ

Từ ví dụ 6.5 ta có

$$\frac{1}{n_1}\mathbf{S}_1 = \begin{bmatrix} 307.227 & 529.409 \\ 529.409 & 1624.609 \end{bmatrix}$$

$$\frac{1}{n_2}\mathbf{S}_2 = \begin{bmatrix} 156.945 & 356.667 \\ 356.667 & 1017.536 \end{bmatrix}$$

$$\left[\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right]^{-1} = (10^{-4})\begin{bmatrix} 59.874 & -20.080 \\ -20.080 & 10.519 \end{bmatrix}$$

So sánh vectơ trung bình từ hai tổng thể

Trường hợp $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$ với cỡ mẫu nhỏ

Khi đó, ta có

$$\frac{1}{n_1}\mathbf{S}_1\left[\frac{1}{n_1}\mathbf{S}_1+\frac{1}{n_2}\mathbf{S}_2\right]^{-1}=\begin{bmatrix}0.776 & -0.60\\ -0.092 & 0.646\end{bmatrix}$$

và

$$\left(\frac{1}{n_1}\mathbf{S}_1\left[\frac{1}{n_1}\mathbf{S}_1+\frac{1}{n_2}\mathbf{S}_2\right]^{-1}\right)^2=\begin{bmatrix}0.608 & -0.085\\ -0.131 & 0.423\end{bmatrix}$$

Hơn nữa

$$\frac{1}{n_2}\mathbf{S}_2\left[\frac{1}{n_1}\mathbf{S}_1+\frac{1}{n_2}\mathbf{S}_2\right]^{-1}=\begin{bmatrix}0.224 & -0.60\\ 0.092 & 0.354\end{bmatrix}$$

và

$$\left(\frac{1}{n_2}\mathbf{S}_1\left[\frac{1}{n_1}\mathbf{S}_1+\frac{1}{n_2}\mathbf{S}_2\right]^{-1}\right)^2=\begin{bmatrix}0.055 & 0.035\\ 0.053 & 0.131\end{bmatrix}$$

So sánh vectơ trung bình từ hai tổng thể

Trường hợp $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$ với cỡ mẫu nhỏ

$$\frac{1}{n_1}\left\{tr\left[\left(\frac{1}{n_1}\mathbf{S}_1\left(\frac{1}{n_1}\mathbf{S}_1+\frac{1}{n_2}\mathbf{S}_2\right)^{-1}\right)^2\right]+\left(tr\left[\frac{1}{n_1}\mathbf{S}_1\left(\frac{1}{n_1}\mathbf{S}_1+\frac{1}{n_2}\mathbf{S}_2\right)^{-1}\right]\right)^2\right\} \approx 0.0678$$

$$\frac{1}{n_2}\left\{tr\left[\left(\frac{1}{n_2}\mathbf{S}_2\left(\frac{1}{n_1}\mathbf{S}_1+\frac{1}{n_2}\mathbf{S}_2\right)^{-1}\right)^2\right]+\left(tr\left[\frac{1}{n_2}\mathbf{S}_2\left(\frac{1}{n_1}\mathbf{S}_1+\frac{1}{n_2}\mathbf{S}_2\right)^{-1}\right]\right)^2\right\} \approx 0.0095$$

Ta có ước lượng của hệ số tự do v là

$$v \approx \frac{2+2^2}{0.0678+0.0095} \approx 77.6$$

So sánh vectơ trung bình từ hai tổng thể

Trường hợp $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$ với cỡ mẫu nhỏ

Với mức ý nghĩa $\alpha = 0.05$, ta có giá trị tới hạn

$$\frac{vp}{v-p+1} F_{p,v-p+1}(0.05) \approx 6.32$$

Từ ví dụ 6.5, giá trị quan sát của thống kê kiểm định T^2 là 15.66. Do đó, giả thuyết $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$ bị bác bỏ với mức ý nghĩa 5%. Phương pháp trên có cùng kết luận với phương pháp mẫu lớn ở ví dụ 6.5.

