

· 讲 座 ·

病例-对照研究中四格表资料精确分析法

上海医科大学 詹绍康

病例-对照研究所得资料中最简单的是四格表。用精确法分析四格表资料，不论观察例数多少，都可获得合理的统计推断。当样本含量较小时，应采用精确法分析，不宜用近似法分析。

相对危险度和比数比

在因果关系的定群研究 (cohort study) 中，暴露于危险因子与不暴露于危险因子的人群发病率或患病率之比称为相对危险度 (relative risk) 或危险比 (risk ratio)，常以RR表示：

$$RR = \frac{\text{暴露人群发病率(或患病率)}}{\text{非暴露人群发病率(或患病率)}} \quad (1)$$

表1 病例-对照研究资料的一般形式

	病例	对照	合计
暴露	a	c	a+c=m ₁
非暴露	b	d	b+d=m ₀
合计	a+b=n ₁	c+d=n ₀	n

然而，在病例-对照研究 (case-control study) 中，由于病例的抽样比例K₁远大于对照的抽样比例K₀，所以a/(a+c)及b/(b+d)并不是发病率，两

表2 说明ψ和RR关系的两个例子

假设的 四格表资料	a	c	50	10	40	10
	b	d	50	90	30	60
从样本算得的ψ值			9		8	
假设的K ₀ /K ₁ 值	总人群发病率		真正的RR		真正的RR	
0.1	0.0909		6.33		6	
0.01	0.0099		8.62		7.73	
0.001	0.000999		8.96		7.97	
0.0001	0.0009999		8.99		7.9972	

表2中两个假设的数字例子显示，如果总人群发病率都为9.09%，则两抽样比例的比值为K₀/K₁=0.1，从样本资料算得的比数比ψ (9和8) 与真正的相对危

险者之比值也不是相对危险度。当K₁和K₀已知时，相对危险度的计算公式为：

$$\widehat{RR} = \frac{a(\frac{K_0}{K_1} \cdot b + d)}{b(\frac{K_0}{K_1} \cdot a + c)} \quad (2)$$

当K₀/K₁极小时， $\frac{K_0}{K_1} \cdot b$ 及 $\frac{K_0}{K_1} \cdot a$ 可近似地看作0而略去不计，这样算得的值称为比数比(odds ratio)。比数比常用符号OR或ψ表示：

$$\widehat{\psi} = \widehat{OR} = \frac{ad}{bc} \quad (3)$$

在实际的流行病学研究中，如果某种疾病的发病率(患病率)很低，病例-对照研究中的病例数n₁和对照例数n₀相差不太悬殊的情况下， $\frac{K_0}{K_1}$ 值显然很小。此时ψ接近于RR，比数比是相对危险度的良好估计值。然而，对于发病率(患病率)不是很低的疾病， $\frac{K_0}{K_1}$ 值并不太小，把 $\frac{K_0}{K_1} \cdot b$ 和 $\frac{K_0}{K_1} \cdot a$ 略而不计会造成ψ与RR较大的差别，也就是说在这种情况下比数比并不是相对危险度的良好估计值。

险度(6.33和6)相差较大。因此，在发病率不太低时用比数比来估计相对危险度应十分谨慎。

对总体比数比的假设作检验

研究者可以假设总体比数比为任何有意义的数值, 然后可对此假设 H_0 作检验。

一、对 $H_0: \Psi = \Psi_0$ 作检验: 在总体比数比 Ψ 等于指定值 Ψ_0 的假设下(严格地说, 还应在四个边缘合计为 n_1, n_0, m_1 和 m_0 的条件下), 四格表左上格取值为 a 的概率, 或者说四个格子内取值为 a, b, c 和 d 的概率 P_a 为:

$$P_a = \frac{C_{n_1}^a C_{n_0}^{m_1-a} \Psi_0^a}{\sum_i C_{n_1}^i C_{n_0}^{m_1-i} \Psi_0^i} \quad (4)$$

