第三章 多元正态分布均值向量和协差阵的检验

第一节 引言

第二节 均值向量的检验

第三节 协差阵的检验

第一节 引言

- 在单一变量的统计分析中，已经给出了正态总体N（μ，σ^2）的均值μ和方差σ^2的各种检验。对于多变量的正态总体N_p（μ，$\sum$），各种实际问题同样要求对μ和$\sum$进行统计推断。
  - 例如，我们要考察全国各省、自治区和直辖市的社会经济发展状况，与全国平均水平相比较有无显著性差异等，就涉及到多元正态总体均值向量的检验问题等。

- 本章类似单一变量统计分析中的各种均值和方差的检验，相应地给出多元统计分析中的各种均值向量和协差阵的检验。

- 其基本思想和步骤均可归纳为：

  第一，提出待检验的假设H_0和H_1；

  第二，给出检验的统计量及其服从的分布；

  第三，给定检验水平，查统计量的分布表，确定相应的临界值，从而得到否定域；

  第四，根据样本观测值计算出统计量的值，看是否落入否定域中，以便对待判假设做出决策（拒绝或接受）。
- 在检验的过程中，关键在于对不同的检验给出不同的统计量，而有关统计量的给出大多用似然比方法得到。由于多变量问题的复杂性，本章只侧重于解释选取统计量的合理性，而不给出推导过程，最后给出几个实例。
- 为了更好的说明检验过程中统计量的分布，本章还要介绍HotellingT^2分布和Wilks分布的定义。

第二节 均值向量的检验

- 一 单一变量检验的回顾及Hotelling T^2分布
- 二 一个正态总体均值向量的检验
- 三 两个正态总体均值向量的检验
- 四 多个正态总体均值向量的检验

一、单一变量检验的回顾及Hotelling T^2分布

- 为了对多元正态总体均值向量作检验，首先需要给出HotellingT^2分布的定义。
- 在单一变量的检验问题中，设 $X_1, X_2, \cdots, X_n$ 来自总体 $N(\mu, \sigma^2)$ 的样本，我们要检验假设

$$H_0: \mu = \mu_0; \quad H_1: \mu \neq \mu_0$$

- 当 σ^2 已知时，用统计量

$$z = \frac{(\bar{X} - \mu_0)}{\sigma}\sqrt{n} \qquad （3.1）$$

其中，$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 为样本均值。当假设成立时，统计量 z 服从正态分布 $z \sim N(0,1)$，从而否定域为 $|z| > z_{\alpha/2}$，$z_{\alpha/2}$ 为 $N(0,1)$ 的上 $\alpha/2$ 分位点。

- 当σ^2未知时，用 $$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \quad (3.2)$$

  作为σ^2的估计量，用统计量：

  $$t=\frac{(\bar{X}-\mu_0)}{S}\sqrt{n} \quad (3.3)$$

  来做检验。当假设成立时，统计量t服从自由度为$n-1$的t分布，从而否定域为$|t|>t_{\alpha/2}(n-1)$，$t_{\alpha/2}(n-1)$为自由度为$n-1$的t分布上的$\alpha/2$分位点。

- 这里我们应该注意到，（3.3）式可以表示为

  $$t^2=\frac{n(\bar{X}-\mu)^2}{S^2}=n(\bar{X}-\mu)'(S^2)^{-1}(\bar{X}-\mu) \quad (3.4)$$

  对于多元变量而言，可以将t分布推广为下面将要介绍的 **Hotelling** T^2分布。

- **定义 3.1** 设 $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，$\boldsymbol{S} \sim W_p(n, \boldsymbol{\Sigma})$ 且 $\boldsymbol{X}$ 与 $\boldsymbol{S}$ 相互独立，$n \geq p$，则称统计量 $T^2 = n\boldsymbol{X}'\boldsymbol{S}^{-1}\boldsymbol{X}$ 的分布为非中心 **Hotelling** T^2 分布，记为 $T^2 \sim T^2(p, n, \boldsymbol{\mu})$。当 $\boldsymbol{\mu} = 0$ 时，称 T^2 服从（中心）**Hotelling** T^2 分布。记为 $T^2(p, n)$。

由于这一统计量的分布首先由 **Harold Hotelling** 提出来的，故称为 **Hotelling** T^2 分布，值得指出的是，我国著名统计学家许宝禄先生在 **1938** 年用不同方法也导出 T^2 分布的密度函数，因表达式很复杂，故略去。

