統計思維

黃文璋

國立高雄大學
應用數學系，統計學研究所

97年 5月30日 於台北大學

前言

馬克吐溫(1907)：

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？

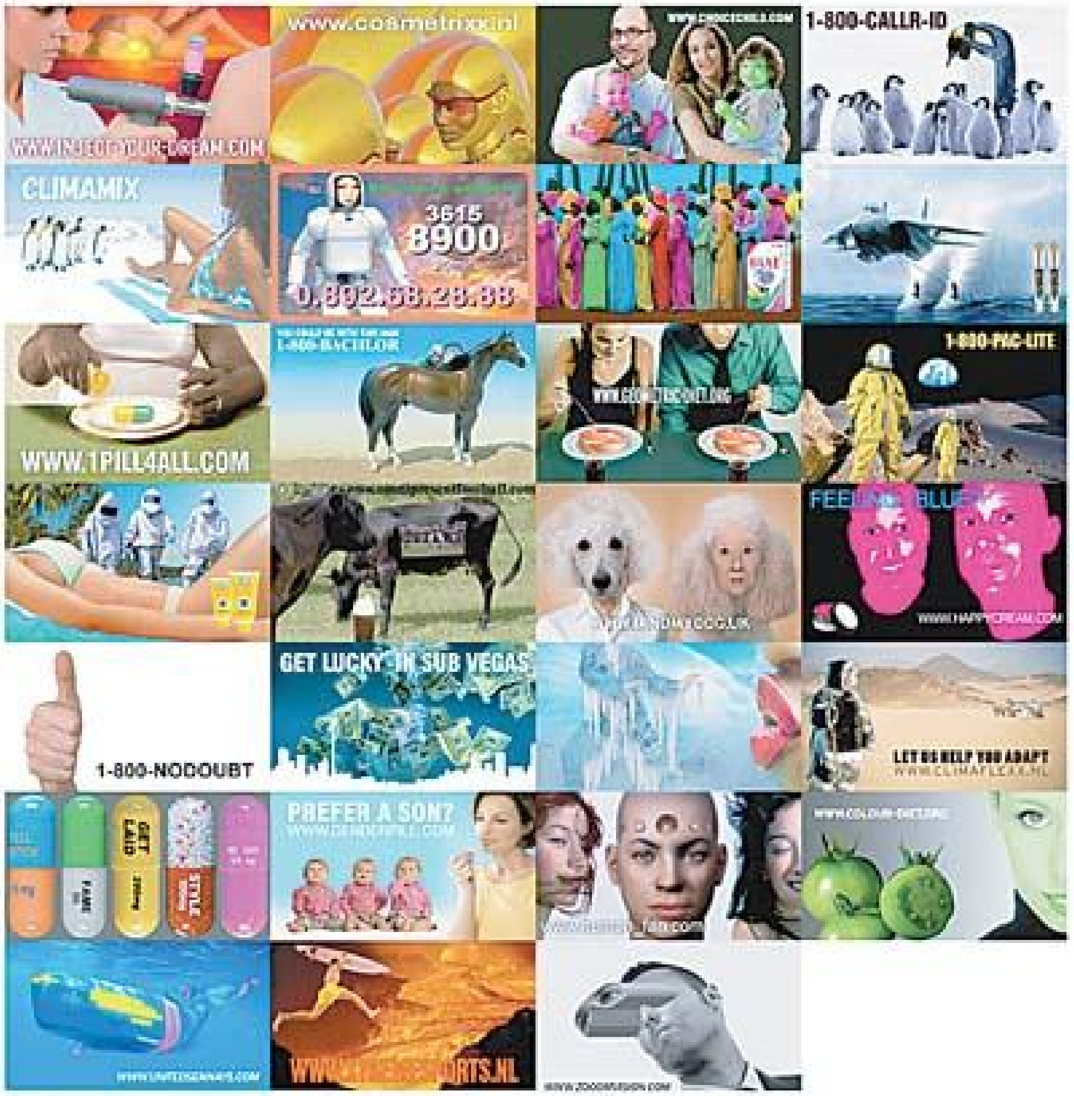
  較有邏輯、計算較精準，…

- 近年來，統計學似乎愈來愈重要。
- 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令人掌握。
- 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
FAME 50
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。

- 一統計方法，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？
- 統計學家的本事要能發揮，就得善用資訊。
- 有些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？

