變異數分析

何謂變異數分析

例：不同教學方式對學生成績的影響

影響學生成績高低的因素很多，譬如說，教學方式、用功程度、老師長相等，都可能會影響同學學習成績。如果我們只考慮一種因素對於學生成績（資料）的影響，稱為單因子變異數分析。如果考慮兩種因素對於學生成績（資料）的影響，稱為雙因子變異數分析。

例：Fisher(1924)研究不同的肥料對作物收成的影響。

- $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$
- $H_1 : H_0$不為真

～之前學的 t 檢定可以讓我們檢查兩個獨立母體的平均數是否有差異，而變異數分析則可以檢驗多個母體平均數是否顯著不同。

變異數分析(analysis of variance; ANOVA) 是一種統計分析的方法，係將一組資料所發生的總變異，依可能發生變異的來源分割為數個部分，每一部分均可歸因於某原因（變異來源）；測度這些不同的變異來源，可了解各種變異是否有顯著差異；若有差異，則表示某一變異來源對資料具有顯著的影響作用，否則便無影響作用。

實驗資料與觀察資料

統計分析所用的資料，可依據資料產生的方式分為「實驗性資料」與「觀察性資料」。觀察性資料(observational Data)是在分析者不介入下，測量紀錄所觀察的資料。實驗性資料 (experimental data) 則是分析者採用實驗設計的方法，經由人為控制干預或安排後所獲得的資料。

實驗法 (experimental method)，是指在控制的情境之下，實驗者有系統的操作 (manipulate) 實驗變數 (又稱因子或自變數)，收集資料以觀察因子對實驗結果 (又稱反應變數、依變數) 的影響。而所謂控制的情境，是指盡可能排除實驗變數以外之一切會影響實驗結果之變數而言。

例子：吃蛋糕是否容易導致肥胖？

- 觀察性資料：訪問 100 位體重過重者，並記錄這些人之中經常吃蛋糕的人數，計算比例。
- 實驗性資料：徵求 100 位體重正常的人，隨機分成兩組，其中一組每天吃蛋糕，另一組不吃。過一個月後，測量所有人的體重。

例子：比較某手機品牌的三種不同廣告策略（A、B、C），對顧客的購買意願的影響是否有明顯差異。

～實驗性資料在數位化的趨勢下，變成現代管理的重要依據，即以證據為基礎的決策。

～因果關係與相關關係不同，實驗性資料在釐清因果關係上有優勢。

常見名詞

- 實驗單位 (experimental unit)：即被實驗的對象，可以是產品、動物、人等等。
- 反應變數 (response)：即實驗結果的變數，也稱做依變數 (dependent variable)。例如某工廠生產A產品的良率。
- 因子 (factor)：即實驗者所操作的變數，也稱為自變數 (independent variable)。例如某工廠生產A產品的溫度與壓力就稱為因子。
- 因子水準 (level)：表示實驗因子的狀態，或簡稱為水準 (level)，若溫度分為高溫、低溫兩種狀態，高溫就是溫度因子的一種水準，而低溫是另一種水準。
- 處理 (treatment)：因子水準的特定組合，稱為處理 (treatment)。若溫度分為兩種水準：高溫、低溫；壓力分為兩種水準：高壓、低壓，則該實驗就有四種處理：(高溫、高壓)、(高溫、低壓)、(低溫、高壓)、(低溫、低壓)。
- 變異數分析 (ANOVA)：即衡量因子對反應變數是否有影響的統計分析方法。

例題 11.1

陳老師想研究三種教學方法對大二學生（共九班）統計學成績的影響，假設三種教學方法為黑板教學、投影片教學、電腦輔助教學，請說明該實驗之實驗單位、反應變數、因子、因子水準、處理分別為何？

解：(1) 實驗單位：即大二學生（共九班）。

(2) 反應變數（依變數）：即統計學成績。因為陳老師想知道教學方法不同時，是否統計學成績亦會隨之不同，所以統計學成績是反應變數。

(3) 因子（自變數）：只有一項，即教學方法。

(4) 因子水準：教學方法為因子，該因子有三個水準，即三種不同的教學方法。

(5) 處理：只有教學方法一個因子，因此處理與因子水準相同，即三種不同的處理方式。

END

進行變異數分析的前提假設

- 常態性假設(Normal distribution)：每個母體之反應變數是常態分配。
- 變異數同質性假設(Homogeneity of variance)：每個母體之反應變數的變異數σ^2均相等。
- 可加性(Additivity)：每個母體抽取之隨機觀察值是獨立樣本。

變異數分析的觀念

- $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$
- $H_1: H_0$不為真

～如果多個母體的平均數相等，那麼所抽出的多個樣本平均數會非常接近。

如果多個樣本平均數愈接近，則愈能支持母體平均數相等的結論；如果多個樣本平均數差異愈大，則愈不支持母體平均數相等的結論。因此，當我們有足夠的證據證明樣本平均數差異很小，則接受H_0，反之接受H_1。請注意：**樣本平均數差異的大小是指樣本平均數的離散程度，即樣本來源是否是來自平均數相同的母體分配。**

首先，如果母體為常態分配$N(\mu, \sigma)$，則樣本數的大小為n時，樣本平均數的抽樣分配為常態分配，$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

其次，若H_0為真，表示所有樣本均來自平均數相同的母體分配，那麼多個樣本平均數抽樣分配的樣本共同變異數$S_{\bar{x}}^2$應為母體變異數σ^2的不偏估計。若H_0不為真，表示多個樣本來自平均數不同的母體分配，那麼多個樣本平均數抽樣分配的樣本共同變異數$S_{\bar{x}}^2$會比較大，將高估母體變異數σ^2。因此，我們可以用F分配做檢定，來推論母體平均數是否相等。

1. 一因子變異數分析

(1)完全隨機化設計

在探討因子對於資料變動的影響時，是以完全隨機的方式處理，稱為完全隨機設計(completely randomized design)。

一因子完全隨機設計是用來檢定某一個因子中，各個水準的影響力是否會造成反應值之間的顯著差異。在此，一個因子的一個水準相當於一個處理。

符號：假設實驗中的主要因子有k個水準，並且隨機指派第i個處理$i = 1, 2, \ldots, k$給實驗單位， 因此$n = n_1 + n_2 + \ldots + n_k$。

