

快速由資料庫上手醫學 論文投稿

醫學研究部 基礎醫學科 生統小組

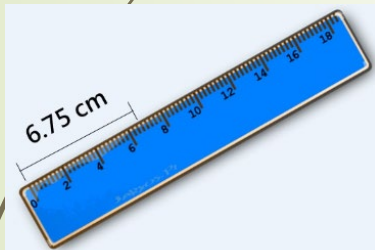
陳韻仔 高級助理研究員

授課日期：112年3月29日

Data Definition

Variable type

Numerical variables



連續變項
Continuous Variables

➤ Height, Weight, IQ, Temperature
(可以有小數點)

間斷變項
Discrete Variables



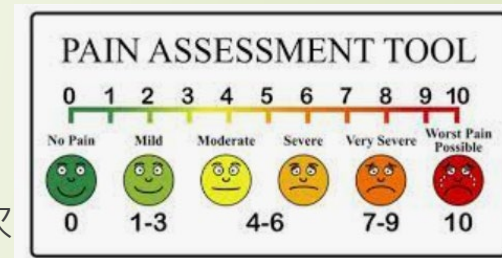
➤ Number of children
(沒有小數點)

Categorical variables

Ordinal 次序變項

(有依程度排序)

➤ Pain severity、考試名次



Nominal 名義變項

(不排序，利用名稱或數值區分)

➤ Race, Gender, Education levels

資料分佈

常態分佈

□ 高斯 (Gaussian) distribution

✓ 平均值 (mean)

- 中心點位置

✓ 標準差 (Std)

- 鐘型曲線形狀

✓ **95% CI of mean: Mean $\pm$ 1.96 Std**

□ 以平均值為中心的對稱曲線

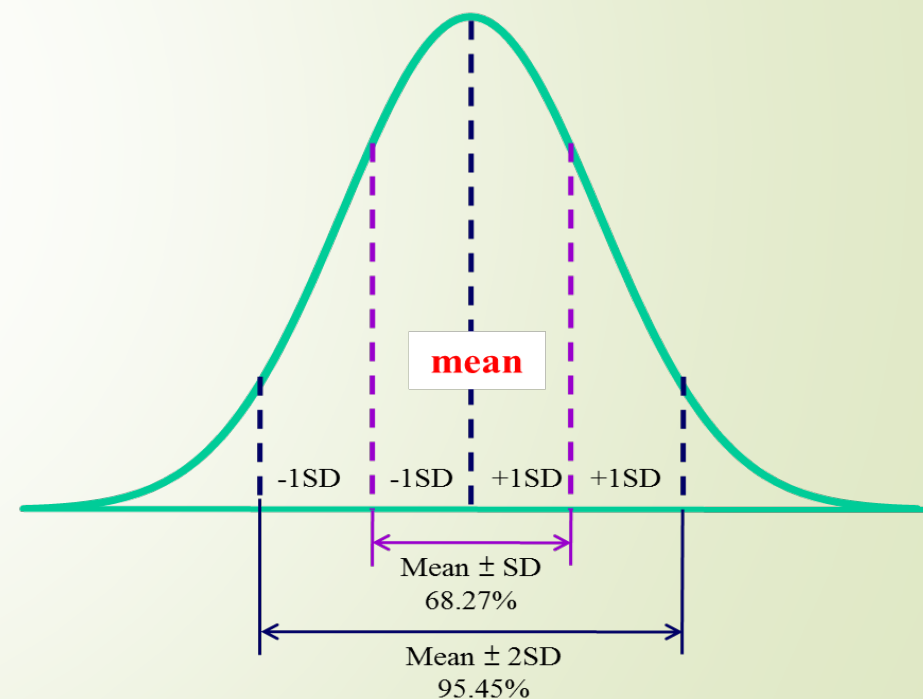
□ Mean = Median = Mode

名詞	簡稱	公式	意義
Standard Deviation	SD	$\sqrt{\frac{\sum (x_i - u)^2}{n}}$	樣本各資料點分布的離散或變異情形
Standard Error	SE	$\frac{\sigma}{\sqrt{n}}$	樣本平均值的估計準確度

常態性檢定→

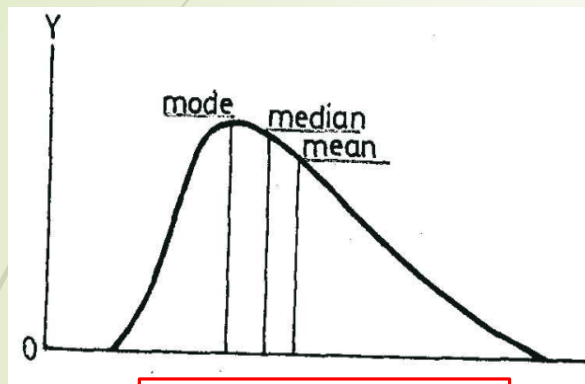
Kolmogorov-Smirnov (K-S) 檢定：樣本數**50個以上**

Shapiro-Wilk (S-W) 檢定：樣本數**50個以下**

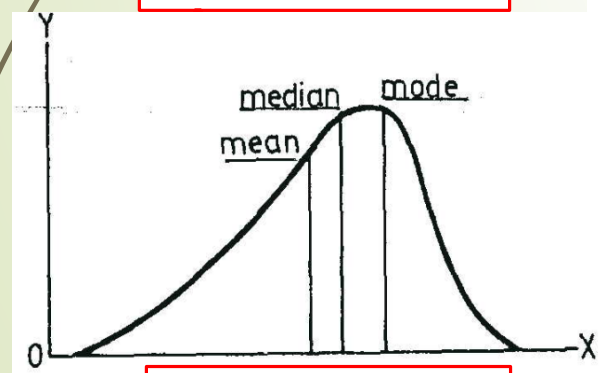


資料分佈

非常態分佈



右偏分布

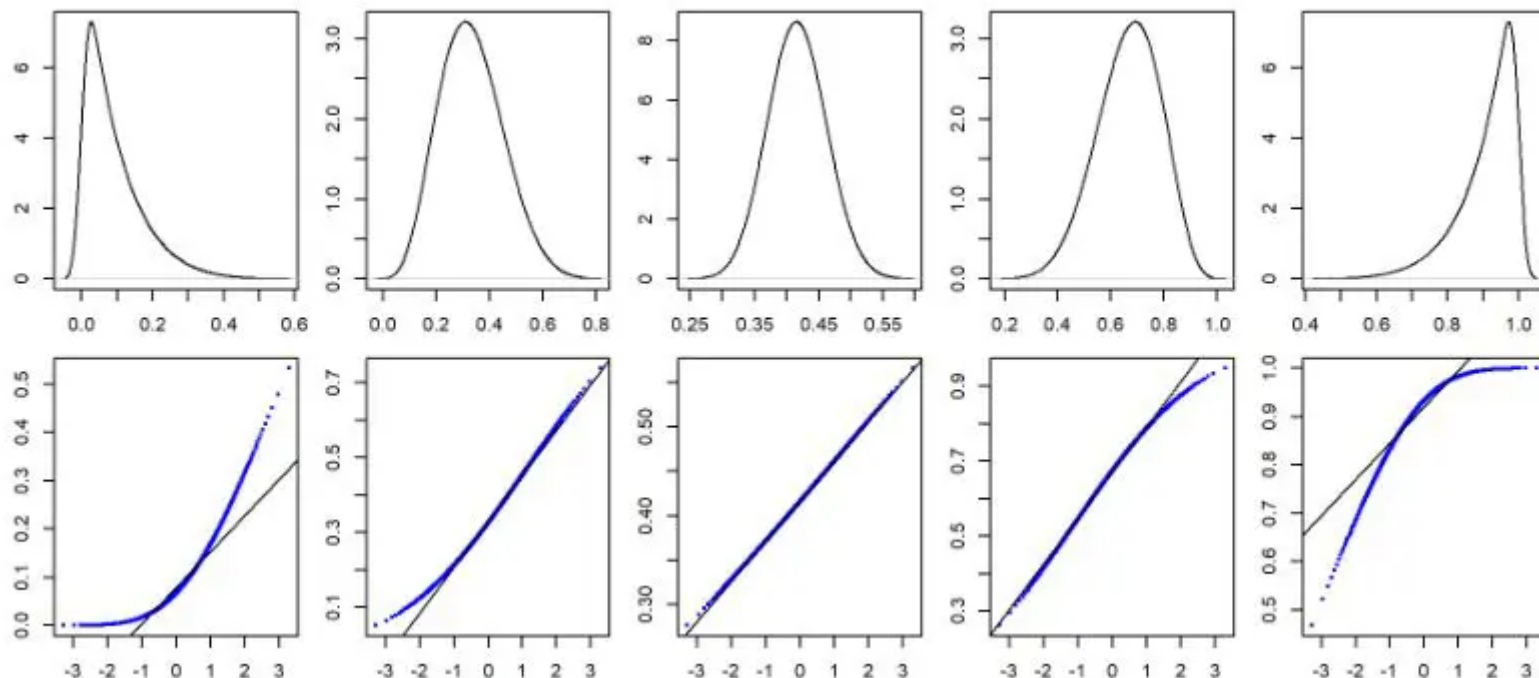


左偏分布

右偏

常態分佈

左偏



Q-Q plot

```
> shapiro.test(wage)
```

Shapiro-Wilk normality test

data: wage

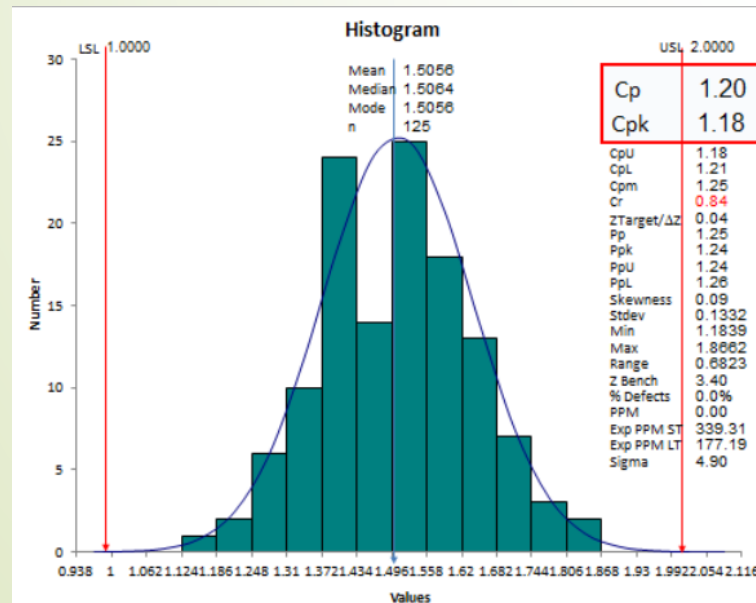
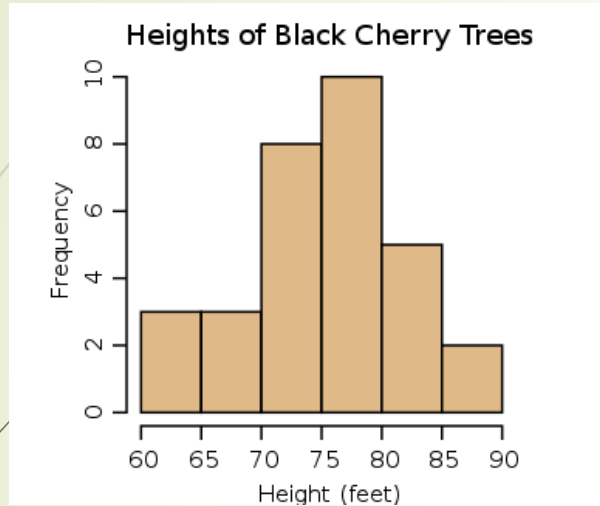
W = 0.89058, p-value < 2.2e-16

非常態分佈

統計圖

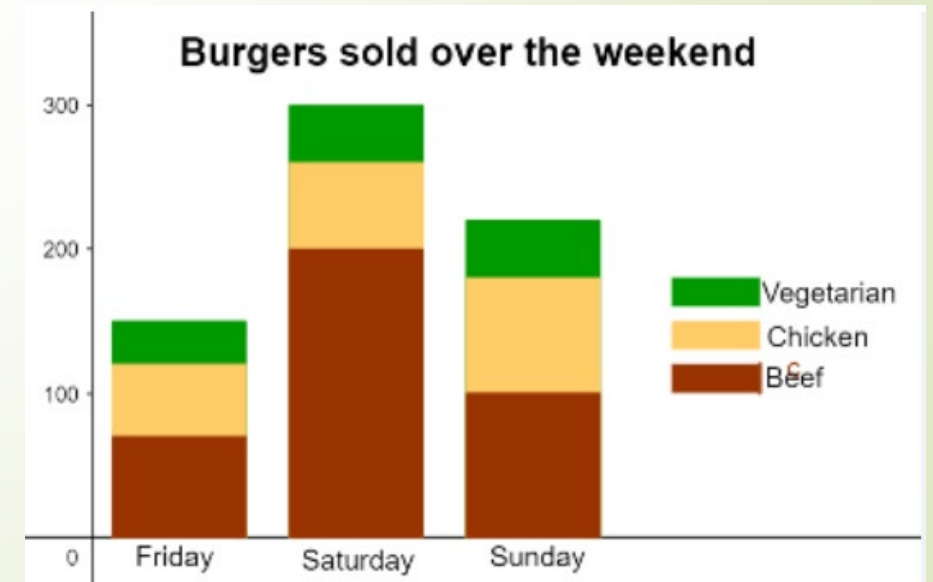
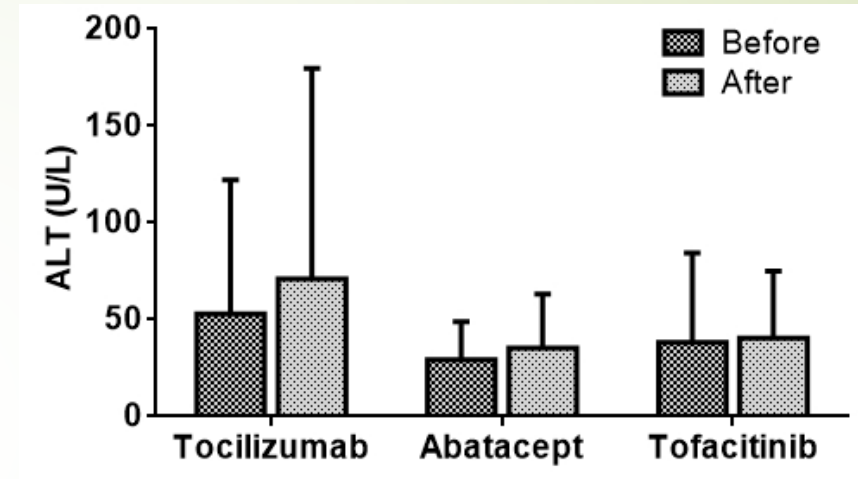
Histogram

直方圖：X axis 為「數值型連續變數」



Bar chart

長條圖：X axis 為「類別型離散變數」



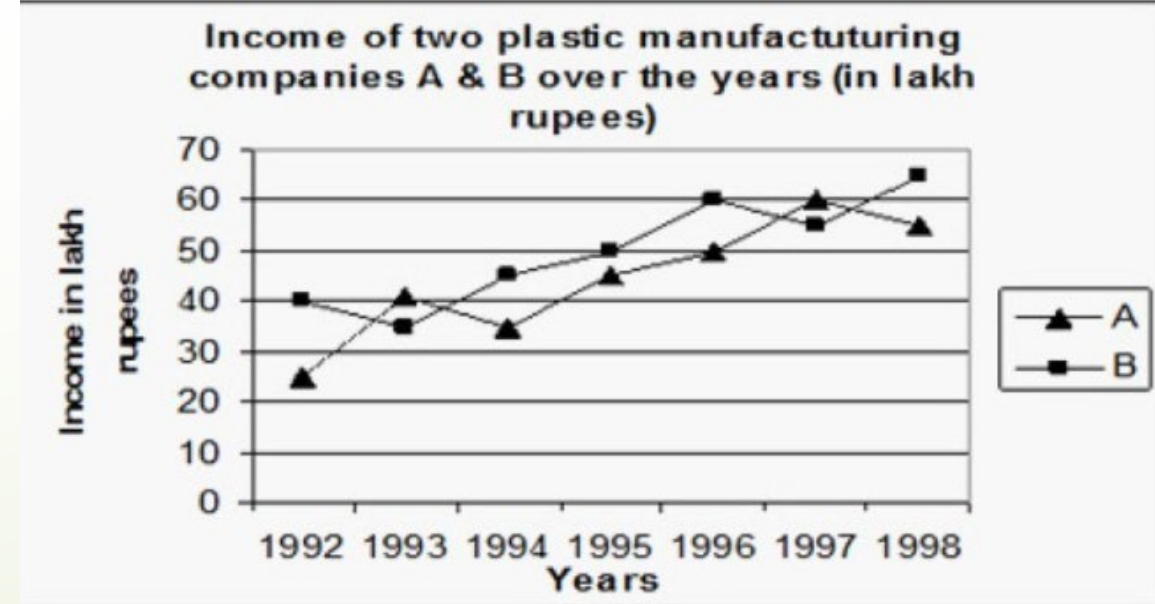
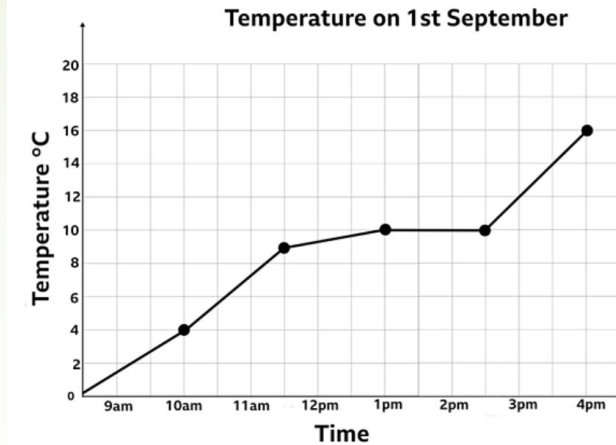
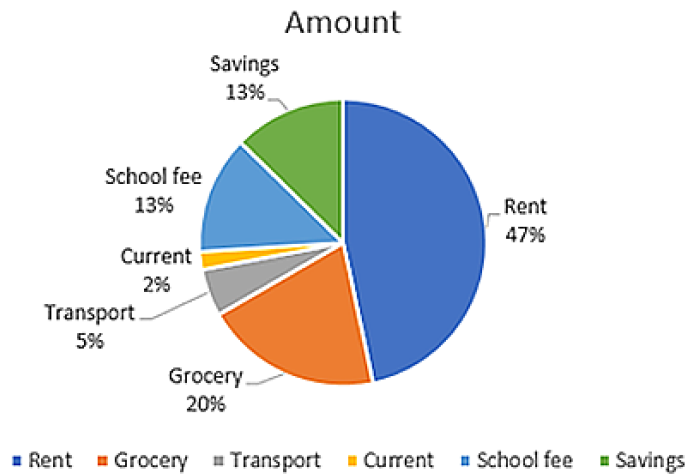
統計圖

