



第 1 章 多元统计分析的理论基础

1.1 多元分布

1.1.1 随机向量

多元统计分析讨论的是多变量总体。把 p 个随机变量放在一起得 $\mathbf{X} = (X_1, X_2, \dots, X_p)$ 是一个 p 维随机向量, 如果同时对 p 个变量作一次观测, 则得到一个样本 $(x_{11}, x_{12}, \dots, x_{1p})$, n 次观测得到的样本排成一个 $n \times p$ 矩阵, 称为样本数据矩阵(或样本资料矩阵), 记为

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{X}_{(1)} \\ \mathbf{X}_{(2)} \\ \vdots \\ \mathbf{X}_{(n)} \end{bmatrix} \stackrel{\text{def}}{=} (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p) \quad (1-1)$$

矩阵 $\mathbf{X}$ 的第 i 行: $\mathbf{X}_{(i)} = (x_{i1}, x_{i2}, \dots, x_{ip}) (i=1, 2, \dots, n)$ 表示对第 i 个样本的观

测值, 是一个 p 维的随机向量。矩阵 $\mathbf{X}$ 的第 j 列: $\mathbf{X}_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix} (j=1, 2, \dots, p)$ 表

示对第 j 个变量的 n 次观测, 是一个 n 维随机向量。

1.1.2 多元分布函数与密度函数

1. 联合分布

设 $\mathbf{X} = (X_1, X_2, \dots, X_p)$ 是 p 维随机向量, 称 p 元函数

$$F(x_1, x_2, \dots, x_p) = P\{X_1 \leq x_1, x_2, \dots, X_p \leq x_p\} \quad (1-2)$$

为 $\mathbf{X}$ 的联合分布函数。

若存在非负函数 $f(x_1, x_2, \dots, x_p)$, 使得随机向量 $\mathbf{X}$ 的联合分布函数对一切 $(x_1, x_2, \dots, x_p) \in \mathbb{R}^p$ 均可表示为

$$F(x_1, x_2, \dots, x_p) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_p} f(t_1, t_2, \dots, t_p) dt_1 dt_2 \dots dt_p \quad (1-3)$$

则称 $f(x_1, x_2, \dots, x_p)$ 为连续型随机向量 $\mathbf{X}$ 的联合概率密度函数, 简称为多元密度函数或密度函数。

多元密度函数 $f(x_1, x_2, \dots, x_p)$ 满足以下两条性质:

- (1) 对一切实数 $x_1, x_2, \dots, x_p, f(x_1, x_2, \dots, x_p) \geq 0$;
- (2) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_p) dx_1 dx_2 \dots dx_p = 1$ 。

2. 边缘分布

称随机向量 $\mathbf{X}$ 的部分分量 $(X_{i_1}, \dots, X_{i_m}) (1 \leq m < p)$ 的分布为边缘分布。

设 $\mathbf{X}^{(1)} = (X_{i_1}, \dots, X_{i_r})$ 为 r 维随机向量, $\mathbf{X}^{(2)} = (X_{i_{r+1}}, \dots, X_{i_p})$ 为 $p-r$ 维随机向量。若 $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$, 则 $\mathbf{X}^{(1)}$ 的边缘分布密度函数为

$$\begin{aligned} f_1(x^{(1)}) &= f_1(x_{i_1}, \dots, x_{i_r}) \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_p) dx_{i_{r+1}} \dots dx_{i_p} \end{aligned} \quad (1-4)$$

$\mathbf{X}^{(2)}$ 的边缘分布密度函数为

$$\begin{aligned} f_2(x^{(2)}) &= f_2(x_{i_{r+1}}, \dots, x_{i_p}) \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_p) dx_{i_1} \dots dx_{i_r} \end{aligned} \quad (1-5)$$

3. 条件分布

称给定 $\mathbf{X}^{(2)}$ 时 $\mathbf{X}^{(1)}$ 的分布为条件分布。当 $\mathbf{X}$ 的密度函数为 $f(x^{(1)}, x^{(2)})$ 、 $\mathbf{X}^{(2)}$ 的密度函数为 $f_2(x^{(2)})$ 时, 则给定 $\mathbf{X}^{(2)}$ 时 $\mathbf{X}^{(1)}$ 的条件密度函数为

$$f_1(x^{(1)} | x^{(2)}) = f(x^{(1)}, x^{(2)}) / f_2(x^{(2)}) \quad (1-6)$$

4. 独立性

设 $X_1, \dots, X_p$ 是 p 个随机变量, X_i 的分布函数记为 $F_i(x_i) (i=1, \dots, p)$,

$F(x_1, \cdots, x_p)$ 是 $\mathbf{X} = (X_1, X_2, \cdots, X_p)$ 的联合分布函数。若对一切实数 $x_1, x_2, \cdots, x_p$,

$$F(x_1, \cdots, x_p) = F_1(x_1) \cdots F_p(x_p) \quad (1-7)$$

均成立, 则称 $X_1, \cdots, X_p$ 相互独立。

在连续型随机变量的情况下, $X_1, \cdots, X_p$ 相互独立, 当且仅当 $\mathbf{X} = (X_1, X_2, \cdots, X_p)$ 的联合密度函数 $f(x_1, x_2, \cdots, x_p)$ 满足: 对一切实数 $x_1, x_2, \cdots, x_p$,

$$f(x_1, x_2, \cdots, x_p) = f_1(x_1) f_2(x_2) \cdots f_p(x_p) \quad (1-8)$$

均成立, 其中 $f_i(x_i)$ 是 $\mathbf{X}_i$ 的密度函数 ($i=1, \cdots, p$)。

1.1.3 随机向量的数字特征

设 $\mathbf{X} = (X_1, X_2, \cdots, X_p), \mathbf{Y} = (Y_1, Y_2, \cdots, Y_q)$ 是两个随机向量。

1. 均值向量

若 $E(X_i) = \mu_i$ 存在, 则称 $E(\mathbf{X}) = (E(X_1), E(X_2), \cdots, E(X_p)) = (\mu_1, \mu_2, \cdots, \mu_p)$ 为随机向量 $\mathbf{X}$ 的均值向量。

2. 协方差阵

若 X_i 和 Y_j 的协方差 $\text{cov}(X_i, Y_j)$ 存在 ($i=1, 2, \cdots, p; j=1, 2, \cdots, q$), 则称

$$\begin{aligned} \text{cov}(\mathbf{X}, \mathbf{Y}) &= E[(\mathbf{X} - E\mathbf{X})(\mathbf{Y} - E\mathbf{Y})] \\ &= \begin{bmatrix} \text{cov}(X_1, Y_1) & \text{cov}(X_1, Y_2) & \cdots & \text{cov}(X_1, Y_q) \\ \text{cov}(X_2, Y_1) & \text{cov}(X_2, Y_2) & \cdots & \text{cov}(X_2, Y_q) \\ \vdots & \vdots & & \vdots \\ \text{cov}(X_p, Y_1) & \text{cov}(X_p, Y_2) & \cdots & \text{cov}(X_p, Y_q) \end{bmatrix} \end{aligned} \quad (1-9)$$

