中华流行病学杂志1994年6月 第15卷第3期

（系列讲座）

现代实用流行病学

第二讲　病例对照研究（续）

章扬熙

[例9]　对例8资料进一步作分层、分级分析，不同性别吸烟量与肺癌关系如表10，试进行分析。

表10　不同性别吸烟量与肺癌的关系

组别		每日吸烟量(支) 0		1～		5～		15～		合计		
		0	Y_0	2.5	Y_1	10	Y_2	20	Y_3			
男	病例	2	$a_{10}=c_1$	33	a_{11}	250	a_{12}	364	a_{13}	649	M_{11}	$\Sigma a_{1j}Y_j=9862.5$
	对照	27	$b_{10}=d_1$	55	b_{11}	293	b_{12}	274	b_{13}	649	M_{01}	$\Sigma N_{1j}Y_j=18410$
i=1	合计	29	N_{10}	85	N_{11}	543	N_{12}	638	N_{13}	1298	T_1	$\Sigma N_{1j}Y_j^2=310050$
	OR	1.00		8.10		11.52		17.93				
女	病例	19	$a_{20}=c_2$	7	a_{21}	19	a_{22}	15	a_{23}	60	M_{12}	$\Sigma a_{2j}Y_j=507.5$
	对照	32	$b_{20}=d_2$	12	b_{21}	10	b_{22}	6	b_{23}	60	M_{02}	$\Sigma N_{2j}Y_j=757.5$
i=2	合计	51	N_{20}	19	N_{21}	29	N_{22}	21	N_{23}	120	T_2	$\Sigma N_{2j}Y_j^2=11418.75$
	OR	1.00		0.98		3.20		4.21				
SRR		1.00		1.66		3.99		5.52				
SRR95%		1.00		1.16～		2.80～		3.87～				
可信区间				2.90		6.97		9.65				

检验剂量反应关系的总趋势时，可用公式40进行扩展的X_{M-EXT}检验，得

$$\chi_{M-EXT}=\frac{\left(9862.5-\frac{649}{1298}\times 18410\right)+\left(507.5-\frac{60}{120}\times 757.5\right)}{\left[\frac{649\times 649}{1298^2(1298-1)}\left(1298\times 310050-18410^2\right)+\frac{60\times 60}{120^2(120-1)}\left(120\times 11418.75-757.5^2\right)\right]^{\frac{1}{2}}}=6.67$$

其界值即$u_{0.05}=1.96$，$u_{0.01}=2.58$，$P<0.01$，说明吸烟量与肺癌有剂量反应关系。于是，可进一步用公式41求线性趋势的斜率，即回归系数，得

$$by=\frac{\left(9862.5-\frac{649}{1298}\times 18410\right)+\left(507.5-\frac{60}{120}\times 757.5\right)}{\left(310050-\frac{18410^2}{1298}\right)+\left(11418.75-\frac{757.5^2}{120}\right)}=0.014$$

把不吸烟组与不同吸烟量组比较，求OR，比如男性每日吸5～支组，$OR=\frac{ad}{bc}=\frac{250\times 27}{293\times 2}=11.52$，余类推。若各层OR一致，可用公式39计算总OR，否则计

本文作者单位：辽宁省卫生防疫站　110005　沈阳市

算标准化率比，本例男女OR已证明相差大，应用公式42求SRR，比如每日吸烟5～支组

$$SRR=\frac{250\times27/293+19\times32/10}{2+19}=3.99$$

再用公式43、44求其95%可信区间，得

$$\left[3.99^{-1}\pm1.96\times\frac{\sqrt{2+19}}{\frac{250\times27}{293}+\frac{19\times32}{10}}\right]^{-1}=2.80\sim6.97$$

