機率與統計：從入門到精通

第 3 章　隨機變數與期望值

Random Variables and Expectation

難度：核心

到目前為止，我們談論不確定性的方式都是「定性」的：事件要麼發生、要麼不發生——「下雨了」、「抽到紅心」、「這封郵件是垃圾郵件」。但科學、工程與機器學習真正要處理的，幾乎都是**數字**：下了幾毫米的雨？模型在這筆資料上的損失是多少？投資組合明天的報酬率是多少？**隨機變數**把數字附到隨機結果上，讓我們第一次能對不確定性「做算術」：取平均、衡量波動、把許多獨立副本加總起來。而隨機變數的兩個摘要數字——描述中心位置的**期望值**，與描述離散程度的**變異數**——是統計學與機器學習的日常詞彙：損失函數的目標、梯度的雜訊、預測的信心，全都用它們書寫。本章把這套詞彙一次建立起來。

學習目標

讀完本章後，你將能夠：

- 定義**隨機變數**，區分**離散**與**連續**兩型，並以機率質量函數（PMF）、機率密度函數（PDF）與累積分布函數（CDF）三種語言描述其分布。
- 解釋為什麼「密度不是機率」：$f(x)$ 可以超過 1、單點機率為 0，只有面積（積分）才是機率。
- 計算並詮釋**期望值**；熟練運用**線性性**（即使變數相依也成立）與**無意識統計學家定律**（LOTUS）。
- 計算**變異數**與**標準差**，運用縮放規則 $\operatorname{Var}[aX+b]=a^2\operatorname{Var}[X]$ 與「獨立和的變異數相加」。
- 說明高階**動差**（偏度、峰度）刻畫的形狀資訊，執行**標準化** $Z=(X-\mu)/\sigma$，並用 Python 模擬驗證本章的理論性質。

3.1 隨機變數：把結果變成數字

第 1 章的樣本空間可以收納任何東西：正反面的序列、撲克牌的花色、明天的天氣。這些結果本身不是數字——你沒辦法把 HT 與 TH 相加，也沒辦法對「紅心」取平均。但幾乎所有定量工作都需要數字：計數、平均、比大小。連接兩個世界的橋樑，就是隨機變數——一條把每個結果標上數字的規則。

定義 3.1　隨機變數

隨機變數（random variable）是一個函數

$$X : \Omega \to \mathbb{R},$$

把樣本空間中的每個結果 ω 對應到一個實數 $X(\omega)$。慣例上以大寫字母 X, Y, Z 表示隨機變數，以小寫字母 x, y, z 表示它們可能取的值。

例 3.1　從結果到數字：三枚硬幣的正面數

擲三枚公正硬幣，樣本空間有 $2^3 = 8$ 個等可能結果。定義 $X =$「正面的個數」，則 X 把每個結果映成一個數字：

結果 ω	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
$X(\omega)$	3	2	2	2	1	1	1	0

「$X = 2$」其實是一個**事件**的縮寫：$\{\omega : X(\omega) = 2\} = \{\text{HHT}, \text{HTH}, \text{THH}\}$，因此

$$\mathbb{P}(X = 2) = \frac{3}{8}.$$

同樣地 $\mathbb{P}(X = 0) = \frac{1}{8}$、$\mathbb{P}(X = 1) = \frac{3}{8}$、$\mathbb{P}(X = 3) = \frac{1}{8}$，四個機率加總為 1——因為事件 $\{X = 0\}, \{X = 1\}, \{X = 2\}, \{X = 3\}$ 互斥且涵蓋整個 Ω。

請注意：隨機變數本身並不「隨機」——它是一個固定的函數，隨機的是餵給它的結果 ω。所有關於 X 的機率敘述都是第 1 章事件語言的縮寫，例如 $\mathbb{P}(a \le X \le b) \equiv \mathbb{P}(\{\omega : a \le X(\omega) \le b\})$；新的是詞彙，不是機率，Kolmogorov 公理與第 1 章的所有規則原封不動地適用。

定義 3.2　離散與連續隨機變數

- **離散**（discrete）隨機變數的可能取值可以列成一份有限或可數無限的清單：骰子的點數（1 到 6）、一篇文章的字數、一小時內客服接到的來電次數。
- **連續**（continuous）隨機變數的可能取值鋪滿實數軸上的一段連續統：身高、氣溫、一次物理測量的誤差。

（也存在混合型，本書聚焦這兩大類。）兩型的記帳方式不同：離散的機率集中在孤立的點上，逐點記錄即可；連續的機率抹在區間上，按區間記帳。以下分別介紹。

定義 3.3　機率質量函數（PMF）

離散隨機變數 X 的**機率質量函數**（probability mass function, PMF）為

$$p(x) = \mathbb{P}(X = x),$$

並滿足兩個條件：$p(x) \ge 0$ 對所有 x，以及 $\sum_x p(x) = 1$（對所有可能取值加總）。知道

PMF 就知道一切：對任何一組值 A，$\mathbb{P}(X \in A) = \sum_{x \in A} p(x)$。

最簡單的例子是公正骰子：X = 點數，每面等可能，$p(x) = \frac{1}{6}$（$x = 1, \dots, 6$，其他值處為 0），總和恰為 1——這是最簡單的「均勻」離散分布。由 PMF 可算任何事件的機率，例如 $\mathbb{P}(X \text{ 為偶數}) = p(2) + p(4) + p(6) = \frac{1}{2}$；下圖畫出這個 PMF 與它的 CDF。

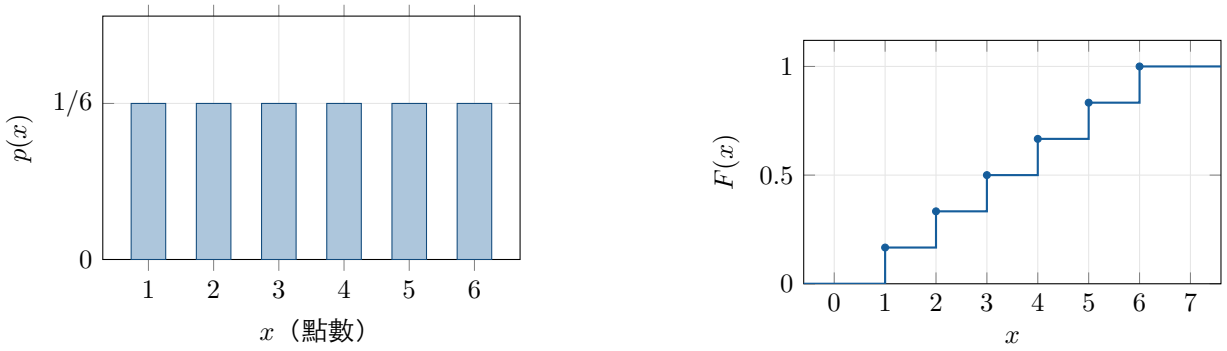


圖 3.1: 公正骰子的 PMF（左）與 CDF（右）。離散分布的機率集中在孤立的點上；CDF 是右連續的階梯函數，每一階跳升的高度正是該點的 PMF 值 1/6。

連續變數的記帳方式完全不同。身高「精確等於」170.000… 公分的機率是零——實數多到任何單一值都分不到正的機率。連續變數的機率不放在點上，而是**抹**在數線上，形成一層「密度」；問機率時，問的是一段區間上累積了多少。

定義 3.4　機率密度函數（PDF）

連續隨機變數 X 的分布由**機率密度函數**（probability density function, PDF）$f(x)$ 描述：對任何 $a \le b$，

$$\mathbb{P}(a \le X \le b) = \int_a^b f(x)\,\mathrm{d}x,$$

並滿足 $f(x) \ge 0$ 與 $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$（總機率為一）。**機率是密度曲線下的面積**：密度本身不是機率，密度乘上寬度（積分）才是。

