软件工程统计方法

方差分析

陈振宇

南京大学软件学院

Email:zychen@software.nju.edu.cn

Homepage:software.nju.edu.cn/zychen

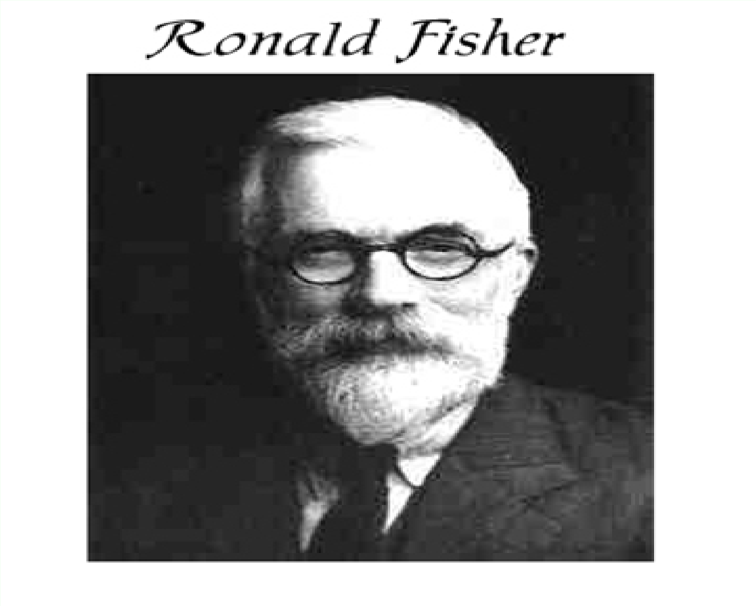
内 容 提 纲

- ❑ 统计学导论
- ❑ 描述统计
- ❑ 概率计算基础
- ❑ 随机变量及其分布
- ❑ 统计量及其抽样分布
- ❑ 参数估计
- ❑ 参数假设检验
- ❑ 非参数假设检验
- ❑ 方差分析
- ❑ 回归分析

1 基本概念

在现实的生产和经营管理中, 经常要分析各种因素对研究对象某些特征值的影响. 方差分析(Analysis of Variance, ANOVA)就是采用数理统计方法对数据进行分析, 以鉴别各种因素对研究对象的某些特征值影响大小的一种有效方法.

由英国著名统计学家R.A.FISHER推导，用于分析试验数据中各个因素对试验指标值影响显著程度的一种统计方法。



第 3 页 共 100 页
返 回
全 屏
关 闭

基 本 概 念

- 指标: 研究对象的特征值, 即所考察的试验(其涵义包括调查, 收集等)结果（如产品品质量、数量、销量、成本等）称为试验指标, 简称指标.
- 因素: 在试验中对所关心的指标有影响的、要加以考察而改变状态的原因称为试验因素, 简称因素.
- 水平: 每个因素可以采取不同的方法或取不同的值, 称之为不同的因素水平, 简称水平.

- 单因素方差分析
- 双因素无重复方差分析
- 双因素等重复方差分析
- 多因素极差分析

例 子

方差分析是为了检验因素的不同水平是否会影响, 通常转化为等方差的样本均值是否相等的问题.

Example 1 试验中选用*5*种不同的除杂方法, 每种方法做*4*次试验, 如下表:

x_{ij}	1	2	3	4	$\overline{x}_{i\cdot}$
A_1	*25.6*	*22.2*	*28.0*	*29.8*	*26.4*
A_2	*24.4*	*30.0*	*29.0*	*27.5*	*27.7*
A_3	*25.0*	*27.7*	*23.0*	*32.2*	*27.0*
A_4	*28.8*	*28.0*	*31.5*	*25.9*	*28.6*
A_5	*20.6*	*21.2*	*22.0*	*21.2*	*21.3*

即检验$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$.

对于m个因素水平$A_1, \cdots, A_m$进行r次试验, 可得$n = mr$个样本观察值. 检验$H_0 : \mu_1 = \cdots = \mu_m$, 传统的假设检验方法(样本两两比较)不再适用.

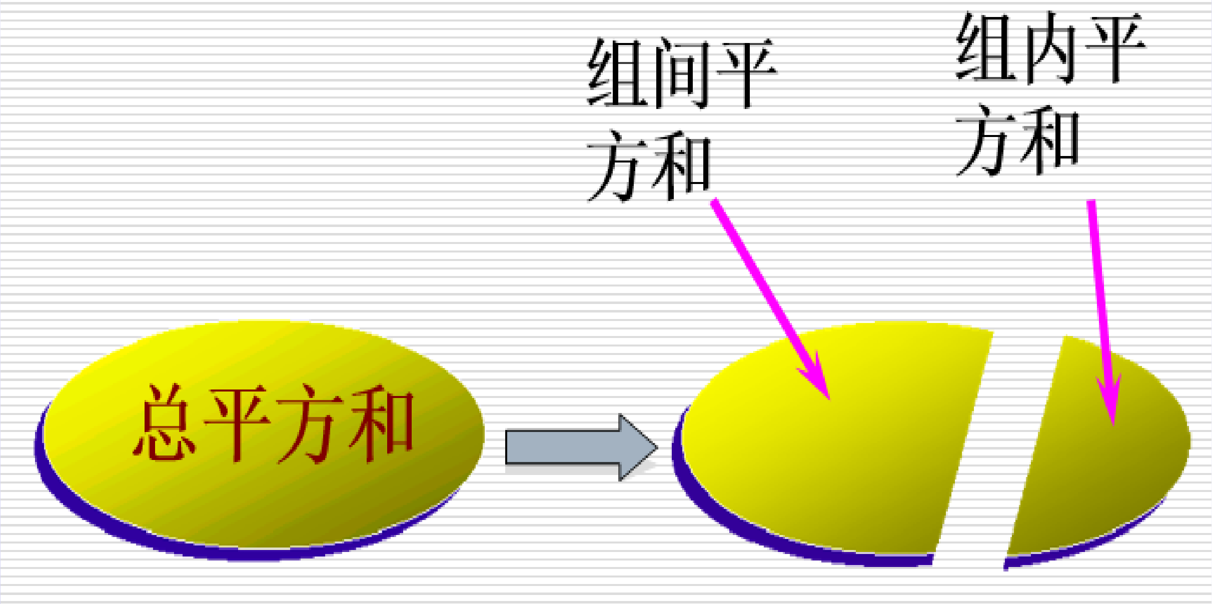
(1) 对于m个因素水平, 需要$\frac{m(m-1)}{2}$假设检验.

(2) 对$2(r-1)$个自由度估计样本均值, 精度不够高.

