第九讲　Hotelling's T^2 检验

2025.3.31

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_{\mathbf{0}})^\top S^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_{\mathbf{0}}) \sim_{H_0} cF_{p,n-p}$$

多元样本均值和样本方差矩阵

样本均值样本方差的无偏性

假设 $\mathbf{x}_1,\ldots,\mathbf{x}_n \in R^p$ iid $\sim(\boldsymbol{\mu},\Sigma)$，即总体均值和方差矩阵分别为 $\boldsymbol{\mu}=E(\mathbf{x}_1),\Sigma=var(\mathbf{x}_1)$，总体未必正态。

$\bar{\mathbf{x}}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$，$S=\frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^\top$

命题1. 假设 $\mathbf{x}_1,\ldots,\mathbf{x}_n \in R^p$ iid $\sim(\boldsymbol{\mu},\Sigma)$, 则样本均值和样本方差分别是总体均值和方差矩阵的无偏估计：

$$E(\bar{\mathbf{x}})=\boldsymbol{\mu},\quad var(\bar{\mathbf{x}})=\frac{\Sigma}{n},\ E(S)=\Sigma.$$

证明：$E(\bar{\mathbf{x}})=\frac{1}{n}\sum_{i=1}^{n}E\mathbf{x}_i=\boldsymbol{\mu},\ \mathrm{var}(\bar{\mathbf{x}})=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(\mathbf{x}_i)=\frac{\Sigma}{n}.$

注意到

$$(n-1)S=\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^\top=\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^\top-n(\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})^\top$$

而 $E(\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^\top)=nE(\mathbf{x}_1-\boldsymbol{\mu})(\mathbf{x}_1-\boldsymbol{\mu})^\top=n\mathrm{var}(\mathbf{x}_1)=n\Sigma$ ，

$E(n(\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})^\top)=n\mathrm{var}(\bar{\mathbf{x}})=\Sigma\ \Rightarrow \mathrm{E}(n-1)S=n\Sigma-\Sigma$

$\Rightarrow E(S)=\Sigma$

多元正态分布参数的极大似然估计

极大似然估计MLE

命题2. 假设$\mathbf{x}_1, \dots, \mathbf{x}_n$ $iid \sim N_p(\boldsymbol{\mu}, \Sigma)$, 则$\boldsymbol{\mu}, \Sigma$的极大似然估计分别为 $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$, $\hat{\Sigma} = S^* = (n-1)S/n$

注：极大似然估计是最优或渐近最优的。$S^* \approx S$。

证明1：似然函数（$\mathbf{x}_1, \dots, \mathbf{x}_n$的联合概率密度）：

$$L(\boldsymbol{\mu}, \Sigma) = p(\mathbf{x}_1, \dots, \mathbf{x}_n)$$

$$= \frac{C}{|\Sigma|^{n/2}} \prod_{i=1}^{n} \exp(-(\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})/2)$$

$$\mathrm{C} = \frac{1}{(2\pi)^{np/2}}$$

$$= \frac{C}{|\Sigma|^{n/2}} \exp\left(-\frac{n}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \Sigma^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) - \frac{1}{2}\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})^\top \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})\right)$$

显然$\boldsymbol{\mu}$的极大点$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$。

下面只需极大化

$$\frac{1}{|\Sigma|^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})^\top \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})\right)$$

$$= |\Omega|^{n/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\Omega \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top)\right)$$

$$\Omega = \Sigma^{-1}$$

$$= |\Omega|^{n/2} \exp\left(-\frac{n}{2}\mathrm{tr}(\Omega S^*)\right),$$

故求Σ的极大似然估计等价于

$$\max_{\Omega>0} f(\Omega)，\ f(\Omega)=\log(|\Omega|)-\mathrm{tr}(\Omega S^*)$$

记$A=\Omega S^*$，其所有特征根 $\lambda_1,\dots,\lambda_p$

$$\begin{aligned} f(\Omega) &= \log(|\Omega|)-\mathrm{tr}(\Omega S^*)=\log(|\Omega S^*|)-\mathrm{tr}(\Omega S^*)-\log(|S^*|) \\ &= \log(|A|)-\mathrm{tr}(A)-\log(|S^*|) \\ &= \textstyle\sum_{i=1}^{p}(\log(\lambda_i)-\lambda_i)-\log(|S^*|) \end{aligned}$$

