數據會說話？

黃文璋

國立高雄大學應用數學系

彰化師範大學

2011年5月13日

1. 前言

- 福爾摩斯：

  “Data! Data! Data!” he cried impatiently.
  “I can’ t make bricks without clay.”

- 福爾摩斯何以能料事如神？
- 得先有數據，就像製磚要有黏土。
- 科學時代，人們依數據做決策，經常有各種統計數據公佈:

  數據會說話。

- 但數據說的話，是否皆有價值？

◆97年1月有一則報導，標題:

O型、射手座、已婚男最易中彩券。

台灣彩券公司，針對前一年422位頭獎，及獎金超過500萬元的高額中獎人分析，發現其中

血型以O型最多，佔四成四；

射手座最多；

男性佔七成；

已婚者佔八成。

◆那些O型、射手座的已婚男看到此報導後，該趕緊去買彩券嗎？

- 台彩在當年1月底，也發佈一則新聞，標題:

  誰最容易中獎？首次一般彩券中獎人調查公開
  獅子座 常常買 一次買500元 最容易中獎。

  指出中獎者中，

  血型以O型最多，佔三成七；

  男性佔五成五；

  已婚者佔五成一。

- 兩份數據有些差異，不過較易中獎的族群大致相同，僅星座在後者變成以獅子座居冠，射手座反而敬陪末座。

◆台彩建議:

消費者可以參考本次調查的結果，做為在過年期間買彩券的小偏方。

- 第二則新聞，是針對四種彩券，96年10月22日至11月30日間，到全台16家中國信託分行兌領，彩券獎金100元至500萬元的中獎人，所進行之問卷調查，總共回收944份。
- 一個多月內，各種彩券中獎人很多，回收問卷數與全部中獎人數相比很少。而去彩券行兌領的人，也沒問到。中獎者也不見得樂意透露個人資料。
- 這類調查，獲得的數據，可能是偏差的。

- 一項調查，想要數據講的話值得參考，必須設計較好的程序，以獲取較客觀的樣本。這點常被忽略。

中獎人以O型最多，很稀奇嗎？

- 台灣O型最多，約佔全體人口的四成。
- 兩項調查，O型中獎分別佔四成四及三成七，並無特殊，根本不值得一提。

中獎人裡以O型最多，與O型最易中獎，一樣嗎？

- 若新北市開出最多頭獎，並不表想中大獎者，就該到新北市購買彩券。因新北市人口在全台各縣市中最多，若賣出最多彩券，導致中獎數最多，只是合理而已。
- 男性佔七成，且已婚者佔八成，與已婚男手氣較佳也是兩回事。

- 但對那一星座較易中獎，可能要更謹慎地分析。
- 如果確有某星座較愛買彩券，這時若此星座中獎人較多，不必大驚小怪。
- 另外，即使各星座的人買彩券之愛好都一樣，中獎比例也很難完全相同，總有高有低。

這是隨機現象的特性。

隨機現象

- 投擲一銅板1萬次，正反面恰好各出現5千次，你相信銅板公正，還是覺得其中必有詐？
- 樂透彩就算開獎機器沒問題，長期下來，便是很難各號碼出現一樣多次。
- 除非做個統計檢定，否則不能輕率認為那一號碼較易出現。

2. 統計與因果關係

- 美國百貨連鎖店統計顧客購物清單，發現尿布與啤酒同時出現的比例很高。原因何在？
- 例1. 100年4月5日有則新聞：

訃聞會說話 韓媒體人壽命短。

南韓圓光大學保健福祉學系教授金鐘仁，從1963年到2011年媒體發佈的3215名人士的訃聞，及統計廳提供的數據，進行分析，在“保健和福祉”上公佈結果。

◆平均壽命：

宗教人士80歲；政治家75歲；教授74歲；企業家73歲；法律界人72歲；高級公務員71歲；演藝人和藝術人70歲；媒體人、體育界人士及作家67歲。

◆例2. 100年4月9日有則新聞：

天天逛街 延年益壽

逛街有益身心健康，是愛血拼的女士編出的嗎？

英國廣播公司(BBC)報導，台灣國家衛生研究院張毓宏博士分析台灣1999至2008年,1856位65歲以上獨居老人後發現，每天逛街與不常逛街者相較，前者存活率高27%(男性高28%,女性高23%)。顯然這種購物療法(retail therapy)對男性健康更有助益。此結果發表於英國Journal of Epidemiology and Community Health。

