

# Stratified Data Analysis

## II. Case-Control Studies

Case control study の場合は下の表のような形になります。

	exposed	unexposed	total
cases	a	b	$M_1$
non-case	c	d	$M_0$
total	$N_1$	$N_0$	T

性別や年齢など confounder variable によって層化すると(stratify)、いくつかの似たような表ができあがります。i = 1, 2, 3, ..., l 番目の strata としますと stratum i は以下のように表されます。

Example age l group

	exposed	unexposed	total
cases	$a_i$	$b_i$	$M_{1i}$
non-case	$c_i$	$d_i$	$M_{0i}$
total	$N_{1i}$	$N_{0i}$	T

1つの臨床研究で confounder はしばしば複数考えられます。例えば避妊薬(OC)の心筋梗塞(MI)発生リスクについて調べた場合、結果は当然年齢に影響されますし、アルコールの量にも影響されるかもしれません。以下の例題をもとに前章と同様、年齢が confounder であるかどうか、それを排除したあとも OC 使用と MI 発生に相関があるかについて検討していきましょう。

MI 発生と OC 使用の関係を示した表です。

年齢	25-29		total
OC	Y	N	N
MI	23	112	135
contro	130	1306	1436
I			
total	153	1418	1571
OR	2.06		

全体で見ると OC 使用者は非使用者に比べ 2.06 倍の頻度で心筋梗塞に罹患しています。しかし、OC は若い人の方が多く使うでしょうし、心筋梗塞は若い人では少ないので、その辺はどうなっているのでしょうか？年齢別にデータを整理してみましょう。

年齢	25-29		total
OC	Y	N	N
MI	4	2	6
contro	62	224	286
I			
total	66	226	292
OR	7.2		

年齢	30-34		total
OC	Y	N	N
MI	9	12	21
contro	33	390	423
I			
total	42	402	444
OR	8.9		

年齢	35-39		total
OC	Y	N	N
MI	4	33	37
contro	26	330	356
I			
total	30	363	393
OR	1.5		

年齢	40-44		total
OC	Y	N	N
MI	6	65	71
contro	9	362	371
I			
total	15	427	442
OR	3.7		

$H_0$  = OC使用とMI発生に関係がない。OR=1  
 $H_A$ =OC使用とMI発生に関係がある。OR  $\neq$  1

基本は

$$\chi^2 = [X - E(X/H_0)]^2 / \text{var}(X/H_0)$$

$X$  = 暴露された中での患者数 =  $a$

$E(X/H_0)$  = confounder が存在しない状況下で暴露された中から発生することが期待される患者数 =  $M_1(N_1/T)$

$\text{Var}(X/H_0) = M_1M_0N_1N_0/T^2(T-1)$  : open cohort と異なるので注意してください。

され confounder を打ち消すために層化(stratification)しましたが、それぞれの表  $i$  を合算しなくてはなりません。基本公式は crude table のものと類似しています。

$$\chi^2 = [X_i - E_i(X_i/H_{0i})]^2 / \text{var}_i(X_i/H_{0i})$$

は最初から  $i$  個の表を全部合算するという意味です。

$X_i$  = 暴露された中での患者数 =  $a_i$

$E_i(X_i/H_{0i})$  = confounder が存在しない状況下で暴露された中から発生することが期待される患者数 =  $M_{1i}(N_{1i}/T_i)$

$\text{Var}_i(X_i/H_{0i}) = M_{1i}M_{0i}N_{1i}N_{0i}/T_i^2(T_i-1)$

自由度は 1 です。

$$X_i = 23$$

$E_i(X_i/H_{i0})$  と  $\text{Var}_i(X_i/H_{0i})$  を表から計算します。

年齢	25-29		total
OC	Y	N	N
MI	4	2	6
contro	62	224	286
I			
total	66	226	292
OR	7.2		
$E_i(X_i/H_{i0})$	$6 \times 66 / 292 =$		
	1.36		
$\text{Var}_i(X_i/H_{0i})$	$6 \times 286 \times 66 \times 226 / 29$		
	$2^2 \times 291 =$		
	1.03		

年齢	30-34		total
OC	Y	N	N
MI	9	12	21
contro	33	390	423
I			
total	42	402	444
OR	8.9		
$E_i(X_i/H_{i0})$	$21 \times 42 / 444 =$		
	1.99		
$\text{Var}_i(X_i/H_{0i})$	$21 \times 423 \times 42 \times 402 / 4$		
	$44^2 \times 443 =$		
	1.72		

