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

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】 本文目的是介绍单因素多水平设计定量资料一元方差分析的前提条件、基本思想、计算公式和SAS实现。前提条件包括独立性、正态性和方差齐性；基本思想的核心是对总离均差平方和的分解；检验统计量F由组间均方除以组内（或称为误差）均方构造而成。方差分析结果是关于某因素全部水平下均值之间差异情况的一个概括性评价，当得出该因素全部均值之间的差异有统计学意义时，需要采取特定的方法对该因素的多个均值进行多重比较。本文借助SAS软件，对两个实例进行方差分析，并采用三种方法对其中一个实例的多个均值之间进行多重比较。

【关键词】 前提条件；离均差平方和；方差分析；多重比较；F分布

中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20220110006

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

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 prerequisites, basic ideas, calculation formulas and the SAS implementation of a single-factor multi-level design quantitative data univariate analysis of variance. The prerequisites included the independence, normality and homogeneity of variance. The core of the basic idea was the decomposition of the sum of squares of the total deviations for the mean. The test statistic F was constructed through the between-group mean square divided by the within-group (or called error) mean square. The result of analysis of variance was a general evaluation of the difference among all means of a factor with the whole levels. When it was found that the difference among all means of the factor was statistically significant, a specific approach needed to be adopted for the multiple comparisons about the multiple means of the factor. With the help of the SAS software, the paper performed the analysis of variances for two examples, and used three approaches to make the multiple comparisons among all means of a factor in one of the examples.

【Keywords】 Prerequisites; Sum of squared deviations from the mean; Analysis of variance; Multiple comparisons; F distribution

单因素多水平设计是生物医学试验研究中使用频率极高的一种设计类型。当观测结果变量为定量变量时，常选用的统计分析方法被简称为“单因素方差分析”。本文着重介绍该设计定量资料一元方差分析的前提条件、基本思想、计算公式和基于SAS软件的实例分析。

1 单因素多水平设计定量资料一元方差分析

1.1 前提条件

第一个前提条件为“独立性”，即全部定量数据中的任何两个数据之间必须相互独立[1]；第二个前提条件为“正态性”，即某因素各水平组定量数据必须抽自正态分布的总体（需要分组进行正态性检验）[2]；第三个前提条件为“方差齐性”，即某因素k个水平组定量数据应抽自方差相等的k个总体（需要对定量资料中每个因素所有水平组的总体方差进行方差齐性检验）[3]。

1.2 方差分析的基本思想与计算公式

单因素多水平设计定量资料一元方差分析的基本思想是关于总离均差平方和的分解，即将全部数据关于总均值的离差平方和分解成组间离均差平方和与组内（或称误差）离均差平方和两部分，自

由度也有类似的分解方法。将各部分离均差平方和除以各自的自由度，就是各项的方差（或称均方）。以组内（或误差）均方为分母，以组间均方为分子，就可以构造出一个检验统计量 F。

对于单因素多水平设计一元定量资料而言，其总离均差平方和 $SS_{总}$ 可按下式分解[4]：

$$SS_{总}=SS_{组间}+SS_{误差} \quad (1)$$

式（1）中，三项离均差平方和的表达式如下：

$$SS_{总}=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij}-\bar{Y}_{..}\right)^2 \quad (2)$$

$$SS_{组间}=\sum_{j=1}^{k}n_j\left(\bar{Y}_{.j}-\bar{Y}_{..}\right)^2 \quad (3)$$

$$SS_{误差}=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij}-\bar{Y}_{..}\right)^2 \quad (4)$$

$$df_{组间}=组数-1,\ df_{误差}=N-1-df_{组间} \quad (5)$$

基于离均差平方和与自由度构造均方 MS，见式（6）、式（7）：

$$MS_{组间}=\frac{SS_{组间}}{df_{组间}} \quad (6)$$

$$MS_{误差}=\frac{SS_{误差}}{df_{误差}} \quad (7)$$

基于均方构造检验统计量 F，见式（8）：

$$F=\frac{MS_{组间}}{MS_{误差}} \quad (8)$$

在式（8）中，F 服从分子自由度为 $df_{组间}$、分母自由度为 $df_{误差}$ 的 F 分布。

若采用手工计算，需要通过查 F 界值表（单侧检验），可得 $F_{(1-\alpha)(df组间,\ df误差)}$，若 $F\geqslant F_{(1-\alpha)(df组间,\ df误差)}$，则 $P\leqslant\alpha$，反之，则 $P>\alpha$。最后，确定 P 值并作出统计推断，再结合专业知识给出专业结论。

