第五章　方差分析与试验设计

第5.1节　单因素方差分析

第5.2节　两因素方差分析

在生产实践和科学试验种，经常发现试验条件的不同，得到的实验结果也不同，有时在相同的试验条件下，也会得到不同的试验结果。试验结果之间的差异是由什么原因造成的呢？是由于试验条件的不同所引起的？还是由于试验误差引起的?

如果是试验条件不同引起的，对试验指标最有利的试验条件应该如何确定?

本章主要介绍单因素试验的方差分析和两因素的方差分析。

第5.1节　单因素方差分析

一、数学模型

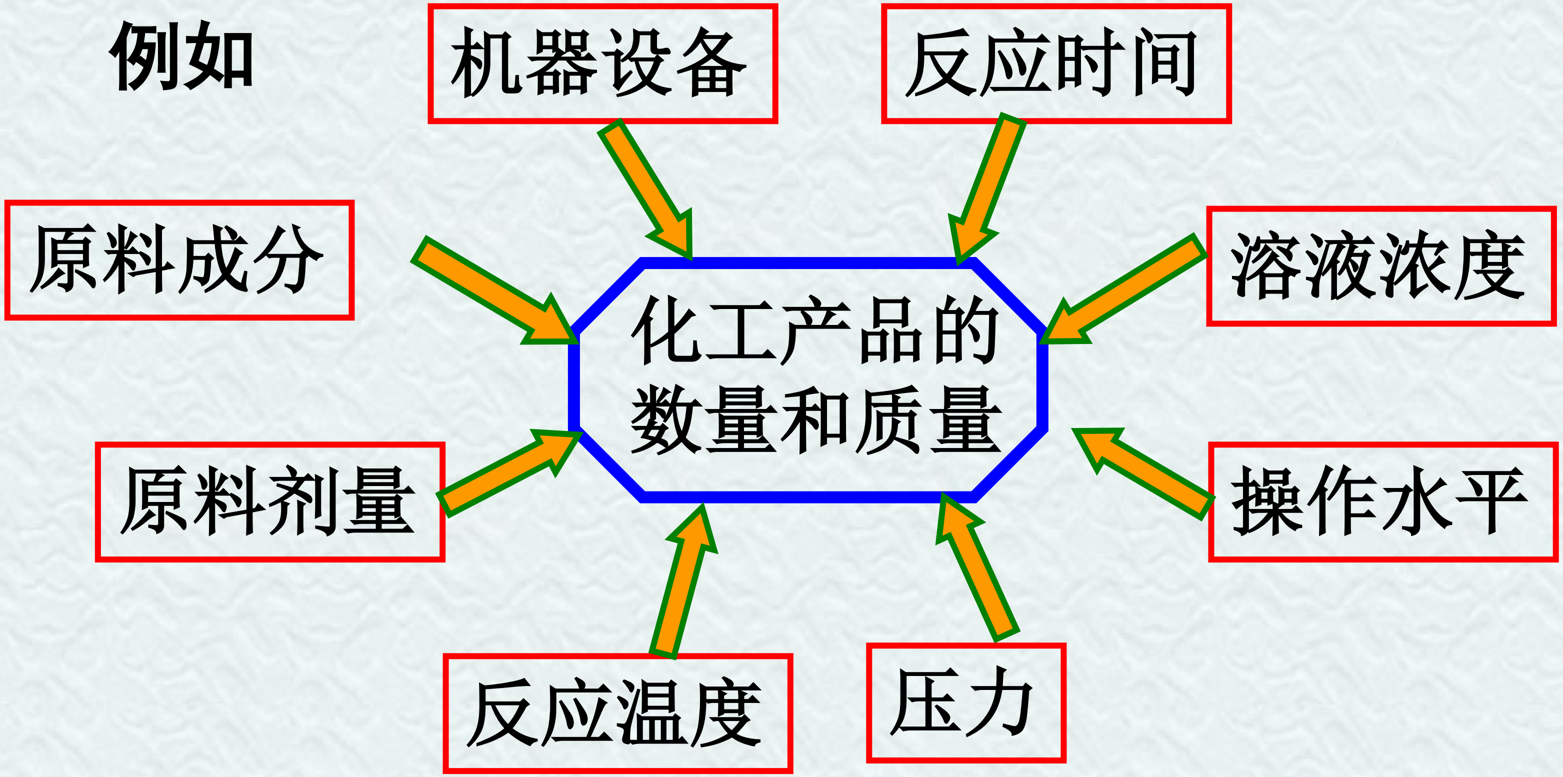
二、离差平方和分解与显著性检验

三、参数估计

一、数学模型

1、问题引入

例如



方差分析——根据试验的结果进行分析,鉴别各个有关因素对试验结果的影响程度.

