课程主页：http://staff.ustc.edu.cn/~ynyang/2022

第十五讲 方差分析

2022.12.9

线性性/可加性：一个变量的效应与其它变量的具体取值无关

因子变量/因素

因子/因素变量(factor)

取值为类别的变量称为因子、因素、属性变量、类别变量，其各个取值称为水平(level)，比如变量 Gender：取值为男、女，2个水平；Dose：取值高、中、低，3个水平（三个水平有次序，ordinal）。

单因素方差分析

根据一个因素（分类指标）将研究对象(随机)分为若干组。

例如：根据剂量分为高、中、低三组。因素：剂量，3个水平，如下表。

剂量	高	中	低

两因素方差分析

根据两个因素交叉分类，将研究对象随机分成双向表格(two-way)。

例如：两个药物*A*, *B*(两个因素)的用药剂量，称作因素*A*和因素*B*,

各有三个水平(高、中、低), 交叉分类, 共9组，如下表。

A \ B	高	中	低
高			
中			
低			

单因素方差分析 (one-way anova)

ANOVA：Analysis of variance

方差分析检验多个正态总体均值相等，是试验设计的重要分析方法。方差分析是线性回归模型中所有自变量为因子变量的特殊情形，虽然简单，但对于理解一般的回归分析思想也有帮助。

单因素方差分析模型：

随机化控制试验，研究对象被随机分配接受K种处理(*treatment*)中的一种。

假设第k组的响应服从（方差相同但均值可能不同的）正态分布：

$$y_{k1},...,y_{kn_k} \text{ iid } \sim N(\mu_k,\sigma^2),\ k=1,...,K,\ n=n_1+...+n_K$$

零假设 $H_0:\ \mu_1=\mu_2=...=\mu_K$

这是多正态总体均值相等性检验，称为单因素方差分析模型。

因为随机化，在处理无效时所有响应的分布相同，在处理有效的情况下，只有均值可能有差异，N和σ^2都相同。

处理	1	…	K
响应	$y_{11}, y_{12}, \cdots, y_{1n_1}$	…	$y_{K1}, y_{K2}, \cdots, y_{Kn_K}$
平均	$\bar{y}_{1\bullet}$	…	$\bar{y}_{K\bullet}$

$$\bar{y}_{k\bullet} = \sum_{i=1}^{n_k} y_{ki} / n_k,$$

$$\bar{y}_{\bullet\bullet} = \left(\sum_{k=1}^{K} \sum_{i=1}^{n_k} y_{ki} \right) / n,$$

$H_0: \mu_1 = \cdots = \mu_K$的检验统计量应该基于$\bar{y}_{1\bullet}, \ldots, \bar{y}_{K\bullet}$之间的差异，如何度量$K$个数的差异？

我们可以两两比较并求平方和 $\sum_{i,j} (\bar{y}_{i\bullet} - \bar{y}_{j\bullet})^2$

也可以将每个$\bar{y}_{i\bullet}$与总平均对比并求平方和 $\sum_i (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$，事实上更合理的可能是加权平方和$SS_{\mathrm{B}} = \sum_i n_i (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$（因为第$i$组有$n_i$个样本，$\bar{y}_{i\bullet}$的方差与$1/n_i$成正比），称为组间平方和。

线性模型形式

单因素方差法分析模型（双下标形式）：

$$y_{ki} = \mu_k + \varepsilon_{ki},\quad \varepsilon_{ki}\ \ iid \sim N(0,\sigma^2),\ i = 1,...,n_k; k = 1,..,K.$$

写成通常的线性模型形式：

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{g}_1\mu_1 + ... + \mathbf{g}_K\mu_K + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(\mathbf{0},\sigma^2 I_n),$$

其中$\mathbf{g}_1,...,\mathbf{g}_K$为各组的示性表示，相互正交，$X = (\mathbf{g}_1,...,\mathbf{g}_K)$. 具体如下：

$$\begin{matrix}\text{第一组}\\ \\ \text{第二组}\\ \\ \\ \text{第}K\text{组}\end{matrix}\begin{pmatrix} y_{11}\\ \vdots\\ y_{1n_1}\\ y_{21}\\ \vdots\\ y_{2n_2}\\ \vdots\\ y_{K1}\\ \vdots\\ y_{Kn_K}\end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0\\ \vdots & \vdots & \dots & 0\\ 1 & 0 & \dots & 0\\ 0 & 1 & \dots & 0\\ \vdots & \vdots & & \\ 0 & 1 & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & 1\\ \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & 0 & 1\end{pmatrix}\begin{pmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_K\end{pmatrix} + \begin{pmatrix}\varepsilon_{11}\\ \vdots\\ \varepsilon_{1n_1}\\ \varepsilon_{21}\\ \vdots\\ \varepsilon_{2n_2}\\ \vdots\\ \varepsilon_{K1}\\ \vdots\\ \varepsilon_{Kn_K}\end{pmatrix}$$

$$= \begin{pmatrix}\mathbf{1}_{n_1} & \mathbf{0} & \dots & \mathbf{0}\\ \mathbf{0} & \mathbf{1}_{n_2} & \dots & \mathbf{0}\\ \vdots & \vdots & \dots & \vdots\\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{1}_{n_K}\end{pmatrix}\begin{pmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_K\end{pmatrix} + \boldsymbol{\varepsilon},$$

$$(\mathbf{g}_1,\ \mathbf{g}_2,\cdots,\mathbf{g}_K)$$

注意$\mathbf{g}_1,\ \mathbf{g}_2,...,\mathbf{g}_K$是各个组的示性变量，为了在数学或计算机上处理双下标模型，我们将K-水平因子变量转换为K个示性变量。

下面求F检验(检验$q=K-1$个自由参数等于0)

$$\hat{\mathbf{y}}=P_{\mathbf{g}_1}\mathbf{y}+\ldots+P_{\mathbf{g}_K}\mathbf{y}=\mathbf{g}_1\bar{y}_{1\bullet}+\ldots+\mathbf{g}_K\bar{y}_{K\bullet}=(\bar{y}_{1\bullet},\ldots\bar{y}_{1\bullet},\bar{y}_{2\bullet},\ldots,\bar{y}_{2\bullet},\ldots,\bar{y}_{K\bullet},\ldots\bar{y}_{K\bullet})^{\mathsf{T}}$$

H_0下模型为：$\mathbf{y}=(\mathbf{g}_1+\ldots+\mathbf{g}_K)\mu_1+\boldsymbol{\varepsilon}=\mathbf{1}\mu_1+\boldsymbol{\varepsilon},\quad V_0=L(\mathbf{1}).$

$$\hat{\mathbf{y}}_0=P_{V_0}\mathbf{y}=(\bar{y}_{\bullet\bullet},\ldots\ldots,\bar{y}_{\bullet\bullet})^{\mathsf{T}}=\mathbf{1}\bar{y}_{\bullet\bullet}$$

$SS_{总}=SS_B+SS_W$

记号：

- 组间平方和：$SS_B=\|\hat{\mathbf{y}}-\hat{\mathbf{y}}_0\|^2=\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\bar{y}_{k\bullet}-\bar{y}_{\bullet\bullet})^2=\sum_{k=1}^{K}n_k(\bar{y}_{k\bullet}-\bar{y}_{\bullet\bullet})^2,$

  度量了各组的中心$\bar{y}_{k\bullet},k=1,\ldots,K$之间的差异（回归平方和）。
- 组内平方和：$SS_W=\|\mathbf{y}-\hat{\mathbf{y}}\|^2=\sum_{k=1}^{K}\sum_{i=1}^{n_k}(y_{ki}-\bar{y}_{k\bullet})^2,$

  度量了组内样本的差异（残差平方和）。
- $SS_{总}=SS_B+SS_W,\quad \mathrm{B:Between\text{-}group,\ W:Within\text{-}group}$

所以$F=\dfrac{\|\hat{\mathbf{y}}-\hat{\mathbf{y}}_0\|^2/q}{\|\mathbf{y}-\hat{\mathbf{y}}\|^2/(n-p)}=\dfrac{SS_B/(K-1)}{SS_W/(n-K)}$

One-way anova的F检验

命题1. 假设模型$y_{k1},...,y_{kn_k}$ iid $\sim N(\mu_k,\sigma^2)$, $k=1,...,K$.

