R 資料分析應用：圖表繪製（一）

尤子芸　副統計分析師

上一期的生統 eNews 教大家安裝 R 軟體，介紹了 R 軟體的使用界面，以及描述性統計與平均數檢定等統計方法，接下來本期的生統 eNews 將為各位介紹如何運用 R 軟體執行最直觀的資料檢視方法：圖表繪製。

當面對一組結構未知的資料時，做資料分析前，需要瀏覽資料，此時利用統計圖表來快速理解資料的特性是相當重要的。圖表繪製又稱為「資料視覺化」，而資料視覺化的目的為使用簡單的方式，降低理解複雜的資料。我們可以利用資料視覺化這樣的方式，清楚且有效地呈現資料的分佈，即使不是統計專家也能看圖說故事，將雜亂的「資料」轉化為有用的「資訊」。

在本期的生統 eNews 中，我們將依序介紹使用 R 軟體繪製次數分配表、列聯表、莖葉圖、2D 散佈圖以及 3D 散佈圖。本章節統一使用來自於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生統課程資料檔中的資料說明檔。

- 範例資料檔案

檔案可從臺北醫學大學管理學院的生物統計研究中心網頁中的生統課程資料檔位置下載 http://biostat.tmu.edu.tw/index.php/course/tmudata，下載檔案名稱 "CVD_All" 源自於基隆地區整合篩檢計畫(KCIS)，其檔案格式為 CSV 檔。

程式碼-1

```
#讀入資料，命名為 CVD
CVD=read.csv("G:/CVD_all.csv")
#步驟 2 偵測資料中是有遺失值
na.fail(CVD)
#步驟 3 若資料中有遺失值，要先去除遺失值
CVD_ok=na.omit(CVD)
#查看讀入之檔案 CVD_ok
View(CVD_ok)
```

資料型態：

ID	心血管疾病	年齡	性別	追蹤時間	腰圍	收縮壓	舒張壓	空腹血糖	高密度脂蛋白	三酸甘油酯	檳榔	飲酒	家族病史	抽菸
1	0	51	1	1	81.00	138.0	87.0	194	47.0	517	0	1	0	1
2	0	52	1	1	79.00	98.0	66.0	101	59.0	186	0	1	0	1
3	0	50	1	3	86.50	135.0	97.0	90	46.0	153	0	1	0	1
4	0	47	1	5	84.00	117.5	88.5	88	50.0	201	0	0	0	0
6	1	55	1	3	94.00	191.0	135.0	200	44.0	995	0	1	0	0
7	0	53	1	4	67.00	134.5	93.0	148	54.5	220	0	1	0	0
8	0	48	1	4	87.00	135.5	97.5	98	61.9	112	0	0	1	0
9	0	51	1	3	74.00	118.0	75.0	75	45.0	169	0	1	0	1
11	0	73	1	1	90.00	131.0	76.0	160	36.0	110	0	0	0	0

在上一期的生統 eNews 中，已經介紹過如何匯入不同格式之檔案，詳細資訊請自行參閱。

- 次數分配表

次數分配表是常見的描述性統計方法，將類別資料依照其組別分組，或將數執行資料依照觀察值的大小分成若干組，計算每一組別的次數、相對次數，與百分比等資訊，以了解資料的分佈情形。

假設研究人員欲了解心血管疾病資料中年齡變項的頻率/次數分配情形，但在資料中年齡變項為連續資料，那麼我們就必須先將年齡變項做分組，此處示範以 10 歲為單位做資料分割，將其分為“小於 10 歲”、“10~20”、“20~30”、“30~40”、“40~50”、“50~60”、“60~70”、“70~80”、“80~90”，以及“超過 90 歲”，共 10 組，分組完後就可以進行次數分配表的製作。

在 R 軟體中，我們將使用 cut()函數來將連續資料轉換成類別資料，以進行次數分配表製作。

程式碼-2

```
#步驟 1. 對年齡變項做分組
CVD_ok$age_gruop=cut(CVD_ok$年齡,
breaks = c(-Inf,10,20,30,40,50,60,70,80,90, Inf), #Inf 為無限值
labels = c("小於 10", "10~20","20~30","30~40","40~50","50~60","60~70",
"70~80","80~90","超過 90"))
    #cut(x, breaks=分組的切點, lables=分組標籤名稱)
#步驟 2. 計算年齡分組次數分配表
n=length(CVD_ok) #計算資料總筆數
f=table(CVD_ok$age_gruop) #table 函數為計算變項之次數分配
rf=f/sum(f) #計算相對次數(relative frequency)
cf=cumsum(f) #計算累計次數
crf=cf/sum(f) #計算累計相對次數
table_agegroup=t(rbind(f, cf, rf,crf)) #使用 rbind 函數將上面計算的數值合併，
再用 t 函數將橫與列欄位做轉置
#查看製作完成的次數分配表：
table_agegroup
```

