巨量資料探勘與統計應用 W12

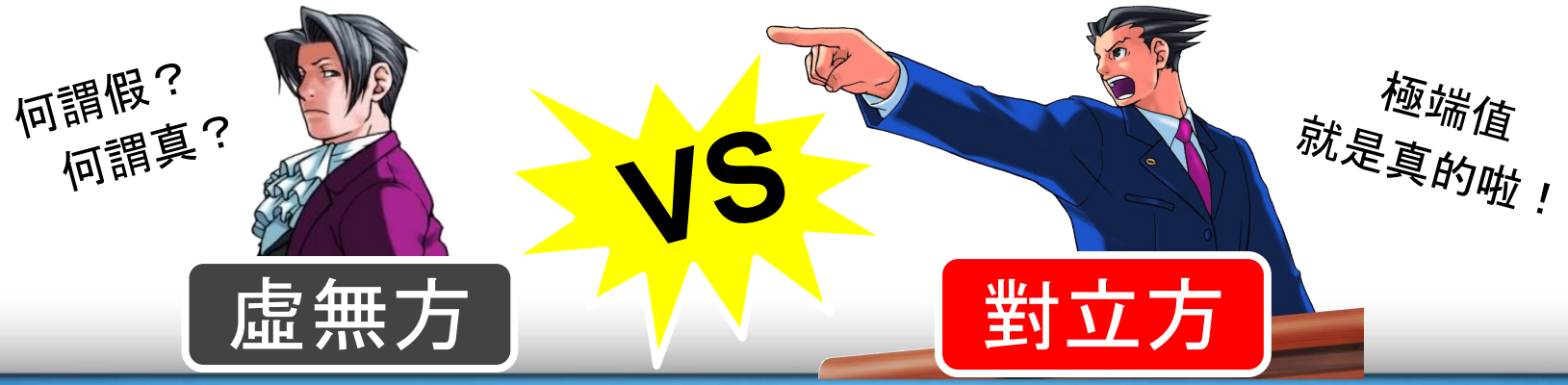
類別變項的相關檢定：卡方獨立性檢定

布丁布丁吃布丁

http://blog.pulipuli.info/

本週課程大綱

1. 類別資料的分析：列聯表與類別變項的相關
2. 類別變項的相關檢定：卡方獨立性檢定
3. 實作：卡方檢定──入學審核有性別歧視？
4. 課堂練習：辛普森詭論──入學審核沒有性別歧視！



Part 1.

類別資料的分析:
列聯表與類別變項的相關

資料蒐集

連續資料

年齡:21

身高:163

收入:32k

類別資料

性別:女性

學歷:大學

閱讀偏好:
專業書籍

問卷設計：類別⇨連續
獨立樣本t檢定

想要知道**居住地區**是否會造成**月收入**的差異？

問卷調查

- 請問您的居住區域：
  □北部 □南部
- 請問您的月收入：
  □ 1. 未滿2萬
  □ 2. 2萬以上，未滿4萬
  □ 3. 4萬以上，未滿6萬
  □ 4. 6萬以上

類別資料

連續資料

問卷設計:連續⇔連續
皮爾森積差相關分析

想要知道**證照數量**跟**月收入**是否有關係？

問卷調查

- 請問您的證照數量:_____
- 請問您的月收入:
  - ☐ 1. 未滿2萬
  - ☐ 2. 2萬以上，未滿4萬
  - ☐ 3. 4萬以上，未滿6萬
  - ☐ 4. 6萬以上

連續資料

連續資料

問卷設計：類別⇔類別
卡方獨立性檢定

想要知道**教育程度**跟**閱讀偏好**是否有關係？

問卷調查

- 請問您的教育程度：
  □ 高中以下 □ 大學以上

- 請問您的閱讀偏好：
  □ 娛樂類書籍
  □ 專業類書籍

類別資料

類別資料

問卷資料整理
製作次數分配表 =列聯表

問卷調查原始資料 → 次數分配表 =列聯表

教育水準	閱讀偏好
大學以上	娛樂類書籍
高中以下	專業類書籍
大學以上	專業類書籍
高中以下	娛樂類書籍
高中以下	專業類書籍

		教育程度	
		高中以下	大學以上
閱讀偏好	娛樂類	100	75
	專業類	75	108

W04 資訊視覺化:統計圖表

列聯表的構造 (1/3)

行變項 (2個)

列變項 (2個)

		教育程度		列總合
		高中以下	大學以上	
閱讀偏好	娛樂類	100	75	175
	專業類	75	108	183
行總合		175	183	358

列聯表的構造 (2/3)

		教育程度		列總合
		高中以下	大學以上	
閱讀偏好	娛樂類	100	75	175
	專業類	75	108	183
行總合		175	183	358

邊際總合

邊際總合

樣本總數

列聯表的構造 (3/3)

2x2列聯表

- 行變項:2個
- 列變項:2個

		教育程度		列總合
		高中以下	大學以上	
閱讀偏好	娛樂類	100	75	175
	專業類	75	108	183
行總合		175		

細格
(2x2=4個)

列聯表的尺寸

問卷調查

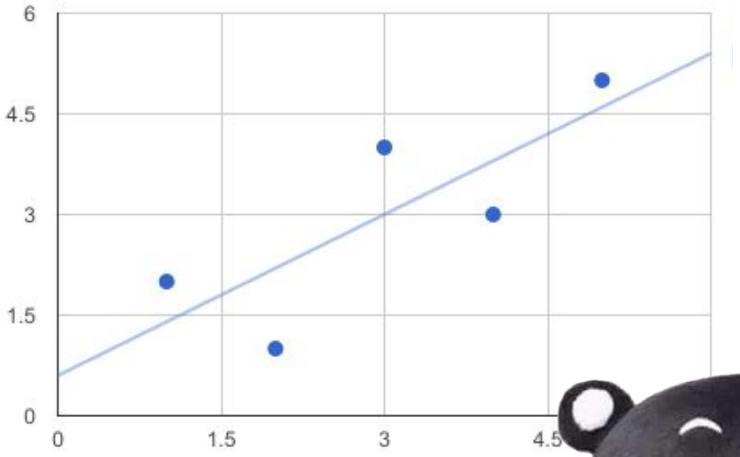
- 請問您的教育程度:
  □ 高中以下 □ 大學以上
- 請問您的閱讀偏好:
  □ 文藝類書籍
  □ 商業類書籍
  □ 科技類書籍
  □ 休閒類書籍

