機率與統計：從入門到精通

第 5 章　聯合分布、共變異數與多元常態分布

Joint Distributions, Covariance, and the Multivariate Gaussian

難度：核心

一張 256×256 的灰階影像是 65,536 個像素組成的向量；一位病患的檢驗報告是數十項指標的組合；一個句子是一串前後呼應的 token。真實世界的資料幾乎從來不是單一數字，而是一整組**互相牽動**的量。要替這樣的資料建模，我們需要三件武器：描述「變數如何一起變動」的**聯合分布**、量化線性共動強度的**共變異數**，以及向量世界的主力分布——**多元常態分布**。本章正是機率論與（姊妹冊）線性代數相遇之處：共變異數矩陣的特徵分解就是主成分分析，多元常態的等密度面就是橢球。

學習目標

讀完本章後，你將能夠：

- 寫出離散與連續的**聯合分布**，以**邊際化**與**條件化**回答機率問題，並判斷獨立性。
- 計算並詮釋**共變異數**與**相關係數**，說清楚「相關不是因果」與「零相關不是獨立」。
- 建構**共變異數矩陣**並證明其對稱半正定；用**特徵分解**找出主軸，連結 PCA 與白化。
- 寫出**多元常態**密度，解讀橢圓輪廓與**馬氏距離**；用**封閉性**與條件分布公式做預測。
- 用 NumPy 抽樣多元常態、估計共變異數矩陣與主軸，以模擬驗證本章公式。

5.1　聯合、邊際與條件分布

到目前為止，我們一次只研究一個隨機變數。但幾乎所有有趣的問題都涉及**多個**變數之間的關係：身高與體重、廣告支出與銷售額、一封郵件的用詞與它是否為垃圾郵件。描述多個隨機變數的工具，就是聯合分布。

定義 5.1　聯合分布

設 X, Y 是同一個隨機試驗中的兩個隨機變數。

- **離散情形**：**聯合機率質量函數**（joint probability mass function）給出每一對取值同時

發生的機率：

$$p(x,y)=\mathbb{P}(X=x,\ Y=y),\qquad \sum_x\sum_y p(x,y)=1.$$

- **連續情形**：**聯合機率密度函數**（joint probability density function）$f(x,y)\ge 0$ 滿足 $\iint f(x,y)\,\mathrm{d}x\,\mathrm{d}y=1$，且對平面上任何區域 A，

$$\mathbb{P}\big((X,Y)\in A\big)=\iint_A f(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

兩者都可以推廣到 n 個變數：$p(x_1,\dots,x_n)$ 或 $f(x_1,\dots,x_n)$。

離散的聯合分布就是一張**表格**，連續的聯合密度則是一座**曲面**（機率是曲面下方的體積）；兩者都是「這組變數同時如何行為」的**完整**描述——邊際、條件、共變異數，一切都能從它算出來。從聯合分布讀出資訊的基本操作恰有兩種：

定義 5.2　邊際分布與條件分布

由聯合分布可以恢復出兩種「衍生」分布：

- **邊際分布**（marginal distribution）：把不關心的變數加總（或積分）掉，

$$p_X(x)=\sum_y p(x,y),\qquad f_X(x)=\int f(x,y)\,\mathrm{d}y.$$

這個操作稱為**邊際化**（marginalization）。

- **條件分布**（conditional distribution）：固定一個變數的值，再重新正規化，

$$p_{Y\mid X}(y\mid x)=\frac{p(x,y)}{p_X(x)},\qquad f_{Y\mid X}(y\mid x)=\frac{f(x,y)}{f_X(x)},$$

（分母須為正）。這個操作稱為**條件化**（conditioning）。

把條件分布的定義移項，就得到第 2 章乘法法則的雙變數版本：$p(x,y)=p_X(x)\,p_{Y\mid X}(y\mid x)$。

定義 5.3　獨立性

隨機變數 X 與 Y **獨立**（independent），若聯合分布在**每一點**都拆成邊際的乘積：

$$p(x,y)=p_X(x)\,p_Y(y)\quad \text{對所有}x,y$$

（連續情形為 $f(x,y)=f_X(x)f_Y(y)$）。等價地，$p_{Y\mid X}(y\mid x)=p_Y(y)$：知道 X 完全不改變對 Y 的看法。只要有一格不滿足乘積關係，兩變數就**相依**（dependent）。

例 5.1　垃圾郵件與「免費」：讀一張 2×3 聯合表

令 $X\in\{0,1\}$ 表示郵件是否為垃圾郵件（$1=$ 垃圾），$Y\in\{0,1,2\}$ 為「免費」一詞出現的次數（超過 2 次以 2 計）。歷史郵件統計出的聯合分布為：

	$y=0$	$y=1$	$y=2$
$x=0$（正常信）	0.42	0.12	0.06
$x=1$（垃圾信）	0.08	0.12	0.20

(a) 求兩個邊際分布。(b) 求垃圾信與正常信中 Y 的條件分布。(c) 看到「免費」出現 2 次，這封信是垃圾信的機率多大？(d) X 與 Y 獨立嗎？

解　**(a)** 六格加總為 1，合法。**橫列**相加得 $p_X=(0.60,0.40)$——四成郵件是垃圾信；**直行**相加得 $p_Y=(0.50,0.24,0.26)$。邊際分布寫在表格「邊緣」，名稱由此而來。

(b) 固定 x 的橫列、除以該列總和：

$$p_{Y|X}(y\mid 1)=\frac{(0.08,0.12,0.20)}{0.40}=(0.20,\ 0.30,\ 0.50),\qquad p_{Y|X}(y\mid 0)=(0.70,\ 0.20,\ 0.10).$$

垃圾信有五成機會出現兩次以上「免費」；正常信只有一成。

(c) 這次固定 $y=2$ 這一直行，除以 $p_Y(2)=0.26$：

$$\mathbb{P}(X=1\mid Y=2)=\frac{0.20}{0.26}=\frac{10}{13}\approx 0.769.$$

沒讀信之前垃圾機率是 0.40，看到兩次「免費」後跳到 0.77——這正是第 2 章的貝氏定理：手上有聯合分布時，貝氏更新只是「換個方向切表格」。

(d) 檢查任何一格：$p(1,2)=0.20\neq p_X(1)\,p_Y(2)=0.40\times 0.26=0.104$，故**相依**——這是好事，正因相依，字詞才對分類有資訊量。

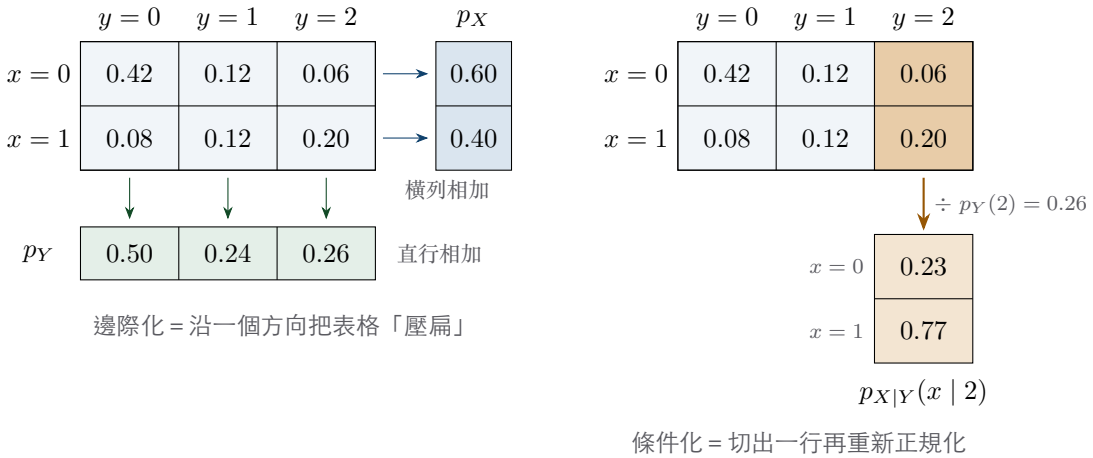


圖 5.1：對聯合分布僅有的兩種基本「查詢」。左：邊際化把不關心的變數加總掉（橫列相加得 p_X，直行相加得 p_Y）。右：條件化固定 $Y=2$ 這一直行，除以 $p_Y(2)=0.26$ 重新正規化，得到 $p_{X|Y}(\cdot\mid 2)=(0.23,0.77)$。

例 5.2　連續情形：三角形上的均勻分布

設 (X, Y) 的聯合密度為

$$f(x, y) = 2, \qquad 0 \le y \le x \le 1$$

（三角形之外為 0；由面積 $\frac{1}{2}$ 可驗證總積分為 1）。求邊際密度、條件密度並判斷獨立性。

解　**邊際化**時要小心積分範圍：固定 x 時 y 只能在 $[0, x]$，固定 y 時 x 只能在 $[y, 1]$，故

$$f_X(x) = \int_0^x 2 \, \mathrm{d}y = 2x, \qquad f_Y(y) = \int_y^1 2 \, \mathrm{d}x = 2(1 - y).$$

