統計方法與應用

中國醫藥大學
生物統計研究所 生統中心
梁文敏 (ext.: 6107)

主題一 資料整理的原則

1. 簡單(不須全放在1張表單)

   →分幾個表單(基本資料1張、定期追蹤紀錄1張，放單一id以便連結…)
2. 用代碼取代文字(sex: Male, Female, M,..)

   →(sex: 1---male, 0---female)
3. 一個變數一個概念(death: 意外事故死亡、心血管疾病死亡、…)

   →分兩個變數(death, d_cause)來表示
4. 將資料分為兩大類型:

   連續型(可以量化者，可計算平均值等指標)、

   類別型(代表特質，不適合計算; 二元、三元、…)

id	sex	hosp_stay	death	death_date	begin_date	BA1_1m	BA1_2m	BA1_3m	…
1	Male	200	死亡	05/22/2010	06/06/2007	**40**	**45**	**50**	…
2	Female	15			08/16/2008	65	70	75	…
3	Fenale	14	意外事故死亡	11/06/2012	11/06/2008	65	60	60	…
…	…	…	…	…	…	…	…	…	…
50	M	13	心血管疾病死亡	11/06/1998	04/22/2010	70	75	80	…

註：hosp_stay:住院天數、

BA1_1m:巴氏量表第一題第 1 個月測值、BA1_2m:巴氏量表第一題第 2 個月測值、

id	sex	hosp_stay	death	d_cause	death_date	begin_date
1	1	200	1	…	05/22/2010	06/06/2007
2	0	15	0	…	…	08/16/2008
3	0	14	1	意外事故	11/06/2012	11/06/2008
…	…	…	…	…	…	03/09/2009
50	1	13	1	心血管疾病	11/06/1998	04/22/2010

註：hosp_stay :住院天數、sex (0:女，1:男) 、death (0:存活，1:死亡) 。

id	time	BA1	BA2	…	BA10
1	1	40	55	…	75
1	2	45	60	…	60
1	3	50	80	…	90
…	…	…		…	
50	12	60	80	…	50

註：time：1---第 2 個月, 2------第 2 個月,… ；。

BA1: 巴氏量表第一題; BA2: 巴氏量表第二題,….

主題二 多花一點時間在資料探索階段、再依序深入

- 理解各變數的角色：清楚知道哪些爲目標變數、哪些爲解釋變數
- 理解各變數的類型：判斷某些連續型變數是否需要分組

  e.g., 年齡以連續型(單位:歲)或

  類別型(1:<65, 2:>=65)表示
- 單一變數的描述性統計(個數(%)、平均值、標準差、中位數、Q1-Q3、IQR、…)

  單一變數→兩兩關係(crude analysis) (過程中清楚知道各變數的角色)
- 調整多個變數的分析(multivariate analysis) (過程中清楚知道各變數的角色)

主題三 統計方法的選取

(1)目標變數的類型

*一般而言此爲最主要的依據

(2)解釋變數的類型

(3)解釋變數的個數

*針對一群中風族群進行研究觀察5年

統計方法的選取：主要由目標變數的類型決定

針對一群中風族群進行研究，觀察 5 年

	常見的統計方法	變數角色	Crude analysis
M1	獨立樣本 t 檢定 (two-sample t test) (Student's t test)	目標變數(Y)	睡眠小時(連續變項)
		解釋變數(X)	性別(二元類別變項)
M2	線性迴歸分析 (linear regression model)	目標變數(Y)	睡眠小時(連續變項)
		解釋變數(X)	性別(二元類別變項)
M3	卡方檢定 (Chi-square test) (chi-square test)	目標變數(Y)	睡眠困擾(二元類別變項)
		解釋變數(X)	性別(二元類別變項)
M4	邏輯斯迴歸分析 (logistic regression model)	目標變數(Y)	睡眠困擾(二元類別資料)
		解釋變數(X)	性別(二元類別)
M5	對數等級檢定 (Log-rank test)	目標變數(Y)	再度中風的時間 (起始點至事件發生時間+設限)
		解釋變數(X)	性別(二元類別)
M6	Cox 迴歸分析 (Cox regression model)	目標變數(Y)	再度中風的時間 (起始點至事件發生時間+設限)
		解釋變數(X)	性別(二元類別)

統計方法的選取：主要由目標變數的類型決定→亦受解釋變數類型影響

針對一群輕微中風的族群進行研究，觀察 5 年

	常見的統計方法	變數角色	Crude analysis
M1	獨立樣本 t 檢定 (two-sample t test) (Student's t test)	目標變數(Y)	睡眠小時(連續變項)
		解釋變數(X)	~~年齡(連續變項)~~→(<65,>=65)
M2	線性迴歸分析 (linear regression model)	目標變數(Y)	睡眠小時(連續變項)
		解釋變數(X)	年齡(連續變項)
M3	卡方檢定 (Chi-square test) (chi-square test)	目標變數(Y)	睡眠困擾(二元類別變項)
		解釋變數(X)	~~年齡(連續變項)~~→(<65,>=65)
M4	邏輯斯迴歸分析 (logistic regression model)	目標變數(Y)	睡眠困擾(二元類別資料)
		解釋變數(X)	年齡(連續變項)
M5	對數等級檢定 (Log-rank test)	目標變數(Y)	再度中風的時間 (起始點至事件發生時間+設限)
		解釋變數(X)	~~年齡(連續變項)~~→(<65,>=65)
M6	Cox 迴歸分析 (Cox regression model)	目標變數(Y)	再度中風的時間 (起始點至事件發生時間+設限)
		解釋變數(X)	年齡(連續變項)

統計方法的選取：主要目標變數類型→解釋變數類型→亦受解釋變數個數影響

針對一群輕微中風的族群進行研究，觀察 5 年

	常見的統計方法	變數角色	Crude analysis	Multivariate analysis
M1	獨立樣本 t 檢定 (two-sample t test) (Student's t test)	目標變數(Y)	睡眠小時(連續變項)	
		解釋變數(X)	性別	~~性別+年齡+壓力+...~~
M2	線性迴歸分析 (linear regression model)	目標變數(Y)	睡眠小時(連續變項)	
		解釋變數(X)	性別	性別+年齡+壓力+...
M3	卡方檢定 (Chi-square test) (chi-square test)	目標變數(Y)	睡眠困擾(二元類別)	
		解釋變數(X)	性別	~~性別+年齡+壓力+...~~
M4	邏輯斯迴歸分析 (logistic regression model)	目標變數(Y)	睡眠困擾(二元類別)	
		解釋變數(X)	性別	性別+年齡+壓力+...
M5	對數等級檢定 (Log-rank test)	目標變數(Y)	再度中風的時間	
		解釋變數(X)	性別	~~性別+年齡+壓力+...~~
M6	Cox 迴歸分析 (Cox regression model)	目標變數(Y)	再度中風的時間	
		解釋變數(X)	性別	性別+年齡+壓力+...

