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

ISBN 978-988-8159-49-9

數學百子櫃系列(十八)　2013/14 中學生統計創意寫作比賽　作品集

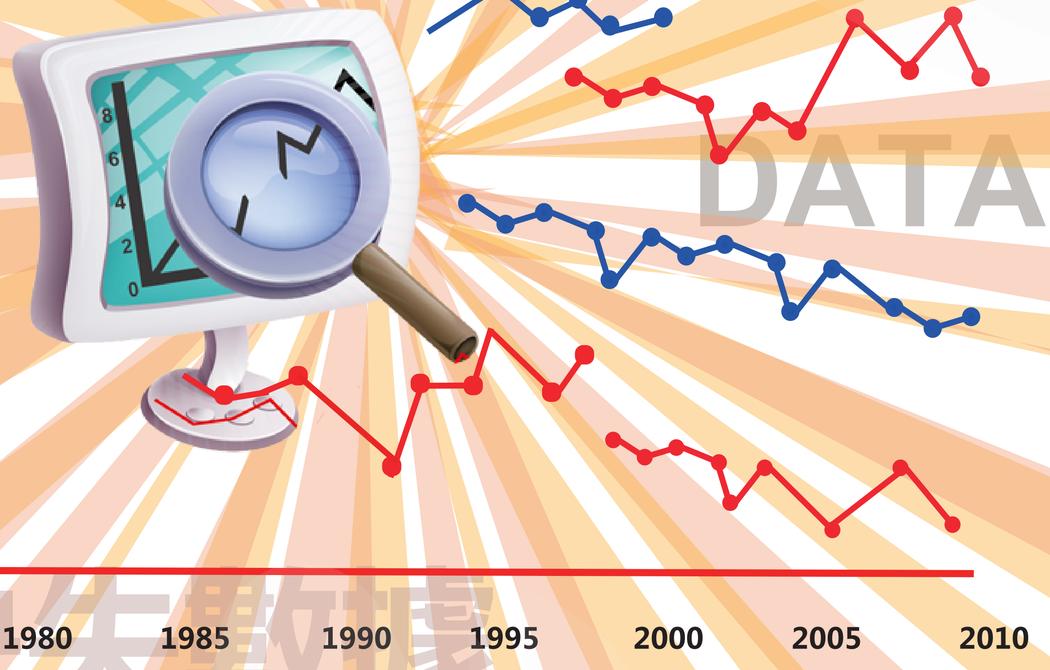
教育局數學教育組

2013/14

數學百子櫃系列 (十八)

中學生統計創意寫作比賽

作品集

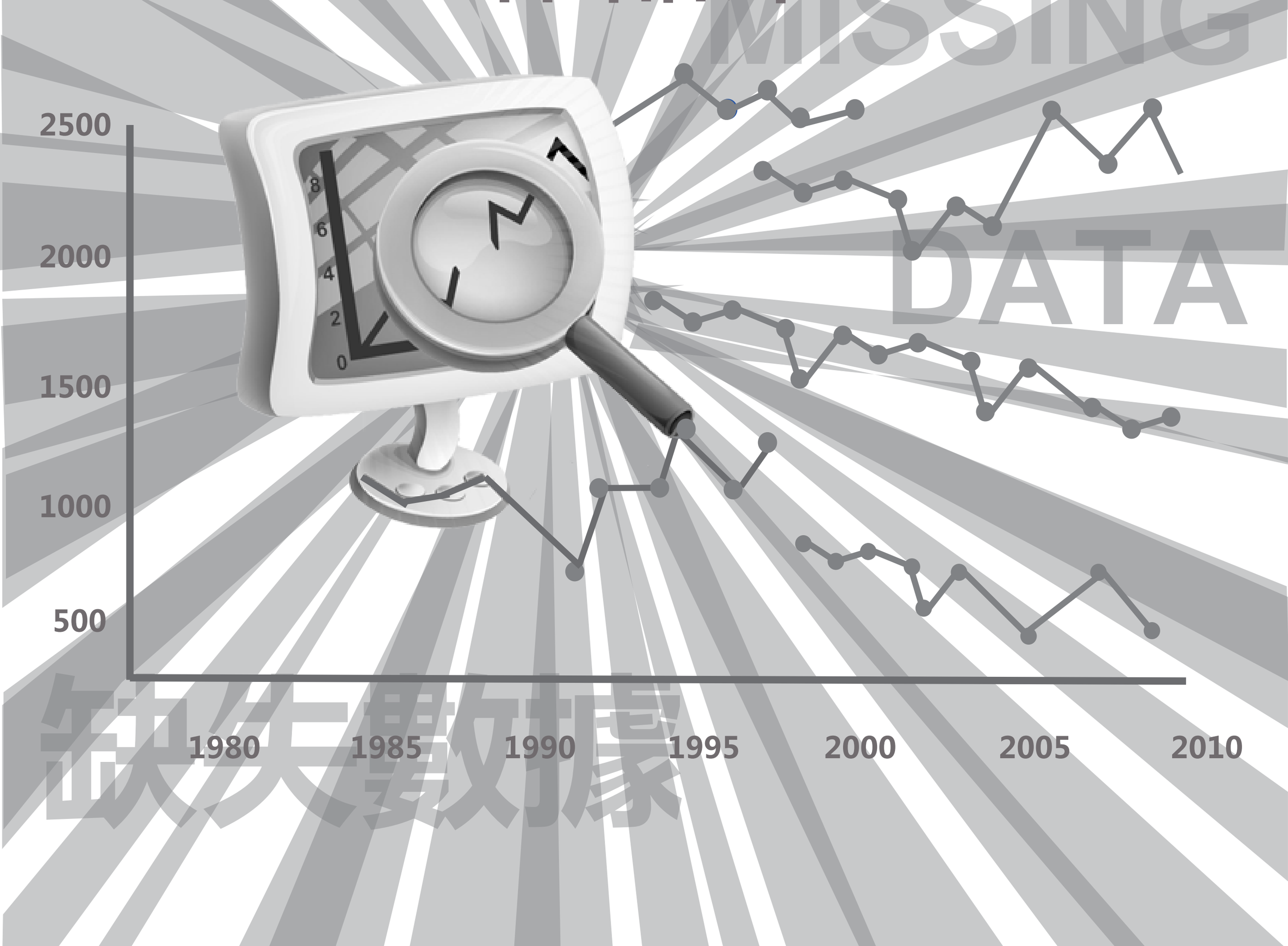


教育局
課程發展處數學教育組

數學百子櫃系列（十八）

2013/14 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

版權

©2014 本書版權屬香港特別行政區政府教育局所有。本書任何部分之文字及圖片等，如未獲版權持有人之書面同意，不得用任何方式抄襲、節錄或翻印作商業用途，亦不得以任何方式透過互聯網發放。

ISBN 978-988-8159-49-9

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《2013/14 中學生統計創意寫作比賽作品集》，是這個系列的第十八冊。本書輯錄的文章，大部分是「2013/14 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 8 篇為「2013/14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 6 篇則為邀請作品，分別由政府統計處的統計師，數學教育組的借調教師，以及香港大學統計及精算學系的教授撰寫。讀者們可隨意選讀各篇獨立的文章。本書的故事，大部分都是顯淺易懂，希望讀者閱讀本書後，感覺有趣，增加統計知識，善用「統計」這項客觀、邏輯和系統性的工具。

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的借調教師致以衷心的感謝。最後，更要多謝這次比賽的籌備委員會主席楊良河博士和評審委員會首席評審員張家俊博士。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，增加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「缺失數據」。多謝香港大學統計及精算學系副教授楊良河博士在比賽簡介會中介紹有關缺失數據的概念，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。本書輯錄今屆所有得獎作品，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。

我們藉此機會感謝籌備委員會和評審委員會全體成員和評審的幫助和支持。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。

籌備委員會主席　楊良河博士
評審委員會首席評審員　張家俊博士

2014年9月29日

目錄

冠軍作品: 跆拳武鬥會, 人多好辦事?

學校名稱：宣道會鄭榮之中學
學生姓名：區兆治
級別：中五
指導教師：朱吉樑

引言

跆拳道近年在世界各地日漸普及，逐漸發展成運動化的體育競技項目，並從 2000 年開始成為奧運會正式比賽項目之一。作為一種標準化的競技，裁判的精確判決故不可少。本文將以各種統計方法，討論裁判人數與判決精確性的關係。

跆拳道搏擊大賽臨近，盧教練和李先生正在討論比賽的詳情。

李：盧教練，你好。我兒子今年第一次參加搏擊比賽，能夠介紹一下比賽的評審方法嗎？

盧：當然可以。比賽以得分制計分。每施加一次有效攻擊便能獲得 1 分，並以最後分數決定勝負。按照慣例，比賽將會使用 3 位裁判，並在過半數的裁判，即 2 人認可得分下，才會正式給分。

李：為什麼要使用 3 位裁判？只用 1 個裁判有問題嗎？

盧：呃……首先，根據世界跆拳道聯盟所頒佈的跆拳道競賽指示，在沒有使用電子護具下，必需使用三位裁判或以上。另外，使用多位裁判，其實是為了使比賽的裁決更精確。事實上，由於搏擊比賽採用即時計分制，裁判需要在選手攻擊時對其有效性作即時判斷，未免有時會「看走眼」。經驗所得，一名裁判在選手攻擊有效時，正確認可該得分的機率為 0.9，不認可的機率為 0.1。

假設每個裁判的判斷都互相獨立，且每個裁判判斷時的機率皆相同。當裁判只有 1 位時，給分的機率= 0.9。

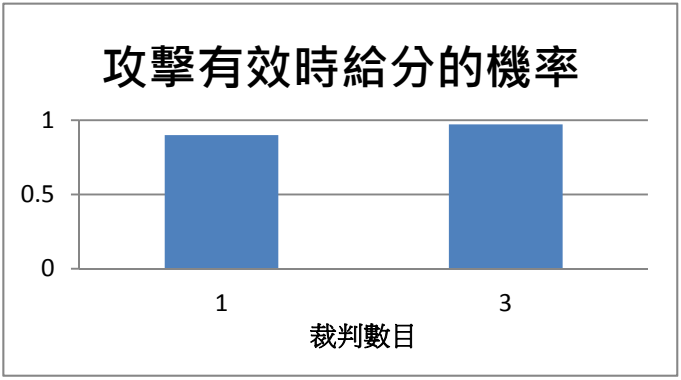
但是，當裁判有 3 位時，給分的機率

$$P(\text{給分}) = P(3\text{位認可得分}) + P(2\text{位認可得分})$$

$$= 0.9^3 + 0.9^2 \times 0.1 \times 3$$

$$= 0.972$$

3 位裁判中有 1 位裁判不給分的情況，共有 3 個組合，所以將其機率乘以 3。



圖一

由是觀之，當有 3 位裁判時，能夠正確給分的機率比只有 1 位裁判時高，判決故更為精確。

李：啊，原來如此，那為什麼不使用更多裁判呢？

是不是裁判越多，判決便越精確？

盧：那卻不一定……。雖然跆拳道的搏擊比賽一般會有 3 至 5 個裁判負責給分。但更多的裁判並不一定代表更準確的判決……

	攻擊有效	攻擊無效
給分	正確給分	錯誤給分
不給分	看走眼	正確不給分

表一

在評估裁判團判決的精準程度之前，我們先看看裁判團錯判的情況(見表一陰影部份)。錯判可分為兩種，分別為在攻擊有效時不給分，即是前述的「看走眼」，以及在攻擊無效時錯誤給分。由於攻擊有效與否是兩種各自獨立的情況，故不宜混為一談，故以下將會分別對攻擊有效與無效的情況進行分析。

情況一　攻擊有效

事實上，多位裁判對某攻擊的判決，可理解成重複進行的伯努利試驗。伯努利試驗即只有成功和失敗兩種結果的試驗。在這情況下，成功代表某一位裁判認可得分，失敗則代表不認可。認可有效攻擊的裁判數目(設該變量為 X) 遵從二項分

佈(Binomial Distribution)，即

$$P(X=k)=C_k^n p^k (1-p)^{n-k} \qquad \ldots\ldots (1)$$

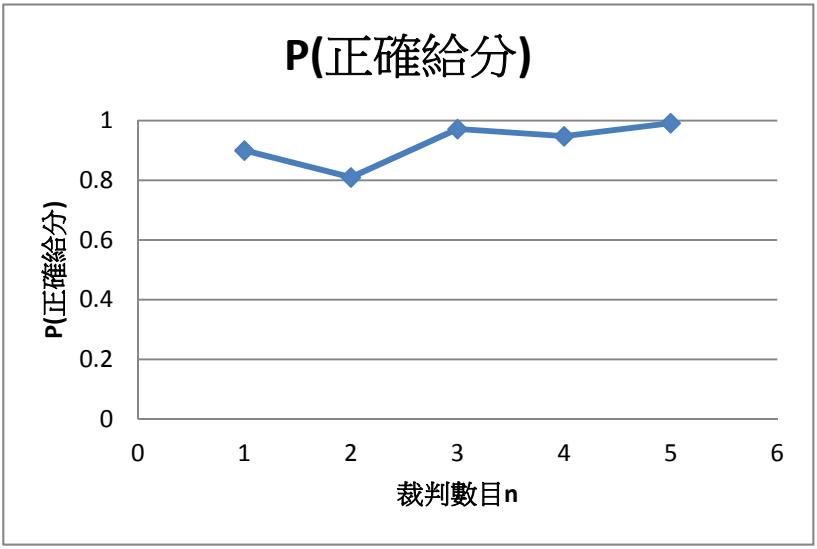
其中，n 為試驗次數，即裁判人數；p 為試驗成功的機率，在這情況下，是在有效攻擊下認可得分的機率。將 $n=\{1,2,3,4,5\}$，$p=0.9$，$k=\{0,1,2,3,4,5\}$，$(n \geq k)$代入(1)式，即可獲得在不同的裁判人數下認可有效攻擊的裁判數目的機率分佈(見下表)：

n＼k	0	1	2	3	4	5	P(給分)
1	0.1	0.9	-	-	-	-	0.9
2	0.01	0.18	0.81	-	-	-	0.81
3	0.001	0.027	0.243	0.729	-	-	0.972
4	1E-04	0.0036	0.0486	0.2916	0.6561	-	0.9477
5	1E-05	0.00045	0.0081	0.0729	0.32805	0.59049	0.99144

表二

從表二可以計算出認可得分的機率。值得注意的，是當 n 為雙數時，仍需要過半數的裁判認可才能給分。(例如，當 n=2，仍需 2 個裁判認可才給分；當 n=4，則需 3 個裁判認可。)

正確給分的機率可繪作下圖。由圖二可見，裁判數目增加，正確給分的機率並非單調上升，而是呈鋸齒狀上升。每當由單數 n 至雙數 n+1 時，得分的機率將稍為下跌。



圖二

情況二　攻擊無效

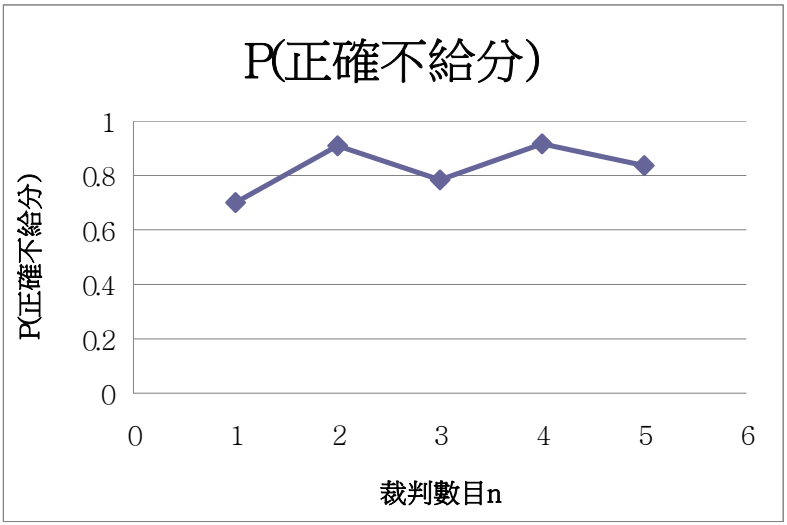
根據過往教練的考核紀錄，貴館的裁判有 0.3 的機率會錯誤地認可無效攻擊。

設認可無效攻擊的裁判數目為 Y，就如 X 一樣，變量 Y 亦遵從二項分佈。代入 $n = \{1,2,3,4,5\}$， $p = 0.3$， $k = \{0,1,2,3,4,5\}$ 代入式(1)，即可獲得在不同的裁判人數下認可無效攻擊的裁判數目的機率分佈(見下表)：

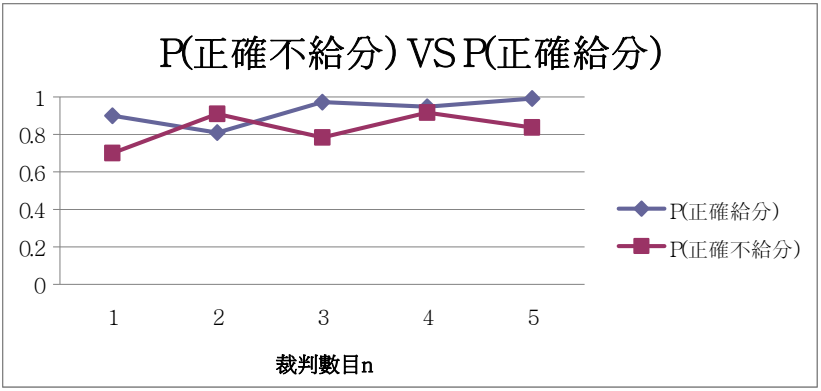
n＼k	0	1	2	3	4	5	P(錯 誤 給分)
1	0.7	0.3	-	-	-	-	0.3
2	0.49	0.42	0.09	-	-	-	0.09
3	0.343	0.441	0.189	0.027	-	-	0.216
4	0.2401	0.4116	0.2646	0.0756	0.0081	-	0.0837
5	0.16807	0.36015	0.3087	0.1323	0.02835	0.00243	0.16308

表三

從表三可以計算出正確不給分的機率 $[1-P(\text{錯誤給分})]$，繪成下圖：



由上圖可見，正確不給分的機率雖然大致隨裁判數目上升，每當由雙數 n 先至單數 $n+1$ 時，正確不給分的機率將稍為下跌。



圖四

事實上，若我們合併圖二及圖三(見圖四)，即可得知，當正確給分的機率上升時，正確不給分的機率便會下降，反之亦然。換句話說，正確給分或不給分的機率隨裁判人數的變化是恰恰相反的，就像一把雙刃刀，當正確給分的概率增加，正確不給分的概率會減少，反之亦然。所以，當要決定裁判人數時，我們需從正確給分和不給分兩者的概率中取捨。

王：那麼，應以甚麼為標準來決定搏擊比賽的裁判數目？

盧：我們可以用綜合正確機率，即綜合在攻擊有效及攻擊無效時，判斷正確的機率。來考慮不同裁判數目下的判決精準程度。

王：讓我來試試計算不同裁判數目下的綜合正確機率吧。假

設在每場比賽中，有 70%的攻擊為有效攻擊，30% 為無效攻擊。

P(裁判正確)

$= P$(給分 $\cap$ 有效攻擊) $+ P$(不給分 $\cap$ 無效攻擊)

$= P$(給分 | 有效攻擊) $\times P$(有效攻擊) $+ P$(不給分 | 無效攻擊) $\times P$(無效攻擊)

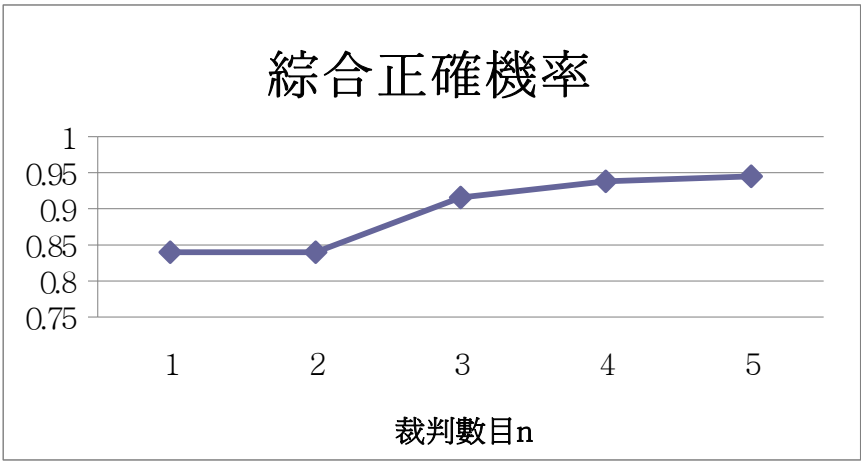
…… (2)

$= P$(給分 | 有效攻擊) $\times 0.7 + P$(不給分 | 無效攻擊) $\times 0.3$ …… (3)

將不同裁判數目的 P(給分 | 有效攻擊)、 P(不給分 | 無效攻擊) 代入(3)式，便可獲得 P(裁判正確)，如下表：

裁判數目 n	P(給分 | 有效攻擊)	P(不給分 | 無效攻擊)	P(判決正確)
1	0.9	0.7	0.84
2	0.81	0.91	0.84
3	0.972	0.784	0.9156
4	0.9477	0.9163	0.93828
5	0.99144	0.83692	0.945084

表四

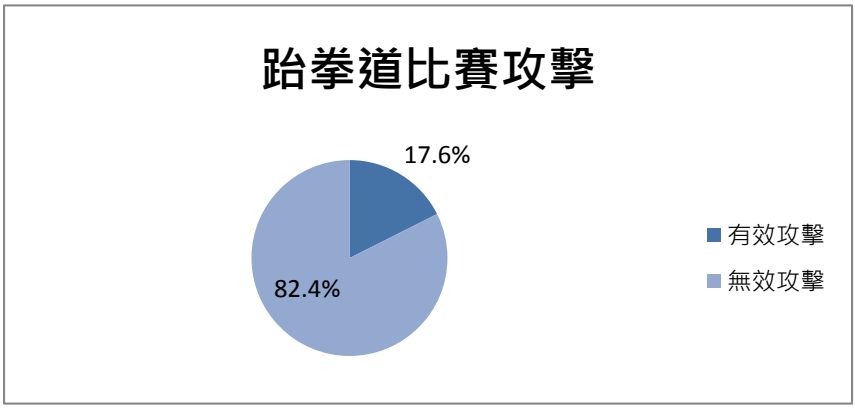


圖五

從圖五可見，在裁判數目增加下，綜合正確機率持續上升。

李：盧教練，我的計算如何？

盧：李先生，你的計算尚算合理。然而，事實上，跆拳道的攻擊中，要做到有效攻擊，並不容易。選手要在多次的無效攻擊下才能成功地對對手作出有效攻擊。因此，有效攻擊及無效攻擊的數目的概率假設分別為 0.7 及 0.3 並不正確。



圖六

＊據 Match Analysis in a University Taekwondo Championship (見參考資料 2)分析，在大學級賽事，男子組重量級的比賽中，每名選手得分及總攻擊數的平均值分別為 5.00 及 28.40。假設該大會無錯判情況，以及每 1 分代表 1 下有效攻擊。計算出有效攻擊佔總攻擊的比例大約為 5/28.40=0.176。

根據相關研究(見圖六)，每場比賽中，每位選手的攻擊，大概只有 17.6%的攻擊屬於有效，其餘 82.4%皆屬無效。

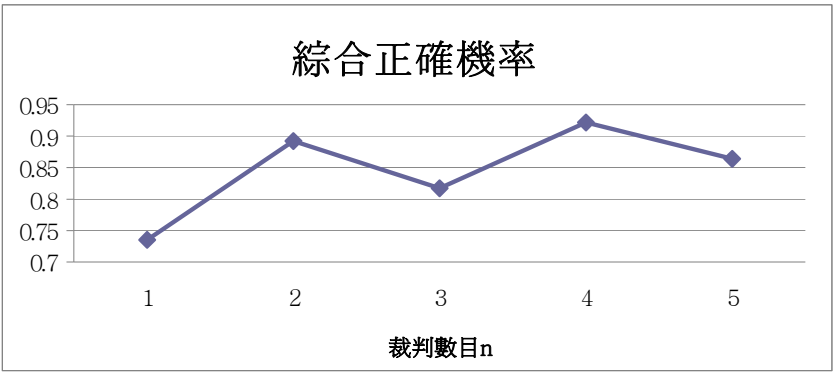
將 $P(\text{有效攻擊}) = 0.176$ 及 $P(\text{無效攻擊}) = 0.824$ 代入(2)式，得：

$$P(\text{裁判正確}) = P(\text{給分}|\text{有效攻擊}) \times 0.176 + P(\text{不給分}|\text{無效攻擊}) \times 0.824 \quad \ldots (4)$$

再將不同裁判數目下的 $P(\text{給分}|\text{有效攻擊})$、$P(\text{不給分}|\text{無效攻擊})$ 代入式(4)，得下表：

裁判數目 n	P(給分|有效攻擊)	P(不給分|無效攻擊)	P(判決正確)
1	0.9	0.7	0.7352
2	0.81	0.91	0.8924
3	0.972	0.784	0.8171
4	0.9477	0.9163	0.9218
5	0.99144	0.83692	0.8641

表五



圖七

根據上圖，綜合正確機率會在裁判人數由雙數 n 至單數 n+1 時下跌。在一般比賽有 3 – 5 位裁判中，在裁判數為 4 人時，判決正確的機率最高，其次為 5 人；在有 3 位裁判時，判決正確的機率最低。

黃：那麼，總結以上，在比賽中使用 4 位裁判應最為合宜，使用 5 人反而會降低判決的準確度！使用 3 位裁判時準確度

最低！既然如此，貴館會不會考慮增加一位裁判，以令判決更準確呢？

盧：這方面……只能說，本館的人力資源實在限啊。希望在不遠的將來，裁判數目的問題能夠有所改善吧！

(~2400 字)

參考資料:

1. 世界跆拳道聯盟 - 跆拳道競賽規則
   http://mtpe.mtwww.mt.au.edu.tw/ezcatfiles/b017/img/img/369/27.doc

2. Coral Falco, Raúl Landeo, Cristina Menescardi, José Luis Bermejo, Isaac Estevan (2012), Match Analysis in a University Taekwondo Championship
   http://www.scirp.org/journal/PaperDownload.aspx?paperID=17208

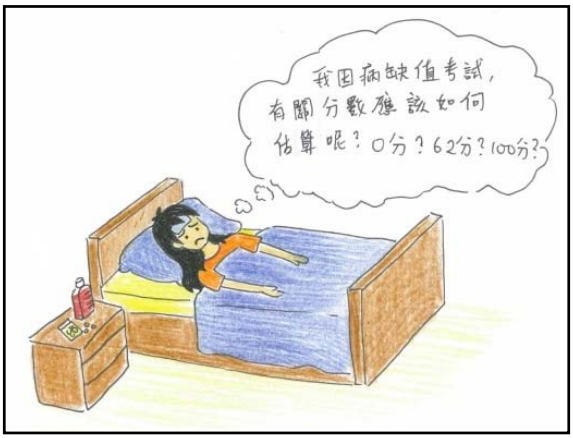
3. 跆拳道- 维基百科，自由的百科全書
   http://zh.wikipedia.org/wiki/%E8%B7%86%E6%8B%B3%E9%81%93

亞軍及最佳專題寫作作品：缺席生的分數捍衛戰

學校名稱：香港神託會培基書院
學生姓名：方綽瑤，許蔚瑩
級別：中五
指導教師：黃智君

引言

考試經常都會有學生因病缺席，令他們錯過了一些計分的機會。那些缺值生分數應如何處理呢？怎樣的分數估計才對其他學生公平呢？筆者將會帶領大家走入校園，透過因缺席下學期考試的經歷，分享自己向老師爭取應有分數而爆發的一場「分數捍衛戰」，並藉以探討估計缺席生成績背後的統計學。



「瑩瑩！因為你缺值下學期的數學科考試，所以你的分數將會是零分！」當我聽到老師的話，立刻從夢中驚醒過來！在下學期考試期間，我因為得了肺炎，不能回校參加數學科考試。雖然我的病情好多了，卻很擔心自己的成績。為公平起見，學校規定不會為缺值生安排補考。為免老師錯誤估計自己的成績，我必須先想想分數估計的方法，然後回校向老師問過清楚明白，誓要捍衛自己應得的分數！