$H_0:\mu_1=\mu_2=...=\mu_K$的$F$检验统计量：

$$F=\frac{SS_B/(K-1)}{SS_W/(n-K)}=\frac{\sum_{k=1}^{K}n_k(\bar{y}_{k\bullet}-\bar{y}_{\bullet\bullet})^2/(K-1)}{\sum_{k=1}^{K}\sum_{j=1}^{n_k}(y_{kj}-\bar{y}_{k\bullet})^2/(n-K)} \sim_{H_0} F_{K-1,n-K}.$$

称为单因素方差分析。

方差分析R函数：aov， lm

```
> aov(响应 ~ 组别, data =... )
> lm(响应 ~ 组别, data =... )
```

*aov*函数输出方差分析表：

来源	因子/分组	残差
平方和	SS_B	SS_W
自由度	$K-1$	$n-K$

特例：两样本t检验

两样本t检验是最简单的单因素方差分析方法。

$K=2$时,方差分析的F统计量：$F=(t)^2 \sim_{H_0} F_{1,n-2}$，其中t为两样本$t$统计量

$$t=\frac{\bar{y}_{1\bullet}-\bar{y}_{2\bullet}}{\sqrt{(1/n_1+1/n_2)\hat{\sigma}^2}},$$

其中$\hat{\sigma}^2=\left(\sum_{j=1}^{n_1}(y_{1j}-\bar{y}_{1\bullet})^2+\sum_{j=1}^{n_2}(y_{2j}-\bar{y}_{2\bullet})^2\right)/(n-2)$。

验证：$\bar{y}_{\bullet\bullet}=(n_1\bar{y}_{1\bullet}+n_2\bar{y}_{2\bullet})/(n_1+n_2),\ n=n_1+n_2$,则

$$F=\frac{\left(n_1(\bar{y}_{1\bullet}-\bar{y}_{\bullet\bullet})^2+n_2(\bar{y}_{2\bullet}-\bar{y}_{\bullet\bullet})^2\right)/(2-1)}{\left(\sum_{j=1}^{n_1}(y_{1j}-\bar{y}_{1\bullet})^2+\sum_{j=1}^{n_2}(y_{2j}-\bar{y}_{2\bullet})^2\right)/(n-2)}=\frac{n_1n_2(\bar{y}_{1\bullet}-\bar{y}_{2\bullet})^2}{n\hat{\sigma}^2}=\left(\frac{\bar{y}_{1\bullet}-\bar{y}_{2\bullet}}{\sqrt{(1/n_1+1/n_2)\hat{\sigma}^2}}\right)^2$$

附：其它(单因素)多总体比较问题

非参数统计：两总体Wilcoxn秩和检验，多总体Kruskal - Wallis检验。

属性数据分析：I×J列联表的Pearson卡方检验。

生存分析：log－rank检验。

例1 (npk数据). 随机在6块(block)地上种植豌豆，每块地上4株。检验各块地上的产量是否相同. 我们需要检验6个地块的产量是否相同。 $H_0 : \mu_1 = ... = \mu_6$

block	1	2	3	4	5	6
产量	49.5,62.8, 46.8,57	59.8,58.5, 55.5,56	...		...	57.2,59, 53.2,56

```
> a  = aov(yield ~ block, data = npk)
> summary( a )
            Df Sum Sq Mean Sq  F value  Pr(>F)
block        5  343.3   68.66    2.318  0.0861 .
Residuals   18  533.1   29.61
```

对比(contrast)：重新参数化

单因素方差分析模型中

$$y_{ki} \sim N(\mu_k, \sigma^2) \iff y_{ki} = \mu_k + \varepsilon_{ki}, \quad \varepsilon_{ki}\ iid \sim N(0, \sigma^2) \qquad (1)$$

零假设 $H_0: \mu_1 = \cdots = \mu_K$是一个复合原假设。为了比较$K$个$\mu$，可重新定义新的参数表达具体的感兴趣的问题，称为对比(contrast)。

Treatment Contrast

假设第一组（处理1 ）是安慰剂组（基准、baseline），其它组使用各种不同药物（或同一药物的各种剂量），那么我们关心的是各种药物或各种剂量相对于安慰剂的作用，重新定义参数

$$\alpha_2 = \mu_2 - \mu_1, \ldots \alpha_K = \mu_K - \mu_1$$

称为各处理相对于基准的效应(effect)或处理对比(treatment contrast).

$H_0: \mu_1 = \cdots = \mu_K \Leftrightarrow H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_K = 0$, 这是一个简单原假设。

为了统一, 不妨记$\alpha_1 \equiv 0$, 则$\mu_k = \mu_1 + \alpha_k$, $k = 1,2, \ldots, K$. 模型(1)改写为

$$y_{ki} = \mu_1 + \alpha_k + \varepsilon_{ki}, \quad \varepsilon_{ki} \sim N(0, \sigma^2)，\ 约束 \alpha_1 \equiv 0$$

Sum Contrast

如果没有一个公认的基准组，各个组对称，那么我们可以将每个处理与某个中心进行比较，比如取中心为 $\bar{\mu} = (\mu_1 + \cdots + \mu_K)/K$，定义$K$个sum对比：

$$\alpha_1 = \mu_1 - \bar{\mu}, \ldots, \alpha_K = \mu_K - \bar{\mu},$$

注意有约束$\alpha_1 + \cdots + \alpha_K = 0$。原假设等价于$H_0: \alpha_1 = \cdots = \alpha_K = 0$。

则单因素方差分析模型改写为

$$y_{ki} = \bar{\mu} + \alpha_k + \varepsilon_{ki}, \quad \varepsilon_{ki} \sim N(0, \sigma^2)，\text{约束}\alpha_1 + \cdots + \alpha_K = 0$$

一般的对比

若$c_1 + \cdots + c_K = 0$, 称$c_1\mu_1 + \cdots + c_K\mu_K$是一个对比$(contrast)$。

对比的定义方式取决于感兴趣的具体问题. 比如trt, sum contrast, 再如，如果K个组代表K个递增的剂量，假设我们关心相邻两种剂量的差异，则可定义

$$\alpha_k = \mu_{k+1} - \mu_k, k = 1,2, \ldots, K-1$$

无论何种约束，模型(1)都可改写为

$$y_{ki} = \mu + \alpha_k + \varepsilon_{ki}, \quad \varepsilon_{ki} \sim N(0, \sigma^2) \qquad (2)$$

