第7章　方差分析模型

吴密霞

北京工业大学统计与数据科学系

E-mail: *wumixia@bjut.edu.cn*

教材

- 吴密霞, 王松桂. 2024. 线性模型引论（第2版）, 科学出版社.

本章目录

- 单向分类模型
  - 模型
  - 参数估计
  - 检验问题
- 两向分类模型
  - 不带交互效应
  - 带交互效应
  - 不平衡数据
- 套分类模型

方差分析模型是应用非常广泛的一类线性模型. 这种模型多有一定的试验设计背景, 因而也称试验设计模型.

两种不同的统计分析方法:

- 平方和分解法. 适用于平衡数据的情形. 将数据总变差平方和按其来源(各种因子和随机误差)进行分解, 得到各因子平方和及误差平方和. 接下来的统计分析是基于各因子平方和与误差平方和大小的比较.
- 回归分析法. 将对各因子效应的约束条件借助于虚拟变量融入模型, 将一般线性模型的估计与检验的结果应用于这种模型.

单向分类模型

设因子A有a个水平，在水平A_i 下作n_i 次重复观测. 记y_{ij}为在第i个水平A_i下第j次的观测值，即模型

$$y_{ij} = \mu + \alpha_i + e_{ij}, \qquad i = 1, 2, \cdots, a, \quad j = 1, 2, \cdots, n_i, \tag{2.1}$$

这里，μ为总平均，α_i为第i个水平的效应，$e_{ij} \sim \mathrm{N}(0, \sigma^2)$,相互独立.

- 边界条件

$$\sum_{i=1}^{a} n_i \alpha_i = 0. \tag{2.2}$$

- 若$n_1 = \cdots = n_a$，则称模型为**平衡的**，(2.1)变为$\sum_{i=1}^{a} \alpha_i = 0$.

单向分类模型的矩阵形式

- 矩阵形式

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e},$$

其中 $\boldsymbol{y} = (y_{11}, \cdots, y_{1n_1}, , y_{21}, \cdots, y_{2n_2}, \cdots, y_{a1}, \cdots, y_{an_a},)'$,

$\boldsymbol{e} = (e_{11}, \cdots, e_{1n_1}, , e_{21}, \cdots, e_{2n_2}, \cdots, e_{a1}, \cdots, e_{an_a})'$,

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_a} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_a} \end{pmatrix}, \quad \boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2, \cdots, \alpha_a)'.$$

设计阵$\mathbf{X}$ 是列降秩的, 即秩小于它的列数; 约束条件$\sum_{i=1}^{a} n_i\alpha_i = 0$

- 因为$\mathrm{rk}(\mathbf{X}) = a$, 所以至多只有$a$个线性无关的可估函数.
- 由于$\mu + \alpha_i, \quad i = 1, \cdots, a$ 都是可估的, 且线性无关, 故任一可估函数都可表示为它们的线性组合:

$$\sum_{i=1}^{a} c_i(\mu + \alpha_i) = \mu \sum_{i=1}^{a} c_i + \sum_{i=1}^{a} c_i\alpha_i. \tag{2.3}$$

- 只包含效应$\alpha_i(i = 1, \cdots, a)$ 而不包含总均值μ的可估函数, 则

$$\sum_{i=1}^{a} c_i\alpha_i \text{可估} \Longleftrightarrow \sum_{i=1}^{a} c_i = 0.$$

对照

称满足条件$\sum_{i=1}^{a} c_i = 0$的函数$\sum_{i=1}^{a} c_i\alpha_i$为一个对照.

定理7.1.1

对于单向分类模型(2.1)

(1) $\sum_{i=1}^{a} c_i\alpha_i$可估$\Longleftrightarrow \sum_{i=1}^{a} c_i\alpha_i$是一个对照, 即$\sum_{i=1}^{a} c_i = 0.$

(2) 对照$\sum_{i=1}^{a} c_i\alpha_i$的BLU估计为$\sum_{i=1}^{a} c_i\bar{y}_{i\cdot}$

- $\mu_i = \mu + \alpha_i$的BLU估计为$\bar{y}_{i\cdot}$, 因为此时模型为单元均值模型：

$$y_{ij} = \mu_i + e_{ij}, \quad i = 1, \cdots, a, \quad j = 1, \cdots, n_i.$$

- 任意$\alpha_i - \alpha_j$, $i \neq j$ 都可估, 其BLU估计为$\widehat{\alpha}_i - \widehat{\alpha}_j = \bar{y}_{i\cdot} - \bar{y}_{j\cdot}$.

单向分类模型的两种约束条件

- 加权均值的单向分类模型 即总体均值为μ_i 的加权和:

$$\mu = \sum_{i=1}^{a} \frac{n_i}{N}\mu_i = E(\bar{y}_{\cdot\cdot}), \quad \text{约束条件} \sum_{i=1}^{a} \frac{n_i}{N}\alpha_i = 0$$

每个个体的均值被视为对总体均值同等重要.

- 无加权均值的单向分类模型 即总体均值为μ_i的简单平均:

$$\mu = \frac{1}{a}\sum_{i=1}^{a} \mu_i, \quad \text{约束条件} \sum_{i=1}^{a} \alpha_i = 0$$

因子A各水平单元下因变量的均值被视为对总体均值同等重要.

注 R软件中函数aov()就是基于约束条件$\sum_{i=1}^{a} \alpha_i = 0$. 令

$$\alpha_1 = -\alpha_2 - \cdots - \alpha_a, \quad \boldsymbol{\beta}_{-1} = (\mu, \alpha_2, \alpha_3, \cdots, \alpha_a)'$$

约简模型(将约束条件代入模型所得)

$$\boldsymbol{y} = \mathbf{X}^* \boldsymbol{\beta}_{-1} + \boldsymbol{e},$$

其中

$$\mathbf{X}^* = \begin{pmatrix} \mathbf{1}_{n_1} & -\mathbf{1}_{n_1} & -\mathbf{1}_{n_1} & \cdots & -\mathbf{1}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{1}_{n_2} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{1}_{n_{a-1}} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_{a-1}} \end{pmatrix} \quad \text{列满秩}$$

注 单向分类模型无论在加权边界约束条件$\sum_{i=1}^{a}\frac{n_i}{N}\alpha_i=0$还是无加权两种边界约束条件$\sum_{i=1}^{a}\alpha_i=0$下, 定理7.1.1都成立.

因为效应对照等于各单元均值对照, 由$\sum\limits_{i=1}^{a}c_i=0$立得

$$\sum_{i=1}^{a}c_i\alpha_i=\sum_{i=1}^{a}c_i(\alpha_i+\mu)=\sum_{i=1}^{a}c_i\mu_i.$$

注 两种定义下的μ: $\sum\limits_{i=1}^{a}\frac{n_i}{N}\mu_i$ 和$\sum\limits_{i=1}^{a}\frac{1}{a}\mu_i$, 它们的BLU估计分别为

$$\sum_{i=1}^{a}\frac{n_i}{N}\bar{y}_{i\cdot}=\bar{y}_{\cdot\cdot},\quad \sum_{i=1}^{a}\frac{1}{a}\bar{y}_{i\cdot}.$$

考察因子A的a个水平效应是否有显著差异, 即检验假设

$$H_0: \quad \alpha_1 = \alpha_2 = \cdots = \alpha_a.$$

等价于

$$H_0: \quad \alpha_1 - \alpha_a = \alpha_2 - \alpha_a = \cdots = \alpha_{a-1} - \alpha_a = 0.$$

其中, $\alpha_i - \alpha_a \ (i = 1, \cdots, a-1)$ 皆可估.

- 若H_0为真, 则约简模型为

$$y_{ij} = \mu + e_{ij}, \quad i = 1, 2, \cdots, a, \quad j = 1, 2, \cdots, n_i. \tag{2.4}$$

- 在H_0下, μ的约束LS估计为

$$\widehat{\mu}_{H_0} = \bar{y}_{..},$$

模型的回归平方和与残差平方和

$$\mathrm{RSS}(\mu) = N\bar{y}_{..}^2, \quad \mathrm{SS}_{He} = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 = \mathrm{SS}_T(\text{ 总平方和}).$$

- H_0不成立时, μ_i的LS估计为

$$\widehat{\mu}_i = \bar{y}_{i\cdot}, \quad i = 1, \cdots, a,$$

模型的回归平方和与残差平方和

$$\mathrm{RSS}(\mu, \alpha) = \sum_{i=1}^{a} y_{i\cdot}^2/n_i, \quad \mathrm{SS}_e = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2.$$

- 因子A的平方和

$$\mathrm{SS}_A = \mathrm{RSS}(\mu,\alpha) - \mathrm{RSS}(\mu) = \sum_{i=1}^{a} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2.$$

- 平方和分解：$\mathrm{SS}_T = \mathrm{SS}_A + \mathrm{SS}_e$.
- 检验假设H_0的F统计量

$$F = \frac{\mathrm{SS}_A/(a-1)}{\mathrm{SS}_e/(N-1)} = \frac{MS_A}{MS_e}, \tag{2.5}$$

其中$MS_A = \mathrm{SS}_A/(a-1)$, $MS_e = \mathrm{SS}_e/(N-a)$分别为因子$A$和误差的均方. 当$H_0$为真时, $F \sim F_{a-1,N-a}$.

Table: 单因素方差分析表

方差源	自由度	平方和	均方	F值
组间差(因子A)	$a-1$	SS_A	$MS_A = SS_A/(a-1)$	$F = \dfrac{MS_A}{MS_e}$
组内差(误差)	$N-a$	SS_e	$MS_e = SS_e/(N-a)$	
总和	$N-1$	$SS_T = SS_A + SS_e$		

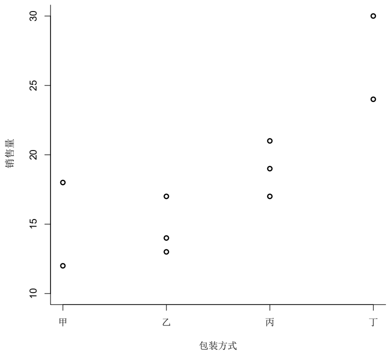
例7.1.1

为了考察儿童对这4种包装方案的喜爱程度, 将甲、丁式包装各2批, 乙、丙式包装各3批, 共10批随机地分给10家食品商店各一批试销, 观察它们的销售量. 选择的这10家食品店所处地段的繁华程度、商店的规模、糖果广告橱窗的布置都相仿. 问当显著性水平为$\alpha = 0.05$时, 儿童对4种包装方式的喜爱程度是否有显著差异.

包装方式	销售量y_{ij}			$y_{i\cdot}$	$y_{i\cdot}^2$
甲	12	18		30	900
乙	14	17	13	44	1936
丙	19	17	21	57	3249
丁	24	30		54	2916

解 设A_1, A_2, A_3,A_4, 分别是甲、乙、丙、丁4种包装.

- 四种包装下的糖果销售量散点图



案例分析

- 采用单向分类模型来分析.

$$\sum_i \sum_j y_{ij}^2 = (12)^2 + (18)^2 + \cdots + (30)^2 = 3689,$$

$$y_{\cdot\cdot} = \sum_i \sum_j y_{ij} = 12 + 18 + \cdots + 30 = 185,$$

$\mathrm{SS}_T = 266.5$, $\mathrm{SS}_e = 52.67$, $\mathrm{SS}_A = 213.83$. 于是

$$F = \frac{\mathrm{SS}_A/3}{\mathrm{SS}_e/6} \approx \frac{71.28}{8.78} \approx 8.12.$$

在显著性水平$\alpha = 0.05$下, 由于P值 $= P(F_{3,6} > 8.12) = 0.0156 < \alpha = 0.05$ 或$F_{3,6}(0.05) = 4.76 < 8.12$, 所以可认为儿童们对糖果的4种包装方式的喜爱程度有显著差异.

