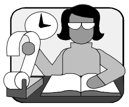


数学实验

Experiments in Mathematics

实验10 方差分析

清 华 大 学 数 学 科 学 系

方差分析

1、方差分析的基本概念、实例

2、单因素方差分析

3、双因素方差分析

4、**MATLAB**统计工具箱**(Statistics Toolbox)**的使用

单因素方差分析示例——灯泡寿命

用4种工艺生产灯泡，从每种工艺制成的灯泡中各抽取若干个测量其寿命(小时)，如表，试推断这几种工艺制成的灯泡寿命是否有显著差异。

序号 \ 工艺	A_1	A_2	A_3	A_4
1	1620	1580	1460	1500
2	1670	1600	1540	1550
3	1700	1640	1620	1610
4	1750	1720		1680
5	1800			
平均	1708	1635	1540	1585

从平均值看A_1最大，但其中最小者比A_4中最大者要小。

数据间的差异有两个原因：不同工艺造成的系统差异；同一工艺内的随机差异。

双因素方差分析示例---小麦产量

为分析4种化肥和3个小麦品种对小麦产量的影响，把试验田等分成24块，对种子和化肥的每一组合种植2块田，产量如下表，问品种、化肥对小麦产量有无显著影响，二者的交互作用对小麦产量有无显著影响。

小麦产量试验数据（公斤）

品种 \ 化肥	A_1	A_2	A_3	A_4
B_1	173, 172	174, 176	177, 179	172, 173
B_2	175, 173	178, 177	174, 175	170, 171
B_3	177, 175	174, 174	174, 173	169, 169

方差分析的基本概念

指标：关心的试验结果

灯泡的寿命，小麦的产量

因素：需要考察、可以控制的条件

灯泡寿命中的**4**个工艺　　小麦产量中的品种和化肥用量

单因素　　双因素

水平：因素所设定的状态

灯泡寿命：工艺**4**水平

小麦产量：品种**3**水平，化肥**4**水平

单因素方差分析（因素**A**）

数学模型

r 个水平$A_1, A_2, \ldots A_r$，A_i下总体$x_i \sim N(\mu_i,\sigma^2), i=1,\cdots r$，$\mu_i,\sigma^2$ 未知。

x_i中抽取容量为 n 的样本$x_{ij} \sim N(\mu_i,\sigma^2), i=1,\cdots,r,\ j=1,\cdots n$且相互独立。

数据表格

	A_1	A_2	$\cdots$	A_r
1	x_{11}	x_{21}	$\cdots$	X_{r1}
2	x_{12}	x_{22}	$\cdots$	x_{r2}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n	x_{1n}	x_{2n}	$\cdots$	x_{rn}

判断A的r个水平对指标有无显著影响，等价于：

假设检验：H_0：$\mu_1=\mu_2=\ldots=\mu_r$；H_1：$\mu_1, \mu_2, \ldots \mu_r$不全相等。

单因素方差分析------示例形式

灯泡寿命数据（若每种工艺灯泡数量相同）

序号 \ 工艺	A_1	A_2	A_3	A_4
1	1620	1580	1460	1500
2	1670	1600	1540	1550
3	1700	1640	1620	1610
4	1750	1720	1840	1680
5	1800	1840	1840	1840

单因素：生产工艺

水平：**4个: A_1 A_2 A_3 A_4**

样本量：每组5个数据

假设检验：H_0：$\mu_1=\mu_2=\mu_3=\mu_4$；H_1：μ_1,μ_2,μ_3,μ_4不全相等。

单因素方差分析----数学模型

$x_{ij}=\mu_i+\varepsilon_{ij},\quad i=1,\cdots r,\ j=1,\cdots n,\ \varepsilon_{ij}\sim N(0,\sigma^2)$ 且相互独立

$\mu=\frac{1}{r}\sum_{i=1}^{r}\mu_i$ ~ 总均值　　$\alpha_i=\mu_i-\mu$ ~ A_i对指标的效应

模型
$$\begin{cases} x_{ij}=\mu+\alpha_i+\varepsilon_{ij} \\ \sum_{i=1}^{r}\alpha_i=0 \\ \varepsilon_{ij}\sim N(0,\sigma^2),\quad i=1,\cdots r,\ j=1,\cdots n, \end{cases}$$

