统计与机器学习

第一章：方差分析

倪 葎

DaSE@ECNU
(lni@dase.ecnu.edu.cn)

目录

1. 单因子方差分析
   - 回顾：二样本独立 t 检验
   - 单因子方差分析的模型及假设
   - 单因子方差分析的检验
   - 单因子方差分析的参数估计

① t分布：

假设 独立 $\begin{cases} X_{1i} \sim N(\mu_1, \sigma^2) \\ X_{2i} \sim N(\mu_2, \sigma^2) \end{cases}$

判断两个分布是否一样

$H_0\ \mu_1 = \mu_2$ vs $H_1\ \mu_1 \neq \mu_2$

回顾：二样本独立 t 检验

概述

在介绍单因子方差分析的问题之前，我们先回顾一类单因子方差分析的特殊情况——二样本独立 t 检验。

- 目的：比较两个方差相等的独立正态分布的均值；
- 数据：

$$\text{样本 } 1: x_{11}, x_{12}, \cdots, x_{1m_1},$$
$$\text{样本 } 2: x_{21}, x_{22}, \cdots, x_{2m_2}.$$

- 假定 x_{ij} 是独立的随机变量，其分布为 $N(\mu_i, \sigma^2)$，
  - μ_i 表示第 i 组的总体均值；
  - σ^2 表示总体方差，是一个未知常数；
  - $i = 1, 2$；
  - $j = 1, 2, \cdots, m_i$；

回顾：二样本独立 t 检验

概述

在介绍单因子方差分析的问题之前，我们先回顾一类单因子方差分析的特殊情况——二样本独立 t 检验。

- 假设检验问题为

$Var(X-Y) = Var(X)+Var(Y) -2Cov(X,Y)$

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

- 检验统计量为

$\bar{x}_1-\bar{x}_2 \sim N(\mu_1-\mu_2, \frac{\sigma^2}{m_1}+\frac{\sigma^2}{m_2})$

独立

$$\boxed{t = \frac{\bar{x}_1 - \bar{x}_2}{s_w\sqrt{\frac{1}{m_1}+\frac{1}{m_2}}}}$$

若两分布相等样本均值差应该不算大

- $\bar{x}_1 = m_1^{-1}\sum_{j=1}^{m_1} x_{1j}$ 表示第一组样本均值；$\bar{x}_1 \sim N(\mu_1, \frac{\sigma^2}{m_1})$
- $s_1^2 = (m_1-1)^{-1}\sum_{j=1}^{m_1}(x_{1j}-\bar{x}_1)^2$ 表示该组样本方差；
- $\bar{x}_2 = m_2^{-1}\sum_{j=1}^{m_2} x_{2j}$ 表示第二组样本样本均值；$\bar{x}_2 \sim N(\mu_2, \frac{\sigma^2}{m_2})$
- $s_2^2 = (m_2-1)^{-1}\sum_{j=1}^{m_2}(x_{2j}-\bar{x}_2)^2$ 表示该组样本方差；

H_0成立时为0

$t(m_1+m_2-2)$

$$\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\hat{\sigma}^2\left(\frac{1}{m_1}+\frac{1}{m_2}\right)}} \sim \cancel{N(0,1)}$$

△自由度？

△论证t、f分布

要证分子分母独立

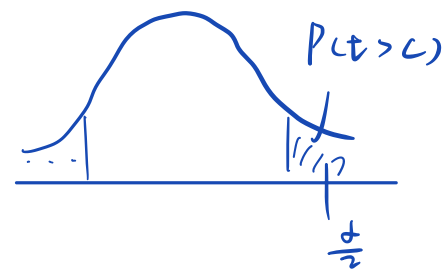
样本均值、方差独立

$$\hat{\sigma}^2=\frac{\sum_{j=1}^{m_1}(X_{1j}-\bar{X}_1)^2+\sum_{j=1}^{m_2}(X_{2j}-\bar{X}_2)^2}{m_1+m_2-2}$$

