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 978-988-8159-24-6

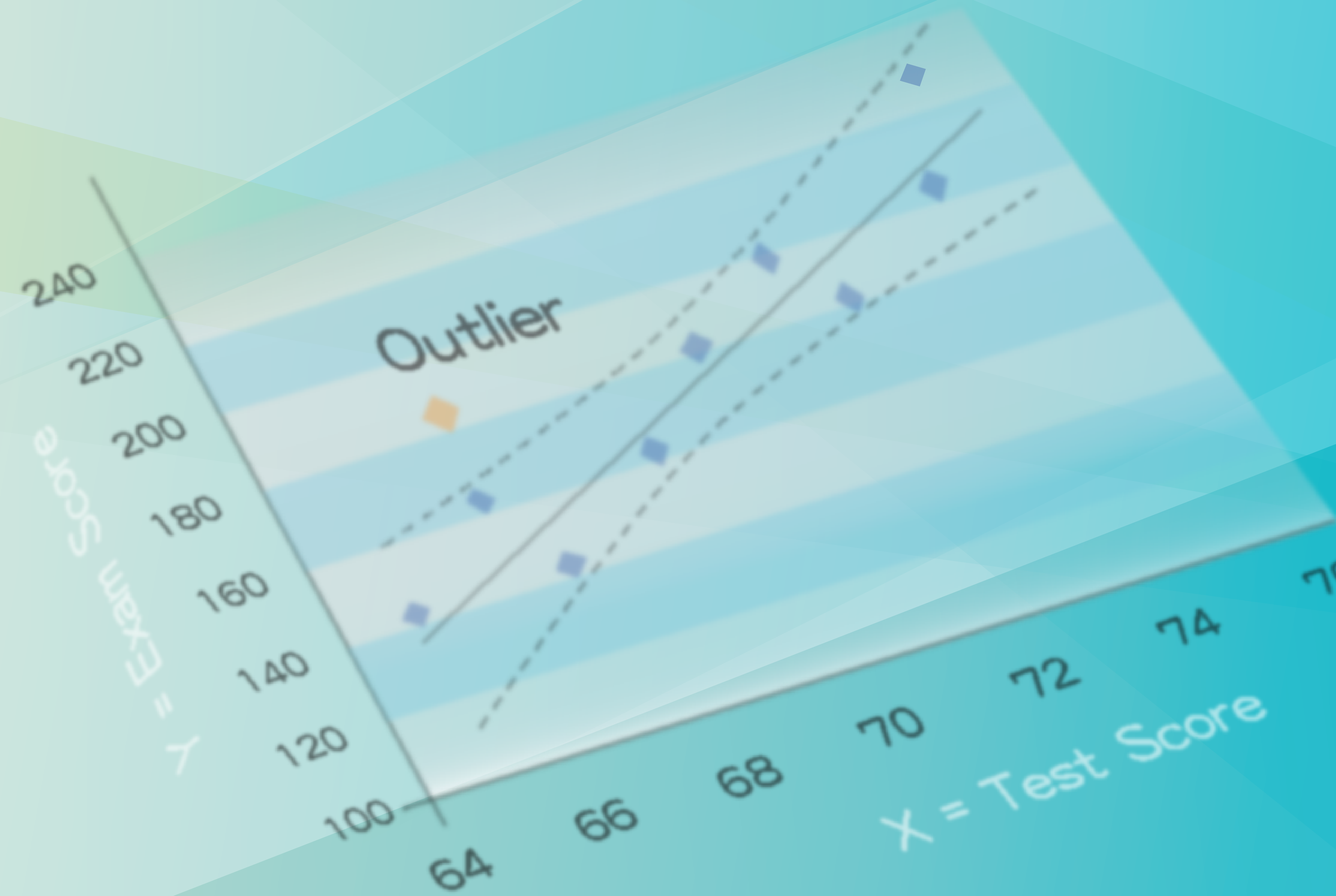
數學百子櫃系列(十五) 2012/13 中學生統計創意寫作比賽 作品集

教育局數學教育組

數學百子櫃系列(十五)

2012/13 中學生統計創意寫作比賽

作品集

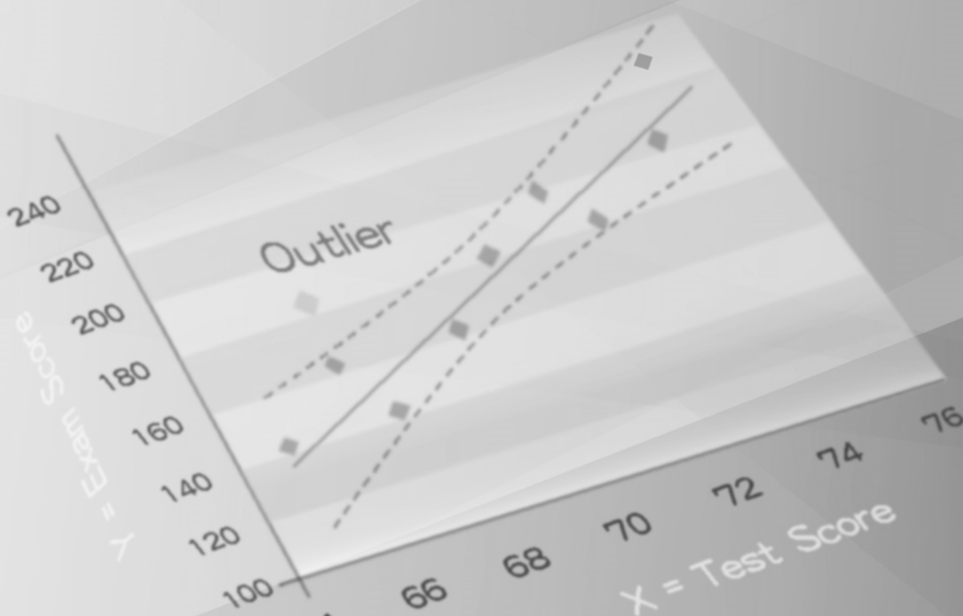


教育局
課程發展處數學教育組

數學百子櫃系列（十五）

2012/13 中學生統計創意寫作比賽

作品集



教育局
課程發展處數學教育組

ISBN 978-988-8159-24-6

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《2012/13 中學生統計創意寫作比賽作品集》，是這個系列的第十五冊。本書輯錄的文章，大部分是「2012/13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「2012/13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及課程主任，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文章。本書的故事大部分顯淺易懂，讓讀者很容易理解基礎統計中的概念。本書的出版，望能豐富統計教材的資源，讓統計教學變得更有趣，學生能就著不同的情境理解如何有效運用統計於生活上。

此書得以順利出版，實有賴籌備委員會各成員所付出的努力。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同工致以衷心的感謝。最後，更要多謝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的鼎力協助，審訂本書的內容，讓學生能夠閱讀更多有趣的文章，增強他們學習統計的興趣。

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

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

教育局課程發展處
數學教育組

前言

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

回顧過去的參賽作品，喜見同學們對統計概念有更深入的認識及掌握如何正確地運用統計。近年，得獎作品的質素亦有所提升。

本年度的比賽專題是「離群值」。多謝香港大學統計及精算學系講師關志威博士在比賽簡介會中介紹有關離群值的概念，並鼓勵同學在離群值這課題上發揮創意。

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。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。

本人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

2013年9月9日

序言

是次「中學生統計創意寫作比賽」共收到近 50 份參賽作品，經由教育局、政府統計處和香港大學統計及精算系內從事統計工作及教育的專業人士所組成的評審團進行評審。評審標準主要為創意和趣味性，及對統計知識的闡釋和正確應用。我謹此鳴謝一眾評審團成員，感激他們公正嚴謹地為我們選出優勝作品。

本書輯錄了今屆所有的得獎作品。能夠從芸芸參賽作品中脫穎而出，這些優勝作品都有其過人之處。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。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。

評審委員會首席評審員　張家俊博士

2013年9月10日

(空白頁)

目錄

(空白頁)

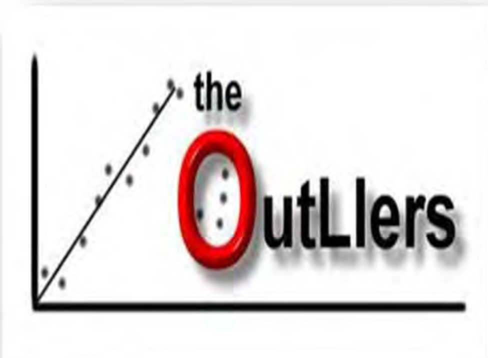
冠軍作品: **Mysteries in Ecology**

An Astonishing Truth or just a Game of Numbers?

School Name: Shau Kei Wan Government Secondary School
Student Name: Wu Tsun Wai, Chan Chui Shan, Yu Sin Ting
Level: Secondary 5
Supervising Teacher: Wong Pui Shan

Introduction

For a set of numerical data, an outlier refers to the number that is markedly smaller or larger than other values. By following the footprints of two ecologists, Kathy and Ernest, we are going to learn more about the causes of outlier and its identifications, as well as the mathematical ways to solve them. Let's begin our journey!

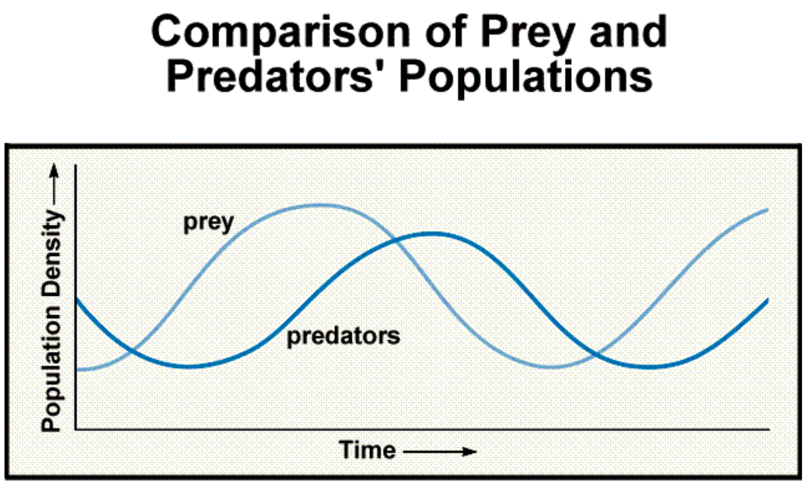


The story begins…

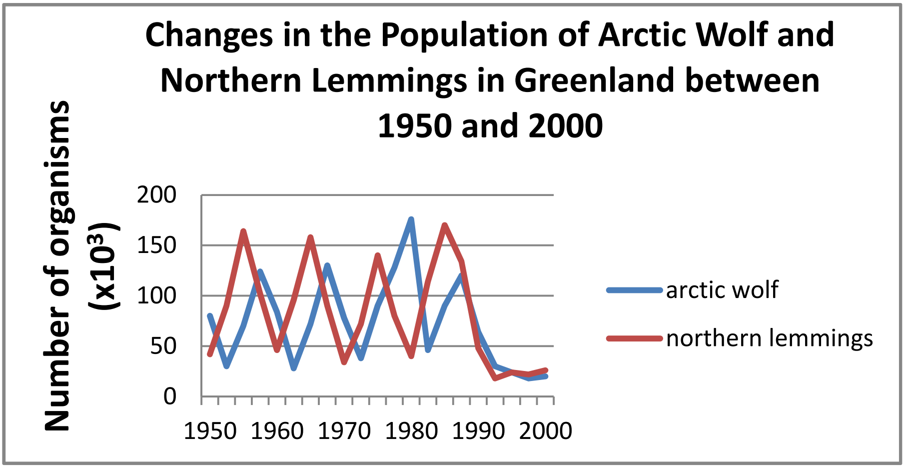
The Problem

Kathy and Ernest were two passionate amateur ecologists working in a research station in the northern part of Greenland. She had been studying the predator-prey relationships of arctic wolf and northern lemmings in Greenland before she found difficulty analysing the data collected.

Kathy: Mmm…. Why is that so strange? I have gathered some data focusing on the fluctuation of the population of arctic wolf and northern lemmings here, but there seems to be a …… slight problem……



Source: *San Diego State University*



Source: Government of Greenland, Ministry of Fisheries, Division on Wildlife Management

**We adopted the above data to estimate the number of arctic wolf and northern lemmings.*

Ernest: In a predator-prey relationship, the usual populations of the prey and predators fluctuate with an apparent time lag. In our study, the wolf and the lemmings are the predator and prey respectively. Nevertheless, the trend of their populations does not proceed as we have expected.

Kathy: The sections representing the change in the population of these two types of animals in the year of 1975 and from 1990 to 2000 just do not make sense. I cannot figure out what the data means.

Ernest: Shall we call for my friend, Herman for help? His aptitude to statistics will surely give us a boost in our progress.

Feeling desperate for help, they turned to Herman for advice.

Kathy: Here is a line graph (P.2) showing the populations of the two species.

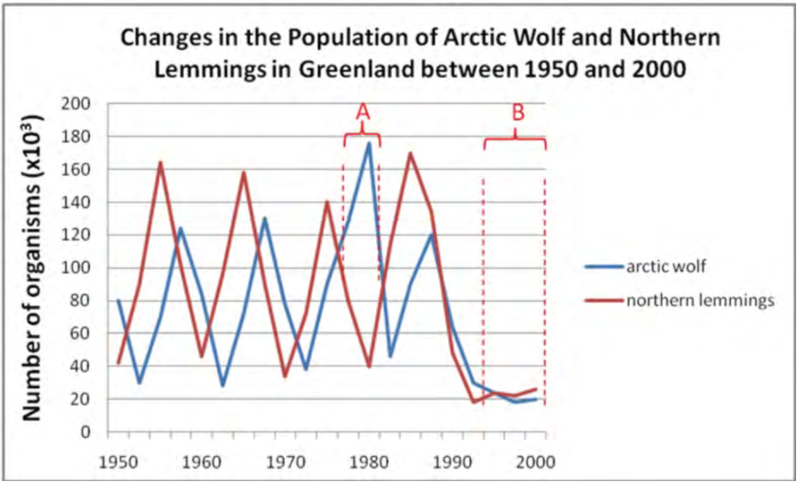
Herman: Let me see……Well, such a graph cannot tell anything but I am sure that your data is not without outliers.

Ernest: Sorry? What are outliers?

Herman: Outlier refers to things or phenomena that lie outside normal experience. In statistics, it is one that appears to deviate markedly from the majority of the samples or data obtained.

Ernest: Oh, I see.

Fundamental Analysis



Kathy: My data shows some unusual changes which I have named those as Region A and Region B.

Ernest: How can we make use of the concepts of outliers?

Herman: Well, first and foremost, I want to have a look of the population of arctic wolf collected in the year of 1980.

Kathy: Alright. Here it is.

Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8	Site 9	Site 10
4	5	3	4	13	8	3	1	2	43
Site 11	Site 12	Site 13	Site 14	Site 15	Site 16	Site 17	Site 18	Site 19	Site 20
1	4	6	5	4	46	3	37	4	5

*Fabricated materials

People at that time used to estimate the population at different sites. The area of each site is about 100 km^2. By calculating the population density (population / km^2), we can find out the approximate number of organisms in one place, i.e. the formula is as follows,

Population density $\times$ Area = Estimated population

- Total Area of Greenland = 2166086 km^2
- Area of Ice-free waters = 410449 km^2
- Excluding the ice-free waters in Greenland, its area is about 1755637 km^2

“Spot the Outliers!!” - Learning the Basics

Herman: By simple calculations, we can obtain the following information:

Mean	Min	Q1	Median	Q3	Max	S.D.
10.05	1	3	4	6.5	46	13.73126

You may see that there are three numerical values that are far beyond the interquartile range, which is a range laying between Q1 and Q3. Despite an apparent observation, we still have dozens of methods to prove outliers.

Kathy: Honestly speaking, Ernest and I neither are interested in statistics nor have the skills to operate complex calculations. Could you suggest a decent method for us to identify outliers with ease?

Herman: By all means! We may adopt a suitable outlier checker. The initial step is to find out Q1, Q3, interquartile range (IQR). The first two have been calculated while the IQR is 6.5-3=3.5. Next, we set a fence. It consists of an upper fence and a lower fence. Together, they

	construct an outlier checker. Data not lying between the fences are considered outliers. One definition of outlier is to multiply the IQR by 1.5, and 3.5×1.5=5.25. After that, we can find out the lower fence and the upper fence.
Ernest:	3-5.25 < 0 and 6.5 + 5.25=11.75. As the data must be positive, the lower fence and the upper one are 0 and 11.75 respectively, right?
Herman:	You got it. Outliers are any values that are 12 or above.
Kathy:	Based on the results, we can now eliminate the outliers obtained at site 5, 10, 16 and 18.
Ernest:	And a new set of information is obtained.

Site 1	*Site 2*	*Site 3*	*Site 4*	~~*Site 5*~~	*Site 6*	*Site 7*	*Site 8*	*Site 9*	~~*Site 10*~~
4	5	3	4	~~13~~	8	3	1	2	~~43~~
Site 11	*Site 12*	*Site 13*	*Site 14*	*Site 15*	~~*Site 16*~~	*Site 17*	~~*Site 18*~~	*Site 19*	*Site 20*
1	4	6	5	4	~~46~~	3	~~37~~	4	5

Mean	*Min*	*Q1*	*Median*	*Q3*	*Max*	*S.D.*
3.875	0	3	4	5	8	1.727534

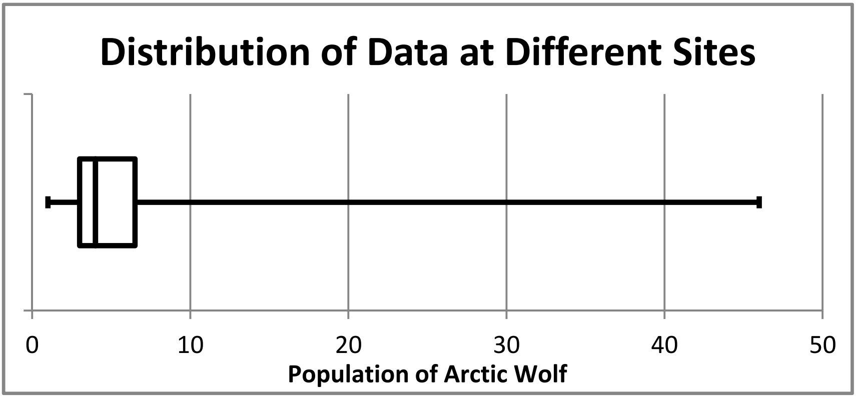
Kathy: Let's compare the data before and after outlier exclusion.

Estimated population before the amendment = $(10.05/100) \times 1755637 \approx 176000$

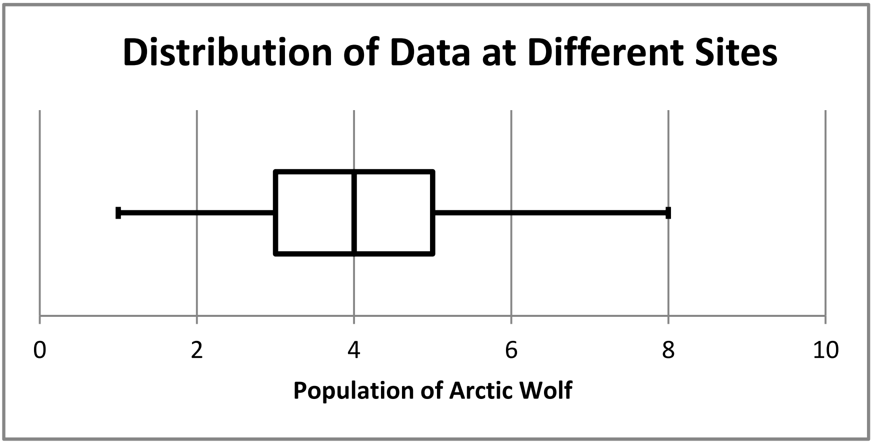
Estimated population after the amendment = $(3.375/100) \times 1755637 \approx 59300$

Wow! That's a big difference!

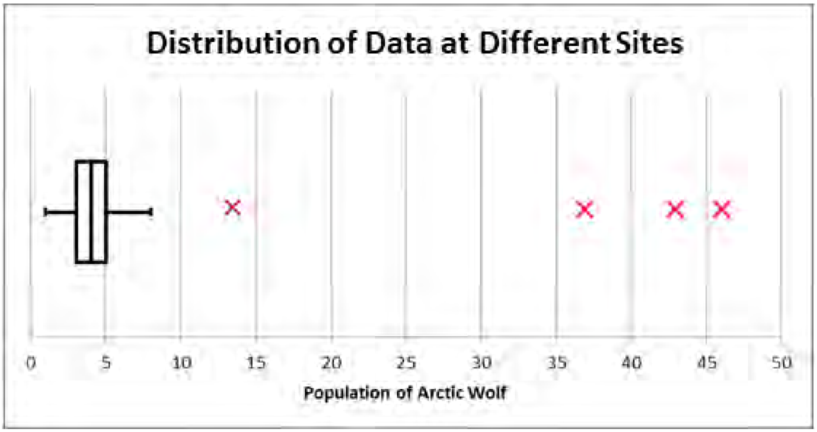
Herman: Yes, without identifying the outliers, the box plot expressing the data will be:



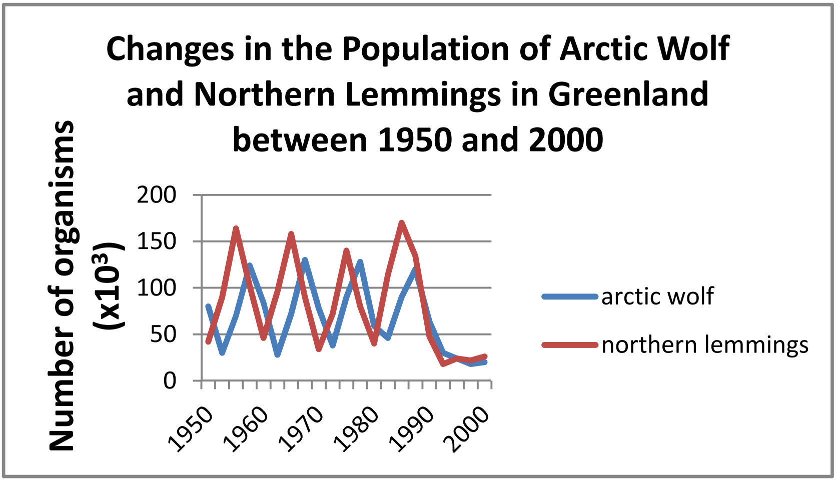
But with an outline checker we can identify the outliers and plot a decent graph.



Nevertheless, as we have excluded the outliers from our analysis, we must state that in the report. In addition, you should show them in the box plot like this:



Ernest: Hey! I have just remade our line graph with the new set of data. This time, the trend looks smooth enough.


Changes in the Population of Arctic Wolf and Northern Lemmings in Greenland between 1950 and 2000
Number of organisms (x10³)
200
150
100
50
0
1950
1960
1970
1980
1990
2000
arctic wolf
northern lemmings

A Suitable Method

Kathy: Thanks Herman! I really appreciate your effort. However, I really want to know why the outliers in the above case can be addressed simply through exclusion.

Herman: That's easy enough. Don't you notice the problem of the original data?

Kathy: Well, the data is not really even, with several extreme cases.

Herman: You got the point. There must be something wrong with the sampling method that people at that time used. Wolves are not fussy eaters. Some research sites might have been chosen with a sheepherding ground which attracts a lot of wolves! (laugh). But anyway, as the number of wolves should be nearly the same in Greenland where everywhere has a similar habitat, I just eliminate the outliers. I think in statistics, you really need to take every case into separate consideration and use the appropriate statistics-analysing tactics.

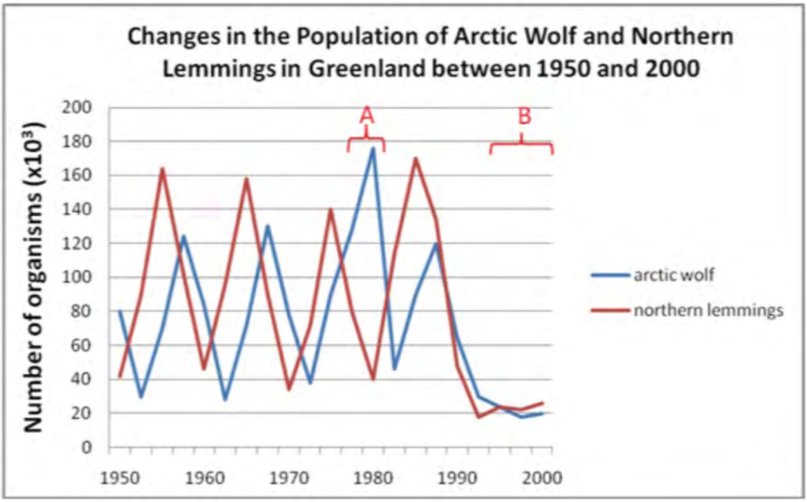
Kathy: How can we eradicate this discrepancy then?

Herman: I am not a professional in ecology. But it is important that people pick random sites for investigations and do increase the size of investigation to minimize sampling errors.

Kathy: I see what you mean. Anyway, now I know that the first function of outliers is to remove inaccuracies in statistics.

Observe. Think. Analyse. Conclude.

Kathy: The first part of the problem is solved. Now it is time we dealt with the second one. The population of lemmings has never proliferated since its decrease in number during the early 90s. What is even stranger is the population of arctic wolves. There is a plunge which suggesting that the ecosystem in Greenland may have experienced some degree of disruption.



Ernest: I have just analysed relevant data using the outlier checker method. Let me show you my work, *Fluctuation of population of northern lemmings from 1850 to 2000*

	Site 1	*Site 2*	*Site 3*	*Site 4*	*Site 5*	*Site 6*	*Site 7*	*Site 8*	*Site 9*	*Site 10*
1985	6	10	13	12	6	9	10	13	12	6
1990	6	2	3	3	1	1	1	4	2	1
1995	5	2	0	2	0	1	1	1	0	0
2000	7	3	0	2	0	1	0	2	0	0
	Site 11	*Site 12*	*Site 13*	*Site 14*	*Site 15*	*Site 16*	*Site 17*	*Site 18*	*Site 19*	*Site 20*
1985	11	12	9	6	8	4	10	13	12	12
1990	3	2	8	2	3	1	5	1	2	4
1995	2	0	7	1	1	0	1	0	0	1
2000	2	0	7	1	1	0	2	0	0	2

Likewise, I obtained a set of information.

	Mean	*Min*	*Q1*	*Median*	*Q3*	*Max*	*S.D.*
1985	9.7	4	7.5	10	12	13	2.758623
1990	2.75	1	1	2	3.25	8	1.840516
1995	1.25	0	0	1	1.25	7	1.757128
2000	1.5	0	0	1	2	7	2.061553

I managed to locate the outliers (as highlighted above) and I think I could calculate the approximate population of lemmings in Greenland, using the above data. Nevertheless, I found that even if I exclude the outliers when plotting the graph, the population of northern lemmings does not show any token of a rebound. Can you give an account for this, Herman?

