1.4 试验设计的基本原则

王正明　易泰河

系统工程学院　军事建模与仿真系

2019 年 11 月 13 日

知识回顾

- 利用一架精度为 σ 的天平称 4 个不同的物体, 能否给出一种精度高于 σ 且只需称 4 次的称重方案?
- 查阅文献, 了解抽样与试验设计的联系和区别.
- 安装 R 和 RStudio, 并利用 “swirl” 包自学 R.
- 请查阅 Fisher, Yates, Box, Taguchi 中某一位的生平并梳理其在试验设计领域的主要贡献.
- 访问宾夕法利亚大学艾伯利理学院试验设计课程网站:

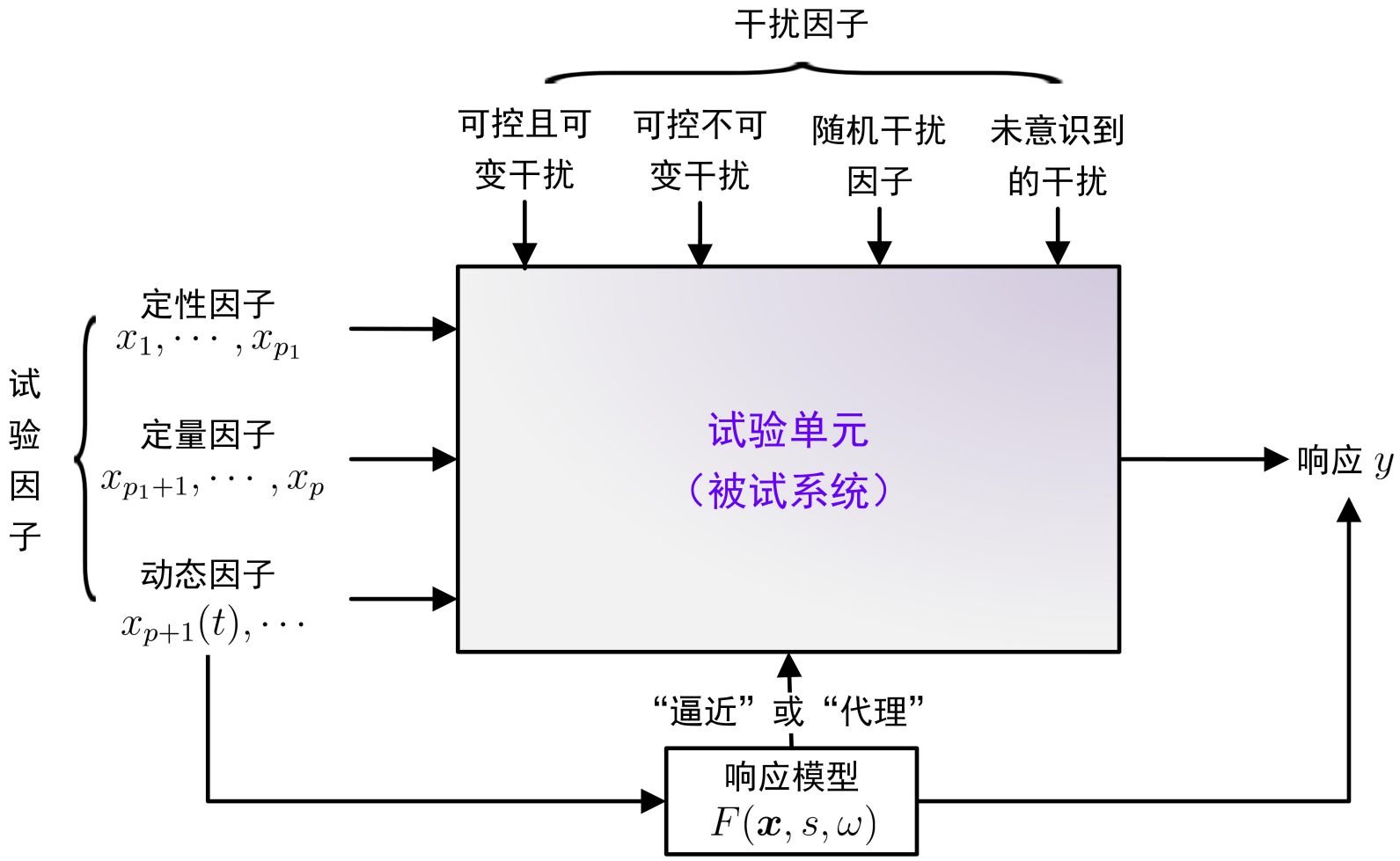
  `https://newonlinecourses.science.psu.edu/stat503/node/1/`

知识回顾

- 推荐使用开源的、免费的 R 语言及其编译器 RStudio.
  - 简单易学, 添加包 “swirl” 一步步地指导初学者学习;
  - 统计学和机器学习领域的研究者在发表论文时, 经常会同时发布一个相应的 R 添加包;
  - 能够绘制很漂亮的统计图形, 对学术研究非常有益;
  - 本课程中的几乎所有方法都能找到相应的添加包.