$$\alpha=P_{H_0}(W)$$

$$=P_{H_0}\left(\underbrace{\boxed{\frac{|\bar{X}_1-\bar{X}_2|}{S_w^2\left(\frac{1}{m_1}+\frac{1}{m_2}\right)}}}_{|t|}>\frac{c}{S_w^2\left(\frac{1}{m_1}+\frac{1}{m_2}\right)}\right) \quad \triangleq c'>0$$

$$=P(\{t<-c'\}\cup\{t>c'\})$$



$$\pm c'=\pm t_{1-\frac{\alpha}{2}}(m_1+m_2-2)$$

第一类错误 vs 第二类错误（越小越好）

回顾：二样本独立 t 检验

概述

- 合方差

$$
\begin{aligned}
s_w^2 &= (m_1+m_2-2)^{-1}((m_1-1)s_1^2+(m_2-1)s_2^2) \\
&= \frac{(m_1-1)}{(m_1+m_2-2)}\cdot s_1^2+\frac{(m_2-1)}{(m_1+m_2-2)}\cdot s_2^2
\end{aligned}
$$

  可看作 s_1^2 和 s_2^2 的加权平均数。

- s_w 是合方差的平方根，即

$$
s_w = \sqrt{s_w^2}
$$

$m_1=m_2$ 时

$$\frac{m-1}{2m-2}=\frac{1}{2}$$

$$\frac{1}{2}s_1^2+\frac{1}{2}s_2^2$$

- 问题：
  - s_w^2 是用来估计什么的?
  - s_w^2 的分布是什么?

回顾：二样本独立 t 检验

概述

- 检验统计量

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_w\sqrt{\frac{1}{m_1} + \frac{1}{m_2}}} \overset{H_0}{\sim} t(m_1 + m_2 - 2)$$

$\sqrt{\frac{1}{x}(s_1^2 + s_2^2)}$

- 二样本独立 t 检验由此得名。
- 特别地，当 $m_1 = m_2 = m$ 时，检验统计量可简化为

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

- 原假设成立时，检验统计量 t 服从自由度为 $2(m-1)$ 的 t 分布；

回顾：二样本独立 t 检验

概述

$W=\{|\bar{x}_1-\bar{x}_2|>c\}$

在显著性水平 α 下，

- 拒绝域法：

$$W=\{|t|\geq t_{1-\alpha/2}(2(m-1))\}$$

其中，$t_\alpha(2m-1))$ 为自由度为 $2(m-1)$ 的 t 分布的 α 分位数。

- p 值法：

$$p=2P(t>|t_0|)$$

其中，t 表示自由度为 $2(m-1)$ 的 t 分布的随机变量，t_0 是通过样本计算的检验统计量；

P值越小 → 做出的判断是拒绝的

回顾：二样本独立 t 检验

例子

- 现有两种为期六周的减肥计划；
- 我们分别用 A 和 B 来表示；
- 选取了 48 名志愿者，随机被分配一种减肥计划，每组有 $m = 24$ 名志愿者；
- 研究者记录了所有志愿者未参加减肥计划时的初始体重，以及参与减肥计划六周后的最终体重；
- 问题：研究者想知道这两种减肥计划的效果是否一致。

