

• 讲座 •

如何正确进行生物医学科研设计

V. 如何合理选择 4 种多因素实验设计 (2)

李长平, 胡良平, 高辉, 金松华, 周诗国, 郭晋

编者按

生物医学科研的研究对象是生物体, 而生物体具有极大的变异性。科研的目的就是要从表面上看似杂乱无章的事物或现象中找出规律性的东西来, 以便正确地解释和揭示事物或现象的变化、发展乃至消亡的规律, 从而达到认识人类社会并促进其发展、认识自然、改造自然以至于征服自然之目的。人类要完成此重任, 仅靠热情和苦干是不够的, 必需加上巧干。所谓巧干, 就是要充分利用人类已经积累起来的宝贵经验, 加上人类特有的创造性思维和勇于实践的精神, 在确立了可行的奋斗目标之后, 对未知事物或现象进行调查或试验。要想使调查或试验得出的结论真实可信, 同时花费的人力、物力、财力和时间又比较少, 就需要有科学完善严谨的科研设计方案的指导。本刊特邀军事医学科学院生物医学统计学咨询中心主任胡良平教授, 以“如何正确进行生物医学科研设计”为题, 撰写系列统计学讲座。希望该系列讲座能对生物医学科研工作者制定科研设计方案有所帮助。

本刊上期刊登了“如何合理选择 4 种多因素实验设计 (1)”, 这里, 接着介绍该文的第二部分, 即“如何合理选择 4 种多因素实验设计 (2)”。

4 正交与均匀设计实施的步骤及注意事项

4.1 正交设计与均匀设计实施的步骤

第一步, 需结合专业知识、文献资料和必要的预实验提供的信息, 确定需要重点考察的实验因素及交互作用项, 确定因素的变化范围以及因素的水平数。

第二步, 根据问题实际, 选择合适的正交表(或均匀表), 并进行表头设计, 将正交表(或均匀表)中的“编码水平”换成实际实验中的“真实水平”, 从而编制出正交设计(或均匀设计)的实验方案。

第三步, 最好采用随机化的方法确定实验的次序, 进行实验并记录相关观测指标的结果。

第四步, 对数据进行方差分析(若是均匀设计需对数据采用多重线性回归分析), 也可进一步探求最优实验组合条件。

4.2 注意事项

4.2.1 正交设计应用中需注意的问题 正交设计应用时要注意合理进行表头设计。所谓表头设计, 是指根据分析的需要, 选用合适的正交表, 将各个因素以及需要考察的交互作用项安排在正交表各列上的过程。在此过程中, 需要把握以下两条设计原则: 一是根据因素个数、水平数合理安排主效应和因素之间的交互作用。正交表选定以后, 即可进行表头设计, 应遵照不混杂的原则, 即不同的因素(包括交互作用)不能占用相同的列, 在做表头设计时, 为了避免混杂, 应先安排有交互作用的因素; 二是设计时须考虑对正交实验数据作方差分析时误差的计算途径——从空白列获得或由重复实验提供的数据中获得^[1]。确定表头设计即确定实验点以后,

还要决定是否要做重复实验。如果分析实验结果时要作方差分析并且表头设计中已无空白列, 尤其是各实验条件下实验结果很不稳定时, 则需要做重复实验, 否则不必做。

【例 1】某研究者欲进行微生物培养基成分优化实验。目的是对响应变量(在本例中是产量)有影响的各因素及各因素间的交互作用的重要性作出分析, 并找出优化的培养基成分。在该研究中需考虑 A、B、C、D、E 共 5 个 3 水平因素, 研究者根据以往经验, 还需考察 A×C、A×B、A×E 的交互作用。请选用合适的设计类型, 并进行具体的设计。