Herman: I am afraid that your worries are becoming a fact. To analyse the situation, you first need to think wisely. Kathy has made a remarkable observation, saying that both populations drop. As the two populations are correlated under a predator-prey relationship, something happened on the natural habitat. I do not think the result is due to human or experimental errors.

Kathy: Perhaps it is due to human activities caused by the development of those places. Pollution and exploitation of the nature may have resulted in this. Otherwise, it may be a spread of disease that has fuelled this.

Ernest: I still have one more question. There are some sites with a constant population. Why is that?

Herman: That is another use of outliers. The outliers in this case unveil that there are some regions unaffected by the drop. In these cases, outliers may carry significances and should not to be ignored. You ought to do some follow-ups with that.

Kathy: Ah! Therefore a good data analysis must be a balanced consideration of the data collected and the real situation. We should keep sober when doing analyses and combine the reality with a suitable analysing method.

A Little Reflection:

Science and mathematics are closely related. Statistics is a big area that both subjects will come across. In this project, we combine elements of both sides into one written report, with a view to stretching our thoughts. This is a topic that we hammered out after a long time of discussion. We hope our effort can inspire everyone and let them know how extensively statistics can be applied in our life. Thank you.

References:

1. Area of Greenland
http://en.wikipedia.org/wiki/Greenland#ref_ddd

2. Government of Greenland: Ministry of Fisheries --Division on Wildlife Management
http://uk.nanoq.gl/

3. Box-and-Whisker Plots: Interquartile Ranges and Outliers
http://www.purplemath.com/modules/boxwhisk3.htm

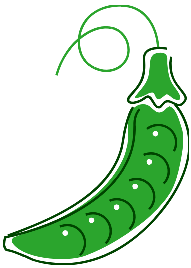
4. Predator-prey relationship between arctic fox and lemming
http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2656.1999.00258.x/full

5. Arctic wolf
http://en.wikipedia.org/wiki/Arctic_wolf

6. *Kjeld Hansen* (Jan 2002). *A Farewell To Greenland's Wildlife*. Baeredygtighed Press.

亞軍作品：如何運用統計學種神奇豌豆

學校名稱：荃灣官立中學
學生姓名：曹惠琪、林建宏、李錦輝
級別：中五
指導老師：陳栢垣老師



引言

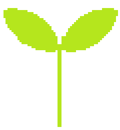
天氣與我們生活息息相關，尤其是從事農業活動的人。我們希望運用統計學的方法，包括幾何分佈、離群值和二項分佈，探討各種天氣的規律。我們使用香港天文台的數據，借一個故事，初窺統計學入門。

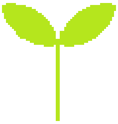
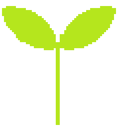
故事發生於 2013 年 6 月，主角是愛麗絲，她和母親住在香港的一處郊區。她們生活非常困苦，家中唯一的財產是一頭母牛，她們只能靠賣牛奶勉強賺錢維生。

一天，愛麗絲在家附近玩耍時，一位身穿白袍的老婦拍了她的肩膀一下，然後說：「小女孩，我把這兩顆神奇的**豌豆種子**送給你。」

愛麗絲歡天喜地拿著兩顆種子，回到家中，打算把它們好好種植，等到收成後便拿去賣，可是她不知道怎樣種植豌豆。

正當愛麗絲為如何種植豌豆而煩惱時，「鈴 — 鈴 —」，她的電話忽然響起，於是她拿起電話一看，發現了以下的短訊：

 豌豆婆婆	小女孩，我是給你神奇豌豆的婆婆。 你懂得種植這種神奇豆子嗎？
 愛麗絲	不懂得啊！可以教我怎樣種植這些豌豆嗎？

 豌豆婆婆	當然可以。這種豌豆生長條件十分特別，它的生長條件只會視乎過去 16 年的天氣數據，生長的月份必須是六月，只會在不是炎熱的晚上（簡稱涼夜）生長，而且必須是連續五晚都是涼夜，才能成功發芽。 如果當中一晚是天氣炎熱的（簡稱熱夜），豌豆便會死掉。 (注：“熱夜”－日最低温度不低於 $28^{o}C$ [1])
 愛麗絲	需要連續五晚都是涼夜！？這根本不可能吧！現在是夏天呢！
 豌豆婆婆	你應該沒有學過幾何分佈 (Geometric Distribution) 吧？ 在講述幾何分佈前，我先向你解釋伯努利試驗(Bernoulli Trial)。伯努利試驗是指只有兩種可能結果的單次隨機試驗。以一個晚上為例，只能是熱夜或不是熱夜，樣本空間就是集合{“熱夜”, “涼夜”}，因此是一種伯努利

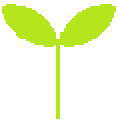
	試驗。 現在講幾何分佈。在伯努利試驗中，X 是得到一次成功所需要的試驗次數。在得到第一次成功之前所經歷的失敗次數 $Y = X - 1$。如果每次試驗的成功概率是 p，那麼經歷 k 次失敗後得到一次成功的概率是 $P(Y = k) = p(1 - p)^k$，其中 $k = 0, 1, 2, 3, \ldots$。 至於 Y 的期望值 $E(Y) = \dfrac{1-p}{p}$。(見附頁 1) 因此我們可以計算出現熱夜前所出現的涼夜數目 Y 的期望值。假設每晚的天氣相互獨立，並參考天文台的數據，過去 16 年(1997–2012 年)每年六月份熱夜日數如下[1]：

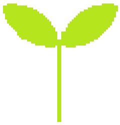
年份	6 月熱夜日數
2012	6
2011	5
2010	5
2009	2
2008	0
2007	3
2006	3
2005	5
2004	7

年份	6 月熱夜日數
2003	5
2002	8
2001	2
2000	6
1999	2
1998	10
1997	4

平均熱夜日數是 4.5625。由於六月有 30 天，因此 2013 年出現熱夜的估計概率為 $\hat{p}=\dfrac{4.5625}{30}\approx 0.152$。假設 $\hat{p}=p$, 那麼出現一晚熱夜前所出現的涼夜數目 Y 的期望值 $\mathrm{E}(Y)=\dfrac{1-\hat{p}}{\hat{p}}\approx$ 5.58 天。這代表如果你今天種下豌豆，在出現一晚熱夜前，該會有約五天半的涼夜，豌豆只需要連續五晚涼夜，而連續五晚涼夜的估計概率為$(1-\hat{p})^5=0.438$，概率有四成以上，因此可以值得一試種下豌豆。

愛麗絲根據豌豆婆婆的指示種下了第一顆種子，雖然往後連續五晚都是涼夜，但是種子沒有發芽，而且死掉了。愛麗絲於是又向豌豆婆婆求救。

 愛麗絲	我種下豌豆後，連續五晚都是涼夜，但是豌豆卻沒有發芽！為甚麼？
 豌豆婆婆	小女孩，雖然最近連續五晚都是涼夜，但這年的六月降雨特別多。神奇豌豆只適合在六月正常的雨量下種植。本月總降雨量有 1400 毫米(此數字為配合故事而虛構)，比 2008 年 6 月總降雨量還要多，因此豌豆不能生長。
 愛麗絲	那如何界定為"不正常"雨量？



豌豆婆婆

這可以用**離群值**來解釋。離群值是指在數據中有一個或幾個數值與其他數值相比差異較大。離群值的判別有很多種方法，其中一種方法是 Chauvenet's criterion (音譯：肖維內)。首先，我們需要計算數據集的平均值 $\bar{x}$ 以及標準差 σ ，如要測試其中的某一數據是否離群值，可以計算該數據的標準分 $\frac{x-\bar{x}}{\sigma}$ ，用正態分佈找出 $P(|Z|>\frac{x-\bar{x}}{\sigma})$ ，若 $P(|Z|>\frac{x-\bar{x}}{\sigma}) < \frac{1}{2n}$ (n 為數據量)， 即代表該數據則可判為離群值 [2]。

翻查香港天文台的資料，以下是香港天文台過去 16 年六月份總雨量的資料 (神奇豌豆生長條件只會視乎過去 16 年的天氣數據 [3])：

年份	總雨量(mm)
2012	261.5
2011	435.6
2010	474.9
2009	341.8
2008	1346.1

年份	總雨量(mm)
2007	490.1
2006	469.2
2005	893.9
2004	144.7
2003	523.5
2002	237.6
2002	237.6
2001	1083.6
2000	443.3
1999	197.4
1998	814.5
1997	783.6

例如，新華網在 2008 年 06 月 27 日報導，該年 6 月「已經成為香港有紀錄以來最多雨的 6 月」[4]。

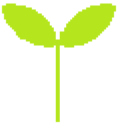
由以往的數據得出，6 月份的平均降雨量為 558.83mm，標準差 σ 為 327.28mm。2008 年 6 月的降雨量是 1346.1mm，標準分是 $\frac{1346.1-558.83}{327.28}=$ 2.40552。

從正態分佈，得出 $P(|Z|>2.40552)=$

	0.0161。 因為 $\frac{1}{2n} = \frac{1}{2(16)} = \frac{1}{32} = 0.03125$，而 $0.0161 < 0.03125$，所以 1346.1 是離群值。本月雨量比 2008 年 6 月時還要多，只能說你太倒霉了。或許你下年 6 月再試試種第二顆種子吧。
 愛麗絲	好吧。☺

2014 年 6 月，愛麗絲成功種植豌豆了，豌豆的藤蔓穿透雲頂。愛麗絲好奇想要往上爬，看看雲上的世界，但她擔心：如果我在攀爬的途中遇上危險，如雷暴的話，要怎麼辦呢？我會被雷擊嗎？

於是她再問豌豆婆婆。

 愛麗絲	豌豆婆婆，我的神奇豌豆長成了，我想看看雲上面的世界，但我擔心在攀爬的途中會遇上雷暴。前幾天一直都是好天氣，那麼接下來的幾天會不會快要打雷呢？
 豌豆婆婆	這要用上二項分佈(Binomial Distribution)了。 二項分佈是 n 個獨立的是與非試驗中成功的次數的離散機率分佈，其中每次試驗的成功概率為 p。 一般的二項分佈是這樣的：假設某事件 X 的發生概率為 p，而試驗做了 n 次。則 n 次中，某事件發生 k 次的概率為 $P(X=k)=C_k^n p^k(1-p)^{n-k}$ 其中 k = 0, 1, 2, …, n. 至於 X 的期望值 $E(X)=np$。(見附頁 2) 根據香港天文台關於 1981-2010 年的資料作為樣本，在 6 月份雷暴的樣本平均日數為 7.03 天，6 月有 30 天。所以樣本概率為

	$p = \frac{7.03}{30} \approx 0.234$。 由於二項分佈樣本的期望值同樣是 *np* (見附頁 3)，假設每天的天氣情況互相獨立，你需要 10 天才能爬上藤蔓頂部，估計在這十天內雷暴的日數為 10×0.234 = 2.34 天。在你爬上藤蔓頂部時，只會有 2.34 日遇上雷暴，風險不算太高。而且，遇上雷暴時，你仍可以嘗試藏在藤蔓裡躲避。你那就不如試一試吧！

愛麗絲計算好日子，終於在天朗氣清的一天爬上了藤蔓，到達了雲的上面。她看到一座宏大的城堡，便向內走去。城堡內坐著一個巨人，巨人正在數著巨大桌子上一袋袋的金幣。愛麗絲爬上桌面，跟巨人表明身份，巨人竟然對愛麗絲的智慧和勇氣深表敬佩，他答應給愛麗絲一袋金幣，並教導她做生意的秘訣。

愛麗絲離開城堡後把通往雲上的藤蔓斬了。愛麗絲嘗試再發短訊給豌豆婆婆，但豌豆婆婆已消失得無影無蹤。她運用從豌豆婆婆和巨人處學到的知識，包括運用統計學預測天氣的狀況而達至相應的用途，成為了一代的女富豪。

附頁:

(i) 計算幾何分佈 P($Y = k$) = $p(1 - p)^k$, k = 0,1,2…,的期望值

$$\mathrm{E}(Y) = \sum_{k=0}^{\infty} k \cdot p(1-p)^k$$

$$= p\sum_{k=0}^{\infty} k(1-p)^k$$

$$= p\sum_{k=1}^{\infty} k(1-p)^{k-1}(1-p)$$

$$= p(1-p)\sum_{k=1}^{\infty} k(1-p)^{k-1}$$

$$= -p(1-p)\sum_{k=1}^{\infty} \frac{d}{dp}(1-p)^k$$

$$= -p(1-p)\frac{d}{dp}\sum_{k=1}^{\infty}(1-p)^k$$

$$= -p(1-p)\frac{d}{dp}\frac{1-p}{1-(1-p)}$$

(注：等比數列無限和為 $S_\infty = \dfrac{a}{1-r}$, 當中 a 為首項，r 為數列相鄰兩數之比)

$$= -p(1-p)\frac{d}{dp}\frac{1-p}{p}$$

$$= -p(1-p)(-\frac{1}{p^2})$$

$$= \frac{1-p}{p}$$

(ii) 計算二項分佈 $P\left(X=k\right)=C_k^n p^k\left(1-p\right)^{n-k}$, k = 0, 1, 2,…, n, 的期望值

$$E(X) = \sum_{k=0}^{n} k\cdot C_k^n p^k\left(1-p\right)^{n-k}$$

$$= \sum_{k=1}^{n} k\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \quad (\text{當 } k=0,\ kC_k^n p^k\left(1-p\right)^{n-k}=0)$$

$$= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!}p^k(1-p)^{n-k}$$

$$= np\sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!}p^{k-1}(1-p)^{n-k}$$

$$= np\sum_{k=0}^{n-1} \frac{(n-1)!}{k!(n-(k+1))!}p^k(1-p)^{n-(k+1)}$$

$$= np\sum_{k=0}^{n-1} \frac{(n-1)!}{k!(n-1-k)!}p^k(1-p)^{n-1-k}$$

$$= np\ (\sum_{k=0}^{n-1} \frac{(n-1)!}{k!(n-1-k)!}p^k(1-p)^{n-1-k} = [p+(1-p)]^n = 1)$$

(iii) 計算二項分佈樣本的期望值

$$\mathrm{E}(S) = \mathrm{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \qquad (n \text{ 為樣本數量})$$

$$= \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

$$= \frac{1}{n} n(np)$$

$$= np$$

參考資料:

1. 自 1885 年(除 1940-1946 年外)香港天文台錄得的熱夜日數，香港天文台
   http://www.hko.gov.hk/cis/statistic/hngtday_statistic_c.htm

2. Chauvenet's Criterion
   http://www.spiritus-temporis.com/chauvenet-s-criterion

3. 每月天氣摘要，香港天文台
   http://www.hko.gov.hk/wxinfo/pastwx/mwsc.htm

4. 香港度過有紀錄以來最多雨的 6 月，新華網
   http://big5.xinhuanet.com/gate/big5/news.xinhuanet.com/newscenter/2008-06/27/content_8452165.htm

5. 自 1947 年香港的雷暴日數，香港天文台
   http://www.hko.gov.hk/cis/statistic/tsday_statistic_c.htm

6. New Trend Senior Secondary Mathematics, Module 1: Calculus and Statistics, Chapter 12 Discrete Probability Distributions, Chapter 13 Continuous Random Variables and Normal Distribution, Chapter 14 Parameter Estimation, Chung Tai Educational Press.

季軍作品: 得籃板，得天下？

學校名稱：順德聯誼總會李兆基中學
學生姓名：煜峻、周嘉豪、梁卓楠
級別：中五
指導老師：許俊江老師

引言

我們在日常生活中會遇到很多很特別的人和事，例如一個極端聰明的人、一場出人意表的球賽、一場意想不到的風暴，往往都吸引著我們。究竟，突出是好，還是壞呢? 越接近平均，是平衡，還是平平無奇？！這，就是我們都經常接觸到的－離群值。

(今日係社際籃球既開幕戰，由文社對信社，最後文社以 17 比 5 贏了信社)

(Cogor 由籃球場走到 Jack 同 Rabbit 身邊，坐下喝水)

Cogor : 啊，真想不到文社竟然贏到這比賽！

Jack : 贏 12 分不算險勝呀！

Cogor ：我們文社連一個籃板都搶不到，竟然能贏得比賽，真神奇啊！

Jack : 又如何？

Rabbit ：你不知道「得籃板，得天下」嗎，在一場比賽裏面，主宰籃板的球隊可以說是主宰了勝利。只要成功控制籃板，就可為隊友製造更加多進攻機會，甚至是防守籃板之後的快速反擊，這往往是勝負關鍵。

Jack : 真的嗎，我就認為命中率比較重要。

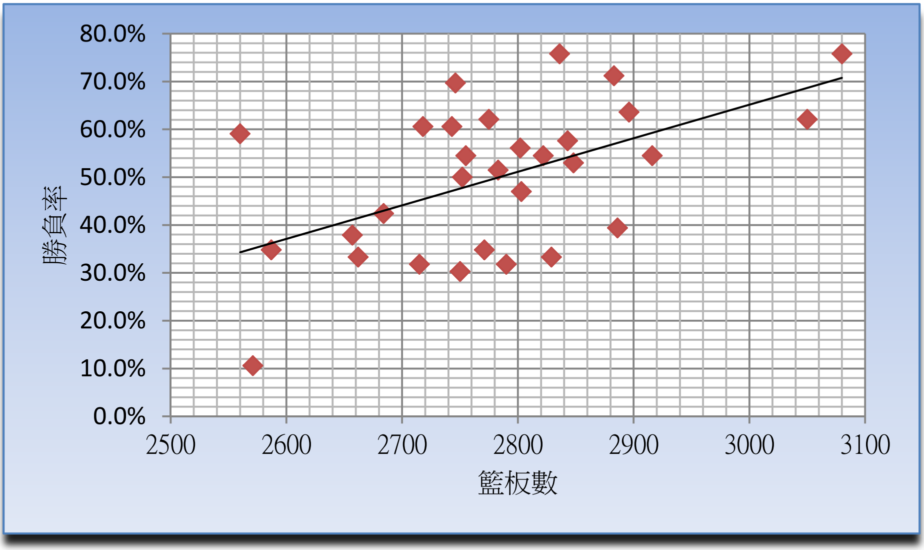
Rabbit ：真的，就像上一季的ＮＢＡ比賽。取得的籃板數目和他們勝出的概率是正相關，取得的籃板數目越多，勝出的概率就越大。你看，就像上季取得最多籃板的公牛一樣，和湖人、雷霆、馬刺等，都有很高的勝出率，是很好的例子。

隊伍	勝負率	籃板數
芝加哥公牛	**75.8%**	**3080**
洛杉磯湖人	**62.1%**	**3050**
猶他爵士	54.5%	2916
印第安那溜馬	63.6%	2896
明尼蘇達灰狼	39.4%	2886
奧克拉荷馬雷霆	**71.2%**	**2883**
費城 76 人	53.0%	2848
丹佛金塊	57.6%	2843
聖安東尼奧馬刺	**75.8%**	**2836**
薩哈拉門托帝王	33.3%	2829

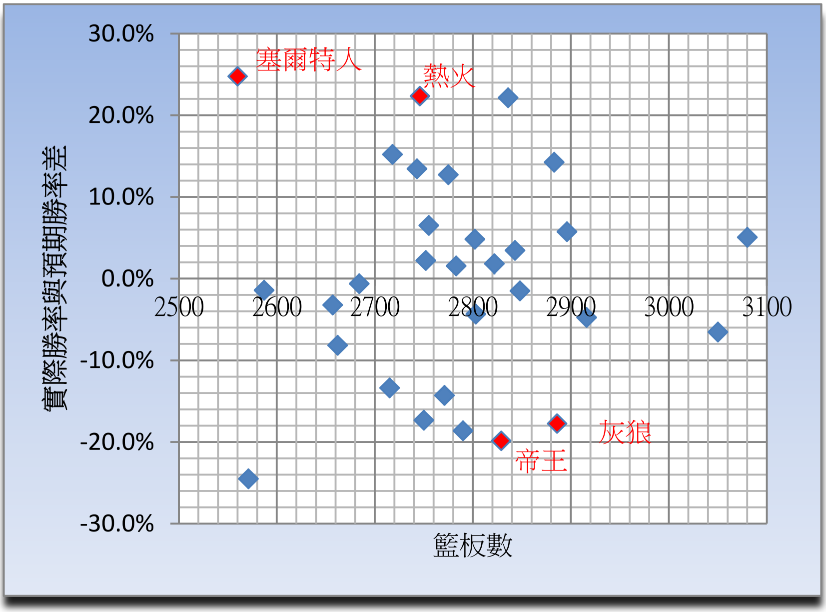
而像上季取得很少籃板的山貓一樣，與勇士、籃網、黃蜂等，都証明取得的籃板數目越少，勝出的概率就越小。

隊伍	勝負率	籃板數
波士頓塞爾特人	59.1%	2560
夏洛特山貓	**10.6%**	**2571**
金州勇士	**34.8%**	**2587**
底特律活塞	**37.9%**	**2657**
布魯克林籃網	**33.3%**	**2662**
波特蘭拓荒者	42.4%	2684
紐奧良黃蜂	**31.8%**	**2715**
亞特蘭大老鷹	60.6%	2718
洛杉磯快艇	60.6%	2743
邁阿密熱火	69.7%	2746

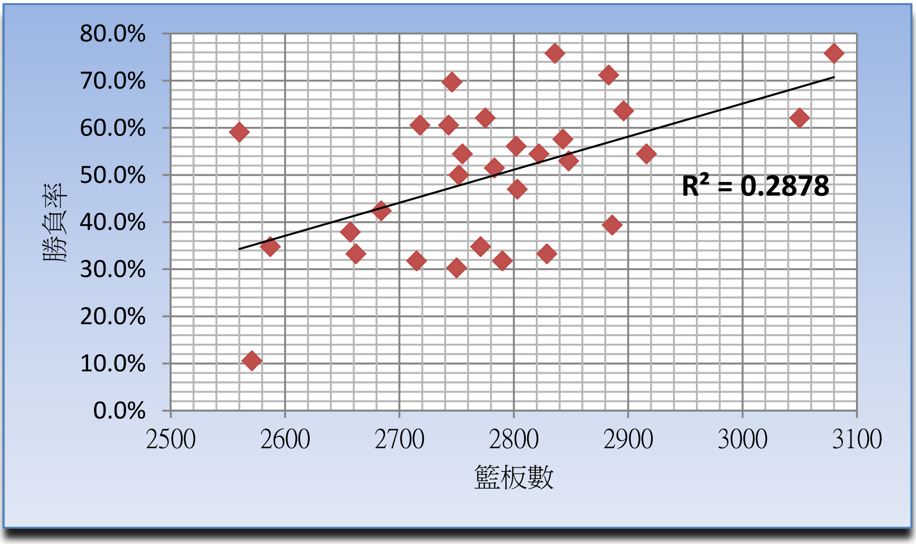
最後，我們就可以得出一個結論，就是籃板數目越多，勝出的概率就越大，籃板數目同勝出的概率正相關。



Jack：但是，亦有例外的！就像灰狼、帝王一樣，即使有多的籃板數目，亦只有低的勝出率；或者像塞爾特人、熱火，他們有高的勝出率但只有很少籃板數目。如果按照你說，根據正比關係的線性資料，便可得出預期的勝出率，再找出真實與預期勝出率的差別，差別越大，便越是離群值。這樣的話，塞爾特人和熱火應該只有約 30-40% 勝出率，但他們實際上還比預期多了約 20%；而灰狼、帝王應該有約 50-60% 勝出率，但他們現在卻比預期少了近 20%！這樣如何解釋，籃板數根本不是和勝出率有絕對關係的！



Cogor ： 這就叫做離群值，不是所有數據都有絕對關係的，所謂的線性只會將所有數據綜合一起畫出一條趨勢線。像塞爾特人、熱火；灰狼、帝王等特別的球隊，會影響整個趨勢線，令他的R平方值降低。所謂的R平方值是介於 0 與 1 之間。R平方值越接近 1，表示籃板數目與勝出的概率相關性越高；越接近 0，表示籃板數目與勝出的概率相關性越低。你看，離群值之多令R平方值下降到只有 0.2878，相關性很小呢！



所以說，自己得到的籃板數目不代表一切，但像剛才社際籃球比賽一樣，信社奪到我們取不到的籃板球，籃板成功搶奪率是 100%！

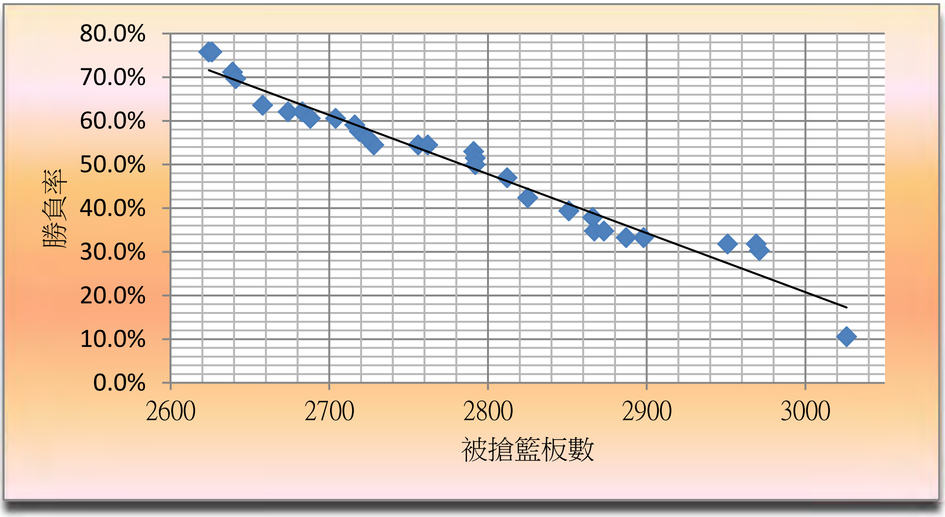
Rabbit ：是的，都要看對方的籃板數量。就如上年ＮＢＡ的情況，被對疊隊伍在自己隊伍上取得最多籃板數量的就是山貓，和巫師、黃蜂、騎士等，都是被對方搶到很多的籃板球，令勝出率下降。

隊伍	勝負率	籃板數	被搶籃板數
夏洛特山貓	**10.6%**	**2571**	**3026**
華盛頓巫師	**30.3%**	**2750**	**2971**
紐奧良黃蜂	**31.8%**	**2715**	**2969**
克里夫蘭騎士	**31.8%**	**2790**	**2951**
布魯克林籃網	33.3%	2662	2898
薩哈拉門托帝王	33.3%	2829	2887
金州勇士	34.8%	2587	2873
多倫多暴龍	34.8%	2771	2867
底特律活塞	37.9%	2657	2866
明尼蘇達灰狼	39.4%	2886	2851

