第五章　方差分析

5.1　单因子试验的方差分析

一、单因素试验

二、平方和的分解

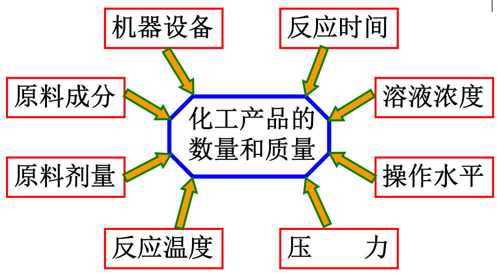
三、S_E, S_A的统计特性

四、假设检验问题的拒绝域

五、未知参数的估计

六、小结

一、单因素（因子）试验



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（因子）——影响试验指标的条件.

因素：可控因素、不可控因素　　通常用A, B, C,…表示

水　平——因素所处的状态或等级.

因子A的r个水平用 A_1, A_2,⋯ A_r表示。

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9.1　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

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

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9.2　电路的响应时间

类型 I	类型 II	类型III	类型IV
19　15	20　40	16　17	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9.3 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: 考察推进器和燃料两因素对射程有无显著的影响.

例1 表9.1 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.

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.

数学模型与分布假设

设因子 A 取 r 个不同的水平 $A_1, A_2, \cdots, A_r$，这相当于有 r 个总体 $X_1, X_2, \cdots, X_n$，又设在水平 A_i 下进行 $n_i(n_i \geqslant 2)$ 次独立试验，相当于从总体 X_i 抽取了容量为 n_i 的样本 $X_{i1}, X_{i2}, \cdots, X_{in_i}(i=1,2,\cdots,r)$，

表 5.2 单因子多水平重复试验数据表

水平 \ 样品号	1	2	…	n_i
A_1	X_{11}	X_{12}	…	X_{1n_1}
A_2	X_{21}	X_{22}	…	X_{2n_2}
⋮	⋮	⋮	…	⋮
A_r	X_{r1}	X_{r2}	…	X_{rn_r}

X_{ij} 就是水平 A_i 下第 j 次重复试验的试验结果数据，

单因素试验方差分析的数学模型

假设 X_{ij} 满足以下数学模型.

$$\begin{cases} X_{ij}=\mu_i+\varepsilon_{ij},\ i=1,2,\cdots,r;\ j=1,2,\cdots,n_i,\ \text{其中} \\ (1)\ E(\varepsilon_{ij})=0,\ D(\varepsilon_{ij})=\sigma_i^2<\infty,\ i=1,2,\cdots,r;\ j=1,2,\cdots,n_i. \\ (2)\ \varepsilon_{ij}\ \text{服从正态分布，且各}\ \varepsilon_{ij}\ \text{相互独立.} \\ (3)\ \sigma_i^2=\sigma^2,\ i=1,2,\cdots,r\ \text{（称为方差齐性或等方差性）.} \end{cases} \quad (5.1.1)$$

易知 $X_{ij} \sim N(\mu_i, \sigma^2)$，$i=1,2,\cdots,r$；$j=1,2,\cdots,n_i$；所有的 X_{ij} 相互独立.

$\varepsilon_{ij}=X_{ij}-\mu_i$ 是水平 A_i 下第 j 次重复试验的试验误差，是不可观测的随机变量，称为随机误差，

μ_i 是总体 X_i 的期望值，其实际意义是水平 A_i 下试验结果数据的理论均值，σ^2 是总体 $X_i(i=1,2,\cdots,r)$的方差.

需要解决的问题

要检验的假设为

H_0：$\mu_1=\mu_2=\cdots=\mu_r$；

H_1：$\mu_1,\mu_2,\cdots,\mu_r$ 中至少有两个不等

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

方差分析就是要决定接受原假设还是拒绝原假设，拒绝 H_0 时，就说明各水平间有显著差异，即因子有显著影响.

数学模型的等价形式

记 $n=\sum_{j=1}^{r}n_j$, $\boxed{\mu=\frac{1}{n}\sum_{j=1}^{r}n_j\mu_j}$. 总平均

$\boxed{\alpha_j=\mu_j-\mu}$, $j=1,2,\cdots,r$.

水平A_j的效应,表示水平A_j下的总体平均值与总平均的差异.