例如：我們想瞭解不同教學方式對學生成績的影響。在試驗中，對於任一個學生，我們是隨機選取其中一種教學方式來施教。

表 11.1　單因子變異數分析模型資料型態

因子 A				
因子水準 A_1	因子水準 A_2	因子水準 A_3	⋯	因子水準 A_k
Y_{11} ⋮ Y_{1n_1}	Y_{21} ⋮ Y_{2n_2}	Y_{31} ⋮ Y_{3n_3}	⋮ ⋮ ⋮	Y_{k1} ⋮ Y_{kn_k}

(2) 母體模式的總變異分解

總變異(total variation) 代表不同的反應值之間差異性(或變異性)的總和指標。

在比較處理所造成的反應值差異時，總變異可以分解為

(1) 組間變異 **(between-group variation):** 總變異之中，由於不同處理的影響力所造成的反應值變異或差異。又名已解釋變動量。

(2) 組內變異 **(within-group variation):** 來自同一處理中觀察值之間的組內變異。又名未解釋變動量。

例如：一群以填鴨方式教統計，另一群以直觀靈活方式教統計。

- 你個人成績與全班平均的差異可以分解成：你個人成績與你所屬族群組平均成績的差異，加上所屬群組平均成績與全班平均成績的差異。
- 所屬群組平均成績與全班平均成績的差異是不同教學法所造成的成績差異，是教學方式這個因子所造成的，是這個因子能解釋的部分，為已解釋變動量。
- 個人成績和所屬群組平均成績的差異，不是教學方式這個因子所能解釋的，可能是個人用功程度、運氣好壞等其他因素所影響，為未解釋變動量。

母體模式：以一個方程式來代表所有反應值的產生方式。

總變異 = 組內變異（未解釋變異）＋ 組間變異（可解釋變異）

$$Y_{ij} - \mu = (Y_{ij} - \mu_i) + (\mu_i - \mu)$$

$$\Rightarrow Y_{ij} - \mu = \varepsilon_{ij} + \alpha_i$$

$$\Rightarrow Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

這個模式代表每一個反應值 Y_{ij} 是由實驗前未受處理影響的平均值μ，加上第 i 個處理的影響效果α_i（代表組間變異），再加上同組實驗單位之間的隨機差異ε_{ij}（代表組內變異），共三個值相加而得最後的測量值。

(3)樣本反應值的總變異分解

相對於母體模式，反應值可以用樣本來表示。

例如：假設40個感冒病人被隨機分成兩組，分別服用新舊兩種感冒藥品。所有病人康復時間（單位：小時）的總變異可分解為以下兩類：

- 不同藥品所造成的變異：如果新舊藥品療效的確有差異，則兩組感冒病人分別服用新舊藥品後，「平均」康復時間應該會有明顯差異，這就是組間變異。
- 服用同一藥品的病人之間的變異：就算是服用同一種藥品，屬於同一組的20個不同病人的康復時間，可能也會因為每個人的居住環境、生活習慣、家族遺傳等因素影響而導致康復速度有差異，這就是組內變異。

例題 11.2

某大學隨機抽取 15 位學生，任意分成三組，每組人數不同，分別實施不同的打字訓練課程，訓練後加以測驗，其每分鐘的字數如下：

訓練課程	每分鐘字數
A	40
A	42
A	27
A	30
B	38
B	41
B	45
B	37
B	44
C	32
C	34
C	28
C	42
C	34
C	36

試問不同的訓練課程下，作業員每分鐘字數是否有差異？

解：對應表 11.1 中的符號，訓練課程為因子 A（變數 A），A、B、C 三種課程為因子水準 A_1、因子水準 A_2、因子水準 A_3，即因子 A 有 3 種可能的情況，$n_1=4$、$n_2=5$、$n_3=6$。

訓練課程（因子 A）		
訓練課程 A（因子水準 A_1）	訓練課程 B（因子水準 A_2）	訓練課程 C（因子水準 A_3）
$Y_{11}=40$ $Y_{12}=42$ $Y_{13}=27$ $Y_{14}=30$	$Y_{21}=38$ $Y_{22}=41$ $Y_{23}=45$ $Y_{24}=37$ $Y_{25}=44$	$Y_{31}=32$ $Y_{32}=34$ $Y_{33}=28$ $Y_{34}=42$ $Y_{35}=34$ $Y_{36}=36$

本例題之資料型態為典型的單因子變異數分析資料型態，在還沒有說明變異數分析的手法前，我們先試著透過統計量與統計圖表解這個問題。

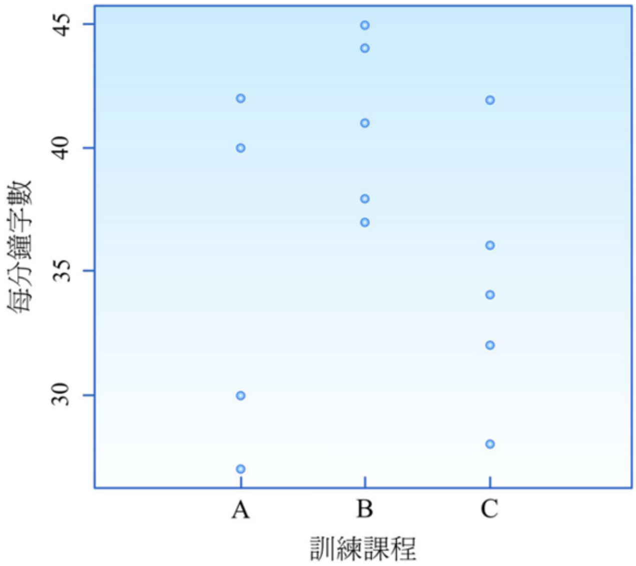
經由資料計算可得

訓練課程 A 的平均數 $\bar{y}_A=34.75$

訓練課程 B 的平均數 $\bar{y}_B=41$

訓練課程 C 的平均數 $\bar{y}_C=34.33$

繪製統計圖，如下：



由三種訓練課程的平均數可看出有差異；由統計圖也可感覺出有差異。其中，訓練課程 B 會有比較好的效果，平均數比較大，變異也比較小。 END

	訓練課程（因子）		
	訓練課程 A	訓練課程 B	訓練課程 C
每分鐘字數	40 42 27 30	38 41 45 37 44	32 34 28 42 34 36
各組母體平均數	μ_A	μ_B	μ_C
各組樣本平均數	$\overline{y}_A = 34.75$	$\overline{y}_B = 41$	$\overline{y}_C = 34.33$

