Medical Statistics 医学统计学

高维列联表资料的统计分析与 SAS 软件实现(一)

葛毅[1]，胡良平[2]

1. 后勤指挥学院，北京 100858

2. 军事医学科学院生物医学统计学咨询中心，北京 100850

关键词：统计学；医学；数据分析，统计；定性资料；SAS 软件

Ge Y, Hu LP. *J Chin Integr Med*. 2009; 7(11): 1086-1089.
Received October 3, 2009; accepted October 23, 2009; published online November 15, 2009.
Indexed/abstracted in and full text link-out at PubMed. Journal title in PubMed: *Zhong Xi Yi Jie He Xue Bao*.
Free full text (HTML and PDF) is available at www.jcimjournal.com.
Forward linking and reference linking via CrossRef.
DOI: 10.3736/jcim20091112

Open Access

Statistical analysis for data of multidimensional contingency table with SAS software package (Part one)

Yi GE[1], Liang-ping HU[2]

1. Command Academy of Logistics, Beijing 100858, China

2. Consulting Center of Biomedical Statistics, Academy of Military Medical Sciences, Beijing 100850, China

Keywords: statistics; medicine; data analysis, statistical; qualitative data; SAS software

在生物医学研究中，往往遇到定性资料，如果某项研究涉及的定性变量个数大于 2，并且用列联表的形式表现出来，则称该列联表为高维列联表。高维列联表的维数由原因和结果变量的个数共同决定，且依据结果变量的性质将高维列联表分为结果变量为二值变量的高维列联表、结果变量为多值有序变量的高维列联表和结果变量为多值名义变量的高维列联表。

1 高维列联表资料统计分析方法的选择

对于高维列联表资料统计分析方法的选择主要依据研究目的，同时还要考虑结果变量的性质。对于结果变量为二值变量的高维列联表，通常可选用 CMH χ^2 检验、多重 logistic 回归分析以及对数线性模型等方法。对于结果变量为二值变量的三维列联表还可选用加权 χ^2 检验(消除掉一个原因变量对结果变量的影响，考察另一个原因变量与结果变量之间是否独立)、M-H χ^2 或 CMH χ^2 检验，即消除掉一个原因变量对结果变量的影响，计算优势比(odds ratio, OR)或相对危险度(relative risk, RR)并对其进行假设检验。对于结果变量为多值有序(或多值名义)变量的高维列联表资料，若研究者考察资料的整体情况，则应选用结果变量为多值有序变量的高维列联表资料的统计分析方法。一般可选择 CMH 校正秩和检验或有序变量多重 logistic 回归分析。在以上的众多方法中，本文主要介绍加权 χ^2 检验和 CMH χ^2 检验的原理及其应用。

加权 χ^2 检验和 CMH χ^2 检验是处理三维列联表资料的常用方法，其主要特点是若根据研究目的，不想用复杂的对数线性模型或 logistic 回归模型来分析三维列联表资料，并且资料又不适合采用简单“合并”方式处理时，就可采用加权 χ^2 或 CMH χ^2 检验。值得注意的是，该检验方法无法回答被合并掉的那个原因变量对结果变量的影响作用有多大，只是评价另一个原因变量对结果变量的影响时，将其对结果变量的影响扣除掉[1-2]。

2 加权 χ^2 检验

2.1 加权 χ^2 检验的原理 加权 χ^2 检验主要用于分析结果变量为二值变量的三维列联表资料。如果被

Correspondence: Prof. Liang-ping HU; Tel: 010-66931130; E-mail: lphu812@sina.com

消除的那个原因变量有 i 个水平，则可将该资料视为 i 个 2×2 列联表资料。设第 i 层 2×2 列联表所对应的观察频数依次为 a、b、c、d，这 4 个观察的频数分别位于第 1 行第 1 列、第 1 行第 2 列、第 2 行第 1 列、第 2 行第 2 列，其合计频数为 n，加权 χ^2 检验公式如下：

$$\chi^2_{加权}=\frac{[\sum[(ad-bc)/n]_i/\sum[(a+b)(c+d)/n]_i]^2}{\sum[(a+b)(c+b)(a+c)(b+d)/n^3]_i/[\sum[(a+b)(c+d)/n]_i]^2}$$

此加权 χ^2 统计量服从 $df=1$ 的 χ^2 分布。其中，若想计算调整一个因素后，另一个因素与结果变量的 OR 或 RR，可采用下式来计算：$OR=\sum(ad/n)_i/\sum(bc/n)_i$，$RR=\sum[a(c+d)/n]_i/\sum[(a+b)c/n]_i$。当发生的事件为罕见事件时，OR 值与 RR 值近似。

$\chi^2_{MH}=\{\sum[(ad-bc)/n]_i\}^2/\sum[efgh/(n-1)/n^2]_i$，$df=1$，其中 $e=a+b$，$f=c+d$，$g=a+c$，$h=b+d$，$n=a+b+c+d$。以上公式被称为 Mantel-Haensel 公式。

