http://staff.ustc.edu.cn/~ynyang/vector

第九讲 多元方差分析MANOVA

2024.4.1

均值、回归系数的检验
都只与协方差矩阵有关

多正态总体的均值相同性检验：MANOVA

多正态问题

模型：g个方差相同的正态总体：

$\mathbf{x}_{11},\ldots,\mathbf{x}_{1n_1}$ $iid \sim N_p(\boldsymbol{\mu}_1,\Sigma)$，样本均值和方差：$\bar{\mathbf{x}}_1, S_1$

…

$\mathbf{x}_{g1},\ldots,\mathbf{x}_{gn_g}$ $iid \sim N_p(\boldsymbol{\mu}_g,\Sigma)$，样本均值和方差：$\bar{\mathbf{x}}_g, S_g$

$H_0:\boldsymbol{\mu}_1=\boldsymbol{\mu}_2=\ldots=\boldsymbol{\mu}_g$，$n=n_1+\ldots+n_g$

记总平均$\bar{\mathbf{x}}=\sum_{k=1}^{g}\sum_{j=1}^{n_l}\mathbf{x}_{kj}/n=\sum_{l=1}^{g}n_k\bar{\mathbf{x}}_k/n$，记$\boldsymbol{\mu}=E(\bar{\mathbf{x}})=\sum_{l=1}^{g}n_l\boldsymbol{\mu}_l/n$，

令$\boldsymbol{\tau}_k=\boldsymbol{\mu}_k-\boldsymbol{\mu}$为水平$k$的效应，即$\boldsymbol{\mu}_k=\boldsymbol{\mu}+\boldsymbol{\tau}_k$.

模型重新表示为：

$\mathbf{x}_{ki}=\boldsymbol{\mu}+\boldsymbol{\tau}_k+\boldsymbol{\varepsilon}_{ki}$，$\boldsymbol{\varepsilon}_{ki}$ $iid \sim N(0,\Sigma), k=1,\ldots,g; i=1,\ldots,n_k$

注意$\boldsymbol{\tau}$'s有约束：$\sum_{l=1}^{g}n_l\boldsymbol{\tau}_l=0$ (sum contrast).

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

模型： $\mathbf{x}_{ki} - \boldsymbol{\mu} = \boldsymbol{\tau}_k + \boldsymbol{\varepsilon}_{ki}$

估计： $\mathbf{x}_{kj} - \bar{\mathbf{x}} = \bar{\mathbf{x}}_k - \bar{\mathbf{x}} + \mathbf{x}_{kj} - \bar{\mathbf{x}}_k$， 两边同时“平方”并求和

$$
\begin{aligned}
T = (n-1)S &= \sum_{k=1}^{g}\sum_{j=1}^{n_l}(\mathbf{x}_{kj} - \bar{\mathbf{x}})(\mathbf{x}_{kj} - \bar{\mathbf{x}})^{\mathsf{T}} \\
&= \sum_{k=1}^{g}\sum_{j=1}^{n_k}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})^{\mathsf{T}} + \sum_{k=1}^{g}\sum_{j=1}^{n_k}(\mathbf{x}_{kj} - \bar{\mathbf{x}}_k)(\mathbf{x}_{kj} - \bar{\mathbf{x}}_k)^{\mathsf{T}} \\
&= \sum_{k=1}^{g}\sum_{j=1}^{n_k}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})^{\mathsf{T}} + \sum_{k=1}^{g}(n_k - 1)S_k \overset{\Delta}{=} B + W
\end{aligned}
$$

B : Between - group

W : Within - group

如果组间变差矩阵B“较大”(相对于总和T或组内W)，则否定H_0.

- $p = 1$时($ANOVA$)，直接比较B和W：

$$F = \frac{B/(g-1)}{W/(n-g)} \overset{H_0}{\sim} F_{g-1,n-g} \Leftrightarrow \frac{B}{B+W} \overset{H_0}{\sim} \text{beta}\left(\frac{g-1}{2}, \frac{n-g}{2}\right)$$

- $p > 1$时，B, W都是$p \times p$方阵，U, V的行列式、trace都可用来检验：

$$U = (B+W)^{-1/2} W (B+W)^{-1/2}, \quad V = B^{-1/2}(B+W)B^{-1/2}, \quad BW^{-1}, \ldots$$