主題四 文章中的第一個表格

- 基本資料描述：單一欄的描述或雙欄的描述最爲常見
- 雙欄的描述：依照人口學特質(男性、女性; <65歲、>=65歲)、醫學資料特質(治療組、對照組; A藥組、B藥組)將資料分欄描述
- 雙欄的描述搭配統計檢定：

  連續型變數---以平均值(SD)表示、並搭配獨立樣本t檢定(M1)最爲常見

  *若資料量小或偏斜、以中位數(Q1-Q3, IQR) 、搭配Wilcoxon sum-rank test

  類別型變數---以個數(%)表示、並搭配卡方檢定(M3)最爲常見

  *若資料量小、卡方檢定不適用時、則以Fisher's exact test取代最爲常見

表 1. 基本資料描述

	一般(n=28) n(%)	肥胖(n=22) n(%)	*P* 值
年齡(歲),M±SD	62.64 ± 14.00	59.55 ± 11.92	0.4504#
住院天數,M±SD	13.04 ± 6.03	37.82 ± 69.50	0.1101#
中位數(Q1-Q3)	12.5(8.0-17.5)	13.0(10.0-17.0)	0.6033◎
睡眠困擾			
無	13(46.43)	3(13.64)	0.0136*
有	15(53.57)	19(86.36)	
五年內再中風			
無	19(67.86)	9(36.36)	0.0266*
有	9(32.14)	14(63.64)	
五年內死亡			
無	26(92.86)	16(72.73)	0.1163#
有	2(7.14)	6(27.27)	

two-sample t test ◎Wilcoxon rank-sun test

*Chi-square test # Fisher's exact test

主題五 文章中的主要分析表格
(常見在表2或表3)

主題五-1 線性迴歸分析

目標變數: 連續變數 (用平均值表示)

解釋變數:連續變數及類別變數皆可

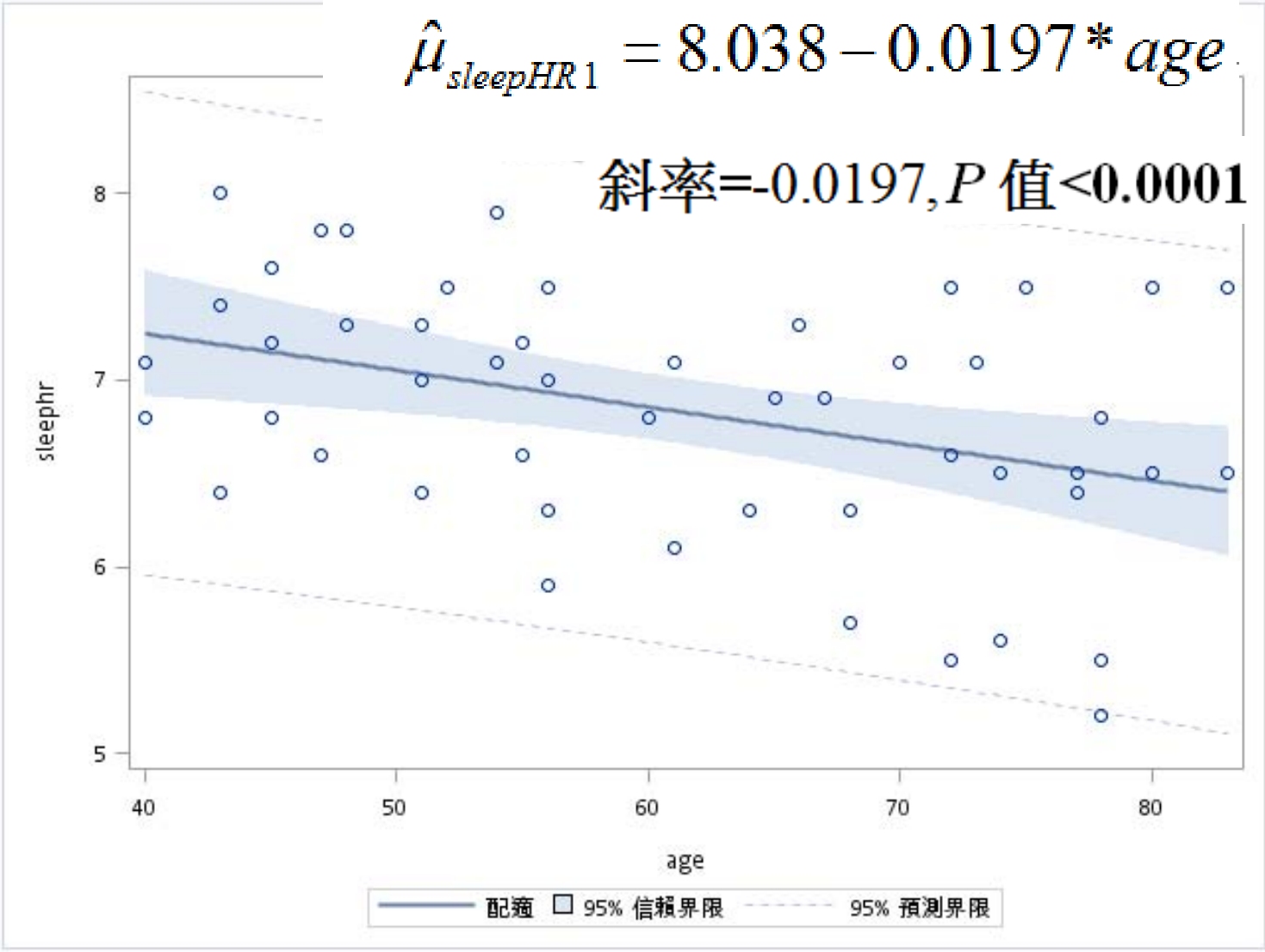
範例:探討睡眠小時的影響因素

參數	估計值	標準誤差	t 值	Pr > |t|
Intercept	8.037662598	0.42393691	18.96	<.0001
age	-0.019686990	0.00678350	-2.90	0.0056

探討睡眠小時與年齡

$$\hat{\mu}_{sleepHR1} = 8.038 - 0.0197 * age$$

斜率=-0.0197, P 值<0.0001



參數	估計值		標準誤差	t 值	Pr > |t|
Intercept	6.331818182	B	0.10590908	59.79	<.0001
obesity 0	0.896753247	B	0.14152696	6.34	<.0001
obesity 1	0.000000000	B			