2 实例与SAS实现

2.1 问题与数据结构

【例 1】根据胆囊纤维化患者胰腺功能（胰蛋白酶分泌量），研究者将患者分为 A、B、C 三组：A 组胰蛋白酶分泌量≤50 U/kg·h^{-1}；B 组胰蛋白酶分泌量为 51~1000 U/kg·h^{-1}；C 组胰蛋白酶分泌量>1000 U/kg·h^{-1}。三组的样本含量分别为 9、10、9 例，测得每位受试对象的蛋白质浓度（mg/mL）如下。A 组：1.7、2.0、2.0、2.2、4.0、4.0、5.0、6.7、7.8；B 组：1.4、2.4、2.4、3.3、4.4、4.7、6.7、7.6、9.5、11.7；C 组：2.9、3.8、4.4、4.7、5.0、5.6、7.4、9.4、10.3[5]。分析三组患者蛋白质浓度平均值之间的差异是否有统计学意义？

【例 2】为研究钙离子对体重的影响，某研究者将 36 只肥胖模型大白鼠随机等分为三组，每组 12 只，分别给予常规剂量钙（0.5%）、中剂量钙（1.0%）和高剂量钙（1.5%）三种不同的高脂饲料，喂养 9 周，测量并计算其喂养前后体重的差值[2]。分析三种不同剂量钙作用下大白鼠体重改变量的均值是否相等？

2.2 对例 1 资料的分析与解答

【分析与解答】这是一个单因素三水平设计一元定量资料，设所需要的 SAS 程序如下：

```
data a;
do i=1 to 10;
do group=1 to 3;
input y @@;
output;
end;
end;
cards;
1.7  1.4  2.9
（此处省略部分数据，见前文）
7.8  9.5  10.3
;
run;
proc glm data=a;
class group;
model y=group/ss3;
means group;
run;
```

【SAS 程序说明】当各组样本含量不等时，以样本含量最多的组为基准，样本含量少的其他组缺少几个数据就用几个“点”填充。

源	自由度	平方和	均方	F	Pr>F
模型	2	19.6273492	9.8136746	1.26	0.2998
误差	25	193.9912222	7.7596489		
校正合计	27	213.6185714			
R 方	变异系数	均方根误差	y 均值		
0.091880	54.54350	2.785615	5.107143		
源	自由度	III 型 SS	均方	F	Pr>F
group	2	19.62734921	9.81367460	1.26	0.2998

【SAS输出结果及解释】

以上输出结果表明：三组患者蛋白质浓度的均值差异无统计学意义（因 F=1.26，df=2，P=0.2998>0.05），故可以认为不同胰蛋白酶分泌量对蛋白质浓度的影响不明显。三组定量资料的箱图见图1。

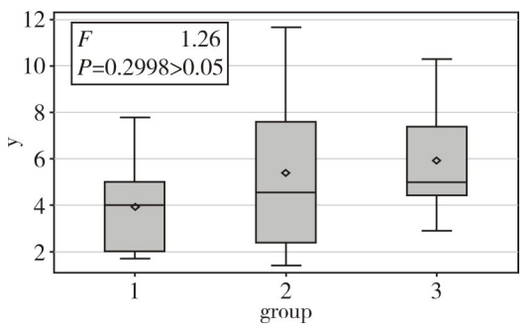

图1 三组定量资料的箱图

图1中，从左到右共有3个长方形，每个长方形中的横线代表各组定量资料的中位数所在的位置；每个长方形中的小正方形代表各组定量资料的平均值所在的位置，各组具体的平均值和标准差如下：

“group”的水平	数目	y	
		均值	标准差
1	9	3.93333333	2.21415898
2	10	5.41000000	3.38179768
3	9	5.94444444	2.54563897

2.3 对例2资料的分析与解答

【分析与解答】这是一个单因素三水平设计一元定量资料，设所需要的SAS程序如下：

```
data a;
do i=1 to 12;
do group=1 to 3;
input y @@;
output;
end;
end;
cards;
332.96 253.21 232.55
297.64 235.87 217.71
312.57 2 69.30 216.15
295.47 258.90 220.72
284.25 254.39 219.46
307.97 200.87 247.47
292.12 227.79 280.24
244.61 237.05 196.01
261.46 216.85 208.24
286.46 238.03 198.41
322.49 238.19 240.35
282.42 243.49 219.56
;
run;
proc glm data=a;
class group;
model y=group/ss3;
means group;
means group/lsd tukey snk;
run;
```

