

一. 成对比较与重复测量设计

(一). 成对比较

① 实验设计

样本	1	2	3	...	n-1	n
X_1	☐	☐	☐		☐	☐
X_2	☐	☐	☐		☐	☐

随机分配

$$\text{let } D_j = X_{j1} - X_{j2} \quad j = 1, 2, \dots, n$$

若已知 D_j 为来自 $N(\delta, \sigma_d^2)$ 的观测值, 则

$$t = \frac{\bar{D} - \delta}{S_d / \sqrt{n}}$$

服从 $t(n-1)$, 其中 $\bar{D} = \frac{1}{n} \sum_{j=1}^n D_j$ $S_d^2 = \frac{1}{n-1} \sum_{j=1}^n (D_j - \bar{D})^2$

② 检验与估计

假设检验:

$$H_0: \delta = 0 \quad \text{v.s.} \quad H_1: \delta \neq 0$$

当 $|t| > t_{\alpha/2}(n-1)$, 拒绝 H_0 .

置信区间:

$$\left(\bar{d} \pm t_{\alpha/2}(n-1) \cdot \frac{S_d}{\sqrt{n}} \right)$$

③ 推广到 P 维

此时

$$D_j = \begin{bmatrix} D_{j1} \\ D_{j2} \\ \vdots \\ D_{jp} \end{bmatrix} = \begin{bmatrix} X_{1j1} \\ X_{1j2} \\ \vdots \\ X_{1jp} \end{bmatrix} - \begin{bmatrix} X_{2j1} \\ X_{2j2} \\ \vdots \\ X_{2jp} \end{bmatrix} \in \mathbb{R}^p$$

结论: 若 $D_1, D_2, \dots, D_n \in \mathbb{R}^p$ 来自总体 $N_p(\delta, \Sigma_d)$ 的随机样本, 则有:

$$T^2 = n(\bar{D} - \delta)^T S_d^{-1} (\bar{D} - \delta) \sim \frac{(n-1)p}{n-p} F(p, n-p)$$

$$\text{其中 } \bar{D} = \frac{1}{n} \sum_{i=1}^n D_i \quad S_d = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})(D_i - \bar{D})^T$$

若大样本, $(n-p \gg 0)$, 则无须正态近似. 此

时 T^2 服从

$$T^2 \sim \chi^2(p)$$

多元的检验与估计

给定对应于式(6-5)中随机变量的观测值之差 $d'_j = [d_{j1}, d_{j2}, \dots, d_{jp}]$, $j=1, 2, \dots, n$, 对总体 $N_p(\delta, \Sigma_d)$, 若 $T^2 = n \bar{d}' S_d^{-1} \bar{d} > \frac{(n-1)p}{(n-p)} F_{p, n-p}(\alpha)$, 则在水平 α 下, 拒绝 $H_0: \delta = 0$ 而接受 $H_1: \delta \neq 0$, 其中 $F_{p, n-p}(\alpha)$ 为自由度为 p 与 $n-p$ 的 F 分布的上 100α 百分位数, 而 $\bar{d}$ 与 S_d 由式(6-8)给出.

δ 的 $100(1-\alpha)\%$ 置信域由所有满足

$$(\bar{d} - \delta)' S_d^{-1} (\bar{d} - \delta) \leq \frac{(n-1)p}{n(n-p)} F_{p, n-p}(\alpha) \quad (6-9)$$

的 δ 组成.

此外, 个别均值差 δ_i 的 $100(1-\alpha)\%$ 联合置信区间为

$$\delta_i: \bar{d}_i \pm \sqrt{\frac{(n-1)p}{(n-p)} F_{p, n-p}(\alpha)} \sqrt{\frac{s_{d_i}^2}{n}} \quad (6-10)$$

此处 $\bar{d}_i$ 为 $\bar{d}$ 的第 i 个分量, $s_{d_i}^2$ 为 S_d 的第 i 个对角元.

当 $n-p$ 充分大时, $[(n-1)p/(n-p)] F_{p, n-p}(\alpha) \doteq \chi_p^2(\alpha)$, 因而不需要正态性假定.

单一均值差的 $100(1-\alpha)\%$ 庞弗罗尼联合置信区间为

$$\delta_i: \bar{d}_i \pm t_{n-1} \left(\frac{\alpha}{2p} \right) \sqrt{\frac{s_{d_i}^2}{n}} \quad (6-10a)$$

其中 $t_{n-1}(\alpha/2p)$ 为自由度为 $n-1$ 的 t 分布的上 $100(\alpha/2p)$ 百分位数.

