CHAPTER

1

集群分析與多元尺度分析

【研究問題】

以企業組織的組織文化、組織知識管理、員工投入程度、生產效益等變因是否可將企業組織分成有意義的群組？上述研究問題的假設檢定中採用的統計方法爲「集群分析」（cluster analysis）。以國中學生的學習動機、學習態度與學業表現等變因，可否將學生分成有意義的群組？此種將觀察值樣本分類爲少數群組的程序稱爲集群分析，分類後群組間特性是互斥而非重疊。

壹、集群分析相關理論

集群分析也是一種多變量分析程序，其目的在於將資料分成幾個相異性最大的群組，而群組間的相似程度最高。研究者如果認為觀察值間並非全部同質，在資料探索分析方面，集群分析是一個非常有用的技巧。由於集群分析時，使用之分析方法不同，結果便有所不同，不同研究者對同一觀察值進行集群分析時，所決定的集群數也未必一致，因而集群分析較偏向於探索性分析方法，在研究應用上，常與區別分析一起使用。

觀察值之集群分析應用與區別分析相似，均在於將獨立分開的觀察值分成不同組別（groups）或將觀察值分類，二者主要差別在於區別分析時，組別特性已知；而集群分析時，觀察值所屬群組特性還未知。此外，在集群分析前，研究者尚不知道獨立觀察值可分為多少個群組（集群），其集群數不知道，而集群的特性也無從得知，集群分析法採用的「數值分類法」（numerical taxonomy），分類的準則並非是研究者事先決定的，此方面就是將計量空間的樣本點加以分組，分組後使在同一群組內的樣本點具有高度的相似性（similarity）／較高的同質性（homogeneity），不同群組間的樣本點則具有高度的異質性（heterogeneity），此種分類法是根據樣本點的計量屬性加以估計分類，是一種「自然分組法」（natural grouping），其關注的焦點不在於估計觀察值在變項上之變異量的差異；相反的，乃是利用觀察值在變項變異量的不同，對觀察值加以分組。

假如有十一個觀察值，根據其在變項的變異量的不同，分為三個群組，其圖示如下：

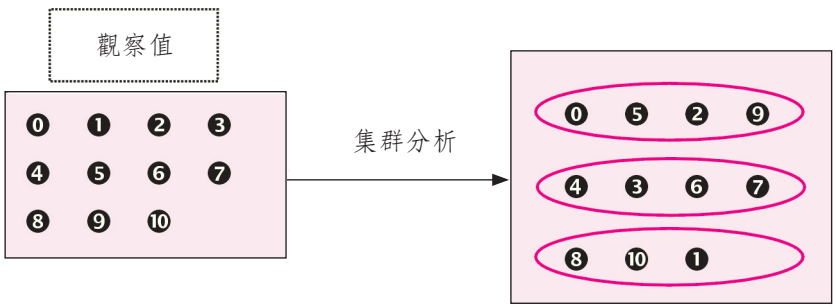

圖1-1

集群分析意義的圖示如下：左邊方框為所有觀察體的分布情形，零散而沒有意義，經由觀察體某些相似的變項性質，將具有類似性質的觀察體合併為一個集群，形成少數有意義而具有某種共同性質的群體。集群分析後，各群組中的觀察值具有最大相似性、各集群間具有最大的相異性。

