機率與統計：從入門到精通

第 14 章　變異數分析

Analysis of Variance (ANOVA)

難度：進階

比較兩組均值，第 8 章的 t 檢定就夠了。比較三組、五組、十組呢？對每一對都做一次 t 檢定，型一錯誤會像複利一樣膨脹——十組就有 45 對，H_0 全真時「至少出現一次偽陽性」的機率高達九成。**變異數分析**（analysis of variance, ANOVA）用一次 F 檢定回答「這些組的均值是否全部相等」，再在整體顯著之後，用控制家族錯誤率的事後比較找出是哪幾組不同。名字裡有「變異數」，問的卻是均值：把總變異拆成「組與組之間」與「組內雜訊」，看前者是否大到後者解釋不了。這套分解也正是第 10 章線性迴歸 F 檢定的特例——虛擬變數迴歸與單因子 ANOVA 是同一件事的兩種寫法。

學習目標

讀完本章後，你將能夠：

- 說明為何對 k 組做兩兩 t 檢定會膨脹型一錯誤，並寫出單因子 ANOVA 的 H_0、H_1。
- 把總平方和分解為 $\mathrm{SST} = \mathrm{SSB} + \mathrm{SSW}$，由均方計算 F 統計量，並讀懂 ANOVA 表。
- 陳述獨立、常態、變異數同質三條假設，並在違反時改用 Welch 或 Kruskal–Wallis。
- 在整體 F 顯著後，用 Bonferroni 與 Tukey HSD 做**事後比較**，並計算效果量 η^2。
- 執行雙因子 ANOVA，先檢定**交互作用**再決定是否解讀主效果。
- 把單因子 ANOVA 改寫成虛擬變數迴歸，看出它與第 10 章的 F 檢定是同一件事。
- 用 SciPy 的 `f_oneway`、`tukey_hsd` 與 `levene` 完成分析。

14.1　為什麼不能連做很多次 t 檢定

第 8 章比較兩組均值用雙樣本 t 檢定。現在有 k 組——三種教學法、五個推薦演算法、十種超參數組合——直覺是「每兩組做一次 t」。k 組有

$$m = \binom{k}{2} = \frac{k(k-1)}{2}$$

對。每一對都在 $\alpha = 0.05$ 下檢定，且假設各檢定獨立、H_0 全部為真，則

$$\mathbb{P}(\text{至少一次偽陽性}) = 1 - (1-\alpha)^m.$$

$k=3$ 時 $m=3$，這個機率已是 14%；$k=5$ 時 $m=10$，升到 40%；$k=10$ 時 $m=45$，高達 90%。你幾乎**注定**會「發現」至少一對其實不存在的差異。第 8 章把這件事叫做家族錯誤率（family-wise error rate, FWER）膨脹；ANOVA 的第一步，就是用**一次**檢定把 FWER 鎖回 α。

例 14.1　十個模型、四十五次兩兩比較

你訓練了 $k=10$ 個模型變體，在同一測試集上得到十個準確率，想知道「有沒有哪個比較準」。若對每一對做 $\alpha=0.05$ 的 t 檢定，即使十個模型的真正表現完全相同，

$$1-0.95^{45}\approx 0.90$$

——九成機率至少出現一顆星號。論文裡「模型 A 顯著優於模型 B」這句話，若沒有先做整體檢定、也沒有校正多重比較，星號的含金量就要打折。

核心觀念

先問整體、再問細節。ANOVA 分兩層：

1. **整體 F 檢定**：虛無假設 $H_0\colon \mu_1=\mu_2=\cdots=\mu_k$（所有組均值相等）。一次檢定、一個 p 值，FWER 維持在 α。
2. **事後比較**（post hoc）：只有整體拒絕 H_0 之後，才去問「是哪幾對不同」，並用 Tukey 或 Bonferroni 把兩兩比較的 FWER 繼續鎖住。

整體不顯著就停——不要再翻兩兩 p 值找「還是有一對過門檻」。那叫 p-hacking，第 8 章已經警告過。

14.2　單因子 ANOVA：把總變異拆成兩塊

14.2.1　模型

定義 14.1　單因子 ANOVA 模型

設因子（factor）有 k 個水準（level），第 i 組有 n_i 筆獨立觀測，$N=\sum_{i=1}^{k} n_i$。寫

$$Y_{ij}=\mu+\tau_i+\varepsilon_{ij},\qquad i=1,\dots,k,\quad j=1,\dots,n_i,$$

其中 μ 是總均值，τ_i 是第 i 組相對總均值的效應（約束 $\sum_i n_i\tau_i=0$），$\varepsilon_{ij}\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,\sigma^2)$。等價地，$Y_{ij}\sim\mathcal{N}(\mu_i,\sigma^2)$ 且各組獨立、變異數相同，$\mu_i=\mu+\tau_i$。

檢定的假設是

$$H_0\colon \tau_1=\cdots=\tau_k=0\quad\big(\text{即}\mu_1=\cdots=\mu_k\big),\qquad H_1\colon \text{至少有一個}\tau_i\neq 0.$$

注意 H_1 **不**指定是哪一組不同，也不指定方向——F 檢定對任何組間差異模式都敏感，因此雖然只看 F 分布的右尾，本質上是雙向問題（第 8 章對 F 的說明）。

14.2.2 平方和的分解

記第 i 組樣本均為 $\bar{Y}_{i\cdot}$、總平均為 $\bar{Y}_{\cdot\cdot}$。每一筆資料離總平均的偏差，可以拆成「它所在那一組離總平均有多遠」加上「它自己離組平均有多遠」：

$$Y_{ij}-\bar{Y}_{\cdot\cdot}=\underbrace{\left(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot}\right)}_{\text{組間}}+\underbrace{\left(Y_{ij}-\bar{Y}_{i\cdot}\right)}_{\text{組內}}.$$

兩邊平方再對所有 i,j 加總，交叉項消失（因為 $\sum_j(Y_{ij}-\bar{Y}_{i\cdot})=0$），得到 ANOVA 的靈魂恆等式：

定理 14.2　平方和分解

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{\cdot\cdot}\right)^2}_{\text{SST 總平方和}}=\underbrace{\sum_{i=1}^{k}n_i\left(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot}\right)^2}_{\text{SSB 組間平方和}}+\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i\cdot}\right)^2}_{\text{SSW 組內平方和}}.$$

自由度跟著拆：$\mathrm{df}_T=N-1$，$\mathrm{df}_B=k-1$，$\mathrm{df}_W=N-k$，且 $(N-1)=(k-1)+(N-k)$。

