機率與統計：從入門到精通

# 第 7 章　估計與最大概似法

Estimation and Maximum Likelihood

難度：進階

到目前為止，我們一直站在機率的立場：**給定模型**，**預測資料**——已知硬幣公正，計算擲出 7 個正面的機率。從本章起，我們正式跨入統計的世界，把箭頭整個反轉：**給定資料**，**推斷模型**——觀察到 7 個正面，回頭猜測這枚硬幣的參數。這個反向問題的核心任務是**估計**（estimation）：用樣本猜測未知參數。而它的主導方法——**最大概似法**——不只是一項統計技術，更是幾乎所有機器學習損失函數背後的推導原理：交叉熵、均方誤差，全都是它的化身。這是全書最重要的章節之一。

學習目標

讀完本章後，你將能夠：

- 解釋為什麼**估計量**本身是隨機變數、什麼是**抽樣分布**，用**偏差**、**變異數**與**均方誤差**（MSE）評價估計量，並證明 $\mathrm{MSE} = \mathrm{Bias}^2 + \mathrm{Var}$。
- 寫出 i.i.d. 模型的**概似函數**與**對數概似**，並清楚分辨「概似」與「機率」的差別。
- 推導白努利、卜瓦松、指數、常態與均勻分布的**最大概似估計**（MLE），包括微積分失效的邊界情形。
- 敘述 MLE 的三大大樣本性質（一致性、漸近常態、漸近有效），以及**費雪資訊**與 **Cramér–Rao 下界**扮演的角色。
- 說明最小化**負對數概似**（NLL）如何生成交叉熵與均方誤差等標準機器學習損失函數。
- 用 Python / SciPy 以數值最佳化求 MLE，並以模擬驗證偏差與校正的理論。

## 7.1 估計量與其品質

先把舞台佈置清楚：假設資料由某個機率模型產生，$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p(x \mid \theta)$，分布的**形式**已知（例如「這是白努利分布」），但**參數** $\theta$ 未知（正面機率是多少？）。統計的第一個任務，是設計「從資料算出 $\theta$ 的猜測值」的程序。前六章固定 $\theta$、對資料做推理；本章固定資料、對 $\theta$ 做推理——方向相反，用的仍是同一套機率語言。

**定義 7.1　估計量、估計值與抽樣分布**

**估計量**（estimator）$\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ 是一個**資料的函數**，用來估計未知參數 $\theta$。把實際觀察到的數字代入後得到的結果稱為**估計值**（estimate）。因為 $\hat{\theta}$ 依賴隨機樣本，它本身是一個**隨機變數**：換一份樣本，就得到不同的估計值。$\hat{\theta}$ 的機率分布稱為**抽樣分布**（sampling distribution）。

**直觀理解**

估計量是**食譜**，估計值是**端出來的菜**：「觀測值加總除以 $n$」是食譜（$\bar{X}$），用今天的資料算出 3.72 是今天端出的菜。明天換一份樣本，同一份食譜會端出不同的菜——結果隨樣本而變，所以它是隨機變數、有自己的分布（抽樣分布）。評價估計量因此不能只看某一次的表現，而要看它**跨所有可能樣本**的整體表現：平均而言準不準？抖動大不大？這正是偏差與變異數要回答的問題。

最經典的估計量是**樣本均值** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$（用來估計期望值 $\mu$）。由期望值的線性，$\mathbb{E}[\bar{X}] = \mu$——平均而言正中目標；再由獨立性，$\mathrm{Var}(\bar{X}) = \sigma^2/n$，其標準差 $\mathrm{SE}(\bar{X}) = \sigma/\sqrt{n}$ 稱為**標準誤**（standard error）：樣本越大，抽樣分布越集中（完整推導見練習 7.5）。注意 $\sqrt{n}$ 的代價——想把精度提高 10 倍，樣本得多 100 倍。有了「估計量是隨機變數」的觀念，就可以定義評價它的三把尺。

**定義 7.2　偏差、變異數與均方誤差**

設 $\hat{\theta}$ 是 $\theta$ 的估計量。

1. **偏差**（bias）：$\mathrm{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$，衡量**系統性的偏移**。若 $\mathrm{Bias}(\hat{\theta}) = 0$，稱 $\hat{\theta}$ 為**不偏估計量**（unbiased estimator）——平均而言正確。
2. **變異數**（variance）：$\mathrm{Var}(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\big]$，衡量估計值**跨樣本的抖動**。
3. **均方誤差**（mean squared error, MSE）：$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta} - \theta)^2\big]$，衡量**總誤差**：離真值平均有多遠（平方尺度）。

三把尺不是獨立的——它們被一條漂亮的恆等式綁在一起。

**定理 7.3　MSE 分解**

對任何估計量 $\hat{\theta}$，

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Bias}(\hat{\theta})^2 + \mathrm{Var}(\hat{\theta}).$$

**證明**　記 $m = \mathbb{E}[\hat{\theta}]$（注意 $\theta$ 是常數、$m$ 也是常數）。在平方裡「加減同一項」：

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta} - m + m - \theta)^2\big] = \mathbb{E}\big[(\hat{\theta} - m)^2\big] + 2(m - \theta)\,\mathbb{E}[\hat{\theta} - m] + (m - \theta)^2.$$

中間的交叉項為零，因為 $\mathbb{E}[\hat{\theta} - m] = \mathbb{E}[\hat{\theta}] - m = 0$。剩下的第一項是 $\mathrm{Var}(\hat{\theta})$，第三項是 $\mathrm{Bias}(\hat{\theta})^2$。 ■

**核心觀念**

$\mathrm{MSE} = \mathrm{Bias}^2 + \mathrm{Var}$ 是**偏差–變異數權衡**（bias–variance trade-off，第 10 章）的種子，這個權衡支配著整個機器學習：更有彈性的估計量（或模型）偏差較低，但對樣本更敏感、變異數更高；更簡單的則相反。因此**最好的估計量不一定不偏**——刻意接受一點偏差，若能砍掉大量變異數，總誤差反而下降，這正是**正則化**（regularization）在做的事。純粹的最大概似天生偏向「低偏差、高變異數」的那一端；各種正則化手段，本質上都是往回拉的配重。

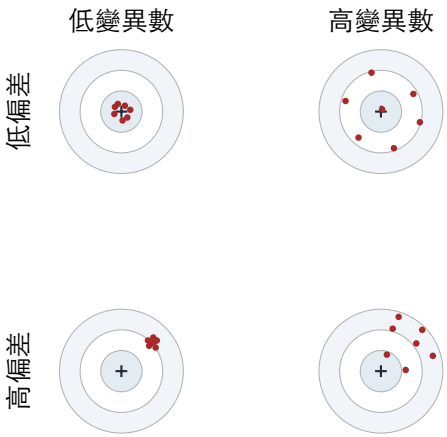



圖 7.1: 偏差與變異數的靶心圖。靶心（十字）是真值 $\theta$，每個紅點是「換一份樣本」得到的一個估計值 $\hat{\theta}$。偏差衡量點群中心偏離靶心多遠；變異數衡量點群本身散得多開。MSE 同時懲罰兩者：右下角最糟，左上角最好，而左下與右上何者較佳，取決於 $\mathrm{Bias}^2$ 與 Var 誰大。

接下來的例子是統計學裡最著名的偏差案例：為什麼樣本變異數要除以 $n-1$ 而不是 $n$？

**例 7.1　除以 $n$ 還是 $n-1$？——Bessel 校正**

設 $X_1, \dots, X_n$ i.i.d.，期望值 $\mu$、變異數 $\sigma^2$ 皆未知。比較兩個 $\sigma^2$ 的估計量：

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \text{與} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

**解**　關鍵是計算 $\mathbb{E}\big[\sum_i (X_i - \bar{X})^2\big]$。先建立一個恆等式：把 $X_i - \mu$ 拆成 $(X_i - \bar{X}) + (\bar{X} - \mu)$，平方後對 $i$ 相加：

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + 2(\bar{X} - \mu)\underbrace{\sum_{i=1}^{n}(X_i - \bar{X})}_{=\,0} + n(\bar{X} - \mu)^2,$$

中間那項為零（樣本對自己均值的偏差總和恆為零），移項即得 $\sum_i (X_i - \bar{X})^2 = \sum_i (X_i - \mu)^2 - n(\bar{X} - \mu)^2$。兩邊取期望值：左邊第一項 $\mathbb{E}[(X_i - \mu)^2] = \sigma^2$ 共 $n$ 項，貢獻 $n\sigma^2$；第二項 $\mathbb{E}[(\bar{X} - \mu)^2] = \mathrm{Var}(\bar{X}) = \sigma^2/n$，貢獻 $n \cdot \sigma^2/n = \sigma^2$。因此

$$\mathbb{E}\Big[\sum_{i=1}^{n}(X_i - \bar{X})^2\Big] = n\sigma^2 - \sigma^2 = (n-1)\,\sigma^2 \quad \Longrightarrow \quad \mathbb{E}[\hat{\sigma}_n^2] = \frac{n-1}{n}\,\sigma^2 < \sigma^2, \quad \mathbb{E}[S^2] = \sigma^2.$$

