統計訓練課程(五)

Survival analysis

- Log rank test
- Survival plot
- Cox model

臺中榮總研究部

生統小組R319 徐倩儀#4047

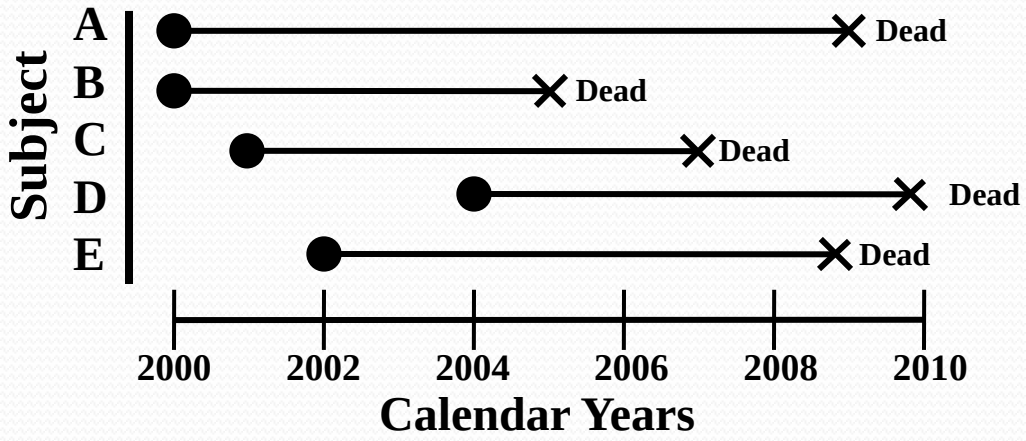
1021004

存活分析概論

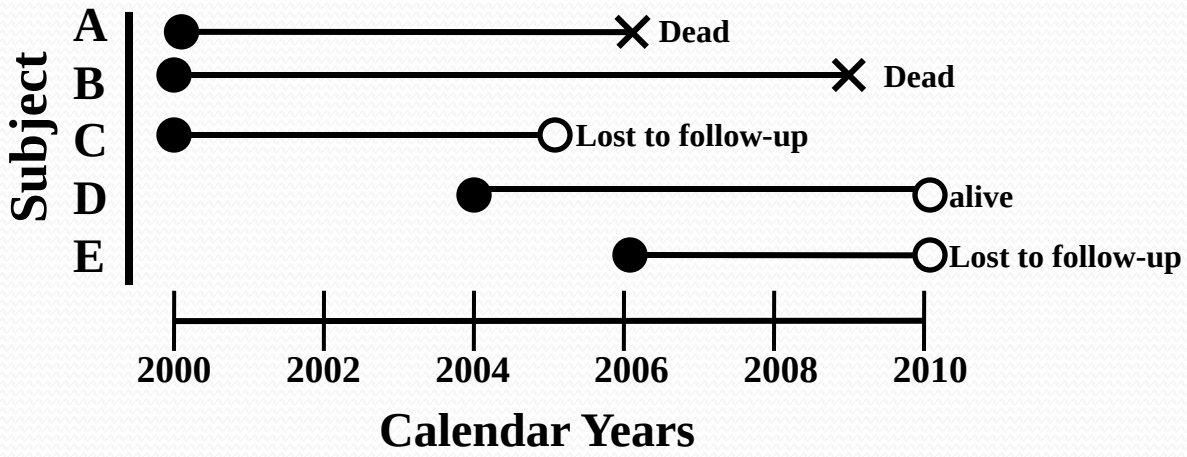
- **依變項(Dependent variable, Outcome variable)**
  - 事件發生時間(**time-to-event**)的資料
- **時間長度**
  - 時間原點(起始時間點)
    - 進入研究時間點、確診癌症時間點
  - 事件發生時間點
    - 疾病發生時間點、死亡時間點
- **事件(event)**
  - 死亡
  - 罹病
  - 復發
  - 提早退出試驗

存活分析概論

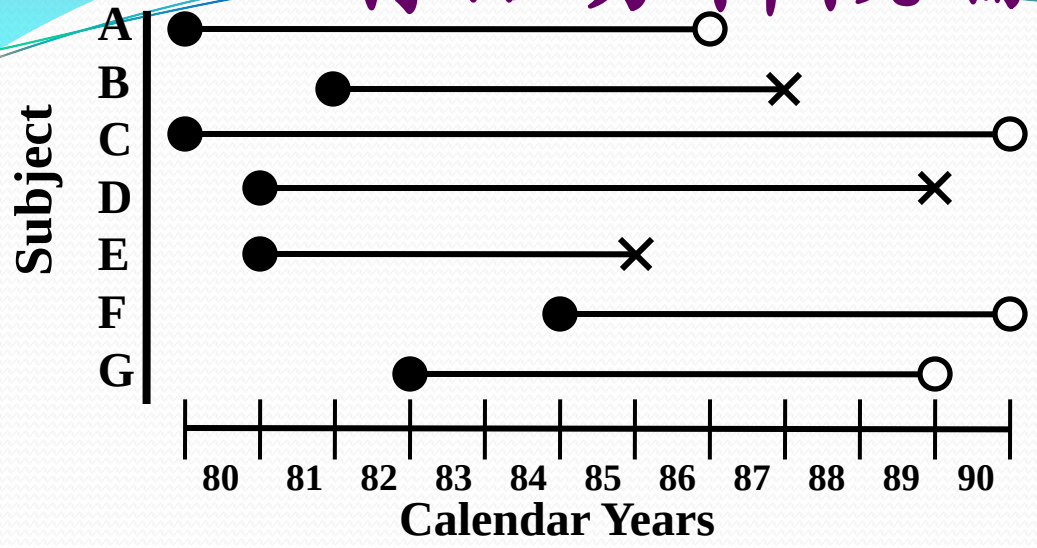
完整資料(Complete data)



設限資料(Censored data)

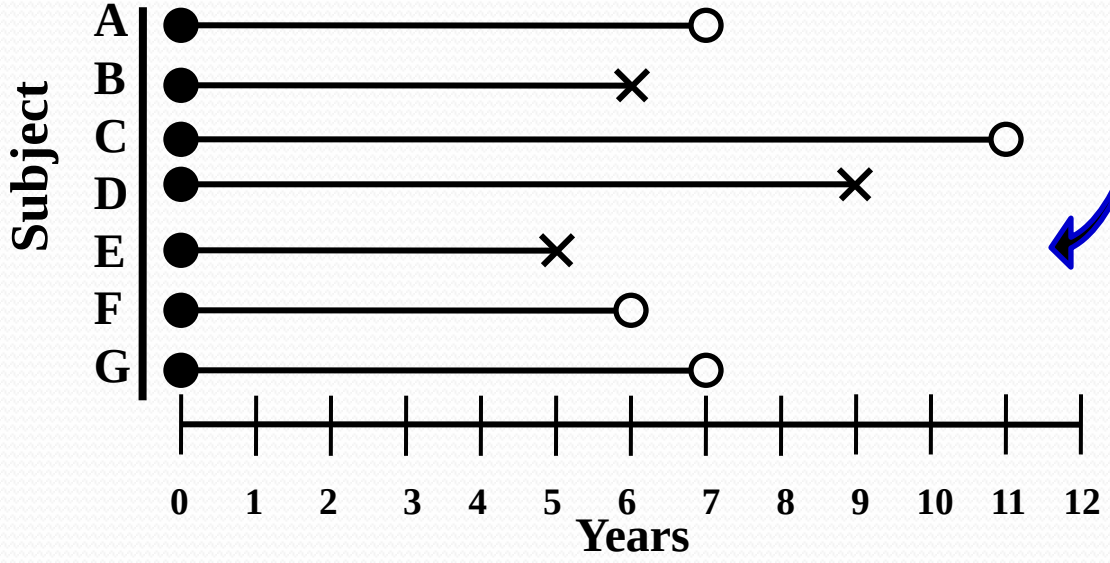


存活分析概論-時間計算



日曆時間(Calendar Time)：
從1980到1990追蹤7位有肺癌的個體情況
(● initial follow-up, × death, ○ alive)

