·讲座·

医学科研中合理运用统计学的方略

——定性资料统计分析方法合理选用(3)

胡良平[1],单彬[1],刘惠刚[2]

(1. 北京军事医学科学院生物医学统计咨询中心,北京 100850; 2. 首都医科大学继续教育学院,北京 100036)

[中图分类号] R195.1 [文献标识码] C [文章编号] 1008 - 8830(2005)03 - 0289 - 03

1 误用独立性检验取代一致性检验

例 1:某研究者在"CAP 过敏原检测系统在高危哮喘儿中的应用"一文中关于统计学处理的表述如下:计量资料结果用均数 ± 标准差($\bar{x} \pm s$)表示,计数资料比较用卡方检验。方法评价:一致性检验采用卡方检验,差异性采用 McNemar 检验,资料见表 1。

表 1 82 例高危哮喘患儿和 40 例支气管肺炎患儿行 CAP 检测的结果

Phadiatop 法	*n*		
尘螨 sIgE 法:	阳性	阴性	合计
阳性	41	13	54
阴性	0	68	68
合计	41	81	122

原作者对如何处理表 1 资料又作了详细说明:Phadiatop 与 sIgE 两种检测方法一致率为 89.3%(109/122)(表 1)。经一致性 χ^2 检验,$\chi^2 = 77.76$,$P < 0.01$;差异性检验,Phadiatop 阳性与 sIgE 阴性的情况多于 Phadiatop 阴性与 sIgE 阳性的情况。差异有显著意义($\chi^2_m = 11.70$,$P < 0.05$)。试辨析原作者在处理此资料时所犯的错误。

对差错的辨析与释疑:本例是一个配对设计的 2×2 表资料,也就是说,用两种检测方法对 122 例患儿的标本都作了检测,按每位患儿的两种检测结果填入配对的四格表中。原作者作了两方面的检验,即差异性检验(计算本身是正确的)和一致性检验(公式用错了!)。

原作者关于两种检测方法之间差异的检验,采用的是 *McNemar* χ^2 检验,仅当两种检测方法中有一种是"金标准"时,才是正确的。若两种检测方法都是被评价的方法,都有较高的假阳性率时,此时的检验无实际意义。

原作者关于两种检测方法的一致性检验,实际采用的是"独立性检验",而文中却声称采用的是"一致性检验"。因为经验算:反映四格表中"行变量与列变量之间"独立性的 χ^2 检验结果为:$\chi^2 = 77.763$,$P < 0.0001$。事实上,检验列联表中两个定性变量之间独立性的 χ^2 检验只能回答"两定性变量之间是否独立",当 $P < 0.05$ 时表明两定性变量之间不独立,说明表中两行上的频数构成是不同的,并不能说明主对角线上的两个频数之和 109(41 + 68)占总频数 122 的比例(即观察的一致率)有统计学意义。而反映两种检测方法测定结果之间一致性高低的方法应该选用 *kappa* 检验,本例中,*kappa* = 0.7786,标准误 $S_{kappa} = 0.0566$,$u = 13.752$,$P < 0.0001$,结论是两种检测方法的检测结果具有一致性。结合前面关于"差异性检验的结果"可知,虽然一致部分有统计学意义,但不一致(即差异)部分也有统计学意义,假定"尘螨 sIgE 法"测定的结果是"金标准",则意味着"Phadiatop 法"检测结果的假阳性率 24.07%(13/54)偏高。

2 随意压缩高维列联表易得出错误的结论

例 2:与例 1 属于同一篇论著中,还有一张统计表,见表 2。原作者认为:高危哮喘儿组与支气管肺炎组相比,两种方法在高危哮喘儿组检测阳性率高于支气管肺炎组,差异有显著意义。试问这样处理表 2 资料犯了什么错误?