表一：上學期及下學期考試分數的比較

學生	上學期考試分數	下學期考試分數
A	100	缺席
B	96	74
C	95	66
D	90	69
E	88	64
F	86	58
G	83	67
H	80	52
I	75	48
J	72	41
K	72	43
L	70	39
M	68	45
N	65	35

學生	上學期考試分數	下學期考試分數
O	64	37
P(筆者)	62	缺席
Q	59	18
R	55	27
平均分	76.67	48.94
標準差	13.25	15.72

方法一：以缺席生的上學期考試分數作取代

若果老師利用缺席生的上學期考試分數取代下學期考試成績，這樣很容易就可以處理缺席生的成績，可大大減輕老師的工作負擔。另外，這方法無需複雜的計算，所以其他師生亦易於理解。可是，從兩次考試的平均值可見上下學期考試的深淺程度不一，下學期考試明顯較深。在這方法下，缺值生(如學生 A)就可以不勞而獲地得到相對較高的分數（即 100 分），這對其他考生非常不公平。由於這方法未有考慮兩次考試的深淺之別，相信老師未必會採納這建議。我還是回校請教黃老師，先了解學校現有的分數估計方法吧！

方法二：以上下學期的名次排序作估計

從黃老師口中得知，學校現有的方法是將所有學生上學期和下學期的考試成績分別地由高至低排序，以缺值生上學期考試成績的名次作為參考，在下學期考試成績的排序中找出相同的名次，並以這個名次在下學期的分數取代缺值生的成

績。最後，學校為保障其他學生的利益及避免考生無故缺值，缺值生的分數會以下學期成績的上四分三位數作為上限。

表二：上學期及下學期考試成績的名次比較

名次	上學期考試分數	下學期考試分數
1	100 (學生 A)	74
2	96	69
3	95	67
4	90	66
5	88	64
6	86	58
7	83	52
8	80	48
9	75	45
10	72	43
11	72	41
12	70	39
13	68	37 (學生 O)
14	65	35
15	64 (學生 O)	27
16	62 (筆者)	18
17	59	
18	55	

以學生 A 為例，他在上學期考試中得到第一名，可是他缺席了下學期考試，所以老師本應以下學期第一名的分數取代他的成績(即 74 分)，但 74 分高出下學期成績的上四分三位數(即 65 分)，所以他最後的估計分數是 65 分。而我在上學期考試中的名次是第十六名，所以老師會以下學期第十六名所得的分數取代我的成績。在這分數估計方法下，我在下學期只能得到 18 分，遠遠低於自己在上學期考試所得的 62 分。當我得知這個估計分數時，我頓時怒火中燒，心情久久平靜不下。我認為這個分數估計方法是極不公平，因為這個估計分數只來自一個有相同名次的學生，未有考慮其他能力相約學生的情況。以上學期的第十四名及第十五名為例，他們上學期的分數(即 65 分及 64 分)與我相約，但他們下學期的分數遠遠高於 18 分！

另外，假設學生 O(上學期的第十五名)在下學期也缺席考試，那麼下學期考試成績就會缺少了第十六至十八名的數據。而我在下學期考試缺席的情況下，便無法以下學期考試第十六名的分數來取代我下學期考試的成績，只能以下學期考試中最低的分數(即 18 分)來取代。在缺值生太多的情況下，能力最弱的一批學生(即上學期的第十六至十八名）的估計分數只能以下學期考試中最低的分數來取代，這做法極不合理！

由此可見，學校現有的分數估計方法沒有考慮到所有學生的整體情況，而缺席生的估計分數只來自一個有相同名次的學生，這做法未免有點以偏概全！為了捍衛自己應得的分數，我決定和黃老師一起研究一個更公平、公正的方法來估計缺席生的成績。

方法三：以上下學期的標準分作估計

在與黃老師的討論中，我想起在中五的數學課學了標準分這個概念，未知能否更合理地估計出缺席生下學期考試的成績呢？

$$\text{學生的標準分} = \frac{\text{學生的分數} - \text{全班的平均分}}{\text{全班的標準差}}$$

當標準分的絕對值越小，即表示學生的分數相對較接近全班的平均分。相反地，標準分的絕對值越大，即表示學生的分數相對遠離近全班的平均分。以上學期的成績為例，

$$\text{學生 A 的標準分} = \frac{100 - 76.67}{13.25} = 1.76$$

$$\text{筆者的標準分} = \frac{62 - 76.67}{13.25} = -1.11$$

1.76 表示學生 A 的分數較平均分高 1.76 個標準差，而–1.11 即表示我的分數較平均分低 1.11 個標準差。假設缺值生在下學期考試有相同的表現，先以他們上學期的標準分作參考，再利用下學期的全班平均分及標準差來估計他們下學期考試的分數。

$$\text{學生 A 的分數估計：} \quad \frac{a - 48.94}{15.72} = 1.76$$

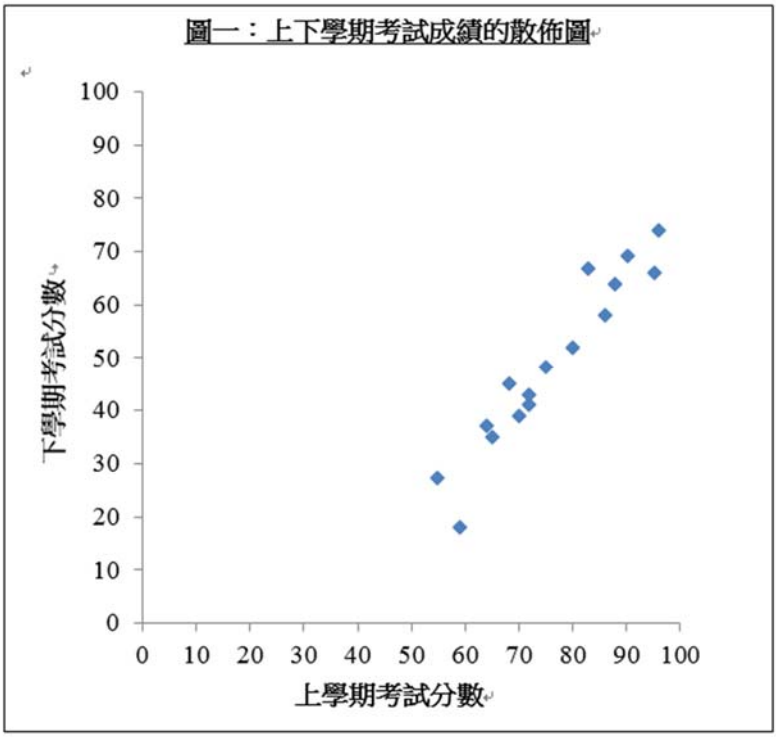
$$a = 76.61$$

筆者的分數估計：
$$\frac{p-48.94}{15.72}=-1.11$$
$$p=31.49$$

我們能估計出學生 A 下學期考試的分數為 76.61，而筆者的估計分數為 31.49。這個方法有考慮兩次考試的深淺之別，並以缺值生與全班平均分之相對差距作估計，所以估計的分數會比較合理。可是，黃老師認為這個方法涉及標準差及標準分的概念，擔心其他師生未必能理解這複雜的計算。因此，我們決定再想一個其他師生較易明白的分數估計方法。

方法四：以「線性迴歸分析」作估計

若上下學期考試成績有關聯的話，我們不就可以使用散佈圖(Scatter Diagram)來作進一步分析嗎？這個方法較圖像化，相信較易令一般師生理解和接受。在散佈圖中，我們可以將上學期考試分數設為橫軸，並將下學期考試分數設為縱軸，然後將十六位同學的數據逐一標示在圖上。



從以上的散佈圖可見，當上學期考試分數增加時，下學期考試分數亦有增加的傾向。那究竟它們之間有甚麼特別關係呢?黃老師指出在統計學上，兩者的關係稱為「正相關」，我們更可以利用一個具體的數值來客觀準確地表達兩者的關係，而那就是相關係數(correlation coefficient)。其實，相關係數是一個可以用作表示二個變數之間的關聯性的數值，而計算方法就加下:

假設現在有兩種相互對應的 n 組數據：

$\{x_1, x_2, ..., x_n\}, \{y_1, y_2, ..., y_n\}$

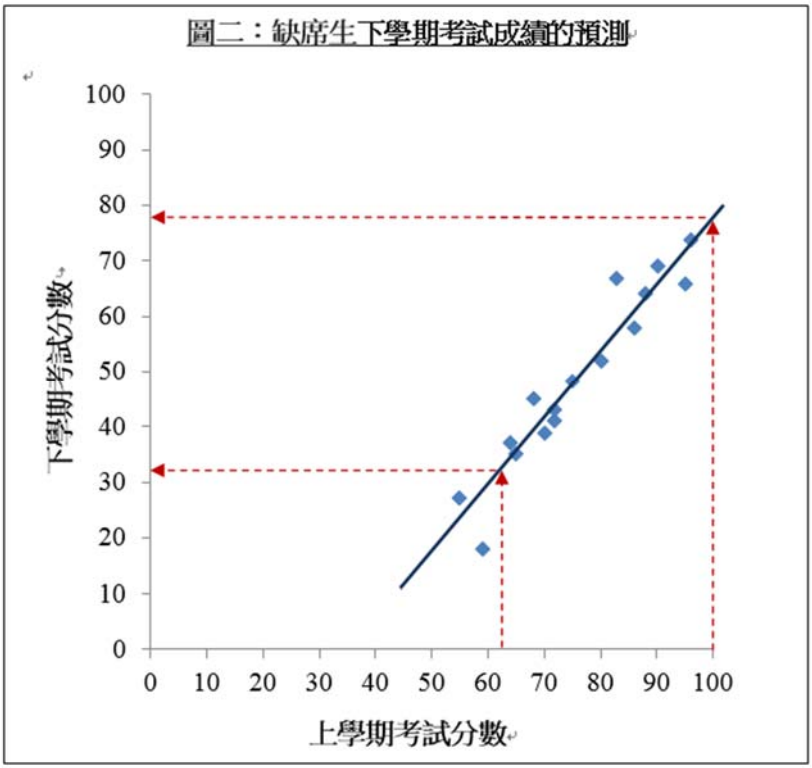
相關係數 $r = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \times \sum \sqrt{(y_i - \bar{y})^2}}$

Σ 是代表「合計值」的符號，而 $\bar{x}$ 和 $\bar{y}$ 則分別代表兩種數據的平均數。

相關係數的數值介乎 -1 至 1 之間。當數值越接近 -1 或 1，關係就越緊密；相反地，數值越接近 0，關係就越薄弱。我們只要利用電腦的試算表(Excel)計算出的需要合計值(詳細列表計算請參考附件一)，再代入公式便可以找出上學期考試和下學期考試的相關係數：

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \times \sum \sqrt{(y_i - \bar{y})^2}} = \frac{2947.13}{\sqrt{2393.75} \times \sqrt{3954.94}} = 0.96$$

由於數值很接近 1，所以兩者的關聯性極強，即表示上下學期考試分數有著極密切的關係。老師更指出在這張散佈圖中，上學期考試成績跟下學期考試成績有著一定的線性關係。在統計學上，這條「隱藏」的直線稱為「迴歸線」。只要我們知道這條「迴歸線」的畫法和其直線方程，就能估計缺席生的下學期考試分數，而這種估計的方法稱為「線性迴歸分析」。



要找出圖中的迴歸線，使用直線方程 $(y = mx + c)$便可。當中 m 代表直線的斜率(slope)，而 c 則代表直線的縱軸截距(y-intercept) 。

$$m = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$= \frac{2947.13}{2393.75}$$
$$= 1.2312$$

因為數據的平均值在迴歸線上，所以我們可以得出以下算式來找出 c 的數值：

$$c = \bar{y} - m\bar{x}$$
$$= 48.94 - (1.2312)(76.13)$$
$$= -44.79$$

因此，上學期考試成績跟下學期考試成績的關係就可用以下的直線方程來表達：

而我們只要將缺值生上學期的分數代入迴歸線的直線方程，便能估計缺席生下學期考試的成績：

學生 A 的分數估計： $y = 1.2312(100) - 44.79$
$= 78.33$

筆者的分數估計： $y = 1.2312(62) - 44.79$
$= 31.54$

通過以上的估計，學生 A 下學期考試的分數為 78.33，而我的估計分數為 31.54，這與利用標準分的估計結果相當接近（即 76.61 分及 31.49 分）。這個分數估計方法既有考慮兩次考試的深淺之別，並以全班學生的成績作線性迴歸分析，而「迴歸線」的圖像表達可令一般師生較易接受。雖然這個方法涉及複雜的計算，但我們可利用電腦輔助計算。因此，黃老師決定接納我的建議，用這個方法來估計缺席生的下學期考試分數。

總結

這場「分數捍衛戰」不但使我得到較公平、合理的估計分數，老師更讚賞我對解決難題的執著和熱誠。通過探討不同的分數估計方法，我明白到缺失數據的處理方法及相關的統計概念。其實，以上的四個方法都各有利弊(詳細列表比較請參閱附件二)，而且缺值生所得的分數只是一個估計值，並不能反映他們在下學期的進步或退步。學校只可因應不同情況作出分數估計，但這只算是一種補救方法。若要知道自己的真正實力，我還是養好身體、努力讀書，與其他同學在下次考試來個真實的比拚吧！

(2485 字)

附件一：相關係數的計算

學生	上學期考試分數	下學期考試分數	離均差		離均差平方		離均差乘交積
	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
A	100	缺席					
B	96	74	19.88	25.06	395.02	628.13	498.12
C	95	66	18.88	17.06	356.27	291.13	322.05
D	90	69	13.88	20.06	192.52	402.50	278.37
E	88	64	11.88	15.06	141.02	226.88	178.87
F	86	58	9.88	9.06	97.52	82.13	89.49
G	83	67	6.88	18.06	47.27	326.25	124.18
H	80	52	3.88	3.06	15.02	9.38	11.87
I	75	48	-1.13	-0.94	1.27	0.88	1.05
J	72	41	-4.13	-7.94	17.02	63.00	32.74

K	72	43	-4.13	-5.94	17.02	35.25	24.49
L	70	39	-6.13	-9.94	37.52	98.75	60.87
M	68	45	-8.13	-3.94	66.02	15.50	31.99
N	65	35	-11.13	-13.94	123.77	194.25	155.05
O	64	37	-12.13	-11.94	147.02	142.50	144.74
P(筆者)	62	缺席					
Q	59	18	-17.13	-30.94	293.27	957.13	529.80
R	55	27	-21.13	-21.94	446.27	481.25	463.43
平均分 (除缺席生)	76.13	48.94	0.00	0.00			
合計值 (除缺值生)	1456.13	831.94	0.00	0.00	2395.75	3954.94	2947.13

上學期考試和下學期考試的相關係數：$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \times \sum \sqrt{(y_i - \bar{y})^2}} = \frac{2947.13}{\sqrt{2393.75} \times \sqrt{3954.94}} = 0.96$

附件二：各種分數估計方法的比較

附件二：各種分數估計方法的比較

方法		缺值生的估計分數*		方法的好處	方法的壞處
		學生A	學生P		
1	以缺值生的上學期考試分數作取代	100	62	此方法無需複雜的計算，所以一般師生較易理解	此方法未有考慮兩次考試的深淺之別，若下學期考試較深，缺值生可以不勞而獲地得到相對較高的分數，這對其他考生不公平。
2	以上下學期的名次排序作估計	74	18		• 此方法的估計分數只來自一個有相同名次的學生，未有考慮其他能力相約學生的情況。 • 如果缺值學生太多，能力最弱的一批學生的估計分數只能以下學期考試中最低的分數取代。
3	以上下學期的標準分作估計	76.61	31.49	此方法有考慮兩次考試的深淺之別，並以缺值生與全班平均分之相對差距作估計。	此方法涉及複雜的計算，所以一般師生較難理解。
4	以「線性迴歸分析」作估計	78.33	31.54	• 此方法有考慮兩次考試的深淺之別，並以全班學生的成績作線性迴歸分析分析。 • 「迴歸線」的圖像化表達可令一般師生較易接受。	

*表中的估計分數未以下學期考試的上四分三位數為上限

參考資料：

1. 高橋信(2010)。世界第一簡單統計學 迴歸分析篇。世茂出版有限公司。
2. 今野紀雄(2010)。3 小時讀通統計(漫畫版)。祥新印刷事業有限公司。
3. 2011/12 中學生統計創意寫作比賽作品集。課程發展處數學教育組(教育局)。

季軍作品：赤壁前傳

學校名稱: 順利天主教中學
學生姓名: 陳天南、劉珈余、曹義銘
級別: 中五
指導教師: 陳靜儀

引言

海戰棋 (Battleships)，西方聞名於世的戰略遊戲；赤壁，中國家喻戶曉的戰場，千古風流人物曾於這裡決一死戰；二龍爭戰決雌雄，赤壁樓船掃地空。中西交匯，一眾三國英雄又能否激起一場概率之爭？本篇將簡單的概率原理融入故事，讓讀者明白概率並非抽象理論，而是生活化的遊戲。

曹操攻下荊州後，欲乘勝追擊，攻打孫權。更連日加緊訓練水軍，準備與東吳決一死戰...

----------東吳水寨(周瑜帳下)----------

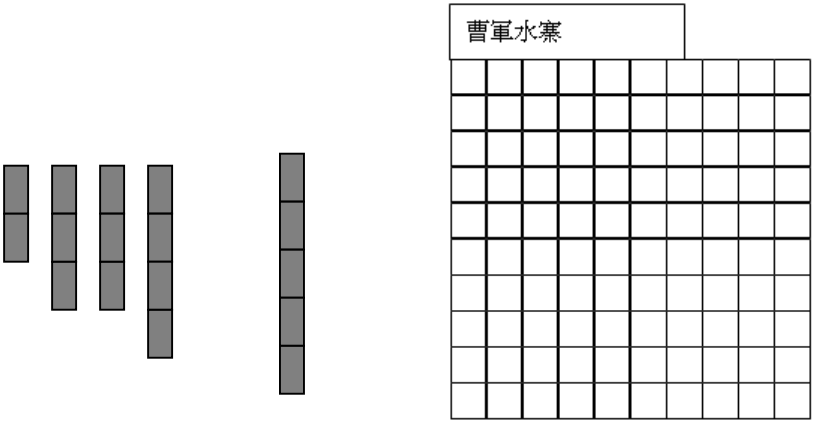
周瑜: 「曹賊要來攻打東吳，實在欺人太甚! 就讓他嘗嘗我們江東水師的厲害，把曹賊殺得片甲不留!」

魯肅: 「都督不用急，還是先等待探子回來，再作打算。」

「探子回報!」

周瑜: 「真是一說曹... 不!我豈能用曹賊名字?... 一說探子， 探子就到。來，快講曹賊水寨的情況! 」

探子: 「小人按曹軍水寨概況，繪畫地圖一幅 :」



探子: 「曹軍水寨可化簡為一個 10 × 10 的正方形， 而曹

軍船隻共分為四種， 其中一種大約為 5 個單位的長度... 」

周瑜: 「這必是曹賊的主船!看我如何將你打敗... 」

魯肅: 「都督還是先讓探子說完吧。」

探子: 「還有一種長 4 個單位的鬥艦，想必是敵方大將張遼的指揮艦。此外， 還有兩艘長 3 個單位的戰船，與一艘長 2 個單位的補給船。」

周瑜: 「嗯... 曹賊水師的船艦數目與我軍相同，而水寨的面積也沒多大差別。但受長江水流所影響，船隻只能縱向或橫向排列。來人，給我傳喚丹陽都尉黃蓋!」

----------不久----------

黃蓋: 「未知都督傳喚末將所謂何事?」

周瑜: 「我們不久後要與曹賊交戰，未知我軍準備如何?」

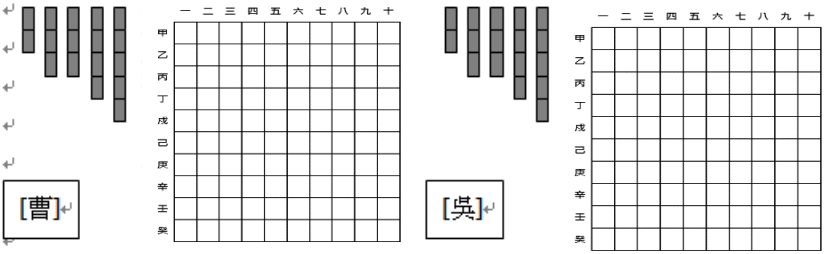
黃蓋: 「近日長江水域一帶異常大霧，若是交戰，雙方皆不能看清對方排陣。我軍不宜與敵人硬碰，以弓弩發箭攻擊曹軍方為上策。」

周瑜: 「我軍現時弓箭數目不多，為免箭矢太快用盡，應每次只向一點攻擊，若曹軍船隻火光烘烘，表示擊中，則乘

勝追擊，不然則先行暫停攻擊，商量下一步對策。」

魯肅: 「將軍所言有理，讓我先把資訊加以整理。」

我軍與曹軍的軍備如下:



1) 雙方皆不能得知對方排列
2) 雙方船隻只能縱向或橫向排列
3) 雙方每次只能攻擊一格位置
4) 若擊中對方船隻，則可獲得多一次攻擊機會，否則讓對方攻擊
5) 攻擊船要擊中對方船隻所有方格才能擊沉對方
6) 先殲滅對方所有船隻者為勝

周瑜: 「有勞子敬整理。究竟我們應如何以最少時間及傷亡擊潰曹賊 ……」

周瑜連日廢寢忘餐，思考擊敗曹軍的計策，最終體力不支，吐血暈倒，臥病不起。此時…

「都督的難題，就由亮去解決吧!」門外傳出一把聲音，正

是諸葛孔明。

魯肅: 「未知孔明有何良計?」

諸葛亮: 「亮近日醉心研究一套命為「概率」的西洋學說，提及戰役必勝之術，定能助都督擬定破敵良策。」

周瑜: 「請先生賜教。」

諸葛亮: 「概率 P(E)，也稱為命中率，是以一個介乎 0 至 1 之間的數字來表示一件事件發生的可能性。P(E)等於 0 表示該事件無可能發生，而 1 則代表該事件必然會發生。」

魯肅:「因此，我軍須盡可能提高每格攻擊位置的命中率。」

諸葛亮: 「對的。而 P(E)的計算方法為（有利結果的數目/可能結果的數目）。」

魯肅: 「由於五首敵船所佔的格數總和為 17 格(5+4+3+3+2)，每一次進攻時每格的命中率該是 $\frac{17}{100}$ 吧!」

周瑜: 「此言差矣。子敬的計算是假設每次攻擊皆為獨立事件，事實上第二次攻擊的命中率 P(F)會受第一次攻擊成功與

否而變動。假如第一次攻擊成功，P(F)為$\frac{16}{99}$；如落空，P(F)即為$\frac{17}{99}$。」

諸葛亮: 「在概率中，當一件事發生與否會對另一件事的概率帶來變動，謂之相關事件。唯都督的計算也有漏洞，由於曹軍中了龐統的連環計，部份船隻的長度多於一單位。以船身佔兩格的補給船為例，假如甲一未能擊中目標，該船的位置必然不是(甲一、甲二)或(甲一、乙一)，故甲二及乙一的命中率會較其他格低。」

周瑜: 「此言有理。我軍如相信第二次的進攻地點選項不受前一次攻擊地點影響，盲目地依照甲一、甲二、甲三的次序攻擊，便中了曹賊的奸計。」

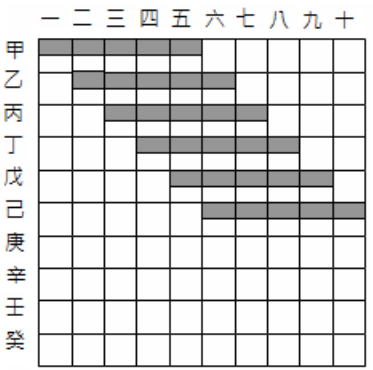
周瑜沉思: 「我定要謹慎地選擇發箭的地點，以一舉殲滅曹軍。」

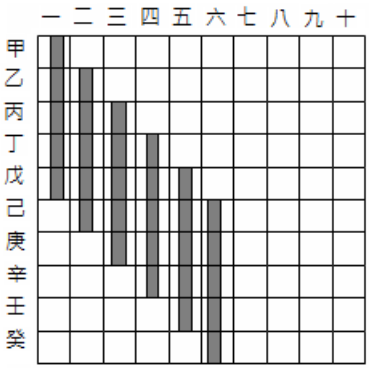
兵法一：先聲奪人

諸葛亮: 「讓我們就對曹軍佈局加以分析。由於 5 個單位的船隻極有可能是曹軍主船，為提高士氣，應先以該船為攻擊對象。」

諸葛亮: 「我軍應以猜測敵軍主船的**排列方法**為目標，故計算命中率方法為(包括該方格的排列方法/可能的排列方法)。

假設 100 格中只有此船，一個橫行及縱行各有 6 個排列方法，故可能排列共有 10×6 + 10× 6 = 120 個。」

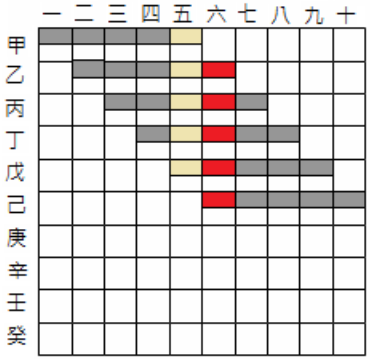


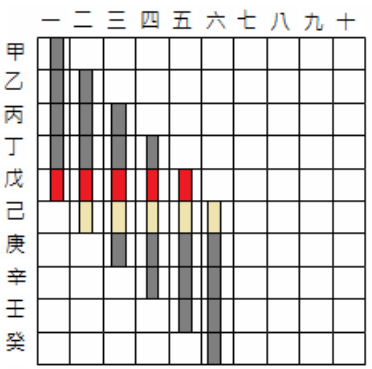


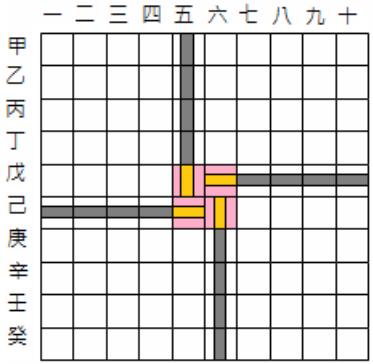
角位(甲一、甲十、癸一、癸十)

周瑜: 「這樣，以甲一為例，包括該方格在內的排列方法有二 (橫行: 甲一至甲五及縱行:甲一至戊一)，故命中率為 $\frac{2}{120}$，即 $\frac{1}{60}$，其他角位的命中率也一樣。」

諸葛亮: 「我們可以發現… 」







中央方格(戊五、戊六、己五、己六)

周瑜: 「若把船放在縱行或橫行的中央， 被箭矢擊中的機會率最高?」

諸葛亮: 「不錯，包括中央方格在內的各有 10 個排列方法，命中率為$\frac{10}{120}$，即$\frac{1}{12}$。」

魯肅: 「那豈不是比角位的命中率的 5 倍嗎?」

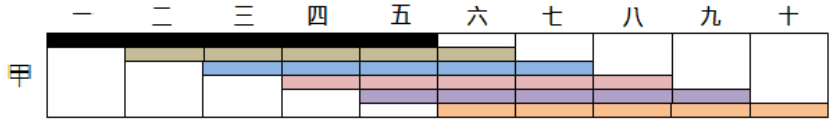
周瑜: 「為先發制人，我們的第一擊就攻擊戊五吧！」

魯肅: 「在下還有一事請教，當我方水師擊中了敵船，下一步該攻擊其上、下，還是左右？」

周瑜: 「這…」

兵法二：乘勝追擊

諸葛亮: 「這就視乎我方擊中的位置。為了方便說明，現假設所有船只能橫放，整個甲列便有六個排列方法。」



諸葛亮: 「如擊中的位置是甲一，繼而攻擊鄰近的格子(甲二)，由於排列方法只有一個，因而命中率為 1，以下列表寫出其他的可能性。」

擊中地點	繼而攻擊地點	排列方法	命中率
甲一	甲二	1	1/1
甲二	甲一	2	1/2
甲二	甲三	2	2/2
甲三	甲二	3	2/3
甲三	甲四	3	3/3
甲四	甲三	4	3/4
甲四	甲五	4	4/4
甲五	甲四	5	4/5
甲五	甲六	5	4/5

