流行病學方法論及實驗
The Methods of Epidemiology and Practices

解決辛普森矛盾之護理魔術

授課教師：陳秀熙 教授
許辰陽

台灣大學流行病學與預防醫學研究所

辛普森矛盾與護理照護效益評估

Can Nightingale’s Sanitary Improvement Reduce Death Rate ?

		死亡		
		是	否	
護理照護介入	是(A)	15	985	1000
	否(B)	45	955	1000
		60	1940	2000

$$RR = \frac{P_A}{P_B} = 0.33$$

Relative Risk
危險對比值

Non-slum Area

護理照護介入	死亡：是	死亡：否	
是(A)	9	891	900
否(B)	3	297	300
	12	1188	1200

$$RR = \frac{P_A}{P_B} = 1$$

Slum Area

護理照護介入	死亡：是	死亡：否	
是(A)	6	94	100
否(B)	42	658	700
	48	752	800

$$RR = \frac{P_A}{P_B} = 1$$

以隨機分派設計移除第三因子影響

Magic Method for Removing Confounding Factor (Socio-economic Status): Randomization

Non-slum Area

護理照護介入	死亡：是	死亡：否	
是(A)	6	494	500
否(B)	18	482	500

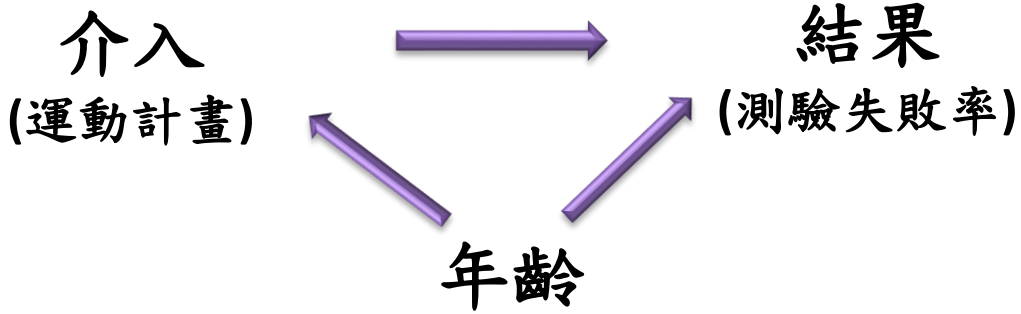
$$RR = \frac{P_A}{P_B} = 0.33$$

slum Area

護理照護介入	死亡：是	死亡：否	
是(A)	9	491	500
否(B)	27	473	500

$$RR = \frac{P_A}{P_B} = 0.33$$

辛普森矛盾之解決方案



消除 Z 之干擾

1. 研究設計使用 隨機分配(Randomization)試驗
2. 研究設計使用 **配對(Matching)**
3. 研究設計使用 **限制 (Restriction)**

4. 統計資料分析使用
   分層分析(Stratified Analysis)
   數學迴歸模式 (Regression Model)

研究設計控制干擾因子的方法

(1) 配對(matching)

Ex. 為研究新型抗生素supramycin和傳統使用之抗生素amoxicillin相比是否較容易導致患者長疹子:

Supramycin 組(暴露組)	Amoxicillin組(非暴露組)
個案1: 使用Supramycin, 無濕疹病史	使用Amoxicillin, 無濕疹病史
個案2: 使用Supramycin, 有濕疹病史	使用Amoxicillin, 有濕疹病史
個案3: 使用Supramycin, 無濕疹病史	使用Amoxicillin, 無濕疹病史
個案4: 使用Supramycin, 無濕疹病史	使用Amoxicillin, 無濕疹病史
……	……

暴露組: 非暴露組= 1:1

1:2

…

1:n

配對因子(干擾因子):濕疹病史
(用藥過敏史、家族濕疹病史)

研究設計控制干擾因子的方法

(1) 配對(matching)

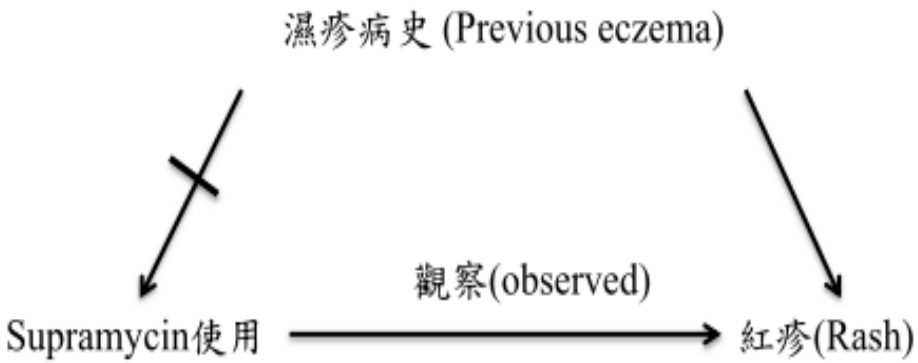
缺點:

1. 研究設計階段即決定配對因子→後續~~新增配對因子~~
2. 多個配對因子→不易配對
3. 配對因子→~~風險因子~~

研究設計控制干擾因子的方法

(2) 限制(restriction)

Ex.為研究新型抗生素supramycin和傳統使用之抗生素amoxicillin相比是否較容易導致患者長疹子→去除有濕疹病史之個案



限制與缺點: 統計檢定力↓, 缺乏外推性

研究設計控制干擾因子的方法

(2) 限制(restriction)

缺點:

1. 統計檢定力下降

2. 缺乏外推性

Darwin College at Cambridge

https://en.wikipedia.org/wiki/Darwin_College,_Cambridge#/media/File:Darwin_college.jpg

However…

運動介入 → 測驗失敗率

年齡、
性別、呼吸道疾病、代謝疾病、遺傳因子
居住區域。。。

→ 消除 Z 之干擾？

跑馬拉松=健康？美國研究說會傷腎

But…

雖然常跑馬拉松的人普遍被視為很健康，但新研究顯示，每跑完26.2哩的馬拉松，對人體的負擔相當重，除了會耗損肌肉、骨頭和心臟，對腎臟也有影響；「就像剛接受開心手術」，研究報告說。

由耶魯醫學院所做的新研究發現，80%剛跑完馬拉松的人的腎臟狀態和剛做完心臟手術的病人相似。研究的主要作者、耶魯大學腎病專家帕瑞克(Chirag Parikh)表示：「我們無法辨別剛跑完馬拉松的跑者和剛接受過心臟手術的患者的區別。」

