如何正确运用χ^2检验——人-时间资料独立性检验与SAS实现

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“未分层人-时间资料”和“分层人-时间资料”的独立性检验与SAS实现方法。在“人-时间资料”中，处理因素各水平组中的样本含量都以“人-年数”来表达。进而，需采用“发病密度”取代通常定性资料分析中的“发病率”。本文详细介绍“未分层人-时间资料”和“分层人-时间资料”的“发病密度”比较的具体方法，并通过两个实例，展示使用SAS软件实现计算的全过程，包括SAS程序代码、SAS输出结果、结果解释和结论陈述。

【关键词】 分层因素；人-时间；发病率；发病密度；独立性检验
中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20210719003

How to use χ^2 test correctly——the independence test for the data of the person-time

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 independence test and the SAS implementation for the "unstratified person-time data" and the "stratified person-time data". In "person-time data", the sample size in each level of the treatment factor was expressed as "person-years". Furthermore, it was necessary to use the "incidence density" to replace the "incidence rate" in the usual qualitative data analysis. The paper introduced the concrete approaches of comparing the "incidence density" of the "unstratified person-time data" and the "stratified person-time data" in detail, and demonstrated the whole process of using SAS software to realize the calculation through two examples, including the SAS program code, the SAS output results, the results explanation and the conclusion statement.

【Keywords】 Stratification factors; Person-time; Incidence rate; Incidence density; Independence test

人们在收集定性资料时，通常会面临如下实际问题：在所考察的处理因素分别处于“暴露”与“非暴露”水平下，观察两组受试对象是否出现某种结局时，发现各组中每位受试对象受到处理因素特定水平（“暴露”或“非暴露”）影响的时间长度可能不相同。这就意味着以各组受试对象的总人数作为计算该组样本发病率的分母是不够合理的，需要同时考虑每个人所经历的“时间长度”，它就是“人-年数”。本文介绍基于“人-年数”为“分母”的定性资料分析方法，即“人-时间资料的独立性检验”。

1 分析人-时间资料所需要的基本知识

1.1 分析人-时间资料涉及到的基本概念

1.1.1 累加发病率（简称“发病率”）

设观察由n个受试对象组成的一个群体在一个确定的时间段（例如一个月或一个季度或一年）内出现患某病的人数为m，则称该病的累加发病率（简称为“发病率”）为$100(m/n)\%$。这里有一个隐含的假定：即所有n个受试对象都被观察了相同时间长度（例如一个月或一个季度或一年）。

1.1.2 人-年数

在现实生活和科研工作中，受试对象被观察的时间长度可能不尽相同，有些受试对象可能分别被观察了3个月、7个月或14个月，如此等等。为了便于分析，不妨将“时间长度”统一折算为“一年”，称为“人-年数”。于是，分别被观察了3、7、14个月的3位受试对象，总共被观察了(3+7+14)/12=2人-年数。

1.1.3 发病密度

一组人群的发病密度（incidence density，ID）定义

为：该组群中发生事件（发生疾病）的人数除以该组群在研究期间累加的人-年（时间）总数[1]。这里的分母是“人-年数”，其取值范围为 0 到∞；而累加发病率的取值范围为 0~1。

1.2 人-时间资料的三种常见形式

第 1 种形式：未分层成组设计的人-时间资料，见表 1[1]。第 2 种形式：分层且含一个定性因素的人-时间资料，见表 2[1]。第 3 种形式：分层且含一个计量因素的人-时间资料，见表 3[2]。

表 1 某地 45~49 岁妇女乳腺癌发病例数与是否使用口服避孕药（OC）的关系

OC 使用与否	乳腺癌例数	人-年数
现在使用	9	2935
从不使用	239	135130

注：OC 代表“口服避孕药”

表 2 绝经后期妇女是否使用 OC 患乳腺癌情况的调查结果

年龄分组	病例数	人-年数	病例数	人-年数	病例数	人-年数
	从不使用 OC		现在使用 OC		过去使用 OC	
39~44 岁	5	4722	12	10199	4	3835
45~49 岁	26	20812	22	14044	12	8921
50~54 岁	129	71746	51	24948	46	26256
55~59 岁	159	73413	72	21576	82	39785
60~64 岁	35	15773	23	4876	29	11965

注：对原表的表达形式作了调整；OC 代表“口服避孕药”

表 3 某地镍精炼工人肺癌死亡情况调查结果

首次雇佣年龄	病例数	人-年数	病例数	人-年数	病例数	人-年数	病例数	人-年数
镍暴露量（mg/kg）：	0		0.5~4.0		4.5~8.0		8.5~12.0	
15.0~19.9 岁	1	502	1	202	1	87	0	30
20.0~27.4 岁	12	957	9	675	5	283	3	118
27.5~34.9 岁	2	452	9	377	2	154	1	73

注：“首次雇佣年龄”为“分层因素”；“镍暴露量”为“计量因素”（即试验因素）

2 基于未分层人-时间资料比较两总体“发病密度”

