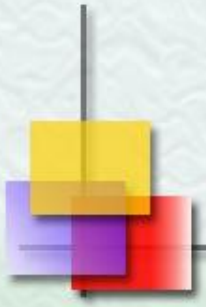
Ch8 方差分析和回归分析

§8.1 方差分析

Analysis of Variance, ANOVA

又称“变异数分析”，是R.A.Fisher发明的，用于检验同方差的若干正态总体均值是否相等的一种统计方法。

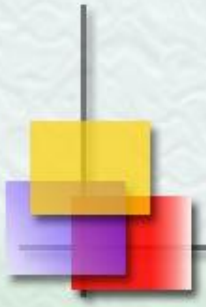


例8.1.1 在饲料养鸡增肥的研究中，某研究所提出三种饲料配方：A_1是以鱼粉为主的饲料，A_2是以槐树粉为主的饲料，A_3是以苜蓿粉为主的饲料。为比较三种饲料的效果，特选 24 只相似的雏鸡随机均分为三组，每组各喂一种饲料，60天后观察它们的重量。试验结果如下表所示：

表8.1.1 鸡饲料试验数据

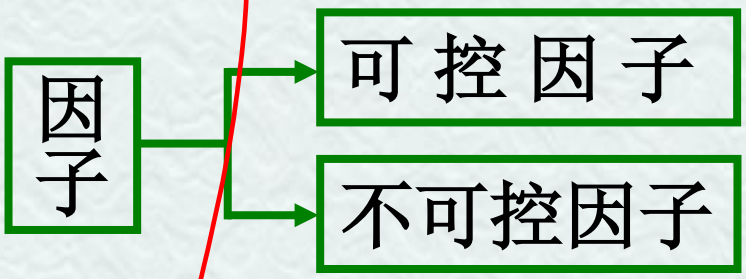
饲料A	鸡 重（克）							
A_1	1073	1009	1060	1001	1002	1012	1009	1028
A_2	1107	1092	990	1109	1090	1074	1122	1001
A_3	1093	1029	1080	1021	1022	1032	1029	1048

本例中，我们要比较的是三种饲料对鸡的增肥作用是否相同。为此，把饲料称为因子，记为A，三种不同的配方称为因子A的三个水平，记为A_1，A_2，A_3，使用配方A_i下第 j只鸡60天后的重量用y_{ij}表示，$i=1, 2, 3$， $j=1, 2,\ldots,10$. 我们的目的是比较三种饲料配方下鸡的平均重量是否相等，为此，需要做一些基本假定，把所研究的问题归结为一个统计问题，然后用方差分析的方法进行解决。



试验指标——试验中要考察的指标.

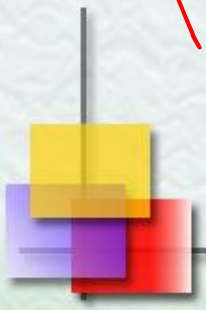
因　　子——影响试验指标的因素.



水　　平——因子(素)所处的状态.

单因子试验——在一项试验中只有一个因子改变.

多因子试验——在一项试验中有多个因子在改变.

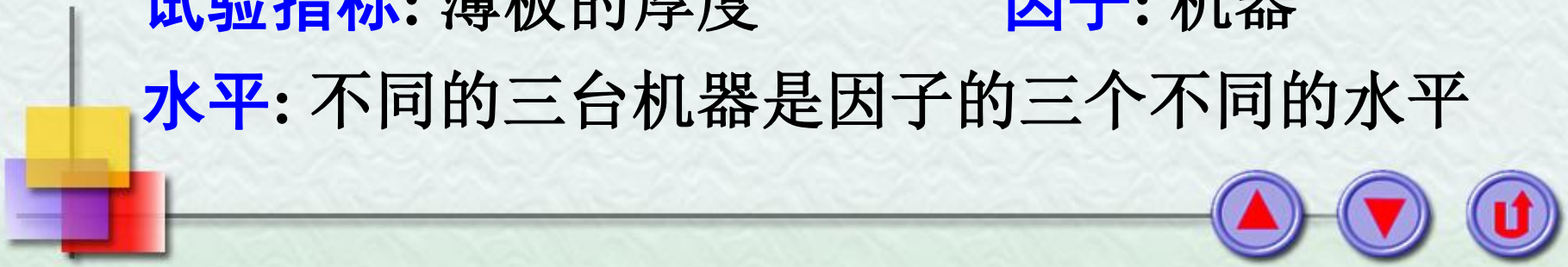


例1：设有三台机器，用来生产规格相同的铝合金薄板.取样,测量薄板的厚度精确至千分之一厘米.得结果如下表所示.

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　　**因子**: 机器

水平: 不同的三台机器是因子的三个不同的水平



假定除机器这一因素外, 其他条件相同, 属于**单因子试验**.

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. 即考察机器这一因子对厚度有无显著的影响.

又如书**P382例8.1**

书例8.2（P398）：

为了考察蒸馏水的pH和硫酸铜溶液浓度对化验血清中白蛋白与球蛋白的影响，对蒸馏水的pH(A)取了4个不同的水平，对硫酸铜浓度(B)取了3个不同的水平，在不同水平组合下各测一次白蛋白与球蛋白之比，列表如下：

浓度(B)		B_1	B_2	B_3
pH(A)	A_1	3.5	2.3	2.0
	A_2	2.6	2.0	1.9
	A_3	2.0	1.5	1.2
	A_4	1.4	0.8	0.3

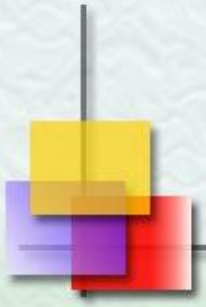
试验指标: 白蛋白与球蛋白之比

因子：蒸馏水的pH和硫酸铜溶液浓度

水平：蒸馏水的pH有4个,硫酸铜浓度有3个

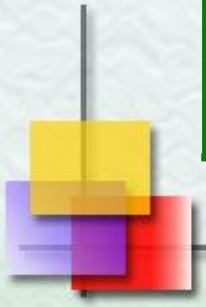
试验目的:考察蒸馏水的pH和硫酸铜溶液浓度对化验血清中白蛋白与球蛋白的影响

双因子无重复试验



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,得火箭射程如下（以海里计）.

