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

第八讲　Hotelling's T^2 检验

2024.3.27

数据矩阵：横着排，竖着看

与Wishart有关的二次型

定理1：假设$W \sim W_p(m)$，$m \geq p$，则对$\forall$常数向量$\mathbf{t} \in S^{p-1}$，有

(1) $\mathbf{t}^\top W \mathbf{t} \sim \chi^2_m$;

(2) $\dfrac{1}{\mathbf{t}^\top W^{-1} \mathbf{t}} \sim \chi^2_{m-p+1}$.

证明：首先，因为$m \geq p$，W可逆。

(1)由第6讲定理1，$W \sim W_p(m, I_p) \Rightarrow \mathbf{t}^\top W \mathbf{t} \sim W_1(m, \mathbf{t}^\top I_p \mathbf{t}) = W_1(m, I_p) = \chi^2_m$。

(2)注意到对任何正交矩阵H,

$$V = HWH^\top \sim W_p(m, HH^\top) = W_p(m), \ \mathbf{t}^\top W^{-1} \mathbf{t} = \mathbf{t}^\top H^\top (HWH^\top)^{-1} H\mathbf{t},$$

取正交矩阵H的第一行为$\mathbf{t}^\top$,即 $H = \begin{pmatrix} \mathbf{t}^\top \\ * \end{pmatrix}$，则$\mathbf{u} = H\mathbf{t} = \begin{pmatrix} \mathbf{t}^\top \mathbf{t} \\ * \end{pmatrix} = \begin{pmatrix} 1 \\ \mathbf{0} \end{pmatrix}$

故 $\mathbf{t}^\top W^{-1} \mathbf{t} = \mathbf{t}^\top H^\top (HWH^\top)^{-1} H\mathbf{t} = \mathbf{u}^\top (HWH^\top)^{-1} \mathbf{u} = \mathbf{u}^\top V^{-1} \mathbf{u}.$

划分$V=\begin{pmatrix} v_{11} & \mathbf{v}_{12} \\ \mathbf{v}_{21} & V_{22} \end{pmatrix}$,其中$v_{11}$是$1\times 1$，则$V^{-1}=\begin{pmatrix} v_{11\bullet 2}^{-1} & * \\ * & * \end{pmatrix}$，所以

$$\mathbf{u}^{\mathsf{T}}V^{-1}\mathbf{u}=(1,\mathbf{0}^{\mathsf{T}})\begin{pmatrix} v_{11\bullet 2}^{-1} & * \\ * & * \end{pmatrix}\begin{pmatrix} 1 \\ \mathbf{0} \end{pmatrix}=v_{11\bullet 2}^{-1}$$

由定理3，　$1/\mathbf{t}^{\mathsf{T}}W^{-1}\mathbf{t}=v_{11\bullet 2}\sim W_1(m-p+1,1)=\chi^2_{m-p+1}$.

推论1：假设$W\sim W_p(m)$, $m\geq p$, $\mathbf{z}\sim N_p(\mathbf{0},I_p)$，$\mathbf{z}\perp\!\!\!\perp W$,

则$\dfrac{\mathbf{z}^{\mathsf{T}}\mathbf{z}}{\mathbf{z}^{\mathsf{T}}W^{-1}\mathbf{z}}\sim\chi^2_{m-p+1}$。

证明：$\mathbf{u}=\mathbf{z}/\|\mathbf{z}\|\sim U(S^{p-1})$,由定理1，给定$\mathbf{z}$条件下

$$\frac{\mathbf{z}^{\mathsf{T}}\mathbf{z}}{\mathbf{z}^{\mathsf{T}}W^{-1}\mathbf{z}}=\left.\frac{1}{\mathbf{u}^{\mathsf{T}}W^{-1}\mathbf{u}}\right|_{\mathbf{z}}\sim\chi^2_{m-p+1},$$

该分布与$\mathbf{z}$无关，所以$\dfrac{\mathbf{z}^{\mathsf{T}}\mathbf{z}}{\mathbf{z}^{\mathsf{T}}W^{-1}\mathbf{z}}\sim\chi^2_{m-p+1}$，且与$\mathbf{z}$独立。