有人敲門，是男是女？

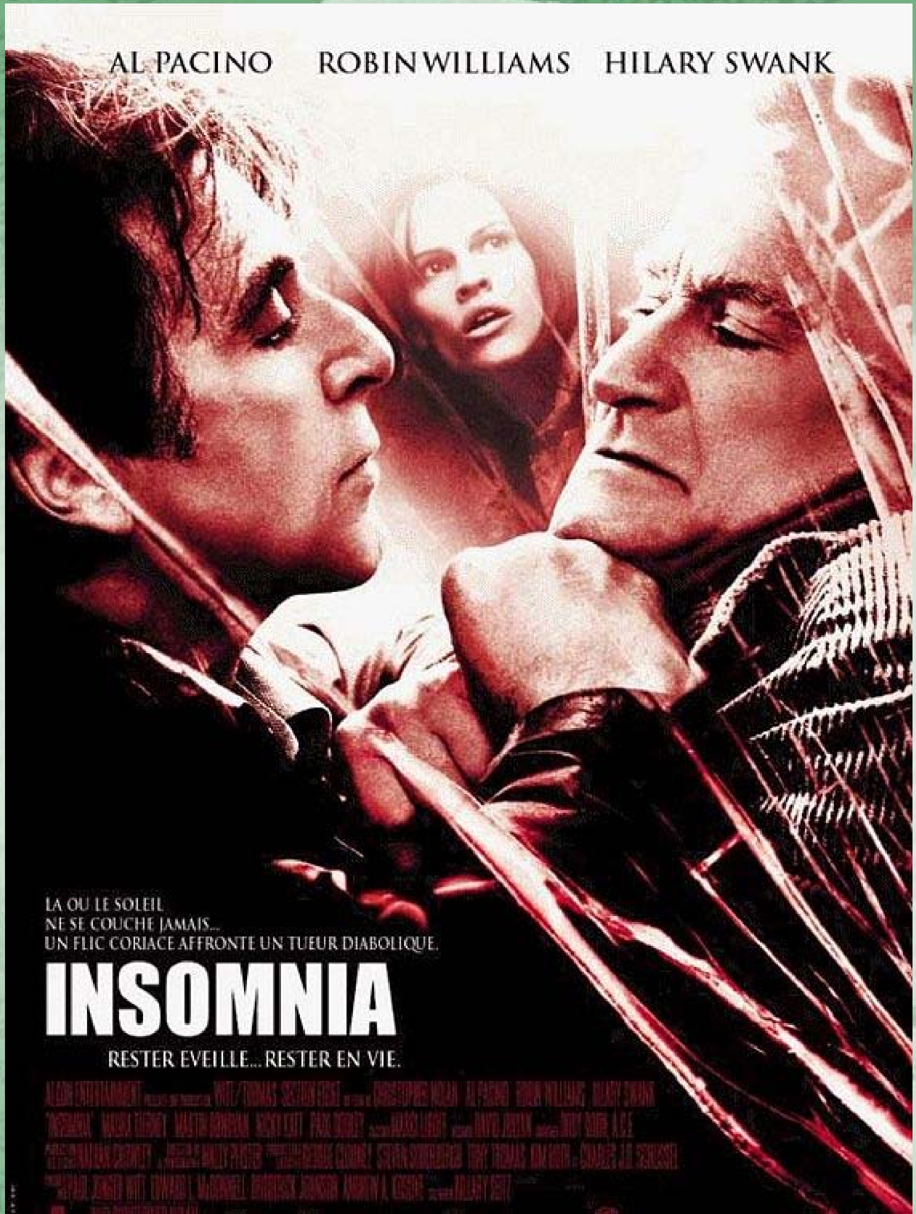
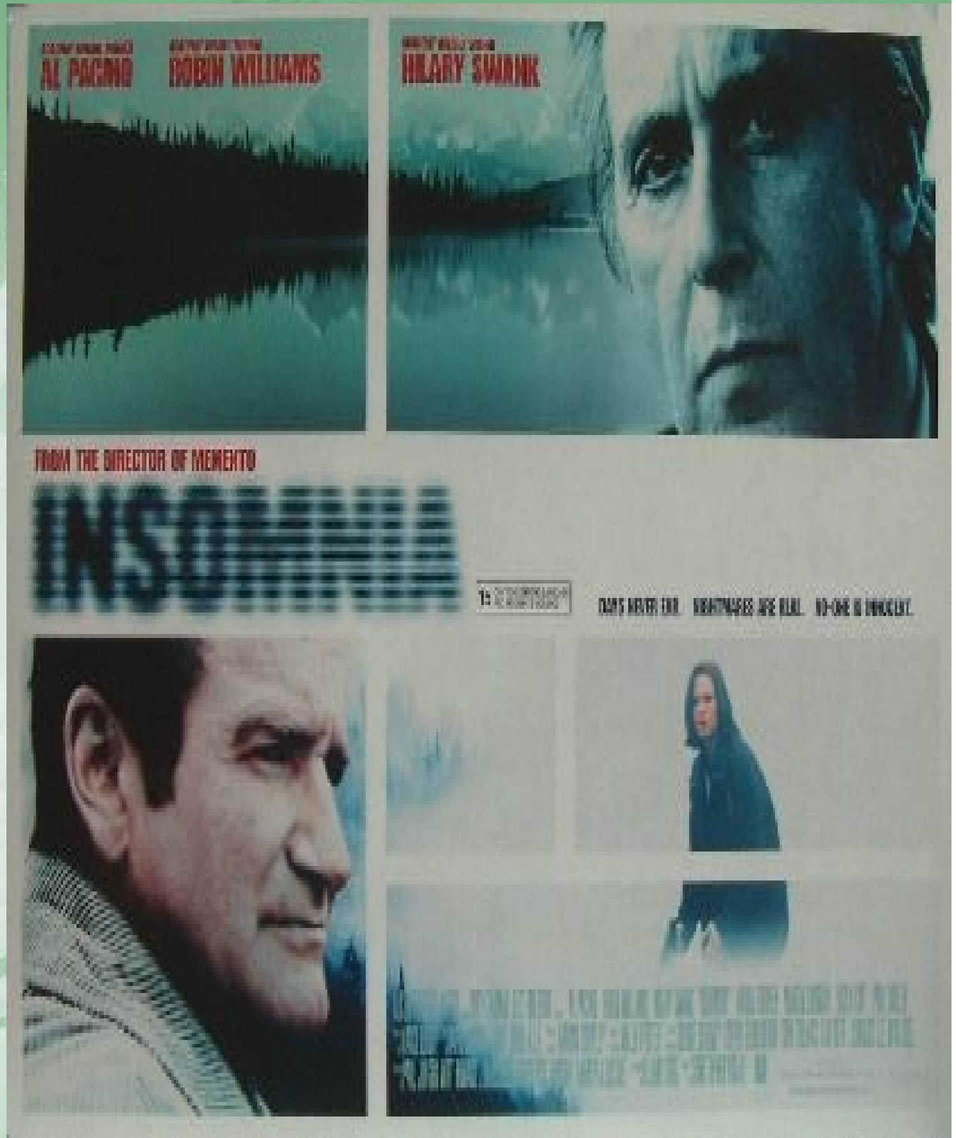
- 機率值會變，是機率的一特性。
- 視新的資訊產生，對一事件機率之判斷，也隨之而變。

- 草進了牛肚 ⇒ 牛奶。
- 資料進了統計學家手中 ⇒ 資訊 ⇒ 決策。

- 有時想儘量減少資訊的透露。
- 號碼鎖、提款卡，或上網的密碼等，要愈難被猜中。

如何防止密碼被猜中？

- 買樂透彩時想要猜中號碼，樂透彩公司則希望儘量不易被猜中。
- 在電影針鋒相對裡，飾演警探的艾爾帕西諾，受到嫌犯羅賓威廉斯的要脅。
- 帕西諾將槍藏在空調的排氣口，但被威廉斯找到拿走。


AL PACINO ROBIN WILLIAMS HILARY SWANK
LA OU LE SOLEIL
NE SE COUCHE JAMAIS...
UN FLIC CORIACE AFFRONTE UN TUEUR DIABOLIQUE.
INSOMNIA
RESTER EVEILLE... RESTER EN VIE.

AL PACINO
ROBIN WILLIAMS
HILARY SWANK
FROM THE DIRECTOR OF MEMENTO
INSOMNIA

- 如何藏東西?

  將適合藏的地點編號，隨機挑一個。

- 樂透彩中獎號碼隨機產生。

- 對於密碼，隨機挑選最難被破解。

- 付出代價:

  密碼自己不易記得。

2. 了解變異

由常數至隨機

- 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，人們將少了努力的動機。
- 光有隨機性，只靠運氣，將令人失去積極認真的企圖心。

- 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 1 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？

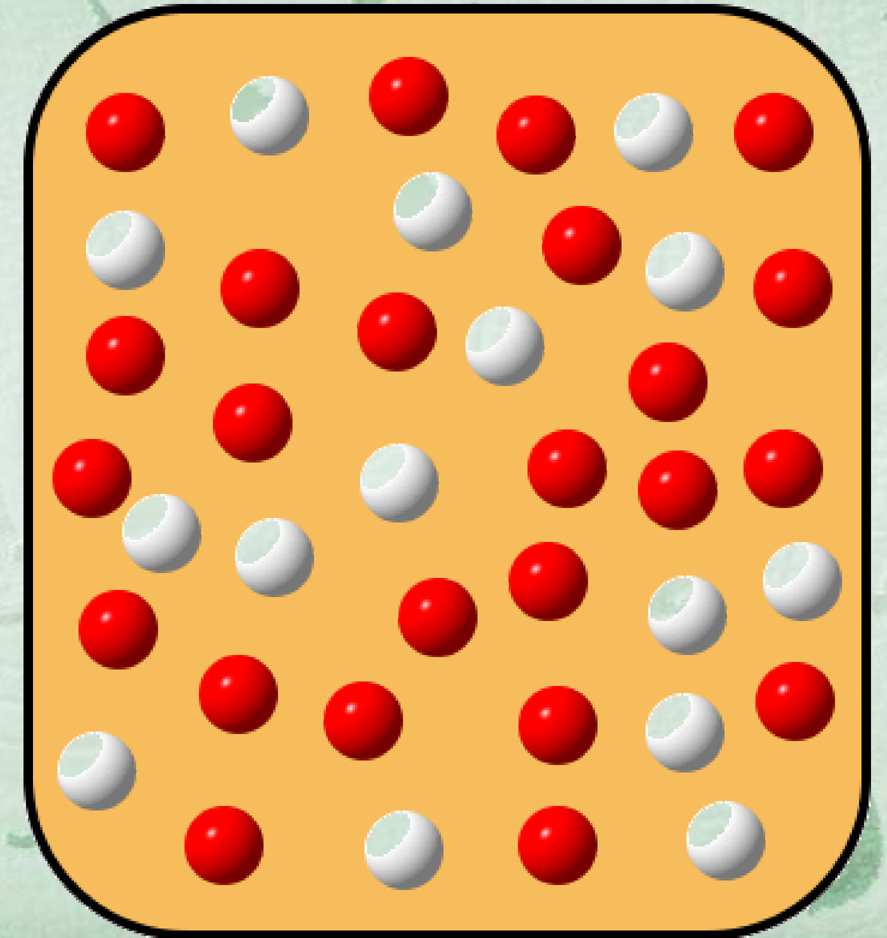
- 數學中常求近似值。
- 必須要能給出誤差大小。
- 中央極限定理指出，量測所得之誤差有常態分佈。
- 誤差大小是隨機的。
- 但誤差之散佈情形，則能描述。