解析: 本实验涉及 5 个 3 水平的实验因素, 还需考察 3 对一级交互作用。若进行析因设计, 则至少需要做 $3^5 \times 2 = 486$ 次实验, 次数太多, 不合适。那么, 考虑选用正交设计。用正交表设计实验时, 表的行数恰好是实验点的个数。若不作重复实验, 总的自由度正好是表的行数减 1 ($f_R = \text{行数} - 1$)。为了保证表头设计满足不混杂的原则, 各因子与交互作用的自由度之和 ($f_{\text{因}}$) 必须小于等于表的自由度即 $\sum f_{\text{因}} \leq f_R$ 。因此采用 $L_{27}(3^{13})$ 正交表安排此实验, 是可行的。因为此表共有 13 列, 5 个因素占用 5 列, 3 对一级交互作用占用 6 列, 还剩 2 个空列用于估计实验误差。将需考察交互作用的因素先安排, 其他因素可在剩下列上随意放, 其表头设计见表 1。

说明: 根据此表的交互作用表可知, 第 1、2 列上的交互作用出现在第 3、4 列上; 第 1、5 列上的交互作用出现在第 6、7 列上; 第 1、9 列上的交互作用出现在第 8、10 列

作者单位: 100850 北京, 军事医学科学院生物医学统计学咨询中心(李长平、胡良平、高辉、周诗国、郭晋); 100850 北京华素制药股份有限公司(金松华)

通讯作者: 胡良平, Email: lphu812@sina.com

表 1 用 $L_{27}(3^3)$ 正交表安排此实验的表头设计

实验 列号	1	2	3	4	5	6	7	8	9	10	11	12	13
编号 因素	A	B	AB ¹	AB ²	C	AC ¹	AC ²	AE ¹	E	AE ²		D	

上。其中 AB¹ 和 AB² 分别代表 A×B 的第一部分和第二部分，其他类似符号含义相同。

表头设计完成后，将完整的 $L_{27}(3^3)$ 正交表中的第 1、2、5、9、12 列摘录出来，它们是仅含单个因素的列，将表中上述 5 列上的“代码水平”转换成因素的“真实水平”，表中每一行是一种特定的实验条件，在每个实验条件下做无重复的（实验结果变异度很小时）实验或有 k 次独立重复的实验（ $k \geq 2$ ）。

【例 2】为优化提取板蓝根的工艺条件，取 20 g 板蓝根饮片，采用正交设计实验方法筛选板蓝根中有效成分总有机酸的最佳提取工艺，考察因素有提取溶剂浓度、溶剂用量、提取次数、提取时间 4 个因素，因素及其水平数见表 2。选用 $L_9(3^4)$ 正交表安排实验，具体结果见表 3。方差分析结果见表 4。研究者根据正交实验结果中各因素的极差大小判定各因素对观测指标影响的重要程度为 $A > C > D > B$ ，并得出提取有效成分总有机酸的最佳工艺为 $A_2B_2C_3D_3$ ，即

表 2 总有机酸提取工艺影响因素水平

水平	因素：	水平取值			
		乙醇浓度 (A, %)	溶剂用量 (B, 倍)	提取次数 (C, 次)	提取时间 (D, h)
1		50	5	1	0.5
2		70	7	2	1.0
3		90	10	3	1.5

表 3 总有机酸提取正交实验结果

实验号	因素：	水平代码				总有机酸含量 (%)
		A	B	C	D	
1		1	1	1	1	2.82
2		1	2	2	2	4.62
3		1	3	3	3	5.22
4		2	1	2	3	4.67
5		2	2	3	1	4.84
6		2	3	1	2	3.81
7		3	1	3	2	3.38
8		3	2	1	3	2.21
9		3	3	2	1	2.65
K_1		12.66	10.88	8.84	10.31	
K_2		13.32	11.77	11.94	10.81	
K_3		8.04	11.68	13.44	12.10	
R		5.28	0.89	4.60	1.79	

表 4 提取总有机酸含量方差分析表

方差来源	离均差平方和	自由度	均方	F	P
A	4.00	2	2.0000	5.71	> 0.10
B	1.00	2	0.5000	1.43	> 0.10
C	3.67	2	1.8350	5.24	> 0.10
D (误差)	0.70	2	0.3500		

