·循证医学中的医学统计学问题·

两两比较的Bonferroni法

伍小英[1,2]， 鲁婧婧[1]， 张晋昕[1]， 李 河[3]

(1. 中山大学公共卫生学院，广州 510080；2 广州市胸科医院医务科，广州 510095；
3. 广东省人民医院，广东省心血管病研究所流行病学研究室，广州 510080)

[摘要] 几组资料的平均水平被初步判定差异有统计学意义后，往往需进一步做两两比较。两两比较时，最常见的错误是将总的检验水准(如α=0.05)直接用于每一次的比较中。Bonferroni法通过修正每次比较的检验水准，控制Ⅰ类错误总的发生概率，是两两比较中最常用的方法之一。本文简要阐述了Bonferroni法的基本原理，并举例说明Bonferroni法的几种应用情况，旨在指导读者运用Bonferroni法正确地进行多组间的两两比较。

[关键词] 两两比较；Bonferroni法
[中图分类号] R195.1 [文献标识码] B [文章编号] 1671-5144(2006)06-0361-03

The Bonferroni Method for Multiple Comparisons

WU Xiao-ying[1,2], LU Jing-jing[1], ZHANG Jin-xin[1], LI He[3]
(1. School of Public Health, Sun Yat-sen University, Guangzhou 510080, China; 2. Department of Medical Management, Guangzhou Chest Hospital, Guangzhou 510095, China; 3. Department of Epidemiology, Guangdong Provincial Cardiovascular Institute, Guangdong Provincial People's Hospital, Guangzhou 510080, China)

Abstract: When the differences of average levels among several groups are statistically significant, the multiple comparisons should be carried out. Using $\alpha=0.05$ directly as significance level in comparison between each pair may result in too high type Ⅰ error probability. Bonferroni method is one of the most common methods for multiple comparisons, which could reduce the accumulated type Ⅰ error probability by adjusting α level. In this report, the fundamental principle of Bonferroni method is explained and examples are presented in some specific situations. This article is supposed to guide researchers to utilize Bonferroni method correctly in multiple comparisons.

Key words: multiple comparisons; Bonferroni method

医学科研中经常遇到比较多个总体水平是否相等的问题。当总的检验拒绝零假设H_0，认为各实验组总体水平不全相等时，研究者时常想回答哪些实验组间的总体水平不同，哪些实验组总体水平更高，哪些较低，这就需要进一步做两两比较。

两两比较中最常见的错误是在每一次比较中，直接沿用总的检验水准(如$\alpha=0.05$)，这样做的结果，是导致总的Ⅰ类错误增大。如果以$\alpha=0.05$为水准分别对m个实际上成立的零假设（即实际情形是H_0为真的时候）进行检验，不犯Ⅰ类错误的概率为$(1-\alpha)^m$，至少出现一次错误的概率(Ⅰ类错误的累积概率)为$1-(1-\alpha)^m$，显然大于0.05[若m=6，$1-(1-\alpha)^m=0.26$，这么大的Ⅰ类错误让人难以接受，且随着m增大，$1-(1-\alpha)^m$将更大]。那么能否通过控制α，降低Ⅰ类错误的累积概率呢？答案是肯定的。Bonferroni提出，设H_0为真，如果进行m次显著性水准为α的假设检验时，犯Ⅰ类错误的累积概率α不超过$m\alpha$，即有Bonferroni不等式$\alpha \leq m\alpha$成立[1,2]。所以令各次比较的显著性水准为$\alpha=0.05/m$，并规定$P\leq 0.05/m$时拒绝H_0，基于这样的做法，就

[作者简介] 伍小英(1974-)，女，广东广州人，主治医师，中山大学公共卫生学院在读MPH硕士研究生，主要研究方向为呼吸系统疾病的诊断与治疗。
[通讯作者] 张晋昕，Tel: 020-87330673-88；E-mail: zhjinx@mail.sysu.edu.cn

可以把Ⅰ类错误的累积概率控制在0.05。这种对检验水准进行修正的方法叫做Bonferroni调整(Bonferroni adjustment)法,简称Bonferroni法。下面举几个例子来说明Bonferroni法在不同假设检验应用场合的具体做法。

例1[3] 表1为随机抽样调查所得健康人和各期矽肺病人的血清粘蛋白含量(mg/dL),请问健康人和各期矽肺病人的血清粘蛋白含量(mg/dL)是否具有不同的水平?

