第 13 章　卡方檢定

Chi-Square Tests

本章處理**類別性資料**（以次數分布呈現），用同一個卡方統計量 $\chi^2 = \sum \frac{(O-E)^2}{E}$ 做三種右尾檢定：(1) **適合度檢定**——單一母體某類別變數是否符合特定分配；(2) **獨立性檢定**——同一母體兩屬性是否獨立（列聯表）；(3) **齊一性檢定**——多母體分配是否相同。差別在抽樣設計與自由度，統計量公式一致。

目錄

1　名詞速查（五欄詞表）

中文術語	English	符號	一句定義	用途／易混
卡方統計量	chi-square statistic	$\chi^2 = \sum \frac{(O-E)^2}{E}$	衡量觀察與期望次數的總差異	三種檢定共用、右尾
觀察／期望次數	observed / expected	O / E	實際次數 / H_0 下應發生的次數	$E \geq 5$ 才可靠
適合度檢定	goodness-of-fit	—	檢定單一母體是否符合特定分配	$E = np_{i0}$
獨立性檢定	test of independence	—	同一母體兩屬性是否獨立	列聯表
齊一性檢定	test of homogeneity	—	多母體分配是否相同	統計量同獨立性
列聯表	contingency table	$r \times c$	兩屬性交叉的次數表	$E_{ij} = \frac{R_i C_j}{n}$
五的規則	rule of five	$E \geq 5$	每格期望次數須 ≥ 5	否則合併格子
自由度	degrees of freedom	df	適合度 $k-1-m$；列聯表 $(r-1)(c-1)$	m ＝估計參數數

2 核心概念：一個統計量，三種檢定

類別資料與卡方統計量 類別資料以各類「次數」呈現。比較**觀察次數** O 與「H_0 為真時應發生的**期望次數** E」。差異 $O-E$ 直接相加為 0，故取平方再除以 E 標準化：

$$\chi^2=\sum\frac{(O-E)^2}{E}\ \approx\ \chi^2_{df}.$$

χ^2 越大代表 O 與 E 差越多 → 越該拒絕 H_0，故三種檢定都是**右尾**：$R=\{\chi^2\geq\chi^2_{\alpha,df}\}$。

三種檢定的分工

- **適合度**：抽**單一**母體，檢定一個類別變數是否符合給定機率分配（如基因比 1:2:1、市佔率不變）。
- **獨立性**：抽**單一**母體同時記錄**兩個**屬性，檢定兩屬性是否獨立（如性別 × 購物平台）。
- **齊一性**：從**多個**母體各抽樣，檢定各母體的分配是否相同（如城市/鄉村對素食接受度）。

獨立性與齊一性的列聯表計算**完全相同**（E_{ij}、df、統計量都一樣），差別只在抽樣設計與結論措辭。

決策：臨界值法與 p-value 法（兩種等價）

卡方檢定一律**右尾**，下面兩種決策方式必得同結論：

- **臨界值法**：拒絕域 $R=\{\chi^2\geq\chi^2_{\alpha,df}\}$；算出的 χ^2_{obs} 落入 R 就拒絕 H_0。
- **p-value 法**：$p=P(\chi^2_{df}\geq\chi^2_{\text{obs}})$，即檢定統計量**右側的面積**；$p\leq\alpha$ 就拒絕 H_0，p 愈小證據愈強。

等價關係：$\chi^2_{\text{obs}}\geq\chi^2_{\alpha,df}\iff p\leq\alpha$。
手算限制：χ^2 表只列特定幾個 α，手算算不出精確 p，只能用表把 p **夾出範圍**（例：χ^2_{obs} 介於 $\chi^2_{0.05,df}$ 與 $\chi^2_{0.025,df}$ 之間 ⇒ $0.025<p<0.05$）；要精確值用 Excel `CHISQ.DIST.RT` 或 R 的 `pchisq`。本章例題沿用臨界值法。

3 適合度檢定（Goodness-of-Fit）

期望次數、統計量與自由度

$$E_i=n\,p_{i0}\ (\text{樣本數}\times H_0\text{機率}),\qquad \chi^2=\sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i}\sim\chi^2_{df}.$$

自由度 $df=k-1-m$（$k=$格數，$m=$由樣本估計的參數個數；機率全給定時 $m=0$，$df=k-1$）。

五的規則（rule of five）

卡方為近似檢定，須**每格期望次數** $E_i\geq5$。若有格子 $E<5$，將相鄰類別**合併**至滿足，合併後 k 變小、df 隨之減少。

個案 1：基因型比例 1:2:1（不拒絕）

理論模式 A:B:C = 1 : 2 : 1，即 $p_A = 0.25, p_B = 0.5, p_C = 0.25$。試驗 100 顆，觀察 $O = 18, 55, 27$。$\alpha = 0.05$。
期望 $E = 100 \times (0.25, 0.5, 0.25) = (25, 50, 25)$。

$$\chi^2 = \frac{(18-25)^2}{25} + \frac{(55-50)^2}{50} + \frac{(27-25)^2}{25} = 1.96 + 0.50 + 0.16 = 2.62.$$

