如何正确运用χ^2检验——高维表资料独立性检验与SAS实现

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】 本文目的是介绍一种特殊高维表（即g×2×2表）资料的独立性检验方法及SAS实现。在SAS软件和统计学教科书中，有三种方法可用于进行高维表资料的独立性检验，分别为广义CMH χ^2检验（简称为“方法1”）、公式中含有权重系数的加权χ^2检验（简称为“方法2”）和公式中没有权重系数的加权χ^2检验（简称为“方法3”）。本文通过公式推导和变形，揭示出“方法2”与“方法3”在本质上是完全相同的加权χ^2检验，但具有不同的表现形式；还揭示出加权χ^2检验统计量的估计值与“方法1”中的CMH χ^2检验统计量的估计值近似相等。本文结合一个实例，介绍基于SAS软件实现g×2×2表资料独立性检验的具体方法，对输出结果进行解释，并做出统计结论和专业结论。

【关键词】 独立性检验；权重；加权χ^2检验；CMH χ^2检验
中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20210719002

How to use χ^2 test correctly——the independence test for the data of a multiway table

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*: *lphu*927@163.*com*）

【**Abstract**】 The purpose of this article was to introduce the methods for the independence test of a special high-dimensional table (ie g×2×2 table) and its SAS implementation. There were three approaches, in the SAS software and the statistical textbooks, which could be used to perform the independence test for the data of a multiway table. The three kinds of methods were the generalized CMH χ^2 test (for short, approach-1), the weighted χ^2 test with the weighted coefficients in its formula (for short, approach-2), and the weighted χ^2 test without the weighted coefficients in its formula (for short, approach-3), respectively. This article revealed that the "approach-2" and "approach-3" were the same weighted χ^2 test essentially, but with different manifestations. It also revealed that the weighted χ^2 test statistic estimation was approximately equal to the CMH χ^2 test statistic estimation in the "approach-1". Based on an example and the SAS software, the article introduced the concrete approaches for the independence test of the g×2×2 table data, explained the output results, and made the statistical and professional conclusions.

【**Keywords**】 Independence test; Weight; Weighted χ^2 test; CMH χ^2 test

对高维表资料进行独立性分析的基本思路是将高维表降为二维表，降维的重要举措就是按一个因素的全部水平或多个因素的全部水平组合对资料进行分层，从而使每层中的资料都是一个二维表资料。一种特殊的高维表就是分层后的二维表为2×2表（即含一个二值的原因变量和一个二值的结果变量），简记为“g×2×2表”。针对独立性检验问题，本文将介绍CMH χ^2检验[1-3]和加权χ^2检验[3-5]两种方法，并通过实例，介绍使用SAS软件[1]实现计算的具体方法。

1 与高维表资料独立性检验有关的概念

1.1 高维表（g×2×2表）资料的表达模式

高维表（g×2×2表）资料的表达模式见表1。

表1 高维表（g×2×2表）资料的第h层2×2表资料的表达模式

危险因素		例 数		
接触与否	是否患病：	患病	未患病	合计
接触		n_{h11}	n_{h12}	$n_{h1\cdot}$
未接触		n_{h21}	n_{h22}	$n_{h2\cdot}$
合计		$n_{h\cdot1}$	$n_{h\cdot2}$	n_h

注：h=1，2，…，g

1.2 高维表资料独立性检验的含义

设高维表资料中有K个因素(或自变量),1个定性的结果变量。除了采用回归分析可以同时考察K个因素对定性结果变量的影响之外,差异性分析的思路是将$K-1$个因素当作分层变量,只研究剩余的一个因素对二值定性结果变量的影响,这被称为将高维表降维后使其成为二维表。显然,在分层变量(它可以是1个因素,也可以是多个因素的水平组合)的每个水平下,都有一张二维表。假定分层变量有g($g\geq 2$)个水平,则有g张2×2表(注:本文不考虑g张$R\times C$表)。研究者关心的是各层2×2表资料中"原因变量"与"结果变量"之间是否独立(不独立时,就意味着存在关联),为了回答这个问题,需要进行高维表资料的独立性检验。在文献[1,3,6-7]中,实现此检验的方法叫做广义CMH χ^2检验;而在文献[4-5]中叫做加权χ^2检验。