這份於28日刊登在「美國腎臟疾病期刊」(American Journal of Kidney Diseases)上的研究，以22名(9男13女)跑了五年以上馬拉松的跑者的血液和尿液為研究樣本。研究發現，絕大多數的跑者在跑完馬拉松後的肌酸(creatine)和炎性蛋白質(inflammatory protein)水平，和患有急性腎損傷的病人類似。

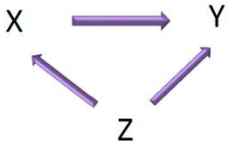
帕瑞克表示，劇烈慢跑在好幾個方面都可能會損害到人的腎臟，脫水會讓腎臟承受沉重的壓力。跑步本身也會提升人體的溫度，造成人全身性的發炎和肌肉的過勞。這些都導致人們的血液產生更多需要被過濾掉的蛋白質，致使腎臟過勞。

之後，人體為了防止過熱和雙腿無法正常活動，會將發揮降溫功能的血液輸送至皮膚和肌肉的表層，減少對腎臟的輸血。連續數小時的劇烈活動、缺水和血液供給的減少，幾乎必然會傷害人的腎臟。

這項新研究對馬拉松跑者而言並非是個噩夢，一般馬拉松跑者的身體狀態在跑完馬拉松的兩天內，能夠恢復正常。就算是那些超級馬拉松賽職業選手的身體，只要沒有什麼潛在的身體問題，最終也能夠完全恢復正常。

帕瑞克指出，倒是那些在熱帶地區、每天要在炎熱天氣下工作10至12小時的人處境令人擔心。這些工人通常都會承受和馬拉松跑者類似的生理負擔，比一般人更容易患得慢性腎臟病。

Simpson's Paradox: Solutions



消除 Z 之干擾

1. 研究設計使用隨機分配(Randomization)試驗
2. 研究設計使用限制 (Restriction)或配對(Matching)
3. 統計資料分析使用分層分析(Stratified Analysis) 或數學迴歸模式 (Regression Model)

https://udn.com/news/index

資料分析方法控制第三因子影響

1. **分層分析**

    粗率

    加權調整

        Mantel Haenszel 方法

2. **迴歸分析**

    – 邏吉思迴歸方法

資料分析方法控制第三因子影響

1.1 粗率(Crude rate)分析

(1)相差風險性(Risk difference, RD)

	E	$\overline{E}$	
D	a	b	m_1
$\overline{D}$	c	d	m_2
	n_1	n_2	n

$$\hat{RD} = \hat{P}_1 - \hat{P}_0 = 0.131 \quad \text{(點估計值)}$$

$$\hat{RD} \pm Z_{\frac{\alpha}{2}} \hat{SE}(\hat{RD}) = \hat{RD} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_0(1-\hat{P}_0)}{n_2}}$$

(大樣本理論)

粗率(Crude rate)分析：相差

(1)相差風險性(Risk difference, RD)

範例:在兒茶酚胺激素(catecholamine, CAT)與冠狀動脈血管疾病(coronary heart disease, CHD)之相關性研究中，資料結構如下，

	High CAT	*Low CAT*	
CHD	27	44	71
No CHD	95	443	538
Total	122	487	609

比較高濃度(P_1)與低濃度(P_0)的茶酚胺激素個體發生冠狀動脈血管疾病的風險：

$$\hat{P}_1 = 0.221 \qquad \hat{P}_0 = 0.090 \qquad \hat{RD} = \hat{P}_1 - \hat{P}_0 = 0.131$$

$$S\hat{E}(R\hat{D}) = \sqrt{\frac{0.22(1-0.22)}{122} + \frac{0.090(1-0.090)}{487}} = 0.040 \qquad 95\%\text{CI}=(0.053\text{-}0.209)$$

粗率(Crude rate)分析: 相比

(2)風險比值或相對風險(Risk Ratio or Relative Risk)

	E	$\overline{E}$	
D	a	b	m_1
$\overline{D}$	c	d	m_2
	n_1	n_2	n

$$\hat{RR} = \hat{P}_1 / \hat{P}_0 \quad \text{(點估計值)}$$

$$S\hat{E}\{\log(R\hat{R})\} \approx \sqrt{\frac{1-\hat{P}_1}{n_1\hat{P}_1} + \frac{1-\hat{P}_0}{n_2\hat{P}_0}}$$

$$\log(\mathrm{R}\hat{\mathrm{R}}) \pm \mathrm{Z}_{\frac{\alpha}{2}} \sqrt{\frac{1-\hat{P}_1}{n_1\hat{P}_1} + \frac{1-\hat{P}_0}{n_2\hat{P}_0}}$$

$$\hat{\mathrm{OR}} = \frac{\hat{P}_1 \ / \ (1-\hat{P}_1)}{\hat{P}_0 / (1-\hat{P}_0)} \quad \text{(點估計值)}$$

$$S\hat{E}\{\log(O\hat{R})\} \approx \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

$$\log(\mathrm{O}\hat{\mathrm{R}}) \pm \mathrm{Z}_{\frac{\alpha}{2}} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

資料分析方法控制第三因子影響

1.2 分層分析(stratified analysis)

(1)2x2聯列表之干擾因子調整

干擾因子F有k個層別

Q1. 調整F因子後的OR=?

Q2. 各層中的OR是否夠接近?

F_1

	E	$\overline{E}$	
D	a_1	b_1	m_{11}
$\overline{D}$	c_1	d_1	m_{21}
	n_{11}	n_{21}	n_1

F_2

	E	$\overline{E}$	
D	a_2	b_2	m_{12}
$\overline{D}$	c_2	d_2	m_{22}
	n_{12}	n_{22}	n_2

.....