定理2：假设$W \sim W_p(m,\Sigma)$, $\mathbf{z} \sim N(\mathbf{0},\Sigma)$, $\Sigma > 0$, $m \ge p$ (W可逆),
假设W与$\mathbf{z}$独立，则 $\left(\dfrac{m-p+1}{p}\right) \times \mathbf{z}^{\mathsf{T}} W^{-1} \mathbf{z} \sim F_{p,m-p+1}$

证明：不妨设$\Sigma = I_p$(否则令$\tilde{\mathbf{z}} = \Sigma^{-1/2}\mathbf{z} \sim N(0, I_p)$,
$\tilde{W} = \Sigma^{-1/2} W \Sigma^{-1/2} \sim W_p(m, I_p)$, $\mathbf{z}^{\mathsf{T}} W^{-1} \mathbf{z} = \tilde{\mathbf{z}}^{\mathsf{T}} \tilde{W}^{-1} \tilde{\mathbf{z}}$)。

由推论1，在$\mathbf{z}$给定的条件下

$$b = \mathbf{z}^{\mathsf{T}}\mathbf{z} / \mathbf{z}^{\mathsf{T}} W^{-1} \mathbf{z} \sim \chi^2_{m-p+1},$$

其分布不依赖于$\mathbf{z}$，所以b与$\mathbf{z}$独立进而与$a = \mathbf{z}^{\mathsf{T}}\mathbf{z} \sim \chi^2_p$独立.

所以由F分布的定义知$\dfrac{a/p}{b/(m-p+1)} \sim F_{p,m-p+1}$，即

$$\frac{\mathbf{z}^{\mathsf{T}}\mathbf{z}/p}{(\mathbf{z}^{\mathsf{T}}\mathbf{z}/\mathbf{z}^{\mathsf{T}}W^{-1}\mathbf{z})/(m-p+1)} = \frac{m-p+1}{p} \times \mathbf{z}^{\mathsf{T}} W^{-1} \mathbf{z} \sim F_{p,m-p+1}$$

定理3：设$\mathbf{x}_1,\ldots,\mathbf{x}_n \ \ iid \sim N_p(\boldsymbol{\mu},\Sigma)$,假设$n>p$,则

(1) $\bar{\mathbf{x}} \sim N_p\left(\boldsymbol{\mu},\frac{1}{n}\Sigma\right)$;

(2) $(n-1)S \sim W_p(n-1,\Sigma)$;

(3) $\bar{\mathbf{x}}$与S独立.

(4) $T^2 = n(\bar{\mathbf{x}}-\boldsymbol{\mu})^\top S^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) \sim \frac{(n-1)p}{n-p}F_{p,n-p}$.

证明：（1）显然$\bar{\mathbf{x}} \sim N_p\left(\boldsymbol{\mu},\frac{1}{n}\Sigma\right)$.

(2) 令$X=(\mathbf{x}_1,\ldots,\mathbf{x}_n)^\top$，我们已知$(n-1)S = X^\top AX$，$A=(I_n-\mathbf{1}\mathbf{1}^\top/n)$。

因为rank$(A)=tr(A)=n-1$，由$Cochran$定理,$(n-1)S = X^\top AX \sim W_p(n-1,\Sigma)$.

(3) 因为$A\mathbf{1}=0$,所以$\mathbf{1}^\top X \perp\!\!\!\perp AX$,

$\Rightarrow \ \bar{\mathbf{x}} = X^\top\mathbf{1}/n \ \perp\!\!\!\perp \ (n-1)S = X^\top AX$

(4) 我们已知： $\mathbf{z} \overset{\Delta}{=} \sqrt{n}(\bar{\mathbf{x}}-\boldsymbol{\mu}) \sim N_p(0,\Sigma)$, $(n-1)S \sim W_p(n-1,\Sigma)$,

且$\bar{\mathbf{x}}$与S独立。由定理2，

$$\mathbf{z}^{\mathsf{T}}\left[(n-1)S\right]^{-1}\mathbf{z} = \frac{1}{n-1} n(\bar{\mathbf{x}}-\boldsymbol{\mu})^{\mathsf{T}} S^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) \sim \frac{p}{n-p} F_{p,n-p}$$

即$n(\bar{\mathbf{x}}-\boldsymbol{\mu})^{\mathsf{T}} S^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) \sim \dfrac{(n-1)p}{n-p} F_{p,n-p}$.

