

如何正确运用 χ^2 检验——两种单向有序二维列联表资料线性趋势 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】 本文目的是介绍两种单向有序二维列联表资料线性趋势 χ^2 检验以及 SAS 与 R 软件实现的方法。两种单向有序二维列联表是指原因变量为 R 值有序变量的“R×2 列联表”和结果变量为 C 值有序变量的“2×C 列联表”, 对它们进行线性趋势检验分别需要采用 Cochran-Armitage's 线性趋势 χ^2 检验和 Lee's 线性趋势 χ^2 检验。

【关键词】 二维列联表; 线性趋势检验; 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/scjsws20210316004

How to use χ^2 test correctly—— χ^2 tests for the data collected from the two kinds of two dimensional contingency tables with an ordinal variable in one direction

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 paper was to introduce the linear trend χ^2 tests for the data of the two kinds of one-way ordinal two-dimensional contingency tables and the implementation of SAS and R software. The two one-way ordered two-dimensional contingency tables referred to the “R×2 contingency table” with R values ordinal variable as the cause variable and the “2×C contingency table” with C values ordinal variable as the outcome variable. Their linear trend tests required the use of the Cochran-Armitage's linear trend χ^2 test and the Lee's linear trend χ^2 test, respectively.

【Keywords】 Two-dimensional contingency table; Linear trend test; χ^2 test; SAS software; R software

对于列联表资料而言, 最常见的统计分析目的是“独立性检验(其本质仍属于差异性检验)”。但针对某些特殊的列联表资料, 其分析目的也可以是“相关分析(例如, 可用 Kendall's Tau-b 秩相关分析处理配对设计扩大形式的二维列联表资料; 可用 Spearman's 秩相关分析处理双向有序二维列联表资料)”, 还可以是“线性趋势检验”(例如, 可分别用 Cochran-Armitage's 线性趋势检验和 Lee's 线性趋势检验处理单向有序二维列联表资料)。本文将介绍两种单向有序二维列联表资料及其线性趋势检验的原理和软件实现的方法。

1 两种单向有序二维列联表资料的实例与统计分析方法选择

1.1 单向有序 R×2 表资料的实例

【例 1】文献[1]中有一个单向有序的 R×2 表资

料, 见表 1。

表 1 家庭月收入(元)与有无自杀意念之间关系的调查结果

家庭月收入(元)	例 数			合计
	有无自杀意念:	有	无	
>10000		58	53	111
5000~10000		164	115	279
3000~5000		263	128	364
1000~3000		173	106	279
<1000		57	45	102
合计		715	447	1162

注: “家庭月收入(元)”是有序原因变量; “有无自杀意念”是二值结果变量

1.2 单向有序 2×C 表资料的实例

【例 2】文献[2]中有一个单向有序的 2×C 表资料, 见表 2。

表2 治疗4周末两组临床疗效观测结果

组别	例数					合计
	疗效:	痊愈	显著进步	进步	无效	
研究组	20	13	7	5	45	
对照组	12	9	10	11	42	
合计	32	22	17	16	87	

注：“组别”是原因变量；“疗效”是多值有序结果变量

1.3 统计分析的目的与统计分析方法的选择

1.3.1 以表1资料为分析对象

若分析目的是考察“家庭月收入”与“有无自杀意念”之间是否独立,需要选择 χ^2 检验或 Fisher’s 精确检验;若分析目的是比较5种家庭月收入的受试对象“有自杀意念的发生率”之间的差异是否具有统计学意义,仍需要选择 χ^2 检验或 Fisher’s 精确检验;若分析目的是考察“有自杀意念的发生率”是否随着“家庭月收入的减少”而呈线性递增或递减趋势,则需选择 Cochran-Armitage’s 线性趋势 χ^2 检验。

【说明】对于表1资料而言,上面描述的前两个分析目的在本质上是完全相同的,故可选用的统计分析方法相同。因篇幅所限,本文只进行 Cochran-Armitage’s 线性趋势 χ^2 检验。

1.3.2 以表2资料为分析对象

若分析目的是考察“组别”与“疗效”之间是否独立,需要选择 χ^2 检验或 Fisher’s 精确检验;若分析目的是比较两组疗效之间的差异是否有统计学意义,需要选择秩和检验;若分析目的是考察研究组在四种“疗效”等级上的“相对比例”从“痊愈”到“无效”是否呈线性递减或递增趋势,需要选择 Lee’s 线性趋势 χ^2 检验。