式中: $0, m_1 - n_0 \leq i \leq m_1, n_1$ 。意即 i 要取遍所有可能取的值, 或者说左上格频数 i 最小是 0 和 $m_1 - n_0$ 中的较大者, 最大是 m_1 和 n_1 中的较小者。如备选假设为 $H_1: \Psi < \Psi_0$, 则应计算

$$P_L = \sum_{i \leq a} P_i$$

如备选假设为 $H_1: \Psi > \Psi_0$, 则应计算

$$P_U = \sum_{i \geq a} P_i$$

然后以 $P(P_L$ 或 $P_U)$ 的大小来作统计推断。若 $P \leq \alpha$, 则在 α 水准上拒绝 H_0 ; 若 $P > \alpha$, 则在 α 水准上不拒绝 H_0 。

表3 胃癌与按时用餐关系的病例-对照研究资料

	胃癌	对照	合计
用餐不按时	3	1	4
用餐按时	3	4	7
合计	6	5	11

例如, 对于表3资料, 如果要对 $\Psi = 2$ 的假设作检验, 备选假设为 $\Psi > 2, \alpha = 0.05$ 。就应计算左上格取值大于及等于3的概率之和 P 。从表3可知 i 的取值范围是 0 与 $4 - 5 = -1$ 中的较大者 0 和 4 与 6 中的较小者 4 之间, 其中符合 $i \geq 3$ 的 i 值是 3 和 4 , 则

$$P_3 = \frac{C_6^3 C_5^1 2^3}{C_6^0 C_5^4 2^0 + C_6^1 C_5^3 2^1 + C_6^2 C_5^2 2^2 + C_6^3 C_5^1 2^3 + C_6^4 C_5^0 2^4}$$

$$= \frac{800}{5 + 120 + 600 + 800 + 240} = 0.4533$$

$$P_4 = \frac{C_6^4 C_5^0 2^4}{C_6^0 C_5^4 2^0 + C_6^1 C_5^3 2^1 + C_6^2 C_5^2 2^2 + C_6^3 C_5^1 2^3 + C_6^4 C_5^0 2^4}$$

$$= \frac{240}{5 + 120 + 600 + 800 + 240} = 0.1360$$

$$P = P_U = P_3 + P_4 = 0.4533 + 0.1360 = 0.5893$$

显然, $P > 0.05$, 不拒绝 $\Psi = 2$ 的假设, 没有理由说总体比数比不是 2 。

二、在特例 $\Psi_0 = 1$ 时对 $H_0: \Psi = \Psi_0$ 作检验: 以 $\Psi_0 = 1$ 代入式(4), 可得

$$P_a = \frac{C_{n_1}^a C_{n_0}^{m_1-a}}{\sum_i C_{n_1}^i C_{n_0}^{m_1-i}} \quad (5)$$

由组合运算规则可知, $C_{n_1+n_0}^{m_1} = \sum_i C_{n_1}^i C_{n_0}^{m_1-i}$, 故式(5)可变成:

$$P_a = \frac{C_{n_1}^a C_{n_0}^{m_1-a}}{C_{n_1+n_0}^{m_1}} = \frac{n_1! n_0! m_1! m_0!}{a! b! c! d! n_1!} \quad (6)$$

这就是众所周知的四格表(在 $H_0: \Psi = 1$ 条件下)计算确切概率的公式(Fisher精确概率法)。对于表3资料: 若 $H_1: \Psi > \Psi_0$, 则:

$$P_3 = \frac{6! 5! 4! 7!}{3! 3! 1! 4! 11!} = 0.3030$$

$$P_4 = \frac{6! 5! 4! 7!}{4! 2! 0! 5! 11!} = 0.0455$$

$$P = P_U = P_3 + P_4 = 0.3030 + 0.0455 = 0.3485$$

同样也是 $P > 0.05$, 不拒绝 $\Psi = 1$ 的假设。没有理由说总体比数比不是 1 。

对总体比数比 Ψ 作估计

一、总体比数比 Ψ 的点估计: 按最大似然估计法, 总体比数比 Ψ 的估计值 $\hat{\Psi}$ 可由方程(7)解得。

$$a = E(a) \quad (7)$$

式中: a 是四格表左上格的观察频数; $E(a)$ 是四格表左上格在比数比为 $\hat{\Psi}$ 时的期望频数, 其计算方法为:

$$E(a) = \sum_i P_i \cdot i \quad 0, m_1 - n_0 \leq i \leq m_1, n_1 \quad (8)$$

仍以表3资料为例, 把有关数值代入式(7), 可得:

$$3 = \frac{C_6^0 C_5^4 \hat{\Psi}^0 \cdot 0 + C_6^1 C_5^3 \hat{\Psi}^1 \cdot 1 + C_6^2 C_5^2 \hat{\Psi}^2 \cdot 2 + C_6^3 C_5^1 \hat{\Psi}^3 \cdot 3 + C_6^4 C_5^0 \hat{\Psi}^4 \cdot 4}{C_6^0 C_5^4 \hat{\Psi}^0 + C_6^1 C_5^3 \hat{\Psi}^1 + C_6^2 C_5^2 \hat{\Psi}^2 + C_6^3 C_5^1 \hat{\Psi}^3 + C_6^4 C_5^0 \hat{\Psi}^4}$$

经整理后得高次方程：

$$\hat{\Psi}^4 - 10\hat{\Psi}^2 - 8\hat{\Psi} - 1 = 0$$

用迭代法解此方程。当 $\hat{\Psi} = 3.5152165$ 时，左式 = 0.00000497，因此可取 $\hat{\Psi} = 3.515$ ，这就是总体比数比的点估计。

二、总体比数比 Ψ 的区间估计：总体比数比 Ψ 的 $100(1-\alpha)\%$ 可信区间的下限和上限可从下列方程中

$$0.025 = \frac{C_6^0 C_5^4 \Psi_U^0 + C_6^1 C_5^3 \Psi_U^1 + C_6^2 C_5^2 \Psi_U^2 + C_6^3 C_5^1 \Psi_U^3}{C_6^0 C_5^4 \Psi_U^0 + C_6^1 C_5^3 \Psi_U^1 + C_6^2 C_5^2 \Psi_U^2 + C_6^3 C_5^1 \Psi_U^3 + C_6^4 C_5^0 \Psi_U^4}$$

和

$$0.025 = \frac{C_6^3 C_5^1 \Psi_L^3 + C_6^4 C_5^0 \Psi_L^4}{C_6^0 C_5^4 \Psi_L^0 + C_6^1 C_5^3 \Psi_L^1 + C_6^2 C_5^2 \Psi_L^2 + C_6^3 C_5^1 \Psi_L^3 + C_6^4 C_5^0 \Psi_L^4}$$

经整理后分别得高次方程：

$$\hat{\Psi}_U^4 - 260\Psi^3 - 390\Psi^2 - 156\Psi_U - 13 = 0$$

和

$$58.5\Psi_L^4 + 390\Psi_L^3 - 15\Psi_L^2 - 6\Psi_L - 0.5 = 0$$

用迭代法可解得

$$261.4937137 < \Psi_U < 261.4937138$$

$$0.1695191 < \Psi_L < 0.1695192$$

所以，认为总体比数比的 95% 可信区间为 0.17 ~ 261.49。

由于样本含量很小，所以 95% 可信区间很宽。随着样本含量增大，可信区间可由宽变窄。

配对四格表资料总体比数比的假设检验

一、对 $H_0: \Psi = \Psi_0$ 作检验：配对四格表资料的一般形式可用表 4 表示。

表 4 配对四格表资料的一般形式

		病 例	
		暴 露	非暴露
对 照	暴 露	a	c
	非暴露	b	d

表 4 中的 a 和 d 不提供疾病和暴露间有无联系的信息。在总体比数比 $\Psi = \Psi_0$ 的假设 H_0 下 (严格地说还应包括在 b+c 固定的条件下)，出现所观察到的 b 和 c 的概率为：