(3) 样本检验的正确率为$(1-\alpha)^m$, m增大时准确率降低.

方 差 分 析

方差分析是要判断因素各水平对指标是否有显著影响, 归结为判断不同总体是否有相同分布的问题. 由于实际中常遇到的是具有正态分布的总体, 同时, 在进行方差分析时, 除了所关心的因素外, 其他条件总是尽可能使其保持一致, 这样就可以认为每个总体的方差是相同的.



2 单因素方差分析

单因素方差分析(One-Way ANOVA)是要判断一个因素各水平对指标是否有显著影响. 对于m个因素水平$A_1, \cdots, A_m$进行r次, 第i个水平第j次试验的观察值为x_{ij}, 则可得下表:

A_i	x_{ij}			$\overline{x}_{i\cdot}$
A_1	x_{11}	$\cdots$	x_{1r}	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
A_m	x_{m1}	$\cdots$	x_{mr}	

我们需要检验$H_0 : \mu_1 = \cdots = \mu_m$?

为了方便计算和论述, 我们采用以下记号:

$$x_{i\cdot} = \sum_{j=1}^{r} x_{ij}, \overline{x}_{i\cdot} = \frac{1}{r}\sum_{j=1}^{r} x_{ij} = \frac{1}{r} x_{i\cdot}$$

$$x_{\cdot\cdot} = \sum_{i=1}^{m}\sum_{j=1}^{r} x_{ij}, \overline{x} = \frac{1}{m}\sum_{j=1}^{r} x_{i\cdot} = \frac{1}{mr}\sum_{i=1}^{m}\sum_{j=1}^{r} x_{ij} = \frac{1}{mr} x_{\cdot\cdot}$$

偏差平方和分解

试验数据的波动使用偏差平方和表示, 每个偏差可表示为$(x_{ij}-\mu)^2$. 使用样本平均值$\overline{x}$估计μ, 则总偏差平方和可表示为

$$S_T=\sum_{i=1}^{m}\sum_{j=1}^{r}(x_{ij}-\overline{x})^2$$

总偏差来自两个方面:

- 因素的水平不同带来的偏差, 称水平间偏差, 表示为$\alpha_i=\mu_i-\mu$. 若使用$\overline{x}_{i\cdot}$估计μ_i, 则水平间偏差平方和可表示为:

$$S_A=r\sum_{i=1}^{m}(\overline{x}_{i\cdot}-\overline{x})^2$$

- 随机性带来的偏差, 称随机误差, 表示为$\varepsilon_{ij}=x_{ij}-\mu_i$. 随机误差采用水平内偏差平方和表示:

$$S_e=\sum_{i=1}^{m}\sum_{j=1}^{r}(x_{ij}-\overline{x}_{i\cdot})^2$$

- 所以总偏差$(x_{ij}-\mu)=\alpha_i+\varepsilon_{ij}$. 所以总偏差平方和分解为:

$$S_T=S_A+S_e$$

偏差平方和化简

在实际应用中, 为了计算简便, 常用下列简便算法求S_T, S_A和S_e.

步骤如下:

1. 计算修正项C:

$$C = \frac{1}{n} x_{\cdot\cdot}^2$$

2. 计算总数据平方和Q_T:

$$Q_T = \sum_{i=1}^{m} \sum_{j=1}^{r} x_{ij}^2$$

3. 计算总偏差平方和$S_T = Q_T - C$.
4. 计算水平间数据平方和Q_A:

$$Q_A = \frac{1}{r} \sum_{i=1}^{m} x_{i\cdot}^2$$

5. 计算水平间偏差平方和$S_A = Q_A - C$.
6. 计算水平内偏差平方和$S_e = S_T - S_A$.

平均偏差平方和

偏差平方和的大小与参与求和的项数有关, 为了比较S_A和S_e的大小, 应消除求和项数的影响.
根据前面的理论可知, S_A和S_e的平均值, 不是它们直接除以对应的求和项数, 而应除以它们对应的自由度.

- S_T的自由度f_T:
  $S_T = \sum_{i=1}^m \sum_{j=1}^r (x_{ij} - \overline{x})^2$具有$n = mr$个数据$x_{ij}$, 存在一个线性约束$\sum_{i=1}^m \sum_{j=1}^r (x_{ij} - \overline{x}) = 0$, 则自由度$f_T = n - 1$.
- S_A的自由度f_A:
  $S_A = r\sum_{i=1}^m (\overline{x}_{i\cdot} - \overline{x})^2$具有$m$个数据$\overline{x}_{i\cdot}$, 存在一个线性约束$\sum_{i=1}^m (\overline{x}_{i\cdot} - \overline{x}) = 0$, 则自由度$f_A = m - 1$.
- S_e的自由度f_e:
  $S_e = \sum_{i=1}^m \sum_{j=1}^r (x_{ij} - \overline{x}_{i\cdot})^2$具有$n$个数据$x_{ij}$, 存在$m$个线性约束$\sum_{j=1}^r (x_{ij} - \overline{x}_{i\cdot}) = 0, j = 1, \cdots, m$ 则自由度$f_e = n - m$.
- 显然有

$$f_T = f_A + f_e$$

显著性检验

S_A代表水平间数据波动, S_e代表水平内数据波动, 通过比较它们的大小来分析不同因素水平对试验数据的影响.

H_0为真时, 所有样本都可以看作来自同一正态总体$N(\mu_i, \sigma^2)$. 因为$V_A = \frac{S_A}{m-1}$, $V_e = \frac{S_e}{n-m}$都是总体方差σ^2的无偏估计. 当H_0成立时, V_A和V_e分别是自由度为$m-1, n-m$的$\mathcal{X}^2$变量, 所以统计量

$$F = \frac{V_A}{V_e} \sim F(m-1, n-m)$$

因此当F接近1时, 我们判定假设H_0成立. S_e代表水平内数据波动(主要体现随机误差), S_A代表水平间的数据波动(即不但有随机误差, 更多体现因素水平差异). 当F比1大较多时, 我们可认为由于因素水平不同带来了试验数据差异. 那么具体如何界定F值的性质呢?

我们通常取$\alpha = 0.01$和$\alpha = 0.05$, 从F分布表查出$F_{0.01}$和$F_{0.05}$的值, 通过样本观察值计算F_0.

- $F_0 \geq F_{0.01}$我们称为高度显著差异, 记为$**$.
- $F_{0.05} \leq F_0 < F_{0.01}$我们称为显著差异, 记为$*$.
- $F_0 < F_{0.05}$我们称为无显著差异.

3 方差分析表

上述计算过程可以总结出一个方差分析表格.