表 2 高危哮喘及支气管肺炎患儿 Phadiatop 法及尘螨 sIgE 法检测结果

组别	*n*	阳性例数		阳性率(%)	
		Phadiatop 法	尘螨 sIgE 法	Phadiatop 法	尘螨 sIgE 法
高危哮喘儿组	82	46	39	56.1	47.6
支气管肺炎组	40	8	2	20.0	5.0
合计	122	54	41	44.3	33.6
两组阳性率比较 χ^2 值				14.20[a]	21.85[a]

a $P < 0.01$

对差错的辨析与释疑:将表 1 资料与表 2 资料对照起来看可知,表 1 是用两种检测方法同时检测 122 位患儿标

[收稿日期]2005-02-04
[作者简介]胡良平(1955 -),男,教授。主攻方向:实验设计与统计学。

本,并按配对设计格式但未按患病类型(高危哮喘、支气管肺炎)分组形成的四格表,而表 2 是同时按患病类型、检测方法(Phadiatop 法、尘螨 sIgE 法)和检测结果(阳性、阴性)分组,这是一个 3 维列联表资料,原作者采用"将 3 维列联表压缩成两个 2 维列联表资料的方法"做了两次 χ^2 检验,显然,未能分析两种检测方法之间的差别是否有统计学意义;更严重的是:不能正确地反映两个影响因素对观测结果的影响情况。

若两种检测方法中有一种方法测定的结果可作为"金标准"时,应将表 2 资料分别按"高危哮喘儿组"与"支气管肺炎组"列出两张类似于表 1 的表格,可以更有效地考察两种测定方法在测定不同类型的患者时的"一致性"与"差异性"情况。表 2 是按成组设计的格式列出了资料,人为扩大了样本含量(注意:表 2 中只给出了阳性数,计算时还需用到阴性数,两次 χ^2 检验的总样本含量为 244,而不是 122,见表 3)。

若实验的结果已弄乱顺序,无法严格按配对设计格式列表,就只能按类似表 2 的格式(即成组设计)列表了,但此时应将原始数据列成表 3 的格式,并采用分析 3 维列联表资料的统计分析方法处理资料为宜。

表 3　高危哮喘及支气管肺炎患儿 Phadiatop 法及尘螨 sIgE 法检测结果(表 2 修改结果)

患儿患病类别	检测方法	n			
		检测结果:	阳性	阴性	合计
高危哮喘儿	Phadiatop 法		46	36	82
	尘螨 sIgE 法		39	43	82
支气管肺炎	Phadiatop 法		8	32	40
	尘螨 sIgE 法		2	38	40
	合计		95	149	244

显然,表 3 是一个"结果变量为二值变量的 3 维列联表资料",可以选用的统计分析方法有三种,即"加权 χ^2 检验"、"多元 Logistic 回归分析"和"对数线性模型"。

若想通过统计学方法消除"患病类型"对结果的影响,着重考察两种检测方法阳性率之间的差别是否具有统计学意义,可考虑选用加权 χ^2 检验。其计算结果为:加权 $\chi^2 = 3.401$, $P = 0.065 > 0.05$,说明消除患病类型的影响后,两种检测方法阳性率之间的差别无统计学意义;当然,还可用加权 χ^2 检验消除检测方法对结果的影响,着重考察两种疾病类型阳性率之间的差别是否具有统计学意义,其计算结果为:加权 $\chi^2 = 35.405$, $P = 2.678 \times 10^{-9} < 0.0001$,说明消除检测方法的影响后,两种疾病类型检测结果的阳性率之间的差别有统计学意义,即高危哮喘儿检测结果的阳性率高于支气管肺炎患儿检测结果的阳性率;而且,高危哮喘儿与支气管肺炎患儿阳性检测结果的比数比 OR = 7.898,其总体 OR 值与 1 之间的差别需用 χ^2_{MH} 检验,$\chi^2_{MH} = 35.115$, $P = 3.108 \times 10^{-9} < 0.0001$,说明高危哮喘儿的阳性检出率是支气管肺炎患儿阳性检出率的7.898 倍。对数线性模型和多元 Logistic 回归分析的结果都与加权 χ^2 检验的结果是一致的,可以直接得到与上面两次加权 χ^2 检验所获得的结论(详细计算结果从略)。

3　随意分割 5 ×2 列联表增大了犯假阳性错误的概率

例 3:原文题目:"儿童低水平铅暴露与神经行为关系的研究",目的:探讨低水平铅暴露对儿童神经行为的影响。依据血铅水平将研究对象分为 5 组,不同组别行为异常率差异有显著性($\chi^2 = 13.695$, $P < 0.01$),进一步两两比较,当血铅≥150 μg/L 后 2 组行为异常率显著高于前 3 组(χ^2 分别为 4.727,6.261,5.168;4.503,5.911,4.928,均 $P < 0.05$),表明血铅≥150 μg/L 时,儿童行为异常率显著增多,资料见表 4。