～ k種處理方式完全隨機化設計的樣本資料結構如下：

	處理方式				
	1	2	3	$\cdots$	k
觀察值	X_{11}	X_{12}	X_{13}	$\cdots$	X_{1k}
	X_{21}	X_{22}	X_{23}	$\cdots$	X_{2k}
	X_{31}	X_{32}	X_{33}	$\cdots$	X_{3k}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	$X_{n_1 1}$	$X_{n_2 2}$	$X_{n_3 3}$	$\cdots$	$X_{n_k k}$
平均數	$\overline{X}_1 = \dfrac{\sum X_{i1}}{n_1}$	$\overline{X}_2 = \dfrac{\sum X_{i2}}{n_2}$	$\overline{X}_3 = \dfrac{\sum X_{i3}}{n_3}$	$\cdots$	$\overline{X}_k = \dfrac{\sum X_{ik}}{n_k}$
平方和	$\sum (X_{i1} - \overline{X}_1)^2$	$\sum (X_{i2} - \overline{X}_2)^2$	$\sum (X_{i3} - \overline{X}_3)^2$	$\cdots$	$\sum (X_{ik} - \overline{X}_k)^2$
總平均數	$\overline{\overline{X}} = \dfrac{n_1\overline{X}_1 + n_2\overline{X}_2 + \cdots + n_k\overline{X}_k}{n_1 + n_2 + \cdots + n_k} = \dfrac{\sum_j^k \sum_i^{n_j} X_{ij}}{n}$				

同時，變異數分析的計算可以分成三種不同的方式：

- **離差(deviation)**：代表任一觀察值X_i與其平均數$\bar{X}$ 的差，即$X_i - \bar{X}$。
- **變異(variation)**：離差的平方，即$(X_i - \bar{X})^2$ ；若將所有變異加總，即構成離差的平方和(sum of square for deviation)。
- **變異數 (variance)**：離差的平方和除以自由度，亦即平方和的均數，又稱均方。例如樣本變異數$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$

～總差異＝ 處理方式差異 + 殘差（隨機差異） = 組間差異 + 組內差異 = 已解釋差異 + 未解釋差異

(1)離差(Deviation)

變異數分析的方法通常係將平均數的總變異，按其發生來源予以分解。

總離差＝ 處理方式離差 + 殘差（隨機離差） = 組間離差 + 組內離差 = 已解釋離差 + 未解釋離差

由於每一觀察值與總平均數之間的差異，一部分係處理方式不同所致，另一部分係相同處理方式下測量上的隨機誤差（亦稱為殘差）所致，故可將之分解 。然後兩邊取平方和，即可得出總變異與各種來源變異的關係。

亦即$X_{ij} - \bar{\bar{X}} = (\bar{X}_j - \bar{\bar{X}}) + (X_{ij} - \bar{X}_j)$

(2) 變異 (Variation)

總變異 = 處理方式的變異 ＋ 殘差變異（隨機變異） = 組間變異 + 組內變異 = 已解釋變異 + 未解釋變異

$\sum_j^k \sum_i^{n_j} (X_{ij} - \bar{\bar{X}})^2 = \sum_j^k n_j (\bar{X}_j - \bar{\bar{X}})^2 + \sum_j^k \sum_i^{n_j} (X_{ij} - \bar{X}_j)^2$

平方和 **(sum of square)**

總平方和**(SST)** = 因子平方和**(SSF)** + 殘差平方和**(SSE)** = 組間平方和**(SSB)** + 組內平方和（**SSW**）

- **總變異**亦稱為總平方和(sum of square for total; SST)。
- **組間變異為不同處理方式的差異**，亦稱因子平方和(sum of squares due to factor, SSF)，或稱為組間平方和(sum of square between; SSB)。
- **組內變異為同一組相同處理方式內的差異**，亦稱殘差平方和(sum of square for error; SSE)，或稱為組內平方和(sum of square within; SSW)。

(3) 變異數 (Variance)

平方和除以自由度為不同變異來源所造成的平均變異，稱為均方。

均方(mean square)

- 總均方(total mean square, MST):
  - $MST = \frac{SST}{n-1}$
- 因子均方(mean square due to factor, MSF):
  - $MSF = MSB = \frac{SSF}{k-1}$
  - 自由度(k-1)的卡方分配除以其自由度(k-1)。
- 誤差均方(mean square due to error, MSE):
  - $MSE = MSW = \frac{SSE}{n-k}$
  - 卡方分配除以其自由度(n-k)

Summary: 完全隨機化設計的**ANOVA**表

變異來源	平方和 (SS)	自由度 (df)	均方 (MS)
處理方式	$SSB = \sum_{j=1}^{k} n_j(\overline{X}_j - \overline{\overline{X}})^2$	$k-1$	$MSB = \frac{SSB}{k-1}$
殘差	$SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_j)^2$	$n-k$	$MSE = \frac{SSE}{n-k}$
總和	$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \overline{\overline{X}})^2$	$n-1$	

