如何正确运用χ^2检验——三种双向无序二维列联表资料的χ^2检验

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】 本文目的是介绍除四格表资料之外的三种双向无序二维列联表资料的χ^2检验以及SAS与R软件实现的方法。三种双向无序二维列联表是指双向无序的“R×2”“2×C”和“R×C”（R与C均大于2）列联表。一般来说，分析它们的目的都是检验列联表中两属性变量是否独立，在满足特定前提条件时，可以选用的简单统计分析方法为χ^2检验；在不满足特定前提条件时，应当选择计算量非常大的Fisher's精确检验法。

【关键词】 二维列联表；前提条件；独立性；χ^2检验；SAS软件；R软件
中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20210316003

How to use χ^2 test correctly—— χ^2 tests for the data collected from the three kinds of two dimensional contingency tables without ordinal variables in two directions

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 χ^2 test and SAS and R software implementation of three kinds of two-way unordered two-dimensional contingency table data except the four-fold table data. The three kinds of the tables referred to the two-way unordered “R×2”“2×C” and “R×C”（both R and C were greater than 2）contingency tables. Generally speaking, the purpose of analyzing them was to test whether the two attribute variables in the contingency table were independent. When certain prerequisites were met, the simple statistical analysis method that could be used was the χ^2 test, when the specific prerequisites were not met, the Fisher's exact test, which required the large amount of calculation, should be selected.

【**Keywords**】 Two-dimensional contingency table; Prerequisites; Independence; χ^2 test; SAS software; R software

在整理临床资料时，若结果变量为定性变量，为了使资料看起来简洁、直观、明了，人们常采取一种特定的表达形式，即“列联表”。若原因变量只有1个，就可以将其整理成“二维列联表资料”；若原因变量有2个，就可以将其整理成“三维列联表资料”；依此类推，若原因变量有k-1个时，就可以将其整理成“k维列联表资料”。一般来说，当k>3时，人们很少采用列联表形式呈现资料，而借用表达结果变量为定量变量时的形式，其常被称为“数据库”格式，即每行代表一个个体的全部信息，而每列代表一个变量（包括一般变量、原因变量和结果变量）及其全部取值。本文着重介绍除四格表资料之外的其他三种双向无序二维列联表资料的形式、适于进行χ^2检验的前提条件、基于SAS和R软件实现统计分析的方法。

1 三种双向无序二维列联表资料的实例

1.1 二维列联表资料

所谓“二维列联表资料”，就是将两个变量（通常都是定性的）分别放置在表格的“横向”与“纵向”（称为两个“维度”）上，设它们分别有“R”个与“C”个水平，它们的全部水平组合数就有“R×C”个，于是，就很自然地将表格（可将其视为一个“二维平面”）划分成“R×C”个“网格”，根据每个个体同时在两个变量上的取值情况，就可以将其归类到某一个“网格”之中，数出各“网格”内的个体数（称其为“频数”），此时的资料就被称为“二维列联表资料”。由此可知，“四格表资料”就是“二维列联表资料”的特例。

1.2 双向无序二维列联表资料的表达模式

设危险因素A有R个水平，结果变量B有C个不同取值，则双向无序二维列联表资料（常简称为$R\times C$表资料）的表达模式见表1。

表1 双向无序$R\times C$表资料的表达模式

危险因素A	例数				
结果变量B：	B_1	B_2	…	Bc	合计
A_1	n_{11}	n_{12}	…	n_{1C}	$n_{1.}$
A_2	n_{21}	n_{22}	…	n_{2C}	$n_{2.}$
…	…	…	…	…	…
A_R	n_{R1}	n_{R2}	…	n_{RC}	$n_{R.}$
合计	$n_{.1}$	$n_{.2}$	…	$n_{.c}$	N

注：危险因素A与结果变量B各有R与C个可能取值，并且它们都是名义变量，如血型变量可取A、B、AB、O几种值，季节变量可取春、夏、秋、冬几个值

在表1中，当R=2、C>2，就简称为“2×C表资料”；当R>2、C=2，就简称为“R×2表资料”；所以，它们都是“$R\times C$表资料”的特例。

1.3 实例

1.3.1 双向无序R×2表资料的实例

【例1】文献[1]中有一个双向无序R×2表资料，见表2。

表2 影响青少年焦虑症状的一个可疑因素的调查结果

第一次了解此次疫情的途径	例数		
焦虑与否：	非焦虑	焦虑	合计
亲朋好友或老师	162	61	223
电视	192	48	240
网站	109	33	142
社交软件（微信、QQ等）	98	35	133
社交平台（微博、知乎等）	147	86	233
杂志	1	1	2
合计	709	264	973

