如何正确运用χ^2检验——人-时间资料率比分析与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“分层人-时间资料”率比分析方法与SAS实现方法，具体内容包括以下四个方面：①分层人-时间资料中第i层率比的点估计和置信区间估计；②不同层间率比的齐性检验；③合并率比的点估计与置信区间估计；④分层人-时间资料中发病密度的线性趋势检验。通过两个实例，展示基于SAS软件实现前述四个分析内容的统计计算过程，包括提供SAS程序代码、对SAS输出结果进行解释，并给出统计结论和专业结论。

【关键词】 分层因素；人-时间；率比；齐性检验；线性趋势检验；置信区间
中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20210719004

How to use χ^2 test correctly——the rate ratio analysis for the data of the person-time and the implementation of SAS software

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 methods of the "layered person-time data" rate ratio analysis and the SAS implementation. The specific contents included the following four aspects: ① the point estimation and the confidence interval estimation of the layer i rate ratio in the stratified person-time data; ② the homogeneity test of the rate ratios among the different layers; ③ the point estimation and the confidence interval estimation of the common rate ratio; ④ the linear trend test of the incidence density in the stratified person-time data. Through two examples, it showed the statistical calculation process based on the SAS software to realize the aforementioned four kinds of the analysis contents, including providing the SAS program code, the explanation of the SAS output results, and giving the statistical and professional conclusions.

【**Keywords**】 Stratification factor; Person-time; Rate ratio; Homogeneity test; Linear trend test; Confidence interval

在分析“人-时间资料”时，若基于“人-年数”，在所考察的处理因素分别处于“暴露”与“非暴露”水平下，求出两组受试对象各自的“发病密度”之后，再求出两“发病密度”之比，就获得了“率比”[1]，它类似于分析通常四格表资料所采用的效应指标“相对危险度”或“优势比”[2-5]；但在文献[6]中，直接将“率比”称为“相对危险度”。本文将介绍与“率比分析”有关的内容以及基于SAS软件实现计算的方法。

1 分层人-时间资料中第i层率比的点估计和置信区间估计

1.1 分层人-时间资料的列表格式

分层人-时间资料的列表格式见表1。

表1 分层人-时间资料的列表格式

分层因素(i)	暴露状态	事件数	人-时间
1	暴露	α_{11}	t_{11}
	非暴露	α_{21}	t_{21}
2	暴露	α_{12}	t_{12}
	非暴露	α_{22}	t_{22}
…	…	…	…
k	暴露	α_{1k}	t_{1k}
	非暴露	α_{2k}	t_{2k}

1.2 第i层率比的点估计及置信区间估计

第i层“率比”是该层中“暴露水平”下的“发病密度(α_{1i}/t_{1i})”与“非暴露水平”下的“发病密度(α_{2i}/t_{2i})”之比值，相当于通常定性资料中的“相对危险度”或“优势比”[2-5]。计算方法见式(1)：

$$RR_i=\frac{a_{1i}/t_{1i}}{a_{2i}/t_{2i}} \quad (1)$$

第 i 层率比 RR_i 的 100(1−α)%置信区间的计算见式(2)：

$$\left[e^{C_1}, e^{C_2}\right] \quad (2)$$

设 $Z=Z_{\left(1-\frac{\alpha}{2}\right)}$，在式(2)中，$C_1$ 与 C_2 的定义分别见式(3)和式(4)：

$$C_1=ln\left(RR_i\right)-Z\times\sqrt{\frac{1}{a_{1i}}+\frac{1}{a_{2i}}} \quad (3)$$

$$C_2=ln\left(RR_i\right)+Z\times\sqrt{\frac{1}{a_{1i}}+\frac{1}{a_{2i}}} \quad (4)$$

2 不同层间率比的齐性检验

为了将各层资料合并起来估计“合并率比”，资料需满足的前提条件是各层率比相等，这就需要进行不同层间率比的齐性检验[1,6]。设分层因素共有 K 个水平，齐性检验的检验假设如下：

H_0：$RR_1=RR_2=\cdots=RR_k$；H_1：至少两个层的 RR_i 不同；α=0.05。

检验统计量见式(5)：

$$\chi^2_{het}=\sum_{i=1}^{K}W_i\left[ln\left(RR_i\right)-lnRR\right]^2\sim\chi^2_{K-1} \quad (5)$$