So sánh nhiều trung bình của tổng thể nhiều chiều (One-Way MANOVA)

- MANOVA viết tắt cho từ Multivariate Analysis of Variance.
- Thông thường, có nhiều hơn hai tổng thể cần được so sánh. Các mẫu ngẫu nhiên lấy từ g tổng thể

$$\begin{array}{ll} \text{Tổng thể 1:} & \mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1} \\ \text{Tổng thể 2:} & \mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2} \\ \vdots & \vdots \\ \text{Tổng thể } g\text{:} & \mathbf{X}_{g1}, \mathbf{X}_{g2}, \ldots, \mathbf{X}_{gn_g} \end{array} \tag{6.31}$$

- Ban đầu MANOVA được sử dụng để kiểm định các vectơ trung bình tổng thể có bằng nhau hay không.

So sánh nhiều trung bình của tổng thể nhiều chiều

Giả định về cấu trúc bộ dữ liệu cho MANOVA

1. $\mathbf{X}_{\ell 1}, \mathbf{X}_{\ell 2}, ..., \mathbf{X}_{\ell n_\ell}$ là một mẫu ngẫu nhiên kích thước n_ℓ lấy từ tổng thể có trung bình $\boldsymbol{\mu}_\ell, \ell = 1, 2, ..., g$. Mẫu ngẫu nhiên lấy từ tổng thể khác nhau thì độc lập.
2. Tất cả tổng thể có cùng ma trận hiệp phương sai $\boldsymbol{\Sigma}$
3. Mỗi tổng thể có phân phối chuẩn nhiều chiều.

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp một chiều

Trong trường hợp một chiều, giả sử $X_{\ell 1}, X_{\ell 2}, \ldots, X_{\ell n_\ell}$ là một mẫu ngẫu nhiên lấy từ tổng thể có phân phối $\mathcal{N}(\mu_\ell, \sigma^2)$, $\ell = 1, 2, \ldots, g$ và các mẫu ngẫu nhiên độc lập với nhau. Ta xét kiểm định giả thuyết $H_0 : \mu_1 = \mu_2 = \ldots = \mu_g$. Khi đó ta có thể tham số hoá như sau

$$\begin{array}{ccccc} \mu_\ell & = & \mu & + & \tau_\ell \\ \begin{pmatrix} \text{trung bình} \\ \text{tổng thể} \\ \text{thứ } \ell \end{pmatrix} & & \begin{pmatrix} \text{trung bình} \\ \text{tổng thể} \end{pmatrix} & & \begin{pmatrix} \text{ảnh hưởng} \\ \text{của tổng thể} \\ \text{(treatment)} \\ \text{thứ } \ell \end{pmatrix} \end{array} \tag{6.32}$$

trong đó $\tau_\ell = \mu_\ell - \mu$.

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp một chiều

Giả thuyết kiểm định lúc này trở thành

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_g = 0$$

Response $X_{\ell j} \sim \mathcal{N}(\mu + \tau_\ell, \sigma^2)$, có thể được biểu diễn thành Khi đó,

$$X_{\ell j} = \underset{\begin{pmatrix}\text{trung bình}\\ \text{tổng thể}\end{pmatrix}}{\mu} + \underset{\begin{pmatrix}\text{ảnh hưởng}\\ \text{của treatment}\end{pmatrix}}{\tau_\ell} + \underset{\begin{pmatrix}\text{error}\\ \text{ngẫu nhiên}\end{pmatrix}}{e_{\ell j}} \tag{6.33}$$

trong đó $e_{\ell j}$ là các biến ngẫu nhiên độc lập có phân phối $\mathcal{N}(0, \sigma^2)$ và $\sum_{\ell=1}^{g} n_\ell \tau_\ell = 0$.

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp một chiều

Như vậy, các giá trị quan trắc của $\mathbf{X}_{\ell j}$ cũng có thể được biểu diễn dưới dạng

$$x_{\ell j} = \overline{x} + (\overline{x}_\ell - \overline{x}) + (x_{\ell j} - \overline{x}_\ell) \tag{6.34}$$

trong đó $\overline{x}$, $\hat{\tau}_\ell = (\overline{x}_\ell - \overline{x})$ và $(x_{\ell j} - \overline{x}_\ell)$ lần lượt là ước lượng của μ, τ_ℓ và $e_{\ell j}$.

