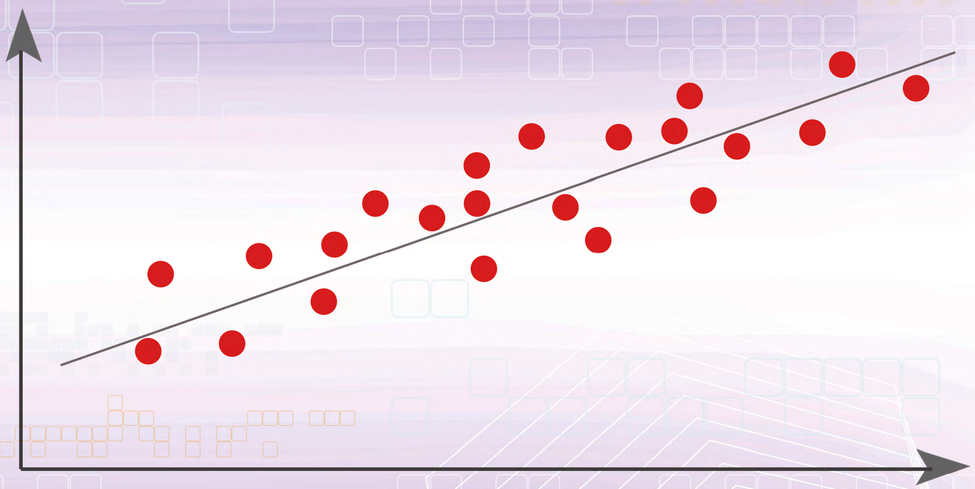
數學百子櫃系列（十九）

2014/15 中學生統計創意寫作比賽作品集



教育局
課程發展處數學教育組

數學百子櫃系列(十九) 2014/15 中學生統計創意寫作比賽 作品集

教育局數學教育組

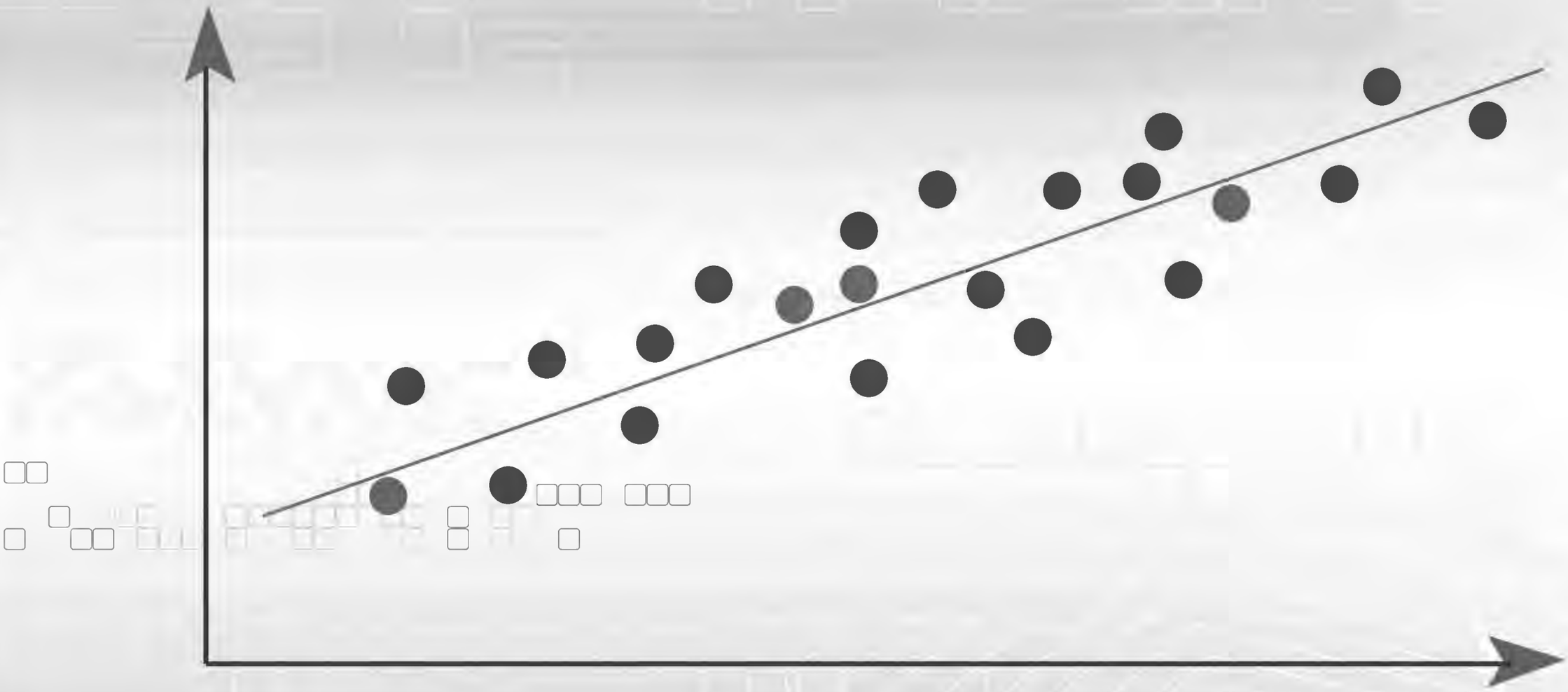
ISBN 978-988-8159-88-8

教育局數學教育組編訂
政府物流服務署印

Prepared by the Mathematics Education Section,
the Education Bureau of the HKSAR
Printed by the Government Logistics Department

數學百子櫃系列（十九）

2014/15 中學生統計創意寫作比賽作品集



教育局
課程發展處數學教育組

ISBN 978–988–8159–88–8

編者的話

為配合香港數學教育的發展，並向教師提供更多的參考資料，課程發展處數學教育組於 2007 年開始邀請大學學者及資深教師撰寫專文，以及蒐集及整理講座資料，輯錄成《數學百子櫃系列》。本書《2014/15 中學生統計創意寫作比賽作品集》，是這個系列的第十九冊。本書輯錄的文章，大部分是「2014/15 中學生統計創意寫作比賽」的優勝作品，由參賽的中學生撰寫。

本書所輯錄的參賽作品嘗試透過統計創意寫作，以簡潔的語言輕鬆地介紹概率和統計的知識。

本書共有 14 篇文章，第 1 至 9 篇為「2014/15 中學生統計創意寫作比賽」的冠軍、亞軍、季軍和優異作品。其餘 5 篇則為邀請作品，分別由政府統計處的統計師，數學教育組的課程主任，以及香港大學統計及精算學系的教授撰寫，供各讀者們閱覽。本書的文章，內容有趣，期望讀者閱讀後能增加統計知識，並能善用「統計」這項客觀、邏輯和系統性的工具決策、解難。

此書得以順利出版，實有賴這次比賽的籌備委員會成員所

付出的努力。在此，謹向撰寫作品的得獎隊伍、政府統計處的統計師、香港大學精算及統計學系的教授和數學教育組同工致以衷心的感謝。最後，更要多謝這次比賽的籌備委員會主席楊良河博士和總評審主任張家俊博士。兩位鼎力協助，審訂本書的內容，讓學生能夠閱讀更多有趣的文章，增強他們學習統計的興趣。

如對本書有任何意見或建議，歡迎以郵寄、電話、傳真或電郵方式聯絡教育局課程發展處數學教育組:

九龍油麻地彌敦道 405 號九龍政府合署 4 樓
教育局課程發展處
總課程發展主任(數學)收
(傳真: 3426 9265 電郵: ccdoma@edb.gov.hk)

教育局課程發展處
數學教育組

前言

香港統計學會一直致力向社會各界推廣對統計的認知。除了每年與教育局合辦「中學生統計習作比賽」（SPC），以鼓勵同學透過團隊合作形式學習正確運用統計數據及增進對社會的認識外，我們於 2009 年再與教育局合作創辦「中學生統計創意寫作比賽」（SCC），旨在鼓勵學生透過創意的手法，以及科學和客觀的精神，用文字表達日常生活所應用的統計概念或利用統計概念創作一個故事。

回顧過去的參賽作品，喜見同學們對統計概念有更深入的認識及掌握如何正確地運用統計。近年，得獎作品的質素亦有所提升。本年度的比賽專題是「線性迴歸」。我們十分感謝香港大學統計及精算學系鍾玉嘉博士在比賽簡介會中介紹有關線性迴歸的概念，並鼓勵同學在這課題上發揮創意。繼承以往的優良成績，今屆的 SCC 收到約 50 份參賽作品，當中不乏精彩之作。文章取材創新，趣味盎然；同學能活學活用各種統計和概率的知識，分析有條有理，見解獨到，言之有物。中學生能有這樣的水平，實在難能可貴，值得欣喜和嘉許。

本書輯錄了今屆所有的得獎作品，藉此嘉許得獎同學所付

出的努力。希望同學能夠從創作或閱讀這些得獎作品中得到啟發，對統計的知識及其運用有更深入和正確的理解。值得一提的是，本書所輯錄的得獎作品均未經任何修改或更正。當中的某些敍述或分析，特別是對線性迴歸這一個較深的課題的運用和演繹，可能有欠準確。常見的問題包括誤將相關關係演繹為因果關係、胡亂應用線性迴歸於各種變量以至過分解讀數據之間的關係、未有探討線性迴歸分析是否適合應用於特定數據（如未經變換的時間序列數據）等。同學們在閱讀本書時請多加留意，遇到疑問或不清楚的地方可向老師請教。

我們藉此機會感謝籌備委員會和評審委員會全體成員對評審的幫助和支持。他們的不遺餘力無疑是有助提高學生對統計的認知和興趣。最後，感謝香港大學統計及精算學系贊助今屆比賽的最佳專題寫作獎，和理大香港專上學院贊助今屆比賽的最佳文章演繹獎。

籌委會主席　楊良河博士

總評審主任　張家俊博士

2015 年 10 月 18 日

目錄

編者的話 ii

前言 iv

目錄 vi

冠軍作品: 《書中自有黃金屋?》……1

亞軍: 探索藝術體操的難度分與完成分的奧妙關係 12

季軍作品: 誰是誰非之「你」快落！我快樂！ 28

優異作品: My Brother vs My Mother.....44

優異作品: 線性迴歸 63

優異作品: Hot Pot Dots 73

優異作品: 成績「分分」跌？「分分」賞！ 83

優異作品: 作弊 94

優異作品: 舊夢不須記?——被遺忘的香港樂壇 101

邀請作品：足印統計 116

邀請作品：車程遠，自然收得貴？—研究港鐵車費訂定的方程式 1119

邀請作品：統計釋疑 128

邀請作品：從球員身價到星系擴張–淺談齊夫定律 134

邀請作品：如何建立一個數據可視化 135

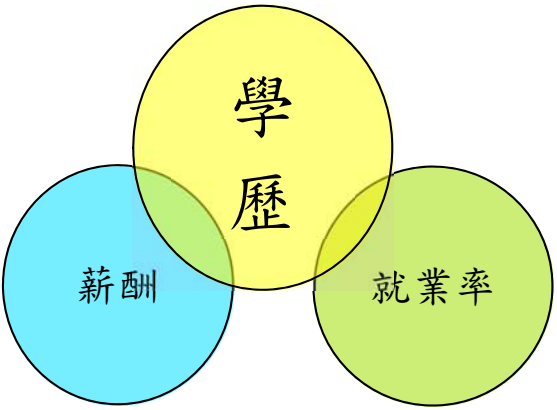
冠軍作品:《書中自有黃金屋?》

學校名稱：保良局何蔭棠中學

學生姓名：何嘉琪，李雯靜，謝富賢

級別：中五

指導教師：陳智仁



引言

古語有云:「書中自有黃金屋，書中自有顏如玉」是出自宋代《勸學詩》，但到了 21 世紀，市場、社會結構和宋代截然不同的情況下，究竟學歷與工作前途有何關係? 「書中自有黃金屋，書中自有顏如玉」在現代又是否成立?

「最低工資於今天正式實施，首次最低工資定於時薪 28 元。」這是 2011 年 5 月 1 日的重要議題。2013 年 5 月 1 日調升至時薪 30 元。直到今天，大家仍然眾說紛紜。恰巧，小明和小美亦正在討論。

「小美，現時香港於最低工資的保障下，我們的出路究竟如何？」「我不看好我們的將來。現在普遍大學生踏足社會工作，月薪跟中學畢業生差不多，有些更要租借劏房居住呢！還有，青年失業日趨嚴重，這是個值得關注的問題。」

「你說得對！我真不明白在這知識型社會上，大學生需花多四年時間和繳交龐大的學費後，畢業後卻與中學畢業生的薪酬相差無幾。」究竟學歷高低與大學畢業生的前途能否扣上關係？現時失業的大學畢業生比比皆是，即便擁有工作，工資也不見得優越。正因如此，我們是次研習的探討方針鎖定學歷對畢業生的前途有何關係。工作的前途是指能否得到優厚的工資和工作穩定性，例如：職位是否容易被替代? 這些都是我們研究的重點。

很多人都認為最低工資實施後，學歷與前途就再沒有關係，花多四年時間和學費升讀大學倒不如早日踏足社會工

作。以最新修定的最低工資下工作每小時的薪金是$30，假設一位中學生畢業後每月工作 25 天，每天 12 小時，他每月薪金便是：25x12x30 = $9,000。而普遍大學生剛畢業入職的薪金也只是大約$10,000 至$12,000，只是相差$1,000 至$3,000 而已。

工資:

「雖然我於四年間沒有收入，可是擁有較高學歷可以找到一份較好的工作！」正當他們正爭執得面紅耳赤時，讓我們看看下列的迴歸圖。（圖 1）最初，一位大學生，初中甚至高中畢業生工作，在最低工資影響下，大家的起薪點均差不多，大約為一萬元。當入職時間越長，年紀愈大，不同學歷的工資起了變化。到了 25–34 歲時，大專畢業生大約有$20000，而初、高中畢業生的工資卻只有$10000–$12000。過了壯年時期後，大專以上的學歷的工資大約保持$18000–$20000，初中及高中畢業生的工資不升反跌。因此迴歸圖反映了不同學歷的人工作，工資幅度亦有不同。大專以上學生的工資會隨着任職時間增長，而中、小學生的工資卻只有輕微的升幅。這是歸究於大專以上的學生擁有較高學歷，因此能從事較少依賴勞動力的工作，例如：管理層，行政總裁；相反，中、小學生由於擁有較低學歷，因此只

能尋找需較多依賴勞動力而低薪的工作，例如：地盤工人。

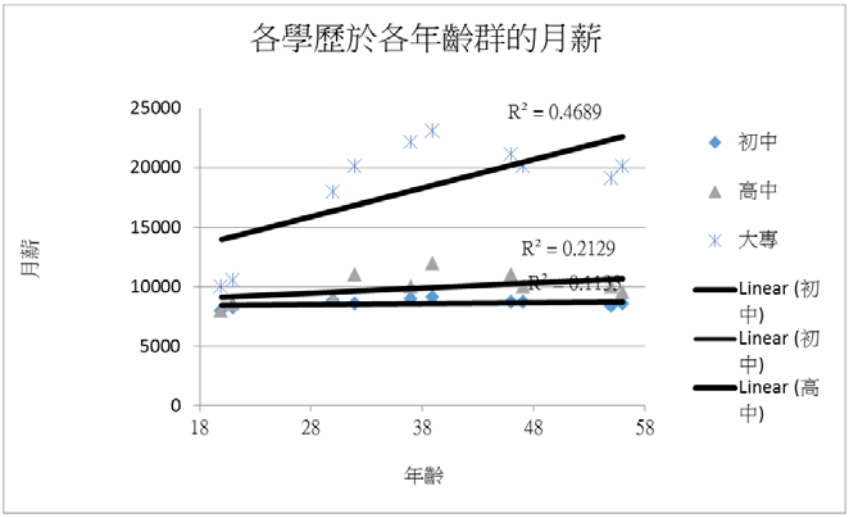


圖 1

失業率:

失業率一向都是值得關注的議題。失業率指一個人願意並有能力為獲取報酬而工作，但尚未找到工作的情況。就讓我們觀察下列的迴歸圖(圖 2)。起初，一位初中及高中畢業生或大專畢業生以上的學生踏足社會第一次尋找工作時，大家的失業率都維持差不多的比率大約 3.5%–4%。可是，當我們觀察迴歸圖的中段和尾段，趨勢各異。一位小學畢業生隨着年紀的增長，失業率日漸增長，到老年時，失業

率是在三種組別的畢業生裏持續高企。然後，當年齡增長，一位中學畢業生的失業率比開端時有明顯下降。可是，仍然比大專以上的學生較多。當大專以上畢業的學生逐漸年長。由此可見，普遍低學歷，小、中學的畢業生於老年失去工作的機會率遠比大專以上的學生為高。這是由於低學歷者只能尋找高勞動力的工作，在壯年時，身體能承受如此重擔，可是，在老年時，勞動力會大大減低，因此，工作並不長久；相反，高學歷者的工作普遍依賴知識，因此，於老年時仍能繼續工作，失業率比低學歷者為低。

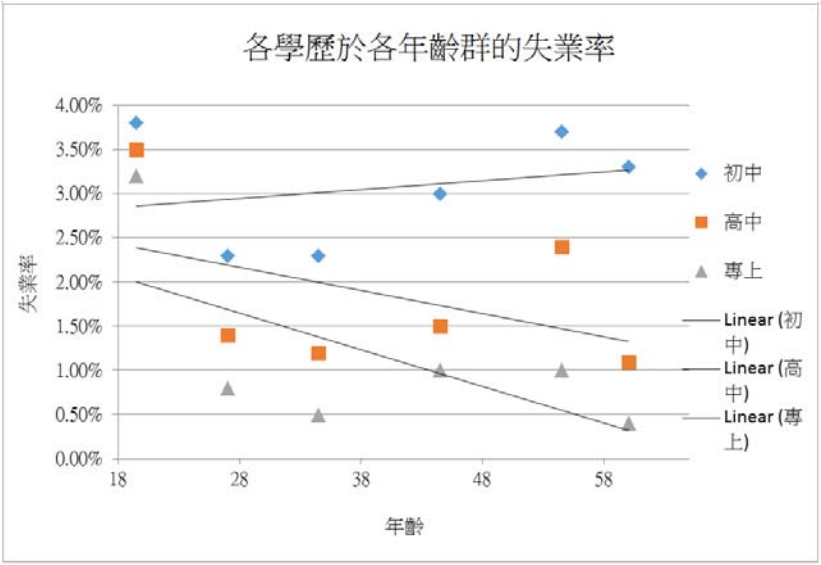


圖 2

「小美，你看着我的迴歸圖，我把三組不同學歷，初中及高中和大專以上畢業生及不同年紀(25 歲至 50 歲以上)作了一個統計，研究了他們隨着年紀增長的薪酬 (圖 3) 和失業率 (圖 4) 的關係。我發現了當三組不同學歷者年紀越大，他們的工資也隨着上升，而三組人的失業率也下降。根據我的迴歸圖可推翻你剛才的結論 — 學歷對工作的前途有正面影響，而我的結論是截然不同，由於迴歸的 R^2 值只是 2.52% 及 5.98%，可見學歷與工資及失業率之間的關係極之薄弱，甚至是毫不相關，兩件是獨立的事件呢！」小明說。

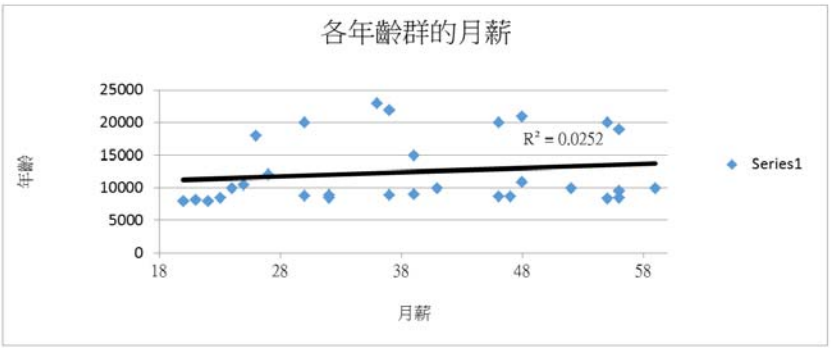


圖 3

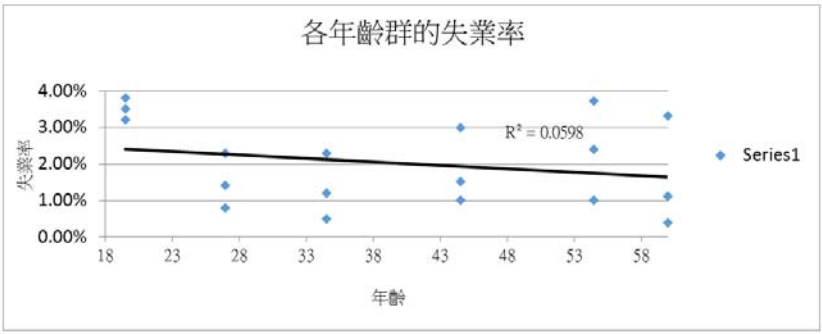


圖 4

「讓我看看你的迴歸圖。你的迴歸圖有很大的錯誤導致你的數據並不可靠。」小美說。

「哪裏出錯了？」小明說。

「你看看我的迴歸圖就會明白了。你把三組不同學歷和不同年齡放在同一個圖表上，然後把三組數據歸納了一起成為一條迴歸線的這個方法並不恰當，因為我們的焦點是研究不同學歷會否影響將來工作的前途，然而，你的迴歸圖把三組學歷當成一組人統計，因而得到的結果會是整體和普遍的情況，未能清楚和獨立看待不同學歷，因此，從你的迴歸圖不能得出最佳的結論。再看看我的迴歸圖(圖 5，6)，我也是把三組數據放在同一圖表上，但我得到了三條不同的迴歸線，這是由於我把小學，中學和大專以上的學生當

了三個獨立事件。」

「那麼你的結論是甚麼？」小明問。

「大專以上的學生隨着年齡逐漸增長失業率會下降，工資會上升，高中畢業生隨着年紀的增長，失業率輕微下降，工資也只有輕微的上升；初中畢業生的情況更糟糕，失業率隨着年齡上升，工資卻下降，而 R^2 值亦於 20%–40%，代表學歷與工資及失業率有直接及緊密的關係，這個才是處理迴歸圖的正確手法和得到準確無誤的結論呢！因此，學歷愈高，工作前途便會更美好。」小美說。

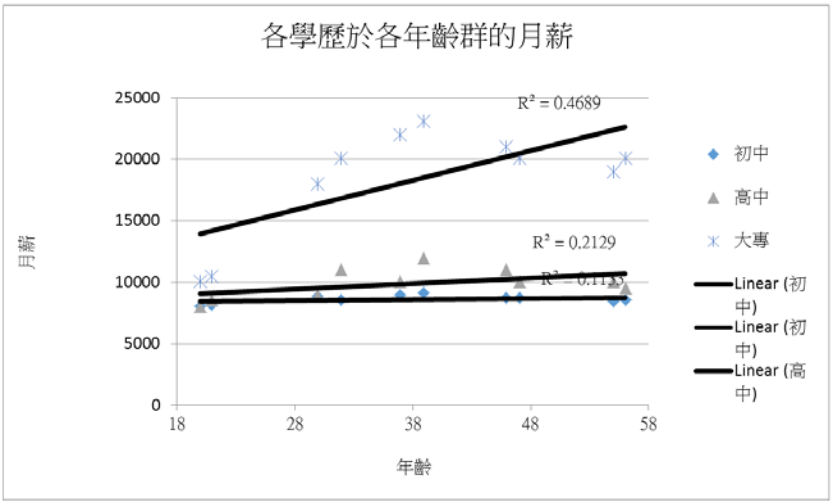


圖 5

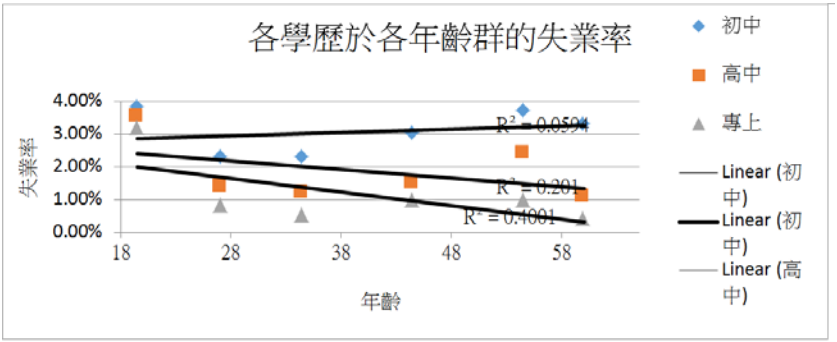


圖 6

各學歷於各年齡工資與失業率中位數

年齡群	**15–24**	**24–34**	**35–44**	**45–54**	學歷
工資	$8000	$8800	$9000	$8700	初中
	$8500	$12000	$15000	$11000	高中
	$10000	$18000	$22000	$20000	大專
失業率	23%	23%	30%	37%	初中
	14%	12%	15%	24%	高中
	8%	5%	1%	1%	專上

