THỐNG KÊ NHIỀU CHIỀU

Chương 6: So sánh các vectơ trung bình nhiều chiều

Đinh Anh Huy - 18110103

Nguyễn Đức Vũ Duy - 18110004

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 $\mathcal{N}_p(\boldsymbol{\delta}, \mathbf{\Sigma_d})$. 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.

Chứng minh

...

Trường hợp n và $n-p$ lớn, ta có

$$\overline{\mathbf{D}} = \frac{1}{n}(\mathbf{D}_1 + \mathbf{D}_2 + ... + \mathbf{D}_p) \sim \mathcal{N}_p(\boldsymbol{\delta}, \frac{1}{n}\mathbf{\Sigma}_d)$$

Suy ra

$$\sqrt{n}(\overline{\mathbf{D}} - \boldsymbol{\delta}) \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma_d})$$

Khi đó

$$n(\overline{\mathbf{D}} - \boldsymbol{\delta})^T \mathbf{S}_{\mathbf{d}}^{-1} (\overline{\mathbf{D}} - \boldsymbol{\delta}) \sim \chi^2_p$$

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

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

Hơn nữa

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

trong đó

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

Chứng minh

Cho $\mathbf{X}_{11}, \mathbf{X}_{12}, ..., \mathbf{X}_{1n_1}$ 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 $\mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. Khi đó, theo kết quả 4.8 ta có

$$\begin{aligned}\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2 &= \frac{1}{n_1}(\mathbf{X}_{11} + \mathbf{X}_{12} + ... + \mathbf{X}_{1n_1}) - \frac{1}{n_2}(\mathbf{X}_{21} + \mathbf{X}_{22} + ... + \mathbf{X}_{2n_2}) \\ &= \frac{1}{n_1}\mathbf{X}_{11} + ... + \frac{1}{n_1}\mathbf{X}_{1n_1} - \frac{1}{n_2}\mathbf{X}_{21} - ... - \frac{1}{n_2}\mathbf{X}_{2n_2}\end{aligned}$$

tuân theo phân phối

$$\mathcal{N}_p\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{\Sigma}\right)$$

Hơn nữa

$$(n_1 - 1)\mathbf{S}_1 \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n_1 - 1) \quad \text{và} \quad (n_2 - 1)\mathbf{S}_2 \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n_2 - 1)$$

Vì $\mathbf{X}_{1j}$ và $\mathbf{X}_{2j}$ với $j = 1, 2, ...$ độc lập với nhau nên $(n_1 - 1)\mathbf{S}_1$ và $(n_2 - 1)\mathbf{S}_2$ cũng độc lập với nhau. Do đó

$$(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n_1 + n_2 - 2)$$

Khi đó

$$\begin{aligned}T^2 &= \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1/2}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^T \mathbf{S}_{pooled}^{-1}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1/2}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)) \\ &= \begin{pmatrix}\text{vectơ ngẫu nhiên} \\ \text{chuẩn nhiều chiều}\end{pmatrix}^T \left(\frac{\text{Ma trận ngẫu nhiên Wishart}}{\text{hệ số tự do}}\right)^{-1}\begin{pmatrix}\text{vectơ ngẫu nhiên} \\ \text{chuẩn nhiều chiều}\end{pmatrix} \\ &= \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})^T\left[\frac{\mathcal{W}_p(\boldsymbol{\Sigma}, n_1 + n_2 - 2)}{n_1 + n_2 - 2}\right]^{-1}\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})\end{aligned}$$

Như vậy thống kê T^2 tuân theo phân phối

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

Hơn nữa

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

trong đó

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

Kết quả 6.3. Cho $c^2 = \dfrac{(n_1+n_2-2)p}{(n_1+n_2-p-1)} F_{p,n_1+n_2-p-1}(\alpha)$. 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

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

Chứng minh

Cho $\mathbf{X}_{11}, \mathbf{X}_{12}, ..., \mathbf{X}_{1n_1}$ 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 $\mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. Khi đó tổ hợp tuyến tính của các quan trắc trong hai mẫu trên là

$$\mathbf{a}^T\mathbf{X}_{1j} = a_1X_{1j1} + a_2X_{1j2} + ... + a_pX_{1jp} \quad \text{và} \quad \mathbf{a}^T\mathbf{X}_{2j} = a_1X_{2j1} + a_2X_{2j2} + ... + a_pX_{2jp}$$

có trung bình mẫu và hiệp phương sai tương ứng là $\mathbf{a}^T\overline{\mathbf{X}}_1, \mathbf{a}^T\mathbf{S}_1\mathbf{a}$ và $\mathbf{a}^T\overline{\mathbf{X}}_2, \mathbf{a}^T\mathbf{S}_2\mathbf{a}$, trong đó $\overline{\mathbf{X}}_1, \mathbf{S}_1$ và $\overline{\mathbf{X}}_2, \mathbf{S}_2$ là trung bình và hiệp phương sai của hai mẫu ban đầu. Khi hai tổng thể ban đầu có cùng ma trận hiệp phương sai $\boldsymbol{\Sigma}$ thì $s^2_{1.\mathbf{a}} = \mathbf{a}^T\mathbf{S}_1\mathbf{a}$ và $s^2_{2.\mathbf{a}} = \mathbf{a}^T\mathbf{S}_2\mathbf{a}$ đều có ước lượng là $\mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}$. Kết hợp hai ước lượng trên ta thu được

$$\begin{aligned} s^2_{\mathbf{a}.pooled} &= \frac{(n_1-1)s^2_{1.\mathbf{a}} + (n_2-1)s^2_{2.\mathbf{a}}}{n_1+n_2-2} \\ &= \mathbf{a}^T\left[\frac{n_1-1}{n_1+n_2-2}\mathbf{S}_1 + \frac{n_2-1}{n_1+n_2-2}\mathbf{S}_2\right]\mathbf{a} \\ &= \mathbf{a}^T\mathbf{S}_{pooled}\mathbf{a} \end{aligned}$$