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp một chiều

Ví dụ 6.7 (Tổng bình phương phân rã cho ANOVA một chiều)

Xét các mẫu độc lập

$$\begin{aligned} &\text{Tổng thể 1}: \quad 9, 6, 9 \\ &\text{Tổng thể 2}: \quad 0, 2 \\ &\text{Tổng thể 3}: \quad 3, 1, 2 \end{aligned}$$

Ta có

$$\begin{array}{ccccccc} \begin{bmatrix} 9 & 6 & 9 \\ 0 & 2 & \\ 3 & 1 & 2 \end{bmatrix} & = & \begin{bmatrix} 4 & 4 & 4 \\ 4 & 4 & \\ 4 & 4 & 4 \end{bmatrix} & + & \begin{bmatrix} 4 & 4 & 4 \\ -3 & -3 & \\ -2 & -2 & -2 \end{bmatrix} & + & \begin{bmatrix} 1 & -2 & 1 \\ -1 & 1 & \\ 1 & -1 & 0 \end{bmatrix} \\ (x_{\ell j}) & = & (\bar{x}) & + & (\bar{x}_\ell - \bar{x}) & + & (x_{\ell j} - \bar{x}_\ell) \end{array}$$

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp một chiều

Phát biểu giả thuyết

$$H_0 : \tau_1 = \tau_2 = \tau_3 = 0$$

Ta sẽ bác bỏ giả thuyết H_0 nếu các treatment đều lớn. Hơn nữa ta có ước lượng của τ_ℓ là $\hat{\tau}_\ell = \overline{x}_\ell - \overline{x}$ thoả mãn $\sum_{\ell=1}^{3} n_\ell \hat{\tau}_\ell = 0$.

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp một chiều

Mặt khác, ta xây dựng vectơ $\mathbf{y} = [9, 6, 9, 0, 2, 3, 1, 2]^T$ là vectơ cột chứa tất cả các quan trắc từ 3 tổng thể. Khi đó ta có độ dài bình phương của vectơ $\mathbf{y}$ là

$$SS_{obs} = 9^2 + 6^2 + 9^2 + 0^2 + 2^2 + 3^2 + 1^2 + 2^2 = 216$$

Tương tự

$$SS_{mean} = 4^2 + 4^2 + 4^2 + 4^2 + 4^2 + 4^2 + 4^2 + 4^2 = 128$$

$$SS_{tr} = 4^2 + 4^2 + 4^2 + (-3)^2 + (-3^2) + (-2)^2 + (-2)^2 + (-2)^2 = 78$$

$$SS_{res} = 1^2 + (-2)^2 + 1^2 + (-1)^2 + 1^2 + 1^2 + (-1)^2 + 0^2 = 10$$

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp một chiều

Các tổng bình phương thoả mãn cùng một phân rã ở (6.34) như các quan trắc. Khi đó

$$\begin{array}{ccccccccc} SS_{obs} & = & SS_{mean} & + & SS_{tr} & + & SS_{res} \\ 216 & = & 128 & + & 78 & + & 10 \end{array}$$

Như vậy, tổng bình phương các giá trị quan trắc của các mẫu được chia thành các thành phần mean, treatment và residual (error).

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp một chiều

Trường hợp tổng quát của ví dụ 6.7

$$\begin{array}{ccccc} \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(x_{\ell j}-\overline{x})^2 & = & \sum_{\ell=1}^{g} n_\ell(\overline{x}_\ell-\overline{x})^2 & + & \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(x_{\ell j}-\overline{x}_\ell)^2 \\ (SS_{cor}) & = & (SS_{tr}) & + & (SS_{res}) \end{array} \tag{6.35}$$

hay

$$\begin{array}{ccccccc} \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}x_{\ell j}^2 & = & (n_1+\ldots+n_g)\overline{x}^2 & + & \sum_{\ell=1}^{g} n_\ell(\overline{x}_\ell-\overline{x})^2 & + & \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(x_{\ell j}-\overline{x}_\ell)^2 \\ (SS_{obs}) & = & (SS_{mean}) & + & (SS_{tr}) & + & (SS_{res}) \end{array} \tag{6.36}$$

