機率統計入門

黃文璋

國立高雄大學統計學研究所

統計研習營

中央研究院統計科學研究所

106年7月5-7日

第一講

統計思維

前言

馬克吐溫(Mark Twain, 1835-1910)在1906年出版的自傳中說：

There are three kinds of lies: lies, damned lies, and statistics.

(有三種謊言：謊言，可惡的謊言，及統計)

◆為什麼要學數學？

生活及專業上會用到一些數學。

◆數學學通會有什麼特質？

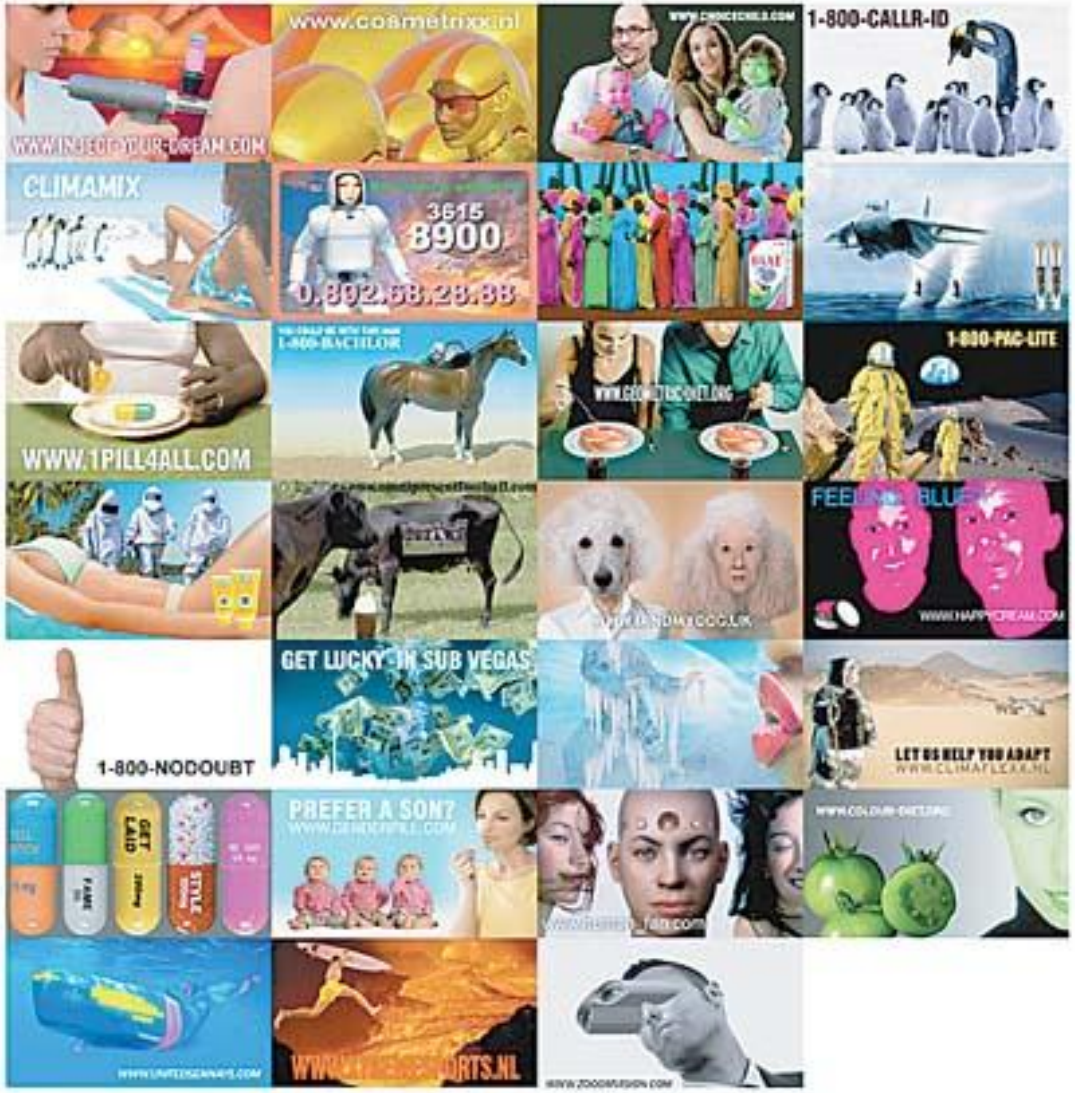
較有邏輯、計算較精準，…

- 近年來，統計學似乎愈來愈重要。
- 中小學數學課程的統計比重增加不少。
- 有人做決策時，非有統計不可。
- 有人對統計嗤之以鼻。
- 有人以為統計就是數學。
- 有人強調統計與數學完全不一樣。

- 統計的內涵似乎不易被人掌握：什麼是很有
  - 統計頭腦？
  - 統計細胞？
  - 統計素養？

PRADA

OBVIOUS CLASSICS #1



上課啦 T恤圖案古怪趣味嘲弄人類恐懼及欲望

PRADA春夏搞起統計學文化

藝術家跨界替時尚品牌設計的限量T恤，往往就是最能表達穿著者個性的搶眼之作。春夏PRADA新推出一系列統計學文化T恤，就算搞不懂藝術家口中什麼統計學和美術的關係，但至少知道這些企鵝、藥丸、機器人圖案真的可愛到不行，透明密封袋包裝也是古怪又趣味。

其實，統計學文化來自名建築師Rem Koolhaas等人組成的團體AMO，他們將統計數據美學作為創作材料，將有趣的統計數字混合各種的圖片，替PRADA創作出Obvious Classic這些圖案T恤，嘲弄人類的恐懼及欲望，共有27款，每件6千元。

(中國時報96年4月27日 徐亦橋)


GET LAID
STYLE
1-800-CALLR-ID
WWW.UNITEDSEAWAYS.COM
LET US HELP YOU ADAPT
WWW.CLIMAFLEXX.NL

統計學究竟在做什麼？

統計學裡所能達到的是：

1. 允許誤差下的機率保證，
2. 允許誤差下的無罪推定。

- 數學裡探討必然性。
- 統計裡處理隨機性。
- 允許誤差，沒有誤差反令人懷疑。
- 統計裡的保證，都是機率式的。
- 通常所能保證的機率，不但不是百分之百，還附有誤差。

例.

有百分之九十五的機率，某飲料的容量，介於326cc至331cc間。

- 很少經由統計去證明那件事必是對的。
- 探索真相？
- 真相留給上帝！
- 在隨機世界，真相常難以大白。
- 一切都是假設，只看你接受那一個。
- 接受或拒絕，採類似刑事訴訟法第154條無罪推定的精神。

周公恐懼流言日，

王莽謙恭下士時，

若使當時身便死，

一生真偽有誰知？

- 驚悚(Primal Fear，1996)由李察吉爾及愛德華諾頓等主演。
- 描述一名律師如何幫助一殺人犯從法庭脫罪。

- 一種統計方法，常對應人們的某種思維。
- 機率及誤差，構成統計思維之兩大支柱。

- 因而發展出統計學裡所著重的幾項要點：
  - 善用資訊
  - 了解變異
  - 相信機率
  - 合理估計
  - 無罪推定
  - 紙上談兵

1. 善用資訊

- 什麼是data?
  - 資料、數據，從調查、實驗或研究中獲得資訊。
  - A general term for observations and measurements collected during any type of scientific investigation.

◆在柯南道爾著的桐山毛櫸山莊，

福爾摩斯說：

Data! data! data!

he cried impatiently.

I can't make bricks without clay.

- 做決策不能沒有data，算命者所倚賴的也是data：
  - 要收集很多人的命運， 並按面相、八字等分類。
- 算命是在做統計實務？
- 讓數據說話：

是否真了解數據所說的話？

單身該多繳稅？

民國101年6月5日，前衛生署長楊志良投書媒體：

我主張單身的朋友多繳一些稅，目的是讓家庭可以少繳一些，減輕負擔、獲得資源，安心養育下一代。 等到這些單身的人步入老年，成了需要被照顧的弱勢，當年他們幫忙養育，後來長大成熟的下一代，就成了有能力的人，回過頭來付出、回饋，擔負起社會正常運轉需要的勞動力，以及繳稅等等，讓年老的單身朋友可以安心養老，…

單身者願意嗎?