- 統計結果，通常無法判定因果關係。
- 在例1裡,研究小組分析，生活規律、不斷修身養性、精神壓力較小，及禁煙禁酒等因素，可能是宗教人平均壽命較長的主因。
- 在例2裡，研究強調，逛街不一定要花大錢，只要到街上和別人打打交道，看看人群，減少孤寂感，就有助於身心健康。
- 研究還指出，逛街比上健身房更能維持健康，因為和正規運動相比，逛街通常不需強烈激勵，或專業人士指導，因此更容易養成習慣。

◆例3. 2004年5月，在美國New Orleans召開的一研討會上，印度的一醫療小組提出報告，從美國人過去半個世紀氣泡飲料的平均消耗，找到和食道癌罹患率提高的關連。

美國人平均一年飲用氣泡飲料的量，過去50年增加了450%，而過去25年來，美國白人男性食道癌罹患率，也呈現明顯增加的趨勢。

氣泡飲料易致癌？

過去50年不只氣泡飲料的飲用量增加，汽車購買，旅遊次數，以及很多其他的消費，都大幅增加。

難道都與食道癌有關？

該小組提出理論基礎，即氣泡飲料會讓胃部膨脹，導致消化液逆流，而這是食道癌產生的原因之一。

- 兩變數看起來有關連，有時可能是其中某些潛在變數所造成。
- 著名的**辛普生詭論**(Simpson's paradox)，也是指因另一變數的介入，而使兩變數的關係反過來。

◆例4. 某大學女學生抱怨，該校研究所對女生較不公平，因全校研究所女生的總錄取率35%，明顯較男生的44%低。該校研究所很多，自其中挑出6個較大的研究所來比較，可得下表(見Freedman et al.(1991))。

6個研究所中，有A、B、D、F等4個研究所，女生錄取率高於男生；男生錄取率高於女生的，僅有C、E等2研究所。但若6個研究所一起看，約有44%的男生被錄取，卻只有30%的女生被錄取，相差高達14%。此矛盾是如何產生的？

所別		申請	錄取	錄取率
A	女	108	89	82%
	男	825	511	65%
B	女	25	17	68%
	男	560	353	63%
C	女	593	201	34%
	男	325	120	37%
D	女	375	131	35%
	男	417	138	33%
E	女	393	94	24%
	男	191	53	28%
F	女	341	24	7%
	男	373	22	6%
合計	女	1835	556	30%
	男	2691	1197	44%

- A、B二研究所較易申請，且有一半以上的男生(共1,385人)申請此二研究所，而C、D、E、F四個研究所較難申請，卻有超過90%的女生(共1,702人)申請。
- 男生較多申請較容易進的研究所，女生較多申請較難進的研究所。
- 因此光由表面上全校研究所總錄取率之高低，來推測其錄取對女生不公，並不恰當。
- 事實上，有可能全校每一研究所都是女生錄取率較高，但全校一起看，卻是男生錄取率較高。

3. 致癌死亡人數多150%？

- 最近有一份報告，題目為

  國光石化營運比六輕石化營運致癌死亡人數多150%。

  由郭珮萱，莊秉潔，薛銘童，林佑勳，胡素婉，陳建仁，江濬如等合著。標題雖很聳動，但由於國光石化實際上尚未開始營運，所以這應只是一預測的結果，而非現況報告。

- 對一項研究，有幾分證據說幾分話。

◆摘要：

目前研究結果顯示，在95%信賴區間下，人民所得、六輕石化所造成之$PM_{2.5}$濃度與全台(不含花東)共322鄉鎮之惡性腫瘤死亡率變化量呈現顯著相關，每增加10μg/m3之$PM_{2.5}$，將使惡性腫瘤標準化死亡率男性增加129%，女性增加176%；另一方面每增加百萬元所得，將降低男性惡性腫瘤標準化死亡率58%、降低女性惡性腫瘤標準化死亡率24%。各關係式之相關係數(r^2)為0.30-0.43。根據這關係計算出全台已因六輕營運所排放之污染造成全癌症標準化死亡人數增加1686人/年，而因六輕所增加之收益(包括石化上中及下游)所造成減少癌症死亡人數每年為35人，總計因六輕計畫每年淨增加癌症死亡人數為1651人/年。

