機率與統計：從入門到精通

第 4 章　常見機率分布

Common Probability Distributions

難度：核心

翻開任何一本統計手冊的附錄，你會看到數十種機率分布，彷彿一座需要死背的動物園。好消息是：我們建模的多數隨機性，其實只需要**少數幾個分布**就能描述。更好的消息是：每個分布都源自一個簡單、反覆出現的「**故事**」——一次擲硬幣、固定時段內罕見事件的計數、許多微小效應之和。學會這些故事，你看到問題就能一眼認出該用哪個分布。本章是常見離散與連續分布的實地指南（field guide），並特別聚焦統計與機器學習中最重要的**高斯分布**：它為何無所不在，以及各分布之間如何交織成一張關係網。

學習目標

讀完本章後，你將能夠：

- 說出每個常見分布背後的「故事」，面對實際情境能選出合適的分布並說明理由。
- 寫出白努利、二項、幾何、卜瓦松、均勻、指數、常態、Beta 等分布的 PMF / PDF，推導代表性的均值與變異數，並運用無記憶性與 z 分數進行計算。
- 解釋高斯分布無所不在的三大理由（中央極限定理、最大熵、解析便利），並熟練使用 68–95–99.7 法則。
- 描繪分布之間的關係網：相加、取極限、離散對應連續、共軛先驗。
- 用 `scipy.stats` 查詢 PMF / PDF / CDF，並以模擬驗證理論計算。

核心觀念

分布 = 故事。每個常見分布都是對某一類隨機情境的「標準答案」：白努利 = 一次是 / 否試驗、二項 = 固定次數中的成功數、幾何 = 等到第一次成功、卜瓦松 = 罕見事件的計數、指數 = 連續的等待、高斯 = 許多微小效應之和。**選分布不是背公式，而是認故事**：先問「資料是由哪種隨機機制產生的？」，答案自然指向正確的分布——在機器學習中，這個選擇將進一步決定你的損失函數（第 7 章）。

4.1 離散分布：從一次擲硬幣到罕見事件的計數

第 3 章教我們用機率質量函數（PMF）描述離散隨機變數；本節的五位主角共用同一個基本零件——白努利試驗——並依「問什麼問題」而分化：問成功幾次是二項、問等多久是幾何、把試驗切到無窮細是卜瓦松、把兩類推廣成 K 類是類別分布。

4.1.1 白努利分布：一次是／否試驗

一切從最簡單的隨機試驗開始：只有兩種結果的**單次**試驗——正面或反面、點擊或不點擊、垃圾郵件或正常郵件。慣例上把我們關心的結果編碼為 1（成功）、另一個編碼為 0（失敗），唯一的參數是成功機率 p。

定義 4.1　白努利分布 Bernoulli(p)

若隨機變數 X 只取兩個值，$\mathbb{P}(X=1)=p$（成功）、$\mathbb{P}(X=0)=1-p$（失敗），則稱 X 服從參數為 p 的**白努利分布**（Bernoulli distribution），記作 $X\sim\mathrm{Ber}(p)$。其 PMF 可寫成一條式子：

$$\mathbb{P}(X=k)=p^{k}(1-p)^{1-k},\qquad k\in\{0,1\}.$$

均值與變異數：

$$\mathbb{E}[X]=p,\qquad \mathrm{Var}(X)=p(1-p).$$

推導只需兩行：$\mathbb{E}[X]=0\cdot(1-p)+1\cdot p=p$；又因 X 只取 0 與 1，有 $X^2=X$，故 $\mathbb{E}[X^2]=p$，$\mathrm{Var}(X)=p-p^2=p(1-p)$。注意變異數在 $p=1/2$ 時最大——五五波是最難預測的局面；p 靠近 0 或 1 時變異數趨近 0，結果幾乎確定。

白努利是**二元分類的原子**：訓練集中一筆有標籤的資料（這封郵件是否為垃圾郵件）就是一次白努利結果，而 logistic 迴歸或神經網路輸出的 $\hat{p}$，正是對白努利參數的估計。它也是本章所有離散分布的積木：接下來的二項、幾何、卜瓦松，都是把白努利試驗以不同方式組裝的結果。

4.1.2 二項分布：固定次數試驗中的成功數

把同一枚硬幣**獨立**地擲 n 次，數一數成功幾次——這就是二項分布的全部故事。關鍵字有二：**固定的試驗次數** n，以及**每次相同且獨立**的成功率 p。

定義 4.2　二項分布 Binomial(n,p)

獨立重複成功率為 p 的白努利試驗 n 次，令 X 為成功總次數，則 X 服從**二項分布**（binomial distribution），記 $X\sim\mathrm{Bin}(n,p)$，其 PMF 為

$$\mathbb{P}(X=k)=\binom{n}{k}p^{k}(1-p)^{n-k},\qquad k=0,1,\dots,n.$$

公式拆成兩塊來讀：$p^{k}(1-p)^{n-k}$ 是「某一種特定順序恰好 k 次成功、$n-k$ 次失敗」的機率（獨立相乘）；$\binom{n}{k}$ 數的是這 k 次成功可以落在哪些位置（第 1 章的組合）。由二項式定理，$\sum_{k=0}^{n}\binom{n}{k}p^{k}(1-p)^{n-k}=\big(p+(1-p)\big)^{n}=1$，機率確實加總為 1——「二項」之名正是這麼來的。

命題 4.3　二項分布的均值與變異數

若 $X \sim \mathrm{Bin}(n,p)$，則

$$\mathbb{E}[X] = np, \qquad \mathrm{Var}(X) = np(1-p).$$

證明　把 X 拆成指示變數之和 $X = X_1 + \cdots + X_n$，其中 $X_i \sim \mathrm{Ber}(p)$ 表示第 i 次是否成功。由期望值的**線性性**（第 3 章，不需要獨立性），$\mathbb{E}[X] = \sum_{i} \mathbb{E}[X_i] = np$；由**獨立性**，變異數可加（第 3 章），$\mathrm{Var}(X) = \sum_{i} \mathrm{Var}(X_i) = np(1-p)$。 ■

「拆成指示變數」的手法比結果本身更值得學——它把看似要硬算的級數化成一行。練習 4.6 請你重走一遍，並想清楚哪一步需要獨立性、哪一步不需要。

例 4.1　品管抽驗

某生產線的瑕疵率為 5%。品管人員隨機抽驗 20 件產品（視為獨立），令 X 為其中的瑕疵品數。求 (a) 完全沒有瑕疵品的機率；(b) 至多一件瑕疵品的機率；(c) X 的均值與標準差。

解　$X \sim \mathrm{Bin}(20, 0.05)$。

(a) $\mathbb{P}(X=0) = (0.95)^{20} \approx 0.3585$。

(b) $\mathbb{P}(X=1) = \binom{20}{1}(0.05)(0.95)^{19} = (0.95)^{19} \approx 0.3774$，故 $\mathbb{P}(X \le 1) \approx 0.3585 + 0.3774 = 0.7358$。

(c) $\mathbb{E}[X] = 20 \times 0.05 = 1$，$\mathrm{Var}(X) = 20 \times 0.05 \times 0.95 = 0.95$，標準差 $\sqrt{0.95} \approx 0.97$。

順帶注意：瑕疵率僅 5%，但「至少中一件」的機率高達 $1 - 0.3585 = 0.6415$——罕見事件重複夠多次就不再罕見，這個直覺在 4.1.4 小節會長成卜瓦松分布。

二項是機器學習裡「數成功」的通用模型：n 位訪客中的轉換數、n 筆預測中的答對數、n 件產品中的瑕疵數。特別地，模型在 n 筆測試資料上答對的題數就是 $\mathrm{Bin}(n, \text{準確率})$——第 8 章將據此替準確率掛上誤差棒。

4.1.3　幾何分布：等到第一次成功

同一台白努利機器，換一個問題：不固定試驗次數，而是**一直試到成功為止**，問第一次成功發生在第幾次——「離散版的等待時間」。

定義 4.4　幾何分布 Geometric(p)