表 4　不同血铅水平行为异常率

组别	血铅(μg/L)	n	异常人数	异常率(%)
Ⅰ	<50	40	3	7.50
Ⅱ	50 ~	56	4	7.14
Ⅲ	100 ~	58	5	8.62
Ⅳ	150 ~	39	10	25.64[a]
Ⅴ	250 ~	18	6	33.33[a]

对差错的辨析与释疑:本资料是一个 5 ×2 列联表资料,分组变量有序而结果变属无序,故可将其视为双向无序的列联表。满足一般 χ^2 检验的条件,故可采用一般 χ^2 检验,总的 $\chi^2 = 15.555$, $P = 0.0037$,说明不同血铅浓度组的异常率之间的差别有统计学意义。研究者在比较任何两个浓度组之间的差别时把原表分割成若干个四格表,增大了犯Ⅰ型错误的概率。

在对 5 个血铅浓度组作总的分析之后,即算出总的 χ^2 值,再对任何两个血铅浓度组构成的四格表资料作检验,共需作 $C = 5 \times 4/2 = 10$ 次检验。为了使做完全部比较犯Ⅰ型错误的概率 α 不增大,需对每次检验的检验水准 α′重新规定[1],即令:$\alpha' = \alpha/2C$(α′为单侧概率),此处,C 为比较的总次数,即 $C = k(k-1)/2$,k 为实验组的组数。本资料若定 $\alpha = 0.05$,则 $\alpha' = 0.05/(2 \times 10) = 0.0025$。当 df = 1 时,与概率为 0.005 对应的临界值为 7.88;与概率为 0.001对应的临界值为 10.83。所以,与概率为0.0025对应的临界值为介于 7.88 ~ 10.83 之间的数。而原文计算出的 6 个四格表资料所对应的 χ^2 值均小于7.88,故尚不能认为其中任何两组间的差别有统计学意义。这说明原文中因所用的统计分析方法欠妥,其结论的可靠性值得怀疑。

4　误用 χ^2 检验分析单向有序的列联表资料

例 4:原文研究目的:探讨黄芪注射液佐治小儿病毒性心肌炎的临床疗效及治疗机制。方法:对近 3 年来收住的 96 例病毒性心肌炎患儿分为观察组 70 例及对照组 26

例,观察组在传统治疗方法的基础上加用黄芪注射液静脉滴注,而对照组只用传统的治疗方法。治疗结果见表 5。

表 5　两组治疗效果比较　例(%)

分组	n	显效	有效	无效	有效率
对照组	26	11	10	5	(80.77)
观察组	70	42	26	2	(97.14)

两组比较 $\chi^2 = 8.64, P < 0.05$

对差错的辨析与释疑:此资料中原因变量(组别)是名义变量,结果变量(疗效)是有序变量,因而属结果变量为有序变量的单向有序列联表资料。因一般 χ^2 检验与变量的有序性没有联系,用一般 χ^2 检验进行分析,得到的结论是两组患者在三个疗效等级上的频数分布是否相同,而不能得出两组疗效之间的差别是否具有统计学意义的结论;原作者很可能是将"疗效"表达成三档,但分析时又将其按两档划分,即分为"有效、无效",将表 5 视为四格表资料来处理,选用一般 χ^2 检验进行分析,这就意味着"表达资料的方法"与"分析资料的方法"二者之间是不吻合的!

适合分析单向有序列联表资料的统计分析方法有秩和检验或 Ridit 分析等。本例采用秩和检验进行统计分析,其结果为:$H_c = 4.0727, P = 0.0436$,可认为两组疗效之间的差别有统计学意义。另外,根据表中所提供的数据,即使按双向无序的 2×3 列联表资料来对待,采用一般 χ^2 检验进行计算,得 $\chi^2 = 8.054, P = 0.018$;若转化成四格表资料来计算,得 $\chi^2 = 7.519, P = 0.0061$。与原作者的计算结果也都不相同[2]。

