均值向量的推断

张伟平

zwp@ustc.edu.cn

Office: 东区管理科研楼 1006

Phone: 63600565

课件 http://staff.ustc.edu.cn/~zwp/

论坛 http://fisher.stat.ustc.edu.cn

简介

1.1 一元均值推断回顾 1
1.2 Hotelling's T^2 统计量 5
1.3 T^2 和似然比检验 11
1.3.1 似然比检验方法回顾 11
1.3.2 多元正态均值检验的似然比检验方法 . 13
1.3.3 协方差已知时均值检验 15
1.4 置信域 17
1.4.1 同时置信区间 19
1.4.2 大样本置信区间 25
1.5 样本存在缺失值时参数的估计 27

从本讲开始, 我们开始介绍一些多元推断技术. 多元分析中的主要特点是需要对 p 个变量同时进行分析. 首先我们看有关于多元均值的一些假设假设问题.

1.1 一元均值推断回顾

假设 $X_1, \ldots, X_n$ 为来自一元总体的简单随机样本, μ 为该总体的均值, 我们常常感兴趣此总体均值是否等于某个已知的常数 μ_0, 即为下述假设检验问题

$$H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu \neq \mu_0$$

距离方法 对此问题, 我们首先找到 μ 的一个相合估计, 然后构造其于 μ_0 之间偏差的某个距离, 当零假设成立时候, 该距离应该很小; 反之, 则应该比较大. 据此可以给出一个检验方法.

- 当样本来自正态总体分布 $N(\mu, \sigma^2)$ 时候, 则样本均值 $\bar{X}$ 和样本标准差 S 分别为 μ 和 σ 的相合估计
- 从而一个合适的检验统计量为

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

- 当 T 值较小时候, $\bar{X}$ 和 μ_0 比较靠近, 因此不能拒绝零假设. 此即为一样本 t 检验方法.
- 当 H_0 成立时, 检验统计量 T 的分布为自由度 $n-1$ 的 t 分布, 从而水平 α 检验的拒绝域为 $|T| \geq t_{n-1}(\alpha/2)$.

这个检验统计量是合理的, 其可以由似然比检验方法导出. 由于似然比

$$LR = \frac{\max_{\sigma>0} L(\mu_0, \sigma^2|\bar{x}, s^2)}{\max_{\mu\in R, \sigma>0} L(\mu, \sigma^2|\bar{x}, s^2)} = \left(1 + \frac{1}{n-1}T^2\right)^{-n/2}$$

- 当 H_0 成立时, LR 的值应靠近 1; 而当 H_1 成立时候, LR 的值应该远小于 1
- 从而似然比检验方法得到的拒绝域有形式 $LR \leq c$, 其中 c 为常数
- 显然, 似然比检验的拒绝域等价于 $|T| \geq c_0$, 在水平 α 下, $c_0 = t_{n-1}(\alpha/2)$.

因此, T 是一个合理的检验统计量. 另一方面,

- 注意 $|T|$ 较大时拒绝 H_0 等价于统计距离

$$T^2 = \frac{(\bar{X} - \mu_0)^2}{S^2/n} = n(\bar{X} - \mu_0)'(S^2)^{-1}(\bar{X} - \mu_0)$$

  太大时拒绝 H_0, 即拒绝域为 $T^2 \geq t^2_{n-1}(\alpha/2) = F_{1,n-1}(\alpha)$.
- 当不能拒绝 H_0 时候, 我们得出 μ_0 距离 $\bar{X}$ 较近 (在 $\bar{X}$ 的标准差单位下), 因此 μ_0 是 μ 的一个合理可能值.

- μ 的合理可能值集包含在 μ 的 $100(1-\alpha)\%$ 置信区间中:

$$\bar{X} - t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}}$$

- 这个置信区间包含了零假设 $H_0 : \mu = \mu_0$ 的所有在水平 α 下不能被拒绝的可能点 μ_0
- 在样本数据被收集前, 此区间为随机区间, 其包含 μ 的概率为 $1-\alpha$
- 当带入具体样本数据后, 此区间为一固定区间, 其要么包含 μ, 要么不包含 μ

当方差 σ^2 已知时候, 类似可知检验统计量为

$$T^2 = \frac{(\bar{X} - \mu_0)^2}{\sigma^2/n} = n(\bar{X} - \mu_0)'(\sigma^2)^{-1}(\bar{X} - \mu_0) \overset{H_0}{\sim} \chi_1^2$$

1.2 Hotelling's T^2 统计量

- 现在考虑 p 维向量 μ_0 是否为 p 维总体均值 μ 的一个合理值, 即

$$H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu \neq \mu_0$$

- 假设样本 $X_1, \ldots, X_n$ 为来自均值向量为 μ, 协方差矩阵为 Σ 的某个多元总体, $\bar{\mathbf{x}}$ 和 S 分别为样本均值向量和样本协方差矩阵