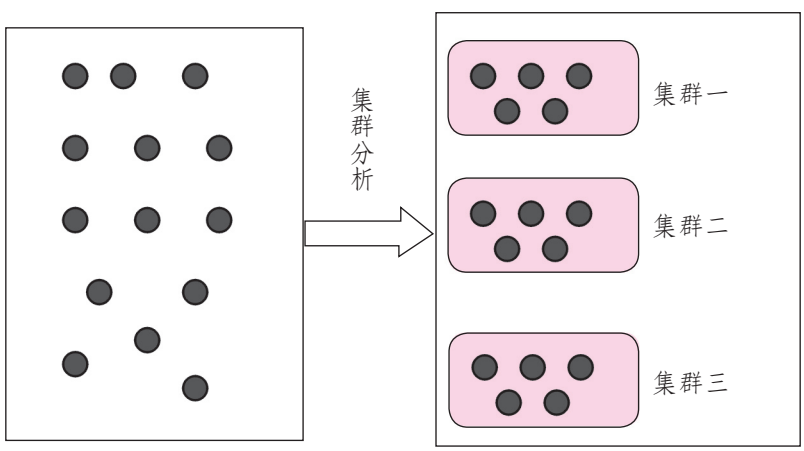

圖1-2

以教育程度及所得二個變項為例，六個樣本觀察值的假設資料如下表（*Sharma, 1996, p. 186*）：

表1-1

個人代號	所得（千元）	教育程度（年數）
S1	5	5
S2	6	6
S3	15	14

（續上頁表）

S4	16	15
S5	25	20
S6	30	19

上表數據依所得及教育程度二個變項繪製之散布圖如下，由二維空間之散布圖中可以看出：代號S1與代號S2為同一群組、代號S3與代號S4為同一群組、代號S5與代號S6為同一群組。集群成員{S5、S6}的所得最高、教育程度的年數也最長；集群成員{S1、S2}的所得最低、教育程度的年數也最短。

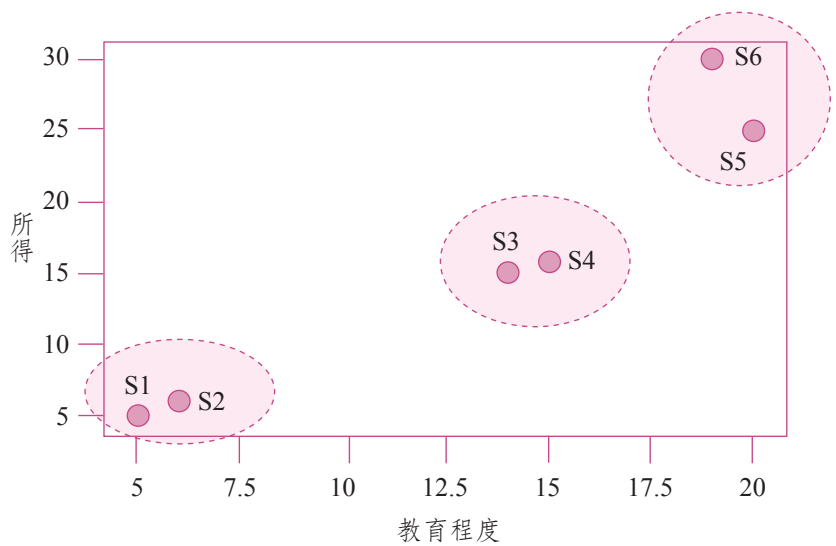

圖1-3

再以某企業組織受訓之二十八位學員的「學習動機」與「學習表現」二個變項來看，二十八名學員測得的數據如下。根據「學習動機」與「學習表現」二個變數測量值所繪製的散布圖如下：

表1-2

編號	學習動機	學習表現	編號	學習動機	學習表現
S1	10	2	S15	6	8
S2	10	8	S16	3	3
S3	8	9	S17	2	2

（續上頁表）

S4	9	10	S18	1	4
S5	8	10	S19	6	7
S6	5	5	S20	7	6
S7	1	3	S21	1	9
S8	2	2	S22	2	10
S9	3	1	S23	9	9
S10	6	6	S24	5	6
S11	3	9	S25	9	2
S12	2	8	S26	10	1
S13	1	10	S27	8	2
S14	10	10	S28	9	1