解得：

$$\alpha/2 = \sum_{i \leq a} P_i(i/n_1, n_0, m_1, m_0; \Psi_U)$$

$$\alpha/2 = \sum_{i \geq a} P_i(i/n_1, n_0, m_1, m_0; \Psi_L) \quad (9)$$

式 (9) 右式表示在 n_1, n_0, m_1, m_0 固定及 Ψ 或 Ψ_L 指定时有关 i 取值的概率之和。

如果对表 3 资料要求总体比数比 Ψ 的 95% 可信区间，即 $\alpha = 0.05$ ，则有

$$P_b = C_{b+c}^b \Pi_0^b (1-\Pi_0)^c \quad (10)$$

$$\Pi_0 = \frac{\Psi_0}{\Psi_0 + 1} \quad (11)$$

若备选假设为 $H_1: \Psi < \Psi_0$ ，则在检验中应计算概率 $P = P_L$ ：

$$P_L = \sum_{i \leq b} P_i \quad (12)$$

若备选假设为 $H_1: \Psi > \Psi_0$ ，则在检验中应计算概率 $P = P_U$ ：

$$P_U = \sum_{i \geq b} P_i \quad (13)$$

如果 $P(P_L \text{ 或 } P_U) \leq \alpha$ ，则在 α 水准上拒绝 H_0 ，否则，不拒绝 H_0 。

例如，某研究者以配对方式调查了胃癌患者和对照者的常吃油炸食品史，得如下资料：

表 5 胃癌与常吃油炸食品关系配对调查结果

		胃 癌	
		常 吃	不常吃
对 照	常 吃	6	4
	不常吃	8	36

假设 $H_0: \Psi = 10$ ，即假设常吃油炸食品者得胃癌的危险是不常吃者的 10 倍，备选假设 $H_1: \Psi < 10$ ， $\alpha = 0.05$ ，作检验时可计算概率 $P = P_L$ 。

$$P_L = C_{12}^8 \cdot 0.9091^8 \cdot 0.0909^4 + C_{12}^7 \cdot 0.9091^7 \cdot 0.0909^5 + C_{12}^6 \cdot 0.9091^6 \cdot 0.0909^6 + C_{12}^5 \cdot 0.9091^5 \cdot 0.0909^7 + C_{12}^4 \cdot 0.9091^4 \cdot 0.0909^8 + C_{12}^3 \cdot 0.9091^3 \cdot 0.0909^9 + C_{12}^2 \cdot 0.9091^2 \cdot 0.0909^{10} + C_{12}^1 \cdot 0.9091 \cdot 0.0909^{11} + C_{12}^0 \cdot 0.0909^{12} = 0.0186$$

或

$$P_L = 1 - (C_{12}^{12} \cdot 0.9091^{12} + C_{12}^{11} \cdot 0.9091^{11} \cdot 0.0909^1 + C_{12}^{10} \cdot 0.9091^{10} \cdot 0.0909^2 + C_{12}^9 \cdot 0.9091^9 \cdot 0.0909^3) = 1 - 0.9814 = 0.0186$$

因 $P = P_L < 0.05$, 故拒绝 $H_0: \Psi = 10$, 认为 $\Psi < 10$ 。

二、在特例 $\Psi_0 = 1$ 时, 对 $H_0: \Psi = \Psi_0$ 作检验:

$$\text{以 } \Psi_0 = 1 \text{ 代入式 (11), 可得 } \Pi_0 = \frac{\Psi_0}{\Psi_0 + 1} = \frac{1}{1+1}$$

$= 0.5$ 。对于表5资料, 假设 $\Psi = 1$, 就是假设胃癌与常吃油炸食品无关。若备选假设为 $H_1: \Psi > 1$, 即胃癌与常吃油炸食品有关。仍取 $\alpha = 0.05$ 。计算检验用概率 $P = P_U$:

$$P_U = C_{12}^8 \cdot 0.5^8 \cdot 0.5^4 + C_{12}^9 \cdot 0.5^9 \cdot 0.5^3 + C_{12}^{10} \cdot 0.5^{10} \cdot 0.5^2 + C_{12}^{11} \cdot 0.5^{11} \cdot 0.5 + 0.5^{12} = 0.1938$$

