機率與統計：從入門到精通

第 9 章　貝氏推論

Bayesian Inference

難度：進階

到目前為止，我們的推論都站在頻率學派這一邊：參數是固定未知的常數，隨機的是資料，估計量的好壞用「重複抽樣下的長期表現」評判。貝氏推論把重點整個翻轉過來：讓參數本身成為**隨機變數**，用機率分布編碼我們對它的信念，再用貝氏定理讓資料更新這些信念。推論的產出不再是單一個點，而是關於未知之物的**完整分布**——點估計、區間、預測全都從中讀出，還能誠實說出「我有多不確定」。這正是現代、注重安全的 AI 所需要的不確定性語言。

學習目標

讀完本章後，你將能夠：

- 寫出貝氏配方「後驗 $\propto$ 概似 $\times$ 先驗」，說出**先驗分布**、**概似**、**證據**與**後驗分布**各自扮演的角色。
- 使用**共軛先驗**（Beta–白努利、常態–常態）以純代數完成後驗更新，並以**偽計數**與**精度**解讀更新公式。
- 從後驗提煉點估計：計算**後驗均值**與**最大後驗估計**（MAP），並解釋「MAP = MLE + 先驗懲罰」如何生出 L2 與 L1 正則化。
- 建構並正確解讀**可信區間**，計算**後驗預測分布**，分辨可信區間與信賴區間陳述的差異。
- 對照頻率學派與貝氏學派各自回答的問題，說明現代機器學習如何務實地兩者並用。
- 用 Python / SciPy 實作 Beta–白努利更新、序貫更新、可信區間與後驗預測。

9.1 貝氏配方：用資料更新信念

第 7 章的最大概似估計（maximum likelihood estimation, MLE）問的是：「哪一個參數值讓觀察到的資料最不令人意外？」它交出一個點 $\hat{\theta}_{\text{MLE}}$；第 8 章再用抽樣分布替這個點掛上誤差條。整套流程的世界觀是頻率學派的：θ 是固定未知的常數，隨機的是資料。

貝氏推論把視角翻轉過來。既然我們不知道 θ，那就誠實地用機率分布描述這份無知：把 θ 當成**隨機變數**。這不是主張參數在物理上會變動，而是把機率當作**信念的語言**（第 1 章的貝氏詮釋）：看資料之前的信念，寫成**先驗分布**（prior distribution）$p(\theta)$；資料的證詞，由**概似**（likelihood）$p(\mathcal{D} \mid \theta)$ 帶進來；第 2 章的貝氏定理則負責把兩者融合成看過資料之後的信念。

定義 9.1　貝氏推論與後驗分布

設參數 θ 有先驗分布 $p(\theta)$，資料 $\mathcal{D}$ 在給定 θ 下依概似 $p(\mathcal{D}\mid\theta)$ 生成。**貝氏推論**（Bayesian inference）計算給定資料下參數的**後驗分布**（posterior distribution）：

$$p(\theta\mid\mathcal{D})=\frac{p(\mathcal{D}\mid\theta)\,p(\theta)}{p(\mathcal{D})},\qquad p(\mathcal{D})=\int p(\mathcal{D}\mid\theta)\,p(\theta)\,\mathrm{d}\theta,$$

分母 $p(\mathcal{D})$ 稱為**證據**（evidence）或**邊際概似**（marginal likelihood）；θ 離散時積分換成加總。一句話：

$$\text{後驗}\;=\;\frac{\text{概似}\times\text{先驗}}{\text{證據}}.$$

證據來自全機率定律：把「資料在各種 θ 之下發生的機率」用先驗加權平均。關鍵觀察是：$p(\mathcal{D})$ **不含** θ——對「哪個 θ 比較可信」這個問題而言，它只是讓後驗總面積等於 1 的歸一化常數。因此實務上人人都寫

$$p(\theta\mid\mathcal{D})\;\propto\;p(\mathcal{D}\mid\theta)\,p(\theta)\qquad(\text{posterior}\propto\text{likelihood}\times\text{prior}),$$

先算出右邊的「形狀」，最後再歸一化——或者像 9.2 節那樣，直接認出這個形狀屬於哪個分布家族，連積分都省了。

核心觀念

從 MLE 到貝氏的轉變，是從「單一最佳 θ」到「**所有合理 θ 的完整分布**」。先驗 $p(\theta)$ 陳述看到資料前的信念；概似讓資料發聲；後驗是刷新後的信念。資料多時概似主導、先驗被沖淡——資料少時先驗讓你保持清醒（並防止過擬合）。你想要的一切——點估計、區間、預測——都從後驗讀出。

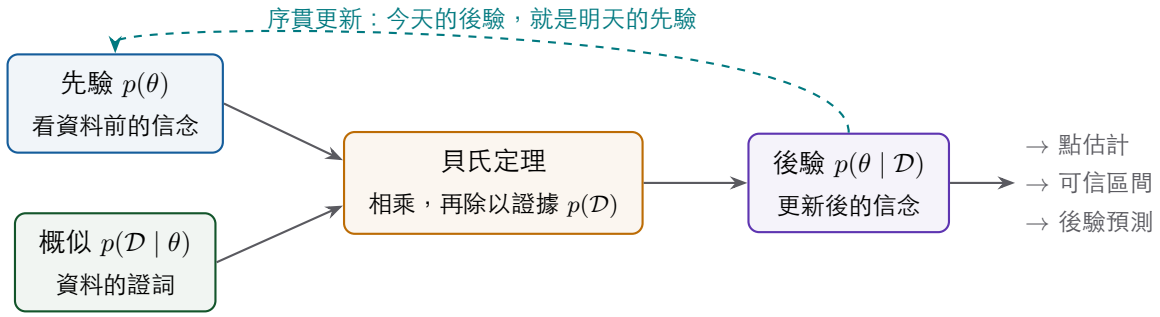


圖 9.1: 貝氏推論的更新迴圈：先驗與概似相乘、歸一化成後驗；一切推論產出都從後驗讀出，而後驗又能充當下一批資料的先驗。

例 9.1　兩個假設的暖身操

一枚來路不明的硬幣只有兩種可能：公正（$p=0.5$）或灌鉛（$p=0.8$），先驗信念各半。擲 10 次，得到 7 正 3 反（順序已記錄）。求兩個假設的後驗機率。

解　對一條特定的正反面序列，概似為 $p^7(1-p)^3$：

$$p(\mathcal{D}\mid 0.5)=0.5^{10}\approx 0.000977,\qquad p(\mathcal{D}\mid 0.8)=0.8^7\times 0.2^3\approx 0.001678.$$

證據是兩者以先驗加權的平均：

$$p(\mathcal{D})=0.5\times 0.000977+0.5\times 0.001678\approx 0.001327.$$

套用貝氏配方：

$$\mathbb{P}(0.5\mid\mathcal{D})=\frac{0.5\times 0.000977}{0.001327}\approx 0.368,\qquad \mathbb{P}(0.8\mid\mathcal{D})=\frac{0.5\times 0.001678}{0.001327}\approx 0.632.$$

資料溫和地偏袒灌鉛假設（概似比約 1.72，尚稱不上鐵證），而整場計算只有「先驗 × 概似，再歸一化」兩個動作。若想對 p 的**所有**可能值保留信念——而非只押兩個候選——就得把先驗升級成 $[0,1]$ 上的連續分布：這正是下一節的主角。

常見陷阱

把概似當成後驗，是本章最要命的方向錯誤。$p(\mathcal{D}\mid\theta)$ 回答「若 θ 為真，看到這筆資料的機率有多大」；$p(\theta\mid\mathcal{D})$ 回答「看過資料後，θ 有多可信」。兩者相差一個先驗與歸一化：概似作為 θ 的函數，通常**連機率分布都不是**（對 θ 積分未必等於 1）。直接把概似讀成信念，等於偷偷假設了均勻先驗；當基本率重要時（罕見疾病篩檢、罕見故障），這就是第 2 章的基本率謬誤穿上參數的外衣重演；法庭版本叫檢察官謬誤——把「若無辜，出現此證據的機率很小」偷換成「出現此證據，無辜的機率很小」。

直觀理解

把貝氏推論想成一台**信念更新機**：先驗是起點，每一筆資料都輕推一次信念，分母只負責把總面積校回 1。推力的大小取決於概似的**對比度**：資料在各種 θ 下都差不多可能，後驗就幾乎不動；資料強烈偏袒某些 θ，後驗便大幅集中。樣本數是槓桿——n 筆獨立資料的概似是 n 個因子連乘，對數尺度上逐筆疊加，而先驗只有固定一票，遲早被壓過。「資料多時概似主導、資料少時先驗坐鎮」的機制就在這裡（9.2 節會算給你看）。