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



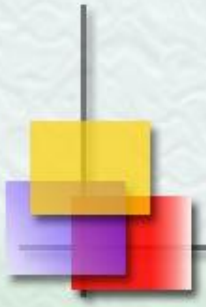
试验指标: 火箭射程

因子: 推进器和燃料

水平: 推进器有3个,燃料有4个

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.

双因子等重复试验　　又如书P402例8.3



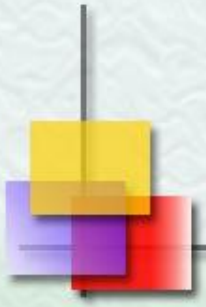
单因子方差分析

一、单因子试验

二、平方和分解

三、S_E, S_A的统计特性

四、假设检验问题



单因子方差分析的统计模型

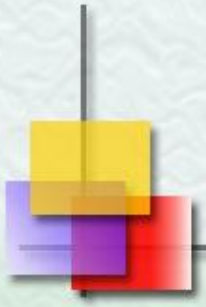
在例8.1.1中我们只考察了一个因子，称其为单因子试验。

通常，在单因子试验中，记因子为 A，设其有r个水平，记为A_1，A_2，…，A_r，在每一水平下考察的指标可以看成一个总体 ，现有r个水平，故有r个总体，假定：

1) 每一总体均为正态总体，记为 $N(\mu_i, \sigma_i^2)$，$i = 1, 2, \ldots, r$；

2) 各总体的方差相同：$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_r^2 = \sigma^2$；

3) 从每一总体中抽取的样本是相互独立的，即所有的试验结果 Y_{ij} 都相互独立。



我们要比较各水平下的均值是否相同，即要对如下的一个假设进行检验:

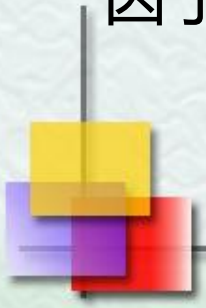
$$H_0: \mu_1 = \mu_2 = \ldots = \mu_r \qquad (8.1.1)$$

备择假设为

H_1: $\mu_1, \mu_2, \ldots, \mu_r$ 不全相等

在不会引起误解的情况下，H_1通常可省略不写。

如果H_0成立，因子A的r个水平均值相同，称因子A的r个水平间没有显著差异，简称因子A不显著；反之，当H_0不成立时，因子A的r个水平均值不全相同，这时称因子A的不同水平间有显著差异，简称因子A显著。



为对假设（8.1.1）进行检验，需要从每一水平下的总体抽取样本，设从第i个水平下的总体获得t个试验结果，令 y_{ij} 表示第i个总体的第j次重复试验结果。共得如下$n=r\times t$个试验结果：

$$y_{ij},\quad i=1,2,\ldots,r,\quad j=1,2,\ldots,t,$$

其中r为水平数，t为重复数，i为水平编号。

在水平A_i下的试验结果Y_{ij}与该水平下的指标均值μ_i一般总是有差距的，

记 $\varepsilon_{ij} = Y_{ij} - \mu_i$，

ε_{ij} 称为随机误差。于是有

$$Y_{ij} = \mu_i + \varepsilon_{ij} \qquad (8.1.2)$$

(8.1.2) 式称为试验结果 Y_{ij} 的数据结构式。

$$
\begin{cases}
Y_{ij} = \mu_i + \varepsilon_{ij}, \\
\varepsilon_{ij} \sim N(0, \sigma^2), \text{各} \varepsilon_{ij} \text{独立}, \\
i = 1, 2, \cdots, r,\ j = 1, 2, \cdots, t, \\
\mu_i \text{与} \sigma^2 \text{均未知}.
\end{cases}
\tag{8.1.3}
$$

单因子方差分析的数学模型

需要解决的问题

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.

记 $\mu=\frac{1}{r}\sum_{i=1}^{r}\mu_i$，表示所有总体均值的平均(一般平均).

$\alpha_i=\mu_i-\mu, i=1,\cdots,r$ 为因子 A 的第 i 个水平的主效应，简称A_i的水平效应.

易见 $\sum_{i=1}^{r}\alpha_i=0.$

则单因子方差分析模型中数据结构式 Y_{ij} 可以表示为

$$Y_{ij}=\mu_i+\varepsilon_{ij}$$
$$=\mu+\alpha_i+\varepsilon_{ij},\ j=1,\cdots,t;\ i=1,\cdots,r.$$

$$\sum_{i=1}^{r}\alpha_i=0.$$

原数学模型 $Y_{ij}=\mu_i+\varepsilon_{ij}$,

$\varepsilon_{ij}\sim N(0,\sigma^2)$, 各 ε_{ij} 独立,

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

μ_i 与 σ^2 均未知.

改写为

$Y_{ij}=\mu+\alpha_i+\varepsilon_{ij}$,

$\varepsilon_{ij}\sim N(0,\sigma^2)$, 各 ε_{ij} 独立,

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

$\sum_{i=1}^{r}\alpha_i=0.$

检验假设 $H_0:\mu_1=\mu_2=\cdots=\mu_r,$

改为检验假设

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

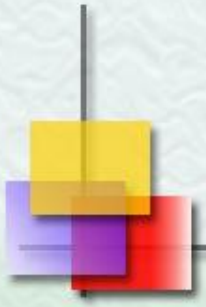
二、平方和分解

注意到引起每个Y_{ij}波动的原因可能有两种:

一是H_0成立时的随机波动;

二是H_0不成立时各水平间的差异引起的.

用平方和分解法将上述两种波动原因表示出来.