除以 $n$ 的版本**系統性偏低**，偏差為 $-\sigma^2/n$；改除 $n-1$（**Bessel 校正**，Bessel's correction）則不偏。直觀原因：$\bar{X}$ 是「追著這份樣本跑」的量（$\bar{X}$ 正是讓 $\sum_i (X_i - a)^2$ 最小的 $a$），所以對 $\bar{X}$ 的偏差平方和永遠不超過對真值 $\mu$ 的，平均少掉的恰好是一個 $\sigma^2$——我們用資料估計了 $\mu$，花掉了一個自由度。極端情形 $n=1$ 最能說明問題：$\hat{\sigma}_1^2 = 0$ 恆成立（一個點對自己毫無散布），顯然荒謬；$S^2$ 的分母 $n-1=0$ 則誠實地拒絕回答。

預告：第 7.3 節會證明，常態分布變異數的**最大概似估計正是除以 $n$ 的版本**——所以 MLE 略有偏差。這是一個具體提醒：**MLE 不自動不偏**。

**常見陷阱**

「不偏」聽起來像道德標準，彷彿有偏就是壞——這是最常見的誤解。兩個矯正：**其一**，不偏但高變異數可以很糟——「丟掉其餘資料、只用 $X_1$ 估計 $\mu$」也不偏，但變異數是 $\bar{X}$ 的 $n$ 倍；**其二**，有偏可以更準——下一個例子就有帶偏差的估計量在 MSE 上獲勝。另有一個實務陷阱：NumPy 的 `np.var()` 預設除以 $n$（`ddof=0`），pandas 的 `.var()` 預設除以 $n-1$（`ddof=1`）——同一份資料、兩個函式庫、兩個答案，混用而不自知是真實發生的分析錯誤。

**例 7.2　接受一點偏差：拉普拉斯平滑**

擲一枚正面機率為 $p$ 的硬幣 $n$ 次，正面次數 $K \sim \mathrm{Bin}(n,p)$，$\mathbb{E}[K] = np$、$\mathrm{Var}(K) = np(1-p)$（第 4 章）。比較兩個估計量：

$$\hat{p} = \frac{K}{n} \qquad \text{與} \qquad \tilde{p} = \frac{K+1}{n+2},$$

後者稱為**拉普拉斯平滑**（Laplace smoothing）：假裝多看了兩次投擲、一正一反，把估計值往 $\frac{1}{2}$ 拉。

**解**　對 $\hat{p}$：$\mathbb{E}[\hat{p}] = p$（不偏），故 $\mathrm{MSE}(\hat{p}) = \mathrm{Var}(\hat{p}) = p(1-p)/n$。對 $\tilde{p}$：

$$\mathrm{Bias}(\tilde{p}) = \frac{np+1}{n+2} - p = \frac{1-2p}{n+2}, \qquad \mathrm{Var}(\tilde{p}) = \frac{np(1-p)}{(n+2)^2},$$

由 MSE 分解得 $\mathrm{MSE}(\tilde{p}) = \big[np(1-p) + (1-2p)^2\big]/(n+2)^2$。代入 $n=10$、$p=0.5$：$\mathrm{MSE}(\hat{p}) = 0.025$ 而 $\mathrm{MSE}(\tilde{p}) = 2.5/144 \approx 0.0174$——帶偏差的 $\tilde{p}$ 總誤差**低了約三成**！代價在邊緣：$p=0.05$ 時 $\mathrm{MSE}(\hat{p}) \approx 0.0048 < \mathrm{MSE}(\tilde{p}) \approx 0.0089$。圖 7.2 畫出完整比較：中央大片區域由 $\tilde{p}$ 獲勝。「往中間拉一點」正是正則化的原型——第 9 章會看到，$\tilde{p}$ 其實是一個貝氏後驗均值。

## 7.2　概似函數

估計的第一步是替「$\theta$ 解釋資料的能力」打分數。工具早就在手上：聯合機率 $p(\mathcal{D} \mid \theta)$。過去我們固定 $\theta$、把它讀成「資料的機率」；現在資料已經觀察到、固定不動，我們把**同一條公式**讀成 $\theta$ 的函數。換個讀法，就換了一個世界。

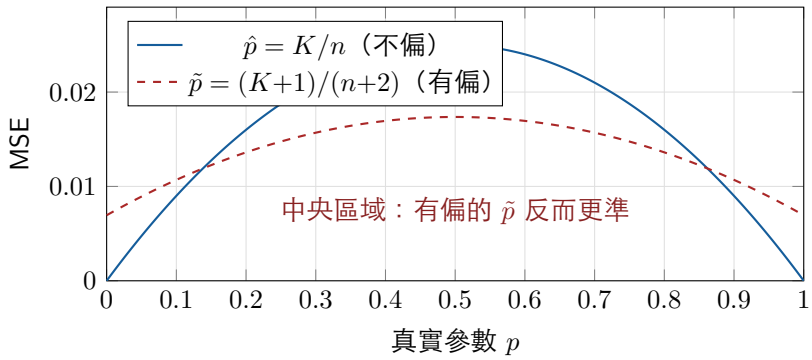



圖 7.2: $n = 10$ 時兩個估計量的 MSE 對照。除了 $p$ 非常接近 0 或 1 的窄帶之外，帶偏差的拉普拉斯平滑估計量全面勝出：它用一點偏差換掉了更多變異數。

**定義 7.4　概似函數與對數概似**

給定觀察資料 $\mathcal{D} = \{x_1, \dots, x_n\}$ 與參數為 $\theta$ 的模型，**概似函數**（likelihood function）定義為「觀察到這份資料的機率，視為 $\theta$ 的函數」；其對數稱為**對數概似**（log-likelihood）。i.i.d. 情形下：

$$L(\theta) = p(\mathcal{D} \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta), \qquad \ell(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln p(x_i \mid \theta).$$

連續資料時 $p(x_i \mid \theta)$ 取機率密度值，$L(\theta)$ 是密度的乘積；「讀成 $\theta$ 的函數」的精神完全相同。

**例 7.3　硬幣資料的概似函數**

擲一枚硬幣 10 次，得到序列「正正反正正正反正正反」——7 個正面、3 個反面。設正面機率為 $p$，寫出概似函數並比較幾個候選 $p$ 值。

解　各次投擲獨立，這個特定序列的機率是每次結果機率的乘積：7 個因子 $p$、3 個因子 $(1-p)$，

$$L(p) = p^{7}(1-p)^{3}.$$

代入幾個候選值（單位 $10^{-3}$）：

| $p$ | 0.3 | 0.5 | 0.7 | 0.9 | 1.0 |
|---|---|---|---|---|---|
| $L(p) \times 10^{3}$ | 0.075 | 0.977 | 2.224 | 0.478 | 0 |

$p = 0.7$ 的解釋力大約是 $p = 0.5$ 的 2.3 倍、$p = 0.3$ 的 30 倍。$p = 1$ 的概似是零：資料裡有反面，「永遠出正面」的模型完全不可能產生這份資料。圖 7.3 畫出完整曲線——一座在 $p = 0.7$ 達到頂峰的單峰山。

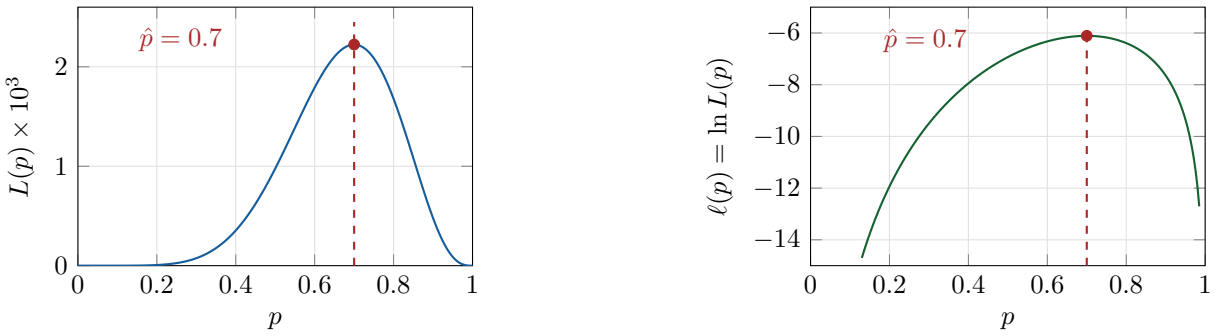



圖 7.3: 左：$n = 10$、$k = 7$ 的概似曲線 $L(p) = p^7(1-p)^3$。右：對數概似 $\ell(p) = 7\ln p + 3\ln(1-p)$。取對數把陡峭的山變成平緩的山，但**山頂位置一動不動**：兩者都在 $\hat{p} = 0.7$ 達到最大。

**常見陷阱**

**概似不是 $\theta$ 上的機率分布。**$L(\theta)$ 對 $\theta$ 積分（或加總）通常不等於 1。例如只擲一次得到正面：$L(p) = p$，$\int_0^1 p\,\mathrm{d}p = \frac{1}{2} \neq 1$。$L(\theta)$ 回答的是「**這個 $\theta$ 多能解釋我看到的資料**」，不是「$\theta$ 本身多可能為真」。想把概似升級成「$\theta$ 的機率」，必須把 $\theta$ 當成隨機變數、引入先驗，經由貝氏定理（第 2 章）得到後驗 $p(\theta \mid \mathcal{D}) \propto L(\theta)\,p(\theta)$——這正是先驗存在要糾正的錯誤：把 $P(\text{資料} \mid \text{假設})$ 誤讀成 $P(\text{假設} \mid \text{資料})$。記住口訣：**機率在資料上；我們只是把它讀成 $\theta$ 的函數。**