原假设成立时（各组均值相同），B,W服从独立的Wishart分布，
U,V服从多元beta分布 (Hsu 1939, Olkin & Rubin 1963)。
拓展到多个独立的Wishart矩阵，会得到多元Dirichlet分布。

若$W \sim W_p(m_1, \Sigma), B \sim W_p(m_2, \Sigma)$,独立，$U = (W+B)^{-1/2} W (W+B)^{-1/2}$
服从多元Beta分布$B_p(m_1/2, m_2/2)$

$$f(U) = \frac{\Gamma_p(m_1/2 + m_2/2)}{\Gamma_p(m_1/2)\Gamma_p(m_2/2)} |U|^{(m_1-p-1)/2} |I_p - U|^{(m_2-p-1)/2}.$$

U是方阵，U的行列式的分布称为Wilks' lambda分布：

$$\Lambda^* \overset{\Delta}{=} |U| = |W| / |W+B| \sim \Lambda_p(m_1, m_2) = \prod_{i=1}^{d} \prod^{m_2} \text{beta}\left(\frac{m_1 - p + i}{2}, \frac{p}{2}\right),$$

该分布非常复杂。下面从似然比检验出发求出其近似逼近，即*Wilks*检验。

似然比检验/Wilks'检验

数理统计的似然理论指出，基于似然的极大似然估计是渐近最优的（CR不等式），似然比检验LRT（以及渐近等价的Score、Wald检验）具有渐近最优性，分布收敛于卡方。

Wilks定理. 假设$\mathbf{x}_i \in R^p, i=1,...,n,$的似然函数为$L(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta,$
原假设$H_0 : \boldsymbol{\theta} \in \Theta_0, \nu = \dim(\Theta), \nu_0 = \dim(\Theta_0),$似然比检验统计量

$$\Lambda = \frac{\max\limits_{\boldsymbol{\theta}\in\Theta_0} L(\boldsymbol{\theta})}{\max\limits_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})} = \frac{L(\hat{\boldsymbol{\theta}}_0)}{L(\hat{\boldsymbol{\theta}})}$$

其中$\hat{\boldsymbol{\theta}}$为$\boldsymbol{\theta}$的极大似然估计，$\hat{\boldsymbol{\theta}}_0$为原假设下的$\boldsymbol{\theta}$的极大似然估计。
记$\nu = \dim(\Theta), \nu_0 = \dim(\Theta_0)$，则$H_0$成立时，在正则(regular)条件下

$$-2\log\Lambda \xrightarrow{d} \chi^2_{\nu-\nu_0}, n \to \infty.$$

Wilks'统计量定义为$\Lambda^* = \Lambda^{2/n}$，$-2\log\Lambda = -n\log\Lambda^*$

对于MANOVA问题（g个正态总体均值检验问题）：

$$\begin{cases} \mathbf{x}_{11},...,\mathbf{x}_{1n_1} \ \ iid \sim N_p(\boldsymbol{\mu}_1,\Sigma), & \text{样本均值和方差：} \bar{\mathbf{x}}_1, S_1 \\ ... \\ \mathbf{x}_{g1},...,\mathbf{x}_{gn_g} \ \ iid \sim N_p(\boldsymbol{\mu}_g,\Sigma), & \text{样本均值和方差：} \bar{\mathbf{x}}_g, S_g \end{cases}$$

$$\Theta = \{\boldsymbol{\mu}_1,...,\boldsymbol{\mu}_g,\Sigma\}, \Theta_0 = \{\boldsymbol{\mu} = \boldsymbol{\mu}_1 = ... = \boldsymbol{\mu}_g,\Sigma\}.$$

- 原假设成立时，极大似然估计为：$\hat{\boldsymbol{\mu}}_0 = \bar{\mathbf{x}},\ \hat{\Sigma}_0 = (B+W)/n,$
- 没有限制时，极大似然估计：$\hat{\boldsymbol{\mu}}_k = \bar{\mathbf{x}}_k, k = 1,...,g, \hat{\Sigma} = W/n,$

$$\Lambda = \frac{\max\limits_{\boldsymbol{\theta}\in\Theta_0} L(\boldsymbol{\theta})}{\max\limits_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})} = \frac{L(\hat{\boldsymbol{\mu}}_0,\hat{\Sigma}_0)}{L(\hat{\boldsymbol{\mu}}_1,...,\hat{\boldsymbol{\mu}}_g,\hat{\Sigma})} = \frac{|\hat{\Sigma}|^{n/2}}{|\hat{\Sigma}_0|^{n/2}} = \frac{|W|^{n/2}}{|B+W|^{n/2}}$$