計算自第0天(**時間零點, Day zero**)至最後追蹤的時間長度



Survival Time
7+
6
11+
9
5
6+
7+

病患研究時間(Patient time)：從1980到1990追蹤7位有肺癌的個體情況
(● initial follow-up, × death, ○ alive)

林建甫（2008)。存活分析。台北市：雙葉書廊。

資料輸入及核對

資料輸入

- 時間設定(統一格式)
  - 2007/4/24或96/4/24
- 欲探討變項
  - 資料屬性歸類
    - 連續型/類別型
      - 年齡、血糖等數值/性別、分期等等

	A	B	C	D
1	First_date	End_date	First_date	End_date
2	2007/4/24	2007/5/30	96/4/24	96/5/30
3	2011/7/18	2011/10/7	100/7/18	100/10/7
4	2010/12/22	2013/3/20	99/12/22	102/3/20
5	2007/3/19	2013/7/18	96/3/19	102/7/18

資料核對

- 計算存活(或追蹤)時間-必**大於0**
  - 年/月/日
- 確認狀態變項
  - 死亡或罹病-類別型(0 /1)
- 輸入值誤差太大
- 資料遺漏

資料輸入及核對-Excel

起始時間點

事件發生時間點

資料核對

研究結束日期

	A	B	C	D	E	F	G	H	I	J	K
1	ID	Gender	Age	First_date	TX	Death_date	Death	last_date	End_date	F_day	F_year
2	1	0	81	2007/4/24	2		0	2013/8/1	2013/8/1	2257	6.271803
3	2	0	86	2011/7/18	2		0	2013/8/1	2013/8/1	733	2.039234
4	3	1	72	2010/12/22	2	2013/3/20	1		2013/3/20	808	2.2423
5	4	0	75	2007/3/19	2	2013/7/18	1		2013/7/18	2279	6.33203
6	5	0	80	2008/1/15	2		0	2013/8/1	2013/8/1	1996	5.542883
7	6	0	84	2007/3/27	2		0	2013/8/1	2013/8/1	2284	6.348455
8	7	0	77	2008/12/28	2	2013/3/15	1		2013/3/15	1517	4.209854
9	8	0	89	2009/9/18	2		0	2013/8/1	2013/8/1	1393	3.869113
10	9	0	83	2010/10/14	2		0	2013/8/1	2013/8/1	1007	2.798084
11	10	1	78	2008/5/26	2	2013/3/11	1		2013/3/11	1725	4.790146
12	11	0	87	2008/9/24	2		0	2013/8/1	2013/8/1	1747	4.850365
13	12	0	78	2007/6/30	3		0	2013/8/1	2013/8/1	2191	6.088385
14	13	0	82	2007/11/26	2	2013/5/29	1		2013/5/29	1983	5.50528
15	14	0	82	2007/11/26	2	2013/5/29	1		2013/5/29	1983	5.50528
16	15	1	83	2007/2/7	2		0	2013/8/1	2013/8/1	2334	6.479859
17	16	0	77	2009/2/4	3	2013/7/1	1		2013/7/1	1587	4.403067
18	17	0	79	2009/10/22	2	2013/7/4	1		2013/7/4	1332	3.699343
19	18	0	56	2007/3/13	3	2013/7/25	1		2013/7/25	2292	6.367618
20	19	0	78	2010/2/26	2	2011/11/10	1		2011/11/10	614	1.70411
21	20	0	83	2011/10/6	2	2012/6/21	1		2012/6/21	255	0.70765

資料輸入及核對-SPSS

變數檢視

檔案(F) 編輯(E) 檢視(V) 資料(D) 轉換(T) 分析(A) 直效行銷(M) 統計圖(G) 效用值(U) 視窗(W) 說明(H)

	名稱	類型	寬度	小數	標記	值	遺漏	欄	對齊	測量
1	ID	數字的	11	0		無	無	11	靠右	名義(N)
2	Gender	數字的	11	0	性別	{0, Male}...	無	11	靠右	名義(N)
3	Age	數字的	11	0	年齡	無	無	11	靠右	尺度(S)
4	First_date	日期	11	0	開始時間	無	無	11	靠右	尺度(S)
5	Death_date	日期	11	0	死亡時間	無	無	11	靠右	尺度(S)
6	last_date	日期	11	0	結束時間	無	無	11	靠右	尺度(S)
7	End_date	日期	11	0	死亡/結束時間	無	無	11	靠右	尺度(S)
8	Death	數字的	11	0	死亡/存活	{0, non-death}...	無	11	靠右	名義(N)
9	TX	數字的	11	0	治療方式	{1, 治療A}...	無	11	靠右	名義(N)
10	F_day	數字的	11	0	追蹤天數	無	無	11	靠右	尺度(S)
11	F_year	數字的	11	9	追蹤年數	無	無	11	靠右	尺度(S)

資料輸入及核對-SPSS

資料檢視

計算存活(或追蹤)時間必**大於0**

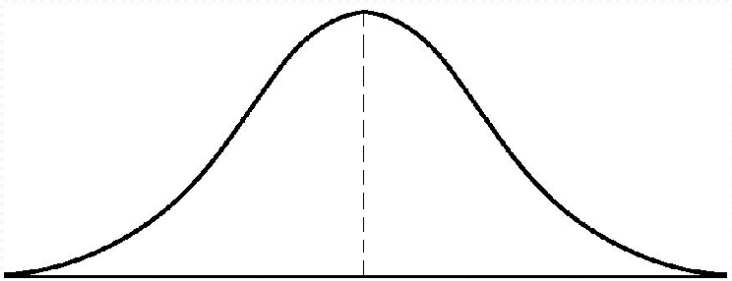
檔案(F) 編輯(E) 檢視(V) 資料(D) 轉換(T) 分析(A) 直效行銷(M) 統計圖(G) 效用值(U) 視窗(W) 說明(H)

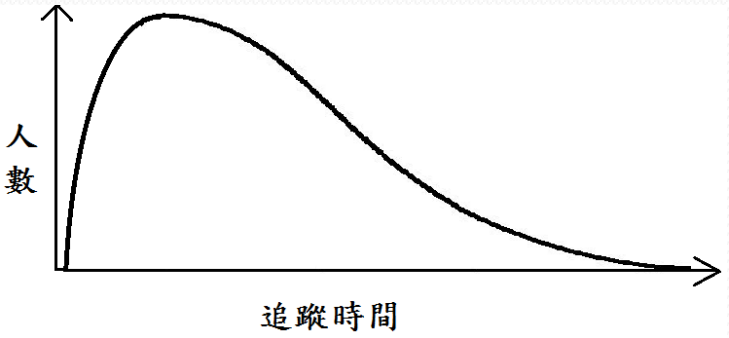
	ID	Gender	Age	First_date	Death_date	last_date	End_date	Death	TX	F_day	F_year
1	1	0	81	24-Apr-2007	.	01-Aug-2013	01-Aug-2013	0	2	2257	6.271803000
2	2	0	86	18-Jul-2011	.	01-Aug-2013	01-Aug-2013	0	2	733	2.039233700
3	3	1	72	22-Dec-2010	20-Mar-2013	.	20-Mar-2013	1	2	808	2.242299800
4	4	0	75	19-Mar-2007	18-Jul-2013	.	18-Jul-2013	1	2	2279	6.332030000
5	5	0	87	03-Aug-2010	24-Jul-2013	.	24-Jul-2013	1	1	1071	2.973306000
6	6	0	80	15-Jan-2008	.	01-Aug-2013	01-Aug-2013	0	2	1996	5.542883400
7	7	0	84	27-Mar-2007	.	01-Aug-2013	01-Aug-2013	0	2	2284	6.348455400
8	8	0	77	28-Dec-2008	15-Mar-2013	.	15-Mar-2013	1	2	1517	4.209854000
9	9	0	89	18-Sep-2009	.	01-Aug-2013	01-Aug-2013	0	2	1393	3.869112700
10	10	0	83	14-Oct-2010	.	01-Aug-2013	01-Aug-2013	0	2	1007	2.798083500
11	11	1	78	26-May-2008	11-Mar-2013	.	11-Mar-2013	1	2	1725	4.790146000
12	12	0	77	09-Jun-2008	24-Jul-2013	.	24-Jul-2013	1	1	1845	5.121350300
13	13	0	87	24-Sep-2008	.	01-Aug-2013	01-Aug-2013	0	2	1747	4.850365000
14	14	0	78	30-Jun-2007	.	01-Aug-2013	01-Aug-2013	0	1	2191	6.088384600
15	15	0	82	26-Nov-2007	29-May-2013	.	29-May-2013	1	2	1983	5.505279500
16	16	0	82	26-Nov-2007	29-May-2013	.	29-May-2013	1	2	1983	5.505279500
17	17	1	83	07-Feb-2007	.	01-Aug-2013	01-Aug-2013	0	2	2334	6.479859400
18	18	0	86	30-Sep-2008	12-Jul-2013	.	12-Jul-2013	1	1	1722	4.779197000
19	19	0	77	04-Feb-2009	01-Jul-2013	.	01-Jul-2013	1	1	1587	4.403066600
20	20	0	79	22-Oct-2009	04-Jul-2013	.	04-Jul-2013	1	2	1332	3.699342700
21	21	0	56	13-Mar-2007	25-Jul-2013	.	25-Jul-2013	1	1	2292	6.367618000
22	22	0	78	26-Feb-2010	10-Nov-2011	.	10-Nov-2011	1	2	614	1.704109500
23	23	1	50	30-Mar-2011	.	01-Aug-2013	01-Aug-2013	0	1	841	2.340328500
24	24	1	61	16-Sep-2009	.	01-Aug-2013	01-Aug-2013	0	1	1395	3.874589200
25	25	1	70	13-Aug-2011	23-Jul-2013	.	23-Jul-2013	1	1	700	1.943430700