(二). 比较各种处理的重复测量设计

针对一个响应变量, q 种处理。对同一样本进行 q 种处理, 得到 q 种结果, 构成这个观测值 $X_j \in \mathbb{R}^q$ 。

目标: 比较 q 种处理的区别。

第 j 个观测值 $X_j = [X_{j1}, X_{j2}, \dots, X_{jq}]^T \in \mathbb{R}^q, j = 1, 2, 3, \dots, n$

考虑 $\mu = E(X_j)$ 的各分量的对比, 即:

$$\begin{bmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \vdots \\ \mu_1 - \mu_q \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{bmatrix} \triangleq C_1 \mu$$

$$\begin{bmatrix} \mu_2 - \mu_1 \\ \mu_3 - \mu_2 \\ \vdots \\ \mu_q - \mu_{q-1} \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{bmatrix} \triangleq C_2 \mu$$

$C_1, C_2 \in \mathbb{R}^{(q-1) \times q}$, 且每行线性独立, 称 C_1, C_2 为对比矩阵。

于是, 比较区别, 假设检验为:

$$H_0: C\mu = 0 \quad \text{v.s.} \quad H_1: C\mu \neq 0$$

构造 T 检验统计量:

$$T^2 = n(C\bar{X} - 0)^T (CSC^T)^{-1} (C\bar{X} - 0)$$

结论:

重复测量设计中各种处理的等效性检验

考虑总体 $N_q(\mu, \Sigma)$. 设 C 为对比矩阵, 若

$$T^2 = n(C\bar{x})'(CSC')^{-1}C\bar{x} > \frac{(n-1)(q-1)}{(n-q+1)} F_{q-1, n-q+1}(\alpha) \quad (6-16)$$

则在显著性水平 α 下拒绝 $H_0: C\mu = 0$ (即处理均值相等) 而接受 $H_1: C\mu \neq 0$.

此处 $F_{q-1, n-q+1}(\alpha)$ 为自由度为 $q-1$ 和 $n-q+1$ 的 F 分布的上 100α 百分位数, 且样本均值向量及样本协方差矩阵分别为

$$\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j \quad \text{和} \quad S = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})(x_j - \bar{x})'$$

结论: T^2 不依赖于 C 的选择

Proof: 对 $\forall$ 对比矩阵 C_1, C_2 有:

$$C_1 = BC_2 \quad B \text{ 非奇异} \quad (\text{证明})$$

$$\begin{aligned} \text{于是 } (C_1 \bar{X})^T (C_1 S C_1^T)^{-1} (C_1 \bar{X}) &= (BC_2 \bar{X})^T (BC_2 S C_2^T B^T)^{-1} (BC_2 \bar{X}) \\ &= (C_2 \bar{X})^T \cdot B^T \cdot (B^T)^{-1} (C_2 S C_2^T)^{-1} B^{-1} \cdot B (C_2 \bar{X}) \\ &= (C_2 \bar{X})^T (C_2 S C_2^T)^{-1} (C_2 \bar{X}) \end{aligned}$$

故 C_1, C_2 等价, T^2 不依赖 C 的选择 $\square$

置信区域: 对对比矩阵 C 的 q -行向量 (对比向量)

$C \in \mathbb{R}^q$, $C\mu$ 的 $100(1-\alpha)\%$ 置信区间由:

$$\underline{n(C\bar{x} - C\mu)^T (CSC^T)^{-1} (C\bar{x} - C\mu)} \quad \begin{matrix} CSC^T \in \mathbb{R}^{(q-1) \times (q-1)} \\ df = n \end{matrix}$$

$$\leq \frac{(n-1)(q-1)}{n-q+1} \cdot F_{\alpha}(q-1, n-q+1)$$

的所有 $C\mu$ 构成。则 $C^T\mu$ 的 $1-\alpha$ CI 为

$$C^T\mu: \left(C^T \bar{x} \pm \sqrt{\frac{(n-1)(q-1)}{n-q+1} F_{\alpha}(q-1, n-q+1)} \sqrt{\frac{C^T S C}{n}} \right)$$

可见习题: 教材 P₂₁₆ 例 6.2

二、两总体均值向量的比较

考虑 2 个随机样本，分别来自总体 1，样本容量为 n_1 ，
总体 2 样本容量为 n_2

具体观测值

样本	汇总统计量	
(总体 1) $x_{11}, x_{12}, \dots, x_{1n_1}$	$\bar{x}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} x_{1j}$	$S_1 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)(x_{1j} - \bar{x}_1)'$
(总体 2) $x_{21}, x_{22}, \dots, x_{2n_2}$	$\bar{x}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} x_{2j}$	$S_2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)(x_{2j} - \bar{x}_2)'$

表中符号的第一个角标(1 或 2)代表总体.

(一). 关于数据结构的假定

- (1) $X_{j1}, X_{j2}, \dots, X_{jn_j}$ 为来自均值向量 μ_j ，协方差阵 Σ_j 的 p 维总体的随机样本， $X_{ji} \in \mathbb{R}^p$ $j=1,2$
- (2) 样本 X_{1i} 与 X_{2i} 均相互独立

(二). n_1 与 n_2 均较小时的进一步假定

- (1) 两总体均为多元正态总体
- (2) $\Sigma_1 = \Sigma_2$

(三). 检验，当 $\Sigma_1 = \Sigma_2$ 时

$$H_0: \mu_1 - \mu_2 = \bar{\delta}_0 \quad \text{v.s.} \quad H_1: \mu_1 - \mu_2 \neq \bar{\delta}_0$$

$$\text{由 } E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

$$\text{由 } X_{1i}, X_{2i} \text{ 相互独立, } \text{Cov}(\bar{X}_1, \bar{X}_2) = 0$$

$$\text{故 } \text{Cov}(\bar{X}_1 - \bar{X}_2) = \text{Cov}(\bar{X}_1) + \text{Cov}(\bar{X}_2) = \frac{1}{n_1} \Sigma_1 + \frac{1}{n_2} \Sigma_2$$