为随机向量 $\mathbf{X}$ 和 $\mathbf{Y}$ 的协方差阵。若 $\text{cov}(\mathbf{X}, \mathbf{Y}) = \mathbf{O}$ (其中 $\mathbf{O}$ 表示零矩阵), 则称 $\mathbf{X}$ 和 $\mathbf{Y}$ 不相关。若 $\mathbf{Y} = \mathbf{X}$, 则 $\text{cov}(\mathbf{X}, \mathbf{X}) = D(\mathbf{X})$ 为 $\mathbf{X}$ 的(协)方差阵。

3. 相关阵

记 $r_{ij} = \frac{\text{cov}(X_i, X_j)}{\sqrt{\text{var}(X_i)} \sqrt{\text{var}(X_j)}} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii} \sigma_{jj}}}$ ($i, j=1, 2, \cdots, p$), 称 $\mathbf{R} = (r_{ij})_{p \times p}$

为 $\mathbf{X}$ 的相关阵, r_{ij} 也称为分量 X_i 与 X_j 之间的(线性)相关系数。

1.2 多元正态分布

1.2.1 多元正态分布的定义和性质

1. 多元正态分布的定义

定义 1(密度定义) 若 p 元随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_p)$ 的概率密度函数为

$$f(x_1, x_2, \dots, x_p) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right\} \quad (1-10)$$

其中, $\boldsymbol{\mu}$ 为 $p \times 1$ 的常数向量, Σ 为正定阵 ($\Sigma > O$), $|\Sigma|$ 为协方差阵 Σ 的行列式, 则称 $\mathbf{X}$ 服从 p 元正态分布, 也称 $\mathbf{X}$ 为 p 元正态变量。记为 $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ 。

当 $p=2$ 时, 可以得到二元正态分布的密度公式。

2. 多元正态分布的性质

性质 1 若 $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$, 则 $E(\mathbf{X}) = \boldsymbol{\mu}$, $D(\mathbf{X}) = \Sigma$ 。

本性质将正态分布的参数 $\boldsymbol{\mu}$ 和 Σ 赋予了明确的统计意义。

性质 2 设 $Y_1, Y_2, \dots, Y_k$ 相互独立, $Y_i \sim N_p(\mu_i, \Sigma_i) (\forall i)$, 则

$$\sum_{i=1}^k Y_i \sim N_p \left(\sum_{i=1}^k \mu_i, \sum_{i=1}^k \Sigma_i \right) \quad (1-11)$$

性质 3 如果正态随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_p)$ 的协方差阵 Σ 是对角阵, 则 $\mathbf{X}$ 的各分量是相互独立的随机变量。

性质 4 多元正态向量 $\mathbf{X} = (X_1, X_2, \dots, X_p)$ 的任意线性变换仍然服从多元正态分布。

1.2.2 多元正态分布均值向量和协方差阵的估计

1. 均值向量的估计

设样本 $\mathbf{X}_{(1)}, \mathbf{X}_{(2)}, \dots, \mathbf{X}_{(n)}$ 间相互独立, 且服从于 p 元正态分布 $N_p(\boldsymbol{\mu}, \Sigma)$, $n > p$, $\Sigma > O$, 则总体参数均值 $\boldsymbol{\mu}$ 的估计量是

$$\begin{aligned}\hat{\mu} &= \bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_{(i)} = \frac{1}{n} \left(\sum_{i=1}^n X_{i1}, \sum_{i=1}^n X_{i2}, \dots, \sum_{i=1}^n X_{ip} \right) \mathbf{X} \\ &= (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)\end{aligned}\quad (1-12)$$

均值向量 μ 的估计量就是样本均值向量, 且 $\hat{\mu}$ 是 μ 的无偏估计。

2. 协方差阵的估计

总体参数协方差阵 Σ 的极大似然估计是

$$\begin{aligned}\hat{\Sigma}_p &= \frac{1}{n} \mathbf{L} = \frac{1}{n} \sum_{i=1}^n (\mathbf{X}_{(i)} - \bar{\mathbf{X}})(\mathbf{X}_{(i)} - \bar{\mathbf{X}})^T \\ &= \frac{1}{n} \begin{bmatrix} \sum_{i=1}^n (X_{i1} - \bar{X}_1)^2 & \sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) & \cdots & \sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{ip} - \bar{X}_p) \\ & \sum_{i=1}^n (X_{i2} - \bar{X}_2)^2 & \cdots & \sum_{i=1}^n (X_{i2} - \bar{X}_2)(X_{ip} - \bar{X}_p) \\ & & \ddots & \vdots \\ & & & \sum_{i=1}^n (X_{ip} - \bar{X}_p)^2 \end{bmatrix}\end{aligned}\quad (1-13)$$

其中 $\mathbf{L}$ 是离差阵, 它是每一个样本(向量)与样本均值(向量)的离差积形成的 n 个 p 阶对称阵的和。同一元相似, $\hat{\Sigma}_p$ 不是 Σ 的无偏估计, 为了得到无偏估计我们常用样本协差阵 $\hat{\Sigma} = \frac{1}{n-1} \mathbf{L}$ 作为总体协差阵的估计。

1.3 多元正态分布均值向量和协方差阵的检验

1.3.1 单总体均值向量的检验

设总体 $\mathbf{X} \sim N_p(\mu, \Sigma)$, 随机样本 $\mathbf{X}_{(\alpha)} (\alpha=1, \dots, n)$ 。检验

$$H_0: \mu = \mu_0 (\mu_0 \text{ 为已知向量}), \quad H_1: \mu \neq \mu_0$$

1. 当 $\Sigma = \Sigma_0$ 已知时均值向量的检验

取检验统计量为

$$T_0^2 = n(\bar{\mathbf{X}} - \mu_0)^T \Sigma_0^{-1} (\bar{\mathbf{X}} - \mu_0) \stackrel{H_0}{\sim} \chi^2(p) \quad (1-14)$$

按传统的检验方法,对给定的显著性水平 α ,查 χ^2 分布临界值表得 λ_α ,使 $P\{T_0^2 > \lambda_\alpha\} = \alpha$,则否定域为 $\{T_0^2 > \lambda_\alpha\}$ 。

由样本值 $x_{(\alpha)}$ ($\alpha=1, \dots, n$),计算 $\bar{X}$ 及 T_0^2 值,若 $T_0^2 > \lambda_\alpha$,则否定 H_0 ,否则 H_0 相容。

利用统计软件还可以通过计算显著性概率值(p 值)给出检验结果,且由此得出的结果更丰富。

假设在 H_0 成立情况下,随机变量 $T_0^2 \sim \chi^2(p)$,由样本值计算得到 T_0^2 的值为 d ,同时可以计算以下概率值

$$p = P\{T_0^2 \geq d\} \quad (1-15)$$

常称此概率值为**显著性概率值**,或简称为 p 值。

对给定的显著性水平 α ,当 $p < \alpha$ 时,则在显著性水平 α 下否定假设 H_0 。在这种情况下,可能犯“以真当假”的第一类错误,且 α 就是犯第一类错误的概率。

当 $p \geq \alpha$ 时,在显著性水平 α 下 H_0 相容。在这种情况下,可能犯“以假当真”的第二类错误,且犯下第二类错误的概率 β 为

$$\beta = P\{T_0^2 \leq \lambda_\alpha \mid \mu = \mu_1 \neq \mu_0\} \quad (1-16)$$