9.2　共軛先驗：讓更新變成代數

貝氏配方概念上只有一行，實作上有兩道關卡：分母積分 $p(\mathcal{D})$ 在高維往往算不動；就算算得動，後驗也未必是叫得出名字的分布。**共軛先驗**讓兩道關卡同時消失：挑一個與概似「合拍」的先驗家族，讓後驗自動留在同一家族——更新化為純代數，舊參數進、新參數出。

定義 9.2　共軛先驗

給定概似 $p(\mathcal{D}\mid\theta)$。若先驗取自分布族 $\mathcal{F}$ 時，後驗必仍屬於 $\mathcal{F}$，則稱 $\mathcal{F}$ 為該概似的**共軛先驗**（conjugate prior）家族。此時貝氏更新化約為家族參數的代數運算，證據積分無需真的動手算。

9.2.1 Beta–白努利共軛：擲硬幣的完整流程

估計成功機率 p（硬幣偏差、點擊率、藥物有效率）時，先驗要住在 $[0,1]$ 上。天選之材是 **Beta 分布**（第 3 章）：$\mathrm{Beta}(\alpha,\beta)$ 的密度為

$$f(p)=\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)},\qquad B(\alpha,\beta)=\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t=\frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)},$$

均值 $\alpha/(\alpha+\beta)$，眾數 $(\alpha-1)/(\alpha+\beta-2)$（$\alpha,\beta>1$ 時），變異數 $\alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$（前兩者的推導見練習 9.5）。$\alpha=\beta=1$ 是均勻分布；調整 α,β 可得單峰、偏斜甚至 U 形——當「機率的機率分布」再合適不過。

定理 9.3　Beta–白努利共軛

設先驗 $p\sim\mathrm{Beta}(\alpha,\beta)$，觀察 n 次獨立白努利試驗（Bernoulli trial）中有 k 次成功。則後驗為

$$p\mid\mathcal{D}\sim\mathrm{Beta}(\alpha+k,\ \beta+n-k).$$

證明　對任一特定序列，概似為 $p^k(1-p)^{n-k}$，故

$$p(p\mid\mathcal{D})\ \propto\ \underbrace{p^{k}(1-p)^{n-k}}_{\text{概似}}\cdot\underbrace{p^{\alpha-1}(1-p)^{\beta-1}}_{\text{先驗的核}}=p^{(\alpha+k)-1}(1-p)^{(\beta+n-k)-1}.$$

這正是 $\mathrm{Beta}(\alpha+k,\beta+n-k)$ 的核；後驗必須積分為 1，所以歸一化常數只能是 $1/B(\alpha+k,\beta+n-k)$，證據積分連算都不用算。 ■

更新規則好記到不可思議：**正面數加進 α、反面數加進 β**。概似的指數直接疊上先驗的指數，像在記帳。

例 9.2　Beta–白努利完整更新

從均勻先驗 $\mathrm{Beta}(1,1)$ 出發，擲 10 次得 7 正 3 反。求後驗及其均值、眾數、變異數，並與 MLE 比較。

解　由定理 9.3，後驗為 $\mathrm{Beta}(1+7,\ 1+3)=\mathrm{Beta}(8,4)$。套公式：

$$\mathbb{E}[p\mid\mathcal{D}]=\frac{8}{12}=\frac{2}{3}\approx0.667,\qquad \text{眾數}=\frac{8-1}{8+4-2}=\frac{7}{10}=0.7,$$

$$\mathrm{Var}(p\mid\mathcal{D})=\frac{8\times4}{12^2\times13}=\frac{32}{1872}\approx0.0171,\qquad \text{標準差}\approx0.131.$$

MLE 為 $\hat p=7/10=0.7$，恰等於後驗眾數——均勻先驗下 MAP 與 MLE 重合（9.3 節）。後驗均值 0.667 則被先驗往 0.5 輕拉了一把：它等於 $(7+1)/(10+2)$，正是 Laplace 平滑

（見下方「直觀理解」）。而完整的答案是整條後驗曲線（圖 9.2 的藍線）：p 在 0.5 附近仍有可觀的密度，證據還不足以蓋棺論定。

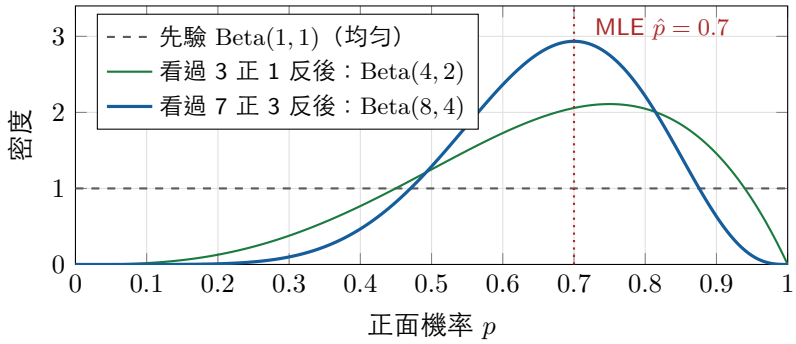


圖 9.2: Beta–白努利的貝氏更新。從均勻先驗出發，看過 3 正 1 反後信念成為 Beta$(4,2)$，累計 7 正 3 反後成為 Beta$(8,4)$：峰越來越高、越來越窄——證據越多，信念越集中。

直觀理解

Beta 先驗的參數可以讀成**偽計數**（pseudo-counts）：Beta(α,β) 就像在看資料之前，已經「想像自己看過」α 次正面與 β 次反面。更新規則「$\alpha+k$、$\beta+(n-k)$」不過是把想像計數與真實計數相加；後驗均值 $(\alpha+k)/(\alpha+\beta+n)$ 是「摻了偽計數的成功比例」。從均勻先驗 Beta$(1,1)$ 出發時，它成為

$$\mathbb{E}[p \mid \mathcal{D}] = \frac{k+1}{n+2}$$

——**Laplace 平滑**（Laplace smoothing）。這正是 naive Bayes 分類器與 n-gram 語言模型裡「每個計數都加 1」的那個 +1：訓練語料裡沒出現過的詞，不會被判零機率，因為均勻先驗早已替每種可能墊了一次偽觀察。「計數為 0」和「不可能」是兩回事——偽計數把這句常識寫進了公式。

例 9.3　序貫更新：後驗當新先驗

資料一筆一筆到來：正、正、反、正。從 Beta$(1,1)$ 出發，每看一筆就把當下的後驗當成下一筆的先驗。驗證最終結果與「一次看完全部資料」相同。

解　逐筆更新（正面把 α 加 1，反面把 β 加 1）：

$$\text{Beta}(1,1) \xrightarrow{\text{正}} \text{Beta}(2,1) \xrightarrow{\text{正}} \text{Beta}(3,1) \xrightarrow{\text{反}} \text{Beta}(3,2) \xrightarrow{\text{正}} \text{Beta}(4,2).$$

批次更新：3 正 1 反一次進帳，$\text{Beta}(1+3,\ 1+1) = \text{Beta}(4,2)$。完全一致。一般性的證明見練習 9.6：概似只透過累計的（正面數、反面數）進入公式，所以資料的**順序**與**分批方式**都無關緊要。

工程意涵：**線上學習**（online learning）免費入手——不存原始資料，只帶著 (α,β) 兩個數字前進，串流資料隨到隨學。圖 9.2 的三條曲線正是這個過程的快照。

命題 9.4　後驗均值的收縮形式

設先驗 $\mathrm{Beta}(\alpha,\beta)$、資料為 n 次中 k 次成功。記先驗均值 $m_0=\alpha/(\alpha+\beta)$、MLE $\hat p=k/n$，則後驗均值是兩者的凸組合：

$$\mathbb{E}[p\mid\mathcal{D}]=\frac{\alpha+k}{\alpha+\beta+n}=w\,m_0+(1-w)\,\hat p,\qquad w=\frac{\alpha+\beta}{\alpha+\beta+n}.$$

證明　分子改寫為 $\alpha+k=(\alpha+\beta)\,m_0+n\,\hat p$（因 $(\alpha+\beta)m_0=\alpha$、$n\hat p=k$），除以 $\alpha+\beta+n$ 即得。 ■