又假设 $\Sigma_1 = \Sigma_2 \triangleq \Sigma$

故 $\text{Cov}(\bar{X}_1 - \bar{X}_2) = (\frac{1}{n_1} + \frac{1}{n_2}) \Sigma$

于是 $\mu_1 - \mu_2$ 的估计为 $\bar{X}_1 - \bar{X}_2$

$\text{Cov}(\bar{X}_1 - \bar{X}_2)$ 的估计为 $(\frac{1}{n_1} + \frac{1}{n_2}) S_p$

其中 $S_p = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2}$ $S_i = \frac{1}{n_i-1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T$

[定理] 考虑 2 个具有相同协方差阵的正态总体, 设 $L(\mu_1, \mu_2, \Sigma)$ 为来自 2 个正态总体 $N_p(\mu_1, \Sigma)$, $N_p(\mu_2, \Sigma)$ 的样本容量分别为 n_1, n_2 的独立样本的似然函数:

$L(\mu_1, \mu_2, \Sigma)$ 的最大值, 当:

$$\mu_1^* = \bar{X}_1 \quad \mu_2^* = \bar{X}_2 \quad \Sigma^* = \frac{1}{n_1+n_2} ((n_1-1)S_1 + (n_2-1)S_2) = (\frac{n_1+n_2-2}{n_1+n_2}) S_p$$

结论: $\frac{df-p+1}{p} \cdot \frac{1}{df} \cdot N_p(0, \Sigma) \cdot \left(\frac{Wdf(\Sigma)}{df} \right)^{-1} N_p(0, \Sigma) \sim F(p, df-p+1)$

结论 6.2 设 $X_{11}, X_{12}, \dots, X_{1n_1}$ 为来自总体 $N_p(\mu_1, \Sigma)$ 的一个容量为 n_1 的随机样本, 而 $X_{21}, X_{22}, \dots, X_{2n_2}$ 则为来自总体 $N_p(\mu_2, \Sigma)$ 的一个容量为 n_2 的独立随机样本, 此时

$$T^2 = [\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_p \right]^{-1} [\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)]$$

的分布为

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

于是有

$$P \left[(\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2))' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_p \right]^{-1} (\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)) \leq c^2 \right] = 1 - \alpha$$

(6-24)

其中,

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

Proof: $\bar{X}_1 - \bar{X}_2 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i} - \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i}$ 服从

$$N_p(\mu_1 - \mu_2, (\frac{1}{n_1} + \frac{1}{n_2}) \Sigma)$$

而 $(n_1-1)S_1$ 与 $(n_2-1)S_2$ 分别服从

$$W_{n_1-1}(\Sigma) \quad W_{n_2-1}(\Sigma)$$

由 X_{1i}, X_{2i} 独立 $\Rightarrow S_1 \perp S_2$ 故

$$(n_1-1)S_1 + (n_2-1)S_2 \sim W_{n_1+n_2-2}(\Sigma)$$

$$\text{因此: } T^2 = (\frac{1}{n_1} + \frac{1}{n_2})^{-\frac{1}{2}} \cdot (\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2))^T \cdot S_p^{-1} \cdot (\frac{1}{n_1} + \frac{1}{n_2})^{-\frac{1}{2}} \cdot (\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2))$$

$$= (\text{多元正态随机向量})^T \cdot \left(\frac{\text{Wishart 阵}}{\text{自由度}} \right)^{-1} (\text{多元正态随机向量})$$

$$= N_p(0, \Sigma)^T \cdot \left(\frac{W_{n_1+n_2-2}(\Sigma)}{n_1+n_2-2} \right)^{-1} \cdot N_p(0, \Sigma)$$

是符合 Hotelling T^2 的定义: 服从 F 分布

p 维总体均值差的 $100(1-\alpha)\%$ 庞弗罗尼联合置信区间为

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

其中的 $t_{n_1+n_2-2}(\alpha/2p)$ 为自由度为 n_1+n_2-2 的 t 分布的上 $100(\alpha/2p)$ 百分位数。

(四) 主轴法

$$\text{方向: } S_p e_i = \lambda_i e_i$$

$$\text{主轴长: } \sqrt{\lambda_i} \cdot \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right) c^2}$$

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

$$\bar{X}_1 - \bar{X}_2 \sim N_p(\mu_1 - \mu_2, (\frac{1}{n_1} + \frac{1}{n_2}) \Sigma)$$

$$(n_1+n_2-2)S_p \sim W_{n_1+n_2-2}(\Sigma)$$

(五) 联合置信区间

结论 6.3 设 $c^2 = [(n_1+n_2-2)p/(n_1+n_2-p-1)]F_{p, n_1+n_2-p-1}(\alpha)$. 此时对任意的 a ,

$$a'(\bar{X}_1 - \bar{X}_2) \pm c \sqrt{a' \left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_p a}$$

将以概率 $1-\alpha$ 包含 $a'(\mu_1 - \mu_2)$. 特别是

$$(\bar{X}_{1i} - \bar{X}_{2i}) \pm c \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right) s_{ii, p}}, i = 1, 2, \dots, p$$

将包含 $\mu_{1i} - \mu_{2i}$.