**直觀理解**

把概似想成一張**排行榜**：每個候選 $\theta$ 都上場，用「若真是我，這份資料出現的機率」當作得分。榜上的分數只有**相對比較**有意義：$L(0.7)/L(0.5) \approx 2.3$ 說「0.7 的解釋力是 0.5 的 2.3 倍」，但單一分數 $L(0.7) = 0.0022$ 本身沒有絕對意義——樣本一多，所有概似都變得極小（一千次投擲的任何特定序列機率都微乎其微），排行榜依然分得出高下。連續資料時分數是密度值，甚至可以大於 1。統計學家因此總在比較概似的**比值**或**差**（對數尺度），而非絕對值。

## 7.3 最大概似估計

排行榜做好了，接下來的動作再自然不過：**選第一名**。

**定義 7.5　最大概似估計**

**最大概似估計**（maximum likelihood estimation, MLE）選擇讓觀察資料最可能出現的參數：

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \sum_{i=1}^{n} \ln p(x_i \mid \theta).$$

第二個等號成立是因為 ln 嚴格遞增：取對數**不改變最大值的位置**，卻把乘積變成和。

**核心觀念**

實務上永遠對**對數概似**做最佳化，理由有三。**（一）數值穩定**：一千筆資料的概似是一千個小於 1 的數的乘積，早在第幾十項就下溢成浮點數的 0；取對數後變成和，安然無恙。**（二）微分容易**：和的導數是導數的和，每筆資料貢獻獨立的一項——這個「逐筆相加」的結構正是隨機梯度下降（SGD）能用小批次訓練神經網路的原因。**（三）位置不變**：ln 嚴格遞增，$\arg\max$ 不動（圖 7.3）。求解只有兩條路：解析法（令 $\ell'(\theta)=0$）與數值法（對 $-\ell(\theta)$ 做梯度下降）——本章例題走前者，深度學習走後者，原理相同。

**直觀理解**

MLE 問了一個漂亮的問題：「**在所有 $\theta$ 的設定中，哪一個會讓我看到的資料最不令人意外？**」然後就選那個。幾何圖像：參數 $\theta$ 像一顆旋鈕，每個位置都對應一個機率分布；轉動旋鈕，直到模型把機率質量**堆在資料實際落下的地方**。資料聚在哪裡，山峰就該建在哪裡。MLE 在簡單模型上恢復的全是常識（頻率、平均值），而同一條原理可以原封不動地推廣到直覺無法處理的模型——邏輯迴歸、隱馬可夫模型、上兆參數的語言模型。常識無法外推，原理可以。

**例 7.4　硬幣的 MLE**

擲硬幣 $n$ 次得 $k$ 個正面，求 $p$ 的最大概似估計。

解　對數概似為 $\ell(p)=k\ln p+(n-k)\ln(1-p)$。令導數為零：

$$\frac{\mathrm{d}\ell}{\mathrm{d}p}=\frac{k}{p}-\frac{n-k}{1-p}=0 \quad\Longrightarrow\quad k(1-p)=(n-k)\,p \quad\Longrightarrow\quad \hat{p}=\frac{k}{n}.$$

二階導數 $\ell''(p)=-k/p^2-(n-k)/(1-p)^2<0$，$\ell$ 處處凹，故 $\hat{p}=k/n$ 確為唯一最大值（$0<k<n$ 時）。MLE 恢復了常識：**觀察頻率**——10 擲 7 正就估 0.7，任何人都會這樣猜；價值在於這個猜法從此有了**原理背書**，而原理可以推廣到心算無法處理的模型。

**例 7.5　卜瓦松分布的 MLE**

某咖啡店記錄每分鐘來客數，8 分鐘的資料為 $3,5,2,4,6,4,3,5$。假設來客數 $X_i \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$，$p(x\mid\lambda)=e^{-\lambda}\lambda^x/x!$，求 $\lambda$ 的 MLE。

解　對數概似為 $\ell(\lambda)=\sum_i\big(-\lambda+x_i\ln\lambda-\ln x_i!\big)=-n\lambda+\big(\sum_i x_i\big)\ln\lambda+$ 常數，其中常數收納與 $\lambda$ 無關的 $-\sum_i\ln x_i!$。令導數為零：

$$\frac{\mathrm{d}\ell}{\mathrm{d}\lambda}=-n+\frac{\sum_i x_i}{\lambda}=0 \quad\Longrightarrow\quad \hat{\lambda}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}.$$

二階導數 $-\sum_i x_i/\lambda^2<0$，確為最大。卜瓦松的 MLE 就是**樣本均值**——與「$\lambda$ 是平均事件率」的詮釋完全吻合。代入資料：$\sum x_i=32$、$n=8$，故 $\hat{\lambda}=4.0$（每分鐘平均 4 位客人）。有了 $\hat{\lambda}$ 就能做**插入式預測**（plug-in prediction）：某一分鐘完全沒有客人的機率估計為 $e^{-4}\approx 0.018$。

**例 7.6　指數分布的 MLE**

測試 5 顆 LED 燈泡的壽命（單位：千小時），得 1.2, 0.7, 2.1, 0.4, 1.6。假設壽命 $X_i \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$，密度 $f(x\mid\lambda)=\lambda e^{-\lambda x}$（$x\ge 0$），求 $\lambda$ 的 MLE。

**解**　對數概似為 $\ell(\lambda)=\sum_i(\ln\lambda-\lambda x_i)=n\ln\lambda-\lambda\sum_i x_i$。令導數為零：

$$\frac{\mathrm{d}\ell}{\mathrm{d}\lambda}=\frac{n}{\lambda}-\sum_{i=1}^{n}x_i=0\quad\Longrightarrow\quad\hat\lambda=\frac{n}{\sum_i x_i}=\frac{1}{\bar x},$$

且 $\ell''(\lambda)=-n/\lambda^2<0$。速率的估計是**平均壽命的倒數**。代入資料：$\sum x_i=6.0$、$\bar x=1.2$，故 $\hat\lambda=1/1.2\approx 0.833$（每千小時 0.833 次故障）。順帶一提 MLE 的**不變性**（invariance）：若想估計的是變換後的量 $g(\theta)$，其 MLE 就是 $g(\hat\theta)$。平均壽命 $\mathbb{E}[X]=1/\lambda$ 的 MLE 因此是 $1/\hat\lambda=\bar x=1.2$ 千小時，不必重新最佳化。

**例 7.7　常態分布的 MLE：兩個參數一起估**

設 $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim}\mathcal{N}(\mu,\sigma^2)$，兩個參數皆未知，求 $(\mu,\sigma^2)$ 的 MLE。

**解**　密度為 $f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\!\big(-\frac{(x-\mu)^2}{2\sigma^2}\big)$，記 $v=\sigma^2$ 以免混淆，對數概似為

$$\ell(\mu,v)=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln v-\frac{1}{2v}\sum_{i=1}^{n}(x_i-\mu)^2.$$

對兩個參數分別取偏導數並令為零。先對 $\mu$：

$$\frac{\partial\ell}{\partial\mu}=\frac{1}{v}\sum_{i=1}^{n}(x_i-\mu)=0\quad\Longrightarrow\quad\hat\mu=\bar x,$$

與 $v$ 無關，乾淨俐落。再對 $v$（代入 $\hat\mu=\bar x$）：

$$\frac{\partial\ell}{\partial v}=-\frac{n}{2v}+\frac{1}{2v^2}\sum_{i=1}^{n}(x_i-\bar x)^2=0\quad\Longrightarrow\quad\hat\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar x)^2.$$

（在 $v=\hat\sigma^2$ 處二階導數為 $-n/(2\hat\sigma^4)<0$，確為最大。）注意分母：MLE 給的是**除以** $n$ 的樣本變異數——正是例 7.1 證明過**偏低**的那個版本，$\mathbb{E}[\hat\sigma^2]=\frac{n-1}{n}\sigma^2$。最大概似原理如此優雅，結果卻自帶偏差：再次確認 **MLE 不保證不偏**，它保證的是「對觀察資料的解釋力最大」。大樣本下 $\frac{n-1}{n}\to 1$，這個偏差隨 $n$ 消失——這是下一節「一致性」的預演。

**例 7.8　均勻分布 $U(0,\theta)$ 的 MLE：微積分失效的一課**

設 $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} U(0,\theta)$，密度 $f(x\mid\theta)=1/\theta$（$0\le x\le\theta$），求 $\theta$ 的 MLE，並計算其偏差。

**解**　概似函數要小心寫：每個因子在 $x_i \le \theta$ 時是 $1/\theta$，否則是 0。所以

$$L(\theta) = \begin{cases} \theta^{-n}, & \theta \ge x_{(n)} := \max_i x_i, \\ 0, & \theta < x_{(n)}. \end{cases}$$