後驗均值把 MLE 往先驗**收縮**（shrinkage）：先驗的發言權 w 由偽觀察數與樣本數的比例決定——$n\to\infty$ 時 $w\to 0$，資料終將勝出；n 小時先驗撐住估計，不讓三五筆資料亂帶風向。這是「先驗防止過擬合」的定量版本。

例 9.4　不同的先驗，同一筆資料

三位分析師拿到同一筆資料（10 擲 7 正），但先驗不同：甲用均勻 $\mathrm{Beta}(1,1)$；乙用強而中立的 $\mathrm{Beta}(10,10)$（先驗均值 0.5，攜帶 20 筆偽觀察）；丙相信硬幣偏向反面，用 $\mathrm{Beta}(2,8)$（先驗均值 0.2）。比較三人的後驗；再看資料增為 1000 擲 700 正時的變化。

解　套定理 9.3 逐一更新：

	$n=10,\ k=7$		$n=1000,\ k=700$	
先驗	後驗	後驗均值	後驗	後驗均值
甲：$\mathrm{Beta}(1,1)$	$\mathrm{Beta}(8,4)$	0.667	$\mathrm{Beta}(701,301)$	0.700
乙：$\mathrm{Beta}(10,10)$	$\mathrm{Beta}(17,13)$	0.567	$\mathrm{Beta}(710,310)$	0.696
丙：$\mathrm{Beta}(2,8)$	$\mathrm{Beta}(9,11)$	0.450	$\mathrm{Beta}(702,308)$	0.695

$n=10$ 時三個後驗均值分歧明顯——先驗還握有話語權：乙的 20 筆偽觀察佔權重 $w=2/3$，丙甚至把估計壓到 0.5 以下。$n=1000$ 時三者幾乎重合——資料淹沒先驗（一般極限見練習 9.8）：命題 9.4 中 $w=(\alpha+\beta)/(\alpha+\beta+n)$ 隨 n 增大而歸零。

常見陷阱

先驗選擇不當有兩種災難。**其一：判零機率等於判死刑。**先驗在某區域給機率 0，任何資料都救不回來——0 乘任何概似還是 0。這是 Cromwell 法則：除非邏輯上不可能，先驗別設 0；其計數版本正是 naive Bayes 的零頻率問題——某詞沒在訓練集出現，整條連乘積被打成 0——Laplace 平滑是解藥。**其二：過強的先驗會輾壓資料。**$\mathrm{Beta}(1000,1000)$ 等於宣稱看過兩千筆偽觀察，真資料要累積數千筆才撼得動；若那份自信沒有根據，推論就被成見綁架。原則：先驗誠實反映知識強度——所知有限就用弱先驗（偽計數小），把麥克風留給資料。

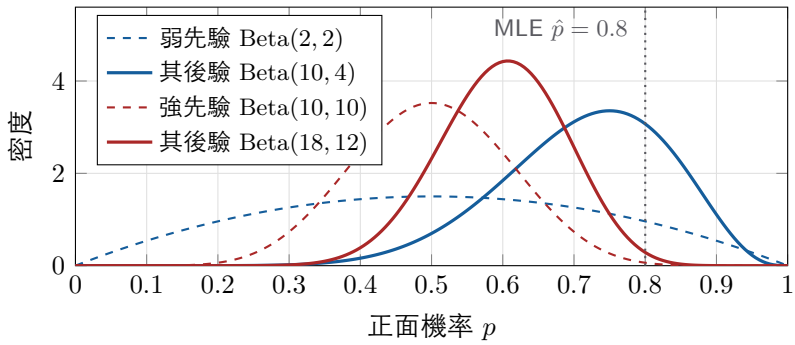


圖 9.3: 先驗強度的影響：同一筆資料（10 擲 8 正，MLE $= 0.8$，點虛線），弱先驗 Beta$(2,2)$（藍虛線）幾乎讓資料說了算，後驗 Beta$(10,4)$ 峰落在 0.75 附近；強先驗 Beta$(10,10)$（紅虛線）握著 20 筆偽觀察，後驗 Beta$(18,12)$ 只肯移到 0.6 附近。先驗越強，資料要花越大力氣才搬得動它。

9.2.2 常態–常態共軛與卡爾曼濾波器

第二個經典共軛配對是：**高斯對自身共軛**——高斯先驗配高斯概似，後驗仍是高斯。敘述之前，先引入一個讓公式變漂亮的詞：**精度**（precision）$=$ 變異數的倒數 $1/\sigma^2$，度量「一份資訊有多值得信賴」——變異數越小、精度越高。

定理 9.5　常態–常態共軛（σ^2 已知）

設先驗 $\theta \sim \mathcal{N}(\mu_0, \tau_0^2)$，資料 $x_1, \dots, x_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$，其中 σ^2 已知。則

$$\theta \mid \mathcal{D} \sim \mathcal{N}(\mu_n, \tau_n^2), \qquad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}, \qquad \mu_n = \tau_n^2 \left(\frac{\mu_0}{\tau_0^2} + \frac{n\bar{x}}{\sigma^2} \right).$$

一句話：**後驗精度 $=$ 先驗精度 $+$ 資料精度；後驗均值 $=$ 以精度為權重的加權平均。**

證明　只追蹤指數中含 θ 的部分：

$$p(\theta \mid \mathcal{D}) \propto \exp\!\Big(-\tfrac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2 \Big) \exp\!\Big(-\tfrac{1}{2\tau_0^2} (\theta - \mu_0)^2 \Big).$$

展開 $\sum (x_i - \theta)^2 = n\theta^2 - 2\theta n\bar{x} +$（與 θ 無關）與 $(\theta - \mu_0)^2 = \theta^2 - 2\theta\mu_0 +$（與 θ 無關），合併同類項：

$$p(\theta \mid \mathcal{D}) \propto \exp\left[-\frac{1}{2} \Big(A\theta^2 - 2B\theta \Big) \right], \qquad A = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2}, \quad B = \frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\tau_0^2}.$$

配方（completing the square）給 $A\theta^2 - 2B\theta = A(\theta - B/A)^2 - B^2/A$，故 $p(\theta \mid \mathcal{D}) \propto \exp\!\big[-(\theta - B/A)^2 / (2/A) \big]$——正是 $\mathcal{N}(B/A,\ 1/A)$ 的核：$\tau_n^2 = 1/A$、$\mu_n = B/A = \tau_n^2 B$，與敘述一致。 ■

例 9.5　精度加權平均：估計真實能力

想估計某位學生的真實能力 θ（以測驗分數為尺度）。母體常模提供先驗 $\theta \sim \mathcal{N}(100, 15^2)$，即 $\mu_0 = 100$、$\tau_0^2 = 225$。單次測驗有雜訊：$x_i \sim \mathcal{N}(\theta, 10^2)$，即 $\sigma^2 = 100$。該生考了 $n = 4$ 次，平均 $\bar{x} = 115$。求後驗。

解　精度相加：

$$\frac{1}{\tau_n^2} = \frac{1}{225} + \frac{4}{100} = \frac{2}{45} \approx 0.0444 \quad\Longrightarrow\quad \tau_n^2 = 22.5,\ \ \tau_n \approx 4.74.$$

精度加權平均：

$$\mu_n = 22.5 \times \left(\frac{100}{225} + \frac{4 \times 115}{100} \right) = 22.5 \times (0.444 + 4.6) = 113.5.$$

換個角度看權重：先驗精度佔 $\frac{1/225}{2/45} = 0.1$、資料佔 0.9，故 $\mu_n = 0.1 \times 100 + 0.9 \times 115 = 113.5$。後驗為 $\mathcal{N}(113.5,\ 22.5)$：估計從樣本平均 115 往母體平均拉回了 1.5 分。直觀：測驗有雜訊，四次平均考出 115 的人，其真實能力多半沒有分數看起來那麼高——這是「回歸均值」現象的貝氏解讀。95% 可信區間見例 9.8。

「高斯進、高斯出」的封閉性可以**反覆串接**：今天的後驗就是明天的先驗，永遠不會跑出家族。這正是線性高斯模型與**卡爾曼濾波器**（Kalman filter）能閉式更新的原因——每個時間步做一次常態–常態更新，「預測」步讓不確定性膨脹、「更新」步用新觀測收窄。自駕車追蹤前車、火箭導航、金融時間序列濾波，底層都是這套代數。

9.3　從後驗提煉點估計：後驗均值與 MAP

後驗是完整分布，但世界常逼你交出一個數字：論文表格要填一格、控制器要一個設定值、神經網路要一組權重。從分布壓縮成數字的方式不只一種——各自在不同的損失下「最優」。

