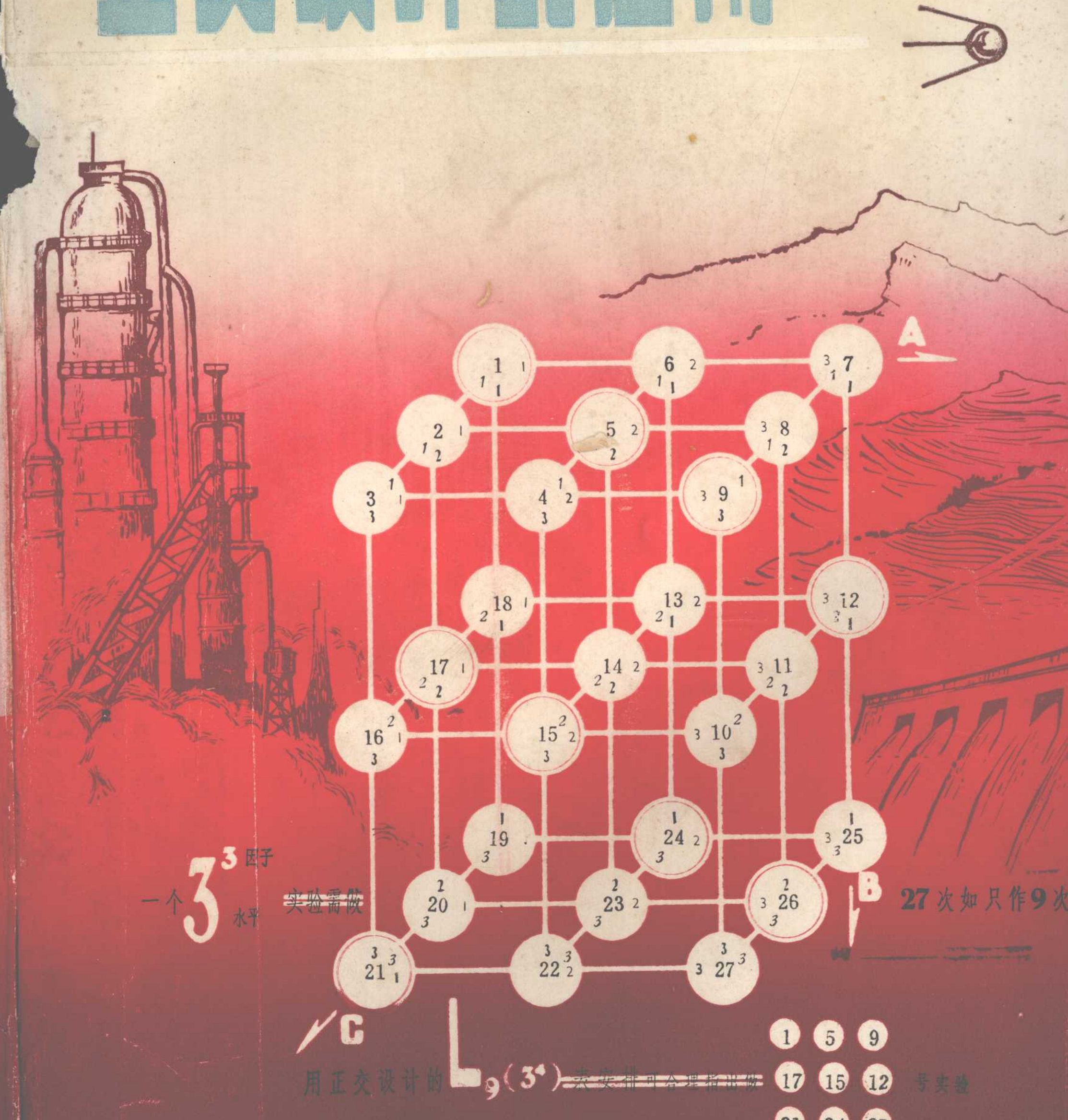
正交设计的应用



江西共大总校七三〇药厂
江西省科学技术情报研究所

毛主席语录

千万不要忘记阶级斗争。

列宁为什么说对资产阶级专政，这个问题要搞清楚。这个问题不搞清楚，就会变修正主义。要使全国知道。

人类总得不断地总结经验，有所发现，有所发明，有所创造，有所前进。

不能把过程中所有的矛盾平均看待，必须把它们区别为主要的和次要的两类，着重于捉住主要矛盾，……。

胸中有“数”。这是说，对情况和问题一定要注意到它的数量方面，要有基本的数量的分析。任何质量都表现为一定的数量，没有数量也就没有质量。

说　明

在毛主席革命路线指引下，在无产阶级文化大革命和批林批孔运动的推动下，广大工农兵群众和科技人员认真学习无产阶级专政理论，以阶级斗争为纲，坚持党的基本路线，积极投入科学实验运动，并涌现出大量科技成果，有力地推动了我国社会主义革命和社会主义建设的发展。

近年来，各条战线在科学实验运动中，都广泛推广应用“优选法”，对多快好省地建设社会主义起了很大的作用。属于“优选法”的另一种正交设计，现在也在全国各地陆续推广和应用。因为它是一种多因素实验设计和实验结果分析相结合的科学实验方法，把设计、实验、分析三者有机地联系起来，对搞好科学实验很有作用。为了使正交设计在我省也能得到广泛的推广和应用，我们特选辑了徐吉民同志在学习和应用正交设计过程中编写的这本小册子，以供同志们参考。

由于编辑水平有限，错误之处，请予批评指正。

一九七五年十二月

内　容　简　介

在生产斗争和科学实验中，经常要做各种实验，如果要使这些实验起到事半功倍的作用，除按有关专业技术严格地进行实验外，实验前的设计和实验后的分析，也是一个重要环节。所以设计、实验和分析三者是缺一不可的。