若令導數為零：$\frac{\mathrm{d}}{\mathrm{d}\theta}\theta^{-n} = -n\theta^{-n-1}$ 在任何 $\theta > 0$ 都不為零——**微積分找不到答案**。改用圖形論證：概似曲線在 $\theta < x_{(n)}$ 時貼地為 0（模型連最大的觀測值都容納不下，資料不可能發生），在 $\theta = x_{(n)}$ 處跳升到 $x_{(n)}^{-n}$，其後隨 $\theta$ 嚴格遞減（模型越大，機率質量攤得越薄，資料越「不特別」）。最大值因此落在**不可微的邊界**上：$\hat\theta_{\mathrm{MLE}} = x_{(n)} = \max_i x_i$。接著算偏差。最大值的分布函數（利用獨立性）為 $\mathbb{P}(X_{(n)} \le t) = (t/\theta)^n$（$0 \le t \le \theta$），密度為 $nt^{n-1}/\theta^n$，故

$$\mathbb{E}[X_{(n)}] = \int_0^\theta t \cdot \frac{n t^{n-1}}{\theta^n}\,\mathrm{d}t = \frac{n}{\theta^n}\cdot\frac{\theta^{n+1}}{n+1} = \frac{n}{n+1}\,\theta.$$

偏差為 $-\theta/(n+1)$：MLE **必然低估**——樣本最大值永遠不會超過真正的上界 $\theta$，只會從下方逼近。修正方法與 MSE 的比較留給練習 7.10：把 $x_{(n)}$ 乘上 $(n+1)/n$ 即得不偏版本。

**常見陷阱**

「令導數為零」不是萬能鑰匙。動筆微分之前，先想清楚概似的**形狀**：它可微嗎？單峰嗎？最大值會不會在參數空間的邊界上？均勻分布（例 7.8）是邊界案例的經典代表。更日常的案例是硬幣資料 $k = 0$（全反面）：$\ell(p) = n\ln(1-p)$ 在 $(0,1)$ 內導數恆負，MLE 落在邊界 $\hat p = 0$——模型從此斷言「正面**永遠**不可能出現」，只因為 10 次裡沒看過。這是純 MLE 在小樣本上**過度擬合**（overfitting）的縮影：對沒見過的事件指定零機率，一旦它日後發生，交叉熵損失直接爆成無窮大。例 7.2 的拉普拉斯平滑正是解藥——語言模型的平滑技術、分類器的 label smoothing，都是同一帖藥方的後代。

## 7.4　MLE 的大樣本性質與費雪資訊

MLE 只是眾多估計方法之一（動差法、最小平方法、貝氏後驗均值都是別的選項），為什麼統計學家與機器學習如此偏愛它？答案藏在大樣本理論裡：當 $n$ 夠大，MLE 幾乎是「你能做到的最好」。

**命題 7.6　MLE 的大樣本性質：統計學家為何愛它**

在正則條件下（分布的支撐集不依賴 $\theta$、對數概似足夠平滑、真值在參數空間內部、模型可識別），最大概似估計 $\hat\theta_n$ 滿足：

1. **一致性**（consistency）：$\hat\theta_n \xrightarrow{P} \theta$——樣本越多，估計收斂到真值；
2. **漸近常態**（asymptotic normality）：$\sqrt{n}\,(\hat\theta_n - \theta) \xrightarrow{d} \mathcal{N}\big(0,\ I(\theta)^{-1}\big)$，其中 $I(\theta)$ 是下面定義的（單筆觀測）費雪資訊；實用寫法是 $\hat\theta_n \approx \mathcal{N}\big(\theta,\ \frac{1}{nI(\theta)}\big)$；

3. **漸近有效**（asymptotic efficiency）：其漸近變異數達到 Cramér–Rao 下界——不偏估計量可能達到的**最小變異數**。

一致性說「終究會對」；漸近常態說「誤差的形狀可預測」（第 8 章將用它建構信賴區間）；漸近有效說「沒有人能做得更好」。注意「正則條件」不是裝飾品：例 7.8 的均勻分布支撐集 $[0,\theta]$ 依賴 $\theta$，違反第一條，其 MLE 的誤差以 $1/n$ 的速度收斂（比常規的 $1/\sqrt{n}$ 還快！）且極限分布不是常態——正則性一破，整套漸近劇本都得重寫。

**定義 7.7　費雪資訊**

單筆觀測的**費雪資訊**（Fisher information）定義為對數概似導數（稱為分數，score）平方的期望值：

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln p(X\mid\theta)\right)^2\right] = -\,\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln p(X\mid\theta)\right],$$

兩種寫法在正則條件下相等。第二種寫法透露幾何意義：$I(\theta)$ 是對數概似在真值附近的**平均曲率**——概似山峰的尖銳程度。$n$ 筆 i.i.d. 觀測的總資訊為 $nI(\theta)$：資料越多，山峰越尖。

**定理 7.8　Cramér–Rao 下界**

在正則條件下，任何**不偏**估計量 $\hat\theta$ 的變異數都不可能低於費雪資訊的倒數：

$$\operatorname{Var}(\hat\theta) \;\geq\; \frac{1}{n\,I(\theta)}.$$

達到此下界的不偏估計量稱為**有效**（efficient）。

**直觀理解**

把對數概似想成一座山，真值 $\theta$ 站在山頂附近。**山越尖，資料對參數的「指認」越明確**：往旁邊挪一點點，解釋力就大幅下降，資料強烈反對；山越平，鄰近的參數解釋力幾乎一樣好，資料無法分辨——不確定性大。費雪資訊量化這份尖銳度，Cramér–Rao 下界把它兌換成硬通貨：**再聰明的（不偏）估計方法，變異數也不可能小於** $1/(nI(\theta))$——精度的極限不取決於方法的聰明，而取決於資料攜帶多少資訊。下界以 $1/n$ 縮小，開根號後正是熟悉的 $1/\sqrt{n}$ 標準誤節奏。

**例 7.9　白努利的費雪資訊與 Cramér–Rao 下界**

設 $X\sim\text{Bernoulli}(p)$，求 $I(p)$，並檢驗 $\hat p = K/n$ 是否有效。

解　單筆觀測的對數機率為 $\ln p(x\mid p) = x\ln p + (1-x)\ln(1-p)$，$x\in\{0,1\}$。一階、二階導數：

$$\frac{\partial}{\partial p}\ln p(x\mid p) = \frac{x}{p} - \frac{1-x}{1-p}, \qquad \frac{\partial^2}{\partial p^2}\ln p(x\mid p) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}.$$

取期望值（$\mathbb{E}[X]=p$、$\mathbb{E}[1-X]=1-p$）：

$$I(p)=-\mathbb{E}\left[\frac{\partial^2}{\partial p^2}\ln p(X\mid p)\right]=\frac{p}{p^2}+\frac{1-p}{(1-p)^2}=\frac{1}{p}+\frac{1}{1-p}=\frac{1}{p(1-p)}.$$

Cramér–Rao 下界因此是 $\operatorname{Var}(\hat p)\ge \frac{1}{nI(p)}=\frac{p(1-p)}{n}$。而例 7.2 已算過 $\hat p=K/n$ 不偏且 $\operatorname{Var}(\hat p)=p(1-p)/n$——**恰好貼在下界上**！觀察頻率不只是常識，它是有效估計量：不偏的世界裡沒有方法比它更準（有限樣本就達到下界是指數族的特權，一般情形 MLE 只保證**漸近**達到）。順帶讀出漸近常態的實用形式：$\hat p\approx\mathcal{N}\big(p,\ \frac{p(1-p)}{n}\big)$，第 8 章的信賴區間將直接建在這條近似上。

## 7.5 從最大概似到機器學習的損失函數

現在收割本章最重要的果實。把 MLE 換一件外衣：最大化 $\ell(\theta)$ 等價於**最小化**它的負值

$$\mathrm{NLL}(\theta)=-\sum_{i=1}^{n}\ln p(x_i\mid\theta),$$

稱為**負對數概似**（negative log-likelihood, NLL）。「最小化某個東西」正是機器學習的母語——NLL 就是損失函數。關鍵在於：**一旦選定機率模型，損失函數就被自動決定**，不需要另外發明。

先看迴歸。假設目標值帶高斯雜訊：$y_i=f(x_i;w)+\varepsilon_i$，$\varepsilon_i\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,\sigma^2)$，其中 $f$ 是任何以 $w$ 為參數的預測函數（線性模型或神經網路都行）。NLL 為

$$\mathrm{NLL}(w)=\frac{n}{2}\ln(2\pi\sigma^2)+\frac{1}{2\sigma^2}\sum_{i=1}^{n}\big(y_i-f(x_i;w)\big)^2.$$

第一項是常數，係數 $\frac{1}{2\sigma^2}$ 不影響最小值位置，所以**對 $w$ 最小化 NLL 等價於最小化平方誤差和**——高斯雜訊假設自動生出**均方誤差**（MSE）損失與最小平方法（完整推導見練習 7.9）。分類的情形則生出交叉熵，這值得一個完整例題。