定義 9.6　後驗均值與最大後驗估計

後驗均值（posterior mean）為 $\hat{\theta}_{\text{mean}} = \mathbb{E}[\theta \mid \mathcal{D}]$。**最大後驗估計**（maximum a posteriori estimate, MAP）為後驗的峰（眾數）：

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta}\, p(\theta \mid \mathcal{D}) = \arg\max_{\theta}\, \big[\log p(\mathcal{D} \mid \theta) + \log p(\theta)\big],$$

第二個等號成立是因為 $\log p(\mathcal{D})$ 不含 θ，不影響取最大值的位置。

命題 9.7　後驗均值最小化期望平方誤差

對任何猜測值 a，

$$\mathbb{E}\big[(\theta - a)^2 \mid \mathcal{D}\big] = \operatorname{Var}(\theta \mid \mathcal{D}) + \big(\mathbb{E}[\theta \mid \mathcal{D}] - a\big)^2,$$

故 $a = \mathbb{E}[\theta \mid \mathcal{D}]$ 是期望平方誤差的唯一最小化者。

證明　記 $m = \mathbb{E}[\theta \mid \mathcal{D}]$。展開

$$(\theta - a)^2 = \big((\theta - m) + (m - a)\big)^2 = (\theta - m)^2 + 2(m - a)(\theta - m) + (m - a)^2,$$

對後驗取期望：交叉項因 $\mathbb{E}[\theta - m \mid \mathcal{D}] = 0$ 而消失，得 $\mathbb{E}[(\theta - a)^2 \mid \mathcal{D}] = \operatorname{Var}(\theta \mid \mathcal{D}) + (m - a)^2$。

第二項非負且僅在 $a = m$ 時為零。 ■

順帶一提：後驗**中位數**最小化期望絕對誤差，對厚尾後驗更穩健；MAP 對應 0–1 損失的極限。三種摘要在對稱單峰後驗（如常態）重合，在偏斜後驗上分家。

例 9.6　同一個後驗，三個點估計

承例 9.2，後驗為 $\mathrm{Beta}(8, 4)$。求後驗均值、中位數與 MAP，並討論選誰。

解　均值 $8/12 \approx 0.667$；MAP（眾數）$= 7/10 = 0.7$；中位數無閉式，數值計算得約 0.676（9.6 節的程式）。三者略有出入，因為 $\mathrm{Beta}(8, 4)$ 往左拖尾（$\alpha > \beta$ 時眾數 $>$ 均值，見練習 9.5）：均值 $0.667 <$ 中位數 $0.676 <$ 眾數 0.700。選誰？在乎平方誤差就報均值；要穩健就報中位數；只想要後驗的峰、或計算資源有限（只需最佳化、不需積分）就用 MAP。特別地，均勻先驗下 $\log p(\theta)$ 是常數，MAP 化約為 MLE——**MLE 其實是戴著均勻先驗面具的 MAP**。

核心觀念

MAP = MLE + 先驗懲罰，而那個懲罰就是正則化。對照定義 9.6：MAP 的目標函數比 MLE 多出一項 $\log p(\theta)$，用最佳化的語言讀，它就是懲罰項。**高斯先驗** $\theta \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I})$ 貢獻 $-\|\theta\|^2/(2\tau^2)$——正是 **L2 正則化 / 權重衰減**（weight decay），即**脊迴歸**（ridge regression）；**Laplace 先驗** $p(\theta_j) \propto e^{-|\theta_j|/b}$ 貢獻 $-\|\theta\|_1/b$——正是 **L1 正則化**，即**套索迴歸**（lasso），把部分參數推成恰好為零、誘導稀疏（練習 9.11）。深度學習中每一個權重懲罰，暗地裡都是「參數應該小」的先驗信念——這是貝氏統計與實務機器學習之間最重要的一座橋。

例 9.7　MAP 就是脊迴歸：完整推導

線性迴歸模型 $\boldsymbol{y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\varepsilon}$，$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$；對權重放高斯先驗 $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I})$。推導 $\hat{\boldsymbol{w}}_{\mathrm{MAP}}$。

解　寫出對數後驗（丟掉與 $\boldsymbol{w}$ 無關的常數）：

$$\log p(\boldsymbol{w} \mid \mathcal{D}) = -\frac{1}{2\sigma^2}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{w}\|^2 \; - \; \frac{1}{2\tau^2}\|\boldsymbol{w}\|^2 \; + \; \text{const.}$$

最大化等價於（乘以 $2\sigma^2$ 後）最小化

$$\|\boldsymbol{y} - \mathbf{X}\boldsymbol{w}\|^2 + \lambda \|\boldsymbol{w}\|^2, \qquad \lambda = \frac{\sigma^2}{\tau^2},$$

這正是**脊迴歸**的目標函數。令梯度為零：$-2\mathbf{X}^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{w}) + 2\lambda \boldsymbol{w} = \mathbf{0}$，解得閉式解

$$\hat{\boldsymbol{w}}_{\mathrm{MAP}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \boldsymbol{y}.$$

對角線上加的那道 $\lambda \mathbf{I}$「山脊」讓矩陣必定可逆——ridge 之名由此而來。讀 $\lambda = \sigma^2/\tau^2$：雜訊越大（資料不可盡信）或先驗越自信「權重該小」（τ^2 小），懲罰越重、收縮越狠。深度學習的權重衰減每步把權重乘上略小於 1 的因子，正是這個高斯先驗在梯度下降中的身影。

9.4 可信區間與後驗預測

第 8 章的信賴區間常被誤讀成「參數有 95% 機率落在裡面」。在貝氏框架裡，這句話終於可以合法說出口——因為 θ 真的有分布。

定義 9.8　可信區間

給定信念水準 γ（例如 95%），**可信區間**（credible interval）是任何滿足

$$\mathbb{P}(\theta \in C \mid \mathcal{D}) = \gamma$$

的區間 C。最常用的是**等尾**（equal-tailed）區間：取後驗的 2.5% 與 97.5% 分位數為端點。另一種是**最高後驗密度區間**（highest posterior density, HPD）：所有滿足條件的區間中最短的那個；後驗對稱時兩者重合。

例 9.8　計算並對照可信區間

(a) 承例 9.2，後驗 $\mathrm{Beta}(8,4)$，求 95% 等尾可信區間。(b) 承例 9.5，後驗 $\mathcal{N}(113.5, 22.5)$，求 95% 可信區間。(c) 就例 9.5 的資料（$n=4$、$\bar{x}=115$、$\sigma=10$）另求頻率學派的 95% 信賴區間，與 (b) 比較。

解　(a) Beta 分位數無閉式，數值計算（9.6 節的 `ppf`）得 $C = [0.390,\ 0.891]$。常態近似 $0.667 \pm 1.96 \times 0.131 = [0.410,\ 0.923]$ 只能當粗估：$\mathrm{Beta}(8,4)$ 左偏，真正的等尾區間整體往左挪。

(b) 常態後驗的分位數是老朋友：

$$113.5 \pm 1.96 \times \sqrt{22.5} = 113.5 \pm 9.30 = [104.2,\ 122.8],$$

於是可以名正言順地說：「給定資料與先驗，這位學生的真實能力有 95% 的機率落在 104.2 到 122.8 之間。」這句話在頻率框架中是禁句，在貝氏框架中是定義。

(c) 信賴區間為 $\bar{x} \pm 1.96\,\sigma/\sqrt{n} = 115 \pm 9.8 = [105.2,\ 124.8]$。貝氏區間中心往先驗均值收縮（$113.5 < 115$）且更窄（寬 18.6 對 19.6）——先驗貢獻了額外精度。再看先驗趨於無知的極限 $\tau_0^2 \to \infty$：由定理 9.5，$\mu_n \to \bar{x} = 115$、$\tau_n^2 \to \sigma^2/n = 25$，可信區間趨於 115 ± 9.8——與信賴區間**數值重合**，卻仍是兩句意義不同的話：一句關於 θ 的信念，一句關於程序的品質。

核心觀念

可信區間 vs. 信賴區間——差別在「什麼是隨機的」。 **可信區間**（貝氏）：資料固定，θ 有分布——「θ 落在 C 內的後驗機率是 95%」。這正是人們**想要**、也常**誤以為**信賴區間在說的話。**信賴區間**（頻率）：θ 固定，區間隨資料而隨機——「產生區間的**程序**，在大量重複使用下有 95% 的次數蓋住真值」。對算出來的那一個區間，θ 要嘛在、要嘛不在，頻率框架不允許替它掛機率。兩句話的差別不是修辭，而是隨機性放在哪裡的立場之別。可信區間的代價也要說清楚：那個 95% 是**建立在先驗之上**的——先驗換了，區間跟著換。

