

完全组内平衡区组设计的性质及应用

杨 林¹, 张建军¹, 廖靖宇¹, 张应山², 吴亚桢¹

(1. 许昌学院数学与统计学院, 河南 许昌 461000)

(2. 华东师范大学金融与统计学院, 上海 200241)

摘 要: 本文研究了完全组内平衡性的相关哲学概念和数学性质. 利用多边矩阵理论, 证明了完全组内平衡区组设计的数学判定条件, 给出了计算机验证完全组内平衡性的方法, 推广了正交表的平衡性质.

关键词: 区组设计; 完全组内平衡; 广义正交表; 区组因子; 再现性

MR(2010) 主题分类号: 62K10

中图分类号: O212.6

文献标识码: A

文章编号: 0255-7797(2014)05-0985-10

1 引言

中国传统文化中的重要统计数据分析理论是象数学理论^[1]. 象数学理论主要用于解决复杂系统的稳定性问题, 它的理论根据是认为: 象或者“道”和人类行为无关. 在中国传统文化的象数学理论中, 比类取象技术是其主要数据分析技术, 这种技术用现代的语言叙述就是试验设计. 按比类取象技术的哲学观点, 试验设计要解决的关键核心问题是要保证其数据分析的结论具有再现性. 也就是说, 不同的人采用不同的试验设计方法对复杂系统的象的研究结论应该是在客观上基本一致的和可以重复再现的, 因为数据分析的客观结论本身和试验工作者的行为以及所用的方法无关. 这样, 不同的试验设计工作者无论采用什么设计表, 无论采用什么公理假设, 无论收集的数据多少, 只要设计表具有若干平衡性, 各个类的数据足够完备, 那么试验数据分析的结论应该是基本一致的和可重复再现的, 因为这些分析结论和试验工作者的行为无关. 也就是说, 论证设计表能否保证数据分析结论具有再现性? 这是试验设计的最基本任务. 为了保证这种再现性, 对研究对象的各种所关注的关系进行公平比较是试验设计的基本方法, 这种公平比较的标准一般称为平衡性. 按象数学理论的哲学含义, 试验设计的平衡性必须具有五方面的含义: 相遇平衡性、组间平衡性、正交平衡性、完全组内平衡性、整体平衡性, 分别对应阴阳五行理论的木、火、土、金、水, 也对应生物能量“人气”方面的生、长、化、收、藏, 具有邻相生、间相克的基本关系属性. 为了保证相应数据分析结论在具有再现性的前提下能够获得最大收获, 本文重点关注完全组内平衡性的哲学和数学方面的定义和性质, 因为这种平衡性是保证试验具有最大收获的最基本条件.

完全组内平衡性是对复杂系统的所有要素 (包括各个子系统和干扰系统内部的所有研究要素) 的功能收益规律的功能认知方面的平衡性, 关注所有要素的功能输出对大系统的影响程度. 这种平衡性要求我们在对复杂系统的所有研究要素的功能输出规律进行认知的时候,

*收稿日期: 2013-09-06

接收日期: 2013-11-26

基金项目: 教育部高等学校博士学科点专项基金资助 (44k550500); 国家自然科学基金资助 (11101381); 河南省高等学校青年骨干教师基金资助 (211GGJS-182); 许昌市科技计划基金资助 (1106008).

作者简介: 杨林 (1982-), 女, 湖北襄阳, 讲师, 主要研究方向: 应用概率统计.

可以不考虑所有要素的实际功能内容, 仅仅当作可以置换安排考虑的类, 即不考虑各个子系统和干扰系统的内部的所有要素的功能输出能力大小、输出功能输出方式的好与坏、输出功能输出效果的高与低等等, 仅仅当作类变量处理, 但要把复杂系统的各个子系统和干扰系统的所有要素的功能输出同时考虑进去, 要把研究的所有要素公平看待, 主要关注的内容是正确识别所有要素的功能输出, 不能对各个子系统或者干扰系统的某些要素的功能输出有所偏好, 要正确把握所有要素的每一个要素. 这种平衡性用中国传统的俗语来说就是: “见好就收”. 用试验设计的语言来说就是: 要求在无公理假设的条件下, 充分考虑干扰因子为区组因子, 不但对各个试验因子的所有水平的重视程度达到最大, 而且对区组因子 (一般作为干扰因子之一) 的各个水平的重视程度也要达到最大 (对区组因子和试验因子的所有水平效应的估计要无偏和方差最小, 保证对区组因子和试验因子的所有水平的不同设计安排方式不影响对各个试验水平的显著性分析结论). 而对试验因子和区组因子的所有水平效应识别的再现性是正确把握复杂系统的所有要素的最基本的要求, 也是试验设计的功能收获阶段认知的基本要求之一. 这种平衡性的基本特点是要求对不同因子 (包括试验因子和区组因子) 的所有水平观察程度不能任意, 要能够使得试验工作者正确识别所有因子的所有水平的效应, 包括正确识别总体均值和交互效应 (相应的估计无偏和方差最小), 要求在干扰条件基本相同的每个试验区组之内对每个试验因子的所有水平都有所观察 (完全性), 并且观察程度相同 (组内平衡), 此时不同区组的试验可以看成对试验因子效应的重复观察, 使得对所有水平效应的估计无偏并且方差最小, 所以这种平衡性被称为完全组内平衡性. 这种完全组内平衡性, 在中国传统文化中称为 “金”, 就像一年的秋季, 是为了研究复杂系统的功能收获阶段的收获大小, 它主要关注世间万物的功能收获机制, 在生物能量 “人气” 方面称为 “收”, 讲究事物功能收获的完全性.