1.3.2 双向无序2×C表资料的实例

【例2】文献[2]中有一个双向无序2×C表资料，见表3。

表3 患者与家属对疾病的主要知晓途径调查结果

组别	例数					
途径：	A	B	C	D	E	合计
患者	444	9	13	6	9	481
家属	446	10	16	2	7	481
合计	890	19	29	8	16	962

注：“途径”代表对疾病的主要知晓途径；A、B、C、D、E分别代表“就诊或住院的医院”“社区医生”“社区宣传”“媒体网络”和“其他”

1.3.3 双向无序R×C表资料的实例

【例3】文献[3]中有一个双向无序R×C表资料（说明：不考虑“时间”的有序性），见表4。

表4 基层精防医护人员K6评定结果

条目	例数				
时间：	所有时间	大部分时间	少部分时间	偶尔	无
A	6	12	28	62	36
B	2	3	6	20	113
C	2	6	16	42	78
D	1	3	10	31	99
E	2	2	9	43	88
F	2	3	3	20	116

注：K6代表“凯斯勒心理困扰量表”的英文缩写；A、B、C、D、E、F分别代表“感到紧张”“感到没有希望”“感到烦躁不安”“感到太沮丧、无法愉快起来”“感到做每一件事情都很费力”和“感到无价值”

1.4 统计分析方法的选择

对于上述呈现的三种双向无序二维列联表资料而言，一般来说，其分析目的是相同的，即检验“两属性变量之间是否独立”。与此分析目的对应的统计分析方法为“χ^2检验”（注意：选用此方法时，资料需满足特定的前提条件，见下文）和“Fisher's精确检验”。

2 双向无序R×C表资料的独立性检验

2.1 检验方法概述

2.1.1 检验假设

H_0：两属性变量之间互相独立；

H_1：两属性变量之间存在关联性。

设置显著性水平为：α=0.05。

2.1.2 检验统计量

Pearson's χ^2检验[4]的检验统计量见下式：

$$\chi^2_P=\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{\left(O_{ij}-T_{ij}\right)^2}{T_{ij}},\ df=(R-1)(C-1) \quad (1)$$

在上式中，O_{ij}、T_{ij}分别代表第(i,j)网格上的观察频数与理论频数；基于概率论中条件概率的计算原理，理论频数T_{ij}可按下式计算：

$$T_{ij}=\frac{\text{第}i\text{行合计值}\times\text{第}j\text{列合计值}}{N} \quad (2)$$

在式(1)中，χ^2_P为服从自由度为df=$(R-1)(C-1)$的χ^2分布。

若基于式(1)进行计算，首先需要计算各网格上的“理论频数”，显然，利用手工计算是极不方便

的。事实上，将式(2)代入式(1)后经过变形，可以推导出不依赖“理论频数”的计算公式，见式(3)：

$$\chi_P^2=N\left(\sum_{j=1}^{C}\sum_{i=1}^{R}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}}-1\right),\ df=(R-1)(C-1) \quad (3)$$

在式(3)中，χ_P^2为服从自由度为$df=(R-1)(C-1)$的χ^2分布；“N”为列联表的总频数、“n_{ij}”为第(i,j)网格中的观测频数、“$n_{i\cdot}$”与“$n_{\cdot j}$”分别代表列联表中第i行与第j列上的“合计值”。

2.1.3 前提条件

$R\times C$表资料χ^2检验的前提条件是理论频数不宜太小，如果太小则有可能产生偏性。关于理论频数小的“界限”说法不一，例如，Cochran将理论频数太小的界定为：有1/5的格子中理论频数小于5，或至少有一个格子中理论频数小于1；Roscoe和Byars认为：若当设定α=0.05时，平均理论频数[即$N/(R\times C)$]小于6，属于理论频数太小；若当设定α=0.01时，平均理论频数小于10，属于理论频数太小[5]。

如果理论频数太小，可采取以下方法进行处理：①增大样本含量，以达到增大理论频数的目的，此属首选方法；②删除理论频数太小的格子所对应的行或列，但此方法会不可避免地损失信息及样本的随机性；③合并相邻的行或列，此法仅当行变量或列变量是有确定顺序的情形才可以。

【说明】当可选用χ^2检验时，也可选用Fisher's精确检验；一般来说，当列联表中理论频数小于5的格子数超过了总格子数的1/5时，应尽可能选用Fisher's精确检验(因篇幅所限，此法的计算原理和公式从略)。

