

如何正确运用 χ^2 检验——横断面设计 四格表资料的 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】 本文目的是介绍横断面设计四格表资料的 χ^2 检验及SAS和R软件实现。具体介绍了3种方法, 即“Pearson's χ^2 检验”“校正的 Pearson's χ^2 检验”和“似然比 χ^2 检验”。内容涉及前述提及的3种统计分析方法的检验假设、计算原理与公式、前提条件、SAS与R软件的实现、结果解释以及结论陈述。

【关键词】 横断面设计; 四格表资料; χ^2 检验; 似然比; SAS软件; R软件

中图分类号: R195.1

文献标识码: A

doi: 10.11886/scjsws20210115004

How to use χ^2 test correctly—— χ^2 test for the data of four-fold tables collected from the cross-sectional design

Hu Chunyan¹, 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: lphu927@163.com)

【Abstract】 The purpose of this article was to introduce the χ^2 test for the data of four-fold tables collected from the cross-sectional design and the realization of SAS and R software. Specifically, three methods were introduced, namely "Pearson's χ^2 test", "corrected Pearson's χ^2 test" and "likelihood ratio χ^2 test". The contents involved the test hypotheses, calculation principles and formulas, prerequisites, implementation of SAS and R software, interpretation of results and statement of conclusions of the aforementioned three statistical analysis methods.

【Keywords】 Cross-sectional design; Four-fold table data; χ^2 test; Likelihood ratio; SAS software; R software

在生物医学和流行病学研究中, 基于横断面设计收集的四格表资料是常见的^[1-2]。根据资料所满足的前提条件不同, 处理四格表资料的统计分析方法有多种, 即“Pearson's χ^2 检验及其精确检验”“校正的 Pearson's χ^2 检验”“似然比 χ^2 检验及其精确检验”和“Fisher's 精确检验”^[3], 因篇幅所限, 本文介绍“Pearson's χ^2 检验”“校正的 Pearson's χ^2 检验”“似然比 χ^2 检验”3种统计分析方法。

1 横断面设计四格表资料的实例与模式

1.1 实例

【例1】在文献[1]中, 根据MINI 5.0中文版自杀模块判断未治疗抑郁障碍患者有无自杀风险, 共6个条目, 评分范围0~33分。评分<6分为无自杀风险, 评分≥6分为有自杀风险。再将两组受试者分别按“性别”“婚姻状况”“有无精神障碍家族史”和“是否伴有精神病性症状”进行划分, 可得4个“横断面

设计四格表资料”, 现将其中一个列在表1中。问: 未治疗抑郁障碍患者“有无自杀风险”与“婚姻状况”之间是否存在关联性?

表1 未治疗抑郁障碍患者“有无自杀风险”与“婚姻状况”之间关系的调查结果

有无自杀风险	例数			
	婚姻状况:	已婚	未婚	合计
有		13	39	52
无		33	32	65
合计		46	71	117

【统计分析方法的选择】两个属性变量(本例为“有无自杀风险”与“婚姻状况”)之间是否存在关联性, 可选用Pearson's χ^2 检验、似然比 χ^2 检验等方法。由于统计学家依据不同的原理和条件推导出了不同的 χ^2 检验统计量, 但它们通常都有一个共同的“检验假设”, 即“两变量之间互相独立或无关联性”; 而且, 所有的 χ^2 检验统计量都必须服从 χ^2 分布。

【例2】在文献[2]中, 为了研究“中山市流动人

口与户籍人口自杀率和自杀方式特征”,研究考察了“户籍状况(分为‘户籍组’与‘流动组’)”与“性别”“民族”“受教育程度”“婚姻状况”和“职业”5个属性变量之间的关系,可得5个列联表。现将其中一个列在表2中。问:中山市2015年-2017年户籍组和流动组自杀案例“户籍状况”与“民族”之间是否存在关联性?