表1 健康人与各期矽肺病人的血清粘蛋白含量 mg/dL

正常人	0期矽肺	Ⅰ期矽肺	Ⅱ期矽肺	Ⅲ期矽肺
64.26	52.01	64.44	74.97	77.11
42.84	67.33	69.63	88.06	82.48
42.48	60.40	69.73	93.47	83.43
48.19	78.91	74.97	94.10	89.01
80.22	74.68	80.44	100.67	97.48
69.61	84.68	80.44	101.14	103.81
48.19	81.14	94.20	113.42	107.10
48.90	94.82	96.39	118.98	178.42

该资料为完全随机设计,采用完全随机设计的单因素方差分析进行计算,计算结果F=8.135,P<0.001,差异有统计学意义,即可以认为正常人和各期矽肺病人的血清粘蛋白含量不全相同。由于研究者更关心的是各期矽肺病人和正常人的血清粘蛋白含量是否居于不同水平,于是进一步做两独立样本比较的t检验。各个分期的矽肺病人均与正常人比较,共需做4次检验,故按照前述思路应将显著性水准调整为$\alpha=0.05/4=0.012\ 5$。通过分析可以认为0期矽肺病人血清粘蛋白含量与正常人血清粘蛋白含量之间的差异无统计学意义,Ⅰ期、Ⅱ期和Ⅲ期矽肺病人的血清粘蛋白含量均高于正常人(见表2)。

表2 各期矽肺病人与正常人血清粘蛋白含量的两两比较

对比组	t值	P值	检验水准修正值α	检验结果
正常人与0期矽肺	-2.69	0.017 7	0.012 5	NS
正常人与Ⅰ期矽肺	-3.61	0.002 8	0.012 5	*
正常人与Ⅱ期矽肺	-6.09	<0.000 1	0.012 5	*
正常人与Ⅲ期矽肺	-3.74	0.002 2	0.012 5	*

注:表中"NS"表示差异无统计学意义;"*"表示P<0.012 5,差异有统计学意义。

由表2可以看到,两两比较中正常人与0期矽肺病人两组样本t检验的结果为P=0.017 7,若以α=0.05为检验水准,会拒绝H_0,但以α=0.012 5为检验水准,则不能拒绝H_0,结论是0期矽肺病人血清粘蛋白含量与正常人血清粘蛋白含量居于相同的水平。

例2[3] 急性病毒性心肌炎、原发性扩张型心肌病患者与正常人的白细胞介素(IL-1)数据如表3所示,请问三组人群的IL-1是否不同?

表3 正常人、急性病毒性心肌炎患者和原发性扩张型心肌病患者的IL-1水平 μg/L

正常人(A)	急性病毒性心肌炎患者(B)	原发性扩张型心肌病患者(C)
0.347	2.612	2.888
0.194	2.799	2.808
0.233	2.693	3.031
0.113	2.420	2.712
0.382	2.814	2.718
0.140	2.948	2.763
0.243	2.978	2.808
0.194	2.674	2.987
0.143	3.009	2.612
0.204	2.468	3.024
0.141	2.487	2.762
0.181	2.900	3.009
0.246	2.798	
0.136		
0.141		

由于三组观察值方差不齐,故不能直接运用方差分析进行计算。应该采用成组设计多个样本比较的秩和检验(Kruskal-Wallis)进行统计分析,得H=28.185 1,P<0.000 1,按α=0.05水准拒绝H_0,可以认为正常人、急性病毒性心肌炎患者和原发性扩张型心肌病患者的IL-1水平不全相同。

要明确究竟是哪些组间的差异有统计学意义还需要进行两两比较。这里有三组人群,设可以组合出的比较次数为m,应进行$m=C_3^2=3$次对比。选定$\alpha=(0.05/m)=(0.05/3)=0.017$,也就是说只有当P值小于0.017时才能拒绝无效假设。分别用两独立样本的秩和检验进行计算,结果可认为急性病毒性心肌炎患者和原发性扩张型心肌病患者IL-1水平均高于正常人,但急性病毒性心肌炎患者与原发性扩张型心肌病患者IL-1水平之间的差异无统计学意义(见表4)。

例3[4] 为探讨幽门螺旋杆菌(helicobacter pylori,Hp)感染与血型的关系,随机选择行胃镜检查的239例胃、十二指肠疾病患者,测定其血型及Hp感染情况(见表5),问血型与Hp感染有无关系?

