機率與統計：從入門到精通

第 6 章　不等式與極限定理

Inequalities and Limit Theorems

難度：核心

為什麼平均更多資料會得到更好的答案？為什麼世界上那麼多量的分布是鐘形的？為什麼可以用一小批樣本的平均去估計整個資料分布上的期望值——這正是 mini-batch 訓練的正當性來源？本章的**極限定理**回答這些問題。馬可夫、柴比雪夫與 Jensen 三個不等式讓我們只憑均值或變異數就能控制機率；**大數法則**保證樣本平均收斂到期望值；**中央極限定理**進一步指出漲落的形狀是高斯、尺度是 $\sigma/\sqrt{n}$。它們合起來是機率（模型）與統計（資料）之間的理論橋樑，也是機器學習之所以行得通的根本原因。

學習目標

讀完本章後，你將能夠：

- 證明並應用**馬可夫不等式**與**柴比雪夫不等式**，只憑均值或變異數界住尾機率，並理解「集中不等式」如何強化它們。
- 使用 **Jensen 不等式**判斷 $\varphi(\mathbb{E}[X])$ 與 $\mathbb{E}[\varphi(X)]$ 的大小關係，並認識它在 KL 散度與 ELBO 中的角色。
- 敘述**大數法則**，說明蒙地卡羅估計與 mini-batch 梯度估計為何可行。
- 敘述**中央極限定理**，用常態分布近似樣本平均，計算標準誤、規劃民調與實驗的樣本數。
- 區分**依機率收斂**、**依分布收斂**與**幾乎必然收斂**三種收斂模式。
- 用 Python 模擬驗證大數法則、中央極限定理與 $1/\sqrt{n}$ 的誤差縮放律。

6.1 機率不等式：馬可夫、柴比雪夫與 Jensen

本章的主角是極限定理，但在抵達「$n \to \infty$」之前，先介紹三個在**有限樣本**就能使用的不等式。它們的共同精神是：即使對分布幾乎一無所知——只知道均值，或再加上變異數——我們仍然能對機率下**保證**。這種「用最少的資訊換一個界」的思想，待會兒會直接用來證明大數法則，也是統計學習理論的起手式。

命題 6.1　馬可夫不等式

設 X 為非負隨機變數（$\mathbb{P}(X \geq 0) = 1$）且 $\mathbb{E}[X]$ 有限。則對任意 $a > 0$，

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

此即**馬可夫不等式**（Markov's inequality）。

證明　關鍵是一條逐點成立的不等式：

$$a \cdot \mathbf{1}\{X \geq a\} \leq X,$$

其中 $\mathbf{1}\{A\}$ 是指示變數（indicator）：事件 A 發生時取 1、否則取 0。驗證：若 $X \geq a$，左邊等於 a，不超過右邊；若 $X < a$，左邊為 0，而右邊非負。兩邊取期望（期望保持不等式），並利用 $\mathbb{E}[\mathbf{1}\{X \geq a\}] = \mathbb{P}(X \geq a)$：

$$a \cdot \mathbb{P}(X \geq a) \leq \mathbb{E}[X],$$

除以 a 即得。 ■

直觀地說：若「$X \geq a$」的機率超過 $\mathbb{E}[X]/a$，那麼光是這一塊質量對期望的貢獻就超過 $a \times \mathbb{E}[X]/a = \mathbb{E}[X]$——與總期望只有 $\mathbb{E}[X]$ 矛盾。均值就這麼多，「高處」能放的機率自然有限。這個界一般很鬆，卻已是「只用均值」所能得到的最佳結果：兩點分布 $\mathbb{P}(X = a) = \mu/a$、$\mathbb{P}(X = 0) = 1 - \mu/a$（期望恰為 μ）使等號成立——質量全押在門檻上，期望被最有效率地拿來製造尾機率。想要更緊的界，必須投入更多資訊（變異數、有界性、指數矩），這正是接下來要做的事。

例 6.1　收入的馬可夫界

某地區居民年收入非負、平均為 5 萬元，除此之外一無所知。年收入至少 25 萬元的人口比例最多是多少？至少 50 萬元呢？

解　以萬元為單位，$X \geq 0$、$\mathbb{E}[X] = 5$。由馬可夫不等式

$$\mathbb{P}(X \geq 25) \leq \frac{5}{25} = \frac{1}{5} = 20\%, \qquad \mathbb{P}(X \geq 50) \leq \frac{5}{50} = 10\%.$$

「收入達到平均的 k 倍」的比例永遠不超過 $1/k$——不需要知道任何分布形狀。實際比例可能遠低於界（例如不到 1%），但沒有任何非負分布能違反它。也要注意界只在 $a > \mathbb{E}[X]$ 時有意義：取 $a = 4$ 得 $\mathbb{P}(X \geq 4) \leq 1.25$，毫無資訊。

只用均值，管不住「離均值多遠」；再加一個變異數，就能同時控制兩側的尾巴。

命題 6.2　柴比雪夫不等式

設 X 的均值 μ 與變異數 σ^2 皆有限。則對任意 $k > 0$，

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2};$$

等價地，對任意 $\varepsilon > 0$，

$$\mathbb{P}(|X-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

此即**柴比雪夫不等式**（Chebyshev's inequality）。

證明　兩邊平方（皆非負）可知事件相等：

$$\{|X-\mu| \geq k\sigma\} = \{(X-\mu)^2 \geq k^2\sigma^2\}.$$

而 $(X-\mu)^2$ 是**非負**隨機變數，期望值正是 $\mathbb{E}[(X-\mu)^2] = \sigma^2$。對它套用馬可夫不等式（門檻 $a = k^2\sigma^2$）：

$$\mathbb{P}\big((X-\mu)^2 \geq k^2\sigma^2\big) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

取 $\varepsilon = k\sigma$ 即得第二形式。 ■

這個證明示範了一個可推廣的套路：對 $|X-\mu|$ 套上一個遞增的非負函數（這裡是平方），再用馬可夫。換別的函數就得到別的不等式——四次方給出衰減更快的界（練習 6.6），指數函數則通往 Chernoff 與 Hoeffding 這一族更強的界。

例 6.2　柴比雪夫的保證 vs. 常態的實際

不論 X 是什麼分布（只要 σ^2 有限），柴比雪夫保證

$$\mathbb{P}(|X-\mu| < 2\sigma) \geq 1 - \frac{1}{4} = 75\%, \qquad \mathbb{P}(|X-\mu| < 3\sigma) \geq 1 - \frac{1}{9} \approx 88.9\%.$$

把它與「X 恰好是常態分布」的實際值並排：

k	柴比雪夫下界（$k\sigma$ 內）	常態分布實際值
1	0%（無資訊）	68.3%
1.5	55.6%	86.6%
2	75%	95.4%
3	88.9%	99.73%
4	93.75%	99.994%

解　差距的來源：柴比雪夫替「所有分布中最壞的那一個」買保險，所以對任何特定的、行為良好的分布必然偏鬆。以 $k=3$ 為例，上界 $1/9 \approx 0.111$ 是常態實際尾機率 0.0027 的約 41 倍（圖 6.1）。但「不管什麼分布都成立」正是它的價值：面對來路不明的資料，它是唯一敢開口的保證。

直觀理解

馬可夫與柴比雪夫的魅力在於**幾乎不需要假設**：不必知道分布的形狀，只要一個均值（或再加一個變異數），就能對尾機率下保證。柴比雪夫說：任何分布放在 k 個標準差之外的質