F_i

	E	$\overline{E}$	
D	a_i	b_i	m_{1i}
$\overline{D}$	c_i	d_i	m_{2i}
	n_{1i}	n_{2i}	n_i

F_k

	E	$\overline{E}$	
D	a_k	b_k	m_{1k}
$\overline{D}$	c_k	d_k	m_{2k}
	n_{1k}	n_{2k}	n_k

1.2 分層分析(stratified analysis)

(2)調整後之勝算比(Adjusted Odds Ratio)

(A)合併加權估計式(Pooled estimate)

F_1

	E	$\overline{E}$	
D	a_1	b_1	m_{11}
$\overline{D}$	c_1	d_1	m_{21}
	n_{11}	n_{21}	n_1

F_2

	E	$\overline{E}$	
D	a_2	b_2	m_{12}
$\overline{D}$	c_2	d_2	m_{22}
	n_{12}	n_{22}	n_2

……

F_i

	E	$\overline{E}$	
D	a_i	b_i	m_{1i}
$\overline{D}$	c_i	d_i	m_{2i}
	n_{1i}	n_{2i}	n_i

F_k

	E	$\overline{E}$	
D	a_k	b_k	m_{1k}
$\overline{D}$	c_k	d_k	m_{2k}
	n_{1k}	n_{2k}	n_k

F的第i個分層: $\hat{OR}_i = \dfrac{a_i d_i}{b_i c_i}$ $\quad \hat{Var}\{\log \hat{OR}_i\} = \sqrt{\dfrac{1}{a_i}+\dfrac{1}{b_i}+\dfrac{1}{c_i}+\dfrac{1}{d_i}}$

$\log \hat{OR}_w = \dfrac{\sum_{i=1}^{k} W_i \log(OR_i)}{\sum_{i=1}^{k} W_i}$ (合併加權的 $\log OR$) , $\quad W_i = \dfrac{1}{Var\{\log \hat{OR}\}_i}$ (以 $\log \hat{OR}$ 變異數之倒數作加權值)

$\hat{OR}_w = \exp(\log \hat{OR}_w)$

1.2 分層分析(stratified analysis)

(2)調整後之勝算比(Adjusted Odds Ratio)

(B)Mantel-Haenszel估計式 (Mantel-Haenszel estimator)

F_1

	E	$\overline{E}$	
D	a_1	b_1	m_{11}
$\overline{D}$	c_1	d_1	m_{21}
	n_{11}	n_{21}	n_1

F_2

	E	$\overline{E}$	
D	a_2	b_2	m_{12}
$\overline{D}$	c_2	d_2	m_{22}
	n_{12}	n_{22}	n_2

……

F_i

	E	$\overline{E}$	
D	a_i	b_i	m_{1i}
$\overline{D}$	c_i	d_i	m_{2i}
	n_{1i}	n_{2i}	n_i

F_k

	E	$\overline{E}$	
D	a_k	b_k	m_{1k}
$\overline{D}$	c_k	d_k	m_{2k}
	n_{1k}	n_{2k}	n_k

各分層加權($w_i = \dfrac{b_i c_i}{n_i}$) 計算： $\mathrm{O}\hat{R}_{MH} = \dfrac{\sum_{i=1}^{k}\left(\dfrac{1}{n_i}\right)a_i d_i}{\sum_{i=1}^{k}\left(\dfrac{1}{n_i}\right)b_i c_i}$

分層分析控制干擾因子實例

(2)調整之勝算比(Adjusted Odds Ratio): **範例**

在研究catecholamine (CAT) 與心臟疾病(CHD)之關係的研究中，年齡的大小(Age)及心電圖是否異常(ECG)為兩個可能的干擾因子

各分層之OR

年輕組($Age < 55$),
心電圖無異常($ECG = 0$)

	High CAT	*Low CAT*	*Total*
CHD	1	17	18
No CHD	7	257	264
Total	8	274	282

$\hat{OR}_1 = 2.16(0.25 \sim 18.58)$

年輕組($Age < 55$),
心電圖有異常($ECG = 1$)

	High CAT	*Low CAT*	*Total*
CHD	3	7	10
No CHD	14	52	66
Total	17	59	76

$\hat{OR}_2 = 1.59(0.36 \sim 6.96)$

年老組($Age \geq 55$),
心電圖無異常($ECG) = 0$

	High CAT	*Low CAT*	*Total*
CHD	9	15	24
No CHD	30	107	137
Total	39	122	161

$\hat{OR}_3 = 2.14(0.85 \sim 5.37)$

年老組($Age \geq 55$),
心電圖有異常 ($ECG = 1$)

	High CAT	*Low CAT*	*Total*
CHD	14	5	19
No CHD	44	27	71
Total	58	32	90

$\hat{OR}_4 = 1.72(0.56 \sim 5.31)$

Crude OR=27x443/95x44=2.86 (1.68~4.85)

年輕組($Age < 55$),
心電圖無異常($ECG = 0$)

	High CAT	Low CAT	Total
CHD	1	17	18
No CHD	7	257	264
Total	8	274	282

$\hat{OR}_1 = 2.16(0.25 \sim 18.58)$

年輕組($Age < 55$),
心電圖有異常($ECG = 1$)

	High CAT	Low CAT	Total
CHD	3	7	10
No CHD	14	52	66
Total	17	59	76

$\hat{OR}_2 = 1.59(0.36 \sim 6.96)$

年老組($Age \geq 55$),
心電圖無異常(ECG) $= 0$

	High CAT	Low CAT	Total
CHD	9	15	24
No CHD	30	107	137
Total	39	122	161

$\hat{OR}_3 = 2.14(0.85 \sim 5.37)$

年老組($Age \geq 55$),
心電圖有異常 ($ECG = 1$)

	High CAT	Low CAT	Total
CHD	14	5	19
No CHD	44	27	71
Total	58	32	90

$\hat{OR}_4 = 1.72(0.56 \sim 5.31)$

(a)合併加權方法 (pooled method) (logit method)

(調整前OR為2.86)

i	W_i	$\hat{OR}_i$	$\log \hat{OR}_i$	$W_i \log(\hat{OR}_i)$
1	0.829	2.16	0.77	0.64
2	1.746	1.59	0.46	0.82
3	4.536	2.14	0.76	3.45
4	3.019	1.72	0.54	1.63

(調整後OR值↓,
因部分相關性被年齡及ECG解釋掉了)

$$W_1 = \frac{1}{\frac{1}{a_1}+\frac{1}{b_1}+\frac{1}{c_1}+\frac{1}{d_1}} = \frac{1}{\frac{1}{1}+\frac{1}{7}+\frac{1}{17}+\frac{1}{257}} = 0.829$$

$$W_2 = 1.746, W_3 = 4.536, W_4 = 3.019$$

$$\log \hat{OR}_w = \frac{6.544}{10.149} = 0.645 \qquad \hat{OR}_w = 1.91$$

$$\hat{Var}\left\{\log(\hat{OR}_w)\right\} = \frac{1}{10.149} = 0.099$$

$$95\%CI = \exp(0.645 \pm 1.96 \times \sqrt{0.099}) = (1.03, 3.53)$$

年 輕組($Age < 55$),
心電圖無異常($ECG = 0$)

