第七章　方差分析模型

Tianxiao Pang

Zhejiang University

December 28, 2023

内容

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, 因而也被称为试验设计模型. 方差分析由英国著名的统计学家Fisher于20世纪20年代提出, 有时也称为F检验.

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, 因而也被称为试验设计模型. 方差分析由英国著名的统计学家Fisher于20世纪20年代提出, 有时也称为F检验.

方差分析的应用范围很广, 可用于社会科学、生物工程、医学研究等各个领域的试验数据分析.

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, 因而也被称为试验设计模型. 方差分析由英国著名的统计学家Fisher于20世纪20年代提出, 有时也称为F检验.

方差分析的应用范围很广, 可用于社会科学、生物工程、医学研究等各个领域的试验数据分析.

本章仅仅介绍最基本的单因素方差分析和两因素方差分析.

方差分析起源于农业田间试验. 假定某个农业试验基地引进了a个小麦品种, 在进行大面积种植之前, 先进行小范围的试验种植, 以便从中挑选出最适合本地区的品种.

单因素方差分析

方差分析起源于农业田间试验. 假定某个农业试验基地引进了a个小麦品种, 在进行大面积种植之前, 先进行小范围的试验种植, 以便从中挑选出最适合本地区的品种.

将一大块田划分成面积相等的几个小块, 其中n_1块种植第1种小麦, n_2块种植第2种小麦, $\cdots$, 等等. 试验的目的是比较小麦的品种, 因此我们感兴趣的只是小麦品种这一个因素. 其它所有因素, 如施肥量、浇水等对这n块田都控制在相同状态下.

单因素方差分析

方差分析起源于农业田间试验. 假定某个农业试验基地引进了a个小麦品种, 在进行大面积种植之前, 先进行小范围的试验种植, 以便从中挑选出最适合本地区的品种.

将一大块田划分成面积相等的几个小块, 其中n_1块种植第1种小麦, n_2块种植第2种小麦, $\cdots$, 等等. 试验的目的是比较小麦的品种, 因此我们感兴趣的只是小麦品种这一个因素. 其它所有因素, 如施肥量、浇水等对这n块田都控制在相同状态下.

在这个例子里, 我们感兴趣的因素只有一个, 即小麦品种. 每个具体的品种, 都称为小麦品种这个因素的一个水平. 现有a个不同的品种, 因此小麦品种这一因素一共有a个水平. 这是单因素a个水平的问题.

记y_{ij}为第i个品种的小麦在第j块田的产量, $i=1,\cdots,a$, $j=1,\cdots,n_i$. 对固定的i, $\{y_{i1},\cdots,y_{in_i}\}$是种植第$i$种小麦的$n_i$块田的产量. 因为除了一些随机误差外, 这$n_i$块田的一切生产条件完全一样. 因此可把它们看做来自正态总体的样本. 即, 可假设$\{y_{ij}\}$相互独立且

$$y_{ij} \sim N(\mu_i, \sigma^2), \ \ i=1,\cdots,a; \ \ j=1,\cdots,n_i. \tag{7.1.1}$$

记y_{ij}为第i个品种的小麦在第j块田的产量, $i=1,\cdots,a$, $j=1,\cdots,n_i$. 对固定的i, $\{y_{i1},\cdots,y_{in_i}\}$是种植第$i$种小麦的$n_i$块田的产量. 因为除了一些随机误差外, 这$n_i$块田的一切生产条件完全一样. 因此可把它们看做来自正态总体的样本. 即, 可假设$\{y_{ij}\}$相互独立且

$$y_{ij}\sim N(\mu_i,\sigma^2),\ \ i=1,\cdots,a;\ \ j=1,\cdots,n_i. \tag{7.1.1}$$

表7.1.1 单因素方差分析问题

水平	总体	样本
1	$N(\mu_1,\sigma^2)$	$y_{11},y_{12},\cdots,y_{1n_1}$
2	$N(\mu_2,\sigma^2)$	$y_{21},y_{22},\cdots,y_{2n_2}$
$\vdots$	$\vdots$	$\vdots$
a	$N(\mu_a,\sigma^2)$	$y_{a1},y_{a2},\cdots,y_{an_a}$

称所考虑的因素为因素A, 假定它有a个水平, 我们的目的是比较这a个水平的差异. 如何把问题模型化?

称所考虑的因素为因素A, 假定它有a个水平, 我们的目的是比较这a个水平的差异. 如何把问题模型化?

将(7.1.1)改写为

$$\begin{cases} y_{ij} = \mu_i + e_{ij}, \\ e_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2), \end{cases} \quad i = 1, \cdots, a; \;\; j = 1, \cdots, n_i, \tag{7.1.2}$$

其中e_{ij}是试验误差.

称所考虑的因素为因素A, 假定它有a个水平, 我们的目的是比较这a个水平的差异. 如何把问题模型化?

将(7.1.1)改写为

$$\begin{cases} y_{ij} = \mu_i + e_{ij}, \\ e_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2), \end{cases} \quad i = 1, \cdots, a; \ \ j = 1, \cdots, n_i, \tag{7.1.2}$$

其中e_{ij}是试验误差.

比较因素A的a个水平的差异归结为比较这a个总体的均值$\mu_1, \cdots, \mu_a$之间的差异. (如何比较? 假设检验)

进一步模型化:

进一步模型化:

记

$$n = \sum_{i=1}^{a} n_i,\ \mu = \frac{1}{n}\sum_{i=1}^{a} n_i\mu_i,\ \alpha_i = \mu_i - \mu,$$

μ为整个样本的均值的总平均, α_i表示第i个水平下的均值与总平均的差异, 它反映了第i个水平对指标y的效应. 因此有

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

进一步模型化:

记

$$n = \sum_{i=1}^{a} n_i,\ \mu = \frac{1}{n}\sum_{i=1}^{a} n_i\mu_i,\ \alpha_i = \mu_i - \mu,$$

μ为整个样本的均值的总平均, α_i表示第i个水平下的均值与总平均的差异, 它反映了第i个水平对指标y的效应. 因此有

$$\sum_{i=1}^{a} n_i\alpha_i = \sum_{i=1}^{a} n_i(\mu_i - \mu) = n\mu - n\mu = 0.$$

因为$\mu_i = \mu + \alpha_i$, 于是(7.1.2)又可以写为

$$\begin{cases} y_{ij} = \mu + \alpha_i + e_{ij}, \\ e_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2), \\ \sum_{i=1}^{a} n_i \alpha_i = 0, \end{cases} \quad i = 1, \cdots, a; \ \ j = 1, \cdots, n_i. \tag{7.1.3}$$

这就是单因素方差分析模型.

因为$\mu_i = \mu + \alpha_i$, 于是(7.1.2)又可以写为

$$\begin{cases} y_{ij} = \mu + \alpha_i + e_{ij}, \\ e_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2), \qquad i = 1, \cdots, a; \ \ j = 1, \cdots, n_i. \\ \sum_{i=1}^{a} n_i \alpha_i = 0, \end{cases} \tag{7.1.3}$$

这就是单因素方差分析模型. 写成矩阵形式即为

$$\begin{cases} \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \ \ \boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n), \\ \boldsymbol{h}'\boldsymbol{\beta} = 0, \end{cases} \tag{7.1.4}$$

其中

$$\begin{aligned} \boldsymbol{Y} &= (y_{11}, \cdots, y_{1n_1}, y_{21}, \cdots, y_{2n_2}, \cdots, y_{a1}, \cdots, y_{an_a})', \\ \boldsymbol{\beta} &= (\mu, \alpha_1, \cdots, \alpha_a)', \\ \boldsymbol{e} &= (e_{11}, \cdots, e_{1n_1}, e_{21}, \cdots, e_{2n_2}, \cdots, e_{a1}, \cdots, e_{an_a})', \\ \boldsymbol{h} &= (0, n_1, n_2, \cdots, n_a)', \end{aligned}$$

以及

$$
\boldsymbol{X} = \begin{array}{l}
\text{第}1\text{行} \\
\vdots \\
\text{第}n_1\text{行} \\
\text{第}n_1+1\text{行} \\
\vdots \\
\text{第}n_1+n_2\text{行} \\
\vdots \\
\text{第}n_1+\cdots+n_{a-1}+1\text{行} \\
\vdots \\
\text{第}n_1+\cdots+n_{a-1}+n_a\text{行}
\end{array}
\begin{pmatrix}
1 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & 0 & \cdots & 0 \\
1 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 0 & \cdots & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 0 & \cdots & 1
\end{pmatrix}.
$$

$$
X = \begin{array}{l}
\text{第}1\text{行} \\
\vdots \\
\text{第}n_1\text{行} \\
\text{第}n_1+1\text{行} \\
\vdots \\
\text{第}n_1+n_2\text{行} \\
\vdots \\
\text{第}n_1+\cdots+n_{a-1}+1\text{行} \\
\vdots \\
\text{第}n_1+\cdots+n_{a-1}+n_a\text{行}
\end{array}
\begin{pmatrix}
1 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & 0 & \cdots & 0 \\
1 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 0 & \cdots & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 0 & \cdots & 1
\end{pmatrix}.
$$

可见, 单因素方差分析模型是一个带约束条件$\boldsymbol{h}'\boldsymbol{\beta}=0$的线性模型.