- 将以上分析结果汇总到方差分析表：

Table: 糖果销售量数据的方差分析表

方差源	自由度	平方和	均方	F值	P值
组间差(因子A)	3	213.83	71.28	8.12	0.0156 *
组内差(误差)	6	52.67	8.78		
总和	9	304			

- 程序

x <- c(12,18,14,17,13,19,17,21,24,30)

A <- factor(c(rep(1,2),rep(2,3),rep(3,3),rep(4,2))) 因子设计阵

X.aov <- aov(x ∼ A)

同时置信区间

- 若$H_0 : \alpha_1 = \cdots = \alpha_a$ 被拒绝, 则可认为因子A的a个水平的效应不全相等. 接下来需考察哪些效应不相等？
- 可通过对$\alpha_i - \alpha_j \ (i \neq j)$ 的同时置信区间来判断.
- 考虑任意m个对照$\sum_i c_i^{(k)} \alpha_i \ (k = 1, \cdots, m)$ 的同时置信区间.

定义：学生化极差分布

设$Z_1, Z_2, \cdots, Z_n \sim N(0, 1), mW^2 \sim \chi^2_m$, 且所有这些随机变量都相互独立, 则称随机变量$q_{n,m} = \frac{\max Z_i - \min Z_i}{W}$ 的分布为参数为n, m的学生化极差分布(studentized range distribution).

记它的上侧α 分位点为$q_{n,m}(\alpha)$.

同时置信区间

- 常见三种同时置信区间：
  - Bonferroni区间 (m个对照$\sum_i c_i^{(k)}\alpha_i$ $(k=1,\cdots,m)$)

$$\sum_i c_i^{(k)}\bar{y}_{i\cdot} \pm t_{N-a}\left(\frac{\alpha}{2m}\right)\hat{\sigma}\sqrt{\sum_i (c_i^{(k)})^2/n_i}, \quad k=1,2,\cdots,m.$$

  - Scheffé区间 (所有对照$\sum_i c_i\alpha_i$, $\sum_i c_i=0$)

$$\sum_i c_i\bar{y}_{i\cdot} \pm \hat{\sigma}\sqrt{(a-1)F_{a-1,N-a}(\alpha)\sum_i \frac{c_i^2}{n_i}},$$

  - Tukey法区间 (所有对照, 平衡数据$n_i=n$)

$$\sum_i c_i\bar{y}_{i\cdot} \pm q_{a,a(n-1)}(\alpha)\frac{\hat{\sigma}}{2\sqrt{n}}\sum_i |c_i|.$$

特别地，

- m个$\alpha_i-\alpha_j\ (i\neq j)$的置信系数为$1-\alpha$的Bonferroni同时置信区间

$$\bar{y}_{i\cdot}-\bar{y}_{j\cdot}\pm t_{N-a}\left(\frac{\alpha}{2m}\right)\widehat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_i}}$$

- 所有$\alpha_i-\alpha_j\ (i\neq j)$的置信系数为$1-\alpha$的Scheffé同时置信区间

$$\bar{y}_{i\cdot}-\bar{y}_{j\cdot}\pm\widehat{\sigma}\sqrt{(a-1)F_{a-1,N-a}(\alpha)\sum_i\frac{c_i^2}{n_i}}$$

- 所有$\alpha_i-\alpha_j\ (i\neq j)$的置信系数为$1-\alpha$的Tukey同时置信区间

$$\bar{y}_{i\cdot}-\bar{y}_{j\cdot}\pm q_{a,N-a}(\alpha)\frac{\widehat{\sigma}}{\sqrt{n}}$$

所有两两效应差共有C_a^2 种（忽略符号）

对平衡单向分类模型, 即$n_i = n\ (i = 1, \cdots, a)$, 我们把Bonferroni同时置信区间、Tukey同时置信区间和Scheffé同时置信区间的的中心都相同, 区间的长度分别为

$$\frac{2\widehat{\sigma}}{\sqrt{n}} t_{N-a}\left(\frac{\alpha}{2m}\right)\sqrt{\sum_i {c_i}^2},$$

$$\frac{\widehat{\sigma}}{\sqrt{n}} q_{a,a(n-1)}(\alpha)\sum_i |c_i|,$$

$$\frac{2\widehat{\sigma}}{\sqrt{n}}\sqrt{(a-1)F_{a-1,a(n-1)}(\alpha)\sum_i c_i^2},$$

于是区间短者为好.

例7.1.2

某单位研制出一种治疗头痛的新药, 现把此新药与阿司匹林和安慰剂(并不是真正的药, 而是生理盐水、葡萄糖剂等)作比较. 观测值为病人服药后头不痛所持续的时间. 数据按药的品种列入下表.

Table: 持续的时间表

药的品种	观测值y_{ij}	n_i	$\bar{y}_{i\cdot}$
安慰剂	0.0, 1.0	2	0.5
新药	2.3, 3.5, 2.8, 2.5	4	2.775
阿司匹林	3.1, 2.7, 3.8	3	3.2

同时置信区间的案例分析

Table: 数据方差分析表

方差源	自由度	平方和	均　方	F值	P值
组间差	2	9.701	4.851	14.94	0.00467
组内差(误差)	6	1.948	0.325		
总和	8	11.649			

R程序

```
x <- c(0.0,1.0,2.3,3.5,2.8,2.5,3.1,2.7,3.8)
A <- factor(c(rep(1,2),rep(2,4),rep(3,3)))
X.aov <- aov(x~A)
summary(X.aov)
```

- 取显著性水平$\alpha = 0.05$, 因为
  $P(F_{2,6} > 14.94) = 0.00467 < 0.05$ (或$F_{2,6} = 5.14 < F = 14.94$),
  所以拒绝原假设, 认为三种药效有显著差异.
- 作新药与安慰剂, 新药与阿司匹林的效应之差$\alpha_i - \alpha_j$, $(i > j)$的同时置信区间. 因为这个例子是非平衡模型, 所以不能应用Tukey方法构造同时置信区间. 我们首先计算

$$t_6\left(\frac{0.05}{2 \times 3}\right)\hat{\sigma}\left(\frac{1}{3}+\frac{1}{4}\right) = 1.430,$$

$$\hat{\sigma}\left(2F_{2,6}\left(0.05\right)\left(\frac{1}{3}+\frac{1}{4}\right)\right)^{1/2} = 1.396.$$

- $\alpha_3-\alpha_2, \alpha_3-\alpha_1$ 和$\alpha_2-\alpha_1$的置信系数为95%的同时置信区间
  - Bonferroni同时置信区间

$$[-1.005,\ 1.855],\quad [0.990,\ 4.410],\quad [0.653,\ 3.897].$$

  - Scheffé 同时置信区间

$$[-0.971,\ 1.821],\quad [1.032,\ 4.368],\quad [0.693,\ 3.857].$$

对于这些结果我们可以得到如下结论:

- 由此可认为新药比安慰剂显著有效(置信区间不包含0), 但新药与阿司匹林的疗效无显著差异(置信区间不包含0).
- 对这个例子, 从Bonferroni区间和Scheffé 区间所得出的结论一致, 但Bonferroni区间比Scheffé区间要长.

在所有其他因子都处于完全控制状态下, 研究因子A和B各个水平对因变量Y的影响.

两向分类模型(无交互)

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \quad i = 1, \cdots, a,\ j = 1, \cdots, b, \tag{2.6}$$

满足约束条件：$\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0,$

- 假设因子A有a个水平: $A_1, \cdots, A_a$, 因子B有b个水平: $B_1, \cdots, B_b$.
- y_{ij}为在A_i和B_j水平下试验的观测值.
- α_i和β_j分别表示水平A_i和B_j的效应, μ表示总平均,
- 假设$e_{ij} \sim N(0, \sigma^2)$, 且相互独立.

两向分类模型(无交互)矩阵表达

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{e},\ \text{这里},\ \boldsymbol{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n),$$

$$\begin{aligned}
\boldsymbol{y} &= (y_{11}, y_{12}, \cdots, y_{1b}, y_{21}, \cdots, y_{2b}, \cdots, y_{a1}, y_{a2}, \cdots, y_{ab})', \\
\boldsymbol{\gamma} &= (\mu, \alpha_1, \alpha_2, \cdots, \alpha_a, \beta_1, \beta_2, \cdots, \beta_b)', \\
\boldsymbol{e} &= (e_{11}, e_{12}, \cdots, e_{1b}, e_{21}, \cdots, e_{2b}, \cdots, e_{a1}, e_{a2}, \cdots, e_{ab})',
\end{aligned}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_b & \mathbf{1}_b & & & \mathbf{I}_b \\ \mathbf{1}_b & & \mathbf{1}_b & & \mathbf{I}_b \\ \vdots & & & \ddots & \vdots \\ \mathbf{1}_b & & & \mathbf{1}_b & \mathbf{I}_b \end{pmatrix} = (\mathbf{1}_{ab},\ \mathbf{I}_a \otimes \mathbf{1}_b,\ \mathbf{1}_a \otimes \mathbf{I}_b),$$

其中$\otimes$ 表示矩阵的Kronecker乘积, $\mathrm{rk}(\mathbf{X}) = a + b - 1 <$ 列数

先导出参数的一组LS解, 再表征所有可估函数.

- 正则方程$\mathbf{X}'\mathbf{X}\gamma = \mathbf{X}'\boldsymbol{y}$为

$$\begin{cases} ab\mu + b\sum_{i=1}^{a}\alpha_i + a\sum_{j=1}^{b}\beta_j = y_{\cdot\cdot}, \\ b\mu + b\alpha_i + \sum_{j=1}^{b}\beta_j = y_{i\cdot}, \quad i = 1, \cdots, a, \\ a\mu + \sum_{i=1}^{a}\alpha_i + a\beta_j = y_{\cdot j}, \quad j = 1, \cdots, b, \end{cases} \tag{2.7}$$

其中$y_{\cdot\cdot} = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}$, $y_{i\cdot} = \sum_{j=1}^{b} y_{ij}$, $y_{\cdot j} = \sum_{i=1}^{a} y_{ij}$.

方程个数为$a+b-1$个，而未知参数有$a+b+1$个.

参数估计

- 引进如下边界条件

$$\sum_{i=1}^{a}\alpha_i = 0, \quad \sum_{j=1}^{b}\beta_j = 0. \tag{2.8}$$

把这两个条件加入到方程组(7.2.3)中. 正则方程组变为

$$\begin{cases} ab\mu = y_{\cdot\cdot}, \\ b\mu + b\alpha_i = y_{i\cdot}, \quad i = 1, \cdots, a, \\ a\mu + a\beta_j = y_{\cdot j}, \quad j = 1, \cdots, b. \end{cases} \tag{2.9}$$

- 解得一组LS解（不能说LS估计, 因为α_i 和β_j不可估）

$$
\begin{aligned}
\hat{\mu} &= \frac{1}{ab} y_{\cdot\cdot} \triangleq \bar{y}_{\cdot\cdot}, \\
\hat{\alpha}_i &= \frac{1}{b} y_{i\cdot} - \widehat{\mu} = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}, \quad i = 1, \cdots, a, \\
\hat{\beta}_j &= \frac{1}{a} y_{\cdot j} - \widehat{\mu} = \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}, \quad j = 1, \cdots, b.
\end{aligned}
$$

- 感兴趣的是分别比较因子A和B各水平的效应, 形如

$$
\sum_{i=1}^{a} c_i \alpha_i \quad 和 \quad \sum_{j=1}^{b} d_j \beta_j,
$$

其中$\sum_{i=1}^{a} c_i = 0$, $\sum_{j=1}^{a} d_i = 0$ 为线性组合可估的充要条件 .

定理7.2.1

对于两向分类模型(7.2.1),

(1) $\sum_{i=1}^{a} c_i\alpha_i$可估当且仅当$\sum_{i=1}^{a} c_i = 0$, 即$\sum_{i=1}^{a} c_i\alpha_i$是一个对照, 这时, 它的BLU估计为$\sum_{i=1}^{a} c_i\bar{y}_{i\cdot}$;

(2) $\sum_{j=1}^{b} d_j\beta_j$ 可估当且仅当$\sum_{j=1}^{b} d_j = 0$, 即$\sum_{j=1}^{b} d_j\beta_j$为一对照, 这时, 它的BLU估计为$\sum_{j=1}^{b} d_i\widehat{\beta}_j = \sum_{j=1}^{b} d_j\bar{y}_{\cdot j}$.