(六) $\Sigma_1 \neq \Sigma_2$ 但大样本

结论 6.4 设样本容量使 $n_1 - p$ 与 $n_2 - p$ 都很大, 此时 $\mu_1 - \mu_2$ 的近似 $100(1-\alpha)\%$ 置信椭圆为满足

$$[\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)]' \left[\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right]^{-1} [\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)] \leq \chi_p^2(\alpha)$$

的所有 $\mu_1 - \mu_2$ 给出, 其中 $\chi_p^2(\alpha)$ 为自由度为 p 的 χ^2 分布的上 100α 百分位数. 而且, 所有线性组合 $a'(\mu_1 - \mu_2)$ 的 $100(1-\alpha)\%$ 联合置信区间为

$$a'(\bar{x}_1 - \bar{x}_2) \pm \sqrt{\chi_p^2(\alpha)} \sqrt{a' \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right) a}$$

特别地, 当 $p=1$ 时, 有

推论:

当 $n_1 = n_2 = n$ 时, $(n-1)/(n+n-2) = 1/2$ 从而

$$\begin{aligned} \frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 &= \frac{1}{n} (S_1 + S_2) \\ &= \frac{(n-1)S_1 + (n-1)S_2}{n+n-2} \left(\frac{1}{n} + \frac{1}{n} \right) \\ &= S_p \left(\frac{1}{n} + \frac{1}{n} \right) \end{aligned}$$

故有结论: 当样本容量相近时, 大样本方法与 $\Sigma_1 = \Sigma_2$ 时情况方法等价。

$n_1 = n_2$ 方差不等效应最小

导致拒绝 H_0 的最关键的线性组合:

$$\begin{aligned} \hat{a} &\propto \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right)^{-1} (\bar{x}_1 - \bar{x}_2) \\ &\quad \text{(或 } S_p^{-1} (\bar{x}_1 - \bar{x}_2) \text{)} \end{aligned}$$

(七) $\Sigma_1 \neq \Sigma_2$ 但小样本

当样本容量不大时对正态总体的 T^2 分布的一个近似

即便两个样本的容量都不大,当总体的协方差矩阵不相等时,我们仍然可以检验 $H_0: \mu_1 - \mu_2 = 0$, 只要这两个总体均是多维正态的. 这种问题经常被称为多变量 Behrens-Fisher 问题. 结果要求两个样本规模 n_1 和 n_2 都要比变量数目 p 大. 这个方法依赖于下述统计量的渐进分布.

$$T^2 = (\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2))' \left[\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right]^{-1} (\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)) \quad (6-27)$$

这个式子和结论 6.4 中的大样本统计是一样的. 然而, 对于较小的样本 (见文献 [15] 和 [19]), 建议采用式 (6-28) 给出的近似逼近, 而不是用 χ^2 近似来求得用来检验 H_0 的临界值.

$$T^2 = \frac{\nu p}{\nu - p + 1} F_{p, \nu - p + 1} \quad (6-28)$$

在式 (6-28) 中, 自由度 ν 是从样本协方差矩阵用式 (6-29) 估计得到的.

$$\nu = \frac{p + p^2}{\sum_{i=1}^2 \frac{1}{n_i} \left\{ \text{tr} \left[\left(\frac{1}{n_i} S_i \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right)^{-1} \right)^2 \right] + \left(\text{tr} \left[\frac{1}{n_i} S_i \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right)^{-1} \right] \right)^2 \right\}} \quad (6-29)$$

这里 $\min(n_1, n_2) \leq \nu \leq n_1 + n_2$. 在单变量情形 ($p=1$), 这个近似就是 Behrens-Fisher 问题通常的 Welch 解.

对中等规模的样本容量和两正态总体, 如果

$$(\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2))' \left[\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right]^{-1} (\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)) > \frac{\nu p}{\nu - p + 1} F_{p, \nu - p + 1}(\alpha)$$

则在近似水平 α 下, 拒绝均值相等假设 $H_0: \mu_1 - \mu_2 = 0$. 上式中, 自由度 ν 由式 (6-29) 给出. 除了临界值 $\chi_p^2(\alpha)$ 被较大的常数 $\frac{\nu p}{\nu - p + 1} F_{p, \nu - p + 1}(\alpha)$ 取代外, 这个过程和结论 6.4 中的大样本检验过程是一致的.

类似的, 近似的 $100(1-\alpha)\%$ 置信区域由所有的满足下式的 $\mu_1 - \mu_2$ 给出.

$$(\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2))' \left[\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right]^{-1} (\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)) \leq \frac{\nu p}{\nu - p + 1} F_{p, \nu - p + 1}(\alpha) \quad (6-30)$$

对于正态的总体, 由式 (6-28) 和式 (6-29) 给出的 T^2 分布近似往往得出合理的结果.