余类推，结果列于表10的末行。从各SRR值显见，吸烟量越大患肺癌的危险度也越大。

在分析性流行病学的研究中，常遇到多因素问题，对于数量较少的多因素问题，可作析因分析。析因分析是按几个研究因素与各因素水平的一切组合状态情况分组，以各因素的基准水平为对比点，计算各状态的OR值，考察各因素的单一作用与交互影响（加法模型$OR_{11}=OR_{10}+OR_{01}-1$与乘法模型$OR_{11}=OR_{10}\cdot OR_{01}$时不存在交互作用），这种方法直观、简易，便于应用。

[例10] 对某造船厂工人进行接触石棉、吸烟两因素与肺癌联系的病例对照研究，结果如表11，试进行分析。

表中OR、S（lnOR）及OR的95%可信区间是分

表11 接触石棉及吸烟与肺癌的关系

接触石棉史	吸烟史	病例	对照	OR	lnOR	S(lnOR)	OR95%可信区间
−	−	50c	203d	1.00	−	−	−
−	+	217	220	4.01	1.3888	0.1187	3.18～5.06
−	++	96	50	7.80	2.0541	0.1880	5.40～11.27
+	−	11	35	1.28	0.2469	0.3527	0.64～2.56
+	+	70	42	6.78	1.9140	0.2074	4.52～10.18
+	++	14	3	18.95	2.9258	0.6401	5.32～65.39

别应用公式29、31、32，求得。从各OR值看，吸烟与接触石棉均为肺癌的危险因素，而且吸烟还有剂量反应关系。接触石棉与吸烟有加法模型的协同作用[6.78＞(4.01+1.28−1)，18.95＞(7.80+1.28−1)]。但是从总体OR的95%可信限看，仅接触石棉者OR为0.64～2.56，说明尚无显著性，同时接触石棉及吸烟者的OR95%可信区间的下限均小于两者单因素OR95%可信区间的上限和减一，也说明不显著，这很可能是由于分层后，有的层例数不够的原因，可改用Logistic模型分析或增加例数作进一步的研究。

配比的病例对照研究统计推断

配比病例对照研究资料应按配比方法进行统计分析，1∶1配比与1∶M配比（M＞1）资料分析公式也不相同。

一、1∶1配比的病例对照研究：

[例11] Sartwell等人研究口服某避孕药史与妇女患血栓栓塞的联系，从美国五个城市选择15～44岁妇女采用1∶1配比方法进行了病例对照研究，结果如表12，试进行分析。

表12 口服避孕药与血栓栓塞的关系

		对照		共配比数
		有服避孕药史	无服避孕药史	
病例	有服避孕药史	10(a)	57(b)	67(a+c)
	无服避孕药史	13(b)	95(d)	108(b+d)
共配比数		23(a+b)	152(c+d)	175(T)

在对比病例与对照口服避孕药史情况的差异，只有b、c有意义，病例服避孕药但对照未服避孕药的配比有57，对病例未服避孕药但对照服避孕药的配比只有13对，χ^2检验用下式