表2 中山市2015年-2017年户籍组和流动组自杀案例
“户籍状况”与“民族”之间关系的调查结果

户籍状况	例数		
	民族:	汉族	其他
户籍组		175	2
流动组		200	14
合计		375	16

【统计分析方法的选择】回答两个属性变量(本例为“户籍状况”与“民族”)之间是否存在关联性,可选用Pearson's χ^2 检验,似然比 χ^2 检验等方法。理由前面已述及,此处从略。

1.2 模式

为了便于采用简单且易于理解的形式呈现横断面设计四格表资料的计算公式,特将表达横断面设计四格表资料的列联表抽象成下面的两种模式,分别见表3和表4。

表3 同时按属性变量A与B对受试对象进行分组计数的模式一

属性变量A	例数		
	属性变量B:	B ₁	B ₂
A ₁		a	b
A ₂		c	d
合计		g=a+c	h=b+d

注:利用该表中的符号便于呈现分析四格表资料的Pearson's χ^2 检验统计量及其校正公式

表4 同时按属性变量A与B对受试对象进行分组计数的模式二

属性变量A	例数		
	属性变量B:	B ₁	B ₂
A ₁		n ₁₁	n ₁₂
A ₂		n ₂₁	n ₂₂
合计		n _{·1}	n _{·2}

注:利用该表中的符号便于呈现分析四格表资料的似然比 χ^2 检验统计量

2 横断面设计四格表资料独立性或关联性分析的原理与计算公式

2.1 独立性与关联性

对于横断面设计四格表资料而言,通常假定两个属性变量之间的地位是平等的且并存的。也就

是说,不认为其中一个是原因变量、另一个是结果变量;也不认为其中一个先出现、另一个后出现。研究者关心的是两个属性变量之间是否存在“关联性”,而与其对立的是“独立性”。在“两属性变量之间具有独立性”的假设成立的条件下,当基于实际数据和统计学原理计算的结果认为应拒绝此假设,则认为“两属性变量之间具有关联性”。于是,其检验假设可表述如下:

H_0 :两属性变量之间相互独立; H_1 :两属性变量之间不独立或有关联性。

2.2 基于独立性假设推导出计算理论频数的公式

基于表3中设定的符号,假定“属性变量A”与“属性变量B”之间互相独立,则依据概率论中计算条件概率的公式可推算出与4个观察频数“a、b、c、d”对应的理论频数“ $T_a = T_{(1,1)}$ 、 $T_b = T_{(1,2)}$ 、 $T_c = T_{(2,1)}$ 和 $T_d = T_{(2,2)}$ ”如下^[4-5]:

$$T_a = T_{(1,1)} = \frac{e \times g}{n} = \frac{(a+b)(a+c)}{a+b+c+d} \quad (1)$$

$$T_b = T_{(1,2)} = \frac{e \times h}{n} = \frac{(a+b)(b+d)}{a+b+c+d} \quad (2)$$

$$T_c = T_{(2,1)} = \frac{f \times g}{n} = \frac{(c+d)(a+c)}{a+b+c+d} \quad (3)$$

$$T_d = T_{(2,2)} = \frac{f \times h}{n} = \frac{(c+d)(b+d)}{a+b+c+d} \quad (4)$$

2.3 基于独立性假设的Pearson's χ^2 检验统计量

K. Pearson于十九世纪末提出用检验统计量来测量二维列联表的各网格中观测频数(O)与理论频数(T)之间的“差异程度”或“吻合度”^[6]。如下式:

$$\chi_p^2 = \sum_{i=1}^R \sum_{j=1}^C \frac{(O_{ij} - T_{ij})^2}{T_{ij}}, \quad df = (R-1)(C-1) \quad (5)$$

在式(5)中, O_{ij} 、 T_{ij} 分别为第(i,j)网格中的“观测频数”与“理论频数”,R、C分别代表二维列联表的“行数”与“列数”,K. Pearson证明了式(5)定义的 χ_p^2 近似地服从自由度为 $df = (R-1)(C-1)$ 的 χ^2 分布。