2 高维表资料独立性检验及SAS实现

2.1 高维表资料加权χ^2检验的具体算法

2.1.1 隐含权重的加权χ^2检验的具体算法

在$g\times 2\times 2$表资料中,设含有g个水平的因素为重要非试验因素,按其分层可得到g个2×2表资料。于是,可按如下的公式将g个2×2表资料整合成一个χ^2检验统计量$\chi^{2[3]}_W$:

$$X_1=\frac{\left\{\sum_h\left[\left(n_{h11}\times n_{h22}-n_{h12}\times n_{h21}\right)/n_h\right]\right\}^2}{\sum_h\left[\left(n_{h11}+n_{h12}\right)\left(n_{h21}+n_{h22}\right)/n_h\right]} \tag{1}$$

$$X_2=\frac{\sum_h\left[\left(n_{h11}+n_{h12}\right)\left(n_{h21}+n_{h22}\right)\left(n_{h11}+n_{h21}\right)\left(n_{h12}+n_{h22}\right)/n_h^3\right]}{\left\{\sum_h\left[\left(n_{h11}+n_{h12}\right)\left(n_{h21}+n_{h22}\right)/n_h\right]\right\}^2} \tag{2}$$

$$\chi^2_W=\frac{X_1}{X_2}=\frac{\left\{\sum_h\left[\left(n_{h11}\times n_{h22}-n_{h12}\times n_{h21}\right)/n_h\right]\right\}^2}{\sum_h\left[\left(n_{h11}+n_{h12}\right)\left(n_{h21}+n_{h22}\right)\left(n_{h11}+n_{h21}\right)\left(n_{h12}+n_{h22}\right)/n_h^3\right]}\sim\chi^2_1 \tag{3}$$

式(3)表明,χ^2_W服从自由度为1的χ^2分布。

【说明】在式(1)、式(2)、式(3)中,看不见反映各层2×2表资料重要性的"权重W_h",故称式(3)为"隐含权重的加权χ^2检验统计量"。

2.1.2 突显权重的加权χ^2检验的具体算法

文献[4-5]提供了另一个突显权重的加权χ^2检验统计量,见式(4)、式(5)、式(6):

$$\chi^2_W=\frac{\bar{d}^2}{S^2_{\bar{d}}}\sim\chi^2_1 \tag{4}$$

式(4)中,$\bar{d}^2$与$S^2_{\bar{d}}$的计算分别见式(5)、式(6):

$$\bar{d}=\frac{\sum_h W_h d_h}{\sum_h W_h} \tag{5}$$

$$S^2_{\bar{d}}=\frac{\sum_h W_h P_h Q_h}{\left(\sum_h W_h\right)^2}=\frac{\sum_h W_h P_h\left(1-P_h\right)}{\left(\sum_h W_h\right)^2} \tag{6}$$

在上面三式中,$\bar{d}$、$S^2_{\bar{d}}$分别代表"合并后2×2表资料中两行上阳性率之差的加权平均值"及其"方差";而W_h、d_h、P_h、Q_h分别代表第h层2×2表资料中的"权重""率差""平均阳性率"和"平均阴性率",其计算分别见下式:

$$W_h=\frac{n_{h1\cdot}\times n_{h2\cdot}}{n_{h1\cdot}+n_{h2\cdot}}=\frac{n_{h1\cdot}\times n_{h2\cdot}}{n_h} \tag{7}$$

$$d_h=P_{h1}-P_{h2}=\frac{n_{h1\cdot}}{n_h}-\frac{n_{h2\cdot}}{n_h} \tag{8}$$

$$P_h=\frac{n_{h11}+n_{h21}}{n_{h1\cdot}+n_{h2\cdot}}=\frac{n_{h11}+n_{h21}}{n_h} \tag{9}$$

$$Q_h=1-P_h=\frac{n_{h12}+n_{h22}}{n_h} \tag{10}$$

将式(5)~式(10)代入式(4),可得到式(11):

$$\chi^2_W=\frac{\bar{d}^2}{S^2_{\bar{d}}}=\frac{\left(\sum_h W_h d_h/\sum_h W_h\right)^2}{\left(\sum_h W_h P_h Q_h\right)/\left(\sum_h W_h\right)^2}\sim\chi^2_1 \tag{11}$$

对式(11)进行变形,得到式(12):

$$\chi^2_W=\frac{\bar{d}^2}{S^2_{\bar{d}}}=\frac{\left[\sum_h\left(n_{h11}\times n_{h2\cdot}-n_{h21}\times n_{h1\cdot}\right)/n_h\right]^2}{\sum_h\left[\left(n_{h11}+n_{h12}\right)\left(n_{h21}+n_{h22}\right)\left(n_{h11}+n_{h21}\right)\left(n_{h12}+n_{h22}\right)/n_h^3\right]}\sim\chi^2_1 \tag{12}$$