$$df(SST) = df(SSB) + df(SSE) \tag{12-5}$$

$$n - 1 = (k-1) + (n-k)$$

變異數分析的統計推論

Question：某因子各個處理的影響力是否有明顯差異？

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$，所有處理(或因子水準)的效果都相同。

變異數分析的基本假設

(1) 每個反應變數的母體均為常態分配。

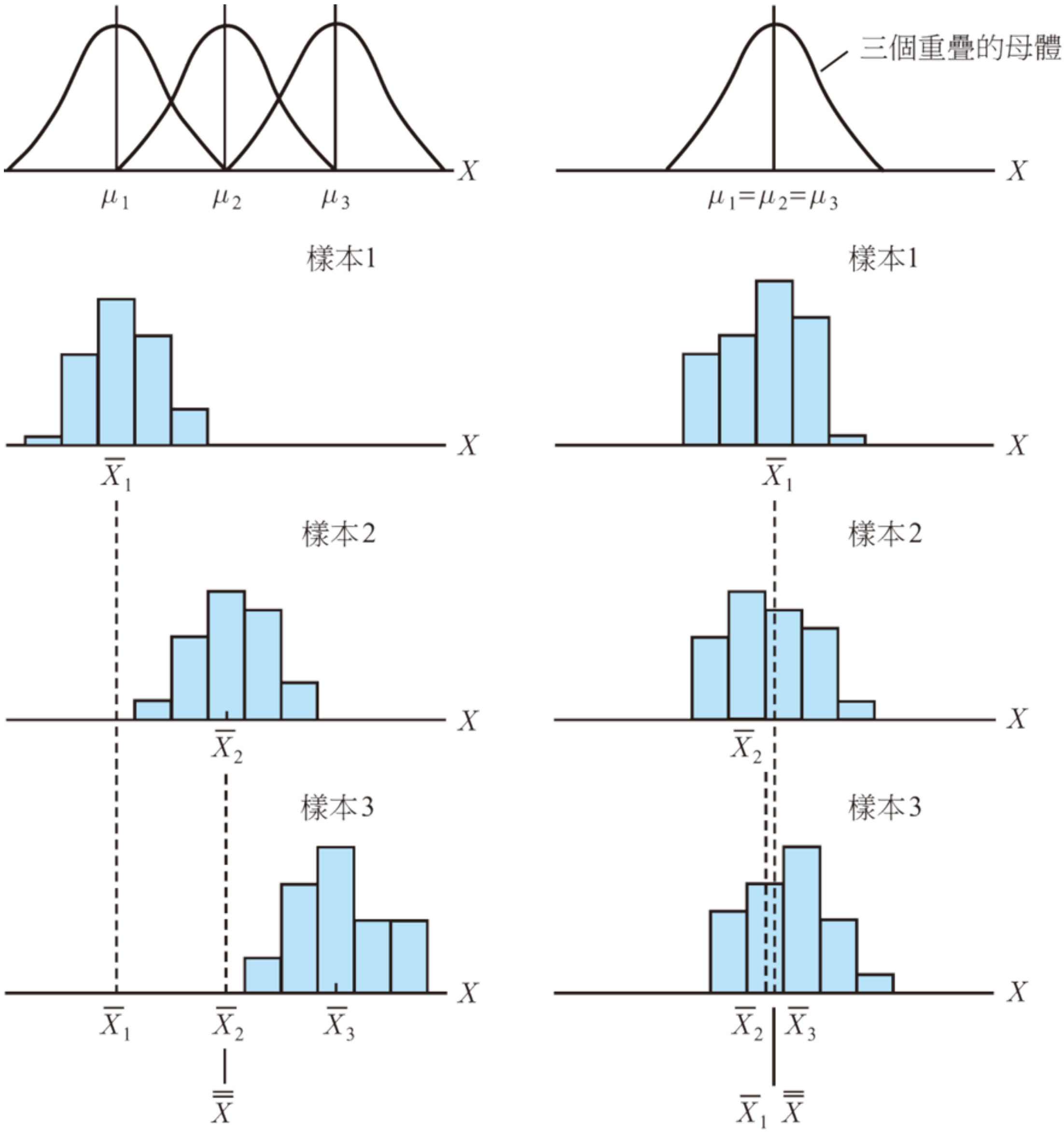
(2) 每個母體的變異數均相等。

(3) 抽自各母體的各組隨機樣本互為獨立。

母體變異數可由兩個獨立的估計值來估計，一個稱為組間變異數，另一個則為組內（或殘差）變異數，此二變異數即我們前面所提及的組間均方與組內（或殘差）均方，皆可用來估計母體變異數。

- 一個表示各組樣本平均數與所有樣本平均數之變異(S_B^2)
- 另一個則表示各觀察值與其處理方式平均數之變異(S_E^2)。

在H0成立與不成立下，樣本平均數$\bar{X}_j$抽樣分配的情形



(a) 具有相同變異數σ^2的三個常態母體，且平均數不同，假設$\mu_1<\mu_2<\mu_3$，H_0不成立。

(b) 具有相同變異數σ^2的三個常態母體，且平均數不同，假設$\mu_1=\mu_2=\mu_3$，H_0成立。

$$\frac{S_B^2}{S_E^2}=\frac{\frac{(k-1)S_B^2}{(k-1)}}{\frac{(n-k)S_E^2}{(n-k)}}=\frac{\frac{\frac{(k-1)S_B^2}{\sigma^2}}{(k-1)}}{\frac{\frac{(n-k)S_E^2}{\sigma^2}}{(n-k)}}=\frac{\frac{\chi_1^2(k-1)}{(k-1)}}{\frac{\chi_2^2(n-k)}{(n-k)}}=F(k-1,n-k)$$

兩個獨立的卡方分配各自除以其自由度後所得的比例式為F分配，且F分配的第一個自由度與第二個自由度分別與分子和分母的自由度相對應。 配合ANOVA表的使用，完全隨機化設計的F檢定統計量為：

$$F=\frac{S_B^2}{S_E^2}=\frac{MSB}{MSE}\geq F_\alpha(k-1,n-k)$$

例題 11.3

陳老師分別對四班各試行一種不同教學方法，期末考後從各班抽出部分學生，其統計學成績的資料如下：

教學方法

第一種	第二種	第三種	第四種
85	75	65	70
94	77	68	72
80	68	80	68
90	88	78	74
	83	76	65
		70	65
		73	

假設此資料適用變異數分析的方法，在顯著水準 $\alpha = 0.01$ 下，試檢定四種教學方法是否有顯著不同？

解：(1) 假設：$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_1 : \mu_i \text{ 不全等，} i = 1, 2, 3, 4 \end{cases}$