有偶者比單身活得久？

民國101年6月16日，內政部公佈：

去年單身者平均死亡年齡為53.68歲，有偶者死亡年齡則為70.12歲，單身者比有偶者短少了16.44年壽命。

對於單身者平均餘命比有偶者少了16歲，戶政司長表示，這趨勢從過去就是如此，國外也有相關研究指出，單身者可能較乏人照顧，心理、環境因素都會影響，才會導致平均餘命比較短。

內政部統計處長解釋，除了心理、生理狀態因素外，因統計納入15至30歲者死亡數據，該年齡層大多未婚，也會拉低單身者平均死亡年齡。

- 統計學家的本事要能發揮，就得善用資訊。
- 有用的資訊不見得容易看出。

THE SMITH FAMILY
What is the probability that the kneeling child is a girl?

跪著的小孩是女孩之機率？

假設生男生女之機率皆為1/2。

答案是1/2？

這張圖片是否提供有助判斷的資訊？

有人敲門，是男是女？

女的機率為

0.5,

0.9,

⋮

。

- 機率值會變，是機率的一特性。
- 視新的資訊產生，對一事件機率之判斷，也隨之而變。
- 亦有人堅持是女生的機率為0.5。

- 草進了牛肚 ⟹ 牛奶。
- 資料進了統計學家手中⟹資訊⟹決策。

2. 了解變異

由常數至隨機

- 小學數學：學生每人重32公斤，求…
- 大學數學：學生體重有一分佈，求…

- 宇宙的運轉，必然性與隨機性交錯進行。
- 對未來，知道大致會發生那些事，但又無法完全掌握。

- 必然性使人們願意事先好好準備。
- 隨機性使人們對未來，充滿著盼望與戒慎恐懼。
- 光有必然性，毫無變異，對未來缺乏盼望，將少了努力的動機。
- 光有隨機性，只靠運氣，將失去企圖心。
- 三分天注定，五分靠打拼，兩分靠運氣。

- 由於變異無可避免的存在，要了解變異，設法減少變異。
- 抽樣調查某產品的良品率。

良品：1表示，不良品：0，

得到 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, 0, …。

以平均 l 出現的次數，估計良品率。

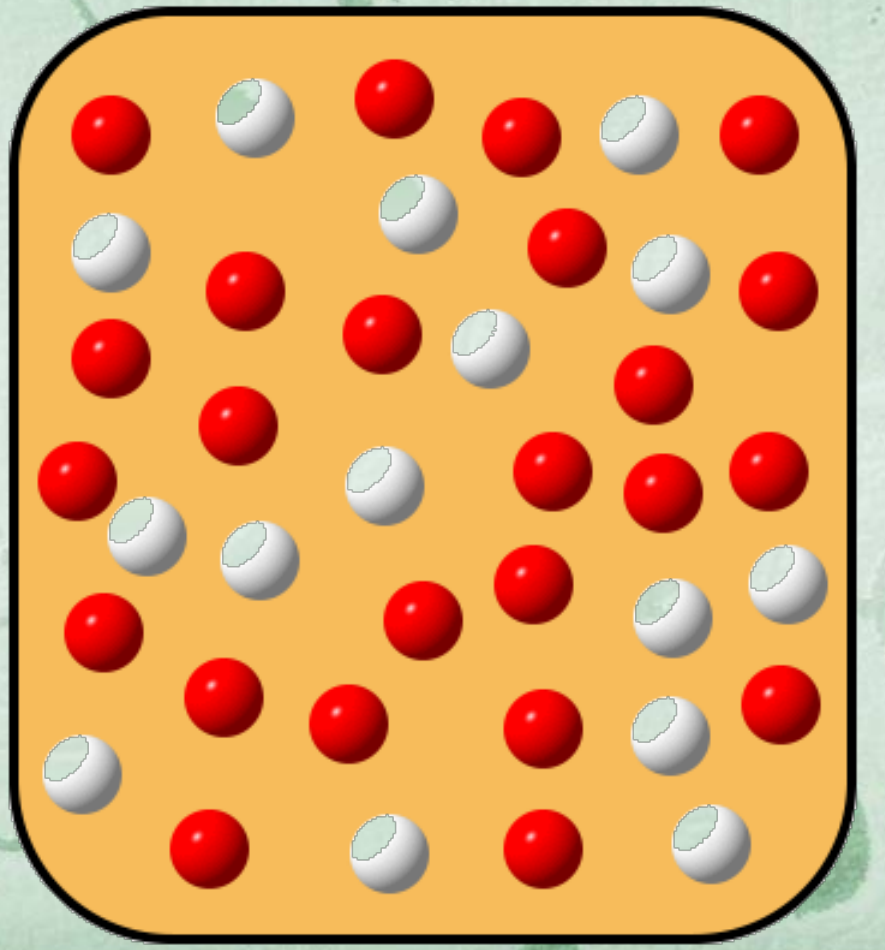
估計要夠精準，樣本數便要較大。

- 雖世事多變，但萬物有常，存在隨機法則。
- 看似沒有規律的0,1數列，其實被大數法則規範。
- 不論樣本數多大，都不能保證前述平均值，剛好等於良品率。
- 誤差究竟有多大？

- 數學中常求近似值。
- 必須要能給出誤差大小。
- 中央極限定理指出，量測所得之誤差有常態分佈。
- 誤差大小是隨機的。
- 但誤差之散佈情形，則能描述。

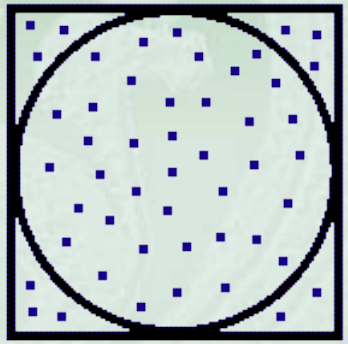
- 統計裡，常在做預測、做估計。
- 做以偏概全的事。
- 若樣本實在太偏差，便是以管窺天，見不到全貌。

