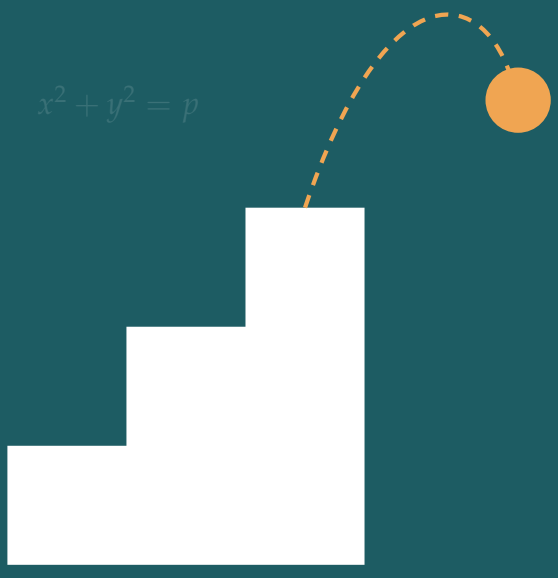
STEP 備試 · 大學數學

Beyond STEP

第 13 章

現實世界中的統計與隨機性



KEITH McNULTY

Beyond STEP

以歷屆 STEP 試題為門
走進大學數學

第 13 章

現實世界中的統計與隨機性

Keith McNulty

Beyond STEP：以歷屆 STEP 試題為門，走進大學數學

第 13 章：現實世界中的統計與隨機性

本冊是 *Beyond STEP* 全書 13 章中的第 13 章。全書、各章的單行本、每一節的可列印工作紙，以及每一道 STEP 試題的詳細解答，均可在 `beyondstep.keithmcnulty.org` 免費閱讀及下載。

2026 年 11 月首次於網上發表。

本 PDF 內凡屬作者本人創作的內容，包括正文、解答、插圖、練習及答案，均按照 Creative Commons CC0 1.0 Universal 授權條款貢獻予公有領域：你可以為任何目的複製、改編及分享，毋須徵求許可。詳見 `creativecommons.org/publicdomain/zero/1.0`。

本書討論的 STEP 試題，版權屬 Cambridge University Press & Assessment 所有。本書並無轉載這些試題：每一題均附連結，通往官方來源的題目，即 STEP Support Programme 的試題資料庫或 OCR 的歷屆試卷。

如有錯漏，歡迎到 `github.com/keithmcnulty/step_book_mcnulty` 提交 issue 指正。

章

13

統計與現實世界中的隨機性

概率由模型推向數據，統計則由數據返回現實世界。

本章內容

13.1	解讀新聞標題	STEP I 1992
13.2	點算田鼠	STEP II 2000
13.3	德國坦克問題	STEP III 1998
13.4	平均數還是中位數？	STEP II 2000
13.5	問尷尬的問題	STEP II 2023
13.6	累積起來的誤差	STEP II 1993
13.7	製造相關性	STEP III 2010
A	練習答案	第 51 頁

從數據中學習

所有模型都是錯的，但有些是有用的。

— 統計學家 George Box

在概率（probability）中，我們知道骰子怎樣被動了手腳，然後問會發生甚麼事。在統計（statistics）中，我們看到已發生的事，然後問骰子必定是怎樣被動了手腳。概率由因推果，統計則由果溯因。對總體的每一個估計、每一次民意調查、每一項臨牀試驗、每一次科學測量，都是一個統計推斷（statistical inference）：根據有限而帶有雜訊的數據，對現實世界作出結論。

與數學的大部分分支相比，統計是一門年輕的學科。它的現代面貌是在二十世紀初由皮爾遜（Karl Pearson）、費雪（Ronald Fisher）、奈曼（Jerzy Neyman）等人建立的，背後的推動力往往是實際問題：釀造啤酒（t 檢驗就是在健力士（Guinness）啤酒廠發明的）、種植農作物、測試藥物。時至今日，隨着數據量爆炸式增長，統計及其近親機器學習（machine learning）已成為數學最重要的應用之一。

統計為何重要

有關藥物、公共衞生、經濟政策、氣候和司法的決定，全都建基於統計證據。誤讀這些證據會帶來實實在在的後果：病人被假陽性（false positive）結果嚇壞（Section 12.2），政策建基於一次偶然的波動(Section 13.1)，電腦系統中的微小誤差不斷累積(Section 13.6)。明白數據能告訴你甚麼、不能告訴你甚麼，以及任何結論有多不確定，是作為一個知情公民的一部分，對科學家、醫生、工程師和經濟學家而言更是專業上的必需。

你在大學會遇到甚麼

統計的入門課程涵蓋估計（estimation，即為未知量找出好的估計值，並說明其精確程度）、假設檢驗（hypothesis testing，即判斷數據是否與某個假設相符）、置信區間（confidence interval）、迴歸（regression，即把一個變量與其他變量聯繫起來），以及實驗和調查的設計。核心概念是*抽樣分佈*（sampling distribution，即估計值本身是一個隨機變量（random variable））、*偏差*（bias）與*方差*（variance），以及*似然*（likelihood）。其後的課程會加入貝葉斯推斷（Bayesian inference）、多元分析（multivariate analysis）、時間序列（time series）和機器學習的方法，而機器學習其實就是規模極其龐大的統計。

貫穿本章的問題

三大問題

1. **是真有其事，還是只是雜訊？**單憑偶然，我們應預期有多大的變動？一項變化要到甚麼程度才值得認真看待？
2. **我們應該怎樣估計？**有了數據，一個未知量的最佳估計是甚麼？我們又怎樣比較不同的方法？
3. **隨機性怎樣進入數據？**數據的收集、保護或處理方式，會對我們的結論造成甚麼影響？

本章先從以批判的眼光閱讀報章上的統計數字開始（Section 13.1），然後轉向估計：用捉放法（capture–recapture）點算田鼠的數目、由序號估計總數，以及比較平均值（mean）與中位數（median）。最後是三道關於隨機性怎樣進入數據的題目：保護私隱的隨機化調查、會累積或互相抵消的捨入誤差（rounding error），以及構造具有指定相關性的變量。

本章的概念

§	課題	概念
13.1	解讀統計數字	比率、自然變異、趨勢
13.2	捉放法	最大似然（maximum likelihood）
13.3	序號	偏差、方差與充分性（sufficiency）
13.4	估計中心位置	效率（efficiency）與穩健性（robustness）
13.5	隨機化回答（randomised response）	混合分佈（mixture）與總方差（total variance）
13.6	捨入誤差	系統誤差（systematic error）與隨機誤差（random error）
13.7	相關（correlation）	楚列斯基（Cholesky）構造法

Section 13.7 會用到協方差（covariance）和相關，這兩個課題已於 2019 年從 STEP 考試範圍中刪去；該節會解釋所需的知識。

怎樣閱讀本章

請記住，每一個估計值都是隨機變量：假如數據有所不同，估計值也會隨之不同。對於本章的每一種方法，都要問一問：重複很多次之後，它的分佈會是怎樣的？平均而言它是否正確？它的變動又有多大？這一個問題，正是統計思維的核心。

13.1 解讀新聞標題

題目

STEP I 1992，第 14 題

繼續閱讀之前，請先看題目（原題為英文）。題目載於 STEP Support Programme 的試題資料庫：

`step.maths.org/questions/92-s1-q14`

引子

統計學（statistics）是從數據中學習的科學，而它的第一課就是：數字本身不會說話。「安全六倍」、「升幅驚人」、「死傷日益慘重」：新聞標題每天都把數字變成論斷，而這些論斷往往是錯的；錯不在數字本身有假，而在於比較的對象不對，或是把尋常的隨機波動（random variation）誤當成趨勢。

這道題完全沒有公式，在 STEP 中甚為罕見。它要求你像統計學家那樣思考：要證明每個標題成立，我們需要知道甚麼？單憑偶然，我們又應預期有多大的波動？答案要用到泊松分佈（Poisson distribution）（Section 12.18），而當中的推理正是本章其餘部分的基礎。

動筆之前

想一想 假如死亡事件以穩定的比率獨立而隨機地發生，每年的總數應有多大的變動？若平均每年為 1080，每年與平均值的典型偏差是多少？

獨立罕見事件的計數近似服從泊松分佈，而平均值為 μ 的泊松計數，其方差（variance）為 μ，標準差（standard deviation）為 $\sqrt{\mu}$。取 $\mu = 1080$，標準差約為 33：每年總數落在約 1015 至 1145 之間，都不足為奇。

逐步解題

階段 1　「坐飛機比步行安全六倍」

> **想一想**　每年有 1080 名行人死亡，180 人死於空難。六倍，究竟是甚麼的六倍？

這個標題比較的是死亡總數，但安全與否關乎的是風險：每單位風險承擔量（exposure）的死亡人數。幾乎人人每天都步行，坐飛機的人卻相對較少，而且只是偶一為之。公平的比較應採用每公里行程、每小時行程或每次行程的死亡人數，而這幾種算法可以得出截然不同的答案（按每公里計，航空旅行極為安全；按每次行程計則沒有那麼出色，因為大多數空難死亡都發生在起飛和降落之時）。題目所給的數字根本不能「證明」兩者孰更安全。此外，「在道路意外中喪生的行人」是個狹窄的類別；整體的道路死亡人數要大得多。

階段 2　「我們的天空更安全：1990 年僅 125 人喪生」

> **想一想**　空難死亡像獨立的罕見事件嗎？你預期每年總數的變動有多大？

空難死亡是成羣出現的：一次墜機就可能奪去數百人的性命。因此每年的總數主要取決於墜機的次數，而這個次數很小，於是總數每年的變動極大。某一年有 125 人死亡，相對於平均的 180 人，與情況毫無改變完全吻合：只要多一次墜機，這個數字就會徹底改觀。一年的數據不能顯示趨勢。

階段 3　「道路死傷日益慘重：星期二 7 人喪生」

> **想一想**　平均每天有多少名行人喪生？若每天的死亡人數服從泊松分佈，一天之內出現 7 人或以上的情況有多常見？

平均每天有 $\frac{1080}{365} \approx 2.96$ 名行人死亡。若每天的計數服從這個平均值的泊松分佈，則 $P(\geq 7) \approx 0.031$：純粹出於偶然，每年就約有 11 天會有 7 名或以上的行人死亡。（而且「道路死傷」想必還包括司機和乘客，那麼一天之內有 7 人死於道路意外就更不值得大驚小怪了。）單單一天的數字，對趨勢毫無說明作用。這個標題從一個變動極大的數列中挑出一個極端值，是典型的謬誤。

階段 4　「升幅驚人：1990 年 1350 名行人喪生」

想一想 把 1350 與泊松標準差比較。這純屬偶然嗎？若不是，還有甚麼可以解釋它？這又足以稱為「升幅驚人」嗎？

這是全題最有分量的部分。在穩定的泊松模型之下，1350 比平均值 1080 高出 270，約為 8 個標準差：這極不可能是偶然所致。所以很可能確實有些事情改變了。但這個標題仍然言過其實，原因有以下幾方面：

- *被平均值掩蓋的趨勢*。1080 這個數字是 16 年的平均值。假如死亡人數由 1974 年的約 900 穩步上升至 1989 年的約 1260，平均值仍然是 1080，而 1990 年的 1350 只不過是延續既有的趨勢，並非突如其來的「驚人」升幅（圖 13.1.1）。我們需要逐年的數字。
- *風險承擔量*。人口增長、步行的人增多、交通更繁忙或人口老化，都會令死亡人數上升，而道路按每次行程計卻沒有變得更危險。按人均計算，或按每公里步行路程計算的比率，才是恰當的量度。
- *定義*。死亡個案的記錄或分類方式一旦改變（例如意外發生後多久之內死亡才計算在內），就可以造成數字的跳升。
- *過度離散*（*over-dispersion*）。現實中每年計數的變動比泊松模型所顯示的要大，因為各種條件（天氣、經濟活動）都在變。對「8 個標準差」這個說法應有所保留。