正交设计，就是用已印好的表格——正交表科学地安排多因素实验，使得实验尽量减少，然后又能通过一些表格化的计算，可以由表及里，去伪存真的分析实验结果。如果把它用到生产实验中，可以找出好的生产条件。因此，正交设计既可用到各个学科进行实验设计和实验结果分析，又是容易掌握的一种符合多快好省的科学实验方法。

本书通过农业、化工、制药、药理、临床、纺织、机械、冶炼等41个例子，较全面地介绍了正交设计方法的应用。可供工农兵、知识青年、中小学教师和一般科技人员参考。

正交设计的应用

第一章 方差分析

方差分析，就是利用各种数据的平方差，将研究对象的变化和其它偶然因素造成的试验误差分开，从而得出正确的结论。

科学试验的目的，是为了通过试验来了解各种因素对产品的性能（或被研究对象的规律性）、成本、产量等的影响。今后，把性能、成本、产量等统称为考察的指标；响影指标的因素统称为因子。因子对指标的影响表现为：因子所处的状态若发生变化就会引起指标数据的变化。今后把因子所处的状态称为水平。

正交设计的计算部分是以方差分析为基础，故在本章着重介绍单因素方差分析，双因素方差分析和重复双因素方差分析。有了这些计算基础，对正交设计中的方差分析的计算就容易理解。

§1 单因素方差分析

分析一批只受一种因素响影的实验数据的方法，叫单因素方差分析（或叫一元方差分析）。这里以笔者研究生长在同一条件下的江西五种薄荷品质优劣为例，这是一个单因素多水平的方差分析。薄荷品种可以看作是一种因素，五种薄荷相当于薄荷品种的五个不同水平。6、8、10、12、14、16、18点钟的七次取样测定薄荷油含量，只看作重复实验，它的每个数据只受薄荷品种的某一水平影响。其实验数据请见表1.1粗线框内的数据。