$$\chi^2=\frac{(b-c)^2}{b+c} \qquad (48)$$

当$b+c<40$时，用连续性校正公式为

$$\chi^2=\frac{(|b-c|-1)^2}{b+c} \qquad (49)$$

本例$\chi^2=\frac{(13-57)^2}{13+57}=27.66$

自由度$df=1$，$P<0.01$，说明病例与对照服避孕药率有差别，可进而用公式50计算比数比

$$OR=\frac{c}{b} \qquad (50)$$

本例$OR=57/13=4.38$

总体OR的95%可信区间用公式11计算，该公式中$\chi=\sqrt{\chi^2}$，χ^2值用不作连续性校正的χ^2值。本例得

$4.38^{1\pm 1.96/\sqrt{27.66}}=2.53\sim7.60$

以上结果说明，有口服避孕药史与患血栓栓塞有联系，其总体OR在2.53～7.60内的概率为95%。

应用公式51计算AR%

$$AR\%=\frac{c-b}{c} \qquad (51)$$

本例 $AR\%=\frac{57-13}{57}=0.7719$

应用公式52、53计算AR%的方差及$100(1-\alpha)\%$可信区间

$$V(AR\%)=\frac{b(b+c)}{c^3} \qquad (52)$$

$$AR\%\pm u_a\sqrt{V(AR\%)} \qquad (53)$$

本例 $V(AR\%)=\frac{13(13+57)}{573}=0.004914$

AR%的95%可信区间为：

$0.7719\pm1.96\sqrt{0.004914}=0.6345\sim0.9093$

这说明在有口服避孕药史的人群中，由于服用口服避孕药导致患血栓栓塞的占血栓栓塞患者总数的77.19%，总体AR%的95%可信区间为63.45%～90.93%

应用公式54计算PAR%

$$PAR\%=\frac{(a+c)(c-b)}{CT} \qquad (54)$$

本例 $PAR\%=\frac{(10+57)(57-13)}{57\times175}=0.2955$

（即29.55%）

应用公式55、56计算PAR%的$100(1-\alpha)\%$可信区间

$$V(PAR\%)=\frac{1}{C^2T^2}\left\{a(c-b)^2+\frac{(c^2+ab)^2}{c}+b(a+c)^2-\frac{(a+c)^2(c-b)^2}{T}\right\} \qquad (55)$$

$$PAR\%\pm u_\alpha\sqrt{V(PAR\%)} \qquad (56)$$

本例 $V(PAR\%)=\frac{1}{57^2\times175^2}\left\{10\times(57-13)^2+\frac{(57^2+10\times13)^2}{57}+13\times(10+57)^2-\frac{(10+57)^2\times(57-13)^2}{175}\right\}=0.00229441$

PAR%的95%可信区间为

$0.2955\pm1.96\sqrt{0.00229441}=0.2016\sim0.3894$

（20.16%～38.94%）

这说明在美国该地患血栓栓塞妇女病例中29.55%可归因于服用该避孕药，总体PAR%的95%可信区间为20.16%～38.94%

二、1：M配比的病例对照研究：以1：2配比的实例来说明。

［例12］ 吸烟与冠心病联系的1：2配比病例对照研究结果如表13，试进行分析

进行χ^2检验，检验假设为两组总体吸烟率相等，

表13 吸烟与冠心病联系的1：2配比资料

病例吸烟史	对照中吸烟史数 0	1	2	合计
1	22($n_{1,0}$)	50($n_{1,1}$)	22($n_{1,2}$)	94
0	21($n_{0,0}$)	34($n_{0,1}$)	13($n_{0,2}$)	68
合计	43	84	35	162

设检验水准为0.05，用公式57求χ^2值，自由度为1（这是1：M配比资料求χ^2值的通用公式）

$$\chi^2=\frac{\left[\sum_{m=1}^{M}(M-m+1)n_{1,m-1}-\sum_{m=1}^{M}m\,n_{0,m}\right]^2}{\sum_{m=1}^{M}(n_{1,m-1}+n_{0,m})\,m(M-m+1)} \quad (57)$$

本例M=2, m=1, 2

$\chi^2=[(2-1+1)\times22+(2-2+1)\times50-(1\times34+2\times13)]^2\div[(22+34)\times1\times(2-1+1)+(50+13)\times2\times(2-2+1)]=4.86$

自由度df=1, $\chi^2_{0.05}=3.84$, $P<0.05$, 拒绝检验假设，说明吸烟与冠心病有联系。

当（总例数$-n_{1,M}-n_{0,0}$）<40时[1]，应用校正χ^2公式，为公式54分子未平方前取绝对值后减去1/2（M+1）后再平方，分母不变。由于χ^2检验说明吸烟与冠心病有联系，可进而用公式58计算比数比(OR)

$$OR_{MH}=\frac{\sum_{m=1}^{M}(M-m+1)n_{1,m-1}}{\sum_{m=1}^{M}m\,n_{0,m}} \quad (58)$$