- 在单一变量统计分析中，若统计量 $t \sim t(n-1)$ 分布，则 $t^2 \sim F(1,n-1)$ 分布，即把 t 分布的统计量转化为 F 统计量来处理，在多元统计分析中 T^2 统计量也具有类似的性质。

- 定理 3.1　若 $\boldsymbol{X} \sim N_p(\boldsymbol{0},\boldsymbol{\Sigma})$，$\boldsymbol{S} \sim W_p(n,\boldsymbol{\Sigma})$ 且 $\boldsymbol{X}$ 与 $\boldsymbol{S}$ 相互独立，令 $T^2 = n\boldsymbol{X}'\boldsymbol{S}^{-1}\boldsymbol{X}$，则

$$\frac{n-p+1}{np}T^2 \sim F(p,n-p+1) \qquad (3.5)$$

在我们后面所介绍的检验问题中，经常会用到这一性质。

二、一个正态总体 均值向量的检验

- 设 $\boldsymbol{X}_{(1)},\boldsymbol{X}_{(2)},\cdots,\boldsymbol{X}_{(n)}$ 是来自 p 维正态总体 $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 的样本，且 $\bar{\boldsymbol{X}}=\frac{1}{n}\sum_{\alpha=1}^{n}\boldsymbol{X}_{(\alpha)}$， $\boldsymbol{S}=\sum_{a=1}^{n}(\boldsymbol{X}_{(a)}-\bar{\boldsymbol{X}})(\boldsymbol{X}_{(a)}-\bar{\boldsymbol{X}})'$。

（一） 协差阵 $\boldsymbol{\Sigma}$ 已知时均值向量的检验

H_0：$\boldsymbol{\mu}=\boldsymbol{\mu}_0$（$\boldsymbol{\mu}_0$ 为已知向量）H_1：$\boldsymbol{\mu}\neq\boldsymbol{\mu}_0$

假设 H_0 成立，检验统计量为

$$T_0^2=n(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)\sim\chi^2(p) \qquad (3.6)$$

给定检验水平 α，查 χ^2 分布表使 $P\left\{T_0^2>\chi_\alpha^2\right\}=\alpha$，可确定出临界值 χ_α^2，再用样本值计算出 T_0^2，若 $T_0^2>\chi_\alpha^2$，则否定 H_0，否则接受 H_0。

- 这里要对统计量的选取做一些解释，为什么该统计量服从 $\chi^2(p)$ 分布。根据二次型分布定理知道，若 $\boldsymbol{X} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$，则 $\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \sim \chi^2(p)$。显然，

$$T_0^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)$$

$$= \sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}\sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) \triangleq \boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$$

其中，$\boldsymbol{Y} = \sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$，因此，

$$T_0^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) \sim \chi^2(p)。$$

（二）协差阵$\boldsymbol{\Sigma}$未知时均值向量的检验

H_0：$\boldsymbol{\mu}=\boldsymbol{\mu}_0$（$\boldsymbol{\mu}_0$为已知向量）$H_1$：$\boldsymbol{\mu}\neq\boldsymbol{\mu}_0$

假设H_0成立，检验统计量为

$$\frac{(n-1)-p+1}{(n-1)p}T^2 \sim F(p,n-p) \qquad (3.7)$$

其中，$T^2=(n-1)[\sqrt{n}(\bar{\mathbf{X}}-\boldsymbol{\mu}_0)'\mathbf{S}^{-1}\sqrt{n}(\bar{\mathbf{X}}-\boldsymbol{\mu}_0)]$

给定检验水平α，查F分布表，使$P\left\{\frac{n-p}{(n-1)p}T^2>F_\alpha\right\}=\alpha$，可确定出临界值$F_\alpha$，再用样本值计算出$T^2$，若$\frac{n-p}{(n-1)p}T^2>F_\alpha$，则否定$H_0$，否则接受$H_0$。

