SXU Report

作业

第四组

山西大学

3-8 题目

- 假定人体尺寸有这样的一般规律：身高 (X_1), 胸围 (X_2) 和上半臂围 (X_3) 的平均尺寸比例是 $6:4:1$. 假设 $X_{(a)}(a=1,\ldots,n)$ 为来自总体 $X=(X_1,X_2,X_3)'$ 的随机样本，并设 $X\sim N_3(\mu,\Sigma)$. 试利用表 3.4 中男婴这一组数据来检验其身高、胸围和上半臂围这三个尺寸 (变量) 是否符合这一规律 (写出假设 H_0, 并导出检验统计量).

表 3.4 婴儿体格测量数据

某地区农村两周岁婴儿的体格测量数据

性别	身高 (X_1)	胸围 (X_2)	上半臂围 (X_3)
男	78	60.6	16.5
	76	58.1	12.5
	92	63.2	14.5
	81	59.0	14.0
	81	60.8	15.5
	84	59.5	14.0
女	80	58.4	14.0
	75	59.2	15.0
	78	60.3	15.0
	75	57.4	13.0
	79	59.5	14.0
	78	58.1	14.5
	75	58.0	12.5
	64	55.5	11.0
	80	59.2	12.5

3-8 答案

- 先将题目中的条件转换为线性形式，然后进行假设

$$\boldsymbol{C} = \begin{pmatrix} 1 & 0 & -6 \\ 0 & 1 & -4 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$$

$$H_0 : \boldsymbol{C\mu} = \mathbf{0}$$

$$H_1 : \boldsymbol{C\mu} \neq \mathbf{0}$$

3-8 答案

- 对于这种多元正态总体间均值向量检验，而又不知道总体的均值向量和协方差矩阵的问题，一般采用霍特林统计量 T^2 进行假设检验

$$T^2 = n\left(\boldsymbol{C}\bar{\boldsymbol{x}}\right)^\top \left(\boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^\top\right)^{-1} \left(\boldsymbol{C}\bar{\boldsymbol{x}}\right)$$

3-8 答案

- 先计算均值向量 μ，这里并不知道总体均值，所以我们用样本均值 $\bar{x}$ 来代替 $C\mu$

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \end{pmatrix} = \begin{pmatrix} 82 \\ 60.2 \\ 14.5 \end{pmatrix}$$

3-8 答案

- 再计算均值向量总体协方差矩阵，同理我们用样本的无偏协方差矩阵来替代

3-8 答案

- 定义离差矩阵 $\boldsymbol{A}$（$n \times p$ 阶）：

$$\boldsymbol{A}_{n\times p} = \begin{pmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2 & \cdots & x_{2p}-\bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & \cdots & x_{np}-\bar{x}_p \end{pmatrix}$$

- 离差交叉乘积和矩阵：

$$\boldsymbol{A}^\top \boldsymbol{A} = \begin{pmatrix} \sum_{i=1}^n (x_{i1}-\bar{x}_1)^2 & \dots & \sum_{i=1}^n (x_{i1}-\bar{x}_1)(x_{ip}-\bar{x}_p) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^n (x_{ip}-\bar{x}_p)(x_{i1}-\bar{x}_1) & \dots & \sum_{i=1}^n (x_{ip}-\bar{x}_p)^2 \end{pmatrix}$$

- 无偏样本协方差矩阵：

$$\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{A}^\top \boldsymbol{A}$$

3-8 答案

- 带入题目中数值得到

$$\boldsymbol{S} = \frac{1}{5}\boldsymbol{A}^{\top}\boldsymbol{A} = \begin{pmatrix} 31.6 & 8.04 & 0.50 \\ 8.04 & 3.172 & 1.31 \\ 0.50 & 1.31 & 1.90 \end{pmatrix}$$

3-8 答案

- 霍特林统计量可以转化为更常见的 F 统计量, 这样更方便计算
$$F = \frac{n-q}{q(n-1)} T^2 \sim F(q, n-q)$$
q 代表约束个数，n 代表样本量，带入题目数值得到 $F = 11.87 > F_{0.05}(2,4) = 6.94$，因此拒绝原假设 H_0，认为男婴三项指标均值不满足 $6:4:1$ 的比例约束。

3-9 题目

- 对单个 p 元正态总体 $N_p(\mu, \Sigma)$ 协方差阵的检验问题，试用似然比原理导出检验 $H_0 : \Sigma = \Sigma_0$ 的似然比统计量及分布.