依六輕石化之劑量反應函數推估國光石化，則若國光石化營運後，其污染會造成全癌症死亡人數達每年4295人，而因國光所增加之收益所造成減少癌症死亡人數每年亦為35人，國光石化營運每年淨增加癌症死亡人數為達4260人/年。國光石化造成癌症死亡人數多於六輕人數將近150%，…

- 須留意事項：

1. 任何迴歸分析(regression analysis)，均應對所推估之迴歸模型做模型診斷(model diagnosis)，以確認模型是否適用。
2. 預測能見度所根據之線性迴歸模型在此是否適用？所得之預測值為負，但能見度豈可為負。
3. 在利用所推估出之迴歸模型做預測時，均應提供預測誤差(prediction error)之大小，以瞭解其預測之可靠程度。

4. 當數據中有明顯離群值(outliers)時，應先檢視這些數據是否合理,是否正確？若不合理或不正確，須先刪除。即使數據無誤，也該考慮離群值對預測所造成之影響。同時並應提供將離群值刪除後之結果，以作比較。即提供在正常情況下(無離群值)的迴歸模型估計。

5. 統計功能有其侷限，二因素間之因果關係是否存在，並非統計分析可以證實。因此不可輕率地對統計上看起來似乎相關的兩個因素，驟下結論說其間有因果關係。一般須有對照組，以多方確認。
6. 應對推估之迴歸模型與預測結果，進行交叉驗證(cross validation)，以確認所提供之分析與預測方法，是可接受的。

民國96-98年台灣各縣市之平均壽命(即零歲之平均餘命)，資料來源內政部統計處。

縣市	平均餘命	縣市	平均餘命
台北縣	79.92	苗栗縣	77.89
新竹市	79.69	臺南縣	77.64
臺中市	79.51	嘉義縣	77.48
澎湖縣	79.30	南投縣	77.03
桃園縣	79.27	雲林縣	76.54
嘉義市	78.77	高雄縣	76.53
新竹縣	78.59	屏東縣	75.75
臺南市	78.54	花蓮縣	74.36
彰化縣	78.39	臺東縣	73.73
基隆市	78.11	臺北市	82.00(96), 81.87(97), 82.50(98)
臺中縣	78.02	高雄市	78.05(96), 78.21(97), 78.61(98)
宜蘭縣	77.94		

註.內政部除北高兩市外，其餘縣市資料提供每三年之統計。

民國98年台灣各縣市每10萬人總死亡人數，及因惡性腫瘤死亡人數，

資料來源內政部衛生署統計室。

縣市	惡性腫瘤死亡人數	縣市	惡性腫瘤死亡人數
雲林縣	265.6	嘉義市	181.5
嘉義縣	240.2	臺北市	174.5
臺東縣	236.5	高雄市	173.6
臺南縣	213.8	臺南市	173.2
澎湖縣	212.1	新竹市	162.8
宜蘭縣	206.8	金門縣	160.3
花蓮縣	202.2	臺中縣	155.6
屏東縣	200.1	臺中市	150.8
南投縣	195.8	新竹縣	148.9
苗栗縣	192.0	連江縣	142.3
高雄縣	191.6	台北縣	135.6
彰化縣	191.0	桃園縣	129.7
基隆市	182.7		

◆造成壽命長短及死亡原因的因素很多。由內政部的資料，台灣有些好山好水縣市的平均壽命(即0歲之平均餘命)，在全台各縣市居於末位。民國98年每10萬人因惡性腫瘤之死亡人數，最低為129.7人，最高為265.6人，高低比達兩倍以上。由其排名之高低，恐也看不出與各縣市石化業化興隆程度之關係。

想依一份不夠嚴謹、號稱用到統計分析的報告，就下一驚人的結論：

國光石化營運比六輕石化營運致癌死亡人數(預測)多150%。

其實是缺乏科學態度的。

- 100年4月19日報導：

  國光石化「健康風險」「經濟分析」爭議落幕

  國光石化爭議健康風險及社會經濟評估及成本效益分析方法論，因環保團體提出聳動數據，引起廣大社會驚嚇，為釐清真相，環保署分別邀請相關學者專家進行學術專業之公開討論會取得共識，環保團體意見正式被否決。…此次誤會發生環工學者在對流行病學等不同領域基本定義與規則的誤解：