$n = rt$ — 总样本量

$\bar{Y} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{t} Y_{ij}$ —样本的总平均

$S_T = \sum_{i=1}^{r}\sum_{j=1}^{t}\left(Y_{ij} - \bar{Y}\right)^2$ —总偏差平方和（总变差）

$\bar{Y}_{i\bullet} = \frac{1}{t}Y_{i\bullet} = \frac{1}{t}\sum_{j=1}^{t} Y_{ij}$ —水平A_i下的样本平均值

$Y_{i\cdot} = \sum_{j=1}^{t} Y_{ij}$

$$S_T = \sum_{i=1}^{r}\sum_{j=1}^{t}\left(Y_{ij} - \overline{Y}\right)^2$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{t}\left[\left(Y_{ij} - \overline{Y}_{i\bullet}\right) + \left(\overline{Y}_{i\bullet} - \overline{Y}\right)\right]^2$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{t}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 + \sum_{i=1}^{r}\sum_{j=1}^{t}\left(\overline{Y}_{i\bullet} - \overline{Y}\right)^2$$

$$+2\sum_{i=1}^{r}\sum_{j=1}^{t}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)\left(\overline{Y}_{i\bullet} - \overline{Y}\right)$$

$$= 0$$

总的偏差平方和可以分解为

$$S_T=\sum_{i=1}^{r}\sum_{j=1}^{t}\left(Y_{ij}-\overline{Y}\right)^2=\sum_{i=1}^{r}\sum_{j=1}^{t}\left(Y_{ij}-\overline{Y}_{i\bullet}\right)^2+\sum_{i=1}^{r}\sum_{j=1}^{t}\left(\overline{Y}_{i\bullet}-\overline{Y}\right)^2$$

$$=\sum_{i=1}^{r}\sum_{j=1}^{t}\left(Y_{ij}-\overline{Y}_{i\bullet}\right)^2+\sum_{i=1}^{r}t\left(\overline{Y}_{i\bullet}-\overline{Y}\right)^2$$

$$=S_e+S_A$$

S_e 表示组内的偏差平方和, S_A 表示组间的偏差平方和 (因子的不同水平引起的偏差平方和).

为了进一步看清 S_e 和 S_A 的意义,注意到

$$Y_{ij}=\mu+\alpha_i+\varepsilon_{ij},$$

$$\bar{Y}_{i\bullet}=\frac{1}{t}\sum_{j=1}^{t}(\mu+\alpha_i+\varepsilon_{ij})=\mu+\alpha_i+\bar{\varepsilon}_{i\bullet},$$

$$\bar{Y}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{t}(\mu+\alpha_i+\varepsilon_{ij})=\mu+\bar{\varepsilon}.$$

所以, $S_e=\sum_{i=1}^{r}\sum_{j=1}^{t}(Y_{ij}-\bar{Y}_{i\bullet})^2=\sum_{i=1}^{r}\sum_{j=1}^{t}(\varepsilon_{ij}-\bar{\varepsilon}_{i\bullet})^2$

称为(随机) 误差的偏差平方和.

$$S_A=\sum_{i=1}^{r}t(\bar{Y}_{i\bullet}-\bar{Y})^2=\sum_{i=1}^{r}t(\alpha_i+\bar{\varepsilon}_{i\bullet}-\bar{\varepsilon})^2$$

称为因子 A 的偏差平方和.

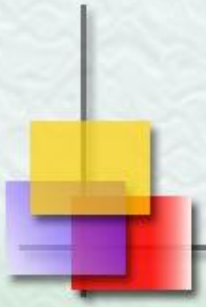
三、S_e，S_A的统计特性

$$S_e=\sum_{i=1}^{r}\sum_{j=1}^{t}\left(Y_{ij}-\bar{Y}_{i\bullet}\right)^2$$

$$=\sum_{j=1}^{t}\left(Y_{1j}-\bar{Y}_{1\bullet}\right)^2+\cdots+\sum_{j=1}^{t}\left(Y_{rj}-\bar{Y}_{r\bullet}\right)^2,$$

$\frac{1}{t-1}\sum_{j=1}^{t}\left(Y_{ij}-\bar{Y}_{i\bullet}\right)^2$是$N(\mu_i,\sigma^2)$的样本方差

$$\frac{1}{\sigma^2}\sum_{j=1}^{t}\left(Y_{ij}-\bar{Y}_{i\bullet}\right)^2\sim\chi^2(t-1).$$



又由于各 Y_{ij} 独立，所以由 χ^2 分布的可加性知

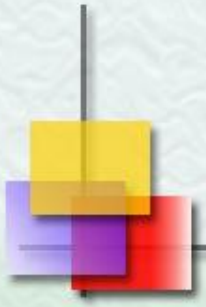
$$S_e/\sigma^2 \sim \chi^2(r(t-1)),$$

即 $$S_e/\sigma^2 \sim \chi^2(n-r), \quad \text{其中} n=rt.$$

根据 χ^2 分布的性质可以得到：

S_e 的自由度为 $n-r$；$$E\left(\frac{S_e}{n-r}\right)=\sigma^2.$$

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



$$S_A=\sum_{i=1}^{r}\sum_{j=1}^{t}\left(\bar{Y}_{i\bullet}-\bar{Y}\right)^2=t\sum_{i=1}^{r}\left(\bar{Y}_{i\bullet}-\bar{Y}\right)^2$$

因为$\sum_{i=1}^{r}\left(\bar{Y}_{i\bullet}-\bar{Y}\right)=\sum_{i=1}^{r}\bar{Y}_{i\bullet}-r\bar{Y}=\sum_{i=1}^{r}\left(\frac{1}{t}\sum_{j=1}^{t}Y_{ij}\right)-r\left(\frac{1}{rt}\sum_{i=1}^{r}\sum_{j=1}^{t}Y_{ij}\right)=0$

所以 S_A 的自由度为 $r-1$.

(*) $Y_{ij}\sim N(\mu_i,\sigma^2), i=1,2,\cdots,r$, 且相互独立,

所以 $\bar{Y}_{i\bullet}=\frac{1}{t}\sum_{j=1}^{t}Y_{ij}\sim N(\mu_i,\frac{\sigma^2}{t})$.