其中各组效应(对比)$\alpha_1, \ldots, \alpha_K$满足一个线性约束。需要注意的是

- H_0的F检验不受具体对比方式的影响，都如命题1所述。

  引入对比概念的目的，一是具体问题的需要，二是方便对研究多因素方差分析。

- 双下标表示简明扼要，既表示了样本标号 i, 也表示了其所在的组号k。
- 但为了在数学上或计算机上处理，我们需要使用示性变量/哑变量表示分组（因子变量），如下页所示。

多下标的记录方式$(a_{ij\cdots k})$称为数组array或张量tensor

对任何一个响应y,其所在组的标号记作G,方差分析模型写成线性模型：

$$y = \mu+\alpha_1 1_{(G=1)} + \cdots + \alpha_K 1_{(G=K)} + \varepsilon \qquad (3)$$

因为有个截距项,诸α有一个线性约束，所以实际上只需要$K-1$个哑变量，比如trt contrast下α_1=0，第一组的哑变量$1_{(G=1)}$多余。

方差分析模型(3)的数据形式：

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{g1} \\ \vdots \\ y_{gn_g} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ \vdots & & & & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & & & & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & & & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & & & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_K \end{pmatrix} + \boldsymbol{\varepsilon}$$

如果是trt contrast约束，即$\alpha_1 = 0$，
实际计算时，删除α_1及设计阵第二列。

因子的哑变量表示

哑变量表示 (dummy coding):

对于一个K水平因子变量，取定其中一类为基准，用其余$K-1$个类的示性函数表示K个类。

例如, 因子变量group取值g1,g2,g3(三个组),以g1为基准,定义

$$x_2 = 1_{(\text{group=g2})}, x_3 = 1_{(\text{group=g3})}$$

则二值变量x_2和x_3完全确定了group:

$$\begin{aligned}
&\text{group} = \text{g1} \quad \Leftrightarrow (x_2, x_3) = (0,0);\\
&\text{group} = \text{g2} \quad \Leftrightarrow (x_2, x_3) = (1,0);\\
&\text{group} = \text{g3} \quad \Leftrightarrow (x_2, x_3) = (0,1).
\end{aligned}$$

group	x_2	x_3
g1	0	0
g2	1	0
g3	0	1

aov($y \sim$ group)，或lm($y \sim$ group)实际上拟合模型：

$$y = a + bx_2 + cx_3 + \varepsilon, \quad x_k = 1_{(\text{group}=k)}$$

其中$\mu_1 = a, \mu_2 = a+b, \mu_3 = a+c \Leftrightarrow a = \mu_1, b = \mu_2 - \mu_1, c = \mu_3 - \mu_1$,

b, c分别称为第2,3组的效应，它们是处理对比(treatment contrast)。

配对t检验

配对设计

配对设计(paired design)：两个相像的研究对象（比如双胞胎），随机选取其中一人服用药物（处理），另一个服用安慰剂（对照）。

第i对样本的响应：(y_{1i}, y_{2i})，下标2表示处理,1表示对照。

$z_i = y_{2i} - y_{1i}$，药效：$\delta = E(z_i)$，$H_0 : \delta = 0$

pair	1	2	…	J	平均
对照	y_{11}	y_{12}	…	y_{1J}	$\bar{y}_{1\bullet}$
处理	y_{21}	y_{22}	…	y_{2J}	$\bar{y}_{2\bullet}$
差	z_1	z_2	…	z_J	$\bar{z}$

同一个pair内的两个个体相似(匹配、相依)，将它们作为独立样本进行两样本t检验是错误的做法，如何进行检验？

基于每对内的差 $z_j = y_{2j} - y_{1j}, j = 1, \ldots, J$ 构造检验

成对t检验 (paired t-test/dependent t-test)

配对样本(y_{1i}, y_{2i})，$i = 1,...,n$ iid，$z_i = y_{2i} - y_{1i}$，$\delta = E(z_i)$，$H_0 : \delta = 0$

假设$z_1,....,z_n$ iid $\sim N(\delta, \sigma^2)$，成对t检验等于基于z_i的单正态总体均值检验：

$$t_{\text{pair}} = \frac{\bar{z}}{s_z/\sqrt{n}} = \frac{(\bar{y}_{2\bullet} - \bar{y}_{1\bullet})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_{2i} - y_{1i} - (\bar{y}_{2\bullet} - \bar{y}_{1\bullet}))^2/n}} \overset{H_0}{\sim} t_{n-1}$$

该检验称为成对 / 配对 t - 检验.

类比两样本t检验（比较处理组和对照组）：

$$t_{\text{two-sample}} = \frac{\bar{y}_{2\bullet} - \bar{y}_{1\bullet}}{\sqrt{(1/n + 1/n)\left(\frac{1}{2n-2}\sum_{j=1}^{n}(y_{1j} - \bar{y}_{1\bullet})^2 + \sum_{j=1}^{n}(y_{2j} - \bar{y}_{2\bullet})^2\right)}}$$

t_{pair}与$t_{\text{two-sample}}$的分子相同，方差(分母)计算方式不同。

事实上，t_{pair}可看作是如下方式得到的：

假设同一对的y_{1i}, y_{2i}共享一个共同的参数μ_i(也可假设为随机变量)

用于刻画第i对两个研究对象的共性，比如μ_i较大/小时，y_{1i}，y_{2i}都较大/小。

$$\begin{cases} y_{1i} = \mu_i + \varepsilon_{1i} \\ y_{2i} = \mu_i + \delta + \varepsilon_{2i} \end{cases}, \quad \varepsilon_{1i}.\varepsilon_{2i}, i = 1,...,n \ \text{iid} \sim N(0,\sigma^2)$$

令$z_i = y_{2i} - y_{1i}$, 则

$$z_1,....,z_n \ \text{iid} \sim N(\delta, 2\sigma^2),$$

消去了μ_i，因此将成对比较问题转化为单正态总体的均值检验$H_0 : \delta = 0,$

这正是上页的成对t-检验的做法。

例2. McNemar检验：二值响应情形下的配对检验。

当响应变量y_{ij}为0,1变量（比如1表示治愈，否则为0),响应变量不可能服从正态分布,但我们依然可用成对t检验来检验处理的效应，只不过此时检验统计量在原假设下近似服从卡方分布。

pair	1	2	…	*J*
对照	y_{11}	y_{12}	…	y_{1J}
处理	y_{21}	y_{22}	…	y_{2J}

$$z_j = y_{2j} - y_{1j} = \begin{cases} 1 & y_{2j}=1, y_{1j}=0 \\ -1 & y_{2j}=0, y_{1j}=1 \\ 0 & y_{2j}=y_{1j} \end{cases}$$

原假设下$\bar{z} \to 0$, $\sqrt{J}\bar{z} \to N(0,2\sigma^2)$,由Slusky引理，知$J \to \infty$时

$$t_{\text{pair}} = \sqrt{J}\bar{z} / \sqrt{(\sum_{j=1}^{J} z_j^{\,2} - J\bar{z}^2)/(J-1)} \approx \sqrt{J}\bar{z} / \sqrt{\sum_{j=1}^{J} z_j^{\,2}/J} = \sum_{j=1}^{J} z_j / \sqrt{\sum_{j=1}^{J} z_j^2} \xrightarrow{d} N(0,1),$$

记$b = \#\{j : y_{2j}=1, y_{1j}=0\}$, $c = \#\{j : y_{2j}=0, y_{1j}=1\}$分别为处理和对照响应不同的pair的个数，则$\sum_{j=1}^{J} z_j = b-c, \sum_{j=1}^{J} z_j^2 = b+c$，$t_{\text{pair}} \approx (b-c)/\sqrt{b+c}$

配对二值(binary)数据可总结为下表，a为两人响应都为0 的pair的个数，b为处理响应为1而对照响应为0的pair的个数，等等。

		处理 0	处理 1
对照	0	a	b
	1	c	d

$$t_{pair} \approx \frac{\sum_{j=1}^{J} z_j}{\sqrt{\sum_{j=1}^{J} z_j^2}} = \frac{(b-c)}{\sqrt{b+c}},$$

McNemar 检验：H_0下，$X^2 = \frac{(b-c)^2}{b+c} \sim_{\text{近似}} \chi_1^2$.

