统计方法与机器学习

第一章 方差分析

倪 葎

Dase@ECNU

SCHOOL OF DATA
SCIENCE & ENGINEERING
数据科学与工程学院

目录

目录

模型与假设

- 有以下两组数据
  - 第1组：$y_{11}, y_{12}, \cdots, y_{1m_1}$；
  - 第2组：$y_{21}, y_{22}, \cdots, y_{2m_2}$。
- 模型与假设：y_{ij} 是独立的正态分布，即

$$y_{ij} \sim N(\mu_i, \sigma^2), i = 1,2, j = 1,2, \cdots, m_i$$

模型与假设

$$y_{ij} \sim N(\mu_i, \sigma^2), i = 1,2, j = 1,2, \cdots, m_i$$

- 注意到，
  - 第 i 组数据是独立同分布正态分布随机变量；
  - 第 i 组数据的均值是相同的，即$E(y_{ij}) = \mu_i$；
  - 不同组数据的均值可能不相同，但方差是相同的；
  - 第 i 组共有 m_i 个数据；
  - 样本量为 $n = m_1 + m_2$。

模型与假设

- 检验问题为

$$H_0: \mu_1 = \mu_2 \quad \text{vs} \quad H_1: \mu_1 \neq \mu_2$$

- 提问：两样本独立t检验的模型与假设的本质是什么？

检验方法

- 检验统计量为

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{s_w\sqrt{\frac{1}{m_1} + \frac{1}{m_2}}}$$

- 其中，
  - $\bar{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}$ 表示第 i 组的样本均值；
  - $s_i^2 = \frac{1}{m_i-1}\sum_{j=1}^{m_i}\left(y_{ij} - \bar{y}_i\right)^2$表示第 i 组的样本方差；
  - $s_w^2 = \frac{m_1-1}{m_1+m_2-2}s_1^2 + \frac{m_2-1}{m_1+m_2-2}s_2^2$ 表示合方差。

检验方法

- 在原假设成立时，检验统计量为

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{s_w\sqrt{\frac{1}{m_1} + \frac{1}{m_2}}} \sim t(m_1 + m_2 - 2)$$

- 两样本独立 t 检验由此得名。
- 提问：为什么检验统计量的分布是 t 分布？
- 特别地，当两组样本量相等时，即 $m_1 = m_2 = m$，检验统计量可写为

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{s_w\sqrt{\frac{2}{m}}} \sim t(2(m-1))$$

检验方法

- 取显著性水平 α；
- **拒绝域法**

$$W=\{|t_0|\geq t_{1-\alpha/2}(m_1+m_2-2)\}$$

其中，$t_\alpha(m_1+m_2-2)$ 表示自由度为 m_1+m_2-2 的 t 分布的 α 分位数。

- ***p*值法**

$$p=2P(t\geq|t_0|)$$

其中，t 表示自由度为 m_1+m_2-2 的 t 分布随机变量。如果 $p<\alpha$，那么拒绝原假设。

例题

- **目标**：比较两种为期6周的减肥计划
- **过程**：
  - 确定 $n = 48$ 名志愿者参与本次实验；
  - 随机被分配一种减肥方案，每种减肥方案有 $m = 24$ 名志愿者；
  - 在减肥计划开始前，测量所有志愿者的体重，记为初始体重；
  - 在减肥计划结束后，测量所有志愿者的体重，记为最终体重；
- **问题**：两种减肥计划的效果是否一致？