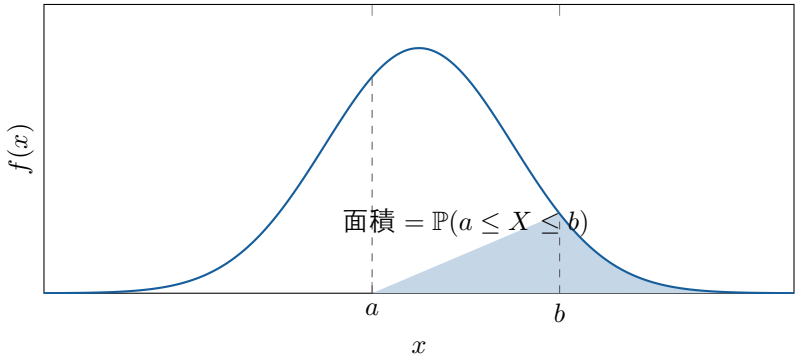


圖 3.2: 連續分布的機率是密度曲線下的面積：$\mathbb{P}(a \le X \le b) = \int_a^b f(x)\,\mathrm{d}x$（圖中陰影面積約 0.62）。曲線的高度 $f(x)$ 本身不是機率——它可以超過 1。

例 3.2　等公車：均勻密度，以及密度大於 1

公車每 10 分鐘一班，你隨機抵達站牌，等待時間 X（分鐘）在 $[0,10]$ 上「均勻」分布：每一段同長的區間分到相同機率。密度必須是常數，而總面積要等於 1，故

$$f(x)=\frac{1}{10},\quad 0\le x\le 10;\qquad f(x)=0,\quad \text{其他}.$$

（驗證：$\int_0^{10}\frac{1}{10}\,\mathrm{d}x=1$。）「等待 2 到 5 分鐘」的機率是面積

$$\mathbb{P}(2\le X\le 5)=\int_2^5\frac{1}{10}\,\mathrm{d}x=\frac{3}{10},$$

與第 1 章「長度比」的直觀完全一致。

解　現在把班距改成半分鐘：$X\sim U(0,\frac{1}{2})$。區間長度只剩 $\frac{1}{2}$，要撐起總面積 1，高度必須是 $f(x)=2$——密度**大於 1**！這完全合法：f 是「機率的濃度」，不是機率本身；任何區間的面積仍不超過 1，例如 $\mathbb{P}(0\le X\le\frac{1}{4})=2\times\frac{1}{4}=\frac{1}{2}$。

常見陷阱

密度不是機率——這是連續分布最經典的誤解，請牢記三件事：

1. $f(x)$ **可以超過** 1（剛剛的 $U(0,\frac{1}{2})$ 密度為 2）。它是機率的「濃度」，單位是「機率除以長度」；就像物理密度可以超過 $1\,\mathrm{g/cm^3}$，質量卻要密度乘體積才得到。
2. 任何單點的機率為零：$\mathbb{P}(X=x)=\int_x^x f(t)\,\mathrm{d}t=0$。「身高恰好 170.000… 公分」的機率是 0，但「169.5 到 170.5 之間」的機率是正的。
3. 因此對連續變數**從不**寫 $\mathbb{P}(X=x)$；一律在區間上工作。小區間 $[x,x+\mathrm{d}x]$ 上的機率近似 $f(x)\,\mathrm{d}x$——密度乘寬度才是機率。

把 $f(x)$ 當成 $\mathbb{P}(X=x)$ 來用，是之後學習貝氏公式與概似函數（第 7 章）時反覆跌倒的地方；現在把觀念扎穩。

離散用 PMF、連續用 PDF，兩套記帳法看似分裂。幸好有一個對兩型通吃的統一描述：不問「機率在哪裡」，改問「從左邊累積到 x，總共收了多少機率」。

定義 3.5　累積分布函數（CDF）

隨機變數 X 的**累積分布函數**（cumulative distribution function, CDF）為

$$F(x)=\mathbb{P}(X\le x),\qquad x\in\mathbb{R}.$$

它對離散與連續變數都有定義，且恆有下列性質：

1. **非遞減**（累積量不會變少），且 $\lim_{x\to-\infty}F(x)=0$、$\lim_{x\to+\infty}F(x)=1$；
2. **區間機率**：$\mathbb{P}(a<X\le b)=F(b)-F(a)$；
3. 離散變數的 F 是右連續的**階梯函數**，在每個可能值 x 處跳升，跳升高度恰為 $p(x)$；

連續變數的 $F(x)=\int_{-\infty}^{x} f(t)\,\mathrm{d}t$ 是連續函數，且在 f 連續處 $F'(x)=f(x)$（微積分基本定理）。

兩個熟面孔立刻可以驗證：骰子的 CDF 是六階的階梯（本節圖右），例如 $F(2.5)=p(1)+p(2)=\frac{1}{3}$；等公車 $X\sim U(0,10)$ 的 CDF 是一道斜坡 $F(x)=x/10$ $(0\le x\le 10$；左側為 0、右側為 1），微分回去 $F'(x)=\frac{1}{10}=f(x)$，密度失而復得，而 $\mathbb{P}(2<X\le 5)=F(5)-F(2)=0.3$ 與直接積分一致。

直觀理解

三種描述是同一個分布的三種語言：**PMF ／ PDF** 回答「機率鋪在哪裡」——離散是一點一點放砝碼，連續是抹一層厚薄不均的奶油；**CDF** 回答「累積到 x 收了多少」。翻譯規則：離散「加總 $\leftrightarrow$ 跳躍」、連續「積分 $\leftrightarrow$ 微分」。知道其中任何一個，就完整決定了分布；之後每認識一個新分布（第 4 章），第一件事都是問：它的 PMF 或 PDF 長什麼樣子？

3.2 期望值：分布的重心

一個分布是一張完整的地圖，但日常決策往往只需要一個數字：「平均而言，會是多少？」保險公司想知道平均理賠金額，賭場想知道每局平均獲利，機器學習想知道模型的平均損失。這個「機率加權的平均值」就是期望值。

定義 3.6　期望值

隨機變數 X 的**期望值**（expected value / expectation），又稱**均值**（mean），記作 $\mathbb{E}[X]$ 或 μ：

$$\mathbb{E}[X]=\sum_{x} x\,p(x)\quad\text{（離散）},\qquad \mathbb{E}[X]=\int_{-\infty}^{\infty} x\,f(x)\,\mathrm{d}x\quad\text{（連續）}.$$

它是把每個可能值 x 依其機率權重 $p(x)$（或密度 $f(x)\,\mathrm{d}x$）加權平均的結果。（技術註記：總和或積分絕對收斂時期望值才存在——本節末的聖彼得堡悖論將展示發散的例子。）

例 3.3　骰子的期望值

擲一顆公正骰子，$X=$ 點數。

$$\mathbb{E}[X]=1\cdot\frac{1}{6}+2\cdot\frac{1}{6}+3\cdot\frac{1}{6}+4\cdot\frac{1}{6}+5\cdot\frac{1}{6}+6\cdot\frac{1}{6}=\frac{21}{6}=3.5.$$

注意：3.5 **不是任何一面的點數！**期望值是分布的重心，不是「最可能出現的值」，也不必是 X 能取到的值——「期望」一詞其實有點誤導，正確讀法是：長期重複擲骰，每一擲**平均**貢獻 3.5 點。

直觀理解

期望值有三種互相補強的讀法：

1. **平衡點（重心）**：把每個機率 $p(x)$ 想成放在數線位置 x 上的砝碼，$\mathbb{E}[X]$ 就是讓整把尺恰好平衡的支點——「力矩和為零」$\sum_x (x-\mu)\,p(x)=0$ 移項正是定義。
2. **長期平均**：重複觀察很多次，樣本平均值會收斂到 $\mathbb{E}[X]$——這是**大數法則**（law of large numbers）的內容（第 6 章）；本章的 Python 實作先讓你親眼見到。
3. **公平價格**：彩金為 X 的賭局，入場費訂 $\mathbb{E}[X]$ 時雙方「長期不賺不賠」；保險精算的「純保費」正是期望理賠額（練習 3.6）。

例 3.4　指示變數：機率是期望值的特例

對任何事件 A，定義**指示變數**（indicator variable）$\mathbf{1}_A(\omega)$：$\omega \in A$ 時取 1，否則取 0。它是最簡單的隨機變數：只取 0 與 1 兩個值，$\mathbb{P}(\mathbf{1}_A = 1) = \mathbb{P}(A)$。其期望值為

$$\mathbb{E}[\mathbf{1}_A] = 1 \cdot \mathbb{P}(A) + 0 \cdot \mathbb{P}(A^c) = \mathbb{P}(A).$$

解　這個一行計算是一座重要的橋：**機率是期望值的特例**；反過來，許多複雜的期望值可以拆成一堆指示變數的和來算（接下來的帽子問題就是經典示範）。機器學習裡，分類器的 0–1 損失正是指示變數「預測錯誤」，其期望值就是錯誤率。