GS1527893K8
DEUTSCHE BUNDESBANK
10
ZEHN DEUTSCHE MARK
10
GS1527893K8


DEUTSCHE BUNDE
10

- 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。

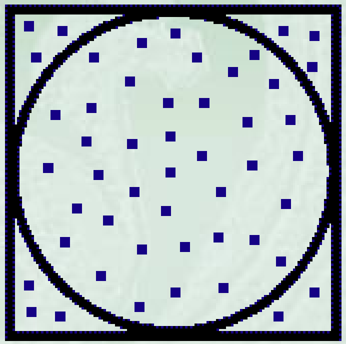
- 袋中白球所佔比例？



- 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。

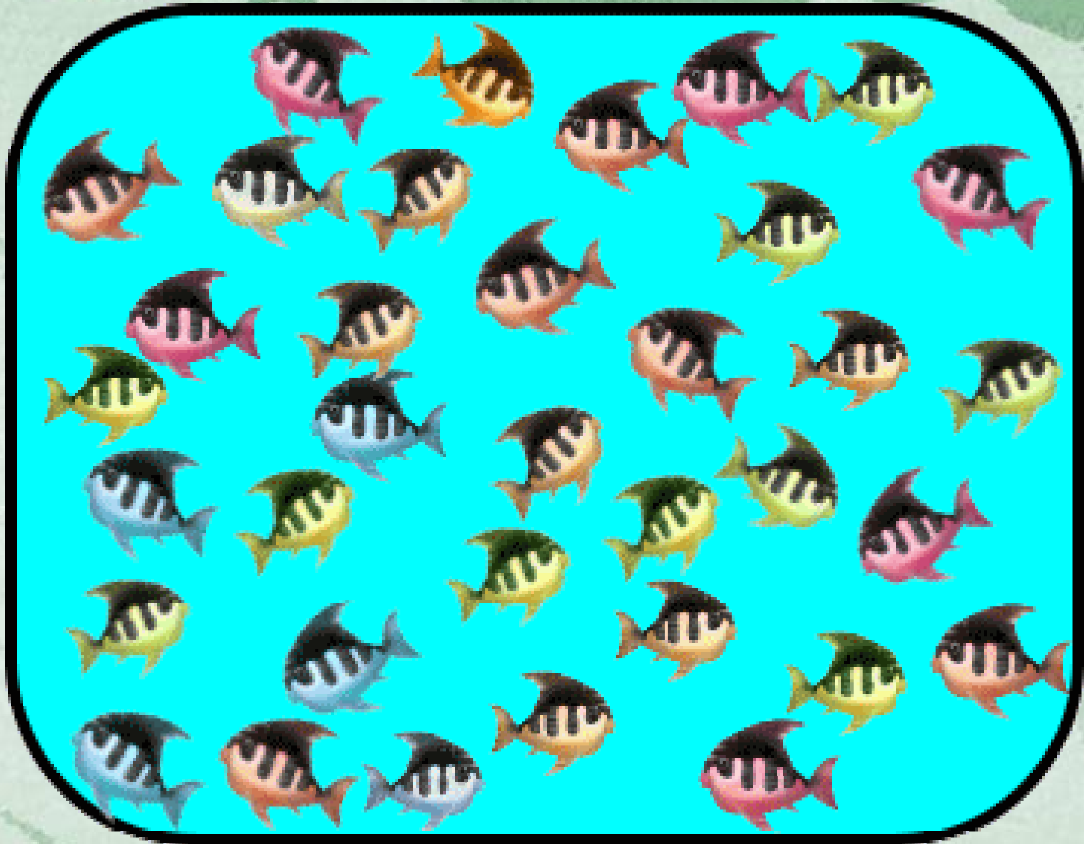
$2r$



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

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

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

- 拉普拉斯曾說：

大部分生活中最重要的疑問，都只是機率的問題。

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%。

- 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。
- 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。
- 以公理化的方式引進機率。

- 事件在發生前才有機率可言。
- 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，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。

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(此機率並不大)，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

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

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

- 以布朗運動的結果來估計：

$$P(\max_{1\le n\le 10{,}000} S_n \le a) \doteq P(\max_{0\le s\le 1} X(s) \le \frac{a}{100})$$

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

其中 Z 有 $N(0,1)$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：

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

當*x*=0.993，機率約為

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

$\Rightarrow$

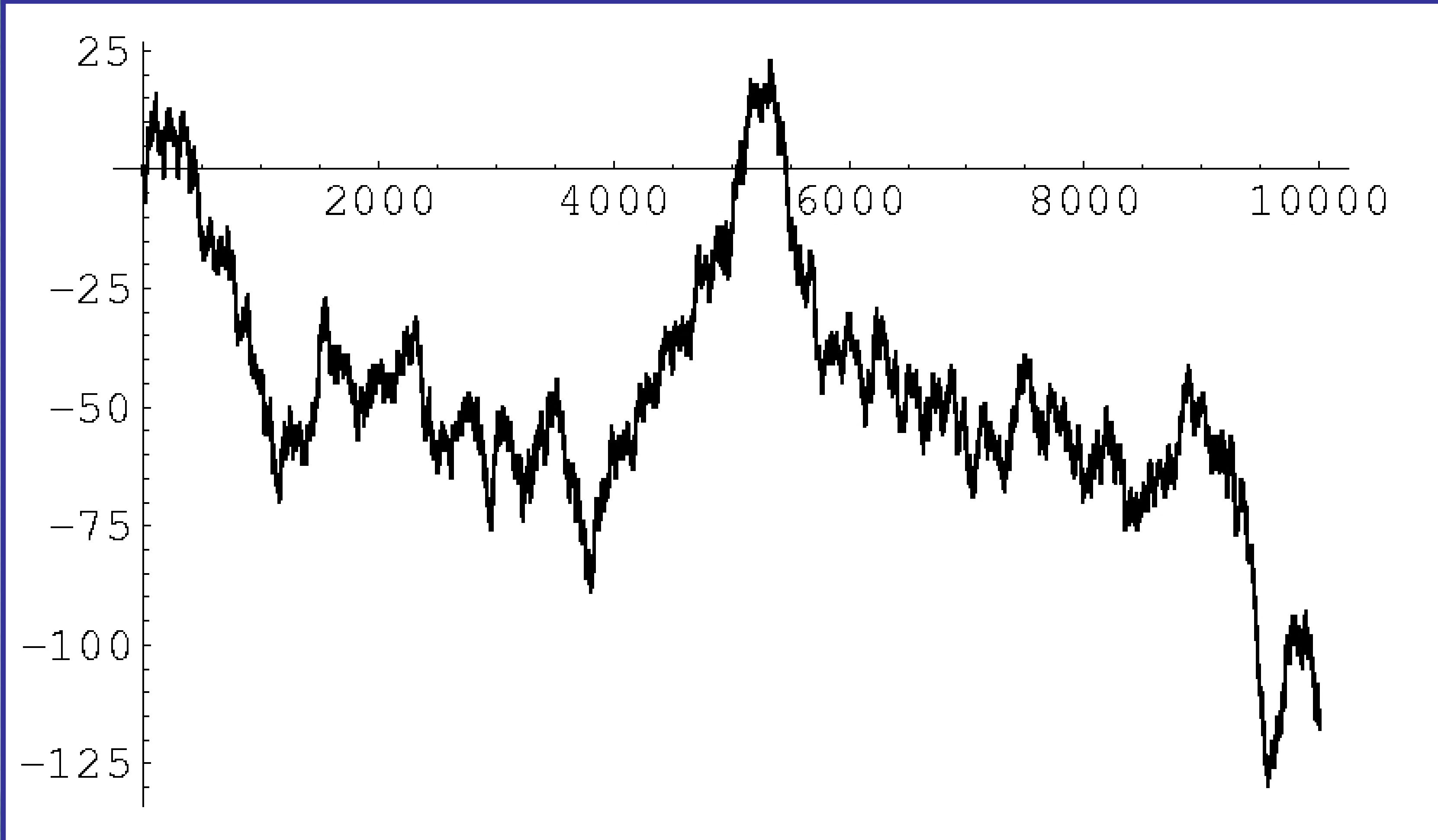
$$P(\text{有一面領先次數} \geq 9930) \doteq 0.0533 \times 2 = 0.1066\ 。$$