統計分析方法

資料特性

非常態分布(右偏分佈)

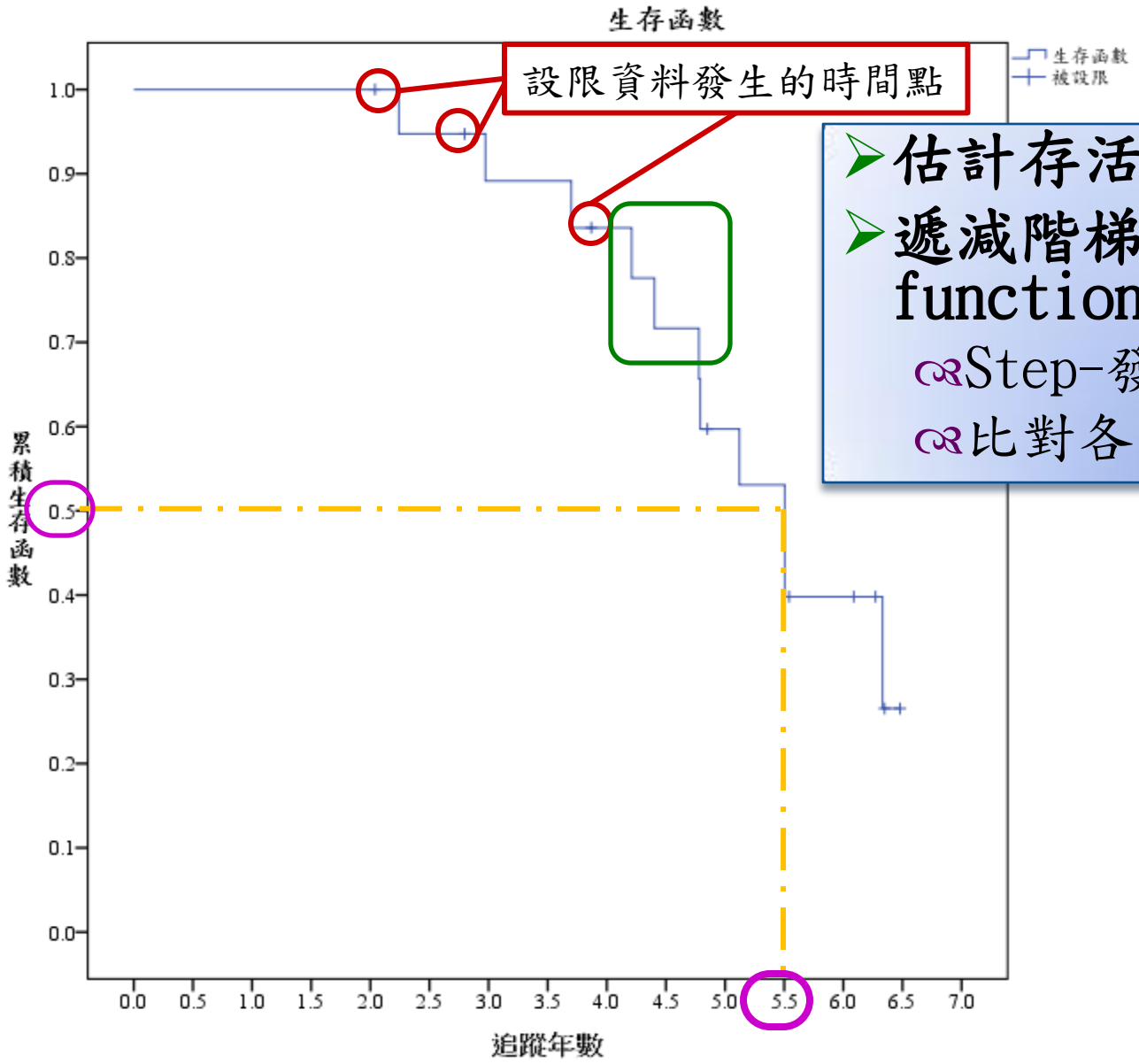




無母數方法

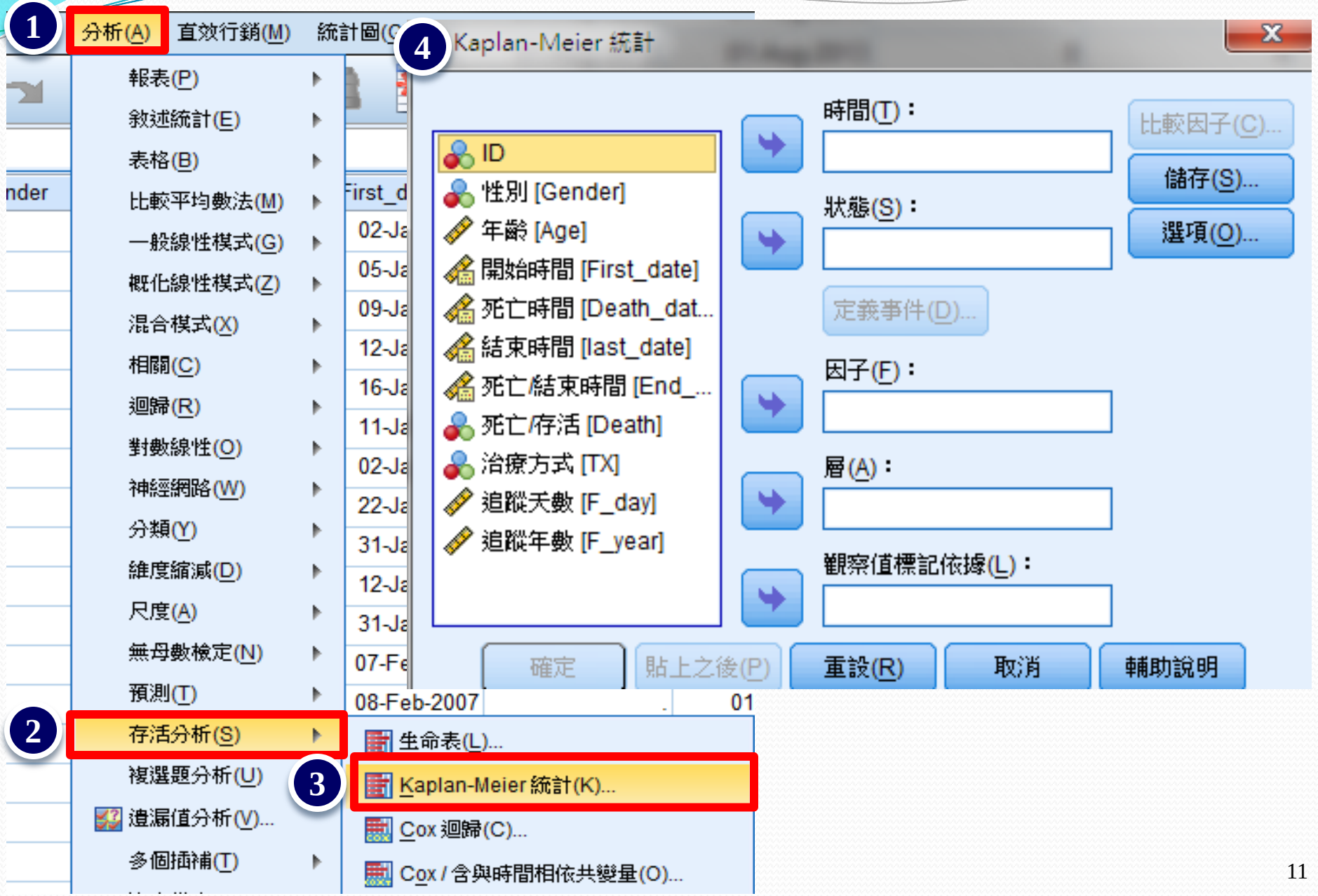
- Kaplan-Meier Survival Analysis
- Log rank test
- Cox Proportional Hazards Model