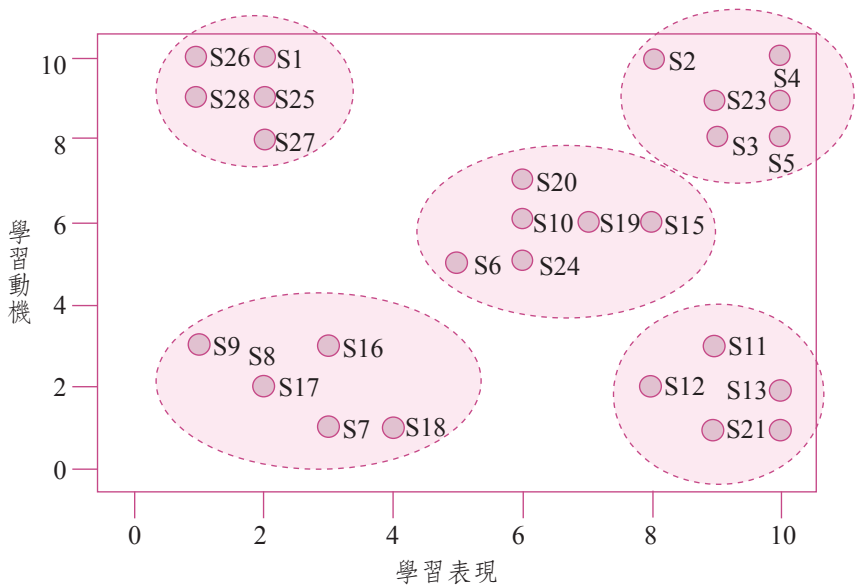

圖1-4

從上面的散布圖可以明顯看出，二十八名學員大致可以分為五個群組：群組[1]為學習動機低，學習表現也低者為{S7、S8、S9、S16、S17、S18}；群組[2]為學習動機低，但學習表現高者為{S11、S12、S13、S21、S22}；群組[3]為學習動機高，但學習表現低者為{S1、S25、S26、S27、S28}；群組[4]為學習動機高，且學習表現也高者為{S2、S3、S4、S5、S14、S23}；群組[5]為學習動機普通，且學習表現中等者為{S6、S10、S15、S19、S20、S24}。

變項的集群分析應用則相似於因素分析（factor analysis），二者進行的程序均在於辨認變項的相關組別。因素分析時，只有一個潛在的理論模式，而集群分析時則蘊涵著一個以上的潛在理論模式。多數實際應用時，二者的主要差別在於因素分析是針對「變項」予以分組；而集群分析則是將「觀察值個體」予以分組，亦即，因素分析時，根據依變項（題項）間之關係密切與否，將變項予以分群（分為幾個層面因素）；而集群分析則較常使用於將變項屬性相似程度較高的觀察值，加以分群，使集群與集群間的異質性達到最大，而同一集群內觀察值同質性很高。因素分析除可將變數分群外，也可與集群分析一樣將觀察值分群，使用因素分析技巧將觀察值分群的方法稱為「Q—因素分析」（Q-factor analysis），學者Sharma（*1996*）建議研究者最好不要使用「Q—因素分析」來進行觀察值的分類，因為這樣會產生其他問題。若是研究者關注的確認潛在因素與潛在因素的指標變項，最好就使用因素分析法，因為這是因素分析統計法被發展出來的主要緣由；如果研究者關注的觀察值／樣本點的分類，最適切的統計方法就是採用集群分析法。

在統計分析程序中，如果集群分析的對象是變項，則變項集群分析結果與變項因素分析結果，往往會有差異出現，其原因在於二者處理變項間關係方式不同，集群分析所採取的是一種「階層式」（hierarchical）的判別，依據個別變項間相關強弱情形逐次合併變項集群，而因素分析在聚合變項時，則是「同時」考量到所有變項間的關係。

集群分析方法，主要有二種，一為「階層式集群分析法」（hierarchical cluster analysis），二為「非階層式集群分析法」，非階層式集群分析法最常使用者為「K組平均法（K-Means集群分析法），如果觀察值的個數較多或資料檔非常龐大（通常觀察值在200個以上），以採用「K-Means集群分析法」較為適宜，因為觀察值數量太多，冰柱圖（icicle plots）與樹狀圖（dendrograms）二種判別圖形，在呈現時會過於分散，不易令人閱讀與解釋。使用「K-Means集群分析法」時，通常要訂定事先集群數目，進行分析次數可能較為多次，研究者可運用全體觀察值中部分數據進行「階層式集群分析法」，以作為決定集群數的參考。如果觀察值樣本不大，則採用「階層式集群分析法」較為適宜。