Line chart

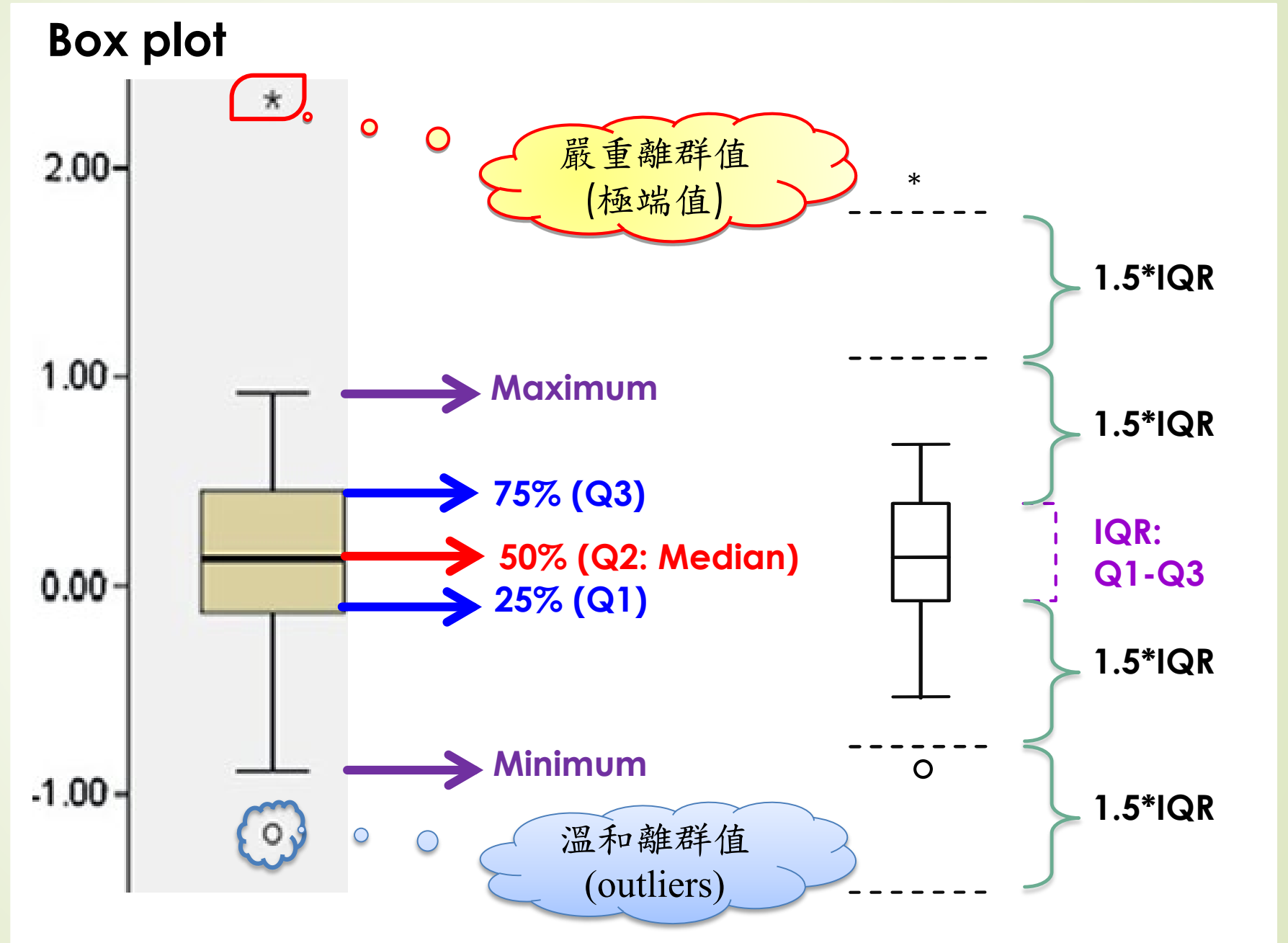
Pie chart

Pie Chart Examples

1	Expenses	Amount
2	Rent	7000
3	Grocery	3000
4	Transport	800
5	Current	300
6	School fee	2000
7	Savings	1900

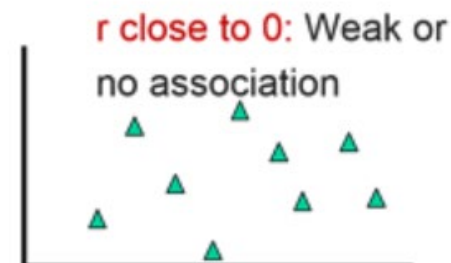
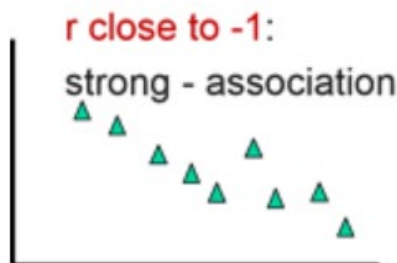
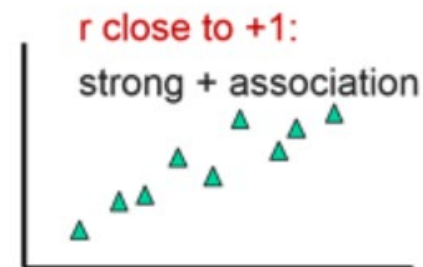
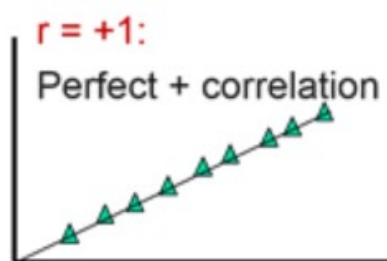
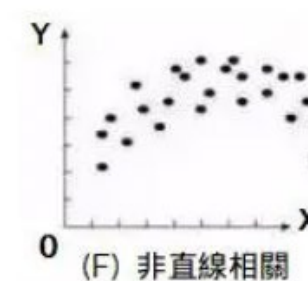
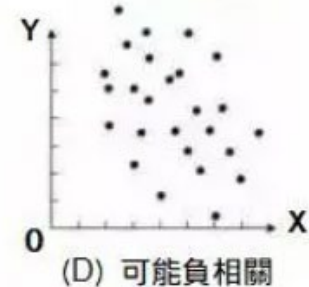
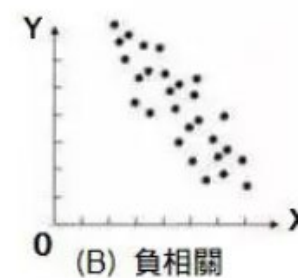
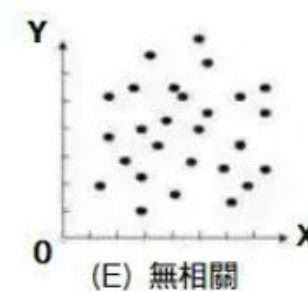
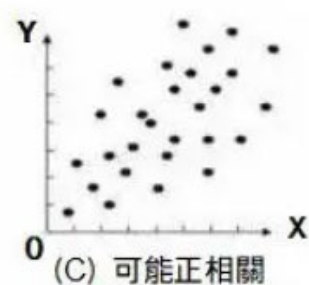
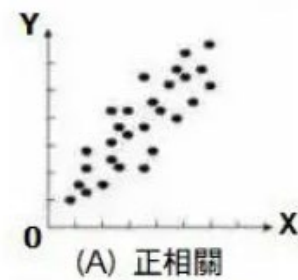
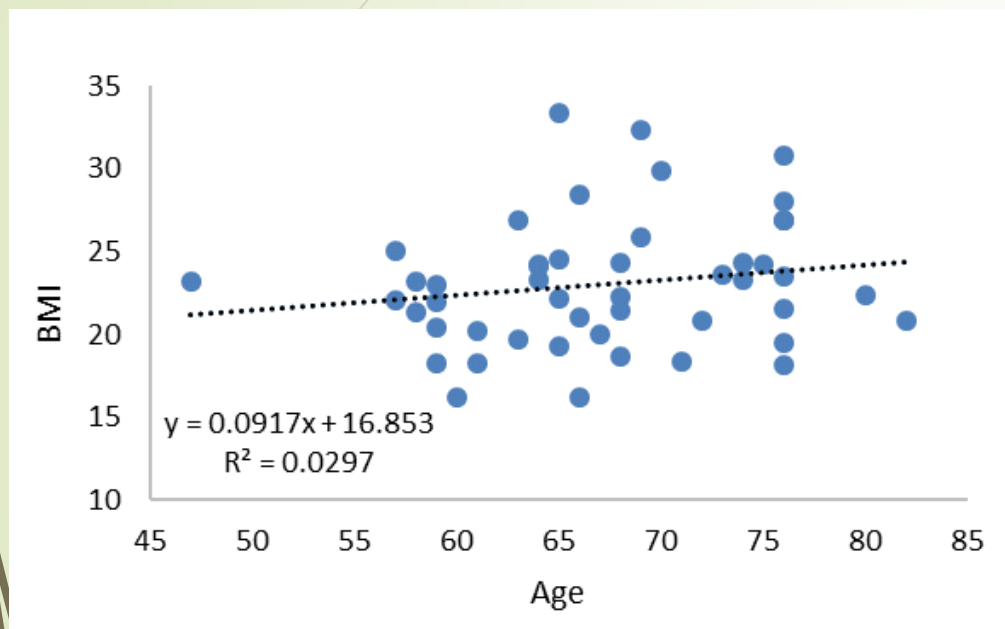


統計圖



統計圖

Scatter plot → Linear regression



資料建檔

欄

變項數值

列

變數名稱

譯碼簿 (編碼簿 coding book)

ID	性別	年齡	婚姻	教育	體重	身高	抽菸	喝酒	日期
001	M	57	已婚	高中	89	186	無	無	2013/4/1
002	F	46	已婚	大專以上	57	160	有	無	2012/6/28
003	F	55	已婚	國中	90	170	無	無	2013/1/21
004	M	65	已婚	大專以上	60	160	有	有	2014/6/20
005	M	60	已婚	國中	58	162	有	有	2013/9/18

資料建檔-譯碼簿

性別	0	F	婚姻	1	已婚	教育	1	小學	抽菸	0	無
	1	M		2	單身		2	國中		1	有
				3	離婚		3	高中	喝酒	0	無
							4	大專以上		1	有

文字型態
改為
數值型態

ID	性別	年齡	婚姻	教育	體重	身高	抽菸	喝酒	日期
001	M	57	已婚	高中	89	186	無	無	2013/4/1
002	F	46	已婚	大專以上	57	160	有	無	2012/6/28
003	F	55	已婚	國中	90	170	無	無	2013/1/21
004	M	65	已婚	大專以上	60	160	有	有	2014/6/20
005	M	60	已婚	國中	58	162	有	有	2013/9/18

ID	性別	年齡	婚姻	教育	體重	身高	抽菸	喝酒	日期
001	1	57	1	3	89	186	0	0	2013/4/1
002	0	46	1	4	57	160	1	0	2012/6/28
003	0	55	1	2	90	170	0	0	2013/1/21
004	1	65	1	4	60	160	1	1	2014/6/20
005	1	60	1	2	58	162	1	1	2013/9/18

轉碼

資料建檔-複選題

請問有無以下疾病？

☐ 1.高血壓

☐ 2.高血脂

☐ 3.心臟病

☐ 4.糖尿病

☐ 5.腎炎、腎徵候群及腎性病變

☐ 6.慢性阻塞性肺病

☐ 7.慢性肝病及肝硬化

☐ 8.其他_____

一個變項數值
一個概念

ID	疾病
001	
002	1,4
003	2,4
004	1,2,4
005	3
006	1



高血壓	高血脂	心臟病	糖尿病
0	0	0	0
1	0	0	1
0	1	0	1
1	1	0	1
0	0	1	0
1	0	0	0

資料建檔-日期格式

ID	日期1	日期2	日期3	日期4	日期5
001	2013/4/1	2015/12/10	20130401	1020401	102/4/1
002	2012/6/28	2015/11/27	20120628	1010628	101/6/28
003	2013/1/21	2015/12/12	20130121	1020121	102/1/21
004	2014/6/20	2015/12/8	20140620	1030620	103/6/20
005	2013/9/18	2015/12/3	20130918	1020918	102/9/18

SUM							=C2-B2	
	A	B	C	D	E	F	G	
1	ID	日期1	日期2	日期3	日期4	日期5	日期2-日期1	
2	001	2013/4/1	2015/12/10	20130401	1020401	102/4/1	=C2-B2	
3	002	2012/6/28	2015/11/27	20120628	1010628	101/6/28	1247	
4	003	2013/1/21	2015/12/12	20130121	1020121	102/1/21	1055	
5	004	2014/6/20	2015/12/8	20140620	1030620	103/6/20	536	
6	005	2013/9/18	2015/12/3	20130918	1020918	102/9/18	806	

日期格式可直接計算
時間差(天數)

資料建檔

- ➡ 簡單且明確建檔
- ➡ 不同檔案各別建立
 - 基本資料
 - 追蹤資料
- ➡ 關鍵字(如病歷號)合併資料

ID	Training mode	Age	BMI	CAT score (Before)	mMRC score (Before)
N01	0	57	25	23	3
N02	0	67	20	27	2
N03	0	63	20	16	3
N04	0	76	24	19	3
N05	0	57	22	19	3
N06	1	66	16	28	3
N07	1	65	25	18	1
N08	1	64	23	12	1

ID	CAT score (After)	mMRC score (After)
N01	10	1
N02	10	1
N03	13	1
N04	4	1
N05	16	3
N06	13	1
N07	16	2
N08	20	2

ID	Training mode	Age	BMI	CAT score (Before)	mMRC score (Before)	CAT score (After)	mMRC score (After)
N01	0	57	25	23	3	10	1
N02	0	67	20	27	2	10	1
N03	0	63	20	16	3	13	1
N04	0	76	24	19	3	4	1
N05	0	57	22	19	3	16	3
N06	1	66	16	28	3	13	1
N07	1	65	25	18	1	16	2
N08	1	64	23	12	1	20	2

Data Type/Statistical Method

Data type	Two groups		≥ 2 groups	Correlation
	Independent	Paired	Independent	Continuous
Numerical				
-Parametric 有母數 需檢定樣本母體 的假設	Independent (Students') t-test	Paired t-test	ANOVA	Pearson correlation
-Non-parametric 無母數 不需檢定樣本母 體的假設	Mann-Whitney U test	Wilcoxon signed-rank test	Kruskal-Wallis test	Spearman rank correlation
Categorical	Chi-square test/ Fisher's exact test	McNemar test	Chi-square test	

Independent (Student's) T-test

- Independent variable : Male and Female (2 groups)
- Dependent variable : BMI 、 Height and Weight.... (Continuous)

Table 1
Demographic data of study participants

	Total (n=29164)	Male (n=12331)	Female (n=16833)	p value
	n %	n %	n %	
BMI (n=25747)	25.20±45.15	24.57±3.98	25.67±59.33	0.041*
Height (cm)	156.72±8.79	163.40±7.03	151.82±6.39	<0.001**
Weight (kg)	60.70±10.64	65.60±10.25	57.11±9.41	<0.001**
Waist circumference (cm)	84.16±9.72	86.99±9.24	82.09±9.54	<0.001**
SBP (mmHg)	134.77±19.23	134.59±18.95	134.90±19.44	0.178
DBP (mmHg)	77.78±11.62	78.30±11.68	77.40±11.56	<0.001**
Pulse pressure (mmHg)	56.99±14.61	56.29±14.33	57.49±14.78	<0.001**

Paired T-test

- Independent variables : **2014 (前)** and **2016 (後)** (**Paired** , 相關樣本)
- Dependent variables : Age 、 BMI 、 Height and Weight....(Continuous)

TABLE 1 | Demographic data of study participants in 2014 and 2016.

	2014 (n = 4,537)		2016 (n = 4,537)		p-value
	n	%	n	%	
Age	71.75	±5.93	73.75	±5.93	–
Gender					1.000
Male	2,207	(48.6%)	2,207	(48.6%)	
Female	2,330	(51.4%)	2,330	(51.4%)	
Smoking	309	(6.8%)	297	(6.5%)	0.119
Drinking	547	(12.1%)	581	(12.8%)	0.862
Exercise	3,237	(71.3%)	3,462	(76.3%)	0.930
BMI	24.49	±3.47	24.41	±3.42	0.004**
Height (cm)	157.67	±8.15	157.59	±8.12	0.063
Weight (kg)	60.99	±10.38	60.78	±10.47	<0.001**
Waist (cm)	83.20	±9.72	84.59	±9.66	<0.001**
SBP (mmHg)	133.91	±18.19	134.63	±18.52	0.007**
DBP (mmHg)	78.14	±10.91	77.21	±11.26	<0.001**
Pulse pressure (mmHg)	55.77	±13.82	57.42	±13.74	<0.001**
Fasting plasma glucose (mg/dl)	104.02	±21.92	104.89	±25.17	0.009**
Triglycerides (mg/dl)	117.03	±66.71	114.85	±67.83	0.024*
HDL cholesterol (mg/dl)	55.61	±15.66	55.60	±15.78	0.922

ANOVA

- Independent variable : CKD stage (**3 groups ↑**)
- Dependent variable : Lab data, pH 、HCO₃ 、PT and INR...(Continuous)

Table 1. Baseline characteristics of all participants divided according to their stages of CKD.

	CKD Stage												Total (<i>n</i> = 12,271)	<i>p</i> - Value	
	I		II		IIIA		IIIB		IV		V				
Laboratory data of blood															
Hb (<i>n</i> = 12,242)	12.30	±2.45	12.36	±2.49	11.98	±2.58	11.56	±2.67	11.05	±2.75	10.13	±2.72	12.08	±2.57	<0.001
Albumin (<i>n</i> = 6940)	3.42	±0.74	3.44	±0.73	3.31	±0.72	3.18	±0.72	3.00	±0.73	3.07	±0.67	3.35	±0.74	<0.001
Calcium (<i>n</i> = 9455)	7.81	±1.71	8.00	±1.70	8.06	±1.62	7.93	±1.78	7.92	±1.67	7.80	±1.86	7.91	±1.71	<0.001
Uric acid (<i>n</i> = 730)	4.98	±2.09	6.42	±2.60	7.26	±2.11	8.34	±2.94	8.73	±3.07	8.38	±2.61	6.67	±2.81	<0.001
Sodium (<i>n</i> = 12,153)	137.36	±5.37	137.91	±5.37	137.42	±6.19	137.56	±6.79	138.07	±7.56	137.20	±6.72	137.59	±5.80	<0.001
Potassium (<i>n</i> = 12,146)	3.98	±0.64	4.03	±0.65	4.10	±0.75	4.23	±0.87	4.34	±0.90	4.46	±1.02	4.07	±0.72	<0.001

Table 1. Cont.

	CKD Stage												Total (n = 12,271)	p- Value	
	I		II		IIIA		IIIB		IV		V				
pH (n = 9348)	6.89	±0.80	6.82	±0.86	6.82	±0.89	6.92	±0.85	7.01	±0.81	7.27	±0.56	6.89	±0.83	<0.001
HCO ₃ ⁻ (n = 5768)	25.14	±4.44	24.55	±4.82	23.68	±5.13	22.55	±5.54	20.31	±5.77	20.51	±5.89	23.84	±5.22	<0.001
PT (n = 9863)	11.49	±3.21	11.79	±5.41	12.00	±4.70	12.91	±7.26	13.58	±7.50	12.80	±6.77	11.93	±5.04	<0.001
INR (n = 9857)	1.09	±0.24	1.12	±0.49	1.15	±0.46	1.23	±0.67	1.30	±0.74	1.20	±0.49	1.14	±0.45	<0.001
Medications															
NSAID	2788	(57.3%)	1858	(49.8%)	608	(40.7%)	427	(37.8%)	204	(31.6%)	91	(22.7%)	5976	(48.7%)	<0.001
ASPIRIN	561	(11.5%)	749	(20.1%)	391	(26.2%)	337	(29.8%)	201	(31.2%)	130	(32.4%)	2369	(19.3%)	<0.001
COX2i	541	(11.1%)	460	(12.3%)	231	(15.5%)	154	(13.6%)	74	(11.5%)	38	(9.5%)	1498	(12.2%)	<0.001
Aminoglycoside	2285	(46.9%)	1651	(44.3%)	590	(39.5%)	407	(36.0%)	218	(33.8%)	114	(28.4%)	5265	(42.9%)	<0.001
Loop diuretics	2230	(45.8%)	1921	(51.5%)	946	(63.3%)	816	(72.1%)	497	(77.1%)	251	(62.6%)	6661	(54.3%)	<0.001

Chi-square test/Fisher's exact test

兩組類別變數的比例分佈是否有差異

觀察值(O)		期望值(E)	
		有病	沒病
抽菸	有病	A	B
	沒病	C	D
		A+B	C+D
不菸	有病	$\frac{(A+B)(A+C)}{N}$	$\frac{(A+B)(B+D)}{N}$
	沒病	$\frac{(A+C)(C+D)}{N}$	$\frac{(B+D)(C+D)}{N}$
		A+C	B+D
		N	

- 每一細格觀察值(O)偏離期望值(E)的程度= $(O-E)/E$, $\chi^2 = \sum (O-E)^2/E$
- 80%以上細格期望值 <5
 - 2 by 2 contingency table 列聯表: Fisher's Exact test

Fisher's exact test with Monte Carlo Simulation
(n by m contingency table, $n > 2$, $m > 2$)

Chi-square test/Fisher's exact test

- Independent variable : CKD stage (Nominal)
- Dependent variable : Heart condition 、 Medications.... (Nominal)

TABLE 1 | Baseline characteristics of older patients with different staging of CKD.

