Analysis of Variance (ANOVA) ~ Basic concept ~

變異數分析基本觀念

如果我們想知道不同品牌的差異……

若我們想要瞭解不同品牌的米、不同栽種法之間對產地產量的差異，今隨機選擇面積相同，條件相似的12塊田地做實驗，得到以下數據資料：

單位：千公斤

栽種法 \ 品牌	池上米	銀川米	越光米
一般施肥	8	3	7
自然農法	10	4	8
溫室栽培	6	5	6
有機施肥	8	4	7

母體？ 因子？ 如何比較差異？

“因子”與母體、樣本之關係

因子：
“栽種法”

因子：
“品牌”

池上米

池上地區
(母體A)

抽樣實驗區(樣本a)

台東池上（池上米）

銀川米

富里地區
(母體B)

抽樣實驗區(樣本b)

花蓮富里（銀川米）

越光米

六甲地區
(母體C)

抽樣實驗區(樣本c)

台南六甲（越光米）

一般施肥

自然農法

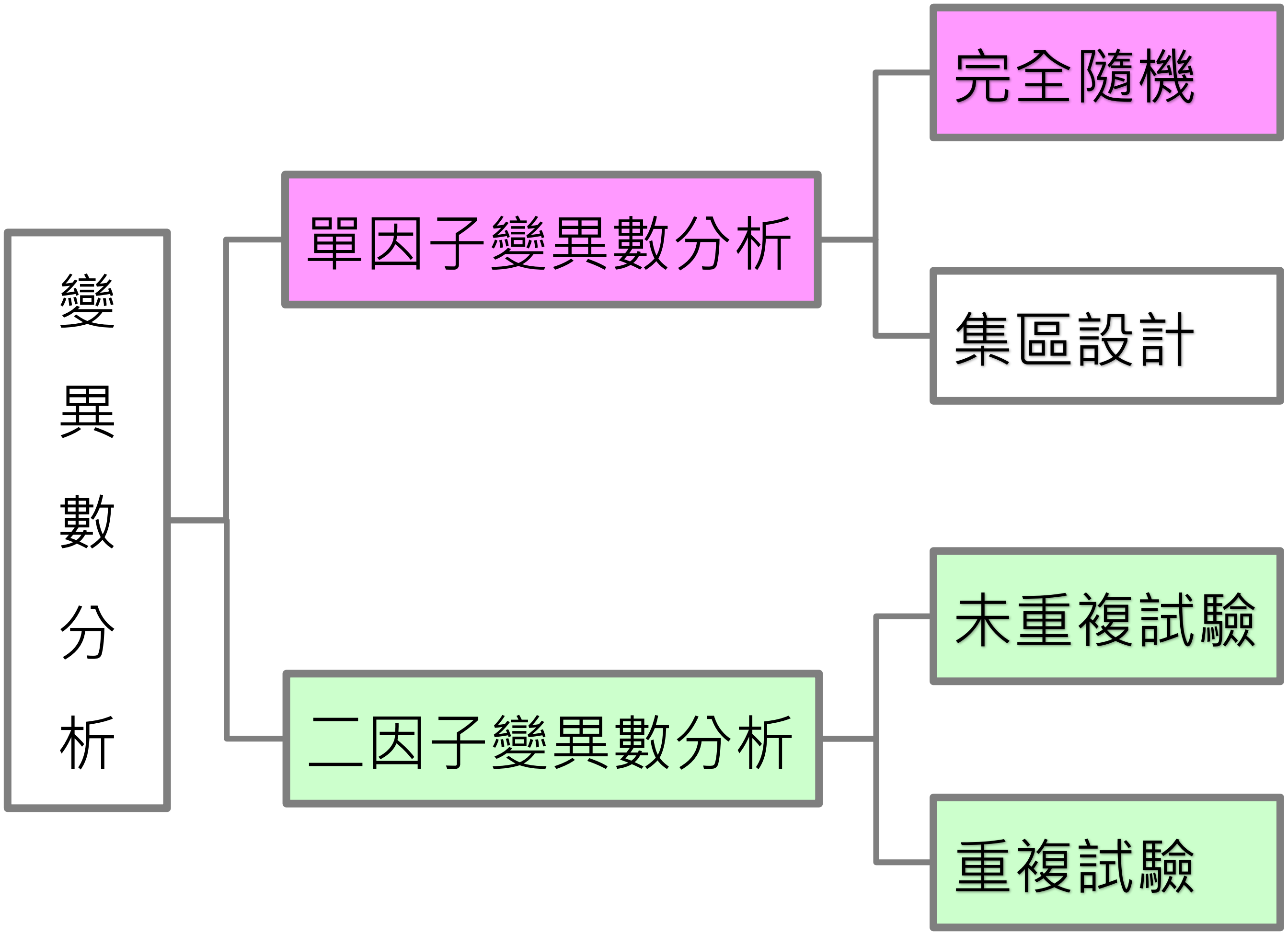
溫室栽培

有機施肥

一	溫
自	有

（兩因子變異數分析）

變異數分析類型



基本觀念

關於變異數分析

●檢定三個或三個上的母體平均數是否相等的檢定方法；或檢定因子（factor）對依變數（dependent variable）是否有影響的統計方法。

→若是利用z檢定或t檢定，兩兩比較檢定，不僅要檢定 C_2^n 次，同時也會導致型 I 錯誤的機率增加，所以必須同時一起檢定。

→檢定3個以上平均數是否相等，也就是與檢定變異數是否有差異，具有相同的意義，所以本章節雖然主要檢定各〝母體的平均數〞，但利用檢定變異數的「F檢定」分析，故稱為「變異數分析」。