表 4 例 2 的组间两两比较的结果

对比组	T 值	P 值	检验水准修正值 α	检验结果
A 组与 B 组	286	<0.000 1	0.017	*
A 组与 C 组	258	<0.000 1	0.017	*
B 组与 C 组	181	0.194 9	0.017	NS

注:表中“NS”表示差异无统计学意义;“*”表示 P<0.017,差异有统计学意义。

表 5 不同血型的胃、十二指肠疾病患者 Hp 感染率比较

血型	Hp 阳性	Hp 阴性	合计	Hp 感染率
A	28	19	47	0.60
B	38	28	66	0.58
O	95	11	106	0.90
AB	10	10	20	0.50
合计	171	68	239	0.72

该资料为 R×C 列联表,选用 χ^2 检验对四组的阳性率进行比较,经计算 χ^2=31.212 5, P<0.000 1,按 α=0.05 水准拒绝 H_0,可认为血型与幽门螺旋杆菌(Hp)感染有关。要具体回答不同血型患者 Hp 感染率间的关系,就可用 Bonferroni 方法来进行两两比较。这里有 4 个处理组,设可以组合出的比较次数为 m,应进行 $m=C_4^2=6$ 次对比。修正检验水准为 $\alpha=(0.05/m)=(0.05/6)=0.008\ 3$,分别对各对比组做两独立样本阳性率比较的 χ^2 检验,结果表明,“A 型”与“O 型”、“B 型”与“O 型”、“AB 型”与“O 型”的 Hp 感染率均不相同,而“A 型”、“B 型”和“AB 型”两两之间 Hp 感染率的差异无统计学意义(见表 6)。结合原数据可知,“O 型”血的 Hp 感染率高于“A 型”、“B 型”和“AB 型”组。

表 6 例 4 的两两组间比较的结果

对比组	χ^2 值	P 值	检验水准修正值 α	检验结果
“A 型”与“B 型”	0.045 1	0.831 7	0.008 3	NS
“A 型”与“O 型”	18.651 1	<0.000 1	0.008 3	*
“A 型”与“AB 型”	0.523 9	0.469 2	0.008 3	NS
“B 型”与“O 型”	23.825 0	<0.000 1	0.008 3	*
“B 型”与“AB 型”	0.357 2	0.550 1	0.008 3	NS
“O 型”与“AB 型”	19.018 9	<0.000 1	0.008 3	*

注:表中“NS”表示差异无统计学意义;“*”表示 P<0.008 3,差异有统计学意义。

例 4[5] 盐酸洛美沙星片剂在三种不同溶液中及不同 pH 值下溶解度(%)的实验数据列于表 7 中。在每一种溶液中溶解度与 pH 值之间有良好的直线关系,分析不同溶液中的溶解度回归直线是否一致。

以 pH 为自变量,各种溶液中的溶解度为因变量拟合直线,三条直线分别为

表 7 盐酸洛美沙星片在三种不同溶液中的溶解度 %

观测号	pH 值	溶解度		
		处方溶液	30%甘油溶液	50%甘油溶液
1	8.0	0.3	0.6	1.0
2	8.5	0.6	1.2	2.0
3	9.0	0.6	2.2	3.8
4	9.5	1.4	4.0	5.2
5	10.0	1.5	5.5	6.6
6	10.5	2.5	6.8	7.5

$$\hat{y}_1=-6.514\ 3+0.137\ 6x$$
$$\hat{y}_2=-20.772\ 4+2.611\ 4x$$
$$\hat{y}_3=-20.862\ 9+2.725\ 7x$$

首先做方差齐性检验,χ^2=0.775 8, P>0.05,故可以认为 3 条回归线的残差相等(此为回归系数间比较的前提条件)。进一步做回归直线的平行性检验,得 F=317.8, P<0.05,可以认为各回归直线彼此不平行。为了说明哪对回归直线不平行,进一步做回归系数间的两两比较。这里有三条回归直线,设可以组合出的比较次数为 m,据此应进行 $m=C_3^2=3$ 次对比。选定 $\alpha=0.05/m=0.05/3=0.017$,按 α=0.017,自由度为 ν=18-3-1=14 得到界值 $t_{0.017(14)}=2.708$,故仅 β_1 和 β_3 间的差异有统计学意义,盐酸洛美沙星片随着 pH 值的变化,其溶解速率在 50%甘油溶液中快于在处方溶液中。而在 30%甘油溶液中和在处方溶液中,盐酸洛美沙星片剂随着 pH 值的变化,其溶解速率的差异无统计学意义;在 50%甘油溶液中与在 30%甘油溶液中,盐酸洛美沙星片随着 pH 值的变化,其溶解速率的差异亦无统计学意义,见表 8。

表 8 例 4 的各回归直线斜率间两两比较的结果

对比组	t	P	检验水准修正值 α	检验结果
处方溶液-30%甘油溶液	-2.650	0.019 0	0.017	NS
处方溶液-50%甘油溶液	-2.772	0.015 0	0.017	*
30%甘油溶液-50%甘油溶液	-0.112	0.912 4	0.017	NS