以上表格為各學歷於各年齡工資與失業率中位數可見無論於各年齡群，大專的工資比高中與初中高而失業率比兩者低。可見學歷越高越能覓得一份高薪及穩定的工作。

最終，我們知道最低工資是無法彌補低學歷者的競爭力，他們將來工作的前途和所走的路決定的高歷者更差，走得更長更辛苦。每個人都希望將來能擁有穩定的生活，而現階段，當學生能做的只有用功及勤力讀書。「書中自有黃金屋。」是成立的和恆久不變的規律。共勉之。

(~2360 字)

參考資料

1. 2008 年半年經濟報告–青少年失業率分析
2. 2014 年第三季經濟報告–勞動力參與和教育程度的關係
3. 工資及薪金總額按季統計報告(2014 年 9 月版)
4. 2012–2014 年按性別、年齡組別、教育程度的每月工資中位數
5. 法定最低工資修訂條文
6. 2014 年 7 月 4 日蘋果日報

亞軍作品：探索藝術體操的難度分與完成分的奧妙關係

學校名稱：順德聯誼總會李兆基中學

學生姓名：蔡佩如、朱倪賢、李晉添

級別：中四

指導教師：許俊江

引言

藝術體操是奧運項目之一，由難度分和完成分來評分，正常來說越高難度越難完成，即是難度分和完成分應成反比，但事實又真的如此嗎？我們來一起探討吧。

蔡佩如（蔡）、朱倪賢（朱)、李晉添（李）

李： 蔡佩如，我妹妹對藝術體操頗有興趣，其實藝術體操是甚麼來的？

蔡： 藝術體操是一項徒手或手持輕器械在音樂伴奏下進行的體育運動項目。

藝術體操在正式比賽時，必須以預先準備好的整套動作的形式出現，比賽場地面積為 12 米x12 米，其周圍至少有一米寬的安全區。運動員可根據自己的個人風格和比賽規則的要求，將所掌握的各種單個動作組合而編排成一套完整的動作（稱為套路）來比賽。現時正式比賽有五種器械，分別為球、圈、繩、棒及絲帶。

藝術體操是以自然性、韻律性動作為基礎，動作優美富藝術感，身體各種關節得到充分的活動，各部肌肉得到均衡的發展。藝術體操不但塑造美的體型，而且陶冶性情。

李： 明白了！聽起來很吸引。我想問比賽是如何計分的？

蔡： 我剛好在看 2014 世界藝術體操錦標賽的圈項賽事，每個參賽者大約有一分半鐘時間去完成比賽，比賽的總得分主要是難度分和完成分的兩項得分總和。難度分是對動作難度的評分，完成分是對能否完美地完成每個動作而作出的評分。你可以看以下首一百名的選手在圈項所得的分數。

R : Rank 排名 D : Difficulty 難度分 E : Execution 完成分 P : Penalty 扣分

T : Total 總分 = D + E – P

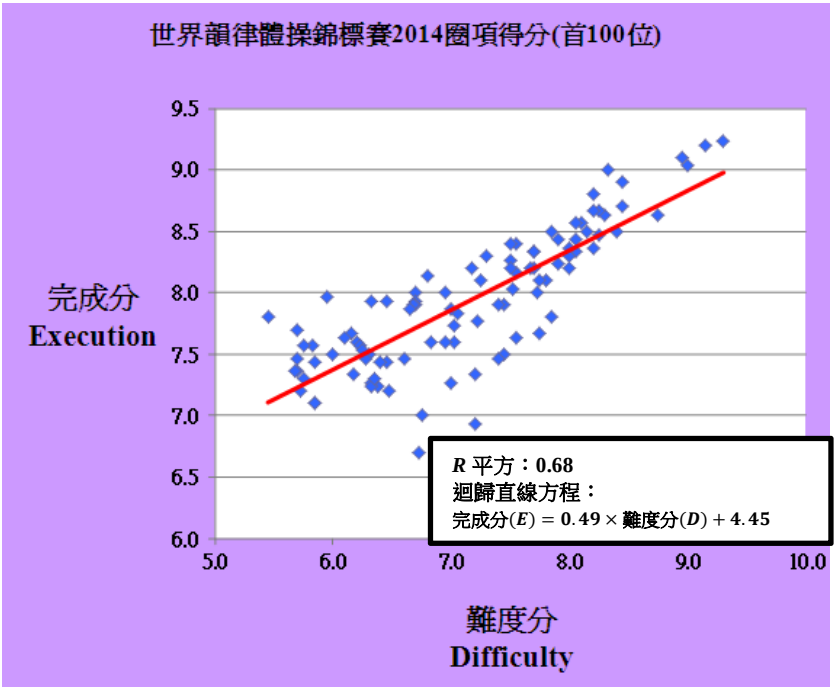
R	D	E	P	T	R	D	E	P	T
1	9.3	9.2	0	18.5	26	7.9	8.2	0	16.1
2	9.2	9.2	0	18.4	27	7.7	8.3	0	16.0
3	9.0	9.1	0	18.1	28	7.7	8.3	0	16.0
4	9.0	9.0	0	18.0	29	7.6	8.4	0	16.0
5	8.8	8.6	0	17.4	30	7.5	8.4	0	15.9
6	8.5	8.9	0	17.4	31	7.7	8.2	0	15.9
7	8.3	9.0	0	17.3	32	7.8	8.1	0	15.9
8	8.5	8.7	0	17.2	33	7.7	8.2	0	15.9
9	8.2	8.8	0	17.0	34	7.8	8.1	0	15.9
10	8.3	8.6	0	16.9	35	7.7	8.0	0	15.7
11	8.3	8.7	0	16.9	36	7.5	8.3	0.05	15.7
12	8.4	8.5	0	16.9	37	7.6	8.2	0	15.7
13	8.2	8.7	0	16.9	38	7.5	8.2	0	15.7
14	8.3	8.5	0	16.7	39	7.3	8.3	0	15.6
15	8.1	8.6	0	16.7	40	7.5	8.0	0	15.6
16	8.2	8.5	0	16.7	41	7.8	7.7	0	15.4
17	8.1	8.6	0	16.6	42	7.2	8.2	0	15.4
18	8.2	8.4	0	16.6	43	7.3	8.1	0	15.4
19	8.1	8.4	0	16.5	44	7.5	7.9	0	15.4
20	8.1	8.3	0	16.4	45	7.9	7.8	0.3	15.4
21	8.0	8.4	0	16.4	46	7.4	7.9	0	15.3
22	7.9	8.5	0	16.4	47	7.6	7.6	0	15.2
23	7.9	8.4	0	16.3	48	7.2	7.8	0	15.0
24	8.0	8.3	0	16.3	49	7.0	8.0	0	15.0
25	8.0	8.2	0	16.2	50	7.5	7.5	0	15.0

R	D	E	P	T	R	D	E	P	T
51	6.8	8.1	0	14.9	76	6.2	7.6	0	13.8
52	7.1	7.8	0	14.9	77	6.3	7.5	0	13.7
53	7.0	7.9	0	14.9	78	6.1	7.6	0	13.7
54	7.4	7.5	0	14.9	79	6.3	7.5	0.05	13.7
55	7.0	7.7	0.05	14.7	80	6.5	7.2	0	13.7
56	6.7	8.0	0	14.7	81	6.4	7.2	0	13.6
57	6.7	7.9	0	14.6	82	6.3	7.3	0	13.6
58	7.0	7.6	0	14.6	83	6.3	7.2	0	13.6
59	6.7	7.9	0	14.6	84	6.4	7.3	0.1	13.6
60	6.7	7.9	0	14.6	85	6.0	7.5	0	13.5
61	7.0	7.6	0	14.6	86	6.8	7.0	0.3	13.5
62	7.2	7.3	0	14.5	87	6.7	6.7	0	13.4
63	6.7	7.9	0	14.5	88	5.7	7.7	0	13.4
64	6.8	7.6	0.05	14.4	89	5.8	7.6	0	13.4
65	6.5	7.9	0.05	14.3	90	5.9	7.4	0	13.3
66	7.0	7.3	0	14.3	91	5.8	7.6	0.05	13.3
67	6.3	7.9	0	14.3	92	5.5	7.8	0	13.3
68	6.6	7.5	0	14.1	93	6.2	7.3	0.3	13.2
69	6.0	8.0	0	13.9	94	5.7	7.5	0	13.2
70	6.5	7.4	0	13.9	95	5.7	7.4	0	13.1
71	6.4	7.4	0	13.8	96	5.8	7.3	0	13.1
72	7.2	6.9	0.3	13.8	97	5.7	7.4	0.05	13.0
73	6.2	7.7	0	13.8	98	5.9	7.1	0	13.0
74	6.2	7.6	0	13.8	99	5.9	7.1	0	13.0
75	6.3	7.5	0	13.8	100	5.7	7.2	0	12.9

* 難度分、完成分及總分經過四捨五入後可能導致加減有誤差

李： 咦，似乎難度分及完成分兩者有明顯的正向關係。我起初還以為難度越高的動作越難完成，兩者會成反比關係。

朱： 是啊，我初中時曾學過散點圖(Scatter diagram)，散點圖是用來表達兩組數據之間的關係，我們可以用散點圖來表達上述的資料。**圖一**共有一百點，每一點代表一個運動員，點的位置是由該運動員所得的難度分和完成分而決定的。



圖一

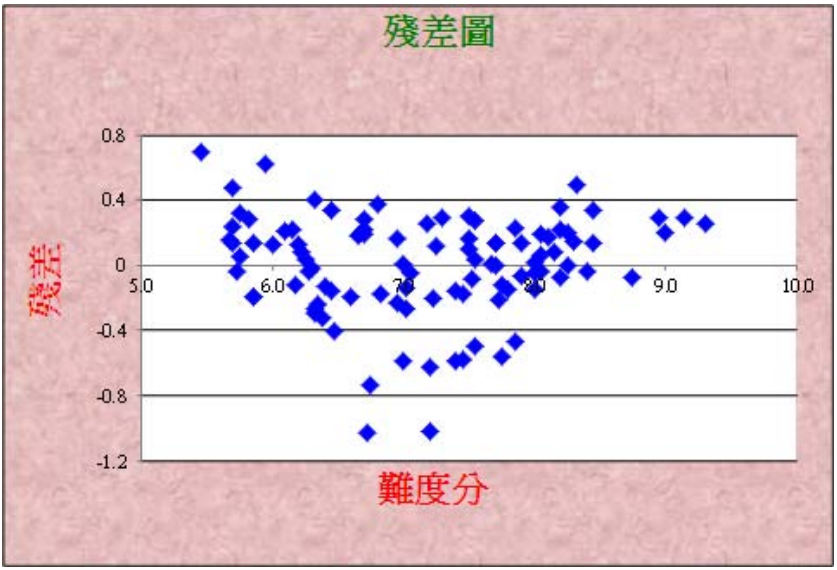
李：　這幅圖清楚表示參與初賽的首 100 名選手的難度分和完成分有明顯的正向關係，難度分愈低的選手，完成分亦相對地低，相反難度分愈高的，完成分亦會較高。咦，但圖中紅色直線是代表甚麼？

朱：　$e_i = E_i - \hat{E}$（i＝第i名運動員、e＝誤差、E＝難度分、$\hat{E}$＝預測的難度分）

迴歸直線能夠將 $\sum e_i^2 = e_1^2 + e_2^2 + e_3^2 + \cdots e_{100}^2$ 最小化。

蔡： 哦？即使不做任何難度亦有 4.45 完成分？那我現在去參賽是不是亦有 4.45 底分？但我見過有運動員的完成分是低過 4.45 分。

朱： 我猜意思是即使運動員選擇很容易的動作，普遍亦會有高於 4.45 的完成分。當然，迴歸直線是由難度分得分大約為 5.5 至 9.5 的運動員而得出，你要起碼得到難度分 5.5 的情況下，應用這條迴歸直線才會比較準確。情況就像你以 0 到 10 歲孩童的身高資料得到的迴歸直線，去預測 60 歲成人身高，你可能會得出高 6 米的巨人，這叫做外推(extrapolation)預測，未必準確。



圖二

朱： 另外，迴歸直線雖然把誤差縮到最小，但是誤差依然存在，迴歸直線還未能解釋 38%數據的變化，我們能利用每一個資料與迴歸直線之間的誤差，再畫出一幅殘差圖（圖二），殘差就是誤差的意思。在殘差圖中我們能看到，難度分較高(> 8)及較低(< 6)的人，誤差較小，但難度分中游的人(6 至 8)，誤差的數值較大。你會如何理解這個現象？

李： 難度越低越容易完成，所以運動員完成分差距不

多。難度越高則越難完成，由於好壞參差，完成分差距較大。但為甚麼難度分最高者，得分差距反而減少？這個我不能理解。

蔡： 我猜那些表演高難度套路的運動員定必是全世界最優秀的，她們經常參與國際賽事，有豐富的經驗及穩定的表現，所以得分差距不大。

李： 原來如此！另外我有一個想法。既然運動員選擇的難度越高，完成分也相對較高，總分亦較高，為甚麼不是所有運動員都選擇難度高的套路？

蔡： 因為不是每個運動員都能夠完成指定難度的套路，所以運動員所選的套路的難度會受到運動員的技術所局限。教練編套路的難度會因應運動員個人的能力及水平，避免編排高難度的動作。運動員未必能在比賽前熟練高難度的套路，此舉亦會增加運動員受傷的機會。

朱： 那就是說，縱使 x 和 y 兩項數據有正向或反向的關係，但它們可能沒有任何因果關係。例如香港出生

率不斷下降，但香港法定最低工資卻不斷增加，你總不能說是因為出生率下降導致最低工資上升。所以縱然兩者有反向的關係，但是兩件事情之間未必有因果關係。

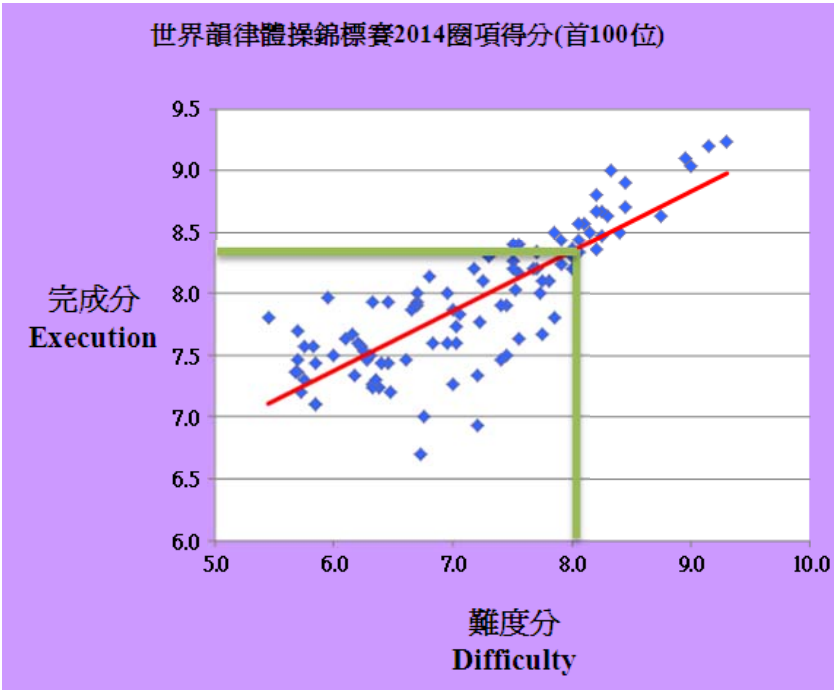
正如這次 x 為難度分，y 為完成分，它們雖然有正向關係，但是兩者都是受到選手個人的技術水平所影響，技術愈高者，便會挑戰高難度的套路，亦能以較高水平完成，並不是由難度分影響完成分。

蔡： 既然 x 的改變不導致 y 的改變，那這幅圖豈不是沒有任何作用？

朱： 那當然不是了。既然你挺熟悉藝術體操，你也知道每一個選手比賽時的套路都是由他們的教練所編排的，而教練會基於選手的技術而安排一套適合他們的套路，教練可以利用這幅圖估計選手能得到的完成分，例如教練安排了一套難度分 8 分的套路，基於這幅圖或利用迴歸直線方程，我們可以計算出完成分為 8.37 分。

$$\text{完成分}(E) = 0.49 \times \text{難道分}(D) + 4.45 = 0.49 \times 8 + 4.45 = 8.37$$

可見，這幅圖也不是完全沒有它的用處的。



朱： 假若運動員得分比預期的完成分較高，則代表該套路對運動員太過輕易，運動員可挑戰更高難度的套路；相反若運動員得分比預期的完成分較低，則代表該套路太難，教練可考慮調整套路的難度。

蔡： 噢！原來如此！

朱： 總結來說，我們可以整合兩項數據，然後用線性迴歸來估計出它們之間的關係，而迴歸直線能否反映 x 影響 y 的程度便要透過計算 R^2，R^2 愈接近 1 代表 x 愈能反映 y 的數值，反之亦然。迴歸直線的 R^2 常常都不等於 1，代表真實數據與直線預測的數值有所差異，而我們可以利用這些誤差畫出一幅殘差圖，能看到不同誤差中的變化。雖然有些時候，兩項數據有正反向關係而沒有因果關係，但是我們也能透過回歸直線預測結果。由此觀之，線性迴歸用途廣泛。正如統計學家 George Box 所言：「Essentially, all models are wrong, but some are useful.」

（~2364 字）

參考資料

1. 2014 世界韻律體操錦標賽
   http://www.rhythmicgymnasticsresults.com/worlds/2014/izmir.html

2. 2014 世界韻律體操錦標賽圈項的首一百名運動員的分數
   http://www.rhythmicgymnasticsresults.com/worlds/2014/individuals/hoopqual.pdf

3. 宜蘭縣政府教育處對女子競技體操的描述
   http://sportmap.ilc.edu.tw/index.php?menu=5&id=25&cid=831

4. Box, G. E. P., and Draper, N. R., (1987), Empirical Model *Building and Response Surfaces*, John Wiley & Sons, New York, NY.

季軍作品：誰是誰非之「你」快落！我快樂！

學校名稱: 宣道會鄭榮之中學

學生姓名: 吳祺欣、黃凱鈴

級別: 中五

指導教師: 朱吉樑

引言

2009 年實施的「一簽多行」政策，令國內訪港旅客上升，有人認為此舉能刺激香港經濟，有利於零售業、旅遊業等行業，亦提供了工作的機會。亦有人認為旅客令香港物價不斷飆升，即使港人工資有所提高，也追不上通貨膨脹，增加了支出上的負擔，生活苦不堪言。到底「一簽多行」對於香港人有何影響？下文將加以分析。

背景：*「誰是誰非」是一個電視節目，主持人就生活所見所聞作訪問，並邀請嘉賓分析問題，探討事情的真相。*

HELEN:「誰是誰非、數理無欺」，歡迎大家收看今日節目，我係主持人 HELEN， 今天得到網民報訊， 提及上水有一群精打細算的家庭主婦，每日早上 8:00 集合開會，分工合作去不同地區，購買最便宜的生活用品，再把「戰利品」一同分享，她們為何要這樣做？ 今天我們找到其中一位參與者，嘗試了解一下！

HELEN: 你好啊！吳女士，不知能否抽一點時間讓我訪問一下你呢？

吳女士: 好呀！不過要好快，因為我要趕去大埔買紙巾、尿片、水果……，之後再落去深水埗幫人買奶粉、米粉、洗衣粉、生粉……。

HELEN: 無問題，請問你為何要每天一大清早便開會，又花那麼多時間去買東西呢？

吳女士: 因為近年香港的物價上升， 一個蘋果由 2 元升到

5 元，一斤菜由 5 元升至 10 元，我們這些低收入人士，如何能負擔？當然要珍惜一分一毫。

HELEN: 你認為近年香港的物價為什麼會上升得那麼急？

吳女士：你望一望火車站便知道，內地旅客大幅增長，上水區影響最嚴重，內地旅客一車一車買貨，導致上水的物價高漲，內地旅客在上水買完貨，又往其他地區買貨，香港貨品的物價也自然不斷上升！我們香港人就成了最大的受害者！

HELEN: 到底吳女士的說話有沒有根據？我們馬上回大本營，跟嘉賓朱教授和黃教授討論一下吧！誰是誰非、數理無欺。」

HELEN: 朱教授和黃教授兩位好。想問一問內地訪客是否對香港的物價造成影響呢？

黃教授：香港的物價升跌，一般會考慮消費物價指數(CONSUMER PRICE INDEX)，簡稱 CPI。不若我們先看一看近年 CPI 變化。(圖一)

不難發現，CPI 不斷上升，而升幅更加一年比一年厲害呢！

朱教授：當然，單從折線圖未必能考慮內地訪客對香港物價的影響，我們可透過散點圖(SCATTER DIAGRAM)，先看看旅客與香港物價的關係。(圖二)

為方便討論：先把以下變數以簡稱代表**:**

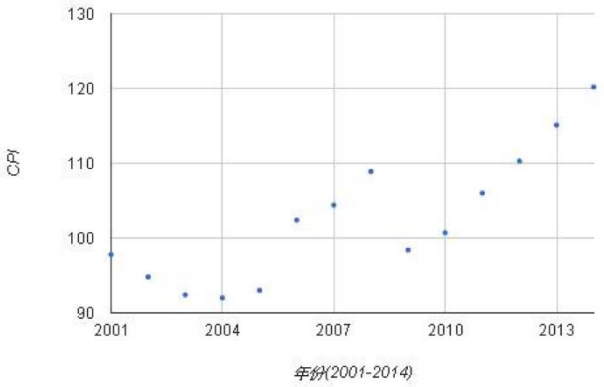
***T*：總旅客數量 *M*：內地旅客數量**

***N*：非內地旅客數量**

另外，數量以「萬」為單位。

圖一: CPI 歷年變化 (2001 – 2014)

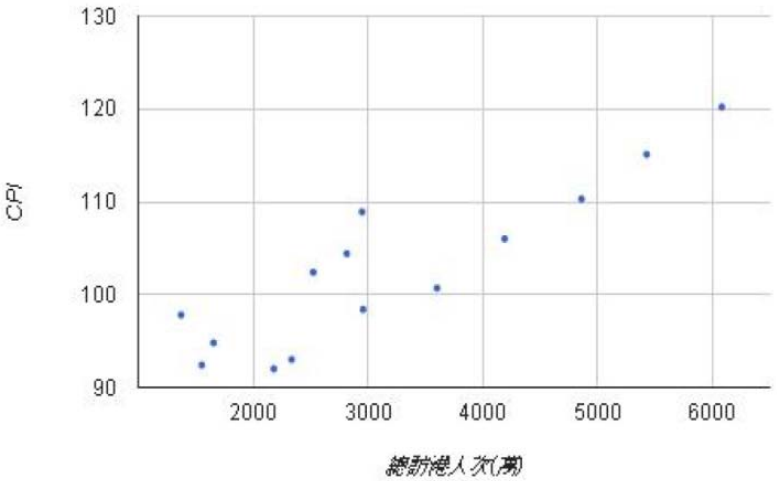
CPI 歷年變化(2001–2014)



朱教授：不難發現，兩者有正關係(POSITIVE RELATIONSHIP)，即訪客量愈大，CPI 亦愈高。

圖二：旅客與香港物價的關係

2001–2014 年總旅客訪港人次(萬) x CPI (綜合)



黃教授：不如進一步以線性迴歸的方法，看看訪客對 CPI 的影響有多大。先定義因變量(DEPENDENT VARIABLE)為 CPI，自變量為 T(總旅客量)，可得出以下結果。(表一)

表一：線性迴歸結果(CPI VS 總旅客量)

2001-2014 年度	CPI vs 總旅客
Model	**CPI**=85.7 (0.0000[1])+0.00532 (0.0000)**T**
P-value	0.0000154285793
R Square	0.8009139752