- 任意$\alpha_i - \alpha_{i'}$, $\beta_j - \beta_{j'}$ 都可估, 它们的BLU估计分别为

$$\widehat{\alpha}_i - \widehat{\alpha}_{i'} = \bar{y}_{i\cdot} - \bar{y}_{i'\cdot}, \quad \widehat{\beta}_j - \widehat{\beta}_{j'} = \bar{y}_{\cdot j} - \bar{y}_{\cdot j'}$$

因子显著性检验

对于两向分类模型, 两个主要感兴趣的问题是考虑

因子A的a个水平效应是否有显著差异, 即检验假设

$$H_{01}:\ \alpha_1 = \alpha_2 = \cdots = \alpha_a, \tag{2.10}$$

因子B的b个水平的效应是否有显著差异, 即检验假设

$$H_{02}:\ \beta_1 = \beta_2 = \cdots = \beta_b. \tag{2.11}$$

- H_{01} 成立, 模型变成单向分类模型

$$y_{ij} = \mu + \beta_j + e_{ij}, \quad i = 1, \cdots, a,\ j = 1, \cdots, b,$$

- H_{02} 成立, 模型变成单向分类模型

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, \cdots, a,\ j = 1, \cdots, b$$

根据5.1节, 两向分类模型(7.2.1)的回归平方和

$$\mathrm{RSS}(\mu,\alpha,\beta)=\frac{y_{\cdot\cdot}^2}{ab}+\left(\sum_{i=1}^{a}\frac{y_{i\cdot}^2}{b}-\frac{y_{\cdot\cdot}^2}{ab}\right)+\left(\sum_{j=1}^{b}\frac{y_{\cdot j}^2}{a}-\frac{y_{\cdot\cdot}^2}{ab}\right).$$

残差平方和为

$$\mathrm{SS}_e=\boldsymbol{y}'\boldsymbol{y}-\mathrm{RSS}(\mu,\alpha,\beta)=\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y}_{\cdot\cdot})^2.$$

- 若H_{01}为真, 约简模型$y_{ij}=\mu+\beta_j+e_{ij}$ 的回归平方和

$$\mathrm{RSS}(\mu,\beta)=\frac{y_{\cdot\cdot}^2}{ab}+\left(\sum_{j=1}^{b}\frac{y_{\cdot j}^2}{a}-\frac{y_{\cdot\cdot}^2}{ab}\right).$$

- 若H_{01}为真, 因子A的平方和为

$$\mathrm{SS}_A = \mathrm{RSS}(\mu, \alpha, \boldsymbol{\beta}) - \mathrm{RSS}(\mu, \boldsymbol{\beta}) = \sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

检验假设H_{01}的F统计量为

$$F_1 = \frac{\mathrm{SS}_A/(a-1)}{\mathrm{SS}_e/(a-1)(b-1)} = \frac{\mathrm{MS}_A}{\mathrm{MS}_e}. \tag{2.12}$$

当H_{01}为真时, $F_1 \sim F_{a-1,(a-1)(b-1)}$.

- 若H_{02}为真, 因子B的平方和为

$$\mathrm{SS}_B = \sum_{j=1}^{b} \frac{y_{\cdot j}^2}{a} - \frac{y_{\cdot\cdot}^2}{ab} = \sum_{i=1}^{a}\sum_{j=1}^{b} (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2.$$

  检验假设H_{02}的F统计量为

$$F_2 = \frac{\mathrm{SS}_B/(b-1)}{\mathrm{SS}_e/(a-1)(b-1)} = \frac{MS_B}{MS_e}.$$

  当H_{02}为真时, $F_2 \sim F_{b-1,(a-1)(b-1)}$.

- 具体的计算可以由R语言中的方差分析函数aov()直接得到.

Table: 无重复试验无交互效应两因素方差分析表

方差源	自由度	平方和	均　方	F值
因子A	$a-1$	SS_A	$MS_A = SS_A/(a-1)$	$F_1 = \dfrac{MS_A}{MS_e}$
因子B	$b-1$	SS_B	$MS_B = SS_B/(b-1)$	$F_2 = \dfrac{MS_B}{MS_e}$
误差	$(a-1)(b-1)$	SS_e	$MS_e = SS_e/(a-1)(b-1)$	
总和	$ab-1$	$SS_T = SS_A + SS_B + SS_e$		

例7.2.1

一种火箭使用了四种燃料、三种推进器进行射程试验. 对于每种燃料与推进器的组合作一次试验, 得到的试验数据见下表, 问各种燃料之间及各种推进器之间有无显著差异?

Table: 火箭试验数据

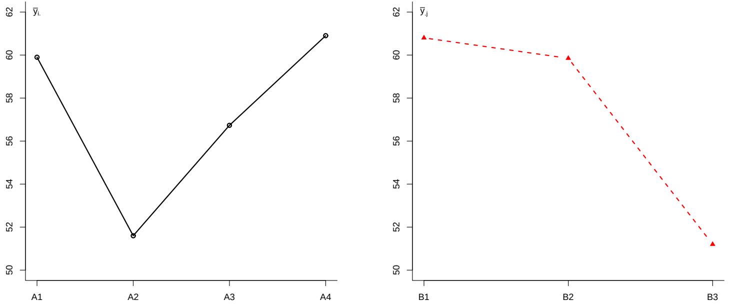
		推进器(B)			
		B_1	B_2	B_3	$y_{i\cdot}$
燃料(A)	A_1	58.2	56.2	65.3	179.7
	A_2	49.1	54.1	51.6	154.8
	A_3	60.1	70.9	39.2	170.2
	A_4	75.8	58.2	48.7	182.7
$y_{\cdot j}$		243.2	239.4	204.8	

- 计算得4 种燃料各自的火箭射程样本均值:

$$(\bar{y}_{1\cdot}, \bar{y}_{2\cdot}, \bar{y}_{3\cdot}, \bar{y}_{4\cdot}) = (59.900, 51.600, 56.733, 60.900)$$

和3 种推进器各自的火箭射程样本均值:

$$(\bar{y}_{\cdot 1}, \bar{y}_{\cdot 2}, \bar{y}_{\cdot 3}) = (60.80, 59.85, 51.20).$$



例7.2.1

- 在显著性水平为$\alpha = 0.05$下检验假设

$$H_{01}: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4,$$

$$H_{02}: \beta_1 = \beta_2 = \beta_3.$$

方差分析计算结果：

Table: 火箭数据方差分析表

方差源	自由度	平方和	均　方	F 值	P 值
因子A	3	157.6	52.53	$F_1 = 0.431$	0.739
因子B	2	223.8	111.92	$F_2 = 0.917$	0.449
误差	6	732.0	122.00		
总和	11	1113.4			

- 因子A对应的F检验的P 值=0.739 > 0.05 或

$$F_{3,6}(0.05) = 4.76 > F_1 = 0.431,$$

接受H_{01}, 认为四种燃料对火箭射程无显著影响.

- 因子B对应的F检验的P值=0.449 > 0.05 （或$F_{2,6}(0.05) = 5.14 > F_2 = 0.917$），接受$H_{02}$.
- 综合两个检验的结果, 我们可以认为各种燃料和各种推进器之间的差异对火箭射程无显著影响.

由于该例为平衡数据, 可用R语言中的方差分析函数aov()，程序如下

```
y <- c(58.2,56.2,65.3,49.1,54.1,51.6,60.1,70.9,39.2,75.8,58.2,48.7)
A <- factor(gl(4,3,12)); B <- factor(gl(3,1,12))
y.aov <- aov(y~A+B)
summary(y.aov)
```

同时置信区间

- 如果经F检验, 假设H_{01}被拒绝, 则表明因子A 的a 个水平的效应不全相等. 和单向分类模型一样, 这时我们希望构造对照$\alpha_i - \alpha_{i'}$的同时置信区间.
- 如果H_{02} 被拒绝, 则表明因子B 的b 个水平的效应不全相等, 于是构造对照$\beta_j - \beta_{j'}$ 的同时置信区间.
- 两者构造类似，只介绍一种.
- 所有形如$\alpha_i - \alpha_{i'}, i \neq i'$ 的对照有$a-1$个线性无关.

同时置信区间

任意m个$\alpha_i - \alpha_{i'}, i \neq i'$ 的置信系数为$1-\alpha$的同时置信区间

- Bonferroni 同时置信区间:

$$(\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}) \pm t_{(a-1)(b-1)}\left(\frac{\alpha}{2m}\right)\hat{\sigma}\sqrt{\frac{2}{b}}.$$

- Scheffé同时置信区间:

$$(\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}) \pm \hat{\sigma}\sqrt{(a-1)F_{a-1,(a-1)(b-1)}(\alpha)\left(\frac{2}{b}\right)}.$$

- Tukey同时置信区间:

$$(\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}) \pm q_{a,(a-1)(b-1)}(\alpha)\frac{\hat{\sigma}}{\sqrt{b}}.$$

例7.2.2

为了考察高温合金中碳的含量(因子A)和锑与铝的含量之和(因子B)对合金强度的影响. 因子A取3 个水平$0.03, 0.04, 0.05$ (上述数字表示碳的含量占合金总量的百分比), 因子B取4个水平$3.3, 3.4, 3.5, 3.6$ (上述数字意义同上). 在每个水平组合下各作一次试验, 试验结果如表7.2.4所示

Table: 合金强度试验数据

		B 锑与铝的含量之和				
		3.3	3.4	3.5	3.5	$y_{i\cdot}$
碳	0.03	63.1	63.9	65.6	66.8	259.4
含	0.04	65.1	66.4	67.8	69.0	268.3
量	0.05	67.2	71.0	71.9	73.5	283.6
$y_{\cdot j}$		195.4	201.3	205.3	209.3	

解 计算的方差分析结果如下：

Table: 方差分析表

方差源	自由度	平方和	均方	F值	P值
因子A	2	74.91	37.46	$F_1 = 70.05$	6.93e-05***
因子B	3	35.17	11.72	$F_2 = 21.92$	0.00124**
误差	6	3.21	0.535		
总和	11	113.29			

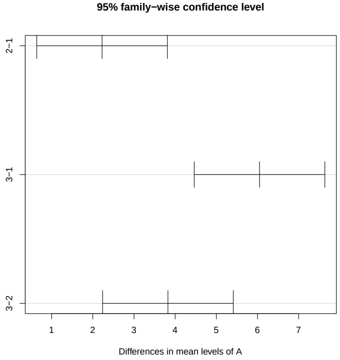
- 当显著性水平$\alpha = 0.05$ 时, 两因子对应的F检验的P值都远远小于0.05, 故因子A的3 个水平之间和因子B的4个水平之间对合金强度的影响都有显著差异.

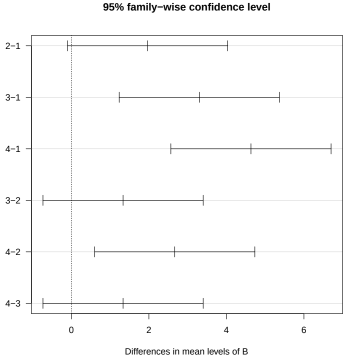
- 构造同时置信区间.

  计算Tukey 同时置信区间. 相应程序如下:

```
y <- c(63.1,63.9,65.6,66.8,65.1,66.4,67.8,69.0,67.2,71.0,71.9,73.5)
A <- factor(gl(3,4,12)); B <- factor(gl(4,1,12))
y.aov <- aov(y~A+B)
TukeyHSD(y.aov)
TukeyHSD(y.aov,"A", ordered = TRUE)
plot(TukeyHSD(y.aov, "A"), las = 1)
TukeyHSD(y.aov, "B", ordered = TRUE)
plot(TukeyHSD(y.aov,"B"), las = 1)
```

- 得到因子A 和B 各自的水平间对照的95%的Tukey同时置信区间





从图可以发现因子B的效应β_1和β_2, β_2和β_3 及β_3和β_4 的差的置信区间包含0点, 故这三对效应不存在显著差异.

用Bonferroni及Scheffé同时置信区间计算, 除了区间不同外, 最后得到的结论和Tukey方法一致.

Tukey可加性检验

针对每个单元仅有一次观测方差分析模型, Tukey (1949) 提出了一种检验交互效应的方法.