- 这里需要解释的是，当 Σ 未知时，自然想到要用样本协差阵 $\frac{1}{n-1}\mathbf{S}$ 取代替 Σ，因 $(n-1)\mathbf{S}^{-1}$ 是 Σ^{-1} 的无偏估计量，而样本离差阵

$$\mathbf{S}=\sum_{a=1}^{n}(\mathbf{X}_{(a)}-\bar{\mathbf{X}})(\mathbf{X}_{(a)}-\bar{\mathbf{X}})'\sim W_p(n-1,\Sigma)$$

$$\sqrt{n}(\bar{\mathbf{X}}-\boldsymbol{\mu}_0)\sim N_p(\mathbf{0},\Sigma)$$

由定义 3.1 知

$$T^2=(n-1)[\sqrt{n}(\bar{\mathbf{X}}-\boldsymbol{\mu}_0)'\mathbf{S}^{-1}\sqrt{n}(\bar{\mathbf{X}}-\boldsymbol{\mu}_0)]\sim T^2(p,n-p)$$

再根据 Hotelling T^2 分布的性质，所以

$$\frac{(n-1)-p+1}{(n-1)p}T^2\sim F(p,n-p)$$

在处理实际问题时，单一变量的检验和多变量检验可以联合使用，多元的检验具有概括和全面考察的特点，而一元的检验容易发现各变量之间的关系和差异，能给人们提供更多的统计分析信息。

三、两个正态总体均值向量的检验

（一）当协差阵相等时，两个正态总体均值向量的检验

设 $\mathbf{X}_{(a)} = (X_{a1}, X_{a2}, \cdots, X_{ap})'$，$a = 1,2,\cdots,n$，为来自 p 维正态总体 $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ 的容量为 n 的样本；$\mathbf{Y}_{(a)} = (Y_{a1}, Y_{a2}, \cdots, Y_{ap})'$，$a = 1,2,\cdots,m$，为来自 p 维正态总体 $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ 的容量为 m 的样本。两组样本相互独立，$n > p, m > p$，且 $\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_{(i)}$，$\bar{\mathbf{Y}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{Y}_{(i)}$。

1．针对有共同已知协差阵的情形

对假设

$$H_0:\ \boldsymbol{\mu}_1=\boldsymbol{\mu}_2 \qquad H_1:\ \boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2$$

进行检验。

对此问题，假设 H_0 成立时，所构造的检验统计量为

$$T_0^2=\frac{n\cdot m}{n+m}(\bar{\mathbf{X}}-\bar{\mathbf{Y}})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}}-\bar{\mathbf{Y}})\sim\chi^2(p) \qquad （3.8）$$

给出检验水平 α，查 $\chi^2(p)$ 分布表使 $P\left\{T_0^2>\chi_\alpha^2\right\}=\alpha$，可确定出临界值 χ_α^2，再用样本值计算出 T_0^2，若 $T_0^2>\chi_\alpha^2$，则否定 H_0，否则接受 H_0。

- 这里，我们应该注意到，在单一变量统计中进行均值相等检验所给出的统计量为

$$z=\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}}\sim N(0,1)$$

显然

$$z^2=\frac{(\bar{X}-\bar{Y})^2}{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}=\frac{n\cdot m}{(n+m)\sigma^2}(\bar{X}-\bar{Y})^2$$

$$=\frac{n\cdot m}{n+m}(\bar{X}-\bar{Y})'(\sigma^2)^{-1}(\bar{X}-\bar{Y})\sim\chi^2(1)$$

此式恰为上边统计量当 $p=1$ 时的情况，不难看出这里给出的检验统计量是单一变量检验情况的推广。

2．针对有共同的未知协差阵的情形

对假设

$$H_0:\ \boldsymbol{\mu}_1=\boldsymbol{\mu}_2 \qquad H_1:\ \boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2$$

进行检验。

对此问题，假设 H_0 成立时，所构造的检验统计量为

$$F=\frac{(n+m-2)-p+1}{(n+m-2)p}T^2\sim F(p,n+m-p-1)$$

（3.9）

其中，

$$T^2=(n+m-2)\left[\sqrt{\frac{n\cdot m}{n+m}}(\bar{\mathbf{X}}-\bar{\mathbf{Y}})\right]'\mathbf{S}^{-1}\left[\sqrt{\frac{n\cdot m}{n+m}}(\bar{\mathbf{X}}-\bar{\mathbf{Y}})\right]$$

$$\mathbf{S}=\mathbf{S}_x+\mathbf{S}_y$$

$$\mathbf{S}_x=\sum_{a=1}^{n}(\mathbf{X}_{(a)}-\bar{\mathbf{X}})(\mathbf{X}_{(a)}-\bar{\mathbf{X}})',\qquad \bar{\mathbf{X}}=(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2,\cdots,\bar{\mathbf{X}}_p)'$$

$$\mathbf{S}_y=\sum_{a=1}^{n}(\mathbf{Y}_{(a)}-\bar{\mathbf{Y}})(\mathbf{Y}_{(a)}-\bar{\mathbf{Y}})',\qquad \bar{\mathbf{Y}}=(\bar{\mathbf{Y}}_1,\bar{\mathbf{Y}}_2,\cdots,\bar{\mathbf{Y}}_p)'$$