原假设为 $H_0:\alpha_1=\alpha_2=\cdots=\alpha_r=0$（略去备选假设）

统计分析

$\bar{x}_i=\frac{1}{n}\sum_{j=1}^{n}x_{ij}$(组平均值)，　$\bar{x}=\frac{1}{r}\sum_{i=1}^{r}\bar{x}_i$(总平均值)

$S=\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2$(总偏差)　　$S=S_A+S_E$(S的分解)

$S_A=\sum_{i=1}^{r}n(\bar{x}_i-\bar{x})^2$(组间平方和)　　$S_E=\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2$(组内平方和)

$ES_E=r(n-1)\sigma^2$，　$ES_A=(r-1)\sigma^2+\sum_{i=1}^{r}n\alpha_i^2$

$H_0:\alpha_1=\alpha_2=\cdots=\alpha_r=0$ (A的r个水平对指标无显著影响)

若 H_0 成立　$\frac{S_A/r-1}{S_E/r(n-1)}\approx 1$，　$F=\frac{S_A/r-1}{S_E/r(n-1)}\sim F(r-1,r(n-1))$

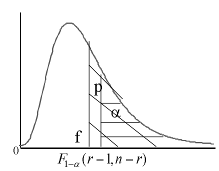
显著性水平：α
检验规则：$F<F_{1-\alpha}(r-1,r(n-1))$时接受 H_0，否则拒绝。

单因素方差分析表

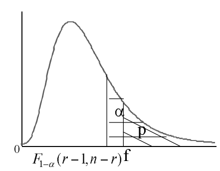
$H_0:\alpha_1=\alpha_2=\cdots=\alpha_r=0$

方差来源	平方和	自由度	平方均值	F 值	概率
因素A	S_A	$r-1$	$\bar{S}_A=S_A/r-1$	$f=\bar{S}_A/\bar{S}_E$	$p=P\{F>f\}$
误差	S_E	$r(n-1)$	$\bar{S}_E=S_E/r(n-1)$		
总和	S	$rn-1$			

f~F的样本值



若$f<F_{1-\alpha}(r-1,n-r)$
(即$p>\alpha$),则接受H_0



若$f>F_{1-\alpha}(r-1,n-r)$
(即$p<\alpha$),则拒绝H_0

单因素方差分析

$H_0:\alpha_1=\alpha_2=\cdots=\alpha_r=0$ (A的r个水平对指标无显著影响)

α=0.01，拒绝 H_0 ---影响非常显著；

α=0.01，不拒绝 H_0，但取α=0.05，拒绝 H_0 ----影响显著；

α=0.05，不拒绝 H_0----无显著影响。

单因素方差分析----MATLAB实现

命令：**p=anova1(x)**

适用于各组样本容量相同的单因素方差分析。

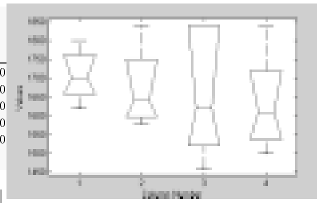
输入：
x为**n**行**r**列的数据矩阵，**n**为样本容量，**r**为水平数。

输出：$p=P\{F>f\}$

方差来源	平方和	自由度	平方均值	F 值
因素A	S_A	$r-1$	$\bar{S}_A=S_A/r-1$	$f=\bar{S}_A/\bar{S}_E$
误差	S_E	$r(n-1)$	$\bar{S}_E=S_E/r(n-1)$	
总和	S	$rn-1$		

表 1　灯泡寿命试验数据

序号＼工艺	A_1	A_2	A_3	A_4
1	1620	1580	1460	1500
2	1670	1600	1540	1550
3	1700	1640	1620	1610
4	1750	1720	1840	1680
5	1800	1840	1840	1840