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp một chiều

Các phép tính tổng bình phương và bậc tự do tương ứng được cho trong bảng ANOVA

Source of variance	Sum of squares (SS)	Degrees of freedom (d.f.)
Treatment	$SS_{tr} = \sum_{\ell=1}^{g} n_\ell(\overline{x}_\ell - \overline{x})^2$	$g - 1$
Residual (error)	$SS_{res} = \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(x_{\ell j} - \overline{x}_\ell)^2$	$\sum_{\ell=1}^{g} n_\ell - g$
Total (corrected for the mean)	$SS_{cor} = \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(x_{\ell j} - \overline{x})^2$	$\sum_{\ell=1}^{g} n_\ell - 1$

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp một chiều

Tổng bình phương của treatment effects: Xét kiểm định $H_0 : \tau_1 = \tau_2 = ... = \tau_g = 0$ với mức ý nghĩa α. Ta bác bỏ H_0 nếu:

$$F = \frac{SS_{tr}/(g-1)}{SS_{res} \Big/ \left(\sum\limits_{\ell=1}^{g} n_\ell - g \right)} > F_{g-1, \Sigma n_\ell - g}(\alpha)$$

trong đó $F_{g-1, \Sigma n_\ell - g}(\alpha)$ là phân vị trên thứ (100α) của phân phối Fisher với bậc tự do là $g-1$ và $\Sigma n_\ell - g$.

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp một chiều

Ví dụ 6.8 (Bảng ANOVA một chiều và kiểm định F cho treatment effects)

Sử dụng thông tin của ví dụ 6.7 ta có bảng ANOVA như sau

Source of variance	Sum of squares	Degrees of freedom
Treatment	$SS_{tr} = 78$	$g - 1 = 2$
Residual	$SS_{res} = 10$	$\sum_{\ell=1}^{g} n_\ell - g = 5$
Total (corrected)	$SS_{cor} = 88$	$\sum_{\ell=1}^{g} n_\ell - 1 = 7$

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp một chiều

Trên mẫu thực nghiệm ta có

$$F = \frac{SS_{tr}/(g-1)}{SS_{res}/(\Sigma n_\ell - g)} = \frac{78/2}{10/5} \approx 19.5$$

Với mức ý nghĩa 1% ta có $F_{g-1,\Sigma n_\ell - g}(0.01) = F_{2,5}(0.01) \approx 13.27$. Vì $F = 19.5 > F_{2,5}(0.01) = 13.27$ nên ta bác bỏ giả thuyết $H_0 : \tau_1 = \tau_2 = \tau_3 = 0$ với mức ý nghĩa 1%.

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp nhiều chiều

Với trường hợp nhiều chiều, MANOVA ta có phân tích sau:

$$\sum_{\ell=1}^{g}\sum_{j=1}^{n_l}(\mathbf{x}_{\ell j}-\overline{x})(\mathbf{x}_{\ell j}-\overline{x})^T$$

$$=\sum_{\ell=1}^{g} n_\ell(\overline{\mathbf{x}}_\ell-\overline{\mathbf{x}})(\overline{\mathbf{x}}_\ell-\overline{\mathbf{x}})^T+\sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(\mathbf{x}_{\ell j}-\overline{x}_\ell)(\mathbf{x}_{\ell j}-\overline{\mathbf{x}}_\ell)^T \quad (6.40)$$

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp nhiều chiều

Ta có thể viết lại tổng bình phương sai số như sau:

$$\begin{aligned}\mathbf{W} &= \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_\ell)(\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_\ell)^T \\ &= (n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + \ldots + (n_g - 1)\mathbf{S}_g \qquad (6.41)\end{aligned}$$

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp nhiều chiều

Ta có bảng MANOVA như sau

Source of variance	Matrix of sum of squares and cross products (SSP)	Degrees of freedom (d.f.)
Treatment	$\mathbf{B} = \sum_{\ell=1}^{g} n_\ell (\bar{\mathbf{x}}_\ell - \bar{\mathbf{x}})(\bar{\mathbf{x}}_\ell - \bar{\mathbf{x}})^T$	$g - 1$
Residual (error)	$\mathbf{W} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_\ell)(\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_\ell)^T$	$\sum_{\ell=1}^{g} n_\ell - g$
Total (corrected for the mean)	$\mathbf{B} + \mathbf{W} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (\mathbf{x}_{\ell j} - \bar{\mathbf{x}})(\mathbf{x}_{\ell j} - \bar{\mathbf{x}})^T$	$\sum_{\ell=1}^{g} n_\ell - 1$