1. 誤將流行病學中累積死亡率認定為死亡風險機率，高估有80倍。
2. 以數學模式模擬值權充實測值納入統計分析中，高估有40~50倍。
3. 未按環保署法規使用三維網格模式，這項高估有2~10倍。

這三項的差異累積造成有5千~4萬倍。

4. 看數據宜有隨機的概念

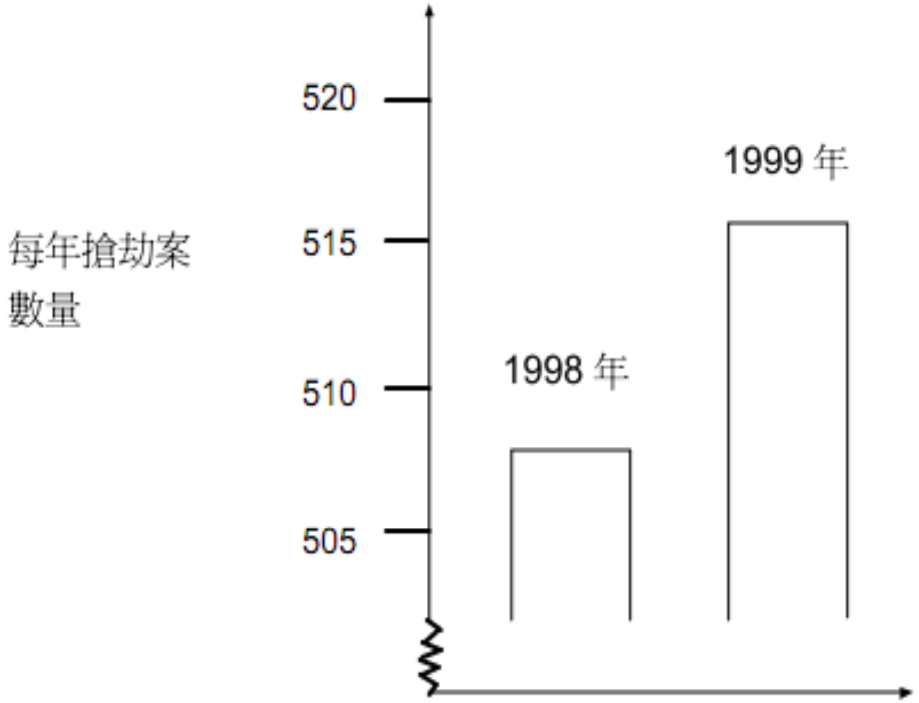
勿僅用數學的角度看數據！

M179 搶劫(1/4)

M179：搶劫

問題 1：搶劫 *M179Q01- 01 02 03 04 11 12 21 22 23 99*

電視主播呈現了下圖並報導：
「從圖表顯示，從 1998 年到 1999 年搶劫案數量有巨幅的上升」。



你認爲這位主播對於上圖的解釋是否合理？請寫出一個理由來支持你的答案。

M179 搶劫(2/4)

搶劫 問題 1 計分

注意：

以下代號中，答案「否」包括所有認爲「該詮釋是不合理的」的句子，而答案「是」則包括所有認爲「該詮釋是合理的」的句子。請不要單憑「是」或「否」來計分，而應看看答案解釋是否合理。

滿分

代號 21： 不，不合理。指出我們看到的只是整個圖表的其中一小部分。

- 不合理，須顯示整個圖表。
- 我不認爲那是合理的詮釋，因爲如果顯示全圖的話，便能看到搶劫案的數目只是輕微上升。
- 不合理，因爲他只用了圖表上方的小部分。如果看到全圖由0到520的情況，便知道上升的幅度不是那麼大。
- 不，那只是因爲該圖表讓人覺得數字巨幅上升。看數字增加並不多。

M179 搶劫(3/4)