聯合密度是「平的」，兩個邊際卻都不均勻：X 偏大、Y 偏小。**條件化**：$f_{Y|X}(y \mid x) = 2/(2x) = 1/x$（$0 \le y \le x$），即 $Y \mid X = x \sim \mathrm{Uniform}(0, x)$，故 $\mathbb{E}[Y \mid X = x] = x/2$——條件期望值是 x 的函數，這正是「迴歸函數」的雛形。**獨立性**：$f(x, y) = 2 \neq f_X(x) f_Y(y) = 4x(1 - y)$，相依；更快的觀察是支撐（support）為三角形而非長方形——Y 的活動範圍隨 X 改變，變數幾乎注定相依。

直觀理解

把聯合分布想成一張表（離散）或一座山丘（連續）。**邊際化**是「壓扁」：沿著不關心的方向把整張表壓成一條，質量全部保留、細節全部犧牲。**條件化**是「切片」：沿著已知資訊切一刀，取出剖面，再放大到總機率為 1（除以邊際機率就是放大倍率）。一壓一切，是對聯合分布僅有的兩種查詢。最後一個提醒：**邊際不足以決定聯合**——「兩枚硬幣完全同步」與「兩枚硬幣獨立」的邊際一模一樣，耦合方式卻天差地遠；變數「如何綁在一起」的資訊只住在聯合分布裡。

核心觀念

每個機率模型終究是一張聯合分布的配方。拆解多變數聯合分布的萬用工具是**連鎖律**（chain rule）：

$$p(x_1, x_2, \ldots, x_n) = p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_1, x_2) \cdots p(x_n \mid x_1, \ldots, x_{n-1}).$$

大型語言模型逐字生成文字，正是沿著這條公式抽樣：每一步從「給定前文的條件分布」取一個 token。而若不做任何假設，n 個二元變數的聯合表有 $2^n - 1$ 個自由參數——$n = 300$ 就超過宇宙原子數。機器學習的核心技藝，正是用獨立性假設、條件分布的參數化（神經網路）或本章的高斯結構，把這張天文數字的表壓縮成學得動的形式。

5.2　共變異數與相關係數

聯合分布是完整的描述，但它是一整張表或一座曲面。實務上我們常想要一個**單一數字**來回答：「X 偏大時，Y 是傾向偏大還是偏小？」這個數字就是共變異數。

定義 5.4　共變異數

隨機變數 X, Y 的**共變異數**（covariance）定義為

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

第二個等號是計算捷徑，把乘積展開、用期望值的線性即得。特別地，$\operatorname{Cov}(X,X) = \operatorname{Var}(X)$：變異數是共變異數的對角特例。

定義的幾何讀法：以均值點 $(\mathbb{E}X, \mathbb{E}Y)$ 為原點劃分四個象限，右上、左下象限的乘積 $(X-\mathbb{E}X)(Y-\mathbb{E}Y)$ 為正，另外兩象限為負；共變異數是這些貢獻的加權平均——**正**表示傾向同漲同跌、**負**表示一漲一跌、**零**表示沒有線性趨勢。

不過共變異數有單位的困擾：把身高從公分改成公尺，Cov 立刻縮小 100 倍，關聯的強度卻沒變。解法是除以兩個標準差，把尺度歸一：

定義 5.5　相關係數

X, Y 的**相關係數**（correlation coefficient）為

$$\operatorname{Corr}(X,Y) = \rho_{XY} = \frac{\operatorname{Cov}(X,Y)}{\sigma_X\,\sigma_Y},$$

其中 σ_X, σ_Y 為標準差。相關係數無單位、對線性尺度變換不變，且恆有

$$-1 \le \rho_{XY} \le 1$$

（證明見練習 5.9）。$\rho = \pm 1$ 若且唯若 Y 是 X 的完美線性函數（斜率正負對應 ρ 的正負）；$\rho = 0$ 稱為**不相關**（uncorrelated）。

例 5.3　手算共變異數與相關係數

用例 5.1 的垃圾郵件聯合表，計算 $\operatorname{Cov}(X,Y)$ 與 $\operatorname{Corr}(X,Y)$。

解　由邊際分布：

$$\mathbb{E}[X] = 0\times 0.60 + 1\times 0.40 = 0.40, \qquad \mathbb{E}[Y] = 0\times 0.50 + 1\times 0.24 + 2\times 0.26 = 0.76.$$

$\mathbb{E}[XY]$ 要用聯合表——只有 $x=1$ 的橫列有貢獻：

$$\mathbb{E}[XY] = 1\cdot 0\cdot 0.08 + 1\cdot 1\cdot 0.12 + 1\cdot 2\cdot 0.20 = 0.52.$$

代入捷徑公式：

$$\operatorname{Cov}(X,Y) = 0.52 - 0.40\times 0.76 = 0.52 - 0.304 = 0.216 > 0,$$

垃圾信與「免費」次數正相關，符合直覺。接著算兩個變異數：

$$\operatorname{Var}(X) = 0.4\times 0.6 = 0.24, \qquad \mathbb{E}[Y^2] = 0\times 0.50 + 1\times 0.24 + 4\times 0.26 = 1.28,$$

$$\mathrm{Var}(Y) = 1.28 - 0.76^2 = 0.7024.$$

因此

$$\mathrm{Corr}(X,Y) = \frac{0.216}{\sqrt{0.24}\,\sqrt{0.7024}} = \frac{0.216}{0.4899 \times 0.8381} \approx 0.526,$$

中等強度的正相關——而且這個數字不依賴 Y 的計數單位。

命題 5.6　共變異數的代數性質

對任何隨機變數與常數 a, b, c, d：

1. **對稱性**：$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$。
2. **雙線性**：$\mathrm{Cov}(aX+b,\ cY+d) = ac\,\mathrm{Cov}(X,Y)$（平移不影響、尺度成倍進出）。
3. **和的變異數**：
$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y).$$
4. **獨立 $\Rightarrow$ 不相關**：若 X, Y 獨立，則 $\mathrm{Cov}(X,Y) = 0$，且 $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$。

證明　(1) 由定義顯然。(2)、(3) 留作練習 5.2（把定義展開即得）。(4) 若獨立，離散情形下

$$\mathbb{E}[XY] = \sum_x \sum_y xy\, p_X(x) p_Y(y) = \Big(\sum_x x\, p_X(x)\Big)\Big(\sum_y y\, p_Y(y)\Big) = \mathbb{E}[X]\,\mathbb{E}[Y],$$

（連續情形把和換成積分），故 $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$；代入 (3) 即得變異數可加。 ■

性質 (4) 的箭頭方向至關重要：獨立保證不相關，但**反過來不成立**。下面是最著名的反例。

例 5.4　零相關但完全相依：$Y = X^2$

設 X 在 $\{-1, 0, 1\}$ 上均勻取值，$Y = X^2$。證明 $\mathrm{Cov}(X,Y) = 0$，但 X, Y 相依。

解　先算需要的期望值：

$$\mathbb{E}[X] = \frac{-1+0+1}{3} = 0, \qquad \mathbb{E}[XY] = \mathbb{E}[X^3] = \frac{(-1)^3 + 0^3 + 1^3}{3} = 0.$$

因此

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0 - 0 \cdot \mathbb{E}[Y] = 0,$$

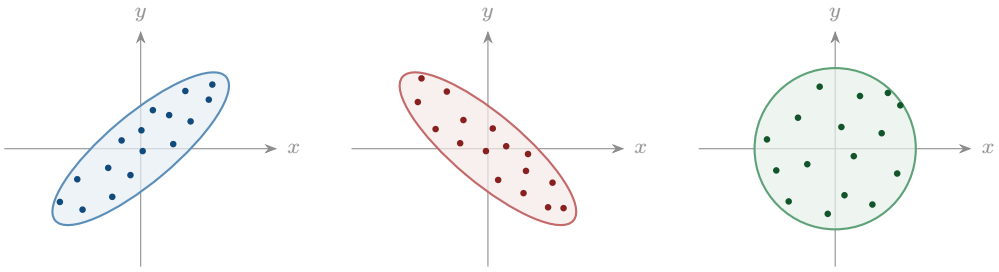
相關係數也是 0。但 Y 完全由 X 決定——這是最強烈的相依！形式驗證：$Y \in \{0,1\}$ 且 $\mathbb{P}(Y=1) = \frac{2}{3}$，而

$$\mathbb{P}(Y = 1 \mid X = 1) = 1 \neq \tfrac{2}{3} = \mathbb{P}(Y=1),$$

知道 X 徹底改變了對 Y 的看法。原因：X 增加時 Y 先降後升，正負貢獻在對稱之下**恰好抵消**——共變異數只偵測線性趨勢，對 U 形關係完全失明。

常見陷阱

兩個必須刻在腦中的警告。**(1) 相關不是因果。**冰淇淋銷量與溺水人數高度正相關，但禁售冰淇淋救不了任何人——共同的原因是夏天。強相關可能來自 X 影響 Y、Y 影響 X，或隱藏的**共同原因**（confounder）推動兩者；分辨因果需要隨機對照試驗或因果推論工具，光看相關永遠不夠。**(2) 零相關不是獨立。**共變異數只量測**線性**關聯；例 5.4 的 $Y = X^2$ 相關為零卻完全相依。正確的蘊涵鏈是：獨立 $\Rightarrow$ 不相關，僅此一個方向。唯一例外在第 5.5 節：若 (X, Y) **聯合**服從多元常態，不相關與獨立等價。



