R-web 資料分析應用：圖表繪製(一)

沈彥廷　副統計分析師

上一期的生統 eNews 向大家介紹了【雲端資料分析暨導引系統】(R-web, http://www.r-web.com.tw)的環境架構以及基本的檔案上傳功能、描述性摘要統計方法等等。那麼緊接著我們這期就來學學如何運用 R-web 進行最直觀的資料檢視方法-『圖表繪製』吧！

面對一組未知結構的資料，善用統計圖表來瞭解資料特性是很重要的。假設您是一個理財專員，手中握有大量股匯市資訊，但該如何在有限時間內向您的客戶進行一個簡單明瞭的簡報，進而展現您的專業、提升客戶的信任感？此時統計圖表比起有如甲骨文般的原始數據就更能清楚且有效的呈現資料的分佈，即使不是統計專家也能夠看圖說故事，將雜亂的『資料』轉化為有用的『資訊』。

我們將在本期中依序介紹 R-web 圖表繪製模組內的次數分配表、列聯表、莖葉圖、2D 散佈圖及 3D 散佈圖。本章節統一使用源自基隆社區為基礎的整合篩檢計畫(Keelung Community-based Integrated Screen Program, KCIS)的心血管疾病資料作為範例資料檔，有關此資料的詳細資訊及變數定義請參閱首期生統 eNews。

➢ 次數分配表

次數分配表是一個常見的敘述性統計方法。將類別資料依照其組別分組，或將數值資料依照觀察值的大小分成若干組，計算每一組的次數、相對次數等資訊，以了解資料的分布情形。

在 R-web 主選單中依序點選【圖表繪製】→【次數分配表】。

操作畫面如上圖所示。首先選擇欲進行分析的資料檔，點選後系統將自動帶出參數設定畫面，接著在步驟二中選入欲計算次數分配的變數(可選擇多個變數)。假設我們想了解心血管疾病資料中年齡及性別變數的頻率/次數分配情形，則可選入 Age、Gender 變數。若在步驟二中所選擇欲計算次數分配的變數皆為類別變數，則可直接點擊【繪製表格】按鈕進行次數分配表的繪製；反之，若在步驟二中所選擇的變數包含數值變數，此時系統將自動帶出步驟三：分組設定。

分組設定中的分割方法共有等間距法、等頻率法、k 組平均數法、使用者自訂四個選項，分割組數則提供 2 至 30 組供選擇。由於此處範例中選擇使用者自訂分割方法，因此還須另行輸入分割點(分割組數 10 組則設定 9 個分割點和資料代碼)，所有參數設定完畢後點擊【繪製表格】按鈕即可繪製次數分配表。

- 資料名稱：CVD
- 變數名稱：Age, Gender
- 計算時間：0.322秒
- 次數分配表：

變數：Age

組別 group	次數 frequency	相對次數 relative frequency	累積次數 cumulative frequency	累積相對次數 cumulative relative frequency
小於10	0	0	0	0
10~20	15	2e-04	15	2e-04
20~30	6246	0.0969	6261	0.0971
30~40	15349	0.238	21610	0.3351
40~50	17832	0.2765	39442	0.6116
50~60	11427	0.1772	50869	0.7888
60~70	8844	0.1371	59713	0.9259
70~80	4720	0.0732	64433	0.9991
80~90	51	8e-04	64484	0.9999
超過90	0	0	64484	0.9999
<NA>	5	1e-04	64489	1
Total	64489	1	-	-

變數：Gender

組別 group	次數 frequency	相對次數 relative frequency
0	40438	0.6271
1	24051	0.3729
Total	64489	1

由上圖可以看出此心血管疾病研究資料受試者年齡主要集中於 30~50 歲之間(超過 50%)；在性別方面，女性受試者則有六成以上的佔比。此外，數值變數 Age 除了次數與相對次數資訊外，亦有累積次數及累積相對次數可供檢視。

➢ 列聯表

列聯表為根據兩個(或以上)的類別變數繪製而成的頻率表。當選擇多個分層變數時，又稱為多維列聯表。在 R-web 內可依序點選主選單中【圖表繪製】→【列聯表】進行繪製。若欲繪製之變數為數值變數，亦可設定切割組數或切割點將數值變數轉換為類別變數進行表格繪製。

參數設定、進階選項設定畫面如上圖及右圖所示。將欲繪製列聯表的變數選入相對應欄位中，其中列變數一及行變數一皆為必選。此外，R-web 亦提供將列聯表轉換為資料框架型態儲存至使用者個人資料檔的功能。