将表3中设定的符号和式(1)~式(4)代入式(5),可得到分析横断面设计四格表资料的专用公式,见式(6)。

$$\chi_p^2 = \frac{n(ad-bc)^2}{efgh} = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}, df = 1 \quad (6)$$

由式(6)定义的 χ_p^2 渐近地服从自由度为 1 的 χ^2 分布。

【前提条件】当四格表资料中 $n > 40$ 且没有 T (理论频数) < 5 时, 可以选用“Pearson's χ^2 检验”。

2.4 基于独立性假设的校正 Pearson's χ^2 检验统计量

在四格表资料中, 只要有一个网格上的理论频数 < 5 , 若直接采用式(6)进行计算, 其结果都会明显偏离 χ^2 分布, 故需要采用下式进行校正。

$$\chi_{cp}^2 = \frac{n(|ad - bc| - 0.5n)^2}{(a+b)(c+d)(a+c)(b+d)},$$

$$df = 1 \quad (7)$$

由式(7)定义的 χ_{cp}^2 渐近地服从自由度为 1 的 χ^2 分布。

【前提条件】当四格表资料中 $n > 40$, 但有 $1 < T$ (理论频数) < 5 时, 可以选用“校正 Pearson's χ^2 检验”。

2.5 基于独立性假设的对数似然比 χ^2 检验统计量

2.5.1 似然比统计量

利用表 4 中设定的符号, 依据观测的结果, 可得任何一个个体落入四格表第 (i, j) 网格中的概率为 $P_{ij} = \frac{n_{ij}}{n}$, $i, j = 1, 2$; 给定假设, H_0 : 两属性变量之间互相独立、 H_1 : 两属性变量之间不独立, 即有关联性, 则可得任何一个个体落入四格表第 (i, j) 网格中的概率为 $Q_{ij} = \frac{T_{ij}}{n}$, $i, j = 1, 2$ 。其中, T_{ij} 为基于“独立性假设”推算的第 (i, j) 网格中的理论频数, 其计算方法见式(1)~式(4)。于是, 基于概率乘法原理, 可获得前述提及的两个条件下的似然函数近似表达式如下:

情形 1: 依据观测的结果, 得似然函数近似表达式:

$$L(P_{ij}, i, j = 1, 2) \propto \prod_{i=1}^2 \prod_{j=1}^2 (P_{ij})^{n_{ij}} \quad (8)$$

情形 2: 依据假设的结果, 得似然函数近似表达式:

$$L(Q_{ij}, i, j = 1, 2) \propto \prod_{i=1}^2 \prod_{j=1}^2 (Q_{ij})^{n_{ij}} \quad (9)$$

上面两式中的“ $\propto$ ”符号代表“呈正比”(例如, A 与 B 呈正比, 可表示为 $A \propto B$)。

用式(8)除以式(9), 得到似然比统计量如下:

$$\lambda = \frac{\prod_{i=1}^2 \prod_{j=1}^2 (P_{ij})^{n_{ij}}}{\prod_{i=1}^2 \prod_{j=1}^2 (Q_{ij})^{n_{ij}}} \quad (10)$$

2.5.2 对数似然比 χ^2 检验统计量

在文献[3, 7]中, 介绍了另一种检验两属性变量关联性或独立性的方法, 叫做“似然比 χ^2 检验”。但实际上, 其检验统计量是对所构造的“似然比统计量[见式(10)]”取对数后乘以 2 的结果。因此, 其真实的名称应该叫做“对数似然比 χ^2 检验统计量”。于是, 对式(10)取自然对数后乘以 2, 可得下式:

$$\chi_{LR}^2 = 2 \sum_{i=1}^2 \sum_{j=1}^2 n_{ij} \ln \frac{n_{ij}}{T_{ij}}, df = 1 \quad (11)$$

在式(11)中, 由 χ_{LR}^2 所定义的检验统计量渐近地服从自由度为 1 的 χ^2 分布。其中, n_{ij} 、 T_{ij} 分别代表四格表中第 (i, j) 网格上的观察频数与理论频数, $i, j = 1, 2$ 。