探討睡眠小時與肥胖

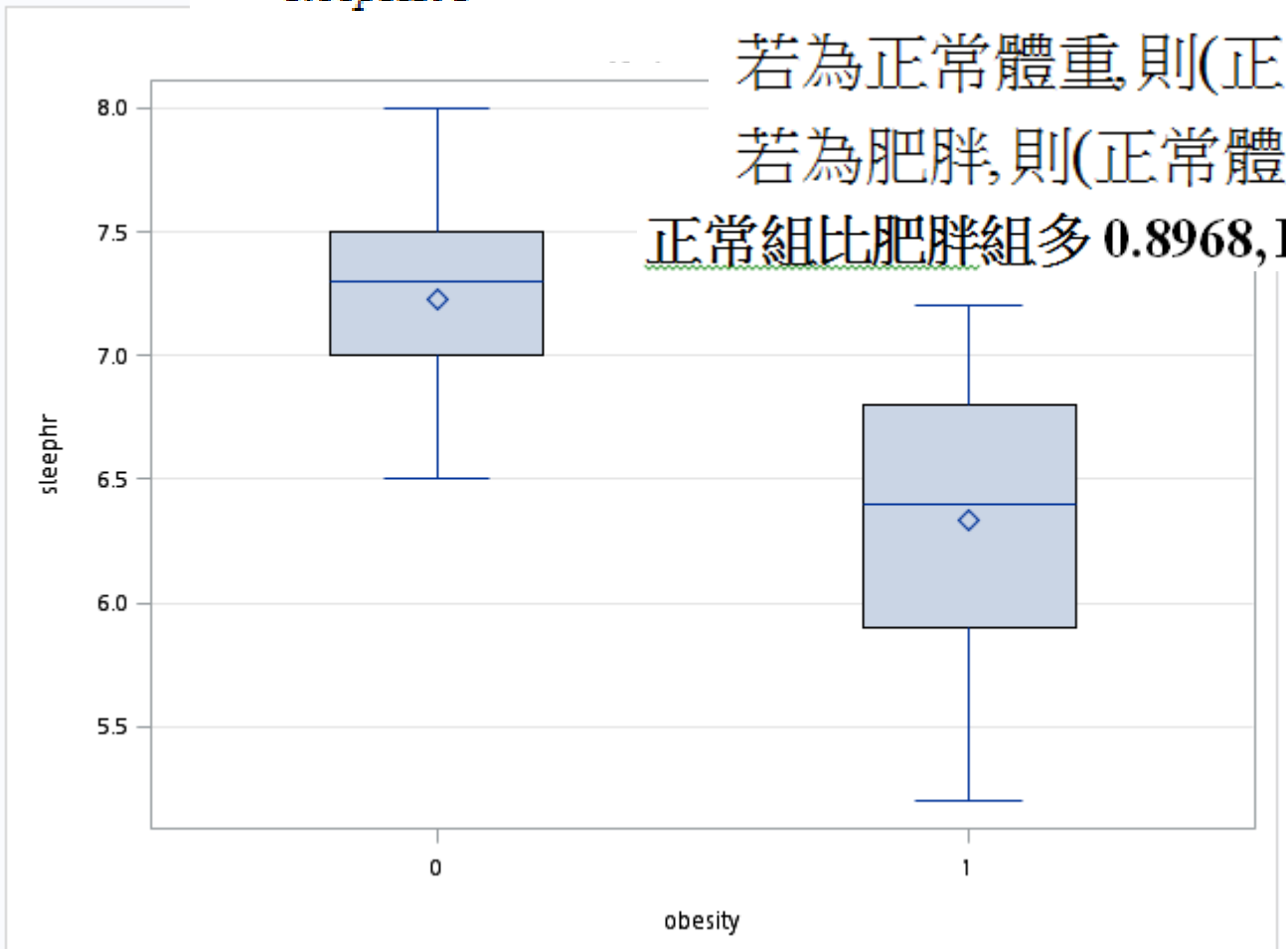
肥胖:分爲正常體重組與肥胖組

$$\hat{\mu}_{sleepHR\,1} = 7.745 + 0.8967(\text{正常體重})$$

若為正常體重,則(正常體重) = 1

若為肥胖,則(正常體重) = 0

正常組比肥胖組多 0.8968, P 值<0.0001



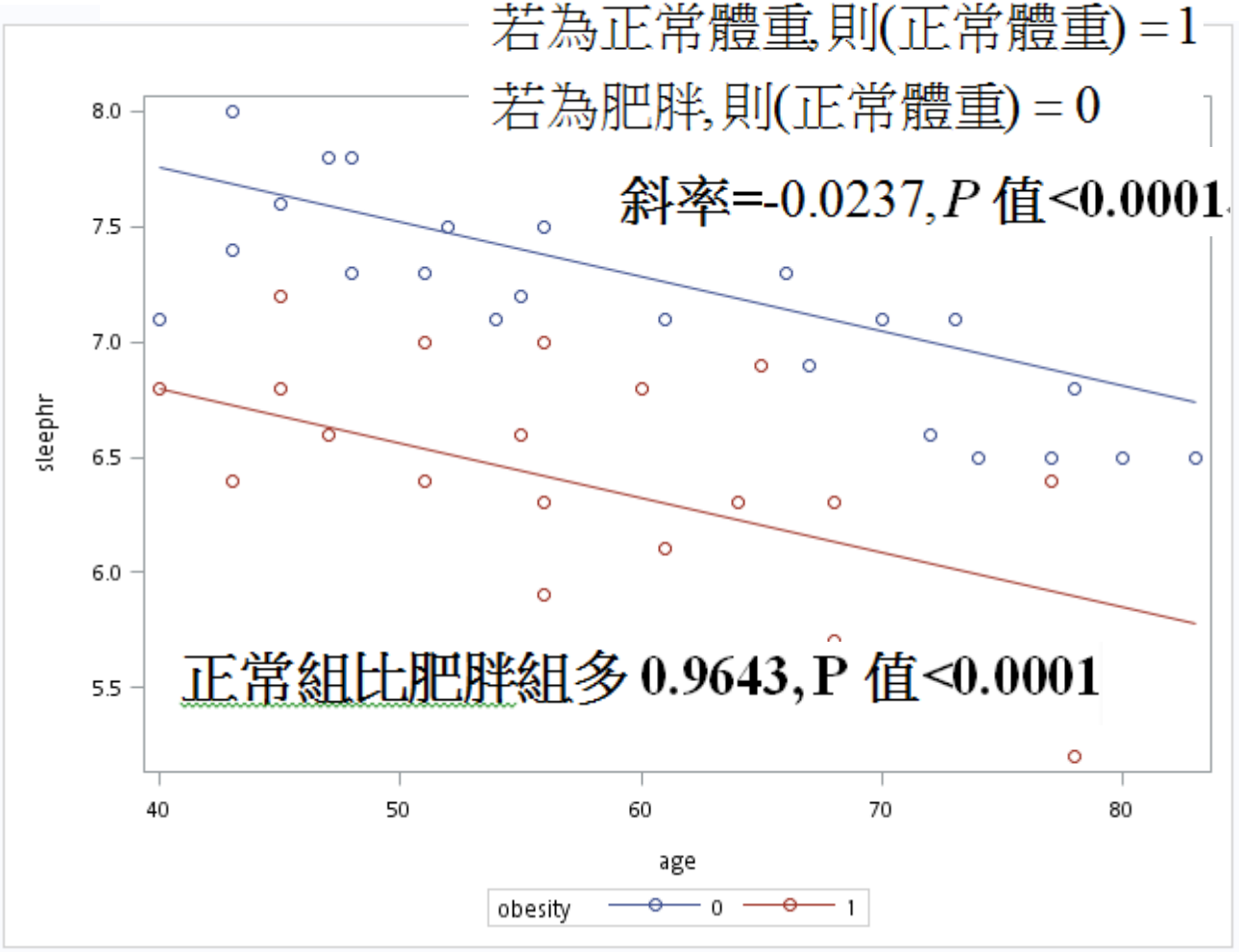
參數	估計值		標準誤差	t 值	Pr > |t|
Intercept	7.744515358	B	0.26912984	28.78	<.0001
age	-0.023724685		0.00429746	-5.52	<.0001
obesity 0	0.964306978	B	0.11206674	8.60	<.0001
obesity 1	0.000000000	B			

