機率與統計：從入門到精通

第 8 章　信賴區間與假設檢定

Confidence Intervals and Hypothesis Testing

難度：進階

單獨的點估計是危險的：它把自身的不確定性藏了起來。報告「模型準確率 91.2%」卻不附任何誤差資訊，就像報告氣溫卻不說單位。本章替估計加上**誤差棒**：**信賴區間**表達估計的精確程度；**假設檢定**判斷觀察到的效應是真實訊號還是機運巧合。這兩件頻率學派的工具，是我們誠實比較模型、負責任地報告結果、避免被雜訊愚弄的方法——包括被自己 benchmark 數字裡的雜訊愚弄。

學習目標

讀完本章後，你將能夠：

- 解釋**抽樣分布**與**標準誤**的意義，計算 $\mathrm{SE}(\bar{X}) = \sigma/\sqrt{n}$，並說明誤差為何只以 $1/\sqrt{n}$ 縮小。
- 建構均值與比例的**信賴區間**，正確陳述「覆蓋率」的頻率意義，避開最常見的誤讀。
- 執行完整的**假設檢定**流程（H_0 / H_1、檢定統計量、拒絕域、p 值、結論），並正確詮釋 p 值。
- 用 2×2 表分析**型一／型二錯誤**，計算**檢定力**並做樣本數規劃；為**多重檢定**做 Bonferroni 或 FDR 校正。
- 依資料型態選用 z、t、卡方或 ANOVA 檢定，並以**自助法**（bootstrap）對任意統計量估計不確定性。
- 用 Python / SciPy 計算信賴區間、執行 t 檢定，並親手實作 bootstrap。

8.1 抽樣分布與標準誤

第 7 章教我們從資料計算**點估計**（point estimate）：樣本平均 $\bar{X}$ 估計母體均值 μ、樣本比例 $\hat{p}$ 估計真實比例 p。但想像把整個實驗重做一次——同一個母體、重新抽 n 筆資料——你會得到**不同的** $\bar{X}$。估計量本身是隨機變數，它有自己的分布。這個分布是本章一切工具的地基。

定義 8.1　抽樣分布與標準誤

估計量 $\hat{\theta} = g(X_1, \dots, X_n)$ 在重複抽樣下的機率分布，稱為它的**抽樣分布**（sampling distribution）。抽樣分布的標準差稱為**標準誤**（standard error, SE）。對樣本平均而言，

$$\mathrm{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}},$$

其中 σ 是母體標準差；實務上 σ 未知，以樣本標準差 s 代入，得估計的標準誤 $\widehat{\mathrm{SE}} = s/\sqrt{n}$。

定理 8.2　樣本平均的抽樣分布

設 $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim}$ 抽自均值 μ、變異數 σ^2 的母體，則

$$\mathbb{E}[\bar{X}] = \mu, \qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n},$$

且由中央極限定理（第 6 章），當 n 夠大時，無論母體是什麼分布，

$$\bar{X} \overset{\text{近似}}{\sim} N\!\left(\mu,\ \frac{\sigma^2}{n}\right).$$

證明　期望值由線性：$\mathbb{E}[\bar{X}] = \frac{1}{n}\sum_i \mathbb{E}[X_i] = \mu$，所以抽樣分布**以真值為中心**。變異數由獨立性（獨立變異數可加）：$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\sum_i \mathrm{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$，取平方根即得 $\mathrm{SE}(\bar{X}) = \sigma/\sqrt{n}$。近似常態性正是第 6 章的中央極限定理。 ■

兩個值得停下來咀嚼的事實：

- **標準誤與母體大小無關**，只與樣本數 n 有關。1000 人的隨機民調，對 2300 萬人的母體與對 3 億人的母體一樣準——這是抽樣理論最反直覺也最實用的結論之一。
- **標準誤只以 $1/\sqrt{n}$ 縮小**。資料變 4 倍，誤差才減半；想多爭取一位小數的精度，資料得變 100 倍。

直觀理解

標準誤就是「估計值與真值的**典型距離**」。$n = 100$、$s = 4$ 時 $\widehat{\mathrm{SE}} = 0.4$：你的樣本平均通常落在真均值 ± 0.4 附近（約 68% 的機率在 ± 1 個 SE 內、95% 在 ± 2 個 SE 內）。而 $1/\sqrt{n}$ 是一道殘酷的匯率：想把誤差減半，資料要 4 倍；想把誤差壓到十分之一，資料要 100 倍。「資料的邊際價值遞減」這句行話，數學出處就在這裡。

常見陷阱

標準差 $\neq$ 標準誤。標準差 s 描述**個別資料點**的散布程度——收再多資料它也不會變小；標準誤 $s/\sqrt{n}$ 描述**樣本平均這個估計量**的不確定性——隨 n 增加而縮小。看到「12.3 ± 0.4」時，務必弄清楚 $\pm$ 後面是 SD 還是 SE，兩者相差 $\sqrt{n}$ 倍：誤把 SD 讀成 SE，會把估計想得太不精確；誤把 SE 讀成 SD，會嚴重低估個體差異。論文圖表的誤差棒兩種畫法都有人

用，不標示清楚就是在製造誤會。

8.2 信賴區間：替估計加上誤差棒

知道了「估計的典型誤差是一個標準誤」，最自然的下一步是把它報告出來：不只說「估計值 12.3」，而是說「12.3，上下大約 0.8 以內」。**信賴區間**把這句話做成一個有精確頻率保證的程序。

定義 8.3　信賴區間與覆蓋率

參數 θ 的 $1-\alpha$ **信賴區間**（confidence interval, CI）是由資料計算出的隨機區間 $[L,\,U]$，滿足：在重複抽樣下，

$$\mathbb{P}\big(L \le \theta \le U\big) = 1-\alpha.$$

這個機率描述的對象是**產生區間的程序**：重複做很多次實驗、每次都依同樣規則算出一條區間，長期而言有 $1-\alpha$ 的比例會蓋住真值，這個比例稱為**覆蓋率**（coverage）。最常用 $\alpha = 0.05$，即 95% 信賴區間。

命題 8.4　大樣本均值的信賴區間

當 n 夠大時，μ 的 $1-\alpha$ 信賴區間為

$$\bar{X} \;\pm\; z_{\alpha/2}\,\frac{s}{\sqrt{n}},$$

其中 $z_{\alpha/2}$ 是標準常態的雙尾臨界值：90% 用 1.645、95% 用 1.960、99% 用 2.576。

證明　由定理 8.2，$Z = (\bar{X}-\mu)/(\sigma/\sqrt{n})$ 近似 $N(0,1)$，故

$$\mathbb{P}\Big(-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\Big) \approx 1-\alpha.$$

把括號內的不等式對 μ 解出，得 $\bar{X} - z_{\alpha/2}\,\sigma/\sqrt{n} \le \mu \le \bar{X} + z_{\alpha/2}\,\sigma/\sqrt{n}$。注意**隨機的是兩個端點**（它們由 $\bar{X}$ 算出），μ 自始至終是固定常數。大 n 時以 s 代 σ 不影響近似；n 小時需改用 t 臨界值（第 8.4 節）。 ■

例 8.1　App 使用時長的信賴區間

你抽樣 $n = 100$ 位使用者，量得每日使用 App 的平均時長 $\bar{x} = 12.3$ 分鐘、樣本標準差 $s = 4.0$ 分鐘。求 μ 的 95% 與 99% 信賴區間。

解　標準誤 $\widehat{\mathrm{SE}} = s/\sqrt{n} = 4.0/\sqrt{100} = 0.4$。

95%：$12.3 \pm 1.96 \times 0.4 = 12.3 \pm 0.78$，即 $[11.52,\ 13.08]$ 分鐘。

99%：$12.3 \pm 2.576 \times 0.4 = 12.3 \pm 1.03$，即 $[11.27,\ 13.33]$ 分鐘。

想要更高的信心，就得撒更寬的網：99% 區間必然比 95% 寬。反過來，若想在同樣信心下把區間縮成一半寬，唯一的辦法是把樣本數變成 4 倍（$n = 400$）。