2.1 未分层成组设计人-时间资料的表达模式

未分层成组设计人-时间资料的表达模式，见表 4。

表 4 未分层成组设计人-时间资料的表达模式

暴露与否	病例数	人-年数
暴露	α_1	t_1
未暴露	α_2	t_2
合计	$\alpha_1+\alpha_2$	t_1+t_2

2.2 检验假设及检验统计量

检验假设可表述如下：$H_0: ID_1=ID_2$；$H_1: ID_1\neq ID_2$；$\alpha=0.05$。

根据资料所满足的前提条件，有两个可供选择的检验统计量[1-2]，分别见式（1）、式（2）：

$$Z=\frac{a_1-E_1-0.5}{\sqrt{V_1}}\sim N(0,1)，\text{如果}a_1>E_1 \quad (1)$$

$$Z=\frac{a_1-E_1+0.5}{\sqrt{V_1}}\sim N(0,1)，\text{如果}a_1\leqslant E_1 \quad (2)$$

在上面两式中，α_1 为表 4 中“暴露水平组”的“病例数”，E_1 和 V_1 分别为 α_1 的“期望频数”和“方差”，其计算分别见式（3）、式（4）：

$$E_1=\frac{(a_1+a_2)t_1}{t_1+t_2} \quad (3)$$

$$V_1=\frac{(a_1+a_2)t_1t_2}{(t_1+t_2)^2} \quad (4)$$

事实上，依据“$Z^2=\chi_1^2$”的统计理论知识[3]，可将式（1）和式（2）合并成式（5）：

$$\chi^2=\frac{(|a_1-E_1|-0.5)^2}{V_1}\sim\chi_1^2 \quad (5)$$

前提条件：这个检验适用于“$V_1\geqslant 5$”。

2.3 两总体发病密度比较的 SAS 实现

【例 1】如表 1 资料，试分析“现在使用 OC”与“从不使用 OC”两组妇女乳腺癌发病密度差异是否有统计学意义。

【分析与解答】设所需要的 SAS 程序[2]如下：

```
data abc;
do i=1 to 1;
input a1 t1 a2 t2 @@;
output;
```

```
end;
cards;
9 2935 239 135130
;
run;
data a;
set abc;
e_a1=(a1+a2)*t1/(t1+t2);
var_a1=(a1+a2)*t1*t2/(t1+t2)**2;
if a1>e_a1 then do;
z=(a1-e_a1-0.5)/sqrt(var_a1);end;
else do;
z=(a1-e_a1+0.5)/sqrt(var_a1);end;
if z<0thendo;p=2*(probnorm(z));end;
else do; p=2*(1-probnorm(z));end;
run;
proc print data=a;
var z p var_a1;
run;
proc sql;
create table b as select
sum(e_a1) as sum_e_a1,
sum(var_a1) as sum_var_a1,
sum(a1) as sum_a1 from a;
run;
quit;
data c;
set b;
chisq=(abs(sum_a1-sum_e_a1)-0.5)**2/
sum_var_a1;
p=1-probchi(chisq,1);
run;
proc print data=c;
var chisq p;
run;
```

【程序说明】第 2 句“do i=1 to 1;”代表该资料只有“一层”(相当于只有一个 4 格表资料);若整个资料有 8 层,此句应修改为“do i=1 to 8;”。

【SAS输出结果及解释】

Z	P	var_a1
1.42105	0.15530	5.15994

以上输出结果是基于标准正态分布理论算得的,V_1=5.15994 为“α_1=9”的方差;而 Z=1.42105、P=0.15530>0.05。

【统计结论和专业结论】上述计算结果说明,某地 45~49 岁妇女使用口服避孕药与不使用口服避孕药的乳腺癌发病密度差异无统计学意义,即可以认为:口服避孕药对该地 45~49 岁妇女是否患乳腺癌没有明显影响。

chisq	p
2.01939	0.15530

以上输出结果是基于 χ^2 分布理论算得的,χ^2=2.01939、P=0.15530>0.05,结论同上,此处从略。

【说明】当自由度为 1 时,$\chi^2=Z^2$,故当只有一个四格表资料时,前面两部分输出结果只需要保留其中任何一个即可。

3 基于分层人-时间资料比较两合并“发病密度”

3.1 资料的表达模式

为节省篇幅,资料的表达模式参见前文表 2 和表 4(假定其代表第“i”层)。值得注意的是:在表 2 中,“年龄分组”可被视为一个“分层因素”(或称为被控制的因素);而“使用 OC 的情况”可被视为该研究的一个试验因素,它有 3 个水平,分别为“从不使用 OC”“现在使用 OC”和“过去使用 OC”。

本文所介绍的方法适用于试验因素具有两个水平,对表 2 资料而言,可以在分层的条件下比较“从不使用 OC”与“现在使用 OC”两个水平下“各层合并后的发病密度”差异是否有统计学意义;也可以比较“从不使用 OC”与“过去使用 OC”两个水平下“各层合并后的发病密度”差异是否有统计学意义。