●通常在人文社會科學的研究中，會針對研究主題來規劃「實驗設計」，來探究或是檢定所推論的結果。其所使用的統計檢定方式，常就以「變異數分析」來進行推論。

關於因子、依變數

- 所謂「因子」(factor)是引起資料發生差異的原因，也有稱之為「獨立變數」（independent variable）、「自變數」、「實驗變數」；通常其產生有可能來自：

  (1)樣本屬性、特質的不同類別(例如〝性別〞可分為男、女；〝職業〞可分為軍、公、教、農、工、商等）；

  (2)來自實驗設計不同的處理類型，通常是想要透過實驗調查結果來瞭解所要釐清的問題。

- 而「依變數」(dependent variable) 是調查者欲觀察到的反應結果(反應變數）。例如，不同栽種法對於農地產量的差異，產量就是依變數；變異數分析就是在檢定某因子（如栽種法）狀況下，不同衡量水準下依變數(產量）的平均數是否相等。

變異數分析的基本假設

- **常態性假設**：每個被分析中的母體皆為常態分配。
- **同質性（homogeneity）假設**：每個被分析的母體，其變異數皆相等。
- **獨立性假設**：在進行變異數分析的抽樣過程中，各個母體分配的樣本皆相互獨立，抽樣方式為簡單隨機抽樣。

滿足以上三個基本前提，就可以利用**F檢定**來做推論。

若是所分析母體無法滿足上面的前提假設，就改用「**無母數統計**」方式進行檢定。

變異數分析的步驟建議

第1步

設立兩個假設：

$$\begin{cases} \mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_{\mathrm{k}} = \mu \\ \mathrm{H}_1 : \mu_1, \mu_2, \cdots\cdots \mu_{\mathrm{k}} \end{cases}$$