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.

例1 设有三台机器, 用来生产规格相同的铝合金薄板.取样,测量薄板的厚度精确至千分之一厘米.得结果如下表所示.

表 铝合金板的厚度

机器I	机器II	机器III
0.236	0.257	0.258
0.238	0.253	0.264
0.248	0.255	0.259
0.245	0.254	0.267
0.243	0.261	0.262

试验指标:薄板的厚度 **因素**:机器

水平:不同的三台机器是因素的三个不同的水平

表　铝合金板的厚度

机器I	机器II	机器III
0.236	0.257	0.258
0.238	0.253	0.264
0.248	0.255	0.259
0.245	0.254	0.267
0.243	0.261	0.262

假定除机器这一因素外,其他条件相同,属于**单因素试验**.

试验目的:考察各台机器所生产的薄板的厚度有无显著的差异.即考察机器这一因素对厚度有无显著的影响.

例2 下表列出了随机选取的、用于计算器的四种类型的电路的响应时间（以毫秒计）.

表 电路的响应时间

类型I	类型II	类型III	类型IV
19	20	16	18
22	21	15	22
20	33	18	19
18	27	26	

试验指标:电路的响应时间 **因素**:电路类型

水平:四种电路类型为因素的四个不同的水平

单因素试验

试验目的:考察电路类型这一因素对响应时间有无显著的影响.

例3 一火箭用四种燃料,三种推进器作射程试验.每种燃料与每种推进器的组合各发射火箭两次,得射程如下（以海里计）.

表 火箭的射程

推进器(B)		B_1	B_2	B_3
燃料(A)	A_1	58.2 52.6	56.2 41.2	65.3 60.8
	A_2	49.1 42.8	54.1 50.5	51.6 48.4
	A_3	60.1 58.3	70.9 73.2	39.2 40.7
	A_4	75.8 71.5	58.2 51.0	48.7 41.4

表 火箭的射程

推进器(*B*)		B_1	B_2	B_3
燃料(*A*)	A_1	58.2 52.6	56.2 41.2	65.3 60.8
	A_2	49.1 42.8	54.1 50.5	51.6 48.4
	A_3	60.1 58.3	70.9 73.2	39.2 40.7
	A_4	75.8 71.5	58.2 51.0	48.7 41.4

试验指标:射程 **因素**:推进器和燃料

水平:推进器有3个,燃料有4个

双因素试验

试验目的:因素对火箭射程的有无显著影响

回顾例1 铝合金板的厚度

机器I	机器II	机器III
0.236	0.257	0.258
0.238	0.253	0.264
0.248	0.255	0.259
0.245	0.254	0.267
0.243	0.261	0.262

问题分析

在因素的不同水平下，试验数据之间存在差异，即使在同一水平下，试验数据之间同样存在差异。那么，这种差异到底是由因素的水平变化引起的？还是由于随机误差的干扰引起的？

在每一个水平下进行独立试验,结果是一个随机变量.将数据看成是来自三个总体的样本值.

设总体均值分别为 μ_1, μ_2, μ_3.

检验假设 $H_0: \mu_1 = \mu_2 = \mu_3$,

$H_1: \mu_1, \mu_2, \mu_3$ 不全相等.

进一步假设各总体均为正态变量,且各总体的方差相等,但参数均未知.

问题——检验同方差的多个正态总体均值是否相等.

解决方法——方差分析法,一种统计方法.

2、数学模型

设因素A有r个水平$A_1, A_2, \cdots, A_r$,在水平$A_i(j=1,2,\cdots,r)$下,进行$n_i(n_i \geq 2)$次独立试验,同时假设

(1)各个水平$A_i(i=1,2,\cdots,r)$下的样本$X_{i1}, X_{i2}, \cdots, X_{in_i}$来自具有相同方差$\sigma^2$,均值分别为$\mu_i(i=1,2,\cdots,r)$的正态总体$N(\mu_i,\sigma^2)$,$\mu_i$与$\sigma^2$均未知;

（2）不同水平A_i下的样本之间相互独立.

因为$X_{ij} \sim N(\mu_i,\sigma^2)$，所以$X_{ij}-\mu_i \sim N(0,\sigma^2)$.

