R-web 資料分析應用：相關暨列聯表分析–列聯表檢定方法

陳逸萱 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） 分析方法中的『相關暨列聯表分析-相關係數』功能。『相關係數』主要用來衡量兩個連續型變數間的線性關聯性高低，但若資料為"類別型變數"，則無法用相關係數來評估。接下來，本期的生統 eNews 將跟大家介紹：檢定兩個類別型變數間是否存在關聯性的『列聯表檢定方法』。

若我們想觀察兩類別變數之間的關聯性，我們可以先將資料整理成『列聯表（Contingency Table）』的形態。假設A類別變數有r個分組，B類別變數有c個分組，計算資料中在此兩個變數產生的$r \times c$個類別組合的樣本次數，即可構成$r \times c$列聯表。列聯表檢定方法依據樣本的特性不同，可分為：卡方獨立性(或稱齊一性)檢定、費雪精確檢定、McNemar檢定，本期的生統eNews將依序跟大家介紹這些方法的應用。本系列分析將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。

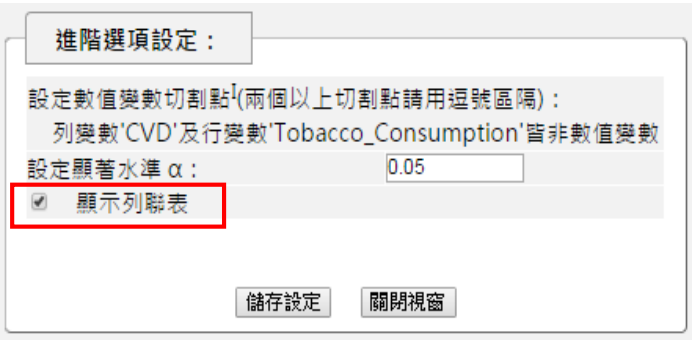
➢ 卡方獨立性檢定（Wilcoxon signed-rank test）

當我們想評估資料中兩類別變數的關聯性，且資料樣本數較大時，即可使用『卡方獨立性檢定』。此方法的概念在比較列聯表中觀察次數和期望次數是否有差異，若兩變數獨立時，觀察次數和期望個數應很接近。以範例資料檔為例，在我們篩選其中有抽菸的族群資料中，"CVD"（個人心血管疾病史）為兩組分類的類別變數，"Tobacco_Consumption"（菸草消

費量）為三組分類的類別變數，我們便可來檢定資料檔中是否罹患心血管疾病與菸草消費量分組是否存在關聯性。

在 R-web 主選單中依序點選【分析方法】→【相關暨列聯表分析】→【卡方獨立性(或稱齊一性)檢定】即可進入分析頁面。

操作畫面如上圖所示。第一步，先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_tobacco”的檔案（篩選好的吸菸者資料），系統將自動帶出參數設定畫面。在步驟二選擇要進行分析的變數，在此設定列變數為”CVD”（個人心血管疾病史）、行變數為”Tobacco_Consumption”（菸草消費量）。最後，點選【進階選項】如右圖，勾選”顯示列聯表”，分析結果便會呈現整理好的列聯表資料，【儲存設定】後即可【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，檢查沒問題即可往下看分析結果。第一個表格為整理好的2×3列聯表；第二個表格顯示檢定統計量與 p 值，本分析之虛無假設為兩變數之間無關聯，而 p-值 0.027441*表顯著，拒絕虛無假設，我們可推論資料中是否罹患心血管疾病與菸草消費量的高低分組有關。在分析結果的列聯表中，藍色框框圈出了各個菸草消費量分組罹患心血管疾病的比例，除了檢定結果告訴我們這個比例在各個菸草消費量分組的分布不同以外，我們還可以觀察到菸草消費量越高的分組（1：每日一包、2：每日兩包、3：每日三包以上），其罹患心血管疾病的比例越高，根據這個現象，研究者可以嘗試再做進一步的分析。

卡方獨立性(或稱齊一性)檢定 - 分析結果

- 分析方法：卡方獨立性(或稱齊一性)檢定
- 資料名稱：cvd_tobacco
- 變數名稱：CVD, Tobacco_Consumption
- 顯著水準：0.05
- 計算時間：0.021秒