例题

序号	计划	体重			序号	计划	体重			序号	计划	体重			序号	计划	体重		
		初始	最终	**差异**			初始	最终	**差异**			初始	最终	**差异**			初始	最终	**差异**
1	A	58.0	54.2	-3.80	13	A	72.0	69	-3.00	25	B	58.0	60.1	2.10	37	B	75.0	72.6	-2.40
2	A	60.0	54	-6.00	14	A	72.0	68.4	-3.60	26	B	58.0	56	-2.00	38	B	75.0	69.2	-5.80
3	A	64.0	63.3	-0.70	15	A	72.0	70.9	-1.10	27	B	59.0	57.3	-1.70	39	B	76.0	72.7	-3.30
4	A	64.0	61.1	-2.90	16	A	74.0	69.5	-4.50	28	B	61.0	56.7	-4.30	40	B	76.0	72.5	-3.50
5	A	65.0	62.2	-2.80	17	A	78.0	73.9	-4.10	29	B	63.0	62.4	-0.60	41	B	77.0	77.5	0.50
6	A	66.0	64	-2.00	18	A	80.0	71	-9.00	30	B	63.0	60.3	-2.70	42	B	78.0	72.7	-5.30
7	A	67.0	65	-2.00	19	A	80.0	77.6	-2.40	31	B	63.0	59.4	-3.60	43	B	78.0	76.3	-1.70
8	A	69.0	60.5	-8.50	20	A	82.0	81.1	-0.90	32	B	65.0	62	-3.00	44	B	79.0	73.6	-5.40
9	A	70.0	68.1	-1.90	21	A	83.0	79.1	-3.90	33	B	66.0	64	-2.00	45	B	79.0	72.9	-6.10
10	A	70.0	66.9	-3.10	22	A	85.0	81.5	-3.50	34	B	68.0	63.8	-4.20	46	B	79.0	71.1	-7.90
11	A	71.0	71.6	0.60	23	A	87.0	81.9	-5.10	35	B	68.0	63.3	-4.70	47	B	80.0	81.4	1.40
12	A	72.0	70.5	-1.50	24	A	88.0	84.5	-3.50	36	B	71.0	66.8	-4.20	48	B	80.0	75.7	-4.30

例题

- 令
  - y_{1j} 表示减肥计划 A 六周前后的体重差异；
  - y_{2j} 表示减肥计划 B 六周前后的体重差异；
- 假设

$$y_{ij} \sim N(\mu_i, \sigma^2), i = 1,2, j = 1,2,\cdots,24$$

- 检验问题为

$$H_0: \mu_1 = \mu_2 \quad \text{vs} \quad H_1: \mu_1 \neq \mu_2$$

例题

- 我们可以计算

$$\bar{y}_1 = -3.300, s_1^2 = 5.0183;$$
$$\bar{y}_2 = -3.1125, s_2^2 = 5.7072;$$

- 合方差为

$$s_w^2 = \frac{1}{2}s_1^2 + \frac{1}{2}s_2^2 = 5.3627$$

- 检验统计量为

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{s_w\sqrt{2/m}} = -0.2805$$

例题

- 取显著性水平 $\alpha = 0.05$；
- **拒绝域法**

$$W = \left\{|t_0| \geq t_{1-\alpha/2}(2(m-1))\right\} = \{|t_0| \geq 2.0129\}$$

- 结论：无法拒绝原假设；
- 可以认为，这两种减肥计划的效果是一致的。

总结与思考

- 两样本独立 t 检验用于比较两个总体均值是否相等的问题；
- 检验统计量

$$\frac{\bar{y}_1 - \bar{y}_2}{s_w\sqrt{\frac{1}{m_1} + \frac{1}{m_2}}} \sim t(m_1 + m_2 - 2)$$

- 临界值法和 p 值法均可用于得到结论。
- 提问：如何比较三个总体均值是否相等？

目录

定义：数据结构

水平	观测到的响应变量				总和	均值
1	y_{11}	y_{12}	$\cdots$	y_{1m}	$y_{1\cdot}$	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	$\cdots$	y_{2m}	$y_{2\cdot}$	$\bar{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{am}	$y_{a\cdot}$	$\bar{y}_{a\cdot}$
求和					$y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot}$

- y_{ij} 表示在第 i 个**水平**下观测到的第 j 个**响应变量**。

因子：引发响应变量大小变化的影响因子
称因子的一种**取值**为一个**水平**或**处理**。

响应变量：所关心的随机变量

定义：数据结构

水平	观测到的响应变量				总和	均值
1	y_{11}	y_{12}	$\cdots$	y_{1m}	$y_{1\cdot}$	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	$\cdots$	y_{2m}	$y_{2\cdot}$	$\bar{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{am}	$y_{a\cdot}$	$\bar{y}_{a\cdot}$
求和					$y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot}$

