Определение скрытых зависимых переменных при анализе зависимостей

Статистический анализ данных

Д.Д. Яшков

МФТИ(ГУ)

09 декабря 2013 г.

Определение

При анализе заболевания часто исследуется влияние конкретного фактора на исход. Скрытой зависимой переменной называется такая переменная, который влияет и на заболевание и на фактор. Возможно два случая:

Confounder	Level 1	High risk for disease High prevalence of exposure
	Level 2	Low risk for disease Low prevalence of exposure

Этот случай называется положительной зависимостью

Confounder	Level 1	High risk for disease Low prevalence of exposure
	Level 2	Low risk for disease High prevalence of exposure

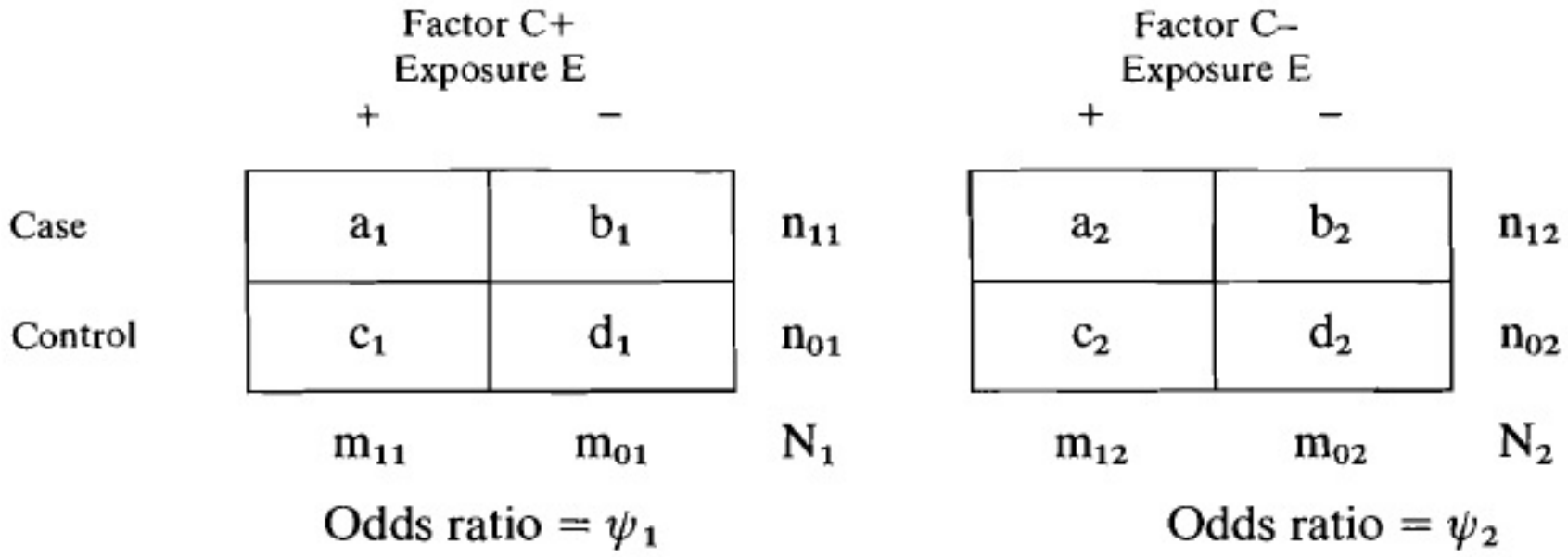
А этот – отрицательной.

	Exposure E +	Exposure E −	
Case	a	b	n_1
Control	c	d	n_0
	m_1	m_0	N

Относительная зависимость фактора E и заболевания выражается как:

$$\psi_p = \frac{ad}{bc}$$

Теперь, предположим что есть фактор C, который потенциально может быть зависимой переменной. Сделаем расслоение по фактору C:



Будем считать, что $\psi_1 = \psi_2 = \psi$. Фактор C будет зависимой переменной, тогда и только тогда, когда он зависит и от E и от заболевания, т.е: $\psi \neq \psi_p$.