諸葛亮: 「綜合上述數據，並加入船隻可橫放及縱放的思量，可歸納出一個更完整的表格去分析擊中地點和最近的邊的距離與其命中率的關係。」

擊中地點和 最近的邊的距離	選擇攻擊與邊距離 較短的格子	選擇攻擊與邊距離 較遠的格子
0	/	1/2
1	1/4	1/2
2	1/3	1/2
3	3/8	1/2
4	2/5	2/5

註: 甲一和邊相離 0 格而甲二和邊相離 1 格

魯肅: 「我明白了！為提高命中率，我們應攻擊與邊距離較遠的格子，若擊中地點和最近的邊的距離是 4，則可隨機攻擊鄰近的格子。」

諸葛亮: 「補充一點，不同船隻的概率應用會稍有出入，當擊中地點和最近的邊的距離為(船身長度−1)便可隨機攻擊，但攻擊與邊距離較遠的格子肯定較有利。」

周瑜: 「若然我軍在第一次攻勢未能... 」

諸葛亮: 「我想都督的疑問是未能擊中曹軍時的對策吧? 」

周瑜: 「嗚... (竟被他看穿了?!)」

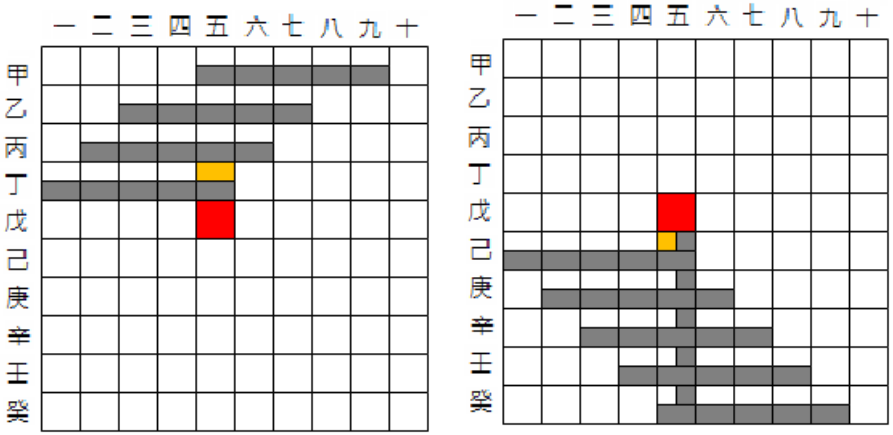
兵法三：屢敗屢戰

諸葛亮: 「先考慮未能在中央方格擊中主船的情況，以戊五一格作解說:」

周瑜:「我軍應否於下一輪攻擊鄰近的方格?」

諸葛亮:「萬萬不可!如下一輪攻擊上方(丁五)，由於所有縱向排列已被排除，只剩橫向的 4 個排列，同樣攻擊左方(戊四)也有 4 個排列。」

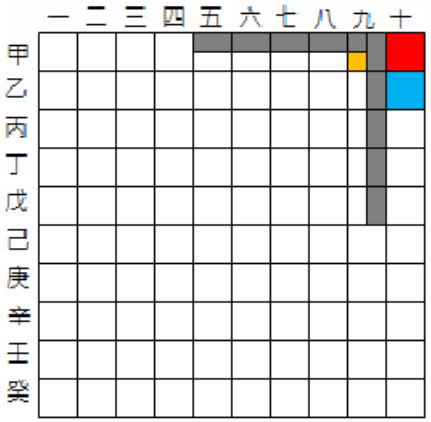
魯肅:「同樣道理，如下一輪攻擊下方(己五)，剩餘 1 個縱向排列及 5 個橫向排列，即剩 6 個排列方法:」

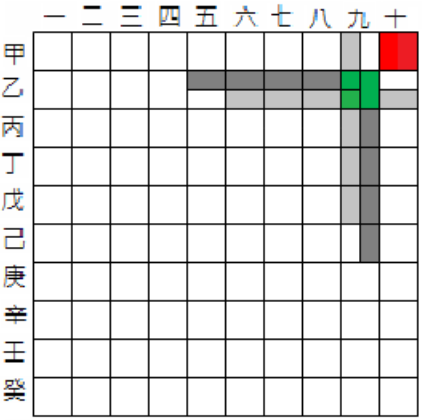


周瑜:「而同樣的排列可能也適用於右方(戊六)。也就是說，由於第一次落空會排除了鄰近方格的排列可能，故攻擊鄰

近方格並非好的戰略?」

諸葛亮:「都督別急下定論，再考慮於四隻角未能擊中的情況，以甲十作解說: 」





諸葛亮: 「可得知甲九左邊(橙色)及乙十下方(藍色)的可行排列各有兩個。」

魯肅: 「但乙九左下方(綠色)則有四個排列！」

周瑜: 「因此，當第一箭未能擊中船時，第二箭射向其斜格為上策。」

魯肅: 「都督，讓我們以此原則制定戰術...」

========== 一個時辰後 ==========

周瑜: 「好！ 黃蓋，快令各船水兵熟練此戰法，我們勝利的日子近啦!」

黃蓋: 「末將領命!」

(2497 字)

參考資料:

1. The Linear Theory of Battleship
   http://thevirtuosi.blogspot.hk/2011/10/linear-theory-of-battleship.html

2. Battleship: Its Value is in the Playing of the Game (Photo)
   http://www.worthpoint.com/blog-entry/battleship-value-playing-game

3. 戰艦遊戲
   http://www.i-gamer.net/play/545.html

4. 海戰棋
   http://taiwanpedia.culture.tw/web/content?ID=24211

5. 三國演義
   http://www.angelibrary.com/oldies/sango/sanguo.html

6. 概率教學網
   http://www.puiching.edu.hk/~maths/student/project/calproject/2001/probability/

優異作品：別讓港鐵的複雜數字欺騙到你

— 票價研究

學校名稱: 香港培正中學
學生姓名: 老綽禧、謝卓熙
指導教師: 梁偉雄









引言

兩鐵合併已經超過七年，但是至今我們仍不難發現票價機制上有着不少潛在問題。我們是次會研究相近港鐵站的問題並嘗試找出票價的計算方法。

(在九龍塘站)
允行: 我們乘搭港鐵回校，應該在哪站下車呢? 旺角站好嗎?

念慈: 旺角東站好一些吧，我們看看誰快一點!

(在香港培正中學)
允行: 竟然你比我快! 如果不是因為扣減了$3.7 導致要增值，我一定比你快!

念慈: 你是說港鐵的費用是$3.7 嗎? 我只用付$3.4 呀!

允行: 真的嗎? 我好像也聽說過在旺角和旺角東的車費有差別。兩站距離不大，但是車費往大部分車站都不一致，且價錢分別很大。

念慈: 對呀! 類似情況也出現在荃灣和荃灣西站。在兩鐵合併後，港鐵公司沒有仔細研究票價，導致不同的問題，例如前往落馬洲在上水出閘再入閘較划算；在荃灣西站前往紅磡的成人車資較前往尖東高，但小童車資相反。以下我們先行研究兩站前往其餘各站的票價。由於港鐵站眾多，我們無法比較所有港鐵站，於是嘗試選取以下的港鐵站作出比較，以盡量覆蓋整個網絡。

允行：根據下表，首兩行分別為旺角或旺角東站前往各站的車資。而第三行我們計算旺角東較旺角站的百分比改變。由於我們是為了比較旺角和旺角東的差異，所以我們可以過海的站也計算。計算此行的平均值，得出 2.4%，因此整體來說兩站票價也相差不多! 比較荃灣站和荃灣西站，平均荃灣西站較荃灣站貴 1.1%，不是太嚴重啦! 因此我們就不用執着吧!

	天水圍	東涌	美孚	九龍	尖東	金鐘	荃灣	荃灣西	九龍塘	大學	沙田圍	牛頭角	將軍澳	北角	柴灣	上水	烏溪沙
旺角	17	15.2	6	4	4.9	10.7	7.4	7.5	4.9	8.9	8	7.4	9	11	13	11.1	9.7
旺角東	17.1	18.2	7.2	6.3	4.9	10.4	8.2	7.6	3.4	6.5	6.7	6.8	9.8	13	14	9.1	9.1
相差%	0.6	19.7	20.0	57.5	0.0	-2.8	10.8	1.3	-30.6	-27.0	-16.3	-8.1	8.9	21.5	9.4	-18.0	-6.2
|相差%|	0.6	19.7	20.0	57.5	0.0	2.8	10.8	1.3	30.6	27.0	16.3	8.1	8.9	21.5	9.4	18.0	6.2

平均票價相差百分比(%)	2.4
票價相差百分比的標準差	21
平均票價相差百分比(%)的絕對值	15.2

	天水圍	東涌	美孚	九龍	尖東	金鐘	旺角	旺角東	九龍塘	大學	沙田圍	牛頭角	將軍澳	北角	柴灣	上水	烏溪沙
荃灣	18	12.8	6	9	9	12.7	7.4	8.2	7.4	10.8	9.9	9	10.7	13	13	14.2	12.8
荃灣西	12.1	16.7	5.8	8.4	9	14.6	7.5	7.6	7.5	10.1	9.6	10.3	10.5	15	15	12.3	11.7
相差%	-32.8	30.5	-3.3	-6.7	0.0	15.0	1.4	-7.3	1.4	-6.5	-3.0	14.4	-1.9	19.7	19.7	-13.4	-8.6
|相差%|	32.8	30.5	3.3	6.7	0.0	15.0	1.4	7.3	1.4	6.5	3.0	14.4	1.9	19.7	19.7	13.4	8.6

平均票價相差百分比;(%)	1.1
票價相差百分比的標準差	15
平均票價相差百分比;(%)的絕對值	10.9

念慈: 你就錯了。看看旺角和旺角東站前往九龍，百分比竟高達 57.5%，相反前往九龍塘卻只-30.6%，計算平均值的話，這些極大值或極小值會互相抵消，令平均值接近 0，造成票價相差不遠的假象。於是我們要計算這些百分比的標準差，以觀察數據分散程度。以旺角和旺角東站為例：

$$\sqrt{\frac{(0.6-2.4)^2+(19.7-2.4)^2+(20-2.4)^2+...+(-18-2.4)^2+(-6.2-2.4)^2}{17-1}} \approx 21$$

而荃灣和荃灣西站為 15，兩者數值都是大的，表示數據分

佈極不平均，有較多極端值。但標準差並不能有效協助我們找出兩站車資問題，我們可以使用絕對值的平均，不理會是哪一個站的票價較高，着重較貴與較便宜的百分比。我們計算第四行(相差百分比的絕對值)的平均，以荃灣和荃灣西站為例：

$$\frac{32.8+30.5+3.3+6.7+0+15+...+14.4+1.9+19.7+19.7+13.4+8.6}{17}\approx 10.9\%$$

而 10.9%及 15.2%皆屬大數值，雖然兩站相差不遠，但礙於兩鐵合併，竟導致票價有如此大差別。

票價	荃灣	大窩口	葵興	葵芳	荔景	美孚	荔枝角	長沙灣	深水埗	太子	旺角	油麻地	佐敦	尖沙咀
荃灣		4	4	4.9	4.9	6	6	7.4	7.4	7.4	7.4	9	9	9
大窩口			4	4	4.9	4.9	6	6	7.4	7.4	7.4	9	9	9
葵興				4	4	4.9	4.9	6	6	7.4	7.4	9	9	9
葵芳					4	4	4.9	4.9	6	6	7.4	7.4	9	9
荔景						4	4	4.9	4.9	6	6	7.4	7.4	9
美孚							4	4	4.9	4.9	6	6	7.2	7.2
荔枝角								4	4	4.9	4.9	6	6	7.2
長沙灣									4	4	4.9	4.9	6	6
深水埗										4	4	4.9	4.9	6
太子											4	4	4.9	4.9
旺角												4	4	4.9
油麻地													4	4
佐敦														4
尖沙咀														

允行：提到港鐵的票價問題，為何不嘗試去尋找港鐵的票價機制呢? 讓我們由時間、站數和距離入手吧! 以荃灣線為例，首先找出每個站前往其他站的車資。

念慈：找到行車時間了! 我真聰明。

時間	荃灣	大窩口	葵興	葵芳	荔景	美孚	荔枝角	長沙灣	深水埗	太子	旺角	油麻地	佐敦	尖沙咀
荃灣		3	5	7	9	12	14	16	17	20	21	23	25	27
大窩口			3	5	7	10	12	14	16	18	20	22	24	25
葵興				3	5	8	10	12	14	16	17	19	21	23
葵芳					3	6	8	10	12	14	16	18	20	21
荔景						4	6	8	10	12	13	15	17	19
美孚							3	5	7	9	11	13	15	16
荔枝角								3	5	7	9	11	13	14
長沙灣									3	5	7	9	11	12
深水埗										3	5	7	9	11
太子											3	4	7	8
旺角												3	5	7
油麻地													3	5
佐敦														3
尖沙咀														

允行：知道你聰明了，還是說正事吧! 有了票價和時間，我們能找出在荃灣線乘搭列車平均每分鐘需多少元。看看我的傑作!

平均每分鐘$	荃灣	大窩口	葵興	葵芳	荔景	美孚	荔枝角	長沙灣	深水埗	太子	旺角	油麻地	佐敦	尖沙咀
荃灣		1.33	0.80	0.70	0.54	0.50	0.43	0.46	0.44	0.37	0.35	0.39	0.36	0.33
大窩口			1.33	0.80	0.70	0.49	0.50	0.43	0.46	0.41	0.37	0.41	0.38	0.36
葵興				1.33	0.80	0.61	0.49	0.50	0.43	0.46	0.44	0.47	0.43	0.39
葵芳					1.33	0.67	0.61	0.49	0.50	0.43	0.46	0.41	0.45	0.43
荔景						1.00	0.67	0.61	0.49	0.50	0.46	0.49	0.44	0.47
美孚							1.33	0.80	0.70	0.54	0.55	0.46	0.48	0.45
荔枝角								1.33	0.80	0.70	0.54	0.55	0.46	0.51
長沙灣									1.33	0.80	0.70	0.54	0.55	0.50
深水埗										1.33	0.80	0.70	0.54	0.55
太子											1.33	1.00	0.70	0.61
旺角												1.33	0.80	0.70
油麻地													1.33	0.80
佐敦														1.33
尖沙咀														
總平均	0.65													

念慈：你是怎樣算出平均每分鐘票價呢?

允行：還說自己聰明。以荃灣到油麻地為例，平均每分鐘

票價等於荃灣到油麻地的票價除以行車時間，因為行車時間(不包括候車時間)可以更準確估計費用，即 $\frac{9}{23} \approx 0.3913...$，為方便計算會取近似值至兩位小數。而總平均則是所有計算出來的平均值相加再除以其數量。我們可得出荃灣線的票價平均每 1 分鐘需付$ 0.65，跟我好好學習吧。

念慈：我早就做好了。我使用類似的方法，找出荃灣線平均每搭乘一站，票價為$1.66。

平均每站$	荃灣	大窩口	葵興	葵芳	荔景	美孚	荔枝角	長沙灣	深水埗	太子	旺角	油麻地	佐敦	尖沙咀
荃灣		4.00	2.00	1.63	1.23	1.20	1.00	1.06	0.93	0.82	0.74	0.82	0.75	0.69
大窩口			4.00	2.00	1.63	1.23	1.20	1.00	1.06	0.93	0.82	0.90	0.82	0.75
葵興				4.00	2.00	1.63	1.23	1.20	1.00	1.06	0.93	1.00	0.90	0.82
葵芳					4.00	2.00	1.63	1.23	1.20	1.00	1.06	0.93	1.00	0.90
荔景						4.00	2.00	1.63	1.23	1.20	1.00	1.06	0.93	1.00
美孚							4.00	2.00	1.63	1.23	1.20	1.00	1.03	0.90
荔枝角								4.00	2.00	1.63	1.23	1.20	1.00	1.03
長沙灣									4.00	2.00	1.63	1.23	1.20	1.00
深水埗										4.00	2.00	1.63	1.23	1.20
太子											4.00	2.00	1.63	1.23
旺角												4.00	2.00	1.63
油麻地													4.00	2.00
佐敦														4.00
尖沙咀														
總平均	1.66													

允行：我不能落後太多，我找出兩站的直線距離(km)並計算每一公里需付$1.95。

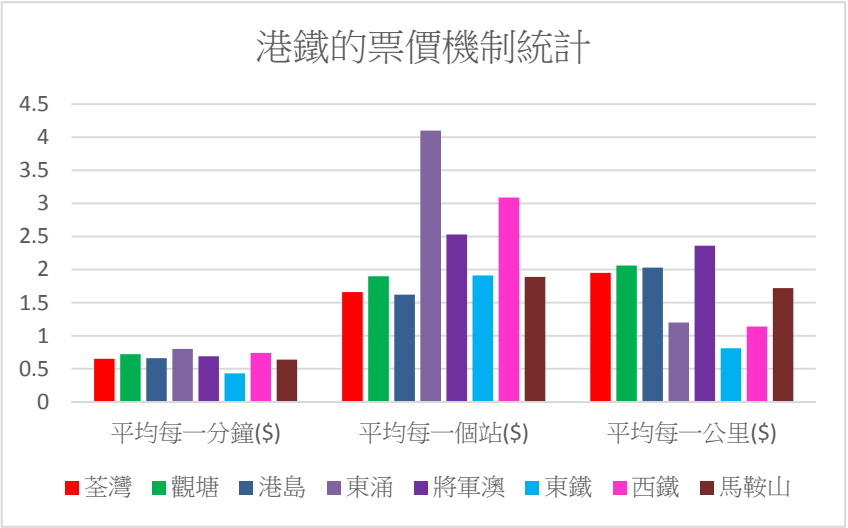
距離	荃灣	大窩口	葵興	葵芳	荔景	美孚	荔枝角	長沙灣	深水埗	太子	旺角	油麻地	佐敦	尖沙咀
荃灣		0.81	1.82	2.14	2.94	4.43	5.11	5.78	6.56	7.53	8.05	8.69	9.46	10.15
大窩口			1.07	1.59	2.49	3.86	4.42	5.05	5.83	6.79	7.33	7.98	8.78	9.48
葵興				0.79	1.71	2.87	3.35	3.98	4.76	5.72	6.26	6.91	7.71	8.42
葵芳					0.94	2.29	3.01	3.75	4.53	5.49	5.98	6.59	7.34	8.02
荔景						1.59	2.57	3.39	4.15	5.07	5.51	6.06	6.75	7.38
美孚							1.2	2.03	2.73	3.59	3.97	4.49	5.16	5.79
荔枝角								0.84	1.58	2.51	2.97	3.58	4.36	5.07
長沙灣									0.79	1.75	2.28	2.95	3.79	4.55
深水埗										0.96	1.52	2.22	3.09	3.87
太子											0.62	1.35	2.25	3.05
旺角												0.73	1.63	2.43
油麻地													0.9	1.7
佐敦														0.8
尖沙咀														

平均每公里$	荃灣	大窩口	葵興	葵芳	荔景	美孚	荔枝角	長沙灣	深水埗	太子	旺角	油麻地	佐敦	尖沙咀
荃灣		4.94	2.20	2.29	1.67	1.35	1.17	1.28	1.13	0.98	0.92	1.04	0.95	0.89
大窩口			3.74	2.52	1.97	1.27	1.36	1.19	1.27	1.09	1.01	1.13	1.03	0.95
葵興				5.06	2.34	1.71	1.46	1.51	1.26	1.29	1.18	1.30	1.17	1.07
葵芳					4.26	1.75	1.63	1.31	1.32	1.09	1.24	1.12	1.23	1.12
荔景						2.52	1.56	1.45	1.18	1.18	1.09	1.22	1.10	1.22
美孚							3.33	1.97	1.79	1.36	1.51	1.34	1.40	1.24
荔枝角								4.76	2.53	1.95	1.65	1.68	1.38	1.42
長沙灣									5.06	2.29	2.15	1.66	1.58	1.32
深水埗										4.17	2.63	2.21	1.59	1.55
太子											6.45	2.96	2.18	1.61
旺角												5.48	2.45	2.02
油麻地													4.44	2.35
佐敦														5.00
尖沙咀														
總平均	1.95													

念慈：你是使用直線距離嗎? 有時候如果我們使用行車距離的話會造成誤差，因為有部份路線雖然直線距離小，但行車距離大。真正的票價機制使用直線距離會較合理。我已經準備好各市區線的數據，就如下表。我們也可發現使用時間計算的結果平均，以站數計算，東鐵、東涌，西鐵線因距離遠，便會較便宜。請不要忽略，我分開各線計算，部份線段因重複而只納入其中一線。同時我認為過海、馬

場、羅湖和落馬洲因有額外收費而沒有計算(附錄一)。

	荃灣	觀塘	港島	東涌	將軍澳	東鐵	西鐵	馬鞍山
平均每一分鐘($)	0.65	0.72	0.66	0.8	0.69	0.43	0.74	0.64
平均每一個站($)	1.66	1.90	1.62	4.1	2.53	1.91	3.09	1.89
平均每一公里($)	1.95	2.06	2.03	1.2	2.36	0.81	1.14	1.72



允行：謝! 不過更詳細去估計整個網絡的票價機制，我們可以使用期望值的概念去比較，先從時間去計算，我們分別使用整段路程的所需時間：

$$\frac{0.65\times 27+0.72\times 20+0.66\times 26+0.8\times 27+0.63\times 14+0.43\times 39+0.74\times 38+0.64\times 18}{27+20+26+27+14+39+38+18}$$

≈ 0.65 (2 dp)

因此平均每乘搭 1 分鐘便需付$0.65，而以下是站數的估

計，我們以各線所包含的港鐵站數量來計算期望值，得出每 1 個站付$2.22:

$$\frac{1.66\times14+1.9\times10+1.62\times14+4.1\times7+2.53\times6+1.91\times11+3.09\times12+1.89\times9}{14+10+14+7+6+11+12+9}$$

$\approx \$2.22$ (2 dp)

念慈：明白，最後分別以各線的行車距離來計算期望值：

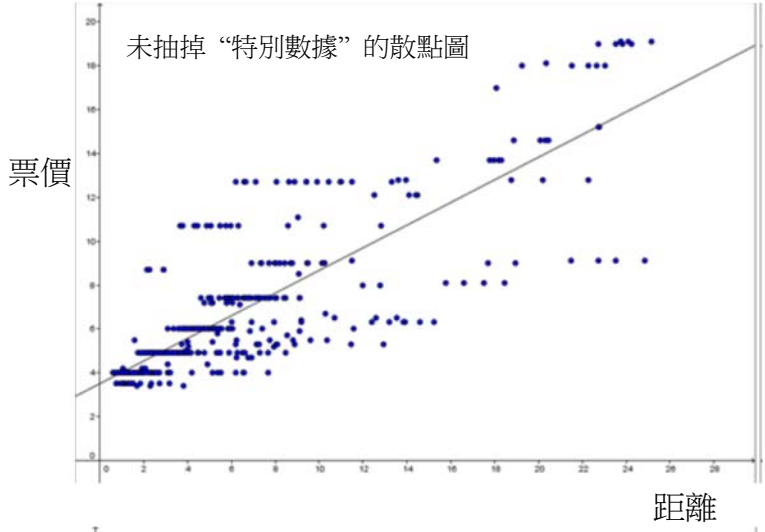
$$\frac{1.95\times14+2.06\times9.7+2.03\times13.3+1.2\times29.1+2.36\times9.9+0.81\times39+1.14\times35.4+1.72\times11.4}{14+9.7+13.3+29.1+9.9+39.1+35.4+11.4}$$

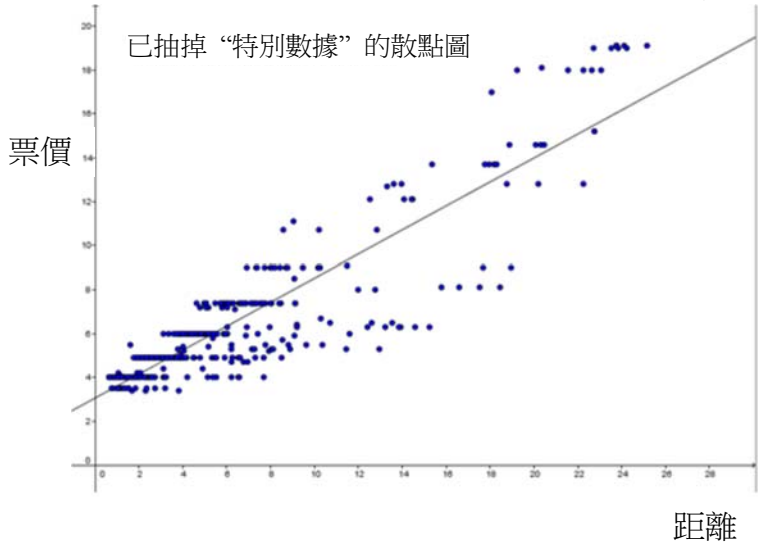
$\approx \$1.38$ (2 dp)

我們就能得出平均每 1km 就需$1.38。另外使用期望值的方法可以減少對整個網絡各線之間估計票價的誤差。由於整個系統有多達 75 個站，我們無法逐一統計，就嘗試分各線統計再計算加權平均數，得出上述 3 個數據，最後找出其平均應為的票價。

允行：你這種方法有少許問題，你代入不同組合的話應該會有頗大誤差。理論上你的方法找出平均應該不大，但港鐵的票價是「短貴長平」的，好像你統計平均時可發現前往表中越右及越上的站會較便宜，票價應是部分常數，而部分則正比於距離或時間或站數。你可以透過散點圖，包括所有的同線組合，找出距離或時間或站數對票價的影響。然而，這時你會發現一些“特別數據”大多是一些票價較貴的特別情況，如過海、馬場和跨境，其票價很大機會是

額外多一個常數。我們可以把這些數據抽掉，使每條路線的票價制度大致正常(票價正比於時間，距離、站數)，才去研究票價機制，用作推測須轉線的組合的真正理論票價。





念慈：以上統計圖得出最適合的線段l_1，其公式為$P = 0.55d + 3.07$，其中 d 代表距離。然後可以使用這 較準確的公式重新研究原始問題─究竟是旺角還是旺角東站的票價不適當。當然我們理應把所有組合放進圖內，但數據繁多，我們只能使用這樣的方法去估計。使用距離來計算是因為大多票價是以距離為大前提，再以時間和站數加以調整，好像從市區到新界，理論上屯門的票價應略低於元朗的票價，但因進入屯門的站數較元朗多，以致屯門和元朗的票價相同。這是常識吧!