其中检验统计量 $T_0^2 \sim \chi^2(p, \delta)$,非中心参数

$$\delta = n(\mu_1 - \mu_0)^T \Sigma_0^{-1} (\mu_1 - \mu_0) \quad (1-17)$$

p 值的直观含义可以这样看,检验统计量 T_0^2 的大小反映 $\bar{X}$ 与 μ_0 的偏差大小,当 H_0 成立时 T_0^2 的值应该较小。现由观测数据计算 T_0^2 值为 d ;当 H_0 成立时统计量 $T_0^2 \sim \chi^2(p)$,由 χ^2 分布可以计算该统计量 $\geq d$ 的概率值(即 p 值)。比如 $p=0.02 < \alpha=0.05$,这时出现一个比较小的概率标准($\alpha=0.05$)还要小的事件 $\{T_0^2 \geq d\}$ 。也就是说,在 $\mu = \mu_0$ 假设下,观测数据中极少情况会出现 T_0^2 的值大于等于 d 值,故在 0.05 显著性水平下有足够的证据否定原假设,即认为 μ 与 μ_0 有显著的差异。

又比如当 $p=0.22 \geq \alpha=0.05$ 时,表示在 $\mu = \mu_0$ 假设下,观测数据中经常会出现 T_0^2 的值大于等于 d 值的情况,故在 0.05 显著性水平下没有足够的证据否定原假设,即认为 μ 与 μ_0 没有显著的差异。

2. 当 Σ 未知时均值向量的检验

考虑统计量

$$\begin{aligned} T^2 &= (n-1)[\sqrt{n}(\bar{X} - \mu_0)]^T \mathbf{A}^{-1} [\sqrt{n}(\bar{X} - \mu_0)] \\ &= (n-1)n(\bar{X} - \mu_0)^T \mathbf{A}^{-1} (\bar{X} - \mu_0) \sim T^2(p, n-1) \end{aligned} \quad (1-18)$$

再利用 T^2 与 F 分布的关系,统计量取为

$$\begin{aligned}
 F &= \frac{(n-1)-p+1}{(n-1)p} T^2 \sim F(p, (n-1)-p+1) \\
 &= F(p, n-p)
 \end{aligned} \tag{1-19}$$

进行检验。

3. 似然比统计量

在数理统计中关于总体参数的假设检验,通常是利用最大似然原理导出似然比统计量来进行检验。在多元统计分析中几乎所有重要的检验都是利用最大似然比原理给出的。下面我们回顾一下最大似然比原理。

设 p 元总体的密度函数为 $f(x, \theta)$, 其中 θ 是未知参数, 且 $\theta \in \Theta$ (参数空间), 又设 Θ_0 是 Θ 的子集, 我们希望对下列假设:

$$H_0: \theta \in \Theta_0, \quad H_1: \theta \in \overline{\Theta_0}$$

作出判断, 这就是假设检验问题。称 H_0 为原假设 (或零假设), H_1 为对立假设 (或备择假设)。

从总体 $\mathbf{X}$ 抽取容量为 n 的样本 $\mathbf{X}_{(t)} (t=1, \dots, n)$ 。把样本的联合密度函数

$$L(x_{(1)}, \dots, x_{(n)}; \theta) = \prod_{t=1}^n f(x_{(t)}; \theta) \tag{1-20}$$

记为 $L(\mathbf{X}; \theta)$, 并称它为样本的似然函数。

引入统计量

$$\lambda = \max_{\theta \in \Theta_0} L(\mathbf{X}; \theta) \mid \max_{\theta \in \Theta} L(\mathbf{X}; \theta) \tag{1-21}$$

它是样本 $\mathbf{X}_{(t)} (t=1, \dots, n)$ 的函数, 常称 λ 为似然比统计量。由于 $\Theta_0 \subset \Theta$, 从而 $0 \leq \lambda \leq 1$ 。

由最大似然比原理知, 如果 λ 取值太小, 说明 H_0 为真时观测到此样本 $\mathbf{X}_{(t)} (t=1, \dots, n)$ 的概率比 H_0 为不真时观测到此样本 $\mathbf{X}_{(t)} (t=1, \dots, n)$ 的概率要小得多, 故有理由认为假设 H_0 不成立, 所以从似然比出发, 以上检验问题的否定域为

$$\{\lambda(\mathbf{X}_{(1)}, \dots, \mathbf{X}_{(n)}) < \lambda_\alpha\} \tag{1-22}$$

按传统的检验方法, λ_α 是由显著性水平 α 确定的临界值, 它满足当 H_0 成立时使得

$$P\{\lambda(\mathbf{X}_{(1)}, \dots, \mathbf{X}_{(n)}) < \lambda_\alpha\} = \alpha \tag{1-23}$$

为了得到 λ_α , 必须研究似然比统计量 λ 的抽样分布。在一些特殊的情况下, 可以得到 λ 的精确分布; 但在很多情况下是得不到 λ 的精确分布的。当样本量很大且满足一定正则条件时, $-2\ln\lambda$ 的抽样分布与 χ^2 分布十分接近。下面不加证明地给出一条很有用的结论。

定理 1 当样本容量 n 很大时,

$$-2\ln\lambda = -2\ln\left[\frac{\max_{\theta\in\Theta_0}L(\mathbf{X};\theta)}{\max_{\theta\in\Theta}L(\mathbf{X};\theta)}\right] \quad (1-24)$$

近似服从自由度为 f 的 χ^2 分布, 其中 $f = \dim\Theta - \dim\Theta_0$ 的维数。

本章将讨论的一些检验问题, 就是利用似然比统计量的近似分布进行检验的方法。

当 Σ 未知时检验均值向量 $\mu = \mu_0$ 的似然比统计量及其分布。

设样本的似然函数为 $L(\mu, \Sigma)$ 。检验均值向量 $\mu = \mu_0$ 的似然比统计量为

$$\lambda = \frac{\max_{\mu=\mu_0, \Sigma>0} L(\mu_0, \Sigma)}{\max_{\mu, \Sigma>0} L(\mu, \Sigma)} \quad (1-25)$$

否定域为

$$\{\lambda < \lambda_\alpha\} \Leftrightarrow \{T^2 > T_\alpha^2\} \Leftrightarrow \{F > F_\alpha\}$$

其中

$$F = \frac{n-p}{p} \frac{T^2}{n-1} \sim F(p, n-p) \quad (1-26)$$

1.3.2 多总体均值向量的检验

1. 两正态总体均值向量的检验

1) 两总体协方差阵相等(但未知)时均值向量的检验

设 $\mathbf{X}_{(\alpha)}$ ($\alpha=1, \dots, n$) 为来自总体 $\mathbf{X} \sim N_{(p)}(\mu^{(1)}, \Sigma)$ 的随机样本; $\mathbf{Y}_{(\alpha)}$ ($\alpha=1, \dots, m$) 为来自样本总体 $\mathbf{Y} \sim N_{(p)}(\mu^{(2)}, \Sigma)$ 的随机样本, 且相互独立, Σ 未知。检验

$$H_0: \mu^{(1)} = \mu^{(2)}, \quad H_1: \mu^{(1)} \neq \mu^{(2)}$$

当 $p=1$ 时, 取检验统计量为

$$t = \frac{(\bar{\mathbf{X}} - \bar{\mathbf{Y}}) \left/ \sqrt{\frac{1}{n} + \frac{1}{m}} \right.}{\sqrt{\left[\sum_{i=1}^n (\mathbf{X}_{(i)} - \bar{\mathbf{X}})^2 + \sum_{j=1}^m (\mathbf{Y}_{(j)} - \bar{\mathbf{Y}})^2 \right] / (n+m-2)}} \sim t(n+m-2) \quad (1-27)$$