進階選項中可透過下拉選單選擇列聯表計算內容為次數或比例，亦可勾選核取項目設定附加計算邊際和。所有參數設定無誤後，點選【繪製表格】即可開始進行計算。

- **資料名稱**：CVD
- **列變數**：Betelnut, FamilyHx
- **行變數**：Alc_Drink, Tobacco
- **分層變數**：CVD
- **儲存位置**： 使用者個人資料檔 - mydata
- **處理時間**：0.04秒

- **列聯表 - 次數**：

CVD = 0

Alc_Drink		0			1			<NA>			Total
Betelnut FamilyHx\Tobacco		0	1	<NA>	0	1	<NA>	0	1	<NA>	
0	0	35234	5493	23	3882	5763	13	147	99	9	50663
	1	1695	272	0	207	295	0	7	5	1	2482
1	0	107	614	0	250	2899	7	3	42	0	3922
	1	5	36	0	10	159	0	0	2	0	212
<NA>	0	34	22	0	28	110	5	15	40	933	1187
	1	3	0	0	1	3	0	1	0	4	12
Total		37078	6437	23	4378	9229	25	173	188	947	58478

CVD = 1

Alc_Drink		0			1			<NA>			Total
Betelnut FamilyHx\Tobacco		0	1	<NA>	0	1	<NA>	0	1	<NA>	
0	0	3609	692	3	364	668	1	9	6	1	5353
	1	127	22	0	16	26	0	0	0	0	191
1	0	14	51	0	19	189	0	0	3	0	276
	1	1	2	0	1	9	0	0	0	0	13
<NA>	0	2	1	1	3	5	0	1	2	162	177
	1	0	0	0	0	1	0	0	0	0	1
Total		3753	768	4	403	898	1	10	11	163	6011

- **資料型態**：

變數名稱	Betelnut	Alc_Drink	FamilyHx	Tobacco	CVD	Freq
變數型態	類別	類別	類別	類別	類別	數值
1.	0	0	0	0	0	35234
2.	1	0	0	0	0	107
3.	<NA>	0	0	0	0	34
4.	0	1	0	0	0	3882
5.	1	1	0	0	0	250
◦			◦			
◦			◦			
◦			◦			
104.	1	1	1	<NA>	1	0
105.	<NA>	1	1	<NA>	1	0
106.	0	<NA>	1	<NA>	1	0
107.	1	<NA>	1	<NA>	1	0
108.	<NA>	<NA>	1	<NA>	1	0

上圖為列聯表繪製結果。在本例中我們使用嚼檳榔習慣(Betelunt)、家

族心血管疾病史(FamilyHx)作為列變數；飲酒習慣(Alc_Drink)、抽菸習慣(Tobacco)作為行變數。由於系統限制行列變數最多僅可各使用兩個變數，因此分層變數：個人心血管疾病史(CVD)即以分表形式呈現。另外，輸出頁面下方則為列聯表經轉換至資料框架之型態檢視。

➢ 莖葉圖

莖葉圖也是一種呈現資料分佈結構的方法，其特色在於呈現方式類似直方圖但又能保留原始數據資料。除了可看出如同直方圖一樣的資料散佈趨勢之外，同時也能更詳細的表現出個別樣本資訊，對於資料量不大的情況下尤其適用。在 R-web 內可依序點選主選單中【圖表繪製】→【莖葉圖】進行繪製。

由於在資料量較大的情況下並不建議使用莖葉圖，因此我們透過分層抽樣取出 100 筆樣本作為此處的範例資料。假設我們想了解腰圍在資料中的分佈情形，僅需在步驟二中選入 Waist 變數即可。在進階選項中，包括莖葉圖單位、圖形寬度皆預設由系統自行選擇合適參數，其中圖

形寬度選項採用 Tukey 法，共有一枝一葉、一枝二葉、一枝五葉可供選擇。使用者也可勾選是否顯示累積次數或刪除遺失值，所有參數設定完畢後，點選【繪製圖形】即可開始進行莖葉圖繪製。

- **資料名稱**：CVD_100
- **變數名稱**：Waist
- **計算時間**：0.14秒

- **莖葉圖**[I]：

```
1 | 2: represents 12
 leaf unit: 1
       n: 99
  11    6. | 55557788999
  24    7* | 0011112223344
  43    7. | 55555666677788888999
 (16)   8* | 0000001122234444
  40    8. | 55556777778899
  26    9* | 00001112233444
  12    9. | 57788
   7   10* | 013
   4   10. | 579
      11* |
      11. |
   1   12* | 1
NA's: 1
```