升幅很可能是真實的，但它是否驚人、成因為何，則需要更多數據才能判斷。

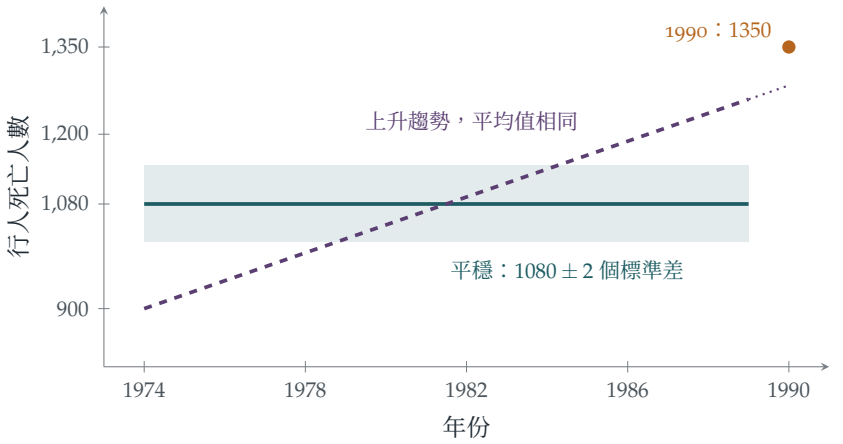


圖 13.1.1 兩段截然不同的歷史，16 年的平均值同樣是 1080。假如死亡人數一直平穩（藍綠色，連同泊松分佈典型波動的範圍帶），1990 年的 1350 就是急劇的跳升。假如死亡人數一直穩步上升（紫紅色），1350 就只是略高於趨勢。單憑平均值，無法分辨這兩種情況。

回顧

- **同類才可相比。**風險需要有分母：每人、每公里或每次行程的死亡人數。
- **認識自然波動。**泊松分佈的變異性，以及成羣出現的事件那大得多的變異性，讓我們知道甚麼時候的變化才值得留意。
- **平均值會掩蓋趨勢。**單憑一個概括的數字，無法描述一個量如何隨時間變化。

背後的數學

訊號與雜訊

每一個統計問題，歸根究底都是要把訊號（signal，即真實的效應或趨勢）與雜訊（noise，即隨機波動）分開。泊松模型為獨立事件計數中的雜訊提供了一把尺：標準差為 $\sqrt{\mu}$。罕見事件的相對波動很大：計數為 9 時，波動約為 ±3，即其大小的三分之一；而計數為 10 000 時，波動只有約 1%。正因如此，小數目（某一天的死亡人數、某一年的墜機次數）才會屢屢被人過度解讀，而小鎮上的癌症「羣組」最終也多半只是偶然。

均值回歸

還有另一個陷阱。道路安全攝影機往往安裝在剛剛經歷了異常惡劣一年的地點。翌年，意外減少，功勞就歸於攝影機。但異常惡劣的一年過後，本來就很可能迎來較為平常的一年：這就是均值回歸（regression to the mean），最早由高爾頓（Francis Galton）在 1880 年代提出。要把攝影機的效果與均值回歸分開，就必須與沒有安裝攝影機的類似地點作比較，這正是對照實驗（controlled experiment）背後的理念。

南丁格爾的統計學

南丁格爾（Florence Nightingale）以克里米亞戰爭中的護士身分聞名於世，同時也是統計學的先驅。她在 1858 年繪製的「玫瑰圖」顯示，死於可預防疾病的士兵遠多於死於創傷的士兵，並說服了政府改革軍隊的衞生狀況。她堅持使用比率（每年每千名士兵的死亡人數）而非原始總數，並堅持同類才可相比，這些正是本題所考核的要點。她是首位獲選為皇家統計學會（Royal Statistical Society）會士的女性。

思維技巧

問一句：「跟甚麼相比？」

在接受一個統計論斷之前，先問清楚：比較的對象是甚麼，分母是甚麼，以及單憑偶然會預期有多大的波動。

到了大學，統計學課程會把這些問題化為精確的方法：假設檢驗（hypothesis test）探究一項觀察結果是否與偶然相符；置信區間（confidence interval）說明一個量測量得有多精確；而實驗設計（design of experiments）則確保比較公平。

延伸練習

1. **泊松分佈的尾部。**設 X 服從平均值為 2.96 的泊松分佈，驗證 $\mathrm{P}(X \geq 7) \approx 0.031$。你預期每年有多少天會有 9 名或以上的行人死亡？

2. **成羣出現的死亡。**假設墜機事故按泊松過程（Poisson process）發生，比率為每年 1.2 宗，而每宗奪去 150 人的性命。求每年空難死亡人數的平均值和標準差。一年之內有 125 人或以下死亡，是否不尋常？有沒有可能某一年恰好有 125 人死亡？

3. **顯著的升幅？**假如 1974–1989 年的數字一直平穩地維持在 1080 ± 33，試（利用正態近似（normal approximation））求單一年份出現 1350 或以上的概率。

4. **擬一個更好的標題。**為四個標題逐一改寫，各擬一個數據真正支持的標題。

延伸閱讀

- **David Spiegelhalter, *The Art of Statistics***（Pelican）。講述如何從數據中學習，並附有大量取自新聞的例子。
- **Tim Harford, *How to Make the World Add Up***（Bridge Street）。思考傳媒中的數字的十條法則。
- **Darrell Huff, *How to Lie with Statistics***（Penguin）。1954 年出版的短篇經典，至今仍不過時。

13.2 點算田鼠

題目

STEP II 2000，第 13 題

繼續閱讀之前，請先看題目（原題為英文）。題目載於 STEP Support Programme 的試題資料庫：

`step.maths.org/questions/00-s2-q13`

引子

一個島上住了多少隻田鼠？逐隻去數是不可能的：牠們會躲藏，會走動，而且長得一模一樣。不過，你可以先捉一些，做上記號，放走，然後再捉一批。假如有 200 隻做了記號，而第二次捉到的 200 隻當中有 11 隻帶有記號，那麼整個種羣中大約有 $\frac{11}{200}$ 帶有記號，所以種羣的大小約為 $200 \times \frac{200}{11} \approx 3636$。

這種*標記重捕法*（capture–recapture）用於估計魚類、鳥類和鯨魚的數量，也用於估計吸毒者、無家者和戰爭傷亡者的人數，即任何無法完全點算的總體。這道題用*最大似然*（maximum likelihood）原理為這個直觀的估計打下穩固的基礎。最大似然是統計估計中最重要的思想：選取使觀察到的數據出現概率最大的那個未知量的值。

動筆之前

想一想 種羣共有 N 隻田鼠，其中 200 隻帶有記號。現隨機抽取 200 隻作為樣本。恰好有 11 隻帶有記號的概率是多少？

從 200 隻有記號的田鼠中選出 11 隻，再從 $N - 200$ 隻沒有記號的田鼠中選出 189 隻，而等可能的樣本共有 $\binom{N}{200}$ 個：

$$p_N = \frac{\binom{200}{11}\binom{N-200}{189}}{\binom{N}{200}}.$$

這就是*超幾何分佈*（hypergeometric distribution）。

逐步解題

階段 1　似然

把含有 N 的二項式係數（binomial coefficient）展開寫出，

$$\binom{N-200}{189} = \frac{(N-200)!}{189!\,(N-389)!}, \qquad \binom{N}{200} = \frac{N!}{200!\,(N-200)!},$$

所以

$$p_N = \binom{200}{11}\frac{200!}{189!}\cdot\frac{((N-200)!)^2}{N!\,(N-389)!} = k\frac{((N-200)!)^2}{N!\,(N-389)!}. \qquad \blacksquare$$

把數據固定，視之為未知量 N 的函數時，p_N 稱為*似然*（likelihood）。

階段 2　對整數求最大值

> **想一想**　N 是整數，所以用微積分並不方便。不如比較 p_{N+1} 與 p_N。兩者之比何時大於 1？

大部分階乘都會互相約去：

$$\frac{p_{N+1}}{p_N} = \frac{(N-199)^2}{(N+1)(N-388)}.$$

當 $(N-199)^2 > (N+1)(N-388)$ 時，這個比大於 1，即 $N^2 - 398N + 39601 > N^2 - 387N - 388$，亦即 $11N < 39989$，$N < 3635.4$。所以 p_N 一直遞增至 $N = 3636$，其後遞減：最大似然估計值為

$$\hat{N} = 3636. \qquad \blacksquare$$

這正是 $\frac{200\times 200}{11} = 3636.4$ 的整數部分，也就是那個直觀的估計。

階段 3　第三個樣本

抽取第二個樣本之後，其中 189 隻沒有記號的田鼠也被做上記號，所以現在共有 389 隻田鼠帶有記號。取 $N = 3636$，則有 $3636 - 389 = 3247$ 隻沒有記號。第三個含 200 隻田鼠的樣本中恰好有 j 隻帶有記號的概率為

$$\frac{\binom{389}{j}\binom{3247}{200-j}}{\binom{3636}{200}}.$$

這些概率（對 $j = 0, 1, \ldots, 200$）之和必定等於 1，由此即得該恆等式。 $\blacksquare$

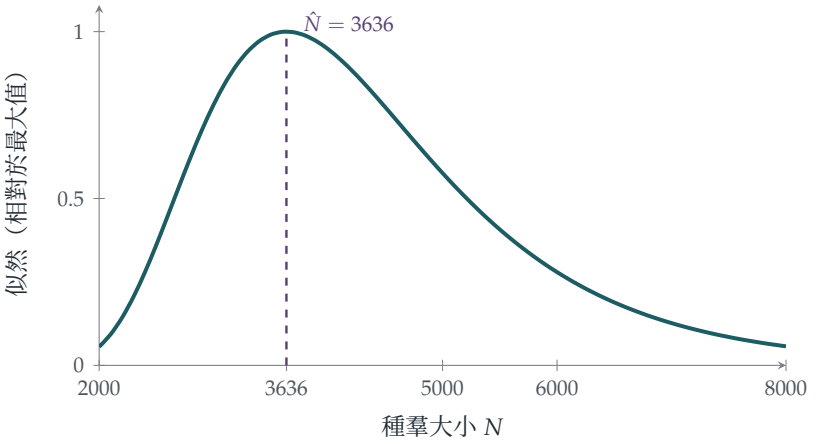


圖 13.2.1 在第二個樣本中發現 11 隻有記號田鼠的似然 p_N，表示為種羣大小 N 的函數，並按比例縮放使其最大值為 1。它在 $N = 3636$ 處達到峯值，但曲線寬闊而且偏斜：由大約 2500 至 6500 的種羣大小，都與數據相當吻合。

回顧

- **把數據的概率看成未知量的函數。**似然把慣常的觀點倒轉過來：數據固定不變，參數（parameter）才是變量。
- **用比值求離散最大值。**比較相鄰兩項，便找出了一個整數變量函數的最大值所在。
- **概率之和為 1。**把一個分佈對所有結果求和，便證明了一個組合恆等式（范德蒙恆等式，Vandermonde's identity）。