即

$$t^2 = \frac{nm}{m+n} (\bar{\mathbf{X}} - \bar{\mathbf{Y}})^T \left[\frac{\sum_{i=1}^n (\mathbf{X}_{(i)} - \bar{\mathbf{X}})^2 + \sum_{j=1}^m (\mathbf{Y}_{(j)} - \bar{\mathbf{Y}})^2}{n+m-2} \right]^{-1} (\bar{\mathbf{X}} - \bar{\mathbf{Y}})$$

$$\sim F(1, n+m-2) \quad (1-28)$$

推广到 p 元总体, 检验统计量的形式类似, 可考虑检验统计量 T^2 。

$$T^2 = \frac{mn}{m+n} (\bar{\mathbf{X}} - \bar{\mathbf{Y}})^T \left(\frac{\mathbf{A}_1 + \mathbf{A}_2}{n+m-2} \right)^{-1} (\bar{\mathbf{X}} - \bar{\mathbf{Y}}) \quad (1-29)$$

其中 $\mathbf{A}_1$ 和 $\mathbf{A}_2$ 是两总体的样本离差阵。 $\mathbf{A}_1$ 是一元统计中的偏差平方和 $\sum (\mathbf{X}_{(i)} - \bar{\mathbf{X}})^2$ 在 p 元情况下的推广。

2) 两总体协方差阵不等时均值向量的检验

在一元统计中 ($p=1$ 时), 当 $\sigma_1^2 \neq \sigma_2^2$ 时, 检验 $H_0: \mu^{(1)} = \mu^{(2)}$ 也没有很好的方法, 以下介绍几种实用的方法。

(1) 当 $n=m$ 时, 作为成对数据进行处理。令

$$\mathbf{Z}_{(i)} = \mathbf{X}_{(i)} - \mathbf{Y}_{(i)} \quad (i=1, \dots, n) \quad (1-30)$$

将两个总体化为单个 p 元总体 $\mathbf{Z}$ 的均值检验问题。

$$H_0: \mu^{(1)} = \mu^{(2)} \Leftrightarrow H_0: \mu_z = \mathbf{O}_p$$

(2) 当 $n \neq m$ 时 (不妨设 $n < m$), 想法也是将其化为单个 p 元新总体的均值检验问题。若只取 n 对数据, 按 (1) 的方法处理又将损失一些信息。改进的办法是利用 $\mathbf{X}_{(i)} (i=1, \dots, n)$ 和 $\mathbf{Y}_{(j)} (j=1, \dots, m)$ 构造新总体 $\mathbf{Z}$ 的样本 $\mathbf{Z}_{(i)}$, 令

$$\mathbf{Z}_{(i)} = \mathbf{X}_{(i)} - \sqrt{\frac{n}{m}} \mathbf{Y}_{(i)} + \frac{1}{\sqrt{nm}} \sum_{j=1}^n \mathbf{Y}_{(j)} - \frac{1}{m} \sum_{j=1}^m \mathbf{Y}_{(j)} \quad (1-31)$$

$$(i=1, 2, \dots, n)$$

则

$$E(\mathbf{Z}_{(i)}) = \mu^{(1)} - \mu^{(2)} \quad (1-32)$$

$$\text{cov}(\mathbf{Z}_{(i)}, \mathbf{Z}_{(j)}) = \begin{cases} \Sigma_1 + \frac{n}{m} \Sigma_2, & i=j \\ \mathbf{O}, & i \neq j \end{cases} \quad (1-33)$$

即

$$\text{cov}(\mathbf{Z}_{(i)}, \mathbf{Z}_{(j)}) = \delta_{ij} \Sigma_2 = \delta_{ij} \left(\Sigma_1 + \frac{n}{m} \Sigma_2 \right)$$

所以 $\mathbf{Z}_{(i)} \sim N_p(\mu^{(1)} - \mu^{(2)}, \Sigma_Z) (i=1, 2, \dots, n)$, 且相互独立。利用前面介绍的单个正态总体均值向量的检验方法进行检验。

(3) 当 Σ_1, Σ_2 相差甚大时, 可以构造近似检验统计量进行检验。

2. 多个正态总体均值向量的检验——多元方差分析

设有 k 个 p 元正态总体 $N_p(\mu^{(t)}, \Sigma) (t=1, 2, \dots, k)$, $X_{(a)}^{(t)} (t=1, \dots, k; a=1, \dots, n_t)$ 是来自 $N_p(\mu^{(t)}, \Sigma)$ 的随机样本, 检验 $H_0: \mu^{(1)} = \dots = \mu^{(k)}$, H_1 : 至少存在 $i \neq j$ 使得 $\mu^{(i)} \neq \mu^{(j)}$ (即 $\mu^{(1)}, \dots, \mu^{(k)}$ 中至少有一对不等)。

当 $p=1$ 时,此检验问题就是一元方差分析问题,比如比较 k 个不同品牌的同类产品中某一个质量指标 X (如耐磨度) 有无显著差异的问题。我们把不同品牌对应不同总体(假定为正态总体),这种多组比较问题就是检验 $H_0: \mu^{(1)} = \dots = \mu^{(k)}$; H_1 : 至少存在 $i \neq j$ 使得 $\mu^{(i)} \neq \mu^{(j)}$ 。从第 i 个总体抽取容量为 n_i ($i = 1, \dots, k; n = n_1 + n_2 + \dots + n_k$) 的随机样本如下:

$$\begin{array}{cccc} \mathbf{X}_{(1)}^{(1)} & \mathbf{X}_{(2)}^{(1)} & \cdots & \mathbf{X}_{(n_1)}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}_1^{(k)} & \mathbf{X}_{(2)}^{(k)} & \cdots & \mathbf{X}_{(n_k)}^{(k)} \end{array}$$

记

$$\bar{\mathbf{X}} = \frac{1}{n} \sum_{t=1}^k \sum_{j=1}^{n_t} \mathbf{X}_{(j)}^{(t)} \quad (1-34)$$

$$\bar{\mathbf{X}}^{(t)} = \frac{1}{n_t} \sum_{j=1}^{n_t} \mathbf{X}_{(j)}^{(t)} \quad (t = 1, \dots, k) \quad (1-35)$$

当 $p=1$ 时,利用一元方差分析的思想来构造检验统计量。

记

$$\text{总偏差平方和 SST} = \sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{X}_{(j)}^{(i)} - \bar{\mathbf{X}})^2 \quad (1-36)$$

$$\text{组内偏差平方和 SSE} = \sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{X}_{(j)}^{(i)} - \bar{\mathbf{X}}^{(i)})^2 \quad (1-37)$$

$$\text{组间偏差平方和 SSA} = \sum_{i=1}^k n_i (\bar{\mathbf{X}}^{(i)} - \bar{\mathbf{X}})^2 \quad (1-38)$$

则有平方和分解公式:

$$\text{SST} = \text{SSA} + \text{SSE} \quad (1-39)$$

直观考察,若 H_0 成立,当总偏差平方和 SST 固定不变时,应有 SSA 小而 SSE 大,因而比值 SSA/SSE 应很小。检验统计量取为

$$F = \frac{\text{SSA}/(k-1)}{\text{SSE}/(n-k)} \sim F(k-1, n-k) \quad (1-40)$$

给定显著性水平 α ,按传统检验方法,查 F 分布临界值表得 F_α 满足: $P\{F > F_\alpha\} = \alpha$,否定域 $W = \{F > F_\alpha\}$ 。