	High CAT	Low CAT	Total
CHD	1	17	18
No CHD	7	257	264
Total	8	274	282

$R\hat{R}_1 = 2.015(0.30 \sim 13.34)$

年輕組($Age < 55$),
心電圖有異常($ECG = 1$)

	High CAT	Low CAT	Total
CHD	3	7	10
No CHD	14	52	66
Total	17	59	76

$R\hat{R}_2 = 1.49(0.43 \sim 5.14)$

年老組($Age \geq 55$),
心電圖無異常(ECG) $= 0$

	High CAT	Low CAT	Total
CHD	9	15	24
No CHD	30	107	137
Total	39	122	161

$R\hat{R}_3 = 1.88(0.89 \sim 3.95)$

年老組($Age \geq 55$),
心電圖有異常 ($ECG = 1$)

	High CAT	Low CAT	Total
CHD	14	5	19
No CHD	44	27	71
Total	58	32	90

$R\hat{R}_4 = 1.54(0.61 \sim 3.90)$

(調整前RR為2.46)

(b)Mantel-Haenzel方法

i	a_i	b_i	c_i	d_i	n_i	$\left(\frac{1}{n_i}\right)a_i d_i$	$\left(\frac{1}{n_i}\right)b_i c_i$
1	1	17	7	257	282	0.911	0.422
2	3	7	14	52	76	2.052	1.289
3	9	15	30	107	161	5.981	2.795
4	14	5	44	27	90	4.200	2.444
						13.145	6.951

$$O\hat{R}_{MH} = \frac{\frac{1\times 257}{282} + \frac{3\times 52}{76} + \frac{9\times 107}{161} + \frac{14\times 27}{90}}{\frac{17\times 7}{282} + \frac{7\times 14}{76} + \frac{9\times 107}{161} + \frac{14\times 27}{90}}$$

$$= \frac{13.145}{6.950}$$

$$= 1.891$$

$$95\%CI = 1.891\times,\div \exp\left(1.96\times\sqrt{0.10014}\right) = (1.017, 3.517)$$

統計分析之概念

	結果變項(Y)性質	樣本統計值平均值	非迴歸分析方法	迴歸分析方法
二組 (X=1/0)	(1)連續變項	$\bar{Y}$(平均值)	獨立 t-test $(\bar{Y}_1 - \bar{Y}_0)$	線性迴歸 Y=a+bx
	(2)二元變項 (Y=1/0)	$\hat{p}$(比例)=d/n d: 事件 n: 總數	X^2 test(2×2 列聯表) 及比例檢定 $\widehat{p_1} - \widehat{p_0}$	羅吉斯迴歸 Logit(P(Y=1|X))=a+bx
	(3)存活時間 (Y=time to event)	存活函數 S(t) 風險函數 h(t)	Hazard Rate $=h_1(t)/h_0(t)$	Cox 迴歸模式(Cox proportional hazard regression model) $h(t)=h_0(t)\exp(a+bx)$
多組 (k=3) X=1/2/3	(1)連續變項	$\bar{Y}_1 - \bar{Y}_2)$/ $\bar{Y}_2 - \bar{Y}_3)$/ $\bar{Y}_1 - \bar{Y}_3)$	One-way Analysis of Variance(ANOVA)	$Y=a+b_1X_1+b_2X_2$ X=1 X=2 X=3 X_1 0 1 0 X_2 0 0 1

迴歸模式之想法

$$Y=a+bx+e \quad e\sim N(0,\sigma^2)$$

$$X=1 \quad \bar{y}_1 = a + b$$

$$X=0 \quad \bar{y}_0 = a$$

$\bar{y}_1 - \bar{y}_0 = b$ (迴歸係數也就是斜率), b 越大, 則兩組平均值差異越大

生物統計課程連結

獨立樣本二元變項(Binary variable)

(1) 兩組樣本比例檢定 (Two-sample proportion test)

$$H_0 : P_1 = P_0 \qquad H_1 : P_1 \neq P_0$$

$$Z = \frac{\left(\hat{P}_1 - \hat{P}_0\right) - 0}{\sqrt{Var\left(\hat{P}_1 - \hat{P}_0\right)}}$$

當 $Z > Z_{1-\alpha/2}$ 或 $Z < Z_{\alpha/2}$，推翻虛無假說

$Z_{\alpha/2} \leq Z \leq Z_{1-\alpha/2}$ ，無法推翻虛無假說

$$Var\left(\hat{P}_1 - \hat{P}_0\right) = \sqrt{\frac{p'q'}{n_1} + \frac{p'q'}{n_0}}$$

$$p' = \frac{n_1 \hat{p}_1 + n_0 \hat{p}_0}{n_1 + n_0}$$

(p_1與p_0加權平均)

Recall

(2) 卡方檢定(Chi-square (χ^2) test)

舉例:某臨床研究欲證明某種新藥是否比舊藥，其造成感染傷口癒合狀況較好。研究者利用隨機分派試驗得到下列結果：

E: Expectation under H_0 is tr

	有療效	沒有療效	總和
新藥	40	20	60
舊藥	16	48	64
總和	56	68	124

$$E_{11} = 60 \times \frac{56}{124}$$

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(40-(56\times 60)/124)^2}{56\times 60/124} + \frac{(20-(60\times 68)/124)^2}{60\times 68/124}$$

$$+ \frac{(16-(56\times 64)/124)^2}{56\times 64/124} + \frac{(48-(64\times 68)/124)^2}{64\times 68/124}$$

$$= 21.71 \quad (p < 0.0001)$$

H_0 :是否有較好的治療效果和使用新藥與舊藥無關

$$\sim \chi^2_{(1)} \qquad \chi^2_{(1)} \equiv Z^2 = \left(\frac{O-E}{\sqrt{V}}\right)^2$$