- y_{ij} 表示在第 i 个水平下观测到的第 j 个响应变量。
- $y_{i\cdot} = \sum_{j=1}^{m} y_{ij}$ 表示在第 i 个水平下响应变量的总和。
- $\bar{y}_{i\cdot} = m^{-1} y_{i\cdot}$表示在第 i 个水平下响应变量的均值。

重复次数：在因子每个水平下，随机变量的个数，记为 m

定义：数据结构

水平	观测到的响应变量				总和	均值
1	y_{11}	y_{12}	$\cdots$	y_{1m}	$y_{1\cdot}$	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	$\cdots$	y_{2m}	$y_{2\cdot}$	$\bar{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{am}	$y_{a\cdot}$	$\bar{y}_{a\cdot}$
求和					$y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot}$

- y_{ij} 表示在第 i 个水平下观测到的第 j 个响应变量。
- $y_{\cdot\cdot} = \sum_{i=1}^{a}\sum_{j=1}^{m} y_{ij}$ 表示所有响应变量的总和。
- $\bar{y}_{\cdot\cdot} = n^{-1}y_{\cdot\cdot}$表示所有响应变量的均值。

样本量：在因子所有水平下，随机变量的个数总和，记为 $n = am$

定义：数据结构

水平	观测到的响应变量				总和	均值
1	y_{11}	y_{12}	$\cdots$	y_{1m}	$y_{1\cdot}$	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	$\cdots$	y_{2m}	$y_{2\cdot}$	$\bar{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{am}	$y_{a\cdot}$	$\bar{y}_{a\cdot}$
求和					$y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot}$

- 特例（两样本独立 t 检验）
- 响应变量：减肥前后的体重差；
- 因子：减肥计划，有 $a = 2$ 个水平；
- 重复次数：每组 $m = 24$ 个数据；
- 样本量： $n = am = 48$ 个数据。

模型：均值模型

- 模型的一般形式为

$$y_{ij} = \mu_i + \varepsilon_{ij}, \qquad i = 1,2,\cdots,a, j = 1,2,\cdots,m$$

  - y_{ij} 表示在第 i 个水平下观测到的第 j 个响应变量；
  - μ_i 表示因子的第 i 个水平下的均值；
  - ε_{ij} 表示随机误差；

模型：**效应模型**

- 令

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

- 模型的另一种形式为

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad i = 1,2,\cdots,a, j = 1,2,\cdots,m$$

  - y_{ij} 表示在第 i 个水平下观测到的第 j 个响应变量；
  - μ 表示总体均值；
  - α_i 表示在第 i 个水平下的效应；
  - ε_{ij} 表示随机误差；

模型

- 均值模型为 $y_{ij} = \mu_i + \varepsilon_{ij},\ i = 1,2,\cdots,a, j = 1,2,\cdots,m$
- 效应模型为 $y_{ij} = \mu + \alpha_i + \varepsilon_{ij},\ i = 1,2,\cdots,a, j = 1,2,\cdots,m$
- 相比于均值模型，效应模型参数个数有增加；
- 通常，需要对参数有约束

$$\sum_{i=1}^{a} \alpha_i = 0$$

- 提问：为什么需要这个约束？这个约束是唯一的吗？
- 这两个模型均仅考虑了一个因子，称为**单因子方差分析模型**。

假设

- 对随机误差的假设

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

  - 随机误差独立同分布；
  - 随机误差均值为零，方差为σ^2。
- 这表明：在不同水平下，响应变量的波动大小是一致的。
- 响应变量是独立的，且为正态分布随机变量，即

$$y_{ij} \sim N(\mu + \alpha_i, \sigma^2)$$

总结

- 单因子方差分析模型有两个形式：均值模型和效应模型。
- 均值模型为

$$y_{ij} = \mu_i + \varepsilon_{ij}, \qquad i = 1,2,\cdots,a, j = 1,2,\cdots,m$$

- 效应模型为

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad i = 1,2,\cdots,a, j = 1,2,\cdots,m$$