(2) 顯著水準：$\alpha = 0.01$

(3) 計算平方和：

$$\overline{y}_{1\bullet} = \frac{349}{4} = 87.25$$

$$\overline{y}_{2\bullet} = \frac{391}{5} = 78.20$$

$$\overline{y}_{3\bullet} = \frac{510}{7} = 72.86$$

$$\overline{y}_{4\bullet} = \frac{414}{6} = 69.00$$

$$\overline{\overline{y}} = \frac{85 + 94 + \cdots + 65 + 65}{4 + 5 + 7 + 6} = 75.64$$

$$\begin{aligned} SSTo &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{\overline{y}})^2 \\ &= (85 - 75.64)^2 + (94 - 75.64)^2 + \cdots + (65 - 75.64)^2 \\ &= 1,485.09 \end{aligned}$$

$$\begin{aligned} SSB &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\overline{y}_{i\bullet} - \overline{\overline{y}})^2 = 4(87.25 - 75.64)^2 + 5(78.2 - 75.64)^2 \\ &\quad + 7(72.86 - 75.64)^2 + 6(69 - 75.64)^2 \\ &= 890.68 \end{aligned}$$

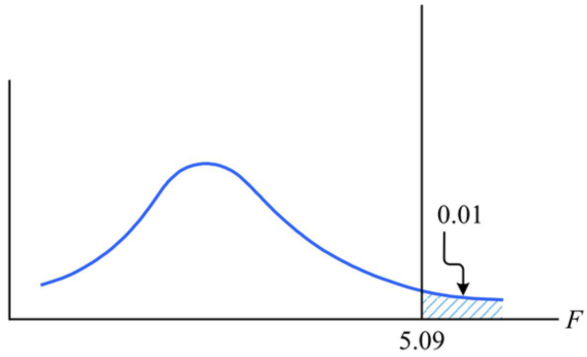
$$SSW = SSTo - SSB = 1485.09 - 890.68 = 594.41$$

(4) 變異數分析表：

變異來源	平方和	自由度	均方和	F 值
因　　子	890.68	3	296.89	8.99
誤　　差	594.41	18	33.02	

誤　　差	594.41	18	33.02	
總　　和	1485.09	21		

(5) 拒絕域：$F > F_{3,18,0.01} = 5.09$



(6) 判斷：$8.99 > 5.09$，所以拒絕 H_0，即四組學生的統計學成績有顯著差異，換言之，四種教學方法有顯著的不同效果。 END

例題 11.5

已知三組資料如下：

第 1 組：$n_1 = 3, \bar{y}_{1\bullet} = 13, s_1^2 = 25$

第 2 組：$n_2 = 5, \bar{y}_{2\bullet} = 14, s_2^2 = 16$

第 3 組：$n_3 = 7, \bar{y}_{3\bullet} = 15, s_3^2 = 9$，其中 $s_i^2 = \dfrac{\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\bullet})^2}{n_i - 1}, i = 1, 2, 3$

設 $\alpha = 0.05$，而資料適合變異數分析，試檢定三個母體的平均數是否相等？

解：(1) 假設：$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 \\ H_1 : \mu_i \text{ 不全等} \end{cases}$

(2) 顯著水準：$\alpha = 0.05$

(3) 計算平方和：

$$\bar{\bar{y}} = \frac{\sum_{i=1}^{k} n_i \bar{y}_{i\bullet}}{\sum_{i=1}^{k} n_i} = \frac{3(13) + 5(14) + 7(15)}{3 + 5 + 7} = 14.27$$

$$SSB = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{i\bullet} - \bar{\bar{y}})^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i\bullet} - \bar{\bar{y}})^2$$

$$= 3(13 - 14.27)^2 + 5(14 - 14.27)^2 + 7(15 - 14.27)^2 = 8.93$$

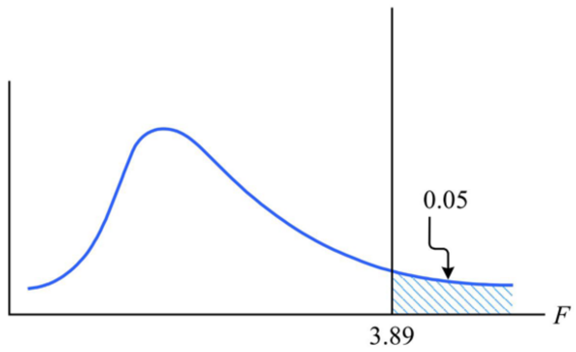
$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\bullet})^2 = \sum_{i=1}^{k}(n_i - 1)s_i^2 = (3-1)25 + (5-1)16 + (7-1)9 = 168$$

$$SSTo = SSB + SSW = 8.93 + 168 = 176.93$$

(4) 變異數分析表：

變異來源	平方和	自由度	均方和	F 值
處　　理	8.93	2	4.47	0.32
誤　　差	168.00	12	14.00	
總　　和	176.93	14		

(5) 拒絕域：$F > F_{2,12,0.05} = 3.89$



(6) 判斷：$0.32 < 3.89$，差異不顯著，所以不拒絕 H_0，即表示三個母體的平均數可能相等。

END

2. 數個平均數的多重比較

如果 F 檢定的結果顯著，變異數分析拒絕虛無假設時，表示 k 組母體平均數不全等。但到底是那些對平均數有顯著差異？若想更進一步知道到底是哪幾對處理間有明顯差異，就需要使用聯合信賴區間 (simultaneous confidence intervals)進行分析。此種「同時」(simultaneous) 比較好幾對平均數的差異方法，稱為多重比較 (multiple comparisons)。

費雪LSD

若變異數分析已經拒絕的虛無假設，費雪最低顯著差異(Least significant difference, LSD)程序可用以決定哪些母體平均數存在差異，是最早且可能應用最廣的兩兩比較方法。

費雪LSD程序試以獨立樣本t檢定之t統計量為基礎，並將估計母體變異數進行修正而得。檢定統計量如下：

$$\bar{X}_i - \bar{X}_j \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

式中$\bar{X}_i$ (i=1,2,...,k)表示來自母體i之樣本平均數，而S_P^2實際上應為共同母體變異數σ^2的綜合估計值，在此我們可用MSE取代，即$S_P = \sqrt{MSE}$。

多重 **t** 聯合信賴區間 = Bonferroni法

類似獨立樣本t檢定。首先，計算兩兩平均數差的組數；其次，為了確保整體鑑定能維持在$1-\alpha$的顯著水準，必須調整每個個別檢定的α值為$\alpha^* = \frac{\alpha}{C_2^k}$。