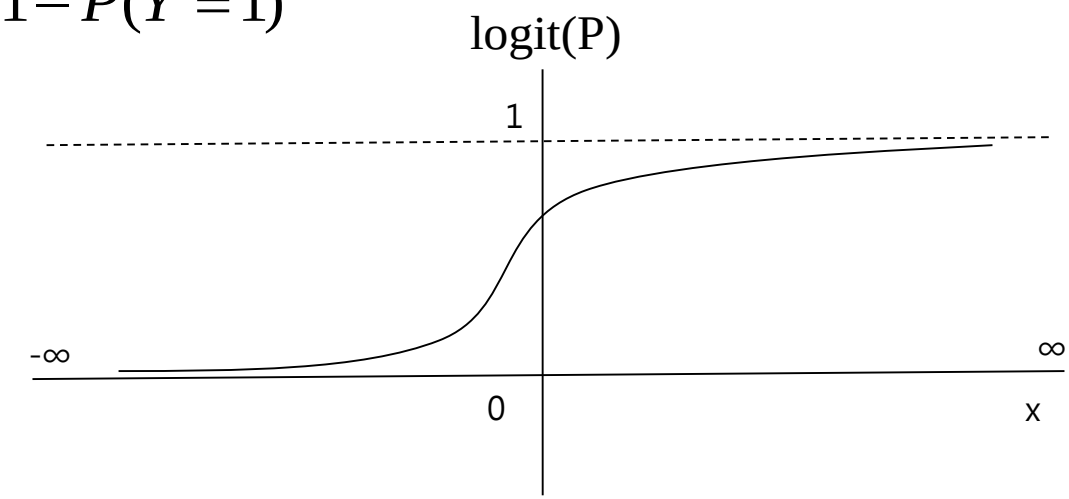
Table A–5. Percentage points or critical values for the χ^2 distribution corresponding to commonly used areas under the curve.

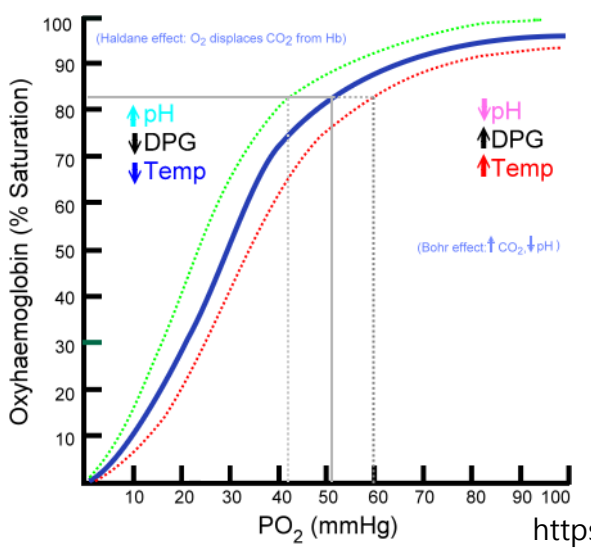
Degrees of Freedom	Area in Upper Tail			
	0.10	0.05	0.01	0.001
1	2.706	3.841	6.635	10.828
2	4.605	5.991	9.210	13.816
3	6.251	7.815	11.345	16.266
4	7.779	9.488	13.277	18.467
5	9.236	11.071	15.086	20.515
6	10.645	12.592	16.812	22.458
7	12.017	14.067	18.475	24.322
8	13.362	15.507	20.090	26.125
9	14.684	16.919	21.666	27.877
10	15.987	18.307	23.209	29.588
11	17.275	19.675	24.725	31.264
12	18.549	21.026	26.217	32.909
13	19.812	22.362	27.688	34.528
14	21.064	23.685	29.141	36.123
15	22.307	24.996	30.578	37.697
16	23.542	26.296	32.000	39.252
17	24.769	27.587	33.409	40.790
18	25.989	28.869	34.805	42.312
19	27.204	30.144	36.191	43.820
20	28.412	31.410	37.566	45.315

Recall

(3)羅吉斯迴歸(Logistic Regression Model)

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right)=\text{logit}(\text{P}(\text{Y}=1))=\text{logit}(\text{P})=\beta_0+\beta_1 X+\varepsilon$$





Sigmoidal shape [edit]

The curve is usually best described by a sigmoid plot, using by a formula of the kind:

$$S(t)=\frac{1}{1+e^{-t}}.$$

A hemoglobin molecule can bind up to four oxygen molecules in a reversible way.

https://en.wikipedia.org/wiki/Oxygen%E2%80%93hemoglobin_dissociation_curve

Recall

(3)羅吉斯迴歸(Logistic Regression Model)

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right)=\text{logit}(\text{P}(\text{Y}=1))=\text{logit}(\text{P})=\beta_0+\beta_1 X+\varepsilon$$

若為兩組比較, 則X代表了組別:

$$\begin{cases} X=1 & log\dfrac{\hat{P}_1}{1-\hat{P}_1}=\hat{\beta}_0+\hat{\beta}_1 \quad (1) \\ X=0 & log\dfrac{\hat{P}_0}{1-\hat{P}_0}=\hat{\beta}_0 \quad (2) \end{cases}$$

(1)-(2) $\log(O\hat{R})=\log\left(\dfrac{\hat{P}_1/1-\hat{P}_1}{\hat{P}_0/1-\hat{P}_0}\right)=\hat{\beta}_1$

勝算比 $\hat{OR}=\exp(\hat{\beta}_1)$

Recall

(3)羅吉斯迴歸(Logistic Regression Model)

比較兩組有無差異 $(H_0: P_1 = P_0)$ 等同於檢定 $H_0: \mathrm{OR} = 1$ 或 $H_0: \beta_1 = 0$

例: 上述新治療方法對於healing及infection之影響

	有療效	沒有療效	總和
新藥	40	20	60
舊藥	16	48	64
總和	56	68	124

$$H_0: \beta_1 = 0 \ \text{ vs. } \mathrm{H}_1: \beta_1 \neq 0$$

羅吉斯迴歸 ➔ $\hat{\beta}_1 = 1.7918$, $\hat{\mathrm{OR}}=\exp(1.7918)=6$

檢定統計值 $\chi^2 = 20.3 > 3.84$, $P < 0.0001$

Mathematical Bridge at Queen’s College, Cambridge

資料分析方法控制干擾因子
2. 羅吉斯迴歸模型, logistic regression model)

探討是否與CHD()有相關

羅吉斯迴歸式:

Y=1: 發生心臟疾病(CHD) ,

Y=0 : 不發生心臟疾病(no CHD)