$df = 3-1-0 = 2$；$R = \{\chi^2 \geq \chi^2_{0.05,2} = 5.991\}$；$2.62 < 5.991 \Rightarrow$ **Do not reject** H_0：無證據說理論模式有誤。

例題 13.1：市場佔有率是否改變（拒絕）

廣告前市佔 A:B: 其他 $= 0.45 : 0.40 : 0.15$。抽 200 人，偏好 $O = 102, 82, 16$。$\alpha = 0.05$。
期望 $E = 200 \times (0.45, 0.40, 0.15) = (90, 80, 30)$。

$$\chi^2 = \frac{(102-90)^2}{90} + \frac{(82-80)^2}{80} + \frac{(16-30)^2}{30} = 1.6 + 0.05 + 6.53 = 8.18.$$

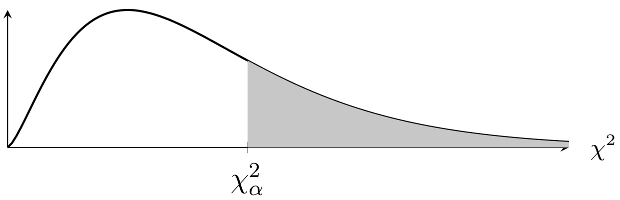
$df = 2$；$R = \{\chi^2 \geq 5.991\}$；$8.18 \geq 5.991 \Rightarrow$ **Reject** H_0：市佔率已顯著改變。

補充：期望次數 < 5 要合併

五公司市佔 $0.40, 0.32, 0.24, 0.02, 0.02$，抽 200 人 $O = 70, 60, 54, 10, 6$。H_0 下 D、E 的期望 $= 200 \times 0.02 = 4 < 5$，違反五的規則 $\rightarrow$ 合併 D、E 為一格（機率 0.04，$E = 8$，$O = 16$）：

$$\chi^2 = \frac{(70-80)^2}{80} + \frac{(60-64)^2}{64} + \frac{(54-48)^2}{48} + \frac{(16-8)^2}{8} = 10.25.$$

合併後 $k = 4$，$df = 4-1 = 3$；$R = \{\chi^2 \geq \chi^2_{0.05,3} = 7.815\}$；$10.25 \geq 7.815 \Rightarrow$ **Reject** H_0。



適合度／獨立性／齊一性皆右尾：$R = \{\chi^2 \geq \chi^2_{\alpha,df}\}$（陰影面積 α）。

4　獨立性檢定與齊一性檢定（列聯表）

列聯表期望次數、統計量與自由度

$r \times c$ 列聯表，第 i 列總和 R_i、第 j 行總和 C_j、總和 n：

$$E_{ij} = \frac{R_i C_j}{n}, \qquad \chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{df}, \qquad df = (r-1)(c-1).$$

推導：為什麼 $E_{ij} = R_i C_j / n$、$df = (r-1)(c-1)$

若兩屬性獨立，$p_{ij} = p_{i\cdot}\, p_{\cdot j}$。以邊際比例估計 $\hat{p}_{i\cdot} = R_i/n$、$\hat{p}_{\cdot j} = C_j/n$，故 $E_{ij} = n\hat{p}_{ij} = n \cdot \frac{R_i}{n} \cdot \frac{C_j}{n} = \frac{R_i C_j}{n}$。

自由度：格數 rc，扣 1（總和限制），再扣估計的邊際參數 $(r-1)+(c-1)$（每組比例和為 1），得 $df = rc - 1 - [(r-1)+(c-1)] = (r-1)(c-1)$。

個案 2：性別 × 購物平台是否獨立（2×3）

隨機訪 395 人：

	Yahoo	Momo	東森	合計
男	30	24	21	75
女	182	104	34	320
合計	212	128	55	395

期望次數 $E_{ij} = R_i C_j/395$，例 $E_{11} = \frac{75\times 212}{395} = 40.25$、$E_{13} = \frac{75\times 55}{395} = 10.44$、$E_{21} = \frac{320\times 212}{395} = 171.75$、$E_{23} = 44.56$。

$$\chi^2 = \frac{(30-40.25)^2}{40.25} + \frac{(24-24.30)^2}{24.30} + \frac{(21-10.44)^2}{10.44} + \cdots + \frac{(34-44.56)^2}{44.56} = 16.40.$$