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp nhiều chiều

Khi đó, ta xét kiểm định:

$$H_0 : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \ldots = \boldsymbol{\tau}_g = \mathbf{0}$$

Ta bác bỏ H_0 nếu tỉ lệ sau rất nhỏ:

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|} = \frac{\left|\sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(\mathbf{x}_{\ell j}-\bar{\mathbf{x}}_\ell)(\mathbf{x}_{\ell j}-\bar{\mathbf{x}}_\ell)^T\right|}{\left|\sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(\mathbf{x}_{\ell j}-\bar{\mathbf{x}}_\ell)(\mathbf{x}_{\ell j}-\bar{\mathbf{x}}_\ell)^T\right|} \tag{6.42}$$

Giá trị $\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}$ gọi là Số Wilk's Lambda.

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp nhiều chiều

Bảng phân phối của Wilk's Lambda:

Table 6.3 Distribution of Wilks' Lambda, $\Lambda^* = |\mathbf{W}|/|\mathbf{B} + \mathbf{W}|$

No. of variables	No. of groups	Sampling distribution for multivariate normal data
$p = 1$	$g \geq 2$	$\left(\frac{\sum n_\ell - g}{g - 1}\right)\left(\frac{1 - \Lambda^*}{\Lambda^*}\right) \sim F_{g-1, \sum n_\ell - g}$
$p = 2$	$g \geq 2$	$\left(\frac{\sum n_\ell - g - 1}{g - 1}\right)\left(\frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2(g-1), 2(\sum n_\ell - g - 1)}$
$p \geq 1$	$g = 2$	$\left(\frac{\sum n_\ell - p - 1}{p}\right)\left(\frac{1 - \Lambda^*}{\Lambda^*}\right) \sim F_{p, \sum n_\ell - p - 1}$
$p \geq 1$	$g = 3$	$\left(\frac{\sum n_\ell - p - 2}{p}\right)\left(\frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2p, 2(\sum n_\ell - p - 2)}$

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp nhiều chiều

Bổ đề

Nếu H_0 đúng và $\Sigma n_\ell = n$ lớn. Khi đó

$$-\left(n - 1 - \frac{(p+g)}{2}\right)\ln(\Lambda^*) \tag{6.43}$$

sẽ xấp xỉ phân phối chi bình phương với bậc tự do $p(g-1)$.

Nghĩa là, với $\Sigma n_\ell = n$ lớn, ta bác bỏ H_0 với mức ý nghĩa α nếu

$$-\left(n - 1 - \frac{(p+g)}{2}\right)\ln(\Lambda^*) > \chi^2_{p(g-1)}(\alpha) \tag{6.44}$$

So sánh nhiều trung bình của tổng thể nhiều chiều

Ví dụ 6.9 (Bảng MANOVA và thống kê Wilks' lambda cho kiểm định ba vectơ trung bình bằng nhau)

Giả sử có một biến bổ sung được quan sát cùng với biến được giới thiệu ở ví dụ 6.7. Kích thước mẫu lần lượt là $n_1 = 3, n_2 = 2$ và $n_3 = 3$. Khi đó ta thu được

$$\left(\begin{array}{ccc} \begin{bmatrix} 9 \\ 3 \end{bmatrix} & \begin{bmatrix} 6 \\ 2 \end{bmatrix} & \begin{bmatrix} 9 \\ 7 \end{bmatrix} \\ \begin{bmatrix} 0 \\ 4 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \\ \begin{bmatrix} 3 \\ 8 \end{bmatrix} & \begin{bmatrix} 1 \\ 9 \end{bmatrix} & \begin{bmatrix} 2 \\ 7 \end{bmatrix} \end{array} \right)$$

với $\bar{\mathbf{x}}_1 = \begin{bmatrix} 8 \\ 4 \end{bmatrix}, \bar{\mathbf{x}}_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \bar{\mathbf{x}}_3 = \begin{bmatrix} 2 \\ 8 \end{bmatrix}$

và $\bar{\mathbf{x}} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp nhiều chiều