- 对该检验问题, 直观上可将一元场合时的距离推广到 p 元场合

$$T^2 = (\bar{\mathbf{x}} - \mu_0)' \left(\frac{1}{n} S\right)^{-1} (\bar{\mathbf{x}} - \mu_0)$$

- 则当 T^2 的值过大时候, 我们拒绝零假设 H_0, 即拒绝域有形式 $T^2 \geq c$, 这里 c 为待定常数

- 由于在零假设下, 当 $n \to \infty$ 时候 $T^2 \to \chi_p^2$, 因此当取 $c = \chi_p^2(\alpha)$ 时我们得到一个渐近水平 α 检验:

$$T^2 \geq \chi_p^2(\alpha) \text{时候拒绝} H_0$$

该假设无需假定总体分布为多元正态分布.

- 当总体分布为多元正态分布 $N_p(\mu, \Sigma)$ 时候, 我们可以得到 T^2 的精确分布. 因此可以得到一个精确的水平 α 检验. 注意到可以找到正交变换使得

$$\sqrt{n}(\bar{\mathbf{x}} - \mu) = z_n, \quad (n-1)S = \sum_{i=1}^{n-1} z_i z_i'$$

其中 $z_1, \ldots, z_n$ 相互独立且同分布于 $N_p(0, \Sigma)$, 因此当 H_0 成立时候

$$T^2 \stackrel{d}{=} z_n' \Big(\frac{1}{n-1} \sum_{i=1}^{n-1} z_i z_i' \Big)^{-1} z_n$$

定理 1. 设样本 $X_1, \ldots, X_n i.i.d \sim N_p(\mu_0, \Sigma)$, 则

$$T^2 = n(\bar{\mathbf{x}} - \mu_0)' S^{-1} (\bar{\mathbf{x}} - \mu_0) \sim \frac{(n-1)p}{n-p} F_{p,n-p}$$

其中 $\bar{\mathbf{x}}$ 和 S 分别表示样本均值和样本协方差矩阵.

证明. 由前面知

$$T^2 \stackrel{d}{=} (n-1) z_n' \Big(\sum_{i=1}^{n-1} z_i z_i' \Big)^{-1} z_n := (n-1) z_n' B^{-1} z_n$$

其中 $z_1, \ldots, z_n i.i.d \sim N_p(0, \Sigma)$. 不难看出, Σ 可以不妨设为 I_p. 取正交矩阵 Q, 其第一行为 $z_n'/\|z_n\|$, 其他行任意. 则 Q 为一随机矩阵. 由正交矩阵的性质易知

$$Qz_n = (\|z_n\|, 0, \ldots, 0)'$$

注意到 $z_n' B^{-1} z_n = (Qz_n)' \Big(\sum_{i=1}^{n-1} (Qz_i)(Qz_i)' \Big)^{-1} (Qz_n)$, 若记 $Y_i =$

$Qz_i, i = 1, \ldots, n$, 以及 $U = \sum_{i=1}^{n-1}(Qz_i)(Qz_i)'$ 则有

$$T^2/(n-1) \stackrel{d}{=} Y_n' U^{-1} Y_n = \|Y_n\|^2 U^{11} = \frac{\|Y_n\|^2}{u_{11\cdot 2}}$$

其中 $1/U^{11} = u_{11\cdot 2} = u_{11} - u_{12}' U_{22}^{-1} u_{21}$.

由于给定 Q 时候, Y_n 和 U 条件独立, 而 $\|Y_n\|^2 = \|Qz_n\|^2 \sim \chi_p^2$ 与 Q 无关. 于是只需证明 $u_{11\cdot 2} \sim \chi_{n-p}^2$, 则可得 Y_n 和 $u_{11\cdot 2}$ 相互独立, 最后由 F 分布定义立得 T^2 的分布.

事实上, 记 $[Qz_1, \ldots, Qz_{n-1}] = (z_{(1)}^*, Z_{(2)}^*)'$, ${z^*}'_{(1)}$ 表示第一行. 则

$$U = \sum_{i=1}^{n-1} z_i^* {z^*}'_i = \begin{pmatrix} {z^*}'_{(1)} \\ {Z^*}'_{(2)} \end{pmatrix} ({z^*}_{(1)}, {Z^*}_{(2)}) = \begin{pmatrix} {z^*}'_{(1)} {z^*}_{(1)} & {z^*}'_{(1)} {Z^*}_{(2)} \\ {Z^*}'_{(2)} {z^*}_{(1)} & {Z^*}'_{(2)} {Z^*}_{(2)} \end{pmatrix}$$