背後的數學

最大似然

最大似然原理由費雪（Ronald Fisher）在 1910 至 1920 年代發展出來，內容是：以使觀察到的數據出現概率最大的那個值，作為未知參數的估計。它是至今應用最廣的估計方法，由以曲線擬合數據，到訓練機器學習中的神經網絡，無一不用；神經網絡的「損失函數」（loss function）通常就是負對數似然。費雪證明了，對於大樣本，最大似然估計值近似服從正態分佈（normal distribution），而且其精確程度不遜於任何合理的估計量（estimator），前提是模型具備某些良好性質，而這些性質對大多數日常模型都成立。不過也有重要的例外。在下一節（Section 13.3）中，未知量是可能出現的最大值，即分佈的一個端點，在那裏最大似然估計值的表現大不相同：它的誤差並不近似服從正態分佈，而且誤差是以 $\frac{1}{n}$ 而非 $\frac{1}{\sqrt{n}}$ 的速度縮小。

這個估計有多精確？

圖 13.2.1 顯示似然曲線很寬闊：11 次重捕所提供的信息並不多。粗略計算一下：第二個樣本中有記號田鼠的數目近似服從泊松分佈（Poisson distribution），平均值為 $\frac{200\times 200}{N}$，所以其標準差（standard deviation）約為 $\sqrt{11}\approx 3.3$，相對誤差約為 30%，而這個誤差會傳遞到 N 的估計上。若要把 N 的估計誤差控制在 10% 以內，生物學家需要大約 100 次重捕。統計學家還會修正這個簡單估計的輕微偏差（bias），採用*查普曼估計量*（Chapman estimator）$\frac{(M+1)(C+1)}{R+1}-1$，其中 M 和 C 是兩次捕捉的數量，R 是重捕到的數目。

拉普拉斯點算法國人口

拉普拉斯（Laplace）在 1802 年用了同樣的想法，在沒有人口普查的情況下估計法國的人口。他知道全國一年內登記的出生人數，並且就一批抽樣的市鎮而言，同時知道其人口和出生人數。他假設人口與出生人數之比在各地都相同，從而估計法國人口約為 28 百萬，甚至給出了不確定程度的度量。後來這個方法在動物種羣研究中被重新發現：C. G. J. Petersen 在 1890 年代於丹麥的峽灣為鰈魚做記號，Frederick Lincoln 則在 1930 年代使用鳥環；這個方法至今仍常被稱為*林肯—彼得森*（Lincoln–Petersen）估計量。

思維技巧

選取使數據最有可能出現的參數

要估計一個未知量，先把你所觀察到的結果的概率寫成該未知量的函數，然後求其最大值。

在大學裏，最大似然是統計推斷（statistical inference）的骨幹，並有一套理論說明估計值有多精確（費雪信息量，Fisher information），以及如何比較互相競爭的模型（似然比檢驗，likelihood ratio test）。

延伸練習

1. **一般公式。**若有 M 隻動物做了記號，而第二個含 C 隻動物的樣本中有 R 隻帶有記號，證明 N 的最大似然估計值是 $\frac{MC}{R}$ 的整數部分。

2. **范德蒙恆等式。**考慮從 a 名女士和 b 名男士中選出 c 人組成委員會的方法數目，以此證明 $\sum_j \binom{a}{j}\binom{b}{c-j}=\binom{a+b}{c}$。

3 **擲硬幣。**一枚硬幣擲出正面的概率 p 未知，擲 n 次後觀察到 h 次正面。證明 p 的最大似然估計值是 $\frac{h}{n}$。

4 **沒有重捕。**如果第二個樣本中沒有任何帶記號的田鼠，最大似然估計值會變成怎樣？這說明了這個方法的甚麼特點？

延伸閱讀

- **David Spiegelhalter, *The Art of Statistics***（Pelican），談估計與不確定性。
- **George Seber, *The Estimation of Animal Abundance***（Blackburn）。標記重捕法的標準參考書。
- **John Rice, *Mathematical Statistics and Data Analysis***（Duxbury）。數理統計的入門課本，內容包括最大似然。

13.3 德國坦克問題

題目

STEP III 1998，第 14 題

繼續閱讀之前，請先看題目（原題為英文）。題目載於 STEP Support Programme 的試題資料庫：

`step.maths.org/questions/98-s3-q14`

引子

第二次世界大戰期間，盟軍情報部門想知道德國每月生產多少輛坦克。間諜的報告顯示每月約 1400 輛。統計學家則另闢蹊徑：繳獲的坦克在變速箱和其他零件上都刻有序號，而這些序號實際上就是從 $1, 2, \ldots, N$ 中抽出的一個隨機樣本（random sample）。單憑這些數字，他們估計產量約為每月 250 輛。按照一般引用的數字，戰後查閱的德國紀錄顯示為 245 輛。統計學家的估計遠比間諜準確。

本題比較兩種利用序號估計 N 的方法。兩者平均而言都是正確的，但其中一種運用數據的效率遠高於另一種；弄清楚是哪一種，就會引出估計理論（estimation theory）的核心思想：在平均而言正確的估計量（estimator）當中，應選取波動最小的一個。

動筆之前

想一想 設 M 是 $[0, N]$ 上 n 個獨立均勻分佈（uniform distribution）取值中的最大者，那麼 $\mathrm{P}(M \le m)$ 是多少？

$M \le m$ 當且僅當每個 $X_i \le m$，所以當 $0 \le m \le N$ 時，$\mathrm{P}(M \le m) = \left(\frac{m}{N}\right)^n$（可與 Section 12.25 比較）。求導後可知，M 的概率密度（density）為 $\frac{nm^{n-1}}{N^n}$。

逐步解題

階段 1　兩者皆無偏

對於 M：

$$\mathrm{E}(M) = \int_0^N m \cdot \frac{n m^{n-1}}{N^n}\,\mathrm{d}m = \frac{n}{n+1}N, \qquad \text{所以} \qquad \mathrm{E}(Z_1) = \frac{n+1}{n} \cdot \frac{n}{n+1}N = N.$$

對於平均值：每個 X_i 的平均值為 $\frac{N}{2}$，所以 $\mathrm{E}(Z_2) = \frac{2}{n} \cdot n \cdot \frac{N}{2} = N$。 ■

兩個估計量都是無偏（unbiased）的：平均而言，它們會給出正確的答案。因子 $\frac{n+1}{n}$ 用來修正這個事實：觀察到的最大序號總是比 N 略小一點。

階段 2　兩者的方差

> **想一想**　求 $\mathrm{E}(M^2)$，從而求 $\mathrm{Var}(M)$。至於 Z_2，單一個 $[0, N]$ 上的均勻變量的方差（variance）是多少？

$\mathrm{E}(M^2) = \int_0^N m^2 \cdot \frac{n m^{n-1}}{N^n}\,\mathrm{d}m = \frac{n}{n+2}N^2$，所以

$$\mathrm{Var}(M) = \frac{n}{n+2}N^2 - \frac{n^2}{(n+1)^2}N^2 = \frac{nN^2}{(n+1)^2(n+2)},$$
$$\mathrm{Var}(Z_1) = \frac{(n+1)^2}{n^2}\mathrm{Var}(M) = \frac{N^2}{n(n+2)}.$$

$[0, N]$ 上的均勻變量的方差為 $\frac{N^2}{12}$，而各個 X_i 互相獨立，所以

$$\mathrm{Var}(Z_2) = \frac{4}{n^2} \cdot n \cdot \frac{N^2}{12} = \frac{N^2}{3n}.$$ ■

階段 3　哪一個較好？

> **想一想**　比較 $\frac{1}{n(n+2)}$ 與 $\frac{1}{3n}$。當 n 增大時，兩者如何變化？

兩者之比為 $\frac{\mathrm{Var}(Z_1)}{\mathrm{Var}(Z_2)} = \frac{3}{n+2}$，當 $n \geq 2$ 時小於 1（當 $n = 1$ 時則相等，此時兩個估計量完全相同）。所以 Z_1 較好，而且 n 愈大就好得愈多：它的標準差（standard deviation）以 $\frac{1}{n}$ 的速度縮小，而 Z_2 的標準差只以 $\frac{1}{\sqrt{n}}$ 的速度縮小（圖 13.3.1）。若有 $n = 10$ 個序號，Z_1 的標準差約為 N 的 9%，而 Z_2 的約為 18%。

選用 Z_1 還有另一個理由：Z_2 有時會得出荒謬的結果。若序號為 10、20 和 300，則 $Z_2 = \frac{2}{3} \cdot 330 = 220$，比我們實際見過的一個序號還要小。而 $Z_1 = \frac{4}{3} \cdot 300 = 400$ 永遠不會小於最大的觀察值。 ■

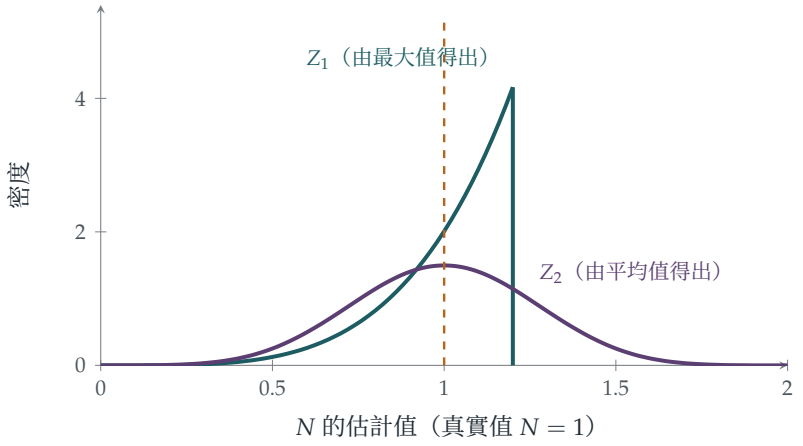


圖 13.3.1 兩個估計量的抽樣分佈（sampling distribution），其中序號數目 $n = 5$，真實值 $N = 1$。兩者的平均值都恰好是 1，但 Z_1（藍綠色）比 Z_2（紫紅色）更緊密地集中在真實值附近，而且它永遠不會低於觀察到的最大序號。

回顧

- **無偏性。**每個估計量都經過縮放，使其期望值（expected value）恰好等於 N。
- **比較方差。**在無偏估計量當中，方差較小的一個較為可靠。
- **選用信息量最大的統計量。**最大值包含了樣本中關於 N 的全部信息；平均值則浪費了其中一大部分。

背後的數學

充分性

為甚麼最大值好這麼多？一旦知道最大值 M，其餘各值不論 N 是多少，都只是在 $[0, M]$ 上均勻分佈：它們再沒有帶來關於 N 的任何信息。最大值是一個*充分統計量*（sufficient statistic）。C. R. Rao 與 David Blackwell 的一條定理（1945–47）指出，任何無偏估計量，只要換成它在給定充分統計量下的期望值，都可以得到改進，或至少不會變差；在這裏，這個做法正好把 Z_2 變成 Z_1。*Lehmann–Scheffé* 定理進一步證明，Z_1 是方差最小的無偏估計量。還要留意，Z_1 的方差以 $\frac{1}{n^2}$ 的速度下降，比大多數估計量的 $\frac{1}{n}$ 更快：這是因為 N 是分佈的端點，數據會擠向這個端點。

離散版本

真實的序號是整數 $1, 2, \dots, N$，並且是以不放回（without replacement）方式抽樣的。這時最佳的無偏估計是 $\hat{N} = m + \frac{m}{k} - 1$，其中 m 是 k 個觀察到的序號中的最大者：即最大值加上各觀察值之間的平均間距。盟軍統計學家運用這類估計，主要根據變速箱和車輪的序號，估計德國的產量約為每月 246 輛坦克；相比之下，其後在德國紀錄中找到的數字是 245 輛，而情報報告所顯示的是 1400 輛。

歷史上的序號

此後這個方法曾多次派上用場，例如分析員曾根據最先售出的幾件產品的序號，估計某款新消費產品的生產數量。製造商汲取了教訓：現在不少製造商都會把序號編碼，使人無法把它們當作簡單的順序編號來解讀。Richard Ruggles 與 Henry Brodie 於 1947 年在 *Journal of the American Statistical Association* 上記述了這項戰時工作。