- 为了可估性, 假设两主效应的交互效应为γ_{ij}, 满足如下关系:

$$\gamma_{ij} = D\alpha_i\beta_j,$$

其中D 是一个常数. 两向分类模型可表示为

$$y_{ij} = \mu + \alpha_i + \beta_j + D\alpha_i\beta_j + e_{ij}, \quad i = 1, \cdots, a, \quad j = 1, \cdots, b.$$

于是两个主效应是否存在交互效应问题等价于检验

$$H_0 : D = 0 \longleftrightarrow H_1 : D \neq 0.$$

Tukey可加性检验

- 交互效应平方和（采用两步估计方法）

$$\mathrm{SS}_{AB} = \frac{\left[\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})y_{ij}\right]^2}{\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \sum_{j=1}^{b}(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2}.$$

- 相应的残差平方和$\mathrm{SS}_e = \mathrm{SS}_T - \mathrm{SS}_A - \mathrm{SS}_B - \mathrm{SS}_{AB}$,
- F 统计量

$$F = \frac{\mathrm{SS}_{AB}}{\mathrm{SS}_e/(ab-a-b)}.$$

当H_0成立时, F倾向于取较小的值, 当它较大时, 就可判定H_1成立. 拒绝域近似为$F > F_{1,ab-a-b}(\alpha)$.

这个检验被称为Tukey的可加性检验或Tukey单自由度检验.

例7.2.3

考察例7.2.1中燃料与推进器对火箭射程是否存在交互效应.

解 本例中, $a=4$, $b=3$. 计算

$$SS_A = 3\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2 = 157.590, \quad SS_B = a\sum_{j=1}^{b}(\bar{y}_{.j}-\bar{y}_{..})^2 = 223.847,$$

$$SS_{AB} = \frac{3\times 4\times 234.734^2}{157.590\times 223.847} = 18.744,$$

$$SS_e = 1113.417 - 157.590 - 223.847 - 18.744 = 713.236.$$

得检验统计量

$$F = \frac{SS_{AB}}{SS_e/(ab-a-b)} = 0.131 < F_{1,5}(0.05) = 6.608,$$

因此认为火箭射程研究中燃料与推进器的交互效应不显著.

两向分类模型中, 要分析交互效应, 需在各个水平组合下作重复试验.

两向分类模型(交互效应存在)

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \tag{7.3.1}$$

$$i = 1, \cdots, a, \quad j = 1, \cdots, b, \quad k = 1, \cdots, c,$$

其中γ_{ij}为水平A_i 和B_j的交互效应.

- 模型未知参数个数为$(a + b + ab + 1)$.
- 设计阵$\mathbf{X}$的秩$\mathrm{rk}(\mathbf{X}) = ab <$ 列数$a + b + ab + 1$.
- 要使得LS解唯一, 需要附加$a + b + 1$ 个独立约束条件.

定理7.3.1

有交互效应的两向分类模型(7.3.1), 下列ab个函数构成了极大线性无关的可估函数组

$$\alpha_i-\alpha_{i+1}+\overline{\gamma}_{i\cdot}-\overline{\gamma}_{(i+1)\cdot},\quad i=1,\cdots,a-1,$$

$$\beta_j-\beta_{j+1}+\overline{\gamma}_{\cdot j}-\overline{\gamma}_{\cdot(j+1)},\quad j=1,\cdots,b-1,$$

$$\delta_{ij}\triangleq\gamma_{ij}-\overline{\gamma}_{i\cdot}-\overline{\gamma}_{\cdot j}+\overline{\gamma}_{\cdot\cdot}\,,\quad i=1,\cdots,a-1,\quad j=1,\cdots,b-1,$$

$$\mu+\frac{1}{a}\sum_{i=1}^{a}\alpha_i+\frac{1}{b}\sum_{j=1}^{b}\beta_j+\overline{\gamma}_{\cdot\cdot}.$$

从关系式$\mu_{ij}=\mu+\alpha_i+\beta_j+\gamma_{ij}$可以看出：

- 当$\delta_{ij}\neq 0$时，α_i并不能反映因子水平A_i的优劣. 因为因子水平A_i的优劣还与因子B的水平有关. 可估函数

$$\overline{\mu}_{i\cdot}=\mu+\alpha_i+\frac{1}{b}\sum_{j=1}^{b}\beta_j+\overline{\gamma}_{i\cdot}$$

和

$$\overline{\mu}_{(i+1)\cdot}-\overline{\mu}_{(i)\cdot}=\alpha_{i+1}-\alpha_i+\overline{\gamma}_{(i+1)\cdot}-\overline{\gamma}_{(i+1)\cdot}.$$

是在因子B的诸水平求平均的意义下, 分别对因子水平A_i优劣的度量和对水平A_{i+1}和A_i 的效应差异的度量.

若取参数边界条件：

$$\sum_{i=1}^{a} \alpha_i = 0, \qquad \sum_{j=1}^{b} \beta_j = 0,$$

$$\sum_{i=1}^{a} \gamma_{ij} = 0, \quad j = 1, \cdots, b, \quad \sum_{j=1}^{b} \gamma_{ij} = 0, \quad i = 1, \cdots, a.$$

注 上边界条件只有$a + b + 1$ 个独立条件(因为$\sum_{i=1}^{a} \sum_{j=1}^{b} \gamma_{ij} = 0$).

- 得参数$\mu, \alpha_i, \beta_j, \gamma_{ij}$的一组特定的LS解：

$$\begin{aligned}
\widehat{\mu} &= \bar{y}_{\cdots} \quad , & & \\
\widehat{\alpha}_i &= \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots}, & & i = 1, \cdots, a, \\
\widehat{\beta}_j &= \bar{y}_{\cdot j \cdot} - \bar{y}_{\cdots}, & & j = 1, \cdots, b, \\
\widehat{\gamma}_{ij} &= \bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}, & & i = 1, \cdots, a, \ j = 1, \cdots, b.
\end{aligned}$$

在参数约束条件下,

- $\delta_{ij} = \gamma_{ij}$ 交互效应.
- 若$\delta_{ij} = 0$, 模型就化为无交互效应的两向分类模型.

  这也说明了δ_{ij} 度量了A_i和B_j的交互效应.
- 下列ab个函数构成了极大线性无关的可估函数组

$$
\begin{aligned}
&\mu \\
&\alpha_i - \alpha_{i+1}, \quad i = 1, \cdots, a-1, \\
&\beta_j - \beta_{j+1}, \quad j = 1, \cdots, b-1, \\
&\gamma_{ij}, \quad i = 1, \cdots, a-1, \quad j = 1, \cdots, b-1
\end{aligned}
$$

假设检验

- 对有交互效应的两向分类模型, 单纯检验$\alpha_1 = \cdots = \alpha_a = 0$与检验$\beta_1 = \cdots = \beta_b = 0$都没有实际意义.
- 一个重要的检验问题是交互效应是否存在. 检验假设:

$$\gamma_{ij} = 0, \quad i = 1, 2, \cdots, a,\ j = 1, 2, \cdots, b.$$

但γ_{ij}不可估. 可以改为检验一个等价的假设

$$H_{01}: \quad \delta_{ij} = 0, \quad i = 1, 2, \cdots, a, \quad j = 1, 2, \cdots, b.$$

- 两向分类模型(交互效应存在)下, 回归平方和为

$$\mathrm{RSS}(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = c \sum_{i=1}^{a} \sum_{j=1}^{b} \bar{y}_{ij\,\cdot}^{2}.$$

残差平方和为

$$\begin{aligned} \mathrm{SS}_e &= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} y_{ijk}^{2} - \mathrm{RSS}(\mu, \alpha, \boldsymbol{,} \gamma) \\ &= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} (y_{ijk} - \bar{y}_{ij\,\cdot})^{2} \end{aligned}$$

- 在假设H_{01}下, 回归平方和为

$$\mathrm{RSS}(\mu,\boldsymbol{\alpha},\boldsymbol{\beta})=\frac{y_{\cdots}^2}{abc}+\left(\sum_{i=1}^{a}\frac{y_{i\cdot\cdot}^2}{bc}-\frac{y_{\cdots}^2}{abc}\right)+\left(\sum_{j=1}^{b}\frac{y_{\cdot j\cdot}^2}{ac}-\frac{y_{\cdots}^2}{abc}\right).$$

交互效应平方和

$$\begin{aligned}\mathrm{SS}_{AB}&=\mathrm{RSS}(\mu,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})-\mathrm{RSS}(\mu,\boldsymbol{\alpha},\boldsymbol{\beta})\\&=\sum_{i=1}^{a}\sum_{j=1}^{b}c(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot j\cdot}+\bar{y}_{\cdots})^2\end{aligned}$$

交互效应的显著性检验

- 检验假设H_{01}的F统计量为

$$F_{AB} = \frac{SS_{H_{01}}/[(a-1)(b-1)]}{SS_e/[ab(c-1)]} = \frac{SS_{AB}/[(a-1)(b-1)]}{SS_e/[ab(c-1)]}. \tag{3.1}$$

当H_{01}成立时，

$$F_{A\times B} \sim F_{(a-1)(b-1),ab(c-1)}.$$

对给定的显著性水平α，若

$$F_{AB} < F_{(a-1)(b-1),ab(c-1)}(\alpha),$$

则我们认为因子A与因子B的相互效应不存在．这时就可以回到7.2节内容去检验因子A和B 的各水平效应的差异．

主因子效应的显著性检验

若交互效应的存在, α_i并不能反映因子水平A_i的优劣, 这是因为因子水平A_i的优劣还与因子B的水平有关. 只能退而求其次, 讨论如下两个假设:

- 在因子B的平均水平意义下, 比较因子A的诸水平优劣, 即检验

$$H_{02}:\ \alpha_1+\overline{\gamma}_{i\cdot}=\cdots=\alpha_a+\overline{\gamma}_{a\cdot};$$

- 在因子A的平均水平意义下, 比较因子B的诸水平优劣, 即检验

$$H_{03}:\ \beta_1+\overline{\gamma}_{\cdot 1}=\cdots=\beta_a+\overline{\gamma}_{\cdot b}$$

主因子效应的显著性检验

- $H_{02}:\ \alpha_1+\overline{\gamma}_{i\cdot}=\cdots=\alpha_a+\overline{\gamma}_{a\cdot}$ 的检验统计量

$$F_A=\frac{SS_{H_{02}}/(a-1)}{SS_e/ab(c-1)}=\frac{SS_A/(a-1)}{SS_e/ab(c-1)}.$$

若H_2为真, $F_A\sim F_{a-1,ab(c-1)}$.

- $H_{03}:\ \beta_1+\overline{\gamma}_{\cdot 1}=\cdots=\beta_a+\overline{\gamma}_{\cdot b}$的检验统计量

$$F_B=\frac{SS_{H_{03}}/(b-1)}{SS_e/ab(c-1)}=\frac{SS_B/(b-1)}{SS_e/ab(c-1)}.$$

若H_3为真, $F_B\sim F_{b-1,ab(c-1)}$.