试验设计一般是为了研究复杂系统的稳定性问题. 考虑复杂系统相应的输出变量 (响应变量) y 和对输出可能有影响的众多相应的输入变量 (自变量) $x_0, x_1, \dots, x_m, \omega$. 相应的统计分析模型可以写成

$$\begin{aligned} y &= f(x_0, \dots, x_m, \omega) = g(x_0, \dots, x_m) + \varepsilon_g, \\ g(x_0, \dots, x_m) &= E(f(x_0, \dots, x_m, \omega) | x_0, \dots, x_m), \\ \varepsilon_g &= f(x_0, \dots, x_m, \omega) - E(f(x_0, \dots, x_m, \omega) | x_0, \dots, x_m), \end{aligned} \quad (1.1)$$

其中 $E(f(x_0, \dots, x_m, \omega) | x_0, \dots, x_m)$ 是统计学中的条件期望. 对于一般的复杂系统, 不但所有函数 f, g, ε_g 是未知的, 而且所有自变量 $x_0, x_1, \dots, x_m, \omega$ 也是未知的. 这种模型称为自由函数模型, 或者简称自由模型.

相应的输入变量 (自变量) $x_0, x_1, \dots, x_m, \omega$ 一般被分成三类: 区组因子 x_0 、试验因子 $x_1, \dots, x_m$ 、未知干扰因子 ω .

区组因子 x_0 也可以称为已知可控的干扰因子, 它的一个水平 (一个区组) 是在已知可控的干扰因子所处条件基本相同时所做的重复试验. 在象数学的比类取象技术中, 每次试验数据一般称为 “象” 数据, 所有数据形成的空间称为 “象空间”. 在许多场合, 重复数据的个数有时受到客观条件的限制, 人类无法控制. 有时因为人类的偏好程度的不同, 对有些区组内的数据可能特别关注, 获得的数据个数很多, 而对另外一些区组内的数据可能不太关注, 获得的数据可能较少, 所以所有这些数据的基本属性可能是不一样的, 或者说它们来源于不同的分布, 但客观结论只能有一个, 它和人类行为无关. 为了保证数据分析结论具有再现性, 合理的

数据分析结论至少应和每个区组内的数据多少基本无关.

试验因子 $x_1, \dots, x_m$ 一般对应人类的可以调整变化因子, 称为可控因子. 由于试验因子 $x_1, \dots, x_m$ 的试验水平对应数据的选取和人类关系很大, 为了排除人类行为对试验数据分析结论的影响, 使得数据分析结论具有再现性, 象数学的比类取象技术要求试验因子必须是没有大小和没有序关系的类变量. 试验因子 $x_1, \dots, x_m$ 关于每个试验输出数据对应的一个水平组合称为一个“卦”, 试验因子对应的所有“卦”形成的空间为“卦空间”.

未知干扰因子 ω 是每种试验设计必须要考虑的因子, 一般情况下它对应试验设计的试验次序的安排, 其水平个数一般和试验次数 n 完全相同. 试验设计也必须设法排除这个未知干扰因子 ω 对试验数据(象)分析结论的影响. 西方试验设计方法一般采用随机化安排试验次序来排除这个干扰因子的影响, 但象数学一般要求用幻方设计表来安排试验次序来排除这个干扰因子的影响, 因为象数学认为随机化本身也和人类行为有关, 很难识别试验数据是否一定是随机化试验获得的.

在复杂系统输入变量(自变量) $x_0, x_1, \dots, x_m, \omega$ 和输出变量(响应变量) y 确定以后, 对应任意的系统函数 f , 条件期望 $E(f(x_0, \dots, x_m, \omega)|x_0, \dots, x_m)$ 将是确定的函数, 记这个条件期望函数为 $g(x_0, x_1, \dots, x_m)$, 为了保证试验数据的分析结论具有再现性, 这个条件期望函数是不能假设的. 同样试验误差 $\varepsilon_g = f(x_0, \dots, x_m, \omega) - E(f(x_0, \dots, x_m, \omega)|x_0, \dots, x_m)$, 也是确定的函数, 也是不能假设的. 尽管这些函数不能假设, 但可以根据系统函数 f 推导出它们所具有的性质:

$$\begin{aligned} E(\varepsilon_g|x_0, \dots, x_m) &= E(f|x_0, \dots, x_m) - E[E(f|x_0, \dots, x_m)] = 0, \\ \text{Var}(\varepsilon_g|x_0, \dots, x_m) &= E(\varepsilon_g^2|x_0, \dots, x_m) = \sigma_{x_0, \dots, x_m}^2 \geq 0, \\ \text{Cov}((g, \varepsilon_g)|x_0, \dots, x_m) &= E[(g - Eg)(\varepsilon_g - E\varepsilon_g)|x_0, \dots, x_m] = 0. \end{aligned}$$

这些性质都不是假设, 只要系统函数 f 保证上述计算有意义, 这些结论都是可以证明的. 可以证明的结论当然不是假设.