**例 7.10　從 MLE 到交叉熵**

設類別標籤 $y_1,\dots,y_n\in\{1,\dots,K\}$ i.i.d. 來自類別分布（categorical distribution），參數為 $q=(q_1,\dots,q_K)$、$\sum_k q_k=1$。(a) 證明 NLL 就是交叉熵；(b) 求 MLE。

解　(a) 記 $n_k$ 為類別 $k$ 出現的次數，$\hat p_k=n_k/n$ 為**經驗分布**。NLL 為

$$\mathrm{NLL}(q)=-\sum_{i=1}^{n}\ln q_{y_i}=-\sum_{k=1}^{K}n_k\ln q_k=n\cdot\Big(-\sum_{k=1}^{K}\hat p_k\ln q_k\Big)=n\cdot H(\hat p,q),$$

其中 $H(\hat p,q)=-\sum_k\hat p_k\ln q_k$ 正是經驗分布與模型分布之間的**交叉熵**（cross-entropy）。最小化 NLL 就是最小化交叉熵，一字不差。

(b) 在限制 $\sum_k q_k=1$ 下最大化 $\sum_k n_k\ln q_k$，用拉格朗日乘數法：

$$\frac{\partial}{\partial q_k}\Big(\sum_j n_j\ln q_j+\lambda\big(1-\sum_j q_j\big)\Big)=\frac{n_k}{q_k}-\lambda=0\quad\Longrightarrow\quad q_k=\frac{n_k}{\lambda}.$$

由 $\sum_k q_k = 1$ 得 $\lambda = \sum_k n_k = n$，故 $\hat{q}_k = n_k/n = \hat{p}_k$：MLE 就是**經驗頻率**——白努利結論（例 7.4）的多類別推廣。更深一層：由恆等式 $H(\hat{p}, q) = H(\hat{p}) + \mathrm{KL}(\hat{p} \,\|\, q)$ 且 $\mathrm{KL} \geq 0$，最小化交叉熵等價於把模型分布往經驗分布**推到 KL 距離為零**。訓練分類器時 $q$ 改由網路輸出 $q_w(y \mid x_i)$ 給出，但目標不變：$\mathrm{NLL}(w) = -\sum_i \ln q_w(y_i \mid x_i)$ 正是深度學習框架裡的 cross-entropy loss。

表 7.1: 選了機率模型，就選了損失函數：NLL 的對照表。

| 觀測資料型態 | 機率模型 | NLL 對應的損失 | 典型應用 |
|---|---|---|---|
| 類別標籤 | 白努利 / 類別分布 | 交叉熵 | 分類器、語言模型 |
| 連續目標 | 高斯 | 均方誤差（MSE） | 迴歸 |
| 連續目標（重尾） | 拉普拉斯 | 平均絕對誤差（MAE） | 穩健迴歸 |
| 非負計數 | 卜瓦松 | 卜瓦松迴歸損失 | 客流、理賠次數 |

計數資料同樣一行走完：若 $y_i \sim \mathrm{Poisson}(\lambda(x_i))$，則 $\mathrm{NLL} = \sum_i \big(\lambda(x_i) - y_i \ln \lambda(x_i)\big) + \text{常數}$，即**卜瓦松迴歸**的損失。表 7.1 整理了這組對應——換一個雜訊模型，就換一個損失，沒有一項是「拍腦袋設計」的。

**這如何驅動機器學習／ AI**

**最大概似是損失函數的來源。**最小化 NLL 等於最大化概似，而機率模型的選擇決定損失長什麼樣：類別標籤配類別分布，NLL 就是**交叉熵**——每一個影像分類器、每一個語言模型（預測下一個詞元的類別分布）都在用它；高斯目標配迴歸，NLL 就是**均方誤差**；卜瓦松計數給出卜瓦松迴歸損失。所以**用交叉熵訓練神經網路，就是在對網路的權重做最大概似估計**——SGD 每一步下降的那座山，就是 NLL 這座山。本章的另一半故事也跟了過來：例 7.2 與均勻分布的教訓告訴我們，純 MLE 在小資料上會過度擬合（對沒見過的事件斷言零機率、對雜訊照單全收），而 $\mathrm{MSE} = \mathrm{Bias}^2 + \mathrm{Var}$ 解釋了為什麼**每個實際系統都會加上正則化**：weight decay、label smoothing、提早停止，全是「刻意引入偏差以砍掉變異數」的配重，而第 9 章會證明它們等價於在 MLE 上疊加**先驗**。MLE 的骨架，撐起了幾乎所有監督式深度學習。

## 7.6 Python 實作

理論之後，動手驗證。下面的程式碼依序示範：用數值最佳化最小化硬幣的 NLL（答案應為 $k/n$）；常態 MLE 的解析解；用模擬讓 Bessel 校正的偏差「現形」；以及均勻分布 MLE 的偏差與修正。

```
import numpy as np
from scipy.optimize import minimize_scalar

```

```
rng = np.random.default_rng(42)

# --- 1. 硬幣 NLL 的數值最佳化：10 擲 7 正 ---
n, k = 10, 7
def nll_coin(p):
    return -(k * np.log(p) + (n - k) * np.log(1 - p))

res = minimize_scalar(nll_coin, bounds=(1e-6, 1 - 1e-6),
                      method="bounded")
print(res.x)                          # -> 約 0.7000 = k/n

# --- 2. 常態 MLE：解析解＝樣本均值與（除以 n 的）變異數 ---
x = rng.normal(loc=5.0, scale=2.0, size=1_000)
mu_hat = x.mean()
var_mle = ((x - mu_hat) ** 2).mean()  # 除以 n：MLE，略偏低
print(mu_hat, var_mle, x.var(ddof=1))  # 對照除以 n-1 的不偏版

# --- 3. 讓偏差現形：n = 5 的小樣本重抽 20 萬次 ---
m, n_small = 200_000, 5
xs = rng.normal(0.0, 2.0, size=(m, n_small))  # 真變異數 = 4
print(xs.var(axis=1, ddof=0).mean())  # 約 3.2 = (n-1)/n * 4
print(xs.var(axis=1, ddof=1).mean())  # 約 4.0：Bessel 校正後不偏

# --- 4. 均勻 U(0, theta) 的 MLE 偏低：E[max] = n*theta/(n+1) ---
theta, n_u = 10.0, 5
u = rng.uniform(0, theta, size=(200_000, n_u))
mx = u.max(axis=1)
print(mx.mean())                      # 約 8.33 = 5*10/6
print(((n_u + 1) / n_u * mx).mean())  # 約 10.0：不偏修正
```

輸出重點：第 1 段的數值解與解析解 $k/n = 0.7$ 在小數點後多位一致——同一座山，爬法不同，山頂相同。第 3 段是例 7.1 的實驗版：$n = 5$ 時除以 $n$ 的估計量平均只有 $3.2 = \frac{n-1}{n} \times 4$，Bessel 校正後回到 4.0。第 4 段印證例 7.8：$\mathbb{E}[X_{(5)}] = \frac{5}{6} \times 10 \approx 8.33$，乘上 $(n+1)/n = 1.2$ 後不偏。

**本章總結**

- **估計量**是資料的函數，因此是隨機變數、有**抽樣分布**；評價它的三把尺是**偏差**（系統性偏移）、**變異數**（跨樣本抖動）與**均方誤差**，且恆有 $\text{MSE} = \text{Bias}^2 + \text{Var}$。
- **不偏不是一切**：接受一點偏差換掉大量變異數常使總誤差下降（拉普拉斯平滑），這是正則化的原型；反向的教訓是樣本變異數需要 **Bessel 校正**（除以 $n-1$）才不偏，而 MLE 給的是除以 $n$ 的偏低版本。

- **概似** $L(\theta) = p(\mathcal{D} \mid \theta)$ 是「資料的機率讀成 $\theta$ 的函數」——它**不是** $\theta$ 的機率分布，對 $\theta$ 積分不必為 1；想談「$\theta$ 的機率」需要先驗與貝氏定理。
- **MLE** 最大化（對數）概似：取對數把乘積變和、數值穩定、最大值位置不變。經典結果：白努利 $\hat{p} = k/n$、卜瓦松 $\hat{\lambda} = \bar{x}$、指數 $\hat{\lambda} = 1/\bar{x}$、常態 $(\bar{x},\ \frac{1}{n}\sum(x_i - \bar{x})^2)$、均勻 $\hat{\theta} = \max_i x_i$（最大值在邊界，微積分失效）。
- 正則條件下 MLE **一致**、**漸近常態**、**漸近有效**；**費雪資訊**度量概似山峰的尖銳度，**Cramér–Rao 下界** $\mathrm{Var} \geq 1/(nI(\theta))$ 給出不偏估計量的精度極限。
- 最小化 **NLL** 就是 MLE，且機率模型決定損失函數：類別分布 → **交叉熵**、高斯 → **MSE**、卜瓦松 → 卜瓦松迴歸損失。用交叉熵訓練神經網路 = 對權重做 MLE。
- 純 MLE 在小資料上會**過度擬合**（零機率斷言、貼合雜訊）；偏差–變異數分解預告了**正則化**（= 先驗，第 9 章；權衡全貌見第 10 章）的必要性。