易见, 检验

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

等价于检验

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

易见, 检验

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

等价于检验

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

若H_0被拒绝, 则接受因素A的各水平的效应之间有显著差异的假设.

下面来推导H_0的检验统计量.

下面来推导H_0的检验统计量. 记

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}.$$

考虑统计量

$$\mathsf{SS}_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2,$$

称SS_T为总离差平方和, 简称为总平方和. 它反映了全部试验数据之间的差异.

对SS_T进行分解:

对SS_T进行分解:

$$
\begin{aligned}
\mathsf{SS}_T &= \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot}+\bar{y}_{i\cdot}-\bar{y})^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2+\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{y}_{i\cdot}-\bar{y})^2 \\
&\quad + 2\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})(\bar{y}_{i\cdot}-\bar{y}),
\end{aligned}
$$

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

对SS_T进行分解:

$$\begin{aligned}\mathsf{SS}_T &= \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot}+\bar{y}_{i\cdot}-\bar{y})^2\\ &= \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2+\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{y}_{i\cdot}-\bar{y})^2\\ &\quad +2\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})(\bar{y}_{i\cdot}-\bar{y}),\end{aligned}$$

其中$\bar{y}_{i\cdot}=\frac{1}{n_i}\sum_{j=1}^{n_i}y_{ij}$. 注意到$\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})=0$, 所以

$$\begin{aligned}\mathsf{SS}_T &= \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2+\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{y}_{i\cdot}-\bar{y})^2\\ &\triangleq \mathsf{SS}_E+\mathsf{SS}_A.\end{aligned}\tag{7.1.5}$$

$\mathsf{SS}_E=\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$反映了随机误差的影响. 因为对固定的$i$, $\{y_{i1},\cdots,y_{in_i}\}$来自同一个正态总体$N(\mu_i,\sigma^2)$. 因此, 它们之间的差异完全是由随机误差所致, 而$\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$正是它们误差大小的度量. 把$a$组这样的离差平方和求和就得到了$\mathsf{SS}_E$. 通常称$\mathsf{SS}_E$为误差平方和或组内平方和.

$\mathsf{SS}_E = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$反映了随机误差的影响. 因为对固定的$i$, $\{y_{i1},\cdots,y_{in_i}\}$来自同一个正态总体$N(\mu_i,\sigma^2)$. 因此, 它们之间的差异完全是由随机误差所致, 而$\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$正是它们误差大小的度量. 把$a$组这样的离差平方和求和就得到了$\mathsf{SS}_E$. 通常称$\mathsf{SS}_E$为误差平方和或组内平方和.

$\mathsf{SS}_A = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{y}_{i\cdot}-\bar{y})^2 = \sum_{i=1}^{a}n_i(\bar{y}_{i\cdot}-\bar{y})^2$. $\bar{y}_{i\cdot}$是第i个总体的样本均值, 它是第i个总体均值μ_i的估计. $\bar{y}$是$\mu=\frac{1}{n}\sum_{i=1}^{a}n_i\mu_i$的估计. 因此, SS_A是a个总体均值$\mu_1,\cdots,\mu_a$之间的差异大小的一个度量. 称SS_A为效应平方和或组间平方和.

$\mathrm{SS}_E = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$反映了随机误差的影响. 因为对固定的$i$, $\{y_{i1},\cdots,y_{in_i}\}$来自同一个正态总体$N(\mu_i,\sigma^2)$. 因此, 它们之间的差异完全是由随机误差所致, 而$\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$正是它们误差大小的度量. 把$a$组这样的离差平方和求和就得到了$\mathrm{SS}_E$. 通常称$\mathrm{SS}_E$为误差平方和或组内平方和.

$\mathrm{SS}_A = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{y}_{i\cdot}-\bar{y})^2 = \sum_{i=1}^{a} n_i(\bar{y}_{i\cdot}-\bar{y})^2$. $\bar{y}_{i\cdot}$是第i个总体的样本均值, 它是第i个总体均值μ_i的估计. $\bar{y}$是$\mu=\frac{1}{n}\sum_{i=1}^{a} n_i\mu_i$的估计. 因此, SS_A是a个总体均值$\mu_1,\cdots,\mu_a$之间的差异大小的一个度量. 称SS_A为效应平方和或组间平方和.

(7.1.5)是平方和分解公式, 它将总平方和按其来源分解成两部分. 一部分是SS_E, 即误差平方和, 是由随机误差引起的. 另一部分是SS_A, 即因素A的平方和, 是由因素A的各水平的差异引起的.

SS_E和SS_A在平均意义下有多大?

SS_E和SS_A在平均意义下有多大?

由于对固定的i, $\{y_{ij}, j = 1, \cdots, n_i\}$为来自$N(\mu_i, \sigma^2)$的样本, 因此

$$\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2/\sigma^2 \sim \chi^2(n_i - 1).$$

所以有

$$\mathsf{E}(\mathsf{SS}_E) = \sum_{i=1}^{a}\mathsf{E}\Big[\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2\Big] =$$

SS_E和SS_A在平均意义下有多大?

由于对固定的i, $\{y_{ij}, j = 1, \cdots, n_i\}$为来自$N(\mu_i, \sigma^2)$的样本, 因此

$$\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2/\sigma^2 \sim \chi^2(n_i - 1).$$

所以有

$$\mathsf{E}(\mathsf{SS}_E) = \sum_{i=1}^{a}\mathsf{E}\Big[\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2\Big] = (n-a)\sigma^2,$$

这说明$\mathsf{SS}_E/(n-a)$是σ^2的一个无偏估计.

另一方面,

$$\mathsf{E}(\mathsf{SS}_A) \quad =$$

另一方面,

$$\begin{aligned}
\mathsf{E}(\mathsf{SS}_A) &= \mathsf{E}\Big[\sum_{i=1}^{a} n_i(\bar{y}_{i\cdot} - \bar{y} - \alpha_i + \alpha_i)^2\Big] \\
&= \sum_{i=1}^{a} n_i\Big[\mathsf{E}(\bar{y}_{i\cdot} - \bar{y} - \alpha_i)^2 + \alpha_i^2\Big] \\
&= \sum_{i=1}^{a} n_i\Big(\frac{\sigma^2}{n_i} - \frac{\sigma^2}{n}\Big) + \sum_{i=1}^{a} n_i\alpha_i^2 \\
&= (a-1)\sigma^2 + \sum_{i=1}^{a} n_i\alpha_i^2.
\end{aligned}$$

所以

$$\mathsf{E}\left[\mathsf{SS}_A/(a-1)\right] = \sigma^2 + \sum_{i=1}^{a} n_i\alpha_i^2/(a-1).$$

从这个结论可以看出, $\mathsf{SS}_A/(a-1)$反映了各水平效应的影响. 当H_0为真时, $\mathsf{SS}_A/(a-1)$是σ^2的无偏估计.

所以
$$\mathsf{E}\left[\mathsf{SS}_A/(a-1)\right]=\sigma^2+\sum_{i=1}^{a}n_i\alpha_i^2/(a-1).$$
从这个结论可以看出, $\mathsf{SS}_A/(a-1)$反映了各水平效应的影响. 当H_0为真时, $\mathsf{SS}_A/(a-1)$是σ^2的无偏估计.

因此, 若H_0为真, 那么
$$F=\frac{\mathsf{SS}_A/(a-1)}{\mathsf{SS}_E/(n-a)}$$
的取值在1附近; 若H_0不真, 则F有变大的趋势. 这启发我们可以通过F的大小来检验H_0. (需要进一步知道SS_E和SS_A的概率分布)

由样本$\{y_{ij}\}$的独立性, 可知

$$\frac{\mathsf{SS}_E}{\sigma^2} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2}{\sigma^2} \sim \chi^2\Big(\sum_{i=1}^{a}(n_i - 1)\Big) = \chi^2(n-a).$$

由样本$\{y_{ij}\}$的独立性, 可知

$$\frac{\mathsf{SS}_E}{\sigma^2} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2}{\sigma^2} \sim \chi^2\left(\sum_{i=1}^{a}(n_i - 1)\right) = \chi^2(n-a).$$

若H_0为真, 那么$\{y_{ij}\}$是独立同分布随机变量序列, 服从$N(\mu, \sigma^2)$. 所以

$$\frac{\mathsf{SS}_T}{\sigma^2} \sim \chi^2(n-1).$$

由样本$\{y_{ij}\}$的独立性, 可知

$$\frac{\mathsf{SS}_E}{\sigma^2} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2}{\sigma^2} \sim \chi^2(\sum_{i=1}^{a}(n_i-1)) = \chi^2(n-a).$$

若H_0为真, 那么$\{y_{ij}\}$是独立同分布随机变量序列, 服从$N(\mu,\sigma^2)$. 所以

$$\frac{\mathsf{SS}_T}{\sigma^2} \sim \chi^2(n-1).$$

为了推导检验统计量在H_0为真时的概率分布, 我们把$\mathsf{SS}_T, \mathsf{SS}_E, \mathsf{SS}_A$都写成正态随机向量的二次型的形式.

回忆

$$\boldsymbol{Y} = (y_{11}, \cdots, y_{1n_1}, \cdots, y_{a1}, \cdots, y_{an_a})'.$$

记$\mathbf{1}_n$为元素全为1的n维的列向量, $\mathbf{1}_{n_i}$为元素全为1的n_i维的列向量, $i = 1, \cdots, a$.

