合理进行均值比较——随机区组设计定量资料多元方差分析

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】 本文目的是介绍与随机区组设计定量资料多元方差分析有关的基本概念、计算方法、一个实例以及SAS实现。基本概念包括区组因素、如何选定区组因素、随机区组设计和不完全随机区组设计；计算方法涉及一般统计量和检验统计量；一个实例涉及“长期饲喂高锌日粮对断奶仔猪免疫机能影响的动物试验及其多元定量资料”。借助SAS实现随机区组设计定量资料的一元方差分析和多元方差分析。并讨论当区组因素对结果的影响无统计学意义时，合理的统计分析方法是不考虑区组因素，直接采用单因素多水平设计一元和多元定量资料方差分析。

【关键词】 区组因素；随机区组设计；平衡不完全区组设计；一元方差分析；多元方差分析
中图分类号：R195.1 **文献标识码**：A doi：10.11886/scjsws20230319003

Reasonably carry out mean value comparison：MANOVA of the quantitative data collected from the randomized block 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 article was to introduce the basic concepts, calculation methods, an example and SAS implementation related to the randomized block design quantitative data multivariate analysis of variance (MANOVA). Basic concepts included block factors, how to select block factors, randomized block design and incomplete randomized block design. Calculation methods involved the general statistics and test statistics. The example involved long-term feeding of high-zinc diets animal experiments and multivariate quantitative data on the effect on the immune function of weaned piglets. With the help of SAS software, the one-way analysis of variance (ANOVA) and MANOVA for the quantitative data in the randomized block design were realized. And it was discussed that when the influence of block factors on the results was not statistically significant, the reasonable statistical analysis method was to directly use single factor multilevel design quantitative data univariate and multivariate ANOVA without considering block factor.

【**Keywords**】 Block factor; Randomizedblock design; Balanced incomplete block design; One-way analysis of variance; Multivariate analysis of variance

由于医学问题的复杂性，在医学试验研究中，当研究者希望重点考查某一个试验因素对结果的影响时，常采取的措施是增加样本含量，并采取随机化方法分配受试对象，以便最大程度地减弱或消除来自受试对象的许多非试验因素对观测结果的影响。然而，通常情况下，受人力、物力、财力和时间的限制，可获取的样本含量非常有限，即便采取了随机化方法分配受试对象，也很难保证来自受试对象的各种非试验因素在各试验组之间处于很好的平衡状态，此时，一个行之有效的方法是采用随机区组设计安排试验。本章将介绍与随机区组设计定量资料多元方差分析有关的基本概念、计算方法、一个医学实例以及SAS实现。

1 基本概念

1.1 区组因素

影响结果变量取值的所有因素可以被统称为影响因素，可以被大致分为试验因素与非试验因素。所谓试验因素，就是研究者特别关注的影响因素，通常是研究者施加给受试对象的，例如试验药物的种类和剂量等；在一个特定的研究项目中，试

验因素之外的所有其他影响因素都被称为非试验因素，其中，有一些来自受试对象本身的条件，例如窝别（动物试验）、性别、血型、职业等，这些因素有时被称为“属性因素（反映受试对象的某种特性）”，有时也被称为“区组因素（具有相同水平的受试对象形成一个小组，例如来自同一窝的4只小鼠形成一个小组，血型相同且患同一种疾病的5例患者形成一个小组）”。

1.2 如何选定区组因素

在一项科学研究中，通常都会涉及“受试对象”。当受试对象是人、动物、样品等时，就需要结合具体的研究问题，找出来自受试对象的所有可能影响观测结果的非试验因素，基于基本常识和专业知识，将所列出的非试验因素按主次排序；再基于实际操作的可行性，选取最重要的几个非试验因素，由它们复合成一个区组因素。在实际应用中，区组因素可能就是非试验因素中最重要的一个。例如，在一项动物试验中，若受试对象是小鼠，则“窝别”就可被确定为一个区组因素，因为“窝别”不仅能体现出生时间和条件的一致性，还能体现“遗传因素”的影响。若再加上“性别”因素，即每窝中只取同一种性别的小鼠，或者每窝中雌性与雄性小鼠数目相等，这样形成的“区组因素”就是由“窝别”与“性别”两个属性因素复合而成的。