$\bar{Y}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{t}Y_{ij}\sim N(\mu,\frac{\sigma^2}{n})$, 其中$\mu=\frac{1}{r}\sum_{i=1}^{r}\mu_i, n=rt$

$$E(S_A)=E\left[t\sum_{i=1}^{r}\left(\overline{Y}_{i\bullet}-\overline{Y}\right)^2\right]=E\left[t\sum_{i=1}^{r}\left(\overline{Y}_{i\bullet}^2-2\overline{Y}_{i\bullet}\overline{Y}+\overline{Y}^2\right)\right]$$

$$=E\left[t\sum_{i=1}^{r}\overline{Y}_{i\bullet}^2-n\overline{Y}^2\right]=t\sum_{i=1}^{r}E(\overline{Y}_{i\bullet}^2)-nE(\overline{Y}^2)$$

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

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

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

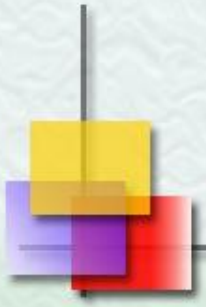
$=0$

四、假设检验问题

检验假设

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

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



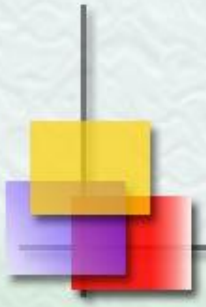
定理8.1.2 在单因子方差分析模型 (8.1.8) 及前述符号下，有

(1) $S_e/\sigma^2 \sim \chi^2(n-r)$，从而

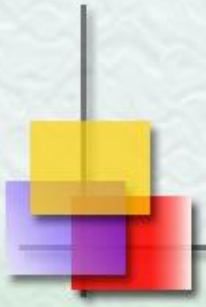
$$E(S_e) = (n-r)\,\sigma^2$$

(2) $ES_A = (r-1)\sigma^2 + t\sum_{i=1}^{r}\alpha_i^2$，进一步，若 H_0 成立，则有 $S_A/\sigma^2 \sim \chi^2(r-1)$

(3) S_A 与 S_e 独立。



定理 8.1 (柯赫伦定理) 设 $X_1, \cdots, X_n$ 为 n 个相互独立的 $N(0,1)$ $r.v.s, Q = \sum_{i=1}^{n} X_i^2$ 为 $\chi^2(n)$ $r.v.$, 若 $Q = Q_1 + \cdots + Q_k$, 其中 Q_i 为某些正态 $r.v.s$ 的平方和, 这些正态 $r.v.s$ 分别是 $X_1, \cdots, X_n$ 的线性组合, 其自由度为 f_i. 则诸 Q_i 相互独立, 且诸 $Q_i \sim \chi^2(f_i)$ 的充要条件是 $\sum_{i=1}^{k} f_i = n.$

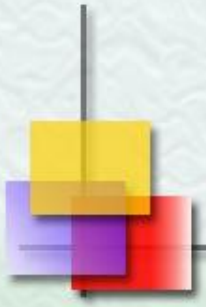


当 H_0 成立时，$\dfrac{1}{\sigma^2}S_e \sim \chi^2(n-r)$，

$$\frac{1}{\sigma^2}S_T \sim \chi^2(n-1)(\text{？})$$

由柯赫伦定理知，$\dfrac{1}{\sigma^2}S_A \sim \chi^2(r-1)$，且与 $\dfrac{1}{\sigma^2}S_e$ 独立.

所以，当 H_0 成立时，$F = \dfrac{S_A/(r-1)}{S_e/(n-r)} \sim F(r-1, n-r)$.

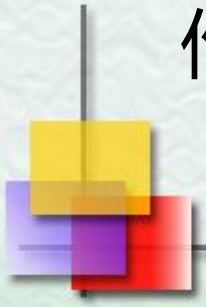


偏差平方和Q的大小与自由度有关，为了便于在偏差平方和间进行比较，统计上引入了均方（Mean Square）的概念，它定义为$MS=Q/f_Q$，其意为平均每个自由度上有多少平方和。

如今要对因子平方和 S_A 与误差平方和 S_e 之间进行比较，用其均方 $MS_A= S_A /f_A$ ， $MS_e= S_e /f_e$ 进行比较更为合理，故可用

$$F=\frac{MS_A}{MS_e}=\frac{S_A/f_A}{S_e/f_e}$$

作为检验H_0的统计量。



因子A 的偏差平方和

当H_0 成立时, $F=\dfrac{S_A/(r-1)}{S_e/(n-r)}\sim F(r-1,n-r).$

随机误差的偏差平方和

拒绝域的类型?

给定显著性水平α, 拒绝域$C=\{F>\lambda\}.$

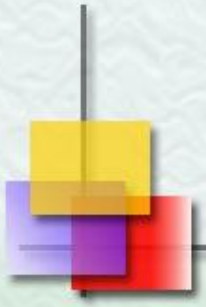
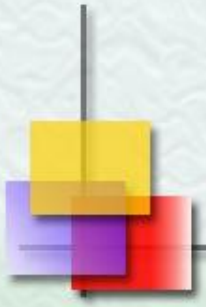


表8.1.3　单因子方差分析表

来源	平方和	自由度	均方和	F比
因子	S_A	$f_A=r-1$	$MS_A=S_A/f_A$	$F=MS_A/MS_e$
误差	S_e	$f_e=n-r$	$MS_e=S_e/f_e$	
总和	S_T	$f_T=n-1$		

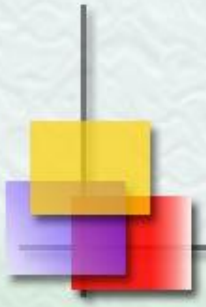


常用的各偏差平方和的计算公式如下

$$S_T=\sum_{i=1}^{r}\sum_{j=1}^{t}\left(Y_{ij}-\overline{Y}\right)^2=\sum_{i=1}^{r}\sum_{j=1}^{t}Y_{ij}^2-n\overline{Y}^2,$$

$$S_A=\sum_{i=1}^{r}\sum_{j=1}^{t}\left(\overline{Y}_{i\bullet}-\overline{Y}\right)^2=t\sum_{i=1}^{r}\overline{Y}_{i\bullet}^2-n\overline{Y}^2=\sum_{i=1}^{r}\frac{Y_{i\bullet}^2}{t}-n\overline{Y}^2,$$

$$\overline{Y}_{i\bullet}=\frac{Y_{i\bullet}}{t}=\frac{\sum_{j=1}^{t}Y_{ij}}{t},\qquad S_e=S_T-S_A.$$



步骤：

1. 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.