Kaplan-Meier analysis



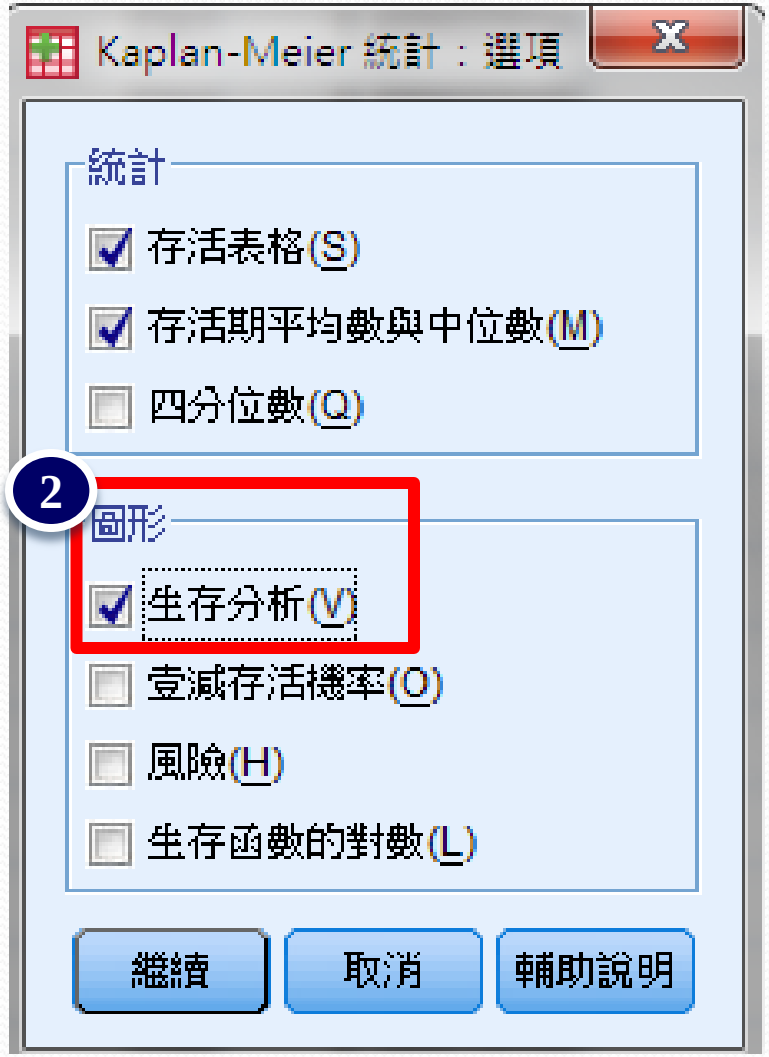
- 估計存活函數
- 遞減階梯函數(Step function)
  - Step-發生事件的時間點
  - 比對各時間點的存活率

Kaplan-Meier analysis



Kaplan-Meier analysis

Kaplan-Meier analysis



Kaplan-Meier analysis 輸出

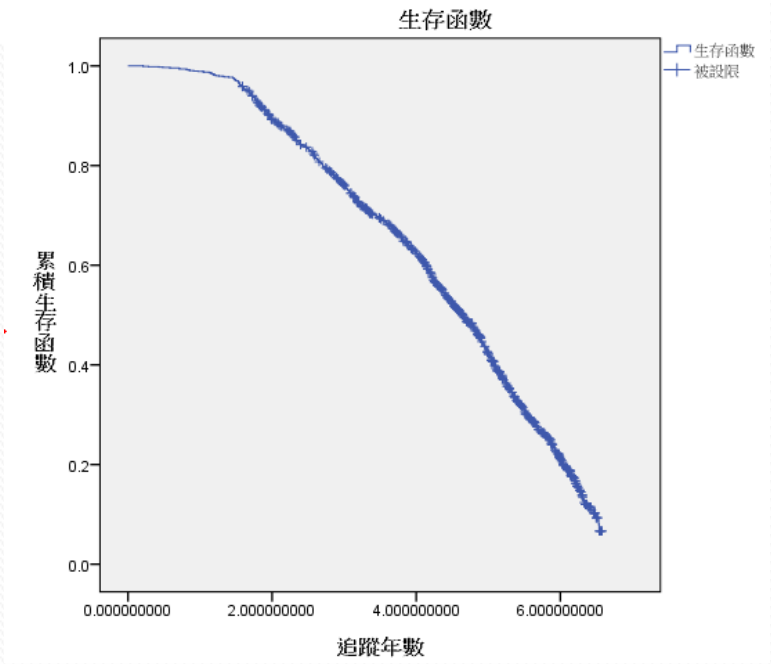
觀察值處理摘要

總個數	事件數量	設限的	
		個數	百分比
1274	867	407	31.9%

存活時間的平均數與中位數

平均數[a]				中位數			
		95% 信賴區間				95% 信賴區間	
估計	標準誤差	下界	上界	估計	標準誤差	下界	上界
4.399	.047	4.306	4.491	4.645	.080	4.489	4.801