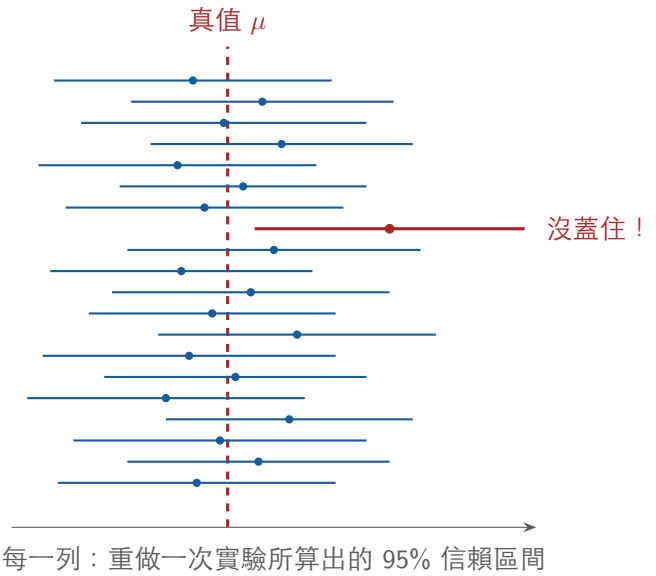


圖 8.1: 信賴區間的正確圖像：真值 μ 固定不動（紅色虛線），隨機的是一條條區間。重複實驗 20 次，約 19 條（95%）蓋住真值、1 條（5%）沒蓋住。「95%」是這個**製造區間的程序**的長期成功率，不屬於任何一條特定的區間。

常見陷阱

信賴區間的詮釋微妙，而且不斷被扭曲。算出 $[11.52,\ 13.08]$ 之後，說「μ 有 95% 的機率落在 11.52 到 13.08 之間」是**錯的**——在頻率觀點裡 μ 是固定常數、不是隨機變數；隨機的是區間本身。這一條特定的區間要嘛蓋住了 μ、要嘛沒蓋住，只是我們不知道是哪一種，沒有「95% 機率」可言。95% 描述的是**程序**的長期覆蓋率：重複實驗一百萬次、每次都算一條 95% CI，約 95 萬條會蓋住真值（圖 8.1）。如果你真正想說的是「參數有 95% 機率在這個區間內」——那是對參數的機率陳述，只有把參數視為隨機變數的貝氏框架才能合法提供：第 9 章的**可信區間**（credible interval）。

直觀理解

信賴區間 = 估計值 ± **誠實的誤差邊際**。區間寬，表示還很不確定（樣本小、變異大）；區間窄，才可以放心。看 benchmark 時這直接翻譯成一條紀律：模型 A 的準確率區間 $[89.9,\ 92.5]$、模型 B 的 $[89.5,\ 92.1]$——兩條區間大幅重疊，表面的贏家很可能只是運氣好。這正是嚴肅的 benchmark 要求報告誤差棒、而不是光禿禿一個數字的原因。

例 8.2　民調支持率的信賴區間

民調訪問 $n = 1000$ 位選民，520 位表態支持候選人甲，樣本支持率 $\hat{p} = 0.52$。求 95% 信賴區間；能宣稱「支持率過半」嗎？

解　樣本比例是 0 / 1 資料的樣本平均，標準誤為 $\sqrt{\hat{p}(1-\hat{p})/n}$（伯努利變異數 $p(1-p)$

代入 $\sqrt{\mathrm{Var}/n}$）：

$$\widehat{\mathrm{SE}} = \sqrt{\frac{0.52 \times 0.48}{1000}} = 0.0158, \qquad 0.52 \pm 1.96 \times 0.0158 = 0.52 \pm 0.031,$$

即 95% CI 為 $[0.489,\ 0.551]$。區間包含 0.5：資料與「支持率其實未過半」完全相容，**不能**宣稱過半。

新聞常說「在 95% 信心水準下，抽樣誤差為 ±3.1 個百分點」——這正是 $n = 1000$、取最保守 $\hat{p} = 0.5$ 時的半寬 $1.96\sqrt{0.25/1000} \approx 0.031$。下次看到民調，你可以自己驗算這個數字。

8.3 假設檢定：效應是真的，還是巧合？

信賴區間回答「參數大概在哪裡」；**假設檢定**回答是非題：效應到底存不存在？新藥有效嗎？新版介面真的比較好嗎？骰子公正嗎？頻率學派對這類問題的回答方式，是讓「沒有效應」這個假設接受資料的審判。

定義 8.5　虛無假設、對立假設與檢定統計量

假設檢定（hypothesis testing）讓兩個對立的主張對決：

- **虛無假設**（null hypothesis）H_0：「沒有效應、沒有差異」——懷疑論者的預設立場；
- **對立假設**（alternative hypothesis）H_1：「效應存在」——我們想證明的主張。

從資料計算**檢定統計量**（test statistic）：它量化資料偏離 H_0 的程度，且在 H_0 為真時分布已知。整個檢定只問一句話：**若 H_0 為真，觀察到的統計量會有多極端？**

直觀理解

假設檢定的邏輯就是法庭。H_0 = 被告無罪（無罪推定），資料 = 證據，拒絕 H_0 = 定罪——需要證據強到「若被告真的無罪，很難出現這些證據」。兩個直接推論：(1) **不定罪 $\neq$ 證明無罪**：可能只是證據不足（樣本太小、檢定力不夠），所以我們說「不拒絕 H_0」而永遠不說「接受 H_0」。(2) 顯著水準 α 是我們願意「冤枉好人」的比率上限——法庭把它設得很低，代價是會放走一些壞人（型二錯誤上升，見第 8.4 節）。

定義 8.6　p 值、顯著水準與拒絕域

p 值（p-value）= **假設 H_0 為真**時，觀察到「至少與實際資料一樣極端」的資料的機率。p 值越小，資料在 H_0 下越令人意外，H_0 越可疑。

事先選定**顯著水準**（significance level）α（慣例 0.05）：若 $p < \alpha$ 就**拒絕** H_0，稱結果**統計顯著**（statistically significant）。等價地，可把「會導致拒絕」的統計量範圍畫成**拒絕域**（rejection region）：$\alpha = 0.05$ 時，雙尾檢定為 $|z| > 1.96$（圖 8.2），單尾檢定為 $z > 1.645$（或 $z < -1.645$）。H_1 是「不等於」用雙尾；H_1 有明確方向（大於 / 小於）用單尾。

一次完整的檢定，永遠是同樣五步：

1. 寫下 H_0 與 H_1，並決定單尾或雙尾；
2. 選定顯著水準 α 與適用的檢定（見第 8.4 節表 8.2）；
3. 由資料計算檢定統計量；
4. 計算 p 值（或比較統計量與臨界值）；
5. 下結論：$p < \alpha$ 則拒絕 H_0；否則**不拒絕** H_0——不是「接受 H_0」，證據不足不等於證明了無效應。

例 8.3　單樣本 z 檢定：包裝重量的宣稱

工廠宣稱麥片平均每包 500 g，長期製程資料顯示重量標準差 $\sigma = 12$ g。你隨機抽 $n = 36$ 包，得 $\bar{x} = 494.5$ g。在 $\alpha = 0.05$ 下檢驗宣稱是否可信。

解　依五步流程：

(1) $H_0\colon \mu = 500$（宣稱屬實）對 $H_1\colon \mu \neq 500$（雙尾——偏高或偏低都違反宣稱）。

(2) σ 已知、$n = 36$，用 z 檢定，$\alpha = 0.05$。

(3) 標準誤 $\mathrm{SE} = \sigma/\sqrt{n} = 12/6 = 2$，

$$z = \frac{\bar{x} - \mu_0}{\mathrm{SE}} = \frac{494.5 - 500}{2} = -2.75,$$

樣本平均比宣稱低了 2.75 個標準誤。

(4) 雙尾 p 值 $= 2\,\mathbb{P}(Z \le -2.75) = 2 \times 0.0030 = 0.006$。

(5) $p = 0.006 < 0.05$，**拒絕** H_0：若宣稱屬實，36 包的平均低到 494.5 g 的機率只有千分之六——我們寧可懷疑宣稱，也不相信自己剛好撞見千分之六的巧合。