例 3.5　指數分布的期望值

指數分布是「等待事件發生的時間」的經典模型（第 4 章正式介紹）：密度為

$$f(x) = \lambda e^{-\lambda x}, \qquad x \geq 0,$$

其中 $\lambda > 0$ 是「速率」參數。求 $\mathbb{E}[X]$。

解　由定義並分部積分（取 $u = x$，$\mathrm{d}v = \lambda e^{-\lambda x}\,\mathrm{d}x$，則 $v = -e^{-\lambda x}$）：

$$\mathbb{E}[X] = \int_0^\infty x\,\lambda e^{-\lambda x}\,\mathrm{d}x = \Big[-x e^{-\lambda x}\Big]_0^\infty + \int_0^\infty e^{-\lambda x}\,\mathrm{d}x = 0 + \Big[-\frac{e^{-\lambda x}}{\lambda}\Big]_0^\infty = \frac{1}{\lambda}.$$

（第一項在 $x \to \infty$ 時因指數衰減壓過線性成長而趨於 0。）詮釋：λ 是單位時間平均發生次數，平均等待時間是其倒數——客服電話以每分鐘 $\lambda = 0.5$ 通進線，兩通之間平均間隔 2 分鐘。

期望值最重要的性質，是它對「加法與縮放」的完美配合——而且完全不需要獨立性。

命題 3.7　期望值的線性性

對任意隨機變數 X, Y 與常數 a, b, c：

$$\mathbb{E}[aX + bY + c] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y] + c.$$

特例：$\mathbb{E}[aX+b]=a\,\mathbb{E}[X]+b$，$\mathbb{E}[X+Y]=\mathbb{E}[X]+\mathbb{E}[Y]$。$X$ **與** Y **不需要獨立**——任意相依都成立。歸納可推廣到任意多項：$\mathbb{E}\big[\sum_i a_iX_i\big]=\sum_i a_i\mathbb{E}[X_i]$。

證明（離散情形）　記聯合機率 $p(x,y)=\mathbb{P}(X=x\text{ 且}Y=y)$（聯合分布的系統理論見第 5 章，這裡只需要「兩件事同時發生的機率」這個直觀意義）。先確認兩個「邊際」事實：固定 x，事件 $\{X=x\}$ 是互斥事件 $\{X=x,Y=y\}$（y 跑遍所有可能值）的聯集，由可加性

$$\sum_y p(x,y)=\mathbb{P}(X=x),\qquad \text{同理}\quad \sum_x p(x,y)=\mathbb{P}(Y=y).$$

於是

$$\begin{aligned}\mathbb{E}[aX+bY+c]&=\sum_x\sum_y (ax+by+c)\,p(x,y)\\&=a\sum_x x\Big[\sum_y p(x,y)\Big]+b\sum_y y\Big[\sum_x p(x,y)\Big]+c\sum_{x,y}p(x,y)\\&=a\sum_x x\,\mathbb{P}(X=x)+b\sum_y y\,\mathbb{P}(Y=y)+c\cdot 1=a\,\mathbb{E}[X]+b\,\mathbb{E}[Y]+c.\end{aligned}$$

整個推導對 $p(x,y)$ 的形狀**沒有任何要求**——沒有一步用到獨立性。連續情形把總和換成積分，同理可證。 ■

核心觀念

線性性是期望值最重要的性質，沒有之一。「和的期望等於期望的和」對**任意相依**的變數都成立——這是機率論裡少見的免費午餐（對照：變異數就沒這麼慷慨，見下一節）。它給出一套實戰模式：想算複雜總量的期望，先拆成簡單零件（常常是指示變數）的和，逐件取期望再加回來——不必知道零件之間如何相關，更不必求出總量的分布。下面的帽子問題讓這套模式一戰成名。

例 3.6　帽子問題：線性性的威力

n 位客人在餐廳寄放帽子，離場時服務生把帽子**隨機**發還（即帽子經過一個均勻隨機的重新排列）。設 $X=$ 拿回自己帽子的人數，求 $\mathbb{E}[X]$。

解　直接求 X 的分布很費勁——第 1 章的錯排問題光是算 $\mathbb{P}(X=0)$ 就動用了完整的排容原理。改走線性性：令 $I_i=\mathbf{1}\{\text{第}i\text{ 位拿回自己的帽子}\}$，則 $X=I_1+\cdots+I_n$。隨機排列下，第 i 位拿到 n 頂帽子中的任何一頂都等可能，故 $\mathbb{E}[I_i]=\frac{1}{n}$。由線性性，

$$\mathbb{E}[X]=\mathbb{E}[I_1]+\cdots+\mathbb{E}[I_n]=n\times\frac{1}{n}=1.$$

無論 10 人還是 10 萬人，**平均恰有 1 人**拿回自己的帽子！請特別注意：I_i 之間**高度相依**（若前 $n-1$ 位都拿對，最後一位必然拿對），但線性性根本不在乎——若堅持先求 X 的完整分布再取期望，是一場排容原理的硬仗；線性性三行結案。

接下來的問題：想算的常常不是 $\mathbb{E}[X]$ 本身，而是 X 某個函數的期望，例如 $\mathbb{E}[X^2]$。照定義，這得先求出新變數 $Y = g(X)$ 的分布——幸好有一條捷徑。

定理 3.8　無意識統計學家定律（LOTUS）

設 g 為函數，則

$$\mathbb{E}[g(X)] = \sum_{x} g(x)\,p(x) \quad \text{（離散）}, \qquad \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\,f(x)\,\mathrm{d}x \quad \text{（連續）}.$$

亦即：**不必**先求出 $g(X)$ 的分布，直接用 X 的分布替 $g(x)$ 加權即可。

證明（離散情形）　記 $Y = g(X)$，其 PMF 為 $p_Y(y) = \mathbb{P}(g(X) = y) = \sum_{x:\,g(x)=y} p(x)$。由期望值定義，

$$\mathbb{E}[Y] = \sum_{y} y\,p_Y(y) = \sum_{y} \sum_{x:\,g(x)=y} y\,p(x) = \sum_{y} \sum_{x:\,g(x)=y} g(x)\,p(x) = \sum_{x} g(x)\,p(x),$$

最後一步成立是因為每個 x 恰好屬於一個群組（它的 y 就是 $g(x)$）。 ∎

綽號「無意識統計學家定律」（law of the unconscious statistician）帶點自嘲：人們常「無意識地」直接寫下 $\sum g(x)\,p(x)$，彷彿那就是定義——嚴格說它需要證明，但結論正確，這個下意識的習慣動作於是被扶正成了定理。

例 3.7　LOTUS 實戰：骰子的平方

擲一顆公正骰子，求 $\mathbb{E}[X^2]$。

解　取 $g(x) = x^2$，由 LOTUS：

$$\mathbb{E}[X^2] = \sum_{x=1}^{6} x^2 \cdot \frac{1}{6} = \frac{1+4+9+16+25+36}{6} = \frac{91}{6} \approx 15.17.$$

對照：$(\mathbb{E}[X])^2 = 3.5^2 = 12.25$。**兩者不相等！**「平方的平均」大於「平均的平方」，差額 $\frac{91}{6} - \frac{49}{4} = \frac{35}{12}$ 正是下一節的主角——變異數。

常見陷阱

$\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$：「先變換再平均」與「先平均再變換」一般**不相等**！唯一保證相等的是線性函數 $g(x) = ax + b$（這正是線性性）。剛剛看到 $\mathbb{E}[X^2] = 15.17 \neq 12.25 = (\mathbb{E}[X])^2$。事實上對凸函數 g 恆有 $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$（Jensen 不等式），差距由分布的離散程度決定。機器學習版警訊：「把平均輸入餵進模型」**不等於**「模型輸出的平均」——用 $\mathbb{E}[X]$ 代入非線性模型冒充 $\mathbb{E}[g(X)]$ 是常見的隱形錯誤。

例 3.8　聖彼得堡悖論：無限的期望值（進階）

反覆擲一枚公正硬幣，直到第一次出現正面為止。若第一次正面出現在第 k 次投擲，你獲得 2^k 元。問：你願意付多少入場費玩這個遊戲？

解　「第一次正面出現在第 k 擲」表示前 $k-1$ 次全是反面、第 k 次是正面，機率為 $(\frac{1}{2})^k$。期望彩金為

$$\mathbb{E}[Y] = \sum_{k=1}^{\infty} 2^k \cdot \left(\frac{1}{2}\right)^k = \sum_{k=1}^{\infty} 1 = +\infty.$$