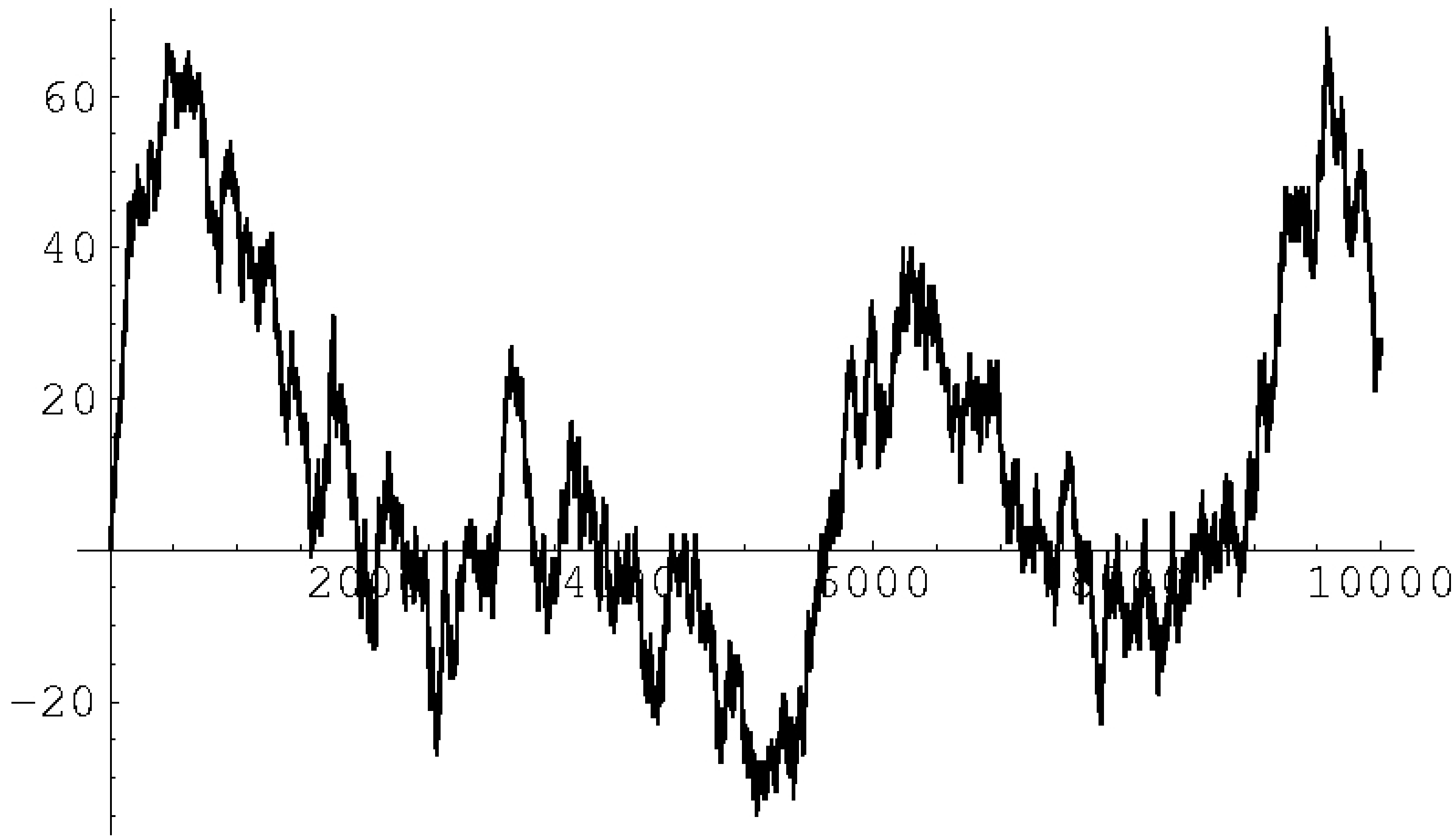
- 底下為五個模擬圖，橫軸為n, 縱軸為S_n，其中$S_i > 0$表正面領先，$S_i < 0$表反面領先。

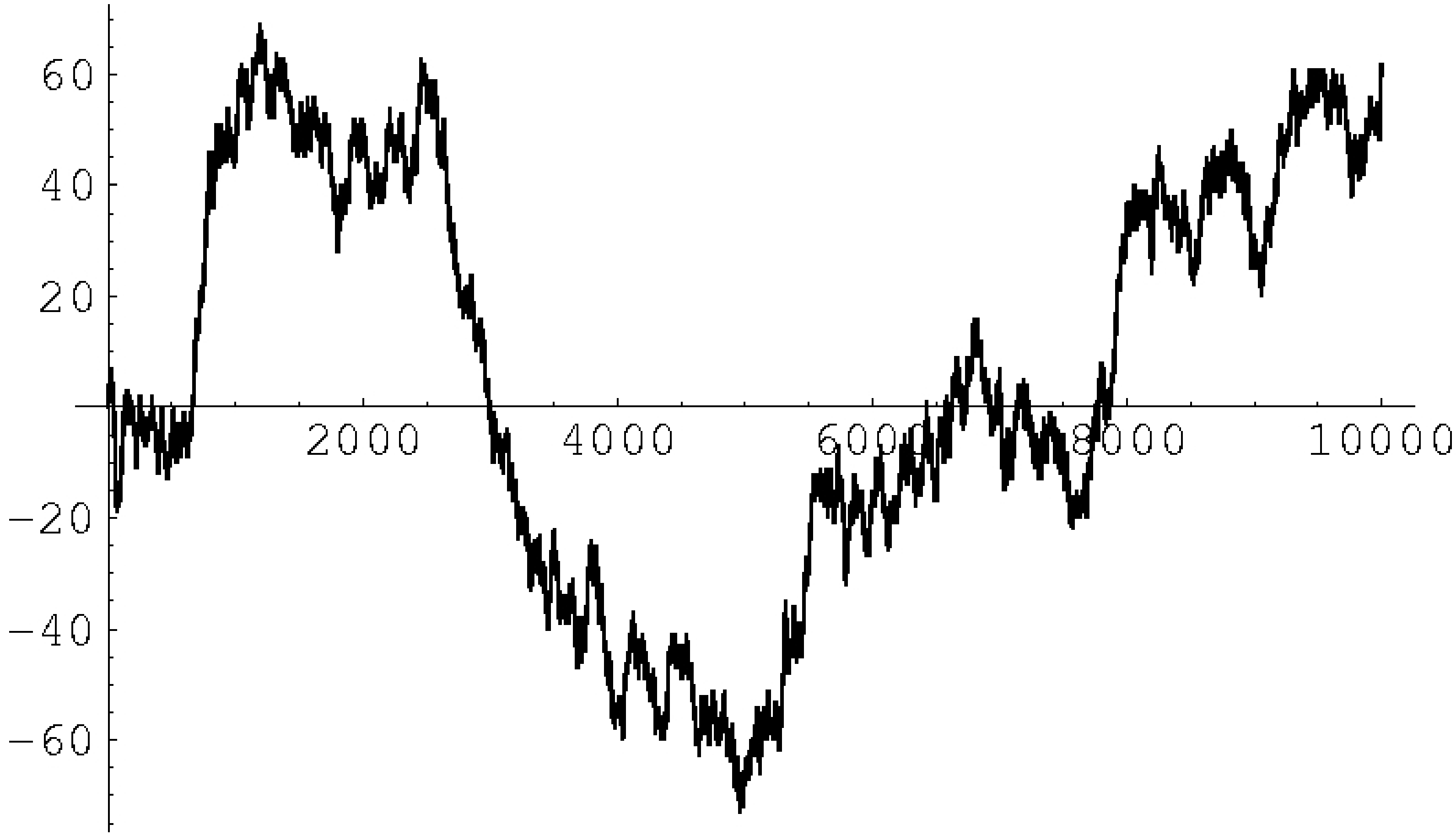
$\{S_n, n \geq 1\}$ 構成一隨機漫步(random walk)。