给定检验水平α，查F分布表，使$p\left\{F>F_\alpha\right\}=\alpha$，可确定出临界值$F_\alpha$，再用样本值计算出$F$，若$F>F_\alpha$，则否定$H_0$，否则接受$H_0$。

- 这里我们需要解释的是，当两个总体的协差阵未知时，自然想到用每个总体的样本协差阵 $\frac{1}{n-1}\mathbf{S}_x$ 和 $\frac{1}{m-1}\mathbf{S}_y$ 去代替，而

$$\mathbf{S}_x=\sum_{a=1}^{n}(\mathbf{X}_{(a)}-\bar{\mathbf{X}})(\mathbf{X}_{(a)}-\bar{\mathbf{X}})'\sim W_p(n-1,\boldsymbol{\Sigma})$$

$$\mathbf{S}_y=\sum_{a=1}^{m}(\mathbf{Y}_{(a)}-\bar{\mathbf{Y}})(\mathbf{Y}_{(a)}-\bar{\mathbf{Y}})'\sim W_p(m-1,\boldsymbol{\Sigma})$$

从而 $\mathbf{S}=\mathbf{S}_x+\mathbf{S}_y\sim W_p(n+m-2,\boldsymbol{\Sigma})$。又由于

$$\sqrt{\frac{n\cdot m}{n+m}}(\bar{\mathbf{X}}-\bar{\mathbf{Y}})\sim N_p(\mathbf{0},\boldsymbol{\Sigma})$$

所以

$$\frac{(n+m-2)-p+1}{(n+m-2)p}T^2\sim F(p,n+m-p+1)$$

下述假设检验统计量的选取和前边统计量的选取思路是一样的，以下只提出待检验的假设，然后给出统计量及其分布，为节省篇幅，不做重复解释。

（二）协差阵不等时，两个正态总体均值向量的检验

设从两个总体$N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$和$N_p(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$中，分别抽取两个样本，即$\mathbf{X}_{(a)}=(\mathbf{X}_{a1},\mathbf{X}_{a2},\cdots,\mathbf{X}_{ap})'$，$a=1,2,\cdots,n$；$\mathbf{Y}_{(a)}=(\mathbf{Y}_{a1},\mathbf{Y}_{a2},\cdots,\mathbf{Y}_{ap})'$，$a=1,2,\cdots,m$，其容量分别为$n$和$m$，且两组样本相互独立，$n>p,m>p$，$\boldsymbol{\Sigma}_1>0$，$\boldsymbol{\Sigma}_2>0$。对假设

$$H_0\text{：}\ \boldsymbol{\mu}_1=\boldsymbol{\mu}_2 \qquad\qquad H_1\text{：}\ \boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2$$

进行检验。

1. 针对$n = m$的情形

令 $\mathbf{Z}_{(i)} = \mathbf{X}_{(i)} - \mathbf{Y}_{(i)}$ $\quad i = 1,2,\cdots,n$

$$\bar{\mathbf{Z}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_{(i)} = \bar{\mathbf{X}} - \bar{\mathbf{Y}}$$

$$\mathbf{S} = \sum_{i=1}^{n}(\mathbf{Z}_{(i)} - \bar{\mathbf{Z}})(\mathbf{Z}_{(i)} - \bar{\mathbf{Z}})'$$

$$= \sum_{i=1}^{n}(\mathbf{X}_{(i)} - \mathbf{Y}_{(i)} - \bar{\mathbf{X}} + \bar{\mathbf{Y}})(\mathbf{X}_{(i)} - \mathbf{Y}_{(i)} - \bar{\mathbf{X}} + \bar{\mathbf{Y}})'$$

