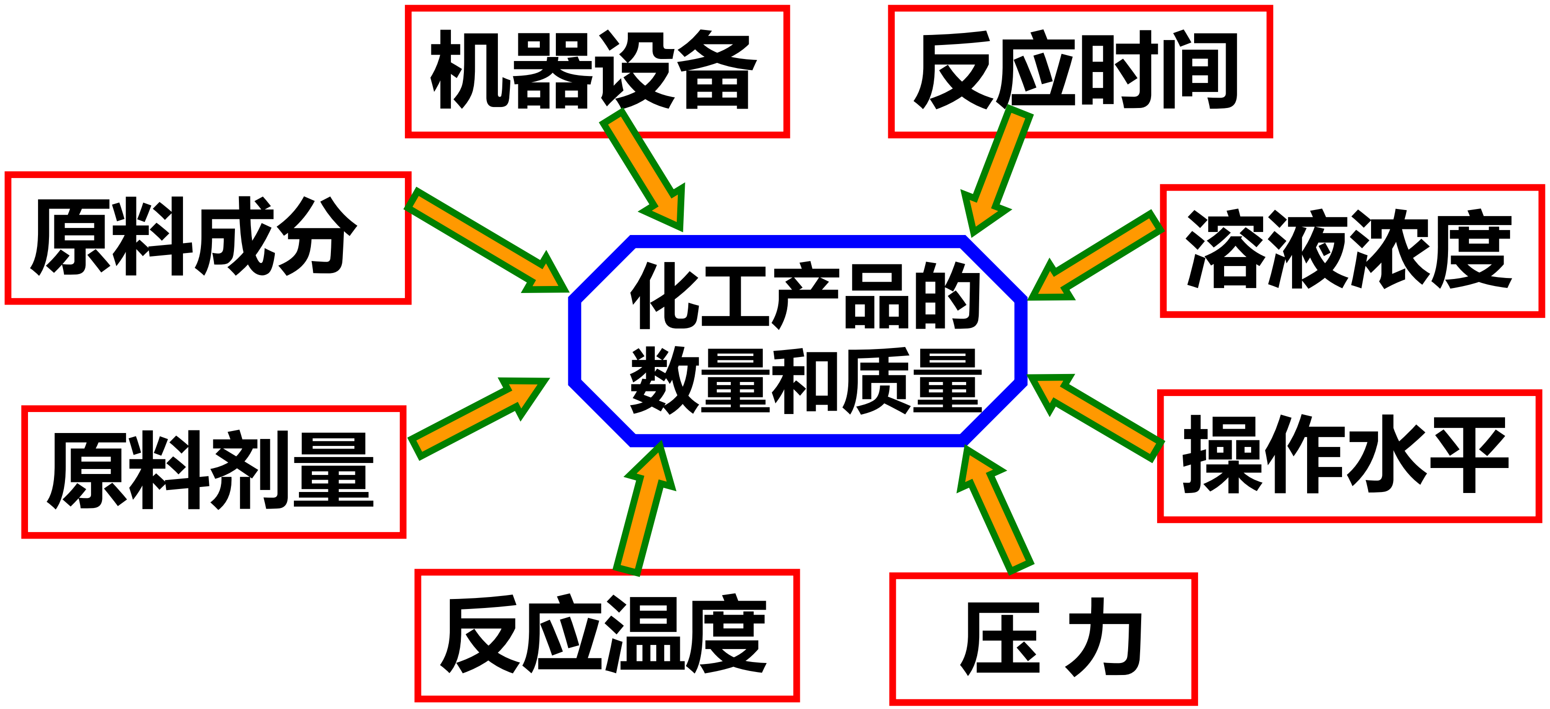
第四章　方差分析

在科学研究、科学实验和工农业生产中, 为了研制新产品，提高产品产量和产品性能，改革工艺条件，降低成本, 需要进行各种试验. 例如



1 试验指标—衡量试验结果的量称为试验指标.

例如：亩产量、强度、硬度等(数量指标). 产品的颜色深浅、气味、光泽等(定性指标, 通常要量化处理). 在进行分析时，试验指标视为可观察的随机变量.

2 试验因素—影响试验指标分布的量称为因素.

例如：化学反应中的温度、压力、催化剂用量，农业试验中的品种、肥料等.

因素既可以是定量的也可以是定性的.