三、多个多元总体均值向量的比较

(单因子多元方差分析)

见书 P226

从 g 个总体分别抽取的随机样本为：

总体 1: $X_{11} \quad X_{12} \quad \dots \quad X_{1n_1}$

总体 2: $X_{21} \quad X_{22} \quad \dots \quad X_{2n_2}$

总体 g : $X_{g1} \quad X_{g2} \quad \dots \quad X_{gn_g}$

目标：多元方差分析首先被用来研究这些总体的均值向量是否相等。若不等，找出哪些分量有显著差异。

(一) 单因子多元方差分析的假设

Assumption 1: 这 g 个总体的均值为 $\mu_l \in \mathbb{R}^p$ ($l=1, 2, \dots, g$)；所有样本均相互独立

Assumption 2: 这 g 个总体均有公共的协方差 Σ

Assumption 3: 这 g 个总体均多元正态总体

(=). 一元方差分析 $\Rightarrow \mu_l \in \mathbb{R} \quad \Sigma = \sigma^2 \in \mathbb{R}$

对 g 个总体的样本

$$X_{l1}, X_{l2}, \dots, X_{ln_l} \sim N(\mu_l, \sigma^2)$$

$l=1, 2, \dots, g \quad X_{li} \in \mathbb{R} \quad \text{一元正态 } N(\mu_l, \sigma^2) \quad \text{同方差 } \sigma^2$

假设检验：

$$H_0: \mu_1 = \mu_2 = \dots = \mu_g \quad \text{v.s.} \quad H_1: \exists k \neq l \quad \mu_k \neq \mu_l$$

重新参数化:

将每个均值 μ_l 改为:

$$\mu_l = \mu + \tau_l$$

其 μ_l 为第 l 总体均值, μ 称为总均值, τ_l 为第 l 个总体(处理)的效应。

新假设检验

$$H_0: \tau_1 = \tau_2 = \dots = \tau_g = 0 \quad \text{v.s.} \quad H_1: \exists l \quad \tau_l \neq 0$$

模型

重新参数化为: $X_{lj} \sim N(\mu + \tau_l, \sigma^2)$ 可表示为:

$$X_{lj} = \mu + \tau_l + e_{lj}$$

其中 e_{lj} 为 i.i.d. $N(0, \sigma^2)$ 随机误差

为了唯一地确定模型参数及它们的最小二乘估计, 通常要附加约束:

$$\sum_{l=1}^g n_l \tau_l = 0$$

观测

在观测值上也做变化:

$$x_{ej} - \bar{x} = (x_{ej} - \bar{x}_e) + (\bar{x}_e - \bar{x})$$

(上式为恒等式) 其中 x_{ej} 为观测值, $\bar{x}$ 为样本总均值, $(\bar{x}_e - \bar{x})$ 为处理效应 τ_e 的估计, $(x_{ej} - \bar{x}_e)$ 为残差. 即:

$\bar{x}$ 为 μ 的估计

$\hat{\tau}_e = (\bar{x}_e - \bar{x})$ 为 τ_e 的估计

$(x_{ej} - \bar{x}_e)$ 为 e_{ej} 的估计

方差分析 ANOVA (-元)

平方求和:

$$\sum_{e=1}^g \sum_{j=1}^{n_e} (x_{ej} - \bar{x})^2 = \sum_{e=1}^g \sum_{j=1}^{n_e} (\bar{x}_e - \bar{x})^2 + \sum_{e=1}^g \sum_{j=1}^{n_e} (x_{ej} - \bar{x}_e)^2$$

$$\text{因为 } \sum_{e=1}^g \sum_{j=1}^{n_e} (\bar{x}_e - \bar{x})(x_{ej} - \bar{x}_e) = \sum_{e=1}^g (\bar{x}_e - \bar{x}) \sum_{j=1}^{n_e} (x_{ej} - \bar{x}_e) = 0$$

最后化简:

$$\sum_{e=1}^g \sum_{j=1}^{n_e} (x_{ej} - \bar{x})^2 = \sum_{e=1}^g n_e (\bar{x}_e - \bar{x})^2 + \sum_{e=1}^g \sum_{j=1}^{n_e} (x_{ej} - \bar{x}_e)^2$$

$$SS_{\text{总}} = SS_{\text{处理}} + SS_{\text{残差}}$$

$$\text{自由度} \quad \sum_{e=1}^g n_e - 1 \quad g - 1 \quad \sum_{e=1}^g n_e - g$$

$$SS_{\text{cor}} = SS_{\text{tr}} + SS_{\text{res}}$$

结论:

$$F = \frac{SS_{\text{tr}} / (g - 1)}{SS_{\text{res}} / (\sum_{e=1}^g n_e - g)} \sim F(g - 1, \sum_{e=1}^g n_e - g)$$

当 $F > F_{\alpha}(g-1, \sum n_e - g)$ 时, 拒绝 H_0 
 $F \text{ 大} \Rightarrow S_{\text{Structure}} \text{ 大} \Rightarrow \text{组间差距大} \Rightarrow \text{显著差异} | \tau_e \neq 0$