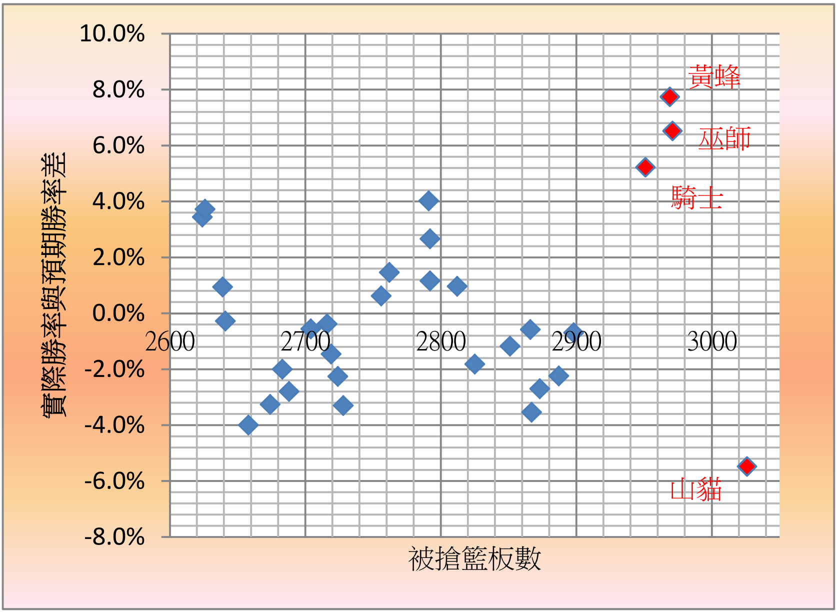
而被別人搶得最少籃板數量的就如公牛，和馬刺、雷霆、熱火等，都是給對方搶到很少的籃板球，令勝出率增加。

隊伍	勝負率	籃板數	被搶籃板數
芝加哥公牛	**75.8%**	**3080**	**2624**
聖安東尼奧馬刺	**75.8%**	**2836**	**2626**
奧克拉荷馬雷霆	**71.2%**	**2883**	**2639**
邁阿密熱火	**69.7%**	**2746**	**2641**
印第安那溜馬	63.6%	2896	2658
洛杉磯湖人	62.1%	3050	2674
孟斐斯灰熊	62.1%	2775	2683
洛杉磯快艇	60.6%	2743	2688
亞特蘭大老鷹	60.6%	2718	2704
波士頓塞爾特人	59.1%	2560	2716

最後，我們又得出一個結論，就是給人搶到籃板數目越少，勝出的概率就越大，被搶籃板數目和勝出的概率是逆相關。



Jack ： 但是，在相若的被搶籃板數下，勝出率也應該相若。但山貓和其他隊伍，如巫師、黃蜂、騎士等，有著完全不同的離群距離呢，照道理他們應與山貓差不多只有約 10% 的勝出率呢，但他們現在比預期多了近 20%，達致約 30%。



(Cogor 與 Rabbit 一時無言而對，互相對視後低頭沉默)

Jack：啊！我想到了。如果把籃板數目與被搶的籃板數目相除，就可以得出他們真正相對的籃板數目，將離群值減少呢！

(Cogor 與 Rabbit 聽到後雙眼發亮，仿似等待下一秒的時間流動)

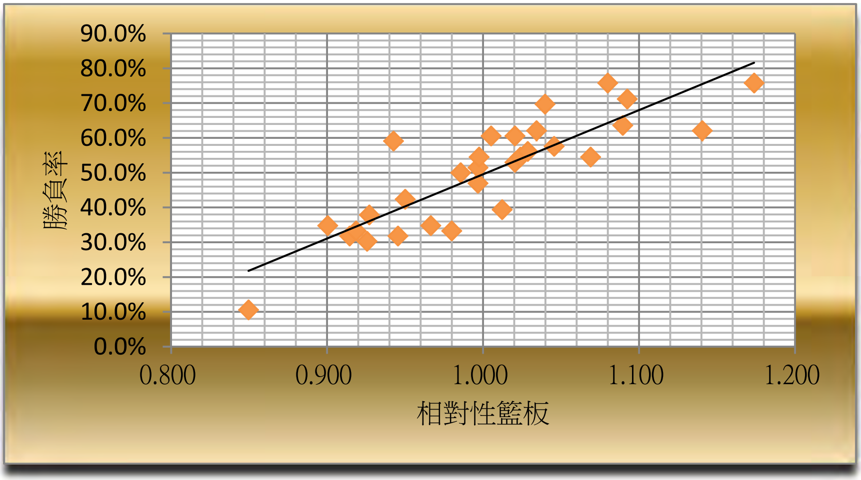
Jack：這次由我來說吧。把籃板數目除被搶的籃板數目就可以得出相對性籃板。之後就可得出一個全新的圖表。

隊伍	勝負率	籃板數	被搶籃板數	相對性籃板
芝加哥公牛	**75.8%**	**3080**	**2624**	**1.174**
洛杉磯湖人	**62.1%**	**3050**	**2674**	**1.141**
奧克拉荷馬雷霆	**71.2%**	**2883**	**2639**	**1.092**
印第安那溜馬	63.6%	2896	2658	1.090
聖安東尼奧馬刺	**75.8%**	**2836**	**2626**	**1.080**
猶他爵士	54.5%	2916	2728	1.069
丹佛金塊	57.6%	2843	2719	1.046
邁阿密熱火	**69.7%**	**2746**	**2641**	**1.040**
孟斐斯灰熊	62.1%	2775	2683	1.034
奧蘭多魔術	56.1%	2802	2724	1.029
達拉斯小牛	54.5%	2822	2756	1.024
洛杉磯快艇	60.6%	2743	2688	1.020
費城 76 人	53.0%	2848	2791	1.020
明尼蘇達灰狼	39.4%	2886	2851	1.012
亞特蘭大老鷹	60.6%	2718	2704	1.005
紐約人	54.5%	2755	2762	0.997
密爾瓦基公鹿	47.0%	2803	2812	0.997
休士頓火箭	51.5%	2783	2792	0.997
鳳凰城太陽	50.0%	2752	2792	0.986
薩哈拉門托帝王	**33.3%**	**2829**	**2887**	**0.980**
多倫多暴龍	34.8%	2771	2867	0.967
波特蘭拓荒者	42.4%	2684	2825	0.950
克里夫蘭騎士	**31.8%**	**2790**	**2951**	**0.945**
波士頓塞爾特人	59.1%	2560	2716	0.943

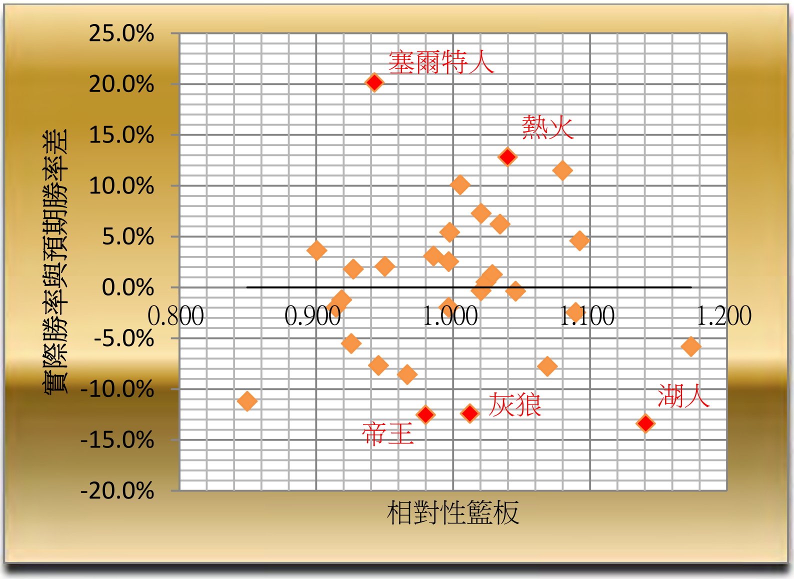
隊伍	勝負率	籃板數	被搶籃板數	相對性籃板
底特律活塞	37.9%	2657	2866	0.927
華盛頓巫師	30.3%	2750	2971	0.926
布魯克林籃網	33.3%	2662	2898	0.919
紐奧良黃蜂	31.8%	2715	2969	0.914
金州勇士	34.8%	2587	2873	0.900
夏洛特山貓	10.6%	2571	3026	0.850

相對性籃板越高，勝出的概率就越高；相反相對性籃板越低，勝出的概率就越小。公牛、湖人、雷霆、馬刺、熱火等，都是相對性籃板很高，自然勝出率就很高；而相反山貓、黃蜂、籃網、騎士、帝王等，都是相對性籃板很低，所以勝出率也遠遠低於水平。

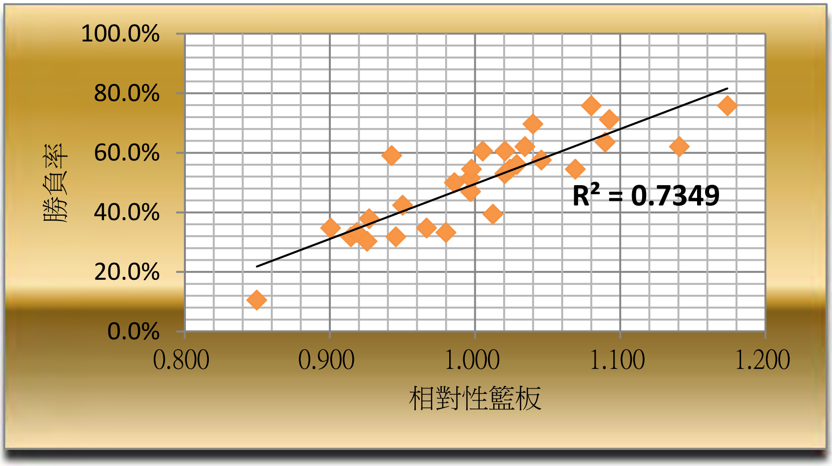
故此，最終的結論就是，相對性籃板越高，勝出的概率就越大，相對性籃板和勝出的概率是正相關。而以這種計算方法就大大減少了之前所說的離群值，達至更加準確。



Rabbit ： 你說得對，這就可以減低之前所說的錯誤，離群值也沒相差得太大呢！雖然依然有像塞爾特人、熱火、湖人、灰狼、帝王等這些不合群的球隊，但相對之前已經更接近平均線呢！



Cogor：無錯，你看，離群值小了令 *R* 平方值上升至 0.7349，相關性大了很多呢！像剛才的比賽，信社把所有籃板球搶到，但依然不能擊敗我們，說到底我們文社就是這改變命運的離群值！其實，我們可以剔除離群值，就可以令 *R* 平方值更高 。但是這就是真實的存在現象嗎？這個世界到處都存在著離群值，天才與白痴，只會把社會拉開得更遠，但把離群值剔除不計後，是否會產生脫離現實的統計結果呢？一條死氣沉沉的劃一統計線，又有甚麼意義呢？統計就是要找出特別的案例，然後加以研究它，找出它異於常人的原因。



Jack：嗯，人也要有高低起伏才算得上人生，我寧可做離群值，讓 *R* 平方值低，做一個突出的人，也不要像被統一化的機械人，有著死板的數據。

Rabbit：對，塞爾特人、熱火就是有很高的命中率，來彌補

他的籃板不足，而灰狼、帝王就是命中率比較低，不能有效得分。每人也有他的強弱項，離群值正正反映這一點。
(吆！這時下一場社際籃球比賽即將開始，由行社對忠社)
Rabbit : 好，就看看相對性籃板是否真的有用！

Cogor, Jack : 可能會出現離群值呢！(看著那行社的籃板王帶領著整隊參差不齊的隊伍，再看看整隊籃球校隊的忠社)

Jack : 忠社，就讓你們成為離群值吧，把 R 平方值拉低也沒所謂的！

Cogor : 才不是呢，籃板王也有決定勝負的一刻！
(Rabbit 看著他們的爭論，笑了笑)

Rabbit : 或許，真的會創造奇蹟呢！

總結：

離群值暗示世上沒有絕對，而突出也非壞事。即使讓整個 R 平方值下跌，也代表著它的獨特性。相對性籃板的增加令勝率有一定的把握，但同時離群的隊伍也因此突破此界限突圍而出，像塞爾特人，是 17 次 NBA 總冠軍，可見籃板亦不是決定一切，也有例外。而你，是願意做鶴立雞群的那位，還是隨波逐流的人呢？！

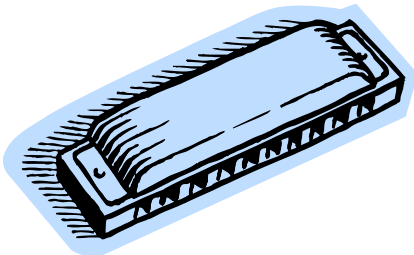
参考資料：

1. NBA 2011-2012 年度各項統計數字
http://stats.nba.com/leagueTeamGeneral.html?Season=2011-12&sortField=OPP_REB&sortOrder=ASC&MeasureType=Base&PerMode=Totals

2. NBA 2011-2012 年度名次及勝出率表
http://www.nba.com/standings/2011/team_record_comparison/conferenceNew_Std_Alp.html

優異作品: 移宮換羽

學校名稱：循道中學
學生姓名：吳嘉賢、謝昭進、黃駿泓
級別：中四
指導老師：楊錫銘老師



音調上下共廿音，

拍子長短為四拍。

排列組合何其多，

試問總數為幾何？

在 1 個四四拍子的小節中，運用八分音符、四分音符、二分音符和全音符，所組成的組合有多少？加上每個音符有 20 個音階，每個音又可以在吹奏時加入重音效果，究竟最後會有多少個組合？

「淅瀝淅瀝—」

好像要把夏日煩悶的熱氣趕走似的，雨一直下個不停，沒有要停下來的跡象。晶瑩剔透的雨點，打落在住宅的玻璃窗上。屋裡很寧靜，和窗外繁忙的街道形成對比；像兩個不同的世界似的，只聽得見雨的淅瀝聲和鉛筆劃過紙張的「沙沙」聲。

炎熱的天氣下，毫無頭緒的作曲家，不由得感到一點煩躁。不知為何，就是沒有靈感湧現。看著眼前還未填上的五線譜，作曲家毅然拿起桌上的十孔口琴，站到窗前吹奏。

「淅瀝淅瀝—」雨點發出清脆的聲響，像奏著一首悅耳動聽的交響樂。

作曲家忽發其想：若是十孔口琴的話，一個 4 拍的小節裡，只用單音，能做出多少個不同的旋律？

「十孔口琴能發出 20 個不同的音，而最快能發出 8 分音符。我先列出在一個 4 拍的小節裡，運用全音符(1)、2 分音符($\frac{1}{2}$)、4 分音符($\frac{1}{4}$)和 8 分音符($\frac{1}{8}$)，節奏能有多少個不同的組合。」作曲家喃喃自語，在計算紙上列出不同的出現模式，「出現模式裡的音符的次序不同的話，也是可以做出不同的節奏，所以必須用到階乘(factorial)。

由於在一個 4 拍的小節裡，全音符(1)的音符數量只有一個，所以計算方式是 1!。

但是運用 2 個 2 分音符的話，音符數量有 2 個，不過由於是 2 個相同的音符，所以是 $\frac{2!}{2!}$。

若是利用 1 個 2 分音符和 2 個 4 分音符的話，音符數量是 3，不過 4 分音符重複了 2 次，所以是 $\frac{3!}{2!}$。

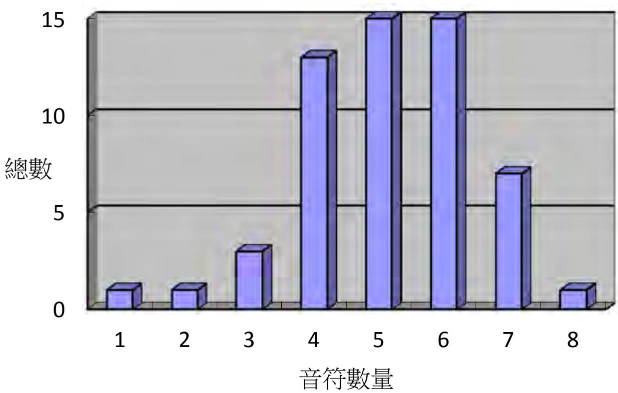
如此類推，共有 56 種不同的節奏。(見表一)

表一：

出現模式	音符數量	計算方式	總數
1	1	$1!$	1
$\frac{1}{2}+\frac{1}{2}$	2	$\frac{2!}{2!}$	1
$\frac{1}{2}+\frac{1}{4}+\frac{1}{4}$	3	$\frac{3!}{2!}$	3
$\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{8}$	4	$\frac{4!}{2!}$	12
$\frac{1}{2}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}$	5	$\frac{5!}{4!}$	5
$\frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{4}$	4	$\frac{4!}{4!}$	1
$\frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{8}+\frac{1}{8}$	5	$\frac{5!}{3!\,2!}$	10
$\frac{1}{4}+\frac{1}{4}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}$	6	$\frac{6!}{2!\,4!}$	15
$\frac{1}{4}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}$	7	$\frac{7!}{6!}$	7
$\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}$	8	$\frac{8!}{8!}$	1

表二：

音符數量	總數
1	1
2	1
3	3
4	13
5	15
6	15
7	7
8	1



作曲家揉了揉眉心，整理出不同音符數量的節奏總數(見表二)：「由於十孔口琴能發出 20 個不同的音，即每個音符共有 20 個不同音階，所以共有 355 億 7010 萬 4420 個不同的旋律。」

$$20^1\times1+20^2\times1+20^3\times3+20^4\times13+20^5\times15+20^6\times15+20^7\times7+20^8\times1$$

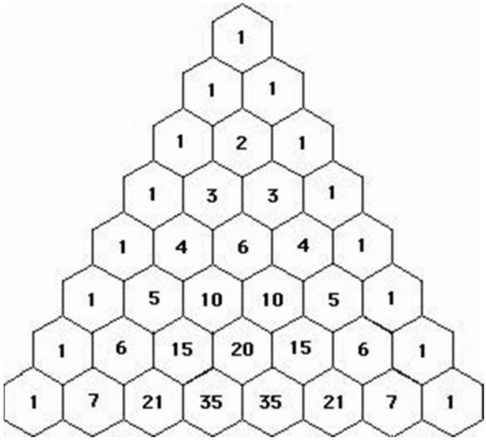
= 35 570 104 420

看著眼前密密麻麻的計算紙，和旁邊毫無進展的五線譜，作曲家嘆了口氣。一想到繳交新曲的限期將近，心情不禁沉重了起來，像石頭般重重地壓了下來。

突然靈機一動：不如加入重音效果？

「一個音符可分為重音或是不重音。因此，我需在不同出現模式的情況下，計算擁有不同重音數目的旋律總數，因為必須運用二項式定理(Binomial theorem)進行計算。」作曲家心裡如此說道，握著筆飛快地寫下公式。

以下恆等式則是計算帕斯卡三角形(Pascal's triangle)中，一行的數字之和。由於音符之間的結合與帕斯卡三角形排列類似，因此可用此恆等式計算出不同組合總數。



$$(1+x)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k}$$

設x為 1，n是音符數量，k是重音數量。

$$2^n = \sum_{k=0}^{n} \binom{n}{k}$$

$\binom{n}{k}$則是組合數學(combinatorics)中的概念，可寫為 C_k^n 或 ${}_nC_k$，公式為 $\frac{n!}{k!(n-k)!}$

在一個 4 拍的小節裡，以 2 個音符數量作例子：

n =2→$2^2 = 4$

出現組合：

$$\frac{1}{2}+\frac{1}{2} \quad \color{red}{\frac{1}{2}}\color{black}{+\frac{1}{2}} \quad \frac{1}{2}+\color{red}{\frac{1}{2}} \quad \color{red}{\frac{1}{2}+\frac{1}{2}}$$

*注: 紅色為有重音組合

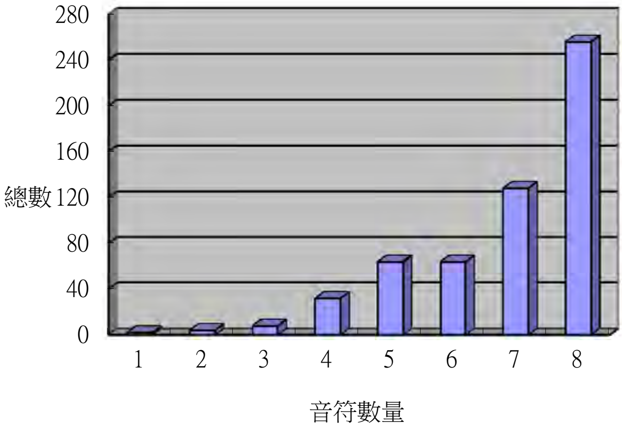
全音符(1)的音符數量只有一個，只可分為 0 個重音和 1 個重音，所以計算方式是$C_0^1 + C_1^1 = 2$；
2 個 2 分音符則是$C_0^2 + C_1^2 + C_2^2 = 4$；
1 個 2 分音符和 2 個 4 分音符便是$C_0^3 + C_1^3 + C_2^3 + C_3^3 = 8$。

從而得出 表三：

出現模式	音符數量	計算方式	總數
1	1	2^1	2
$\frac{1}{2}+\frac{1}{2}$	2	2^2	4
$\frac{1}{2}+\frac{1}{4}+\frac{1}{4}$	3	2^3	8
$\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{8}$	4	2^4	16
$\frac{1}{2}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}$	5	2^5	32
$\frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{4}$	4	2^4	16
$\frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{8}+\frac{1}{8}$	5	2^5	32
$\frac{1}{4}+\frac{1}{4}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}$	6	2^6	64
$\frac{1}{4}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}$	7	2^7	128
$\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}$	8	2^8	256

表四：

音符數量	總數
1	2
2	4
3	8
4	32
5	64
6	64
7	128
8	256



「和之前一樣，十孔口琴能發出 20 個不同的音，所以共有 6 兆 7217 億 4598 萬 5640 個不同的旋律。」

$$20^1\times 2+20^2\times 4+20^3\times 8+20^4\times 32+20^5\times 64+20^6\times 64+20^7\times 128+20^8\times 256$$
$$= 6\ 721\ 745\ 985\ 640$$

「相比之下，加了重音後的可能性多了 6 686 175 881 220 個可能性呢！」

呼了口氣，放下算草紙，重新拿起五線譜。作曲家苦惱著：到底使用哪一個旋律作為下一個小節的呢？

呼嘯而過的汽車聲，喚起了一個想法：生活在繁盛都市的我們，已多久沒有感受到片刻的安詳和寧靜？

拿起筆，在五線譜上寫下了一個全休止符。

優異作品: 一血定生死

學校名稱：香港神託會培基書院
學生姓名：陳伊琳、楊凱婷、黃樂勤
級別：中五
指導老師：區松安老師

引言

相信不少人都有捐血的經驗，還記得首次捐血目的何在? 大部分人也是想知道自己的血型，而紅十字會人員則會根據血型而捐給合適人士。然而，在不知道血型的情況下把血型輸給別人是很危險的。倘若遇上危急情況，那你捐還是不捐呢?