3.2 检验假设及检验统计量

检验假设可表述如下:H_0:合并 ID_1=合并 ID_2;H_1:合并 ID_1≠合并 ID_2;α=0.05。

检验统计量[1-2]见式(5):

$$\chi^2=\frac{\left[|A-E(A)|-0.5\right]^2}{Var(A)}\sim\chi_1^2 \tag{5}$$

式(5)中,各符号的含义如下:$A=\sum_{i=1}^{k}a_{1i}$ 为所有层中暴露组发生病例的观察频数之和;$E(A)=\sum_{i=1}^{k}E(a_{1i})$ 为所有层中暴露组发生病例的理论频数之和;$E(a_{1i})=(a_{1i}+a_{2i})t_{1i}/(t_{1i}+t_{2i})$ 为与观察频数 α_{1i} 对应的理论频数;$Var(A)=\sum_{i=1}^{k}Var(a_{1i})$ 为所有层中暴露组发

生病例的观察频数的方差之和；$Var(a_{1i})=(a_{1i}+a_{2i})t_{1i}t_{2i}/(t_{1i}+t_{2i})^2$ 为观察频数 α_{1i} 的方差。在上面各式中，“k”为分层因素的水平数。

前提条件：①假定各层发病密度之比[$RR_i=(\alpha_{1i}/t_{1i})/(\alpha_{2i}/t_{2i})$，$i=1,2,\cdots,k$]相等；②$Var(A)\geq5$。

3.3 两总体合并发病密度比较的SAS实现

【例2】如表2资料，试分析按年龄分组且在“从不使用OC”与“现在使用OC”两个条件下，合并的妇女乳腺癌发病密度差异是否有统计学意义。

【分析与解答】设所需要的SAS程序[2]如下：

```
data abc;
do i=1 to 5;
input a1 t1 a2 t2 @@;
output;
end;
cards;
5 4722 12 10199
26 20812 22 14044
129 71746 51 24948
159 73413 72 21576
35 15773 23 4876
;
run;
```

后面紧接其他SAS程序语句，具体内容与“第2.3节”中自“data a;”到最后完全相同，为节省篇幅，此处从略。

【SAS输出结果及解释】

观测	z	p	var_a1
1	0.06261	0.95008	3.6774
2	−0.63566	0.52500	11.5476
3	−0.69133	0.48936	34.4593
4	−2.98840	0.00280	40.5517
5	−2.72193	0.00649	10.4618

以上输出的是表2中5个“年龄分组(层)”各自的计算结果，其中，Z、P 和Var_a1分别代表“检验统计量”“P值”和“各层 α_1 的方差”。由“P值”列可知，只有在最后两个年龄组中，“从不使用OC”与“现在使用OC”两个条件下，妇女乳腺癌发病密度差异有统计学意义。

chisq	p
12.8219	0.000342577

以上输出的是表2中5个“年龄分组(层)”合并后的计算结果，即 $\chi^2=12.8219$，$P=0.000343<0.01$。

【统计结论和专业结论】上述计算结果说明，在“从不使用OC”与“现在使用OC”两个条件下，合并后的妇女乳腺癌发病密度差异有统计学意义。从表2中的实际数据可知，“现在使用OC”者的乳腺癌发病密度比“从不使用OC”者的乳腺癌发病密度大。

4 讨论与小结

4.1 讨论

采用“人-年数”取代“总样本含量”是人们在处理定性资料时，严格遵照“实事求是”原则的一个具体体现，是统计学的一个微小进步。然而，在实际科研工作中，精准地获得各组受试对象的“人-年数”是十分困难的事，尤其是在观察时期较长、回顾性研究且各组样本含量较大的情境中。因此，应尽可能事先制订出相对完善的研究设计方案，并严格执行研究设计方案(包括“标准操作规程方案”和“实时精准质量控制方案”等)[4-5]，以确保所获得的科研数据是精准可靠的[6-7]。

4.2 小结

本文介绍了与“人-时间资料”有关的基本知识、基于未分层人-时间资料比较两总体“发病密度”和基于分层人-时间资料比较两合并“发病密度”等内容；通过两个实例，介绍了基于SAS软件实现前述两种场合下的统计计算方法，对SAS输出结果进行解释，并做出了统计结论和专业结论。

参考文献

[1] 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，2004：648-704.

[2] 胡良平. 面向问题的统计学——(2)多因素设计与线性模型分析[M]. 北京：人民卫生出版社，2012：376-388.

[3] 胡纯严，胡良平. 如何正确运用χ^2检验——χ^2分布及相关内容[J]. 四川精神卫生，2021，34(1)：39-43.

[4] 胡良平，黄国平. 医学科研设计方法与关键技术[M]. 成都：四川大学出版社，2017：1-213.

[5] 胡良平. 课题设计方法与数据分析——关键技术与标准模板[M]. 北京：军事医学科学出版社，2014：1-136.

[6] 胡良平，胡纯严，鲍晓蕾. 应用数理统计[M]. 北京：电子工业出版社，2015：1-36.

[7] 胡良平. 临床统计学——临床课题统计解读[M]. 郑州：河南科学技术出版社，2019：1-39.

(收稿日期：2021-07-09)

(本文编辑：戴浩然)