对式(12)做进一步变形,可得到式(13):

$$\chi^2_W=\frac{\bar{d}^2}{S^2_{\bar{d}}}=\frac{\left[\sum_h\left(n_{h11}\times n_{h22}-n_{h21}\times n_{h12}\right)/n_h\right]^2}{\sum_h\left[\left(n_{h11}+n_{h12}\right)\left(n_{h21}+n_{h22}\right)\left(n_{h11}+n_{h21}\right)\left(n_{h12}+n_{h22}\right)/n_h^3\right]}\sim\chi^2_1 \tag{13}$$

比较式(3)与式(13)可知,它们是完全相同的。

【说明】“隐含权重的加权χ^2检验统计量”实际上是在原本有“权重”的式(11)的基础上,将g×2×2表资料中“各层原始数据以及行合计和列合计”代入公式中的有关变量并进行变形后的“结果或形式”。在本质上,只有一个“突显权重的加权χ^2检验统计量”。

2.2 高维表资料CMH χ^2检验的具体算法

文献[1-2]介绍了广义CMH χ^2检验统计量及其三种变形。下面再介绍一种类似于加权χ^2检验统计量的CMH χ^2检验统计量[3],见式(14):

$$\chi^2_{CMH}=\frac{\left(\sum_{h=1}^{q}n_{h11}-\sum_{h=1}^{q}m_{h11}\right)^2}{\sum_{h=1}^{q}v_{h11}}\sim\chi^2_1 \tag{14}$$

在式(14)中,n_{h11}、m_{h11}和v_{h11}分别为第h层2×2表资料中第(1,1)格上的“观察频数”“期望频数或理论频数”和“方差”,后两项的计算分别见式(15)、式(16):

$$E\left(n_{h11}|H_0\right)=\frac{n_{h1\cdot}\times n_{h\cdot1}}{n_h}=m_{h11} \tag{15}$$

$$V\left(n_{h11}|H_0\right)=\frac{n_{h1\cdot}\times n_{h2\cdot}\times n_{h\cdot1}\times n_{h\cdot2}}{n_h^2\left(n_h-1\right)}=v_{h11} \tag{16}$$

在上面两式中,H_0为该假设检验的无效假设或称为零假设,其具体表述如下。

H_0:在各层2×2表资料中,行、列两变量间互相独立。

将式(15)和式(16)代入式(14)中后再变形,得到式(17):

$$\chi^2_{CMH}=\frac{\left[\sum_h\left(n_{h11}\times n_{h22}-n_{h21}\times n_{h12}\right)/n_h\right]^2}{\sum_h\left[\left(n_{h11}+n_{h12}\right)\left(n_{h21}+n_{h22}\right)\left(n_{h11}+n_{h21}\right)\left(n_{h12}+n_{h22}\right)/n_h^2\left(n_h-1\right)\right]} \tag{17}$$

对式(13)与式(17)进行比较:二者存在微小的差别,即分母上分别为“n_h^3”与“$n_h^2\left(n_h-1\right)$”,故可以认为:$\chi^2_W\approx\chi^2_{CMH}$。

2.3 高维表资料独立性检验的SAS实现

2.3.1 问题与数据

【例1】文献[5]提供了如下资料,试分析新疗法与旧疗法的治愈率是否相等。见表2。

表2 新疗法与旧疗法对某疾病的效果

组 别	新疗法组例数		旧疗法组例数	
	治愈	未愈	治愈	未愈
成人	32	8	49	21
儿童	40	40	12	18

2.3.2 多项研究中两关键变量之间独立性检验的SAS实现

【例2】沿用例1中的“问题与数据”,通常设“组别”为“分层因素”,研究者关心的是“治疗方法”与“治疗结果”之间是否存在关联性。与其等价的表述或假设是:H_0:“治疗方法”与“治疗结果”之间互相独立;H_1:“治疗方法”与“治疗结果”之间不独立。试基于表2资料,检验前面给出的“检验假设”。