回忆

$$\boldsymbol{Y} = (y_{11}, \cdots, y_{1n_1}, \cdots, y_{a1}, \cdots, y_{an_a})'.$$

记$\mathbf{1}_n$为元素全为1的n维的列向量, $\mathbf{1}_{n_i}$为元素全为1的n_i维的列向量, $i = 1, \cdots, a$. 则可知, 若H_0为真,

$$\boldsymbol{Y} \sim N(\mu \mathbf{1}_n, \sigma^2 \boldsymbol{I}_n), \ \frac{\boldsymbol{Y}}{\sigma} \sim N(\frac{\mu}{\sigma}\mathbf{1}_n, \boldsymbol{I}_n),$$

$$\begin{cases} \mathsf{SS}_T = \boldsymbol{Y}'(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n')\boldsymbol{Y} =: \boldsymbol{Y}'\boldsymbol{C}\boldsymbol{Y}, \\ \mathsf{SS}_E = \boldsymbol{Y}'(\boldsymbol{I}_n - \mathrm{diag}(\frac{1}{n_1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}', \cdots, \frac{1}{n_a}\mathbf{1}_{n_a}\mathbf{1}_{n_a}'))\boldsymbol{Y} =: \boldsymbol{Y}'\boldsymbol{C}_1\boldsymbol{Y}, \\ \mathsf{SS}_A = \boldsymbol{Y}'(\mathrm{diag}(\frac{1}{n_1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}', \cdots, \frac{1}{n_a}\mathbf{1}_{n_a}\mathbf{1}_{n_a}') - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n')\boldsymbol{Y} =: \boldsymbol{Y}'\boldsymbol{C}_2\boldsymbol{Y}. \end{cases}$$

回忆

$$\boldsymbol{Y}=(y_{11},\cdots,y_{1n_1},\cdots,y_{a1},\cdots,y_{an_a})'.$$

记$\mathbf{1}_n$为元素全为1的n维的列向量, $\mathbf{1}_{n_i}$为元素全为1的n_i维的列向量, $i=1,\cdots,a$. 则可知, 若H_0为真,

$$\boldsymbol{Y}\sim N(\mu\mathbf{1}_n,\sigma^2\boldsymbol{I}_n),\ \frac{\boldsymbol{Y}}{\sigma}\sim N(\frac{\mu}{\sigma}\mathbf{1}_n,\boldsymbol{I}_n),$$

$$\begin{cases}\mathsf{SS}_T=\boldsymbol{Y}'(\boldsymbol{I}_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n')\boldsymbol{Y}=:\boldsymbol{Y}'\boldsymbol{C}\boldsymbol{Y},\\ \mathsf{SS}_E=\boldsymbol{Y}'(\boldsymbol{I}_n-\mathsf{diag}(\frac{1}{n_1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}',\cdots,\frac{1}{n_a}\mathbf{1}_{n_a}\mathbf{1}_{n_a}'))\boldsymbol{Y}=:\boldsymbol{Y}'\boldsymbol{C}_1\boldsymbol{Y},\\ \mathsf{SS}_A=\boldsymbol{Y}'(\mathsf{diag}(\frac{1}{n_1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}',\cdots,\frac{1}{n_a}\mathbf{1}_{n_a}\mathbf{1}_{n_a}')-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n')\boldsymbol{Y}=:\boldsymbol{Y}'\boldsymbol{C}_2\boldsymbol{Y}.\end{cases}$$

注 第二个等式成立是因为

$$\boldsymbol{I}_n-\mathsf{diag}(\frac{1}{n_1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}',\cdots,\frac{1}{n_a}\mathbf{1}_{n_a}\mathbf{1}_{n_a}')=\mathsf{diag}(\boldsymbol{I}_{n_1}-\frac{1}{n_1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}',\cdots,\boldsymbol{I}_{n_a}-\frac{1}{n_a}\mathbf{1}_{n_a}\mathbf{1}_{n_a}').$$

现已知:
(1) $\boldsymbol{C}=\boldsymbol{C}_1+\boldsymbol{C}_2$;
(2) $\boldsymbol{C}_2$为非负定矩阵(因为它是对称幂等矩阵, 特征根非0即1);
(3) $\frac{\mathsf{SS}_E}{\sigma^2}=(\frac{\boldsymbol{Y}}{\sigma})'\boldsymbol{C}_1(\frac{\boldsymbol{Y}}{\sigma})\sim\chi^2(n-a)$;
(4) 若H_0为真, $\frac{\mathsf{SS}_T}{\sigma^2}=(\frac{\boldsymbol{Y}}{\sigma})'\boldsymbol{C}(\frac{\boldsymbol{Y}}{\sigma})\sim\chi^2(n-1)$;

现已知:

(1) $\boldsymbol{C} = \boldsymbol{C}_1 + \boldsymbol{C}_2$;

(2) $\boldsymbol{C}_2$为非负定矩阵(因为它是对称幂等矩阵, 特征根非0即1);

(3) $\frac{\mathsf{SS}_E}{\sigma^2} = (\frac{\boldsymbol{Y}}{\sigma})'\boldsymbol{C}_1(\frac{\boldsymbol{Y}}{\sigma}) \sim \chi^2(n-a)$;

(4) 若H_0为真, $\frac{\mathsf{SS}_T}{\sigma^2} = (\frac{\boldsymbol{Y}}{\sigma})'\boldsymbol{C}(\frac{\boldsymbol{Y}}{\sigma}) \sim \chi^2(n-1)$;

则根据定理2.4.4, 若H_0为真, 有

$$\frac{\mathsf{SS}_A}{\sigma^2} = \frac{\mathsf{SS}_T}{\sigma^2} - \frac{\mathsf{SS}_E}{\sigma^2} = (\frac{\boldsymbol{Y}}{\sigma})'\boldsymbol{C}_2(\frac{\boldsymbol{Y}}{\sigma}) \sim \chi^2(a-1, \lambda)$$

且SS_A与SS_E独立, 其中非中心参数

$$\lambda = \left(\frac{\mu}{\sigma}\mathbf{1}_n\right)' \boldsymbol{C}_2 \left(\frac{\mu}{\sigma}\mathbf{1}_n\right) = 0.$$

最后, 可知若H_0为真, 那么检验统计量

$$F=\frac{\mathsf{SS}_A/(a-1)}{\mathsf{SS}_E/(n-a)}\sim F(a-1,n-a). \tag{7.1.6}$$

最后, 可知若H_0为真, 那么检验统计量

$$F = \frac{\mathsf{SS}_A/(a-1)}{\mathsf{SS}_E/(n-a)} \sim F(a-1, n-a). \tag{7.1.6}$$

给定显著性水平α, 假设检验的拒绝域为

$$\{F > F_\alpha(a-1, n-a)\}.$$

最后, 可知若H_0为真, 那么检验统计量

$$F = \frac{\mathsf{SS}_A/(a-1)}{\mathsf{SS}_E/(n-a)} \sim F(a-1, n-a). \tag{7.1.6}$$

给定显著性水平α, 假设检验的拒绝域为

$$\{F > F_\alpha(a-1, n-a)\}.$$

表7.1.2 方差分析表

方差来源	平方和	自由度	均方	F比
因素A 误差	SS_A SS_E	$a-1$ $n-a$	$\mathsf{MS}_A = \frac{\mathsf{SS}_A}{a-1}$ $\mathsf{MS}_E = \frac{\mathsf{SS}_E}{n-a}$	$F = \frac{\mathsf{MS}_A}{\mathsf{MS}_E}$
总和	SS_T	$n-1$		

例7.1.1 设有三个小麦品种, 经试种得每公顷产量数据如下表(单位: kg/hm^2).

表7.1.3: 小麦品种试验数据

品种＼试验号	1	2	3	4	5
1	4350	4650	4080	4275	
2	4125	3720	3810	3960	3930
3	4695	4245	4620		

问: 不同品种的小麦产量之间有无显著的差异?

计算与分析过程:

$\bar{y} = 4205$; $\bar{y}_{1\cdot} = 4339$, $n_1 = 4$; $\bar{y}_{2\cdot} = 3909$, $n_2 = 5$; $\bar{y}_{3\cdot} = 4520$, $n_3 = 3$; 总平方和$\mathsf{SS}_T = 1186800$, 自由度为$n-1 = 12-1 = 11$; 效应平方和$\mathsf{SS}_A = 807311.25$, 自由度为$a-1 = 2$; 误差平方和$\mathsf{SS}_E = 379488.75$, 自由度为$n-a = 12-3 = 9$. F检验统计量的样本值为

$$F = \frac{807311.25/2}{379488.75/9} = 9.57.$$

计算与分析过程:

$\bar{y} = 4205$; $\bar{y}_{1\cdot} = 4339$, $n_1 = 4$; $\bar{y}_{2\cdot} = 3909$, $n_2 = 5$; $\bar{y}_{3\cdot} = 4520$, $n_3 = 3$; 总平方和$\mathrm{SS}_T = 1186800$, 自由度为$n - 1 = 12 - 1 = 11$; 效应平方和$\mathrm{SS}_A = 807311.25$, 自由度为$a - 1 = 2$; 误差平方和$\mathrm{SS}_E = 379488.75$, 自由度为$n - a = 12 - 3 = 9$. F检验统计量的样本值为

$$F = \frac{807311.25/2}{379488.75/9} = 9.57.$$

表7.1.4 小麦品种的方差分析表

方差来源	平方和	自由度	均方	F比
因素A	807311.25	2	403655.62	9.57
误差	379488.75	9	42165.42	
总和	1186800	11		

查表得$F_{0.05}(2, 9) = 4.26 < 9.57$. 所以认为小麦品种的效应具有显著的差异.