X: catecholamine (CAT) 之高低

若為世代研究:
X: Age (≥55 vs <55);
ECG (不正常 vs 正常)

$$\text{logit}\ \ P(CHD=1\,|\,\mathbf{X}) = \alpha + \beta_1 \times CAT + \beta_2 \times EKG + \beta_3 \times Age$$

$$P(Y=1\,|\,\mathbf{X}) = \frac{\exp(\alpha + \sum \beta_k x_k)}{1+\exp(\alpha + \sum \beta_k x_k)} \Leftrightarrow \text{logit}\ \ P(Y=1\,|\,\mathbf{X}) = \alpha + \sum_{k=1}^{K} \beta_k x_k$$

羅吉斯迴歸模型 (logistic regression model)

模型1: $Log(P/(1-P))=\alpha+\beta_1\times(CAT=High)$
模型2: $Log(P/(1-P))=\alpha+\beta_1\times(CAT=High)+\beta_2\times(ECG=Abnormal)$

利用羅吉斯迴歸得到對CHD風險之**迴歸係數(標準差)**估計結果列表如下

變數	分層	粗率 (模型 1)
CAT	High vs Low	1.05 (0.2693)

羅吉斯迴歸模型 (logistic regression model)

模型1: $Log(P/(1-P))=\alpha+\beta_1\times(CAT=High)$
模型2: $Log(P/(1-P))=\alpha+\beta_1\times(CAT=High)+\beta_2\times(ECG=Abnormal)$
模型3: $Log(P/(1-P))=\alpha+\beta_1\times(CAT=High)+\beta_2\times(Age\geq 55)$

利用羅吉斯迴歸得到對CHD風險之**迴歸係數(標準差)**估計結果列表如下

變數	分層	粗率 (模型 1)	調整 心電圖 (模型 2)
CAT	High vs Low	1.05 (0.2693)	0.90 (0.2971)
ECG	Abnormal vs Normal		0.36 (0.2904)

羅吉斯迴歸模型 (logistic regression model)

模型1: $Log(P/(1-P))=\alpha+\beta_1\times(CAT=High)$
模型2: $Log(P/(1-P))=\alpha+\beta_1\times(CAT=High)+\beta_2\times(ECG=Abnormal)$
模型3: $Log(P/(1-P))=\alpha+\beta_1\times(CAT=High)+\beta_2\times(Age\geq 55)$
模型4: $Log(P/(1-P))=\alpha+\beta_1\times(CAT=High)+\beta_2\times(ECG=Abnormal)+\beta_3\times(Age\geq 55)$

利用羅吉斯迴歸得到對CHD風險之**迴歸係數(標準差)**估計結果列表如下

變數	分層	粗率 (模型 1)	調整 心電圖 (模型 2)	調整年齡 (模型 3)
CAT	High vs Low	1.05 (0.2693)	0.90 (0.2971)	0.77 (0.2955)
ECG	Abnormal vs Normal		0.36 (0.2904)	
Age	≥55 vs <55			0.62 (0.2840)

羅吉斯迴歸模型 (logistic regression model)

模型1: $Log(P/(1-P))=\alpha+\beta_1\times(CAT=High)$
模型2: $Log(P/(1-P))=\alpha+\beta_1\times(CAT=High)+\beta_2\times(ECG=Abnormal)$
模型3: $Log(P/(1-P))=\alpha+\beta_1\times(CAT=High)+\beta_2\times(Age\geq 55)$
模型4: $Log(P/(1-P))=\alpha+\beta_1\times(CAT=High)+\beta_2\times(ECG=Abnormal)+\beta_3\times(Age\geq 55)$

利用羅吉斯迴歸得到對CHD風險之**迴歸係數(標準差)**估計結果列表如下

變數	分層	粗率 (模型 1)	調整 心電圖 (模型 2)	調整年齡 (模型 3)	調整年齡 及 心電圖 (模型 4)
CAT	High vs Low	1.05 (0.2693)	0.90 (0.2971)	0.77 (0.2955)	0. 62 (0.3193)
ECG	Abnormal vs Normal		0.36 (0.2904)		0.36 (0.2904)
Age	≥55 vs <55			0.62 (0.2840)	0.62 (0.2838)

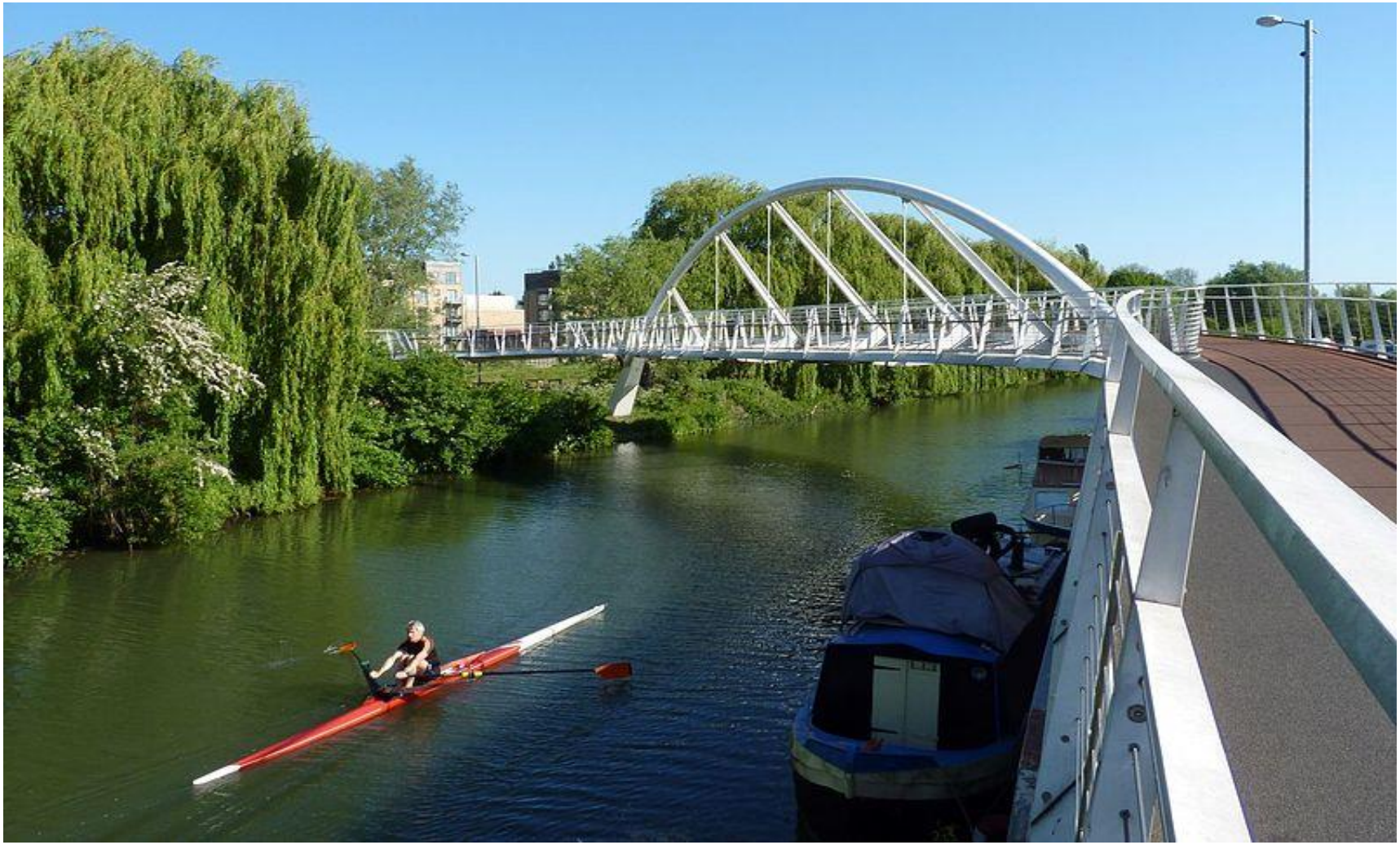
羅吉斯迴歸模型 (logistic regression model)

