·科研方法专题·

如何正确运用方差分析——正交设计定量资料一元方差分析与SAS实现

胡纯严[1]，胡良平[1,2*]

（1. 军事科学院研究生院，北京 100850；
2. 世界中医药学会联合会临床科研统计学专业委员会，北京 100029
*通信作者：胡良平，E-mail：lphu927@163.com）

【摘要】 本文目的是介绍正交设计及其定量资料的方差分析和SAS实现。从自由度角度来划分，正交设计可分为饱和正交设计与非饱和正交设计两类；从因素的水平数角度来划分，正交设计又可分为同水平正交设计与混合水平正交设计两类；从规范化角度来划分，正交设计还可分为标准正交设计与非标准正交设计两类。对来自标准正交设计的定量资料，可采用常规方法进行方差分析；而对来自非标准正交设计的定量资料，需要对方差分析方法做必要的改进。本文基于3个实例，借助SAS软件实现了无重复试验和有重复试验标准正交设计定量资料方差分析。

【关键词】 正交设计；试验点；最优水平组合；方差分析；*F*分布
中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20220510004

How to use analysis of variance correctly——an analysis of variance for the univariate quantitative data collected from the orthogonal design

Hu Chunyan[1], *Hu Liangping*[1,2*]

（1. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*
Corresponding author: Hu Liangping, E-mail: lphu927@163.com）

【Abstract】 The purpose of this paper was to introduce the orthogonal design and its quantitative data analysis of variance and the SAS implementation. From the perspective of degrees of freedom, the orthogonal design could be divided into the saturated orthogonal design and the unsaturated orthogonal design. From the perspective of the number of factor levels, the orthogonal design could be divided into the same level orthogonal design and the mixed level orthogonal design. From the perspective of normalization, the orthogonal design could also be divided into the standard orthogonal design and the non-standard orthogonal design. Quantitative data from the standard orthogonal designs could be analyzed by the conventional methods, while quantitative data from the non-standard orthogonal designs needed to be improved. Based on three examples, this paper realized the quantitative data analysis of variance with the standard orthogonal design without repeated experiments and with repeated experiments by means of the SAS software.

【Keywords】 Orthogonal design; Experimental point; Optimal level combination; Analysis of variance; *F* distribution

正交设计是安排多因素的一种高效方法，在满足某些特定条件时，正确地运用正交设计，不仅可以获得比较精准的试验结果，而且所需的样本含量较少。本文将介绍正交设计的基本概念、设计方法、计算公式以及基于SAS软件实现正交设计定量资料方差分析的方法。

1 基本概念

1.1 正交

在几何学上，正交的意思是两条直线互相垂直。在代数学上，若两个向量的内积（或称数量积）等于0，就定义这两个向量是互相正交的[1]。假设有两个向量如下：a=(-1,-1,-1,-1,1,1,1,1)，b=(-1,1,-1,1,-1,1,-1,1)。将它们对应位置上的数字相乘后求和，可记为(a,b)=-1×(-1)+(-1)×1+……+1×1=0，它就是向量a与向量b的“内积”。于是，在数学上，称上面两串数字排成的序列或向量是互相正交的。

在统计学上，统计学家构造出一系列类似于上面给出的“向量”，将它们排列在一张表格中，表格

中任何两列数据的“内积”都等于0，这种表格就叫做正交表[2]。以$L_8(2^7)$正交表为例，其中“L”代表正交表，数字“8”代表此表有8行，数字“2”代表每列有2个不同的水平，数字“7”代表此表有7列。$L_8(2^7)$正交表见表1。表1共有7列，每列可以代表一个试验因素的2个水平各重复出现4次的一种排列；实际使用时，每列可以安排一个二水平因素，每行代表全部因素的一种水平组合结果，例如，第1行所有因素都取“1水平”。若用“-1”取代表中的“2”，则表1中的任何两列的“内积”都等于0。