记$X_{ij}-\mu_i=\varepsilon_{ij}$,表示随机误差,那么$X_{ij}$可写成

$$\left.\begin{array}{l} X_{ij} = \mu_i + \varepsilon_{ij}, \\ \varepsilon_{ij} \sim N(0,\sigma^2), \varepsilon_{ij}\text{相互独立}, \\ j = 1,2,\cdots,n_i, i = 1,2,\cdots,r, \\ \mu_i\text{与}\sigma^2\text{均未知} \end{array}\right\}$$

上式就是单因素试验方差分析的数学模型

需要解决的问题

1.检验假设 $H_0: \mu_1 = \mu_2 = \cdots = \mu_r,$

$H_1: \mu_1, \mu_2, \cdots, \mu_r$不全相等.

2.估计未知参数$\mu_1, \mu_2, \cdots, \mu_r, \sigma^2$.

3、数学模型的等价形式

设$n=\sum_{i=1}^{n} n_i$,　记 $\boxed{\mu=\frac{1}{n}\sum_{i=1}^{r} n_i\mu_i}$

总平均

称为水平A_i的效应,表示水平A_i下的总体均值与总平均的差异.

$\boxed{\alpha_i=\mu_i-\mu,}$　　$i=1,2,\cdots,r$

且 $n_1\alpha_1+n_2\alpha_2+\cdots+n_r\alpha_r=0.$

$$\sum_{i=1}^{r} n_i\alpha_i=\sum_{i=1}^{r} n_i(\mu_i-\mu)$$

$$=n\mu-\left(\sum_{i=1}^{r} n_i\right)\mu=0$$

原数学模型

$$\left.\begin{aligned}&X_{ij}=\mu_i+\varepsilon_{ij},\\&\varepsilon_{ij}\sim N(0,\sigma^2),\varepsilon_{ij}\text{相互独立},\\&j=1,2,\cdots,n_i,i=1,2,\cdots,r,\\&\mu_i\text{与}\sigma^2\text{均未知}\end{aligned}\right\}$$

改写为

$$\left.\begin{aligned}&X_{ij}=\mu+\alpha_i+\varepsilon_{ij},\\&\varepsilon_{ij}\sim N(0,\sigma^2),\varepsilon_{ij}\text{相互独立},\\&i=1,2,\cdots,n_i,i=1,2,\cdots,r,\\&\sum_{j=1}^{r}n_i\alpha_i=0\end{aligned}\right\}$$

检验假设

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_r,$$

$$H_1: \mu_1, \mu_2, \cdots, \mu_r \text{至少有两个不相等.}$$

等价于

检验假设

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0,$$

$$H_1: \alpha_1, \alpha_2, \cdots, \alpha_r \text{至少有1个不为0.}$$

二、离差平方和的分解与显著性检验

为了进行上述假设检验，需要进行一系列变化。

1、总离差平方和分解公式

$$\bar{X}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}X_{ij},i=1,\cdots,r$$ —水平A_i下的样本平均值，称为组内平均

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij}$$ —总平均

$$Q_T=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2$$ —总离差平方和（总变差）

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

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

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

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

$$= 0$$

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

$$= Q_E + Q_A$$

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

—组内离差平方和，反映试验误差引起的数据波动

$$Q_A = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{X}_i - \bar{X})^2 = \sum_{i=1}^{r} n_i(\bar{X}_i - \bar{X})^2$$

$$= \sum_{i=1}^{r} n_i \bar{X}_i^{\,2} - n\bar{X}^2$$

—组间离差平方和，反映了因素水平变化和试验误差引起的数据波动

2、分解式的统计特性

令$\bar{\varepsilon}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\varepsilon_{ij}$, $\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\varepsilon_{ij}$得

$$Q_E = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(\mu+\alpha_i+\varepsilon_{ij}-\mu-\alpha_i-\bar{\varepsilon}_i\right)^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(\varepsilon_{ij}-\bar{\varepsilon}_i\right)^2$$

$$Q_A = \sum_{i=1}^{r}n_i\left(\mu+\alpha_i+\bar{\varepsilon}_i-\mu-\bar{\varepsilon}\right)^2 = \sum_{i=1}^{r}n_i\left(\alpha_i+\bar{\varepsilon}_i-\bar{\varepsilon}\right)^2$$

$$= \sum_{i=1}^{r}n_i\alpha_i^2 + \sum_{i=1}^{r}n_i\left(\bar{\varepsilon}_i-\bar{\varepsilon}\right)^2 + 2\sum_{i=1}^{r}n_i\alpha_i\left(\bar{\varepsilon}_i-\bar{\varepsilon}\right)$$