2.2 *加权 χ^2 检验的应用* 有一项关于齐多夫定(zidovudine，AZT)药物是否能延缓 AIDS 病人症状进展的研究[3]。有 338 名感染 AIDS 病毒后免疫力开始下降的志愿者接受试验，他们被随机地分为使用 AZT 组及不使用 AZT 组，在 3 年内看其 AIDS 症状是否加重。数据见表 1。

表 1 不同种族的 AIDS 病人使用 AZT 后的症状是否加重

种族	AZT 用否	人数	
	症状加重与否：	是	否
白种人	用	14	93
	否	32	81
黑人	用	11	52
	否	12	43

由表 1 可以看出，该列联表资料涉及 2 个原因变量，即种族、是否用药，结果变量为症状是否加重。该资料类型属于结果变量为二值变量的三维列联表资料，对于此表不应采用简单合并方法来实现该三维列联表资料的压缩。所谓简单合并方法就是不考虑种族的影响，把是否用 AZT 药物组中两组相对应的症状加重的人数相加以及症状没有加重的人数相加，最后合并成一个四格表资料再进行分析。这样做是不对的！因为种族在两药物组各自内部构成是不同的，也就是说种族与药物种类或疗效之间并不是独立的。为了选用较简单的分析方法且还能消除“种族”对结果变量的影响，可以选用加权 χ^2 检验或 M-H χ^2 检验。

采用加权 χ^2 检验的结果如下(用 SAS 统计软件包计算)：W-chis $q=6.802$，W-$p=0.009\ 105\ 581\ 7$，RR=0.561，MH-chis $q=6.762$，MH-$p=0.009\ 311\ 933\ 6$。其中，W-chis q 为加权 χ^2 值，W-p 为对应的 P 值，即加权 $\chi^2=6.802$，$P=0.009\ 105\ 581\ 7$；RR=0.561，MH-chis q 为 MH χ^2，MH-p 为其对应的 P 值，即 MH $\chi^2=6.762$，$P=0.009\ 311\ 933\ 6$。

专业结论 消除种族的影响后，发现是否用 AZT 药物与 AIDS 的症状加重之间的关联有统计学意义。相对危险度 RR=0.561，说明使用 AZT 药物对 AIDS 症状的加重有保护作用。

上例除了用加权 χ^2 检验或 M-H χ^2 检验外，还可以用 CMH χ^2 检验、logistic 回归分析。其检验的结果类似，logistic 回归分析结果中种族因素无统计学意义，是否用 AZT 药物有统计学意义，其 RR=0.487。需要说明的是，logistic 回归分析所计算的 RR 值与加权 χ^2 检验所计算的不一样，这除了两种计算方法不一样外(logistic 回归分析方法将在后续的文章中介绍)，更重要的是看 logistic 回归模型拟合的好坏(具体判断方法将在后续的文章中介绍)。从本例上看，用 logistic 回归分析该资料，其模型拟合较好，$P=0.24$。但是对于有些三维列联表资料，其 logistic 模型的拟合结果不是很好，此时便显现出加权 χ^2 检验或 M-H χ^2 检验的优势了。

3 CMH χ^2 检验的应用

3.1 *CMH χ^2 检验的原理* CMH 统计分析(Cochran-Mantel-Haenszel statistics)也称为 CMH 校正的秩和检验，它是在 MH 统计分析方法的基础上发展并提出来的，现在统称为扩展的 MH 卡方统计量。它是在考虑控制分层混杂因素影响的前提下，根据 R×C 表格中行变量与列变量的属性不同，给出下例 3 种检验统计量。(1)当行变量与列变量均为有序变量时，计算非零相关统计量(nonzero correlation)，这样可以获得类似秩相关分析的检验结果，但不能给出秩相关系数，也未体现出相关的方向；条件如不满足时，计算该统计量则没有意义。(2)当行变量为无序变量而列变量为有序变量时，计算行平均分差异统计量(row mean scores

difference),也称为方差分析统计量(有别于定量资料方差分析),可以获得校正的 CMH 秩和检验结果。(3)当行变量与列变量均为无序变量或行变量是有序变量而列变量为无序变量时,计算一般关联统计量(general association)。

值得注意的是,当结果变量为二值变量时,用 CMH 检验得出的以上 3 个统计量数值相等,CMH χ^2 即第 3 种统计量。其检验假设的步骤如下:H_0,每层中原因变量和结果变量之间不存在关联;H_1,至少有一层原因变量和结果变量之间存在某种关联。当仅有一层时,该 CMH 统计量与 Pearson χ^2 统计量的关系为:$\chi^2_{CMH}=[(n-1)/n]\chi^2$,其中 n 为总例数;当有多层时,该统计量为层修正的 Pearson χ^2 统计量。当然,相似的校正也能够通过对各层 Pearson χ^2 统计量求和而得到,但是这种校正方法需要每层的样本含量都要足够大,而 CMH 统计量仅仅需要总的样本含量比较大。

3.2 CMH χ^2 检验的应用

3.2.1 结果变量为二值变量时的应用 如表 1 中的资料,除了用加权卡方检验及 logistic 回归分析以外,同样还可以用 CMH χ^2 检验进行分析。由于该资料属于结果变量为二值变量的三维列联表资料,据前述内容可知,可采用一般关联统计量,其值为 6.762 4,P=0.009 3。从统计量的计算结果上看与加权卡方接近。专业结论为:消除种族的影响后,是否用 AZT 药物,AIDS 症状加重的差别有统计学意义。RR=0.561,说明使用 AZT 药物有利于减轻 AIDS 的症状。