2x4列聯表

		教育程度	
		高中以下	大學以上
閱讀偏好	文藝類	100	75
	商業類	75	108
	科技類	30	57

相關？

連續變項的相關

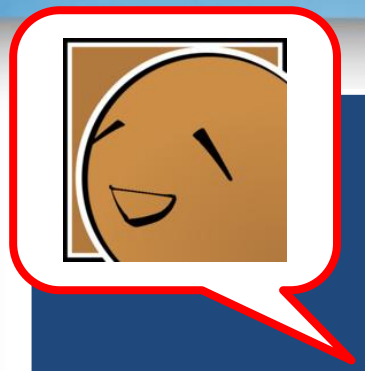


類別變項的相關？

		教育程度	
		高中以下	大學以上
閱讀偏好	娛樂類	100	75
	專業類	75	108

相關係數：0

無相關



		教育程度		列總合
		高中以下	大學以上	
閱讀偏好	娛樂類	25	25	50
	專業類	25	25	50
行總合		50	50	100

教育程度

跟

閱讀偏好

無相關

相關係數: 0.2

低度相關

A縣市

		教育程度		列總合
		高中以下	大學以上	
閱讀偏好	娛樂類	30	20	50
	專業類	20	30	50
行總合		50	50	100

高中以下看娛樂類
大學以上看專業類

低度相關

相關係數: 0.6

中度相關

B縣市

		教育程度		列總合
		高中以下	大學以上	
閱讀偏好	娛樂類	40	10	50
	專業類	10	40	50
行總合		50	50	100

高中以下多看娛樂類

大學以上多看專業類

中度相關

相關係數: 0.8

高度相關

C縣市

		教育程度		列總合
		高中以下	大學以上	
閱讀偏好	娛樂類	45	5	50
	專業類	5	45	50
行總合		50	50	100

高中以下
幾乎都看娛樂類

大學以上
幾乎都看專業類

高度相關

相關係數：1
完全相關

幻想縣市

		教育程度		列總合
		高中以下	大學以上	
閱讀偏好	娛樂類	50	0	50
	專業類	0	50	50
行總合		50	50	100

高中以下
全部都看娛樂類

大學以上
全部都看專業類

完全相關

類別變項的相關

無相關

不能用教育程度
來預測閱讀偏好

		教育程度	
		高中以下	大學以上
閱讀偏好	娛樂類	25	25
	專業類	25	25

完全相關

可以用教育程度
來預測閱讀偏好

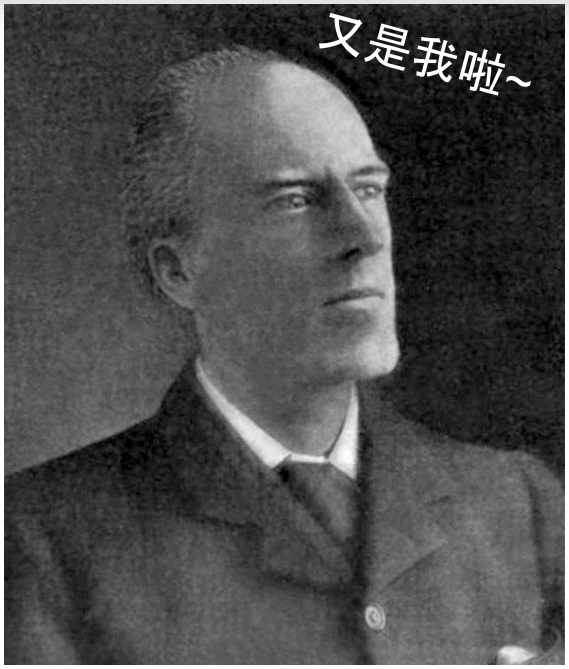
		教育程度	
		高中以下	大學以上
閱讀偏好	娛樂類	50	0
	專業類	0	50

Part 2.

類別變項的相關檢定:
卡方獨立性檢定

Pearson's chi-squared test
皮爾森卡方檢定 χ^2

(唸作chi-squared = 開·史克威爾)