證明　令 $a_i=\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot}$、$e_{ij}=Y_{ij}-\bar{Y}_{i\cdot}$，則 $Y_{ij}-\bar{Y}_{\cdot\cdot}=a_i+e_{ij}$。平方加總：

$$\sum_{i,j}(a_i+e_{ij})^2=\sum_{i,j}a_i^2+\sum_{i,j}e_{ij}^2+2\sum_{i}a_i\sum_{j}e_{ij}.$$

內層 $\sum_j e_{ij}=0$，交叉項為 0。又 $\sum_j a_i^2=n_i a_i^2$，即得 $\mathrm{SST}=\mathrm{SSB}+\mathrm{SSW}$。自由度：總平均用掉 1 個約束；組間有 k 個組平均、1 個總平均約束；組內每組用掉 1 個組平均，共 k 個約束。■

直觀理解

把每位學生的成績想成「離全班平均有多遠」。這段距離有兩個來源：**他上的那門課平均就比較高**（組間——教學法的效應），以及**就算同一門課，學生之間本來就有高低**（組內——殘差、雜訊）。ANOVA 問的是：組間那一塊，是否大到組內雜訊解釋不了？名字叫「分析變異數」，真正比較的是均值——均值差得越開，SSB 越大。

14.2.3 F 統計量

均方（mean square）就是平方和除以自由度，形式上像變異數：

$$\mathrm{MSB}=\frac{\mathrm{SSB}}{k-1},\qquad \mathrm{MSW}=\frac{\mathrm{SSW}}{N-k}.$$

MSW 永遠估計 σ^2（組內雜訊），與 H_0 真假無關。MSB 在 H_0 為真時也估計 σ^2；一旦組均值真的不同，它會額外含進 $\sum n_i\tau_i^2$ 那一項，因而偏大。於是它們的比

$$F=\frac{\mathrm{MSB}}{\mathrm{MSW}}$$

在 H_0 下應靠近 1；遠大於 1 就是「組間差異大過雜訊」的證據。

定理 14.3　H_0 下 F 的抽樣分布

在定義 14.1 的假設下，若 H_0 為真，則

$$\frac{\mathrm{SSB}}{\sigma^2} \sim \chi^2_{k-1}, \qquad \frac{\mathrm{SSW}}{\sigma^2} \sim \chi^2_{N-k},$$

兩者獨立，因而

$$F = \frac{\mathrm{MSB}}{\mathrm{MSW}} \sim F_{k-1,\,N-k}.$$

拒絕域為右尾：$F \geq F_{\alpha;\,k-1,\,N-k}$，或等價地 $p = \mathbb{P}(F_{k-1,N-k} \geq F_{\mathrm{obs}})$ 小於 α。

F 分布由兩個自由度塑造：分子 $\mathrm{df}_1 = k-1$ 小、分母 $\mathrm{df}_2 = N-k$ 大時，密度在 1 附近堆得較尖（圖 14.1）。它只取正值、右偏，期望值（當 $\mathrm{df}_2 > 2$）為 $\mathrm{df}_2/(\mathrm{df}_2-2)$，略大於 1——所以「$F \approx 1$」是 H_0 下的基準，不是「$F \approx 0$」。

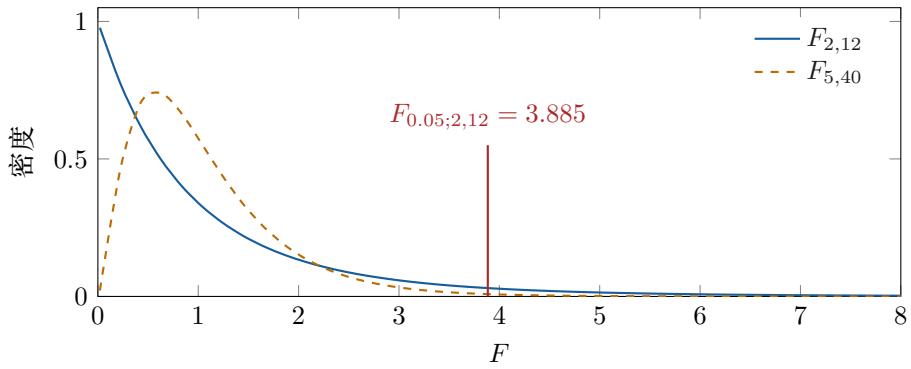


圖 14.1：F 分布右偏、只取正值。自由度越大，密度越集中在 1 附近。單因子 ANOVA 的拒絕域是右尾：觀測 F 若落在臨界值右側，拒絕「各組均值相等」。

命題 14.4　$k=2$ 時 ANOVA 就是 t 檢定

兩組、變異數相同的雙樣本 t 檢定滿足

$$F_{1,\,N-2} = t^2_{N-2}.$$

兩個 p 值完全相同。ANOVA 是 t 檢定往 $k \geq 3$ 的唯一自然延伸。

證明（等組人數 $n_1 = n_2 = n$）　總平均 $\bar{Y}_{\cdot\cdot} = (\bar{Y}_1 + \bar{Y}_2)/2$，故 $\bar{Y}_1 - \bar{Y}_{\cdot\cdot} = (\bar{Y}_1 - \bar{Y}_2)/2$。於是

$$\mathrm{SSB} = n \cdot 2 \cdot \left(\frac{\bar{Y}_1 - \bar{Y}_2}{2}\right)^2 = \frac{n}{2}(\bar{Y}_1 - \bar{Y}_2)^2,$$

$\mathrm{MSB} = \mathrm{SSB}$（因為 $k-1=1$）。合併變異數 $s_p^2 = \mathrm{MSW}$，而

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_p^2(1/n + 1/n)}} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{2s_p^2/n}},$$

平方即 $t^2 = \mathrm{MSB}/\mathrm{MSW} = F$。臨界值也吻合：$t^2_{0.025,\,\nu} = F_{0.05;\,1,\,\nu}$（例如 $\nu = 8$ 時 $2.306^2 = 5.318$）。 ■

14.3　完整例題：三種教學法

十五位學生隨機分到三種教學法，期末分數如下。這組數字足夠小，可以手算每一個平方；請務必自己走一遍，後面的事後比較、效果量、虛擬變數迴歸都建立在同一張表上。

教學法		分數					組平均
A	講授	82	86	79	83	90	$\bar{y}_A = 84$
B	影片	75	78	72	80	70	$\bar{y}_B = 75$
C	練習	88	91	85	89	92	$\bar{y}_C = 89$

總平均 $\bar{y}_{\cdot\cdot} = (84 + 75 + 89)/3 = 82.667$（或 $1240/15$）。圖 14.2 把十五個點畫出來：組與組之間的水平位移是 SSB 的來源，各組內部的垂直散開是 SSW 的來源。