【说明】因篇幅所限,本文只进行 Lee’s 线性趋势 χ^2 检验。

2 单向有序 R×2 表资料的线性趋势检验

2.1 单向有序 R×2 表资料的表达模式

单向有序 R×2 表资料的表达模式见表3。在表3中,假定“B₁”代表“阳性”、“B₂”代表“阴性”,则各行上都可以计算出一个“阳性率 $\pi_i(i=1,2,\dots,R)$ ”;设 $P = \frac{n_{+1}}{n}$,则 P 就是第1列的“阳性率”,或称为各行上阳性率的平均率。

表3 R×2 列联表资料的表达模式

因素 A	例数			合计
	指标 B:	B ₁	B ₂	
A ₁		n ₁₁	n ₁₂	n ₁₊
A ₂	
...	
A _R		n _{R1}	n _{R2}	n _{R+}
合计		n ₊₁	n ₊₂	n

注:因素 A 是原因变量,指标 B 为二值结果变量;n₁₊和 n₊₁分别表示第一行例数之和与第一列例数之和,其他符号含义类似

2.2 检验方法概述

2.2.1 检验假设

$$H_0: \pi_1 = \pi_2 = \dots = \pi_R;$$

$$H_{1a}: \pi_1 < \pi_2 < \dots < \pi_R \text{ 或 } H_{1b}: \pi_1 > \pi_2 > \dots > \pi_R.$$

设置显著性水平为: $\alpha = 0.05$ 。

2.2.2 检验统计量

Cochran-Armitage’s χ^2 检验的检验统计量^[3]见下式:

$$z = \frac{\sum_{i=1}^R n_{i1}(x_i - \bar{x})}{\sqrt{P(1-P) \sum_{i=1}^R n_{i+}(x_i - \bar{x})^2}} \quad (1)$$

$$\text{式(1)中, } P = \frac{n_{+1}}{n}, \bar{x} = \frac{\sum_{i=1}^R n_{i+}x_i}{n}, x_i (i=1, 2, \dots, R)$$

代表第 i 行上因素 A 的取值(假定因素 A 为数值型变量或有序变量);Z 渐近地服从标准正态分布,由 χ^2 分布的定义^[4]可知,式(1)中的“Z”平方为服从自由度为1的 χ^2 分布,即有下式成立:

$$Z^2 = \chi_1^2 \quad (2)$$

2.3 基于 SAS 软件实现计算

【例3】沿用例1的资料,试基于 SAS 软件检验“有自杀意念的发生率”是否随着“家庭月收入的减少”而呈线性递增或递减趋势。

【分析与解答】为了回答所提出的问题,可选用 Cochran-Armitage’s χ^2 检验,设所需要的 SAS 程序如下:

```
data a;
do a=1 to 5;
do b=1 to 2;
input f @@;
output;
end;
```

```
end;
cards;
58 53
164 115
263 128
173 106
57 45
;
run;
proc freq data=a;
tables a*b/trend scores=table;
weight f;
exact trend;
run;
```

【程序说明】“tables 语句”中的选项“trend”要求采用近似方法进行线性趋势检验；而“exact trend;”语句则要求采用精确方法进行线性趋势检验。

【SAS 输出结果及解释】

Cochran-Armitage 趋势检验	
统计量(Z)	0.9895
渐近检验	
单侧 Pr>Z	0.1612
双侧 Pr> Z	0.3224
精确检验	
单侧 Pr>=Z	0.1681
双侧 Pr>= Z	0.3249

以上结果表明:SAS 给出的检验统计量为 Z(它服从标准正态分布),但 Z 的平方就是自由度为 1 的 χ^2 检验统计量的值;无论采用渐近检验还是精确检验,也无论选用单侧检验还是双侧检验,对应的 P 值均大于 0.05,即“有自杀意念的发生率”不会随着“家庭月收入的减少”而呈线性递增或递减趋势。

2.4 基于 R 软件实现计算

【例 4】沿用例 1 的资料,试基于 R 软件检验“有自杀意念的发生率”是否随着“家庭月收入的减少”而呈线性递增或递减趋势。

【分析与解答】为了回答所提出的问题,可选用 Cochran-Armitage's χ^2 检验,设所需要的 R 程序如下:

```
> x<- c(58,164,263,173,57)
> n<- c(111,279,354,279,102)
> prop.trend.test(x,n)
```

【程序说明】第 1 行输入表 1 资料的第 1 列数

据;第 2 行输入表 1 资料的纵向合计列数据;实现 Cochran-Armitage's χ^2 检验的 R 函数为“prop.trend.test()”。

【R 输出结果及解释】

```
Chi-squared Test for Trend in Proportions
data: x out of n
using scores: 1 2 3 4 5
```