「階層式集群分析法」又可分為「凝聚法」或「聚合法」（algorithms method）與「分割法」（divisive method）二種，「凝聚法」又稱為「凝聚階層法」（agglomerative hierarchical methods），「分割法」又稱「分割階層法」（divisive hierarchical methods）：

一、聚合法

開始時，先計算N個觀察體在P個分類變項上之相似性資料，以得到一個N×N的相似性矩陣，此時將所有個別觀察體視為一個群組（N群），之後將二個相似性最大的觀察體合併為新的一個群組（變成N-1群），依次兩兩配對的方式，執行N-1次的合併，讓所有觀察體變成一個集群。此種方法有一個特性，即二個觀察體一旦被合併為同一個群組時，則之後所有合併的程序，這二個觀察體必會在同一個集群內。凝聚法依其計算方法的不同，包括單一連結法（single linkage）、完全連結法（complete linkage）、平均連結法（average linkage）、形心連結法（centroid liknage）與華德最小變異法（minimum variance method）等。

二、分割法

分割法的分類程序剛好與凝聚法相反，開始時將所有觀察值視為一個集群，之後依觀察值相異性最大或相似性最小者抽離出來（多數計算此觀察體與集群內其他觀察體的平均距離），將觀察體分割成二個集群，接著分別計算大集群中每個觀察體與集群內及集群外觀察體的平均距離，根據平均距離將大集群中的觀察體分割出來，依此步驟分割成三個集群、四個集群……，直到每一個觀察體單獨成一個集群為止。此種方割法較為複雜，因而在階層式集群分析法中較少使用。分割法與聚合法的集群程序，均可以以雙構面的樹狀圖表示，從樹狀圖中可以得知集群分割或聚合的詳細步驟。

非階層式集群分析法使用時，必須先將所有觀察體資料粗分為K個集群，與階層式集群分析法相較之下，集群組數目的分類剛好相反，階層式集群分析是根據觀察體間的相似性，逐步進行群組的合併與分類；而非階層式集群分析必須事先決定集群數目（K個群組），因而此種分類法又稱

為「K組平均集群法」。非階層式集群分析法的步驟如下（*Sharma, 1996, p. 202*；呂金河，*2005*）：

1. 選定K個初始集群的形心（centroids）或種子點（seeds），其中K是假定想要分群集群數目。
2. 計算每個觀察體到各集群形心距離遠近，將每一個觀察體分派到離其最近集群。
3. 根據事先假定的調整規則，重新分配或重新配置每一個觀察值到K組集群中。
4. 如果重新分配資料點能滿足調整規則條件，則重複步驟2、步驟3，直到資料點無法重新配置。

多數非層次法聚合結果會依下列二個條件而異：一為初始K個群集形心或種子點的設定；二為重新分派觀察值的調整規則。可見初始種子點的設定與分類結果有密不可分的關係，初始種子點的設定之常用方法有以下六種：

1. 選取前K個沒有遺漏值的觀察值作為初始集群的形心或種子點。
2. 先選取第一個沒有遺漏值的觀察值作為第一個集群的種子，第二個集群的種子則選取與第一個種子的距離超過某個特定標準者，第三個集群的種子則選取與先前二個種子的距離超過某個既定標準者。依此方法，直到選出K個集群的的種子。
3. 以隨機方式選出K個沒有遺漏值的觀察值作為集群的形心或種子。
4. 先選擇K個種子，次則依照某種特定規則（規則如種子間的距離儘可能夠遠）重新調整種子。
5. 使用簡單合理的方式確認集群的形心，使形心間的距離儘可能夠遠。
6. 使用研究者提供的種子。

