NAG アドインとクラスター分析

このドキュメントでは、「Excel NAG 統計解析アドイン」（以下「NAG アドイン」と略称）を用いて行うことができる（階層的な）クラスター分析の方法を簡単に説明し、分析の結果として樹形図を作成するまでの具体的な操作例を示します。

目　　次

1　クラスター分析とは？

クラスター分析とは、対象（個体の場合もあるし、変数の場合もある）の集合に対して、対象間の測度（非類似度（距離など）又は類似度（相関係数など））※を定義することで、類似した対象が同じ部分集合（クラスターと呼ぶ）になるように分類する方法の総称です。方法を大別すると、階層的な方法と非階層的な方法に分けられます。

※　非類似度はその値が小さいほど類似性が高く、類似度はその値が大きいほど類似性が高いと考える。

このドキュメントでは、前者（階層的な方法）を取り上げます。以下「クラスター分析」と言った場合は「階層的な方法」のことを意味し、対象としては個体を、測度としては距離を想定します。また、一つの個体から成るクラスターを単に「個体」と呼ぶこともあります（文脈上混乱は生じないと思います）。

階層的な方法は、結果として樹形図（樹状図、デンドログラムとも言う）が得られる方法で、事前にクラスター数を定めることなく、個体間の階層的構造を求めることができます。この樹形図を適当な高さ（距離）で切断することにより個体をいくつかのクラスターに分け、そして各クラスターに含まれる個体を調べることによりクラスター（クラスター間）の特徴を把握することが分析の目的です。
（この「いくつのクラスターに分けるか」及び「クラスターの特徴は何か」については、数学的に定まった方法があるわけではなく、分析者自身の判断に委ねられます。）

個体	変数（観測値）	
	X_1	X_2
A	5	4
B	1	2
C	4	4
D	2	1
E	5	2

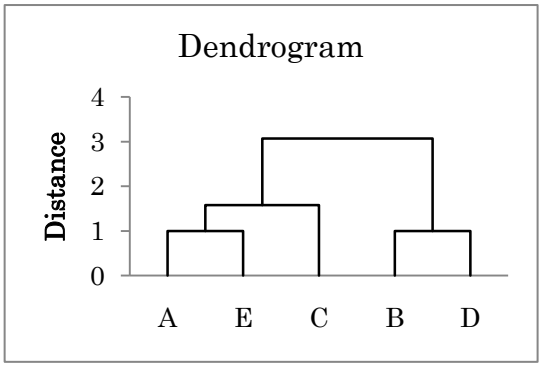


図１　（階層的な）クラスター分析のイメージ

2 距離の定義

個体間（個体と個体の間）及びクラスター間（クラスター（特に個体の場合もある）とクラスターの間）の距離の定義を説明します。

2.1 個体間の距離

n 個の各個体が p 変数の観測値 $\{x_{ji},\ j=1,\cdots,n,\ i=1,\cdots,p\}$ を持っているとします。個体 j と個体 k の間の距離を d_{jk} と書いて、以下の様に定義します。

$$d_{jk}=\left\{\sum_{i=1}^{p} D\left(x_{ji}/s_i, x_{ki}/s_i\right)\right\}^{\alpha}$$

ここで、s_i は i 番目の変数の標準化、D は適当な関数です。

「NAG アドイン」では、以下の３つの距離の定義（関数 D）が利用できます。

(a) ユークリッド距離： $D(u,v)=(u-v)^2,\ \alpha=\frac{1}{2}$
(b) ユークリッド平方距離： $D(u,v)=(u-v)^2,\ \alpha=1$
(c) 絶対距離（city block 距離）： $D(u,v)=|u-v|,\ \alpha=1$

また、以下の３つの標準化が利用できます。

(a) 標準偏差： $s_i=\sqrt{\sum_{j=1}^{n}\left(x_{ji}-\bar{x}\right)^2/(n-1)}$
(b) 範囲（レンジ）： $s_i=\max\left(x_{1i},x_{2i},\cdots,x_{ni}\right)-\min\left(x_{1i},x_{2i},\cdots,x_{ni}\right)$
(c) 標準化しない： $s_i=1$

2.2 クラスター間の距離

それぞれ n_i, n_j, n_k 個の個体から成る i, j, k の３つのクラスターがあり、各クラスター間の距離は d_{ij}, d_{ik}, d_{jk} として与えられているとします。
（特に i, j, k が個体の場合、$n_i = 1$, $n_j = 1$, $n_k = 1$ とします。）

j クラスターと k クラスターが統合されて jk クラスターと成った時に、i クラスターと jk クラスターの間の距離を $d_{i,jk}$ と書くとすると、「NAG アドイン」では、以下の６つの距離の定義が利用できます。
（このクラスター間の距離の定義それぞれがクラスター分析の方法に対応します。）