式(5)中，RR_i 为第 i 层率比[见式(1)]；W_i 为第 i 层权重，RR 为合并率比，其计算分别见式(6)、式(7)：

$$W_i=1/Var\left[ln\left(RR_i\right)\right] \quad (6)$$

$$RR=e^{C} \quad (7)$$

在式(6)中，$Var[ln(RR_i)]$ 为 $ln(RR_i)$ 的方差，计算方法见式(8)：

$$Var\left[ln\left(RR_i\right)\right]\approx\frac{1}{a_{1i}}+\frac{1}{a_{2i}} \quad (8)$$

在式(7)中，指数 C 的计算方法见式(9)：

$$C=\frac{\sum_{i=1}^{K}W_i ln\left(RR_i\right)}{\sum_{i=1}^{K}W_i} \quad (9)$$

3 合并率比的点估计与置信区间估计

当各层率比满足齐性要求时，可以对合并率比进行点估计和置信区间估计[1,6]。合并率比的点估计见式(7)和式(9)；合并率比的 100(1−α)%置信区间估计见式(10)：

$$\left[e^{C_1}, e^{C_2}\right] \quad (10)$$

设 $Z=Z_{\left(1-\frac{\alpha}{2}\right)}$，在式(10)中，$C_1$ 与 C_2 的定义分别见式(11)、式(12)：

$$C_1=ln(RR)-Z\times\sqrt{1/\sum_{i=1}^{K}W_i} \quad (11)$$

$$C_2=ln(RR)+Z\times\sqrt{1/\sum_{i=1}^{K}W_i} \quad (12)$$

4 分层人-时间资料率比分析与SAS实现

4.1 分层人-时间资料的实例

文献[1]提供了一个分层人-时间资料，见表2。

表2 绝经后期妇女是否使用口服避孕药(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”

4.2 分层人-时间资料各层率比齐性检验

【例1】试分析表2资料中“过去使用OC”与“从不使用OC”两组妇女各年龄组中，乳腺癌发病的率比是否满足齐性要求。

【分析与解答】设所需要的SAS程序[6-7]如下：

```
data abc;
do i=1 to 5;
input a1 t1 a2 t2 @@;
RR_a=(a1/t1)/(a2/t2);
w=1/(1/a1+1/a2);
wlnRR_a=w*log(RR_a);
output;
end;
cards;
5 4722 4 3835
26 20812 12 8921
129 71746 46 26256
159 73413 82 39785
```

```
35 15773 29 11965
;
run;
proc sql;
create table a as select sum(w) as sum_w,
sum(wlnRR_a) as sum_wlnRR_a
from abc;
run;
quit;
data b; set a;
c=sum_wlnRR_a/sum_w;RR=exp(c);
run;
proc sql;
create table c as select *
from abc, b;
run;
quit;
data d;
set c;
y=w*(log(RR_a)-log(RR))**2;
run;
proc sql;
create table e as select sum(y)
as chisq from d;
run;
data f;
set e;
p=1-probchi(chisq,4);
proc print;
var chisq p;
run;
```

【SAS输出结果及解释】

chisq	p
0.30263	0.98964

以上输出结果为：χ^2_{het}=0.30263、P=0.98964，说明各层率比满足齐性要求。

4.3 分层人-时间资料各层和合并率比的点估计与置信区间估计

【例2】如表2资料，试分析“过去使用OC”与“从不使用OC”两组妇女各年龄组中乳腺癌发病的率比点估计及置信区间估计；进而求“过去使用OC”与“从不使用OC”两组妇女乳腺癌发病合并率比的点估计及置信区间估计。

【分析与解答】设所需要的SAS程序[6-7]如下：

```
%let alpha=0.05;
data abc;
do i=1 to 5;
input a1 t1 a2 t2 @@;
output;
end;
cards;
5 4722 4 3835
26 20812 12 8921
129 71746 46 26256
159 73413 82 39785
35 15773 29 11965
;
run;
data a;
set abc;
RR_a=(a1/t1)/(a2/t2);
c_low=exp(log(RR_a)-probit(1-&alpha/2)*
sqrt(1/a1+1/a2));
c_up=exp(log(RR_a)+probit(1-&alpha/2)*
sqrt(1/a1+1/a2));
w=1/(1/a1+1/a2);
wlnRR_a=w*log(RR_a);
run;
proc print data=a;
var RR_a c_low c_up;
run;
proc sql;
create table b as select sum(w) as sum_w,
sum(wlnRR_a) as sum_wlnRR_a from a;
run;
quit;
data c;
set b;
c=sum_wlnRR_a/sum_w;
RR=exp(c);
c_low=exp(log(RR)-probit(1-&alpha/2)*
sqrt(1/sum_w));
c_up=exp(log(RR)+probit(1-&alpha/2)*
sqrt(1/sum_w));
run;
proc print data=c;
var RR c_low c_up;
run;
```