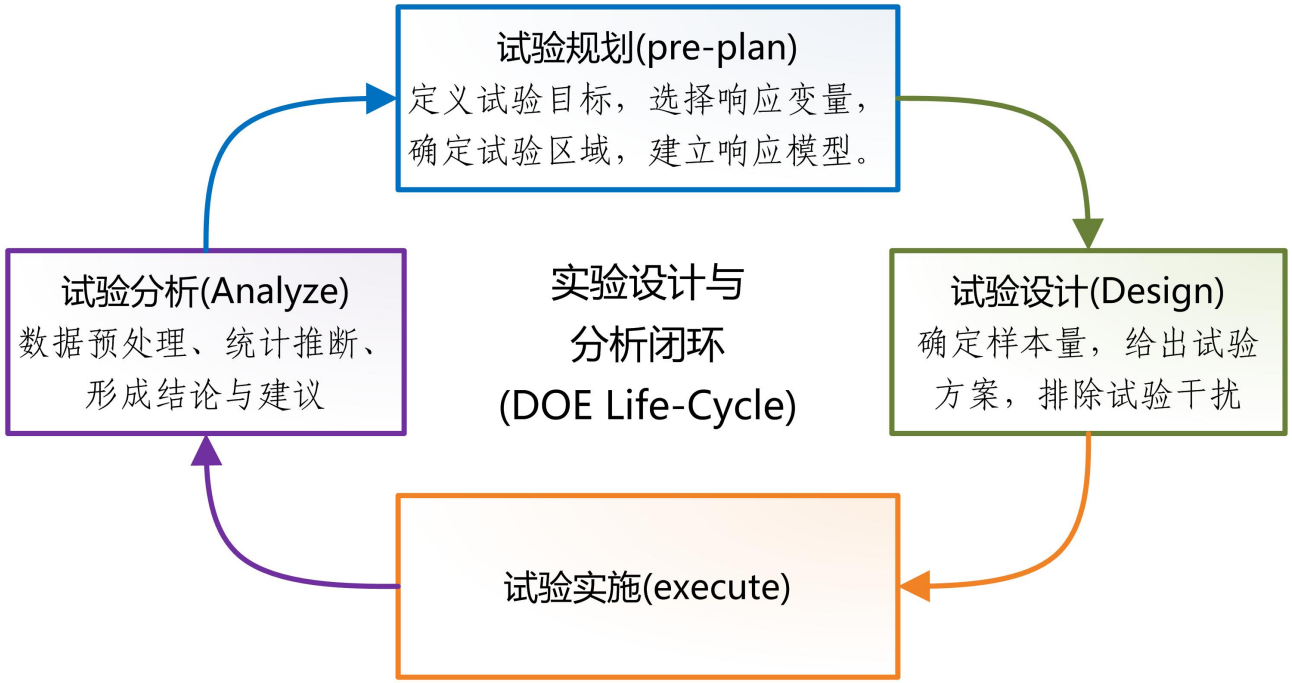
知识回顾

- 试验的要素



- 误差控制技术: 重复、随机化、区组.

知识回顾



试验设计:

(1) 确定试验次数, 资源消耗与精度构成一对均衡关系.

(2) 确定设计的准则, 并依据该准则从 $\mathcal{X}$ 中确定试验方案 $\xi_n = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$.

(3) 控制误差, 降低噪声因子的影响.