思維技巧

以抽樣分佈評價估計量

估計量是一個隨機變量（*random variable*）。要問它平均而言是否正確（偏差），以及它的波動有多大（方差），然後選取誤差較小的一個。

在大學，這會引出有效估計（efficient estimation）的理論：Cramér–Rao 下界（Cramér–Rao bound）指出無偏估計量的方差最小可以小到甚麼程度，而充分性則說明如何達到這個下界。

延伸練習

1. **均方誤差（mean squared error）。**有偏估計量 M 滿足 $\mathrm{E}(M-N)^2 = \mathrm{Var}(M) + \big(\mathrm{E}(M) - N\big)^2$。計算此值，並與 $\mathrm{Var}(Z_1)$ 比較。求使 $\mathrm{E}(cM - N)^2$ 取最小值的常數 c。

2. **最小值也可以。**證明 $(n+1)\min\{X_i\}$ 也是 N 的無偏估計量，並求它的方差。這個估計量好用嗎？

3. **離散公式。**對於序號 $1, \dots, N$，以不放回方式抽取 k 個，其中最大者為 m。證明 $\mathrm{P}(\max \le m) = \binom{m}{k} / \binom{N}{k}$，並驗證當 $k = 1$ 時 $m + \frac{m}{k} - 1$ 是無偏的。

4. **坦克。**若 5 輛繳獲的坦克的序號為 19、40、42、60、98，試估計坦克的生產數量。

延伸閱讀

- **Richard Ruggles & Henry Brodie, "An empirical approach to economic intelligence in World War II"**，*Journal of the American Statistical Association* 42 (1947)。最初的記述。
- **John Rice, *Mathematical Statistics and Data Analysis***（Duxbury），論及偏差、方差與充分性。
- **David Spiegelhalter, *The Art of Statistics***（Pelican）。

13.4 平均數還是中位數？

題目

STEP II 2000，第 14 題

繼續閱讀之前，請先看題目（原題為英文）。題目載於 STEP Support Programme 的試題資料庫：

`step.maths.org/questions/00-s2-q14`

引子

要估計一組量度數據的中心，你可以取平均值（mean），也可以取中位數（median）。哪一個較好？答案視乎數據而定。假如量度結果偶爾會出現離譜的誤差，中位數就可靠得多：一個荒謬的數值足以把平均值拉到任何地方，卻幾乎動搖不了中位數。至於規規矩矩的數據，通常是平均值勝出。本題要量度的，正是在數據服從均勻分佈（uniform distribution）時，平均值究竟勝出多少：中位數的方差（variance）是平均值的三倍，所以要令中位數與平均值同樣精確，就需要三倍數量的觀測值。

這個比較引入了*效率*（efficiency）的概念，它是選擇統計方法時的主要考慮之一；同時也用到 Beta 積分（beta integral），這個積分在 Section 4.4 也出現過。

動筆之前

? **想一想** 為甚麼中位數的密度（density）與 $y^n(1-y)^n$ 成正比？（要 Y 接近 y，$2n+1$ 個值之中須有多少個在其下，又有多少個大於 y？）

中位數要落在 y，就必須有一個值在 y，n 個值在其下（各自的概率為 y），另有 n 個值在其上（各自的概率為 $1-y$）。這就得出因子 $y^n(1-y)^n$；常數則是數算哪些值扮演哪個角色的選法數目。

逐步解題

階段 1　常數

密度的積分必須等於 1：

$$k\int_0^1 y^n(1-y)^n\,\mathrm{d}y = k\cdot\frac{n!\,n!}{(2n+1)!} = 1, \qquad k = \frac{(2n+1)!}{(n!)^2}. \qquad \blacksquare$$

階段 2　平均值與方差

密度關於 $\frac{1}{2}$ 對稱（把 y 換成 $1-y$ 即可看出），所以 $\mathrm{E}(Y) = \frac{1}{2}$。

> **想一想**　利用 Beta 積分計算 $\mathrm{E}(Y^2) = k\int_0^1 y^{n+2}(1-y)^n\,\mathrm{d}y$，並化簡當中的階乘。

$$\mathrm{E}(Y^2) = \frac{(2n+1)!}{(n!)^2}\cdot\frac{(n+2)!\,n!}{(2n+3)!} = \frac{(n+2)(n+1)}{(2n+3)(2n+2)} = \frac{n+2}{2(2n+3)}.$$

因此

$$\mathrm{Var}(Y) = \frac{n+2}{2(2n+3)} - \frac{1}{4} = \frac{2(n+2)-(2n+3)}{4(2n+3)} = \frac{1}{4(2n+3)}. \qquad \blacksquare$$

階段 3　與平均值比較

> **想一想**　$\bar{X}$ 是 $2n+1$ 個獨立均勻變量的平均，它的平均值和方差是多少？

每個 X_i 的平均值為 $\frac{1}{2}$，方差為 $\frac{1}{12}$，所以 $\mathrm{E}(\bar{X}) = \frac{1}{2}$，而 $\mathrm{Var}(\bar{X}) = \frac{1}{12(2n+1)}$。當 n 很大時，

$$\mathrm{Var}(Y) \approx \frac{1}{8n}, \qquad \mathrm{Var}(\bar{X}) \approx \frac{1}{24n}.$$

> **想一想**　假設 Y 和 $\bar{X}$ 都服從正態分佈（normal distribution），平均值為 $\frac{1}{2}$，標準差（standard deviation）分別為 σ_Y 和 $\sigma_{\bar{X}}$，試用標準正態分佈函數 Φ 表示這兩個概率。

$\mathrm{P}\big(|Y-\frac{1}{2}| < \frac{d}{\sqrt{n}}\big) = 2\Phi\big(\frac{d}{\sqrt{n}\,\sigma_Y}\big) - 1$，而 $\sqrt{n}\,\sigma_Y \to \frac{1}{\sqrt{8}}$；$\bar{X}$ 的情況相似，其中 $\sqrt{n}\,\sigma_{\bar{X}} \to \frac{1}{\sqrt{24}}$。所以兩個概率分別趨向 $2\Phi(d\sqrt{8}) - 1$ 和 $2\Phi(d\sqrt{24}) - 1$。由於 Φ 是遞增函數，而且 $\sqrt{24} > \sqrt{8}$，後者較大，因此對所有足夠大的 n，不等式都成立。 ■平均值較有可能接近真正的中心。

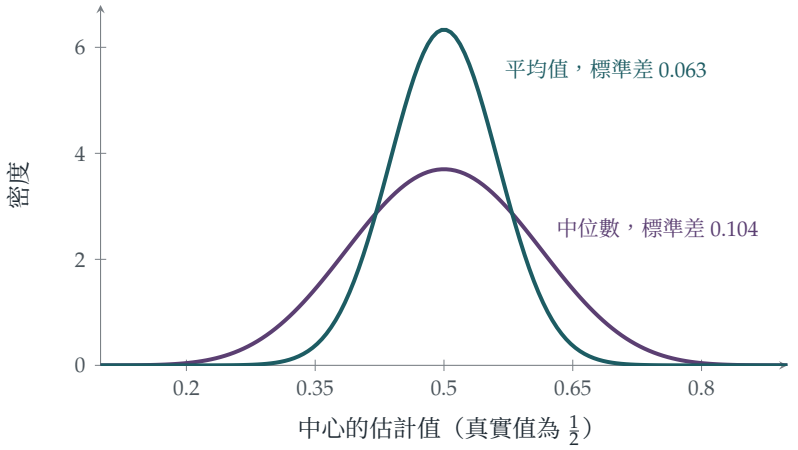


圖 13.4.1 21 個獨立均勻觀測值的中位數（紫紅色）和平均值（藍綠色）的抽樣分佈（sampling distribution）。兩者都以真實值 $\frac{1}{2}$ 為中心，但平均值集中得多：它的方差大約是中位數方差的三分之一。

回顧

- **Beta 積分。** $\int_0^1 y^r(1-y)^s \, \mathrm{d}y$ 這一條公式，同時給出了歸一化常數（normalising constant）和各階矩（moments）。
- **對稱性給出平均值。** 求 $\mathrm{E}(Y)$ 根本不用積分。
- **在合適的尺度上比較離散程度。** 把標準差乘以 $\sqrt{n}$，比較的結果便與 n 無關。

背後的數學

效率

兩個無偏估計量（unbiased estimator）的*相對效率*（relative efficiency）是它們方差之比；它說明其中一個需要多少倍的觀測值，才能與另一個看齊。對於均勻分佈的數據，中位數相對於平均值的效率是 $\frac{1}{3}$。對於正態分佈的數據，這個數字是 $\frac{2}{\pi} \approx 0.64$：中位數大約需要多 57% 的觀測值。但遇上厚尾（heavy-tailed）數據，形勢就逆轉了。對於柯西（Cauchy）分佈（Section 12.22），平均值毫無用處，因為 n 個觀測值的平均並不比單獨一個觀測值好，而中位數卻表現良好。對於拉普拉斯（Laplace）分佈（Section 12.24），中位數的效率是平均值的兩倍。

穩健統計學

中位數是*穩健*（robust）的：即使把接近一半的數據改成任意數值，它也不會移動太遠。平均值卻沒有這種保障：一個錯誤的數值就足以把它毀掉。在 1960 及 1970 年代，John Tukey、Peter Huber 等人發展出*穩健統計學*（robust statistics），即一些估計量，對乾淨的數據而言效率幾乎與平均值一樣高，卻又能抵禦離羣值（outlier），例如*截尾平均值*（trimmed mean）（捨去最大和最小的 10%，再取餘下數據的平均），跳水、體操等運動項目的評分就採用這種方法。

拉普拉斯與中位數

拉普拉斯在 1774 年證明，假如誤差服從雙指數（拉普拉斯）分佈，則真實值的最佳估計是中位數；高斯（Gauss）在 1809 年證明，假如誤差服從正態分佈，則最佳估計是平均值，他並由此得出最小二乘法（method of least squares）。誰是誰非，取決於誤差的表現如何；每遇上一種新的數據，統計學家至今仍會為這個問題爭論一番。

思維技巧

以變異程度比較不同方法

兩種方法可能平均而言都正確；要比較的是它們的答案變動有多大，以及數據增加時這種變動收縮得有多快。

在大學裏，漸近效率（asymptotic efficiency），即比較大樣本下的方差，是在不同統計方法之間作取捨的標準工具；與之並用的還有穩健性，以應付模型可能出錯的情況。

延伸練習

1. **中位數的密度。**用數算的方法推導密度 $\frac{(2n+1)!}{(n!)^2}y^n(1-y)^n$：一個值落在 $[y, y+\delta y]$ 內，n 個在其下，n 個在其上。

2. **三個觀測值。**當 $n=1$（三個均勻觀測值）時，求 $\mathrm{Var}(Y)$ 和 $\mathrm{Var}(\bar{X})$ 的準確值。哪一個較小？

3. **中列數。**中心的另一個估計量是中列數（midrange）$\frac{1}{2}(\max+\min)$。證明對於均勻分佈的數據，若有 $m=2n+1$ 個觀測值，它的方差是 $\frac{1}{2(m+1)(m+2)}$。它與平均值相比如何？（試與 Section 13.3 比較。）

4 **一個離羣值。**十一個量度結果為 9.8、9.9、10.0、10.0、10.1、10.1、10.2、10.2、10.3、10.4 和 104（打字錯誤）。比較平均值和中位數。

延伸閱讀

- **John Rice, *Mathematical Statistics and Data Analysis***（Duxbury），論及順序統計量（order statistics）和效率。
- **Rand Wilcox, *Fundamentals of Modern Statistical Methods***（Springer）。為非專家講解穩健方法為何重要。
- **Stephen Stigler, *The Seven Pillars of Statistical Wisdom***（Harvard University Press）。從歷史角度看數據的匯總、平均值和中位數。