(1) 代表系數的 p–value

黃教授：從表一的結果中，*P* 值(*P*–VALUE)少於 5%，可知總旅客量對 CPI 在統計上有顯著(SIGNIFICANT)的影響，再觀察其決定系數(R^2)，可知約有 80%的 CPI 受到總體旅客的直線關係所影響，由此，可知旅客量的多少對 CPI 是有影響的。再從線性模型中，可知每增加一萬名旅客，CPI 將約有 0.005 的上升。

朱教授：為了更進一步考慮內地旅客的影響，我們可分別考慮內地與非內地旅客量對 CPI 的影響。(表二)

表二: 線性迴歸結果(CPI VS 內地旅客量與非內地旅客)

2001-2014 年度	CPI vs 內地旅客量	CPI vs 非內地旅客量
Model	**CPI**=90.7 (0.0000)+0.00592(0.0000)**M**	**CPI**=63.7 (0.0000)+0.0334(0.0006)**N**
P-value	0.00002895474874	0.0006286478823
R Square	0.7793099966	0.636453805

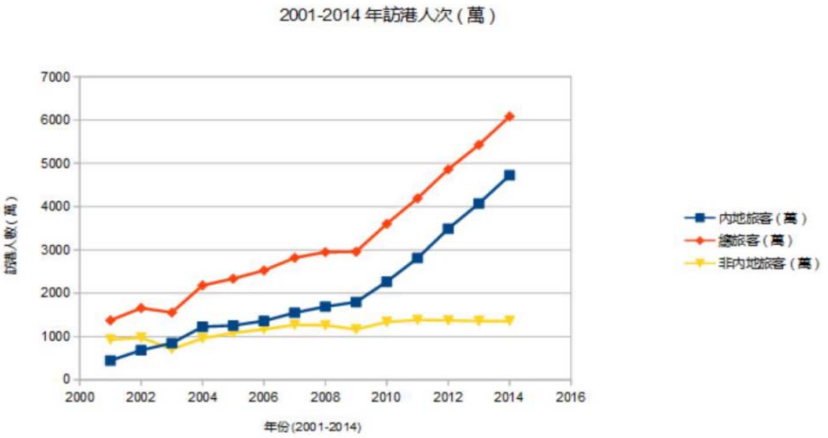
朱教授：通過相同的分析，可從表二中得知，內地與非內地旅客量都對 CPI 有顯著的影響，當考慮兩類旅客的啤打值(*b*)，每增加一萬名旅客，非內地旅客量對 CPI 有更大影響 (0.0334 > 0.00592)。

HELEN: 這樣看來，也不可只怪內地旅客把物價推高呀！

黃教授： 哈哈！你也太魯莽了吧！難道你忘記了在 2009 年才開始的「一簽多行」政策嗎？

朱教授：黃教授說得對，我們應該分開兩個階段計算，你先看看圖三，內地與非內地旅客的數量有明顯的不同，旅客量自 2009 年起便由內地旅客主導。

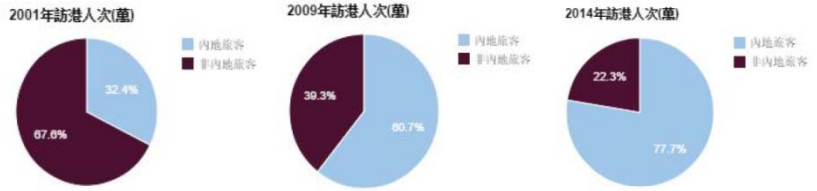
圖三: 旅客量的變化(2001–2014)



從 2009 年起，政府開放了一簽多行，准許中國大陸居民申請一年多次訪港個人遊。

朱教授：試試再看看圖四中，內地與非內地旅客量於 2001, 2009 及 2014 年的分佈，可見內地旅客量的比重愈見重要。因此，旅客量在 2001 年至 2008 年及 2009 年至 2014 年有明顯的分別，所以，較合理的做法，是分開兩個時段作比較的。

圖四: 旅客量的變化(2001–2014)



黃教授：從表三中可見，在未實施「一簽多行」(2001–2008)前，CPI 與內地旅客量未有顯著(INSIGNIFICANT)的關係(P–VALUE>5%)，即內地旅客數量不會影響香港的 CPI。反而，非內地旅客量對 CPI 有顯著影響(P–VALUE <5%)。

表三: 線性迴歸結果(CPI VS 內地旅客量與非內地旅客)

2001-2008 年度	CPI vs 內地旅客量	CPI vs 非內地旅客量
Model	NIL[3]	**CPI**=70.2 (0.0000)+0.0269(0.0156)**N**
P-value	0.1076153276	0.01562623738
R Square	0.3732393032	0.6501578991

朱教授：事實上，當考慮 2001–2008 時段時，總旅客量亦在統計上對 CPI 沒有影響呢(P–VALUE > 5%)！(表四)亦即內地旅客量的影響甚至令總旅客量當中的關係被掩藏了。

表四: 線性迴歸結果(CPI VS 總旅客量)

2001-2008	CPI vs 總旅客量
P-value	0.05066227233
R Square	0.497490544

黃教授：再來，看看 2009 – 2014 的分析，情況正正相反。(表五)

表五: 線性迴歸結果(CPI VS 內地旅客量與非內地旅客)

2009-2014 年度	CPI vs 內地旅客量	CPI vs 非內地旅客量
Model	**CPI**=84.3 (0.0000)+0.0075(0.0000)**M**	NIL
P-value	0.000004523611532	0.1998140003
R Square	0.9965288337	0.3703899339

因 *P* 值>5%，此模型未能解釋兩者關係，故不作表示。

黃教授：當實施一簽多行(2009–2014)後，內地訪港人數對香港的 CPI 會有顯著影響(SIGNIFICANT)，而決定系數(R^2)更達到 99.7%，即內地旅客的增長對香港的消費物價指數影響頗大呢！

HELEN: 不過，即使影響了物價，但也不一定影響了香港人的生活吧！

朱教授：這是一個好問題，我們可以試試考慮香港人的快樂感覺與 CPI 的關係呀！

HELEN: 教授，快樂一詞十分抽象，我們可以如何做呢？

朱教授：事實上，香港自 2005 年開始，亦有進行「快樂指數」的調查，在調查中引入有關「關愛」、「智慧」、「堅毅」和「行動」的問題，以了解有助產生導致快樂的「心理效益」的因素。

HELEN: 教授，是否可以再用迴歸方法考慮這個問題呢？

朱教授：當然可以，不過今次的變量在考慮上有一點不同。因為我們想探討 CPI 的變化如何影響香港人的快樂感覺，所以定義因變量(DEPENDENT VARIABLE)為「快樂指數」，而自變量(INDEPENDENT VARIABLE)為 CPI。

黃教授：由於時間關係，我已經把快樂指數與 CPI 於 2009 – 2014 年間的資料作了迴歸分析，結果可見表六。

表六: 線性迴歸結果(快樂指數 VS CPI)

2009 – 2014 年度	快樂指數 X CPI
P–value	0.7542455346 (> 5%)

HELEN: 以我了解，由於 *P* 值 大於 5%，即兩者沒有顯著關係，對不對？即 CPI 的變化，在統計上不會對快樂指數有影響！

黃教授：HELEN，你真聰明。

朱教授：不過，我們還可以再作分析。由於快樂指數中，也有分不同的收入級別的快樂指數，我們可以看看 CPI 對不同收入級別的人，會否有不同影響。大家看看表七。

表七: 線性迴歸結果(不同收入級別的快樂指數 VS CPI)

快樂指數(收入級別)	P-value
$1 - $9999	0.4828069348
$10,000 - $ 19999	0.03563023564 (<5%)
$20,000 - $ 29999	0.7651202976
$30,000 - $ 39999	0.5828257231
$40,000 or above	0.9307791604

表八：線性迴歸結果(「中下」級的中產人士快樂指數 VS CPI)

2009-2014 年度	快樂指數 ($10,000 - $ 19999) vs CPI
Model	快樂指數=97.0 (0.0005) - 0.263 (0.0356)CPI
P-value	0.035563023569 (<5%)
R Square	0.7083614735

根據香港政府統計處的定義，「中下」的中產階級一般是指家庭人數眾多，收入總和每月約一至兩萬元的公屋住戶。

黃教授：很有趣的結果呢！只有$10,000 – $19999 的組別，即「中下」級的中產人士，快樂指數會受到 CPI 的影響。而且啤打值(*b*)是負值(參看表八)，即 CPI 對快樂指數有相反的影響，即 CPI 上升會令到快樂指數下降！

HELEN: 為甚麼只有這一組別對 CPI 特別敏感呢？

朱教授：其實也不難理解，因為收入較高者對價格變動影響不大，舉例一個月入$40,000 的人，每月花在飲食上為$5000，即使食物價格升了 10%，對其支出也只是增加了$500，對日常生活影響不會太大。但對於月入$20,000 的人，即使每月花在飲食上相對較少，假設為$3000，在相同情況下 10%的增長是$300，相對的百分負擔則較明顯了

$$\left(\frac{300}{20000}\times 100\% = 1.5\%\right) \text{ VS } \left(\frac{500}{40000}\times 100\% = 1.25\%\right)$$

HELEN: 為什麼低收入組別又沒有影響呢？而一簽多行政策應否改變呢？

黃教授：這就是和政府綜援支助有關，低收入者得到入息補助，為經濟上無法自給的人士提供安全網，使他們的入息達到一定水平，以應付生活上的基本需要。

朱教授：所以吳女士應是「中下」級的中產代表，所以生活比較艱苦，要運用智慧生存 。至於，修改一簽多行政策，未必是最合適及有效的方法，反而政府應就受影響對象，即「中下」級的中產，增加支援，例如醫療津貼、 教育基金等，從而改善他們的生活水平，增加他們的快樂指數，不是更可行嗎?

HELEN: 多謝兩位教授。本集時間差不多了，下集再見!
(~2400 字)

附件

Year	CPI(綜合)	內地旅客(萬)	非內地旅客(萬)	訪港人次(總旅客)(萬)
2001	97.8	445	928	1373
2002	94.8	683	973	1656
2003	92.4	847	707	1554
2004	92	1225	956	2181
2005	93	1254	1082	2336
2006	102.4	1359	1166	2525
2007	104.4	1549	1268	2817
2008	108.9	1690	1260	2950
2009	98.4	1795	1164	2959
2010	100.7	2268	1335	3603
2011	106	2810	1382	4192
2012	110.3	3490	1372	4862
2013	115.1	4074	1356	5430
2014	120.2	4725	1359	6084

	快樂指數 ALL	CPI(綜合)
2009	70.6	98.4
2010	70.1	100.7
2011	71.3	106
2012	70.3	110.3
2013	70.5	115.1
2014	70.5	120.2

年份	CPI(綜合)	快樂指數 ($1 - $9,999)	快樂指數 ($10,000- $19,999)	快樂指數 ($20,000- $29,999)	快樂指數 ($30,000- $39,999)	快樂指數 $40,000 or above
2009	98.4	63.85	70.42	71.89	70.13	72.82
2010	100.7	66.21	69.57	70.58	70.6	73.29
2011	106	73.1	71	73.2	74.1	72.4
2012	110.3	68.4	68	70.8	66.9	73.8
2013	115.1	65.31	68.24	71.56	70.14	73.3
2014	120.2	70.2	63.7	72.2	69.4	72.7

參考資料

1. 香港政府統計處

   http://www.statistics.gov.hk/

2. 消費物價指數(CPI)

   http://www.censtatd.gov.hk/hkstat/sub/sp270_tc.jsp?productCode=B1060002

3. 訪港旅客數字

   http://partnernet.hktb.com/filemanager/intranet/ViS_Stat/ViS_Stat_C/ViS_C_2014/Tourism_Stat_12_2014_0.pdf

4. 香港快樂指數調查

   http://commons.ln.edu.hk/hkhi/

5. 綜合社會保障援助(綜援)計劃

   http://www.swd.gov.hk/tc/index/site_pubsvc/page_socsecu/sub_comprehens/

6. 中產階級

   https://zh.wikipedia.org/wiki/%E4%B8%AD%E7%94%A2%E9%9A%8E%E7%B4%9A

優異作品: My Brother vs My Mother

School Name: Sha Tin Government Secondary School

Name of Student: Tong Yung Ching

Level: Secondary 5

Supervising Teacher: Chan Sai Hung

Introduction

Every day we produce statistics and correlation, even when you are taking a shower.

Life is full of figures and numbers. If we carefully select them and think about them for a moment, they will tell us some truths and possible reasons.

Struggling between a lovely brother and a caring mother, Darren has to judge whom he supports. At midnight, only regression can inspire him.

Darren frowned at a pile of envelopes on the desk, then he glanced at his mother's long face and his brother's red one. The unpleasant confrontation between mother and brother were staged endlessly every day. Yet he did not want to show his stand. He chose to be the neutralizing state during the Cold War period.

The door was slammed shut with a raucous clang. Mother hissed. Still, Darren could not put up with this anymore. Sneaking back to his own room, he dialed a number.

D: Hello? Can I talk to Alicia?

A: Hello, Darren. It's midnight. What's the matter?

D: I'm sorry about this, but I can't really sleep.

A: You sound terrible. Are you sick?

D: I wish I were. Instead mother and brother have been bickering for a couple of months. I'm fed up with that.

A: I guess your brother is six years old now. Is he facing some

sort of problems?

D: You know there is a persistent inflation in Hong Kong. We're a small family; however, we're paying more and more water bill. Since brother bathed on his own, he's been staying in the washroom longer and longer. The water keeps running as he takes a shower.

A: So your mother blamed him for wasting water and increasing charges.

D: I have nothing to say. I don't know if it's true that brother contributes to our cost of living anyway. You must be thinking my family is ridiculous. Now brother sometimes baths three times a day. Sometimes he baths for more than fifteen minutes.

A: Three times and fifteen minutes, you sure?

D: Yes. I did count it. He insisted that he needed everything to be clean, neat and tidy. If I talk to him, I will only make him more repulsive. On the other hand, mother is upset because I didn't speak for her as being the elder son at home.

A: Your situation is quite embarrassing. Well, the only solution is to find out who is right so that you can talk to them both with strong reasons.

D: How could I do that?

A: You can make use of simple linear regression. In fact, it's the maths homework I'm working on.

D: Recession?

A: Not recession, it's regression. A regression studies if any association exists between two factors, which means "are the two related?" or "how are they related?". Its analysis is a kind of prediction derived from the historical relationship between an independent and a dependent variable.

D: If I process a linear regression, I will know if it is my brother's fault and I can help cope with the argument. Alicia, can you teach me now please?

A: Um…alright. You're so enthusiastic suddenly. Start by gathering the information we need, the water bill in the past twelve months and the mean time your brother spent in bathing each month.

D: No problem. I've been very sensitive to the figures.

A: As we're looking at whether increasing bathing time leads to an increasing water bill, the water bill should be taken as the dependent variable y.

Table 1: *x* and *y* variables

month	water bill ($) → variable *y*	Mean time spent for bathing/taking shower in minutes→variable *x*
10	210	2
11	222	2
12	222	3
1	246	4
2	227	6
3	276	8
4	279	10
5	280	13
6	300	15
7	333	16
8	319	19
9	363	20
Total	3277	118
Mean	273.10	9.83

A: 20 minutes! How could your brother bath that long?

D: Who knows? But I'm still very confused about what the regression is.

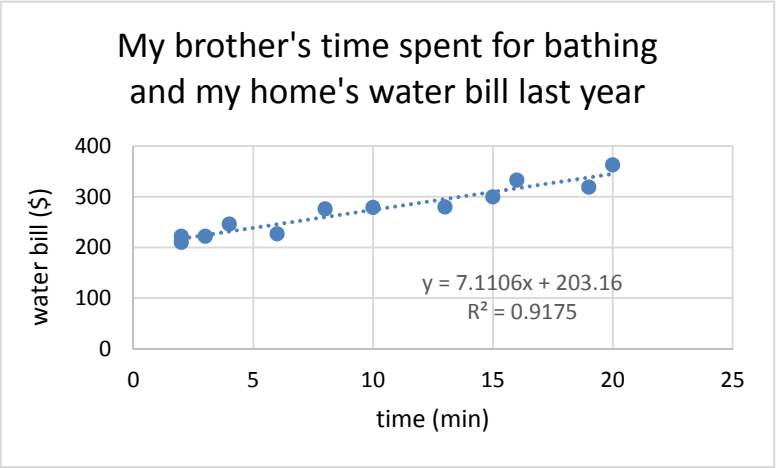
A: Don't worry. We can do it step by step. To study how the two are related, first you need to construct a scatter plot with bivariate observations (x, y).

D: Let me do it right now with my Excel.

A: In simplicity, the regression formula involves generating a best–fitting straight line in the form of $y = ax + b$ through all points. However, we can get a formula even if there is no relationship among variables. So it's important to examine the trend of the points before carrying regression analysis or it will lead to bias or other misleading outcomes.

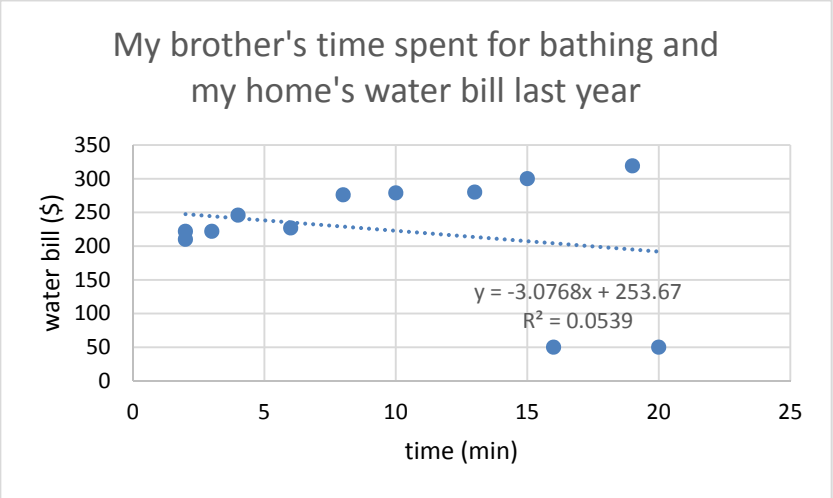
D: Good news. The points appear to cluster linearly.

Graph 1: scatter plot and the regression line



A: Moving on, we need to exclude any outliers that diverge significantly beyond the original observation range. Note that an outlier can greatly affect the slope and *y*–intercept, and affects the accuracy of the regression line then.

Graph 2: effect of outliers



D: That's true! Excel gives the formula $y = 7.1106x + 203.16$. Where does it come from?

A: As you see some points will lie above or below the line. Therefore, we need to minimize the average vertical distance between each of the predicted value of y ($\hat{y}$ on the line) and the real value of y (data points) as it represents the errors of prediction. We call the vertical distances ($y - \hat{y}$ or $\hat{y} - y$) residuals. Both the sum and the mean of the residuals equal to zero.

D: Then negative residuals may be resulted.

A: Therefore, we make use of the Ordinary Least Squares

Principle to get the smallest sum of squares of all residuals $\sum_{i=1}^{12}(y_i - ax_i - b)^2$ to eliminate the cancelling effects of positive and negative distances.

D: The formula looks sophisticated to me……

A: Calm down! This symbol $\sum$ named Sigma simply sums up the things going after it. The 12 above is total number of data points and *i*=1 below indicates to repeat and to run this formula from 1 to 12. The whole expression helps add up all.

D: That's clever.

A: Basically, for all regression formula, we have

$$\begin{cases} a = \dfrac{\sum_{i=1}^{12}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{12}(x_i - \bar{x})^2} \\ b = \bar{y} - a\bar{x} \end{cases}$$, where b is the constant and a is the coefficient of x

D: It sounds like some terrible mathematics!

A: Statistics is not as horrible as you think. Just like linear regression which is a very useful tool at workplace to draw up commercial decisions. For example, a company might want to decide how many advertisements should be placed on a

street. Don't give up!

D: You're right. I'm carrying out an important mission. Come on, please continue.

A: Remember the trend line must pass through $(\bar{x}, \bar{y})$ = (9.83, 273.10), the averages of variables.

$\because y = ax + b$

$= a\,x + (\bar{y} - \bar{x}\,a)$

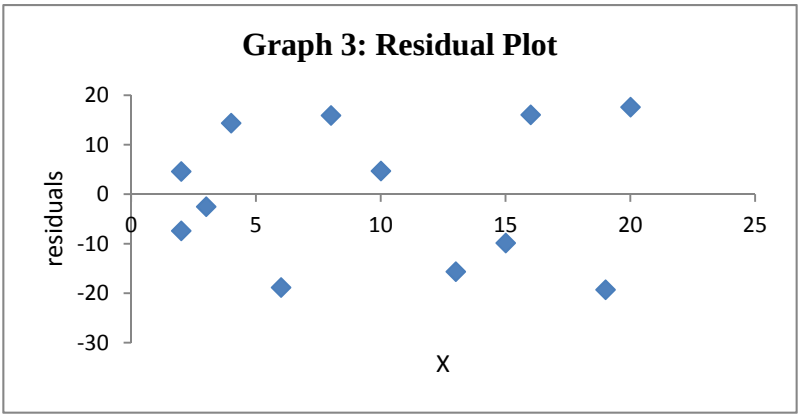
$= a\,(x - \bar{x}) + \bar{y}$

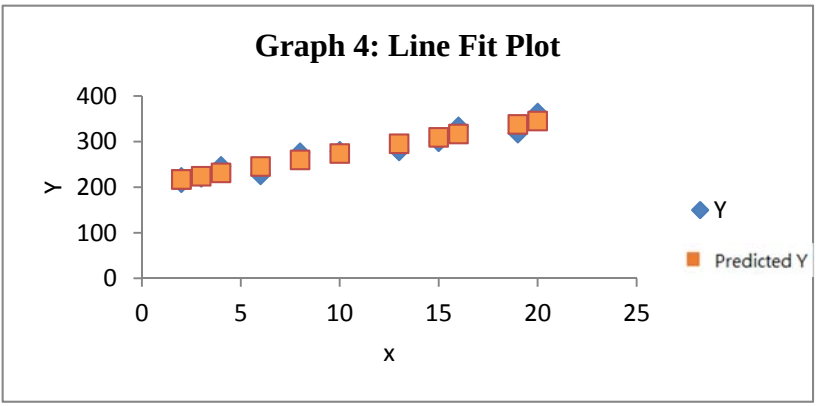
$= a\,(\bar{x} - \bar{x}) + \bar{y}$ —— Substitute x into $\bar{x}$

$= a \times 0 + \bar{y}$

$= \bar{y}$

Nowadays, researchers seldom calculate by hand. Excel can give us a detailed output. The random pattern along the *x*–axis in the residual plot also indicates that this linear regression model is appropriate for the data.





D: By observation, actual data and predicted data are quite close as well. I see. According to the result, I can deduce that…

A: Hold on. We can't rush to draw a conclusion when we haven't considered how reliable the data given is.

D: How can we test for the credibility of the statistics?

A: Well… now, we can focus on some essential elements of interpreting regression coefficients.

Output Summary (Graph 6)

Regression Statistics	
Multiple R	0.957837
R Square	0.917451
Adjusted R Square	0.909196
Standard Error	14.83352
Observations	12

	Coefficients	Standard Error	t Stat	P-value	Lower95%	Upper95%
Intercept	203.1623	7.894622	25.73427	1.8E-10	185.572	220.7526
Variable x	7.110613	0.674483	10.54232	9.78E-07	5.607771	8.613455

A: You will notice that Coefficients÷Standard Error = *t* stat. The last column gives an interval of the slope.ie. It is 95% confidence that *b* lies between 5.607771 and 8.613455. Multiple R is the absolute value of Pearson's coefficient of correlation between the predicted value and the real value of correlation between the predicted value and the real value of *y*. For your reference:

$$b = r \times \frac{\sqrt{\sum_{i=1}^{12}(y-\bar{y})^2}}{\sqrt{\sum_{i=1}^{12}(x_i-\bar{x})^2}} \quad \text{and} \quad r = \frac{\sum_{i=1}^{12}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{12}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{12}(y_i-\bar{y})^2}},$$

where $-1 \le r \le 1$

D: What is the correlation?