表1 $L_8(2^7)$正交表

Table 1 $L_8(2^7)$ Orthogonal table

试验号 列号：	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

1.2 正交设计

正交设计是利用一系列规格化的正交表来安排各试验因素及其水平组合的方法。正交表是经过严格的数学推导编制而成，正交表中的每一行代表全部试验因素的一种水平组合，称为一个试验点；正交表中的每一列代表一个试验因素（或交互作用项）及其水平的排列和重复出现的情况，将全部因素及其拟考察的交互作用项恰当地安排在正交表的表头上，就叫做正交设计。

1.3 正交设计的特点

正交设计有4个突出特点：①由正交表挑选出来的试验点在空间上具有“均匀分散性”；②由正交表挑选出来的试验点在统计分析时具有“整齐可比性”；③某些未包括在正交表中的试验点，可以通过统计分析将其发现；④在某些假设（例如某些因素之间的交互作用不存在或可以忽略不计，特定试验条件下试验结果的稳定性很好）成立的条件下，正交设计比其他多因素设计（排除均匀设计）所需的样本含量更少。

1.4 正交设计误差处理

在采用正交设计安排试验时，应预先考虑好如何估计试验误差：在正交表中留有空白列（可用于估计不同试验条件之间所产生的试验误差，被称为第一类误差），或者在挑选出来的试验点上进行m次独立重复试验（$m\geq2$，可用于估计相同试验条件下不同重复试验次数之间所产生的试验误差，被称为第二类误差）。一般来说，第一类误差大于第二类误差。在对资料进行方差分析时，若只有第一类误差，就以其均方为分母，构造F检验统计量；若有两类误差，就需要将它们的离均差平方和、自由度分别进行合并，用合并后的离均差平方和除以合并后的自由度，构造出误差的均方。误差项的自由度越大（意味着所需要的样本含量越大），所得结论的可信度就越高。

1.5 正交表的种类

正交表可分为两大类：第一类是饱和正交表，第二类是非饱和正交表，也称不完备正交表[3]。在不做重复试验的饱和正交表中，总自由度=各列自由度之和=表的行数－1，各列自由度=各列中不同水平数－1。总自由度不等于各列自由度之和的正交表，就是非饱和正交表。例如，在$L_8(2^7)$正交表中，$df_{总}$=8－1=7，每一列的自由度=2－1=1，7列自由度之和=7，满足“总自由度=各列自由度之和”的要求，故$L_8(2^7)$正交表是饱和正交表。

根据正交表中各列的水平数是否相等，还可将正交表分为同水平正交表与混合水平正交表[4-5]。常见的同水平正交表有：二水平的正交表[$L_4(2^3)$、$L_8(2^7)$、$L_{16}(2^{15})$、$L_{32}(2^{31})$、$L_{64}(2^{63})$]，三水平的正交表[$L_9(3^4)$、$L_{27}(3^{13})$、$L_{81}(3^{40})$]，四水平的正交表[$L_{16}(4^5)$、$L_{64}(4^{21})$]，五水平的正交表[$L_{25}(5^6)$、$L_{125}(5^{31})$]。常见混合水平的正交表有：$L_8(4\times2^4)$、$L_{12}(3\times2^4)$、$L_{16}(4^2\times2^9)$、$L_{16}(4^3\times2^6)$、$L_{18}(2\times3^7)$、$L_{18}(6\times3^6)$、$L_{27}(9\times3^9)$、$L_{32}(4^5\times2^{16})$、$L_{36}(6\times3^{12})$、$L_{32}(16\times2^{16})$、$L_{24}(3\times4\times2^4)$、$L_{36}(6^2\times3^5\times2)$、$L_{32}(8\times4^6\times2^6)$。

有些混合水平正交表是非饱和正交表，例如，在$L_{12}(3\times2^4)$正交表中，总自由度=12－1=11，而各列自由度之和=(3－1)+4×(2－1)=6。