这里$SS_A=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(\overline{y}_{i\cdot\cdot}-\overline{y}_{\cdots})^2,\quad SS_B\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(\overline{y}_{\cdot j\cdot}-\overline{y}_{\cdots})^2.$

Table: 有交互效应两因素方差分析表

方差源	自由度	平方和	均方	F值
因子A	$a-1$	SS_A	$\mathrm{MS}_A=\mathrm{SS}_A/(a-1)$	$F_A=\frac{MS_A}{MS_e}$
因子B	$b-1$	SS_B	$\mathrm{MS}_B=\mathrm{SS}_B/(b-1)$	$F_B=\frac{MS_B}{MS_e}$
交互效应 $(A\times B)$	$(a-1)(b-1)$	SS_{AB}	$\mathrm{MS}_{AB}=\frac{\mathrm{SS}_{\mathrm{AB}}}{(a-1)(b-1)}$	$F_{AB}=\frac{\mathrm{MS}_{AB}}{\mathrm{MS}_e}$
误差	$ab(c-1)$	SS_e	$\mathrm{MS}_e=\mathrm{SS}_e/ab(c-1)$	
总和	$abc-1$	$\sum_{i,j,k}(y_{ijk}-y_{\ldots})^2/abc$		

例7.3.1

为了考察某种电池的最大输出电压受板极材料与使用电池的环境温度的影响, 材料类型(因子A)取3个水平(即3种不同的材料), 温度也取3个水平, 每个水平组合下重复4次试验. 电池试验数据如下

		温度 B						
		B_1	(15°)	B_2	(25°)	B_3	(35°)	$y_{i\cdot\cdot}$
材料类型 A	A_1	130	155	34	40	20	70	1098
		174	180	80	75	82	58	
		(639)		(229)		(230)		
	A_2	150	188	136	122	25	70	1300
		159	126	106	115	58	45	
		(623)		(479)		(198)		
	A_3	138	110	174	120	96	104	1501
		168	160	150	139	82	60	
		(576)		(583)		(342)		
$y_{\cdot j\cdot}$		1838		1291		770		$3899 = y_{\cdot\cdot\cdot}$

方差分析表

- 电池试验数据的方差分析表

方差源	自由度	平方和	均　方	F值	P值
因子A	2	6767.06	3383.53	$F_A = 6.727$	0.004261
因子B	2	47535.39	23767.70	$F_B = 47.253$	1.52e-09
交互效应 $(A \times B)$	4	13180.44	3295.11	$F_{AB} = 6.551$	0.000807
误差	27	13580.75	502.99		
总和	35	81063.64			

- 由于$F_{2,27}(0.05) = 3.35$, $F_{4,27}(0.05) = 2.73$, 所以因子A、因子B以及交互效应$A \times B$ 在显著性水平$\alpha = 0.05$下都是显著的.

```
y <- c(130,155,174,180,34,40,80,75, 20,70,82,58,150,188,
159,126, 136,122,106,115, 25,70,58,45,138,110,168,
160,174,120,150,139,96,104,82,60)
A <- factor(gl(3,12,36))
B <- factor(gl(3,4,36))
y.aov <- aov(y~A+B+A:B)
summary(y.aov)
```

注1 对于非平衡两向分类模型(n_{ij}不全相同), 可估函数的表征及检验统计量就没有以上简洁形式.

- 非平衡数据出现的常见原因:

  (1) 在观察性研究中通常很少或根本无法控制单元样本量.

  (2) 在试验性研究中也会遇到各单元样本量不等的情况(如记录遗失).

  (3) 无论是观察性研究还是试验性研究，研究者都可能在试验成本较低的处理单元采样较多，成本高的单元采样较少.

例如, 一家包装食品制造商希望测量其早餐谷物食品中玉米糖浆改为低热量甜味剂（因子A）对消费者产品评价的影响. 三类消费者（因素B：儿童、成年女性和成年男性）被认为是重要的. 据了解, 约60% 的消费者是儿童，20% 是成年男性，20% 是成年女性. 因此, 按比例抽样可为最重要的消费者群体提供更高的精确度.

非平衡数据下两向分类模型

$$y_{ijk} = \mu+\alpha_i+\beta_j+\gamma_{ij}+e_{ijk}, \;\; i=1,\cdots,a, \;\; j=1,\cdots,b, \;\; k=1,\cdots,n_{ij}, \tag{3.2}$$

这里n_{ij}不全相等($n_{ij}>0$).

- 最小二乘正则方程组不再简单易解
- 各因子效应平方和不再独立
- 它们的和不再等于总变差平方和

- 在参数边界约束条件下,

$$\alpha_a = -\sum_{i=1}^{a-1}\alpha_i,\quad \beta_b = -\sum_{j=1}^{b-1}\beta_j,\quad \gamma_{ib} = -\sum_{j=1}^{b-1}\gamma_{ij},\quad \gamma_{aj} = -\sum_{i=1}^{a-1}\gamma_{ij}.$$

为了表示模型的回归项, 我们引入虚拟变量,

$$A_{ijkg} = \begin{cases} 1, & i = g, \\ -1, & i = a, \\ 0, & i \neq l, a, \end{cases} \qquad g = 1, \cdots, a-1,$$

$$B_{ijkh} = \begin{cases} 1, & j = h, \\ -1, & j = b, \\ 0, & j \neq h, b, \end{cases} \qquad h = 1, \cdots, b-1.$$

- 在边界约束条件下上方差分析模型等价于约简回归模型

$$y_{ijk} = \mu + \sum_{g=1}^{a-1} A_{ijkg}\alpha_g + \sum_{h=1}^{b-1} B_{ijkh}\beta_h + \sum_{g=1}^{a-1}\sum_{h=1}^{b-1} A_{ijkg}B_{ijkh}\gamma_{gh} + e_{ijk},$$

其中未知回归参数个数为$1+(a-1)+(b-1)+(a-1)(b-1) = ab$，且以上$ab$个参数皆可估. 记

- SS_e 为全模型的残差平方和;
  $SS_e(A)$ 为模型$y_{ijk} = \mu + \sum_{g=1}^{a-1} A_{ijkg}\alpha_g + e_{ijk}$的残差平方和;
  $SS_e(A,B)$ 为模型$y_{ijk} = \mu + \sum_{g=1}^{a-1} A_{ijkg}\alpha_g + \sum_{h=1}^{b-1} B_{ijkh}\beta_h + e_{ijk}$的残差平方和;
  $SS_e(AB)$ 为模型$y_{ijk} = \mu + \sum_{i=1}^{a-1}\sum_{i=1}^{b-1} A_{ijk}B_{ijk}\gamma_{ij} + e_{ijk}$的残差平方和;
  $SS_e(A,AB)$ 为去掉B因子($SS_e(B,AB)$ 为去掉A因子)后的残差平方和.

非平衡数据下方差分析中平方和三种类型定义：
I类型(序贯型)；II类型(分层型)；III类型(边界型)

- I类型(序贯型)

首先, 求A对y的影响, 令因子A的平方和为

$$R(\alpha|\mu) = SS(A) = SS_T - SS_e(A) = \sum_{i=1}^{a} n_{i\cdot}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2;$$

其次, 控制A，求B对y的影响, 令因子B的调整平方和为

$$R(\beta|\mu,\alpha) = SS_e(A) - SS_e(A,B);$$

最后, 控制A和B的主效应, 求A与B的交互效应对y的影响, 令

$$R(\gamma|\mu,\alpha,\beta) = SS_e(A,B) - SS_e.$$

R软件中的aov()默认使用的是I 型平方和.

$H_{01}: \alpha_1 = \alpha_2 = \cdots = \alpha_a$的检验统计量为

$$F_1^{(I)} = \frac{R(\boldsymbol{\alpha}|\mu)/(a-1)}{\mathrm{SS}_e/(N-ab)}$$

$H_{02}: \beta_1 = \beta_2 = \cdots = \beta_b$ 的检验统计量为

$$F_2^{(I)} = \frac{R(\boldsymbol{\beta}|\mu, \boldsymbol{\alpha})/(b-1)}{\mathrm{SS}_e/(N-ab)}$$

H_{03} : 所有的$\gamma_{ij} = 0$ 的检验统计量为

$$F_3 = \frac{R(\boldsymbol{\gamma}|\mu, \boldsymbol{\alpha}, \boldsymbol{\beta})/(a-1)(b-1)}{\mathrm{SS}_e/(N-ab)}.$$

- 对于非平衡数据情形, I型平方和仍满足平方和分解:

$$\mathrm{SS}_T = R(\boldsymbol{\alpha}|\mu) + R(\boldsymbol{\beta}|\mu, \boldsymbol{\alpha}) + R(\boldsymbol{\gamma}|\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}) + \mathrm{SS}_e.$$

- $R(\boldsymbol{\alpha}|\mu)$, $R(\boldsymbol{\beta}|\mu, \boldsymbol{\alpha})$和$R(\boldsymbol{\gamma}|\mu, \boldsymbol{\alpha}, \boldsymbol{\beta})$都与残差平方和$\mathrm{SS}_e$独立.
- $F_1^{(I)}$ 和$F_2^{(I)}$ 也不适合作为H_{01}和H_{02}的检验统计量,

  (因为$R(\boldsymbol{\alpha}|\mu)$ 和$R(\boldsymbol{\beta}|\mu, \boldsymbol{\alpha})$的零分布是非中心的$\chi^2$).
- 关于交互效应的显著性检验, F_3仍适合作为检验统计量.

I型平方和适合用于平衡数据的情形以及不平衡数据单向分类模型的情形，这也是R软件中的函数aov()的适用范围.

- II型平方和（分层型）

  效应根据同水平或低水平的效应做调整. 对因子A的平方和也根据因子B做调整, 令其调整平方和为

$$R(\alpha|\mu,\beta) = SS_e(B) - SS_e(A,B),$$

  关于因子B调整平方和以及A与B的交互效应的调整平方和与I型平方和相同.

II型平方和适合平衡数据的方差分析模型、无交互效应的方差分析模型以及纯线性回归模型.

- III型平方和（边界型）

  令因子A、B的调整平方和分别为

$$R(\alpha|\mu,\beta,\gamma)=\mathrm{SS}_e(B,AB)-\mathrm{SS}_e,$$

$$R(\beta|\mu,\alpha,\gamma)=\mathrm{SS}_e(A,AB)-\mathrm{SS}_e,$$

  当H_{01}成立时，调整的统计量

$$F_1=\frac{R(\alpha|\mu,\beta,\gamma)/(a-1)}{\mathrm{SS}_e/(N-ab)}\sim F_{(a-1),(N-ab)};$$

  当H_{02}成立时，调整的统计量

$$F_2=\frac{R(\beta|\mu,\alpha,\gamma)/(b-1)}{\mathrm{SS}_e/(N-ab)}\sim F_{(b-1),(N-ab)}$$

III型平方和适用于不平衡数据下大部分方差分析模型.

方差分析中三种类型平方和的定义

Table: 方差分析中三种类型平方和的定义

效应平方和	I 型平方和	II型平方和	III型平方和
A	$\mathrm{R}(\boldsymbol{\alpha}\vert\mu)$	$\mathrm{R}(\boldsymbol{\alpha}\vert\mu,\boldsymbol{\beta})$	$\mathrm{R}(\boldsymbol{\alpha}\vert\mu,\boldsymbol{\beta},\gamma)$
B	$\mathrm{R}(\boldsymbol{\beta}\vert\mu,\boldsymbol{\alpha})$	$\mathrm{R}(\boldsymbol{\beta}\vert\mu,\boldsymbol{\alpha})$	$\mathrm{R}(\boldsymbol{\beta}\vert\mu,\boldsymbol{\alpha},\gamma)$
AB	$\mathrm{R}(\gamma\vert\mu,\boldsymbol{\alpha},\boldsymbol{\beta})$	$\mathrm{R}(\gamma\vert\mu,\boldsymbol{\alpha},\boldsymbol{\beta})$	$\mathrm{R}(\gamma\vert\mu,\boldsymbol{\alpha},\boldsymbol{\beta})$

- I 型平方和的模型是“从无到有”, 逐步增加的.
- III型平方和都是用全模型与缺少某效应的模型比较.
- 平衡设计下，三种类型等价.
- 不平衡设计下, III型平方和适用于大部分情形.

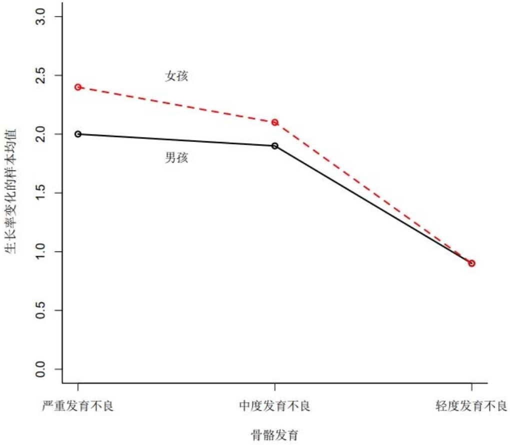
例7.3.2

一家临床研究中心为生长激素缺乏、尚未进入青春期的矮小儿童注射合成生长激素. 研究人员对儿童的性别（因素A）和骨骼发育（因素B）对激素诱导生长速度的影响很感兴趣. 儿童的骨骼发育分为三类：严重发育不良、中度发育不良和轻度发育不良. 每个性别-骨骼发育组随机抽取3 名儿童. 我们关注的反应变量（Y）是生长激素治疗期间的生长速度与治疗前正常生长速度之间的差异.