A: You can treat correlation as an index of strength of a relationship between two variables. Correlation does not limit to cause–and–effect only. r^2 is the coefficient of determination. It expresses the proportion of the variation in *y* which is explained by variation in *x*:

$$r^2 = \frac{\left[\sum_{i=1}^{12}(x_i-\bar{x})(y_i-\bar{y})\right]^2}{\sqrt{\sum_{i=1}^{12}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{12}(y_i-\bar{y})^2}}, \text{ where } 0 \le r^2 \le 1$$

D: In the third term, how is the *r* square adjusted?

A: Adjusted r^2 is obtained by the computer after adding some extra predicted data points to fit the model, i.e. to penalize adding more data to fit the model. It is always smaller than the original r^2. Comparing the two, you will know whether the regression model is genuinely better or if it has more predicted values only.

$r^2{}_{adj} = 1 - \left[(1 - r^2) \frac{n-1}{n-2} \right]$, where n = number of observations =12

It reflects both the number of explanatory variables in the model and the sample size. You get it? Darren? Hey……are you still here? Don't fall asleep!

D: Oh, yes. At least I get a framework of the whole picture. In general, the higher the value of r^2 is, the higher the accuracy of the regression formula can be obtained.

A: Right. There is no strict rule regarding the minimum value of r^2, you can consider 0.5 as the benchmark. In our formula, the coefficient of determination equals 0.917451 which implies about 91% of the variation in water bill can be explained by its relationship to the bathing habit of your brother.

D: Isn't it quite correlated? I'm so disappointed. Mother's word is not nonsense.

A: Still, the goodness of how well the calculated linear equation suits our data cannot solely be justified by r^2. Don't be sad. Let me introduce a simple method, Hypothesis Testing, to you. If our formula passes the test, it is recognized. Before that, we need to understand what a statistical hypothesis is.

D: Hypothesis? I heard of it during one of my science lessons. Scientists suggested a hypothesis before designing experiments to try to prove the propositions they asserted. If they fail to do so, the hypothesis will be rejected and they have to keep bringing up other possible theories for testing.

A: The idea here is similar. Since it's too hard and costly to examine whether all the data is real, we often test on a random sample instead.

D: I suppose there are only two cases; the statistics are convincing or else.

A: True. It's important that we state two clear assumptions with no equivocal area. One is the null hypothesis (denoted by H_0) means that the sample observations result purely by chance, while the alternative hypothesis (denoted by H_a) suggests the

opposite.

H_0 **: The slope of the regression line is equal to zero.**

H_a **: The slope of the regression line is not equal to zero.**

Applying it to our situation, if H_0 is true, the water bill is the same no matter how long your brother baths. Then your brother does not have any responsibility in raising your home's charges because no relation between them exists.

If H_a is true, the water bill is probably influenced by sort of a hidden factor behind (say corresponding to the time.)

D: Now the two hypotheses are mutually exclusive. If one is true, the other must be false.

A: To evaluate the null hypothesis, one of the approaches is the P–value method. P–value is to let us know how our data fit with the null.

D: That is easy. In the following step, we can simply reject the null hypothesis when the P–value is unusually low.

A: You're not entirely wrong. Any significance level can be used from 0 to 1. Usually, we choose 0.01(99%), 0.05(95%) or 0.10(90%). When the P–value is smaller than the significance level, the null hypothesis is rejected.

D: Pick 0.01 as our significance level. Back to our regression output, the P–value of the intercept is 0.00000000018 and that

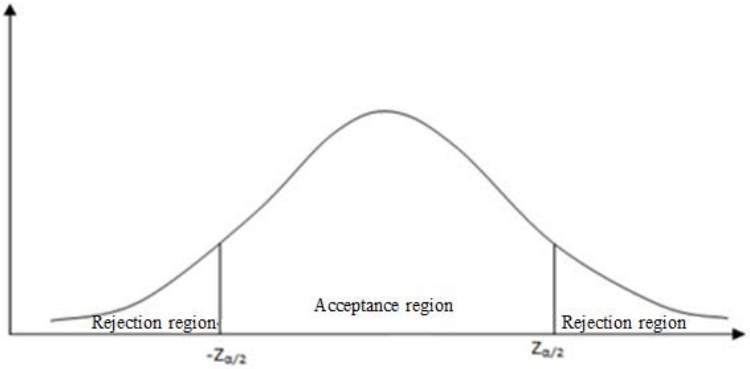
of the distance is 0.000000978. Both are much smaller than the significance level. We cannot accept the null hypothesis.

A: Indeed. The decision rule is alike in another approach. Imagine we have included the data of IQ scores of the entire population in the world within a generalized distribution graph; it will appear as a mountain–like diagram. And the mean value will equal the IQ score with the highest distribution. Do you know why?

D: Is it because most people score in the median range?

A: In fact, two-thirds of the population scores IQ between 85 and 115 while only about 5% score a very high or a very low IQ. There is then a region of acceptance represented by the area under the normal distribution curve. Those who score exceptionally high or low IQ will fall into the critical region (region of rejection).

Graph 5: Acceptance and rejection regions for two–tailed testing



D: So if the test statistic falls within the region of acceptance, the null hypothesis is not rejected; otherwise, it is rejected.

A: Yes. Here, we use a two-tailed testing. Both extremely high and low values of the test statistics indicate unusual situations.

D: To sum up, the more time my brother spent in taking a shower, the more fee of the water bill we have to pay. Using our formula, he will only take a longer shower later, thus results in a larger payment.

A: Your conclusion is sound only under certain conditions. Be aware that your brother might attack the weaknesses of your statement.

D: I know he always have a lot of excuses. I'd better think of some counter arguments beforehand.

A: Based on the simple linear regression result, we only consider the period around this year, where most of our data points lie. Extrapolation, meaning to predict the values of water bill or your brother's consumption time in the far future, lacks a ground due to many other fluctuations.

D: You're right. Our information is very limited. We've just carried out a simple linear regression on two variables while neglecting other possible factors. In reality, the Water Supplies Department calculates on a tariff structure consisting of four tiers with progressively increasing prices. Brother would like to point out he is not the only one using water. I also drink and bath. Mother also consumes water for cleansing, cooking, drinking, planting and so on. How can I persuade him to cut down on using water when bathing?

A: Relation does not necessarily imply causation. Coincidence exists all the time in daily life. There are often multiple factors ending up with a certain phenomenon.

Then here, for instance, one possible form of the equations may be $y = a_1x_1 + a_2x_2 + ...+ b$ where x_1, $x_{2, \ldots}$ etc. represent different factors.

Nevertheless, you can emphasise the significance of one of the causes. Tell your brother that no other family members takes a bath as frequent and as long as him. In this case, your brother is likely the dominating factor among all.

D: I've learnt a lot today. I've never thought about statistics can be so down-to-earth. Using simple linear regression as a way to look at daily issues is so fascinating and meaningful. From your voice, I can hear that you're exhausted. Thank you for listening to me, Alicia. We should really sleep now. It's 2 a.m. already.

A: It's Okay. I do hope your brother will know how to protect the environment and conserve natural resources. It's school day tomorrow. See you!

D: See you!

(2498 words)

References:

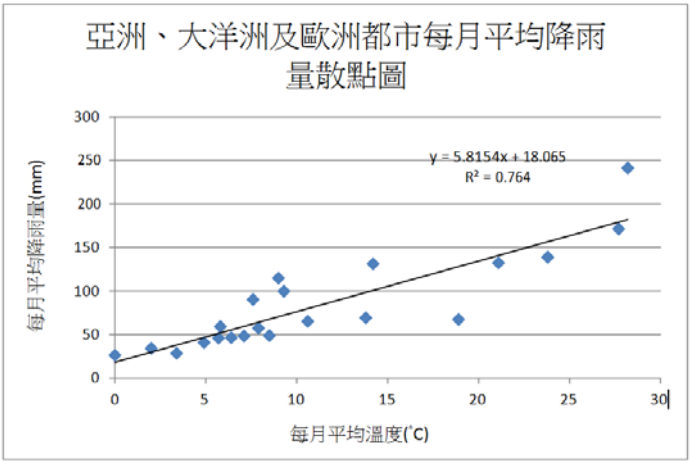
1. 《世界第一簡單統計學》高橋信著
2. Jim, F. (2014, April 17). How to Correctly Interpret P Values. Retrieved from http://blog.minitab.com/blog/adventures-in-statistics/how-to-correctly-interpret-p-values
3. http://stattrek.com/
4. 2012/13 中學生統計創意寫作比賽作品集

優異作品：線性迴歸

學校名稱：沙田崇真中學

學生姓名：周炯然

指導教師：李建中



引言

你知道降雨量與溫度的關係嗎?就讓珉荻和紫凝為你解開謎團吧! 他們會搜集世界上不同地區的降雨量和溫度的資料,然後運用線性迴歸的概念，告訴你他們的發現。

我的名字是珉荻，紫凝是我的同班同學，而我們是好朋友。在一個上課天的小息，窗外下起雨來，我突然聽到歌聲「天下起雨來，雨，不停落下來・・・」 我望向聲音傳來的方向，看見紫凝正望著那白濛濛的窗外，雨點一點一滴地打在窗戶上，而她正隨意地唱起歌來，我走向她然後問道:「紫凝，為甚麼自己一個 人在看著窗外唱歌呢?」她的歌聲被我的問題打斷了，然後答道:「沒甚麼啊，只是在想為甚麼天會下著滂沱大雨，有時卻是和風細雨，雨點的大小都各有不同，而我正在嘗試從觀察雨中尋找著答案呢。」我聽完後笑了一笑，然後說:「你認為窗外的雨點會與你交談嗎?其實降雨量是有關於天氣的，尤其是該地區的氣溫。」我看見紫凝的眼神變得充滿疑慮，我又說:「如果想了解多點關

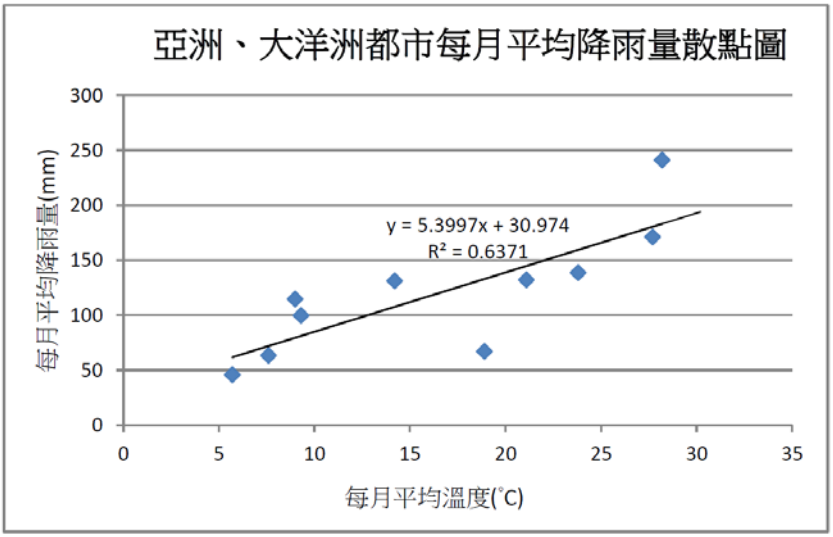
於降雨與溫度的關係的話，其實問我沒有甚麼用，哈哈!倒不如我們放學後一起找鄭老師解答你的疑問吧!」紫凝笑語:「好!鄭老師既是地理老師又是數學老師，必定能解答我們的問題。」

就在當天的放學後，我與紫凝一起到教員室找鄭老師。看見鄭老師後，紫凝疑惑地問:「鄭老師!鄭老師!究竟為甚麼雨有時下得很大，有時卻只有很小呢? 珉荻說降雨量是與氣溫有關，兩者真的有關係嗎?」鄭老師回答道:「珉荻你真聰明，兩者確實是有關聯的。」我笑而不語，紫凝便再問:「兩者實際上有甚麼關係呢?」鄭老師說:「你們看來真的很有興趣，與其我就這樣告訴你們答案，不如你們回家作資料搜集，然後明天把資料帶回來，讓我透過資料分析，再詳細跟你們說明一下降雨量與氣溫之間的關係吧」我和紫凝同時點頭並說:「嗯!」鄭老師說:「這樣吧，世界各地的氣溫及雨勢都大有不同，我們集中研究一下亞洲、大洋洲及歐洲地區吧，紫凝作亞洲及大洋洲的資料搜集，而珉荻則作歐洲的資料搜集。你們找那些地區的都市每月平均降雨量及每月平均氣溫作標準，好讓我們一起分析。」我和紫凝異口同聲地說:「知道，鄭老師再見!」

另一天放學後，我和紫凝興高采烈地拿著搜集的資料去找鄭老師。把資料交給鄭老師後，鄭老師叫我們等一等。稍後，鄭老師拿著數張印著圖表及數據的紙張到教員室外的桌子上和我們一起坐下，並說:「你們都做得很好，昨天的熱誠果然沒有倒退，依然對於這問題有好奇心，懷著追問到底的精神。在說明降雨量與氣溫之間的關係前，需要教導你們一個統計概念 —— 線性迴歸。」「線性迴歸?」我們道。鄭老師便加以解釋:「對，線性迴歸是一個統計學的概念，是透過自變數(x)和因變數(y)之間的相關關係去預測自變數對於因變數的影響，而兩者之間的關係叫作線性關係。自變數及因變數可以組成一組數據 (x, y)，而每組數據可以建立一個散點圖，並畫出一條迴歸直線，這條迴歸直線是一條自變數及因變數之間的趨勢線。我分別為你們的

數據作了一個圖表，以紫凝的亞洲、大洋洲都市每月平均降雨量散點圖先作說明吧。」鄭老師先拿首由紫凝搜集的數據所製成的圖表。

紫凝的亞洲、大洋洲都市每月平均降雨量散點圖及數據圖表:



亞洲、大洋洲都市	每月平均溫度(°C)	每月平均降雨量(mm)
首爾	5.7	45.8
伊斯坦堡	7.6	63.2
東京	9	114.5
大阪	9.3	99.5
福州	14.2	131
香港	18.9	66.9
雪梨	21.1	132.1
布里斯班	23.8	138.5
新加坡	27.7	171
吉隆坡	28.2	240.9

資料來源:台灣交通部中央氣象局

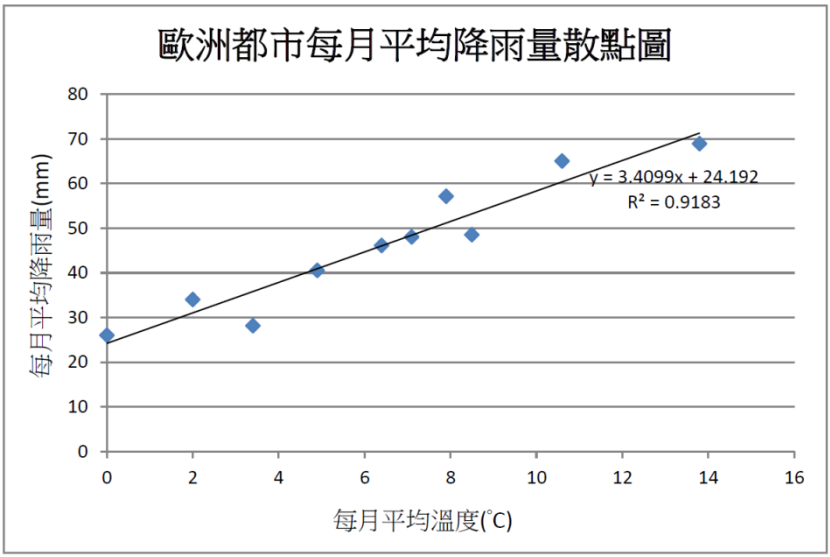
鄭老師說:「根據紫凝的數據，畫出亞洲、大洋洲都市每月平均降雨量散點圖，以每月平均溫度(°C)作自變數(x)而每月平均降雨量(mm)則作因變數(y)，把一組一組的數字以散點型式繪圖，然後畫出在圖中最適切的迴歸線叫作迴歸直線，繼而得出方程式。$y = 5.3997x + 30.974$ 是這圖的迴歸直線的方程式。而 X 與 Y 存在一定的線性相關關係，如果

R^2 的數值越接近 1，X 與 Y 有越高度的線性關係。圖中 R^2 有 0.6371 的數值，證明亞洲及大洋洲都市每月平均溫度(°*C*)與每月平均降雨量(*mm*)有著中度的線性關係。然而，首爾、香港、東京、新加坡等都市有著不同的氣候，並不能只因為氣溫的不同而衡量降雨量，因為當中有因着其他的變數所限制，例如: 相對濕度、大氣壓力、風勢所影響。話雖如此，但我們能夠透過線性迴歸，簡單的憑著每月平均溫度(°*C*)去預測每月平均降雨量(*mm*)，譬如以雪梨為例，雪梨的每月平均溫度(°*C*)是 21.1，而根據迴歸直線預測每月平均降雨 量(*mm*)是大約 140*mm*，實際數據 132.1*mm*。雖然當中有少許誤差，但仍能透過簡單的統計，找出當中的預測降雨量。」

紫凝說:「原來線性迴歸是一種統計方法去尋找兩者之間的關係並作出預測，我明白了。根據數據及迴歸線所顯示，溫度越高，降雨量亦會因而增加，氣溫與降雨量是有著正相關關係。珉荻，歐洲的都市亦會因氣溫上升而導致降雨量增加嗎?」

此時，老師拿出圖表及數據，說:「珉荻，你嘗試解釋一下當中的關係吧。」 我(珉荻)的歐洲都市每月平均降雨量散

點圖及數據圖表:



歐洲都市	每月平均溫度(°C)	每月平均降雨量(mm)
斯德哥爾摩	0	26
哥本哈根	2	34
布拉格	3.4	28.1
柏林	4.9	40.5
維也納	6.4	46.1
倫敦	7.1	48
巴黎	7.9	57.1
威尼斯	8.5	48.5
羅馬	10.6	65
里斯本	13.8	68.9

資料來源:台灣交通部中央氣象局

我硬著頭皮嘗試解釋，道:「根據這個歐洲都市每月平均降雨量散點圖，當中的迴歸直線有著 $y = 3.4099x + 24.192$ 的方程及 $R^2 = 0.9183$。由於 R^2 的數值是 0.9183，說明歐洲都市的每月平均溫度(*°C*)與每月平均降雨量(*mm*)有著高度的線性關係，並且歐洲都市的迴歸線比亞洲、大洋洲都市的迴歸線更為準確，是圖表中最適切(*fit*)的迴歸線。以我最喜

愛的城市倫敦為例，倫敦的每月平均溫度 是 7.1°*C*，根據圖表及迴歸線的預測，倫敦的每月平均降雨量會在大約 47*mm*-48*mm*，而實際每月平均降雨量正正是 48*mm*，可見線性迴歸的統計圖可以透過自變數(每月平均溫度(°*C*))預測到因變數(每月平均降雨量(*mm*))的大概數值。整體來看，歐洲都市的降雨量亦會隨著氣溫的上升而增加，可見兩者之間是有正相關關係的。」

鄭老師點一點頭，說:「珉荻，我相信你都已經掌握到甚麼是線性迴歸，解釋得很詳盡。紫凝，你有沒有透過線性迴歸了解到氣溫與降雨量的關係?」紫凝說:「當然有，知道原來都市的氣溫與降雨量是成正比例的!鄭老師不但解答了我對於氣溫與降雨量的難題，還好好的教我們甚麼是線性迴歸，日後想研究甚麼有關聯的變數及關係，可以自己找資料並使用線性迴歸的概念來尋找它們之間的關係。」鄭老師又說:「沒錯，你們有甚麼困難也可以再來找我喔! 」我和紫凝微笑而且感激地說:「多謝鄭老師!我們會好好的記住老師的教導!鄭老師再見!」我和紫凝道別後便揮手離去，心中還存著一顆感謝老師的心。

(~2293 字)

優異作品: Hot Pot Dots

School Name: HEEP YUNN SCHOOL

Name of Students: Cindy Kat, Karen Lee, Hebi Wong

Supervising Teacher: Mr Y.C. WOO

Introduction:

During winter, it definitely makes one's day to have a hot pot with all sorts of delicacies. However, with the temperature on the rise, it is a reasonable prediction that the sales of hot pots will drop.

Miss Wong: Let's have some good food! It's on me by the way.

Cindy and Karen: Thank you Miss Wong for your professional advice along the way.

Miss Wong: I'm really proud that you won a business proposal writing competition. The presentation and the floor time just now must have been nerve-racking but you totally nailed it. Well done guys! You are almost like professionals now. OK… What do you guys want to order?

Karen: Am I too greedy to want a hot pot?

Cindy: I don't feel like having one because the temperature is much higher than yesterday. Rice with two choices of sides will do for me.

Karen: Right! Not many customers are ordering hot pots here though it's the best choice in this restaurant in my opinion.

Miss Wong: Both of you are quite observant. Now, this is a real-life example of what we learn in BAFS. One of the operational functions is about forecasting demand. A smart manager would consider different factors, like the temperature, and then adjust the amount of certain dishes accordingly. Like me, I would want chilled noodles with such weather.

Cindy: It's only natural that there is a positive relationship between the temperature and the number of orders of hot pots.

Karen: Don't be sure! It's unprofessional and unscientific to jump to the conclusion.

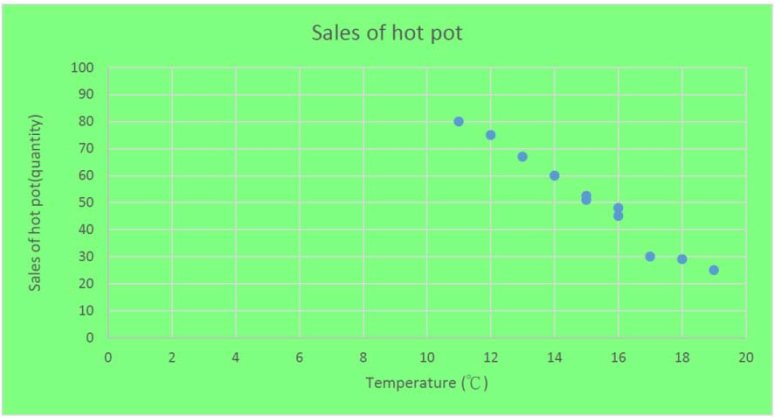
Miss Wong: We should also collect data about rice with two choices of sides and chilled noodles. But let's order our lunch first. Waitress!

After ordering lunch, they also managed to get hold of the sales data.

Miss Wong: Actually we can make use of scatter diagrams to figure the relationship between the temperature and the sales of various type of food. I think that's well within your ability!

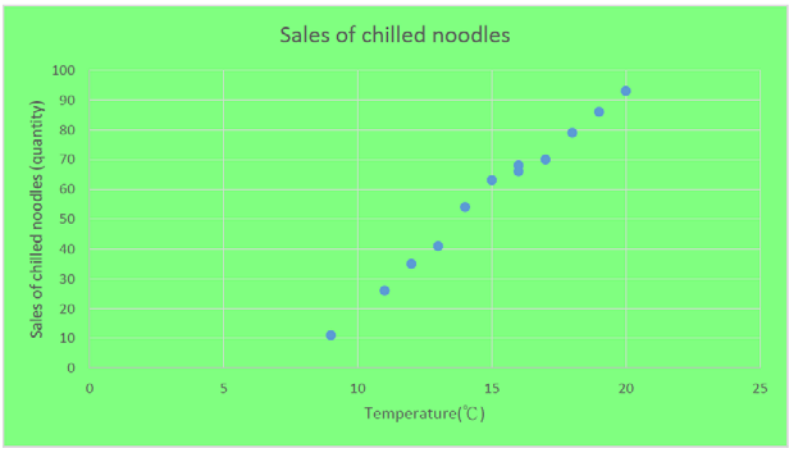
Karen and Cindy: OK!

Date	(℃) i.e. Xi		Sales	
		Hot pot i.e. Yi	Chilled noodles	Rice with two choices of sides
1	20	19	93	83
2	19	25	86	78
3	17	30	70	101
4	18	29	79	88
5	16	45	66	96
6	15	51	63	94
7	16	48	68	86
8	14	60	54	75
9	12	75	35	95
10	13	67	41	110
11	11	80	26	108
12	9	101	11	94
Total	180	630	692	1108
Average	15	52.5	57.66667	92.33333



Karen: The higher the temperature, the lower the sales of hot pots, right?

Cindy: There seems to be no relationship between the temperature and the sales of rice with two choices of sides…



Karen: Oh good! Indeed, the higher the temperature is, the more the quantity of chilled noodles sold.