- 列聯表(CVD*Tobacco_Consumption)[I]：

		Tobacco_Consumption			合計 Total
		1	2	3	
CVD	0	13021 80.54 89.28 90.40	1420 8.78 9.74 88.97	144 0.89 0.99 85.71	14585
	1	1383 8.55 87.37 9.60	176 1.09 11.12 11.03	24 0.15 1.52 14.29	1583
合計 Total		14404	1596	168	16168

I：列聯表內容為觀察值個數 / 百分比 / 列百分比 / 行百分比

- 卡方獨立性(或稱齊一性)檢定：

虛無假設：兩變數之間無關聯		
卡方檢定統計量 chi-square statistics	自由度 d.f.	p-值[I] p-value
7.1914	2	0.027441 *

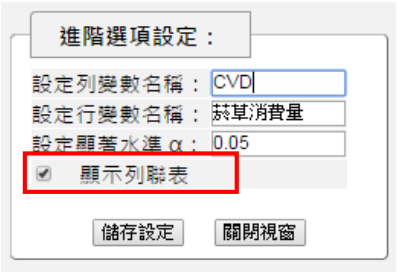
I：顯著性代碼： '***' : < 0.001, '**' : < 0.01, '*' : < 0.05, '#' : < 0.1

- 分析結果建議：由於檢定結果P-值為(0.027441) < 顯著水準0.05，因此可拒絕虛無假設。

➢ 費雪精確檢定（Fisher's exact test）

當資料樣本數較小（以樣本筆數<30 為區分標準）時，卡方獨立性檢定的 p 值較不可靠，此時我們可改用『費雪精確檢定』來檢定兩類別變數的關聯性。費雪精確檢定是透過"超幾何分配"的公式來檢定兩變數的相關性，比起卡方獨立性檢定較精確，但是樣本數很大時會耗費較久的運算時間。比照前面的例子，我們可以嘗試用費雪精確檢定來檢定是否罹患心血管疾病與菸草消費量分組是否存在關聯性，雖然此範例的樣本數夠大，我們仍可大略比較兩方法的差異。

在 R-web 主選單中依序點選【分析方法】→【相關暨列聯表分析】→【費雪精確檢定】即可進入分析頁面。



在此例中，我們可以透過前面得到的列聯表數值來進行分析，操作畫面如上圖所示。首先，選擇”以列聯表型態直接輸入資料”，並調整列聯表為:2 列*3 行,點選”輸入資料”後，系統將自動帶出列聯表的空白格式。接者，將列聯表中兩變數的類別項目名稱與對應觀察個數填入，完成後點選【進階選項】如左圖，在此依據自己需求設定行、列變數名稱，勾選”顯示列聯表”，分析結果便會呈現整理好的列聯表資料,【儲存設定】後即可【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，檢查沒問題即可往下看分析結果。第一個表格為2 × 3列聯表；第二個表格顯示費雪精確檢定的 p 值，本分析之虛無假設為兩變數之間無關聯，而 p-值 0.028289*表顯著，拒絕虛無假設，我們可推論資料中是否罹患心血管疾病與菸草消費量的高低分組有關。此分析結果與前面卡方獨立性檢定的趨勢

相同，我們可知在大樣本的情況下，兩方法可得到相同的結論。

費雪精確檢定 - 分析結果

- 分析方法：費雪精確檢定
- 資料名稱：自行輸入資料
- 變數名稱：CVD, 菸草消費量
- 顯著水準：0.05
- 計算時間：0.016秒

- 列聯表(CVD*菸草消費量)[I]：

		菸草消費量			合計 Total
		1	2	3	
CVD	無CVD	13021 80.54 89.28 90.40	1420 8.78 9.74 88.97	144 0.89 0.99 85.71	14585
	有CVD	1383 8.55 87.37 9.60	176 1.09 11.12 11.03	24 0.15 1.52 14.29	1583
合計 Total		14404	1596	168	16168