令$\frac{\partial f}{\partial\lambda_i}=\frac{1}{\lambda_i}-1=0 \Rightarrow \lambda_i=1, i=1,\dots,p$

$\Rightarrow A=I_p \Rightarrow$最优解$\hat{\Omega}=S^{*-1}$，$\hat{\Sigma}=S^*$

方差矩阵的优化目标函数

当Σ或$\Omega=\Sigma^{-1}$不是完全未知、具有参数结构时，比如$\Omega=\Omega(\boldsymbol{\theta})$，似然方法极大化：

$$\max_{\Omega(\boldsymbol{\theta})>0}\big(\log(|\Omega(\boldsymbol{\theta})|)-\mathrm{tr}(\Omega(\boldsymbol{\theta})S^*)\big),$$

其中一般用S替代S^*。

证明2. 不限制Ω对称，$f(\Omega) = \log(|\Omega|) - \mathrm{tr}(\Omega S^*)$对$\Omega$求导

$$\frac{\partial f}{\partial \Omega} = \frac{\partial \log(|\Omega|)}{\partial \Omega} - \frac{\partial \mathrm{tr}(\Omega S^*)}{\partial \Omega} = \Omega^{-1} - S^*$$

令之为0，得$\hat{\Omega} = {S^*}^{-1}$，$\hat{\Sigma} = S^*$

矩阵、向量导数：

(1)A对称,则$\dfrac{\partial \mathbf{x}^{\mathrm{T}} A \mathbf{x}}{\partial \mathbf{x}} = 2A\mathbf{x}$

(2)X，A为$p \times p$矩阵，则$\dfrac{\partial \log|X|}{\partial X} = X^{-1}$，$\dfrac{\partial \mathrm{tr}(AX)}{\partial X} = A^{\mathrm{T}}$.

一些矩阵微商公式表

$y = f(X)$	$\frac{\partial y}{\partial X}$
$a'x$	a
$x'Ax$	$2Ax$
$\mid X \mid$	$\begin{cases} \mid X \mid (X^{-1})', & \\ \mid X \mid (2X^{-1} - \mathrm{diag}(X^{-1})) & (X\text{对称}) \end{cases}$
$\mid AXB \mid$	$\mid AXB \mid A'((AXB)^{-1})'B'$
$\ln \mid AXB \mid$	$A'((AXB)^{-1})'B'$
$\ln \mid X'AX \mid$	$2AX(X'AX)^{-1}$ $(A\text{ 对称})$
$\mathrm{tr}(XAX')$	$X(A + A')$
$\mathrm{tr}(AXB)$	$A'B'$
$\mathrm{tr}(X'AXB)$	$AXB + A'XB'$
$\ln \mid X \mid$	$2X^{-1} - \mathrm{diag}(X^{-1})$ $(X\text{ 对称})$
$\mathrm{tr}(AX)$	$A + A' - \mathrm{diag}(A)$ $(X\text{ 对称})$

参见王松桂等（2004）线性模型引论 P49

http://staff.ustc.edu.cn/~ynyang/vector/books/Wang-linear-model.pdf

正态随机向量的线性组合

第4讲多元正态性质罗列中漏掉了如下性质（至目前还没用到过）

命题3. 假设$\mathbf{x}_i \sim N_p(\boldsymbol{\mu}_i, \Sigma_i), i = 1,2$独立，则
$$\mathbf{x}_1 + \mathbf{x}_2 \sim N_p(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \Sigma_1 + \Sigma_2)$$

证明1：因为$\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma) \Leftrightarrow$ 矩母函数
$$E exp(\mathbf{t}^\top \mathbf{x}) = \exp(\mathbf{t}^\top \boldsymbol{\mu} + \mathbf{t}^\top \Sigma \mathbf{t}/2), \forall \mathbf{t} \in R^p,$$
则
$$E exp(\mathbf{t}^\top (\mathbf{x}_1 + \mathbf{x}_2)) = E exp(\mathbf{t}^\top \mathbf{x}_1) E exp(\mathbf{t}^\top \mathbf{x}_2)$$
$$= \exp(\exp(\mathbf{t}^\top (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + \mathbf{t}^\top (\Sigma_1 + \Sigma_2)\mathbf{t}/2)) \,.$$
这说明$\mathbf{x}_1 + \mathbf{x}_2 \sim N_p(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \Sigma_1 + \Sigma_2)$。

证明2： 由$\mathbf{x}_1, \mathbf{x}_2$的联合概率密度可知$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim N_{2p}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}\right)$,

由多元正态性质4(第4讲)，$\mathbf{x}_1 + \mathbf{x}_2 = (I_p, I_p) \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim N_p(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \Sigma_1 + \Sigma_2)$.