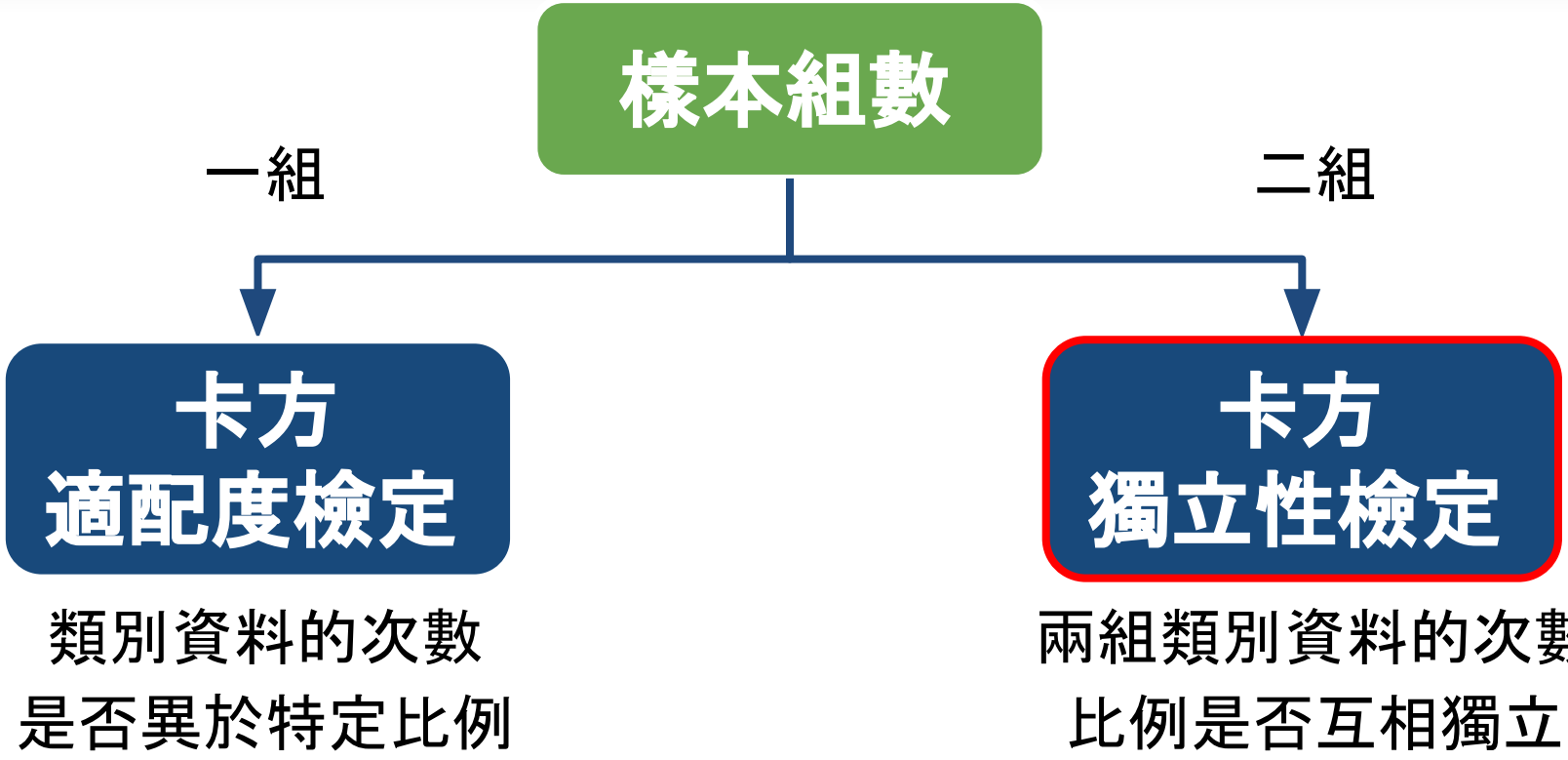


卡爾·皮爾森
(1857-1936)
英國數學家

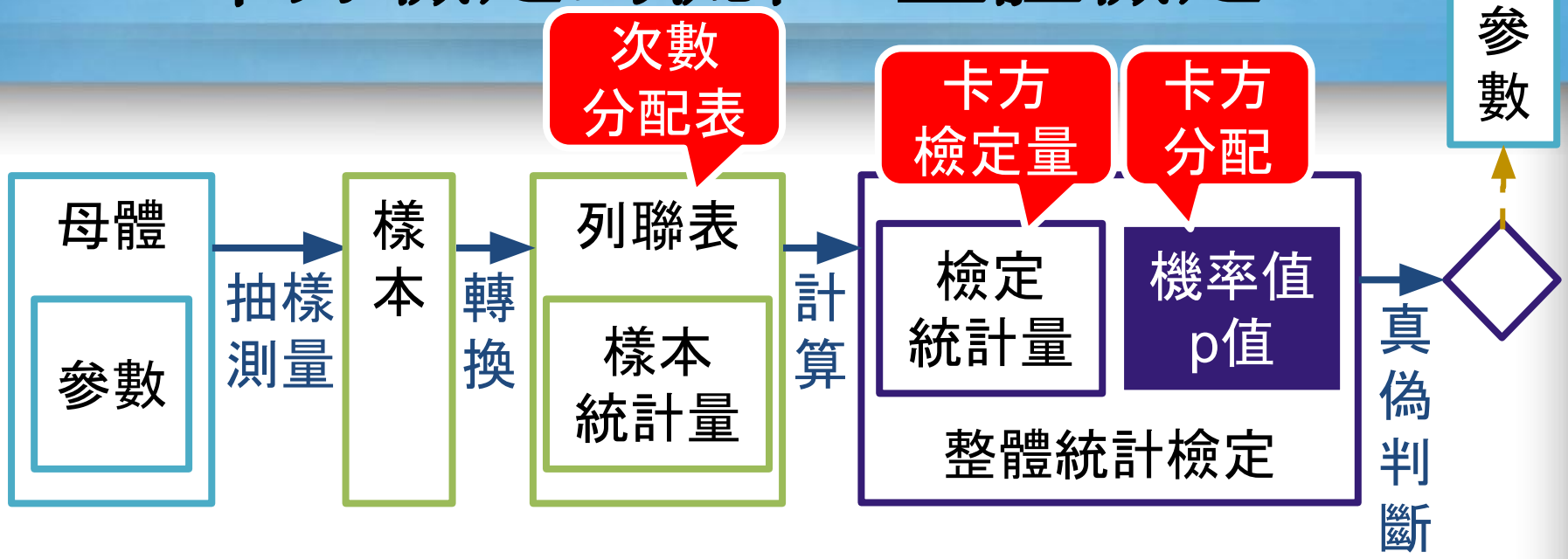
- 1900年發表
- 卡方分配:類別資料會遵循的機率分配
- 卡方檢定:(大樣本)
  - 皮爾森卡方檢定
- 卡方檢定:(小樣本)
  - 費雪爾正確概率檢定(1922)
  - 葉氏連續型校正(1934)

