混淆因素分层分析方法

天津医学院流行病学教研室 武光林

混淆因素调整方法包括分层分析方法、配比方法和多因素分析方法。本文重点讨论混淆因素分层分析常用统计学方法。

两暴露水平资料的混淆因素调整方法

应用最广泛的是Mantel-Haenszel检验方法〔1～3〕，近年来已在国内逐渐应用。下面再介绍另一种应用较为广泛的Woolf方法。

这种方法也建立在分层方法的基础上。把研究的因素（暴露因素）按双分变量或者说按两个暴露水平分析，然后按混淆因素分为K层，形成K个亚组，表1为第i亚组2×2表的一般形式。

表 1 第i亚组病例和对照暴露频数

暴露	病例	对照	合计
有	a_i	b_i	m_{1i}
无	c_i	d_i	m_{2i}
合计	n_{1i}	n_{2i}	n_i

假定每个2×2表亚组的比数比为ψ_i，那么调整混淆因素后Woolf比数比为：

$$\hat{\psi}_w = \exp\left[\sum_{i=1}^{k} V_i^{-1} \ln \hat{\psi}_i \Big/ \sum_{i=1}^{k} V_i^{-1}\right] \quad \cdots\cdots\cdots 1$$

式中$V_i = (a_i + c_i)/a_i c_i + (b_i + d_i)/b_i d_i$

假定$\ln\hat{\psi}_w$为近似的正态分布，显著性检验的统计值χ^2_w的计算公式为：

$$\chi^2_w = W(\ln\hat{\psi}_w)^2 \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots 2$$

式中$w = \sum_{i=1}^{k} V_i^{-1}$，自由度df＝1。

$\hat{\psi}_w$的95%可信限为：

$$\underline{\psi}_w, \overline{\psi}_w = \hat{\psi}_w \exp(\pm 1.96/\sqrt{W}) \quad \cdots\cdots\cdots\cdots 3$$

表2是一个心肌梗塞（MI）和口服避孕药(OC)的病例对照研究资料。新近应用OC为暴露因素，按年龄分层分析MI-OC的关系。应用Woolf方法计算按年龄调整后比数比及其95%可信限$\hat{\psi}_w$＝4.2(2.51，6.99)，χ^2_w＝30.48（df＝1）；而按Mantel-Haenszel方法计算$\hat{\psi}_{MH}$＝3.97（2.28，6.93），χ^2_{MH}＝27.97（df＝1）。可见两种计算方法非常接近，均表明MI-OC有关。但是Woolf方法在计算时有一定的限制。由于公式中涉及到V_i，当2×2亚表出现0值就不能计算，所以Mantel-Haenszel方法应用更为广泛。

表 2 按年龄调整后心肌梗塞和应用口服避孕药的关系

应用OC	年龄组（岁）25～29 病例	25～29 对照	30～34 病例	30～34 对照	35～39 病例	35～39 对照	40～44 病例	40～44 对照	45～49 病例	45～49 对照
是	4	62	9	33	4	26	6	9	6	5
否	2	224	12	390	33	330	65	362	93	301

多暴露水平资料混淆因素调整方法

我们前面提到的Mantel-Haenszel检验方法和Woolf方法对混淆因素调整有一个显著的特点，即均建立在2×2表的基础上。对于混淆因素可以同时调整几个，而每个混淆因素亦可分为K个水平。但对暴露因素则仅仅是双分的，只能分为两个暴露水平。但是在某些变量与疾病关系分析中，是否存在剂量反应关系是我们需要考虑的重要问题。若按混淆因素分层适当，应当在每个层次内消除了混淆作用。如果粗比数比分析存在剂量反应关系，那么在对混淆因素调整后，暴露因素的这种关系是否继续存在显然是一个需要深入分析的问题。这用前面介绍的方法是不能解决的，而Meittinen提出的相对危险度标化方法〔4〕和Mantel提出的扩大的Mantel-Haenszel检验程序〔5〕恰恰解决了这样一个问题。

