12.卡方檢定
Chi-square Tests

鄒慶士 (Ching-Shih Tsou)
台北商業技術學院資訊與決策科學所
E-mail：cstsou@mail.ntcb.edu.tw

卡方檢定

- 類別資料的統計檢定
- 設Z_i為iid N(0,1)的隨機變數，因Z_i^2服從自由度1的卡方分配，故由卡方分配之相加性得知：

$$\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$$

- 當自由度越大時，卡方分配越接近常態分配
- 常見的卡方檢定類型
  - 適合度檢定(Chi-squared goodness of t tests)
  - 獨立性檢定(Chi-squared tests of independence)
  - 齊一性檢定(Chi-squared tests for homogeneity)

適合度檢定

- 利用樣本資料檢查母體是否為某一特定分布的統計方法
- 實驗中得到的次數稱為觀察次數: f_i
- 根據虛無假設推論出的次數稱為期望次數: e_i
- 卡方值計算如下

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

獨立性檢定

- 獨立性檢定是用來檢定兩個屬性間是否獨立的統計方法
- 檢定量

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

- r: 橫列個數，c: 縱行個數，f_ij: 樣本觀察次數
- e_ij: 估計理論次數，自由度=(r-1)(c-1)

齊一性檢定

- 齊一性檢定是檢定兩個或兩個以上母體的某一特性的分布(各類別的比例)是否齊一或相近
- 檢定量 $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$
- r:橫列個數，c:縱行個數，f_ij:樣本觀察次數
- e_ij: 估計理論次數，自由度=(r-1)(c-1)

Some Insights

- 若樣本數由100 增加為500，且在各組之次數依比例(5 倍)加大，則卡方值將增大5 倍。因此可知，若樣本數增加，將使卡方值加大，而卡方值變大，則易於拒絕虛無假設。

Example: Is the die fair?

- 假設擲骰子150次，我們得到下列各個點數出現的次數分配：

face	1	2	3	4	5	6
Number of rolls	22	21	22	27	22	36

- 假設我們預期擲骰子的次數出現機率是公平的，也就是每面出現的機率都是1/6或是每面平均出現25次。但第6面出現了36次，這是巧合還是另有其他原因？

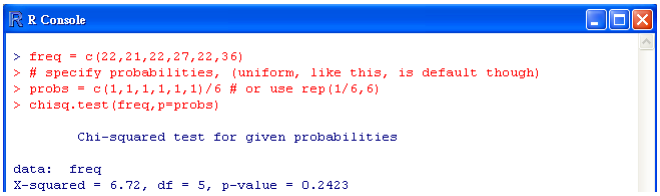
- 這個問題的關鍵在於實際出現次數與期望出現次數間的差距。假設我們將擲骰子實際出現的次數定義成f_i，期望出現的次數定義成e_i 這樣我們就可以將卡方分配統計量定義為：

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_i - e_i)^2}{e_i}$$

- 也就是計算(實際出現次數-期望出現次數)^2/期望出現次數的加總。如果算出來的值很大，那麼就表示實際出現次數與期望出現次數間有很大的差距。反之，則很小。

- 如果期望次數都不小於1，且大多數(80%以上)的期望次數都大於5的。那麼前述的卡方分配統計量就會服從自由度為n-1的卡方分配。
- 虛無假設是各點數出現機率相等，也就是1/6；而對立假設是指點數機率不全為1/6。此例中期望次數全部都符合其值必須大於5次的條件。

- H0：骰子每面出現的機率都是1/6
- H1：骰子每面出現的機率不全是1/6



- 假設顯著水準為90%，此處的p-value 0.2423明顯大於0.1，所以結論是無法拒絕虛無假設，也就是骰子各面出現的機率都是1/6。

Example: Letter distributions

- 英文中最常出現的5個字母為E、T、N、R、O，據統計這些字母的分佈如下表，此訊息可以幫助我們解決密文的相關問題。

letter	E	T	N	R	O
freq.	29	21	17	17	16

- 今有某篇文章其上述字母出現的次數如下表：

letter	E	T	N	R	O
freq.	100	110	80	55	14

- 請使用卡方分配來檢驗字母出現的比例是不是πE= .29, πT= .21, πN=.17, πR= .17, πO= .16 。

- H_0: 字母E、T、N、R、O出現的機率分別為29%、21%、17%、17%及16%，亦即可判定此篇文章是由英文所組成。
- H1: 字母E、T、N、R、O出現的機率不全為29%、21%、17%、17%及16%，亦即此篇文章不是由英文所組成的。

```
R Console

> x = c(100,110,80,55,14)
> probs = c(29, 21, 17, 17, 16)/100
> chisq.test(x,p=probs)

        Chi-squared test for given probabilities

data:  x
X-squared = 55.3955, df = 4, p-value = 2.685e-11
```


- 結論：

  假設顯著水準為90%，此處p-value值為2.685e-11，小於0.1。所以結論是拒絕虛無假設，也就是這篇文章中，字母E、T、N、R、O出現的機率不全是29%、21%、17%、17%、16%，因此這篇文章不是用英文寫的，可能是用其它的語言撰寫的。