海庭和衣林是一對好姐妹，常結伴同去遊玩，這回她們挑戰極限，打算往非洲森林歷險去。衣林本是個貪玩之人，一到目的地便四處亂竄，怎料樂極生悲，一股勁兒摔到溪澗旁的大石頭，摔得頭破血流，失血過多而昏倒。海庭頓時不知所措，在訊號不通、設備簡陋、寥無人煙的情況下，衣林的命運將會如何？

「衣林你醒醒啊！衣林……」海庭大聲叫喊著，手足無措。

「你這樣叫喚，她也不會甦醒的！」一名青年男子突然從後冒出，滑稽的說道。

「你是誰？你是香港人？」海庭疑惑問道。

「是的！冒昧了，我叫洛芹，是位醫生。我一生愛好瀟灑，幾年來浪跡天涯為救萬人……」洛芹滔滔不絕的自我介紹著。

海庭一見香港人就生安慰，也不管他什麼來歷，一開口就哭道：「洛芹醫生，求求你救救衣林吧！她流了好多血！嗚嗚……」

「哈！遇到我算你走運！拔刀相助正是我的座右銘。來！讓我看看！」洛芹一腔熱血的道。

「糟了！病人大量失血，需要立即輸血！如果不盡快輸血，病人很大機會死亡！」洛芹緊張地道。

「那快把我的血輸給她吧！我和她都是女孩，應該合適的！」海庭不假思索的搶著說。

「你懂不懂醫學常識的？輸血是根據血型，與性別無關的！如果血型不一致，就會產生生溶血反應，衣林會死的！」洛芹氣急敗壞的反駁海庭。

「啊！對哦，我記得從前上生物課時老師曾說過，血型分為四種：A，B，AB 及 O！」海庭恍然大悟。但海庭不隔半秒又道：「可是我不知道衣林是什麼血型……我們只有 $\frac{1}{4}$ 機會把相同的血型輸給她！」

「據我所知，香港人的血型分布不是一樣的，例如有 6.40% 的香港人是 AB 型，41.83%的香港人是 O 型，已經超過 $\frac{1}{4}$ 了！（見表一）」洛芹說。

血型	血型分布
A	26.30%
B	25.48%
AB	6.40%
O	41.83%

表一：香港人血型的分佈 (資料來源: 維基百科(血型))

「另外，輸血未必只能輸給相同血型才可以（見表二）。例如， AB 型血人士是萬用受血者，所有血型都適合的。相反，O 型血人士只能夠接受 O 型血人士捐輸。所以絕對不是 $\frac{1}{4}$ 機會呢!」洛芹解釋道。

		輸血的血型			
		A	B	AB	O
受血者	A	✓	✗	✗	✓
	B	✗	✓	✗	✓
	AB	✓	✓	✓	✓
	O	✗	✗	✗	✓

表二：血型的相容性

「管他的，現在我不知道衣林是甚麼血型，那有多少機會救活她呀？」海庭問道。

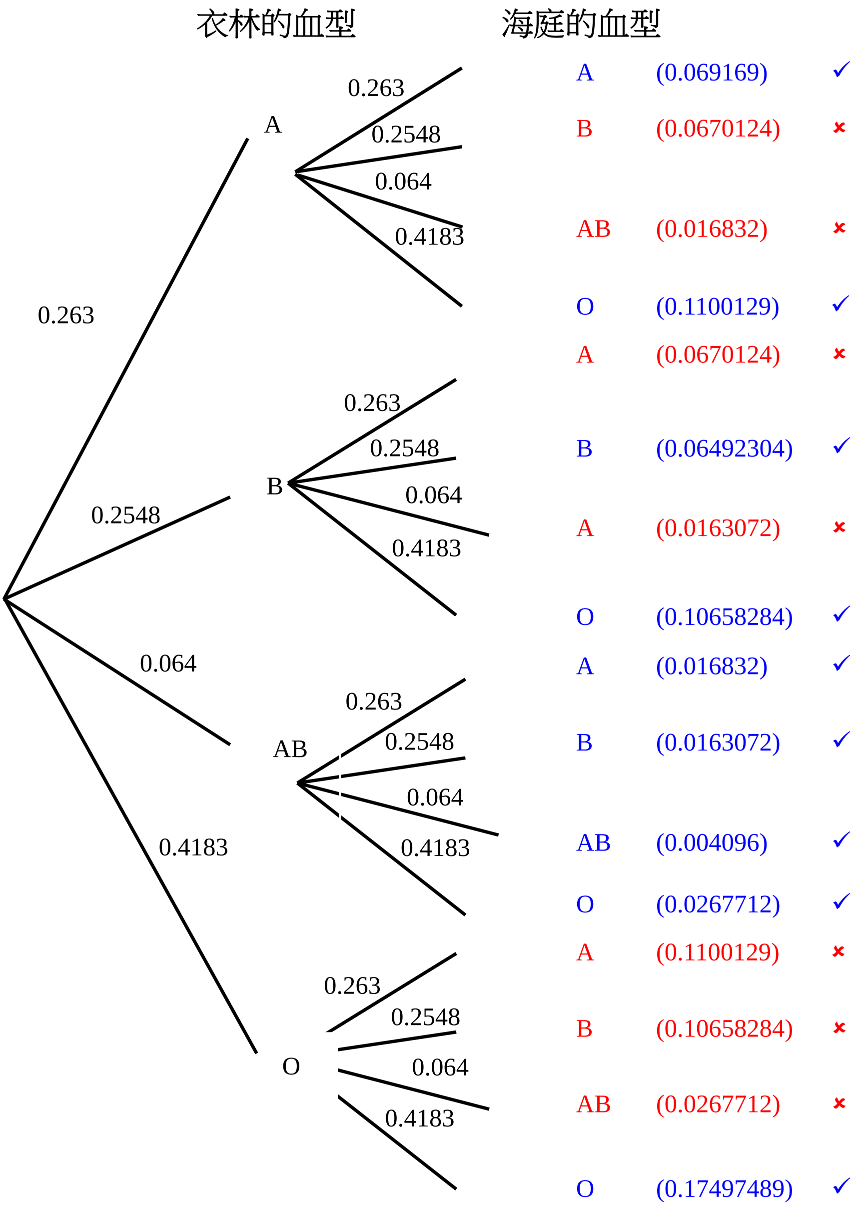
「既然如此，我們來算算其概率吧」洛芹回答。

假設衣林是 A 血型，已知香港有 26.3%人是 A 型血人士(見表 1)， 只能接受 A 型血和 O 型血 (見表二) 。如衣林是 A 型及捐血者也是 A 型，其概率是 $0.263 \times 0.263 = 0.069169$ 。另外衣林也能接受 O 型血，其概率是 $0.263 \times 0.4183 = 0.1100129$ 。如此類推，衣林能夠接受相容性的血型的概率是：

P (海庭是 A | 衣林是 A) P (衣林是 A) +
P (海庭是 O | 衣林是 A) P (衣林是 A) +
P (海庭是 B | 衣林是 B) P (衣林是 B) +
P (海庭是 O | 衣林是 B) P (衣林是 B) +
P (海庭是 A | 衣林是 AB) P (衣林是 AB) +
P (海庭是 B | 衣林是 AB) P (衣林是 AB) +
P (海庭是 AB | 衣林是 AB) P (衣林是 AB) +
P (海庭是 O | 衣林是 AB) P (衣林是 AB) +
P (海庭是 O | 衣林是 O) P (衣林是 O)

= (0.263)(0.263) + (0.4183)(0.263) + (0.2548)(0.2548) +
(0.4183) (0.2548) + (0.263) (0.064) + (0.2548) (0.064) +
(0.064)(0.064) + (0.4183) (0.064) + (0.4183)(0.4183)

= 58.97%


衣林的血型
海庭的血型
A
B
AB
O
0.263
0.2548
0.064
0.4183
A (0.069169) ✓
B (0.0670124) ×
AB (0.016832) ×
O (0.1100129) ✓
A (0.0670124) ×
B (0.06492304) ✓
A (0.0163072) ×
O (0.10658284) ✓
A (0.016832) ✓
B (0.0163072) ✓
AB (0.004096) ✓
O (0.0267712) ✓
A (0.1100129) ×
B (0.10658284) ×
AB (0.0267712) ×
O (0.17497489) ✓

圖一：樹形圖

海庭聽到衣林只約有一半機會輸入相同的血型，心急如焚，正要昏過去之際，靈光一閃：「啊！雖然我不知道自己的血型，不過我記得衣林父母的血型呀！這會不會有幫助呢？」

「肯定有！這能推測衣林的血型，從而收窄範圍。實在太好了！」洛芹化悲為喜的說道。

「我清楚記得衣林爸媽都是 AB 型的，他們還常開玩笑擁有相同的血型非常有緣。」海庭興高采烈的回答說。

「讓我先看看衣林血型的可能性吧。」洛芹說。

父母血型	子女血型	機會率
AB × AB	A	25%
	B	25%
	AB	50%

表三：父母血型與子女血型的關係

「那麼衣林的血型一定只會是 A，B 或 AB，而不會是 O 了。」洛芹繼續。

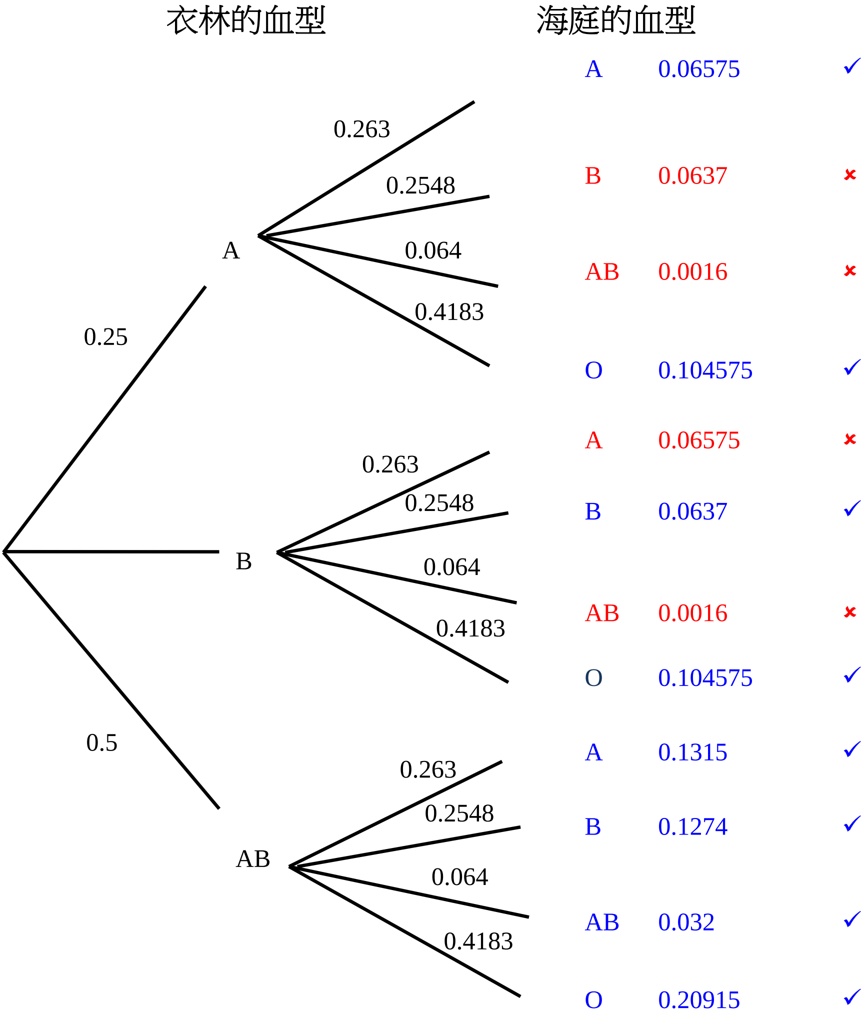
「讓我再重新計算一次輸給她相同血型的概率：」

P (海庭是 A | 衣林是 A) P (衣林是 A) +
P (海庭是 O | 衣林是 A) P (衣林是 A) +
P (海庭是 B | 衣林是 B) P (衣林是 B) +

P (海庭是 O | 衣林是 B) P(衣林是 B) +
P (海庭是 A | 衣林是 AB) P (衣林是 AB) +
P (海庭是 B | 衣林是 AB) P (衣林是 AB) +
P (海庭是 AB | 衣林是 AB) P (衣林是 AB) +
P (海庭是 O | 衣林是 AB) P (衣林是 AB) +

= (0.263)(0.25) + (0.4183)(0.25) + (0.2548)(0.25) +
(0.4183)(0.25) + (0.263)(0.5) + (0.2548)(0.5) + (0.064)(0.5)
+ (0.4183)(0.5)

= 83.87%


衣林的血型
海庭的血型
0.25
0.5
A
B
AB
0.263
0.2548
0.064
0.4183
0.263
0.2548
0.064
0.4183
0.263
0.2548
0.064
0.4183
A 0.06575 ✓
B 0.0637 ×
AB 0.0016 ×
O 0.104575 ✓
A 0.06575 ×
B 0.0637 ✓
AB 0.0016 ×
O 0.104575 ✓
A 0.1315 ✓
B 0.1274 ✓
AB 0.032 ✓
O 0.20915 ✓

圖二：樹形圖

「那代表了什麼？」海庭問。

「但當我們推算衣林的血型後，能推算出她能接受正確血型的概率是 83.87%，其實也是一個比較高的概率啊！怎樣？你決定捐血嗎？」洛芹解釋。

於是海庭二話不說大力地點頭答應輸血。半小時後，捐血程序完成，但衣林依然昏迷不醒，二人憂心忡忡。不一會兒嘈吵的汽車聲傳入樹林中，原來是砍伐樹木的車駛過，精通非洲語的洛芹游說下，他們把衣林載往市區醫院救治。大家才安心下來。

第二天，衣林終於醒了。她不見海庭，又身處陌生地方，正在徬徨之際，洛芹見狀，上前問道：「你還好吧！」她揉揉雙眼，定睛望著洛芹，問道：「你是誰？你是香港人？」洛芹開始滔滔不絕自我介紹……

參考資料：

1. 維基百科
   http://zh.wikipedia.org/wiki/%E8%A1%80%E5%9E%8B

2. 血型之組合
   http://home.dar.com.tw/wendu-28.html

3. 香港紅十字會
   http://www5.ha.org.hk/rcbts/template?series=5&article=11

優異作品: "Detective ‧ Banker"

School Name: Good Hope School
Student Name: Sharon Pun, Vivian Lee, Pamela Wong
Level: Secondary 4
Supervising Teacher: Mr Kenneth Tang

Abstract

The story begins with Jimmy, a banker responsible for middle-class clients, being scolded by his boss for not meeting the expected monthly sales volume for the financial products of the bank. Having failed to deal with his existing clients, he considered recruiting new ones. When reviewing the financial information of some potential clients, Jimmy unexpectedly discovered some strange cases which triggered his further investigation.

Why was there a deposit of a tremendous amount of money in an account suddenly? What caused transactions of considerable sums of money into an account regularly?

Let's begin this mysterious journey with Jimmy!

My Diary

2013

Sebtember

26th September, 2013

Dear Diary,

Dull sky and heavy clouds… It's been a mad day! I was called to the boss' office in the morning. He warned me that I was not reaching the expected sales volume for the bank's financial products for months (again!). The man was really serious this time, constantly warning me that I'd be FIRED if my sales volume stays low.

All I've got in my life is this job as a banker in the King's Bank; all I know is how to promote investment plans to the clients of "Superb Account". I've been doing this for 7 years now, so how could I live if I lost my job? I just couldn't imagine how to live a life of disaster in unemployment!

I realized how severe the case was, so I put extra effort into looking for potential clients, so that they would upgrade to the Superb Account, the first essential step to increasing my sales volume. I have a few in mind, but it's really late now, so I'll look into them one by one tomorrow.

Good night,

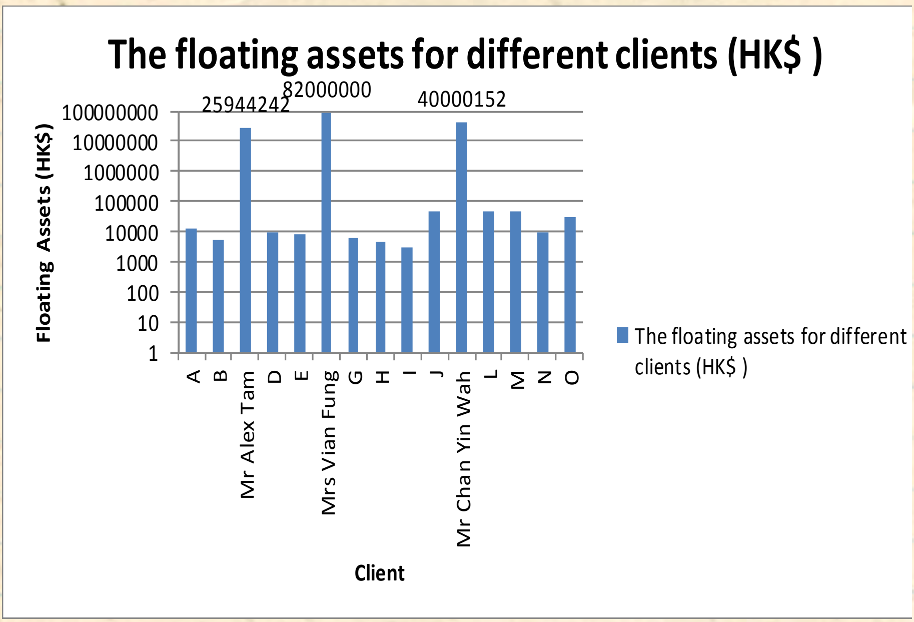
Jimmy

27th September, 2013

Dear Diary,

I continued searching hard for potential clients in the normal accounts dedicated to me. They are randomly taken from all clients holding normal accounts in the region of my bank. I looked into sheets and sheets of statistics showing the floating asset levels of those clients dedicated to me.

Client	The floating assets of clients (HK$)
A	12000
B	5000
Mr Alex Tam	25944242
D	10000
E	8000
Mrs Vian Fung	82000000
G	6004
H	4543
I	3145
J	45894
Mr Chan Yin Wah	40000152
L	45924
M	45968
N	9894
O	32420



Then I found something eye-catching. There are three clients with particularly high floating asset levels. They must be the ones I should look for!

Good night anyway,

Jimmy

28th September, 2013

Dear Diary,

It was superb to have a meeting with Keith in Fantasy Bar just now. We haven't seen each other for years; actually, since when we both graduated from school! He's now a statistical officer working for the government – so talented is he!

We chatted about what we've been doing lately and I told him about my discovery a few days ago (without going into details of the personal information of the clients, of course). He suggested the three distinctly high asset levels might be outliers, from a statistics perspective. I learnt from him three statistical ways to show outliers. Let me check to see whether they are actually outliers. I can then understand more about the data before any further action!

Anyway, I am really exhausted after a hectic working day; perhaps I'll do all this tomorrow.

Goodnight,

Jimmy

29th September, 2013

Dear Diary,

I'd describe today as a success in proving the existence of outliers in my set of data.

I tried out one of the methods "Mean± 3 Standard Deviation" today.

To begin with, I first needed to understand the meaning of Standard Deviation (SD):

In statistics and probability theory, **standard deviation** (represented by the symbol sigma, **σ**) shows how much variation or dispersion exists for the data from the average (mean), or expected value. A low standard deviation indicates that the data tend to be very close to the mean; high standard deviation indicates that the data points are spread out over a large range of values.

The formula for standard deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2},$$

In this formula, μ is the value of the mean, N is the sample size, and x_i represents each data value from i=1 to i=N. The ∑ symbol indicates the operation of sum $(x_1 - \mu)^2 + (x_2 - \mu)^2 + (x_3 - \mu)^2 + (x_4 - \mu)^2 + (x_5 - \mu)^2 . . . + (x_N - \mu)^2$.

I took my data, except the three distinct data, as the whole population and calculated the mean as well as the standard deviation. A range was calculated based on the formula "Mean ± 3Standard Deviation". Any data that is out of this range is then regarded as an outlier.

Floating assets of different clients (HK$)

Client	Floating Assets (HK$)
A	12000
B	5000
D	10000
E	8000
G	6004
H	4543
I	3145
J	45894
L	45924
M	45968
N	9894
O	32420

Mean	19066
SD	17114.1
Mean+3SD	70408.3
Mean -3SD	-32276.3

Therefore the resulting range was between \$-32276.3 and \$70408.3, while the three suspected outliers were \$25944242, \$82000000 and \$40000152 respectively, and the three lied far beyond the range. Therefore, it seemed that the three were indeed outliers!

Keith mentioned that the above method was most useful when the data followed a normal distribution but I was not certain that my data followed a normal distribution. As a result, I tried another method: "median± 3Median Absolute Deviation (MAD)" to verify the above result. Similarly, I first needed to understand what is meant by MAD:

Median absolute deviation is a robust measure of the variability of a univariate sample of quantitative data.

For a univariate data set $X_1, X_2, ..., X_n$, the MAD is defined as the median of the absolute deviations from the data's median:

$$MAD = \underset{i}{median}|x_i - median|$$

that is, starting with the deviations of every datum from the median of all data, median absolute deviation is the median of the absolute values of the deviations. For example, consider the data (1, 1, 2, 2, 4, 6, 9). It has a median value of 2. The absolute deviations from 2 are (1, 1, 0, 0, 2, 4, 7) respectively which in turn have a median value of 1.So the median absolute deviation for this data is 1.

Using similar techniques as in the first method, I calculated the range for "median± 3MAD" as follows .

Client	Floating assets (HK$)	Difference from median
A	12000	2053
B	5000	4947
D	10000	53
E	8000	1947
G	6004	3943
H	4543	5404
I	3145	6802
J	45894	35947
L	45924	35977
M	45968	36021
N	9894	53
O	32420	22473

*Median of Floating Assets = 9947

Median of Floating Assets	**9947**
MAD	**5175.5**
Median + 3MAD	**25473.5**
Median - 3MAD	**-5489.5**

Again, the three suspected outliers fall out of the range of $-5489.5 to $25473.5, so they must really be outliers!

Using these two methods only tells us that these three are outliers. However, I want to know "how outlying are the outliers?"

It's late now so I'll try the last method tomorrow!

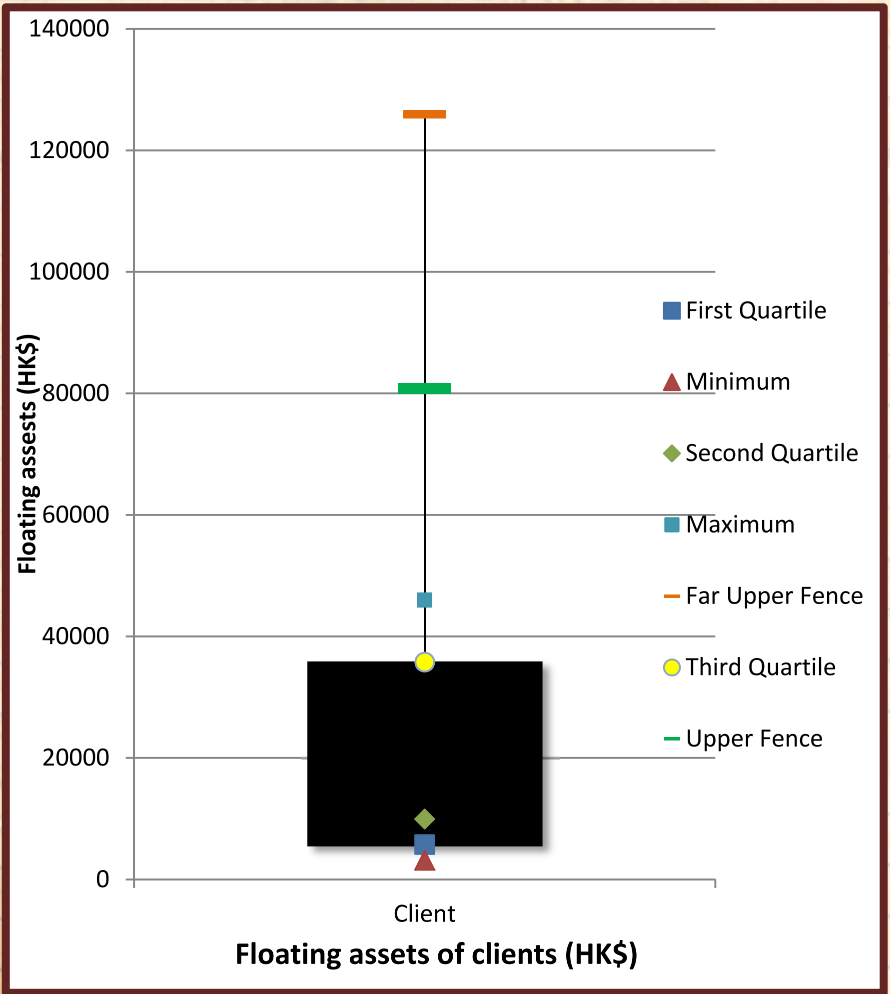
Love,

Jimmy

30th September, 2013

Dear Diary,

I tried using the last method suggested by Keith to verify the outliers today. I divided the data into 4 quartiles, then identified the maximum and minimum data values, and made a box plot. This method classifies the outliers into two categories by setting two reference points: the upper fence and far upper fence. The upper fence is a value equal to the sum of the upper quartile and 1.5(upper quartile – lower quartile), whereas the far upper fence is a value equal to the sum of the upper quartile and 3(upper quartile – lower quartile).



With reference to the graph, the three distinct values all exceeded the far upper fence. It means that they are all far outliers, i.e. extreme outliers.

So it is pretty certain that the three distinct asset levels are all potential clients of the Superb Account. I will investigate their cases in greater detail and approach them individually.

Love,

Jimmy

October

2th October, 2013

Dear Diary,

Today I was reviewing Mr. Alex Tam, a junior clerk's, financial information,. However, when I was studying his money assets, I found that there was a single deposit of $25900000 made to his account on 26 September, 2013 and before that his floating asset level was much lower. I started to wonder what must have happened to him recently.

I called Mr. Tam to invite him to come to the bank to discuss his financial plans. In our conversation, completely to my surprise, he said he had no idea about the $25900000 at all!

How could that be possible? I quickly checked the bank records, made some phone calls and found out at last…

It was actually a silly mistake whereby another customer had mistakenly typed the last digit of their account number wrongly, such that a great sum of money was accidentally deposited into the bank account of Mr. Tam.

Jimmy

9th October, 2013

Dear Diary,

It was a good sunny day and my mind was filled with sunshine as well. I investigated another outlier that I found in my set of data. I phoned Mrs. Fung, a middle-aged lady who had an amount of $ 82000000 deposited to her bank account a few months ago. After the previous case, I was more than careful when talking to her. It turned out that she had won the lottery in Canada months ago!

How lucky she was… I wish I were her … then I wouldn't be sitting here worrying about my career. But at least Mrs. Fung appreciated my effort and agreed to visit me. I hope that she can help improve my performance figures a bit…

Wish me luck!

Jimmy

10th October, 2013

Dear Diary,

Today was finally time to verify and investigate the last outlier. I started the day full of hope. However, I ended it in despair and annoyance.

When I dialed the phone number of Mr. Chan, the one who had an outlying floating asset level, I had mixed feelings about the possible outcome. It took a while until Mr. Chan answered the phone. It was such an ugly voice with a rude tone, and he hung up before I had even finished my speech! This man was completely ruining my day!

I then calmed myself down and convinced myself that he was just too busy and in a bad mood right at that time. After a few hours, I plucked up the courage to call him again. And guess what? It ended with the same result, may be even worse!

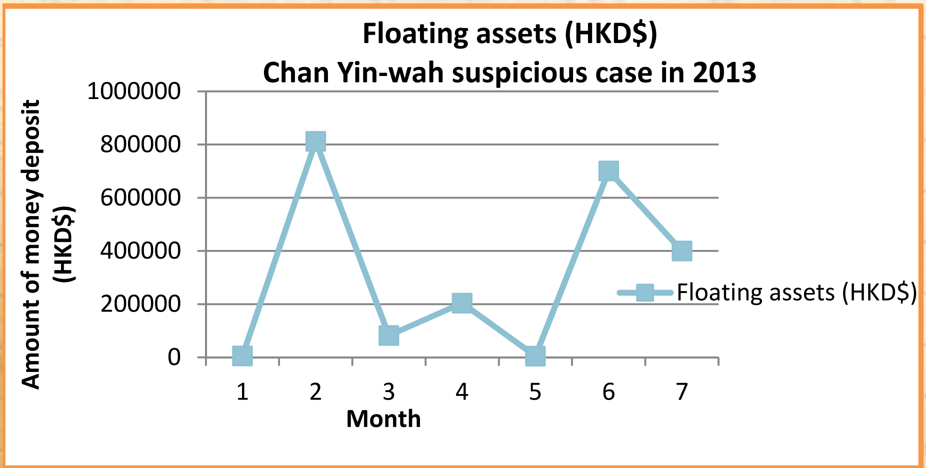
Jimmy

12th October, 2013

Dear Diary,

'I AM NOT GOING TO ENTERTAIN THIS MR CHAN ANYMORE', I swore.

But when my boss came into my mind, I thought it might be a good idea to do some more preparation before I called him again. I looked in detail at his account and discovered wild fluctuations of his floating assets over time.



It looked strange… Well he might be running a business and have an irregular cash flow, or may be money laundering. But the point is, I am confused about what I should do next.

Jimmy

17th October, 2013

Dear Diary,

I've been through a big struggle today. Finally I decided to share my thoughts with my boss and seek his opinion about the best way to deal with Mr. Chan's case. To my surprise, my boss asked me not to follow-up this case myself, but to send him all the relevant information about Mr. Chan. As such, I was too busy to think of the consequences of losing a potential client.

Sweet dreams,

Jimmy

November

3rd November, 2013

Dear Dairy,

Today was the best day of 2013 by far! The case of Mr. Chan turned out to be a suspicious money-laundering case! He was taken to court and part of the evidence used in the court had been prepared by me days before. I am glad that my hard work paid off after all and helped towards maintaining a harmonious society.

Better still, Mrs. Fung was convinced by my financial plan and invested a considerable amount of money in some financial products. There would be no problem in meeting my expected sales volume for months. Because of this, my boss praised me several times!

I hope to make good use of my skills and experience to continue to contribute to the company and society in the future!

