

· 科研方法专题 ·

如何正确运用 χ^2 检验——病例对照设计 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】 本文目的是介绍病例对照设计四格表资料的 χ^2 检验、优势比的假设检验和区间估计以及使用 SAS 与 R 软件实现的方法。针对病例对照设计的特点, 重点介绍优势比的概念、假设检验和区间估计方法。对 SAS 与 R 软件计算的结果作出解释, 并给出统计结论和专业结论。

【关键词】 病例对照设计; χ^2 检验; 优势比; 置信区间; SAS 软件; R 软件

中图分类号: R195.1

文献标识码: A

doi: 10.11886/scjsws20210316002

How to use χ^2 test correctly—— χ^2 test for the data of four-fold tables collected from the case-control 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 paper was to introduce the χ^2 test for the data of four-fold tables collected from the case-control design, the hypothesis test and interval estimation of odds ratio, and the implementation of SAS and R software. In view of the characteristics of case-control design, the concept of odds ratio, hypothesis testing and interval estimation were emphasized. It was given that the interpretation of the results calculated by SAS and R software, as well as the statistical conclusion and professional.

【Keywords】 Case-control design; χ^2 test; Odds ratio; Confidence interval; SAS software; R software

在流行病学研究或临床研究中, 为了较快地找到可能导致某种疾病发生的原因, 研究者需要开展“病例对照研究”。也就是找到与病例组条件接近的对照组, 基于临床专业知识, 提出各种可能导致所研究疾病发生的全部可疑因素, 回顾性调查病例组与对照组受试对象接触各种可疑因素的情况, 计算并比较两组受试对象接触各种可疑因素的比例, 从而为确定导致所研究疾病发生的可能危险因素提供线索。

若假定每个可疑危险因素只分为“接触”与“未接触”两个水平, 此时, 基于病例对照设计所收集的资料就可以简化成“病例对照设计四格表资料(参见下文表 1 和表 2)”。对其进行统计分析主要包括两个方面: 其一, 检验“患病与否”与“接触与否”之间是否存在关联性(所需要的统计分析方法与处理横断面设计四格表资料的统计分析方法相同, 参见文献[1]); 其二, 检验“优势比(Odds ratio, OR)是否

等于 1”。本文着重介绍优势比的概念、对总体 OR 的 Mantel-Haenszel's χ^2 检验(简称 MH χ^2 检验)、对总体 OR 的区间估计方法以及使用 SAS 和 R 软件实现计算的方法。

1 病例对照设计的概念及其四格表资料的实例

1.1 概念

所谓“病例对照设计”, 就是依据专业知识和基本常识, 提出一些“重要的非试验因素(如年龄、性别、职业、生活环境、生活方式等)”, 针对现有“病例组”具备的条件, 寻找在所述重要的非试验因素方面与“病例组”非常接近但未患所研究疾病的另一组人群, 他们被称为“对照组”; 再依据专业知识, 提出一些可能导致所研究疾病发生的“可疑影响因素”, 让前述提及的两组受试对象回忆之前是否接

触过这些“可疑影响因素”。于是,计算出两组人群对每种“可疑影响因素”的“接触比例”,进而依据统计学原理对其进行假设检验。

1.2 病例对照设计四格表资料的表达模式

从前面的“概念”可知,在病例对照设计资料的收集过程中,先有“患病”与“未患病”的结果分档,后有“接触危险因素(简称‘接触’)”与“未接触危险因素(简称‘未接触’)”的原因分组,故其四格表资料的表达模式见表1。

表1 n个受试对象病例对照研究结果的表达模式

危险因素A	例 数		
	结果变量B: 患病	未患病	合计
接触	a	b	e=a+b
未接触	c	d	f=c+d
合计	g=a+c	h=b+d	n=a+b+c+d