【前提条件】当四格表资料中, 无论是观察频数还是理论频数中都没有出现“零”时, 可以选用“对数似然比 χ^2 检验”。

3 横断面设计四格表资料关联性或独立性分析的软件实现

3.1 基于 SAS 实现关联性或独立性分析

沿用例 1 中的“问题与数据”, 试进行关联性或独立性分析。设所需要的 SAS 程序如下^[3]:

```
data a;
do a=1 to 2;
do b=1 to 2;
input f @@;
output;
end;
end;
cards;
13 39
33 32
;
run;
proc freq;
weight f;
tables a*b / chisq;
run;
```

【SAS 主要输出结果及解释】

统计量	自由度	值	概率
卡方	1	8.0406	0.0046
似然比卡方检验	1	8.2365	0.0041
连续调整卡方	1	6.9968	0.0082

第 1 行与第 3 行分别为“Pearson's χ^2 检验”与“校正 Pearson's χ^2 检验”的结果,第 2 行为“似然比 χ^2 检验(严格地说,为对数似然比 χ^2 检验)”的结果。3 种检验方法所得的 P 值都小于 0.01,都拒绝“ H_0 : 有自杀风险与婚姻状况之间互相独立”,接受“ H_1 : 有自杀风险与婚姻状况之间存在关联性”。

【结论】因为“有自杀风险”的“已婚比例为 $(13/52) \times 100\% = 25.00\%$ ”,而“无自杀风险”的“已婚比例为 $(33/65) \times 100\% = 50.77\%$ ”,这提示“有自杀风险”的未治疗抑郁障碍患者的结婚比例明显低于“无自杀风险”的未治疗抑郁障碍患者的结婚比例。

3.2 基于 R 实现关联性或独立性分析

设所需要的 R 程序如下^[8-9]:

```
>rownum<- c(13,39)
>colnum<- c(33,32)
>chisq.test(rbind(rownum,colnum),correct=
FALSE)
>chisq.test(rbind(rownum,colnum))
```

【程序说明】“>”代表 R 软件运行环境中的“提示符”,上文的 R 程序中共有 4 个提示符,说明共有 4 个 R 语句;第一句组织第一个“行向量”,代表四格表资料的第 1 行数据;第二句组织第二个“行向量”,代表四格表资料的第 2 行数据;第三句和第四句都是调用函数“chisq.test()”进行 χ^2 检验,其区别在于:选项“correct=FALSE”代表不进行“校正”,而该选项的默认值为“校正”,等价于写“correct=TRUE”;而选项“rbind()”是将两个行向量合并在一起,即构成两行两列的四格表数据。

【R 主要输出结果及解释】

```
Pearson's Chi-squared test
data:rbind(rownum,colnum)
X-squared=8.0406,df=1,p-value=0.004574
以上为未进行连续性校正的  $\chi^2$  检验结果。
Pearson's Chi-squared test with Yates' continuity
correction
data:rbind(rownum,colnum)
X-squared=6.9968,df=1,p-value=0.008166
```

以上为进行了连续性校正的 χ^2 检验结果。

以上 R 输出结果与 SAS 输出的结果是相同的(见上文 SAS 输出结果中的第 1 行与第 3 行)。

【结论】参见前文 SAS 输出结果及结论,此处从略。

4 讨论与小结

4.1 讨论

4.1.1 横断面设计的概念

所谓横断面设计,就是依据研究目的确定了一个研究的总体,在一个时点或时段上,研究者对该总体中的全部个体(属于全面调查或普查)或按某种规则随机抽取的部分个体进行调查或测量。这种做法就被称为“横断面研究设计”,简称为“横断面设计”。