Characteristics	CKD stage 3A (n = 147)	CKD stage 3B (n = 101)	CKD stage 4 (n = 52)	CKD stage 5 (n = 31)	p-value
Age, years	83.1 (77.1–86.3)	81.8 (76.0–86.7)	83.2 (79.0–87.1)	78.8 (73.3–83.6)	0.043
Male	101 (68.7)	71 (70.3)	39 (75.0)	21 (67.7)	0.846
Heart condition					0.001
Non-heart failure	119 (81.0)	84 (83.2)	36 (69.2)	15 (48.4)	
HFpEF	15 (10.2)	12 (11.9)	10 (19.2)	8 (25.8)	
HFrEF	13 (8.8)	5 (4.9)	6 (11.5)	8 (25.8)	
Medications					
Diuretics	108 (73.5)	72 (71.3)	47 (90.4)	24 (77.4)	0.054
MRA	31 (21.1)	16 (15.8)	18 (34.6)	15 (48.4)	0.001
β-blocker	88 (59.9)	60 (59.4)	41 (78.9)	20 (64.5)	0.078
ACEI or ARB	110 (74.8)	77 (76.2)	43 (82.7)	19 (61.3)	0.183
Anti-platelet agents	95 (64.6)	66 (65.4)	44 (84.6)	21 (67.7)	0.052
Anti-coagulants	24 (16.3)	13 (12.9)	19 (36.5)	6 (19.4)	0.003

Example: Dementia in Heart Failure

Values are mean $\pm$ SD or (n (%)).

^aComparison for the no dementia and dementia subgroups.

^bComparison for the 3 dementia types.

Mean $\pm$ SD: Student's t-test

N (%): Chi-square test

Mean $\pm$ SD: One-way ANOVA

N (%): Chi-square test

TABLE 1 Baseline Characteristics of All Patients and by Dementia Classifications

	Overall (N = 202,121)	No Dementia (n = 179,976)	Dementia (n = 22,145)	P Value ^a	Alzheimer's Disease (n = 5,936)	Vascular Dementia (n = 4,008)	Unspecified Dementia (n = 12,201)	P Value ^b
Age, y	75.3 $\pm$ 13.0	74.7 $\pm$ 13.3	80.7 $\pm$ 8.6	<0.001	81.2 $\pm$ 8.3	78.5 $\pm$ 0.0	81.1 $\pm$ 8.5	<0.001
Male	98,361 (48.7)	90,378 (50.2)	7,983 (36.0)	<0.001	1,855 (31.2)	1,640 (40.9)	4,488 (36.8)	<0.001
Smoking	44,928 (22.2)	40,361 (22.4)	4,567 (20.6)	<0.001	1,192 (20.1)	794 (19.8)	2,581 (21.2)	<0.001
Alcohol	5,083 (2.5)	4,561 (2.5)	522 (2.4)	0.118	147 (2.5)	105 (2.6)	270 (2.2)	0.170
Comorbidities								
Hypertension	90,932 (45.0)	79,490 (44.2)	11,442 (51.7)	<0.001	2,877 (48.5)	2,300 (57.4)	6,265 (51.3)	<0.001
Diabetes	55,696 (27.6)	49,675 (27.6)	6,021 (27.2)	0.198	1,461 (24.6)	1,249 (31.2)	3,311 (27.1)	<0.001
Obesity	1,937 (1.0)	1,787 (1.0)	150 (0.7)	<0.001	27 (0.5)	35 (0.9)	88 (0.7)	<0.001
Stroke	24,247 (12.0)	20,629 (11.5)	3,618 (16.3)	<0.001	664 (11.2)	1,149 (28.7)	1,805 (14.8)	<0.001
Anemia	28,314 (14.0)	24,794 (13.8)	3,520 (15.9)	<0.001	1,038 (17.5)	643 (16.0)	1,839 (15.1)	<0.001
Coronary artery disease	68,984 (34.1)	61,350 (34.1)	7,634 (34.5)	0.257	1,929 (32.5)	1,457 (36.4)	4,248 (34.8)	<0.001
Peripheral vascular disease	26,694 (13.2)	22,469 (12.5)	4,225 (19.1)	<0.001	743 (12.5)	1,253 (31.3)	2,229 (18.3)	<0.001
Dyslipidemia	22,266 (11.0)	19,939 (11.1)	2,327 (10.5)	0.011	515 (8.7)	534 (13.3)	1,278 (10.5)	<0.001
Atrial fibrillation	50,007 (24.7)	43,817 (24.3)	6,190 (28.0)	<0.001	1,484 (25.0)	1,269 (31.7)	3,437 (28.2)	<0.001
Other arrhythmia	12,578 (6.2)	11,093 (6.2)	1,485 (6.7)	0.002	411 (6.9)	270 (6.7)	804 (6.6)	0.013
Cirrhosis	2,104 (1.0)	1,947 (1.1)	157 (0.7)	<0.001	39 (0.7)	35 (0.9)	83 (0.7)	<0.001
Chronic renal failure	21,983 (10.9)	20,037 (11.1)	1,946 (8.8)	<0.001	447 (7.5)	378 (9.4)	1,121 (9.2)	<0.001
Other chronic kidney diseases	34,837 (17.2)	31,936 (17.7)	2,901 (13.1)	<0.001	686 (11.6)	593 (14.8)	1,622 (13.3)	<0.001

型一誤差 v.s 型二誤差

Type		
Type I Error	虛無假說事實上成立， 但計算出來的檢定統計量落在拒絕區域， 使其拒絕虛無假設	假設： 原本wafer size是59奈米 (nm)， 樣本標準差為1，在正常製程底下wafer size 將服從 ($\mu=59, \sigma=1$) 的常態機率分布
Type II Error	虛無假設事實上不成立，但計算 出來的的檢定統計量沒有落在拒絕區域，也就是落在接受區域， 使其不拒絕虛無假設	顯著水準alpha 5% 就是型一錯誤 (Type I Error) 的機率， 因為在製程正常的機率分布底下，還是 “有可能” 會落在拒絕區域，只是機率很低

多重比較 (post hoc)

	Methods	方法比較		
各組樣本數相同	Bonferroni	● 利用原先的顯著水準α除以檢定次數	● Type I error 過度保守 ● 檢定力偏低 ● 較不易顯著	● 事前處理 ● Any contrasts
	Fisher's LSD: Least Squares Difference	● 最原始，不多做修正 ● 利用整體顯著水準α與檢定的P-value作比較	● 適合檢定次數少 ● 最容易顯著 ● Type I error $\uparrow$	● 事後處理 ● Pairwise
	Tukey's Honestly Significant Difference (HSD)	● 測試所有可能的pairwise的平均值，並決定是否其中之一為0 ● 需符合常態分佈	● 只適合兩兩比較 (只想測試某兩組時)	● 事後處理 ● Pairwise ● Any contrasts (可，但是poor)
各組樣本數不同	Scheffe	● 同時測試所有可能的contrast，並視是否其中之一為0 ● 可用於非常態分佈	● 臨界值最大，最不容易顯著 ● 較不容易犯 Type II error ● 想測試所有差異時	● 事後處理 ● Any contrasts

McNemar's test

■ 兩相依樣本

- 僅適用於 2×2 表格
- 課程前後，認知問答對與錯是否有差異

■ McNemar's test注意事項

- 非一致觀察值加總 < 25
- binomial test

課程前後，認知問答對與錯是否有差異？

	Q1_1 (n=50)		Q2_2 (n=50)		p value
	n	%	n	%	
對	30	60%	20	40%	0.110
錯	20	40%	30	60%	

McNemar test. * $p < 0.05$, ** $p < 0.01$

ID	Q1_1	Q1_2
01	0	0
02	0	0
03	0	1
...	...	...
...	...	...
48	0	0
49	0	1
50	0	0

Correlation

➤ Pearson correlation

- Normal distribution
- Continuous

➤ Spearman rank correlation

- Skewed distribution
- Ordinal/ Continuous

Table 4

Correlations of ANA, anti-dsDNA antibody, C3, C4 and subcomponents of PANSS and UKU in schizophrenia patients with SLE or autoimmune features.

	ANA titer		Anti-dsDNA antibody		C3		C4	
	r_s	p	r_s	p	r_s	p	r_s	p
N4 (social withdraw)	0.108	0.402	-0.010	0.940	-0.258	0.047*	-0.148	0.257
P3 (hallucinatory behavior)	-0.031	0.811	-0.041	0.754	-0.213	0.103	-0.285	0.027*
G9 (unusual thought content)	0.160	0.213	-0.172	0.181	-0.166	0.204	-0.296	0.022*
P1 (delusion)	0.097	0.452	-0.019	0.885	-0.271	0.036*	-0.158	0.229
N6 (lack of spontaneity)	-0.051	0.693	-0.007	0.958	-0.318	0.013*	-0.005	0.969
N1 (blunted affect)	0.046	0.721	-0.104	0.422	-0.327	0.011*	-0.036	0.783
G5 (mannerisms/posturing)	0.016	0.899	-0.093	0.472	-0.189	0.148	-0.063	0.635
P2 (conceptual disorganization)	0.387	0.002**	-0.011	0.933	-0.225	0.083	-0.200	0.125
UKU (increased salivation)	0.187	0.146	0.055	0.671	-0.325	0.011*	-0.162	0.217

Spearman's rho coefficient.

PANSS: Positive and Negative Syndrome Scale.

UKU: Udvalg for Kliniske Undersøgelser rating scale.

* $p < 0.05$.

** $p < 0.01$.

說明：(相關係數數值範圍為-1~1)

相關係數	相關程度
1	完全相關
0.7~0.99	高度相關
0.5~0.69	中度相關
0.25~0.49	低度相關
0~0.24	無相關

Logistic regression

- 主要在探討以自變項 (Independent variable) 的變化來預測或解釋結果變項 (Dependent variable) 的變化。
- 目的要建立變數間的因果關係，以利預測結果
- 依變項為二元分類資料 (0 or 1) → 預測 1 的機率
 - 年齡與有無特定慢性病的關係
 - 吸菸與有無肺癌的關係
 - 嚼檳榔與有無口腔癌的關係
 - 醫師年資與手術成功與否的關係
- 自變項為類別變數或連續變數

Logistic regression

- 單變項迴歸表示式

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 X$$

- 多變項迴歸表示式

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 X + \beta_2 X_2 + \dots + \beta_k X_k$$

- 自變數 (X) 對依變數 (Y) 的影響是以指數的方式做變動，因此不需要常態分配的假設
- 使用「最大概似函數估計法 (Maximum Likelihood Estimation)」做參數估計

Odds ratio (OR)

Exposure	Outcome	
	Lung ca	Control
Smoking	a	b
Non Smoking	c	d

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Odds ratio	意義
OR=1	抽菸與肺癌無關
OR>1	抽菸可能為肺癌的危險因子
OR<1	抽菸可能為肺癌的保護因子

Logistic regression example

- 經由ELISA 和CIA探討dsDNA抗體檢測不一致的風險因素

Table 5. Logistic regression of risk factors for inconsistency of anti-dsDNA antibody tests by ELISA and CIA.

	Univariate			Multivariable		
	Odds Ratio	95%CI	p-Value	Odds Ratio	95%CI	p-Value
Age	1.03	(1.01–1.05)	0.010 *			
Female	Reference					
Male	1.75	(0.72–4.22)	0.215			
Creatine < 1.4 (mg/dL)	Reference					
Creatine ≥ 1.4 (mg/dL)	1.67	(0.81–3.47)	0.165			
UPCR < 500 (mg/g)	Reference					
UPCR ≥ 500 (mg/g)	0.66	(0.37–1.17)	0.157			
ANA < 1:80	Reference					
ANA 1:80–1:640	1.47	(0.30–7.14)	0.632			
ANA ≥ 1:640	0.76	(0.16–3.58)	0.724			
ANA Homogeneous (n = 281)	0.59	(0.26–1.34)	0.206			
C3 ≥ 87 (mg/dL)	Reference			Reference		
C3 < 87 (mg/dL)	0.40	(0.24–0.67)	0.001 **	0.93	(0.40–2.13)	0.861
C4 ≥ 19 (mg/dL)	Reference			Reference		
C4 < 19 (mg/dL)	0.45	(0.27–0.77)	0.003 **	1.04	(0.43–2.51)	0.930
CIC < 10.8 (μg Eq/mL)	Reference			Reference		
CIC ≥ 10.8 (μg Eq/mL)	0.44	(0.24–0.79)	0.006 **	0.42	(0.18–0.94)	0.036 *
Anti-dsDNA antibody by CIA (IU/mL)	0.98	(0.98–0.99)	<0.001 **			
Anti-dsDNA antibody by ELISA (WHO units/mL)	0.98	(0.98–0.99)	<0.001 **	0.98	(0.98–0.99)	<0.001 **
SLEDAI < 4	Reference			Reference		
SLEDAI ≥ 4	0.22	(0.12–0.39)	<0.001 **	0.33	(0.14–0.79)	0.013 *

Logistic regression. * $p < 0.05$, ** $p < 0.01$. Anti-dsDNA antibodies: anti-double-stranded DNA antibodies; UPCR: urine protein/creatinine ratio; ANA: anti-nuclear antibodies; CIC: anti-C1q circulating immune complexes antibody; CIA: chemiluminescent immunoassay; ELISA: enzyme-linked immunosorbent assay; SLEDAI: Systemic Lupus Erythematosus Disease Activity Index.

ROC curve

AUC	Discrimination
0.5	沒有鑑別力
0.7+	可接受的鑑別力
0.8+	好的鑑別力
0.9+	非常好的鑑別力

■ 使用時機

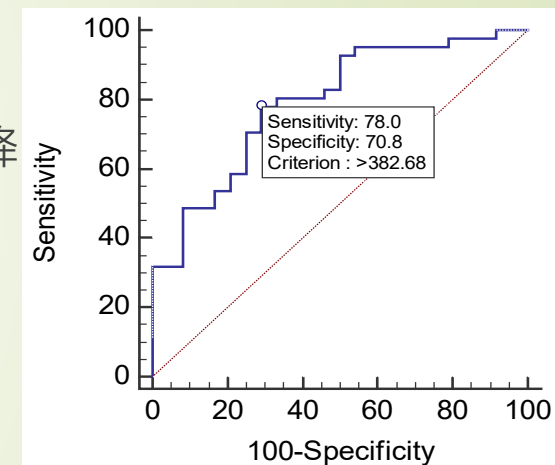
- 當開發新的檢驗方法，無法決定臨界值 (Cut-off value)
- 利用連續數值預測結果 (二元分類)
 - ✓ 比較不同Biomarker (A β 42/t-tau) 預測認知功能有無下降

■ 目的

- 將連續數值決定臨界值
- 用來比較不同工具的好壞

■ 最適切點 (Youden Index = Sensitivity + Specificity - 1)

- Sensitivity + Specificity 最大的點
- 0-1 之間，越接近 1 越好



診斷工具指標

	Gold Standard Test		Total
	Disease	Control	
Screen test			
Positive	a	b	a+b
Negative	c	d	c+d
Total	a+c	b+d	a+b+c+d

$$\text{Sensitivity} = \frac{a}{a+c}$$

$$\text{Specificity} = \frac{d}{b+d}$$

$$\text{False positive} = \frac{b}{b+d}$$

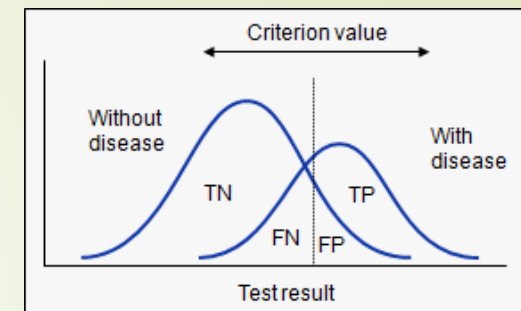
$$\text{False negative} = \frac{c}{a+c}$$

$$\text{Positive Predictive Value} = \frac{a}{a+b}$$

$$\text{Negative Predictive Value} = \frac{d}{c+d}$$

$$\text{Likelihood ratio positive} = \frac{\text{Sensitivity}}{1 - \text{specificity}}$$

$$\text{Likelihood ratio negative} = \frac{1 - \text{sensitivity}}{\text{specificity}}$$



Ref: <https://www.medcalc.org/manual/roc-curves.php>

概似比的臨床意義

Likelihood ratio	Interpretation
>10	Strong evidence to rule in disease
5~10	Moderate evidence to rule in disease
2~5	Weak evidence to rule in disease
0.5~2.0	No signification change in the likelihood of disease
0.2~0.5	Weak evidence to rule out disease
0.1~0.2	Moderate evidence to rule out disease
<0.1	Strong evidence to rule out disease

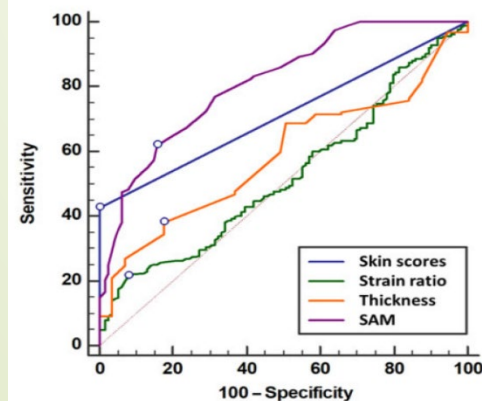
ROC example

► 利用高光譜影像評估皮膚狀況是否預測硬皮症

TABLE 3 ROC analysis of skin assessments in the diagnosis of scleroderma

Variables	AUC (95% CI)	P	Optimal cutoff	Sensitivity, %	Specificity, %	Accuracy, %	PPV, %	NPV, %
Skin scores ^a	0.712 (0.658, 0.763)	<0.001	>0	42.47	100.00	64.33	100.00	51.58
Strain ratio	0.522 (0.464, 0.580)	0.510	≤0.88	21.51	92.11	48.33	81.63	41.83
Thickness	0.585 (0.527, 0.642)	0.009	>1.31	38.17	82.46	55.00	78.02	44.98
SAM	0.812 (0.763, 0.854)	<0.001	>0.025	61.83	84.21	70.33	86.47	57.49

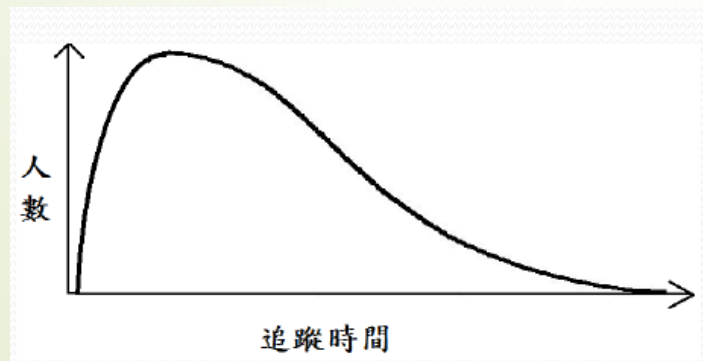
FIG. 3 Comparisons of scleroderma diagnostic accuracy among tests



Rheumatology 59.11 (2020): 3201-3210.