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

$X_{ij}=\mu_i+\varepsilon_{ij}$，$i=1,2,\cdots,r$；$j=1,2,\cdots,n_i$，

$\varepsilon_{ij}\sim N(0,\sigma^2)$，且 ε_{ij}，$i=1,2,\cdots,r,j=1,2,\cdots,n_i$

相互独立

因此，模型（5.1.1）可改写为

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

故假设式

H_0：$\mu_1=\mu_2=\cdots=\mu_r$；

H_1：$\mu_1,\mu_2,\cdots,\mu_r$ 中至少有两个不等

等价于

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

H_1：$\alpha_i,i=1,2,\cdots,r$ 不全为零

二、平方和的分解

$i=1,2,\cdots,r$；第 i 个总体 X_i 的样本均值和样本方差

$$\bar{X}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}X_{ij}\qquad S_i^2=\frac{1}{n_i-1}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2$$

$$n=\sum_{i=1}^{r}n_i$$

样本的总均值 $\bar{X}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij}=\frac{1}{n}\sum_{i=1}^{r}n_i\bar{X}_i$

总的离差平方和 $S_T=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2$

它反映了全体样本 X_{ij} 的波动程度的大小.

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

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

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

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

$=0$

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

$$= S_E + S_A$$

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

—误差平方和，组内平方和

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

—因子平方和，组间平方和

三、S_E，S_A 的统计特性

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

$$\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2\Big/\sigma^2\sim\chi^2(n_i-1).$$

又由于各 X_{ij} 独立，所以由 χ^2 分布的可加性知

$$S_E/\sigma^2\sim\chi^2\left(\sum_{i=1}^{r}(n_i-1)\right),$$

即 $$S_E/\sigma^2\sim\chi^2(n-r),\text{其中}n=\sum_{i=1}^{r}n_i.$$

根据 χ^2 分布的性质可以得到

S_E 的自由度是 $n-r$；

$$E(S_E)=(n-r)\sigma^2.$$

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

因 $$\sum_{i=1}^{r}\sqrt{n_i}[\sqrt{n_i}(\bar{X}_i-\bar{X})]=\sum_{i=1}^{r}n_i(\bar{X}_i-\bar{X})$$

$$=\sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij}-n\bar{X}=0$$

所以 S_A 的自由度为 $r-1$.

又因 $\mu=\frac{1}{n}\sum_{i=1}^{r}n_i\mu_i$，$\bar{X}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij}$，$X_{ij}$ 相互独立

所以 $\bar{X}\sim N(\mu,\sigma^2/n)$.

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

$$E(S_A)=E[\sum_{i=1}^{r}n_i\bar{X}_i^2-n\bar{X}^2]$$

$$=\sum_{i=1}^{r}n_iE(\bar{X}_i^2)-nE(\bar{X}^2)$$

$$=\sum_{i=1}^{r}n_i[\frac{\sigma^2}{n_i}+(\mu+\alpha_i)^2]-n[\frac{\sigma^2}{n}+\mu^2]$$

$$=(r-1)\sigma^2+2\mu\underbrace{\sum_{i=1}^{r}n_i\alpha_i}_{=0}+n\mu^2+\sum_{i=1}^{r}n_i\alpha_i^2-n\mu^2$$

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

S_A与S_E独立，H_0为真时， $S_A/\sigma^2\sim\chi^2(r-1)$.

四、假设检验问题的拒绝域

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

H_1：$\alpha_i, i=1,2,\cdots,r$ 不全为零

H_0为真时， $S_A/\sigma^2\sim\chi^2(r-1)$. $E(\frac{S_A}{r-1})=\sigma^2$

H_1是真的

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

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

即不管 H_0 是否是真，$S_E/(n-r)$ 都是 σ^2 的无偏估计

$$F=\frac{S_A/(r-1)}{S_E/(n-r)}.$$

1. 分子和分母相互独立;

2. 分母 S_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.