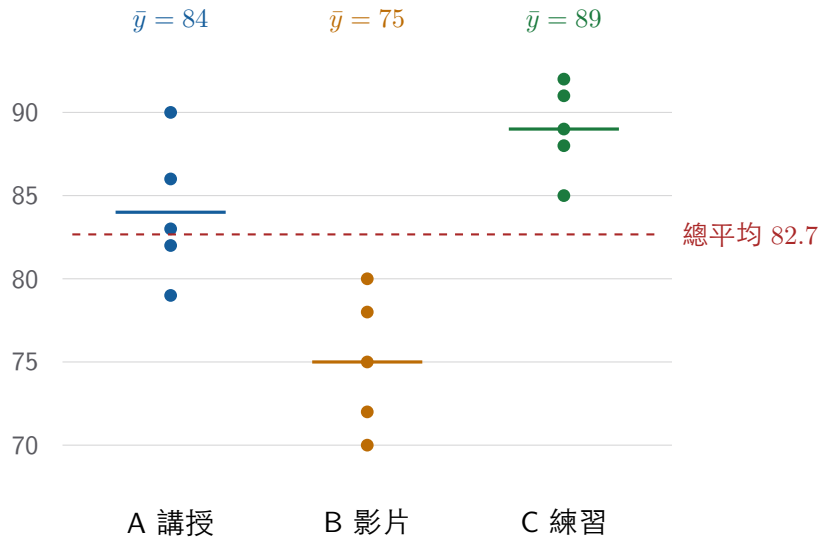


圖 14.2: 三種教學法的原始分數（點）與組平均（橫線）。虛線是總平均。組平均離虛線越遠，SSB 越大；點離自己那條橫線越遠，SSW 越大。

例 14.2　手算一張完整的 ANOVA 表

解　組內平方和——每組對自己的平均：

$$A: (-2)^2 + 2^2 + (-5)^2 + (-1)^2 + 6^2 = 70,$$
$$B: 0^2 + 3^2 + (-3)^2 + 5^2 + (-5)^2 = 68,$$
$$C: (-1)^2 + 2^2 + (-4)^2 + 0^2 + 3^2 = 30,$$

故 $\text{SSW} = 70 + 68 + 30 = 168$。組間平方和——組平均對總平均，再乘組人數：

$$\text{SSB} = 5\big[(84 - 82.667)^2 + (75 - 82.667)^2 + (89 - 82.667)^2\big] = 503.33.$$

核對：$\text{SST} = 168 + 503.33 = 671.33$。均方與 F：

$$\text{MSB} = \frac{503.33}{2} = 251.67, \qquad \text{MSW} = \frac{168}{12} = 14, \qquad F = \frac{251.67}{14} = 17.98.$$

臨界值 $F_{0.05;\,2,12} = 3.885$。觀測 $F = 17.98$ 遠超過；$p = \mathbb{P}(F_{2,12} \geq 17.98) = 0.00025$。拒絕 H_0：三種教學法的均值並非全部相等。

來源	SS	df	MS	F	p
組間（教學法）	503.33	2	251.67	17.98	0.00025
組內（誤差）	168.00	12	14.00		
總計	671.33	14			

這張表就是所有統計軟體 ANOVA 輸出的骨架：來源、平方和、自由度、均方、F、p。學會手算一次，之後看軟體輸出只是同一張表的排版。

定義 14.5　效果量 η^2 與 ω^2

$$\eta^2 = \frac{\mathrm{SSB}}{\mathrm{SST}}, \qquad \omega^2 = \frac{\mathrm{SSB} - (k-1)\,\mathrm{MSW}}{\mathrm{SST} + \mathrm{MSW}}.$$

η^2 是「總變異裡有多大比例可被組別解釋」，等價於把組別當預測變數的迴歸 R^2（第 10 章）。它略偏高；ω^2 做了偏誤修正。Cohen 的粗略標尺：$\eta^2 \approx 0.01$ 小、0.06 中、0.14 大——但標尺不能取代領域判斷。

本例 $\eta^2 = 503.33/671.33 = 0.750$，$\omega^2 = 0.694$：約四分之三的分數變異來自教學法，效應非常大。$p$ 值只告訴你「不是純雜訊」，η^2 才告訴你「大到有沒有實務意義」。樣本一大，p 可以很小而 η^2 接近 0——練習 14.11 專門練這件事。

14.4　假設、穩健性與非參數替代

單因子 ANOVA 的 F 精確服從 $F_{k-1,N-k}$，依賴三條假設：

1. **獨立**：不同觀測、不同組之間獨立。同一個人重複測量、時間序列、叢集抽樣都會破壞它——這是最致命的一條，幾乎無法靠大樣本挽救。
2. **常態**：每組的殘差 ε_{ij} 近似常態。F 檢定對輕微偏態相當穩健；嚴重離群值或明顯雙峰時，改用 Kruskal–Wallis 等級檢定。
3. **變異數同質**（homoscedasticity）：$\sigma_1^2 = \cdots = \sigma_k^2$。可用 Levene 檢定；若拒絕且組人數很不均衡，改用 Welch ANOVA（不合併各組變異數，類似不假設等變異的 t 檢定）。

例 14.3　對教學法資料做假設檢查

解　Levene 檢定比較各組 $|Y_{ij} - \bar{Y}_{i\cdot}|$ 的均值，得 $W = 0.44$、$p = 0.66$，不拒絕同質性——三組樣本變異數 17.5、17.0、7.5，在 $n_i = 5$ 下這種差距完全正常。Shapiro–Wilk 對三組分別得 $p = 0.93$、0.83、0.83，沒有常態性警訊。非參數的 Kruskal–Wallis 得 $H = 10.26$、$p = 0.0059$，與參數 F 的結論一致：拒絕「各組分布相同」。小樣本下兩條路都走一遍，是

好習慣。

常見陷阱

獨立性比常態性重要得多。把同一位使用者的十次點擊當成十筆獨立樣本、把同一測試集上十個模型的準確率當成十組獨立實驗，都會讓 p 值樂觀得離譜——因為 MSW 低估了真正的不確定性。ANOVA 的「隨機分派到各組」不是寫在假設清單裡的客套話，是讓獨立性成立的實驗設計。觀察資料（沒有隨機化）頂多能說「組平均有差異」，不能說「因子造成差異」。

14.5　事後比較：是哪幾組不同

整體 F 顯著只告訴你「不是全部相等」。三種教學法有三對：A vs. B、A vs. C、B vs. C。若直接做三次 $\alpha = 0.05$ 的 t 檢定，FWER 變成 $1 - 0.95^3 \approx 0.14$。事後比較的任務是：比較所有想比的對，同時把 FWER 鎖在 α。