⑷ 最短距離法： $d_{i,jk} = \min\left(d_{ij}, d_{ik}\right)$
⑸ 最長距離法： $d_{i,jk} = \max\left(d_{ij}, d_{ik}\right)$
⑹ 群平均法： $d_{i,jk} = \frac{n_j}{n_j+n_k} d_{ij} + \frac{n_k}{n_j+n_k} d_{ik}$
⑺ 重心法： $d_{i,jk} = \frac{n_j}{n_j+n_k} d_{ij} + \frac{n_k}{n_j+n_k} d_{ik} - \frac{n_j n_k}{\left(n_j+n_k\right)^2} d_{jk}$
⑻ メディアン法： $d_{i,jk} = \frac{1}{2} d_{ij} + \frac{1}{2} d_{ik} - \frac{1}{4} d_{jk}$
⑼ 最小分散法： $d_{i,jk} = \left\{\left(n_i + n_j\right) d_{ij} + \left(n_i + n_k\right) d_{ik} - n_i d_{jk}\right\} / \left(n_i + n_j + n_k\right)$

3 クラスター分析の流れ

クラスター分析は、以下の様なステップで行います。

① 個体間の距離の定義、及び、クラスター間の距離の定義（クラスター分析の方法）を選択する。

② 個体間（全ての組合せ）の距離を計算する。
（この時点で、個体それぞれを（最小の）クラスターと見なす。）

③ 距離が最小となる個体と個体を統合して１つのクラスターとする。

④ 新しく形成されたクラスターと既存のクラスター（特に個体の場合もある）の間の距離を計算し、（全ての組合せから）距離が最小のものを統合して新たなクラスターとする。これを全てのクラスターが統合され単一のクラスター（全体集合）になるまで繰り返す。

⑤ クラスターの統合過程に基づいて樹形図を作成する。

⑥ 樹形図を適当な高さ（距離）で切断し、個体をいくつかのクラスターに分ける。

⑦ 各クラスターに含まれる個体を調べ、クラスター（クラスター間）の特徴を把握する。

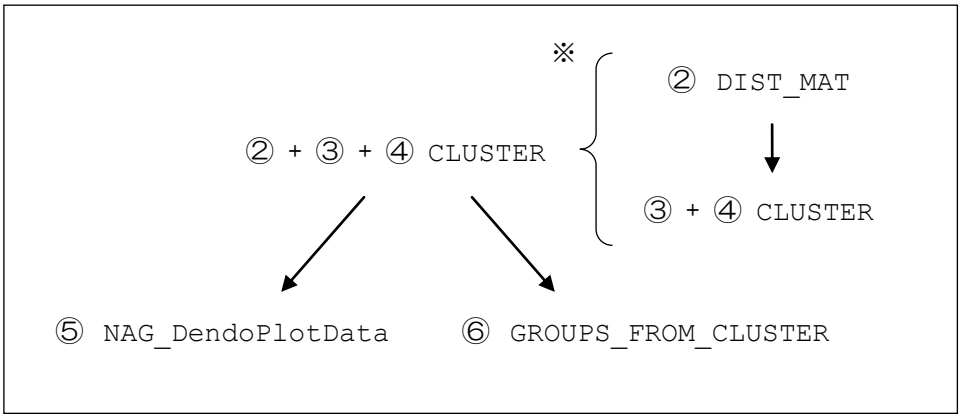


図２　「NAG アドイン」の関数との関係

※　関数 DIST_MAT を用いて、個体間（全ての組合せ）の距離（距離行列）を別途計算することもでき、その結果（距離行列）を入力（引数）として関数 CLUSTER でクラスター分析を行うことも可能です。（このドキュメントでは、関数 DIST_MAT に関する操作例は割愛します。）

4 「NAG アドイン」を用いた分析例

以下の表１のデータ（５つの個体がそれぞれ２変数を持っている）に対して、実際に「NAG アドイン」を用いたクラスター分析の例を示します。
（本章の説明は Excel 2007 に「NAG アドイン」がインストールされていることが前提となります。また、Excel 2003 をご利用の方は適宜読み替えてください。）

距離の定義には、個体間の距離に「ユークリッド距離」、クラスター間の距離（クラスター分析の方法）に「最短距離法」を用いることにします。

表１　５段階評価による成績

生徒	教科	
	国語	数学
A	5	4
B	1	2
C	4	4
D	2	1
E	5	2

この分析例では、以下の３つの関数を使用します。

《 CLUSTER 》
表１の様な行列データを入力（引数）としてクラスター分析を行います。
各種距離の定義を指定でき、計算結果としてクラスターの統合過程と樹形図を描く為の情報が出力されます。

《 NAG_DendoPlotData 》
関数 CLUSTER の計算結果を入力（引数）として樹形図を作成します。
結果は、Excel のグラフとして出力されます。