$P_U > 0.05$, 不拒绝 H_0 , 没有理由说胃癌与常吃油炸食品有关。

配对四格表资料对总体比数比 ψ 作估计

一、总体比数比 Ψ 的点估计: 由配对的病例-对照研究四格表资料对总体比数比 Ψ 的最大似然估计值 $\hat{\Psi}$ 为:

$$\hat{\Psi} = \frac{b}{c} \quad (14)$$

对于表4资料, 可算得:

$$\hat{\Psi} = \frac{8}{4} = 2$$

估计常吃油炸食品者得胃癌的危险是不常吃者的2倍。

二、总体比数比 Ψ 的区间估计: 式(11)所表达的 Ψ 与 Π 的函数关系也可写成式(15)形式:

$$\Psi = \frac{\Pi}{1-\Pi} \quad (15)$$

式中:

$$\Pi = \frac{\text{配对四格表中病例有暴露而对照无暴露的对子数}}{\text{配对四格表中病例和对照的暴露情况不一致的对子总数}}$$

实际的配对病例-对照研究中, $P = \frac{b}{b+c}$ 就是 Π 的估计

值 $\hat{\Pi}$, 以 P 可求出 Π 的可信区间 $\Pi_L \sim \Pi_U$, 再以 Π_L 和代入式(15), 可求得 Ψ 的可信区间 $\Psi \sim \Psi$ 。

Π 的 $100(1-\alpha)\%$ 可信区间的下限 Π_L 和上限可从以二项分布原理制成的百分比(比例)可信限表查得, 也可由下列公式算得:

$$\Pi_L = \frac{b}{b + (c+1)F_{\alpha(2c+2, 2b)}} \quad (16)$$

$$\Pi_U = \frac{(b+1)F_{\alpha(2b+2, 2c)}}{c + (b+1)F_{\alpha(2b+2, 2c)}} \quad (17)$$

式中 $F_{\alpha(2c+2, 2b)}$ 及 $F_{\alpha(2b+2, 2c)}$ 是指自由度为 $2c+2, 2b$ 或 $2b+2, 2c$ 时的双侧概率为 α 的 F 值, 所以应查方差分析用 F 值表中的 $F_{\alpha/2}$, 或查方差齐性检验用 F 值表中的 F_{α} 。

对于表5数据, 若要计算95%可信区间, 可查方差分析用 F 值表的 $F_{0.025}$ 或方差齐性检验用表的 $F_{0.05}$ 。以自由度 $2c+2 = 2 \cdot 4 + 2 = 10$ 及 $2b = 2 \cdot 8 = 16$ 在方差齐性检验用表中查得 $F_{0.05}(10, 16) = 2.99$, 以自由度 $2b+2 = 2 \cdot 8 + 2 = 18$ 及 $2c = 2 \cdot 4 = 8$ 在方差齐性检验用表中查得 $F_{0.05}(18, 8) = 4.06$, 因而:

$$\Pi_L = \frac{8}{8 + (4+1)2.99} = 0.3486$$

$$\Pi_U = \frac{9 \cdot 4.06}{4 + (8+1)4.06} = 0.9013$$

以 Π_L 与 Π_U 代入公式(15), 可得:

$$\Psi_L = \frac{\Pi_L}{1-\Pi_L} = \frac{0.3486}{1-0.3486} = 0.5352$$

$$\Psi_U = \frac{\Pi_U}{1-\Pi_U} = \frac{0.9013}{1-0.9013} = 9.1317$$

因此由表5资料算得总体比数比的95%可信区间为 $0.54 \sim 9.13$ 。

如果用查表的方法, 可用 $n = b+c = 12, x = b = 8$, 从百分比可信限表直接查得 Π 的95%可信区间为 $0.35 \sim 0.90$, Π 的99%可信区间为 $0.27 \sim 0.94$, 再以公式(15)求得比数比 Ψ 的95%可信区间为 $0.54 \sim 9.00$, 99%可信区间为 $0.37 \sim 15.67$ 。