Với các quan trắc của biến thứ nhất, ta thu được các kết quả sau từ phần thảo luận ANOVA một chiều

$$\begin{bmatrix} 9 & 6 & 9 \\ 0 & 2 & \\ 3 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 4 & 4 & 4 \\ 4 & 4 & \\ 4 & 4 & 4 \end{bmatrix} + \begin{bmatrix} 4 & 4 & 4 \\ -3 & -3 & \\ -2 & -2 & -2 \end{bmatrix} + \begin{bmatrix} 1 & -2 & 1 \\ -1 & 1 & \\ 1 & -1 & 0 \end{bmatrix}$$

$$(\text{observation}) = (\text{mean}) + \begin{pmatrix} \text{treatment} \\ \text{effect} \end{pmatrix} + (\text{residual})$$

và

$$SS_{obs} = SS_{mean} + SS_{tr} + SS_{res}$$

$$216 = 128 + 78 + 10$$

$$\text{Total SS (corrected)} = SS_{obs} - SS_{mean} = 216 - 128 = 88$$

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp nhiều chiều

Lặp lại quá trình trên cho các quan trắc của biến thứ hai ta được

$$\begin{bmatrix} 3 & 2 & 7 \\ 4 & 0 & \\ 8 & 9 & 7 \end{bmatrix} = \begin{bmatrix} 5 & 5 & 5 \\ 5 & 5 & \\ 5 & 5 & 5 \end{bmatrix} + \begin{bmatrix} -1 & -1 & -1 \\ -3 & -3 & \\ 3 & 3 & 3 \end{bmatrix} + \begin{bmatrix} -1 & -2 & 3 \\ 2 & -2 & \\ 0 & 1 & -1 \end{bmatrix}$$

$$(\text{observation}) = (\text{mean}) + \begin{pmatrix} \text{treatment} \\ \text{effect} \end{pmatrix} + (\text{residual})$$

và

$$SS_{obs} = SS_{mean} + SS_{tr} + SS_{res}$$

$$272 = 200 + 48 + 24$$

$$\text{Total SS (corrected)} = SS_{obs} - SS_{mean} = 272 - 200 = 72$$

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp nhiều chiều

Nhân theo từng dòng của các ma trận tương ứng của hai biến, ta thu được

$$\begin{aligned}
\text{Mean:} \quad & 4 \times 5 + 4 \times 5 + \ldots + 4 \times 5 = 160 \\
\text{Treatment:} \quad & 3 \times 4 \times (-1) + 2 \times (-3) \times (-3) + 3 \times (-2) \times 3 = -12 \\
\text{Residual:} \quad & 1 \times (-1) + (-2) \times (-2) + 1 \times 3 + \ldots + 0 \times (-1) = 1 \\
\text{Total:} \quad & 9 \times 3 + 6 \times 2 + 9 \times 7 + \ldots + 2 \times 7 = 149
\end{aligned}$$

$$\begin{aligned}
\text{Total (corrected) cross product} &= \text{total cross product} - \text{mean cross produ} \\
&= 149 - 160 = -11
\end{aligned}$$

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp nhiều chiều

Khi đó, ta có bảng MANOVA có dạng như sau

Source of variance	Matrix of sum of squares and cross products	Degrees of freedom (d.f.)
Treatment	$\begin{bmatrix} 78 & -12 \\ -12 & 48 \end{bmatrix}$	$3-1=2$
Residual	$\begin{bmatrix} 10 & 1 \\ 1 & 24 \end{bmatrix}$	$3+2+3-3=5$
Total (corrected)	$\begin{bmatrix} 88 & -11 \\ -11 & 72 \end{bmatrix}$	7

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp nhiều chiều

Phương trình (6.40) được xác định bằng

$$\begin{bmatrix} 88 & -11 \\ -11 & 72 \end{bmatrix} = \begin{bmatrix} 78 & -12 \\ -12 & 48 \end{bmatrix} + \begin{bmatrix} 10 & 1 \\ 1 & 24 \end{bmatrix}$$

Khi đó ta có

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|} = \frac{\begin{vmatrix} 10 & 1 \\ 1 & 24 \end{vmatrix}}{\begin{vmatrix} 88 & -11 \\ -11 & 72 \end{vmatrix}} \approx 0.0385$$