可控因素—化学反应中, 温度在实验室内可控,

不可控因素—农业试验中, 日平均气温不可控.

影响试验指标的因素一般很多, 进行方差分析时，我们总是在可控因素中选取, 并且尽量少而精.

3 因素水平—因素在试验中处于的状态称为因素水平.

例如：化学反应中的温度可在一定范围内变化,

假设选取几种特殊的状态：80^{0}C, 90^{0}C, 100^{0}C. 温度就成为3水平的因素. 当然这种假设要根据经验和专业知识.

单因素试验—只考虑一个因素的试验,
双因素试验—考虑二个因素的试验，
多因素试验—考虑多个因素的试验.

单因素试验通常有均分法、0.618分法、分数法等，多因素试验最常用的是正交试验法.

4 方差分析—就是根据试验得到的数据进行分析

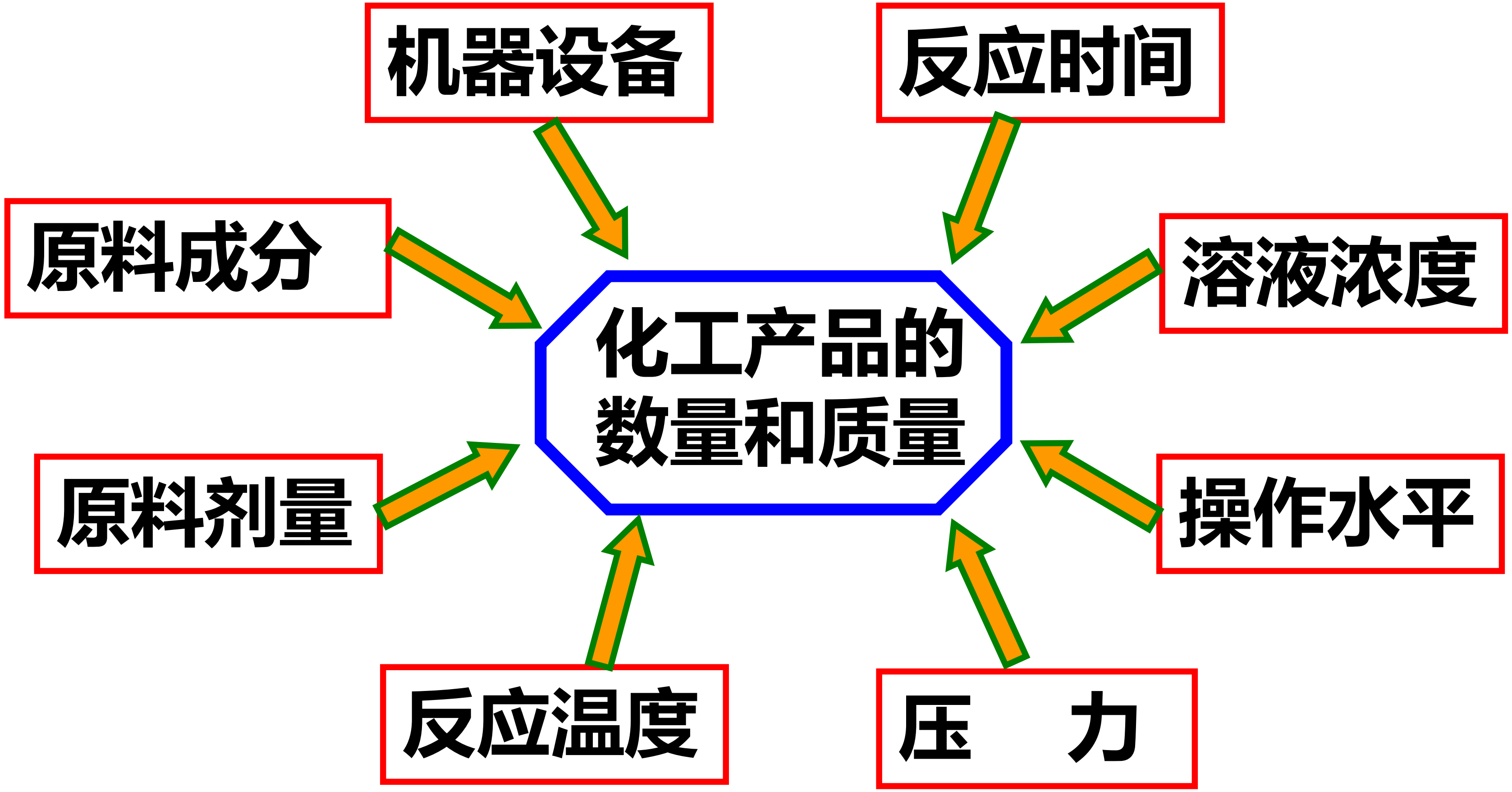
方差分析的目的：判断哪些因素对试验指标有

显著性影响；影响显著因素的好水平是什么；因素之间是否有交互作用；一个因素的各个水平之间是否有显著性差异.