探討睡眠小時與年齡及肥胖

$$\hat{\mu}_{sleepHR1} = 7.745 - 0.0237 * age + 0.9643(\text{正常體重})$$

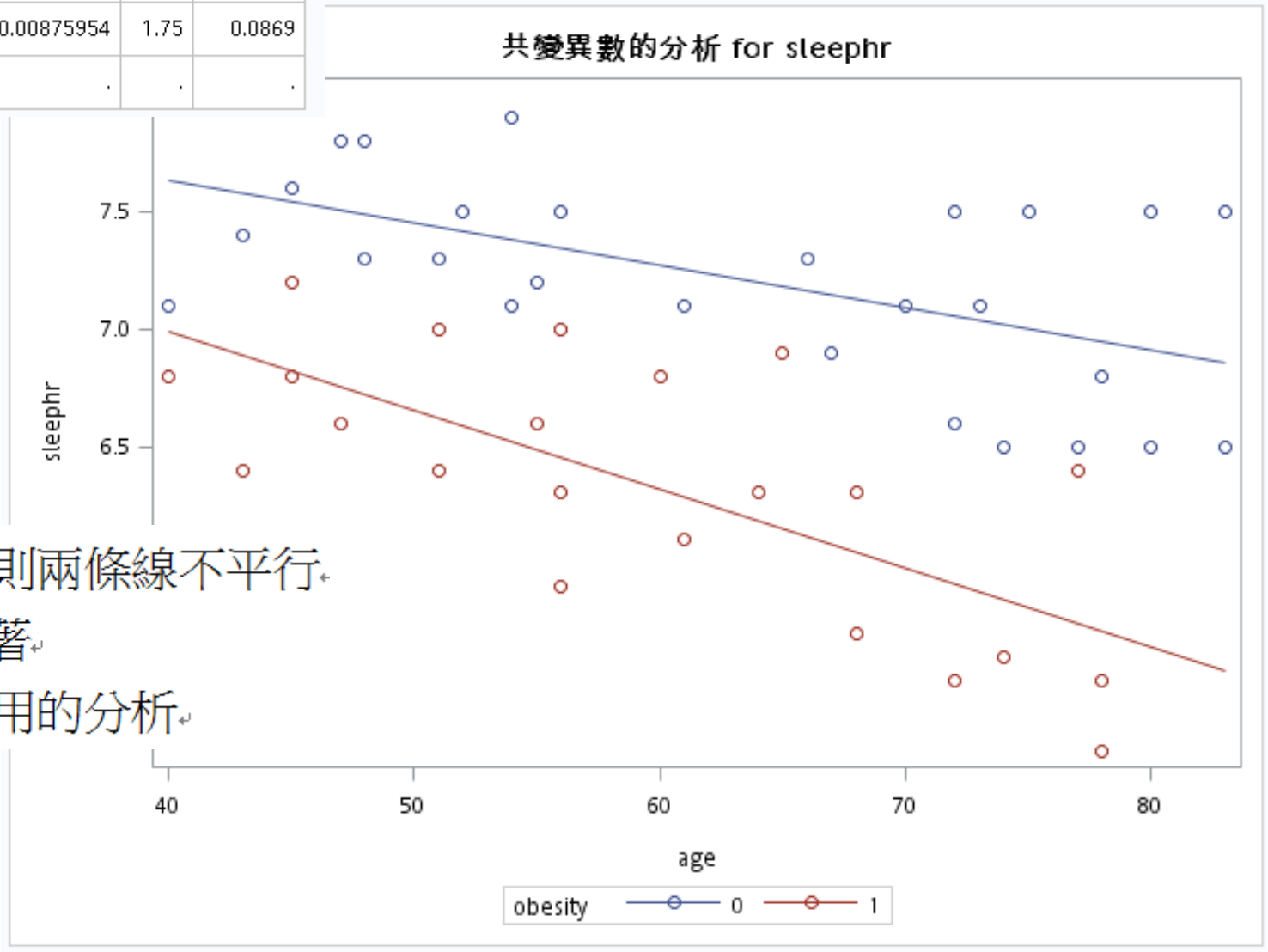
若為正常體重,則(正常體重) = 1

若為肥胖,則(正常體重) = 0



參數	估計值		標準誤差	t 值	Pr > |t|
Intercept	8.327968940	B	0.42496910	19.60	<.0001
age	-0.033523143	B	0.00700411	-4.79	<.0001
obesity 0	0.036005553	B	0.54180529	0.07	0.9473
obesity 1	0.000000000	B	.	.	.
age*obesity 0	0.015325498	B	0.00875954	1.75	0.0869
age*obesity 1	0.000000000	B	.	.	.

$$\hat{\mu}_{sleepHR1}$$
$$= 8.3280 - 0.0335 * age + 0.036(\text{正常體重})$$
$$+ 0.0153 * age * (\text{正常體重})$$



加入交互作用則兩條線不平行

P=0.0869 不顯著

故用無交互作用的分析

表 以線性迴歸模式分析睡眠之影響因素

	單變項分析		多變項分析	
	迴歸係數	*P* 值[$]	迴歸係數	*P* 值[#]
年齡(歲、每增加一歲)	-0.0197	<0.0001	-0.0237	<.0001
體重				
正常體重	0.8968	<0.0001	0.9643	<0.001
肥胖				

[$] *P* 值係根據單變項迴歸分析

[#] *P* 值係根據多變項迴歸分析

主題五-2 邏輯斯迴歸分析

目標變數: 二元類別變數(用勝算表示)

解釋變數:連續變數及類別變數皆可

範例:探討再中風的影響因素

邏輯斯迴歸分析

- 用勝算來表示再中風的風險
- 勝算=再中風率 / (1-再中風率)
  解釋爲再中風率的勝算

範例:探討再中風與肥胖

例如: 肥胖組：再中風率=0.64
再中風率勝算=0.64/0.36=1.75
正常組：再中風率=0.32
再中風率勝算=0.32/0.68=0.47

肥胖組再中風率的勝算/正常組再中風率的勝算
=1.75/0.47=3.69 (勝算比, OR– Odds ratio)
OR > 1,肥胖組再中風風險高,

通常會計算95%CI(信賴區間) , 若95%CI 不包括1,
表示此相關具有統計上顯著的意義

Crude analysis

最大概度估計值的分析						
參數		DF	估計值	標準誤差	Wald 卡方	Pr > ChiSq
Intercept		1	-0.7471	0.4046	3.4086	0.0649
obesity	1	1	1.3067	0.6001	4.7406	0.0295