回顾：二样本独立 t 检验

例子

表 1.1: 减肥计划的数据

序号	减肥计划	体重		序号	减肥计划	体重	
		初始	最终			初始	最终
1	A	58	54.2	25	B	58	60.1
2	A	60	54.0	26	B	58	56.0
3	A	64	63.3	27	B	59	57.3
4	A	64	61.1	28	B	61	56.7
5	A	65	62.2	29	B	63	62.4
6	A	66	64.0	30	B	63	60.3
7	A	67	65.0	31	B	63	59.4
8	A	69	60.5	32	B	65	62.0
9	A	70	68.1	33	B	66	64.0
10	A	70	66.9	34	B	68	63.8
11	A	71	71.6	35	B	68	63.3
12	A	72	70.5	36	B	71	66.8
13	A	72	69.0	37	B	75	72.6
14	A	72	68.4	38	B	75	69.2
15	A	72	70.9	39	B	76	72.7
16	A	74	69.5	40	B	76	72.5
17	A	78	73.9	41	B	77	77.5
18	A	80	71.0	42	B	78	72.7
19	A	80	77.6	43	B	78	76.3
20	A	82	81.1	44	B	79	73.6
21	A	83	79.1	45	B	79	72.9
22	A	85	81.5	46	B	79	71.1
23	A	87	81.9	47	B	80	81.4
24	A	88	84.5	48	B	80	75.7

回顾：二样本独立 t 检验

例子

表 1.2: 计算后减肥计划的数据

序号	减肥计划	体重 初始	体重 最终	体重 差异	序号	减肥计划	体重 初始	体重 最终	体重 差异
1	A	58	54.2	-3.8	25	B	58	60.1	2.1
2	A	60	54.0	-6.0	26	B	58	56.0	-2.0
3	A	64	63.3	-0.7	27	B	59	57.3	-1.7
4	A	64	61.1	-2.9	28	B	61	56.7	-4.3
5	A	65	62.2	-2.8	29	B	63	62.4	-0.6
6	A	66	64.0	-2.0	30	B	63	60.3	-2.7
7	A	67	65.0	-2.0	31	B	63	59.4	-3.6
8	A	69	60.5	-8.5	32	B	65	62.0	-3.0
9	A	70	68.1	-1.9	33	B	66	64.0	-2.0
10	A	70	66.9	-3.1	34	B	68	63.8	-4.2
11	A	71	71.6	0.6	35	B	68	63.3	-4.7
12	A	72	70.5	-1.5	36	B	71	66.8	-4.2
13	A	72	69.0	-3.0	37	B	75	72.6	-2.4
14	A	72	68.4	-3.6	38	B	75	69.2	-5.8
15	A	72	70.9	-1.1	39	B	76	72.7	-3.3
16	A	74	69.5	-4.5	40	B	76	72.5	-3.5
17	A	78	73.9	-4.1	41	B	77	77.5	0.5
18	A	80	71.0	-9.0	42	B	78	72.7	-5.3
19	A	80	77.6	-2.4	43	B	78	76.3	-1.7
20	A	82	81.1	-0.9	44	B	79	73.6	-5.4
21	A	83	79.1	-3.9	45	B	79	72.9	-6.1
22	A	85	81.5	-3.5	46	B	79	71.1	-7.9
23	A	87	81.9	-5.1	47	B	80	81.4	1.4
24	A	88	84.5	-3.5	48	B	80	75.7	-4.3

回顾：二样本独立 t 检验

例子

- 令 x_{1j} 表示减肥计划 A 六周前后的体重差异，x_{2j} 表示减肥计划 B 六周前后的体重差异， $j = 1, 2, \cdots, 24$;
- 假设

$$x_{ij} \overset{\text{独立}}{\sim} N(\mu_i, \sigma^2), \quad i = 1, 2, j = 1, 2, \cdots, 24.$$

- 检验问题为

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

- 我们可以计算

$$\bar{x}_1 = -3.3000, \quad s_1^2 = 5.0183;$$
$$\bar{x}_2 = -3.1125, \quad s_2^2 = 5.7072;$$

回顾：二样本独立 t 检验

例子

- 合方差为

$$s_w^2 = 5.3627$$

- 检验统计量为

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_w\sqrt{\frac{2}{m}}} = -0.2805.$$

- 取显著性水平 $\alpha = 0.05$。
- 拒绝域为

$$\{|t| \geq t_{1-\alpha/2}(2m-2)\} = \{|t| \geq 2.0129\}$$

- 我们认为这两种减肥计划的效果是一致的。

单因子方差分析的模型及假设

动机

- 如果需要比较三种减肥方式是否一致?