注:表中“NS”表示差异无统计学意义;“*”表示 P<0.017,差异有统计学意义。

从以上几个例子可以看出 Bonferroni 法的应用非常广泛,其实 Bonferroni 法不仅可以用于以上几种情形,它的思想适用于所有的两两比较。值得注意的是,这并不是说 Bonferroni 法可以取代其它的两两比较方法。可以证明 Bonferroni 法调整后,总的犯Ⅰ类错误的概率 $1-(1-\alpha/m)^m<m\cdot\alpha/m=\alpha$(此处的 α/m 为每次比较时所取的检验 (下转第 374 页)

人［如一项或多项ADL依赖并伴有严重合并病或老年病症候群(如重度痴呆)］,只能作为姑息治疗的候选人;(3)中度功能损害的病人(如介于1组和2组之间)不能耐受延长生命的治愈性治疗但可从一些特别的药理途径上获益（如初始剂量减小,后续剂量增加到可耐受为止的化疗方法)。以上三组病人可能死于肿瘤或在生存时发生肿瘤并发症(见SAO-1)。第四组人群指的是那些期望寿命很短以至于不会发生肿瘤相关疾病的病人;这些病人可以接受症状处理和支持治疗(见NCCN支持治疗临床指引)。

可以耐受治愈性治疗方案的老年患者有特殊的需要。总的来说,年龄不是外科手术风险首要考虑的因素。然而,合并使用放化疗时要谨慎小心;必要时调整化疗的剂量。化疗可以导致一系列的问题(如神经毒性、心脏毒性、黏膜炎),但可以通过借用NCCN老年肿瘤临床指引上的推荐的专业方法而达到减少或预防的目的(见SAO-2)。

NCCN老年肿瘤专家组成员

* Lodovico Balducci, MD/Chair
H. Lee Moffitt Cancer Center & Research Institute at the University of South Florida
Harvey Jay Cohen, MD
Duke Comprehensive Cancer Center
Paul F. Engstrom, MD
Fox Chase Cancer Center
David S. Ettinger, MD
The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins
Leo I. Gordon, MD
Robert H. Lurie Comprehensive Cancer Center of Northwestern University
Jeffrey Halter, MD
University of Michigan Comprehensive Cancer Center
Krystyna Kiel, MD
Robert H. Lurie Comprehensive Cancer Center of Northwestern University
Andrew Kneier, PhD
UCSF Comprehensive Cancer Center
Dean Lim, MD
City of Hope Cancer Center
Stephen H. Petersdorf, MD
Fred Hutchinson Cancer Research Center/ Seattle Cancer Care Alliance
Ronnie Rosenthal, MD
Consultant
Rebecca Silliman, MD, PhD
Consultant
Julie M. Vose, MD
UNMC Eppley Cancer Center at The Nebraska Medical Center
Michael J. Walker, MD
Arthur G. James Cancer Hospital & Richard J.Solove Research Institute at The Ohio State University
Babu Zachariah, MD
H. Lee Moffitt Cancer Center & Research Institute at the University of South Florida

* 写作委员会成员

［收稿日期］ 2006-10-19

(上接第363页)

水准),且m值越大,结果越保守[2]。当多组间比较次数不多时,Bonferroni法的效果较好。但是,当组间比较次数较多(例如10次以上)时,检验水准过低,导致结果过于保守,犯Ⅱ类错误的概率增加[6]。顺便指出,与其它几种常用两两比较方法相比,Bonferroni法比LSD法、Duncan法、SNK法偏于保守,不过,它比Tukey法、Scheffe法要敏感。

［参 考 文 献］

[1] Bland JM, Altman DG. Multiple significance tests: the Bonferroni method [J]. Br Med J, 1995, 310(6986): 1073-1075.
[2] Heiberger RM, Holland B. Statistical Analysis and Data Display. Springer[M]. 2004: 155-168.
[3] 倪宗瓒. 医学统计学实习指导 [M]. 北京: 高等教育出版社, 2004: 118.
[4] 方积乾. 医学统计学与电脑实验 [M]. 第2版. 上海: 上海科学技术出版社, 2001: 8-9.
[5] 柳 青(总主编: 徐天和). 中国医学统计百科全书. 多元统计分册[M]. 北京: 人民卫生出版社, 2004: 96-97.
[6] Perneger TV. What's wrong with Bonferroni adjustments [J]. Br Med J, 1998, 316(7139): 1236-1238.

［收稿日期］ 2006-09-11