1.3 随机区组设计

随机区组设计，也称为随机完全区组设计[1-2]，其具体做法如下：第一步，确定一个试验因素及其水平数，设水平数为k；第二步，确定评价指标（一般包括定量和定性的指标）；第三步，确定符合研究目的且具有同质性的某种受试对象；第四步，基于专业知识和基本常识，确定“区组因素（可以是一个属性因素，也可以是多个属性因素复合而成的）”；第五步，将所选取的全部受试对象（严格地说，应基于统计学和专业要求，估计出合适的样本含量）按事先确定的“区组因素”形成多个区组或小组，每个区组或小组中受试对象的个数必须是k的整数倍（通常就是k）；第六步，将每个区组中的受试对象完全随机地均分入k个试验组中去。

1.4 不完全随机区组设计

不完全随机区组设计，也称为平衡不完全随机区组设计[1-2]。此设计常用于每个区组内的试验单位数小于试验因素水平数的试验研究场合。例如，研究者希望考查4种药物对脚气病的疗效。受试对象是双脚患有脚气病的患者，若将每位患者视为一个“区组”，显然，每人只有两只脚，不便接受4种药物治疗。按基本常识来考虑试验安排，从4种药物中取2种来组成一个患者的治疗方案，共有6种方案，即至少需要6名患者，才能确保所有方案都可以被实施，并且，任何两种药物被使用的次数应相等，这就是“平衡”之义；又由于每个区组（一个人）中只有2个试验单位（两只脚），不能完全接受4种药物治疗，故称之为“不完全随机区组”。根据试验因素的水平数和每个区组内的试验单位数的不同，需要符合一定的要求，方可实现不完全区组设计。具体做法如下：设r为试验因素每个水平重复施加的次数，v为试验因素的水平数，a为每个区组中受试对象的个数，b为区组的个数，λ为试验因素任何两个水平同时出现的区组数。不完全随机区组设计，要求以上所提及的5个参数之间应满足下列两个条件：①rv=ab；②λ=r(a−1)/(v−1)必须是整数。

2 计算方法

2.1 一般统计量

在随机区组设计中，设A为试验因素，其水平数为a；B为区组因素，其水平数为b；试验中的总例数$n=ab$；需要观测的结果变量的个数为m。用X_{ij}表示试验因素的第i个水平与区组因素的第j个水平组合下的观测向量，$i=1,2,\cdots,a$，$j=1,2,\cdots,b$；$\bar{X}$表示所有观测的样本均值向量；$\bar{X}_{i\cdot}$表示试验因素第i个水平下，也就是第i个处理组的样本均值向量；$\bar{X}_{\cdot j}$表示区组因素第j个水平下，也就是第j个区组的样本均值向量。

用多元方差分析处理随机区组设计的定量资料时，全部观测的总变异由总离均差矩阵T表示，处理间变异由试验因素A的离均差矩阵H_A表示，区组间变异由区组因素B的离均差矩阵H_B表示，误差变异由误差离均差矩阵E表示。当总变异被分解为处理间变异、区组间变异与误差变异时，相应的总离均差矩阵被分解为试验因素A的离均差矩阵、区组因素B的离均差矩阵与误差离均差矩阵，见式(1)。

$$T = H_A + H_B + E \tag{1}$$

总离均差矩阵、试验因素A的离均差矩阵、区组因素B的离均差矩阵与误差离均差矩阵的表示分别见式(2)、式(3)、式(4)、式(5)。

$$T = \sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij} - \bar{X})(X_{ij} - \bar{X})' \tag{2}$$

$$H_A = b\sum_{i=1}^{a}(\bar{X}_{i\cdot} - \bar{X})(\bar{X}_{i\cdot} - \bar{X})' \quad (3)$$

$$H_B = a\sum_{j=1}^{b}(\bar{X}_{\cdot j} - \bar{X})(\bar{X}_{\cdot j} - \bar{X})' \quad (4)$$

$$E=\sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}-\bar{X})(X_{ij}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}-\bar{X})' \quad (5)$$

上述呈现的内容可以被总结为表1。

表1 随机区组设计定量资料多元方差分析公式汇总

Table 1 Summary of multivariate analysis of variance formulas for quantitative datain randomized block design

变异来源	自由度	离均差矩阵
总变异	$n-1$	式(1)或式(2)
处理间	a－1	式(3)
区组间	$b-1$	式(4)
误差	$(a-1)(b-1)$	式(5)