4.1.2 横断面设计四格表资料的特点

在表 1 和表 2 资料中,都有两个“属性变量”,通常,研究者同时按两个属性变量的水平对受试对象进行分组计数,而且,两个属性变量都是从受试对象身上“提取”出来的,换句话说,属性变量是受试对象“自带的”属性(例如性别、民族);在多数场合下,两个属性变量之间不存在明显的“原因变量”与“结果变量”之分,也就是说,它们之间更常见的是“相互关系”,而不是“因果关系”。

若一个变量为“药物种类(药物 A 与药物 B)(研究者施加给受试对象的因素)”,另一个变量为“治疗效果(‘有效’与‘无效’)”,此时,前者可视为“原因变量”,后者可视为“结果变量”。一般来说,从患者“服药”到出现“疗效”的时间间隔不长时,可视为一个“时点或时段”,这样的四格表资料仍可被称为“横断面设计四格表资料”。然而,当前述所指的“时间间隔”很长时,就不适合叫做“横断面设计四格表资料”了。

值得注意的是,基于“药物种类”与“治疗效果”而形成的四格表资料与流行病学上常见的“队列设计四格表资料(例如调查‘是否吸烟’与‘是否患肺癌’之间关系所得的资料)”是十分相似的。二者的区别在于以下两点:其一,原因变量的来源不同。“药物种类”是研究者施加给受试对象的影响因素,而“吸烟与否”是受试对象自身的行为习惯;其二,获得结果变量取值的时间长度不同,通常,“服药”与“疗效”之间时间间隔不会很长;而“吸烟与否”与“患肺癌与否”之间时间间隔通常会比较长。

4.1.3 横断面设计四格表资料统计分析方法的合理选择

在检验横断面设计四格表资料两属性变量之间“独立性假设”的方法中,除了本文介绍的3种方法(运用时,请检查前提条件是否满足)之外,还有Fisher's精确检验、Pearson's χ^2 检验的精确算法、似然比 χ^2 检验的精确算法^[3],因篇幅所限,此处从略。

在运用SAS/STAT的FREQ过程时,若使用“exact chisq/mc;”语句,输出结果中会呈现4种精确检验的结果^[3],其中,除了前面提及的3种精确检验结果之外,还给出了“Mantel-Haenszel's χ^2 检验及其精确检验结果”。然而,这种检验方法适用于“队列设计”或“病例对照设计”四格表资料,其对应的检验假设为:“ $H_0: RR = 1, H_1: RR \neq 1$ ”(RR为相对危险度)或“ $H_0: OR = 1, H_1: OR \neq 1$ ”(OR为优势比)^[10],因这些内容与本文的目的不吻合,故未做介绍。

4.2 小结

本文呈现了横断面设计四格表资料的实例和模式,详细给出了3种检验横断面设计四格表资料“关联性或独立性”假设的检验方法的原理和计算公式;基于SAS和R软件实现了对横断面设计四格表资料的3种假设检验,对软件输出的结果做出了解释,并给出了统计和专业结论。

参考文献

- [1] 徐海婷,刘嫣然,吕婧,等.未治疗抑郁障碍患者自杀风险与认知情绪调节策略的关系[J].四川精神卫生,2020,33(1):44-48.
- [2] 邓兰芳,余金鸣,黄彩英,等.中山市流动人口与户籍人口自杀率和自杀方式特征分析[J].四川精神卫生,2020,33(1):67-70,75.
- [3] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 2997-3216.
- [4] 茆诗松,程依明,濮晓龙.概率论与数理统计教程[M].北京:高等教育出版社,2009:38-47.
- [5] 魏宗舒.概率论与数理统计教程[M].北京:高等教育出版社,2008:32-40.
- [6] 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.
- [7] 伊冯·M·毕晓普,斯蒂芬·E·芬伯格,保罗·W·霍兰德.离散多元分析理论与实践[M].张尧庭,译.北京:中国统计出版社,1998:437-451.
- [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.现代医学统计学[M].北京:科学出版社,2020:244-257.
- [10] 颜虹.医学统计学[M].北京:人民卫生出版社,2008:95-100,163-166.

(收稿日期:2021-01-15)

(本文编辑:戴浩然)