利用上表中估計得到的迴歸係數可以計算勝算比 (odds ratio, OR)如下:

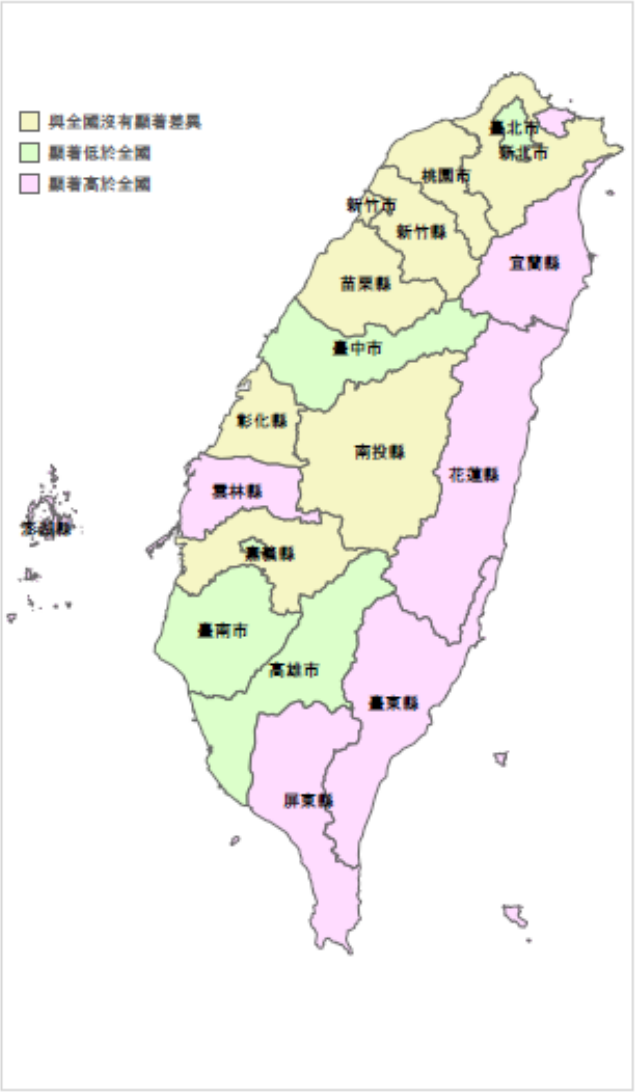
變數	分層	粗率 (模型 1)	調整心電圖 (模型 2)	調整年齡 (模型 3)	調整年齡及心電圖 (模型 4)
CAT	High vs Low	2.86 (1.689, 4.85)	2.45 (1.37, 4.38)	2.17 (1.22, 3.87)	1.86 (1.00, 3.48)
ECG	Abnormal vs Normal	---	1.44 (0.82, 2.54)	---	1.44 (0.81, 2.54)
Age	≥55 vs <55	---	---	1.85 (1.06, 3.24)	1.85 (1.06, 3.23)

調整後OR (aOR)= 1.91 (分層分析中的合併加權方法)
1.89 (Mantel-Haenszel方法)
1.86 (羅吉斯迴歸方法)

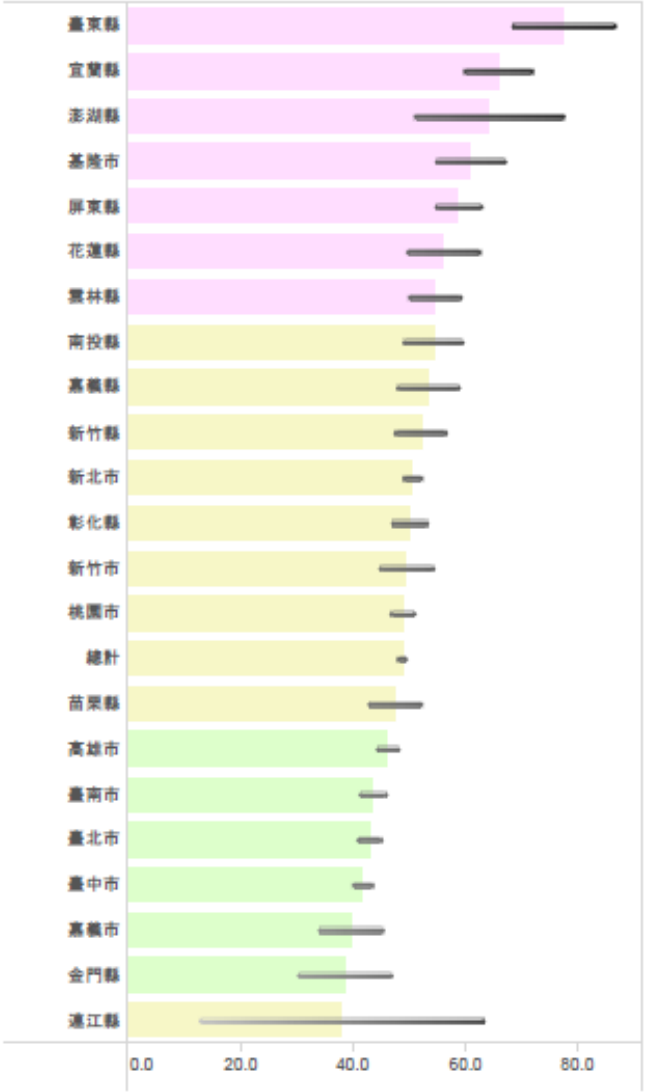
Cambridge Riverside Bridge



102年-104年平均
標準化死亡率地圖



標準化死亡率(單位:每十萬人口)