## 7.7 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，「挑戰 · 程式」需要動手寫程式。

**練習 7.1** (基礎)　概念辨析：判斷下列敘述的對錯並簡述理由。(a) 概似函數 $L(\theta)$ 是 $\theta$ 的機率密度，對 $\theta$ 積分必為 1。(b) 固定 $\theta$ 時，$p(\mathcal{D} \mid \theta)$ 對所有可能的資料 $\mathcal{D}$ 加總（或積分）為 1。(c) $L(0.7) > L(0.5)$ 表示參數 0.7 比 0.5 更能解釋觀察到的資料。(d) $L(0.7) > L(0.5)$ 表示「$\theta = 0.7$ 為真」的機率比「$\theta = 0.5$ 為真」高。(e) 想把概似轉成「$\theta$ 的機率」，需要引入先驗並使用貝氏定理。

**練習 7.2** (基礎)　擲一枚硬幣 12 次，得到 9 個正面。(a) 寫出概似函數 $L(p)$ 與對數概似 $\ell(p)$。(b) 求 $p$ 的最大概似估計。(c) 計算概似比 $L(0.75)/L(0.5)$，並說明這個數字的直觀意義。

**練習 7.3** (基礎)　某網站記錄每分鐘的進站人數，10 分鐘的資料為 $2, 1, 0, 3, 2, 4, 1, 2, 0, 3$，假設人數 $\overset{\text{i.i.d.}}{\sim}$ Poisson($\lambda$)。(a) 從對數概似出發，推導 $\hat{\lambda}_{\text{MLE}} = \bar{x}$。(b) 代入資料求 $\hat{\lambda}$。(c) 用插入式預測估計「某一分鐘無人進站」與「某一分鐘至少 3 人進站」的機率。(d) 計算樣本變異數並與 $\hat{\lambda}$ 比較，說明這個比較為什麼是卜瓦松模型的「體檢」。

**練習 7.4** (基礎)　客服中心記錄相鄰兩通來電的間隔時間（分鐘）：2.1, 0.6, 1.4, 3.2, 0.7，假設間隔 $\overset{\text{i.i.d.}}{\sim}$ Exp($\lambda$)。(a) 推導並計算 $\hat{\lambda}_{\text{MLE}}$。(b) 利用 MLE 的不變性，寫出「平均間隔時間」的 MLE。(c) 估計「下一通電話要等超過 2 分鐘」的機率。

**練習 7.5** (基礎)　設 $X_1, \dots, X_n$ i.i.d.，期望值 $\mu$、變異數 $\sigma^2$。(a) 證明樣本均值 $\bar{X}$ 是 $\mu$ 的不偏估計量，並指出證明中是否用到獨立性。(b) 證明 $\operatorname{Var}(\bar{X}) = \sigma^2/n$，並指出這一步何處用到獨立性。(c) 利用柴比雪夫不等式（第 6 章）證明 $\bar{X}$ 是一致估計量，並解釋標準誤的 $1/\sqrt{n}$ 收斂速度在實務上的含意。

**練習 7.6** (進階)　幾何分布 $\mathbb{P}(X = x) = (1-p)^{x-1}p$（$x = 1, 2, \dots$）描述「第幾次試驗才首次成功」。(a) 對 i.i.d. 樣本 $x_1, \dots, x_n$ 推導 $p$ 的 MLE。(b) 某手機遊戲玩家記錄五次「抽到稀有角色所需的抽數」：$4, 1, 7, 2, 6$，求 $\hat{p}$。(c) 用插入式預測估計「超過 5 抽仍未中」的機率。

**練習 7.7** (進階)　(a) 從定義 $\operatorname{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$ 出發，完整證明 $\operatorname{MSE} = \operatorname{Bias}^2 + \operatorname{Var}$，並指出交叉項為何消失。(b) 設 $X_1, \dots, X_n$ i.i.d.、期望值 $\mu$、變異數 $\sigma^2$，考慮收縮估計量 $\hat{\mu}_a = a\bar{X}$（$0 \le a \le 1$）。求 $\operatorname{MSE}(\hat{\mu}_a)$，證明使 MSE 最小的 $a^* = \dfrac{\mu^2}{\mu^2 + \sigma^2/n}$ 恆小於 1，並討論為什麼這個最佳收縮量在實務上「看得到吃不到」。

**練習 7.8** (進階)　設 $X_1, \dots, X_n$ i.i.d.、期望值 $\mu$、變異數 $\sigma^2$，考慮加權平均 $\hat{\mu}_w = \sum_{i=1}^{n} w_i X_i$，其中 $\sum_i w_i = 1$。(a) 證明 $\hat{\mu}_w$ 對任何這樣的權重都不偏。(b) 求 $\operatorname{Var}(\hat{\mu}_w)$。(c) 證明在 $\sum_i w_i = 1$ 的限制下，等權重 $w_i = 1/n$（即 $\bar{X}$）使變異數最小。(d) 若 $n = 3$、權重取 $(0.5, 0.3, 0.2)$，變異數比 $\bar{X}$ 大多少（百分比）？

**練習 7.9** (進階)　迴歸模型 $y_i = f(x_i; w) + \varepsilon_i$，$\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$。(a) 寫出 NLL，並證明對 $w$ 最小化 NLL 等價於最小化 $\sum_i (y_i - f(x_i; w))^2$（最小平方法）。(b) 把 $\sigma^2$ 也當成參數做 MLE，證明 $\hat{\sigma}^2 = \frac{1}{n}\sum_i (y_i - f(x_i; \hat{w}))^2$，即訓練殘差的均方。(c) 若雜訊改為拉普拉斯分布 $p(\varepsilon) = \frac{1}{2b} e^{-|\varepsilon|/b}$，證明 MLE 等價於最小化平均絕對誤差（MAE），並說明這對離群值的意義。

**練習 7.10** (挑戰)　延續例 7.8：$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} U(0, \theta)$，$\hat{\theta} = X_{(n)} = \max_i X_i$。(a) 推導 $X_{(n)}$ 的分布函數與密度函數。(b) 計算 $\mathbb{E}[X_{(n)}]$ 與偏差。(c) 證明 $\tilde{\theta} = \frac{n+1}{n} X_{(n)}$ 不偏。(d) 計算兩者的 MSE 並比較：$\operatorname{MSE}(\hat{\theta})$ 與 $\operatorname{MSE}(\tilde{\theta})$ 誰小？(e) 對照例 7.2（加偏差降 MSE）與本題（去偏差降 MSE），說明決定勝負的是什麼。

**練習 7.11** (挑戰)　(a) 計算卜瓦松分布 $\operatorname{Poisson}(\lambda)$ 的費雪資訊 $I(\lambda)$，分別用「分數平方的期望值」與「二階導數的負期望值」兩種定義各算一次，確認相等。(b) 寫出 $n$ 筆觀測下不偏估計量的 Cramér–Rao 下界。(c) 驗證 $\hat{\lambda} = \bar{X}$ 達到下界，因此是有效估計量。(d) 寫出 $\hat{\lambda}$ 的漸近分布，並說明它如何支撐「$\hat{\lambda} \pm 2\sqrt{\hat{\lambda}/n}$」這類粗略誤差報告。

**練習 7.12** (挑戰 · 程式)　用 Python 完成：(a) 以 `scipy.optimize.minimize_scalar` 數值最小化「10 擲 7 正」的 NLL，確認答案與 $k/n$ 一致。(b) 繪製概似曲線 $L(p) = p^7(1-p)^3$ 並標出最大值位置。(c) 產生 500 筆 $\mathcal{N}(3, 1.5^2)$ 樣本，用 `scipy.optimize.minimize` 同時對 $(\mu, \ln\sigma)$ 最小化高斯 NLL，對照解析解 $\bar{x}$ 與 $\frac{1}{n}\sum(x_i-\bar{x})^2$。提示：以 $\ln\sigma$ 為變數可把「$\sigma>0$」的限制式化於無形。

## 7.8 練習題詳解

**練習 7.1 詳解**　(a) **錯**。概似只是把 $p(\mathcal{D}\mid\theta)$ 讀成 $\theta$ 的函數，沒有任何機制保證對 $\theta$ 積分為 1。反例：擲一次硬幣得正面，$L(p)=p$，$\int_0^1 p\,\mathrm{d}p=\frac{1}{2}$。
(b) **對**。固定 $\theta$ 時 $p(\cdot\mid\theta)$ 是**資料空間**上的機率分布，總和（積分）為 1——機率在資料上。
(c) **對**。這正是概似的定義性讀法：在 $p=0.7$ 的模型下，觀察到這份資料的機率較高，解釋力較強。
(d) **錯**。「$\theta$ 為真的機率」需要把 $\theta$ 視為隨機變數。概似高不等於後驗機率高——若先驗認為 $\theta=0.7$ 極不可能，後驗仍可能偏向 0.5。把 $P(\text{資料}\mid\text{假設})$ 誤讀成 $P(\text{假設}\mid\text{資料})$，正是第 2 章醫療檢驗例子裡的經典錯誤。
(e) **對**。貝氏定理給出 $p(\theta\mid\mathcal{D})\propto L(\theta)\,p(\theta)$：概似乘上先驗、再正規化，才成為 $\theta$ 上的合法機率分布（第 9 章的主題）。