且满足

$$\sum_{i=1}^{a} \alpha_i = 0$$

目录

假设

- **均值模型**为

$$y_{ij} = \mu_i + \varepsilon_{ij}, \qquad i = 1,2,\cdots,a, j = 1,2,\cdots,m$$

- 对应的假设问题为

$$H_0\colon \mu_1 = \mu_2 = \cdots = \mu_a$$

vs H_1:存在两种水平 i 和 i' 下均值不相等，即 $\mu_i \neq \mu_{i'}$

- **效应模型**为

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad i = 1,2,\cdots,a, j = 1,2,\cdots,m$$

- 对应的假设问题为

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

vs H_1:存在第 i 个水平不为零，即 $\alpha_i \neq 0$

这两种模型和假设是等价的。

检验统计量

- 两样本独立 t 检验的检验统计量为

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{1}{m}(s_1^2 + s_2^2)}}$$

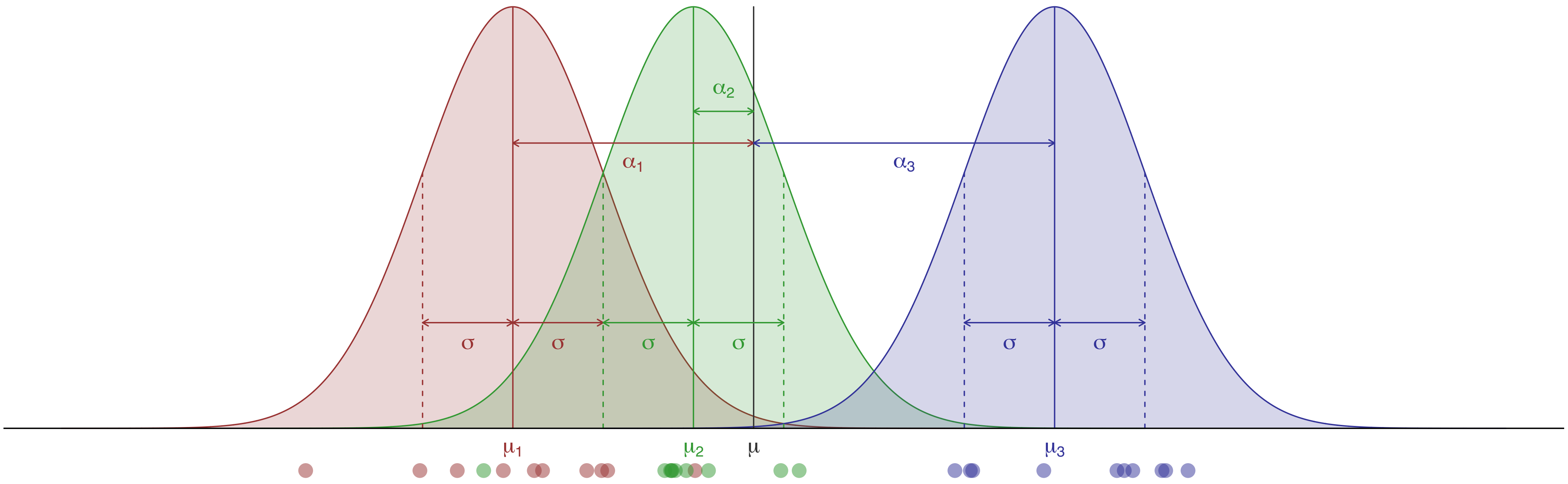
分母：组内差异

分子：组间差异

- 本质上，比较两组样本均值差异与数据波动的大小。
- 如果比值比较大，即两组样本均值的差异比数据波动大得多，那么数据中有足够的证据支撑这两组数据的均值是不一致的。
- 提问：多个总体均值比较时，统计量应该是怎样的？

检验统计量

- 特例：因子水平数取 $a = 3$。



检验统计量

- **重要性质**：平方和分解公式为

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

检验统计量

- **重要性质**：平方和分解公式为

$$\sum_{i=1}^{a}\sum_{j=1}^{m}\left(y_{ij}-\bar{y}_{\cdot\cdot}\right)^2 = m\sum_{i=1}^{a}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2 + \sum_{i=1}^{a}\sum_{j=1}^{m}\left(y_{ij}-\bar{y}_{i\cdot}\right)^2$$