一、标化相对危险度方法：如表3，如果我们以参照组为标准进行标化，那么计算第i暴露水平SRR_i的公式为：

$$S\hat{R}R_i = (\sum_j a_{ij}d_j/C_{ij})/b \quad \cdots\cdots 4$$

$\sum_j$表示各层次总和，下标ij分别表示在第i水平和第j层。

表 3　第j层标化相对危险度公式中符号意义

层次	组别	危险因素范畴	
		第i水平	参照组
第j层	病例	a_{ij}	b_{ij}
	对照	c_{ij}	d_{ij}
合计	病例	a_i	b
	对照	c_i	d

表 4 的资料是分析产次和乳腺癌关系的病例对照研究。按粗相对危险度（CRR）分析发现随着产次的增加对乳腺癌相对危险度降低。为了反映产次和乳腺癌的真正关系，我们通过标化相对危险度方法控制初产年龄的混淆作用。结果显示：各产次范围的SRR均与参照组接近，说明产次与乳腺癌无关联。

表 4　按初产年龄调整后生产次数和乳腺癌的关系

初产年龄（岁）	组别	生产次数			合计
		1*	2～3	4～9	
<20	病例	2	6	10	18
	对照	24	47	50	121
	合计	26	53	60	139
20～29	病例	36	144	68	248
	对照	129	428	312	869
	合计	165	572	380	1117
≥30	病例	39	47	22	108
	对照	80	144	32	256
	合计	119	191	54	364
合计	病例	77	197	100	374
	对照	233	619	394	1246
	CRR	1.00	0.96	0.77	
	SRR_i	1.00	0.94	1.14	

* 参照组

这里我们应用了流行病学常用指标相对危险度。从统计学的角度讲，前面的公式只有应用比数比（odds ratio）的概念才普遍适用。所以标化相对危险度也可以称为标化比数比。

二、扩大的Mantel-Haenszel检验方法：标化相对危险度的明显好处是比较直观。如上例，经标化的相对危险度与参照组相比均接近 1，显示产次与乳腺癌的关系是由于初产年龄混淆的结果。如果暴露-疾病之间无关联，调整混淆因素后的SRR理论上应该等于 1。但是由于调查资料是样本，存在着机会问题，不可能完全出现这种情况，所以需要做显著性检验。这就是下面介绍的扩大的Mantel-Haenszel χ^2检验方法。

1.基本公式：假定我们对所研究的因素分层为K个有序水平，把其按某一变量（混淆变量）分为i层，即形成i个 2 ×K表，那么第i层就形成第i个 2 ×K表（表 5）。

表 5　第i个2×K连续表图表说明

	所研究因素水平j					合计
	0	1	2	…	K－1	
	Y_0	Y_1	Y_2	…	Y_{k-1}	
病例	A_{0i}	A_{1i}	A_{2i}	…	A_{k-1i}	N_{1i}
对照	B_{0i}	B_{1i}	B_{2i}	…	B_{k-1i}	N_{2i}
合计	M_{0i}	M_{1i}	M_{2i}	…	M_{k-1i}	T_i

所研究水平（暴露水平）即以j表示，分层数以表示，病例和对照分别为A_{ji}和B_{ji}，则下标 j 表示水平数，下标i表示层数，Yj 为暴露水平的标度，规定χ值病例为 1，对照为 0。

①单一的 2 ×K连续表χ^2的计算公式（df=1）：

$$\chi^2=\frac{\left[\sum_j A_{ji}Y_j - E\left(\sum_j A_{ji}Y_j\right)\right]^2}{V\left(\sum_j A_{ji}Y_j\right)} = \frac{\left[\sum_j A_{ji}Y_j - \frac{N_{1i}}{T_i}\left(\sum_j M_{ji}Y_j\right)\right]^2}{\frac{N_{1i}N_{2i}}{T_i^2(T_i-1)}\left[T_i\sum_j M_{ji}Y_j^2-\left(\sum_j M_{ji}Y_j\right)^2\right]} \quad \cdots\cdots 5$$

此为第i个 2 ×K表线性趋势的χ^2检验公式，$\sum_j A_{ji}Y_i$为不同暴露水平Yj与相应水平病例数相乘的合计数，$E(\sum_j A_{ji}Y_j)$和$V(\sum_j A_{ji}Y_i)$分别表示$\sum_j A_{ji}Y_j$的预期值和方差。目的在于检验暴露-疾病二者之间是否存在剂量反应关系。