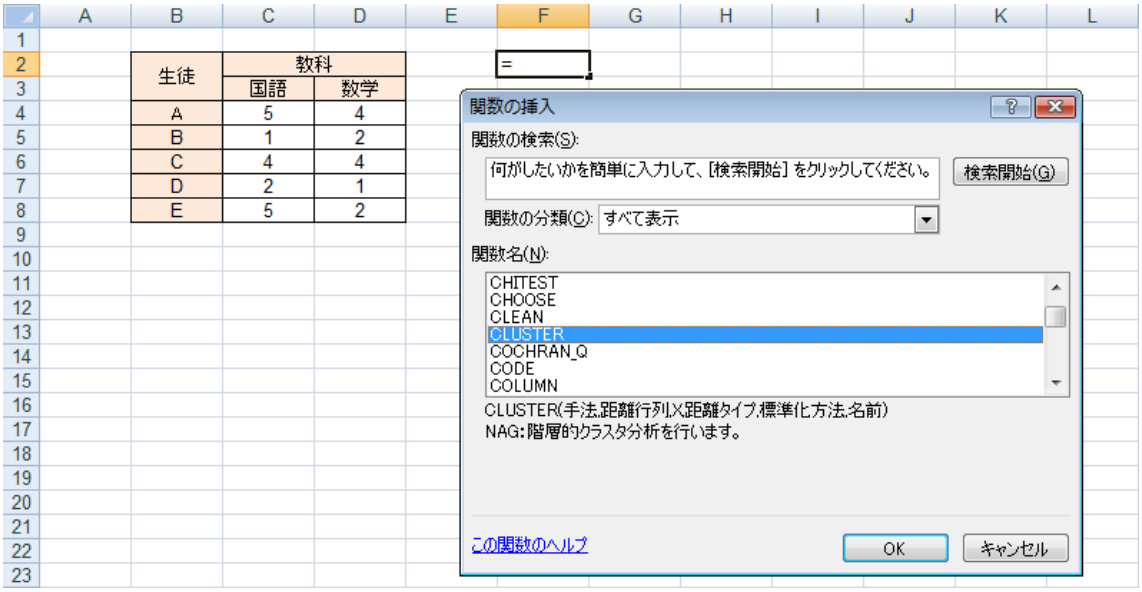
《 GROUPS_FROM_CLUSTER 》
関数 CLUSTER の計算結果を入力（引数）としてクラスター分けを行います。
クラスター数、又は、クラスター距離（高さ）のどちらかを指定して、どの個体が同じクラスターに属しているかを示すインデックス（指標）を出力します。

4.1　クラスター分析（CLUSTER）

Excel の表に上記の表１のデータを入力します。

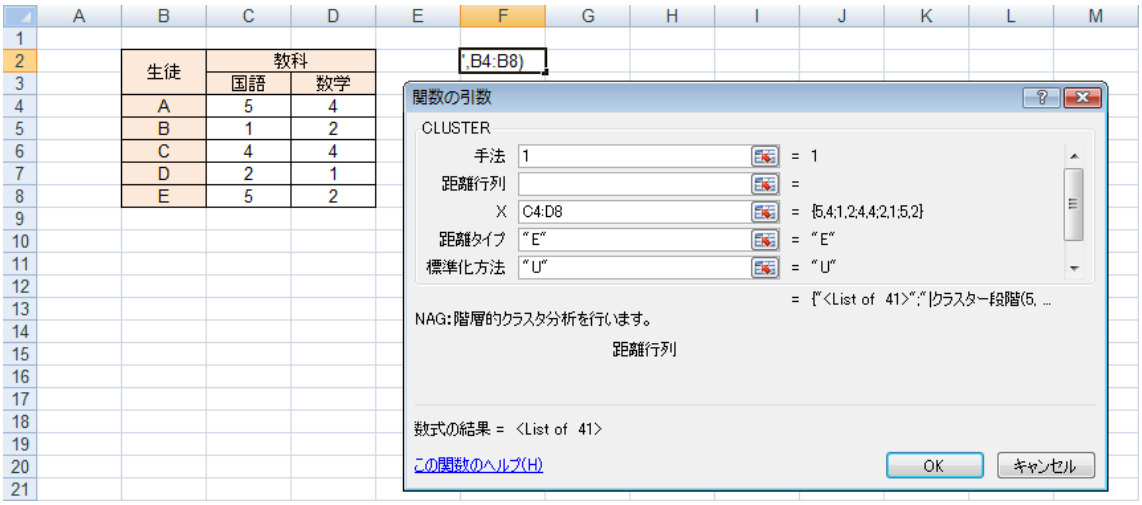
	A	B	C	D	E
1					
2		生徒	教科		
3			国語	数学	
4		A	5	4	
5		B	1	2	
6		C	4	4	
7		D	2	1	
8		E	5	2	
9					

