如何正确运用方差分析——多个均值之间的多重比较

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】 本文目的是介绍多个均值之间的多重比较方法与SAS实现。多重比较方法可细分为成对比较、所有处理组与一个对照组比较、将每个处理组平均值与全部组平均值进行比较、近似和基于模拟的方法、多阶段检验和贝叶斯方法。除贝叶斯方法外，其他多重比较方法之间的区别在于控制的误差类型不同。误差类型大体可分为以下三类：比较误差率、试验误差率和最大试验误差率。基于控制不同误差率所构造出来的多重比较方法，在得出结论的推理强度上是不相同的。本文使用SAS软件对实例进行分析，并对输出结果作出解释。

【关键词】 方差分析；多重比较；比较误差率；试验误差率；极差分布
中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20220110007

How to use analysis of variance correctly——the multiple comparisons among the multiple means

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 the paper was to introduce the multiple comparison method among multiple means and the SAS implementation. The multiple comparison approaches could be subdivided into the pairwise comparisons, the comparisons of all treatment groups with a control group, the comparisons of the mean of each treatment group with the average of all groups, the approximate and simulation-based approach, the multi-stage testing and Bayesian method. Except for the Bayesian approach, the difference between other multiple comparison methods lied in the types of error that were controlled. Error types could be roughly divided into the following three categories, the comparisonwise error rate, the experimentwise error rate and the maximum experimentwise error rate. The multiple comparison methods constructed based on the control of different error rates were not all the same in the strength of inference to draw conclusions. This paper used the SAS software to analyze the examples and explained the output results.

【Keywords】 Analysis of variance; Multiple comparisons; Comparisonwise error rate; Experimentwise error rate; Range distribution

基于均值比较的方差分析的结果是对定量资料中每个试验因素各水平下均值之间是否存在差异的一个概括性结论，若某因素各水平下均值之间的差异无统计学意义，就不需要对该因素各水平下均值做进一步比较了；反之，则需要进行多重比较。本文针对多重比较问题进行探讨，阐释各种多重比较方法之间的异同点以及使用时的指导思想和参考建议。

1 多重比较

1.1 概述

当比较两个以上的平均值时，方差分析（或称F检验）会反映这些平均值之间的差异是否有统计学意义，但它不会反映哪些平均值与其他平均值不同。多重比较的目的是比较三种及以上“处理”（例如不同药物、不同受试者）的平均效应，以确定哪些处理更好、哪些更差，以及比较处理程度，同时控制做出错误决定的概率。SAS/STAT的GLM过程中的MEANS和LSMEANS语句提供多种多重比较的方法[1]。

多重比较过程可以通过两种方式进行分类：根据它们所做的比较和它们所提供的推理强度。根据所做的比较，GLM过程提供了两种类型[1-2]：①所有平均值对之间的比较；②对照与其他所有方法之

间的比较。推理的强度表示当一个检验有统计学意义时，可推断出的关于平均值结构的内容；它与多重比较过程控制的错误率类型有关。GLM过程中可用的多重比较过程按从弱到强的顺序提供以下推理类型之一：①单次比较，均值之间的差异，未针对多次比较进行调整；②不均匀性，表示均值彼此不同；③不等，哪些均值之间是不同的；④区间，展示均值差异的联合置信区间。

在PROC GLM过程中，用两张表给出了可用于所有成对比较、所有处理组与对照组比较的多重比较过程，以及MEANS和LSMEANS语句中关于各种多重比较方法的选项[1]。因篇幅所限，此处从略。

1.2 成对比较

本节讨论的所有方法均取决于标准化成对差异，见式(1)：

$$t_{ij}=\left(\bar{y}_i-\bar{y}_j\right)/\hat{\sigma}_{ij} \tag{1}$$

在式(1)中，i和j是两组的标记，$\bar{y}_i$和$\bar{y}_j$是i组和j组的平均值或LSMEANS(即最小平方平均值)，$\hat{\sigma}_{ij}$是$\bar{y}_i-\bar{y}_j$估计方差的平方根。