- 总偏差平方和 SS_T，与因子无关。
- 组间偏差平方和 SS_A，表示不同水平下数据均值与所有数据总体均值之间的差异，既包含因子取不同水平引起的数据差异，又包含数据波动对其影响。
- 组内偏差平方和 SS_E，表示同一水平下数据与其均值的差异，是由于数据波动引起的。

检验统计量

- **重要性质** $SS_T = SS_A + SS_E$ **的证明**

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

$$=m\sum_{i=1}^{a}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2+\sum_{i=1}^{a}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})^2+\boxed{2\sum_{i=1}^{a}\sum_{j=1}^{m}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})(y_{ij}-\bar{y}_{i\cdot})}$$

其中交叉项为零, 这是因为

$$\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})=y_{i\cdot}-m\bar{y}_{i\cdot}=y_{i\cdot}-y_{i\cdot}=0$$

检验统计量

- **重要性质**：平方和分解公式为

$$SS_T = SS_A + SS_E$$

- 给定一组数据，总偏差平方和 SS_T 是不会变的。
- 如果原假设成立，那么 SS_A 仅受到数据波动的影响。
- 一个直观的想法是：构造一个比值

$$\frac{SS_A}{SS_T}$$

- 如果比值越大，那么越有证据支撑备择假设；
- 反之，认为原假设更为合理。

检验统计量

- **重要性质**：平方和分解公式为

$$SS_T = SS_A + SS_E$$

- $\frac{SS_A}{SS_E}$ 随 $\frac{SS_A}{SS_T}$ 增大而增大的。
- 在单因子方差分析模型中，所使用的检验统计量是

$$\frac{SS_A}{SS_E}$$

- 提问：如何通过这个检验统计量进行检验呢？

检验统计量

定理1.1 考虑单因子方差分析模型，有以下三个重要的结论：

- 组内偏差平方和的分布为

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

- 组间偏差平方和的期望为

$$E(SS_A) = (a-1)\sigma^2 + m\sum_{i=1}^{a} \alpha_i^2$$

特别地，在原假设成立时，有

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

- 组间偏差平方和 SS_A 与组内偏差平方和 SS_E 独立。

检验统计量

定理1.1 考虑单因子方差分析模型，有以下三个重要的结论：

- 组内偏差平方和的分布为

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

- 组间偏差平方和的期望为

$$E(SS_A) = (a-1)\sigma^2 + m\sum_{i=1}^{a} \alpha_i^2$$

特别地，在原假设成立时，有

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

- 组间偏差平方和 SS_A 与组内偏差平方和 SS_E 独立。

检验统计量

定理1.1的证明

$$\boxed{y_{ij} = \mu + \alpha_i + \varepsilon_{ij}}$$

$$
\begin{aligned}
SS_E &= \sum_{i=1}^{a}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i\cdot})^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{m}\left((\mu + \alpha_i + \varepsilon_{ij}) - m^{-1}\sum_{j=1}^{m}(\mu + \alpha_i + \varepsilon_{ij})\right)^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{m}\left((\cancel{\mu} + \cancel{\alpha_i} + \varepsilon_{ij}) - (\cancel{\mu} + \cancel{\alpha_i} + m^{-1}\sum_{j=1}^{m}\varepsilon_{ij})\right)^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{m}(\varepsilon_{ij} - \bar{\varepsilon}_{i\cdot})^2
\end{aligned}
$$

检验统计量

定理1.1的证明

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{m}(\varepsilon_{ij} - \bar{\varepsilon}_{i.})^2$$

ε_{ij} 是独立同分布的，且

$\varepsilon_{ij} \sim N(0, \sigma^2)$

- $\varepsilon_{i1}, \cdots \varepsilon_{im}$ 可以看作独立同分布的样本，而 $\bar{\varepsilon}_{i.}$ 可以看作其样本均值。
- 于是有，

$$\sigma^{-2}\sum_{j=1}^{m}(\varepsilon_{ij} - \bar{\varepsilon}_{i.})^2 \sim \chi^2(m-1), i = 1,2,\cdots,a$$