交叉檢查：μ 的 95% CI 為 $494.5 \pm 1.96 \times 2 = [490.6,\ 498.4]$，不含 500，與拒絕的結論一致（見下方核心觀念）。

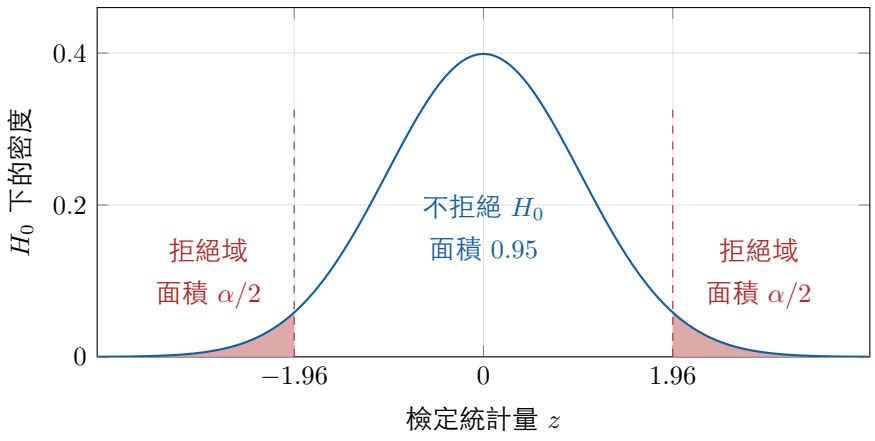


圖 8.2: 雙尾檢定的拒絕域（$\alpha = 0.05$）：若 H_0 為真，統計量落入兩條紅色尾巴的機率僅 5%。當觀察到 $|z| > 1.96$，我們寧可懷疑 H_0，也不相信自己剛好撞見了 5% 的巧合。

核心觀念

信賴區間與檢定是同一枚硬幣的兩面。「μ_0 落在 95% 信賴區間之外」 $\Longleftrightarrow$ 「雙尾檢定在 $\alpha = 0.05$ 下拒絕 $H_0\colon \mu = \mu_0$」。信賴區間一次回答了所有檢定：它就是「所有不會被資料拒絕的參數值」所成的集合。這也是實務上偏愛報告區間的原因——一條區間比一個「顯著 / 不顯著」的二元標籤資訊量大得多：它同時顯示效應的方向、大小與不確定性。

常見陷阱

p 值可能是科學中被誤解最多的數字。$p = 0.03$ **不是**「H_0 為真的機率是 3%」，**不是**「結果是僥倖的機率是 3%」，也**不是**「有 97% 的機率能重現」。它是 $\mathbb{P}(\text{資料至少這麼極端} \mid H_0)$——一個「給定假設，資料多意外」的陳述，而不是「給定資料，假設多可信」的陳述。想把 $\mathbb{P}(\text{資料} \mid H_0)$ 反轉成 $\mathbb{P}(H_0 \mid \text{資料})$，需要貝氏定理與先驗（第 2 章、第 9 章）；沒有先驗，這個反轉在邏輯上不可能。另外要小心 **p-hacking**（p 值操縱）：測試夠多東西，總有一些會純憑機運跨過 0.05——第 8.4 節詳談。

例 8.4　兩個模型的準確率差異是真的嗎？

你的新模型在 $n = 2000$ 筆的測試集上準確率 91.2%，基線模型 90.8%。「提升 0.4 個百分點」是真本事，還是雜訊？

解　把兩個準確率視為比例 $\hat{p}_A = 0.912$、$\hat{p}_B = 0.908$，差異 $\hat{d} = 0.004$。差的標準誤（獨立近似）：

$$\mathrm{SE}(\hat{d}) = \sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n} + \frac{\hat{p}_B(1-\hat{p}_B)}{n}} = \sqrt{\frac{0.912 \times 0.088 + 0.908 \times 0.092}{2000}} = 0.00905.$$

檢定 $H_0\colon d = 0$：$z = 0.004/0.00905 = 0.44$，$p = 2\,\mathbb{P}(Z > 0.44) \approx 0.66$——在 H_0 下，這種程度的差異出現機率高達三分之二。差異的 95% CI 為

$$0.004 \pm 1.96 \times 0.00905 = [-0.014,\ 0.022],$$

大剌剌地跨過零。**結論：什麼都還不能說**——真實差異可能是 -1.4 到 $+2.2$ 個百分點之間的任何值。要可靠偵測 0.4 個百分點的差異，測試集需要約 8 萬筆（例 8.5 會算給你看）。技術註記：兩模型作答的是**同一份**測試集，嚴謹的比較應使用配對檢定（如 McNemar 檢定），通常更靈敏；但此例結論不變——0.4 個百分點在 $n = 2000$ 下遠低於偵測門檻。

8.4　錯誤、檢定力與多重檢定

8.4.1　兩種錯誤與檢定力

檢定是在雜訊中做二元決策，永遠冒著兩種風險。真實情況（H_0 真或假）配上我們的決策（拒絕或不拒絕），構成表 8.1 的 2×2 格局。

表 8.1: 檢定的兩種錯誤：每個格子都有名字，兩種錯誤的機率此消彼長。

	未拒絕 H_0	拒絕 H_0
H_0 為真（無效應）	正確（機率 $1-\alpha$）	**型一錯誤**（偽陽性），機率 α
H_0 為假（有效應）	**型二錯誤**（偽陰性），機率 β	正確：**檢定力** $1-\beta$

定義 8.7　型一錯誤、型二錯誤與檢定力

- **型一錯誤**（Type I error）：H_0 為真卻拒絕——「看見不存在的效應」（偽陽性，false positive）。其機率就是顯著水準 α，由我們直接設定。
- **型二錯誤**（Type II error）：H_0 為假卻未拒絕——「漏掉真實的效應」（偽陰性，false negative），機率記為 β。
- **檢定力**（power）$= 1-\beta$：真效應存在時，成功偵測到它的機率。檢定力隨**樣本數**與**效應大小**（effect size）增加。

兩種錯誤此消彼長：把 α 壓低等於把臨界值外推，拒絕變難，β 隨之上升（圖 8.3）。想同時壓低兩者，唯一的辦法是取得更多資訊——更大的樣本、更精準的測量、或更大的真實效應。

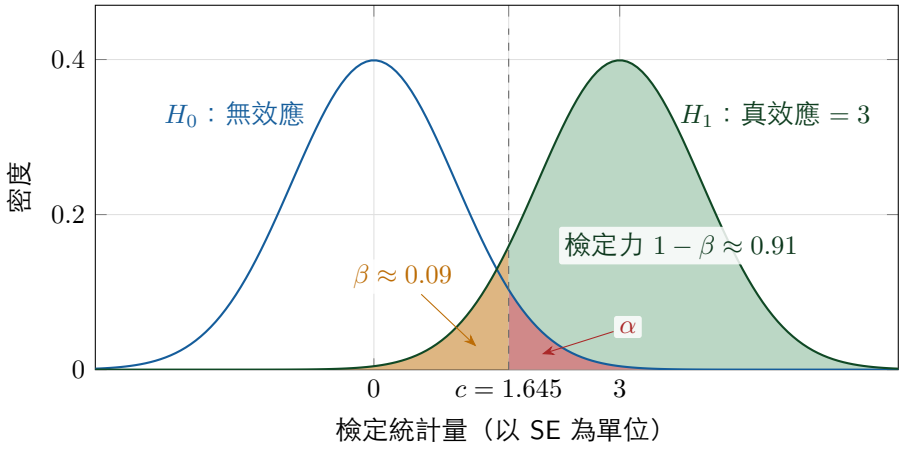


圖 8.3: 單尾檢定（$\alpha = 0.05$，臨界值 $c = 1.645$）下的兩種錯誤：紅色 = 型一錯誤率 α（H_0 為真卻越過 c）；琥珀色 = 型二錯誤率 β（效應存在卻沒越過 c）；綠色 = 檢定力 $1-\beta$。把 c 右移可壓低 α，但 β 隨之變大；唯有增加樣本數（縮小 SE、把兩座山峰拉開）能同時改善兩者。