推广到 k 个 p 元正态总体 $N_p(\mu^{(i)}, \Sigma)$ (假定 k 个总体的协方差阵相等,且记为 Σ),记第 i 个 p 元总体的数据阵为

$$\mathbf{X}^{(i)} = \begin{bmatrix} X_{11}^{(i)} & \cdots & X_{1p}^{(i)} \\ \vdots & \ddots & \vdots \\ X_{n_i 1}^{(i)} & \cdots & X_{n_i p}^{(i)} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{(1)}^{(i)\text{T}} \\ \vdots \\ \mathbf{X}_{(n_i)}^{(i)\text{T}} \end{bmatrix} \quad (i = 1, \dots, k) \quad (1-41)$$

对总离差阵 $\mathbf{T}$ 进行分解:

$$\begin{aligned}
 \mathbf{T} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{X}_{(j)}^{(i)} - \bar{\mathbf{X}})(\mathbf{X}_{(j)}^{(i)} - \bar{\mathbf{X}})^{\mathrm{T}} \\
 &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{X}_{(j)}^{(i)} - \bar{\mathbf{X}}^{(i)} + \bar{\mathbf{X}}^{(i)} - \bar{\mathbf{X}})(\mathbf{X}_{(j)}^{(i)} - \bar{\mathbf{X}}^{(i)} + \bar{\mathbf{X}}^{(i)} - \bar{\mathbf{X}})^{\mathrm{T}} \\
 &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{X}_{(j)}^{(i)} - \bar{\mathbf{X}}^{(i)})(\mathbf{X}_{(j)}^{(i)} - \bar{\mathbf{X}}^{(i)})^{\mathrm{T}} + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{\mathbf{X}}^{(i)} - \bar{\mathbf{X}})(\bar{\mathbf{X}}^{(i)} - \bar{\mathbf{X}})^{\mathrm{T}} \\
 &= \sum_{i=1}^k \mathbf{A}_i + \sum_{i=1}^k n_i (\bar{\mathbf{X}}^{(i)} - \bar{\mathbf{X}})(\bar{\mathbf{X}}^{(i)} - \bar{\mathbf{X}})^{\mathrm{T}} \\
 &= \mathbf{A} + \mathbf{B}
 \end{aligned} \tag{1-42}$$

其中, $\mathbf{A} = \sum_{i=1}^k \mathbf{A}_i$ 称为组内离差阵, $\mathbf{B} = \sum_{i=1}^k n_i (\bar{\mathbf{X}}^{(i)} - \bar{\mathbf{X}})(\bar{\mathbf{X}}^{(i)} - \bar{\mathbf{X}})^{\mathrm{T}}$ 称为组间离差阵。

根据直观想法及用似然比原理得到检验 H_0 的统计量为

$$\Lambda = \frac{|\mathbf{A}|}{|\mathbf{A} + \mathbf{B}|} = \frac{|\mathbf{A}|}{|\mathbf{T}|} \tag{1-43}$$

易见:

(1) 因 $\mathbf{A}_i \sim W_p(n_i - 1, \Sigma)$ 且相互独立 ($i = 1, 2, \dots, k$), 由可加性可得

$$\mathbf{A} = \sum_{i=1}^k \mathbf{A}_i \sim W_p(n - k, \Sigma) \quad (n = n_1 + \dots + n_k)$$

(2) 在 H_0 下, $\mathbf{T} \sim W_p(n - 1, \Sigma)$;

(3) 还可以证明在 H_0 下, $\mathbf{B} \sim W_p(k - 1, \Sigma)$, 且 $\mathbf{B}$ 与 $\mathbf{A}$ 相互独立。

根据 Λ 分布的定义, 可知

$$\Lambda = \frac{|\mathbf{A}|}{|\mathbf{A} + \mathbf{B}|} \sim \Lambda(p, n - k, k - 1) \tag{1-44}$$

给定显著性水平 α , 查威尔克斯分布临界值表, 可得 λ_α , 使

$$P\{\Lambda < \lambda_\alpha\} = \alpha \tag{1-45}$$

故否定域 $W = \{\Lambda < \lambda_\alpha\}$ 。当手头没有威尔克斯临界表时, 可用 χ^2 分布或 F 分布来近似, 即由 Λ 的函数的近似分布进行检验。

1.3.3 协方差阵的检验

1. 单个 p 元正态总体协方差阵的检验

设 $\mathbf{X}_{(\alpha)}$ ($\alpha = 1, \dots, n$) 为来自 p 元正态总体 $N_p(\mu, \Sigma)$ ($\Sigma > \mathbf{O}$ 未知) 的随机样

本, 检验

$$H_0: \Sigma = \Sigma_0 (\Sigma_0 > \mathbf{O} \text{ 为已知矩阵}), \quad H_1: \Sigma \neq \Sigma_0$$

1) 当 $\Sigma_0 = \mathbf{I}_p$ 时检验 $H_0: \Sigma = \mathbf{I}_p, H_1: \Sigma \neq \mathbf{I}_p$

利用似然比原理来导出似然比统计量 λ_1

$$\lambda_1 = \max_{\mu} L(\mu, \mathbf{I}_p) \bigg/ \max_{\mu, \Sigma > \mathbf{O}} L(\mu, \Sigma) \quad (1-46)$$

当 $\Sigma = \mathbf{I}_p$ 成立时, 似然函数 $L(\mu, \mathbf{I}_p)$ 在 $\mu = \bar{\mathbf{X}}$ 时达最大值, 因此

$$\begin{aligned} \lambda_1 \text{ 表示式的分子} &= L(\bar{\mathbf{X}}, \mathbf{I}_p) \\ &= (2\pi)^{-np/2} |\mathbf{I}_p|^{-n/2} \exp\left[-\frac{1}{2} \text{tr}(\mathbf{I}_p^{-1} \mathbf{A})\right] \end{aligned} \quad (1-47)$$

$$\begin{aligned} \lambda_1 \text{ 表示式的分母} &= L\left(\bar{\mathbf{X}}, \frac{1}{n} \mathbf{A}\right) = (2\pi)^{-np/2} \left|\frac{1}{n} \mathbf{A}\right|^{-n/2} e^{-np/2} \\ &= (2\pi)^{-np/2} \left(\frac{e}{n}\right)^{-np/2} |\mathbf{A}|^{-n/2} \end{aligned} \quad (1-48)$$

所以似然比统计量

$$\lambda_1 = \exp\left\{-\frac{1}{2} \text{tr}(\mathbf{A})\right\} |\mathbf{A}|^{n/2} \left(\frac{e}{n}\right)^{np/2} \quad (1-49)$$

其中,

$$\mathbf{A} = \sum_{a=1}^n (\mathbf{X}_{(a)} - \bar{\mathbf{X}})(\mathbf{X}_{(a)} - \bar{\mathbf{X}})^T \quad (1-50)$$

利用定理 1.3.1 可知, 当 n 很大且 H_0 成立时, $\xi = -2 \ln \lambda_1$ 的近似分布为

$\chi^2\left(\frac{p(p+1)}{2}\right)$, 利用检验统计量 ξ 来构造检验方法。

2) 当 $\Sigma_0 \neq \mathbf{I}_p$ 时检验 $H_0: \Sigma = \Sigma_0, H_1: \Sigma \neq \Sigma_0$

