《统计方法与机器学习》作业一 方差分析

10215501438 韩弈文

Q1

考虑一个因子有4种不同的水平，在各个水平下，我们进行了6次重复实验。已计算$SS_T=10,\ SS_E=2.5$，请写出完整的ANOVA表。

解：因为此因子有4种不同的水平，所以$a=4$.

因为在各个水平下进行了6次重复实验，所以$m=6$.

因为$SS_T=10,\ SS_E=2.5$，所以$SS_A=SS_T-SS_E=10-2.5=7.5$.

因子的自由度$df_A=a-1=4-1=3$，误差的自由度$df_E=n-a=am-a=4\times6-4=20$.

因子的均方和$MS_A=\frac{SS_A}{a-1}=\frac{7.5}{3}=2.5$，误差的均方和$MS_E=\frac{SS_E}{n-a}=\frac{2.5}{20}=0.125$.

$F_A=\frac{MS_A}{MS_E}=\frac{2.5}{0.125}=20$.

所以可以得到方差分析表为

表1 方差分析表

来源	平方和SS	自由度df	均方和MS	F值
因子A	7.5	3	2.5	20
误差E	2.5	20	0.125	
总和	10	23		

Q2

假设我们有两组独立的数据

第一种：$x_1,x_2,\cdots,x_m$

第二种：$y_1,y_2,\cdots,y_m$

假定$x_i\overset{i.i.d}{\sim}N(\mu_1,\sigma^2)$且$y_i\overset{i.i.d}{\sim}N(\mu_2,\sigma^2)$。其中，$\sigma^2$是未知常数。检验问题为

$$H_0:\mu_1=\mu_2\quad vs\quad H_1:\mu_1\neq\mu_2.$$

用单因子方差分析模型来解决假设检验的问题：

证明：在这种情况下，单因子方差分析模型与二样本独立t检验是等价的。（提示：考虑两个检验统计量分布之间的关系）

解：

- 单因子方差分析

  因子数$a=2$，实验次数为m.

  $x_.=\sum\limits_{j=1}^m x_j,\quad y_.=\sum\limits_{j=1}^m y_j$

  $\bar{x}_.=\frac{x_.}{m},\quad \bar{y}_.=\frac{y_.}{m}$

  记$z=x_.+y_.,\quad \bar{z}=\frac{z}{2m}$

则$SS_A = m[(\bar{x}_. - \bar{z})^2 + (\bar{y}_. - \bar{z})^2]$, $SS_E = \sum\limits_{j=1}^{m} [(x_j - \bar{x}_.)^2 + (y_j - \bar{y}_.)^2]$

$F_A = \frac{SS_A/(2-1)}{SS_E/(2m-2)} = \frac{SS_A}{SS_E/(2m-2)} \sim F(1, 2m-2)$

- 二样本独立t检验

  $s_w^2 = \frac{1}{2m-2} \sum\limits_{j=1}^{m} [(x_j - \bar{x}_.)^2 + (y_j - \bar{y}_.)^2]$

  $t = \frac{\bar{x}_. - \bar{y}_.}{s_w \sqrt{\frac{2}{m}}} \sim t(2m-2)$

对于$F(1, 2m-2)$，随机变量可以写为$F = \frac{X_1}{X_2/(2m-2)}$的形式，其中$X_1 \sim \chi^2(1)$, $X_2 \sim \chi^2(2m-2)$.

对于$t(2m-2)$，随机变量可以写为$t = \frac{X_1}{\sqrt{X_2/(2m-2)}}$的形式，其中$X_1 \sim N(0,1)$, $X_2 \sim \chi^2(2m-2)$.

$t^2 = \frac{X_1^2}{X_2/(2m-2)}$，其中$X_1^2 \sim \chi^2(1)$, $X_2 \sim \chi^2(2m-2)$.

可见$F = t^2$.

对于相同的显著性水平α，$P(F_A > c) = \alpha$, $P(t > c^2) = \alpha$，只需要设置相应的临界值就可以达到相同的显著性水平。

Q3

假设我们有数据如下：

第1组：$y_{11}, y_{12}, \cdots, y_{1m_1}$

第2组：$y_{21}, y_{22}, \cdots, y_{2m_2}$

$\vdots$

第a组：$y_{a1}, y_{a2}, \cdots, y_{am_a}$

注：这组数据中每组的重复次数是不相等的。

- 写出符合此数据的单因子方差分析模型；
- 写出原假设与备择假设；
- 写出检验统计量；
- 写出方差分析表。

符号说明：

- $y_{i.} = \sum\limits_{j=1}^{m_i} y_{ij}$, $i = 1, 2, \cdots, a$
- $\bar{y}_{i.} = \frac{y_{i.}}{m}$, $i = 1, 2, \cdots, a$
- $y_{..} = \sum\limits_{i=1}^{a} y_{i.}$
- $\bar{y}_{..} = \frac{y_{..}}{n}$, $n = \sum\limits_{i=1}^{a} m_i$