样本均值和方差的分布

命题4. 假设$\mathbf{x}_1, \dots, \mathbf{x}_n$ $iid \sim N_p(\boldsymbol{\mu}, \Sigma)$, 则

(1) $\bar{\mathbf{x}} \sim N_p\left(\boldsymbol{\mu}, \frac{\Sigma}{n}\right)$; (2) $(n-1)S \sim W_p(n-1, \Sigma)$; (3) $\bar{\mathbf{x}} \perp\!\!\!\perp S$.

证明:(1) 由引理1,$\bar{\mathbf{x}} \sim N_p(\boldsymbol{\mu}, \Sigma/n)$.

(2) 参见第8讲定理3,

记$X = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$, $Z = (\mathbf{x}_1 - \boldsymbol{\mu}, \dots, \mathbf{x}_n - \boldsymbol{\mu})^\top = X - \mathbf{1}\boldsymbol{\mu}^\top$,

其中$\mathbf{x}_1 - \boldsymbol{\mu}, \dots, \mathbf{x}_n - \boldsymbol{\mu}$ $iid \sim N_p(\mathbf{0}, \Sigma)$,则

$$(n-1)S = X^\top\left(I_n - \frac{\mathbf{1}\mathbf{1}^\top}{\boldsymbol{n}}\right)X = Z^\top\left(I_n - \frac{\mathbf{1}\mathbf{1}^\top}{\boldsymbol{n}}\right)Z$$

由Cochran定理$(n-1)S \sim W_p(n-1, \Sigma)$。

(3) 因为$\bar{\mathbf{x}} = \frac{X^\top \mathbf{1}}{n} = \frac{Z^\top \mathbf{1}}{n} + \boldsymbol{\mu}$, $\left(I_n - \frac{\mathbf{1}\mathbf{1}^\top}{\boldsymbol{n}}\right)\mathbf{1} = \mathbf{0}$

$\Rightarrow \mathbf{1}^\top Z \perp\!\!\!\perp \left(I_n - \frac{\mathbf{1}\mathbf{1}^\top}{\boldsymbol{n}}\right)Z$

第8讲引理2. $AB^\top = 0 \Rightarrow AZ \perp\!\!\!\perp BZ$。

$\Rightarrow \mathbf{1}^\top Z \perp\!\!\!\perp Z^\top\left(I_n - \frac{\mathbf{1}\mathbf{1}^\top}{\boldsymbol{n}}\right)Z = (n-1)S$,

所以$\bar{\mathbf{x}} \perp\!\!\!\perp S$。

多元正态模型下的显著性检验

多元分析中方差矩阵完全未知（或几乎完全未知):

假设$\mathbf{x}_1, \ldots, \mathbf{x}_n\ iid \sim N_p(\boldsymbol{\mu}, \Sigma)$。$\bar{\mathbf{x}} \sim N_p(\boldsymbol{\mu}, \Sigma/n)$, $\mathrm{var}(\bar{\mathbf{x}}) = \Sigma/n$完全未知

以马氏距离度量$\bar{\mathbf{x}}$ 和$\boldsymbol{\mu}$之间的差距

$$T(\boldsymbol{\mu}, \Sigma) = (\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \mathrm{var}(\bar{\mathbf{x}})^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) = n(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \Sigma^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \sim \chi_p^2$$

为了检验H_0：$\boldsymbol{\mu} = \boldsymbol{\mu_0}$（$\boldsymbol{\mu_0}$已知）， $T(\boldsymbol{\mu}, \Sigma)$中$\boldsymbol{\mu}$换成$\boldsymbol{\mu_0}$，并代入$\Sigma$的估计$S$，得到Wald检验统计量(Hotelling's T^2)

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu_0})^\top S^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu_0}) \sim_{H_0} cF_{p,n-p}$$（定理1）