R程序:

```
wheat=data.frame(
X=c(4350,4650,4080,4275,4125,3720,3810,3960,3930,4695,4245,4620),
A=factor(rep(1:3,c(4,5,3)))   (产生因子变量A)
)
wheat.aov=aov(X~A,data=wheat)
summary(wheat.aov)
```

R程序:

```
wheat=data.frame(
X=c(4350,4650,4080,4275,4125,3720,3810,3960,3930,4695,4245,4620),
A=factor(rep(1:3,c(4,5,3)))   (产生因子变量A)
)
wheat.aov=aov(X~A,data=wheat)
summary(wheat.aov)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
A            2 807311  403656   9.573 0.00591 **
Residuals    9 379489   42165
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

$\{y_{ij}\}$的方差齐性?

$\{y_{ij}\}$的方差齐性?

对方差齐性的检验可使用Bartlett检验方法:

bartlett.test(X~A,data=wheat)

```
> bartlett.test(X~A,data=wheat)

        Bartlett test of homogeneity of variances

data:  X by A
Bartlett's K-squared = 0.68263, df = 2, p-value = 0.7108
```

由于p-value大于0.05, 所以接受方差齐性假设.

如果F检验的结论是拒绝原假设，则表明从现在掌握的数据看，我们有理由认为因素A的a个水平效应之间有显著的差异，也就是说，$\mu_1, \cdots, \mu_a$不完全相同．这时，需要对每一对μ_i和μ_j之间的差异程度作出估计．这就要对效应之差$\mu_i - \mu_j$作区间估计，或者对$H_0: \mu_i = \mu_j$进行假设检验．

如果F检验的结论是拒绝原假设，则表明从现在掌握的数据看，我们有理由认为因素A的a个水平效应之间有显著的差异，也就是说，$\mu_1, \cdots, \mu_a$不完全相同. 这时，需要对每一对μ_i和μ_j之间的差异程度作出估计. 这就要对效应之差$\mu_i - \mu_j$作区间估计，或者对$H_0: \mu_i = \mu_j$进行假设检验.

不难看出:

$$\frac{(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) - (\mu_i - \mu_j)}{\sigma\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim N(0, 1); \tag{7.1.7}$$

记$\hat{\sigma}^2 = \mathsf{SS}_E/(n-a)$，那么

$$\frac{(n-a)\hat{\sigma}^2}{\sigma^2} = \frac{\mathsf{SS}_E}{\sigma^2} \sim \chi^2(n-a); \tag{7.1.8}$$

且

$$(\bar{y}_{i\cdot} - \bar{y}_{j\cdot})\text{与}\mathsf{SS}_E\text{相互独立}.$$

所以

$$\frac{(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) - (\mu_i - \mu_j)}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim$$

所以

$$\frac{(\bar{y}_{i\cdot}-\bar{y}_{j\cdot})-(\mu_i-\mu_j)}{\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}}\sim t(n-a).$$

因此在$H_0:\mu_i=\mu_j$成立时, 检验统计量

$$t_{ij}=\frac{\bar{y}_{i\cdot}-\bar{y}_{j\cdot}}{\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}}\sim t(n-a).$$

给定显著性水平α, 检验的拒绝域为

$$W=\{|t_{ij}|>t_{\frac{\alpha}{2}}(n-a)\}.$$

或者用区间估计的方法进行统计推断. $\mu_i-\mu_j$的置信水平为$1-\alpha$的置信区间为

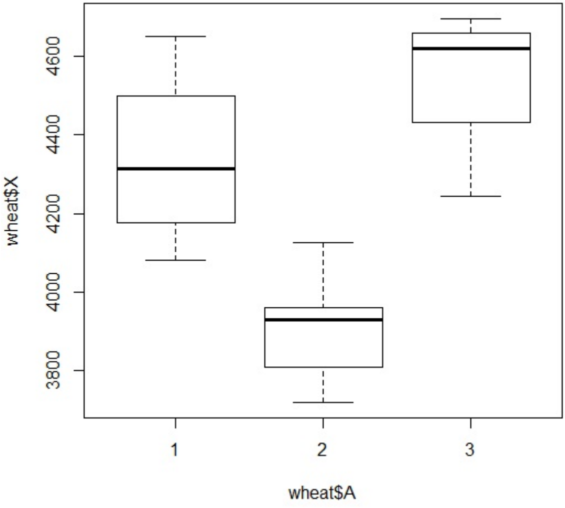
$$\left(\bar{y}_{i\cdot}-\bar{y}_{j\cdot}-\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}t_{\frac{\alpha}{2}}(n-a),\bar{y}_{i\cdot}-\bar{y}_{j\cdot}+\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}t_{\frac{\alpha}{2}}(n-a)\right).$$

或者用区间估计的方法进行统计推断. $\mu_i-\mu_j$的置信水平为$1-\alpha$的置信区间为

$$\left(\bar{y}_{i\cdot}-\bar{y}_{j\cdot}-\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}t_{\frac{\alpha}{2}}(n-a),\bar{y}_{i\cdot}-\bar{y}_{j\cdot}+\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}t_{\frac{\alpha}{2}}(n-a)\right).$$

如果这个区间包含0, 则表明我们以概率$1-\alpha$断言μ_i与μ_j没有显著差异; 如果整个区间落在0的左边, 则我们以概率$1-\alpha$断言μ_i小于μ_j; 如果整个区间落在0的右边, 则我们以概率$1-\alpha$断言μ_i大于μ_j.

例7.1.1续 μ_1, μ_2, μ_3之间的差异.
R程序: plot(wheat$X~wheat$A)



首先计算各个因子间的均值, 再用多重t检验的方法进行统计推断.

首先计算各个因子间的均值, 再用多重t检验的方法进行统计推断.

```
> mu=tapply(wheat$X,wheat$A,mean)
> mu
      1       2       3 
4338.75 3909.00 4520.00 
> 
```

说明: tapply用于分组统计.

```
> pairwise.t.test(wheat$X,wheat$A,p.adjust.method="none")

        Pairwise comparisons using t tests with pooled SD

data:  wheat$X and wheat$A

  1      2
2 0.0123 -
3 0.2776 0.0028

P value adjustment method: none
> |
```

因此我们以95%的把握断言：μ_1和μ_2有显著差异，μ_2和μ_3有显著差异，μ_1和μ_3无显著差异.

但要注意的是, 现在进行的是多个假设检验, 这些p-value是不可靠的, 为什么呢?

但要注意的是, 现在进行的是多个假设检验, 这些p-value是不可靠的, 为什么呢?

假设μ_1, μ_2, μ_3之间并无显著差别, 那么两两比较的显著性检验的p-value在区间$[0, 1]$上呈现均匀分布. 因此, 至少出现一个虚假的显著性结果的概率为

$$1 - 0.95^3 = 0.142.$$

若同时进行6个两两比较的显著性检验, 则至少出现一个虚假的显著性结果的概率上升为

$$1 - 0.95^6 = 0.265.$$

所以有必要对p-value进行修正.

其中的一个修正方法是Bonferroni修正法. 假设需要进行两两比较的个数是m个, 那么Bonferroni修正法把满足$m \times p$-value$< \alpha$的结果才称为是统计显著的(即有显著性差异).

显然, Bonferroni修正法过于保守了. 另一种常用的修正方法是FDR(False Discovery Rate)方法, 它由Benjamini和Hochberg在1995年提出. 该方法先对m个p-value进行排序:

$$p_{(1)} \leq \cdots \leq p_{(m)},$$

然后选择满足

$$p_{(i)} \leq \alpha i/m$$

的最大i值, 并把$p_{(1)}, \cdots, p_{(i)}$所对应的结果称为是统计显著的(即有显著性差异).

再回到前面的例子, 我们对两两t检验的p-value进行修正.

pairwise.t.test(wheat$X,wheat$A,p.adjust.method=“bonferroni”)
pairwise.t.test(wheat$X,wheat$A,p.adjust.method=“fdr”)

```
> pairwise.t.test(wheat$X,wheat$A,p.adjust.method="bonferroni")

        Pairwise comparisons using t tests with pooled SD

data:  wheat$X and wheat$A

  1      2
2 0.0370 -
3 0.8327 0.0083
```

```
> pairwise.t.test(wheat$X,wheat$A,p.adjust.method="fdr")