例 8.5　檢定力與樣本數規劃

新藥宣稱能降血壓。設計單尾 z 檢定：$H_0\colon \mu = 0$（無效）對 $H_1\colon \mu > 0$，$\alpha = 0.05$，個體反應的標準差已知 $\sigma = 15$ mmHg，假設真實效應為 $\mu = 5$ mmHg。

(a) 若 $n = 36$，檢定力多大？　**(b)** 要達到 90% 檢定力，需要多少受試者？　**(c)** 回到例 8.4：要以 80% 檢定力偵測 0.4 個百分點的準確率差異，測試集要多大？

解　**(a)** $\mathrm{SE} = 15/\sqrt{36} = 2.5$，拒絕條件為 $\bar{X} > z_{0.05}\,\mathrm{SE} = 1.645 \times 2.5 = 4.11$。當真值

$\mu = 5$：

$$\text{檢定力} = \mathbb{P}(\bar{X} > 4.11 \mid \mu = 5) = \mathbb{P}\Big(Z > \frac{4.11 - 5}{2.5}\Big) = \mathbb{P}(Z > -0.355) = 0.64.$$

即使藥真的有效，36 人的試驗仍有約 36% 的機率什麼都測不到。

(b) 要求檢定力 $1-\beta$ 的一般公式：拒絕門檻 $z_\alpha\,\sigma/\sqrt{n}$ 必須落在 H_1 分布（中心 δ）往左 z_β 個標準誤處，即 $\delta\sqrt{n}/\sigma = z_\alpha + z_\beta$，整理得

$$n = \Big(\frac{(z_\alpha + z_\beta)\,\sigma}{\delta}\Big)^2 = \Big(\frac{(1.645 + 1.282)\times 15}{5}\Big)^2 = 77.1 \quad\Longrightarrow\quad n = 78.$$

(c) 比例差的版本同理（雙尾用 $z_{\alpha/2}$）：

$$n = \frac{(z_{\alpha/2} + z_\beta)^2\,\big[p_A(1-p_A) + p_B(1-p_B)\big]}{d^2} = \frac{(1.960 + 0.842)^2 \times 0.1638}{0.004^2} \approx 80{,}000.$$

$n = 2000$ 的測試集對這種毫米級的差異完全無能為力——例 8.4 的區間為何那麼寬，答案在此。**先算檢定力、再做實驗**，才不會做一場注定看不見答案的實驗。

8.4.2 常見檢定一覽：z、t、卡方與 ANOVA

五步流程是通用骨架，換掉「檢定統計量與其分布」就得到不同的檢定。表 8.2 是最常用的四種。

表 8.2: 四種最常用的檢定：依資料型態與問題選擇。

檢定	典型問題	檢定統計量	H_0 下的分布
z 檢定	均值或比例；大樣本或 σ 已知	$z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$N(0,1)$
t 檢定	均值；σ 未知、小樣本；兩組均值比較	$t = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}}$	t_{n-1}（厚尾）
卡方檢定	類別計數：適合度、獨立性	$\chi^2 = \displaystyle\sum_i \frac{(O_i - E_i)^2}{E_i}$	χ^2_{k-1}
ANOVA	三組以上均值是否全部相等	$F = \dfrac{\text{組間均方}}{\text{組內均方}}$	$F_{k-1,\,N-k}$

例 8.6　小樣本：t 檢定對照 z 檢定

咖啡店宣稱手沖出杯溫度平均 65°C。你量了 $n = 9$ 杯：$\bar{x} = 62.5$、$s = 3.0$。小樣本且 σ 未知——用 t 檢定，$\alpha = 0.05$（雙尾）。

解　$\widehat{\mathrm{SE}} = s/\sqrt{n} = 3/3 = 1$，

$$t = \frac{62.5 - 65}{1} = -2.5, \qquad \text{自由度} n - 1 = 8.$$

t 分布的臨界值 $t_{0.025,8}=2.306$，比 z 的 1.96 大——t 的厚尾在替「用 s 估計 σ 的額外不確定性」買單。$|t|=2.5>2.306$，拒絕 H_0；精確 p 值 $=2\,\mathbb{P}(T_8\le -2.5)=0.037$。
對照：若錯把它當 z 檢定，$p=2\,\mathbb{P}(Z\le -2.5)=0.012$——比正確值小了三倍，小樣本誤用 z 會系統性誇大證據強度。n 超過 30 之後 t_{n-1} 與 $N(0,1)$ 漸趨重合，差異才可忽略。小樣本的信賴區間同樣改用 t 臨界值：$62.5\pm 2.306\times 1=[60.2,\ 64.8]$，不含 65，與檢定結論一致。

例 8.7　卡方適合度檢定：骰子公正嗎？

擲一顆骰子 120 次，各點數出現次數如下。有人看到「1 出現 25 次、3 只出現 15 次」就嚷嚷骰子被動了手腳。用 $\alpha=0.05$ 檢定。

點數	1	2	3	4	5	6
觀察次數 O_i	25	17	15	23	24	16
期望次數 E_i	20	20	20	20	20	20
$(O_i-E_i)^2/E_i$	1.25	0.45	1.25	0.45	0.80	0.80

解　H_0：骰子公正（每面機率 1/6），此時每面期望次數 $E_i=120/6=20$。**卡方統計量**把每格的偏離平方後除以期望次數再加總：

$$\chi^2=\sum_{i=1}^{6}\frac{(O_i-E_i)^2}{E_i}=1.25+0.45+1.25+0.45+0.80+0.80=5.0.$$

H_0 下 χ^2 近似服從自由度 $k-1=5$ 的卡方分布。臨界值 $\chi^2_{0.05,5}=11.07$，而 $5.0<11.07$；$p=\mathbb{P}(\chi^2_5\ge 5.0)=0.42$。**不拒絕** H_0：這種程度的高低起伏，公正骰子隨隨便便就擲得出來——光用眼睛看原始計數，很容易把純雜訊看成陰謀。
同款工具的另一用途是**獨立性檢定**：把兩個類別變數列成列聯表，檢定「版位與點擊是否獨立」之類的問題，統計量同形，自由度改為 $(r-1)(c-1)$。

表 8.2 的最後一列是**變異數分析**（analysis of variance, ANOVA）：要比較**三組以上**的均值（$H_0\colon \mu_1=\mu_2=\cdots=\mu_k$）時，ANOVA 計算

$$F=\frac{\text{組間變異（各組均值彼此差多遠）}}{\text{組內變異（每組內部的自然雜訊）}},$$

在 H_0 下服從 F 分布。F 遠大於 1，表示組間差異大到組內雜訊解釋不了，拒絕「全部相等」。為什麼不乾脆對每一對組各做一次 t 檢定？因為 k 組有 $\binom{k}{2}$ 對——$k=10$ 就是 45 個檢定。這正是下一小節的主題：檢定做多了，機運自己會送你「顯著」。

8.4.3 多重檢定：機運自己會製造顯著

每個檢定都自帶 $\alpha = 0.05$ 的偽陽性率。單獨看無傷大雅；一次做 m 個，災難就來了。假設 20 個虛無假設**全部為真**（沒有任何真效應）且檢定彼此獨立：

$$\mathbb{P}(\text{至少一個「顯著」}) = 1 - 0.95^{20} = 0.64.$$

將近三分之二的機率，你會「發現」至少一個純屬虛構的效應；平均而言，每 20 個檢定送你 $20 \times 0.05 = 1$ 個偽陽性。基因體學一次檢定兩萬個基因、A/B 平台同時跑數百個實驗、ML 研究者對同一個測試集嘗試幾十組超參數——不做校正的話，「顯著發現」是可以量產的。