核心觀念

兩條最常用的路：

- **Bonferroni**：做 m 次比較，每次用 α/m。由聯集上界，FWER $\le \alpha$，不需獨立性（第 8 章）。簡單、偏保守，m 大時檢定力差。
- **Tukey HSD**（honestly significant difference）：專門為「所有兩兩比較」設計，用學生化全距（studentized range）分布，比 Bonferroni 更有檢定力，且精確控制 FWER。等組人數時

$$\mathrm{HSD} = q_{\alpha;\,k,\,N-k}\sqrt{\frac{\mathrm{MSW}}{n}},$$

  兩組平均差的絕對值超過 HSD 就顯著。組人數不等時改用 Tukey–Kramer，把 n 換成調和形式 $2/(1/n_i + 1/n_j)$。

還有兩種情境不該用「全部兩兩」：若事前只關心「每一種處理 vs. 對照組」，用 Dunnett；若想檢驗任意對比（contrast），用 Scheffe。

例 14.4　教學法的 Bonferroni 與 Tukey

三對、$m = 3$，Bonferroni 門檻 $\alpha/3 = 0.0167$。用合併變異數 MSW $= 14$ 的雙樣本 t 檢定（與 ANOVA 同一誤差項）：

對照	$\lvert\bar{y}_i - \bar{y}_j\rvert$	未校正 p	Bonferroni p	Tukey p
A vs. B	9	0.0090	0.027	0.0066
A vs. C	5	0.056	0.167	0.129
B vs. C	14	0.00023	0.00068	0.00019

解　Tukey 的 $q_{0.05;\,3,12} = 3.773$，

$$\text{HSD} = 3.773 \times \sqrt{14/5} = 6.31.$$

$|\bar{y}_A - \bar{y}_B| = 9 > 6.31$、$|\bar{y}_B - \bar{y}_C| = 14 > 6.31$，顯著；$|\bar{y}_A - \bar{y}_C| = 5 < 6.31$，不顯著。與 Bonferroni 結論相同：影片（B）明顯低於講授與練習；講授與練習之間，這份 $n_i = 5$ 的資料分辨不了。注意 A vs. C 的未校正 $p = 0.056$ 在 $\alpha = 0.05$ 下「差一點」——若你先看了資料再決定只報這一對，就是典型的 p-hacking；校正之後它安分地不顯著。

常見陷阱

整體不顯著，就不要做兩兩比較。有人 ANOVA 的 $p = 0.12$，卻在九對裡找到一對未校正 $p = 0.04$ 然後大書特書。那一對是在 H_0 下純機運也會出現的星號。相反的陷阱：整體顯著也不等於每一對都顯著——本例就是「B 拖後腿，A 與 C 分不出來」。報告時兩層都要寫：整體 F 與事後哪些對不同。

14.6　雙因子 ANOVA 與交互作用

一個因子不夠時，實驗往往同時操作兩個：肥料種類 × 水量、模型架構 × 學習率、藥物 × 性別。**雙因子 ANOVA**（two-way ANOVA）一次回答三個問題：因子 A 有沒有主效果？因子 B 有沒有主效果？兩者有沒有**交互作用**——A 的效果是否取決於 B 處於哪個水準。

定義 14.6　雙因子 ANOVA 模型

因子 A 有 a 個水準、因子 B 有 b 個水準，每個組合（cell）有 n 個重複。寫

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},$$

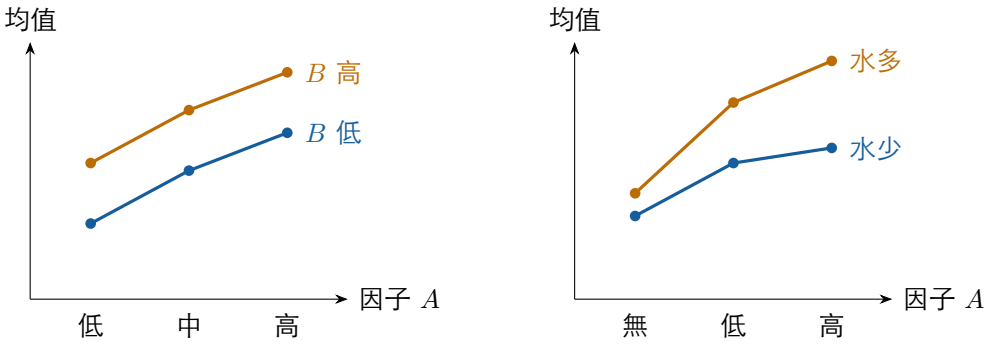
$\varepsilon_{ijk} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$。平方和多拆兩塊：

$$\text{SST} = \text{SS}_A + \text{SS}_B + \text{SS}_{AB} + \text{SSE},$$

自由度分別為 $a-1$、$b-1$、$(a-1)(b-1)$、$ab(n-1)$（總計 $abn-1$）。三個 F：

$$F_A = \frac{\text{MS}_A}{\text{MSE}}, \quad F_B = \frac{\text{MS}_B}{\text{MSE}}, \quad F_{AB} = \frac{\text{MS}_{AB}}{\text{MSE}}.$$

永遠先看交互作用。若 F_{AB} 顯著，主效果是把另一因子平均掉之後的邊際均值——可能沒有任何一個子群真正長那樣。此時應改看**簡單主效果**（simple main effects）：固定 B 的某一水準，再比較 A。若交互作用不顯著，才把 SS_{AB} 視為雜訊，解讀兩個主效果。



（a）平行：無交互作用　　（b）不平行：有交互作用

圖 14.3: 交互作用圖。兩條線平行，表示 A 的效果與 B 無關；兩條線斜率不同（甚至交叉），表示「A 有沒有效」取決於 B——主效果的單一數字會誤導。

例 14.5　肥料 × 水量：交互作用顯著

每格 $n = 3$ 盆植物的產量：

肥料 \ 水	少	多
無	12, 14, 13（均 13）	15, 16, 14（均 15）
低	18, 19, 17（均 18）	24, 26, 25（均 25）
高	20, 21, 19（均 20）	28, 30, 29（均 29）

格子均值畫在圖 14.3(b)：水從「少」加到「多」，無肥只 +2、低肥 +7、高肥 +9——水量的效益隨肥料增加，這就是交互作用。總平均 $\bar{y} = 20$。平方和（手算或程式，見第 14.8 節）：

$$\mathrm{SS}_A = 351, \quad \mathrm{SS}_B = 162, \quad \mathrm{SS}_{AB} = 39, \quad \mathrm{SSE} = 12, \quad \mathrm{SST} = 564.$$

來源	SS	df	MS	F	p
肥料 A	351	2	175.5	175.5	$< 10^{-8}$
水量 B	162	1	162.0	162.0	$< 10^{-7}$
交互 AB	39	2	19.5	19.5	0.00017
誤差	12	12	1.0		
總計	564	17			

