·科研方法专题·

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

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】 本文目的是介绍随机完全区组设计定量资料一元方差分析的模型、计算公式和SAS实现。在计算中,涉及F_A和F_B两个检验统计量,其中,A代表试验因素,B代表区组因素(即重要的非试验因素)。一般来说,假定随机区组设计中的两个因素之间不存在交互作用或交互作用可以忽略不计,故不需要评价交互作用项是否具有统计学意义,因此,在两因素各水平组合下可以不做重复试验。本文借助SAS软件,分别对无和有重复试验数据的两个实例进行随机完全区组设计定量资料一元方差分析,并给出计算结果,作出统计结论和专业结论。

【关键词】 方差分析;F检验;定量资料;设计类型;随机区组设计;交互作用
中图分类号:R195.1 **文献标识码**:A doi:10.11886/scjsws20220310001

How to use analysis of variance correctly——an analysis of variance for the univariate quantitative data collected from the randomized complete 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 paper was to introduce the model, calculation formulas and the SAS implementation of the univariate analysis of variance for the quantitative data with randomized complete block design. In the calculation, two test statistics were involved, namely F_A and F_B. Among them, the subscript "A" represented the experimental factor, and the subscript "B" represented the block factor (i.e., the important non-experimental factor). In general, it was assumed that there was no or negligible interaction between the two factors in a randomized block design, so there was no need to assess whether the interaction term was statistically significant. Therefore, it was not necessary to do repeated experiments under each combination of two factors. With the help of SAS software, this paper conducted the analysis of variance for the quantitative data with randomized complete block design for two instances without and with repeated experiments, gave the calculation results, and made the statistical and professional conclusions.

【**Keywords**】 Analysis of variance; *F* test; Quantitative data; Design type; Randomized block design; Interaction

随机完全区组设计简称为随机区组设计或配伍组设计[1-2],它是考察一个试验因素和一个区组因素对定量观测结果影响的一个节省样本含量的试验设计方法。本文将介绍该设计类型的要点、定量资料一元方差分析的模型和计算公式,以及基于SAS软件实现定量资料一元方差分析的具体方法。

1 随机完全区组设计的要点

在单因素多水平的试验研究场合中,若全部受试对象可以按某种重要的属性(例如动物的窝别、样品的批次、患者的血型、受试对象的工作车间等)被分成几个小组,则此时就可采用随机完全区组设计取代单因素多水平设计,以便排除区组因素对结果变量的影响。

随机完全区组设计的具体实施方法:基于定量观测指标,依据研究目的和专业知识,确定试验因素及其水平,并找出对定量观测指标影响最明显且来自受试对象的一个属性变量(也叫区组因素),将属性变量取值(即水平)相同的受试对象划分为一个大组;设试验因素有r个水平,区组因素有s个水

平。先从依据研究目的确定的具有同质性的总体中随机抽取s组受试对象，应确保每组受试对象的个数≥r；再从每组中随机抽取r个受试对象并随机分配进入r个试验组中；最后，从每个受试对象身上测定定量观测指标的数值。随机完全区组设计的呈现模式见表1。

表1 随机完全区组设计一元定量资料的呈现模式

A	定量观测指标：Y_{ij}				
B：	1	2	…	s	均值
1	Y_{11}	Y_{12}	…	Y_{1s}	$\bar{Y}_{1.}$
2	Y_{21}	Y_{22}	…	Y_{2s}	$\bar{Y}_{2.}$
…	…	…	…	…	…
r	Y_{s1}	Y_{s2}	…	Y_{rs}	$\bar{Y}_{r.}$
均值	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$	…	$\bar{Y}_{.s}$	$\bar{Y}_{..}$

注：A为试验因素，B为区组因素

2 随机完全区组设计定量资料一元方差分析

2.1 方差分析的模型

假定试验因素A与区组因素B之间的交互作用不存在或可以忽略不计，于是，它们之间各水平组合条件下可以不做重复试验；又假定试验因素A和区组因素B分别有r和s个水平。随机完全区组设计定量资料一元方差分析模型[3-4]见式(1)：

$$\begin{cases} Y_{ij}=\mu+a_i+b_j+\varepsilon_{ij} \quad i=1,2,\cdots,r \quad j=1,2,\cdots,s \\ \sum_{i=1}^{r}a_i=0 \quad \sum_{j=1}^{s}b_j=0 \\ \varepsilon_{ij}\sim \text{iid} \quad N(0,\sigma^2) \end{cases} \tag{1}$$