5　结果变量的资料类型究竟是什么

例 5:原文题目:"复方地西泮灌肠剂预防小儿热性惊厥复发的临床研究",为探讨复方地西泮灌肠剂预防小儿热性惊厥(FC)复发的作用。原作者把 91 例 FC 患儿随机分成复方地西泮组(A 组)、地西泮组(B 组)和布洛芬组(C 组)。在比较 3 组惊厥复发率时得到表 5 资料。

表 6　3 组 FC 患儿惊厥复发率比较

组别	例数	复发 1 次(例)	复发 2 次(例)	复发总例次	复发率(%)
A 组	32	1	0	1	3.1
B 组	29	4	2	6	20.7
C 组	30	5	3	8	26.7

3 组复发总例次比较,$\chi^2 = 39.052, P < 0.05$。

对差错的辨析与释疑:三种不同药物治疗组的治疗结果都分为未复发(复发 0 次)、复发 1 次和复发 2 次,因而结果变量应视为有序的,而组别因素各水平之间没有顺序,所以该资料应属于"结果变量(即复发次数)为有序变量的单向有序的 2 维列联表资料"。原作者用 χ^2 检验,无法回答三种不同药物的治疗结果之间的差别是否具有统计学意义,只能回答三个药物组的患者在不同复发次数上的人数分布是否相同,并且该资料有超过 1/5 的格子上的理论频数小于 5,不能用 χ^2 检验,只能用 Fisher 的精确检验,$P = 0.078$,尚不能认为三个药物组的患者在不同复发次数上的人数分布是不相同的;

若按"复发、未复发"划分进行 χ^2 检验,则得 χ^2 检验 $= 6.781, P = 0.0337$,说明原作者计算结果有误;另外,本例中"复发总例次"较小,而在某些问题中,各行上的"例次"值有可能超过各组总人数,这说明用"总复发率"为观测指标是不够合理的。

对于表 5 这样的"结果变量为有序变量的单向有序的 2 维列联表资料"可供选用的统计分析方法有:秩和检验、Ridit 分析和有序变量的 Logistic 回归分析。本例选用秩和检验,得:$H_c = 6.8438, P = 0.0327$。

应当注意的是:计算前应列出各组"复发 0 次"、"复发 1 次"和"复发 2 次"的人数,然后,再用秩和检验进行计算。

专业结论为:不同药物治疗组的治疗结果之间的差别具有统计学意义,将各组秩和的平均值代入近似的 t 检验,进行两两比较[3],得结果为:

对比组	平均秩之差	检验统计量 t 值	对应的 P 值
A 组 VS B 组	-8.070	5.265	<0.01
A 组 VS C 组	-10.873	7.156	<0.01
B 组 VS C 组	-2.803	1.800	>0.05

由结果可知复方地西泮组(A 组)的疗效与地西泮组(B 组)和布洛芬组(C 组)的疗效之间的差别都有统计学意义,结合表 5 中的复发例次可认为复方地西泮组的疗效好于地西泮组和布洛芬组的疗效;地西泮组和布洛芬组的疗效之间的差别无统计学意义,不能认为两者疗效不同。

再仔细看表 5 的观测结果,对每位患者而言是复发几次,其取值只有"0,1,2"三种,显然,本例中的"复发次数"是一个定量观测指标,只不过其取值的种类较少而已。事实上,表 5 资料更确切的判定结果应叫做"单因素 3 水平设计定量资料",由于各组只有 3 种可能的取值,其分布规律可能偏离正态分布较远,选用上面的秩和检验处理此资料是比较合理的。

值得一提的是:虽然都选中了秩和检验,但对资料中结果变量资料类型的判定结果却是绝然不同的,开始是将结果变量视为"多值有序变量",后来是将其视为"定量变量"。当定量变量的取值接近正态分布且组间总体方差相等时,是可以进行单因素 3 水平设计定量资料方差分析的,此时,若判定为"多值有序变量"则是很不恰当的了!

[参　考　文　献]

[1] 胡良平. 现代统计学与 SAS 应用[M]. 北京:军事医学科学出版社,2002,188-189.
[2] 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,2003,189-217,298-320.
[3] 郭祖超. 高等医药院校研究生教材医学统计学[M]. 北京:人民军医出版社,1999,71-74.

(本文编辑:王霞)