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

由Wilks定理，$-2\log(\Lambda) = -n\log\left(\frac{|W|}{|B+W|}\right) \xrightarrow{d} \chi^2_{(g-1)p}$。

One-way MANOVA：Wilks'检验

令Wilks检验统计量(Wilks' Lambda）

$$\Lambda^* = \frac{\det(W)}{\det(W+B)}$$

则 H_0成立时 $T = -n\log\Lambda^* \xrightarrow{d} \chi^2_{(g-1)p}$， $n \to \infty$

当$g>2$且$p>1$时：当$T \geq \chi^2_{p(g-1)}(\alpha)$时，在$\alpha$水平下否定原假设。

Bartlett修正：$T_2 = -(n-1-\frac{p+g}{2})\log\Lambda^* \xrightarrow{d} \chi^2_{(g-1)p}$,

注意：当$g=1,2$或$p=1$时，没必要应用上述近似检验。

$g=2$时，$-2\log\Lambda = n\log(1+T^2/(n_1+n_2-2))$，其中

$$T^2 = \frac{n_1 n_2}{n}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathsf{T}} S_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$服从cF分布(精确分布)

$p=1$时，B/W服从cF分布(精确分布).

$g=2$时，Hotelling T2检验与Wilks检验等价, 即Hotelling检验实际上也是只与协方差矩阵有关（作业）。单正态总体情形也是如此：

例1. 设$\mathbf{x}_1,...,\mathbf{x}_n \ \ iid \sim N_p(\boldsymbol{\mu},\Sigma)$, $\quad H_0:\boldsymbol{\mu}=\boldsymbol{\mu}_0$（$\boldsymbol{\mu}_0$已知）,则Wilks$\Lambda^*$等价于$T^2$：

$$\Lambda^*=\frac{1}{1+T^2/(n-1)},\quad T^2=n(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^{\mathsf{T}}S^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)$$

证：$\hat{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^{\mathsf{T}}$,原假设下，$\hat{\Sigma}_0=\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu}_0)(\mathbf{x}_i-\boldsymbol{\mu}_0)^{\mathsf{T}}$,故

分解$\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu}_0)(\mathbf{x}_i-\boldsymbol{\mu}_0)^{\mathsf{T}}=\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^{\mathsf{T}}+n(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^{\mathsf{T}}$,

$$\text{所以}\Lambda^*=\frac{\det(\hat{\Sigma})}{\det(\hat{\Sigma}_0)}=\frac{\det\left(\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^{\mathsf{T}}\right)}{\det\left(\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^{\mathsf{T}}+n(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^{\mathsf{T}}\right)}$$

$$=\frac{\det(S)}{\det\left(S+n(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^{\mathsf{T}}/(n-1)\right)}=\frac{1}{\det\left(I_p+n(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^{\mathsf{T}}S^{-1}/(n-1)\right)}$$

$$=\frac{1}{1+n(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^{\mathsf{T}}S^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)/(n-1)}=\frac{1}{1+T^2/(n-1)}.$$

其它检验

因为B, W都是矩阵，Wilks Λ^*

$$\Lambda^* = \frac{|\mathrm{W}|}{|\mathrm{W}+\mathrm{B}|} = \frac{1}{|\mathrm{I}_p + \mathrm{BW}^{-1}|},$$

以行列式度量B相对于W是否接近0，其它检验方法还有：

- Lawley - Hotelling trace: $tr(\mathrm{BW}^{-1})$
- Phillai trace: $tr(\mathrm{B}(\mathrm{B}+\mathrm{W})^{-1})$
- Roy's largest root(最大特征根): $\lambda_{\max}(\mathrm{W}(\mathrm{B}+\mathrm{W})^{-1})$

例2 (例6.10， Johnson and Wichern p233). 威斯康星州卫生和社会服务部需要给养老院（nursing home）提供补贴, 补贴多少依据护理等级、护理成本或职工工资水平等。养老院分私营、非盈利经营和国企等三种所有权形式。我们希望了解养老院的运行成本是否与所有权形式有关，共统计了 $p = 4$种人力成本（护理、膳食、设备运行和维护、清洁维护）。各组数据的均值方法如下：