在式(1)中，Y_{ij}为两因素A、B的(i,j)水平组合下定量观测结果，μ为全部条件下定量观测结果的总体平均值，a_i为试验因素A的第i个水平的效应，b_j为区组因素B的第j个水平的效应，它们满足以下关系式：

$$\mu=\frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s}\mu_{ij} \tag{2}$$

$$a_i=\mu_{i.}-\mu \tag{3}$$

$$b_j=\mu_{.j}-\mu \tag{4}$$

$$\mu_{i.}=\frac{1}{s}\sum_{j=1}^{s}\mu_{ij} \tag{5}$$

$$\mu_{.j}=\frac{1}{r}\sum_{i=1}^{r}\mu_{ij} \tag{6}$$

在以上各式中，μ_{ij}为两因素A、B的(i,j)水平组合下定量观测结果的总体平均值。

基于最大似然法，可得到式(2)、式(3)、式(4)的最大似然估计值，分别见式(7)、式(8)、式(9)：

$$\hat{\mu}=\bar{Y} \tag{7}$$

$$\hat{a}_i=\bar{Y}_{i.}-\bar{Y},i=1,2,\cdots,r \tag{8}$$

$$\hat{b}_j=\bar{Y}_{.j}-\bar{Y},j=1,2,\cdots,s \tag{9}$$

为检验一切μ_{ij}是否相等，可改为检验以下两个假设。

检验试验因素A的无效假设与备择假设分别见式(10)、式(11)。

$$H_{0A}:a_1=a_2=\cdots=a_r=0 \tag{10}$$

$$H_{1A}:a_1,a_2,\cdots,a_r\text{不全为}0 \tag{11}$$

检验区组因素B的无效假设与备择假设分别见式(12)、式(13)。

$$H_{0B}:b_1=b_2=\cdots=b_s=0 \tag{12}$$

$$H_{1B}:b_1,b_2,\cdots,b_s\text{不全为}0 \tag{13}$$

2.2 方差分析的公式

从前面式(1)的第1个式子可以看出，随机完全区组设计定量资料一元方差分析模型共有$r\times s$个，即便将式(7)、式(8)、式(9)代入其中，获得基于观测数据的全部模型的最终计算结果，仍然没有解决所需要回答的问题：即试验因素A各水平对观测结果的影响差别是否有统计学意义[对应的检验假设为前文的式(10)和式(11)]？区组因素B各水平对观测结果的影响差别是否有统计学意义[对应的检验假设为前文的式(12)和式(13)]？为了回答这两个问题，需要构造出两个F检验统计量，分别见式(14)、式(15)：

$$F_A=\frac{MS_A}{MS_E}=\frac{SS_A/df_A}{SS_E/df_E} \tag{14}$$

$$F_B=\frac{MS_B}{MS_E}=\frac{SS_B/df_B}{SS_E/df_E} \tag{15}$$

在式(14)和式(15)中，MS_A、MS_B、MS_E分别代表试验因素A、区组因素B和试验误差E的均方；SS_T、SS_A、SS_B、SS_E分别代表全部数据(简称“T”)、试验因素A、区组因素B和试验误差E的离均差平方和；而df_T、df_A、df_B、df_E分别代表总变异T、试验因素A、区组因素B和试验误差E的自由度。各项离均差平方和的计算公式如下：

$$SS_T=\sum_{i=1}^{r}\sum_{j=1}^{s}\left(Y_{ij}-\bar{Y}\right)^2=\sum_{i=1}^{r}\sum_{j=1}^{s}Y_{ij}^2-\frac{1}{n}\left(\sum_{i=1}^{r}\sum_{j=1}^{s}Y_{ij}\right)^2 \tag{16}$$

$$SS_A=\sum_{i=1}^{r}s\left(\bar{Y}_{i.}-\bar{Y}\right)^2=\sum_{i=1}^{r}\frac{Y_{i.}^2}{s}-\frac{1}{n}\left(\sum_{i=1}^{r}\sum_{j=1}^{s}Y_{ij}\right)^2 \tag{17}$$

$$SS_B=\sum_{j=1}^{s}r\left(\bar{Y}_{.j}-\bar{Y}\right)^2=\sum_{j=1}^{s}\frac{Y_{.j}^2}{r}-\frac{1}{n}\left(\sum_{i=1}^{r}\sum_{j=1}^{s}Y_{ij}\right)^2 \tag{18}$$