Survival analysis

- 將『時間』變項列入分析的統計方法
- 從一時間點至事件(event)發生的時間(time to event)，稱為survival time
 - Start time- 研究起始時間/確診癌症時間
 - End time- 研究結束時間/疾病發生時間/死亡時間
- 資料特性：資料通常不是常態分配

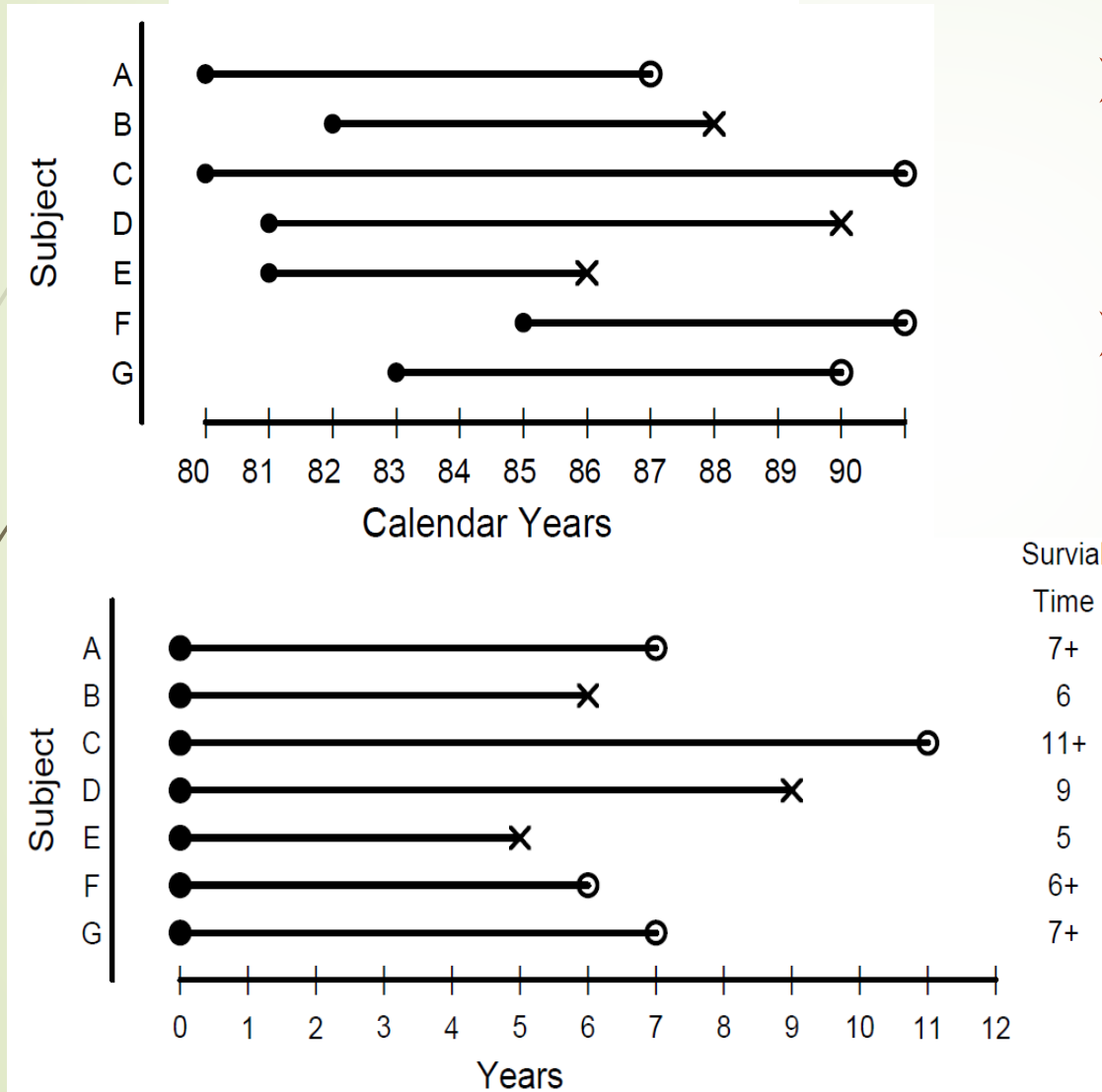


資料分類

- 完整資料(complete data)
 - 觀察期間提供事件發生的時間點
- 設限資料(censored data)
 - 觀察期間失去聯絡或觀察結束時仍未發生事件
- 競爭死因資料(Competing cause of deaths)
 - 觀察期間死於其他死因之時間

Survival time

(● initial follow-up, × death, ○ alive.)



- Survival time= 事件發生時間 – 追蹤的起始時間 (需大於0)
- 個案BDE是完整資料；
個案ACFG為設限資料

存活資料建檔

基本資料

存活狀態

起訖時間

分析整理

ID	Sex	Age	Tx	event (death)	Start_date	Death_date	follow_date	End_date	follow_year
Case1	0	45	0	0	2012/5/8		2016/7/15	2016/7/15	4.19
Case2	1	55	0	1	2013/12/9	2015/1/16		2015/1/16	1.10
Case3	0	56	0	0	2012/5/10		2014/3/12	2014/3/12	1.84
Case4	1	51	1	1	2014/9/11	2016/7/18		2016/7/18	1.85
Case5	0	62	1	1	2012/8/12	2016/12/19		2016/12/19	4.35
Case6	1	70	0	1	2018/7/13	2019/7/20		2019/7/20	1.02
Case7	1	58	0	0	2015/5/14		2016/8/21	2016/8/21	1.27
Case8	0	66	0	0	2018/8/15		2019/7/22	2019/7/22	0.93
Case9	0	60	1	0	2019/5/16		2019/7/23	2019/7/23	0.19
Case10	0	63	1	0	2017/10/17		2018/7/24	2018/7/24	0.77
Case11	0	66	1	0	2016/3/20		2017/7/25	2017/7/25	1.35
Case12	0	69	1	0	2014/8/22		2016/7/26	2016/7/26	1.93

- 日期格式統一西元年
- 可分組比較存活曲線差異，如Sex (F vs M), Age (<60 vs >=60)
- 事件(event): 0=存活;1=死亡
- 未死亡個案End_date，研究者需設定結束時間(建議可用最後看診日期)。

Definition of Time to Event Variables

指標	事件	設限
1.無疾病進展存活期 (progression free survival, PFS)	疾病進展或死亡	無疾病進展
2.無疾病存活期 (disease free survival, DFS)	疾病復發或死亡	無疾病復發
3.治療至疾病進展時間 (time to progression, TTP)	疾病進展	無疾病進展、無疾病狀態死亡(如車禍或不明原因)
4.治療至治療失敗時間 (time to treatment failure, TTF)	疾病進展、疾病復發、藥物副作用、死亡	無疾病進展或復發
5.總存活時間 (overall survival, OS)	死亡	存活
6. 疾病特定存活期 (disease specific survival, DSS)	由疾病本身所導致的死亡(車禍或其他原因要歸為設限資料)	存活、因其他原因死亡

存活分析 - 統計方法

■ 無母數分析

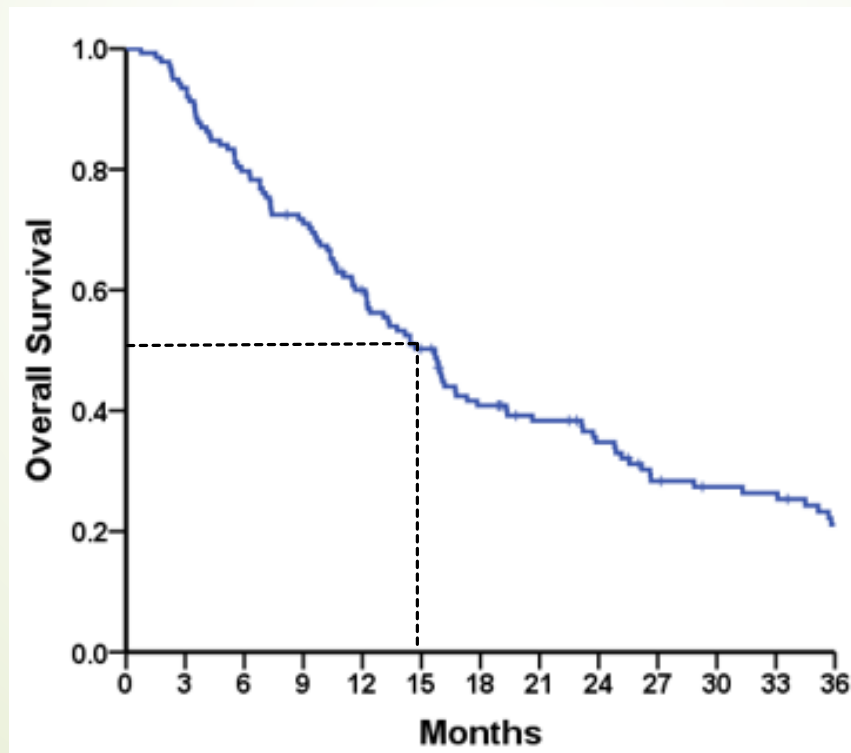
- Kaplan-Meier
- Log-rank test
- Cox Proportional Hazards Model

2.2. Statistical Analysis

The most important censored point of this study was the first episode of UBR. UBR-free survival was defined as the interval between the date of NUx and the date of the first episode of UBR. Survival data was analyzed using both the Kaplan–Meier method and the logrank test. Univariate and multivariate analyses by Cox’s proportional hazards model were used to determine the relevance between each of the clinicopathological factors and UBR. P-values less than 0.05 were defined as statistically significant. All statistical analyses were performed with SPSS (Statistical Package for the Social Sciences, version 22.0, IBM, NY, USA).

Kaplan-Meier

- 常用來估計存活曲線的方法，此方法用每一個事件發生時間點及設限點來設定區間
- 可估計存活中位數及不同時間點的存活率



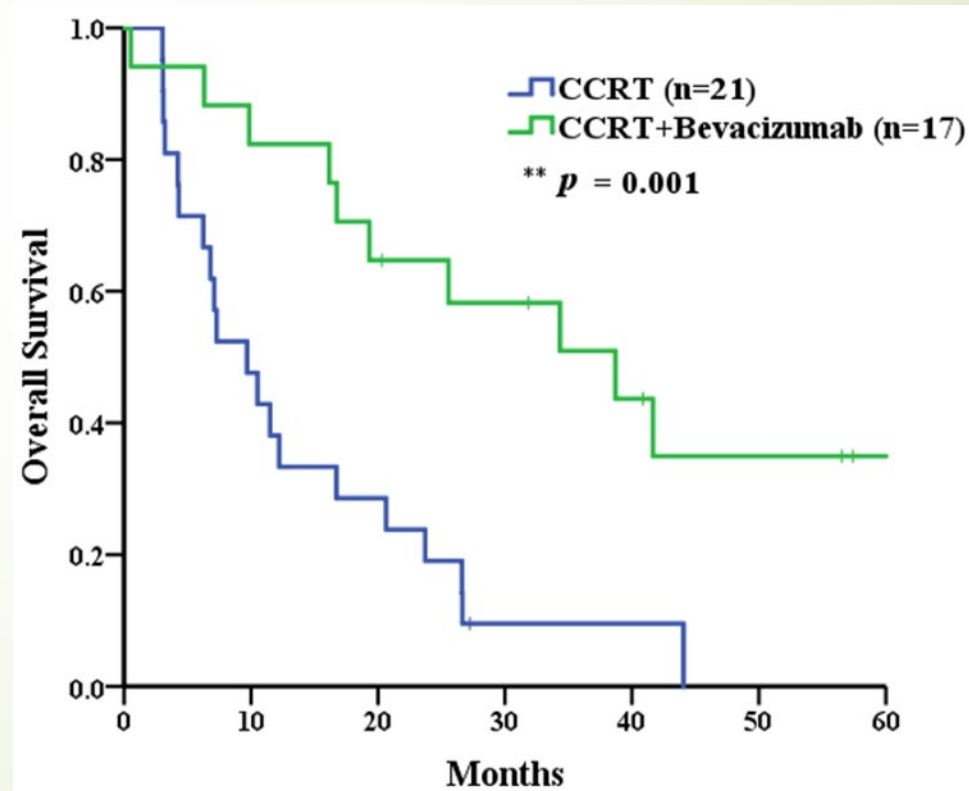
Kaplan-Meier estimates

ID	fu_time (month)	status
1	2	0
2	2	0
3	2	1
4	3	0
5	3	0
6	4	1
7	4	0
8	5	0
9	6	1
10	6	0

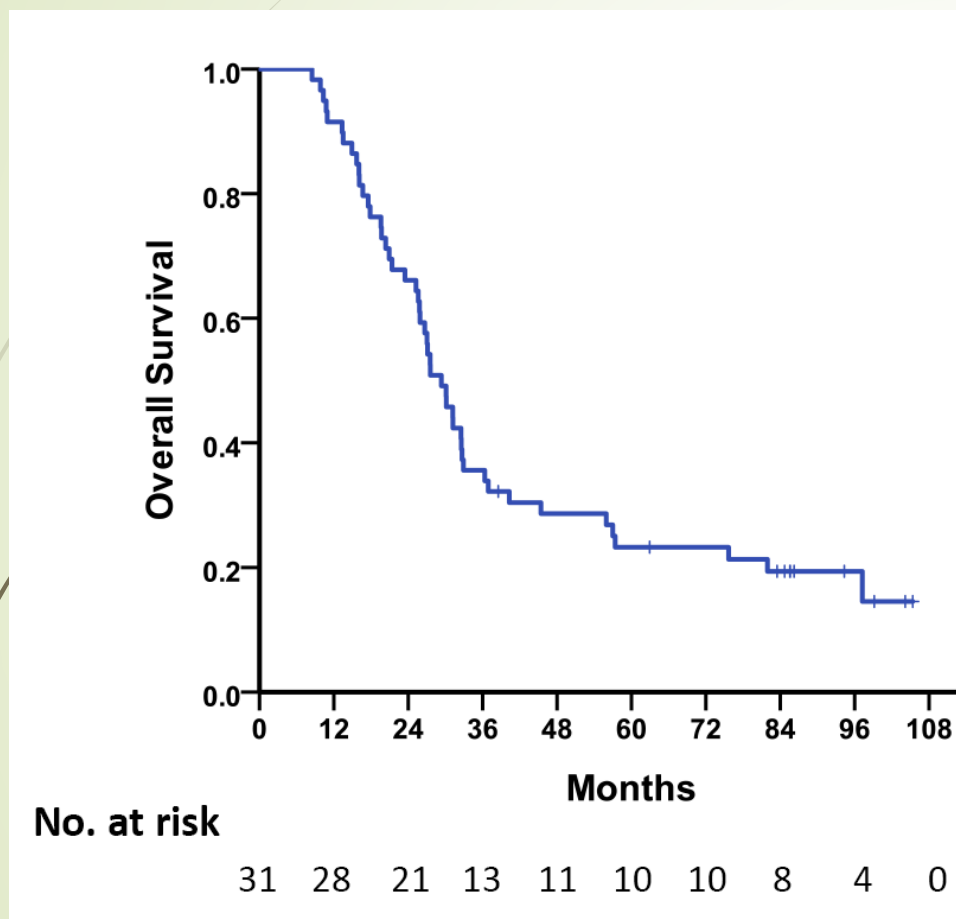
month	Number at risk	Number of events	Number of censored	Conditional Probability	Survival Function
1	10	0	0	$10/10 = 1.00$	1.00
2	10	1	2	$9/10 = 0.90$	$0.90 * 1.00 = 0.90$
3	7	0	2	$7/7 = 1.00$	$1.00 * 0.90 = 0.90$
4	5	1	1	$4/5 = 0.80$	$0.80 * 0.90 = 0.72$
5	3	0	1	$3/3 = 1.00$	$1.00 * 0.72 = 0.72$
6	2	1	1	$1/2 = 0.50$	$0.50 * 0.72 = 0.36$

Log-rank test

- Kaplan-Meier法僅能了解不同組別的存活曲線分佈
- 比較組別間是否差異
 - H0: 兩條存活曲線相同
 - H1: 兩條存活曲線不相同



Survival curve



	Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
			Estimate	Std. Error		
12M	8.467	Death	.968	.032	1	30
	9.833	Death	.935	.044	2	29
	10.767	Death	.903	.053	3	28
4	14.900	Death	.871	.060	4	27
5	15.667	Death	.839	.066	5	26
6	16.667	Death	.806	.071	6	25
7	17.533	Death	.774	.075	7	24
8	17.833	Death	.742	.079	8	23
24M	20.933	Death	.710	.082	9	22
	23.467	Death	.677	.084	10	21
11	25.233	Death	.645	.086	11	20
12	25.800	Death	.613	.087	12	19
13	27.033	Death	.581	.089	13	18
14	27.533	Death	.548	.089	14	17
15	27.567	Death	.516	.090	15	16
16	29.333	Death	.484	.090	16	15
36M	30.133	Death	.452	.089	17	14
	32.500	Death	.419	.089	18	13
48M	40.300	Death	.387	.087	19	12
	45.400	Death	.355	.086	20	11
60M	55.933	Death	.323	.084	21	10
	75.667	Death	.290	.082	22	9
84M	81.933	Death	.258	.079	23	8
	85.567	Survival	.	.	23	7
96M	85.600	Survival	.	.	23	6
	86.233	Survival	.	.	23	5
	94.333	Survival	.	.	23	4
28	97.233	Death	.194	.081	24	3
29	99.167	Survival	.	.	24	2
30	104.167	Survival	.	.	24	1
31	105.367	Survival	.	.	24	0

Cox Proportional Hazards Model

- 評估多個變數對存活(Time to event)的影響
- 可使用類別/連續變數呈現危險因子，並估算出這些危險因子對outcome的影響

$$\log \frac{h(t)}{h_0(t)} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Hazard ratio (HR)	Explanation
HR=1	Two groups have the same survival experience
HR>1	Survival is better in the control group
HR<1	Survival is better in the treatment group

Kaplan-Meier Example

- 比較在第三期子宮內膜癌術後治療，Sandwich組和CT alone兩組治療預後的差異

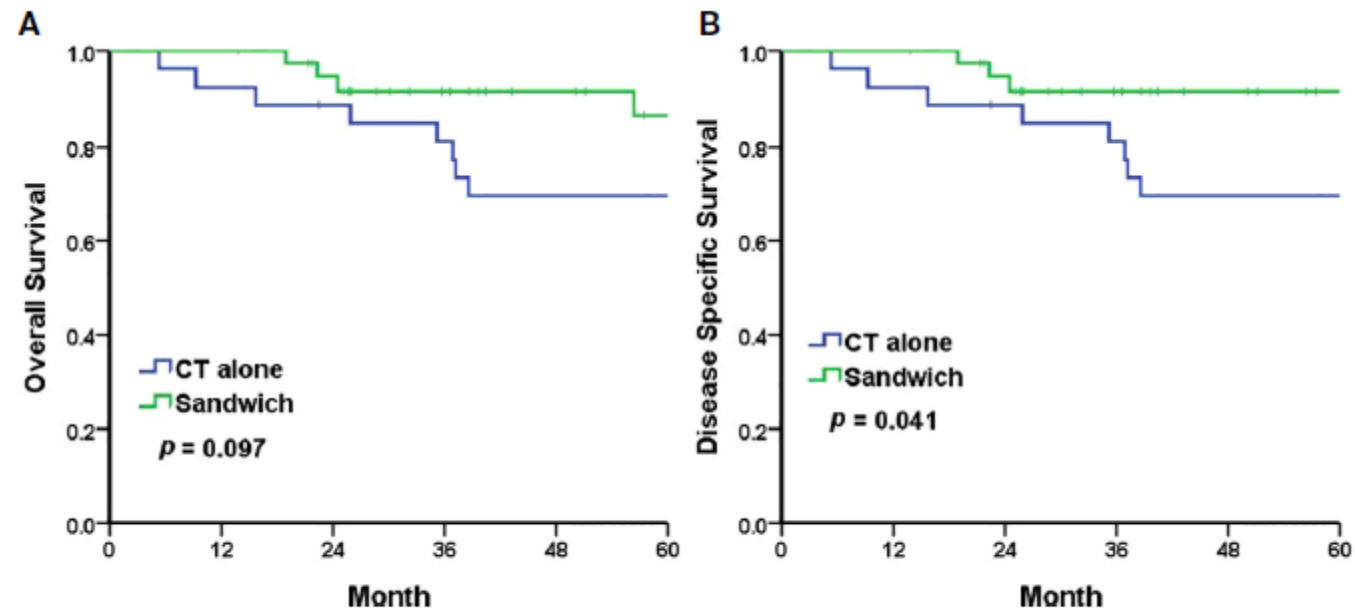


FIGURE 2

The Kaplan–Meier survival curves for 5-year overall survival (A) and 5-year disease-specific survival (B). CT chemotherapy.

Outcomes of "sandwich" chemoradiotherapy compared with chemotherapy alone for the adjuvant treatment of FIGO stage III endometrial cancer. *Frontiers in oncology*, 12 (2022), 946113-946113.