- X^2中没有a, d : 两个研究对象响应都是0或者都是1的pair不提供处理效应的信息。
- McNemar检验是Cochran - Mantel - Haenszel(CMH)检验的特殊情况,

CMH检验是综合多个列联表数据的独立性检验，与两因素方差分析方法类似。

两因素方差分析

两因素方差分析中，决定分组的因素/因子有两个，可以都是处理（比如各种压力和各种温度组合下的工业试验），但通常一个因素是处理，另一个因素是区组，即区组设计。

随机区组设计 (Randomized block design)

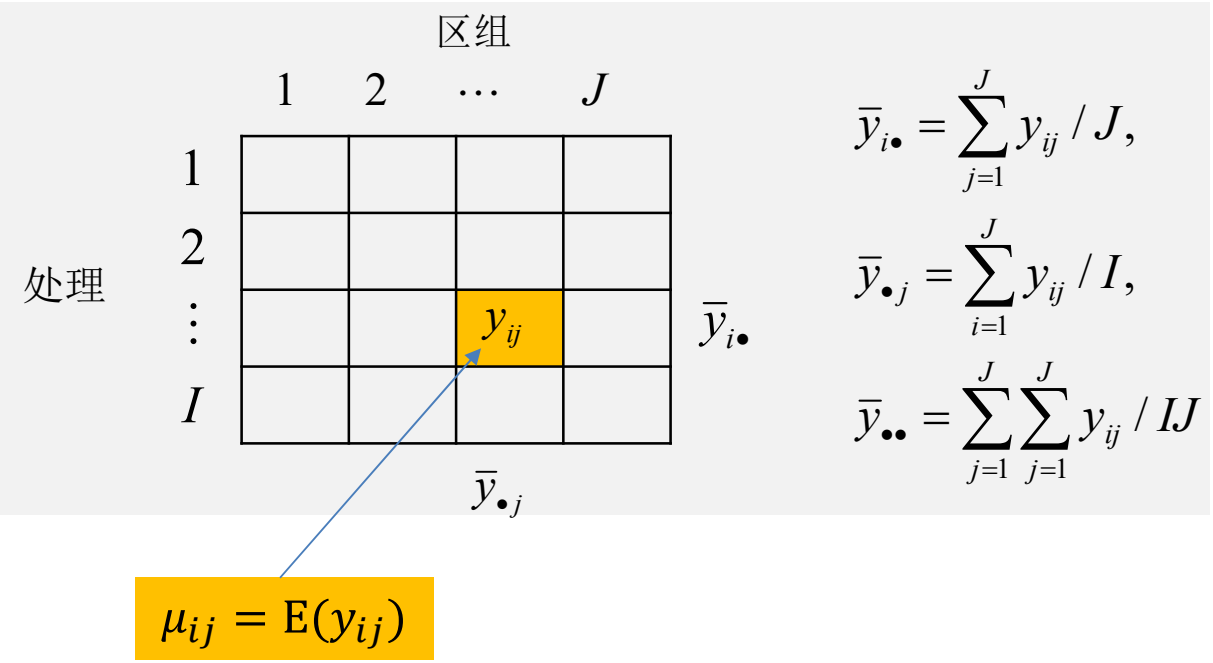
随机化试验设计中，当研究对象差异较大时，为了避免处理组与对照组出现严重的不均衡，通常先将研究对象分成内部个体相似的区组或层(blocks/strata), 在区组内随机化分配处理。

block：Fisher田间试验的地块；
strata：生物、社会学常用

配对设计是最简单的区组设计，每一个pair就是一个区组。区组内的两个研究对象相似， 但接受不同的处理。

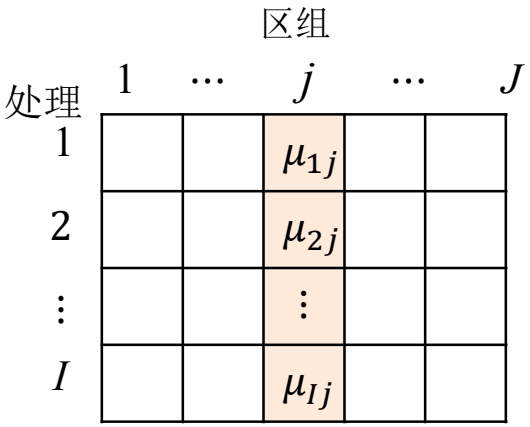
假设两个因子各有I,J个水平，交叉分类得到$I \times J$表，格子(i,j)内的观测数目为n_{ij}。

我们主要以每个格子内只有一次测量$(n_{ij}=1)$的情形为例（这也是最常见或最重要的情形），介绍两因素可加方差分析方法。假设$y_{ij} \sim N(\mu_{ij}, \sigma^2)$, 所有$y_{ij}$独立。



我们希望检验所有区组 内，不同的处理的效果都相同：

$$H_0: \mu_{1j} = \mu_{2j} = \cdots = \mu_{Ij}, \qquad 1 \le j \le J$$



每个区组内只有I个样本，但有$I+1$个参数，如何检验？

答案：假设$\{\mu_{ij}\}$有简单结构/模型-假设各列之间存在共性(一致性)。

处理效应的一致性假设：
$\mu_{ij} - \mu_{i'j}$与所在区组j无关，只与处理水平i, i'有关

以$I=J=2$为例，我们说明效应一致性假设等价于可加性或线性性

	区组1	区组2
处理1	μ_{11}	μ_{12}
处理2	μ_{21}	μ_{22}

假设处理效应的一致性：
$\mu_{21}-\mu_{11}=\mu_{22}-\mu_{12}\triangleq\alpha$
行间差为常数

	区组1	区组2
处理1	μ_{11}	μ_{12}
处理2	$\mu_{11}+\alpha$	$\mu_{12}+\alpha$

α称为处理2的效应

列间差也是常数，即区组效应也一致：
$\mu_{22}-\mu_{21}=(\mu_{12}+\alpha)-(\mu_{11}+\alpha)$
$=\mu_{12}-\mu_{11}\triangleq\beta$

	区组1	区组2
处理1	μ_{11}	$\mu_{11}+\beta$
处理2	$\mu_{11}+\alpha$	$\mu_{11}+\alpha+\beta$