13.5 問尷尬的問題

題目

STEP II 2023，第 11 題

繼續閱讀之前，請先看題目（原題為英文）。題目載於 OCR 提供下載的 2023 年 STEP 試卷（zip 檔案）內的 STEP II 試卷：

`www.ocr.org.uk/Images/707102-2023-past-papers.zip`

引子

怎樣才能知道有多少比例的學生曾在考試中作弊，或者有多少比例的僱員曾偷取公司財物？直接問的話，很多人都會說謊。1965 年，統計學家 Stanley Warner 想出了一個巧妙的解決辦法：隨機化回答（randomised response）。每個人私下擲一枚硬幣或一顆骰子，由結果決定他是如實回答那條敏感問題，還是回答與之相反的問題。訪問員永遠不知道是哪一種情況，所以任何一個人的答案都不足以成為對他不利的證據；然而，由於隨機化的概率（probability）是已知的，整體的比例仍然可以估計出來。

本題探討的正是這背後的數學。讓每個人獨立地（independently）隨機化（第 (ii) 部分），所得的總數性質良好，足以確定該比例。用同一次擲硬幣為所有人隨機化（第 (iii) 部分），所得的總數雖然平均值（mean）相同，分佈（distribution）卻截然不同，呈雙峯形狀，幾乎不能告訴我們任何事。隨機性放在哪裏，至關重要。

動筆之前

想一想 Y 以概率 p 等於 X_1，以概率 q 等於 X_2。$\mathrm{E}(Y)$ 是甚麼？$\mathrm{E}(Y^2)$ 又是甚麼？

以選中哪一個變量為條件，可得 $\mathrm{E}(Y) = p\mu_1 + q\mu_2$，以及 $\mathrm{E}(Y^2) = p\,\mathrm{E}(X_1^2) + q\,\mathrm{E}(X_2^2) = p(\sigma_1^2 + \mu_1^2) + q(\sigma_2^2 + \mu_2^2)$。

逐步解題

階段 1　混合分佈的方差

$$\mathrm{Var}(Y) = p(\sigma_1^2 + \mu_1^2) + q(\sigma_2^2 + \mu_2^2) - (p\mu_1 + q\mu_2)^2.$$

含 μ 的各項為 $p\mu_1^2 + q\mu_2^2 - p^2\mu_1^2 - 2pq\mu_1\mu_2 - q^2\mu_2^2 = pq\mu_1^2 + pq\mu_2^2 - 2pq\mu_1\mu_2$，當中用到 $p - p^2 = pq$ 和 $q - q^2 = pq$。因此

$$\mathrm{Var}(Y) = p\sigma_1^2 + q\sigma_2^2 + pq(\mu_1 - \mu_2)^2. \qquad \blacksquare$$

混合分佈（mixture）的方差（variance），等於各成分內部方差的平均，再加上各成分平均值之間的方差。

階段 2　獨立隨機化

> **想一想**　$\mathrm{P}(B = 1)$ 是多少？B 的 n 個值是否獨立？

$\mathrm{P}(B = 1) = \frac{1}{2}\cdot\frac{1}{6} + \frac{1}{2}\cdot\frac{5}{6} = \frac{1}{2}$。這 n 個值互相獨立，每個的成功概率都是 $\frac{1}{2}$，所以 $Z_1 \sim \mathrm{B}(n, \frac{1}{2})$，其平均值為 $\frac{n}{2}$，標準差（standard deviation）為 $\frac{\sqrt{n}}{2}$。

與平均值相差不超過 10%，即是 $|Z_1 - \frac{n}{2}| < \frac{n}{20}$。以標準單位表示，這就是 $|Z| < \frac{n/20}{\sqrt{n}/2} = \frac{\sqrt{n}}{10}$，而此值趨向無窮大，所以該概率趨向 1。 $\blacksquare$

階段 3　所有人共用一枚硬幣

> **想一想**　現在只擲一次硬幣，便決定全部 n 顆骰子的情況。已知擲得正面，Z_2 的分佈是甚麼？已知擲得反面呢？試用第 (i) 部分的結果。

已知擲得正面，$Z_2 \sim \mathrm{B}(n, \frac{1}{6})$：平均值為 $\frac{n}{6}$，方差為 $\frac{5n}{36}$。已知擲得反面，$Z_2 \sim \mathrm{B}(n, \frac{5}{6})$：平均值為 $\frac{5n}{6}$，方差為 $\frac{5n}{36}$。在第 (i) 部分取 $p = q = \frac{1}{2}$，得

$$\mathrm{E}(Z_2) = \frac{n}{2}, \qquad \mathrm{Var}(Z_2) = \frac{5n}{36} + \frac{1}{4}\left(\frac{4n}{6}\right)^2 = \frac{5n}{36} + \frac{n^2}{9}.$$

> **想一想**　大部分的概率集中在哪裏？在平均值 $\frac{n}{2}$ 附近的概率多嗎？

Z_2 有兩個峯，分別在 $\frac{n}{6}$ 和 $\frac{5n}{6}$ 附近，每個峯的闊度約為 $\sqrt{n}$（圖 13.5.1），而兩峯之間幾乎甚麼都沒有。具有相同平均值和方差的正態分佈（normal distribution）則只有一個峯，中心位於 $\frac{n}{2}$，正正是 Z_2 幾乎從不出現的位置。要 Z_2 與 $\frac{n}{2}$ 相差不超過 10%，一個服從 $\mathrm{B}(n, \frac{1}{6})$ 或 $\mathrm{B}(n, \frac{5}{6})$ 的變量必須落在 $[0.45n, 0.55n]$ 之內。對 $\mathrm{B}(n, \frac{1}{6})$ 而言，這比其平均值

高出至少 $0.28n$，約為 $\frac{0.28n}{0.37\sqrt{n}} \approx 0.75\sqrt{n}$ 個標準差，所以該概率趨向 0；$\mathrm{B}(n, \frac{5}{6})$ 的情況亦類似。 ■

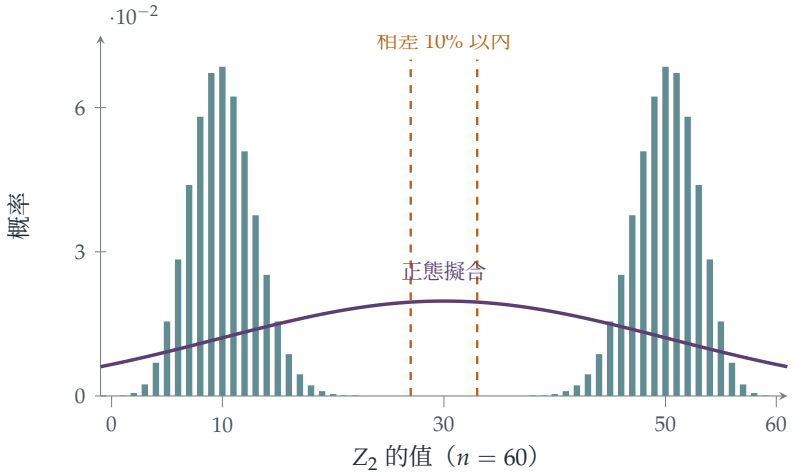


圖 13.5.1 $n = 60$ 顆骰子時 Z_2 的分佈（長條）：兩個峯分別位於 10 和 50，來自那一次擲硬幣的兩種結果。具有相同平均值（30）和方差的正態分佈（紫紅色）擬合得很差，而且幾乎沒有概率落在與平均值相差 10% 以內的範圍（兩條琥珀色線之間）。

回顧

- **以隱藏的選擇為條件。**混合分佈的平均值和方差，都是以選中哪一個成分為條件而求得的。
- **內部方差加上組間方差。**額外的一項 $pq(\mu_1 - \mu_2)^2$ 量度了兩個成分之間的離散程度。
- **平均值和方差並非全部。**Z_2 與 Z_1 的平均值相同，但形狀完全不同。

背後的數學

隨機化回答

在 Warner 的方案中，每位受訪者私下作出選擇，以概率 p 回答「你是否屬於 A 組？」，以概率 $1 - p$ 回答「你是否不屬於 A 組？」。若屬於 A 的人所佔比例為 π，則答「是」的概率為

$$\lambda = p\pi + (1 - p)(1 - \pi),$$

因此由觀察所得答「是」的比例 $\hat{\lambda}$，我們可以估計 $\hat{\pi} = \frac{\hat{\lambda} - (1-p)}{2p-1}$（前提是 $p \neq \frac{1}{2}$）。本題第 (ii) 部分正是這種情況，由每個人自己的骰子來決定。私隱的代價是精確度：除以

$2p - 1$ 會令方差增大，所以隨機化調查所需的受訪者比直接調查更多。不過每個人都是獨立地隨機化的，所以大數定律（law of large numbers）能夠發揮作用。第 (iii) 部分顯示了共用隨機化時會出甚麼問題：那一枚硬幣所帶來的不確定性，永遠不會被平均抵銷。

差分私隱

隨機化回答是*差分私隱*（differential privacy）的前身。差分私隱是 Cynthia Dwork 及其同事於 2006 年提出的一個關於私隱的數學定義：如果加入或移除任何一個人的數據，只會令任何輸出的概率改變至多一個很小的倍數，該程序便算是保障私隱的。科技公司用它來收集使用統計數據，美國人口普查局亦在 2020 年人口普查中採用了它，在公佈的數字上加入經仔細校準的隨機雜訊，令任何個人的答案都無法被還原。

全方差定律

第 (i) 部分的恆等式是*全方差定律*（law of total variance）的一個特例：$\operatorname{Var}(Y) = \operatorname{E}\big(\operatorname{Var}(Y \mid C)\big) + \operatorname{Var}\big(\operatorname{E}(Y \mid C)\big)$，其中 C 是隱藏的選擇。它是*方差分析*（analysis of variance，ANOVA）的基礎。方差分析由費雪（Ronald Fisher）於 1920 年代在 Rothamsted 為農業實驗而引入，它把農作物產量的變異分拆為兩部分：可由肥料或品種解釋的部分，以及未能解釋的部分。

思維技巧

要看整個分佈，而不只是它的概括數值

平均值和方差能很好地描述單峯分佈，但對混合分佈卻可能造成嚴重誤導。要問一問：概率實際上落在哪裏。

在大學裏，混合模型（mixture model）用來描述由幾個羣組組成的總體，而全方差定律則是統計學的標準工具，從調查設計到階層模型（hierarchical model）都會用到。

延伸練習

1 **雙胞胎。**試做 STEP II 2005，第 12 題（`step.maths.org/questions/05-s2-q12`）。這其實是一道喬裝了的隨機化回答謎題：在甚麼情況下，一個答案完全不能透露硬幣的結果？

2 **估計 π。**在一項 $p = \frac{3}{4}$ 的 Warner 調查中，有 400 人作答，其中 160 人答「是」。試估計 A 組所佔的比例 π，並求出你的近似標準誤差（standard error）。

3 **全方差。**設 C 為取多個值的離散隱藏變量，證明全方差定律。

4 **共用硬幣。**在第 (iii) 部分中，$Z_2 > \frac{n}{2}$ 的概率是多少？試解釋為何當 n 很大時，這個概率接近 $\frac{1}{2}$，儘管 Z_2 幾乎從不接近 $\frac{n}{2}$。

延伸閱讀

- **Stanley Warner, "Randomized response: a survey technique for eliminating evasive answer bias"**，*Journal of the American Statistical Association* 60 (1965)。
- **Cynthia Dwork & Aaron Roth, *The Algorithmic Foundations of Differential Privacy***（可於網上免費閱讀）。現代理論；開首幾章不難讀懂。
- **David Spiegelhalter, *The Art of Statistics***（Pelican），談調查及其陷阱。