本例$OR_{MH}=\frac{(2-1+1)\times22+(2-2+1)\times50}{1\times34+2\times13}=1.57$

总体OR的95%可信区间仍用公式11计算，本例为

$1.57(1\pm1.96/\sqrt{4.86})=1.05\sim2.34$

病例对照研究的Logistic回归模型分析

在现代流行病学研究中，常遇到多种致病因素相互联系、相互制约的复杂病因链，对其采用选择对象、配比等方法，化多因素问题为单因素，则会产生孤立的研究单因素不能认识因素间的交互作用之弊。所以多因素问题宜采用多因素的研究方法，分层分析仅适用因素较少、样本例数较多的情况，而且当因素是连续性变量时，只能用等级分层方法，从而损失不少信息。有人曾用线性回归模型来研究多个暴露因素与疾病的联系，但估计的发病率可能出现小于0或大于1的不合理情况。二十世纪六十年代，在流行病学领域开始应用了Logistic回归模型，它弥补了M-H分层分析法和线性回归的不足，经过近年的发展，现已成为分析性流行病学研究中的重要方法。

一、Logistic回归模型概述：设影响疾病发病的研究因素有p个，即X_1, X_2, ……, Xp，发病概率为P，Logistic模型可以表达为

$$P=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots\cdots+\beta pXp)}} \quad (59)$$

又设Logit $P=\ln\frac{P}{1-P}$，即发病率与未发病数的比数的自然对数，则Logistic模型又可以表达成线性形式

$$\text{Logit } P=\beta_0+\beta_1X_1+\beta_2X_2+\cdots\cdots\beta pXp \quad (60)$$

式中βk（k=1, 2, ……, p）为偏回归系数。对这些参数的估计，通常采用最大似然法，以使βk不限于多元正态分布的假定，而推广至指数密度族的一些子集，自变量的类别和分布不受限制，具有广泛的应用价值。这个计算是应用专用程序在电子计算机上进行的。计算步骤为：

（一）构造一个似然函数L，为了简化计算，再取其自然对数值lnL；

（二）对这个对数似然函数lnL求其各参数β的一阶偏导数；

（三）通常是应用Newton-Raphson迭代法，解下列非线性方程组，求出满足下列方程组的各$\hat{\beta}k$，即为各参数βk的最大似然估计值。

$$\left[\frac{\partial\ln L}{\partial\beta_k}\right]_{\beta=\hat{\beta}}=0$$

（四）各参数估计值$\hat{\beta}k$的方差、协方差矩阵为迭代终止时得到的信息矩阵的逆矩阵，即目标函数二阶偏导数的负值组成的矩阵的逆矩阵。应用这些信息，可对$\hat{\beta}k$进行假设检验和参数的区间估计。

（五）应用Logistic回归模型进行分析，目的在于把因素效应有显著意义的变量纳入模型之中。通常采用阶梯式配合技术，即由最简单的模型开始，直到把所有显著意义的变量全部纳入模型之中，最后把有显著意义的变量间交互影响项也纳入模型之中。各模型对资料拟合的好坏，可进行拟合优度检验，包括直接检验方法和参数的假设检验两类。

1.直接检验方法：将每个个体的发病概率估计值从小到大排列，采用10分位分组方法分为10组（每组合计人数为T/10）进行C检验或按发病概率P（Y=

1|X）等距分为10组（各组段上限依次为0.1, 0.2, …, 1.0）进行H检验，

$$C=\sum_{k=0}^{1}\sum_{l=1}^{10}\frac{(O_{kl}-E_{kl})^2}{E_{kl}} \quad (61)$$

$$O_{1L}=\sum_{j\in D_1} Yj, \quad O_{0l}=\sum_{j\in D_1}(1-Yj) \quad (62, 63)$$

$$E_{1L}=\sum_{j\in D_1} P(Y=1|X) \quad (64)$$

$$E_{0L}=\sum_{j\in D_1}\Big(1-P(Y=1|Xj)\Big) \quad (65)$$

式中O_{11}为实际病例数，O_{01}为实际非病例数，E_{11}为期望病例数，F_{01}为期望非病例数、自由度为组数减二，10−2=8。H检验公式与C检验相同，C分布与H分布近似卡方分布，自由度为组数减二。