β称为区组2的效应

所以处理效应一致性⇔区组效应一致性⇔可加性(两个因素互不干扰)：

$\mu_{21}=\mu_{11}+\alpha$； $\mu_{12}=\mu_{11}+\beta$ $\mu_{22}=\mu_{21}+\beta=\mu_{11}+\alpha+\beta$

即当一个因素固定而另一个因素从水平1到2时，均值增加α或β。当两个因素同时从水平1到2时，均值增加$\alpha+\beta$（可加）。

处理效应的一致性⇔区组效应的一致性⇔可加模型

一般的$I \times J$情形类似。以处理1为基准(baseline), 假设处理的一致性：
处理 i $(1 \le i \le I)$ 的效应 $\mu_{ij} - \mu_{1j}$与所在区组 j 无关, 记$\alpha_i = \mu_{ij} - \mu_{1j}$，称之为处理 i 的效应。

		区组1		区组j		
处理	1	μ_{11}		μ_{1j}		
	2	$\mu_{11} + \alpha_2$		$\mu_{1j} + \alpha_2$		
	⋮	⋮		⋮		
	I	$\mu_{11} + \alpha_I$		$\mu_{1j} + \alpha_I$		

对任何一行i（处理i），区组1和区组j的均值之差，

$$\mu_{ij} - \mu_{i1} = (\mu_{1j} + \alpha_i) - (\mu_{11} + \alpha_i) = \mu_{1j} - \mu_{11} \triangleq \beta_j,$$

与i无关（即区组效应也有一致性），记之为β_j, 称为区组j的效应，所以

$$\mu_{ij} = \mu_{i1} + \beta_j = \mu_{11} + \alpha_i + \beta_j$$

这称为可加效应模型（处理和区组效应可加）。因此效应一致等价于可加模型。

两因素可加方差分析模型

假设处理i,区组j的响应$y_{ij} \sim N(\mu_{ij}, \sigma^2)$，其中

$$\mu_{ij} = \mu + \alpha_i + \beta_j \quad (\alpha's和\beta's各有一个约束)$$

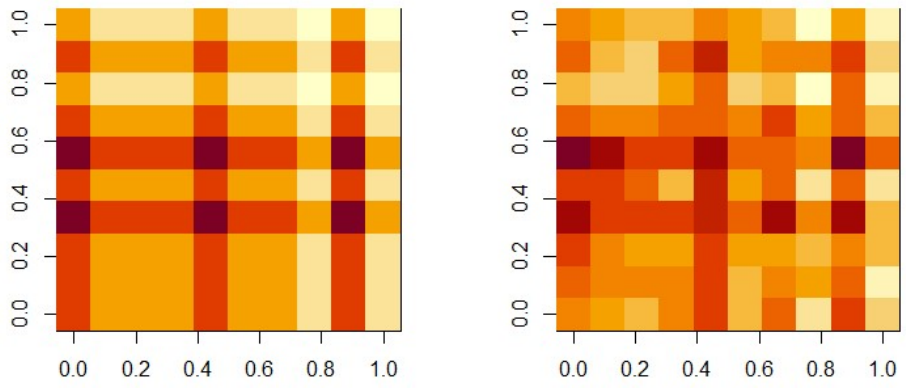
称为两因素可加方差分析模型 ($n_{ij} \equiv 1$情形)。

零假设为$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_I = 0$。

两因素可加模型是方差 分析中最重要的模型和 方法.

左图：可加模型 $\mu_{ij} = \mu + \alpha_i + \beta_j$，每两行之差为常数， 每两列也是。

右图偏离可加，交互模型 $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$.



可加方差分析模型$y_{ij}=\mu+\alpha_i+\beta_j+\varepsilon_{ij}$是线性模型的特殊形式，

例如$I=J=2$,时，模型的矩阵-向量形式：

$$\begin{pmatrix} y_{11} \\ y_{21} \\ y_{12} \\ y_{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \boldsymbol{\varepsilon}$$

约束：

treatment contrast : $\alpha_1=0$， $\beta_1=0$

sum contrast : $\alpha_1+\alpha_2=0$， $\beta_1+\beta_2=0$

约束LS估计

可加模型下，参数个数为：$I+J-1$（$I-1$个α,$J-1$个β，一个σ^2）$<n=IJ$。

参数可估。效应的LS估计依赖于具体的效应定义，我们仅演示sum contrast约束情形（此时求解最简单），假设约束$\Sigma\alpha_i=\Sigma\beta_j=0$,

对误差平方和 $f=\sum_{i,j}(y_{ij}-\mu-\alpha_i-\beta_j)^2$，求导得正则方程：

$$\begin{cases}\partial f/\partial\mu = \sum_i\sum_j(y_{ij}-\mu-\alpha_i-\beta_j)=0,\ IJ\mu=\sum_i\sum_j y_{ij}=IJ\bar{y}_{\bullet\bullet}\Rightarrow\hat{\mu}=\bar{y}_{\bullet\bullet}\\ \partial f/\partial\beta_j = \sum_i(y_{ij}-\mu-\alpha_i-\beta_j)=0,\ I\bar{y}_{\bullet j}-I\mu-I\beta_j=0\Rightarrow\hat{\beta}_j=\bar{y}_{\bullet j}-\bar{y}_{\bullet\bullet}\\ \partial f/\partial\alpha_i = \sum_j(y_{ij}-\mu-\alpha_i-\beta_j)=0,\ J\bar{y}_{i\bullet}-J\mu-J\alpha_i=0\Rightarrow\hat{\alpha}_i=\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet}\end{cases}$$

$\Rightarrow$ 全模型下拟合值：$\hat{y}_{ij}=\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j=\bar{y}_{\bullet\bullet}+(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})+(\bar{y}_{\bullet j}-\bar{y}_{\bullet\bullet})$

零假设下，各处理(各行)无差异，$\tilde{\mu}=\bar{y}_{\bullet\bullet},\tilde{\beta}_j=\bar{y}_{\bullet j}-\bar{y}_{\bullet\bullet}$.，拟合值：

$$\tilde{y}_{ij}=\tilde{\mu}+\tilde{\beta}_j=\bar{y}_{\bullet\bullet}+(\bar{y}_{\bullet j}-\bar{y}_{\bullet\bullet})$$

平方和分解

$$\begin{aligned}SS_{总}&=\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\bar{y}_{\bullet\bullet})^2=\sum_{i=1}^{I}\sum_{j=1}^{J}\left((y_{ij}-\hat{y}_{ij})+(\hat{y}_{ij}-\tilde{y}_{ij})+(\tilde{y}_{ij}-\bar{y}_{\bullet\bullet})\right)^2\\ &=\sum_{i=1}^{I}\sum_{j=1}^{J}\left((y_{ij}-\bar{y}_{i\bullet}-\bar{y}_{\bullet j}+\bar{y}_{\bullet\bullet})+(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})+(\bar{y}_{\bullet j}-\bar{y}_{\bullet\bullet})\right)^2\\ &=\sum_{i,j}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2+\sum_{i,j}(\bar{y}_{\bullet j}-\bar{y}_{\bullet\bullet})^2+\sum_{i,j}(y_{ij}-\bar{y}_{i\bullet}-\bar{y}_{\bullet j}+\bar{y}_{\bullet\bullet})^2\overset{\Delta}{=}SS_{处理}+SS_{区组}+SS_{组内}\end{aligned}$$