表 1·1　　　　**五种薄荷含薄荷油量**

X 薄荷油含量		A 薄荷品种					ΣX	$(\Sigma X)^2$	
		一号	二号	三号	四号	五号			
B 采样时间	6点钟	-50	31	-1	-65	-95	-180	32400	
	8点钟	30	99	59	-21	-70	97	9409	
	10点钟	-80	15	97	-49	-72	-89	7921	
	12点钟	98	75	18	-41	-21	129	16641	$=82713$ (ΣT_j^2)
	14点钟	28	146	43	-40	-51	126	15876	
	16点钟	-62	153	-14	-70	-12	-5	25	
	18点钟	-24	93	36	-88	-38	-21	441	
ΣX(各数相加)		-60	612	238	-374	-359	57(T)		
$(\Sigma X)^2$(各数相加后平方)		3600	374544	56644	139876	128881	703545(ΣT_j^2)		
ΣX^2(各数平方后相加)		24608	69691	16556	22988	23739	157582(R)		

表 1.1 粗线框内的数据，是五个薄荷品种的七次取样测定的数值，粗线框下端的各种数据，是五种薄荷测定数据的对应计算数值，ΣX 是对应的各数相加合计，$(\Sigma X)^2$ 是 ΣX 对应合计的平方，ΣX^2 是框内对应数据平方后相加的合计，五个薄荷品种所对应的 ΣX^2 相加，就是全部试验数据平方后相加的合计，用 R 表示，它是计算总的离均

各种符号说明

粗线框内各数据均×100－200

$T=\Sigma X$(全部测定数据之和)

$CT=P=\dfrac{T^2}{N}$(全部测定数据和的平方后的均数，称不变项)

$R=\Sigma X^2$(全部测定数据平方后的合计)

$N=$全部数据个数

$Q_A=\Sigma T_i^2\times\dfrac{1}{nb}$(薄荷品种平方和的均数)　na为薄荷水平数

$Q_B=\Sigma T'^2_j\times\dfrac{1}{nb}$(取样时间平方和的均数)　nb为取样水平数

平方差和的基本数（$S_T=R-CT$）。五个薄荷品种对应的 ΣX 相加的合计，用 T 表示，它是计算总均数$\left(\hat{\mu}=\frac{T}{N}\right)$和不变项$\left(CT=\frac{T^2}{N}\right)$的基本数。有了这些数据，单因素方差分析就有了依据。如果把不同时间取样不看作只是一个试验的重复，而是作为一个因素，七次取样时间是这个因子的七个水平，以了解不同时间测定的含量是否有差异，于是可在表的右侧照表下方那样增加 ΣX 和 $(\Sigma X)^2$ 的计算，就能对其进行双因素方差分析。

方差分析中的总离均差平方和（用 S_T 表示），它是由各个试验数据平方后的合计（ΣX^2 在表 1.1 用 R 表示），再减去各个试验数据合计（T）的平方的均数$\left(\frac{T^2}{N}\right)$，$\left(\text{用}CT\text{表示，即}CT=\frac{T^2}{N}\right)$。以薄荷为例，$S_T$ 中包含有不同薄荷品种含量的差异（S_A），也包含有不同取样测定的差异（S_B）和试验中各种原因造成的试验误差（S_e）。

薄荷品种间的含油量离均差平方和（S_A），它是由每种薄荷的七次试验的合计平方 $(\Sigma X)^2$ 之后相加（ΣT_i^2）除以取样次数，再减去不变项（CT）算得。S_A 刻划了不同薄荷品种含量的差异，也有试验中种种原因造成的试验误差（S_e）。

不同取样测定的离均差平方和（S_B），它是由每次取样的五个薄荷品种的合计平方 $(\Sigma X)^2$ 之后合计（ΣT_j^2），除以薄荷品种个数，再减去不变项（CT）算得。S_B 刻划了不同时间取样差异，也包含有试验中种种原因造成的试验误差（S_e）。在本节单因素方差分析中，只把 S_B 当作重复试验，不了解其差异，把由不同取样引起的差异，放在试验误差中去了。