Wikipedia: https://goo.gl/q3BmoH

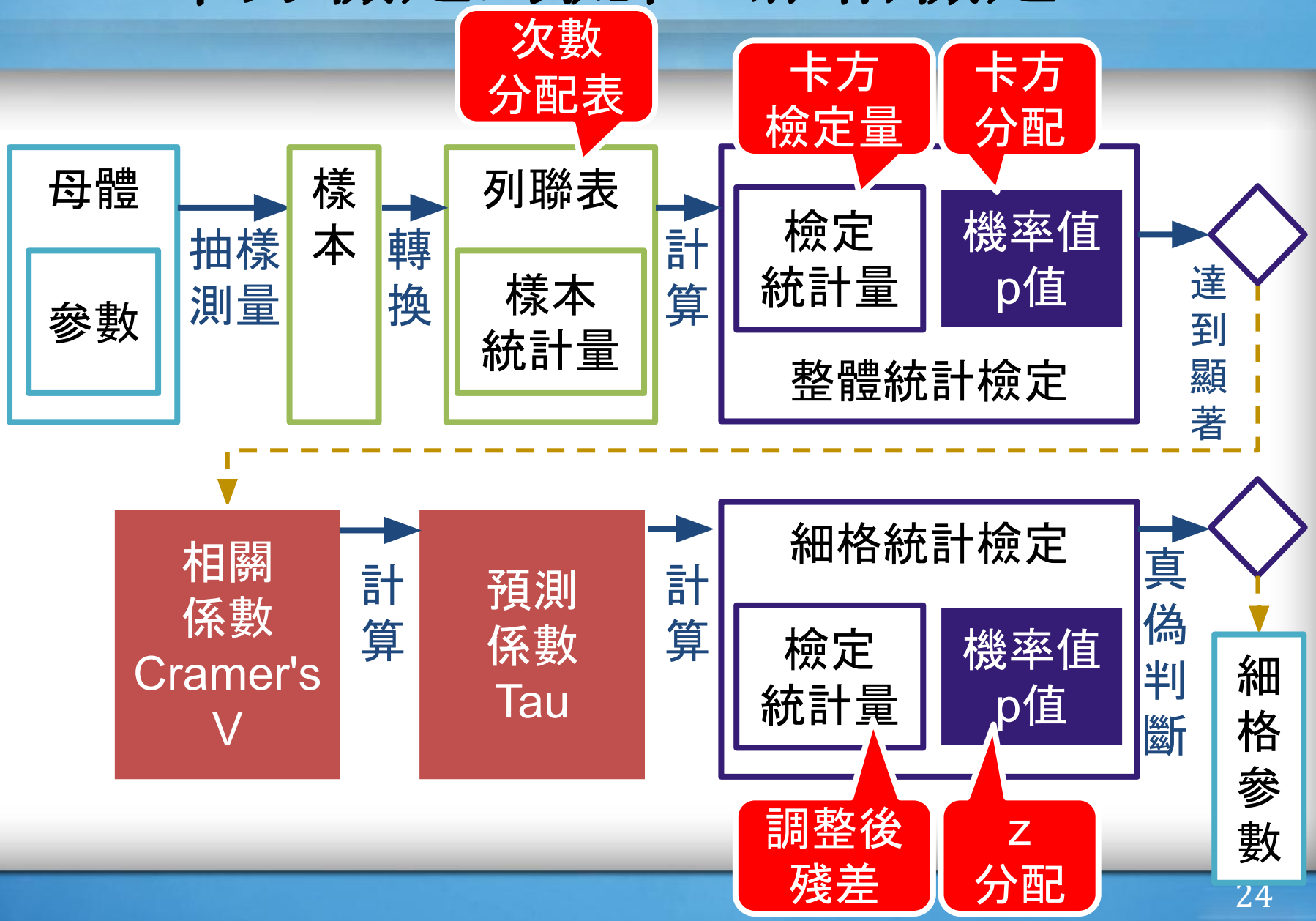
卡方檢定家族


樣本組數
一組
二組
卡方
適配度檢定
卡方
獨立性檢定
類別資料的次數
是否異於特定比例
兩組類別資料的次數
比例是否互相獨立

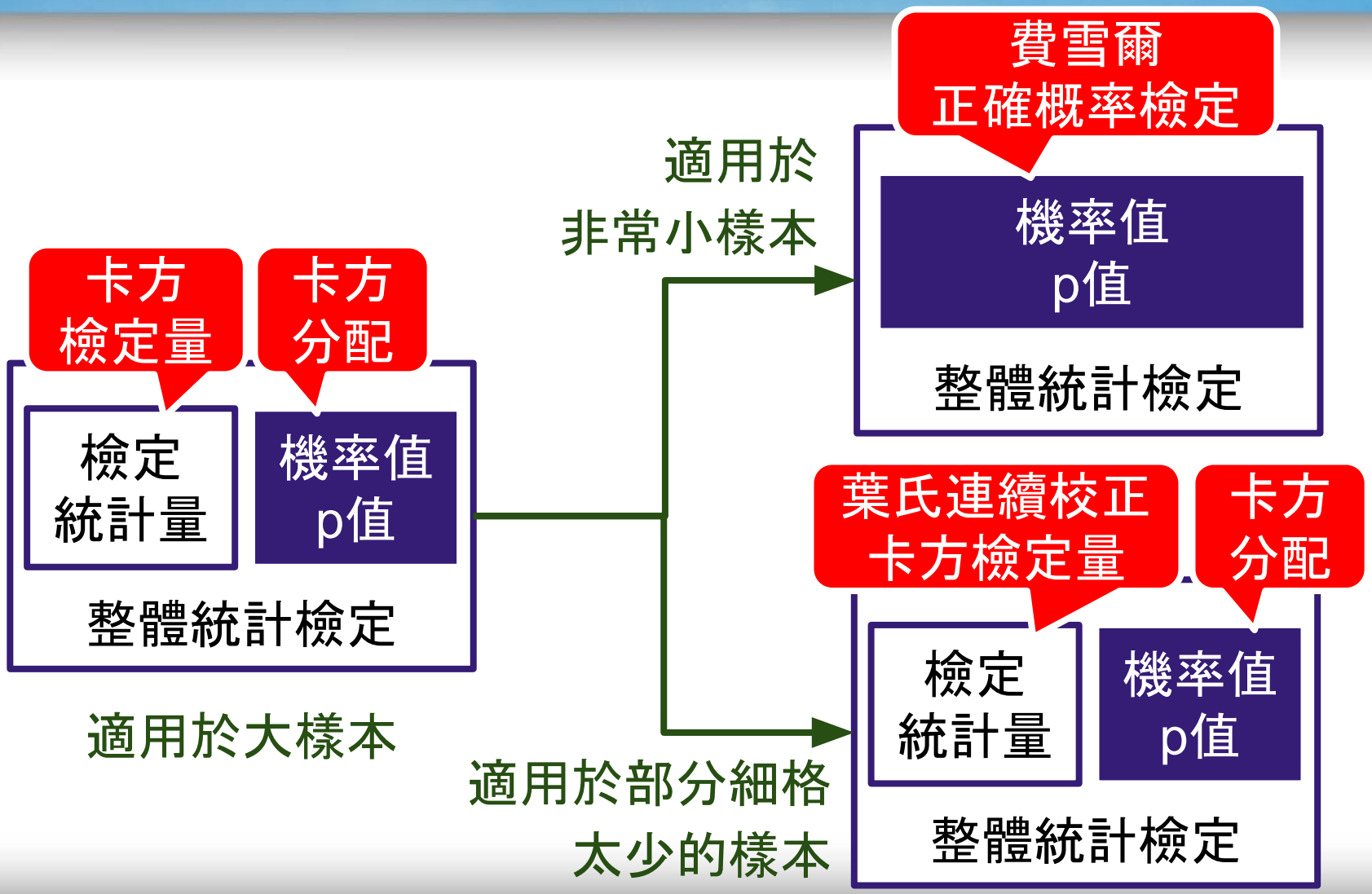
卡方檢定的流程:整體檢定



卡方檢定的流程:細格檢定



卡方檢定的小樣本校正


卡方
檢定量
卡方
分配
檢定
統計量
機率值
p值
整體統計檢定
適用於大樣本
適用於
非常小樣本
費雪爾
正確概率檢定
機率值
p值
整體統計檢定
適用於部分細格
太少的樣本
葉氏連續校正
卡方檢定量
卡方
分配
檢定
統計量
機率值
p值
整體統計檢定

卡方檢定詳細公式參考資料

- 卡方檢定統計量:
  - 邱皓政(2005)。統計原理與分析技術:SPSS中文視窗版操作實務詳析。
- 葉氏連續性校正:
  - Wikipedia: https://goo.gl/2cbGxv
- 費雪爾正確概率檢定:
  - Ina Parks S. Howell的Fisher's Exact Test An Example http://www2.fiu.edu/~howellip/Fisher.pdf
- 相關係數Cramer's V值跟預測係數Tau值:
  - 邱皓政(2005)。統計原理與分析技術:SPSS中文視窗版操作實務詳析。
- 細格統計檢定之調整後標準化殘差:
  - 邱皓政(2005)。統計原理與分析技術:SPSS中文視窗版操作實務詳析。

列聯表的期望個數計算 $\hat{\mu}_{ij}$ (1/2)

		教育程度(j)		
		高中以下	大學以上	列總合
閱讀偏好(i)	娛樂類	100	75	175
	專業類	75	108	183
		175	183	358

細格個數 n_{ij}

列總合人數 $n_{i.}$

樣本總數 N

行總合人數 $n_{.j}$

列聯表的期望個數計算 $\hat{\mu}_{ij}$ (2/2)

$$\hat{\mu}_{ij} = \frac{n_{i.} \times n_{.j}}{N}$$

$$= \frac{183 \times 175}{358} = 89.46$$

殘差與標準化殘差

殘差Δ (delta)

$$\Delta_{ij} = n_{ij} - \hat{\mu}_{ij}$$

標準化殘差 Δ'

$$\Delta'_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$$

殘差越大,
表示該細格個數
的關聯性越高

計算皮爾森卡方 值
Pearson χ^2 (1/2)

先計算每一個細格的標準化殘差Δ'

		教育程度	
		高中以下	大學以上
閱讀偏好	娛樂類	1.6	-1.5
	專業類	-1.5	1.5

將所有細格的標準化殘差平方後加總, 即得到Pearson $\chi 2$

$$\chi 2 = (1.6)^2+(-1.5)^2+(-1.5)^2+(1.5)^2=9.31$$

計算皮爾森卡方 值
Pearson χ^2 (2/2)

將所有細格的標準化殘差平方後加總, 即得到Pearson $\chi 2$

$$\chi^2 = \sum \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

$= (1.6)^2+(-1.5)^2+(-1.5)^2+(1.5)^2=9.31$

葉氏連續性校正

為了避免樣本個數過少的時候, 卡方值的誤差容易顯著的問題

$$\chi^2_{Yates} = \sum \frac{(|n_{ij} - \hat{\mu}_{ij}| - 0.5)^2}{\hat{\mu}_{ij}}$$

使用葉氏連續性校正的條件:

- 只有在2x2列聯表
- 有細格的期望個數小於5個
- 樣本總數超過20個

後面的算法跟卡方值一樣

Wikipedia: https://goo.gl/2cbGxv

自由度 df

		教育程度(j)	
		高中以下	大學以上
閱讀偏好(i)	娛樂類	1.6	-1.5
	專業類	-1.5	1.5

自由度df

= (列變項個數-1)