从而

$$u_{11\cdot 2} \stackrel{d}{=} {z^*}'_{(1)} (I_{n-1} - {Z^*}_{(2)} ({Z^*}'_{(2)} {Z^*}_{(2)})^{-1} {Z^*}'_{(2)}) {z^*}_{(1)}$$

在给定 Q 时候, 注意到 $I_{n-1} - {Z^*}_{(2)}({Z^*}'_{(2)}{Z^*}_{(2)})^{-1}{Z^*}'_{(2)}$ 为幂等阵, 其秩为 $n-p$. 而 $z^*_{(1)}$ 的分量为 $i.i.d$ 标准正态随机变量, 因此给定 ${Z^*}_{(2)}$ 时候 $u_{11\cdot 2} \overset{d}{=} \chi^2_{n-p}$ 其分布与 ${Z^*}_{(2)}$ 和 Q 无关.

综上即得

$$\frac{(n-p)T^2}{(n-1)p} = \frac{\|z_n\|^2/p}{u_{11\cdot 2}/(n-p)} \sim F_{p,n-p}.$$

□

- 当总体分布为多元正态分布 $N_p(\mu, \Sigma)$ 时候,

$$\begin{aligned} &P\Big(\frac{(n-p)T^2}{(n-1)p} \geq F_{p,n-p}(\alpha)\Big) \\ &= P\Big(n(\bar{\mathbf{x}}-\mu)'S^{-1}(\bar{\mathbf{x}}-\mu) \geq \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\Big) = \alpha, \end{aligned}$$

于是一个精确的水平 α 检验的拒绝域为 $T^2 \geq \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)$.

T^2 的线性变换不变性

对 X 测量单位的如下线性变换

$$Y_{p\times 1} = C_{p\times p}X_{p\times 1} + d_{p\times 1}, \quad C\text{非奇异}$$

Hotelling's T^2 统计量具有不变性. 事实上, 若观测到样本 $X_1, \ldots, X_n$, 则

$$\bar{\mathbf{y}} = C\bar{\mathbf{x}} + d, S_y = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{\mathbf{y}})(Y_i - \bar{\mathbf{y}})' = CS_xC'$$

而 $\mu_{y,0} = C\mu_0 + d$, 因此

$$\begin{aligned} T_y^2 &= n(\bar{\mathbf{y}} - \mu_{y,0})'S_y^{-1}(\bar{\mathbf{y}} - \mu_{y,0}) \\ &= n(\bar{\mathbf{x}} - \mu_0)'C'(C')^{-1}S_x^{-1}C^{-1}C(\bar{\mathbf{x}} - \mu_0) = T_x^2 \end{aligned}$$

1.3 T^2 和似然比检验

当总体分布为多元正态分布 $N_p(\mu, \Sigma)$ 时, 显然 $\bar{\mathbf{x}}, S$ 为 μ, Σ 的充分完备统计量. 因此, 一个优良的检验统计量应该是 $\bar{\mathbf{x}}, S$ 的函数. 下面证明可以由似然比检验方法导出 Hotelling's T^2 检验统计量是一个优良的检验统计量.

1.3.1 似然比检验方法回顾

假设 θ 为总体感兴趣的参数向量, 其取值空间为 Θ. 样本 $X_1, \ldots, X_n$ 的联合概率函数为 $f_\theta(x_1, \ldots, x_n)$(联合密度函数或者联合分布律), 则似然函数为 $L(\theta) = f_\theta(x_1, \ldots, x_n)$. 注意似然函数在参数真值处达到极大, 因此对假设检验问题

$$H_0 : \theta \in \Theta_0 \leftrightarrow H_i : \theta \notin \Theta_0$$

一个自然合理的检验为: 当

$$LRT = \frac{\max\limits_{\theta\in\Theta_0} L(\theta)}{\max\limits_{\theta\in\Theta} L(\theta)} < c$$

时拒绝 H_0, 其中 c 为待定常数.

在给定检验水平 α 下, 为确定常数 c, 我们需要使用似然比检验统计量 LRT 在 H_0 下的分布. 若该分布不可得, 则可以考虑其近似分布:

引理 1. *当样本量 $n \to \infty$ 时, 在 H_0 下*

$$-2log(LRT) \xrightarrow{d} \chi^2_{df}$$

其中 $df = dim(\Theta) - \dim(\Theta_0)$.

于是假设 $H_0 \leftrightarrow H_1$ 的一个近似水平 α 检验拒绝域为 $-2log(LRT) > \chi^2_{df}(\alpha)$.