对于简单算术平均值、加权算术平均值和由参数估计量$1_i'b$和$1_j'b$的线性组合定义的最小平方平均值，其对应的$\hat{\sigma}_{ij}$计算公式分别见式(2)、式(3)、式(4)：

$$\hat{\sigma}_{ij}=s^2\left(1/n_i+1/n_j\right) \tag{2}$$

$$\hat{\sigma}_{ij}=s^2\left(1/w_i+1/w_j\right) \tag{3}$$

$$\hat{\sigma}_{ij}=s^2 1_i'(X'X)^{-}1_j' \tag{4}$$

式(2)中，n_i和n_j分别是i组和j组的样本大小，s^2是误差的均方(或方差)，具有自由度v。式(3)中，w_i和w_j分别是i组和j组中的权重之和。

此外，所有的方法都根据形式$|t_{ij}|\geq c(\alpha)$的假设检验进行了讨论。其中，$c(\alpha)$是一些常数，取决于事先确定的显著性水平α。这样的检验可以转变成为置信区间的形式，见式(5)：

$$\left(\bar{y}_i-\bar{y}_j\right)-\hat{\sigma}_{ij}c(\alpha)\leqslant\mu_i-\mu_j\leqslant\left(\bar{y}_i-\bar{y}_j\right)+\hat{\sigma}_{ij}c(\alpha) \tag{5}$$

在多重比较的过程中，涉及以下基本概念：完全零假设(即所有总体平均值相等)下的试验误差率和部分零假设(即某些平均值相等，但其他平均值不等)下的试验误差率。常用的缩略语有：单次比较误差率(CER)、完全零假设下的试验误差率(EERC)以及在任何完全或部分零假设下的最大试验错误率(MEER)。这些误差率与不同的推理强度有关：单次检验控制CER，均值不均匀性检验控制EERC，产生置信区间的检验控制MEER。初步F检验控制EERC，但不控制MEER。

通过将CER设置为足够小的值，可以在α水准上控制MEER。Bonferroni不等式已广泛用于此目的。如果CER=α/c(c为比较的总次数)，那么MEER就小于α。如果有下式成立，则具有MEER<α的Bonferroni t检验(MEANS语句中的BON选项，LSMEANS语句中的ADJUST=BON)就表明两均值之间的差异有统计学意义：

$$|t_{ij}|\geq t(\varepsilon;\ \nu) \tag{6}$$

在式(6)中，$\varepsilon=\dfrac{2\alpha}{k(k-1)}$，$k$代表参与比较的均值个数。

一个更严格的界限由式(7)给出：

$$\mathrm{CER}=1-(1-\alpha)^{1/c} \tag{7}$$

对于任何一组c次比较，同样可以保证MEER<α。因此，由SIDAK选项提供的Sidak t检验由式(6)给出，但ε由式(8)给出：

$$\varepsilon=1-(1-\alpha)^{\frac{2}{k(k-1)}} \tag{8}$$

统计学家还提出了许多其他“成对比较”方法[1]，因篇幅所限，此处从略。

1.3 所有处理组与一个对照组比较

平均值比较的一种特殊情况是，需要检验的唯一比较是一组新处理和一个单一对照之间的比较[3]。在这种情况下，可以通过使用仅限于检验与单个控制平均值比较的方法来获得更好的功效。Dunnett针对这种情况提出了一种检验，如果下式成立，该检验表明所考察的平均值与对照组平均值之间的差异有统计学意义：

$$|t_{i0}|\geq d\left(1-\alpha;\ k,\nu,\rho_1,\cdots,\rho_{k-1}\right) \tag{9}$$

其中，$\bar{y}_0$是对照组平均值[参见式(1)]，$d(1-\alpha,k,\nu,\rho_1,\cdots,\rho_{k-1})$是$k$个平均值与对照组平均值比较且具有自由度为$v$、相关系数为$\rho_1,\cdots,\rho_{k-1}\left[\rho_i=n_i/\left(n_0+n_i\right)\right]$的“多对一比较的$t$统计量”的临界值。出现相关项是因为每个处理组都与同一对照组进行比较。Dunnett的检验将MEER保持在不超过规定的水平α。