允行：說得對! 我們分別將距離代入你的公式，並計算我們的票價與真正票價的分別。很容易發現分佈仍不平均，而且誤差不一，顯示這仍不準確。以下兩圖的形式跟最初的部份格式相似，不多解釋了，我們已減少誤差了。

旺角東站	天水圍	東涌	美孚	九龍	尖東	荃灣	荃灣西	九龍塘	大學	沙田圍	牛頭角	將軍澳	上水	烏溪沙
旺角至該站的票價($)	17.10	18.20	7.20	6.30	4.90	8.20	7.60	3.40	6.50	6.70	6.80	9.80	9.10	9.10
旺角與該站直線距離(km)	22.28	24.10	4.12	2.28	2.99	8.05	8.27	1.69	10.84	6.54	4.82	9.15	20.47	14.00
使用公式的理論票價($)	15.32	16.33	5.34	4.32	4.71	7.50	7.62	4.00	9.03	6.67	5.72	8.10	14.33	10.77
理論票價與實際票價的%差異	11.59	11.49	34.9	45.7	3.93	9.37	-0.24	-14.99	-28.03	0.49	18.86	20.95	-36.49	-15.51
l理論票價與實際票價的%差異l	11.59	11.49	34.9	45.7	3.93	9.37	0.24	14.99	28.03	0.49	18.86	20.95	36.49	15.51
旺角站														
旺角至該站的票價($)	17.00	15.20	6.00	4.00	4.90	7.40	7.50	4.90	8.90	8.00	7.40	9.00	11.10	9.70
旺角與該站直線距離(km)	22.22	23.73	3.97	1.85	2.72	8.05	8.21	2.08	11.25	6.96	5.12	9.43	20.70	14.44
使用公式的理論票價($)	15.29	16.12	5.25	4.09	4.57	7.50	7.59	4.21	9.26	6.90	5.89	8.26	14.46	11.01
理論票價與實際票價的%差異	11.18	-5.72	14.2	-2.14	7.31	-1.3	-1.13	16.28	-3.86	15.98	25.72	9.01	-23.21	-11.91
l理論票價與實際票價的%差異l	11.18	5.72	14.2	2.14	7.31	1.3	1.13	16.28	3.86	15.98	25.72	9.01	23.21	11.91

	旺角東	旺角
平均百分比差異	4.43	3.60
百分比的標準差	22.84	13
平均百分比差異(絕對值)	18.04	10.64

荃灣西站	天水圍	東涌	美孚	九龍	尖東	旺角	旺角東	九龍塘	大學	沙田圍	牛頭角	將軍澳	上水	烏溪沙
荃灣西至該站的票價($)	12.1	16.7	5.8	8.4	9	7.5	7.6	7.5	10.1	9.6	10.3	10.5	12.3	11.7
荃灣西與該站直線距離(km)	14.01	19.46	4.37	8.89	10.52	8.21	8.27	7.65	11.45	8.83	12.68	16.89	14.92	15.37
使用公式的理論票價($)	10.78	13.77	5.47	7.96	8.86	7.59	7.62	7.28	9.37	7.93	10.04	12.36	11.28	11.52
理論票價與實際票價的%差異	12.29	21.25	5.97	5.53	1.63	-1.13	-0.24	3.06	7.82	21.113	2.55	-15.05	9.08	1.53
理論票價與實際票價的%差異I	12.29	21.25	5.97	5.53	1.63	1.1	0.24	3.06	7.82	21.11	2.55	15.05	9.08	1.53
荃灣站														
荃灣至該站的票價($)	18	12.8	6	9	9	7.4	8.2	7.4	10.8	9.9	9	10.7	14.2	12.8
荃灣與該站直線距離(km)	14.30	20.46	4.43	8.93	10.5	8.05	8.05	7.24	10.46	7.96	12.26	16.4	14.24	14.37
使用公式的理論票價($)	10.94	14.32	5.51	7.98	8.85	7.50	7.50	7.05	8.82	7.45	9.81	12.09	10.90	10.97
理論票價與實際票價的%差異	64.61	-10.63	8.96	12.76	1.75	-1.3	9.37	4.93	22.41	32.92	-8.28	-11.5	30.25	16.64
理論票價與實際票價的%差異I	64.61	10.63	8.96	12.76	1.75	1.3	9.37	4.93	22.41	32.92	8.28	11.5	30.25	16.64
	荃灣西	荃灣												
平均百分比差異	5.39	12.35												
百分比的標準差	9	21												
平均百分比差異(絕對值)	7.73	16.88												

念慈：說得太好啦! 我們在這個討論，先發現有些組合的票價制度有不妥之處，我們認為港鐵使用平均每 km 的價錢去計算 $(P = kd)$，因為這才對短途公平。我們又使用短貴長平的方式去尋求車資 $(P = k + hd)$。我們一直使用各線的內部組合再推算全系統的所有組合。但是我們代入兩者後，發現也有點參差，如上表荃灣及旺角的平均票價差異，但有些站挺準確(荃灣西至旺角東)，有些則很大誤差。當然整個研究暫不考慮過海及跨境，這些日後再研究。總括而言，港鐵並沒有一個完善的票價制度和準確的票價計算方法，很多時候只有一個原則：搭遠了、搭多了就要貴一點。我們提議港鐵可參考其他城市如首爾 $(P = k + hd)$，建立良好而可供參考的公式。

(2499 字)

參考資料：

1. 香港鐵路有限公司 www.mtr.com.hk/
2. www.distancefromto.net/
3. 維基百科(港鐵) https://zh.wikipedia.org/wiki/港鐵
4. 維基百科(地鐵)http://zh-yue.wikipedia.org/wiki/香港地鐵

使用軟件:

1. Geogebra www.geogebra.org
2. mathtype http://www.dessci.com/en/products/mathtype/

(註: 篇幅所限，附錄一與附錄二於網上版刊登)

優異作品：Please help me to find the quantity!

School Name: Good Hope School
Names of Students: Lo Hau Yan, Lo Suet Ching, Wong Hiu Wai
Supervising Teacher: Mr. Kenneth Tang

Introduction

John, a new member of the EPD staff, finds that the data of the quantities of exported recyclable polystyrene and paper are missing. Being hot-blooded, John is desperate to find out the solution so he contacts an expert for advice. Through calculation, John learns more about Statistics and Mathematics, and finds out different data through a series of methods. Will the newly-found data be accurate or just another big error?

To: chriswong@amail.com.hk
From: johnli@amail.com.hk
Subject: Seeking a helping hand

Dear Mr. Wong,

Good morning. My name is John Li, a new officer in the Environmental Protection Department. I am currently working under Edward Chong, and he has requested me to do a report about the monitoring of solid waste in Hong Kong. It is required to use the data collected by the department.

As I am still new to the post, Mr. Chong introduced you to me, saying that you are an expert in this area and might be able to give some guidance and advice. Would you please help me when I come across problems?

Yours sincerely,
John Li

To: johnli@amail.com.hk From: chriswong@amail.com.hk Subject: Nice to meet you
Dear John, Good afternoon and nice to meet you. I am surprised that Edward would introduce me to you. And of course, you can find me at any time. We may continue to communicate through emails. Best wishes, Chris

To: chriswong@amail.com.hk From: johnli@amail.com.hk Subject: Problems
Dear Chris, Hello! I am doing analysis about the data of the quantities of exported recyclable materials by type. However, I realised there are some problems: the quantities of tin, glass and paper in 2009 and polyethylene and polystyrene and copolymers in 2010 are missing. Could you please kindly tell me why? Regards, John

To: johnli@amail.com.hk
From: chriswong@amail.com.hk
Subject: Quick answer

Dear John,

Hi! It's normal for the data of tin and glass to be missing, since we didn't export any of these. But for paper and polyethylene and polystyrene & copolymers, it is not right. Data should be collected. As I recall, the computer system had a malfunction when we were inputting the data. Perhaps it was lost at that time.

Best wishes,
Chris

To: chriswong@amail.com.hk
From: johnli@amail.com.hk
Subject: Resolution

Dear Chris,

Hi! Since the lost data of polyethylene is handled by other colleagues, I just need to mainly focus on paper and polystyrene & copolymers. Can I use the median to substitute the missing data? The median is a numerical value which separates the higher half of the data from the lower half. The median can be found by listing the data in ascending order. If

there is an odd number of terms, we can simply obtain the median by getting the middle value, but if there is an even number of terms, we take the mean of the two middle values, thus the median is obtained.

I arranged the records of paper and polystyrene & copolymers respectively in ascending order of quantity.

Paper:

Year	Quantity	Year	Quantity
1998	392157	2005	792458
1997	439831	2006	934041
1999	533741	2008	1091196
2002	592830	2007	1101969
2003	633307	2012	1162294
2000	644061	2010	1194535
2001	657336	2011	1278366
2004	731446	2009	

We have 15 data items in total, so we take the eighth number as the median. We can thus substitute it into the missing data.
The figure for paper in 2009 is now 731446.

Polystyrene and copolymers:

Year	Quantity	Year	Quantity
2012	1411	1999	30100
2003	5693	2011	34192
2004	11784	1997	47950
2001	18445	2009	48562
2002	18799	2007	54397
2006	18846	1998	58634
2000	25498	2008	89167
2005	27960	2010	

We also take the eighth number and substitute it into the table, replacing the missing data.

So, the missing figure for polystyrene & copolymers in 2010 is 27960.

Am I right?

Yours truly,
John

To: johnli@amail.com.hk
From: chriswong@amail.com.hk
Subject: Another way

Dear John,

This could be one of the ways. However, this method does not take all data into account. Would the median be unsuitable in this case?

To make it more accurate, you may try other methods.

Best wishes,
Chris

To: chriswong@amail.com.hk
From: johnli@amail.com.hk
Subject: Problem solved! (Resolution 2)

Dear Chris,

You are right! But actually, we have just found out the missing data. Mr. Chong took a look at the records and also discovered that the data was missing. He had already asked a senior member to input it back. He found the answer by using average. He taught me that to average is to divide the sum of the terms by the number of terms.

There is a formula:
$A=S/N$

A stands for average, while S stands for the sum of the terms and N stands for the number of the terms. Let me give you an example. There are 5 numbers, 3, 4, 6, 5 and 2. First, to find out *S*, we have to add them altogether. The sum of these 5 numbers is 20. Now we find *N*. We have a total of 5 numbers, so it is 5. By dividing 20 by 5, we can get *A*, which is 4.

By arranging $A=S/N$ to $A=a+b+c+d+e+\ldots/N$, (*a,b,c,d,e* stands for different terms) and substituting the numbers of the past 3 years into it, I can find out the missing data.

Year	Polystyrene & copolymers	Paper	Year	Polystyrene & copolymers	Paper
1997	47950	439831	2005	27960	792458
1998	58634	392157	2006	18846	934041
1999	30100	533741	2007	54397	1101969
2000	25498	644061	2008	89167	1091196
2001	18445	657336	2009	48562	
2002	18799	592830	2010		1194535
2003	5693	633307	2011	34192	1278366
2004	11784	731446	2012	1411	1162294

Average of paper:

(*439831+392157+533741+644061+657336+592830+633307+731446+792458+934041+1101969+1091196+1194535+1278366+1162294*) */15*

$= \frac{\mathbf{12179568}}{\mathbf{15}}$

=811971 (*tons*)(*to the nearest integer*)

Average of polystyrene and copolymers:

(*47950+58634+30100+25498+18445+18799+5693+11784+27960+18846+54397+89167+48562+34192+1411*) */15*

$= \frac{491438}{15}$

=32763(*tons*)(*to the nearest integer*)

So we substitute the two averages into the table, and the missing figure for paper is now 811971 tons, and that of polystyrene and copolymers is 32763 tons.

Regards,
John

To: johnli@amail.com.hk From: chriswong@amail.com.hk Subject: Wait!
Dear John, Did you notice that average is actually very similar to the median? The average is actually affected by extreme data. They are both inaccurate in this case. It's your show time, young man! Show Edward you are better and more intelligent than the senior staff and give him a good impression! Try to find another way to arrive at an answer. I'll support you! Best wishes, Chris

To: chriswong@amail.com.hk From: johnli@amail.com.hk Subject: Could you help?
Dear Chris, Thank you for your support and encouragement. Nevertheless, to be honest, I'm weak at mathematics. I only know the previous methods. Could please kindly help me one more time? I'd be very GRATEFUL! Best regards, John

To: johnli@amail.com.hk
From: chriswong@amail.com.hk
Subject: Linear Regression

Dear John,

OK, I can help! Have you heard of a resolution called linear regression? It is relatively more complicated, but I think you are smart enough to handle it. This is the formula:

$$y = ax + b$$

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i\right)$$

Yours,
Chris

To: chriswong@amail.com.hk From: johnli@amail.com.hk Subject: I find it!
Dear Chris, I used your method and found out the correlation between paper and polystyrene and copolymers. With the missing data, it is 0.079115312…… Am I correct? Can I carry on the calculation? Best regards, John

To: johnli@amail.com.hk From: chriswong@amail.com.hk Subject: Linear Regression reminder
Dear John, I am sorry, John. I forgot to tell you that the closer the correlation to 1, the more the data are related, i.e. more accurate. I think this is too close to 0 and the error is big. Perhaps you can try out more sets of data? Yours, Chris

To: chriswong@amail.com.hk
From: johnli@amail.com.hk
Subject: I will try

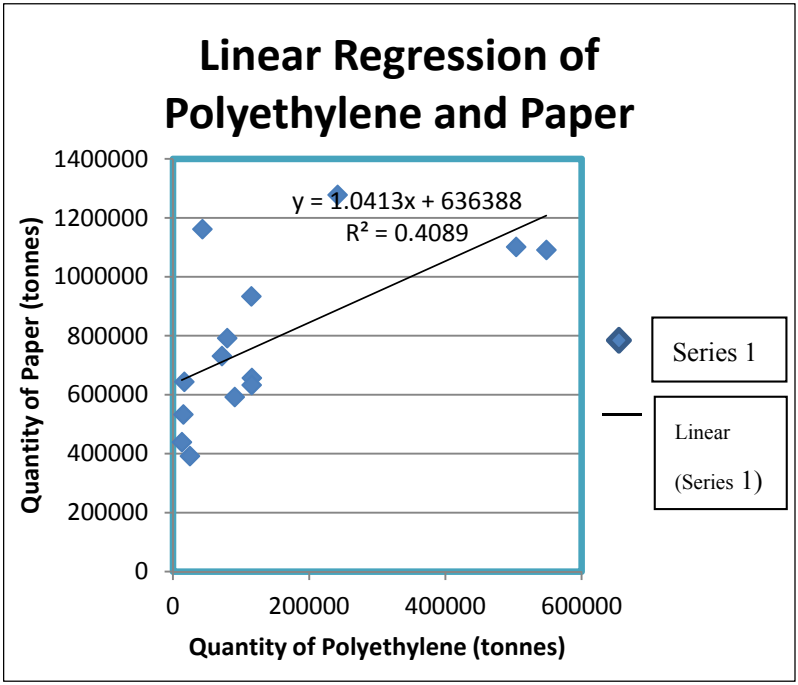
Dear Chris,

Oh my goodness! It sounds this “trial and error” will take a long time! But anyway, with your support, I believe I can make it!

Best regards,
John

To: chriswong@amail.com.hk
From: johnli@amail.com.hk
Subject: Resolution 3 – Linear Regression

Year	Polyethylene	Paper	Year	Polyethylene	Paper
1997	12992	439831	2005	79320	792458
1998	24694	392157	2006	115011	934041
1999	15108	533741	2007	503731	1101969
2000	16476	644061	2008	548064	1091196
2001	115653	657336	2009	333691	
2002	90556	592830	2010		1194535
2003	115438	633307	2011	241442	1278366
2004	71675	731446	2012	43097	1162294



This time, to play safe, I changed the data of polystyrene & copolymers to polyethylene, which has a positive relationship with paper, having a correlation of 0.639415035…… , which is quite good. After finding the data of polyethylene, we can find polystyrene and copolymers with polyethylene with linear regression.

Using Linear Regression to find out the missing data.

$y=ax+b$

To find out *a*, I used the formula you gave to me. To make it simple, I calculated it in Excel, by plotting a scatter diagram, and the equation $y=ax+b$ comes out as $y=1.0414x+636388$ after calculation.

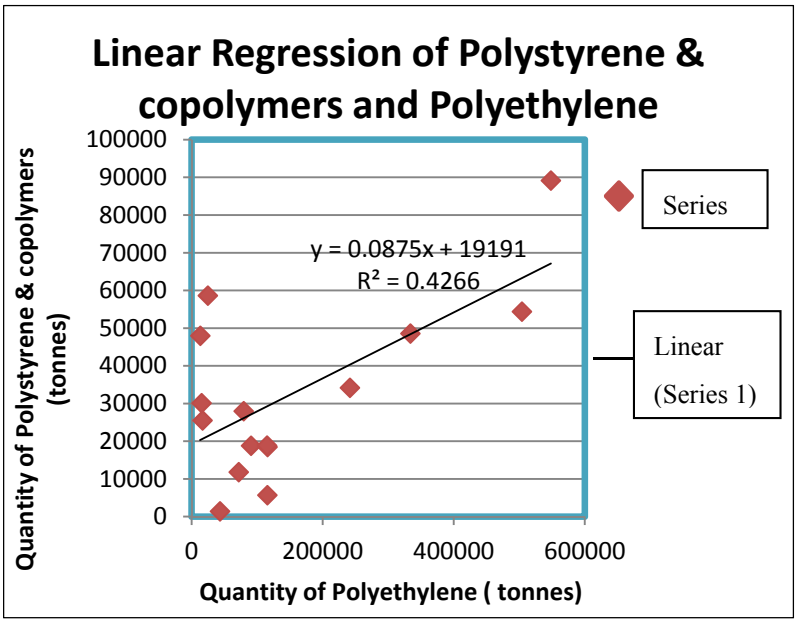
By setting the missing data of polyethylene in 2010 as *x* and the data of paper in 2010 as *y*,

$1194535=1.0413x+636388$
$558147=1.0413x$
$x=536004.3785$
x=536004(tons)(to the nearest integer)

By setting the data of polyethylene in 2009 as *x* and the missing data of paper in 2009 as *y*,
$y=1.0413\times333691+636388$
y=983864(tons)(to the nearest integer)

Therefore, the missing data of paper is 983893 tons and that of polyethylene is 536004 tons.

Year	Polyethylene	Polystyrene & copolymers	Year	Polyethylene	Polystyrene & copolymers
1997	12992	47950	2005	79320	27960
1998	24694	58634	2006	115011	18846
1999	15108	30100	2007	503731	54397
2000	16476	25498	2008	548064	89167
2001	115653	18445	2009	333691	48562
2002	90556	18799	2010	535958	
2003	115438	5693	2011	241442	34192
2004	71675	11784	2012	43097	1411



Using the equation $y=ax+b$,
it comes out as $y = 0.0875x + 19191$ after calculation.

By setting the newly found data of polyethylene in 2010 as x and the data of polystyrene & copolymers in 2010 as y,

$y=0.0875\times535958+19191$
$y=66087$(*tons*)(*to the nearest integer*)

Therefore, the missing data of polystyrene & copolymers in 2010 is 66087 tons.

Am I correct?

Regards,
John

To: johnli@amail.com.hk
From: chriswong@amail.com.hk
Subject: Good Job!

Dear John,

Well done this time! As this method can clearly show the trend of the quantities of the exported recyclable materials, it can help to find a more accurate number. I am sure Edward will be pleased.

Yours,
Chris

To: chriswong@amail.com.hk
From: johnli@amail.com.hk
Subject: Thank You!

Dear Chris,
Thank you for your guidance through out the project! After I calculated the missing data, the IT department also helped us to recover the lost data. Luckily, my answers are quite similar to the original ones and my boss is glad:

Polyethylene: 521,804 tons
Paper:1027229 tons
Polystyrene & copolymers: 89,472 tons

By the way, you also raised my interest in mathematics. I will try to learn more about different kinds of mathematics and of course, try my best to work for the EPD!

Best wishes,
John

(1553 words)

References:

1. https://www.wastereduction.gov.hk/en/materials/info/msw2012.pdf
2. https://www.wastereduction.gov.hk/en/materials/info/msw2011.pdf
3. https://www.wastereduction.gov.hk/en/materials/info/msw2010.pdf
4. https://www.wastereduction.gov.hk/en/materials/info/msw2009.pdf
5. https://www.wastereduction.gov.hk/en/materials/info/msw2008.pdf
6. https://www.wastereduction.gov.hk/en/materials/info/msw2007.pdf
7. https://www.wastereduction.gov.hk/en/materials/info/msw2006.pdf
8. https://www.wastereduction.gov.hk/en/materials/info/msw2005.pdf
9. https://www.wastereduction.gov.hk/en/materials/info/msw2004.pdf
10. https://www.wastereduction.gov.hk/en/materials/info/msw2003.pdf
11. https://www.wastereduction.gov.hk/en/materials/info/msw2002.pdf
12. https://www.wastereduction.gov.hk/en/materials/info/msw2001.pdf
13. https://www.wastereduction.gov.hk/en/materials/info/msw2000.pdf
14. https://www.wastereduction.gov.hk/en/materials/info/msw1999.pdf
15. https://www.wastereduction.gov.hk/en/materials/info/msw1998.pdf
16. https://www.wastereduction.gov.hk/en/materials/info/msw1997.pdf
17. http://en.wikipedia.org/wiki/Simple_linear_regression

優異作品: 假波風雲

學校名稱：東華三院吳祥川紀念中學
學生姓名：莫子聰
級別：中五
指導老師：劉漢昌

引言

在很多統計調查中，受訪者要向訪問員回答一些敏感或令人尷尬的問題。究竟這些調查是怎樣進行呢？如何令受訪者誠實回答？本人透過一個虛構的故事來解釋一個特別的統計的技巧。

香城最近出現「打假波事件」(打假波是指於球類運動中利用不同形式作弊的行為)。早前，一隊球隊於足球比賽中連輸數場 (其中一場涉及故意的擺烏龍入球)，現處於聯賽榜榜尾。該球隊部份球員被懷疑涉嫌打假波，須接受廉潔公署調查。

香城足球總會（下稱足總）的高層為打假波一事連日開會，尋求對策。以下是他們開會的情況。與會者包括黃主席、陳秘書、李幹事和張主任。

黃主席：「一個月前香城康體局局長要我們徹查本地足球員打假波的情況。陳秘書，調查的進展如何？」

陳秘書：「三星期前，我們將問卷放到所有香城足球隊員的儲物櫃中。所有問卷都是不記名的。問卷的目的是要受訪者回答有沒有參與打假波。受訪者填寫問卷後只須用附上的回郵信封寄回我們的郵政信箱就可以。主席，這是我們收回問卷的結果。」

黃主席閱畢問卷結果，輕嘆一聲:「你們自己看看問卷調查結果吧。」

組別	總人數	回應數量	回應率
甲組	5310	477	8.98%
乙組	3650	285	7.81%
丙組	1610	64	3.98%
全體球員	10570	833	7.88%

李幹事看過後，說：「全體球員問卷的回應率只有 7.88%，難怪主席你會愁眉苦臉。」

黃主席：「由於回覆這些問卷屬自願性質，所以回應率很低。」

李幹事：「不如我們派統計員直接訪問球員。我相信這樣有助提升回應率。」

陳秘書：「我相信如果直接向球員查詢，大部份受訪者都會拒絕回答。受訪者就算曾經打假波，都會回答「沒有」。沒有球員會在別人面前承認自己曾打假波。這樣，調查的準確度會受到影響。」

這時，財務總監張主任氣喘吁吁地走進會議室，上氣不接下氣地說：「對不起，今天因為塞車遲到，你們談到那裡？」李幹事把剛才關於向球員調查打假波的討論告訴張主任，並說：「這真是一個兩難的局面。」

黃主席說：「張主任，你好像在大學時期曾經修讀過統計和概率。對於這個問題，你有甚麼對策？」

張主任聽完後，說：「對於這個困難，我們可以採用「隨機化回答。」利用「隨機化回答」，受訪者會覺得他們的私隱受到保障，而我們可從訪問中獲取數據。」黃主席及其他與會者聽到後，都異口同聲說：「甚麼是隨機化回答？」

張主任回答：「讓我解釋一下吧。首先，我們在問卷中要設有以下兩條問題：

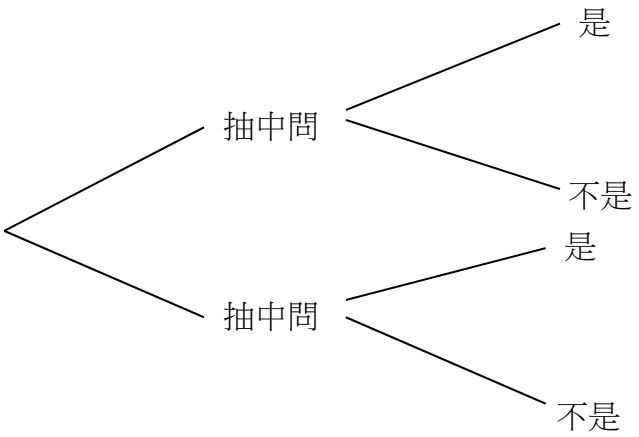
問題 A： 你曾參與打假波活動。
請回答「是」或「不是」。
問題 B： 你未曾參與任何打假波活動。
請回答「是」或「不是」。

李幹事：「張主任，這兩條問題十分相似。請問有何作用？」

張主任：「等我繼續解釋。隨機化回答的設計，是要受訪者根據一個隨機結果而回答問題 A 或問題 B。由於訪問員不知道受訪者回答那一條問題，受訪者會較放心回答問題。」

陳秘書：「當受訪者回答「是」時，可代表他曾參與過打假波活動，亦可代表他未曾參與過任何打假波活動，對嗎？」

張主任：「正確。我用以下的樹形圖解釋一下吧。



如果用 "*P*(是)" 代表受訪者回答「是」的概率，則有

$$P\text{ (是)} = P\text{ (是|抽中問題 A)} \times P\text{(抽中問題 A)} + P\text{ (是|抽中問題 B)} \times P\text{(抽中問題 B)}$$

假設我們訪問了 N 個球員，當中 n 個回答「是」，而曾參與打假波的球員佔全體球員的比是 p，那麼我們有

$$\frac{n}{N} = p \times P\text{(抽中問題 A)} + (1-p) \times P\text{(抽中問題 B)}$$

李幹事：「我們的訪問員可以知道抽中問題 A 和抽中問題 B 的概率嗎？」

張主任：「可以的。我們只須寫一個隨機數產生程式，該程式能產生 1 至 100 的整數 (包括首尾兩個數)。若受訪者產生的是 4 的倍數，則回答問題 A，否則回答問題 B。

這樣我們就可令到 $P(\text{抽中問題A}) = \frac{1}{4}$，而

$P(\text{抽中問題 B}) = \frac{3}{4}$。」

黃主席：「張主任，我大致明白你的做法。我總結一下你調查的步驟吧。

(1) 訪問員會帶備一部裝有上述隨機數產生程式的平板電腦。

(2) 訪問員須預先告訴受訪者，他不會觀看受訪者按動平板電腦的過程，令受訪者放心回答問題。受訪者完成回答後，可按動平板電腦的還原制，確保訪問員不知道隨機數的結果。

(3) 受訪者啟動隨機數產生程式。根據隨機數結果，受訪者回答問題 A 或問題 B。

(4) 假設我們訪問了 N 個球員，當中 n 的回答「是」，而曾參與打假波的球員佔全體球員的比是 p，那麼我們有 $\frac{n}{N} = \frac{1}{4} \times p + \frac{3}{4}(1-p)$，由此可得出 $p = 2\left(\frac{3}{4} - \frac{n}{N}\right)$。」

陳秘書：「那麼，我們明天開始調查吧！」

李幹事：「且慢。早前我們用了不少時間在收集自願性問卷調查的回覆。香城康體局局長限我們兩星期內交回報告。香城約有 10000 個全職球員。若果要逐一訪問所有的球員，我恐怕時間不夠。」

張主任：「我們可選取 1000 人做訪問。香城甲組、乙組和丙組的球員的人數大概是 **10: 7 : 3** 。根據這個比，我們可從甲組、乙組和丙組隨機選取 **500** 、**350** 和 **150** 人做訪問。這樣做的好處除了省時外，亦可了解不同組別球員打假波的行為。」

一星期後，足總完成調查工作。足總高層再次開會檢視統計結果。

組別	訪問人數 (N)	回答「是」的人數 (n)	曾打假波的概率 (p)
甲組	500	363	0.048
乙組	350	249	0.077
丙組	150	95	0.233

陳秘書：「主席，這是今次訪問的結果。」

黃主席：「看來香城球員「打假波」的情況頗普遍，尤其是丙組球員，有達兩成球員曾打假波。李幹事，你給我了解一下情況。」再次調查後，黃主席發現丙組球員「打假波」的主要原因是球員薪酬偏低。於是，足總高層開始研究提升球員薪酬來解決「打假波」這問題。

(字數：1952)

參考資料：

1. 隨機化回答：
http://en.wikipedia.org/wiki/Randomized_response

2. 傑出統計師的三十四年數字人生:
http://www.csb.gov.hk/hkgcsb/csn/csn72/72/c/persioners_1a.html

優異作品: Data Missing – Health Missing – Job Missing

School Name: St. Paul's Co-educational College
Names of Students: Lai Rachel, Fong Pak Yui
Level: Secondary 4
Supervising Teacher: Mr. Cheng Kam Muk

Introduction

Missing data is unavoidable in a large-scale investigation of any kind. They can adversely affect the investigation results, causing deviation that make the actual results completely opposite to the expected results. Everybody knows this simple fact; but how can we treat the data properly without being told off as a "liar"?

