区组设计

区　组

把全部试验单元分为若干个组，使得每个组内各试验单元之间的差异尽可能的小，这样的组被称为**区组**.

如何建立区组被称为**区组设计**.

在区组设计中，因子的水平被称为**处理**.

区组设计的例子

例 3.1.1 有 4 种杀虫剂 A_1, A_2, A_3, A_4，它们被称为 4 种处理.

为了比较 4 种杀虫剂对棉田中害虫的杀虫效果高低，特选了 20 块田，每块 1 亩. 如何安排试验呢?

•**随机化设计**：将 20 块田随机的均分为 4 组，分别实施 4 种处理. 其数据分析用单因子方差分析.

评论：20 块棉田试验单元间总有差异，如害虫多少，植物长势，土地肥沃程度等. 这些差别对杀虫效果会带来影响，从而对比较产生干扰.

假如此种差异很微小，实施随机化设计是妥当的.

假如此种差异不可忽略，就要采取随机化区组设计.

- **随机化区组设计**：分二步进行.

第一步， 将 20 块棉田按差异大小排序，将害虫最多的 4 块棉田分为第 1 个区组，将害虫最少的 4 块棉田分为第 5 个区组，其它按序入组.

第二步，在 1 个区组内随机的实施一种杀虫剂.

区组 1	区组 2	区组 3	区组 4	区组 5
A_1	A_2	A_4	A_2	A_3
A_3	A_4	A_1	A_3	A_2
A_4	A_1	A_2	A_4	A_1
A_2	A_3	A_3	A_1	A_4

特点：每个处理（一种杀虫剂）在每

个区组内各种处理也仅出现一

随机化区组设计的一般定义

设(某因子)有 v 个处理需要比较，有 n 个试验单元用于试验.

第一步：把 n 个试验单元均分为 k 个组（$k=n/v$），使每个组内的试验单元尽可能相似，这样的组称为区组.

第二步：在每个区组内对各试验单元以随机方式实施不同处理．这样的设计称为**随机化区组设计**.

若区组大小＝处理个数 v，这样的设计称为**随机化完全区组设计**.

若区组大小＜处理个数 v，这样的设计称为**随机化不完全区组设计**.

以上各种设计都是**平衡的**，若各区组大小不尽相同，称为**不平衡区组设计**.

例： 比较两种人造物质的鞋底（A_1, A_2）的磨损多少.

• 一个人的两只脚是一个合乎情理的区组. 因为一个人的左右鞋的磨损情况是近似相同的. 不同人之间的磨损情况是有差异的.

• 若有 6 人参加试验，且 6 人的鞋的尺码相同，则需对每种人造物质的鞋各制造 3 双，这样有 6 只左鞋和 6 只右鞋，它们外形相同，每个人随机的从中各选一只左鞋和右鞋. 这就完成了随机化完全区组设计.

• 一个月后收回，分别测量其磨损量，然后进行数据分析.

• 讨论：若有三种人造物质的鞋底（A_1, A_2, A_3），那就要采用随机化不完全区组设计.

随机化完全区组设计的数据

在随机化完全区组设计中一般假定有 v 个处理和 b 个区组，共需进行 $n=v\times b$ 次试验，记 y_{ij} 表示第 i 个处理在第 j 个区组内进行试验所得到的观察值.

区组 / 处理	1	2	…	b	(处理)和	均值
1	y_{11}	y_{12}	…	y_{1b}	T_1	$\overline{T}_1$
2	y_{21}	y_{22}	…	y_{2b}	T_2	$\overline{T}_2$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
v	y_{v1}	y_{v2}	…	y_{vb}	T_v	$\overline{T}_v$
(区组)和	B_1	B_2	…	B_b	T	
均值	$\overline{B}_1$	$\overline{B}_2$	…	$\overline{B}_b$		$\overline{y}$

其中：

$$T=\sum_{i=1}^{v}\sum_{j=1}^{b}y_{ij}$$

$$\overline{y}=T/vb$$

例 3.1.3　化学制剂对布料有侵蚀作用，会降低布料的抗拉强度．某工程师研究出一种能抗化学制剂的新型布料，为考察其抗侵蚀作用，特选定 4 种化学制剂和 5 匹布．考虑到布匹间的差异，特在每匹布的中部切取 4 段布料组成一个区组，用随机化完全区组设计安排试验．试验数据如下：

表 3.1.2　例 3.1.3 的试验数据（原始数据−70）

区组 处理	1	2	3	4	5	T_i	$\overline{T}_i$
B_j						T	
$\overline{B}_j$						$\overline{y}$	

1	3
2	3
3	5
4	5

16　1　18　11　-4

随机化完全区组设计的统计模型

在 v 个处理和 b 个区组场合的统计模型如下：

$$y_{ij} = \mu + a_i + b_j + \varepsilon_{ij},\quad i = 1, ,2 \square ,v, j = 1,2, \square ,b$$

其中

y_{ij} 一第 i 个处理在第 j 个区组内的试验结果.