- 袋中白球所佔比例？



◆**例.** 如何估計圓周率π？

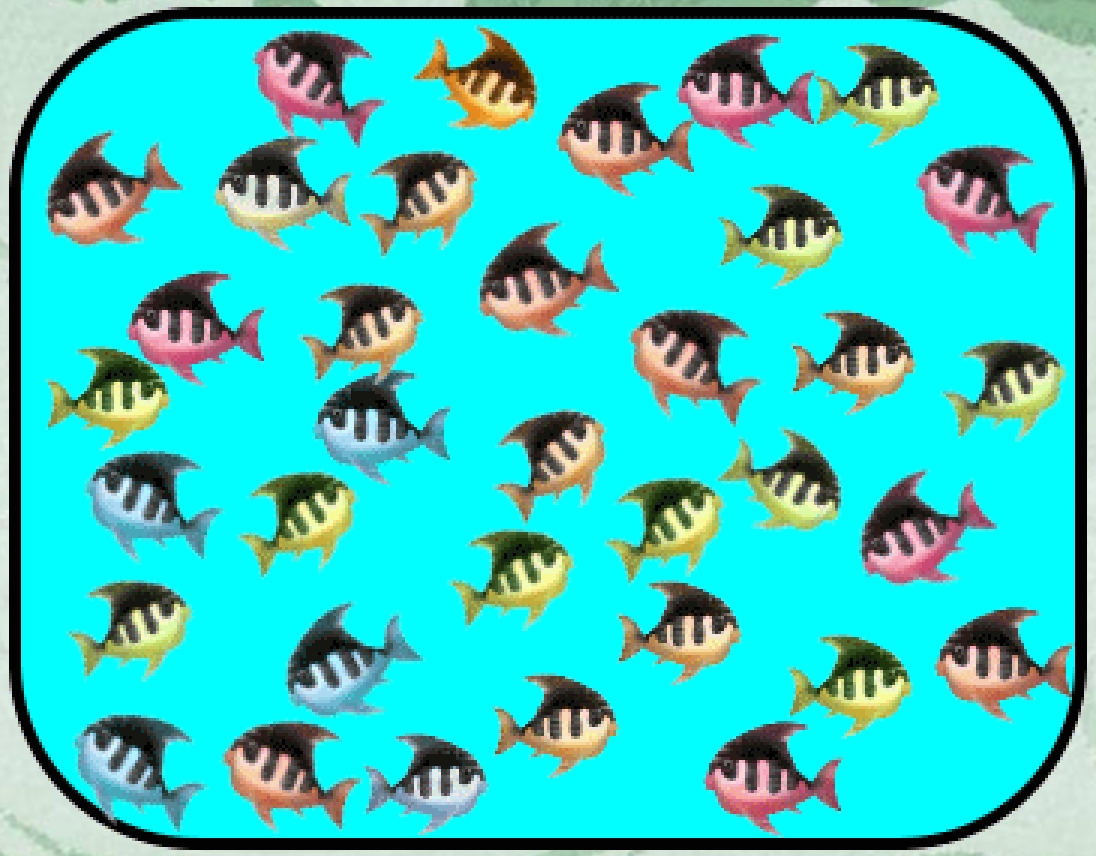
隨機灑芝麻，n粒落進正方形內，a粒落進圓中。



$$\frac{\pi}{4} = \frac{\pi r^2}{4r^2} = \frac{a}{n}$$

$$\pi = \frac{4a}{n}$$

估計圓周率π



池中有多少魚？

◆調查若與人有關，不容易做：

人會改變想法，
不見得會與調查者合作，
不同群體的人想法差異很大。

3. 相信機率

機率是統計上騙人的東西，許多事情要重複做100次才有機率可言。懷孕不可能100次，每次懷孕生雙胞胎機率是1/89，但單次懷孕生雙胞胎機率若不是0%，就是100%。就好像問我，50元銅幣丟到地上一次，是蘭花機率有多少？事實上，50元銅幣丟到地上，不是總統府，就是蘭花。如果丟到地上 100次，那麼機率就會接近50%。如果丟到地上1次，蘭花的機率，若不是0%，就是100%。

(取自網路)

- 機率的意義是什麼？
- 投擲骰子，或抽籤，常以相同的可能性來解釋。
- 棒球的打擊率，常以相對頻率來解釋。

理論基礎：大數法則。

針對的現象：可以重覆觀測。

- 無法重覆觀測時，常用主觀機率：

追女孩成功的機率。

- 以公理化的方式引進機率。

- 事件在發生前，或知道結果前，才會談機率。
- 觀測一事件，
  - 結果是不發生，或發生，
  - 而非機率0%，或100%。

◆有二銅板：

銅板 A 出現正面之機率為0.3，

銅板 B 出現正面的機率為0.2。

問：

- 0.3是什麼意思？丟10次會得到3次正面？丟10,000次得到3,000次正面？
- 0.3＞0.2，若二銅板各丟10次，銅板A之正面數＞銅板B？
- 兩銅板各丟一次，共得到一正面之機率為0.3＋0.2＝0.5？

有些機率沒有想像的小

例.

每23人中有兩人生日相同之機率>0.5。

每40人中有兩人生日相同之機率約0.891。

每56人中有兩人生日相同之機率約0.988。

每64人中有兩人生日相同之機率約0.997。

連生2兒子後，較容易生出女兒？

- 有位婦女已連生2個兒子，很想生個女兒。
- 試一下，那有運氣那麼壞？

約有一半的人成功，約有一半的人失敗。
失敗者中，有些還會再試一次。
又有約一半的人成功。

⋮

- 賭戲對賭客並不公平，何以許多人一上了賭台就下不來？

- 情況不利 ⇒ 那有運氣那麼壞，該轉運了。
  - 再玩若仍輸 ⇒ 下次更該贏了。
  - 若幸運贏了 ⇒ 開始翻身了。
- 若情況有利 ⇒ 手氣正順，怎可停止？
- 除非是一直輸贏不太多(機率不大)，讓人覺得此賭戲沒趣。

- 新聞媒體多半只報導有人樂透彩中大獎，或在賭場大贏的新聞。
- 人有選擇性記憶的傾向。在賭之前向神明祈求，大部分的時候沒有效果。但若贏了，可能真覺得神明聽了自己的祈求。

以投擲銅板為例

- 持續投擲一公正銅板10,000次，令

$$S_n = X_1 + \cdots + X_n,\ n = 1, \cdots, 10,000 \ ,$$

$$P(X_i = 1) = P(X_i = -1) = \frac{1}{2} \ 。$$

◆以布朗運動(Brownian motion)的結果來估計：

$$P(\max_{1\le n\le 10,000} S_n \le a) \approx P(\max_{1\le s\le 1} X(s) \le \frac{a}{100})$$

$$= P(|Z| \le \frac{a}{100}) \approx \begin{cases} 0.0796, a = 10, \\ 0.1586, a = 20, \\ 0.3830, a = 50, \\ 0.6826, a = 100, \end{cases}$$

其中 Z 有 $\mathcal{N}(0,1)$ 分佈，$\{X(t), t \ge 0\}$ 表一標準的布朗運動。

仍以布朗運動的結果來估計：

$$P(\frac{\text{正面領先次數}}{\text{投擲次數}} \geq x) \approx 1 - \frac{2}{\pi}\arcsin\sqrt{x}。$$

當$x = 0.993$，機率約為

$$1 - \frac{2}{\pi}\arcsin\sqrt{0.993} \approx 0.0533。$$

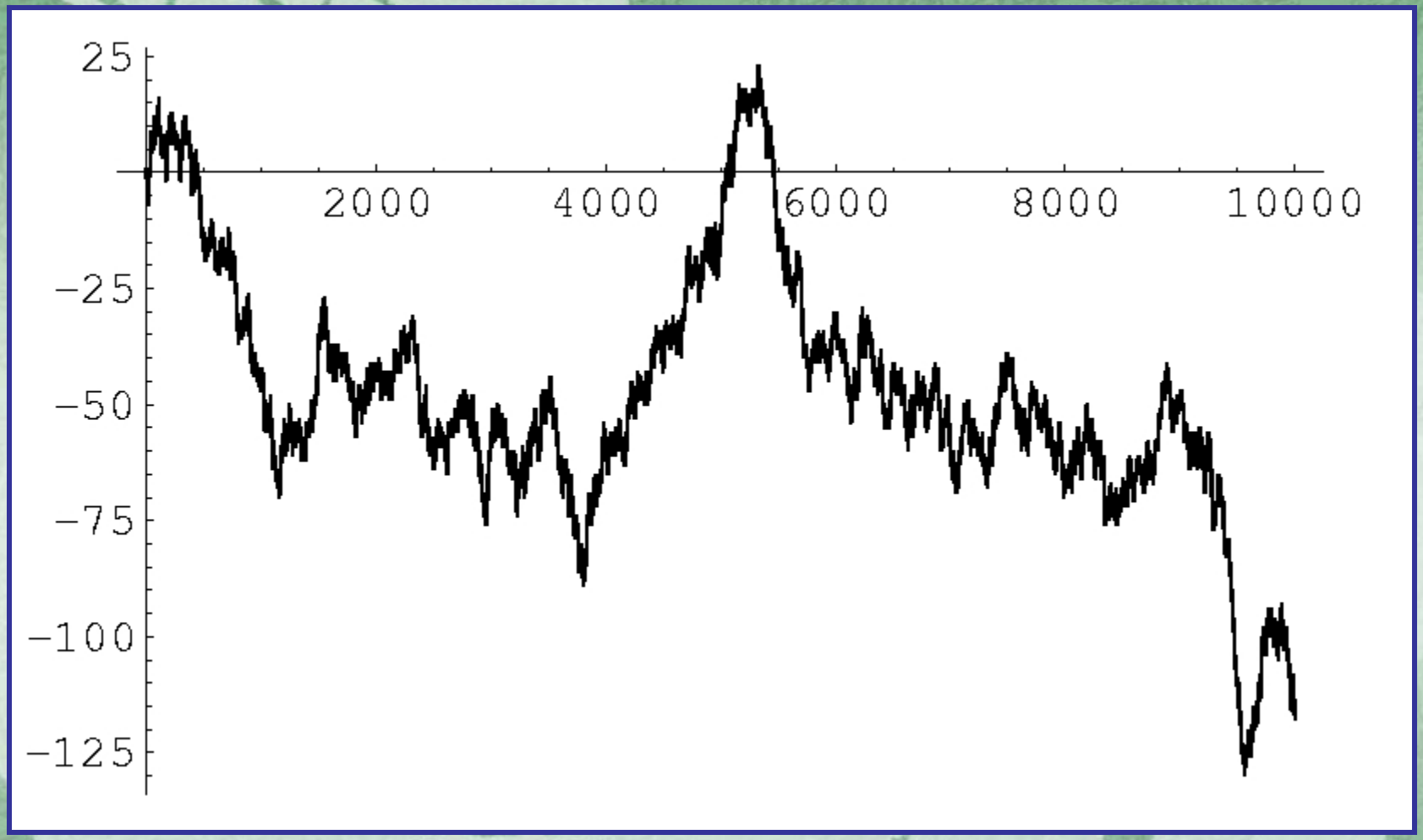
$$\Rightarrow P(\text{有一面領先次數} \geq 9930) \approx 0.0533 \cdot 2 = 0.1066。$$