由于原因不明引起的试验误差的离均差平方和（S_e），它是由总的离均差平方和，减去各种因素的离均差平方和之后剩余的数值。在单因素方差分析中只减去 S_A，在双因素分析中则减去 S_A 和 S_B，在三因素分析中则再减去一个 S_c，四因素、五因素以此类推。S_e 刻划了试验误差的大小。其数愈小，说明试验条件控制较好，试验精密度较高。

根据以上各原则，现就表 1.1 各数据进行单因素方差分析的各项计算。

1.计算各种离均差平方和——S

$$S_T(\text{总离均差平方和})=\Sigma X^2-\frac{T^2}{N}=R-CT$$

$$=157582-\frac{(57)^2}{5\times7}=157582-99.46=157482.54$$

$$S_A(\text{薄荷 品种 离均差平方和})=\Sigma T_i^2\times\frac{1}{nb}-\frac{T^2}{N}=Q_A-CT$$

$$=703545\times\frac{1}{7}-99.46=100406.9$$

$$S_e(\text{试验误差的离均差平方和})=S_T-S_A=R-Q_A$$

$$=157482.54-100406.9=57075.64$$

要比较各类数据差异大小，不能直接用S，因为S的大小与参与计算的数目的个数有关。只有用离均差均方($\overline{S}$)才有意义，所以第二步就要计算其自由度(f)和均方($\overline{S}$)。

2.计算各类自由度——f

$$f_T(\text{总自由度})=n_a\times n_b-1=N-1=7\times5-1=34$$

$$f_a(\text{薄荷品种自由度})=n_a-1=5-1=4$$

$$f_e(\text{误差自由度})=f_T-f_a=34-4=30$$

3.计算各种均方——$\overline{S}$

$$\overline{S}_A(\text{薄荷品种离均差均方})=S_a\times\frac{1}{f_a}$$

$$=100406.9\times\frac{1}{4}=25117.3$$

$$\overline{S}_e(\text{误差离均差均方})=S_e\times\frac{1}{f_e}=5710756.4\times\frac{1}{30}=1902.52$$

前面已经谈到，在S_A中既包含有不同薄荷品种含薄荷油量的差异，也包含有试验引起的误差，所以，如果$\overline{S}_A$与$\overline{S}_e$相近，则说明$\overline{S}_A$不是不同薄荷品种含薄荷油不同引起的，而是试验误差造成的。如果$\overline{S}_A$大于$\overline{S}_e$几倍，即F值越大，则说明$\overline{S}_A$不单是实验引起的误差，而且有不同薄荷品种所含薄荷油不同而引起的。究竟F值多大，才说明

薄荷品种间所含薄荷油量达到显著(0.05水准)，或者高度显著（0.01水准）的差异。统计学证明，它与参与实验样本数大小有关，此值可从F值表查得。

4.计算F值：

$$F_A=\overline{S}_A/\overline{S}_e=\frac{25117.3}{1902.52}=13.20$$（此值说明$\overline{S}_A$比$\overline{S}_e$大13.20倍）

5.查F表：

因$\overline{S}_A$的$f_a=4$，相当于F表的f因，$\overline{S}_e$的$f_e=30$相当于F表的f_e，这时如以$\alpha=0.05$水准（显著水平）衡量，可根据这两者自由度查$\alpha=0.05$的F表，表上所示$F_{0.05}(4,30)=2.69$；如查$\alpha=0.01F$表，表上所示$F_{0.01}(4,30)=4.02$，也就是说如计算的F值等于或大于2.69时，或F值等于或大于4.02时，则说明此因素的水平间有显著或高度显著差异。现在计算的$F_A=13.20>F_{0.01}(4,30)=4.02$，所以说明薄荷品种间含油量有高度显著差异。也就是说薄荷品种间含油量的差异不是试验误差造成的，而是薄荷品种的不同造成的。于是方差分析到此可告结束。并可将这些数据列成方差分析表公布。