$$SS_E = SS_T - SS_A - SS_B \quad (19)$$

各项自由度的计算公式如下：

$$df_T = n - 1 \quad (20)$$

$$df_A = r - 1 \quad (21)$$

$$df_B = s - 1 \quad (22)$$

$$df_E = df_T - df_A - df_B \quad (23)$$

将以上主要公式汇集在一张表中，见表2。

表2　随机完全区组设计两因素各水平组合下进行一次试验的方差分析表

变异来源	离均差平方和(SS)	自由度(df)	均方(MS)	F
试验因素A	SS_A	$r-1$	SS_A/df_A	MS_A/MS_E
区组因素B	SS_B	$s-1$	SS_B/df_B	MS_B/MS_E
误差E	SS_E	$df_T-df_A-df_B$	SS_E/df_E	–
总和T	SS_T	$n-1$	–	–

3　随机完全区组设计一元定量资料的实例与SAS实现

3.1　实例与数据结构

【例1】为探索丹参对肢体缺血再灌注损伤的影响，研究者将30只纯种新西兰实验用大白兔按窝别分为10个区组，每个区组的3只大白兔（来自同一窝）随机接受三种不同处理，即在松止血带前分别给予丹参2 mL/kg（A_1）、丹参1 mL/kg（A_2）、生理盐水2 mL/kg（A_3），并分别测定松止血带前、后1小时内大白兔血中白蛋白含量（g/L），计算白蛋白的减少量[5]，见表3。问三种处理的平均值之间差异是否有统计学意义。

表3　随机完全区组设计下三种处理后大白兔血中白蛋白减少量（g/L）

窝　别	白蛋白减少量		
处理因素A：	A_1	A_2	A_3
1	2.21	2.91	4.25
2	2.32	2.64	4.56
…	…	…	…
10	3.42	2.86	4.23

注：详细数据见后文SAS程序

【例2】某研究者希望研究三种减肥方案（试验因素）的效果，分别从两个工作地点（即区组因素，其两个水平分别为“办公室”与“车间”）中随机选取女工15名，再将每个工作地点中的15名女工随机均分入三种减肥方案组中。试验开始之前，测定她们的空腹体重；在接受试验一定时间之后，再测定空腹体重，用第1次测定结果减去第2次测定结果，得到体重改变量（正值为体重减少量，负值为体重增加量）[6]，见表4。问哪种减肥方案效果最佳？

表4　三种减肥方案中办公室和车间女工的体重改变量（磅）

工作地点	体重改变量														
减肥方案：	食疗					运动					食疗+运动				
办公室	6	2	10	−1	8	3	4	−2	6	−2	8	12	7	10	5
车间	3	15	4	8	6	−4	6	8	−2	3	15	8	10	16	3

注：减肥方案为试验因素，工作地点为区组因素；两因素各水平组合下均做了5次独立重复试验

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】设例1资料所需要的SAS程序如下：

```
data abc;
do B=1 to 10;
do A=1 to 3;
input Y @@;
output;
end; end;
cards;
2.21 2.91 4.25
2.32 2.64 4.56
3.15 3.67 4.33
1.86 3.29 3.89
2.56 2.45 3.78
1.98 2.74 4.62
2.37 3.15 4.71
2.88 3.44 3.56
3.05 2.61 3.77
3.42 2.86 4.23
;
run;
proc glm data=abc;
class A B;
model Y=A B/ss3;
means A/lsd snk tukey;
run;
```

【SAS输出结果及解释】

源	自由度	平方和	均方	F	Pr>F
模型	11	15.259 386 67	1.387 216 97	6.61	0.000 2
误差	18	3.778 493 33	0.209 916 30		
校正合计	29	19.037 880 00			

源	自由度	III型SS	均方	F	Pr>F
A	2	13.701 840 00	6.850 920 00	32.64	<0.000 1
B	9	1.557 546 67	0.173 060 74	0.82	0.602 4

以上第一部分为随机完全区组设计定量资料一元方差分析总模型的输出结果，F=6.61，P=0.000 2，表明方差分析模型具有统计学意义（误差项的自由度=18）。

以上第二部分输出的是随机完全区组设计定量资料一元方差分析的主要结果，结果表明：处理因素A对白蛋白减少量的影响是不同的（F=32.64，P<0.000 1）；而窝别因素对白蛋白减少量的影响无统计学意义，即窝别对结果的影响可以忽略不计。