**練習 7.2 詳解**　(a) 每個正面貢獻因子 $p$、反面貢獻 $(1-p)$：$L(p)=p^9(1-p)^3$，$\ell(p)=9\ln p+3\ln(1-p)$。（若把 $L$ 定義為二項機率會多一個係數 $\binom{12}{9}$，它不依賴 $p$，不影響任何結論。）
(b) 令導數為零：

$$\ell'(p)=\frac{9}{p}-\frac{3}{1-p}=0 \quad\Longrightarrow\quad 9(1-p)=3p \quad\Longrightarrow\quad \hat{p}=\frac{9}{12}=0.75,$$

且 $\ell''(p)<0$ 保證是最大值。與公式 $k/n$ 一致。
(c) 直接計算：$L(0.75)=0.75^9\times0.25^3\approx1.173\times10^{-3}$，$L(0.5)=0.5^{12}\approx2.441\times10^{-4}$，兩者之比約為 **4.8**。意義：這份資料由 $p=0.75$ 的硬幣產生的機率，是由公正硬幣產生的約 4.8 倍——資料對前者的支持強了近五倍，但並非壓倒性。這種**概似比**（likelihood ratio）正是第 8 章假設檢定與模型比較的核心統計量。

**練習 7.3 詳解**　(a) 對數概似（丟掉與 $\lambda$ 無關的 $-\sum_i\ln x_i!$）：

$$\ell(\lambda)=-n\lambda+\Big(\sum_i x_i\Big)\ln\lambda,\qquad \ell'(\lambda)=-n+\frac{\sum_i x_i}{\lambda}=0\;\Longrightarrow\;\hat{\lambda}=\bar{x},$$

且 $\ell''(\lambda)=-\sum_i x_i/\lambda^2<0$。
(b) $\sum_i x_i=2+1+0+3+2+4+1+2+0+3=18$，$n=10$，故 $\hat{\lambda}=1.8$（人 / 分鐘）。

(c) 插入 $\hat\lambda = 1.8$：

$$\hat{\mathbb{P}}(X=0) = e^{-1.8} \approx 0.165, \qquad \hat{\mathbb{P}}(X \ge 3) = 1 - e^{-1.8}\Big(1 + 1.8 + \tfrac{1.8^2}{2}\Big) \approx 1 - 0.731 = 0.269.$$

(d) 樣本均值為 1.8，偏差平方和為 15.6，故 $S^2 = 15.6/9 \approx 1.73$（除以 $n$ 則為 1.56）。卜瓦松分布的指紋是**期望值等於變異數**，1.73 與 1.8 相當接近，模型通過體檢。若樣本變異數遠大於樣本均值（過度離散，overdispersion），就該懷疑卜瓦松假設本身——例如來客其實成群抵達。

**練習 7.4 詳解**　(a) 由例 7.6 的推導，$\ell(\lambda) = n\ln\lambda - \lambda\sum_i x_i$，令導數為零得

$$\hat\lambda = \frac{n}{\sum_i x_i} = \frac{5}{2.1 + 0.6 + 1.4 + 3.2 + 0.7} = \frac{5}{8.0} = 0.625 \,(\text{通 / 分鐘}).$$

(b) 平均間隔為 $\mathbb{E}[X] = 1/\lambda$，由 MLE 的不變性，其 MLE 為 $1/\hat\lambda = \bar x = 1.6$ 分鐘——不必重新最佳化。
(c) 指數分布的存活函數為 $\mathbb{P}(X > t) = e^{-\lambda t}$，插入估計值：$\hat{\mathbb{P}}(X > 2) = e^{-0.625\times 2} = e^{-1.25} \approx 0.287$。大約每三通來電中，有一通會讓客服等超過兩分鐘。

**練習 7.5 詳解**　(a) 期望值的線性（**不需要**獨立性）：$\mathbb{E}[\bar X] = \frac{1}{n}\sum_i \mathbb{E}[X_i] = \frac{n\mu}{n} = \mu$，故 $\bar X$ 不偏。
(b) 這一步**需要**獨立性（其實只需兩兩不相關）：獨立時變異數對和可加，

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

若觀測正相關，交叉的共變異數項為正，真實變異數會比 $\sigma^2/n$ 大——重複測量同一個對象不等於多了獨立樣本。
(c) 柴比雪夫不等式給出，對任意 $\epsilon > 0$：$\mathbb{P}\big(|\bar X - \mu| \ge \epsilon\big) \le \frac{\mathrm{Var}(\bar X)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$（$n \to \infty$），即 $\bar X \xrightarrow{P} \mu$：一致性成立（這正是弱大數法則）。標準誤 $\sigma/\sqrt{n}$ 的含意：**樣本數要乘 100，精度才乘 10**——收集資料的邊際報酬遞減，是所有實驗設計都要面對的成本曲線。

**練習 7.6 詳解**　(a) 概似與對數概似：

$$L(p) = \prod_{i=1}^{n}(1-p)^{x_i-1}p = p^n(1-p)^{\sum_i x_i - n}, \qquad \ell(p) = n\ln p + \Big(\sum_i x_i - n\Big)\ln(1-p).$$

令導數為零：

$$\frac{n}{p} - \frac{\sum_i x_i - n}{1-p} = 0 \implies n(1-p) = p\Big(\sum_i x_i - n\Big) \implies n = p\sum_i x_i \implies \hat p = \frac{1}{\bar x}.$$

與直覺吻合：幾何分布的期望值是 $1/p$，MLE 把樣本均值倒過來。
(b) $\sum_i x_i = 4 + 1 + 7 + 2 + 6 = 20$，$\bar x = 4$，故 $\hat p = 0.25$。

(c)「超過 5 抽仍未中」表示前 5 抽全失敗：$\hat{\mathbb{P}}(X>5)=(1-\hat{p})^5=0.75^5\approx 0.237$——將近四分之一的玩家會經歷這種「非洲時刻」，即使機率設定完全誠實。

**練習 7.7 詳解**　(a) 記 $m=\mathbb{E}[\hat\theta]$（常數；$\theta$ 在頻率觀點下也是常數）。加減 $m$：

$$\mathbb{E}\big[(\hat\theta-\theta)^2\big]=\mathbb{E}\big[(\hat\theta-m)^2\big]+2(m-\theta)\,\mathbb{E}[\hat\theta-m]+(m-\theta)^2.$$

交叉項中 $\mathbb{E}[\hat\theta-m]=0$（隨機變數減自身期望值），故消失；剩下 $\mathrm{Var}(\hat\theta)+\mathrm{Bias}(\hat\theta)^2$。■
(b) $\mathbb{E}[\hat\mu_a]=a\mu$，故 $\mathrm{Bias}=(a-1)\mu$；$\mathrm{Var}(\hat\mu_a)=a^2\sigma^2/n$。由 (a)，對 $a$ 微分令為零（$\frac{2a\sigma^2}{n}+2(a-1)\mu^2=0$）：

$$\mathrm{MSE}(a)=\frac{a^2\sigma^2}{n}+(a-1)^2\mu^2\quad\Longrightarrow\quad a^*=\frac{\mu^2}{\mu^2+\sigma^2/n}<1\quad(\text{只要}\sigma^2>0),$$

且二階導數 $2\sigma^2/n+2\mu^2>0$ 確為最小。代回可得 $\mathrm{MSE}(a^*)=\dfrac{(\sigma^2/n)\,\mu^2}{\mu^2+\sigma^2/n}<\dfrac{\sigma^2}{n}=\mathrm{MSE}(\bar X)$：**略往 0 收縮永遠比純 $\bar X$ 好一點**。例如 $\mu=1$、$\sigma^2=1$、$n=10$：$a^*\approx 0.909$，MSE 從 0.100 降到 0.091。「看得到吃不到」：$a^*$ 依賴**未知的** $\mu$，不知道真值就算不出最佳收縮量。實務上把收縮強度當超參數、用驗證資料挑選（第 10 章），或由先驗誘導（第 9 章）——這正是正則化係數需要調參的根本原因。

**練習 7.8 詳解**　(a) $\mathbb{E}[\hat\mu_w]=\sum_i w_i\mathbb{E}[X_i]=\mu\sum_i w_i=\mu$：只要權重和為 1 就不偏。
(b) 由獨立性：$\mathrm{Var}(\hat\mu_w)=\sum_i w_i^2\,\mathrm{Var}(X_i)=\sigma^2\sum_i w_i^2$。
(c) 由柯西–施瓦茲不等式（或拉格朗日乘數法）：

$$1=\Big(\sum_{i=1}^n w_i\cdot 1\Big)^2\le n\sum_{i=1}^n w_i^2\quad\Longrightarrow\quad\sum_{i=1}^n w_i^2\ge\frac{1}{n},$$

等號成立若且唯若所有 $w_i$ 相等，即 $w_i=1/n$。因此 $\bar X$ 是所有不偏線性組合中變異數最小的——「最佳線性不偏估計量」（BLUE）思想的雛形，也是高斯–馬可夫定理的前奏。
(d) $\sum_i w_i^2=0.25+0.09+0.04=0.38$，而等權重為 $1/3\approx 0.333$，比值 $0.38/(1/3)=1.14$：變異數比 $\bar X$ 大 14%。權重越不均，浪費越多——除非某些觀測本身品質更高（變異數更小），等權重才會被打敗。

