卡方檢定

許明樹

臨床上當資料型態為類別變項V.S. 類別變項時使用，通常我們會將類別資料綜合表現於列聯表(Contingency table)，又稱交叉表(cross tabulation)，列聯表的表格用在計算各分類的發生次數，生物統計學上我們可用卡方檢定來檢定

以下情況：

一、獨立性檢定(test for independence)

檢定兩變數是否獨立(相關)。

二、齊一性檢定(test for homogeneity)

檢定抽自於k個不同母體的樣本，在一個類別變數的幾個不同類別中，其分配比例是否相同。

在 SAS 中計算卡方檢定語法如下：

```
PROC FREQ <options> ;
    BY variables;
    EXACT statistic-options </ computation-options> ;
    OUTPUT <OUT=SAS-data-set> output-options;
    TABLES requests </ options> ;
    TEST options;
    WEIGHT variable </ option> ;
```

列聯表(Contingency table)

為幫助大家學習如何使用，將以 SAS/STAT(R) 9.3 User's Guide The FREQ Procedure 中的作為例子，文中舉例將會 Example 3.1 提供的歐洲的兩個地區兒童的眼睛與頭髮的顏色資料修改後進行講解。

當你需要自行創造或是輸入資料時可參考以下語法：

```
data Color; /*新建一個資料，名稱為Color*/
input Region Eyes $ Hair $ Count @@;
    label Eyes   ='Eye Color'
          Hair   ='Hair Color'
          Region='Geographic Region';
    datalines;
1 blue  black  23  1 blue  red      7  1 blue  black 24
1 blue  black  11  1 green black   19  1 green red     7
2 green black  23  2 green black   56  2 green red    42
```

/*新建三個變項 文字後面須加上”$”字號，@@是在跟SAS說下面的資料輸入模式為Region Eyes $ Hair $ Count的循環，以空格做區隔*/

```
2 green black 53   2 green black    54   2 green black   13
1 green black 18   1 green black    14   1 green black    34
1 green red        5   1 green black 41   1 green black    40
1 green black    3   2 blue   black    46   2 blue   red      21
2 blue   black 44   2 blue   black    40   2 blue   black    6
2 green black    50   2 green red      31   2 green black 37
;
run; /* Geographic Region 分為 1, 2 兩區，eyes=眼睛顏色，hair=頭髮顏色，
count=次數，raw data 如 fig1.*/
```

Fig1. Raw data

	Geographic Region	Eye Color	Hair Color	Count
1	1	blue	black	23
2	1	blue	red	7
3	1	blue	black	24
4	1	blue	black	11
5	1	green	black	19
6	1	green	red	7
7	2	green	black	23
8	2	green	black	56
9	2	green	red	42
10	2	green	black	53
11	2	green	black	54
12	2	green	black	13
13	1	green	black	18
14	1	green	black	14
15	1	green	black	34
16	1	green	red	5
17	1	green	black	41
18	1	green	black	40
19	1	green	black	3
20	2	blue	black	46
21	2	blue	red	21
22	2	blue	black	44
23	2	blue	black	40
24	2	blue	black	6
25	2	green	black	50
26	2	green	red	31
27	2	green	black	37

```
proc freq data=Color;
    tables Eyes   Hair   Eyes*Hair ; /* 當想看列聯表時須於兩變項中加上 * 號*/
    weight Count;
run;
```

/* weight 是因為上述資料已經分組好並且歸好人數，如果你沒有下指令 weight 當你在 table 打 eyes 時 SAS 會去針對 eyes 變項有幾次去做計算(ex:blue 9 次，但下了 weight Count 時，SAS 將會去看 count 的次數去做計算，此時的 blue=222 次，如 fig2., fig3.所示，使用者也可以自行用 SAS 體驗其差異)*/

Fig2.有 weigh 的結果

次數
百分比
列百分比
欄百分比

Table of Eyes by Hair			
	Hair(Hair Color)		
Eyes(Eye Color)	black	red	總計
blue	194 25.46 87.39 29.89	28 3.67 12.61 24.78	222 29.13
green	455 59.71 84.26 70.11	85 11.15 15.74 75.22	540 70.87
總計	649 85.17	113 14.83	762 100.00

Fig.3 無 weight 的結果

次數
百分比
列百分比
欄百分比

Table of Eyes by Hair			
	Hair(Hair Color)		
Eyes(Eye Color)	black	red	總計
blue	7 25.93 77.78 33.33	2 7.41 22.22 33.33	9 33.33
green	14 51.85 77.78 66.67	4 14.81 22.22 66.67	18 66.67
總計	21 77.78	6 22.22	27 100.00

獨立性檢定(test for independence)

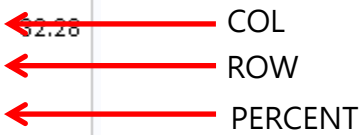
用於檢定兩變數是否獨立(相關)，在此以example 3.1提供的歐洲的兩個地區兒童眼睛顏色與地區是否有相關性為例進行分析與講解，假設檢定如下：

$$\begin{cases} H_0：兒童眼睛顏色與地區無相關 \\ H_1：兒童眼睛顏色與地區有相關 \end{cases}$$

```
proc freq data=Color order=data;
    tables Region*eyes /chisq;
/*fig.4中 COL(行的百分比)、ROW(列的百分比)、PERCENT(總百分比)如果不需要此類型的百分比結果，可以加入noCOL、noROW、noPERCENT，例如我不想要列的百分比，則語法為tables Region*eyes /chisq noROW; ，使用者可以自己體驗語法增減後的差異*/
    weight Count;
    run;
```

fig.4 Chi-Square

次數 百分比 列百分比 欄百分比	Table of Region by Eyes		
Region(Geographic Region)	Eyes(Eye Color)		
	blue	green	總計
1	65	181	246
	8.53	23.75	32.28
	26.42	73.58	
	29.28	33.52	
2	157	359	516
	20.60	47.11	67.72
	30.43	69.57	
	70.72	66.48	
總計	222	540	762
	29.13	70.87	100.00