正相關（$\rho \approx 0.9$）　　負相關（$\rho \approx -0.9$）　　零相關（$\rho \approx 0$）

圖 5.2: 資料雲的形狀透露相關係數的正負：正相關沿「右上—左下」拉長、負相關沿「左上—右下」、零相關沒有傾斜的長軸。零相關只排除**線性**趨勢——U 形、環形等非線性關聯照樣可以有 $\rho = 0$。

5.3 共變異數矩陣：通往線性代數的橋

三個以上的變數怎麼辦？n 個變數有 n 個變異數、$\binom{n}{2}$ 個兩兩共變異數，最自然的收納方式是一個矩陣。從這一節開始，機率論正式與線性代數會師。

定義 5.7　隨機向量與共變異數矩陣

把 n 個隨機變數疊成**隨機向量**（random vector）$\boldsymbol{x} = (X_1, X_2, \dots, X_n)^{\top}$。其**均值向量**為 $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}] = (\mathbb{E}X_1, \dots, \mathbb{E}X_n)^{\top}$，**共變異數矩陣**（covariance matrix）為

$$\Sigma = \mathbb{E}\big[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{\top}\big], \qquad \Sigma_{ij} = \operatorname{Cov}(X_i, X_j).$$

對角線元素 $\Sigma_{ii} = \operatorname{Var}(X_i)$ 是各變數的變異數；非對角線元素記錄每一對變數的共動。

命題 5.8　共變異數矩陣的性質

Σ 恆為**對稱**（symmetric）且**半正定**（positive semidefinite）矩陣：

$$\Sigma = \Sigma^{\top}, \qquad \boldsymbol{a}^{\top}\Sigma\,\boldsymbol{a} \geq 0 \quad \text{對所有} \boldsymbol{a} \in \mathbb{R}^n.$$

此外，對任何矩陣 A 與向量 $\boldsymbol{b}$，線性變換 $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{b}$ 的共變異數矩陣為

$$\operatorname{Cov}(A\boldsymbol{x} + \boldsymbol{b}) = A\,\Sigma\,A^{\top}.$$

證明　對稱性來自 $\mathrm{Cov}(X_i,X_j)=\mathrm{Cov}(X_j,X_i)$。半正定性有一個漂亮的一行證明：

$$\boldsymbol{a}^{\top}\Sigma\boldsymbol{a}=\mathbb{E}\big[\boldsymbol{a}^{\top}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{a}\big]=\mathbb{E}\big[\big((\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{a}\big)^{2}\big]=\mathrm{Var}(\boldsymbol{a}^{\top}\boldsymbol{x})\ge 0,$$

換句話說：任何投資組合（任何線性組合）的變異數都不可能是負的。線性變換公式同樣是定義的直接展開：$\boldsymbol{y}-\mathbb{E}\boldsymbol{y}=A(\boldsymbol{x}-\boldsymbol{\mu})$，故 $\mathrm{Cov}(\boldsymbol{y})=\mathbb{E}[A(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\top}A^{\top}]=A\Sigma A^{\top}$。 ■

對稱半正定矩陣正是線性代數中性質最好的一類矩陣。姊妹冊的**譜定理**（spectral theorem）告訴我們：

定理 5.9　特徵分解與主軸

任何共變異數矩陣可分解為

$$\Sigma=Q\Lambda Q^{\top}=\sum_{i=1}^{n}\lambda_i\,\boldsymbol{v}_i\boldsymbol{v}_i^{\top},$$

其中 $Q=[\boldsymbol{v}_1\ \cdots\ \boldsymbol{v}_n]$ 的直行是互相垂直的單位**特徵向量**（eigenvector），$\Lambda=\mathrm{diag}(\lambda_1,\dots,\lambda_n)$ 收納對應的**特徵值**（eigenvalue），半正定保證 $\lambda_i\ge 0$。機率的詮釋：特徵向量是資料雲的**主軸**方向，且 $\mathrm{Var}(\boldsymbol{v}_i^{\top}\boldsymbol{x})=\boldsymbol{v}_i^{\top}\Sigma\boldsymbol{v}_i=\lambda_i$——特徵值就是**沿主軸的變異數**；$\boldsymbol{v}_1$（最大特徵值的方向）是所有方向中變異最大的一個，即**主成分分析**（principal component analysis, PCA）的第一主成分。

把資料投影到前 k 個特徵向量上，就以最小的資訊犧牲把 n 維壓到 k 維——PCA 是最常用的降維方法，而它整件事就是對 Σ 做特徵分解。

例 5.5　2×2 特徵分解：手算 PCA 方向

求

$$\Sigma=\begin{pmatrix}3 & 1.5\\ 1.5 & 1\end{pmatrix}$$

的特徵值、特徵向量，以及第一主成分解釋的變異比例。

解　**特徵值**：解特徵方程式 $\det(\Sigma-\lambda I)=(3-\lambda)(1-\lambda)-1.5^2=\lambda^2-4\lambda+0.75=0$，得

$$\lambda=\frac{4\pm\sqrt{16-3}}{2}=\frac{4\pm\sqrt{13}}{2}\quad\Longrightarrow\quad\lambda_1\approx 3.8028,\qquad\lambda_2\approx 0.1972.$$

驗算：$\lambda_1+\lambda_2=4=\mathrm{tr}\,\Sigma$、$\lambda_1\lambda_2=0.75=\det\Sigma$，正確。

特徵向量：對 λ_1，解 $(\Sigma-\lambda_1 I)\boldsymbol{v}=\boldsymbol{0}$ 的第一列 $(3-3.8028)v_1+1.5v_2=0$，得 $v_2=0.5352\,v_1$，故

$$\boldsymbol{v}_1\propto(1,\ 0.5352)^{\top}\quad\Longrightarrow\quad\hat{\boldsymbol{v}}_1=(0.8817,\ 0.4719)^{\top},$$

方向角 $\theta=\arctan(0.5352)\approx 28.2^{\circ}$。$\boldsymbol{v}_2$ 與之垂直：$\hat{\boldsymbol{v}}_2=(-0.4719,\ 0.8817)^{\top}$（可驗證 $\Sigma\hat{\boldsymbol{v}}_2=\lambda_2\hat{\boldsymbol{v}}_2$）。

變異比例：第一主成分解釋 $\lambda_1/(\lambda_1+\lambda_2)=3.8028/4\approx 95.1\%$ 的總變異。也就是說，把二維資料投影到 $\hat{\boldsymbol{v}}_1$ 這條線上，只丟失約 4.9% 的變異——降維幾乎是免費的。

特徵分解還送我們一個常用的預處理技術：**白化**（whitening）。定義 $\boldsymbol{z}=\Lambda^{-1/2}Q^{\top}(\boldsymbol{x}-\boldsymbol{\mu})$，

即「先旋轉到主軸座標、再把每一軸除以自己的標準差」。由線性變換公式，

$$\mathrm{Cov}(\boldsymbol{z}) = \Lambda^{-1/2} Q^{\top} \Sigma\, Q \Lambda^{-1/2} = \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I\,:$$

白化後的資料各分量變異數為 1、兩兩不相關，像「白雜訊」一樣沒有方向偏好；許多演算法以它作為預處理的第一步。

直觀理解

把 Σ 讀成**資料雲的形狀說明書**：對角線說「每個軸各自伸展多寬」，非對角線說「雲往哪個斜方向傾」；特徵分解把說明書翻譯成最直白的語言——「長軸朝 $\boldsymbol{v}_1$、伸展 $\sqrt{\lambda_1}$；短軸朝 $\boldsymbol{v}_2$、伸展 $\sqrt{\lambda_2}$」。也別忘了成本：Σ 有 $n(n+1)/2$ 個參數，高維時往往估不動，實務上常用對角近似（忽略相關）或低秩近似（只留前幾個主成分）。