μ 一总均值，是待估参数.

a_i 一第 i 个处理的效应，且满足 $a_1 + a_2 + \square + a_v = 0$

b_j 一第 j 个区组的效应，且满足 $b_1 + b_2 + \square + b_b = 0$

ε_{ij} 一试验误差，服从 $N(0,\sigma^2)$. σ^2 为误差方差.

由此可见：

$$y_{ij} \sim N(\mu + a_i + b_j, \sigma^2)$$

参数估计

利用最小二乘法，可获得各种效应的最小二乘估计.

$$\hat{\mu} = \bar{y}$$

$$\hat{a}_i = \bar{T}_i - \bar{y},\quad i = 1,2,\square\ ,v$$

$$\hat{b}_j = \bar{B}_j - \bar{y},\quad j = 1,2,\square\ ,b.$$

由此可得各拟合值 $\hat{y}_{ij}$ 与残差 e_{ij}：

$$\hat{y}_{ij} = \hat{\mu} + \hat{a}_i + \hat{b}_j = \bar{T}_i + \bar{B}_j - \bar{y}$$

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{T}_i - \bar{B}_j + \bar{y}$$

例 由表 3.1.2 上的诸均值，容易获得各效应估计如下：

$$\hat{\mu} = 2.1, \hat{a}_1 = -1.5, \hat{a}_2 = -0.9, \hat{a}_3 = 0.3, \hat{a}_4 = 2.1,$$

$$\hat{b}_1 = 1.9, \hat{b}_2 = -1.85, \hat{b}_3 = 2.4, \hat{b}_4 = 0.65, \hat{b}_5 = -3.1.$$

总平方和分解公式

在模型(3.1.1)中，全部 vb 个观察值的总平方和有如下分解式：

$$S_T = \sum_{i=1}^{v}\sum_{j=1}^{b}(y_{ij} - \bar{y})^2, \qquad f_T = vb - 1$$

$$= b\sum_{i=1}^{v}(\bar{T}_i - \bar{y})^2 + v\sum_{j=1}^{b}(\bar{B}_j - \bar{y})^2 + \sum_{i=1}^{v}\sum_{j=1}^{b}(y_{ij} - \bar{T}_i - \bar{B}_j + \bar{y})^2$$

处理平方和：

$$S_A = b\sum_{i=1}^{v}(\bar{T}_i - \bar{y})^2, \quad f_A = v - 1$$

区组平方和：

$$S_B = v\sum_{j=1}^{b}(\bar{B}_j - \bar{y})^2, \quad f_B = b - 1$$

误差平方和：

$$S_e = \sum_{i=1}^{v}\sum_{j=1}^{b}(y_{ij} - \bar{T}_i - \bar{B}_j + \bar{y})^2$$

则有

$$S_T = S_A + S_B + S_e, \quad f_T = f_A + f_B + f_e$$

注意：设立区组的目的，就是把区组平方和从总平方和分解出来，免其对处理平方和与误差平方和的干扰，从而加强以后判断的准确性.

各平方和的简化计算公式

$$S_T = \sum_{i=1}^{v}\sum_{j=1}^{b} y_{ij}^2 - \frac{T^2}{vb}, \qquad f_T = vb\text{-}1$$

$$S_A = \frac{T_1^2 + T_2^2 + \cdots + T_v^2}{b} - \frac{T^2}{vb}, \quad f_A = v\text{-}1$$

$$S_B = \frac{B_1^2 + B_2^2 + \cdots + B_b^2}{v} - \frac{T^2}{vb}, \quad f_B = b\text{-}1$$

$$S_e = S_T - S_A - S_B, \qquad f_e = (v\text{-}1)(b-1)$$

平方和的性质

- 可以证明：平方和的期望分别为

$$E(S_A)=(v-1)\sigma^2+b\sum_{i=1}^{v}a_i^2,$$

$$E(S_A)=(b-1)\sigma^2+v\sum_{j=1}^{b}b_j^2,$$

$$E(S_e)=(v-1)(b-1)\sigma^2,$$

由此可见，

$$\hat{\sigma}^2=MS_e=S_e/f_e$$

是误差方差 σ^2 的无偏估计.

- 可以证明：在诸处理效应皆为零时，

$$S_A/\sigma^2\sim\chi^2(v-1)$$

在诸区组效应皆为零时，

$$S_B/\sigma^2\sim\chi^2(b-1)$$

$$S_e/\sigma^2\sim\chi^2((v-1)(b-1))$$

，且三者相互独立.