代號 22： 不，不合理。用比率或百分比的數字作論據，論點正確。

- 不，不合理。與總數500比較，10不是一個巨幅的增加。
- 不，不合理。計算百分比，約只有2%的增加。
- 不，多了8宗搶劫案，即上升了1.5%。我認爲那不是很多！
- 不，今年只多了8或9宗，與507宗比較，那不是很大的數字。

代號 23： 要有趨勢的數據資料才可作出判斷。

- 我們不能說增加是否巨幅。若1997年的搶劫案數目與1998年的相同，那麼我們可以說1999年有巨幅增加。
- 有多「巨幅」，我們無從得知。因爲至少需要有兩個改變，才可判別哪個大，哪個小。

部份分數

代號 11： 不，不合理，但欠缺詳細解釋。

- 只有討論搶劫案的實際增加數字，但沒有將它與總數比較。
- 不合理。搶劫案數目大約增加了10宗。用「巨幅」一詞去形容搶劫案數目增加的真實情況不正確。搶劫案數目只大約增加了10宗，我不會稱之爲「巨幅」。
- 由508至515不是一個大增加。
- 不，因爲8或9不是一個大數目。
- 有點不合理。由508 至515 是有增加，但不是巨幅的增加。

M179 搶劫(4/4)

注意：

由於圖表的比例尺不是太清楚，因此如果搶劫案增加的數字在5至15之間，可以接受。

代號 12： 不，不合理。方法正確但有輕微計算錯誤。

- 方法和結論皆正確，但計算出來的百分比是0.03%。

零分

代號 01： 不。表示不合理，但沒有提供解釋、沒有充分解釋或解釋不正確。

- 不，我不同意。
- 主播不應用「巨幅」這個字眼。
- 不，這是不合理的。主播（記者）經常喜歡誇大。

代號 02： 是。基於圖表的形狀，因而指出搶劫案數字雙倍增加。

- 是，圖形的高度雙倍增加。
- 是，搶劫案數字差不多雙倍增加。

代號 03：是。沒有提供解釋，或提供代號02以外的解釋。

代號 04：其他答案

代號 99：沒有作答

劉禹錫 烏衣巷：

朱雀橋邊野草花，烏衣巷口夕陽斜；
舊時王謝堂前燕，飛入尋常百姓家。

- 住在昔日王導、謝安豪宅中的新主人，甚至同一巷子中的居民，想必都不會是太尋常的老百姓。
- 上例中解釋是否合理，並非看增加的數值之大小，而是看發生機率之大小。
- 一件事是否尋常，也依是機率發生之大小。
- 統計裡用顯著一詞。

- 假設 $X \sim B(520, 0.977)$ 。

  $np = 508.04$ ，

  $\sqrt{np(1-p)} \approx 3.418$ 。

  如此得

  $$P(X \geq 515) \approx P\left(\frac{X - np}{\sqrt{np(1-p)}} \geq \frac{515 - 508.04}{3.418}\right)$$

  $$\approx P(Z \geq 2.036)$$

  $$\approx 0.0209$$ 。

  此機率夠小，所以該主播之解釋並無不妥。

◆對隨機現象，人們平常理解誤差的存在。

99年4月19日，台大公布甄選榜單，報載醫學系20個名額中，有近半名額9位為女生，寫下歷史紀錄。台大教務長表示，女生錄取比例增加是剛好，台大招生的立場，向來就是找最適合、最有能力的學生，不會考慮學生性別。

◆只要近半，並不需女生錄取正好一半，才認定未考慮學生性別：

此講法正確否？

應由各性別的錄取率，而非由各性別的錄取人數，來判定兩性錄取是否有異。

某版高中選修數學(I)，交叉分析之例。

例題 2

某大學熱門科系的入學方式分成「學校推薦」和「個人申請」兩種，去年度經由此兩方式提出入學許可者的審核結果雙向表如下：

	推薦	申請
錄取	24	36
不錄取	36	54

(1) 求錄取的學生中，推薦和申請者所占的比例.

(2) 求推薦錄取率，申請錄取率和總錄取率.

解：(1) 先求雙向表中各列與各行總和：

	推薦	申請	總和
錄取	24	36	60
不錄取	36	54	90
總和	60	90	150

所有錄取的 60 名學生中，

經由「推薦」入學者的比例為 $\frac{24}{60}=40\%$.