本章的内容

- 单因素方差分析
- 双因素方差分析
  - 无交互作用的双因素方差分析
  - 有交互作用的双因素方差分析

一、单因素试验



试验指标——试验中要考察的指标.

因　　素——影响试验指标的条件.

- 因素
  - 可　控因素
  - 不可控因素

水　　平——因素所处的状态.

单因素试验——在一项试验中只有一个因素改变.

多因素试验——在一项试验中有多个因素在改变.

第一节　单因素方差分析

例4.1某灯泡厂用4种不同配料方案制成的的灯丝生产了4批灯泡，从每批灯泡中随机第抽取n_i $(i=1,2,3,4)$个灯泡，测得其使用寿命（小时）如下表. 试问4批灯泡的使用寿命是否有显著差异？

灯丝 \ 寿命 \ 灯泡	1	2	3	4	5	6	7	8
1	1600	1610	1650	1680	1700	1700	1780	
2	1500	1640	1400	1700	1750			
3	1640	1500	1600	1620	1640	1600	1740	1800
4	1510	1520	1530	1570	1640	1680		

本例中，试验指标是灯泡寿命，因素是配料方案, 有**4**种不同的配料方案(**4**个水平), 是单因素4水平的试验问题.

经过假设检验，如果认为4种灯丝配料方案有显著性差异, 工厂将采用使灯泡寿命长的那种配料方案, 如果无显著差异, 将采用最便宜的那种配料方案.

这相当于有四个独立的总体，从各个总体分别抽取样本，检验四个总体的均值是否相等.

对于单因素试验, 设单因素A有 r 个水平$A_1, A_2, \cdots\cdots, A_r$, 每种水平条件下的试验指标视为一个总体. 这相当于有 r 个总体$X_1, X_2, \cdots\cdots, X_r$. 因素水平$A_i$条件下重复进行$n_i$次试验，相当于从总体$X_i$抽取了 n_i 个样本. 记为$X_{i1}, X_{i2}, \cdots\cdots, X_{in_i}$. 令$n=\sum_{i=1}^{r} n_i$，它表示样本的总容量.

$$\bar{X}_{i\cdot}=\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}, \qquad \bar{X}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} X_{ij}.$$

样本观测数据如下表

因素水平	样本观测数据	总体
A_1	$X_{11}, X_{12}, \cdots\cdots, X_{1n_1}$	$X_1 \sim N(\mu_1, \sigma^2)$
A_2	$X_{21}, X_{22}, \cdots\cdots, X_{2n_2}$	$X_2 \sim N(\mu_2, \sigma^2)$
$\cdots$	$\cdots\cdots\cdots\cdots\cdots\cdots$	$\cdots$
A_r	$X_{r1}, X_{r2}, \cdots\cdots, X_{rn_r}$	$X_r \sim N(\mu_r, \sigma^2)$
r个	理论分析时视为随机变量	r个

注：因素A 以外的其他因素, 对指标的影响是相同的, 因此可以认为 r 个总体的方差是相等的. 因素A 对指标的影响体现在各总体均值μ_i的不同.

（一）单因素方差分析模型

$$\begin{cases} X_{ij} \sim N(\mu_i, \sigma^2), \quad i = 1,2,\cdots,r. \quad j = 1,2,\cdots,n_i. \\ X_{ij}\text{独立，} \ \mu_i, \sigma^2\text{未知.} \end{cases}$$

此模型的假设检验

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_r, \quad H_1: \mu_1, \mu_2, \cdots, \mu_r\text{不全相等.}$$

$$\boxed{\text{数据}X_{ij}\text{的差异}} \begin{cases} \text{原因} \qquad \boxed{\text{因素A}} \oplus \boxed{\text{其他因素}} \\ \text{大小} \qquad S_T = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2 \end{cases}$$