完全平衡区组设计要考虑的设计问题是: 在上述自由模型下, 如何选取试验因子的各个水平在每一个区组内的平衡性质, 才能保证在试验数据分析结论具有再现性的条件下, 使得试验设计的收获最大? 此收获最大的含义用统计语言来说就是: 所有估计是无偏的和方差最小的.

在正交表的每个区组内, 各个水平都出现并且出现的次数相同. 这种性质保证对每个试验因子的各个水平, 在任何区组内都可以公平地进行比较. 在这种条件下, 对系统的每个试验因子的各个水平, 仅仅利用一个区组内的数据就可以公平比较. 多个区组试验相当于试验比较的重复, 重复性是保证试验具有再现性的常用试验条件之一. 它不仅使得试验因子各个水平效应估计精确, 也使得区组因子的各个区组水平效应也估计精确, 试验分析结论不受试验因子和区组因子表头设计排列安排方式的任何影响, 这是最好的一种试验收获状态, 也是比较难以满足的一种试验条件. 正因为这种性质保证在各个区组内近似相同的试验条件下, 可以在每个区组内正确比较各个试验因子的各个水平之间的显著性, 取得的是试验所有因子的最大的收获, 所以这种性质是一种收获平衡性, 在五行哲学中称为金, 表示万事之收获阶段, 就像一年中的秋季. 这种平衡保证收获最大, 符合完全组内平衡的哲学含义. 本文把正交表的这种平衡性质推广到一般具有组间平衡性质的广义正交表, 给出它们的数学定义, 并从区组设计表和数据分析的角度反复分析它们的优良性质.

2 区组设计的完全组内平衡性

定义 2.1 ^[5] 对于任意一个水平 $x, x \in \{1, \dots, v\}$, 设其在区组设计

$$D_{b \times k}(v) = [B_1, \dots, B_b]^T = (b_{ij})$$

的第 i 行 B_i^T 出现的次数为 $r_i(x) = \{j : b_{ij} = x\}$. 称 $r_*(x) = \sum_{i=1}^b r_i(x)$ 为水平 x 在整个区组设计中的重复次数, 简称为水平 x 的重复数. 称 $\lambda_*(x, y) = \sum_{i=1}^b r_i(x)r_i(y)$ 为不同水平组合 (x, y) 的相遇次数, 简称相遇数. 并称 $r(x) = \sum_{i=1}^b \frac{r_i(x)}{k_i}$ 为水平 x 在整个区组设计中的重复程度, 简称为水平 x 的重复度, 称 $\lambda(x, y) = \sum_{i=1}^b \frac{r_i(x)r_i(y)}{k_i}$ 为不同水平组合 (x, y) 的相遇程度, 简称相遇度. 并称 $\mathbf{r} = (r(1), \dots, r(v))$ 为重复度行向量, 也称 $\mathbf{k} = (k_1, \dots, k_b)$ 为区组大小列向量, 也称 $\mathbf{\Lambda} = (\lambda(x, y))_{v \times v}$ 为相遇度矩阵.

定义 2.2 ^[5] 一个区组设计称为完全的, 如果在每一个区组内包含设计的所有水平. 一个区组设计称为不完全的, 如果存在一个区组内不包含设计的所有水平.

一个区组设计称为组内平衡的, 如果在每一个区组内的元素出现的次数相同.

一个区组设计称为完全组内平衡, 如果这个区组设计既是完全的, 又是组内平衡的.

上述完全组内平衡的定义保证在相应区组设计的每个区组内的数据不但含有所有水平之间的比较关系的所有信息, 而且所有比较关系的观察程度也是相同的, 所以, 在不考虑区组水平干扰大小的条件下, 仅仅使用一个区组内的数据就可以做出试验的分析结论. 多个区组仅仅相当于对分析结论的在多种区组水平大小干扰的条件下重复再现, 这种重复再现性保证并对区组因子 (已知可控干扰) 的水平干扰程度可以有效精确估计.

上述关于区组设计的不完全性的定义与组合数学内的平衡不完全区组设计 (BIBD) 的定义不同, 前者仅仅要求存在一个区组内不包含所有水平即可, 而后者要求所有区组内都不包含所有水平.

例 1 设 $B_1 = (1, 2)^T$, $B_2 = (1, 2, 1)^T$, 那么 B_1, B_2 就组成了一个区组设计, 记为

$$D_{3 \times (2,3)^T}(2) = [B_1, B_2]^T = \begin{pmatrix} 1 & 2 \\ 1 & 2 & 1 \end{pmatrix},$$

由于所有水平 1, 2 在各个区组内出现, 所以区组设计 $D_{3 \times (2,3)^T}(2)$ 是完全的.