期望值是**無限大**——照「公平價格」的邏輯你該願意付任何金額，但實際上多數人不願付超過幾十元。這個張力稱為**聖彼得堡悖論**（St. Petersburg paradox），癥結在於巨額彩金全靠機率指數小的「尾巴事件」撐起（有一半機率你只拿到 2 元）。兩個經典消解：其一，賭場資本有限，彩金一封頂，期望值立刻塌縮到幾十元的量級（練習 3.9 會算出：即使賭場備有十億元，公平入場費也只有 31 元）；其二，Daniel Bernoulli（1738）主張金錢的滿足感是凹函數，以對數效用計算 $\mathbb{E}[\ln Y] = \sum_{k\ge 1} 2^{-k}\, k \ln 2 = \ln 4$——只相當於「確定拿到 4 元」。教訓：期望值是強大的摘要，但在**重尾**（heavy-tailed）分布下可能失真、甚至不存在；報告平均值之前，先看看分布的尾巴。

這如何驅動機器學習／ AI

機器學習的目標函數，幾乎都是期望值。訓練資料是從真實世界的分布中抽出的隨機變數；模型在一筆隨機資料上的損失 $\text{loss}(\theta; X)$ 因此也是隨機變數。學習的正式目標是最小化**風險**（risk）——期望損失 $R(\theta) = \mathbb{E}[\text{loss}(\theta; X)]$。實務上真實分布未知、這個期望算不出來，於是用一個 mini-batch 的**平均損失**來估計它：這正是「長期平均趨近期望值」的工程化，第 6 章的大數法則將給出理論保證。而線性性天天在訓練迴圈裡上工：批次損失是逐樣本損失的平均，批次梯度因此是逐樣本梯度的平均——「逐筆算梯度再相加」的合法性正來自線性性。

3.3　變異數與標準差：離散程度的度量

兩位基金經理的年報酬率期望值都是 5%：一位做定存加碼，報酬幾乎總在 4% 到 6% 之間；另一位重押單一飆股，可能 $+40\%$ 也可能 -30%。期望值一模一樣，風險天差地遠——「中心位置」這個摘要顯然不夠，我們需要第二個數字，量化「圍繞中心的散布程度」。

定義 3.9　變異數與標準差

隨機變數 X（記 $\mu = \mathbb{E}[X]$）的**變異數**（variance）為「偏離均值的平方」的期望：

$$\text{Var}[X] = \mathbb{E}\big[(X-\mu)^2\big] \ge 0,$$

其正平方根 $\sigma = \sqrt{\text{Var}[X]}$ 稱為**標準差**（standard deviation）。

為什麼要「平方」？偏差 $X-\mu$ 的期望恆為零——線性性立刻給出 $\mathbb{E}[X-\mu] = \mu - \mu = 0$，正負偏離完美抵銷，平均偏差量不出任何東西。先平方再平均，正負不再抵銷，而且大偏離會被放大計較。代價是單位變成 X 的平方（元變元平方、公分變平方公分），不便解讀——開根號回到原單位，得到標準差：X 的「典型偏離量」。報告結果時多半引用 σ；推導計算時多半操作 Var。

命題 3.10　變異數的計算公式

$$\operatorname{Var}[X] = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2.$$

由 $\operatorname{Var}[X] \geq 0$ 立得副產品：恆有 $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$。

證明　展開 $(X-\mu)^2 = X^2 - 2\mu X + \mu^2$，用線性性逐項取期望（注意 μ 是常數）：

$$\operatorname{Var}[X] = \mathbb{E}[X^2] - 2\mu\,\mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - 2\mu^2 + \mu^2 = \mathbb{E}[X^2] - \mu^2. \qquad \blacksquare$$

實務分工：**定義式**（偏差平方的平均）負責「看懂意義」；**計算式** $\mathbb{E}[X^2]-\mu^2$ 負責「動手算」——$\mathbb{E}[X^2]$ 用 LOTUS 直接加權，往往省下大量整理工夫。

例 3.9　骰子的變異數

擲一顆公正骰子，求 $\operatorname{Var}[X]$ 與 σ。

解　用計算式最快。前面已算得 $\mathbb{E}[X] = 3.5$、$\mathbb{E}[X^2] = \frac{91}{6}$，故

$$\operatorname{Var}[X] = \frac{91}{6} - (3.5)^2 = \frac{364}{24} - \frac{294}{24} = \frac{70}{24} = \frac{35}{12} \approx 2.917, \qquad \sigma = \sqrt{\tfrac{35}{12}} \approx 1.708.$$

用定義式驗算亦得同值：六個偏差 $\pm 2.5, \pm 1.5, \pm 0.5$ 的平方平均為 $17.5/6 = 35/12$。✓ 解讀：擲骰子的結果「典型上」偏離均值 3.5 約 1.7 點。

例 3.10　均勻分布的期望值與變異數

設 $X \sim U(a,b)$，即 $f(x) = \frac{1}{b-a}$（$a \leq x \leq b$）。求 $\mathbb{E}[X]$ 與 $\operatorname{Var}[X]$。

解　期望值的積分：

$$\mathbb{E}[X] = \int_a^b \frac{x}{b-a}\,\mathrm{d}x = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2},$$

正是區間**中點**——對稱分布的重心落在對稱軸上（等公車 $U(0,10)$ 平均等 5 分鐘）。再用 LOTUS 算二次項（用因式分解 $b^3 - a^3 = (b-a)(a^2+ab+b^2)$）：

$$\mathbb{E}[X^2] = \int_a^b \frac{x^2}{b-a}\,\mathrm{d}x = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3},$$

代入計算式並通分：

$$\operatorname{Var}[X] = \frac{a^2+ab+b^2}{3} - \Big(\frac{a+b}{2}\Big)^2 = \frac{a^2-2ab+b^2}{12} = \frac{(b-a)^2}{12}.$$

特例 $U(0,1)$：$\operatorname{Var} = \frac{1}{12}$。觀察：變異數只依賴區間**長度** $b-a$，與位置無關——整段平移，離散程度不變，正是縮放規則的預演。

命題 3.11　變異數的縮放規則與獨立和

1. **縮放規則**：對任意常數 a, b，

$$\operatorname{Var}[aX+b] = a^2 \operatorname{Var}[X], \qquad \text{因此} \quad \sigma_{aX+b} = |a|\,\sigma_X.$$

（平移 $+b$ 完全不改變離散度；縮放 a 以**平方**進入變異數。）

2. **獨立和**：若 X 與 Y **獨立**，則

$$\operatorname{Var}[X+Y] = \operatorname{Var}[X] + \operatorname{Var}[Y].$$

證明　(1) 記 $\mu = \mathbb{E}[X]$。線性性給 $\mathbb{E}[aX+b] = a\mu + b$，於是

$$(aX+b) - \mathbb{E}[aX+b] = a(X-\mu), \qquad \operatorname{Var}[aX+b] = \mathbb{E}\big[a^2(X-\mu)^2\big] = a^2 \operatorname{Var}[X].$$

開根號時 $\sqrt{a^2} = |a|$（標準差不能為負），故 $\sigma_{aX+b} = |a|\sigma_X$。

(2) 的關鍵是展開式

$$\operatorname{Var}[X+Y] = \operatorname{Var}[X] + \operatorname{Var}[Y] + 2\big(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\big),$$

練習 3.5 會引導你逐步完成。交叉項 $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ 稱為**共變異數**（covariance，第 5 章的主角）；當 X, Y 獨立時 $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$（嚴謹證明見第 5 章），交叉項歸零，變異數相加。■

解讀第 (2) 條：獨立來源的「波動」以**平方**的方式疊加（$\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$，帶著畢氏定理的味道）——這是第 6 章 $\sqrt{n}$ 律與中央極限定理的根源，也是練習 3.7 批次平均雜訊分析的全部數學。