2 正交设计定量资料一元方差分析的计算公式

正交表的种类很多，且因素的个数、水平数以及需要考察的交互作用项数都会随着具体问题发生变化，还涉及是否进行重复试验。因此，无法给出一个统一的正交设计定量资料一元方差分析计

算公式。现假定在$L_8(4\times2^4)$的前3列上分别安排因素A(4水平)、因素B(2水平)和因素C(2水平),扼要介绍正交设计定量资料一元方差分析的计算思路。计算正交表中每一列上的离均差平方和的方法,与单因素两水平或多水平设计定量资料一元方差分析计算组间离均差平方和的方法相同,可参见文献[6]。

第一步,计算最后一列试验结果Y的总离均差平方和。

$$SS_T=\sum_{i=1}^{8}\left(Y_i-\bar{Y}\right)^2=\sum_{i=1}^{8}Y_i^2-\frac{1}{8}\left(\sum_{i=1}^{8}Y_i\right)^2 \quad (1)$$

第二步,计算A、B、C列试验结果Y的总离均差平方和。

$$SS_A=2\sum_{j=1}^{4}\left(\bar{Y}_j-\bar{Y}\right)^2 \quad (2)$$

$$SS_B=4\sum_{k=1}^{2}\left(\bar{Y}_k-\bar{Y}\right)^2 \quad (3)$$

$$SS_C=4\sum_{l=1}^{2}\left(\bar{Y}_l-\bar{Y}\right)^2 \quad (4)$$

第三步,计算误差E的离均差平方和。

$$SS_E=SS_T-(SS_A+SS_B+SS_C) \quad (5)$$

第四步,计算总自由度和因素A、B、C的自由度以及误差E的自由度。

$$df_T=8-1=7 \quad (6)$$

$$df_A=4-1=3 \quad (7)$$

$$df_B=df_C=2-1=1 \quad (8)$$

$$df_E=7-(3+1+1)=2 \quad (9)$$

第五步,计算误差的均方和因素A、B、C的均方。

$$MS_E=SS_E/df_E \quad (10)$$

$$MS_A=SS_A/df_A \quad (11)$$

$$MS_B=SS_B/df_B \quad (12)$$

$$MS_C=SS_C/df_C \quad (13)$$

第六步,计算与因素A、B、C对应的检验统计量F的值。

$$F_A=MS_A/MS_E \quad (14)$$

$$F_B=MS_B/MS_E \quad (15)$$

$$F_C=MS_C/MS_E \quad (16)$$

若基于手工计算,需根据计算检验统计量F值的分子和分母的自由度去查方差分析用的F临界值表,再比较检验统计量F值与其对应的F临界值的大小,做出统计推断。因篇幅所限,此处从略。本文将基于SAS软件实现方差分析[7]。

3　正交设计定量资料一元方差分析的实例与SAS实现

3.1　实例与数据结构

【例1】某化工厂为了提高产品的收率,根据具体情况和经验确定试验条件如下:因素A为反应温度(A_1=30℃,A_2=40℃);因素B为反应时间(B_1=60 min,B_2=90 min);因素C为两种原料比(C_1=1∶1,C_2=1.5∶1);因素D为搅拌速度(D_1=慢,D_2=快)。选用$L_8(2^7)$正交表安排上述4个因素,试验结果见表2[4]。试分析哪些因素对收率Y的影响有统计学意义。

表2　4个试验因素对产品收率影响的试验结果

Table 2　Experimental results of the influence of four experimental factors on the product yield

试验号	1 (A)	2 (B)	3	4 (C)	5	6	7 (D)	收率 (%)
1	1(30℃)	1(60 min)		1(1∶1)			1(慢)	75
2	1(30℃)	1(60 min)		2(1.5∶1)			2(快)	84
3	1(30℃)	2(90 min)		1(1∶1)			2(快)	81
4	1(30℃)	2(90 min)		2(1.5∶1)			1(慢)	83
5	2(40℃)	1(60 min)		1(1∶1)			2(快)	80
6	2(40℃)	1(60 min)		2(1.5∶1)			1(慢)	84
7	2(40℃)	2(90 min)		1(1∶1)			1(慢)	72
8	2(40℃)	2(90 min)		2(1.5∶1)			2(快)	77