a. 估計如果經過設限，則限制為最大的存活時間。



Kaplan-Meier analysis 輸出

完整資料(Complete data)-1

觀察值處理摘要

總個數	事件數量	設限的	
		個數	百分比
867	867	0	.0%

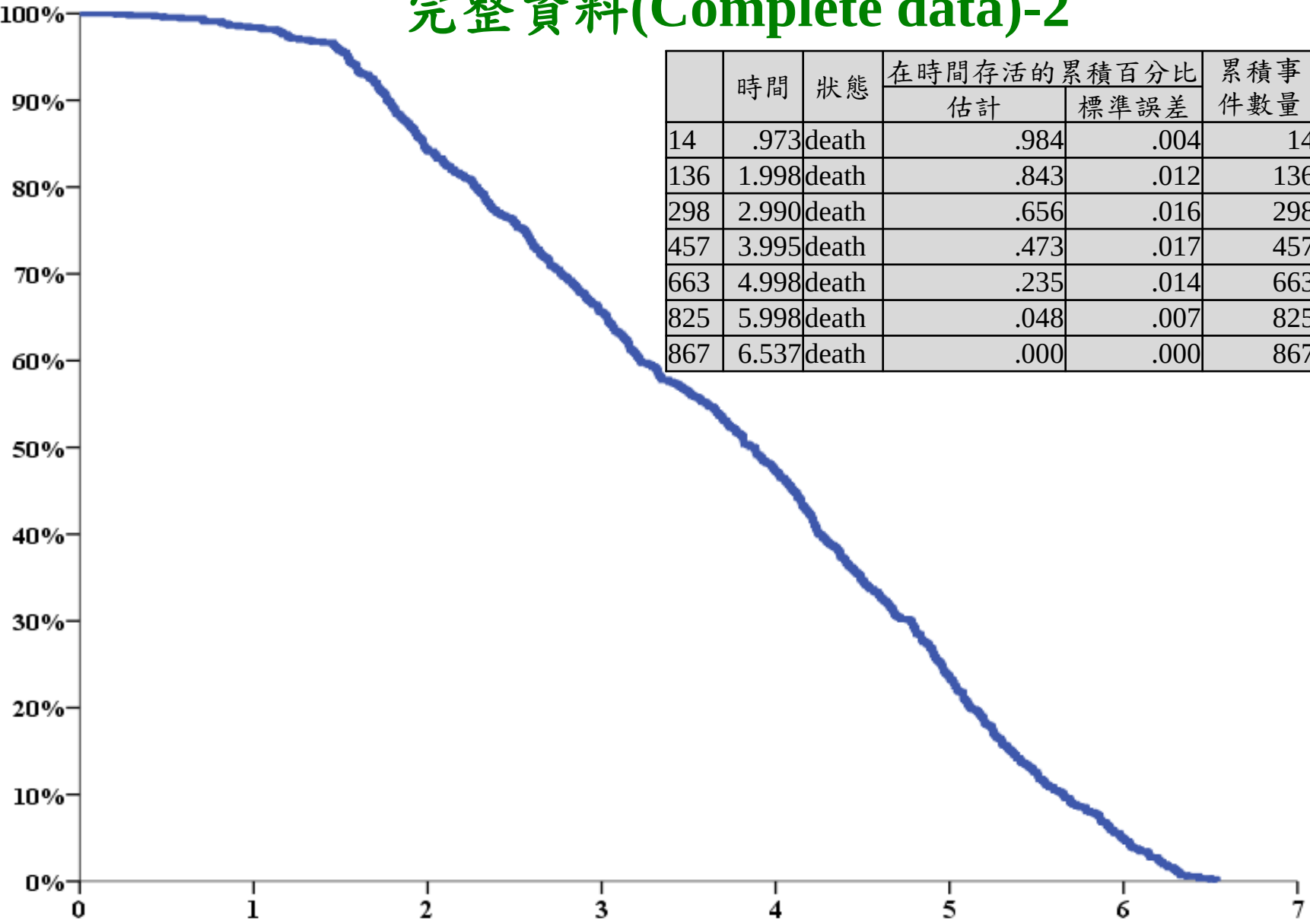
存活時間的平均數與中位數

平均數[a]				中位數			
		95% 信賴區間				95% 信賴區間	
估計	標準誤差	下界	上界	估計	標準誤差	下界	上界
3.743	.049	3.647	3.839	3.877	.085	3.710	4.044

a. 估計如果經過設限，則限制為最大的存活時間。

Kaplan-Meier analysis 輸出

完整資料(Complete data)-2



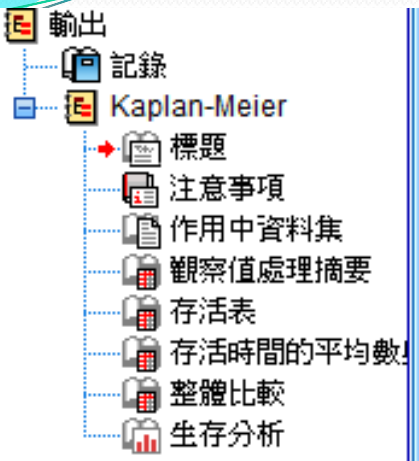
	時間	狀態	在時間存活的累積百分比		累積事件數量	剩餘的觀察值個數
			估計	標準誤差		
14	.973	death	.984	.004	14	853
136	1.998	death	.843	.012	136	731
298	2.990	death	.656	.016	298	569
457	3.995	death	.473	.017	457	410
663	4.998	death	.235	.014	663	204
825	5.998	death	.048	.007	825	42
867	6.537	death	.000	.000	867	0

Log rank test

- **單變量概念**
  - 找出個別與存活狀態（或某事件）的顯著危險因子（考慮時間的因素）
- **檢定兩條(或多條)存活曲線是否相同**
  - 僅能提供P value

Log rank test-性別

Log rank test-性別1



Kaplan-Meier

[資料集1] D:\2013.10.04_Survival\Survival_data.sav

觀察值處理摘要

性別	總個數	事件數量	設限的	
			個數	百分比
Male	765	512	253	33.1%
Female	509	355	154	30.3%
整體	1274	867	407	31.9%

存活表

性別		時間	狀態	在時間存活的累積百分比		累積事件數量	剩餘的觀察值個數
				估計	標準誤差		
Male	1	.211	death	.999	.001	1	764
	2	.285	death	.997	.002	2	763
	3	.432	death	.996	.002	3	762
	4	.481	death	.995	.003	4	761
	5	.708	death	.	.	5	760
	6	.708	death	.992	.003	6	759

Log rank test-性別2

存活時間的平均數與中位數

性別	平均數[a]				中位數			
			95% 信賴區間				95% 信賴區間	
	估計	標準誤差	下界	上界	估計	標準誤差	下界	上界
Male	4.433	.060	4.315	4.551	4.782	.097	4.591	4.972
Female	4.344	.075	4.196	4.491	4.491	.138	4.220	4.762
整體	4.399	.047	4.306	4.491	4.645	.080	4.489	4.801