- $SS_{处理}=\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2=J\sum_{i=1}^{I}(\alpha_i-\bar{\alpha}+\bar{\varepsilon}_{i\bullet}-\bar{\varepsilon}_{\bullet\bullet})^2$,处理的差异, 消除了区组效应
- $SS_{组内}=\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\bar{y}_{i\bullet}-\bar{y}_{\bullet j}+\bar{y}_{\bullet\bullet})^2$， 残差平方和,消除了行、列效应

两因素方差分析F检验

命题2. I个处理水平,J个区组, 假设(i,j)水平下的测量y_{ij}满足可加模型

$$y_{ij}=\mu+\alpha_i+\beta_j+\varepsilon_{ij},\ \varepsilon_{ij}\ \text{iid}\sim N(0,\sigma^2),\ （\alpha's和\beta's各有一个约束）$$

$H_0:\alpha_1=\ldots=\alpha_I=0$ 的F检验为：

$$F=\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2/(J-1)}{\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\bar{y}_{i\bullet}-\bar{y}_{\bullet j}+\bar{y}_{\bullet\bullet})^2/(I-1)(J-1)}\overset{H_0}{\sim}F_{I-1,(I-1)(J-1)}$$

我们只需要验证自由度，$F=\dfrac{SS_{处理}/q}{SS_{组内}/(n-p)}$,

$q=$待检验的自由参数的个数$=I-1$（注意$\alpha's$有一个约束），

$p=$效应(回归系数)个数$=(I-1)+(J-1)+1=I+J-1$

$n=$样本量$=IJ,\quad n-p=IJ-I-J+1=(I-1)(J-1).$

配对设计：
$I=2, n_{ij}\equiv 1$

当$I=2, n_{ij}=1$时，F检验为：$F=J\bar{z}^2/s_z^2 \sim F_{1,J-1}$,

其中 $t_{pair}=\sqrt{J}\bar{z}/s_z \sim t_{J-1}$ 为配对t检验统计量,

所以配对t检验是两因素方差分析的特殊情况.

验证：注意$\bar{y}_{\bullet\bullet}=(\bar{y}_{1\bullet}+\bar{y}_{2\bullet})/2 \Rightarrow \bar{y}_{1\bullet}-\bar{y}_{\bullet\bullet}=(\bar{y}_{1\bullet}-\bar{y}_{2\bullet})/2=-\bar{y}_{2\bullet}+\bar{y}_{\bullet\bullet}$

$SS_{处理}=J(\bar{y}_{1\bullet}-\bar{y}_{\bullet\bullet})^2+J(\bar{y}_{2\bullet}-\bar{y}_{\bullet\bullet})^2=J(\bar{y}_{1\bullet}-\bar{y}_{2\bullet})^2/2=J\bar{z}^2/2,$

其中$z_j=y_{2j}-y_{1j}\sim N(\alpha,2\sigma^2),\ \bar{z}=\bar{y}_{2\bullet}-\bar{y}_{1\bullet}$,其余略。

例3 ($I = J = 2$)

α：处理效应，β：区组效应

$$\begin{cases} y_{11} = \mu + \varepsilon_{11} \\ y_{21} = \mu + \alpha + \varepsilon_{21} \\ y_{12} = \mu + \beta + \varepsilon_{12} \\ y_{22} = \mu + \beta + \alpha + \varepsilon_{22} \end{cases}, \quad \varepsilon_{ij} \ iid \sim N(0, \sigma^2)$$

		pair 1	pair 2
对照	1	y_{11}	y_{12}
处理	2	y_{21}	y_{22}

令$z_j = y_{2j} - y_{1j}, j = 1,2$

$$H_0 : \alpha = 0, \ F = \left(\frac{y_{21} - y_{11} + y_{22} - y_{12}}{y_{21} - y_{11} - y_{22} + y_{12}} \right)^2 = \left(\frac{z_1 + z_2}{z_1 - z_2} \right)^2 \sim_{H_0} F_{1,1}$$

$$t = \frac{z_1 + z_2}{z_1 - z_2} \sim_{H_0} t_1 ,$$

交互作用

仍以2×2情况为例，考察可加性或效应不一致情形下的模型。以处理1和区组1为基准。假设如果处理效应不一致，即$\mu_{22}-\mu_{12} \neq \mu_{21}-\mu_{11}$.

	区组1	区组2
处理1	μ_{11}	μ_{12}
处理2	μ_{21}	μ_{22}

在基准区组1，假设处理效应为$\alpha = \mu_{21} - \mu_{11}$

在基准处理组1，假设区组效应为$\beta = \mu_{12} - \mu_{11}$

	区组1	区组2
处理1	μ_{11}	$\mu_{11} + \beta$
处理2	$\mu_{11} + \alpha$	μ_{22}

记效应之差$\gamma = \gamma_{22} = (\mu_{22}-\mu_{12}) - (\mu_{21} - \mu_{11}) = \mu_{22} - \alpha - \beta - \mu_{11}$
则$\mu_{22} = \mu_{11} + \alpha + \beta + \gamma$。

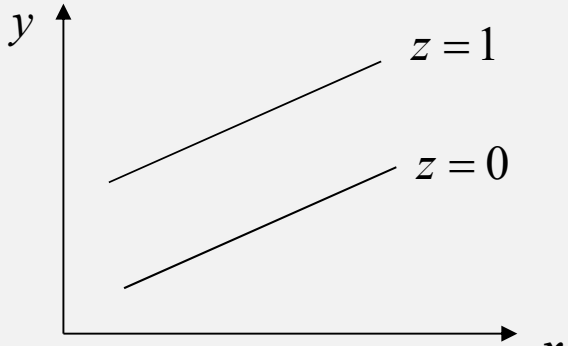
注意$\gamma = 0$即$\mu_{22} = \mu_{11} + \alpha + \beta$时是可加模型，所以$\gamma$代表了“当两个因素都从水平1变到2的时候，$\mu_{22}$偏离可加的程度”， γ称为交互作用。

交互作用：一个因素的效应与另一个因素的具体水平有关

线性模型⇔效应一致性⇔可加性

线性模型$y=a+bx+cz+\varepsilon$意味着自变量效应的可加性：

不论自变量z取值如何，y与x的关系(回归系数)都是常数b.



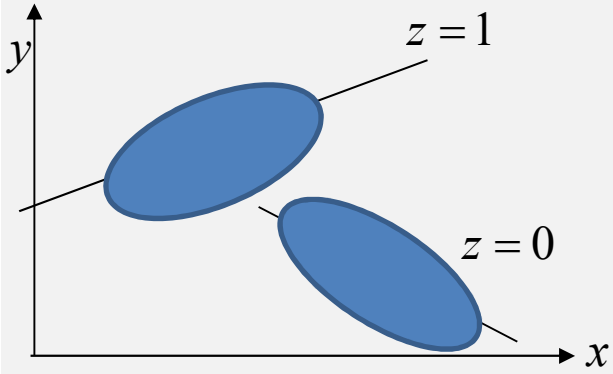
可加模型$y=a+bx+cz+\varepsilon$

x的效应与z无关：

$z=0$时，$y=a+\underline{\underline{bx}}+\varepsilon$

$z=1$时，$y=(a+c)+\underline{\underline{bx}}+\varepsilon$

交互作用：偏离可加性



交互模型$y=a+bx+cz+\gamma xz+\varepsilon$

交互作用：x的效应随z变化：

$z=0$时，$y=a+\underline{\underline{bx}}+\varepsilon$

$z=1$时，$y=(a+c)+\underline{\underline{(b+\gamma)x}}+\varepsilon$

附1: 平衡设计下的两因素可加方差分析 $(n_{ij} \equiv K)$

类似地，对于一般的平衡设计$n_{ij} \equiv K$，有类似的结果

（对于n_{ij}不全相同的情形，结果也类似，在此不再赘述）

ANOVA F 检验

命题3: （两因素方差分析,可加情形,平衡设计： $n_{ij} \equiv K$）

原假设$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_I = 0$的$F$检验统计量为

$$F = \frac{SS_{处理}/(I-1)}{SS_{组内}/(n-I-J+1)} \overset{H_0}{\sim} F_{I-1,n-I-J+1}, \quad n = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} = IJK$$