1.3.2 多元正态均值检验的似然比检验方法

当样本 $X_1, \ldots, X_n i.i.d \sim N_p(\mu, \Sigma)$ 时候, $\theta = (\mu, \Sigma)$, 似然函数为

$$L(\theta, \Sigma) = (2\pi)^{-n/2}|\Sigma|^{-n/2} exp\Big[-\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)'\Sigma^{-1}(X_i - \mu)\Big]$$

注意到假设检验问题

$$H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu \neq \mu_0$$

等价于

$$H_0 : \mu = \mu_0, \Sigma > 0 \leftrightarrow H_1 : \mu \neq \mu_0, \Sigma > 0$$

于是 $\Theta_0 = (\mu_0, \Sigma)$, $\Theta = (\mu, \Sigma)$, 注意到 $\bar{\mathbf{x}}, \hat{\Sigma}$ 为 μ, Σ 的极大似然估计, 因此

$$\max_{\theta \in \Theta} L(\mu, \Sigma) = \max_{\mu, \Sigma > 0} L(\mu, \Sigma) = (2\pi)^{-n/2}|\hat{\Sigma}|^{-n/2} e^{-np/2}$$

而

$$\max_{\theta\in\Theta_0} L(\mu,\Sigma) = \max_{\mu_0,\Sigma>0} L(\mu,\Sigma) = (2\pi)^{-n/2}|\hat{\Sigma}_0|^{-n/2}e^{-np/2}$$

其中 $\hat{\Sigma}_0 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)(X_i-\mu_0)'$.

因此

$$LRT = \frac{\max_{\theta\in\Theta_0} L(\theta)}{\max_{\theta\in\Theta} L(\theta)} = \Big(\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_0|}\Big)^{n/2}$$

注意到

$$\hat{\Sigma}_0 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{\mathbf{x}})(X_i-\bar{\mathbf{x}})' + (\bar{\mathbf{x}}-\mu_0)(\bar{\mathbf{x}}-\mu_0)' = \hat{\Sigma} + CC'$$

其中 $C = \bar{\mathbf{x}} - \mu_0$, 以及事实 $|I+AB| = |I+BA|$ 易知

$$\begin{aligned} |\hat{\Sigma}_0| &= |\hat{\Sigma}|\cdot|I+\hat{\Sigma}^{-1/2}CC'\hat{\Sigma}^{-1/2}| = |\hat{\Sigma}|\cdot(1+C'\hat{\Sigma}^{-1}C) \\ &= |\hat{\Sigma}|\cdot(1+\frac{T^2}{n-1}). \end{aligned}$$

所以

$$LRT = \left(1 + \frac{T^2}{n-1}\right)^{-n/2}$$

从而拒绝域

$$LRT < c \Longleftrightarrow T^2 > c$$

这说明 Hotelling's T^2 统计量是一个良好的检验统计量.

由前面 T^2 在 H_0 下的分布知水平 α 检验的拒绝域为

$$T^2 > c = \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha).$$

1.3.3 协方差已知时均值检验

当协方差矩阵 Σ 已知时, 类似于 T^2 检验统计量, 此时假设检验问题

$$H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu \neq \mu_0$$

一个合理的检验统计量为

$$Z^2 = n(\bar{\mathbf{x}} - \mu_0)'\Sigma^{-1}(\bar{\mathbf{x}} - \mu_0)$$

- 若总体分布未知, 则由中心极限定理知

$$Z^2 \xrightarrow{d} \chi_p^2$$

因此可得一个渐近水平 α 检验拒绝域: $Z^2 > \chi_p^2(\alpha)$.

- 若设样本 $X_1, \ldots, X_n i.i.d \sim N_p(\mu, \Sigma)$ 时候, 则

$$Z^2 \sim \chi_p^2$$

因此可得一个精确的水平 α 检验拒绝域: $Z^2 > \chi_p^2(\alpha)$. 并可以证明该检验等价于似然比检验.

1.4 置信域

一元参数置信区间推广到多元场合即为置信域:

设参数 $\theta \in R^p$，则称由样本 $\mathbf{X}$ 所确定的 p 维区域 $R(\mathbf{X})$ 称为 θ 的 $1-\alpha$ 置信域，如果

$$P(\theta \in R(\mathbf{X})) = 1-\alpha.$$

Definition

对多元正态均值, 很容易得到一个 $1-\alpha$ 置信域:

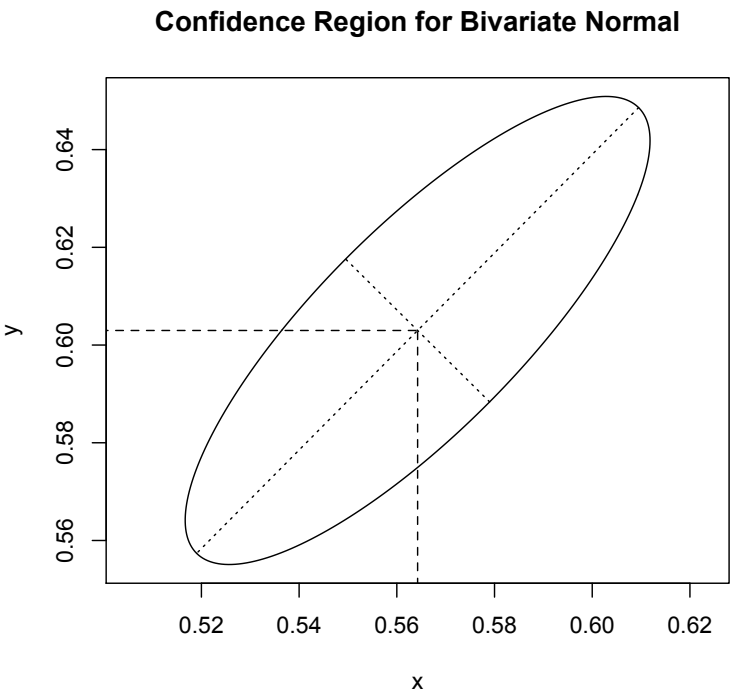
定理 2. 设样本 $X_1, \ldots, X_n i.i.d \sim N_p(\mu, \Sigma)$, 其中 $\mu, \Sigma > 0$ 均未知, $\bar{\mathbf{x}}, S$ 如前定义, 则 μ 的一个 $1-\alpha$ 置信域为

$$\left\{ \mu \in R^p \middle| n(\bar{\mathbf{x}}-\mu)' S^{-1} (\bar{\mathbf{x}}-\mu) \leq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha) \right\}$$

显然, 置信域是以 $\bar{\mathbf{x}}$ 为中心, 椭球的轴为

$$\pm\sqrt{\lambda_i}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)}\ \mathbf{e}_i, \quad \text{其中} \quad S\mathbf{e}_i = \lambda_i\mathbf{e}_i, i = 1, 2, \ldots, p$$

例 课本例 5.3: 表 4.1 和 4.2 数据的 0.25 幂次后可以认为服从二元正态分布, 左图为其均值的 95% 置信域观测值.



1.4.1 同时置信区间

若设样本 $X_1, \ldots, X_n i.i.d \sim N_p(\mu, \Sigma)$, 则对所有的 $\mathbf{a} \neq 0 \in R^p$, $\mathbf{a}'\mu$ 的同时 $1-\alpha$ 置信区间如何得到?

- 对任意固定的 $\mathbf{a} \neq 0 \in R^p$, 在多元正态总体下, $Y_j = \mathbf{a}'X_j \sim N_1(\mathbf{a}'\mu, \mathbf{a}'\Sigma\mathbf{a}), j = 1, \ldots, n$
- 从而由

$$t = \frac{\bar{Y} - \mathbf{a}'\mu}{s_{\bar{y}}} = \frac{\mathbf{a}'\bar{\mathbf{x}} - \mathbf{a}'\mu}{\sqrt{\frac{1}{n}\mathbf{a}'S\mathbf{a}}} \sim t_{n-1}$$

可得 $\mathbf{a}'\mu$ 的 $1-\alpha$ 置信区间

$$\bar{Y} - t_{n-1}(\alpha/2)\frac{s_{\bar{y}}}{\sqrt{n}} \leq \mathbf{a}'\mu \leq \bar{Y} + t_{n-1}(\alpha/2)\frac{s_{\bar{y}}}{\sqrt{n}}$$

即为

$$\mathbf{a}'\bar{\mathbf{x}} - t_{n-1}(\alpha/2)\frac{\sqrt{\frac{1}{n}\mathbf{a}'S\mathbf{a}}}{\sqrt{n}} \leq \mathbf{a}'\mu \leq \mathbf{a}'\bar{\mathbf{x}} + t_{n-1}(\alpha/2)\frac{\sqrt{\frac{1}{n}\mathbf{a}'S\mathbf{a}}}{\sqrt{n}}$$

进一步

$$\left|\frac{\mathbf{a}'\bar{\mathbf{x}}-\mathbf{a}'\mu}{\sqrt{\frac{1}{n}\mathbf{a}'S\mathbf{a}}}\right| \leq t_{n-1}(\alpha/2)$$

- 从而, 如果对所有的 $\mathbf{a} \neq 0 \in R^p$, 如果

$$\max_{\mathbf{a}\neq 0\in R^p}\left|\frac{\mathbf{a}'\bar{\mathbf{x}}-\mathbf{a}'\mu}{\sqrt{\frac{1}{n}\mathbf{a}'S\mathbf{a}}}\right| \leq c$$

等价于

$$\max_{\mathbf{a}\neq 0\in R^p}\left[\frac{\mathbf{a}'\bar{\mathbf{x}}-\mathbf{a}'\mu}{\sqrt{\frac{1}{n}\mathbf{a}'S\mathbf{a}}}\right]^2 = \max_{\mathbf{a}\neq 0\in R^p}\frac{n\mathbf{a}'(\bar{\mathbf{x}}-\mu)(\bar{\mathbf{x}}-\mu)'\mathbf{a}}{\mathbf{a}'S\mathbf{a}} \leq c^2$$

则可以得到其同时置信区间.