②i个 2 ×K连续表总χ^2检验公式（df＝1），

$$\chi^2_{sum}=\frac{\left[\sum_i\sum_j A_{ji}Y_j-\sum_i E(\sum_j A_{ji}Y_j)\right]^2}{\sum_i V(\sum_j A_{ji}Y_j)} \quad \cdots\cdots\cdots\cdots 6$$

此公式则为按混淆因素分层后暴露-疾病之间剂量反应关系χ^2检验公式。

③暴露水平测量值Y_{ji}：可以把Y_j按照所研究变量的暴露范围中点确定，如把吸烟按照每天吸烟支数0、1～10、11～20、21～30、≥31分层，那么可以把Y_j分别定为0、5、15、25、35。或者让$Y_j=j$，即把Y_j按0、1、2、3、4处理。也可利用Y_j专用公式，可参考文献〔5〕这里不再介绍。

2.基本计算步骤：

单一的2×K表χ^2检验：

①$\sum_j A_{ji}Y_j$

②$\sum_j M_{ji}Y_j$

③$\sum_j M_{ji}Y_j^2$

④$\frac{N_{1i}}{T_i}$

⑤$\frac{N_{1i}N_{2i}}{T_i^2(T_i-1)}$

⑥$E(\sum_j A_{ji}Y_j)=$②④

⑦$V(\sum_j A_{ji}Y_j)=$⑤$(T_i$③$-$②$^2)$

⑧$\chi^2=\frac{(①-⑥)^2}{⑦}$

计算i个2×K表的总χ^2值：

$$\chi^2_{sum}=\frac{(\sum_i ①-\sum_i ⑥)^2}{\sum_i ⑦}$$

应用这些步骤计算起来相当方便。我们仍以表4的资料为例。首先确定$Y_j=j$，定为0、1、2三个水平，然后把初产年龄三个层次分别按i=1、2、3计算单一的2×K表χ^2值。例如在初产年龄＜20岁层次内（i=1），计算：

①$\sum_j A_{ji}Y_j=2\times0+6\times1+10\times2=26$

②$\sum_j M_{ji}Y_j=26\times0+53\times1+60\times2=173$

③$\sum_j M_{ji}Y_j=26\times0+53\times1+60\times2^2=293$

然后按照①～⑧中其它步骤计算$\chi^2=1.47$。各层次计算结果列于表6。

表6 表4中3个2×K表主要计算结果

i	$\sum_j A_{ji}Y_j$ ①	$E(\sum_j A_{ji}Y_j)$ ⑥=②④	$V(\sum_j A_{ji}Y_j)$ ⑦=⑤(T_i③−②^2)	χ^2_i ⑧=$\frac{(①-⑥)^2}{⑦}$
1	26	22.403	8.820	1.47
2	280	295.731	88.069	2.81
3	91	88.714	33.771	0.15
$\sum_i$	397	406.848	130.660	—

在此基础上计算3个2×K表的总χ^2值：

$$\chi^2_{sum}=\frac{(397-406.848)^2}{130.660}=0.74$$

所以在消除了初产年龄混淆作用之后χ^2值为0.74，P>0.05（df=1），产次与乳腺癌无关联，这与前面标化相对危险度分析结果完全一致。

前面介绍的两种方法均为多暴露水平资料的混淆因素调整方法。和Mantel-Haenszel检验方法比较，SRR相当于ψ_{MH}，但ψ_{MH}表示两暴露水平的调整比数比，而SRR表示和暴露水平0比较的各暴露水平的调整比数比。SRR的主要优点是，既可以与CRR比较，也可以看出SRR与暴露水平之间的关系变化。但是在一般情况下需做统计学检验。扩大的Mantel-Haenszel检验方法相当于χ^2_{MH}。所不同的是，χ^2_{MH}用于检验ψ_{MH}是否有显著的统计学关系，扩大的Mantel-Haenszel检验方法用于检验按混淆因素分层后暴露-疾病之间是否存在剂量反应关系，是一种线性趋势χ^2检验方法，作为统计学推断，这是一种主要方法。但是作为流行病学分析，和标化相对危险度方法结合使用可能更全面一些。