其中$SS_{处理} = \sum_{i,j,k}(\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2, SS_{组内} = \sum_{i,j,k}(y_{ijk} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})^2$。

其中我们有方差分解公式：

$$SS_{总} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \bar{y}_{\bullet\bullet\bullet})^2 \overset{\Delta}{=} SS_{处理} + SS_{区组} + SS_{组内}$$

$$= \sum_{i,j,k}(\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 + \sum_{i,j,k}(\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 + \sum_{i,j,k}(y_{ijk} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})^2$$

附2. 交互作用模型

假设两个因素(x,z)都是0-1水平,组合(i,j)格子的均值$\mu_{ij}, i,j=0,1$

饱和saturated模型/完全模型:
不对均值做任何结构假设
(4个参数: $\mu_{00},\mu_{10},\mu_{01},\mu_{11}$).
$H_0: \mu_{00}=\mu_{10}, \mu_{01}=\mu_{11}$.

x \ z	0	1
0	μ_{00}	μ_{01}
1	μ_{10}	μ_{11}

重新参数化(treatment contrast):

$\mu=\mu_{00}$, $\alpha_1=\mu_{10}-\mu_{00}$, $\beta_1=\mu_{01}-\mu_{00}$

$\gamma_{11}=(\mu_{11}-\mu_{01})-(\mu_{10}-\mu_{00})$

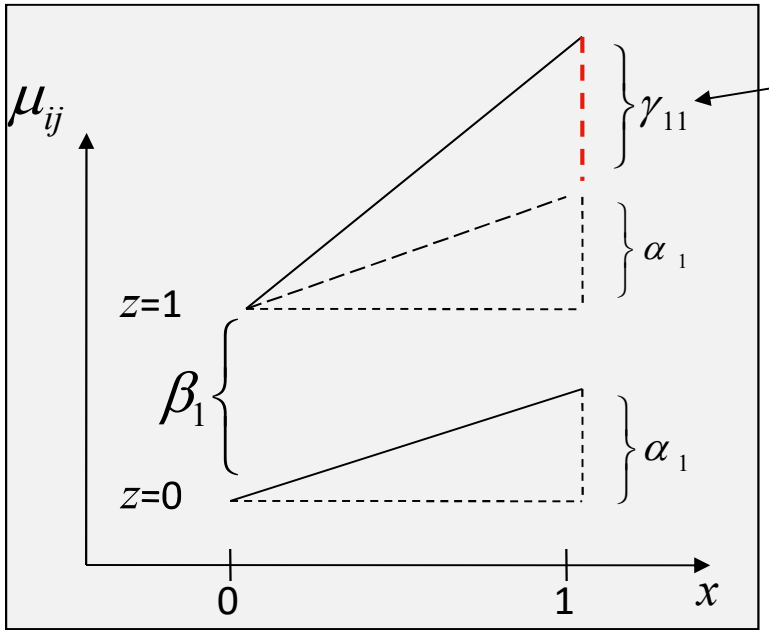
$=(\mu_{11}-\mu_{10})-(\mu_{01}-\mu_{00})$

	0	1
0	μ	$\mu+\beta_1$
1	$\mu+\alpha_1$	$\mu+\alpha_1+\beta_1+\gamma_{11}$

所以饱和模型可重写为交互作用模型：

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

其中我们记 $\alpha_0 = 0,\ \beta_0 = 0,\ \gamma_{i0} = \gamma_{0j} = 0$。



交互效应

$$\gamma_{11} = (\mu_{11} - \mu_{01}) - (\mu_{10} - \mu_{00}) = (\mu_{11} - \mu_{10}) - (\mu_{01} - \mu_{00})$$

$\gamma_{11} = 0 \Leftrightarrow$ 可加

两因素交互作用模型(一般情形)

设两个因子各I,J水平,假设饱和模型/交互作用模型($n_{ij} \equiv K > 1$):

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

其中参数α, β, γ个数分别是$I-1, J-1, (I-1)(J-1)$个,满足$I+J+1$个约束。

treatment effect contrast 满足约束：$\alpha_1 = 0, \beta_1 = 0, \gamma_{1j} = \gamma_{i1} = 0,$

sum contrast 满足约束：$\sum_{i=1}^{I} \alpha_i = 0, \sum_{j=1}^{J} \beta_j = 0, \sum_{i=1}^{I} \gamma_{ij} = 0, \sum_{j=1}^{J} \gamma_{ij} = 0.$

sum contrast $\Leftrightarrow \{\mu_{ij}, i, j\}$重写参数化为

$$\mu = \sum \mu_{ij} / IJ, \alpha_i = \bar{\mu}_{i\bullet} - \mu = \sum_j \mu_{ij} / J - \mu,$$

$$\beta_j = \bar{\mu}_{\bullet j} - \mu = \sum_i \mu_{ij} / I - \mu, \gamma_{ij} = \mu_{ij} - \alpha_i - \beta_j + \mu.$$

LS估计

对误差平方和 $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk}-\mu-\alpha_i-\beta_j-\gamma_{ij})^2$求导得正则方程：

$$\begin{cases} IJK\mu+JK\sum_{i=1}^{I}\alpha_k+IK\sum_{j=1}^{J}\beta_j+K\sum_{i=1}^{I}\sum_{j=1}^{J}\gamma_{ij}=y_{\bullet\bullet\bullet}, \\ JK\mu+JK\alpha_k+K\sum_{j=1}^{J}\beta_j+K\sum_{j=1}^{J}\gamma_{ij}=y_{i\bullet\bullet}, \\ IK\mu+K\sum_{i=1}^{I}\alpha_i+IK\beta_j+K\sum_{i=1}^{I}\gamma_{ij}=\bar{y}_{\bullet j\bullet}, \\ K(\mu+\alpha_i+\beta_j+\gamma_{ij})=y_{ij\bullet} \end{cases}$$

sum contrast 表示下，参数满足约束 ：

$$\sum_{i=1}^{I}\alpha_i=0,\sum_{j=1}^{J}\beta_j=0,\sum_{i=1}^{I}\gamma_{ij}=0,\sum_{j=1}^{J}\gamma_{ij}=0,$$

此时， LS估计具有简单形式 ：

$$\hat{\mu}=\bar{y}_{\bullet\bullet\bullet},\quad \hat{\alpha}_i=\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet\bullet\bullet},\quad \hat{\beta}_j=\bar{y}_{\bullet j\bullet}-\bar{y}_{\bullet\bullet\bullet},$$

$$\hat{\gamma}_{ij}=\bar{y}_{ij\bullet}-\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet j\bullet}+\bar{y}_{\bullet\bullet\bullet}$$