2.2 基于SAS软件实现计算

2.2.1 分析例1资料

【例4】沿用例1的资料，试检验两属性变量之间是否存在独立性。

【分析与解答】回答两属性变量之间是否存在独立性的统计分析方法可选用χ^2检验，设所需要的SAS程序如下[6]：

```
DATA a;
DO A=1 TO 6;
DO B=1 TO 2;
INPUT F @@; OUTPUT;
END;
END;
CARDS;
162 61
192 48
109 33
98 35
147 86
1 1
;
RUN;
PROC FREQ data=a;
WEIGHT F;
TABLES A*B/CHISQ;
RUN;
```

【说明】若希望采用Fisher's精确检验，需在前述的FREQ过程步中的“TABLES语句”之后增加一句，即“exact fisher;”。

【SAS输出结果及解释】

统计量	自由度	值	概率
卡方	5	19.1096	0.0018
似然比卡方检验	5	18.7972	0.0021
Mantel-Haenszel卡方	1	8.7891	0.0030
Phi系数		0.1401	
列联系数		0.1388	
Cramer V		0.1401	

【统计学结论】因χ_P^2=19.1096，P=0.0018，说明应拒绝“H_0：两属性变量之间互相独立”，接受“H_1：两属性变量之间存在关联性”。

【专业结论】第一次了解此次疫情的途径不同，发生焦虑的比例是不同的，具体地说，发生焦虑比例从高到低依次为“杂志”>“社交平台”>“亲朋好友或老师”>“社交软件”>“网站”>“电视”。值得一提的是：“杂志”所对应的样本含量为2，样本过小，抽样误差很大，应将该行数据删除重新计算。

重新计算的结果为：χ_P^2=18.5998，P=0.0009，结论同上，即发生焦虑比例从高到低依次为“社交平台”>“亲朋好友或老师”>“社交软件”>“网站”>“电视”。

2.2.2 分析例2资料

【例5】沿用例2的资料，试检验两属性变量之间是否存在独立性。

【分析与解答】回答两属性变量之间是否存在独立性的统计分析方法可选用χ^2检验，设所需要的

SAS程序如下：

```
DATA b;
DO A=1 TO 2;
DO B=1 TO 5;
INPUT F @@; OUTPUT;
END;
END;
CARDS;
444 9 13 6 9
446 10 16 2 7
;
RUN;
PROC FREQ data=b;
WEIGHT F;
TABLES A*B/CHISQ;
RUN;
```

【SAS输出结果及解释】

χ^2_P=2.6175，P=0.6237，说明“患者与家属对疾病的主要知晓途径基本相同”。

【说明】因篇幅所限，用SAS软件分析例3数据从略，读者可借助前面的SAS程序自己去完成，但需要修改原始数据的“行数”与“列数”（位于SAS程序中两个“DO语句”）。

2.3 基于R软件实现计算

【例6】沿用例3的资料，试检验两属性变量之间是否存在独立性。

【分析与解答】回答两属性变量之间是否存在独立性的统计分析方法可选用χ^2检验，设所需要的R程序如下[7-8]：

```
r1<- c(6,12,28,62,36)
r2<- c(2,3,6,20,113)
r3<- c(2,6,16,42,78)
r4<- c(1,3,10,31,99)
r5<- c(2,2,9,43,88)
r6<- c(2,3,3,20,116)
chisq.test(rbind(r1,r2,r3,r4,r5,r6))
```

【R输出结果及解释】

```
        Pearson's Chi-squared test
data: rbind(r1, r2, r3, r4, r5, r6)
X-squared = 139.12, df = 20, p-value < 2.2e-16
        Warning message:
In chisq.test(rbind(r1, r2, r3, r4, r5, r6)) :
        Chi-squared近似算法有可能不准
```

结果中，χ^2_P=139.12，P<0.0001，说明基层精防医护人员所涉及的“条目”与“时间”之间不独立，即存在一定程度的关联性。具体地说，从表4最后3列的各列中发现有3个最大值，分别为“28”“62”和“116”，这3个数对应的“行数”分别为“A”“A”和“F”，也就是说，基层精防医护人员会在“少部分时间”或“偶尔”“感到紧张”；从来不会（指表4中最后一列的纵标目“无”）“感到无价值”。

遗憾的是：R软件给出了“警告信息”，表明此资料不符合进行χ^2检验的前提条件（具体地说，资料中小于5的理论频数的个数超过了总格子数的1/5），故χ^2近似算法给出的结果可能不准确。

将上面R程序中的最后一行换成下面的语句，就可对此资料进行Fisher's精确检验：

```
fisher.test(rbind(r1,r2,r3,r4,r5,r6),simulate.p.value=TRUE)
```

【说明】R软件中采用蒙特卡罗（monte carlo）模拟计算方法。

【R输出结果及解释】

```
Fisher's Exact Test for Count Data with simulated
        p-value (based on 2000 replicates)
data: rbind(r1, r2, r3, r4, r5, r6)
        p-value = 0.0004998
alternative hypothesis: two.sided
```