薄荷品种方差分析表

方差来源	离均差平方和	自由度	均　　方	F	显著性
薄荷品种	100406.9	4	25117.3	13.20	**
误　　差	57075.64	30	1902.52		
总　　计	157482.54	34			

$F_{0.05}(4,30)=2.69$　　$F_{0.01}(4,30)=4.02$

§2　双因素方差分析

分析一批受两种因素响影的实验数据的方法，叫双因素方差分析（或叫二元方差分析）。在前一节中介绍的那个例子是单因素方差分析，因为例一中只有薄荷品种作为一个因素，而不同取样时间只看作薄荷品种的重复实验，而没有把它看作另一因素的不同水平，所以在

计算时未算它的离均差平方和，而是把它当作误差放在误差中一道计算了。如果现在不但要了解这五个薄荷品种的含薄荷油量是否有显著差异？而且要知道不同取样时间其薄荷油量是否有显著差异？那在计算时不但要算出薄荷品种A因素 离均差平方和（S_A），而且要算出取样时间B因素离均差平方和（S_B）及实验误差离均差平方和（S_e）。这种方差分析就叫双因素方差分析。

在上述单因素计算基础上，像计算S_A一样增加计算一个S_B，再在计算S_e时，从S_T中多减去一个S_B，即成下述结果。

1.各种S：

$S_T=157482.54$(同例一)

$S_A=100406.9$(同例一)

根据表 1.1 右侧各数据可算出S_B。

$$S_B=\frac{\sum T_j^2}{na}-\frac{(\sum X)^2}{N}=Q_B-CT=82713\times\frac{1}{5}-\frac{(57)^2}{35}$$

$$=16443.7$$

$$S_e=S_T-S_A-S_B=157482.54-100406.9-16443.7$$

$$=40631.94$$

2.各种f：

$f_T=34$(同例一)

$f_a=4$(同例一)

$f_b=nb-1=7-1=6$

$f_e=f_T-f_a-f_b=24$

3.各种$\overline{S}$：

$\overline{S}_A=25117.3$(同例一)

$$\overline{S}_B=S_B\times\frac{1}{f_b}=16443.74\times\frac{1}{6}=2740.62$$

$$\overline{S}_e=S_e\times\frac{1}{f_e}=40631.94\times\frac{1}{24}=1692.99$$

4.各种F：

$$F_A=\frac{\overline{S}_A}{\overline{S}_e}=\frac{25117.3}{1692.99}=14.84$$

$$F_B=\frac{\overline{S}_B}{\overline{S}_e}=\frac{2740.62}{1692.99}=1.62$$

5.查F表：

当$f_a=4$，$f_e=24$ $\alpha=0.01$时，其$F=4.22$，而$F_A=14.84$，所以A因素$P<0.01$，即薄荷品种间的含薄荷油量有高度显著差异（以＊＊表示）。

当$f_b=6$，$f_e=24$，$\alpha=0.05$时，其$F=2.51$，而$F_B=1.62$，所以B因素$P>0.05$，即薄荷不同取样时间其含薄荷油量无显著差异。于是可以列出方差分析表。

方差来源	离均差平方和	自由度	均方	F	显著性
A薄荷品种	100406.9	4	25117.3	14.84	**
B采样时间	16443.7	6	2740.62	1.62	
误差	40631.94	24	1692.99		
总计	157482.54	34			

$F_{0.05}(6,24)=2.51$ $F_{0.01}(4,24)=4.22$

§3 重复双因素方差分析

为了提高试验的精密度，减少试验误差的干扰，经常对同一个试验重复几次。如果在双因素试验中重复几次，就叫重复双因素试验，分析这种试验数据的方法，也就叫重复双因素方差分析。在多因素实验中如重复几次，就叫重复多因素试验，分析这种数据的方法也就叫重复多因素方差分析。计算重复双因素和重复多因素方差分析方法基本一致，而与没有重复的方差分析则有所不同。