单因子方差分析的模型及假设

定义

- **响应变量**：我们关心的随机变量，一般用 y 表示；
- **因子**：引发响应变量 y 大小变化的因素，一般用大写字母表示，例如：A，有 a 种不同的取值，通常 $a \geq 2$；称因子 A 的一种取值为一个水平或一个处理；
- **重复次数**：在因子 A 每个水平下，随机变量的个数，记为 m；
- 在例子（两种减肥方案的比较）中，
  - 减肥计划前后的体重差作为响应变量；
  - 减肥计划为所关心的因子，$a = 2$；
  - 每组有 24 名志愿者，即 $m = 24$；
  - 样本量 $n = am = 48$；

单因子方差分析的模型及假设

定义

- 数据的结构为

水平	观测到的响应				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}	$y_{a\cdot}$	$\bar{y}_{a\cdot}$
汇总					$y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot}$

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

单因子方差分析的模型及假设

定义

- 这些符号之间的关系为

$$y_{i\cdot} = \sum_{j=1}^{m} y_{ij} \quad \bar{y}_{i\cdot} = \frac{y_{i\cdot}}{m} \quad i = 1, 2, \cdots, a$$

$$y_{\cdot\cdot} = \sum_{i=1}^{a} \sum_{j=1}^{m} y_{ij} \quad \bar{y}_{\cdot\cdot} = \frac{y_{\cdot\cdot}}{n}$$

模型本身就是一种假设

单因子方差分析的模型及假设

模型：均值模型

$X \sim N(\mu, \sigma^2)$　$x_1, \cdots, x_n$ 是样本（独立同分布）

$x_i = \mu + \varepsilon_i$，$\varepsilon_i \sim N(0, \sigma^2)$

- 方差分析模型的一般形式为

随机项

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

  - y_{ij} 表示在因子的第 i 种水平下所观测到的第 j 个响应变量；
  - μ_i 表示因子的第 i 个水平下的均值；
  - ε_{ij} 是随机误差；通常认为随机误差的期望为零，即 $\boxed{E(\varepsilon_{ij}) = 0}$。 if $E(\varepsilon_{ij}) = \eta \neq 0$ $\Rightarrow \tilde{\varepsilon}_{ij} = \varepsilon_{ij} - \eta$, $E(\tilde{\varepsilon}_{ij}) = 0$
- 很明显的结果为 $E(y_{ij}) = \mu_i, j = 1, 2, \cdots, m$
- 称这个模型为**均值模型**。

单因子方差分析的模型及假设

模型：效应模型

- 令

$$\boxed{\mu_i = \mu + \alpha_i}, i = 1, 2, \cdots, a.$$

- 方差分析模型的另一种形式为

替换得

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

- 称这个模型为**效应模型**。

$\begin{cases} a+1 \text{ 方程} \\ a+2 \text{ 变量} \end{cases} \Rightarrow$ 无穷多解

单因子方差分析的模型及假设

说明

? 无穷多解

- 相比于均值模型，效应模型参数个数有所增加。
- 为了避免参数无法识别的问题，我们通常需要对参数 $(\mu, \alpha_1, \alpha_2, \cdots, \alpha_a)$ 给出一个合理的约束。
- 最常用的约束之一为

$$\sum_{i=1}^{n} \alpha_i = 0.$$

- 这表明了因子 A 的各个水平的效应在零附近波动，且所有效应的总和为零。

单因子方差分析的模型及假设

说明

- 在效应模型中，

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

  在因子的第 i 个水平下的均值可以划分为两部分，
  - 其一为总体均值 $\mu = a^{-1}\sum_{i=1}^{a}\mu_i$，
  - 其二为第 i 个水平的$\boxed{\text{效应}}$ α_i，也就是说，各个水平的效应是各个水平的均值与总体均值的偏差。
- 因为在均值模型（或效应模型）中仅考虑了一个因子，所以，称这两个模型为**单因子方差分析模型**。

