如何正确运用χ^2检验——配对设计四格表资料的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】 本文目的是介绍配对设计四格表资料的McNemar's χ^2检验及SAS和R软件实现。首先，提出配对设计四格表资料存在3种情形，即①特设“金标准”的配对设计四格表资料，值得进行统计分析；②缺乏“金标准”的配对设计四格表资料，不值得进行统计分析；③隐含“金标准”的配对设计四格表资料，值得进行统计分析。其次，以第1种情形的“问题与数据”为统计分析的对象，分别采用SAS与R软件进行差异性分析，得出了计算结果，对结果作了解释，并给出了统计和专业结论。

【关键词】 配对设计；四格表资料；McNemar'sχ^2检验；金标准；SAS软件；R软件
中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20210115003

How to use χ^2 test correctly——χ^2 test for the data of four-fold table collected from the matched pairs design

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*: *lphu927@163.com*）

【Abstract】 The purpose of this paper was to introduce the McNemar's χ^2 test and SAS and R software implementation of four-fold table data collected from the matched pairs design. Firstly, it was proposed that there were three situations for the data of four-fold table of the paired design, namely ①the data of paired design four-fold table with the special "gold standard" was worthy of statistical analysis; ②the data of four-fold table of the paired design without the special "gold standard" was not worthy of statistical analysis; ③the data of four-fold table collected from the matched pairs design with implicit "gold standard" was worthy of statistical analysis. Secondly, taking the "problems and data" in the first case as the object of statistical analysis, SAS and R software were used to analyze the differences, the calculation results were given and explained, and the statistical and professional conclusions were also made.

【Keywords】 Matched pairs design; Four-fold table data; McNemar's χ^2 test; Gold standard; SAS software; R software

在诊断医学研究中，研究者常需要使用两种方法（或两位评价者）测定同一批样品并按配对的方式把“二值”结果呈现出来，表达此种资料的表格叫做“配对设计四格表资料或配对设计2×2表资料”[1]。所谓“二值”结果，即检测结果只有两种，通常分为阳性（用“+”表示）与阴性（用“-”表示）。本文介绍对配对设计四格表资料进行差异性分析的基本原理和基于SAS与R软件实现统计计算的方法。

1 配对设计四格表资料的3种情形

1.1 特设“金标准”的配对设计四格表资料

【例1】设有一种能准确诊断血友病的方法（称为金标准），用它对34名血友病隐性携带的女性患者和34名健康妇女检测的结果作为标准对照，对每位受试者再用欲比较优劣的试验方法检测。两种方法对每位受试者的样品检测的结果按配对的形式整理成表1的形式[2]。问：表1资料是否值得分析？

表1 两种检测方法对同一组受试者检测的结果

试验法检测结果	例数		
金标准检测：	+	-	合计
+	31	4	35
-	3	30	33
合计	34	34	68

注：+表示检测结果阳性；-代表检测结果阴性

【解答】对于表1资料而言，有“金标准”检测方法作为标准对照，可以明确地判定试验检测方法的优劣。这种四格表资料称为特设“金标准”的配对设计四格表资料，可以对其进行统计分析。

【统计分析方法的选择】对于特设金标准的配

对设计四格表资料，有两种统计分析方法：其一，检验两种检测方法检测结果不一致部分的差别是否具有统计学意义（简称“差异性检验”），可用McNemar's χ^2检验来实现；其二，检验两种检测方法检测结果一致部分的数量是否具有统计学意义（简称“一致性检验”），可用Kappa检验（或称一致性检验）法，具体实现见文献[3]。

1.2 缺乏“金标准”的配对设计四格表资料

【例2】设有两种探针，分别叫做“生物探针”和“P探针”。用它们同时检测每份样品中是否具有某种物质，某研究者将86份样品的检测结果以表2的形式呈现出来。事实表明，任何一种探针的检测结果正确与否，是未知的。问：表2资料是否值得分析？

表2 两种探针同时检测的结果

生物探针	例数			
检测结果	P探针检测结果：	+	−	合计
+		40	4	44
−		3	39	42
合计		43	43	86

注：+表示检测结果阳性；−代表检测结果阴性

【解答】在表2资料中，两种检测方法不知何者为优，用任何一种方法检测都可能出现假阳性或假阴性结果，比较它们检测结果不一致的两个频数（或率）之间的差别是否具有统计学意义，无论统计分析结果是什么，都不能说明任何问题。也就是说，对于缺乏“金标准”的配对设计四格表资料，不值得做统计分析。