Miss Wong: Actually, we can use more professional and scientific methods to prove the correlation.

Date	Temperature(℃)	The sales of hot pot
	Xi	Yi
1	20	19
2	19	25
3	17	30
4	18	29
5	16	45
6	15	51
7	16	48
8	14	60
9	12	75
10	13	67
11	11	80
12	9	101
Total	180	630
Average	15	52.5

Relationship between temperature and the sales of hot pots

Date	Xi- $\bar{x}$		Yi- $\bar{y}$	$(Xi - \bar{x})^2$	$(Yi - \bar{y})^2$	(Xi- $\bar{x}$)(Yi- $\bar{y}$)
1		5	-33.5	25	1125.3	-167.25
2		4	-27.5	16	756.25	-110
3		2	-22.5	4	506.25	-45
4		3	-23.5	9	552.5	-70.5
5		1	-7.5	1	56.25	-7.5
6		0	-1.5	0	2.25	0
7		1	-4.5	1	20.25	-4.5
8		-1	7.5	1	56.25	-7.5
9		-3	22.5	9	506.25	-67.5
10		-2	14.5	4	210.25	-29
11		-4	27.5	16	756.25	-110
12		-6	48.5	36	2352.3	-291
Total		0	0	122	6900.3	-909.75
Average		0	0	10.16667	575.0208	-75.8125

Miss Wong: we can input the figures into the formula to arrive at the correlation coefficient.

$$r = \frac{\sum_{i=1}^{12}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{12}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{12}(y_i - \bar{y})^2}}$$

$$r = \frac{-75.8125}{\sqrt{1760.9}\sqrt{10.16667}}$$

$$r = -0.991535259$$

Miss Wong: Using exactly the same methodology, we can arrive at the other two data.

	Sales		
	Hot Pot	Chilled noodles	Rice with two choices of sides
Correlation coefficient relative to the temperature	-0.9915	0.9944	-0.5026

Miss Wong: In fact, the value of correlation coefficient lies between -1 to 1. The stronger the correlation between the two variables, meaning the scattered dots form a straight line, the value will either be closer to -1 or 1. Conversely, the closer the value of

the correlation coefficient is to 0, the weaker the correlation between the two variables. This is it!

Negative correlation		No correlation		Positive correlation
-1	~	0	~	1

Karen: Oh I understand!

Karen and Cindy: Thank you Miss Wong.

優異作品: 成績「分分」跌？「分分」賞！

學校名稱：香港紅十字會雅麗珊郡主學校

學生姓名：郭曉明

指導教師：馬文俊

引言

特殊學校的學生相對體弱多病，所以請假較頻密，心理也較敏感，家長會擔心子女因心理健康而影響學習。本文將探討在「手術」、「覆診」、「病假」、「事假」、「被讚」和「被罵」這六個原因中，哪個對成績的影響最大。

今天是一年一度由卡娃伊學校主辦的攤位遊戲日，有不少家長前來湊熱鬧、玩遊戲。其中由數學主任郭老師主持的「終極猜成績」最受歡迎，吸引不少家長參加，遊戲玩法是從「手術」、「覆診」、「病假」、「事假」、「被罵」、「被讚」這六個原因中，猜猜哪個對學生的成績影響最大。家長們都議論紛紛，其中陳太和馬太正鬧得面紅耳赤。

陳太：「肯定是手術！因為手術後要休養一段時間，成績一定下跌！」

馬太：「我認為是病假。病痛多自然無心向學，不影響成績才怪！」

正當她們吵個不停時，郭老師突然出現。

郭老師：「你們各有道理，但是要知道哪個原因影響最大的話，一定要深入探究才行。各位家長如果有空的話，請到會議室，我會為大家詳細講解！」

家長們眼見仍有餘閒，於是紛紛到會議室。

在會議室，郭老師手持數據說道：「如果要探究哪個原因的影響最大，口講並不準確，要有實質數據支持才可以。這些數據是從學校資料庫和問卷中收集，而對象學生是從本校學生中隨機選出的。」

眾家長聽郭老師說得頭頭是道，便更加留心聆聽。

郭老師續說：「我們今次主要採用簡單線性迴歸的方法來進行分析，即是利用線性方程歸納及描述變量之間系統性的關係。剛才我所說的六個原因都是自變量，它們會影響因變量，而成績改變則是因變量，它會受到自變量的影響而變化。」

陳太說道：「雖然我不太明白你在說甚麼，但是我仍然覺得手術的影響較大。」

馬太在旁說道：「對啊！我仍然覺得病假的影響較大，你會不會覺得我們很麻煩呢？」

郭老師說道：「當然……不會！大家不明白的話，我非常樂意為大家解釋，就先利用散點圖分析它們影響吧！根據圖 1 和圖 2，當做手術的次數越多，成績會隨着退步，而病假越多，成績都會隨着下降。」

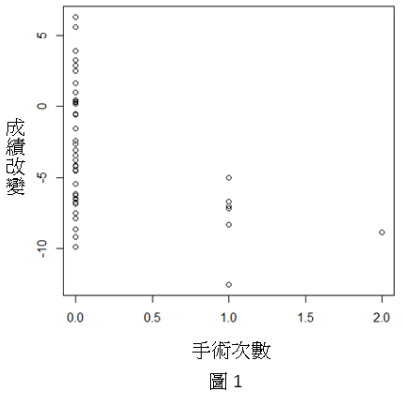


圖 1

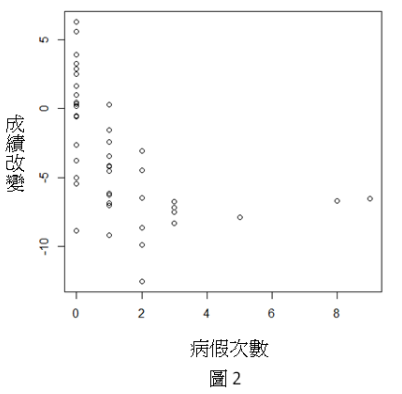


圖 2

平時沉默寡言的譚太忽然作聲：「那覆診和事假呢？」

郭老師再說：「根據圖 3，其實覆診的影響也差不多，即是覆診次數越多，成績亦會下跌，但是圖 4 顯示事假對成績改變的影響真的很少，總括而言，手術、病假和覆診都會影響成績，而事假的影響則不太明顯。」

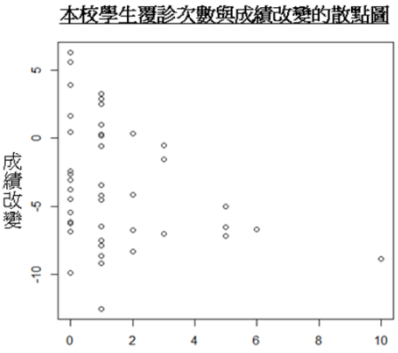


圖 3

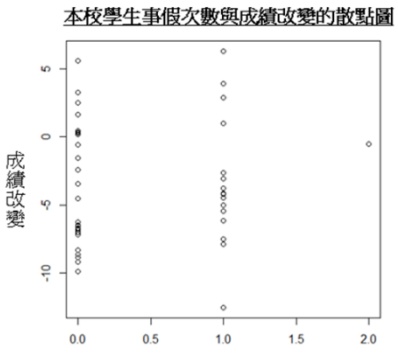


圖 4

在場眾太太齊聲問：「那手術、病假和覆診的影響究竟有多大呢？快點說吧！」

郭老師沒好氣地說：「其實它們的影響可以透過下列迴歸方程顯示出來：

手術的成績改變 = - 2.70 - 2.48 × *手術次數* ，

即每做手術一次，成績跌 2.48 分；

病假的成績改變 = - 1.98 - 1.12 × *病假次數* ，

即每請一次病假，成績跌 1.12 分；

覆診的成績改變 $= -2.49 - 0.678 \times$ *覆診次數* ，

即每覆一次診，成績跌 0.678 分；

事假的成績改變 $= -3.80 + 0.723 \times$ *事假次數* ，

即每請一次事假，成績升 0.723 分。」

這時家長們嘩然，均問道：「請一次事假竟然升 0.723 分！如果我為蝦女請百多天病假，豈不是全科一百分？還需要讀書嗎？郭老師，你真懂開玩笑的。」

郭老師：「其實下這個結論前，我們還要利用假設檢定來確認一下各個自變量的影響是否顯著，請大家看看表 1 中的資料，以事假為例，考慮：

零假設(H_0)：　　事假的影響 $= 0$

對立假設(H_a)：　事假的影響 $\neq 0$

利用 t 檢定，取顯著性水平(α)為 0.1，如果事假的 p 值大於 0.05，就要接受 H_0；相反事假的 p 值小於 0.05，就不可以接受 H_0。而事假的 p 值是 0.577，所以應該接受 H_0，即事假的影響是 0。」

	自變量的 p 值	R^2 值
手術與成績改變的模型	0.004	0.1835
病假與成績改變的模型	0.001	0.2425
覆診與成績改變的模型	0.045	0.0946
事假與成績改變的模型	0.577	0.00766

表 1

郭老師續說：「我們還可利用決定系數(R^2)的值來判斷模型是否合適。如果 R^2 值越接近 1，代表因變量的改變可以用這個模型的自變量來表達，即是模型越好；而事假的 R^2 值是 0.00766，與 1 的距離頗遠，所以成績的改變不能通過這個模型由事假次數表達。總括而言，請事假會令成績上升這件事，根本說不過去！」

馬太：「哦！那我明白了！根據表 1 中的資料，手術、病假和覆診的 p 值都比 0.05 小，由此可知，這三個自變量對成績的影響都顯著。但是它們的 R^2 值太小，所以成績的改變不完全能通過這三個模型來表達。郭老師，我是否很聰明呢？」

郭老師苦笑道：「其實.....總括而言，這幾個模型都不算太好，所以我們要找到更好的因變量才能更準確地描述成績

的改變，而我所說的就是被罵和被讚這兩個因變量。」

眾太太充滿疑惑地問：「成績改變與被罵和被讚又有何關係呢？」

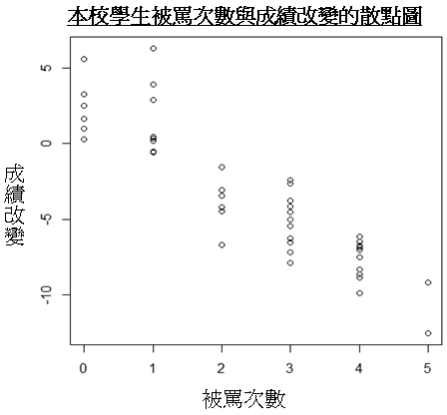


圖 5

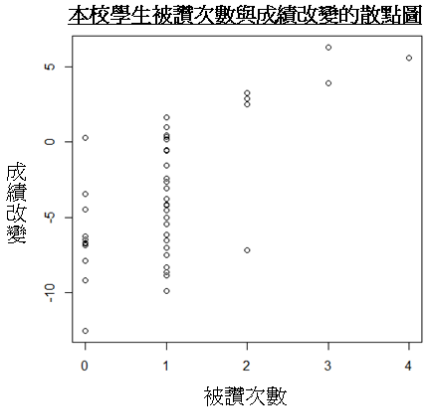


圖 6

郭老師不耐煩地說：「你們再看清楚點吧！從圖 5 和圖 6 來看，學生被罵的次數越多，成績就會越差；學生被讚的次數越多，成績就越好，利用以下的迴歸方程可以顯示出更清晰的關係：

被罵的成績改變 = 3.00 – 2.71 × *被罵次數*，

即被罵一次，成績跌 2.71 分；

被讚的成績改變 = –6.90 + 3.40 × *被讚次數*，

即被讚一次，成績升 3.40 分。

除了以上的迴歸方程之外，我們再看看表 2 中兩個模型的 p 值和 R^2 值吧！表中顯示這兩個模型的 R^2 值較剛才高，且它們的 p 值小於 0.05，即是說這兩個自變量對成績的影響較顯著。」

	自變量的 p 值	R^2 值
被罵與成績改變的模型	<0.0001	0.816
被讚與成績改變的模型	<0.0001	0.436

表 2

陳太問道：「既然已有兩個不錯的模型，可否不再理會其他自變量呢？」

郭老師補充道：「雖然是可以只利用被罵和被讚這兩個自變量來建立模型，但是完全忽視其他自變量的影響亦不一定是一個好方案。所以我們可以嘗試將所有自變量放進同一個模型中，綜合來看它們對成績改變的影響有多大。」

整個會議室忽然響起一句：「哦，我們『明白』了！」

郭老師無奈地說：「算了，我繼續說吧！其實當所有自變量放進同一個模型時，成績改變可以由以下迴歸方程來表達：

$$\textit{成績改變} = 0.46 + 0.0981 \times \textit{覆診次數} - 1.92 \times \textit{手術次數} - 0.378 \times \textit{病假次數}$$

$$-0.0587 \times \textit{事假次數} \ -1.96 \times \textit{被罵次數} \ +1.49 \times \textit{被讚次數}$$

這個模型的 R^2 值是 0.9303，所以成績改變可以很有效地由這個模型表達出來。但是根據表 3，覆診和事假的 p 值大於 0.05，即是說它們的影響並不顯著。如果剔除這兩個自變量，我們可以得到以下迴歸方程：

$$\textit{成績改變} = 0.558 - 1.55 \times \textit{手術次數} - 0.347 \times \textit{病假次數} \\ -1.99 \times \textit{被罵次數} + 1.48 \times \textit{被讚次數}$$

即是說：每做一次手術、請一次病假和被人罵一次，成績分別會跌 1.55 分、0.347 分和 1.99 分；而每被人讚一次，成績則會升 1.48 分。還有，這個模型的 R^2 值是 0.9297，亦非常接近 1，即是成績改變可以透過這個模型的自變量來表達，而表 4 顯示所有自變量的 p 值都小於 0.05，所以它們的影響都是顯著的。」

自變量	覆診次數	手術次數	病假次數	事假次數	被罵次數	被讚次數
p 值	0.578	0.0208	0.0035	0.876	<0.0001	<0.0001

表 3

自變量	手術次數	病假次數	被罵次數	被讚次數
p 值	0.00185	0.00238	<0.0001	<0.0001

表 4

經過郭老師詳細的解說，一眾家長終於有所領悟，她們都在思索解決方法。

馬太先發言：「做手術這樣重要、又關乎健康的事，怎可以胡亂改時間！我看都是請求醫生可否在長假期，例如暑假裏才做，否則真是沒辦法啊！」

陳太接著說：「對啊！我都只好叫女兒早點兒睡，多菜少肉，這才有強健的體魄，少請病假。還有，叫她勤力點讀書，交足功課，這被人罵的機會都會減少。」

譚太再說：「經常被人讚有難度，真的要又乖又深得老師同學喜愛才可，不過不要緊！我兒子在學校很乖很聽話，應該不常被罵。希望他的成績有所進步吧！」

郭老師見家長們挺滿意自己的分析，喜出望外，並答應來年繼續與她們一同研究數學，探索新事物。
(~2393 字)

參考資料

1. Mendenhall W. and Sincich T.(2011). *A Second Course In Statistics: Regression Analysis (7th Edition)*. Pearson.

2. 維基百科；迴歸分析. Retrieved from http://zh.wikipedia.org/wiki/%E8%BF%B4%E6%AD%B8%E5%88%86%E6%9E%90

3. 香港紅十字會雅麗珊郡主學校學生資料庫

優異作品: 作弊

學校名稱：筲箕灣官立中學

學生姓名：傅德熙

級別 ： 中三

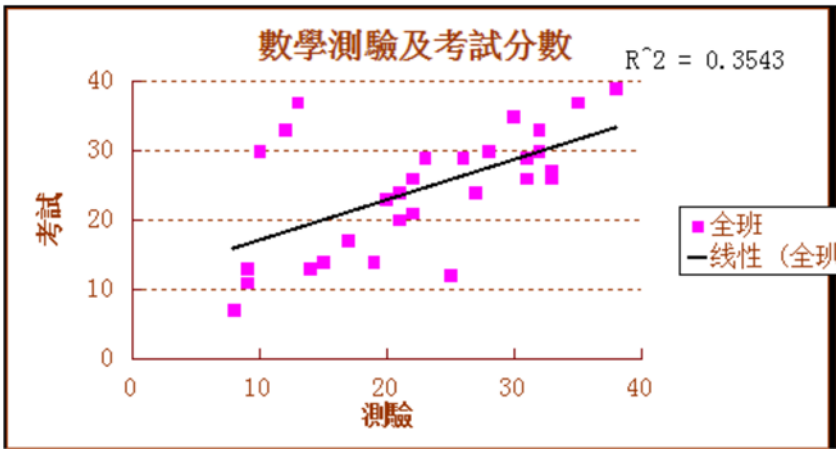
指導教師 ： 黃佩珊

引言

正當老師批改完學生的試卷，把試卷分數輸入電腦中，突然老師發現部分學生的成績突飛猛進。究竟是因為努力讀書還是其他原因呢? 如何利用線性迴歸找出答案?

2014 年 6 月，期末考試後的第一天。當學生們沉醉在考試結束的喜悅時，老師們卻要留在教員室，埋頭在試卷堆裡搏鬥。

「啊！韓英同學這次拿到 39 分呢（滿分 40）！他平時一直在做練習，這次考試應該難不倒他。」賈老師如此說。評改全班同學的試卷以後，賈老師一如往常把所有同學的分數輸入電腦，然後畫成散點圖，看看同學們與上次測驗相比有沒有進步。



圖一：3D 班數學測驗及考試分數分佈散點圖

$y = 0.584672x + 11.16979$

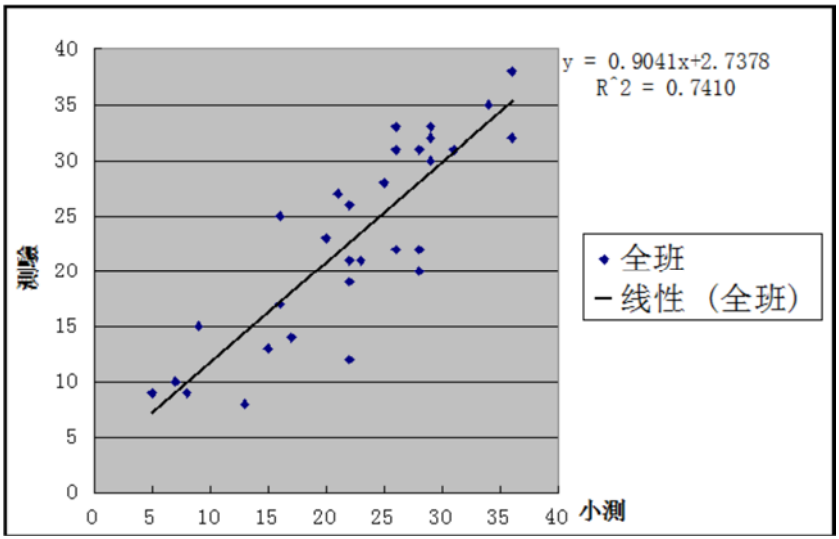
其實班中每個人的實力似乎都在學期初便顯示出來，除了數個因為補習和自己努力不懈而令成績一直提高，或是因為懶惰而令成績退步的考生外，大部分考生的測驗成績與其上一次的成績相差都在 5 分之內。而且這次考試的 R^2 是 0.3543（比較低，接近 0），這表示測驗與考試成績的線性關係已不太強，這也代表着考生的成績也越發波動。

賈老師看見這次的結果，他又皺一皺眉頭。「如我所料，這次同學們的考試成績進步了一點。儘管每次測驗和考試的分數之間沒有甚麼必然的關係，但是據我所知，班上成績最差的，也是最愛生事的幾位男生，數學成績也好不到哪裡去，成績中上的那些表現平穩，為甚麼這次有幾個點在左上角冒了出來？難道有人來了個突飛猛進？」

成績波動？突飛猛進？作為一個數學老師，他當然希望教導班上每一個人好好學習數學，令他們的成績穩步上升，斷不希望看到學生們的成績受太多外在因素影響而波動，一次考得好，一次考得差。有人數學成績突然飆升，賈老師理應為他感到高興，但是他此刻的臉上卻融合着疑惑和苦惱。「為甚麼會這樣呢？」

於是他把上次測驗與小測的分數圖表拿出來研究一下。

圖二：3D 班數學測驗及小測分數分佈散點圖



「為甚麼會這樣？明明沒有人在數學科的功課和小測上有不尋常的進步，為甚麼偏偏考試的時候才出現了那麼多的離群值？」

看看小測和測驗成績的圖表，R^2 為 0.7410（接近 1），這代表同學的表現相對地穩定，

他決定把試卷重改一次，發現 4 號、8 號和 12 號同學的分數都相當高，和上一次對比，進步了 20 多分（這裡就有點不

尋常），而且他們都有一個共同點：測驗和小測的成績都在合格線下，但考試的成績卻異常的高，有的還達到了全級首十名！

於是，賈老師把他們的試卷抽出來，分析他們的答題模式。他發現第 3 到第 7 題，還有第 15，21 和 33 題，（整份試卷都是多項選擇題，每題有 4 個選項）這些同學都給出了相同的答案，而且其中有一題他們都不約而同做錯了！

題目	3	4*	5	6	7	15	21	33
他們的答案	D	C	C	A	C	B	B	A
正確答案	D	B	C	A	C	B	B	A

賈老師如此想：8 題選擇題均得出相同答案的機會率是 $\left(\frac{1}{4}\right)^8 = \frac{1}{65536}$，六萬多分之一，機會率這麼小，他們三個竟然得出相同的答案！於是他決定隔天叫他們三個來見面。可是經過一輪面談，4 號、8 號和 12 號都否認作弊，並異口同聲說卷子是自己做的，答案相同純屬巧合。

賈老師回到教員室，苦苦思索解決方法。他一方面不想冤枉學生，另一方面又不想任由作弊這一惡行發生，他忽然想到：何不看看其他題目的答對率？

題目	3	4	5	6	7	15	21	33
答對率(%)	50	86.67	36.67	43.33	30	96.67	100	76.67

賈老師看到其中 4 道題目均有相當高的答對率(75%以上)，而剩下 4 道題目出現相同選項的機會率則為：$\left(\frac{1}{4}\right)^4 = \frac{1}{256}$。

「不！真正的答對率應該是 50%×13.33% (因為這題他們都答錯了，所以是(1–86.67%)×36.67%×43.33%×30%×96.67%×1×76.67%，那就是 0.00294，這麼低的機會率，說明了他們絕對有可能作弊！」

他又看了看他們的答題規律，發現 4 號同學第 8 到 14 題、8 號同學第 16 到 20 題、12 號同學第 35 到 40 題全部都答對了。

4 號同學答對的機會率：$\left(\frac{1}{4}\right)^7 = \frac{1}{16384}$

8 號：$\left(\frac{1}{4}\right)^5 = \frac{1}{1024}$

12 號：$\left(\frac{1}{4}\right)^6 = \frac{1}{4096}$

已知這些題目的答對率均在 40-60%之間，而且他們的已有知識，根本無法解答這幾道問題。

賈老師再仔細看看他們的答題紙，發現上面有着不少擦痕，跟 11 號韓英同學答題紙上的痕跡完全一樣！而且 4 號、8 號和 12 號都跟韓英的坐位非常接近，絕對可以趁老師不為意的時候偷瞄他的答案！

1	5	9	13	17	21	25	29
2	6	10	14	18	22	26	30
3	7	11	15	19	23	27	空
4	8	12	16	20	24	28	空

（考試座位表）

賈老師決定在卷子發下去的那一天再叫那三人來教員室，審問他們作弊的原因。

（~1748 字）

優異作品：舊夢不須記?——被遺忘的香港樂壇

學校名稱：香港培正中學

學生姓名： 蔡依彤，李曉盈

級別 ： 中三

指導教師：梁偉雄老師

引言

近年，香港樂壇明顯沒有七、八十年代那麼輝煌，很多人亦忘記了當時的光輝歲月。我們是次將會嘗試找出香港樂壇光輝不再的原因，加以分析，並希望藉此勾起人們心中那段漸漸被時間洪流沖走的記憶。

美鈴: 「結束感情沒有罪，變心不是你不對……」

媽媽: 美鈴你在唱甚麼?

美鈴: 古巨基的 《眼睛不能沒眼淚》啊！你不是沒聽過吧？

媽媽: 甚麼變心有罪啊？眼睛流眼淚啊？歌名又長，歌詞又詞不達意，這些歌這麼奇怪，怎會流行啊？以前的流行歌比現在的簡短易記多了，內容又夠深刻。唉，真可惜香港現在的樂壇已經今非昔比!