- 检验诸处理效应皆为零时，所用的检验统计量是

$$F=\frac{MS_A}{MS_e}=\frac{S_A/f_A}{S_e/f_e}\sim F(f_A,f_e)$$

随机化完全区组设计的方差分析表

表 3.1.3　随机化完全区组设计的方差分析表

来源	平方和	自由度	均方和	F 比
处理	$S_A=\frac{1}{b}\sum_{i=1}^{v}T_i^2-\frac{T^2}{vb}$	$f_A=v-1$	$MS_A=S_A/f_A$	$F=\frac{MS_A}{MS_e}$
区组	$S_B=\frac{1}{v}\sum_{j=1}^{b}B_j^2-\frac{T^2}{vb}$	$f_B=b-1$	$MS_B=S_B/f_B$	
误差	$S_e=S_T-S_A-S_B$	$f_e=(v-1)(b-1)$	$MS_e=S_e/f_e$	
总和	$S_T=\sum_{i=1}^{v}\sum_{j=1}^{b}y_{ij}^2-\frac{T^2}{vb}$	$f_T=vb-1$		

例 3.1.4 在化学制剂对布料抗拉强度的试验（例 3.1.3）中，按表 3.1.2 上的数据可算得各平方和及其自由度：

表 3.1.4 例 3.1.4 的方差分析表

来源	平方和	自由度	均方和	F 比
处理	37.8	3	12.6	14.16
区组	91.3	4	22.83	——
误差	10.7	12	0.89	
总和	139.8	19		

给定显著性水平 α =0.05，查其临界值 $F_{0.95}(3,12) = 3.49$，由于 F>3.49，故拒绝 H_0，即四种化学制剂对新型布料的抗拉强度的影响有显著差异，还需改进布料设计．试验误差的方差 σ^2 的估计为 $\hat{\sigma}^2 = 0.89$ $\hat{\sigma} = 0.94$

注释一. 假如不设立区组，则区组平方和并入误差平方和. 数据仍然可按单因子方差分析处理，所得方差分析表如下：

表 3.1.5　把区组从设计中剔除后的不正确分析

来源	平方和	自由度	均方和	F 比
处理	37.8	3	12.6	1.97
误差	102.0	16	6.38	
总和	139.8	19		

对给定 α

=0.05，4 种处理间没有显著差异.

这一错误结论是没有重视区组作用而导致的.

所以在试验中，凡是试验单元间有较大差异时，应运用区组概念去减少数据中的误差.

注释二．可以把区组看作另一个因子

•我们的注意力总是放在 v 个处理间是否有显著差异上．区组就象一个垃圾桶，把区组平方和分解出来就可以了．

•若还要关注区组平方和的大小，即考察区组间是否存在显著差异，可把区组也看作一个因子．要检验如下一对假设：

$$H_0\text{：} b_1 = b_2 = \cdots = b_b = 0$$

$$H_1\text{：诸} b_i \text{中至少有一个不为零}$$

在此假设下，检验统计量为

$$F = \frac{MS_B}{MS_e}$$

•尽管对此检验的合理性存在着争论，但从双因子方差分析看，再一次使用 F 检验也未尝不可，把检验结果作为一种参考也是有价值的．

注释三．随机效应问题，有三种情况

- 仅仅处理效应是随机的.
- 仅仅区组效应是随机的.
- 处理效应和区组效应都是随机的.

这里仅讨论第二种情况，其它情况可作类似讨论.

随机区组效应场合下的统计模型

其中 $$y_{ij} = \mu + a_i + b_j + \varepsilon_{ij},\quad i = 1,,2\square\ ,v,\ j = 1,2,\square\ ,b$$

y_{ij} —第 i 个处理在第 j 个区组内的观察值.

μ

—总均值.

a_i

— 第 i 个固定处理的效应，且满足 $$a_1 + a_2 + \square\ + a_v = 0$$

b_j . b_j $N(0,\sigma_b^2)$

σ_b^2

—第 j 个随机区组的效应，诸 是来正态分布 的一个样本，其中

ε_{ij} $\varepsilon_{ij} \sim N(0,\sigma^2)$ 是区组效应的方差分量. ε_{ij} b_j σ^2

—试验误差， ，诸 与诸 相互独立， 是误

在上述假设下有

其中

$$y_{ij} \sim N(\mu + a_i, \sigma_b^2 + \sigma^2), \quad i = 1,2,\square,v, \; j = 1,\ldots,b$$

μ 和诸 a_i 的最小二乘估计是

$$\hat{\mu} = \bar{y}, \qquad \hat{a}_i = \bar{T}_i - \bar{y}, \quad i = 1,2,\square,v$$

方差分量 σ_b^2 与 σ^2 的无偏估计为

$$\hat{\sigma}_b^2 = \frac{MS_B - MS_e}{v}, \qquad \hat{\sigma}^2 = MS_e$$

这是因为各平方和的期望值如下：

$$E(S_A) = (v-1)\sigma^2 + b\sum_{i=1}^{v} a_i^2$$

$$E(S_A) = (b-1)\sigma^2 + v(b-1)\sigma_b^2$$

$$E(S_e) = (v-1)(b-1)\sigma^2$$

假设检验问题

在随机区组效应场合的模型中，常需作如下二个检验.