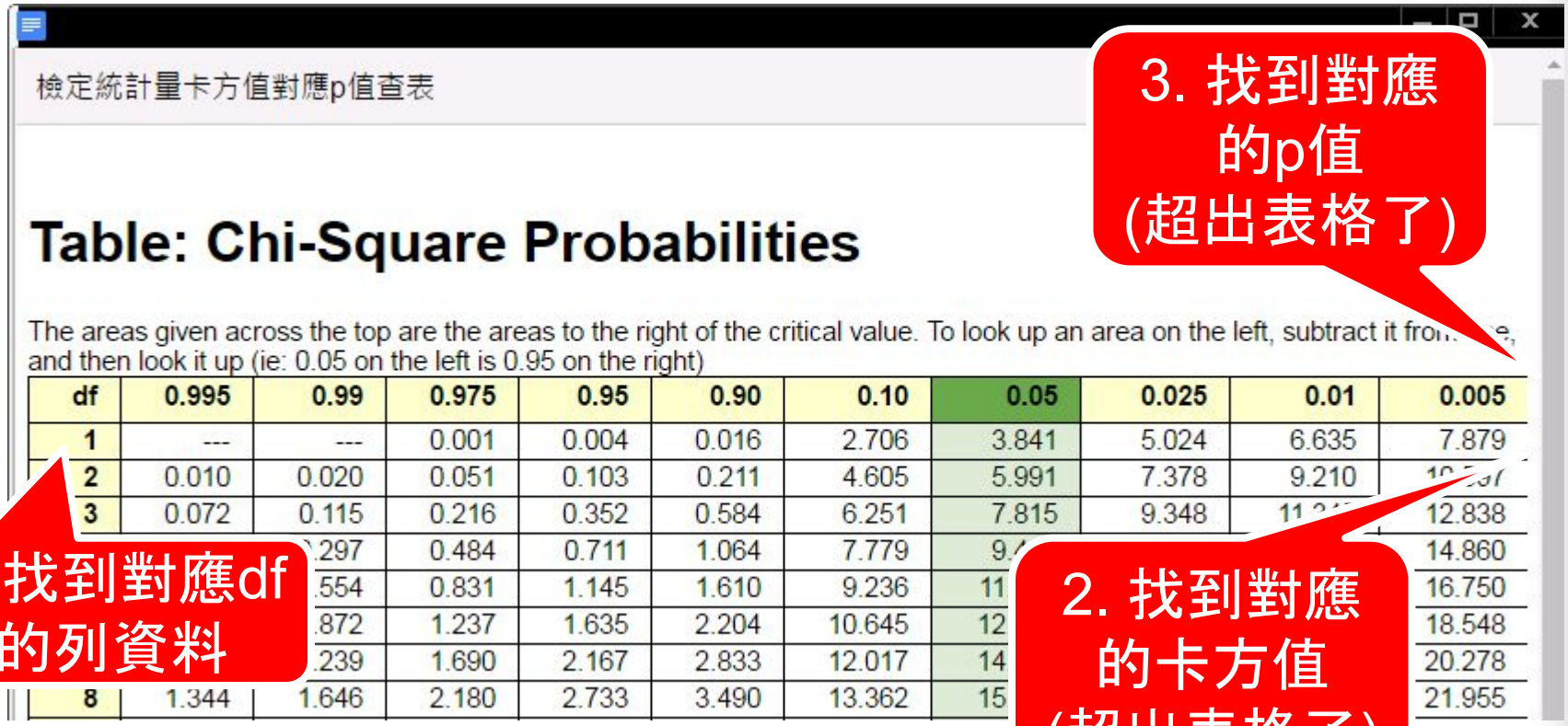
✖(欄變項個數-1)

=(2-1)✖(2-1)=1

查詢卡方值的機率
p值 (1/2)

檢定統計量
卡方值對應p
值查表

查詢卡方值的機率
p值 (2/2)



檢定統計量卡方值對應p值查表

Table: Chi-Square Probabilities

The areas given across the top are the areas to the right of the critical value. To look up an area on the left, subtract it fro… e, and then look it up (ie: 0.05 on the left is 0.95 on the right)

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	…
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	…	12.838
…	…	.297	0.484	0.711	1.064	7.779	9.…	…	…	14.860
…	…	.554	0.831	1.145	1.610	9.236	11	…	…	16.750
…	…	.872	1.237	1.635	2.204	10.645	12	…	…	18.548
…	…	.239	1.690	2.167	2.833	12.017	14	…	…	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15	…	…	21.955

1. 找到對應df的列資料

2. 找到對應的卡方值(超出表格了)

3. 找到對應的p值(超出表格了)

費雪爾正確概率檢定

	通過	未通過	列總合
甲班	8	14	22
乙班	1	3	4
行總合	9	17	26

以列變項為i，行變項為j，代號如下：

	j = 1	j = 2	
i = 1	n_{11}	n_{12}	$n_{1.}$
i = 2	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

調整後的標準化殘差

調整後殘差 adjΔ' (1/2)

		教育程度		列總合	列人數比例
		高中以下	大學以上		
閱讀偏好	娛樂類	100	75	175	0.49
	專業類	75	108	183	0.51
行總合		175	183	358	1
行人數百分比		0.49			1

列人數比例
$P_{i.}=n_{i.}/N$

列總合人數
$P_{.j}=n_{.j}/N$

調整後的標準化殘差

調整後殘差 adjΔ' (2/2)

$$adj\Delta'_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - P_{i.})(1 - P_{.j})}}$$

- 調整後殘差遵從標準化常態分佈
- 當調整後殘差的絕對值大於1.96時
  - 即表示p值小於0.05，達到α=0.05的顯著水準
  - 意思是該細格個顯著大於/小於期望個數

相關係數
Cramer'V係數

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

k = min(行變項數量,列變項數量)

- 相關係數Cramer'V係數介於0~1之間
  - 越接近1，表示行變項與列變項越相關
    (請參考先前投影片對列聯表相關的說明)
- 可以用來標準化行變項與列變項之間的相關程度

預測係數
Tau(τ_y)係數

$$\tau_y = 1 - \frac{\sum\left(\frac{n_{ij}(n_{i.} - n_{ij})}{n_{i.}}\right)}{\sum\left(\frac{n_{.j}(N - n_{.j})}{N}\right)}$$

- 預測係數Tau值可以計算以i變項預測j變項的正確比例
- 預測係數Tau值有方向性:
  - 以上公式是以i變項預測j變項的正確比例
  - 將i與j對換, 即可改成以j變項預測i變項
- 在2x2列聯表中, 兩種方向得到的數值都一樣

卡方檢定的限制

		Lag 1 (t)			Lag 0 總數
		A	B	C	
Lag 0 (g)	A	1	2	1	4
	B	0	1	2	3
	C	2	0	0	2
Lag 1 總數		3	3	3	9

卡方檢定適用場合

- 樣本數量最好在30以上
- 0細格數量不可超過細格總數的1/4

不一定所有的列聯表都適用卡方檢定

卡方檢定計算器


Pearson Correlation
Input
文字框輸入
選擇檔案輸入
Google試算表共用網址
請上傳CSV檔案：(範例檔案下載)
選擇檔案
未選擇任何檔案
小數點位數
3
顯示顯著性跟個數
分析結果:
petallength
sepalwidth
卡方檢定
計算器

檢定主修科系與父母的社經地位

統計數據證實：沒有父母當靠山，你會選那
死的主修

2015-08-03 薇琪

1780 年，John Adams 寫了一封信給他太太 Abigail。在這封信中，Adam
子及孫子未來將投入的工作。Adams 本身花了很多的時間在精通政治學及
個改革性的必需品），他期望他的孩子可以學習能夠促進國家建設的學科，
運、商業。Adams 的盤算是這些實用的科目往後才能讓他孩子或後代子孫
詞、音樂、建築、雕塑、壁毯及瓷器的空間。