第2步

建立變異數分析表（ANOVA table），並利用「F檢定」做檢定：

$$F^* = \frac{MSA}{MSE}$$

第3步

下結論並做決策或推論：

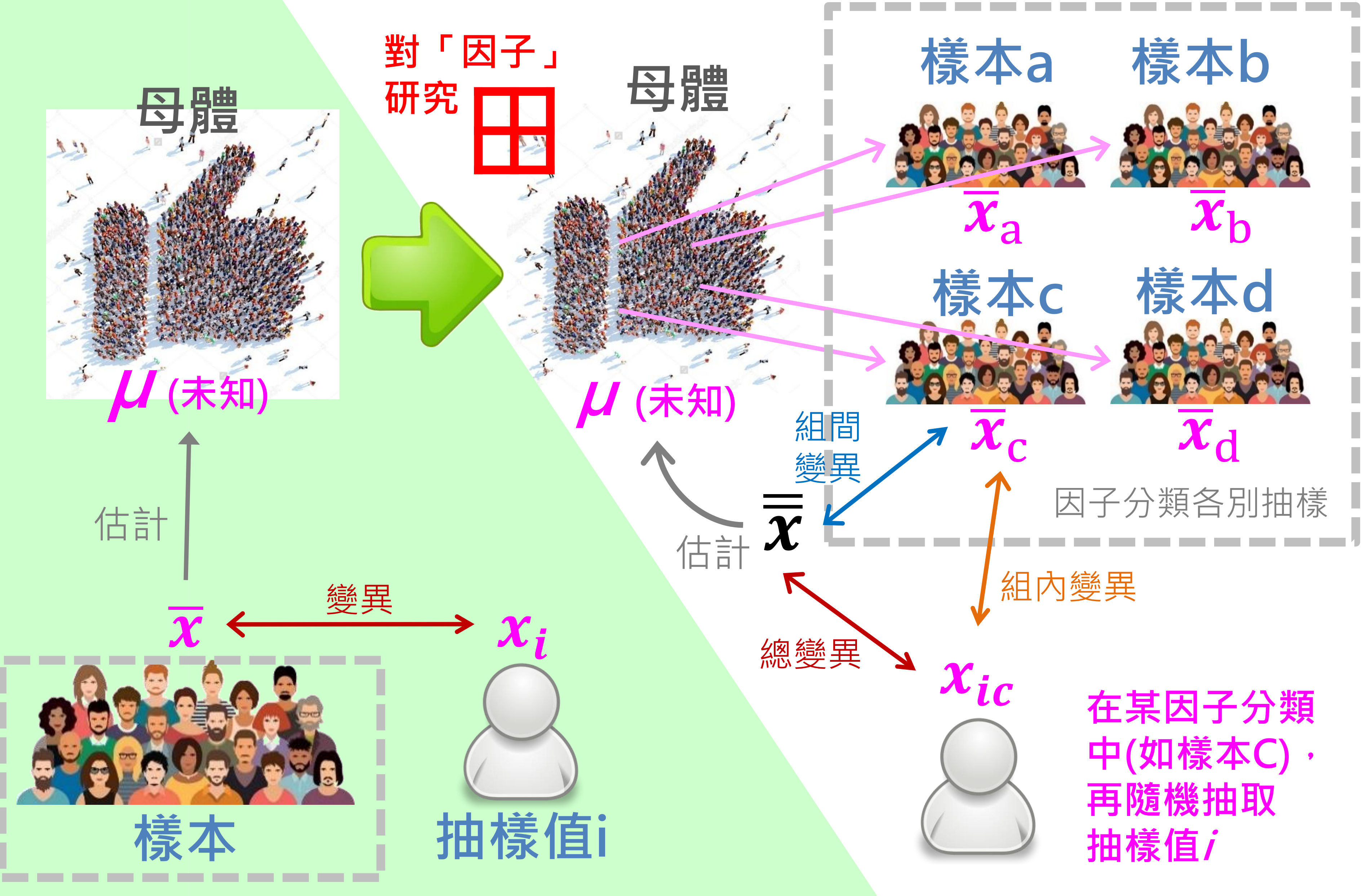
$$F^* > F_{k-1, n_T-k, \alpha} \rightarrow \text{拒絕}\mathrm{H}_0$$