【SAS输出结果及解释】

观测	RR_a	c_low	c_up
1	1.01519	0.27261	3.78053
2	0.92874	0.46863	1.84058
3	1.02627	0.73297	1.43694
4	1.05082	0.80501	1.37169
5	0.91552	0.55966	1.49765

以上输出结果是5个年龄组各自率比(RR_a)及其95%置信区间的下限(c_low)与上限(c_up)的估计值,5个置信区间都包含1,说明各年龄组“过去使用OC”与“从不使用OC”的妇女乳腺癌发病率比与1之间差异无统计学意义,即各层中“过去使用OC”与“从不使用OC”的妇女乳腺癌发病密度相等。

RR	c_low	c_up
1.01399	0.84414	1.21801

以上输出结果是合并率比(*RR*)及其95%置信区间下限(c_low)和上限(c_up)的估计值。

【统计结论和专业结论】由以上输出可知,置信区间包含1,说明“过去使用OC”与“从不使用OC”两组妇女乳腺癌发病率比与1之间差异无统计学意义,即整体而言,“过去使用OC”与“从不使用OC”的妇女乳腺癌发病密度相等。

5 分层人-时间资料中发病密度的线性趋势检验

5.1 分层人-时间资料中发病密度线性趋势检验的资料列表格式

假设有一个暴露变量(或危险因素)E,而E有k个水平,第j个暴露水平组用得分S_j表示,这个S_j可以是该组内平均暴露水平,如果没有明显的计分方法,也可以用整数1,2,…,k代表k个计分。见表3。

表3 有k个暴露水平的分层的人-时间资料的列表格式

分层因素(i)		病例数	人-年数	病例数	人-年数	病例数	人-年数	病例数	人-年数
	暴露水平(j):	1		2		…		k	
1		α_{11}	t_{11}	α_{12}	t_{12}	…	…	α_{1k}	t_{1k}
2		α_{21}	t_{21}	α_{22}	t_{22}	…	…	α_{2k}	t_{2k}
…		…	…	…	…	…	…	…	…
s		α_{s1}	t_{s1}	α_{s2}	t_{s2}	…	…	α_{sk}	t_{sk}

5.2 分层人-时间资料中发病密度线性趋势检验的计算公式

基于表3中的符号,设p_{ij}代表第i层中第j暴露水平上的真实发病密度;$\hat{p}_{ij}$代表第i层中第j暴露水平上的观察发病密度[1,6]。

假设$ln(p_{ij})=\alpha_i+\beta S_j$,使用双侧检验及显著性水平为$\alpha$。

第一步:建立检验假设,确定检验水准。

$H_0:\beta=0;H_1:\beta\neq0;\alpha=0.05$。

第二步:计算检验统计量。

$$Z=\frac{\hat{\beta}}{se\left(\hat{\beta}\right)}\sim N(0,1) \tag{13}$$

上式中分子与分母的计算分别见式(14)、式(15):

$$\hat{\beta}=\frac{L_{xy}}{L_{xx}} \tag{14}$$

$$se\left(\hat{\beta}\right)=\frac{1}{\sqrt{L_{xx}}} \tag{15}$$

上面两式中,L_{xy}和L_{xx}的计算分别见式(16)、式(17):

$$L_{xy}=A-B\times C \tag{16}$$

$$L_{xx}=D-E \tag{17}$$

在式(16)中,A、B、C的计算公式分别见式(18)、式(19)、式(20):

$$A=\sum_{i=1}^{s}\sum_{j=1}^{k}w_{ij}S_j ln\left(\hat{p}_{ij}\right) \tag{18}$$

$$B=\sum_{i=1}^{s}\sum_{j=1}^{k}w_{ij}S_j \tag{19}$$

$$C=\frac{\sum_{i=1}^{s}\sum_{j=1}^{k}w_{ij}ln\left(\hat{p}_{ij}\right)}{\sum_{i=1}^{s}\sum_{j=1}^{k}w_{ij}} \tag{20}$$

在式(17)中,D、E的计算分别见式(21)、式(22):

$$D=\sum_{i=1}^{s}\sum_{j=1}^{k}w_{ij}S_j^2 \tag{21}$$

$$E=\frac{\left(\sum_{i=1}^{s}\sum_{j=1}^{k}w_{ij}S_j\right)^2}{\sum_{i=1}^{s}\sum_{j=1}^{k}w_{ij}} \tag{22}$$