適当なセル（例えば F2）を選択し、関数 CLUSTER を挿入します。
（Excel の標準関数と同様に「数式｜関数の挿入」から行います。）



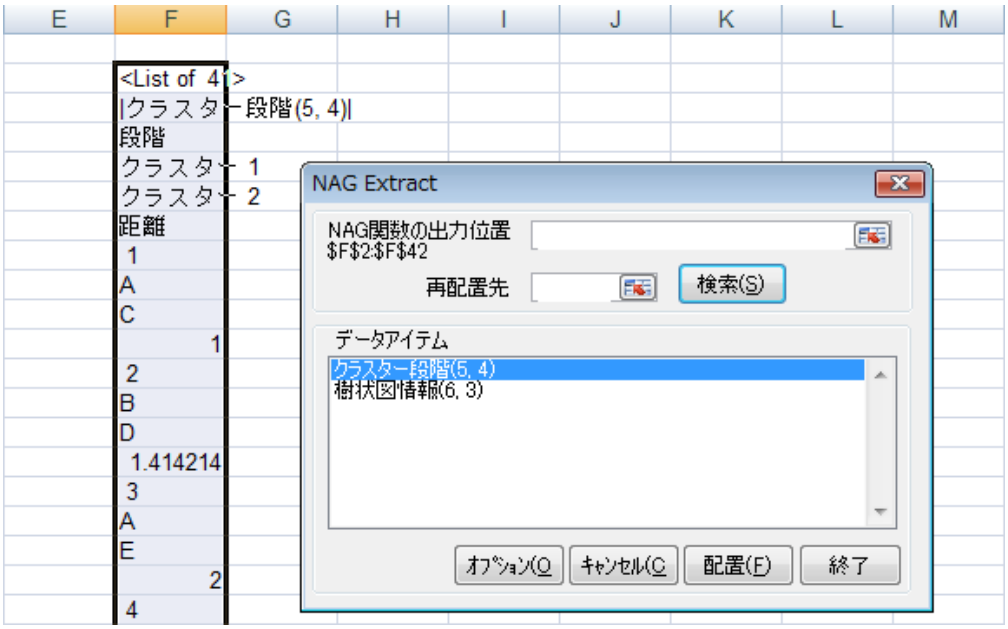
関数の引数を入力するウィンドウが表示されます。
各引数を以下の様に設定し、「OK」ボタンを押します。
（各引数の詳細は「この関数のヘルプ」をご参照ください。）

「手法」 ⇒ 最短距離法 ⇒ 1
「距離行列」 ⇒ 今回は距離行列を使わない ⇒ 設定なし
「X」 ⇒ 変数データ（成績） ⇒ C4:D8
「距離タイプ」 ⇒ ユークリッド距離 ⇒ E
「標準化方法」 ⇒ 今回は標準化を行わない ⇒ U
「名前」 ⇒ 個体の名前（生徒） ⇒ B4:B8

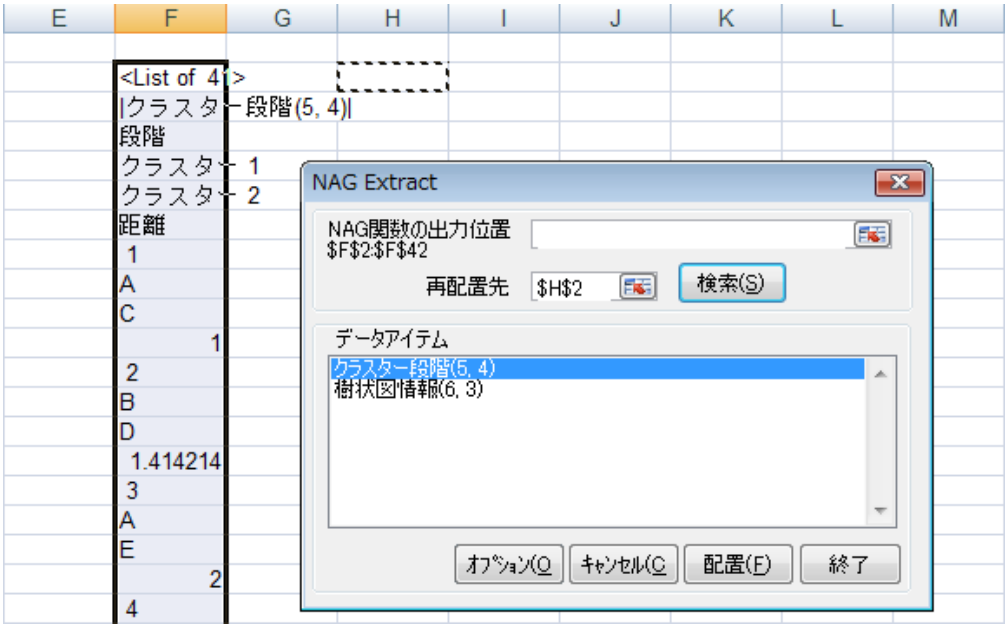


関数を挿入したセル（F2）に〈List of 41〉などと出力されれば計算は完了です。
引き続き、この選択状態のままで Ctrl + E を押します。

関数を挿入したセル（F2）の下方向に出力結果が展開され、NAG Extract（出力結果再配置ツール）ウィンドウが開きます。



「データアイテム」から「クラスター段階」選択し、「再配置先」として適当なセル（例えば H2）をクリックし、「終了」ボタンを押します。



出力結果（この場合は「クラスター段階」）が整理された状態で配置されます。

H	I	J	K
	クラスター段階		
段階	クラスター 1	クラスター 2	距離
1	A	C	1
2	B	D	1.414213562
3	A	E	2
4	A	B	3.16227766

他のデータアイテムの配置も同様です。関数を挿入したセル（F2）を選択し、Ctrl + E を押し NAG Extract ウィンドウを開きます。
「データアイテム」で今度は「樹状図情報」選択し、「再配置先」として適当なセル（例えば H9）をクリックし、「終了」ボタンを押します。
出力結果（この場合は「樹状図情報」）が整理された状態で配置されます。

H	I	J	K
	クラスター段階		
段階	クラスター 1	クラスター 2	距離
1	A	C	1
2	B	D	1.414213562
3	A	E	2
4	A	B	3.16227766
	樹状図情報		
指標	オブジェクト	距離	
1	A	1	
3	C	2	
5	E	3.16227766	
2	B	1.414213562	
4	D	3.16227766	