因 $\Sigma_0 > \mathbf{O}$, 存在非退化矩阵 $\mathbf{D}_{p \times p}$, 使 $\mathbf{D} \Sigma_0 \mathbf{D}^T = \mathbf{I}_p$ 。令

$$\mathbf{Y}_{(a)} = \mathbf{D} \mathbf{X}_{(a)}, \quad \alpha = 1, \dots, n \quad (1-51)$$

则

$$\mathbf{Y}_{(a)} \sim N_p(\mathbf{D} \mu, \mathbf{D} \Sigma \mathbf{D}^T) \stackrel{\text{def}}{=} N_p(\mu^*, \Sigma^*) \quad (1-52)$$

检验

$$H_0: \Sigma = \Sigma_0 \Leftrightarrow H_0: \Sigma^* = \mathbf{I}_p$$

从新样本 $\mathbf{Y}_{(a)} (\alpha = 1, \dots, n)$ 出发, 检验 $H_0: \Sigma^* = \mathbf{I}_p$ 的似然比统计量取为

(以下记 $\exp(\text{tr} \mathbf{A}) \stackrel{\text{def}}{=} \text{etr}(\mathbf{A}))$:

$$\lambda_2 = \exp\left\{-\frac{1}{2} \text{tr}(\mathbf{A}^*)\right\} |\mathbf{A}^*|^{n/2} \left(\frac{e}{n}\right)^{np/2}$$

$$= \text{etr}\left(-\frac{1}{2}\mathbf{A}^*\right) |\mathbf{A}^*|^{n/2} \left(\frac{\mathbf{e}}{n}\right)^{np/2} \quad (1-53)$$

其中, $\mathbf{A}^* = \sum_{a=1}^n (\mathbf{Y}_{(a)} - \bar{\mathbf{Y}})(\mathbf{Y}_{(a)} - \bar{\mathbf{Y}})^T = \mathbf{DAD}^T$ 。若注意到 $\mathbf{D}\Sigma_0\mathbf{D}^T = \mathbf{I}_p$, 则似然比统计量 λ_2 还可以表示为

$$\lambda_2 = \exp\left(-\frac{1}{2}\mathbf{A}\Sigma_0^{-1}\right) |\mathbf{A}\Sigma_0^{-1}|^{n/2} \left(\frac{\mathbf{e}}{n}\right)^{np/2} \quad (1-54)$$

研究似然比统计量 λ_2 的抽样分布是很困难的, 通常根据定理 1.3.1 由 λ_2 的近似分布来构造检验法。

样本容量 n 很大, 在 H_0 成立时, $-2\ln\lambda_2$ 的极限分布为 $\chi^2\left(\frac{p(p+1)}{2}\right)$ 。除此之外, 在不同试用范围下还有其他近似分布可用来构造检验法。

3) 检验 $H_0: \Sigma = \sigma^2 \Sigma_0$ (σ^2 未知)

当 $\Sigma_0 = \mathbf{I}_p$ 时此检验常称为**球性检验**。以下利用似然比原理来导出似然比统计量 λ_3 。

$$\lambda_3 = \max_{\mu, \sigma^2 > 0} L(\mu, \sigma^2 \Sigma_0) \Big/ \max_{\mu, \Sigma > \mathbf{O}} L(\mu, \Sigma) \quad (1-55)$$

当 σ^2 给定时, 似然函数 $L(\mu, \sigma^2 \Sigma_0)$ 在 $\mu = \bar{\mathbf{X}}$ 时达最大值, 且

$$L(\bar{\mathbf{X}}, \sigma^2 \Sigma_0) = (2\pi)^{-np/2} (\sigma^2)^{-np/2} |\Sigma_0|^{-n/2} \text{etr}\left(-\frac{1}{2\sigma^2}(\Sigma_0^{-1}\mathbf{A})\right) \quad (1-56)$$

令

$$\begin{aligned} & \frac{\partial L(\bar{\mathbf{X}}, \sigma^2 \Sigma_0)}{\partial \sigma^2} \\ &= (2\pi)^{-np/2} (\sigma^2)^{-np/2-2} |\Sigma_0|^{-n/2} \text{etr}\left(-\frac{1}{2\sigma^2}(\Sigma_0^{-1}\mathbf{A})\right) \left[-\frac{np}{2}\sigma^2 + \frac{1}{2}\text{tr}(\Sigma_0^{-1}\mathbf{A})\right] \\ &= 0 \end{aligned} \quad (1-57)$$

可得出 $\sigma^2 = \frac{1}{np}\text{tr}(\Sigma_0^{-1}\mathbf{A})$ 。从而有

$$\lambda_3 \text{ 表示式的分子} = (2\pi)^{-np/2} \left(\frac{1}{np}\text{tr}(\Sigma_0^{-1}\mathbf{A})\right)^{-np/2} |\Sigma_0|^{-n/2} \mathbf{e}^{-np/2} \quad (1-58)$$

$$\lambda_3 \text{ 表示式的分母} = L\left(\bar{\mathbf{X}}, \frac{1}{n}\mathbf{A}\right) = (2\pi)^{-np/2} \left(\frac{\mathbf{e}}{n}\right)^{-np/2} |\mathbf{A}|^{-n/2} \quad (1-59)$$

所以似然比统计量

$$\lambda_3 = \frac{|\Sigma_0^{-1}\mathbf{A}|^{n/2}}{[\text{tr}(\Sigma_0^{-1}\mathbf{A})/p]^{np/2}} \quad (1-60)$$

或等价于

$$W = (\lambda_3)^{2/n} = \frac{|\Sigma_0^{-1}\mathbf{A}| P^p}{[\text{tr}(\Sigma_0^{-1}\mathbf{A})]^p} \quad (1-61)$$

当样本容量 n 很大, 在 H_0 为真时有以下近似分布:

$$-\left((n-1) - \frac{2p^2 + p + 2}{6p}\right) \ln W \text{ 近似为 } \chi^2\left(\frac{p(p+1)}{2} - 1\right)$$

2. 多总体协方差阵的检验

设有 k 个总体 $N_p(\mu^{(t)}, \Sigma_t) (t=1, \dots, k)$, $\mathbf{X}_{(a)}^{(t)} (t=1, \dots, k; a=1, \dots, n)$ 为来自第 t 个总体 $N_p(\mu^{(t)}, \Sigma_t)$ 的随机样本, 记 $n = \sum_{i=1}^k n_i$ 。检验 $H_0: \Sigma_1 = \Sigma_2 = \dots = \Sigma_k \stackrel{\text{def}}{=} \Sigma$, $H_1: \Sigma_1, \Sigma_2, \dots, \Sigma_k$ 不全相等。

样本 $\mathbf{X}_{(a)}^{(t)}$ 的似然函数为

$$L(\mu^{(1)}, \Sigma_1, \dots, \mu^{(k)}, \Sigma_k) = \prod_{t=1}^k L(\mu^{(t)}, \Sigma_k) \quad (1-62)$$