$$F^* \leq F_{k-1, n_T-k, \alpha} \rightarrow \text{不拒絕}\mathrm{H}_0$$

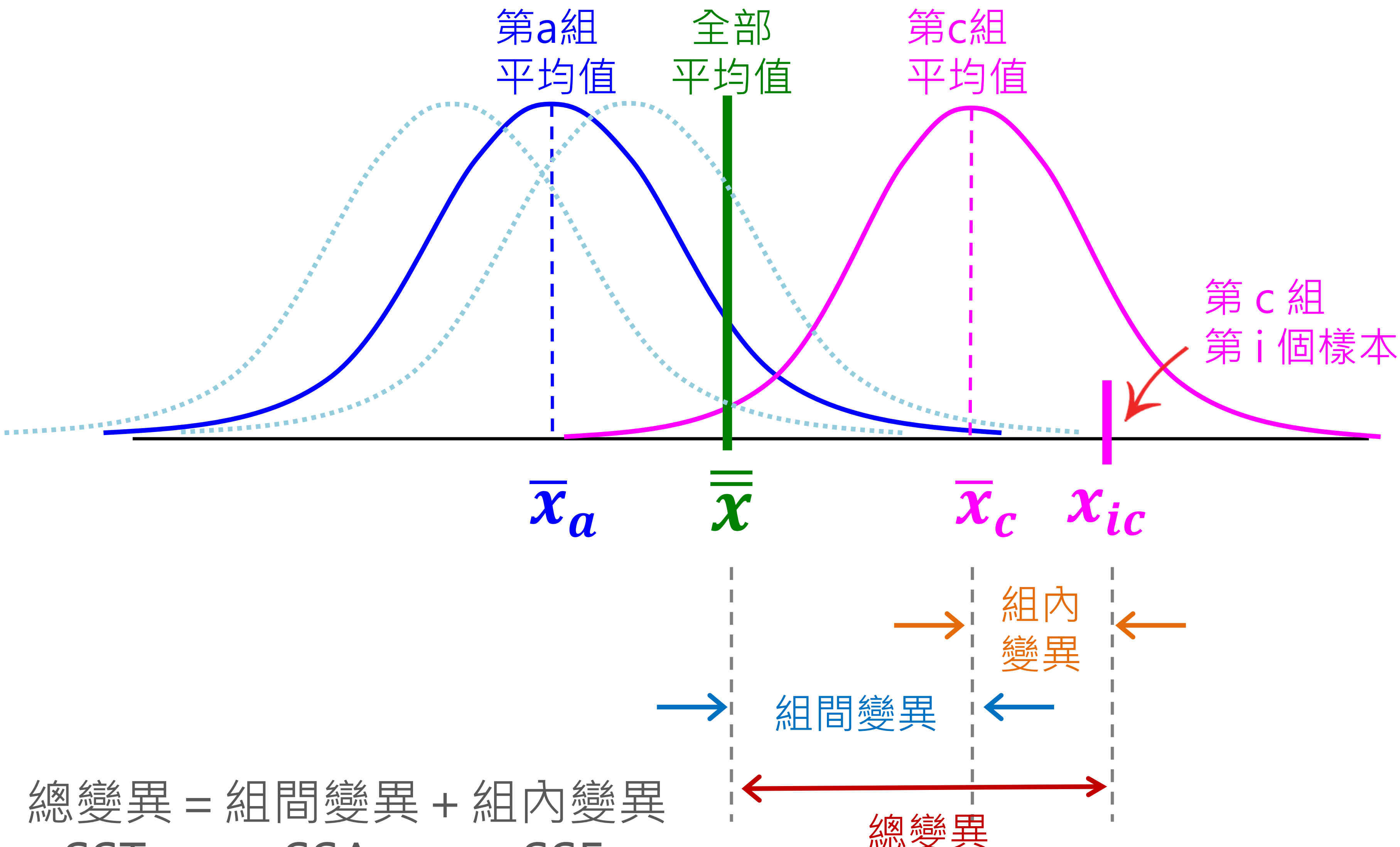
再談觀念
~ 單因子變異數分析

組間變異（Among-group variation）
組內變異（Within-group variation）

組內與組間變異的關連性



組內與組間變異的示意圖



總變異 = 組間變異 + 組內變異

SST = SSA + SSE

組內與組間變異的推導

$$x_{ij} - \bar{\bar{x}} = (\bar{x}_{\mathrm{j}} - \bar{\bar{x}}) + (x_{ij} - \bar{x}_{\mathrm{j}})$$

單一個抽樣值的變異（誤差）

$$\Rightarrow \sum_{\mathrm{j}=1}^{\mathrm{k}} \sum_{\mathrm{i}=1}^{\mathrm{n_j}} (x_{ij} - \bar{\bar{x}})^2 = \sum_{\mathrm{j}=1}^{\mathrm{k}} \sum_{\mathrm{i}=1}^{\mathrm{n_j}} \left[(\bar{x}_{\mathrm{j}} - \bar{\bar{x}}) + (x_{ij} - \bar{x}_{\mathrm{j}})\right]^2$$