獨立重複成功率為 p 的白努利試驗，令 X 為**第一次成功發生時的試驗次數**，則 X 服從**幾何分布**（geometric distribution），記 $X \sim \mathrm{Geom}(p)$：

$$\mathbb{P}(X = k) = (1-p)^{k-1} p, \qquad k = 1, 2, 3, \ldots$$

尾機率為 $\mathbb{P}(X > k) = (1-p)^{k}$（前 k 次全部失敗）。均值與變異數：

$$\mathbb{E}[X] = \frac{1}{p}, \qquad \mathrm{Var}(X) = \frac{1-p}{p^2}.$$

PMF 的讀法：前 $k-1$ 次全失敗、第 k 次成功，獨立相乘。注意另一種慣例把 X 定義為「第一次成功前的失敗次數」（支撐從 0 開始），查表或用套件前務必確認；`scipy.stats.geom` 採用本書「試驗次數」的慣例。

均值 $1/p$ 有個漂亮的「第一步分析」推導：第一次試驗要嘛成功（機率 p，此時 $X=1$），要嘛失敗（機率 $1-p$，已花 1 次，且一切重新開始），故

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

成功率 2% 的抽卡平均要抽 50 次——倒數關係直觀好記，但小心：幾何分布**右偏**，均值被長尾拉高，「抽滿 50 次仍落空」的機率其實還有三分之一以上（練習 4.3 會算給你看）。推導中「失敗後一切重來」正是幾何分布的招牌：**無記憶性**（memorylessness），$\mathbb{P}(X>m+n \mid X>m) = \mathbb{P}(X>n)$——已連續失敗 m 次，對「還要再等多久」毫無影響。證明留作練習 4.8；連續版本（指數分布）將在例 4.4 登場。

直觀理解

二項與幾何是**同一台白努利機器上的兩個對偶問題**：二項固定試驗次數 n、問成功幾次；幾何固定成功次數（1 次）、問要試幾回。拿到題目先自問：「固定的是次數，還是目標？」把目標推廣到「第 r 次成功」就得到**負二項分布**（negative binomial distribution）。

4.1.4 卜瓦松分布：罕見事件的計數器

客服電話、網站點擊、基因突變、一頁書裡的錯字——這些計數有共同特徵：事件**罕見**、彼此**獨立**、發生在連續的時間或空間中，而且**沒有自然的試驗次數上限**（你沒辦法說一小時裡有幾次「接到電話的機會」）。卜瓦松分布正是為這種「無 n 可言」的計數而生，唯一的參數是平均發生次數 λ。

定義 4.5　卜瓦松分布 Poisson(λ)

若 X 的 PMF 為

$$\mathbb{P}(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}, \qquad k = 0,1,2,\dots$$

其中 $\lambda>0$ 為固定區間內事件的**平均發生次數**，則稱 X 服從**卜瓦松分布**（Poisson distribution），記 $X\sim \mathrm{Poi}(\lambda)$。其均值與變異數相等：

$$\mathbb{E}[X] = \mathrm{Var}(X) = \lambda.$$

機率加總為 1 靠的是泰勒級數 $\sum_{k\ge 0}\lambda^k/k! = e^{\lambda}$。均值的推導只有一行——關鍵是 $k/k! = 1/(k-1)!$：

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k\,\frac{\lambda^k e^{-\lambda}}{k!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda}e^{\lambda} = \lambda.$$

變異數同為 λ 的推導是練習 4.7。「均值 = 變異數」是卜瓦松的指紋，稍後會看到它是二項的 $np(1-p)$ 在 $p\to 0$ 極限下留下的遺跡。

例 4.2　客服中心的忙碌時刻

某客服中心平均每小時接到 3 通投訴電話，假設來電符合卜瓦松模型。求一小時內 (a) 恰好接到 3 通的機率；(b) 至少接到 2 通的機率。

解　令 $X \sim \mathrm{Poi}(3)$，其中 $e^{-3} \approx 0.049787$。

(a) $\mathbb{P}(X=3) = \dfrac{3^3 e^{-3}}{3!} = 4.5\, e^{-3} \approx 0.2240$。

(b)「至少 2 通」包含無限多種情況，直接加總算不完；改用**補集**（第 1 章的老朋友）：

$$\mathbb{P}(X \geq 2) = 1 - \mathbb{P}(X=0) - \mathbb{P}(X=1) = 1 - e^{-3} - 3e^{-3} = 1 - 4e^{-3} \approx 0.8009.$$

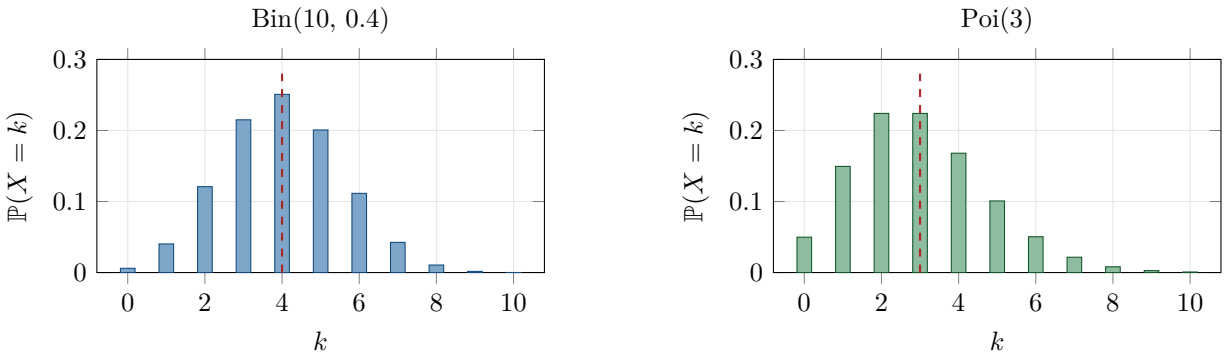


圖 4.1：左：Bin(10, 0.4) 的 PMF，質量集中在均值 $np = 4$（虛線）附近，因 $p < 0.5$ 略向右偏。右：Poi(3) 的 PMF；λ 為整數時 $\mathbb{P}(X=\lambda-1) = \mathbb{P}(X=\lambda)$，故 $k = 2, 3$ 兩根長條等高並列眾數。兩者形狀相似並非巧合——見定理 4.6。

定理 4.6　罕見事件定律：二項的卜瓦松極限

設 $X_n \sim \mathrm{Bin}(n, p_n)$。若 $n \to \infty$、$p_n \to 0$ 且 $np_n = \lambda$ 保持固定，則對每個固定的 k，

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

即二項分布收斂到卜瓦松分布 $\mathrm{Poi}(\lambda)$。（證明見練習 4.10。）

直觀理解

把一小時切成 $n = 3600$ 個一秒的小格子，每一秒「有沒有電話進來」近似一次成功率極低的白努利試驗（$p = \lambda/n$），整小時的電話數就是 $\mathrm{Bin}(n, \lambda/n)$。格子切得越細，近似越精確，極限就是卜瓦松——**卜瓦松是「無窮多次、無窮小機率」的二項**。n 已被送到無窮大，所以它沒有參數 n，只留下平均次數 λ。

例 4.3　卜瓦松近似有多準？

某本書共 100 頁，每頁獨立地以機率 0.03 含有錯字（每頁至多計一次）。含錯字的頁數 $X \sim \mathrm{Bin}(100, 0.03)$。用 $\lambda = np = 3$ 的卜瓦松近似它，並逐點比較。

解　兩個 PMF 在 $k = 0, \dots, 6$ 的值如下（涵蓋約 97% 的機率質量）：

k	0	1	2	3	4	5	6
Bin(100, 0.03)	0.0476	0.1471	0.2252	0.2275	0.1706	0.1013	0.0496
Poi(3)	0.0498	0.1494	0.2240	0.2240	0.1680	0.1008	0.0504

逐點差距最大僅約 0.0035。經驗法則：$n \geq 100$ 且 $p \leq 0.05$ 時卜瓦松近似已相當精確；p 越小、n 越大越準。好處是只剩一個參數、不涉及巨大的組合數，在 n 極大的場合（如「全臺一天的信用卡盜刷筆數」）尤其方便。

常見陷阱

二項還是卜瓦松？（1）**有自然上限用二項，沒有就用卜瓦松**：「100 位訪客中點擊廣告的人數」最多 100 人，是 $\mathrm{Bin}(100, p)$；「一小時內的總點擊數」沒有天花板，用 $\mathrm{Poi}(\lambda)$。n 大、p 小時兩者幾乎相同（定理 4.6），但 p 不小時卜瓦松會明顯失準——二項的變異數 $np(1-p)$ 嚴格小於均值，卜瓦松卻硬說兩者相等。（2）λ **必須隨觀察窗縮放**：平均每分鐘 2 通電話，5 分鐘內的通數是 $\mathrm{Poi}(10)$ 而非 $\mathrm{Poi}(2)$——使用前先確認 λ 對應的區間長度與題目一致。