聯立信賴區間（小樣本）

$m = C_2^k$ 個（k 表示處理方式的個數）兩平均數差 $(\mu_i - \mu_j)$ 之 $100(1-\alpha)\%$ 聯立信賴區間為：

$$(\overline{X}_i - \overline{X}_j) \pm t_{\frac{\alpha}{2m}} \cdot S_P \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \qquad (12\text{-}9)$$

其中 $S_P = \sqrt{MSE}$，$m =$ 兩兩平均數差的組數，且 $t_{\frac{\alpha}{2m}}$ 表示自由度為 $n-k$ 之 t 分配的上界點 $\frac{\alpha}{2m}$。

使用此方法，保證所有 m 組（兩兩平均數差）信賴區間同時成立的機率至少為 $(1-\alpha)$。

Scheffe 聯合信賴區間

註：Scheffe公式算出的信賴區間寬度比較寬，也就是更保守。除非效果之間的差異夠大，否則不輕易拒絕 H0。

例題 11.4

為了解價格對銷售量的影響，乃進行一項試銷實驗，得銷售量如下：

高價格	50	43	48
中價格	54	60	51
低價格	72	66	60

假設本資料適合變異數分析，試在 $\alpha = 0.05$ 下，價格改變是否會影響到銷售量的不同？

解：(1) 假設：$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 \\ H_1 : \mu_i \text{ 不全等} \end{cases}$

(2) 顯著水準：$\alpha = 0.05$

(3) 計算平方和：

$$\bar{y}_{1\bullet} = \frac{50+43+48}{3} = 47,\quad \bar{y}_{2\bullet} = \frac{54+60+51}{3} = 55$$

$$\bar{y}_{3\bullet} = \frac{72+66+60}{3} = 66,\quad \bar{\bar{y}} = \frac{50+43+\cdots+66+60}{9} = 56$$

$$SSTo = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{\bar{y}})^2 = (50-56)^2 + (43-56)^2 + \cdots + (60-56)^2 = 686$$

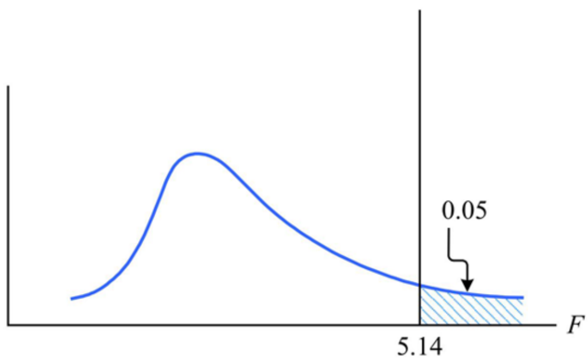
$$SSB = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{i\bullet} - \bar{\bar{y}})^2 = 3(47-56)^2 + 3(55-56)^2 + 3(66-56)^2 = 546$$

$$SSW = SSTo - SSB = 686 - 546 = 140$$

(4) 變異數分析表：

變異來源	平方和	自由度	均方和	F 值
價　　格	546	2	273	11.70
誤　　差	140	6	23.33	
總　　和	686	8		

(5) 拒絕域：$F > F_{2,6,0.05} = 5.14$



(6) 判斷：$F = 11.70 > 5.14$，在拒絕域中，差異顯著，所以拒絕 H_0，即表示不同價格會影響銷售量的不同。

END

例題 11.6 （承例題 11.4）

設信賴係數 $1-\alpha=0.95$：

(1) 在例題 11.4 的結論後，是否需要進行多重比較？若需要，試以薛費法 (Scheffe' method) 求各母體平均數差之同時信賴區間的估計，並說明檢定上的涵義。

(2) 若只考慮高價格與中價格的比較，$\mu_1-\mu_2$ 的 95% 信賴區間。

(3) 若只考慮高價格的資料，μ_1 的 95% 信賴區間。

解：(1) 變異數分析之 F 值顯著，表示三種價格間銷量有差異，可進行多重比較以確定哪兩個母體平均數對有差異。其聯合信賴區間為：（共 $C_2^3=3$ 對）

$$\mu_1-\mu_2:\quad (47-55)\pm\sqrt{(3-1)5.14}\sqrt{23.33}\sqrt{\frac{1}{3}+\frac{1}{3}}=-8\pm12.64$$

$$\mu_1-\mu_3:\quad (47-66)\pm\sqrt{(3-1)5.14}\sqrt{23.33}\sqrt{\frac{1}{3}+\frac{1}{3}}=-19\pm12.64$$

$$\mu_2-\mu_3:\quad (55-66)\pm\sqrt{(3-1)5.14}\sqrt{23.33}\sqrt{\frac{1}{3}+\frac{1}{3}}=-11\pm12.64$$

列表如下：

價格水準	高價格	中價格	低價格
高 價 格	—	-8 ± 12.64	-19 ± 12.64*
中 價 格		—	-11 ± 12.64
低 價 格			—

* 表示有顯著差異

由上表可知，$\mu_1-\mu_3$ 之信賴區間不包括 0，可能 μ_1 與 μ_3 不相等，即有顯著差異，而 $\mu_1-\mu_2$ 與 $\mu_2-\mu_3$ 皆包括 0，無顯著差異，即 $\mu_1=\mu_2$ 和 $\mu_2=\mu_3$ 可能發生。

(2) 個別 $\mu_1-\mu_2$ 的 95% 信賴區間：$(t_{6,0.975}=2.45)$

$$\mu_1-\mu_2:\quad (47-55)\pm2.45\sqrt{23.33}\sqrt{\frac{1}{3}+\frac{1}{3}}=-8\pm9.66$$

$$-17.66\le\mu_1-\mu_2\le1.66$$

$\mu_1-\mu_2$ 的 95% 信賴區間包括 0，因此 μ_1 與 μ_2 可能相等。

(3) 個別 μ_1 的 95% 信賴區間：

$$\mu_1:\quad 47\pm2.45\sqrt{\frac{23.33}{3}}=47\pm6.83$$

$$40.17\le\mu_1\le53.83$$

END

3. 二因子變異數分析

		實驗因子 A			
		因子水準 A_1	因子水準 A_2	…	因子水準 A_c
集區變數 B	集區 B_1	Y_{11}	Y_{12}	…	Y_{1c}
	集區 B_2	Y_{21}	Y_{22}	…	Y_{2c}
	⋮	⋮	⋮	⋮	⋮
	集區 B_r	Y_{r1}	Y_{r2}	…	Y_{rc}