- 根据卡方分布的可加性，有

$$\frac{SS_E}{\sigma^2} \sim \chi^2(a(m-1))$$

$n - a = a(m-1)$

检验统计量

定理1.1 考虑单因子方差分析模型，有以下三个重要的结论：

- 组内偏差平方和的分布为

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

- 组间偏差平方和的期望为

$$E(SS_A) = (a-1)\sigma^2 + m\sum_{i=1}^{a}\alpha_i^2$$

特别地，在原假设成立时，有

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

- 组间偏差平方和 SS_A 与组内偏差平方和 SS_E 独立。

检验统计量

定理1.1的证明

$$
\begin{aligned}
SS_A &= m\sum_{i=1}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2 \\
&= m\sum_{i=1}^{a}\left(\frac{1}{m}\sum_{j=1}^{m}(\mu+\alpha_i+\varepsilon_{ij})-\frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{m}(\mu+\alpha_i+\varepsilon_{ij})\right)^2 \\
&= m\sum_{i=1}^{a}(\alpha_i+\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot})^2 \\
&= m\sum_{i=1}^{a}\left(\alpha_i^2+(\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot})^2+2\alpha_i(\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot})\right) \\
&= m\sum_{i=1}^{a}\alpha_i^2+m\sum_{i=1}^{a}(\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot})^2+2m\sum_{i=1}^{a}\alpha_i(\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot})
\end{aligned}
$$

检验统计量

定理1.1的证明

- 因为ε_{ij}是独立同分布的，且

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

所以有

$$\bar{\varepsilon}_{i\cdot} = \frac{1}{m}\sum_{j=1}^{m} \varepsilon_{ij} \sim N\left(0, \frac{\sigma^2}{m}\right)$$

和

$$\bar{\varepsilon}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{m} \varepsilon_{ij} \sim N\left(0, \frac{\sigma^2}{n}\right)$$

检验统计量

定理1.1的证明

$$SS_A = m\sum_{i=1}^{a}\alpha_i^2 + m\sum_{i=1}^{a}(\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})^2 + 2m\sum_{i=1}^{a}\alpha_i\,(\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})$$

- 于是，

$$E(SS_A) = m\sum_{i=1}^{a}\alpha_i^2 + mE\left(\sum_{i=1}^{a}(\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})^2\right)$$

- 这是因为交叉项的期望为零，即

$$E\left(2m\sum_{i=1}^{a}\alpha_i\,(\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})\right) = 2m\sum_{i=1}^{a}\alpha_i\,E(\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot}) = 0$$

检验统计量

定理1.1的证明

- $\bar{\varepsilon}_{i\cdot}$ 可以看作第 i 组的随机误差的均值。
- 各组之间随机误差均是相互独立的。于是， $\bar{\varepsilon}_{1\cdot}, \bar{\varepsilon}_{2\cdot}, \cdots, \bar{\varepsilon}_{a\cdot}$ 是相互独立的。
- 并且，

$$\bar{\varepsilon}_{\cdot\cdot} = \frac{1}{a}\sum_{i=1}^{a} \bar{\varepsilon}_{i\cdot}$$

- 所以，

$$\frac{1}{\sigma^2/m}\sum_{i=1}^{a}(\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})^2 \sim \chi^2(a-1)$$

检验统计量

定理1.1的证明

- 在原假设成立时，$\alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$，有

$$\frac{SS_A}{\sigma^2} = \frac{m\sum_{i=1}^{a}(\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})^2}{\sigma^2} \sim \chi^2(a-1)$$

提问：为什么组间偏差平方和的期望为

$$E(SS_A) = m\sum_{i=1}^{a}\alpha_i^2 + (a-1)\sigma^2\ ?$$

检验统计量

定理1.1 考虑单因子方差分析模型，有以下三个重要的结论：

- 组内偏差平方和的分布为

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

- 组间偏差平方和的期望为

$$E(SS_A) = (a-1)\sigma^2 + m\sum_{i=1}^{a}\alpha_i^2$$

特别地，在原假设成立时，有

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

- 组间偏差平方和 SS_A 与组内偏差平方和 SS_E 独立。

检验统计量

定理1.1的证明

- 因为

$$SS_A = m\sum_{i=1}^{a}(\alpha_i + \bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})^2$$