1.4 试验设计的基本原则

1.4.1 重复

1.4.2 随机化

1.4.3 区组

重复: 处理 $\boldsymbol{x}_i$ 处重复试验 m 次:

$$y_{ij} = \mu(\boldsymbol{x}_i) + \varepsilon_j, \quad \varepsilon_j \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \quad j = 1, \cdots, m.$$

1. 提供 $\mu(\boldsymbol{x}_i)$ 的更精确的估计:

$$\hat{\mu}(\boldsymbol{x}_i) = \bar{y}_{i\cdot} := \frac{1}{m}\sum_{j=1}^{m} y_{ij} \sim N(\mu(\boldsymbol{x}_i), \sigma^2/m).$$

2. 提供方差 σ^2 的估计:

$$\hat{\sigma}^2 = \frac{1}{m}\sum_{j=1}^{m} (y_{ij} - \bar{y})^2 .$$

3. 使得比较不同的处理成为可能.

重复: 处理 $\boldsymbol{x}_i$ 处重复试验 m 次:

$$y_{ij} = \mu(\boldsymbol{x}_i) + \varepsilon_j, \quad \varepsilon_j \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \quad j = 1, \cdots, m.$$

1. 提供 $\mu(\boldsymbol{x}_i)$ 的更精确的估计:

$$\hat{\mu}(\boldsymbol{x}_i) = \bar{y}_{i\cdot} := \frac{1}{m}\sum_{j=1}^{m} y_{ij} \sim N(\mu(\boldsymbol{x}_i), \sigma^2/m).$$

2. 提供方差 σ^2 的估计:

$$\hat{\sigma}^2 = \frac{1}{m}\sum_{j=1}^{m} (y_{ij} - \bar{y})^2 .$$

3. 使得比较不同的处理成为可能.

1.4 试验设计的基本原则

1.4.1 重复

1.4.2 随机化

1.4.3 区组

- 随机化(randomization): 试验单元的分配和试验的次序随机确定.
  1. 使各次试验结果互相独立, 便于采用各种统计方法分析试验数据;
  2. 消除或降低试验人员尚未意识到的噪声因子的影响;
  3. 抵消系统偏差, 提高对试验误差估计的准确度;
  4. 使得归纳因果关系成为可能.

- 设考察某个 3 水平试验因子, 每个水平重复 4 次,

```
set.seed(7638)
f <- factor(rep( c("a", "b", "c"), each = 4))
fac <- sample(f, 12)
eu <- 1:12
plan <- data.frame( runs = eu, levels = fac)
write.csv(plan, file = "Plan.csv", row.names =
    FALSE)
```

- sample() 的使用方式为

```
sample(x, size, replace = FALSE, prob = NULL)
```

- getwd() 找到当前工作目录;
- setwd() 修改工作目录.

例 (绿茶叶酸测定试验)

考察不同产地绿茶的叶酸含量是否有显著差异, 试验因子是绿茶的产地, 选四个不同的产地 A_1, A_2, A_3, A_4 构成四个水平. 重复数相等的设计称为平衡设计, 重复数不等的设计称为不平衡设计:

因子 A 的水平	试验编号						
A_1	1	2	3	4	5	6	7
A_2	8	9	10	11	12		
A_3	13	14	15	16	17	18	
A_4	19	20	21	22	23	24	

如果试验按次序一天内就完成, 若测得产地 A_4 的叶酸含量较低, 这可能是产地的原因, 也可能是由于试验员劳累引起的. 在 1 到 24 个试验号中一个接一个的随机抽取, 抽取所得的序列就是实际进行试验的次序. 这样安排的单因子试验在试验设计中称为不平衡完全随机设计.

1.4 试验设计的基本原则

1.4.1 重复

1.4.2 随机化

1.4.3 区组

- 如何消除试验中已知且可控因子的影响?
  - 有的因子可调整在固定的水平上
  - 试验单元之间的差异可能是造成试验结果波动的重要因子之一.
- 把“相近”的一组试验单元放在一起, 称为一个区组, 区组之间允许有较大的差异.
  1. 完全随机区组设计
  2. 拉丁方设计
  3. 平衡不完全区组设计
- 区组不是试验因子

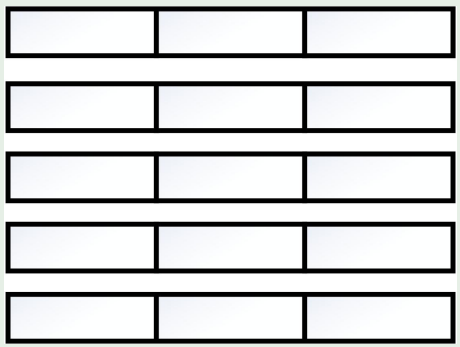
- 如何消除试验中已知且可控因子的影响?
  - 有的因子可调整在固定的水平上
  - 试验单元之间的差异可能是造成试验结果波动的重要因子之一.
- 把“相近”的一组试验单元放在一起, 称为一个区组, 区组之间允许有较大的差异.
  1. 完全随机区组设计
  2. 拉丁方设计
  3. 平衡不完全区组设计
- 区组不是试验因子