為算出總誤差，兩邊取平方和，此誤差即稱為「變異」

$$\Rightarrow \quad = \sum_{\mathrm{j}=1}^{\mathrm{k}} \sum_{\mathrm{i}=1}^{\mathrm{n_j}} (\bar{x}_{\mathrm{j}} - \bar{\bar{x}})^2 + 2 \sum_{\mathrm{j}=1}^{\mathrm{k}} \sum_{\mathrm{i}=1}^{\mathrm{n_j}} (\bar{x}_{\mathrm{j}} - \bar{\bar{x}})(x_{ij} - \bar{x}_{\mathrm{j}}) + \sum_{\mathrm{j}=1}^{\mathrm{k}} \sum_{\mathrm{i}=1}^{\mathrm{n_j}} (x_{ij} - \bar{x}_{\mathrm{j}})^2$$

其中，$\sum_{\mathrm{j}=1}^{\mathrm{k}} \sum_{\mathrm{i}=1}^{\mathrm{n_j}} (\bar{x}_{\mathrm{j}} - \bar{\bar{x}})(x_{ij} - \bar{x}_{\mathrm{j}}) = \sum_{\mathrm{j}=1}^{\mathrm{k}} (\bar{x}_{\mathrm{j}} - \bar{\bar{x}}) \sum_{\mathrm{i}=1}^{\mathrm{n_j}} (x_{ij} - \bar{x}_{\mathrm{j}}) = \sum_{\mathrm{j}=1}^{\mathrm{k}} (\bar{x}_{\mathrm{j}} - \bar{\bar{x}}) \times 0 = 0$

$$\Rightarrow \quad = \sum_{\mathrm{j}=1}^{\mathrm{k}} \sum_{\mathrm{i}=1}^{\mathrm{n_j}} (\bar{x}_{\mathrm{j}} - \bar{\bar{x}})^2 + \sum_{\mathrm{j}=1}^{\mathrm{k}} \sum_{\mathrm{i}=1}^{\mathrm{n_j}} (x_{ij} - \bar{x}_{\mathrm{j}})^2$$

組內與組間變異的推導

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{\bar{x}})^2=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{x}_j-\bar{\bar{x}})^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2$$

SST(SSTO)	SSA(SSB, SSF, SSTR)	SSE(SSW)
sum of squares total	sum of squares among groups sum of square due to factor sum of square treatment	sum of squares error sum of square within groups
總變異	因子之變異 (組間變異)	隨機變異 (組內變異)

總變異 = 組間變異 + 組內變異

SST = SSA + SSE

各變異的算式關連

SST (總變異，sum of squares total)

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}-\bar{\bar{x}}\right)^2 = n_T \cdot \left[\frac{1}{n_T}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}-\bar{\bar{x}}\right)^2\right] = n_T \cdot \boldsymbol{\sigma}_T^2 = (n_T - 1)S_T^2$$

SSA(組間變異，sum of squares among groups)

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\bar{x}_j-\bar{\bar{x}}\right)^2 = \sum_{j=1}^{k} n_j\bar{x}_j^2 - n_T\bar{\bar{x}}^2 = n \times k \times \left[\frac{1}{k}\sum_{j=1}^{k}\bar{x}_j^2 - \bar{\bar{x}}^2\right] = n_T \cdot \boldsymbol{\sigma}_{\bar{x}}^2$$

(假設每個抽樣樣本數都相同，$n=n_j$)

SSE (組內變異，sum of squares error)

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2 = \sum_{j=1}^{k} n_j\left[\frac{1}{n_j}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2\right] = \sum_{j=1}^{k} n_j \cdot \boldsymbol{\sigma}_j^2 = \sum_{j=1}^{k}(n_j-1)\cdot S_j^2$$

母體變異數的估計值

SSA ⟶ MSA

(組間變異) (樣本間均方)

(mean square between)

$$MSA = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{x}_j - \bar{\bar{x}})^2}{k-1} = \frac{SSA}{k-1}$$

自由度

SSE ⟶ MSE

(組內變異) (樣本內均方)

(mean square within)

$$MSE = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2}{n_T - k} = \frac{SSE}{n_T - k}$$

自由度

$$F^* = \frac{MSA}{MSE} \sim F_{k-1,\ n_T-k}$$

變異數分析表

以「單因子變異數分析」（完全隨機）為例

變異來源	平方和	自由度	平均平方和	F檢定
因子(組間)	SSA	k-1	MSA	$F^* = \dfrac{MSA}{MSE}$
隨機(組內)	SSE	n_T-k	MSE	
總和	SST	n_T-1		