So sánh nhiều trung bình của tổng thể nhiều chiều

Trường hợp nhiều chiều

Ta xét kiểm định

$$H_0 : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \boldsymbol{\tau}_3 = \mathbf{0}$$

với mức ý nghĩa 1% Vì $p = 2$ và $g = 3$, theo bảng 6.3, ta sẽ chọn trường hợp 2. Ta có:

$$(\frac{1 - \sqrt{\Lambda*}}{\sqrt{\Lambda*}})\frac{(\sum n_l - g - 1)}{(g-1)} = (\frac{1 - \sqrt{0.0385}}{\sqrt{0.0385}})\frac{8-3-1}{3-2} = 8.19$$

Phân vị trên cần tìm là:

$$F_{(2(g-1),2(\sum n_l - g - 1)}(0.01) =_{4,8} (0.01) = 7.01$$

Ta có, $8.19 > 7.01$ nên ta bác bỏ H_0 với mức ý nghĩa 1%. Từ đó, các véctơ trung bình của các tổng thể không bằng nhau.

Ví dụ 6.10 (Phân tích nhiều biến cho dữ liệu viện dưỡng lão Wisconsin)

Nhóm	Số lượng quan trắc	Véctơ trung bình mẫu
l=1(private)	$n_1 = 271$	$\overline{x_1}' = \begin{bmatrix} 2.066 & .480 & .082 & .360 \end{bmatrix}$
l=2 (nonprofit)	$n_2 = 138$	$\overline{x_2}' = \begin{bmatrix} 2.167 & .596 & .124 & .418 \end{bmatrix}$
l=3(government)	$n_3 = 107$	$\overline{x_3}' = \begin{bmatrix} 2.273 & .521 & .125 & .383 \end{bmatrix}$

Ma trận hiệp phương sai mẫu là:

$$S_1 = \begin{bmatrix} .291 & & & \\ -.001 & .011 & & \\ .002 & .000 & .001 & \\ .010 & .003 & .000 & .010 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} .561 & & & \\ .011 & .025 & & \\ .001 & .004 & .005 & \\ .037 & .007 & .002 & .019 \end{bmatrix}$$

$$S_3 = \begin{bmatrix} .261 & & & \\ .030 & .017 & & \\ .003 & -.000 & .004 & \\ .018 & .006 & .001 & .013 \end{bmatrix}$$

Ta tìm ma trận W.

$$
\begin{aligned}
W &= (n_1-1)S_1 + (n_2-1)S_2 + (n_3-1)S_3 \\
&= \begin{bmatrix} 182.962 & & & \\ 4.408 & 8.200 & & \\ 1.695 & .633 & 1.484 & \\ 9.581 & 2.428 & .394 & 6.538 \end{bmatrix}
\end{aligned}
$$

Và,

$$
\overline{x} = \frac{n_1\overline{x_1} + n_2\overline{x_2} + n_3\overline{x_3}}{n_1+n_2+n_3} = \begin{bmatrix} 2.136 \\ .519 \\ .102 \\ .380 \end{bmatrix}
$$

Hơn nữa,

$$
B = \sum_{l=1}^{3} n_l(\overline{x_l}-\overline{x})(\overline{x_l}-\overline{x})' = \begin{bmatrix} 3.475 & & & \\ 1.111 & 1.225 & & \\ .821 & .453 & .235 & \\ .584 & .610 & .230 & .304 \end{bmatrix}
$$

Từ đây, ta tính được:

$$\Lambda * = \frac{|W|}{|B+W|} = .7714$$

Ta cũng có:

$$\left(\frac{\sum n_l - p - 2}{p}\right)\left(\frac{1-\sqrt{\Lambda *}}{\sqrt{\Lambda *}}\right) = \left(\frac{516-4-2}{4}\right)\left(\frac{1-\sqrt{.7714}}{\sqrt{.7714}}\right) = 17.67$$

Cho độ tin cậy $\alpha = 0.01$, ta có:

$F_{2p,2(\sum n_l - p - 2)}(\alpha) = F_{8,1020}(.01) = 2.51$. Do $17.67 > 2.51$ nên ta bác bỏ H_0 với mức ý nghĩa 1%. Nghĩa là giá tiền trung bình phụ thuộc vào loại sở hữu.