由于在区组 $B_2 = (1, 2, 1)^T$ 内, 水平 1 出现 2 次, 而水平 2 出现一次, 所以区组设计 $D_{3 \times (2,3)^T}(2)$ 不是组内平衡的.

但在区组 $B_1 = (1, 2)^T$ 内, 水平 1 和 2 都出现 1 次, 所以区组设计 $D_{3 \times (2,3)^T}(2)$ 关于区组 B_1 可以说是组内平衡的.

由于 $r(1) = \frac{1}{2} + \frac{2}{3} = \frac{7}{6}$, $r(2) = \frac{1}{2} + \frac{1}{3} = \frac{5}{6}$, 所以 $D_{3 \times (2,3)^T}(2)$ 不是组间平衡的. 但 $\mathbf{r} = (r(1), r(2))$ 的极差为 $R_c = \max(\mathbf{r}) - \min(\mathbf{r}) = \frac{1}{3} = \frac{1}{\max_{1 \leq i \leq b} k_i}$. 从均匀性的角度考虑, 这是二水平设计两个区组 5 次试验中, 水平重复度频率分布最好的情况, 或者说是最接近组间平衡的条件. 在实际应用中, 这种水平重复度频率分布的区组设计也是可以考虑的.

3 完全组内平衡区组设计的性质

记区组设计 $D_{b \times k}(v) = [B_1, \dots, B_b]^T$ 的第 i 个区组 B_i 的示性函数为

$$I_{B_i}(x) = \begin{cases} 1, & \text{水平 } x \text{ 在区组 } B_i \text{ 出现;} \\ 0, & \text{水平 } x \text{ 在区组 } B_i \text{ 不出现.} \end{cases}$$

根据这个示性函数可以方便地推导完全组内平衡区组设计的性质.

性质 3.1 对于任意区组设计 $D_{b \times k}(v)$, 我们有

1. 区组设计 $D_{b \times k}(v)$ 是组内平衡的当且仅当 $\frac{r_i(x)}{k_i} = \frac{I_{B_i}(x)}{c_i}$ 其中 $c_i = \sum_{y=1}^v I_{B_i}(y)$ 为区组 B_i 包含不同水平的个数, 此相当于 $r_i(x) = r_i I_{B_i}(x)$, 其中 $r_i = \frac{k_i}{c_i}$ 为区组 B_i 内水平的重复次数.

2. 区组设计 $D_{b \times k}(v)$ 是完全组内平衡的当且仅当 $\frac{r_i(x)}{k_i} = \frac{1}{v}$. 此时, 区组设计 $D_{b \times k}(v)$ 是相遇平衡和组间平衡的.

3. 在 $k_i = k_1, \forall i \in \{1, \dots, b\}$ 的条件下, 如果区组设计 $D_{b \times k}(v)$ 是完全组内平衡的, 那么区组设计 $D_{b \times k}(v)$ 是相遇平衡和组间平衡的, 并且存在置换矩阵 $\sigma_1, \dots, \sigma_b$ 使得

$$D_{b \times k}(v) = [\sigma_1((v) \otimes \mathbf{1}_c), \dots, \sigma_b((v) \otimes \mathbf{1}_c)]^T,$$

其中 $(v) = (1, \dots, v)^T$.

证 1. 区组设计 $D_{b \times k}(v)$ 是组内平衡的, 当且仅当存在整数 r_i 使得 $r_i(x) = r_i I_{B_i}(x)$, 设 $c_i = \sum_{y=1}^v I_{B_i}(y)$, 其为区组 B_i 内不同水平的个数, 那么此条件 $r_i(x) = r_i I_{B_i}(x)$ 相当于区组设计

$D_{b \times k}(v)$ 在 B_i 的元素的重复数为 $r_i = \frac{\sum_{y=1}^v r_i(y)}{c_i} = \frac{k_i}{c_i}$.

2. 区组设计 $D_{b \times k}(v)$ 是完全的相当于 $I_{B_i}(x) = 1, \forall x$, 即 $c_i = v$. 由于 1. 可知: 区组设计 $D_{b \times k}(v)$ 是完全组内平衡的当且仅当 $\frac{r_i(x)}{k_i} = \frac{I_{B_i}(x)}{c_i} = \frac{1}{v}$. 此时,

$$\lambda(x, y) = \sum_{i=1}^b \frac{r_i(x)r_i(y)}{k_i} = \sum_{i=1}^b \frac{k_i}{v^2} = \frac{n}{v^2},$$

$$r(x) = \sum_{i=1}^b \frac{r_i(x)}{k_i} = \sum_{i=1}^b \frac{1}{v} = \frac{b}{v}.$$

即区组设计 $D_{b \times k}(v)$ 是相遇平衡和组间平衡的.

3. 由 2 知, 区组设计 $D_{b \times k}(v)$ 是相遇平衡和组间平衡的. 记 $r_i(x) = k_1/v = c$, 那么存在置换矩阵 σ_i 使得 $B_i = \sigma_i((v) \otimes \mathbf{1}_c)$. 即证.