注意

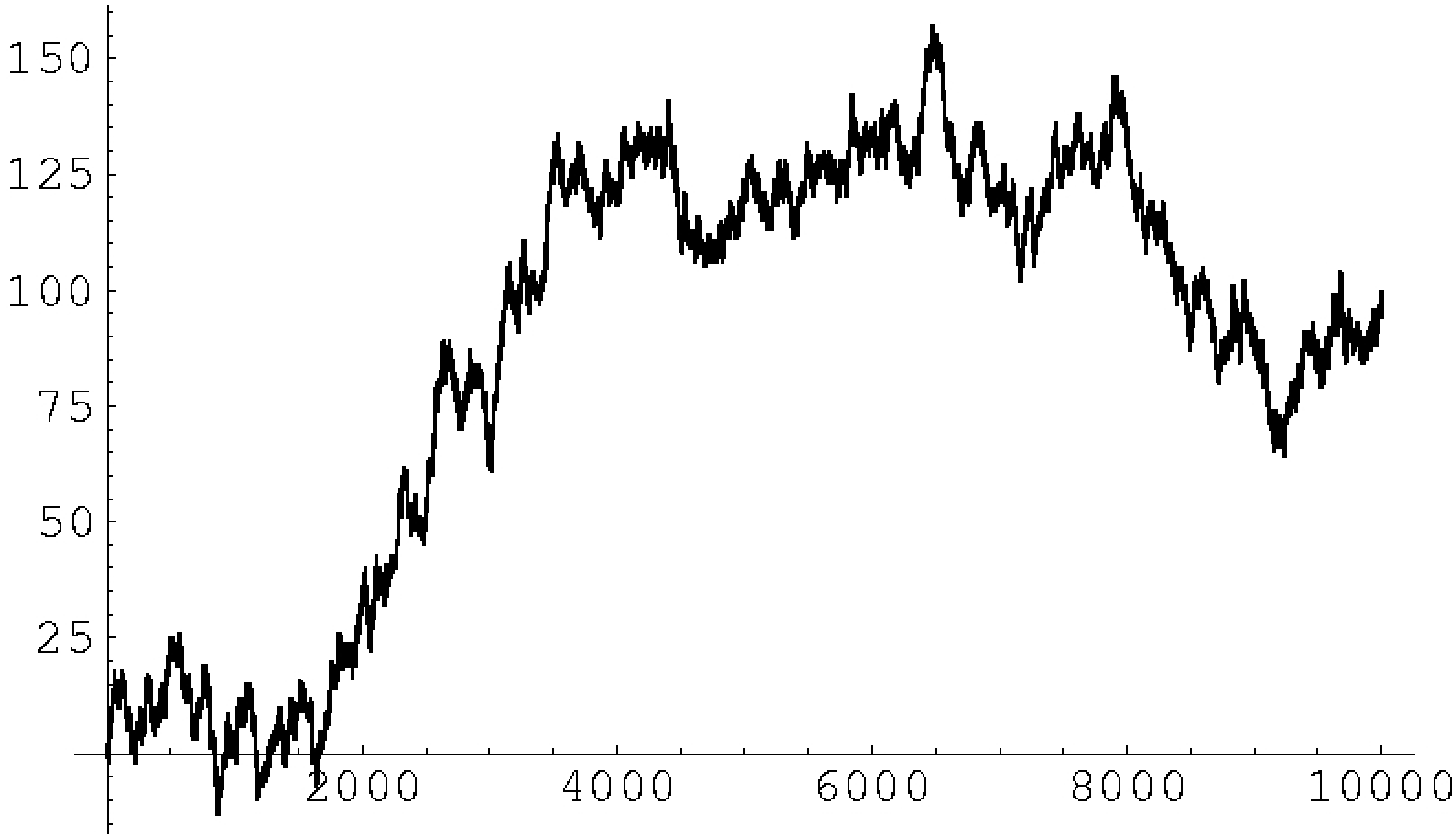
$$E(S_n) = 0,\ \forall n \geq 1$$ 。

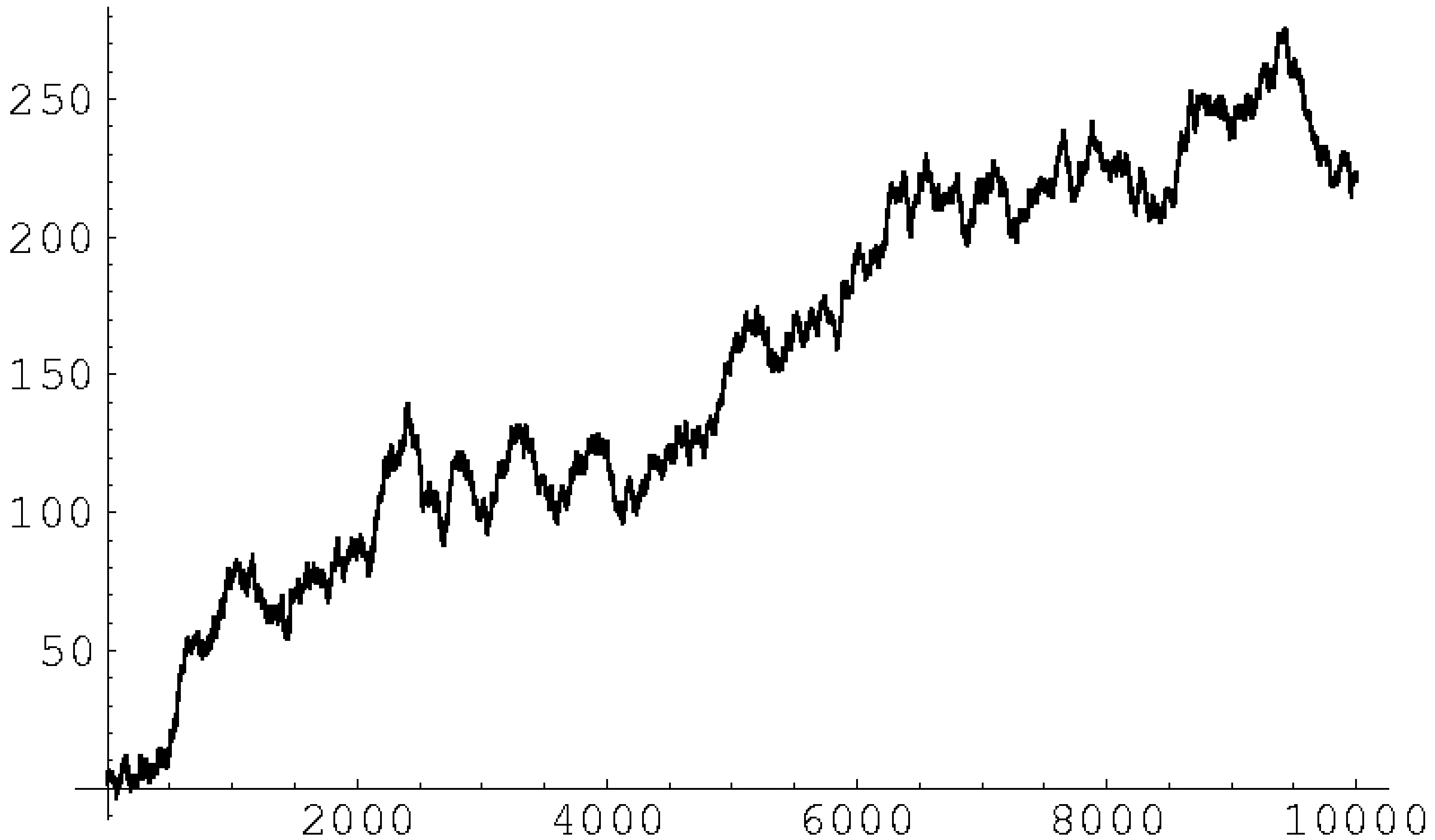

25
-25
-50
-75
-100
-125
2000
4000
6000
8000
10000


60
40
20
-20
2000
4000
6000
8000
10000

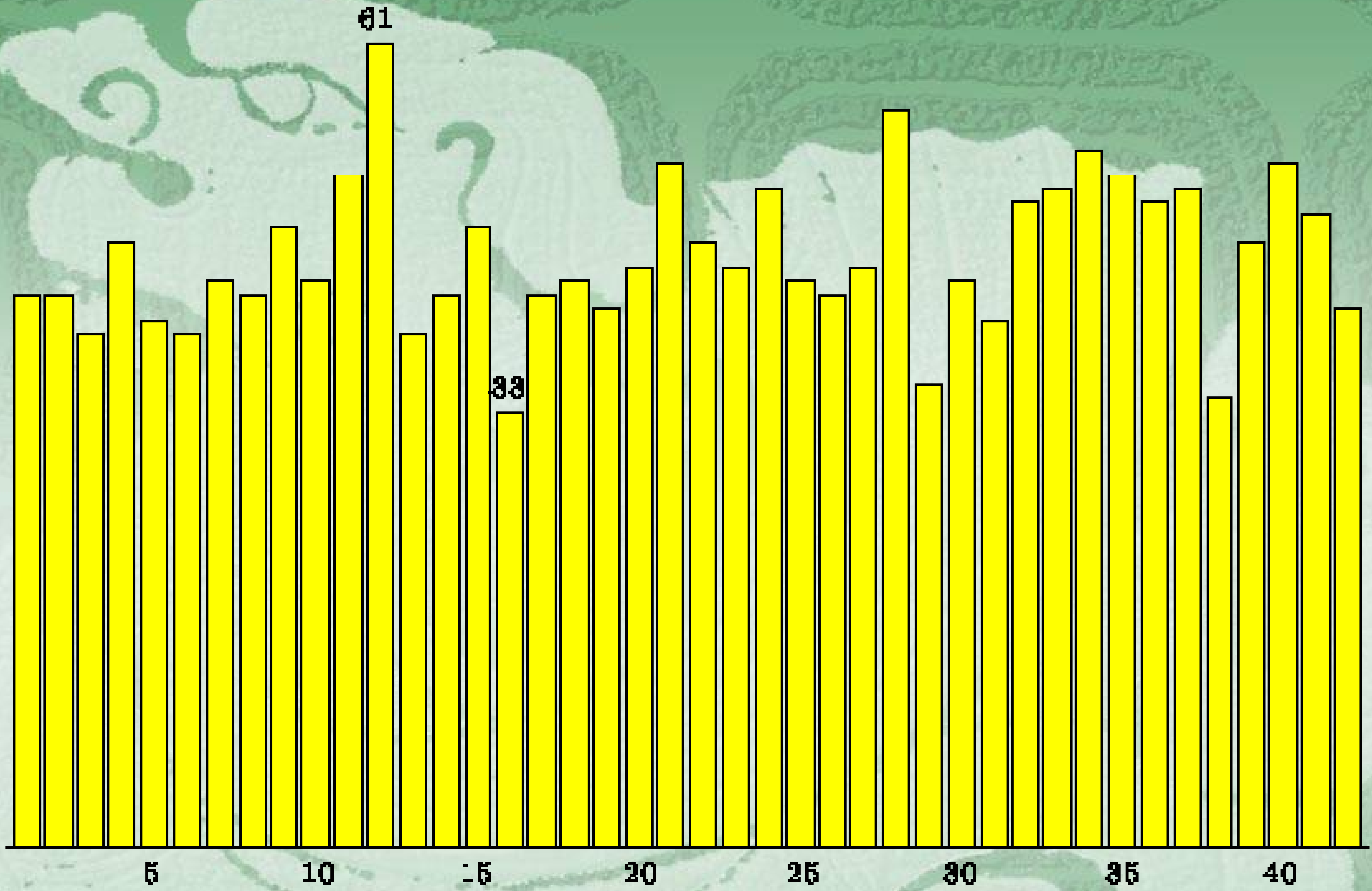


60
40
20
-20
-40
-60
2000
4000
6000
8000
10000


150
125
100
75
50
25
2000
4000
6000
8000
10000


250
200
150
100
50
2000
4000
6000
8000
10000

均匀分佈下的不均匀



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。

- 美國有人兩次中樂透彩頭獎。
- 20杯奶茶說對先放奶或放茶之機率

$$\frac{1}{2^{20}} = \frac{1}{1,048,576}\text{。}$$

對全台灣每人測試，約有23人可全說對。

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次時，贏的機率 $\approx 4 \times 10^{-10}$。