解　$F_{AB} = 19.5 > F_{0.05;\,2,12} = 3.885$，交互作用顯著。結論不該寫成「肥料有效、水也有效」這種把兩者拆開的句子，而該寫：「水的增產效果在有施肥時明顯較大；無肥時多澆水幾乎沒幫助。」這正是農業與藥物試驗關心的協同效應（synergy）。η^2：A 佔 $351/564 = 0.62$，B 佔 0.29，交互佔 0.069——主效果仍主導，但交互作用不是雜訊。

直觀理解

交互作用就是第 5 章「獨立」的反面：兩個因子的聯合效應 ≠ 兩個主效果相加。圖 14.3(a) 裡，把 B 從低調到高，不論 A 在哪一檔都是往上平移同一段——可加。圖 14.3(b) 裡平移的距離在變。機器學習裡到處都是交互：dropout 在小模型有用、在大模型有害；學習率的最佳值取決於 batch size。只報告「平均而言 dropout 提升 0.3 個百分點」，可能把兩個相反的子效應沖銷成一句沒人能用的話。

14.7 ANOVA 就是虛擬變數迴歸

第 10 章的線性迴歸 $y = \beta_0 + \beta_1 x + \varepsilon$ 裡，x 是連續的。若 x 是「哪一組」這個類別，只要把它編成 $k-1$ 個**虛擬變數**（dummy variables），迴歸與 ANOVA 輸出同一個 F。

例 14.6　教學法的虛擬變數編碼

以 A 組為參照，令 $D_B = 1$ 若該學生在 B 組（否則 0），D_C 同理。模型

$$Y = \beta_0 + \beta_B D_B + \beta_C D_C + \varepsilon$$

的最小平方解就是各組平均：

$$\hat\beta_0 = \bar y_A = 84, \qquad \hat\beta_B = \bar y_B - \bar y_A = -9, \qquad \hat\beta_C = \bar y_C - \bar y_A = 5.$$

迴歸的解釋平方和 SSR = 503.33 等於 SSB，殘差平方和等於 SSW = 168，$R^2 = \eta^2 = 0.750$。檢定 $H_0\colon \beta_B = \beta_C = 0$（兩個虛擬係數都是 0）的迴歸 F 為

$$F = \frac{(\mathrm{SST} - \mathrm{SSE})/(k-1)}{\mathrm{SSE}/(N-k)} = \frac{503.33/2}{168/12} = 17.98,$$

與 ANOVA 的 F **數字到小數都相同**。所謂「類別預測變數的迴歸」，就是 ANOVA；所謂「ANOVA」，就是沒有連續共變項的迴歸。把連續共變項加進來，就是**共變數分析**（ANCOVA）——同一條最小平方，只是設計矩陣多了幾欄。

這如何驅動機器學習／ AI

這條等式在機器學習裡有三個直接後果。**(1) A/B/n 測試**：同時上線 k 個頁面變體，正確流程是一次 ANOVA（或它的廣義線性版，轉換率用邏輯斯），再 Tukey 找出勝過對照的變體——不是對每一對轉換率做未校正 z 檢定。**(2) 消融實驗**（ablation）：拿掉特徵組 A / B / C 之後損失有沒有變，是 k 組均值比較；同一測試集上的損失高度相依，樸素 ANOVA 會過於樂觀，實務上用配對、重複量數或 cluster-bootstrap。**(3) 巢狀模型 F 檢定**：完整模型相對精簡模型是否顯著改善，公式

$$F = \frac{(\mathrm{SSE}_{\text{small}} - \mathrm{SSE}_{\text{full}})/q}{\mathrm{SSE}_{\text{full}}/(n-p)}$$

就是 ANOVA 分解——q 是多出來的參數個數。第 10 章「加一層、加一組特徵值不值得」問的正是這個 F。超參數搜尋若把驗證集翻了幾十遍再挑最小的損失，又回到本章開頭的多重比較：沒有校正的「最佳組態」含有選擇偏差。

重複量數（repeated measures）ANOVA 是同一條思路的配對版：每位受試者經歷所有水準，人與人之間的差異被抽成獨立的誤差項，組內雜訊變小、檢定力上升——正如配對 t 檢定比雙樣本 t 更敏銳。代價是多一條**球面性**（sphericity）假設；違反時用 Greenhouse–Geisser 校正，或改用混合效應模型。本書不展開計算，但你應知道：同一個人、同一模型、同一測試集上的多次測量，不是獨立樣本。

14.8 Python 實作

SciPy 已內建單因子 ANOVA、Levene、Tukey HSD 與 Kruskal–Wallis。雙因子沒有一列式介面，下面用定義直接算平方和——看懂這段，任何多因子表你都能自己拆。

```
import numpy as np
from scipy import stats

A = np.array([82., 86, 79, 83, 90])   # 講授
B = np.array([75., 78, 72, 80, 70])   # 影片
C = np.array([88., 91, 85, 89, 92])   # 練習

F, p = stats.f_oneway(A, B, C)
print(round(F, 2), round(p, 5))       # -> 17.98  0.00025

print(stats.levene(A, B, C).pvalue)   # -> 0.66，不拒絕同質性
print(stats.kruskal(A, B, C).pvalue)  # -> 0.0059，結論一致

th = stats.tukey_hsd(A, B, C)
print(th)                             # A-B、B-C 顯著；A-C 否
```

雙因子：把 6 個格子做成形狀 (a, b, n) 的陣列，依定義累加。

```
# 軸 0 = 肥料（無/低/高），軸 1 = 水（少/多），軸 2 = 重複
Y = np.array([
    [[12, 14, 13], [15, 16, 14]],
    [[18, 19, 17], [24, 26, 25]],
    [[20, 21, 19], [28, 30, 29]],
], dtype=float)
a, b, n = Y.shape
grand = Y.mean()
A_mean = Y.mean(axis=(1, 2))          # 各肥料的邊際平均
```

```
B_mean = Y.mean(axis=(0, 2))
cell = Y.mean(axis=2)
SSA = b * n * ((A_mean - grand)**2).sum()
SSB = a * n * ((B_mean - grand)**2).sum()
SSAB = n * ((cell - A_mean[:, None]
             - B_mean[None, :] + grand)**2).sum()
SSE = ((Y - cell[:, :, None])**2).sum()
MSE = SSE / (a * b * (n - 1))
FA, FB = (SSA / (a - 1)) / MSE, (SSB / (b - 1)) / MSE
FAB = (SSAB / ((a - 1) * (b - 1))) / MSE
pA = 1 - stats.f.cdf(FA, a - 1, a * b * (n - 1))
pB = 1 - stats.f.cdf(FB, b - 1, a * b * (n - 1))
pAB = 1 - stats.f.cdf(FAB, (a - 1) * (b - 1), a * b * (n - 1))
print(round(FA, 1), round(FB, 1), round(FAB, 1))
print(pA, pB, pAB)                    # 三個 p 都 << 0.05
```