1.4 将每个处理组均值与全部组平均值进行比较

平均值分析(ANOM)是一种比较组平均值并以图形方式显示比较结果的方法[1]。如果某组的均值与总体平均值差异有统计学意义，则判断均值不同，并根据多次比较调整显著性水平。总平均值作

为LSMEANS的加权平均值计算，权重与方差成反比。如果在LSMEANS语句中使用PDIFF=ANOM选项，则该方法将显示用于检验每个LSMEANS和平均LSMEANS之间差异的P值（默认情况下，针对多次比较进行了调整）。SAS/QC软件中的ANOM过程显示表格和图形，用于分析各种响应类型的平均值。对于单因素设计，PDIFF=ANOM比较的置信区间等同于PROC ANOVA的结果。不同之处在于，PROC GLM直接显示差异的置信区间，而PROC ANOVA的图形输出将其显示为总体平均值周围的决策界限。

1.5 近似和基于模拟的方法

Tukey、Dunnett和Nelson的检验都基于相同的一般分位数计算[1]：

$$q'(1-\alpha,\nu,R)=q\in P[\max(|t_1|,\cdots,|t_n|)>q]=\alpha \quad (10)$$

在式（10）中，t_i（i=1，2，…，n）服从自由度为v、相关系数矩阵为R的联合多元t分布。一般来说，评估$q'(1-\alpha,v,R)$需要对（n+1）重积分进行重复的数值计算，这通常是很难解决的。但在Tukey检验中，当R具有一定的对称性时，问题会简化为可行的2重积分，在Dunnett和Nelson检验中，则会简化为因子分析结构。在以下两种情况下，R矩阵具有精确计算Tukey检验所需的对称性：①t_i是具有相同方差的k个不相关均值形成的$k(k-1)/2$对均值之间的学生化差量；②t_i是方差平衡设计（例如平衡不完全区组设计）中k个LSMEANS形成的$k(k-1)/2$对均值之间的学生化差量。

1.6 多阶段检验

可以使用到目前为止讨论的所有方法来获得同时的置信区间。通过牺牲同步估计功能，使用多阶段检验（MST）获得更大功效的同步检验[4-5]。MST有上升和下降两种类型。SAS/STAT软件中提供了使用更广泛的下降方法[1]。逐步下降MST首先在一个水平γ_k上检验所有平均值的均匀性。如果检验结果为拒绝，则$k-1$个平均值的每个子集都在一个水平γ_{k-1}进行检验；否则，程序将停止。一般来说，如果一组p个均值的同质性假设在该水平γ_p上被拒绝，则在该水平γ_{k-1}上检验$p-1$个均值的每个子集；否则，p个均值集被认为差异无统计学意义，且其子集均不进行检验。已提出的多种MST在子集检验所依据的水平γ_p和统计量上有所不同。显然，下降MST的EERC不大于γ_k，CER不大于γ_2，但MEER是γ_p（p=2，…，k）的一个复杂函数。

对于不相等的单元格大小，PROC GLM使用单元格大小的调和平均值作为公共样本大小。然而，由于产生的运行特性可能不理想，建议仅在平衡情况下使用MST。当样本大小相等时，使用极差统计量可以按升序或降序排列均值，并仅检验连续子集。但如果指定F统计量，则无法使用此快捷方式。因此，仅实施基于极差的MST。通常情况下，报告MST结果的方法是按这样的顺序书写平均值，并绘制平行于齐次子集平均值列表的线。这种表示形式也便于在单元格大小相同的情况下进行成对比较。