- 如何消除试验中已知且可控因子的影响?
  - 有的因子可调整在固定的水平上
  - 试验单元之间的差异可能是造成试验结果波动的重要因子之一.
- 把“相近”的一组试验单元放在一起, 称为一个区组, 区组之间允许有较大的差异.
  1. 完全随机区组设计
  2. 拉丁方设计
  3. 平衡不完全区组设计
- 区组不是试验因子

(1) 完全随机区组设计 (RCBD)

例 (完全随机区组设计)

比较 3 个品种水稻的产量是否存在区别. 现取 5 个不同土壤条件的地区, 每个地区各选 3 块面积和形状都非常接近的试验田. 若每块试验田安排一个品种, 如何安排试验?

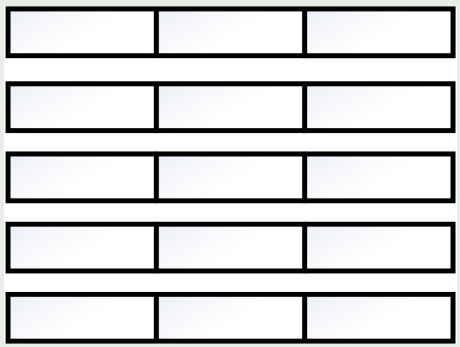


把每个地区的三个试验单元作为一个区组, 每个区组内三个品种各占一块地. 区组内哪个品种占哪块地随机分配. RCBD 要求试验次数是因子水平数的整数倍.

(1) 完全随机区组设计 (RCBD)

例 (完全随机区组设计)

比较 3 个品种水稻的产量是否存在区别. 现取 5 个不同土壤条件的地区, 每个地区各选 3 块面积和形状都非常接近的试验田. 若每块试验田安排一个品种, 如何安排试验?



把每个地区的三个试验单元作为一个区组, 每个区组内三个品种各占一块地. 区组内哪个品种占哪块地随机分配. RCBD 要求试验次数是因子水平数的整数倍.

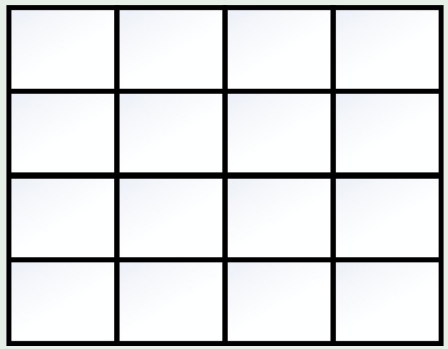
(1) 完全随机区组设计 (RCBD)

```
install.packages("agricolae")
library(agricolae)
treat <- c(1,2,3)
outdesign <- design.rcbd(treat, 5, seed = 11)
rcb <- outdesign$book
levels(rcb$block) <- c("block1", "block2", "block3",
    "block4", "block5")
```

(2) 拉丁方设计

例 (双向区组设计)

有 4 个玉米品种, 在一块长方形的试验田上进行试验, 将其按横向和竖向各 4 等分, 共分为 16 个长方块, 每个品种占 4 块. 土壤肥沃程度和其他条件沿横竖两个方向都有差异, 如何安排试验?

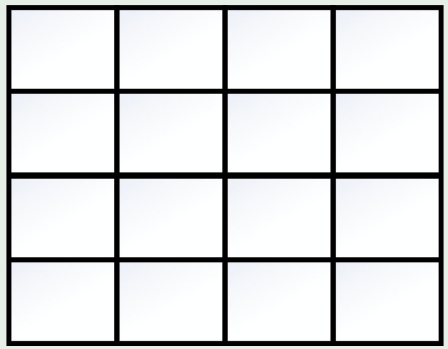


双向区组设计: 每一横向的 4 块试验田中每个品种都占一块, 且每一竖向的 4 块试验田中每个品种也都占一块.

(2) 拉丁方设计

例 (双向区组设计)

有 4 个玉米品种, 在一块长方形的试验田上进行试验, 将其按横向和竖向各 4 等分, 共分为 16 个长方块, 每个品种占 4 块. 土壤肥沃程度和其他条件沿横竖两个方向都有差异, 如何安排试验?