一旦K個初始種子確定後，接著要將其他剩餘的N-K個觀察值分派到距離最近的種子點（可以形成一個集群）。非層次法聚合重新調整分群的方

法有以下三種規則（*Sharma, 1996, p. 203*）：

1. 重新計算每個集群的形心，將每個觀察值分派到距離最接近的形心集群中。在分派觀察值到K個集群的過程中，不更新形心的數值，直到所有觀察值都分派完，才重新計算集群的形心。若後一次集群形心與前一次集群形心的改變值大於某個收斂標準（convergence criterion），則重複之前的步驟，重新計算集群的形心並重新分派觀察值到距離最接近的形心集群中，直到後一次集群形心與前一次集群形心的改變值小於某個收斂標準。
2. 分群完後，重新計算每個集群的形心，並將每個觀察值分派到距離最接近的形心集群中，在分派觀察值到K個集群的過程中，每次均會重新計算觀察值加入集群前後二次形心的改變值。重複此步驟，直到形心的改變值小於某個特定的收斂的標準值。
3. 重新分派觀察值，以讓某種統計準則數值為最小值，這些方法一般稱為「爬坡法」（hill-climbing method），常用的統計準則或客觀函數為最小值的方法有以下四種：

⑴組內SSCP矩陣的跡（trace）（最小的ESS）（方陣中左上至右下主對主線元素的總和稱為矩陣的跡）。

⑵組內SSCP矩陣的行列式（determinant）值。

【備註】：行列式爲一個數值，是一個方陣所構成的面積或容積，求方陣的行列式可使用試算表函數「=MDETERM（陣列範圍）」求得。

$$\begin{vmatrix} 5 & 7 \\ 4 & 6 \end{vmatrix} = 5\times 6-(7\times 4) = 2 \text{、} \left[\begin{vmatrix} 2 & 3 & 3 \\ 2 & 5 & 7 \\ 3 & 4 & 6 \end{vmatrix}\right] = 10$$

⑶$W^{-1}B$矩陣的跡，其中W、B分別代表組內SSCP矩陣（$SSCP_W$）與組間SSCP矩陣。W^{-1}矩陣為矩陣W的反矩陣（inverse），W矩陣與其反矩陣有以下性質：$W^{-1}W = WW^{-1} = I$，矩陣I為單元矩陣（主對角線元素為1，其餘元素為0）。

【備註】：反矩陣（inverse matrix）的求法可以試算表函數語法

「=MINVERSE（原矩陣範圍）」求得，求出結果時要同時按『Ctrl』+『Shift』+『Enter』鍵。

$$\begin{bmatrix}3 & 4\\5 & 6\end{bmatrix}\begin{bmatrix}-3 & 2\\2.5 & -1.5\end{bmatrix}=\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$$，若$W=\begin{bmatrix}3 & 4\\5 & 6\end{bmatrix}$，則其反矩陣為$W^{-1}=\begin{bmatrix}-3 & 2\\2.5 & -1.5\end{bmatrix}$

⑷$W^{-1}B$矩陣的最大特徵值。

階層式集群分析法中，根據觀察值或變項間距離，將最相似物件結合在一起，以逐次聚合的方式（agglomerative clustering）將觀察值分組。計算觀察值相似性最常用的方法是歐基里得距離平方法（square Euclidean distance）。歐基里得距離平方法在計算觀值的相異程度時，會隨著測量單位不同而不同。在集群分析中，如果階層集群分析法與非階層集群分析法統合運用，則稱為「二階段集群分析法」。在非階層集群分析法中最讓研究者困惑者，為研究者必須主觀事前決定集群的數目，此種主觀的認定有時並非十分適切；若是研究者利用多次測試，從中發掘一組的集群數，有時時間又不許可；此外，在階層集群分析法中，以相似性聚合集群方法，之前被分派到同一集群的觀察體，在之後的分群中，會被一直歸於同一集群內，為獲得最佳的集群數，研究者可使用二階段集群分析法（two-stage cluster analysis）。

在二階段集群分析法中，研究者可採用相似性聚合的方法（如華德法或平均連結法），來獲得集群的群數與起始點，再以K組平均數法，以之前獲得的集群數作為K組平均數法中起始假定或主觀認定的集群數，此種統合方法，可以解決在非階層集群分析法中主觀假定集群數的問題，此外，也可以解決階層集群分析法中原觀察體無法重新分派或重新配置到其他集群的問題。

當研究者進行集群分析時，要考量到以下幾點（*SPSS Inc., 1998*）：

三、標準化程序

變項間單位如果不同，原始數值較大的變項，在距離測量演算程序的