• 对固定处理效应要建立的一对假设是：

$$H_0\text{：}\ a_1 = a_2 = \cdots = a_v = 0$$

H_1：诸 a_i 不全为零.

• 对随机区组效应要建立的一对假设是：

$$H_{01}\text{：}\ \sigma_b^2 = 0$$

• 对 H_{11}：$\sigma_b^2 > 0$

H_0 与 H_1 所用的检验统计量是：

$$F = \frac{MS_A}{MS_e}$$

，拒绝域

$$W = \{F > F_{1-\alpha}(f_A, f_e)\}$$

• 对 H_{01} 与 H_{11} 所用的检验统计量是：

$$F = \frac{MS_B}{MS_e}$$

，拒绝域

$$W = \{F > F_{1-\alpha}(f_B, f_e)\}$$

注释四．多重比较问题

• 在固定处理效应场合，不论区组效应是否是随机的，只要处理效应间有显著差异，都应对处理效应施行多重比较.

所用方法同§2.3，但要注意：

随机化完全区组设计中，若并且方差分析确认

Δ 把重复数改为区组数 b，水平数改为处理数.

Δ 误差自由度 $f_e = (v-1)(b-1)$

例 3.1.6 （略）

注释五．模型的适合性

在随机化完全区组设计场合（不管诸效应中是否有随机的），模型中都有正态性和方差齐性等两个问题.

- 在缺少重复情况下，对误差方差齐性的检验还缺少方法．只能从产生数据的过程来判断．若数据是在相同的或类似的环境下产生的，常可以认为误差方差近似达到齐性.

- 关于正态性诊断仍可借助残差分析进行.

例 3.1.7　在例 3.1.3 的 20 个数据中，利用残差计算公式

$$e_{ij} = y_{ij} - \overline{T}_i - \overline{B}_j + \overline{y}$$

可算得 20 个残差，它们（按从小到大次序）是：

-1.35	-1.3	-1.1	-0.8	-0.5	-0.35	-0.25	-0.1	0.05	0.0
0.15	0.15	0.25	0.4	0.4	0.5	0.7	0.9	0.9	1.45

从其正态概率图看出：没有非正态性的严重标志，可认为该组残差近似为正态分布.

图 3.1.3　例 3.1.3 的残差的正态概率图

平衡不完全区组设计(BIB设计)

在随机化完全区组设计中若省去部分试验，余下部分试验就组成一个不完全区组设计.

表3.2.1a 完全区组设计

区组 / 处理	1	2	3	4
1	y_{11}	y_{12}	y_{13}	y_{14}
	y_{21}	y_{22}	y_{23}	y_{24}
	y_{31}	y_{32}	y_{33}	y_{34}
	y_{41}	y_{42}	y_{43}	y_{44}

表3.2.1b 不完全区组设计

区组 / 处理	1	2	3	4
1	y_{11}	y_{12}	y_{13}	
2	y_{21}	y_{22}		y_{24}
3	y_{31}	y_{32}	y_{33}	y_{34}
4		y_{42}		y_{44}

评论：减少总试验
这就不能随
织不完全区

表3.2.1c　不完全区组设计

处理 \ 区组	1	2	3	4
1	y_{11}	y_{12}		y_{14}
2		y_{22}	y_{23}	y_{24}
3	y_{31}	y_{32}	y_{33}	
4	y_{41}		y_{43}	y_{44}

要求：
- 每个区组含有的处理数相等，都为3个；
- 每个处理在不同区组中出现次数相等，都为3次；
- 每对处理在同一区组内相遇次数相等，都为2次.

BIB设计的一般定义

将v个处理安排到b个区组的一个不完全区组设计称为**平衡不完全区组设计(BIB设计)**，假如该设计满足下列三个条件：

1．每个区组都含k个不同处理，k称为区组大小．

2．每个处理都在r个不同区组中出现，r称为处理重复数．

3．任一对处理在 λ λ

个不同区组中相遇，

数．

从这个定义可以看出，

- 一个BIB设计中的v个处理可以得到公平
- 任一BIB设计由五个设计参数v，k，r，b

λ

定理3.2.1 BIB设计存在的必要条件是：在5个设计参数间同时有下列三个关系式：

（1） $vr = bk$

（2）

$$r(k-1) = \lambda(v-1)$$

（3）

$$b \geq v, \quad r \geq k$$

至今人们已经找到很多 BIB 设计.

附表 9 对

$$4 \leq v \leq 10 \quad r \leq 10$$