4.1.5 類別分布與多項分布：從二選一到多選一

白努利處理「二選一」，但真實世界常是「K 選一」：手寫數字是 0 到 9 的其中一個、下一個詞是整個詞彙表中的任何一個。把白努利與二項各自推廣到 K 類，就得到類別分布與多項分布。

定義 4.7　類別分布與多項分布

設單次試驗有 K 種互斥結果，機率分別為 $p_1, \dots, p_K$（$p_k \geq 0$，$\sum_k p_k = 1$）。

- **類別分布**（categorical distribution）$\mathrm{Cat}(p_1, \dots, p_K)$：單次試驗的結果 $X \in \{1, \dots, K\}$，$\mathbb{P}(X = k) = p_k$。它是白努利的 K 類推廣（$K = 2$ 時退回白努利）。
- **多項分布**（multinomial distribution）$\mathrm{Mult}(n; p_1, \dots, p_K)$：獨立重複 n 次，令 N_k 為結果 k 出現的次數，則

$$\mathbb{P}(N_1 = n_1, \dots, N_K = n_K) = \frac{n!}{n_1! \cdots n_K!} p_1^{n_1} \cdots p_K^{n_K}, \qquad \sum_k n_k = n.$$

  它是二項的 K 類推廣；每個邊際皆為二項：$N_k \sim \mathrm{Bin}(n, p_k)$。

機器學習慣用 **one-hot 編碼**表示類別結果：$Y = (0, \dots, 1, \dots, 0)$，1 在第 X 個位置，此時 $\mathbb{E}[Y] = (p_1, \dots, p_K)$——期望值就是機率向量本身。softmax 層輸出的正是一個合法的類別分布，所以類別分布是**每個 softmax 分類器的目標分布**：把標籤建模為類別分布並最大化概似，損失

恰好是交叉熵（第 7 章）。多項分布則是詞袋（bag-of-words）模型的主角：一篇 n 個詞的文件，其詞頻向量就是一次 $\mathrm{Mult}(n;\, p_{1:K})$ 的抽樣。

表 4.1: 離散分布總覽：故事、PMF、均值與變異數。

分布	PMF $\mathbb{P}(X=k)$	均值	變異數	一句話故事
白努利 $\mathrm{Ber}(p)$	$p^k(1-p)^{1-k}$，$k\in\{0,1\}$	p	$p(1-p)$	一次是 / 否試驗
二項 $\mathrm{Bin}(n,p)$	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$	n 次試驗的成功數
幾何 $\mathrm{Geom}(p)$	$(1-p)^{k-1}p$，$k\geq 1$	$1/p$	$(1-p)/p^2$	等到第一次成功
卜瓦松 $\mathrm{Poi}(\lambda)$	$\lambda^k e^{-\lambda}/k!$	λ	λ	罕見事件的計數
類別 $\mathrm{Cat}(p_{1:K})$	p_k，$k\in\{1,\dots,K\}$	—	—	K 選一

4.2 連續分布：從均勻亂數到鐘形曲線

離散分布數「次數」，連續分布量「大小」：等待時間、身高、誤差、比率。回顧第 3 章：連續隨機變數由機率密度函數（PDF）f 描述，$\mathbb{P}(a\leq X\leq b)=\int_a^b f(x)\,\mathrm{d}x$，任何單點的機率都是 0。故事依然是主角。

4.2.1 均勻分布：範圍內等可能

最無知也最誠實的分布：只知道值落在 $[a,b]$ 裡，此外毫無偏好。密度是常數，機率就是長度的比例。

定義 4.8　均勻分布 Uniform(a,b)

若 X 的 PDF 在區間 $[a,b]$ 上為常數

$$f(x)=\frac{1}{b-a},\qquad a\leq x\leq b$$

（區間外為 0），則稱 X 服從**均勻分布**（uniform distribution），記 $X\sim\mathrm{U}(a,b)$。對任何子區間 $[c,d]\subseteq[a,b]$，$\mathbb{P}(c\leq X\leq d)=\frac{d-c}{b-a}$。均值與變異數：

$$\mathbb{E}[X]=\frac{a+b}{2},\qquad \mathrm{Var}(X)=\frac{(b-a)^2}{12}.$$

均值是對稱中心，一望即知；變異數的 12 來自 $\mathbb{E}[X^2]=\frac{b^3-a^3}{3(b-a)}$ 減去均值平方。均勻分布有兩個遠比外表重要的角色：（1）**亂數生成的基礎**——電腦產生的「隨機數」`rng.random()` 就是 $\mathrm{U}(0,1)$ 的樣本，幾乎所有其他分布的抽樣都由它變換而來（例如 $-\ln(U)/\lambda\sim\mathrm{Exp}(\lambda)$）；（2）**無資訊先驗**的原型——對 $[0,1]$ 上的未知機率毫無成見時，貝氏統計常以 $\mathrm{U}(0,1)$（即 $\mathrm{Beta}(1,1)$）作為起點（第 9 章）。

4.2.2 指數分布：連續時間的等待

幾何分布問「第一次成功在第幾次」，指數分布問連續版：「第一件事發生在什麼時刻」。等下一通電話、下一次地震、下一個放射性衰變——凡是以固定平均速率、彼此獨立地隨機發生的事件，其等待時間就是指數分布。

定義 4.9　指數分布 Exponential(λ)

若 $X \geq 0$ 的 PDF 為

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

則稱 X 服從率參數為 λ 的**指數分布**（exponential distribution），記 $X \sim \mathrm{Exp}(\lambda)$。其存活函數（survival function）、均值與變異數為

$$\mathbb{P}(X > t) = e^{-\lambda t}, \qquad \mathbb{E}[X] = \frac{1}{\lambda}, \qquad \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

均值用一次分部積分即得：

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

參數 λ 是「**率**」：每單位時間平均發生 λ 件事，平均等待時間 $1/\lambda$——率越高、等越短。寫程式時小心：NumPy 與 SciPy 的指數分布吃的參數是 `scale` $= 1/\lambda$（均值）而非 λ，這是統計程式最常見的 bug 之一。

指數與卜瓦松是同一枚硬幣的兩面，合稱**卜瓦松過程**（Poisson process）：事件以率 λ 隨機發生時，「長度 t 的窗口內的事件數」服從 $\mathrm{Poi}(\lambda t)$，「相鄰兩件事的間隔」服從 $\mathrm{Exp}(\lambda)$。兩者的橋樑只有一行：$\mathbb{P}(\text{等待} > t) = \mathbb{P}(\text{窗口內零事件}) = e^{-\lambda t}$，正是指數的存活函數。4.5 節的模擬將直接利用這件事：用指數間隔累加出到達時刻，數出來的計數自動服從卜瓦松。

例 4.4　指數分布的無記憶性

設 $X \sim \mathrm{Exp}(\lambda)$。證明對任意 $s, t \geq 0$，

$$\mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t),$$

並解釋這個性質的意義與適用限制。

解　事件 $\{X > s+t\}$ 包含於 $\{X > s\}$，由條件機率的定義（第 2 章）：

$$\mathbb{P}(X > s+t \mid X > s) = \frac{\mathbb{P}(X > s+t)}{\mathbb{P}(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbb{P}(X > t). \qquad \blacksquare$$

意義：「已經等了 s 分鐘」對「還要再等多久」毫無影響——指數分布不累積等待、不老化。適用於**沒有磨損機制**的事件（放射性衰變、卜瓦松到達的間隔）；不適用於會老化或有節奏的事物（輪胎壽命、班距固定的捷運）。推論：無論已等多久，剩餘的期望等待時間永遠還是 $1/\lambda$。