Exposition

"Rachel! If you do not produce satisfactory results in this investigation, you'll have no chance in passing the probation!" Rachel's boss yelled furiously. One can never blame him, however; the business of the fitness centre had been declining significantly and Rachel was to investigate the problem of obesity among their clients and thus making suitable packages for them.

Rachel mumbled several "yes" s and went out of the boss's office, horrified. She did not know where to begin, so she searched from the internet and received some data on local and overseas researches.

She discovered that BMI was often used as an indicator of obesity:

Body mass index (BMI) is a simple index of weight-for-height that is commonly used to classify overweight and obesity in adults. It is defined as a person's weight in kilograms divided by the square of his height in meters (kg/m^2).

- a BMI greater than or equal to 23 is overweight
- a BMI greater than or equal to 25 is obesity

Source: Department of Health

Using this as a benchmark, she generated a list on what she

needed to find:

Key figures to report in the health study
☐ Trend in the percentage of obese individuals
☐ Comparison of the obese trends between males and females
☐ Comparison of the obese trends between people of different ages
☐ Comparison of the obese trends between people of different professions

Apart from the asking the height and weight of the clients to calculate the BMI, which is a numerical estimation towards obesity, Rachel also decided to investigate into the lifestyles of the clients. She set a questionnaire so as to know what were the main causes of obesity and how the fitness company could apply different methods to pinpoint on such causes and solve the problem.

How Balanced are You?

1. What is your height? (in cm)

2. What is your weight? (in kg)

3. What is your age?
   ☐ ≥40 ☐ <40

4. What is your gender?
   ☐ Male ☐ Female

5. What is your occupation?

6. How much time do you usually spend in sports per week?
   ☐ >6 hours ☐ 5-6 hours ☐ 3-4 hours
   ☐1-2 hours ☐<1 hour

7. What kind of sports do you usually take? (please list):

8. How many times do you usually take in snacks per day?
   ☐More than three times ☐three times ☐twice
   ☐once ☐never

9. What kind of food would you take in as snacks?
   ☐ Sweets ☐Seaweed ☐ Potato Chips ☐ Biscuits

Afterwards she asked her colleagues to conduct the survey when the clients visited the centre during a period of two weeks. After some hard work of her colleagues, she finally got all the data from the company's clients and did a rough summary on the results. The numerical data is as follows:

	Males aged under 40	**Males aged over 40**	**Females aged under 40**	**Females aged over 40**
Underweight (BMI <18.5)	4 (5%)	3 (3.2%)	10 (10.3%)	16 (12.7%)
Normal (BMI 18.5 – 23.0)	32 (40%)	38 (40.4%)	42 (43.3%)	68 (54.0%)
Overweight (BMI 23.0 – 25.0)	21 (26.3%)	27 (2.1%)	18 (18.6%)	11 (8.7%)
Obese (BMI ≥25)	15 (18.8%)	19 (20.2%)	7 (7.2%)	5 (4.0%)
Missing	8 (10%)	7 (7.4%)	20 (20.6%)	26 (20.6%)
Total	80	94	97	126

To generate the trend in the percentage of obese individuals, Rachel decided to compare the data with a health study conducted by the Centre of Health Protection two years ago. It was therefore also listed here:

Classification (BMI)	Male Number (%)	Female Number (%)	Overall Number (%)
Underweight (BMI <18.5)	58 (6.3%)	159 (14.3%)	217 (10.6%)
Normal (BMI 18.5-23.0)	412 (44.5%)	611 (54.8%)	1 023 (50.1%)
Overweight (BMI 23.0-25.0)	198 (21.4%)	166 (14.9%)	365 (17.9%)
Obese (BMI ≥25.0)	239 (25.8%)	144 (12.9%)	383 (18.8%)
Unknown/Missing	19 (2.1%)	34 (3.1%)	53 (2.6%)
Total	926 (100.0%)	1 115 (100.0%)	2 041 (100.0%)

Source: Centre for Health Protection, HKSAR, April 2012

The Problem

However, a problem arose from the results of the study conducted by the Centre of Health Protection: how could the results provided by the Government be "unknown" or "missing"? Rachel had not supposed that such a thing would occur while conducting a survey. She was even more appalled to find that the results received from her colleagues was even more serious, with 15.4%.

"It is impossible for a person to have no height or no weight, isn't it? This is really a vital question..." she frowned.

The Reasons accounting for the Problem

Rachel called Pius, a secondary schoolmate of hers. Pius was working in the Census and Statistics Department, so she knew that he was the perfect person to answer her questions.

Pius and Rachel met on the next day at a cafe, while Pius introduced to her some of his favourite tea. Rachel then lamented on what she had come across – the missing information of the data collection.

"Oh, I see. This must be the problem due to missing data." Pius hypothesized, "Missing data refers to unknown values or absences in the values of a data collection. They could greatly affect your report if you do not treat them in the correct way."

"Oh right..." she rolled her eyes. "So what are the exact reasons for the occurrence of missing data? My colleagues only explained with vague reasons like 'the participants do not want to reveal their weight'. "

"Well, this is precisely why missing data occur!" Pius replied, "missing data is caused by the respondents themselves. For example, if one person was too busy to come to the centre for the two whole weeks, and another refused to reveal his weight, there would be no stored data in the corresponding collection. What do you expect about their results then, Rachel? "

"I have no idea about those who did not turn up at the fitness centre, but for those who did not answer, I guess their heights or weights are usually more embarrassing." said Rachel.

After some discussion, they generated some more possible reasons for missing data:

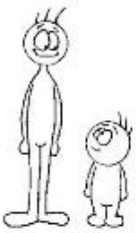
Possibilities of Missing Data: Missing not at random (NMAR)

What is meant by NMAR is that the data is missing because of the quantity it wants to collect (i.e. the dependent variable itself). In this case, it is more specifically the heights and weights of the participants. Since NMAR is caused due to certain preferences in giving away data from participants, the remaining data is usually biased.

What if they are really obese and do not wish to give away their weights such that they could avoid being teased at?

Or is it that they do not want to leave their sedate lifestyle so they deliberately left their 'weight' blank?

Could it be that they are so short that they fear showing the other people their real height?



All of the above speculations are due to the preferences of the participants. Therefore, the data is missing not at random.

Possibilities of Missing Data: Missing completely at random (MCAR)

This means that the data was missing not in relation to the independent and dependent variables.

The participants did not turn up at the fitness centre, but there is no explanation why. However, this leads to missing data since they should originally turn up but they did not.

Another possibility is that Rachel's colleagues forgot to take the heights and weights of the participants. Therefore after the fitness programme, the row for heights and weights was left blank.

All these data were missing not due to the bias of the participants, but due to the setting and the external conditions. Often, MCAR data are less significant in creating a bias; so somehow they tend to be omitted.

Source: Missing Data and How to Deal: An overview of missing data by M. Humphries

Rachel then talked about the missing data in the questionnaire. Handing to him the questionnaire she designed, she explained," I have also asked guiding questions to find out how people develop

obesity. Nevertheless, most respondents simply left them blank!"

6. How much time do you usually spend in sports per week?
   ☐ ≥4 hours ☐ 3-4 hours ☐ 2-3 hours
   ☐1-2 hours ☐≤1 hour

7. What kind of sports do you usually take? (please list):

8. How many times do you usually take in snacks per day?
   ☐More than three times ☐three times
   ☐twice ☐once ☐never

9. What kind of food would you take in as snacks?
   ☐ Sweets ☐Seaweed ☐ Potato Chips ☐ Biscuits

"Now I understand." With a cheeky smile, he responded, "Not saying that I have perfect hindsight, but you should have told them the aim of the investigation is to **provide a tailor-made course, or even provide discounts** for them. That way, they would surely accept the questionnaire. Also, instead of asking your colleagues to inquire the clients' height and weight, you can simply **measure them directly**. Not only all the data -- if the person shows up, that is -- can be collected, you can also reduce the errors arising from variances between each balance or ruler. Besides, you can also add a '**others**' in list questions instead of giving rigid closed answers or open-ended answers. This way, they will have no excuse to say that they have no choice in the

answers.”

With a slight pause to make sure Rachel understood, he proceeded: "Open-ended questions frequently receive blank answers, right? This is also another reason why missing data occur. You cannot sort them, but they are still counted. Moreover, for question 9, if the respondents failed to find a suitable option, or if they had answered ‘never’ in question 8, the answer would be blank. This is a default in the design of the questionnaire, more commonly known as ‘skip patterns’, and it is sometimes inevitable because it is applicable only to some people."

9. What kind of food would you take in as snacks?
   ☐ Sweets ☐Seaweed ☐ Potato Chips ☐ Biscuits
   ☐ Others (please specify: _____)

Method I. Listwise Deletion

“What? It seems like I have made a disaster!” Rachel was startled. "Calm down... for clarification, can I collect some of your data?" Pius said calmly.

Rachel then randomly selected 20 sets of data collected from one of the centres for him:

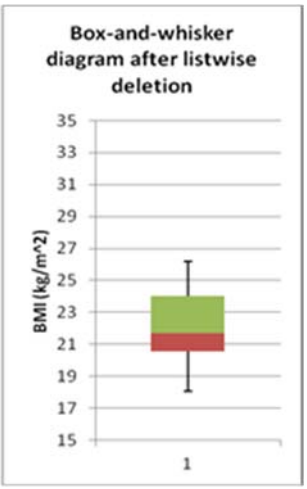
Client	Gender	Age	Height (m)	Weight (kg)	BMI (kg/m^2)
1	F	48	1.51	48	21.05
2	F	18	1.63	53	19.95
3	M	21	1.68	58	20.54
4	F	41	1.31	31	18.06
5	M	46	1.84	79	23.33
6	F	53	1.66	N/A	N/A
7	M	28	N/A	N/A	N/A
8	F	37	1.54	49	20.66
9	M	52	1.86	84	24.28
10	M	30	1.93	80	21.48
11	F	20	1.34	N/A	N/A
12	M	45	1.67	73	26.18
13	F	35	1.69	70	24.51
14	F	39	1.70	66	22.84
15	F	64	N/A	71	N/A
16	M	41	1.85	75	21.91

Client	Gender	Age	Height (m)	Weight (kg)	BMI (kg/m^2)
17	M	22	1.83	86	25.68
18	F	33	1.61	N/A	N/A
19	F	59	N/A	N/A	N/A
20	M	19	1.42	37	18.35

**fabricated materials*

"To generate results, I simply removed those with missing data
(i.e. Clients 6, 7, 11, 15, 18, 19)
and arrived with the following:

1. The Mean = 22.06 kg/m^2
2. Standard Deviation = 2.436 kg/m^2
3. A Box-and-whisker diagram
   (as on the right).



Just by referring to the results, we can see that the mean BMI is much lower than the definition of the World Health Organisation. Furthermore, in the random sample, only two out of fourteen people are obese, that is 14.3%, indicated in yellow. "

"Well Rachel, you have done quite a good job! Listwise deletion, which you have done, is a method in removing missing data. Although this is rather easy, you can see that the data can be

utterly different from the reality. Here if I re-input the data as follows:

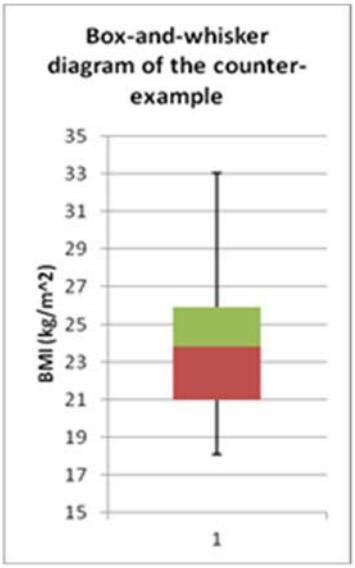
Client	Gender	Age	Height (m)	Weight (kg)	BMI (kg/m^2)
1	F	48	1.51	48	21.05
2	F	18	1.63	53	19.95
3	M	21	1.68	58	20.54
4	F	41	1.31	31	18.06
5	M	46	1.84	79	23.33
6	F	53	1.66	91	33.02
7	M	28	1.59	76	30.02
8	F	37	1.54	49	20.66
9	M	52	1.86	84	24.28
10	M	30	1.93	80	21.48
11	F	20	1.34	44	24.50
12	M	45	1.67	73	26.18
13	F	35	1.69	70	24.51
14	F	39	1.70	66	22.84
15	F	64	1.55	71	29.55
16	M	41	1.85	75	21.91
17	M	22	1.83	86	25.68
18	F	33	1.61	67	25.85
19	F	59	1.37	49	26.11
20	M	19	1.42	37	18.35

Mean: 23.89 kg/m^2

Standard Deviation: 3.82 kg/m^2

Percentage of people overweight: $\frac{7}{20} \times 100\% = 35\%$

Comparing the two sets of data, we can see a difference in mean BMI of over 1 kg/m^2! In addition, there may be a great loss of data after listwise deletion, in this case we have 61, and this may leave with us a blurred, misleading image of the actual situation," he stated. "Only if the missing data is MCAR, listwise deletion could be a good method in doing so."



Method II. Imputing the Mean

"The second way is to impute data like what I have just done. You can straightforwardly input the mean or the median in the missing spaces," said Pius.

"But of course, putting the mean height and weight would not be that rational. Imagine a 1.34 m person weighing 64 kg (which is the mean weight of the 15 clients)! Wouldn't that be funny!

"Therefore, we must consider **the relationship between variables**. Putting the original mean BMI (22.06 kg/m^2) would be a better option.

To calculate the heights and weights, we can derive them from the mean BMI.

Take Client 6 as an example. Since she is 1.66 m tall while her BMI is 22.06 kg/m^2, her weight is (22.06 × 1.66^2) kg = 60.789 kg."

Client	Gender	Age	Height (m)	Weight (kg)	BMI (kg/m^2)
6	F	53	1.66	60.79	22.06
7	M	28	N/A	N/A	N/A
11	F	20	1.34	39.61	22.06
15	F	64	1.79	71	22.06
18	F	33	1.61	57.18	22.06
19	F	59	N/A	N/A	N/A

**The boxes in green signify the iterated mean BMI; the blue boxes are the heights and weights calculated from it. Clients 7 and 19 cannot have the missing values substituted because their set of data is completely missing.*

Mean: 22.06 kg/m^2

Standard Deviation: 2.1486 kg/m^2

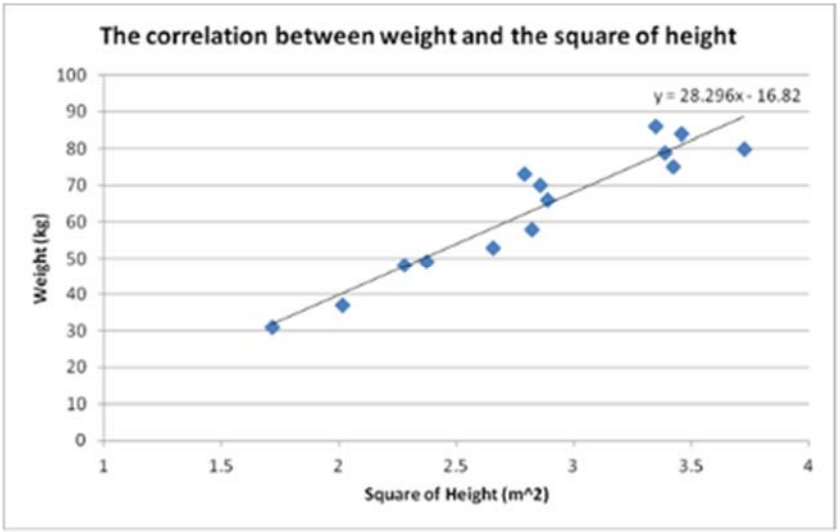
Percentage of people overweight: $\frac{2}{18} \times 100\% = 11\%$

“The bad thing about this method is that the bias still cannot be minimised. The variance is substantially weakened.” Pius said.

Method III. The Linear Regression Model

Rachel smiled, but she didn’t seem contented with inputting the mean either.

Pius further offered help. "How about using the linear regression model? If we generate a diagram of the known heights and weights of the respondents like this, we can arrive at a **best fit straight line**.”



And we impute the data as follows, which accords to the formula (y=28.296x – 16.82) given by the best fit line of the graph:

Client	H^2 (m^2)	W (kg)	BMI (kg/m^2)
6	$1.66^2=2.7556$	61.15	22.19
11	$1.34^2=1.7956$	33.99	18.93
15	$\sqrt{3.1036}=1.76$	71	22.88
18	$1.61^2=2.5921$	56.53	21.81

Mean: $21.92 kg/m^2$

Standard Deviation: $2.276\ kg/m^2$

Percentage of people overweight: $\frac{2}{18} \times 100\% = 11\%$

"As for the missing data like clients 7 and 19, you can just take any point on the best fit line. This may make the whole piece of data closer to the average. Moreover, you can find a formula on which you can depend to match different sets of data." Pius said as he nodded his head. "However, we are assuming that we could fit all the data on a straight line. In this case, the variance is weakened and the bias still exists, although it is already slightly reduced."

Method IV. Multiple Imputation

"Oh yes, Rachel. I've been recently studying on a widely accepted method in missing data, and your investigation has come right into the place," Pius said happily.

"What? What is it?" Rachel was more than eager to know.

"It is called multiple imputation," mused Pius, "by generating several regression models like the previous one, we can have sets of data making use of all variables.

"However, the whole computation may seem a little complicated, but actually it is very suitable for a person who is greatly in need of it like you. "By using statistical software named **SPSS** as stated here, the data will be analyzed and imputed according to the different variables. The software then generates regression models, characterized by vague regression lines. Therefore we can arrive at these sets:

Dataset 1

Client	Height (m)	Weight(kg)	BMI (kg/m^2)
6	1.66	71.1	25.80
7	1.82	76.6	23.13
11	1.34	39.4	21.94
15	1.69	71	24.86
18	1.61	59.9	23.11
19	1.84	76.1	22.48

Dataset 2

Client	Height (m)	Weight(kg)	BMI (kg/m^2)
6	1.66	57.2	20.76
7	1.80	69.9	21.57
11	1.34	33.3	18.55
15	1.75	71	23.18
18	1.61	52.3	20.18
19	1.75	72.8	23.77

Dataset 3

Client	Height (m)	Weight(kg)	BMI (kg/m^2)
6	1.66	59.4	21.56
7	1.47	52.1	24.11
11	1.34	36.8	20.49
15	1.69	71	24.86
18	1.61	67.2	25.92
19	1.67	60.2	21.59

Dataset 4

Client	Height (m)	Weight(kg)	BMI (kg/m^2)
6	1.66	60.4	21.92
7	1.73	64.1	21.42
11	1.34	33.5	18.66
15	1.73	71	23.72
18	1.61	57.1	22.03
19	1.78	74.3	23.45

Dataset 5

Client	Height (m)	Weight(kg)	BMI (kg/m^2)
6	1.66	60.7	22.03
7	1.61	66.0	25.46
11	1.34	35.8	19.94
15	1.81	71	21.67
18	1.61	54.4	20.99
19	1.72	63.5	21.46

Afterwards, you can take the mean by combining the five sets of data:

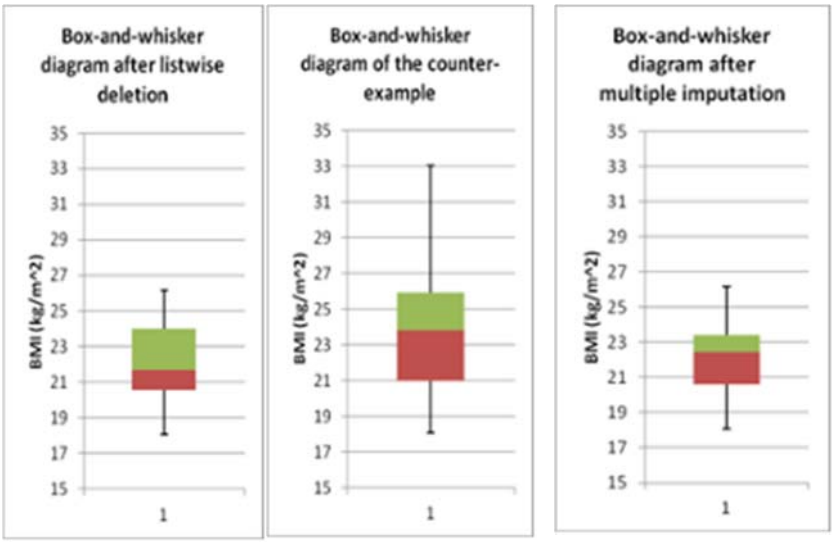
Client	BMI (kg/m^2)
6	22.414
7	23.138
11	19.916
15	23.658
18	22.446
19	22.550

Overall, the following is therefore obtained.

Mean: 22.15 kg/m^2

Standard Deviation: 2.142 kg/m^2

Percentage of people overweight: $\frac{2}{20} \times 100\% = 10\%$



Practise the same steps over the rest of the data, and soon you will have everything done. Although it may be a little time consuming, technology can extend a helping hand. Moreover, I deeply believe that the time consumed is much worth it."

Rachel was filled with excitement and gratitude. "Thank you! I'll modify the data by imputation, most likely by the multiple imputation. I shall generate a more accurate result and secure my job. How can I ever thank you?"

“Good luck! Let us have another cup of tea next time!” waved Pius before he left.

Denouement

It was two weeks later. Rachel's boss had read the report and was surprised by the progress she made.

"Thank you. I now have a better idea on obesity and there must be a way to revive our company. Impressive! Congratulations, you passed the probation period. What's more, I will offer you a pay rise!"

Rachel was delighted, not just because she could get her dream job, and most importantly due to the fact that she finally understood that the “mathematics” that she put away years ago would really prove useful in a field completely irrelevant to mathematics.

Missing Data seems to be quite of a trivial problem in statistics, but somehow it can make the whole result quite biased. Deletion may be a simple approach to the problem, but imputation is a more modern and safe method in tackling missing data. Often, imputation involves more complicated methods, but it is more reliable in reducing the bias. But deep inside, she knew that none of the methods could generate 100% accurate data, but only

estimations. Improving the collection of data would be more essential in solving this problem rather than the method used to remove missing data. *I can improve my method of collection to avoid missing data,* Rachel thought to herself, *but at the meanwhile, why don't we have some tea?*

(2486 words)

References:

1. Centre of Health Protection, Hong Kong
   http://www.chp.gov.hk/en/content/9/25/8802.html
   http://www.chp.gov.hk/en/data/1/10/280/1331.html

2. Melissa Humphries, *Missing Data and How to Deal: An overview of missing data*

優異作品: 換樂無窮

學校名稱：香港四邑商工總會黃棣珊紀念中學

學生姓名：陳明詩、李茵、朱梓琛

冒險天地內 ー

「哇! 這個公仔好可愛啊哈哈，但是要好多張票才可以換到啊!」小怡望著那隻要 500 張票才能換到的毛公仔，不禁一陣慨歎。

這時，小明走近小怡悄悄地說：「聽說這部幸運機和擲彩虹可以賺取最多的票啊，不如我們就來試試看這個傳聞到底是不是真的吧! 要是真的話，你很快就可以得到心愛的毛公仔了!」

小怡聽了以後一陣歡喜，於是，他們倆買了 30 個金幣來「先打頭陣」，去試一試傳聞的真假。

很快，他們帶著剛換到手的金幣，興高采烈地跑到了擲彩虹的面前。看著那色彩斑斕的彩虹一個個地展現在眼前，再加上獎品換領處的各個公仔，他們都感到眼前一亮。

「哇! 真的好多公仔啊! 好想快點得到啊，到底是不是真的有很大機會的? 我不想浪費錢啊」小怡雖然很開心，但是仍然有所顧慮。

「俗話說，心急吃不了熱豆腐，不如我們先看看規則吧。」小明冷靜地說。

「1. 一切中獎以由高空垂直的角度向下望為準。」
「2. 手不可以超過圍欄。」
「3. 金幣必須擲於彩虹的顏色內，如有擲界，一律作廢。」

小怡說：「原來扔個硬幣都有這麼多規矩的...那麼我想問扔中每一種顏色各有幾多票呀?」

小明說：「這裡已經寫了。紅色可以獲得 20 票，黃色可以獲得 100 票，綠色可以獲得 500 票，藍色可以獲得 1000 票。哇! 如果你一次就擲中綠色，那你就可以換到你喜歡的公仔啦。」

「那事不宜遲，我地快點開始吧!!」

擲了很多次後 ……

「啊! 我終於擲到有顏色的地方了，哈哈!」小明高興地說。

「等一等，你剛才不是有看規則的嗎? 你有沒有見到你的硬幣踏到黑線上?」小怡見到後有些懷疑地說。

「是嗎? 讓我問一問工作人員吧。」 「麻煩問一下我這個硬幣可不可以得獎啊?」

工作人員：「不好意思啊先生。你這個硬幣是踏界的。不要緊啦，再接再厲啊!」

在以上情況下，雖然小明扔中了有顏色的區域，但是因為碰到了黑邊，即踏界，所以不作中獎論。由此可見，要中獎，硬幣的任何一處地方都不可以碰到黑邊的任何一處，硬幣的圓心要掉落一定的範圍內，才能獲獎。

「怎麼這樣啊? 我還以為有機會得到公仔呢!」小怡抱怨說。

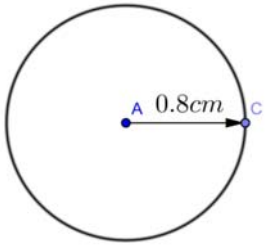
「啊…現在看起來好像很難啊。難道真的沒有機會嗎?」小明說。

聽到這兩位朋友的疑問，又是我們數學三人組出場的時候了。我們是無處不在的，專門為大家解決生活上各種數學難題。

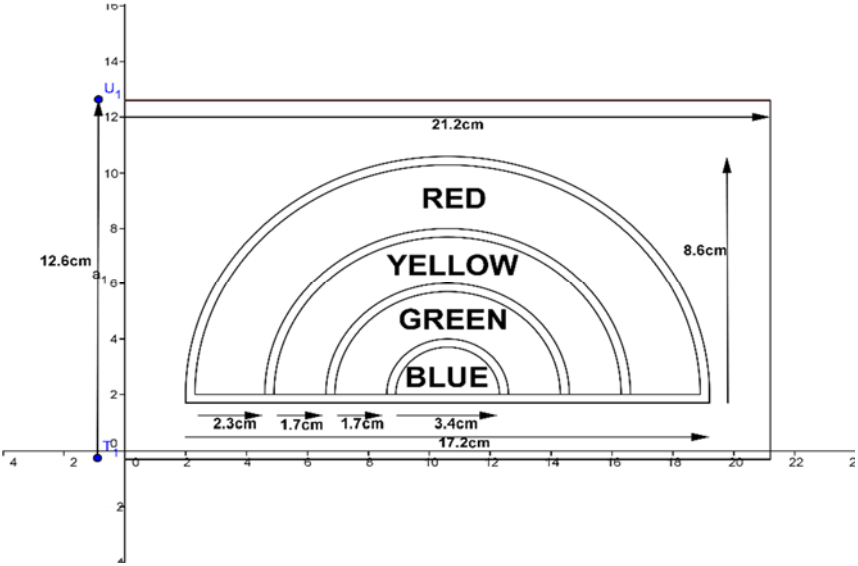
我們問他們：「怎麼了?有需要幫忙的地方嗎?」

小怡說：「我們很想得到冒險樂園裡的獎品，但是在擲彩虹這個遊戲裡，似乎沒有機會了! 你們幫我們看一看，我們有多大機會獲獎呢?」

玩擲彩虹的遊戲，中獎的概率是多少?
要解答以上的問題，首先，我們要知道用來擲彩虹的硬幣，它的半徑是多少。



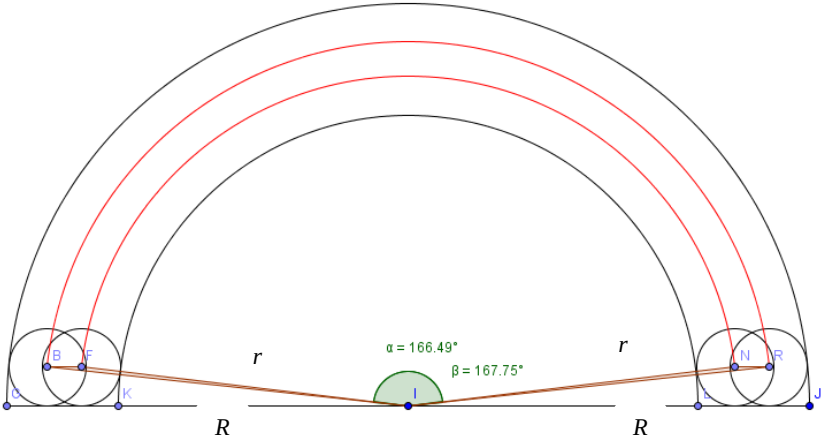
上圓的圓代表用來擲彩虹的硬幣，正如上圖所示，硬幣的半徑為 0.8cm。其次，每次擲幣只有一枚，但彩虹板上有多個相同大小，距離相若的白色區域。因此我們可以假設每次投幣時，圓心都會落在以下的矩形內。