Cheers,

Jimmy

"All the News That's Fit to Print"

The Hong Kong Times

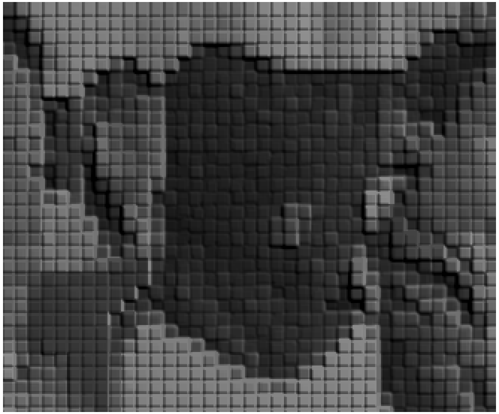
Late Edition

Today: mostly cloudy, a passing afternoon shower, cooler, high 16℃. Tonight: cloudy, low 13℃. Tomorrow: mostly cloudy, a couple of showers, high 19℃. Weather map, Page D8

24-year-old man jailed for 10 years for laundering HK$100 million in Hong Kong

Case discovered by a banker when studying customers' information

by Wong Yee-wah Joyce and Chan Hing-lam Roger



Convicted money launderer Chan Yin-wah, 24, has been sentenced to 10 years imprisonment for persistent laundering in recent months.

Magistrate Tang Lok-in said Chan's crime was serious as he laundered for a Triad society for 4 months, which would have an unbearable impact on society. She mentioned that she could not remember a case where the laundering activity had lasted for so long.

According to evidence provided by the Hong Kong Police Force, Chan helped an unnamed drug lord to launder $100 million in total within a 6-month period.

After Chan successfully finished each task, he received $10000, suspected to be the main motive for his criminal activity.

Jimmy Leung Chun-yuen, one of the key witnesses of the case, discovered the whole incident. He works as a banker in King's Bank and accidentally discovered sudden changes in Chan's floating assets. He reported the case to his senior, who later sent the case to the police for further investigation.

Leung is being tipped to receive a Good Citizen Award in January 2014.

References:

1. Money Laundering

“Money laundering” covers all kinds of methods used to change the identity of illegally obtained money (i.e. crime proceeds) so that it appears to have originated from a legitimate source.

http://www.nd.gov.hk/pdf/moneylaundering/AML_eng_part3.pdf

2. Determining suspicious transactions

To understand the financial background of clients. To confirm the source of assets deposited to clients’ bank accounts etc.

http://www.sfc.hk/web/doc/TC/intermediaries/supervision/prevention/Module5.pdf

PowerPoint presentation conducted by the Narcotics Divisions, Security Bureau, and Security and Futures Commission

http://www.nd.gov.hk/pdf/moneylaundering/AML_eng_part3.pdf

3. Standard Deviation

http://en.wikipedia.org/wiki/Standard_deviation

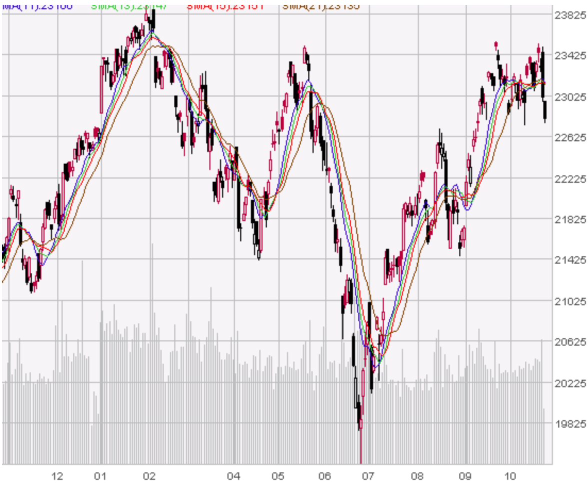
http://www.mathmotivation.com/symbolic/standard-deviation.html

4. Median Absolute Deviation

http://en.wikipedia.org/wiki/Median_absolute_deviation

優異作品: 估市? 股市 – 從過去估計將來

學校名稱：香港道教聯合會圓玄學院第三中學
學生姓名：蘇浩賢、汪嘉恩
級別：中五
指導老師：伍健佳老師



引言

在電影《奪命金》中，主角利用陰陽燭的升跌數目估計將來股市升跌，令他反敗為勝。我們能否從過去估計將來？我倆在這次習作中，利用電影中的方法進行統計及預測，看看是否估中股市升跌。

占士：　你對剛才所看的電影《奪命金》感覺如何?

露絲：　我的天呀！劉青雲所飾演的主角運氣真的太好，在危急關頭，單憑過去數天股票的升跌，便猜中將來股票的走勢，令他賺取巨額財富，胡亂猜測也給他猜中了！

占士：　露絲，那只是電影戲劇效果而已，難道你認為世界上真的有白雪公主和像王子的青蛙嗎？拜託，請別這麼天真。

露絲：　可是，在現實中，又能否利用過去的股市情況來預計未來的升跌呢？

占士：　就讓我們嘗試用劇中的方法做一個統計吧！

現在，我手上有 2010 年至 2011 的股市走勢圖，這一年大約有兩百多天開市，我們可以從走勢圖上的陰陽燭看出升、跌，空心（即陽燭），低開高收，代表升；相反，實心（即陰燭），高開低收，代表跌。〔見附件一〕根據這圖上的統計結果，連續 5 天零升市日數（即連續 5 天下跌），下一天會升的有 4 次，會跌的則有 2 次。由此可見，當下一次股市有連續 5 天跌的時候，買升而贏的機會有 67%，可是買跌而贏的概率則只有 33%，在同樣的情況下，聰明的我在下次當然會買升！

占士： 不錯！做得好！可是在 5 天統計，升市日數為 1、2、3、4 或者全升，那接下來的第 6 天的情況又如何呢？

露絲： 分別很大。在每 5 天內，當中有 4 天跌，而只有 1 天升的時候，接下來的第 6 天會升的次數有 39 次，跌的有 31 次，雖然這一天升、跌的次數非常的接近，分別為 56% 和 44%，可是我依然會選擇買升。在 5 天的統計下，升市日數為 2，第 6 天會升和跌的百分比分別是有 44%及 56%；升市日數為 3，第 6 天升跌的百分比分別是 39%及 61%；升市日數為 4，第 6 天升、跌的日數相差頗大，分別為 37%和 63%；最後，升市日數為 5 的時候，下一天是跌的機會是 100%，根本沒有機會升。顯而易見，由 5 天統計的升市日數為 2 上升至 5，第 6 天跌的百分比都比升的機會高，比你更聰明的我，如果將來出現五日內，出現兩次、三次、四次或五次升，我一定根據這些結果去買跌。哈哈！這次不就像電影般發大財啦！ （見表一）

表一　　五天升跌日數統計

升市日數	0	1	2	3	4	5
第 **6** 日為升	4 67%	39 56%	54 44%	33 39%	11 37%	0 0%
第 **6** 日為跌	2 33%	31 44%	69 56%	52 61%	19 63%	1 100%
小結	買升	買升	買跌	買跌	買跌	買跌

占士：　你這自作聰明這性格永遠不變，我們是不能這樣下定論的。要做到穩賺不賠，那就要多做 10 天的統計、20 天的統計，甚至完成 50 天的統計。重複照著 5 天統計的方法計算第 11、21 和 51 天在不同的升市日數升、跌的結果，才做一個明智的投資者（見表二至表四）。但你是否只跟隨統計結果就去買下一日股市的升或跌呢？

表二　十天升跌日數統計

升市日數	2	3	4	5	6	7
第 **11** 日為升	8 57%	31 48%	39 46%	34 41%	21 42%	5 33%
第 **11** 日為跌	6 43%	34 52%	45 54%	48 59%	29 58%	10 67%
小結	買升	買跌	買跌	買跌	買跌	買跌

表三　二十天升跌日數統計

升市日數	5	6	7	8	9	10	11	12	13
第 **21** 日為升	5 83%	10 43%	17 44%	27 47%	29 35%	23 62%	17 49%	5 25%	0 0%
第 **21** 日為跌	1 17%	13 57%	22 56%	30 53%	53 65%	14 38%	18 51%	15 75%	1 100%
小結	買升	買跌	買跌	買跌	買跌	買升	買跌	買跌	買跌

表四　　五十天升跌日數統計

升市日數	17	18	19	20	21	22	23	24	25	26	27
第**51**日為升	4 67%	9 47%	13 43%	14 47%	16 50%	21 54%	16 42%	17 40%	9 38%	3 50%	1 33%
第**51**日為跌	2 33%	10 53%	17 57%	16 53%	16 50%	18 46%	22 58%	26 60%	15 63%	3 50%	2 67%
小結	買升	買跌	買跌	買跌	不作估計	買升	買跌	買跌	買跌	不作估計	買跌

露絲：　當然不會，下一步我們就要像科學科般做實驗，測試統計結果是否正確。我們希望做到以下結果：預期升市，下日真的是升市；或者預期跌市，下日真的是跌市。

真實結果 ／ 統計結果	下日升	下日跌
預期升	✓	✗
預期跌	✗	✓

現在我們就以 2011 至 12 年的圖來測試〔見附件二〕。就 2011–12 年的 5 天市況統計，得知升市日數為 0 的同樣情況出現了 18 次，而在這 18 次當中跟統計數字買升中的百分比卻只有 38.9%，可是對比統計 2010–11 年的數據，出現升的機會應該有 67%，可見兩者之間出現了極大的落差，發現結果竟然跟我們預期的相反（見表五）。

表五　　五天統計後的實驗結果

升市日數	0	1	2	3	4	5
出現次數	18	55	84	80	23	2
預期	買升	買升	買跌	買跌	買跌	買跌
中的次數	7 38.9%	21 38.2%	42 50.0%	57 58.8%	13 56.5%	2 100.0%
準確否?	不準確	不準確	打和	準確	準確	準確

同樣地，利用 2011 至 12 年的 10、20 和 50 天統計，我們都能發現不同的升市日數測試出來的結果都有所不同，有的準確，有的卻不準確，甚至會出現在統計時沒有的升市日數的統計。（見表六至表八）

表六　　十天統計後的實驗結果

升市日數	0	1	2	3	4	5	6	7	8
出現次數	1	10	22	48	52	56	51	14	4
預期	統計時 沒有出現	統計時 沒有出現	買升	買跌	買跌	買跌	買跌	買跌	統計時 沒有出現
中的次數	不能估計	不能估計	8 36.4%	32 66.7%	27 51.9%	28 50.0%	29 56.9%	9 64.3%	不能估計
準確否?	不能統計	不能統計	不準確	準確	準確	打和	準確	準確	不能統計

表七　　二十天統計後的實驗結果

升市日數	3	4	5	6	7	8
出現次數	2	8	17	23	25	35
預期	統計時 沒有出現	統計時 沒有出現	買升	買跌	買跌	買跌
中的次數	不能估計	不能估計	9 52.9%	14 60.9%	16 64.0%	24 68.6%
準確否?	不能統計	不能統計	準	準	準	準
升市日數	9	10	11	12	13	14
出現次數	28	39	44	20	5	2
預期	買跌	買升	買跌	買跌	買跌	統計時沒有出現
中的次數	15 53.6%	24 61.5%	23 52.3%	15 75.0%	3 60.0%	不能估計
準確否?	準確	準確	準確	準確	準確	不能統計

表八　　五十天統計後的實驗結果

升市日數	13	14	15	16	17	18	19	20	21
出現次數	1	5	18	26	14	9	6	9	21
小結	統計時 沒有出現	統計時 沒有出現	統計時 沒有出現	統計時 沒有出現	買升	買跌	買跌	買跌	不能預期
中的次數	不能估計	不能估計	不能估計	不能估計	2 14.3%	5 55.6%	3 50.0%	5 55.6%	沒有估計
準確否?	不能統計	不能統計	不能統計	不能統計	唔準	準確	打和	準確	不能統計
升市日數	22	23	24	25	26	27	28	29	
出現次數	16	9	4	16	13	33	17	1	
小結	買跌	買升	買跌	買跌	不能預期	買跌	統計時 沒有出現	統計時 沒有出現	
中的次數	7 43.8%	7 77.8%	2 50.0%	8 50.0%	沒有估計	15 45.5%	不能估計	不能估計	
準確否?	不準確	準確	打和	打和	不能統計	不準確	不能統計	不能統計	

占士：　等一等！剛才的資料很詳細地列明 5、10、20 和 50 天內不同的升市日數，雖然下一天的升、跌機會，而且經過測試之後，我們都知道以往的統計數據與實際發生的有一定的出入，但我見到二十天的統計是預計得最好，扣除原本在第一次統計沒有出現的情況外，其他日子的預期都可以使到準確的次數比不準確的多，我們的進財大計有很大機會成功…

露絲：　不要太早發夢！即使使用二十天的統計中出現準確的機會較高，但我們還沒有計算升、跌的幅度。股市不同賭大小般有固定的賠率，若我倆真的用二十天的結果買股市的升跌，買中時只能賺蠅頭小利，買錯時卻賠上大錢，真是「贏粒糖，輸間廠」。

占士：　要計算升、跌的幅度？以我倆的能力沒有可能做到。要成為股市中的最終贏家，不能單單依靠升跌這樣表面的資訊作表面的功夫，電影中的橋段真的不能作準。

參考資料:

ChartNexus[提供統計期時間內的陰陽燭圖]

http://www.chartnexus.com/

維基百科

http://zh.wikipedia.org/wiki/%E9%99%B0%E9%99%BD%E7%87%AD

附件一 **2010 年 1 月 10 日至 2011 年 4 月 21 日的股市升跌統計**

	當日升跌	五日升跌日數統計	十天升跌日數統計	二十天日數升跌統計	五十天升跌日數統計
2010 01 11	跌				
2010 01 12	跌				
2010 01 13	跌				
2010 01 14	跌				
2010 01 15	跌	0			
2010 01 18	升	1			
2010 01 19	升	2			
2010 01 20	跌	2			
2010 01 21	跌	2			
2010 01 22	升	3	3		
2010 01 25	升	3	4		
2010 01 26	跌	2	4		
2010 01 27	跌	2	4		
2010 01 28	升	3	5		
2010 01 29	升	3	6		
2010 02 01	升	3	6		
2010 02 02	跌	3	5		
2010 02 03	升	4	6		
2010 02 04	跌	3	6		
2010 02 05	跌	2	5	8	
2010 02 08	跌	1	4	8	
2010 02 09	升	2	5	9	
2010 02 10	跌	1	5	9	
2010 02 11	升	2	5	10	
2010 02 12	升	3	5	11	
2010 02 17	跌	3	4	10	
2010 02 18	跌	2	4	9	
2010 02 19	跌	2	3	9	
2010 02 22	跌	1	3	9	
2010 02 23	升	1	4	9	
2010 02 24	升	2	5	9	
2010 02 25	跌	2	4	9	
2010 02 26	跌	2	4	9	
2010 03 01	升	3	4	9	

	當日升跌	五日升跌日數統計	十天升跌日數統計	二十天日數升跌統計	五十天升跌日數統計
2010 03 02	跌	2	3	8	
2010 03 03	跌	1	3	7	
2010 03 04	跌	1	3	7	
2010 03 05	跌	1	3	6	
2010 03 08	升	1	4	7	
2010 03 09	升	2	4	8	
2010 03 10	跌	2	3	8	
2010 03 11	跌	2	3	7	
2010 03 12	跌	2	3	7	
2010 03 15	升	2	3	7	
2010 03 16	跌	1	3	6	
2010 03 17	升	2	4	7	
2010 03 18	跌	2	4	7	
2010 03 19	跌	2	4	7	
2010 03 22	跌	1	3	7	
2010 03 23	跌	1	2	6	18
2010 03 24	跌	0	2	5	18
2010 03 25	跌	0	2	5	18
2010 03 26	升	1	3	6	19
2010 03 29	升	2	3	6	20
2010 03 30	升	3	4	7	21
2010 03 31	跌	3	3	7	20
2010 04 01	升	4	4	8	20
2010 04 07	升	4	5	9	21
2010 04 08	升	4	6	9	22
2010 04 09	升	4	7	9	22
2010 04 12	跌	4	7	9	21
2010 04 13	跌	3	7	9	21
2010 04 14	跌	2	6	9	21
2010 04 15	跌	1	5	8	20
2010 04 16	跌	0	4	8	19
2010 04 19	跌	0	4	7	18
2010 04 20	升	1	4	8	19
2010 04 21	跌	1	3	8	18
2010 04 22	升	2	3	9	19

	當日升跌	五日升跌日數統計	十天升跌日數統計	二十天日數升跌統計	五十天升跌日數統計
2010 04 23	跌	2	2	9	19
2010 04 26	升	3	3	10	20
2010 04 29	跌	2	4	9	19
2010 04 30	升	3	5	9	19
2010 05 03	跌	2	5	9	19
2010 05 04	跌	2	4	8	19
2010 05 05	跌	1	4	7	19
2010 05 06	跌	1	3	6	19
2010 05 07	升	1	4	6	19
2010 05 10	升	2	4	7	19
2010 05 11	跌	2	4	7	19
2010 05 12	升	3	4	8	20
2010 05 13	跌	3	4	8	19
2010 05 14	跌	2	3	8	19
2010 05 17	跌	1	3	8	19
2010 05 18	升	2	4	8	20
2010 05 19	升	2	5	9	21
2010 05 20	跌	2	5	8	20
2010 05 24	升	3	5	9	20
2010 05 25	跌	3	4	8	20
2010 05 26	升	3	5	9	21
2010 05 27	升	3	5	9	22
2010 05 28	跌	3	5	9	21
2010 05 31	升	3	6	9	22
2010 06 01	跌	3	6	9	21
2010 06 02	跌	2	5	9	21
2010 06 03	跌	1	4	9	21
2010 06 04	升	2	5	10	22
2010 06 07	升	2	5	10	23
2010 06 08	升	3	6	10	24
2010 06 09	升	4	6	11	25
2010 06 10	升	5	6	11	25
2010 06 11	跌	4	6	11	24
2010 06 14	跌	3	5	11	23
2010 06 15	升	3	6	12	24

	當日升跌	五日升跌日數統計	十天升跌日數統計	二十天日數升跌統計	五十天升跌日數統計
2010 06 17	跌	2	6	11	23
2010 06 18	升	2	7	11	23
2010 06 21	升	3	7	12	23
2010 06 22	跌	3	6	11	22
2010 06 23	升	3	6	12	23
2010 06 24	跌	3	5	11	23
2010 06 25	升	3	5	11	24
2010 06 28	跌	2	5	11	24
2010 06 29	跌	2	5	10	24
2010 06 30	升	2	5	11	25
2010 07 02	跌	2	5	11	24
2010 07 05	升	2	5	12	25
2010 07 06	升	3	5	12	25
2010 07 07	跌	3	5	11	25
2010 07 08	跌	2	4	10	24
2010 07 09	升	3	5	10	25
2010 07 12	跌	2	4	9	24
2010 07 13	跌	1	4	9	24
2010 07 14	跌	1	4	9	23
2010 07 15	跌	1	3	8	23
2010 07 16	跌	0	3	8	23
2010 07 19	升	1	3	8	24
2010 07 20	升	2	3	8	25
2010 07 21	升	3	4	9	25
2010 07 22	升	4	5	9	25
2010 07 23	跌	4	4	9	25
2010 07 26	跌	3	4	8	24
2010 07 27	跌	2	4	8	24
2010 07 28	升	2	5	9	25
2010 07 29	跌	1	5	8	25
2010 07 30	升	2	6	9	25
2010 08 02	升	3	6	9	25
2010 08 03	跌	3	5	8	25
2010 08 04	跌	2	4	8	24
2010 08 05	跌	2	3	8	24

	當日升跌	五日升跌日數統計	十天升跌日數統計	二十天日數升跌統計	五十天升跌日數統計
2010 08 06	升	2	4	8	24
2010 08 09	升	2	5	9	24
2010 08 10	跌	2	5	9	24
2010 08 11	跌	2	4	9	23
2010 08 12	升	3	5	10	24
2010 08 13	跌	2	4	10	24
2010 08 16	升	2	4	10	25
2010 08 17	升	3	5	10	25
2010 08 18	跌	3	5	9	24
2010 08 19	跌	2	5	8	23
2010 08 20	升	3	5	9	23
2010 08 23	跌	2	4	9	22
2010 08 24	跌	1	4	9	22
2010 08 25	升	2	5	9	23
2010 08 26	跌	2	4	9	22
2010 08 27	升	2	5	9	23
2010 08 30	跌	2	4	8	22
2010 08 31	跌	2	3	8	21
2010 09 01	升	2	4	9	22
2010 09 02	跌	2	4	9	21
2010 09 03	升	2	4	9	22
2010 09 06	升	3	5	9	22
2010 09 07	跌	3	5	9	22
2010 09 08	跌	2	4	9	22
2010 09 09	跌	2	4	8	21
2010 09 10	跌	1	3	8	21
2010 09 13	升	1	4	8	21
2010 09 14	跌	1	4	7	20
2010 09 15	跌	1	3	7	20
2010 09 16	跌	1	3	7	20
2010 09 17	升	2	3	7	20
2010 09 20	升	2	3	8	21
2010 09 21	跌	2	3	8	21
2010 09 22	跌	2	3	7	21
2010 09 24	升	3	4	8	22

	當日升跌	五日升跌日數統計	十天升跌日數統計	二十天日數升跌統計	五十天升跌日數統計
2010 09 27	跌	2	4	7	22
2010 09 28	跌	1	3	7	21
2010 09 29	升	2	4	8	21
2010 09 30	升	3	5	8	21
2010 10 04	升	3	6	9	21
2010 10 05	升	4	6	9	22
2010 10 06	跌	4	5	8	22
2010 10 07	跌	3	5	8	22
2010 10 08	升	3	6	9	22
2010 10 11	升	3	6	10	23
2010 10 12	跌	2	6	10	22
2010 10 13	升	3	7	10	22
2010 10 14	升	4	7	11	23
2010 10 15	升	4	7	12	24
2010 10 18	跌	3	6	12	24
2010 10 19	升	4	6	12	24
2010 10 20	升	4	7	12	24
2010 10 21	跌	3	7	12	24
2010 10 22	跌	2	6	12	24
2010 10 25	升	3	6	12	24
2010 10 26	跌	2	6	12	24
2010 10 27	跌	1	5	12	23
2010 10 28	跌	1	4	11	22
2010 10 29	跌	1	3	10	22
2010 11 01	升	1	4	10	23
2010 11 02	升	2	4	10	23
2010 11 03	升	3	4	11	24
2010 11 04	升	4	5	12	25
2010 11 05	跌	4	5	11	24
2010 11 08	升	4	5	11	25
2010 11 09	跌	3	5	11	24
2010 11 10	跌	2	5	10	24
2010 11 11	升	2	6	10	25
2010 11 12	跌	2	6	9	24
2010 11 15	跌	1	5	9	24

	當日升跌	五日升跌日數統計	十天升跌日數統計	二十天日數升跌統計	五十天升跌日數統計
2010 11 16	跌	1	4	8	23
2010 11 17	跌	1	3	7	22
2010 11 18	升	1	3	8	23
2010 11 19	跌	1	3	8	23
2010 11 22	升	2	3	8	24
2010 11 23	跌	2	3	8	24
2010 11 24	跌	2	3	8	23
2010 11 25	跌	1	2	8	23
2010 11 26	跌	1	2	8	23
2010 11 29	升	1	3	8	24
2010 11 30	跌	1	3	7	23
2010 12 01	升	2	4	7	23
2010 12 02	跌	2	3	6	23
2010 12 03	跌	2	3	6	23
2010 12 06	跌	1	2	5	22
2010 12 07	升	2	3	6	23
2010 12 08	跌	1	3	6	23
2010 12 09	跌	1	3	5	22
2010 12 10	升	2	4	6	22
2010 12 13	跌	2	3	6	21
2010 12 14	跌	1	3	6	20
2010 12 15	跌	1	2	6	20
2010 12 16	跌	1	2	5	20
2010 12 17	升	1	3	6	20
2010 12 20	跌	1	3	5	19
2010 12 21	升	2	3	6	20
2010 12 22	跌	2	3	6	19
2010 12 23	跌	2	3	6	18
2010 12 24	跌	1	2	6	17
2010 12 28	升	2	3	6	18
2010 12 29	升	2	4	7	18
2010 12 30	升	3	5	7	18
2010 12 31	跌	3	5	7	18
2011 01 03	升	4	5	8	19
2011 01 04	升	4	6	9	19

	當日升跌	五日升跌日數統計	十天升跌日數統計	二十天日數升跌統計	五十天升跌日數統計
2011 01 05	升	4	6	9	20
2011 01 06	跌	3	6	9	20
2011 01 07	跌	3	6	9	20
2011 01 10	跌	2	6	8	20
2011 01 11	升	2	6	9	20
2011 01 12	升	2	6	10	20
2011 01 13	跌	2	5	10	19
2011 01 14	跌	2	5	10	18
2011 01 17	跌	2	4	9	18
2011 01 18	跌	1	3	9	17
2011 01 19	升	1	3	9	18
2011 01 20	跌	1	3	9	18
2011 01 21	跌	1	3	9	17
2011 01 24	跌	1	3	9	17
2011 01 25	跌	1	2	8	17
2011 01 26	升	1	2	8	18
2011 01 27	跌	1	2	7	18
2011 01 28	跌	1	2	7	17
2011 01 31	升	2	3	7	18
2011 02 01	升	3	4	7	18
2011 02 02	升	3	4	7	19
2011 02 07	跌	3	4	7	19
2011 02 08	跌	3	4	7	19
2011 02 09	跌	2	4	7	19
2011 02 10	跌	1	4	6	18
2011 02 11	升	1	4	6	19
2011 02 14	升	2	5	7	19
2011 02 15	跌	2	5	7	19
2011 02 16	升	3	5	8	20
2011 02 17	升	4	5	9	21
2011 02 18	升	4	5	9	21
2011 02 21	跌	3	5	9	21
2011 02 22	跌	3	5	9	21
2011 02 23	跌	2	5	9	20
2011 02 24	跌	1	5	9	20

	當日升跌	五日升跌日數統計	十天升跌日數統計	二十天日數升跌統計	五十天升跌日數統計
2011 02 25	升	1	5	9	21
2011 02 28	升	2	5	10	22
2011 03 01	升	3	6	11	23
2011 03 02	跌	3	5	10	22
2011 03 03	升	4	5	10	23
2011 03 04	跌	3	4	9	22
2011 03 07	升	3	5	10	23
2011 03 08	升	3	6	11	24
2011 03 09	跌	3	6	11	24
2011 03 10	跌	2	6	11	23
2011 03 11	跌	2	5	10	22
2011 03 14	升	2	5	10	22
2011 03 15	跌	1	4	10	22
2011 03 16	跌	1	4	9	21
2011 03 17	升	2	4	9	21
2011 03 18	跌	2	4	8	20
2011 03 21	升	2	4	9	21
2011 03 22	升	3	4	10	22
2011 03 23	升	4	5	11	23
2011 03 24	跌	3	5	11	22
2011 03 25	升	4	6	11	22
2011 03 28	跌	3	5	10	22
2011 03 29	升	3	6	10	23
2011 03 30	升	3	7	11	24
2011 03 31	跌	3	6	10	24
2011 04 01	升	3	7	11	24
2011 04 04	升	4	7	11	25
2011 04 06	升	4	7	11	26
2011 04 07	跌	3	6	11	26
2011 04 08	升	4	7	12	27
2011 04 11	跌	3	6	12	26
2011 04 12	跌	2	6	11	26
2011 04 13	升	2	6	12	27
2011 04 14	升	3	6	13	27
2011 04 15	跌	2	6	12	26