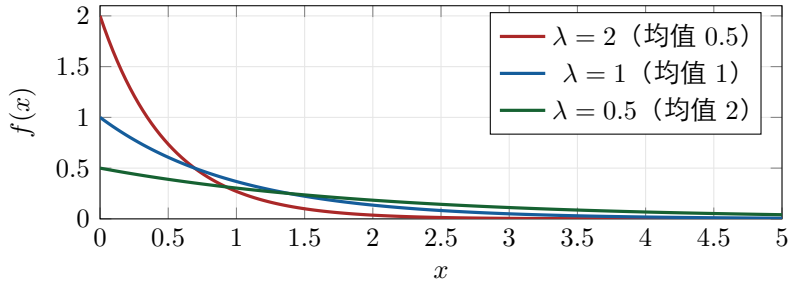


圖 4.2: 指數分布的密度 $f(x)=\lambda e^{-\lambda x}$：率 λ 越大，密度在 0 附近越高、衰減越快，平均等待 $1/\lambda$ 越短；三條曲線下的面積都是 1。

4.2.3 常態（高斯）分布：鐘形曲線

終於輪到主角。身高、測量誤差、考試分數的分布都近似這條鐘形曲線；它憑什麼無所不在，4.3 節專門回答，這裡先把它定義清楚、學會計算。

定義 4.10　常態分布 Normal(μ,σ^2)

若 X 的 PDF 為

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\!\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),\qquad x\in\mathbb{R},$$

則稱 X 服從均值 μ、變異數 σ^2 的**常態分布**（normal distribution），又稱**高斯分布**（Gaussian distribution），記 $X\sim N(\mu,\sigma^2)$。$\mu=0$、$\sigma=1$ 時稱**標準常態**（standard normal），其 CDF 記為 $\Phi(z)$。

幾何性質：曲線對稱於 μ（均值＝中位數＝眾數）；σ 控制寬度；**反曲點恰在** $\mu\pm\sigma$——「一個標準差」在圖上看得見。尾部以 e^{-x^2} 的速度衰減，比指數更快，屬「輕尾」。CDF 沒有初等閉式，所有常態機率都靠標準常態的 Φ 查表或軟體——而標準化保證一張表就夠用。

命題 4.11　標準化與 z 分數

若 $X\sim N(\mu,\sigma^2)$，則

$$Z=\frac{X-\mu}{\sigma}\sim N(0,1),\qquad \text{因此}\qquad \mathbb{P}(X\le x)=\Phi\!\left(\frac{x-\mu}{\sigma}\right).$$

把 x 換算成「距離均值幾個標準差」的數字 $z=(x-\mu)/\sigma$，稱為 z **分數**（z-score）。

實際流程永遠是兩步：算 z 分數、查 Φ。例 4.6 將示範「不查表」的心算版（68–95–99.7 法則），精確版交給 4.5 節的 `norm.cdf`。

4.2.4 Beta 分布：機率的分布

到目前為止，參數 p 都被當成已知常數。但真實世界裡 p 往往正是我們最想知道的未知數：新廣告的點擊率、新藥的治癒率。既然不確定，就給它一個分布——一個定義在 $[0,1]$ 上、專門描述「機率本身」的分布。這就是 Beta。

定義 4.12　Beta 分布 Beta(α,β)

若 $X\in[0,1]$ 的 PDF 為

$$f(x)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)},\qquad B(\alpha,\beta)=\frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)},$$

其中 $\alpha,\beta>0$，則稱 X 服從 **Beta 分布**，記 $X\sim\mathrm{Beta}(\alpha,\beta)$。均值、眾數（需 $\alpha,\beta>1$）與變異數：

$$\mathbb{E}[X]=\frac{\alpha}{\alpha+\beta},\qquad \text{眾數}=\frac{\alpha-1}{\alpha+\beta-2},\qquad \mathrm{Var}(X)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

這裡 Γ 是階乘的推廣（正整數時 $\Gamma(n)=(n-1)!$），$B(\alpha,\beta)$ 只是讓密度積分為 1 的正規化常數。形狀完全由 (α,β) 決定，見圖 4.3——同一族分布可以是平的、山形的、偏斜的，甚至 U 形的。

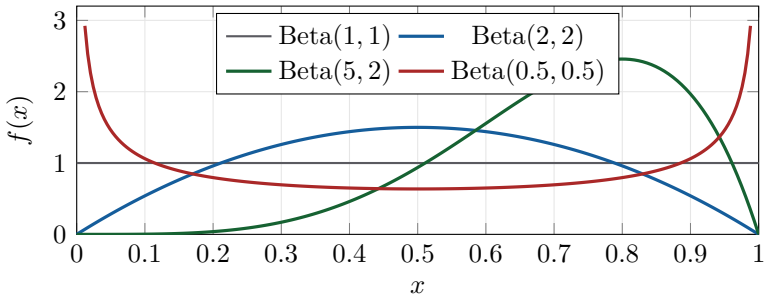


圖 4.3: Beta 分布的四種面貌：Beta$(1,1)$ 是均勻分布；Beta$(2,2)$ 溫和偏好中間；Beta$(5,2)$ 偏向高機率；Beta$(0.5,0.5)$ 呈 U 形、密度在兩端趨向無窮大（圖中截斷顯示）。

例 4.5　把 Beta 讀成「偽計數」

把參數讀成 $\alpha-1$ 次「虛擬成功」與 $\beta-1$ 次「虛擬失敗」（合稱**偽計數**，pseudo-counts），對照圖 4.3 解讀四組參數的形狀，並計算 Beta$(5,2)$ 的均值與眾數。

解

- Beta$(1,1)$：0 成功、0 失敗——毫無資訊，退化為均勻分布 U$(0,1)$。
- Beta$(2,2)$：各 1 次——山形，溫和偏好 $p\approx0.5$，但不排除兩端。
- Beta$(5,2)$：4 成功、1 失敗——右移單峰。均值 $5/7\approx0.714$，眾數 $(5-1)/(5+2-2)=0.8$。
- Beta$(0.5,0.5)$：偽計數是「負的」——U 形，把賭注押在兩個極端（p 接近 0 或 1）。

偽計數的威力在**更新**：以 Beta(α,β) 為先驗、觀察到 s 次成功與 f 次失敗後，後驗恰好是 Beta$(\alpha+s,\ \beta+f)$——真實計數直接加在偽計數上，分布家族不變。這種性質稱為**共軛**（conjugacy）：Beta 是白努利 / 二項參數的**共軛先驗**（conjugate prior），第 9 章完整展開，4.5 節的程式會先讓你看一眼數字。

常見陷阱

密度可以大於 1！機率質量 $\mathbb{P}(X=k)$ 永遠不超過 1，但機率**密度** $f(x)$ 沒有這個限制：$\mathrm{U}(0,0.5)$ 的密度恆為 2；$N(\mu,0.1^2)$ 的峰值約 3.99；圖 4.3 裡 $\mathrm{Beta}(0.5,0.5)$ 的密度在兩端甚至衝向無窮大。這都完全合法：密度是「**每單位長度**的機率」，真正必須等於 1 的是曲線下總面積 $\int f(x)\,\mathrm{d}x$——底越窄，峰就可以越高。同一個誤解的另一面，是把 $f(x)$ 當成 $\mathbb{P}(X=x)$：連續分布中任何單點的機率都是 0，有意義的是區間，$\mathbb{P}(a\le X\le b)=\int_a^b f(x)\,\mathrm{d}x$。

4.2.5 Gamma、卡方與 t 分布：推論的配角

最後三個分布在本章戲份不多，卻是第 8 章統計推論的要角，先認識故事即可。

定義 4.13　Gamma、卡方與 t 分布

- **Gamma 分布** $\mathrm{Gamma}(\alpha,\lambda)$：PDF 為 $f(x)=\dfrac{\lambda^\alpha x^{\alpha-1}e^{-\lambda x}}{\Gamma(\alpha)}$（$x>0$），均值 α/λ、變異數 α/λ^2。當 $\alpha=k$ 為正整數時，它是 k 個獨立 $\mathrm{Exp}(\lambda)$ 之和——「等到第 k 件事發生的總時間」。
- **卡方分布**（chi-squared distribution）χ^2_k：k 個獨立標準常態的平方和 $Z_1^2+\cdots+Z_k^2$ 的分布，等同於 $\mathrm{Gamma}(k/2,\,1/2)$；均值 k、變異數 $2k$。k 稱為**自由度**（degrees of freedom）。
- **t 分布**（Student's t）t_ν：比值 $T=\dfrac{Z}{\sqrt{\chi^2_\nu/\nu}}$（分子分母獨立）的分布。鐘形、對稱於 0，但尾部比常態厚；$\nu\to\infty$ 時趨近 $N(0,1)$。