I : 0,1,2,3,4以'*'表示；5,6,7,8,9以'.'表示

上圖為莖葉圖繪製結果。圖中『|』符號表示莖葉的分界，且在本例中系統所選擇之最佳莖葉圖單位為 1、最佳圖形寬度為一枝二葉，因此圖中 6.|5 即表示 65、7*|0 則代表 70，以此類推。圖形最左側則為累積次數，括弧處為中位數發生位置。

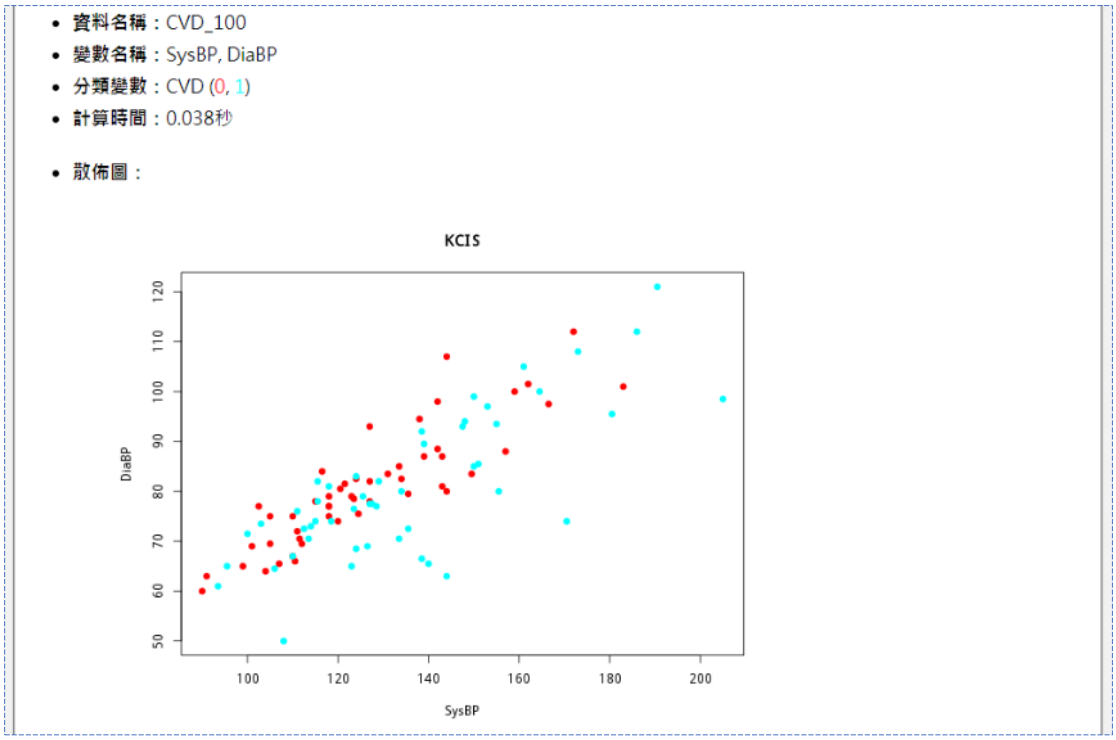
➢ 2D 散佈圖

2D 散佈圖可用以將兩個可能相關之數值變數分別置於座標圖上的 X 與 Y 軸，用圖點標示各資料點的位置，可初步觀察兩變數間的相關性。在 R-web 中，2D 散佈圖共包含一般散佈圖、散佈圖矩陣、條件散佈圖，其中散佈圖矩陣及條件散佈圖僅是一般散佈圖的延伸應用，因此為節省版面空

間，我們在此僅為各位讀者介紹一般散佈圖功能。依序點選主選單中【圖表繪製】→【散佈圖】→【2D 散佈圖】→【一般散佈圖】進行繪製。

上圖及右圖為散佈圖的參數設定和進階選項設定畫面。在步驟二中選入欲繪製散佈圖的變數，亦可選擇是否設定分類變數，例如此例中我們選擇 CVD 作為分類變數，則在輸出圖形中將以顏色區分不同分類的圖點。在進階選項中，可另行依使用者需求更動圖點顏色、圖點符號、主標題、雙軸範圍等設定。所有參數設定無誤後，點選【繪製圖形】即可開始繪製。

散佈圖繪製結果如下頁圖所示。由圖可大致看出心臟收縮壓(SysBP)與心臟舒張壓(DiaBP)大致呈線性相關，且若以個人血管疾病史(CVD)作分層檢視，在相同舒張壓水平下，曾患有心血管疾病者有出現部分觀察值為收縮壓偏高的現象。