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。

- NBA職業籃球賽，以一球季的勝率，決定那些隊參加季後賽。
- 以相對頻率來估計的想法 $\Rightarrow$ 動差法。

- 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

$\Rightarrow$ 貝氏法。

信賴區間在信賴什麼？

- 政治人物慣說我對XXX是百分之百信賴。
- 對一參數 p ，以單一的值來估計 p ，信心不太夠。
- 以一隨機區間來估計 p ，此區間表 p 可能落在的範圍。
- 信心水準指 p 會落在此區間的機率。
- 就估計的觀點，一區間較一點更可信賴。
- 給出一估計之明確的信心水準，反而使此區間更可信賴。

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，對於死因：

    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。

- 考試有20道選擇題，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？

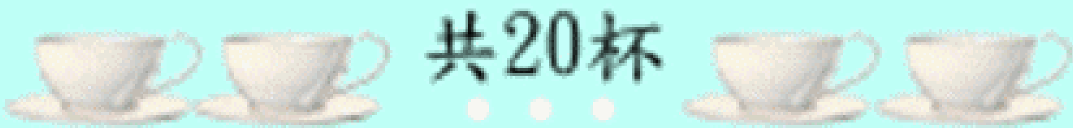
- 假設檢定，是對隨機現象，做決策之一重要依據。
- 現代統計學之創始者費雪 (R.A. Fisher)提出下述故事：

