變異數分析 ANOVA

概念、單因子變異數分析、事後比較

變異數分析的必要性

- 在此之前，我們學的 t檢定
  - 重點在於兩個母體的檢定
  - 但是，三個母體呢？四個呢？以上呢？

  - t 檢定不太適用
    - 為什麼？
    - 三個母體我們必須檢定3 $\left(C_2^3 = \frac{3!}{2!(3-2)!} = 3\right)$次，假設信心水準為 95%，因為有三次，故信心水準變成$95\% \times 95\% \times 95\% = 85.74\%$
  - 當然有解決方法，$x^3 = 95\% \Longrightarrow x \doteqdot 98.30\%$

變異數分析的概念

- 變異數分析 ANOVA (ANalysis Of VAriance, F檢定)
  - 專有名詞
    - 水準：男生/女生，偏遠/離島/市區，甲/乙/丙/丁
    - 因子：性別，地區，班級
      - 不同班級(甲/乙/丙/丁)的數學成績是否有差異？
      - 代表一個因子(班級)，有四個水準(甲/乙/丙/丁)，四個水準是否有差異
  - 這裡，我們要討論的是「單因子變異數分析」(One-way ANOVA)
  - 主要是採用變異數來檢定差異

變異數分析的概念 (續)

- 名詞
  - 組間平方和(SS_b)：每一水準和總平均的差異的平方和
    - $SS_b=$ (每個水準 − 總平均$)^2$的總和 $=\sum_j n_j\left(\bar{x}_j-\bar{\bar{x}}\right)^2$
  - 組內平方和(SS_w)：每一水準內資料和水準的差異的平方和
    - $SS_w=$ (每個資料 − 其水準$)^2$的總和 $=\sum_j\sum_i\left(x_{ij}-\bar{x}_j\right)^2$
  - 總變異=組間平方和+組內平方和
- $F=\frac{MSB}{MSE}=\frac{\frac{SS_B}{df_B}}{\frac{SS_W}{df_W}}$
  - 我們希望，組間差異愈大愈好，為什麼？
    - 代表不同水準有非常大的差異
  - $df_B=$ 水準個數 − 1；$df_w=$ 總個數 − 水準個數

範例一

- 來自不同地區學生，其對台東大學滿意度有所差異？
  - 滿意度為1~5點量表，5代表非常滿意，1代表非常不滿意

偏遠(組1)	鄉鎮(組2)	城市(組3)
5,3,5,4,4	3,4,3,4,5	2,3,2,3,1
$\bar{X}_1 = 4.2$	$\bar{X}_2 = 3.8$	$\bar{X}_3 = 2.2$

  - $\bar{\bar{x}} = \frac{5+3+5+4+4+3+4+3+4+5+2+3+2+3+1}{5+5+5} = 3.4$
  - $SS_b = \sum_j n_j\left(\bar{x}_j - \bar{\bar{x}}\right)^2 = 5 \times (4.2-3.4)^2 + 5 \times (3.8-3.4)^2 + 5 \times (2.2-3.4)^2 = 11.2$
  - $SS_w = \sum_j \sum_i \left(x_{ij} - \bar{x}_j\right)^2 = [(5-4.2)^2 + (3-4.2)^2 + (5-4.2)^2 + (4-4.2)^2 + (4-4.2)^2] + [(3-3.8)^2 + (4-3.8)^2 + (3-3.8)^2 + (4-3.8)^2 + (5-3.8)^2] + [(2-2.2)^2 + (3-2.2)^2 + (2-2.2)^2 + (3-2.2)^2 + (1-2.2)^2] = 8.4$

範例一 (續)

- $$F = \frac{MSB}{MSE} = \frac{\frac{SS_b}{df_b}}{\frac{SS_w}{df_w}} = \frac{\frac{11.2}{3-1}}{\frac{8.4}{15-3}} = \frac{5.6}{0.7} = 8$$

變異數摘要表

	平方和(*SS*)	自由度(*df*)	平均平方和(*MS*)	F檢定
組間	11.2	2	5.6	8
組內	8.4	12	0.7	
總和	19.6	14		

變異數分析的其他概念

- 基本假設
  - 每個水準的母體均為常態分配
  - 每個母體的變異數均相等
  - 各水準的樣本互為獨立

- 單因子變異數分析
  - 自變數：一個，就是因子，但因子有多個水準
  - 依變數：一個，例如對台東大學滿意度
  - 如果
    - 兩個自變數(以上)，稱為二(多)因子變異數分析
    - 兩個依變數以上，稱為多變量分析

範例二

- 某工廠欲瞭解4部機器的性能。但機器會因為無法測知的因素導致每小時平均產量不同。試問，這4部機器的產量是否有顯著差異？

機器	A	B	C	D
產量	10, 15, 8, 12, 15	14, 18, 21, 15	17, 16, 14, 15, 17, 15, 18	12, 15, 17, 15, 16, 15

  - 虛無假設：各機器的平均產量相同 $(\mu_A = \mu_B = \mu_C = \mu_D)$
  - 對立假設： 某兩機器的平均產量有所不同
  - 採用單因子變異數分析
  - 信心水準：95%

範例二 (續)