輸出 $F_A = 175.5$、$F_B = 162.0$、$F_{AB} = 19.5$，與例 14.5 的表一致。把 `Y` 換成你自己的試驗資料即可。

本章總結

- 對 k 組做 $\binom{k}{2}$ 次未校正 t 檢定會讓 FWER 膨脹；ANOVA 用**一次** F 檢定 $H_0\colon \mu_1 = \cdots = \mu_k$。
- 靈魂恆等式 $\mathrm{SST} = \mathrm{SSB} + \mathrm{SSW}$；$F = \mathrm{MSB}/\mathrm{MSW} \sim F_{k-1,N-k}$（$H_0$、常態、同質、獨立之下）。$F$ 遠大於 1 才拒絕。
- $k = 2$ 時 $F = t^2$，兩個 p 值相同。
- 整體顯著之後才做 Bonferroni 或 Tukey HSD；Tukey 專為全部兩兩比較而設，等組人數時 $|\bar{y}_i - \bar{y}_j| > \mathrm{HSD}$ 即顯著。
- 效果量 $\eta^2 = \mathrm{SSB}/\mathrm{SST}$ 等於虛擬變數迴歸的 R^2；p 小不代表效應大。
- 雙因子先檢定交互作用：顯著則解讀簡單主效果，不要只報邊際均值。
- 單因子 ANOVA = 虛擬變數迴歸的 F 檢定；加上共變項就是 ANCOVA；巢狀模型比較仍是同一個 F。
- 獨立性最要命。重複量數、同一測試集上的多模型比較，都不是獨立樣本。
- ML 實戰：A/B/n、消融、超參數搜尋，都是多組比較——先整體、後校正、報告效果量。

14.9 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 14.1 (基礎)　四種飼料、每種 $n = 6$ 隻動物，$N = 24$。已知 SSB $= 48$、SSW $= 80$。(a) 寫出完整 ANOVA 表（SS、df、MS、F）。(b) $\alpha = 0.05$ 的臨界值為 $F_{0.05;\,3,20} = 3.098$，下結論。(c) 計算 η^2，並用一句話解釋它。

練習 14.2 (基礎)　三種睡前飲品對睡眠時數（小時）的影響，每組 $n = 5$：咖啡 $\{5, 6, 4, 5, 5\}$、安慰劑 $\{7, 8, 6, 7, 7\}$、茶 $\{6, 7, 6, 5, 6\}$。(a) 計算各組平均與總平均。(b) 計算 SSB、SSW、F。(c) 已知 $F_{0.05;\,2,12} = 3.885$，下結論並給出 $p = 0.0028$ 的解讀。

練習 14.3 (基礎)　關於單因子 ANOVA，判斷正誤並說明：(a)「F 顯著表示每一對組平均都有差異。」(b)「H_0 為真時 F 應該靠近 0。」(c)「組內變異數很大時，比較難拒絕 H_0。」(d)「$p = 0.04$ 表示效應很大。」(e)「各組人數必須相等才能做 ANOVA。」

練習 14.4 (進階)　由 $Y_{ij} - \bar{Y}_{\cdot\cdot} = (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}) + (Y_{ij} - \bar{Y}_{i\cdot})$ 出發，完整證明 SST $=$ SSB $+$ SSW，特別說明交叉項為何為 0。並驗證自由度 $(N-1) = (k-1) + (N-k)$ 的「約束計數」直觀。

練習 14.5 (進階)　兩組分數 $X = \{10, 12, 11, 13, 14\}$、$Y = \{8, 9, 7, 10, 6\}$。(a) 做雙樣本等變異 t 檢定，報 t、p。(b) 做單因子 ANOVA，報 F、p。(c) 驗證 $F = t^2$，並說明兩個 p 值為何相同。(d) 若改用「變異數不同」的 Welch t，還能指望 $F = t^2$ 嗎？為什麼？

練習 14.6 (進階)　(a) 列出單因子 ANOVA 的三條假設，並指出哪一條被違反時後果最嚴重。(b) Levene 檢定的 $p = 0.40$ 應如何解讀？「不拒絕」是否等於「證明各組變異數相等」？(c) 何時改用 Kruskal–Wallis？它的 H_0 與 ANOVA 的 H_0 有何細微差別？

練習 14.7 (進階)　$k = 5$ 組，你打算做全部兩兩比較。(a) m 是多少？未校正時 FWER 是多少（獨立、H_0 全真、$\alpha = 0.05$）？(b) Bonferroni 的單一比較門檻是多少？(c) 為什麼 Tukey 在「全部兩兩」這個任務上通常比 Bonferroni 更有檢定力？(d) 若你事前只想比較四種新藥與一種安慰劑，應該用 Tukey 還是 Dunnett？

練習 14.8 (進階)　三個模型在五個摺疊上的準確率（%）：基準 $\{72, 74, 71, 73, 75\}$、dropout $\{78, 80, 77, 79, 81\}$、mixup $\{79, 81, 78, 80, 82\}$。(a) 計算三組平均、SSB、SSW、F，並與 $F_{0.05;\,2,12} = 3.885$ 比較。(b) Tukey 的 $q_{0.05;\,3,12} = 3.773$，計算 HSD 並判斷哪幾對顯著。(c) 用一句話寫給沒學過統計的工程師聽。

練習 14.9 (進階)　某 2×2 藥物試驗的細胞均值（療效分數）：男 / 安慰劑 50、男 / 藥物 62、女 / 安慰劑 52、女 / 藥物 54。每格人數相同。(a) 畫出交互作用圖（兩條線：男、女；

橫軸：安慰劑–藥物）。(b) 藥物在男性與女性身上的簡單效果各是多少？(c) 主效果「藥物平均提升 $(12+2)/2=7$ 分」這句話有什麼問題？(d) 若交互作用檢定顯著，你的報告應怎麼寫？

練習 14.10 (挑戰)　把練習 14.2 的三組編成虛擬變數（以咖啡為參照）。(a) 寫出迴歸模型，並不看軟體、直接由組平均讀出 $\hat\beta_0$、$\hat\beta_{\text{安慰劑}}$、$\hat\beta_{\text{茶}}$。(b) 說明檢定 $H_0\colon \beta_{\text{安慰劑}}=\beta_{\text{茶}}=0$ 的迴歸 F 為何等於練習 14.2 的 ANOVA F。(c) 若再加入連續共變項「年齡」，模型叫什麼？F 檢定的 H_0 變成什麼意思？

練習 14.11 (挑戰)　三組、$n=10000$ 人 / 組。樣本平均恰好為 100.00、100.20、100.10，$\text{MSW}=25$。(a) 計算 SSB、F。(b) 自由度很大時 $F_{0.05;\,2,\infty}\approx 3.00$，下結論。(c) 計算 η^2。(d) 為什麼「統計顯著」與「實務顯著」在這裡分道揚鑣？若這是三種推薦演算法的點擊率（百分比），你會怎麼向產品經理報告？