彩虹的長度為 17.2 cm，高 8.6 cm。而彩虹以外的地方則是白色的區域，該區域長 21.2 cm，闊 12.6 cm，黑邊闊 0.3 cm。而彩虹各種顏色的闊度如下，紅色的闊 2.3 cm，黃色的闊 1.7 cm，綠色同樣闊 1.7 cm。

由於彩虹的各種顏色有不同的闊度，假設圓心的位置落在區域上任何一點的機會是相同的，我們將以圓心可以掉落的位置綜合成面積，來計算出擲中彩虹每種不同顏色的不同概率，最後再計算出擲彩虹中獎的概率。

首先，是彩虹紅色的部分。



如上圖所示，我們分別把 4 個硬幣放到彩虹的兩端上，並貼近最底的黑邊。其中兩個硬幣的圓貼近外面的黑邊，另外兩個硬幣的圓則貼近裡面的黑邊。

然後，我們把兩個貼近外面黑邊的硬幣的圓心相連，形成一條弧線 (BR)。同樣，再把兩個貼近裡面黑邊的硬幣的圓心相連，形成另一條弧線 (FN)。最後，把兩端的兩個硬幣的圓心相連(BF 和 NR)。這時，根據圖表所示，紅色的區域就是我們期望硬幣的圓心可以掉落的範圍，因為圓心只要在這個區域內，就不會碰到黑邊，換句話說，可以中獎。

紅色區域的面積

$$=\frac{(\pi)(R^2)(\beta)}{360°}-\frac{(\pi)(r^2)(\alpha)}{360°}-\frac{1}{2}(r)(n)sin\left(\frac{\beta-\alpha}{2}\right)-\frac{1}{2}(r)(n)sin\left(\frac{\beta-\alpha}{2}\right)$$

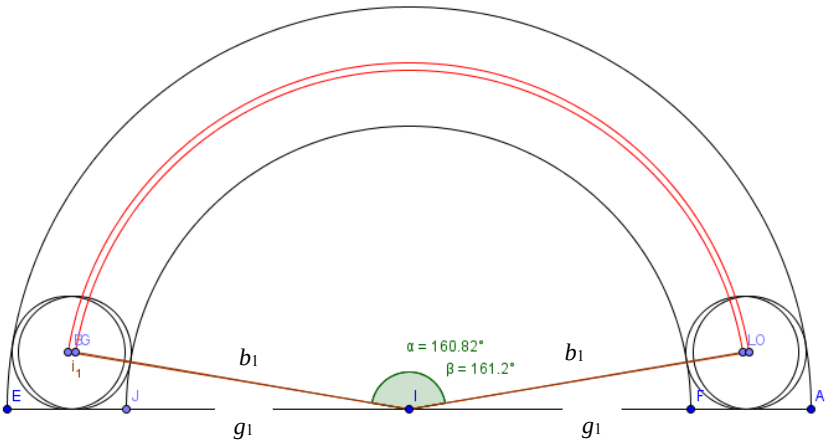
$$= \frac{(7.5)^2(\pi)(167.75°)}{360°} - \frac{(6.8)^2(\pi)(166.49°)}{360°} - \frac{1}{2}(6.8)(7.5)sin\left(\frac{167.75° - 166.49°}{2}\right)$$
$$- \frac{1}{2}(6.8)(7.5)sin\left(\frac{167.75° - 166.49°}{2}\right)$$
$$\approx 14.601cm^2$$

擲中彩虹紅色部分的概率 $= \frac{\mathbf{14.601}}{\mathbf{(21.2)(12.6)}}$

$$= \frac{\mathbf{14.601}}{\mathbf{267.12}}$$
$$\approx 0.054661801$$

彩虹其他顏色的部分，我們會以相同的方法，找出擲中該區域的概率。

以下是彩虹黃色的部分。



同樣，根據圖表所示，黃色的區域就是我們期望硬幣的圓心可以掉落的位置，因為圓心只要在這個區域內，就不會

碰到黑邊。

黃色區域的面積

$$= \frac{(\pi)(g_1)(\beta)}{360°} - \frac{(\pi)(b_1)(\alpha)}{360°} - \frac{1}{2}(b_1)(g_1)sin\left(\frac{\beta-\alpha}{2}\right) - \frac{1}{2}(b_1)(g_1)sin\left(\frac{\beta-\alpha}{2}\right)$$

$$= \frac{(4.9)^2(\pi)(161.20°)}{360°} - \frac{(4.8)^2(\pi)(160.82°)}{360°} - \frac{1}{2}(4.8)(4.9)sin\left(\frac{161.20°-160.82°}{2}\right)$$

$$-\frac{1}{2}(4.8)(4.9)sin\left(\frac{161.20°-160.82°}{2}\right)$$

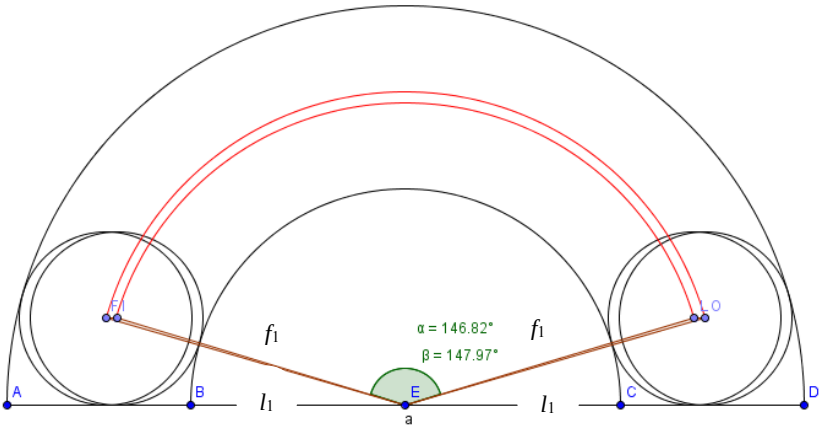
$$\approx 1.363cm^2$$

擲中彩虹黃色部分的概率 $\approx \frac{1.363}{(21.2)(12.6)}$

$$\approx \frac{1.363}{267.12}$$

$$\approx 0.005102575$$

以下是彩虹綠色的部分。



綠色區域的面積

$$=\frac{(\pi)(f_1^{\,2})(\beta)}{360°}-\frac{(\pi)(l_1^{\,2})(\alpha)}{360°}-\frac{1}{2}(f_1)(l_1)sin\left(\frac{\beta-\alpha}{2}\right)-\frac{1}{2}(f_1)(l_1)sin\left(\frac{\beta-\alpha}{2}\right)$$

$$=\frac{(2.9)^2(\pi)(147.97°)}{360°}-\frac{(2.8)^2(\pi)(146.82°)}{360°}-\frac{1}{2}(2.8)(2.9)sin\left(\frac{147.97°-146.82°}{2}\right)$$

$$-\frac{1}{2}(2.8)(2.9)sin\left(\frac{147.97°-146.82°}{2}\right)$$

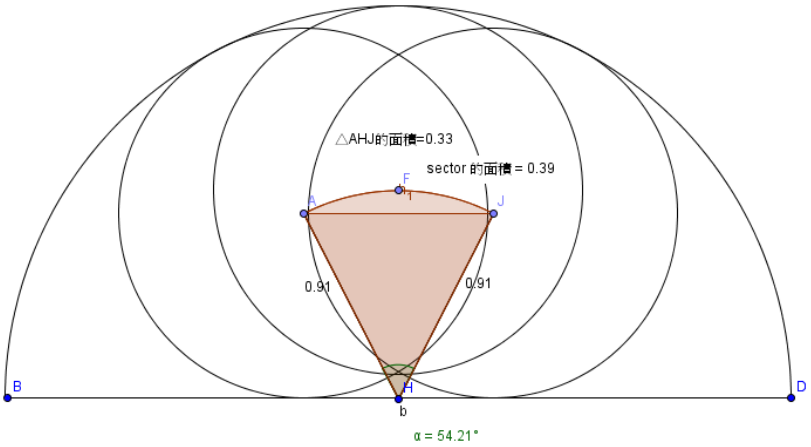
$$\approx 0.733cm^2$$

擲中彩虹綠色部分的概率 $\approx \frac{0.733}{(21.2)(12.6)}$

$$\approx \frac{0.733}{267.12}$$

$$\approx 0.002744085$$

以下是彩虹藍色的部分。



由於彩虹藍色的部分是一個半圓，硬幣的圓心可以掉落的

範圍是一個弓形(AFJ)。

根據數學軟體 GeoGebra，弓形 AFJ 的面積
= 扇形(AFJH)的面積－ΔAHJ 的面積
= 0.391750504 - 0.335863241
= 0.055887262 cm^2

擲中彩虹藍色部分的概率 $= \dfrac{0.06}{(21.2)(12.6)}$

$= \dfrac{0.06}{267.12}$

$= 0.000224618$

綜合擲中彩虹各種顏色的概率，中獎的概率如下。

中獎的概率 $= 0.054661801 + 0.005102575 + 0.002744085 +$
0.000224618
≈ 0.062733079

因此，我們得出以下的期望值。
$= 0.054661801 \times 20 + 0.005102575 \times 100 + 0.002744085 \times 500 + 0.000224618 \times 1000$
≈ 3.0 票
≈ 3 票

由此可見，擲中彩虹紅色部分的概率比擲中其他顏色的遠遠多出十多倍，更甚是二百多倍。

小結

從以上的結果得知，在擲彩虹中，擲中彩虹各種顏色的概率非常小，擲中藍色的概率是微乎其微，比擲中其他顏色的遠遠多出三十倍，更甚是四百多倍。雖然擲中彩虹紅色部分的機會有大約二十分之一，但是我們卻不能避免其他意外的情況出現。例如，我們這次的實驗是假設了硬幣不會掉落兩側的坑內，也不會掉落人行的通道上。擲中彩虹紅色部分的機會是四種顏色中最多的，而藍色是最少的。

「原來玩彩虹贏的機會也不是很大呢，幾乎等於沒有啊。」小怡說。

「那麼我們還浪費了 15 個金幣呢，不過也謝謝你們的計算。如果不是你們，我們還一直在浪費金錢呢!」小明說。

「哈哈! 不用謝。既然玩彩虹不能得到獎，那就看看在樂園裡有沒有什麼其他的遊戲有多些機會吧!」三人組中的小菌說。

「我們一開始不是說過幸運機也可以獲得多獎勵嗎? 不如我們去試一試吧!」小明說。

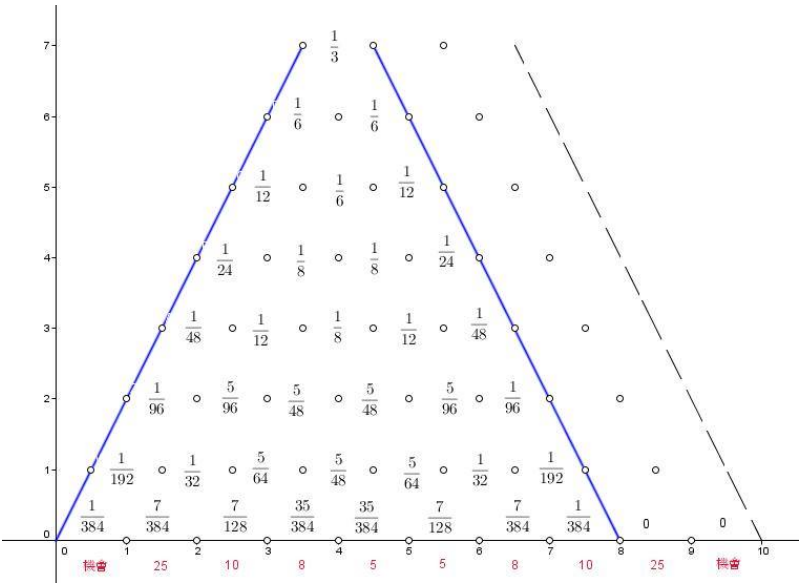
於是，我們五個人就一齊到了幸運機的面前。

這部幸運機分成了 10 個空格，各有兩個機會(即 40 票)、25 票、10 票、8 票、5 票的格子對稱地分成兩邊。因為搖擺條只會落在左、中、右和隨意四個方向，並且維持相同的停留時間，我們可以分別演繹四種不同的個案，來看看落在哪個方向能獲取最多的票數。

個案一

當搖擺條落在左邊時，就佔了整個搖擺時間的三分之一。

所以，這枚硬幣只能去到右邊的 10 票，而不能碰及 25 票和機會票。



每個障礙物都使得落入最下面空格的概率產生變化，而且從左邊入，覆蓋的三角形面也相對較少。

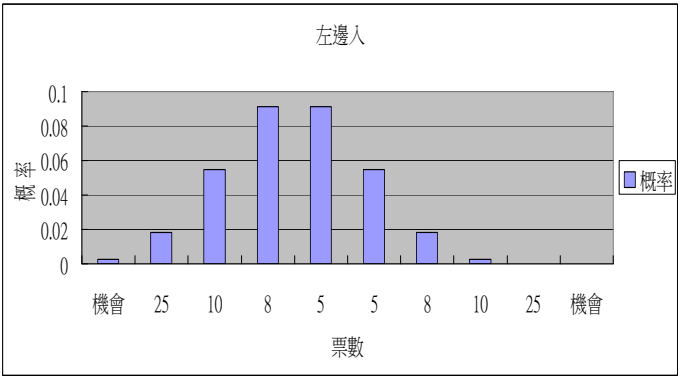
例如：

落入機會票格的概率：

$=(\frac{1}{192})(\frac{1}{2})$

$=\frac{1}{384}$

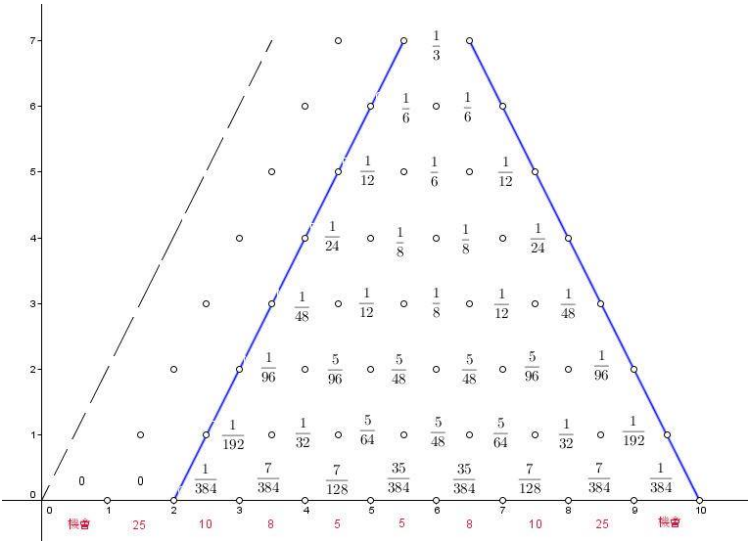
以下是當硬幣從左邊入時，到達票數格的概率。



從上圖可見，右邊的 25 票和機會票是沒可能落中的，而他們的概率全給了左邊的機會票，令落中機會票的概率上升。

個案二

當搖擺條落在右邊時，也佔了整條搖擺時間的三分之一。所以這枚硬幣只能去到左邊的 10 票，而不能碰及 25 票和機會票。

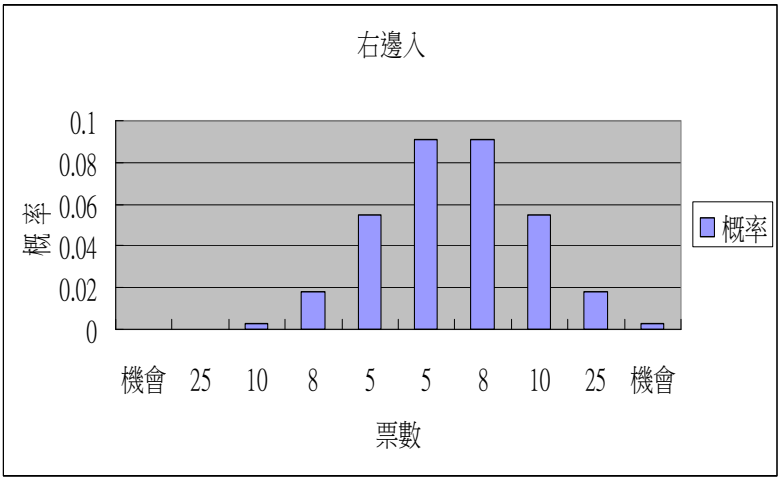


同樣的，落入機會票格的概率：

$= (\frac{1}{192})(\frac{1}{2})$

$= \frac{1}{384}$

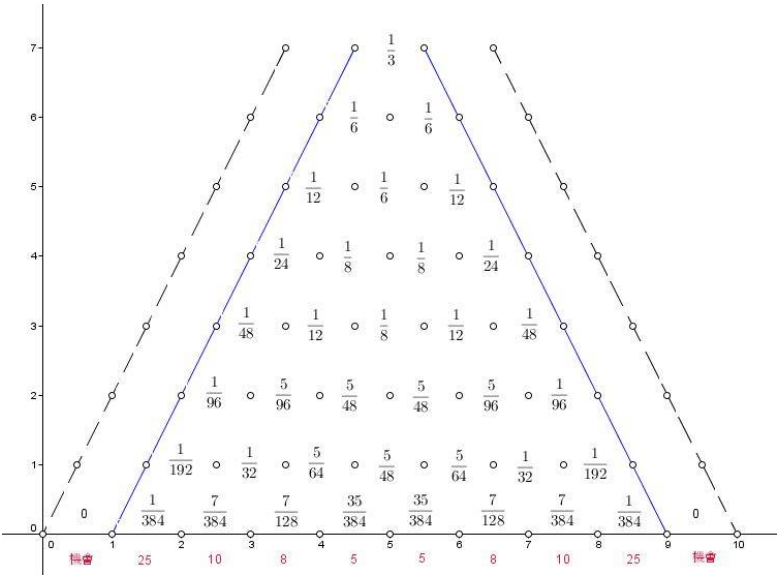
下圖是當硬幣從右邊入時，落入各個票數格的概率。



與個案一一樣，不過是左右位置的調換，從右邊入時，左邊的 25 票和機會票的概率都是 0。

個案三

當搖擺條從中間落下，同樣的也佔了整個搖擺時間的三分之一。但是，這一次的硬幣卻能隨意地落入中間的八個票數格內，而不能落入兩邊的兩個機會票格內，並且兩邊的概率都是對稱的和相同的。



這一次的落入機會票格的概率：

= 0

並且兩邊的機會票的概率是對等的。

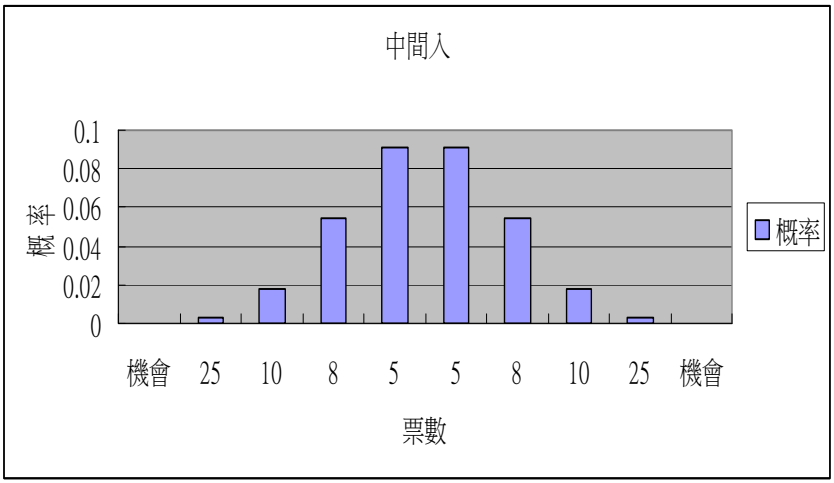
又例如：

落入 5 票格的概率：

$$=\left(\frac{1}{2}\left[\left(\frac{5}{64}\right)+\left(\frac{5}{48}\right)\right]\right)$$

$$=\frac{35}{384}$$

兩個中間的 5 票都是有著一樣的概率，兩個旁邊的機會票也是有著一樣的概率，所以其他對稱的票數也是有一樣的概率，也就是說，這個硬幣只要從中間入的話，兩邊的概率是對等的，落入中間的機會也會較大。
就如下表：

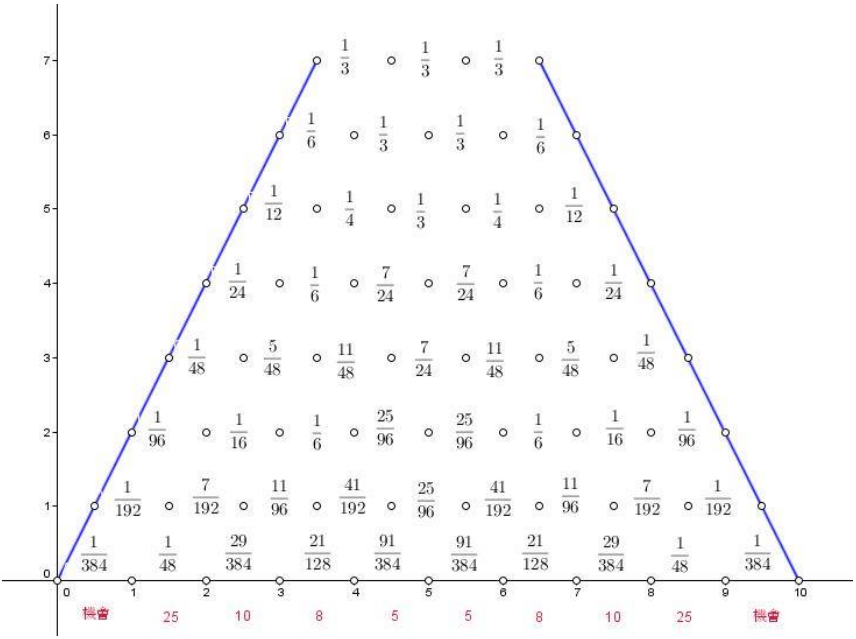


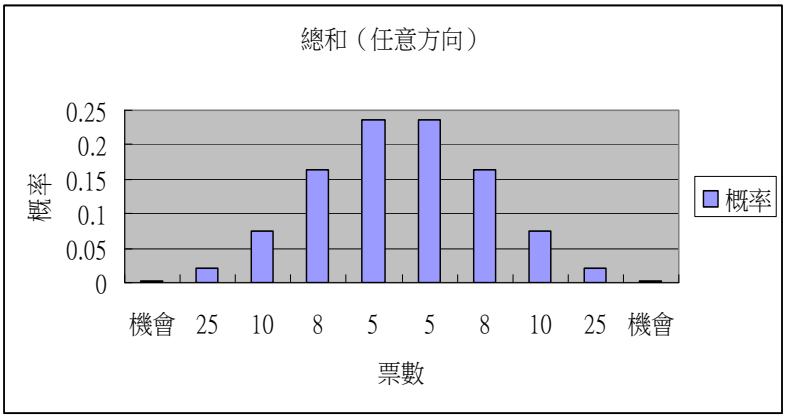
如果從中間放入硬幣的話，絕大的機會會落入 5 票，而落入機會票（40 票）的機會是零。

來看看三個個案作一個比較，可見從左右兩邊放入硬幣有更大的機會獲得更多的機會票，若從中間投入硬幣，就較大機會獲得 5 票。因為從左右兩邊放入硬幣時，另外一邊的 25 票和機會票會減少至 0，而它們的概率就會加至那一邊入的機會票中，所以就會增加某一邊的機會票的概率。

個案四

如果我們任意地放入硬幣，那麼這個個案的總概率就會等於 1，看看下表，這是隨意放入一個硬幣，有可能得到的票數的概率，這些平均數告訴我們始終是會有很大的概率獲得 5 票。但是，總的來說，這樣任意地落入硬幣所獲得的機會票的概率，仍然比固定地從中間落入的獲得機會票的概率多。





期望值

雖然從每一邊進入票數格的時間各佔三分之一，但是實際上我們未必每一次都可以進入自己想入的那一邊，所以，我們可以運用個案四的圖表來看看隨意地放入硬幣可以得到的期望值是多少。

$$E(any)=2[(40)(\frac{1}{384})+(25)(\frac{1}{48})+(10)(\frac{29}{384})+(8)(\frac{21}{128})+(5)(\frac{91}{384})]$$

$$= 7.755$$

但是，一些常玩幸運機的達人，他們是可以很準確地控制幸運機的搖擺條所落下的方向，從而得出單一個方向的期望值。因為每個方向各佔搖擺時間的三分之一，所以我們需要把每個方向的期望值乘三，才會得出單一個方向的期望值。

$$E(\text{left})=3[(40)(\frac{1}{384}) + (25)(\frac{7}{384}) + (10)(\frac{7}{128}) + (8)(\frac{35}{384}) + (5)(\frac{35}{384}) + (5)(\frac{7}{128}) + (8)(\frac{7}{384}) + (10)(\frac{1}{384}) + (25)(0) + (40)(0)] = 8.2109375$$

$$E(\text{right})=3[(40)(0) + (25)(0) + (10)(\frac{1}{384}) + (8)(\frac{7}{384}) + (5)(\frac{7}{128}) + (5)(\frac{35}{384}) + (8)(\frac{35}{384}) + (10)(\frac{7}{128}) + (25)(\frac{7}{384}) + (40)(\frac{1}{384})] = 8.2109375$$

$$E(\text{middle}) = (3)(2)[(25)(\frac{1}{384}) + (10)(\frac{7}{384}) + (8)(\frac{7}{128}) + (5)(\frac{35}{384})] = 6.84375$$

按照以上的分析，我們可以看到如果任意地放入硬幣，它所票數的概率大約是 7 票，相比之下，因為左右兩邊的期望值都是一樣的，所以從左右兩邊落入硬幣會有較大的概率獲得更多的票數，即是有 8 票。而從中間放入硬幣的話，僅能獲得 7 票，是最少的票數。所以，如果你們想要獲得最多的票數，就要等到搖擺條從左邊或右邊落入票數格內，其次是任意地投入硬幣和中間落入硬幣。這種方式才能幫助你們用更高的概率獲得更多的票數。

小結

現實生活中的應用並不等於理論中得到的結果，若我們矇著眼睛任意地去放入硬幣，一定要用多幾倍的硬幣去投(500 個或是 1000 個或是 3000 個)，實踐的結果才能和理論概率得出相似的結果。另外，對於這遊戲來說，從左右兩邊放入硬幣，才能得到最多的票數。

總結

綜合兩個實驗，我們發現……

在擲彩虹中，人們普遍認為扔中顏色的區域就起碼有 20 票，但這是一個錯誤的期望，我們得出任意地擲彩虹的期望值是 3 票，與人們所想的有著天淵之別。所以，我們不建議玩擲彩虹。

在幸運機中，獲得票數和獎勵的概率是 1，即是一定會獲得票數，比擲彩虹的概率，約十七分之一高得多。我們在幸運機中，期望獲得的票數有 7-8 票，比擲彩虹高約 2-3 倍。

	擲彩虹	幸運機
獲取票（任何數量）的概率	$\frac{1}{17}$	1
期望獲得的票數（票）	3	7-8

但是，這個理論只是我們計算得出的結果，並不等於實驗獲得的票數。只有在長期的觀察或者嘗試很多次以後，才會得出和實踐結果接近的答案。在票價值相等的情況下，我們比較推薦大家玩幸運機。