$df = (2-1)(3-1) = 2$；$R = \{\chi^2 \geq \chi^2_{0.05,2} = 5.991\}$；$16.40 \geq 5.991 \Rightarrow$ **Reject** H_0：性別與購物平台**有關聯**（不獨立）。

例題 13.2：大學學位 × MBA 主修（4×3）

抽 152 位 MBA 學生，4 種學位 × 3 種主修。期望次數同樣用 $E_{ij} = R_i C_j/152$（如 $E_{\text{BA, 會計}} = \frac{60\times 61}{152} = 24.08$）。算得

$$\chi^2 = \frac{(31-24.08)^2}{24.08} + \frac{(13-17.37)^2}{17.37} + \cdots = 14.7.$$

$df = (4-1)(3-1) = 6$；$R = \{\chi^2 \geq \chi^2_{0.05,6} = 12.592\}$；$14.7 \geq 12.592 \Rightarrow$ **Reject** H_0：學位與主修方向有關。

獨立性 vs 齊一性　兩者列聯表算法（E_{ij}、χ^2、$df = (r-1)(c-1)$）完全相同，只差抽樣設計：獨立性是**一個**母體記錄兩屬性（邊際皆隨機）；齊一性是**多個**母體各抽固定樣本，檢定分配是否相同（一邊邊際是事先固定的）。

5 公式整理（總表）

檢定	期望次數 E	自由度 df
適合度	$E_i = n\,p_{i0}$	$k-1-m$（m =估計參數數）
獨立性	$E_{ij} = R_i C_j/n$	$(r-1)(c-1)$
齊一性	$E_{ij} = R_i C_j/n$	$(r-1)(c-1)$

共同：$\chi^2 = \sum \frac{(O-E)^2}{E}$，右尾 $R = \{\chi^2 \geq \chi^2_{\alpha,df}\}$，須 $E \geq 5$。

6 易錯點總整理

卡方檢定

- **一律右尾**：χ^2 大才拒絕；沒有雙尾、左尾。
- **自由度別記錯**：適合度 $k-1-m$（機率全給定 $m=0$）；列聯表 $(r-1)(c-1)$，不是 $rc-1$。
- **五的規則**：每格 $E \geq 5$；不足要合併格子，合併後 k（或 r, c）變小、df 跟著變。
- **期望次數來源**：適合度 $E = np_{i0}$；列聯表 $E_{ij} = R_i C_j / n$（列總 × 行總 ÷ 總和）。
- 統計量用**次數**（O, E），不是比例或機率。
- 獨立性與齊一性「算法相同」，但 H_0 措辭不同（獨立 vs 分配相同）。

7 計算跳板（數值代入演練）

類型	代入	結果
適合度 E	$200 \times (0.45, 0.40, 0.15)$	$= (90, 80, 30)$
適合度 χ^2	$\frac{(102-90)^2}{90} + \frac{(82-80)^2}{80} + \frac{(16-30)^2}{30}$	$= 8.18$ ($\geq \chi^2_{0.05,2} = 5.991$，拒絕)
列聯表 E_{ij}	$E_{11} = \frac{75\times212}{395}$	$= 40.25$
列聯表 df	$(2-1)(3-1)$	$= 2$（2×3 表）
查臨界值	$\chi^2_{0.05,2}$；$\chi^2_{0.05,6}$	$= 5.991$；$= 12.592$

寫字區：自選一題，從假設寫到結論（八步驟）＋列期望次數表

8 自我檢查

1. 三種卡方檢定共用的統計量是什麼？為什麼都是右尾？
2. 適合度檢定的期望次數怎麼算？自由度公式為何要扣「估計參數個數」？
3. 4×3 列聯表獨立性檢定的自由度是多少？期望次數怎麼算？
4. 某格期望次數 $= 3$，違反什麼規則？該怎麼處理？對 df 有何影響？
5. 獨立性檢定與齊一性檢定的計算差在哪？抽樣設計差在哪？

寫字區：作答

參考答案： 1. $\chi^2 = \sum(O-E)^2/E$；χ^2 大表示 O 與 E 差大、越該拒絕，故右尾。 2. $E_i = np_{i0}$；若機率須由樣本估計，每估一個參數就少一個自由度（$df = k-1-m$）。 3. $df = (4-1)(3-1) = 6$；$E_{ij} = R_iC_j/n$。 4. 違反五的規則（$E \geq 5$）；合併相鄰類別，合併後格數變少、df 減少。 5. 計算（E_{ij}、χ^2、$df = (r-1)(c-1)$）完全相同；獨立性抽單一母體記兩屬性，齊一性自多母體各抽固定樣本。