Степень зависимости

Степень скрытой зависимости определяется по следующей формуле:

$$w = \frac{\psi_c p_1 + (1 - p_1)}{\psi_c p_2 + (1 - p_2)},$$

где $p_1 = \frac{c_1}{c_1 + c_2}$, $p_2 = \frac{d_1}{d_1 + d_2}$.
Далее будет приведена таблица значений w при различных p_1, p_2, ψ_c.

Table 3.4 Confounding risk ratios associated with varying relative risk (ψ'_c), frequency of occurrence of the confounding variable among controls exposed to E (p_1) and not exposed to E (p_2)

$\psi'_c = 2$

Value of p_2	Value of p_1 0.1	0.3	0.5	0.8
0.1	1	1.18	1.36	1.64
0.3	0.85	1	1.15	1.38
0.5	0.75	0.87	1	1.20
0.8	0.61	0.72	0.83	1

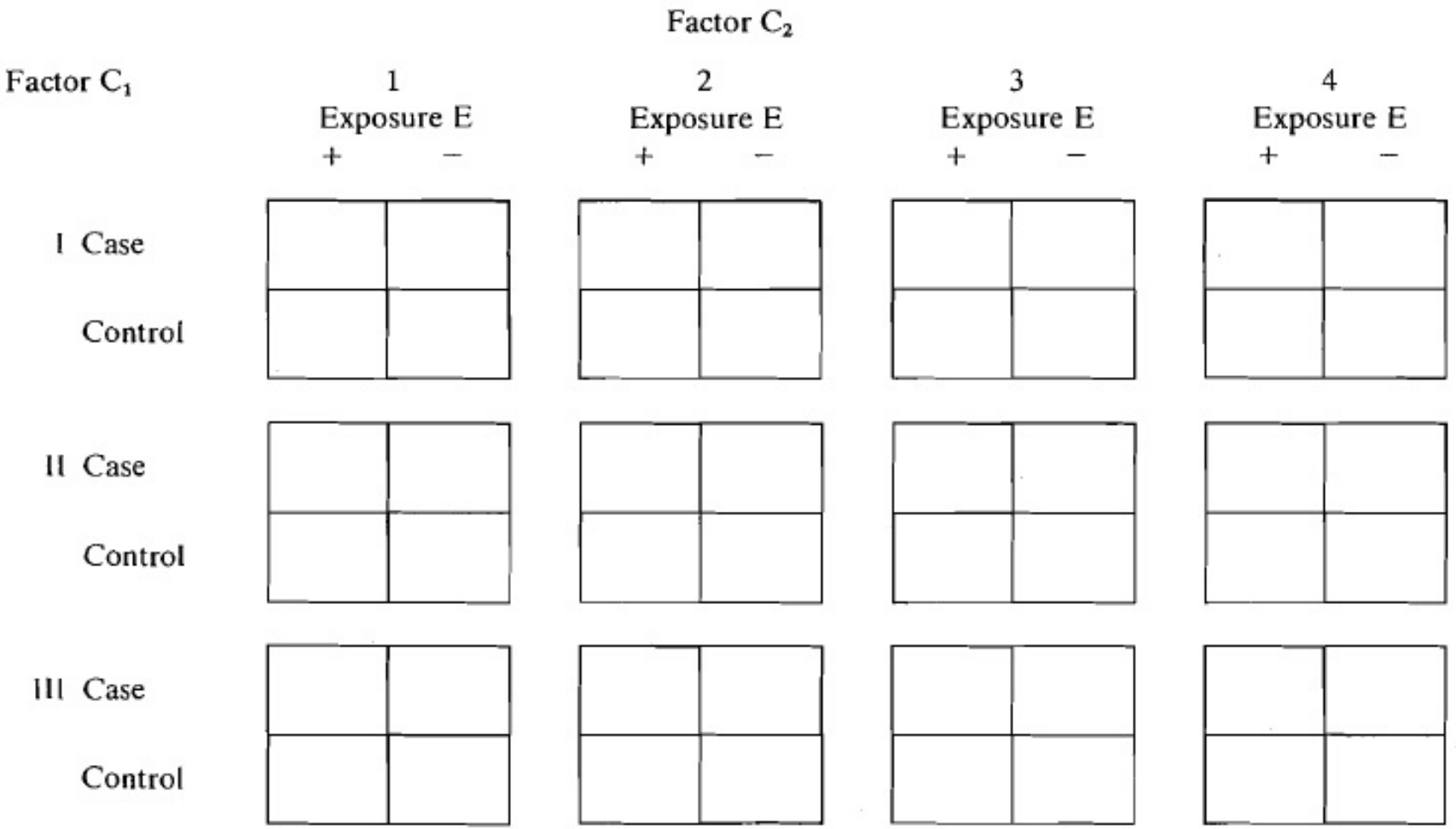
$\psi'_c = 5$

Value of p_2	Value of p_1 0.1	0.3	0.5	0.8
0.1	1	1.57	2.14	3.00
0.3	0.64	1	1.36	1.91
0.5	0.47	0.73	1	1.40
0.8	0.33	0.52	0.71	1

$\psi'_c = 10$

Value of p_2	Value of p_1 0.1	0.3	0.5	0.8
0.1	1	1.95	2.80	4.32
0.3	0.51	1	1.49	2.22
0.5	0.35	0.67	1	1.49
0.8	0.23	0.45	0.67	1

Кол-во переменных больше одной, или у переменной более одного уровня



Степень расслоения

Пусть скрытая зависимая переменная C имеет K уровней и после расслоения исходных данных на K таблиц мы посчитаем относительные риски $r_1, \ldots, r_k$, и каждый уровень появляется с частотой p_{1i} среди всех неболеющих, у которых значение фактора $E = +$. p_{2i} – аналогично с $E = -$. Тогда степень зависимости фактора C считается следующим образом:

$$w = \frac{\sum_{k=1}^{K} p_{1k} r_k}{\sum_{k=1}^{K} p_{2k} r_k}$$

Группировка

Пусть теперь K слоев мы сгруппировали в J групп. Тогда соответственно мы получим:

$$w^* = \frac{\sum_{j=1}^{J} p_{1j}^* r_j^*}{\sum_{j=1}^{J} p_{2j}^* r_j^*} \tag{1}$$

Можно рассмотреть ещё более подробно, если посчитать внутри каждой из групп w_j^*, тогда можно выразить w через w_j^*:

$$w = \frac{\sum_{j=1}^{J} w_j^* p_{1j}^* r_j^*}{\sum_{j=1}^{J} p_{2j}^* r_j^*}$$

Отношение w и w^* описывает остаточный эффект после группировки.

Пример

Table 3.5 Residual confounding effects after various degrees of stratification by cigarette consumption

Average daily cigarette consumption	Risk ratio[a]	p_{2i}[a] %	p_{1i} %	Grouping, with residual confounding risk ratio within each group						
				I	II	III	IV	V	VI	VII
0	1	38	10	1.0	1.0	1.39	1.39	1.0	1.57	1.31
1–4	5.6	3	2	0.97	1.05			1.37		
5–9	3.2	8	7							
10–14	9.4	10	9	1.01		1.01				
15–19	11.3	10	12							
20–24	23.2	15	15	1.01	1.16	1.01	1.28		1.16	1.34
25–29	24.9	7	15							
30–34	38.2	5	15	1.02		1.02				
35–40	50.7	4	15							
Confounding risk ratio = 1.93				1.91	1.68	1.86	1.51	1.41	1.61	1.44

[a] Data adapted from Doll and Peto (1978)