Table: 生长激素治疗期间的生长速度与治疗前正常生长速度的差(厘米/月)

性别（因素A）	骨骼发育（因素B）严重发育不良	中度发育不良	轻度发育不良
男	1.4, 2.4, 2.2	2.1, 1.7	0.7, 1.1
女	2.4	2.5, 1.8, 2.0	0.5, 0.9, 1.3

- 3类不同骨骼发育下男孩和女孩各自生长率变化的样本均值



解 考虑两向分类模型

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}.$$

由于不平衡数据，采用虚拟变量A_{ijk1}, B_{ijk1},B_{ijk2}, 得回归模型

$$y_{ijk} = \mu + A_{ijk1}\alpha_1 + B_{ijk1}\beta_1 + B_{ijk2}\beta_2 + A_{ijk1}B_{ijk1}\gamma_{11} + A_{ijk1}B_{ijk2}\gamma_{12} + e_{ijk}$$

记$\boldsymbol{y} = (1.4, 2.4, 2.2, 2.1, 1.7, 0.7, 1.1, 2.4, 2.5, 1.8, 2.0, 0.5, 0.9, 1.3)'$, α_1,β_1, β_2, γ_{11},γ_{12}相应的虚拟变量向量分别为

$(1, 1, 1, 1, 1, 1, 1, -1, -1, -1, -1, -1, -1, -1)' \triangleq \boldsymbol{x}_1$,

$(1, 1, 1, 0, 0, -1, -1, 1, 0, 0, 0, -1, -1, -1)' \triangleq \boldsymbol{x}_2$,

$(0, 0, 0, 1, 1, -1, -1, 0, 1, 1, 1, -1, -1, -1)' \triangleq \boldsymbol{x}_3$,

$(1, 1, 1, 0, 0, -1, -1, -1, 0, 0, 0, 1, 1, 1)' \triangleq \boldsymbol{x}_4$,

$(0, 0, 0, 1, 1, -1, -1, 0, -1, -1, -1, 1, 1, 1)' \triangleq \boldsymbol{x}_5$.

例7.3.2

记

$$\mathbf{X} = (\mathbf{1}_{14}, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5),$$

$$\boldsymbol{\theta} = (\mu, \alpha_1, \beta_1, \beta_2, \gamma_{11}, \gamma_{12})',$$

$\boldsymbol{e}$为相应的误差向量. 则线性模型的矩阵形式为

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{e}.$$

- 方差分析模型下关于交互效应、A、B主因子效应的显著性检验问题就转化成检验线性模型的某些回归系数是否为零的问题.

- 关于交互效应的显著性检验. 原假设和备择假设等价于

$$H_0:\gamma_{11}=\gamma_{12}=0 \longleftrightarrow H_1:\gamma_{11}\text{ 和}\gamma_{11}\text{不全为零}.$$

H_0成立, 约简模型为

$$y_{ijk}=\mu+A_{ijk1}\alpha_1+B_{ijk1}\beta_1+B_{ijk2}\beta_2+e_{ijk}. \tag{3.3}$$

分别计算原模型和约简模型下的LS 残差平方和, 分别为

$$\mathrm{SS}_e=\boldsymbol{y}'(\mathbf{I}_{14}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{y}=1.3000,$$

$$\mathrm{SS}_e(A,B)=\boldsymbol{y}'(\mathbf{I}_{14}-\mathbf{X}_0(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0')\boldsymbol{y}=1.3754,$$

其中$\mathbf{X}_0=(\mathbf{1}_{14},\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3)$. 它们的自由度分别为8和10.

- 于是检验统计量

$$F_3 = \frac{(\mathrm{SS}_e(A,B) - \mathrm{SS}_e)/(10-8)}{\mathrm{SS}_e/8} = 0.2321.$$

在显著性水平$\alpha = 0.05$下, 由于$F_3 = 0.2321 < F_{2,8}(0.05) = 4.46$, 因此接收$H_0$, 认为儿童的性别($A$) 和骨骼发育($B$) 的交互效应对激素诱导生长速度的影响不显著.

该检验统计量的P值为0.7980.

- 关于主因子效应的显著性检验.

  检验儿童的性别A 或骨骼发育B对激素诱导生长速度的影响是否显著. 原假设和备择假设等价于

$$H_{01}:\alpha_1=0\longleftrightarrow H_{11}:\alpha_1\neq 0,$$

$$H_{02}:\beta_1=\beta_2=0\longleftrightarrow H_{12}:\beta_1 \text{ 和}\beta_2\text{不全为零}.$$

于是在H_{01}, H_{02}下的相应约简模型分别

$$y_{ijk}=\mu+B_{ijk1}\beta_1+B_{ijk2}\beta_2+A_{ijk1}B_{ijk1}\gamma_{11}+A_{ijk1}B_{ijk2}\gamma_{12}+e_{ijk},$$

$$`y_{ijk}=\mu+A_{ijk1}\alpha_1+A_{ijk1}B_{ijk1}\gamma_{11}+A_{ijk1}B_{ijk2}\gamma_{12}+e_{ijk}.$$

计算约简模型(??)和(??)下的LS残差平方和, 分别得

$$\mathrm{SS}_e(B,AB)=\boldsymbol{y}'(\mathbf{I}_{14}-\mathbf{X}_{-A}(\mathbf{X}'_{-A}\mathbf{X}_{-A})^{-1}\mathbf{X}'_{-A})\boldsymbol{y}=1.4200,$$

$$\mathrm{SS}_e(A,AB)=\boldsymbol{y}'(\mathbf{I}_{14}-\mathbf{X}_{-B}(\mathbf{X}'_{-B}\mathbf{X}_{-B})^{-1}\mathbf{X}'_{-B})\boldsymbol{y}=5.4897,$$

其中$\mathbf{X}_{-A}=(\mathbf{1}_{14},\boldsymbol{x}_2,\boldsymbol{x}_3,\boldsymbol{x}_4,\boldsymbol{x}_5)$, $\mathbf{X}_{-B}=(\mathbf{1}_{14},\boldsymbol{x}_1,\boldsymbol{x}_4,\boldsymbol{x}_5)$. 它们的自由度分别为9和10. 计算$H_{01}$和$H_{02}$的检验统计量的值

$$F_1=\frac{\mathrm{SS}_e(B,AB)-\mathrm{SS}_e}{\mathrm{SS}_e/8}=0.7385<F_{1,8}(0.05)=5.32,$$

$$F_2=\frac{(\mathrm{SS}_e(A,AB)-\mathrm{SS}_e)/2}{\mathrm{SS}_e/8}=12.8914>F_{2,8}(0.05)=4.46.$$

Table: 生长激素数据的方差分析

方差源	自由度	平方和	均方	F值	P值
性别(A)	1	0.1200	0.1200	0.7385	0.4152
骨骼发育(B)	2	4.1897	2.0948	12.8914	0.0031
A和B交互	2	0.0754	0.0377	0.2321	0.7980
残差	8	1.3000	0.1625		

因此, 在显著性水平$\alpha = 0.05$下，可以认为儿童的性别对激素诱导生长速度的影响不显著，但儿童的骨骼发育对激素诱导生长速度影响的显著.

在方差分析模型的检验中, 一般首先检验交互效应是否显著, 如果不显著, 则在检验主因子效应的显著性时可直接基于无交互效应的方差分析模型, 从而可提高检验的精度.

- 例7.3.2中由F_3检验结果表明A和B的交互效应不显著后, A和B的显著性检验就可基于模型:

$$y_{ijk} = \mu + A_{ijk1}\alpha_1 + B_{ijk1}\beta_1 + B_{ijk2}\beta_2 + e_{ijk}. \tag{3.4}$$

计算在H_{01}、H_{02}下的相应约简模型分别为

$$y_{ijk} = \mu + B_{ijk1}\beta_1 + B_{ijk2}\beta_2 + e_{ijk}$$

$$y_{ijk} = \mu + A_{ijk1}\alpha_1 + e_{ijk}$$

LS残差平方和分别为

$$\mathrm{SS}_e^*(B) = 1.468, \quad \mathrm{SS}_e^*(A) = 5.771.$$

从而计算得

$$\mathrm{R}(\boldsymbol{\alpha}|\mu, \boldsymbol{\beta}) = \mathrm{SS}_e^*(B) - \mathrm{SS}_e(A, B) = 0.0926$$

$$\mathrm{R}(\boldsymbol{\beta}|\mu, \boldsymbol{\alpha}) = \mathrm{SS}_e^*(A) - \mathrm{SS}_e(A, B) = 4.3960$$

$$\mathrm{SS}_e^* = \mathrm{SS}_e^*(A, B) = 1.3754$$

它们的自由度分别为1, 2, 10. 检验统计量的值分别为

$$F_1^* = \frac{\mathrm{R}(\boldsymbol{\alpha}|\mu, \boldsymbol{\beta})}{\mathrm{SS}_e^*/10} = 0.673 < F_{1,10}(0.05) = 4.9646,$$

$$F_2^* = \frac{\mathrm{R}(\boldsymbol{\beta}|\mu, \boldsymbol{\alpha})/2}{\mathrm{SS}_e^*/10} = 15.980 > F_{2,10}(0.05) = 4.1028.$$

例7.3.2

- 结果仍显示儿童的性别(A) 对激素诱导生长速度的影响不显著，儿童的骨骼发育(B) 对激素诱导生长速度有显著影响.
- 不过剔除不显著的交互效应后，因子A显著性检验统计量的P值更大，而因子B的检验统计量的P值更小，即表明关于因子A和B的显著性检验的精度提高.

程序

```
install.packages("car"); library(car)
A<-factor( c(1,1,1, 1,1,1,1,2,2,2, 2,2,2,2) )
B<-factor(c(1,1,1, 2,2,3,3,1,2,2,2,3,3,3))
ylm<- lm(y~A+B)
Anova(ylm, type=2)  ♯ 适用于无交互效应
```

- 两向分类模型, 又称为交叉分类模型, 因子A和B的任意两个水平都可以相遇.

两级套分类模型

因子A的每个水平只与B的水平中某些水平相遇.

例如在化工试验中, 要比较甲, 乙两种催化剂, 同时还要选择每种催化剂所适应的温度. 往往不同的催化剂所要求的温度不同. 如果对催化剂甲选择的温度是200℃, 220℃ 和240℃, 而对催化剂乙选择的温度是150℃, 170℃ 和190℃

我们称催化剂是一级因素, 温度是二级因素.

两级套分类模型

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + e_{ijk}, \quad i = 1, \cdots, a, \quad j = 1, \cdots b_i, \quad k = 1, \cdots, n_{ij},$$

这里μ, α, e_{ijk}的意义和以前讨论的各种模型都相同. 并假定$e_{ijk} \sim N(0, \sigma^2)$, 所有$e_{ijk}$相互独立, 且称$\beta_{j(i)}$为水平$B_{j(i)}$的效应.

- $\mu_{ij} = \mu + \alpha_i + \beta_{j(i)}$ 可估.
- 参数函数$\beta_{j(i)} - \beta_{j'(i)}$, 对一切 $i, j \neq j'$ 都可估.
- $\alpha_i - \alpha_{i'}, i \neq i'$ 不可估计.
- 平衡数据下$(b_i = b, n_{ij} = c)$, $(\alpha_i + \overline{\beta}_{\cdot(i)}) - (\alpha_{i'} + \overline{\beta}_{\cdot(i')})$ 可估

- 首先考虑二级因子诸水平效应是否相等的假设

$$H_{01}: \quad \beta_{1(i)} = \cdots = \beta_{b_i(i)}, \quad i = 1, \cdots, a.$$

当假设H_{01}成立时, $\beta_{j(i)}$只与i 有关与j无关, 故记作β_i. 于是约简模型为

$$y_{ijk} = \mu + \alpha_i + \beta_i + e_{ijk} \triangleq \mu^0 + \alpha_i^0 + e_{ijk},$$

这里$\mu^0 = \mu,\ \alpha_i^0 = \alpha_i + \beta_i$. 这个单向分类模型的残差平方和

$$\mathrm{SS}_{He} = \sum_{i=1}^{a}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{i\cdot\cdot})^2.$$

全模型的残差平方和

$$\mathrm{SS}_e = \sum_{i=1}^{a}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij\cdot})^2.$$

其自由度等于$n_{\cdot\cdot} - m$, $m = \sum_i b_i$.