X-squared = 1.0414, df = 1, p-value = 0.3075

以上结果表明: $\chi^2=1.0414, P=0.3075$ (注意:此计算结果与 SAS 输出结果略有差别),结论同上(参见第 2.3 节),此处从略。

3 单向有序 2×C 表资料的线性趋势检验

3.1 单向有序 2×C 表资料的表达模式

单向有序 2×C 表资料的表达模式见表 4。

表 4 结果变量为有序变量 2×C 列联表资料的表达模式

因素 A	例 数					合计
	指标 B:	B ₁	B ₂	...	B _C	
A ₁		n ₁₁	n _{1C}	n ₁₊
A ₂		n ₂₁	n _{2C}	n ₂₊
合计		n ₊₁	n _{+C}	n

注:因素 A 是二值原因变量,指标 B 为多值有序的结果变量,通常由定量指标分段而形成

3.2 检验方法概述

3.2.1 检验假设

$$H_0: \pi_1 = \pi_2 = \dots = \pi_C;$$

$$H_{1a}: \pi_1 < \pi_2 < \dots < \pi_C \text{ 或 } H_{1b}: \pi_1 > \pi_2 > \dots > \pi_C.$$

设置显著性水平为: $\alpha = 0.05$ 。

【说明】 $\pi_j (j=1, 2, \dots, C)$ 代表第 j 列上的总体的发生率; H_{1a} 代表各列上的“总体率”呈线性递增变化趋势; H_{1b} 代表各列上的“总体率”呈线性递减变化趋势。

3.2.2 检验统计量

Lee's 线性趋势检验统计量^[5]见下式:

$$z_{linear} = \frac{\sum_{j=1}^C x_j (y_{1j} - n_{1+} w_j)}{\sqrt{n_{1+} \left[\sum_{j=1}^C w_j (x_j - \bar{x})^2 \right]}} \quad (3)$$

这里 $\bar{x} = \sum_{j=1}^C w_j x_j, w_j = \frac{n_{+j}}{n} (j = 1, 2, \dots, C)$ (w_j 为第 j 列上的权重系数), $x_j (j=1, 2, \dots, C)$ 代表第 j 列上指标 B 的取值(假定指标 B 为数值型变量或有序变

量)。如果用下标序号 j 代替公式中的 x_j (通常为各组段上的组中值), 相当于秩和检验中的编秩, 则该统计量称为单调趋势检验统计量 $Z_{monotone}$ 。这些统计量都渐近服从标准正态分布。

由 χ^2 分布的定义^[4]可知, 上式中的“ Z_{Linear} ”平方为服从自由度为 1 的 χ^2 分布, 即有下式成立:

$$\chi_1^2 = Z_{Linear}^2 \quad (4)$$

3.3 基于 SAS 软件实现计算

【例 4】沿用例 2 的资料, 试基于 SAS 软件检验研究组在四种“疗效”等级上的“相对比例”从“痊愈”到“无效”是否呈线性递减或递增趋势。

【分析与解答】为了回答所提出的问题, 可选用 Lee's χ^2 检验, 设所需要的 SAS 程序如下:

```
data abc;
%let n=87;
%let n1=45;
input x y1 y;
w=y/&n;
num=x*(y1-&n1*w);
cards;
1 20 32
2 13 22
3 7 17
4 5 16
;
run;
proc univariate data=abc noprint;
var num;
output out=aaa sum=sum_num;
run;
proc univariate data=abc noprint;
weight w;
var x;
output out=bbb css=css_x;
run;
data abc;
merge aaa bbb;
nwx=sqrt(&n1*css_x);
z=sum_num/nwx;
if Z>=0 then P=1-PROBNORM(Z);
else if Z<0 then P=PROBNORM(Z);
file print;
put #2 @10 'Z value' @30 'P Value';
put #4 @10 z @30 p;
```

run;

【程序说明】“ $n=87$ ”代表表 2 资料中的总频数; “ $n1=45$ ”代表表 2 资料中第 1 行的合计频数; “cards 语句”后的 3 列数据分别是: 第 1 列为 4 个疗效等级的“分值”; 第 2 列为表 2 资料中的第 1 行频数; 第 3 列为表 2 资料中的横向合计行上的 4 个频数。

【SAS 输出结果及解释】

Z	P
-1.565746313	0.0587040096

结果表明: $Z=-1.565746$ ($\chi^2=2.452$), $P=0.058704$, 说明研究组在四种“疗效”等级上的“相对比例”从“痊愈”到“无效”不呈线性递减或递增趋势。