Cox regression Example

5-year disease-specific survival						
	Univariate			Multivariable		
	HR	95% CI	p-value	HR	95% CI	p-value
Age group						
<60	Reference					
≥60	1.34	(0.36-5.07)	0.665			
FIGO stage						
IIIA and IIIB and IIIC1	Reference					
IIIC2	1.45	(0.44-4.75)	0.541			
Histology grading						
Grades 1 and 2	Reference			Reference		
Grade 3	8.70	(1.11-68.01)	0.039*	9.16	(1.17-71.70)	0.035*
Treatment						
CT alone	Reference			Reference		
Sandwich	0.27	(0.07-1.04)	0.056	0.23	(0.06-0.87)	0.030*
LVI						
Absent	Reference					
Present	3.82	(0.49-29.89)	0.201			
Deep myometrial invasion						
Absent	Reference			Reference		
Present	7.85	(1.00-61.34)	0.050	9.44	(1.20-74.15)	0.033*

Outcomes of "sandwich" chemoradiotherapy compared with chemotherapy alone for the adjuvant treatment of FIGO stage III endometrial cancer. *Frontiers in oncology*, 12 (2022) , 946113-946113.

SPSS操作步驟

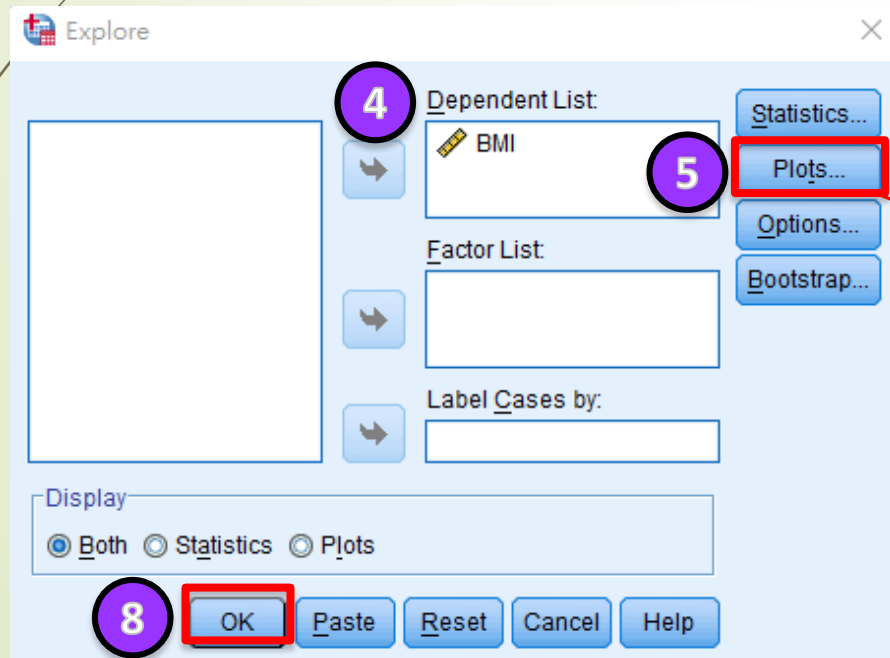
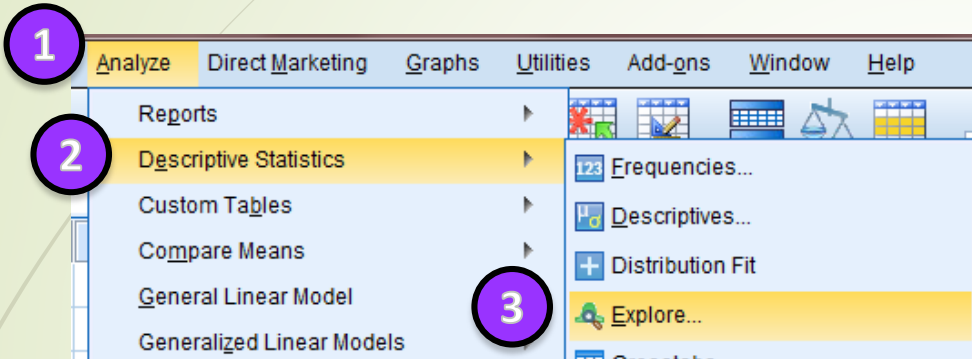
醫學研究部 基礎醫學科 生統小組

陳韻仔 高級助理研究員

授課日期：112年3月29日

Test of Normality

分析 > 描述性統計 > 探索



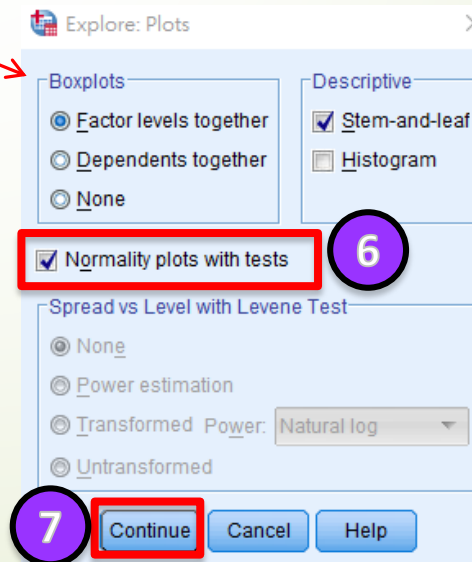
For N > 50 For N < 50

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
BMI	.076	40	.200*	.981	40	.710

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

$p < 0.05$ 代表資料呈非常態分佈



Independent T-test

SPSS dataset

caseno	Sex	SBP	DBP	pluse_pressure	WC	Height	Weight	BMI
1	1	150	83	67	95.50	165.60	76.70	27.97
2	1	135	68	67	99.06	157.00	69.30	28.11
3	1	131	77	54	93.00	163.00	72.50	27.29
4	1	144	93	51	97.00	169.30	76.90	26.83
5	1	149	78	71	91.95	163.50	70.00	26.19
6	1	140	70	70	91.44	174.00	76.00	25.10
7	1	140	71	69	93.00	170.50	77.30	26.59
8	1	110	91	19	92.00	177.40	78.00	24.78
9	1	133	80	53	99.06	158.10	75.40	30.13
10	1	135	91	44	96.00	173.50	79.00	26.24
11	1	130	78	52	91.44	166.00	82.00	29.76
12	1	140	90	50	102.00	156.00	66.00	27.12

Value Labels

Value: Label:

Sex

1 = "Male"
2 = "Female"

Add Change Remove

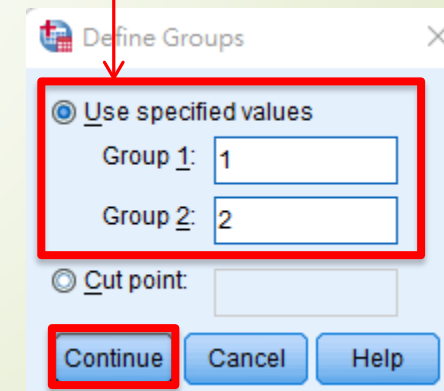
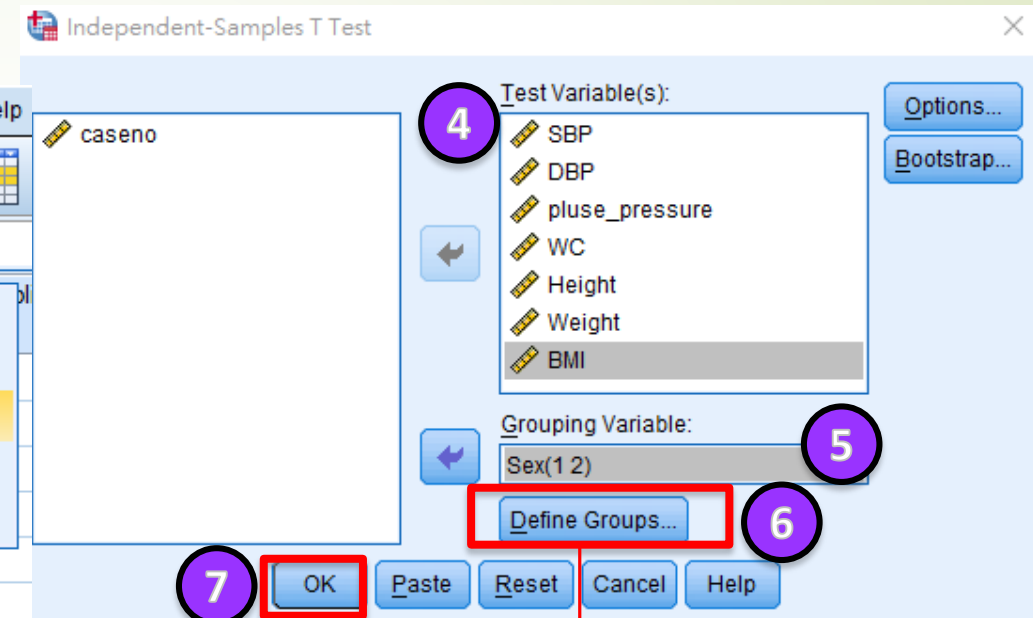
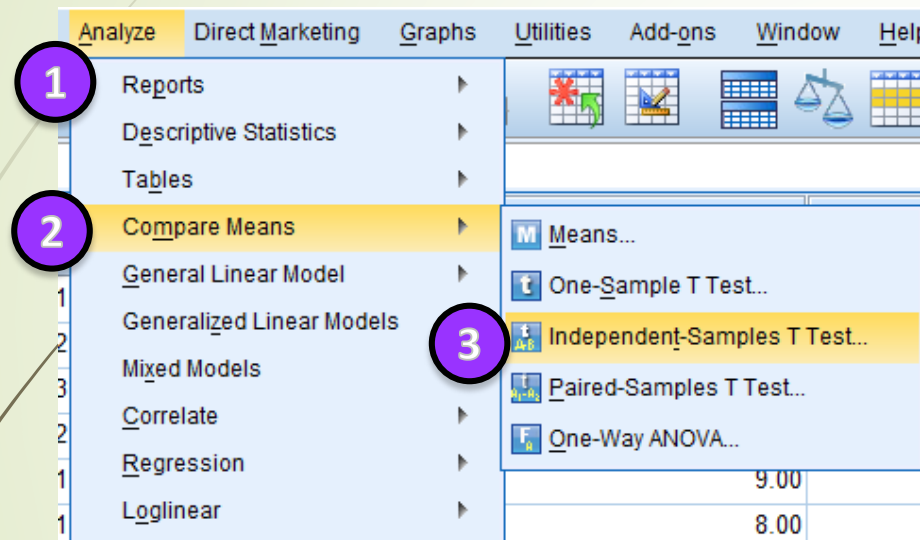
Spelling...

OK Cancel Help

Independent T-test

SPSS: Analyze

分析 > 比較平均數法 > 獨立樣本T檢定



Independent T-test SPSS Output

Group Statistics					
	Sex	N	Mean	Std. Deviation	Std. Error Mean
SBP	Male	12331	134.59	18.950	.171
	Female	16833	134.90	19.438	.150
DBP	Male	12331	78.30	11.683	.105
	Female	16833	77.40	11.556	.089
pluse_pressure	Male	12331	56.29	14.332	.129
	Female	16833	57.49	14.784	.114
WC	Male	12331	86.9879	9.23906	.08320
	Female	16833	82.0870	9.54373	.07356

在舒張壓 (DSP) 的部分，
男性和女性有統計差異
男生平均高於女生 ($P < 0.001$)

Levene檢定

- 變異數相同 ($P > 0.05$)
- 變異數不相同 ($P < 0.05$)

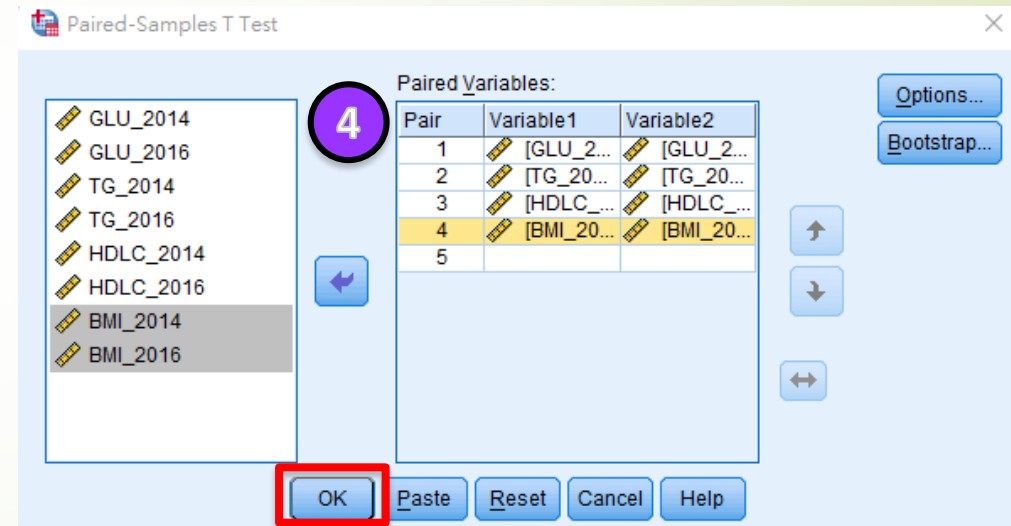
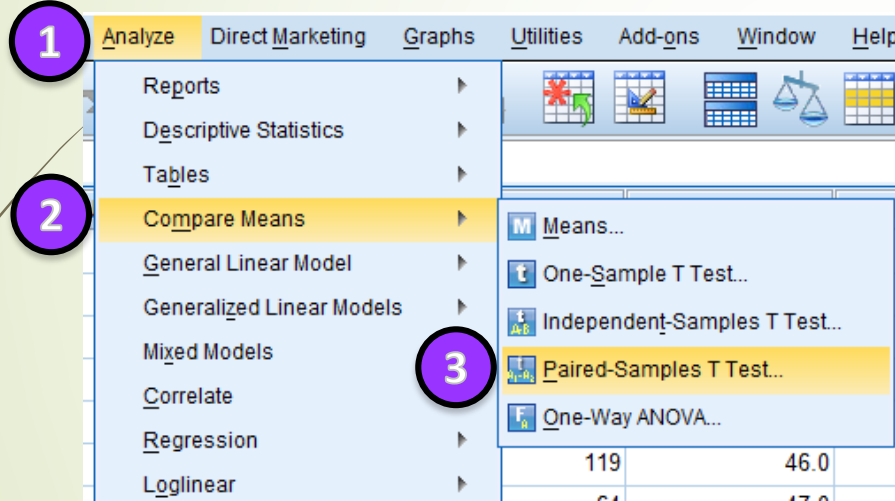
Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
SBP	Equal variances assumed	6.621	.010	-1.346	29162		-.307	.228	-.754	.140
	Equal variances not assumed			-1.351	26938.705	.177	-.307	.227	-.752	.138
DBP	Equal variances assumed	6.533	.011	6.495	29162		.894	.138	.624	1.164
	Equal variances not assumed			6.484	26402.928	.000	.894	.138	.624	1.164
pluse_pressure	Equal variances assumed	18.439	.000	-6.940	29162		-1.201	.173	-1.540	-.862
	Equal variances not assumed			-6.974	27018.184	.000	-1.201	.172	-1.538	-.863

Paired T-test SPSS Dataset

		Before		After					
caseno	GLU_2014	GLU_2016	TG_2014	TG_2016	HDLC_2014	HDLC_2016	BMI_2014	BMI_2016	
1	91	91	245	180	34.0	42.0	27.68	27.27	
2	103	224	89	104	45.0	35.0	19.15	20.16	
3	92	96	96	56	54.0	47.0	20.94	23.57	
4	97	96	123	92	57.0	57.0	25.54	25.57	
5	88	307	125	106	42.0	87.0	18.97	17.10	
6	94	94	101	119	46.0	47.0	21.50	21.79	
7	92	91	49	64	47.0	50.0	25.67	26.58	
8	151	89	83	69	39.0	47.0	24.61	26.37	
9	97	97	120	59	42.0	44.0	23.95	24.31	
10	149	156	142	128	50.0	50.0	25.54	23.84	
11	113	124	337	104	58.0	80.0	23.28	23.74	
12	99	92	114	76	67.0	60.0	20.20	20.57	

Paired T-test SPSS Analysis

分析>比較平均數法>成對樣本T檢定



Paired T-test SPSS Output

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	GLU_2014	104.02	4537	21.923	.325
	GLU_2016	104.89	4537	25.166	.374
Pair 2	TG_2014	117.03	4537	66.712	.990
	TG_2016	114.85	4537	67.833	1.007
Pair 3	HDLC_2014	55.615	4537	15.6628	.2325
	HDLC_2016	55.598	4537	15.7802	.2343
Pair 4	BMI_2014	24.4877	4537	3.46738	.05148
	BMI_2016	24.4141	4537	3.41873	.05076

在Glucose的部分，
2014 和 2016 有統計差異
2016 平均高於 2014 (P = 0.009)
Mean difference: 0.869

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	GLU_2014 - GLU_2016	-.869	22.503	.334	-1.524	-.214	-2.600	4536	.009
Pair 2	TG_2014 - TG_2016	2.186	65.013	.965	.294	4.078	2.265	4536	.024
Pair 3	HDLC_2014 - HDLC_2016	.0164	11.3191	.1680	-.3131	.3458	.098	4536	.922
Pair 4	BMI_2014 - BMI_2016	.07361	1.73773	.02580	.02303	.12418	2.853	4536	.004

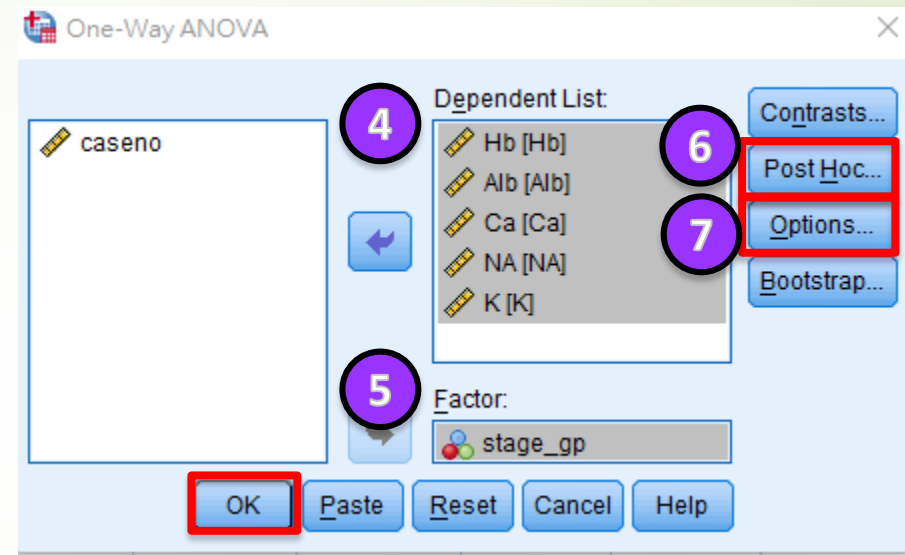
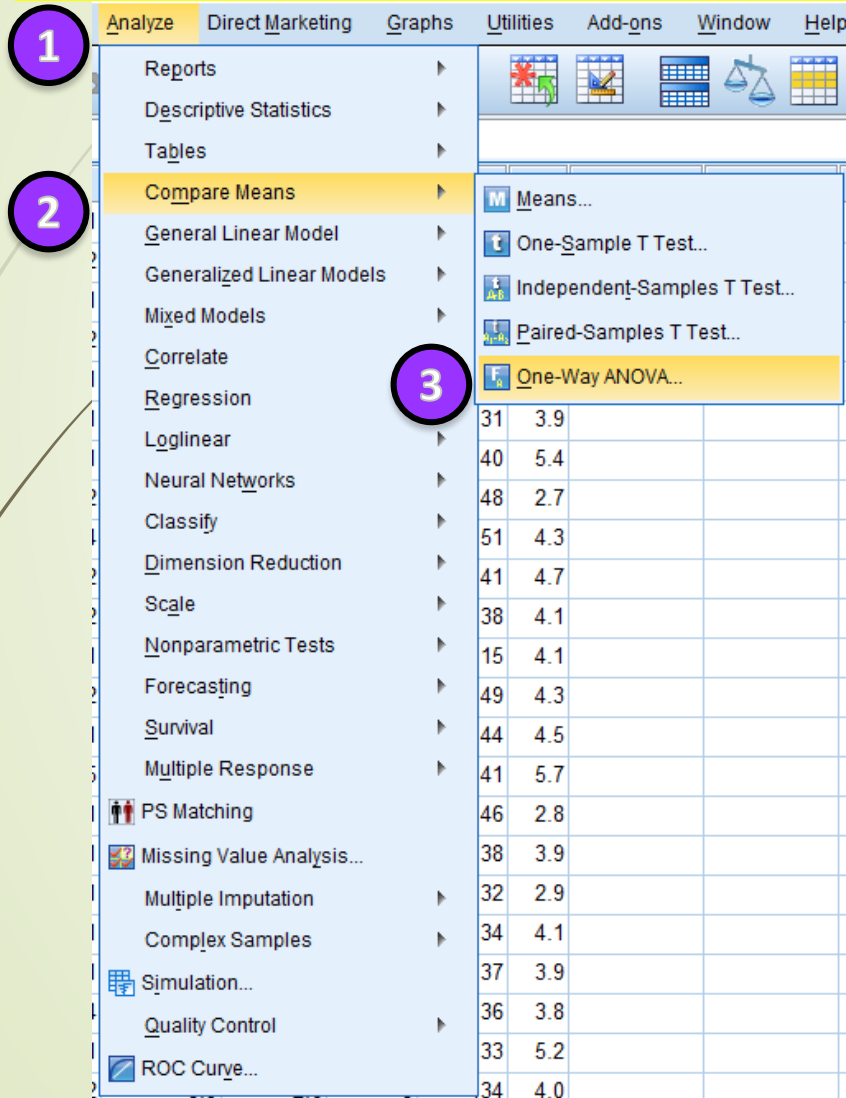
ANOVA

