單因子變異數分析

One-Way ANOVA

1 問題概述與理論基礎

- 單因子 ANOVA 的主要議題是:

  ▷ 按照某個「因子」不同而分成的 k 群母體，其平均值是否相同?

  虛無假設為 $\rightarrow$ $\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k$

  ▷ ANOVA = Analysis of Variance = 變異數分析

  使用這 k 群「變異數」的差異，來對其「平均值」進行分析 $\rightarrow$ F 檢定

- ANOVA 與四大檢定的連結:

	單一母體 one population	兩個母體 two populations
平均值 mean	H_0: $\mu = 5$ t-統計量	H_0: $\mu_1 = \mu_2$ t-統計量
變異數 variance	H_0: $\sigma^2 = 10$ χ^2-統計量	H_0: $\sigma_1^2 = \sigma_2^2$ F-統計量

  ▷ 可視為兩個母體平均值檢定的一般化版本 (可用於「兩個以上」母體)

- ANOVA 與相關分析的差異:

	自變數		應變數	
	個數		個數	
相關分析	一個	屬量	一個	屬量
變異數分析	多個	屬質	一個	屬量

  ▷ ANOVA 按自變數個數不同，分為單因子 ANOVA 和多因子 ANOVA

  ▷ 「相關分析」與「單迴歸分析」也是一體的兩面:

$$\text{單迴歸的式子：}\ y = \alpha + \beta x + u\text{，其中迴歸係數 } \beta = \rho_{xy}\frac{\sigma_y}{\sigma_x}$$

- 從一個簡單的例子了解問題本質:

  ABC 食品公司生產各類型的休閒食品，它在桃園、台中、高雄均設有工廠。公司管理人員想瞭解三個工廠的產品重量是否相同，以確保公司產品品質的穩定性。於是管理人員自三間工廠各隨機抽取了 5 個產品重量的資料如下表，想用以檢定這三間工廠的產品重量是否相同。

15 個產品重量

觀察值	桃園	臺中	高雄
1	86	78	100
2	78	90	86
3	68	89	88
4	90	67	80
5	68	88	76
樣本平均數	78	82	86
樣本變異數	102	91.5	84

▷ 虛無假設為 $\rightarrow$ H_0: $\mu_1 = \mu_2 = \mu_3$

對立假設為 $\rightarrow$ H_1: 以上三者不全相等 (前述 H_0 的反敘述)

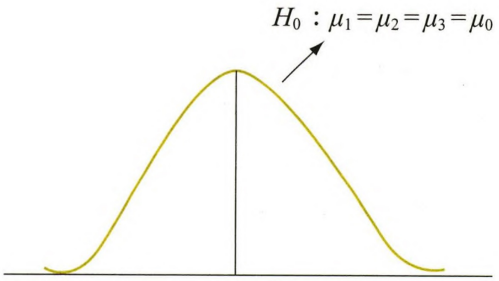

圖 12.1 AVOVA 中虛無假設成立的狀況

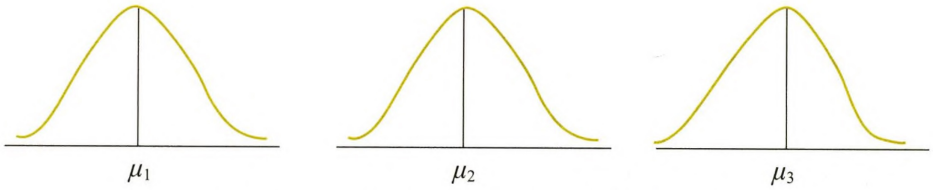

圖 12.2 AVOVA 中對立假設成立的狀況

▷ 思考: 為什麼檢定「平均值」要透過「變異數」?

- 定義: 變異 = 平方和 (Sum of Squares)

▷ 總變異 (Sum of Squares Total) $\rightarrow$ $\text{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2$

▷ 組間變異 (Sum of Squares Between) $\rightarrow$ $\text{SSB} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$

▷ 組內變異 (Sum of Squares Error) $\rightarrow$ $\text{SSE} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$

- 上述三者存在以下關係:

$$\text{SST} = \text{SSB} + \text{SSE}$$
$$(\text{總變異} = \text{組間變異} + \text{組內變異})$$

  ▷ 思考: SST 固定，如果 SSB 很大，但 SSE 很小，代表什麼情況?
  (組間差異相對大，組內差異相對小)

  ▷ 思考: SST 固定，如果 SSB 很小，但 SSE 很大，又代表什麼情況?
  (組間差異相對小，組內差異相對小)

  ▷ 以上那一種情形，比較傾向拒絕 H_0? (當然是前者!)