为了了解有重复试验的方差分析概念，我们引江锡基氏对娥氏小蘖的不同生长期和地区的小蘖硷含量研究数据来着重介绍重复双因素方差分析。江氏试验数据见表1.2粗线框内。

表 1·2　　不同生长期和地区的娥氏小蘖硷含量（重复合计×100—100）

两次实验合计 X_1 / 生长区 A / 生长期 B	A_1	A_2	A_5	A_6	A_7	A_8	A_9	A_{10}	ΣX_1（B数相加）	$(\Sigma X_1)^2$（B数相加后平方）
营养期B_1	-14	-10	-52	-25	-4	-110	-12	-102	-329	108200
开花期B_2	-8	6	-32	-10	0	-69	75	-41	-79	6241
果落期B_3	-19	75	0	152	80	-24	124	108	496	246000
ΣX_1（A各数相加）	-41	71	-84	117	76	-203	187	-35	88(T)	360441(ΣT_j^2)
$(\Sigma X_1)^2$（各数相加后平方）	1681	5641	6561	13689	4489	41210	34970	1225	107227(ΣT_i^2)	
Σ_1^2（各数平方后相加）	621	5761	3728	23825	6416	17557	21149	23741	102898(R_1)	

$R=\Sigma X^2$（合并前各实验值平方和）$=53312$

$R_1=\Sigma X_1^2$（合并后各实试值平方和）$=102898$

$$W=\frac{1}{k}\Sigma X_1^2=\frac{1}{2}\times 102898=51449$$

$\Sigma T_i^2=A$的ΣX_1平方和　$\Sigma T_j^2=B$的ΣX_1平方和

$N=8(a)\times 3(b)\times 2(k)$（即重复数）$=48$

$$CT=P=\frac{T^2}{N}=\frac{(88)^2}{48}=165.5$$

$$Q_A=\Sigma T_i^2\times\frac{1}{bk}=107227\times\frac{1}{3\times 2}=17871.17$$

$$Q_B=\Sigma T_j^2\times\frac{1}{ak}=360441\times\frac{1}{8\times 2}=22527.56$$

重复双因素计算，和双因素计算基本一致，它的S_T仍然是各实验数值平方和（ΣX^2），减去总实验值平方的均数$\left(\frac{(\Sigma X)^2}{N}\right)$。$S_T$就是各种因素$S_A$、$S_B$、$S_{AB}$和$S_e$的离均差平方和合计。所不同的是：计算$S_A$、$S_B$、$S_{AB}$时，要将重复实验值相加($X_1$表示)，然后按双因素那样计算(即表1.2)，同时，在计算时要注意增加除以重复数k，它的$\Sigma X_1^2\times\frac{1}{k}-\frac{(\Sigma X_1)^2}{N}$算出的$S_{T1}$值，只包含有$S_A$、$S_B$和$S_{AB}$，而$S_e$就不在其中。把这个概念弄清了，进行下列计算就容易理解了。这种计算方法，同样适合于重复多因素方差分析。