SPSS Dataset

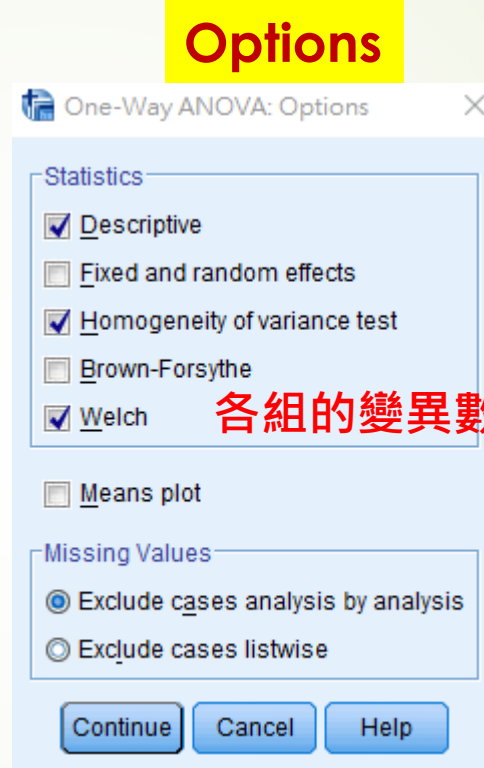
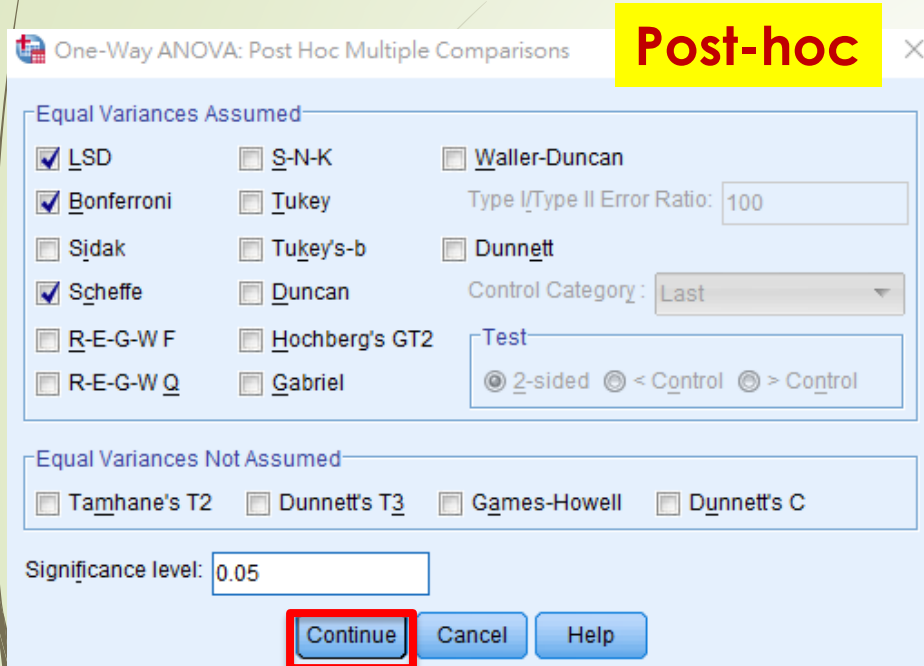
caseno	stage_gp	Hb	Alb	Ca	NA	K
2	1	9.6	1.8	4	135	4.3
3	2	8.3	.	4	131	5.6
4	1	7.5	2.2	8	138	4.1
5	2	10.0	2.6	8	136	3.6
6	1	9.1	3.1	6	140	4.7
8	1	14.5	2.6	8	131	3.9
9	1	10.8	2.7	8	140	5.4
12	2	14.5	3.8	8	148	2.7
13	4	10.9	2.1	9	151	4.3
14	2	10.6	2.7	4	141	4.7
15	2	9.6	3.1	8	138	4.1
16	1	15.6	4.1	8	115	4.1

ANOVA: SPSS Analysis-1

分析 > 比較平均數法 > 單因子變異數分析



ANOVA: SPSS Analysis-2



各組的變異數不相同 (要改用 Welch ANOVA)

Brown-Forsythe 統計量：
檢定組別平均數的相等性。
當變數異相同的假設不成立時，
慣用這個統計量，而不使用 F 統計量。

ANOVA: SPSS Output-1

Descriptives									
		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
Hb	>=90	4860	12.301	2.4542	.0352	12.232	12.370	4.3	20.1
	60-90	3722	12.365	2.4885	.0408	12.285	12.445	3.1	19.0
	45-60	1489	11.984	2.5768	.0668	11.853	12.115	4.2	19.3
	30-45	1128	11.562	2.6699	.0795	11.406	11.718	4.0	19.0
	15-30	644	11.045	2.7465	.1082	10.833	11.258	4.7	19.1
	<15	399	10.133	2.7236	.1363	9.865	10.401	3.4	17.5
	Total	12242	12.077	2.5737	.0233	12.031	12.122	3.1	20.1
Alb	>=90	2674	3.418	.7383	.0143	3.390	3.446	.2	5.9
	60-90	2028	3.444	.7257	.0161	3.413	3.476	.8	5.4
	45-60	855	3.306	.7158	.0245	3.258	3.354	.4	5.4
	30-45	680	3.179	.7195	.0276	3.124	3.233	1.1	5.8
	15-30	427	2.999	.7309	.0354	2.929	3.068	.5	6.7
	<15	276	3.071	.6682	.0402	2.992	3.150	1.0	5.2
	Total	6940	3.349	.7391	.0089	3.331	3.366	.2	6.7

在Alb的部分，
不同CKD stage程度有統計差異
Stage II平均高於其他 ($P < 0.001$)

Test of Homogeneity of Variances				
	Levene Statistic	df1	df2	Sig.
Hb	15.688	5	12236	.000
Alb	2.103	5	6934	.062

Alb, $P > 0.05$ 三組變異數無差異(同質性)
Hb, 違反同質性假設

ANOVA					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	126.191	5	25.238	47.763	.000
Within Groups	3663.970	6934	.528		
Total	3790.161	6939			

Robust Tests of Equality of Means				
Hb	Statistic ^a	df1	df2	Sig.
Welch	85.179	5	2215.976	.000

a. Asymptotically F distributed.

Hb用Welch結果, 取代ANOVA
 $P < 0.001$

ANOVA: SPSS Output-2

Multiple Comparisons						
Dependent Variable: Alb						
Bonferroni						
(I) stage_gp	(J) stage_gp	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
>=90	60-90	-.0260	.0214	1.000	-.089	.037
	45-60	.1121*	.0286	.001	.028	.196
	30-45	.2396*	.0312	.000	.148	.331
	15-30	.4193*	.0379	.000	.308	.531
	<15	.3472*	.0460	.000	.212	.482
60-90	>=90	.0260	.0214	1.000	-.037	.089
	45-60	.1381*	.0296	.000	.051	.225
	30-45	.2657*	.0322	.000	.171	.360
	15-30	.4454*	.0387	.000	.332	.559
	<15	.3732*	.0466	.000	.236	.510
45-60	>=90	-.1121*	.0286	.001	-.196	-.028
	60-90	-.1381*	.0296	.000	-.225	-.051
	30-45	.1276*	.0374	.010	.018	.237
	15-30	.3073*	.0431	.000	.181	.434
	<15	.2351*	.0503	.000	.087	.383
30-45	>=90	-.2396*	.0312	.000	-.331	-.148
	60-90	-.2657*	.0322	.000	-.360	-.171
	45-60	-.1276*	.0374	.010	-.237	-.018
	15-30	.1797*	.0449	.001	.048	.311
	<15	.1075	.0519	.574	-.045	.260
15-30	>=90	-.4193*	.0379	.000	-.531	-.308
	60-90	-.4454*	.0387	.000	-.559	-.332
	45-60	-.3073*	.0431	.000	-.434	-.181
	30-45	-.1797*	.0449	.001	-.311	-.048
	<15	-.0722	.0561	1.000	-.237	.093
<15	>=90	-.3472*	.0460	.000	-.482	-.212
	60-90	-.3732*	.0466	.000	-.510	-.236
	45-60	-.2351*	.0503	.000	-.383	-.087
	30-45	-.1075	.0519	.574	-.260	.045
	15-30	.0722	.0561	1.000	-.093	.237

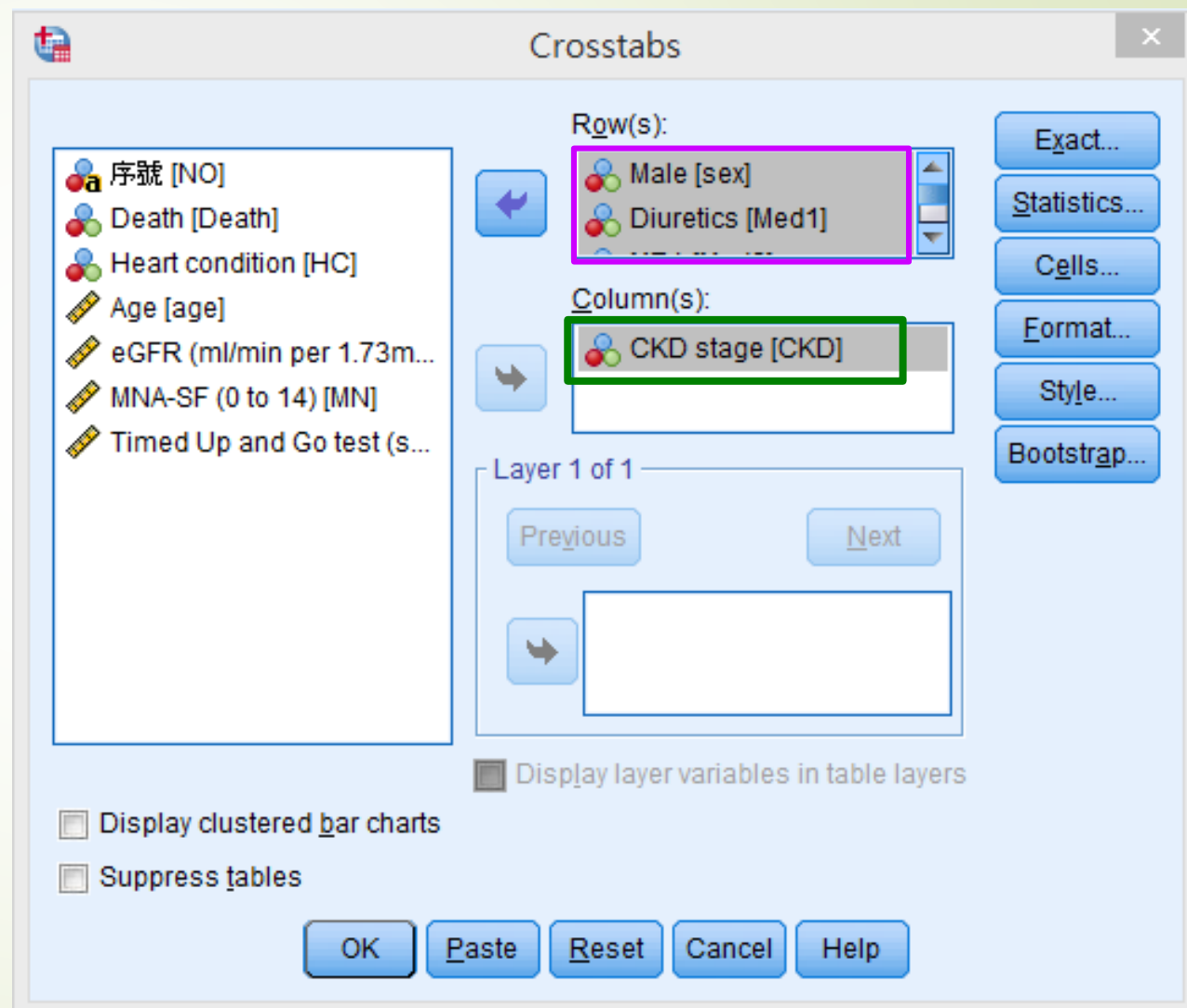
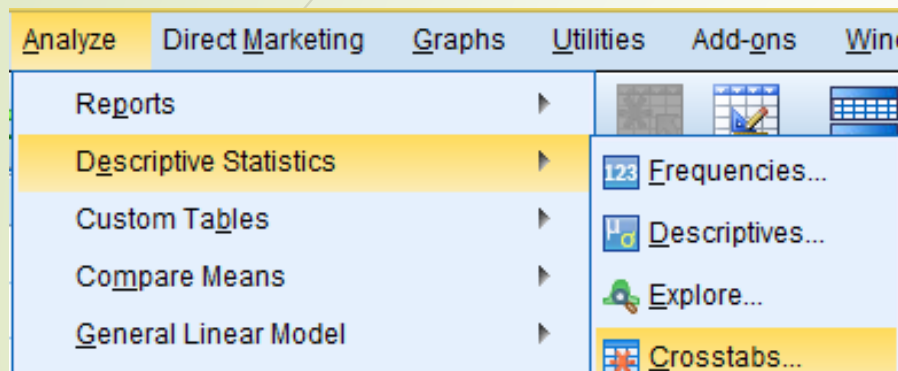
*. The mean difference is significant at the 0.05 level.

在Alb的事後檢定， $P < 0.05$
 Stage I vs Stage III A;
 Stage I vs Stage III B;
 Stage I vs Stage IV;
 Stage I vs Stage V;
 Stage II vs Stage III A;
 Stage II vs Stage III B;
 Stage II vs Stage IV;
 Stage II vs Stage V;

.....

Chi-square Test-1

分析 > 描述性統計 > 交叉表



Chi-square Test-2

Crosstabs: Statistics

3 ☒ **Chi-square**

☐ Correlations

Nominal

- ☐ Contingency coefficient
- ☐ Phi and Cramer's V
- ☐ Lambda
- ☐ Uncertainty coefficient

Ordinal

- ☐ Gamma
- ☐ Somers' d
- ☐ Kendall's tau-b
- ☐ Kendall's tau-c

Nominal by Interval

- ☐ Eta
- ☐ Kappa
- ☐ Risk
- ☐ McNemar

☐ Cochran's and Mantel-Haenszel statistics

Test common odds ratio equals: 1

Continue Cancel Help

Crosstabs: Cell Display

4

Counts

- ☒ Observed
- ☐ Expected
- ☐ Hide small counts

Less than 5

z-test

- ☐ Compare column proportions
- ☐ Adjust p-values (Bonferroni method)

Percentages

- ☐ Row
- ☒ Column
- ☐ Total

Residuals

- ☐ Unstandardized
- ☐ Standardized
- ☐ Adjusted standardized

Noninteger Weights

- ☒ Round cell counts
- ☐ Round case weights
- ☐ Truncate cell counts
- ☐ Truncate case weights
- ☐ No adjustments

Continue Cancel Help

Chi-square Test-3

Crosstabs

Row(s): Male [sex], Diuretics [Med1]

Column(s): CKD stage [CKD]

Layer 1 of 1

Previous Next

☐ Display clustered bar charts

☐ Suppress tables

OK Paste Reset Cancel Help

Exact... Statistics... Cells... Format... Style... Bootstrap...

Exact Tests

☐ Asymptotic only

☐ Monte Carlo

Confidence level: 99 %

Number of samples: 10000

☒ Exact

☒ Time limit per test: 5 minutes

Exact method will be used instead of Monte Carlo when computational limits allow.

For nonasymptotic methods, cell counts are always rounded or truncated in computing the test statistics.

Continue Cancel Help

Chi-square Test SPSS Output

Crosstab

			CKD stage				Total
			CKD 3A	CKD 3B	CKD 4	CKD 5	
Heart condition	Non-HF	Count	119	84	36	15	254
		% within CKD stage	81.0%	83.2%	69.2%	48.4%	76.7%
	HFpEF	Count	15	12	10	8	45
		% within CKD stage	10.2%	11.9%	19.2%	25.8%	13.6%
	HFrEF	Count	13	5	6	8	32
		% within CKD stage	8.8%	5.0%	11.5%	25.8%	9.7%
	Total	Count	147	101	52	31	331
		% within CKD stage	100.0%	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	21.558 ^a	6	.001
Likelihood Ratio	19.173	6	.004
Linear-by-Linear Association	11.965	1	.001
N of Valid Cases	331		

a. 2 cells (16.7%) have expected count less than 5. The minimum expected count is 3.00.

test * Death Crosstabulation

			Death		Total
			Alive	Death	
test	No	Count	30	4	34
		% within Death	10.5%	9.1%	10.3%
	Yes	Count	257	40	297
		% within Death	89.5%	90.9%	89.7%
Total		Count	287	44	331
		% within Death	100.0%	100.0%	100.0%

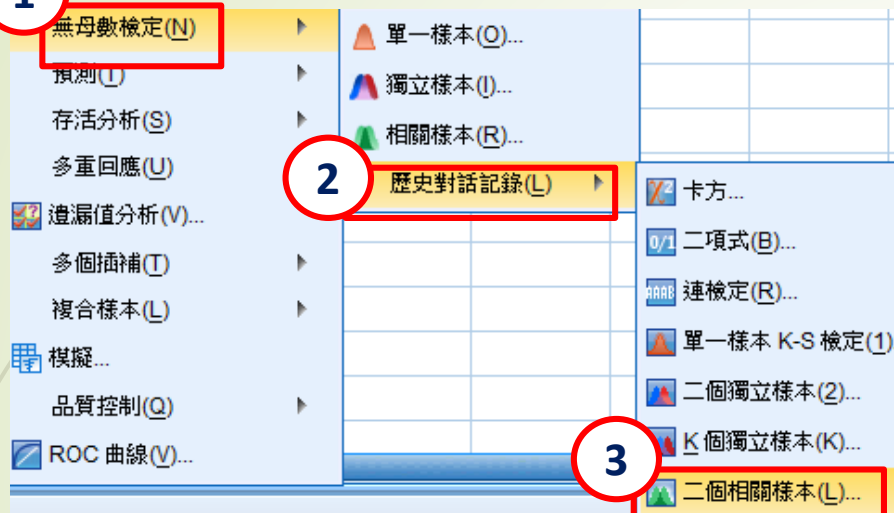
Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.077 ^a	1	.782		
Continuity Correction ^b	.000	1	.992		
Likelihood Ratio	.079	1	.778		
Fisher's Exact Test				1.000	.517
Linear-by-Linear Association	.077	1	.782		
N of Valid Cases	331				

a. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4.52.