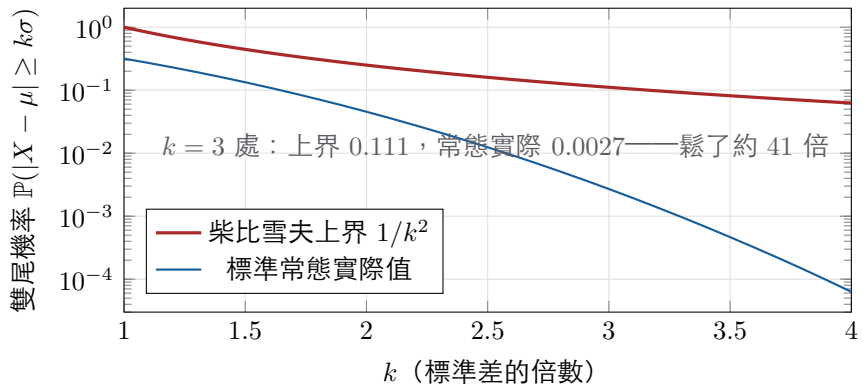


圖 6.1: 柴比雪夫上界與標準常態實際尾機率的對照（縱軸為對數座標）。上界以多項式速度 $1/k^2$ 衰減，常態尾巴則以指數速度崩落；界雖鬆，卻對任何分布都成立。

量都不可能超過 $1/k^2$——不管什麼形狀，超過 2σ 的最多 25%。代價是界通常很鬆（常態實際只有 4.6%）。對**有界**隨機變數的平均，更銳利的**集中不等式**（concentration inequality）如 Hoeffding 不等式

$$\mathbb{P}\big(|\bar{X}_n-\mu|\geq\varepsilon\big)\leq 2\exp\!\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right),\qquad X_i\in[a,b],$$

把多項式衰減 $1/k^2$ 加強為**指數**衰減。這一族不等式是學習理論中**泛化界**（generalization bound）的骨幹：界住「訓練誤差與真實誤差的偏離」隨樣本數消失的速度。練習 6.6 (c) 給出它們共同的出發點。

第三個不等式換一個舞台：它比較的不是尾機率，而是「先取期望再套函數」與「先套函數再取期望」誰大誰小。回顧：函數 φ 稱為**凸函數**（convex function），若其圖形上任兩點的連線都位於圖形上方；二階可微時等價於 $\varphi''(x)\geq 0$。不等號反向者稱為**凹函數**（concave function）。例如 x^2、x^4、e^x、$|x|$、$1/x$（$x>0$）都是凸函數；$\log x$、$\sqrt{x}$（$x>0$）都是凹函數。

命題 6.3　Jensen 不等式

設 φ 為凸函數，且 $\mathbb{E}[X]$ 與 $\mathbb{E}[\varphi(X)]$ 都存在。則

$$\varphi\big(\mathbb{E}[X]\big)\leq\mathbb{E}\big[\varphi(X)\big].$$

若 φ 為凹函數，不等號反向；若 φ 嚴格凸（strictly convex）且 X 不是常數，則不等式嚴格成立。此即 **Jensen 不等式**（Jensen's inequality）。

證明　凸函數在任一點 m 都有一條**支撐線**（supporting line）：存在斜率 c 使得

$$\varphi(x)\geq\varphi(m)+c\,(x-m)\qquad\text{對所有}x$$

（φ 在 m 可微時取 $c=\varphi'(m)$，支撐線就是切線；凸性正是「圖形位於每條切線上方」）。取

$m=\mathbb{E}[X]$，把上式中的 x 換成隨機變數 X，兩邊取期望：

$$\mathbb{E}[\varphi(X)] \geq \varphi(m)+c\,\big(\mathbb{E}[X]-m\big)=\varphi\big(\mathbb{E}[X]\big).$$

凹函數的情形對 $-\varphi$ 套用即可。 ■

一個好記的口訣：**凸函數之下，「先平均」小於「後平均」**。直觀原因：平均會削平波動，而凸函數在兩端翹起、獎勵波動，所以先套函數再平均得到的值比較大。

例 6.3　Jensen 的三個經典應用

判斷下列各式的不等號方向，並說明理由。

(a) $\mathbb{E}[X^2]$ 與 $(\mathbb{E}[X])^2$；　(b) $\mathbb{E}[\log X]$ 與 $\log\mathbb{E}[X]$（$X>0$）；　(c) 算術平均與幾何平均。

解　(a) $\varphi(x)=x^2$ 是凸函數，故

$$\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2.$$

移項就是 $\mathrm{Var}(X)=\mathbb{E}[X^2]-(\mathbb{E}[X])^2\geq 0$：「變異數非負」其實是 Jensen 不等式的特例。

(b) log 是凹函數，方向反轉：

$$\mathbb{E}[\log X] \leq \log\mathbb{E}[X].$$

數值檢查：X 等機率取 1 或 4，則 $\mathbb{E}[\log X]=\frac{0+\log 4}{2}=\log 2\approx 0.693$，而 $\log\mathbb{E}[X]=\log 2.5\approx 0.916$，方向正確。

(c) 設 $a_1,\dots,a_n>0$，令 X 均勻地取值於 $\{a_1,\dots,a_n\}$。由 (b)：

$$\log\Big(\prod_i a_i\Big)^{1/n}=\frac{1}{n}\sum_i \log a_i=\mathbb{E}[\log X]\leq\log\mathbb{E}[X]=\log\Big(\frac{1}{n}\sum_i a_i\Big),$$

兩邊取指數即得**算幾不等式**：幾何平均 $\leq$ 算術平均，等號在所有 a_i 相等時成立。

Jensen 不等式在機器學習中的分量遠超它樸素的外表：**KL 散度的非負性** $\mathrm{KL}(p\,\|\,q)\geq 0$ 是 Jensen 對凹函數 log 的直接應用（練習 6.11），而變分推論的 **ELBO**（證據下界）正是用 Jensen 把難算的 $\log p(x)$ 換成可優化的下界——那是第 12 章的主戲，入場券在本章發放。

6.2　大數法則：為什麼平均會收斂

「多量幾次、取平均」是比機率論更古老的直覺：天文學家重複測量星體位置、工匠重複量同一段尺寸，再取平均以抵銷誤差。大數法則把這個直覺升級成定理。

設 $X_1,X_2,\dots \overset{\text{i.i.d.}}{\sim}$ 某分布，均值 μ、變異數 σ^2，樣本平均為

$$\bar{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i.$$

兩個立即的計算（期望的線性；獨立變數的變異數相加，見第 5 章）：

$$\mathbb{E}[\bar{X}_n]=\mu,\qquad \mathrm{Var}(\bar{X}_n)=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i)=\frac{\sigma^2}{n}.$$

第一式說 $\bar{X}_n$「瞄得準」（無偏），第二式說它「越縮越緊」：變異數以 $1/n$ 的速度消失。第二式的平方根重要到值得一個名字。

定義 6.4　標準誤

樣本平均 $\bar{X}_n$ 的標準差稱為**標準誤**（standard error）：

$$\mathrm{SE}(\bar{X}_n) = \sqrt{\mathrm{Var}(\bar{X}_n)} = \frac{\sigma}{\sqrt{n}}.$$