練習 14.12 (挑戰·程式)　用 NumPy 模擬單因子 ANOVA 的型一錯誤。(a) 在 H_0 全真之下：三組各 $n=20$，皆抽自 $N(0,1)$，重複 $M=5000$ 次實驗。每次做 `f_oneway`，估計 $P(p<0.05)$，應接近 0.05。(b) 同一批實驗，對三對做未校正雙樣本 t 檢定，估計「至少一對 $p<0.05$」的比例，應接近 $1-0.95^3\approx 0.143$。(c) 改用 `tukey_hsd`，估計「至少一對 Tukey $p<0.05$」的比例，應接近 0.05——這正是 HSD 在控制的 FWER。(d) 把 (a) 的一組改成 $N(0,9)$（變異數異質）再跑一次，觀察名義 $\alpha=0.05$ 被破壞了多少。

14.10 練習題詳解

練習 14.1 詳解　(a) $\text{df}_B=4-1=3$，$\text{df}_W=24-4=20$，$\text{df}_T=23$。$\text{MSB}=48/3=16$，$\text{MSW}=80/20=4$，$F=16/4=4$。$\text{SST}=128$。

來源	SS	df	MS	F
組間	48	3	16	4
組內	80	20	4	
總計	128	23		

(b) $4>3.098$，拒絕 H_0。精確 $p=\mathbb{P}(F_{3,20}\ge 4)=0.022$。四種飼料的均值並非全部相等。
(c) $\eta^2=48/128=0.375$：總變異裡約 38% 可被飼料種類解釋，屬大效應。

練習 14.2 詳解　(a) 咖啡 $\bar y=5$，安慰劑 $\bar y=7$，茶 $\bar y=6$，總平均 $\bar y_{\cdot\cdot}=6$。
(b) $\text{SSB}=5\big[(5-6)^2+(7-6)^2+(6-6)^2\big]=10$。

$$\text{SSW}=(0+1+1+0+0)+(0+1+1+0+0)+(0+1+0+1+0)=6.$$

MSB $= 10/2 = 5$，MSW $= 6/12 = 0.5$，$F = 5/0.5 = 10$。
(c) $10 > 3.885$，拒絕 H_0。$p = 0.0028$ 的意思是：若三種飲品的母體均值真的相同，抽到「組間差異至少這麼大」的資料，機率只有 0.28%。它不是「H_0 為真的機率」。

練習 14.3 詳解　(a) **錯**。F 只說「至少一組不同」。可能是一組離群、其餘相等，必須事後比較才知道是哪幾對。
(b) **錯**。H_0 下 MSB 與 MSW 都估計 σ^2，F 靠近 **1**（精確地說，期望值 $\mathrm{df}_W/(\mathrm{df}_W - 2)$ 略大於 1）。靠近 0 反而古怪，暗示 MSB 小得不尋常。
(c) **對**。MSW 是 F 的分母，組內雜訊大，F 被壓小，檢定力下降。減雜訊（配對、共變項、更好的測量）與加樣本，是兩條提升檢定力的路。
(d) **錯**。p 值混合了效應、雜訊與樣本數。大樣本可以把微小差異打成 $p = 0.04$；看 η^2 或組平均差的信賴區間。
(e) **錯**。不平衡設計（各 n_i 不同）完全合法，公式用 n_i 加權即可。不平衡時同質性假設被違反的後果比較嚴重，這時才較需要 Welch。

練習 14.4 詳解　令 $a_i = \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}$、$e_{ij} = Y_{ij} - \bar{Y}_{i\cdot}$。則 $Y_{ij} - \bar{Y}_{\cdot\cdot} = a_i + e_{ij}$，

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(a_i + e_{ij})^2 = \sum_{i,j} a_i^2 + \sum_{i,j} e_{ij}^2 + 2\sum_{i=1}^{k} a_i \sum_{j=1}^{n_i} e_{ij}.$$

對固定的 i，$\sum_j e_{ij} = \sum_j (Y_{ij} - \bar{Y}_{i\cdot}) = n_i\bar{Y}_{i\cdot} - n_i\bar{Y}_{i\cdot} = 0$，故交叉項為 0。又 $\sum_j a_i^2 = n_i a_i^2$，即 SST $=$ SSB $+$ SSW。
自由度：計算 SST 時，把 N 個偏差加總前已強制它們的和為 0（因為減了總平均），故 $\mathrm{df}_T = N - 1$。SSB 用了 k 個組平均，但它們對總平均加權平均必須還原 $\bar{Y}_{\cdot\cdot}$，1 個約束，$\mathrm{df}_B = k - 1$。SSW 每組減了自己的平均，k 個約束，$\mathrm{df}_W = N - k$。兩者相加還原 $N - 1$。

練習 14.5 詳解　(a) $\bar{x} = 12$、$\bar{y} = 8$。合併變異數 $s_p^2 =$ MSW。雙樣本 $t = 4$，雙尾 $p = 0.00395$（$\mathrm{df} = 8$）。
(b) SSB $= 5\big[(12-10)^2 + (8-10)^2\big] = 40$，MSB $= 40$。可算 SSW $= 20$，MSW $= 20/8 = 2.5$，$F = 40/2.5 = 16$，$p = 0.00395$。
(c) $t^2 = 4^2 = 16 = F$，兩個 p 值相同，因為 $\{t : |t| \geq 4\}$ 與 $\{F : F \geq 16\}$ 在 t_8 與 $F_{1,8}$ 下是同一事件（命題 14.4）。
(d) **不能**。Welch t 不合併兩組變異數，分母不再是 ANOVA 的 MSW；Welch ANOVA 才是它的 k 組對應。經典 ANOVA 與等變異 t 共享同一誤差項，這才有 $F = t^2$。

練習 14.6 詳解　(a) 獨立、常態、變異數同質。獨立性被違反最嚴重：它讓 MSW 系統性低估不確定性，p 值全面失效，大樣本救不了。常態性在每組人數不太小時相當穩健；同質性在平衡設計下也還穩健，不平衡加上異質才危險。

(b)「若各組母體變異數相等，得到至少這麼極端的 Levene 統計量」的機率是 0.40——沒有證據拒絕同質性。**不拒絕 $\neq$ 證明相等**（第 8 章）：可能只是樣本太小、檢定力不足。把它當成「可以安心用經典 ANOVA 的綠燈」，同時用眼睛看一下各組 s_i。
(c) 明顯離群、嚴重偏態、或應變數本質是順序的。Kruskal–Wallis 的 H_0 是「各組來自同一分布」（更強：不只均值，整個分布），拒絕時可能是散布不同而非位置不同。它用等級，不受單調變換影響。