方差来源	偏差平方和	自由度	方差	F值	F_α	显著性
因素(水平间)	S_A	$f_A = m - 1$	$V_A = \frac{S_A}{f_A}$	$F_A = \frac{V_A}{V_e}$	$F_{0.05}$	
误差(水平内)	S_e	$f_e = n - m$	$V_e = \frac{S_e}{f_e}$		$F_{0.01}$	
总和	S_T	$n - 1$				

例 子

我们检验$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$. 样本数据及其计算如下:

A_i	1	2	3	4	$x_{i\cdot}$	$x_{i\cdot}^2$
A_1	25.6	22.2	28.0	29.8	105.6	11151.36
A_2	24.4	30.0	29.0	27.5	110.9	12298.81
A_3	25.0	27.7	23.0	32.2	107.9	11642.41
A_4	28.8	28.0	31.5	25.9	114.2	13041.64
A_5	20.6	21.2	22.0	21.2	85.0	7225.00
					$\sum_i$:523.6	$\sum_i$:55359.22
$\sum_i x_{ij}^2$	3129.52	3393.57	3630.25	3801.38	$\sum_j$:13954.72	

$$C = \frac{1}{n}x_{\cdot\cdot}^2 = \frac{523.6^2}{20} = 13707.85$$

$$S_T = Q_T - C = 13954.72 - 13707.85 = 246.87$$

$$S_A = Q_A - C = \frac{55359.22}{4} - C = 13839.81 - 13707.85 = 131.96$$

$$S_e = S_T - S_A = 246.87 - 131.96 = 114.91$$

例 子

- 自由度:

$$f_A = m - 1 = 4, f_e = n - m = 15$$

- 方差:

$$V_A = \frac{S_A}{f_A} = \frac{131.96}{4} = 32.99, V_e = \frac{S_e}{f_e} = \frac{114.91}{15} = 7.66$$

- F值:

$$F = \frac{V_A}{V_e} = \frac{32.99}{7.66} = 4.3$$

方差分析表计算如下:

方差来源	偏差平方和	自由度	方差	F值	F_α	显著性
因素(水平间)	131.96	4	32.99	4.3	$F_{0.05}(4,15)=3.05$	*
误差(水平内)	114.91	15	7.66		$F_{0.01}(4,15)=4.89$	
总和	246.87	19				

$F_{0.05}(4,15) < F < F_{0.01}(4,15)$, 故拒绝$H_0$, 即不同除杂方法对除杂效果有显著影响.

数 据 预 处 理

有些情况, 我们可以对原数据作如下预处理,

$$x'_{ij} = \frac{x_{ij} - a}{b}, b \neq 0$$

来简化计算. 可以证明(略), 这样的线性变换不影响最终F的值.

例如, 本例中$x'_{ij} = x_{ij} - 25$, 可得下表:

A_i	1	2	3	4
A_1	0.6	-2.8	3.0	4.8
A_2	-0.6	5.0	4.0	2.5
A_3	0	2.7	-2.0	7.2
A_4	3.8	3.0	6.5	0.9
A_5	-4.4	-3.8	-3.0	-3.8

4 多重比较

- 方差分析是通过F检验讨论组间变异在总变异中的作用,借以对两组以上的平均数进行差异显著性检验, 得到的是一个整体的结论。
- 如果F检验不显著，说明实验中的自变量（因素）对因变量没有显著影响，检验就此结束。
- 如果F检验的结果显著，表明多组平均数两两比较中至少有一对平均数间的差异达到了显著水平，至于哪一对并没有回答。需要专门的分析技术，即方差分析中的多重比较。
  - 最小显著差异法(LSD)

多重比较

最小显著差异法, Least Significant Different (LSD)
对m组中的两组的平均数进行比较，当样本容量均为n时，有

$$\frac{|\overline{x}_i - \overline{x}_j|}{\sqrt{\frac{2}{n}S_e}} \sim t(df)$$

$$LSD_\alpha = t_{\alpha/2}\sqrt{\frac{2}{n}S_e}$$

多重比较

$$LSD_\alpha = t_{\alpha/2}\sqrt{\frac{2}{n}S_e}$$

$$t_{0.025}(15) = 2.1314$$

$$\sqrt{\frac{2}{n}S_e} = \sqrt{\frac{2}{4}7.66} = 1.957$$

$$LSD_{0.05} = 4.17$$

$$|\overline{x}_1 - \overline{x}_2| = |26.4 - 27.7| = 0.8 < 4.17$$

$$|\overline{x}_1 - \overline{x}_5| = |26.4 - 21.3| = 5.1 > 4.17$$

练 习

某商场以不同销售方式A_1, A_2, A_3销售电器, 连续三天的销售量如下. 试分析销售方式是否有显著影响.

销售方式	第一天	第二天	第三天
A_1	56	50	53
A_2	56	47	53
A_3	50	50	53

备注:计算结果保留2位小数.

5 双因素无重复方差分析

同时考虑因素A和B对试验结果的影响, 因素A取$A_1, \cdots, A_a$共a个水平, 因素B取$B_1, \cdots, B_b$共b个水平. A和B两因素的每种水平搭配A_iB_j各进行一次独立试验, 共进行$n = ab$次试验, 试验数据为$x_{ij}(i = 1, 2, \cdots, a; j = 1, 2\cdots, b)$. 记

$$x_{i\cdot} = \sum_{j=1}^{b} x_{ij}(i = 1, 2, \cdots, a),\ \overline{x}_{i\cdot} = \frac{1}{b}x_{i\cdot}$$

$$x_{\cdot j} = \sum_{i=1}^{a} x_{ij}(j = 1, 2, \cdots, b),\ \overline{x}_{\cdot j} = \frac{1}{a}x_{\cdot j}$$

$$x_{\cdot\cdot} = \sum_{i=1}^{a}\sum_{j=1}^{b} x_{ij},\ \overline{x}_{\cdot\cdot} = \frac{1}{n}x_{\cdot\cdot}, n = ab$$

方差分析前提条件:

1. 在每个因素每一个水平上试验的结果是一个正态随机变量(x_{ij}).
2. 相同因素的不同水平对应的正态总体的方差是相等的, 具有方差齐性.
3. 各因素总体是认为是相互独立的, 样本与样本之间也是相互独立的.

偏差平方和分解

- 总偏差平方和表示

$$S_T = \sum_{i=1}^{m}\sum_{j=1}^{r}(x_{ij} - \overline{x})^2$$

- A因素水平间偏差平方和

$$S_A = b\sum_{i=1}^{m}(\overline{x}_{i\cdot} - \overline{x})^2$$

- B因素水平间偏差平方和

$$S_B = a\sum_{j=1}^{m}(\overline{x}_{\cdot j} - \overline{x})^2$$

- 水平内的偏差平方和

$$S_e = \sum_{i=1}^{m}\sum_{j=1}^{r}(x_{ij} - \overline{x}_{i\cdot} - \overline{x}_{\cdot j} + \overline{x})^2$$

$$S_e = S_T - S_A - S_B$$

偏差平方和化简

在实际应用中, 为了计算简便, 常用下列简便算法求S_T, S_A,S_B和S_e.