注:表头上的"1~7"分别代表7列的编号

【例2】某电解腐蚀试验考察4个三水平因素(各因素的具体含义从略),交互作用均不考虑。试验指标为产品质量,按照规定标准进行综合评分。选用$L_9(3^4)$正交表安排试验,在每个试验条件下进行3次重复试验。试验安排和试验结果见表3[3]。试分析4个因素对试验结果的影响是否有统计学意义。

表3　4个三水平因素对电解腐蚀作用的正交设计及试验结果

Table 3　Orthogonal design and the experimental results of four three-level factors on the role of the electrolytic corrosion

试验号	A (1)	B (2)	C (3)	D (4)	$y_i=(X_i-70)/5$		
					y_1	y_2	y_3
1	1	1	1	1	−1	−2	0
2	1	2	2	2	0	−1	3
3	1	3	3	3	−1	0	2
4	2	1	2	3	−3	−2	2
5	2	2	3	1	−4	−5	0
6	2	3	1	2	−6	−6	−6
7	3	1	3	2	4	0	−1
8	3	2	1	3	3	3	2
9	3	3	2	1	−4	−1	−1

注:表头上的"(1)~(4)"分别代表正交表上的4个列号;各行上最后3个数据是重复试验的结果,原文为简化计算,对原始数据X进行了变量变换,用y_1、y_2、y_3表示

【例3】某农业科研所为了提高小麦产量，拟考察A、B、C三个因素对小麦亩产的影响。各因素各水平的含义如下：因素A为品种，有4个水平，1~4分别代表泰山4号、济6矮、矮秆早、139号；因素B为施肥量，有2个水平，1和2分别代表高肥和低肥；因素C为播种期，有2个水平，1和2分别代表早播和晚播。试验安排和试验结果见表4[5]。试分析三个因素对小麦亩产Y的影响是否有统计学意义？

表4 3个因素对小麦亩产影响的正交设计及试验结果

Table 4 Orthogonal design and the experimental results of the influence of three factors on the wheat yield per mu

试验号	A (1)	B (2)	C (3)	(4)	(5)	亩产(斤)
1	1	1	1	1	1	980
2	1	2	2	2	2	690
3	2	1	1	2	2	1 035
4	2	2	2	1	1	760
5	3	1	2	1	2	830
6	3	2	1	2	1	985
7	4	1	2	2	1	660
8	4	2	1	1	2	890

注：表头上的“(1)~(5)”分别代表5个列号

3.2 用SAS实现方差分析

3.2.1 对例1的分析与解答

【分析与解答】所需要的SAS程序如下：

```
data a1;
input A B C D y;
cards;
1 1 1 1 75
1 1 2 2 84
(详细数据见表2的后6行)
;
run;
proc anova data=a1;
class A B C D;
model y=A B C D;
/* model y=A B C A*B;
model y=C A*B;
model y=C; */
run;
```

【SAS程序说明】在SAS过程步中，有四个“model语句”，第一个是可执行的，后三个放在注释语句“/* */”中是不可执行的。每次运行SAS程序，只能让一个“model语句”处于可执行状态。

【SAS输出结果及解释】

由第一个“model语句”的输出结果可知，4个因素均无统计学意义，其中，因素D的作用最小。因篇幅所限，具体输出结果从略，下同。

由第二个“model语句”的输出结果可知，因素C（F=11.54，P=0.042 6）和交互作用项A*B（F=11.54，P=0.04 26）具有统计学意义，而因素A（F=2.88，P=0.188 0）和因素B（F=2.88，P=0.188 0）均无统计学意义。

由第三个“model语句”的输出结果可知，因素C（F=11.54，P=0.042 6）具有统计学意义，而交互作用项A*B（F=5.77，P=0.092 0）无统计学意义。