(二). 多元方差分析 MANOVA

模型

多元情形: 即对 g 个总体的随机样本 $X_{ej} \in \mathbb{R}^p$
 此时分解为:

$$X_{ej} = \mu + \tau_e + e_{ej}$$

其中 $e_{ej} \text{ iid} \sim N_p(0, \Sigma)$, 仍有约束 $\sum_{e=1}^g n_e \tau_e = 0$

观测

类似地分解观测值 $x_{ej} \in \mathbb{R}^p$

$$x_{ej} - \bar{x} = (\bar{x}_e - \bar{x}) + (\bar{x}_{ej} - \bar{x}_e)$$

方差分析

平方和表示为:

$$\begin{aligned} \sum_{e=1}^g \sum_{j=1}^{n_e} (x_{ej} - \bar{x})(x_{ej} - \bar{x})^T &= \sum_{e=1}^g n_e (\bar{x}_e - \bar{x})(\bar{x}_e - \bar{x})^T \\ &\quad + \underbrace{\sum_{e=1}^g \sum_{j=1}^{n_e} (x_{ej} - \bar{x}_e)(x_{ej} - \bar{x}_e)^T}_{\downarrow (n_e-1) S_e} \\ &= \underbrace{B}_{\text{组间}} + \underbrace{W}_{\substack{\text{组内} \\ \downarrow \\ W = \sum_{e=1}^g (n_e-1) S_e}} \end{aligned}$$

补充* (了解即可)

$$W = \sum_{e=1}^g (n_e-1) S_e$$

观测值向量可按模型所暗示的方式分解如下：

$$\begin{array}{c} \mathbf{x}_{lj} \\ \text{(观测值)} \end{array} = \begin{array}{c} \bar{\mathbf{x}} \\ \left[\begin{array}{c} \text{样本总} \\ \text{均值} \hat{\boldsymbol{\mu}} \end{array} \right] \end{array} + \begin{array}{c} (\bar{\mathbf{x}}_l - \bar{\mathbf{x}}) \\ \left[\begin{array}{c} \text{处理效应} \\ \text{的估计量} \hat{\boldsymbol{\tau}}_l \end{array} \right] \end{array} + \begin{array}{c} (\mathbf{x}_{lj} - \bar{\mathbf{x}}_l) \\ \left[\begin{array}{c} \text{残差} \\ \hat{\mathbf{e}}_{lj} \end{array} \right] \end{array} \quad (6-39)$$

分解式(6-39)导致与一元情形相类似的多元平方和分解式. 首先, 注意到乘积 $(\mathbf{x}_{lj} - \bar{\mathbf{x}})(\mathbf{x}_{lj} - \bar{\mathbf{x}})'$ 可写成

$$\begin{aligned} (\mathbf{x}_{lj} - \bar{\mathbf{x}})(\mathbf{x}_{lj} - \bar{\mathbf{x}})' &= [(\mathbf{x}_{lj} - \bar{\mathbf{x}}_l) + (\bar{\mathbf{x}}_l - \bar{\mathbf{x}})][(\mathbf{x}_{lj} - \bar{\mathbf{x}}_l) + (\bar{\mathbf{x}}_l - \bar{\mathbf{x}})]' \\ &= (\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)(\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)' + (\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})' \\ &\quad + (\bar{\mathbf{x}}_l - \bar{\mathbf{x}})(\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)' + (\bar{\mathbf{x}}_l - \bar{\mathbf{x}})(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})' \end{aligned}$$

因为 $\sum_{j=1}^{n_l} (\mathbf{x}_{lj} - \bar{\mathbf{x}}_l) = \mathbf{0}$, 所以对 j 求和后, 中间两项均为零矩阵. 于是将交叉乘积对 l 和 j 求和后可得:

$$\begin{array}{c} \star \\ \sum_{l=1}^g \sum_{j=1}^{n_l} (\mathbf{x}_{lj} - \bar{\mathbf{x}})(\mathbf{x}_{lj} - \bar{\mathbf{x}})' = \sum_{l=1}^g n_l (\bar{\mathbf{x}}_l - \bar{\mathbf{x}})(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})' + \sum_{l=1}^g \sum_{j=1}^{n_l} (\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)(\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)' \end{array}$$

$$\begin{array}{cc} \text{(总平方和及交叉乘积和)} & \left[\begin{array}{c} \text{处理平方和及} \\ \text{交叉乘积和(或} \\ \text{组间平方和)} \end{array} \right] & \left[\begin{array}{c} \text{残差平方和及} \\ \text{交叉乘积和(或} \\ \text{组内平方和)} \end{array} \right] \end{array} \quad (6-40)$$

组内平方和可表示为

$$\begin{aligned} \rightarrow W &= \sum_{l=1}^g \sum_{j=1}^{n_l} (\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)(\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)' \\ &= (n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + \cdots + (n_g - 1)\mathbf{S}_g \end{aligned} \quad (6-41)$$

其中 $\mathbf{S}_l$ 为第 l 个样本的样本协方差矩阵. 它是二样本时 $(n_1 + n_2 - 2)\mathbf{S}_p$ 的推广, 在处理效应的存在性检验中, 它起着决定性作用.