實務中 σ 未知，以樣本標準差 s 代入得估計值 $\widehat{\mathrm{SE}} = s/\sqrt{n}$。標準誤回答的問題是：「用 $\bar{X}_n$ 估計 μ 時，典型誤差有多大？」

注意分母是 $\sqrt{n}$ 而不是 n：樣本翻 4 倍，誤差才減半；想把誤差降到十分之一，樣本要翻 100 倍（圖 6.2）。

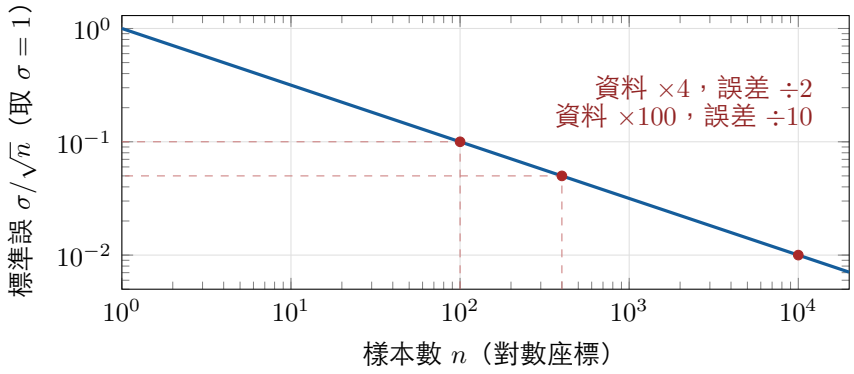


圖 6.2: 標準誤 $\sigma/\sqrt{n}$ 在雙對數座標下是一條斜率 $-1/2$ 的直線。想多換一位小數的精度，就要付一百倍的資料——統計學的基本匯率。

變異數縮小到零，直觀上 $\bar{X}_n$ 就被「釘」在 μ 附近。把這句話說嚴謹，就是大數法則。

定理 6.5　大數法則

設 $X_1, X_2, \ldots$ 為 i.i.d.（獨立且同分布）隨機變數，期望值 $\mu = \mathbb{E}[X_i]$ 存在且有限。則對任意 $\varepsilon > 0$，

$$\lim_{n\to\infty} \mathbb{P}\big(|\bar{X}_n - \mu| \geq \varepsilon\big) = 0,$$

即樣本平均**依機率收斂**到 μ，記作 $\bar{X}_n \xrightarrow{p} \mu$。此為**（弱）大數法則**（law of large numbers, LLN）。

當變異數有限時，證明只是柴比雪夫不等式的一行應用，而且還附贈收斂速率——把 $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$ 代進柴比雪夫即可，完整推導留作練習 6.10。一般情形（只假設期望存在）結論不變，證明較深。6.4 節還會提到更強的「幾乎必然」版本：連單一條樣本路徑本身都以機率 1 收斂。

核心觀念

大數法則是統計學的**基石保證**：更多資料的平均，收斂到真值。它讓第 1 章的頻率詮釋自洽——「長期頻率趨近機率」正是大數法則套在指示變數上的特例：取 $X_i = \mathbf{1}\{\text{第}i\text{次試驗中}A\text{發生}\}$，則 $\bar{X}_n$ 是發生頻率，而 $\mathbb{E}[X_i] = \mathbb{P}(A)$，故頻率 $\xrightarrow{p} \mathbb{P}(A)$。它也授權**蒙地卡羅估計**：想算的期望值沒有解析解？抽樣、取平均就對了。每一次你用一批樣本的平均去估計損失、梯度或積分，背後都是大數法則在撐腰。

大數法則最重要的應用模式，叫做**蒙地卡羅方法**（Monte Carlo method）：想計算的量若能寫成期望值，就用樣本平均近似——

$$\mathbb{E}[g(X)] \approx \frac{1}{N}\sum_{i=1}^{N} g(x_i),$$

其中 $x_1, \dots, x_N$ 是依 X 的分布抽出的獨立樣本。任何積分都能改寫成期望（$\int g(x)f(x)\,\mathrm{d}x = \mathbb{E}_f[g(X)]$），所以蒙地卡羅本質上是一台「用亂數算積分」的機器。它的標準誤以 $1/\sqrt{N}$ 縮小，而且——關鍵——**與維度無關**：積分是 1 維還是 1000 維都一樣。這讓它成為高維積分幾乎唯一可行的辦法，也因此貫穿貝氏計算與深度學習。

例 6.4　用亂數擲飛鏢估 π

在單位正方形 $[0,1]^2$ 內均勻撒 N 個點，落在四分之一圓 $x^2 + y^2 \le 1$ 內的比例應接近其面積 $\pi/4$。這能把 π「量」出來嗎？精度如何？

解　設第 i 個點為 (u_i, v_i)，命中指示變數 $H_i = \mathbf{1}\{u_i^2 + v_i^2 \le 1\}$，則 $H_1, \dots, H_N \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$，$p = \pi/4$。由大數法則，估計量

$$\hat{\pi} = 4\bar{H}_N = \frac{4}{N}\sum_{i=1}^{N} H_i \xrightarrow{p} 4p = \pi,$$

且 $\mathbb{E}[\hat{\pi}] = \pi$（無偏）。精度由標準誤決定：$\mathrm{Var}(H_i) = p(1-p) \approx 0.1685$，故

$$\mathrm{SE}(\hat{\pi}) = 4\sqrt{\frac{p(1-p)}{N}} \approx \frac{1.642}{\sqrt{N}}.$$

$N = 10^4$ 時 $\mathrm{SE} \approx 0.016$——大約只有小數點後兩位可信。想把 SE 壓到 0.005，需要

$$N \ge \left(\frac{1.642}{0.005}\right)^2 \approx 1.08 \times 10^5.$$

下一節的中央極限定理會再補一句：這個誤差近似常態，因此 $\hat{\pi} \pm 2\,\mathrm{SE}$ 大約就是 95% 的信賴範圍。

6.3　中央極限定理：為什麼世界是鐘形的

大數法則說樣本平均「會收斂」；**中央極限定理**（central limit theorem, CLT）說「怎麼收斂」——漲落的形狀是什麼、尺度多大——並且順手解答了一個古老謎題：為什麼世界上到處都是鐘

形曲線。

定理 6.6　中央極限定理

設 $X_1, X_2, \dots$ 為 i.i.d. 隨機變數，$\mathbb{E}[X_i] = \mu$、$\mathrm{Var}(X_i) = \sigma^2$，且 $0 < \sigma^2 < \infty$。令 $S_n = X_1 + \cdots + X_n$。則標準化樣本平均

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

依分布收斂到標準常態：對每個 $z \in \mathbb{R}$，

$$\lim_{n\to\infty} \mathbb{P}(Z_n \le z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,\mathrm{d}t.$$

等價的實用讀法：n 夠大時，

$$\bar{X}_n \overset{\text{近似}}{\sim} N\!\left(\mu, \frac{\sigma^2}{n}\right), \qquad S_n \overset{\text{近似}}{\sim} N\!\left(n\mu, n\sigma^2\right).$$

定理最不可思議之處在於**結論與 X_i 的分布無關**：擲硬幣的 0/1、右偏的指數等待時間、離散的骰子點數……只要 i.i.d. 且變異數有限，標準化平均一律流向同一個 $N(0,1)$（證明需要特徵函數工具，超出本書範圍；圖 6.3 展示收斂實況）。實務上「n 多大才夠」取決於原分布的偏斜程度：經驗法則 $n \gtrsim 30$ 通常安全，偏斜或重尾越重，需要的 n 越大。