是 $\bar{\varepsilon}_{1\cdot}, \bar{\varepsilon}_{2\cdot}, \cdots, \bar{\varepsilon}_{a\cdot}$ 的函数。

- 而

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{m}\left(\varepsilon_{ij} - \bar{\varepsilon}_{i\cdot}\right)^2$$

$\sum_{j=1}^{m}\left(\varepsilon_{ij} - \bar{\varepsilon}_{i\cdot}\right)^2$与 $\bar{\varepsilon}_{i\cdot}$ 相互独立

- 所以，SS_A 和 SS_E 相互独立。

检验统计量

定理1.1 考虑单因子方差分析模型，有以下三个重要的结论：

- 组内偏差平方和的分布为

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

- 组间偏差平方和的期望为

$$E(SS_A) = (a-1)\sigma^2 + m\sum_{i=1}^{a} \alpha_i^2$$

特别地，在原假设成立时，有

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

- 组间偏差平方和 SS_A 与组内偏差平方和 SS_E 独立。

检验方法

- 在原假设成立时，检验统计量

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

- 取显著性水平 α；
- **拒绝域法**

$$W = \{F_A \geq F_{1-\alpha}(a-1, n-a)\}$$

其中，$F_\alpha(a-1, n-a)$ 表示自由度分别为 $a-1$ 和 $n-a$ 的 F 分布的 α 分位数。

- ***p*值法**

$$p = P(F \geq F_A)$$

其中，F 表示自由度分别为 $a-1$ 和 $n-a$ 的 F 分布随机变量。如果 $p < \alpha$，那么拒绝原假设。

总结

方差分析表

来源	平方和	自由度	均方	F 值	p 值
因子	SS_A	$a-1$	$MS_A=\dfrac{SS_A}{a-1}$	$F_A=\dfrac{MS_A}{MS_E}$	$p=P(F\geq F_A)$
误差	SS_E	$n-a$	$MS_E=\dfrac{SS_E}{n-a}$		
总和	SS_T	$n-1$			

目录

1. 单因子方差分析（One-way Analysis of Variance）
   I. 两样本独立 t 检验
   II. 单因子方差分析模型与假设
   III. 单因子方差分析检验
   IV. 单因子方差分析参数估计

点估计

- **估计方法**：极大似然估计
- **分布假定**：正态分布

$$y_{ij} \sim N(\mu + \alpha_i, \sigma^2), i = 1,2,\cdots,a, j = 1,2,\cdots,m$$

- **待估参数**：所需要估计的参数为

$$(\mu, \alpha_1, \alpha_2, \cdots, \alpha_a, \sigma^2)^\top$$

- 提问：有多少参数需要估计？

点估计

- 似然函数：

$$L(\mu, \alpha_1, \alpha_2, \cdots, \alpha_a, \sigma^2) = \prod_{i=1}^{a} \prod_{j=1}^{m} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_{ij} - \mu - \alpha_i)^2}{2\sigma^2} \right\} \right\}$$

- 对数似然函数：

$$\begin{aligned} l(\mu, \alpha_1, \alpha_2, \cdots, \alpha_a, \sigma^2) &= \ln L(\mu, \alpha_1, \alpha_2, \cdots, \alpha_a, \sigma^2) \\ &= -\frac{n}{2} \ln(2\pi\sigma^2) - \sum_{i=1}^{a} \sum_{j=1}^{m} \frac{(y_{ij} - \mu - \alpha_i)^2}{2\sigma^2} \end{aligned}$$

点估计

- 对各个参数求偏导得似然方程，即

$$\begin{cases} \dfrac{\partial l}{\partial \mu} = \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{a}\sum_{j=1}^{m}(y_{ij}-\mu-\alpha_i)=0 \\ \dfrac{\partial l}{\partial \alpha_i} = \dfrac{1}{\sigma^2}\displaystyle\sum_{j=1}^{m}(y_{ij}-\mu-\alpha_i)=0, \quad i=1,2,\cdots,a \\ \dfrac{\partial l}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{a}\sum_{j=1}^{m}(y_{ij}-\mu-\alpha_i)^2=0 \end{cases}$$