- 令$Q_T=\sum_{i=1}^{a}\sum_{j=1}^{b}x_{ij}^2$, 修正项$C=\frac{1}{n}x_{\cdot\cdot}^2$, 则

$$S_T=\sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij}-\overline{x})^2=\sum_{i=1}^{a}\sum_{j=1}^{b}x_{ij}^2-\frac{1}{n}x^2=Q_T-C$$

- 再令$Q_A=\frac{1}{b}\sum_{i=1}^{a}x_{i\cdot}^2$, 则

$$S_A=b\sum_{i=1}^{a}(\overline{x}_{i\cdot}-\overline{x})^2=\frac{1}{b}\sum_{i=1}^{a}x_{i\cdot}^2-\frac{1}{n}x^2=Q_A-C$$

- 再令$Q_B=\frac{1}{a}\sum_{j=1}^{b}x_{\cdot j}^2$, 则

$$S_B=a\sum_{j=1}^{b}(\overline{x}_{\cdot j}-\overline{x})^2=\frac{1}{a}\sum_{j=1}^{b}x_{\cdot j}^2-\frac{1}{n}x^2=Q_B-C$$

$$S_e=S_T-S_A-S_B=Q_T-Q_A-Q_B+C$$

平 均 偏 差 平 方 和

偏差平方和的大小与参与求和的项数有关, 为了比较S_A和S_e的大小和S_B和S_e的大小, 应求平均消除求和项数的影响.

- S_T的自由度f_T:
  $S_T = \sum_{i=1}^a \sum_{j=1}^b (x_{ij} - \overline{x})^2$具有$n = ab$个数据$x_{ij}$, 存在一个线性约束$\sum_{i=1}^a \sum_{j=1}^b (x_{ij} - \overline{x}) = 0$, 则自由度$f_T = n - 1$.
- S_A的自由度f_A:
  $S_A = b\sum_{i=1}^a (\overline{x}_{i\cdot} - \overline{x})^2$具有$a$个数据$\overline{x}_{i\cdot}$, 存在一个线性约束$\sum_{i=1}^a (\overline{x}_{i\cdot} - \overline{x}) = 0$, 则自由度$f_A = a - 1$.
- S_B的自由度f_B:
  $S_B = a\sum_{j=1}^b (\overline{x}_{\cdot j} - \overline{x})^2$具有$b$个数据$\overline{x}_{\cdot j}$, 存在一个线性约束$\sum_{j=1}^b (\overline{x}_{\cdot j} - \overline{x}) = 0$, 则自由度$f_B = b - 1$.
- S_e的自由度$f_e = (a-1)(b-1)$.

$$f_T = f_A + f_B + f_e$$

显 著 性 检 验

S_A和S_B代表水平间数据波动, S_e代表水平内数据波动, 通过比较它们的大小来分析不同因素水平对试验数据的影响.

H_0为真时, 所有样本都可以看作来自同一正态总体$N(\mu,\sigma^2)$. 因为$V_A = \frac{S_A}{a-1}$,$V_B = \frac{S_B}{b-1}$, $V_e = \frac{S_e}{(a-1)(b-1)}$都是总体方差$\sigma^2$的无偏估计. 当$H_0$成立时, V_A, V_B和V_e分别是自由度为$a-1, b-1$和$(a-1)(b-1)$的$\mathcal{X}^2$变量, 所以统计量

$$F_A = \frac{V_A}{V_e} \sim F(a-1,(a-1)(b-1))$$

同样，S_B和S_e分别是自由度为$b-1$, $(a-1)(b-1)$的$\mathcal{X}^2$变量, 所以统计量

$$F_B = \frac{V_B}{V_e} \sim F(b-1,(a-1)(b-1))$$

所以拒绝域为$F_A \geq F_\alpha(a-1,(a-1)(b-1))$, $F_B \geq F_\alpha(b-1,(a-1)(b-1))$.

6 方差分析表

上述计算过程可以总结出一个方差分析表格.

来源	平方和	自由度	方差	F值	F_α	显著性
A	S_A	f_A	$V_A=\frac{S_A}{f_A}$	$F_A=\frac{V_A}{V_e}$	查表	
B	S_B	f_B	$V_B=\frac{S_B}{f_B}$	$F_B=\frac{V_B}{V_e}$	查表	
e	S_e	f_e	$V_e=\frac{S_e}{f_e}$			
总和	S_T	f_T				

例子

Example 2 原来检验果汁中含铅量有三种方法A_1, A_2, A_3, 现研究出另外一种快速检验法A_4, 能否用A_4替代前3种方法? 我们考虑不同的果汁B : $B_1, \cdots B_6$, 进行双因素交叉试验. 试验结果如下:

$A_i \mid B_i$	B_1	B_2	B_3	B_4	B_5	B_6	$x_{i\cdot}$
A_1	0.05	0.46	0.12	0.16	0.84	1.30	2.93
A_2	0.08	0.38	0.40	0.10	0.92	1.57	3.45
A_3	0.11	0.43	0.05	0.10	0.94	1.10	2.73
A_4	0.11	0.44	0.08	0.03	0.93	1.15	2.74
$x_{\cdot j}$	0.35	1.71	0.65	0.39	3.63	5.12	$x_{\cdot\cdot} = 11.85$

例 子

计算各偏差平方和: 计算各偏差平方和:

$$C=\frac{x_{\cdot\cdot}^2}{n}=5.851, Q_T=\sum_i\sum_j x_{ij}^2=10.9853, S_T=Q_T-C=5.1343$$

$$Q_A=\frac{1}{6}\sum_i x_{i\cdot}^2=5.908, S_A=Q_A-C=0.057$$

$$Q_B=\frac{1}{4}\sum_j x_{\cdot j}^2=10.753, S_B=Q_B-C=4.902$$

$$S_e=S_T-S_A-S_B=0.173$$

来源	平方和	自由度	方差	F值	F_α	显著性
A	0.057	3	0.019	1.63	$F_{0.05}(3,15)=3.29$	
B	4.902	5	0.980	83.98	$F_{0.05}(5,15)=4.56$	**
e	0.175	15	0.012			
总和	5.134	23				

试验结论如下:

- A_4可以替代以前的旧方法;
- 不同类型果汁对试验结果影响显著.