回归分析中的多参数检验中方差矩阵除了一个参数σ^2，完全已知：

$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N\left(0, \sigma^2 I_p\right), \hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2 (X^\top X)^{-1})$，$\mathrm{var}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 (X^\top X)^{-1}$，
σ^2未知，$(X^\top X)^{-1}$已知

以马氏距离度量$\hat{\boldsymbol{\beta}}$与$\boldsymbol{\beta}$的差距

$$T(\boldsymbol{\beta}, \sigma^2) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathrm{var}\left(\hat{\boldsymbol{\beta}}\right)^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (X^\top X)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)/\sigma^2 \sim \chi_p^2$$

与多元正态均值检验问题不同，这里方差矩阵$\mathrm{var}\left(\hat{\beta}\right) = \sigma^2 (X^\top X)^{-1}$ 只有一个未知σ^2，代入σ^2的估计即得到

$$F = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (X^\top X/\hat{\sigma}^2)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim_{H_0} F_{p,n-p}$$

回忆第8讲推论4. 假设$W \sim W_p(m)$，$\mathbf{z} \sim N_p(\mathbf{0}, I_p), m \geq p$,
假设$W \perp\!\!\!\perp \mathbf{z}$，则$\frac{\mathbf{z}^\top \mathbf{z}}{\mathbf{z}^\top W^{-1} \mathbf{z}} \sim \chi^2_{m-p+1}$ (且与$\mathbf{z}$独立)

对于非标准的Wishart和正态分布，我们有

引理1. 假设$W \sim W_p(m, \Sigma), \mathbf{z} \sim N_p(\mathbf{0}, \Sigma)$，$W \perp\!\!\!\perp \mathbf{z}, m \geq p$，$\Sigma > 0$, 则

$$\frac{\mathbf{z}^\top \Sigma^{-1} \mathbf{z}}{\mathbf{z}^\top W^{-1} \mathbf{z}} \sim \chi^2_{m-p+1},$$

且$\frac{\mathbf{z}^\top \Sigma^{-1} \mathbf{z}}{\mathbf{z}^\top W^{-1} \mathbf{z}}$与$\mathbf{z}$独立。

证：令$\tilde{\mathbf{z}} = \Sigma^{-1/2} \mathbf{z} \sim N_p(\mathbf{0}, I_p)$，$\widetilde{W} = \Sigma^{-1/2} W \Sigma^{-1/2} \sim W_p(m, I_p)$,
则由第8讲推论4，$\frac{\mathbf{z}^\top \Sigma^{-1} \mathbf{z}}{\mathbf{z}^\top W^{-1} \mathbf{z}} = \frac{\tilde{\mathbf{z}}^\top \tilde{\mathbf{z}}}{\tilde{\mathbf{z}}^\top \widetilde{W}^{-1} \tilde{\mathbf{z}}} \sim \chi^2_{m-p+1}$。

定理1. 假设$W \sim W_p(m, \Sigma)$，$\mathbf{z} \sim N_p(\mathbf{0}, \Sigma), \Sigma > 0, m \geq p$, 假设$W \perp\!\!\!\perp \mathbf{z}$,
则$\mathbf{z}^\top W^{-1} \mathbf{z} \sim \frac{p}{m-p+1} F_{p, m-p+1}$

证明： 由引理1， $\frac{\mathbf{z}^\top\Sigma^{-1}\mathbf{z}}{\mathbf{z}^\top W^{-1}\mathbf{z}}\sim\chi^2_{m-p+1}$,与$\mathbf{z}$独立, 而 $\mathbf{z}^\top\Sigma^{-1}\mathbf{z}\sim\chi^2_p$,
由F分布的定义

$$\frac{m-p+1}{p}\mathbf{z}^\top W^{-1}\mathbf{z}=\frac{\mathbf{z}^\top\Sigma^{-1}\mathbf{z}/p}{\frac{\mathbf{z}^\top\Sigma^{-1}\mathbf{z}}{\mathbf{z}^\top W^{-1}\mathbf{z}}/(m-p+1)}\sim F_{p,m-p+1}.$$

定理2. 假设$\mathbf{x}_1,\ldots,\mathbf{x}_n\ iid\ \sim N_p(\boldsymbol{\mu},\Sigma), n>p, \Sigma>0$, 则

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

证明： $\bar{\mathbf{x}}\sim N_p(\boldsymbol{\mu},\Sigma/n)\Rightarrow\mathbf{z}\triangleq\sqrt{n}(\bar{\mathbf{x}}-\boldsymbol{\mu})\sim N_p(\mathbf{0},\Sigma)$,