检验H_{01}的F统计量为

$$F_1 = \frac{\mathrm{SS}_{He} - \mathrm{SS}_e/(m-a)}{\mathrm{SS}_e/(n_{\cdot\cdot} - m)}.$$

当H_1为真时, $F_1 \sim F_{m-a,\, n_{\cdot\cdot}-m}$.

- 一级因子A诸水平效应相等性检验:

$$H_{02}: \ \alpha_1 + \overline{\beta}_{\cdot(1)} = \cdots = \alpha_a + \overline{\beta}_{\cdot(a)}.$$

因$\alpha_1 = \cdots = \alpha_a$是不可检验假设, 故在因子$B$诸水平平均的意义下考虑

考虑平衡数据情形($b_i = b$, $n_{ij} = c$).

若H_{02}成立, $\alpha_i + \overline{\beta}_{\cdot(i)}$与$i$无关. 可模型可被改写为

$$y_{ijk} = \mu^* + \beta^*_{j(i)} + e_{ijk},$$

其中$\mu^* = \mu + \alpha_i + \overline{\beta}_{\cdot(i)}$, $\beta^*_{j(i)} = \beta_{j(i)} - \overline{\beta}_{\cdot(i)}$, 满足$\sum_{j=1}^{b} \beta^*_{j(i)} = 0$.

应用Lagrange 乘子法, 可以求出μ^*和$\beta^*_{j(i)}$的约束LS解,

$$\widehat{\mu}^* = \bar{y}_{\dots}, \quad \widehat{\beta}^*_{j(i)} = \bar{y}_{ij\,\cdot} - \bar{y}_{i\cdot\cdot}.$$

因子A各水平下所有观测值的平均对总平均的变差平方和

$$\mathrm{SS}_{H_{02}} = \mathrm{RSS}(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}) - \mathrm{RSS}(\mu^*, \boldsymbol{\beta}^*) = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\dots})^2,$$

其自由度为$a-1$, $\mathrm{RSS}(\mu^*, \boldsymbol{\beta}^*)$为约简模型的回归平方和.

检验假设H_{02}的F统计量

$$F_2 = \frac{\mathrm{SS}_{H_{02}}/(a-1)}{\mathrm{SS}_e/ab(c-1)},$$

当H_{02}为真时, $F_2 \sim F_{a-1,\,ab(c-1)}$.

例7.4.1

比较甲,乙,丙,丁四种催化剂,每种催化剂要求的温度范围不完全相同.
对每种催化剂, 温度都取了三个水平(℃):

甲(A_1): 50, 55, 60, 乙(A_2): 70, 80, 90,
丙(A_3): 55, 65, 75, 丁(A_4): 90, 95, 100,

Table: 催化剂数据表

		催化剂			
		A_1	A_2	A_3	A_4
温度	$B_{1(i)}$	85, 89	82, 84	65, 61	67, 71
	$B_{2(i)}$	72, 70	91, 88	59, 62	75, 78
	$B_{3(i)}$	70, 67	85, 83	60, 56	85, 89

解 平衡数据, 可采用函数AOV()完成计算, 程序如下：

```
y<-c(85,89,82,84,65,61,67,71,72,70,91,88,59,62,75,78,70,67,
85,83,60,56,85,89)
B<-factor(gl(3,8,24))；A<-factor(gl(4,2,24))
y.aov<-aov(y~A/B)    或aov(y~A+A:B) 套结构
summary(y.aov)
```

Table: 方差分析表

方差源	自由度	平方和	均　方	F 值	P值
催化剂A	3	1960.5	653.5	$F_1 = 122.5$	2.89e-09
温度(催化剂) $B(A)$	8	804.0	100.5	$F_2 = 18.84$	1.11e-05
误　差	12	64.0	5.3		
总　和	23	2828.5			

- 从表中可以看出,

$$P(F_{3,12} > 122.5) = 2.89e-09, \quad P(F_{8,12} > 18.84) = 1.11e-05,$$

所以在显著性水平$\alpha = 0.01$, 我们拒绝两个原假设, 即认为

  - 对这四种催化剂, 温度不同水平的差异是显著的;
  - 就三种温度平均说来, 四种催化剂的差异也都是显著的.

- 进一步可通过分别作两组可估函数

$$\beta_{j(i)} - \beta_{j'(i)} \ \ j \neq j', \quad (\alpha_i + \beta_{\cdot(i)}) - (\alpha_i' + \beta_{\cdot(i')}), i = 1, 2, 3, 4,$$

的同时置信区间, 搞清楚对固定的催化剂是哪些温度之间有差异, 以及在四种催化剂中, 哪些催化剂之间有显著的差异. (自练)

误差正态性及方差齐性检验

- 前面方差分析要求观测误差向量e 满足
  - 诸分量相互独立;
  - 正态性;
  - 方差齐性(即每个观测值的方差相等).

  其中相互独立性一般容易满足

  只要在试验过程中随机化得到很好的实现.

主要讨论

- 误差的正态性检验
- 误差的方差齐性检验

- 正态性检验

考虑单向分类模型

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, \cdots, a, \quad j = 1, \cdots, n_i,$$

这里诸e_{ij}相互独立, $\mathrm{Var}(e_{ij}) = \sigma^2$.

  - 当样本量n_i都较大时, 可采用第6章介绍的基于残差的正态检验的方法.
  - 当样本量较小时, 这些方法不再适用.

下面主要介绍一种针对小样本的正态检验方法.

记第i水平下第j次观测的残差为

$$\hat{e}_{ij} = y_{ij} - \bar{y}_{i\cdot}, \quad i = 1, \cdots, a, \quad j = 1, \cdots, n_i.$$

- 若e_{ij}服从正态分布$N(0, \sigma^2)$, 则

$$E(\hat{e}_{ij}) = 0, \quad \mathrm{Var}(\hat{e}_{ij}) = \frac{n_i - 1}{n_i}\sigma^2,$$

$$\mathrm{Cov}(\hat{e}_{ij}, \hat{e}_{i'j'}) = \begin{cases} 0, & i \neq i', \\ -\sigma^2/n_i, & i = i',\ j \neq j. \end{cases}$$

在同一水平下残差方差相同但不独立; 在不同水平下残差方差不等, 但相互独立.

- 若作如下线性变换

$$z_{il} = \sqrt{\frac{l}{l+1}}\left(\frac{1}{l}\sum_{j=1}^{l}\hat{e}_{ij} - \hat{e}_{i,l+1}\right) = \sqrt{\frac{l}{l+1}}\left(\frac{1}{l}\sum_{j=1}^{l}y_{ij} - y_{i,l+1}\right),$$

$$l = 1, 2, \cdots, n_i - 1, \quad i = 1, 2, \cdots, a.$$

将$N = \sum_i n_i$ 个残差变为$N - a$ 个z_{il}, 且有

$$E(z_{il}) = 0, \quad \mathrm{Var}(z_{il}) = \sigma^2, \quad \mathrm{Cov}(z_{il}, z_{i'l'}) = 0.$$

可通过$\{z_{ij}\}$ 检验误差分布的正态性.

注 方差分析法对总体分布偏离正态分布有较好的稳健性.

注 但当总体分布偏离正态分布较大时, 建议使用非参数法.

- 对单向分类模型, 若误差方差不相等, 则模型可表示为

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, \cdots, a, \quad j = 1, \cdots, n_i.$$

这里诸e_{ij}相互独立, 且$e_{ij} \sim N(0, \sigma_i^2)$. 检验假设

$$H_0: \quad \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_a^2 \Longleftrightarrow H_1: \quad \sigma_1^2, \sigma_2^2, \cdots, \sigma_a^2 \text{ 不全相等}.$$

与线性模型下的异方差检验不同

- 线性模型下的异方差检验: 不同次观测对应的误差方差不等
- 方差分析模型的方差齐性检验: 是在假设同一水平的n_i 次重复观测所对应的误差方差相等下, 检验不同水平间的误差方差σ_i^2是否相等.

因此, 方差齐性检验有它特有的检验方法.

- 记第i个水平的误差平方和和均方

$$\mathrm{SS}_{e_i} = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2, \quad \mathrm{MS}_{e_i} = \mathrm{SS}_{e_i}/(n_i - 1).$$

在正态性假设下, $\mathrm{SS}_{e_i} \sim \sigma_i^2 \chi^2_{n_i - 1}$.

- 记$N = \sum_{i=1}^{a} n_i$为总样本量.
- 记第i个水平下第j次观测的残差为

$$\hat{e}_{ij} = y_{ij} - \bar{y}_{i\cdot}, \quad i = 1, \cdots, a, \;\; j = 1, \cdots, n_i,$$

这里$\bar{y}_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}/n_i$.

1. Levene 检验

   由Levene (1960)提出一种适用性较广的方差齐性检验方法. 令

$$d_{ij} = |\hat{e}_{ij}|, \quad i = 1, \cdots, a, \quad j = 1, \cdots, n_i.$$

注意到

$$E(d_{ij}) = \sigma_i \sqrt{\frac{2(n_i - 1)}{n_i \pi}}, \quad i = 1, \cdots, a, \quad j = 1, \cdots, n_i,$$

平衡数据即$n_1 = \cdots = n_a = n$ 时, $E(d_{ij}) = \sigma_i \sqrt{2(n-1)/(n\pi)}$.

当平衡数据情形($n_i = n$) 或每个n_i都较大, 可把d_{ij}看作观察值, 当作单因素试验数据处理.

计算组内和组间平方和

$$\mathrm{SS}_W^2=\sum_{i=1}^{a}\sum_{j=1}^{n_i}(d_{ij}-\overline{d}_{i\cdot})^2,\quad \mathrm{SS}_B^2=\sum_{i=1}^{a}n_i(\overline{d}_{i\cdot}-\overline{d}_{\cdot\cdot})^2$$

构造F检验统计量:

$$L=\frac{\mathrm{SS}_B^2/(a-1)}{\mathrm{SS}_W^2/(N-a)},\tag{3.5}$$

这里, $\overline{d}_{i\cdot}=\sum_{j=1}^{n_i}d_{ij}/n_i$, $\overline{d}_{\cdot\cdot}=\sum_{i=1}^{a}\sum_{j=1}^{n_i}d_{ij}/N$.

在原假设H_0成立的条件下, L 近似服从$F_{a-1,N-a}$.

若$L>F_{a-1,N-a}(\alpha)$, 就拒绝H_0, 认为各水平间的方差不全相等.

注1 Levene 检验对总体分布偏离正态分布有较好的稳健性. 但要求n_i不能太小.

注2 文献中提出了一系列改进的Levene型检验.

- Brown-Forsythe检验（更稳健） 即用

$$z_{ij} = |y_{ij} - m_i|$$

替换d_{ij} 所得的Levene检验, 其中m_i 为第i组数据的中位数.

- O'Brien检验（针对小、中等样本量的不平衡样本） 即用

$$u_{ij} = d_{ij}/\sqrt{(n_i - 1)/n_i} \qquad (\text{在}H_0\text{下}u_{ij}\text{方差相等})$$

替换d_{ij} 所得的Levene 检验.

2. Hartley 检验

由Hartley(1950)提出的, 也称最大F比检验. 统计量为

$$F_{\max} = \frac{\max_i(\mathrm{MS}_{e_i})}{\min_i(\mathrm{MS}_{e_i})}.$$

当计算的值超过临界值(Pearson 和Hartley, 1966)时, 将拒绝H_0.

注1 若$a = 2$时, 则在H_0下, 基于$F_{\max}$的检验是精确F检验.

注2 严格上, Hartley检验只适用于样本量相等的情形, 对于样本量不相等情形, 若它们间差别不大时, 可近似用

注3 Hartley检验对于正态性的偏离十分敏感. 当总体服从正态分布时, 用Hartley检验比Levene检验具有更高功效, 但严重偏离正态时, 犯第一类错误的概率会膨胀.