3-9 答案

- 似然比定义为有约束条件下的似然函数最大值与无约束条件下似然函数最大值之比。
- 似然比检验的思想是：如果参数约束是有效的，那么加上这样的约束不应该引起似然函数最大值的大幅度降低。也就是说似然比检验的实质是在比较有约束条件下的似然函数最大值与无约束条件下似然函数最大值。
- 设无约束条件下似然函数最大值之比为 L_0，无约束条件下为 L_1，似然比为 $\lambda = \frac{L_0}{L_1}$，我们需要对这个统计量进行检验。

3-9 答案

- 假设有 n 个样本，先计算无约束条件下（或者说备则假设下）的极大似然估计

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\Sigma} = S = \frac{1}{n}A$$

多元正态总体的似然函数为

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu})\right\}$$

3-9 答案

- 我们先来带入数值计算指数项，这里用到矩阵迹的性质：对于任意向量 u_i 和矩阵 B，有
  $\sum_{i=1}^{n} u_i^T B u_i = \operatorname{tr}\left(B \sum_{i=1}^{n} u_i u_i^T\right)$
  令 $u_i = X_i - \bar{X}$，$B = A^{-1}$，则 $\sum_{i=1}^{n} u_i u_i^T = A$（离差阵的定义），因此：
  $\sum_{i=1}^{n}(X_i - \bar{X})^T A^{-1}(X_i - \bar{X}) = \operatorname{tr}(A^{-1}A) = \operatorname{tr}(I_p) = p$
  所以指数项次数为
  $-\frac{1}{2}n \cdot \sum_{i=1}^{n}(X_i - \bar{X})^T A^{-1}(X_i - \bar{X}) = -\frac{1}{2}np$

3-9 答案

- 接着计算行列式，因为 $\hat{\Sigma} = \frac{1}{n}A$，利用行列式性质 $|kA| = k^p|A|$（A 是 p 阶矩阵）:
  $|\hat{\Sigma}|^{-\frac{n}{2}} = \left(\left(\frac{1}{n}\right)^p |A|\right)^{-\frac{n}{2}} = n^{\frac{np}{2}}|A|^{-\frac{n}{2}}$

3-9 答案

- 所以我们最后得到

$$L_1(\hat{\mu}, \hat{\Sigma}) = (2\pi)^{-\frac{np}{2}} n^{\frac{np}{2}} |A|^{-\frac{n}{2}} e^{-\frac{np}{2}}$$

3-9 答案

- 接着来计算有约束的极大似然估计，与上述类似，在 H_0 下，Σ 固定为已知正定阵 Σ_0，仅需估计 μ，仍为 $\bar{X}$。代入似然函数得 H_0 下的似然极大值：
$L_0(\hat{\mu}, \Sigma_0) = (2\pi)^{-\frac{np}{2}} |\Sigma_0|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Sigma_0^{-1} A)\right\}$

3-9 答案

- 最后代入两个极大值并化简：

$$\lambda = \left(\frac{|A/n|}{|\Sigma_0|}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\left(\mathrm{tr}(\Sigma_0^{-1}A) - np\right)\right\}$$
$$= \left(\frac{|S|}{|\Sigma_0|}\right)^{\frac{n}{2}} \exp\left\{-\frac{n}{2}\left(\mathrm{tr}(\Sigma_0^{-1}S) - p\right)\right\}$$

取对数似然比便于检验：

$$-2\ln\lambda = n\left[\mathrm{tr}(\Sigma_0^{-1}S) - \ln\left|\Sigma_0^{-1}S\right| - p\right]$$

而威尔克斯定理给出大样本情况下，这个统计量服从一个卡方分布，卡方分布的参数为无约束的自由度减去约束后的自由度，这里的假设给定了协方差矩阵，所以协方差矩阵参数的个数就是卡方分布的参数，故
$-2\ln\lambda \sim \chi^2\left(\frac{p(p+1)}{2}\right)$

1. 3-8
2. 3-9
3. 3-11
4. 3-12

3-11 题目

- 表 3.4 给出 15 名两周岁婴儿的身高 (X_1), 胸围 (X_2) 和上半臂围 (X_3) 的测量数据。
  假设男婴的测量数据 $X_{(a)}(a=1,\ldots,6)$ 为来自总体 $N_3(\mu^{(1)},\Sigma)$ 的随机样本; 女婴的测量数据 $Y_{(a)}(a=1,\ldots,9)$ 为来自总体 $N_3(\mu^{(2)},\Sigma)$ 的随机样本。
  试利用表 3.4 中的数据检验:

$$H_0:\mu^{(1)}=\mu^{(2)}\quad(\alpha=0.05).$$

3-11 答案

- 与 3-8 题的理由一样，这里我们选择霍特林统计量 T^2 来进行假设检验，这里我们用合并协方差矩阵来两个样本的均值向量的差来计算，这样就能体现出两个样本均值向量间的差异，幸运的是这里两样本来自两个独立的正态总体，计算比较方便