Phát biểu giả thuyết

$$H_0 : \mathbf{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathbf{a}^T\boldsymbol{\delta}_0 \quad \text{và} \quad H_1 : \mathbf{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \neq \mathbf{a}^T\boldsymbol{\delta}_0$$

Ta xét thống kê t^2 cho hai mẫu đơn biến

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

Xét *bổ đề Maximization*: Cho $\underset{(p\times p)}{\mathbf{B}}$ là ma trận xác định dương và $\underset{(p\times 1)}{\mathbf{d}}$ là một vectơ bất kỳ. Khi đó, với một vectơ khác không tuỳ ý $\underset{(p\times 1)}{\mathbf{x}}$ thì

$$\max_{x\neq 0}\frac{(\mathbf{x}^T\mathbf{d})^2}{\mathbf{x}^T\mathbf{B}\mathbf{x}} = \mathbf{d}^T\mathbf{B}^{-1}\mathbf{d}$$

với giá trị cực đại đạt được khi $\underset{(p\times1)}{\mathbf{x}} = c\underset{(p\times p)(p\times 1)}{\mathbf{B}^{-1}\, d}$ với mọi hằng số $c \neq 0$.

Theo bổ đề trên với $\mathbf{d} = (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))$ và $\mathbf{B} = (1/n_1 + 1/n_2)\mathbf{S}_{pooled}$ ta có

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

với mọi $\mathbf{a} \neq \mathbf{0}$. Do đó

$$\begin{aligned} 1 - \alpha &= P[T^2 \leq c^2] = P[t_{\mathbf{a}}^2 \leq c^2, \forall \mathbf{a}] \\ &= P\left[\left|\mathbf{a}^T(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) - \mathbf{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right| \leq c\sqrt{\mathbf{a}^T\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{pooled}\mathbf{a}}, \forall \mathbf{a}\right] \end{aligned}$$

trong đó

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

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

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

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

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

Chứng minh

Cho $\mathbf{X}_{11}, \mathbf{X}_{12}, ..., \mathbf{X}_{1n_1}$ 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 $\mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. Khi đó ta có

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

và

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

Theo định lý giới hạn trung tâm, ta có $\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2$ sẽ xấp xỉ về phân phối $\mathcal{N}_p[\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, n_1^{-1}\boldsymbol{\Sigma}_1 + n_2^{-1}\boldsymbol{\Sigma}_2]$. Nếu $\boldsymbol{\Sigma}_1$ và $\boldsymbol{\Sigma}_2$ đều được biết trước thì bình phương khoảng cách thống kê từ $\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2$ đến $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ sẽ là

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

Theo kết quả 4.7, bình phương khoảng cách ở trên sẽ xấp xỉ về phân phối χ^2_p. Khi n_1 và n_2 đều lớn, với xác suất cao, $\mathbf{S}_1$ sẽ càng gần với $\mathbf{\Sigma}_1$ và $\mathbf{S}_2$ sẽ gần với $\mathbf{\Sigma}_2$.

...

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:

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

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.

Chứng minh Do τ_{ki} là phần tử thứ i của τ_k và τ_k được ước lượng bởi $\overline{x_k} - \overline{x}$. Nên ta có ước lượng sau:

$$\hat{\tau_{ki}} = \overline{x_{ki}} - \overline{x_i}$$

Khi đó,

$$\hat{\tau_{ki}} - \hat{\tau_{li}} = \overline{x_{ki}} - \overline{x_{li}}$$

Ta nhận thấy rằng,

$$var(\hat{\tau_{ki}} - \hat{\tau_{li}}) = var(\overline{x_{ki}} - \overline{x_{li}}) = (\frac{1}{n_k} + \frac{1}{n_l})\sigma_{11}$$

Khi đó, $var(\overline{x_{ki}} - \overline{x_{li}})$ được ước lượng bằng cách chia từng phần tử của W bởi bậc tự do của nó, nghĩa là phần tử ở đường chéo thứ i của $var(\overline{x_{ki}} - \overline{x_{li}})$ sẽ là:

$$var(\hat{\overline{x_{ki}}} - \overline{x_{li}}) = (\frac{1}{n_k} + \frac{1}{n_l})\frac{\omega_{ii}}{n-g}$$

Trong đó, ω_{11} là phần tử ở đường chéo thứ i và $n = n_1 + n_2 + ... + n_g$.
Giờ đây, ta có p biến và số cặp k, l có thể có sẽ là $C_2^g = \dfrac{g(g-1)}{2}$. Do đó, mỗi khoảng student cho 2 mẫu sẽ có giá trị tới hạn là $t_{n-g}(\dfrac{\alpha}{2.\dfrac{pg(g-1)}{2}}) = t_{n-g}(\dfrac{\alpha}{pg(g-1)})$.

Từ tất cả điều trên, ta thế vào phương pháp Bonferroni ta thu được khoảng tin cậy đồng thời với mức tin cậy $1-\alpha$ cho $\tau_{ki} - \tau_{li}$ sẽ thuộc vào:

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

Đây là điều phải chứng minh.