讨　论

疾病和某些因素之间的关系是非常复杂的。疾病的发病原因往往是多因素的，有关因素和一些无关因素之间又是彼此联系相互影响的。一些混淆因素常常掩盖某些因素与疾病的关系，如果不加控制就不能发现彼此间的真正关系。所以仅对资料进行一般的分析是不够的。

我们上面讨论了Mantel-Haenszel检验程序及在其基础上发展的方法，也介绍了Woolf检验方法Miettinen标化相对危险度方法、这些方法解决了一个或若干混淆因素同时调整问题，也解决了把暴露水

平分为两个或多于两个水平的分析问题。这些方法是建立在分层分析方法的基础上的。分层分析方法在病例对照研究中得到广泛的应用，同样也适用于队列研究分析〔6〕。

参考文献

1. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies. J Natl Cancer Inst 1959; 22: 719.
2. Hauck WW. The large sample variance of the Mantel-Haenszel estimator of a common odds ratio. Biometrics 1979; 35: 812.
3. Schlesselman JJ. Case-control studies: Design, conduct, analysis. New York, Oxford University Press, 1982.
4. Miettinen OS. Standization of risk ratios. Am J Epidemiol 1972; 96: 383.
5. Mantel N: Chi-square tests with one degree of fredom extension of the Mantel-Haenszel procedue. J Am Statist Assoc 1963; 58: 690.
6. Breslow NE. Elementary methods of cohort analysis. Int J Epidemiol 1984, 13: 112.

（本文承蒙耿贯一教授审阅，谨此致谢）

一起由侵袭性大肠杆菌引起的食物中毒

北京市西城区卫生防疫站　张志果　杨好施　罗　寒　陆秀娟　马艳华

1986年11月份我市西城区某大学院校，因就餐快点盒饭而导致一起76人食物中毒，经流行病学调查及病原学鉴定，证实为侵袭性大肠杆菌所引起。现报告于下：

一、流行病学调查：该校因有在外演出任务，于17点30分晚餐进食快点盒饭。演出人数为80人。进食快点盒饭者76人，其中四人未吃快点盒饭。快点盒饭为不透明塑料包装，内有米饭3～4两，菜为胡萝卜土豆烧牛肉，并有白色熟鸡块。进餐盒饭时，即有感馊味，约八小时后，即翌日凌晨1点开始出现症状，进食者，无一幸免，全体出现中毒症状，仅轻重不同而已。其中一人较重者。仅进食四分之一饭盒，即腹泻5～7次。中毒患者中，女性较男性为重，而未进食盒饭的四人，无一人出现症状，在校所有学生职工，也无一人出现症状，全部正常。76位患者经治疗，无一例死亡。

二、临床症状：发病开始首先出现头痛、头晕、恶心，继之于腹部脐周出现阵发性腹痛，而后出现腹泻，多者5～7次，一般为3～4次。腹泻呈粘液状稀便，便镜检，有的可见白细胞。腹部有里急后重感，似菌痢样。无呕吐者；体温正常。

三、采取样品及实验室检查：根据中华人民共和国食品卫生检验方法规定进行采样，计剩馀快点盒饭二件，中毒患者大便五份、血二份，送交实验室进行病原菌鉴定，实验室经培养，挑选可凝菌落，进行生化学及血清学鉴定：于快点盒饭及中毒患者大便中，同时分离出侵袭性大肠杆菌$O28acK_{73}$和$O112acK_{66}$两型侵袭性大肠杆菌，并且作了初期患者血清与恢复期患者血清抗体滴度测定，与$O28acK_{73}$和O1 12acK分离的两型菌株，其抗体滴度分别增长16倍和32倍，证实该两型菌株确属为该次引起食物中毒的病原菌。

四、小结：这次发生的食物中毒十分典型。首先人数集中，凡进食快点盒饭者，全部出现中毒症状，而病发时间也集中，未检出其它病原菌。

从本文资料来看，应加强执行有关食品卫生管理法规，尤当认真对盒饭以及熟食肉类的管理，是完全必要的，以保证人民的健康。