年齢	35-39		total
OC	Y	N	N
MI	4	33	37
contro	26	330	356
I			
total	30	363	393
OR	1.5		
$E_i(X_i/H_{i0})$	$37 \times 30 / 393 =$		
	2.82		
$\text{Var}_i(X_i/H_{0i})$	$37 \times 356 \times 30 \times 363 / 3$		
	$93^2 \times 392 =$		
	2.37		

年齢	40-44		total
OC	Y	N	N
MI	6	65	71
contro	9	362	371
I			
total	15	427	442
OR	3.7		
$E_i(X_i/H_{i0})$	$71 \times 15 / 442 =$		
	2.41		
$Var_i(X_i/H_{i0})$	$71 \times 371 \times 15 \times 427 / 4$		
	$42^2 \times 441 =$		
	1.96		

$$E_i(X_i/H_{i0}) = 8.58$$

$$Var_i(X_i/H_{i0}) = 7.08$$

よって

$$\chi^2 = [X_i - E_i(X_i/H_{i0})]^2 / var_i(X_i/H_{i0}) = (23 - 8.58)^2 / 7.08 = 29.4$$

$$Pr(\chi^2 > 29.4) = 0.000001$$

$H_0$ は棄却され、喫煙と冠動脈疾患死亡率の間に関連があると結論できます。もちろん年齢についてはある程度controlしていますが、さらに年齢層を5年でなく更にこまかくすることもできるので、confounderが残存していないとも限りません(residual confounder)。また他のconfounder(飲酒、性別など)、selection bias, information biasが存在しないことが前提となります。

年齢で層化していない場合のp valueは0.003であったのに対して、年齢(confounder)で層化した後のp valueは0.000001と大分小さくなっています。年齢が上がれば上がるほどOC使用が減る傾向にあり、年齢が高くなればなる程MIの発症も増える傾向にあります。つまりconfounderはnegativeの方向に働くので、confounderを調整しないと結果が過小評価されてしまいます。

さてそれでは何倍に増加させるのでしょうか？

Open cohortの場合は

$$w_i = b_i N_{1i} / T_i$$

ですが、case control studyにおけるMantel-Haenszel weightsは以下のようになります。

$$w_i = b_i c_{1i} / T_i$$

他のweightの方法もありますが、現在ではMHを先に述べた理由からMHを使用します。

$$OR_{MH} = w_i OR_i / \sum w_i = b_i c_{1i} / T_i (a_i d_i / b_i c_i) / ( \sum b_i c_i / T_i ) = ( a_i d_i / T_i ) / ( \sum b_i c_i / T_i )$$

$$95\%CI = \ln(OR_{MH}) \pm 1.96 \sqrt{var(X)}$$

$var(X)$  = computer programによって算出すると0.075となります。

$$\ln^{(4.00)} \pm 1.96 \sqrt{0.075} = e^{(0.847, 1.921)} = (2.33, 6.83)$$

1986年 Robins, Greenland, & Breslow (RGB)の公式が発表されるまでは下記公式を使用していました。しかし(必ずしもそうではないのですが)、この公式では本当の  $\text{var}(\ln OR)$  を過小評価し誤った結果に導く可能性があります。SASはまた古い公式を使用しているため、問題があります。

$$e^{\ln(ORMH) \pm 1.96 \cdot [\ln(ORMH)]^{2/3} / \sqrt{2}} \cdot \dots$$

$$= (2.42, 6.58)$$

年齢	25-29		total
OC	Y	N	N
MI	4	2	6
control	62	224	286
total	66	226	292
OR	7.2		
Var ln(OR)	0.771		
$W_i$	1.298		
$w_i [\ln(OR_i) - \ln(OR_{MH})]^2$	0.4578		

年齢	30-34		total
OC	Y	N	N
MI	9	12	21
control	33	390	423
total	42	402	444
OR	8.9		
Var ln(OR)	0.227		
$W_i$	4.399		
$w_i [\ln(OR_i) - \ln(OR_{MH})]^2$	2.803		

年齢	35-39		total
OC	Y	N	N
MI	4	33	37
control	26	330	356
total	30	363	393
OR	1.5		
Var ln(OR)	0.322		
$W_i$	3.108		
$w_i [\ln(OR_i) - \ln(OR_{MH})]^2$	2.822		

年齢	40-44		total
OC	Y	N	N
MI	6	65	71
control	9	362	371
total	15	427	442
OR	3.7		
Var ln(OR)	0.295		
$W_i$	3.379		