        Pairwise comparisons using t tests with pooled SD

data:  wheat$X and wheat$A

  1      2
2 0.0185 -
3 0.2776 0.0083
```

可以看出, 若取$\alpha = 0.05$, Bonferroni方法和FDR方法仍认为μ_1和μ_2有显著差异, μ_2和μ_3有显著差异. 但若取$\alpha = 0.03$, 则Bonferroni方法只认为μ_2和μ_3有显著差异, 而FDR方法仍维持$\alpha = 0.05$时的判断.

在一项实际试验中, 往往有这样的情况, 研究者本想考察某个因素对指标的影响, 但是由于客观条件的限制, 还有个别因素不可能在所有试验中把它们控制在完全相同的状态. 譬如, 在上一节开始讨论的小麦品种的例子中, 实验者要研究的是“小麦品种”这一因素对产量的影响. 但在实际中可能会出现这样的情况: 很难找到一大块田, 其土质肥沃程度完全一样. 因此“土质”就成为另一个因素不可避免地进入了试验, 导致了两因素的试验问题.

在农业试验中解决这个问题的方法是采用所谓的区组设计. 它的做法是, 先把一块田(大块)分成若干块(小块), 譬如b(小)块, 使得同一大块田里的各小块的土质肥沃程度基本上保持一致. 在试验设计中, 称这种大块(田)为区组, 然后把每一个区组又分成若干小块(田), 称为试验单元. 现在有a个小麦品种, 一个方便的做法就是把每个区组(大块)分成a个试验单元(小块). 在每一个试验单元上种植一种小麦.

在农业试验中解决这个问题的方法是采用所谓的区组设计. 它的做法是, 先把一块田(大块)分成若干块(小块), 譬如b(小)块, 使得同一大块田里的各小块的土质肥沃程度基本上保持一致. 在试验设计中, 称这种大块(田)为区组, 然后把每一个区组又分成若干小块(田), 称为试验单元. 现在有a个小麦品种, 一个方便的做法就是把每个区组(大块)分成a个试验单元(小块). 在每一个试验单元上种植一种小麦.

若用y_{ij}表示在第j个区组中种植第i种小麦的那个试验单元的产量, 则y_{ij}就可表为

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \quad i = 1, \cdots, a, \quad j = 1, \cdots, b, \qquad (7.2.1)$$

这里μ称为总平均, α_i为第i个小麦品种的效应, β_j为第j个区组的效应, e_{ij}为随机误差.

在实际应用中, 更多的情况是研究者所感兴趣的问题本身就是两因素的.

在实际应用中, 更多的情况是研究者所感兴趣的问题本身就是两因素的.

例如, 在一项工业试验中, 影响产品质量的因素是反应温度和反应压力. 实验者的目的是选择最好的生产条件, 若反应温度有a个水平, 反应压力有b个水平, 记y_{ij}为在反应温度处于第i个水平和反应压力处于第j个水平时产品质量的指标值, 那么y_{ij}也有表达式(7.2.1).

在实际应用中, 更多的情况是研究者所感兴趣的问题本身就是两因素的.

例如, 在一项工业试验中, 影响产品质量的因素是反应温度和反应压力. 实验者的目的是选择最好的生产条件, 若反应温度有a个水平, 反应压力有b个水平, 记y_{ij}为在反应温度处于第i个水平和反应压力处于第j个水平时产品质量的指标值, 那么y_{ij}也有表达式(7.2.1).

若影响产品质量的因素除了反应温度和反应压力外, 还有反应时间和催化剂这两个因素, 当我们把任何两个因素控制在某一状态而研究剩余两个因素对产品质量的影响时, 这同样导致一个两因素的问题.

考虑一般的两因素试验问题, 将这两个因素分别记为A和B. 假定因素A有a个不同的水平, 记为$A_1, \cdots, A_a$, 而因素B有b个不同的水平, 记为$B_1, \cdots, B_b$.

考虑一般的两因素试验问题，将这两个因素分别记为A和B. 假定因素A有a个不同的水平，记为$A_1, \cdots, A_a$, 而因素B有b个不同的水平，记为$B_1, \cdots, B_b$.

在因素A和B的各个水平的组合下做c次试验. 记y_{ijk}为在水平组合(A_i, B_j)下的第k次试验的指标值. 对固定的i和j, $\{y_{ij1}, \cdots, y_{ijc}\}$都是在水平组合$(A_i, B_j)$下的指标观测值，可以把它们看成来自同一个正态总体的样本，均值为μ_{ij}. 于是

$$\text{给定} i, j, \quad y_{ijk} \overset{i.i.d.}{\sim} N(\mu_{ij}, \sigma^2), \quad k = 1, \cdots, c. \tag{7.2.2}$$

表7.2.1 两因素方差分析问题

$A_i \diagdown B_j$	B_1	B_2	$\cdots$	B_b
A_1	$y_{111}, y_{112}, \cdots, y_{11c}$	$y_{121}, y_{122} \cdots, y_{12c}$	$\cdots$	$y_{1b1}, y_{1b2} \cdots, y_{1bc}$
A_2	$y_{211}, y_{212}, \cdots, y_{21c}$	$y_{221}, y_{222} \cdots, y_{22c}$	$\cdots$	$y_{2b1}, y_{2b2} \cdots, y_{2bc}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_a	$y_{a11}, y_{a12}, \cdots, y_{a1c}$	$y_{a21}, y_{a22} \cdots, y_{a2c}$	$\cdots$	$y_{ab1}, y_{ab2} \cdots, y_{abc}$

进一步模型化: 将(7.2.2)改写成

$$\begin{cases} y_{ijk} = \mu_{ij} + e_{ijk}, \ i = 1, \cdots, a, \ j = 1, \cdots, b, \ k = 1, \cdots, c, \\ e_{ijk} \overset{i.i.d.}{\sim} N(0, \sigma^2). \end{cases} \tag{7.2.3}$$

为进行统计分析, 将μ_{ij}做适当的分解. 记

$$\begin{aligned} \mu =& \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}, \ \ \bar{\mu}_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b}\mu_{ij}, \ \ \bar{\mu}_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a}\mu_{ij}, \\ \alpha_i =& \bar{\mu}_{i\cdot} - \mu, \ \ i = 1, \cdots, a, \\ \beta_j =& \bar{\mu}_{\cdot j} - \mu, \ \ j = 1, \cdots, b, \\ \gamma_{ij} =& \mu_{ij} - \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot j} + \mu, \ \ i = 1, \cdots, a; \ \ j = 1, \cdots, b, \end{aligned}$$

其中μ为总平均, α_i为因素A的水平A_i的效应, β_j为因素B的水平B_j的效应, γ_{ij}可写为

$$\begin{aligned}\gamma_{ij} =&\mu_{ij} - (\bar{\mu}_{i\cdot} - \mu) - (\bar{\mu}_{\cdot j} - \mu) - \mu \\ =&(\mu_{ij} - \mu) - \alpha_i - \beta_j,\end{aligned}$$

表示A_i和B_j的交互效应. 通常把因素A和B对试验指标的交互效应设想为某一因素的效应, 称这个因素为A与B的交互作用, 记为$A \times B$. 不难验证

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

注意到μ_{ij}可改写为$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$, 因此(7.2.3)可写成

$$
\begin{cases}
y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \\
\qquad\qquad i = 1, \cdots, a; \;\; j = 1, \cdots, b; \;\; k = 1, \cdots, c, \\
e_{ijk} \overset{i.i.d.}{\sim} N(0, \sigma^2), \\
\sum_{i=1}^{a} \alpha_i = 0, \;\; \sum_{j=1}^{b} \beta_j = 0, \;\; \sum_{i=1}^{a} \sum_{j=1}^{b} \gamma_{ij} = 0.
\end{cases}
\tag{7.2.4}
$$

这就是两因素方差分析模型.

注意到μ_{ij}可改写为$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$, 因此(7.2.3)可写成

$$\begin{cases} y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \\ \qquad\qquad i = 1, \cdots, a; \;\; j = 1, \cdots, b; \;\; k = 1, \cdots, c, \\ e_{ijk} \overset{i.i.d.}{\sim} N(0, \sigma^2), \\ \sum_{i=1}^a \alpha_i = 0, \;\; \sum_{j=1}^b \beta_j = 0, \;\; \sum_{i=1}^a \sum_{j=1}^b \gamma_{ij} = 0. \end{cases} \tag{7.2.4}$$

这就是两因素方差分析模型.

注 其实, $\sum_{i=1}^a \sum_{j=1}^b \gamma_{ij} = 0$应写成

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

这里共有$a + b - 1$个约束.

无交互效应的情形

假设$\gamma_{ij} = 0, i = 1, \cdots, a, j = 1, \cdots, b$, 即不存在交互效应. 此时, 只考虑每种水平组合下试验次数为$c = 1$的情形.

假设$\gamma_{ij}=0, i=1,\cdots,a, j=1,\cdots,b$, 即不存在交互效应. 此时, 只考虑每种水平组合下试验次数为$c=1$的情形.

模型(7.2.4)可写为

$$\begin{cases} y_{ij}=\mu+\alpha_i+\beta_j+e_{ij}, \quad i=1,\cdots,a, \ j=1,\cdots,b, \\ e_{ij} \overset{i.i.d.}{\sim} N(0,\sigma^2), \\ \sum_{i=1}^a \alpha_i=0, \ \ \sum_{j=1}^b \beta_j=0. \end{cases} \tag{7.2.5}$$

这就是无交互效应的两因素方差分析模型.

试验的目的是考察因素A或B的各个水平对试验指标的影响有无显著差异, 即检验

$$H_1 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$$

和

$$H_2 : \beta_1 = \beta_2 = \cdots = \beta_b = 0.$$

试验的目的是考察因素A或B的各个水平对试验指标的影响有无显著差异, 即检验

$$H_1 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$$

和

$$H_2 : \beta_1 = \beta_2 = \cdots = \beta_b = 0.$$

我们采用与单因素方差分析模型类似的方法导出检验统计量.

记

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

$$\mathsf{SS}_T = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y})^2.$$

SS_T是全部试验数据的离差平方和, 称为总平方和.

记

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