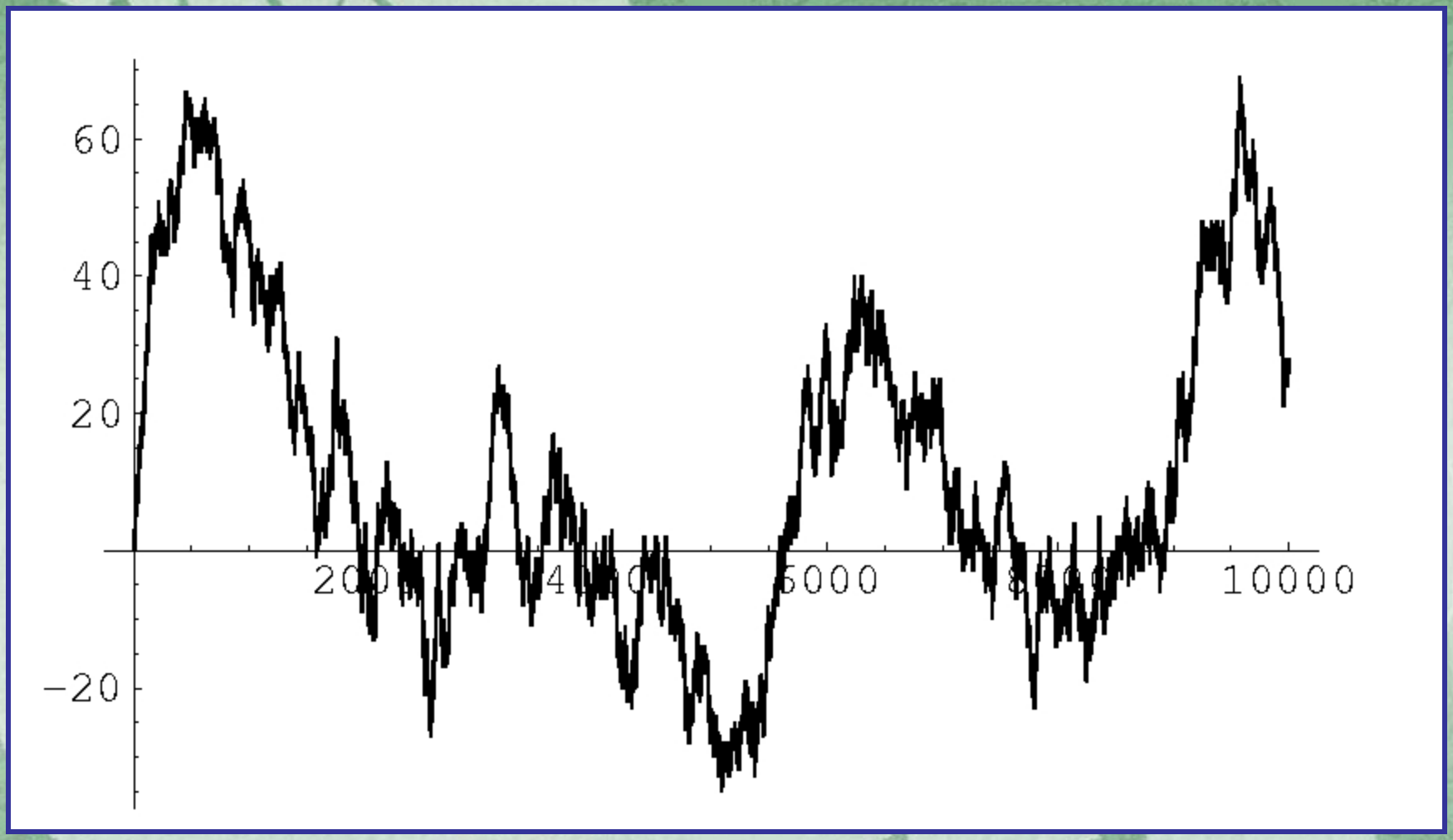
- 底下為五個模擬圖，橫軸為n, 縱軸為S_n，其中$S_i > 0$表正面領先，$S_i < 0$表反面領先。

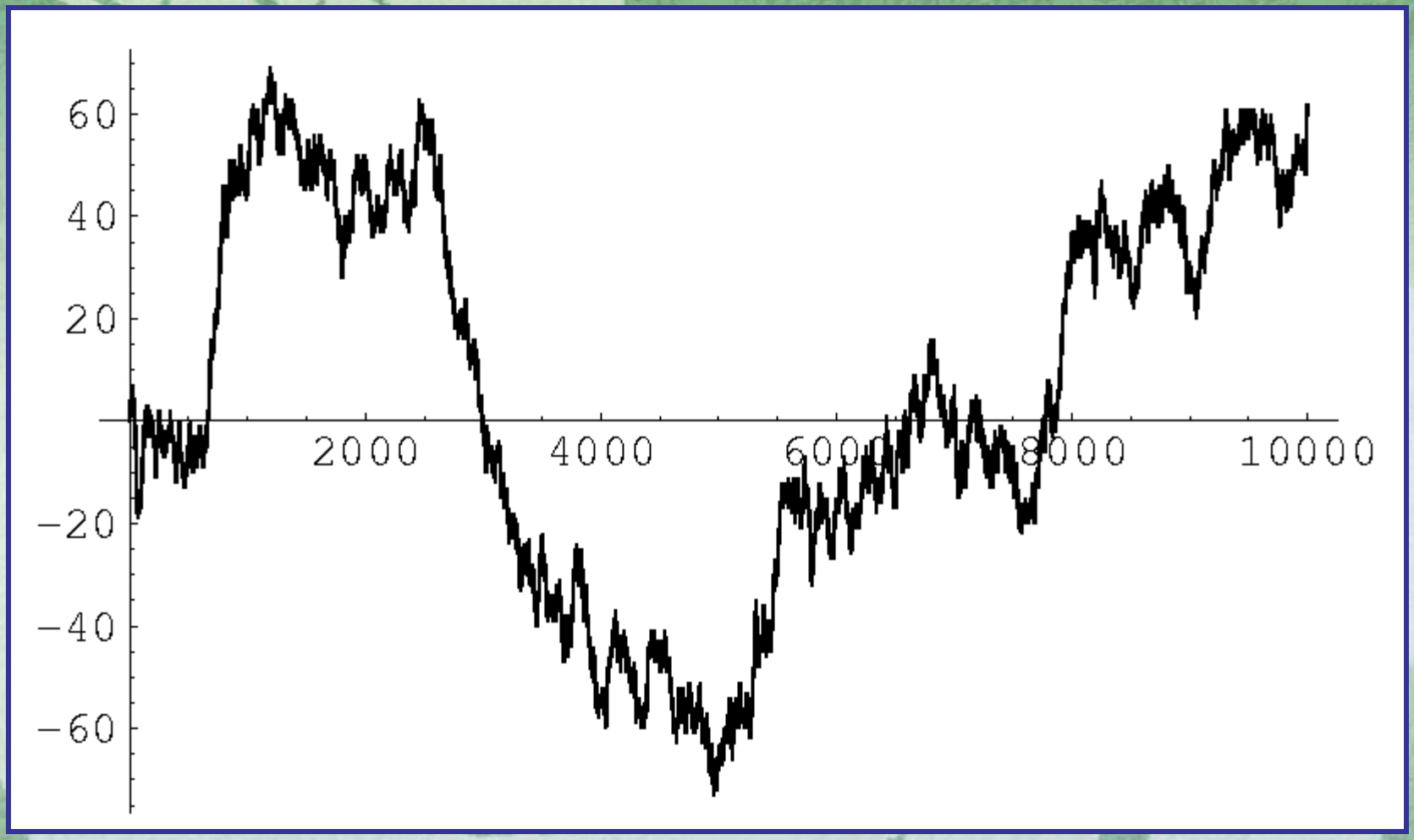
$\{S_n, n \geq 1\}$ 構成一隨機漫步(random walk)。