假设H_0成立时，构造检验统计量为

$$F = \frac{(n-p)n}{p}\bar{\mathbf{Z}}'\mathbf{S}^{-1}\bar{\mathbf{Z}} \sim F(p, n-p) \qquad (3.10)$$

2. 针对$n \neq m$的情形

在此，我们不妨假设$n < m$，令

$$\mathbf{Z}_{(i)} = \mathbf{X}_{(i)} - \sqrt{\frac{n}{m}}\mathbf{Y}_{(i)} + \frac{1}{\sqrt{n \cdot m}}\sum_{j=1}^{n}\mathbf{Y}_{(i)} - \frac{1}{m}\sum_{j=1}^{n}\mathbf{Y}_{(i)}$$

$$i = 1,2,\cdots,n$$

$$\overline{Z} = \frac{1}{n}\sum_{i=1}^{n}Z_{(i)} = \overline{X} - \overline{Y}$$

$$\mathbf{S}=\sum_{i=1}^{n}(\mathbf{Z}_{(i)}-\bar{\mathbf{Z}})(\mathbf{Z}_{(i)}-\bar{\mathbf{Z}})'$$

$$=\sum_{i=1}^{n}\left[(\mathbf{X}_{(i)}-\bar{\mathbf{X}})-\sqrt{\frac{n}{m}}(\mathbf{Y}_{(i)}-\frac{1}{n}\sum_{j=1}^{n}\mathbf{Y}_{(j)})\right]$$

$$\cdot\left[(\mathbf{X}_{(i)}-\bar{\mathbf{X}})-\sqrt{\frac{n}{m}}(\mathbf{Y}_{(i)}-\frac{1}{n}\sum_{j=1}^{n}\mathbf{Y}_{(j)})\right]'$$

假设 H_0 成立时，构造检验统计量为

$$F=\frac{(n-p)n}{p}\bar{\mathbf{Z}}'\mathbf{S}^{-1}\bar{\mathbf{Z}}\sim F(p,n-p) \qquad (3.11)$$

四、多个正态总体均值向量的检验

- 解决多个正态总体均值向量的检验问题，实际上应用到多元方差分析的知识。多元方差分析是单因素方差分析直接的推广。为了容易理解多元方差分析方法，我们有必要先回顾单因素方差分析方法。

（一）单因素方差分析的基本思想及 Wilks 分布

设 k 个正态总体分别为 $N(\mu_1,\sigma^2)$，$\cdots$，$N(\mu_k,\sigma^2)$，从 k 个总体取 n_i 个独立样本如下：

$$X_1^{(1)},X_2^{(1)},\cdots,X_{n_1}^{(1)}$$
$$\cdots\cdots$$
$$X_1^{(k)},X_2^{(k)},\cdots,X_{n_k}^{(k)}$$

H_0：$\mu_1=\mu_2=\cdots=\mu_k$　　　H_1：至少存在 $i\neq j$ 使 $\mu_i\neq\mu_j$

假设 H_0 成立时，构造检验统计量为

$$F=\frac{SSA/(k-1)}{SSE/(n-k)}\sim F(k-1,n-k) \qquad (3.11)$$

多元统计

- 这里 $SSA=\sum_{i=1}^{k}n_i(\bar{X}_i-\bar{X})^2$ 称为组间平方和；$SSE=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_j^{(i)}-\bar{X}_i)^2$

  称为组内平方和；$SST=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_j^{(i)}-\bar{X})^2$ 称为总平方和。其中

$$\bar{X}_i=\frac{1}{n_i}\sum_{j=1}^{n}X_j^{(i)}$$

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_j^{(i)} \qquad n=n_1+\cdots+n_k$$

给定检验水平 α，查 F 分布表，使 $p\{F>F_\alpha\}=\alpha$，可确定出临界值 F_α，再用样本值计算出 F 值，若 $F>F_\alpha$，则否定 H_0，否则接受 H_0。

- 定义 3.2　若 $\mathbf{X} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$，则称协差阵的行列式 $|\boldsymbol{\Sigma}|$ 为 $\mathbf{X}$ 的广义方差。称 $\left|\frac{1}{n}\mathbf{S}\right|$ 为样本广义方差。其中 $\mathbf{S} = \sum_{a=1}^{n}(\mathbf{X}_{(a)} - \bar{\mathbf{X}})(\mathbf{X}_{(a)} - \bar{\mathbf{X}})'$。