当 $r_i(x) \leq 1, k_i = k_1 < v, \forall i \in \{1, \dots, b\}$ 时, 对应的相遇平衡、组间平衡、组内平衡, 但所有区组都不完全的区组设计就是通常的平衡不完全区组 (BIB) 设计.

当 $r_i(x) = 1, k_i = v = b, \forall i \in \{1, \dots, b\}$ 时, 对应的相遇平衡、组间平衡、完全组内平衡的区组设计就是通常的拉丁方设计 [2,4].

当 $r_i(x) = c, k_i = vc, \forall i \in \{1, \dots, b\}$ 时, 对应的相遇平衡、组间平衡、完全组内平衡的区组设计就是通常的一种拉丁矩阵设计 [2,4].

当 $r_i(x) \leq 1, k_i < v, \forall i \in \{1, \dots, b\}, \exists k_i \neq k_{i'}$ 时, 对应的相遇平衡、组间平衡、但所有区组都不完全的区组设计的区组设计就是通常的部分平衡不完全区组 (PBIB) 设计的一种定义形式. 另外的部分平衡不完全区组 (PBIB) 设计的相遇平衡和组间平衡的定义形式为

$$\lambda_*(x, y) = \sum_{i=1}^b r_i(x)r_i(y) = \lambda_* > 0, r_*(x) = \sum_{i=1}^b r_i(x) = r_* > 0.$$

因为这种相遇平衡和组间平衡与区组大小的分解式 $n = \sum_{i=1}^b k_i$ 有关, 难以保证试验设计的数据分析结论具有再现性, 所以, 我们不推荐这种形式的部分平衡不完全区组 (PBIB) 设计.

区组设计 $D_{b \times k}(v)$ 是完全组内平衡的也是判断其是否相遇平衡和组间平衡的方便条件.

例 2 设区组设计为 $D_{3 \times 21_3}^1(2) = \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{pmatrix}$, 由于它是完全组内平衡的, 所以它是相遇

平衡的, 也是组间平衡的.

4 完全组内平衡区组设计的判定定理

定理 4.1 (完全组内平衡区组设计的判定定理) 设区组设计 $D = D_{b \times k}(v)$ 的关联矩阵为 $C = \text{Design}(\text{Vec}(D))$, 设 $n \times n$ 投影矩阵 $\mathcal{P}_k = \text{block}(P_{k_1}, \dots, P_{k_b}), \tau_k = I_n - \mathcal{P}_k$, 广义关联矩阵为 $W = \tau_k C$, 那么

1. 区组设计 $D_{b \times k}(v)$ 是组内平衡的当且仅当 $\mathcal{P}_k C$ 的各个行向量的非零元素都相同.
2. 区组设计 $D_{b \times k}(v)$ 是完全的当且仅当 $\mathcal{P}_k C$ 没有零元素 0.

证 1. 设区组设计 $D = D_{b \times k}(v)$ 的关联矩阵 $C = \text{Design}(\text{Vec}(D))$ 的第 i 行块为 $C_{(i)}$. 按区组设计 $D = D_{b \times k}(v)$ 的关联矩阵 $C = \text{Design}(\text{Vec}(D))$ 的定义可知, $C_{(i)}$ 的各个列的和就是 $(r_i(1), \dots, r_i(v))$, 即 $r_i(x) = C_{ix}^T \mathbf{1}_{k_i} = C_{ix}^T C_{ix}$, 此处 C_{ix} 是 $C_{(i)}$ 的第 x 列. 这样, 对水平 x , 记 $e_x(v)$ 是 v 维向量空间 R^v 的标准基向量, 则矩阵 $P_{k_i} C_{(i)}$ 的 x 列向量是

$$P_{k_i} C_{(i)} e_x(v) = \mathbf{1}_{k_i} \cdot \frac{1}{k_i} \cdot \mathbf{1}_{k_i}^T C_{ix} = \mathbf{1}_{k_i} \cdot \frac{r_i(x)}{k_i}.$$

如果区组设计 $D = D_{b \times k}(v)$ 是组内平衡的, 那么

$$P_{k_i} C_{(i)} e_x(v) = \mathbf{1}_{k_i} \cdot \frac{r_i}{k_i} I_{B_i}(x) = c_i \mathbf{1}_{k_i} I_{B_i}(x), \forall x = 1, \dots, v.$$

这说明 $\mathcal{P}_k C$ 的各个行向量的非零元素都相同, 即至多只有两个值, 其中一个值是 0, 另一个值是区组 B_i 包含不同水平的个数.

如果 $\mathcal{P}_k C$ 的各个行向量的非零元素都相同, 那么所有非零的 $\frac{r_i(x)}{k_i}, \forall x = 1, \dots, v$, 是常数, 即 $r_i(x) = r_i I_{B_i}(x)$, 这说明区组设计 $D = D_{b \times k}(v)$ 是组内平衡的.

2. 由 1 我们知道: $P_{k_i} C_{(i)} e_x(v) = \mathbf{1}_{k_i} \frac{r_i(x)}{k_i}$, 它非零当且仅当水平 x 在区组 B_i 出现, 即 $P_{k_i} C_{(i)} e_x(v)$ 没有零元素. 由于 i 的任意性, 所以 $\mathcal{P}_k C$ 没有零元素和区组设计的完全性等价. 即证.