由定理3，我们可以构造均值的置信椭球如下

置信椭球

设$\mathbf{x}_1,\dots,\mathbf{x}_n$ $iid \sim N_p(\boldsymbol{\mu},\Sigma), n > p$,

则$\boldsymbol{\mu}$的置信水平$1-\alpha$的置信椭球为：

$$\left\{\boldsymbol{\mu}: \ n(\bar{\mathbf{x}}-\boldsymbol{\mu})^{\mathsf{T}} S^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) \le \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)\right\}$$

单正态总体的Hotelling's T^2检验

设$\mathbf{x}_1,\ldots,\mathbf{x}_n$ $iid \sim N_p(\boldsymbol{\mu},\Sigma)$, $n>p$。$H_0:\boldsymbol{\mu}=\boldsymbol{\mu}_0$($\boldsymbol{\mu}_0$已知)

Hotelling T^2 检验统计量

$$T^2=n(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^{\mathrm{T}}S^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)\overset{H_0}{\sim}\frac{(n-1)p}{n-p}F_{p,n-p},$$

当$T^2>\dfrac{(n-1)p}{n-p}F_{p,n-p}(\alpha)$,在$\alpha$水平下否定原假设。

注：$p=1$时,T^2是student's t的平方：

$$T^2=\left(\sqrt{n}(\bar{x}-\mu_0)/s\right)^2=t^2\sim_{\mathrm{H}_0}F_{1,n-1},\quad t=\sqrt{n}(\bar{x}-\mu_0)/s\sim_{\mathrm{H}_0}t_{n-1}。$$

单正态总体的Hotelling T^2检验主要用于成对设计

成对Hotelling's T^2检验

假设成对设计中第i对研究对象的测量为$(\mathbf{x}_{i1},\mathbf{x}_{i2}), i=1,\ldots,n,$

假设差异$\mathbf{d}_i=\mathbf{x}_{i1}-\mathbf{x}_{i2}, i=1,\ldots,n$ iid $\sim N_p(\boldsymbol{\delta},\Sigma)$,

原假设为$H_0:\boldsymbol{\delta}=\mathbf{0}$。成对Hotelling's T^2检验:

$$T^2=n\bar{\mathbf{d}}^{\mathrm{T}}S^{-1}\bar{\mathbf{d}}\geq\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\text{时否定}H_0。$$

例6.1. 城市废水处理后，在排往河流之前需要检测其有害成分是否超标。废水处理公司有自己的检测系统(commercial lab), 为了考察其水质检测是否可靠，威斯康星州水文局的实验室也进行了检测。11分处理后的样品各被分成两份，由公司和水文局独立检测其生化需氧量指标BOD和悬浮物（浓度）指标SS。 数据如下， 两个实验室的检测结果是否一致 ？

TABLE 6.1 EFFLUENT DATA

Sample j	Commercial lab x_{1j1} (BOD)	Commercial lab x_{1j2} (SS)	State lab of hygiene x_{2j1} (BOD)	State lab of hygiene 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

- BOD（biochemical oxygen demand）：水中有机微生物分解有机污染物质所消耗的氧气量，单位mg/L（无污染：1mg/L，污水处理后正常值20mg/L）；
- SS（suspended solids）：直径大于2μm的难以过滤的无机悬浮物的直径长度，单位：μm (micron) 。

这是一个成对比较问题，我们要检验指标均值之差是否为**0**。两个实验室测得的指标之差($\boldsymbol{d}_j, j=1,\dots,11$)如下表：

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

$\boldsymbol{d}_j$, $j=1,\dots,11$, 的样本均值和样本协方差：

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

Hotelling T^2检验统计量：

$$T^2 = 11[-9.36, \quad 13.27]\begin{bmatrix} .0055 & -.0012 \\ -.0012 & .0026 \end{bmatrix}\begin{bmatrix} -9.36 \\ 13.27 \end{bmatrix} = 13.6 \; > 9.47 = \frac{p(n-1)}{n-p}F_{p,n-p}(\alpha)$$

在0.05水平下拒绝原假设, pvalue = 0.021.