核心觀念

多重檢定必須校正。兩條主流路線：

- **Bonferroni 校正**：做 m 個檢定，把每個檢定的門檻改成 α/m。由聯集上界（第 1 章的排容原理放寬成不等式），「至少一個偽陽性」的機率（family-wise error rate, FWER）$\le m \times \alpha/m = \alpha$，**不需要任何獨立性假設**。簡單保險，但 m 大時門檻嚴苛、檢定力大失。
- **偽發現率**（false discovery rate, FDR）：改為控制「被宣告顯著者當中，偽陽性的期望比例」（Benjamini–Hochberg 程序）。容忍少量偽陽性以換取檢定力，是基因體學、大型 A/B 平台等大規模檢定的實務標準。

常見陷阱

***p*-hacking（*p* 值操縱）**：不斷嘗試不同的資料子集、變數轉換、模型設定，直到某個 $p < 0.05$ 才停手，並且**只**報告那一個。名義上的 0.05 早已失守——你其實做了幾十次沒有申報的檢定。這是科學重現性危機的核心成因之一。防範之道：事先登錄假設與分析計畫、對嘗試次數誠實校正、把最終評估留給一份從頭到尾沒碰過的保留資料。ML 的翻版：對著同一個測試集反覆調參，等於把測試集用成了驗證集，報出來的成績注定過度樂觀——第 7 章的過度配適，換個舞台重演。

8.5 重抽法：自助法（bootstrap）

均值有現成的標準誤公式 $\sigma/\sqrt{n}$——因為均值夠簡單。中位數呢？相關係數呢？模型的 AUC 呢？這些統計量的抽樣分布沒有好用的公式。Efron 於 1979 年提出的**自助法**（bootstrap）給出近乎萬用的答案：**沒有公式，就用重抽模擬。**

定義 8.8　自助法

設資料為 $x_1, \dots, x_n$，關心統計量 $\hat{\theta} = g(x_1, \dots, x_n)$。自助法的程序：

1. 從原資料**有放回**地抽 n 筆，得到一個 **bootstrap 樣本**（有些點重複出現、有些缺席）；

2. 對它重算統計量，得到一個 **bootstrap 複本** $\hat{\theta}^*$；
3. 重複 B 次（慣例 $B = 1000$ 到 10000），得 $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$；
4. 複本的分布近似 $\hat{\theta}$ 的抽樣分布：複本的標準差就是 **bootstrap 標準誤**；複本的第 2.5 與 97.5 百分位，就是 95% 的**百分位法信賴區間**。

例 8.8　手動走一遍：中位數的 bootstrap

量測 7 次 API 反應時間（ms）：$x = (12, 15, 9, 21, 14, 32, 11)$，排序後為 $9, 11, 12, 14, 15, 21, 32$，樣本中位數 $\hat{\theta} = 14$。中位數沒有簡單的 SE 公式，用 bootstrap。每次從索引 $\{1, \ldots, 7\}$ 有放回抽 7 個，前三個複本如下：

b	抽到的索引	排序後的 bootstrap 樣本	中位數 $\hat{\theta}^*_b$
1	$(2, 3, 3, 6, 1, 7, 5)$	9, 9, 11, 12, 14, 15, 32	12
2	$(4, 4, 1, 5, 2, 6, 3)$	9, 12, 14, 15, 21, 21, 32	15
3	$(7, 5, 5, 1, 3, 2, 4)$	9, 11, 12, 14, 14, 15, 21	14

注意每個複本裡有些觀測值出現兩次、有些完全沒被抽到——這正是「有放回」的效果。把過程重複 $B = 10000$ 次（程式見第 8.6 節），複本的標準差給出 bootstrap 標準誤約 3.2；取第 2.5 與 97.5 百分位，得 95% 信賴區間約 $[11,\ 21]$。區間明顯不對稱（14 離 11 近、離 21 遠）：右側的 32 是離群值，中位數往上晃的空間比往下大——bootstrap 自動捕捉了這種不對稱，公式法很難做到。

直觀理解

自助法的核心是**替身原理**（plug-in principle）。理想上我們想「從母體重抽很多組資料」來看估計量怎麼晃動，但母體拿不到、實驗只做得起一次。手上的樣本是母體的最佳替身——於是「從樣本中重抽樣本」來模擬「重做實驗」。**有放回**是靈魂：它讓每個 bootstrap 樣本各不相同，重現了抽樣的隨機性。代價是樣本必須能代表母體：n 太小、替身失真，bootstrap 跟著失真；極端統計量（如最大值：複本的最大值永遠不會超過原樣本的最大值）與強相依資料（如時間序列，需改用區塊 bootstrap）也會讓樸素版失效。

這如何驅動機器學習／ AI

本章工具正是 ML 實務者「避免自欺」的整套方法論。**(1) 誠實的模型比較**：例 8.4 的教訓——91.2% 對 90.8% 在 $n = 2000$ 下信賴區間跨零，表面的贏家很可能只是運氣，所以嚴肅的 benchmark 與排行榜要求誤差棒與顯著性檢定；論文報告多個隨機種子的平均與區間，正是在估計「訓練隨機性」造成的抽樣分布。**(2) A/B 測試就是假設檢定**：新版模型上不上線，由雙比例 z 檢定之類的程序把關（練習 8.11），科技公司每天同時跑數百個——多重檢定校正不是學術潔癖，是日常。**(3) bootstrap 無所不在**：對 AUC、F1 這類沒有公

式的指標給信賴區間；而 bagging（bootstrap aggregating）與隨機森林更直接把 bootstrap 當引擎——重抽資料訓練多個模型再平均，用「模擬的重複實驗」壓低變異。**(4) 檢定力分析決定實驗規模**：上線 A/B 前先算「要偵測預期效應需要多少流量」（例 8.5(c)），避免跑一場注定看不見答案的實驗。**(5) 重現性危機**：「刷 SOTA」文化——同一測試集反覆調參、只報最好的種子、測幾十個變體挑最好的發表——本質是未校正的多重比較加上選擇效應，會系統性膨脹表面收益。誤差棒、保留測試集、多種子報告，是這門手藝的職業道德。

8.6 Python 實作

下面把本章四件主力工具各跑一遍：均值的信賴區間、雙樣本 t 檢定、兩模型準確率差異的區間，以及中位數的 bootstrap。

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

rng = np.random.default_rng(42)

# --- 1. 均值的 95% 信賴區間（例 8.1 的數字）---
n, xbar, s = 100, 12.3, 4.0
se = s / np.sqrt(n)                          # 標準誤 = 0.4
z = stats.norm.ppf(0.975)                    # 1.95996...，即慣用的 1.96
print(xbar - z * se, xbar + z * se)          # -> 11.516, 13.084

# --- 2. 雙樣本 t 檢定：A、B 兩組均值是否相同 ---
a = rng.normal(5.0, 1.0, size=40)            # A 組（真均值 5.0）
b = rng.normal(5.6, 1.0, size=40)            # B 組（真均值 5.6）
t_stat, p = stats.ttest_ind(a, b)
print(t_stat, p)                             # -> t = -3.30, p = 0.0015
# p < 0.05：拒絕「兩組均值相同」；方向看 t 的正負

# --- 3. 兩模型準確率差異的 95% CI（例 8.4）---
n_test = 2000
pa, pb = 0.912, 0.908
d = pa - pb
se_d = np.sqrt(pa*(1-pa)/n_test + pb*(1-pb)/n_test)
print(d - 1.96*se_d, d + 1.96*se_d)          # -> -0.0137, 0.0217（跨零！）