預測新資料時，貝氏的做法同樣是「別急著挑一個 θ」。

定義 9.9　後驗預測分布

給定 θ 時新觀測 x_{new} 與 $\mathcal{D}$ 條件獨立，則**後驗預測分布**（posterior predictive distribution）為

$$p(x_{\text{new}} \mid \mathcal{D}) = \int p(x_{\text{new}} \mid \theta)\, p(\theta \mid \mathcal{D})\, \mathrm{d}\theta,$$

即「每個 θ 的預測」以後驗為權重的平均。這一步把參數的不確定性**邊際化**（marginalize）進預測，而不是假裝某個點估計就是真理——貝氏預測因此更誠實、往往也更**校準**（calibrated）：說 80% 就大約中 80%。

例 9.9　Laplace 接續律與常態預測

(a) Beta–白努利：後驗 $\mathrm{Beta}(\alpha+k, \beta+n-k)$ 下，下一次試驗成功的機率是多少？(b) 常態–常態：例 9.5 之後，再考一次會考幾分？

解　(a) 對白努利分布，$p(x_{\text{new}}=1 \mid p) = p$，故

$$\mathbb{P}(x_{\text{new}}=1 \mid \mathcal{D}) = \int_0^1 p \cdot p(p \mid \mathcal{D})\, \mathrm{d}p = \mathbb{E}[p \mid \mathcal{D}] = \frac{\alpha+k}{\alpha+\beta+n}.$$

均勻先驗 $(\alpha=\beta=1)$ 時為 $(k+1)/(n+2)$——**Laplace 接續律**（rule of succession），Laplace 當年用它估算「太陽明天照常升起」的機率。例 9.2 的資料（7 正 3 反）給 $8/12 = 2/3$。(b) 寫 $x_{\text{new}} = \theta + \varepsilon$，其中 $\theta \mid \mathcal{D} \sim \mathcal{N}(\mu_n, \tau_n^2)$ 與 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ 獨立。獨立常態相加仍是常態（第 5 章），變異數相加：

$$x_{\text{new}} \mid \mathcal{D} \sim \mathcal{N}\big(\mu_n,\ \tau_n^2 + \sigma^2\big) = \mathcal{N}(113.5,\ 22.5 + 100) = \mathcal{N}(113.5,\ 122.5),$$

標準差約 11.1。若把 μ_n 當真值插回去（plug-in），只得 $\mathcal{N}(113.5, 100)$、標準差 10——少算的 22.5 正是「我們其實還不確定 θ」的那一份，忽略它，預測就會過度自信。

9.5　貝氏與頻率：兩種提問方式

學到這裡，值得停下來把兩個學派並排看清楚。爭論「誰才對」往往搞錯重點——它們回答的根本是不同的問題。

直觀理解

頻率學派問：「什麼樣的**程序**，在大量重複使用下有良好的長期錯誤率？」它交付品質保證書：型一錯誤率 5%、區間覆蓋率 95%——對「程序」負責。**貝氏學派問**：「**給定這一筆資料**與我的先驗，我該相信什麼？」它交付一張信念地圖：後驗——對「這一次」負責。機器學習日常兩種都有：評估流水線（「這套 A/B 測試流程長期可靠嗎？」）是頻率式的，線上決策（「依目前證據，這個版本較好的機率多大？」）是貝氏式的。熟練的人不站隊，只挑對工具。

表 9.1: 頻率學派與貝氏學派的對照：不是誰對誰錯，而是回答不同的問題。

	頻率學派	貝氏學派
參數 θ	固定未知的常數	隨機變數，帶有分布
資料 $\mathcal{D}$	隨機（想像重複抽樣）	固定（就是眼前這一筆）
機率的意義	長期頻率	信念的程度
需要先驗嗎	不需要	需要（也因此能吸收領域知識）
區間陳述	信賴區間：程序覆蓋率 95%	可信區間：θ 落入的機率 95%
代表方法	MLE、假設檢定、p 值	後驗、MAP、貝氏因子
計算負擔	通常較輕	證據積分常需近似（MCMC、變分推論）
擅長場景	大樣本、標準化檢定流程	小樣本、先驗知識、不確定性傳播

頻率方法不需要先驗，靠程序保證主導了科學研究的假設檢定與臨床試驗規範，大樣本、標準流程下高效可靠。貝氏方法優雅處理小樣本與先驗知識，讓不確定性一路傳進下游（後驗 → 後驗預測 → 決策）；代價是必須選先驗，且證據積分在高維往往算不動——第 12 章的**變分推論**（variational inference）正是為了讓它變可行而生。現代機器學習務實地兩者兼用：**訓練**用帶懲罰的 MLE（9.3 節：偽裝的 MAP），**思想**上讓貝氏供應正則化、不確定性估計與生成模型的骨架。

這如何驅動機器學習／ AI

貝氏思維貫穿機器學習，即使沒有人開口說「貝氏」：

- **每個正則項都是先驗**：權重衰減 = 高斯先驗（L2、脊迴歸）；L1 稀疏 = Laplace 先驗（lasso）；label smoothing = 在標籤分布上放先驗——不准模型把訓練標籤當成 100% 確定。
- **dropout ≈ 貝氏模型平均**（Bayesian model averaging）：每一步都在訓練一個隨機子網路，整體近似「對許多模型的預測取後驗加權平均」。
- **校準的不確定性**：貝氏神經網路（Bayesian neural network）給權重掛上後驗；MC dropout 與深度集成（deep ensembles）用取樣近似後驗預測分布——在醫療診斷、自駕與主動學習（active learning）裡，「知道自己不知道」是安全關鍵。
- **變分推論**（第 12 章）把算不動的證據積分變成最佳化問題，也是變分自編碼器（VAE）的訓練原理；**高斯過程**（Gaussian process）則是完全貝氏的非參數模型——後驗均值給預測、後驗變異數給誤差條。

當一個系統說「我 80% 確定」而且是認真的，它的底下就有一個後驗。

9.6 Python 實作

理論之後，動手驗證。以下程式完整走一遍本章流程：Beta–白努利共軛更新、可信區間、序貫更新、後驗預測，以及常態–常態的精度加權平均。

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

# --- 1. Beta-白努利共軛更新：先驗 Beta(1,1)，資料 7 正 3 反 ---
alpha0, beta0 = 1, 1                          # 均勻先驗（偽計數各 1）
k, n = 7, 10
alpha1, beta1 = alpha0 + k, beta0 + n - k     # 後驗 Beta(8, 4)：純代數
post = stats.beta(alpha1, beta1)
print(post.mean())                # 0.6667 = (7+1)/(10+2)：Laplace 平滑
print((alpha1 - 1) / (alpha1 + beta1 - 2))  # 眾數 0.7 = MLE = MAP
print(post.std())                 # 0.1307：後驗自帶不確定性刻度

# --- 2. 95% 等尾可信區間：直接讀後驗分位數 ---
print(post.ppf([0.025, 0.975]))   # [0.390, 0.891]

# --- 3. 序貫更新 = 批次更新：今天的後驗是明天的先驗 ---
a, b = 1, 1
for x in [1, 1, 0, 1, 1, 1, 0, 1, 1, 0]:    # 同樣 7 正 3 反，逐筆餵入
    a, b = a + x, b + (1 - x)
print((a, b))                     # (8, 4)：與一次全部更新完全相同