Ta cũng nhận thấy $\sum n_l = n = 516$, nghĩa là cỡ mẫu lớn nên ta có thể sử dụng bổ đề để kiểm định với mức ý nghĩa $\alpha = .01$. Ta có:

$$-(n-1-(p+g)/2)ln\left(\frac{|W|}{|B+W|}\right) = -511.5ln(.7714) = 132.76$$

Ta cũng có $\chi^2_{p(g-1)}(.01) = \chi^2_8(.01) = 20.09$. Do $132.76 > 20.09$ nên ta cũng bác bỏ H_0 với mức ý nghĩa 1%.

Đặt τ_{ki} là thành phần thứ i của τ_k.

Ta có: $\hat{\tau}_k = \bar{x}_k - \bar{x}$ là ước lượng cho τ_k.

Ta cũng có Độ chênh lệch giữa hai trung bình mẫu k và l độc lập là:

$$\hat{\tau}_{ki} - \hat{\tau}_{li} = \overline{x}_{ki} - \overline{x}_{li}$$

Ta nhận thấy,

$$Var(\hat{\tau}_{ki} - \hat{\tau}_{li}) = Var(\bar{X}_{ki} - \bar{X}_{li}) = \left(\frac{1}{n_k} + \frac{1}{n_l}\right)\sigma_{11}$$

Ta cũng có, $Var(\bar{X}_{ki} - \bar{X}_{li})$ được ước lượng bởi:

$$Var(\bar{X}\hat{}_{ki} - \bar{X}_{li}) = \left(\frac{1}{n_k} + \frac{1}{n_l}\right)\frac{\omega_{11}}{n-g}$$

trong đó, ω_{11} là phần tử thứ i trên đường chéo của ma trận W và $n = n_1 + ... + n_g$

Ta có p biến và C_2^g cặp k, l cho chênh lệch giữa chúng. Do đó, khoảng student cho 2 mẫu sẽ có giá trị tới hạn là

$$t_{n-g}\left(\frac{\alpha}{2pC_2^g}\right) = t_{n-g}\left(\frac{\alpha}{pg(g-1)}\right)$$

Kết quả 6.5

Đặt $n = \sum_{k=1}^{g} n_k$. Với độ tin cậy ít nhất $(1-\alpha)$, $\tau_{ki} - \tau_{li}$ thuộc vào khoảng:

$$\bar{x_{ki}} - \bar{x_{li}} \pm t_{n-g}\left(\frac{\alpha}{pg(g-1)}\right)\sqrt{\frac{\omega_{ii}}{n-g}\left(\frac{1}{n_k}+\frac{1}{n_l}\right)}$$

Với mọi phần tử $i = 1, 2, ..., p$ và với tất cả $l < k = 1, ..., g$. ω_{11} là phần tử thứ i trên đường chéo của ma trận W.

Ví dụ 6.11: Khoảng tin cậy đồng thời cho chênh lệch giữa các treatment của dữ liệu viện dưỡng lão

Sử dụng bộ dữ liệu từ ví dụ 6.10, ta có:

$$\hat{\tau}_1 = (\overline{x_1} - \overline{x}) = \begin{bmatrix} -.070 \\ -.039 \\ -.020 \\ -.020 \end{bmatrix} \qquad \hat{\tau}_3 = (\overline{x_3} - \overline{x}) = \begin{bmatrix} .137 \\ .002 \\ .023 \\ .003 \end{bmatrix}$$

$$W = \begin{bmatrix} 182.962 & & & \\ 4.408 & 8.200 & & \\ 1.695 & .633 & 1.484 & \\ 9.581 & 2.428 & .394 & 6.538 \end{bmatrix}$$

Từ đây, ta có khoảng tin cậy cho $\tau_{13} - \tau_{33}$ với mức ý nghĩa 5% là:

$$\begin{aligned} & \hat{\tau_{13}} - \hat{\tau_{33}} \pm t_{n-g}\left(\frac{\alpha}{pg(g-1)}\right)\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_3}\frac{\omega_{33}}{n-g}\right.} \\ = \quad & t_{513}\left(\frac{0.05}{4.3.2}\right)\sqrt{\left(\frac{1}{271} + \frac{1}{107}\right)\frac{1.484}{516-3}} \\ = \quad & -.043 \pm .018 \end{aligned}$$