2. H_0成立时, 选取检验统计量

$$F=\frac{S_A/(r-1)}{S_e/(n-r)}\sim F(r-1,n-r)$$

3. 给定显著性水平 α, 拒绝域 $C=\{F\geq F_{1-\alpha}(r-1,n-r)\}.$

4. 根据样本数据计算F的值，并判断其是否落入拒绝域中.

例2：

某灯泡厂用四种不同的配料方案制成的灯丝生产了四批灯泡，在每批灯泡中随机抽取若干个测得其使用寿命如下：(小时)

	A_1	A_2	A_3	A_4
1	1600	1580	1460	1510
2	1610	1640	1550	1520
3	1650	1640	1600	1530
4	1680	1700	1620	1570
5	1700	1750	1640	1680
6	1720	1800	1740	1600

$$Y_i \sim N(\mu_i, \sigma^2), i = 1,2,3,4$$

检验：$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

问由这四种灯丝生产的灯泡平均寿命有无显著差异？$\alpha = 0.05$

假设各灯泡的寿命都独立地服从正态分布，且方差相等.

假设 $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

	A_1	A_2	A_3	A_4
1	1600	1580	1460	1510
2	1610	1640	1550	1520
3	1650	1640	1600	1530
4	1680	1700	1620	1570
5	1700	1750	1640	1680
6	1720	1800	1740	1600
$y_{i\bullet}$	9960	10110	9610	9410
$y_{i\bullet}^2$	99201600	102212100	92352100	88548100

$$S_A = \sum_{i=1}^{r} \frac{y_{i\bullet}^2}{t} - n\bar{y}^2$$

$$= 51145.8$$

$$S_e = S_T - S_A$$

$$= 108916.7$$

$$\sum_{i,j} y_{ij} = 39090$$

$$\sum_{i} y_{i\bullet}^2 = 382313900$$

$r = 4, t = 6, n = 24.$

$$S_T = \sum_{i=1}^{r}\sum_{j=1}^{t} y_{ij}^2 - n\bar{y}^2 = 63827900 - 24 \times 1628.75^2 = 160062.5$$

方差分析表

来源	平方和	自由度	均方和	F 比
A	51145.8	$r-1=3$	17048.6	3.13
e	108916.7	$n-r=20$	5445.835	
总和	160062.5			$F_{0.95}(3,20)=3.10$

由于 $F=\dfrac{S_A/(r-1)}{S_e/(n-r)}=3.13>F_{0.95}(3,20)=3.10$

所以拒绝 H_0, 可以认为在 $\alpha=0.05$ 的显著性水平下, 不同配方所生产的灯泡平均寿命有显著的差异.

即可以认为配方是影响灯泡寿命的显著因素.

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.

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

假设各薄板的厚度都独立地服从正态分布,且方差相等.

取$\alpha = 0.05$, 问各机器生产的薄板厚度有显著差异?

解 检验假设$H_0:\mu_1=\mu_2=\mu_3$，　$H_1:\mu_1,\mu_2,\mu_3$不全相等.

$r=3,t=5,n=15,$

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

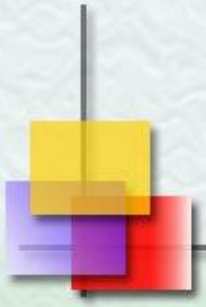
方差分析表

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

$$F = 32.92 > F_{0.95}(2,12) = 3.89.$$

在水平0.05下拒绝 H_0.

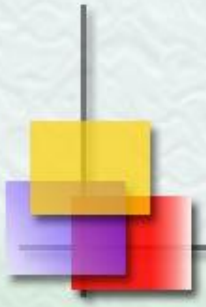
各机器生产的薄板厚度有显著差异.



当因子在第i个水平下做了t_i次试验时，$i=1,\cdots,r$，修改公式如下 $S_T=\sum_{i=1}^{r}\sum_{j=1}^{t_i}Y_{ij}^2-n\bar{Y}^2$，

$$S_A=\sum_{i=1}^{r}\frac{Y_{i\bullet}^2}{t_i}-n\bar{Y}^2,\ S_e=S_T-S_A,$$

其中$n=\sum_{i=1}^{r}t_i$，$Y_{i\bullet}=\sum_{j=1}^{t_i}Y_{ij}$，其余部分均不变.



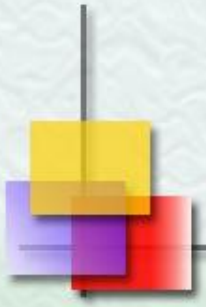
练习: 设用3台机器 A, B 和 C 制造同一产品,
对每台机器观察 5 天的日产量如下:

A : 41 48 41 57 49 51

B : 65 57 54 72 64

C : 45 51 56 48

试问: 在日产量上, 各台机器是否有差别? $\alpha = 0.05$



方差分析：单因素方差分析

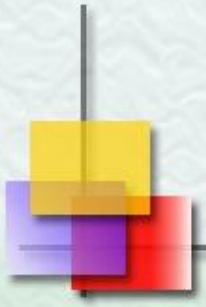
SUMMARY

组	观测数	求和	平均	方差
列 1	6	9960	1660	2360
列 2	6	10110	1685	6550
列 3	6	9610	1601.667	8736.667
列 4	6	9410	1568.333	4136.667

方差分析

差异源	SS	df	MS	F	P-value	F crit
组间	51145.83	3	17048.61	3.130579	0.048513	3.098391
组内	108916.7	20	5445.833			
总计	160062.5	23				

SS： Stdev Square 偏差平方和
df： degree of freedom 自由度
MS: Mean Square 均方
F crit: Critical Values of F F的临界值