直觀理解

中央極限定理是數學中最驚人的事實之一：把許多獨立隨機效應**加總**，結果就是鐘形——不管個別效應長什麼樣（均勻、偏斜、離散……只要變異數有限）。這解釋了鐘形曲線在自然界的霸權：測量誤差是無數小干擾之和、身高是眾多基因與環境因素之和、樣本平均是 n 個觀測之和除以 n——所以它們都近似常態。另一個必須內化的細節是 $\sqrt{n}$：平均的標準誤以 $1/\sqrt{n}$ 縮小，想把誤差**減半**需要 4 倍資料，想多一位小數的精度需要 100 倍資料——這是統計的基本（有時很痛的）匯率。

CLT 讓我們能替「用平均做的估計」標上精度，代表性應用是樣本數規劃：

例 6.5　民調要問多少人？

想估計候選人的支持率 p（未知），隨機訪問 n 人，以樣本比例 $\hat{p} = \bar{X}_n$（$X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(p)$）作估計。要求 95% 信心下誤差不超過 ± 3 個百分點，n 至少要多大？

解　由 CLT，$\hat{p}$ 近似 $N\!\left(p, \frac{p(1-p)}{n}\right)$。標準常態有 95% 的機率落在 ± 1.96 之內，故 95% 誤差界（margin of error）為

$$m = 1.96\sqrt{\frac{p(1-p)}{n}}.$$

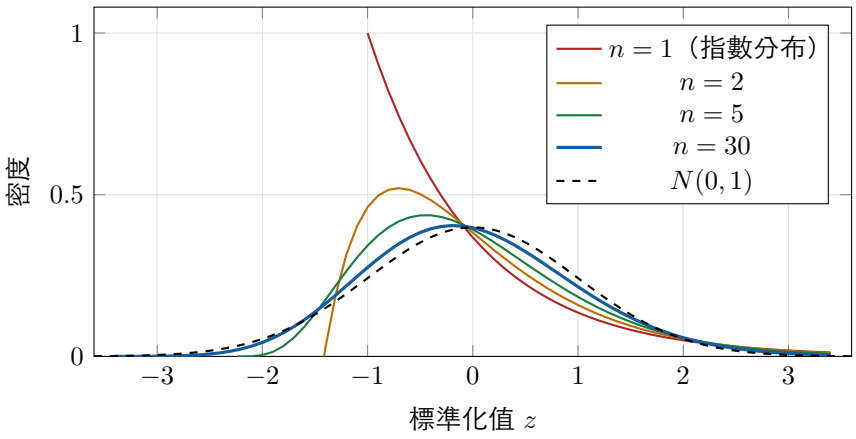


圖 6.3: 中央極限定理實況：n 個獨立 Exp(1) 之和（即 $\Gamma(n,1)$ 分布）標準化後的密度。$n=1$ 嚴重右偏且在左端點跳躍；$n=5$ 已見鐘形雛形；$n=30$ 與標準常態（黑色虛線）幾乎重合。

p 未知，但二次式 $p(1-p)$ 在 $p=\frac{1}{2}$ 時最大、值為 $\frac{1}{4}$，因此無論真實支持率為何，

$$m \le \frac{1.96}{2\sqrt{n}}.$$

令其 ≤ 0.03：$\sqrt{n} \ge \frac{1.96}{0.06} \approx 32.67$，得 $n \ge 1067.1$，取 $n = 1068$。這正是台灣新聞稿標準句式「有效樣本 1,068 份，在 95% 信心水準下抽樣誤差為 ±3.0 個百分點」的出處。值得再看一眼的重點：所需樣本數**與母體大小無關**——母體是兩千三百萬人還是三億人，精度都只由 n 決定。

CLT 也常用來近似**離散**分布。若 $X \sim \mathrm{Bin}(n,p)$（第 4 章），則 X 是 n 個 Bernoulli 變數之和，故 X 近似 $N\big(np,\, np(1-p)\big)$。因為 X 只取整數值，把整數 k「攤」成區間 $[k-0.5,\, k+0.5]$ 的**連續性校正**（continuity correction）能明顯提高精度：

$$\mathbb{P}(a \le X \le b) \approx \Phi\left(\frac{b+0.5-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-0.5-np}{\sqrt{np(1-p)}}\right).$$

練習 6.7 會看到：校正把近似誤差從第二位小數壓到第四位。

常見陷阱

中央極限定理需要有限變異數。Cauchy 分布 $f(x) = \frac{1}{\pi(1+x^2)}$ 的尾巴重到連期望值都不存在：n 個 i.i.d. Cauchy 的樣本平均**仍然是同一個 Cauchy 分布**——平均一百萬筆和只看一筆一樣糟，大數法則與中央極限定理雙雙失效。實務教訓：重尾資料（金融報酬、社群網路的度數、極端災損）取平均前務必先檢查尾巴，必要時改用中位數等穩健統計量。第二個常見錯誤是把 $1/\sqrt{n}$ 記成 $1/n$：誤差棒不會因為資料翻倍就減半——只縮小 $\sqrt{2} \approx 1.41$ 倍。規劃樣本數時請用平方律：精度要求 $\times k$，樣本數 $\times k^2$。

6.4 收斂模式簡述

本章已經出現兩種「收斂」：大數法則的收斂與中央極限定理的收斂其實是不同的概念。機率學家區分三種主要模式——區別對證明很重要，對使用者而言，知道地圖長什麼樣子即可。

定義 6.7　三種收斂模式

設 $X_1, X_2, \dots$ 與 X 為隨機變數。

1. **依機率收斂**（convergence in probability），記作 $X_n \xrightarrow{p} X$：對任意 $\varepsilon > 0$，$\mathbb{P}(|X_n - X| \geq \varepsilon) \to 0$——大偏差的機率消失。
2. **幾乎必然收斂**（almost sure convergence），記作 $X_n \xrightarrow{\text{a.s.}} X$：$\mathbb{P}\big(\lim_{n\to\infty} X_n = X\big) = 1$——除了機率 0 的例外，每一條樣本路徑本身都收斂。
3. **依分布收斂**（convergence in distribution），記作 $X_n \xrightarrow{d} X$：在 F_X 的每個連續點上 $F_{X_n}(x) \to F_X(x)$——分布的形狀逼近極限分布。

三者由強到弱構成蘊涵鏈：

$$\text{幾乎必然收斂} \implies \text{依機率收斂} \implies \text{依分布收斂},$$

反向一般不成立（唯一例外：極限是常數時，依分布收斂可升級為依機率收斂）。三大定理各據一格：**強大數法則**（strong law of large numbers, Kolmogorov）說只要 $\mathbb{E}|X_1| < \infty$，就有 $\bar{X}_n \xrightarrow{\text{a.s.}} \mu$；弱大數法則是依機率收斂；中央極限定理是依分布收斂。用擲硬幣感受強弱之別：弱版本說「正面比例大幅偏離 1/2 的機率越來越小」，強版本說「你手上這一條無窮拋擲路徑的正面比例，以機率 1 收斂到 1/2」。

核心觀念

三種收斂模式的精細區別對「證明定理」很重要；對「使用定理」的我們，需要記住的圖像只有一幅：**樣本平均收斂到期望值（大數法則），其漲落是尺度為 $\sigma/\sqrt{n}$ 的近似高斯（中央極限定理）**。前者告訴你估計終將正確，後者告訴你此刻的誤差多大、長什麼形狀。