拒绝域形如 $F=\frac{S_A/(r-1)}{S_E/(n-r)}\geq k.$

所以 H_0 为真时，

$$S_A/\sigma^2 \sim \chi^2(r-1),\ S_E/\sigma^2 \sim \chi^2(n-r),$$

$$\frac{S_A/(r-1)}{S_E/(n-r)}=\frac{S_A/\sigma^2}{(r-1)}\Big/\frac{S_E/\sigma^2}{(n-r)}\sim F(r-1,n-r).$$

给定检验水平 α 后

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

H_1：$\alpha_i, i=1,2,\cdots,r$ 不全为零

拒绝域为: $F=\frac{S_A/(r-1)}{S_E/(n-r)}\geq F_{1-\alpha}(r-1,n-r).$

表 5.3 方差分析表

方差来源	平方和	自由度	均方	F值	显著性
因子	$S_A=\sum_{i=1}^{r} n_i(\overline{X}_i-\overline{X})^2$	$r-1$	$S_A/(r-1)$	$F=\frac{S_A/(r-1)}{S_E/(n-r)}$	
误差	$S_E=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)^2$	$n-r$	$S_E/(n-r)$		
总和	$S_T=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2$	$n-1$			

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

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

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij}=\frac{1}{n}\sum_{i=1}^{r}n_i\overline{X}_i$$

例4 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9.1 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

取$\alpha=0.05$,检验假设

$$H_0:\mu_1=\mu_2=\mu_3,\quad H_1:\mu_1,\mu_2,\mu_3\text{不全相等}.$$

解 $r=3, n_1=n_2=n_3=5, n=15,$

$S_T=0.00124533,\ S_A=0.00105333,\ S_E=0.000192.$

方差分析表

方差来源	平方和	自由度	均　方	F 比
因　素A	0.00105333	2	0.00052667	32.92
误　　差	0.000192	12	0.000016	
总　　和	0.00124533	14		

$F=32.92>F_{1-0.05}(2,12)=3.89.$ 在水平 **0.05** 下拒绝 H_0.

各机器生产的薄板厚度有显著差异.

五、未知参数的估计

$$\begin{cases}X_{ij}=\mu_i+\varepsilon_{ij},\ i=1,2,\cdots,r;\ j=1,2,\cdots,n_i,\ \text{其中}\\ (1)\ E(\varepsilon_{ij})=0,\ D(\varepsilon_{ij})=\sigma_i^2<\infty,\ i=1,2,\cdots,r;\ j=1,2,\cdots,n_i.\\ (2)\ \varepsilon_{ij}\text{服从正态分布，且各}\ \varepsilon_{ij}\ \text{相互独立.}\\ (3)\ \sigma_i^2=\sigma^2,\ i=1,2,\cdots,r\ \text{（称为方差齐性或等方差性）.}\end{cases}$$

（1）点估计. $\mu=\frac{1}{n}\sum_{i=1}^{r}n_i\mu_i\quad \alpha_i=\mu_i-\mu, i=1,2,\cdots,r.$

$\hat{\mu}=\overline{X}$ 是 μ 的无偏估计，$\hat{\mu}_j=\overline{X}_j$ 是 μ_j 的无偏估计，

$\hat{\alpha}_i=\overline{X}_i-\overline{X}$ 是 α_i 的无偏估计，

$\hat{\sigma}^2=S_E/(n-r)$ 是 σ^2 的无偏估计

(2) 区间估计.

为挑选效应最大的水平（称之为优水平），有必要对均值差 $\mu_i-\mu_j$ 做区间估计.

记 $\theta=\sum_{i=1}^{r}a_i\mu_i$，其中，$a_i$ 为常数，且 $\sum_{i=1}^{r}a_r=0$.

$\hat{\theta}=\sum_{i=1}^{r}a_i\bar{X}_i$ 是 θ 的无偏估计.

$$X_{ij}\sim N(\mu_i,\sigma^2), i=1,2,\cdots,r, j=1,2,\cdots,n_r$$

$$\bar{X}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}X_{ij}\sim N(\mu_i,\sigma^2/n_i), i=1,2,\cdots,r.$$

$$E\Big(\sum_{i=1}^{r}a_i\bar{X}_i\Big)=\sum_{i=1}^{r}a_i\mu_i,\ D\Big(\sum_{i=1}^{r}a_i\bar{X}_i\Big)=\sigma^2\sum_{i=1}^{r}\frac{a_i^2}{n_i},$$

$$U=\frac{\sum_{i=1}^{r}a_i\bar{X}_i-\sum_{i=1}^{r}a_i\mu_i}{\sqrt{\sigma^2\sum_{i=1}^{r}a_i^2/n_i}}\sim N(0,1)$$

$S_E/\sigma^2\sim\chi^2(n-r)$，且 S_E 与各 $\bar{X}_i$ 相互独立，

S_E 与 $\sum_{i=1}^{r}a_i\bar{X}_i$ 独立.