5.4　多元常態分布

單變數世界的主角是常態分布 $\mathcal{N}(\mu, \sigma^2)$（第 4 章）；向量世界的主角是它的推廣——多元常態：「均值」升級成均值向量、「變異數」升級成共變異數矩陣。

定義 5.10　多元常態分布

隨機向量 $\boldsymbol{x} \in \mathbb{R}^n$ 服從**多元常態分布** (multivariate normal / Gaussian distribution) $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$（$\Sigma$ 為正定矩陣），若其密度為

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\,|\Sigma|^{1/2}} \exp\!\Big(-\tfrac{1}{2}\,(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big),$$

其中 $|\Sigma|$ 是行列式。$n=1$ 時退回熟悉的 $\mathcal{N}(\mu, \sigma^2)$。

$n=2$ 時把 $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ 代入，可展開成常用的**二維常態**形式：

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\!\left(-\frac{z_1^2 - 2\rho z_1 z_2 + z_2^2}{2(1-\rho^2)}\right), \qquad z_i = \frac{x_i - \mu_i}{\sigma_i}.$$

從這條式子立刻看出一件單變數世界看不到的事：當 $\rho = 0$ 時，密度拆成 z_1 的函數乘 z_2 的函數——**對聯合常態而言，不相關就等於獨立**。這是第 5.2 節警告「零相關 $\neq$ 獨立」的唯一著名例外，成立的關鍵是「聯合」二字。

密度的幾何全部藏在指數裡的二次式。集合 $\{\boldsymbol{x} : (\boldsymbol{x}-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c^2\}$ 是以 $\boldsymbol{\mu}$ 為中心的**橢圓（橢球）**：代入 $\Sigma = Q\Lambda Q^{\top}$ 可知它沿特徵向量 $\boldsymbol{v}_i$ 定向、半軸長為 $c\sqrt{\lambda_i}$。三種特例值得記住——**球形** $\Sigma = \sigma^2 I$：正圓輪廓，密度只依賴歐氏距離；**對角** Σ：與座標軸對齊的橢圓，各分量獨立；**一般** Σ：橢圓被拉伸並旋轉，方向由特徵向量決定（圖 5.3）。

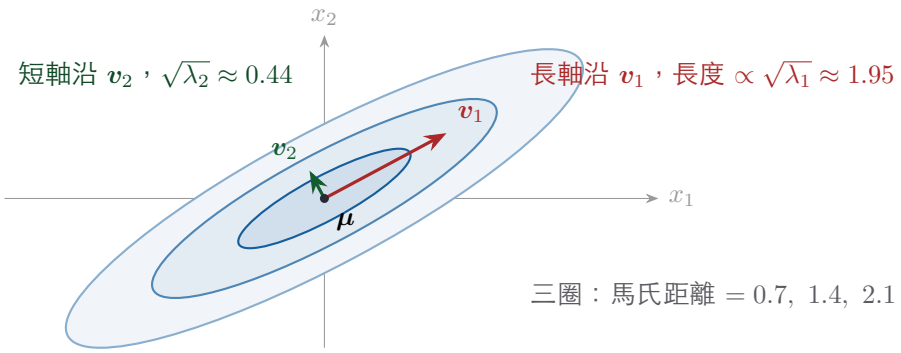


圖 5.3: 多元常態 $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$（取例 5.5 的 Σ）的等密度輪廓：以 $\boldsymbol{\mu}$ 為中心的同心橢圓，沿 Σ 的特徵向量定向，半軸長正比於 $\sqrt{\lambda_i}$。$\lambda_1 \approx 3.80 \gg \lambda_2 \approx 0.20$，橢圓因此又扁又斜——95% 的變異集中在 $\boldsymbol{v}_1$ 方向。

核心觀念

多元常態**完全**由兩個物件決定：均值向量 $\boldsymbol{\mu}$（中心在哪）與共變異數矩陣 Σ（雲長什麼形狀）。指數中的二次式是**馬氏距離的平方**——沿各主軸「以標準差為單位」度量的距離，密度沿其等高線遞減。它是唯一在邊際化、條件化、線性變換下都能**完全解析求解**的向量連續分布——在機器學習中無所不在，不是因為世界都是高斯的，而是因為高斯算得動、又足夠有表達力。

定義 5.11　馬氏距離

點 $\boldsymbol{x}$ 到中心 $\boldsymbol{\mu}$（相對於共變異數 Σ）的**馬氏距離**（Mahalanobis distance）為

$$d_\Sigma(\boldsymbol{x}, \boldsymbol{\mu}) = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}.$$

當 $\Sigma = \sigma^2 I$ 時退化為 $\|\boldsymbol{x}-\boldsymbol{\mu}\|/\sigma$：以標準差為單位的歐氏距離。更漂亮的觀點：馬氏距離就是**白化後的歐氏距離**——令 $\boldsymbol{z} = \Lambda^{-1/2} Q^\top (\boldsymbol{x}-\boldsymbol{\mu})$，即有 $\|\boldsymbol{z}\| = d_\Sigma(\boldsymbol{x}, \boldsymbol{\mu})$。

例 5.6　看起來近，其實遠：馬氏距離 vs. 歐氏距離

設 $\boldsymbol{\mu} = \mathbf{0}$、$\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$（兩變數高度正相關）。比較點 $\boldsymbol{a} = (1,1)^\top$ 與 $\boldsymbol{b} = (1,-1)^\top$ 到中心的兩種距離。

解　兩點的**歐氏距離**相同：$\|\boldsymbol{a}\| = \|\boldsymbol{b}\| = \sqrt{2} \approx 1.414$。馬氏距離則需要

$$\Sigma^{-1} = \frac{1}{1-0.9^2} \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix} = \frac{1}{0.19} \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}.$$

對 $\boldsymbol{a} = (1,1)^\top$，二次式為 $(1 - 2\times 0.9 + 1)/0.19 = 0.2/0.19 \approx 1.053$；對 $\boldsymbol{b} = (1,-1)^\top$ 則是 $(1 + 2\times 0.9 + 1)/0.19 = 3.8/0.19 = 20$。故

$$d_\Sigma(\boldsymbol{a}, \mathbf{0}) = \sqrt{1.053} \approx 1.026, \qquad d_\Sigma(\boldsymbol{b}, \mathbf{0}) = \sqrt{20} \approx 4.472.$$

歐氏眼中等距的兩點，馬氏眼中差了 4.4 倍！$\rho = 0.9$ 的資料雲沿 $(1,1)$ 方向拉得很長（變異數 $1+\rho = 1.9$）、沿 $(1,-1)$ 方向極窄（$1-\rho = 0.1$）：$\boldsymbol{a}$ 順著雲走是尋常樣本，$\boldsymbol{b}$ 逆著雲走是重大異常。距離必須用資料自己的尺度來量——這正是馬氏距離驅動異常偵測的原理。

常見陷阱

「每個分量各自是常態」**並不保證**整個向量是多元常態——「聯合高斯」是比「邊際高斯」強得多的條件。經典反例：$X \sim \mathcal{N}(0,1)$，S 為與 X 獨立的隨機正負號，令 $Y = SX$。可驗證 $Y \sim \mathcal{N}(0,1)$ 且 $\mathrm{Cov}(X,Y) = \mathbb{E}[S]\,\mathbb{E}[X^2] = 0$，但 $|Y| = |X|$ 永遠成立——強烈相依，且 $X+Y$ 有一半機率恰等於 0，根本不是常態。「不相關 $\Rightarrow$ 獨立」只適用於**聯合**常態；看到兩個邊際都是鐘形就套用高斯性質，是真實發生過的建模事故。

5.5 封閉性：邊際、條件與線性變換

多元常態之所以是機器學習的主力，關鍵不只是密度公式，而是它的**封閉性**（closure）：對聯合分布做我們關心的每一種操作，結果仍是多元常態。

定理 5.12　多元常態的封閉性

設 $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$。

1. **線性變換**：對任何矩陣 A 與向量 $\boldsymbol{b}$，

$$A\boldsymbol{x} + \boldsymbol{b} \sim \mathcal{N}\big(A\boldsymbol{\mu} + \boldsymbol{b},\ A\Sigma A^{\top}\big).$$

   特例：任何線性組合 $\boldsymbol{a}^{\top}\boldsymbol{x}$ 都是一維常態；兩個獨立常態向量之和仍是常態。

2. **邊際化**：任何子向量仍是多元常態，參數就是 $\boldsymbol{\mu}$ 與 Σ 中對應的區塊。

3. **條件化**：給定部分分量的值，其餘分量的條件分布仍是多元常態（公式見定理 5.13）。

4. **不相關 $\iff$ 獨立**：聯合常態的兩個分量若共變異數為零，則它們獨立。

封閉性 (1) 也解釋了電腦如何抽樣多元常態：取 $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, I)$，找任何滿足 $LL^{\top} = \Sigma$ 的 L（如 Cholesky 分解），則 $\boldsymbol{\mu} + L\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$——VAE 的重參數化技巧（reparameterization trick）用的正是這條式子。