模型建立

1. 均值模型

   $y_{ij} = \mu_i + \epsilon_{ij}$, $i = 1, 2, \cdots, a$, $j = 1, 2, \cdots, m_i$

μ_i：因子的第i个水平下的均值

ϵ_{ij}：随机误差，一般$E(\epsilon_{ij})=0$

可以得到$E(y_{ij})=\mu_i,\quad j=1,2,\cdots,m_i$

2. 效应模型

   $\mu_i=\mu+\alpha_i,\quad i=1,2,\cdots,a$

   $y_{ij}=\mu+\alpha_i+\epsilon_{ij},\quad i=1,2,\cdots,a,\quad j=1,2,\cdots,m_i$

   通过设置$\sum\limits_{i=1}^{n}m_i\alpha_i=0$，来解决参数的无法识别问题

模型的前提假设

1. $\epsilon_{ij}\overset{i.i.d}{\sim}N(0,\sigma^2)$，即各总体方差相同
2. $y_{ij}\sim N(\mu_i,\sigma_i^2),\quad i=1,2,\cdots,a$
3. y_{ij}相互独立

假设检验

1. 均值模型

   H_0：$\mu_1=\mu_2=\cdots=\mu_a$

   H_1：$\mu_1,\mu_2,\cdots,\mu_a$不全相等

2. 效应模型

   H_0：$\alpha_1=\alpha_2=\cdots=\alpha_a=0$

   H_1：$\alpha_1,\alpha_2,\cdots,\alpha_a$不全等于0

模型求解

1. 平方和分解公式

$$\begin{aligned}SS_T=&\sum_{i=1}^{a}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_{..})^2\\=&\sum_{i=1}^{a}\sum_{j=1}^{m_i}((\bar{y}_{i.}-\bar{y}_{..})+(y_{ij}-\bar{y}_{i.}))^2\\=&\sum_{i=1}^{a}m_i(\bar{y}_{i.}-\bar{y}_{..})^2+\sum_{i=1}^{a}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_{i.})^2+2\sum_{i=1}^{a}\sum_{j=1}^{m_i}(\bar{y}_{i.}-\bar{y}_{..})(y_{ij}-\bar{y}_{i.})\end{aligned}$$

而$\sum\limits_{j=1}^{m_i}(y_{ij}-\bar{y}_{i.})=y_{i.}-m_i\bar{y}_{i.}=0$，

所以得到

$$SS_T=\sum_{i=1}^{a}m_i(\bar{y}_{i.}-\bar{y}_{..})^2+\sum_{i=1}^{a}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_{i.})^2$$

组间偏差平方和与组内偏差平方和分别为

$$SS_A=\sum_{i=1}^{a}m_i(\bar{y}_{i.}-\bar{y}_{..})^2$$

$$SS_E=\sum_{i=1}^{a}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_{i.})^2$$

2. 求解$\frac{SS_E}{\sigma^2}$的分布

$$
\begin{aligned}
SS_E =&\sum_{i=1}^{a}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_{i.})^2\\
=&\sum_{i=1}^{a}\sum_{j=1}^{m_i}((\mu+\alpha_i+\epsilon_{ij})-\frac{1}{m_i}\sum_{j=1}^{m_i}(\mu+\alpha_i+\epsilon_{ij}))^2\\
=&\sum_{i=1}^{a}\sum_{j=1}^{m_i}(\epsilon_{ij}-\bar{\epsilon}_{i.})^2
\end{aligned}
$$

记$s_i=\frac{1}{m_i-1}\sum\limits_{j=1}^{m_i}(\epsilon_{ij}-\bar{\epsilon}_{i.})^2$

可得$\frac{(m_i-1)s_i}{\sigma^2}\sim\chi^2(m_i-1)$

于是$\frac{\sum\limits_{i=1}^{a}(m_i-1)s_i}{\sigma^2}\sim\chi^2(\sum\limits_{i=1}^{a}(m_i-1))$

也即$\frac{SS_E}{\sigma^2}\sim\chi^2(n-a)$

3. 求解$\frac{SS_A}{\sigma^2}$的分布

$$
\begin{aligned}
SS_A =&\sum_{i=1}^{a}m_i(\bar{y}_{i.}-\bar{y}_{..})^2\\
=&\sum_{i=1}^{a}m_i(\frac{1}{m_i}\sum_{j=1}^{m_i}(\mu+\alpha_i+\epsilon_{ij})-\frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{m_i}(\mu+\alpha_i+\epsilon_{ij}))^2\\
=&\sum_{i=1}^{a}m_i(\alpha_i+\bar{\epsilon}_{i.}-\bar{\epsilon}_{..})^2\\
=&\sum_{i=1}^{a}m_i(\alpha_i^2+(\bar{\epsilon}_{i.}-\bar{\epsilon}_{..})^2+2\alpha_i(\bar{\epsilon}_{i.}-\bar{\epsilon}_{..}))\\
=&\sum_{i=1}^{a}m_i\alpha_i^2+\sum_{i=1}^{a}m_i(\bar{\epsilon}_{i.}-\bar{\epsilon}_{..})^2+\sum_{i=1}^{a}2m_i\alpha_i(\bar{\epsilon}_{i.}-\bar{\epsilon}_{..})
\end{aligned}
$$