「雖然兩個遊戲都有獎勵，但似乎不是很多呢。」小怡說。

「是啊。而且有時候還很講運氣啊。」小明說。

「那現在經過了我們的分析後，你們會玩哪一個呢？」小菌說。

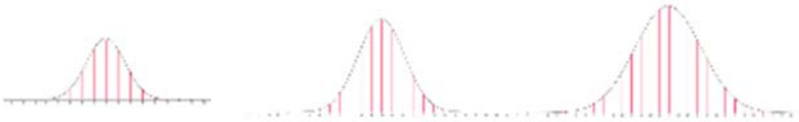
「幸運機！」小怡和小明異口同聲地說道。

邀請作品：68、95、99.7

2012 至 2013 年度借調老師：朱立夫

高中的同學當修讀統計課時，其中一個重要的課題是正態分布。因為正態分布能描述很多日常生活中的統計數據分布，例如一所學校內的學生高度，或一間工廠生產的金屬棒長度等。而正態分布的其中一項性質，就是分別約有 68% 、 95% 和 99.7% 的數據在距平均數一個、二個和三個標準差的範圍內。這三個數字是如何計算出來呢？有修讀單元一的同學可能知道查表便可以得出，但在歷史上，當正態分布還未完全成熟時，已經有數學家利用數學分析的方法找出這些數字了。

在17世紀數學家柏斯卡(Pascal)和費馬(Fermat)的書信中，他們引入了組合方法來處理概率問題，而計算組合數 C_r^n 便成了計算古典概率的重要步驟，但是當 n 的數值很大時，計算計算組合數 C_r^n 便很困難了。到了 18 世紀初，當時數學家棣莫弗(De Moivre)對於二項系數 C_r^n 的近似值作研究，他的研究能幫助計算出當 n 的數值很大時， $\frac{C_{n/2}^n}{2^n}$ 的近似值。事實上，二項分布和正態分布有著很密切的關係，考慮以下各圖像：



$n = 10$　$p = 0.5$　$n = 20$　$p = 0.5$　$n = 50$　$p = 0.5$

各圖像都是表示某個二項分布在 $p = 0.5$ 下的情況，而當 n 愈大，圖像愈接近一個完整的正態分布。在計算 C_r^n 時，需要計算階乘，但大數的階乘是很難計算出來的，所以棣莫弗希望找出一條式子可以方便快捷地計算出大數值 n 下 $\frac{C_{n/2}^n}{2^n}$ 的近似值，這個數其實是二項分布中的中間項，從這個中間項開始，他更可以估計其他項的數值。他的研究結果可以在他在 1738 年出版的著作：The Doctrine of Chances 中找到，現在將其中一些結果向大家作介紹。

我們都不會否認計算加數比計算乘數容易，只要利用對數函數便可把乘數變為加數，所以棣莫弗先考慮化簡 $\ln\frac{C_m^{2m}}{2^{2m}}$，此處的 $n = 2m$。利用牛頓的展式：

$$\ln\left(\frac{1+x}{1-x}\right) = 2\left(\frac{x}{1} + \frac{x^3}{3} + \frac{x^5}{5} + \cdots\right)$$，可先得出：

$$\ln\frac{C_m^{2m}}{2^{2m}}=\ln\frac{(m+1)(m+2)\cdots(2m)}{1\times2\times\cdots\times m\times2^{2m}}$$

$$=\ln\frac{(m+1)(m+2)\cdots(2m-1)}{1\times2\times\cdots\times(m-1)\times2^{2m-1}}$$

$$=\ln\left(\frac{m+1}{m-1}\times\frac{m+2}{m-2}\times\cdots\times\frac{2m-1}{1}\times\frac{1}{2^{2m-1}}\right)$$

$$=\ln\frac{1+\frac{1}{m}}{1-\frac{1}{m}}+\ln\frac{1+\frac{2}{m}}{1-\frac{2}{m}}+\cdots+\ln\frac{1+\frac{m-1}{m}}{1-\frac{m-1}{m}}-(2m-1)\ln2$$

$$=2\left[\frac{\left(\frac{1}{m}\right)}{1}+\frac{\left(\frac{1}{m}\right)^3}{3}+\frac{\left(\frac{1}{m}\right)^5}{5}+\cdots\right]+2\left[\frac{\left(\frac{2}{m}\right)}{1}+\frac{\left(\frac{2}{m}\right)^3}{3}+\frac{\left(\frac{2}{m}\right)^5}{5}+\cdots\right]$$

$$+\cdots+2\left[\frac{\left(\frac{m-1}{m}\right)}{1}+\frac{\left(\frac{m-1}{m}\right)^3}{3}+\frac{\left(\frac{m-1}{m}\right)^5}{5}+\cdots\right]-(2m-1)\ln2$$

$$=\frac{2}{m}\left[1+2+\cdots+(m-1)\right]+\frac{2}{3m^3}\left[1^3+2^3+\cdots+(m-1)^3\right]$$

$$+\frac{2}{5m^5}\left[1^5+2^5+\cdots+(m-1)^5\right]+\cdots-(2m-1)\ln2$$

要再化簡下去，我們需要知道如何計算 $1^k+2^k+\cdots+n^k$ ，這是個很古老的數學問題，數學家雅各布·白努利(Jacob Bernoulli)給出公式：

$$1^k+2^k+\cdots+n^k=\frac{n^{k+1}}{k+1}+\frac{n^k}{2}+\frac{1}{2}C_1^kB_2n^{k-1}+\frac{1}{4}C_3^kB_4n^{k-3}+\cdots\ ,$$

而 $B_2=\frac{1}{6}$ ， $B_4=-\frac{1}{30}$ ，….等是白努利數。跟著的步驟需要利用到大學程度的微積分，計算在此從略，而棣莫弗得出了結果：

$\ln\frac{C_m^{2m}}{2^{2m}}\approx\ln 2-\frac{1}{2}\ln(2m)-\left(1-\frac{1}{12}+\frac{1}{360}-\frac{1}{1260}+\cdots\right)$。當時另一位數學家史特靈(Stirling)計算出：$\ln\sqrt{2\pi}=1-\frac{1}{12}+\frac{1}{360}-\frac{1}{1260}+\cdots$，所以棣莫弗的估計可寫成一個漂亮的表示式：$\frac{C_{n/2}^n}{2^n}\approx\frac{2}{\sqrt{2\pi n}}$。

下一步棣莫弗想找出其他項的近似值，他考慮 $\frac{C_{\ell+n/2}^n}{2^n}$ 的數值，而 $-\frac{n}{2}\le\ell\le\frac{n}{2}$。他利用相同的數學技巧，得出：

$\ln\frac{C_{\ell+n/2}^n}{C_{n/2}^n}\approx-\frac{2\ell^2}{n}$ ，所以 $\frac{C_{\ell+n/2}^n}{C_{n/2}^n}\approx e^{-\frac{2\ell^2}{n}}=1-\frac{2\ell^2}{n}+\frac{4\ell^4}{2n^2}-\cdots$ 。

最後利用積分方法，可知：

$$\frac{C_{n/2}^n}{2^n}+\frac{C_{1+n/2}^n}{2^n}+\cdots+\frac{C_{\ell+n/2}^n}{2^n}\approx\frac{C_{n/2}^n}{2^n}\times\left(\ell-\frac{2\ell^3}{3n}+\frac{4\ell^5}{2\times5n^2}-\cdots\right)$$

$$\approx\frac{2}{\sqrt{2\pi n}}\left(\ell-\frac{2\ell^3}{3n}+\frac{2\ell^5}{5n^2}-\cdots\right)$$

。

從二項分布的性質，我們知道平均值是 np 而標準差是 $\sqrt{np(1-p)}$ 。若 $p=0.5$ ，平均值便是 $n/2$ ，而標準差則是 $\frac{1}{2}\sqrt{n}$ 。今我們想知道在距平均數一個、二個和三個標準差的範圍內的數據百分比，我們可以分別代入 $\ell=\frac{1}{2}\sqrt{n}$ 、 $\ell=\sqrt{n}$ 和 $\ell=\frac{3}{2}\sqrt{n}$ 而得知。例如當 $\ell=\frac{1}{2}\sqrt{n}$ ，

$$\frac{C_{n/2}^n}{2^n}+\frac{C_{1+n/2}^n}{2^n}+\cdots+\frac{C_{\ell+n/2}^n}{2^n}\approx\frac{2}{\sqrt{2\pi n}}\left(\ell-\frac{2\ell^3}{3n}+\frac{2\ell^5}{5n^2}-\cdots\right)$$

$$=\frac{2}{\sqrt{2\pi n}}\left(\frac{\sqrt{n}}{2}-\frac{2n\sqrt{n}}{3\times8n}+\frac{2n^2\sqrt{n}}{5\times32n^2}-\cdots\right)$$

$$=\frac{2}{\sqrt{2\pi}}\left(\frac{1}{2}-\frac{1}{3\times4}+\frac{1}{5\times16}-\cdots\right)$$

$$\approx0.341344$$

因此在距平均數一個標準差的範圍內的數據百分比約有 0.682688 。同理，棣莫弗計算出其餘兩個百分比的近似值為 0.95428 及 0.99874 。

利用現代的方法，這三個百分比的準確值應該是 0.682689 、0.95450 及 0.99730 。雖然棣莫弗計算這些百分率時出現了誤差，但他的估算已經相當不錯。同時，我們看見歷史上一些重要的數學結果，是需要不同數學家一同合作努力多年，才可以得出來，不是簡單的按下計算機便知道的！

參考資料：

1. Hald, Anders. *A History of Probability and Statistics and Their Applications Before 1750*.
   JohnWiley & Sons: New York. 1990.

2. Stigler, Stephen M. *The History of Statistics: the Measurement of Uncertainty before 1900*.
   Belknap Press of the Harvard University Press: Cambridge, MA. 1986.

邀請作品：分賭注問題

2012 至 2013 年度借調老師：朱立夫

分賭注問題是數學發展歷史上的一個著名題目。不少數學家對這個問題都感到興趣，他們想出不同的解答，經過近 200 年的努力，最終被數學家費馬[1]和柏斯卡[2]用組合的方法得出合理結果，而且為概率論奠下了基礎。

分賭注問題[3]：有 2 位賭徒甲和乙進行一連串比賽，每勝一局便可得 1 分，誰先得到 4 分便可以拿走各人先前給出的 24 個作賭注的金幣。但當甲得到 2 分而乙得到 1 分時，比賽突然終止。這時他們各人應該拿回多少個金幣才算合理呢？

Pacioli 是在文獻紀錄上最早提出解答的人，在 1494 年他簡單地按賭徒已得的分數作為分金幣的準則，所以他認為甲應得到所有金幣的 $\frac{2}{3}$ ，即甲和乙分別得回 32 和 16 個金幣。但後來有很多人批評這種分賭注的方法，舉例來說，如果他們先前協議的不是用 4 分來定勝負，而是用 100 分，那麼明顯地甲得的 2 分和乙得的 1 分相比起 100 分來都是很少的，甲不應該可得回比乙多出 1 倍的金幣。其後 Cardano

[1] 費馬 (Pierre de Fermat, 1601-1665) ，法國業餘數學家。
[2] 柏斯卡 (Blaise Pascal, 1623-1662) ，法國數學家。
[3] 分賭注問題又稱得點問題 (Problem of Points)

(1539) 、 Tartaglia (1556) 和 Forestani (1603) 等數學家都提出不同的解答，他們都考慮到各賭徒已勝出的局數和總局數，可惜他們的計算都是一些主觀想法，欠缺數學證明。及後有一賭徒 Chevalier de Mere (1654) 把這個問題向柏斯卡求教，柏斯卡和費馬互相討論，最後他們用不同的方法得到相同結果，而且他們的方法成了日後計算理論概率的基礎。

在一封 1654 年費馬寫給柏斯卡的信中，費馬利用了組合的方法來解決分賭注問題。他考慮到如果甲和乙繼續進行比賽，他們最多需要進行 $2+3-1=4$ 局便能分出勝負。這 4 局可以有以下的勝負組合：

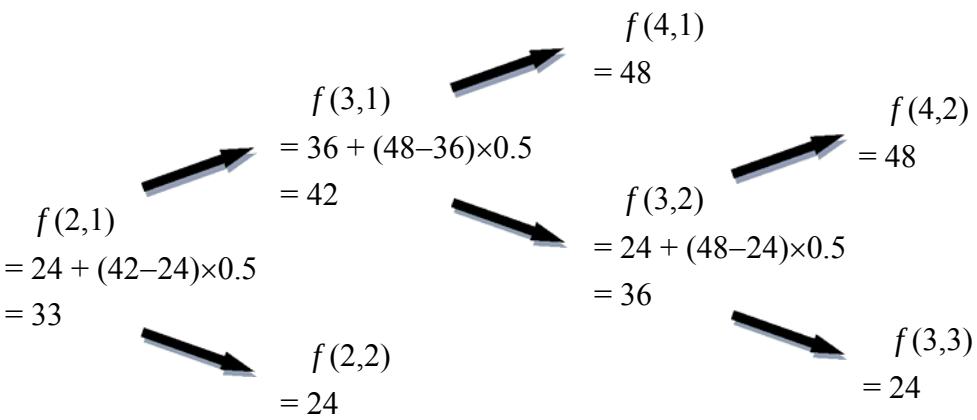
第一局	甲	甲	甲	甲	甲	甲	甲	甲	乙	乙	乙	乙	乙	乙	乙	乙
第二局	甲	甲	甲	甲	乙	乙	乙	乙	甲	甲	甲	甲	乙	乙	乙	乙
第三局	甲	甲	乙	乙	甲	甲	乙	乙	甲	甲	乙	乙	甲	甲	乙	乙
第四局	甲	乙	甲	乙	甲	乙	甲	乙	甲	乙	甲	乙	甲	乙	甲	乙
最後勝出者	**甲**	**甲**	**甲**	**甲**	**甲**	**甲**	**甲**	**乙**	**甲**	**甲**	**甲**	**乙**	**甲**	**乙**	**乙**	**乙**

在這 16 個可能的情況下，甲佔 11 個而乙佔 5 個，所以費馬認為金幣需按比例 11:5 來分配，即是甲可得回 33 個金幣。其實在這裡費馬已經引入了「等概」這個概念了。

在 1654 年 7 月 29 日柏斯卡的回信中，柏斯卡對費馬的方法表示認同，但他卻批評這種列舉法很煩瑣，所以他提出了

另一種較簡潔的解法。柏斯卡先考慮到如果甲和乙分別已經得到 3 分和 2 分，那麼下一局完成後，便會出現甲勝出或相方打和的兩種情況。若果是打和，甲應得回 24 個金幣，這是甲最起碼可以得到金幣的數目；但若果甲勝出，甲便可以獲得所有 48 個金幣了。利用「期望」值的想法，甲在這一刻(即甲和乙分別已經得到 3 分和 2 分)應得到的金幣數目是 $24+(48-24)\times\frac{1}{2}=36$。利用這個結果，我們可以推回甲和乙分別得到 2 分和 1 分的情況下，甲應得到金幣的數目。

用現代的語言，設 $f(m,n)$ 為在甲和乙分別已經得到 m 分和 n 分下，甲應得到金幣的數目，即剛才已計算出 $f(3,2)=36$。作以下樹形圖便可以方便找出 $f(2,1)$ 來：



這個結果和費馬的相同，即甲應可得到 33 個金幣。事實上，這種遞迴關係和二項係數之間的關係相似，而當時柏斯卡剛完成了一本關於二項係數的著作，所以他再利用二項係數的性質得出了以下的表達式：

$$
\begin{aligned}
f(2,1) &= 24 + 24 \times \frac{\frac{1}{2}C_2^4}{\frac{1}{2}C_2^4 + C_3^4 + C_4^4} \\
&= 24 + 24 \times \frac{3}{2^3} \\
&= 24 + 24 \times \frac{1 \times 3}{2 \times 4}
\end{aligned}
$$

第二項中的分數更有著一個漂亮的一般表達式：

$$\frac{\frac{1}{2}C_2^4}{\frac{1}{2}C_2^4 + C_3^4 + C_4^4} = \frac{1 \times 3}{2 \times 4}$$

甲尚欠 2 局而乙尚欠 3 局

$$\frac{\frac{1}{2}C_3^6}{\frac{1}{2}C_3^6 + C_4^6 + C_5^6 + C_6^6} = \frac{1 \times 3 \times 5}{2 \times 4 \times 6}$$

甲尚欠 3 局而乙尚欠 4 局

$$\frac{\frac{1}{2}C_4^8}{\frac{1}{2}C_4^8 + C_5^8 + C_6^8 + C_7^8 + C_8^8} = \frac{1 \times 3 \times 5 \times 7}{2 \times 4 \times 6 \times 8}$$

$\vdots$

甲尚欠 4 局而乙尚欠 5 局

相信柏斯卡只是從代數式中觀察出以上的結果，但其實我們可以用組合理論來理解這些表達式的。由於甲和乙還可以進行4局比賽，即是可產生勝負的組合數目有 2^4 。有利於甲的組合中可分為：甲勝 2 局、甲勝 3 局和甲勝 4 局的情況，這些組合所佔的數目為 $C_2^4 + C_3^4 + C_4^4$ 。所以

$$
\begin{aligned}
f(2,1) &= 48 \times \frac{C_2^4 + C_3^4 + C_4^4}{2^4} \\
&= 24 \times \frac{C_2^4 + C_3^4 + C_4^4}{2^3} \\
&= 24 \times \frac{\frac{1}{2}C_2^4 + \frac{1}{2}C_2^4 + C_3^4 + C_4^4}{\frac{1}{2}C_2^4 + C_3^4 + C_4^4} \\
&= 24 + 24 \times \frac{\frac{1}{2}C_2^4}{\frac{1}{2}C_2^4 + C_3^4 + C_4^4} \\
&= 24 + 24 \times \frac{3}{8}
\end{aligned}
$$

這個想法可帶出一般結果：在甲尚欠 m 分和乙尚欠 n 分的情況下，甲應可得到金幣的數目的比例是 $\frac{C_m^{m+n-1} + C_{m+1}^{m+n-1} + \cdots + C_{m+n-1}^{m+n-1}}{2^{m+n-1}}$ ，而這便是我們現在所熟悉的二項分佈。所以，最關鍵的地方是要考慮甲乙雙方尚要勝出的局數，才能公平地分配賭注。

在 1654 年 8 月 24 日柏斯卡給費馬的另一封信中，柏斯卡再次批評費馬的方法，因為柏斯卡認為用他的「期望」值想法可以處理更一般的情況，但費馬的方法卻不適用於 3 人進行比賽的情況，他舉出以下的例子說明他的論點：

有 3 人甲、乙和丙分別尚欠 1 分、2 分和 2 分便可獲勝，他們需要進行最多 3 局作出勝負，每局勝負的可能組合如下：

第一局	甲	甲	甲	甲	甲	甲	甲	甲	甲
第二局	甲	甲	甲	乙	乙	乙	丙	丙	丙
第三局	甲	乙	丙	甲	乙	丙	甲	乙	丙
最後勝出者	**甲**	**甲**	**甲**	**甲**	**甲**	**甲**	**甲**	**甲**	**甲**
					乙				
									丙
第一局	乙	乙	乙	乙	乙	乙	乙	乙	乙
第二局	甲	甲	甲	乙	乙	乙	丙	丙	丙
第三局	甲	乙	丙	甲	乙	丙	甲	乙	丙
最後勝出者	**甲**	**甲**	**甲**	**甲**			**甲**		
		乙		**乙**	**乙**	**乙**		**乙**	
									丙
第一局	丙	丙	丙	丙	丙	丙	丙	丙	丙
第二局	甲	甲	甲	乙	乙	乙	丙	丙	丙
第三局	甲	乙	丙	甲	乙	丙	甲	乙	丙
最後勝出者	**甲**	**甲**	**甲**	**甲**			**甲**		
					乙				
			丙			**丙**	**丙**	**丙**	**丙**

若單考慮對各人有利的結果，他們各人得回金幣的比例會是 19:7:7 。但有些情況會對 2 個人都同時有利的，若考慮平分，在這種觀點下，得回金幣的比例會是 16:5.5:5.5 ，但是柏斯卡利用自己的方法卻計算出 17:5:5 。若要從上表得出柏斯卡的結果，我們要在對有 2 個有利者的情況中選擇先勝出者：

第一局	甲	甲	甲	甲	甲	甲	甲	甲	甲
第二局	甲	甲	甲	乙	乙	乙	丙	丙	丙
第三局	甲	乙	丙	甲	乙	丙	甲	乙	丙
最後勝出者	**甲**	**甲**	**甲**	**甲**	**甲**	**甲**	**甲**	**甲**	**甲**
					~~乙~~				
									~~丙~~
第一局	乙	乙	乙	乙	乙	乙	乙	乙	乙
第二局	甲	甲	甲	乙	乙	乙	丙	丙	丙
第三局	甲	乙	丙	甲	乙	丙	甲	乙	丙
最後勝出者	**甲**	**甲**	**甲**	~~**甲**~~			**甲**		
		~~乙~~		**乙**	**乙**	**乙**		**乙**	
									丙
第一局	丙	丙	丙	丙	丙	丙	丙	丙	丙
第二局	甲	甲	甲	乙	乙	乙	丙	丙	丙
第三局	甲	乙	丙	甲	乙	丙	甲	乙	丙
最後勝出者	**甲**	**甲**	**甲**	**甲**			~~**甲**~~		
					乙				
			~~丙~~			**丙**	**丙**	**丙**	**丙**

所以最後柏斯卡的結論是，費馬的組合列舉方法不能處理這類情況。

費馬在 1654 年 9 月 25 日回信，信中指出柏斯卡修訂他的方法後得出的結果是正確的，但他卻不認同自己的方法錯誤，而費馬改用另一方法計算出比例 $17:5:5$ 。他考慮有一 3 面骰子，每一個面分別代表甲乙丙 3 人。對於甲來說，若要擲 1 次便勝出，在 3 個可能結果中佔 1 個；若要擲 2 次才勝出，在 9 個可能出現的結果組合中佔 2 個 (即“乙甲”和“丙甲”)；若要擲 3 次才勝出，在 27 個可能的結果組合中佔 2 個 (即“乙丙甲”和“丙乙甲”)。所以甲應得到所有賭注的 $\frac{1}{3}+\frac{2}{9}+\frac{2}{27}=\frac{17}{27}$ ，結果和柏斯卡是一致的。這個想法可帶出另一個 2 人情況下的一般結果：在甲尚欠 m 分和乙尚欠 n 分的情況下，甲應可得到的金幣數目的比例是

$$\frac{C_{m-1}^{m-1}}{2^m}+\frac{C_{m-1}^{m}}{2^{m+1}}+\cdots+\frac{C_{m-1}^{m+n-2}}{2^{m+n-1}}\ ,$$

而這便是負二項分佈。讀者可嘗試證明等式：

$$\frac{C_{m}^{m+n-1}+C_{m+1}^{m+n-1}+\cdots+C_{m+n-1}^{m+n-1}}{2^{m+n-1}}=\frac{C_{m-1}^{m-1}}{2^m}+\frac{C_{m-1}^{m}}{2^{m+1}}+\cdots+\frac{C_{m-1}^{m+n-2}}{2^{m+n-1}}\ 。$$

柏斯卡和費馬得出的分賭注比例，其實是各賭徒最終獲勝的概率。他們的書信中還有更詳細討論分賭注問題的其他性質，也有討論另外兩個投擲骰子的問題。他們的結果除了平息了分賭注問題的爭論，更誕生了一門新的數學理論：概率論。而柏斯卡和費馬的成功，在於他們不單考慮

各賭徒已經勝出的局數，更考慮到若果比賽繼續下去，兩名賭徒各自勝出的機會。某些數學家都有引入各賭徒獲勝機會的想法，不過他們只從直觀給予每局獲勝機會的數值，例如 Cardano 認為勝出第一局的機會是勝出第二局的機會的一倍，但這種計算沒有甚麼根據。另外，柏斯卡和費馬都應用組合方法來得出分賭注規則，而且柏斯卡還利用到二項係數的性質來計算組合數目，這成了日後計算理論概率的重要技巧。

參考資料：

1. Hald, Anders. *A History of Probability and Statistics and Their Applications Before 1750*. John Wiley & Sons: New York. 1990.
2. David Eugene Smith. *A Source Book in Mathematics II*. Dover Publications, Inc.: New York. 1959.