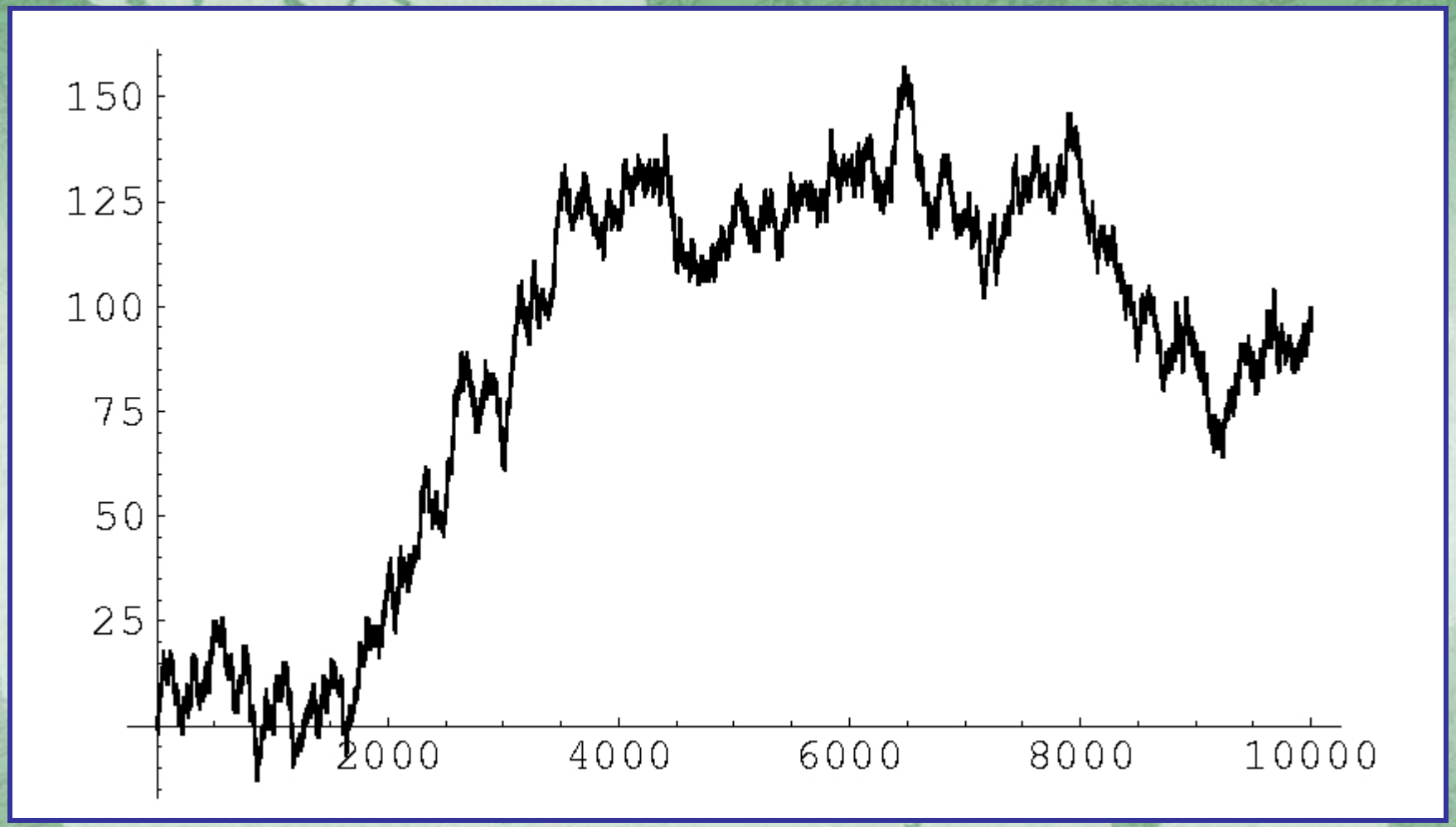
注意

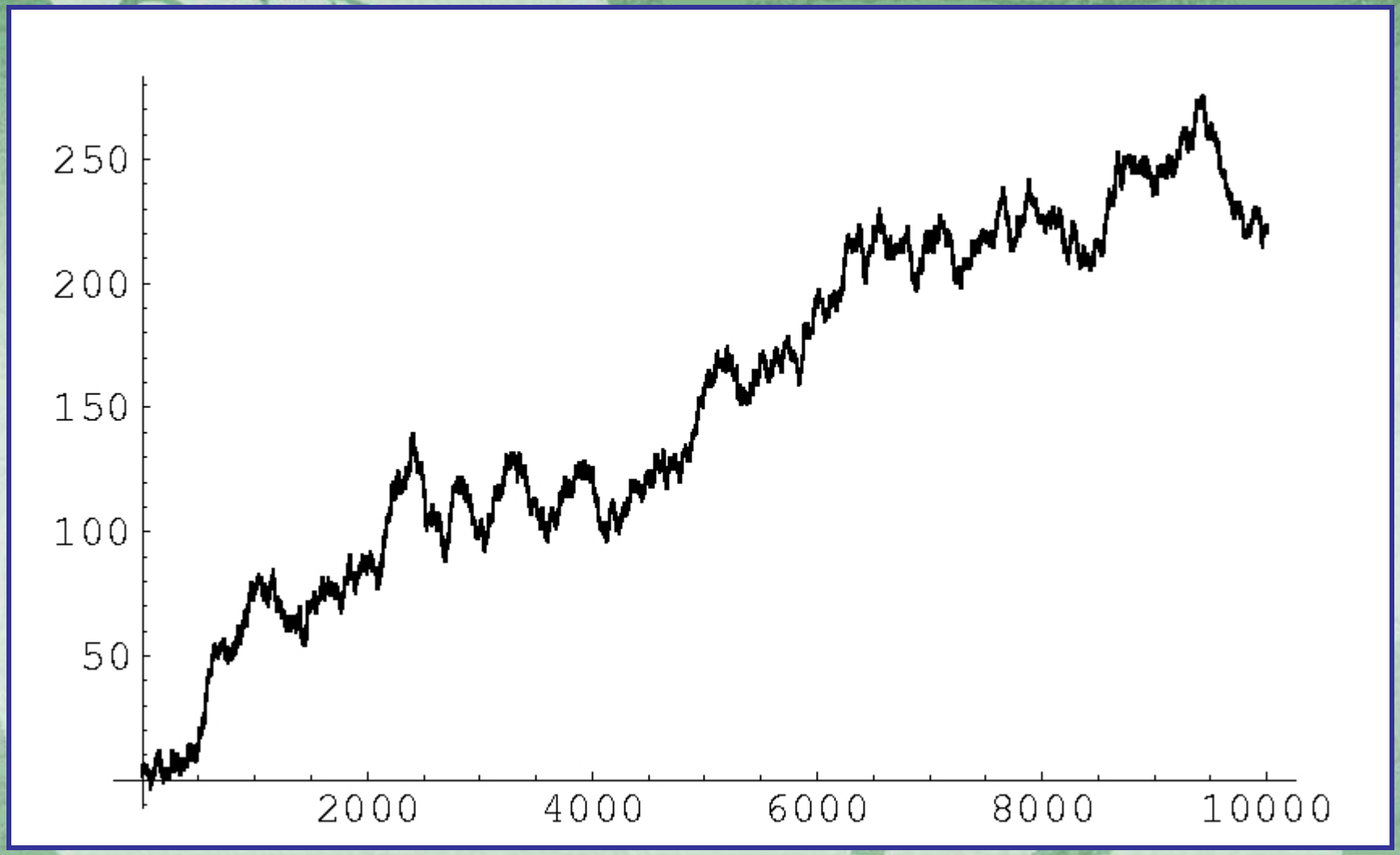
$$E(S_n) = 0, \ \forall n \geq 1 \ 。$$


25
2000
4000
6000
8000
10000
−25
−50
−75
−100
−125


60
40
20
−20
10000


60
40
20
2000
4000
6000
8000
10000
−20
−40
−60


150
125
100
75
50
25
2000
4000
6000
8000
10000


250
200
150
100
50
2000
4000
6000
8000
10000

均勻分佈下的不均勻

- 做芝麻餅，希望芝麻很均勻，學過均勻分佈，隨機撒？

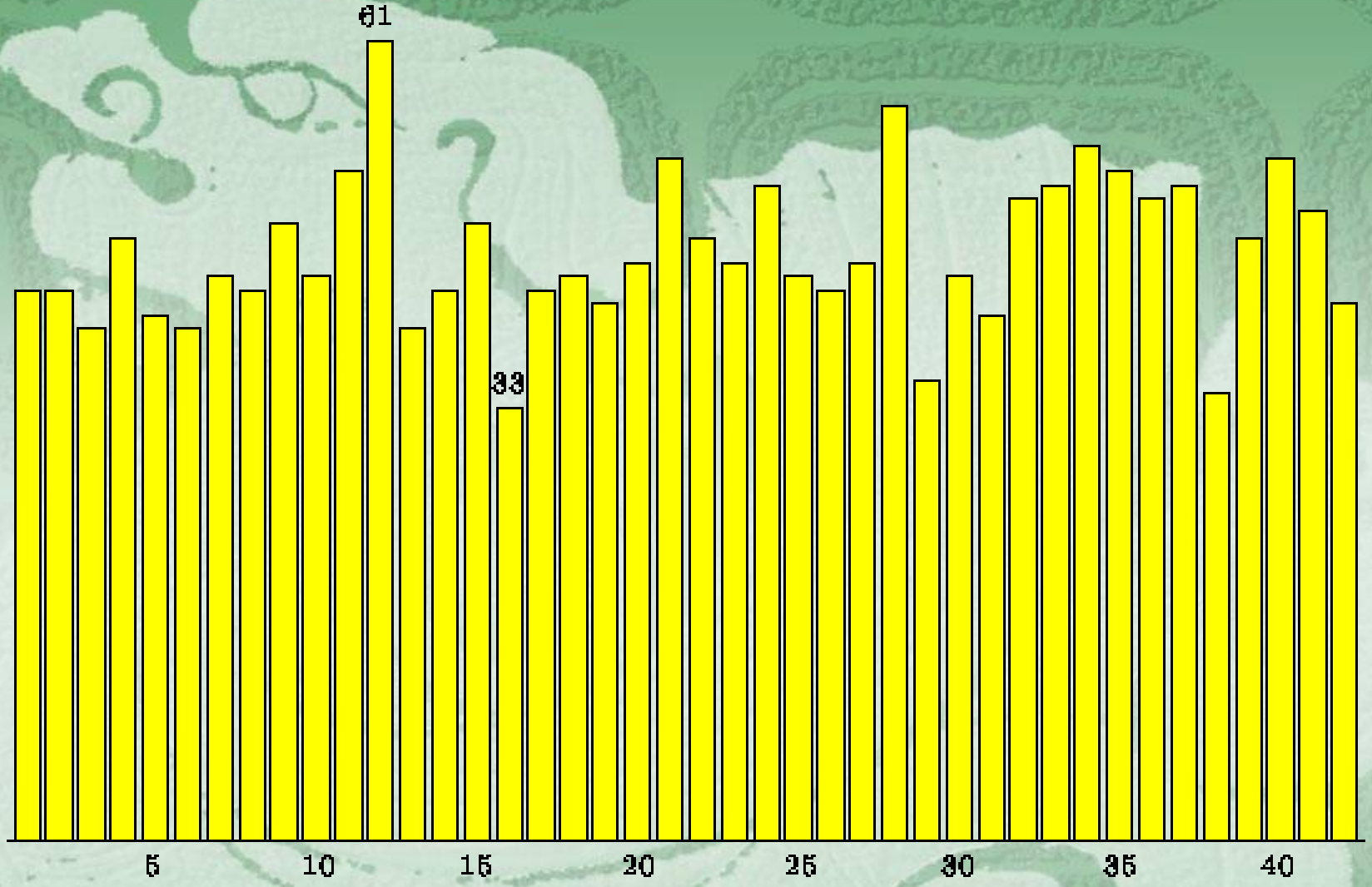

jodie foster / anthony hopkins / scott glenn
the silence of the lambs
from the terrifying best seller

◆在電影沈默的羔羊(The Silence of the Lambs, 1991)裡：

Doesn’t this random scattering site scenes desperately random, like the elaborate on bad liar.

這些隨機散佈的地點，不是極度地隨機嗎？就像差勁的騙子精心設計的謊言。

- 看起來很隨機，反而會像精心設計的謊言！老師點名如果是隨機地點，是很難每次都點不同的人。樂透彩，如果每個號碼出現的頻率都一樣，或連續幾期開出的號碼都不一樣，反而才該懷疑。



91.1.22~94.1.21共314期，樂透彩頭獎號碼出現頻率，平均≒44.857

絕對誤差變大，

相對誤差變小。

巨數法則
(law of truly large numbers)

- 小機率遇上大樣本，發生就不令人驚訝。

小機率事件屢屢發生

- 某公司員工一萬人，年終摸彩，頭獎一名，中獎比投擲銅板連得13個正面還難：

$$\frac{1}{2^{13}} = \frac{1}{8{,}192},$$

但每年都有人中頭獎。

- 美國紐約時報曾在第一版(1986年2月14日)報導一位名叫Adams的女士第二度贏得紐澤西州的樂透彩頭獎。1985年10月24日，她第一次得390萬美元，第二次則得150萬美元。這是紐澤西州第一次有人得到兩次百萬美元以上獎金的樂透彩。

- 第一次樂透彩是39取6，中頭獎之機率為

$$\frac{1}{\binom{39}{6}}=\frac{1}{3{,}262{,}623}。$$

- 第二次樂透彩是42取6，中頭獎之機率為

$$\frac{1}{\binom{42}{6}}=\frac{1}{5{,}245{,}786}。$$

◆樂透彩發行商說一生中中兩次頭獎之機率為

$$\frac{1}{3{,}262{,}623}\cdot\frac{1}{5{,}245{,}786}\approx\frac{1}{1.7115\cdot 10^{13}},$$

