·讲座·

如何合理选择统计分析方法处理实验资料（V）

胡良平

编者按

生物医学期刊是宣传和反映生物医学科研与临床研究成果的重要媒体，是培养年轻科研工作者的摇篮，也是一个国家科研实力的重要象征。期刊中学术论文的质量是期刊存在的重要保证，而学术论文质量高低的重要标志之一是科研设计和统计分析质量的高低。本刊在2007年中，拟邀请军事医学科学院生物医学统计学咨询中心主任、博士生导师胡良平教授，以“如何合理选择统计分析方法处理实验资料”为题，撰写6篇文章，每期发表1篇，较系统地介绍在生物医学论文写作中，如何正确地应用医学统计学知识，从而提高学术论文的质量。需要指出的是，论文中统计学应用正确，并不能说明科研课题做得一定正确。广大作者和读者更应高度重视科研工作之前的科研设计的质量。事实上，由一个错误的科研设计产生出来的实验结果，即使其论文写得再漂亮，统计分析方法用得再正确，对于一个国家科技事业的发展和人才培养都是有害无益的。

上期文章介绍了如何合理选择统计分析方法处理四格表资料，本文将介绍如何合理选择统计分析方法处理R×C表资料和高维列联表资料。

1 双向无序的R×C表资料及对应的统计分析方法

双向无序的R×C表资料，顾名思义，就是表中2个定性变量都是名义变量，见表1和表2。

表1 某医院3年间4种甲状腺疾病在四季中发病人数的分布情况

甲状腺病	患者例数				
分类	季节：	春	夏	秋	冬	合计
甲亢		411	451	294	284	1440
亚甲炎		249	329	331	204	1113
甲低		60	61	59	52	232
甲状腺肿瘤		45	50	46	40	181
合计		765	891	730	580	2966

表2 心律失常种类与心肌梗死部位关系的调查结果

缓慢心律	患者例数				
失常种类	部位：	下壁	前壁	真后壁	心内膜下	合计
窦性过缓		8	7	2	1	18
被动心律		1	1	0	0	2
房室阻滞		6	3	1	1	11
束支阻滞		1	16	1	0	18
合计		16	27	4	2	49

上述2张表都属于双向无序的R×C表资料[1-2]，但表1中没有小于5的理论频数，故可以选用一般的χ^2检验公式计算；而表2中小于5的理论频数的格子数超过了总格子数的1/5，若仍选用一般的χ^2检验公式计算，将会增大犯假阳性错误的概率，故应改用Fisher的精确检验法。

2 结果变量为有序变量的单向有序的R×C表资料及对应的统计分析方法

单向有序的R×C表资料，要特别强调结果变量是有序的，见表3；若仅原因变量是有序的单向有序的R×C表资料，仍应将其视为“双向无序的R×C表资料”，见表4。

表3 3种药物治疗某病患者疗效的观察结果

药物	患者例数				
种类	疗效：	治愈	显效	好转	无效	合计
A		15	49	31	5	100
B		4	9	50	22	85
C		1	15	45	24	85
合计		20	73	126	51	270

表4 CAM-1的表达与食管癌TNM分期的关系

食管癌	患者例数			
TNM分期	基因表达情况：	表达	未表达	合计
Ⅱa		3	4	7
Ⅱb		8	2	10
Ⅲ		21	2	23
合计		32	8	40

上述的2张表虽然都属于单向有序的列联表，但表3中的结果变量“疗效”是有序的，而表4中的原因变量“食管癌TNM分期”是有序的[1-2]。前者应称为“结果变量为有序变量的单向有序R×C列联表”，可以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

作者单位：100850 北京，军事医学科学院生物医学统计学咨询中心

析；而后者应被视为“双向无序的 R×C 列联表”，因列联表内小于 5 的理论频数的格子数超过了总格子数的 1/5，故宜选用 Fisher 的精确检验法。

3 双向有序且属性不同的 R × C 表资料及对应的统计分析方法

当 R×C 列联表中的 2 个定性变量都是有序变量，且它们的属性（一个变量为年龄，而另一个变量为疗效，显然它们反映了事物的不同方面，称为属性）不同，此时，称这样的列联表资料为双向有序且属性不同的 R×C 列联表资料，见表 5。