最後に、出力結果を簡単に説明します。

「クラスター段階」とは、クラスターの統合過程を示します。
段階１で個体 A と個体 C が統合されます。
（このクラスターは改めて A と名づけられます。）
段階２で個体 B と個体 D が統合されます。
（このクラスターは改めて B と名づけられます。）
段階３でクラスターA と個体 E が統合されます。
（このクラスターは改めて A と名づけられます。）
段階４でクラスターA とクラスターB が統合されます。
（これで全てのクラスターが統合されたので終了です。）

「樹状図情報」は樹形図（樹状図、デンドログラムとも言う）を描く為の情報です。
この情報から樹形図を描くルールは以下の様になります。

① 「オブジェクト」にある個体名（A, C, E, B, D）を左から右に並べる。
② その個体名の上に「距離」で示される長さの垂直な線を引く。
③ 左から順に線の最上部から右に向かって水平線を引き次の線にぶつかった所で止める。

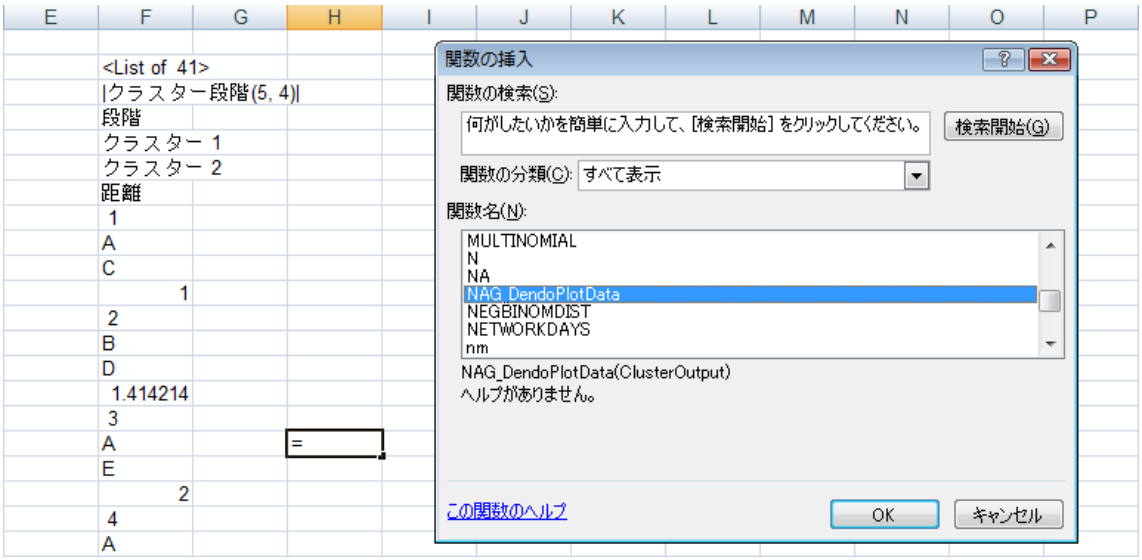
この作図に関しましては、Excel のグラフ機能を用いた樹形図を描く為の関数
`NAG_DendoPlotData` を弊社から別途無償にて提供しています。
詳細は次節 4.2 をご参照ください。

以上で関数 `CLUSTER` によるクラスター分析は終了です。

4.2　樹形図の作成（NAG_DendoPlotData）

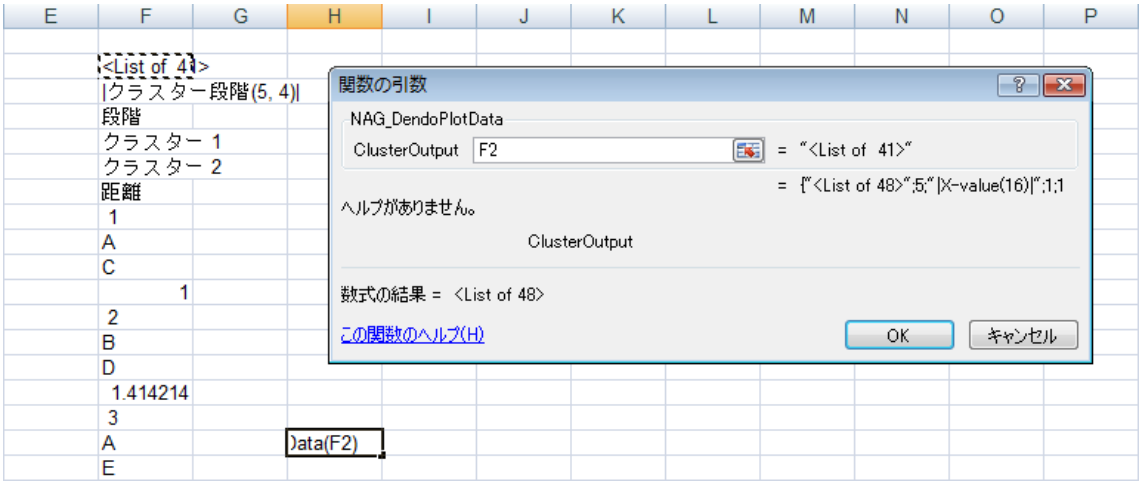
この節では 4.1 節の操作が完了し「樹状図情報」が得られていることが前提となります。