【SAS程序说明】第2个“means语句”给出了三个选择项，分别代表三种多重比较的方法，在GLM过程中，类似的比较方法还有很多，此处从略。

【SAS输出结果及解释】

源	自由度	平方和	均方	F	Pr>F
模型	2	31320.13272	15660.06636	31.49	<0.0001
误差	33	16410.01756	497.27326		
校正合计	35	47730.15028			

R方	变异系数	均方根误差	y均值
0.656192	8.830340	22.29962	252.5342

源	自由度	III型SS	均方	F	Pr>F
group	2	31320.13272	15660.06636	31.49	<0.0001

以上输出结果表明：三组大白鼠体重改变量的均值差异有统计学意义（因 F=31.49，df=2，P<0.0001），故可以认为饲料中钙剂量不同，大白鼠体重增加量也不同。饲料中钙剂量越高，大白鼠体重增加量越少。饲料中钙剂量由低到高对应的三组大白鼠体重增加量的箱图见图2。

在图2中，从左到右有3个长方形，每个长方形中的横线代表各组定量资料的中位数所在的位置；

每个长方形中的小正方形代表各组定量资料的平均值所在的位置，各组具体的平均值和标准差如下：

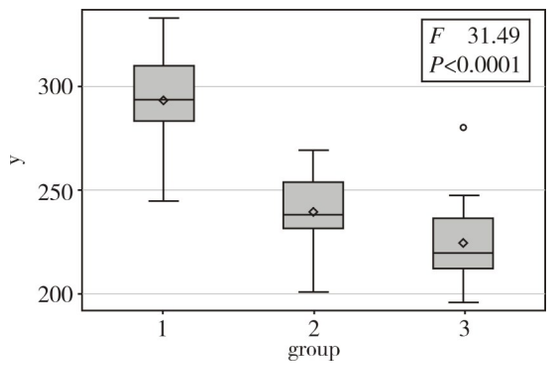

图2　饲料中钙剂量由低到高对应的三组大白鼠体重增加量的箱图

“group”的水平	数目	y 均值	y 标准差
1	12	293.368333	24.6206834
2	12	239.495000	18.7215867
3	12	224.739167	23.1331778

t Tests (LSD) for y	
Alpha	0.05
Error Degrees of Freedom	33
Error Mean Square	497.2733
Critical Value of t	2.03452
Least Significant Difference	18.522

注：此检验控制Ⅰ型比较误差率，不是试验误差率。

基于LSD法（即成组设计一元定量资料 t 检验）进行均值之间两两比较的结果见图3。

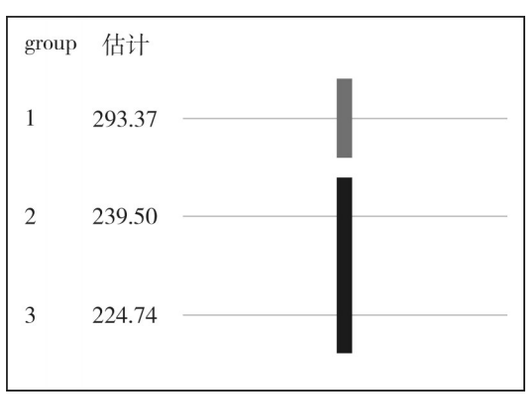

图3　基于LSD法进行均值之间两两比较的结果

图3中，第1组（$\overline{Y_1}$=293.37）与第2、3两组（$\overline{Y_2}$=239.50、$\overline{Y_3}$=224.74）之间的差异有统计学意义，而第2组与第3组之间差异无统计学意义。

Student-Newman-Keuls Test for y		
Alpha	0.05	
Error Degrees of Freedom	33	
Error Mean Square	497.2733	
Number of Means	2	3
Critical Range	18.521522	22.338804

注：此检验控制Ⅰ型试验误差率，对应的假设为完全无效假设，但不是部分无效假设。

基于SNK法进行均值之间两两比较的结果见图4。

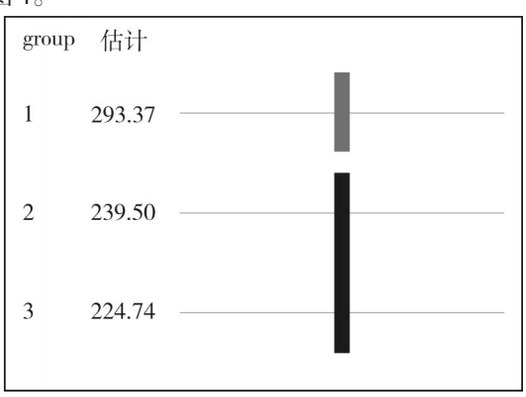

图4　基于SNK法进行均值之间两两比较的结果

对输出结果的解释和结论同上（见图3后面的解释），此处从略。

Tukey's Studentized Range (HSD) Test for y	
Alpha	0.05
Error Degrees of Freedom	33
Error Mean Square	497.2733
Critical Value of Studentized Range	3.47019
Minimum Significant Difference	22.339