注:a、b、c、d分别代表“患病组接触危险因素人数”“未患病组接触危险因素人数”“患病组未接触危险因素人数”和“未患病组未接触危险因素人数”;e、f分别为“第1行合计人数”与“第2行合计人数”;g、h分别为“第1列合计人数”与“第2列合计人数”;n代表四格表中的“总人数”

1.3 实例

【例1】在文献[2]中,作者根据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。假定作者收集资料时,首先关注的是“是否伴有精神病性症状”,并将其视为“结果变量”。在此基础上,对“有精神病性症状”与“无精神病性症状”的两组受试对象进行回顾性调查或测量,从而获得两组受试对象中各自“有自杀风险者所占的比例”,具体数据见表2。

表2 未治疗抑郁障碍患者“有无精神病性症状”与“有无自杀风险”之间关系的调查结果

有无自杀风险	例 数		合计
	有无精神病性症状: 有	无	
有	11	41	52
无	4	61	65
合计	15	102	117

注:假定收集资料时,先将全部受试对象按“有无精神病性症状”分为“第一组,有,15例”“第二组,无,102例”;再对两组受试对象进行回顾性调查,按“有无自杀风险”分为“第一组,有,11例;无,4例”“第二组,有,41例;无,61例”

对表2资料进行统计分析涉及以下两个问题:①“有无精神病性症状”与“有无自杀风险”之间是否存在独立性;②“有精神病性症状”与“无精神病性症状”的优势比是否等于1。

【统计分析方法的选择】回答第1个问题,可选用“Pearson’s χ^2 检验”或“校正的Pearson’s χ^2 检验”或“似然比 χ^2 检验”或“Fisher’s精确检验”。因为这些检验方法所对应的零假设均为“两属性变量互相独立”;回答第2个问题,需要先计算“OR”,然后选用“Mantel-Haenszel’s χ^2 检验”,该检验的检验假设为:“ $H_0:OR=1, H_1:OR \neq 1$ ”。因为MH χ^2 检验所对应的零假设为“优势比等于1”。

1.4 独立性检验方法

虽然病例对照设计四格表资料中的两个变量有“原因变量”与“结果变量”之分,但在回答两变量之间是否存在“独立性”时,仍可将其视为“两属性变量”之间关系的研究问题,故可采用的统计分析方法与横断面设计四格表资料是完全相同的,具体方法参见文献[1]。因篇幅所限,此处不再赘述。

2 优势比的概念、假设检验与区间估计

2.1 概念

OR是两个人群odd值之比^[3-4],这两个人群分别受到一个可疑危险因素两个水平(例如接触与未接触)的影响。具体地说,OR是病例组的“ $odd_{\text{病例}}$ ”与对照组的“ $odd_{\text{对照}}$ ”之比,它揭示了“危险因素的两个水平”在两组中的相对作用强度大小。若 $OR>1$,表明“危险因素”对“病例组”的影响大于对“对照组”的影响;反之亦然。OR的计算公式如下:

$$OR = \frac{odd_{\text{病例}}}{odd_{\text{对照}}} = \frac{(a/g)/(c/g)}{(b/h)/(d/h)} = \frac{a/c}{b/d} = \frac{ad}{bc} \quad (1)$$

前提条件:两个人群应处于相同的地理环境之中,生活在相同的时间区间内;每个人群都有足够大的样本含量(两个人群的样本含量之差越小越好);对照组在一切重要的非试验因素方面与病例组越接近越好;应避免带有诱导性的提问;应尽可能保证在回顾性调查过程中,每位受试对象通过回忆给出的回答越准确越好。

2.2 假设检验

OR是一个一般的样本统计量,存在抽样误差,要想知道总体中的OR是否为1,就需要对其进行假设检验,即

$$H_0:OR=1; H_1:OR \neq 1 \quad (2)$$

对式(2)进行假设检验所需要的检验统计量为 Mantel-Haenszel's χ^2 , 一般记为 χ^2_{MH} , 见下式:

$$\chi^2_{MH} = \frac{(n-1)(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad (3)$$