3.2.2 结果变量为多值有序变量时的应用 有一项关于不同政党对于意识形态的看法的调查,其看法的结果分为非常自由到非常保守 5 个等级,具体资料见表 2[4]。

表 2 不同性别及政党对政治意识形态的看法

性别	政党	人数				
	看法:	非常自由	轻度自由	中等	轻度保守	非常保守
女	民主党	44	47	118	23	32
	共和党	18	28	86	39	48
男	民主党	36	34	53	18	23
	共和党	12	18	62	45	51

该调查研究的主要目的是考察不同政党的成员对意识形态的看法是否一致。由表 2 可以看出,该列联表资料涉及 2 个原因变量,即性别和政党,结果变量为看法。其中结果变量分了 5 个等级(非常自由、轻度自由、中等、轻度保守、非常保守),这 5 个等级之间存在一定的顺序性,因此该资料的类型为结果变量为多值有序变量的三维列联表资料。依据该调查目的,可以将"性别"看作是一个分层因素,这样可以考虑用 CMH χ^2 检验进行处理。

计算结果如下:因为结果变量为多值有序变量,因此采用 CMH χ^2 检验时应选取行平均分差异统计量,其值为 χ^2_{CMH}=50.181 9,P<0.000 1。专业结论为:在消除了性别的影响后,民主党与共和党成员在意识形态上的看法是不一样的,从各行构成比来看,民主党对意识形态的看法相对自由一些。

3.2.3 结果变量为多值名义变量时的应用 有一项关于老年人信仰的调查,主要目的是考察不同性别的老年人的信仰状况,具体数据见表 3[5]。

在该研究中原因变量为种族和性别,其中种族和性别为二值名义变量;结果变量为信仰状况,有 3 个水平,分别为有信仰、不确定、没有信仰,属于多值名义变量。其主要的目的为考察不同性别的老年人的信仰状况的构成情况是否有差别,因此可以认为种族为一个分层因素,需要平衡掉,进而可以考虑采用 CMH χ^2 检验。由于其原因变量为二值变量,结果变量为多值名义变量,因此在采用 CMH χ^2 检验时应选取一般关联统计量。其计算结果为:χ^2_{CMH}=7.231 9,P=0.026 9。

因此,可以认为,在控制了种族这个因素之后,性别与信仰状况之间存在一定的关联。结合列联表中的行百分数可以看到,相对来说,老年人中女性比男性有信仰的人数要多,换句话讲,老年人中女性比男性更容易有信仰。

表 3 不同性别老年人的信仰状况

种族	性别	人数		
	信仰状况:	有信仰	不确定	没有信仰
白种人	女	371	49	74
	男	250	45	71
黑人	女	64	9	15
	男	25	5	13

从上述 3 个例子当中,可以看出 CMH χ^2 检验可以处理结果变量为二值变量、多值有序变量、多值名义变量的高维列联表。它可以控制一个原因变

量，而考察其他原因变量同结果变量之间的联系。虽然 logistic 回归模型及其他模型也能处理这类资料，但是需要对模型的拟合效果进行判定，只有拟合效果好的模型才有应用价值。更重要的是当其中某一个原因变量的频数很少，但总的频数较多时 CMH χ^2 将显现出它的检验效能，所以如果维数不高，且想控制一个因素时，可以考虑使用 CMH χ^2 检验进行分析。

加权 χ^2 检验及 CMH χ^2 检验都是处理高维列联表资料的常用方法，它们的主要特点就是能够控制一个或多个原因变量，特别适合控制分层因素（通常为重要的非实验因素）。但需要说明的是，当各层比较组间的变化趋势一致时，上述方法比较有效；反之，则此方法不容易检测出差别来，此时应单独考察各层或采用其他方法进行分析。

REFERENCES

1 Hu LP, Liu HG, Li ZJ. Design of scientific researches in ecsomatics and statistical analysis. Beijing: People's Military Medical Press, 2004: 132-143. Chinese.
胡良平，刘惠刚，李子建. 检验医学科研设计与统计分析. 北京：人民军医出版社. 2004：132-143.

2 Hu LP. Applied course of statistical analysis by Windows 6.12 & 8.0. Beijing: Press of Military Medical Sciences, 2001: 335-354. Chinese.
胡良平. Windows SAS 6.12 & 8.0 实用统计分析教程. 北京：军事医学科学出版社. 2001：335-354.

3 Stein MD, Piette J, Mor V, Wachtel TJ, Fleishman J, Mayer KH, Carpenter CC. Differences in access to zidovudine (AZT) among symptomatic HIV-infected persons. J Gen Intern Med. 1991; 6(1): 35-40.

4 Agresti A. Categorical data analysis. New York: John Wiley & Sons, Inc. 2002: 165-178.

5 Agresti A. An introduction to categorical data analysis. New York: John Wiley & Sons, Inc. 2006: 114-120.