一个影响output的因素

单因子方差分析的模型及假设

假设

- 随机误差的假定：

$$\varepsilon_{ij} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$$

  - 独立同分布；
  - 以均值为零，方差为 σ^2 正态分布的随机变量；
- 这表明：不同水平下，响应变量的波动大小是一致的；
- 观测到的数据是相互独立且均服从正态分布，即

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

单因子方差分析的模型及假设

总结

- 单因子方差分析的模型为

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},\ \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2),\ \begin{cases} i = 1, 2, \cdots, a \\ j = 1, 2, \cdots, m \end{cases}$$

s.t.

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

单因子方差分析的检验

假设

- 均值模型的假设

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_a$$
H_1：存在在两种水平 i, j 下的均值不相等，即$\mu_i \neq \mu_j$.

- 效应模型的假设

因为$\sum\alpha_i=0$

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$$
H_1：存在第 i 个水平不为零，即$\alpha_i \neq 0$.

- 这两种假设都是正确且等价的，只是针对不同的模型而提出的。

单因子方差分析的检验

回顾：二样本 t 检验

- 检验统计量为

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

→ 均值差异

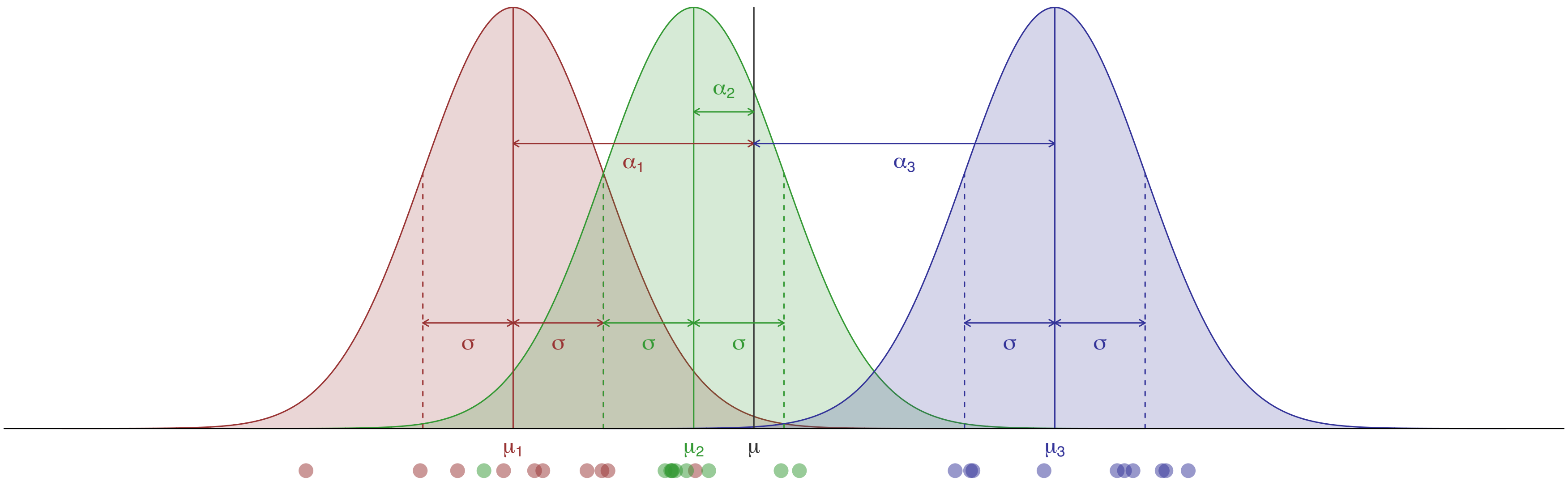
→ 数据波动

- 本质上比较的是两组样本均值差异与数据波动的大小。
- 相比于数据的波动，两组样本均值的差异大得多，那么我们才能有足够的证据支撑说明这两组数据的均值是不一致的。