- 由 $Cauchy-Schwarz$ 不等式知

$$\frac{n\mathbf{a}'(\mathbf{x}-\mu)(\mathbf{x}-\mu)'\mathbf{a}}{\mathbf{a}'S\mathbf{a}} \leq T^2 \sim \frac{p(n-1)}{(n-p)}F_{p,n-p}$$

且等号在 $\mathbf{a} = dS^{-1}(\bar{\mathbf{x}} - \mu)$ 时达到, 其中 d 为任意非零常数.

- 因此, 取 $c^2 = \frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)$, 则由

$$\{\frac{n\mathbf{a}'(\bar{\mathbf{x}} - \mu)(\bar{\mathbf{x}} - \mu)'\mathbf{a}}{\mathbf{a}'S\mathbf{a}} \leq c^2\} \supseteq \{T^2 \leq c^2\}$$

知

$$P\Big(\frac{n\mathbf{a}'(\bar{\mathbf{x}} - \mu)(\bar{\mathbf{x}} - \mu)'\mathbf{a}}{\mathbf{a}'S\mathbf{a}} \leq c^2\Big) \geq P(T^2 \leq c^2) = 1 - \alpha.$$

因此对所有的 $\mathbf{a} \neq 0 \in R^p$, $\mathbf{a}'\mu$ 的同时 $1 - \alpha$ 置信区间为

$$\left[\mathbf{a}'\bar{\mathbf{x}} - \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p}(\alpha)\mathbf{a}'S\mathbf{a}}, \quad \mathbf{a}'\bar{\mathbf{x}} + \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p}(\alpha)\mathbf{a}'S\mathbf{a}}\right].$$

- 由置信区间与假设检验的等价性, 以及假设 $H_0 : \mu = \mu_0$ 等价于 $H_0 : \mathbf{a}'mu = \mathbf{a}'mu_0, \forall \mathbf{a} \neq 0$, 可得假设 $H_0 : \mu = \mu_0$ 的水平 α 检验为 $T^2 > \frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)$.

μ 的分量 $\mu_1, \ldots, \mu_p$ 的同时 $1-\alpha$ 置信区间

当取 $\mathbf{a} = (1, 0, \ldots, 0)'$, $\mathbf{a} = (0, 1, 0, \ldots, 0)'$, 以此类推, $\mathbf{a} = (0, \ldots, 0, 1)'$ 时候, 可得下面 p 个区间同时成立的概率为 $1-\alpha$.

$$C_1 : \bar{x}_1 - \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p}(\alpha) s_{11}} \le \mu_1 \le \bar{x}_1 + \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p}(\alpha) s_{11}}$$

$$C_2 : \bar{x}_2 - \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p}(\alpha) s_{22}} \le \mu_2 \le \bar{x}_2 + \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p}(\alpha) s_{22}}$$

$$\vdots$$

$$C_p : \bar{x}_p - \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p}(\alpha) s_{pp}} \le \mu_p \le \bar{x}_p + \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p}(\alpha) s_{pp}}$$

即

$$P(\prod_{i=1}^{p} C_i) \ge 1 - \alpha.$$

多重比较的 Bonferroni 方法

注意到 μ_i 的边际 $1-\beta$ 置信区间:

$$B_i(\beta): \bar{x}_i - t_{n-1}(\beta/2)\frac{s_{ii}}{\sqrt{n}} \le \mu_i \le \bar{x}_i + t_{n-1}(\beta/2)\frac{s_{ii}}{\sqrt{n}} \qquad i = 1, \ldots, p$$

即 $P(B_i(\beta)) = 1-\beta, i = 1, \ldots, p.$

由

$$P\Big(\prod_{i=1}^{p} B_i\Big) = 1 - P\Big(\sum_{i=1}^{p} \bar{B}_i\Big) \ge 1 - \sum_{i=1}^{p} P(\bar{B}_i)$$

于是, 若取 $\beta = \alpha/p$, 则

$$P\Big(\prod_{i=1}^{p} B_i(\alpha/p)\Big) \ge 1 - \sum_{i=1}^{p} \frac{\alpha}{p} = 1 - \alpha.$$