假设检验：H_0：$\mu_1=\mu_2=\mu_3=\mu_4$

结果：**p =0.8351**

结论：接受H_0, 不同工艺无显著影响。

ANOVA Table

Source	SS	df	MS	F
Columns	1.368e+004	3	4560	0.2855
Errors	2.555e+005	16	1.597e+004	
Total	2.692e+005	19		

调用lamp.m (case 1)

命令：**p=anova1(x,group)**　假设检验：H_0：$\mu_1=\mu_2=\mu_3=\mu_4$

用于各组样本容量不同的单因素方差分析。

表 2　灯泡寿命试验数据

序号＼工艺	A_1	A_2	A_3	A_4
1	1620	1580	1460	1500
2	1670	1600	1540	1550
3	1700	1640	1620	1610
4	1750	1720		1680
5	1800			

ANOVA Table

Source	SS	df	MS	F
Columns	6.282e+004	3	2.094e+004	4.061
Errors	6.188e+004	12	5157	
Total	1.247e+005	15		

输入：要求**r**个水平样本值顺序排放，

x=[$x_{11}\ldots x_{1n1}\ x_{21}\ldots x_{2n2}\ldots x_{r1}\ldots x_{nr}$]

group=[1 1 1 1 1 2 2 2 2 3 3 3 4 4 4 4]

结果：**p =0.0331**

结论：拒绝H_0, 不同工艺有显著影响。

调用 lamp.m (case 2)

多重比较

假设检验：H_0：$\mu_1=\mu_2=\mu_3=\mu_4$

结论：拒绝H_0, 不同工艺有显著影响。

序号＼工艺	A_1	A_2	A_3	A_4
1	1620	1580	1460	1500
2	1670	1600	1540	1550
3	1700	1640	1620	1610
4	1750	1720		1680
5	1800			
平均	1708	1635	1540	1585

哪几种工艺有显著影响？　两总体的假设检验 (ttest2)

原假设	$\mu_1=\mu_2$	$\mu_1=\mu_3$	$\mu_1=\mu_4$
H	0	1	1
p	0.1459	0.0202	0.0408

A_1与A_3, A_4有显著差异（α=0.05），但与A_2无显著差异

双因素方差分析

为考察某指标（如小麦产量）受两个因素A（化肥）, B（品种）影响的显著性，

将A,B各划分几个水平；

每个水平组合作若干次试验；

对试验数据进行方差分析；

检验因素A,B是否分别对指标有显著影响，

以及两因素是否对指标有显著的交互影响。

双因素方差分析

数学模型　**两个因素：A和B**

A：r个水平$A_1, A_2, \ldots A_r$，　B：s个水平$B_1, B_2, \ldots B_s$

水平组合（A_i, B_j）：

总体 $x_{ij} \sim N(\mu_{ij},\sigma^2),\ i=1,\cdots r,\ j=1,\cdots s$

（A_i, B_j）下作了t个试验，所得结果记作x_{ijk}，x_{ijk} 服从$N(\mu_{ij},\sigma^2), i=1,\cdots r,\ j=1,\cdots s, k=1,\cdots t$，且相互独立

	A_1	A_2	…	A_r
B_1	$x_{111},\ldots x_{11t}$	$x_{211}\ldots x_{21t}$	…	$x_{r11},\ldots x_{r1t}$
B_2	$x_{121},\ldots x_{12t}$	$x_{221},\ldots x_{22t}$	…	$x_{r21},\ldots x_{r2t}$
…	…	…	…	…
B_s	$x_{1s1},\ldots x_{1st}$	$x_{2s1},\ldots x_{2st}$	…	$x_{rs1},\ldots x_{rst}$

数学模型　$x_{ijk} \sim N(\mu_{ij},\sigma^2)$

x_{ijk} 分解：$x_{ijk}=\mu_{ij}+\varepsilon_{ijk}$　$\varepsilon_{ijk}\sim N(0,\sigma^2)$，相互独立

