R-web 資料分析應用：相關暨列聯表分析 – 相關係數

蔡靜雯 副統計分析師

生統eNews【雲端資料分析暨導引系統】（R-web, http://www.r-web.com.tw）截至目前為止，介紹了圖表繪製和多種資料特性(平均數、中位數、變異數...)的檢定方法，本期將接著介紹分析方法中的「相關暨列聯表分析-相關係數」。欲了解兩個連續型變數間的關係，除了可以用散佈圖來表達，還可以用數值性指標來衡量兩個變數間的相關程度和其相關方向(正相關或負相關)，常用的相關性指標有皮爾生相關係數(Pearson's correlation coefficient) 和斯皮爾曼等級相關係數(Spearman's rank correlation coefficient)，以下將分別對這兩種相關係數做介紹，並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，示範相關係數的使用方法，有關此資料的詳細資訊及變數定義請參閱首期生統eNews。

➢ 皮爾生相關係數(Pearson's correlation coefficient)

皮爾生相關係數是用來測量兩連續型變數 X 和 Y 之間的**線性關係**。母體的相關係數通常以 ρ 表示，其定義為

$$\rho = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}} \quad ,$$

其中，$Cov(X,Y) = E(XY) - E(X)E(Y)$。在實務的應用上，母體的相關係數 ρ 通常都是未知的，便用樣本皮爾生相關係數 r_{xy} 來估計，其定義為

$$r_{xy}=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}}$$。

皮爾生相關係數特性：

1. $-1<r_{xy}<1$：由於母體相關係數ρ是以樣本相關係數r_{xy}來估計，若$|\rho|$真正值等於 1，$|r_{xy}|$的值雖然未必等於 1，但會很靠近 1。因此，當$|r_{xy}|$很接近 1 時，便可接受 X 和 Y 之間具有線性關係存在。$|r_{xy}|$越接近 1，則表示 X 和 Y 兩變數間的直線關係越強。
2. $r_{xy}=1$：表示 X 和 Y 兩變數間有完全正相關的線性關係。

   $r_{xy}=-1$：表示 X 和 Y 兩變數間有完全負相關的線性關係。

   $r_{xy}>0$：表示 X 和 Y 兩變數間存在正相關的線性關係。

   $r_{xy}<0$：表示 X 和 Y 兩變數間存在負相關的線性關係。

   $r_{xy}=0$或很接近 0：則表示 X 和 Y 兩變數間不具有線性關係，但並不代表沒有其他關係存在

除了計算出樣本相關係數來估計母體相關係數外，通常會對檢定母體相關係數是否為 0 感到興趣。若兩變數的樣本資料所來自的母體為常態分配，則虛無假設和對立假設分別為$H_0:\rho=0$ vs. $H_0:\rho\neq 0$，在虛無假設下，其檢定統計量為

$$T=r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}}\sim t_{n-2}$$，

其中$n-2$為t分配的自由度，在顯著水準為α下，拒絕域為

$C=\left\{|T|>t_{\frac{\alpha}{2}}(n-2)\right\}$，或用 p 值檢定方法 $p=2\times P\left(|T|>t_{\frac{\varepsilon}{2}}(n-2)\right)$，決定是否拒絕虛無假設。

➢ 斯皮爾曼等級相關係數(Spearman's rank correlation coefficient)

斯皮爾曼等級相關係數是依據 X 和 Y 兩變數資料，分別依大小排序後的兩列成對等級(rank)，再以各對等級差來進行計算，是一種無母數方法，其定義為

$$r_s=\frac{\sum_{i=1}^{n}\left(R_{X_i}-\overline{R}_X\right)\left(R_{Y_i}-\overline{R}_Y\right)}{\sqrt{\sum_{i=1}^{n}\left(R_{X_i}-\overline{R}_X\right)^2\sum_{i=1}^{n}\left(R_{Y_i}-\overline{R}_Y\right)^2}}$$