1.3 隐含“金标准”的配对设计四格表资料

【例3】假定有甲、乙两种培养基，同时用这两种培养基对同一批痰液标本进行培养，培养的结果以表3的形式呈现出来。问：表3资料是否值得分析？

表3 两种培养基对同一批痰液标本同时培养的结果

甲培养基	例数			
培养结果	乙培养基结果：	+	−	合计
+		36	34	70
−		0	135	135
合计		36	169	205

注：+表示检测结果阳性；−代表检测结果阴性

【解答】相同的痰液标本中，若甲培养基能培养出阳性结果，而乙培养基却培养出阴性结果，则表明甲培养基优于乙培养基，这种阳性结果就是“真阳性”，而不会出现假阳性。此时，比较两种培养基培养的结果不一致的样品数之间的差别是否具有统计学意义，是有价值的。这种四格表资料称为隐含“金标准”的配对设计四格表资料，值得做统计分析。

【统计分析方法的选择】有两种可供选用的统计分析方法，即“差异性检验”和“一致性检验”，详见前面“例1”的“统计分析方法的选择”部分，此处不再赘述。

2 配对设计四格表资料差异性检验

2.1 配对设计四格表资料差异性检验的原理

2.1.1 配对设计四格表资料的一般表达形式

配对设计四格表资料的一般表达形式见表4。

表4 配对设计四格表资料的表达形式

处理方法A	例数			
所得结果	*B：	+	−	合计
+		a	$b(T_b)$	$n_1=a+b$
−		$c(T_c)$	d	$n_2=c+d$
合计		$m_1=a+c$	$m_2=b+d$	$n=a+b+c+d$

注：*B代表“处理方法B所得结果”；+表示检测结果阳性；−代表检测结果阴性；a、b、c、d分别为结果的观察频数；T_b与T_c分别代表b与c的理论频数；n_1、n_2分别代表“第一行合计频数”与“第二行合计频数”；m_1、m_2分别代表“第一列合计频数”与“第二列合计频数”；n代表四格表中的“总频数”

2.1.2 配对设计四格表资料差异性检验

2.1.2.1 建立检验假设

$H_0: T_b = T_c, H_1: T_b \neq T_c, \alpha=0.05$

【说明】T_b、T_c分别代表“b”与“c”的理论频数。

2.1.2.2 构建差异性检验的检验统计量

配对设计四格表资料差异性检验的检验统计量为McNemar's χ^2检验统计量[4]，公式如下：

若$b+c\geq40$时可应用未校正的公式：

$$\chi^2 = \frac{(b-c)^2}{b+c}, df = 1 \quad (1)$$

若$b+c<40$时应用连续性校正公式：

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c}, df = 1 \quad (2)$$

以上两式定义的χ^2检验统计量均服从自由度为1的χ^2分布χ_1^2。

【说明】McNemar's χ^2精确检验方法见文献[5-6]；

McNemar's χ^2非参数检验方法见文献[7],因篇幅所限,此处从略。

2.2 配对设计四格表资料差异性检验

2.2.1 基于SAS实现差异性检验

【例4】沿用例1中的“问题与数据”,试基于SAS进行差异性检验。设所需要的SAS程序如下[6,9]:

```
data a;
do a=1 to 2;
do b=1 to 2;
input f @@;
output;
end;
end;
cards;
31 4
3 30
;
run;
proc freq;
weight f;
tables a*b / agree;
exact mcnem;
run;
```

【程序说明】McNemar's χ^2检验需要通过“tables语句”中的选项“agree”来指定;“exact语句”中的选项“mcnem”是为了求McNemar's χ^2检验的精确概率。

【SAS主要输出结果及解释】

McNemar 检验

统计量 (S)	0.1429
自由度	1
渐近 Pr > S	0.7055
精确 Pr >= S	1.0000

McNemar's χ^2=0.1429,近似的概率值P=0.7055;精确的概率值P=1.00>0.05,应接受$H_0: T_b = T_c$,即两种检测方法检测结果不一致的频数之间的差别无统计学意义。