（补充） 方差齐性检验

在进行方差分析时要求r个方差相等，这称为方差齐性。理论研究表明，当正态性假定不满足时对F检验影响较小,即F检验对正态性的偏离具有一定的稳健性，而F检验对方差齐性的偏离较为敏感。所以r个方差的齐性检验就显得十分必要。

所谓方差齐性检验是对如下一对假设作出检验：

$$H_0:\ \sigma_1^2=\sigma_2^2=\cdots\sigma_r^2 \quad vs \quad H_1:\ 诸\sigma_i^2不全相等 \tag{8.3.1}$$

很多统计学家提出了一些很好的检验方法，这里介绍几个最常用的检验，它们是：

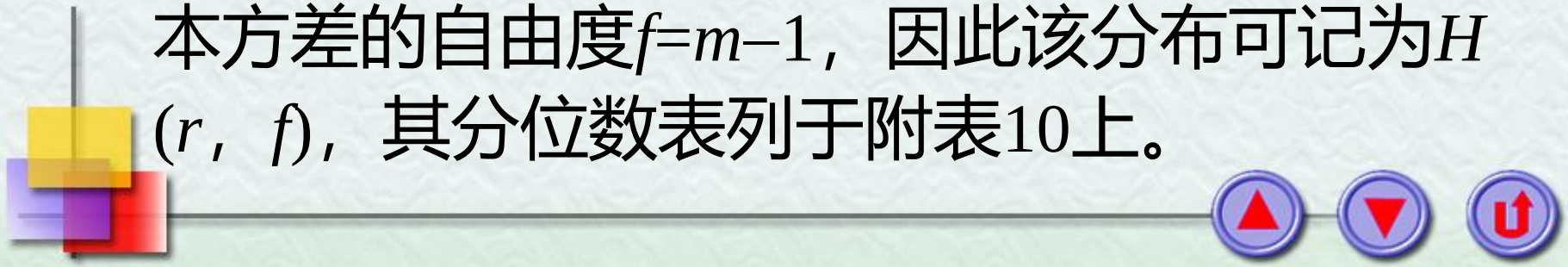
- **Hartley检验，仅适用于样本量相等的场合；**
- **Bartlett检验，可用于样本量相等或不等的场合，但是每个样本量不得低于5；**
- **修正的Bartlett检验，在样本量较小或较大、相等或不等场合均可使用。**

8.3.1 Hartley（哈特利）检验

当各水平下试验重复次数相等时，即$m_1=m_2=\ldots=m_r=m$,Hartley提出检验方差相等的检验统计量：

$$H=\frac{\max\left\{s_1^2,s_2^2,\cdots,s_r^2\right\}}{\min\left\{s_1^2,s_2^2,\cdots,s_r^2\right\}} \tag{8.3.2}$$

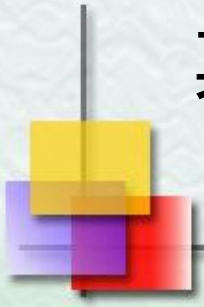
这个统计量的分布无明显的表达式，但在诸方差相等条件下，可通过随机模拟方法获得H分布的分位数，该分布依赖于水平数r和样本方差的自由度$f=m-1$，因此该分布可记为$H(r，f)$，其分位数表列于附表10上。



直观上看，当H_0成立，即诸方差相等（$\sigma_1^2=\sigma_2^2=\ldots=\sigma_r^2$）时，$H$的值应接近于1，当$H$的值较大时，诸方差间的差异就大，$H$愈大，诸方差间的差异就愈大，这时应拒绝 (8.3.1) 中的H_0。由此可知，对给定的显著性水平α，检验H_0的拒绝域为

$$W=\{H > H_{1-\alpha}(r, f)\} \tag{8.3.3}$$

其中$H_{1-\alpha}(r, f)$为H分布的$1-\alpha$分位数。



例8.3.1 有四种不同牌号的铁锈防护剂（简称防锈剂），现要比较其防锈能力。数据见表8.3.1。

这是一个重复次数相等的单因子试验。我们考虑用方差分析方法对之进行比较分析，为此，首先要进行方差齐性检验。

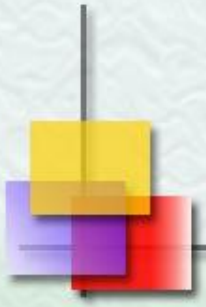


表 8.3.1　防锈能力数据及计算表

因子 A(防锈剂)		A_1	A_2	A_3	A_4
数据 y_{ij}	1	43.9	89.8	68.4	36.2
	2	39.0	87.1	69.3	45.2
	3	46.7	92.7	68.5	40.7
	4	43.8	90.6	66.4	40.5
	5	44.2	87.7	70.0	39.3
	6	47.7	92.4	68.1	40.3
	7	43.6	86.1	70.6	43.2
	8	38.9	88.1	65.2	38.7
	9	43.6	90.8	63.8	40.9
	10	40.0	89.1	69.2	39.7
和 T_i		431.4	894.4	679.5	404.7
均值 $\bar{y}_{i\cdot}$		43.14	89.44	67.95	40.47
组内平方和 Q_i		81.00	44.28	42.33	53.42

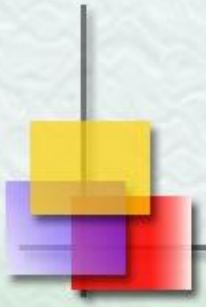
本例中，四个样本方差可由表8.3.1中诸Q_i求出，即

$$s_1^2=\frac{81.00}{9}=9.00,\quad s_2^2=\frac{44.28}{9}=4.92,$$

$$s_3^2=\frac{42.33}{9}=4.70,\quad s_4^2=\frac{53.42}{9}=5.94$$

由此可得统计量H的值 $H=\dfrac{9.00}{4.70}=1.9149$

在α=0.05时，由附表10查得$H_{0.95}(4,9)$=6.31，由于H<6.31，所以应该保留原假设H_0，即认为四个总体方差间无显著差异。



ANOVA

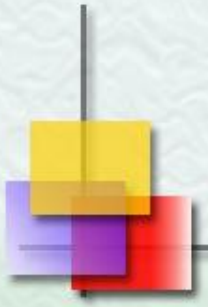
数据

			平方和	自由度	均方	F	显著性
组间	（组合）		15953.466	3	5317.822	866.118	.000
	线性项	对比	435.125	1	435.125	70.869	.000
		偏差	15518.341	2	7759.171	1263.743	.000
组内			221.034	36	6.140		
总计			16174.500	39			