	當日升跌	五日升跌日數統計	十天升跌日數統計	二十天日數升跌統計	五十天升跌日數統計
2011 04 18	跌	2	5	12	25
2011 04 19	跌	2	4	11	25
2011 04 20	升	2	4	11	26
2011 04 21	升				

附件二　**2011 年 4 月 26 日至 2012 年 5 月 25 日的股市升跌統計**

	當日升跌	五日升跌日數統計	十天升跌日數統計	二十天日數升跌統計	五十天升跌日數統計
2011 04 26	升				
2011 04 27	跌				
2011 04 28	跌				
2011 04 29	跌				
2011 05 03	跌	1			
2011 05 04	跌	0			
2011 05 05	升	1			
2011 05 06	升	2			
2011 05 09	升	3			
2011 05 11	跌	3	4		
2011 05 12	跌	3	3		
2011 05 13	升	3	4		
2011 05 16	跌	2	4		
2011 05 17	跌	1	4		
2011 05 18	升	2	5		
2011 05 19	升	3	6		
2011 05 20	跌	2	5		
2011 05 23	跌	2	4		
2011 05 24	升	3	4		
2011 05 25	升	3	5	9	
2011 05 26	升	3	6	9	
2011 05 27	升	4	6	10	
2011 05 30	升	5	7	11	
2011 05 31	升	5	8	12	
2011 06 01	跌	4	7	12	
2011 06 02	跌	3	6	12	
2011 06 03	跌	2	6	11	
2011 06 07	升	2	7	11	
2011 06 08	跌	1	6	10	
2011 06 09	跌	1	5	10	
2011 06 10	跌	1	4	10	
2011 06 13	升	2	4	10	
2011 06 14	升	2	4	11	
2011 06 15	跌	2	3	11	

	當日升跌	五日升跌日數統計	十天升跌日數統計	二十天日數升跌統計	五十天升跌日數統計
2011 06 16	跌	2	3	10	
2011 06 17	跌	2	3	9	
2011 06 20	跌	1	3	9	
2011 06 21	升	1	3	10	
2011 06 22	跌	1	3	9	
2011 06 23	跌	1	3	8	
2011 06 24	升	2	4	8	
2011 06 27	升	3	4	8	
2011 06 28	跌	2	3	7	
2011 06 29	跌	2	3	6	
2011 06 30	升	3	4	7	
2011 07 04	跌	2	4	7	
2011 07 05	跌	1	4	7	
2011 07 06	跌	1	3	6	
2011 07 07	跌	1	3	6	
2011 07 08	升	1	4	7	21
2011 07 11	跌	1	3	7	20
2011 07 12	跌	1	2	6	20
2011 07 13	升	2	3	6	21
2011 07 14	跌	2	3	6	21
2011 07 15	跌	1	2	6	21
2011 07 18	升	2	3	7	22
2011 07 19	升	3	4	8	22
2011 07 20	跌	2	4	7	21
2011 07 21	跌	2	4	7	20
2011 07 22	升	3	4	8	21
2011 07 25	跌	2	4	7	21
2011 07 26	升	2	5	7	21
2011 07 27	升	3	5	8	22
2011 07 28	升	4	6	9	23
2011 07 29	跌	3	6	8	22
2011 08 01	跌	3	5	8	21
2011 08 02	跌	2	4	8	21
2011 08 03	跌	1	4	8	21
2011 08 04	跌	0	4	8	20

	當日升跌	五日升跌日數統計	十天升跌日數統計	二十天日數升跌統計	五十天升跌日數統計
2011 08 05	跌	0	3	7	19
2011 08 08	升	1	4	8	19
2011 08 09	升	2	4	9	19
2011 08 10	跌	2	3	8	18
2011 08 11	升	3	3	9	18
2011 08 12	跌	3	3	9	18
2011 08 15	升	3	4	9	19
2011 08 16	跌	2	4	8	19
2011 08 17	跌	2	4	8	18
2011 08 18	跌	1	4	8	18
2011 08 19	跌	1	4	7	18
2011 08 22	跌	0	3	7	18
2011 08 23	升	1	3	7	18
2011 08 24	跌	1	3	6	17
2011 08 25	跌	1	2	5	17
2011 08 26	跌	1	2	5	17
2011 08 29	跌	1	1	5	17
2011 08 30	跌	0	1	5	17
2011 08 31	升	1	2	6	17
2011 09 01	跌	1	2	6	17
2011 09 02	跌	1	2	6	17
2011 09 05	跌	1	2	5	16
2011 09 06	升	2	2	5	16
2011 09 07	升	2	3	6	17
2011 09 08	跌	2	3	5	17
2011 09 09	跌	2	3	5	16
2011 09 12	跌	2	3	4	16
2011 09 14	跌	1	3	4	16
2011 09 15	跌	0	2	4	16
2011 09 16	跌	0	2	4	16
2011 09 19	跌	0	2	4	15
2011 09 20	跌	0	2	4	15
2011 09 21	跌	0	1	3	15
2011 09 22	跌	0	0	3	14
2011 09 23	升	1	1	4	15

	當日升跌	五日升跌日數統計	十天升跌日數統計	二十天日數升跌統計	五十天升跌日數統計
2011 09 26	跌	1	1	4	15
2011 09 27	升	2	2	5	15
2011 09 28	升	3	3	6	15
2011 09 30	跌	3	3	5	15
2011 10 03	跌	2	3	5	15
2011 10 04	跌	2	3	5	14
2011 10 06	升	2	4	6	15
2011 10 07	升	2	5	6	15
2011 10 10	跌	2	5	5	14
2011 10 11	跌	2	4	5	13
2011 10 12	升	3	5	6	14
2011 10 13	升	3	5	7	15
2011 10 14	跌	2	4	7	15
2011 10 17	升	3	5	8	16
2011 10 18	跌	3	5	8	16
2011 10 19	跌	2	5	8	16
2011 10 20	跌	1	4	8	15
2011 10 21	跌	1	3	8	14
2011 10 24	升	1	4	9	15
2011 10 25	升	2	5	9	15
2011 10 26	升	3	5	10	16
2011 10 27	升	4	5	10	16
2011 10 28	跌	4	5	9	16
2011 10 31	跌	3	4	9	16
2011 11 01	跌	2	4	9	16
2011 11 02	升	2	5	10	17
2011 11 03	跌	1	5	9	17
2011 11 04	跌	1	5	8	16
2011 11 07	跌	1	4	8	16
2011 11 08	跌	1	3	8	16
2011 11 09	跌	0	2	7	16
2011 11 10	跌	0	1	6	16
2011 11 11	跌	0	1	6	16
2011 11 14	跌	0	1	5	15
2011 11 15	升	1	2	6	16

	當日升跌	五日升跌日數統計	十天升跌日數統計	二十天日數升跌統計	五十天升跌日數統計
2011 11 16	跌	1	1	6	16
2011 11 17	升	2	2	7	17
2011 11 18	跌	2	2	7	16
2011 11 21	跌	2	2	6	15
2011 11 22	升	2	3	6	16
2011 11 23	跌	2	3	5	16
2011 11 24	升	2	4	5	17
2011 11 25	升	3	5	6	18
2011 11 28	升	4	6	7	19
2011 11 29	升	4	6	8	20
2011 11 30	跌	4	6	7	20
2011 12 01	跌	3	5	7	20
2011 12 02	升	3	6	8	21
2011 12 05	升	3	7	9	22
2011 12 06	跌	2	6	9	21
2011 12 07	升	3	7	10	22
2011 12 08	跌	3	6	10	21
2011 12 09	跌	2	5	10	20
2011 12 12	跌	1	4	10	20
2011 12 13	升	2	4	10	21
2011 12 14	升	2	5	11	22
2011 12 15	跌	2	5	10	21
2011 12 16	升	3	5	11	21
2011 12 19	跌	3	4	11	21
2011 12 20	升	3	5	11	22
2011 12 21	跌	2	4	11	21
2011 12 22	升	3	5	11	21
2011 12 23	升	3	6	11	22
2011 12 28	跌	3	6	10	21
2011 12 29	升	3	6	10	22
2011 12 30	跌	3	5	10	22
2012 01 03	升	3	6	11	23
2012 01 04	跌	2	5	10	23
2012 01 05	升	3	6	10	23
2012 01 06	跌	2	5	10	22

	當日升跌	五日升跌日數統計	十天升跌日數統計	二十天日數升跌統計	五十天升跌日數統計
2012 01 10	升	3	6	11	22
2012 01 11	升	4	6	12	23
2012 01 12	跌	3	6	12	23
2012 01 13	跌	3	5	11	23
2012 01 16	跌	2	5	10	22
2012 01 17	升	2	5	11	23
2012 01 18	升	2	6	11	24
2012 01 19	升	3	6	12	25
2012 01 20	跌	3	6	11	25
2012 01 26	跌	3	5	11	25
2012 01 27	升	3	5	11	26
2012 01 30	跌	2	4	10	26
2012 01 31	升	2	5	11	27
2012 02 01	跌	2	5	10	26
2012 02 02	升	3	6	11	27
2012 02 03	升	3	6	11	27
2012 02 06	跌	3	5	11	27
2012 02 07	跌	2	4	10	27
2012 02 08	升	3	5	11	27
2012 02 09	升	3	6	11	28
2012 02 10	跌	2	5	10	27
2012 02 13	升	3	6	10	27
2012 02 14	升	4	6	11	27
2012 02 15	升	4	7	12	27
2012 02 16	跌	3	6	12	27
2012 02 20	跌	3	5	11	27
2012 02 21	升	3	6	11	27
2012 02 22	升	3	7	11	27
2012 02 23	升	3	7	12	28
2012 02 24	跌	3	6	12	27
2012 02 27	跌	3	6	11	27
2012 02 28	升	3	6	12	28
2012 02 29	升	3	6	12	29
2012 03 01	跌	2	5	12	28
2012 03 02	跌	2	5	11	27

	當日升跌	五日升跌日數統計	十天升跌日數統計	二十天日數升跌統計	五十天升跌日數統計
2012 03 05	跌	2	5	10	27
2012 03 06	跌	1	4	10	26
2012 03 07	升	1	4	11	27
2012 03 08	升	2	4	11	27
2012 03 09	升	3	5	11	28
2012 03 12	升	4	6	12	28
2012 03 13	升	5	6	12	28
2012 03 14	跌	4	5	11	28
2012 03 15	升	4	6	11	28
2012 03 16	跌	3	6	11	28
2012 03 19	跌	2	6	11	27
2012 03 20	跌	1	6	10	27
2012 03 21	跌	1	5	9	26
2012 03 22	升	1	5	9	27
2012 03 23	跌	1	4	9	26
2012 03 26	跌	1	3	9	25
2012 03 27	升	2	3	9	25
2012 03 28	跌	2	3	8	25
2012 03 29	跌	1	2	8	25
2012 03 30	升	2	3	9	26
2012 04 02	跌	2	3	9	25
2012 04 03	升	2	4	10	25
2012 04 05	升	3	5	10	25
2012 04 06	升	4	5	10	26
2012 04 09	升	4	6	10	27
2012 04 10	跌	4	6	9	26
2012 04 11	升	4	6	9	27
2012 04 12	升	4	7	10	27
2012 04 13	升	4	8	10	28
2012 04 16	升	4	8	11	28
2012 04 17	跌	4	8	11	27
2012 04 18	跌	3	7	11	27
2012 04 19	升	3	7	12	28
2012 04 20	升	3	7	12	28
2012 04 23	跌	2	6	12	27

	當日升跌	五日升跌日數統計	十天升跌日數統計	二十天日數升跌統計	五十天升跌日數統計
2012 04 24	升	3	7	13	28
2012 04 25	跌	3	6	12	27
2012 04 26	升	3	6	13	27
2012 04 27	跌	2	5	13	26
2012 04 30	升	3	5	13	27
2012 05 02	升	3	6	14	28
2012 05 03	升	4	7	14	28
2012 05 04	跌	3	6	13	27
2012 05 07	跌	3	5	12	26
2012 05 08	跌	2	5	11	26
2012 05 09	跌	1	4	11	26
2012 05 10	跌	0	4	10	25
2012 05 11	跌	0	3	9	24
2012 05 14	跌	0	3	8	24
2012 05 15	升	1	3	8	25
2012 05 16	跌	1	2	8	25
2012 05 17	跌	1	1	8	25
2012 05 18	升	2	2	8	25
2012 05 21	升	3	3	8	25
2012 05 22	跌	2	3	8	24
2012 05 23	跌	2	3	7	23
2012 05 24	跌	2	3	7	22
2012 05 25	跌				

優異作品:「NBA 勝利因素」

學校名稱：聖公會呂明才中學
學生姓名：李普然、馮耀鋒
級別：中五
指導老師：盧偉業老師

引言

在籃球比賽中有很多不同的因素影響球隊的得分例如籃板次數、助攻次數、偷球次數、失誤次數。今次我們會研究這些因素如何影響一場籃球的分數。

首先,我們選擇了兩隊 NBA 球隊的數據來分析，這隊就是「波士頓塞爾特人隊」和「印地安納溜馬隊」。為甚麼我們會選擇這兩隊做比較? 第一，兩隊球隊的得分排名相若(分別排聯盟 18 位和 21 位)。第二，兩隊球隊的風格不同。波士頓塞爾特人隊的籃板次數不多(排名聯盟的 29 位)但助攻次數多(排名聯盟的 4 位)而印地安溜馬隊則相反，籃板次數多(排名聯盟的 1 位)但助攻次數不多(排名聯盟的 26 位)。我們想了解究竟有甚麼因素會影響球隊得分。

因素考慮

我們利用線性迴歸方法研究籃板次數、助攻次數、偷球次數、失誤次數對於球隊得分的影響。設 Y 為比賽得分，X_1，X_2，X_3，X_4 為籃板次數、助攻次數、偷球次數、失誤次數。我們觀察最近五十場比賽，得到 50 組 (x_{n1}，x_{n2}，x_{n3}, x_{n4} , y_n) 觀察數據。(n 是比賽場數, n = 1, 2, …50)。

線性迴歸模型

利用線性迴歸模型表達:

$Y = X\beta + \varepsilon$, 其中

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{pmatrix}, i = 1,2,\ldots p,\ X = \begin{pmatrix} 1 | X_1 | X_2 | \ldots | X_p \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

其中 p 為參數數目，β 為變量，ε 為隨機變量。

當中我們要假設

1. 每個隨機變量 (ε_i) 為互相獨立和正態分佈
2. 每個隨機變量為期望值為零的隨機變量
3. 變量與變量之間相互獨立

我們考慮不同的參數組合以找出不同的模型。

印地安納溜馬隊

參數	模型
X_1	$Y = 92.12927 + 0.017779X_1$
X_2	$Y =62.02959 + 1.532061\ X_2$
X_3	$Y = 92.34735 + 0.085794\ X_3$
X_4	$Y =95.48218 - 0.178993\ X_4$
$X_1\ X_2$	$Y = 57.75426+0.087668\ X_1+1.545532\ X_2$
$X_1\ X_3$	$Y =91.54868+0.017559\ X_1+0.085332\ X_3$
$X_1\ X_4$	$Y = 94.38437+0.025058\ X_1-0.182271\ X_4$
$X_2\ X_3$	$Y =61.10551+1.533305\ X_2+0.12988\ X_3$
$X_2\ X_4$	$Y =63.1459+1.526845\ X_2-0.071221\ X_4$
$X_3\ X_4$	$Y = 95.03969+0.057886\ X_3-0.176047\ X_4$
$X_1\ X_2\ X_3$	$Y =56.85725+0.087392\ X_1+1.546716\ X_2+0.127969\ X_3$
$X_1\ X_2\ X_4$	$Y = 58.8942+0.090699\ X_1+1.539981\ X_2-0.082157\ X_4$
$X_1\ X_3\ X_4$	$Y =93.96238+0.024796\ X_1+0.056711\ X_3-0.17935\ X_4$
$X_2\ X_3\ X_4$	$Y = 62.1991+1.528443\ X_2+0.119432\ X_3-0.065029\ X_4$
$X_1\ X_2\ X_3\ X_4$	$Y =57.9933+0.090226\ X_1+1.54146\ X_2+0.115679\ X_3-0.076102\ X_4$

波士頓塞爾特人隊

參數	模型
X_1	$Y = 78.63815+0.424242\ X_1$
X_2	$Y =66.84296+1.232781\ X_2$
X_3	$Y = 89.94413+0.643317\ X_3$
X_4	$Y =93.45636+0.15586\ X_4$
$X_1\ X_2$	$Y = 57.8413+0.266364\ X_1+1.161943\ X_2$
$X_1\ X_3$	$Y =74.6839+0.399892\ X_1+0.559779\ X_3$
$X_1\ X_4$	$Y = 79.30072+0.432826\ X_1-0.072872\ X_4$
$X_2\ X_3$	$Y =63.09611+1.207672\ X_2+0.492136\ X_3$
$X_2\ X_4$	$Y = 63.60019+1.238319\ X_2+0.225555\ X_4$
$X_3\ X_4$	$Y =88.10363+0.638829\ X_3+0.136194\ X_4$
$X_1\ X_2\ X_3$	$Y = 55.00445+0.249365\ X_1+1.143603\ X_2+0.448064\ X_3$
$X_1\ X_2\ X_4$	$Y =56.96484+0.255491\ X_1+1.16695\ X_29+0.086521\ X_4$
$X_1\ X_3\ X_4$	$Y = 75.38423+0.408996\ X_1+0.560433\ X_3-0.077532\ X_4$
$X_2\ X_3\ X_4$	$Y =60.14589+1.213196\ X_2+0.48455\ X_3+0.209224\ X_4$
$X_1\ X_2\ X_3\ X_4$	$Y = 54.19864+0.239322\ X_1+1.148302\ X_2+0.446929\ X_3+0.080257\ X_4$

我們如何運用以上模型?

例如我們考慮用 4 個參數的模型:

Y =57.99933 + 0.090226 X_1 + **<u>1.54146 X_2</u>** + 0.115679 X_3 - 0.076102 X_4

假設籃板次數、偷球次數、失誤次數的數據(X_1 , X_3 , X_4)不變，如果助攻次數增加一次(X_2)，其得分(Y) 會增加 1.54146 分

分析決定系數

雖然我們找到以上 15 個不同的模型，但是哪一個模型最「好」，能夠解釋最多的數據呢?

如何量度呢? 我們可以計算模型的決定系數 (R^2)。

$$R^2 = \frac{SSR}{SST}, \text{ 其中 } SSR = \sum\left(\hat{y}_i - \bar{y}\right)^2, SST = \sum\left(y_i - \bar{y}\right)^2$$

其中 y_i 是觀察值，

$\hat{y}_i$ 是迴歸模式在 y_i 點的預測值，

$\bar{y}$ 是全部 y_i 平均值。

如果 R^2 愈接近 1，愈能代表此迴歸模型能夠解釋全體 y_i 變異量。

我們用以上的計算方式來計出每個組合當中的決定系數，
結果如下：

印地安納溜馬隊

參數	決定系數	排序
X_1	0.000155	15
X_2	0.338431	8
X_3	0.000354	14
X_4	0.004793	12
$X_1 X_2$	0.34217	4
$X_1 X_3$	0.000505	13
$X_1 X_4$	0.005099	10
$X_2 X_3$	0.339241	6
$X_2 X_4$	0.339186	7
$X_3 X_4$	0.004953	11
$X_1 X_2 X_3$	0.342957	3
$X_1 X_2 X_4$	0.343171	2
$X_1 X_3 X_4$	0.005252	9
$X_2 X_3 X_4$	0.339865	5
$X_1 X_2 X_3 X_4$	0.343808	1

波士頓塞爾特人隊

參數	決定系數	排序
X_1	0.0687	12
X_2	0.29638	8
X_3	0.032894	14
X_4	0.002067	15
$X_1 X_2$	0.322483	4
$X_1 X_3$	0.09338	10
$X_1 X_4$	0.069124	11
$X_2 X_3$	0.315508	6
$X_2 X_4$	0.300702	7
$X_3 X_4$	0.034471	13
$X_1 X_2 X_3$	0.338232	2
$X_1 X_2 X_4$	0.323076	3
$X_1 X_3 X_4$	0.093859	9
$X_2 X_3 X_4$	0.319222	5
$X_1 X_2 X_3 X_4$	0.338742	1

自此可見，大約 33% – 34% 的數據可由全模型(full model) 解釋，似乎全模型是最好的。

調整 R^2 值

我們可發現如果使用的參數愈多，模型的 R^2 會愈大。這是當然的，因為數據明顯多了。為了減去過多的參數，我們可考慮調整 R^2 值:

$$R_a^2 = 1 - \frac{\dfrac{SSE}{n-p-1}}{\dfrac{SST}{n-1}}$$

其中 除去 $n - p - 1$ 的目的是為了消去參數數目。我們用以上的計算方式來計出每個組合當中的調整決定系數，結果如下：

印地安納溜馬隊

參數	決定系數	排序	調整決定系數	排序
X_1	0.000155	15	-0.02025	11
X_2	0.338431	8	0.32493	1
X_3	0.000354	14	-0.02005	10
X_4	0.004793	12	-0.01552	9
$X_1 X_2$	0.34217	4	0.314761	2
$X_1 X_3$	0.000505	13	-0.04114	14
$X_1 X_4$	0.005099	10	-0.03636	12
$X_2 X_3$	0.339241	6	0.31171	3
$X_2 X_4$	0.339186	7	0.311652	4
$X_3 X_4$	0.004953	11	-0.03651	13
$X_1 X_2 X_3$	0.342957	3	0.301018	6
$X_1 X_2 X_4$	0.343171	2	0.301245	5
$X_1 X_3 X_4$	0.005252	9	-0.05824	15
$X_2 X_3 X_4$	0.339865	5	0.297729	7
$X_1 X_2 X_3 X_4$	0.343808	1	0.286748	8

波士頓塞爾特人隊

參數	決定系數	排序	調整決定系數	排序
X_1	0.0687	12	0.049694	10
X_2	0.29638	8	0.28202	4
X_3	0.032894	14	0.013158	13
X_4	0.002067	15	-0.0183	15
$X_1 X_2$	0.322483	4	0.294253	2
$X_1 X_3$	0.09338	10	0.055604	9
$X_1 X_4$	0.069124	11	0.030337	12
$X_2 X_3$	0.315508	6	0.286987	3
$X_2 X_4$	0.300702	7	0.271565	8
$X_3 X_4$	0.034471	13	-0.00576	14
$X_1 X_2 X_3$	0.338232	2	0.295992	1
$X_1 X_2 X_4$	0.323076	3	0.279868	6
$X_1 X_3 X_4$	0.093859	9	0.036021	11
$X_2 X_3 X_4$	0.319222	5	0.275768	7
$X_1 X_2 X_3 X_4$	0.338742	1	0.281241	5

可見，經過調整後，印地安納溜馬隊用 $X2$ 模型反而比用 ($X1$ $X2$ $X3$ $X4$) 好，波士頓塞爾特人隊用($X1$ $X2$ $X3$)模型反而比 ($X1$ $X2$ $X3$ $X4$) 好。由此可見參數不一定愈多愈好。

另外可見，就算印地安納溜馬隊籃板次數多(排名聯盟的 1 位)，但是對於得分的貢獻不大，反而只考慮助攻次數竟然比考慮全部因素更好!

限制和誤差

我們的模型是考慮籃板次數、助攻次數、偷球次數、失誤次數是獨立的因素。不過事實上它們當中可能有互相影響的情況出現。

另外，我們可能還有其他因素未考慮的，例如主場能力，球員身體質素和賽程都會對球隊得分有影響。

邀請作品：抄襲疑雲

2008/09 學年借調教師：張應祥

數學教育組提供延伸閱讀的內容

胡老師任教中四甲班地理學科。一天，她批改該班作業時，發覺其中兩名學生*黃統計*和*藍創意*的工作紙，全部十題是非題的答案皆相同。她經過仔細計算，得出兩份全部十題是非題的答案皆相同的概率為 $\frac{1}{2^{10}}$ ，即 $\frac{1}{1024}$ ，數值非常渺小。她覺得事有蹺蹊，於是召見該兩名學生詢問;可是兩名學生異口同聲地否認曾抄襲他人的答案。胡老師不敢貿然下決定，唯有先打發他們離去，囑咐他們翌日再來見她。

胡老師一方面不欲冤枉學生，另一方面又不希望助長學生抄襲功課風氣，可惜苦苦思量良久，也未能想出解決方法。恰巧任教經濟學科的陳老師經過，見胡老師愁眉深鎖，於是詢問底蘊。陳老師得悉事情後，回應說:「有甚麼稀奇呢! 倘若十題的是非題當中，有七題非常簡易，所有學生皆能答對。餘下三題的答案組合總共只有 $2^3 = 8$ 個。既然如此，那兩名學生，全部十題是非題的答案皆相同的概率實為 $\frac{1}{2^3}$ 即 $\frac{1}{8}$ ，並非如妳所計算的概率 $\frac{1}{1024}$ 那麼

小。『放過』他們吧!」

胡老師聽畢陳老師的分析後，覺得有點道理但對假設全部十題是非題裡，有七題的難度在全部學生能力範圍之內，故所有學生皆能答對，則覺得有點兒戲。她於是向任教數學科的馬老師請教。聽過胡老師的困惑後，馬老師首先指出胡老師的算法，未有充份考慮工作紙的隱藏數據；建議她可參考陳老師的看法，全面地計算各種可能後再行推斷。他隨而列舉出『非常簡易』題目的各種總數下，其餘所有題目的答案組合數目:

『非常簡易』題目的總數	0	1	2	3	4	5
其餘所有題目的答案組合數目	2^{10} = 1024	2^{9} = 512	2^{8}= 256	2^{7} = 128	2^{6} = 64	2^{5} = 32
『非常簡易』題目的總數	6	7	8	9	10	
其餘所有題目的答案組合數目	2^{4} = 16	2^{3} = 8	2^{2} = 4	2^{1} = 2		

瀏覽過上表後，胡老師如釋重負，感覺摸索出解決困惑的線索。但思索了一會後，她開始又變得迷惘起來，心想「用這方法去分析，總勝過我胡亂瞎猜。可是，當我判斷某題目是否『非常簡易』時，我須計算該題目的答對率是否 100%；究竟這做法是否恰當? 哪麼我又應否數算該兩名學生的答案呢?」

於是，她再次向馬老師請教。馬老師解釋說:「當然我們可降低答對率要求，但這可能會用上主觀看法來決定門檻。再者，若考慮該兩名學生的答案應否包括在內，分析方法就變得複雜起來了。其實只要該班學生抄襲功課風氣不泛濫，兼且那兩名學生沒有抄襲其他學生的答案，現時建議的『非常簡易』題目判斷標準 — 答對率達 100%是可行的。」

聽了馬老師的解釋，胡老師立即埋頭苦幹，計算那十題是非題的答對率，赫然發現總共有五題的答對率達 100%。她立即恍然大悟，並大叫起來: 「怪不得有兩名學生的是非題答案全然一樣，或許*黃統計*和*藍創意*真的沒有說謊!」

胡老師如此大叫，究竟中四甲班最少有多少名學生?