練習 14.7 詳解　(a) $m = \binom{5}{2} = 10$。FWER $= 1 - 0.95^{10} = 0.401$。十次獨立的「公平硬幣偏 5%」實驗，四成機率至少出現一顆星號。
(b) $0.05/10 = 0.005$。
(c) Bonferroni 用聯集上界，對任意相依結構都有效，因而對「全部兩兩」這種正相關的家族偏保守。Tukey 用學生化全距——精確描述「k 個組平均裡最大減最小、再除以共同標準誤」的分布，把家族裡的相依吃進臨界值，所以在這個特定任務上檢定力更高、FWER 仍精確等於 α。
(d) **Dunnett**。Tukey 為全部 $\binom{5}{2} = 10$ 對付錢；Dunnett 只為 4 對「處理 vs. 對照」付錢，臨界值較小、檢定力較高。事前計劃比事後「全部都比」便宜，這是多重比較最重要的實務教訓之一。

練習 14.8 詳解　(a) 平均 73、79、80，總平均 77.333。

$$\text{SSB} = 5\big[(73-77.333)^2 + (79-77.333)^2 + (80-77.333)^2\big] = 143.33,$$

SSW $= 30$（每組對自己平均的平方和都是 10），$F = (143.33/2)/(30/12) = 28.67 \gg 3.885$，$p = 2.7 \times 10^{-5}$。整體：三個模型的均值並非全部相等。
(b) HSD $= 3.773 \times \sqrt{30/12/5} = 3.773 \times \sqrt{0.5} = 2.668$。$|73-79| = 6$、$|73-80| = 7$ 都 > 2.668，顯著；$|79-80| = 1 < 2.668$，不顯著。dropout 與 mixup 都勝過基準，彼此分不出來。
(c)「在這五個摺疊上，兩個改進版都明顯優於基準；至於 dropout 和 mixup 誰比較好，這份資料沒有分辨力。」

練習 14.9 詳解　(a) 橫軸安慰劑–藥物。男性那條從 50 升到 62（陡）；女性那條從 52 升到 54（幾乎平）。兩條線不平行。
(b) 男性簡單效果 +12；女性 +2。
(c) 邊際主效果 +7 是兩個子群的平均，沒有任何一個性別真正提升 7 分。它掩蓋了「對男性很有效、對女性幾乎無效」這個臨床上最重要的事實。
(d) 先報交互作用顯著，再分層報簡單效果與信賴區間：「藥物使男性療效分數平均提升 12 分，女性僅 2 分；兩性反應不同，不宜用單一『平均療效』概括。」若要決定是否開藥，應進入「性別 × 藥物」的簡單效果檢定或事前對比，而不是看邊際 F_B。

練習 14.10 詳解　(a) $Y=\beta_0+\beta_P D_P+\beta_T D_T+\varepsilon$，其中 D_P、D_T 分別是安慰劑、茶的指示變數。$\hat\beta_0=\bar y_{\text{咖啡}}=5$，$\hat\beta_P=7-5=2$，$\hat\beta_T=6-5=1$。虛擬係數就是「該組相對參照組的平均差」。

(b) $H_0:\beta_P=\beta_T=0$ 等價於「三組平均都等於參照組平均」，也就是三組平均相等。迴歸的 $\text{SSR}=\text{SSB}=10$，$\text{SSE}=\text{SSW}=6$，

$$F=\frac{10/2}{6/12}=10$$

與練習 14.2 相同。設計矩陣的行空間就是「組平均可以任意」的那個 k 維空間，兩個 F 是同一投影的兩種寫法。

(c) 叫做共變數分析（ANCOVA）。此時針對飲品的 F 變成：**在年齡相同的人當中**，三種飲品的均值是否仍不同。它把年齡能解釋的變異從 SSW 裡抽走，通常讓誤差更小、檢定更敏銳；同時校正了「分組時年齡碰巧不均衡」的混淆。

練習 14.11 詳解　(a) 總平均 100.10。

$$\text{SSB}=10000\big[(100.00-100.10)^2+(100.20-100.10)^2+0^2\big]=200,$$

$\text{MSB}=100$，$F=100/25=4$。$\text{df}_W=29997$。

(b) $4>3.00$，拒絕 H_0。大樣本把 0.2 個單位的差異打成統計顯著。

(c) $\text{SSW}=29997\times 25=749925$，$\text{SST}=750125$，$\eta^2=200/750125\approx 0.00027$。總變異裡只有 0.027% 來自組別。

(d) p 值隨 n 趨於把任何非零差異判為顯著——這是一致性，不是 bug。但產品經理要的是「值得改版的效果」。0.2 個百分點的點擊率、$\eta^2\approx 0.0003$，大概遠小於改版成本與風險。報告應同時給：$F=4$（$p\approx 0.018$）、各組平均、η^2、以及平均差的信賴區間。顯著性是「訊號超過雜訊」；重要性是「訊號超過決策門檻」。

練習 14.12 詳解　以下是一份可直接執行、註解對應 (a)–(d) 的草稿。你的估計值會因種子而略有起伏，$M=5000$ 時標準誤約 $\sqrt{0.05\times 0.95/5000}\approx 0.003$，所以 (a) 落在 0.04–0.06、(b) 落在 0.13–0.16、(c) 落在 0.04–0.06 都算成功。

```
import numpy as np
from scipy import stats

rng = np.random.default_rng(42)
M, n = 5000, 20
rej_f = rej_t = rej_tukey = 0
for _ in range(M):
    g = rng.normal(0, 1, size=(3, n))
    if stats.f_oneway(*g).pvalue < 0.05:
```

```
        rej_f += 1
    ps = [stats.ttest_ind(g[i], g[j]).pvalue
          for i, j in [(0, 1), (0, 2), (1, 2)]]
    if min(ps) < 0.05:
        rej_t += 1
    th = stats.tukey_hsd(*g)
    if np.min(th.pvalue) < 0.05:      # 對角線為 1，min 取自兩兩比較
        rej_tukey += 1
print(rej_f / M, rej_t / M, rej_tukey / M)
```

(a) 應 ≈ 0.05：名義顯著水準被守住，表示 F 檢定在假設成立時校準正確。(b) 應 ≈ 0.14：未校正兩兩比較的 FWER 膨脹，與 $1-0.95^3$ 吻合。(c) 應 ≈ 0.05：Tukey 把「至少一對被宣告顯著」的機率拉回 α，這正是 HSD 存在的理由。(d) 把 `g[2] = rng.normal(0, 3, n)` 再跑 (a)，平衡設計下 F 仍大致穩健（估計值會略偏離 0.05，但通常不至於翻倍）；若再改成 $n=(5,5,50)$ 這種嚴重不平衡，異質性就會明顯拉高或壓低型一錯誤——這是教科書建議「不平衡時改 Welch」的原因。