2.3 区间估计

2.3.1 概述

由于 OR 是一个一般的样本统计量, 通常, 人们需要知道与其对应的总体参数所在的范围, 这就是总体优势比的区间估计问题。在 SAS/STAT 的 FREQ 过程中^[5], 给出了三种置信区间的计算方法, 即“渐近置信区间法(简称 WALD 法)”“评分置信区间法”和“精确置信区间法”^[4-5]。

2.3.2 渐近置信区间法

OR 的渐近 $100(1-\alpha)\%$ 置信区间的计算公式:

$$\left[OR \times \exp(-z\sqrt{v}), OR \times \exp(z\sqrt{v}) \right] \quad (4)$$

在式(4)中, z 是标准正态分布曲线下左侧面积为 $100(1-\alpha/2)\%$ 处横坐标轴上的分位数; 而 v 为对数优势比的方差, 见下式:

$$v = Var(\ln OR) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \quad (5)$$

2.3.3 评分置信区间法

Miettinen 和 Nurminen 提出的关于优势比 OR 的评分置信限是通过对 OR 插入评分检验来计算的^[5]。

设 $\theta = OR = \frac{ad}{bc}$, 在“ $H_0: \theta = 1$ ”成立的条件下, 基于评分的 χ^2 检验统计量见下式:

$$Q(\theta) = \frac{n-1}{n} \left[e \cdot (\hat{P}_1 - \tilde{P}_2) \right]^2 \left[\frac{1}{e \cdot \tilde{P}_1 (1 - \tilde{P}_1)} + \frac{1}{f \cdot \tilde{P}_2 (1 - \tilde{P}_2)} \right]^{-1} \quad (6)$$

在上式中, 第 1 项“ $\frac{n-1}{n}$ ”为偏移校正因子; “ e ”与“ f ”分别代表四格表资料的第 1 行、第 2 行上的合计值; \hat{P}_1 是第 1 行上观测到的风险(即比例); 在限定“ $\theta = OR = \frac{ad}{bc}$ ”的条件下, \tilde{P}_1 与 \tilde{P}_2 分别是第 1 行与第 2 行上风险的最大似然估计值。

由上式定义的检验统计量 $Q(\theta)$ 可以构造出下面的接受域:

$$\left[\theta: Q(\theta) < \chi^2_{1(1-\alpha)} \right] \quad (7)$$

满足式(7)的所有 θ 值构成 OR 的 $100(1-\alpha)\%$ 置信区间。

在式(7)中, $\chi^2_{1(1-\alpha)}$ 是具有一个自由度的 χ^2 分布的 $100(1-\alpha)\%$ 分位数。SAS 中的 FREQ 过程通过迭代计算求取 OR 的置信限。

在限定“ $\theta = OR = \frac{ad}{bc}$ ”的条件下, P_1 与 P_2 的最大似然估计可按下式计算:

$$\tilde{P}_2 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad (8)$$

$$\tilde{P}_1 = \frac{\tilde{P}_2 \theta}{1 + \tilde{P}_2 (\theta - 1)} \quad (9)$$

在上面两式中, $a = f \cdot (\theta - 1)$; $b = e \cdot \theta + f - \hat{P}_{\cdot 1} (\theta - 1)$; $c = -\hat{P}_{\cdot 1}$ 。其中, “ e ”与“ f ”分别代表四格表资料的第 1 行、第 2 行上的合计值。

2.3.4 精确置信区间法

基于 Thomas 和 Gart 提出的方法^[5], 用迭代的方法求解下列两个方程, 可以获得置信限的下限和上限值 ϕ_1 和 ϕ_2 :

$$\frac{\sum_{i=a}^g \binom{e}{i} \binom{f}{g-i} \phi_1^i}{\sum_{i=0}^g \binom{e}{i} \binom{f}{g-i} \phi_1^i} = \frac{\alpha}{2} \quad (10)$$

$$\frac{\sum_{i=0}^a \binom{e}{i} \binom{f}{g-i} \phi_2^i}{\sum_{i=0}^g \binom{e}{i} \binom{f}{g-i} \phi_2^i} = \frac{\alpha}{2} \quad (11)$$