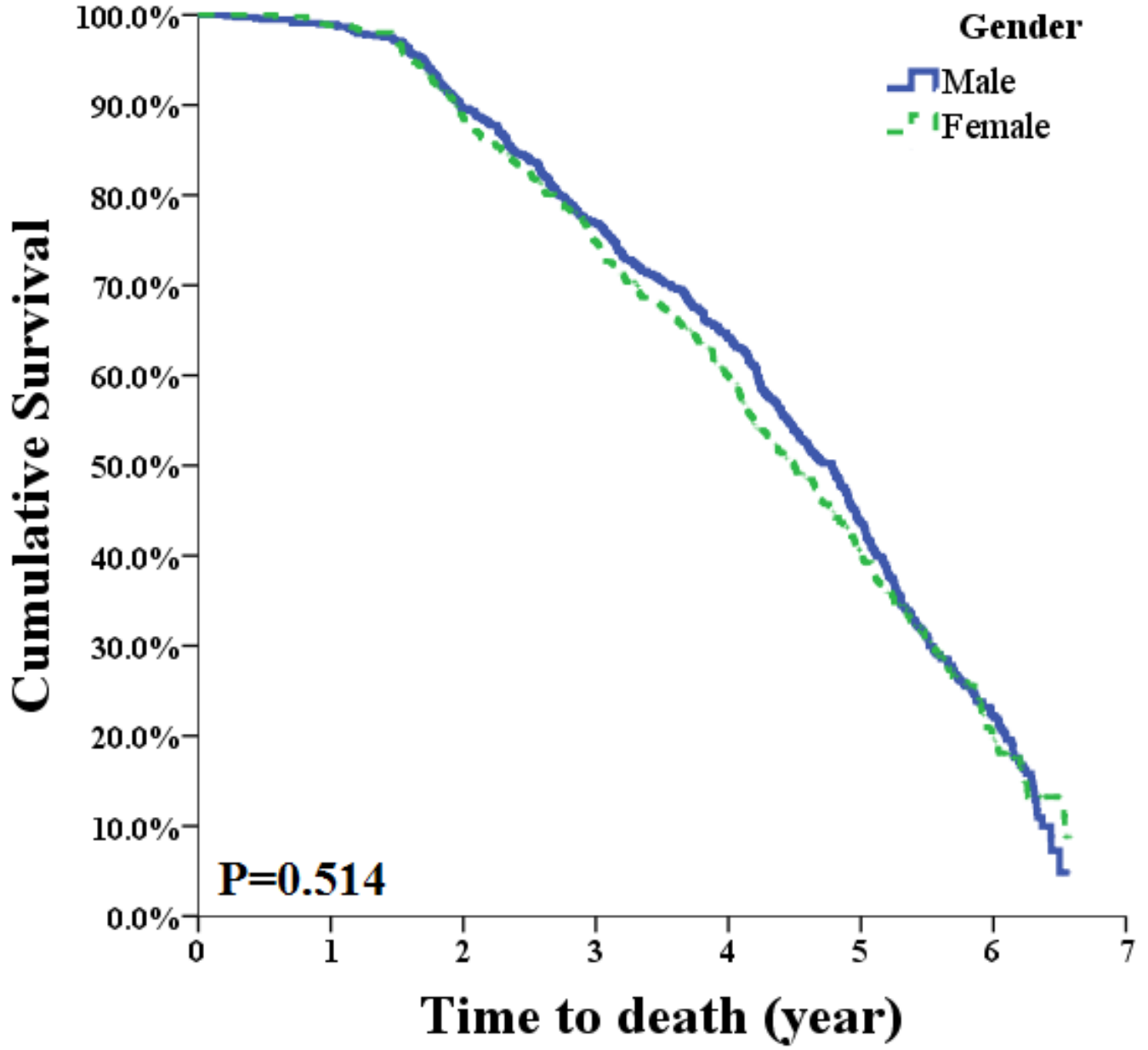
a. 估計如果經過設限，則限制為最大的存活時間。

整體比較

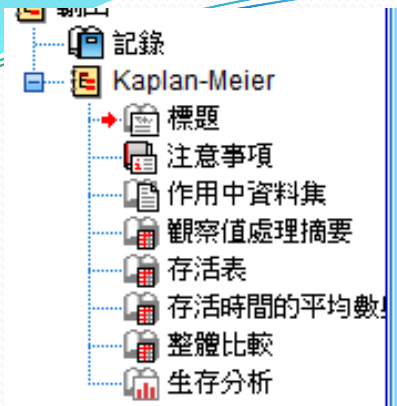
	卡方統計量	自由度	顯著性
Log Rank (Mantel-Cox)	.425	1	.514

性別 不同水準的存活分配等式檢定。

Log rank test-性別3 (Survival curve)



Log rank test-治療方式1

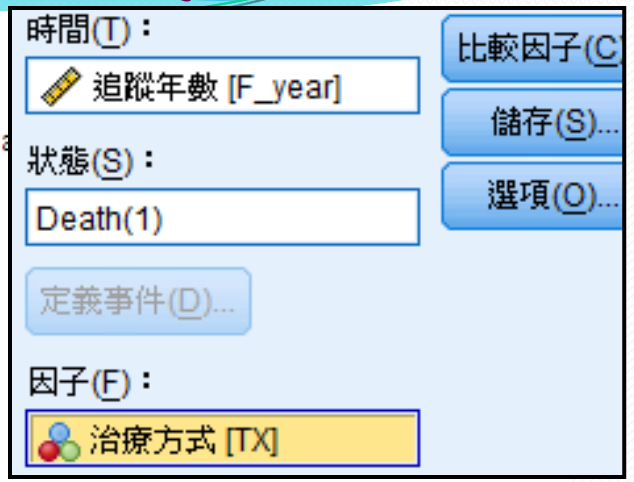


Kaplan-Meier

[資料集1] D:\2013.10.04_Survival\Survival_data.sa

觀察值處理摘要

治療方式	總個數	事件數量	設限的	
			個數	百分比
治療A	589	387	202	34.3%
治療B	274	155	119	43.4%
治療C	411	325	86	20.9%
整體	1274	867	407	31.9%



存活表

治療方式		時間	狀態	在時間存活的累積百分比		累積事件數量	剩餘的觀察值個數
				估計	標準誤差		
治療A	1	.211	death	.998	.002	1	588
	2	1.040	death	.997	.002	2	587
	3	1.053	death	.995	.003	3	586
	4	1.166	death	.993	.003	4	585
	5	1.193	death	.992	.004	5	584
	6	1.197	death	.990	.004	6	583
	7	1.198	death	.988	.004	7	582
	8	1.398	death	.986	.005	8	581
	9	1.467	death	.985	.005	9	580

Log rank test-治療方式2

存活時間的平均數與中位數

治療方式	平均數[a]				中位數			
			95% 信賴區間				95% 信賴區間	
	估計	標準誤差	下界	上界	估計	標準誤差	下界	上界
治療A	4.584	.067	4.452	4.716	4.934	.094	4.749	5.119
治療B	4.323	.116	4.095	4.550	4.507	.230	4.056	4.958
治療C	4.206	.077	4.055	4.357	4.315	.106	4.109	4.522
整體	4.399	.047	4.306	4.491	4.645	.080	4.489	4.801