	A	B	C	D
n	5	4	7	6
$\bar{X}$	12	17	16	15
$SS = \sum(X_i - \bar{X})^2$	38	30	12	14

- $\bar{\bar{X}} = 15$
- $SS_b = 5 \times (12-15)^2 + 4 \times (17-15)^2 + 7 \times (16-15)^2 + 6 \times (15-15)^2 = 68$
- $SS_w = 38 + 30 + 12 + 14 = 94$
- $MSB = \frac{SS_b}{df_b} = \frac{68}{4-1} = 22.67$
- $MSE = \frac{SS_w}{df_w} = \frac{94}{22-4} = 5.22$

範例二 (續)

- $F = \frac{MSB}{MSE} = \frac{22.67}{5.22} = 4.34 > F_{\alpha}(3{,}18) = 3.1599$

變異數摘要表

	平方和(SS)	自由度(df)	平均平方和(MS)	F檢定
組間	68	3	22.67	4.34
組內	94	18	5.22	
總和	162	21		

- 結論：落在拒絕區，拒絕虛無假設。在信心水準95%下，有兩部機器的平均產量有顯著差異。

事後比較 Post hoc

- ANOVA的問題
  - 它拒絕了虛無假設「各水準間沒有差異」，但是
    - 是大家都有差異嗎？是其中兩個有差異嗎？是其中幾個有差異嗎？
    - 以範例二來看，$\mu_A < \mu_D < \mu_C < \mu_B$，所以大家都有差異。對嗎？
      - 沒錯啊！平均數不一樣啊！
      - 不幸地，我們真的不知道！
- 事後比較
  - 在ANOVA之後，再進行比較(就字面上來看)
  - 檢定兩兩水準的平均數
  - 有很多方法
    - Bonforrini法：$(\bar{X}_i - \bar{X}_j) \pm t_{\frac{\alpha}{2m}}(df_w) \times S_p \times \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$
    - Scheffe法：$(\bar{X}_i - \bar{X}_j) \pm \sqrt{df_b \times F_\alpha(df_b, df_w)} \times S_p \times \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$

$m = C_2^k$

$S_p = \sqrt{MSE}$

範例三：Bonforrini法

- 接續範例二，試問是那些機器有顯著差異？採用Bonforrini法。
- $t_{\frac{\alpha}{2m}}(df_w) \times S_p = t_{\frac{.05}{2\times 6}}(18) \times \sqrt{5.22} = 2.878 \times \sqrt{5.22} = 6.5755$
- $\mu_A - \mu_B$: $(12-17) \pm 6.5755 \times \sqrt{\frac{1}{5}+\frac{1}{4}} = (-9.411, -0.589)$
- $\mu_A - \mu_C$: $(12-16) \pm 6.5755 \times \sqrt{\frac{1}{5}+\frac{1}{7}} = (-7.851, -0.149)$
- $\mu_A - \mu_D$: $(12-15) \pm 6.5755 \times \sqrt{\frac{1}{5}+\frac{1}{6}} = (-6.982, 0.982)$
- $\mu_B - \mu_C$: $(17-16) \pm 6.5755 \times \sqrt{\frac{1}{4}+\frac{1}{7}} = (-3.122, 5.122)$

範例三：Bonforrini法 (續)

- $\mu_B - \mu_D$: $(17-15) \pm 6.5755 \times \sqrt{\frac{1}{4}+\frac{1}{6}} = (-2.245, 6.245)$
- $\mu_C - \mu_D$: $(16-15) \pm 6.5755 \times \sqrt{\frac{1}{7}+\frac{1}{6}} = (-2.659, 4.659)$
- 結論：在信心水準95%下，
  - A機器的產量顯著低於B機器的產量
  - A機器的產量顯著低於C機器的產量
  - 其餘沒有顯著差異

範例四：Scheffe法

- 接續範例二，試問是那些機器有顯著差異？採用Scheffe法。
- $\sqrt{df_b \times F_\alpha(df_b, df_w)} \times S_p = \sqrt{3 \times 4.34} \times \sqrt{5.22} = 8.2441$
- $\mu_A - \mu_B$: $(12-17) \pm 8.2441 \times \sqrt{\frac{1}{5}+\frac{1}{4}} = (-10.530, 0.530)$
- $\mu_A - \mu_C$: $(12-16) \pm 8.2441 \times \sqrt{\frac{1}{5}+\frac{1}{7}} = (-8.827,\ \ 0.827)$
- $\mu_A - \mu_D$: $(12-15) \pm 8.2441 \times \sqrt{\frac{1}{5}+\frac{1}{6}} = (-7.992, 1.992)$
- $\mu_B - \mu_C$: $(17-16) \pm 8.2441 \times \sqrt{\frac{1}{4}+\frac{1}{7}} = (-4.167, 6.167)$
- $\mu_B - \mu_D$: $(17-15) \pm 8.2441 \times \sqrt{\frac{1}{4}+\frac{1}{6}} = (-3.322, 7.322)$
- $\mu_C - \mu_D$: $(16-15) \pm 8.2441 \times \sqrt{\frac{1}{7}+\frac{1}{6}} = (-3.587, 5.587)$
- 結論：接受虛無假設。在信心水準95%下，所有機器皆沒有顯著差異。

為什麼和ANOVA
結果不一樣？