$$\mathsf{SS}_T=\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y})^2.$$

SS_T是全部试验数据的离差平方和, 称为总平方和. 对其进行分解得

$$\begin{aligned}\mathsf{SS}_T&=\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y}+\bar{y}_{i\cdot}-\bar{y}+\bar{y}_{\cdot j}-\bar{y})^2\\&=\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y})^2+\sum_{i=1}^{a}b(\bar{y}_{i\cdot}-\bar{y})^2+\sum_{j=1}^{b}a(\bar{y}_{\cdot j}-\bar{y})^2\\&\triangleq\mathsf{SS}_E+\mathsf{SS}_A+\mathsf{SS}_B.\end{aligned}$$

$\sum_{i=1}^{a}(\bar{y}_{i\cdot}-\bar{y})^2$反映了$\bar{y}_{1\cdot},\cdots,\bar{y}_{a\cdot}$差异的程度. 因为$\bar{y}_{i\cdot}$是水平$A_i$下所有观测值的平均, 所以这种差异是由于因素$A$的不同水平所引起的. 因此称$\mathsf{SS}_A$为因素$A$的平方和. 类似地, 称$\mathsf{SS}_B$为因素$B$的平方和.

SS_E反映了试验的随机误差的影响, 称为误差平方和. 事实上, 可写

$$\begin{aligned}\mathsf{SS}_E &= \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y})^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b}[y_{ij}-(\bar{y}_{i\cdot}-\bar{y})-(\bar{y}_{\cdot j}-\bar{y})-\bar{y}]^2,\end{aligned}$$

而$\mathsf{E}(\bar{y}_{i\cdot}-\bar{y})=\bar{\mu}_{i\cdot}-\mu=\alpha_i$, $\mathsf{E}(\bar{y}_{\cdot j}-\bar{y})=\bar{\mu}_{\cdot j}-\mu=\beta_j$, $\mathsf{E}(\bar{y})=\mu$, 所以可把$y_{ij}-(\bar{y}_{i\cdot}-\bar{y})-(\bar{y}_{\cdot j}-\bar{y})-\bar{y}$近似看作$y_{ij}-\mu-\alpha_i-\beta_j=e_{ij}$.

可以证明:

$$\mathsf{SS}_E/\sigma^2 \sim \chi^2((a-1)(b-1));$$

当H_1成立时,

$$\mathsf{SS}_A/\sigma^2 \sim \chi^2(a-1)\text{且与}\mathsf{SS}_E\text{独立};$$

当H_2成立时,

$$\mathsf{SS}_B/\sigma^2 \sim \chi^2(b-1)\text{且与}\mathsf{SS}_E\text{独立}.$$

注 (1) 证明$\mathsf{SS}_E/\sigma^2 \sim \chi^2((a-1)(b-1))$的关键是把$\mathsf{SS}_E$写成

$$\begin{aligned}\mathsf{SS}_E &= \sum_{i=1}^{a}\sum_{j=1}^{b}(e_{ij}-\bar{e}_{i\cdot}-\bar{e}_{\cdot j}+\bar{e})^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b}(e_{ij}-\bar{e}_{i\cdot})^2-\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{e}_{\cdot j}-\bar{e})^2,\end{aligned}$$

其中

$$\bar{e}=\frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}e_{ij},\ \ \bar{e}_{i\cdot}=\frac{1}{b}\sum_{j=1}^{b}e_{ij},\ \ \bar{e}_{\cdot j}=\frac{1}{a}\sum_{i=1}^{a}e_{ij}.$$

注 (1) 证明$\mathsf{SS}_E/\sigma^2 \sim \chi^2((a-1)(b-1))$的关键是把$\mathsf{SS}_E$写成

$$\begin{aligned}\mathsf{SS}_E &= \sum_{i=1}^{a}\sum_{j=1}^{b}(e_{ij}-\bar{e}_{i\cdot}-\bar{e}_{\cdot j}+\bar{e})^2\\ &= \sum_{i=1}^{a}\sum_{j=1}^{b}(e_{ij}-\bar{e}_{i\cdot})^2-\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{e}_{\cdot j}-\bar{e})^2,\end{aligned}$$

其中

$$\bar{e}=\frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}e_{ij},\quad \bar{e}_{i\cdot}=\frac{1}{b}\sum_{j=1}^{b}e_{ij},\quad \bar{e}_{\cdot j}=\frac{1}{a}\sum_{i=1}^{a}e_{ij}.$$

因此$\sum_{i=1}^{a}\sum_{j=1}^{b}(e_{ij}-\bar{e}_{i\cdot})^2$有如下的平方和分解式:

$$\sum_{i=1}^{a}\sum_{j=1}^{b}(e_{ij}-\bar{e}_{i\cdot})^2=\sum_{i=1}^{a}\sum_{j=1}^{b}(e_{ij}-\bar{e}_{i\cdot}-\bar{e}_{\cdot j}+\bar{e})^2+\sum_{j=1}^{b}a(\bar{e}_{\cdot j}-\bar{e})^2.$$

(2) 证明当H_1成立时, $\mathsf{SS}_A/\sigma^2 \sim \chi^2(a-1)$的关键是写$\mathsf{SS}_A$为

$$\mathsf{SS}_A = \sum_{i=1}^{a} b(\bar{y}_{i\cdot} - \bar{y})^2 = \sum_{i=1}^{a} (\sqrt{b}\bar{y}_{i\cdot} - \sqrt{b}\bar{y})^2,$$

其中$\sqrt{b}\bar{y}_{i\cdot} \sim N(\mu + \alpha_i, \sigma^2)$. 所以当$H_1$成立时, $\sqrt{b}\bar{y}_{i\cdot} \sim N(\mu, \sigma^2)$, $\{\sqrt{b}\bar{y}_{i\cdot}, i = 1, \cdots, a\}$为来自$N(\mu, \sigma^2)$的样本.

(2) 证明当H_1成立时, $\mathsf{SS}_A/\sigma^2 \sim \chi^2(a-1)$的关键是写$\mathsf{SS}_A$为

$$\mathsf{SS}_A = \sum_{i=1}^{a} b(\bar{y}_{i\cdot} - \bar{y})^2 = \sum_{i=1}^{a} (\sqrt{b}\bar{y}_{i\cdot} - \sqrt{b}\bar{y})^2,$$

其中$\sqrt{b}\bar{y}_{i\cdot} \sim N(\mu+\alpha_i, \sigma^2)$. 所以当$H_1$成立时, $\sqrt{b}\bar{y}_{i\cdot} \sim N(\mu, \sigma^2)$, $\{\sqrt{b}\bar{y}_{i\cdot}, i = 1, \cdots, a\}$为来自$N(\mu, \sigma^2)$的样本.

(3) 证明当H_2成立时, $\mathsf{SS}_B/\sigma^2 \sim \chi^2(b-1)$的关键是写$\mathsf{SS}_B$为

$$\mathsf{SS}_B = \sum_{j=1}^{b} a(\bar{y}_{\cdot j} - \bar{y})^2 = \sum_{j=1}^{b} (\sqrt{a}\bar{y}_{\cdot j} - \sqrt{a}\bar{y})^2,$$

其中$\sqrt{a}\bar{y}_{\cdot j} \sim N(\mu+\beta_j, \sigma^2)$. 所以当$H_2$成立时, $\sqrt{a}\bar{y}_{\cdot j} \sim N(\mu, \sigma^2)$, $\{\sqrt{a}\bar{y}_{\cdot j}, j = 1, \cdots, b\}$为来自$N(\mu, \sigma^2)$的样本.

于是当H_1成立时,

$$F_A = \frac{\mathsf{SS}_A/(a-1)}{\mathsf{SS}_E/[(a-1)(b-1)]} \sim F(a-1,(a-1)(b-1)). \tag{7.2.6}$$

给定显著性水平α, 假设检验的拒绝域为

$$W = \{F_A > F_\alpha(a-1,(a-1)(b-1))\}.$$

同理, 当H_2成立时,

$$F_B = \frac{\mathsf{SS}_B/(b-1)}{\mathsf{SS}_E/[(a-1)(b-1)]} \sim F(b-1,(a-1)(b-1)). \qquad (7.2.7)$$

给定显著性水平α, 假设检验的拒绝域为

$$W = \{F_B > F_\alpha(b-1,(a-1)(b-1))\}.$$

表7.2.2 无交互效应的两因素方差分析表

方差来源	平方和	自由度	均方	F比
因素A	SS_A	$a-1$	$\mathrm{MS}_A=\frac{\mathrm{SS}_A}{a-1}$	$F_A=\frac{\mathrm{MS}_A}{\mathrm{MS}_E}$
因素B	SS_B	$b-1$	$\mathrm{MS}_B=\frac{\mathrm{SS}_B}{b-1}$	$F_B=\frac{\mathrm{MS}_B}{\mathrm{MS}_E}$
误差	SS_E	$(a-1)(b-1)$	$\mathrm{MS}_E=\frac{\mathrm{SS}_E}{(a-1)(b-1)}$	
总计	SS_T	$ab-1$		

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. 对于每种燃料和推进器的组合做一次试验, 得到试验数据如下表. 问各种燃料之间及各种推进器之间有无显著差异?

表7.2.3 火箭试验数据

燃料A\推进器B	B_1	B_2	B_3
A_1	58.2	56.2	65.3
A_2	49.1	54.1	51.6
A_3	60.1	70.9	39.23
A_4	75.8	58.2	48.7

这是一个两因素试验且不考虑交互效应. 记燃料为因素A, 它有4个水平, 水平效应为$\alpha_1, \cdots, \alpha_4$; 推进器为因素$B$, 它有3个水平, 水平效应为$\beta_1, \beta_2, \beta_3$. 在显著性水平$\alpha = 0.05$下检验

$$\begin{aligned} H_1 &: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0, \\ H_2 &: \beta_1 = \beta_2 = \beta_3 = 0. \end{aligned}$$

经计算(略)得

表7.2.4 火箭试验的方差分析表

方差来源	平方和	自由度	均方	F比
因素A	157.59	3	52.53	F_A =0.4306
因素B	223.85	2	111.93	F_B =0.9175
误差	731.98	6	122.00	
总计	1113.42	11		

经计算(略)得

表7.2.4 火箭试验的方差分析表

方差来源	平方和	自由度	均方	F比
因素A	157.59	3	52.53	F_A =0.4306
因素B	223.85	2	111.93	F_B =0.9175
误差	731.98	6	122.00	
总计	1113.42	11		

因为$F_{0.05}(3,6) = 4.76, F_{0.05}(2,6) = 5.14$, 而$F_A = 0.4306 < 4.76$, $F_B = 0.9175 < 5.14$, 所以接受H_1和H_2, 即认为各种燃料和各种推进器之间的差异对于火箭射程均无显著影响.

R程序:

R程序:

```
rocket=data.frame(
Y=c(58.2,56.2,65.3,49.1,54.1,51.6,60.1,70.9,39.23,75.8,58.2,48.7),
A=gl(4,3),
B=gl(3,1,12)
)
rocket.aov=aov(Y~A+B,data=rocket)
summary(rocket.aov)
```

说明: gl(4,3)用来产生因子, 该因子有4个水平, 每个水平重复3次. gl(3,1,12)用来产生因子, 该因子有3个水平, 每个水平重复1次, 循环产生12个因子值.

```
> summary(rocket.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
A            3  157.6   52.52   0.431  0.739
B            2  223.5  111.74   0.917  0.449
Residuals    6  731.3  121.88
> 
```

如果经过F检验, H_1被拒绝, 那么在这种情况下, 认为因素A的a个水平效应$\alpha_1, \cdots, \alpha_a$不全相同. 此时希望比较$\alpha_i$的大小, 这需要做$H_0: \alpha_i = \alpha_k$的假设检验或者$\alpha_i - \alpha_k$区间估计.

如果经过F检验, H_1被拒绝, 那么在这种情况下, 认为因素A的a个水平效应$\alpha_1, \cdots, \alpha_a$不全相同. 此时希望比较$\alpha_i$的大小, 这需要做$H_0: \alpha_i = \alpha_k$的假设检验或者$\alpha_i - \alpha_k$区间估计.

因为$y_{ij} \sim N(\mu + \alpha_i + \beta_j, \sigma^2)$, 利用$\sum_{j=1}^{b} \beta_j = 0$易知

$$\bar{y}_{i\cdot} \sim N(\mu + \alpha_i, \frac{\sigma^2}{b}), \quad i = 1, \cdots, a.$$

于是

$$\bar{y}_{i\cdot} - \bar{y}_{k\cdot} \sim N(\alpha_i - \alpha_k, \frac{2\sigma^2}{b}). \tag{7.2.8}$$

注意到

$$\frac{\mathsf{SS}_E}{\sigma^2} \sim \chi^2((a-1)(b-1)),$$

且与$\bar{y}_{i\cdot} - \bar{y}_{k\cdot}$独立. 因此对固定的$i, k$, $H_0 : \alpha_i = \alpha_k$的检验统计量

$$t_{ik} = \frac{\sqrt{b}(\bar{y}_{i\cdot} - \bar{y}_{k\cdot})}{\sqrt{2}\hat{\sigma}} \overset{H_0}{\sim} t((a-1)(b-1)).$$

给定显著性水平α, 检验的拒绝域为

$$W = \left\{ |t_{ik}| > t_{\frac{\alpha}{2}}((a-1)(b-1)) \right\}.$$

或者考虑区间估计. $\alpha_i - \alpha_k$的置信水平为$1-\alpha$的置信区间为

$$\Big(\bar{y}_{i\cdot} - \bar{y}_{k\cdot} - \sqrt{\frac{2}{b}}\hat{\sigma} t_{\frac{\alpha}{2}}((a-1)(b-1)),\quad \bar{y}_{i\cdot} - \bar{y}_{k\cdot} + \sqrt{\frac{2}{b}}\hat{\sigma} t_{\frac{\alpha}{2}}((a-1)(b-1))\Big). \tag{7.2.9}$$

或者考虑区间估计. $\alpha_i - \alpha_k$的置信水平为$1-\alpha$的置信区间为

$$\left(\bar{y}_{i\cdot} - \bar{y}_{k\cdot} - \sqrt{\frac{2}{b}}\hat{\sigma} t_{\frac{\alpha}{2}}((a-1)(b-1)),\right.$$
$$\left.\bar{y}_{i\cdot} - \bar{y}_{k\cdot} + \sqrt{\frac{2}{b}}\hat{\sigma} t_{\frac{\alpha}{2}}((a-1)(b-1))\right). \quad (7.2.9)$$

如果这个区间包含0, 则表明我们以概率$1-\alpha$断言α_i和α_k没有显著差异. 如果整个区间落在0的左边, 则我们以概率$1-\alpha$断言α_i小于α_k. 如果整个区间落在0的右边, 则我们以概率$1-\alpha$断言α_i大于α_k.

若经过F检验, 假设H_2被拒绝, 类似于以上的讨论, 可以构建$H_0: \beta_j = \beta_k$的检验统计量

$$t_{jk} = \frac{\sqrt{a}(\bar{y}_{\cdot j} - \bar{y}_{\cdot k})}{\sqrt{2}\hat{\sigma}} \overset{H_0}{\sim} t((a-1)(b-1)).$$

给定显著性水平α, 检验的拒绝域为

$$W = \left\{ |t_{jk}| > t_{\frac{\alpha}{2}}((a-1)(b-1)) \right\}.$$

或者构建$\beta_j - \beta_k$的置信水平为$1-\alpha$的置信区间:

$$\Big(\bar{y}_{\cdot j} - \bar{y}_{\cdot k} - \sqrt{\frac{2}{a}}\hat{\sigma} t_{\frac{\alpha}{2}}((a-1)(b-1)),$$
$$\bar{y}_{\cdot j} - \bar{y}_{\cdot k} + \sqrt{\frac{2}{a}}\hat{\sigma} t_{\frac{\alpha}{2}}((a-1)(b-1))\Big). \quad (7.2.10)$$

或者构建$\beta_j - \beta_k$的置信水平为$1-\alpha$的置信区间:

$$\Big(\bar{y}_{\cdot j} - \bar{y}_{\cdot k} - \sqrt{\frac{2}{a}}\hat{\sigma} t_{\frac{\alpha}{2}}((a-1)(b-1)),$$
$$\bar{y}_{\cdot j} - \bar{y}_{\cdot k} + \sqrt{\frac{2}{a}}\hat{\sigma} t_{\frac{\alpha}{2}}((a-1)(b-1))\Big). \quad (7.2.10)$$

如果这个区间包含0, 则表明可以概率$1-\alpha$断言β_j和β_k没有显著差异. 如果整个区间落在0的左边, 则可以概率$1-\alpha$断言β_j小于β_k. 如果整个区间落在0的右边, 则可以概率$1-\alpha$断言β_j大于β_k.

与单因素方差分析类似, 做多重假设检验的时候, 需要对p-value进行调整. 对刚才的例子, 若H_1被拒绝(其实没有被拒绝), 则还需进行多重假设检验, 分析哪些水平之间有显著差异.

pairwise.t.test(rocket\$Y,rocket\$A,p.adjust.method="fdr")

```
> pairwise.t.test(rocket$Y,rocket$A,p.adjust.method="fdr")

        Pairwise comparisons using t tests with pooled SD

data:  rocket$Y and rocket$A

  1    2    3
2 0.88 -    -
3 0.88 0.88 -
4 0.91 0.88 0.88
```

同样地, 若H_2被拒绝(其实没有被拒绝), 也还需进行多重假设检验, 分析哪些水平之间有显著差异.

pairwise.t.test(rocket$Y,rocket$B,p.adjust.method= “fdr”)

```
> pairwise.t.test(rocket$Y,rocket$B,p.adjust.method="fdr")