定理 4.1 给我们了一个方便的用计算机验证一个区组设计是否组内平衡和完全性的条件.

例 3 设区组设计为 $D_{3 \times 21_3}^1(3) = \begin{pmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 1 \end{pmatrix}$, 如下程序可以用计算机来验证这个设计是

否组内平衡和完全的:

```
proc iml; d0={1 1, 2 2, 3 3}; d={1 2, 2 3, 3 1};
C0=Design(d0[loc(d0)]); C=Design(d[loc(d)]); P=C0*ginv(C0'*C0)*C0';
PC=P*C; v=ncol(C); n=nrow(c0); do i=1 to n; r=PC[i, ]; r=r[loc(r)];
cc=cc//nrow(r); RC=RC//(max(r)-min(r)); end; print PC cc v RC;
run;quit;
```

计算可知

$$PC = \mathcal{P}_k C = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0 & 0.5 \end{pmatrix}, CC = \begin{pmatrix} 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \end{pmatrix}, v = 3, RC = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

程序中 CC 是区组内不同水平的个数, 所以 $CC \geq v$ 是区组设计为完全的判定条件, 另外 RC 是各个行非零元素组成的向量 r 的极差, 所以 $RC = 0$ 是区组设计为组内平衡的判定条件. 程序中矩阵运算 $//$ 是矩阵并行运算. 用定理 4.1 可知这个区组设计是组内平衡的, 但不是完全的.

将程序中的区组设计 D 换成例 2 内的区组设计 $D_{3 \times 21_3}^1(2)$, 即将程序中的区组设计 D 换成 $D = \{1\ 2, 1\ 2, 1\ 2\}$, 计算可知

$$PC = \mathcal{P}_k C = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \\ 0.5 & 0.5 \\ 0.5 & 0.5 \\ 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, CC = \begin{pmatrix} 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \end{pmatrix}, v = 2, RC = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

其中 $CC = 21_6 \geq 2$ 成立说明区组设计 $D_{3 \times 21_3}^1(2)$ 是完全的, $RC = 0_6$ 说明区组设计 $D_{3 \times 21_3}^1(2)$ 是组内平衡的.

5 完全组内平衡区组设计的数据分析性质

象数学试验设计的一般模型和试验设计表无关, 都是不自生模型, 需要根据自由模型 (1.1) 派生.

在自由模型 (1.1) 中, 对于一般的系统函数 $f(x_0, \dots, x_m, \omega)$, 考虑条件期望 $g(x_0, \dots, x_m) = E(f|x_0, \dots, x_m)$, 设 $M = \{j_1, \dots, j_t\} \subseteq \Omega = \{0, \dots, m\}$, 记 $x_M = (x_{j_1}, \dots, x_{j_t})$, $|M| = t$, 可以进行如下分解

$$g(x_0, \cdots, x_m) = \sum_{M \subseteq \Omega} \tau_{x_M}. \quad (5.1)$$

其中 $\tau_{x_M} = \sum_{N: N \subseteq M} (-1)^{|M|-|N|} E(g|x_N) = \sum_{N: N \subseteq M} (-1)^{|M|-|N|} E(f|x_N), \forall M \subseteq \Omega$.

在系统函数平方可积的基本条件下, 可以证明的基本定理是

$$E\tau_\emptyset = Eg = Ef, E\tau_{x_M} = 0, \forall M \neq \emptyset, \quad (5.2)$$

$$\text{Var}(g) = \sum_{\emptyset \neq M \subseteq \Omega} \text{Var}(\tau_{x_M}), \quad (5.3)$$

其中方差 $\text{Var}(g) = E(g - Eg)^2, \text{Var}(\tau_{x_M}) = E\tau_{x_M}^2 \geq 0, \forall M \neq \emptyset$.

按不自生逻辑, 为了东西方试验设计模型的接轨, 根据模型 (1.1), 记 $\tau_{x_j} = E(f|x_j) - Ef, j = 0, \cdots, m$, 较为合理的试验设计派生模型如下.

设一复杂系统的输出变量数据为 (y_{ij}) , 可用如下模型来表示

$$y_{ij} = \mu + \alpha_i + \beta_{1b_{ij}^1} + \cdots + \beta_{mb_{ij}^m} + \varepsilon_{ij}, \quad i = 1, \cdots, b, j = 1, \cdots, k_i, \\ \varepsilon_{ij} \stackrel{i.i.d.}{\sim} N(0, \sigma^2), \sum_{i=1}^b \alpha_i = 0, \sum_{x=1}^{v_t} \beta_{tx} = 0, \quad (5.4)$$