问题：S_T的分解 $\Rightarrow$ 与因素A有关的项 $\oplus$ 与其他因素有关的项

比较大小

引入几个记号，并对模型作等价变形

$$\begin{cases} \varepsilon_{ij} = X_{ij} - \mu_i, \quad n = \sum_{i=1}^{r} n_i, \quad \mu = \frac{1}{n}\sum_{i=1}^{r} n_i \mu_i. \\ \alpha_i = \mu_i - \mu, \qquad i = 1, 2, \cdots, r. \end{cases}$$

ε_{ij}是随机误差项，它是其他因素对总体X_i(因素水平A_i的条件下)的第j个样本的影响. μ称为一般平均, α_i称为因素水平A_i的效应. 得到如下等价模型.

单因素方差分析模型
$$\begin{cases} X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2) \\ \quad i = 1, 2, \cdots, r, \quad j = 1, 2, \cdots, n_i . \\ \sum_{i=1}^{r} n_i \alpha_i = 0. \\ \mu, \alpha_i, \sigma^2 \text{未知}. \end{cases}$$

模型的假设检验

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0, \quad H_1: \alpha_1, \cdots, \alpha_r \text{不全为零}.$$

（二）平方和分解公式

$$S_T = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left[(X_{ij} - \bar{X}_{i\cdot}) + (\bar{X}_{i\cdot} - \bar{X}) \right]^2$$

$$=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})^2+\sum_{i=1}^{r}n_i(\bar{X}_{i\cdot}-\bar{X})^2$$

$$+2\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})(\bar{X}_{i\cdot}-\bar{X})=S_e+S_A$$

其中 $S_e=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})^2,\qquad S_A=\sum_{i=1}^{r}n_i(\bar{X}_{i\cdot}-\bar{X})^2.$

$$\begin{cases} X_{ij}\sim N(\mu+\alpha_i,\sigma^2), \\ \bar{X}\sim N\left(\mu,\dfrac{\sigma^2}{n}\right), \\ \bar{X}_{i\cdot}\sim N\left(\mu+\alpha_i,\dfrac{1}{n_i}\sigma^2\right), \end{cases} \Longrightarrow \begin{cases} X_{ij}-\bar{X}\sim N\left(\alpha_i,\dfrac{n-1}{n}\sigma^2\right), \\ \bar{X}_{i\cdot}-\bar{X}\sim N\left(\alpha_i,\dfrac{n-n_i}{nn_i}\sigma^2\right), \\ X_{ij}-\bar{X}_{i\cdot}\sim N\left(0,\ \dfrac{n_i-1}{n_i}\sigma^2\right) \end{cases}$$

$$\begin{cases} E(S_e)=(n-r)\sigma^2, \\ E(S_A)=(r-1)\sigma^2+\sum_{i=1}^{r} n_i\alpha_i^2 \end{cases}$$

$$\begin{cases} E(S_e/(n-r))=\sigma^2, \\ E(S_A/(r-1))=\sigma^2+\dfrac{1}{r-1}\sum_{i=1}^{r} n_i\alpha_i^2 > \sigma^2. \end{cases}$$

S_e反映误差的波动，称为误差的偏差平方和；S_A称为因素的偏差平方和. 在假设H_0成立下，S_A反映误差的波动；若假设H_0不成立下, 则S_A反映了不同效应之间的差异(含误差).

$$S_T=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2$$

的分解

$$S_A=\sum_{i=1}^{r}n_i(\bar{X}_{i\cdot}-\bar{X})^2 \longrightarrow \boxed{\text{与因素A有关的项}}$$

$$\oplus$$

$$S_e=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})^2 \longrightarrow \boxed{\text{与其他因素有关的项}}$$

因素 A

$$S_A=\sum_{i=1}^{r}n_i(\bar{X}_{i\cdot}-\bar{X})^2$$

数据 X_{ij} 的差异的原因

$$S_T=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2$$

其他因素

$$S_e=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})^2$$

（三） 检验统计量与否定域

(1) S_e的分布 $\dfrac{1}{\sigma^2}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})^2 \sim \chi^2(n_i-1)$,

$$\frac{S_e}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})^2 \sim \chi^2(n-r)$$

$$\Rightarrow E\left(\frac{S_e}{\sigma^2}\right)=n-r,\quad E(S_e)=(n-r)\sigma^2.$$