延伸閱讀:

以上故事是否似曾相識? 怎樣分析考試結果，找出某位學生有否抄襲其他同學? 現在我們嘗試以單項選擇題作探討。國際性的考試題目常有單項選擇題，好處是客觀性強，考核範圍廣，評卷迅速等優點。魔鬼經濟學(FREAKONOMICS)[1] 的兩位作者 Steven D. Levitt 和 Stephen J. Dubner 曾於書中提及一個懷疑作弊的案例，並指出如何透過分析學生答題模式來找出懷疑作弊的學生。

該次考試共有 45 道選擇題，每道題有四種選項。作者透過電腦程式發現某班的 22 名學生中有 15 名學生於考卷的後半部分給出了 6 個連續的正確答案，因而引起他的懷疑，理由最少有四方面。

其一，那些問題設於考卷的後半部分，難度較大。其二，那些學生並沒有在考卷的其他部分連續答對 6 道題目。其三，該 15 名學生在前半部分的答案幾乎沒有任何重複之處。其四，那些學生當中有 3 位至少在前半部分有 3 道問題同時留空。

書中還有提及其他懷疑學生們作弊的原因，不在此詳述。為何作者有此懷疑? 或者我們用概率解釋一下。假設某位學生考試時*沒有任何知識*，他只是隨機選取答案，選答正確選項的概率為 0.25。現有 10 條選擇題，答對 6 題的概率為 $C_6^{10}\left(\frac{1}{4}\right)^6\left(\frac{3}{4}\right)^4 = 0.01622$。0.01622 可以理解為若該位學生沒有抄襲其他同學，該事件發生的概率。這個假設是否正確? 一般學生會為考試作準備，因此揀選正確選項的概率會高很多，同時題目的深淺度影響選答是否正確，這個計算方法未免過分簡化現實情況。

究竟有沒有較客觀的準則判定學生有否抄襲? 其實有很多學者發表論文，提出找出可疑學生的方法，很有參考價值。讓我簡述他們的做法。

根據 Harpp and Hogan 1993 的一篇論文[2]，他們主要是比較兩位學生的答案模式找出作弊行為。他們會逐一列出兩位同學每題答案的選項。當兩位同學同時留空某一道選擇題，標示作**空錯** (**EIC**)。當兩位學生都答錯且選項相同，標示作**共錯** (**EEIC**)。數算 **EIC** 和 **EEIC** 的總數，計算 **EIC** 與 **EEIC** 的比值。若 $\frac{\text{EIC}}{\text{EEIC}} > 0.75$ ，則兩位學生作弊的機會很高。為何比值是 0.75? 定出這個數字是根據當時所得的樣本資料。學者們發現小於這個比值的兩位學生並不相鄰，大於這個比值的則通常相鄰，因此訂下這個作弊準則。

建基以上兩位學者的研究， Leda Nath 和 Michael Lovaglia[3] 提出改良版本，利用概率評估方法找出作弊者。他們首先把各同學的考卷編號，鄰座同學的考卷號碼被編為連續號碼，方便日後分析。現簡述該方法的理念。若每題選擇題有四個選項，兩位鄰座的同學均不懂作答，而正確答案為 C。假設他們隨機選答每題的錯誤答案，並選答每題的選項時的决定相互獨立。其中一位可能選取 A，另一位則選取 B，或他們同時選取 B。若選擇題中有兩個同樣吸引的錯誤選項，則兩位同學選取該項的概率均為 0.5。若有三個這樣的錯誤選項，則兩位同學選取該項的概率則轉為 0.333。兩位學者假設兩位同學誤選的概率通常為 0.5，即只有兩個同樣吸引的錯誤選項。他們發現

很多時抄襲他人的同學會有一連串的錯誤答案。假設他們連續選取 10 個錯誤的答案，答案模式一樣，這個概率等於 $\frac{1}{2^{10}}=\frac{1}{1024}$ ，少於 0.001，因此引起懷疑，需要與該些同學們面談，找出真相。

兩位學者運用電腦軟件比較鄰座兩位同學的 **EEIC** 總數並計算其發生的概率。電腦程式數算全班同學揀選每題錯誤選項的概率。例如，現有 100 位同學，在某道選擇題中，50 位選取正確答案 A，25 位選取錯誤答案 B，15 位揀選了錯誤答案 C，10 位同學則揀選了錯誤答案 D。在選取錯誤答案的 50 位同學中，選取了 B 為答案的概率為 0.5，選取了 C 為答案的概率為 0.3。電腦軟件會把涉及 **EEIC** 的所有答案的概率相乘，因假設學生揀選每題的選項為獨立事件，數值若少於 0.001 則須再作研究。

內地學者趙世明及張穎則提出利用**錯同率**找出選擇題的作弊者。錯同率是指 A 考生和 B 考生都答錯且選項相同的題目數佔各自答錯題目總數的比例。

A 考生和 B 考生都答錯且選項相同的題目數佔 A 考生答錯題目總數的比例稱為錯同率 A，A 考生和 B 考生都答錯且選項相同的題目數佔 B 考生答錯題目總數的比例稱為錯同率 B。

二考生之間的錯同率 A 和錯同率 B 稱為**雙向錯同率**。如果錯同率為正態分布的話，判定作弊的依據應該是**雙向錯同率**在統計學意義上明顯高於正常考試情況下的平均錯同率。錯同率達到多少時我們才可以認為其明顯高於正常或平均水平呢？對這方面有興趣的讀者，不妨參考資料[4]找出答案。

讀者可能有一連串的疑問，或者學生們是好朋友，一起溫習並得出一些錯誤的觀念，所以誤選的情況相若。其實，以上各方法都有利弊。本文只是簡略介紹各學者在這方面的研究，有興趣的讀者可參考各資料深入瞭解。

參考資料:

1. Levitt, S. D., & Dubner, S. J. (2005). *Freakonomics: A rogue economist explores the hidden side of everything*. New York: William Morrow.

2. Harpp, D. N., and J. J. Hogan. (1993). *Crime in the Classroom: Detection and Prevention of Cheating on Multiple-Choice Exams*. Journal of Chemical Education 70 (4): 306-11.

3. Leda Nath; Michael Lovaglia (2009). *Cheating on Multiple-Choice Exams: Monitoring, Assessment, and an Optional Assignment*. College Teaching *57* (1): 3-8.

4. 張穎、趙世明.(2004). 多選題作弊雷同判定標準的常模研究[J]. 考試研究。

邀請作品：「樣本方差」除以「n-1」的疑問

某位中學數學教師提出以下的問題:

「在高中數學課程延伸部分的單元一內，提出總體方差的公式是除以 n 而樣本方差的公式則要除以 n-1，有何較好的方法可在課堂內清楚地向學生們解說這概念? 為何計算樣本方差時是除以 n-1? 」

就著這個問題本人參考不同資料，經綜合、整理後給出一些想法及建議，望能對教師有幫助。為何樣本方差的公式是除以$n-1$? 市面上一般數學課本的處理手法是用**無偏估計量**的概念作解釋。何謂無偏估計量? 現實生活中，很多時我們會利用*樣本統計量*估計未知*總體參數*，自然希望*樣本統計量*與未知*總體參數*相差不大。當量度這個"相差"的概念時，教師需要考慮兩點:

1. 樣本統計量是否集結在我們想估計的未知總體參數旁?
2. 從不同的樣本所獲得的樣本統計量變動是否很大?

由於抽取樣本是隨機的，不能保證樣本統計量每次都與未知總體參數十分接近，"相差"有時大一點，有時小一點，但總的來說，我們希望它與未知總體參數沒有差別才好，體現這一要求的準則就是無偏性準則。假設總體 X 的平均

值為 μ ，方差為σ^2 ，現有隨機樣本 (x_1, x_2, x_3, … , x_n) ，其樣本平均值為 $\bar{x}$ 。若樣本方差為 $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ ，則樣本方差 s^2 是 σ^2 的無偏估計量，記作 $\mathrm{E}\left(s^2\right)=\sigma^2$ 。對於這個概念，一般課本均會附上一些數學證明步驟。單從以上的數式，學生們可能懂得運算，但要明白其背後的理念也有一定的困難，同學只好記著公式便算。除了課本上的證明外，教師可以補上以下的運算以證明若樣本方差 s^2 除以 n 則會低估總體方差 σ^2:

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu\right)^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[\left(x_i-\bar{x}\right)+\left(\bar{x}-\mu\right)\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\left(x_i-\bar{x}\right)^2+2\left(x_i-\bar{x}\right)\left(\bar{x}-\mu\right)+\left(\bar{x}-\mu\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 \\
&\geq \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2
\end{aligned}
$$

當然，這種做法也有缺點，學生可能會問，我們知道若 $s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ 是低估了總體方差 σ^2 ，但怎知 $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ 又有否高估了總體方差σ^2呢?

其實這個疑問在台灣的教學界亦引起很大迴響。丁村成先生於*數學傳播*發表的一篇有關的文章 [1]。他提出用以下的方法引入樣本方差較易讓學生理解、明白。假設我們有樣

本資料(x_1, x_2, x_3, … , x_n)。$x_i - \bar{x}$ 可表示某資料與平均值 $\bar{x}$ 的差距，若根據此式把資料總和，由於正負值互相抵消，最後所得的總和只是零。資料間的離散程度根本無法表示出來。因此，他提出採用 $\frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$ 及 $\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ 表示一組樣本資料的離散程度。為了減省運算上的麻煩，方差 $\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ 的表示方式會較佳。此外，要兼顧方差與原來資料的單位不同，我們最後定義樣本標準差為 $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$ 。因此，樣本方差採取除以 n 的定義，不論在教與學都比較符合中學生的思維。葉連昌先生所發表的另一篇文章 [2] 也同意這種說法，但台灣課程與香港的一樣，要求樣本方差為 $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ ，學生們又如何明白這公式，和理解除以 n-1 的理由呢? 現建議教師解說這概念時提出以下的各點:

1. 樣本方差的概念於高中數學課程延伸部分***抽樣分布和點估計***的學習單位出現，帶有「統計抽樣」的意義，其目的是想藉由抽取之「樣本」所提供的訊息來推算

「總體」的狀況，是用「樣本統計量」來估計「總體參數」的情況，重點在於「以少推多」、「以小看大」。當你抽取一個樣本時，該樣本就是你所擁有的全部訊息，我們需要利用這訊息推斷總體的特性。

2. 如葉連昌先生所言，「樣本方差」並不是被抽取之樣本資料的方差。它應該還是「總體」的方差，因它是藉由「**樣本**」來估計總體的方差，才稱之為「**樣本**方差」的，可參考附錄所建議的學生實踐活動讓他們明白其理念。

3. 根據丁村成先生所言，教師可嘗試考慮一個極端的情況，設 $n = 1$，即只有一個資料，根據 $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ 的公式，我們得出答案為 0，這表示只有一個資料，其離散度為 0。若考慮 $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$，則變為 $\frac{0}{0}$，它表示只有一個資料時，討論其離散變得毫無意義。教師可與學生討論以上兩種情況何者較為正確?

4. 教師可設計一個簡單、可行的情境，利用電腦軟件去展示樣本方差 $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ 是總體方差 σ^2 的較

佳估計。假設總體由數字 1，3，4，6，8 和 9 所組成，其總體方差是 $\sigma^2 = \dfrac{\sum x^2 - n\bar{x}^2}{n} = \dfrac{207 - 6\left(\frac{31}{6}\right)^2}{6} = 7.81$。若我們從中隨機選取一個容量為 3 的樣本，每次抽取某數字後均放回，及每個數字被選中的概率均等。例如，現有樣本 3、4、8，則 $s^2 = \dfrac{\sum x^2 - n\bar{x}^2}{n-1} = \dfrac{89 - 3\left(\frac{15}{3}\right)^2}{2} = 7$。我們若以另一公式 $s^2 = \dfrac{\sum x^2 - n\bar{x}^2}{n} = \dfrac{89 - 3\left(\frac{15}{3}\right)^2}{3} = 4.667$ 計算樣本方差，則明顯地與總體方差的數值相差較大。外地學者 Pete Brayne 設計了一個 EXCEL 軟件 [4]，可以讓學生輸入數據實踐。學生透過這活動能求出總體方差與不同樣本的樣本方差的數值，體會它們之間的異同，看到那種估計公式較佳。當然，學生們亦應明白這並不是一個證明方法。

假設於中學文憑試延伸部分單元一給出一條題目: 現有一組資料，要求找出該資料的方差，學生是否需要判定那是總體的資料還是樣本的資料，因而在計算方差時決定除以 n 還是除以 n-1。 若題目並沒有具體說出情境，我認為兩者皆可，你又會怎樣回答呢?

參考資料:

1. 丁村成（民 96）。從方差除以 n 或 n-1 談起 — 數學傳播，第二十九卷第一期。

2. 葉連昌 「方差」為什麼除以「n-1」
http://w3.cmgsh.tp.edu.tw/~math/05.doc

3. 鄭惟厚（民 96）。你不能不懂的統計常識。

4. Activities for the A level Statistics Classroom
http://www.making-statistics-vital.co.uk/

5. Michael Shaughnessy; Beth L Chance (2005). *Statistical Questions from the CLASSROOM* Reston, VA : National Council of Teachers of Mathematics.

建議的教學活動:

把三張尺吋相同的咭紙放在盒內，咭紙上面分別寫上數字 1、2 及 3。設它們代表總體，則總體平均值: $\mu=\frac{1+2+3}{3}=2$ ，總體方差: $\sigma^2=\frac{(1-2)^2+(2-2)^2+(3-2)^2}{3}=\frac{2}{3}$

學生們須從中隨機抽取兩張咭紙，第一次抽取一張咭紙後放回盒內再抽取另一張，組成一個樣本。兩張咭紙所組成的樣本總數共有 $3^2=9$ 個。學生們分別就著不同的樣本組合計算樣本平均值和樣本方差，結果詳列如下:

樣本組合	樣本平均值 $\bar{x}=\frac{\sum x}{n}$	樣本方差 $s^2=\frac{\sum(x-\bar{x})^2}{n-1}$	另一公式 $k^2=\frac{\sum(x-\bar{x})^2}{n}$
(1,1)	$\frac{1+1}{2}=1$	$\frac{(1-1)^2+(1-1)^2}{2-1}=0$	$\frac{(1-1)^2+(1-1)^2}{2}=0$

樣本組合	樣本平均值 $\bar{x}=\frac{\sum x}{n}$	樣本方差 $s^2=\frac{\sum(x-\bar{x})^2}{n-1}$	另一公式 $k^2=\frac{\sum(x-\bar{x})^2}{n}$
(1,2)	$\frac{1+2}{2}=1.5$	$\frac{(1-1.5)^2+(2-1.5)^2}{2-1}=0.5$	$\frac{(1-1.5)^2+(2-1.5)^2}{2}=0.25$
(1,3)	$\frac{1+3}{2}=2$	$\frac{(1-2)^2+(3-2)^2}{2-1}=2$	$\frac{(1-2)^2+(3-2)^2}{2}=1$
(2,1)	$\frac{2+1}{2}=1.5$	$\frac{(2-1.5)^2+(1-1.5)^2}{2-1}=0.5$	$\frac{(2-1.5)^2+(1-1.5)^2}{2}=0.25$
(2,2)	$\frac{2+2}{2}=2$	$\frac{(2-2)^2+(2-2)^2}{2-1}=0$	$\frac{(2-2)^2+(2-2)^2}{2}=0$
(2,3)	$\frac{2+3}{2}=2.5$	$\frac{(2-2.5)^2+(3-2.5)^2}{2-1}=0.5$	$\frac{(2-2.5)^2+(3-2.5)^2}{2}=0.25$

樣本組合	樣本平均值 $\bar{x}=\dfrac{\sum x}{n}$	樣本方差 $s^2=\dfrac{\sum(x-\bar{x})^2}{n-1}$	另一公式 $k^2=\dfrac{\sum(x-\bar{x})^2}{n}$
(3,1)	$\dfrac{3+1}{2}=2$	$\dfrac{(3-2)^2+(1-2)^2}{2-1}=2$	$\dfrac{(3-2)^2+(1-2)^2}{2}=1$
(3,2)	$\dfrac{3+2}{2}=2.5$	$\dfrac{(3-2.5)^2+(2-2.5)^2}{2-1}=0.5$	$\dfrac{(3-2.5)^2+(2-2.5)^2}{2}=0.25$
(3,3)	$\dfrac{3+3}{2}=3$	$\dfrac{(3-3)^2+(3-3)^2}{2-1}=0$	$\dfrac{(3-3)^2+(3-3)^2}{2}=0$

9 個樣本平均值 $\bar{x}$ 的平均值 $\dfrac{1+1.5+2+1.5+2+2.5+2+2.5+3}{9}=2$ 等於總體平均值

9 個樣本方差 s^2 的平均值 $\dfrac{0+0.5+2+0.5+0+0.5+2+0.5+0}{9}=\dfrac{2}{3}$ 等於總體方差

9 個樣本 k^2 的平均值 $\dfrac{0+0.25+1+0.25+0+0.25+1+0.25+0}{9}=\dfrac{1}{3}$ 則小於總體方差

邀請作品：購物天堂還是洛陽紙貴？

大家去世界各地旅遊時可能會有這樣的體會吧！倫敦、巴黎的生活指數高得很，在香港能見到的連鎖快餐店，食物售價都比香港的貴一大截。東南亞大部份國家如泰國、馬來西亞，市內物價比較相宜。十多年前大家都說東京、大阪物價高昂，現在到日本購物，跟香港的價格差距已拉近不少。

這都是一些很概括的個人經驗，而不同人的體會也不盡相同。打個比喻，愛在外地搜購化妝品、護膚品的朋友，自然對不同國家的化妝品價格較為敏感，而饞嘴的旅人會比較在意不同地方食店的收費。財經雜誌《經濟學人》便以連鎖快餐店麥當勞售賣的巨無霸，作為全球主要城市購買力平價（Purchasing Power Parity）的指標。《經濟學人》便以此作為一個指標，以兩地物價差距的比例，與兩地貨幣的幣值差距比例作比較，從而得出那些貨幣幣值被高估或低估的結論。

巨無霸價格與貨幣幣值

假設於某月全美國巨無霸的平均售價為 4.56 美元，而在中國，經市場匯率折算平均價格為 2.61 美元，那「巨無霸指數」顯示人民幣的匯率被低估了 43%。根據購買力平價的經濟理論，長遠來說，兩種貨幣之間的兌換率會趨向使同一件貨物在兩地的價值一致。

另一個角度看，這兩地之間物價水平的差距，可以用來真正的比較兩地的生活水平。如果兩個地方的人均收入，經市場匯率折算後，水平相若，那又是否可以下定論，兩地的生活水準都相若呢？正如上面提到的例子，假設兩個國家人民平均收入水平相若，可是平均物價卻差 43%，物價稍低的一方，人民可以收入換得的物品或服務較多，生活質素（最少在物質上）便較另一地的優勝了。

比較全球各地物價的客觀指標

然而，用巨無霸這單一的貨品作環球性的比較，始終太過概括、把實情過分簡化。世界銀行便由 1968 年首次進行了一個大規模的全球性價格統計調查－國際比較方案，以估算購買力平價。

購買力平價可以看成一個兌換率，是指以基準經濟體系一個貨幣單位可以購買當地的貨品和服務，相當於另一個經濟體系購買同等貨品和服務所需的該經濟體系貨幣單位的數量。購買力平價調整了經濟體系間的價格水平差異，並提供一個量度產量的指標（例如本地生產總值），以便就宏觀經濟環境進行有意義的跨經濟體系比較。

簡單來說，要計算購買力平價，首先要在不同的經濟體系訂定一系列可比較及具代表性的項目，並搜集相關的價格數據，然後把各項目的價格比例以相關的本地生產總值開支權數作匯總。

根據世界銀行的報告，146 個經濟體系參與 2005 年的一輪國際比較方案，從不同經濟體系搜集超過 1000 項貨品和服務的價格。

香港真的是一個購物天堂？

根據 2005 國際比較方案的結果，香港的物價算得上很高嗎？

表一：根據2005國際比較方案，選定經濟體系的價格水平指數

經濟體系	價格水平指數 (美國=100)
丹麥	142
日本	118
英國	118
澳洲	106
加拿大	100
美國	100
韓國	77
中國香港	73
新加坡	65
中國台灣	60
中國內地	42
印度	33

如上表所示，雖然香港的價格水平比較一些亞太地區為高（如

新加坡、中國台灣等)，但相對那些跟香港收入水平相若的經濟體系（如英國、澳洲)，香港的價格水平則明顯較低。所以，在發達地區當中，香港的物價不算高吧！

世界銀行於2011年開展了新一輪國際比較方案的工作。當發佈結果時，大家便可以知道2011年各地價格水平的資料了！

不過，大家應用這些統計數據時應留意，由於國際比例方案涵蓋的是本地生產總值所有開支組成部分的價格水平，除了一般消費品及服務，還包括政府服務、設備及建築工程等，因此價格水平指數的高低並不只代表各地的消費水平的高低。

參考資料：

1. 《經濟學人》創立的巨無霸指數

http://www.economist.com/content/big-mac-index

2. 2005 國際比例方案結果

http://web.worldbank.org/WBSITE/EXTERNAL/DATASTATISTICS/ICPEXT/0,,contentMDK:22438220~menuPK:6748067~pagePK:60002244~piPK:62002388~theSitePK:270065,00.html

3. 2011 國際比例方案

http://siteresources.worldbank.org/ICPEXT/Resources/ICP_2011.html

邀請作品：檢定養魚秘訣的小實驗

志明在客廳裡，對著他昨天買來的魚苗沉思。志明爸爸見狀便上前向志明問個究竟。原來志明今天早上到學校參加興趣小組，與同學交流養魚心得。其中一位同學告訴他坊間流傳令魚苗加快生長的秘訣。

第一個秘訣是把水質穩定劑加進魚缸裡；第二個秘訣是以剪碎了的紅蟲代替一般魚糧去餵飼魚苗。志明對同學提供的秘訣半信半疑，想驗証一下這兩個秘訣能否使志明的魚苗加快生長。他也想知道若果兩個秘技同時使用時會否事半功倍。

於是，爸爸鼓勵志明去做一個實驗，測試水質穩定劑和飼料各自對魚苗生長的效應。志明對爸爸的提議滿有興趣，埋首地準備實驗所需要的用品。

首先，志明將 16 條身長相同的魚苗，隨機地放到 4 個形狀大小相同的魚缸裡。在實驗期間，志明對 4 個魚缸裡的魚苗採用不同的飼養方法。以下是實驗期內志明使用的飼養方法的比較。

魚苗飼養方法的比較

	水質穩定劑 (因素 A)	飼料 (因素 B)
魚缸甲 (1)	沒有 (-)	一般魚糧 (-)
魚缸乙 (A)	有 (+)	一般魚糧 (-)
魚缸丙 (B)	沒有 (-)	剪碎了的紅蟲 (+)
魚缸丁 (AB)	有 (+)	剪碎了的紅蟲 (+)

兩個月後，志明為 16 條魚苗逐一量度並記錄了牠們的長度，並計算每條魚苗在實驗期間身長的變化。

表一：實驗期間魚苗身長的變化

		使用水質穩定劑	飼料	實驗期間魚苗身長的變化 (厘米)				
				魚苗 1	魚苗 2	魚苗 3	魚苗 4	平均 $(\bar{y}_{ij\bullet})$
魚缸甲	(1)	-	-	4.5	3.4	3.7	4.4	4
魚缸乙	A	+	-	4.6	4.4	4.1	3.3	4.1
魚缸丙	B	-	+	5.8	5.6	5.0	5.6	5.5
魚缸丁	AB	+	+	6.8	6.1	6.7	6.0	6.4
							總平均	5

表二：魚苗身長變化的平均值

		魚苗身長變化的平均值(厘米)
有使用水質穩定劑	魚缸乙、魚缸丁	5.25
沒有使用水質穩定劑	魚缸甲、魚缸丙	4.75
用紅蟲餵飼	魚缸甲、魚缸乙	5.95
用一般魚糧餵飼	魚缸丙、魚缸丁	4.05

表二顯示有使用水質穩定劑的魚缸內的魚苗，平均身長變化是 5.25 厘米，比沒有使用水質穩定劑的魚苗平均身長變化多出 0.5 厘米。另一方面，用紅蟲餵飼的魚苗平均身長變化是 5.95 厘米，比用一般魚糧餵飼的魚苗平均身長變化多出 1.9 厘米。

以平均值作比較下，似乎使用水質穩定劑和用紅蟲餵飼這兩個秘訣都有效。

表一顯示在四個魚缸中，同時使用兩個秘訣的魚缸丁內的魚苗，平均身長變化最大，平均有 6.4 厘米的增長。志明心想，魚缸丁的效果最大，似乎同時使用兩個秘訣對魚苗生長事半功倍吧？

爸爸告訴志明，比較魚苗身長變化平均值具參考價值。不過，純粹比較魚苗身長變化平均值，未能推斷養魚秘訣對魚苗生長在統計學上有否顯著的效應。

統計學上的析因設計(factorial design)可以幫助志明去驗証這一點。析因設計是統計學上其中一種實驗設計(experimental design)。在日常生活中，科研人員會用實驗設計的技巧，去測試新藥物對病人的藥效。

其實，爸爸教志明做的實驗，已經運用了析因設計。在這個 2×2 析因設計實驗中，水質穩定劑和飼料(因素 A 和 B)皆是不變項因素(independent variables)。當中，因素 A 和因素 B 各自只有兩種程度，以加號(+)及減號(-)來標示。換言之，兩個因素的不同程度可以組成 4 個組合，分別是(-)(-)、(+)(-)、(-)(+)和(+)(+)，以(1), A, B 和 AB 來標示。

在實驗期內，魚苗生長的變化，稱為依變項 (dependent variable)。至於其他因素，例如魚苗的品種、家居環境因素、飼料的份量和換水的頻率，則維持一致，目的是確保這個實驗只反映水質穩定劑和飼料這兩個不變項因素對魚苗生長的效應。

去分析析因設計的實驗結果，志明需要學習運用統計模型，作出適當的虛假設(null hypothesis)，接著用方差分析(ANOVA)來進行假設檢驗。方差分析是用來探討各個魚缸內的魚苗之間的差異是否顯著。〈想了解多些分析方法的詳情，可閱讀關於的實驗設計與方差分析的參考資料。〉

參考資料：

1. Statistics Glossary – design of experiments & ANOVA (http://www.stats.gla.ac.uk/steps/glossary/anova.html)

2. Kuiper S, Sklar J. Practicing Statistics: Guided Investigations for the Second Course. Pearson Education. (http://www.pearsonhighered.com/kuiper1einfo/assets/pdf/Kuiper_Ch04.pdf)

Invited Article: An Experiment in Statistics

Dr. Tony S.T. WONG
Department of Statistics & Actuarial Science
The University of Hong Kong

Introduction

Students who have studied statistics are expected to recognise the concept of confidence interval. In most textbooks, the idea of confidence interval is that an interval estimate of an unknown population parameter is a random interval constructed so that it has a given probability of including the parameter (Miller and Miller, 2004). In the old days that information technology is very limited, students may rely on imagination to gain insight on the idea with the help of some real life examples. For brevity, we will restrict the discussion of confidence interval in the case of normal population.