	f	cf	rf	crf
小於10	0	0	0.000000000	0.000000000
10~20	322	322	0.005649123	0.005649123
20~30	6536	6858	0.114666667	0.120315789
30~40	14274	21132	0.250421053	0.370736842
40~50	15723	36855	0.275842105	0.646578947
50~60	9699	46554	0.170157895	0.816736842
60~70	7181	53735	0.125982456	0.942719298
70~80	3265	57000	0.057280702	1.000000000
80~90	0	57000	0.000000000	1.000000000
超過90	0	57000	0.000000000	1.000000000

- 列聯表

列連表為根據兩個(或以上)的類別變數繪製而成的次數分配表。當選擇多個分層變數時，又稱為多維列聯表。

假設研究人員想了解 KCIS 此筆資料中，參與研究者之個人心血管疾病史其性別之分布，那麼我們就來製作一個使用個人心血管疾病史作為行變數，性別作為列變數繪製列聯表。

程式碼-3.1

```
table_cvd_gender= table(CVD_ok$性別, CVD_ok$心血管疾病)
#table(列變數, 行變數)
```

程式碼-3.2

```
colnames(table2)=c("CVD_no","CVD_yes") #行變數命名
rownames(table2)=c("Female","Male") #列變數命名
#查看列聯表
table_cvd_gender
```

```
        CVD_no CVD_yes
Female   33205    3053
Male     18851    1891
```

➢ 莖葉圖

莖葉圖也是一種呈現資料分布結構的方法，其特色在於呈現方式類似直方圖，卻又能保留原始數據資料。除了可以看出像直方圖一般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，對於資料量不大的狀況下尤其適用。

由於在資料量較大的情況下，並不建議使用莖葉圖，因此我們匯入原資料(CVD_all.CSV)的前 100 筆樣本作為此處的範例資料。

程式碼-4

```
#取 100 筆樣本
cvd_100=read.table("I:/我的雲端硬碟/201806_enews/CVD_All.csv", header =
TRUE, sep = ",",nrows = 100)
#繪置莖葉圖
stem(cvd_100$腰圍) #stem(x)
```

```
  6 | 7
  7 | 0013444
  7 | 56778899999
  8 | 000011111122333333333344444444
  8 | 55556666677778888888999999
  9 | 0000022233444
  9 | 6678889
 10 | 33
 10 | 59
```

假設研究人員想知道此 100 筆樣本中，腰圍的資料分布情況，故繪製了上圖的莖葉圖，由圖所示，「|」代表莖葉的分界，且 R 軟體中的預設單位為 1，所以圖中的 6|7 即表示 67，7|0 即表示 70，以此類推。

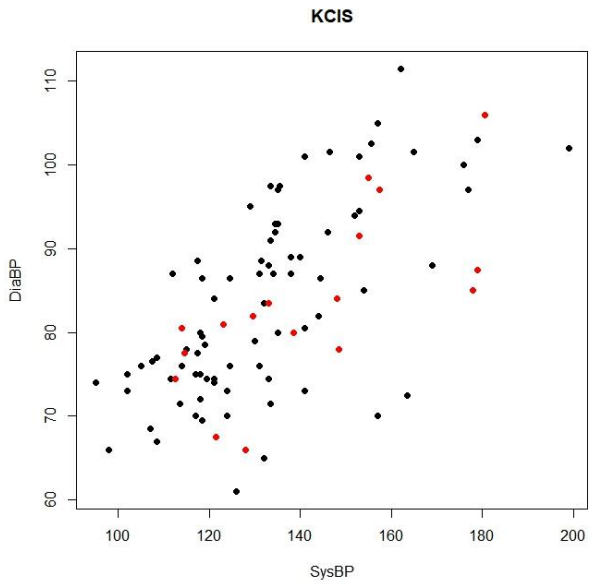
➢ 2D 散佈圖

2D 散佈圖可用來將兩個可能相關的數值變項分別置於座標圖上的 X 軸與 Y 軸，用圓點標示每個資料點的位置，可初步觀察兩變數之間的相關性。為了能讓散佈圖能較清楚的表示出每個資料點的位置，此處將使用 100 筆樣本的範例資料。