美鈴: 那些都你的錯覺而已，以前的歌名、歌曲長度和現在的都沒甚麼大分別啊！你看我的統計：

表一:

平均值	1984-1993	1994-2003	2004-2013
歌名字數	4.05	4.15	3.51
歌曲時間	4.34	4.07	4.13

表二:

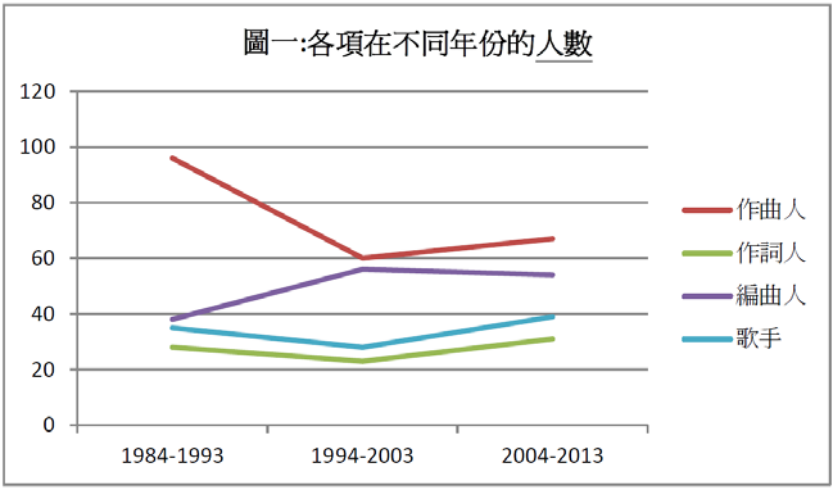
歌名字數	1984-1993	1994-2003	2004-2013
中位數	4	4	4
眾數	4	4	4

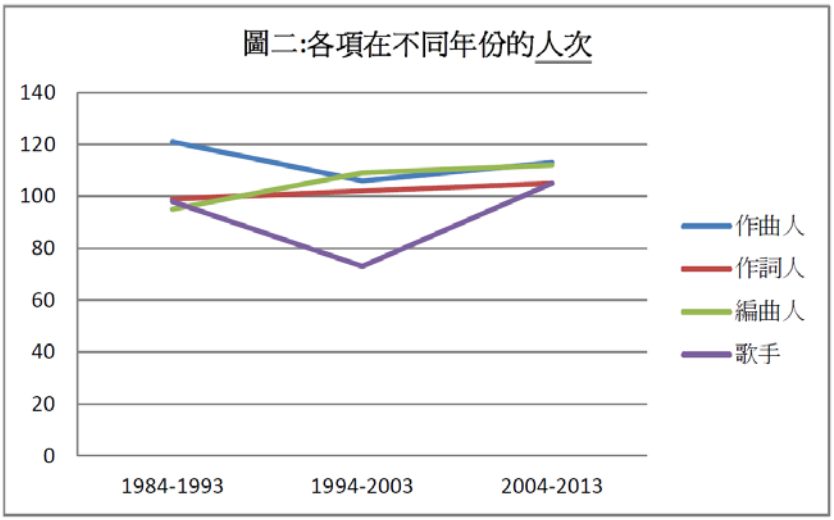
美鈴: 我假設歌曲時間和歌名字數是影響香港樂壇的原因，因此做了這些表。這些圖表的資料是來自 1984–2013 十大勁歌金曲獎得獎歌的時間和歌名字數。根據上表，我們可以看見歌名字數有下降的趨勢；時間方面沒有特別的趨勢。

媽媽: 做得不錯！這些資料可見原來一直以來，歌曲的歌名字數和時間長度都沒有太大的改變。唉，那爲何香港的樂壇會衰落到如此地步？我聽說黃霑臨終前曾做過這方面的研究，美玲不如你去找一找？

（幾天後）

美鈴：媽媽，我研讀過黃霑的博士學位論文後，我發現有幾點很值得我們留意的。例如: 他在論文中提及「科技機器令作曲人增多，卻沒有令水準提高，反而很多時候，令水準下降」[1]，讓我來分析一下他的結論是否正確。





美鈴: 這些資料是根據 1984–2013 年十大勁歌金曲獎所得出的。每十年為一組。「人數」就是統計每組有多少個不同的作

曲、作詞、編曲人和歌手;「人次」是統計他們曾出現的次數。可見以上兩個圖表各項目的趨勢都大致相同。歌手、編曲和作詞人人數和人次都是隨著時間增加而遞增;但作曲人人數和人次則是隨著時間增加而遞減。

媽媽: 咦! 那麼黃霑的說法「作曲人增多」豈不就錯了?

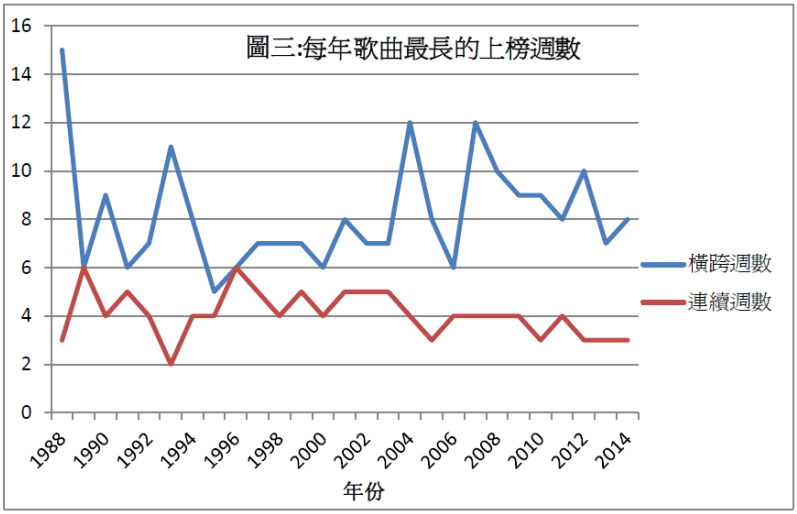
美鈴: 對!你也可以這麼說，但由於他的論文研究止於 1997 年，所以你下定論時只看第一至第二組好了。

媽媽: 哦! 那就是說「作曲人增多」可能並非導致香港樂壇水準下降的原因。

美鈴: 我也可以根據這表作一些推論，1984 至 2013 整體的作曲人人數下降，曲子的風格因而差不多; 歌手、編曲和作詞人的人數上升，卻使更多的歌曲面世。這樣，差不多的曲子一方面使樂迷感乏味，另一方面卻因太多歌曲而使競爭劇烈……

媽媽: 哦!

美鈴: 黃霑也曾說過:「因為競爭劇烈，歌曲的壽命也相繼縮短。以前，一首大家歡迎的歌，會流行三至六個月。到 90 年代中葉之後，一首歌，可以三四個星期仍然播放，已算是大熱了。又讓我來分析一下吧!



根據圖三，歌曲最長的橫跨上榜周數並沒有甚麼明顯的趨勢;而歌曲最長的連續上榜周數則可看見有輕微下降的趨勢，但也不是太明顯。不過我們卻可以看見「橫跨上榜週數」和「連續上榜週數」相差越來越多。你認為這代表了甚麼？

媽媽:你的數據好像有一些極端的數值，你要小心分析。

美鈴:我明白，這情況使用平均值並不恰當，因有極端數值影響，所以我同時用了中位數去表達。

表三:

	1988-1992	1993-1997	1998-2002	2003-2007	2008-2012
最長橫跨上榜週數平均值	8.6	7.4	7	9	9.2
最長連續上榜週數平均值	4.4	4.2	4.6	4	3.6
最長橫跨上榜週數中位數	7	7	7	8	9
最長連續上榜週數中位數	4	4	5	4	4

媽媽: 這是不是代表了近年歌曲流行的時段的間距較以前遠？如一首歌在今週很流行，在下週卻不流行，但相隔一個月後又從新流行起來。這使近年的歌曲橫跨上榜周數上升，連續上榜周數卻同時下降。以前歌曲的橫跨上榜周數雖然不及現在多，但「橫跨上榜周數」和「連續上榜周數」卻相差很少，由此可見以前流行歌的浪潮持續得較久，不像現在斷斷續續的。所以我贊同黃霑說法，歌曲壽命在某程度上是在縮短。

美鈴: 對了！

媽媽: 雖然如此，天王劉德華的歌仍然是百聽不厭，歌曲壽命還是這麼長。

美鈴: 等等，誰說劉德華是天王? 我的最愛容祖兒才是真正的天后呢！

媽媽: 天后怎會是容祖兒呢？你不要空口講白話！

美鈴: 我才沒有空口講白話！我是根據十大勁歌金曲獎裏歌手出現的次數推斷的！你看表四：

表四:十大勁歌金曲獎歌手次數分佈

名次\歌手(次數)	1984-1993	1994-2003	2004-2013
1	譚詠麟/張學友(11)	劉德華(10)	容祖兒(14)
2	梅艷芳(9)	許志安/郭富城(8)	古巨基(8)
3	葉蒨文(6)	陳慧琳/鄭秀文(7)	楊千嬅(6)
人次	98	73	105

媽媽: 你的表四好明顯有誤導成分！來看我的表五！

表五:1984-2013 年十大勁歌金曲獎出現次數最多的歌手的百分數

名次	歌手的百分數
1	劉德華 (13.7%)
2	容祖兒 (13.3%)
3	譚詠麟/張學友 (11.2%)
4	許志安/郭富城 (11.0%)
5	陳慧琳/鄭秀文 (9.59%)
6	梅艷芳 (9.18%)
7	古巨基 (7.62%)
8	葉蒨文 (6.12%)
9	楊千嬅 (5.71%)

美鈴:表五的資料你是怎麼得出的?

媽媽:首先你要將該歌手出現的次數除以該時段的總人數，再乘以 100%就行了。以你最愛的容祖兒為例就是：

14/105 × 100% =13.3%（三位有效數字）

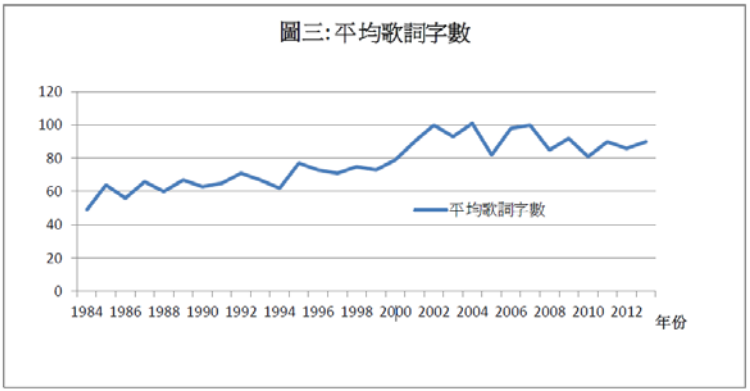
可見真正的天王應該是我的最愛

美鈴: 好吧！我們這兩代人有很大的代溝……

媽媽: 呵呵!黃霑也曾這樣說呢!

美鈴: (小聲地)又是黃霑！黃霑！黃霑……

媽媽: 他說「 .....年輕歌迷口味，自然和成熟歌迷不一樣。這一代的年輕人習慣奇特，喜歡聽多字的歌。認為歌詞寫得密，才可表達意思。因此在 90 年代後期流行的歌，旋律音調繁促，用音又多又密，幾乎令唱歌的人無法轉一口氣.....」(p.187) 我們也可用圖表來分析。



美鈴: 虛心受教！這些資料又是怎麼得出的？

媽媽: 很好！每年的資料是根據每年十大勁歌金曲獎得獎歌的

歌詞字數而得來的，先統計它們各自共有多少歌詞，再找出十首歌歌詞的總和，然後再找他們的平均值。假設 1984 年的歌詞總和是 500，便將 500/10=50，我們便能得出一年的歌詞字數。那你現在告訴我透過這表，你看見甚麼？

美鈴: 從以上的圖表，可看見在 1984–2013 年間，流行曲平均的歌詞字數有明顯的上升趨勢。所以黃霑的結論沒錯，看來我們的口味真的很不一樣……

媽媽: 好吧！你也可以這樣說。所以我還是最愛劉德華，還有我年輕時的歌。

美鈴: 哎！你年輕時不是有句歌詞說「舊夢不須記，逝去種種昨日經已死…舊事也不須記，事過境遷以後不再提起…」嗎? 都是忘記那些舊歌，聽聽越難越愛吧！

媽媽: 唉！雖然香港樂壇「紅日再不會昇、熱心漸似冰、從此星沉天際」，但「上海灘」又有誰能真正忘記呢？

–難忘燦爛誰會忘得了上海灘–

（~2400 字）

附錄一

第一組-1984-1993 年度十大勁歌金曲獎					
名次	歌曲	作曲/作詞/編曲	時間	歌名字數	平均每分鐘歌詞字數 *準確至整數
1984					
1	零時十分 葉倩文	林子祥/林振強/Chirs Babida	3:24	4	38
2	愛在深秋 譚詠麟	李鎬俊/林敏聰/盧東尼	4:00	4	55
3	似水流年 梅艷芳	喜多郎/鄭國江/黎小田	5:01	4	32
4	天籟 關正傑	盧冠廷/卡龍/--	3:55	2	32
5	再度孤獨 甄妮	伊滕薰/林振強/--	4:35	4	50
6	無敵是愛 許冠傑、甄妮	顧嘉輝/許冠傑/徐日勤	3:20	4	71
7	偶遇 林志美	李雅桑/鄭國江/--	3:22	2	33
8	Monica 張國榮	Nobody/黎彼得/--	3:47	1	93
9	愛的根源 譚詠麟	陳輩立/林敏聰/盧東尼	4:23	4	37
10	幻影 譚詠麟	林敏怡/林敏聰/林敏怡	4:29	2	49
1985					
1	愛情陷阱 譚詠麟	芹澤廣明/潘源良/入江純	3:50	4	45
2	情已逝 張學友	來生孝夫/潘源良/盧東尼	4:46	3	76
3	不羈的風 張國榮	Yoshiyuki Osawa/林振強/羅迪	4:37	4	92
4	聽不到的說話 呂方	杉真理/向雪懷/黎學斌	5:37	6	34

5	誰可相依 蘇芮	林敏怡/潘源良/林敏怡	4:49	4	40
6	日本娃娃 許冠傑	許冠傑/許冠傑/盧東尼	3:18	4	115
7	壞女孩 梅艷芳	C.DORE、J.LITTMAN/林振強/羅迪	4:46	3	102
8	暴風女神 Lorelei 譚詠麟	芹澤廣明/林振強/入江純	4:13	5	32
9	雨夜的浪漫 譚詠麟	鈴木キサブロー/向雪懷/盧東尼	4:38	5	36
10	十分十二时* 林子祥	--/--/鍾定一、黃良昇	10:16	5	--
*由於這首得獎歌是一首串燒歌，即組合了多首歌曲，有多個作曲和作詞人，而時間方面也較久，所以這資料將不會被統計。					
1986					
1	將冰山劈開 梅艷芳	M Cretu,H Kemmler /黎彼得/黃良昇	5:01	5	36
2	遙遠的她 張學友	谷村新司/潘源良/盧東尼	4:39	4	73
3	當年情 張國榮	顧嘉輝/黃霑/顧嘉輝	4:20	3	65
4	千個太陽 葉德嫻、陳潔靈	M.Gore,D.Pitchford/林振強/Romeo Diaz	3:52	4	62
5	千億個夜晚 林子祥	Foster, Keane, Jackson, Wakefiel/潘偉源/ Chris Babida	5:26	5	75
6	幾許風雨 羅文	Choo Seho/小美/杜自持	4:08	4	56
7	無言感激 譚詠麟	神林早人,深澤德/小美/盧東尼	4:47	4	49
8	朋友 譚詠麟	芹澤廣明/向雪懷/入江純	4:47	2	27
9	有誰共鳴 張國榮	谷村新司/小美/超增熹	4:06	4	61
10	夢伴 梅艷芳	Kisaburo Suzuki/林敏聰/黎小田	3:28	2	60
1987					
1	知心當玩偶	譚詠麟/陳少琪/盧東尼	4:21	5	46

	譚詠麟				
2	灰色 林憶蓮	Tambi Femando Iris Femando, Wayne Brown/林振強/杜自持	3:16	2	112
3	別人的歌 Raidas	黃耀光/林夕/黃耀光	3:30	4	99
4	烈酒紅唇 梅艷芳	倫永亮/潘偉源/倫永亮	3:31	4	103
5	流下眼淚前 徐小鳳	Keith, Peters/鄭國江/盧東尼	4:57	5	53
6	無邊的思憶 譚詠麟	網倉一也/盧永強/入江純	5:21	5	44
7	Don't Say Goodbye 譚詠麟	大津彰，鈴木キサブロー/黃真/盧東尼	4:43	3	38
8	太陽星辰 張學友	德永英明/儂儂/倫永亮	5:10	4	51
9	海誓山盟 林子祥	林敏怡/潘源良/林敏怡	4:31	4	37
10	無心睡眠 張國榮	郭小霖/林敏聰/船山基紀	3:21	4	73
		1988			
1	貼身 張國榮	盧東尼/陳少琪/盧東尼	3:29	2	80
2	Stand By Me 梅艷芳	King, Leiber, Stoller/陳少琪/趙文海	3:57	3	70
3	祝福 葉倩文	梁弘志/潘偉源/蘇德華	3:59	2	42
4	真的漢子 林子祥	林子祥/鄭國江/--	4:54	4	39
5	優女 陳慧嫻	M. L. Diego, M. T. Diego/林振強/盧東尼	3:47	2	99
6	大地 Beyond	黃家駒/劉卓輝/--	4:21	2	75
7	胭脂扣 梅艷芳	黎小田/鄧景生/黎小田	2:21	3	51
8	煙雨凄迷 陳百強	黎小田/潘偉源/黎小田	4:29	4	34
9	沉默是金 張國榮、許冠傑	張國榮/許冠傑/鮑比達	3:52	4	78

參考資料

1. 黃霑的博士學位論文. Retrieved from http://hub.hku.hk/bitstream/10722/31835/6/FullText.pdf?accept=1
2. 維基百科「香港四台冠軍歌曲清單」. Retrieved from http://zh.wikipedia.org/wiki/Category:%E9%A6%99%E6%B8%AF%E5%9B%9B%E5%8F%B0%E5%86%A0%E8%BB%8D%E6%AD%8C%E6%9B%B2%E5%88%97%E8%A1%A8
3. 維基百科 1984–2013 年「十大勁歌金曲獎」. Retrieved from http://zh.wikipedia.org/wiki/Category:%E5%8D%81%E5%A4%A7%E5%8B%81%E6%AD%8C%E9%87%91%E6%9B%B2%E9%A0%92%E7%8D%8E%E5%85%B8%E7%A6%AE
4. 魔鏡歌詞網 http://mojim.com/
5. 阿波羅娛樂 http://tw.aboluowang.com/2015/0324/532441.html
7. 發藏網 http://mingxing.facang.com/gangtai/17722.html

8. 和訊圖片

http://house.hexun.com/2014－03－26/163373925_24.html

9. 好戲網

http://www.mask9.com/node/48377

10. 頭條日報

http://news.stheadline.com/figure/index_r.asp?id=253

11. Time Out Hong Kong

http://www.timeout.com.hk/around-town/features/14771/beyond.html

12. 搜狗百科

http://baike.sogou.com/v111100.htm

13. 虎扑論壇

http://bbs.hupu.com/6533142.html

邀請作品：足印統計

香港有許多風景優美的沙灘，一步一步地經過長長的海岸線，就留下一串串足印。這些足印蘊藏著一些有用資訊，例如走過沙灘的人數和路徑，但亦包含了不少雜訊，有時凌亂或被海水沖去的足印都可能令人產生錯覺。如果能夠利用統計好好管理和分析這些足印數據，可能會有一些用處。

當主體「流動」的時候，就會產生足印。現實生活中，船隻隨著航道航行，車輛在公路上行駛，都會產生大量的數據。這些數據會不會被記錄，或怎樣被記錄，就視乎它的用途。以電子道路收費為例，用戶需要把將一個收發器放在車輛上，當車輛經過收費點時，收發器便會對收費點的感應器發出訊號，系統便會記錄車輛經過的日期及時間，以作繳費用途。但是，若於每一街道都裝上一個感應器，那麼每部車輛的「流動」情況都可以被記錄得一清二楚，像一個十分精確的定位系統。這當然是一個誇張的例子，但在八達通卡被普遍使用的香港，卡主的行蹤甚至購物記錄都曾被使用作為協助警方調查。

除了物理的「流動」外，我們瀏覽網頁的時候，在按「上一頁」和「下一頁」之間，都記錄著許多「虛擬足印」。以往，這些「虛擬足印」會用在登入管理等的範疇。例如，網站會使用的 HTTP Cookies 來識別用戶。現在，一些網上購物網站會記錄和分析用戶的「虛擬足印」，從而理解顧客的消費行為。例如，當用戶在瀏覽甲項商品的時候，他們總會按下同一個頁面的相關商品欄內的乙項商品。網上購物網站會設法從「虛擬足印」中找出這些關連，並且按照這些關連去設計網站，使用戶能夠更快更準的得知合適商品的資訊。隨著尋找成本降低，顧客的消費額便有可能提高。

雖然這些足印數據用途廣泛，但同時存在許多私隱的顧慮，畢竟沒有人希望自己的一舉一動會受監視。所以，怎樣記錄、儲存、使用這些足印數據等等，都應該更受規範和監管。我們處理這些足印數據，亦應該留意相關的規定。

參考資料

1. Professor Paul CHEUNG, Invited Keynote Lecture, "Using Location Information For Better Planning and Decision Support: Integrating Big Data, Official Statistics, Geo–information", International Workshop on Integrating Geospatial and Statistical Information, organized by United Nations Initiative on Global Geospatial Information Management (UN–GGIM), 9–12 June 2014, Beijing, China. Retrieved from https://sites.google.com/site/paulcheungpolo/lectures–and–presentations
2. Lee Rainie, Sara Kiesler, Ruogu Kang and Mary Madden, "Anonymity, Privacy, and Security Online", September 2013 http://www.pewinternet.org/2013/09/05/anonymity–privacy–and–security–online/
3. Autotoll Limited https://www.autotoll.com.hk/aboutautotoll.php
4. 2015 Google – Privacy Policy – Terms of Service – Program Policies https://support.google.com/mail/answer/6603?hl=en

邀請作品：車程遠，自然收費高？—

研究港鐵車費訂定的方程式

我每天上班都會乘搭港鐵。由沙田往金鐘，雖然要轉兩次車，分別乘三條線(東鐵線、觀塘線和荃灣線)，但只消半小時就到了。而每程收費$15.1，現在還有第二程九折優惠，每天的交通費尚算合理吧！每天來回均從相同的地點出發，坐著相同的交通工具，到相同的目的地上班，可有想過港鐵公司是如何訂定每程的車資呢？

根據港鐵公司的資料，港鐵的車資是依據一個票價結構，按車程距離釐訂。按車程距離釐訂似乎很合理，走多遠的路便收多少錢。但一般人又會覺得似乎不是那麼簡單，肯定有其他因素影響怎訂價的，最起碼過海與否已經是一大考慮，否則怎會相差一個站(由沙田往金鐘還是尖沙咀)，車費會相差超過$6 ($15.1 和 $9.0 的分別)。