工艺参数为：乙醇浓度 70%，用量为药材 7 倍量，提取 3 次，每次 1.5 h。请问：这样安排正交实验是否最合适？

分析：研究者考察了 4 个 3 水平的实验因素对总有机酸含量的影响，选用 $L_9(3^4)$ 正交表进行表头设计并安排实验，这样 4 个实验因素就将 $L_9(3^4)$ 正交表的 4 列占满，没有空列了，且研究者在每个实验条件下也没有做独立重复实验，这违反了上述表头设计的第二条原则，将导致误差项无法估计。研究者以均方最小的 D 因素的效应作为误差项来检验其它 3 个因素的效应不是最合适的，这样就无法估计 D 因素对观测指标的影响了。

此外，若没有专业知识、预实验或相关参考文献为依据，可认为各因素之间的全部或部分交互作用可以忽略不计的话，在实验设计时还是要考虑可能存在的交互作用。这一点在寻找最佳实验组合条件时尤为重要。以本例来说，若 A、B、C、D 这 4 个因素间不存在交互作用，可以采用传统的“比一比、看一看”的直接法来确定最优提取条件；但若这 4 个因素间的部分交互作用有统计学意义的话，则不能仅凭极值大小和因素各水平的合计数大小来确定最优实验条件了，此时可从线性统计模型出发比较几种较重要的水平组合条件下平均强度的估计值^[1]，也可采用遗传算法来进行最优实验条件的选择^[2]，或者根据考察交互作用的多少选取较大的正交表来进行该实验，如选择 $L_{27}(3^3)$ 正交表。

值得注意的是，如果没有预实验或明确的专业知识为依据，可认为相同实验条件下观测指标的取值离散度很小时，在应用正交设计和均匀设计时都应该在挑选的各种实验条件下进行 2 次以上的独立重复实验，以减少生物个体变异对观测指标的影响。

4.2.2 均匀设计应用中需注意的问题

(1)第一，均匀设计更适合于筛选因素，最好再采用正交设计或析因设计得出更稳定的结论。

均匀设计因其所需实验次数少，可节省大量的科研时间、科研经费而广受科研工作者的欢迎和青睐，尤其是在药物处方、生产工艺等领域应用甚广。但由于均匀设计定量资料分析时的回归方程不唯一，导致了实验结论可能因人而异，不够稳定。所以，最好是采用均匀设计进行实验因素的筛选，再利用正交设计或析因设计对这些筛选出来的实验因素做进一步的确证性实验，从而达到预期的研究目的。

(2)第二，分析时应结合专业知识和统计学知识，选择合适的回归模型，避免得出与专业不符的结论。

在均匀实验设计中，要采用多重回归分析方法进行统计分析。对于特定的实验资料，若自变量的个数以及对哪些变量采用何种变量变换都事先确定的条件下，“最优”回归方程应该是产生于在专业知识和统计学知识两方面综合考虑的前提下得出的一个最合适的回归方程。具体地说，它应符合以下标准^[3]：其一，整个回归方程以及全部回归参数具有统计学意义；其二，各回归系数的正、负号与其所对应的自变量在专业上的含义是吻合的；其三，在自变量一切可能的取值区间（以实验设计时确定的区域为基准）内，因变量的

预测值符合专业知识的要求；其四，对“较优”的若干个回归方程进行残差分析，尽可能取残差图上散点分布很理想（散点随机分布、无异常点、无明显的变化趋势）、总残差平方和最小且回归方程中所包含的自变量个数最少的回归方程。当然，由于全部需要考察的自变量的个数难以确定，构建“最优”回归方程只能作为一个理论上追求的目标，而更实事求是的提法或许是构建“较优”回归方程，即在一定的范围内或条件下找到一个最好的变量组合。如果因素的最佳水平组合不在实验方案中，适当的追加实验是必要的^[4]。

(3)第三，用重复实验的均数取代实验结果的做法不妥。

用重复实验的均数取代实验结果的做法不妥，它无法做“拟合不足”或称“失拟”的检验；另外，在做变量筛选中，当“自变量总数超过实验点数”时，不应采用经典的逐步回归选变量法。这虽然是国内外很流行的方法，但不妥^[5]。比较正确的做法是：尽可能充分利用各次重复实验的数据，并进行“失拟检验”；结合专业知识和预实验的经验，尽可能引入对观测结果可能有影响的平方项和交叉乘积项；尽可能多采用几种筛选自变量的方法（如前进法、后退法、逐步回归法、最优回归子集法）筛选变量，并利用上面介绍的评价标准综合评价，从而获得在专业和统计学两方面考察都比较理想的回归分析结果。