$$\frac{\sum_{i=1}^{r}a_i\bar{X}_i-\sum_{i=1}^{r}a_i\mu_i}{\sqrt{\Big(\sum_{i=1}^{r}a_i^2/n_i\Big)S_E/(n-r)}}\sim t(n-r)$$

给定置信水平 $1-\alpha$，得

$$\sum_{i=1}^{r}a_i\bar{X}_i-t_{1-\frac{\alpha}{2}}(n-r)\sqrt{\Big(\sum_{i=1}^{r}a_i^2/n_i\Big)S_E/(n-r)}\leqslant\sum_{i=1}^{r}a_i\mu_i$$

$$\leqslant\sum_{i=1}^{r}a_i\bar{X}_i+t_{1-\frac{\alpha}{2}}(n-r)\sqrt{\Big(\sum_{i=1}^{r}a_i^2/n_i\Big)S_E/(n-r)}$$

取 $a_i=1$，其余的 $a_j=0$，得 μ_i 的区间估计为

$$\bar{X}_i-t_{1-\frac{\alpha}{2}}(n-r)\sqrt{\frac{S_E}{n-r}/n_i},\quad \bar{X}_i+t_{1-\frac{\alpha}{2}}(n-r)\sqrt{\frac{S_E}{n-r}/n_i}$$

取 $a_i=1$，$a_j=-1$，其余的 $a_k=0$，得均值差 $\mu_i-\mu_j$ 的区间估计

$$\Big[\bar{X}_i-\bar{X}_j-t_{1-\frac{\alpha}{2}}(n-r)\sqrt{\frac{S_E}{n-r}\Big(\frac{1}{n_i}+\frac{1}{n_j}\Big)},$$

$$\bar{X}_i-\bar{X}_j+t_{1-\frac{\alpha}{2}}(n-r)\sqrt{\frac{S_E}{n-r}\Big(\frac{1}{n_i}+\frac{1}{n_j}\Big)}\Big]$$

例 5.1.4 某化工厂用钡泥制取硝酸钡试验中，考虑到溶钡的溶出率随酸度的增大而提高，今将酸度从 pH=4 降至 pH=1，每次作 4 次试验，测得废水中硝酸钡的含量（%）见表 5.4.

表 5.4 废水中硝酸钡含量/%

pH值 \ 样品	1	2	3	4
4	6.17	6.73	6.45	6.53
3	5.89	5.73	5.50	5.61
2	5.01	5.19	5.37	5.26
1	4.28	4.75	4.79	4.50

试问溶钡酸度对废水中硝酸钡含量是否有显著影响?

解 假设在一定的溶钡酸度下硝酸钡的含量服从正态分布，各总体的均值分别为 μ_1,μ_2,μ_3,μ_4，需检验

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

由表 5.4 的数据计算所得结果列于下面的方差分析表中，

方差来源	平方和	自由度	均方	F值	显著性
因子	$S_A=7.6211$	3	2.5404	63.07	* *
误差	$S_E=0.4831$	12	0.0403		
总和	$S_T=8.1042$	15			

取检验水平 $\alpha=0.01$，$F_{0.99}(3,12)=5.95$ 小于 F 值=63.07，故拒绝 H_0，即认为溶钡酸度对废水中硝酸钡含量的影响是高度显著的.

各总体均值的估计值，经计算得 $\hat{\mu}_1 = \overline{X}_1 = \frac{1}{4}\sum_{j=1}^{4} X_{ij} = \frac{1}{4} \times 25.88 = 6.47$，$\hat{\mu}_2 = \overline{X}_2 = 5.68$，$\hat{\mu}_3 = 5.21$，$\hat{\mu}_4 = 4.58$.

均值差 $\mu_1 - \mu_2$ 的 95% 置信区间为

$$\left[\overline{X}_1 - \overline{X}_2 \pm t_{1-\frac{\alpha}{2}}(n-r)\sqrt{\frac{S_E}{n-r}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right]$$

$$=\left[6.47-5.68\pm 2.179\sqrt{\frac{2}{4}\times 0.0403}\right]$$

$$=[0.48, 1.10]$$

由于置信区间不含零，故可知在水平 $\alpha=0.05$ 上，μ_1 显著地大于 μ_2. 若所得均值差 $\mu_i - \mu_j$ 的置信区间含零，则认为 μ_i 与 μ_j 无显著差异.

六、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.

作业

P144： 2、3、7