還是觀念

～雙因子變異數分析

A因子組間變異（Among-group variation due to factor A）

B因子組間變異（ Among-group variation due to factor B）

兩因子間相互影響變異（Interaction-effect variation between factor A and B）

組內變異（Within-group variation）

雙因子變異數分析(未重複實驗)

- 探討兩個因子對於依變數之間的關係。
- 要做假設檢定，就是檢定A因子或B因子的衡量水準對依變數是否會造成顯著的影響。
- 所謂「未重複實驗」，實際的操作方式是：

在抽樣時，每個因子配對的衡量水準下，各抽取一個樣本做觀察。

假設我們要研究不同學區（北、中、南、東）與不同入學方式（繁星、申請、指考）對於商用統計學成績是否有影響，所以兩個因子（學區、入學方式）總共組成12種配對：（北，繁星）、（中，繁星）、（南，繁星）、.....、（南，指考）、（東，指考）。每種配對抽取一個樣本，總共就有12個樣本，此即為〝未重複實驗〞。

雙因子統計模式的建立(未重複實驗)

抽樣資料

因子	A_1	A_2	…	A_c	B平均
B_1	x_{11}	x_{12}	…	x_{1c}	$\bar{B}_1$
B_2	x_{21}	x_{22}	…	x_{2c}	$\bar{B}_2$
:	:	:	…	:	:
B_r	x_{r1}	x_{r2}	…	x_{rc}	$\bar{B}_r$
A平均	$\bar{A}_1$	$\bar{A}_2$	…	$\bar{A}_c$	$\bar{\bar{x}}$

其中 $\bar{A}_j = \dfrac{\sum_{j=1}^{r} x_{ij}}{r}$ $\quad \bar{\bar{x}} = \dfrac{\sum_{j=1}^{r}\sum_{i=1}^{c} x_{ij}}{r \times c} = \dfrac{\sum_{j=1}^{r}\sum_{i=1}^{c} x_{ij}}{n_T}$

$$\bar{B}_i = \frac{\sum_{i=1}^{c} x_{ij}}{c}$$

建立雙因子的假設檢定(未重複實驗)

A因子的假設檢定：

$$\begin{cases} H_0：\mu_{A_1} = \mu_{A_2} = \cdots = \mu_{A_c} \\ H_1：\mu_{A_1}, \mu_{A_2}, \cdots\cdots \mu_{A_c} \text{ 不全相等} \end{cases}$$

B因子的假設檢定：

$$\begin{cases} H_0：\mu_{\mathrm{B}_1} = \mu_{\mathrm{B}_2} = \cdots = \mu_{\mathrm{B}_c} \\ H_1：\mu_{\mathrm{B}_1}, \mu_{\mathrm{B}_2}, \cdots\cdots \mu_{\mathrm{B}_c} \text{ 不全相等} \end{cases}$$

雙因子各變異的分解(未重複實驗)

總變異 = A因子組間變異 + B因子組間變異 + 組內變異

SST = SSA + SSB + SSE

$$\sum_{j=1}^{r}\sum_{i=1}^{c}\left(x_{ij}-\bar{\bar{x}}\right)^2=\sum_{j=1}^{r}\sum_{i=1}^{c}\left(\bar{A}_j-\bar{\bar{x}}\right)^2+\sum_{j=1}^{r}\sum_{i=1}^{c}\left(\bar{B}_i-\bar{\bar{x}}\right)^2+\sum_{j=1}^{r}\sum_{i=1}^{c}\left(x_{ij}-\bar{A}_j-\bar{B}_i+\bar{\bar{x}}\right)^2$$

$$SST=\sum_{j=1}^{r}\sum_{i=1}^{c}\left(x_{ij}-\bar{\bar{x}}\right)^2=\sum_{j=1}^{r}\sum_{i=1}^{c}{x_{ij}}^2-n_T\bar{\bar{x}}^2=n_T\sigma_T^2$$

$$SSA=\sum_{j=1}^{r}\sum_{i=1}^{c}\left(\bar{A}_j-\bar{\bar{x}}\right)^2=\sum_{j=1}^{r}\sum_{i=1}^{c}\bar{A}_j^2-n_T\bar{\bar{x}}^2=n_T\sigma_{\bar{A}_j}^2$$

$$SSB=\sum_{j=1}^{r}\sum_{i=1}^{c}\left(\bar{B}_i-\bar{\bar{x}}\right)^2=\sum_{j=1}^{r}\sum_{i=1}^{c}\bar{B}_i^2-n_T\bar{\bar{x}}^2=n_T\sigma_{\bar{B}_i}^2$$