- **定义 3.3**　若 $\mathbf{A}_1 \sim W_p(n_1, \boldsymbol{\Sigma}), n_1 \geq p$，$\mathbf{A}_2 \sim W_p(n_2, \boldsymbol{\Sigma}), \boldsymbol{\Sigma} > 0$，且 $\mathbf{A}_1$ 和 $\mathbf{A}_2$ 相互独立，则称

$$\Lambda = \frac{|\mathbf{A}_1|}{|\mathbf{A}_1 + \mathbf{A}_2|}$$

为 **Wilks** 统计量，Λ 的分布称为 **Wilks** 分布，简记为 $\Lambda \sim \Lambda(p, n_1, n_2)$，其中 n_1，n_2 为自由度。

- 这里我们需要说明的是，在实际应用中经常把 Λ 统计量化为 T^2 统计量进而化为 F 统计量，利用 F 统计量来解决多元统计分析中有关检验问题。表 **3.1** 列举常见的一些情形。

表 3.1　　Λ 与 F 统计量的关系

p	n_1	n_2	F 统计量及分别
任意	任意	1	$\frac{n_1-p+1}{p}\cdot\frac{1-\Lambda(p,n_1,1)}{\Lambda(p,n_1,1)}\sim F(p,n_1-p+1)$
任意	任意	2	$\frac{n_1-p}{p}\cdot\frac{1-\sqrt{\Lambda(p,n_1,2)}}{\sqrt{\Lambda(p,n_1,2)}}\sim F(2p,2(n_1-p))$
1	任意	任意	$\frac{n_1}{n_2}\cdot\frac{1-\Lambda(1,n_1,n_2)}{\Lambda(1,n_1,n_2)}\sim F(n_2,n_1)$
2	任意	任意	$\frac{n_1-1}{n_2}\cdot\frac{1-\sqrt{\Lambda(2,n_1,n_2)}}{\sqrt{\Lambda(2,n_1,n_2)}}\sim F(2n_2,2(n_1-1))$

以上几个关系式说明对一些特殊的 Λ 统计量可以化为 F 统计量，而当 $n_2>2$，$p>2$ 时，可用 χ^2 统计量或 F 统计量来近似表示，后面给出。

（二）多元方差分析法

设有k个p维正态总体$N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma})$，$\cdots$, $N_p(\boldsymbol{\mu}_k,\boldsymbol{\Sigma})$，从每个总体抽取独立样本个数分别为$n_1,n_2,\cdots,n_k$，$n_1+\cdots+n_k=n$，每个样品观测$p$个指标得观测数据如下：

第一个总体：$\mathbf{X}_i^{(1)}=(X_{i1}^{(1)},X_{i2}^{(1)},\cdots,X_{ip}^{(1)})$，$i=1,2,\cdots,n_1$

第二个总体：$\mathbf{X}_i^{(2)}=(X_{i1}^{(2)},X_{i2}^{(2)},\cdots,X_{ip}^{(2)})$，$i=1,2,\cdots,n_2$

……　　……　　……

第k个总体：$\mathbf{X}_i^{(k)}=(X_{i1}^{(k)},X_{i2}^{(k)},\cdots,X_{ip}^{(k)})$，$i=1,2,\cdots,n_k$

全部样品的总均值向量：

$$\bar{\mathbf{X}}_{1\times p}=\frac{1}{n}\sum_{r=1}^{k}\sum_{i=1}^{n_r}\mathbf{X}_{\mathbf{i}}^{(r)}=(\bar{\mathbf{X}}_1,\bar{\mathbf{X}}_2,\cdots,\bar{\mathbf{X}}_p)$$

- 各总体样品的均值向量：

$$\bar{\mathbf{X}}^{(r)}_{1\times p}=\frac{1}{n}\sum_{i=1}^{n_r}\mathbf{X}_i^{(r)}\triangleq(\bar{\mathbf{X}}_1^{(r)},\bar{\mathbf{X}}_2^{(r)},\cdots,\bar{\mathbf{X}}_p^{(r)})\text{，}\quad r=1,2,\cdots,k$$