約17兆分之一。

◆這樣算對嗎？

- 上述計算是假設Adams只在兩週買彩券：玩法39取6，及玩法42取6，各買一張。
- 事實上Adams每週買好幾張且買了好幾年。而且在第一次中頭獎後，便增加買的張數。若在39取6的玩法裡，及在42取6的玩法裡，每週各買2張，則每週大約有百萬分之一的機率中頭獎。在4年(約200期，每週一期)裡，一次頭獎皆未中的機率約為

$$\left(1-\frac{1}{1,000,000}\right)^{200} \approx e^{-\frac{200}{1,000,000}} = e^{-\frac{1}{5,000}}。$$

- 中一次頭獎的機率約為

$$\frac{1}{5,000}e^{-\frac{1}{5,000}} \approx \frac{1}{5,000}。$$

- 中兩次頭獎的機率約為

$$\frac{1}{2}\left(\frac{1}{5,000}\right)^2 e^{-\frac{1}{5,000}} \approx \frac{1}{50,000,000},$$

約5千萬分之一。

- 至於一個人一生(以30年，1,500期計)會中兩次頭獎之機率則約為

$$\frac{1}{2}\left(\frac{1,500}{1,000,000}\right)e^{-\frac{1,500}{1,000,000}} \approx 0.000001125,$$

超過百萬分之一。

- 不論是5千萬分之一，或百萬分之一的機率，都很小。但紐澤西州有841萬人，若其中有1百萬人，一生中每期皆如前述方法買彩券，則會有人一生中至少中兩次頭獎之機率便很大了：

平均有$1,000,000 \cdot 0.000001125 = 1.125$(人)，

機率約 $1-\left(1-0.000001125\right)^{1,000,000}$

$\approx 1-e^{-1.125} \approx 0.6753$。

何況美國有50州！

◆若全美有5千萬人，每期皆如前述方法買彩券，則大約4年，便會有人中兩次頭獎了：

$$\text{平均有} 50,000,000 \cdot \frac{1}{50,000,000} = 1(\text{人})，$$

$$\text{機率約} 1-\left(1-\frac{1}{50,000,000}\right)^{50,000,000}$$

$$\approx 1-e^{-1} \approx 0.6322。$$

◆1998年，Humphries第二度贏得賓州樂透彩頭獎，兩次合計中680萬美元的獎金。

千萬不要小看大數的威力！

機率是千秋的事

- 毛澤東：

  一萬年太久，只爭朝夕。

- 對於機率：

  不爭一時而爭千秋。

- 觀測次數夠多後，機率的威力就顯現。

- 銅板，出現正面機率為0.6。投擲若干次，那一面出現較多次便贏：

要選那一面？

- 假設選反面：
  - 投擲1次，有0.4的機率贏。
  - 投擲10次，約有0.166的機率贏。
  - 投擲100次，贏的機率僅約0.016。
  - 投擲1,000次，贏的機率約 $4 \cdot 10^{-10}$。