定理 5.13　條件分布公式

把 $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ 分割成兩塊：

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

則觀測到 $\boldsymbol{x}_2$ 之後，

$$\boldsymbol{x}_1 \mid \boldsymbol{x}_2 \sim \mathcal{N}\big(\boldsymbol{\mu}_{1|2},\ \Sigma_{1|2}\big), \qquad \boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \qquad \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

二維特例（X_1, X_2 聯合常態、相關係數 ρ）：

$$X_1 \mid X_2 = x_2 \ \sim\ \mathcal{N}\Big(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2),\ \ \sigma_1^2(1-\rho^2)\Big).$$

（證明用配方法把聯合密度整理成「條件 × 邊際」，冗長但直接，此處省略。）三個值得咀嚼的讀法：**條件均值是 x_2 的線性函數**，觀測值高於均值多少，預測就按比例 $\rho\sigma_1/\sigma_2$（迴歸係數 $\Sigma_{12}\Sigma_{22}^{-1}$）調整——這是線性迴歸的分布論起源；**條件變異數不依賴觀測值**且 $\sigma_1^2(1-\rho^2) \le \sigma_1^2$——資訊永不增加不確定性，$\rho^2$ 正是「被解釋掉」的變異比例；$\rho = 0$ 時條件分布退回邊際分布——再次印證聯合常態下不相關即獨立。

例 5.7　父子身高與迴歸向平均

假設父親身高 $F \sim \mathcal{N}(172, 6^2)$、兒子身高 $S \sim \mathcal{N}(174, 6^2)$（單位 cm），兩者聯合常態、相關係數 $\rho = 0.5$。(a) 已知父親身高 190，預測兒子身高。(b) 兒子超過 190 的機率多大？

解　**(a)** 套用條件公式：

$$\mathbb{E}[S \mid F = 190] = 174 + 0.5 \times \frac{6}{6} \times (190 - 172) = 174 + 9 = 183,$$

而 $\mathrm{Var}(S \mid F = 190) = 36(1 - 0.25) = 27$，標準差 $\sqrt{27} \approx 5.2$。注意這裡的驚奇：父親比平均高了 18 cm，兒子卻預期只比平均高 9 cm——極端的父親傾向生出**沒那麼極端**的兒子。這個現象稱為**迴歸向平均**（regression to the mean），是 Galton 在 1886 年研究身高遺傳時發現的，「迴歸」一詞就從這裡進入統計學。

(b) 標準化：$\mathbb{P}(S > 190 \mid F = 190) = \mathbb{P}\big(Z > (190 - 183)/5.2\big) = \mathbb{P}(Z > 1.35) \approx 0.089$——即使父親身高 190，兒子超過 190 的機率也不到一成。

常見誤解：迴歸向平均**不是**「身高一代代收斂到平均」——條件化是雙向的（兒子很高時，父親的條件期望同樣較靠近平均），收縮只發生在條件預測裡。

核心觀念

封閉性是高斯的超能力。取邊際、取條件、疊加雜訊、線性投影——只要從高斯出發，結果永遠留在高斯家族，參數更新只需線性代數。**高斯過程**把函數視為無限維高斯向量，觀測幾點後直接用條件公式給出預測曲線與不確定帶；**卡爾曼濾波器**每個時刻用線性變換封閉性做預測、用條件公式吸收新觀測。這些方法能**精確**求解、不需近似，全靠封閉性；一旦離開高斯家族，貝氏更新幾乎都得動用近似推論。

這如何驅動機器學習／ AI

共變異數矩陣與多元常態位居古典與現代機器學習的心臟。**PCA**——最常用的降維方法——就是 Σ 的特徵分解：保留頂部特徵向量，把上千維的特徵壓到幾十維；**白化**把資料變換到 $\Sigma = I$，是標準預處理。**VAE** 的潛在先驗是 $\mathcal{N}(\mathbf{0}, I)$，訓練用 $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\varepsilon}$ 重參數化——正是線性變換封閉性；**擴散模型**的前向過程反覆疊加高斯雜訊，把圖片分布推向 $\mathcal{N}(\mathbf{0}, I)$，再學習逆轉。**高斯過程**以多元常態作為函數的先驗、**卡爾曼濾波器**以它追蹤動態系統、**高斯判別分析**假設每類是一個高斯。**馬氏距離**驅動異常偵測：對正常資料擬合 $(\boldsymbol{\mu}, \Sigma)$，新樣本馬氏距離過大即報警。當模型回報「預測的共變異數」，它正是在用本章的語言表達**相關的不確定性**：哪些量會一起錯、往哪個方向錯。

5.6 Python 實作

理論之後，動手驗證。以下程式示範：抽樣相關的二維高斯並恢復 Σ 與主軸、零相關但相依的反例，以及條件常態的「迴歸向平均」。

```python
import numpy as np

rng = np.random.default_rng(42)

# --- 1. 抽樣相關二維高斯，恢復共變異數矩陣與主軸 ---
mu = np.array([0.0, 0.0])
Sigma = np.array([[3.0, 1.5],
                  [1.5, 1.0]])
X = rng.multivariate_normal(mu, Sigma, size=100_000)

S = np.cov(X.T)                # 樣本共變異數矩陣
print(np.round(S, 3))          # 接近 [[3, 1.5], [1.5, 1]]

lam, Q = np.linalg.eigh(S)     # 特徵分解（特徵值由小到大）
print(np.round(lam, 3))        # 約 [0.197, 3.803] = (4 -+ sqrt(13))/2
print(np.round(Q[:, 1], 3))    # 第一主軸 [0.882, 0.472]（符號可翻轉）

# --- 2. 零相關但不獨立：Y = X^2 ---
x = rng.uniform(-1.0, 1.0, size=1_000_000)
y = x ** 2
print(np.corrcoef(x, y)[0, 1])           # 約 0：相關係數看不見
print(np.corrcoef(np.abs(x), y)[0, 1])   # 約 0.97：關聯其實極強

# --- 3. 條件常態 = 迴歸向平均：父高 190，兒子平均約 183 ---
C = np.array([[36.0, 18.0],
              [18.0, 36.0]])             # Cov = 0.5 * 6 * 6 = 18
```

```
FS = rng.multivariate_normal([172.0, 174.0], C, size=2_000_000)
sel = np.abs(FS[:, 0] - 190.0) < 0.5      # 父親身高約 190 的家庭
print(FS[sel, 1].mean())                  # 約 183 = 174 + 0.5*(190-172)
```

幾個輸出值得對照理論：區塊 1 的樣本特徵值 $\approx (0.197, 3.803)$ 與例 5.5 的 $(4 \mp \sqrt{13})/2$ 吻合（特徵向量正負號可整體翻轉，屬正常現象）；區塊 2 並列了本章的核心教訓——同一組資料，$\mathrm{Corr}(X, Y) \approx 0$ 而 $\mathrm{Corr}(|X|, Y) \approx 0.97$，「零相關」與「無關聯」是兩回事；區塊 3 用模擬驗證條件公式：父親約 190 cm 的家庭，兒子平均正是理論值 183 cm，而非天真的 192 cm。

本章總結

- **聯合分布**（表格或密度曲面）是多變數行為的完整描述；對它只有兩種基本查詢：**邊際化**（把不關心的變數加總 / 積分掉）與**條件化**（切片再重新正規化）。邊際分布不足以決定聯合分布。
- **共變異數** $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ 度量線性共動；**相關係數**把它正規化到 $[-1, 1]$。獨立 $\Rightarrow$ 不相關，但**反之不成立**（$Y = X^2$ 反例）；**相關也不是因果**（共同原因）。
- **共變異數矩陣** $\Sigma = \mathbb{E}[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top]$ 對稱、半正定（$\boldsymbol{a}^\top \Sigma \boldsymbol{a} = \mathrm{Var}(\boldsymbol{a}^\top \boldsymbol{x}) \geq 0$）；其特徵向量是資料雲的主軸、特徵值是沿主軸的變異數——特徵分解就是 **PCA**，$\Lambda^{-1/2} Q^\top$ 就是**白化**。
- **多元常態** $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ 完全由中心與形狀兩個參數決定；等密度輪廓是沿特徵向量定向的橢球，半軸長 $\propto \sqrt{\lambda_i}$；指數中的二次式是**馬氏距離**的平方——白化後的歐氏距離。
- 多元常態的**封閉性**：線性變換、邊際、條件之後仍是高斯；條件公式 $\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ 是高斯過程與卡爾曼濾波器可精確求解的引擎，也帶出**迴歸向平均**。
- 聯合常態下不相關 $\iff$ 獨立——但「邊際皆常態」不等於「聯合常態」。這套語言支撐 PCA、白化、VAE 與擴散模型的先驗、高斯過程、卡爾曼濾波器與馬氏距離異常偵測。