单因子方差分析的检验

图示

- 以 $a = 3$ 个水平的因子为例，



~~两两之间作差~~
和 σ^2 同阶

→ 三组取平均，每组与平均比较
和 σ 同阶

单因子方差分析的检验

平方和分解公式

- 总偏差平方和 SS_T 可拆分为两部分，即

引入 $\bar{y}_{i\cdot}$

直接与总平均作差

组间差 组内差

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

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

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

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

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

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

组间平方和 组内平方和

单因子方差分析的检验

平方和分解公式

- 交叉项为零，这是因为

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

单因子方差分析的检验

平方和分解公式

- 平方和分解公式 $SS_T = SS_{tr} + SS_E$

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

- 第一项为组间偏差平方和 SS_A，即

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

SS_A 表示了不同水平下数据的平均值与所有数据的总平均值之间的偏差平方和，既包含了因子 A 取不同水平引起的数据差异，又包含了随机误差对它的影响；

单因子方差分析的检验

平方和分解公式

- 平方和分解公式

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

- 第二项为组内偏差平方和 SS_E，即

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

SS_E 表示同一水平下数据 y_{ij} 与其平均值 $\bar{y}_{i\cdot}$ 的差异，是由于试验误差引起的。

单因子方差分析的检验

检验统计量

○想清楚为什么要
除以总偏差平方和

- 平方和分解公式简记为

$$SS_T = SS_A + SS_E$$

各因子影响相同.

- 对于给定的一组数据，总偏差平方和 SS_T 是不变的。
- 如果原假设成立，SS_A 仅仅受到随机误差方差的影响，取值应该不大。是因为每组的样本均值 $\bar{y}_{i\cdot}$ 是 μ_i 的一个合理的估计，也应该取值接近。
- 一个直观的想法是比较比值

可以关注因素a重要与否

$$SS_A/SS_T,$$

如果这个比值越大，我们越有证据支持备择假设；反之我们认为原假设更为合理。

单因子方差分析的检验

检验统计量

$$\frac{1}{1+\frac{SS_E}{SS_A}} = \frac{SS_A}{SS_A+SS_E} = \frac{SS_A}{SS_T}$$

- 根据平方和分解公式

$$SS_T = SS_A + SS_E$$

$$\frac{1}{1+\frac{1}{x}}$$ 增函数

- SS_A/SS_E 随 SS_A/SS_T 增大而增大的。
- 在单因子方差分析模型中，我们所构造的检验统计量是基于

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

- 问题：SS_A 和 SS_E 的分布是什么?

。搞清楚卡方分布!!!

单因子方差分析的检验

定理

在单因子方差分析模型中，我们有：

- 组内偏差平方和的分布为

正态分布的组内偏差平方和？（自己推一遍）

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

- 在原假设 H_0 成立时，组间偏差平方和的分布为

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

- 组间偏差平方和与组内偏差平方和独立。

我们先看看这个定理有什么用？**点击这里**。

单因子方差分析的检验

定理（第一部分）

在单因子方差分析模型中，我们有：

- 组内偏差平方和的分布为

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

单因子方差分析的检验

证明：定理（第一部分）

- 根据单因子方差分析模型

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

- SS_E 可以写为

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

单因子方差分析的检验

证明：定理（第一部分）

- 由于 ε_{ij} 是独立同分布的正态随机变量，即

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

- 在因子 A 的第 i 个水平下，$\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{im}$ 可以看作来自正态总体 $N(0, \sigma^2)$ 的一组样本量为 m 的样本，而 $\bar{\varepsilon}_{i\cdot} = m^{-1}\sum_{j=1}^{m} \varepsilon_{ij}$ 可以看作这组样本的样本均值；