# --- 4. 中位數的 95% bootstrap CI（沒有公式可用）---
xdat = np.array([12, 15, 9, 21, 14, 32, 11])
B = 10_000
meds = np.array([np.median(rng.choice(xdat, size=xdat.size))
                 for _ in range(B)])         # choice 預設「有放回」
```

```
print(np.median(xdat))                    # 樣本中位數 14.0
print(np.percentile(meds, [2.5, 97.5]))   # -> [11., 21.]
print(meds.std(ddof=1))                   # bootstrap SE，約 3.2
```

輸出解讀：第 2 段 $p = 0.0015$ 遠小於 0.05——兩組真均值確實相差 0.6，且 $n = 40$ 的檢定力足以看見它。第 3 段重現例 8.4：區間跨零，0.4 個百分點的「提升」不可宣稱。第 4 段的 bootstrap 區間 $[11, 21]$ 與例 8.8 一致；把 `B` 改成 1000 或 100000 重跑幾次，可以觀察區間端點隨 B 增大而趨於穩定。

本章總結

- 估計量有**抽樣分布**；**標準誤** $\mathrm{SE}(\bar{X}) = \sigma/\sqrt{n}$（以 $s/\sqrt{n}$ 估計）是估計與真值的典型距離，且只以 $1/\sqrt{n}$ 縮小——精度加倍，資料四倍。
- **信賴區間** $\bar{X} \pm z_{\alpha/2}\, s/\sqrt{n}$ 的「95%」指的是**程序的長期覆蓋率**，不是「這條區間含真值的機率」；想對參數做機率陳述，要用第 9 章的貝氏可信區間。
- **假設檢定**讓資料審判 H_0（無效應）；p **值** $= \mathbb{P}(\text{資料至少這麼極端} \mid H_0)$，**不是** $\mathbb{P}(H_0 \mid \text{資料})$——反轉方向需要貝氏定理與先驗。不拒絕 $\neq$ 接受。
- 兩種風險：**型一錯誤**（α，偽陽性）與**型二錯誤**（β，偽陰性）此消彼長；**檢定力** $1-\beta$ 隨樣本數與效應大小增加——實驗前先做檢定力分析與樣本數規劃。
- 工具箱：z 檢定（大樣本或 σ 已知）、t 檢定（小樣本，厚尾替估計 σ 的不確定性買單）、卡方檢定（類別計數）、ANOVA（多組均值）。
- **多重檢定**不校正，機運自己會製造顯著（20 個檢定約 64% 出現偽陽性）：Bonferroni 用 α/m 控 FWER，FDR 控偽發現比例；隨時提防 p-hacking。
- **自助法**：有放回重抽 B 次、重算統計量，複本分布近似抽樣分布——沒有公式也能算 SE 與 CI，並且是 bagging / 隨機森林的引擎。
- ML 實戰：benchmark 要誤差棒、A/B 測試就是檢定、指標用 bootstrap 給區間、實驗規模靠檢定力分析——這是誠實報告結果的基本功。

8.7 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 8.1 (基礎)　測試 $n = 64$ 顆電池，樣本平均壽命 $\bar{x} = 305$ 小時、樣本標準差 $s = 24$ 小時。(a) 求均值壽命的 95% 信賴區間。(b) 求 99% 信賴區間。(c) 解釋 99% 區間為何較寬；若想把 95% 區間的半寬縮成一半，樣本數要增為多少？

練習 8.2 (基礎)　某研究（$n = 50$）測得某藥物的平均血壓降幅，95% 信賴區間為 $[4.2,\ 6.8]$ mmHg。判斷下列敘述何者正確、何者錯誤，並說明理由：(a)「真均值 μ 有 95% 的機率落在 4.2 到 6.8 之間。」(b)「若重複整個研究非常多次、每次都計算 95% 信賴區間，長期約有 95% 的區間會蓋住 μ。」(c)「95% 的病人血壓降幅落在 4.2 到 6.8 之間。」(d)「若再做一次同樣的研究，新的樣本平均有 95% 的機率落在 4.2 到 6.8 之間。」

練習 8.3 (基礎)　民調訪問 $n = 900$ 人，468 人支持某政策。(a) 求樣本支持率與其標準誤。(b) 求 95% 信賴區間。(c) 能否宣稱「支持率過半」？(d) 若要求 95% 信賴區間的半寬不超過 2 個百分點（以最保守的 $p = 0.5$ 規劃），至少要訪問多少人？

練習 8.4 (進階)　前導研究顯示某測量的標準差約 $\sigma = 15$。(a) 要讓均值的 95% 信賴區間半寬不超過 2，樣本數至少多少？(b) 半寬不超過 1 呢？(c) 由 (a)(b) 歸納「精度與樣本數」的一般規律，並指出它與 $1/\sqrt{n}$ 的關係。

練習 8.5 (基礎)　飲料標示每瓶 330 ml，長期製程標準差 $\sigma = 4$ ml。抽 $n = 25$ 瓶，得 $\bar{x} = 328.2$ ml。在 $\alpha = 0.05$ 下做雙尾 z 檢定。(a) 寫出 H_0、H_1 並計算檢定統計量。(b) 計算 p 值並下結論。(c) 用 95% 信賴區間交叉驗證你的結論。

練習 8.6 (進階)　新教學法宣稱能提升成績。$n = 16$ 位學生的前後測分數差平均 $\bar{x} = 3.5$ 分、$s = 6.0$ 分。(a) 在 $\alpha = 0.05$ 下做單尾 t 檢定（$H_1\colon \mu > 0$），寫出統計量、自由度、臨界值與結論。(b) 用 t 表把 p 值夾在兩個常用分位之間，並給出軟體算出的精確值。(c) 若誤用 z 檢定，p 值會變多少？說明差異的來源。

練習 8.7 (進階)　某雙尾 z 檢定得到 $p = 0.03$。判斷下列敘述何者正確、何者錯誤，並說明理由：(a)「H_0 為真的機率是 3%。」(b)「假設 H_0 為真，觀察到至少與本次資料一樣極端的結果的機率是 3%。」(c)「若重做一次實驗，有 97% 的機率會再次得到顯著結果。」(d)「p 這麼小，表示效應在實務上很大、很重要。」(e)「在 $\alpha = 0.05$ 下應拒絕 H_0；在 $\alpha = 0.01$ 下不應拒絕。」

練習 8.8 (進階)　垃圾郵件過濾器對每封信做一次「檢定」，H_0：「這是正常郵件」。(a) 描述此情境的型一錯誤與型二錯誤各是什麼，後果為何。(b) 對多數使用者而言哪種錯誤較嚴重？這對 α 的選擇有何含意？(c) 在模型與資料固定下，把 α 調低會對 β 造成什麼影響？為什麼？(d) 若想同時降低兩種錯誤，該怎麼辦？

練習 8.9 (進階)　新產品有四種口味，隨機訪問 $n = 200$ 位消費者的偏好：A 口味 62 人、B 口味 41 人、C 口味 55 人、D 口味 42 人。在 $\alpha = 0.05$ 下檢定「四種口味受歡迎程度相同」。(a) 寫出期望次數。(b) 計算卡方統計量。(c) 給出自由度與臨界值，下結論。(d) 已知

$p \approx 0.099$：若 $\alpha = 0.10$ 結論會如何？這說明了什麼？

練習 8.10 (進階)　你一次做 $m = 20$ 個檢定，所有虛無假設其實都為真，每個檢定的顯著水準為 $\alpha = 0.05$。(a) 若檢定彼此獨立，「至少出現一個偽陽性」的機率是多少？平均會出現幾個偽陽性？(b) 寫出 Bonferroni 校正後的單一檢定門檻，並用聯集上界證明校正後 FWER ≤ 0.05（不需獨立性）。(c) 在獨立假設下計算校正後的精確 FWER。(d) Bonferroni 的代價是什麼？FDR 又是如何折衷的？

練習 8.11 (挑戰)　A/B 測試：舊版頁面 $n_A = 4000$ 人中 200 人轉換（$\hat{p}_A = 5.0\%$），新版 $n_B = 4000$ 人中 250 人轉換（$\hat{p}_B = 6.25\%$）。(a) 用合併比例（pooled proportion）做雙尾 z 檢定（$\alpha = 0.05$）：寫出合併比例、標準誤、z 與 p 值，下結論。(b) 求轉換率差異的 95% 信賴區間（用未合併的標準誤），並確認與 (a) 一致。(c) 從商業角度解讀：相對提升多少？區間的兩端各代表什麼情境？