2.2 检验统计量

根据上述离均差矩阵，可以计算 Wilks′ λ 检验统计量，对试验因素 A 和区组因素 B 进行检验的检验统计量分别见式(6)和式(7)。

$$\lambda_A = \frac{|E|}{|H_A + E|} \quad (6)$$

$$\lambda_B = \frac{|E|}{|H_B + E|} \quad (7)$$

在式(6)和式(7)中，|*|代表求“*”的行列式的值，算得 Wilks′ λ 检验统计量之后，再将其进一步转化为 F 统计量(转换公式参见文献[3-4])，就可以实现对试验因素和区组因素的检验。

3 实例与SAS实现

3.1 问题与数据结构

3.1.1 一个实例及数据

【例1】为了研究长期饲喂高锌日粮对断奶仔猪免疫机能的影响，根据窝别和性别，将60头仔猪分为20个区组，每个区组内的3只仔猪随机分配到A、B、C三个试验组中。三组均饲喂基础日粮，在此基础上，B组增加3 000 mg/kg氧化锌，C组增加500 mg/kg氧化锌。于断奶后第70天检测三组仔猪的血清免疫球蛋白G(IgG)、免疫球蛋白A(IgA)和免疫球蛋白M(IgM)水平。资料见表2[5]。假定资料满足参数检验的前提条件，试比较三组仔猪的血清IgG、IgA和IgM水平差异有无统计学意义。

表2 各组仔猪的血清IgG、IgA和IgM水平(g/L)

Table 2 Serum IgG, IgA and IgM levels of piglets in each group

区 组	IgG水平			IgA水平			IgM水平		
	A组	B组	C组	A组	B组	C组	A组	B组	C组
1	0.383	0.222	0.211	0.138	0.078	0.090	0.325	0.273	0.218
2	0.465	0.224	0.179	0.137	0.069	0.092	0.276	0.214	0.218
3	0.409	0.202	0.200	0.151	0.081	0.084	0.281	0.263	0.174
4	0.444	0.138	0.198	0.072	0.075	0.090	0.310	0.222	0.260
5	0.458	0.142	0.246	0.112	0.087	0.078	0.282	0.234	0.253
6	0.425	0.217	0.247	0.138	0.066	0.084	0.236	0.190	0.229
7	0.368	0.133	0.222	0.136	0.087	0.075	0.265	0.260	0.223
8	0.385	0.207	0.228	0.123	0.086	0.068	0.281	0.275	0.289
9	0.363	0.184	0.185	0.130	0.090	0.081	0.296	0.216	0.256
10	0.428	0.119	0.227	0.109	0.070	0.075	0.286	0.268	0.251
11	0.408	0.156	0.198	0.139	0.074	0.071	0.301	0.255	0.186
12	0.414	0.187	0.148	0.123	0.085	0.086	0.272	0.262	0.269
13	0.332	0.169	0.177	0.136	0.071	0.095	0.252	0.236	0.227
14	0.452	0.191	0.209	0.134	0.081	0.081	0.275	0.216	0.218
15	0.369	0.214	0.200	0.111	0.078	0.070	0.265	0.289	0.245
16	0.460	0.159	0.196	0.108	0.075	0.076	0.286	0.273	0.238
17	0.371	0.225	0.224	0.108	0.086	0.084	0.328	0.252	0.236
18	0.412	0.165	0.205	0.112	0.069	0.071	0.290	0.265	0.202
19	0.415	0.185	0.198	0.147	0.087	0.086	0.269	0.292	0.170
20	0.471	0.193	0.268	0.121	0.097	0.083	0.310	0.235	0.184

注：区组是由窝别与性别两个属性因素复合而成的一个因素，即每个区组内的3只仔猪来自同一窝且性别相同；IgG，免疫球蛋白G；IgA，免疫球蛋白A；IgM，免疫球蛋白M；A、B、C分别代表3种不同的饲料种类

3.1.2 对数据结构的分析

在表2中，区组因素是一个重要的非试验因素，而饲料种类是一个试验因素，血清IgG、IgA和IgM是3个定量指标，每行上有3只仔猪，故总样本含量为60。该资料应为随机区组设计三元定量资料。