這如何驅動機器學習／ AI

極限定理是日常機器學習的**沉默正當化**。**Mini-batch SGD**：我們真正想要的梯度是整個資料分布上的期望 $\mathbb{E}[\nabla_\theta \ell(\theta; X)]$，實際計算的是一小批（mini-batch）樣本的平均梯度——大數法則保證這個替換合法（批次平均是真梯度的無偏估計，且隨批次放大而收斂），中央極限定理進一步給出梯度雜訊的量級 $\propto 1/\sqrt{B}$（B 為批次大小）——這正是大批次訓練「越大越穩、但報酬遞減」的原因。**蒙地卡羅估計**：VAE 的重參數化技巧、dropout 當作近似貝氏推論、強化學習的期望報酬估計，全都是「用樣本平均近似期望」，收斂由大數法則背書。**誤差棒與信賴區間**：論文表格裡準確率旁邊的 $\pm$ 值，多半就是 CLT 的 $\hat{\sigma}/\sqrt{n}$（第 8 章）。**泛化界**：Hoeffding、Bernstein 等集中不等式界住「訓練誤差與真實誤差的偏離」，是統計學習理論的骨幹。最後，CLT 還解釋了為什麼「假設雜訊是高斯」經常安全：被平均、被加總過的量，依法則本來就長成鐘形。

例 6.6　mini-batch 的雜訊帳本

把訓練集視為資料分布，真梯度為 $g^* = \mathbb{E}[\nabla_\theta \ell]$。已知某一個參數座標上，單一樣本梯度的標準差為 $\sigma_g = 4.0$。批次梯度取 B 個樣本的平均：$\hat{g}_B = \frac{1}{B}\sum_{i=1}^{B} \nabla_\theta \ell_i$。(a) $B = 32$ 與 $B = 128$ 時，$\hat{g}_B$ 在該座標的標準誤各是多少？(b) 想把 $B = 32$ 的雜訊壓到四分之一，批次要多大？

解　(a) 批次平均正是樣本平均，故 $\mathrm{SE}(\hat{g}_B) = \sigma_g/\sqrt{B}$：

$$B = 32: \ \frac{4}{\sqrt{32}} \approx 0.707, \qquad B = 128: \ \frac{4}{\sqrt{128}} \approx 0.354.$$

批次 $\times 4$，雜訊恰好 $\div 2$。(b) 雜訊 $\div 4$ 需要批次 $\times 16$：$B = 32 \times 16 = 512$。

討論：這就是「大批次讓梯度更平滑、但報酬遞減」的定量內容——每一步的計算成本隨 B **線性**成長，雜訊卻只以 $1/\sqrt{B}$ 下降。CLT 還告訴我們 B 夠大時 $\hat{g}_B$ 近似常態，許多把 SGD 雜訊建模成高斯的理論分析正是以此為據。

6.5　Python 實作

理論之後，動手驗證：大數法則的收斂、中央極限定理的 $1/\sqrt{n}$ 縮放與常態形狀、柴比雪夫的普遍性，以及蒙地卡羅估 π。

```python
import numpy as np

rng = np.random.default_rng(42)

# --- 1. 大數法則：Exp(1) 的樣本平均隨 n 增大而逼近真均值 1 ---
for n in [10, 100, 10_000, 1_000_000]:
    x = rng.exponential(1.0, size=n)
    print(n, round(x.mean(), 4))          # 逐步逼近 1.0

# --- 2. CLT 的 1/sqrt(n) 縮放：n 個 Exp(1) 之平均的標準差 ---
M = 10_000                                # 每個 n 重複 M 次
for n in [1, 10, 100, 1000]:
    means = rng.exponential(1.0, size=(M, n)).mean(axis=1)
    print(n, round(means.std(), 4))       # 約 1, 0.316, 0.1, 0.0316

# --- 3. 標準化樣本平均近似 N(0,1)：檢查 95% 覆蓋率 ---
n = 1000
means = rng.exponential(1.0, size=(M, n)).mean(axis=1)
z = (means - 1.0) * np.sqrt(n)            # sigma = 1
print(round(np.mean(np.abs(z) <= 1.96), 4))   # 約 0.95

# --- 4. 蒙地卡羅估 pi：誤差以 1/sqrt(N) 縮小 ---
```

```python
for N in [100, 10_000, 1_000_000]:
    u, v = rng.uniform(size=N), rng.uniform(size=N)
    pi_hat = 4 * np.mean(u**2 + v**2 <= 1)
    print(N, round(pi_hat, 4))            # 標準誤約 1.64/sqrt(N)
```

第 2 段的輸出正是理論值 $1/\sqrt{n}$ 的數列 1, 0.316, 0.1, 0.0316：樣本平均的漲落以 $1/\sqrt{n}$ 縮小。第 3 段驗證形狀：雖然 Exp(1) 嚴重右偏，1000 個的平均標準化後落在 ± 1.96 內的比例已非常接近 95%。第 4 段的估 π 誤差依 $1.64/\sqrt{N}$ 遞減：N 每乘 100，多換到一位小數。讀者可自行加一段驗證柴比雪夫的普遍性：對常態、指數、均勻樣本各算「距均值 2 個標準差以外」的比例，會得到約 4.6%、5.0%、0%——全部乖乖躺在上界 25% 之下。

本章總結

- **馬可夫不等式**：非負 X、$a>0$ 時 $\mathbb{P}(X \ge a) \le \mathbb{E}[X]/a$——只用均值就界住尾巴。**柴比雪夫不等式**（馬可夫套在 $(X-\mu)^2$ 上）：$\mathbb{P}(|X-\mu| \ge k\sigma) \le 1/k^2$——任何分布放在 2σ 之外的質量最多 25%。
- **集中不等式**（Hoeffding、Bernstein）對有界變數的平均把 $1/k^2$ 加強為指數衰減，是統計學習理論中**泛化界**的骨幹：界住訓練誤差與真實誤差的偏離。
- **Jensen 不等式**：凸 φ 時 $\varphi(\mathbb{E}X) \le \mathbb{E}[\varphi(X)]$（凹者反向）。變異數非負、算幾不等式、$\mathrm{KL} \ge 0$ 與 ELBO 全是它的特例。
- **大數法則**：$\bar{X}_n$ 依機率（甚至幾乎必然）收斂到 μ——蒙地卡羅估計與 mini-batch 梯度估計的理論基礎。
- **中央極限定理**：$(\bar{X}_n-\mu)/(\sigma/\sqrt{n})$ 依分布收斂到 $N(0,1)$；n 大時 $\bar{X}_n \approx N(\mu, \sigma^2/n)$。**標準誤** $\sigma/\sqrt{n}$：誤差減半需要 4 倍資料，精度多一位小數需要 100 倍資料。
- 收斂模式地圖：幾乎必然收斂 $\Rightarrow$ 依機率收斂 $\Rightarrow$ 依分布收斂；強大數法則、弱大數法則與 CLT 各據一格。CLT 需要**有限變異數**——Cauchy 分布的樣本平均永不收斂。
- 合起來，極限定理正當化了：SGD 的批次梯度（雜訊 $\propto 1/\sqrt{B}$）、蒙地卡羅目標（重參數化、RL 報酬）、以及指標的信賴區間與誤差棒。