为节省篇幅，下面仅给出采用TUKEY法对三种处理下的三个均值进行两两比较的结果，见图1。由图1可看出：处理组1、2、3的均值分别为2.580、2.976和4.170；两两比较结果显示，处理组1与组2的均值之间差异无统计学意义，而它们与处理组3的均值差异均有统计学意义。说明相对于生理盐水而言，大白兔接受1 mL/kg或2 mL/kg的丹参处理后，白蛋白的含量明显下降。

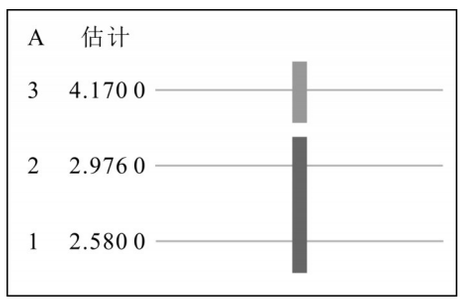

图1 采用TUKEY法对三种处理下的三个均值进行两两比较的结果

源	自由度	平方和	均方	F	Pr>F
模型	2	13.701 840 00	6.850 920 00	34.67	<0.000 1
误差	27	5.336 040 00	0.197 631 11		
校正合计	29	19.037 880 00			

以上为单因素三水平设计定量资料一元方差分析总模型的输出结果，F=34.67，P<0.000 1，表明方差分析模型具有统计学意义（误差项的自由度=27）。

以上第一行结果是单因素三水平设计定量资料一元方差分析的主要结果，结果表明：处理因素A对白蛋白减少量的影响是不同的（F=34.67，P<0.000 1）。

采用TUKEY法对三种处理下的三个均值进行两两比较的结果同图1，不再赘述。

3.2.2 对例2的分析与解答

【分析与解答】设例2资料所需要的SAS程序如下：

```
data abc;
do B=1 to 2;
do A=1 to 3;
do R=1 to 5;
input Y @@;
output;
end; end; end;
cards;
6 2 10 -1 8
3 4 -2 6 -2
8 12 7 10 5
3 15 4 8 6
-4 6 8 -2 3
15 8 10 16 3
```

由于窝别因素对结果的影响可以忽略不计，故本例资料采用单因素三水平设计定量资料一元方差分析更合适，这样可以增大误差项的自由度，使分析结果更加稳定。可采用如下SAS过程步：

```
proc glm data=abc;
class A;
model Y=A/ss3;
means A/lsd snk tukey;
run;
```

【SAS输出结果及解释】

```
;
run;
proc glm data=abc;
class A B;
model Y=A B/ss3;
means A/SNK TUKEY;
run;
proc glm data=abc;
class A;
```

源	自由度	平方和	均方	F	Pr>F
模型	3	292.500 000 0	97.500 000 0	5.33	0.005 4
误差	26	475.666 666 7	18.294 871 8		
校正合计	29	768.166 666 7			

源	自由度	III型SS	均方	F	Pr>F
A	2	274.866 666 7	137.433 333 3	7.51	0.002 7
B	1	17.633 333 3	17.633 333 3	0.96	0.335 3

以上第一部分结果表明：随机完全区组设计定量资料一元方差分析模型有统计学意义（F=5.33，P=0.005 4），误差项的自由度为26。

以上第二部分结果表明：减肥方案A的3个水平组均值之间差别有统计学意义（F=7.51，P=0.002 7），表明不同减肥方案的减肥效果是不同的；而工作地点B之间差异无统计学意义（F=0.96，P=0.335 3），表明不同工作地点对减肥效果的影响可忽略不计。

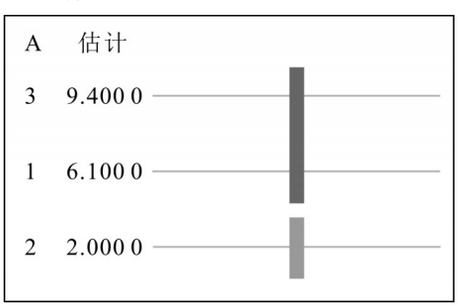

图2　采用SNK法对三种减肥方案下的三个均值进行两两比较的结果

由图2可知：方案3与方案1的均值之间差异无统计学意义，而方案3和方案1的均值与方案2的均值之间差异均有统计学意义。方案3、1、2对应的体重减少量的均值依次为9.4、6.1和2.0磅。

```
model Y=A/ss3;
means A/SNK TUKEY;
run;
```

【SAS程序说明】第1个过程步是进行随机完全区组设计定量资料一元方差分析；而第2个过程步是进行单因素（指因素A：减肥方案）三水平设计定量资料一元方差分析（前提条件是区组因素无统计学意义，否则，不可以使用第2个过程步）。