在上面两式中, a, e, f, g 分别代表四格表资料中“第 1 行第 1 列网格上的频数”“第 1 行合计值”“第 2 行合计值”和“第 1 列合计值”; 小括号内代表“组合数”, 例如: $\binom{n}{m}$ 代表 $C_n^m = \frac{n!}{m!(n-m)!}$, 其中, “!” 读作“阶乘”, 例如: $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$ 。

3 病例对照设计四格表资料统计分析的软件实现

3.1 问题与数据

【例 2】沿用例 1 中的“问题与数据”, 试完成下列 4 项任务: ① 检验两变量(即“有无精神病性症状”与“有无自杀风险”)之间的关联性; ② 计算优势比 (OR); ③ 对 OR 进行假设检验, 即“ $H_0: OR=1, H_1: OR \neq 1$ ”; ④ 对 OR 进行区间估计。

3.2 基于 SAS 实现独立性检验、计算 OR 值及求其置信区间

所需要的 SAS 程序如下^[5-6]：

```
data a;
do a=1 to 2;
do b=1 to 2;
input f @@;
output;
end;
end;
cards;
11 41
4 61
;
run;
proc freq; /*过程步 1*/
weight f;
tables a*b / chisq or;
run;
proc freq; /*过程步 2*/
weight f;
tables a*b / or(cl=wald);
run;
proc freq; /*过程步 3*/
weight f;
tables a*b / or(cl=score);
run;
proc freq; /*过程步 4*/
weight f;
tables a*b / or(cl=exact);
exact or;
run;
```

【程序说明】第一个过程步对四格表资料进行“关联性或独立性”检验(包括 χ^2 检验、校正 χ^2_{CP} 检验、 χ^2_{LR} 检验和 Fisher's 精确检验);计算 OR 值,还包括对 OR 是否等于 1 的 χ^2_{MH} 检验。第二个过程步对四格表资料进行“OR 值计算”,并基于“渐近置信区间法(即 WALD 法)”求“OR 值的 95% 置信区间”。第三个过程步对四格表资料进行“OR 值计算”,并基于“评分法”求“OR 值的 95% 置信区间”。第四个过程步对四格表资料进行“OR 值计算”,并基于“精确置信区间法(简称精确法)”求“OR 值的 95% 置信区间”。

【SAS 主要输出结果及解释】

第一部分,“关联性或独立性”检验结果见表 3。

表 3 例 1 中病例对照设计四格表资料关联性或独立性检验结果

检验统计量或方法	值	渐近概率	精确概率
χ^2	5.8156	0.0159	-
χ^2_{CP}	4.5510	0.0329	-
χ^2_{LR}	5.8966	0.0152	-
Fisher's 法	-	-	0.0242

第二部分,优势比的计算结果及其假设检验的结果:OR=4.0915, χ^2_{MH} =5.7659, 渐近概率=0.0163。

第三部分,基于三种方法估计总体优势比 OR 值的 95% 置信区间的结果如下。渐近法:[1.2190, 13.7329];评分法:[1.2704, 13.0550];精确法:[1.1022, 18.6182]。

【结论】由表 3 结果可知,“有无精神病性症状”与“有无自杀风险”之间是不独立的,具体地说,“有精神病性症状”者比“无精神病性症状”者具有更大的“自杀风险”,其 OR=4.0915;总体中 OR 值的 95% 置信区间随着计算方法的改变略有变化,即渐近法:[1.2190, 13.7329];评分法:[1.2704, 13.0550];精确法:[1.1022, 18.6182]。