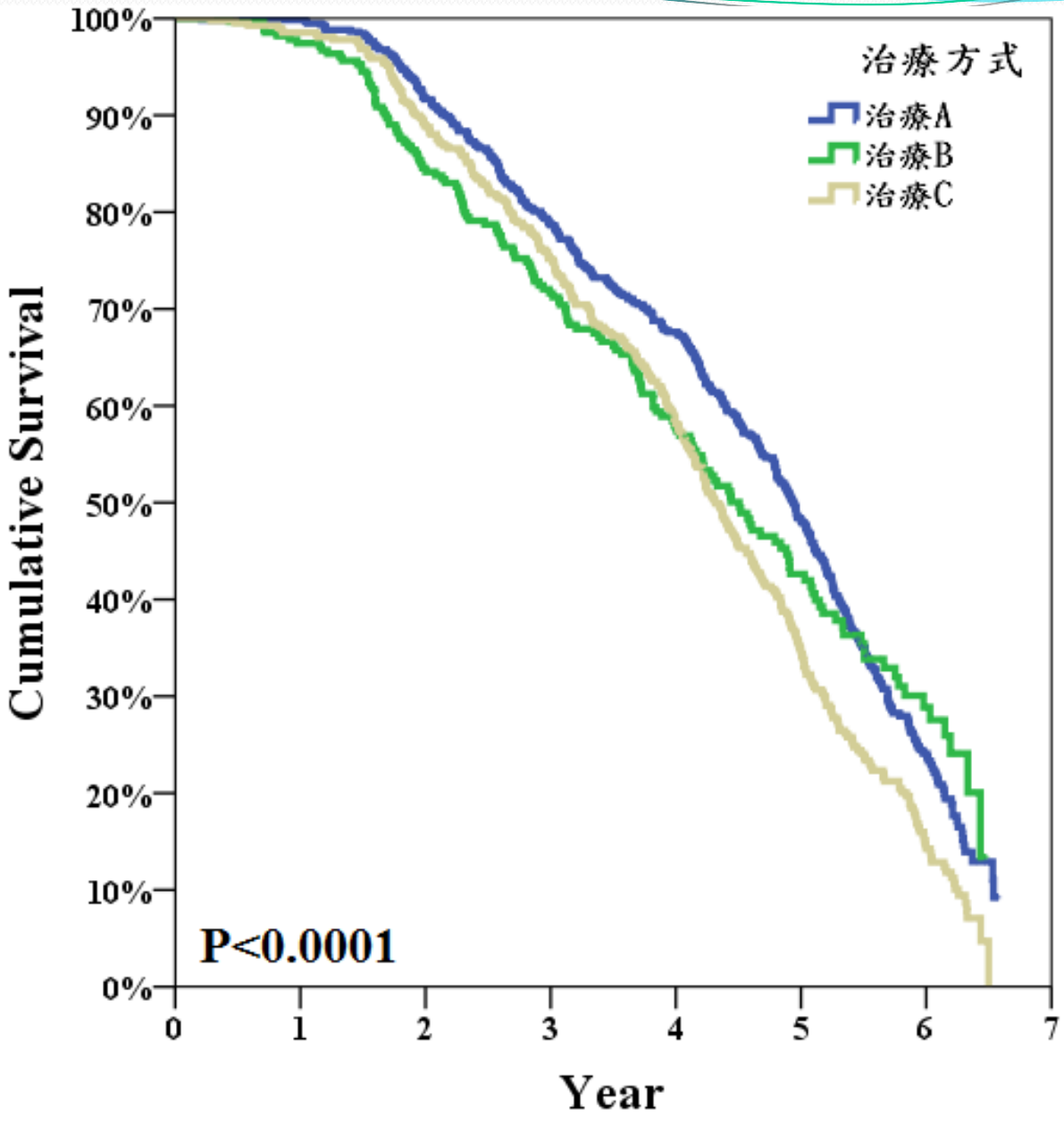
a. 估計如果經過設限，則限制為最大的存活時間。

整體比較

	卡方統計量	自由度	顯著性
Log Rank (Mantel-Cox)	16.425	2	.000

治療方式 不同水準的存活分配等式檢定。

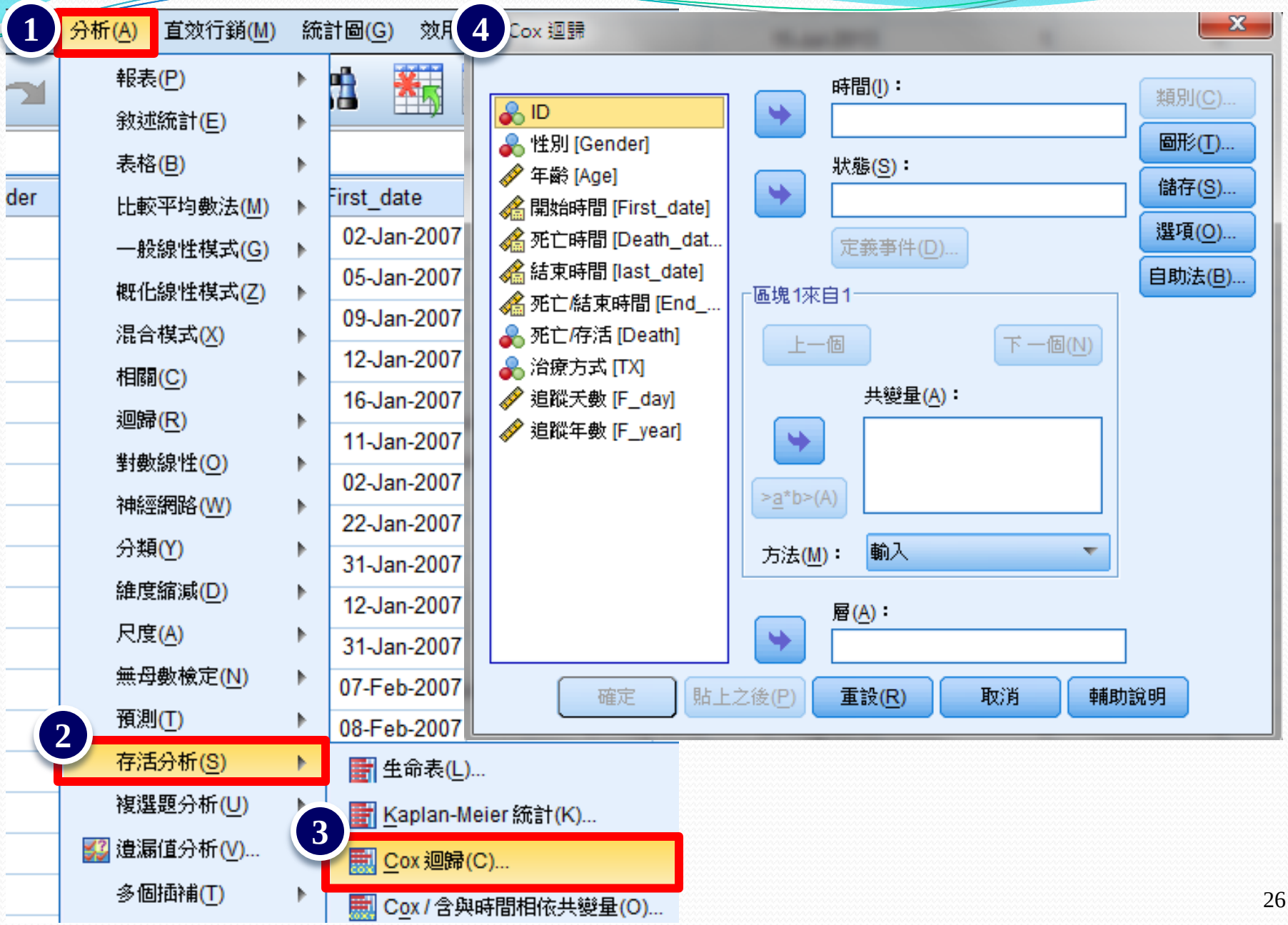
Log rank test-治療方式3 (Survival curve)