        Pairwise comparisons using t tests with pooled SD

data:  rocket$Y and rocket$B

  1    2
2 0.90 -
3 0.37 0.37
```

有交互效应的情形

若要考虑因素A, B之间的交互作用$A \times B$, 那么在各水平组合下需要做重复试验. 设每种组合下试验次数均为$c(c > 1)$. 此时对应的统计模型就是前面的(7.2.4):

$$
\begin{cases}
y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \\
\qquad\qquad i = 1, \cdots, a; \;\; j = 1, \cdots, b; \;\; k = 1, \cdots, c, \\
e_{ijk} \overset{i.i.d.}{\sim} N(0, \sigma^2), \\
\sum_{i=1}^{a} \alpha_i = 0, \;\; \sum_{j=1}^{b} \beta_j = 0, \;\; \sum_{i=1}^{a} \sum_{j=1}^{b} \gamma_{ij} = 0.
\end{cases}
$$

有交互效应的情形

若要考虑因素A, B之间的交互作用$A \times B$, 那么在各水平组合下需要做重复试验. 设每种组合下试验次数均为$c(c > 1)$. 此时对应的统计模型就是前面的(7.2.4):

$$\begin{cases} y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \\ \qquad\qquad i = 1, \cdots, a; \quad j = 1, \cdots, b; \quad k = 1, \cdots, c, \\ e_{ijk} \overset{i.i.d.}{\sim} N(0, \sigma^2), \\ \sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0, \quad \sum_{i=1}^{a} \sum_{j=1}^{b} \gamma_{ij} = 0. \end{cases}$$

在这样的模型下, α_i并不能反映水平A_i的优劣. 这是因为在交互效应存在的情况下, 因子水平A_i的优劣还与因子B的水平有关系. 因此, 对这样的模型, 直接检验$\alpha_1 = \cdots = \alpha_a = 0$与$\beta_1 = \cdots = \beta_b = 0$都是没有实际意义的. 一个重要的检验问题是交互效应是否存在, 即检验

$$H_3 : \gamma_{ij} = 0, \quad i = 1, \cdots, a; \quad j = 1, \cdots, b.$$

有交互效应的情形

若要考虑因素A, B之间的交互作用$A \times B$, 那么在各水平组合下需要做重复试验. 设每种组合下试验次数均为$c(c > 1)$. 此时对应的统计模型就是前面的(7.2.4):

$$\begin{cases} y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \\ \qquad\qquad i = 1, \cdots, a; \;\; j = 1, \cdots, b; \;\; k = 1, \cdots, c, \\ e_{ijk} \overset{i.i.d.}{\sim} N(0, \sigma^2), \\ \sum_{i=1}^a \alpha_i = 0, \;\; \sum_{j=1}^b \beta_j = 0, \;\; \sum_{i=1}^a \sum_{j=1}^b \gamma_{ij} = 0. \end{cases}$$

在这样的模型下, α_i并不能反映水平A_i的优劣. 这是因为在交互效应存在的情况下, 因子水平A_i的优劣还与因子B的水平有关系. 因此, 对这样的模型, 直接检验$\alpha_1 = \cdots = \alpha_a = 0$与$\beta_1 = \cdots = \beta_b = 0$都是没有实际意义的. 一个重要的检验问题是交互效应是否存在, 即检验

$$H_3 : \gamma_{ij} = 0, \;\; i = 1, \cdots, a; \;\; j = 1, \cdots, b.$$

若H_3被接受, 检验$\alpha_1 = \cdots = \alpha_a = 0$与检验$\beta_1 = \cdots = \beta_b = 0$才有意义.

引进记号:

$$\bar{y} = \frac{1}{abc}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} y_{ijk}, \quad \bar{y}_{ij\cdot} = \frac{1}{c}\sum_{k=1}^{c} y_{ijk},$$

$$\bar{y}_{i\cdot\cdot} = \frac{1}{bc}\sum_{j=1}^{b}\sum_{k=1}^{c} y_{ijk}, \quad \bar{y}_{\cdot j\cdot} = \frac{1}{ac}\sum_{i=1}^{a}\sum_{k=1}^{c} y_{ijk}.$$

作平方和分解:

$$
\begin{aligned}
\mathsf{SS}_T &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(y_{ijk}-\bar{y})^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(y_{ijk}-\bar{y}_{ij\cdot}+\bar{y}_{i\cdot\cdot}-\bar{y}+\bar{y}_{\cdot j\cdot}-\bar{y}+\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot j\cdot}+\bar{y})^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(y_{ijk}-\bar{y}_{ij\cdot})^2+bc\sum_{i=1}^{a}(\bar{y}_{i\cdot\cdot}-\bar{y})^2 \\
&\quad + ac\sum_{j=1}^{b}(\bar{y}_{\cdot j\cdot}-\bar{y})^2+c\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot j\cdot}+\bar{y})^2 \\
&\triangleq \mathsf{SS}_E+\mathsf{SS}_A+\mathsf{SS}_B+\mathsf{SS}_{A\times B},
\end{aligned}
$$

其中

$$\mathrm{SS}_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(y_{ijk} - \bar{y}_{ij\cdot})^2,$$

$$\mathrm{SS}_A = bc\sum_{i=1}^{a}(\bar{y}_{i\cdot\cdot} - \bar{y})^2,$$

$$\mathrm{SS}_B = ac\sum_{j=1}^{b}(\bar{y}_{\cdot j\cdot} - \bar{y})^2,$$

$$\mathrm{SS}_{A\times B} = c\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y})^2.$$

称SS_E为误差平方和, SS_A为因素A的平方和, SS_B为因素B的平方和, $\mathrm{SS}_{A\times B}$为交互作用的平方和.

类似于以前的讨论，可以证明当H_3成立时，

$$\begin{aligned} F_{A\times B} &= \frac{\mathsf{SS}_{A\times B}/[(a-1)(b-1)]}{\mathsf{SS}_E/[ab(c-1)]} \\ &\overset{H_3}{\sim} F((a-1)(b-1), ab(c-1)). \end{aligned} \tag{7.2.11}$$

给定显著性水平α，假设检验的拒绝域为

$$W = \{F_{A\times B} > F_\alpha((a-1)(b-1), ab(c-1))\}.$$

表7.2.5 关于交互效应的两因素方差分析表

方差来源	平方和	自由度	均方	F比
因素A	SS_A	$a-1$	$\mathsf{MS}_A=\frac{\mathsf{SS}_A}{a-1}$	$F_A=\frac{\mathsf{MS}_A}{\mathsf{MS}_E}$
因素B	SS_B	$b-1$	$\mathsf{MS}_B=\frac{\mathsf{SS}_B}{b-1}$	$F_B=\frac{\mathsf{MS}_B}{\mathsf{MS}_E}$
交互效应$A\times B$	$\mathsf{SS}_{A\times B}$	$(a-1)(b-1)$	$\mathsf{MS}_{A\times B}=\frac{\mathsf{SS}_{A\times B}}{(a-1)(b-1)}$	$F_{A\times B}=\frac{\mathsf{MS}_{A\times B}}{\mathsf{MS}_E}$
误差	SS_E	$ab(c-1)$	$\mathsf{MS}_E=\frac{\mathsf{SS}_E}{ab(c-1)}$	
总计	SS_T	$abc-1$		

例7.2.2 研究树种与地理位置对松树生长的影响, 对4个地区的3种同龄松树的直径进行测量得到数据如下表所示. A_1, A_2, A_3表示三个不同的树种, B_1, B_2, B_3, B_4表示4个不同的地区. 对每一种水平组合, 进行了5次测量, 对比试验结果进行方差分析.

表7.2.6 3种同龄松树的直径测量数据(单位: cm)

	B_1	B_2	B_3	B_4
A_1	23,25,21,14,15	20,17,11,26,21	16,19,13,16,24	20,21,18,27,24
A_2	28,30,19,17,22	26,24,21,25,26	19,18,19,20,25	26,26,28,29,23
A_3	18,15,23,18,10	21,25,12,12,22	19,23,22,14,13	22,13,12,22,19

R程序:

R程序:

```
tree=data.frame(
A=gl(3,20,60),
B=gl(4,5,60),
Y=c(23, 25, 21, 14, 15, 20, 17, 11, 26, 21,
16, 19, 13, 16, 24, 20, 21, 18, 27, 24,
28, 30, 19, 17, 22, 26, 24, 21, 25, 26,
19, 18, 19, 20, 25, 26, 26, 28, 29, 23,
18, 15, 23, 18, 10, 21, 25, 12, 12, 22,
19, 23, 22, 14, 13, 22, 13, 12, 22, 19)
)
tree.aov=aov(Y~A+B+A:B, data=tree)
summary(tree.aov)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
A            2  352.5  176.27   8.959 0.000494 ***
B            3   87.5   29.17   1.483 0.231077    
A:B          6   71.7   11.96   0.608 0.722890    
Residuals   48  944.4   19.68                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

```
          Df Sum Sq Mean Sq F value   Pr(>F)    
A          2  352.5  176.27   8.959 0.000494 ***
B          3   87.5   29.17   1.483 0.231077    
A:B        6   71.7   11.96   0.608 0.722890    
Residuals 48  944.4   19.68                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

可见在显著性水平$\alpha = 0.05$下, 树种和地理位置的交互效应不显著, 树种(因素A)的效应是有显著差异的, 而地理位置(因素B)的效应无显著差异.

来看一下多重假设检验(跟前面一样, 需要对p-value进行调整).

pairwise.t.test(tree$Y,tree$A,p.adjust.method=“fdr”)

```
> pairwise.t.test(tree$Y,tree$A,p.adjust.method="fdr")

        Pairwise comparisons using t tests with pooled SD

data:  tree$Y and tree$A

  1       2
2 0.00852 -
3 0.20103 0.00032
```