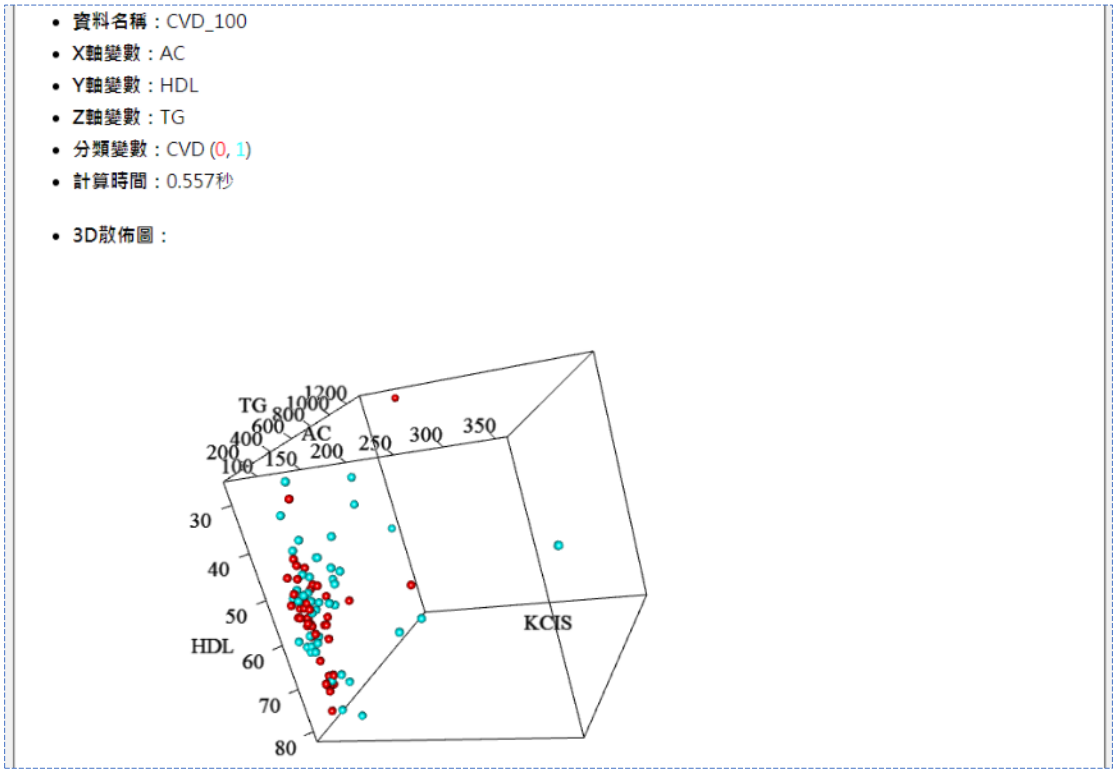
➢ 3D 散佈圖

若研究人員想同時觀察空腹葡萄糖、高密度脂蛋白、三酸甘油酯三者之間的相關性，那麼前面所介紹僅以二維平面呈現的散佈圖就顯得不敷使用，此時 3D 散佈圖便派上用場了！

3D 散佈圖提供繪製人類視覺上最高維度(三度空間，3-dimension)的散佈圖，使用者可利用同時以三個變數所繪製的圖形解釋資料，大幅提升我們了解資料特徵的能力。在 R-web 中使用時可依序點選主選單中【圖表繪製】→【散佈圖】→【3D 散佈圖】進行繪製。

下頁圖例為 3D 散佈圖之操作步驟及進階選項設定說明。首先在步驟一中選擇資料檔後系統將自動帶出參數設定畫面，在此我們仍以抽樣資料檔作為示範。在步驟二中分別選入欲繪製散佈圖的 X、Y、Z 軸變數，並視需求選擇是否加入分類變數和進階選項中的圖點大小、主標題。所有參數設定完畢後，點選【繪製圖形】即可開始繪製。

下圖為 3D 散佈圖繪製結果，在 R-web 中還可以利用滑鼠滾輪調整圖形大小，或拖曳變換視角從不同角度認識資料喔！



本期的生統 eNews 礙於篇幅就介紹到此為止，此次我們分別向各位讀者介紹了 R-web 中的次數分配表、列聯表、莖葉圖、2D 散佈圖及 3D 散佈圖，希望大家能有所收穫。下一期的生統 eNews 將繼續為大家介紹其餘常用繪圖功能如曲線(面)圖、直方圖、長條圖、圓餅圖、盒鬚圖在 R-web 中的應用，那麼我們下回見囉！