最著名的MST是Duncan（Duncan选项）和Student-Newman-Keuls（SNK选项）方法。Duncan的方法见式（11），SNK方法见式（12）。

$$\gamma_p=1-(1-\alpha)^{p-1} \quad (11)$$

$$\gamma_p=\alpha \quad (12)$$

统计学家还提出了一些其他“多阶段检验”方法，因篇幅所限，此处从略。

1.7 贝叶斯方法

在加性损失下最小化贝叶斯风险，而不是控制Ⅰ型错误率[1,6]。对于每对总体均值μ_i和μ_j，定义了零假设(H_0^{ij})和备择假设(H_a^{ij})，见式（13）、式（14）：

$$H_0^{ij}:\mu_i-\mu_j\leqslant 0 \quad (13)$$

$$H_a^{ij}:\mu_i-\mu_j>0 \quad (14)$$

对于任何(i,j)对，设d_0表示有利于H_0^{ij}的决策；设d_a表示有利于H_a^{ij}的决策；设$\delta=\mu_i-\mu_j$。与(i,j)对的决策对应的损失函数见式（15）、式（16）：

$$L(\delta_0|\delta)=\begin{cases}0 & \delta\leqslant 0\\ \delta & \delta>0\end{cases} \quad (15)$$

$$L(\delta_a|\delta)=\begin{cases}-k\delta & \delta\leqslant 0\\ \delta & \delta>0\end{cases} \quad (16)$$

在式（16）中，k表示指定的常数，而不是平均值的个数。涉及所有平均值对的联合决策的损失是每个单独决策的损失之和。假设总体均值具有未知方差的正态先验分布，均方差的对数具有一致的先验分布。对于(i,j)对，如果下式成立，则拒绝零假设：

$$\bar{y}_i-\bar{y}_j\geqslant t_B s\sqrt{\frac{2}{n}} \quad (17)$$

在式（17）中，t_B是贝叶斯t值，取决于k、单因素方差分析的F统计量和F的自由度。t_B的值是F的递减函数，因此，随着F的增加，Waller-Duncan检验（由Waller选项指定）变得更自由。

2 实例与SAS实现

2.1 问题与数据结构

【例1】为了研究某种降血脂新药的临床疗效，按统一纳入标准选择120例高血脂患者，采用完全随机设计方法将患者等分为四组（A组：安慰剂组；B组：2.4 g组；C组：4.8 g组；D组：7.2 g组），每组30例，进行双盲法试验。6周后检测患者低密度脂蛋白含量（单位：mmol/L）作为定量试验结果，具体数据见后面的SAS程序（此处从略）[4]。问四个药物组患者的低密度脂蛋白含量总体均值之间的差异是否有统计学意义？

2.2 分析与解答

【分析与解答】本例属于单因素四水平设计一元定量资料，可采用相应设计定量资料的方差分析；若四组均值之间的差异有统计学意义，还需要对四个均值进行多重比较。设所需要的SAS程序如下：

```
data a;
input group$ n @@;
do i=1 to n; input y @@; output; end;
cards;
A 30
3.53 4.59 4.34 2.66 3.59 3.13 2.64 2.56
3.50 3.25 3.30 4.04 3.53 3.56 3.85 4.07
3.52 3.93 4.19 2.96 1.37 3.93 2.33 2.98
4.00 3.55 2.96 4.30 4.16 2.59
（其他三组数据从略，详见文献[4]）
;
run;
proc glm data=a;
class group;
model y=group/ss3;
means group;
means group/SNK;
means group/WALLER;
run;
proc glm data=a;
class group;
model y=group/ss3;
means group/DUNNETT ('A');
run;
```

【SAS程序说明】第1个过程步的作用是进行四组间两两比较；第2个过程步的作用是以“A组”为对照组，其他组都与A组比较。

【SAS输出结果及解释】

变异来源	自由度	平方和	均方	F	Pr>F
模型	3	32.15603000	10.71867667	24.88	<0.0001
误差	116	49.96702000	0.43075017		
校正合计	119	82.12305000			