# --- 4. 後驗預測：下一擲為正面的機率（Laplace 接續律）---
print(alpha1 / (alpha1 + beta1))  # 8/12 = 0.6667

# --- 5. 常態-常態共軛：精度相加、精度加權平均 ---
mu0, tau0_sq = 100.0, 225.0       # 先驗 N(100, 15^2)
sigma_sq, m, xbar = 100.0, 4, 115.0         # 4 筆觀測，平均 115
prec = 1 / tau0_sq + m / sigma_sq           # 後驗精度 = 先驗 + 資料
tau_sq = 1 / prec
mu_post = tau_sq * (mu0 / tau0_sq + m * xbar / sigma_sq)
print(mu_post, tau_sq)            # 113.5, 22.5
# MAP 備忘：權重先驗 w ~ N(0, tau^2 I) 時，最大化後驗等價於
# 最小化「平方誤差 + (sigma^2/tau^2)||w||^2」——正是脊迴歸（ridge）。
```

依序印出 0.6667（後驗均值 = Laplace 平滑）、0.7（眾數 = MAP = MLE）、0.1307（後驗標準差）、$[0.390, 0.891]$（可信區間）、$(8, 4)$（序貫 = 批次）、0.6667（後驗預測）、113.5 與 22.5（常態後驗）——每個數字都對應本章的一個公式：共軛先驗讓「推論」跟「查表」一樣便宜。

本章總結

- **貝氏配方**：後驗 $\propto$ 概似 $\times$ 先驗；證據 $p(\mathcal{D})$ 是歸一化常數。推論的產出從「單一最佳 θ」升級為「所有合理 θ 的完整分布」。

- **共軛先驗**讓更新變成代數：Beta–白努利把正反面數加進 (α, β)——先驗參數是**偽計數**，均勻先驗的後驗均值 $(k+1)/(n+2)$ 就是 **Laplace 平滑**；常態–常態則是**精度相加、精度加權平均**，串接起來就是卡爾曼濾波器。
- 後驗均值是先驗均值與 MLE 的凸組合：n 小時先驗坐鎮（正則化、防過擬合），$n \to \infty$ 時資料淹沒先驗、後驗集中到 MLE。
- **點估計**：後驗均值最小化平方誤差、中位數最小化絕對誤差、MAP 是後驗的峰。**MAP = MLE + 先驗懲罰**：高斯先驗 → L2 / 權重衰減 / 脊迴歸；Laplace 先驗 → L1 / lasso 稀疏——**正則化就是先驗**。
- **可信區間**的意義（「θ 有 95% 機率在裡面」）正是信賴區間常被誤解成的那個；差別在於把隨機性放在參數還是程序上。
- **後驗預測分布**把參數不確定性邊際化進預測，比 plug-in 誠實、更校準——低估不確定性是「把點估計當真理」的通病。
- 兩學派回答不同的問題；現代 ML 兩者並用：訓練是帶懲罰的 MLE（MAP），不確定性與生成建模交給貝氏思想（dropout、深度集成、貝氏神經網路、變分推論、高斯過程）。

9.7 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 9.1 (基礎)　從先驗 $p \sim \text{Beta}(2,2)$ 出發，擲硬幣 10 次得 6 正 4 反。(a) 寫出後驗分布。(b) 計算後驗均值、眾數與標準差。(c) 與 MLE $\hat{p} = 0.6$ 比較，用「收縮」解釋差異的方向。

練習 9.2 (基礎)　一枚硬幣只有兩種可能：公正（$p = 0.5$，先驗機率 0.75）或灌鉛（$p = 0.9$，先驗機率 0.25）。連擲 3 次都是正面。(a) 求兩個假設的後驗機率。(b) 求「下一擲仍為正面」的後驗預測機率。

練習 9.3 (基礎)　新藥在 10 位受試者中對 9 位有效。採均勻先驗 $\text{Beta}(1,1)$。(a) 寫出有效率 p 的後驗。(b) 用 Laplace 接續律求「下一位受試者有效」的機率。(c) 若 10 位全部有效，MLE 會給出什麼預測？Laplace 接續律呢？哪個更合理，為什麼？

練習 9.4 (基礎)　你想用 Beta 先驗編碼這樣的信念：「成功率大約 0.3，把握程度相當於看過 20 次試驗」。(a) 給出適當的 $\text{Beta}(\alpha, \beta)$。(b) 隨後觀察 50 次中 20 次成功，寫出後驗並計算後驗均值。(c) 用命題 9.4 的凸組合公式驗證 (b) 的結果。

練習 9.5 (進階)　設 $p \sim \mathrm{Beta}(\alpha,\beta)$。(a) 由積分推導 $\mathbb{E}[p]=\alpha/(\alpha+\beta)$（提示：$\Gamma(z+1)=z\Gamma(z)$）。(b) 對 $\alpha,\beta>1$，推導眾數為 $(\alpha-1)/(\alpha+\beta-2)$。(c) 證明 $\alpha>\beta$ 時眾數大於均值。

練習 9.6 (進階)　資料分兩批到來：第一批 n_1 次中 k_1 次成功，第二批 n_2 次中 k_2 次成功。從先驗 $\mathrm{Beta}(\alpha,\beta)$ 出發，證明「先用第一批更新、把所得後驗當先驗、再用第二批更新」與「兩批合併一次更新」給出完全相同的後驗，並說明這對線上學習系統的意義。

練習 9.7 (進階)　感測器量測某管線的真實溫度 θ。先驗 $\theta\sim\mathcal{N}(20,4)$；儀器單次讀值 $x_i\sim\mathcal{N}(\theta,1)$。今取得 $n=5$ 筆讀值，平均 $\bar{x}=22$。(a) 求後驗分布。(b) 求 95% 等尾可信區間。(c) 求頻率學派的 95% 信賴區間，比較兩區間的中心與寬度並解釋差異。

練習 9.8 (進階)　「先驗會不會被資料淹沒？」(a) 把 $\mathrm{Beta}(\alpha,\beta)$ 先驗下的後驗均值寫成先驗均值 m_0 與 MLE $\hat{p}=k/n$ 的凸組合。(b) 固定 $\hat{p}$、令 $n\to\infty$，證明後驗均值收斂到 $\hat{p}$。(c) 證明後驗變異數滿足 $\mathrm{Var}(p\mid\mathcal{D})\le\frac{1}{4(\alpha+\beta+n+1)}$，並解讀其極限行為。

練習 9.9 (進階)　判斷下列敘述正確或錯誤，並說明理由。(a)「由這筆資料算出 95% 信賴區間 $[0.42,0.61]$，所以 θ 有 95% 的機率落在 $[0.42,0.61]$ 內。」(b)「由這筆資料與先驗算出 95% 可信區間 $[0.42,0.61]$，所以在此先驗下 θ 有 95% 的機率落在其中。」(c)「同一筆資料下，可信區間與信賴區間必然是不同的區間。」(d)「可信區間的陳述依賴先驗的選擇。」

練習 9.10 (挑戰)　以二項概似 $p(k\mid p)=\binom{n}{k}p^k(1-p)^{n-k}$ 與先驗 $p\sim\mathrm{Beta}(\alpha,\beta)$ 完整驗證共軛性：(a) 計算證據 $p(k)=\int_0^1 p(k\mid p)\,p(p)\,\mathrm{d}p$（結果稱為 Beta–二項分布）。(b) 寫出後驗密度並確認它是 $\mathrm{Beta}(\alpha+k,\beta+n-k)$。(c) 說明為何二項係數 $\binom{n}{k}$ 對後驗完全沒有影響。

練習 9.11 (挑戰)　單一觀測 $x\sim\mathcal{N}(\theta,\sigma^2)$，先驗為 Laplace 分布 $p(\theta)=\frac{1}{2b}e^{-|\theta|/b}$。(a) 寫出 MAP 的目標函數（負對數後驗）。(b) 令 $\lambda=\sigma^2/b$，證明

$$\hat{\theta}_{\mathrm{MAP}}=\mathrm{sign}(x)\,\max(|x|-\lambda,\;0)$$

（軟閾值函數）。(c) 與高斯先驗 $\theta\sim\mathcal{N}(0,\tau^2)$ 的 MAP $\hat{\theta}=\tau^2x/(\tau^2+\sigma^2)$ 比較：為何 Laplace 先驗能把估計推成「恰好為零」而高斯先驗不能？這與 lasso 的稀疏性有何關係？

練習 9.12 (挑戰·程式)　用 `scipy.stats.beta` 完成：先驗 $\mathrm{Beta}(2,2)$，觀察 20 次中 13 次成功。(a) 計算後驗參數、後驗均值與眾數。(b) 求 95% 等尾可信區間。(c) 求「下一次成功」的後驗預測機率。(d) 把 13 個成功與 7 個失敗打亂順序逐筆更新，驗證與批次更新一致。(e)（選做）畫出先驗與後驗密度曲線。

9.8 練習題詳解

練習 9.1 詳解　(a) 由定理 9.3：後驗為 $\text{Beta}(2+6,\ 2+4) = \text{Beta}(8, 6)$。
(b) 套 Beta 分布公式：

$$\text{均值} = \frac{8}{14} = \frac{4}{7} \approx 0.571, \qquad \text{眾數} = \frac{8-1}{8+6-2} = \frac{7}{12} \approx 0.583,$$

$$\text{Var}(p \mid \mathcal{D}) = \frac{8 \times 6}{14^2 \times 15} = \frac{48}{2940} \approx 0.0163, \qquad \text{標準差} \approx 0.128.$$