		父母地位	
		低收入	高收入
孩子主修	藝術家	30	65
	醫生	72	41
	圖書館員	35	32

※數字為虛構

https://buzzorange.com/vidaorange/2015/08/03/rich-family-affect-your-major/

檢定醫療方法的實驗結果

		醫療方法	
		玩俄羅斯方塊	沒有玩
傷患創傷記憶	順利康復	25	8
	沒有幫助	11	27

※數字為虛構

檢定告白傳說的真假

		告白場所	
		聖誕樹下	其他地方
告白結果	成功	40	31
	失敗	10	19

※數字為虛構

檢定告白傳說的真假

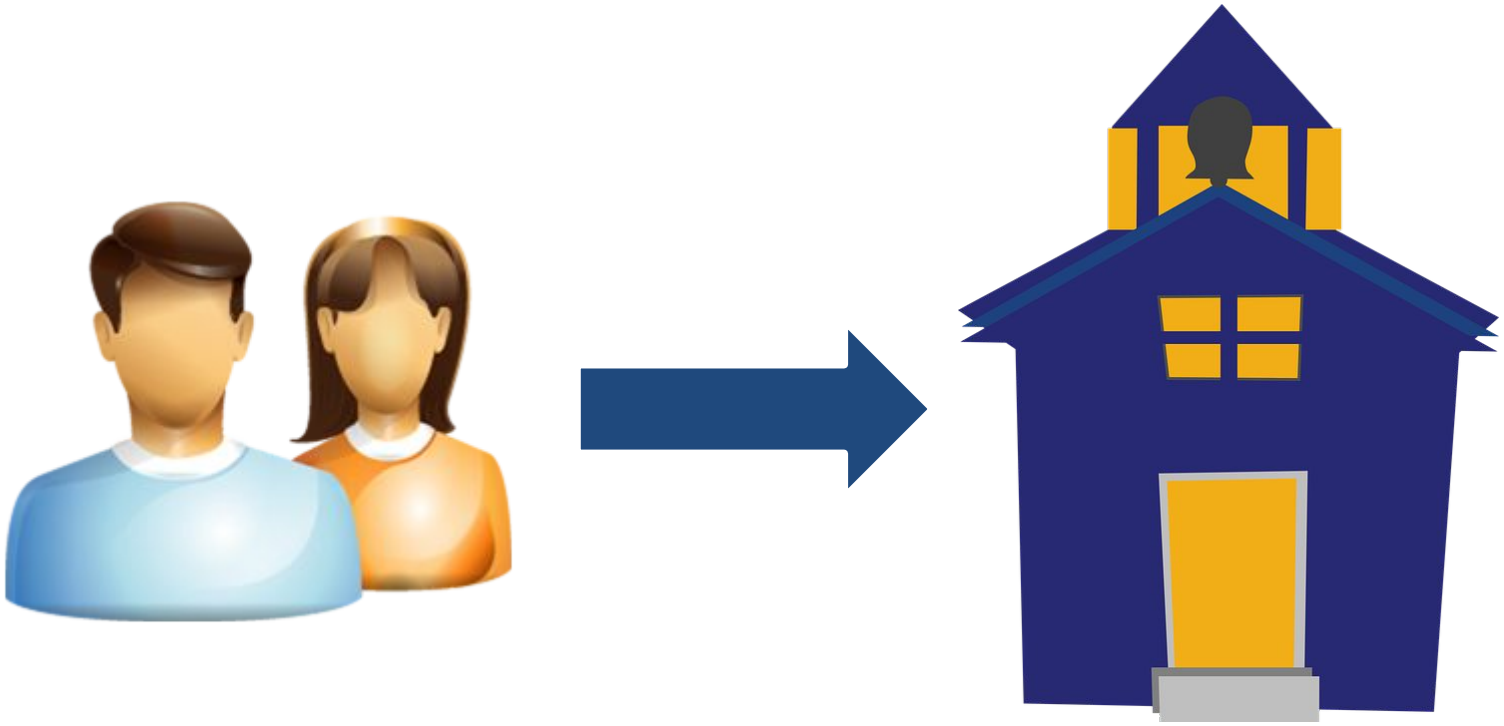
為虛構

Part 3.