【SAS输出结果及解释】

源	自由度	平方和	均方	F	Pr>F
模型	2	274.866 666 7	137.433 333 3	7.52	0.002 5
误差	27	493.300 000 0	18.270 370 4		
校正合计	29	768.166 666 7			

以上为单因素三水平设计定量资料一元方差分析总模型的输出结果，F=7.52，P=0.002 5，表明方差分析模型有统计学意义（误差项的自由度=27）。

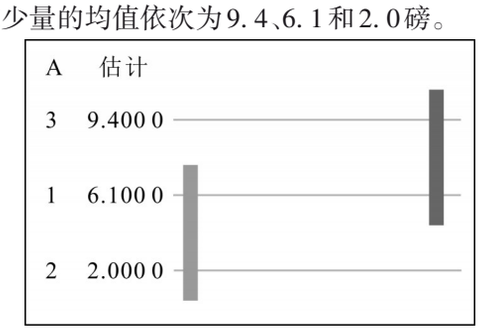

图3　采用TUKEY法对三种减肥方案下的三个均值进行两两比较的结果

由图3可知：方案3与方案1的差异无统计学意义，而方案1与方案2的差异也无统计学意义，但方案3与方案2的差异有统计学意义。

由于工作地点因素对结果的影响可以忽略不计，故本例资料采用单因素三水平设计定量资料一元方差分析（见前文的第2个SAS过程步）更合适，这样可以增大误差项的自由度，使分析结果更加稳定。其SAS输出结果如下：

均方	F	Pr>F
137.433 333 3	7.52	0.002 5
18.270 370 4		

以上第一行是单因素三水平设计定量资料一元方差分析的主要结果，结果表明：减肥方案因素A对体重减少量的影响是不同的（F=7.52，P=0.002 5）。

对减肥方案因素 A 的三个水平下的均值进行两两比较，基于 SNK 法和 TUKEY 法得到的结果分别与图 2 和图 3 的结果相同，为节省篇幅，此处从略。

4 讨论与小结

4.1 讨论

随机完全区组设计最适合用于区组因素对定量结果具有不可忽视的影响的试验研究场合，在实际的试验研究中，有时可能同时存在多个重要的非试验因素，在设计试验时，可以将它们复合成一个区组因素。

从方差分析的角度来看，方差分析方法对定量资料的前提条件要求很苛刻，无论试验研究中涉及多少个因素，要求每个因素都必须满足“独立性”“正态性”和“方差齐性”三个前提条件。值得注意的是，独立性是针对整个试验资料中任何两个定量数据而言的，即任何两个定量数据之间是互相独立的；正态性是针对任何一个因素的某一个水平而言的，即该因素每个特定水平下定量资料应服从正态分布；而方差齐性则是针对任何一个因素的全部水平而言的，即该因素各水平下总体方差应相等。仅当前述提及的三个前提条件都满足时，方差分析的结果才是正确的。否则，建议采用混合效应模型分析方法处理资料[4]。

在例 2 的两两比较的分析结果中，SNK 法与 TUKEY 法给出的结果略有不同，其原因在于这两种方法控制的误差类型不同[7-8]。相对来说，TUKEY 法给出的结果可信度更高。

4.2 小结

本文概述了随机完全区组设计的要点，介绍了随机完全区组设计定量资料的方差分析模型和计算公式，借助 SAS 软件对两个实例进行了分析，对输出结果作出了解释，并给出了统计结论和专业结论。

参考文献

[1] Dean A, Voss D. Design and analysis of experiments[M]. 北京：世界图书出版公司, 2010：295-338.

[2] 徐勇勇. 中华医学统计百科全书 医学研究与临床统计设计分册[M]. 北京：中国统计出版社, 2013：37-38.

[3] 茆诗松. 统计手册[M]. 北京：科学出版社, 2003：438-443.

[4] Littell RC, Milliken GA, Stroup WW, et al. SAS system for mixed models[M]. Cary, NC: SAS institute Inc, 1996：2-5.

[5] 方积乾. 卫生统计学[M]. 7 版. 北京：人民卫生出版社, 2012：132-134.

[6] 胡良平. 科研设计与统计分析[M]. 北京：军事医学科学出版社, 2012：230-232.

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

[8] 胡纯严, 胡良平. 如何正确运用方差分析：多个均值之间的多重比较[J]. 四川精神卫生, 2022, 35(1)：21-25.

（收稿日期：2022-03-10）

（本文编辑：戴浩然）

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。