適当なセル（例えば H17）を選択し、関数 NAG_DendoPlotData を挿入します。
（Excel の標準関数と同様に「数式｜関数の挿入」から行います。）

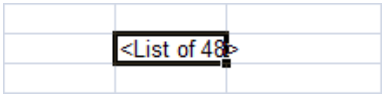


関数の引数を入力するウィンドウが表示されます。
引数を以下の様に設定し、「OK」ボタンを押します。

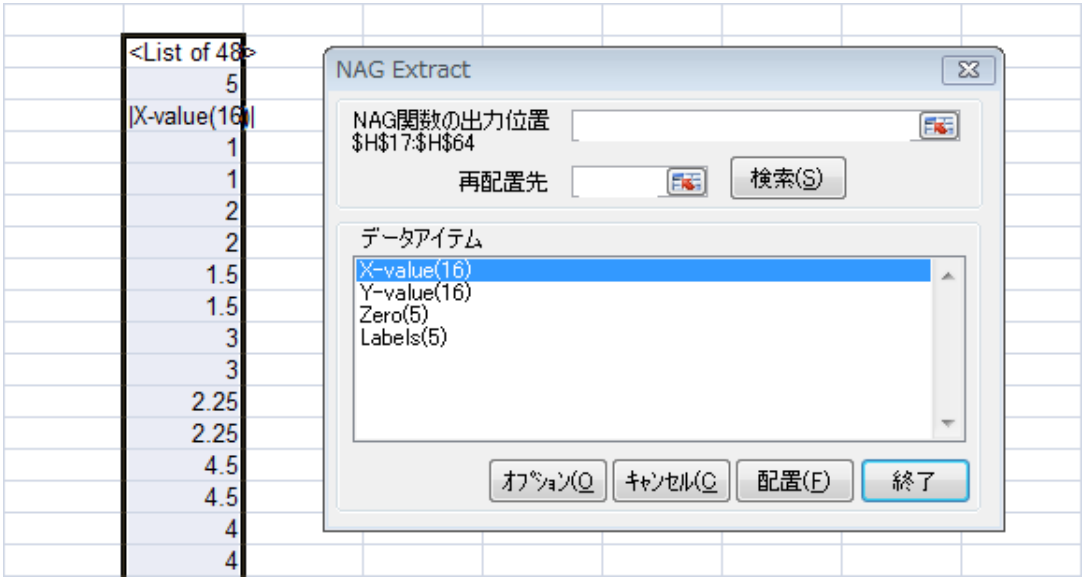
"ClusterOutput" ⇒ 関数 CLUSTER を配置したセルを選択する ⇒ F2



関数を挿入したセル（H17）に 〈List of 48〉 などと出力されれば計算は完了です。
引き続き、この選択状態のままで Ctrl + E を押します。

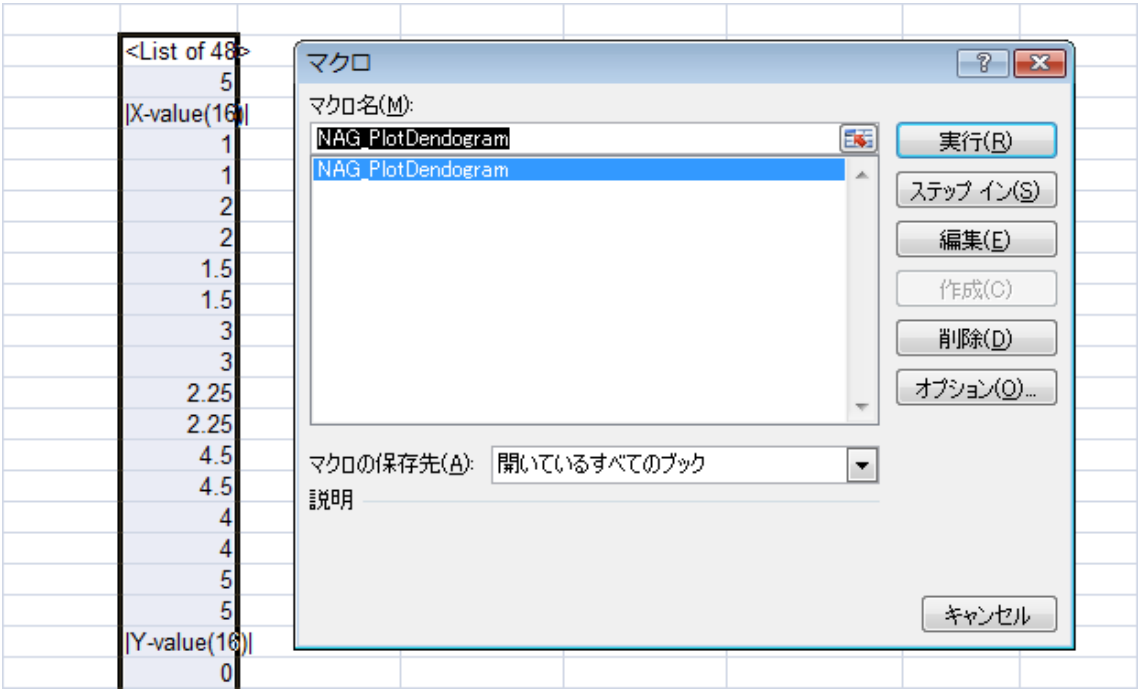


関数を挿入したセル（H17）の下方向に出力結果が展開され、NAG Extract（出力結果再配置ツール）ウィンドウが開きます。
ここでは何もせず「終了」ボタンを押します（出力結果を展開するのが目的です）。