13.6 累積起來的誤差

題目

STEP II 1993，第 15 題

繼續閱讀之前，請先看題目（原題為英文）。題目載於 STEP Support Programme 的試題資料庫：

`step.maths.org/questions/93-s2-q15`

引子

1982 年，溫哥華證券交易所推出了一個新指數，起始值為 1000。每當有股價變動，指數便會重新計算，並截斷（truncate）至小數點後三位。每次截斷平均損失 0.0005，看來微不足道。可是每天大約要重新計算 3000 次，二十二個月後，儘管大市上升，指數卻只剩下約 520。經正確重算後，指數一夜之間躍升至約 1100。

這道題目解釋了箇中原因。截斷誤差（truncation error）全部朝同一方向，因此會穩步累積，與運算次數成正比。捨入誤差（rounding error）則有正有負，大部分互相抵銷：它們的增長速度只相當於運算次數的平方根，這正是中心極限定理（central limit theorem）（Section 12.24）所預測的隨機漫步（random walk）行為。

動筆之前

? **想一想** 若某數的小數部分 U 在 $[0,1)$ 上均勻分佈（uniform），把該數截斷時誤差是多少？捨入時又是多少？兩者的平均值（mean）和方差（variance）各是多少？

截斷是以 $\lfloor a \rfloor$ 取代 a，誤差（真確值減近似值）為 U，在 $[0,1)$ 上均勻分佈：平均值為 $\frac{1}{2}$，方差為 $\frac{1}{12}$。捨入所得的誤差在 $(-\frac{1}{2}, \frac{1}{2}]$ 上均勻分佈：平均值為 0，方差為 $\frac{1}{12}$。

逐步解題

階段 1　總誤差

加了 1500 個數之後，總誤差是 1500 個獨立（independent）誤差之和，根據中心極限定理，它近似於正態分佈（normal distribution）：

- LEP（捨入）：平均值為 0，方差為 $\frac{1500}{12} = 125$，標準差（standard deviation）為 $\sqrt{125} \approx 11.18$。
- VOZ（截斷）：平均值為 $\frac{1500}{2} = 750$，標準差同樣是 11.18。

階段 2　出現大誤差的概率

> **想一想**　對 LEP 而言，若 $E \sim \mathrm{N}(0, 125)$，$\mathrm{P}(|E| > 15)$ 是多少？對 VOZ 而言，誤差幾乎必然落在哪裏？

對 LEP 而言，$\frac{15}{11.18} \approx 1.342$，所以

$$\mathrm{P}(|E| > 15) \approx 2\big(1 - \Phi(1.342)\big) \approx 2 \times 0.090 = 0.18.$$

對 VOZ 而言，誤差約為 750 ± 11；其絕對值小於 15 的概率基本上是零，所以 $\mathrm{P}(|E| > 15) \approx 1$。 ■

階段 3　可以做多少次加法？

> **想一想**　做了 n 次加法後，LEP 的誤差的標準差為 $\sqrt{n/12}$。何時有 $\mathrm{P}(|E| < 10) \geq 0.9$？

我們需要 $\frac{10}{\sqrt{n/12}} \geq 1.645$（因為 $\mathrm{P}(|Z| < 1.645) = 0.9$），即 $\frac{n}{12} \leq \left(\frac{10}{1.645}\right)^2 \approx 36.95$，所以 $n \leq 443$。LEP 可以做大約 443 次加法。 ■

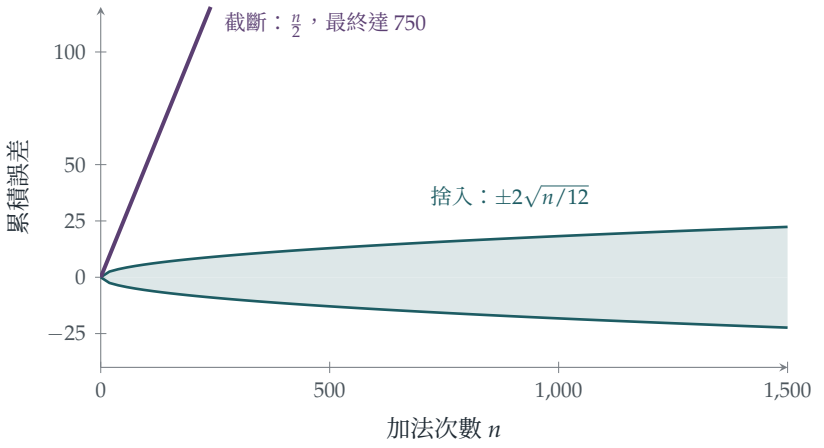


圖 13.6.1 誤差如何累積。截斷誤差（紫紅色）全部同號，與 n 成正比增長：做了 1500 次加法後，總誤差約為 750。捨入誤差（藍綠色帶狀區域，即零的兩側各兩個標準差）平均而言互相抵銷，增長速度只相當於 $\sqrt{n}$：做了 1500 次加法後，誤差通常在約 ±22 之內。

回顧

- **為每個誤差建立模型。**小數部分均勻分佈，令每個誤差都有簡單的分佈。
- **把平均值和方差相加。**平均值可以相加（得出偏差 $\frac{n}{2}$），而對於獨立的誤差，方差同樣可以相加。
- **正態近似。**中心極限定理把 1500 個均勻誤差之和化為正態分佈。

背後的數學

偏差與方差

截斷是*有偏*（biased）的：其誤差的平均值不是零，而偏差（bias）會線性累積。捨入是*無偏*（unbiased）的，其誤差的表現有如隨機漫步，增長速度相當於 $\sqrt{n}$。經過一百萬次運算後，兩者的分別就是數量級為 500 000 的誤差與數量級為 300 的誤差之別。這與實驗中系統誤差（systematic error，例如儀器校準失準）和隨機測量誤差（random error）之間的區別如出一轍：把多次測量結果取平均，可以減少隨機誤差，對系統誤差卻毫無作用。

浮點數

所有電腦都只能以有限的精確度儲存實數，每次算術運算都要把結果捨入。現代電腦遵從 IEEE 754 標準，把結果捨入至最接近的可表示數；若剛好位於兩數正中，則取末位為偶數的一個，目的正是令捨入平均而言無偏。即使如此，把大量數字相加時若不小心處理，仍會損失準確度；William Kahan 的*補償求和*（compensated summation）算法（1965 年）會記錄失去的低位數字，再把它們加回去。1991 年波斯灣戰爭期間，一個愛國者導彈連未能攔截一枚來襲導彈，原因是它的時鐘以十分之一秒為單位計時，而這個數無法以二進制準確儲存；連續運作 100 小時後，截斷誤差已增至約三分之一秒，足以令目標位置的估算偏差超過半公里。

溫哥華指數

溫哥華證券交易所指數於 1982 年 1 月設立，起始值為 1000.000。到 1983 年 11 月，儘管大市普遍上升，指數卻已跌至約 524。顧問發現，指數每天重新計算數千次，而每次都截斷（而非捨入）至小數點後三位。經過一個周末的正確重算，指數重開時約為 1098。這件事已成為數值計算課程中的經典警示。

思維技巧

分清系統誤差與隨機誤差

正負號一致的誤差會線性累積；平均值爲零的獨立誤差則按平方根的速度累積。先認清你面對的是哪一種。

在大學，數值分析（numerical analysis）研究誤差如何在計算過程中傳播，統計學則區分偏差與方差；「偏差–方差權衡」（bias–variance trade-off）是現代機器學習的核心概念。

延伸練習

1. **修正 VOZ。**若 VOZ 在最後把 $\frac{n}{2}$ 加到答案上，其誤差的分佈是甚麼？它現在是否與 LEP 一樣好？

2. **捨入至小數點後一位。**若把各數捨入至小數點後一位，而被捨去的數字均勻分佈，求把 10 000 個數相加時總誤差的標準差。

3 **溫哥華指數。**若指數每天重新計算 3000 次，每次截斷平均損失 0.0005，那麼在一年 250 個交易日內，指數會損失多少？

4 **精確分佈。**求兩個捨入誤差之和的精確密度函數（density，呈三角形）。若是三個誤差之和，試繪出其草圖。它多快便開始貌似正態分佈？

延伸閱讀

- **Nicholas Higham, *Accuracy and Stability of Numerical Algorithms***（SIAM）。關於捨入誤差的權威著作；第一章淺白易讀。
- **David Goldberg, "What every computer scientist should know about floating-point arithmetic"**, *ACM Computing Surveys* 23 (1991)。經典之作，可在網上免費閱讀。
- **Tim Harford, *How to Make the World Add Up***（Bridge Street），探討數據中的系統誤差。

13.7 製造相關性

題目

STEP III 2010，第 13 題

繼續閱讀之前，請先看題目（原題為英文）。題目載於 STEP Support Programme 的試題資料庫：

`step.maths.org/questions/10-s3-q13`

引子

一間銀行想知道，手上的股票投資組合(portfolio)有多大機會虧蝕。股價往往一起變動：一隻股票下跌，其他股票也傾向跟着下跌。要模擬數以千計的可能未來，銀行必須產生一批隨機數，而這些隨機數之間的相關程度要恰到好處。如果手上只有互相獨立（independent）的隨機變量（random variable），怎樣才能製造出具有指定相關係數（correlation）的隨機變量呢？

這道題就三個變量的情形，一步一步給出了配方。每個新變量都是先前各個獨立材料的組合，再加上一份全新的獨立材料，用來補足方差（variance）。這個構造其實是線性代數(linear algebra)中一個著名結果的化身，即科列斯基分解(Cholesky decomposition)，它同時揭示了一個隱藏的限制：並非每一組相關係數都有可能出現。

關於課程範圍的說明

協方差（covariance）和相關係數已於 2019 年從 STEP 的考試範圍中刪去。協方差是 $\mathrm{Cov}(U,V) = \mathrm{E}(UV) - \mathrm{E}(U)\mathrm{E}(V)$；獨立變量的協方差為零，而且協方差對每個變元都是線性的：$\mathrm{Cov}(aU+bV,W) = a\,\mathrm{Cov}(U,W) + b\,\mathrm{Cov}(V,W)$。此外 $\mathrm{Var}(U) = \mathrm{Cov}(U,U)$。相關係數必定介乎 -1 與 1 之間。

動筆之前

想一想　Z_1 和 Z_2 互相獨立。它們的協方差是多少？相關係數又是多少？

獨立變量滿足 $\mathrm{E}(Z_1Z_2) = \mathrm{E}(Z_1)\,\mathrm{E}(Z_2)$。所以協方差是 0，而 $\mathrm{Corr}(Z_1, Z_2) = 0$。

逐步解題

階段 1　兩個變量

$\mathrm{E}(Y_2) = 0$。由於 Z_1 和 Z_2 互相獨立，方差均為 1，

$$\mathrm{Var}(Y_2) = \rho_{12}^2 + (1-\rho_{12}^2) = 1, \qquad \mathrm{Cov}(Y_1, Y_2) = \mathrm{Cov}(Z_1, \rho_{12}Z_1 + \sqrt{1-\rho_{12}^2}\,Z_2) = \rho_{12}.$$

兩個方差都是 1，所以 $\mathrm{Corr}(Y_1, Y_2) = \rho_{12}$。 ■

想一想　把 Z_1 和 Z_2 想像成互相垂直的單位向量。Y_2 在哪裏？Y_1 與 Y_2 之間夾角的餘弦是多少？

$Y_2 = \rho_{12}Z_1 + \sqrt{1-\rho_{12}^2}\,Z_2$ 是一個單位向量，與 $Z_1 = Y_1$ 的夾角為 θ，其中 $\cos\theta = \rho_{12}$。把變量看成向量，相關係數就是它們之間夾角的餘弦（圖 13.7.1）。