**練習 7.9 詳解**　(a) 每筆觀測的密度為 $\frac{1}{\sqrt{2\pi\sigma^2}}\exp\!\big(-\frac{(y_i-f(x_i;w))^2}{2\sigma^2}\big)$，故

$$\mathrm{NLL}(w,\sigma^2)=\frac{n}{2}\ln(2\pi\sigma^2)+\frac{1}{2\sigma^2}\sum_{i=1}^n\big(y_i-f(x_i;w)\big)^2.$$

固定 $\sigma^2$ 時，第一項是常數、第二項的係數 $\frac{1}{2\sigma^2}$ 為正常數，因此 $\arg\min_w\mathrm{NLL}=\arg\min_w\sum_i\big(y_i-f(x_i;w)\big)^2$：**最小化 NLL = 最小平方法 = MSE 損失。假設雜訊是高斯的，就自動得到均方誤差。**

(b) 記 $\mathrm{SSE}=\sum_i (y_i-f(x_i;\hat w))^2$，對 $v=\sigma^2$ 微分：

$$\frac{\partial \,\mathrm{NLL}}{\partial v}=\frac{n}{2v}-\frac{\mathrm{SSE}}{2v^2}=0 \quad\Longrightarrow\quad \hat\sigma^2=\frac{\mathrm{SSE}}{n},$$

即**訓練殘差的均方**——模型對自身誤差大小的 MLE，正是訓練集上的 MSE 值本身（與例 7.7 的結構完全平行，同樣略偏低）。

(c) 拉普拉斯雜訊下：

$$\mathrm{NLL}(w)=n\ln(2b)+\frac{1}{b}\sum_{i=1}^{n}\bigl|y_i-f(x_i;w)\bigr|,$$

對 $w$ 最小化等價於最小化 $\sum_i |y_i-f(x_i;w)|$，即 **MAE（L1）損失**。拉普拉斯尾巴比高斯厚，把「偶爾的巨大誤差」視為正常，不像平方損失那樣被離群值挾持——L1 迴歸的穩健性根源於雜訊模型。**選噪音模型就是選損失函數。**

**練習 7.10 詳解**　(a) 由獨立性，$\mathbb{P}(X_{(n)}\le t)=\mathbb{P}(\text{所有}X_i\le t)=\prod_i \mathbb{P}(X_i\le t)=(t/\theta)^n$ $(0\le t\le\theta)$。微分得密度

$$f_{X_{(n)}}(t)=\frac{n\,t^{n-1}}{\theta^n},\qquad 0\le t\le\theta.$$

(b) $\mathbb{E}[X_{(n)}]=\int_0^\theta t\cdot\frac{nt^{n-1}}{\theta^n}\,\mathrm{d}t=\frac{n}{\theta^n}\cdot\frac{\theta^{n+1}}{n+1}=\frac{n}{n+1}\theta$，故 $\mathrm{Bias}=-\frac{\theta}{n+1}$。

(c) $\mathbb{E}[\tilde\theta]=\frac{n+1}{n}\mathbb{E}[X_{(n)}]=\frac{n+1}{n}\cdot\frac{n}{n+1}\theta=\theta$：不偏。■

(d) 先算二階動差 $\mathbb{E}[X_{(n)}^2]=\int_0^\theta t^2\frac{nt^{n-1}}{\theta^n}\,\mathrm{d}t=\frac{n}{n+2}\theta^2$，於是 $\mathrm{Var}(X_{(n)})=\frac{n}{n+2}\theta^2-\bigl(\frac{n}{n+1}\bigr)^2\theta^2=\frac{n\,\theta^2}{(n+1)^2(n+2)}$。MLE 的 MSE（偏差平方加變異數）與修正版的 MSE（不偏，故等於自身變異數 $(\frac{n+1}{n})^2\,\mathrm{Var}(X_{(n)})$）分別為

$$\mathrm{MSE}(\hat\theta)=\frac{\theta^2}{(n+1)^2}+\frac{n\theta^2}{(n+1)^2(n+2)}=\frac{2\theta^2}{(n+1)(n+2)},\qquad \mathrm{MSE}(\tilde\theta)=\frac{\theta^2}{n(n+2)}.$$

比值 $\mathrm{MSE}(\hat\theta)/\mathrm{MSE}(\tilde\theta)=\frac{2n}{n+1}>1$（$n\ge 2$），所以**修正版全面獲勝**：去掉偏差的同時 MSE 也變小（$n=5$ 時 $\theta^2/21\approx 0.048\theta^2$ 對 $\theta^2/35\approx 0.029\theta^2$）。

(e) 兩題的判準是同一條式子：$\mathrm{MSE}=\mathrm{Bias}^2+\mathrm{Var}$。例 7.2 中加偏差能**大幅壓低變異數**（分母從 $n^2$ 變 $(n+2)^2$），交易划算；本題中 $X_{(n)}$ 的偏差是「純損失」——乘上 $\frac{n+1}{n}$ 放大的變異數增幅，小於偏差平方的降幅。偏差本身無所謂好壞，**值不值得，永遠看 $\mathrm{Bias}^2$ 與 Var 的交換匯率**。

**練習 7.11 詳解**　(a) $\ln p(x\mid\lambda)=-\lambda+x\ln\lambda-\ln x!$。分數 $\frac{\partial}{\partial\lambda}\ln p=\frac{x}{\lambda}-1$，二階導數 $\frac{\partial^2}{\partial\lambda^2}\ln p=-\frac{x}{\lambda^2}$。

定義一（分數平方）：$\mathbb{E}\bigl[(\frac{X}{\lambda}-1)^2\bigr]=\dfrac{\mathbb{E}[(X-\lambda)^2]}{\lambda^2}=\dfrac{\mathrm{Var}(X)}{\lambda^2}=\dfrac{\lambda}{\lambda^2}=\dfrac{1}{\lambda}$。

定義二（負二階導）：$-\mathbb{E}\bigl[-\frac{X}{\lambda^2}\bigr]=\dfrac{\mathbb{E}[X]}{\lambda^2}=\dfrac{1}{\lambda}$。兩者一致：$I(\lambda)=1/\lambda$。

(b) $n$ 筆觀測的 Cramér–Rao 下界：$\mathrm{Var}(\hat\lambda)\ge\dfrac{1}{nI(\lambda)}=\dfrac{\lambda}{n}$。

(c) $\bar{X}$ 不偏（$\mathbb{E}[\bar{X}] = \lambda$），且 $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n = \lambda/n$——恰好等於下界，故 $\bar{X}$ 是有效估計量：不偏世界裡的精度冠軍。
(d) 由漸近常態（即中央極限定理）：$\hat{\lambda} \approx \mathcal{N}\big(\lambda,\ \lambda/n\big)$，標準誤約 $\sqrt{\lambda/n}$，實務上以 $\sqrt{\hat{\lambda}/n}$ 估計。常態約 95% 的質量落在 $\pm 2$ 個標準差內，故「$\hat{\lambda} \pm 2\sqrt{\hat{\lambda}/n}$」是粗略的 95% 誤差帶——第 8 章會把它鍛造成正式的信賴區間。

**練習 7.12 詳解**　參考程式與說明：

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.optimize import minimize_scalar, minimize

# (a) 硬幣 NLL 的數值最佳化
n, k = 10, 7
nll = lambda p: -(k * np.log(p) + (n - k) * np.log(1 - p))
res = minimize_scalar(nll, bounds=(1e-6, 1 - 1e-6), method="bounded")
print(res.x)                        # 0.69999...，與 k/n = 0.7 一致
# (b) 概似曲線與最大值
ps = np.linspace(0.001, 0.999, 400)
plt.plot(ps, ps ** k * (1 - ps) ** (n - k))
plt.axvline(res.x, linestyle="--", color="red")
plt.xlabel("p"); plt.ylabel("L(p)")
plt.savefig("likelihood_curve.png", dpi=120)
# (c) 高斯兩參數 MLE：以 log-sigma 重參數化
x = np.random.default_rng(1).normal(3.0, 1.5, size=500)
def nll_gauss(par):
    mu, log_s = par
    s2 = np.exp(2 * log_s)          # sigma^2 > 0 自動成立
    return 0.5 * np.sum(np.log(2 * np.pi * s2)
                        + (x - mu) ** 2 / s2)

r = minimize(nll_gauss, x0=np.array([0.0, 0.0]))
print(r.x[0], np.exp(2 * r.x[1]))  # 數值最佳化的 MLE
print(x.mean(), x.var(ddof=0))     # 解析 MLE：應幾乎相同
```

說明：(a) 有界一維最佳化直接回到解析解 0.7。(b) 峰頂與虛線重合於 $p = 0.7$，即圖 7.3 的程式版。(c) 關鍵是**重參數化**：以 $\ln\sigma$ 為變數，$\sigma^2 = e^{2\ln\sigma}$ 自動為正，受限問題變成無限制問題。最後兩行輸出幾乎相同：數值與解析 MLE 是同一個東西。

下一章：假設檢定與信賴區間——量化估計的不確定性，讓資料對假說做出裁決。