表 5 地方性甲状腺肿患者各年龄组疗效的观察结果

年龄	患者例数				
（岁）	疗效： 治愈	显效	好转	无效	合计
11～	35	1	1	3	40
20～	32	8	9	2	51
30～	17	13	12	2	44
40～	15	10	8	2	35
50～	10	11	23	5	49
合计	109	43	53	14	219

表 5 中，原因变量（年龄）与结果变量（疗效）都是有序变量，它们的属性是完全不同的，因此，该表被称为双向有序且属性不同的 R×C 列联表资料。对于这样的资料应当采用何种统计分析方法处理合适呢？不能一概而论，应视具体的分析目的而定。一般来说，有以下 4 个可能的分析目的[2]。

(1)只关心各年龄组患者治疗结果之间的差异是否具有统计学意义，此时，年龄的有序性就变得无关紧要了，可将此时的“双向有序 R×C 列联表资料”视为“结果变量为有序变量的单向有序 R×C 列联表资料”，可以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。

(2)若希望考察年龄与疗效之间是否存在线性相关关系，此时，需要选用处理定性资料的相关分析方法，通常采用 Spearman 秩相关分析方法。

(3)若 2 个有序变量之间的相关关系具有统计学意义，研究者希望进一步了解这 2 个有序变量之间的变化关系是呈直线关系还是呈某种曲线关系，此时宜选用线性趋势检验。

(4)若希望考察列联表中各行上的频数分布是否相同，宜选用一般 χ^2 检验或 Fisher 的精确检验（若列联表内小于 5 的理论频数的格子数超过了总格子数的 1/5）。

4 双向有序且属性相同的 R × C 表资料及对应的统计分析方法

当 R×C 列联表中的 2 个定性变量都是有序变量，且它们的属性相同，则称这样的列联表资料为双向有序且属性相同的 R×C 列联表资料，见表 6 和表 7。

表 6 446 例流行性出血热患者病情转化情况

早期	患者例数			
分度	最后定型： 轻型	中型	重型	合计
轻度	111	21	1	133
中度	5	163	20	188
重度	0	1	124	125
合计	116	185	145	446

表 7 100 例脑肿瘤患者临床诊断与 CT 诊断的结果

临床诊断	患者例数			
结果	CT 诊断结果： 检出	疑惑	未检出	合计
检出	60	4	2	66
疑惑	4	12	3	19
未检出	3	3	9	15
合计	67	19	14	100

表 6 和表 7 中，2 个定性变量都是结果变量且都是有序变量，它们的属性是完全相同的，而且，各定性变量的水平数也是相同的，因此，这样的表被称为双向有序且属性相同的“方形”列联表资料，简称为“方表”[1-2]。对表 6 而言，研究者希望看前、后 2 个不同时间点上诊断的结果是否具有一致性，而对表 7 而言，研究者希望考察 2 种方法诊断的结果是否具有一致性。它们的本质是相同的，都是希望回答 2 种检测方法检测结果是否具有一致性的问题。这样的资料实际上就是配对设计 2×2 列联表资料的“扩大”，只不过在处理配对设计 2×2 列联表资料时，人们更关心的是 2 种检测方法检测的结果不一致部分的数量之间的差异是否具有统计学意义，而在处理“方表”资料时，人们更关心的是 2 种检测方法检测的结果之间是否具有一致性，故常用的统计分析方法叫做一致性检验或称为 Kappa 检验。

5 高维列联表资料及对应的统计分析方法

当列联表中包含的定性变量的格子数大于等于 3 时，就称为高维列联表资料。见表 8～10。

上面的 3 张表都属于三维列联表，但仔细考察它们的内部情况时，不难发现它们是有区别的[1-2]。对表 8 而言，“接受护理地点”和“产前护理量”都属于原因变量，而“婴儿存活状况”是结果变量，它是“二值变量”，这样的资料可以选用“多重 logistic 回归分析方法”或“对数线性模型”处理，还可以用加权的方法消除“接受护理地点”对结果的

表 8 孕妇在 2 个诊所接受产前护理量与婴儿存活情况的观察结果

接受护	产前	婴儿例数		死亡率
理地点	护理量	婴儿存活状况： 死	活	(%)
诊所 A	少	3	176	1.676
	多	4	293	1.347
诊所 B	少	17	197	7.944
	多	2	23	8.000