It is well-known that an $100(1-\alpha)\%$ equal-tailed confidence interval for the mean μ of a normal population with known variance σ^2 is

$$C.I. = \left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right),$$

where $\bar{x}$ is the sample mean (Miller and Miller, 2004). Below are three true-or-false questions extracted from a self-study quiz from Levine et. al. (2010):

1. The confidence interval is centered about the point estimate of the sample and indicates the probability that the sample statistic is located within the interval.

2. α is the designation for the proportion of values outside the confidence interval (in the tail(s) of the normal distribution) so that 1- α is the confidence interval.

3. The further the confidence interval differs from the sample mean (point estimate), the less confident you can be that the population mean falls within that interval of values.

Students may have difficulties in responding to these questions. The reason seems to be the confusion arising from

(i) the known population parameter,

which plays no role in the formula of $C.I.$. Routine calculations may also lead to the misconception that confidence interval is a constant real line. This is in contrary to the fact that

(ii) the confidence interval is random.

In addition, students are guided to draw a conclusion that with $100(1-\alpha)\%$ confidence the population parameter lies between the $C.I.$. The following interpretation usually accompanies with the conclusion:

(iii) if we could draw a large number of samples from the population and find a confidence interval for μ for each sample, then about $100(1-\alpha)\%$ would contain μ.

It happens that points (i), (ii) and (iii) are inevitably vague because students lack the experience. The situation is similar to the Benedict's test for reducing sugar. Without an experiment of heating Benedict's solution with grape juice leading to formation of red precipitates, we can hardly visualise the process and result. A concrete idea has yet developed in our mind. We are aware of the importance of experiments in scientific learning but the experimental approach to statistics will be difficult for some of us to swallow. Some may insist that statistics is similar to mathematics or equivalent to applied mathematics which is all about proof and calculation. Kurt (1951) made an epochal remark:

If mathematics describes an objective world just like physics, there is no reason why inductive methods should not be applied in mathematics just the same as in physics.

We acknowledge that statistics and mathematics are not all about formal proof, and experiments should have a place in the fields. Borwein and Bailey (2008) noted that

one of the more valuable benefits of the computer-based experimental approach in modern mathematics is its value in rejecting false conjectures. A single computational example can save countless hours of human effort that would otherwise be spent attempting to prove false notions.

Experiments in mathematics appear to be a very powerful tool. We also treasure the fact that the experimental methodology to make statistics more tangible, lively and fun for students. Some may then ponder the spacious laboratory and apparatuses we need. Test tubes and Bunsen burners are not required but a computer which is available in all schools and most households. The purpose of this article is to consolidate the idea of confidence interval through experimental demonstration. Microsoft Excel is one of the many popular computer softwares, offering a platform for statistical experiments. It is able to provide many population distributions and carry out random sampling.

Method

In order to carry out the experiment, readers are suggested to have Microsoft Excel ready and follow the steps. The Microsoft Excel may be viewed as a laboratory and each spreadsheet as a survey. Each cell in the spreadsheet contains an observation from a population. The command RAND() plays an important role for random selection. It gives a random number between zero and one. The use of RAND() mimics the situation that a random selection takes place. For example, a candidate with identity 0.1234 is selected if an output of RAND() is 0.1234. To begin with, we consider a normal population with mean μ=10 and variance σ^2=2^2. The sample size is fixed at $n = 20$. The parameters σ^2 and n are assumed known and may change in different experimental setups. The parameter μ is a crucial experimental factor which we can control in order to gain insight. It is in fact an unknown quantity of interest which we want to make inference by drawing a random sample.

Step 1: Generate a random sample of size 20 from the normal distribution $N\left(10, 2^2\right)$.

(a) Type in cell A1 the formula

$$= NORMINV\left(RAND(\ ),10,2\right).$$

(b) Drag from cell A1 to cell T1.

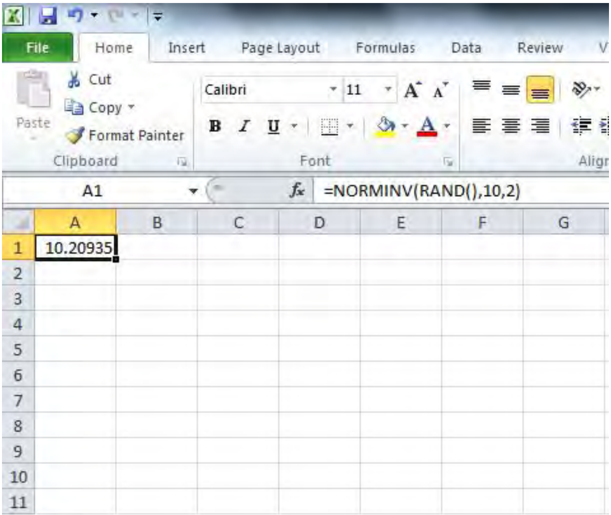


Figure 1. Generate a random variable from $N\left(10,2^2\right)$

Each of these 20 cells contains a randomly selected observation from the normal population $N\left(10,2^2\right)$. The command returns a random variable from $N\left(\mu,\sigma^2\right)$ with identity RAND(). We can revise the concept of random sample. It is defined as a set of independent and identically distributed random variables.

Independence means that the occurrence of an event does not affect the occurrence of another. Each cell has its particular value of RAND() which causes no effect on another. This guarantees independence. Identical distribution refers to the fact that the observations come from the same population. Each cell is governed by two input parameters $\mu = 10$ and $\sigma^2 = 2^2$ which ensure homogenous population. If the value of μ in cell T1 is carelessly input as 100, only the first 19 observations are obtained from the same population but the twentieth is not.

Step 2: Construct a 95% confidence interval for μ.

(a) Compute the sample mean of the random sample. Type in cell $V1$ the formula

$$= AVERAGE\left(A1:T1\right).$$

(b) Construct the lower confidence limit. Type in cell $X1$ the formula

$$= V1\text{-}1.96 * 2 / SQRT\left(20\right).$$

(c) Construct the upper confidence limit. Type in cell $Y1$ the formula

$$= V1+1.96 * 2 / SQRT\left(20\right).$$

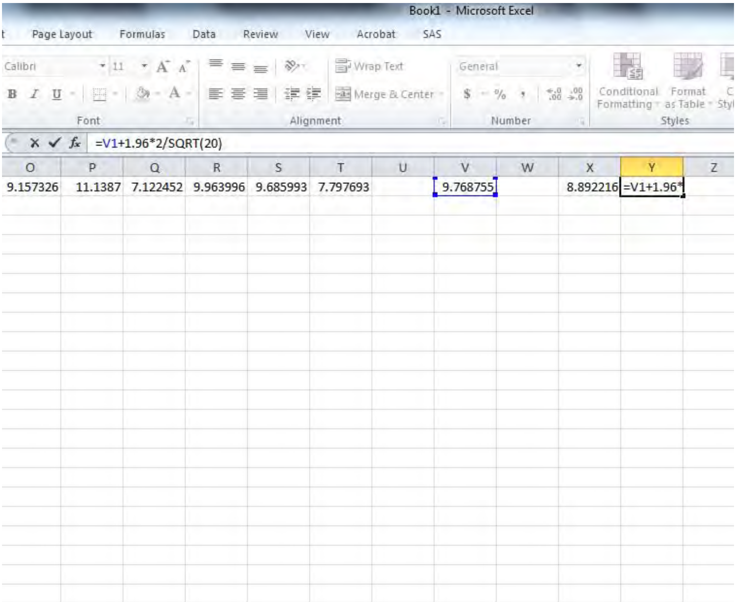
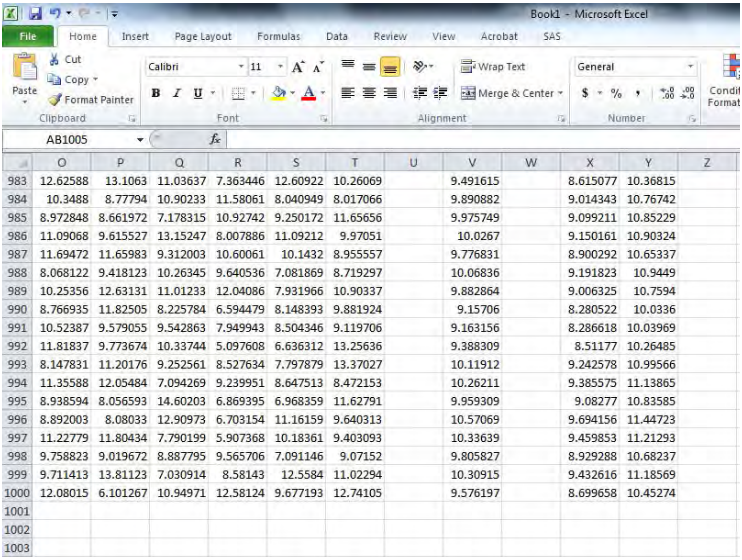


Figure 2. Construct an upper confidence limit

It is not surprising that students can follow quite well until Step 2. They can think of a similar situation of obtaining the height of 20 randomly selected students in a class of 40. The idea of confidence interval will be better illustrated if we have k confidence intervals. It is absolutely a tedious exercise if students have to keep the random selection of 20 students for k=1000 times. In this tiny laboratory, a simple click will do.

Step 3: Construct 1000 confidence intervals for μ.

(a) Drag from row A1-Y1 to row A1000-Y1000.



	O	P	Q	R	S	T	U	V	W	X	Y	Z
983	12.62588	13.1063	11.03637	7.363446	12.60922	10.26069		9.491615		8.615077	10.36815	
984	10.3488	8.77794	10.90233	11.58061	8.040949	8.017066		9.890882		9.014343	10.76742	
985	8.972848	8.661972	7.178315	10.92742	9.250172	11.65656		9.975749		9.099211	10.85229	
986	11.09068	9.615527	13.15247	8.007886	11.09212	9.97051		10.0267		9.150161	10.90324	
987	11.69472	11.65983	9.312003	10.60061	10.1432	8.955557		9.776831		8.900292	10.65337	
988	8.068122	9.418123	10.26345	9.640536	7.081869	8.719297		10.06836		9.191823	10.9449	
989	10.25356	12.63131	11.01233	12.04086	7.931966	10.90337		9.882864		9.006325	10.7594	
990	8.766935	11.82505	8.225784	6.594479	8.148393	9.881924		9.15706		8.280522	10.0336	
991	10.52387	9.579055	9.542863	7.949943	8.504346	9.119706		9.163156		8.286618	10.03969	
992	11.81837	9.773674	10.33744	5.097608	6.636312	13.25636		9.388309		8.51177	10.26485	
993	8.147831	11.20176	9.252561	8.527634	7.797879	13.37027		10.11912		9.242578	10.99566	
994	11.35588	12.05484	7.094269	9.239951	8.647513	8.472153		10.26211		9.385575	11.13865	
995	8.938594	8.056593	14.60203	6.869395	6.968359	11.62791		9.959309		9.08277	10.83585	
996	8.892003	8.08033	12.90973	6.703154	11.16159	9.640313		10.57069		9.694156	11.44723	
997	11.22779	11.80434	7.790199	5.907368	10.18361	9.403093		10.33639		9.459853	11.21293	
998	9.758823	9.019672	8.887795	9.565706	7.091146	9.07152		9.805827		8.929288	10.68237	
999	9.711413	13.81123	7.030914	8.58143	12.5584	11.02294		10.30915		9.432616	11.18569	
1000	12.08015	6.101267	10.94971	12.58124	9.677193	12.74105		9.576197		8.699658	10.45274	
1001												
1002												
1003												

Figure 3: Construct 1000 confidence intervals for μ

The first 20 entries in each row represent a random sample of size 20. The sample mean of each random sample is calculated along column V. Each random sample yields a confidence interval along columns X and Y. The significance of the experiment lies on the visualization of the 1000 confidence intervals. We can refer to point (i) about the parameter μ. Although it is not involved in the calculation of confidence intervals, it controls the values of each individual in the random sample. The inference on μ by confidence interval will be illustrated in Step 4. Point (ii) is also obvious now because the values in columns X and Y are different. If we are asked why the 1000 confidence intervals are different from each other, the answer is straight-forward that the sample observations are different. In short, confidence interval is a random interval because it depends on the sample which is in nature random, too. Point (iii) will be demonstrated in Step 4.

Step 4: Count the number of times μ falls in the confidence intervals.

(a) Check whether the unknown parameter $\mu = 10$ falls in each of the confidence intervals. Type in cell Z1

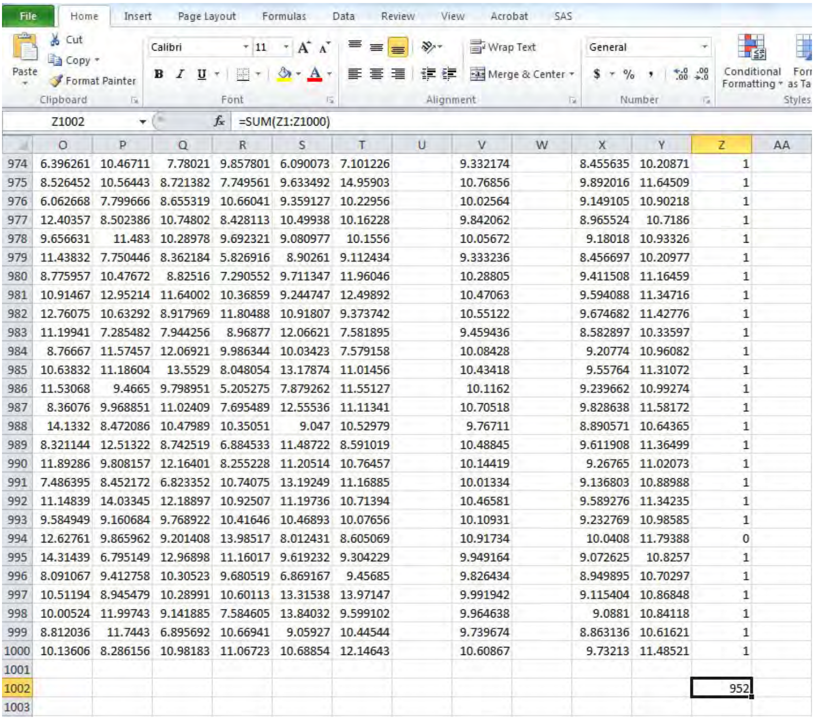
$$= IF\left(AND\left(X1<10,10<Y1\right),1,0\right),$$

where IF returns 1 if the argument is true and 0 if it is not, and the command *AND* returns as true for the intersection of two events. Here, we have 1 to indicate that μ falls in the confidence interval and 0 to indicate that it does not.

(b) Drag from cell Z1 to cell Z1000.

(c) Type in cell Z1002

$$= SUM\left(Z1:Z1000\right).$$



Z1002 =SUM(Z1:Z1000)

	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
974	6.396261	10.46711	7.78021	9.857801	6.090073	7.101226		9.332174		8.455635	10.20871	1	
975	8.526452	10.56443	8.721382	7.749561	9.633492	14.95903		10.76856		9.892016	11.64509	1	
976	6.062668	7.799666	8.655319	10.66041	9.359127	10.22956		10.02564		9.149105	10.90218	1	
977	12.40357	8.502386	10.74802	8.428113	10.49938	10.16228		9.842062		8.965524	10.7186	1	
978	9.656631	11.483	10.28978	9.692321	9.080977	10.1556		10.05672		9.18018	10.93326	1	
979	11.43832	7.750446	8.362184	5.826916	8.90261	9.112434		9.333236		8.456697	10.20977	1	
980	8.775957	10.47672	8.82516	7.290552	9.711347	11.96046		10.28805		9.411508	11.16459	1	
981	10.91467	12.95214	11.64002	10.36859	9.244747	12.49892		10.47063		9.594088	11.34716	1	
982	12.76075	10.63292	8.917969	11.80488	10.91807	9.373742		10.55122		9.674682	11.42776	1	
983	11.19941	7.285482	7.944256	8.96877	12.06621	7.581895		9.459436		8.582897	10.33597	1	
984	8.76667	11.57457	12.06921	9.986344	10.03423	7.579158		10.08428		9.20774	10.96082	1	
985	10.63832	11.18604	13.5529	8.048054	13.17874	11.01456		10.43418		9.55764	11.31072	1	
986	11.53068	9.4665	9.798951	5.205275	7.879262	11.55127		10.1162		9.239662	10.99274	1	
987	8.36076	9.968851	11.02409	7.695489	12.55536	11.11341		10.70518		9.828638	11.58172	1	
988	14.1332	8.472086	10.47989	10.35051	9.047	10.52979		9.76711		8.890571	10.64365	1	
989	8.321144	12.51322	8.742519	6.884533	11.48722	8.591019		10.48845		9.611908	11.36499	1	
990	11.89286	9.808157	12.16401	8.255228	11.20514	10.76457		10.14419		9.26765	11.02073	1	
991	7.486395	8.452172	6.823352	10.74075	13.19249	11.16885		10.01334		9.136803	10.88988	1	
992	11.14839	14.03345	12.18897	10.92507	11.19736	10.71394		10.46581		9.589276	11.34235	1	
993	9.584949	9.160684	9.768922	10.41646	10.46893	10.07656		10.10931		9.232769	10.98585	1	
994	12.62761	9.865962	9.201408	13.98517	8.012431	8.605069		10.91734		10.0408	11.79388	0	
995	14.31439	6.795149	12.96898	11.16017	9.619232	9.304229		9.949164		9.072625	10.8257	1	
996	8.091067	9.412758	10.30523	9.680519	6.869167	9.45685		9.826434		8.949895	10.70297	1	
997	10.51194	8.945479	10.28991	10.60113	13.31538	13.97147		9.991942		9.115404	10.86848	1	
998	10.00524	11.99743	9.141885	7.584605	13.84032	9.599102		9.964638		9.0881	10.84118	1	
999	8.812036	11.7443	6.895692	10.66941	9.05927	10.44544		9.739674		8.863136	10.61621	1	
1000	10.13606	8.286156	10.98183	11.06723	10.68854	12.14643		10.60867		9.73213	11.48521	1	
1001													
1002												952	
1003													

Figure 4: Count the number of times μ falls in the confidence intervals

Point (iii) infers that about 950 confidence intervals would contain μ=10. What is your output in cell Z1002? As seen in Figure 4, the value is 952 which is just about 950.

Conclusion

It is well known that experiment is an indispensable component in studying a science subject. In statistics, it provides a compelling way to generate understanding and insight because of the fun that randomness shares with us. Now we can give affirmative answers to questions 1, 2 and 3.

Reference:

1. Borwein, J. and Bailey, D. (2008). *Mathematics by Experiments: Plausible Reasoning in the 21st Century*, 2nd Edition, Massachusetts: A K Peters, Ltd.

2. Gödel, K. (1951). Some Basic Theorems on the Foundations. Unpublished work. http://philoscience.unibe.ch/documents/kursarchiv/SS03/G%C3%B6del1951.pdf

3. Levine, D. M., Krehbiel, T. C. and Berenson, M. K. (2010). *Business Statistics: First Course*, 5th Edition, N.J.: Pearson Prentice Hall.

4. http://wps.prenhall.com/bp_levine_firstcours_4/36/9389/2403630.cw/-/2403649/index.html

5. Miller, I. and Miller, M. (2004). *John E. Freund's Mathematical Statistics with Applications*, 7th Edition, N.J.: Pearson Prentice Hall.

Invited Article: The Road to Fairness

This article considers two pure mathematical problems on how unfairness can be turned to fairness. We leave fairness/ unfairness issues due to historic/ cultural/ financial reasons to sociologists/ philosophers/ politicians/ economists.

The first problem is classical and is rather elementary; the second one, which is taken from the lovely book by *The Art of Mathematics* by Béla Bollobás [1], is more challenging.

Problem A: Coins have many uses apart from being a medium of payment to facilitate transactions. For instance, coin tossing is a simple and quick way of settling disputes in a civilized way or deciding between two options. To serve such purposes, the coin used has to be fair so that there is an equal chance of getting a head or a tail. Suppose a coin is given but we are not sure whether it is fair. What can we do? How can a possibly unfair coin be used "in a fair way"?

Problem B: We all know that by rolling two fair dice, the probability of getting a sum of 7 is the highest and that of 2 (or 12) is the smallest. Is it possible to create two (possibly different) loaded dice so that each of the sums 2, 3, . . ., 12 comes up with the same probability 1/11?

[1] Bollobás, B., 2006. *The Art of Mathematics,* Cambridge University Press

- *Give some thought before reading the solution!*

Problem A Solution

Consider the following simple procedure:

- Step 1: toss the coin twice and record the two results (e.g. Tail-Tail, or Head-Tail)
- Step 2: if the two results are the same, repeat Step 1; otherwise, if they are different, take the result of the first toss.

The examples below may help you better understand this procedure:

HT →**H**
HH → TH →**T**
HH → TT → TH → **T**

Using this procedure, the chance of getting a head or a tail will be the same. Try yourself by repeating the procedures many times, and you will obtain roughly the same number of H and T. Without using any calculation, can you explain why this procedure works?

Problem B Solution

Solution 1 Here, we describe an interesting approach by Béla Bollobás to tackle this question. Suppose that such two dice really exist. We use a_i , i = 1, . . . ,6, to denote the probability that the first dice shows i , and b_i , i = 1, … ,6, to denote the probability that the second dice shows i. Of course, we have $a_1 + . . . + a_6 = 1$ and $b_1 + . . . + b_6 = 1$. Next we define two functions by

$$P_1(t) = a_1 + a_2t + . . . + a_6t^5,$$

and

$$P_2(t) = b_1 + b_2t + . . . + b_6t^5.$$

We call these generating functions for the two sets of probabilities $\{a_i\}$ and $\{b_i\}$. They are polynomials of degree 5 because their leading coefficients a_6 and b_6 are strictly positive. Why should we consider these two functions? Let's consider the product of $P_1(t)$ and $P_2(t)$. By expanding the product and collecting the terms, we obtain

$$P_1(t)P_2(t) = a_1b_1 + (a_1b_2 + a_2b_1)t + (a_1b_3 + a_2b_2 + a_3b_1)t^2 + . . . + a_6b_6t^{10} :$$

Obviously, this is a polynomial of degree 10 because $a_6b_6 \neq 0$. The coefficients of this new polynomial has a very interesting property：

- the coefficient of t^0 is a_1b_1, which equals the probability of getting a sum of 2 when rolling these two dice;

- the coefficient of t^1 is $a_1b_2 + a_2b_1$, which equals the probability of getting a sum of 3 when rolling these two dice；
- the coefficient of t^2 is $a_1b_3 + a_2b_2 + a_3b_1$, which equals the probability of getting a sum of 4 when rolling these two dice；
- $\vdots$
- the coefficient of t^9 is $a_5b_6 + a_6b_5$, which equals the probability of getting a sum of 11 when rolling these two dice；
- the coefficient of t^{10} is a_6b_6, which equals the probability of getting a sum of 12 when rolling these two dice；

Can you see the pattern here? In fact, for $n = 0, 1, . . . , 10$ the coefficient of t^n of the polynomial $P_1(t)P_2(t)$ equals the probability of getting a sum of $n + 2$ when rolling these two dice.

Going back to our question, we want to design two dice such that each of the sums 2, 3, . . . , 12 comes up with the same probability of 1/11. This means that we have to find $a_1, . . . , a_6$ and $b_1, . . . , b_6$ such that

$$\mathrm{P}_1(t)\mathrm{P}_2(t) = \frac{1}{11} + \frac{1}{11}t + \frac{1}{11}t^2 + \cdots + \frac{1}{11}t^{10}$$

By using the identity

$$t^{\mathrm{m}+1} - 1 \equiv (t-1)(t^{\mathrm{m}} + t^{\mathrm{m}-1} + \cdots + t^2 + \mathrm{t} + 1)$$

we can rewrite the equation above as

$$11P_1(t)P_2(t) = \frac{t^{11}-1}{t-1}, \;\; t \neq 1.$$

As $P_1(t)$ (and also $P_2(t)$) is a polynomial of degree 5, the LHS of equation (1) has at least one real root (recall that complex roots of a polynomial always come in pair, one conjugate of the other, therefore any polynomial of odd degree must have at least one real root). On the other hand, the polynomial $t^{11} - 1$ has 11 roots (called roots of unity, all lying on the unit circle on the complex plane), among which 10 of them are complex numbers (a total of 5 conjugate pairs) and the only real root is $t = 1$, hence $\frac{(t^{11}-1)}{(t-1)}$ has 10 complex roots but no real root. Therefore, equation (1) cannot hold, no matter how we choose $a_1, \dots, a_6$ and $b_1, \dots, b_6$, and hence we can conclude that it is impossible to create two (possibly different) loaded dice so that each of the sums 2, 3, . . . , 12 comes up with the same probability.

Solution 2 Next we outline a more elementary way of solving this question, which makes use of the AM-GM inequality: for any positive numbers x and y, the arithmetic mean is always greater than the geometric mean:

$$\frac{x+y}{2} \geq \sqrt{xy}\,.$$

Suppose that the two dice (represented by $\{a_i\}$ and $\{b_i\}$) exist, the unknown probabilities $a_1, \ldots, a_6$; $b_1, \ldots, b_6$ should satisfy the following system of equations:

$$\begin{aligned} a_1b_1 &= 1/11, \\ &\vdots \\ a_1b_6 + a_2b_5 + \ldots + a_5b_2 + a_6b_1 &= 1/11, \\ &\vdots \\ a_6b_6 &= 1/11. \end{aligned}$$

We want to show that this system is *inconsistent*, meaning that it does not have any solution. To this end, observe that

$$\frac{1}{11} = a_1b_6 + a_2b_5 + \cdots + a_5b_2 + a_6b_1 \geq a_1b_6 + a_6b_1.$$

Now apply the AM-GM inequality to the two positive numbers a_1b_6 and a_6b_1, what do you get? Can you notice any inconsistency here?

二零一二至一三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
籌備委員會成員	陳秀騰先生，教育局
	陳家豪先生，政府統計處
	蔡佩霖女士，政府統計處
	楊琬婷女士，政府統計處

數學百子櫃系列 作者

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