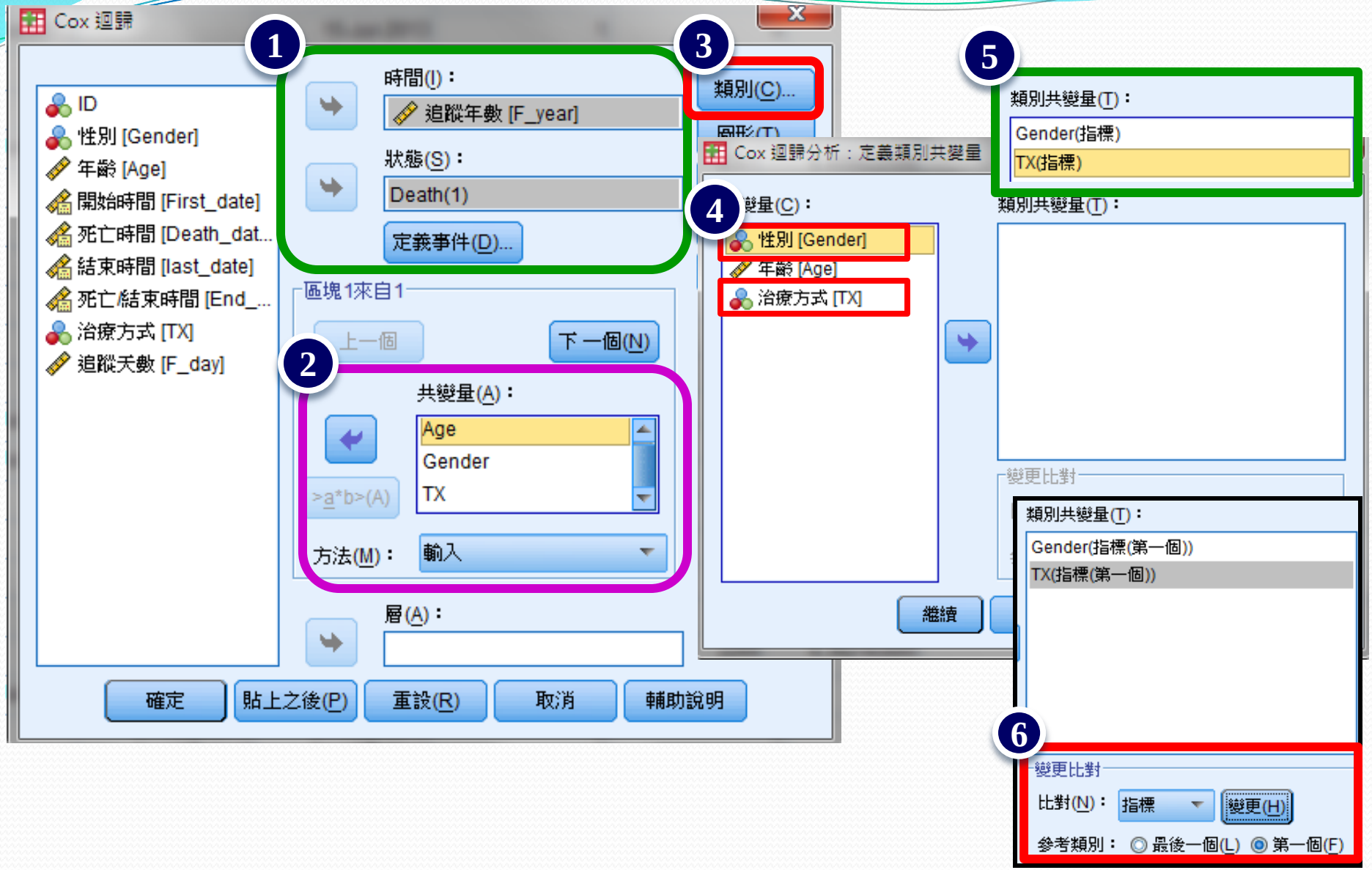
Cox model

- **Cox Proportional Hazards Model**
- **評估多個共變量同時存在同一個模型中**
  - 多變量概念
    - 單變量P值顯著的變數
    - P值不顯著，但對研究主題重要的變數
      - 年齡、性別等等
- **自變項-連續型或類別型**
- **估計危險比(Hazards Ratio)**

Cox model


1
分析(A)
直效行銷(M)
統計圖(G)
報表(P)
敘述統計(E)
表格(B)
比較平均數法(M)
一般線性模式(G)
概化線性模式(Z)
混合模式(X)
相關(C)
迴歸(R)
對數線性(O)
神經網路(W)
分類(Y)
維度縮減(D)
尺度(A)
無母數檢定(N)
預測(T)
2
存活分析(S)
複選題分析(U)
遺漏值分析(V)...
多個插補(T)
生命表(L)...
Kaplan-Meier 統計(K)...
3
Cox 迴歸(C)...
Cox / 含與時間相依共變量(O)...
First_date
02-Jan-2007
05-Jan-2007
09-Jan-2007
12-Jan-2007
16-Jan-2007
11-Jan-2007
02-Jan-2007
22-Jan-2007
31-Jan-2007
12-Jan-2007
31-Jan-2007
07-Feb-2007
08-Feb-2007
4
Cox 迴歸
ID
性別 [Gender]
年齡 [Age]
開始時間 [First_date]
死亡時間 [Death_dat...
結束時間 [last_date]
死亡/結束時間 [End_...
死亡/存活 [Death]
治療方式 [TX]
追蹤天數 [F_day]
追蹤年數 [F_year]
時間(I)：
狀態(S)：
定義事件(D)...
區塊 1來自1
上一個
下一個(N)
共變量(A)：
>a*b>(A)
方法(M)：
輸入
層(A)：
類別(C)...
圖形(T)...
儲存(S)...
選項(O)...
自助法(B)...
確定
貼上之後(P)
重設(R)
取消
輔助說明

Cox model


Cox 迴歸
ID
性別 [Gender]
年齡 [Age]
開始時間 [First_date]
死亡時間 [Death_dat...
結束時間 [last_date]
死亡/結束時間 [End_...
治療方式 [TX]
追蹤天數 [F_day]
時間(I)：
追蹤年數 [F_year]
狀態(S)：
Death(1)
定義事件(D)...
區塊 1來自 1
上一個
下一個(N)
共變量(A)：
Age
Gender
TX
>a*b>(A)
方法(M)：
輸入
層(A)：
確定
貼上之後(P)
重設(R)
取消
輔助說明
類別(C)...
Cox 迴歸分析：定義類別共變量
變量(C)：
性別 [Gender]
年齡 [Age]
治療方式 [TX]
繼續
類別共變量(T)：
Gender(指標)
TX(指標)
變更比對
類別共變量(T)：
Gender(指標(第一個))
TX(指標(第一個))
變更比對
比對(N)：
指標
變更(H)
參考類別：
最後一個(L)
第一個(F)
1
2
3
4
5
6

Cox model

Cox 迴歸
ID
性別 [Gender]
年齡 [Age]
開始時間 [First_date]
死亡時間 [Death_dat...
結束時間 [last_date]
死亡/結束時間 [End_...
治療方式 [TX]
追蹤天數 [F_day]
時間(I)：
追蹤年數 [F_year]
狀態(S)：
Death(1)
定義事件(D)...
類別(C)...
圖形(T)...
儲存(S)...
選項(O)...
自助法(B)...
1
區塊 1來自 1
上一個
下一個(N)
共變量(A)：
Age
Gender(Cat)
TX(Cat)
>a*b>(A)
方法(M)：
輸入
層(A)：
確定
貼上之後(P)
重設(R)
取消
輔助說明
Cox 迴歸分析：選項
模式統計量
exp (B) 的 C1 (C)
95
%
估計值相關性(R)
顯示模式訊息
在每一步驟(E)
在最後步驟(L)
逐步之機率
輸入(E)：
.05
移除(A)：
.10
最大疊代(I)：
20
顯示基準線函數(B)
繼續
取消
輔助說明
2
Cox 迴歸分析：選項
模式統計量
exp (B) 的 C1 (C)
95
%
估計值相關性(R)
顯示模式訊息
在每一步驟(E)
在最後步驟(L)
逐步之機率
輸入(E)：
.05
移除(A)：
.10
最大疊代(I)：
20
顯示基準線函數(B)
繼續
取消
輔助說明

Cox model

觀察值處理摘要

		N	百分比
分析中可用的觀察值	事件[a]	867	68.1%
	被設限的	407	31.9%
	總和	1274	100.0%
刪除的觀察值	有遺漏值的觀察值	0	.0%
	含有負數時間的觀察值；	0	.0%
	對分層觀察值		
	總和		
總和			

a. 依變數：追蹤年數

類別變數編碼[b,c]

		次數	(1)	(2)
Gender[a]	0=Male	765	0	
	1=Female	509	1	
TX[a]	1=治療A	589	0	0
	2=治療B	274	1	0
	3=治療C	411	0	1

a. 指標參數編碼g

b. 類別變數： Gender (性別)

c. 類別變數： TX (治療方式)

參考組

以相對於存活來說，風險最低的當參考組(0)

Cox model

方程式中的變數

	B	SE	Wald	自由度	顯著性。	Exp(B)	Exp(B) 的 95.0% CI	
							下界	上界
Age	-.003	.003	1.082	1	.298	.997	.991	1.003
Gender	.042	.070	.359	1	.549	1.043	.910	1.195
TX			17.195	2	.000			
TX(1)	.136	.100	1.863	1	.172	1.146	.942	1.392
TX(2)	.318	.077	17.113	1	.000	1.374	1.182	1.598

	Univariate analysis			Multivariate analysis		
	HR	95 % CI	*P* -value	HR	95 % CI	*P* -value
Age	0.999	(0.994- 1.004)	0.7272	0.997	(0.991- 1.003)	0.2981
Gender						
Male		ref.			ref.	
Female	1.046	(0.913- 1.198)	0.5144	1.043	(0.910- 1.195)	0.5490
TX						
治療A		ref.			ref.	
治療B	1.107	(0.918- 1.334)	0.2875	1.146	(0.942- 1.392)	0.1723
治療C	1.354	(1.168- 1.569)	<0.0001**	1.374	(1.182- 1.598)	<0.0001**

Cox regression. HR, Hazards Ratio. *P<0.05 **P<0.01

Outcome：得到肺癌與否

範例：研究變項含有：得到肺癌與否、吸菸年數、疾病史、性別等等變項

新增研究變項：
收案時間
確診肺癌時間
研究結束時間

Logistic Regression or Cox Regression ?!

- Logistic regression
  - 沒有考慮時間
- Cox regression
  - 考慮時間(Censored data)
  - HR含有baseline風險概念

Thank you

For your attention!!