$\mu=\frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s}\mu_{ij}$ ~ 总均值

$\mu_j=\frac{1}{r}\sum_{i=1}^{r}\mu_{ij},\ \beta_j=\mu_j-\mu$

$\beta_j \sim B_j$对指标的效应

$\mu_i=\frac{1}{s}\sum_{j=1}^{s}\mu_{ij},\ \alpha_i=\mu_i-\mu$

$\alpha_i \sim A_i$对指标的效应

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

$\gamma_{ij}\sim A_i, B_j$对指标的交互效应

模型

$$\begin{cases} x_{ijk}=\mu+\alpha_i+\beta_j+\gamma_{ij}+\varepsilon_{ikj} \\ \sum_{i=1}^{r}\alpha_i=0,\ \sum_{j=1}^{s}\beta_j=0,\ \sum_{i=1}^{r}\gamma_{ij}=\sum_{j=1}^{s}\gamma_{ij}=0 \\ \varepsilon_{ijk}\sim N(0,\sigma^2),\quad i=1,\cdots r,\ j=1,\cdots s,\ k=1,\cdots t \end{cases}$$

假设检验

原假设
$H_{01}:\alpha_i=0\ (i=1,\cdots r);$
$H_{02}:\beta_j=0\ (j=1,\cdots s);$
$H_{03}:\gamma_{ij}=0\ (i=1,\cdots r,\ j=1,\cdots s)$

统计分析

$$\bar{x}_{ij}=\frac{1}{t}\sum_{k=1}^{t}x_{ijk}\qquad \bar{x}_{i\cdot}=\frac{1}{s}\sum_{j=1}^{s}\bar{x}_{ij},\ \bar{x}_{\cdot j}=\frac{1}{r}\sum_{i=1}^{r}\bar{x}_{ij}\qquad \bar{x}=\frac{1}{r}\sum_{i=1}^{r}\bar{x}_{i\cdot}=\frac{1}{s}\sum_{j=1}^{s}\bar{x}_{\cdot j}$$

(总偏差) $S=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(x_{ijk}-\bar{x})^2 \Rightarrow S=S_A+S_B+S_{AB}+S_E$

$$\begin{cases} S_A=st\sum_{i=1}^{r}(\bar{x}_{i\cdot}-\bar{x})^2, & S_B=rt\sum_{j=1}^{s}(\bar{x}_{\cdot j}-\bar{x})^2 \\ S_{AB}=t\sum_{i=1}^{r}\sum_{j=1}^{s}(\bar{x}_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x})^2, & S_E=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(x_{ijk}-\bar{x}_{ij})^2 \end{cases}$$

统计分析 $H_{01}:\alpha_i=0,\ H_{02}:\beta_j=0,\ H_{03}:\gamma_{ij}=0$

$$ES_A=(r-1)\sigma^2+st\sum_{i=1}^{r}\alpha_i^2\qquad ES_B=(s-1)\sigma^2+rt\sum_{j=1}^{s}\beta_j^2$$

$$ES_{AB}=(r-1)(s-1)\sigma^2+t\sum_{i=1}^{r}\sum_{j=1}^{s}\gamma_{ij}^2\qquad ES_E=rs(t-1)\sigma^2$$

H_{01}成立 $F_A=\dfrac{S_A/r-1}{S_E/rs(t-1)}\sim F(r-1,rs(t-1))$ $\quad F_A<F_{1-\alpha}(r-1,rs(t-1))$ 接受H_{01}

H_{02}成立 $F_B=\dfrac{S_B/s-1}{S_E/rs(t-1)}\sim F(s-1,rs(t-1))$ $\quad F_B<F_{1-\alpha}(s-1,rs(t-1))$ 接受H_{02}

H_{03}成立 $F_{AB}=\dfrac{S_{AB}/(r-1)(s-1)}{S_E/rs(t-1)}\sim F((r-1)(s-1),rs(t-1))$