勝算比估計值			
效果	點估計值	95% Wald 信賴界限	
obesity 1 與 0 之間的關係	3.694	1.139	11.976

$$\log\frac{P}{1-P} = -0.747 + 1.307 * (\text{肥胖})$$

$$OR = \exp(1.307) = 3.694,$$

$$95\%CI = (1.139, 11.976),\ P = 0.0295$$

範例:探討再中風與肥胖

	正常體重組	肥胖組
總人數	28	22
再中風人數	9	14
a=再中風率	0.32	0.64
b=1-再中風率	0.68	0.36
再中風勝算=a/b	0.47	1.75
肥胖組勝算/正常組勝算		3.69

ODDS RATIO= 3.69

Risk Ratio=0.64/0.32= 1.98

Crude analysis

最大概度估計值的分析						
參數		DF	估計值	標準誤差	Wald 卡方	Pr > ChiSq
Intercept		1	-0.2075	0.3734	0.3089	0.5783
age1	1	1	0.1124	0.5747	0.0382	0.8450

勝算比估計值			
效果	點估計值	95% Wald 信賴界限	
age1 1 與 0 之間的關係	1.119	0.363	3.452

$$\log\frac{P}{1-P} = -0.2075 + 0.1124*(\text{年齡} > 65)$$

$$OR = \exp(0.1124) = 1.119,$$

$$95\%CI = (0.363, 3.452), P = 0.8450$$

範例:探討再中風與年齡

	<=65歲組	大於65歲組
總人數	29	21
再中風人數	13	10
a=再中風率	0.45	0.48
b=1-再中風率	0.55	0.52
再中風勝算=a/b	0.81	0.91
肥胖組勝算/正常組勝算		1.12

ODDS RATIO= 1.12

Risk Ratio=0.86/0.29= 1.06

Multivariate analysis

最大概度估計值的分析						
參數		DF	估計值	標準誤差	Wald 卡方	Pr > ChiSq
Intercept		1	-0.9558	0.5272	3.2871	0.0698
obesity	1	1	1.3924	0.6217	5.0159	0.0251
age1	1	1	0.4026	0.6263	0.4133	0.5203

勝算比估計值			
效果	點估計值	95% Wald 信賴界限	
obesity 1 與 0 之間的關係	4.025	1.190	13.612
age1 1 與 0 之間的關係	1.496	0.438	5.104

$$\log\frac{P}{1-P} = -0.9558 + 1.3924*(\text{肥胖}) + 0.4026*(\text{年齡} > 65\text{歲})$$

肥胖*vs*正常：得再中風的勝算比 = 4.025(顯著)

年齡每增1歲：得再中風的勝算比 = 1.495(不顯著)

範例:探討再中風與肥胖及年齡

表 以邏輯斯迴歸模式分析再中風之影響因素

	單變項分析		多變項分析	
	OR(95%CI)	*P* 值	OR(95%CI)	*P* 值
體重				
正常體重	1		1	
肥胖	3.69(1.14-11.98)	0.0295	4.03(1.19-13.61)	0.0251
年齡				
<=65 歲	1		1	
>65 歲	1.12(0.36-3.45)	0.8450	1.50(0.44-5.10)	0.5203

敏感度分析---利用分層分析

>65歲組	正常體重組	肥胖組
總人數	14	7
再中風人數	4	6
a=再中風率	0.29	0.86
b=1-再中風率	0.71	0.14
再中風勝算=a/b	0.40	6.00
肥胖組勝算/正常組勝算		15.00

ODDS RATIO= 15.00

Risk Ratio=0.86/0.29= 3.00

<=65歲組	正常體重組	肥胖組
總人數	14	15
再中風人數	5	8
a=再中風率	0.36	0.53
b=1-再中風率	0.64	0.47
再中風勝算=a/b	0.56	1.14
肥胖組勝算/正常組勝算		2.06

ODDS RATIO= 2.06

Risk Ratio=0.86/0.29= 1.49

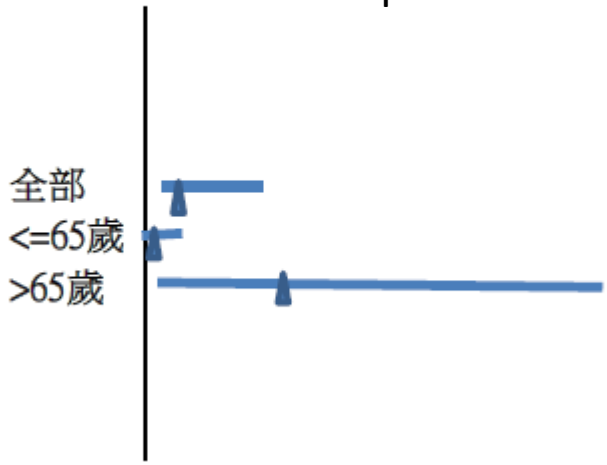
n=21 勝算比估計值

效果	點估計值	95% Wald 信賴界限	
obesity 1 與 0 之間的關係	15.000	1.342	167.638

n=29 勝算比估計值

效果	點估計值	95% Wald 信賴界限	
obesity 1 與 0 之間的關係	2.057	0.463	9.140

Forest plot



敏感度分析---利用分層分析

表 以邏輯斯迴歸模式分析再中風之影響因素

	單變項分析		多變項分析	
	OR(95%CI)	*P* 值	OR(95%CI)	*P* 值
體重				
正常體重	1		1	
肥胖	3.69(1.14-11.98)	0.0295	4.03(1.19-13.61)	0.0251
年齡				
<=65 歲	1		1	
>65 歲	1.12(0.36-3.45)	0.8450	1.50(0.44-5.10)	0.5203