5.7 練習題

建議先獨立完成再對照詳解；最後一題為程式實作。

練習 5.1 (基礎)　某網路商店統計顧客「是否使用折價券」（$X = 1$ 表示使用）與「單筆購買件數」$Y \in \{1, 2, 3\}$ 的聯合分布：

	$y = 1$	$y = 2$	$y = 3$
$x = 0$	0.15	0.25	0.10
$x = 1$	0.05	0.15	0.30

(a) 求兩個邊際分布。(b) 求條件分布 $p_{Y\mid X}(y\mid 1)$ 與 $p_{Y\mid X}(y\mid 0)$，並計算 $\mathbb{E}[Y\mid X=1]$ 與 $\mathbb{E}[Y\mid X=0]$。(c) X 與 Y 獨立嗎？(d) 計算 $\mathbb{E}[XY]$、$\mathrm{Cov}(X,Y)$ 與 $\mathrm{Corr}(X,Y)$。(e) 由 (d) 的正號能否斷言「使用折價券導致顧客多買」？

練習 5.2 (基礎)　證明共變異數的兩條代數性質：(a) $\mathrm{Cov}(aX+b,\,cY+d)=ac\,\mathrm{Cov}(X,Y)$；(b) $\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\,\mathrm{Cov}(X,Y)$。(c) 由 (a) 推論：相關係數在正尺度變換（$a,c>0$）下不變；並說明 $\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)$ 成立的充要條件是「不相關」而非「獨立」。

練習 5.3 (進階)　設 (X,Y) 的聯合密度為 $f(x,y)=x+y$ $(0\le x,y\le 1)$。(a) 驗證 f 是合法的聯合密度。(b) 求邊際密度 f_X, f_Y。(c) 求條件密度 $f_{Y\mid X}(y\mid x)$，並寫出 $x=0$ 時的特例。(d) 判斷 X,Y 是否獨立。(e) 計算 $\mathrm{Cov}(X,Y)$ 與 $\mathrm{Corr}(X,Y)$。

練習 5.4 (進階)　對共變異數矩陣 $\Sigma=\left(\begin{smallmatrix}5 & 2\\ 2 & 2\end{smallmatrix}\right)$：(a) 求特徵值與單位特徵向量。(b) 驗證兩特徵向量互相垂直，並驗證 $\Sigma=\lambda_1 \boldsymbol{v}_1\boldsymbol{v}_1^{\top}+\lambda_2\boldsymbol{v}_2\boldsymbol{v}_2^{\top}$。(c) 第一主成分方向為何？它解釋總變異的多少比例？

練習 5.5 (進階)　(a) 設 $\boldsymbol{\mu}=\mathbf{0}$、$\Sigma=\mathrm{diag}(9,1)$。計算點 $(3,0)$ 與 $(0,3)$ 的歐氏距離與馬氏距離，並解釋兩者排序為何不同。(b) 改取 $\Sigma=\left(\begin{smallmatrix}1 & 0.8\\ 0.8 & 1\end{smallmatrix}\right)$，計算點 $(1,1)$ 與 $(1,-1)$ 的馬氏距離。(c) 證明當 $\Sigma=\sigma^2 I$ 時，馬氏距離退化為 $\|\boldsymbol{x}-\boldsymbol{\mu}\|/\sigma$。

練習 5.6 (進階)　某課程期中考 $X_2\sim\mathcal{N}(65,8^2)$、期末考 $X_1\sim\mathcal{N}(70,10^2)$，兩者聯合常態、$\rho=0.6$。一位學生期中考得 81 分。(a) 求期末成績的條件分布。(b) 求該生期末超過 90 分的機率。(c) 若完全不知道期中成績，期末超過 90 分的機率是多少？比較兩者。(d) 條件標準差與非條件標準差的比值是多少？它依賴觀測到的分數嗎？

練習 5.7 (進階)　對 $\Sigma=\left(\begin{smallmatrix}3 & 1\\ 1 & 3\end{smallmatrix}\right)$：(a) 求特徵分解 $\Sigma=Q\Lambda Q^{\top}$。(b) 寫出白化矩陣 $W=\Lambda^{-1/2}Q^{\top}$。(c) 驗證 $W\Sigma W^{\top}=I$。(d) 白化矩陣唯一嗎？請說明。

練習 5.8 (進階)　概念題。(a) 「各縣市的冰淇淋銷量與溺水人數正相關」與「小學生的鞋號與識字量正相關」各隱藏著什麼共同原因？若據此禁售冰淇淋或給孩子買大鞋，會發生什麼事？(b) 為什麼隨機對照試驗能夠切斷共同原因的影響、讓相關反映因果？(c)「$\mathrm{Corr}(X,Y)=0$」能否推出「X,Y 無關聯」？請舉反例。(d) 在哪一類分布家族中，零相關確實蘊涵獨立？

練習 5.9 (挑戰)　證明 $-1\le\mathrm{Corr}(X,Y)\le 1$。提示：考慮標準化變數 $X^{*}=(X-\mathbb{E}X)/\sigma_X$、$Y^{*}=(Y-\mathbb{E}Y)/\sigma_Y$，並利用 $\mathrm{Var}(X^{*}\pm Y^{*})\ge 0$。並證明：$\rho=\pm 1$ 若且唯若 Y 幾乎必然是

X 的線性函數（斜率符號與 ρ 一致）。

練習 5.10 (挑戰)　(a) 證明任何共變異數矩陣 Σ 都是半正定的：對所有 $\boldsymbol{a} \in \mathbb{R}^n$，$\boldsymbol{a}^\top \Sigma \boldsymbol{a} \geq 0$。(b) 證明 Σ 奇異（存在 $\boldsymbol{a} \neq \mathbf{0}$ 使 $\boldsymbol{a}^\top \Sigma \boldsymbol{a} = 0$）若且唯若某個非平凡線性組合 $\boldsymbol{a}^\top \boldsymbol{x}$ 幾乎必然是常數，並舉一個具體例子。(c) 說明奇異的 Σ 對馬氏距離與多元常態密度造成什麼麻煩，實務上如何補救。

練習 5.11 (挑戰 · 程式)　用 NumPy 完成：(a) 從 $\mathcal{N}\big(\mathbf{0}, \left(\begin{smallmatrix} 3 & 1.5 \\ 1.5 & 1 \end{smallmatrix}\right)\big)$ 抽取 200,000 個樣本，用 `np.cov` 估計共變異數矩陣、`np.linalg.eigh` 求特徵值，並與例 5.5 的理論值 $(4 \pm \sqrt{13})/2$ 比較。(b) 讓 X 均勻取值於 $\{-1, 0, 1\}$、$Y = X^2$，用 `np.corrcoef` 驗證相關約為 0，再用條件比例（例如 $\mathbb{P}(Y = 1 \mid X = 1)$ 與 $\mathbb{P}(Y = 1)$）驗證兩者相依。

5.8 練習題詳解

練習 5.1 詳解　(a) 橫列相加：$p_X(0) = 0.15{+}0.25{+}0.10 = 0.50$，$p_X(1) = 0.05{+}0.15{+}0.30 = 0.50$。直行相加：$p_Y(1) = 0.20$、$p_Y(2) = 0.40$、$p_Y(3) = 0.40$。（檢查：各自總和為 1。）
(b) 用折價券的顧客：

$$p_{Y\mid X}(y \mid 1) = \frac{(0.05, 0.15, 0.30)}{0.50} = (0.1,\ 0.3,\ 0.6), \qquad \mathbb{E}[Y \mid X = 1] = 0.1 + 0.6 + 1.8 = 2.5.$$

不用折價券的顧客：

$$p_{Y\mid X}(y \mid 0) = \frac{(0.15, 0.25, 0.10)}{0.50} = (0.3,\ 0.5,\ 0.2), \qquad \mathbb{E}[Y \mid X = 0] = 0.3 + 1.0 + 0.6 = 1.9.$$

用券顧客平均多買 0.6 件。
(c) 檢查一格即可：$p(1,3) = 0.30$，而 $p_X(1)\,p_Y(3) = 0.50 \times 0.40 = 0.20 \neq 0.30$，故**不獨立**（獨立需要每一格都等於乘積，一格反例即可否定）。
(d) $\mathbb{E}[X] = 0.5$，$\mathbb{E}[Y] = 1 \times 0.2 + 2 \times 0.4 + 3 \times 0.4 = 2.2$；$\mathbb{E}[XY]$ 只有 $x = 1$ 橫列有貢獻：

$$\mathbb{E}[XY] = 1 \times 0.05 + 2 \times 0.15 + 3 \times 0.30 = 1.25, \qquad \operatorname{Cov}(X, Y) = 1.25 - 0.5 \times 2.2 = 0.15 > 0.$$