描述

数据

	个案数	平均值	标准 偏差	标准 错误	平均值的 95% 置信区间		最小值	最大值
					下限	上限		
A1	10	43.1400	3.00007	.94871	40.9939	45.2861	38.90	47.70
A2	10	89.4400	2.21821	.70146	87.8532	91.0268	86.10	92.70
A3	10	67.9500	2.16859	.68577	66.3987	69.5013	63.80	70.60
A4	10	40.4700	2.43632	.77043	38.7272	42.2128	36.20	45.20
总计	40	60.2500	20.36494	3.21998	53.7370	66.7630	36.20	92.70



8.3.2 Bartlett检验

在单因子方差分析中有r个样本，设第i个样本方差为：

$$s_i^2=\frac{1}{m_i-1}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_i)^2=\frac{Q_i}{f_i},\quad i=1,2,\cdots,r$$

由于几何平均数总不会超过算术平均数，故有$GMS_e\le MS_e$，其中

$$MS_e=\frac{1}{f_e}\sum_{i=1}^{r}Q_i=\sum_{i=1}^{r}\frac{f_i}{f_e}s_i^2,\quad GMS_e=\left[\left(s_1^2\right)^{f_1}\left(s_2^2\right)^{f_2}\cdots\left(s_r^2\right)^{f_r}\right]^{1/f_e}$$

等号成立当且仅当诸s_i^2彼此相等，若诸s_i^2间的差异愈大，则此两个平均值相差也愈大。

由此可见，在比值GMS_e/MS_e较大时，就意味着诸样本方差差异较大，从而检验（8.3.1）表示的一对假设的拒绝域应是

$$W=\{\ln GMS_e/MS_{e>}>d\}$$

(8.3.4)

Bartlett证明了，在大样本场合

$$B=\frac{f_e}{C}(\ln MS_e-\ln GMS_e)\sim\chi^2(r-1)$$

其中

$$C=1+\frac{1}{3(r-1)}\left[\sum_{i=1}^{r}\frac{1}{f_i}-\frac{1}{f_e}\right]$$

检验统计量

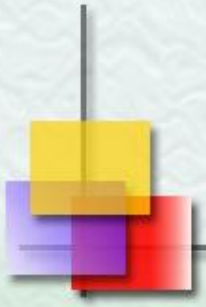
$$B=\frac{1}{C}\left[f_e \ln MS_e-\sum_{i=1}^{r} f_i \ln s_i^2\right]$$

检验的拒绝域为

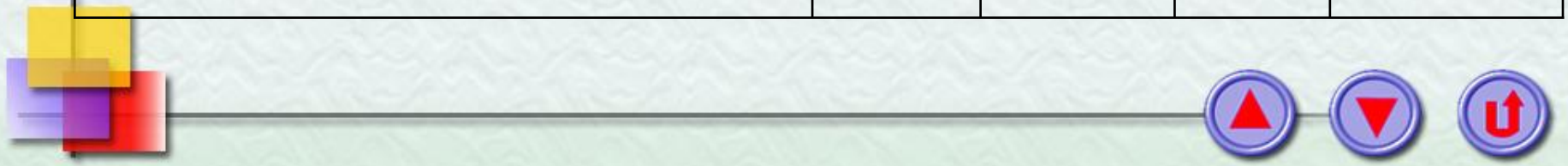
$$W=\{B>\chi_{1-\alpha}^2(r-1)\} \qquad (8.3.8)$$

考虑到这里χ^2分布是近似分布，在诸样本量m_i均不小于5时使用上述检验是适当的。

例8.3.2 为研究各产地的绿茶的叶酸含量是否有显著差异，特选四个产地绿茶，其中A_1制作了7个样品，A_2制作了5个样品，A_3与A_4各制作了6个样品，共有24个样品，按随机次序测试其叶酸含量，测试结果如表8.3.3所示。



水平	数据	重复数	和	均值	组内平方和
A_1	7.9 6.2 6.6 8.6 8.9 10.1 9.6	$m_1=7$	$T_1=57.9$	8.27	$Q_1=12.83$
A_2	5.7 7.5 9.8 6.1 8.4	$m_2=5$	$T_2=37.5$	7.50	$Q_2=11.30$
A_3	6.4 7.1 7.9 4.5 5.0 4.0	$m_3=6$	$T_3=34.9$	5.82	$Q_3=12.03$
A_4	6.8 7.5 5.0 5.3 6.1 7.4	$m_4=6$	$T_4=38.1$	6.35	$Q_4=5.61$
		$n=24$	$T=168.4$		$S_e=41.77$



来源	平方和	自由度	均方	F比	p值
因子A	23.50	3	7.83	3.75	0.0435
误差e	41.77	20	2.09		
和T	65.27	23			

显著性水平α =0.05，查表知$F_{0.95}(3,20)$ =3.10。由于F>3.10，故应拒绝原假设H_0，即认为四种绿茶的叶酸平均含量有显著差异。

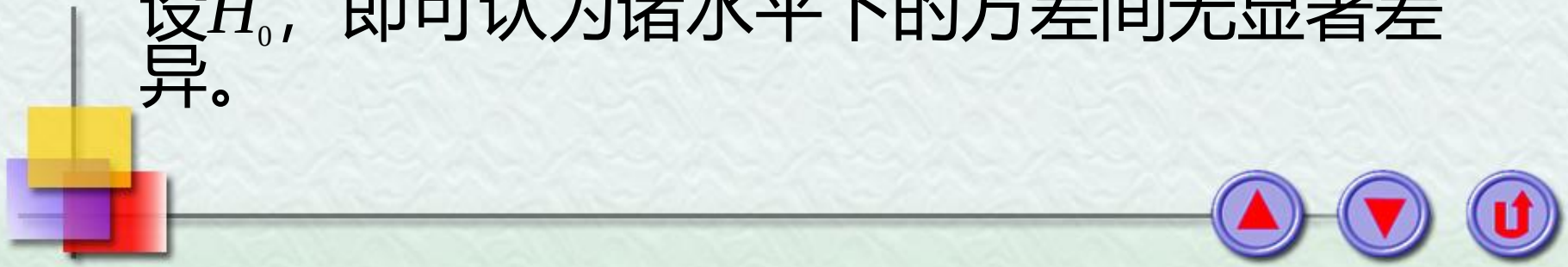
为能进行方差分析，首先要进行方差齐性检验，从表8.3.3中数据可求得$s_1^2=2.14$, $s_2^2=2.83$, $s_3^2=2.41$, $s_4^2=1.12$，再从表8.3.4上查得$MS_e=2.09$，由(8.3.6)，可求得