死因別
- 全死因
- 惡性腫瘤
- 心臟疾病
- 腦血管疾病
- 肺炎
- 糖尿病
- 事故傷害
- 慢性下呼吸道疾病
- 高血壓性疾病
- 腎病變
- 慢性肝病及肝硬化
- 自殺

http://iiqsw.mohw.gov.tw/dataviscategory.aspx?dtype=1

干擾因子與年齡標準化

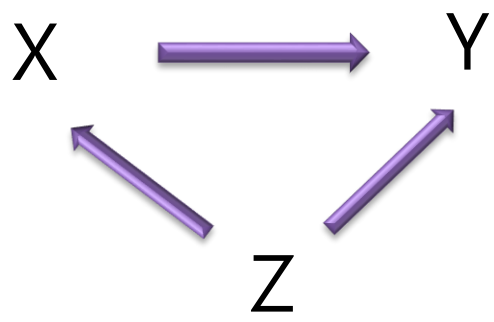
兩區域之心血管疾病(cardiovascular disease)發生風險是否不同?

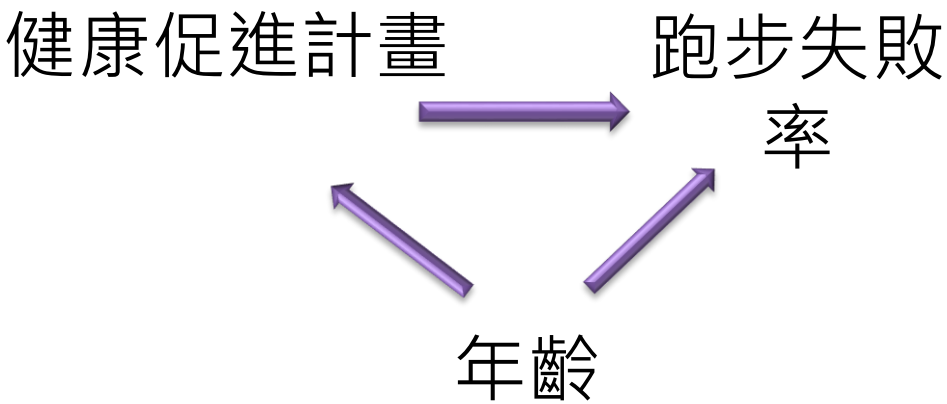
標準人口		區域 1			區域2			
年齡層	人口數	事件數	人口數	發生率	事件數	人口數	發生率	發生率之比
40-49	126,300	18	30,000	0.0006	45	90,000	0.0005	1.20
50-59	99,200	108	60,000	0.0018	300	150,000	0.002	0.90
60-69	66,800	1,050	210,000	0.005	360	60,000	0.006	0.83
總和	292,300	1,176	300,000	0.00392	705	300,000	0.00235	**1.67**

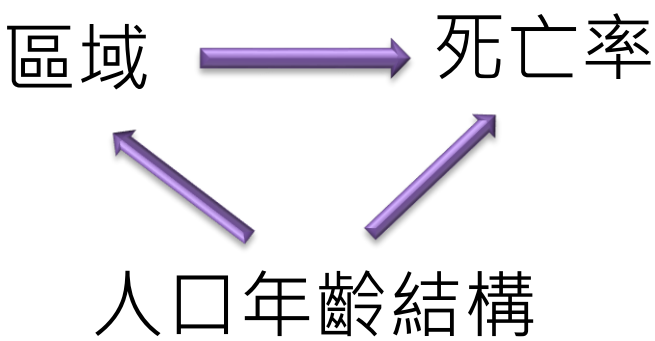
粗率　　：區域1 > 區域2

年齡別發生率: 區域1 ≈ 區域2

年齡標準化與辛普森矛盾







以年齡標準化方法調整干擾因子

年齡組	標準族群	研究族群		
	人口數	人口數	事件數	事件比例
1	N_1	n_1	r_1	p_1
⋮	⋮	⋮	⋮	⋮
i	N_i	n_i	r_i	p_i
⋮	⋮	⋮	⋮	⋮
k	N_k	n_k	r_k	p_k

$$p' = \frac{\sum N_i p_i}{\sum N_i}$$

粗發生率之比

標準人口		區域 1			區域2			
年齡層	人口數	事件數	人口數	發生率	事件數	人口數	發生率	發生率之比
40-49	126,300	18	30,000	0.0006	45	90,000	0.0005	1.20
50-59	99,200	108	60,000	0.0018	300	150,000	0.002	0.90
60-69	66,800	1,050	210,000	0.005	360	60,000	0.006	0.83
總和	292,300	1,176	300,000	0.00392	705	300,000	0.00235	**1.67**

粗發生率之比$(crude\ risk\ ratio)$:

$$\frac{1176/300000}{705/300000}=\frac{0.00392}{0.00235}=1.67$$

直接標準化方法比較兩區域之心血管疾病發生的風險

標準人口		區域 1			區域2			
年齡層	人口數	事件數	人口數	發生率	事件數	人口數	發生率	發生率之比
40-49	126,300	18	30,000	0.0006	45	90,000	0.0005	1.20
50-59	99,200	108	60,000	0.0018	300	150,000	0.002	0.90
60-69	66,800	1,050	210,000	0.005	360	60,000	0.006	0.83
總和	292,300	1,176	300,000	0.00392	705	300,000	0.00235	1.67

$$\text{區域1}: \frac{126300 \times 0.0006 + 99200 \times 0.0018 + 66800 \times 0.005}{292300} = 0.00201$$

$$\text{區域 2}: \frac{126300 \times 0.0005 + 99200 \times 0.002 + 66800 \times 0.006}{292300} = 0.00227$$

$$\text{年齡標準化發生率之比}: \frac{0.00201}{0.00227} = 0.89$$