其中 i.i.d 表示独立同分布, $n = k_1 + \cdots + k_b$ 是试验次数, $\mu = Ef, \alpha_i = \tau_{x_0=x_0(i)}, \beta_{tx} = \tau_{x_t=x_t(x)}$ 为参数, 满足和设计表无关的约束 $\frac{1}{b} \sum_{i=1}^b \alpha_i = \frac{1}{b} \sum_{i=1}^b \tau_{x_0=x_0(i)} \approx E\tau_{x_0} = 0, \frac{1}{v_t} \sum_{x=1}^{v_t} \beta_{tx} = \frac{1}{v_t} \sum_{x=1}^{v_t} \tau_{x_t=x_t(x)} \approx E\tau_{x_t} = 0$. 区组设计是 $D_{b \times k}^t(v_t) = (b_{ij}^t), 1 \leq b_{ij}^t \leq v_t, t = 0, \cdots, m$, 其中 $D_{b \times k}^0(v_0) = (b_{ij}^0) = [11_{k_1}, \cdots, b1_{k_b}]^T, v_0 = b, b_{ij}^0 = i$ 是区组示性表. 相应的广义正交表为

$$GL_n(b^1; v_1 \cdots v_m) = [a_0, \cdots, a_m] = [\text{Vec}(D_{b \times k}^0(v_0)), \cdots, \text{Vec}(D_{b \times k}^m(v_m))].$$

派生模型 (5.4) 要解决的问题是和试验设计无关的自由模型 (1.1) 的问题. 用统计方法给出的结论, 完全可以用自由模型 (1.1) 来模拟检验试验设计所用方法的合理性.

根据文献 [6-8] 的研究结论, 如果区组设计是相遇平衡和正交平衡的, 即相应的行列设计为广义正交表, 那么对上述象数学试验设计派生模型的数据分析结论具有再现性. 根据引言的完全组内平衡的哲学含义的分析, 完全组内平衡条件应该使得总体平均值和区组因子的水平估计无偏方差最小.

同样根据文献 [6-8] 的研究结论, 在模型 (5.4) 的条件下, 如果相应的行列设计表为广义正交表, 那么对参数 μ 和 α_i , 其传统估计一般使用的是不正确的数据总平均

$$\hat{\mu} = \bar{y}_{**} = \frac{1}{n} \sum_{i=1}^b \sum_{j=1}^{k_i} y_{ij} = \frac{1}{n} \sum_{i=1}^b k_i \bar{y}_{i*}, \bar{y}_{i*} = \frac{1}{k_i} \sum_{j=1}^{k_i} y_{ij},$$

和区组数据平均与数据总平均的差

$$\hat{\alpha}_i = \bar{y}_{i*} - \frac{1}{n} \sum_{i=1}^b k_i \bar{y}_{i*} = \bar{y}_{i*} - \bar{y}_{**}, i = 1, \cdots, b.$$

这些估计不但和区组设计的区组大小的分解式 $n = k_1 + \cdots + k_b$ 有关, 而且也受到试验因子水平效应大小的影响, 无法保证试验数据的分析结论具有再现性. 一般广义正交表估计的正确形式应该是

$$\begin{aligned}\hat{\mu} &= \bar{y}_{**}^w - \sum_{t=1}^v \sum_{x=1}^{v_t} \left(\frac{r^t(x)}{b} - \frac{1}{v_t} \right) \hat{\beta}_{tx} \sim N(\mu, \sigma^2/\lambda_\mu), \\ \hat{\alpha}_i &= \bar{y}_{i*} - \bar{y}_{**}^w - \sum_{t=1}^v \sum_{x=1}^{v_t} \left(\frac{r_i^t(x)}{k_i} - \frac{r^t(x)}{b} \right) \hat{\beta}_{tx}, \hat{\alpha} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_b)^T \sim N_b(\alpha, \tau_b \sigma^2/(\lambda_\alpha b)), \\ \hat{\beta}_{tx} &= \frac{1}{\lambda_t v_t} \sum_{i=1}^b r_i(x) (\bar{y}_{i*}(x) - \bar{y}_{i*}), \hat{\beta}_t = (\hat{\beta}_{t1}, \cdots, \hat{\beta}_{tv_t})^T \sim N_{v_t}(\beta_t, \tau_{v_t} \sigma^2/(\lambda_t v_t)),\end{aligned}$$

其中

$$\begin{aligned}\lambda_\mu &= 1/\left(\frac{\bar{W}}{b} + \sum_{t=1}^v \frac{1}{\lambda_t v_t} \sum_{x=1}^{v_t} \left(\frac{r^t(x)}{b} - \frac{1}{v_t} \right)^2 \right), \bar{W} = \frac{1}{b} \sum_{i=1}^b \frac{1}{k_i}, \\ \lambda_\alpha &= 1/\left(b\bar{W} + \sum_{t>1}^v \frac{b}{(b-1)\lambda_t v_t} \sum_{x=1}^{v_t} \left(\frac{r_i^t(x)}{k_i} - \frac{r^t(x)}{b} \right)^2 \right), \\ \bar{y}_{**}^w &= \frac{1}{b} \sum_{i=1}^b \bar{y}_{i*}, \bar{y}_{i*} = \frac{1}{k_i} \sum_{j=1}^{k_i} y_{ij}, \bar{y}_{i*}(x) = \frac{1}{r_i(x)} \sum_{j: b_{ij}^t=x} y_{ij}, \\ r_i(x) &= |\{j : b_{ij}^t = x\}|, r^t(x) = \sum_{i=1}^b \frac{r_i^t(x)}{k_i}, \lambda_t = \sum_{i=1}^b \frac{r_i^t(x)r_i^t(y)}{k_i}, \forall x \neq y.\end{aligned}$$

$\tau_r = I_r - P_r$, I_r 是 r 阶单位矩阵, $P_r = \frac{1}{r} 1_r 1_r^T$ 是秩为 1 的投影矩阵, 1_r 是元素都是 1 的 r 维列向量.