6.6 練習題

建議先獨立完成再對照詳解。題目依難度標示為**基礎**、**進階**、**挑戰**三級；標「程式」者需要動手寫程式。

練習 6.1 (基礎)　某線上服務的平均回應時間為 200 ms（回應時間非負，分布未知）。(a) 用馬可夫不等式界住「回應超過 1000 ms」的機率。(b) 界住「回應超過 500 ms」的機率。(c) 對「回應超過 100 ms」馬可夫不等式給出什麼？這說明馬可夫界什麼時候才有用？

練習 6.2 (基礎)　某測驗分數 X 的均值為 100、標準差為 15，分布未知。(a) 用柴比雪夫不等式給「分數落在 70 到 130 之間」機率的下界。(b) 給「分數與 100 相差至少 45」機率的上界。(c) 若事後得知 X 近似常態分布，(a)、(b) 對應的實際機率各是多少？柴比雪夫的界各鬆了幾倍？

練習 6.3 (基礎)　用 Jensen 不等式判斷下列每一組的大小關係（或說明無法判斷），設所有期望值都存在：(a) $\mathbb{E}[e^X]$ 與 $e^{\mathbb{E}[X]}$；(b) $\mathbb{E}[\sqrt{X}]$ 與 $\sqrt{\mathbb{E}[X]}$（$X \ge 0$）；(c) $\mathbb{E}[1/X]$ 與 $1/\mathbb{E}[X]$（$X > 0$）；(d) $\mathbb{E}[X^4]$ 與 $(\mathbb{E}[X])^4$；(e) $\mathbb{E}[|X|]$ 與 $|\mathbb{E}[X]|$；(f) $\mathbb{E}[X^3]$ 與 $(\mathbb{E}[X])^3$（X 可取任何實數）。

練習 6.4 (基礎)　標準誤的匯率：(a) 想把樣本平均的標準誤降到目前的十分之一，資料量要變成幾倍？(b) 樣本數從 50 增加到 200，標準誤如何變化？(c) 某估計在 $n = 400$ 時標準誤為 0.06；要把標準誤降到 0.02，總共需要多少樣本？

練習 6.5 (進階)　感測器單次讀值的標準差為 $\sigma = 2$（單位 °C），對真實溫度 μ 無偏。我們想讓樣本平均與 μ 相差至少 0.5 °C 的機率不超過 5%。(a) 只用柴比雪夫不等式，n 至少要多大？(b) 改用中央極限定理近似，n 至少要多大？(c) 比較兩個答案並解釋差異的來源。

練習 6.6 (進階)　(a) 不看正文，由馬可夫不等式推導柴比雪夫不等式。(b) 同樣的手法證明四階矩版本：

$$\mathbb{P}(|X - \mu| \ge \varepsilon) \le \frac{\mathbb{E}[(X-\mu)^4]}{\varepsilon^4}.$$

(c) 敘述這個手法的一般模式：對任意嚴格遞增的非負函數 g，$\mathbb{P}(|Z| \ge t)$ 可以被什麼量界住？取 $g(x) = e^{\lambda x}$（$\lambda > 0$）會得到什麼形式的界？

練習 6.7 (進階)　擲一枚公正硬幣 100 次，令 X 為正面次數。(a) 用中央極限定理（不做連續性校正）近似 $\mathbb{P}(45 \le X \le 55)$。(b) 加上連續性校正重算一次。(c) 已知精確值為 0.7287（電腦計算），比較兩個近似的誤差。

練習 6.8 (進階)　(a) 某民調訪問 $n = 1200$ 人，得到支持率 $\hat{p} = 0.48$。求 95% 信賴區間，並以「正負幾個百分點」敘述。由這份民調能否排除「真實支持率過半」？(b) 想把 95% 誤差界控制在 ± 2 個百分點以內（對任何 p 都成立），需要多少樣本？(c) ± 1 個百分點呢？由 (b)(c) 總結「誤差減半」的代價。

練習 6.9 (進階)　想用蒙地卡羅法估計 $I = \int_0^1 e^{-x^2}\,\mathrm{d}x$（被積函數沒有初等反導函數）。(a) 說明如何把 I 寫成期望值、給出估計量，並解釋為什麼大數法則保證它收斂。(b) 先證明：取值於 $[0,1]$ 的隨機變數，其變異數必 $\le 1/4$。據此求出保證標準誤 ≤ 0.001 的樣本數 N。

(c) 已知 $\mathbb{E}[g(U)] \approx 0.7468$、$\mathbb{E}[g(U)^2] = \int_0^1 e^{-2x^2}\,\mathrm{d}x \approx 0.5981$，求實際變異數與實際所需的 N，並與 (b) 的保守界比較。

練習 6.10 (挑戰)　弱大數法則的完整證明。設 $X_1, \dots, X_n$ i.i.d.，均值 μ、變異數 $\sigma^2 < \infty$。(a) 證明 $\mathbb{E}[\bar X_n] = \mu$ 與 $\mathrm{Var}(\bar X_n) = \sigma^2/n$，並指出每一步用到什麼性質。(b) 用柴比雪夫不等式證明：對任意 $\varepsilon > 0$，

$$\mathbb{P}\big(|\bar X_n - \mu| \ge \varepsilon\big) \le \frac{\sigma^2}{n\varepsilon^2},$$

並下結論 $\bar X_n \xrightarrow{p} \mu$。(c) 設 $\sigma = 10$、$\varepsilon = 1$，要讓 (b) 的上界不超過 0.01，n 至少要多大？

練習 6.11 (挑戰)　Jensen 不等式的兩個機器學習推論。(a) 證明 KL 散度非負：對同一可數集合上的機率分布 p, q（$p_i > 0$ 時 $q_i > 0$），

$$\mathrm{KL}(p \,\|\, q) = \sum_i p_i \log \frac{p_i}{q_i} \ge 0,$$

且等號成立若且唯若 $p = q$。(b) 證明 ELBO 是 $\log p(x)$ 的下界：對任意滿足支撐條件的分布 $q(z)$，

$$\log p(x) \ge \mathbb{E}_q\big[\log p(x, Z) - \log q(Z)\big].$$

提示：$p(x) = \sum_z q(z)\frac{p(x,z)}{q(z)}$，對凹函數 $\log$ 用 Jensen。

練習 6.12 (挑戰 · 程式)　用 NumPy 驗證中央極限定理與 $1/\sqrt{n}$ 縮放。(a) 對 $n = 1, 10, 100, 1000$，各生成 $M = 10^4$ 組「n 個 Exp(1) 樣本的平均」，計算每組樣本平均的標準差，驗證接近 $1/\sqrt{n}$（即 $1,\ 0.316,\ 0.1,\ 0.0316$）。(b) 把 $n = 1000$ 的樣本平均標準化為 $z = (\bar x - 1)\sqrt{n}$，計算落在 ± 1 與 ± 1.96 內的比例，與理論值 68.3%、95% 對照。(c)（選做）畫出 z 的直方圖並疊上標準常態密度，觀察鐘形。

6.7　練習題詳解

練習 6.1 詳解　回應時間 $T \ge 0$、$\mathbb{E}[T] = 200$ ms。

(a) 馬可夫不等式：

$$\mathbb{P}(T \ge 1000) \le \frac{200}{1000} = 0.2.$$

(b) 同法：

$$\mathbb{P}(T \ge 500) \le \frac{200}{500} = 0.4.$$

(c) $\mathbb{P}(T \ge 100) \le 200/100 = 2$——上界超過 1，毫無資訊（任何機率本來就 ≤ 1）。馬可夫界只有在門檻 $a > \mathbb{E}[T]$ 時才非平凡：它回答的是「比平均差很多倍」的機率。