表9 甲、乙医院用A、B、C3种药医治某病的疗效观察

医院名称	药物名称	患者例数 疗效:	治愈	显效	好转	无效	合计
甲	A		15	49	31	5	100
	B		4	9	50	22	85
	C		1	15	45	24	85
乙	A		36	115	184	47	382
	B		4	18	44	35	101
	C		1	9	25	4	39

表10 在某地随机抽查4种职业的人按性别与血型分类的结果

性别	职业	调查人数 血型:	A	B	AB	O	合计
女	P_1		25	46	14	23	108
	P_2		28	52	11	31	122
	P_3		34	49	17	28	128
	P_4		45	31	18	34	128
男	P_1		54	48	24	36	162
	P_2		42	51	21	43	157
	P_3		65	52	32	49	198
	P_4		37	41	22	39	139

注：这是一个假设的例子

影响，仅考察“产前护理量”对“婴儿存活率”的影响，即采用加权 χ^2 检验处理资料；对表9而言，“医院名称”和“药物名称”都属于原因变量，而“疗效”是结果变量，它是多值有序变量，这样的资料可以选用“有序变量的多重logistic回归分析方法”处理；对表10而言，表中虽有3个定性变量，但以其中任何一个作为结果变量都不太妥当。若想以“血型”为结果变量，它却是一个多值名义变量，分析这种资料的统计分析方法可以选用“对数线性模型”或“多项logit模型（属于多重logistic回归模型的一种扩展模型）”予以处理。

参考文献

[1] Hu LP, Li ZJ. Fundamental of medical statistics and discrimination of typical misuse. Beijing: Press of Military Medical Sciences, 2003: 1-247. (in Chinese)
胡良平, 李子建. 医学统计学基础与典型错误辨析. 北京: 军事医学科学出版社, 2003:1-247.

[2] Hu LP, Liu HG, Li ZJ, et al. Scientific research design and statistical analysis of laboratory medicine. Beijing: People's Military Medical Press, 2004:86-250. (in Chinese)
胡良平, 刘惠刚, 李子建, 等. 检验医学科研设计与统计分析. 北京: 人民军医出版社, 2004:86-250.

·协会之窗·

中国医药生物技术协会常务理事扩大会议在青岛召开

2007年9月7日，中国医药生物技术协会第3届第10次常务理事扩大会议在青岛召开。

会议由中国医药生物技术协会彭玉理事长主持。首先，由刘海林副理事长兼秘书长作题为“2007年上半年工作汇报和下半年工作计划”的报告。在去年的理事会及今春的常务理事会上，协会制订了今年的工作目标，提出了“六个一工程”，即“办一个精品会议，创一份权威期刊，买一处永久会所，发展100家单位会员和1000名个人会员，筹备一个富有朝气、开拓进取、团结务实的新一届协会领导集体”。协会2007年上半年的工作均围绕这六个方面展开。今年下半年的主要工作是坚持不懈地把年初提出的目标完成好，主要抓好4件事：①筹备换届选举工作，包括成立筹备换届选举领导小组、组织文件起草、会章修改等小组，全面启动筹备工作；②在协会组织建设方面，着重发展企业/单位会员，增加企业会员在协会中的比重；③杂志编辑部要充分发挥编委的作用，努力扩大稿源，探索扩大发行和广告招商的新思路，扩大杂志的影响，提升及效益；④认真学习领会国务院办公厅《关于加快推进行业协会商会改革和发展的若干意见》的精神，抓住机遇，增强紧迫感，积极争取政府部门授权开展行业统计、行业资质认证，积极开展新技术和新产品鉴定、推广和开拓国际市场等服务。

其后，与会代表就刘海林秘书长的报告发表了意见和建议，肯定了协会上半年的工作，认为行业协会的特点越来越鲜明，为产、学、研的结合发挥了桥梁、中介作用，促进了横向技术交流。希望协会继续增强服务意识，充分发挥行业指导作用，按照国务院有关文件的指示精神，推动制定行业标准，推动一批企业的发展，促进科学家和企业家结合，培植我国医药生物领域内的龙头企业。

本次会议上，经全体与会理事审议通过，增补辽宁锦州奥鸿药业有限责任公司、沈阳解放军第四二三医院、深圳市赛百诺基因技术有限公司为协会的常务理事单位。

（文勇）