而$W\triangleq(n-1)S\sim W_p(n-1,\Sigma)$，

由定理1，

$$\mathbf{z}^\top W^{-1}\mathbf{z}=\frac{1}{n-1}n(\bar{\mathbf{x}}-\boldsymbol{\mu})^\top S^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})\sim\frac{p}{n-p}F_{p,n-p}$$

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

由定理1，我们可以构造均值的置信椭球如下

置信椭球

假设$\mathbf{x}_1, \ldots, \mathbf{x}_n$ $iid \sim N_p(\boldsymbol{\mu}, \Sigma), n > p, \Sigma > 0$，则$\boldsymbol{\mu}$的置信水平$1-\alpha$的置信椭球

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

向量$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$椭球区域不容易解释，通常人们更习惯于使用同时置信区间，即$\mu_1, \ldots, \mu_p$每个参数的置信区间。

Bonferroni 同时置信区间

由于$x_{1k}, \ldots, x_{nk}$ $iid \sim N_1(\mu_k, \sigma_{kk})$，样本均值$\bar{x}_k$（$\bar{\mathbf{x}}$的$k$分量），样本方差$s_{kk}$（$S$的$(k,k)$元）

$$\Sigma = (\sigma_{ij})$$
$$S = (s_{ij})$$

$$\frac{\sqrt{n}(\bar{x}_k - \mu_k)}{s_{kk}} \sim t_{n-1}$$

μ_k的置信水平$1-\alpha/p$的置信区间

$$\left\{\mu_k \in R^1: \ \left|\frac{\sqrt{n}(\bar{x}_k - \mu_k)}{s_{kk}}\right| \le t_{n-1}(\alpha/p)\right\}$$

所有这些置信区间称为$\mu_1, \ldots, \mu_p$的Bonferroni同时置信区间(置信水平$1-\alpha$)。

Hotelling T^2 检验 (单总体)

假设$\mathbf{x}_1, \ldots, \mathbf{x}_n$ $iid \sim N_p(\boldsymbol{\mu}, \Sigma), n > p, \Sigma > 0$. 零假设$H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$ ($\boldsymbol{\mu}_0$已知)的Hotelling T^2 检验统计量

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

当 $T^2 > \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$时否定$H_0$（水平α）。

注1：一元情形$(p = 1)$, T^2是student's t的平方

$$T^2 = [t]^2 \overset{H_0}{\sim} F_{1,n-1}, \quad t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s} \overset{H_0}{\sim} t_{n-1}$$

注2：单个正态总体的Hotelling T^2检验主要应用于成对数据。 假设第 i 对(pair)测量$(\mathbf{x}_{i1}, \mathbf{x}_{i2}), i = 1, \ldots, n$, $\boldsymbol{\delta} = E(\mathbf{x}_{i2}) - E(\mathbf{x}_{i1})$, 零假设

$$H_0: \boldsymbol{\delta} = \mathbf{0}$$

若假设$\mathbf{d}_i = \mathbf{x}_{i2} - \mathbf{x}_{i1}, i = 1, \ldots, n$ $iid \sim N_p(\boldsymbol{\delta}, \Sigma)$, 成对Hotelling T^2检验：

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

例6.1 (Johnson&Wichern). 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个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$2\mu m$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($\mathbf{d}_j, j = 1, \ldots, 11$)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

$\mathbf{d}_j$, $j = 1, \ldots, 11$, 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：

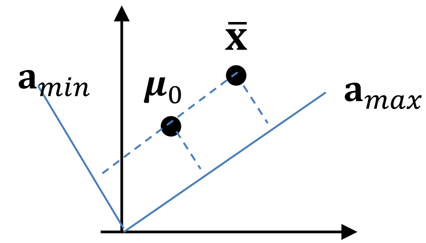
$$T^2 = 11[-9.36, \quad 13.27] \begin{bmatrix} .0055 & -.0012 \\ -.0012 & .0026 \end{bmatrix} \begin{bmatrix} -9.36 \\ 13.27 \end{bmatrix} = 13.6 \ > 9.47 = \frac{p(n-1)}{n-p} F_{p,n-p}(\alpha)$$

在0.05水平下拒绝原假设, $\text{pvalue} = 0.021$.