$SSE = SST - SSA - SSB$

其中 $n_T=rc$

雙因子變異數分析表(未重複實驗)

變異來源	平方和	自由度	平均平方和	F檢定
A因子	SSA	c-1	MSA= $\frac{SSA}{c-1}$	$F_A^* = \frac{MSA}{MSE}$
B因子	SSB	r-1	MSB= $\frac{SSB}{r-1}$	
隨機	SSE	(c-1)(r-1)	MSE= $\frac{SSE}{(c-1)(r-1)}$	$F_B^* = \frac{MSB}{MSE}$
總和	SST	n_T-1		

檢定A因子是否有影響： $\mathrm{F_A^*} > F_{\mathrm{r-1,(r-1)(c-1)},\,\alpha} \rightarrow$ 拒絕H_0

檢定B因子是否有影響： $\mathrm{F_B^*} > F_{\mathrm{c-1,(r-1)(c-1)},\,\alpha} \rightarrow$ 拒絕H_0

雙因子變異數分析(重複實驗)

- 探討兩個因子對於依變數之間的關係。
- 要做假設檢定，就是檢定A因子或B因子的衡量水準對依變數是否會造成顯著的影響。
- 所謂「重複實驗」，實際的操作方式是：

  在抽樣時，每個因子配對的衡量水準下，各抽取 n 個樣本做觀察。

假設我們要研究不同學區（北、中、南、東）與不同入學方式（繁星、申請、指考）對於商用統計學成績是否有影響，所以兩個因子（學區、入學方式）總共組成12種配對：（北，繁星）、（中，繁星）、（南，繁星）、.....、（南，指考）、（東，指考）。每種配對抽取10 個樣本，總共就有12×10個樣本，此即為〝重複實驗〞。

雙因子統計模式的建立(重複實驗)

因子	A_1	A_2	…	A_c	*B平均*
B_1	x_{111} x_{112} $\vdots$ x_{11n}	x_{121} x_{122} $\vdots$ x_{12n}	…	x_{1c1} x_{1c2} $\vdots$ x_{1cn}	$\bar{B}_1$
	$\overline{B_1A_1}$	$\overline{B_1A_2}$	…	$\overline{B_1A_c}$	
B_2	x_{211} $\vdots$ x_{21n}	x_{221} $\vdots$ x_{22n}	…	x_{2c1} $\vdots$ x_{2cn}	$\bar{B}_2$
	$\overline{B_2A_1}$	$\overline{B_2A_2}$	…	$\overline{B_2A_c}$	
$\vdots$	$\vdots$	$\vdots$	…	$\vdots$	$\vdots$
B_r	x_{r11} $\vdots$ x_{r1n}	x_{r21} $\vdots$ x_{r2n}	…	x_{rc1} $\vdots$ x_{rcn}	$\bar{B}_r$
	$\overline{B_rA_1}$	$\overline{B_rA_2}$	…	$\overline{B_rA_c}$	
A平均	$\bar{A}_1$	$\bar{A}_2$	…	$\bar{A}_c$	$\bar{\bar{x}}$

細格平均數
(cell means)

每列有c×n筆資料

每行有r×n筆資料

雙因子統計模式的建立(重複實驗)

其中 c = A因子的類別數　　　n = 重複數

r = B因子的類別數　　　n_T = 總樣本數 = c×r×n

x_{ijk} = A因子第 j 個水準、B因子第 i 個水準的第 k 個重複觀察值

$$\bar{A}_j = \frac{\sum_{\mathrm{i}=1}^{r}\sum_{\mathrm{k}=1}^{\mathrm{n}} x_{ij\mathrm{k}}}{r \times \mathrm{n}}$$