變異數：$\operatorname{Var}(X) = 0.5 \times 0.5 = 0.25$；$\mathbb{E}[Y^2] = 1 \times 0.2 + 4 \times 0.4 + 9 \times 0.4 = 5.4$，故 $\operatorname{Var}(Y) = 5.4 - 2.2^2 = 0.56$。因此

$$\operatorname{Corr}(X, Y) = \frac{0.15}{\sqrt{0.25}\sqrt{0.56}} = \frac{0.15}{0.5 \times 0.7483} \approx 0.401.$$

(e) **不能**。正號只說明「使用折價券」與「多買」同向出現：可能是折價券刺激購買，也可能本來就打算大買的顧客更會去找折價券（自我選擇），或兩者同受促銷檔期驅動。區分這些解釋需要隨機對照試驗（隨機發放折價券），見練習 5.8。

練習 5.2 詳解　(a) 記 $\mu_X = \mathbb{E}X$、$\mu_Y = \mathbb{E}Y$。則 $\mathbb{E}[aX+b] = a\mu_X + b$，故

$$\begin{aligned}\operatorname{Cov}(aX+b,\ cY+d) &= \mathbb{E}\big[(aX+b-a\mu_X-b)(cY+d-c\mu_Y-d)\big] \\ &= \mathbb{E}\big[ac(X-\mu_X)(Y-\mu_Y)\big] = ac\,\operatorname{Cov}(X,Y).\end{aligned}$$

平移常數 b, d 完全消失；尺度 a, c 各自提出一次。
(b) 展開平方：

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

(c) 由 (a)，當 $a, c > 0$：

$$\operatorname{Corr}(aX+b,\ cY+d) = \frac{ac\,\operatorname{Cov}(X,Y)}{(a\sigma_X)(c\sigma_Y)} = \operatorname{Corr}(X,Y),$$

（若 $ac<0$ 則變號）——相關係數量的是「關聯的形狀」，與單位無關。由 (b)，$\operatorname{Var}(X+Y) = \operatorname{Var}(X)+\operatorname{Var}(Y)$ 若且唯若 $\operatorname{Cov}(X,Y)=0$：**不相關**就夠了，獨立是充分但非必要條件。

練習 5.3 詳解　(a) 在 $[0,1]^2$ 上 $f = x+y \ge 0$，且 $\int_0^1\!\int_0^1 (x+y)\,\mathrm{d}y\,\mathrm{d}x = \int_0^1 (x+\tfrac12)\,\mathrm{d}x = 1$。✓
(b) $f_X(x) = \int_0^1 (x+y)\,\mathrm{d}y = x + \tfrac12$　$(0 \le x \le 1)$；由對稱性 $f_Y(y) = y + \tfrac12$。
(c) $f_{Y\mid X}(y\mid x) = \dfrac{x+y}{x+\frac12}$　$(0 \le y \le 1)$。特例 $x=0$：$f_{Y\mid X}(y\mid 0) = 2y$——一條往右上傾的三角形密度（檢查 $\int_0^1 2y\,\mathrm{d}y = 1$）。
(d) $f_X(x)f_Y(y) = (x+\tfrac12)(y+\tfrac12) \ne x+y$（例如在 $(0,0)$：左邊 $=\tfrac14$、右邊 $=0$），故**相依**。
(e) 先算各期望值：$\mathbb{E}[X] = \int_0^1 x(x+\tfrac12)\,\mathrm{d}x = \tfrac13 + \tfrac14 = \tfrac{7}{12} = \mathbb{E}[Y]$，

$$\mathbb{E}[XY] = \int_0^1\!\int_0^1 xy(x+y)\,\mathrm{d}x\,\mathrm{d}y = \int x^2\,\mathrm{d}x \int y\,\mathrm{d}y + \int x\,\mathrm{d}x \int y^2\,\mathrm{d}y = \tfrac13\cdot\tfrac12 + \tfrac12\cdot\tfrac13 = \tfrac13,$$

故 $\operatorname{Cov}(X,Y) = \tfrac13 - (\tfrac{7}{12})^2 = \tfrac{48}{144} - \tfrac{49}{144} = -\tfrac{1}{144}$。再算 $\mathbb{E}[X^2] = \int_0^1 x^2(x+\tfrac12)\,\mathrm{d}x = \tfrac14 + \tfrac16 = \tfrac{5}{12}$，得 $\operatorname{Var}(X) = \tfrac{5}{12} - \tfrac{49}{144} = \tfrac{11}{144} = \operatorname{Var}(Y)$。因此

$$\operatorname{Corr}(X,Y) = \frac{-1/144}{11/144} = -\frac{1}{11} \approx -0.091.$$

負號說得通：密度偏好 $x+y$ 大的角落，一個變數大時另一個可以稍小。驗證：$\mathbb{E}[Y\mid X=x] = (3x+2)/(6x+3)$ 是 x 的遞減函數，與負相關一致。

練習 5.4 詳解　(a) 特徵方程式：

$$\det\begin{pmatrix} 5-\lambda & 2 \\ 2 & 2-\lambda \end{pmatrix} = \lambda^2 - 7\lambda + (10-4) = (\lambda-6)(\lambda-1) = 0,$$

得 $\lambda_1 = 6$、$\lambda_2 = 1$。對 $\lambda_1 = 6$：$(5-6)v_1 + 2v_2 = 0 \Rightarrow v_1 = 2v_2$，取

$$\boldsymbol{v}_1 = \frac{1}{\sqrt{5}}\begin{pmatrix} 2 \\ 1 \end{pmatrix}; \qquad \text{對}\lambda_2 = 1:\ 4v_1 + 2v_2 = 0 \Rightarrow \boldsymbol{v}_2 = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 \\ -2 \end{pmatrix}.$$

(b) $\boldsymbol{v}_1^\top \boldsymbol{v}_2 = \frac{1}{5}(2\times 1 + 1\times(-2)) = 0$，垂直。重建：

$$6\,\boldsymbol{v}_1\boldsymbol{v}_1^\top + 1\,\boldsymbol{v}_2\boldsymbol{v}_2^\top = \frac{6}{5}\begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix} + \frac{1}{5}\begin{pmatrix} 1 & -2 \\ -2 & 4 \end{pmatrix} = \frac{1}{5}\begin{pmatrix} 25 & 10 \\ 10 & 10 \end{pmatrix} = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}.\checkmark$$

(c) 第一主成分方向是 $\boldsymbol{v}_1 = (2,1)^\top/\sqrt{5}$（斜率 $\frac{1}{2}$ 的直線方向），解釋比例 $\lambda_1/(\lambda_1+\lambda_2) = 6/7 \approx 85.7\%$：投影到這條線上保留約六分之五的變異。

練習 5.5 詳解　(a) 兩點的歐氏距離都是 3。馬氏距離：$d\big((3,0)\big) = \sqrt{3^2/9} = 1$、$d\big((0,3)\big) = \sqrt{3^2/1} = 3$。第一軸標準差是 3，$(3,0)$ 只是「一個標準差」的尋常波動；第二軸標準差是 1，$(0,3)$ 是三個標準差的稀有事件——馬氏距離用「每個方向自己的尺度」重新量距離。
(b) $\det\Sigma = 1 - 0.64 = 0.36$，

$$\Sigma^{-1} = \frac{1}{0.36}\begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}.$$

對 $(1,1)$：$d^2 = \dfrac{1 - 1.6 + 1}{0.36} = \dfrac{0.4}{0.36} = \dfrac{10}{9}$，故 $d = \sqrt{10}/3 \approx 1.054$。對 $(1,-1)$：$d^2 = \dfrac{1 + 1.6 + 1}{0.36} = \dfrac{3.6}{0.36} = 10$，故 $d = \sqrt{10} \approx 3.162$。兩點歐氏距離同為 $\sqrt{2}$，馬氏距離卻差 3 倍：$(1,1)$ 順著正相關的資料雲、$(1,-1)$ 逆著雲。
(c) $\Sigma = \sigma^2 I$ 時 $\Sigma^{-1} = I/\sigma^2$，故

$$d_\Sigma(\boldsymbol{x},\boldsymbol{\mu}) = \sqrt{\frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{x}-\boldsymbol{\mu})}{\sigma^2}} = \frac{\|\boldsymbol{x}-\boldsymbol{\mu}\|}{\sigma}.$$

練習 5.6 詳解　(a) 套用條件公式（定理 5.13 的二維特例）：

$$\mathbb{E}[X_1 \mid X_2 = 81] = 70 + 0.6 \times \frac{10}{8} \times (81 - 65) = 70 + 0.75 \times 16 = 82,$$