双向区组设计: 每一横向的 4 块试验田中每个品种都占一块, 且每一竖向的 4 块试验田中每个品种也都占一块.

(2) 拉丁方设计

- 一个 n 阶拉丁方设计为一个由 n 个拉丁字母 $n \times n$ 的方阵, 每个字母在每行只出现一次, 在每列也只出现一次.
- 可借助包 `agricolae` 中的函数 `design.lsd()` 来实现.
- n 阶拉丁方设计唯一吗?

1	2	3	4	4	2	1	3
2	1	4	3	2	1	3	4
3	4	1	2	1	3	4	2
4	3	2	1	3	4	2	1

(2) 拉丁方设计

- 一个 n 阶拉丁方设计为一个由 n 个拉丁字母 $n \times n$ 的方阵, 每个字母在每行只出现一次, 在每列也只出现一次.
- 可借助包 `agricolae` 中的函数 `design.lsd()` 来实现.
- n 阶拉丁方设计唯一吗?

1	2	3	4	4	2	1	3
2	1	4	3	2	1	3	4
3	4	1	2	1	3	4	2
4	3	2	1	3	4	2	1

(2) 拉丁方设计

定义

称 n 阶拉丁方称为左循环拉丁方, 若其第 $i+1$ 行 $\boldsymbol{x}_{i+1}$ 可由第 i 行 $\boldsymbol{x}_i$ 通过左移算子 L 得到, 即

$$\boldsymbol{x}_{i+1} = L\boldsymbol{x}_i, \quad i = 1, \cdots, n-1,$$

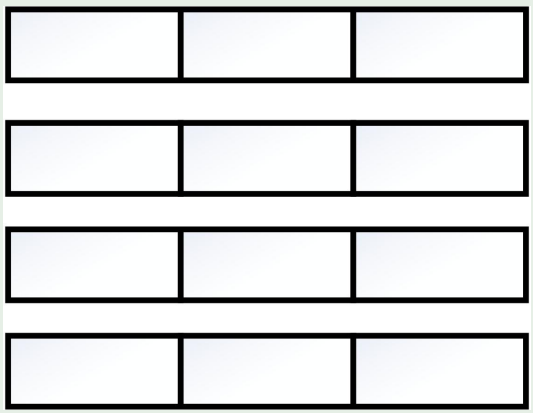
$$L(a_1, a_2, \cdots, a_n) = (a_2, a_3, \cdots, a_n, a_1).$$

类似地可以定义右循环拉丁方.

(3) 平衡不完全区组设计 (BIBD)

例 (平衡不完全区组设计)

比较 4 个水稻品种 A_1, A_2, A_3, A_4 的产量. 取 4 个不同土壤条件的地区 B_1, B_2, B_3, B_4 作为 4 个区组, 各选 3 块面积和水土条件都非常接近的试验田. 如何安排试验?



(3) 平衡不完全区组设计 (BIBD)

水稻品种	地区			
	B_1	B_2	B_3	B_4
A_1	1	1	1	–
A_2	1	1	–	1
A_3	1	–	1	1
A_4	–	1	1	1

- 随机性: 区组内部随机化;
- 平衡性: 每个区组中都含 3 个水平, 每个水平都在 3 个区组中出现, 任一对水平在同一区组内同时出现的次数都是 2;
- 区组内不完全性.

(3) 平衡不完全区组设计 (BIBD)

- BIBD 中有 5 个参数 (q, t, b, r, λ):
  - q 为水平数,
  - t 为每区组所含的试验单元数,
  - b 为区组数目,
  - r 为每个水平的试验次数,
  - λ 为任一对水平在同一区组内出现的次数.

- BIBD 存在的必要条件:

$$bt = qr, \quad \lambda(q-1) = r(t-1), \quad b \geq q.$$

- 添加包 `daewr` 中的函数 `BIBsize()` 可以用于计算满足条件的 λ 和 r.

总结

1. 随机噪声因子采用 _____ 的方法来处理;
2. 未知的噪声因子的影响采用 _____ 的方法处理;
3. 已知且可以改变的噪声因子的影响采用 _____ 的方式来处理;
4. 已知但不能任意改变的噪声因子采用 _____ 的技术来处理;
5. 三类区组设计 _____, _____, _____.

习题

1. 给出一个五阶右循环拉丁方设计.
2. 根据随机误差方差的估计, 分析样本量与精度之间的关系.