两正态总体的均值相同性检验：两样本Hotelling's T^2检验

与单个正态总体的定理1类似，两正态问题有以下结论：

定理4.假设两个多元正态总体（方差相同但均值可以不同）：

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

$\mathbf{x}_{21},\dots,\mathbf{x}_{2n_2}\ iid \sim N_p(\boldsymbol{\mu}_2,\Sigma)$，样本均值和方差为$\bar{\mathbf{x}}_2, S_2$。

令 $S_{\text{pooled}} = \dfrac{1}{n_1+n_2-2}\left((n_1-1)S_1+(n_2-1)S_2\right)$，总样本量$n=n_1+n_2$，则

$$T^2 = \frac{1}{1/n_1+1/n_2}\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right)^{\mathrm{T}} S_{\text{pooled}}^{-1}\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right) \sim \frac{(n-2)p}{n-p-1}F_{p,n-p-1}$$

$$\bar{\mathbf{x}}_1=\sum_{i=1}^{n_1}\mathbf{x}_{1i}/n_1,\quad \bar{\mathbf{x}}_2=\sum_{i=1}^{n_1}\mathbf{x}_{2i}/n_1,$$

$$S_1=\sum_{i=1}^{n_1}(\mathbf{x}_{1i}-\bar{\mathbf{x}}_1)(\mathbf{x}_{1i}-\bar{\mathbf{x}}_1)^{\mathrm{T}}/(n_1-1), S_2=\sum_{i=1}^{n_2}(\mathbf{x}_{2i}-\bar{\mathbf{x}}_2)(\mathbf{x}_{2i}-\bar{\mathbf{x}}_2)^{\mathrm{T}}/(n_2-1)$$

证明：

$$\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2 \sim N_p\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2,\left(1/n_1+1/n_2\right)\Sigma\right),$$

$$\mathbf{z}=(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2))/(1/n_1+1/n_2)\sim N_p(\mathbf{0},\Sigma),$$

由定理6,$(n_1-1)S_1\sim W_p(n_1-1,\Sigma)$, $(n_2-1)S_2\sim W_p(n_2-1,\Sigma)$,两者独立

$$\Rightarrow W=(n_1-1)S_1+(n_2-1)S_2\sim W_p(n-2,\Sigma)$$

另外，$\mathbf{z}$与W独立

由定理2,$\left(\dfrac{n-2-p+1}{p}\right)\times\mathbf{z}^{\mathrm{T}}W^{-1}\mathbf{z}=\left(\dfrac{n-p-1}{(n-2)p}\right)T^2\sim F_{p,n-p-1}$

定理4结果可用于构造置信域，也可以用于构造两样本Hotelling’s T^2检验。

置信域

$\boldsymbol{\delta}=\boldsymbol{\mu}_1-\boldsymbol{\mu}_2$的置信度为$1-\alpha$置信域（置信椭球）为

$$\left\{\boldsymbol{\delta}:\ \frac{1}{1/n_1+1/n_2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2-\boldsymbol{\delta})^{\mathrm{T}}S_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2-\boldsymbol{\delta})\le\frac{(n-2)p}{n-p-1}F_{p,n-p-1}(\alpha)\right\}$$

两样本Hotelling's T^2检验是应用最为广泛的统计方法之一。

两样本Hotelling's T^2检验

两样本Hotelling's T^2检验： $H_0: \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_0$ ($\boldsymbol{\delta}_0$已知)

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

$\mathbf{x}_{21},\ldots,\mathbf{x}_{2n_2}$ $iid \sim N_p(\boldsymbol{\mu}_2,\Sigma)$，样本均值和方差：$\bar{\mathbf{x}}_2, S_2$

记 $S_{\text{pooled}} = \dfrac{1}{n-2}\left((n_1-1)S_1 + (n_2-1)S_2\right)$, $n = n_1 + n_2$.

$$\text{Hotelling's } T^2 = \frac{1}{1/n_1 + 1/n_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta}_0)^{\mathrm{T}} S_{\text{pooled}}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta}_0)$$

当$T^2 \geq \dfrac{(n-2)p}{n-p-1} F_{p,n-p-1}(\alpha)$时，拒绝$H_0$（水平$\alpha$）。

Harold Hotelling (1895 – 1973) was an American mathematical statistician and an influential economic theorist, known for Hotelling's law, Hotelling's lemma, and Hotelling's rule in economics, as well as Hotelling's T2 distribution in statistics. He also developed and named the principal component analysis method widely used in finance, statistics and computer science.