階段 2　第三個變量

想一想　試以 a、b、c 寫出 $\mathrm{Var}(Y_3) = 1$、$\mathrm{Cov}(Y_1, Y_3) = \rho_{13}$、$\mathrm{Cov}(Y_2, Y_3) = \rho_{23}$ 這三個條件。

利用線性和獨立性：

$$a^2 + b^2 + c^2 = 1, \qquad a = \rho_{13}, \qquad \rho_{12}a + \sqrt{1-\rho_{12}^2}\,b = \rho_{23}.$$

所以

$$a = \rho_{13}, \qquad b = \frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}}, \qquad c = \sqrt{1 - \rho_{13}^2 - \frac{(\rho_{23} - \rho_{12}\rho_{13})^2}{1-\rho_{12}^2}}.$$ ■

（由於 $c \geq 0$，這裏取非負平方根。）

階段 3　任意的均值和方差

縮放和平移都不會改變相關係數，所以

$$X_i = \mu_i + \sigma_i Y_i \qquad (i = 1, 2, 3)$$

的均值為 μ_i，方差為 σ_i^2，相關係數為 ρ_{ij}。 ■

> **停一停，試一試** c 的公式要求平方根內的量非負。試取 $\rho_{12} = \rho_{13} = \rho_{23} = -0.6$。哪裏出了問題？三個變量不可能兩兩之間都有很強的負相關，為甚麼這是合理的？

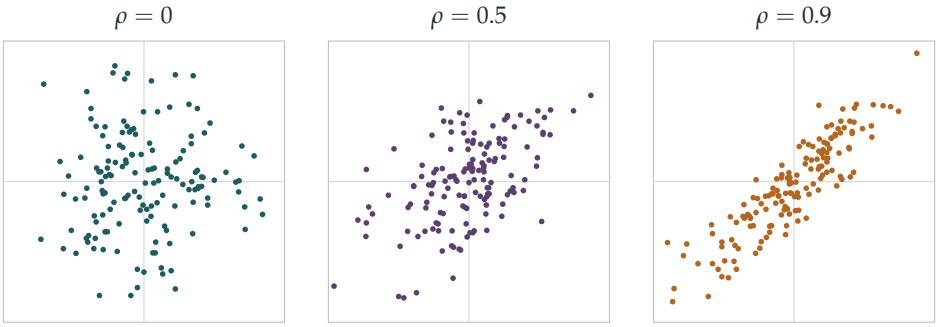


圖 13.7.1 (Y_1, Y_2) 的樣本，由獨立的標準正態（standard normal）變量 Z_1, Z_2 按配方 $Y_1 = Z_1, Y_2 = \rho Z_1 + \sqrt{1-\rho^2}\, Z_2$ 製成，分別對應 $\rho = 0$、0.5 和 0.9。同樣的獨立材料，可以造出我們任意選定的相關係數。

回顧

- **每次只構造一個變量。**每個新變量都從先前的材料中取用適當的份量，再加入一份全新的獨立材料。
- **協方差是線性的。**所有條件都化成了關於係數的簡單方程。
- **相關係數是餘弦。**把標準化後的變量看成單位向量，相關係數就變成了夾角。

背後的數學

科列斯基分解

用矩陣形式來寫，這個構造就是 $\mathbf{Y} = L\mathbf{Z}$，其中的矩陣是下三角（lower triangular）矩陣

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \rho_{12} & \sqrt{1-\rho_{12}^2} & 0 \\ a & b & c \end{pmatrix},$$

而 $\mathbf{Y}$ 的相關矩陣（correlation matrix）是 $R = LL^{\mathsf{T}}$。把一個對稱矩陣寫成 LL^{T} 的形式，就是科列斯基分解，由法國軍事工程師科列斯基（André-Louis Cholesky）約於 1910 年

發現，用來求解大地測量中的方程組（他在第一次世界大戰中陣亡，這項成果在他身後才發表）。只要 R 是正定（positive definite）的，這個分解就行得通，而計算過程正是我們剛才用手做的：逐行求解。在金融、工程和氣候模型中，這是模擬相關變量的標準方法。

哪些相關係數是可能的？

c 的平方根內的數必須非負。展開後，條件是

$$1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23} \geq 0,$$

這正是相關矩陣的行列式（determinant）非負的條件。從幾何上看：三個單位向量兩兩之間的夾角為 θ_{ij}（其中 $\cos\theta_{ij} = \rho_{ij}$），它們必須真的能在空間中拼合起來。如果 $\rho_{12} = \rho_{13} = \rho_{23} = -0.6$，各個夾角都約為 127°，而三個兩兩相隔 127° 的向量，夾角之和就要超過 360°：這是不可能的。一般而言，一個矩陣要成為相關矩陣，必須是半正定（positive semi-definite）的，即所有特徵值（eigenvalue）均非負（Section 7.2）。如果各個相關係數是根據不完整的數據分別估計出來的，這個條件就可能不成立，統計學家於是要找出最接近的有效相關矩陣，這個問題由 Nicholas Higham 在 2002 年解決。

高爾頓、皮爾遜與相關係數

高爾頓（Francis Galton）在 1888 年研究父母與子女身高的關係時引入了相關的概念，而皮爾遜（Karl Pearson）在 1896 年給出了本題所用的積矩（product moment）公式。高爾頓還注意到均值回歸（regression to the mean）的現象（Section 13.1）。皮爾遜於 1911 年在倫敦大學學院（University College London）創立了全世界第一個大學統計學系。「相關不等於因果」這句警告，幾乎與相關係數本身一樣古老。

思維技巧

逐步構造，每次加入全新的獨立材料

要構造具有指定關係的對象，可以逐個加入，每一個都是先前各個對象的組合，再加上一些全新而獨立的東西。

這是線性代數中的格拉姆–施密特（Gram–Schmidt）正交化過程倒過來做，也就是科列斯基分解。在大學裏，它會出現在最小二乘法（least squares）、多元正態分佈（multivariate normal distribution）的理論，以及隨機過程的模擬之中。

延伸練習

1 **不可能的相關係數。**證明：除非 $\rho \geq -\frac{1}{2}$，否則三個隨機變量不可能兩兩之間的相關係數都是 ρ。對於 n 個變量，相應的結論是甚麼？

2 **檢查行列式。**證明 $1 - \rho_{13}^2 - b^2 \geq 0$ 等價於 $\det R \geq 0$，其中 R 是 3×3 相關矩陣。

3 **投資組合的風險。**兩隻股票回報的標準差（standard deviation）分別為 $\sigma_1 = 0.2$ 和 $\sigma_2 = 0.3$，相關係數為 ρ。若投資組合中兩隻股票各佔一半，求 $\rho = -1$、0 和 1 時該組合的標準差。分散投資達到了甚麼效果？

4 **還原材料。**給定題目中的 Y_1、Y_2，試以 Y_1 和 Y_2 表示 Z_2。（這就是格拉姆–施密特過程中的一步。）

延伸閱讀

- **Gilbert Strang,** ***Introduction to Linear Algebra***（Wellesley-Cambridge），講述正定矩陣和科列斯基分解。
- **Paul Glasserman,** ***Monte Carlo Methods in Financial Engineering***（Springer）。講述實際應用中如何模擬相關變量；內容較深，但開頭幾章不難讀。
- **Theodore Porter,** ***Karl Pearson: The Scientific Life in a Statistical Age***（Princeton University Press）。

我們學到了甚麼

本章開首說過，概率（probability）是由模型走向數據，而統計（statistics）則是由數據回到現實世界。我們提出了三個問題：這是真有其事，還是純屬雜訊？應該怎樣估計？隨機性是怎樣進入數據的？以下是本章七道題目所揭示的答案。

1. 是真有其事，還是純屬雜訊？

在 Section 13.1 中，四則新聞標題把數字變成了論斷。比較總數而不理會分母、單憑一年或一日的數據就看出趨勢、把趨勢隱藏在平均數之中，這些全都是錯誤；而泊松模型（Poisson model）則提供了一把尺，用來量度計數數據的自然變異。上升了 8 個標準差（standard deviation），幾乎肯定是真有其事；某個星期二有七人死亡，則幾乎肯定不是。再加上均值回歸（regression to the mean）和公平比較的必要，整幅圖畫便完整了。

2. 應該怎樣估計？

有三道題目發展了估計的理論。

- **最大似然（maximum likelihood）**（Section 13.2）：選取令觀察所得數據出現的概率最大的未知數值。在捉放法（capture–recapture）中，它給出了自然的估計值 $\frac{200\times200}{11}\approx 3636$，而寬闊的似然函數則顯示出這個估計有多不確定。
- **偏差（bias）與方差（variance）**（Section 13.3）：潛艇數目有兩個無偏估計量（unbiased estimator），一個以最大值為基礎，另一個以平均值為基礎。最大值是充分統計量（sufficient statistic），精確得多，其方差按 $\frac{1}{n^2}$ 而非 $\frac{1}{n}$ 的速度縮小。同樣的推理，在戰時估計德軍坦克數量一事上，勝過了情報部門。
- **效率（efficiency）與穩健性（robustness）**（Section 13.4）：對於均勻分佈的數據，平均值的效率是中位數（median）的三倍；但對於重尾（heavy-tailed）數據，則是中位數勝出，而且它不受極端誤差影響，十分穩健。

3. 隨機性是怎樣進入數據的？

隨機性可以是刻意加入的：隨機化回答（randomised response）（Section 13.5）既保障每位受訪者的私隱，又仍然可以對總體作出估計，條件是每個人各自獨立地進行隨機化；假如所有人共用同一枚硬幣，便會得出一個雙峯的混合分佈，幾乎不能告訴我們任何東

西。隨機性也會在處理數據的過程中進入：四捨五入的誤差會互相抵銷，只按 $\sqrt{n}$ 的速度增長，而截尾誤差（truncation error）則會線性地累積，溫哥華證券交易所（Vancouver Stock Exchange）便親身領教過（Section 13.6）。隨機性還可以經人手設計出來：把獨立的成分以 Cholesky 構造法（Cholesky construction）組合起來，便可產生具有任何一組有效相關係數（correlation）的變量（Section 13.7），但並非每一組相關係數都是有效的。

4. 統計學的工具箱

方法	在哪裏用過
泊松變異	新聞標題（13.1）
似然及其最大值	田鼠（13.2）
無偏性與方差	序號（13.3）；中位數（13.4）
充分統計量	序號（13.3）
正態近似（normal approximation）	13.4、13.5、13.6
全方差定律（law of total variance）	隨機化回答（13.5）
偏差與隨機誤差	四捨五入（13.6）
協方差（covariance）的線性性質	相關係數（13.7）

下一步往哪裏去

- 貝葉斯定理（Bayes' theorem）（Section 12.2）是貝葉斯統計學的基礎，而貝葉斯統計學正是本章各種方法以外最主要的另一套方法。
- 中心極限定理（central limit theorem）（Section 12.24）為本章處處用到的正態近似提供了理據；柯西分佈（Cauchy distribution）（Section 12.22）則顯示這些近似在甚麼時候失效。
- 相關矩陣（correlation matrix）是半正定矩陣（positive semi-definite matrix），可以用 Section 7.2 的特徵向量（eigenvector）來研究。

如果你想在入讀大學之前再走遠一點

- **David Spiegelhalter, *The Art of Statistics***：為一般讀者而寫的統計思維入門書之中最好的一本。
- **Tim Harford, *How to Make the World Add Up***：教你怎樣解讀新聞中的數字。
- **John Rice, *Mathematical Statistics and Data Analysis***：大學數理統計的第一門課程。
- **Stephen Stigler, *The Seven Pillars of Statistical Wisdom***：統計學的重大思想及