I：列聯表內容為觀察值個數 / 百分比 / 列百分比 / 行百分比

- 費雪列聯表檢定：

虛無假設：兩變數之間無關聯
p-值[I] p-value
0.028289 *

I：顯著性代碼： '***' : < 0.001, '**' : < 0.01, '*' : < 0.05, '#' : < 0.1

- 分析結果建議：由於檢定結果P-值為(0.028289) < 顯著水準0.05，因此可拒絕虛無假設。

➢ McNemar 檢定（McNemar's test）

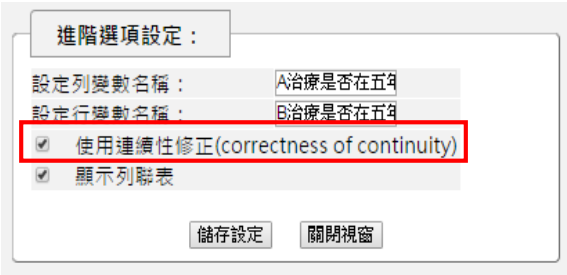
當我們想比較類別為兩類的配對(matched pairs)資料，我們可以將資料轉換為成對資料的列聯表，並用『McNemar 檢定』進行分析。由於範例資料並非配對資料，在這邊我們改用生統教科書中的例子[1]來說明：某一臨床試驗欲比較 A 和 B 兩種乳癌化療藥物的療效，納入了 621 對經過年齡配對的乳癌病人（共 1242 人），分別給予 A 藥和 B 藥的治療，而後觀察這

些病人五年的存活狀況，觀察的結果整理成下表：有 90 對的病人無論進行 A 治療或 B 治療五年內皆死亡，而有 510 對的病人五年內皆存活；有 16 對的病人進行 A 治療者在五年內存活，但進行 B 治療者在五年內死亡；另有 5 對的病人進行 B 治療者在五年內存活，但進行 A 治療者在五年內死亡。

		進行 B 治療的病人		
		是否在五年內死亡		Total
進行 A 治療的病人		No	Yes	
是否在五年內死亡	No	510	16	526
	Yes	5	90	95
Total		515	106	621

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

操作畫面如上圖所示，先選擇”以列聯表型態直接輸入資料”，點選”輸



入資料”後，系統將自動帶出列聯表的空白格式。而後參考本例的成對列聯表，將兩變數的類別項目名稱與對應觀察個數填入，完成後點選【進階選項】如右圖，在此可依據自己需求設定行、列變數名稱及是否”顯示列聯表”，若樣本數較小或有細格(cell)數≤ 5時，建議勾選”使用連續性修正(correctness of continuity)”，【儲存設定】後即可【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。第一個表格為成對的2 × 2列聯表；第二個表格顯示 McNemar 檢定的 p 值，本分析之虛無假設為兩變數之間無關聯，而 p-值 0.029096*表顯著，拒絕虛無假設，我們可推論五年存活狀況與 A、B 治療種類有關。此資料中我們感興趣的為存活狀況不一致的配對，即下圖藍色框框圈出的 21（15 + 6）對病人，其中進行 A 治療者在五年內存活、但進行 B 治療者在五年內死亡的 16 對病人占多數，我們可以推論 A 治療的療效比 B 治療好。

McNemar檢定 - 分析結果

- 分析方法：McNemar檢定
- 資料名稱：自行輸入資料
- 變數名稱：A治療是否在五年內存活, B治療是否在五年內存活
- 計算時間：0.004秒
- 列聯表(A治療是否在五年內存活*B治療是否在五年內存活)[I]：

		B治療是否在五年內存活 No	B治療是否在五年內存活 Yes	合計 Total
A治療是否在五年內存活	No	510 82.13 96.96 99.03	16 2.58 3.04 15.09	526
	Yes	5 0.81 5.26 0.97	90 14.49 94.74 84.91	95
合計 Total		515	106	621

I：列聯表內容為觀察值個數 / 百分比 / 列百分比 / 行百分比

- McNemar檢定：

虛無假設：兩變數之間無關聯		
卡方檢定統計量[I] chi-square statistics	自由度 d.f.	p-值[II] p-value
4.7619	1	0.029096 *

I：使用連續性修正

II：顯著性代碼： '***' : < 0.001, '**' : < 0.01, '*' : < 0.05, '#' : < 0.1

本期生統 eNews 的介紹到此告一段落，這次介紹了列聯表檢定的三種方法：卡方獨立性(或稱齊一性)檢定、費雪精確檢定、McNemar 檢定，希望大家能更加熟悉這些檢定方法的使用時機與操作方式。下一期的生統 eNews 將為大家介紹更進階的分析方法－『迴歸分析』，敬請期待！

參考資料

1. Bernard Rosner(2010), *Fundamentals of Biostatistics*, 7th Edition. 373-377