實作:卡方檢定

入學審核有性別歧視？

某大學的入學招生



招生統計資料列聯表

		性別	
		男性	女性
報考結果	通過	100	75
	不通過	75	108
報考人數總合		175	183

男生都通過？女生都沒過？

我們強烈質疑

貴校入學審核有性別歧視！

圖片來源：https://kknews.cc/baby/rek2brr.html

實作學習單

實作:
卡方檢定
.docx

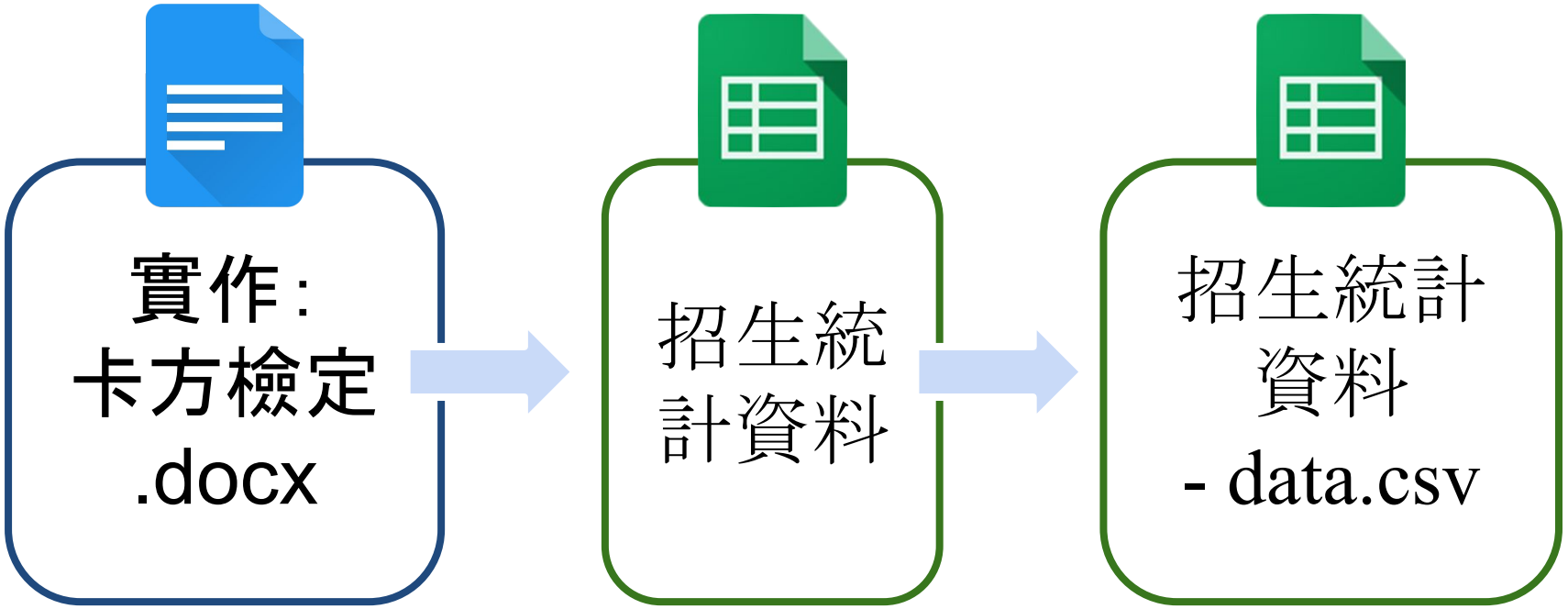
W12-a. 卡方檢定

1. 下載CSV檔案

2. 上傳CSV檔案到「卡方檢定計算器」
3. 解讀報表:卡方檢定結果

4. 撰寫結論

1. 下載CSV檔案



列聯表資料格式 說明

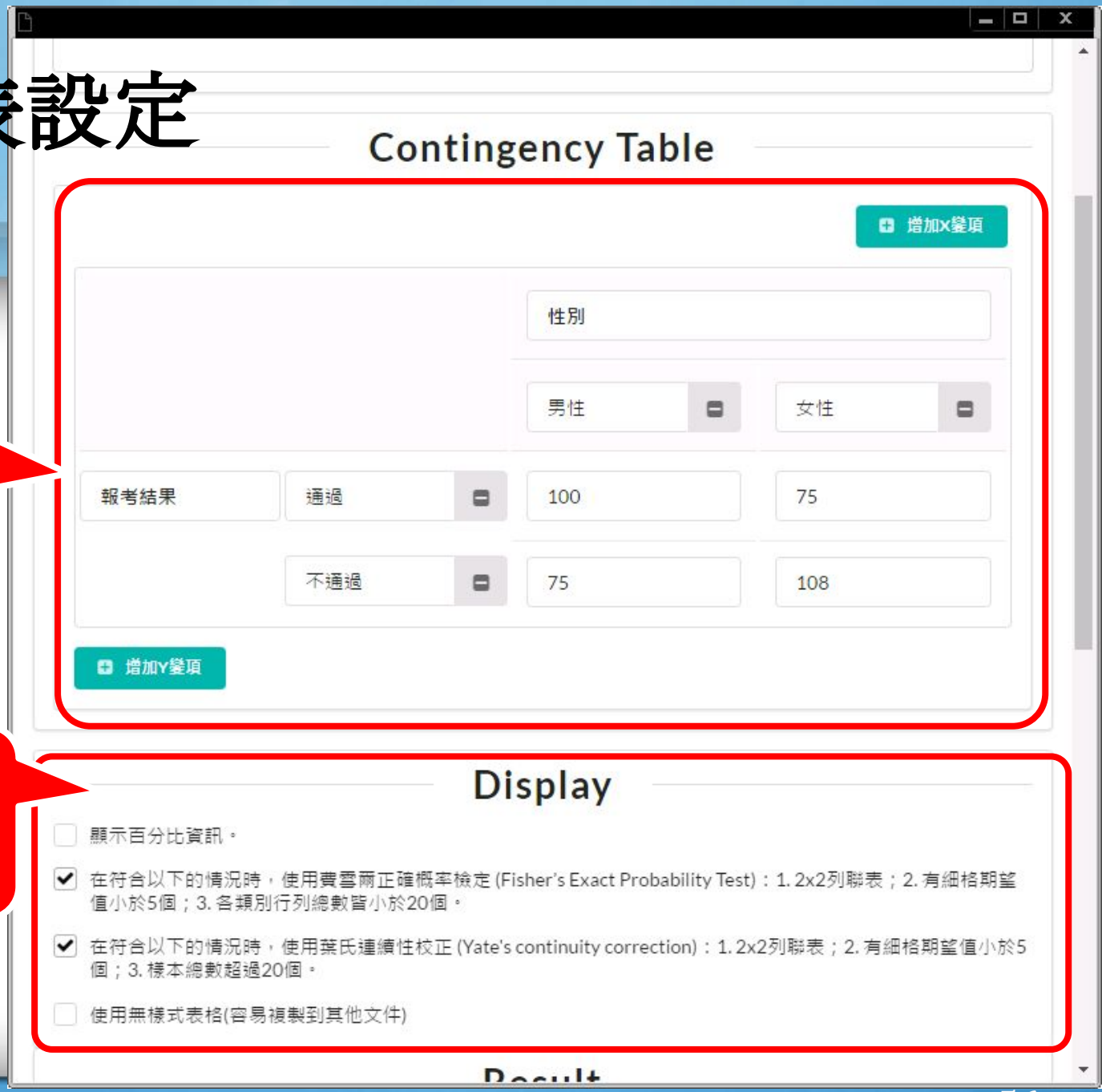
變項類別:變項名稱

- 行變項
- 列變項

2. 上傳CSV檔案

招生統計資料
- data.csv

列聯表設定



可手動
輸入數值

報表內容與
檢定設定

3. 解讀報表

列聯表
統計結果

卡方檢定
結果

使用無樣式表格(容易複製到其他文件)

Result

			性別		
			男性	女性	列總合
報考結果	通過	個數	100	75	175
		期望個數	85.6	89.5	175
		殘差	14.5	-14.4	
		標準化殘差	1.6	-1.5	
		調整後殘差	3.1	-3.0	
	不通過	個數	75	108	183
		期望個數	89.5	93.6	183
		殘差	-14.4	14.5	
		標準化殘差	-1.5	1.5	
		調整後殘差	-3.0	3.1	
行總合		個數	175	183	358
		期望個數	175	183	358

卡方檢定結果：

- 卡方檢定統計量χ^2 = 9.349，機率值p = 0.003，達到顯著水準α = 0.05，因此拒絕虛無假設，接受對立假設。表示「性別」的不同對「報考結果」有顯著的影響。
- 「性別」跟「報考結果」之關聯係數Cramer's V值(介於0~1之間)為 0.162，屬於低度相關。
- Goodman與Kruskal的預測係數Tau值的分析：
  - 以「性別」來預測「報考結果」的正確比例為2.612%。
  - 以「報考結果」來預測「性別」的正確比例為2.612%。
- 細格之調整後殘差分析：
  - 「男性」中「通過」之調整後殘差為3.1，表示觀察個數顯著高於期望個數。
  - 「女性」中「通過」之調整後殘差為-3，表示觀察個數顯著低於期望個數。
  - 「男性」中「不通過」之調整後殘差為-3，表示觀察個數顯著低於期望個數。
  - 「女性」中「不通過」之調整後殘差為3.1，表示觀察個數顯著高於期望個數。

列聯表統計結果

			性別		
			男性	女性	列總合
報考結果	通過	個數	100	75	175
		期望個數	85.6	89.5	175
		殘差	14.5	-14.4	
		標準化殘差	1.6	-1.5	
		調整後殘差	3.1	-3.0	
	不通過	個數	75	108	183
		期望個數	89.5	93.6	183
		殘差	-14.4	14.5	
		標準化殘差	-1.5	1.5	
		調整後殘差	-3.0	3.1	
行總合		個數	175	183	358
		期望個數	175	183	358

調整後殘差絕對值 > 1.96 表示該細格的個數顯著的多/少 (以黃底標示)

3-1. 整體統計檢定

整體統計檢定

有顯著！

- 卡方檢定統計量$\chi 2 = 9.349$，p值 = 0.003，達到$\alpha = 0.05$的顯著水準，因此拒絕虛無假設，接受對立假設。
- 表示「性別」的不同對「報考結果」有顯著的影響。

可以繼續往下分析

3-2. 相關係數與預測係數分析

相關係數

- 「性別」跟「報考結果」之相關係數Cramer's V值(介於0~1之間)為 0.162，屬於低度相關。

預測係數

- Goodman與Kruskal的預測係數Tau值的分析：
  - 以「性別」來預測「報考結果」的正確比例為2.612%。
  - 以「報考結果」來預測「性別」的正確比例為2.612%。

相關係數Cramer's V值的範圍
(介於0~1之間)

(跟積差相關係數r一樣)

相關係數範圍	變項之間關聯程度
1.00	完全相關
0.70 - 0.99	高度相關
0.40 - .069	中度相關
0.10 - 0.39	低度相關
0.10 以下	無相關