2.参数的假设检验：对模型中新加入参数的假设检验有三种。

①似然比检验：通过对比两个包含不同参数个数的lnL值，对模型中的参数进行假设检验。求似然比统计量G

$$G=2(\ln L_{t+f}-\ln L_t) \quad (66)$$

上式中t为原包含在模型中的参数个数，f为所加入的参数的个数，统计量G服从自由度为f的χ^2分布。通过χ^2检验考察f个参数贡献有无显著意义。有显著性者纳入模型，不显著者剔除出模型。

②Z检验：计算模型中各因素的偏回归系数$\hat{\beta}_k$及其方差$V(\hat{\beta}_k)$，进而用公式67求标准化偏回归系数的估计值$Z=\hat{\beta}_k/\sqrt{V(\hat{\beta}_k)}$，在大样本中该值呈标准正态分布（u布），$H_0:\beta_k=0$，对Z值进行检验，不显著则剔出模型，显著则纳入模型。

③积分统计量（score statistic）：又称记分检验或梯度检验。设对数似然函数的一阶偏导数矢量为$S=S(\beta_0,\beta)$，二阶偏导数矩阵的负值为信息矩阵$I=I(\beta_0,\beta)$，若检验假设$H_0:\beta_1=0$，用下式计算S

$$S=[S(\beta_0,O)]^T[-I(\beta_0,O)]^{-1}[S(\beta_0,O)] \quad (68)$$

S分布近似χ^2分布，自由度亦为新引入的参数个数.这种方法计算较繁，稳定性较差，国内很少应用。

（六）比数比（odds ratio）的计算：设有q个暴露因素E，P_1个混杂因素V，P_2个效应修正变量W（交互作用），Logistic模型可表示为

$$\text{logit } P=\alpha+\sum_{k=1}^{q}\beta_kE_k+\sum_{i=1}^{p1}r_iV_i+\sum_{k=1}^{q}\sum_{j=1}^{p2}\delta jE_kW_j$$

诸暴露因素暴露条件E^*与E^{**}的比数比为

$$OR(E^*, E^{**})=\exp\Big[\sum_{k=1}^{q}(E^*_k-E^{**}_k)\cdot(\beta k+\sum_{j=1}^{p2}\delta_{kj}W_j)\Big] \quad (69)$$

当暴露因素只有一个，即q=1，E^*与E^{**}的比数比为

$$OR(E^*, E^{**})=\exp\Big[(E^*-E^{**})\cdot(\beta+\sum_{j=1}^{p2}\delta jWi)\Big] \quad (70)$$

若不存在效应修正变量，即δ=0，仅考虑单一暴露变量E，两组暴露条件E^*和E^{**}之差等于1时，比数比为

$$OR(E^*,E^{**})=\exp[(\beta)(E^*-E^{**})]=\exp(\beta) \quad (71)$$

二、Logistic模型适用条件和在病例对照研究中的应用

（一）适用条件：第一，因变量必须是两项分类（0，1型）的数据，或必须限于0～1之间的数据。第二，每个自变量X与因变量P之间呈单调上升或下降的S形曲线，或者说自变量X与LogitP之间呈线性关系。自变量的效应呈相乘的统计关系。

（二）Logistic模型在病例对照研究中的应用：Logistic模型有两种，即非条件Logistic模型与条件Logistic模型，前者可用于非配比的病例对照研究，后者用于配比的病例对照研究。在应用非条件Logistic模型时，对于分层资料必须引入反映层次的变量。由于这两种模型的似然函数不同，所以计算也不一样，通常各有其计算机专用程序，在电子计算机上运算。程序的使用方法可参考有关的使用说明。