似然比统计量 λ_4 为

$$\lambda_4 = \max_{\mu^{(i)}, \Sigma > \mathbf{O}} L(\mu^{(1)}, \dots, \mu^{(k)}, \Sigma) \bigg/ \max_{\mu^{(i)}, \Sigma_i > \mathbf{O}} L(\mu^{(1)}, \Sigma_1, \dots, \mu^{(k)}, \Sigma_k) \quad (1-63)$$

从而有

$$\lambda_4 \text{ 表示式的分母} = (2\pi)^{-np/2} e^{-np/2} \prod_{t=1}^k \left| \frac{\mathbf{A}_t}{n_t} \right|^{-n_t/2} \quad (1-64)$$

$$\lambda_4 \text{ 表示式的分子} = (2\pi)^{-np/2} e^{-np/2} \left| \frac{\mathbf{A}}{n} \right|^{-n/2} \quad (1-65)$$

(其中 $\mathbf{A} = \mathbf{A}_1 + \dots + \mathbf{A}_t$)

则似然比统计量 λ_4 为

$$\lambda_4 = \left| \frac{\mathbf{A}}{n} \right|^{-n/2} \bigg/ \prod_{t=1}^k \left| \frac{\mathbf{A}_t}{n_t} \right|^{-n_t/2} \quad (1-66)$$

根据无偏性的要求进行修正, 将 λ_4 中的 n_i 用 $n_i - 1$ 替代, n 用 $n - k$ 替代。然后对 λ_4 取对数, 可得到统计量

$$M = -2 \ln \lambda_4^* = (n - k) \ln \left| \frac{\mathbf{A}}{n - k} \right| - \sum_{t=1}^k (n_t - 1) \ln \left| \frac{\mathbf{A}_t}{n_t - 1} \right| \quad (1-67)$$

当样本容量 n 很大时, 在 H_0 为真时 M 有以下近似分布:

$$(1 - d)M = -2(1 - d) \ln \lambda_4^* \sim \chi^2(f) \quad (1-68)$$

其中

$$f = \frac{1}{2} p(p+1)(k-1) \quad (1-69)$$

$$d = \begin{cases} \frac{2p^2 + 3p - 1}{6(p+1)(k-1)} \left[\prod_{i=1}^k \frac{1}{n_i - 1} - \frac{1}{n-k} \right], & \text{当 } n_i \text{ 不全相等} \\ \frac{(2p^2 + 3p - 1)(k+1)}{6(p+1)(n-k)}, & \text{当 } n_i \text{ 全相等} \end{cases} \quad (1-70)$$

3. 多个正态总体的均值向量和协方差阵同时检验

设有 k 个总体 $n = \sum_{t=1}^k n_t$ ($t=1, \dots, k$), $\mathbf{X}_{(\alpha)}^{(t)}$ ($t=1, \dots, k; \alpha=1, \dots, n_t$) 为来自第 t 个总体 $N_p(\mu^{(t)}, \Sigma_t)$ 的随机样本。检验

$$H_0: \mu^{(1)} = \mu^{(2)} = \dots = \mu^{(k)} \text{ 且 } \Sigma_1 = \Sigma_2 = \dots = \Sigma_k,$$

$$H_1: \mu^{(i)} (i=1, \dots, k) \text{ 或 } \Sigma_i (i=1, \dots, k) \text{ 至少有一对不相等。}$$

记

$$\begin{aligned} \bar{\mathbf{X}}^{(t)} &= \frac{1}{n_t} \sum_{j=1}^{n_t} \mathbf{X}_{(j)}^{(t)}, \quad \bar{\mathbf{X}} = \frac{1}{n} \sum_{t=1}^k \sum_{j=1}^{n_t} \mathbf{X}_{(j)}^{(t)}, \quad n = \sum_{t=1}^k n_t \\ \mathbf{A}_t &= \sum_{j=1}^{n_t} (\mathbf{X}_{(j)}^{(t)} - \bar{\mathbf{X}}^{(t)}) (\mathbf{X}_{(j)}^{(t)} - \bar{\mathbf{X}}^{(t)})^T, \quad \mathbf{A} = \sum_{t=1}^k \mathbf{A}_t \\ \mathbf{T} &= \sum_{t=1}^k \sum_{j=1}^{n_t} (\mathbf{X}_{(j)}^{(t)} - \bar{\mathbf{X}}) (\mathbf{X}_{(j)}^{(t)} - \bar{\mathbf{X}})^T \\ &= \mathbf{A} + \sum_{t=1}^k n_t (\bar{\mathbf{X}}^{(t)} - \bar{\mathbf{X}}) (\bar{\mathbf{X}}^{(t)} - \bar{\mathbf{X}}) \end{aligned} \quad (1-71)$$

则检验以上假设 H_0 的似然比统计量为

$$\lambda_5 = \frac{\prod_{t=1}^k |\mathbf{A}_t|^{n_t/2}}{|\mathbf{T}|^{n/2}} \cdot \frac{n^{np/2}}{\prod_{t=1}^k n_t^{n_t p/2}} \quad (1-72)$$

若用 Δ 表示当协方差阵均相同时检验 k 个总体均值向量是否相等的似然比估计量, 将发现这里的似然比统计量 $\lambda_5 = \Delta \cdot \lambda_4$. 在实际应用中我们采用类似的修正方法, 在 λ_5 中用 $n_t - 1$ 代替 n_t , 用 $n - k$ 代替 n . 修正后的统计量记为 λ_5^* ,

$$\lambda_5^* = \frac{\prod_{t=1}^k |\mathbf{A}_t|^{n_t/2}}{|\mathbf{T}|^{n/2}} \cdot \frac{(n-k)^{\frac{(n-k)p}{2}}}{\prod_{t=1}^k (n_t-1)^{\frac{(n_t-1)p}{2}}} \quad (1-73)$$

当样本容量 n 很大, 在 H_0 为真时 λ_5^* 有以下近似分布:

$$-2(1-b) \ln \lambda_5^* \sim \chi^2(f) \quad (1-74)$$

其中

$$f = \frac{1}{2}p(p+3)(k-1) \tag{1-75}$$

$$b = \left(\sum_{i=1}^k \frac{1}{n_i - 1} - \frac{1}{n - k} \right) \left(\frac{2p^2 + 3p - 1}{6(p+3)(k-1)} \right) - \frac{p - k + 2}{(n - k)(p + 3)} \tag{1-76}$$

1.3.4 多元正态分布均值向量和协方差阵检验的上机实现

本书上机实现主要依靠 SPSS 19.0 版本来实现,下面利用 6.3 节中从《国际统计年鉴》搜集到的 2011 年 131 个国家的人文发展指数数据和类别数据,介绍多元正态分布均值向量和协方差阵检验的上机实现。在上面的数据中,不同的类别可以看作是不同的总体,因此,131 个国家的数据分别来自 4 个总体,下面尝试对 4 个不同类别的国家人文发展状况进行比较。

1. 单变量正态性检验

在进行比较分析之前,首先要对各变量是否遵从多元正态分布进行检验。然而遗憾的是,多元正态性检验在常见的统计软件中并不容易实现。在实际工作中,人们往往借助于考察每一个变量的结果来对向量的分布做出判断;并且,当数据量较大,且没有明显的证据表明所得数据不遵从多元正态时,通常认为数据来自多元正态总体。SPSS 软件提供了对单变量进行正态性检验的功能。