此处
$$\bar{\mathbf{X}}_j^{(r)}=\frac{1}{n_r}\sum_{i=1}^{n_r}X_{ij}^{(r)}$$

类似一元方差分析办法，将诸平方和变成了离差阵即：

$$\mathbf{A}=\sum_{r=1}^{k}n_r(\bar{\mathbf{X}}^{(r)}-\bar{\mathbf{X}})'(\bar{\mathbf{X}}^{(r)}-\bar{\mathbf{X}})$$

$$\mathbf{E}=\sum_{r=1}^{k}\sum_{i=1}^{n_r}(\bar{\mathbf{X}}_i^{(r)}-\bar{\mathbf{X}}^{(r)})'(\bar{\mathbf{X}}_i^{(r)}-\bar{\mathbf{X}}^{(r)})$$

$$\mathbf{T}=\sum_{r=1}^{k}\sum_{i=1}^{n_r}(\bar{\mathbf{X}}_i^{(r)}-\bar{\mathbf{X}})'(\bar{\mathbf{X}}_i^{(r)}-\bar{\mathbf{X}})$$

这里，我们称 $\mathbf{A}$ 为组间离差阵；$\mathbf{E}$ 为组内离差阵；$\mathbf{T}$ 为总离差阵。很显然有 $\mathbf{T}=\mathbf{A}+\mathbf{E}$。

- 我们的问题是检验假设

$$H_0:\ \boldsymbol{\mu}_1=\boldsymbol{\mu}_2=\cdots=\boldsymbol{\mu}_k \qquad H_1:\ 至少存在 i\neq j 使 \boldsymbol{\mu}_i\neq\boldsymbol{\mu}_j$$

用似然比原则构成的检验统计量为

$$\Lambda=\frac{|\mathbf{E}|}{|\mathbf{T}|}=\frac{|\mathbf{E}|}{|\mathbf{A}+\mathbf{E}|}\sim\Lambda(p,n-k,k-1) \qquad \textbf{(3.13)}$$

给定检验水平α，查 **Wilks** 分布表，确定临界值，然后作出统计判断。在这里我们特别要注意，**Wilks** 分布表可用χ^2分布或F分布来近似。

巴特莱特（**Bartlett**）提出了用χ^2分布来近似。设$\Lambda\sim\Lambda(p,n,m)$，令

$$V=-(n+m-(p+m+1)/2)\ln\Lambda=\ln\Lambda^{1/t} \qquad \textbf{(3.14)}$$

则V近似服从$\chi^2(pm)$分布。其中，$t=n+m-(p+m+1)/2$。

- **Rao** 后来又研究用 F 分布来近似。设 $\Lambda \sim \Lambda(p,n,m)$，令

$$R=\frac{1-\Lambda^{1/L}}{\Lambda^{1/L}}\cdot\frac{tL-2\lambda}{pm} \qquad (3.15)$$

则 R 近似服从 $F(pm,tL-2\lambda)$，这里 $tL-2\lambda$ 不一定为整数，可用与它最近的整数来作为 F 的自由度，且 $\min(p,m)>2$。其中，

$$t=n+m-(p+m+1)/2$$

$$L=\left(\frac{p^2m^2-4}{p^2+m^2-5}\right)^{1/2}$$

$$\lambda=\frac{pm-2}{4}$$

第三节 协差阵的检验

一 一个正态总体协差阵的检验

二 多个协差阵相等检验

一、一个正态总体协差阵的检验

- 设 $\mathbf{X}_{(a)}=(X_{a1},X_{a2},\cdots,X_{ap})'$ $(a=1,2,\cdots,n)$ 来自 p 维正态总体 $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 的样本，$\boldsymbol{\Sigma}$ 未知，且 $\boldsymbol{\Sigma}>0$。

首先，我们考虑检验假设

$$H_0:\ \boldsymbol{\Sigma}=\mathbf{I}_p \qquad H_1:\ \boldsymbol{\Sigma}\neq\mathbf{I}_p$$

所构造的检验统计量为

$$\lambda=\exp\left\{-\frac{1}{2}tr\mathbf{S}\right\}|\mathbf{S}|^{n/2}\left(\frac{e}{n}\right)^{np/2} \tag{3.16}$$

其中

$$\mathbf{S}=\sum_{a=1}^{n}(\mathbf{X}_{(a)}-\bar{\mathbf{X}})(\mathbf{X}_{(a)}-\bar{\mathbf{X}})'$$

然后，我们考虑检验假设

$$H_0:\ \boldsymbol{\Sigma}=\boldsymbol{\Sigma}_0\neq\mathbf{I}_p \qquad H_1:\ \boldsymbol{\Sigma}\neq\boldsymbol{\Sigma}_0\neq\mathbf{I}_p$$