多元检验与一元检验

多元假设检验问题能否转化/降维为一元？我们下面考虑单个正态总体的均值检验问题（两总体以及其它多元检验都类似）。假设$\mathbf{x}_1,\dots,\mathbf{x}_n\ iid \sim N_p(\boldsymbol{\mu},\Sigma)$，零假设$H_0:\boldsymbol{\mu}=\boldsymbol{\mu}_0$ （$\boldsymbol{\mu}_0$已知)。

考察样本均值$\bar{\mathbf{x}}$是否远离$\boldsymbol{\mu}_0$。作为p维空间的两个点，在某些方向($\mathbf{a}_{max}$)上$\bar{\mathbf{x}}$远离$\boldsymbol{\mu}_0$，而在另外一些方向($\mathbf{a}_{min}$)上两者不容易区分。



取常数向量$\mathbf{a}\in S^{p-1}$, 在 $\mathbf{a}$方向投影得投影坐标

$$\mathbf{a}^\top\mathbf{x}_1,\dots,\mathbf{a}^\top\mathbf{x}_n\ iid \sim N_1(\mathbf{a}^\top\boldsymbol{\mu},\mathbf{a}^\top\Sigma\mathbf{a})$$

基于这些一元正态的样本，$H_{0\mathbf{a}}:\mathbf{a}^\top\boldsymbol{\mu}=\mathbf{a}^\top\boldsymbol{\mu}_0$的t检验

$$t(\mathbf{a})=\frac{\sqrt{n}(\mathbf{a}^\top\bar{\mathbf{x}}-\mathbf{a}^\top\boldsymbol{\mu}_0)}{\sqrt{\mathbf{a}^\top S\mathbf{a}}}$$

H_0成立时, $t(\mathbf{a})\sim t_{n-1}$。当$|t(\mathbf{a})|>t_{n-1}(\alpha/2)$时否定$H_{0\mathbf{a}}$ (水平α)。

因为一般不知道哪个方向最容易区分$\bar{\mathbf{x}}$和$\boldsymbol{\mu}_0$，我们尝试所有方向并取最大的一个：$t_{max}=\max\limits_{\mathbf{a}\in S^{p-1}}|t(\mathbf{a})|$, 可以证明(作业)

$$t_{max}=\sqrt{n(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^\top S^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)},\quad t_{max}^2=\text{Hotelling's } T^2,$$

所以多元Hotelling's T^2检验⇔最大一元t检验。

注：基于完全相同的思想，可得Scheffe同时置信区间。

两样本Hotelling's T^2检验

考虑两个正态总体：

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

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

令$S_p = \big((n_1-1)S_1 + (n_2-1)S_2\big)/(n_1+n_2-2)$, 定义

$$T^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^\top S^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))$$

与单个正态总体的定理2类似，两正态问题有以下结论：

定理3. 假设$\Sigma > 0, n > p+1, T^2$如上定义， 则 $T^2 \sim \frac{(n-2)p}{n-p-1} F_{p,n-p-1}$。

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

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

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

由定理1， $\mathbf{z}^\top W^{-1}\mathbf{z} \sim \frac{p}{n-p-1}F_{p,n-p-1}, T^2 = (n-2)\,\mathbf{z}^\top W^{-1}\mathbf{z} \sim \frac{(n-2)p}{n-p-1}F_{p,n-p-1}$

定理3结果可用于构造置信域和构造两样本Hotelling's T^2检验。

置信域

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

$$\left\{\boldsymbol{\delta}: \ \frac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta})^\top S^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta}) \le \frac{(n-2)p}{n-p-1} F_{p,n-p-1}(\alpha)\right\}$$

Bonferoni同时置信区间与单正态情形类似构建（略）。

两样本Hotelling's T^2检验是应用最为广泛的 检验方法之一.

两样本 Hotelling T^2检验

零假设H_0: $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_0$($\boldsymbol{\delta}_0$已知)的Hotelling T^2 检验统计量

$$T^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta}_0)^\top S^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta}_0) \sim_{H_0} \frac{(n-2)p}{n-p-1} F_{p,n-p-1}$$

当 $T^2 > \frac{(n-2)p}{n-p-1} F_{p,n-p-1}(\alpha)$时否定$H_0$（水平α）。

Harold Hotelling (1895-1973) was an American 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.