根据区组设计的完全组内平衡性质可知如下条件是一定成立的,

$$\frac{r^t(x)}{b} - \frac{1}{v_t} = 0, \frac{r_i^t(x)}{k_i} - \frac{r^t(x)}{b} = 0, \forall x \in \{1, \cdots, v_t\}, t = 1, \cdots, m.$$

利用这个条件, 可以把具有完全组内平衡性质的广义正交表的数据分析结论总结成如下定理.

定理 5.1 考虑模型 (5.4). 在广义正交表的基础上, 如果相应的区组设计也满足完全组内平衡性质, 那么对总体均值、区组水平效应、试验因子水平效应的估计为

$$\hat{\mu} = \frac{1}{b} \sum_{i=1}^b \bar{y}_{i*} = \bar{y}_{**}^w, \hat{\alpha}_i = \bar{y}_{i*} - \frac{1}{b} \sum_{i=1}^b \bar{y}_{i*} = \bar{y}_{i*} - \bar{y}_{**}^w, \hat{\beta}_{tx} = \frac{1}{b} \sum_{i=1}^b r_i(x) (\bar{y}_{i*}(x) - \bar{y}_{i*}).$$

所有估计是无偏的和方差最小的, 并且在 n 比较大时, 其与区组大小 k_i 基本无关的, 且具有试验数据分析结论再现性, 使得试验收获最大.

6 结论

本文首先对试验设计的完全组内平衡性概念进行了哲学层面的分析. 并根据完全组内平衡区组设计的哲学分析结论, 对区组设计的组合条件的数学定义和数学判定方法进行了重点研究. 其数学定义方法不但具有完全组内平衡哲学方面的意义, 也具有组合分析方面的重要意义, 其数学判定方法比较简单, 可以用 SAS 程序快速验证. 具有完全组内平衡性质的广义正交表的数据分析结论, 不但能够使得试验因子的各个水平效应的估计无偏和方差最小, 而且能够使得对试验的总体均值和区组水平效应的估计是无偏的和方差最小的, 并且在数据个数较大时, 其余区组大小 k_i 基本无关, 具有试验数据分析结论再现性. 因此, 我们推荐在试验设计中, 在可能的条件下, 要尽可能的采用这种具有完全组内平衡性质的广义正交表. 一般地, 我们把这种设计表称为近似正交表.

参 考 文 献

- [1] Zhang Yingshan, Shao Weilan. Image mathematics-mathematical intervening principle based on “Yin Yang Wu Xing” theory in traditional chinese mathematics[J]. Applied Mathematics, 2012, 3(6): 617–636.
- [2] 张应山. 多边矩阵理论 [M]. 北京: 中国统计出版社, 1993.
- [3] 茆诗松, 周纪芃, 陈颖. 试验设计 [M]. 北京: 中国统计出版社, 2004.
- [4] 张应山, 卢一强. 广义拉丁矩阵的计数 [J]. 河南师范大学学报, 1997, 25(1): 9–11.
- [5] 罗纯, 潘长缘. 穷举法寻找正交平衡区组设计 [J]. 应用概率统计, 2011, 27(1): 1–13.
- [6] 吴亚桢, 廖靖宇, 张应山, 田萍, 张建军. 正交平衡区组设计统计分析模型的参数估计 [J]. 数学的实践与认识, 2012, 42(2): 200–208.
- [7] 吴亚桢, 廖靖宇, 张应山, 田萍, 张建军. 正交平衡区组设计统计分析模型参数估计的分布特征研究 [J]. 数学的实践与认识, 2012, 42(3): 212–221.
- [8] 吴亚桢, 廖靖宇, 张应山, 田萍, 张建军. 正交平衡区组设计统计分析模型参数估计的矩阵表达 [J]. 数学的实践与认识, 2012, 42(16): 185–189.

PROPERTIES AND APPLICATION OF BLOCK DESIGNS WITH COMPLETE-INTERNAL BALANCE

YANG Lin¹, ZHANG Jian-jun¹, LIAO Jing-yu¹, ZHANG Ying-shan², WU Ya-zhen¹

(1.School of Mathematics and Statistics, Xuchang University, Henan Xuchang 461000, China)

(2.School of Finance and Statistics, East China Normal University, Shanghai 200241, China)

Abstract: In this paper, we study the complete-internal balanced in philosophy of related concepts and mathematical properties. By theory of multilateral matrices, the completely mathematical determination conditions for the complete-internal balanced block design is proved, and the method of computer validation is given. The balance of orthogonal table are extended.

Keywords: block design; complete-internal balanced; generalized orthogonal array; block factor; reproducibility

2010 MR Subject Classification: 62K10