Group	Number of observations	Sample mean vectors
$\ell = 1$ (private)	$n_1 = 271$	$\bar{\mathbf{x}}_1 = \begin{bmatrix} 2.066 \\ .480 \\ .082 \\ .360 \end{bmatrix}$; $\bar{\mathbf{x}}_2 = \begin{bmatrix} 2.167 \\ .596 \\ .124 \\ .418 \end{bmatrix}$; $\bar{\mathbf{x}}_3 = \begin{bmatrix} 2.273 \\ .521 \\ .125 \\ .383 \end{bmatrix}$
$\ell = 2$ (nonprofit)	$n_2 = 138$	
$\ell = 3$ (government)	$n_3 = 107$	
	$\sum_{\ell=1}^{3} n_\ell = 516$	

Sample covariance matrices

$$\mathbf{S}_1 = \begin{bmatrix} .291 & & & \\ -.001 & .011 & & \\ .002 & .000 & .001 & \\ .010 & .003 & .000 & .010 \end{bmatrix}; \quad \mathbf{S}_2 = \begin{bmatrix} .561 & & & \\ .011 & .025 & & \\ .001 & .004 & .005 & \\ .037 & .007 & .002 & .019 \end{bmatrix};$$

$$\mathbf{S}_3 = \begin{bmatrix} .261 & & & \\ .030 & .017 & & \\ .003 & -.000 & .004 & \\ .018 & .006 & .001 & .013 \end{bmatrix}$$

由上页数据计算得到总平均$\bar{\mathbf{x}}$，组间和组内“平方和”B, W:

$$\bar{\mathbf{x}} = \frac{n_1\bar{\mathbf{x}}_1 + n_2\bar{\mathbf{x}}_2 + n_3\bar{\mathbf{x}}_3}{n_1 + n_2 + n_3} = \begin{bmatrix} 2.136 \\ .519 \\ .102 \\ .380 \end{bmatrix}$$

$$\mathbf{B} = \sum_{\ell=1}^{3} n_\ell(\bar{\mathbf{x}}_\ell - \bar{\mathbf{x}})(\bar{\mathbf{x}}_\ell - \bar{\mathbf{x}})' = \begin{bmatrix} 3.475 & & & \\ 1.111 & 1.225 & & \\ .821 & .453 & .235 & \\ .584 & .610 & .230 & .304 \end{bmatrix}$$

$$\mathbf{W} = (n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + (n_3 - 1)\mathbf{S}_3 = \begin{bmatrix} 182.962 & & & \\ 4.408 & 8.200 & & \\ 1.695 & .633 & 1.484 & \\ 9.581 & 2.428 & .394 & 6.538 \end{bmatrix}$$

$$\Lambda^* = \frac{\det(\mathrm{W})}{\det(\mathrm{W}+\mathrm{B})} = 0.7714, \quad -n\log\Lambda^* = 132.8 > \chi^2_{p(g-1)}(\alpha) = \chi^2_8(0.01) = 20.09,$$

拒绝原假设。$pvalue = P(\chi^2_{p(g-1)} \geq 132.8) = 0.001$

总结：正态均值检验

g个总体：$\mathbf{x}_{k1},\dots,\mathbf{x}_{kn_k}\sim N_p(\boldsymbol{\mu}_k,\Sigma), k=1,\dots,g;$
零假设：$H_0:\boldsymbol{\mu}_1=\cdots=\boldsymbol{\mu}_g$

	精确检验($g\le 2$)		近似检验($g>2$)	
	单总体 ($g=1$)	两总体 ($g=2$)	多总体 ($g>2$)	$g=\infty$
统计量	$T^2=n\bar{\mathbf{x}}^\top S^{-1}\bar{\mathbf{x}}$	$T^2=\frac{n_1n_2}{n_1+n_2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^\top S^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)$	$-n\log\frac{\lvert W\rvert}{\lvert W+B\rvert}$	
零分布	$\frac{(n-1)p}{n-p}F_{p,n-p}$	$\frac{(n-2)p}{n-p-1}F_{p,n-p-1}$	$\chi^2_{(g-1)p}$	
多元 ($p>1$)	Hotelling T^2检验（F检验）	Hotelling T^2检验（F检验）	MANOVA（卡方检验）	多元线性模型（卡方检验）
一元 ($p=1$)	F/t检验	F/t检验	ANOVA（F检验）	一元线性模型（F检验）

以上所有检验都只与协方差矩阵有关