从而得到 $\mu_1, \ldots, \mu_p$ 的同时 $1-\alpha$ 置信区间

$$B_1(\alpha/p), \ldots, B_p(\alpha/p)$$

μ 的分量差 $\mu_i - \mu_k (i \neq k)$ 的同时 $1-\alpha$ 置信区间

当取 $\mathbf{a} = (0, \ldots, a_i, 0, \ldots, 0, a_k, 0, \ldots, 0)'$, 其中 $a_i = 1, a_k = -1$ 时候, $\mathbf{a}'\mu = \mu_i - \mu_k$, $\mathbf{a}' S \mathbf{a} = s_{ii} - 2s_{ik} + s_{kk}$, 因此可得 $\mu_i - \mu_k (i \neq k)$ 的同时置信区间为

$$\bar{x}_i - \bar{x}_k \pm \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{ii} - 2s_{ik} + s_{kk}}{n}}$$

Bonferroni 同时置信区间

注意到 $\bar{x}_i - \bar{x}_k \sim N(\mu_i - \mu_k, \sigma^*_{ik}), \sigma^*_{ik} = \sigma_{ii} - 2\sigma_{ik} + \sigma_{kk}, i \neq k$ 以及 $(n-1)(s_{ii} - 2s_{ik} + s_{kk}))/\sigma^*_{ik} \sim \chi^2_{n-1}$, 且两者相互独立, 因此

$$\frac{\sqrt{n}(\bar{x}_i - \bar{x}_k - (\mu_i - \mu_k)}{\sqrt{s_{ii} - 2s_{ik} + s_{kk}}} \sim t_{n-1}$$

因此可得 $\mu_i - \mu_k (i \neq k)$ 的 Bonferroni 同时置信区间为

$$\left| \frac{\sqrt{n}(\bar{x}_i - \bar{x}_k - (\mu_i - \mu_k))}{\sqrt{s_{ii} - 2s_{ik} + s_{kk}}} \right| \leq t_{n-1}(\alpha/p), i \neq k$$

1.4.2 大样本置信区间

当样本量 n 很大时, 对均值向量的推断可以不假设多元正态成立. 由大样本理论知

$$P[n(\bar{\mathbf{x}}-\mu)'S^{-1}(\bar{\mathbf{x}}-\mu) \le \chi_p^2(\alpha)] \approx 1-\alpha$$

因此

定理 3. 设 $X_1,\ldots,X_n$ 为来自均值 μ, 协方差矩阵 Σ 的总体的简单随机样本, 其中 $\mu \in R^p, \Sigma_{p\times p} > 0$ 均未知, 令 $\bar{\mathbf{x}}, S$ 如前定义, 则当 $n-p$ 很大时候, μ 的一个渐近 $1-\alpha$ 置信域为

$$\{\mu \in R^p | n(\bar{\mathbf{x}}-\mu)'S^{-1}(\bar{\mathbf{x}}-\mu) \le \chi_p^2(\alpha)\}$$

由置信区间和假设检验的等价性, 此时假设 $H_0: \mu = \mu_0 \leftrightarrow H_1: \mu \ne \mu_0$ 的一个渐近水平 α 检验为

当 $n(\bar{\mathbf{x}}-\mu_0)'S^{-1}(\bar{\mathbf{x}}-\mu_0) > \chi_p^2(\alpha)$时拒绝$H_0$.

定理 4. 设 $X_1, \ldots, X_n$ 为来自均值 μ, 协方差矩阵 Σ 的总体的简单随机样本, 其中 $\mu \in R^p, \Sigma_{p\times p} > 0$ 均未知, 令 $\bar{\mathbf{x}}, S$ 如前定义, 则当 $n-p$ 很大时候, $\mathbf{a}'\mu$ 的一个渐近 $1-\alpha$ 置信域为

$$\mathbf{a}'\bar{\mathbf{x}} - \sqrt{\chi_p^2(\alpha)}\sqrt{\frac{\mathbf{a}'S\mathbf{a}}{n}} \le \mathbf{a}'\mu \le \mathbf{a}'\bar{\mathbf{x}} + \sqrt{\chi_p^2(\alpha)}\sqrt{\frac{\mathbf{a}'S\mathbf{a}}{n}}$$

其中 $\mathbf{a}$ 为任意非零 p 维向量.

因此, $\mu_1, \mu_2, \ldots, \mu_p$ 的同时渐近 $1-\alpha$ 置信区间为

$$\bar{x}_i \pm \sqrt{\chi_p^2(\alpha)}\sqrt{\frac{s_{ii}}{n}}, i = 1, \ldots, p.$$

1.5 样本存在缺失值时参数的估计

在对多元变量进行观测时, 常常会遇到其某些分量缺失 (missing) 的情况. 处理这类不完全观测值问题的最好方法在很大程度上依赖于试验本身. 如果丢弃有缺失值的样本, 仅使用完全观测到的样本进行统计推断, 其有效性与缺失发生的机制有关.