图 1-1 单变量正态性检验选择渠道

对上面的数据,选择“分析”→“描述统计”→“探索”命令(见图 1-1),进入“探索”对话框(见图 1-2),可以看到反映国家人文发展状况的所有变量名及变量标签均出现在左边的列表框中,将人文发展指数、出生时预期寿命、预期受教育年限、平均受教育年限、教育指数、收入指数、健康指数、性别不平等指数、预期寿命指数 9 个变量选入“因变量列表”框中,单击右侧的“绘制”按钮进入图形对话框(见图 1-3),选中“带检验的正态图”复选框,单击“继续”、“确定”按钮,则可以得到有关正态性检验的结果(见表 1-1)。



图 1-2 “探索”对话框(因变量选择)

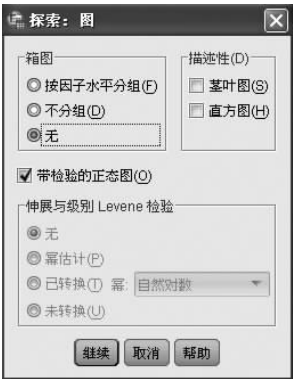


图 1-3 “探索: 图”对话框

表 1-1 正态性检验表

类 别	Kolmogorov-Smirnov ^①			Shapiro-Wilk		
	统计量	df	Sig.	统计量	df	Sig.
人文发展指数	0.100	131	0.003	0.949	131	0.000
出生时预期寿命	0.177	131	0.000	0.896	131	0.000
受教育年限预期	0.063	131	0.200 ^②	0.979	131	0.040
受教育年限平均	0.080	131	0.038	0.966	131	0.002
教育指数	0.076	131	0.062	0.965	131	0.002
收入指数	0.072	131	0.093	0.971	131	0.006
健康指数	0.176	131	0.000	0.896	131	0.000
性别不平等指数	0.084	131	0.024	0.958	131	0.000
预期寿命指数	0.177	131	0.000	0.896	131	0.000

注: ① Lilliefors 显著水平修正。

② 这是真实显著水平的下限。

表 1-1 给出了对每一个变量进行正态性检验的结果,根据 Kolmogorov-Smirnov 统计量的 Sig. 值可以看到,人文发展指数、出生时预期寿命、平均受教育年限、健康指数、性别不平等指数、预期寿命指数均明显不遵从正态分布。因此,在下面的分析中,我们只对预期受教育年限、教育指数、收入指数这 3 个指标进行比较,并认为这 4 个变量组成的向量遵从正态分布(尽管事实上也许并非如此)。进一步,我们将类别变量放入因子列表,对预期受教育年限、教育指数、收入指数这 3 个指标进行正态性检验(见图 1-4),运行结果见表 1-2。



图 1-4 “探索”对话框(因子列表选择)

结果显示,除收入指数中的第 2 类和第 4 类数据不服从正态分布外,其余类别及预期受教育年限、教育指数的各类别均服从正态分布。

表 1-2 带分类变量的正态性检验

类 别		Kolmogorov-Smirnov ^①			Shapiro-Wilk		
		统计量	df	Sig.	统计量	df	Sig.
受教育年限预期	1	0.107	36	0.200 ^②	0.967	36	0.347
	2	0.104	31	0.200 ^②	0.951	31	0.164
	3	0.098	32	0.200 ^②	0.982	32	0.867
	4	0.097	32	0.200 ^②	0.966	32	0.391
教育指数	1	0.124	36	0.178	0.965	36	0.306
	2	0.085	31	0.200 ^②	0.977	31	0.711
	3	0.103	32	0.200 ^②	0.949	32	0.133
	4	0.076	32	0.200 ^②	0.985	32	0.923

续表

类 别		Kolmogorov-Smirnov ^①			Shapiro-Wilk		
		统计量	df	Sig.	统计量	df	Sig.
收入指数	1	0.116	36	0.200 ^②	0.962	36	0.249
	2	0.159	31	0.045	0.926	31	0.035
	3	0.078	32	0.200 ^②	0.988	32	0.968
	4	0.203	32	0.002	0.907	32	0.009

注：① Lilliefors 显著水平修正。
② 这是真实显著水平的下限。

2. 多元正态分布均值与方差的检验

SPSS 的一般线性模型模块可以完成多元正态分布有关均值与方差的检验。依次选择“分析”→“一般线性模型”→“多变量”命令进入“多变量”对话框（见图 1-5），将预期受教育年限、教育指数、收入指数这 3 个指标选入“因变量”列表框，将类别选入“固定因子”列表框，单击“确定”按钮，则可以得到表 1-3～表 1-5 的结果。

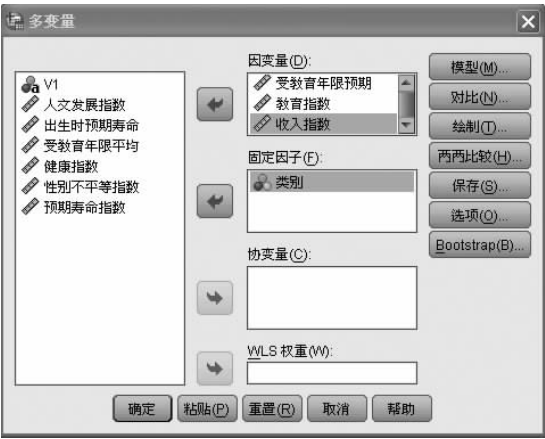


图 1-5 “多变量”对话框

表 1-3 主体间因子

类别	N	类别	N
1	36	3	32
2	31	4	32

表 1-4 多变量检验^③

效 应		值	<i>F</i>	假设 df	误差 df	Sig.
截距	Pillai 的跟踪	0.994	6817.051 ^①	3.000	125.000	0.000
	Wilks 的 lambda	0.006	6817.051 ^①	3.000	125.000	0.000
	Hotelling 的跟踪	163.609	6817.051 ^①	3.000	125.000	0.000
	Roy 的最大根	163.609	6817.051 ^①	3.000	125.000	0.000
类别	Pillai 的跟踪	0.928	18.948	9.000	381.000	0.000
	Wilks 的 lambda	0.077	62.961	9.000	304.368	0.000
	Hotelling 的跟踪	11.857	162.927	9.000	371.000	0.000
	Roy 的最大根	11.852	501.727 ^②	3.000	127.000	0.000

注：① 精确统计量。
② 该统计量是 *F* 的上限，它产生了一个关于显著性级别的下限。
③ 设计：截距＋类别。

表 1-5 主体间效应的检验

源	因变量	Ⅲ型平方和	df	均方	<i>F</i>	Sig.
校正模型	受教育年限预期	1002.368 ^①	3	334.123	175.307	0.000
	教育指数	4.402 ^②	3	1.467	181.968	0.000
	收入指数	3.921 ^③	3	1.307	269.109	0.000
截距	受教育年限预期	19985.559	1	19985.559	10485.975	0.000
	教育指数	53.572	1	53.572	6644.339	0.000
	收入指数	44.942	1	44.942	9253.056	0.000
类别	受教育年限预期	1002.368	3	334.123	175.307	0.000
	教育指数	4.402	3	1.467	181.968	0.000
	收入指数	3.921	3	1.307	269.109	0.000
误差	受教育年限预期	242.053	127	1.906		
	教育指数	1.024	127	0.008		
	收入指数	0.617	127	0.005		
总计	受教育年限预期	21610.800	131			
	教育指数	60.274	131			
	收入指数	50.568	131			