(2) S_A的分布

若H_0成立，利用柯赫伦定理可以证明

$$\frac{S_T}{\sigma^2}\sim\chi^2(n-1),\qquad \frac{S_A}{\sigma^2}\sim\chi^2(r-1).$$

(3)检验统计量与否定域

$$\text{检验统计量}: F=\frac{S_A/r-1}{S_e/n-r}\overset{H_{0A}}{\sim}F(r-1,n-r)$$

$$\text{否定域}: W=\left\{(x_1,\cdots,x_n): F>F_\alpha(r-1\ ,\ n-r)\right\}$$

若$F>F_\alpha(r-1\ ,\ n-r)$，则认为因素取不同水平对指标影响显著.

$F>F_{0.01}(r-1\ ,\ n-r)$ 认为因素的影响高度显著，用$**$表示；

$F_{0.05}<F\le F_{0.01}$,认为因素的影响显著, 用$*$表示；

$F_{0.1}<F\le F_{0.05}$，认为因素有一定显著, 用$(*)$表示；

$F\le F_{0.1}$，认为因素的影响不显著, 无表示.

（四）单因素方差分析表

上述的分析过程可列成下表的形式

来源	平方和	自由度	均方和	F比	临界值	显著性
因素A	S_A	r-1	S_A/ r-1	$F=\dfrac{S_A/r-1}{S_e/n-r}$	F_α	
误差e	S_e	n-r	S_e/ n-r			
总和	S_T	n-1				

数据S_e，S_A，S_T的计算常按下列顺序进行

$$S_T=\sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij}^{\ 2}-\frac{1}{n}(\sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij})^2=\sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij}^{\ 2}-n\bar{X}^2$$

$$S_A=\sum_{i=1}^{r}\frac{1}{n_i}(\sum_{j=1}^{n_i}X_{ij})^2-\frac{1}{n}(\sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij})^2=\sum_{i=1}^{r}n_i\bar{X}_{i\cdot}^2-n\bar{X}^2$$

$$S_e=S_T-S_A.$$

当数据X_{ij}较复杂时, 可对其进行线性变换

$$Y_{ij}=b(X_{ij}-a)$$

其中a,b为适当的常数且$b\neq 0$，使得新数据Y_{ij}较简单. 根据Y_{ij}计算的F值不变，特别$b=1$时，S_e,S_A及S_T的值也不变.

在例1中，$r=4$，n_i分别为7,5,8,6. $n=\sum_{i=1}^{4} n_i=26$.计算得

来源	平方和	自由度	均方和	F比	临界值	显著性
因素A	39776	3	13258	1.638	$F_{0.1}=$ 2.35	
误差e	178089	22	8095			
总和	217865	25				

由于$F < F_{0.1}$，故接受H_0，即认为灯丝的配料方案对灯泡的使用寿命无显著性影响，显著性栏里不用表示.

（五） 未知参数的估计

(1) 点估计

$$\hat{\mu}_i=\bar{X}_{i\cdot}\quad,\quad \hat{\mu}=\bar{X}\quad,\quad \hat{\alpha}_i=\bar{X}_{i\cdot}-\bar{X}\quad,\quad \hat{\sigma}^2=\frac{S_e}{n-r}$$

分别为未知参数μ_i ,μ ,α_i ,σ^2的点估计，且它们都是无偏估计. $i=1,2,\cdots,r.$

(2) 区间估计

由于$S_e/\sigma^2\sim\chi^2(n-r)$，所以$\sigma^2$的置信度为$1-\alpha$的置信区间为$\left(\dfrac{S_e}{\chi^2_{\alpha/2}(n-r)},\ \dfrac{S_e}{\chi^2_{1-\alpha/2}(n-r)}\right).$

可以证明

$$\frac{\bar{X}_{i\cdot}-\mu_i}{\sqrt{S_e/n_i(n-r)}}\sim t(n-r),$$

所以μ_i的置信度为$1-\alpha$的置信区间为

$$\left[\bar{X}_{i\cdot}-\delta,\ \ \bar{X}_{i\cdot}+\delta\right].$$

其中 $\delta=t_{\alpha/2}(n-r)\sqrt{S_e/n_i(n-r)}.\quad i=1,2,\cdots,r.$

练习题:

推导出未知参数$\mu,\ \alpha_i$的置信度为$1-\alpha$的置信区间.