讓我以東鐵線為簡單例子，試試拆解車費是如何訂定的。附件一表列了東鐵線的單程成人八達通車費，而附件二則是各站之間的路軌距離。如果純粹以一個站的距離來看(例

如紅磡至旺角東)，就可以得出以下少許結論：

(一) 無論距離多短，最少都收費 $3.6。

(二) 來往羅湖的車費是特別貴的。

(三) 最遠的車程收費 $4.2，反而九龍塘至大圍車程較短，但車資較貴($5.7)。

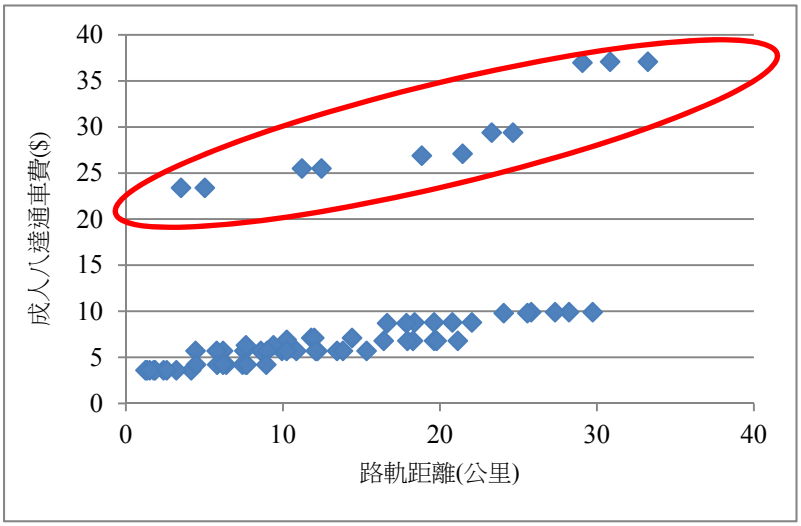
表一：東鐵線每個單站之間的距離與車費

來往	路軌距離 (公里)	車費 (港幣)
紅磡 – 旺角東	2.40	3.6
旺角東 – 九龍塘	1.76	3.6
九龍塘 – 大圍	4.43	5.7
大圍 – 沙田	1.35	3.6
沙田 – 火炭	1.86	3.6
火炭 – 大學	2.60	3.6
大學 – 大埔墟	6.39	4.2
大埔墟 – 太和	1.24	3.6
太和 – 粉嶺	6.19	4.2
粉嶺 – 上水	1.51	3.6
上水 – 羅湖	3.51	23.4

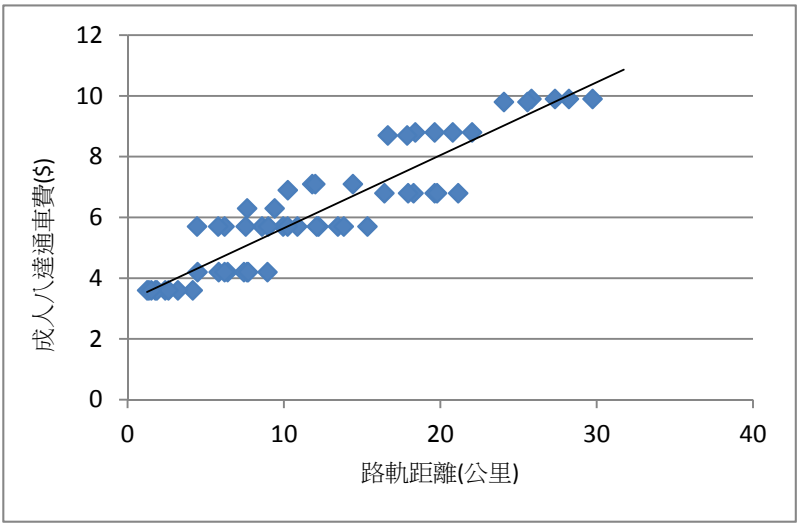
有了這基礎，我們就知道車資釐訂大概的準則了。當我們

把所有站與站的組合放在一起，可得以下圖一。紅圈所示為來往羅湖站的數據，價格比其他的高出好幾倍，似乎並不適合和其他數據一併分析。

圖一：東鐵線車費與路軌距離的關係(撇除馬場站外的所有車站組合)



圖二：東鐵線車費與路軌距離的關係(撇除馬場、羅湖站外的所有車站組合)



如果將羅湖站的數據先行拿走，剩下的好像呈現一個簡單的線性關係。用簡單的線性迴歸分析(simple linear regression)，以車費為 *Y*、路軌距離為 *X* 的話，會得出以下結果：

車費 $= 3.25 + 0.24 \times$ 路軌距離，調整 *R* 平方 $= 0.85$

雖然結果大致證明兩者的線性關係，但亦可能引來疑問，指某些車程的特殊因素如何反映在車費上。在再進一步探究之前，不如先解讀一下上述迴歸分析的結果。根據得出的方程式，我們可以想像車費分為基本和浮動兩部分組

成。基本車費可指為無論車程多短，最低限度都要收取的費用($3.25)，而浮動車費則與車程距離掛勾成正比例(以公里計算車程的 0.24 倍)。這種考慮實際上是符合商業原則。因為從營運成本角度出發，有些成本是與列車行走的長短沒太大關係，例如車站控制室或票務的運作、公司行政管理等固定成本。而另一種則是與行車有關的，譬如列車行走耗用的電力、路軌與相關設備的損耗折舊、列車的維修保養等的成本。從迴歸分析中，我們可以估計港鐵公司背後的計價方程式。

那麼，從九龍塘站至大圍站，要穿過獅子山，從表一都可看到隧道相關的額外成本都反映了在票價。我們可以用迴歸分析，量化這個因素嗎？答案是可以的，但牽涉到較複雜的概念，包括複合線性迴歸分析(multiple linear regression)和類別型變數(categorical variable)。簡單一點說，複合線性迴歸分析是牽涉兩個或以上變數(variables)的線性迴歸分析。以車費為例，我們可以試試尋找一下它與 (i)路軌距離和 (ii)車程中有沒有經過隧道的線性關係。

類別型變數其實是應用在處理剛才例子中(ii)的因素。因為「車程中有沒有經過隧道」這件事為二元選擇，即非「有」

即「無」，這已經並非線性關係了。要將這種變數放進迴歸模型，我們便要重新定義一下(ii)這個變數：如果那車程是有經過隧道的，變數的值就設為 1，否則就設為 0。舉例說，從紅磡至大學站的那個數據點，車費(Y)為\$7.1，而路軌距離(X_1)為 14.4 公里，經過隧道與否(X_2)為 1。如果是從沙田至大學站，車費(Y)為\$4.2，而路軌距離(X_1)為 4.46 公里，經過隧道與否(X_2)為 0。將這組數據再進行複合線性迴歸分析，結果會是：

車費 $= 3.11 + 0.20 \times$ 路軌距離 $+ 1.52 \times$ 經過隧道與否，
調整 $R^2 = 0.97$

大家可會留意到，這個迴歸模型的調整 R^2 比第一個模型高出很多。這是因為新加入的變數有助解釋到車費水平的差異，而平均來說，車程有經過隧道的，車費比沒有的高約 \$1.5。

如是者，當加入其他因素為變數和其他港鐵路線的車費，便可以逐步試出不同港鐵路線，車費定價的原則了！

附件一：東鐵線的單程成人車費(使用八達通)

單位：港幣

	紅磡	旺角東	九龍塘	大圍	沙田	火炭	馬場	大學	大埔墟	太和	粉嶺	上水	羅湖	落馬洲
紅磡		3.6	3.6	5.7	5.7	7.1	13.6	7.1	8.8	8.8	9.9	9.9	37.1	37.1
旺角東			3.6	5.7	5.7	6.3	13.1	7.1	8.8	8.8	9.9	9.9	37.1	37.1
九龍塘				5.7	5.7	6.3	13	6.9	8.7	8.7	9.8	9.8	37	37
大圍					3.6	3.6	7.1	4.2	5.7	5.7	6.8	6.8	29.4	29.4
沙田						3.6	7.1	4.2	5.7	5.7	6.8	6.8	29.4	29.4
火炭							6.8	3.6	5.7	5.7	6.8	6.8	27.1	27.1
馬場								6.8	7.3	7.3	10.7	10.7	25	25
大學									4.2	4.2	5.7	5.7	26.9	26.9
大埔墟										3.6	4.2	4.2	25.5	25.5
太和											4.2	4.2	25.5	25.5
粉嶺												3.6	23.4	23.4
上水													23.4	23.4
羅湖														23.4
落馬洲														

資料來源：港鐵公司車費表[http://www.mtr.com.hk/archive/en/tickets/octopus_fare201506.pdf]

附件二：東鐵線各站之間的路軌距離(公里)

單位：公里

	紅磡	旺角東	九龍塘	大圍	沙田	火炭	大學	大埔墟	太和	粉嶺	上水	羅湖
紅磡		2.40	4.16	8.59	9.94	11.8	14.40	20.79	22.03	28.22	29.73	33.24
旺角東			1.76	6.19	7.54	9.40	12.00	18.39	19.63	25.82	27.33	30.84
九龍塘				4.43	5.78	7.64	10.24	16.63	17.87	24.06	25.57	29.08
大圍					1.35	3.21	5.81	12.20	13.44	19.63	21.14	24.65
沙田						1.86	4.46	10.85	12.09	18.28	19.79	23.30
火炭							2.60	8.99	10.23	16.42	17.93	21.44
大學								6.39	7.63	13.82	15.33	18.84
大埔墟									1.24	7.43	8.94	12.45
太和										6.19	7.70	11.21
粉嶺											1.51	5.02
上水												3.51
羅湖												

資料來源：由於港鐵公司不願提供資料，以上距離以 Google Map 距離功能估算。註：

1. 來往羅湖與落馬洲站的車費相同，因此在分析中撇除了落馬洲站。
2. 馬場站為特別車站，只在賽馬日開放，亦從今次分析中撇除。

參考資料

1. 港鐵公司

   http://www.mtr.com.hk/ch/customer/main/index.html

2. Google Map

   https://www.google.com.hk/maps

3. 立法會交通事務委員會港鐵西港島綫的票價 [文件編號：CB(1)203/14–15(01)]

   http://www.legco.gov.hk/yr14–15/chinese/panels/tp/papers/tp20141125cb1–203–1–c.pdf

邀請作品：統計釋疑

二十一世紀，網絡知識充斥。垂手可得的網絡知識可靠嗎? 對我們生活又有何影響呢? 近來常聽聞「大數據」這字詞，「大數據」究竟能否幫助我們改善生活呢? 試想一想，你在日常生活中是用何種方法解決問題? 是否單憑直覺或一般推理判斷便能做到? 你有否嘗試運用統計知識來支持你的論點和決策呢? 你又懂得如何操作嗎?

部分同學認為學習統計是件苦差。統計學教育是以黑板與粉筆，或是只用紙筆來理解計算公式。學生通常都是以手工計算數十筆左右的資料，然後進行分析。同學們可能會覺得教科書的內容過於嚴肅刻板，所述說「統計的運用」的示例，亦較人工化，著重運算過程，較難顯示透過運用數據作決策的威力。市面上其實也有一些素材豐富的書本，但起點較高，未必適合一般中學生閱讀。本文的主要目的是向學生介紹兩個在歷史上應用統計知識釋疑的事件，引發學生對統計學產生興趣。

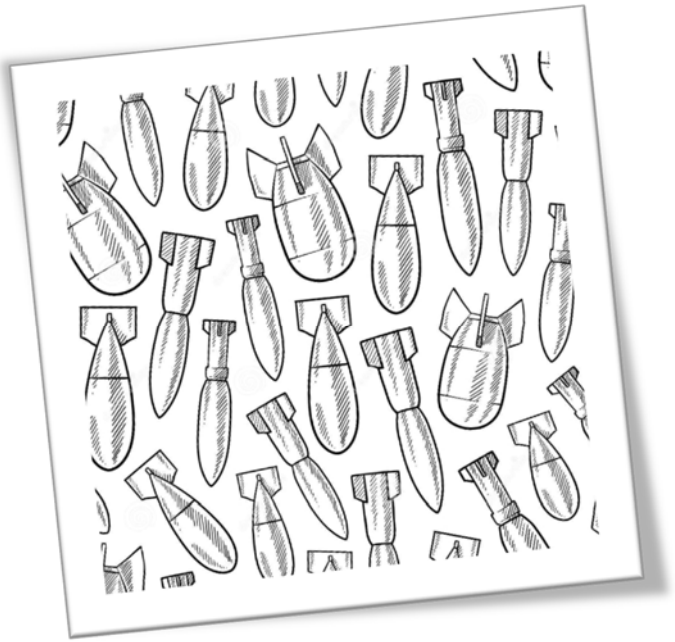
今年適逢抗戰勝利七十周年，就以二次大戰作背景，說一個故事吧。這個故事涉及英國與德國的戰事，看看統計知識如何幫助英國政府決策，抵禦德軍的攻擊。兩軍交戰，武器先進者佔優。二戰期間，德軍製造威力強大的新型彈道飛彈 V–2 火箭，向倫敦和英國的東南部大約發射了 1500

枚，殺害了 7250 人。彈道飛彈 V–2 火箭的速度比聲音更快，這意味受害人未聽見飛彈來臨已爆炸，造成嚴重傷亡。同學們若想略知當時的情況，可以上網搜尋相關網頁及知識，或 https://www.youtube.com/watch?v=MCdlBc__3kg 觀看 YOUTUBE 短片。

為了抵禦敵人的彈道飛彈，英軍需要部署防衛武器。怎樣佈防才適合呢? 英國政府當時有一個疑問，究竟德軍的彈道飛彈 V–2 火箭能瞄準目標發射，或只是隨機亂射，英軍的佈防大有不同。英國政府於是請求統計學家克拉克 (R.D. Clarke) 幫助，找出答案。統計學家克拉克所用的方法並不深奧，香港中學數學課程延伸部分單元一也有提及，就是利用卜松分佈(Poisson distribution)來分析飛彈的分佈點是否隨機事件。他首先把倫敦南區以 0.25 平方公里的一小區塊分為 576 塊，並根據統計資料，在每一小區塊數算飛彈落下的數目，結果如下表所示:

小區塊內的飛彈數目	實際小區塊的數目
0	229
1	211
2	93
3	35
4	7
5 或 以上	1
總數	576

從他所收集的數據，總共有 537 枚飛彈落在 576 塊區塊內，於是得出每一區塊的平均飛彈數目 λ 為 $\frac{537}{576} = 0.9323$。因此，克拉克要找出飛彈的分佈與平均值為 0.9323 的泊松分佈相符嗎？若在每個小區塊的飛彈數目 (X) 依循 Po(0.9323)，則 $\mathrm{P}(X = x) = \frac{0.9323^{x} e^{-0.9323}}{x!}$, $x = 0, 1, 2, ...$。根據這公式，可以計算飛彈落在不同小區塊的數目，並與實際數目比較。



小區塊內的飛彈數目	實際小區塊的數目	期望小區塊的數目 (Po(0.9323))
0	229	$\frac{0.9323^0 e^{-0.9323}}{0!} \times 576 = 226.7$
1	211	$\frac{0.9323^1 e^{-0.9323}}{1!} \times 576 = 211.4$
2	93	$\frac{0.9323^2 e^{-0.9323}}{2!} \times 576 = 98.5$
3	35	$\frac{0.9323^3 e^{-0.9323}}{3!} \times 576 = 30.6$
4	7	$\frac{0.9323^4 e^{-0.9323}}{4!} \times 576 = 7.1$
5 或 以上	1	$\left(1 - \frac{0.9323^0 e^{-0.9323}}{0!} - \frac{0.9323^1 e^{-0.9323}}{1!} - \frac{0.9323^2 e^{-0.9323}}{2!} - \frac{0.9323^3 e^{-0.9323}}{3!} - \frac{0.9323^4 e^{-0.9323}}{4!}\right) \times 576 = 1.6$
總數	576	

可以看出，數據的差異很小，符合泊松分佈的模型。雖然看來飛彈的落點是高度聚集在一些方格內，但統計告訴我們，這實際上是服從一個隨機、獨立事件的分佈規律。飛彈無法瞄準目標，德軍飛彈密集落在幾個目標區的說法不攻自破。克拉克同時也運用 χ^2 之適合度考驗 (The goodness of fit test) 對觀察數據與期望數據進行測定，結論相同。篇幅所限，此處無法細說。

有關另一故事是這樣的: 相信讀者聽過莎士比亞的名字，或看過其作品，如 *羅密歐與朱麗葉*、*王子復仇記* 等。一位美國學者泰勒於 1985 年 11 月在英國牛津大學的一所圖書館，找到一首詩，就稱為泰勒詩。有一派學者認為是莎士比亞的作品，另一派則認為這首詩在用字遣詞與風格上都和莎士比亞其他作品分別很大。兩派學者為這首詩的真偽爭論不休，大打筆戰。兩個月後，當時的兩位統計學者 Efron 與 Thisted 在一本 Science 雜誌上刊登了一篇文章，嘗試運用統計方法，鑑定這首詩是否為莎士比亞所作。

他們的想法是這樣的: 各人寫作有其用字習慣，特別是對於罕用字。每位作者使用的習慣，差異可以更大。 於是，他們將莎士比亞已知的所有作品輸入電腦中，並計算莎士比亞所用過的全部字數，及每一個字使用過的次數。莎士比亞全部作品之總字數為 884647，其中有 14376 個相異字只出現一次，4343 個相異字只出現 2 次，全部作品總共有 31534 個相異字，當中有 846 個字出現次數超過 100。那些在總作

品中，出現次數較低的，就當成莎士比亞的罕用字。這首泰勒詩相當短，共有 429 個字。若為莎士比亞所寫，根據統計總作品的資料，他們估計會有幾個字，在總作品中從未出現，只出現 1 次，2 次，…，99 次，都給出估計值。研究發現，實際情況與估計非常吻合。

任務完成? Efron 與 Thisted 認為這樣做還不夠。他們想到，會不會那時代的詩人，用字習慣都類似? 正如現代的年輕人受互聯網的影響，談話中多涉及「網上潮語」。於是，Efron 及 Thisted 亦拿幾位與莎士比亞同時代的詩人作品來比較，這樣就更保險了。兩位學者再找了三位與莎士比亞同時代的詩人，各取其一首詩，及另取四首莎士比亞所作的詩，與這首泰勒詩比較。經過 3 種統計檢定(詳情無法在此介紹性的短文表述)，發現對前三首詩，若假設為莎士比亞的作品，罕用字出現次數之實際值與估計值皆不吻合。而所挑選的四首莎士比亞的詩，雖偶有不合，但總的來說是可接受的。兩位統計學家的分析，並無法證明泰勒詩為莎士比亞所寫，但在罕用字之使用情況，卻與莎士比亞的總作品非常吻合。一場文學上的爭論，經統計學家發聲後，迅速平息，又一次證明統計的用途無所不在。

參考資料

1. Thomas H. Davenport, Jinho Kim(2013) *Keeping up with the quants: your guide to understanding and using analytics*. Harvard Business Review Press.

2. V– 1 Flying Bomb (2011) Fieseler Fi 103 (Vergeltungswaffe). Retrieved from: http://www.youtube.com/watch?NR=1&v=QY308O42Ur4&feature=endscreen

3. 莎士比亞新詩真偽之鑑定。取自: http://www.stat.nuk.edu.tw/cbme/math/appreciation/pdf/17–content.pdf

邀請作品：從球員身價到星系擴張 —

淺談齊夫定律

香港大學統計及精算學系 鍾玉嘉博士

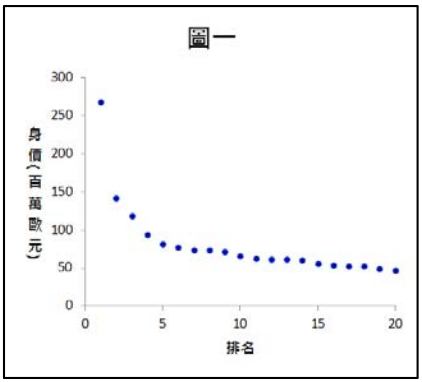
引言

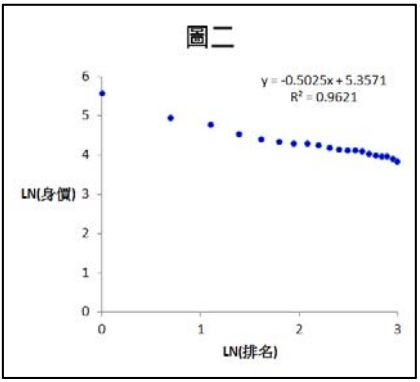
很多愛好觀賞足球比賽的朋友都常慨嘆：現今球壇在高度商業化之下土豪當道，足球早已從單純的運動項目轉變成為一個資本運作的商業項目。據歐洲足球統計權威機構 CIES (International Centre for Sports Studies) 於 2015 年 6 月發表的一份報告，全球最高身價的足球員三甲依次為效力於巴塞隆拿的美斯(2.55 億至 2.81 億歐元)，車路士的夏薩特(1.35 億至 1.49 億歐元)，與及皇家馬德里的 C.朗拿度(1.13 億至 1.25 億歐元)。球員的身價達到天文數字之巨，除了一些基本因素如球員在綠茵場上的超卓表現之外，更受著複雜的市場因素影響，但當中又有否一些簡單的規律可循呢？

身價排名	球員	估計轉會費(百萬歐元)	中間值
1	美斯	255.3 – 280.8	268.05
2	夏薩特	135.4 – 148.9	142.15
3	C.朗拿度	113.3 – 124.7	119.00
4	尼馬	89.6 – 98.5	94.05
5	阿古路	78.2 – 86.0	82.10
6	史達寧	73.7 – 81.0	77.35
7	普巴	70.3 – 77.3	73.80
8	迪亞高哥斯達	69.9 – 76.9	73.40
9	阿歷斯山齊士	67.8 – 74.6	71.20
10	占士洛迪古斯	62.5 – 68.8	65.65
11	蘇亞雷斯	60.0 – 66.0	63.00
12	基沙文	59.0 – 64.9	61.95
13	法比加斯	58.7 – 64.5	61.60
14	艾斯高	57.2 – 62.9	60.05
15	哈利簡尼	53.5 – 58.9	56.20
16	加里夫巴爾	51.3 – 56.4	53.85
17	古天奴	50.4 – 55.5	52.95
18	古圖奧斯	50.3 – 55.3	52.80
19	奧斯卡	47.5 – 52.3	49.90
20	賓斯馬	44.6 – 49.1	46.85

(*來源: CIES Football Observatory Monthly Report, Issue no. 6 – June 2015*)

上表依次羅列了全球身價最高二十位職業足球員的估計轉會費(身價)。為了簡單起見，以下圖表所用之身價只取其中間值。





圖一是首二十位球員身價與排名關係的散點圖。理所當然，排名愈後的球員身價愈低，但其減少的速度卻並非線性，尤以首幾位與餘下球員之間的差異為甚。假若我們以對數轉換後的變量來表達兩者之間的關係，便可以得出圖二中近乎直線的散點圖，相關係數 $r = -0.9808$ 表示其有著相當強的線性關係，用簡單線性迴歸分析，可得出方程： $\ln(\text{身價}) = 5.3571 - 0.5025 \times (\text{排名})$

與及轉換成原來的變量後，一個更簡單的近似關係式：

$$\text{身價} \alpha \frac{1}{\text{排名}^{0.5}}$$

換句話說，球員的身價與其排名的平方根成反比。這種特

殊關係，就是描述很多實際現象都很貼切的經驗法則之一：**齊夫定律**。

齊夫定律

齊夫定律 (Zipf's Law) 是以美國語言學家喬治・金斯利・齊夫 (George Kingsley Zipf) 命名的實驗定律。在上世紀四十年代，齊夫發現在愛爾蘭現代主義作家詹姆斯・喬伊斯 (James Joyce) 所著的長篇小說《尤利西斯》(Ulysses) 中，作者所用的詞匯，其出現的頻率隱隱然暗藏著簡單的規律。例如，"I" 一詞在書中共出現 2653 次，排名第 10；"say" 一詞出現 265 次，排名第 100；"bag" 一詞只出現 26 次，排名第 1000。明顯地，這些排名和頻率的乘積接近相同，因此，他大膽假設在大規模的語言系統中，詞匯的使用次數 (f) 與其排序 (k) 近乎成反比：$f = \frac{C}{k}$

換句話說，最常見的詞匯出現次數大約是第二名的兩倍，也大約是第三名的三倍，如此類推。在布朗語料庫 (Brown Corpus) 中，使用次數最多的是 "the" (69971)，差不多剛好是第二名 "of"(36411) 的兩倍。這種規律，後來在很多語言系統中得到驗證，而為著更靈活地描述這些系統的規律，我們會加入一個大於零的參數 s，從而得出一般性的齊夫