練習 8.12 (進階)　關於自助法：(a) 寫出 bootstrap 的四個步驟，並解釋為何必須「有放回」抽樣、每個 bootstrap 樣本的大小為何必須是 n。(b) 資料為 $\{3, 7, 8, 12, 15\}$（中位數 8）。以索引 $(2, 5, 1, 2, 4)$ 與 $(4, 4, 3, 5, 1)$ 產生兩個 bootstrap 樣本，各算中位數複本。(c) 說明百分位法如何從 B 個複本得到 95% 信賴區間。(d) 舉出一個 bootstrap 會失效的情境並說明原因。

練習 8.13 (挑戰·程式)　半導體廠比較新舊製程的晶圓良率（%）。用 NumPy 模擬：舊製程 30 片抽自 $N(72, 8^2)$、新製程 30 片抽自 $N(76, 8^2)$。(a) 用 `scipy.stats.ttest_ind` 做雙樣本 t 檢定。(b) 手寫 bootstrap（$B = 10000$）估計「新舊均值差」的 95% 信賴區間。(c) 檢查兩種方法的結論是否一致，並解釋為何樣本均值差可能與真實差異（4 個百分點）相去甚遠。

8.8 練習題詳解

練習 8.1 詳解　標準誤 $\widehat{\mathrm{SE}} = s/\sqrt{n} = 24/\sqrt{64} = 3$。
(a) 95%：$305 \pm 1.96 \times 3 = 305 \pm 5.88$，即 $[299.1,\ 310.9]$ 小時。
(b) 99%：$305 \pm 2.576 \times 3 = 305 \pm 7.73$，即 $[297.3,\ 312.7]$ 小時。
(c) 更高的信心要求更寬的網：臨界值從 1.96 升到 2.576，區間隨之變寬——同一份資料下，「更確定蓋住真值」只能靠「說得更保守」換來。半寬 $= z\,s/\sqrt{n}$ 與 $\sqrt{n}$ 成反比，要把半寬 5.88 縮成 2.94，需 $\sqrt{n}$ 加倍，即 $n = 4 \times 64 = 256$ 顆。

練習 8.2 詳解　(a) **錯**。頻率觀點中 μ 是固定常數，不是隨機變數；這條算出來的區間要嘛蓋住 μ 要嘛沒蓋住，只是我們不知道是哪一種。「95%」屬於產生區間的程序，不屬於這條

特定區間。
(b) **對**。這正是覆蓋率的定義——信賴區間唯一合法的機率詮釋。
(c) **錯**。信賴區間描述的是**均值**的不確定性，不是個別資料的散布。個別病人的降幅散布由標準差（約 s）描述，遠比均值的區間寬；描述「單一新觀測值會落在哪」需要的是預測區間。
(d) **錯**。舊區間與新樣本平均都是隨機的。設母體標準差為 σ，則 $\bar{X}_{\text{新}} - \bar{X}_{\text{舊}} \sim N(0,\, 2\sigma^2/n)$，故

$$\mathbb{P}\big(|\bar{X}_{\text{新}} - \bar{X}_{\text{舊}}| \le 1.96\,\sigma/\sqrt{n}\big) = \mathbb{P}\big(|Z| \le 1.96/\sqrt{2}\big) = \mathbb{P}(|Z| \le 1.386) \approx 0.83,$$

只有約 83%，不是 95%。

練習 8.3 詳解　(a) $\hat{p} = 468/900 = 0.52$，$\widehat{\text{SE}} = \sqrt{0.52 \times 0.48/900} = 0.0167$。
(b) $0.52 \pm 1.96 \times 0.0167 = 0.52 \pm 0.033$，即 $[0.487,\ 0.553]$。
(c) 不能。區間包含 0.5，資料與「支持率恰為一半、甚至略低」相容——證據不足以排除未過半。
(d) 要求 $1.96\sqrt{0.5 \times 0.5/n} \le 0.02$，即 $\sqrt{n} \ge 1.96 \times 0.5/0.02 = 49$，得 $n \ge 2401$。半寬從 3.3 個百分點壓到 2 個百分點，受訪人數要從 900 增到 2401——精度的代價是平方級的。

練習 8.4 詳解　半寬條件 $z_{\alpha/2}\,\sigma/\sqrt{n} \le h$ 解出 $n \ge (z_{\alpha/2}\,\sigma/h)^2$。
(a) $n \ge (1.96 \times 15/2)^2 = 14.7^2 = 216.1$，取 $n = 217$。
(b) $n \ge (1.96 \times 15/1)^2 = 29.4^2 = 864.4$，取 $n = 865$。
(c) 半寬減半、樣本數變 4 倍（$865 \approx 4 \times 217$，差在進位）。一般規律：$n \propto 1/h^2$。根源是 $\text{SE} \propto 1/\sqrt{n}$——精度是用樣本數的**平方**買的，這就是「多一位小數要 100 倍資料」的出處。

練習 8.5 詳解　(a) $H_0\colon \mu = 330$ 對 $H_1\colon \mu \ne 330$。標準誤 $\text{SE} = 4/\sqrt{25} = 0.8$，

$$z = \frac{328.2 - 330}{0.8} = -2.25.$$

(b) 雙尾 $p = 2\,\mathbb{P}(Z \le -2.25) = 2 \times 0.0122 = 0.024 < 0.05$，拒絕 H_0：平均填充量顯著低於標示值（低了 2.25 個標準誤）。
(c) 95% CI：$328.2 \pm 1.96 \times 0.8 = [326.6,\ 329.8]$，不含 330——與 (b) 的拒絕一致（區間與雙尾檢定的對偶性）。

練習 8.6 詳解　(a) $H_0\colon \mu = 0$ 對 $H_1\colon \mu > 0$。$\widehat{\text{SE}} = 6/\sqrt{16} = 1.5$，

$$t = \frac{3.5}{1.5} = 2.33, \qquad \text{自由度} 15.$$

單尾臨界值 $t_{0.05,15} = 1.753$；$2.33 > 1.753$，拒絕 H_0：有證據顯示成績提升。
(b) 由 t 表：$t_{0.025,15} = 2.131 < 2.33 < t_{0.01,15} = 2.602$，故 $0.01 < p < 0.025$；軟體給出精確值 $p = 0.017$。

(c) 誤用 z：$p = \mathbb{P}(Z > 2.33) = 0.0098$，約為正確值的 0.6 倍。差異來自 s 估計 σ 的額外不確定性：$n = 16$ 時 s 本身晃動不小，t_{15} 用較厚的尾巴替這件事買單；z 假裝 σ 精確已知，因此高估了證據強度。

練習 8.7 詳解　(a) **錯**。p 值以「H_0 為真」為前提計算，邏輯上不可能反過來告訴你這個前提的機率；$\mathbb{P}(H_0 \mid \text{資料})$ 需要貝氏定理與先驗。
(b) **對**。這正是 p 值的定義。
(c) **錯**。重現機率取決於真實效應大小與檢定力，與 $1 - p$ 無關；若真效應恰在檢定力 50% 的邊緣，重現機率大約只有一半。
(d) **錯**。p 值混合了效應大小與樣本數：n 夠大時，微不足道的效應也能得到極小的 p。統計顯著 $\neq$ 實務重要，效應大小要看估計值與信賴區間。
(e) **對**。$0.01 < p = 0.03 < 0.05$：在 $\alpha = 0.05$ 下顯著、在 $\alpha = 0.01$ 下不顯著。同一份資料、不同門檻、不同結論——這正是報告 p 值本身（而非只報「顯著」）的價值。

練習 8.8 詳解　(a) 型一錯誤：正常郵件被判為垃圾（誤攔）——重要信件（面試通知、帳單）消失在垃圾匣，使用者甚至不會發現。型二錯誤：垃圾郵件被放行——收件匣多幾封廣告。