  ▷ 可見: 利用組間和組內變異的相對大小，可以用來判斷平均值是否相同!
  (這就是 ANOVA 分析的原理 → 用變異數來分析平均值)

- ANOVA 的檢定統計量:

  ▷ 檢定統計量的正式定義為

$$F\text{檢定統計量} = \frac{\text{標準化的組間變異}}{\text{標準化的組內變異}} = \frac{\dfrac{\text{SSB}}{k-1}}{\dfrac{\text{SSE}}{n-k}} = \frac{\text{MSB}}{\text{MSE}} \sim F_{(k-1,n-k)}$$

  (此處的 n 為各組樣本數的總和，k 為組數)

  ▷ 追加定義: 分子 = 組間均方 MSB (Mean Square Between) $= \dfrac{\text{SSB}}{k-1}$

  ▷ 追加定義: 分母 = 組內均方 MSE (Mean Square Error) $= \dfrac{\text{SSE}}{n-k}$

  ▷ 所以可看出: ANOVA 分析是進行 F 檢定!
  (它是透過變異數的相對差異來對平均值的相同與否進行檢定!)

  ▷ 思考: 在六大檢定中，F 檢定用來檢定什麼? (答: 比較兩變異數看是否相等)
  又為何此處 ANOVA 也是用 F 檢定? (答: 因為在比較組間和組內的變異數)

- ANOVA Table

變異來源	平方和 (SS)	自由度 (df)	均方 (MS)	F	p-value (significance)	臨界值 (critical value)
組間	SSB	$k-1$	MSB	$\frac{\text{MSB}}{\text{MSE}}$	(軟體產生)	(軟體產生)
組內	SSE	$n-k$	MSE			
總共	SST					

- 事後檢定 (post hoc test) (或稱後測，多重比較)

  ▷ ANOVA 只能檢定多組的平均值是否相等

  ▷ 如果拒絕 H_0，表示平均值不全相等，我們並不知道是那幾組不相等

  ▷ 因此需進行兩兩比較，此即為事後檢定 (後測)

  (常用的方法為 Tukey 檢定和 Scheffe 檢定，此處略)

- 進行 ANOVA 檢定的前提假設

  1. 每組 (處理 = treatment) 樣本背後的母體服從常態分配
  2. 每組背後的常態母體，變異數均相同 (在此基礎下檢定平均值是否相等)

     → 有時為求慎重，會先做同質性檢定 (例如 Levene's test，Bartlett's test)

     以確定此一前提成立，此類同質性檢定的虛無假設為

     $$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

  3. 每組樣本皆由各組母體中隨機取得

2 兩組樣本的單因子 ANOVA

- 此時用以前四大檢定中的 t 檢定，和用現在單因子 ANOVA 的 F 檢定，將會得到一樣的結果

  ▷ 上述的 t 檢定指獨立 (非成對) 樣本，且在變異數相同的前提之下

  (亦即，與 ANOVA F 檢定的前提相同)

  ▷ 可觀察到 $F = t^2$ 的數學關係

  ▷ 兩種檢定的 p-value 會完全一樣 (對 t 指雙尾，對 F 指單尾)

- 例題 1:

  在上市公司中，獲利率和產業之間是否有關? 若是，我們在投資股票時就要慎選產業; 若否，就要看各公司的表現來決定投資對象。在本例中，我們以石化業和電子業 ROE (股東權益報酬率) 的隨機抽樣資料，以 ANOVA 來分析產業和報酬率的關係:

  石化業與電子業歷年產業平均 ROE

年度	80	81	82	83	84	85	86	87	88	89
石化	0.05	0.06	−0.02	−0.05	−0.01	0.00	0.00	0.00	0.02	0.00
電子	0.06	0.02	0.04	0.03	0.07	0.01	0.04	0.07	0.08	0.00

  ▷ 請自行觀察 $F = t^2$ 的關係

  ▷ 請自行確定採用「平均值檢定」和「ANOVA 檢定」會得到一樣的檢定結果。