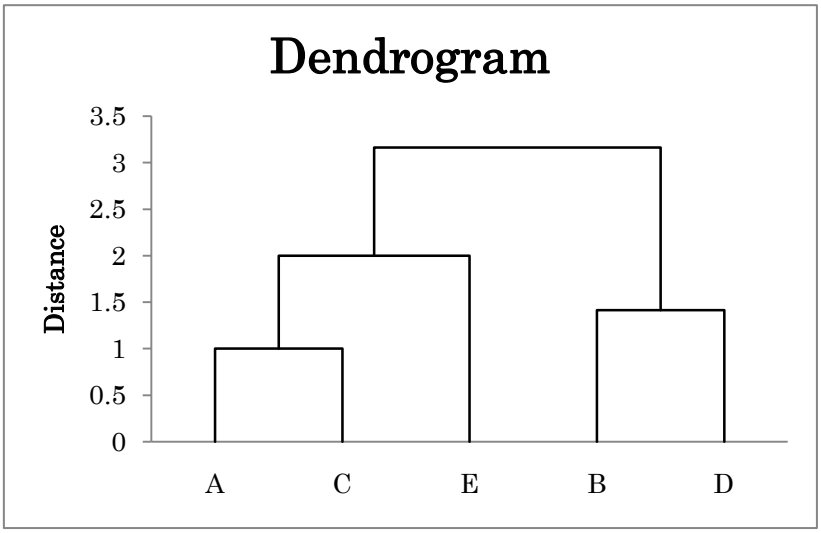


引き続き、関数を挿入したセル（H17）を選択した状態のままで Excel のメニュー「表示｜マクロ」からマクロウィンドウを開きます。

“NAG_PlotDendogram” を選択し、「実行」ボタンを押します。



Excel のグラフとして樹形図が出力されます。

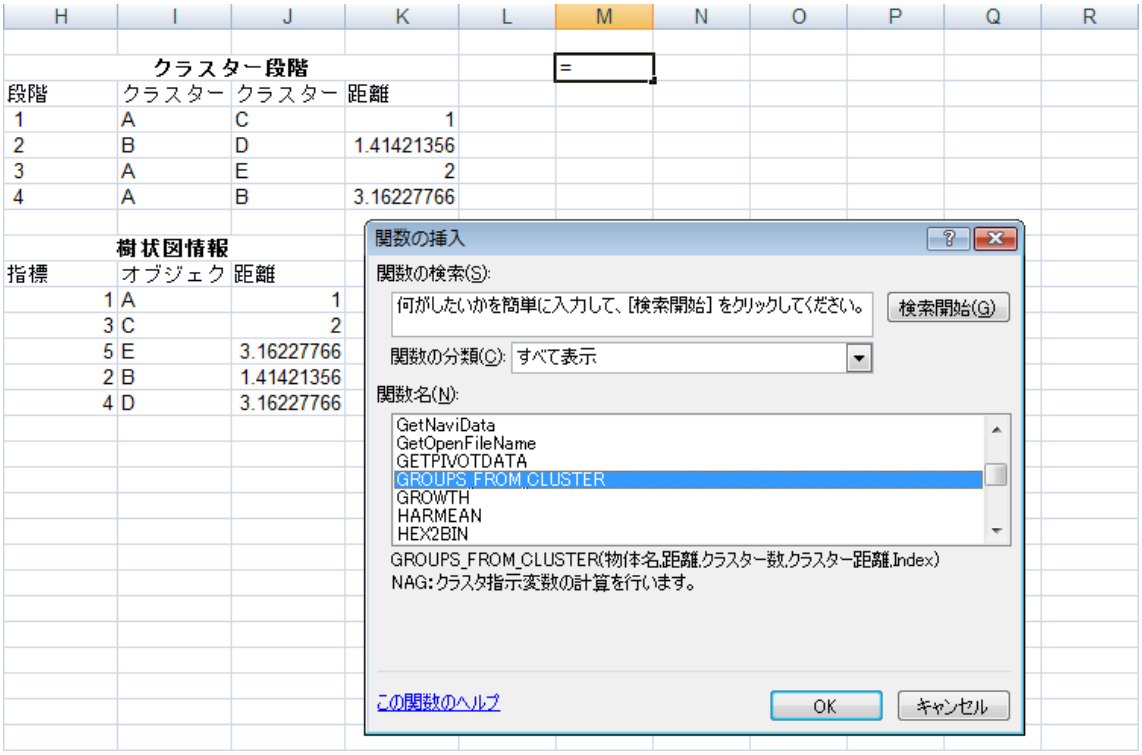


以上で関数 NAG_DendoPlotData による樹形図の作成は終了です。

4.3 クラスター分け（GROUPS_FROM_CLUSTER）

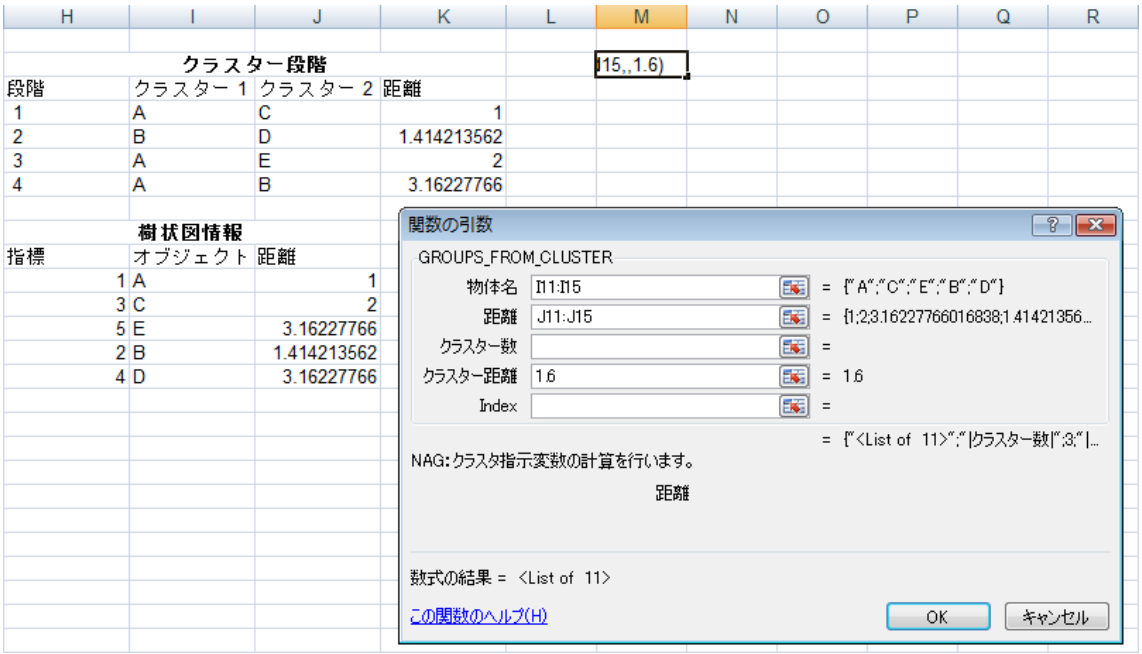
この節では 4.1 節の操作が完了し「樹状図情報」が得られていることが前提となります。