4. 合理估計

包公審錢案

所羅門王的智慧

發展出？

最大概似法

依發生機率之最大者來決定估計值。

- 日常生活常以此法來做決策：
  - 教室的玻璃窗破了，小明平常最喜歡亂丟東西，最可能是他。
  - 警方辦案，從有前科者開始調查。

公平嗎？

- 不從最調皮，有前科者開始查，從其他人開始查，不是更不合理？

- NBA30支職業籃球賽，依全球季的勝率，決定那16隊參加季後賽，及排定主場優勢。
- 以相對頻率來估計的想法 $\Longrightarrow$ 動差法。

$X_1, X_2, \cdots, X_n \sim$ i.i.d. $Ber(p)$，

以

$$\frac{1}{n}\sum_{i=1}^{n} X_i$$ 估計p。

- 一表人才、彬彬有禮、姣好外表、學歷不錯
  ⟹ 無言推薦！
- 遲到、草率，… ⟹ 開始扣分。
- 有時依據過去經驗，或主觀上的認定，會有一些事先的看法。
- 再依觀測後的結果，調整原先的看法

貝氏法。

◆其他估計法

最小平方法，…，

稍後再談。

5. 無罪推定

- 人們常求公平或公正。
- 兩人分蛋糕，如何分？

你切我選。

如何設計出讓人覺得較公正的判決機制？

◆歐陽修的瀧岡阡表：

➢其父生前為官批文，對於死囚：

求其生而不得，則死者與我皆無恨也。

◆民國92年，刑事訴訟法修正為無罪推定原則：

被告未經審判證明有罪之前，推定其為無罪。

◆考試有20道單選題，每題4選項。

二情境：

1. 如果你們二位沒作弊，怎會有15題對的一樣？
2. 如果你們二位沒作弊，怎會有15題錯的一樣？

- 假設檢定，是對隨機現象，做決策之一重要依據。
- 現代統計學之創始者費雪(R.A. Fisher, 1890-1962)曾提出下述故事：

  在1920年代後期，有位女士宣稱，奶茶的調製順序對風味有很大的影響。即

  茶加進牛奶，和牛奶加進茶，喝起來完全不同。

摩登神農氏

先牛奶？先茶？

共20杯

台灣兩千三百萬人，如果每人每天猜一遍，每天約有23人會猜對！

懷孕者之尿液可使種子提前發芽

- 土耳其國立安卡拉大學醫學院婦科系教授庫克表示，早在西元前2200至2000年，古埃及人已能不靠藥劑檢驗出女性是否懷孕。

提早發芽

懷孕

未提早發芽

未懷孕

- 怎樣算提前？
- 古埃及人如何操作？

- 在費雪的故事裡，若只拿一杯奶茶讓那位女士喝，她說對先放茶或先放牛奶，會相信她真有分辨能力嗎？
  - 兩次皆說對呢?
  - 連續10次皆說對呢？
  - 20次中錯一次呢?
  - 20次中錯兩次呢?

  分辨能力指她每杯講對的機率大於$\frac{1}{2}$？
- 我們對犯錯有些容忍度，程度多大，因人、因情況而異。

- 在數學裡，一命題，一旦被證明是對的，就毫無疑問地成立。
- 數學上可以寫

  假設

  $n \geq 3$為一整數，且x, y, z皆不為 0 ，(A)

  試證

  $x^n+y^n=z^n$無整數解。(B)
- 在隨機世界，事件往往不知真偽。

- 到底該女士能否分辨奶茶是先放奶還是先放茶，即使20次皆說對，仍有人不信她有此能力。
- 因此我們不會說：

  試證某女士有分辨奶茶是先放奶或先放茶的能力，

  或：

  試證某女士無分辨…。

- 數學家因相信在條件A下，B是對的，於是去證明。
- 對奶茶問題，我們相信什麼？由於該女士希望人家相信她有分辨能力，因此

  假設該女士無分辨能力。
- 然後拿20杯讓她分辨。先設定一能忍受的推論錯誤機率α，再求出在無分辨能力的假設下，講對次數會這麼多的機率。
- 如果機率小於α(即這是較不尋常)，則拒絕原假設，即接受該女士有分辨能力，否則接受原假設。

假設檢定

- 波蘭人奈曼(J. Neyman, 1894-1981)及英國人皮爾生(E.S. Pearson, 1895-1980)，1933年，給出奈曼–皮爾生引理。
- 其架構中，

  虛無假設(H_0)：通常表現況，
  對立假設(H_a)：表我們傾向相信，或希望它是對的。

◆虛無假設(現況)要被保護，除非證據夠強，否則不輕易推翻：

朝令夕改非統計的精神。

例.

虛無假設：

- 樂透彩沒有做弊，
- 喝綠茶不能減肥，
- 模特兒A沒有服用毒品。
- 某縣長沒有A錢。

對立假設：

- 樂透彩作弊，
- 喝綠茶能減肥，
- 模特兒A服用毒品。
- 某縣長有A錢。

- 何以稱為虛無假設？
  - 為一空的假設。
  - 接受虛無假設表實驗失敗。
  - 大家有興趣的是對立假設。
  - 常報導接受虛無假設的媒體少有人要看。

- 民國97年各高中推薦在校成績前 1 %的學生參加繁星計畫。常春藤高中，有254名高三學生。報上說：

  理論上，可有 3位學生獲推薦，結果該校共推薦31人，且最後有11人錄取，明顯高過正常比例。

- 明顯高過正常比例，用統計語言來說，為一顯著事件。

- 統計的功能只能到此，究竟成績是否作假，須調出原始成績才知。

◆人們常提到關係，有兩種關係是數學中常出現的：

因果關係，

函數關係。

◆統計裡尚有一重要的關係：

獨立。

獨立

- 某女孩找對象，開出條件：

1. 35-55歲(機率2/3)，
2. 年收入80萬元以上(機率1/2)，
3. 有趣、熱心、體貼(機率1/2)，
4. 身高170cm以上(機率1/2)，
5. 碩士以上學歷(機率1/2)，
6. 聰明、好看(機率1/3)，
7. 不抽菸、不喝酒(機率2/3)，
8. 喜愛音樂、大自然(機率1/4)，
9. 會照顧家庭(機率1/2)。

$$P(\text{找到})=\frac{1}{864}。$$

- 每週相親1個，平均要

$$864/52\approx16.6年。$$

合理嗎？

- 統計裡常在做估計、做預測。那種情況下最難估計(預測)？

獨立！

- 樂透彩開獎，各次頭獎號碼為獨立。

統計與因果關係

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

計聞會說話 韓媒體人壽命短。

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

◆平均壽命：

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

◆例2. 100年4月9日有則新聞：

天天逛街 延年益壽

逛街有益身心健康，是愛血拼的女士編出的嗎？

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

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

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

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

氣泡飲料易致癌？

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

都與食道癌有關？

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

◆例4. 1972-1974年，英國進行一有關甲狀腺疾病與心臟病的研究。二十年後做後續的追蹤研究。Appleton(1996)一文對其中婦女抽煙與死亡的數據做一些分析，部分數據列在表1：

不抽煙者之死亡率明顯地較高。

經統計檢定，得到不抽煙者之死亡率高於抽煙者之死亡率的推論。

表1 婦女抽煙與死亡數據

	抽煙	不抽煙	總數
死亡	139	230	369
存活	443	502	945
總數	582	732	1,314
死亡率	23.88%	31.42%	

- 統計上的推論，只是證實抽煙與不抽煙兩群人的死亡率不同，且不抽煙那群人之死亡率較高，並未證實抽煙是造成死亡率較低的原因。
- 可樂銷售量較大時，到醫院腸胃科就診人數常亦增加？

- 仔細分析數據，有一重要的變數不能忽略：

  年齡。

- 初次調查時，高齡者中較少是抽煙的。

- 對年齡分群得表2。

表2 對年齡分群之抽煙與死亡的數據

年齡分群	18-24		25-34		35-44		45-54		55-64		65-74		≥ 75	
	+	−	+	−	+	−	+	−	+	−	+	−	+	−
死亡人數	2	1	3	5	14	7	27	12	51	40	29	101	13	64
存活人數	53	61	121	152	95	114	103	66	64	81	7	28	0	0