【分析与解答】

解法一,采用加权χ^2检验。设所需要的SAS程序如下:

```
data abc;
do k=1 to 2;
input a b c d;
n=a+b+c+d; e=a+b; f=c+d; g=a+c; h=b+d;
d1=(a*d-b*c)/n;i=c+d;d2=e*f/n; pd=e*f*g*h;
v1=pd/n**3;k2=pd/(n-1)/n**2;j=a*f;k=c*e;
l=j/n;m=k/n;
output;
end;
cards;
32 8 49 21
40 40 12 18
;
run;
proc means noprint;
var d1 d2 v1 k2 l m;
output out=aaa sum=d1 d2 v1 k2 l m; run;
data a2; set aaa;
v2=d2**2; md=d1/d2; v=v1/v2; k1=d1**2;
wchisq=round(md**2/v,0.001); wp=1-probchi
```

```
(wchisq,1);
    rr=round(l/m, 0.001); mhchisq=round(k1/k2, 0.001);
    mhp=1-probchi(mhchisq,1);
    file print;
    put #3 @5'W-chisq=' wchisq @35'W-p=' wp
    #6 @5'RR=' rr @15'MH-chisq=' mhchisq @35'MH-p=' mhp;
    run;
```

【SAS输出结果及解释】

W-chisq=2.153　　W-p=0.1422916183

以上输出的结果是：χ^2_W=2.153，P=0.142292。

【统计结论】由以上输出结果可知：χ^2_W=2.153，P=0.142292>0.05，说明“治疗方法”与“治疗结果”之间的关联性无统计学意义。

【专业结论】在消除年龄因素的影响之后，可以认为：新疗法与旧疗法对应的治愈率相等（说明“治疗方法”与“治疗结果”之间互相独立）。

解法二，采用CMH χ^2检验。设所需要的SAS程序如下：

```
data a;
do age=1 to 2;
do treat=1 to 2;
do x=1 to 2;
input f @@;
output;
end;
cards;
32  8
49  21
40  40
12  18
;
run;
proc freq;
tables age*treat*x/cmh;
weight f;
run;
```

【SAS输出结果及解释】

Cochran-Mantel-Haenszel统计量（基于表评分）

统计量	备择假设	自由度	值	概率
1	非零相关	1	2.1334	0.1441
2	行评分均值差异	1	2.1334	0.1441
3	一般关联	1	2.1334	0.1441

以上输出的结果是：χ^2_{CMH}=2.1334，P=0.1441。

【统计结论】由以上输出结果可知：χ^2_{CMH}=2.1334，P=0.1441>0.05，说明“治疗方法”与“治疗结果”之间的关联性无统计学意义。

【专业结论】在消除年龄因素的影响之后，可以认为：新疗法与旧疗法的治愈率相等。

3　讨论与小结

3.1　讨论

本文所介绍的统计分析方法主要适用于g×2×2表资料，而不适用于g×R×C表资料（R与C中至少有一个大于2）；本法的优点是适用面较宽，即不论分层后的2×2表资料来自何种设计类型，均可使用；检验假设可笼统表述为：在分层后的各2×2表资料中，H_0：“原因变量”与“结果变量”之间互相独立，H_1：“原因变量”与“结果变量”之间不独立；从公式推导的最终结果可知，对于前述的“检验假设”而言，加权χ^2检验统计量的数值与CMH χ^2检验统计量的数值接近相等。

3.2　小结

本文针对g×2×2表资料独立性检验问题，呈现了两种不同形式的加权χ^2检验公式和CMH χ^2检验公式，通过公式推导和变形，揭示出两种不同形式的加权χ^2检验公式是完全相同的；同时，还揭示出加权χ^2检验统计量与CMH χ^2检验统计量在数值上是接近相等的。通过一个实例，展示了基于SAS软件实现加权χ^2检验和CMH χ^2检验的全过程，并对SAS输出结果进行解释，做出统计结论和专业结论。

参考文献

[1] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 1109-1204, 2997-3216, 6007-6303, 7991-8092.

[2] 胡纯严，胡良平. 如何正确运用χ^2检验——三种R×C列联表资料的CMH χ^2检验[J]. 四川精神卫生，2021，34(2)：121-125.

[3] 胡良平. 医学统计学运用三型理论分析定量与定性资料[M]. 北京：人民军医出版社，2009：325-334.

[4] 金丕焕. 医用统计方法[M]. 上海：上海医科大学出版社，1993：172-177.

[5] 胡良平. 现代统计学与SAS应用[M]. 北京：军事医学科学出版社，1996：191-197.

[6] 胡良平. 现代医学统计学[M]. 北京：科学出版社，2020：276-285.

[7] 约瑟夫·阿德勒. R语言核心技术手册[M]. 2版. 刘思喆，李舰，陈钢，等译. 北京：电子工业出版社，2014：410-416.

（收稿日期：2021-07-19）

（本文编辑：戴浩然）