7 双因素等重复方差分析

同时考虑因素A和B对试验结果的影响, 因素A取$A_1,\cdots,A_a$共a个水平, 因素B取$B_1,\cdots,B_b$共b个水平. A和B两因素的每种水平搭配A_iB_j各进行r次独立试验, 共进行$N=abr$次试验, 试验数据为$x_{ijk}(i=1,\cdots,a;j=1,\cdots,b;k=1,\cdots,r)$, 表示因素$A$取$i$水平, 因素$B$取$j$水平时, 第$k$次的试验结果. 记

$$x_{ij\cdot}=\sum_{k=1}^{r}x_{ijk}(i=1,2,\cdots,a;j=1,2\cdots,b),\overline{x}_{ij\cdot}=\frac{1}{r}x_{ij\cdot}$$

$$x_{i\cdot\cdot}=\sum_{j=1}^{b}\sum_{k=1}^{r}x_{ijk}(i=1,2,\cdots,a),\overline{x}_{i\cdot\cdot}=\frac{1}{br}x_{i\cdot\cdot}$$

$$x_{\cdot j\cdot}=\sum_{i=1}^{a}\sum_{k=1}^{r}x_{ijk}(j=1,2,\cdots,b),\overline{x}_{\cdot j\cdot}=\frac{1}{ar}x_{\cdot j\cdot}$$

$$\overline{x}=\frac{1}{abr}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}x_{ijk}$$

偏差平方和分解

- 总偏差平方和: $S_T=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(x_{ijk}-\overline{x})^2$
- A因素水平间偏差平方和: $S_A=br\sum_{i=1}^{m}(\overline{x}_{i\cdot\cdot}-\overline{x})^2$
- B因素水平间偏差平方和: $S_B=ar\sum_{j=1}^{m}(\overline{x}_{\cdot j\cdot}-\overline{x})^2$
- 随机误差偏差平方和:

$$S_e=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(x_{ijk}-\overline{x}_{ij\cdot})^2$$

- A,B交互作用波动:

$$S_{AB}=r\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{x}_{ij\cdot}-\overline{x}_{i\cdot\cdot}-\overline{x}_{\cdot j\cdot}+\overline{x})^2$$

$$S_T=S_A+S_B+S_{AB}+S_e$$

偏差平方和化简

- 令$Q_T=\sum_{i=1}^a\sum_{j=1}^b\sum_{k=1}^r x_{ijk}^2$, 修正项$C=\frac{1}{n}x_{\cdots}^2$, 则

$$S_T=\sum_{i=1}^a\sum_{j=1}^b\sum_{k=1}^r(x_{ijk}-\overline{x})^2=Q_T-C$$

- 再令$Q_A=\frac{1}{br}\sum_{i=1}^a x_{i\cdot\cdot}^2$, 则

$$S_A=br\sum_{i=1}^m(\overline{x}_{i\cdot\cdot}-\overline{x})^2=Q_A-C$$

- 再令$Q_B=\frac{1}{ar}\sum_{j=1}^b x_{\cdot j\cdot}^2$, 则

$$S_B=ar\sum_{j=1}^m(\overline{x}_{\cdot j\cdot}-\overline{x})^2=Q_B-C$$

- 再令$Q_{AB}=\frac{1}{r}\sum_{i=1}^a\sum_{j=1}^b x_{ij\cdot}^2$, 则

$$S_{AB}=r\sum_{i=1}^a\sum_{j=1}^b(\overline{x}_{ij\cdot}-\overline{x}_{i\cdot\cdot}-\overline{x}_{\cdot j\cdot}+\overline{x})^2=Q_{AB}-Q_A-Q_B+C$$

$$S_e=S_A+S_B+S_{AB}-S_T=Q_T-Q_{AB}$$

平均偏差平方和

偏差平方和的大小与参与求和的项数有关，为了比较S_A和S_e的大小和S_B和S_e的大小，应求平均消除求和项数的影响.

- S_T的自由度f_T:
  $f_T = n - 1.$
- S_A的自由度f_A:
  $f_A = a - 1.$
- S_B的自由度f_B:
  $f_B = b - 1.$
- S_{AB}的自由度f_{AB}:
  $f_{AB} = (a-1)(b-1).$
- S_e的自由度f_e:
  $f_e = ab(r-1).$

显然有

$$f_T = f_A + f_B + f_{AB} + f_e$$

显著性检验

S_A, S_B代表水平间数据波动, S_e代表水平内数据波动, 通过比较它们的大小来分析不同因素水平对试验数据的影响.

H_0为真时, 所有样本都可以看作来自同一正态总体$N(\mu, \sigma^2)$. 因为$V_A = \frac{S_A}{a-1}$,$V_B = \frac{S_B}{b-1}$, $V_{AB} = \frac{S_{AB}}{(a-1)(b-1)}$, $V_e = \frac{S_e}{(a-1)(b-1)}$都是总体方差$\sigma^2$的无偏估计. 当$H_0$成立时, V_A, V_B, V_AB和V_e分别是自由度为$a-1$, $b-1$, $(a-1)(b-1)$和$ab(r-1)$的$\mathcal{X}^2$变量, 所以统计量

$$F_A = \frac{V_A}{V_e} \sim F(a-1, ab(r-1))$$

同样，S_B和S_e分别是自由度为$b-1$, $ab(r-1)$的$\mathcal{X}^2$变量, 所以统计量

$$F_B = \frac{V_B}{V_e} \sim F(b-1, ab(r-1))$$

同样，S_{AB}和S_e分别是自由度为$(a-1)(b-1)$, $ab(r-1)$的$\mathcal{X}^2$变量, 所以统计量

$$F_{AB} = \frac{S_{AB}}{V_{AB}} \sim F((a-1)(b-1), ab(r-1))$$

8 方差分析表

上述计算过程可以总结出一个方差分析表格.