3.4 基于 R 软件实现计算

【例 5】沿用例 2 的资料, 试基于 R 软件检验研究组在四种“疗效”等级上的“相对比例”从“痊愈”到“无效”是否呈线性递减或递增趋势。

【分析与解答】为了回答所提出的问题, 可选用 Lee's χ^2 检验, 设所需要的 R 程序^[6-7]如下:

```
x<- c(1,2,3,4)
r1<- c(20,13,7,5)
r2<- c(12,9,10,11)
r3<- r1+r2
n<- sum(r3)
w<- r3/n
n1<- sum(r1)
nw<- n1*w
fz<- sum(x*(r1-nw))
xb<- sum(w*x)/sum(w)
wx<- sum(w*(x-xb)^2)
fm<- sqrt(n1*wx)
z<- fz/fm
if (z<=0) p=pnorm(z, mean=0, sd=1) else p=1-
pnorm(z, mean=0, sd=1)
z
p
```

【程序说明】以上程序各语句之前都省略了 R 软件提示符“>”; 第 1 行为 4 个疗效等级的“分值”; 第 2 行为表 2 资料中的第 1 行频数; 第 3 行为表 2 资料中的第 2 行频数。

【R 输出结果及解释】

Z	P
-1.565746	0.05870401

结果表明: $Z=-1.565746(\chi^2=2.452)$, $P=0.058704$,说明研究组在四种“疗效”等级上的“相对比例”从“痊愈”到“无效”不呈线性递减或递增趋势。

4 讨论与小结

4.1 讨论

单向有序 $R \times 2$ 表资料的线性趋势检验(简称“前者”)与单向有序 $2 \times C$ 表资料的线性趋势检验(简称“后者”)在以下两个方面存在区别。其一,变量的性质及其水平数不同:前者的“原因变量”为“ R 值有序变量($R > 2$)”、“结果变量”为“二值变量”;而后的“原因变量”为“二值变量”、“结果变量”为“ C 值有序变量($C > 2$)”。其二,构建线性趋势检验统计量的统计学原理不同。前者是利用加权回归分析的思想^[8],将由 $R \times 2$ 列联表资料计算得到的总 χ^2 值分解为“线性回归分量A”和“偏离线性回归分量B”两部分。若A有统计学意义、B无统计学意义,说明原因变量与结果变量之间存在线性关系;若A与B都有统计学意义,说明原因变量与结果变量之间可能存在某种非线性关系。而后者是基于多项分布原理进行推导,构造出检验统计量^[5,8],因篇幅所限,此处从略。

尽管SAS软件在给出Cochran-Armitage's线性趋势检验结果时,呈现了单侧检验和双侧检验两种结果。但由于其备择假设有两种可能情况,而且,在一个实际问题中,只能选择其一(递增趋势或递减趋势),故基于常识可知,线性趋势检验更适合选择“单侧检验”,而不是双侧检验。

值得一提的是:在对以上两种列联表资料进行线性趋势检验时,都涉及到如何给有序变量的各水

平进行赋值。一般来说,直接赋值“1、2、3、……”即可;若有专业知识为依据,可给有序变量的各水平赋值为非连续的自然数,例如1、3、8、15、27、……对于同一个资料不同的赋值方法,所得的计算结果会略有差别,但一般不会明显改变最终的结论。

4.2 小结

本文呈现了两种单向有序二维列联表资料的实例和模式,给出了对其进行线性趋势检验的原理和计算公式;基于SAS和R软件实现了统计计算,对统计软件的输出结果进行了解释,并做出了统计结论和专业结论。

参考文献

- [1] 缪楹,徐晓津,周勇杰,等.社会支持对重性抑郁障碍患者自杀意念的影响[J].四川精神卫生,2020,33(2):142-145.
- [2] 徐彩霞,汤超华,俞洋,等.高频rTMS联合度洛西汀对青年首发广泛性焦虑障碍的疗效及认知功能的影响[J].四川精神卫生,2020,33(4):326-329.
- [3] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 2997-3216.
- [4] 方开泰,许建伦.统计分布[M].北京:科学出版社,1987:136-211.
- [5] Armitage P, Colton T. Encyclopedia of biostatistics[M]. 2nd. John Wiley & Sons, 2005: 5971-5978.
- [6] 胡良平.现代医学统计学[M].北京:科学出版社,2020:258-264.
- [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.
- [8] 胡良平.面向问题的统计学——(1)科研设计与统计分析[M].北京:军事医学科学出版社,2012:483-499,513-525.

(收稿日期:2021-03-16)

(本文编辑:陈霞)