由第四个“model语句”的输出结果可知，当模型中仅有因素C时，它对定量观测结果的影响变得无统计学意义（F=3.41，P=0.114 4）。

由以上分析过程和结果可知：多因素方差分析结果是相对的，不是一成不变的。因为随着模型中待分析的项的改变（在本质上，是模型误差的均方在改变），各因素及交互作用项的统计显著性也会发生变化。

3.2.2 对例2的分析与解答

【分析与解答】所需要的SAS程序如下：

```
data a2;
input A B C D y1 y2 y3;
y=y1; output;
y=y2; output;
y=y3; output;
cards;
1 1 1 1 -1 -2 0
1 2 2 2 0 -1 3
(详细数据见表3的后7行)
;
run;
proc anova data=a2;
class A B C D;
model y=A B C D;
means A B C D;
run;
proc anova data=a2;
class A B D;
model y=A B D;
run;
```

【SAS输出结果及解释】

由第一个过程步的输出结果可知，因素C（F=0.54，P=0.591 0）对试验结果的影响无统计学意义，因素A（F=11.20，P=0.000 7）、因素B（F=5.17，P=0.016 9）、因素D（F=5.01，P=0.018 6）对试验结果的影响均有统计学意义。因篇幅所限，具体输出结果从略，下同。

由第二个过程步的输出结果可知，因素A（F=11.74，P=0.000 4）、因素B（F=5.41，P=0.013 2）、因素D（F=5.25，P=0.014 7）对试验结果的影响均有统计学意义。

因为结果的平均值越大越好（为节省篇幅，4个因素各水平下的均值计算结果从略），故因素A应取第三个水平，因素B应取第二个水平，因素C应取第三个水平，因素D应取第三个水平。

3.2.3 对例3的分析与解答

【分析与解答】所需要的SAS程序如下：

```
data a3;
input A B C Y;
cards;
1 1 1 980
1 2 2 690
（详细数据见表4的后6行）
;
run;
proc anova data=a3;
class A B C;
model Y=A B C;
/* model Y=A C;
model Y=C; */
means A B C;
run;
```

【SAS程序说明】在SAS过程步中，有三个“model语句”，第一个是可执行的，后两个放在注释语句“/* */”中是不可执行的。每次运行SAS程序，只能让一个“model语句”处在可执行状态。

【SAS输出结果及解释】

由第一个过程步的输出结果可知，仅因素C（F=154.27，P=0.006 4）对产量的影响具有统计学意义。

由第二个过程步的输出结果可知，在淘汰掉P值最大的因素B之后，仍然只有因素C（F=61.39，P=0.004 3）对产量的影响具有统计学意义。

由第三个过程步的输出结果可知，在淘汰掉因素A和因素B之后，因素C（F=23.98，P=0.002 7）对产量的影响仍有统计学意义。

4 讨论与小结

4.1 讨论

标准正交表对因素的水平数有严格要求，但在某些实际问题中，很难完全与之符合，此时，就需要对原先的正交表进行改造。例如，多个因素有3个水平，少数因素有2个水平，又没有相应的混合水平正交表可以使用。此时，可采用“拟水平法”将二水平因素改造成三水平因素[5]。在有的试验中，根据研究者经验，在安排方案时已知某些因素之间存在依赖关系，一个因素用量或水平的选取需要随另一个因素用量或水平而定，这时就需要采用“活动水平法”[5]；在有的试验中，按正交表做了一些试验后，用趋势图分析发现某一因素对试验指标的影响存在某种变化趋势，研究者认为有必要对该因素添加新的水平并追加几次试验，以便对它的影响有更全面的了解，此时，就需要采用“部分追加法”[5]。

值得注意的是，对标准正交表进行改动后，对定量资料进行方差分析的方法也需相应改变，否则，会影响计算结果的精确度；当正交表中的评价指标或结果变量有多个时，若它们之间存在一定的关联，就需要同时对它们进行分析。常规的统计分析方法是正交设计定量资料多元方差分析[8]，还有一种统计分析方法是“广义方差分析”[5]。因篇幅所限，本文不予赘述。