(第 j 行的總平均)

$$\bar{B}_i = \frac{\sum_{\mathrm{j}=1}^{\mathrm{c}}\sum_{\mathrm{k}=1}^{\mathrm{n}} x_{ij\mathrm{k}}}{c \times \mathrm{n}}$$

(第 i 行的總平均)

$$\overline{B_i A_j} = \frac{\sum_{k=1}^{n} x_{ijk}}{n}$$

(細格的平均數)

$$\bar{\bar{x}} = \frac{\sum_{\mathrm{i}=1}^{r}\sum_{\mathrm{j}=1}^{c}\sum_{\mathrm{k}=1}^{\mathrm{n}} x_{ij\mathrm{k}}}{r \times c \times \mathrm{n}}$$

(總平均)

建立雙因子的假設檢定(重複實驗)

A因子的假設檢定：

$$\begin{cases} H_0：\mu_{A_1}=\mu_{A_2}=\cdots=\mu_{A_c} \\ H_1：\mu_{A_1},\mu_{A_2},\cdots\cdots\mu_{A_c}\text{ 不全相等} \end{cases}$$

B因子的假設檢定：

$$\begin{cases} H_0：\mu_{\mathrm{B}_1}=\mu_{\mathrm{B}_2}=\cdots=\mu_{\mathrm{B}_c} \\ H_1：\mu_{\mathrm{B}_1},\mu_{\mathrm{B}_2},\cdots\cdots\mu_{\mathrm{B}_c}\text{ 不全相等} \end{cases}$$

AB因子交互效果檢定：

$$\begin{cases} H_0：A，B\text{因子具交互影響} \\ H_1：A，B\text{因子不具交互影響} \end{cases}$$

雙因子各變異的分解(重複實驗)

總變異 = A因子變異 + B因子變異 + A、B交互影響變異 + 組內變異

SST = SSA + SSB + SSAB + SSE

$$SST=\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}\left(x_{ijk}-\bar{\bar{x}}\right)^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}{x_{ijk}}^2-n_T\bar{\bar{x}}^2=n_T\sigma_T^2$$

$$SSA=\mathrm{rn}\sum_{j=1}^{c}\left(\bar{A}_j-\bar{\bar{x}}\right)^2=\mathrm{rn}\sum_{j=1}^{c}\bar{A}_j^2-n_T\bar{\bar{x}}^2=n_T\sigma_{\bar{A}_j}^2$$

$$SSB=\mathrm{cn}\sum_{i=1}^{r}\left(\overline{\mathrm{B}}_\mathrm{i}-\bar{\bar{x}}\right)^2=\mathrm{cn}\sum_{i=1}^{r}\overline{\mathrm{B}}_\mathrm{i}^2-n_T\bar{\bar{x}}^2=n_T\sigma_{\overline{\mathrm{B}}_\mathrm{i}}^2$$

$$SSE=\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}\left(x_{ijk}-\overline{\mathrm{A_jB_i}}\right)^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}\mathrm{x}_{\mathrm{ijk}}^2-\mathrm{n}\sum_{i=1}^{r}\sum_{j=1}^{c}\overline{\mathrm{A_jB_i}}^2$$

SSAB = SST - SSA – SSB - SSE

其中 $n_T=rcn$

雙因子變異數分析表(重複實驗)

變異來源	平方和	自由度	平均平方和	F檢定
A因子	SSA	c-1	MSA=$\frac{SSA}{c-1}$	$F_A^* = \frac{MSA}{MSE}$
B因子	SSB	r-1	MSB=$\frac{SSB}{r-1}$	$F_B^* = \frac{MSB}{MSE}$
交互影響	SSAB	(c-1)(r-1)	MSAB=$\frac{SSAB}{(c-1)(r-1)}$	$F_{AB}^* = \frac{MSAB}{MSE}$
隨機	SSE	rc(n-1)	MSE=$\frac{SSE}{rc(n-1)}$	
總和	SST	n_T-1		

檢定A因子是否有影響： $F_A^* > F_{c-1,rc(n-1),\alpha} \rightarrow$ 拒絕H_0

檢定B因子是否有影響： $F_B^* > F_{r-1,rc(n-1),\alpha} \rightarrow$ 拒絕H_0

檢定AB因子是否有交互影響：$F_{AB}^* > F_{(c-1)(r-1),rc(n-1),\alpha} \rightarrow$ 拒絕H_0

The End