＋：抽煙

－：不抽煙

- 將65歲以上之資料去掉得表3。則抽煙者之死亡率便高於不抽煙者了，亦通過統計檢定。

表3 不滿65歲婦女抽煙與死亡數據

	抽煙	不抽煙	總數
死亡	97	65	162
存活	436	474	910
總數	533	539	1,072
死亡率	18.20%	12.06%	

◆數據會說話，但若所取得原始數據之品質便不佳，或對數據的處理過程有太大瑕疵，甚至對數據的解讀有誤，所說出的話自然不會太正確。

統計功能有其侷限，二因素間是否有因果關係，並非統計分析可以證實。因此不可輕率地對統計上看起來似乎相關的兩個因素，驟下結論說其間有因果關係。一般須有對照組，以多方確認。

迴歸分析

- 迴歸一詞，為19世紀英國統計學家高爾頓(Francis Galton，1822-1911)首先引進，以描述諸如父、子身高二變數間的關係。以x_i表第i個父親的身高，Y_i表其兒子的身高，$i=1,2,\cdots,n$。高爾頓想建立

$$Y=Q(x)+\varepsilon。$$

即以父親的身高x，來預測兒子的身高Y。

- 父親身高不視為隨機變數，所以小寫。理想狀況是$Y=Q(x)$，但不可能，會有誤差ε，ε為一隨機變數。若ε不太大，則以$Q(x)$來估計Y，便不太離譜。

- 可能的$Q(x)$：

$$Q(x)=\alpha+\beta x,$$

α, β為常數，此稱為簡單線性迴歸。

$$Y_i=\alpha+\beta x_i+\varepsilon_i,\ i=1,2,\cdots,n,$$

可假設$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$相互獨立，且$E(\varepsilon_i)=0$，$\mathrm{Var}(\varepsilon_i)=\sigma^2,\ i=1,2,\cdots,n$。

- α, β如何估計？

最小平方法

使 $\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2$ 最小，

即誤差平方和最小。解出

$$\hat{\beta}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

$$\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}。$$

以 $\hat{\alpha}+\hat{\beta}x_i$ 估計 Y_i。

◆亦有

$$Q(x)=\beta_0+\beta_1 x+\beta_2 x^2,$$

$$Q(x)=\beta_0+\beta_1 x+\cdots+\beta_r x^r,$$

$$\vdots$$

6. 紙上談兵

- 收集數據前，要先有實驗設計。
- 經模擬產生數據。
- 林覺民，在 與妻訣別書中，寫不盡對愛妻的不捨。 最後說：

紙短情長，所未盡者尚有幾萬千，
汝可以模擬得之。

結語

◆十二年國教規劃五個領域的素養：

語文、數學、科學、數位、教養/美感。

- 數學知識的四大領域：

1. 變化與關係
2. 空間與形狀
3. 數量
4. 不確定性與數據

統計究竟能做什麼？

- 1985年11月14日，Gary Taylor在牛津大學的圖書館找到一首可能是莎士比亞(W. Shakespeare, 1564-1616)的詩。就稱泰勒詩，有 429字。
- 不少專家認為泰勒詩，用字遣詞與韻味風格，都異於莎士比亞其他作品。

- 統計學者也介入這場紛爭。
- 1986年1 月24日Science 雜誌，刊登

莎士比亞的新詩—向統計學禮讚

(Shakespeare's new poem : an ode to statistics)

介紹Efron及Thisted，以統計的方法判定泰勒詩，是否為莎士比亞所作。

◆莎士比亞總作品中

共有884,647個字,

其中

有31,534個相異字，

有14,376個相異字只出現1次，

有4,343 個相異字只出現2次。

⋮

◆在總作品中，罕用字的使用非常普遍。

◆Efron 與 Thisted 估計

莎士比亞尚認識11,460±150個字。

- 發表：

  Did Shakespeare write a newly-discovered poem?

- 若泰勒詩為莎士比亞所作，估計有6.97±2.64個新字，實際有9個。
- 估計曾出現1次的字有4.21±2.05個，實際為7。
- 估計曾出現2次的字有3.33±1.83個，實際為5。

- 一直到曾出現100 次的字，估計與實際值，吻合程度皆相當驚人。
- 用統計術語來說：

  不能拒絕此詩為莎士比亞所做之假設。

- Efron及Thisted 也對另3位與莎士比亞同時代的詩人，各取1首詩，及另取4首莎士比亞的詩，與這首泰勒詩做比較。
- 經過3種統計檢定，發現對前3首，罕用字出現次數，與莎士比亞所的頻率皆不吻合。
- 雖然挑選的4首莎士比亞的詩偶而有不吻合處，總的來說是可接受的。

◆統計證明泰勒詩為莎士比亞所作？

非也！

但統計讓文學家接受泰勒詩為莎士比亞所作！

緣起

- 1940年代。生物學家 Williams向費雪提出一個似乎不可能回答的問題：

  Williams曾前往馬來西亞採集蝴蝶，他把自己共見過幾種(species)蝴蝶，以及每種各見過的次數，都告訴費雪：

  想知道馬來西亞的蝴蝶，他沒見過的有多少？

◆此問題似乎毫無頭緒。

◆只要假設

蝴蝶依每一種隻數的比例，隨機地被捕捉，

統計學家便有辦法估計。

- Efron 與 Thisted把費雪所用的方法拿來分析莎士比亞的作品。想回答

  若發現一新作品，如何經由統計分析其中用字出現頻率，以決定此作品是否為莎士比亞所作？

- 1976年發表：

  Estimating the number of unseen species: How many words did Shakespeare know?

- 在生態學中估計某生物未見到的種數。
- 在該文中，尚未見到的種，卻是莎士比亞知道但不曾用過的字。
- 蝴蝶的種 ⇔ 相異字

統計是倚天劍？

統計的判定，
是否一出手，
就令人臣服？

- 克萊門斯(W. R. Clemens, 1962-)，活躍美國職棒大聯盟二十餘年，共獲得七座賽揚獎(Cy Young Award)，為史上最多。
- 前訓練員麥克納米，指控他在職業生涯的後期，服用類固醇及生長激素。
- 美國國會進行調查。

- 克萊門斯請專家為他整理出一份45頁的報告，包含38個圖表。
- 4 位美國賓州大學的統計及經濟教授，認為這份報告，對讓人相信克萊門斯的無辜，並無說服力。

- 在克萊門斯將他自己，與在1993年46歲時退休的名投手萊恩進行對比。
- 兩人都是在40多歲時達到顛峰。
- 與另外兩位同時代的投手詹森，及席林相比，結果也類似。
- 數據會說話，還克萊門斯清白？

- 4位專家指出，報告中僅將克萊門斯與那些在職業生涯第二階段獲得成功的投手相比，而非與所有跟克萊門斯一樣，在菜鳥階段就揚名立萬的投手相比。
- 在統計學裡，稱此為選擇偏差。
- 比較的對象精心挑選，克萊門斯的數據，沒有顯得不正常就不稀奇了。

- 4 位專家分析自1968年起，大聯盟31位優異的投手之生涯投球數據。
- 相對於這較大的比較群，克萊門斯的數據，就顯得異常。
- 一般投手，都是初期逐漸成長，30歲左右達到高峰，約自35歲起，逐漸走下坡。
- 但克萊門斯的表現在快30歲時下降，而在35歲至40歲間又成長，極不尋常。

(2008年2月10日 紐約時報)

- 有三種謊言：謊言，可惡的謊言，及統計。
- 統計為何被當做謊言？

- 統計的結論要有價值，其中每一程序，從設計，取樣到分析，都要儘量客觀。
- 統計學家會犯錯，因所有保證都是機率式的，並附帶一定的犯錯機率。決策若不願犯錯，後果不見得就好：

法官不錯放一個之後果？

法官不錯關一個之後果？

- 機率理論告訴我們，如果統計分析是遵循該有的程序，則長期下來，犯錯次數的比例，差不多就是所設定的犯錯機率，乃可容忍。
- 分析過程中，若有偏差，則即使工程再浩大，得到的結論，不但無法取信真正的專家，被當成謊言不說，有時還給自己製造出極不利的後果。

- 統計，正如我們的思維，客觀至上，否則便是自欺欺人。
- 反之我們的思維若是統計式的，便是極客觀的。

謝謝各位！