卡方的均值有個一行推導：$\mathbb{E}[Z^2]=\mathrm{Var}(Z)=1$，由線性性 $\mathbb{E}[\chi^2_k]=k$；它描述「k 個獨立誤差的能量總和」，是樣本變異數與適合度檢定的核心（第 8 章）。t 分布的故事：想用樣本均值推論 μ 卻不知道 σ，只好用樣本標準差 s 代替——多出來的不確定性讓尾巴變厚，厚多少由自由度 ν 控制（$\nu=30$ 時已與常態幾乎重合；極端情形 t_1 是柯西分布，尾巴厚到連均值都不存在）。t 檢定與小樣本信賴區間正是它的主場。

表 4.2: 連續分布總覽：故事、PDF、均值與變異數。

分布	PDF $f(x)$	支撐	均值	變異數	一句話故事
均勻 $\mathrm{U}(a,b)$	$\frac{1}{b-a}$	$[a,b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	範圍內毫無偏好
指數 $\mathrm{Exp}(\lambda)$	$\lambda e^{-\lambda x}$	$[0,\infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	等第一件事發生
常態 $N(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$	$\mathbb{R}$	μ	σ^2	許多小效應之和
$\mathrm{Beta}(\alpha,\beta)$	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$	$[0,1]$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	機率本身的分布
$\mathrm{Gamma}(\alpha,\lambda)$	$\frac{\lambda^\alpha x^{\alpha-1}e^{-\lambda x}}{\Gamma(\alpha)}$	$(0,\infty)$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$	等 k 件事的總時間
卡方 χ^2_k	$=\mathrm{Gamma}(k/2,1/2)$	$(0,\infty)$	k	$2k$	k 個標準常態平方和
t_ν	$\propto(1+x^2/\nu)^{-(\nu+1)/2}$	$\mathbb{R}$	0（$\nu>1$）	$\frac{\nu}{\nu-2}$（$\nu>2$）	厚尾鐘形、小樣本

4.3 高斯為何無所不在

打開任何一篇實驗科學論文，誤差棒假設常態；打開任何一本機器學習教科書，雜訊預設高斯。一個分布憑什麼享有這種地位？三個彼此獨立的理由。

核心觀念

高斯分布稱霸統計與機器學習，原因有三：

1. **中央極限定理**（central limit theorem, CLT；第 6 章）：許多獨立小效應的**和（或平均）**近似高斯，無論每個效應本身服從什麼分布。測量雜訊、聚合行為、mini-batch 的平均梯度，都是這種和。
2. **最大熵**（maximum entropy）：在「均值為 μ、變異數為 σ^2」的所有分布中，高斯的熵最大——它是最分散、**假設最少**的選擇。只知道前兩個動差時，選高斯等於不多說任何沒有根據的話。
3. **數學便利**：高斯家族在線性變換、相加、條件化、邊際化之下都封閉（見下方命題），對數密度是簡單的二次式——這正是「平方誤差損失對應高斯雜訊」的原因（第 7 章）。

加法選中了高斯。擲一顆骰子，點數分布是平的；兩顆之和成三角形；十顆之和已近乎鐘形——**和會遺忘出身**，無論原料是什麼分布，加總夠多份之後都趨向高斯（第 6 章將精確化）。而「許多小效應之和」正是世界的常態：測量誤差、身高、mini-batch 平均梯度都是這種和。高斯不是被人選中的，是被加法選中的。

最大熵是誠實。只知道均值與變異數時，最大熵原理要你選「此約束下最分散、假設最少」的分布，可以證明答案恰是 $N(\mu, \sigma^2)$。同一原理也解釋本章另外兩位成員：給定範圍 $[a, b]$ 時最大熵是均勻，給定正支撐與均值時是指數——三個「預設選項」地位相同：**只編碼已知，其餘保持最大的不可知**。

解析便利。高斯怎麼運算都還是高斯：

命題 4.14　高斯的封閉性

設 $X \sim N(\mu_X, \sigma_X^2)$ 與 $Y \sim N(\mu_Y, \sigma_Y^2)$ 獨立，a, b 為常數。則

$$aX + b \sim N\big(a\mu_X + b,\ a^2\sigma_X^2\big), \qquad X + Y \sim N\big(\mu_X + \mu_Y,\ \sigma_X^2 + \sigma_Y^2\big).$$

更進一步，聯合高斯向量的邊際分布與條件分布仍是高斯（證明見第 5 章）。

且它的對數密度是二次式：$\ln f(x) = -\frac{(x-\mu)^2}{2\sigma^2} + \text{常數}$。第 7 章將看到：把觀測建模為「真值加高斯雜訊」再做最大概似，得到的正是最小平方法——**平方誤差不是任意選擇，而是高斯假設的直接後果**。卡爾曼濾波、高斯過程、變分自編碼器，全都仰賴這種「運算封閉 + 二次對數」的便利。

核心觀念

68–95–99.7 法則（經驗法則）：若 $X \sim N(\mu, \sigma^2)$，則

$$\mathbb{P}(|X-\mu| \le \sigma) \approx 68.3\%, \qquad \mathbb{P}(|X-\mu| \le 2\sigma) \approx 95.4\%, \qquad \mathbb{P}(|X-\mu| \le 3\sigma) \approx 99.7\%.$$

這三個數字值得背下來：實驗報告的「2σ 誤差棒」、民調的 95% 信賴區間（第 8 章）、工廠的 3σ 品管，全都源於此。反過來讀：偏離均值超過 3σ 的觀測，每千次只該出現約 3 次——它要嘛是異常值，要嘛你的模型錯了。

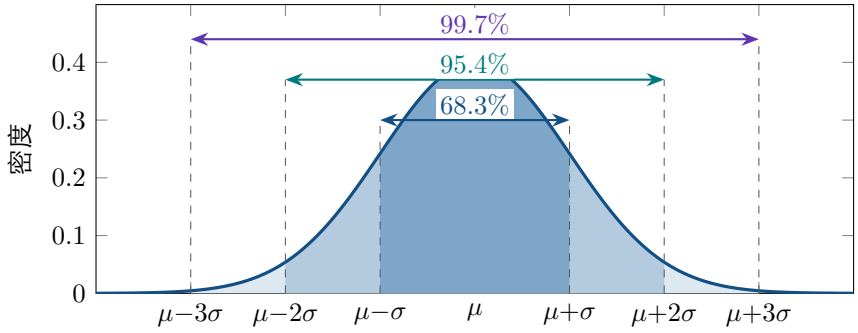


圖 4.4：常態分布的 68–95–99.7 法則：落在 $\mu \pm \sigma$、$\mu \pm 2\sigma$、$\mu \pm 3\sigma$ 內的機率分別約 68.3%、95.4%、99.7%；3σ 之外每側僅約 0.13%。

例 4.6　用 68–95–99.7 法則心算常態機率

假設某國成年男性身高 $X \sim N(172, 6^2)$（單位：公分）。心算下列各值：(a) $\mathbb{P}(166 \le X \le 178)$；(b) $\mathbb{P}(X > 184)$；(c) $\mathbb{P}(X < 154)$；(d) 某大學有 2000 名男學生，期望有幾人高於 184 公分？

解　先把邊界換算成 z 分數（距離均值幾個 σ）：

(a) $166 = \mu - \sigma$、$178 = \mu + \sigma$，故 $\mathbb{P} \approx 68.3\%$。

(b) $184 = \mu + 2\sigma$。$\pm 2\sigma$ 之外約 $1 - 0.954 = 0.046$，由對稱性右尾佔一半：$\mathbb{P}(X > 184) \approx 0.023$。

(c) $154 = \mu - 3\sigma$，左尾約 $(1 - 0.997)/2 = 0.0015$，即千分之 1.5。

(d) 每位學生獨立地以 $p \approx 0.023$ 高於 184 公分，人數服從 $\mathrm{Bin}(2000, 0.023)$，期望約 $2000 \times 0.023 = 46$ 人。注意這一步把常態與二項**串接**了起來——分布不是孤島，這正是下一節的主題。