表 以邏輯斯迴歸模式分析再中風之影響因素

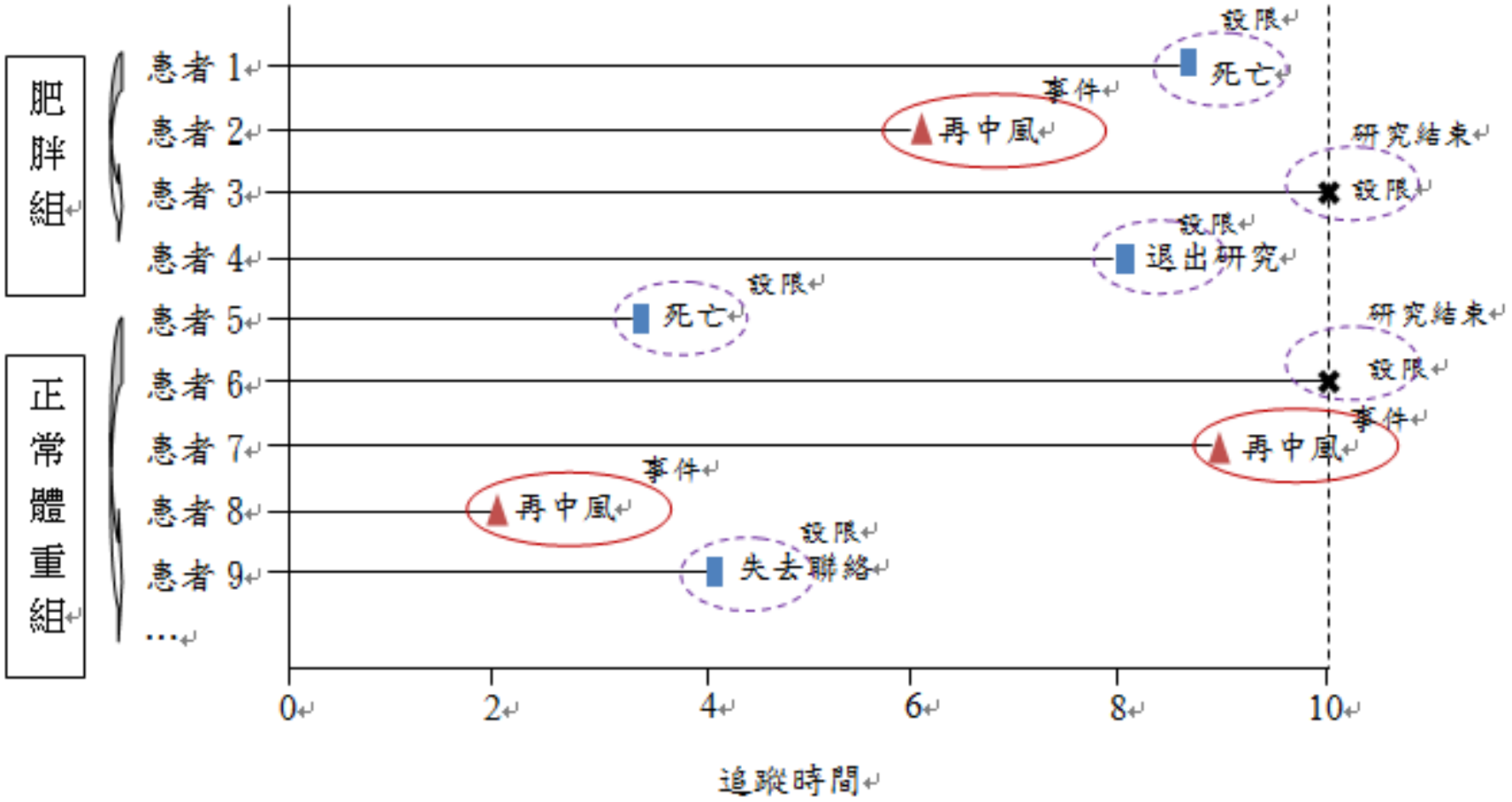
	<=65 歲		>65 歲	
	OR(95%CI)	*P* 值	OR(95%CI)	*P* 值
體重				
正常體重	1		1	
肥胖	2.06(0.46-9.14)	0.3432	15.00(1.34-167.64)	0.0279

肥胖與再中風的關係,在不同年齡層的結果差異很大,顯示Multivariate analysis 的結果並不穩健,應採分層分析或....

主題五-3 COX迴歸分析

目標變數:
從起始點到事件發生的時間+設 限狀態
**用風險(hazard)表示事件生危險性
解釋變數:連續變數及類別變數皆可

範例:探討再中風發生風險的影響因素



- 目標變數: 起始點到事件發生的時間+設限狀態 (用風險(hazard)表示)
- 在每一個時間點計算再中風的風險(hazard)、是一種條件機率、每時段從新計算當時處於危險中的人數(no. at risk)
- 風險~ (某時段再中風人數/某時段處於危險中的人數)

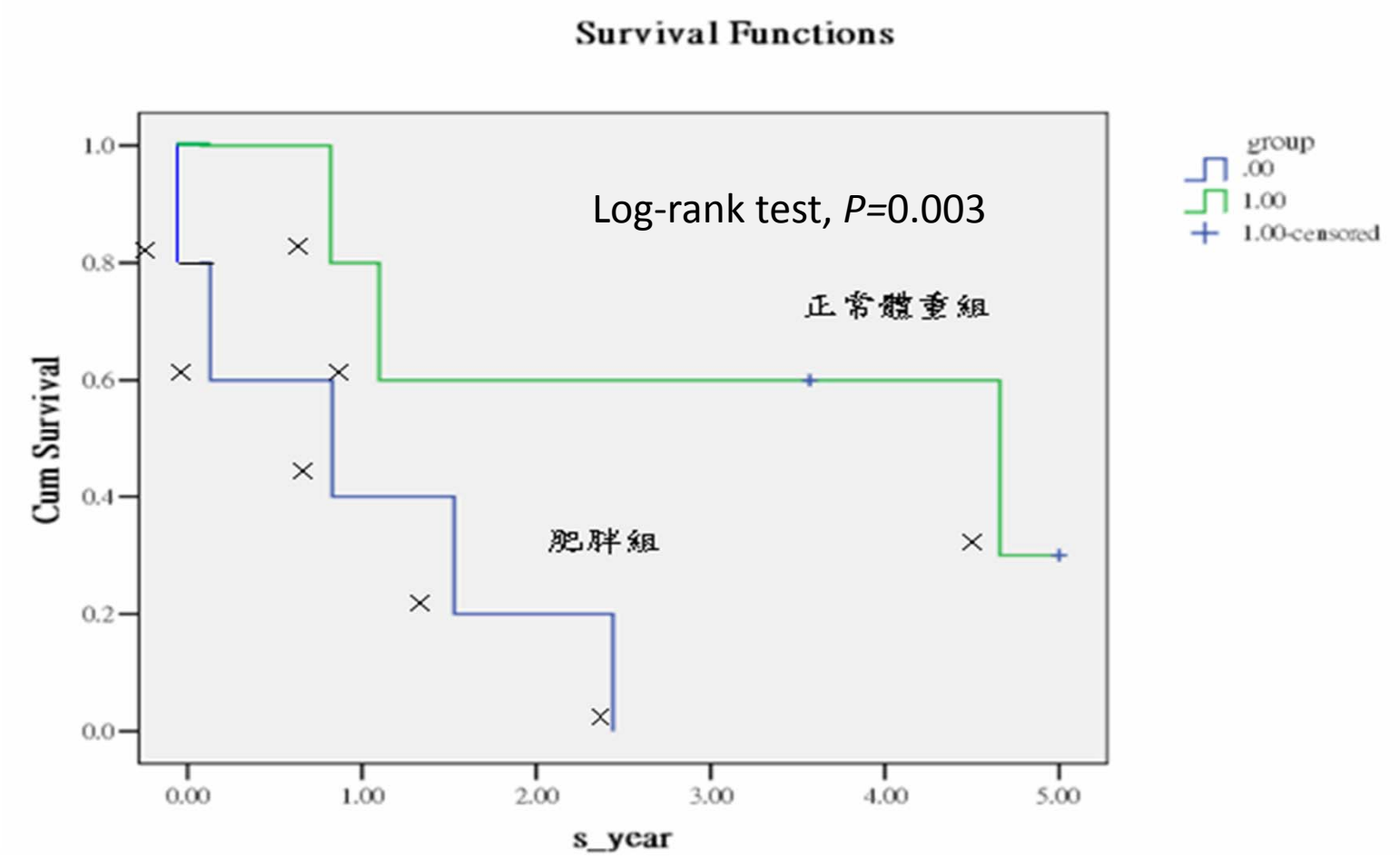
表 以 COX 迴歸模式分析再中風之影響因素

	單變項分析		多變項分析	
	HR(95%CI)	*P* 值[$]	HR(95%CI)	*P* 值[#]
年齡(歲、每增加一歲)	1.2(1.1-1.3)	<0.0001	1.3(1.2-1.4)	<.0001
體重				
正常體重	1		1	
肥胖	2.3(2.1-2.6)	<0.0001	2.1(1.9-2.4)	<0.001