来源	平方和	自由度	方差	F值	F_α	显著性
A	S_A	f_A	$V_A = \frac{S_A}{f_A}$	$F_A = \frac{V_A}{V_e}$	查表	
B	S_B	f_B	$V_B = \frac{S_B}{f_B}$	$F_B = \frac{V_B}{V_e}$	查表	
AB	S_{AB}	f_{AB}	$V_{AB} = \frac{S_{AB}}{f_{AB}}$	$F_{AB} = \frac{V_{AB}}{V_e}$	查表	
e	S_e	f_e	$V_e = \frac{S_e}{f_e}$			
总和	S_T	f_T				

例 子

Example 3 火箭试验中进行燃料A_1, A_2, A_3, A_4和推进器B_1, B_2, B_3的搭配试验, 各试验2次. 试验结果如下:

$A_i \| B_i$	B_1	B_2	B_3	$x_{i\cdot\cdot}$
A_1	58.2 52.6	56.2 41.2	65.3 60.8	334.3
A_2	49.1 42.8	54.1 50.5	51.6 48.4	296.5
A_3	60.1 58.3	70.9 73.2	39.2 40.7	342.4
A_4	75.8 71.5	58.2 51.0	48.7 41.4	346.6
$x_{\cdot j\cdot}$	468.4	455.3	396.1	$x_{\cdots} = 1319.8$

例 子

计算各偏差平方和:

$$C=\frac{x_{\cdots}^2}{n}=72578.00, Q_T=\sum_i\sum_j\sum_k x_{ijk}^2=75216.30, S_T=Q_T-C=2638.30$$

$$Q_A=\frac{1}{6}\sum_i x_{i\cdot\cdot}^2=72839.68, S_A=Q_A-C=261.68$$

$$Q_B=\frac{1}{8}\sum_j x_{\cdot j\cdot}^2=72948.98, S_B=Q_B-C=370.98$$

$$Q_{AB}=\frac{1}{2}\sum_i\sum_j x_{ij\cdot}^2=74979.35, S_e=Q_T-Q_{AB}=236.95$$

$$S_{AB}=S_T-S_A-S_B-S_e=1768.69$$

来源	平方和	自由度	方差	F值	F_α	显著性
A	261.68	3	87.23	4.42	$F_{0.05}(3,12)=3.49$	*
B	370.98	2	185.49	9.39	$F_{0.01}(2,12)=3.88$	**
AB	1768.69	6	294.78	14.90	$F_{0.01}(6,12)=3.00$	**
e	236.95	12	19.75			
总和	2638.30	23				

9 正交实验设计

对于单因素或两因素试验, 因其因素少, 试验的设计、实施与分析都比较简单. 但在实际工作中, 常常需要同时考察3个或3个以上的实验因素, 若进行全面实验, 则实验的规模将很大, 往往因实验条件的限制而难于实施.

正交实验设计(Orthogonal Experimental Design)是研究多因素多水平的一种设计方法, 它是根据正交性从全面实验中挑选出部分有代表性的点进行实验, 这些有代表性的点具备了“均匀分散, 齐整可比” 的特点, 正交实验设计是析因设计的主要方法.

基 本 概 念

正交实验设计是利用正交表来安排与分析多因素实验的一种设计方法. 它是由实验因素的全部水平组合中, 挑选部分有代表性的水平组合进行实验的, 通过对这部分实验结果的分析了解全面实验的情况, 找出最优的水平组合.

Example 4 例如, 要考察增稠剂用量、pH值和杀菌温度对豆奶稳定性的影响. 每个因素设置3个水平进行实验.A因素是增稠剂用量, 设A_1、A_2、A_3 3个水平; B因素是pH值, 设B_1、B_2、B_3 3个水平; C因素为杀菌温度, 设C_1、C_2、C_3 3个水平.

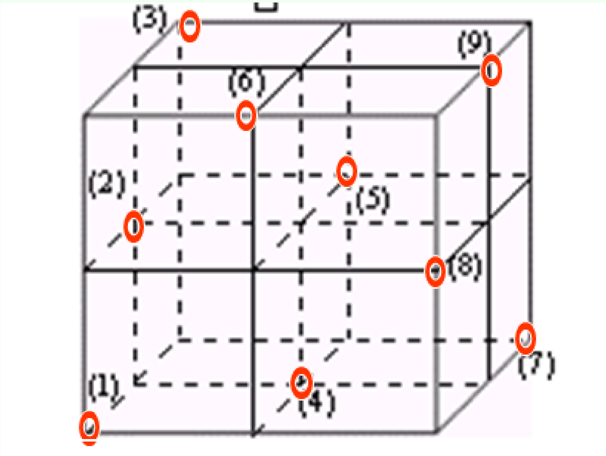
这是一个3因素3水平的实验, 各因素的水平之间全部可能组合有27种.

全面实验: 可以分析各因素的效应, 交互作用, 也可选出最优水平组合. 但全面实验包含的水平组合数较多, 工作量大, 在有些情况下无法完成. 3因素3水平的全面实验水平组合数为$3^3=27$, 4 因素3水平的全面实验水平组合数为$3^4=81$, 10因素10水平的全面实验水平组合数为10^{10}.

10 正 交 表

如对于上述3因素3水平实验, 若不考虑交互作用, 可利用正交表$L_9(3^3)$安排, 实验方案仅包含9个水平组合, 就能反映实验方案包含27个水平组合的全面实验的情况, 找出最佳的生产条件.

$$\begin{array}{lll}(1)A_1B_1C_1 & (2)A_1B_2C_2 & (3)A_1B_3C_3\\(4)A_2B_1C_2 & (5)A_2B_2C_3 & (6)A_2B_3C_1\\(7)A_3B_1C_3 & (8)A_3B_2C_1 & (9)A_3B_3C_2\end{array}$$



正 交 表

正交表使用$L_n(m_1 \times \cdots \times m_k)$标记. L是Latin的第一个字母, n为行数, k个因素, 第i个因素的水平数位m_i. 下表为$L_9(3^3)$