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：

把茶加進牛奶裡，和把牛奶加進茶裡，喝起來完全不同。

摩登神農氏

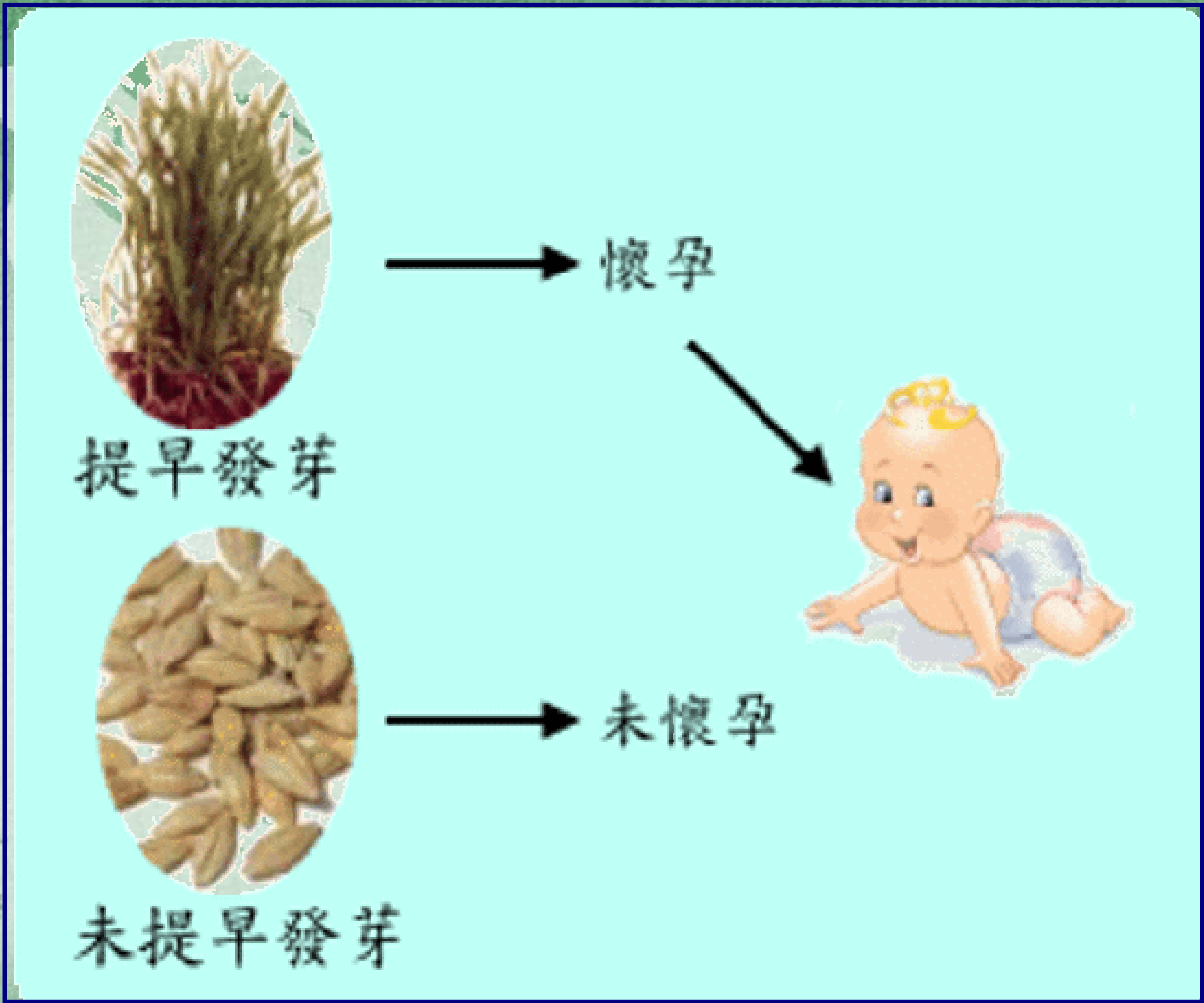
先牛奶？先茶？



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。


懷孕
提早發芽
未懷孕
未提早發芽

- 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?

在此分辨能力指她每杯講對的機率大於隨機猜的機率 $\frac{1}{2}$。

- 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，一件事往往不知真偽。

- 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還先放茶的能力，

  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，再求在無分辨能力的假設下，講對次數會這麼多的機率有多大？
- 如果機率小於α(即這是較不尋常)，則拒絕原假設，否則接受原假設。

- 波蘭人奈曼(J. Neyman)及英國人皮爾生 (E.S. Pearson)，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。

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。
- 巴紐案官員沒有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。
- 巴紐案官員有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沒人要看。

例.投擲一銅板若干次，正面數出現比率為 50.114%，僅比50%略多，是否不足以推翻此銅板為公正？

解.結論為何與投擲數 n 有關。

設 $X_1,\cdots,X_n$ 為 i.i.d. $Ber(p)$ 分佈之r.v.'s。

$E(X)=p$，$\mathrm{Var}(X)=p(1-p)$。 欲檢定

H_0：p=0.5 ，H_a：p>0.5。

拒絕域為 $\{\overline{X}_n>c\}$。由中央極限定理，

$$\frac{\sqrt{n}(\overline{X}_n-p)}{\sqrt{p(1-p)}}\approx N(0,1)。$$

現觀測到 $\overline{X}_n = 0.50114$。

(1) n = 13,000,000。

$$p\text{-值} = P(\overline{X_n} \geq 0.50114 \mid H_0\text{爲真})$$

$$\approx P\left(Z \geq \frac{\sqrt{13{,}000{,}000}(0.50114 - 0.5)}{\sqrt{0.5 \cdot 0.5}}\right)$$

$$\approx P(Z \geq 8.22)$$

$$\approx 1.03 \cdot 10^{-16}\text{。}$$

此值微乎其微，故拒絕H_0，即認為銅板並非公正。

(2) $n = 1{,}000{,}000$。

$$p\text{–值} = P\left(Z \ \geq \ \frac{\sqrt{1{,}000{,}000}\,(0.50114 - 0.5)}{\sqrt{0.5 \cdot 0.5}}\right)$$

$$= P(Z \ \geq \ 2.28)$$

$$\approx \ 0.0113 \ 。$$

若 $\alpha > 0.0113$，則拒絕 H_0，否則接受 H_0。

(3) $n = 100{,}000$。

$$p-\text{值} = P\left(Z \geq \frac{\sqrt{100{,}000}(0.50114 - 0.5)}{\sqrt{0.5 \cdot 0.5}}\right)$$

$$= P(Z \geq 0.721)$$

$$\approx 0.2355$$ 。

大部分的情況，皆接受H_0。

- 50.114%與50%差異是否夠大，與投擲數*n*有關！

  ***n*愈大，此差異就可能夠大，為顯著，**

  ***n*較小時，此差異可能不夠大，為不顯著。**

- 顯著與否，依發生機率大小，而非觀測值大小。

- 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，成績是否作假，須調出原始成績才知。

6. 紙上談兵

- 實驗設計。
- 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

- 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
(Shakeapeare's new poem : an ode to statistics)，

介紹Efron及Thisted，以統計的方法鑑定泰勒詩, 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。

統計是倚天劍？

統計的判定，
是否一出手，
就令人臣服？

- 克萊門斯，活躍美國職棒大聯盟二十餘年，共獲得七座賽揚獎。

- 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。

- 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！

例.你交了一新朋友。問他“有幾個小孩”，他說“有兩個”。問他“有女孩嗎?”他答“有”。

問：他亦有一男孩之機率為何？

解.兩個小孩的性別：

男男、男女、女男、女女。

已知有一女孩：

男女、女男、女女。

會有男孩：

男女，女男。

$\Rightarrow$ 另一小孩為男孩之機率為2/3，另一小孩為女孩之機率為1/3。

- 以模擬來說明。

- 產生1,000,000組(i, j)，i , j=0,1。
  1表女孩，0表男孩。
  無女孩：250,820組，
  有女孩：749,180組，
  有二女孩：249,131組。

$$\frac{\text{有二女孩}}{\text{有女孩}} = \frac{249,131}{749,180} \doteq 0.332538241 \doteq \frac{1}{3}\text{。}$$