【结论】就本例而言,试验法的假阳性例数(或率)与假阴性例数(或率)接近相等。

2.2.2 基于R实现差异性检验

设所需要的R程序如下[8-9]:

```
>Performance <-
matrix(c(31, 4, 3, 30),
nrow = 2,
dimnames = list("Test Method" = c("Positive",
"Negative"),
"Gold Standard" = c("Positive", "Negative")))
>Performance
>mcnemar.test(Performance)
```

【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;第一句的目的是创建一个名为Performance的矩阵,通过“<-”实现赋值(说明:“<-”的作用类似于“=”);第二句要求系统给出所创建的矩阵;第三句调用实现McNemar's χ^2检验的函数mcnemar.test()。圆括号内的参数Performance就是以矩阵形式呈现的配对设计四格表资料(包括横标目与纵标目以及表内部的4个频数)。

【R主要输出结果及解释】

```
Gold Standard
Test Method Positive Negative
Positive 31 3
Negative 4 30
McNemar's Chi-squared test with continuity correction
data: Performance
McNemar's chi-squared=0, df=1, p-value=1
```

第一部分结果显示出已经成功创建的矩阵,实际上就是本例中的配对设计四格表资料。

第二部分结果:χ^2=0,df=1,P=1。

【注意】用R计算所得的结果“χ^2=0”与用SAS计算所得的结果“χ^2=0.1429”不同,因为R软件中是采用校正公式(2)计算的;而SAS软件中是采用未校正公式(1)计算的。

【结论】就本例而言,试验法的假阳性例数(或率)与假阴性例数(或率)接近相等。

3 讨论与小结

3.1 讨论

配对设计四格表资料差异性检验也被称为“对称性检验”,当检验结果为P>0.05时,表明表中的“b(严格地说,应是T_b)(假阳性频数)”与“c(严格地说,应是T_c)(假阴性频数)”之间的差别无统计学意义,也可理解成这两个位置上的频数关于“主对角线(从左上角到右下角的连线)”对称。此时,很容易

误解成“试验法与金标准法检测结果之间无差别，可用试验法取代金标准法”。也就是说，对配对设计四格表资料进行McNemar's χ^2检验，只能回答试验法自身的“假阳性频数(或率)”与“假阴性频数(或率)”之间的差别是否具有统计学意义，若“$P>0.05$”，表明试验法检测出现“假阳性结果”与出现“假阴性结果”机会均等；若“$P<0.05$”，表明试验法检测出现“假阳性结果”与出现“假阴性结果”机会不均等。此时，若“$b>c$”，则表明试验法出现“假阳性结果”的概率明显大于其出现“假阴性结果”的概率，反之亦然。

若问“试验法可否取代金标准法”，则需要采用“kappa检验或称一致性检验”，当检验结果为“$P<0.05$”，并且，“样本一致率”大于“专业上要求的一致率”时，才可以认为：“试验法可以取代金标准法”[3]。

3.2 小结

本文展示了配对设计四格表资料的3种情形，其中，特设“金标准”的配对设计四格表资料不仅是值得进行统计分析的，也是最有实用价值的；基于SAS与R软件实现了配对设计四格表资料McNemar's χ^2检验；针对此种“差异性检验”的结果，如何进行正确地解读，如何陈述专业结论，都做了深入地阐释。

参考文献

[1] 周晓华，南希·A·奥布乔夫斯基，唐娜·k·麦克利什. 诊断医学统计学[M]. 宇传华，译. 北京：人民卫生出版社，2005：42-78，117-120，218-229.

[2] 胡良平，王琪. 定性资料统计分析及应用[M]. 北京：电子工业出版社，2016：2-18.

[3] 胡纯严，胡良平. 如何正确运用Z检验——定性资料一致性Z检验及SAS实现[J]. 四川精神卫生，2020，33(6)：522-526.

[4] 孙振球. 医学统计学[M]. 北京：人民卫生出版社，2002：106-107.

[5] 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，2004：359-365.

[6] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary，NC：SAS Institute Inc，2018：2997-3216.

[7] 詹姆斯·J·希金斯. 现代非参数统计概论[M]. 北京：中国统计出版社，2005：186-188.

[8] 胡良平. 现代医学统计学[M]. 北京：科学出版社，2020：244-257.

[9] 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：2021-01-15）

（本文编辑：戴浩然）