3-3. 細格統計檢定

細格統計檢定

- 細格統計檢定分析:
  - 「男性」中「通過」之調整後殘差為3.1,
    表示觀察個數顯著高於期望個數。
  - 「女性」中「通過」之調整後殘差為-3,
    表示觀察個數顯著低於期望個數。
  - 「男性」中「不通過」之調整後殘差為-3,
    表示觀察個數顯著低於期望個數。
  - 「女性」中「不通過」之調整後殘差為3.1,
    表示觀察個數顯著高於期望個數。

4. 撰寫結論:結論寫作框架

研究目的
XXXX(行變項)X(列變項)XXXXX
樣本敘述統計量
整體統計檢定
達到顯著水準　未達顯著水準
相關係數與預測係數分析
細格統計檢定

4-1. 研究目的 (框架)

研究目的

- 本研究使用卡方檢定分析(行變項)的差異對於(列變項)是否有所影響。

			性別		
			男性	女性	列總合
報考結果	通過	個數	100	75	175
		期望個數	85.6	89.5	175
		殘差	14.5	-14.4	
		標準化殘差	1.6	-1.5	
		調整後殘差	3.1	-3.0	
	不通過	個數	75	108	183

4-1. 研究目的 (填空)

研究目的

- 本研究使用卡方檢定分析性別的差異對於報考結果是否有所影響。

			性別		
			男性	女性	列總合
報考結果	通過	個數	100	75	175
		期望個數	85.6	89.5	175
		殘差	14.5	-14.4	
		標準化殘差	1.6	-1.5	
		調整後殘差	3.1	-3.0	
	不通過	個數	75	108	183

4-2. 樣本敘述統計量 (框架)

樣本敘述統計量

- 樣本數總共為(樣本總數)。

			性別		
			男性	女性	列總合
報考結果	通過	個數	100	75	175
		期望個數	85.6	89.5	175
		殘差	14.5	-14.4	
		標準化殘差	1.6	-1.5	
		調整後殘差	3.1	-3.0	
	不通過	個數	75	108	183
		期望個數	89.5	93.6	183
		殘差	-14.4	14.5	
		標準化殘差	-1.5	1.5	
		調整後殘差	-3.0	3.1	
行總合		個數	175	183	358
		期望個數	175	183	358

4-2. 樣本敘述統計量 (填空)

樣本敘述統計量

- 樣本數總共為358位。

			性別		
			男性	女性	列總合
報考結果	通過	個數	100	75	175
		期望個數	85.6	89.5	175
		殘差	14.5	-14.4	
		標準化殘差	1.6	-1.5	
		調整後殘差	3.1	-3.0	
	不通過	個數	75	108	183
		期望個數	89.5	93.6	183
		殘差	-14.4	14.5	
		標準化殘差	-1.5	1.5	
		調整後殘差	-3.0	3.1	
行總合		個數	175	183	358
		期望個數	175	183	358

4-3. 整體統計檢定

整體統計檢定

p < α, 達到顯著水準的寫法 (⇦本實作的情況)

- 卡方檢定統計量χ2 = 9.349，p值 = 0.003，達到α = 0.05 的顯著水準,
- 因此拒絕虛無假設，接受對立假設。表示「性別」的不同對「報考結果」有顯著的影響。

p >= α, 未達顯著水準的寫法

- 卡方檢定統計量χ2 = 0.008，p值 = 0.931，未達 α = 0.05 的顯著水準,
- 因此無法拒絕虛無假設。表示「性別」的不同對「報考結果」並沒有顯著的影響。

4-4. 相關係數與預測係數分析

(如果整體統計檢定達到顯著才寫)

相關係數

- 「性別」跟「報考結果」之相關係數Cramer's V值(介於0~1之間)為 0.162，屬於低度相關。

預測係數

- Goodman與Kruskal的預測係數Tau值的分析結果顯示:
  - 以「性別」來預測「報考結果」的正確比例為2.612%。
  - 以「報考結果」來預測「性別」的正確比例為2.612%。

4-5. 細格統計檢定 (冗長)

細格統計檢定 (如果整體統計檢定達到顯著才寫)

- 細格統計檢定分析結果顯示:
  - 「男性」中「通過」之調整後殘差為3.1,
    表示觀察個數顯著高於期望個數。
  - 「女性」中「通過」之調整後殘差為-3,
    表示觀察個數顯著低於期望個數。
  - 「男性」中「不通過」之調整後殘差為-3,
    表示觀察個數顯著低於期望個數。
  - 「女性」中「不通過」之調整後殘差為3.1,
    表示觀察個數顯著高於期望個數。

過於冗長

4-5. 細格統計檢定 (摘要)

細格統計檢定 (若卡方檢定有顯著，才進行以下分析)

- 細格統計檢定分析結果顯示:
  - 「男性」中「通過」之調整後殘差為3.1,
    表示觀察個數顯著高於期望個數。
  - 反之,「女性」中「通過」之調整後殘差為-3,
    表示觀察個數顯著低於期望個數。
  - ~~「男性」中「不通過」之調整後殘差為-3,~~
    ~~表示觀察個數顯著低於期望個數。~~
  - ~~「女性」中「不通過」之調整後殘差為3.1,~~
    ~~表示觀察個數顯著高於期望個數。~~

W12-a. 卡方檢定

實作啦！

			性別		
			男性	女性	列總合
報考結果	通過	個數	100	75	175
		期望個數	85.6	89.5	175
		殘差	14.5	-14.4	
		標準化殘差	1.6	-1.5	
		調整後殘差	3.1	-3.0	
	不通過	個數	75	108	183
		期望個數	89.5	93.6	183
		殘差	-14.4	14.5	
		標準化殘差	-1.5	1.5	
		調整後殘差	-3.0	3.1	
行總合		個數	175	183	358
		期望個數	175	183	358

Part 4.

課堂練習:辛普森詭論

入學審核沒有性別歧視！

不是這個辛普森

校長辛普森出來講話了

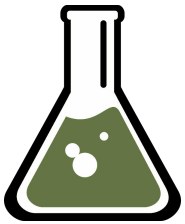
		性別	
		男性	女性
報考結果	通過	100	75
	不通過	75	108
報考人數總合		175	183

「這個是假象！」

列聯表雙向表⇨三向表

全校

		性別	
		男性	女性
報考結果	通過	100	75
	不通過	75	108
報考人數總合		175	183



理學院

		性別	
		男性	女性
報考結果	通過	75	25
	不通過	25	8
報考人數總合		100	33

文學院

		性別	
		男性	女性
報考結果	通過	25	50
	不通過	50	100
報考人數總合		75	150

招生統計資料
依學院分

理學院的入學列聯表

		性別	
		男性	女性
報考結果	通過	75	25
	不通過	25	8
報考人數總合		100	33

文學院的入學列聯表

		性別	
		男性	女性
報考結果	通過	25	50
	不通過	50	100
報考人數總合		75	150

W12-b. 辛普森詭論

1. 下載CSV檔案

2. 上傳CSV檔案到「卡方檢定計算器」
3. 解讀報表:卡方檢定結果

4. 撰寫結論

辛普森詭論
合併分組之後的隱藏變數

- 1951年英國統計學家辛普森發現一種現象:
  「當兩組資料合併成一組時, 相關的本質可能會改變, 甚至轉換方向」
- 發生原因:
  - 兩組資料量差異很大
  - 資料比例分配相反:不同學院通過與否的比例差別很大
- 警示:
  - 若貿然將資料加總, 可能無法反應真實情況
  - 可能另有造成主要影響的**隱藏變項:學院別**

Thank you for your attention