3-11 答案

- T^2 的公式为 $T^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$
  我们先计算两个样本的均值向量，$\bar{\mathbf{x}}_1$ 和 $\bar{\mathbf{x}}_2$

3-11 答案

- 再分别计算两个样本的离差阵，再相加，最后除以两组自由度的和即可
  1. 合并离差阵 $\boldsymbol{A}_p = \boldsymbol{A}_1 + \boldsymbol{A}_2$
  2. 合并协方差矩阵 $\boldsymbol{S}_p = \frac{\boldsymbol{A}_p}{n_1+n_2-2}$

  其实可以理解成合并协方差矩阵是两个样本的协方差矩阵的加权平均，$\boldsymbol{S}_p = \frac{(n_1-1)\boldsymbol{S}_1+(n_2-1)\boldsymbol{S}_2}{n_1+n_2-2}$

3-11 答案

- 接着计算一下 ${A_p}^{-1}$ 来得到 ${S_p}^{-1}$

$$A_p = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$$

余子式矩阵：

$$\begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}$$

3-11 答案

- $$A_{11} = \begin{vmatrix} e & f \\ h & i \end{vmatrix} = ei - fh$$

$$A_{12} = -(di - fg) = fg - di$$

伴随矩阵：

$$A^* = \begin{pmatrix} A_{11} & A_{21} & A_{31} \\ A_{12} & A_{22} & A_{32} \\ A_{13} & A_{23} & A_{33} \end{pmatrix}$$

$$A^{-1} = \frac{1}{|A|}A^*$$

3-11 答案

- 最后将 T^2 转换成 F 统计量，
  $F = \frac{(n_1+n_2-p-1)}{p(n_1+n_2-2)} T^2 \sim F(p, n_1+n_2-p-1)$
  经过计算得
  $$F_{0.05}(3, 11) = 3.59$$
  $$F = 1.5 < 3.59$$
  因此我们接受假设 H_0

3-12 题目

地质勘探中，在 A,B,C 三个地区采集了一些岩石，测量其部分化学成分，其数据见表 3.5。假定这三个地区岩石的成分遵从 $N_3(\mu^{(i)}, \Sigma_i)$ $(i = 1, 2, 3)$ $(\alpha = 0.05)$。

1. 检验 $H_0 : \Sigma_1 = \Sigma_2 = \Sigma_3$； $H_1 : \Sigma_1, \Sigma_2, \Sigma_3$ 不全等；
2. 检验 $H_0 : \mu^{(1)} = \mu^{(2)}$， $H_1 : \mu^{(1)} \neq \mu^{(2)}$；
3. 检验 $H_0 : \mu^{(1)} = \mu^{(2)} = \mu^{(3)}$， H_1 : 存在 $i \neq j$，使 $\mu^{(i)} \neq \mu^{(j)}$；
4. 检验三种化学成分相互独立。

表 3.5 岩石部分化学成分数据

地区	SiO_2	FeO	K_2O
A 地区	47.22	5.06	0.10
A 地区	47.45	4.35	0.15
A 地区	47.52	6.85	0.12
A 地区	47.86	4.19	0.17
A 地区	47.31	7.57	0.18
B 地区	54.33	6.22	0.12
B 地区	56.17	3.31	0.15
B 地区	54.40	2.43	0.22
B 地区	52.62	5.92	0.12
C 地区	43.12	10.33	0.05
C 地区	42.05	9.67	0.08
C 地区	42.50	9.62	0.02
C 地区	40.77	9.68	0.04

3-12-(1) 答案

- 对于这种多个多元正态总体协方差阵的齐性检验的问题，通常使用 M 统计量进行检验，为了构建 M 统计量，我们需要计算各个样本分别的协方差矩阵 S_i 和他们的合并协方差矩阵 S_p，然后代入 M 的公式中：
  $M = df \cdot \ln|\boldsymbol{S}_p| - \sum_{i=1}^{3} df_i \cdot \ln|\boldsymbol{S}_i|$
  其中 df 代表自由度