由于$\varepsilon_{ij} \sim N\left(0,\sigma^2\right)$, $\bar{\varepsilon}_i \sim N\left(0,\dfrac{\sigma^2}{n_i}\right)$, $\bar{\varepsilon} \sim N\left(0,\dfrac{\sigma^2}{n}\right)$

$$i = 1,2,\cdots,r, j = 1,2,\cdots,n_i$$

回忆$\dfrac{nS_n^2}{\sigma^2} \sim \chi^2(n-1)$

$$EQ_E = E\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(\varepsilon_{ij} - \bar{\varepsilon}_i\right)^2 = \sum_{i=1}^{r}\left(n_i - 1\right)\sigma^2 = (n-r)\sigma^2$$

$$EQ_A = \sum_{i=1}^{r} n_i\alpha_i^2 + E\left[\sum_{i=1}^{r} n_i\bar{\varepsilon}_i^2 - n\bar{\varepsilon}^2\right] + 2\sum_{i=1}^{r} n_i\alpha_i\left[E(\bar{\varepsilon}_i) - E(\bar{\varepsilon})\right]$$

$$= \sum_{i=1}^{r} n_i\alpha_i^2 + \left(r-1\right)\sigma^2$$

所以

$$E\left(\frac{Q_E}{n-r}\right)=\sigma^2$$

$$E\left(\frac{Q_A}{r-1}\right)=\sigma^2+\frac{1}{r-1}\sum_{i=1}^{r}n_i\alpha_i^2$$

3、构造统计量及统计量的分布

当H_0成立时，即$\alpha_1=\alpha_2=\cdots=\alpha_r=0$时，

$$E\frac{Q_A}{r-1}=E\frac{Q_E}{n-r}=\sigma^2,\text{ 反之，}E\frac{Q_A}{r-1}\geq E\frac{Q_E}{n-r}$$

从而当H_0不成立时,比值

$$\frac{Q_A/(r-1)}{Q_E/(n-r)}\overset{\text{def}}{=}\frac{\overline{Q}_A}{\overline{Q}_E}=F$$

有偏大的趋势,所以F可作为检验H_0的统计量。

下面我们先求出在H_0成立条件下,统计量F的概率分布.

当H_0成立时,所有的α_i都等于零, 所以

$$X_{ij}=\mu+\varepsilon_{ij}$$

代入Q_T,Q_E,和Q_A的表达式,可得:

$$Q_T=\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(\varepsilon_{ij}-\bar{\varepsilon}\right)^2=\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(\varepsilon_{ij}-\bar{\varepsilon}_i\right)^2+\sum_{i=1}^{r}\left[\sqrt{n_i}\left(\bar{\varepsilon}_i-\bar{\varepsilon}\right)\right]^2$$

又因为$\sum_{i=1}^{r}\sum_{j=1}^{n_i}\varepsilon_{ij}^2=\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(\varepsilon_{ij}-\bar{\varepsilon}\right)^2+n\bar{\varepsilon}^2$

故$\sum_{i=1}^{r}\sum_{j=1}^{n_i}\varepsilon_{ij}^2=\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(\varepsilon_{ij}-\bar{\varepsilon}_i\right)^2+\sum_{i=1}^{r}\left[\sqrt{n_i}\left(\bar{\varepsilon}_i-\bar{\varepsilon}\right)\right]^2+\left(\sqrt{n}\bar{\varepsilon}\right)^2$

上式两边同除以σ^2, 左边$\dfrac{1}{\sigma^2}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\varepsilon_{ij}^2$是自由度为$n$ 的χ^2变量，

右边三项分别为:

$$\frac{1}{\sigma^2}Q_E = \frac{1}{\sigma^2}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(\varepsilon_{ij} - \overline{\varepsilon}_i\right)^2$$有r个约束条件

$$\sum_{j=1}^{n_i}\left(\varepsilon_{ij} - \overline{\varepsilon}_i\right) = 0\left(i = 1,2,\cdots,r\right)$$所以二次型$\dfrac{1}{\sigma^2}Q_E$的秩为$n-r$.

$$\frac{1}{\sigma^2}Q_A = \frac{1}{\sigma^2}\sum_{i=1}^{r}\left[\sqrt{n_i}\left(\overline{\varepsilon}_i - \overline{\varepsilon}\right)\right]^2$$有一个约束条件

$$\sum_{i=1}^{r}n_i\left(\overline{\varepsilon}_i - \overline{\varepsilon}\right) = 0,$$所以$\dfrac{Q_E}{\sigma^2}$的秩为$r-1$.