經由「申請」入學者的比例為 $\frac{36}{60}=60\%$.

(2) 推薦者的錄取率為 $\frac{24}{60}=40\%$,

申請者的錄取率為 $\frac{36}{90}=40\%$,

總錄取率 $\frac{60}{150}=40\%$, 三者相同. ☒

在例題 2 第(1)小題中，所有錄取者中，由「推薦」而錄取者占 40%，「申請」而錄取者占 60%，是否可解讀為申請比推薦容易呢？這個錯覺是因為申請的人數比推薦的人數多所造成的．事實上，由第(2)小題得知，推薦與申請的錄取率都是 40%，我們合理的推測：「入學方式」和「通過與否」並沒有關聯．

雙向表的目的既是探討兩事件的關聯性是否存在，換成機率的語言就是兩事件是否獨立，上述的例子，在所有參加推薦和申請的學生中，設 A 表示參加推薦者，B 表示錄取者，則 $P(B \mid A)$ 表示推薦者的錄取率，而 $P(B)$ 表示總錄取率．在這個例子中，

$$P(B \mid A) = P(B).$$

即兩事件 A，B 為獨立事件．換句話說，在雙向表中，若某一列的各數據在其所在的行中所占比例皆相同，兩特性就沒有關聯．至於比例不相等時是否就代表有關聯呢？統計學上有更深入的檢定辦法，留待日後再學習之．

◆有兩點必須指出：

(一)由男女錄取率的相等與否，來判定錄取與男女性別是否有關，並不正確。

這非統計思維。除非事先設定男女錄取率一定要相同(這時男女錄取標準，就很難相同)，否則即使用抽籤(這時錄取與否總該跟性別無關)，來決定錄取名單，都不能保證抽出的男女錄取率相同。

(二)不應將觀測值視為機率：

$$P(A|B) = 40\%$$

是錯的概念。

這點人們平常其實大都了解。例如，投擲銅板100次，出現52次正面，並不會將0.52當做正面出現的機率(為估計值)；也不會將一次民調的支持率，當做候選人的得票率。但統計只要一進入高中數學課本，就連常識都失去了。

- 甚至只要一進入書中，該有的機率知識也不見了。

數據過度完美，是好還是不好？

孟德爾實驗

孟德爾(Gregor Johann Mondel)將圓黃(round yellow)種子的豌豆，與縐綠(wrinkled green)種子的豌豆雜交。依其理論，會生長出圓黃、圓綠、縐黃及縐綠種子的後代之比率，應分別為

$$\frac{9}{16}=56.25\%\ ，\frac{3}{16}=18.75\%\ ，$$

$$\frac{3}{16}=18.75\%\ ，\frac{1}{16}=6.25\%\ 。$$

經由一組有556個樣本的實驗，得到下表。

	圓黃	圓綠	縐黃	縐綠	合計
後代數	315	108	101	32	556
觀測比率	56.65%	19.42%	18.17%	5.76%	100%
預期比率	56.25%	18.75%	18.75%	6.25%	100%

豌豆觀測到的後代比率，與預期比率有些差異。

經過卡方檢定，即使α值大到0.90，都無法拒絕

虛無假設：孟德爾的理論為正確。

是否無庸置疑接受孟德爾的理論？

- 此實驗結果與預期太吻合(fit too well)，引起費雪(R. A. Fisher, 1890-1962)的懷疑：

  認為孟德爾可能重覆做實驗，直到結果看起來很好才停止，只公佈結果較好的那組數據。

- 對於隨機實驗，若結果與理論值過於一致，反而會讓人懷疑做假：

  完美的數據，絕對的懷疑！

  與數學思維大異其趣。

◆有人看到數據就胡亂說話，有人則從數據取得的程序便有問題。一旦數據不可靠，說的話還可信嗎？甚至不少人對統計方法不求甚解，往往得到一些經不起任何考驗的結果。因此統計一方面被視為做決策之重要依據，一方面又常讓人嗤之以鼻，落入曾任英國首相迪斯雷利(Benjamin Disraeli，1804-1881)所說的：

有三種謊言：謊言，可惡的謊言，及統計。

◆這句話由於美國著名小說家馬克吐溫(1835-1910)在其自傳中引用，而廣為流傳。

謝謝各位！