无法观测

- 那么

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

卡方分布. 自由度 = 样本量 -1

而且不同水平下的偏差平方和是相互独立的。

单因子方差分析的检验

证明：定理（第一部分）

- 根据卡方分布的可加性，我们有

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

- 注意到，$a(m-1) = am - a = n - a$；

单因子方差分析的检验

定理（第二部分）

在单因子方差分析模型中，我们有：

- 在原假设 H_0 成立时，组间偏差平方和的分布为

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

单因子方差分析的检验

证明：定理（第二部分）

- 组间偏差平方和 SS_A 可写为

推分布

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

$n = am$

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

$\alpha_i - \frac{m\sum_{i=1}^{a}\alpha_i}{n} = \alpha_i$

$\frac{1}{m}\sum_{j=1}^{m}\varepsilon_{ij} - \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{m}$

$= \bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot}$

$=0$?不确定

$=0$

原假设成立时 → $\alpha_1 = \alpha_2 = \cdots \alpha_a$

$\mu_i = \mu + \alpha_i$

○还有吗?

$\alpha_1+\alpha_2+\cdots\alpha_n=0$

$\Rightarrow \alpha_1=\cdots=\alpha_n=0$

$\mu_1=\mu_2=\cdots$

$\Rightarrow \alpha_1=\alpha_2=\cdots$

单因子方差分析的检验

证明：定理（第二部分）

- 因为 $\varepsilon_{ij} \sim N(0, \sigma^2)$ 且相互独立，所以

$$\bar{\varepsilon}_{i\cdot} = m^{-1}\sum_{j=1}^{m}\varepsilon_{ij} \sim N(0, \sigma^2 m^{-1}) \quad \text{和} \quad \bar{\varepsilon}_{\cdot\cdot} = n^{-1}\sum_{i=1}^{a}\sum_{j=1}^{m}\varepsilon_{ij} \sim N(0, \sigma^2 n^{-1}).$$

- 于是，交叉项的期望为

常数

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

- 那么，我们有

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

单因子方差分析的检验

证明：定理（第二部分）

- $\bar{\varepsilon}_{i\cdot}$ 是第 i 个水平下随机误差的样本均值，因为不同水平下的随机误差是相互独立的，所以，这些随机误差的样本均值 $\bar{\varepsilon}_{1\cdot}, \bar{\varepsilon}_{2\cdot}, \cdots, \bar{\varepsilon}_{a\cdot}$ 是相互独立的。

$\bar{\varepsilon}_{i\cdot}$ 独立同分布

- 而

$$\bar{\varepsilon}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{m}\varepsilon_{ij} = \frac{1}{a}\sum_{i=1}^{a}\bar{\varepsilon}_{i\cdot}$$

$=am$

$\bar{\varepsilon}_{\cdot\cdot}$ 的算术平均数

可以看作 ~~a 个~~ $\bar{\varepsilon}_{1\cdot}, \bar{\varepsilon}_{2\cdot}, \cdots, \bar{\varepsilon}_{a\cdot}$ 的样本均值。

a 个不同正态分布的样本均值.

- 于是，

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

单因子方差分析的检验

证明：定理（第二部分）

- 在原假设 H_0 成立时，即 $\alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$，我们有

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

推论

- 组间偏差平方和的期望为

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

单因子方差分析的检验

定理（第三部分）

在单因子方差分析模型中，我们有：

- 组间偏差平方和与组内偏差平方和独立，即

$$SS_A \perp SS_E.$$

单因子方差分析的检验

证明：定理（第三部分）

- 因为
$$SS_A = m\sum_{i=1}^{a}(\alpha_i + \bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})^2$$
可以是 $\bar{\varepsilon}_{1\cdot}, \bar{\varepsilon}_{2\cdot}, \cdots, \bar{\varepsilon}_{a\cdot}$ 的函数。
- 同时，我们知道 $\sum_{j=1}^{m}(\varepsilon_{ij} - \bar{\varepsilon}_{i\cdot})^2$ 与 $\bar{\varepsilon}_{i\cdot}$ 是相互独立的，而且因子不同水平下的随机误差是相互独立的。
- 因此，SS_A 与 SS_E 独立。