【例 3】刺五加是我国北方地区特产常用药材之一，它的各个部位均含有黄酮类化合物，其中叶中含量最高。黄酮类化合物具有抗自由基、抗氧化等多种生理活性及药理作用。现欲研究高压提取刺五加叶中总黄酮的最优工艺，需考虑因素与水平如表 5 所示。

表 5 本例中的因素及其具体水平

水平	A: 提取压力 (Mpa)	B: 乙醇浓度 (%)	C: 溶剂用量 (ml/g)
1	350	35	35
2	400	40	40
3	450	45	45
4	500	50	50
5	550	55	55

现欲通过实验找出影响总黄酮提取率(%)的全部因素的最佳水平组合及因素的主次顺序，通过预实验发现，因素之间的交互作用可以忽略不计，现希望能最大限度地减少实验次数，请安排合适的实验设计方案。

分析：该实验涉及到 3 个因素，每个因素均有 5 个水

平。若采用析因设计，其实验次数至少为 $5^3 \times 2 = 250$ 次，太多了。若选用正交设计，至少要做 25 次（供选择的正交表为 $L_{25}(5^6)$ ）。现要求最大限度的减少实验次数，可选择均匀设计。选用 $U_5(5^4)$ 安排实验，将 3 个因素分别放在 1、2、4 列上（注意均匀设计表表头的安排要根据均匀设计表的使用表来安排实验因素），实验方案如表 6 所示。

表 6 用 $U_5(5^4)$ 均匀表安排 3 个实验因素

实验号	水平代码及其取值			
	列号及因素:	1 (A)	2 (B)	3 (C)
1		2 (350)	2 (40)	4 (50)
2		3 (400)	4 (50)	3 (45)
3		3 (450)	1 (35)	2 (40)
4		4 (500)	3 (45)	1 (35)
5		5 (550)	5 (55)	5 (55)

如果条件允许，也可选用 $U_{10}(10^{10})$ 表安排实验，将 3 个因素分别放在 1、5、7 列上，将其中的 6~10 水平分别依次按 1~5 水平使用，具体实验方案从略。当然，也可直接选用水平数为 5 但行数为 5 的 2 倍或 3 倍的均匀设计表。

参考文献

[1] Wang RA. Design and analysis of medical experiments. Beijing: Peking Univ Med Press. 2000, 47-60. (in Chinese)
王仁安. 医学实验设计与统计分析. 北京: 北京医科大学出版社. 2000, 47-60.

[2] Qiu LX, Liu GF, He DW, et al. Application of genetic algorithm in selecting the best experimental condition for orthogonal experimental design. Mod Preve Med, 2004, 31(4):493-495. (in Chinese)
仇丽霞, 刘桂芬, 何大卫, 等. 遗传算法在正交试验设计最优条件选择中的应用. 现代预防医学, 2004, 31(4):493-495.

[3] Hu LP, Li ZJ, Liu HG, et al. Creating best regression subsets for the experimental data of the uniform design. Chin J Hosp Statistics, 2004, 11(4):315-318. (in Chinese)
胡良平, 李子建, 刘惠刚, 等. 均匀设计实验资料最优回归子集的构建. 中国医院统计, 2004, 11(4):315-318.

[4] Guo DX, Qiu LX, Zhang MD. Uniform design method and it's application. J Mathematical Med, 2005, 18(1):69-71. (in Chinese)
郭东星, 仇丽霞, 张满栋. 均匀设计方法及其应用. 数理医学杂志, 2005, 18(1):69-71.

[5] Sun SG. Statistical analysis with repeated trials in uniform design. Appl Statistics Manag, 2000, 19(2):24-29. (in Chinese)
孙尚拱. 均匀设计中有重复试验的统计分析. 数理统计与管理, 2000, 19(2):24-29.