(b) 多數使用者認為型一較嚴重：漏掉一封錄取通知的代價，遠大於多刪幾封廣告。因此過濾器把 α（誤攔率）壓得很低，寧可放行可疑郵件。
(c) α 調低等於把「判為垃圾」的門檻提高：更難拒絕 H_0，垃圾郵件更容易通過，β 上升。固定資訊量下兩種錯誤此消彼長，這是決策理論的鐵律。
(d) 增加資訊：更多訓練資料、更好的特徵（寄件者信譽、連結分析）、更強的模型——相當於把兩個類別的「分布」拉得更開，兩種錯誤才能同時下降。

練習 8.9 詳解　(a) 均勻之下 $E_i = 200/4 = 50$（每種口味）。
(b)

$$\chi^2 = \frac{(62-50)^2}{50} + \frac{(41-50)^2}{50} + \frac{(55-50)^2}{50} + \frac{(42-50)^2}{50} = 2.88 + 1.62 + 0.50 + 1.28 = 6.28.$$

(c) 自由度 $k - 1 = 3$，臨界值 $\chi^2_{0.05,3} = 7.815$。$6.28 < 7.815$，不拒絕 H_0：證據不足以說偏好不均。
(d) $\alpha = 0.10$ 的臨界值為 $\chi^2_{0.10,3} = 6.251 < 6.28$，會拒絕。同一份資料在不同 α 下結論翻轉，說明 6.28 落在「灰色地帶」——此時直接報告 $p \approx 0.099$ 讓讀者自行判斷，比只丟出「顯著 / 不顯著」誠實得多；更謹慎的做法是蒐集更多資料再說。

練習 8.10 詳解　(a) 獨立時

$$\mathbb{P}(\text{至少一個偽陽性}) = 1 - (1 - 0.05)^{20} = 1 - 0.358 = 0.64.$$

偽陽性個數的期望值 $= 20 \times 0.05 = 1$ 個（期望值可加，不需獨立性）。
(b) Bonferroni：每個檢定改用 $\alpha' = 0.05/20 = 0.0025$。設 R_i 為「第 i 個檢定偽陽性」，由聯集上界

$$\mathbb{P}\Big(\bigcup_{i=1}^{20} R_i\Big) \le \sum_{i=1}^{20} \mathbb{P}(R_i) = 20 \times 0.0025 = 0.05,$$

不論檢定之間相關與否，FWER 都被壓在 0.05 以下。
(c) 獨立時精確值 $= 1 - (1 - 0.0025)^{20} = 1 - 0.9975^{20} = 0.049 \le 0.05$，與上界幾乎貼合。
(d) 代價：門檻從 0.05 掉到 0.0025，個別檢定的檢定力大幅下降，真效應容易被漏掉（型二錯誤暴增）；m 上千時幾乎測不到任何東西。FDR（Benjamini–Hochberg）改而控制「被宣告顯著者中偽陽性的期望比例」，容忍少量偽陽性換回檢定力，是大規模檢定的實務折衷。

練習 8.11 詳解　(a) $H_0\colon p_A = p_B$。合併比例 $\hat{p} = (200 + 250)/8000 = 0.05625$，

$$\mathrm{SE} = \sqrt{\hat{p}(1-\hat{p})\Big(\frac{1}{4000} + \frac{1}{4000}\Big)} = \sqrt{0.05625 \times 0.94375 \times 0.0005} = 0.00515,$$

$$z = \frac{0.0625 - 0.05}{0.00515} = 2.43, \qquad p = 2\,\mathbb{P}(Z > 2.43) = 0.015 < 0.05.$$

拒絕 H_0：新版轉換率顯著較高。
(b) 未合併標準誤

$$\mathrm{SE} = \sqrt{\frac{0.05 \times 0.95}{4000} + \frac{0.0625 \times 0.9375}{4000}} = 0.00515,$$

差異的 95% CI：$0.0125 \pm 1.96 \times 0.00515 = [0.0024,\ 0.0226]$。區間不含 0，與 (a) 的拒絕一致。
(c) 相對提升 $0.0125/0.05 = 25\%$。但區間告訴我們：真實提升可能低至 0.24 個百分點（相對約 5%，可能不值得上線成本），也可能高至 2.26 個百分點（相對約 45%，大勝）。上線決策應對照整條區間與改版成本，而不是只看「顯著」二字。

練習 8.12 詳解　(a) 步驟：(1) 從 n 筆原資料**有放回**抽 n 筆成 bootstrap 樣本；(2) 重算統計量得複本 $\hat{\theta}^*$；(3) 重複 B 次；(4) 用複本的標準差當標準誤、複本的百分位當信賴區間。「有放回」是為了讓每次重抽結果不同，模擬抽樣的隨機性——若不放回，每個「重抽樣本」都恰是原資料，毫無變化。樣本大小取 n，是因為估計量的不確定性依賴樣本數（$\mathrm{SE} \propto 1/\sqrt{n}$）：用別的大小去重抽，模擬到的是「別的樣本數下」的不確定性。
(b) 索引 $(2,5,1,2,4)$ 取出 $(7,15,3,7,12)$，排序 $3,7,7,12,15$，中位數 $\hat{\theta}^*_1 = 7$。索引 $(4,4,3,5,1)$ 取出 $(12,12,8,15,3)$，排序 $3,8,12,12,15$，中位數 $\hat{\theta}^*_2 = 12$。
(c) 把 B 個複本由小到大排序，取第 2.5 百分位為下端點、第 97.5 百分位為上端點（$B = 10000$ 時即第 250 與第 9750 個值）——中間恰包住 95% 的複本。
(d) 例：統計量為樣本**最大值**時失效——複本的最大值永遠不超過原樣本的最大值，重抽根本探索不到「更大的值」，抽樣分布的右尾完全缺失。其他失效情境：n 太小（替身失真）、

時間序列等強相依資料（獨立重抽破壞相依結構，需改用區塊 bootstrap）。

練習 8.13 詳解　參考程式與輸出：

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

rng = np.random.default_rng(0)

# 模擬資料：舊、新製程各 30 片晶圓的良率（%）
old = rng.normal(72.0, 8.0, size=30)
new = rng.normal(76.0, 8.0, size=30)
print(new.mean() - old.mean())            # -> 7.18（樣本均值差）

# (a) 雙樣本 t 檢定
t_stat, p = stats.ttest_ind(new, old)
print(t_stat, p)                          # -> t = 3.905, p = 0.0002

# (b) 均值差的 bootstrap 95% CI
B = 10_000
diffs = np.empty(B)
for i in range(B):
    diffs[i] = (rng.choice(new, new.size).mean()
                - rng.choice(old, old.size).mean())
print(np.percentile(diffs, [2.5, 97.5]))  # -> [3.72, 10.82]
```

(a) $t = 3.905$、$p = 0.0002 < 0.05$：拒絕「兩製程良率相同」。(b) bootstrap 區間 $[3.7,\ 10.8]$。(c) 兩者一致：區間不含 0 $\iff$ 檢定拒絕。注意樣本均值差 7.18 與真實差異 4 相去不小——這正是抽樣波動：SE $\approx \sqrt{8^2/30 + 8^2/30} \approx 2.07$，偏離 1.5 個標準誤稀鬆平常，寬達 7 個百分點的信賴區間誠實地反映了這件事。每次執行 bootstrap 的端點會略有不同，屬正常現象（B 越大越穩定）。

下一章：貝氏統計——把機率當作信念，讓資料負責更新它。