4.2 小结

本文介绍了正交设计的基本概念、具体实施方法以及定量资料一元方差分析的计算公式，通过三个实例，展示了二水平、三水平和混合水平正交设计的具体实施方法；还给出了无重复试验和有重复试验定量资料方差分析的SAS实现方法。

参考文献

[1] 程云鹏. 矩阵论[M]. 2版. 西安：西北工业大学出版社，2000：87.
Cheng Y. Matrix theory[M]. 2nd edition. Xi′an：Northwestern Polytechnical University Press，2000：87.

[2] 方开泰，马长兴. 正交与均匀试验设计[M]. 北京：科学出版社，2001：35-82.
Fang K，Ma C. Orthogonal and uniform experimental design[M]. Beijing：Science Press，2001：35-82.

[3] 任露泉. 试验优化设计与分析[M]. 2版. 北京：高等教育出版社，2003：79-110.
Ren L. Experimental optimization design and analysis[M]. 2nd edition. Beijing: Higher Education Press, 2003: 79-110.
[4] 黄志宏，方积乾. 数理统计方法[M]. 北京：人民卫生出版社，1987：203-218.
Huang Z, Fang J. Mathematical statistical methods[M]. Beijing: People's Medical Publishing House, 1987: 203-218.
[5] 姬振豫. 正交设计的方法与理论[M]. 香港：世界科技出版社，2001：1-164.
Ji Z. The Method and theory of orthogonal design[M]. Hong Kong: World Science and Technology Press, 2001: 1-164.
[6] 胡纯严，胡良平. 如何正确运用方差分析：单因素多水平设计定量资料一元方差分析[J]. 四川精神卫生，2022，35(1)：16-20.
Hu C, Hu L. How to use analysis of variance correctly: an analysis of variance for the univariate quantitative data collected from the design of a single factor with multi-level[J]. Sichuan Mental Health, 2022, 35(1): 16-20.
[7] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 1053-1108.
[8] 胡良平. 医学统计学：运用三型理论进行多元统计分析[M]. 北京：人民军医出版社，2010：171-176.
Hu L. Medical statistics: multivariate statistical analysis using three type theory[M]. Beijing: People's Military Medical Press, 2010: 171-176.

（收稿日期：2022-05-10）
（本文编辑：陈 霞）

科研方法专题策划人——胡良平教授简介

胡良平，男，1955年8月出生，教授，博士生导师，曾任军事医学科学院研究生部医学统计学教研室主任和生物医学统计学咨询中心主任、国际一般系统论研究会中国分会概率统计系统专业理事会常务理事、中国生物医学统计学会副会长、北京大学口腔医学院客座教授和《中华医学杂志》等10余种杂志编委；现任世界中医药学会联合会临床科研统计学专业委员会会长、国家食品药品监督管理局评审专家和3种医学杂志编委；主编统计学专著48部、参编统计学专著10部；发表第一作者和通信作者学术论文300余篇、发表合作论文130余篇；获军队科技成果和省部级科技成果多项；参加并完成三项国家标准的撰写工作、参加三项国家科技重大专项课题研究工作。在从事统计学工作的30年中，为几千名研究生、医学科研人员、临床医生和杂志编辑讲授生物医学统计学，在全国各地作统计学学术报告100余场，举办数十期全国统计学培训班，培养20多名统计学专业硕士和博士研究生。近几年来，参加国家级新药和医疗器械项目评审数十项、参加100多项全军重大重点课题的统计学检查工作。归纳并提炼出有利于透过现象看本质的“八性”和“八思维”的统计学思想，独创了逆向统计学教学法和三型理论。擅长于科研课题的研究设计、复杂科研资料的统计分析和SAS与R软件实现、各种层次的统计学教学培训和咨询工作。