4.4　分布之間的關係網

本章的分布不是互不相干的詞條，而是一張關係緊密的網（圖 4.5）：箭頭代表相加、取極限、平方、離散對應連續、共軛。把這張網記住，目錄就變成了理解。

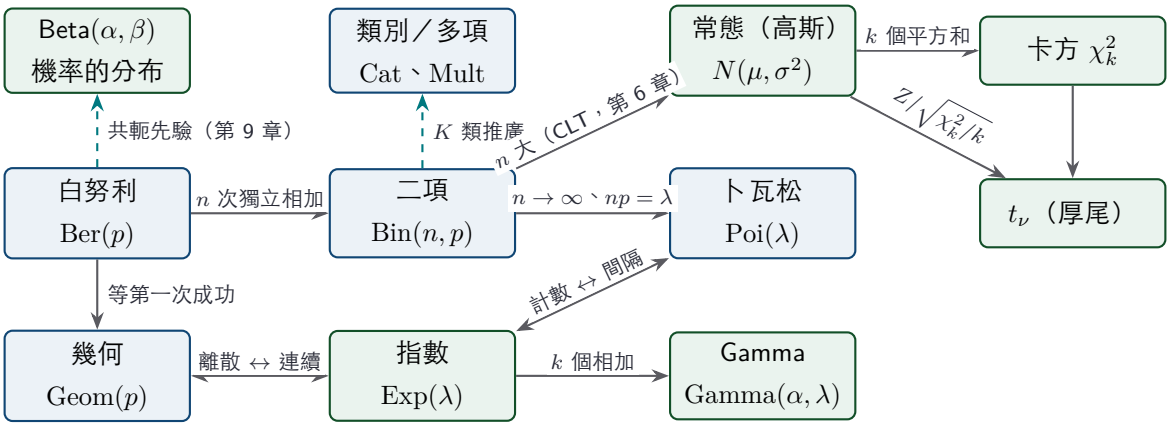


圖 4.5: 分布關係網。藍框為離散分布、綠框為連續分布；虛線箭頭表示推廣或先驗關係。

直觀理解

別把本章讀成「目錄」，要讀成「**地圖**」。知道「白努利相加成二項」，$\mathbb{E} = np$ 就不必背；知道「卜瓦松是罕見事件的二項極限」，就明白它為何均值等於變異數（$np(1-p) \to \lambda$，因為 $p \to 0$）；知道「指數是幾何的連續版」，無記憶性就一點也不神祕。日後遇到新分布（負二項、Dirichlet、對數常態……），第一個問題永遠是：**它掛在這張網的哪裡？**

這如何驅動機器學習／ AI

選分布 = 選模型 = 選損失函數。第 7 章將證明「最大化概似」與「最小化損失」是同一件事的兩面，於是本章每個分布都對應機器學習的一個預設選項：

- 標籤建模為**類別分布**，得到**交叉熵**（cross-entropy）損失——分類的預設；二元版（白努利）就是 binary cross-entropy 與 logistic 迴歸。
- 迴歸目標建模為**高斯**（真值加高斯雜訊），得到**均方誤差**（mean squared error, MSE）——迴歸的預設；換成厚尾的 t 或 Laplace 分布，就得到對異常值更穩健的損失。
- 計數建模為**卜瓦松**，得到卜瓦松迴歸——預測點擊數、理賠數、缺陷數的標準工具。
- **Beta** 與**高斯**分別是白努利比率與高斯均值的**共軛先驗**（第 9 章），支撐貝氏 A/B 測試與 Thompson 抽樣；**白努利**遮罩是 dropout 的機制；**高斯**是 VAE 的潛在先驗、也是擴散模型逐步添加的雜訊；語言模型每一步預測的下一個詞，都是一個巨大的**類別分布**。

選定一個分布，就是在告訴模型：「資料是由這種隨機性生成的。」選對了，損失函數與不確定性估計都自動就位；選錯了（例如把厚尾資料硬塞給高斯），模型會用整個訓練過程替這個錯誤買單。

4.5 Python 實作

照例，動手驗證。scipy.stats 對每個分布提供統一介面：pmf / pdf、cdf、sf（存活函數，即 1 − CDF）、rvs（抽樣）。以下依序驗證代表性計算、68–95–99.7 法則、卜瓦松極限，以及「指數間隔生成卜瓦松計數」的模擬。

```python
import numpy as np
from scipy import stats

rng = np.random.default_rng(42)

# --- 1. pmf / pdf / cdf / sf：查分布的四把鑰匙 ---
print(stats.binom.pmf(4, 10, 0.4))    # 0.2508  圖 4.1 左 k=4 的長條
print(stats.poisson.pmf(3, 3))        # 0.2240  圖 4.1 右 k=3 的長條
print(stats.geom.sf(50, 0.02))        # 0.3642  練習 4.3：50 抽全落空
print(stats.expon.sf(15, scale=10))   # 0.2231  Exp(均值10) 的 P(T>15)
print(stats.norm.cdf(178, 172, 6)
      - stats.norm.cdf(166, 172, 6))  # 0.6827  例 4.6(a)

# --- 2. 68-95-99.7 法則：用 norm.cdf 驗證 ---
for k in [1, 2, 3]:
    p = stats.norm.cdf(k) - stats.norm.cdf(-k)
    print(k, round(p, 4))             # 0.6827 / 0.9545 / 0.9973

# --- 3. 二項的卜瓦松極限：n=100, p=0.03, lam=3（例 4.3）---
for k in range(5):
    pb = stats.binom.pmf(k, 100, 0.03)
    pp = stats.poisson.pmf(k, 3)
    print(k, round(pb, 4), round(pp, 4))

# --- 4. 卜瓦松過程：指數間隔累加 -> 單位時間的到達數 ---
lam = 3.0
gaps = rng.exponential(scale=1/lam, size=(100_000, 30))
counts = (np.cumsum(gaps, axis=1) <= 1.0).sum(axis=1)
print(counts.mean(), counts.var())    # 均值與變異數都約 3
print((counts >= 2).mean())           # 約 0.8009，對照例 4.2(b)

# --- 5. Beta 共軛更新：Beta(2,2) 先驗 + 7 成功 3 失敗 ---
a, b = 2 + 7, 2 + 3                   # 後驗 Beta(9,5)（第 9 章）
print(stats.beta.mean(a, b))          # 0.6429  後驗均值
print((a - 1) / (a + b - 2))          # 0.6667  後驗眾數
```

輸出解讀：第 1 段五個數字分別對應圖 4.1、練習 4.3、練習 4.4 與例 4.6 的手算結果；第 3 段兩欄幾乎相同，複現例 4.3 的對照表。第 4 段最有意思：只用指數亂數就造出了卜瓦松計數——均值與變異數都接近 3（卜瓦松的指紋），$\mathbb{P}(N \geq 2) \approx 0.8009$ 與例 4.2 的補集計算吻合。模

擬是最誠實的驗算：手算若有錯，十萬次抽樣不會替你圓謊。

本章總結

- **每個分布是一個故事**：白努利 = 一次是 / 否試驗；二項 = n 次試驗的成功數；幾何 = 等到第一次成功；卜瓦松 = 固定區間內罕見事件的計數；均勻 = 範圍內等可能；指數 = 連續的等待時間；高斯 = 許多微小效應之和；Beta = 機率本身的分布。**選分布就是認故事。**
- 代表性結果與手法：二項的 $\mathbb{E} = np$、$\mathrm{Var} = np(1-p)$（指示變數 + 線性性）；卜瓦松的 $\mathbb{E} = \mathrm{Var} = \lambda$（級數）；幾何與指數具**無記憶性**；「至少」型問題用補集。
- **高斯稱霸**的三大理由：中央極限定理（和 / 平均趨向高斯）、最大熵（給定均值與變異數下假設最少）、解析便利（線性運算封閉、對數密度為二次式）。務必記住 **68–95–99.7**。
- 分布是**相連的網**而非孤立條目：白努利相加成二項；二項的極限是卜瓦松（罕見事件）與高斯（CLT）；指數是幾何的連續版、與卜瓦松合組卜瓦松過程；高斯平方和成卡方；Beta 是白努利比率的共軛先驗。
- **選分布 = 選損失**：類別 → 交叉熵、高斯 → MSE、卜瓦松 → 卜瓦松迴歸。實作套路：`scipy.stats` 的 `pmf` / `cdf` / `sf` / `rvs`；密度可以大於 1，機率是面積；模擬是最誠實的驗算。

4.6 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 4.1 (基礎)　某籃球員罰球命中率為 0.8，投 10 球，各球獨立。令 X 為命中數。(a) X 服從什麼分布？寫出它的 PMF。(b) 求恰好命中 8 球的機率。(c) 求至少命中 9 球的機率。(d) 求 X 的均值、變異數與標準差。