$\frac{1}{\sigma^2}\left(\sqrt{n}\bar{\varepsilon}\right)^2$的自由度是1.它们的自由度之和为$\left(n-r\right)+$

$\left(r-1\right)+1=n$，由定理1.7（柯赫伦分解定理）的充分

条件知,$\frac{Q_E}{\sigma^2}\sim\chi^2(n-r),\frac{Q_A}{\sigma^2}\sim\chi^2(r-1)$且相互独立。

所以在H_0成立的条件下,

$$F=\frac{Q_A/\sigma^2\left(r-1\right)}{Q_E/\sigma^2\left(n-r\right)}=\frac{Q_A/\left(r-1\right)}{Q_E/\left(n-r\right)}=\frac{\bar{Q}_A}{\bar{Q}_E}\sim F\left(r-1,n-r\right)$$

4、假设检验问题的拒绝域

检验假设 $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0,$

$H_1: \alpha_1, \alpha_2, \cdots, \alpha_r$不全为零.

H_0为真时， $Q_A / \sigma^2 \sim \chi^2(r-1)$

$$E\left(\frac{Q_A}{r-1}\right) = \sigma^2, \text{即} \frac{Q_A}{r-1} \text{是} \sigma^2 \text{的无偏估计.}$$

H_1为真时， $\sum_{i=1}^{r} n_i \alpha_i^2 > 0,$

$$E\left(\frac{Q_A}{r-1}\right) = \sigma^2 + \frac{1}{r-1}\sum_{i=1}^{r} n_i \alpha_i^2 > \sigma^2.$$

因为$E(Q_E)=(n-r)\sigma^2$，所以$E(\frac{Q_E}{n-r})=\sigma^2$，

即不管H_0是否为真,$Q_E/(n-r)$都是σ^2的无偏估计.

$$F=\frac{Q_A/(r-1)}{Q_E/(n-r)}.$$

1.分子和分母相互独立;

2.分母Q_E的数学期望始终是σ^2;

3.H_0为真时,分子的期望为σ^2,H_0不真时,分子取值有偏大的趋势.

于是拒绝域为 $F=\frac{Q_A/(r-1)}{Q_E/(n-r)}\geq F_\alpha(r-1,n-r).$

单因素试验方差分析表

方差来源	平方和	自由度	平均离差	F 比
组　间	Q_A	$r-1$	$\bar{Q}_A=\dfrac{Q_A}{r-1}$	$F=\bar{Q}_A/\bar{Q}_E$
组　内	Q_E	$n-r$	$\bar{Q}_E=\dfrac{Q_E}{n-r}$	
总　和	Q_T	$n-1$		

为了计算方便，记 $Q=\sum_{i=1}^{r}\dfrac{1}{n_i}(\sum_{j=1}^{n_i}X_{ij})^2,$

$P=\dfrac{1}{n}(\sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij})^2,\quad R=\sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij}^2,$则

$$Q_A=Q-P,\quad Q_E=R-Q,\quad Q_T=R-P$$

例4(p156例5.3) 某种型号化油器的原喉管结构油耗较大，为节省能源，设想了两种改进方案以降低油耗指标一比油耗，现对用各种结构的喉管制造的化油器分别测得如下表数据

指标 水平	比 油 耗							
A_1：原结构	231.0	232.8	227.6	228.3	224.7	225.5	229.3	230.3
A_2：改进方案I	222.8	224.5	218.5	220.2				
A_3：改进方案II	224.3	226.1	221.4	223.6				

在显著性水平$\alpha = 0.01$条件下进行方差分析,判断喉管的结构对比油耗的影响是否显著.

解 为了计算方便，首先每个原始数据减220，通过计算P,R,Q,计算其离差平方和.

其中 $r=3, n_1=8, n_2=n_3=4, n=16,\quad Q_A'=155.64$

$Q_T'=240.98,\quad Q_E'=85.34$

方差分析表

方差来源	平方和	自由度	均　方	F　比
组　内	**155.64**	**2**	**77.82**	**11.86**
组　间	**85.34**	**13**	**6.56**	
总　和	**240.98**	**15**		

$F=11.86>F_{0.01}(2,13)=6.70.$ 在水平0.01下拒绝H_0.