練習 6.2 詳解　$\mu = 100$、$\sigma = 15$。

(a) 區間 $(70, 130)$ 即 $|X - 100| < 30 = 2\sigma$（$k = 2$）：

$$\mathbb{P}(70 < X < 130) \geq 1 - \frac{1}{2^2} = \frac{3}{4} = 75\%.$$

(b) $45 = 3\sigma$（$k = 3$）：

$$\mathbb{P}(|X - 100| \geq 45) \leq \frac{1}{3^2} = \frac{1}{9} \approx 0.111.$$

(c) 常態分布的實際值：2σ 內為 95.45%（即 2σ 外 4.55%）；3σ 外為 0.27%。比較「外側」機率：

$$\frac{0.25}{0.0455} \approx 5.5 \text{ 倍}, \qquad \frac{0.1111}{0.0027} \approx 41 \text{ 倍}.$$

柴比雪夫是對「所有分布的最壞情況」的保證，k 越大，它與輕尾分布（如常態）的差距越懸殊。

練習 6.3 詳解　逐一檢查凸性（二階可微時看 φ'' 的正負）。

(a) $(e^x)'' = e^x > 0$，凸：$\mathbb{E}[e^X] \geq e^{\mathbb{E}[X]}$。（此式是矩母函數理論與 Chernoff 界的起點。）

(b) $(\sqrt{x})'' = -\frac{1}{4}x^{-3/2} < 0$，凹：$\mathbb{E}[\sqrt{X}] \leq \sqrt{\mathbb{E}[X]}$。

(c) $(1/x)'' = 2/x^3 > 0$（$x > 0$），凸：$\mathbb{E}[1/X] \geq 1/\mathbb{E}[X]$。

(d) $(x^4)'' = 12x^2 \geq 0$，凸：$\mathbb{E}[X^4] \geq (\mathbb{E}[X])^4$。

(e) $|x|$ 凸（三角不等式可直接驗證定義，雖在 0 不可微）：$\mathbb{E}[|X|] \geq |\mathbb{E}[X]|$。

(f) $(x^3)'' = 6x$ 會變號：x^3 在整條實數線上**既非凸也非凹**，Jensen 不適用，兩個方向都可能發生。例：X 等機率取 $\{-2, 1\}$，則 $\mathbb{E}[X^3] = \frac{-8+1}{2} = -3.5 < (\mathbb{E}X)^3 = (-0.5)^3 = -0.125$；改取 $\{2, -1\}$，則 $\mathbb{E}[X^3] = 3.5 > (0.5)^3 = 0.125$。（若額外已知 $X \geq 0$，x^3 在 $[0, \infty)$ 上凸，方向即為 $\geq$。）

練習 6.4 詳解　一切由 $\text{SE} = \sigma/\sqrt{n}$ 推出：$\text{SE} \propto 1/\sqrt{n}$，故「精度 $\times k$ 需要樣本 $\times k^2$」。

(a) 標準誤 $\div 10$ 需要 $n \times 10^2 = 100$ 倍資料。

(b) n 變成 4 倍（$50 \to 200$），標準誤變成 $1/\sqrt{4} = 1/2$：減半。

(c) 標準誤要從 0.06 降到 0.02，即 $\div 3$，需要 $n \times 9$：

$$n = 400 \times 9 = 3600.$$

練習 6.5 詳解　$\text{Var}(\bar{X}_n) = \sigma^2/n = 4/n$。

(a) 柴比雪夫（$\varepsilon = 0.5$）：

$$\mathbb{P}(|\bar{X}_n - \mu| \geq 0.5) \leq \frac{4/n}{0.5^2} = \frac{16}{n} \leq 0.05 \quad \Longleftrightarrow \quad n \geq 320.$$

(b) CLT：$\bar{X}_n$ 近似 $N(\mu, 4/n)$，要求 0.5 至少是 1.96 個標準誤：

$$1.96 \cdot \frac{2}{\sqrt{n}} \leq 0.5 \quad \Longleftrightarrow \quad \sqrt{n} \geq 7.84 \quad \Longleftrightarrow \quad n \geq 61.5,$$

取 $n = 62$。
(c) 柴比雪夫的答案約是 CLT 的 5 倍。差異來源：柴比雪夫對**任何**分布成立（最壞情況保證），CLT 則利用「平均近似常態」這項額外資訊，換得小得多的樣本數。工程上常以 CLT 規劃、以柴比雪夫做保底。

練習 6.6 詳解　(a) 事件等價：$\{|X-\mu| \ge \varepsilon\} = \{(X-\mu)^2 \ge \varepsilon^2\}$。$(X-\mu)^2$ 非負且期望為 σ^2，馬可夫（門檻 ε^2）給出

$$\mathbb{P}(|X-\mu| \ge \varepsilon) = \mathbb{P}\big((X-\mu)^2 \ge \varepsilon^2\big) \le \frac{\sigma^2}{\varepsilon^2}.$$

取 $\varepsilon = k\sigma$ 得 $1/k^2$ 形式。
(b) 同法，改用四次方（仍是遞增變換下的等價事件）：

$$\mathbb{P}(|X-\mu| \ge \varepsilon) = \mathbb{P}\big((X-\mu)^4 \ge \varepsilon^4\big) \le \frac{\mathbb{E}[(X-\mu)^4]}{\varepsilon^4}.$$

若四階矩有限，此界對大 ε 的衰減（ε^{-4}）比柴比雪夫（ε^{-2}）更快。
(c) 一般模式：g 嚴格遞增且非負時，$\{|Z| \ge t\} = \{g(|Z|) \ge g(t)\}$，故

$$\mathbb{P}(|Z| \ge t) \le \frac{\mathbb{E}[g(|Z|)]}{g(t)}.$$

「換一個 g，換一個界。」取 $g(x) = e^{\lambda x}$ 得

$$\mathbb{P}(Z \ge t) \le e^{-\lambda t}\,\mathbb{E}[e^{\lambda Z}],$$

再對 $\lambda > 0$ 取最小值，就是 **Chernoff 界**——對指數矩有限的變數給出指數衰減的尾界，Hoeffding 不等式正是把它用在有界變數的和上得到的。

練習 6.7 詳解　$X \sim \mathrm{Bin}(100, 0.5)$：$\mu = np = 50$，$\sigma = \sqrt{np(1-p)} = \sqrt{25} = 5$。
(a) 不做校正：

$$\mathbb{P}(45 \le X \le 55) \approx \mathbb{P}\left(\frac{45-50}{5} \le Z \le \frac{55-50}{5}\right) = \mathbb{P}(-1 \le Z \le 1) = 2\Phi(1) - 1 \approx 0.6827.$$

(b) 連續性校正把區間攤成 $[44.5, 55.5]$：

$$\mathbb{P}(44.5 \le X \le 55.5) \approx \mathbb{P}\left(-\frac{5.5}{5} \le Z \le \frac{5.5}{5}\right) = 2\Phi(1.1) - 1 \approx 0.7287.$$

(c) 與精確值 0.7287 比較：無校正誤差約 0.046（第二位小數就錯），校正後誤差約 8×10^{-5}。直觀原因：端點 $\mathbb{P}(X = 45) \approx 0.048$ 的質量，不校正等於各被「切掉一半」。