- **MCAR: Missing Completely At Random**. 缺失与否与其他所有量都无关, 即 $P(Missing|Obervered, Unobserved) = P(Missing)$
- **MAR: Missing At Random**. 缺失与否仅依赖于观测到的量, 即 $P(Missing|Obervered, Unobserved) = P(Missing|Observed)$
- **MNAR: Missing Not At Random**. 缺失与否与观测到和未观察到的量均有关.

对 MAR, 仅依赖观测到的完全样本进行推断将导致估计量严重的偏倚. 但当数据的缺失机制是 MCAR, 则可以有效的处理这种问题.

对缺失数据的处理已经有一些有效的方法, 我们这里介绍一种从不完全数据出发计算极大似然估计的一般方法: EM 算法.

EM algorithm(Expectation-Maximization)

记观测到的量为 X, 缺失量为 Z, 完全数据为 $Y = (X, Z)$, 待估参数为 θ. 则我们有的信息是观测数据的对数似然 $l(\theta|x) = logL(\theta|x)$, 最大化此似然或者说是求 θ 的极大似然估计是我们的目标. EM 算法通过迭代方式来寻求最大化 $l(\theta|x)$ 的解.

记 $\theta^{(t)}$ 表示在第 t 次迭代后对数似然的最大值点, $t = 0, 1, 2, \cdots$. 定义 $Q(\theta|\theta^{(t)})$ 为在观测到 $X = x$, 以及在参数 $\theta = \theta^{(t)}$ 的条件下完

全数据的对数似然函数的期望, 此期望是对 $f_{Z|X}(z|x,\theta^{(t)})$ 计算. 即

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E\{l(\theta|Y)|x,\theta^{(t)}\} \\ &= E\{logf(x,Z|\theta)|x,\theta^{(t)}\} \\ &= \int logf(x,z|\theta)f(z|x,\theta^{(t)})dz \end{aligned}$$

最后一式强调一旦我们给定了 $X=x$, Z 就是 Y 唯一的随机部分.

EM 算法从 $\theta^{(0)}$ 开始, 然后在两步之间交替. 该算法概括如下:

(1) E 步: 计算 $Q(\theta|\theta^{(t)}) = E[l(\theta|Y)|x,\theta^{(t)}]$.
(2) M 步: 关于 θ 最大化 $Q(\theta|\theta^{(t)})$, 并记 $\theta^{(t+1)}$ 表示此时的最大值点.
(3) 返回到 E 步, 直至收敛准则达到.

假设样本 $X_1, \ldots, X_n i.i.d \sim N_p(\mu, \Sigma)$, $\mu, \Sigma > 0$ 未知. $X_i = (X_i^{(1)'}, X_i^{(2)'})'$, 其中 $X_i^{(1)}$ 为缺失的分量, $X_i^{(2)}$ 为观测到的分量, 相应样本值记为 $, x_i^{(2)}, i = 1, 2, \ldots, n$. 则完全数据下的对数似然函数

$$l(\mu, \Sigma) \propto -\frac{n}{2} log|\Sigma| - \frac{1}{2} \sum_{i=1}^{n} tr[\Sigma^{-1}(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)']$$

注意到 $X_i^{(1)}|x_i^{(2)}, \mu^{(t)}, \Sigma^{(t)} \sim N(\mu_{11\cdot 2}^{(t)}, \Sigma_{11\cdot 2}^{(t)})$, $\mu_{11\cdot 2}^{(t)} = \mu_1^{(t)} + \Sigma_{12}^{(t)} \Sigma_{22}^{-1(t)} (x_i^{(2)} - \mu_2^{(t)})$, 因此

$$E[(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)'|x_i^{(2)}, \mu^{(t)}, \Sigma^{(t)}] = A_i^{(t)} + (z_i^{(t)} - \mu)(z_i^{(t)} - \mu)'$$

其中 $z_i^{(t)} = (\mu_{11\cdot 2}^{(t)'}, x_i^{(2)'})'$. 以及

$$A_i^{(t)} = \begin{pmatrix} \Sigma_{i,11\cdot 2}^{(t)} & 0 \\ 0 & 0 \end{pmatrix}$$

于是

$$\begin{aligned} Q(\mu,\Sigma) &= E[l(\mu,\Sigma)|x_1^{(2)},\ldots,x_n^{(2)},\mu^{(t)},\Sigma^{(t)}] \\ &\approx -\frac{n}{2}log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n} tr[\Sigma^{-1}[A_i^{(t)} + (z_i^{(t)} - \mu)(z_i^{(t)} - \mu)']] \end{aligned}$$

因此最大化 $Q(\mu,\Sigma)$ 得到

$$\mu^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} z_i^{(t)},$$

$$\Sigma^{(t+1)} = \frac{1}{n}[\sum_{i=1}^{n}(z_i^{(t)} - \mu^{(t+1)})(z_i^{(t)} - \mu^{(t+1)})' + A_i^{(t)}].$$