McNemar's Test

1



2

歷史對話記錄(L)

3

二個相關樣本(L)...



Before/After

Before

Q1_1 & Q1_2

	Q1_2	
Q1_1	No	Yes
No	9	21
Yes	11	9

After

檢定統計量^a

	Q1_1 & Q1_2
N	50
卡方檢定 ^b	2.531
漸近顯著性	.112

a. McNemar 檢定

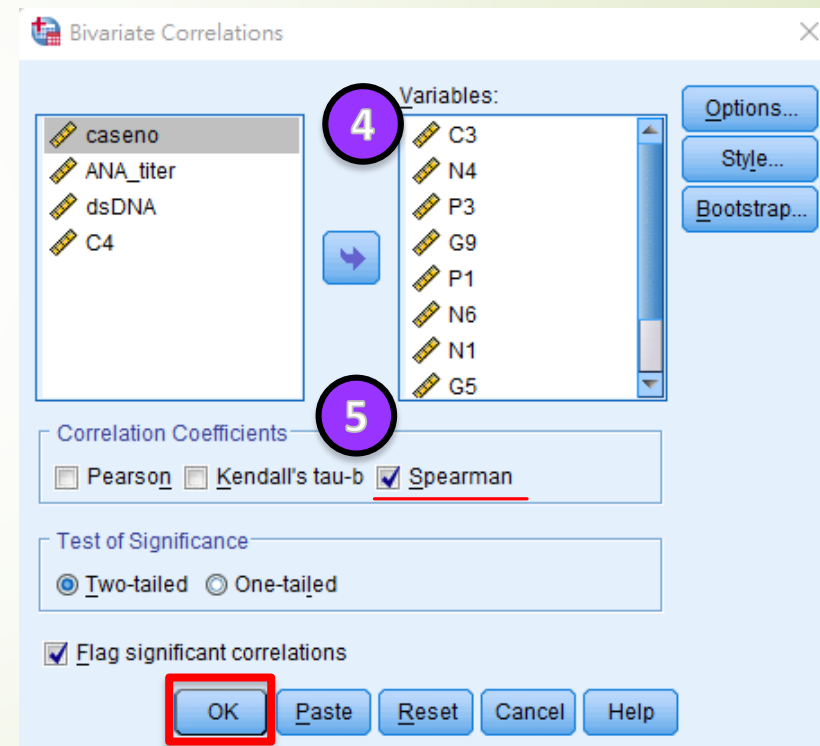
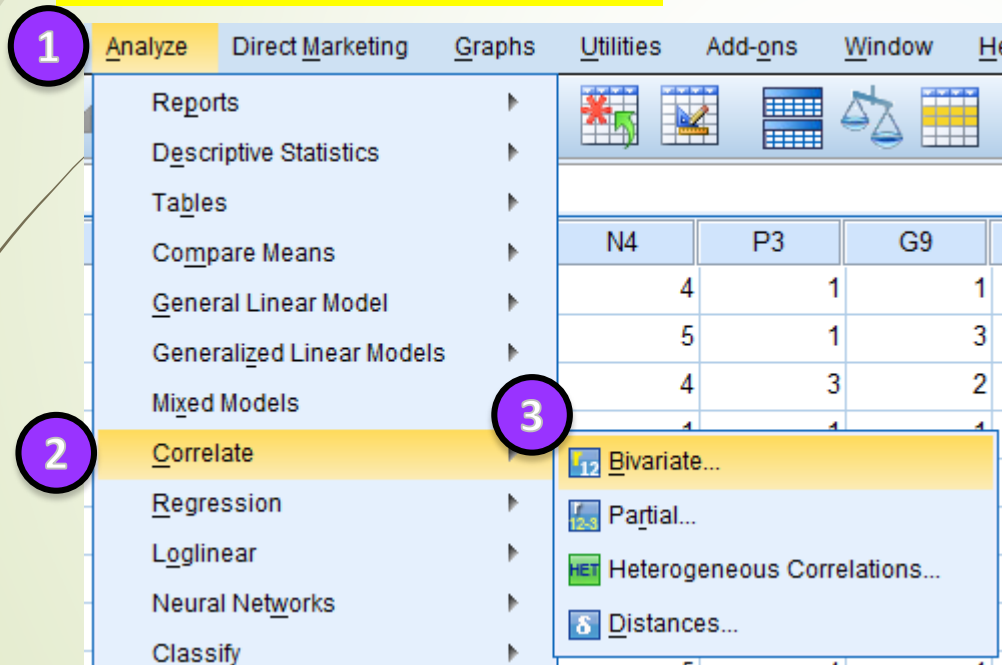
b. 已更正連續性

Correlation SPSS Dataset

	caseno	ANA_titer	dsDNA	C3	C4	N4	P3	G9	P1
1	1	1280	77.946	97.6	23.3	4	1	1	1
2	2	160	21.197	140.7	28.2	5	1	3	2
3	3	160	68.895	122.4	37.6	4	3	2	1
4	4	160	40.632	114.3	24.1	1	1	1	1
5	5	160	6.306	113.7	18.8	5	2	2	2
6	6	320	166.537	63.7	9.5	4	2	3	3
7	7	640	207.814	84.4	10.9	6	2	2	2
8	8	80	141.294	95.3	20.6	5	6	1	3
9	9	640	19.552	107.0	19.1	5	4	1	1
10	10	160	15.728	113.9	21.4	2	4	2	2
11	11	320	161.745	107.3	19.4	2	1	1	1
12	12	80	220.212	138.6	27.9	4	1	1	1

Correlation SPSS Analysis

分析 > 相關 > 雙變數



Correlation: SPSS Output

Correlations												
			C3	N4	P3	G9	P1	N6	N1	G5	P2	UKU
Spearman's rho	C3	Correlation Coefficient	1.000	-.258*	-.213	-.166	-.271*	-.318*	-.327*	-.189	-.225	-.325*
		Sig. (2-tailed)	.	.047	.103	.204	.036	.013	.011	.148	.083	.011
		N	60	60	60	60	60	60	60	60	60	60
	N4	Correlation Coefficient	-.258*	1.000	.159	.200	.290*	.404**	.541**	.102	.382**	.016
		Sig. (2-tailed)	.047	.	.217	.119	.022	.001	.000	.429	.002	.899
		N	60	62	62	62	62	62	62	62	62	62
	P3	Correlation Coefficient	-.213	.159	1.000	.342**	.439**	.000	.010	.294*	.290*	.177
		Sig. (2-tailed)	.103	.217	.	.007	.000	1.000	.939	.020	.022	.170
		N	60	62	62	62	62	62	62	62	62	62
	G9	Correlation Coefficient	-.166	.200	.342**	1.000	.791**	-.047	.101	.315*	.309*	.085
		Sig. (2-tailed)	.204	.119	.007	.	.000	.716	.435	.013	.014	.512
		N	60	62	62	62	62	62	62	62	62	62
	P1	Correlation Coefficient	-.271*	.290*	.439**	.791**	1.000	.078	.217	.383**	.410**	.176
		Sig. (2-tailed)	.036	.022	.000	.000	.	.547	.091	.002	.001	.172
		N	60	62	62	62	62	62	62	62	62	62
	N6	Correlation Coefficient	-.318*	.404**	.000	-.047	.078	1.000	.616**	.273*	-.026	-.193
		Sig. (2-tailed)	.013	.001	1.000	.716	.547	.	.000	.032	.843	.133
		N	60	62	62	62	62	62	62	62	62	62
	N1	Correlation Coefficient	-.327*	.541**	.010	.101	.217	.616**	1.000	.067	.218	-.012
		Sig. (2-tailed)	.011	.000	.939	.435	.091	.000	.	.606	.088	.926
		N	60	62	62	62	62	62	62	62	62	62
	G5	Correlation Coefficient	-.189	.102	.294*	.315*	.383**	.273*	.067	1.000	.128	.033
		Sig. (2-tailed)	.148	.429	.020	.013	.002	.032	.606	.	.321	.801
		N	60	62	62	62	62	62	62	62	62	62
	P2	Correlation Coefficient	-.225	.382**	.290*	.309*	.410**	-.026	.218	.128	1.000	.257*
		Sig. (2-tailed)	.083	.002	.022	.014	.001	.843	.088	.321	.	.043
		N	60	62	62	62	62	62	62	62	62	62
	UKU	Correlation Coefficient	-.325*	.016	.177	.085	.176	-.193	-.012	.033	.257*	1.000
		Sig. (2-tailed)	.011	.899	.170	.512	.172	.133	.926	.801	.043	.
		N	60	62	62	62	62	62	62	62	62	62

*. Correlation is significant at the 0.05 level (2-tailed).

**.

Logistic Regression: SPSS Dataset

Outcome (0 or 1)

	caseno	Inconsistency	Age	Sex	Creatinine_1.4gp
1	1	0	69.4	0	0
2	2	0	53.4	0	0
3	3	0	55.4	0	0
4	4	0	36.1	0	0
5	5	0	35.9	0	0
6	6	0	48.7	0	0
7	7	0	66.1	0	0
8	8	0	48.9	0	0
9	9	0	37.1	1	0
10	10	0	42.1	0	0
11	11	0	57.2	0	0
12	12	0	40.7	0	0

Variable	Code
Sex	
Female	0
Male	1
Creatinine_1.4gp	
<1.4	0
≥1.4	1
UPCR_500gp	
<500	0
≥500	1
ANA_gp	
<1:80	1
1:80-1:640	2
≥1:640	3
C3_87gp	
<87	0
≥87	1
C4_19gp	
<19	0
≥19	1
CIC_10.8gp	
<10.8	0
≥10.8	1
SLEDAI_4gp	
<4	0
≥4	1

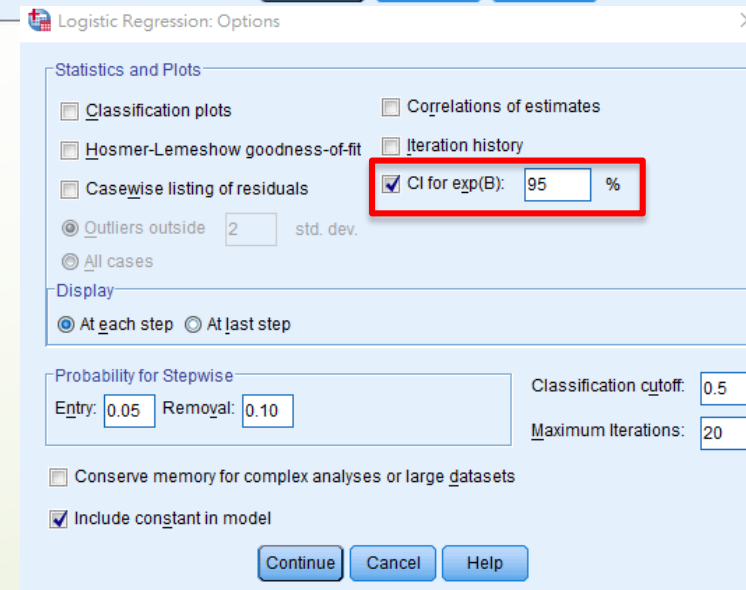
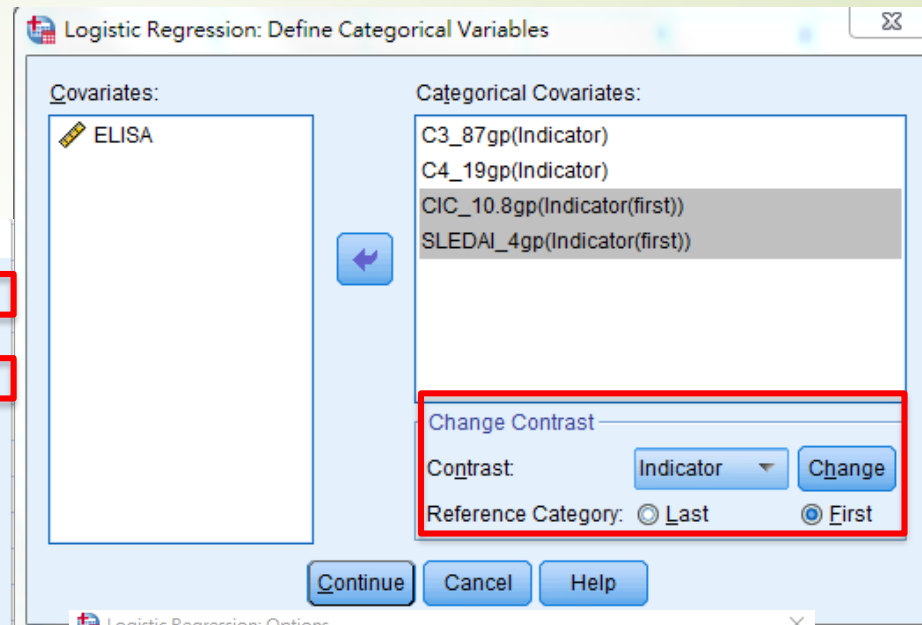
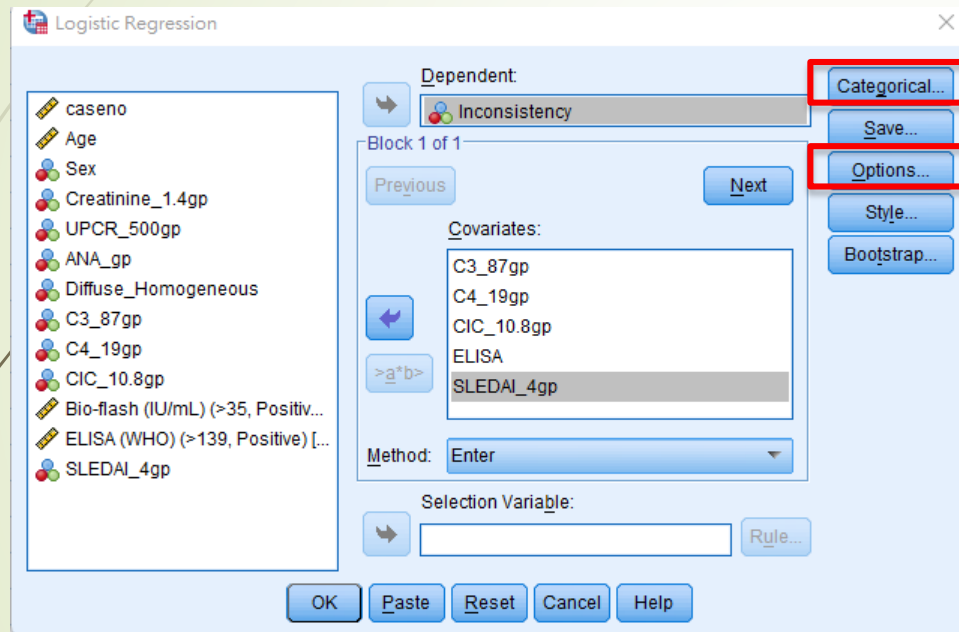
Logistic Regression: SPSS Analysis-1

分析>迴歸>二元Logistic

The screenshot shows the SPSS software interface. The 'Analyze' menu is open, and the 'Regression' option is selected, which has opened a submenu. In this submenu, 'Binary Logistic...' is highlighted. Other options in the 'Regression' submenu include 'Automatic Linear Modeling...', 'Linear...', 'Residual Heteroscedasticity Test...', 'Regression Relative Importance', 'Curve Estimation...', 'Tobit Regression...', 'Robust Regression...', 'Quantile Regression...', 'Partial Least Squares...', 'Heckman Regression', 'Multinomial Logistic...', and 'Ordinal...'. In the background, a data view is partially visible with columns labeled '_1.4gp', 'UPCR_500gp', 'ANA_gp', and 'Diff'.

_1.4gp	UPCR_500gp	ANA_gp	Diff
0	0	.	
0	1	2	
0	0	3	
0	0		

Logistic Regression: SPSS Analysis-2



Logistic Regression: SPSS Output

OR: odds ratio

Variables in the Equation								
	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Step 1 ^a								
C3_87gp(1)	.074	.424	.031	1	.861	1.077	.469	2.473
C4_19gp(1)	-.040	.449	.008	1	.930	.961	.399	2.315
CIC_10.8gp(1)	-.878	.419	4.400	1	.036	.415	.183	.944
ELISA	-.017	.003	42.615	1	.000	.983	.978	.988
SLEDAI_4gp(1)	-1.114	.449	6.150	1	.013	.328	.136	.792
Constant	4.064	.749	29.416	1	.000	58.196		

a. Variable(s) entered on step 1: C3_87gp, C4_19gp, CIC_10.8gp, ELISA, SLEDAI_4gp.

經多變項調整後，

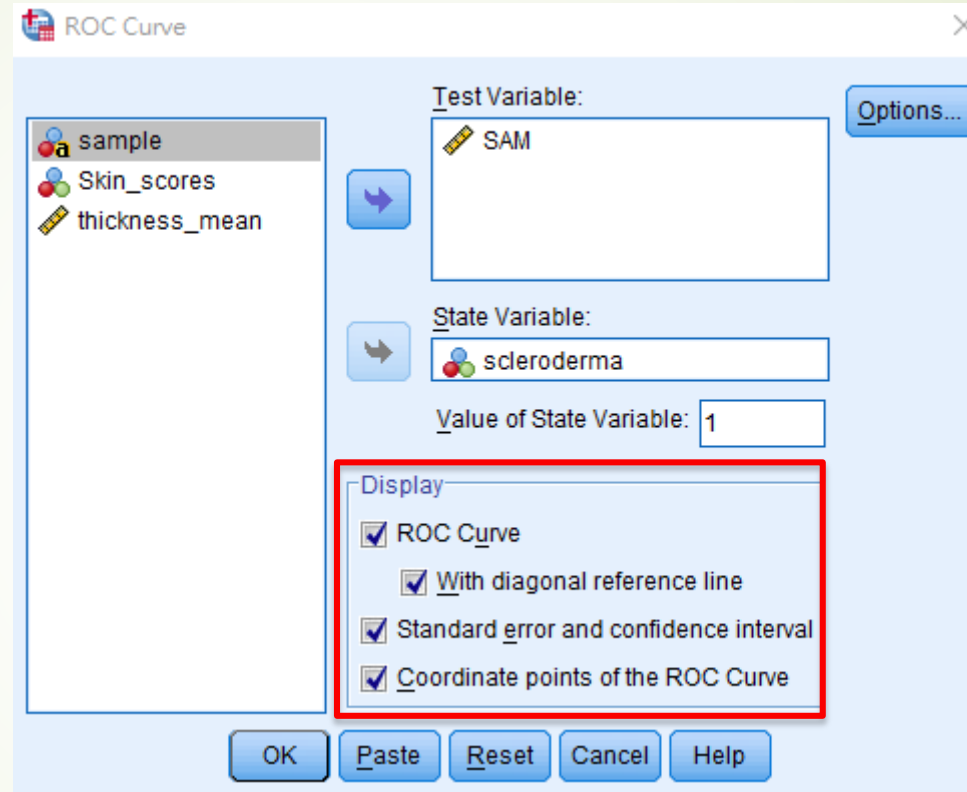
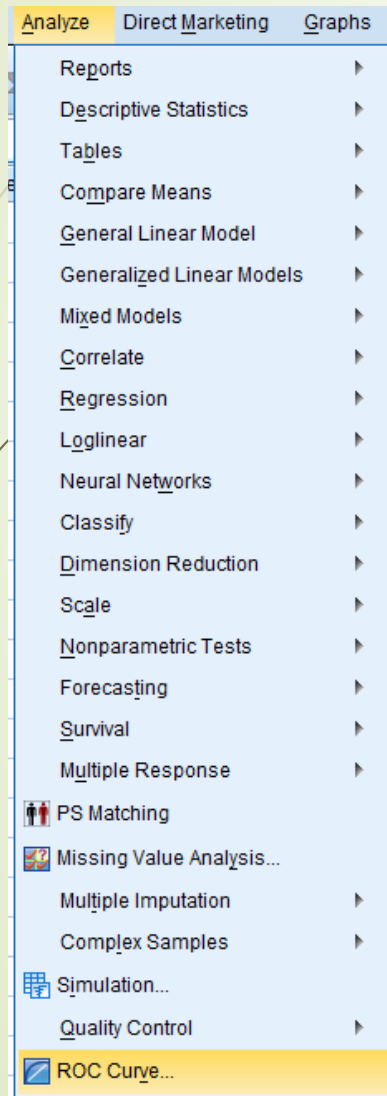
SLEDAI ≥ 4 比 < 4 有 0.328 倍的風險發生不一致情形且達統計差異(P = 0.013)，代表 SLEDAI ≥ 4 比 < 4 ，較不會發生不一致情形。