源	自由度	III型SS	均方	F	Pr>F
group	3	32.15603000	10.71867667	24.88	<0.0001

以上结果表明：四个均值之间的差异有统计学意义。

基于SNK法的分析结果见图1。

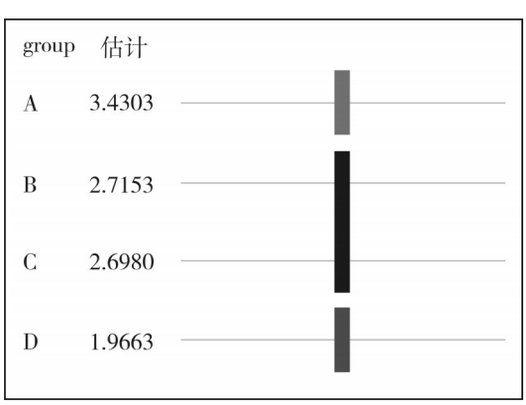

图1 基于SNK法的分析结果

由图1可看出，仅B组与C组均值之间差异无统计学意义，其他任何两组均值之间的差异均有统计学意义。WALLER法的分析结果与图1的结果相同，此处从略。

group比较	均值间差值	联合95%置信区间		
B−A	−0.7150	−1.1183	−0.3117	***
C−A	−0.7323	−1.1357	−0.3290	***
D−A	−1.4640	−1.8673	−1.0607	***

【说明】以上比较的显著性水平α=0.05。B、C、D组与A组均值比较的结果，差异均有统计学意义。

【专业结论】降血脂新药的三个剂量均能降低高血脂患者低密度脂蛋白含量；7.2 g剂量效果最好，2.4 g剂量与4.8 g剂量之间的差异不明显。

3　讨论与小结

3.1　讨论

多重比较是方差分析之后不可缺少的内容。然而，由于多重比较的方法非常多，特别是各方法控制的误差类型不同，导致结果的推论强度不同[1]。使用者在选择这些多重比较方法时，可参考SAS软件所给出的建议：如果对几个孤立的比较感兴趣，并且不关心多重推断的影响，可以重复使用 t 检验[7-8]或Fisher无保护LSD法；如果对所有成对比较或与对照组的所有比较感兴趣，则应分别使用Tukey检验和Dunnett检验，以便做出可能最强的推断；如果对推理要求较弱，特别是如果不需要均值差异的置信区间，则应使用REGWQ法。如果同意贝叶斯方法以及Waller和Duncan的假设，应该使用Waller-Duncan检验。

当各水平组样本含量不相等时，多次比较也会导致违反直觉的结果。例如，考虑A、B、C、D四个因素，以A>B>C>D为样本均值，A和D各有两个观测值，B和C各有10 000个观测值，B和C的差异可能有统计学意义，而A和D之间的差异则可能没有统计学意义。

3.2　小结

本文详细介绍了SAS/STAT的GLM过程中可以实现的所有多重比较方法，根据结果的推论强度，它们可以被划分成4类，共11种；另外，还有贝叶斯方法。借助SAS软件对一个实例进行了方差分析，并给出了采用SNK法、Waller法（即贝叶斯法）和Dunnett法进行多重比较的结果。

参考文献

[1] SAS Institute Inc. SAS/STAT®15. 1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 3957-4142.

[2] 胡良平. 现代统计学与SAS应用[M]. 北京：军事医学科学出版社，1996：146-151.

[3] 孙振球. 医学统计学[M]. 北京：人民卫生出版社，2002：64-67.

[4] 颜虹. 医学统计学[M]. 北京：人民卫生出版社，2005：54-56，141-143.

[5] 方积乾. 卫生统计学[M]. 7版. 北京：人民卫生出版社，2012：134-138.

[6] 伯杰. 统计决策论及贝叶斯分析[M]. 贾乃光，译. 北京：中国统计出版社，1998：8-37.

[7] 张洪璐，刘媛媛，李长平，等. 如何正确运用 t 检验：两算术均值比较一般差异性 t 检验及SAS实现[J]. 四川精神卫生，2020，33(3)：217-221.

[8] 于泽洋，刘媛媛，李长平，等. 如何正确运用 t 检验：两几何均值比较一般差异性 t 检验及SAS实现[J]. 四川精神卫生，2020，33(3)：222-225.

（收稿日期：2022-01-10）
（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名誉会长、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。