練習 4.2 (基礎)　某電商網站尖峰時段平均每分鐘出現 2 筆訂單，假設訂單到達符合卜瓦松模型。(a) 求某一分鐘內沒有任何訂單的機率。(b) 求某一分鐘內至少 3 筆訂單的機率。(c) 接下來 5 分鐘內的訂單數服從什麼分布？「至少一筆」的機率是多少？

練習 4.3 (基礎)　某手機遊戲抽卡，每抽獨立，抽中 SSR 的機率為 0.02。令 X 為第一次抽中 SSR 所需的抽數。(a) 寫出 X 的分布與 PMF。(b) 求前 10 抽全部落空的機率。(c) 求期望抽數。(d) 求抽了 50 抽（恰為期望值）仍未抽中的機率，並評論這個數字說明了什麼。

練習 4.4 (基礎)　某路口計程車的到達符合卜瓦松過程，平均每 10 分鐘一輛。令 T 為你開始等待到第一輛車出現的時間。(a) T 服從什麼分布？求等待超過 15 分鐘的機率。(b) 求 5 分鐘內等到車的機率。(c) 求等待時間的中位數，並解釋它為何小於均值 10 分鐘。(d) 你已經等了 10 分鐘還沒有車，「接下來還要再等超過 15 分鐘」的機率是多少？

練習 4.5 (基礎)　某大型測驗的分數近似 $N(70,\,8^2)$。用 68–95–99.7 法則心算：(a) 分數落在 62 到 78 之間的比例；(b) 高於 86 分的比例；(c) 落在 54 到 86 之間的比例；(d) 要贏過約 99.85% 的考生，至少需要幾分？

練習 4.6 (進階)　設 $X \sim \mathrm{Bin}(n,p)$。(a) 把 X 寫成指示變數之和，利用期望值的線性性證明 $\mathbb{E}[X] = np$。(b) 證明單一白努利變數的變異數為 $p(1-p)$，並利用獨立性推出 $\mathrm{Var}(X) = np(1-p)$。(c) 若各次試驗仍各自服從 $\mathrm{Ber}(p)$ 但彼此不獨立（例如正相關），(a) 與 (b) 的結論哪個仍然成立？哪個會失效？

練習 4.7 (進階)　設 $X \sim \mathrm{Poi}(\lambda)$。本章正文已推導 $\mathbb{E}[X] = \lambda$。(a) 用級數證明 $\mathbb{E}[X(X-1)] = \lambda^2$。(b) 由此推出 $\mathrm{Var}(X) = \lambda$。(c) 說明為何「樣本均值 $\approx$ 樣本變異數」可當作卜瓦松模型的快速健檢，以及真實計數資料常見的偏離方向。

練習 4.8 (進階)　設 $X \sim \mathrm{Geom}(p)$。(a) 證明 $\mathbb{P}(X > k) = (1-p)^k$。(b) 證明無記憶性：$\mathbb{P}(X > m+n \mid X > m) = \mathbb{P}(X > n)$。(c) 指數分布滿足完全相同的等式（例 4.4）。說明在什麼意義下指數是幾何的「連續表親」，並用無記憶性解釋為何「已經連輸十把」對下一把毫無幫助（賭徒謬誤）。

練習 4.9 (進階)　替下列每個情境選擇最合適的分布（從本章出場者中挑選），寫出參數的意義，用該分布的「故事」說明理由，並在必要時指出模型可能失準之處：(a) 一小時內通過某收費站的車輛數；(b) 100 位收到促銷信的顧客中實際下單的人數；(c) 一顆不會老化的電子零件從現在到故障的時間；(d) 擲一顆公正骰子，直到第一次出現 6 所需的次數；(e) 1000 位成年女性的身高；(f) 一個新上線廣告的未知點擊率。

練習 4.10 (挑戰)　證明定理 4.6（罕見事件定律）：設 $X_n \sim \mathrm{Bin}(n,\,\lambda/n)$，$\lambda > 0$ 固定。證明對每個固定的 k，

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

提示：把 PMF 拆成數個因子分別取極限，可使用 $\left(1-\lambda/n\right)^n \to e^{-\lambda}$。並順帶說明：這個極限如何解釋卜瓦松「均值 = 變異數」的特徵？

練習 4.11 (挑戰 · 程式)　用 `scipy.stats` 與 NumPy 完成：(a) 對 $k = 0, \dots, 6$ 印出 $\mathrm{Bin}(100, 0.03)$ 與 $\mathrm{Poi}(3)$ 的 PMF，驗證例 4.3 的表格；(b) 模擬卜瓦松過程：產生大量 $\mathrm{Exp}(3)$ 間隔並累加，統計單位時間到達數的均值、變異數與 $\mathbb{P}(N \ge 2)$，對照例 4.2；(c) 用 `norm.cdf` 驗證 68–95–99.7 法則。

4.7 練習題詳解

練習 4.1 詳解　(a) 10 次**固定次數、獨立且同命中率**的是 / 否試驗，數成功次數，故 $X \sim \mathrm{Bin}(10,\, 0.8)$，$\mathbb{P}(X = k) = \binom{10}{k}(0.8)^k(0.2)^{10-k}$（$k = 0, \dots, 10$）。
(b) $\mathbb{P}(X = 8) = \binom{10}{8}(0.8)^8(0.2)^2 = 45 \times 0.16777 \times 0.04 \approx 0.3020$。
(c) 至少 9 球即 $k = 9$ 或 10：$\mathbb{P}(X \ge 9) = \binom{10}{9}(0.8)^9(0.2) + (0.8)^{10} \approx 0.2684 + 0.1074 = 0.3758$。
(d) $\mathbb{E}[X] = np = 8$，$\mathrm{Var}(X) = np(1-p) = 1.6$，標準差 $\sqrt{1.6} \approx 1.26$——典型表現約「8 ± 1.3 球」。

練習 4.2 詳解　令 X 為一分鐘內的訂單數，$X \sim \mathrm{Poi}(2)$。
(a) $\mathbb{P}(X = 0) = e^{-2} \approx 0.1353$。
(b)「至少 3 筆」用補集：$\mathbb{P}(X \ge 3) = 1 - e^{-2}\big(1 + 2 + \frac{2^2}{2}\big) = 1 - 5e^{-2} \approx 0.3233$。
(c) 卜瓦松的 λ 隨觀察窗長度縮放（4.1.4 小節的陷阱提醒）：5 分鐘內的訂單數 $Y \sim \mathrm{Poi}(10)$，$\mathbb{P}(Y \ge 1) = 1 - e^{-10} \approx 0.99995$，幾乎必然發生。

練習 4.3 詳解　(a) 每抽是獨立、同機率 $p = 0.02$ 的白努利試驗，X 是第一次成功的試驗數，故 $X \sim \mathrm{Geom}(0.02)$，$\mathbb{P}(X = k) = (0.98)^{k-1}(0.02)$（$k = 1, 2, \dots$）。
(b) 前 10 抽全部落空即 $\mathbb{P}(X > 10) = (0.98)^{10} \approx 0.8171$。
(c) $\mathbb{E}[X] = 1/p = 50$ 抽。
(d) $\mathbb{P}(X > 50) = (0.98)^{50} = e^{50 \ln 0.98} \approx 0.3642$。抽滿「期望值」50 抽，仍有超過三分之一的機率一無所獲——幾何分布右偏，均值被長尾拉高；中位數其實只有 35 抽（$(0.98)^{34} \approx 0.503 > 0.5 > (0.98)^{35}$）。

練習 4.4 詳解　(a) 卜瓦松過程的等待時間服從指數分布。平均每 10 分鐘一輛表示率 $\lambda = 0.1$（每分鐘），故 $T \sim \mathrm{Exp}(0.1)$，$\mathbb{P}(T > 15) = e^{-1.5} \approx 0.2231$。
(b) $\mathbb{P}(T \le 5) = 1 - e^{-0.5} \approx 0.3935$。
(c) 中位數 m 滿足 $e^{-0.1m} = 0.5$，故 $m = 10 \ln 2 \approx 6.93$ 分鐘。中位數小於均值，因為指數分布**右偏**：多數等待很短，偶爾的漫長等待把平均拉高。
(d) 由無記憶性（例 4.4），已等 10 分鐘不改變剩餘等待的分布：$\mathbb{P}(T > 25 \mid T > 10) = \mathbb{P}(T > 15) = e^{-1.5} \approx 0.2231$，與 (a) 完全相同。