(c) MLE 為 0.6，後驗均值 0.571 介於先驗均值 0.5 與 MLE 之間。由命題 9.4：$\mathbb{E}[p \mid \mathcal{D}] = \frac{4}{14} \times 0.5 + \frac{10}{14} \times 0.6 = \frac{4}{7}$。先驗攜帶 4 筆偽觀察、資料有 10 筆真觀察，估計被往 0.5 收縮，但資料權重（10/14）較大，結果偏向 MLE。

練習 9.2 詳解　(a) 概似：$\mathbb{P}(\text{三連正} \mid 0.5) = 0.5^3 = 0.125$；$\mathbb{P}(\text{三連正} \mid 0.9) = 0.9^3 = 0.729$。先驗乘概似：

$$0.75 \times 0.125 = 0.09375, \qquad 0.25 \times 0.729 = 0.18225,$$

證據 $p(\mathcal{D}) = 0.09375 + 0.18225 = 0.276$。歸一化：

$$\mathbb{P}(\text{公正} \mid \mathcal{D}) = \frac{0.09375}{0.276} \approx 0.340, \qquad \mathbb{P}(p = 0.9 \mid \mathcal{D}) = \frac{0.18225}{0.276} \approx 0.660.$$

三個正面就把信念從「3 比 1 賭公正」翻成「約 2 比 1 賭灌鉛」。
(b) 後驗預測 = 對兩個假設的預測取後驗加權平均：

$$\mathbb{P}(\text{下一擲為正} \mid \mathcal{D}) = 0.340 \times 0.5 + 0.660 \times 0.9 \approx 0.764.$$

注意它不等於任何單一假設的預測——不確定性被誠實地平均進來了。

練習 9.3 詳解　(a) $\text{Beta}(1+9,\ 1+1) = \text{Beta}(10, 2)$。
(b) Laplace 接續律（例 9.9）：

$$\mathbb{P}(\text{下一位有效} \mid \mathcal{D}) = \frac{k+1}{n+2} = \frac{10}{12} = \frac{5}{6} \approx 0.833.$$

（等於後驗均值 10/12。）
(c) 全部有效時 MLE 為 $10/10 = 1$：宣稱下一位**必定**有效——僅憑 10 筆資料就排除失敗，顯然過度自信（零頻率問題）。Laplace 接續律給 $(10+1)/(10+2) = 11/12 \approx 0.917$：仍然看好，但為沒見過的失敗保留餘地。偽計數讓「沒觀察到」不被誤讀成「不可能」，因此更合理。

練習 9.4 詳解　(a) 需要先驗均值 $\alpha/(\alpha+\beta) = 0.3$ 且偽觀察總數 $\alpha + \beta = 20$，解得 $\alpha = 6$、$\beta = 14$，即 $\text{Beta}(6, 14)$。

(b) 觀察 $k=20$、$n=50$：後驗 $\mathrm{Beta}(6+20,\ 14+30)=\mathrm{Beta}(26,44)$，後驗均值

$$\mathbb{E}[p\mid\mathcal{D}]=\frac{26}{70}=\frac{13}{35}\approx 0.371.$$

(c) 凸組合公式：$w=\frac{\alpha+\beta}{\alpha+\beta+n}=\frac{20}{70}=\frac{2}{7}$，MLE $\hat p=0.4$，故 $w\,m_0+(1-w)\,\hat p=\frac{2}{7}\times 0.3+\frac{5}{7}\times 0.4=\frac{2.6}{7}\approx 0.371$，與 (b) 一致（精確值同為 26/70）。

練習 9.5 詳解　(a) 由定義與 Beta 函數的性質：

$$\mathbb{E}[p]=\int_0^1 p\cdot\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\,\mathrm{d}p=\frac{1}{B(\alpha,\beta)}\int_0^1 p^{\alpha}(1-p)^{\beta-1}\,\mathrm{d}p=\frac{B(\alpha+1,\beta)}{B(\alpha,\beta)}.$$

以 Γ 展開並用 $\Gamma(z+1)=z\,\Gamma(z)$：

$$\frac{B(\alpha+1,\beta)}{B(\alpha,\beta)}=\frac{\Gamma(\alpha+1)\,\Gamma(\beta)}{\Gamma(\alpha+\beta+1)}\cdot\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}=\frac{\alpha\,\Gamma(\alpha)}{\Gamma(\alpha)}\cdot\frac{\Gamma(\alpha+\beta)}{(\alpha+\beta)\,\Gamma(\alpha+\beta)}=\frac{\alpha}{\alpha+\beta}.$$

(b) 對數密度（略去常數）為 $\ell(p)=(\alpha-1)\ln p+(\beta-1)\ln(1-p)$，

$$\ell'(p)=\frac{\alpha-1}{p}-\frac{\beta-1}{1-p}=0\iff(\alpha-1)(1-p)=(\beta-1)p\iff p^{*}=\frac{\alpha-1}{\alpha+\beta-2}.$$

且 $\ell''(p)=-\dfrac{\alpha-1}{p^2}-\dfrac{\beta-1}{(1-p)^2}<0$（$\alpha,\beta>1$），故 p^{*} 確為唯一極大點。
(c) 直接相減、通分：

$$\text{眾數}-\text{均值}=\frac{\alpha-1}{\alpha+\beta-2}-\frac{\alpha}{\alpha+\beta}=\frac{(\alpha-1)(\alpha+\beta)-\alpha(\alpha+\beta-2)}{(\alpha+\beta-2)(\alpha+\beta)}=\frac{\alpha-\beta}{(\alpha+\beta-2)(\alpha+\beta)},$$

分母為正（$\alpha+\beta>2$），故 $\alpha>\beta$ 時眾數 $>$ 均值：分布往左拖尾，均值被尾巴拉低——例 9.6 的 $\mathrm{Beta}(8,4)$ 正是實例。

練習 9.6 詳解　第一批更新後（定理 9.3）：

$$p\mid\mathcal{D}_1\sim\mathrm{Beta}(\alpha+k_1,\ \beta+n_1-k_1).$$

把它當作先驗、餵入第二批（k_2 成功、n_2-k_2 失敗），再次套定理 9.3：

$$p\mid\mathcal{D}_1,\mathcal{D}_2\sim\mathrm{Beta}\big(\alpha+k_1+k_2,\ \beta+(n_1-k_1)+(n_2-k_2)\big).$$

另一方面，兩批合併是 n_1+n_2 次中 k_1+k_2 次成功，一次更新給出

$$\mathrm{Beta}\big(\alpha+k_1+k_2,\ \beta+(n_1+n_2)-(k_1+k_2)\big),$$

兩者完全相同。■
根本原因：獨立試驗的概似連乘後指數相加，$p^{k_1}(1-p)^{n_1-k_1}\cdot p^{k_2}(1-p)^{n_2-k_2}=p^{k_1+k_2}(1-p)^{(n_1+n_2)-(k_1+k_2)}$，而加法滿足交換律與結合律，分幾批、什麼順序都不影響。對線上學習系統的意義：不必保存歷史資料，只需維護 (α,β) 兩個數字，串流資料隨到隨學，結果與存下全部資料重算一次完全相同——記憶體與計算都是常數級。

練習 9.7 詳解　(a) 先驗精度 $1/\tau_0^2 = 1/4 = 0.25$，資料精度 $n/\sigma^2 = 5/1 = 5$。由定理 9.5：

$$\frac{1}{\tau_n^2} = 0.25 + 5 = 5.25 = \frac{21}{4} \implies \tau_n^2 = \frac{4}{21} \approx 0.190, \quad \tau_n \approx 0.436,$$

$$\mu_n = \tau_n^2 \left(\frac{20}{4} + \frac{5 \times 22}{1} \right) = \frac{4}{21} \times (5 + 110) = \frac{460}{21} \approx 21.905.$$

後驗為 $\mathcal{N}(21.905,\ 0.190)$。
(b) 95% 等尾可信區間：

$$21.905 \pm 1.96 \times 0.436 = 21.905 \pm 0.855 = [21.05,\ 22.76].$$

(c) 頻率學派信賴區間：

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 22 \pm 1.96 \times \frac{1}{\sqrt{5}} = 22 \pm 0.877 = [21.12,\ 22.88].$$

比較：貝氏區間的中心從 22 被先驗往 20 拉到 21.905（先驗權重 $0.25/5.25 \approx 4.8\%$）；寬度較窄（1.71 對 1.75），因為先驗貢獻了額外精度（$5.25 > 5$）。另外別忘了兩者陳述不同：(b) 說「θ 有 95% 機率在裡面」，(c) 說「這套程序長期有 95% 的覆蓋率」。

練習 9.8 詳解　(a) 命題 9.4（推導重述）：