[$] *P* 值係根據單變項迴歸分析

[#] *P* 值係根據多變項迴歸分析

- 在每一個時間點(或時段)計算再中風的風險(hazard)、是一種條件機率、每時段從新計算當時處於危險中的人數(no. at risk)←每次計算時、將設限或事件發生者從no. at risk 扣除
- 每一個點(或時段)的存活率可由hazard計算得之



- 利用Kaplan-Meier方法繪製存活曲線圖
- 利用Log-rank test 進行檢定兩條存活曲線是否相等

主題六 進階延伸的應用

1.考慮競爭風險的存活分析
基本概念

- 當目標爲探討再中風的風險
- 若這群人的死亡率高、則某些人可能因爲先死亡而無法有再中風的發生
- 我們稱死亡爲再中風的競爭風險
- 一般存活分析、若病患在觀察時間內死亡、則將其設爲設限狀況、計算風險時會將死亡個案排除於分母中、造成風險的高估
- 要考慮競爭風險的論點：既然個案已死亡就不可能發生再中風、故不該將死亡個案排除於分母中、造成風險的高估

***風險~ (某時段再中風人數/某時段處於危險中的人數)

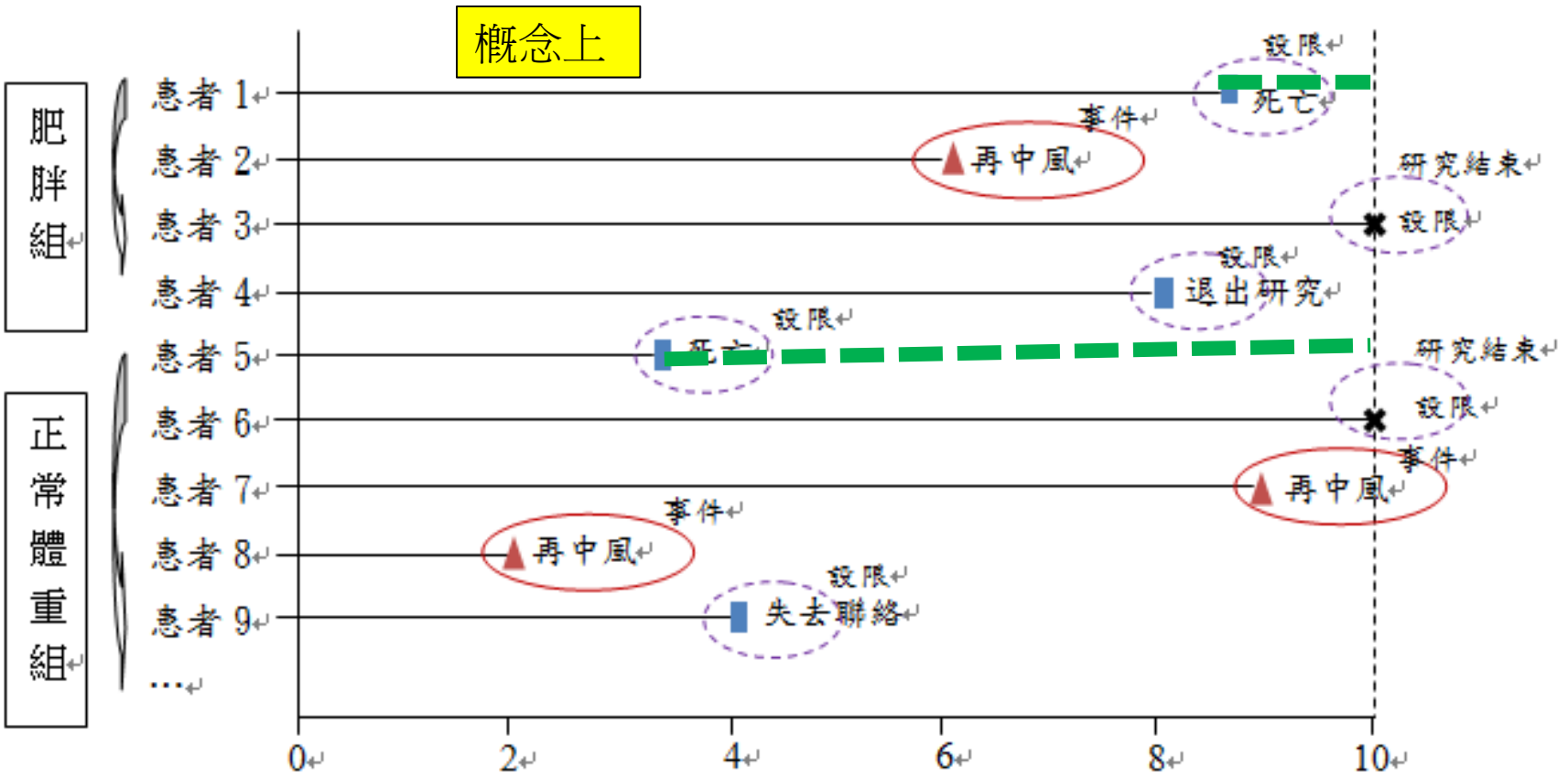
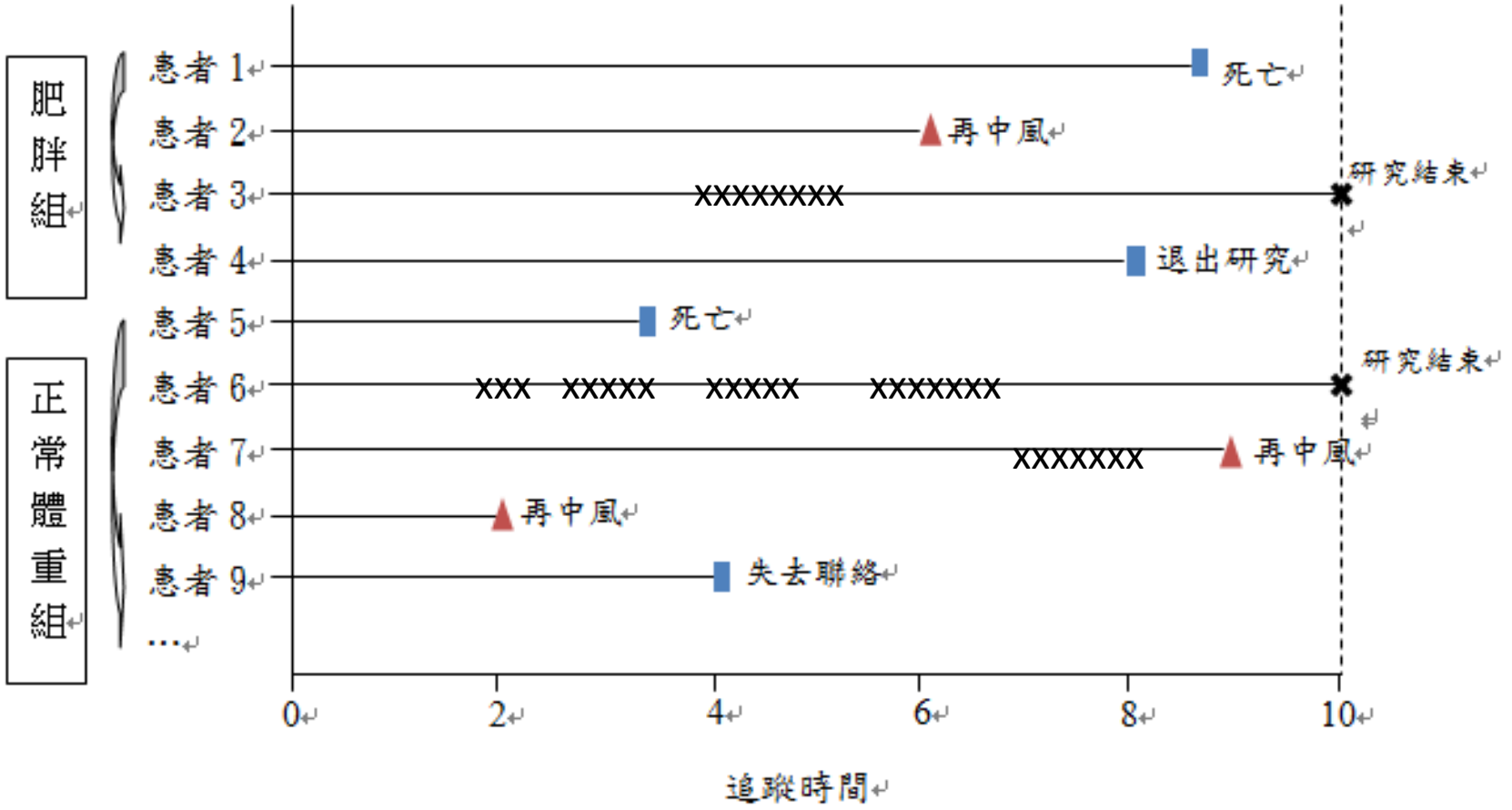