喉管结构对比油耗有显著影响，其中改进方案一好.

三、参数估计

由于 $E(Q_E/(n-r))=\sigma^2$ $\qquad \hat{\sigma}^2=Q_E/(n-r)$

$E(\bar{X})=\mu$ $\qquad \hat{\mu}=\bar{X}$

$E(\bar{X}_i)=\mu_i, i=1,2,\cdots,r$ $\qquad \hat{\mu}_i=\bar{X}_i$

$E(\bar{X}_i-\bar{X})=\mu_i-\mu, i=1,2,\cdots,r$ $\qquad \hat{\alpha}_i=\bar{X}_i-\bar{X}$

无偏估计

若拒绝H_0,需对两总体$N(\mu_i,\sigma^2)$, $N(\mu_k,\sigma^2)$的均值差$\mu_i-\mu_k=\alpha_i-\alpha_k$作出区间估计.

因为$E(\bar{X}_i-\bar{X}_k)=\mu_i-\mu_k$,

$$D(\bar{X}_i-\bar{X}_k)=\sigma^2\left(\frac{1}{n_i}+\frac{1}{n_k}\right),$$

$\bar{X}_i - \bar{X}_k$与$\hat{\sigma}^2 = \bar{Q}_E = Q_E / (n-r)$独立,

所以
$$\frac{(\bar{X}_i - \bar{X}_k) - (\mu_i - \mu_k)}{\sqrt{\bar{Q}_E\left(\frac{1}{n_i} + \frac{1}{n_k}\right)}}$$
$$= \frac{(\bar{X}_i - \bar{X}_k) - (\mu_k - \mu_k)}{\sigma\sqrt{1/n_i + 1/n_k}} \Bigg/ \sqrt{\frac{Q_E}{\sigma^2} \Big/ (n-r)} \sim t(n-r).$$

均值差$\mu_i - \mu_k = \alpha_i - \alpha_k$的置信水平为$1-\alpha$的置信区间为

$$\left(\bar{X}_i - \bar{X}_k \pm t_{\alpha/2}(n-r)\sqrt{\bar{Q}_E\left(\frac{1}{n_i} + \frac{1}{n_k}\right)}\right).$$

例5(p163例5.4) 有5种油菜品种，分别在4块试验田上种植，所得亩产量如下表：

田块 / 品种	1	2	3	4
A_1	256	222	280	298
A_2	244	300	290	275
A_3	250	277	230	322
A_4	288	280	315	259
A_5	206	212	220	212

1、不同品种对亩产量有无显著影响

2、求$\mu_1 - \mu_5$的置信度为0.95的置信区间.

解 令X_{ij}表示第i个品种在第j块试验田的亩产量，

$i=1,2,\cdots,5,\ \ n_1=n_2=\cdots=n_5=4, n=20,$计算

$$R=\sum_{i=1}^{5}\sum_{j=1}^{4}x_{ij}^2=1395472, Q=\sum_{i=1}^{5}\frac{1}{4}(\sum_{j=1}^{4}x_{ij})^2=1383980.5$$

$$P=\frac{1}{20}(\sum_{i=1}^{5}\sum_{j=1}^{5}x_{ij})^2=1370784.8,\quad Q_T=R-P=24687.2$$

$$Q_A=Q-P=13195.7,\quad Q_E=Q_T-Q_A=11491.5$$

$$F=\frac{Q_A/4}{Q_E/15}=4.31>F_{0.05}(4,15)=3.06$$

即品种对产量有显著影响.

又因为 $\bar{x}_1 = 264, \quad \bar{x}_5 = 212.5$

$$t_{0.025}(n-r) = t_{0.025}(15) = 2.1315,$$

$$\bar{Q}_E = Q_E / n - r = 11491.5/15 = 766.1$$

$$t_{0.025}(15)\sqrt{\bar{Q}_E\left(\frac{1}{n_1}+\frac{1}{n_5}\right)} = 2.1315\times 19.6$$

$\mu_1 - \mu_5$的置信水平为0.95的置信区间为

$$(264 - 212.5 \pm 2.1315\times 19.6) = (9.7,\ 93.3).$$

四、小结

1.随机试验:单因素试验、多因素试验

2.单因素试验方差分析步骤

(1)建立数学模型;

(2)分解平方和;

(3)研究统计特性;

(4)进行假设检验;

(5)估计未知参数.

Thank You!