其中，R_{X_i} R_{Y_i} 分別為兩變數資料的等級，$\overline{R}_X$ $\overline{R}_Y$ 分別為兩變數等級的平均值。

斯皮爾曼等級相關係數特性：

1. 斯皮爾曼等級相關係數 r_s 界於 1 和 −1 之間。

2. $r_s=1$：表示 X 和 Y 兩變數完全正相關。

   $r_s=-1$：表示 X 和 Y 兩變數完全負相關。

   $r_s>0$：表示 X 和 Y 兩變數間存在正相關。

   $r_s<0$：表示 X 和 Y 兩變數間存在負相關。

   $r_{xy}=0$ 或很接近 0：則表示 X 和 Y 兩變數間不具有相關性。

若兩變數的樣本資料所來自的母體不是常態分配，或是資料中具有極端值，或是一個變數會隨著另一個變數增加(減少)的趨勢為非線性相關，此時，就適合用斯皮爾曼等級相關係數來對母體相關係數是否為 0 做檢

定。其虛無假設和對立假設分別為 $H_0:\rho=0$ vs. $H_0:\rho\neq 0$，在虛無假設下，其檢定統計量為

$$T=\frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}\sim t_{n-2}\ ,$$

其中 $n-2$ 為 t 分配的自由度，在顯著水準為 α 下，拒絕域為 $C=\left\{|T|>t_{\frac{\alpha}{2}}(n-2)\right\}$，或用 p 值檢定方法 $p=2\times P\left(|T|>t_{\frac{\varepsilon}{2}}(n-2)\right)$，決定是否拒絕虛無假設。

➢ 範例應用與 R-web 操作方式

皮爾生相關係數

想了解 KCIS 範例資料檔中，年齡和腰圍是否存在相關性？若有相關，其相關程度是高還是低？相關方向為何？

在 R-web 主選單中依序點選【分析方法】→【相關暨列聯表分析】→【皮爾生相關係數】即可進入分析頁面。

操作畫面如上圖所示。首先，在步驟一：資料匯入的地方選擇要進行分析的資料檔，點選”使用者個人資料檔”後選擇”CVD”(KCIS 範例資料檔)的檔案。接著，在步驟二：參數設定中，選擇要進行分析的變數”Age”(年齡)、”Waist”(腰圍)。

接著，點選【進階選項】如下圖，選擇檢定方向設定為”雙尾”；若有需要敘述統計量的資訊，可勾選顯示樣本敘述統計量，點選後儲存設定，即可【開始分析】。

下圖為分析結果，左上方可以看到檢定的變數及相關設定，檢查沒問題後即可看分析結果。第一個表格為樣本”Age”和”Waist”的敘述統計量；第二個表格顯示皮爾生相關係數矩陣，每一個格子內的值依序為皮爾生相關係數、P-值和樣本數，”Age”和”Waist”的相關性資訊可看矩陣的右上方或左下方的格子，這裡”Age”和”Waist”的皮爾生相關係數為 0.347，表示年齡和腰圍存在低度正相關。P-值為 0，可拒絕虛無假設，表示資料中年齡和腰圍的相關係數顯著不為 0。

皮爾生相關係數 - 分析結果

- **分析方法**：皮爾生相關係數
- **資料名稱**：CVD
- **變數名稱**：Age, Waist
- **虛無假設**：相關係數ρ = 0 (雙尾檢定)
- **計算時間**：0.382秒

- **樣本敘述統計量**[I]：

變數名稱 Variable	樣本數 Count	平均數 Mean	中位數 Median	最小值 Minimum	最大值 Maximum	標準差 Std. dev.
Age	64484	46.82	45	19	80	13.8959
Waist	62852	78.3391	78	37	179	10.6747

I：樣本敘述統計量皆不包含遺失值

- **皮爾生相關係數矩陣**[I]：

	Age	Waist
Age	1.000 0.000 64484	0.347 0.000 62847
Waist	0.347 0.000 62847	1.000 0.000 62852

I：表格內容為皮爾生相關係數 / P-值 / 樣本數

若要同時看多個變數間的相關程度，例如：同時看年齡與腰圍、心臟收縮壓、心臟舒張壓、空腹葡萄糖、高密度脂蛋白和三酸甘油酯的相關性，可在步驟二：參數設定中，同時選擇多個要進行分析的變數 Age、Waist、SysBP、DiaBP、AC、HDL 和 TG，操作畫面如下圖所示