因为 $\boldsymbol{\Sigma}_0>0$，所以存在 $\mathbf{D}$ $(|\mathbf{D}|\neq 0)$，使得 $\mathbf{D}\boldsymbol{\Sigma}_0\mathbf{D}'=\mathbf{I}_p$。

- 令 $\mathbf{Y}_{(a)} = \mathbf{D}\mathbf{X}_{(a)} \qquad a = 1,2,\cdots,n$

则 $\mathbf{Y}_{(a)} \sim N_p(\mathbf{D\mu}, \mathbf{D\Sigma D'}) = N_p(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$

因此，检验$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$等价于检验$\boldsymbol{\Sigma}^* = \mathbf{I}_p$

此时构造检验统计量为

$$\lambda = \exp\left\{-\frac{1}{2}tr\mathbf{S}^*\right\}\left|\mathbf{S}^*\right|^{n/2}\left(\frac{e}{n}\right)^{np/2} \qquad (3.17)$$

其中 $\mathbf{S}^* = \sum_{a=1}^{n}(\mathbf{Y}_{(a)} - \bar{\mathbf{Y}})(\mathbf{Y}_{(a)} - \bar{\mathbf{Y}})'$

给定检验水平α，因为直接由λ分布计算临界值λ_0很困难，所以通常采用λ的近似分布。

在H_0成立时，$-2\ln\lambda$极限分布是$\chi^2_{p(p+1)/2}$分布。因此当$n >> p$，由样本值计算出λ值，若$-2\ln\lambda > \chi^2_\alpha$即$\lambda < e^{-\chi^2_\alpha/2}$，则拒绝$H_0$，否则接受$H_0$。

二、多个协差阵相等检验

- 设有 k 个正态总体分别为 $N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$， $\cdots$, $N_p(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$，$\boldsymbol{\Sigma}_i>0$ 且未知， $i=1,\cdots,k$ 。从 k 个总体分别取 n_i 个样本

$$\mathbf{X}_{(a)}^{(i)}=(X_{a1}^{(i)},X_{a2}^{(i)},\cdots,X_{ap}^{(i)})' \quad i=1,\cdots,k\text{；}\ a=1,\cdots,n_i$$

这里 $\sum_{i=1}^{k} n_i=n$ 为总样本容量。

我们考虑检验假设

$$H_0\text{:}\ \boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\cdots=\boldsymbol{\Sigma}_k \qquad H_1\text{:}\{\boldsymbol{\Sigma}_i\}\text{不全相等}$$

- 构造检验统计量为

$$\lambda_k = n^{np/2}\prod_{i=1}^{k}|\mathbf{S}_i|^{n_i/2} \Big/ |\mathbf{S}|^{n/2}\prod_{i=1}^{k} n_i^{pn_i/2} \qquad (3.18)$$

其中

$$\mathbf{S} = \sum_{i=1}^{k}\mathbf{S}_i$$

$$\mathbf{S}_i = \sum_{\alpha=1}^{n_i}(X_{(a)}^{(i)} - \bar{\mathbf{X}}^{(i)})(X_{(a)}^{(i)} - \bar{\mathbf{X}}^{(i)})'$$

$$\bar{\mathbf{X}}^{(i)} = \frac{1}{n_i}\sum_{\alpha=1}^{n_i}\mathbf{X}_{(a)}^{(i)}$$

多元统计

- 巴特莱特（Bartlett）建议 ，将 n_i 改为 n_i-1，从而 n 变为 $n-k$，变换以后的 λ_k 记为 λ_k'，称为修正的统计量，则 $-2\ln\lambda_k'$ 近似分布 $\chi_f^2/(1-\boldsymbol{D})$。 其中

$$f=\frac{1}{2}p(p+1)(k-1)$$

$$\mathbf{D}=\begin{cases}\dfrac{2p^2+3p-1}{6(p+1)(k-1)}\left(\displaystyle\sum_{i=1}^{k}\frac{1}{n_i-1}-\frac{1}{n-k}\right), & \text{至少有一对} \quad n_i\neq n_j\\[2ex] \dfrac{(2p^2+3p-1)(k+1)}{6(p+1)(n-k)}, & n_1=n_2=\cdots=n_k\end{cases}$$

本章结束