单因子方差分析的检验

检验统计量

- 检验统计量为

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

$\mu_1 = \mu_2 = \cdots = \mu_n$

- 在原假设 H_0 成立下服从自由度分别为 $a-1$ 和 $n-a$ 的 F 分布，即 $F_A \sim F(a-1, n-a)$。
- 在显著性水平 α 下，如果

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

那么，我们会拒绝原假设，其中 $F_\alpha(a-1, n-a)$ 是自由度分别为 $a-1$ 和 $n-a$ 的 F 分布的 α 分位数。

单因子方差分析的检验

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	组数 $a-1$	$MS_A = \frac{SS_A}{a-1}$	$F_A = \frac{MS_A}{MS_E}$
误差 E	SS_E	样本量← $n-a$	$MS_E = \frac{SS_E}{n-a}$	
总和	SS_T	$n-1$	方差的估计	

越小越拒绝原假设

单因子方差分析的检验

p 值的计算

- 计算 p 值来进行判断，即

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

其中，F_A 是通过样本计算而得的检验统计量，F 为一个自由度为 $a-1$ 和 $n-a$ 的 F 分布的随机变量。

- 如果 $p_A < \alpha$，那么我们会拒绝原假设；否则，我们无法拒绝原假设。

单因子方差分析的参数估计

点估计

- 由于

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

- 可以采用极大似然估计来估计参数

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

单因子方差分析的参数估计

点估计

- 似然函数为 $L(\mu, \alpha_1, \alpha_2, \cdots, \alpha_a, \sigma^2) =$

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

- 其对数似然函数为

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

单因子方差分析的参数估计

点估计

- 对各个参数求偏导，得似然方程为

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

- 我们可以发现，上述的 $a+2$ 个方程中有 1 个方程是多余的。（问题：为什么?）

单因子方差分析的参数估计

点估计

- 效应模型的约束

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

单因子方差分析的参数估计

点估计

估计时不是无偏估计？

- 于是，我们可以求出各参数的极大似然估计为

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

- 由极大似然估计的不变性，各个水平的均值 μ_i 的极大似然估计为

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

- 因为 $E(SS_E) = \sigma^2(n-a)$，所以，$\hat{\sigma}^2_{\text{MLE}}$ 并不是 σ^2 的一个无偏估计，而常用 $\hat{\sigma}^2 = \frac{SS_E}{n-a} = MS_E$。

单因子方差分析的参数估计

区间估计

- 讨论各水平均值 μ_i 的置信区间。
- 由于

$$\begin{aligned}\bar{y}_{i\cdot} &= \frac{1}{m}\sum_{j=1}^{m} y_{ij} = \mu + \alpha_i + \bar{\varepsilon}_{i\cdot} \sim N(\mu+\alpha_i, \sigma^2 m^{-1}) \\ \frac{SS_E}{\sigma^2} &\sim \chi^2(n-a),\end{aligned}$$

- 而且 $\bar{y}_{1\cdot}, \bar{y}_{2\cdot}, \cdots, \bar{y}_{a\cdot}$ 均与 SS_E 相互独立,
- 所以,

$$\frac{\sqrt{m}(\bar{y}_{i\cdot} - \mu_i)}{\sqrt{SS_E/(n-a)}} \sim t(n-a), \quad i = 1, 2, \cdots, a.$$

单因子方差分析的参数估计

区间估计

- 于是，因子 A 的第 i 个水平的均值 μ_i 的 $1-\alpha$ 置信区间为

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

其中，$t_\alpha(n-a)$ 为自由度为 $n-a$ 的 t 分布的分位数，而 $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$。