常見陷阱

變異數的三大誤區，每一條都常見於考卷與程式碼：

1. **以為「差的變異數是變異數相減」。**錯：若 X, Y 獨立，$\operatorname{Var}[X-Y] = \operatorname{Var}[X] + (-1)^2 \operatorname{Var}[Y] = \operatorname{Var}[X] + \operatorname{Var}[Y]$——相減時不確定性照樣**疊加**（況且相減可能得出負數，根本不可能是變異數）。
2. **混淆 $2X$ 與「兩個獨立副本之和」。**$\operatorname{Var}[2X] = 4\operatorname{Var}[X]$，但獨立副本之和 $X_1 + X_2$ 的變異數是 $2\operatorname{Var}[X]$——「完全相關」與「獨立」差了一倍。
3. **相依時照抄相加公式。**極端反例：$Y = -X$ 時 $X+Y \equiv 0$，變異數是 0 而非 $2\operatorname{Var}[X]$。期望值的線性性**無條件**成立；變異數的相加性要用**獨立性**來換。

最後補一個尺度感：σ 與 X 同單位，可直接讀作「典型偏離」——常見分布的絕大部分機率落在 $\mu \pm 2\sigma$ 到 $\mu \pm 3\sigma$ 之間（第 4 章常態分布給出精確版：68–95–99.7 法則）。看到「平均 3.5、標準差 1.7」，腦中應自動浮現「數值大致在 1.8 到 5.2 之間打轉」的畫面。

3.4　高階動差與標準化

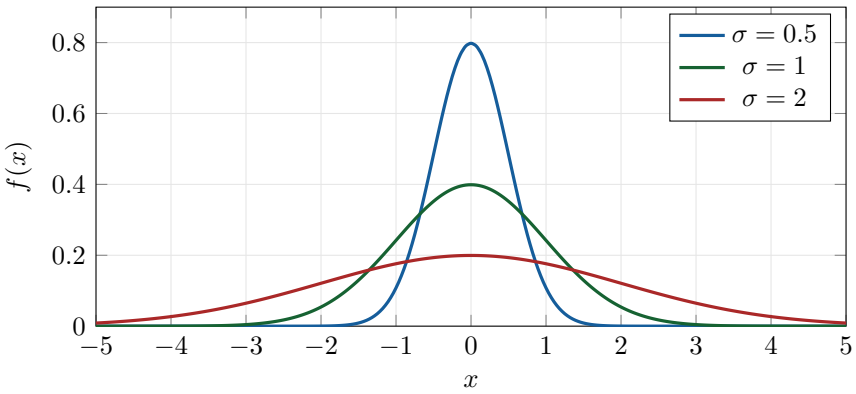


圖 3.3: 三個同均值（$\mu = 0$）、不同標準差的鐘形分布。中心位置由期望值決定，寬度由標準差決定：σ 越大，曲線越攤平，離散程度越高（總面積恆為 1）。

均值與變異數其實是一個更大家族的前兩位成員。它們抓住了分布的中心與寬度，卻抓不住**形狀**：分布對稱嗎？尾巴厚不厚？極端值常不常出現？回答這些問題需要更高階的「動差」。

定義 3.12　動差、偏度與峰度

隨機變數 X（記 $\mu = \mathbb{E}[X]$、$\sigma^2 = \mathrm{Var}[X]$）的 k 階**動差**（moment）為 $\mathbb{E}[X^k]$，k 階**中心動差**（central moment）為 $\mathbb{E}[(X-\mu)^k]$：均值是一階動差，變異數是二階中心動差。標準化的三、四階中心動差刻畫**形狀**：

- **偏度**（skewness）$\gamma_1 = \mathbb{E}[(X-\mu)^3]/\sigma^3$ 衡量**不對稱**。三次方保留正負號：右尾較長（如所得分布，少數人拉高右尾）給出正偏，左尾較長（如簡單考題的分數分布）給出負偏，對稱分布為 0。
- **峰度**（kurtosis）$\gamma_2 = \mathbb{E}[(X-\mu)^4]/\sigma^4$ 衡量**尾部厚度**：四次方極度放大大偏離，峰度高表示極端值的機率不可忽略。常態分布的峰度恰為 3，常作基準（$\gamma_2 - 3$ 稱為超額峰度）；金融資產報酬的峰度往往遠大於 3——市場崩盤這類「黑天鵝」比常態模型預測的更常發生。

（除以 σ 的冪次讓兩者成為**無單位**的純形狀量，不受量尺影響——這正是接下來「標準化」的思想。）

高階動差告訴我們：不同變數住在不同的單位與量級裡，直接比較沒有意義。把它們搬到同一把尺上的操作，是統計學最常用的手勢之一。

定義 3.13　標準化

設 $\mu = \mathbb{E}[X]$、$\sigma = \sqrt{\mathrm{Var}[X]} > 0$。$X$ 的**標準化**（standardization）為

$$Z = \frac{X - \mu}{\sigma}.$$

Z 稱為 z **分數**（z-score）：以標準差為單位，度量 X 離均值幾「步」。

標準化的效果兩行就能驗證：把 $Z = \frac{1}{\sigma}X - \frac{\mu}{\sigma}$ 看成 $aX + b$（$a = \frac{1}{\sigma}$、$b = -\frac{\mu}{\sigma}$），線性性給 $\mathbb{E}[Z] = (\mathbb{E}[X] - \mu)/\sigma = 0$，縮放規則給 $\mathrm{Var}[Z] = \mathrm{Var}[X]/\sigma^2 = 1$：**均值** 0、**變異數** 1，正是標準化的全部目的。標準化是線性變換，只做平移與縮放，**不改變分布的形狀**（偏度、峰度原封不動）；它做的是「行囊歸零」：中心搬到 0、尺度縮到 1。$z = +2$ 讀作「高於平均 2 個標準差」——無論原始單位是分數、公分還是新台幣。

例 3.11　z 分數：哪一科考得比較好？

小潔這次段考數學 82 分、英文 88 分。全班的分數分布：數學 $\mu = 70$、$\sigma = 8$；英文 $\mu = 80$、$\sigma = 10$。哪一科的**相對**表現較好？

解　分別標準化：

$$z_{\text{數}} = \frac{82 - 70}{8} = 1.5, \qquad z_{\text{英}} = \frac{88 - 80}{10} = 0.8.$$

數學高於全班平均 1.5 個標準差，英文只高 0.8 個——雖然英文的原始分數較高，數學的相對表現明顯更突出。z 分數的用途正是**跨尺度比較**：不同科目、不同單位、不同量級的變數，標準化之後都躺在同一把無單位的尺上。

之後各章反覆出現的手勢——常態表查值（第 4 章）、中央極限定理的表述（第 6 章）、檢定統計量與信賴區間（第 8 章）——都是同一個動作的變奏：**先標準化，再比較**。

這如何驅動機器學習／ AI

把本章詞彙逐條翻譯成機器學習的日常：

- **資料是隨機變數**：每一筆訓練樣本、每一個標籤，都是從某個分布抽出的隨機變數；「資料集」是它的多次實現。
- **目標是期望值**：訓練目標是風險（期望損失）$\mathbb{E}[\text{loss}]$，用 mini-batch 平均損失估計；線性性讓「逐樣本梯度取平均」成為合法的批次梯度。
- **變異數是雜訊的語言**：mini-batch 梯度的變異數就是 SGD 的「梯度雜訊」，隨批次大小以 $1/B$ 縮小（練習 3.7）；偏差–變異數權衡（第 10 章）用變異數量測模型對訓練資料的敏感度；許多模型也輸出預測變異數作為**不確定性**。
- **標準化是標配**：特徵正規化把每個特徵 z 分數化，讓梯度下降不被不同量綱拉扯；batch normalization 與 layer normalization 本質上就是對神經元活化值做 $(X - \mu)/\sigma$（再配兩個可學參數縮放平移回去），讓每一層輸入保持穩定尺度——深網路因此才「練得動」。
- **模型輸出也是隨機變數**：分類器輸出類別的分布、語言模型輸出下一個詞的分布；「點預測」是均值、「信心」是變異數——本章「兩數摘要」的直接應用。