$w_i [\ln(OR_i) - \ln(OR_{MH})]^2$	0.0175
------------------------------------	--------

$$OR_{MH} = (a_i d_i / T_i) / (b_i c_i / T_i) = (3.07 + 7.91 + 3.36 + 4.91) / (0.42 + 0.89 + 2.18 + 1.32) = 19.25 / 4.81 = 4.00$$

$$95\%CI = \ln(OR_{MH}) \pm 1.96 \sqrt{\text{var}(X)}$$

$$\text{var}(X) = 1/a + 1/b + 1/c + 1/d$$

$$\chi^2 = 0.4578 + 2.803 + 2.822 + 0.0175 = 6.10$$

$$\Pr [\chi^2_{i-1=3} > 6.10] = 0.12$$

各年齢層のORは同じである (homogeneity) とする仮説は棄却できず、effect modification の証明はなりません。Effect modification を統計学的にいうには相当の違いがないと証明しにくいとされています。つまり、 $H_0$ が棄却されない理由は、effect modification が存在しないというよりは、パワーが足りない場合が多いのです。もしも $H_0$ が棄却された場合には全部を合わせた表を示すことは意味がなく、それぞれの表毎に（ここでは年齢層毎に）OR、95%CIを示すべきでしょう。

上の例では年齢を confounder として調べてきました。いくら不必要に confounder を選んでも bias になることはありませんが、efficiency を落とす事になります。

### STATAによる解析

STATA を用いて経口避妊薬使用 (OC) と心筋梗塞発症 (MI) の関連を解析します。23 人は OC を使用し MI に罹患、130 人は OC を使用しなかったが MI に罹患、112 人は OC を使用したが MI を発症せず、1306 人は OC を使用せず MI を発症しなかった人々です。

```
. input Y X count
      Y      X      count
1.  1  1  23
2.  1  0 130
3.  0  1 112
4.  0  0 1306
5. end
```

Case control study として指示をだします。

```
. cc Y X [freq=count], woolf
```

	X		Proportion	
	Exposed	Unexposed	Total	Exposed
Cases	23	130	153	0.1503
Controls	112	1306	1418	0.0790
Total	135	1436	1571	0.0859
	Point estimate		[95% Conf. Interval]	
Odds ratio	2.063049		1.272071	3.345861 (Woolf)
Attr. frac. ex.	.5152806		.2138805	.7011232 (Woolf)
Attr. frac. pop	.0774605			

chi2(1) = 8.95 Pr>chi2 = 0.0028

オッズ比(Odds ratio) とその 95%CI が算出されました。その範囲は 1 を含んでいないので有意です。カイ 2 乗検定でも p は 0.0028 と有意差の存在を示しています。つまり OC 使用は MI 発症を助長するといえます。

次に層化したデータを用いて計算します。ここでは年齢層に従って 25-29 歳を Z=0、30-34 歳を Z=1、35-39 歳を Z=2、40-44 歳を Z=3 としてデータを入力します。

```
. input Z Y X count

          Z          Y          X          count
1.  0  1  1  4
2.  0  1  0  2
3.  0  0  0  1  62
4.  0  0  0  224
5.  1  1  1  9
6.  1  1  0  12
7.  1  0  1  33
8.  1  0  0  390
9.  2  1  1  4
10. 2  1  0  33
11. 2  0  1  26
12. 2  0  0  330
13. 3  1  1  6
14. 3  1  0  65
15. 3  0  1  9
16. 3  0  0  362
17. end
```

計算を指示します。

```
. cc X Y [freq=count], woolf by(Z)
```

Z	OR	[95% Conf. Interval]		M-H Weight
0	.	.	.	0 (Woolf)
1	8.863636	3.481632	22.5653	.8918919 (Woolf)
2	1.538462	.5060878	4.676786	2.183206 (Woolf)
3	3.712821	1.278377	10.78323	1.323529 (Woolf)
-----				
Crude	3.947085	2.368698	6.577233	(Woolf)
M-H combined	4.563768	2.694459	7.729929	
-----				
Test of homogeneity (M-H)	chi2(2) =	5.76	Pr>chi2 =	0.0562
Test that combined OR = 1:				
	Mantel-Haenszel	chi2(1) =	39.99	
		Pr>chi2 =	0.0000	

年齢層によって OR が違うことがよくわかります。しかし、若年者では MI 発症頻度が少ないので、M-H Wight は小さくなっています。Crude OR は 3.9 であるのに対して M-H combined は 4.6 と大きくなっています。このことは年齢が confounder として OR を実際の値より小さくしていた、即ち我々は OR を過小評価していたこととなります。年齢が上がると OC 使用は減り、一方 MI 発症は増えるのでこのような状態になります。crude OR と比較しておよそ adjusted OR が 10% 以上異なる場合は confounder が存在する可能性が高くなります。もし層化した結果 Crude OR と Adjusted OR がほとんど変わらない場合 confounder の存在を否定して層化をやめた方がよいと思われます。そのままでもバイアスを生じませんが、1つのマス数が減るので効率が下がります。