假設研究人員想知道參與研究者之個人心血管疾病史作為分層，心臟收縮壓(SysBP)與心臟舒張壓(DiaBP)之相關性。

程式碼-5

```
#繪製 2D 散佈圖
plot(cvd_100$收縮壓[cvd_100$心血管疾病==0],
     cvd_100$舒張壓[cvd_100$心血管疾病==0],pch=19,col="Black",
     xlab = "SysBP",ylab = "DiaBP",main="KCIS")
     #plot(x, y, pch=符號, col=顏色, xlab=x 軸名稱, ylab=y 軸名稱, main=散佈圖名稱)
points(cvd_100$收縮壓[cvd_100$心血管疾病==1],
       cvd_100$舒張壓[cvd_100$心血管疾病==1],pch=19,col="Red")
     #在繪製的散佈圖中加入其他資料點 point(x, y, pch=符號, col=顏色)
```



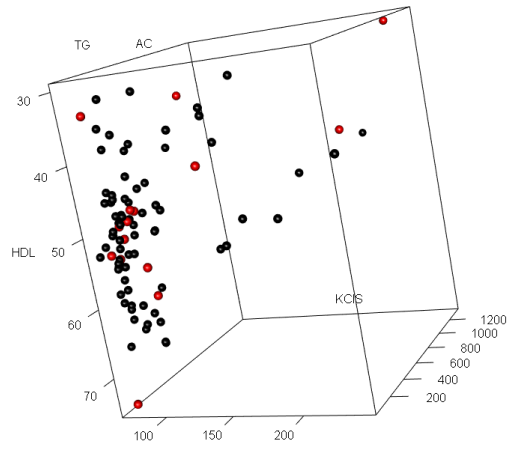
從上圖可大致看出心臟收縮壓(SysBP)與心臟舒張壓(DiaBP)大致呈現性關係，且以個人心血管疾病史(CVD)做分層檢視，在相同的舒張壓水平下，曾患有心血管疾病者有出現部分觀察值為收縮壓偏高的現象。

➢ 3D 散佈圖

若研究人員想同時觀察空腹血糖、高密度脂蛋白、三酸甘油脂三者之間的相關性，那麼前面所介紹僅以二維平面呈現的散佈圖就會顯得不敷使用，此時，我們就能使用 3D 散佈圖來了解資料分佈的趨勢了。3D 散佈圖提供繪製視覺上最高維度(三度空間，3-dimension)的散佈圖，以提供使用者可以同時利用三個變數所繪製之圖形解釋資料，大幅提升了解資料特徵之能力。進行繪製 3D 散佈圖之前，必須先安裝"rgl" package。

程式碼-6

```
#安裝"rgl"
install.packages("rgl")
library("rgl")
#繪製 3D 散佈圖
plot3d(cvd_100$空腹血糖[cvd_100$心血管疾病==0],
cvd_100$高密度脂蛋白[cvd_100$心血管疾病==0],
cvd_100$三酸甘油酯[cvd_100$心血管疾病==0],
xlab="AC",ylab="HDL",zlab="TG",col="Black",type="s",size=1,main="KCIS")
#plot3d(x, y, z, xlab=x 軸名稱, ylab=y 軸名稱, zlab=z 軸名稱, col=顏色, type=
資料點形狀, size= 資料點大小, main=3D 散佈圖名稱)
plot3d(cvd_100$空腹血糖[cvd_100$心血管疾病==1],
cvd_100$高密度脂蛋白[cvd_100$心血管疾病==1],
cvd_100$三酸甘油酯[cvd_100$心血管疾病==1],
col="Red",type="s",size=1,add=T)
#plot3d(x, y, z, col=顏色, type= 資料點形狀, size= 資料點大小, add=將此資料
點加入繪製好的 3D 散佈圖)
```



上圖為 3D 散佈圖的繪製結果，在 R 軟體中可以利用滑鼠滾輪調整圖大小，或拖曳變換視角從不同角度認識資料。

- 參考資料

1. R 軟體 應用統計方法 陳景祥編著 東華書局
2. Package 'rgl' - CRAN-R - R Project
   https://cran.r-project.org/web/packages/rgl/rgl.pdf