3.5 Python 實作

理論之後，動手驗證。第一段程式用公正骰子示範：由 PMF 直接計算期望值與變異數（LOTUS 的數值版），再用百萬次模擬對照——長期平均確實趨近期望值，這是第 6 章大數法則的預告片。

```
import numpy as np

rng = np.random.default_rng(42)

# --- 1. 公正骰子：用 PMF 算理論值 ---
faces = np.arange(1, 7)                # 可能取值 1..6
pmf = np.full(6, 1/6)                  # 每面機率 1/6
mu = np.sum(faces * pmf)               # E[X]：機率加權平均
EX2 = np.sum(faces**2 * pmf)           # LOTUS：E[X^2]，不必求 X^2 的分布
var = EX2 - mu**2                      # Var = E[X^2] - (E[X])^2
print(mu, round(EX2, 4), round(var, 4))   # 3.5 15.1667 2.9167 (= 35/12)

# --- 2. 模擬對照：長期平均趨近期望值（第 6 章大數法則）---
rolls = rng.integers(1, 7, size=1_000_000)
print(round(rolls.mean(), 4), round(rolls.var(), 4))  # 約 3.5 與 2.917
```

第二段程式驗證兩個核心性質：線性性在**相依**變數上照樣成立，以及標準化把樣本壓成「均值 0、標準差 1」。至於批次平均如何以 $1/B$ 壓低變異數（「加大 batch 可以壓低梯度雜訊」的數學原型），理論見練習 3.7，程式驗證留給練習 3.10。

```
# --- 3. 線性性：X 與 Y 高度相依，期望值照樣拆開 ---
N = 1_000_000
X = rng.normal(2.0, 1.0, size=N)
Y = 3 * X + rng.normal(0.0, 1.0, size=N)   # Y 由 X 生成：強相依
print(round(np.mean(2*X - Y + 5), 4))      # 模擬 E[2X-Y+5]：約 3.0
print(round(2*X.mean() - Y.mean() + 5, 4)) # 線性性拆開：2*2-6+5 = 3

# --- 4. 標準化：z 分數的均值為 0、標準差為 1 ---
Z = (X - X.mean()) / X.std()
print(round(Z.mean(), 6), round(Z.std(), 6))   # 0.0 與 1.0
```

本章總結

- **隨機變數**是把結果變成數字的函數 $X : \Omega \to \mathbb{R}$；「$X = k$」是事件的縮寫，第 1 章的公理與規則全部沿用。
- 離散用 **PMF**（非負、總和 1）、連續用 **PDF**（非負、積分 1）描述分布；**CDF** $F(x) = \mathbb{P}(X \le x)$ 是兩型共用的統一語言——跳躍對應 PMF、斜率對應 PDF。

- **密度不是機率**：$f(x)$ 可以大於 1、單點機率為 0，只有面積 $\int_a^b f$ 才是機率；連續變數永遠在區間上工作。
- **期望值**是機率加權平均：分布的平衡點、長期平均的極限（大數法則，第 6 章）、公平價格；它不必是 X 能取到的值（骰子的 3.5），重尾時甚至可能不存在（聖彼得堡悖論）。
- **線性性** $\mathbb{E}[aX+bY+c]=a\mathbb{E}[X]+b\mathbb{E}[Y]+c$ 對**相依**變數照樣成立；搭配**指示變數**（$\mathbb{E}[\mathbf{1}_A]=\mathbb{P}(A)$）是計算複雜期望的第一利器（帽子問題）。**LOTUS** 讓 $\mathbb{E}[g(X)]$ 不必先求 $g(X)$ 的分布；但注意 $\mathbb{E}[g(X)]\neq g(\mathbb{E}[X])$。
- **變異數** $\mathrm{Var}[X]=\mathbb{E}[(X-\mu)^2]=\mathbb{E}[X^2]-(\mathbb{E}[X])^2$，標準差 σ 與 X 同單位。**平移不變、縮放平方**（$\mathrm{Var}[aX+b]=a^2\,\mathrm{Var}[X]$）；**獨立**時變異數相加（相減時也是加）——相依時交叉項（共變異數，第 5 章）出場。
- 高階**動差**補足形狀資訊：**偏度**量不對稱、**峰度**量尾厚。**標準化** $Z=(X-\mu)/\sigma$ 得均值 0、變異數 1，是特徵正規化與 batch / layer normalization 的數學核心。
- 機器學習對照表：資料是隨機變數；**風險 = 期望損失**是訓練目標，mini-batch 平均是它的估計；梯度雜訊的變異數隨批次大小以 $1/B$ 下降；模型輸出用均值（預測）與變異數（信心）摘要。

3.6 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。

練習 3.1 (基礎)　離散隨機變數 X 的 PMF 為 $p(x)=cx$，$x\in\{1,2,3,4\}$，其他值處為 0。(a) 求常數 c。(b) 求 $\mathbb{P}(X\geq 3)$。(c) 求 $\mathbb{E}[X]$ 與 $\mathrm{Var}[X]$。(d) 寫出完整的 CDF $F(x)$（含所有區段），並指出每個跳躍的高度。

練習 3.2 (基礎)　連續隨機變數 X 的密度為 $f(x)=cx$（$0\leq x\leq 2$），其他值處為 0。(a) 求常數 c。(b) 求 CDF $F(x)$，並計算 $\mathbb{P}(X\leq 1)$。(c) 求 $\mathbb{E}[X]$ 與 $\mathrm{Var}[X]$。

練習 3.3 (進階)　房間裡有 n 個人，生日獨立且均勻分布在 365 天中。對每一對人 (i,j)（$i<j$）定義指示變數 $I_{ij}=\mathbf{1}\{\text{第}i\text{ 人與第}j\text{ 人同一天生日}\}$，令 $X=\sum_{i<j}I_{ij}$ 為同生日的**配對總數**。(a) 求 $\mathbb{E}[I_{ij}]$。(b) 用線性性求 $\mathbb{E}[X]$ 的公式。(c) 代入 $n=23$ 與 $n=50$ 算出數值，並與第 1 章生日問題的結果（23 人時「至少一對同生日」的機率約 0.507）對照，說明「期望配對數」與「至少一對的機率」是兩個不同的摘要。

練習 3.4 (進階)　由變異數的定義出發，證明 $\operatorname{Var}[aX+b] = a^2\operatorname{Var}[X]$，並說明 $\sigma_{aX+b} = |a|\,\sigma_X$ 中的絕對值是從哪裡來的。

練習 3.5 (進階)　設 X, Y 為隨機變數，並承認以下事實：當 X, Y **獨立**時 $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$（第 5 章證明）。(a) 展開證明一般公式 $\operatorname{Var}[X+Y] = \operatorname{Var}[X] + \operatorname{Var}[Y] + 2\big(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\big)$，並由此推出獨立時 $\operatorname{Var}[X+Y] = \operatorname{Var}[X] + \operatorname{Var}[Y]$。(b) 獨立時 $\operatorname{Var}[X-Y]$ 是多少？(c) 取 $Y = -X$（$\operatorname{Var}[X] > 0$），用 (a) 的一般公式驗證 $\operatorname{Var}[X+Y] = 0$，並說明這對「相依時變異數不能直接相加」的警告有何啟示。

練習 3.6 (進階)　某保險公司承保一年期意外險：一年內發生事故的機率為 0.002，發生時理賠 50 萬元（至多一次）。(a) 求期望理賠金額（即「純保費」）。(b) 若保費訂為 $1{,}500$ 元，保險公司每張保單的期望利潤是多少？(c) 從保戶的角度：不投保時，損失 L 的期望值與標準差各是多少？投保後呢？據此說明一張「期望值上吃虧」的保單，為什麼對保戶仍然可能是理性的選擇。

練習 3.7 (進階)　訓練集中隨機抽出的單一樣本損失 L 滿足 $\mathbb{E}[L] = \mu$、$\operatorname{Var}[L] = \sigma^2$。一個 mini-batch 含 B 筆獨立樣本，批次平均損失為 $\bar{L} = \frac{1}{B}\sum_{i=1}^{B} L_i$。(a) 求 $\mathbb{E}[\bar{L}]$ 與 $\operatorname{Var}[\bar{L}]$。(b) 批次大小從 32 加到 128，批次平均的標準差（雜訊）變成原來的幾倍？(c) 想把雜訊標準差再減半，B 要開到多大？由此說明「加大 batch 的報酬遞減」現象。