練習 6.8 詳解　(a) 標準誤與誤差界：

$$\widehat{\mathrm{SE}} = \sqrt{\frac{\hat p(1-\hat p)}{n}} = \sqrt{\frac{0.48\times 0.52}{1200}} \approx 0.0144, \qquad m = 1.96\times 0.0144 \approx 0.028.$$

95% 信賴區間 0.48 ± 0.028，即 $[45.2\%,\ 50.8\%]$——「支持率四成八，誤差正負二點八個百分點」。因為區間上端超過 50%，**不能**排除真實支持率過半。
(b) 最壞情況 $p(1-p)\le 1/4$：

$$n \ge \frac{1.96^2\times 0.25}{0.02^2} = \frac{0.9604}{0.0004} = 2401.$$

(c) 同法 $n \ge 0.9604/0.0001 = 9604$。從 $\pm 2\%$ 到 $\pm 1\%$，樣本數 $2401 \to 9604$，恰為 4 倍：**誤差減半、成本四倍**，又是 $1/\sqrt{n}$ 匯率。

練習 6.9 詳解　(a) 令 $U\sim \mathrm{Uniform}(0,1)$、$g(x)=e^{-x^2}$，則

$$I = \int_0^1 e^{-x^2}\,\mathrm{d}x = \mathbb{E}[g(U)], \qquad \hat I_N = \frac{1}{N}\sum_{i=1}^{N} g(u_i), \quad u_i \overset{\text{i.i.d.}}{\sim} \mathrm{Uniform}(0,1).$$

$\hat I_N$ 是 i.i.d. 變數 $g(u_i)$ 的樣本平均且 $\mathbb{E}[g(U)] = I$ 有限，大數法則保證 $\hat I_N \xrightarrow{p} I$（且無偏）。
(b) 設 $Y\in[0,1]$。則 $Y^2 \le Y$，故 $\mathbb{E}[Y^2]\le \mathbb{E}[Y]$，於是

$$\mathrm{Var}(Y) = \mathbb{E}[Y^2] - (\mathbb{E}Y)^2 \le \mathbb{E}[Y] - (\mathbb{E}Y)^2 = \mathbb{E}[Y]\big(1-\mathbb{E}[Y]\big) \le \frac{1}{4},$$

最後一步因 $t(1-t)$ 在 $t=1/2$ 取最大值 $1/4$。本題 $g(U)\in[e^{-1},1]\subset[0,1]$，故 $\mathrm{Var}(g(U))\le 1/4$、$\mathrm{SE}(\hat I_N)\le \frac{1}{2\sqrt N}$。要 $\mathrm{SE}\le 0.001$：

$$\frac{1}{2\sqrt N} \le 0.001 \quad\Longleftrightarrow\quad \sqrt N \ge 500 \quad\Longleftrightarrow\quad N \ge 250{,}000.$$

(c) 實際變異數

$$\mathrm{Var}(g(U)) = \mathbb{E}[g^2] - (\mathbb{E}g)^2 \approx 0.5981 - 0.7468^2 \approx 0.0404, \qquad \mathrm{sd}\approx 0.201.$$

所需樣本 $N \ge 0.0404/0.001^2 \approx 40{,}400$（約四萬）。保守界 250,000 約多花 6 倍——「完全不看 g 形狀」的代價；實務上常先以小規模試驗估計變異數，再決定 N。

練習 6.10 詳解　(a) 期望的**線性**（不需要獨立性）：

$$\mathbb{E}[\bar X_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{n\mu}{n} = \mu.$$

變異數：常數平方提出（$\mathrm{Var}(cY) = c^2\,\mathrm{Var}\,Y$），再用**獨立**變數變異數可加：

$$\mathrm{Var}(\bar X_n) = \frac{1}{n^2}\,\mathrm{Var}\Big(\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

(b) 對 $\bar X_n$（均值 μ、變異數 σ^2/n）套柴比雪夫的 ε 形式：

$$\mathbb{P}\big(|\bar X_n-\mu|\ge\varepsilon\big)\le\frac{\operatorname{Var}(\bar X_n)}{\varepsilon^2}=\frac{\sigma^2}{n\varepsilon^2}\xrightarrow[n\to\infty]{}0.$$

依定義，$\bar X_n\xrightarrow{p}\mu$：弱大數法則得證，且界還附贈速率 $O(1/n)$。■
(c) 要 $\frac{\sigma^2}{n\varepsilon^2}=\frac{100}{n}\le 0.01$，即 $n\ge 100/0.01=10{,}000$。

練習 6.11 詳解　(a) 因為 log 是凹函數，Jensen 給 $\mathbb{E}[\log Y]\le\log\mathbb{E}[Y]$。令 $Y=q_i/p_i$（在分布 p 之下取值的隨機變數）：

$$-\operatorname{KL}(p\,\|\,q)=\sum_i p_i\log\frac{q_i}{p_i}\le\log\sum_i p_i\cdot\frac{q_i}{p_i}=\log\sum_i q_i=\log 1=0,$$

故 $\operatorname{KL}(p\,\|\,q)\ge 0$。等號條件：log 是**嚴格**凹函數，等號成立若且唯若 $Y=q_i/p_i$ 是常數 c；由 $\sum_i q_i=c\sum_i p_i$ 得 $c=1$，即 $p=q$。（此結果亦稱 Gibbs 不等式。）
(b) 由全機率與提示，$p(x)=\sum_z q(z)\frac{p(x,z)}{q(z)}=\mathbb{E}_q\big[\frac{p(x,Z)}{q(Z)}\big]$。再用 Jensen（log 凹）：

$$\log p(x)=\log\mathbb{E}_q\left[\frac{p(x,Z)}{q(Z)}\right]\ge\mathbb{E}_q\left[\log\frac{p(x,Z)}{q(Z)}\right]=\mathbb{E}_q\big[\log p(x,Z)-\log q(Z)\big].$$

右邊就是 **ELBO**（evidence lower bound）；兩邊之差恰為 $\operatorname{KL}\big(q(z)\,\|\,p(z\mid x)\big)\ge 0$——變分推論（第 12 章）就建立在這兩行 Jensen 上。

練習 6.12 詳解　參考程式與輸出：

```python
import numpy as np
rng = np.random.default_rng(0)
M = 10_000
# (a) 樣本平均的標準差應接近 1/sqrt(n)
for n in [1, 10, 100, 1000]:
    means = rng.exponential(1.0, size=(M, n)).mean(axis=1)
    print(n, round(means.std(), 4), round(1 / n**0.5, 4))
# (b) n=1000 的標準化樣本平均 vs N(0,1)
n = 1000
z = (rng.exponential(1.0, size=(M, n)).mean(axis=1) - 1) * n**0.5
print(round(np.mean(np.abs(z) <= 1.00), 4))   # 理論值 0.6827
print(round(np.mean(np.abs(z) <= 1.96), 4))   # 理論值 0.9500
```

(a) 兩欄應逐列吻合（誤差約在第三位小數）。(b) 比例應接近 0.683 與 0.950：右偏的 Exp(1) 平均後已幾乎服從 $N(0,1)$。(c) 直方圖疊上常態密度即為圖 6.3 的模擬版。

下一章：參數估計與最大概似——把「樣本平均收斂到期望值」升級成一整套估計理論。