練習 4.5 詳解　$\mu=70$、$\sigma=8$，逐一換算 z 分數：
(a) $62=\mu-\sigma$、$78=\mu+\sigma$，比例約 68.3%。
(b) $86=\mu+2\sigma$。$\pm2\sigma$ 之外共約 4.6%，右尾佔一半：約 2.3%。
(c) $54=\mu-2\sigma$、$86=\mu+2\sigma$，比例約 95.4%。
(d) 「贏過 99.85%」表示落在最右端約 0.15%。由法則，3σ 之外兩尾合計約 0.3%、單尾約 0.15%，故需要 $\mu+3\sigma=70+24=94$ 分以上。

練習 4.6 詳解　(a) 令 $X_i\sim\mathrm{Ber}(p)$ 為第 i 次試驗的指示變數，則 $X=X_1+\cdots+X_n$。期望值的線性性對**任何**隨機變數都成立（不需獨立性）：$\mathbb{E}[X]=\sum_{i=1}^{n}\mathbb{E}[X_i]=np$。
(b) 因 $X_i^2=X_i$（0 與 1 的平方是自身），$\mathbb{E}[X_i^2]=p$，故 $\mathrm{Var}(X_i)=p-p^2=p(1-p)$。各 X_i **獨立**時變異數可加：$\mathrm{Var}(X)=\sum_i\mathrm{Var}(X_i)=np(1-p)$。
(c) (a) 仍成立：線性性不需要獨立性，即使試驗相關，$\mathbb{E}[X]=np$ 依舊。(b) 會失效：相關時 $\mathrm{Var}(X)=\sum_i\mathrm{Var}(X_i)+\sum_{i\neq j}\mathrm{Cov}(X_i,X_j)$，正相關（例如疾病的家庭群聚感染）會使變異數大於 $np(1-p)$——即**過度離散**（overdispersion），二項模型會低估波動。

練習 4.7 詳解　(a) 級數推導，關鍵是 $k(k-1)/k!=1/(k-2)!$（$k\geq2$；$k=0,1$ 的項為 0）：

$$\mathbb{E}[X(X-1)]=\sum_{k=2}^{\infty}k(k-1)\frac{\lambda^k e^{-\lambda}}{k!}=\lambda^2e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}=\lambda^2e^{-\lambda}e^{\lambda}=\lambda^2.$$

(b) 由 $\mathbb{E}[X(X-1)]=\mathbb{E}[X^2]-\mathbb{E}[X]$ 得 $\mathbb{E}[X^2]=\lambda^2+\lambda$，故 $\mathrm{Var}(X)=\lambda^2+\lambda-\lambda^2=\lambda$。 ■
(c) 「均值 = 變異數」是卜瓦松的內建承諾，不是資料的普遍事實。拿到計數資料先比較兩個樣本量：接近則卜瓦松是合理起點；若變異數明顯大於均值（最常見的偏離，源於群聚到達或率的異質性），即**過度離散**，應改用負二項等更有彈性的模型。

練習 4.8 詳解　(a) 「$X>k$」等價於「前 k 次全部失敗」，由獨立性相乘得 $(1-p)^k$。（級數驗證：$\sum_{j=k+1}^{\infty}(1-p)^{j-1}p=p(1-p)^k\cdot\frac{1}{p}=(1-p)^k$，幾何級數求和。）
(b) 事件 $\{X>m+n\}$ 包含於 $\{X>m\}$，故

$$\mathbb{P}(X>m+n\mid X>m)=\frac{(1-p)^{m+n}}{(1-p)^m}=(1-p)^n=\mathbb{P}(X>n).\qquad\blacksquare$$

(c) 兩者的尾機率都是時間的指數衰減：幾何是 $(1-p)^k=e^{k\ln(1-p)}$（離散步伐），指數是 $e^{-\lambda t}$（連續時間）——把指數分布只在整數時刻取樣就得到幾何；兩者也分別是各自世界中**唯一**無記憶的分布。賭徒謬誤：由 (b)，「已連輸 10 把」之下接下來的分布與剛坐下時一模一樣——獨立試驗沒有帳本，不會「欠」你一次成功。

練習 4.9 詳解　(a) **卜瓦松** $\mathrm{Poi}(\lambda)$，λ 為每小時平均車流量：連續時間的計數、無自然上限。若尖峰離峰混用同一個 λ、或車輛成隊通過（不獨立），會失準（過度離散）。

(b) **二項** $\mathrm{Bin}(100,p)$，p 為單一顧客的下單率：固定 $n=100$ 的獨立是 / 否試驗。若 p 很小可用 $\mathrm{Poi}(100p)$ 近似。
(c) **指數** $\mathrm{Exp}(\lambda)$，$1/\lambda$ 為平均壽命：「不會老化」正是無記憶性的白話文。會磨損的零件應改用 Weibull 或 Gamma。
(d) **幾何** $\mathrm{Geom}(1/6)$：獨立重複試驗等第一次成功，期望擲 6 次。
(e) **常態** $N(\mu,\sigma^2)$：身高是大量基因與環境小效應之和，中央極限定理使其趨向高斯。
(f) **Beta** $\mathrm{Beta}(\alpha,\beta)$：點擊率是 $[0,1]$ 上的未知機率，Beta 是「機率的分布」、白努利資料的共軛先驗。

練習 4.10 詳解　記 $p=\lambda/n$，把 PMF 整理成四個因子：

$$\mathbb{P}(X_n=k)=\binom{n}{k}\Big(\frac{\lambda}{n}\Big)^k\Big(1-\frac{\lambda}{n}\Big)^{n-k}=\underbrace{\frac{\lambda^k}{k!}}_{\text{(i)}}\cdot\underbrace{\frac{n(n-1)\cdots(n-k+1)}{n^k}}_{\text{(ii)}}\cdot\underbrace{\Big(1-\frac{\lambda}{n}\Big)^{n}}_{\text{(iii)}}\cdot\underbrace{\Big(1-\frac{\lambda}{n}\Big)^{-k}}_{\text{(iv)}}.$$

逐一取 $n\to\infty$ 的極限（k 固定）：(i) 與 n 無關；(ii) $=\prod_{j=0}^{k-1}(1-j/n)\to 1$；(iii) $\to e^{-\lambda}$（取對數：$n\ln(1-\lambda/n)=n\big(-\lambda/n+O(n^{-2})\big)\to-\lambda$）；(iv) $\to 1$（底趨近 1、次方固定）。四者相乘即得

$$\lim_{n\to\infty}\mathbb{P}(X_n=k)=\frac{\lambda^k e^{-\lambda}}{k!}.\qquad\blacksquare$$

「均值 = 變異數」的由來：X_n 的均值恆為 λ，變異數 $\lambda(1-\lambda/n)\to\lambda$——極限把因子 $(1-p)$ 送到 1，變異數在極限中追平均值。

練習 4.11 詳解　參考程式與輸出：

```python
import numpy as np
from scipy import stats
rng = np.random.default_rng(0)
for k in range(7):                    # (a) 二項 vs 卜瓦松對照
    print(k, round(stats.binom.pmf(k, 100, 0.03), 4),
          round(stats.poisson.pmf(k, 3), 4))
gaps = rng.exponential(scale=1/3, size=(200_000, 40))  # (b) Exp 間隔
counts = (np.cumsum(gaps, axis=1) <= 1.0).sum(axis=1)  # 單位時間到達數
print(counts.mean(), counts.var())    # 皆約 3：均值 = 變異數
print((counts >= 2).mean())           # 約 0.801（理論 0.8009）
for k in [1, 2, 3]:                   # (c) 68-95-99.7
    print(k, round(stats.norm.cdf(k) - stats.norm.cdf(-k), 4))
```

說明：(a) 兩欄與例 4.3 一致，逐點差距不超過 0.0035。(b) 模擬得到均值與變異數皆 ≈ 3.00、$\mathbb{P}(N\ge 2)\approx 0.801$，與例 4.2 的 $1-4e^{-3}\approx 0.8009$ 吻合。(c) 輸出 0.6827、0.9545、0.9973。

下一章：多個隨機變數——聯合分布、共變異數與相關。