当原假设成立时，$\alpha_i=0,\quad i=1,2,\cdots,a$

所以$SS_A=\sum\limits_{i=1}^{a}m_i(\bar{\epsilon}_{i.}-\bar{\epsilon}_{..})^2=\sum\limits_{i=1}^{a}(\sqrt{m_i}\bar{\epsilon}_{i.}-\sqrt{m_i}\bar{\epsilon}_{..})^2.$

而$\frac{\sum\limits_{i=1}^{a}(\sqrt{m_i}\bar{\epsilon}_{i.}-\sqrt{m_i}\bar{\epsilon}_{..})^2}{\sigma^2}\sim\chi^2(a-1).$

得到$\frac{SS_A}{\sigma^2}\sim\chi^2(a-1)$

综上所述，可以得到：

- 单因子方差分析模型：

$$
\begin{cases}
y_{ij}=\mu+\alpha_i+\epsilon_{ij}, & i=1,2,\cdots,a,\quad j=1,2,\cdots,m_i\\
\sum\limits_{i=1}^{r}m_i\alpha_i=0\\
\epsilon_{ij}\text{相互独立，且}\epsilon\sim N(0,\sigma^2)
\end{cases}
$$

- 原假设与备择假设：

$$
H_0:\ \alpha_1=\alpha_2=\cdots=\alpha_a=0\quad v.s.\quad H_1:\ \exists i,\ s.t.\ \alpha_i\neq 0
$$

- 检验统计量：

$$F=\frac{SS_A/(a-1)}{SS_E/(n-a)}$$

$$SS_A=\sum_{i=1}^{a} m_i(\bar{y}_{i.}-\bar{y}_{..})^2$$

$$SS_E=\sum_{i=1}^{a}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_{i.})^2$$

- 方差分析表

表2 $Q3$方差分析表

来源	平方和SS	自由度df	均方和MS	F值
因子A	SS_A	$a-1$	$\frac{SS_A}{a-1}$	$\frac{MS_A}{MS_E}$
误差E	SS_E	$n-a$	$\frac{SS_E}{n-a}$	
总和	SS_A+SS_E	$n-1$		

Q4

有七种人造纤维，每种抽4根测其强度，得每种纤维的平均强度及标准差如下：

组号	均值	标准差
1	6.3	0.81
2	6.2	0.92
3	6.7	1.22
4	6.8	0.74
5	6.5	0.88
6	7.0	0.58
7	7.1	1.05

假设各种纤维的强度服从等方差的正态分布：

- 试问七种纤维强度间有无显著差异（取$\alpha=0.05$）
- 根据第一小问的结果，回答：
    - 若各种纤维之间的强度间无显著差异，则给出平均强度的置信水平为0.95的置信区间
    - 若各种纤维的强度间有显著差异，请进一步在$\alpha=0.05$下进行多重比较，并指出哪种纤维平均强度最大，同时给出该种纤维平均强度的置信水平为0.95的置信区间。

解：$m=4,a=7$

$\bar{y}_{..}=\frac{1}{n}\sum\limits_{i=1}^{a} m\bar{y}_{i.}=\frac{233}{35}$

$SS_A=a\sum\limits_{i=1}^{m}(\bar{y}_{i.}-\bar{y}_{..})^2$

$SS_E=\sum\limits_{i=1}^{a}(m-1)\sigma_i^2$

$SS_A=\frac{488}{175},\quad SS_E=17.2554$

$F=\frac{SS_A/(a-1)}{SS_E/(n-a)}=\frac{488/175/6}{17.2554/21}=0.56562$

通过查表得到$F_{0.95}(6,21)=2.5$

于是可以得到$F < F_{0.95}(6,21)$，所以接受原假设，即七种纤维强度间无显著差异

$t_{1-\frac{\alpha}{2}}(28-1) = 2.0518$

$\hat{\sigma}^2 = \frac{SS_T}{n-1} = \frac{20.04397}{27} = 0.74237$

置信区间为$[\frac{233}{35} - 2.0518 * 0.86161/\sqrt{28}, \frac{233}{35} + 2.0518 * 0.86161/\sqrt{28}]$

即[6.32305，6.99124]