3.1.3 创建SAS数据集

设所需要的SAS数据步（注意：各列数据顺序与表2中不一样）程序如下[6]。

```
data a1;
do block=1 to 20;
do treat='A','B','C';
input IgG IgA IgM @@;
output;
end;end;
cards;
0.383 0.138 0.325 0.222 0.078 0.273 0.211
0.090 0.218
0.465 0.137 0.276 0.224 0.069 0.214 0.179
0.092 0.218
（此处省略多行数据）
0.415 0.147 0.269 0.185 0.087 0.292 0.198
0.086 0.170
0.471 0.121 0.310 0.193 0.097 0.235 0.268
0.083 0.184
;
run;
```

【变量说明】block代表区组因素，treat代表饲料种类，IgG、IgA、IgM代表3个定量结果变量。

3.2 用SAS实现统计分析

考虑区组因素对结果的影响，设所需要的SAS过程步程序如下：

```
proc glm data=a1;
class block treat;
model IgG IgA IgM=block treat/ss3;
manova H=block treat;
run;quit;
```

【SAS输出结果及解释】对block而言，Wilks'λ=0.263，F=1.070，分子和分母的自由度分别为57和108，P=0.373，说明由3个定量指标组成的均值向量在20个区组之间的差异无统计学意义；对treat而言，Wilks'λ=0.028，F=59.450，分子和分母的自由度分别为6和72，P<0.001，说明由3个定量指标组成的均值向量在三组之间的差异有统计学意义。

不考虑区组因素对结果的影响，设所需要的SAS过程步程序如下：

```
proc glm data=a1;
class treat;
model IgG IgA IgM=treat/ss3;
contrast 'A vs B' treat 1 -1 0;
contrast 'A vs C' treat 1 0 -1;
contrast 'B vs C' treat 0 1 -1;
manova H=treat;
means treat;
run;quit;
```

【SAS输出结果及解释】第一部分，一元方差分析的结果如下。

对IgG而言，3个水平组之间整体比较，F=280.890，分子和分母的自由度分别为2和57，P<0.001；A组与B组比较，F=469.510，分子和分母的自由度分别为1和57，P<0.001；A组与C组比较，F=366.830，分子和分母的自由度分别为1和57，P<0.001；B组与C组比较，F=6.330，分子和分母的自由度分别为1和57，P=0.015。整体和3个水平组之间两两比较结果均有统计学意义。

对IgA而言，3个水平组之间整体比较，F=82.750，分子和分母的自由度分别为2和57，P<0.001；A组与B组比较，F=128.020，分子和分母的自由度分别为1和57，P<0.001；A组与C组比较，F=120.120，分子和分母的自由度分别为1和57，P<0.001；B组与C组比较，F=0.130，分子和分母的自由度分别为1和57，P=0.724。整体和3个水平组之间两两比较结果中，仅B组与C组之间差异无统计学意义。

对IgM而言，3个水平组之间整体比较，F=21.170，分子和分母的自由度分别为2和57，P<0.001；A组与B组比较，F=15.530，分子和分母的自由度分别为1和57，P<0.001；A组与C组比较，F=41.670，分子和分母的自由度分别为1和57，P<0.001；B组与C组比较，F=6.320，分子和分母的自由度分别为1和57，P=0.015。整体和3个水平组之间两两比较结果均有统计学意义。

第二部分，三元方差分析的结果如下。

试验因素treat的3个水平组之间整体比较：Wilks'λ=0.049，F=64.520，分子和分母的自由度分别为6和110，P<0.001，说明由3个定量指标组成的均值向量在3个水平组之间的差异有统计学意义。

试验因素treat的3个水平组均值向量之间两两

比较：A 组与 B 组比较，Wilks′λ=0.071，F=240.500，分子和分母的自由度分别为 3 和 55，P<0.001；A 组与 C 组比较，Wilks′λ=0.079，F=212.50，分子和分母的自由度分别为 3 和 55，P<0.001；B 组与 C 组比较，Wilks′λ=0.821，F=4.000，分子和分母的自由度分别为 3 和 55，P=0.012。

由 IgG、IgA、IgM 这 3 个定量指标组成的均值向量$\left[\overline{\mathrm{IgG}},\ \overline{\mathrm{IgA}},\ \overline{\mathrm{IgM}}\right]'$在 A、B、C 三组中的计算结果分别为 $[0.412, 0.124, 0.284]'$、$[0.182, 0.080, 0.250]'$、$[0.208, 0.081, 0.227]'$。