- 回答“沒有” ⇒ 兩個小孩皆為男孩。
- 若問“有男孩嗎?” 回答“有” ⇒ 另一個小孩為女孩之機率亦為2/3。
- 小孩是男是女的可能性皆為1/2：

  事前機率。
- 獲得一些資訊(知道其中有一女孩)後，另一小孩是男或是女的機率改變了：

  事後機率。
- 不少家庭都是兩個小孩。若知道其中一個是男孩，則猜另一個是女孩;若知其中一個為女孩，便猜另一個是男孩：

  猜中的比率很容易超過一半。

問：不論問有男孩或有女孩，只要答“有”，
另一小孩為異性之機率皆為2/3，合理嗎？

問：如果問題改為「老大是女孩嗎？」
結果有何不同？

- 福爾摩斯根據一些蛛絲馬跡來推測，一張清晰的圖片，是會顯示一些資訊。

最近新出現的網路芳鄰電腦病毒，主要是入侵密碼可以輕易被破解的電腦伺服主機。因此，防毒公司呼籲網路使用者，取個特殊的密碼，並經常更換，才是避免被電腦病毒入侵破壞的最好辦法。…。

防毒軟體公司對此表示，破解密碼已成為駭客快速入侵企業網路的模式。而且，駭客通常是使用字典攻擊法進行攻擊。這種字典攻擊法，就是以特定程式將所有字典上的單字逐一嘗試，破解密碼。而這隻網路芳鄰電腦病毒，就是採用字典攻擊法滲透上萬部電腦系統。

因此，防毒軟體公司呼籲網路使用者，…，取個沒有邏輯可循的密碼，避免使用個人或親朋好友的生日或電話號碼，英文字或是純數字組合，…。(中廣新聞網92年3月29日)

- 簡單計算：

  1.英文字典有為數十萬左右的單字。若以英文單字為密碼，很可能採常見字，不超過兩萬個。

  2.若以26個英文字母加上0,1,…,9等10個數字混和編碼，共36個。則長度為6的字母數字串，組合數共有

$$36^6=2,176,782,336(\text{種})。$$

$$\frac{2,176,782,336}{20,000} \doteq 108,839(\text{倍})。$$

- 若字典攻擊法平均一天可破解，隨機編碼，平均要十萬多天

$$\frac{100,000}{365} \doteq 298(年)$$

才能破解。

- 若採長度為7的字母數字串：

$$\frac{36^7}{20,000} \doteq 3,918,208(天),$$

$$\frac{3,918,208}{365} \doteq 10,734.8(年)。$$

Is Your Password Predictable?

你的密碼成了公開的祕密？

Before you read . . .

日常生活中，多少都會用得到密碼，你是用什麼方法來挑選密碼？用生日？寵物名字？還是紀念日？小心別讓你的密碼成了公開的秘密！據研究顯示，密碼不僅反映出個性，還與潛意識有關，因此，若你選的密碼是明顯與個人事物相關的數字，便很容易遭人破解。現在你是不是考慮要更改密碼了呢？

CNN Anchor

[When] you pick a computer password, using the first word that comes to mind might not be the best option. Psychologists say that's because passwords typically reflect a user's personality, making some relatively easy to predict, as Andrew Brown found out.

Andrew Brown, CNN Correspondent

Just think of all the things on your desk. They may mean nothing at all. They may also help someone **crack** your computer password and then, **masquerading**[1] as you, send e-mails, access files, even **plunder**[2] your online bank account.

CNN 主播

挑選電腦密碼時，使用第一個想到的字可能不是最好的選擇。心理學家表示，因為密碼往往反映出使用者的個性，所以，相對地比較容易讓人猜出來。安德魯・布朗有了以下的發現。

CNN 特派員 安德魯・布朗

想想你桌上所有的東西，這些東西也許根本無關緊要，但它們卻有可能幫別人破解你的電腦密碼，然後冒充你寄發電子郵件、存取檔案，甚至盜用你在線上銀行的戶頭。

1. masquerade [ˌmæskəˈred] *v.* 冒充；偽裝
2. plunder [ˈplʌndɚ] *v.* 侵吞；掠奪

- 金融機構建議線上客戶使用隨機挑選的字詞，並要經常更換密碼以增強安全性。而且當選擇密碼時，不要只鍵入password這個字，畢竟，要破解這種密碼並不是太難。(CNN互動英文2002年5月號)

- 對一隨機現象，

  提出猜想，

  將猜想表為統計假設，

  ⇒ 接受或拒絕統計假設。

  此過程稱為假設檢定，

  得到的推論，稱為統計推論。

- 統計假設與一般數學中的假設不同。
  ◇在數學裡：

$$\text{假設}x > y\text{。}$$

  並未涉及任何隨機的量，

  此非統計假設。

  ◇令 μ 表北銀大樂透1號出現的機率，則

$$\mu > \frac{6}{49},$$

  為一統計假設。
- 數學中的假設不須去檢驗是否為真。
- 統計中的假設並不是要證明是否為真，而是要判定該接受或拒絕。

- 取一組隨機樣本，並利用此組樣本，當做是否接受某一假設之證據。
- 如果證據與假設所陳述的不合，或者說吻合的機率很低，便拒絕該假設，否則便接受該假設。
- 數據會說話，但不論方法多好，對一統計假設所做的推論，是可能有錯的。
- 數學證明不能有錯，統計推論允許犯錯！換一組樣本，結論可能便相反。
- 在無法避免犯錯下，只能以較好的方法減小犯錯的機率。

- 北銀大樂透頭獎號碼中，1號出現的機率是否大於$\frac{6}{49}$？若傾向相信答案是肯定的，則取

$$H_0：\mu = \frac{6}{49}，H_a：\mu > \frac{6}{49}。$$

- 在奶茶問題，取

$$H_0：\mu = \frac{1}{2}，H_a：\mu > \frac{1}{2}。$$

- 虛無假設是被保護的，除非證據夠強，否則不輕易推翻。
- 對於現況不輕易推翻，會使人們在做決策時更謹慎：

朝令夕改非假設檢定的精神！

- 假設檢定裡，多大的錯誤機率可忍受？
- 有兩種錯誤的機率：
  ◇虛無假設為真卻拒絕(第一型錯誤)，
  ◇虛無假設不真卻接受(第二型錯誤)。

	H_0為真	H_0不真
接受H_0	正確	第二型錯誤
拒絕H_0	第一型錯誤	正確

- 當樣本數固定，兩型錯誤的機率， 有一減小另一必增大。

- 第一型錯誤較嚴重，通常的作法：
  ◇先控制第一型錯誤的機率不要超過某一 α 值，
  ◇然後使第二型錯誤的機率愈小愈好。

- 在樂透彩號碼的隨機性檢定裡，先假設各號碼出現的機率相同，再看於此假設下會出現這麼異常的結果之機率是否夠小，以判定該不該推翻機率相同的假設。
- 通過檢定並不表號碼真的是隨機產生，只是說尚無不合；但若不通過，大約便不相信號碼為隨機產生。

- 統計假設的架構，與刑事訴訟法中的無罪推定原則(第154條)類似：

  被告未經審判證明有罪確定前,

  推定其為無罪。

- 先相信虛無假設，然後看實驗出現的數據合不合理。
- 以機率大小來判定合理性。
- 若會出現這種數據的機率很小，便認為不合理，由是拒絕虛無假設。

- 統計裡的推論與數學中的證明不同。

- 假設投擲一銅板100次，皆得到正面。在合理的 α 下(只要 $\alpha > \frac{1}{2^{100}}$)，會拒絕

$$H_0 \text{：此為公正銅板。}$$

但公正的銅板還是有可能(只是機率很小) 出現如此極端之結果。

問：統計檢定中，α 太小(如 $\alpha = \frac{1}{2^{100}}$)是否恰當？

例. 某大公司有一萬名員工，年終摸彩，總經理秘書中頭獎。作一檢定：

H_0：沒有作弊，

H_a：有作弊。

在H_0之下，隨機地抽獎，秘書中頭獎機率為0.0001，只要$\alpha>0.0001$，會拒絕H_0。

合理嗎？全公司每一員工都不該中頭獎？

- 一觀測值若使H_0被拒絕，此觀測結果便稱為顯著。
- 顯著與否，並非依觀測值的大小，而是依發生機率之大小。

- 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。
- 只要假設蝴蝶是

  ★依照每一種之隻數的比例，

  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