定律：

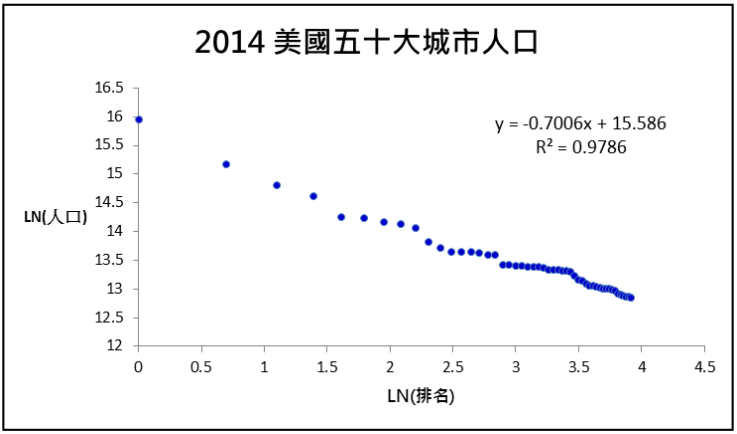
$$f = \frac{C}{k^s}$$

在上式中兩邊作對數變換後可得出以下線性方程：

$$\ln(f) = \ln(C) - s \times \ln(k)$$

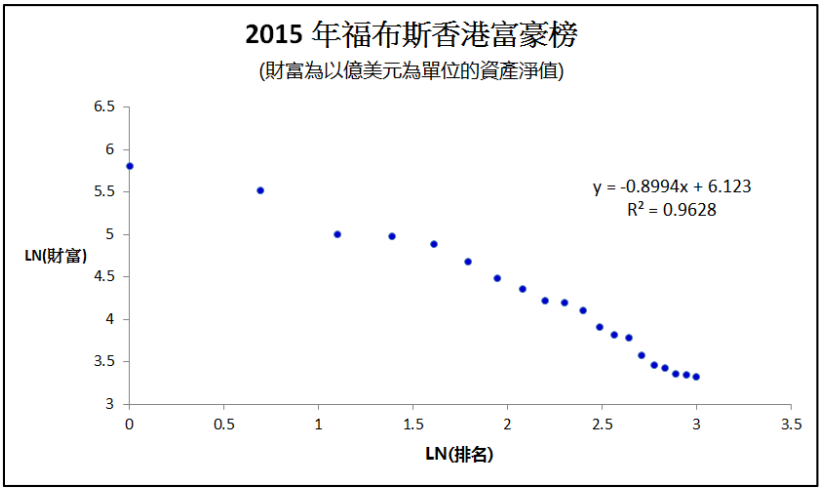
因此要驗證齊夫定律很容易，只需觀察縱橫軸坐標分別為 $\ln(f)$和 $\ln(k)$的散點圖，如其所有的點都近乎落於一條直線中，那麼它就遵從齊夫定律，而最適切迴歸直線的斜率就是參數s的估計值。

有趣的是，齊夫定律不只限於在語言學裡觀察得到，更可推而廣之適用於許多大規模個體互動的系統中，例如城市人口的研究。下圖是根據美國人口普查局在 2014 年估計的城市人口數字所作的散點圖，從其近似一直線的關係可見，城市人口數據和齊夫定律是相當吻合的。事實上，在不同地域和社會環境下的人口數據，也可觀察得到類似的結果。



(數據來源: United States Census Bureau)

另一個有趣的例子是社會中的財富分佈。下圖是根據 2015 年福布斯香港富豪榜，首二十位富豪的資產淨值對其排名的散點圖，從中也可以窺探到齊夫定律的蹤跡。



(數據來源: Forbes)

值得一提的是，如果我們假設齊夫定律適用於描述一個人口為 N 的社會的財富分佈，那麼最富有的 n 個人口所擁有的財富，佔整體社會總財富的比率大約可以以下式計算：

$$\text{比率} = \sum_{k=1}^{n} \frac{C}{k^s} \Big/ \sum_{k=1}^{N} \frac{C}{k^s} = \sum_{k=1}^{n} \frac{1}{k^s} \Big/ \sum_{k=1}^{N} \frac{1}{k^s}$$

以上圖中估算的 $s \approx 0.9$，以及

$N = 7000000, n = 0.2 \times N = 1400000$ 代入上式，可得出比率約為 0.815。換句話說，大約有 81.5% 的財富是被掌握在 20% 人手中，這和意大利經濟學家維弗雷多・帕雷托 (Vifredo Pareto) 早於 1896 年提出的二八定律(或稱 80/20 法則)相當吻合。而於數理統計學中，齊夫分佈可被視為帕雷托分佈的一個離散特例，故此兩者衍生出的定律本質上會有著不少共通的地方。

除了以上兩個應用於描述社會現像的例子，多年來不少學者都嘗試在其他範疇所得的數據中尋找類似的規律，而齊夫定律似乎也適用於揭示其他系統的內在秩序，這包括電影票房、暢銷貨品的銷量、網站點擊率等等的經濟數據；甚至也適用於描述月球環型山大小、太陽耀班的強度、地震級數、DNA 密碼等等的自然現象。

齊夫定律的解釋

齊夫定律是一個經驗法則，本來只是純粹由整理數據所得，背後沒有複雜的理論支持，但卻出奇地能夠和現實世界不同範疇的現象擬合，因此引起不少學者的興趣，嘗試去解釋為甚麼會出現這樣的一個規律。齊夫自己就提出了一個「最少努力原則」(Principle of Least Effort)：人類天生有惰性，面對一個問題要解決時，會盡量用最少努力來得到最大的效益。因此書寫時不會遍及使用所有詞彙，而只會頻繁地集中使用有限的某幾個；在選擇城市居住時會傾向於大城市，因其舒適且便利。當然這都只是齊夫的猜想，沒有經過任何嚴謹的數學證明，故此也無法用以解釋其他自然現象中的規律。

另有意見認為，齊夫定律並非魔法，而只是源於隨機過程的自然結果。以英語為例，假設我們從包括 a–z 和空格的 27 個字元中不斷獨立地隨機抽出字元來組成文本，再把空格之間的連串字母視為一個單詞，例如 “dfs_sa_ouoie_coivoer_i_yen_zw” 這「隨機句子」中就出現了 dfs, sa, ouoie, coivoer, i, yen, zw 這七個單詞。隨機產生的詞彙愈長，出現的次數就會愈少，且呈指數曲線下降。把所有可能產生的單詞依據其出現次數排序後，其頻率和排

名就會呈現齊夫定律所描述的關係, 且其 s 值接近為 1。生物訊息學家李問天教授於 1992 年為此提供了一個簡單的數學證明，並斷言齊夫定律可能只是一個統計假像，未必是語言系統的特性。不過，這數學證明只用於解釋語言系統的規律，並不適用於其他社會學和自然科學的範疇。

在 2015 年初，美國兩位天文學家，亨利・林 (Henry Lin) 與亞伯拉罕・勒伯 (Abraham Loeb) 提出了一個統一理論，嘗試把在城市發展、病毒擴散、銀河擴張等等觀察到的齊夫定律整合起來。他們先在二維平面建立一個人口密度變化的數學模型，再假設每個超密度城市的空間增減為隨機漫步過程，以此和早期宇宙物質密度變化的原理作出對比，發現兩者是一致的，因此天文學已有的數學工具可以應用到城市人口變化的數學模型上，從而輾轉推導與觀察數據相符的齊夫定律。推而廣之，其他與人口密度有關的現象 (如病毒擴散) 都可作如是觀。他們更相信，其統一理論可以解釋一切大規模人口互動的經驗法則，且與宇宙的發展是相同的。這種說法，有點像中國古代「天人合一」的哲學思想可堪玩味。

參考資料及延伸閱讀

1. Zipf GK (1949). Human behavior and the principle of least effort. *Journal of Consulting Psychology*, **13**:224–224.

2. Newman MEJ (2005). Power laws, Pareto distributions and Zipf's Law. *Contemporary Physics*, **46**:323–351.

3. Li WT (1992). Random texts exhibit Zipf–law–like word–frequency distribution. *IEEE Transactions on Information Theory*, **38**: 1842–1845.

4. Lin H and Loeb A (2015). A unifying theory for scaling laws of human populations. arXiv:1501.00738v2.

5. Saichev A, Malevergne Y, Sornette D (2010). *Theory of Zipf's Law and Beyond (Lecture Notes in Economics and Mathematical Systems).* Springer.

邀請作品：如何建立一個數據可視化

香港大學統計及精算學系 雷照盛博士

引言

數據可視化 (Data Visualization) 現今成為一個炙手可熱的題目，在多種不同的媒體中，我們都可發現它的存在，例如紐約時報(New York Times)利用數據可視化展示美國各州自 1900 以來的人口流動狀況[1]; 時代雜誌(Times Magazine)利用數據可視化的技術強調美國各州人口的差異[2]; 英國衛報(The Guardian)展示全球各國於 2007–12 的死刑人數數字[3]; 非政府組織(Non–Governmental Organization, NGO)利用數據可視化的技術把群體需要的訊息傳送到大眾當中[4]。

面對現今的大數據時代(Big Data Era)，數據可視化變得更為重要。由於需要面對複雜的數據結構，以及大數據的海量、高速和不確定性，有效的數據可視化可有助從大數據中找出重要資訊；與此同時，數據可視化也可以把大數據

1. http://www.nytimes.com/interactive/2014/08/13/upshot/where-people-in-each-state-were-born.html
2. http://content.time.com/time/interactive/0,31813,1549966,00.html
3. http://www.theguardian.com/news/datablog/2011/mar/29/death-penalty-countries-world
4. http://mastersofmedia.hum.uva.nl/2013/03/21/what-is-data-visualizations-goal-ngos-and-real-impact/

中的訊息有效地傳達到讀者群中。

這篇文章的主要目的是討論如何建立一個有效的數據可視化，以致把訊息有效地傳達到讀者群中。

根據 Illinsky and Steele (2011)，一個有效的數據可視化由以下四個元素組成：(1) 清晰的目標； (2) 合適的內容； (3) 正確的架構； (4) 有效用的格式。事實上，這四個元素是連續和有先後次序 (Sequential) 的，

下面將會解釋每個元素：

(1) 目標

在建立一個數據可視化之前，定立目標聲明 (Statement of Goal) 是重要的，因為一個明確的目標有助去決定放置合適的內容在數據可視化內。現在以一個簡單的例子闡述：

聲明 1：

「表示實驗結果。」

聲明 2 :

「表示在不同的外來刺激下，當中重要基因的變異率。」

明顯地，聲明 2 能夠指出一個比較明確的目標，以致能夠決定以後在數據可視化內放置相關內容。在訂立目標聲明的時候，其實可透過提出以下兩方面的問題找到答案：

(a) 從建立數據可視化的作者角度來看，

- 為何需要建立這個數據可視化？
- 誰是這個數據可視化的讀者？
- 讀者從這個數據可視化中領會到甚麼呢？

(b) 從觀看數據可視化的讀者角度來看，

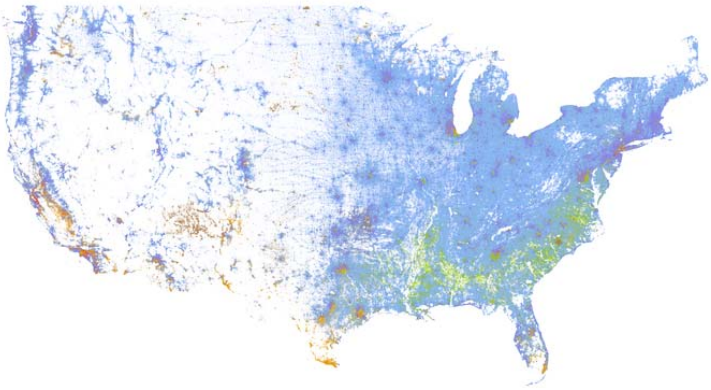
- 為何對方建立這個數據可視化？
- 誰是這個數據可視化的對象？
- 對方希望讀者從這個數據可視化中領會到甚麼呢？

(2) 內容

假若沒有定立一個目標聲明，當建立一個數據可視化的時候，一個可能出現的情況是把所有的數據和資訊放在一個數據可視化裡面，然而這個數據可視化可能會變得混亂一片。用一個 CNN 在 2013 年報道的作為例子，根據美國 2010 年的人口普查，整個地區的人口為 341,817,095，如果在一幅美加的地圖上用一點去代表一個人，這樣就會有超過三億點在一幅地圖上同時出現（如下圖），只能夠表示那一個地區的人口比較密集，那一個地區的人口比較稀疏，沒有太多的資訊能夠表達。

來源:http://edition.cnn.com/2013/01/22/tech/new–interactive–dot–map–aims–to–show–every–person–in–the–u–s/index.html

假若現在建立這個數據可視化的目標是表達不同的地區的種族分佈 (Racial Distribution)[5]，加入了種族 (Race) 這個因素，以不同顏色的代表不同種族，就會產生以下的數據可視化。



來源：http://demographics.coopercenter.org/DotMap/index.html

當決定把那一種內容放進一個數據可視化裡面，以下的問題是需要考慮的：

- 現在有那種數據可展示的？
- 現在有那種變量關係(Variable Relationship)可展示的？
- 根據以前所定立的目標聲明，需要展示的數據和變量關係。
- 有否把多餘的數據和變量關係包含在這個數據可視化內？

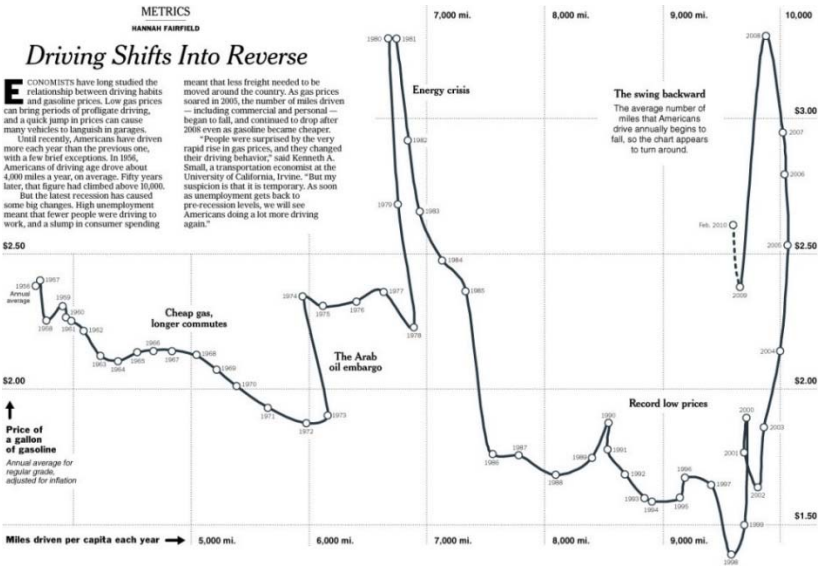
5. http://demographics.coopercenter.org/DotMap/index.html

(3) 架構

在前面的步驟，已經決定了需要展示的數據和變量關係以後，現在要決定的是數據可視化內的架構，需要考慮的問題是

- 如何以最佳狀態顯示重要的數據和變量關係？
- 選擇合適的編排 (layout) 和 軸線 (axes)
- 根據前面所定立的目標和內容

例如以下在紐約時報 2010 年報道 1956-2009 有關駕駛習慣與汽油價格的關係。傳統地，我們會使用時間序列圖 (Time Series Plot) 去表達汽油價格隨著時間的變化，以及使用散佈圖 (Scatter Plot) 去表達汽油價格和駕駛里數的關係。但下面的數據可視化為了表達汽油價格當中的重大轉變和駕駛習慣的改變，把內容編排 (layout) 和 軸線 (axes) 放在一個非傳統的設定之內。特別是顯示了 1973-74 石油禁運，1980-81 能源危機，2005-10 油價反彈。

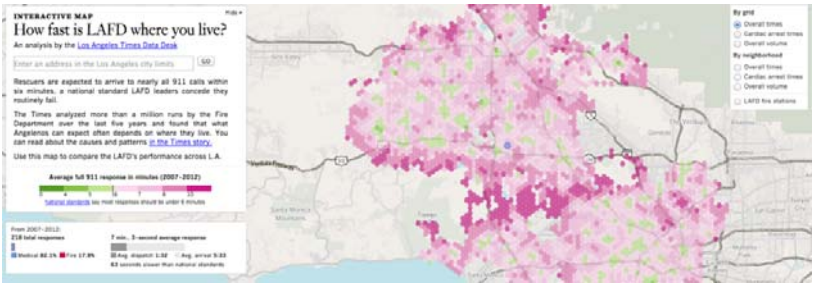


來源: http://www.nytimes.com/imagepages/2010/05/02/business/02metrics.html

(4) 格式

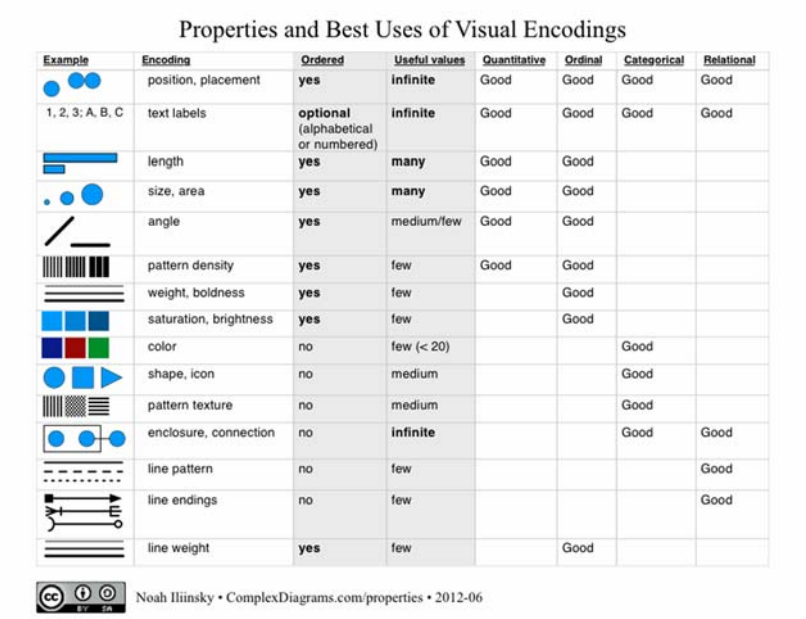
建立數據可視化的最後步驟是決定格式：根據前面所決定的目標、內容和架構，決定使用那一種格式去顯示數據和變量關係，特別是把重要的數據和變量關係強調出來。下面的例子是洛杉磯時報 (LA Times) 顯示 2007-2012 期間當地消防局對火警的平均回應時間 (Response Time)，越接近深粉紅色表示越長的回應時間，即狀況越危險；越接近深綠色表示越短的回應時間，即狀況越安全。在這個數據視覺化的例子中，顏色的選擇提供了一個很直接的表示，做

讀者容易明白。另外，除了對火警的回應時間以外，這個數據視覺化也可獨立地展示其他變量的數據，例如消防局對心臟停頓個案的平均回應時間和消防局地理位置。



來源：http://graphics.latimes.com/how–fast–is–lafd/

事實上，格式上的決定對應不同性質的變量會有不同的選擇，如 Illinsky and Steele (2011) 提供以下的建議：



Properties and Best Uses of Visual Encodings

Example	Encoding	Ordered	Useful values	Quantitative	Ordinal	Categorical	Relational
	position, placement	**yes**	**infinite**	Good	Good	Good	Good
1, 2, 3; A, B, C	text labels	**optional** (alphabetical or numbered)	**infinite**	Good	Good	Good	Good
	length	**yes**	**many**	Good	Good		
	size, area	**yes**	**many**	Good	Good		
	angle	**yes**	medium/few	Good	Good		
	pattern density	**yes**	few	Good	Good		
	weight, boldness	**yes**	few		Good		
	saturation, brightness	**yes**	few		Good		
	color	no	few (< 20)			Good	
	shape, icon	no	medium			Good	
	pattern texture	no	medium			Good	
	enclosure, connection	no	**infinite**			Good	Good
	line pattern	no	few				Good
	line endings	no	few				Good
	line weight	**yes**	few		Good		

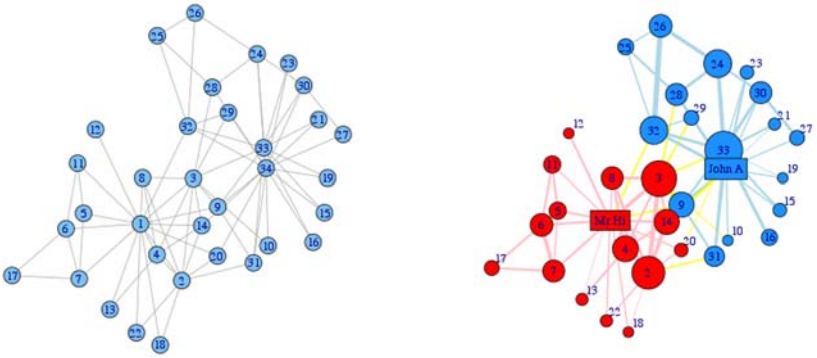
Noah Iliinsky • ComplexDiagrams.com/properties • 2012-06

來源：http://complexdiagrams.com/properties

三. 例子

現在考慮 Zachary's Karate Club 的資料集作為一個實際例子，這個資料集記載了 1970–1972 年間一所大學內的空手道學會裡面 32 位會員之間的互動而成為的社交網絡 (Social Network)，由於當中一位行政人員 John A 和一位導師 "Mr. Hi" 發生衝突，導致該空手道學會分裂，一半會員跟隨 Mr. Hi 成立了一個新的空手道學會，其餘的會員有部分找到了新的導師和有部分會員放棄了學空手道。此資料集的數據

現可在互聯網上公開下載[6]。



一般地，社交網絡的數據會以社交網絡圖 (Social Network Diagram) 顯示出來。以左上方的圖為例，每一個圓形代表一位會員；再加上 Mr. Hi 和 John A 分別以 1 號和 34 號表達，所以有 34 個圓形出現在網絡圖上。然而，儘管在左上圖當中已經用了網絡線去表示學會內人士的關連，但是每位會員與這兩個關鍵人物的關係並未能強調出來。

假若加入了格式的考慮，這個空手道學會的人脈關係可更清楚地表達在上右圖。關鍵人物分別以長方形表示 — 紅色和藍色分別表示分裂後兩個新空手道學會的形成，粗幼的紅藍網絡線表示會員與關鍵人物的強弱關係。從上右圖可發現大部分會員在分裂之後只與自己學會的會員聯繫，

6. https://networkdata.ics.uci.edu/data.php?id=105

少數會員仍有交叉聯繫，並用黃色表示出來。以上社交網絡圖是利用 R 語言編程產生出來。

四. 總結

最後，設計數據視覺化是一個熱門的題目。當中要考慮的因素非常多，特別是面對大數據的資料集，建立一個有效的數據視覺化是一個充滿挑戰性的工作。一方面，製作者需要更有效的演算法，因為面對大數據的資料集，快速地產生圖像是一個困難的工作。另外，複雜的數據結構需要更有效的技術把重要資訊顯示出來，所以，數據視覺化是一個充滿挑戰的研究課題。

參考資料

1. Illinsky, N. and J. Steele (2011), *Designing Data Visualizations: Intentional Communication from Data to Display*. Sebtastopol, CA: O'Reilly.

二零一四至一五年度中學生統計創意寫作比賽的籌備委員會：

主席	楊良河博士，香港大學統計及精算學系
總評審主任	張家俊博士，香港大學統計及精算學系
籌委會成員	陳秀騰先生，教育局
	陳家豪先生，政府統計處
	陳健昌先生，政府統計處
	陳浩榮先生，政府統計處
	李妍慧女士，政府統計處

數學百子櫃系列	作者
(一) 漫談數學學與教—新高中數學課程必修部分	張家麟、黃毅英、韓藝詩
(二) 漫談數學學與教新高中數學課程延伸部分單元一	韓藝詩、黃毅英、張家麟
(三) 漫談數學學與教新高中數學課程延伸部分單元二	黃毅英、張家麟、韓藝詩
(四) 談天說地話數學	梁子傑
(五) 數學的應用: 區像處理—矩陣世紀	陳漢夫
(六) 數學的應用: 投資組合及市場效率	楊良河
(七) 數學的應用: 基因及蛋白的分析	徐國榮
(八) 概率萬花筒	蕭文強、林建
(九) 數學中年漢的自述	劉松基
(十) 中學生統計創意寫作比賽 2009 作品集	
(十一) 從「微積分簡介」看數學觀與數學教學觀	張家麟、黃毅英
(十二) 2010/11 中學生統計創意寫作比賽作品集	
(十三) 2011/12 中學生統計創意寫作比賽作品集	
(十四) 數學教師不怕被學生難倒了! — 中小學數學教師所需的數學知識	黃毅英、張僑平
(十五) 2012/13 中學生統計創意寫作比賽作品集	
(十六) 尺規作圖實例、題解和證明	孔德偉
(十七) 摺紙與數學	阮華剛、譚志良
(十八) 2013/14 中學生統計創意寫作比賽作品集	