此时，平方和分解为：

$$SS_{总}=\sum_{i,j,k}\left(\hat{\alpha}_i^{\,2}+\hat{\beta}_j^{\,2}+\hat{\gamma}_{ij}^{\,2}+[y_{ijk}-(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j+\hat{\gamma}_{ij})]^2\right)$$

$$=\sum_{i,j,k}(\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet\bullet\bullet})^2+\sum_{i,j,k}(\bar{y}_{\bullet j\bullet}-\bar{y}_{\bullet\bullet\bullet})^2+\sum_{i,j,k}(\bar{y}_{ij\bullet}-\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet j\bullet}+\bar{y}_{\bullet\bullet\bullet})^2$$

$$+\sum_{i,j,k}(y_{ijk}-\bar{y}_{ij\bullet})^2$$

$$=SS_{处理}+SS_{区组}+SS_{交互}+SS_W$$

可加性检验

可加性检验：

$H_0:\gamma_{ij}=0,\ \ i=1,...,I,\ j=1,..,J$ （no interaction)

$$F_{交互}=\frac{SS_{交互}/(J-1)(I-1)}{SS_W/(IJK-IJ)}\sim_{H_0}F_{(I-1)(J-1),IJK-IJ}$$

其中$SS_{交互}=\sum_{i,j,k}(\bar{y}_{ij\bullet}-\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet j\bullet}+\bar{y}_{\bullet\bullet\bullet})^2,\ \ SS_W=\sum_{i,j,k}(y_{ijk}-\bar{y}_{ij\bullet})^2$

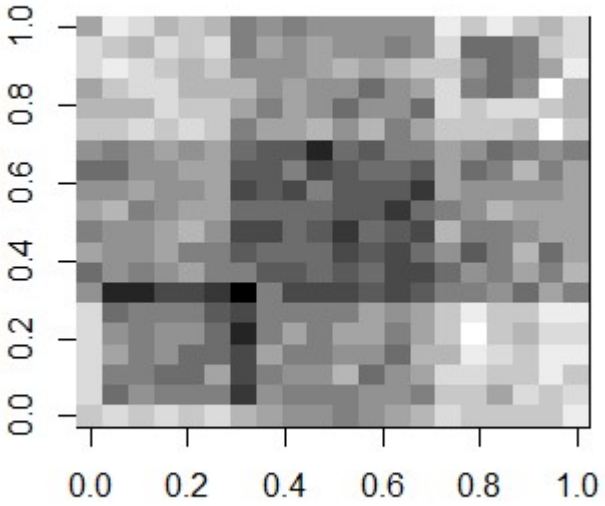
左图: 原数据矩阵，

右图: 消除可加模型（原数据减去行、列效应)后的残差,

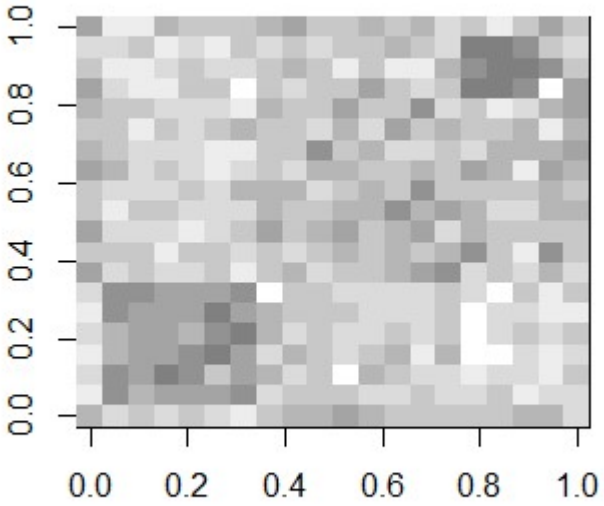
可加模型下，残差应该近似地服从$(0,\sigma^2)$分布，

左下和右上角说明数据偏离可加模型，存在交互项。

y_{ij}



$y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j = y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet}$



Tukey 1-df 可加性检验

解决交互作用效应参数太多的问题

两因素，I, J水平，响应$y_{ijk}, k = 1,2,..., n_{ij}$ iid $\sim N(\mu_{ij}, \sigma^2)$,其中

$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$.

$H_0 : \gamma_{ij} = 0, i = 1,..., I, j = 1,.., J,$

可加性检验的自由度为$(I-1)(J-1), n - IJ$

- 当I, J 较大时，检验的参数个数/自由度为$(I-1)(J-1)$太大,功效不大。
- 另外一方面，如果$n_{ij} \equiv 1$（没有重复测量), $RSS = 0$,如何检验可加性？

Tukey1- df可加性检验解决了上述两个问题。

Tukey模型：

假设交互作用效应 $\gamma_{ij} = \lambda \alpha_i \beta_j$， 即 $\mu_{ij} = \mu + \alpha_i + \beta_j + \lambda \alpha_i \beta_j$

$H_0 : \lambda = 0$

算法（$n_{ij} \equiv 1$）：

(1)计算$\hat{\mu} = \bar{y}_{\bullet\bullet}$，$\hat{\alpha}_i = \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}$，$\hat{\beta}_j = \bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet}$，$\hat{\gamma}_{ij} = y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet}$

(2)令$w_{ij} = \hat{\alpha}_i \hat{\beta}_j$，拟合$y_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \lambda w_{ij} + \varepsilon_{ij}$，即无截距的简单回归

$$\hat{\gamma}_{ij} = \lambda w_{ij} + \varepsilon_{ij}$$

得LS估计：$\hat{\lambda} = \dfrac{\sum_i \sum_j w_{ij} y_{ij}}{\sum_i \sum_j w_{ij}^2} = \dfrac{\sum_i \sum_j \hat{\alpha}_i \hat{\beta}_j y_{ij}}{\sum_i \hat{\alpha}_i^2 \sum_j \hat{\beta}_j^2}$

（3）计算交互作用平方和,残差平方和：

$$SS_{\text{int}} = \sum_i \sum_j (\hat{\lambda} \hat{\alpha}_i \hat{\beta}_j)^2,\ SS_W = SS_{\text{总}} - \left(J \sum_i \hat{\alpha}_i^2 + I \sum_j \hat{\beta}_j^2 + SS_{\text{int}} \right).$$

(4）计算F检验统计量：$F_{\text{Tukey}} = \dfrac{SS_{\text{int}}}{SS_W/(IJ - I - J)} \sim_{H_0} F_{1, IJ-I-J}$，

自由度为 1，$IJ-I-J$

附3：非参数检验和列联表检验

非参数检验和列联表检验与方差分析的对应关系

方法 总体分布 零分布	**方差分析** 正态分布 F分布	**非参数检验** 任何分布 卡方或置换	**属性数据** 多项分布 卡方或置换/Fisher
两总体（处理: 2水平）	两样本t检验	Wilcoxn秩和检验	Pearson卡方检验
多总体（处理: 多水平）	单因素方差分析	Kruskal-Wallis检验	Pearson卡方检验
多总体（处理×分层）	两因素方差分析	Friedman检验	CMH检验
配对 (2水平处理×分层)	配对t检验	Wilconx符号秩检验/Fisher符号检验	McNemar检验
无限总体	回归模型	非参数回归	Logistic回归/多项回归