獨立性檢定

Chi-Squared Tests of Independence

獨立性檢定

H_0(虛無假設)：戴安全帶&損傷程度為獨立(無關係)

H_1(對立假設)：戴安全帶&損傷程度為不獨立(有關係)

For example:

根據以下數據，探討乘客是否繫安全帶與受傷程度的關係。

損傷程度

有無繫上安全帶

		Injury Level			
		None	minimal	minor	major
Seat Belt	Yes	12,813	647	359	42
	No	65,963	4,000	2,642	303

請問兩者是獨立的嗎?
還是安全帶確實有發揮減低傷害的作用呢?
利用卡方檢定就可回答此問題。

		Injury Level			
		None	minimal	minor	major
Seat Belt	Yes	12,813	647	359	42
	No	65,963	4,000	2,642	303

計算此格的預期數

H_0下兩者為獨立的情況下，我們可以利用邊際機率計算期望次數。

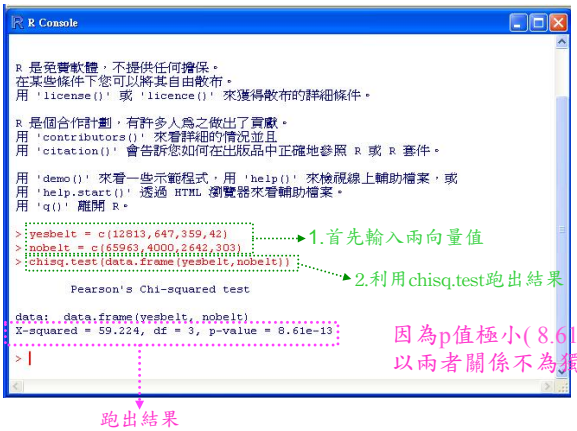
P(none and yes) = P(none)P(yes)

先估計(under H_0)P(none & yes)=P(none)P(yes)，再計算期望次數

>> ◯ = $\frac{12813+65963}{12813+647+359+42+65963+4000+2642+303}$ * $\frac{12813+647+359+42}{12813+647+359+42+65963+4000+2642+303}$

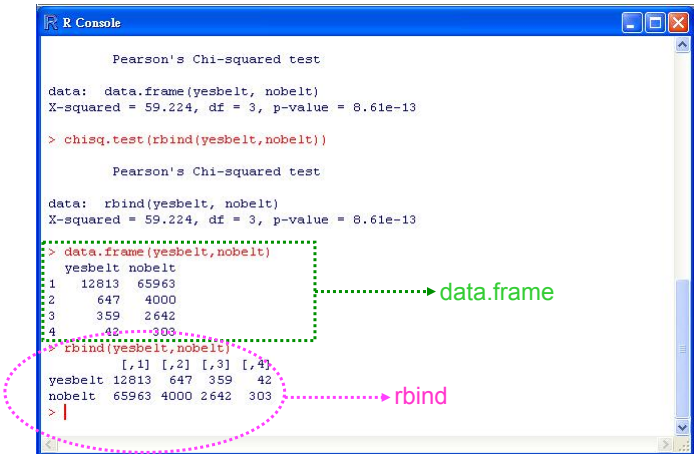
=0.91 x 0.16 = 0.1456〔P(none & yes)的機率〕

>> none & yes的期望次數 = 0.1456(期望機率) * 86769(總數) = 12633.57#



結論：拒絕H_0，所以繫安全帶跟受傷程度是有關係的。

補充：



同質(或齊一)性檢定

Chi-Squared Tests of Homogeneity

齊一性檢定

- 獨立性檢定是用來確認行與列是否是獨立的，而同質性檢定則是用來檢視各列是否來自相同的母體分配。一般來說，若每一列的分配相同，則各行的比例應該相近，卡方統計量可以幫助我們回答這個問題。

Example：
A difference in distributions ?

- 同質性檢定檢驗各列資料是否來自相同的分配，我們先從不同的分配抽取隨機資料。投擲一個公平的骰子200次和一個不公平的骰子100次，看卡方檢定如何能檢定出兩組資料來自不同的分配。

Example：
A difference in distributions？

```
> die.fair = sample(1:6,200,p=c(1,1,1,1,1,1)/6,replace=T)
> die.bias = sample(1:6,100,p=c(.5,.5,1,1,1,2)/6,replace=T)
> res.fair = table(die.fair);res.bias = table(die.bias)
> rbind(res.fair,res.bias)
          1  2  3  4  5  6
res.fair 38 26 26 34 31 45
res.bias 12  4 17 17 18 32
```


Example：
A difference in distributions？

- 不公平的骰子六點出現的次數遠多於預期，兩點出現的次數則遠低於預期。以卡方檢定進行同質性檢定的程序與獨立性檢定的程序類似，仍然是根據邊際機率來計算各cell的期望次數。

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i}$$

Example：
A difference in distributions？

- 虛無假設為兩組資料的分配相同(同質)，檢定統計量服從5個自由度的卡方分配，也就是列數減1乘上行數減1。我們同樣以chisq.test函數來完成這項任務：

```
> chisq.test(rbind(res.fair,res.bias))
    Pearson's Chi-squared test
data:  rbind(res.fair, res.bias)
X-squared = 10.7034, df = 5, p-value = 0.05759
```


The End