分析結果如下圖，年齡與其他變數的皮爾生相關係數，除了”HDL”(高密度脂蛋白)為負相關，其他變數皆為正相關。*p* 值都很小，可拒絕虛無假設，表示資料中年齡與腰圍、心臟收縮壓、心臟舒張壓、空腹葡萄糖、高

密度脂蛋白和三酸甘油酯的相關係數皆顯著不為 0。

皮爾生相關係數 - 分析結果

- 分析方法：皮爾生相關係數
- 資料名稱：CVD
- 變數名稱：Age, Waist, SysBP, DiaBP, AC, HDL, TG
- 虛無假設：相關係數ρ = 0 (雙尾檢定)
- 計算時間：0.23秒

- 皮爾生相關係數矩陣[I]：

	Age	Waist	SysBP	DiaBP	AC	HDL	TG
Age	1.000	0.347	0.420	0.256	0.220	-0.012	0.129
	0.000	0.000	0.000	0.000	0.000	0.002	0.000
	64484	62847	63251	63240	60973	60079	60886
Waist	0.347	1.000	0.426	0.399	0.200	-0.399	0.323
	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	62847	62852	62383	62376	59651	59574	59563
SysBP	0.420	0.426	1.000	0.743	0.191	-0.163	0.219
	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	63251	62383	63256	63205	59992	59620	59904
DiaBP	0.256	0.399	0.743	1.000	0.130	-0.172	0.220
	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	63240	62376	63205	63245	59977	59607	59889
AC	0.220	0.200	0.191	0.130	1.000	-0.108	0.235
	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	60973	59651	59992	59977	60978	60064	60867
HDL	-0.012	-0.399	-0.163	-0.172	-0.108	1.000	-0.359
	0.002	0.000	0.000	0.000	0.000	0.000	0.000
	60079	59574	59620	59607	60064	60084	59976
TG	0.129	0.323	0.219	0.220	0.235	-0.359	1.000
	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	60886	59563	59904	59889	60867	59976	60891

I：表格內容為皮爾生相關係數 / P-值 / 樣本數

斯皮爾曼等級相關係數

沿用 KCIS 範例資料檔中相同的變數，使用斯皮爾曼等級相關係數看年齡與腰圍、心臟收縮壓、心臟舒張壓、空腹葡萄糖、高密度脂蛋白和三酸甘油酯的相關性。

在 R-web 主選單中依序點選【分析方法】→【相關暨列聯表分析】→【斯皮爾曼等級相關係數】即可進入分析頁面。

操作畫面如上圖所示。同樣，在步驟一：資料匯入的地方選擇要進行分析的資料檔，點選”使用者個人資料檔”後選擇”CVD”(KCIS 範例資料檔)的檔案。接著，在步驟二：參數設定中，選擇要進行分析的變數”Age”(年齡)、”Waist”(腰圍)。”SysBP”(心臟收縮壓)、”DiaBP”(心臟舒張壓)、”AC”(空腹葡萄糖)、”HDL”(高密度脂蛋白)和”TG”(和三酸甘油酯)。

接著，點選【進階選項】如下圖，選擇檢定方向設定為”雙尾”；若有需要敘述統計量的資訊，可勾選顯示樣本敘述統計量，點選後儲存設定，即可【開始分析】。

分析結果如下圖，同樣先確認左上方欲檢定的變數及相關設定是否正確，檢查沒問題後即可看分析結果，斯皮爾曼等級相關係數結果和皮爾生相關係數分析結果差不多，年齡與其他變數的斯皮爾曼等級相關係數，除了”HDL”(高密度脂蛋白)為負相關，其他變數皆為正相關。*p* 值都很小，可