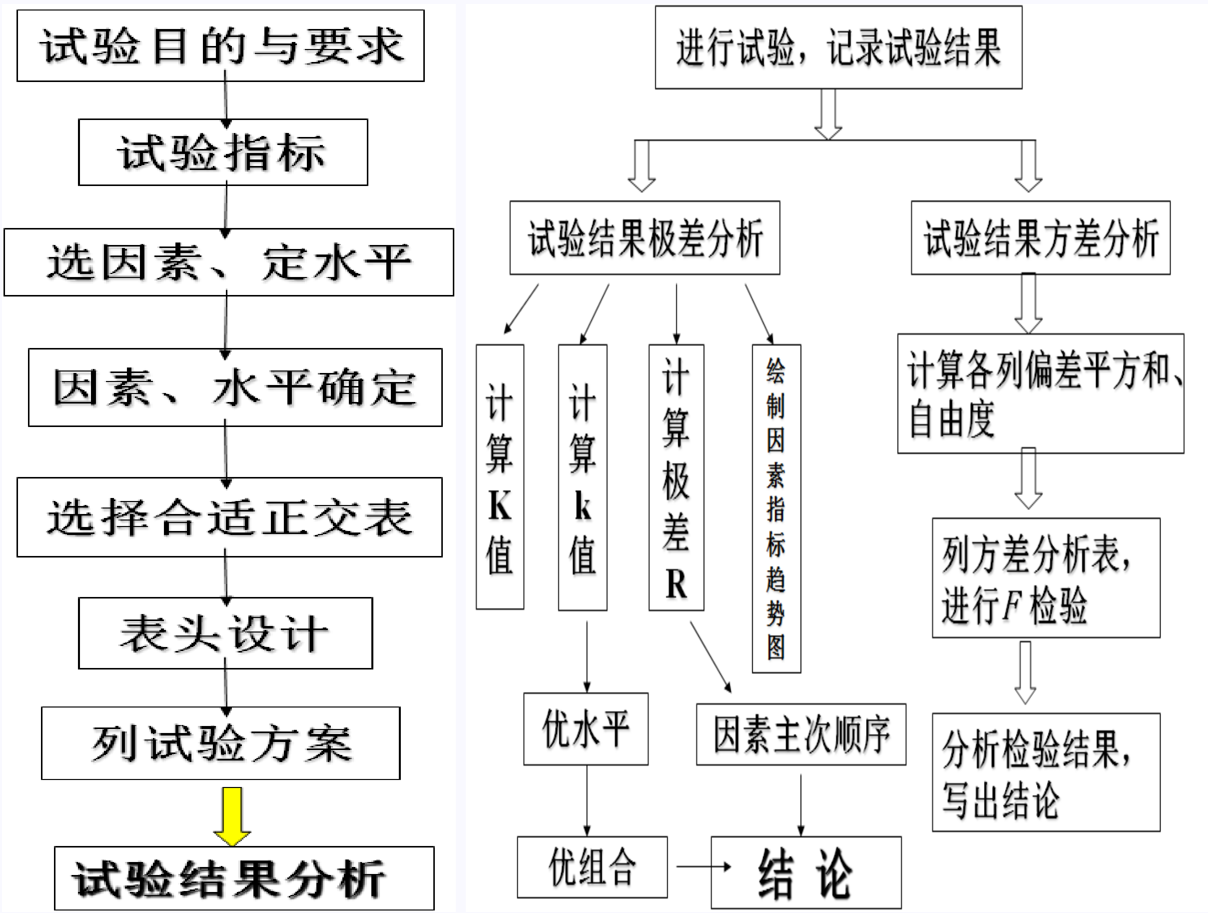
No.	A	B	C
(1)	1	1	1
(2)	1	2	2
(3)	1	3	3
(4)	2	1	2
(5)	2	2	3
(6)	2	3	1
(7)	3	1	3
(8)	3	2	1
(9)	3	3	2

正交表特点:

- 任一列各水平都出现, 且出现次数相等.
- 任两列的各种组合都出现, 且出现次数相等.

使用正交表安排的实验, 具有均衡分散和整齐可比的特点.

11 正交实验分析


试验目的与要求
试验指标
选因素、定水平
因素、水平确定
选择合适正交表
表头设计
列试验方案
试验结果分析
进行试验，记录试验结果
试验结果极差分析
试验结果方差分析
计算K值
计算k值
计算极差R
绘制因素指标趋势图
优水平
因素主次顺序
优组合
结 论
计算各列偏差平方和、自由度
列方差分析表，进行F检验
分析检验结果，写出结论

正 交 实 验 分 析

- 分析方法:
  - 极差分析
  - 方差分析
- 分析目的:
  - 分清各因素及其交互作用的主次顺序;
  - 判断因素对实验指标影响的显著程度;
  - 找出实验因素的优水平和实验范围内的最优组合;
  - 找出指标随因素变化的规律和趋势, 为进一步实验指明方向;
  - 了解各因素之间的交互作用情况;
  - 估计实验误差的大小.

例 子

例 如: 要 考 察 增 稠 剂 用 量A_1, A_2, A_3、pH值B_1, B_2, B_3和 杀 菌 温 度C_1, C_2, C_3对豆奶稳定性的影响. 每个因素设置3个水平进行实验.

No.	A	B	C	实验结果
(1)	1	1	1	6.25
(2)	1	2	2	4.97
(3)	1	3	3	4.54
(4)	2	1	2	7.53
(5)	2	2	3	5.54
(6)	2	3	1	5.50
(7)	3	1	3	11.4
(8)	3	2	1	10.9
(9)	3	3	2	8.95

我们能否判断组合(7)最好?

12 极差分析

极差分析(R法)计算简便, 直观, 简单易懂, 是正交实验结果分析最常用方法. 以上例为实例来说明极差分析.

基本步骤如下:

- 计算K_{mj}: 第j列m水平实验指标和.

$$K_{mj} = \sum_m x_{ij}$$

  - 判断优水平
  - 判断最优组合
- 计算R_j:

$$R_j = max(K_{mj}) - min(K_{mj})$$

  - 判断因素主次

例子

No.	A	B	C	实验结果
(1)	1	1	1	6.25
(2)	1	2	2	4.97
(3)	1	3	3	4.54
(4)	2	1	2	7.53
(5)	2	2	3	5.54
(6)	2	3	1	5.50
(7)	3	1	3	11.4
(8)	3	2	1	10.9
(9)	3	3	2	8.95
K_{1j}	15.76	25.18	22.65	
K_{2j}	18.57	21.41	21.45	
K_{3j}	31.25	18.99	21.48	
优水平	A_3	B_1	C_1	
R_j	15.49	6.19	1.2	

结果分析

- $R_A > R_B > R_C$, 所以因素主次顺序为ABC.
- 最优组合可能为A_3, B_1, C_1.
- 根据K_{mj}走势图可以调整各因素水平.

13 方差分析

方差分析基本思想是将数据的总波动分解成因素水平间波动和随机误差波动两部分, 构造F统计量, 作F检验, 判断因素作用是否显著.

1. 偏差平方和分解

$$S_T = \sum_j S_j + S_e$$

2. 自由度分解

$$f_T = \sum_j f_j + f_e, f_T = n - 1, f_j = m - 1$$

3. 方差计算

$$V_j = \frac{S_j}{f_j}, V_e = \frac{S_e}{f_e}$$

4. 构造统计量

$$F_j = \frac{V_j}{V_e} \sim F(f_j, f_e)$$

化简偏差平方和

$$C = \frac{1}{n}(\sum_{i=1}^{n} x_i)^2$$

$$Q_T = \sum_{i=1}^{n} x_i^2$$

$$Q_j = \frac{1}{r}\sum_{i=1}^{m} K_{ij}^2$$

- 总偏差平方和

$$S_T = Q_T - C$$

- 因素j偏差平方和

$$S_j = Q_j - C$$

- 误差e偏差平方和

$$S_e = S_T - \sum_j S_j$$

例 子

No.	A	B	C	实验结果	平方值
(1)	1	1	1	6.25	39.06
(2)	1	2	2	4.97	24.70
(3)	1	3	3	4.54	20.61
(4)	2	1	2	7.53	56.70
(5)	2	2	3	5.54	30.69
(6)	2	3	1	5.50	30.25
(7)	3	1	3	11.4	129.96
(8)	3	2	1	10.9	118.81
(9)	3	3	2	8.95	80.10
K_{1j}	15.67	25.18	22.65		
K_{2j}	18.57	21.41	21.45		
K_{3j}	31.25	18.9	21.39		
$\sum$				65.58	530.88
K_{1j}^2	248.38	634.03	513.02		
K_{2j}^2	344.84	458.39	460.10		
K_{3j}^2	976.56	360.62	461.39		