而 $\mathrm{Var}(X_1 \mid X_2 = 81) = 10^2(1 - 0.6^2) = 64$，即 $X_1 \mid X_2 = 81 \sim \mathcal{N}(82, 8^2)$。
(b) $\mathbb{P}(X_1 > 90 \mid X_2 = 81) = \mathbb{P}\big(Z > (90-82)/8\big) = \mathbb{P}(Z > 1) \approx 0.159$。
(c) 無條件時 $X_1 \sim \mathcal{N}(70, 10^2)$，故 $\mathbb{P}(X_1 > 90) = \mathbb{P}(Z > 2) \approx 0.023$。知道期中考得好，把「期末破 90」的機率從 2.3% 拉到 15.9%——條件化就是吸收資訊。

(d) 比值 $= \sqrt{1-\rho^2} = \sqrt{0.64} = 0.8$，與觀測到的分數**無關**：不論期中幾分，不確定性都縮小同樣比例——高斯世界的特色，一般分布未必如此。

練習 5.7 詳解　(a) 特徵方程式 $\det(\Sigma - \lambda I) = (3-\lambda)^2 - 1 = 0$，得 $\lambda_1 = 4$、$\lambda_2 = 2$；特徵向量分別為 $(1,1)^\top/\sqrt{2}$ 與 $(1,-1)^\top/\sqrt{2}$（驗證：$\Sigma(1,1)^\top = (4,4)^\top$、$\Sigma(1,-1)^\top = (2,-2)^\top$）。故 $Q = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix}1 & 1\\ 1 & -1\end{smallmatrix}\right)$、$\Lambda = \left(\begin{smallmatrix}4 & 0\\ 0 & 2\end{smallmatrix}\right)$。

(b)

$$W = \Lambda^{-1/2} Q^\top = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{\sqrt{2}} \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix}.$$

(c) 利用 $Q^\top \Sigma Q = \Lambda$：$W\Sigma W^\top = \Lambda^{-1/2}(Q^\top \Sigma Q)\Lambda^{-1/2} = \Lambda^{-1/2}\Lambda\Lambda^{-1/2} = I$。✓（也可數值驗算：$\boldsymbol{w}_1^\top \Sigma \boldsymbol{w}_1 = \frac{1}{8}(1,1)\Sigma(1,1)^\top = \frac{1}{8} \times 8 = 1$。）

(d) **不唯一**：若 W 是白化矩陣、U 是任意正交矩陣，則 $(UW)\Sigma(UW)^\top = UIU^\top = I$——白化後的資料再怎麼旋轉仍是白的。常見選擇：PCA 白化（本題的 $\Lambda^{-1/2}Q^\top$）與 ZCA 白化（$\Sigma^{-1/2} = Q\Lambda^{-1/2}Q^\top$）。

練習 5.8 詳解　(a) 冰淇淋與溺水的共同原因是**夏季高溫**：天氣熱，吃冰與玩水的人同時變多。鞋號與識字量的共同原因是**年齡**：孩子越大，腳越大、識字也越多。介入下游變數不會改變共同原因——禁售冰淇淋不會讓夏天消失，買大鞋不會讓孩子早慧；相關照舊，因果從未存在。

(b) 隨機對照試驗把「誰接受處理」交給**隨機化**決定，處理組與對照組在其他所有因素（含看不見的共同原因）上分布相同，兩組結果的差異只能來自處理本身——隨機化切斷了「共同原因 → 處理」這條路。

(c) 不能。零相關只排除**線性**關聯：例 5.4 的 $Y = X^2$ 相關為零，但 Y 由 X 完全決定；改算 $|X|$ 與 Y 的相關，會發現關聯極強。

(d) **聯合多元常態**家族：其密度在 $\rho = 0$ 時剛好因式分解，零相關即獨立。前提是「聯合」常態——邊際皆常態但聯合非常態時（第 5.4 節陷阱框的 $Y = SX$），零相關仍可能高度相依。

練習 5.9 詳解　標準化 $X^* = (X - \mathbb{E}X)/\sigma_X$、$Y^* = (Y - \mathbb{E}Y)/\sigma_Y$，則 $\mathrm{Var}(X^*) = \mathrm{Var}(Y^*) = 1$，且 $\mathrm{Cov}(X^*, Y^*) = \rho$（用練習 5.2 的雙線性）。由變異數非負與和的公式：

$$0 \le \mathrm{Var}(X^* \pm Y^*) = 1 + 1 \pm 2\rho = 2(1 \pm \rho) \quad \Longrightarrow \quad \rho \ge -1 \text{ 且} \rho \le 1.$$

（這其實是 Cauchy–Schwarz 不等式 $|\mathbb{E}[UV]| \le \sqrt{\mathbb{E}[U^2]\mathbb{E}[V^2]}$ 取 $U = X - \mathbb{E}X$、$V = Y - \mathbb{E}Y$ 的機率版本。）

等號條件：$\rho = 1$ 時 $\mathrm{Var}(X^* - Y^*) = 0$，而變異數為零的隨機變數幾乎必然等於其均值（此處為 0），故 $Y^* = X^*$ 幾乎必然，展開得 $Y = \mathbb{E}Y + \frac{\sigma_Y}{\sigma_X}(X - \mathbb{E}X)$：一條斜率為正的直線；同理 $\rho = -1$ 給出斜率為負的直線。反之若 $Y = aX + b$（$a \neq 0$），由雙線性 $\rho = a\sigma_X^2/(\sigma_X \cdot |a|\sigma_X) = \mathrm{sign}(a)$。故 $|\rho| = 1$ 恰好刻畫完美線性關係。

練習 5.10 詳解　(a) 對任意 $\boldsymbol{a}$：

$$\boldsymbol{a}^\top \Sigma \boldsymbol{a} = \boldsymbol{a}^\top \mathbb{E}\big[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\big]\boldsymbol{a} = \mathbb{E}\big[\big(\boldsymbol{a}^\top(\boldsymbol{x}-\boldsymbol{\mu})\big)^2\big] \geq 0,$$

中間把常數向量 $\boldsymbol{a}$ 移進期望值（線性），平方項非負故期望值非負。又 $\mathbb{E}[\boldsymbol{a}^\top(\boldsymbol{x}-\boldsymbol{\mu})] = 0$，所以這個量正是 $\mathrm{Var}(\boldsymbol{a}^\top\boldsymbol{x})$：任何線性組合的變異數都非負。
(b) 由 (a) 的等式，$\boldsymbol{a}^\top\Sigma\boldsymbol{a} = 0 \iff \mathrm{Var}(\boldsymbol{a}^\top\boldsymbol{x}) = 0 \iff \boldsymbol{a}^\top\boldsymbol{x}$ 幾乎必然等於常數——資料被線性等式束縛、整團躺在低維超平面上。例：$X_2 = 2X_1$、$\mathrm{Var}(X_1) = 1$，則 $\Sigma = \left(\begin{smallmatrix}1 & 2\\ 2 & 4\end{smallmatrix}\right)$ 奇異，取 $\boldsymbol{a} = (2,-1)^\top$ 得 $\mathrm{Var}(2X_1 - X_2) = 0$。
(c) 奇異時 Σ^{-1} 不存在：馬氏距離與多元常態密度都無法定義。補救：刪除共線的冗餘特徵、用偽逆（pseudo-inverse），或**正則化**——以 $\Sigma + \varepsilon I$ 取代 Σ；樣本數少於維度時樣本共變異數必然奇異，正則化幾乎是標準動作。

練習 5.11 詳解　參考程式與輸出：

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

# (a) 抽樣、估計共變異數、特徵分解
Sigma = np.array([[3.0, 1.5], [1.5, 1.0]])
X = rng.multivariate_normal([0.0, 0.0], Sigma, size=200_000)
S = np.cov(X.T)
lam, Q = np.linalg.eigh(S)
print(np.round(S, 3))        # 約 [[3.003, 1.500], [1.500, 0.999]]
print(np.round(lam, 4))      # 約 [0.1972, 3.8028] = (4 -+ sqrt(13))/2

# (b) 零相關但相依
x = rng.choice([-1, 0, 1], size=1_000_000)
y = x ** 2
print(round(float(np.corrcoef(x, y)[0, 1]), 4))  # 約 0.00x
print(y[x == 1].mean(), y.mean())  # P(Y=1|X=1) = 1.0，P(Y=1) 約 0.667
```

(a) 樣本共變異數與真值偏差約 10^{-2} 量級（以 $1/\sqrt{N}$ 收斂）；特徵值與 $(4 \pm \sqrt{13})/2$ 一致。
(b) 樣本相關約 0，但 $\mathbb{P}(Y{=}1 \mid X{=}1) = 1 \neq \mathbb{P}(Y{=}1) \approx \frac{2}{3}$，鐵證相依——相關係數的沉默，不代表變數之間沒有故事。

下一章：極限定理——大數法則與中央極限定理，「平均」為何是統計學最可靠的武器。