邀請作品：多吃巧克力更易獲諾貝爾獎？

—— 認識「關聯」與「因果關係」的分別

美國哥倫比亞大學一名學者 Franz H. Mersserli 曾發表一項學術研究結果，指人均巧克力消費量越高的國家，按人口平均計算的諾貝爾獎得主人數就越多。換言之，吃巧克力與獲取諾貝爾獎成正比。在被納入研究的 23 個國家當中，瑞士的人均巧克力消費量與人均諾貝爾獎得主人數皆是最多。該學者亦指出，巧克力中含有一種物質，能夠減慢與年齡相關的大腦衰老，亦能增強腦部活動能力。

在闡釋有關的數據分析時，有一點須要注意的是，兩件有關聯的事件不一定存在著因果關係。「關聯」與「因果關係」是兩個相似但不同的概念。若兩件事件有關聯，其發展變化的方向與強度均存在一定的聯繫。另一方面，若兩件事存有因果關係，換言之「原因」導致「結果」，兩者發生的時序不能顛倒。

以上述例子為例，人均巧克力消費量與人均諾貝爾獎得主人數兩者之間有正比關係，只代表兩者之間有關聯性。學者不能單憑觀察性研究的結果去引證多吃巧克力能有助人們進行優秀的學術研究，從而增加得到諾貝爾獎之類的學界殊榮的機會。

由於該研究的數據反映不了「吃巧克力」與「取得諾貝爾

獎」發生的時序，無法證明「吃巧克力」發生在「取得諾貝爾獎」之前，因此，從時序性的角度來看，該研究的數據結果不足以證明人們吃朱古力是取得諾貝爾獎的原因。

另一例子，某城市夏季時的冰淇淋銷量比冬季時錄得大幅上升，同期，該城市游泳池入場人數亦上升。吃冰淇淋與游泳兩者之間或存在關聯性。可是，我們不能因此推斷，吃冰淇淋與游泳兩者之間存在因果關係。

我們進行統計分析時，盡可能把所有有關聯的變項一併考慮。就此例而言，其他有關聯的變項包括室外氣溫、冰淇淋價格、游泳價格等。在進行關聯分析時，建議先控制其他變項不變，避免分析結果受其他變項的變化所影響。在全面考慮各個變項後，你會發現，室外氣溫與冰淇淋銷量兩者的關聯，比吃冰淇淋與游泳之間的關聯更加強。

參考資料：

1. Mersserli, Franz H. (2012). “Chocolate Consumption, Cognitive Function, and Nobel Laureates” *The New England Journal of Medicine* 367:1562-1564 http://www.nejm.org/doi/full/10.1056/NEJMon1211064

2. Eat chocolate, win the Nobel Prize? (Reuters) http://www.reuters.com/article/2012/10/10/us-chocolate-nobels-idUSBRE8991SS20121010

邀請作品：位置左右決定？

—— 有關地理位置的統計

你有沒有想過為何不同品牌的便利店或咖啡店，明明是直接競爭對手，但它們店鋪的位置卻又如此鄰近？在商業區，幾乎每隔幾條街、甚至可能只是相隔幾步，就能夠找到不同品牌的店鋪售賣相近的貨物。

我們可以將企業的競爭行為歸納成不同範疇：（一）貨物的種類及質素；（二）貨物的售價；（三）其他的交易成本，包括地理優勢。

以便利店為例，由於不同品牌經營的店鋪售賣的貨物的種類及質素是大同小異，而售價亦相近，所以地理優勢顯得尤其重要。如果可以搶到顧客容易到達的位置開業，或者令到顧客不用花費很久來搜尋便可得知店鋪的位置，都會使交易成本降低，間接提升競爭力和增加收益。

因此，不同品牌的便利店在選擇店鋪位置時都會份外留神，一方面希望得到人流多或「就腳」的位置，但又不願意分薄利潤。與此同時，便利店亦會列出店鋪的位置，務求令顧客容易得知。

有關地理位置的統計在這些時候便大派用場。現在利用電子地圖（例如 Google Map），不少品牌的便利店或咖啡店都

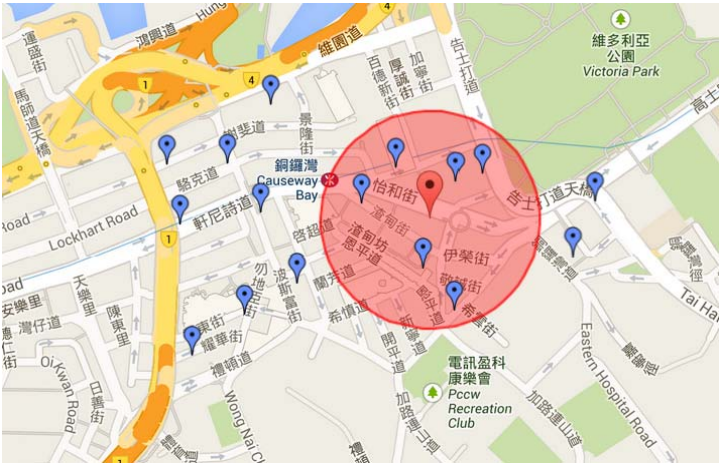
能夠在電子地圖上「點出」它們店鋪的位置，讓顧客不但得知店鋪的位置，更可以知道怎樣去和有多遠，從而選取在相隔幾條街的店鋪中，哪一間比較容易到達。在電子地圖上，一個位置可以透過人造衛星定位，並以緯度和經度表示，類似中學時候學過的 XY 座標，而現在亦有不少網站能夠替地址尋找其緯度和經度（即 Geocode）。

另一方面，正正因為能夠得知這些店鋪的位置，企業可以在選擇店鋪開業時，先去分析在某個位置方圓特定距離之內，到底有多少競爭對手已經營業的店鋪。以下面虛構例子為例，地圖上藍色的是競爭對手已經營業的店鋪，而企業現在打算從兩個可行位置選擇一個開業，便可以考慮它們各自方圓 200 米內，有多少個藍色標記。從圖中所見，**位置一**有 6 個藍色標記，而**位置二**只有 3 個，企業就要想想**位置一**所多出的人流或者較便宜的租金，能否抵消較多競爭對手的不利因素。

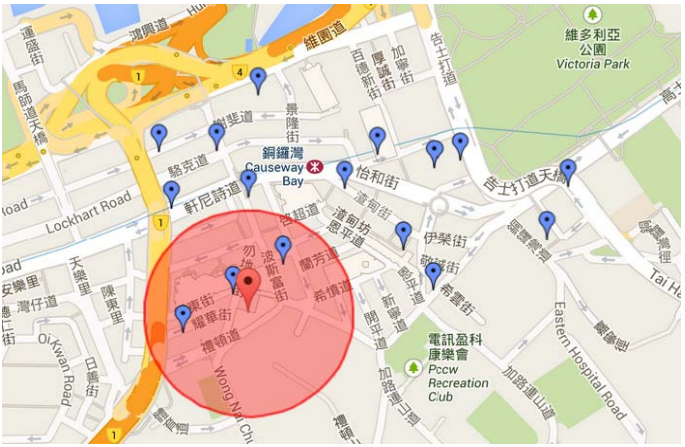
以相近的概念舉一反三，我們可以想出有關地理位置的統計還有不少應用的地方。但是在應用時，要留意我們有興趣的數據（例如競爭對手店鋪的數量），並不單與地理位置有關係、甚至未必與地理位置有關係，所以分析時要儘可能考慮其他有關聯的變項，否則便會把世界簡化為「地理位置決定一切」。與此同時，我們亦應注意地理資料和其他統計數據一樣，未必能夠輕易地得到，而且亦可能有量度誤差，例如在偏遠的地區，這些情況特別嚴重。「缺失數據」是本年度的主題，如果存在這些問題，我們想想能夠

用哪些統計方法處理？

位置一方圓 200 米內有 6 個藍色標記（即競爭對手已經營業的店鋪）



位置二方圓 200 米內有 3 個藍色標記（即競爭對手已經營業的店鋪）



參考資料：

1. Google Maps JavaScript API 第 3 版
   https://developers.google.com/maps/documentation/javascript/?hl=zh-tw

2. 我哋喺邊 | 7-Eleven® Hong Kong
   http://www.7-eleven.com.hk/store-locator.aspx

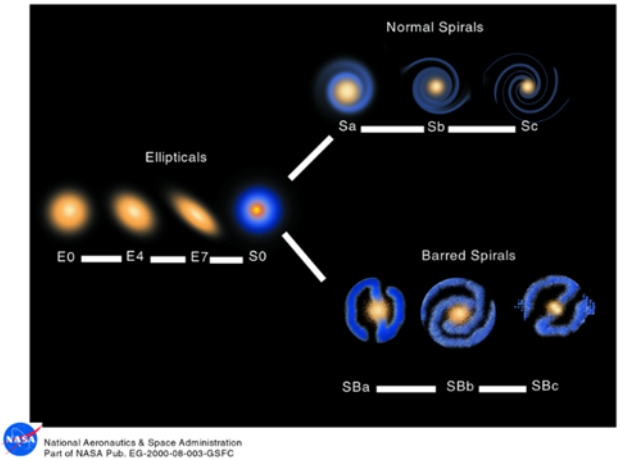
3. kyle's Blog – Google Map API-繪製雷達圖
   http://kylesheng.blogspot.hk/2013/09/google-map-api.html

邀請作品：群集分析與天文學

香港大學統計及精算系　關志威博士

引子

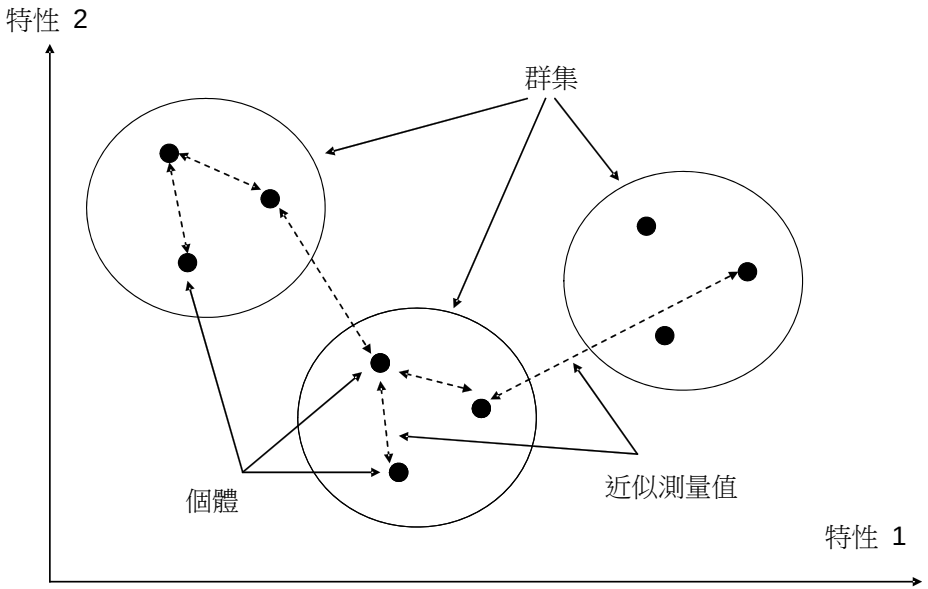
自二十世紀九十年代，統計學開始廣泛地應用在天文研究上。如天文圖像處理，星系分類及貝葉斯宇宙學等。其中一種應用的統計方法是群集分析。天文學家須要了解大量恒星，星系及天文群體的特性。他們會把天體分類為獨立的類別。每個類別所包含的天體都擁有相似的特質。 A.J. Cannon 把上百萬張的低解像度的恒星光譜照片分類。E. Hubble 將星系影像分類成橢圓，螺旋及棒旋型態的音叉圖。星系分類有助限制基本宇宙參數及探究於不同質量的暗物質暈中，星系形成的效率。在這文章，我們會介紹統計學中的群集分析及其簡單應用於天文學上。



(http://imagine.gsfc.nasa.gov/docs/teachers/galaxies/transparencies/trans3.html)

群集分析

群集分析目的是把個體分成不同群集。根據預定的方法分類後，每群集裡的個體都會很近似，而群集與群集之卻有著很大的分別。如下圖示：



群集分析是需要依據個體的一些變量。這些變量既能代表個體的特性，亦要可以用作個體與個體之間的比較。要形成群集結構，需要採用一個量度”接近”或”近似”的測量值，一般採用距離量度。一些常用距離為：

- 歐氏距離

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_p - y_p)^2}$$

- 歐氏距離平方

$$d(\mathbf{x},\mathbf{y}) = (x_1 - y_1)^2 + \cdots + (x_p - y_p)^2$$

- 統計學距離

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \mathbf{A} (\mathbf{x} - \mathbf{y})}$$

其中 **A** 一般為常數矩陣，如樣本共變異數矩陣的逆矩陣。

近似測量值容易被衡量所影響。數值變化大的變量比數值變化小的變量會對近似測量值做成較大影響。所以一般會把變量標準化。第 i 個數據的第 j 個變量標準化後是

$$z_{ij} = (x_{ij} - \bar{x}_j) / s_j$$

當中 $\bar{x}_j$ 是第 j 個變量的平均值，s_j 是第 j 個變量的標準方差。

其中一種群集分析是階層式分群法。階層式分群法不須要觀察所有可行的分群而找出一些合理的結果。盡管現代的電腦記憶容量大，運算快，也很少會檢視所有可能的分群結果。階層式分群法是以一連串的合併或分拆進行分群。這裡我們介紹合併式方法。

合併式方法首先把每一個個體各自成為一個群集。然後把接近的群集合併成一個新群集，直至所有個體合併到一個大群集裡。群集結果可以用一個二維的樹狀圖表示。階層式分群法其中一種演算法是連結法。步驟如下：

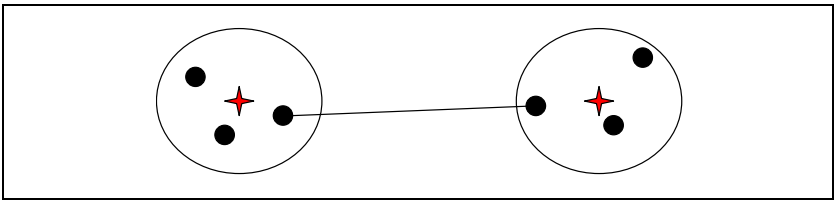
(1) 首先，N 個群集各自包含一個個體。然後計算出一個 $N \times N$ 的距離矩陣。
(2) 在矩陣中找出最接近的兩個群集。假設最接近的兩個群集，U 跟 V，的距離為 d_{UV}。
(3) 把群集 U 跟 V 合併成新群集 (UV)。更新距離矩陣。方法是先刪除群集 U 及群集 V 的行和列，然後加入群集 (UV) 跟其他群集間的距離的行和列。
(4) 重複第二及第三步 $N - 1$ 次至所有個體都包含在同一個群集內。

個體間的距離之前已經定義，但如何定義群集間的距離？不同的連結法有不同的定義。一些連結法舉例如下。單一連結法中，群集的距離定義為兩個群集中最接近兩點間的距離：

$$d(\mathrm{U},\mathrm{V}) = \min_{x_i \in \mathrm{U}, y_j \in \mathrm{V}} (x_i, y_j)。$$

由此，新群集 (UV) 跟群集 W 的距離可推算為

$$d_{(UV)W} = \min\{d_{UW}, d_{VW}\}。$$



例子

4 個個體的距離為

$$\mathbf{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

1. 第一步
   - 合併群集 1 及群集 2。
   - 新距離矩陣為

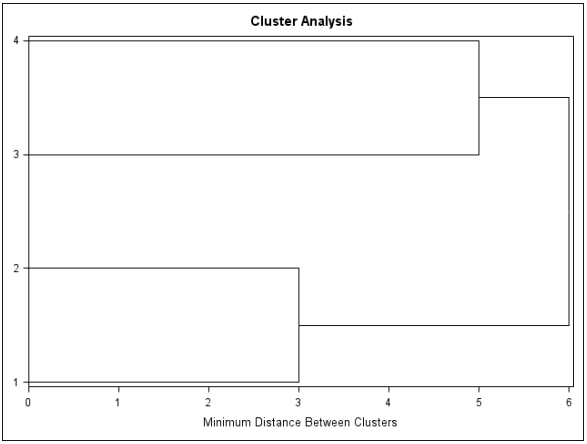
$$\mathbf{D} = \begin{array}{c} \\ (12) \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} (12) & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & 5 & 0 \end{bmatrix} \end{array}$$

2. 第二步
   - 合併群集 3 及群集 4
   - 新距離矩陣為

$$\mathbf{D} = \begin{array}{c} \\ (12) \\ (34) \end{array} \begin{array}{c} \begin{array}{cc} (12) & (34) \end{array} \\ \begin{bmatrix} 0 & \\ 6 & 0 \end{bmatrix} \end{array}$$

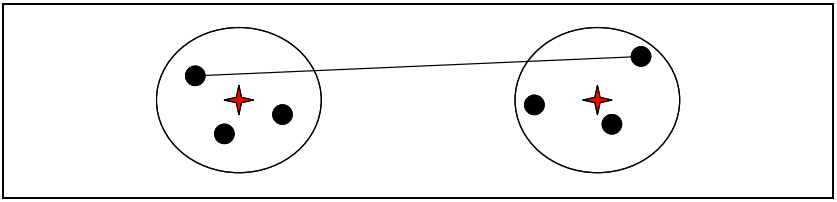
3. 第三步
   - 合併群集(12)及群集(34)成為群集(1234)

樹狀圖

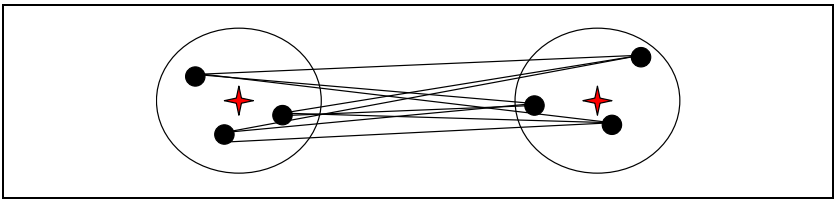


完整連結法中，群集間的距離定義為兩個群集中最遠兩點間的距離，

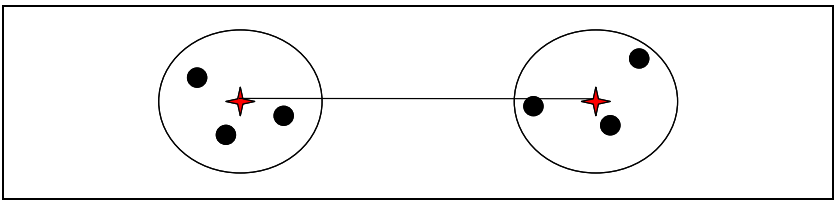
$$d(\mathrm{U},\mathrm{V}) = \max_{x_i \in \mathrm{U}, y_j \in \mathrm{V}} (x_i, y_j)。$$



還有平均連結法，群集間的距離定義為兩個群集間各點與各點間距離的平均值，



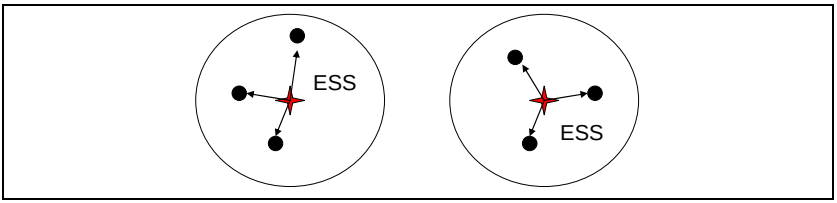
及中心連結法，群集間的距離定義為兩個群集中心點的距離。



另一類演算是法沃德法。首先定義群集內平方和為

$$ESS = \sum_{K} \sum_{i \in K} \sum_{j=1}^{p} \left(x_{ijK} - \bar{x}_{.jK} \right)^2$$

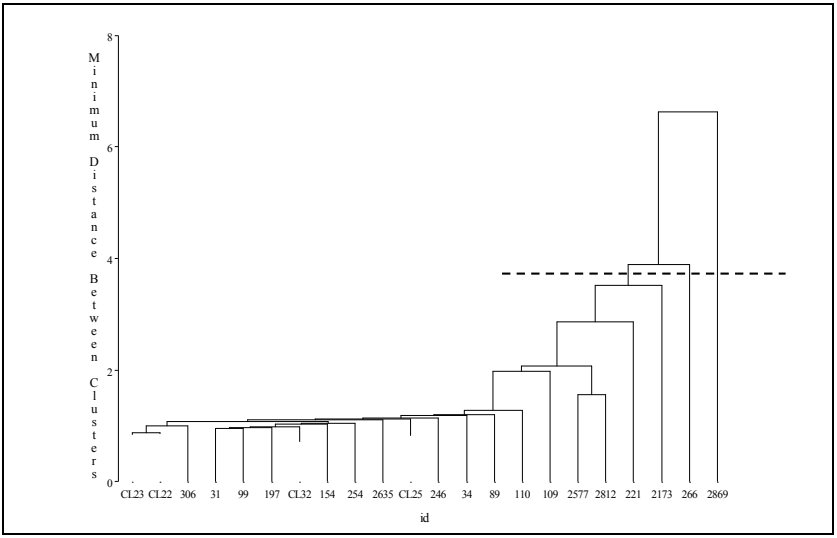
其中 $\bar{x}_{.jK}$ 是第 K 個群集中，第 j 個變量的平均值。合併哪一對群集，基於其合併後是否可把群集內平方和減至最少。其實沃德法也可視為一種連結法。



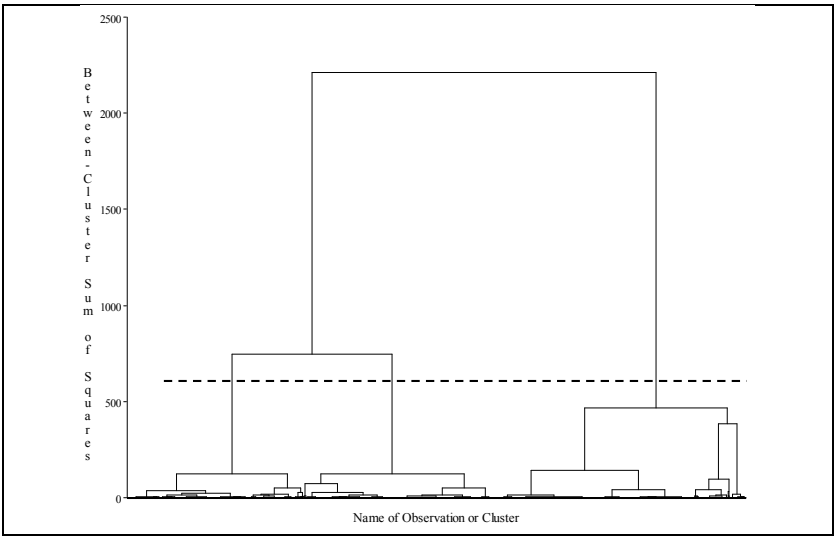
天文學上的應用

我們採用一個包含 1290 個星體的樣本(Feigelson & Babu, 2012)。星體的特性包括測光星表提供的五種測光波段：u (紫外線)， g (綠)， r (紅)， i 和 z (甚近紅外線) 波段。由於同一類的星體距離地球有著巨大差別，視星等並不是一個可靠的星體類別的指標。所以，我們採用五種波段的光度比例以消除距離的差別。由五種測光波段，我們得到四個顏色指標。而我們的分類研究，便建基於這四維空間。

我們的資料集包括四個變量－UG：紫綠比，GR：綠紅比，RI：紅紅比和 IZ：紅外比。我們把這四個變量標準化，然後採用歐氏距離。首先我們採用單一連結法，以下是部份的樹狀圖。



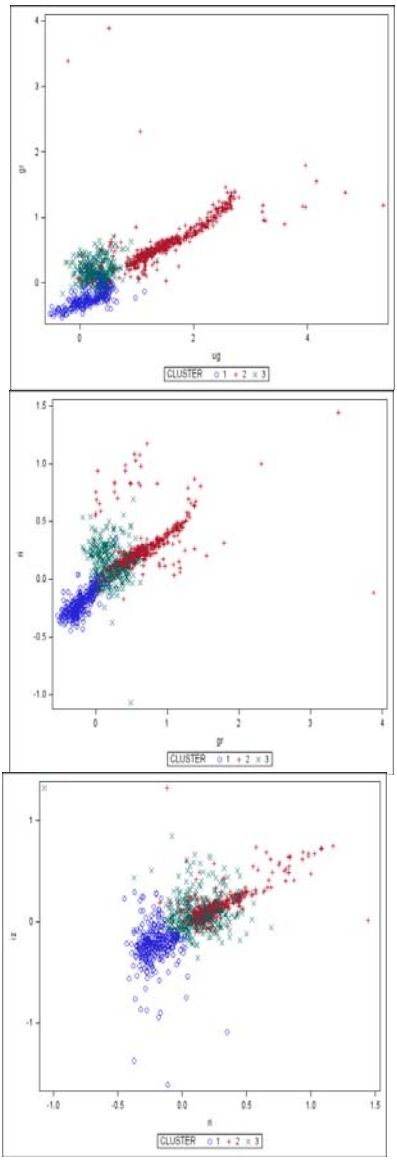
如果採納三個群集的分類，群集 1 有 1288 個星體，而群集 2 及 3 卻只有 1 個星體。由此可以看到單一連結法的一個普遍問題：單一連結法偏向於形成一個大群集。現在，我們採用沃德法，樹狀圖如下。



我們得到一個相對合理的分類。如果採納三個群集的分類，我們得到以下結果。

		平均值			
群集	數量	UG	GR	RI	IZ
1	394	0.21	-0.20	-0.20	-0.23
2	529	1.53	0.62	0.27	0.15
3	367	0.30	0.17	0.12	0.09

群集 1 所有顏色指標相對較小，群集 2 所有顏色指標相對較大，而群集 3 顏色指標在兩者之間。以下是顏色指標二維圖。群集分析清析地把星體分作 3 個群集。大至上，群集 1 是白矮星，群集 2 是類星體而群集 3 是主序星。


CLUSTER 1 2 3
CLUSTER 1 2 3
CLUSTER 1 2 3

參考資料：

1. Feigelson, E.D. & Babu, G.J. (2012) *Modern Statistical Methods for Astronomy*. Cambridge UP.
2. Johnson, R.A. & Wichern, D.W. (2007) *Applied Multivariate Statistical Analysis, 6th ed*. Pearson.
3. NASA. http://www.nasa.gov.

邀請作品：淺談大數據中的統計分析

香港大學統計及精算學系　楊良河博士

隨著近年科技不斷發展，例如社交網路的出現，與日常生活有關的資訊日趨電腦化，電郵、網誌、微博等非結構化的信息量高速增長，大數據（Big Data）的概念應運而生。難怪財爺於 2014 年初宣讀財政預算案中，亦提及「資訊科技的迅速發展，將世界帶到指尖。處理和分析資訊的能力，成為現代大型企業競爭優勢的重要一環。……政府會研究進一步使用物聯網(Internet of Things)、感應器(sensors)和大數據分析(big data analytics)技術，更有效地管理我們的城市。」

其實早在 2012 年初，美國政府已宣佈一項"大數據研究與開發倡議"（Big Data Research and Development Initiative）計劃。可以說現今社會已邁進大數據時代。在大數據時代中，掌握大數據就是掌握機遇，關鍵是能否從大數據中挖掘出潛在的有效訊息。

以下是一個真實個案。事件講述一位父親走進一間美國巨型連鎖超級市場 Target，嚷著要見經理。他緊握著一封信件，非常生氣地說道：「我的女兒收到了這推廣郵件！她還在唸高中，而你們竟然寄給她嬰兒服裝和嬰兒床的優惠券！你們是否要鼓勵她懷孕呢？」
超市經理見那位父親來勢洶洶，便立即向他口頭道歉。幾

天後，經理再遇到這位父親時，對方一反其氣焰，羞愧地告訴經理：「我已跟女兒詳談，原來有些家事我還未知道，那就是她將於幾個月後分娩。我應向你道歉才是。」原來女兒所收的郵件,是總公司憑大數據分析,預測某客人將會分娩, 而自動寄出推銷有關產品的郵件!

處理大數據的核心其實是懂得正確運用分析數據的工具。分析數據的過程離不開兩大步驟：(1) 數據搜集及處理；(2) 數據建模及測試。以 Target 為例，超市的統計師先識別孕婦經常購買的產品，並且收集每一位女客戶的個人資料及購物數據，跟著建立最有效的模型來進行分析，試圖找出懷孕者的消費購物規律，例如確立婦女在懷孕初期傾向於購買鈣，鎂或鋅的補充劑，以及無味的潤膚露。最後模型會為每位女客戶計算出一個"懷孕預測"的得分及估計她的預產日期。Target 就根據這些預測，在預測懷孕的某階段(如六個月後) 發送有關優惠券，鼓勵她們消費。

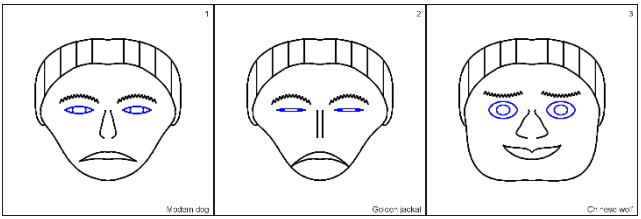
那麼 Target 是如何建立模型來預測懷孕？筆者相信 Target 是使用某一種分類方法(classification method)來預測誰是孕婦，並同時利用變項選取法(variable selection)找出哪些產品是最有預測懷孕的能力。常用的分類方法包括 Logistic 回歸，神經網絡，支持向量機器(SVM)和隨機森林(random forest) 等。Target 實際上用那一種方法來識別懷孕的客戶，這當然是商業機密。

另一項分析大數據的工具是數據視覺化(Data visualization)。由於數據變量的數目相當龐大，我們常常無法或無暇去理解枯燥的數據和複雜的分析結果。在這種情況下，數據視覺化就是傳達信息最快捷便利的方法。

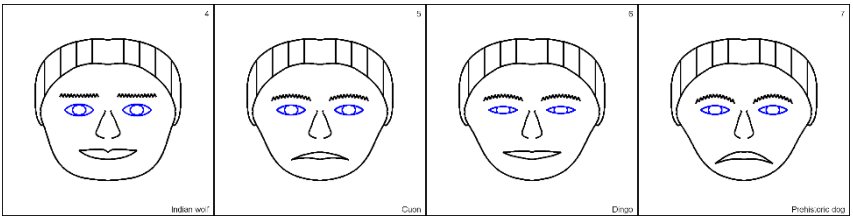
最簡單的數據視覺化即為傳統的統計圖表（statistical chart），比如散點圖（scatter plot）、直方圖（histogram）、圓形圖（pie chart）、折線圖（line chart）等。這些圖都只能表述一至兩個變量的數據。當變量數目很多時，我們如何能用圖去表示它呢? 早在 1973 年，統計學家 Herman Chernoff 就提出了利用類似人的面孔的 Chernoff 臉譜圖（Chernoff face）來呈現多維數據的方法。

我們以泰國考古隊發掘出的史前狗(prehistoric dog)骨頭為例。為了找出史前狗(prehistoric dog)的類別，研究員就搜集現存六種的犬科動物標本，然後量度每個標本的多维數據, 例如下顎部位的大小，下顎骨的闊度和高度等。圖一是根據每個品種下顎數據的平均值所繪制的 Chernoff 臉譜圖 (Chernoff faces), 其中(g) 為史前狗, (a) 到(f) 是已知品種 。

圖一. 七種犬科動物標本中下顎數據的 Chernoff 臉譜圖



(a) (b) (c)



(d) (e) (f) (g)

臉上的每個特徵就代表一個下顎變量的平均值。例如眼的大小代表下顎骨闊度的大小，鼻就代表下顎骨的高度，而面形就代表第一只臼齒的闊度等。由圖一可見史前狗(g)最類似(a)，即是泰國現今農村可見的村狗。

以上只是簡單介紹一些分析大數據的重要技術，而大數據產業才剛剛起步，相信未來會有更多創新的大數據分析技術應運而生，令大數據分析百花齊放。

二零一三至一四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
	楊琬婷女士，食物及衛生局

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