1.计算各种S：

$$S_T=\Sigma X^2-\frac{(\Sigma X_1)^2}{N}=R-CT=53312-165.5=53146.5$$

$$S_A=\Sigma T_i^2\times\frac{1}{bk}-CT=Q_A-CT=17871.17-165.5$$
$$=17705.67$$

$$S_B=\Sigma T_j^2\times\frac{1}{ak}-CT=Q_B-CT=22527.56-165.5$$
$$=22362.06$$

$$S_{T_1}=R_1\frac{1}{k}-CT=W-CT=51449-165.5=51283.5$$

$$S_{A\times B}=S_{T_1}-S_A-S_B=51283.5-17705.67-223620.6$$
$$=11215.77$$

$$S_e=\Sigma X^2-\Sigma X_1^2\times\frac{1}{k}=R-W=53312-51449=1863$$

为了验算上述S是否正确，S_e还有下列算法，如两数值符合，则说明计算没有误差。

$$S_e=S_T-S_A-S_B-S_{AB}=53146.5-17705.67$$
$$-22362.06-11215.77=1863$$

2.计算各种自由度——f

$$f_T = na \times nb \times k - 1 = 8 \times 3 \times 2 - 1 = 47$$

$$f_a = na - 1 = 8 - 1 = 7$$

$$f_b = nb - 1 = 3 - 1 = 2$$

$$f_{ab} = (na - 1) \times (nb - 1) = 7 \times 2 = 14$$

$$f_e = f_T - f_a - f_b - f_{ab} = 47 - 7 - 2 - 14 = 24$$

3.计算各种均方——$\overline{S}$

$$\overline{S}_A = S_A \times \frac{1}{f_a} = 17705.67 \times \frac{1}{7} = 2529.38$$

$$\overline{S}_B = S_B \times \frac{1}{f_b} = 22362.06 \times \frac{1}{2} = 11181.03$$

$$\overline{S}_{AB} = S_{AB} \times \frac{1}{f_{ab}} = 11215.77 \times \frac{1}{14} = 801.13$$

$$\overline{S}_e = S_e \times \frac{1}{f_e} = 1863 \times \frac{1}{24} = 77.64$$

4.计算各种F:

$$F_A = \overline{S}_A / \overline{S}_e = 2529.38/77.64 = 32.58$$

$$F_B = \overline{S}_B / \overline{S}_e = 11181.03/77.64 = 144.01$$

$$F_{A\times B} = \overline{S}_{A\times B} / \overline{S}_e = 801.13/77.64 = 10.31$$

5.F检验——查F表:

当 $f_a=7$，$f_e=24$ $\alpha=0.01$时，其$F=3.50$，而$F_A=32.58$，所以A因素$P<0.01$。即娥氏小蘗的不同栽培地区对娥氏小蘗硷含量有高度显著影响。

当 $f_b=2$，$f_e=24$ $\alpha=0.01$时，其$F=5.61$，而$F_B=144.01$，所以B因素$P<0.01$。即娥氏小蘗的不同采收期，对娥氏小蘗的小蘗硷含量有高度显著影响。

当 $f_{AB}=14$，$f_e=24$，$\alpha=0.01$时，其$F=2.93$，而$F_{AB}=10.31$，所以$A\times B$交互作用$P<0.01$。即娥氏小蘗采收期与生长地区有高度显著交互影响。现列表如下：

娥氏小蘖不同生长期及生长区的小蘖含量方差分析

方差来源	离均差平方和	自由度	均方	F值	显著性
A不同生长区	17705.67	7	2529.38	32.58	* *
B不同生长期	22362.06	2	11181.03	144.01	* *
AB两者交互作用	11215.77	14	801.13	10.31	* *
误差	1863	24	77.64		
总计	531465	47			

$F_{0.01}(7,24)=3.50$

$F_{0.01}(2,24)=5.61$

$F_{0.01}(14,24)=2.93$

§4 多重比较

在前几节单因素和重复双因素分析中，它们的水平数（K）均超过二。以例1.1来讲，薄荷品种有五个薄荷油含量均数。又如例1.2：以生长地区来讲，它就有8个小蘖硷含量均数；以生长期来讲，它有三个小蘖硷含量均数。前节方差分析中F检验证明：薄荷品种和娥氏小蘖生长期、生长地区差异均显著或高度显著，但它并不能告诉我们这些均数之间差异均显著，而有可能有些差异尚不显著。如要进一步弄清这个问题，就要进行多重比较（或称多均数比较）。

在例1.1薄荷选种工作中，本来经方差分析显著后，如果确定从中取两个品种，一般选薄荷油均数较高的二号和三号薄荷就可以。但因为三号薄荷没有一号薄荷抗病强，想选一号薄荷，但不知道一号与三号薄荷含油量差异有没有达到显著水平。这时多重比较就发挥作用了。

又如，在研究娥氏小蘖采收期和生长地区差异时，经F检验，它们均显著，如果要问那几块地区对娥氏小蘖硷含量最有利？那几块地区是居第二位？这也要通过多重比较才行。下面就这个例子进行计算。