練習 3.8 (挑戰)　**幾何分布的期望值**：獨立重複試驗，每次成功機率為 $p \in (0,1)$，令 $X =$ 第一次成功所需的試驗次數，其 PMF 為 $\mathbb{P}(X=k) = (1-p)^{k-1}p$（$k = 1, 2, \dots$）。(a) 驗證 PMF 總和為 1。(b) 用等比級數逐項微分的技巧證明 $\mathbb{E}[X] = 1/p$。(c) 改用「首步分析」：解釋遞迴式 $\mathbb{E}[X] = 1 + (1-p)\mathbb{E}[X]$ 的直觀意義，並由它重新解出 $\mathbb{E}[X]$。(d) $p = 1/6$ 時，擲骰子平均要幾次才會見到第一個 6？

練習 3.9 (挑戰)　**有限資本下的聖彼得堡賭局**：延續例 3.8 的賭局，但賭場資本有限，最多只能支付 2^m 元——若第一次正面出現在第 k 擲，實際支付 $\min(2^k, 2^m)$ 元。(a) 證明期望彩金為 $m+1$ 元。(b) 若賭場備有約十億元的資本（$2^{30} \approx 1.07 \times 10^9$），公平入場費是多少？(c) 這個結果對「期望值無限大」的悖論給出了什麼樣的消解？

練習 3.10 (挑戰 · 程式)　用 NumPy 完成以下模擬並與理論對照：(a) 模擬 10^6 次擲骰子，對照均值 3.5 與變異數 $35/12 \approx 2.917$。(b) 模擬幾何分布（$p = 1/6$），對照練習 3.8 的結論 $\mathbb{E}[X] = 6$。(c) 對批次大小 $B \in \{4, 16, 64, 256\}$，模擬批次平均的標準差，並對照理論值 $\sigma/\sqrt{B}$（$\sigma = \sqrt{35/12}$）。

3.7 練習題詳解

練習 3.1 詳解　(a) PMF 總和必須為 1：

$$\sum_{x=1}^{4} c\,x = c\,(1+2+3+4) = 10c = 1 \quad\Longrightarrow\quad c = \frac{1}{10}.$$

(b) $\mathbb{P}(X \ge 3) = p(3) + p(4) = 0.3 + 0.4 = 0.7$。
(c) 期望值與二次動差（LOTUS）：

$$\mathbb{E}[X] = \frac{1\cdot 1 + 2\cdot 2 + 3\cdot 3 + 4\cdot 4}{10} = \frac{1+4+9+16}{10} = 3, \qquad \mathbb{E}[X^2] = \frac{1+8+27+64}{10} = 10.$$

故 $\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = 10 - 9 = 1$（$\sigma = 1$）。注意這個 PMF 把較大的機率放在較大的值上，所以均值 3 高於「四值中點」2.5。
(d) CDF 為右連續的階梯函數：$x < 1$ 時 $F(x) = 0$，在 $[1,2)$、$[2,3)$、$[3,4)$ 上依序為 0.1、0.3、0.6，$x \ge 4$ 後為 1；跳躍高度 $0.1, 0.2, 0.3, 0.4$ 恰好是 PMF 的四個值。

練習 3.2 詳解　(a) 總面積為 1：

$$\int_0^2 c\,x\,\mathrm{d}x = c\cdot\frac{x^2}{2}\Big|_0^2 = 2c = 1 \quad\Longrightarrow\quad c = \frac{1}{2}.$$

(b) 對 $0 \le x \le 2$：

$$F(x) = \int_0^x \frac{t}{2}\,\mathrm{d}t = \frac{x^2}{4},$$

（$x < 0$ 時為 0、$x > 2$ 時為 1）。故 $\mathbb{P}(X \le 1) = F(1) = \frac{1}{4}$。驗算：$F'(x) = x/2 = f(x)$。
(c)

$$\mathbb{E}[X] = \int_0^2 x\cdot\frac{x}{2}\,\mathrm{d}x = \frac{x^3}{6}\Big|_0^2 = \frac{4}{3}, \qquad \mathbb{E}[X^2] = \int_0^2 x^2\cdot\frac{x}{2}\,\mathrm{d}x = \frac{x^4}{8}\Big|_0^2 = 2,$$

故 $\mathrm{Var}[X] = 2 - (\frac{4}{3})^2 = \frac{2}{9} \approx 0.222$、$\sigma = \sqrt{2/9} \approx 0.471$。密度向右遞增、機率往右堆，均值 $\frac{4}{3}$ 落在區間中點的右邊——合理。

練習 3.3 詳解　(a) 固定第 i 人的生日，第 j 人與其相同的機率為 $\frac{1}{365}$，故 $\mathbb{E}[I_{ij}] = \mathbb{P}(I_{ij} = 1) = \frac{1}{365}$。
(b) 配對數共 $\binom{n}{2}$ 項。各 I_{ij} 之間並不獨立（I_{12} 與 I_{13} 共享第 1 人的生日），但線性性照用：

$$\mathbb{E}[X] = \sum_{i<j} \mathbb{E}[I_{ij}] = \binom{n}{2}\cdot\frac{1}{365}.$$

(c) 代入數值：$n = 23$ 時 $\mathbb{E}[X] = \frac{253}{365} \approx 0.693$；$n = 50$ 時 $\mathbb{E}[X] = \frac{1225}{365} \approx 3.36$。對照第 1 章：23 人時「至少一對」的機率約 0.507，而期望配對數約 0.693——兩個數字回答不同的問題。機率 $\mathbb{P}(X \ge 1)$ 只在乎「有沒有」，期望值 $\mathbb{E}[X]$ 則把「兩對、三對」的情形也計入貢獻，所

以可以大於機率（50 人時平均已超過 3 對，機率卻不能超過 1）。同時注意難度懸殊：$\mathbb{E}[X]$ 用線性性一行搞定，$\mathbb{P}(X \ge 1)$ 卻需要第 1 章的整串連乘——期望值往往比分布容易得多。

練習 3.4 詳解　記 $\mu = \mathbb{E}[X]$，$Y = aX + b$。第一步（線性性）：$\mathbb{E}[Y] = a\mu + b$。第二步：中心化的偏差為 $Y - \mathbb{E}[Y] = (aX + b) - (a\mu + b) = a(X - \mu)$，常數 b 在相減時消失——這就是「平移不改變離散度」的代數原因。第三步（定義）：

$$\operatorname{Var}[Y] = \mathbb{E}\big[(Y - \mathbb{E}[Y])^2\big] = \mathbb{E}\big[a^2(X-\mu)^2\big] = a^2\,\mathbb{E}[(X-\mu)^2] = a^2 \operatorname{Var}[X].$$

取平方根：$\sigma_Y = \sqrt{a^2 \operatorname{Var}[X]} = \sqrt{a^2}\,\sigma_X = |a|\,\sigma_X$。絕對值來自 $\sqrt{a^2} = |a|$：當 $a < 0$ 時（例如 $Y = -X$），分布左右翻轉但離散程度不變，標準差必須仍是正的 $|a|\sigma_X$，而不是負數 $a\,\sigma_X$。

練習 3.5 詳解　(a) 記 $\mu_X = \mathbb{E}[X]$、$\mu_Y = \mathbb{E}[Y]$。由線性性 $\mathbb{E}[X+Y] = \mu_X + \mu_Y$，代入計算公式：

$$\begin{aligned}
\operatorname{Var}[X+Y] &= \mathbb{E}[(X+Y)^2] - (\mu_X + \mu_Y)^2 \\
&= \mathbb{E}[X^2 + 2XY + Y^2] - (\mu_X^2 + 2\mu_X\mu_Y + \mu_Y^2) \\
&= \big(\mathbb{E}[X^2] - \mu_X^2\big) + \big(\mathbb{E}[Y^2] - \mu_Y^2\big) + 2\big(\mathbb{E}[XY] - \mu_X\mu_Y\big) \\
&= \operatorname{Var}[X] + \operatorname{Var}[Y] + 2\big(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\big).
\end{aligned}$$