表 以 COX 迴歸模式分析再中風之影響因素

	Cox's model Caused-specific Hazard 多變項分析		Fine and Gray's Model Subdistribution Hazard 多變項分析	
	HR(95%CI)	*P* 值#	HR(95%CI)	*P* 值#
年齡(歲、每增加一歲)	1.3(1.2-1.4)	<.0001	1.1(1.05-1.3)	.0051
體重				
正常體重	1		1	
肥胖	2.1(1.9-2.4)	<0.001	1.9(1.7-2.3)	0.0034

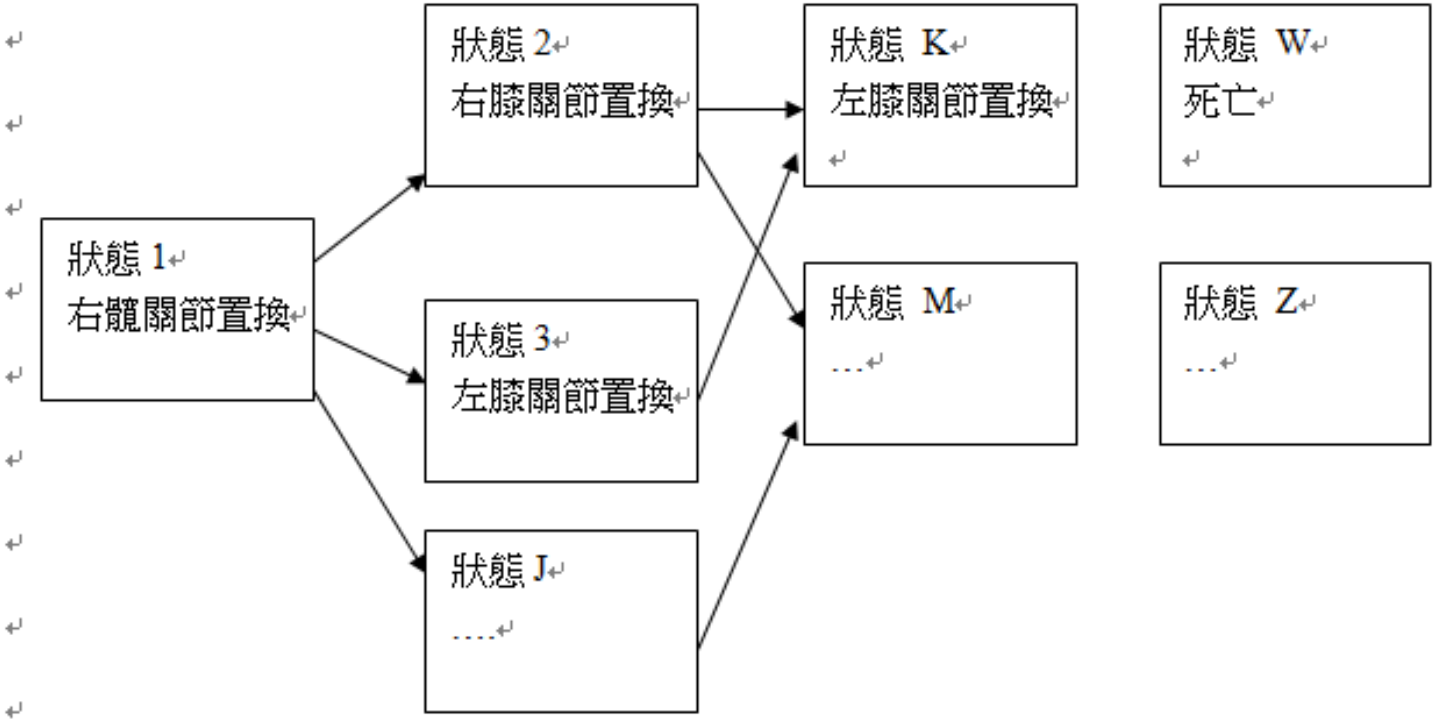
2.考慮隨時間變動的共變數的存活分析基本概念

- 一般存活分析的解釋變項通常是研究一開始就存在的變項（例如性別、年齡、有無肥胖）
- 若我們感興趣的變項是追蹤期間才可能會發生的變項，例如以針灸治療介入看是否能預防再中風，則可用**時間相依**共**變數**之存活分析(Cox proportional hazard model with time-dependent covariate)



1. **xxxx:**針灸介入、可當作時間相依的二元類別變數、亦可計算累計日數的影響
2. 可用來避免**immortal time bias**

3.考慮多重狀態(multi-state)的存活分析基本概念



右髖關節置換→左膝關節置換的風險 (state1→state3) > 右髖關節置換→右膝關節置換的風險 (state1→state2)

結語

~~~~ 統計方法在研究上的應用

主要取決於

研究者想要表達的是甚麼概念~~~~~

謝謝聆聽!!!
~~~~