在上面各式中,$w_{ij}=\alpha_{ij}$=第i层第j暴露水平上的病例数。

第三步:根据检验统计量的值确定P值,并作出统计推断和专业结论。

5.3 分层人-时间资料中发病密度线性趋势性检验的SAS实现

【例3】某地镍精炼工人肺癌死亡资料如表4所示[6]，试问镍精炼工人的肺癌死亡密度是否随着镍暴露量的增加呈增高趋势？

表4 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)

【分析与解答】设所需要的SAS程序如下：

```
data abc;
do i=1 to 3;
do j=1 to 4;
input w t @@;
p=w/t;
wslnp=w*j*log(p);
ws=w*j;
wlnp=w*log(p);
ws2=w*(j**2);
output;
end;
end;
cards;
1 502 1 202 1 87 0 30
12 957 9 675 5 283 3 118
2 452 9 377 2 154 1 73
;
run;
proc sql;
create table b as select sum(wslnp)
as sum_wslnp,sum(ws) as sum_ws,
sum(wlnp) as sum_wlnp,sum(w) as sum_w,
sum(ws2) as sum_ws2 from abc;
run;
quit;
data c;
set b;
l_xy=sum_wslnp-sum_ws*sum_wlnp/sum_w;
l_xx=sum_ws2-(sum_ws**2)/sum_w;
beta=l_xy/l_xx;
se_beta=1/sqrt(l_xx);
z=beta/se_beta;
if z<0 then do;
p=2*probnorm(z);
end;
else do;
p=2*(1-probnorm(z));
end;
run;
proc print;
var z p beta;
run;
```

【SAS输出结果及解释】

z	p	beta
1.58907	0.11204	0.25453

【统计结论和专业结论】由以上SAS输出结果可知，Z=1.58907，P=0.11204>0.05，说明$\hat{\beta}$=0.25453与0之间差异无统计学意义；可以认为：肺癌死亡密度不随镍暴露量的增加呈线性增高趋势。

6 讨论与小结

6.1 讨论

对人-时间资料进行分析，与对通常的定性资料[2-5]进行分析是大同小异的。从效应指标上来看，用"发病密度"取代了"发病率"，用"率比"取代了"相对危险度(适用于队列研究设计四格表资料)"和/或"优势比(适用于病例对照研究设计四格表资料)"。稍有不同的是：在分析通常的定性资料时，若各层间满足或不满足齐性要求，分别由基于"固定效应模型"或"随机效应模型"导出的公式去估计"合并或共同"相对危险度或优势比及其置信区间[2-5,8-13]；而分析人-时间资料时，若各层间满足齐性要求，有方法估计"合并或共同"率比及其置信区间[1,6]；而当各层间不满足齐性要求时，目前尚没有方法估计"合并或共同"率比及其置信区间。

6.2 小结

本文介绍了“分层人-时间资料中第i层率比的点估计和置信区间估计”“不同层间率比的齐性检验”“合并率比的点估计与置信区间估计”和“分层人-时间资料线性趋势性检验”等方法；通过两个实例，介绍了基于SAS软件实现前述各种场合下的统计计算的过程，对SAS输出结果进行解释，并做出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] 方积乾. 卫生统计学[M]. 7版. 北京：人民卫生出版社，2012：434-455.

[3] 方积乾，陆盈. 现代医学统计学[M]. 北京：人民卫生出版社，2002：150-209.

[4] 万崇华，罗家洪. 高级医学统计学[M]. 北京：科学出版社，2014：391-411.

[5] 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，2016：1-68.

[6] 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.

[7] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary，NC：SAS Institute Inc，2018：1109-1204，2997-3216，6007-6303，7991-8092.

[8] 赵仲堂. 流行病学研究方法与应用[M]. 2版. 北京：科学出版社，2005：545-565.

[9] 曾先涛. 应用STATA做Meta分析[M]. 北京：军事医学科学出版社，2014：22-57.

[10] 罗杰，冷卫东. 系统评价/Meta分析理论与实践[M]. 北京：军事医学科学出版社，2013：248-286.

[11] 胡纯严，胡良平. 如何正确运用χ^2检验——高维表资料齐性检验与SAS实现[J]. 四川精神卫生，2021，34(3)：202-207.

[12] 胡纯严，胡良平. 如何正确运用χ^2检验——高维表资料优势比分析与SAS实现[J]. 四川精神卫生，2021，34(3)：208-213.

[13] 胡纯严，胡良平. 如何正确运用χ^2检验——高维表资料相对危险度分析与SAS实现[J]. 四川精神卫生，2021，34(3)：214-219.

（收稿日期：2021-07-19）

（本文编辑：戴浩然）