獨立時 $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$，交叉項歸零，得 $\operatorname{Var}[X+Y] = \operatorname{Var}[X] + \operatorname{Var}[Y]$。
(b) 寫 $X - Y = X + (-Y)$。由縮放規則 $\operatorname{Var}[-Y] = (-1)^2 \operatorname{Var}[Y] = \operatorname{Var}[Y]$，且 X 與 $-Y$ 仍獨立，故 $\operatorname{Var}[X-Y] = \operatorname{Var}[X] + \operatorname{Var}[Y]$——相減時不確定性**照樣疊加**，絕不是相減。
(c) 取 $Y = -X$：$X + Y \equiv 0$ 是常數，$\operatorname{Var}[X+Y] = 0$。用 (a) 的一般公式驗證：$\mathbb{E}[XY] = -\mathbb{E}[X^2]$、$\mu_X\mu_Y = -\mu_X^2$，交叉項為 $2(-\mathbb{E}[X^2] + \mu_X^2) = -2\operatorname{Var}[X]$，總和 $= \operatorname{Var}[X] + \operatorname{Var}[X] - 2\operatorname{Var}[X] = 0$。✓ 啟示：相依時交叉項（2 Cov，第 5 章）可以吃掉甚至反轉相加性——完全負相關時波動互相抵銷，這正是投資「對沖」的數學原理。

練習 3.6 詳解　(a) 理賠金額 L 取 500,000（機率 0.002）或 0（機率 0.998）：$\mathbb{E}[L] = 500{,}000 \times 0.002 = 1{,}000$ 元。這就是「純保費」：長期而言每張保單平均賠出 1,000 元。
(b) 期望利潤 $= 1{,}500 - \mathbb{E}[L] = 1{,}500 - 1{,}000 = 500$ 元。
(c) 不投保時：$\mathbb{E}[L] = 1{,}000$ 元，而

$$\mathbb{E}[L^2] = (500{,}000)^2 \times 0.002 = 2.5 \times 10^{11} \times 0.002 = 5 \times 10^8,$$

$$\operatorname{Var}[L] = 5 \times 10^8 - (1{,}000)^2 = 4.99 \times 10^8, \qquad \sigma_L \approx 22{,}338 \text{ 元}.$$

投保後：損失固定為保費 1,500 元，標準差為 0。比較：保戶在期望值上多付 500 元（1,500 對 1,000），換得的是把標準差從約 2.2 萬元壓到 0——把「小機率、毀滅性」的損失換成

「確定、可負擔」的小額支出。對風險趨避的人（用聖彼得堡例題的語言：效用函數是凹的），這 500 元的「風險溢價」買到的安穩完全值得。另一邊，保險公司承保上萬張**獨立**保單，總理賠的**相對**波動以 $1/\sqrt{n}$ 縮小（與練習 3.7 同一條數學）——保戶買確定性、公司賣大數法則，保險市場於焉成立。

練習 3.7 詳解　(a) 線性性（期望值部分連獨立性都不需要）：

$$\mathbb{E}[\bar{L}] = \frac{1}{B}\sum_{i=1}^{B}\mathbb{E}[L_i] = \frac{1}{B}\cdot B\mu = \mu,$$

批次平均是風險的**不偏**估計。變異數部分用獨立和與縮放規則：

$$\mathrm{Var}[\bar{L}] = \frac{1}{B^2}\,\mathrm{Var}\Big[\sum_{i=1}^{B} L_i\Big] = \frac{1}{B^2}\cdot B\sigma^2 = \frac{\sigma^2}{B}, \qquad \mathrm{SE}(\bar{L}) = \frac{\sigma}{\sqrt{B}}.$$

(b) B 從 32 變成 128 是放大 4 倍，標準差變為 $\frac{1}{\sqrt{4}} = \frac{1}{2}$ 倍——雜訊減半。
(c) 想再減半，需要再放大 4 倍：$B = 512$。每次「雜訊減半」都要付出「樣本翻四倍」的代價——雜訊以 $1/\sqrt{B}$ 下降，是平方根級的報酬遞減；算力成本卻隨 B 線性上升，所以 batch size 不是越大越好。這條 $\sigma/\sqrt{B}$ 正是第 6 章「標準誤」的原型。

練習 3.8 詳解　記 $q = 1 - p \in (0, 1)$。
(a) 等比級數（首項 p、公比 q）：

$$\sum_{k=1}^{\infty} q^{k-1}p = p\cdot\frac{1}{1-q} = \frac{p}{p} = 1. \checkmark$$

(b) 由定義 $\mathbb{E}[X] = \sum_{k=1}^{\infty} k\,q^{k-1}p$。關鍵技巧：對等比級數 $\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$（$|q| < 1$）逐項微分（冪級數在收斂半徑內可逐項微分）：

$$\sum_{k=1}^{\infty} k\,q^{k-1} = \frac{\mathrm{d}}{\mathrm{d}q}\Big(\frac{1}{1-q}\Big) = \frac{1}{(1-q)^2},$$

代回得 $\mathbb{E}[X] = p/(1-q)^2 = p/p^2 = 1/p$。
(c) 首步分析：看第一次試驗。以機率 p 直接成功，$X = 1$；以機率 q 失敗——此時已經花掉一次，而「接下來還要等多久」與從頭開始沒有差別（試驗獨立、成功機率不變），期望上還要 $\mathbb{E}[X]$ 次。故

$$\mathbb{E}[X] = p\cdot 1 + q\,\big(1 + \mathbb{E}[X]\big) = 1 + q\,\mathbb{E}[X] \quad\Longrightarrow\quad (1-q)\,\mathbb{E}[X] = 1 \quad\Longrightarrow\quad \mathbb{E}[X] = \frac{1}{p}.$$

兩法答案一致；遞迴法用到的「失敗後重新開始」直觀，正是第 4 章幾何分布**無記憶性**的預告。
(d) $p = \frac{1}{6}$：$\mathbb{E}[X] = 6$——平均擲 6 次才見到第一個 6，與「每次成功機率 $\frac{1}{6}$，平均要 $\frac{1}{1/6}$ 次」的直覺相符。

練習 3.9 詳解　(a) 實付彩金為 $\min(2^k, 2^m)$：$k \le m$ 時照付 2^k，$k > m$ 時封頂 2^m。期望值拆成兩段：

$$\mathbb{E}[Y] = \sum_{k=1}^{m} 2^{-k} \cdot 2^k + \sum_{k=m+1}^{\infty} 2^{-k} \cdot 2^m = \sum_{k=1}^{m} 1 + 2^m \sum_{k=m+1}^{\infty} 2^{-k}.$$

第二個級數是首項 $2^{-(m+1)}$、公比 $\frac{1}{2}$ 的等比級數，總和為 2^{-m}，故 $\mathbb{E}[Y] = m + 2^m \cdot 2^{-m} = m + 1$。

(b) $2^{30} \approx 1.07 \times 10^9$（約十億元）時，$m = 30$，公平入場費 $\mathbb{E}[Y] = 31$ 元。

(c) 「無限的期望值」完全由「機率指數小、金額指數大」的無限尾巴撐起：每層尾巴各貢獻 1 元，層數無限就發散。現實資本一封頂，期望值立刻塌縮到 $m + 1$——資本每翻倍，公平價格只加 1 元（對數成長），即使賭場富可敵國，公平入場費也不過幾十元，與直覺報價同一量級。教訓：重尾分布下，期望值可能是被稀有巨額事件撐起的誤導性摘要，中位數往往更誠實。

練習 3.10 詳解　參考程式與輸出：

```python
import numpy as np
rng = np.random.default_rng(0)
N = 1_000_000

# (a) 骰子：模擬 vs 理論 (3.5 與 35/12 = 2.9167)
rolls = rng.integers(1, 7, size=N)
print(round(rolls.mean(), 4), round(rolls.var(), 4))

# (b) 幾何分布：第一次成功所需次數，理論均值 1/p = 6
p = 1/6
draws = rng.geometric(p, size=N)
print(round(draws.mean(), 4))            # 約 6.0

# (c) 批次平均的標準差 vs 理論 sigma / sqrt(B)
sigma = np.sqrt(35/12)
for B in [4, 16, 64, 256]:
    m = rng.integers(1, 7, size=(100_000, B)).mean(axis=1)
    print(B, round(m.std(), 4), round(sigma/np.sqrt(B), 4))
```

預期輸出（不同隨機種子有 10^{-3} 量級的波動）：(a) 約 3.4997 與 2.9166，吻合理論值 3.5 與 2.9167。(b) 約 6.00，印證 $\mathbb{E}[X] = 1/p$。(c) 四行輸出從 (4, 0.854, 0.8539) 一路到 (256, 0.107, 0.1067)，模擬標準差緊貼理論曲線 $\sigma/\sqrt{B}$：B 每翻 4 倍、雜訊減半，親眼見證練習 3.7 的報酬遞減律。

下一章：常見的機率分布——伯努利、二項、常態與指數，機率世界的標準零件庫。