同一元情形相似, 可通过比较处理与残差的平方和及交叉乘积和的相对大小, 来检验关于无处理效应的假设

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

等价的做法是, 也可以考虑残差与总和(修正后的)平方和及交叉乘积和的相对大小. 我们将得出检验统计量的计算用 MANOVA 表列出.

总体均值向量比较的 MANOVA 表

变差来源	平方和及交叉乘积和矩阵(SSP)	自由度(d. f.)
处理	$\mathbf{B} = \sum_{l=1}^g n_l (\bar{\mathbf{x}}_l - \bar{\mathbf{x}})(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})'$	$g - 1$
$\star$ 残差(误差)	$\mathbf{W} = \sum_{l=1}^g \sum_{j=1}^{n_l} (\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)(\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)'$	$\sum_{l=1}^g n_l - g$
总和(修正后的)	$\mathbf{B} + \mathbf{W} = \sum_{l=1}^g \sum_{j=1}^{n_l} (\mathbf{x}_{lj} - \bar{\mathbf{x}})(\mathbf{x}_{lj} - \bar{\mathbf{x}})'$	$\sum_{l=1}^g n_l - 1$

此表与一元情形下的 ANOVA 表形式完全相同,只是 ANOVA 表中标量平方和在此处已被相应的向量分量的平方和所代替.例如 $(\bar{x}_l - \bar{x})^2$ 被换成了 $(\bar{x}_l - \bar{x})(\bar{x}_l - \bar{x})'$.各种自由度与一元几何相对应,因而也与某种包含威沙特密度在内的多元分布原理相对应(见文献[1]).

假设 $H_0: \tau_1 = \tau_2 = \cdots = \tau_k = \mathbf{0}$ 的一个检验涉及广义方差,若广义方差比



$$\Lambda^* = \frac{|W|}{|B+W|} = \frac{\left| \sum_{l=1}^k \sum_{j=1}^{n_l} (x_{lj} - \bar{x}_l)(x_{lj} - \bar{x}_l)' \right|}{\left| \sum_{l=1}^k \sum_{j=1}^{n_l} (x_{lj} - \bar{x})(x_{lj} - \bar{x})' \right|}$$

(6-42)

Wilks

过小,则拒绝 H_0 . $\Lambda^* = |W|/|B+W|$ 这个量最早是由威尔克斯提出的(见文献[25]),它与一元情形下假设“ H_0 : 无处理效应”的 F 检验的等价形式(6-33)相对应. 威尔克斯^① Λ 统计量的优点是使用方便,而且与似然比准则密切相关.对于表 6.3 所列的一些特殊情形,可导出 Λ^* 的精确分布.对于其他情形及大样本情形,经巴特利特(见文献[4])修正的 Λ^* 可用于检验 H_0 .

表 6.3 威尔克斯 $\Lambda^* = |W|/|B+W|$ 的分布

变量数	组数	多元正态数据的抽样分布
$p=1$	$g \geq 2$	$\left(\frac{\sum n_l - g}{g-1}\right) \left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{g-1, \sum n_l - g}$
$p=2$	$g \geq 2$	$\left(\frac{\sum n_l - g - 1}{g-1}\right) \left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2(g-1), 2(\sum n_l - g - 1)}$
$p \geq 1$	$g=2$	$\left(\frac{\sum n_l - p - 1}{p}\right) \left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{p, \sum n_l - p - 1}$
$p \geq 1$	$g=3$	$\left(\frac{\sum n_l - p - 2}{p}\right) \left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2p, 2(\sum n_l - p - 2)}$

巴特利特(见文献[4])证明了,若假设 H_0 为真且 $\sum n_l = n$ 充分大,则

$$-\left(n-1-\frac{(p+g)}{2}\right) \ln \Lambda^* = -\left(n-1-\frac{(p+g)}{2}\right) \ln \left(\frac{|W|}{|B+W|}\right)$$
(6-43)

近似服从自由度为 $p(g-1)$ 的 χ^2 分布,因此,当 $\sum n_l = n$ 充分大时,若

$$-\left(n-1-\frac{(p+g)}{2}\right) \ln \left(\frac{|W|}{|B+W|}\right) > \chi^2_{p(g-1)}(\alpha)$$
(6-44)

我们就以显著性水平 α 拒绝 H_0 ,其中的 $\chi^2_{p(g-1)}(\alpha)$ 为自由度为 $p(g-1)$ 的 χ^2 分布的上 100α 百分位数.