拒絕虛無假設，表示資料中年齡與腰圍、心臟收縮壓、心臟舒張壓、空腹葡萄糖、高密度脂蛋白和三酸甘油酯的相關係數都顯著不為0。

斯皮爾曼等級相關係數 - 分析結果

- 分析方法：斯皮爾曼等級相關係數
- 資料名稱：CVD
- 變數名稱：Age, Waist, SysBP, DiaBP, AC, HDL, TG
- 虛無假設：相關係數ρ = 0 (雙尾檢定)
- 計算時間：2.216秒

- 樣本敘述統計量[I]：

變數名稱 Variable	樣本數 Count	平均數 Mean	中位數 Median	最小值 Minimum	最大值 Maximum	標準差 Std. dev.
Age	64484	46.82	45	19	80	13.8959
Waist	62852	78.3391	78	37	179	10.6747
SysBP	63256	123.2666	120.5	70	276	20.8228
DiaBP	63245	78.0701	77	40	140	11.9757
AC	60978	93.1598	87	49	606	28.9281
HDL	60084	57.3059	57	10	154	12.1861
TG	60891	121.073	92	11	4137	111.0751

I：樣本敘述統計量皆不包含遺失值

- 斯皮爾曼相關係數矩陣[I]：

	Age	Waist	SysBP	DiaBP	AC	HDL	TG
Age	1.000	0.364	0.421	0.284	0.267	-0.033	0.248
	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	64484	62847	63251	63240	60973	60079	60886
Waist	0.364	1.000	0.458	0.418	0.232	-0.450	0.481
	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	62847	62852	62383	62376	59651	59574	59563
SysBP	0.421	0.458	1.000	0.742	0.255	-0.210	0.322
	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	63251	62383	63256	63205	59992	59620	59904
DiaBP	0.284	0.418	0.742	1.000	0.172	-0.198	0.309
	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	63240	62376	63205	63245	59977	59607	59889
AC	0.267	0.232	0.255	0.172	1.000	-0.154	0.238
	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	60973	59651	59992	59977	60978	60064	60867
HDL	-0.033	-0.450	-0.210	-0.198	-0.154	1.000	-0.456
	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	60079	59574	59620	59607	60064	60084	59976
TG	0.248	0.481	0.322	0.309	0.238	-0.456	1.000
	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	60886	59563	59904	59889	60867	59976	60891