注：此检验控制Ⅰ型试验误差率，但一般来说，此法比REGWQ法具有更高的Ⅱ型误差率。

基于TUKEY法进行均值之间两两比较的结果见图5。

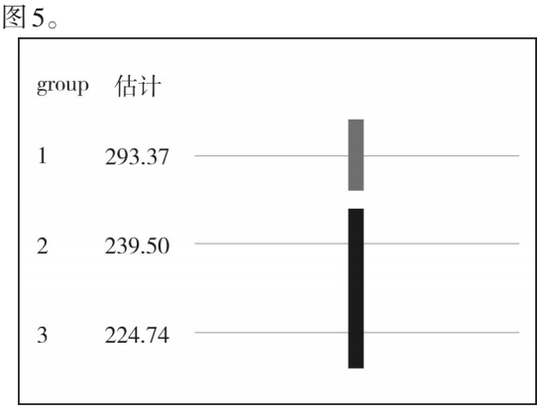

图5　基于TUKEY法进行均值之间两两比较的结果

对输出结果的解释和结论同上(见图3后面的解释),此处从略。

3 讨论与小结

3.1 讨论

严格地说,用多次t检验取代方差分析的做法欠妥。事实上,当采用多次t检验进行$n(n\geq3)$个均值两两比较时,可知比较的次数有$c=n!/2!(n-2)!$。比较的次数越多,在无效假设为真时,拒绝无效假设的Ⅰ类错误概率也越大。设每次检验水准即犯Ⅰ类错误概率为α,累积Ⅰ类错误的概率为α',则对多个均值进行c次检验时,根据概率乘法原理,其累积Ⅰ类错误概率与c的关系见式(9)[6]:

$$\alpha'=1-(1-\alpha)^c \quad (9)$$

例如,设$\alpha=0.05$,$c=3$,其累积的Ⅰ类错误概率为:$\alpha'=1-(1-0.05)^3=0.0143$。可见,用多次$t$检验取代方差分析,意味着Ⅰ类错误概率会增大,即出现假阳性的可能性会增加。因此,不应该用t检验取代方差分析。

一般来说,对单因素多水平设计定量资料进行一元方差分析后,若结论是各水平组的均值差异有统计学意义,这是一个概括性的结论,它并不意味着任何两个平均值之间的差异都有统计学意义。欲知详情,应对多个均值进行多重比较。然而,对多个均值进行两两比较的方法很多,其区别是不同方法控制的误差类型不同。详见文献[7-8]。

进行方差分析前,需检查定量资料是否满足三个前提条件。因篇幅所限,本文在分析实例时,假定资料满足方差分析所需要的前提条件。在实际应用中,应严格检查给定资料是否满足前提条件。否则,方差分析的结果可能不准确。

3.2 小结

本文介绍了与单因素多水平设计定量资料一元方差分析有关的主要内容,包括前提条件、基本思想和计算公式。借助SAS软件对两个实例进行了方差分析,还采用三种两两比较的方法(即LSD法、SNK法和TUKEY法)对例2中的三个均值进行了分析。最后,在讨论中阐明了不适合采用多次t检验取代方差分析的理由。

参考文献

[1] 胡良平, 李子建. 医学统计学基础与典型错误辨析[M]. 北京: 军事医学科学出版社, 2003: 164.
[2] 方积乾. 卫生统计学[M]. 7版. 北京: 人民卫生出版社, 2012: 129-139.
[3] 颜虹. 医学统计学[M]. 北京: 人民卫生出版社, 2005: 133-150.
[4] 黄志宏, 方积乾. 数理统计方法[M]. 北京: 人民卫生出版社, 1987: 141-151.
[5] 伯纳德·罗斯纳. 生物统计学基础[M]. 孙尚拱, 译. 北京: 科学出版社, 2004: 545.
[6] 胡良平. 面向问题的统计学: (1)科研设计与统计基础[M]. 北京: 人民卫生出版社, 2012: 407-420.
[7] 杨树勤. 中国医学百科全书: 医学统计学[M]. 上海: 上海科学技术出版社, 1985: 108-112.
[8] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 3957-4142.

(收稿日期:2022-01-10)
(本文编辑:陈 霞)