- 提问：以上似然方程有多少个？

点估计

- 对各个参数求偏导得似然方程，即

$$\begin{cases} \dfrac{\partial l}{\partial \mu} = \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{a}\sum_{j=1}^{m}(y_{ij}-\mu-\alpha_i)=0 \\ \dfrac{\partial l}{\partial \alpha_i} = \dfrac{1}{\sigma^2}\displaystyle\sum_{j=1}^{m}(y_{ij}-\mu-\alpha_i)=0, \quad i=1,2,\cdots,a \\ \dfrac{\partial l}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{a}\sum_{j=1}^{m}(y_{ij}-\mu-\alpha_i)^2=0 \end{cases}$$

- 还有一个方程：

$$\sum_{i=1}^{a}\alpha_i=0$$

点估计

- 各参数的极大似然估计为

$$\begin{cases} \hat{\mu} = \bar{y}_{\cdot\cdot} \\ \hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}, \quad i = 1, 2, \cdots, a \\ \hat{\sigma}^2_{\mathrm{MLE}} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{a} \sum_{j=1}^{m} (y_{ij} - \bar{y}_{i\cdot})^2 = \dfrac{SS_E}{n} \end{cases}$$

- μ_i 的极大似然估计为

$$\hat{\mu}_i = \bar{y}_{i\cdot}$$

- σ^2 的无偏估计为

$$\hat{\sigma}^2 = \frac{SS_E}{n-a} = MS_E$$

区间估计

- **本质问题**：求各个水平 μ_i 的置信区间。
- **枢轴量法**
- μ_i 的点估计为

$$\bar{y}_{i\cdot} = \frac{1}{m}\sum_{j=1}^{m} y_{ij} = \mu + \alpha_i + \bar{\varepsilon}_{i\cdot}$$

- 其分布为

$$\bar{y}_{i\cdot} \sim N\left(\mu + \alpha_i, \frac{\sigma^2}{m}\right)$$

- 因为存在冗余参数 σ^2，需要使用其估计 $\hat{\sigma}^2 = \frac{SS_E}{n-a}$ 代替。

区间估计

- 因为

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

且 $\bar{y}_{1\cdot}, \bar{y}_{2\cdot}, \cdots, \bar{y}_{a\cdot}$ 与 SS_E 相互独立。

- 枢轴量为

$$\frac{\sqrt{m}(\bar{y}_{i\cdot} - \mu_i)}{\sqrt{\hat{\sigma}^2}} \sim t(n-a), i = 1,2,\cdots,a$$

- 因此， μ_i 的 $1-\alpha$ 置信区间为

$$\left[\bar{y}_{i\cdot} - t_{1-\alpha/2}(n-a)\ \frac{\hat{\sigma}}{\sqrt{m}}, \bar{y}_{i\cdot} + t_{1-\alpha/2}(n-a)\ \frac{\hat{\sigma}}{\sqrt{m}}\right]$$

总结

- 各参数的点估计为

$$\begin{cases} \hat{\mu} = \bar{y}_{\cdot\cdot} \\ \hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}, \quad i = 1, 2, \cdots, a \\ \hat{\sigma}^2_{\text{MLE}} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{a}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i\cdot})^2 = \dfrac{SS_E}{n} \end{cases}$$

- μ_i 的点估计为 $\bar{y}_{i\cdot}$
- μ_i 的 $1-\alpha$ 置信区间为

$$\left[\bar{y}_{i\cdot} - t_{1-\alpha/2}(n-a)\frac{\hat{\sigma}}{\sqrt{m}}, \bar{y}_{i\cdot} + t_{1-\alpha/2}(n-a)\frac{\hat{\sigma}}{\sqrt{m}}\right]$$

谢谢

SCHOOL OF DATA
SCIENCE & ENGINEERING
数据科学与工程学院