四、效应 τ_k 的置信区间 $\Rightarrow$ 判断两总体间边际比较

设 $\tau_{ki} \in \mathbb{R}$ 为 $\tau_k \in \mathbb{R}^p$ 的第 i 个分量。由于

$$\hat{\tau}_k = \bar{x}_k - \bar{x} \in \mathbb{R}^p$$

为 τ_k 的估计量，故有：

$$\hat{\tau}_{ki} = \bar{x}_{ki} - \bar{x}_i \in \mathbb{R}$$

为 τ_{ki} 的估计。

于是有：

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

求方差：

$$\bar{x}_{ki} \sim N(\mu_k, \frac{1}{n_k} \sigma_{ii})$$

$$\text{Var}(\hat{\tau}_{ki} - \hat{\tau}_{li}) = \text{Var}(\bar{x}_{ki} - \bar{x}_{li}) = (\frac{1}{n_k} + \frac{1}{n_l}) \sigma_{ii}$$

而 σ_{ii} 为 $\Sigma \in \mathbb{R}^{p \times p}$ 的 (i, i) 元素，未知，故估计：

$$\hat{\text{Var}}(\hat{\tau}_{ki} - \hat{\tau}_{li}) = (\frac{1}{n_k} + \frac{1}{n_l}) \frac{w_{ii}}{n-g} \quad (n-g) \text{ 为 } W \text{ 自由度}$$

其中 w_{ii} 为 $W \in \mathbb{R}^{p \times p}$ 的 (i, i) 元素， $n = \sum_{l=1}^g n_l$

$$W = \sum (n_l - 1) S_l$$

求置信区间

首先有 p 个分量，任取 2 个 τ_{li} 有 $\binom{g}{2}$ 个，由

Bonferroni CI : $t_{\alpha/2m}(n-g)$

$$m = p \cdot \binom{g}{2} = pg(g-1)/2$$

结论:

结论 6.5 设 $n = \sum_{k=1}^g n_k$. 对于模型(6-38), 对所有的分量 $i=1, \dots, p$ 及所有差值 $l < k = 1, \dots, g$, 以至少为 $1-\alpha$ 的置信度有

$$\tau_{ki} - \tau_{li} \text{ 属于 } \bar{x}_{ki} - \bar{x}_{li} \pm t_{n-g} \left(\frac{\alpha}{pg(g-1)} \right) \sqrt{\frac{w_{ii}}{n-g} \left(\frac{1}{n_k} + \frac{1}{n_l} \right)}$$

其中的 w_{ii} 为 W 的第 i 个对角元素.

五、协方差阵 Σ 相同性的检验

对 g 个总体, 检验:

$$H_0: \Sigma_1 = \Sigma_2 = \dots = \Sigma_g = \Sigma$$

(1) 经验法则

对 $H_0: \Sigma_1 = \Sigma_2$

考虑比较: $\sigma_{1,ii} > 4\sigma_{2,ii}$ 或 $\sigma_{2,ii} > 4\sigma_{1,ii}$

则拒绝 H_0

(2) Box M test *

对 g 个总体的样本: 假设

$$X_{j1}, X_{j2}, \dots, X_{jn_j} \text{ iid } \sim N_p(\mu_j, \Sigma_j)$$

$j=1, 2, \dots, g$

$$H_0: \Sigma_1 = \Sigma_2 = \dots = \Sigma_g = \Sigma$$

构造似然比：

$$\Lambda^* = \frac{\prod_{j=1}^g |S_j|^{(n_j-1)/2}}{|S_{\text{pooled}}|^{(n-g)/2}}$$

$$\text{其中 } S_{\text{pooled}} = \frac{1}{\sum_e (n_e - 1)} \left(\sum_e (n_e - 1) S_e \right)$$

记

$$M = \left[\sum_e (n_e - 1) \right] \cdot \ln |S_{\text{pooled}}| - \sum_e (n_e - 1) \cdot \ln |S_e|$$

当 H_0 成立：

$$(1-c) M \stackrel{a}{\sim} \chi^2(V)$$

$$V = \frac{1}{2} p(p+1)(g-1) \quad c = \left(\sum_e \frac{1}{n_e - 1} - \frac{1}{n-g} \right) \frac{2p^2 + 3p - 1}{6(p+1)(g-1)}$$

当 $(1-c) M > \chi^2_{\alpha}(V)$ 拒绝 H_0

$C_1, C_2 \in \mathbb{R}^{(q-1) \times q}$ 且 $C_i \cdot \mathbf{1} = \mathbf{0}$ 其中 $\mathbf{1} = (1, \dots, 1)^T \in \mathbb{R}^q$

且 $\text{rank}(C_i) = q-1$ 。证明：

$\exists$ 非奇异 $B \in \mathbb{R}^{(q-1) \times (q-1)}$ s.t. $C_1 = BC_2$

Proof: 由 $C_i \cdot \mathbf{1} = \mathbf{0}$ $\text{rank}(C_i) = q-1$ 故 C_i 规范化

$$C_i = \begin{bmatrix} 1 & c_{i2} & \cdots & 0 \\ & 1 & c_{i3} & \vdots & 0 \\ & & \ddots & \vdots & \vdots \\ & & & 1 & 0 \end{bmatrix}$$

故 C_1, C_2 可以通过列变换转换, 即

$$\exists B \text{ s.t. } C_1 = BC_2$$

$$\text{而 } q-1 = \text{rank}(C_1) = \text{rank}(BC_2)$$

$$\leq \min \{ \text{rank}(B), \text{rank}(C_2) \}$$

$$= \min \{ \text{rank}(B), q-1 \}$$

$$\Rightarrow q-1 \geq \text{rank}(B) \geq q-1$$

$$\Rightarrow \text{rank}(B) = q-1 \Rightarrow \text{非奇异} \quad \square$$