其來源。

George Box 說過，所有模型都是錯的，但有些是有用的。本章嘗試說明怎樣令模型變得有用：知道應該預期多大的變異，以方法在多次重複之下的表現來評價方法，並且思考數據是怎樣得來的。正是這些習慣，把數字變成知識。

最後的話

全書到此結束。在這十三章裏，我們以 STEP 題目為門徑，走進大學一年級課程的數學：數論與分析、微積分與複數、代數與線性代數、微分方程與動力系統、組合數學、力學、概率與統計。每一道題目都是一份邀請，請你追問它背後藏着甚麼，而答案幾乎每一次都是一整門學科。如果你曾經手執鉛筆，認真做過其中哪怕只是幾節，思考頁邊提出的問題，並嘗試每節末的習題，你就已經開始像數學家那樣思考了。餘下的旅程，就由你自己去走。

延伸練習答案

以下只是簡短的答案和提示，並非完整解答：這裏只給出結果和關鍵步驟，方便你核對自己的做法。每道練習都應先認真做過，然後才看答案。

13.1 解讀新聞標題

1. $1 - \sum_{k=0}^{6} \mathrm{e}^{-2.96}\frac{2.96^k}{k!} \approx 0.0315$，所以一年之中大約有 11 天會出現七宗或以上的死亡個案。$\mathrm{P}(X \geq 9) \approx 0.0035$：大約一年一天。

2. 死亡人數為 $150K$，其中 K 服從平均值（mean）為 1.2 的泊松分佈（Poisson distribution）：平均值為 180，方差（variance）為 $150^2 \times 1.2 = 27\,000$，標準差（standard deviation）約為 164。死亡人數為 125 或以下的年份，就是沒有發生空難的年份，其概率為 $e^{-1.2} \approx 0.30$：完全是尋常事。但在這個模型中，死亡人數不可能恰好是 125，因為死亡人數只能是 150 的倍數。平均值和標準差都顯示不出這一點：個案扎堆出現，令分佈變得一團一團，全部概率都集中在 0、150、300 等數值上。

3. $z = \frac{1349.5-1080}{33} \approx 8.2$：概率的數量級為 10^{-16}。穩定不變的發生率無法解釋 1350 這個數字；必定是有些事情真的改變了（又或者是統計死亡人數的方法改變了）。

4. 例如：(i)「單憑總數，無法判斷乘飛機與步行哪樣較安全」；(ii)「1990 年空難死亡人數屬正常範圍」；(iii)「星期二七名行人死亡：不幸的一天，但這樣的日子一年約有十一次」；(iv)「1990 年行人死亡人數急升：必須追查原因」。

13.2 點算田鼠

1. $\frac{p_N}{p_{N-1}} = \frac{(N-M)(N-C)}{N(N-M-C+R)}$，而這個比值不小於 1，當且僅當 $NR \leq MC$。所以 p_N 一直遞增至 $N = \frac{MC}{R}$，然後遞減，而最大值就在該數的整數部分取得。

2. 從 a 名女性和 b 名男性中選出 c 人組成委員會，按其中女性的人數 j 來數：對每個 j，共有 $\binom{a}{j}\binom{b}{c-j}$ 種。

3. $\frac{\mathrm{d}}{\mathrm{d}p}\big(h \ln p + (n-h)\ln(1-p)\big) = \frac{h}{p} - \frac{n-h}{1-p} = 0$ 給出 $p = \frac{h}{n}$。

4. 當 $R = 0$ 時，p_N 永遠遞增，估計值是無限大。這個方法只有在第二個樣本中有足夠多已標記的動物時才行得通；重複捕獲的數目少的話，估計值會非常不穩定。

13.3 德國坦克問題

1. 設 M 為最大值，則 $\mathrm{E}(M) = \frac{nN}{n+1}$ 及 $\mathrm{Var}(M) = \frac{nN^2}{(n+1)^2(n+2)}$，所以 $\mathrm{E}(M-N)^2 = \frac{2N^2}{(n+1)(n+2)}$。這是 $\mathrm{Var}(Z_1) = \frac{N^2}{n(n+2)}$ 的 $\frac{2n}{n+1}$ 倍：較差。最佳的倍數是 $c = \frac{\mathrm{NE}(M)}{\mathrm{E}(M^2)} = \frac{n+2}{n+1}$，其均方誤差（mean squared error）為 $\frac{N^2}{(n+1)^2}$，比 Z_1 稍佳。

2. $\mathrm{E}(\min) = \frac{N}{n+1}$，而 min 的方差與 M 的相同，所以 $\mathrm{Var}\big((n+1)\min\big) = \frac{nN^2}{n+2}$。這個方差不會隨 n 增大而縮小：是個差劣的估計量（estimator）。

3. 最大值不超過 m，當且僅當全部 k 個數都來自 $\{1,\dots,m\}$。當 $k=1$ 時，估計量是 $2X-1$，其平均值為 $2\cdot\frac{N+1}{2}-1 = N$。

4. $98 + \frac{98}{5} - 1 = 116.6$：大約 117 輛坦克。

13.4 平均數還是中位數？

1. 先選定哪一個值是中位數（median）（有 $2n+1$ 種選法），再選定其餘各值中哪 n 個在它之下（有 $\binom{2n}{n}$ 種選法）。所求概率為 $(2n+1)\binom{2n}{n}y^n(1-y)^n\,\delta y = \frac{(2n+1)!}{(n!)^2}y^n(1-y)^n\,\delta y$。

2. $\mathrm{Var}(Y) = \frac{1}{4(2n+3)} = \frac{1}{20}$ 及 $\mathrm{Var}(\bar{X}) = \frac{1}{12(2n+1)} = \frac{1}{36}$。平均值較佳。

3. $\mathrm{Var}(\max) = \mathrm{Var}(\min) = \frac{m}{(m+1)^2(m+2)}$ 及 $\mathrm{Cov}(\max,\min) = \frac{1}{(m+1)^2(m+2)}$，由此得出 $\frac{1}{2(m+1)(m+2)}$。它縮小的速度與 $\frac{1}{m^2}$ 同階，遠快於平均值的 $\frac{1}{12m}$：對於均勻分佈（uniform distribution）的數據，資訊都在極端值之中，與坦克問題的情況一樣。

4. 平均值約為 18.6，被一個打字錯誤完全破壞；中位數是 10.1，不受影響。

13.5 問尷尬的問題

1. 設 $t = ap + bq$，即答案屬實的概率。回答「是」的概率為 $\frac{1}{2}t + \frac{1}{2}(1-t) = \frac{1}{2}$，所以 $\mathrm{P}(\text{正面} \mid \text{是}) = t$，而它等於 $\frac{1}{2}$，當且僅當 $2t = 1$。若由同一名孿生兒回答兩次，則不論硬幣擲出的是正面還是反面，先答「是」後答「否」的概率都是 $pa(1-a)+qb(1-b)$，所以後驗概率（posterior probability）是 $\frac{1}{2}$。兩次都答「是」的概率分別為 pa^2+qb^2（正面）和 $p(1-a)^2+q(1-b)^2$（反面），兩者相等當且僅當 $2(ap+bq)=1$。

2. $\mathrm{P}(\text{是}) = \frac{3}{4}\pi + \frac{1}{4}(1-\pi)$，所以 $\hat{\pi} = \frac{0.4-0.25}{0.5} = 0.3$。標準誤（standard error）為 $\frac{\sqrt{0.4\times0.6/400}}{0.5} \approx 0.049$。

3. $\mathrm{E}(Y^2) = \sum_c \mathrm{P}(C=c)\big(\mathrm{Var}(Y\mid c) + \mathrm{E}(Y\mid c)^2\big)$；減去 $\mathrm{E}(Y)^2 = \big(\sum_c \mathrm{P}(C=c)\mathrm{E}(Y\mid c)\big)^2$ 便得 $\mathrm{E}(\mathrm{Var}(Y\mid C)) + \mathrm{Var}(\mathrm{E}(Y\mid C))$。

4. Z_2 接近 $\frac{n}{6}$（正面）或 $\frac{5n}{6}$（反面），兩種情況的概率各為 $\frac{1}{2}$。所以 $\mathrm{P}(Z_2 > \frac{n}{2}) \to \frac{1}{2}$，儘管 Z_2 幾乎從不接近 $\frac{n}{2}$。

13.6 累積起來的誤差

1. 每個截尾誤差（truncation error）都在 $(-1, 0]$ 上均勻分佈；各加上 $\frac{1}{2}$ 後，便在 $(-\frac{1}{2}, \frac{1}{2}]$ 上均勻分佈，與四捨五入(rounding)的情況完全一樣。所以修正後的 VOZ 與 LEP 一樣好。

2. 每個誤差都在 $(-0.05, 0.05)$ 上均勻分佈，方差為 $\frac{0.1^2}{12}$，所以總和的標準差為 $\sqrt{10\,000 \times 0.01/12} \approx 2.9$。

3. $3000 \times 250 \times 0.0005 = 375$ 點。

4. 兩個誤差時，是 $[-1, 1]$ 上的 $1 - |x|$。三個誤差時：當 $|x| \le \frac{1}{2}$ 時為 $\frac{3}{4} - x^2$，當 $\frac{1}{2} \le |x| \le \frac{3}{2}$ 時為 $\frac{1}{2}\left(\frac{3}{2} - |x|\right)^2$，已經很接近正態曲線（normal curve）。

13.7 製造相關性

1. 對於標準化變量（standardised variables），$\mathrm{Var}(X_1 + X_2 + X_3) = 3 + 6\rho \ge 0$，所以 $\rho \ge -\frac{1}{2}$。對於 n 個變量，$n + n(n-1)\rho \ge 0$ 給出 $\rho \ge -\frac{1}{n-1}$。

2. 取 $b = \frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}}$，則 $(1-\rho_{12}^2)(1-\rho_{13}^2-b^2) = 1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23} = \det R$。由於 $1 - \rho_{12}^2 > 0$，這兩個條件等價。

3. $\mathrm{Var} = \frac{1}{4}(0.04 + 0.09 + 0.12\rho)$：當 $\rho = -1$、0、1 時，標準差分別為 0.05、0.18 和 0.25。除非兩隻股票完全相關（perfectly correlated），否則這個組合的風險低於兩者風險的平均值（0.25）。

4. $Z_2 = \dfrac{Y_2 - \rho_{12} Y_1}{\sqrt{1-\rho_{12}^2}}$。

Beyond STEP

STEP 是大多數中學生一生中應考過最艱深的數學考試，也是最出色的考試之一：題目簡短、陌生而深刻，當中最精彩的，其實是喬裝了的大學數學小品。

本書精選最有趣的 STEP 試題，由 1987 年首次開考之前的樣本試卷（specimen papers），到 2025 年的試卷，並以每一道題目作為一扇門。每道題目都會仔細解答，沿途提出優秀學生會問的問題，然後帶你穿過這扇門：這個想法從何而來，到了大學會演變成甚麼，以及接下來該讀些甚麼。

本書為正在準備 STEP、有天分又有志投身數學的學生而寫。他們想要的不止是模擬試卷：他們想開始像數學系本科生那樣思考，並一探前方有甚麼正等着自己。

- 共 13 章，涵蓋純數學（pure mathematics）、力學（mechanics）、概率（probability）與統計（statistics）
- 136 道 STEP 歷屆試題，每題均附引導式解答
- 每道題目背後的大學數學
- 附答案的延伸練習，以及延伸閱讀

全書共 13 章，本冊為第 13 章。 全書各章及所有詳細解答均可免費閱讀、下載及分享，網址為

beyondstep.keithmcnulty.org

文字內容已貢獻予公有領域（CC0 1.0）。STEP 試題 © Cambridge University Press & Assessment。