隨機化區集設計

隨機集區設計 (Randomized block design, RBD) 是指依據某一外在干擾變數，將實驗單位分成若干「集區」(block)，然後再觀察實驗因子之處理效果。

- 實驗因子之處理影響效果，即各水準處理間是否有顯著差異。
- 集區變數之集區影響效果，即各集區間是否有顯著差異。

隨機化區集設計之資料結構

區集	處理方式					列或區集和
	1	2	3	$\cdots$	k	
1	X_{11}	X_{12}	X_{13}	$\cdots$	X_{1k}	$B_1 = \sum_{j=1}^{k} X_{1j}$
2	X_{21}	X_{22}	X_{22}	$\cdots$	X_{2k}	$B_2 = \sum_{j=1}^{k} X_{2j}$
3	X_{31}	X_{32}	X_{33}	$\cdots$	X_{3k}	$B_3 = \sum_{j=1}^{k} X_{3j}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
b	X_{b1}	X_{b2}	X_{b3}	$\cdots$	X_{bk}	$B_b = \sum_{j=1}^{k} X_{bj}$
行或處理和	$T_1 = \sum_{i=1}^{b} X_{i1}$	$T_2 = \sum_{i=1}^{b} X_{i2}$	$T_3 = \sum_{i=1}^{b} X_{i3}$	$\cdots$	$T_k = \sum_{i=1}^{b} X_{ik}$	總和 $T = \sum_{j=1}^{k} \sum_{i=1}^{b} X_{ij}$

- $B_i = \sum_{j=1}^{k} X_{ij} =$ 第i個區集的觀察值總和
- $\bar{B}_i = \frac{B_i}{k} =$ 第i個區集的樣本平均數
- $T_j = \sum_{i=1} X_{ij} =$ 第j種處理方式的觀察值總和
- $\bar{T}_j = \frac{T_j}{b} =$ 第j種處理方式之樣本平均數
- $T = \sum_{j=1}^{k} \sum_{i=1}^{b} X_{ij} =$ 所有觀察值的總和
- $\bar{\bar{T}} = \frac{T}{bk} =$ 總平均數 $=$ 所有樣本的平均數

總差異可分解如下：

總差異 = 處理方式差異 + 區集因素差異 + 殘差

$$X_{ij} - \bar{\bar{X}} = (\bar{T}_j - \bar{\bar{T}}) + (\bar{B}_i - \bar{\bar{X}}) + (X_{ij} - \bar{T}_j - \bar{B}_i + \bar{\bar{X}})$$

總變異 = 處理方式變異 + 區集因素變異 + 殘差變異

$$\sum_{j=1}^{k}\sum_{i=1}^{b}(X_{ij}-\bar{\bar{X}})^2 = b\sum_{j=1}^{k}(\bar{T}_j-\bar{\bar{X}})^2 + k\sum_{i=1}^{b}(\bar{B}_i-\bar{\bar{X}})^2 + \sum_{j=1}^{k}\sum_{i=1}^{b}(X_{ij}-\bar{T}_j-\bar{B}_i +$$

總平方和（**SST**）= 處理方式平方和（**SSTR**）＋區集因素平方和（**SSBL**）＋殘差平方和（**SSE**）

$SST = \sum_j\sum_i(X_{ij}-\bar{\bar{X}})^2 = \sum_j\sum_i X_{ij}^2 - \frac{T^2}{n}$, 自由度為$n-1$

$SSTR = b\sum j(\bar{T}_j-\bar{\bar{T}})^2 = \frac{\sum T_j^2}{b} - \frac{T^2}{n}$, 自由度為$k-1$

$SSBL = b\sum j(\bar{B}_i-\bar{\bar{T}})^2 = \frac{\sum B_i^2}{k} - \frac{T^2}{n}$, 自由度為$b-1$

$SSE = SST - SSTR - SSBL$, 自由度為$(k-1)(b-1)$

Summary: 隨機化區集設計之**ANOVA**表

變異來源	平方和 (SS)	自由度 (df)	均方 (MS)	F 比值
處理方式	$SSTR$	$k-1$	$MSTR=\frac{SSTR}{k-1}$	$F_1=\frac{MSTR}{MSE}$
區集因素	$SSBL$	$b-1$	$MSBL=\frac{SSBL}{b-1}$	$F_2=\frac{MSBL}{MSE}$
殘差	SSE	$(k-1)(b-1)$	$MSE=\frac{SSE}{(k-1)(b-1)}$	
總和	SST	$n-1=kb-1$		

兩個虛無假設：處理方式與區集因素的F檢定

隨機化區集設計的 F 檢定

$$H_0:\mu_1=\mu_2=\cdots=\mu_k$$
$$H_0':\mu_1'=\mu_2'=\cdots=\mu_b'$$

若 $F_1=\frac{MSTR}{MSE}\geq F_\alpha[k-1,\,(k-1)(b-1)]$，則拒絕 H_0；

若 $F_2=\frac{MSBL}{MSE}\geq F_\alpha[b-1,\,(k-1)(b-1)]$，則拒絕 H_0'。

例題 11.8

某公司為了解不同價格調整是否會影響銷售量，乃在台北市進行一項實驗。為考慮到不同商店類型可能會影響銷售量，於是選擇四種不同類型的商店各 3 家，每家只按一種價格水準出售產品，經一星期後，各家商店的銷售量如下表所示：

商店類型（集區）	價　格		
	高	中	低
超　　市	54	60	78
超　　商	54	63	60
雜 貨 店	51	54	69
冷 飲 店	45	63	81

假設該資料符合變異數分析的假設，$\alpha = 0.05$，試求：

(a) 不同的價格調整，對銷售量是否有顯著影響？

(b) 不同的商店類型，對銷售量是否有顯著影響？

解：(1) 假設：
$$\begin{cases} H_0 : \mu_{\bullet 1} = \mu_{\bullet 2} = \mu_{\bullet 3} \text{（價格變數）} \\ H_1 : \mu_{\bullet j} \text{ 不全等} \end{cases}$$
$$\begin{cases} H_0' : \mu_{1\bullet} = \mu_{2\bullet} = \mu_{3\bullet} = \mu_{4\bullet} \text{（集區變數）} \\ H_1' : \mu_{i\bullet} \text{ 不全等} \end{cases}$$