ROC: SPSS Dataset

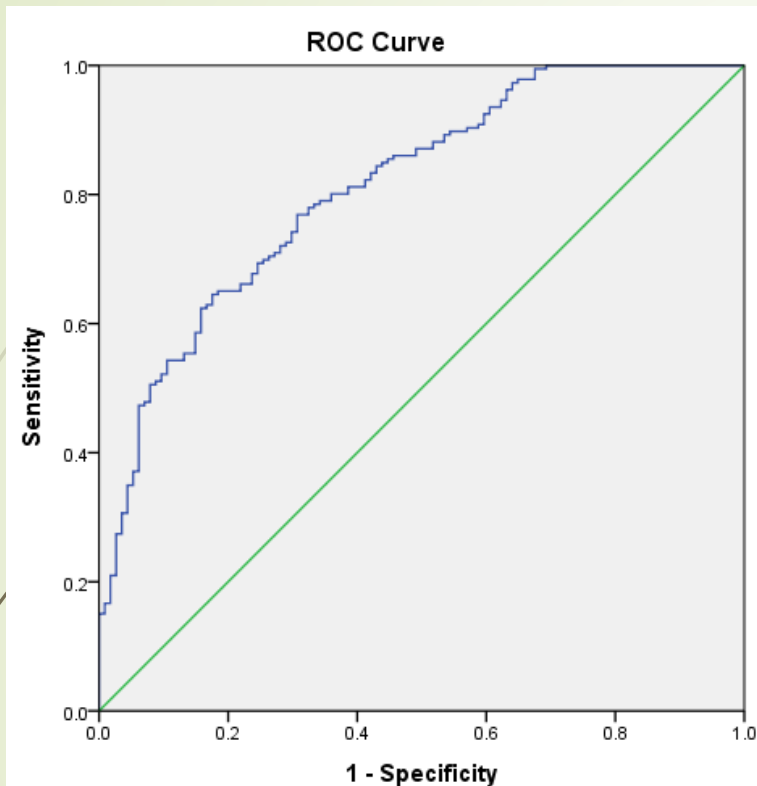
	caseno	scleroderma	Skin_scores	thickness_mean	SAM	Strainratio
1	1	1.00	.00	.175	.016	1.02
2	2	1.00	.00	.129	.014	5.07
3	3	1.00	.00	.142	.022	4.53
4	4	1.00	.00	.127	.011	1.12
5	5	1.00	.00	.131	.015	1.45
6	6	1.00	.00	.148	.014	7.93
7	7	1.00	2.00	.255	.041	.19
8	8	1.00	2.00	.253	.038	3.16
9	9	1.00	2.00	.216	.038	6.15
10	10	1.00	2.00	.216	.044	1.91
11	11	1.00	2.00	.203	.040	6.77
12	12	1.00	2.00	.161	.012	5.81

ROC: SPSS Analysis

分析>ROC曲線



ROC: SPSS Output



AUC

Area Under the Curve

Test Result Variable(s): SAM

Area	Std. Error ^a	Asymptotic Sig. ^b	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.812	.025	.000	.764	.861

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Coordinates of the Curve

Test Result Variable(s): SAM

Positive if Greater Than or Equal To ^a	Sensitivity	1 - Specificity
.0000	1.000	1.000
.0045	1.000	.991
.0055	1.000	.974
.0065	1.000	.930
.0075	1.000	.886
.0085	1.000	.833
.0095	1.000	.789
.0105	1.000	.711
.0115	.989	.675
.0125	.973	.640
.0135	.935	.614
.0145	.903	.579
.0155	.892	.544
.0165	.860	.491
.0175	.833	.421
.0185	.823	.412
.0195	.769	.316
.0205	.726	.289
.0215	.715	.281

Youden Index

- **Youden Index** = Sensitivity + Specificity - 1
- 新增 Specificity 欄位, 公式為 $1 - (1 - \text{specificity})$
- 將 $\text{sen} + \text{spe} - 1$ 欄 由大至小排序

**Cut-off
Point**

Coordinates of the Curve			Youden Index	
Test Result \ SAM				
Positive if Greater Than or Equal To ^a	Sensitivity	1 - Specificity	spe	sen+spe-1
.0255	.618	.158	0.842	0.460
.0195	.769	.316	0.684	0.453
.0245	.651	.202	0.798	0.449
.0235	.661	.219	0.781	0.442
.0205	.726	.289	0.711	0.436
.0225	.672	.237	0.763	0.435
.0215	.715	.281	0.719	0.434
.0265	.575	.149	0.851	0.426
.0285	.516	.096	0.904	0.420
.0275	.548	.132	0.868	0.417
.0175	.833	.421	0.579	0.412
.0305	.473	.061	0.939	0.412
.0185	.823	.412	0.588	0.410
.0295	.484	.079	0.921	0.405
.0315	.435	.061	0.939	0.374
.0165	.860	.491	0.509	0.369
.0325	.419	.061	0.939	0.358

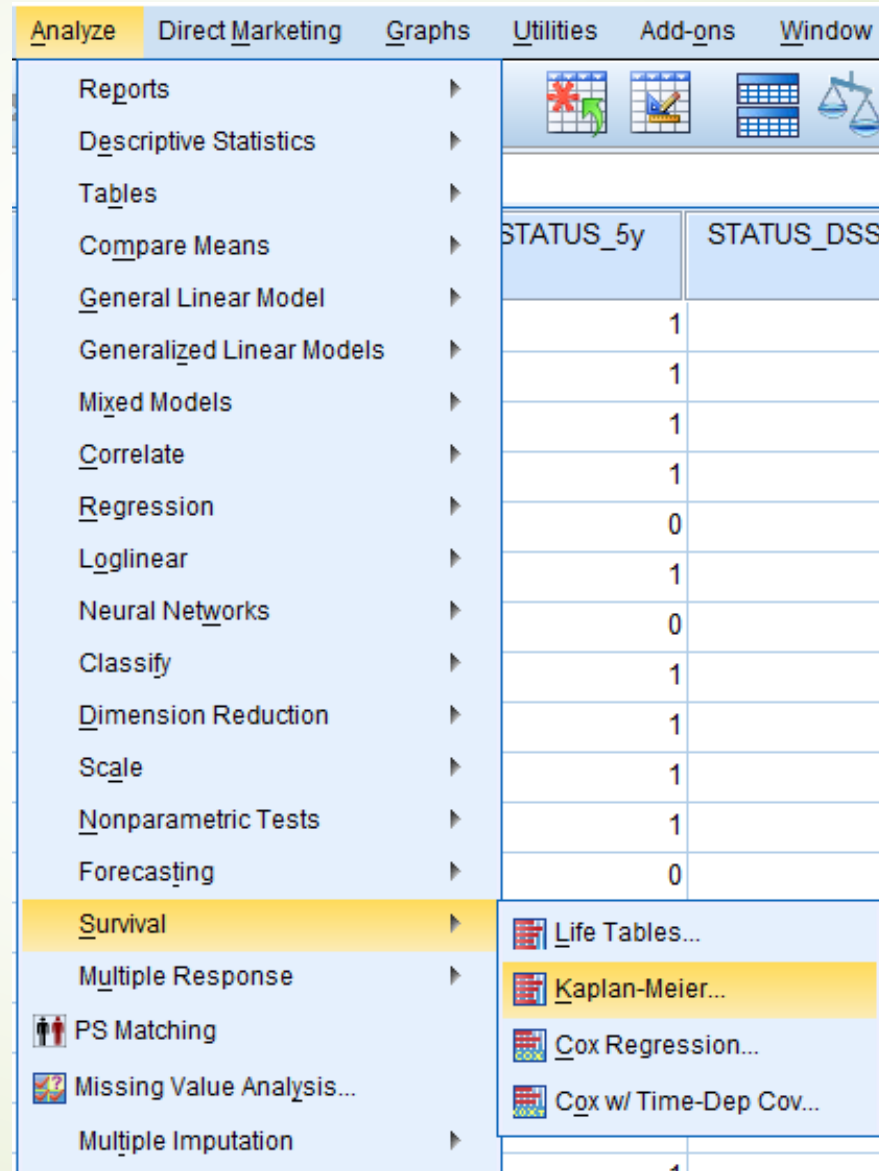
KM Plot and Cox Regression

SPSS Dataset

ID	group	OP_date	End_Date	STATUS_5y	STATUS_DSS_5y	Month_OS	Age_gp	Hist_grading
1	1	2016/04/20	2021/09/09	1	0	64.66	0	2
2	1	2012/04/27	2020/02/20	1	0	93.80	1	1
3	1	2015/12/30	2021/01/19	1	0	60.68	0	1
4	0	2010/11/12	2016/04/02	1	0	64.66	0	2
5	1	2012/11/21	2014/10/01	0	1	22.31	0	2
6	1	2013/10/08	2020/02/12	1	0	76.16	0	1
7	0	2006/03/29	2008/05/25	0	1	25.89	0	1
8	1	2012/06/19	2020/12/23	1	0	102.14	1	2
9	1	2018/10/29	2021/10/20	1	0	35.71	1	1
10	1	2019/08/27	2021/10/18	1	0	25.72	0	2
11	0	2010/02/05	2017/11/24	1	0	93.60	0	2
12	0	2011/02/24	2014/05/14	0	1	38.60	0	2
13	1	2017/08/25	2021/01/05	1	0	40.38	1	1
14	1	2012/04/20	2015/07/10	1	0	38.64	0	2
15	0	2005/10/06	2015/02/18	1	0	112.43	0	2
16	1	2018/06/25	2021/10/14	1	0	39.66	0	2
17	1	2015/04/08	2021/09/28	1	0	77.70	1	1
18	1	2014/10/17	2021/10/27	1	0	84.34	0	1
19	0	2012/08/03	2013/05/12	0	1	9.26	0	2
20	1	2013/01/14	2017/09/26	0	0	56.38	0	2

Kaplan-Meier Survival Analysis

分析>存活分析
>Kaplan-Meier 統計



Kaplan-Meier Survival Analysis

The image displays the SPSS Kaplan-Meier Survival Analysis workflow with four dialog boxes and red annotations:

- Kaplan-Meier (Main Dialog):**
 - Time:** Month_OS
 - Status:** STATUS_DSS_5y(1)
 - Define Event...** (Annotated with a red box and arrow pointing to the Define Event dialog)
 - Factor:** group
 - Strata:** (Empty)
 - Label Cases by:** (Empty)
- Kaplan-Meier: Compare Factor Levels:**
 - Test Statistics:** ☒ Log rank, ☐ Breslow, ☐ Tarone-Ware
 - Linear trend for factor levels:** ☐ Pooled over strata, ☐ Pairwise over strata, ☐ For each stratum, ☐ Pairwise for each stratum
- Kaplan-Meier: Define Event For Status ...:**
 - Value(s) indicating event has occurred:** ☒ Single value: 1, ☐ Range of values: through, ☐ List of values:
- Kaplan-Meier: Options:**
 - Statistics:** ☒ Survival table(s), ☒ Mean and median survival, ☐ Quartiles
 - Plots:** ☒ Survival, ☐ One minus survival, ☐ Hazard, ☐ Log Survival

Red boxes highlight the **Define Event...**, **Compare Factor...**, **Options...**, **Log rank**, **Single value: 1**, and **Survival** options. Red arrows indicate the flow from the main dialog to the sub-dialogs.

Kaplan-Meier Output

Case Processing Summary

group	Total N	N of Events	Censored	
			N	Percent
CT alone	27	8	19	70.4%
Sandwich	39	3	36	92.3%
Overall	66	11	55	83.3%

Disease specific survival

	Total	DSS	Censored		Survival rate (%)			p for log rank
			n	%	1y	3y	5y	
Treatment								
CT alone	27	8	19	70.4%	92.6%	81.2%	69.6%	0.041
Sandwich	39	3	36	92.3%	100.0%	97.4%	91.8%	

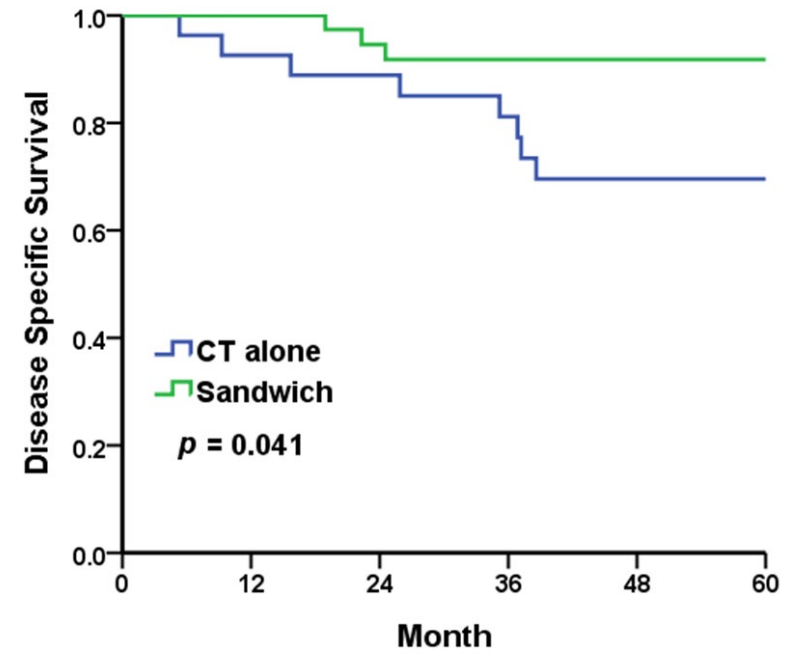
Survival Table

group	Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases	
			Estimate	Std. Error			
CT alone	1	5.322	DSS	.963	.036	1	26
	2	9.265	DSS	.926	.050	2	25
	3	15.704	DSS	.889	.060	3	24
	4	22.472	no-DSS	.	.	3	23
	5	25.889	DSS	.850	.069	4	22
	6	35.187	DSS	.812	.076	5	21
	7	36.862	DSS	.773	.082	6	20
	8	37.191	DSS	.734	.086	7	19
	9	38.604	DSS	.696	.090	8	18
	10	63.639	no-DSS	.	.	8	17
	11	64.526	no-DSS	.	.	8	16
	12	64.657	no-DSS	.	.	8	15
	13	72.279	no-DSS	.	.	8	14
	14	85.290	no-DSS	.	.	8	13
	15	89.823	no-DSS	.	.	8	12

Overall Comparisons

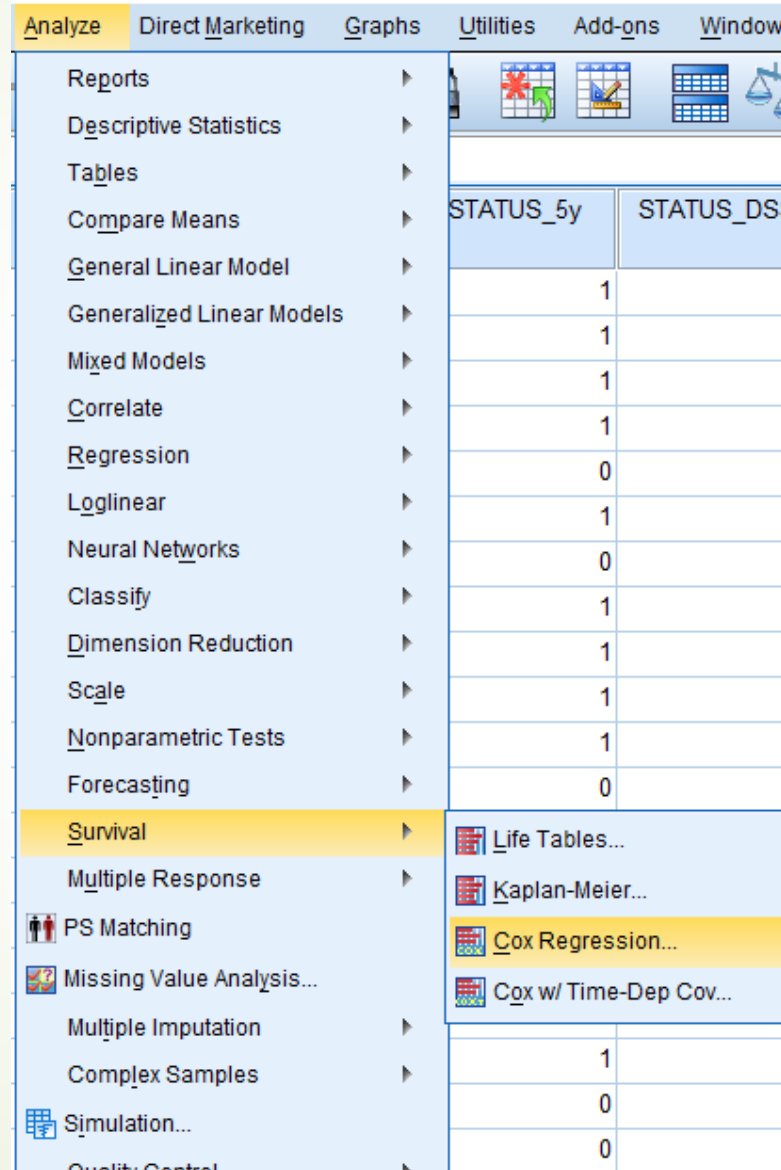
	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	4.169	1	.041

Test of equality of survival distributions for the different levels of group.



Cox Proportional Hazards Regression

分析>存活分析>Cox迴歸



Cox Proportional Hazards Regression

The image displays three screenshots of the SPSS Cox Regression dialog boxes, illustrating the configuration of categorical covariates and model options.

Top Left: Cox Regression Dialog

- Time:** Month_OS
- Status:** STATUS_DSS_5y(1)
- Block 1 of 1:** Previous, Next
- Covariates:** Hist_grading(Cat), group(Cat), Deepmyometriuminv...
- Method:** Enter
- Strata:**
- Buttons:** OK, Paste, Reset, Cancel, Help

Top Right: Cox Regression: Define Categorical Covariates Dialog

- Covariates:**
- Categorical Covariates:** Hist_grading(Indicator(first)), group(Indicator(first)), Deepmyometriuminv...
- Change Contrast:** Contrast: Indicator, Reference Category: Last, First
- Buttons:** Continue, Cancel, Help

Bottom Right: Cox Regression: Options Dialog

- Model Statistics:** ☒ CI for exp(B) 95 %
- Correlation of estimates:** ☐
- Display model information:** ☒ At each step, ☐ At last step
- Probability for Stepwise:** Entry: .05, Removal: .10
- Maximum Iterations:** 20
- Display baseline function:** ☐
- Buttons:** Continue, Cancel, Help

Red boxes and arrows highlight the **Categorical...** button in the top left dialog, the **Options...** button in the top left dialog, and the **Model Statistics** section in the bottom right dialog.

Cox Model-Output

Categorical Variable Codings^{a,c,d}

		Frequency	(1)
group ^b	0=CT alone	27	0
	1=Sandwich	39	1
Hist_grading ^b	1=Grade 1 and 2	29	0
	2=Grade 3	37	1
Deepmyometriuminvasion ^b	0=Absent	27	0
	1=Present	39	1

a. Category variable: group

b. Indicator Parameter Coding

c. Category variable: Hist_grading

d. Category variable: Deepmyometriuminvasion (Deep myometrium invasion)

經多變項調整後，

Sandwich相較於CT alone降低 23% 的風險死於子宮內膜癌且有統計差異

(HR: 0.229, 95% CI: 0.060-0.867; P = 0.030)

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Hist_grading	2.215	1.050	4.452	1	.035	9.161	1.170	71.701
group	-1.474	.679	4.711	1	.030	.229	.060	.867
Deepmyometriuminvasion	2.245	1.051	4.559	1	.033	9.442	1.202	74.150

感謝您的聆聽！



生統小組：統計方法教育訓練