3.3 基于 R 实现独立性检验以及 OR 值的计算、假设检验与区间估计

所需要的 R 程序如下^[6-7]：

```
>oddratio<-
matrix(c(11,41,4,61),
byrow=T,
nrow=2,
dimnames=list(wy=c("my","rg"),
qc=c("qc","wqc")))
>fisher.test(oddratio,alternative="two.sided")
>chisq.test(oddratio)
```

【程序说明】“>”代表 R 软件运行环境中的“提示符”,上面的 R 程序中共有 3 个提示符,说明共有 3 个 R 语句;第一句将四格表资料组织成一个矩阵;第二句调用函数 fisher.test(), 此函数将采用 Fisher's 精确检验实现四格表资料的“独立性检验”,同时,还计算优势比 OR 的值,并采用精确法求总体 OR 值的 95% 置信区间;第三句调用函数 chisq.test(), 此函数将采用校正公式进行 Pearson's χ^2 检验(即实现两属性变量之间的独立性检验)。

【R 主要输出结果及解释】

Fisher's Exact Test for Count Data

data: oddratio

p-value = 0.02423

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval

1. 102154 18. 612957

sample estimates:

odds ratio

4. 042456

以上结果表明,采用 Fisher's 精确检验对四格表资料中两属性变量进行独立性检验得到的结果为: $P=0.02423$,优势比 $OR=4.042456$ (注意:与前面基于 SAS 计算得出的 $OR=4.0915$ 略有出入),基于精确法求得总体 OR 的 95% 置信区间为 $[1.1022, 18.6130]$ 。

Pearson's Chi-squared test with Yates' continuity correction

data: oddratio

X-squared=4. 551, df=1, p-value=0. 0329

以上为进行了连续性校正的 Pearson's χ^2 检验结果,即 $\chi^2_{cp}=4.551, P=0.0329$ 。

【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

在 SAS/STAT 的 FREQ 过程中^[5],增加了一些新的方法求总体优势比 OR 值的置信区间,其中,精确置信限的计算方法计算得出的下限值与其他两种方法计算的下限值比较接近,而上限值比其他两种方法计算得出的上限值大得多,其原因有待进一步查证。

“病例对照研究”是一种“由果溯因”的研究,它只能为探索疾病的“因果关系”提供线索,而不能提供确凿的证据。原因主要有以下几点:其一,很难保证两组人群具有高度的可比性;其二,很难确定每位受试对象接触可疑影响因素的强度(包括作用时间长短和作用量的大小);其三,很难保证每位受试对象回忆结果的准确性;其四,很难保证真正的

危险因素全部都被研究者提出来并加以研究。

在实际科研工作中,无论基于队列研究设计还是病例对照研究设计收集的资料通常都是多因素多结局的资料,一般来说,常将多结局变量分解成一个一个的结局变量,但应该将多个原因变量同时纳入统计分析(即选用多重回归分析或其他多因素统计分析),这样不仅可以克服单因素分析的许多弊端,还有利于考察因素之间的交互作用是否对结果变量具有不可忽视的影响,使所获得的结果和结论更可靠。

4.2 小结

本文交代了病例对照设计的概念、呈现了病例对照设计四格表资料的实例和模式,给出了优势比的概念、计算公式和区间估计方法;基于 SAS 和 R 软件实现了对病例对照设计四格表资料的多种假设检验及对优势比 OR 值的计算和区间估计,对软件输出的结果做出了解释,并给出了统计结论和专业结论。

参考文献

- [1] 胡纯严,胡良平. 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[J]. 四川精神卫生, 2021, 34(1): 48-52.
- [2] 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.
- [3] 颜虹. 医学统计学[M]. 北京: 人民卫生出版社, 2005: 95-100, 163-166.
- [4] 伯纳德·罗斯纳. 生物统计学基础[M]. 孙尚拱,译. 北京: 科学出版社, 2004: 551-579.
- [5] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 2997-3216.
- [6] 胡良平. 现代医学统计学[M]. 北京: 科学出版社, 2020: 244-257.
- [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:2021-03-16)

(本文编辑:陈霞)