$$C=1+\frac{1}{3(4-1)}\left[\left(\frac{1}{6}+\frac{1}{4}+\frac{1}{5}+\frac{1}{5}\right)-\frac{1}{20}\right]=1.0852$$

再由(8.3.7)，还可求得Bartlett检验统计量的值

$$B=\frac{1}{1.0852}\left[20\times\ln 2.09-(6\times\ln 2.14+4\times\ln 2.83+5\times\ln 2.41+5\times\ln 1.12)\right]=0.970$$

对给定的显著性水平$\alpha=0.05$，查表知$\chi^2_{0.95}(4-1)=7.815$。由于$B<7.815$，故应保留原假设H_0，即可认为诸水平下的方差间无显著差异。



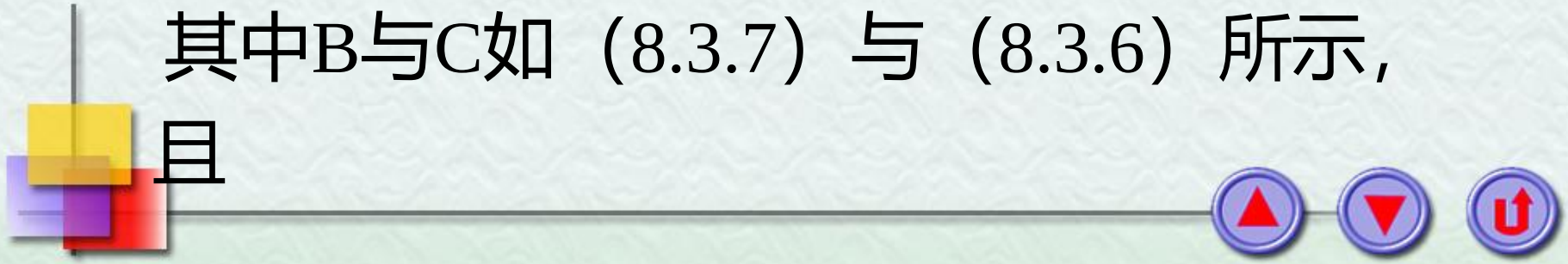
8.3.3 修正的Bartlett检验

针对样本量低于5时不能使用Bartlett检验的缺点，Box提出修正的Bartlett检验统计量

$$B' = \frac{f_2 BC}{f_1 (A - BC)}$$

$$f_1 = r - 1, \quad f_2 = \frac{r+1}{(C-1)^2}, \quad A = \frac{f_2}{2 - C + 2/f_2} \tag{8.3.9}$$

其中B与C如（8.3.7）与（8.3.6）所示，且



在原假设H_0：$\sigma_1^2=\sigma_2^2=\ldots=\sigma_r^2$成立下，Box还证明了统计量 B 的近似分布是F分布$F(f_1, f_2)$，对给定的显著性水平α，该检验的拒绝域为

$$W=\{B'>F_{1-\alpha}(f_1, f_2)\} \tag{8.3.10}$$

其中f_2的值可能不是整数，这时可通过对F分布的分位数表施行内插法得到分位数。

例8.3.3 对例8.3.2中的绿茶叶酸含量的数据，我们用修正的Bartlett检验再一次对等方差性作出检验。

在例8.3.2中已求得：C=1.0856，B=0.970，还可求得：

$$f_1 = 4-1 = 3$$

$$f_2 = \frac{4+1}{(1.0852-1)^2} = 688.8$$

$$A = \frac{688.8}{2-1.0852+2/688.8} = 750.6$$

$$B' = \frac{688.8\times 0.970\times 1.0852}{3(750.6-0.970\times 1.0852)} = 0.322$$

对给定的显著性水平α =0.05，在F分布的分位数表上可查得 $F_{0.95}(3,682.4)= F_{0.95}(3,\infty)=2.60$

由于$B' < 2.60$，故接受原假设H_0，即认为四个水平下的方差间无显著差异。

$\sigma_1^2=\sigma_2^2=\sigma^2$未知,关于$\mu_1,\mu_2$的假设检验

$$H_0:\mu_1-\mu_2=0 \quad \leftrightarrow \quad H_1:\mu_1-\mu_2\neq 0$$

检验统计量:

$$T=\frac{\bar{X}-\bar{Y}}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}},\text{其中 } S_w=\sqrt{\frac{(n_1-1)S_1^{\,2}+(n_2-1)S_2^{\,2}}{n_1+n_2-2}}.$$

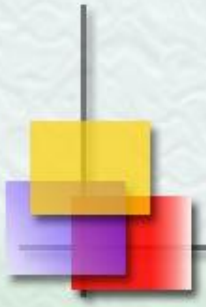
当H_0为真时, $T\sim t(n_1+n_2-2)$.

方差分析

检验假设 $H_0: \mu_1 = \mu_2 = \cdots = \mu_r = 0,$

$H_1: \mu_1, \mu_2, \cdots, \mu_r$不全为零.

当 H_0 成立时, $F = \dfrac{S_A/(r-1)}{S_e/(n-r)} \sim F(r-1, n-r).$



思考：

1. 等方差的两个正态总体的均值差检验和两个水平的单因子方差分析有什么关系？

作业： p.428 8.1，8.4，8.6