以上输出内容表明：经过2000次重复模拟计算，得到P=0.0004998，说明该列联表资料中两个属性变量之间不独立，结论同上，此处不再赘述。

3 讨论与小结

3.1 讨论

本文所介绍的三种双向无序列联表资料的检验假设是相同的，具体内容见“第2.1.1节”；当检验结果为拒绝“H_0”、接受“H_1”时，其统计结论也是相同的，即“表中两属性变量之间不独立”，更确切地表述为：“各行或各列上的频数分布规律是不同的”。在实践中，人们可能会提出更具体的要求，例如，究竟哪些行或列上的频数分布规律是不同的、哪些行或列上的频数分布规律是相同的？对这个问题的回答涉及两方面的内容，其一，列联表资料的多重比较问题[9]；其二，χ^2值的非精确分割与精确分割问题[10-11]。因篇幅所限，此处从略。

若用SAS软件对本文中例3资料进行“Fisher's精确检验”，计算的时间可能需要数个小时；但采用

R 软件进行计算（见本文例 6），大约只需要 2 秒钟。之所以会产生如此大差距的根本原因是两种统计软件所采用的计算方法不同，前者是基于超几何分布原理推算出来的计算方法；而后者采用的是蒙特卡罗模拟计算方法。

3.2　小结

本文呈现了三种双向无序二维列联表资料的实例，介绍了基于 Pearson's χ^2 检验的计算原理和采用 SAS 与 R 软件实现独立性检验的具体方法和结果解释。最后，提出了两个与“χ^2 检验”有关且有待进一步讨论的问题。

参考文献

[1] 郭鹏飞，李欣，刘帅，等. 新冠肺炎流行期间安徽省青少年焦虑现状及影响因素[J]. 四川精神卫生，2020，33(6)：501-505.

[2] 易海，杨琼花，杜育如，等. 重性精神疾病社区管理服务背景下湛江地区患者及家属对疾病知晓情况的分析[J]. 四川精神卫生，2020，33(1)：57-60.

[3] 范箬馨，杨先梅，黄明金，等. 新冠肺炎疫情防控中基层精防医护人员心理健康状况及需求调查[J]. 四川精神卫生，2020，33(3)：207-210.

[4] 黄志宏，方积乾. 数理统计分析[M]. 北京：人民卫生出版社，1987：116-124.

[5] 胡良平. 正确实施科研设计与统计分析[M]. 北京：人民军医出版社，2011：274.

[6] SAS Institute Inc. SAS/STAT®15. 1 user's guide[M]. Cary，NC：SAS Institute Inc，2018：2997-3216.

[7] 胡良平. 现代医学统计学[M]. 北京：科学出版社，2020：265-275.

[8] 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.

[9] 胡良平. 科研设计与统计分析[M]. 北京：军事医学科学出版社，2012：346-348.

[10] 金丕焕. 医用统计方法[M]. 上海：上海医科大学出版社，1993：170-187.

[11] 胡良平. 现代统计学与 SAS 应用[M]. 北京：军事医学科学出版社，1996：164-190.

（收稿日期：2021-03-16）

（本文编辑：陈　霞）

科研方法专题策划人——胡良平教授简介

胡良平，男，1955 年 8 月出生，教授，博士生导师，曾任军事医学科学院研究生部医学统计学教研室主任和生物医学统计学咨询中心主任、国际一般系统论研究会中国分会概率统计系统专业理事会常务理事、中国生物医学统计学会副会长、北京大学口腔医学院客座教授和《中华医学杂志》等 10 余种杂志编委；现任世界中医药学会联合会临床科研统计学专业委员会会长、国家食品药品监督管理局评审专家和 3 种医学杂志编委；主编统计学专著 48 部、参编统计学专著 10 部；发表第一作者和通信作者学术论文 300 余篇、发表合作论文 130 余篇；获军队科技成果和省部级科技成果多项；参加并完成三项国家标准的撰写工作、参加三项国家科技重大专项课题研究工作。在从事统计学工作的 30 年中，为几千名研究生、医学科研人员、临床医生和杂志编辑讲授生物医学统计学，在全国各地作统计学学术报告 100 余场，举办数十期全国统计学培训班，培养 20 多名统计学专业硕士和博士研究生。近几年来，参加国家级新药和医疗器械项目评审数十项、参加 100 多项全军重大重点课题的统计学检查工作。归纳并提炼出有利于透过现象看本质的“八性”和“八思维”的统计学思想，独创了逆向统计学教学法和三型理论。擅长于科研课题的研究设计、复杂科研资料的统计分析和 SAS 与 R 软件实现、各种层次的统计学教学培训和咨询工作。