$F_{AB}<F_{1-\alpha}((r-1)(s-1),rs(t-1))$ 接受H_{03}

随机变量F_A, F_B和F_{AB}代入样本值后，分别记成 f_A，f_B 和f_{AB}

方差分析表

方差来源	平方和	自由度	平方均值	F 值	概率
因素A	S_A	$r-1$	$\bar{S}_A=S_A/r-1$	$f_A=\bar{S}_A/\bar{S}_E$	$P_A=P(F_A>f_A)$
因素B	S_B	$s-1$	$\bar{S}_B=S_B/s-1$	$f_B=\bar{S}_B/\bar{S}_E$	$P_B=P(F_B>f_B)$
因素A×B	S_{AB}	$(r-1)(s-1)$	$\bar{S}_{AB}=S_{AB}/(r-1)(s-1)$	$f_{AB}=\bar{S}_{AB}/\bar{S}_E$	$P_{AB}=P(F_{AB}>f_{AB})$
误差	S_E	$rs(t-1)$	$\bar{S}_E=S_E/rs(t-1)$		
总和	S	$rst-1$			

显著性水平 α

检验规则

$f_A<F_{1-\alpha}(r-1,rs(t-1))$(即$P_A>\alpha$)时接受 H_{01}，否则拒绝 H_{01}；

$f_B<F_{1-\alpha}(s-1,rs(t-1))$($P_B>\alpha$)时接受 H_{02}，否则拒绝 H_{02}；

$f_{AB}<F_{1-\alpha}((r-1)(s-1),rs(t-1))$($P_{AB}>\alpha$)时接受 H_{03}，否则拒绝 H_{03}。

双因素方差分析-----无交互影响情况

根据经验或某种分析能够事先断定两因素之间没有交互影响，每组试验就不必重复，t=1.

模型

$$\begin{cases} x_{ijk}=\mu+\alpha_i+\beta_j+\gamma_{ij}+\varepsilon_{ikj} \\ \sum_{i=1}^{r}\alpha_i=0,\ \sum_{j=1}^{s}\beta_j=0,\ \sum_{i=1}^{r}\gamma_{ij}=\sum_{j=1}^{s}\gamma_{ij}=0 \\ \varepsilon_{ijk}\sim N(0,\sigma^2),\ i=1,\cdots r,\ j=1,\cdots s,\ k=1,\cdots t \end{cases} \Rightarrow \begin{cases} x_{ij}=\mu+\alpha_i+\beta_j+\varepsilon_{ij} \\ \sum_{i=1}^{r}\alpha_i=0,\ \sum_{j=1}^{s}\beta_j=0 \\ \varepsilon_{ij}\sim N(0,\sigma^2) \end{cases}$$

假设检验: $H_{01}:\alpha_i=0\ (i=1,\cdots,r);\quad H_{02}:\beta_j=0\ \ (j=1,\cdots,s)$

统计分析 $H_{01}:\alpha_i=0\ (i=1,\cdots,r);\quad H_{02}:\beta_j=0\ \ (j=1,\cdots,s)$

$$S=\sum_{i=1}^{r}\sum_{j=1}^{s}(x_{ij}-\bar{x})^2,\qquad S=S_A+S_B+S_E$$

$$S_A=s\sum_{i=1}^{r}(\bar{x}_{i\cdot}-\bar{x})^2,\quad S_B=r\sum_{j=1}^{s}(\bar{x}_{\cdot j}-\bar{x})^2,\quad S_E=\sum_{i=1}^{r}\sum_{j=1}^{s}(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x})^2$$

$$ES_E=(r-1)(s-1)\sigma^2,\ ES_A=(r-1)\sigma^2+s\sum_{i=1}^{r}\alpha_i^2,\ ES_B=(s-1)\sigma^2+r\sum_{j=1}^{s}\beta_j^2$$

H_{01}成立时， $F_A=\dfrac{S_A/r-1}{S_E/(r-1)(s-1)}\sim F(r-1,(r-1)(s-1))$

当H_{02}成立时， $F_B=\dfrac{S_B/s-1}{S_E/(r-1)(s-1)}\sim F(s-1,(r-1)(s-1))$

统计分析

方差分析表

方差来源	平方和	自由度	平方均值	F 值	概率
因素A	S_A	$r-1$	$\bar{S}_A=S_A/r-1$	$f_A=\bar{S}_A/\bar{S}_E$	$P_A=P(F_A>f_A)$
因素B	S_B	$s-1$	$\bar{S}_B=S_B/s-1$	$f_B=\bar{S}_B/\bar{S}_E$	$P_B=P(F_B>f_B)$
误差	S_E	$(r-1)(s-1)$	$\bar{S}_E=S_E/(r-1)(s-1)$		
总和	S	$rs-1$			