几点说明

一、从统计学上讲, 四格表

a	c	m_1
b	d	m_0
n_1	n_0	n

有三种情况：

1. n_1, n_0, m_1, m_0 是固定的；
2. n_1 和 n_0 是固定的；
3. n 是固定的。

在一些流行病学研究中，可能是调查含量为 n_1 和 n_0 的两个独立样本，那么所得四格表资料应属第(2)种情况。例如，随机抽查甲乙两方法免疫接种者 n_1 和 n_0 人，比较其中发生副反应者的比例；或者随机抽查大肠癌患者与对照者 n_1 和 n_0 人，比较其中常吃绿色蔬菜者的比例。此时，只有 n_1 和 n_0 是固定的，其中的阳性数（发生副反应或常吃绿色蔬菜的人数）都是随机的，两组的阳性数之和也是随机的。而两组的阳性数在理论上分别服从于总体比例为 Π_1 和 Π_0 的两独立二项分布。因此，如果要用这种四格表资料对假设 $H_0: \Pi_1 = \Pi_0$ 或 $H_0: \Psi = \Psi_0$ 作检验，就应计算两个独立二项分布的联合概率。然而，经数理统计学者研究，即使属于上述第(2)种情况的四格表，仍可按第(1)种情

况来作统计处理，即用超几何分布的方法计算精确概率以作假设检验。

二、流行病学研究中的四格表资料，常见的有定群研究和病例-对照研究两类。对于定群研究资料，可估计相对危险度RR，并对相对危险度的假设作检验；对于病例-对照研究，可估计比数比 Ψ ，并对比数比的假设作检验。当发病率或患病率很低时，比数比 Ψ 是相对危险度RR的良好估计值。在实际工作中，定群研究往往以随访形式获得资料，所以不常以简单的四格表来表示研究结果，而用适合于对随访资料作分析的生存分析(survival analysis)等方法来作统计分析。

三、四格表资料以精确法来分析，其优点是“精确”——可获“确切”的概率，而不是一种近似的概率。精确法的计算过程往往比较繁复。如果样本含量不小，或者四格表中各格子里的观察数都不太小时，用精确法来分析，计算工作量很大。而此时用近似法来分析不仅近似得较好，而且会大大减少计算工作量。所以精确分析法往往用于样本含量较小的情况。

普种流脑菌苗对于控制流脑周期性流行高峰年的观察

江西萍乡市卫生防疫站 包林

为了控制八十年代流脑流行，1985、1986两年我们对部分人群进行了流脑多糖菌苗接种，1987年又进行了大面积接种，并对发病及免疫情况进行了观察。

根据多年的流脑监测表明，我市的流脑流行菌群为A群，带菌以B群为主。据疫情统计全市1985年流脑发病211例，发病率为16.72/10万，出现了4个疫情爆发点，1986年发病146例，发病率为11.50/10万，出现了3个疫情爆发点。从我市流脑流行为8~10年的规律分析及疫情监测表明，1985年为流脑上升年，1986~87年为流脑流行年。为了打破及削平流脑流行的高峰，降低发病率1987年我们对全市15岁以下的儿童普遍接种A群流脑多糖菌苗，接种率为91%。

免疫后共观察215 610人，其中免疫组和对照组观察人数分别为122 326人和93 284人，两组病例数和发病率为1、0.81/10万和29、31/10万，效果指数为1:38.3，保护效价为97.4%，两组差别显著($\chi^2=34.86$,

$P<0.005$)。

1987年流脑发病32例，发病率为2.45/10万，比1986年下降了78.7%，由预测的流行年转为低散发年。经流行病学调查核实1987年流脑发病者中有接种史1人，占病例数的3.1%，无接种史占病例数的96.9%，死亡1例(无接种史)，病死率为3.1%，发病年龄在15岁以下的占病例的90.6%。

流脑流行年大面积接种流脑疫苗后，发病率大幅度下降，打破了流行规律，削平了流行期高峰，病例显高度散发分布。1987年流脑监测发现人群流脑A群带菌率与流行强度不成正比，这说明普遍免疫后人群A群抗体水平的高低并不影响A群的带菌率。普种流脑菌苗后，发病年龄集中在15岁以下，提示我们今后流脑菌苗接种的重点仍然是15岁以下的儿童。

(参加工作的还有：欧书勇、罗云江、陈金霞、苏接发、陆志雄、刘清、刘秀贞)