$$\mathbb{E}[p \mid \mathcal{D}] = \frac{\alpha + k}{\alpha + \beta + n} = \frac{(\alpha + \beta)\, m_0 + n\, \hat{p}}{\alpha + \beta + n} = w\, m_0 + (1 - w)\, \hat{p}, \qquad w = \frac{\alpha + \beta}{\alpha + \beta + n}.$$

(b) α, β 固定時 $w = \dfrac{\alpha + \beta}{\alpha + \beta + n} \to 0$（$n \to \infty$），故

$$\mathbb{E}[p \mid \mathcal{D}] = w\, m_0 + (1 - w)\, \hat{p} \ \to\ 0 \cdot m_0 + 1 \cdot \hat{p} = \hat{p}.$$

先驗的影響以 $O(1/n)$ 的速度消失。
(c) 記 $a = \alpha + k$、$b = \beta + n - k$，則 $a + b = \alpha + \beta + n$。由算幾不等式 $ab \le (a+b)^2/4$：

$$\mathrm{Var}(p \mid \mathcal{D}) = \frac{ab}{(a+b)^2(a+b+1)} \le \frac{(a+b)^2/4}{(a+b)^2(a+b+1)} = \frac{1}{4(\alpha + \beta + n + 1)} \ \longrightarrow\ 0.$$

解讀：不論先驗為何（只要沒把真值判零機率），後驗終將**集中**——均值收斂到 MLE、變異數歸零，塌縮成真值附近的尖峰。先驗的角色是資料少時的正則化，不是永遠的枷鎖。

練習 9.9 詳解　(a) **錯誤**。頻率框架中 θ 是固定常數；對算出的特定區間，θ 要嘛在、要嘛不在，沒有「95% 的機率」可言。95% 描述的是**程序**：大量重複「抽樣、算區間」，約 95% 的區間會蓋住 θ。
(b) **正確**。這正是可信區間的定義：$\mathbb{P}(\theta \in [0.42, 0.61] \mid \mathcal{D}) = 0.95$——在貝氏框架中 θ 有後驗分布，這句話合法而且就是它的本意。

(c) **錯誤**。兩者可以數值重合：例 9.8(c) 中，當常態模型的先驗趨於平坦（$\tau_0^2 \to \infty$），可信區間收斂到與信賴區間相同的端點。但即使數值相同，兩句陳述的意義仍然不同（一句關於 θ 的信念、一句關於程序的覆蓋率）。
(d) **正確**。後驗由先驗與概似共同決定；換一個先驗，同一筆資料會給出不同的可信區間。這是貝氏方法「透明的主觀性」：先驗攤開來被檢視，既是代價也是誠實。

練習 9.10 詳解　(a) 把常數提出積分：

$$p(k) = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} \cdot \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\,\mathrm{d}p = \frac{\binom{n}{k}}{B(\alpha,\beta)} \int_0^1 p^{\alpha+k-1}(1-p)^{\beta+n-k-1}\,\mathrm{d}p.$$

剩下的積分正是 Beta 函數的定義：

$$p(k) = \binom{n}{k} \frac{B(\alpha+k,\ \beta+n-k)}{B(\alpha,\beta)}, \qquad k = 0, 1, \ldots, n,$$

此即 **Beta–二項分布**（Beta-binomial distribution）。
(b) 後驗 = 先驗 × 概似 ÷ 證據：

$$p(p \mid k) = \frac{\binom{n}{k} p^k (1-p)^{n-k} \cdot p^{\alpha-1}(1-p)^{\beta-1} / B(\alpha,\beta)}{\binom{n}{k}\, B(\alpha+k, \beta+n-k)/B(\alpha,\beta)} = \frac{p^{\alpha+k-1}(1-p)^{\beta+n-k-1}}{B(\alpha+k,\ \beta+n-k)},$$

恰為 $\mathrm{Beta}(\alpha+k,\ \beta+n-k)$ 的密度。共軛性成立，且歸一化常數自動就位。■
(c) $\binom{n}{k}$ 不含 p，在分子與分母（證據）同時出現而完全消去。因此「把資料當成特定序列」（概似 $p^k(1-p)^{n-k}$）或「只記錄成功總數」（概似多乘 $\binom{n}{k}$），後驗一模一樣：概似中不依賴參數的因子對推論毫無影響。

練習 9.11 詳解　(a) 負對數後驗（略去與 θ 無關的常數）：

$$f(\theta) = -\log p(\theta \mid x) + \text{const} = \frac{(x-\theta)^2}{2\sigma^2} + \frac{|\theta|}{b},$$

MAP 即 $\arg\min_\theta f(\theta)$。$f$ 是凸函數（凸加凸），最小值唯一。
(b) 分三種情形。**情形一** $\theta > 0$：$f'(\theta) = \dfrac{\theta - x}{\sigma^2} + \dfrac{1}{b} = 0$ 給 $\theta = x - \sigma^2/b = x - \lambda$，落在 $\theta > 0$ 區域需 $x > \lambda$。**情形二** $\theta < 0$：同理 $\theta = x + \lambda$，需 $x < -\lambda$。**情形三** $|x| \le \lambda$：檢查 $\theta = 0$ 的單邊導數，

$$f'(0^+) = \frac{-x}{\sigma^2} + \frac{1}{b} \ge 0 \iff x \le \lambda, \qquad f'(0^-) = \frac{-x}{\sigma^2} - \frac{1}{b} \le 0 \iff x \ge -\lambda,$$

兩者同時成立，凸函數在 0 處左減右增，故 $\hat\theta = 0$。三種情形合寫即軟閾值（soft thresholding）：

$$\hat\theta_{\mathrm{MAP}} = \mathrm{sign}(x)\ \max(|x| - \lambda,\ 0), \qquad \lambda = \sigma^2/b. \qquad \blacksquare$$

(c) 高斯先驗的 MAP 是**等比例縮小**：只要 $x \neq 0$，$\hat\theta = \tau^2 x/(\tau^2+\sigma^2)$ 永遠不為零。Laplace 先驗的密度在原點有「尖角」（不可微），懲罰 $|\theta|$ 在 0 處的次梯度形成**死區**：凡 $|x| \le \lambda$ 的

觀測都被壓成 $\hat{\theta} = 0$。推廣到迴歸即 lasso 的稀疏性——訊號弱的係數恰好歸零（自動特徵選擇），訊號強的只被平移 λ。L2 收縮、L1 篩選，根源都在先驗的形狀。

練習 9.12 詳解　參考程式與輸出：

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

a0, b0 = 2, 2                       # 先驗 Beta(2,2)
k, n = 13, 20
a1, b1 = a0 + k, b0 + n - k         # (a) 後驗參數
post = stats.beta(a1, b1)

print((a1, b1))                     # (15, 9)
print(post.mean())                  # 0.625（後驗均值）
print((a1 - 1) / (a1 + b1 - 2))     # 0.6364（眾數 = MAP）
print(post.ppf([0.025, 0.975]))     # (b) [0.4273, 0.8029]
print(a1 / (a1 + b1))               # (c) 0.625（後驗預測）
# (d) 打亂順序逐筆更新
rng = np.random.default_rng(0)
xs = np.array([1]*13 + [0]*7)
rng.shuffle(xs)
a, b = a0, b0
for x in xs.tolist():
    a, b = a + x, b + (1 - x)
print((a, b))                       # (15, 9)：與批次一致
# (e) 畫先驗與後驗
import matplotlib.pyplot as plt
grid = np.linspace(0, 1, 400)
plt.plot(grid, stats.beta(a0, b0).pdf(grid), "--", label="prior")
plt.plot(grid, post.pdf(grid), label="posterior")
plt.legend(); plt.xlabel("p"); plt.show()
```

逐項對照：(a) 後驗 $\text{Beta}(15, 9)$，均值 0.625、眾數 ≈ 0.636。(b) 可信區間 $[0.427,\ 0.803]$。(c) 後驗預測 0.625，等於後驗均值。(d) 任何順序都得到 $(15, 9)$——序貫與批次一致。(e) 後驗比先驗高聳、集中於 0.6 附近：13 比 7 的證據把「一半一半」的先驗推向「六成多」。

後驗在手，不確定性有了形狀——下一步，是讓它在更大的模型與決策裡工作。