检验规则

$f_A<F_{1-\alpha}(r-1,(r-1)(s-1))$(即$P_A>\alpha$)时接受$H_{01}$，否则拒绝$H_{01}$；

$f_B<F_{1-\alpha}(s-1,(r-1)(s-1))$(即$P_B>\alpha$)时接受$H_{02}$，否则拒绝$H_{02}$。

MATLAB实现 双因素方差分析(无交互作用)

命令：**p=anova2(x)**

输入： **x为s行r列的数据矩阵。**

输出: $p = P\{F_A > f_A\}, P\{F_B > f_B\}$

方差来源	平方和	自由度	平方均值	F 值
因素A	S_A	$r-1$	$\bar{S}_A = S_A / r-1$	$f_A = \bar{S}_A / \bar{S}_E$
因素B	S_B	$s-1$	$\bar{S}_B = S_B / s$	$f_B = \bar{S}_B / \bar{S}_E$
误差	S_E	$(r-1)(s-1)$	$\bar{S}_E = S_E / (r-1)(s-1)$	
总和	S	$rs-1$		

MATLAB实现 双因素方差分析(有交互作用)

命令：**p=anova2(x,rep)**

输入： **x为st行r列的数据矩阵, 每一个试验水平(A_i, B_j)的t个数据按列排列, rep=t重复试验的次数** 。

输出： $p = P\{F_A > f_A\},\ P\{F_B > f_B\},\ P\{F_{AB} > f_{AB}\}$

方差来源	平方和	自由度	平方均值	F 值
因素A	S_A	$r-1$	$\bar{S}_A = S_A / r-1$	$f_A = \bar{S}_A / \bar{S}_E$
因素B	S_B	$s-1$	$\bar{S}_B = S_B / s$	$f_B = \bar{S}_B / \bar{S}_E$
因素A×B	S_{AB}	$(r-1)(s-1)$	$\bar{S}_{AB} = S_{AB} / (r-1)(s-1)$	$f_{AB} = \bar{S}_{AB} / \bar{S}_E$
误差	S_E	$rs(t-1)$	$\bar{S}_E = S_E / rs(t-1)$	
总和	S	$rst-1$		

双因素方差分析----MATLAB实现

例：小麦产量

问品种、化肥及二者的交互作用对小麦产量有无显著影响。

化肥 / 品种	A_1	A_2	A_3	A_4
B_1	173, 172	174, 176	177, 179	172, 173
B_2	175, 173	178, 177	174, 175	170, 171
B_3	177, 175	174, 174	174, 173	169, 169

假设检验

$$H_{01}: \alpha_i = 0 \quad (i = 1, \cdots 4);$$
$$H_{02}: \beta_j = 0 \quad (j = 1, \cdots 3);$$
$$H_{03}: \gamma_{ij} = 0 \ (i = 1, \cdots 4,\ j = 1, \cdots 3)$$

双因素方差分析----MATLAB实现

```
%read data from 'wheatdata.m'
x=dlmread('wheatdata.m','\t');
p=anova2(x,2)
```

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

ANOVA Table

Source	SS	df	MS	F
Columns	90.83	3	30.28	33.03
Rows	8.083	2	4.042	4.409
Interaction	51.92	6	8.653	9.439
Error	11	12	0.9167	
Total	161.8	23		

结论:

因素**A(化肥)**和交互作用**AB**影响非常显著, 因素**B(品种)**显著.

演示 wheat.m

目的

1、了解方差分析的基本原理

2、根据问题的要求提出模型

3、对已经确定的模型，确定参数、使用MATLAB

作业 1), 4) , 6) *

1. 目的。 2. 内容（对每一题）：模型（对应用题）；算法设计；计算结果；结果分析；附程序（必要时加说明语句）。 3. 收获和建议。