(2) 顯著水準：$\alpha = 0.05$

(3) 拒絕域：$F_c > F_{2,6,0.05} = 5.14$

$F_r > F_{3,6,0.05} = 4.76$

(4) 計算：

① 行：$\bar{Y}_{\bullet 1} = \dfrac{204}{4} = 51$, $\bar{Y}_{\bullet 2} = \dfrac{240}{4} = 60$, $\bar{Y}_{\bullet 3} = \dfrac{288}{4} = 72$

② 列：$\bar{Y}_{1\bullet} = \dfrac{192}{3} = 64$, $\bar{Y}_{2\bullet} = \dfrac{177}{3} = 59$, $\bar{Y}_{3\bullet} = \dfrac{174}{3} = 58$, $\bar{Y}_{4\bullet} = \dfrac{189}{3} = 63$

③ 總平均：$\bar{\bar{Y}} = \dfrac{54 + 54 + \cdots + 69 + 81}{12} = 61$

$$SSTo = (54-61)^2 + (60-61)^2 + \cdots + (81-61)^2 = 1{,}266$$
$$SSc = 4[(51-61)^2 + (60-61)^2 + (72-61)^2] = 888$$
$$SSr = 3[(64-61)^2 + (59-61)^2 + (58-61)^2 + (63-61)^2] = 78$$
$$SSE = 1{,}266 - 888 - 78 = 300$$

(5) 變異數分析表：

變異來源	平方和	自由度	均方和	*F* 值
價格變數	888	2	444	8.88
集區變數	78	3	26	0.52
誤　　差	300	6	50	
總　　和	1266	11		

(6) 判斷：

(a) $8.88 > 5.14$，差異顯著，拒絕 H_0，即不同價格水準，其銷售量有顯著差異。

(b) $0.52 < 4.76$，差異不顯著，不拒絕 H_0'，商店類型不同，其銷售量無顯著差異。 END

例題 11.9

下列資料為四位學生英文、數學、會計學、統計學之期末考成績，設資料符合變異數分析的假設，在 $\alpha = 0.05$ 之下，試檢定：

(a) 各科之難易程度是否相同？

(b) 學生之能力程度是否相等？

學生＼科目	英　文	數　學	統計學	會計學
1	78	62	71	77
2	57	49	62	60
3	69	78	72	83
4	71	66	59	67

解：(1) 假設：

$$\begin{cases} H_0 : \mu_{\bullet 1} = \mu_{\bullet 2} = \mu_{\bullet 3} = \mu_{\bullet 4} \text{ (科目變數)} \\ H_1 : \mu_{\bullet j} \text{ 不全等} \end{cases}$$

$$\begin{cases} H_0' : \mu_{1\bullet} = \mu_{2\bullet} = \mu_{3\bullet} = \mu_{4\bullet} \text{ (學生變數)} \\ H_1' : \mu_{i\bullet} \text{ 不全等} \end{cases}$$

(2) 顯著水準：$\alpha = 0.05$

(3) 拒絕域：$F_c > F_{3,9,0.05} = 3.86$

$F_r > F_{3,9,0.05} = 3.86$

(4) 計算：

① 行：$\overline{Y}_{\bullet 1} = \dfrac{275}{4} = 68.75$, $\overline{Y}_{\bullet 2} = \dfrac{255}{4} = 63.75$

$\overline{Y}_{\bullet 3} = \dfrac{264}{4} = 66$, $\overline{Y}_{\bullet 4} = \dfrac{287}{4} = 71.75$

② 列：$\overline{Y}_{1\bullet} = \dfrac{288}{4} = 72$, $\overline{Y}_{2\bullet} = \dfrac{228}{4} = 57$

$\overline{Y}_{3\bullet} = \dfrac{302}{4} = 75.5$, $\overline{Y}_{4\bullet} = \dfrac{263}{4} = 65.75$

③ 總平均：$\overline{\overline{Y}} = \dfrac{78+57+\cdots+83+67}{16} = 67.56$

$$SSTo = (78-67.56)^2 + (62-67.56)^2 + \cdots + (67-67.56)^2 = 1241.94$$

$$SSc = 4[(68.75-67.56)^2 + (63.75-67.56)^2 + (66-67.56)^2 + (71.75-67.56)^2] = 143.69$$

$$SSr = 4[(72-67.56)^2 + (57-67.56)^2 + (75.5-67.56)^2 + (65.75-67.56)^2] = 790.19$$

$$SSE = 1,241.94 - 143.69 - 790.19 = 308.06$$

(5) 變異數分析表：

變異來源	平方和	自由度	均方和	*F* 值
科　　目	143.69	3	47.90	1.40
學　　生	790.19	3	263.40	7.70
誤　　差	308.06	9	34.23	

總　　和	1241.94	15		

(6) 判斷：

(a) $1.40 < 3.86$，不在拒絕域，差異不顯著，不拒絕 H_0，即科目難易程度並無顯著不同。

(b) $7.70 > 3.86$，在拒絕域中，差異顯著，拒絕 H_0'，即學生能力程度有顯著的不同。

END

```python

import pandas as pd
import statsmodels.stats.anova as anova
from statsmodels.formula.api import ols

PSID=pd.read_csv('PSID.csv')
PSID.head(3)

#進行單因子變異數分析

model = ols('earnings ~ C(educatn)', data = PSID.dropna()).fit()
table1 = anova.anova_lm(model)
print(table1)

#進行雙因子變異數分析

model = ols('earnings ~ C(married) + C(educatn)', data = PSID.dropna()).fit()
table2 = anova.anova_lm(model)
print(table2)

#進行析因變異數分析（交叉變項）

model = ols('earnings ~ C(married)*C(educatn)',\
            data = PSID.dropna()).fit()
table3 = anova.anova_lm(model)
print(table3)
```