3. Bartlett检验

检验统计量为

$$B = \frac{1}{c}\left((N-a)\ln(\mathrm{MS}_e) - \sum_{i=1}^{a}(n_i-1)\ln(\mathrm{MS}_{e_i})\right),$$

其中

$$c = 1 + \frac{1}{3(a-1)}\left\{\left(\sum_{i=1}^{a}\frac{1}{n_i-1}\right) - \frac{1}{N-a}\right\}.$$

在误差正态假定下，若H_0成立，则$B \sim \chi^2_{a-1}$.

当$B > \chi^2_{a-1}(\alpha)$，将拒绝H_0.

- Bartlett检验的优点是不要求各水平重复观测次数n_i相等或相近，但它的缺点是对误差非正态性是很敏感.

4. 最大方差检验法(Cochran法)

因为Hartley法和Bartlett法对较小的SS_{e_i}值很敏感, 所以当SS_{e_i}中存在一个值为0或者很小时, Hartley法和Bartlett法均不能使用. 但是当重复数n_i较小时, 这种情况是经常会出现的.

令统计量为

$$C = \frac{\max_i(MS_{e_i})}{\sum_{i=1}^{a} MS_{e_i}}. \tag{3.6}$$

统计量C的临界值表可见项可风和吴启光(1989) (p.579∼580, 表4). 当计算的C大于表中相对应的临界值, 就拒绝方差相等的零假设, 认为方差不全相等.

- Cochran的方差齐性检验法只适用于平衡数据情形.
- 在实际中，当n较小时, Cochran法应用较为广泛.

针对异方差的处理办法:

- 加权LS方法. 第i个水平的第j次观测的权为

$$w_{ij} = \frac{1}{\mathrm{MS}_{e_i}},$$

  即第i个水平的样本方差的倒数.

- 数据变化法. 可采用第6 章介绍的方差稳定化变换、Box-Cox变换等. 变换后的数据可能仍不是正态的. 对于统计推断的结果来说, 方差齐性的要求远比正态性假设更为重要.

- 非参数法. 检验a个不同处理/水平对应的子总体均值或中位数是否相等的问题, 可以采用非参数秩F检验, 其检验统计量为

- $$F_R = \frac{(N-a)\sum_{i=1}^{a} n_i(\bar{R}_{i\cdot} - \bar{R}_{\cdot\cdot})^2}{(a-1)\sum_{i=1}^{a}\sum_{j=1}^{n_i}(R_{ij} - \bar{R}_{i\cdot})^2},$$

其中$N = \sum_{i=1}^{a} n_i$, R_{ij}是y_{ij}在集合$\{y_{ij}; i = 1, \cdots, a, j = 1, \cdots, n_i\}$的元素从小到大排列后所处的位置(即$y_{ij}$的秩),

$$\bar{R}_{i\cdot} = \sum_{j=1}^{n_i} R_{ij}/n_i, \quad \bar{R}_{\cdot\cdot} = \sum_{i=1}^{a}\sum_{j=1}^{n_i} R_{ij}/N.$$

若样本量n_i不是很小, 且

$$F_R > F_{a-1,N-a}(\alpha),$$

拒绝原假设，认为a个子总体的均值不全相等，否则接收子总体均值全相等的假设.

例7.5.1 (饲料对比试验)

为发展我国机械化养鸡, 某研究所根据我国的资源情况, 研究了三种饲料配方： 第一种, 以鱼粉为主的鸡饲料; 第二种, 以槐树粉, 苜蓿粉为主, 加少量鱼粉; 第三种, 以槐树粉、苜蓿粉为主, 加少量化学药品. 后两种是他们研制的新配方. 为比较三种饲料在养鸡增肥上的效果, 各喂养10只母雏鸡, 于60天后观察它们的重量.

饲料	鸡重(克)									
第一种	1073	1058	1071	1037	1066	1026	1053	1049	1065	1051
第二种	1016	1058	1038	1042	1020	1045	1044	1061	1034	1049
第三种	1084	1069	1106	1078	1075	1090	1079	1094	1111	1092

案例分析

- 在这项试验中, 60天的鸡重是指标; 因素是饲料.
- 在试验方案中共取了三个水平, 试验的目的是要比较三种饲料在养鸡增肥的效果上有何差别.
- 需要作均值间的比较检验, 因为这是一个单因素方差分析问题, 可用7.1 节介绍的方法进行各均值相等性检验, 然后作所有两两均值差的同时置信区间以进一步得到每两种喂养效应间有无显著差异.
- 在作均值的相关推断前, 需要先对模型误差作正态检验和方差齐性检验.

案例分析

设模型

$$y_{ij} = \mu_i + e_{ij}, \quad i = 1, 2, 3, \quad j = 1, 2, \cdots, 10,$$

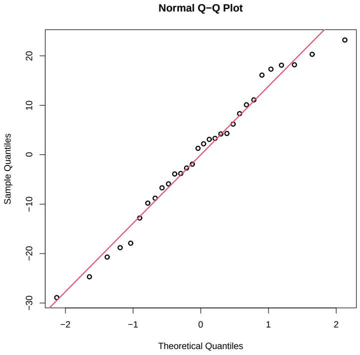
其中y_{ij}为采用第i种饲料喂养的第j只鸡60天的重量.

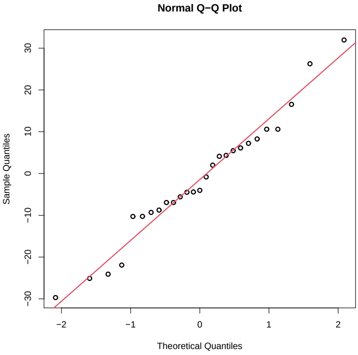
- 首先对随机误差e_{ij}作正态性检验和方差齐性检验.

```
y1<-c(1073,1058,1071,1037,1066,1026,1053,1049,1065,1051)
y2<-c(1016,1058,1038,1042,1020,1045,1044,1061,1034,1049)
y3<-c(1084,1069,1106,1078,1075,1090,1079,1094,1111,1092)
y <- c(y1,y2,y3);
A<-factor(gl(3,10,30))
lm.reg<-lm(y~A)
y.res = residuals(lm.reg)    ♯ 残差ê_ij
```

```
n1<-length(y1); n2<-length(y2); n3<-length(y3)
z1<-c(rep(0,n1-1)); z2<-c(rep(0,n2-1)); z3<-c(rep(0,n3-1))
s1<-0; s2<-0; s3<-0
for (l in 1:(n1-1))
s1<-s1+y1[l]; z1[l]=(s1/l-y1[l+1])*sqrt(l/(l+1))
for (l in 1:(n2-1))
s2<-s2+y2[l]; z2[l]=(s2/l-y2[l+1])*sqrt(l/(l+1))
for (l in 1:(n3-1))
s3<-s3+y3[l]; z3[l]=(s3/l-y3[l+1])*sqrt(l/(l+1))
z<-c(z1,z2,z3); z  ♯ 基于残差构造独立等方差序列
plot(lm.reg,2);  ♯ 残差的Q-Q图
shapiro.test(y.res)  ♯ p-value = 0.6474
shapiro.test(z)  ♯ p-value = 0.6486
qqnorm(z, main = "Normal Q-Q Plot", xlab = "Theoretical Quantiles", ylab =
"Sample Quantiles", plot.it = TRUE, datax = FALSE, lwd=2) abline(mean(z),
sd(z), col=2, lwd=2)  ♯ 数据Z的Q-Q图
```

- 正态性检验





$\hat{e}_{ij}$的正态QQ图　　z_{ij}的正态QQ图

QQ图和Shapiro-Wilk检验结果: 可接受这批数据服从正态性.

- 方差齐性检验

  因数据服从正态分布且平衡, 以上方差齐性检验方法都可用.

  - Levene检验法

    leveneTest(y.res, A, center = mean) ♯ p-value=0.9666

  - Brown-Forsythe检验

    leveneTest(y.res, A, center = median) ♯ p-value= 0.959

  - Harley最大F比法

$$F_{\max} = \max_i(\mathrm{MS}_{e_i})/\min_i(\mathrm{MS}_{e_i}) = \frac{226}{182} = 1.24 < F_{\max}(0.05, 3, 9) = 5.34.$$

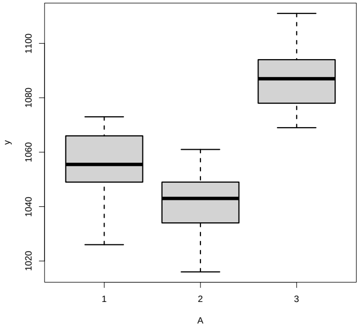
  - Bartlett的χ^2检验法

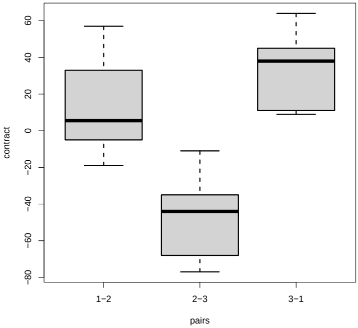
    bartlett.test(y, A) ♯ p-value = 0.9492

  - Cochran检验法.

$$C = \frac{\max(\mathrm{MS}_{e_i})}{\mathrm{MS}_{e_1} + \mathrm{MS}_{e_2} + \mathrm{MS}_{e_3}} = \frac{226}{619} \approx 0.365 < C_{9,3}(0.05) = 0.6167.$$

- 检验假设$H_{01}: \mu_1 = \mu_2 = \mu_3$ 和 $H_{02}: \mu_i = \mu_j \ (i \neq j)$.
  - 先通过Box图直观了解三组数据$\{y_{il}, \cdots, y_{in}\}$, $i = 1, 2, 3$, 以及两两组差数据$\{y_{1j} - y_{2j}, \cdots, y_{1n} - y_{2n}\}$, $\{y_{2j} - y_{3j}, \cdots, y_{2n} - y_{3n}\}$, $\{y_{3j} - y_{1j}, \cdots, y_{3n} - y_{1n}\}$的各自均值情况.





- 采用方差分析进一步检验假设H_{01}：三组均值都相等

  相应程序和运行结果如下:

```
A<-factor(gl(3,10,30))
y.aov <- aov(y~ A)
summary(y.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	2	11675	5837	28.3	2.36e-07 ***
Residuals	27	5569	206		

结果表明: 拒绝均值全相等的假设, 认为三种饲料下60天的鸡重的均值存在显著差异.

- 两两均值差的置信区间

计算得三组均值的估计为

$$\widehat{\mu}_1 = \bar{y}_{1\cdot} = 1054.9, \quad \widehat{\mu}_2 = \bar{y}_{2\cdot} = 1040.7, \quad \widehat{\mu}_3 = \bar{y}_{3\cdot} = 1087.8,$$

误差方差σ^2的估计为$\widehat{\sigma}^2 = \mathrm{MS}_e = 206$. 由于$e_{ij} \sim N(0, \sigma^2)$, 立得

$$\frac{(\widehat{\mu}_i - \widehat{\mu}_j) - (\mu_i - \mu_i)}{\widehat{\sigma}/\sqrt{5}} \sim t_{27}, \quad i \neq j.$$

$\alpha = 0.05$, 查表得$t_{27}(\alpha/2) = t_{27}(0.025) = 2.052$,

$t_{27}(\alpha/6) \approx 2.552$, $F_{2,27}(0.05) = 3.354$, $q_{2,27}(0.05) = 3.506$.

案例分析

- $\mu_1-\mu_2$, $\mu_2-\mu_3$ 和$\mu_3-\mu_1$ 各自的置信区间：

  [1.022, 27.378], [-60.278, -33.922], [19.722, 46.078]

  三个对照的置信区间都没有包括0, 表明: 在显著性水平$\alpha=0.05$下, 均值两两之间都存在显著差异.

- 两两均值对照的置信系数为95%的同时置信区间

均值对照	Bonferroni区间	Scheffé区间	Tukey区间
$\mu_1-\mu_2$	[-2.193, 30.593]	[-2.435, 30.835]	[-1.724 30.124]
$\mu_2-\mu_3$	[-63.493, -30.707]	[-63.735, -30.465]	[-63.024 -31.176]
$\mu_3-\mu_1$	[16.507, 49.293]	[16.265, 49.535]	[16.976 48.824]

Tukey区间的长度最短, Scheffé区间的长度最长.