例 子

1. 修正项C

$$C=\frac{65.58^2}{9}=477.86$$

2. 总偏差平方和

$$S_T=530.88-477.86=53.02$$

3. 因素偏差平方和

$$S_A=\frac{248.38+344.84+976.56}{3}-477.86=45.4$$

$$S_B=\frac{634.03+458.29+360.62}{3}-477.86=6.49$$

$$S_C=\frac{513.02+460.10+461.39}{3}-477.86=0.31$$

4. 误差偏差平方和

$$S_e=53.02-45.4-6.49-0.31=0.82$$

例 子

计算方差:

$$V_A=\frac{45.4}{2}=22.7, V_B=\frac{6.49}{2}=3.23, V_C=\frac{0.31}{2}=0.155, V_e=\frac{0.82}{2}=0.41$$

计算F值:

$$F_A=\frac{22.7}{0.41}=55.37, F_B=\frac{3.24}{0.41}=7.90, F_C=\frac{0.16}{0.41}=0.0.39$$

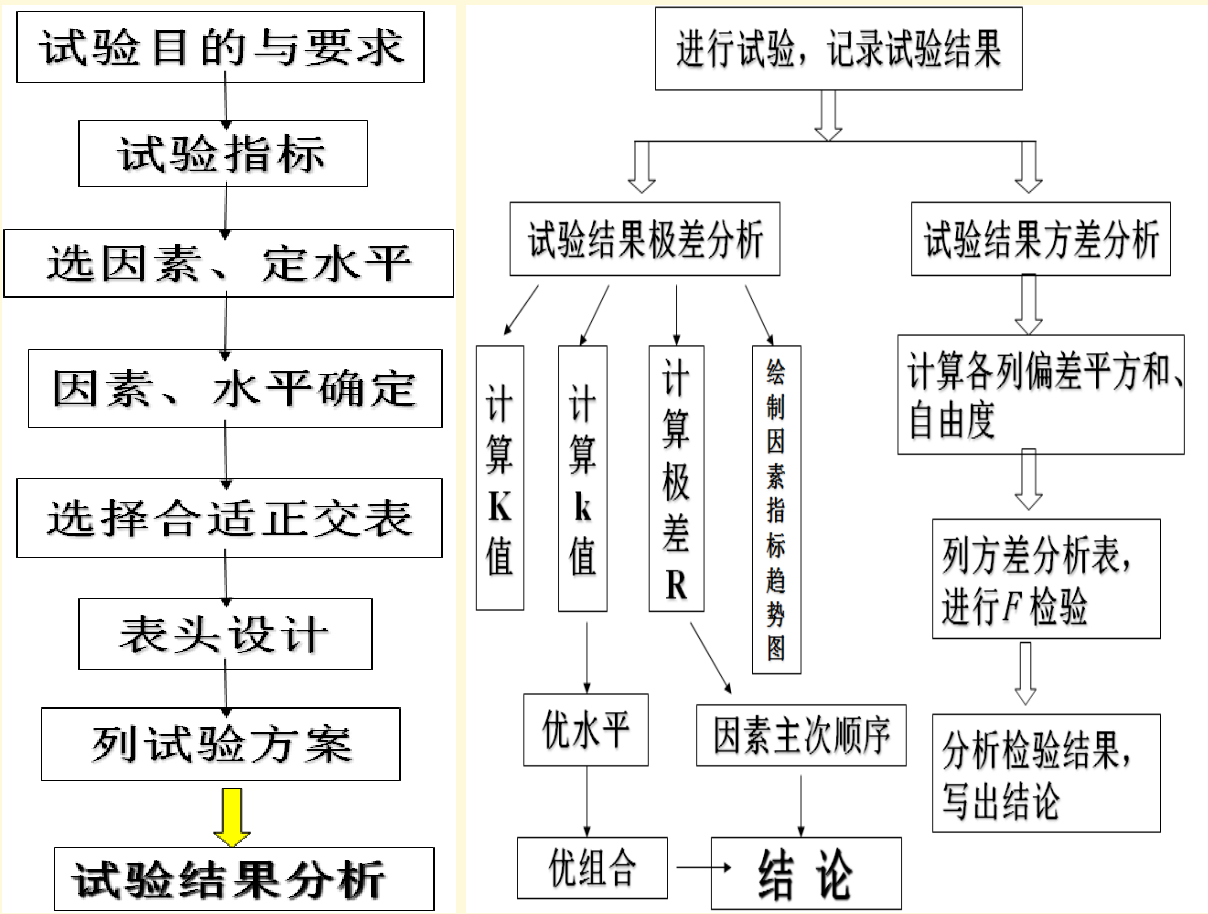
来源	平方和	自由度	方差	F值	F_α	显著性
A	45.4	2	22.7	55.37	$F_{0.05}(2,2)=19.00$	*
B	6.49	2	3.24	7.90	$F_{0.01}(2,2)=99.01$	
C	0.31	2	0.16	0.39		
e	0.83	2	0.41			
总和	53.02	8				

- 根据F值, 因素影响顺序为$A>B>C$, 因素主次顺序ABC.
- 实验指标越大越好, 因素A,B有显著影响, 所以确定优选水平为A_3和B_1, 因素C对实验结果几乎没有影响, 可以通过其它方法(如经济原因)选定.

例 子

No.	A	B	C	实验结果
(1)	1	1	1	1365
(2)	1	2	2	1395
(3)	1	3	3	1385
(4)	2	1	2	1390
(5)	2	2	3	1395
(6)	2	3	1	1380
(7)	3	1	3	1390
(8)	3	2	1	1390
(9)	3	3	2	1410

正交实验分析


试验目的与要求
试验指标
选因素、定水平
因素、水平确定
选择合适正交表
表头设计
列试验方案
试验结果分析
进行试验，记录试验结果
试验结果极差分析
试验结果方差分析
计算K值
计算k值
计算极差R
绘制因素指标趋势图
优水平
因素主次顺序
优组合
结论
计算各列偏差平方和、自由度
列方差分析表，进行F检验
分析检验结果，写出结论