【结论】对于 IgG、IgA、IgM 这 3 个定量指标而言，为了在试验因素 treat 的 3 个水平组之间进行整体比较和两两比较，无论是进行一元方差分析还是进行多元方差分析，除了 IgA 在 B 组与 C 组之间进行一元方差分析得到的结果差异无统计学意义外，其他所有的比较（包括一元分析与三元分析、整体比较与两两比较）均有统计学意义。具体地说，仅食用基础日粮的 A 组仔猪的 IgG、IgA、IgM 的值均较高；食用基础日粮和 500 mg/kg 氧化锌的 C 组仔猪的 IgG、IgA、IgM 的值明显低于 A 组；而食用基础日粮和 3 000 mg/kg 氧化锌的 B 组仔猪的 IgG、IgA、IgM 的值明显低于 A 组，但 IgG 的值低于 C 组、IgM 的值高于 C 组。

4 讨论与小结

4.1 讨论

在试验设计阶段，是否需要选用随机区组设计取代单因素多水平设计，取决于是否能够找到可能影响结果变量取值的重要非试验因素作为“区组因素”，以及能否将受试对象按区组因素形成多个不同的小组；在对定量资料进行差异性分析时，是否一定要选用随机区组设计定量资料多元方差分析取代单因素多水平设计定量资料多元方差分析，取决于区组因素对定量结果变量的影响程度。若区组因素对定量结果变量的影响有统计学意义，则必须采用随机区组设计定量资料的一元和多元方差分析；反之，就应选用单因素多水平设计定量资料一元和多元方差分析，此时，误差项的自由度增大了，方差分析结果的可靠性增加。一元方差分析是多元方差分析的基础，有关随机区组设计和平衡不完全随机区组设计定量资料一元方差分析的细节，可参阅文献[7-8]。

4.2 小结

本文介绍了与随机区组设计定量资料多元方差分析有关的基本概念、计算方法、一个实例及其 SAS 实现。基本概念包括区组因素、如何选定区组因素、随机区组设计和不完全随机区组设计；计算方法包括一般统计量和检验统计量；实例涉及“长期饲喂高锌日粮对断奶仔猪免疫机能影响的动物试验及其多元定量资料”。基于 SAS 实现了随机区组设计定量资料的多元方差分析，并讨论了如何正确看待区组因素在定量资料差异性分析中的作用。

参考文献

[1] Montgomery DC. Design and analysis of experiments[M]. 6 版. 北京：人民邮电出版社，2007：119-159.
Montgomery DC. Design and analysis of experiments[M]. 6th edition. Beijing：Posts & Telecom Press，2007：119-159.

[2] Dean A，Voss D. Design and analysis of experiments[M]. 北京：世界图书出版公司，2010：295-386.
Dean A，Voss D. Design and analysis of experiments[M]. Beijing：World Book Publishing Company，2010：295-386.

[3] Wilks SS. Certain generalization in the analysis of variance[J]. Biometrika，1932，24(3-4)：471-494.

[4] Rao CR. An asymptotic expansion of the distribution of Wilks′λ criterion[J]. Bull Inst Internat Statist，1951，33(Part Ⅱ)：177-180.

[5] 胡良平. 面向问题的统计学：(3)试验设计与多元统计分析[M]. 北京：人民卫生出版社，2012：390-400.
Hu LP. Problem-oriented statistics：(3) experimental design and multivariate statistical analysis[M]. Beijing：People′s Medical Publishing House，2012：390-400.

[6] SAS Institute Inc. SAS/STAT®15.1 user′s guide[M]. Cary，NC：SAS Institute Inc，2018：3957-4142.

[7] 胡纯严，胡良平. 如何正确运用方差分析：随机完全区组设计定量资料一元方差分析[J]. 四川精神卫生，2022，35(2)：97-102.
Hu CY，Hu LP. How to use analysis of variance correctly：an analysis of variance for the univariate quantitative data collected from the randomized complete block design[J]. Sichuan Mental Health，2022，35(2)：97-102.

[8] 胡纯严，胡良平. 如何正确运用方差分析：平衡不完全区组设计定量资料一元方差分析[J]. 四川精神卫生，2022，35(2)：103-107.
Hu CY，Hu LP. How to use analysis of variance correctly：an analysis of variance for the univariate quantitative data collected from the balanced incomplete block design[J]. Sichuan Mental Health，2022，35(2)：103-107.

（收稿日期：2023-03-19）
（本文编辑：陈 霞）