I：表格內容為斯皮爾曼等級相關係數 / P-值 / 樣本數

皮爾生相關係數和斯皮爾曼等級相關係數的比較

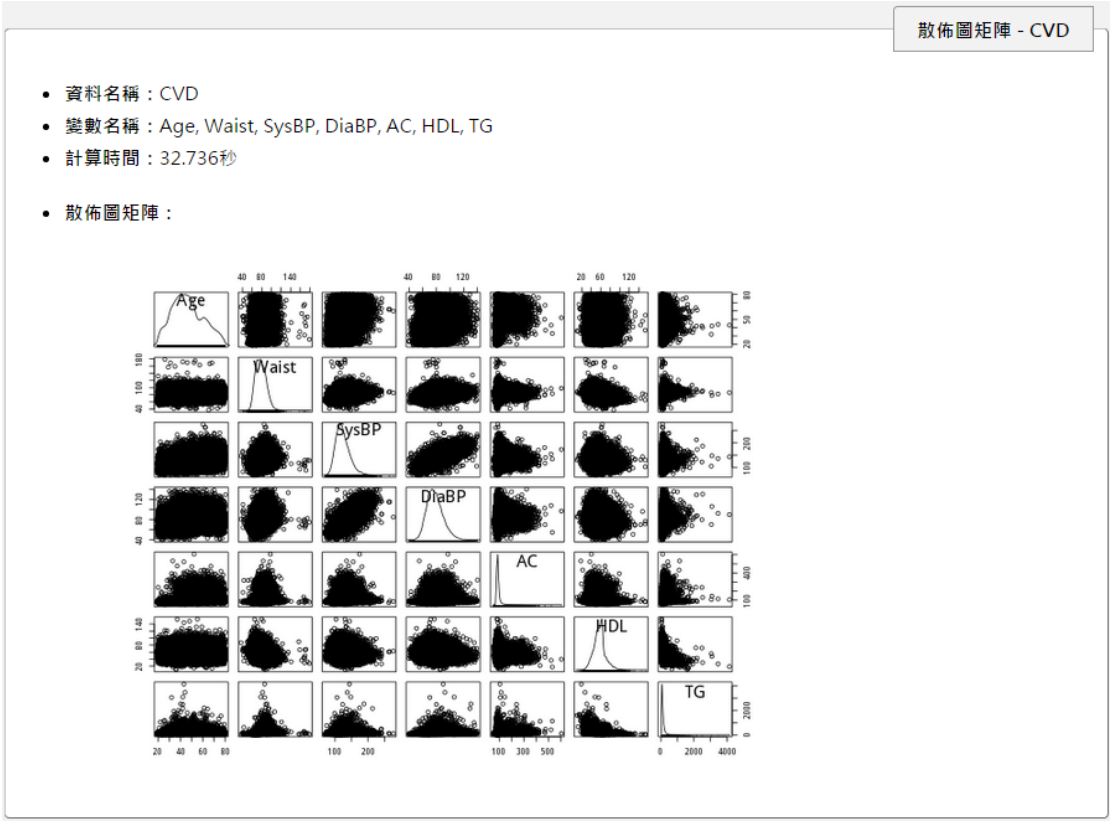
- 皮爾生相關係數矩陣[I]：

		Age	Waist	SysBP	DiaBP	AC	HDL	TG
Age		1.000	0.347	0.420	0.256	0.220	-0.012	0.129
		0.000	0.000	0.000	0.000	0.000	0.002	0.000
		64484	62847	63251	63240	60973	60079	60886

- 斯皮爾曼相關係數矩陣[I]：

		Age	Waist	SysBP	DiaBP	AC	HDL	TG
Age		1.000	0.364	0.421	0.284	0.267	-0.033	0.248
		0.000	0.000	0.000	0.000	0.000	0.000	0.000
		64484	62847	63251	63240	60973	60079	60886

從以上兩個圖表可以看出，年齡和其他六個變數的皮爾生相關係數值和斯皮爾曼等級相關係數值，除了”AC”和”TG”這兩個變數相差稍微比較大，其他變數的差異都很小，有可能是”AC”和”TG”這兩個變數不為常態分配或有極端值。運用前幾期介紹過的散佈圖，回頭看這幾個變數的資料分佈情況。從下圖結果可以清楚看到，”AC”和”TG”這兩個變數的分配的確是比較不符合常態分配的圖形。建議在做相關性分析前，可以先畫散佈圖，從散佈圖的大概情況來初步決定要用皮爾生相關係數或斯皮爾曼等級相關係數分析。皮爾生相關係數主要是測量符合常態分配下的兩變數間是否有線性關係，當兩變數間有相關，但資料不符合常態或非線性關係或是有極端值，此時斯皮爾曼等級相關係數就是較為適合的一個方式。



本期生統 eNews 的介紹到此，這次介紹了兩種相關係數以衡量連續型變數間的相關性，兩種相關係數有不同的使用時機，希望大家能清楚了解且能更加熟練操作方式，根據不同的資料型態找到合適的分析方法。下一期生統 eNews 將為大家介紹分析方法中的「相關暨列聯表分析-檢定方法」，更深入探討變數間的關係，敬請期待！

參考資料

1. 華泰書局，現代統計學 第十章 相關係數
2. Higgins, *Introduction to Modern Nonparametric Statistics*, 1st Edition. 153-158
3. Woodward, M(2004).*Epidemiology-Study Design and Data Analysis*, 2nd Edition.Chapman & Hall/CRC, London. 456-459
4. 斯皮爾曼等級相關（Spearman Rank Correlation）

http://amebse.nchu.edu.tw/new_page_517.htm

5. Pearson 相關係数和 Spearman 秩相關系数介绍

   http://wenku.baidu.com/view/ad01681cb7360b4c2e3f64fd.html

6. 皮爾森相關係數與斯皮爾曼相關係數。

   http://wenku.baidu.com/view/ad01681cb7360b4c2e3f64fd.html