3-12-(1) 答案

- 然后我们需要计算校正因子（1-d），其中 d 为：

$$d=\begin{cases}\dfrac{2p^2+3p-1}{6(p+1)(k-1)}\left(\displaystyle\sum_{i=1}^{k}\frac{1}{n_i-1}-\frac{1}{n-k}\right), & n_i \text{ 不全等},\\ \dfrac{(2p^2+3p-1)(k+1)}{6(p+1)(n-k)}, & n_i \text{ 全相等}.\end{cases}$$

这一题属于第一种情况

3-12-(1) 答案

- (1-d)M 近似服从一个卡方分布，如果在假设下 M 大于这个卡方分布的临界值那我们就拒绝假设，卡方假设自由度为
  $f = \frac{1}{2}p(p+1)(k-1) = \frac{1}{2} \times 3 \times 4 \times 2 = 12$
  $\chi^2 = 2.157 < \chi^2_{0.05}(12) = 21.026$
  故我们可以接受假设 H_0。

3-12-(2) 答案

- 与 3-11 题一致，我们使用霍特林统计量进行检验，最后得到

$$F_{0.05}(3,5) = 5.41$$

$$F = 32.09 > 3.59$$

故我们拒绝 H_0

3-12-(3) 答案

- 我们接受第一问的假设，三个总体的协方差矩阵相同，选用 Wilks' Λ 统计量进行检验

$$\Lambda = \frac{|W|}{|W+B|}$$

其中 W 是组内离差阵（$W=\sum_{i=1}^{3} A_i$），B 是组间离差阵

$$\boldsymbol{B}=\sum_{i=1}^{k} n_i\left(\bar{X}^{(i)}-\bar{X}\right)\left(\bar{X}^{(i)}-\bar{X}\right)^T$$

其中 k 为组数，此题为 3。
$\bar{X}$ 为总样本均值向量：

$$\bar{X}=\frac{n_1\bar{X}^{(1)}+n_2\bar{X}^{(2)}+n_3\bar{X}^{(3)}}{n_1+n_2+n_3}=\frac{5\bar{X}^{(1)}+4\bar{X}^{(2)}+4\bar{X}^{(3)}}{13}$$

其实是一个对样本均值向量的加权平均。

3-12-(3) 答案

- 接下来我们需要将 Λ 转换成更常见的统计量来进行检验，这里有两种做法，根据威尔克斯定理大样本情况下似然比统计量近似服从一个卡方分布，不过这里的样本数目不大，我们可以用另一种方法，将 Λ 转换成 F 统计量，

$$F=\frac{n-k-p+1}{p}\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}$$

带入题目计算得 $F=16.91>F_{0.05}(6,15)=2.79$
故我们拒绝原假设 H_0

3-12-(4) 答案

- 由于三个总体都服从多元正态分布，所以题目可以写为
  $H_0=$ 三个协方差矩阵是对角矩阵
  $H_1=$ 三个协方差矩阵不全是对角矩阵
  我们可以接受第一问的假设，认为三个协方差矩阵相同，计算更方便，当然，也可以单独对三个样本进行检验
  $H_0=$ 协方差矩阵是对角矩阵
  $H_1=$ 协方差矩阵不是对角矩阵

3-12-(4) 答案

- 我们可以选用似然统计来进行检验，根据 3-9 一样的方法，我们可以构造出一个似然统计量 $V = \frac{|S|}{\prod_{i=1}^{p} s_{ii}} = \frac{|A|}{\prod_{i=1}^{p} a_{ii}}$，其中 a_{ii} 是总离差阵 A 的对角元素
  我们可以复用第（3）问的结果，总离差阵 $A = W + B$

3-12-(4) 答案

- 接着我们处理统计量，让它可以近似服从卡方分布，从而用卡方分布的临界值来进行检验
  $\xi = -b\ln V \sim \chi^2(f)$,
  $b = n - \frac{3}{2} - \frac{p^3-\sum p_i^3}{3(p^2-\sum p_i^2)}$，验证变量独立时可以用简化公式
  $b = n - \frac{3}{2} - \frac{p+1}{3}$
  $f = \frac{1}{2}\left[p(p+1) - \sum_{i=1}^{k} p_i(p_i+1)\right]$
  最后得到 $\xi = 26.68 > \chi^2_{0.05}(3) = 7.815$
  故我们拒绝原假设 H_0，认为三种化学成分不相互独立。

Thanks!