表格 Eyes-Region*s 的統計值

統計值	自由度	值	機率
卡方	1	1.2933	0.2554
概度比卡方	1	1.3068	0.2530
連續性調整卡方	1	1.1066	0.2928
Mantel-Haenszel 卡方	1	1.2916	0.2558
Phi 係數		-0.0412	
列聯係數		0.0412	
Cramer V		-0.0412	

經過上述語法(**proc freq...**)後將會得到一個卡方的結果「Table of Region by Eyes」此為 Region, eyes 兩變項所組成的列聯表，「表格 Eyes-Region*s 的統計值」中統計值的卡方的機率(p-value)=0.2554，結論，居住地區與兒童眼睛顏色是沒有相關性的(p>0.05)。

齊一性檢定(test for homogeneity)

當檢定抽自於k個不同母體的樣本，在一個類別變數的幾個不同類別中，欲檢驗分配比例是否相同可使用齊一性檢定(test for homogeneity)，在此以example 3.1提供的歐洲的兩個地區兒童頭髮的顏色資料進行分析與講解，以下為我們的假設檢定：

$$\begin{cases} H_0：\text{相同地區之髮色分布情況一致} \\ H_1：\text{相同地區之髮色分布情況不一致} \end{cases}$$

```
proc sort data=Color; /*針對color這個data進行排序*/
    by Region;/*以Region此變項進行排序*/
run;
proc freq data=Color order=data;
    tables Hair / chisq testp=(50 50); /*testp中的總數需為100，但其中的比例可自行定義*/
    weight Count;
    by Region; /*結果將會分成region=1, region=2 兩塊*/
      /*此處的by，需要上述的proc sort，先進行sort完後才能夠正確的執行*/
run;   /*chisq 此時是檢定是否符合testp中各50%的百分比設定*/
```

Fig.5未進行排序的獨立性檢定

Geographic Region=1

Hair Color			
Hair	次數	百分比	檢定 百分比
black	77	84.62	50.00
red	14	15.38	50.00

卡方檢定 適用於指定的比例	
卡方	43.6154
DF	1
Pr > ChiSq	<.0001

Fig.6 error info.

```
138  proc freq data=Color order=data;
139     tables Hair / nocum chisq testp=(50 50);
140     weight Count;
141     by Region;
142  run;

ERROR: Data set WORK.COLOR is not sorted in 遞增 sequence. The current BY group has Geographic
       Region = 2 and the next BY group has Geographic Region = 1.
NOTE: SAS 系統已因為錯誤而停止處理此步驟。
NOTE: 已從資料集 WORK.COLOR. 讀取 13 個觀測值
NOTE: 已使用 PROCEDURE FREQ (總處理時間):
      實際時間           0.25 秒
      CPU 時間           0.07 秒
```

Fig.7 進行排序後的獨立性檢定

Geographic Region=1

Hair Color			
Hair	次數	百分比	檢定 百分比
black	227	92.28	50.00
red	19	7.72	50.00

卡方檢定 適用於指定的比例	
卡方	175.8699
DF	1
Pr > ChiSq	<.0001

Geographic Region=2

Hair Color			
Hair	次數	百分比	檢定 百分比
black	422	81.78	50.00
red	94	18.22	50.00

卡方檢定 適用於指定的比例	
卡方	208.4961
DF	1
Pr > ChiSq	<.0001

Fig5. 是沒有針對變項Region進行排序的結果，可以看到只有Geographic Region=1的結果總數只有91，且沒有Geographic Region=2的結果，如fig6. 所示因為沒有排序而導致sas出現錯誤，而當你進行排序(**proc sort...**)以後，執行此檢定，就可以順利地得到Geographic Region=1, 2的結果，**關鍵在於當你proc freq 的過程中如果要下by進行分組的結果時，要先對by的特定變項先進行排序**，使用者們可以觀察fig.1與fig.8 中 Geographic Region欄位的差異，齊一性檢定的結果在fig.7中的「卡方檢定其適用於指定的比例」中顯示，若Pr>Chisq為<0.05則是顯著的差異，觀察fig.6 中Geographic Region=1、Geographic Region=2的Pr>Chisq 為p<0.0001，顯示Geographic Region=1, 2 的結果皆不符合50%、50%的比例。

Fig8.排序後的raw data

	Geographic Region	Eye Color	Hair Color	Count
1	1	blue	black	23
2	1	blue	red	7
3	1	blue	black	24
4	1	blue	black	11
5	1	green	black	19
6	1	green	red	7
7	1	green	black	18
8	1	green	black	14
9	1	green	black	34
10	1	green	red	5
11	1	green	black	41
12	1	green	black	40
13	1	green	black	3
14	2	green	black	23
15	2	green	black	56
16	2	green	red	42
17	2	green	black	53
18	2	green	black	54
19	2	green	black	13
20	2	blue	black	46
21	2	blue	red	21
22	2	blue	black	44
23	2	blue	black	40
24	2	blue	black	6
25	2	green	black	50
26	2	green	red	31
27	2	green	black	37

以上是對於卡方檢定在SAS中應用的介紹，希望能對使用者在使用上有幫助。

Reference

1. SAS/STAT(R) 9.3 User's Guide The FREQ Procedure