適当なセル（例えば M2）を選択し、関数 GROUPS_FROM_CLUSTER を挿入します。
（Excel の標準関数と同様に「数式｜関数の挿入」から行います。）



関数の引数を入力するウィンドウが表示されます。
各引数を以下の様に設定し、「OK」ボタンを押します。
（各引数の詳細は「この関数のヘルプ」をご参照ください。）

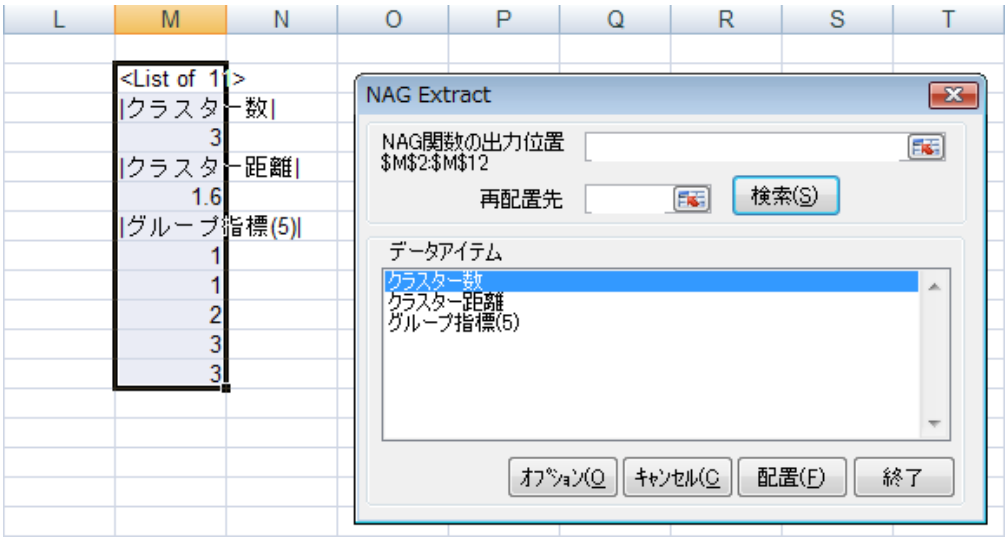
「物体名」 ⇒ 樹状図情報の「オブジェクト」 ⇒ I11:I15
「距離」 ⇒ 樹状樹情報の「距離」 ⇒ J11:J15
「クラスター数」 ⇒ 今回は「クラスター距離」を用いる ⇒ 設定なし
「クラスター距離」 ⇒ 樹形図を高さ 1.6 で切断する ⇒ 1.6
「Index」 ⇒ 出力される指標は樹形図での個体の順番（デフォルト） ⇒ 設定なし



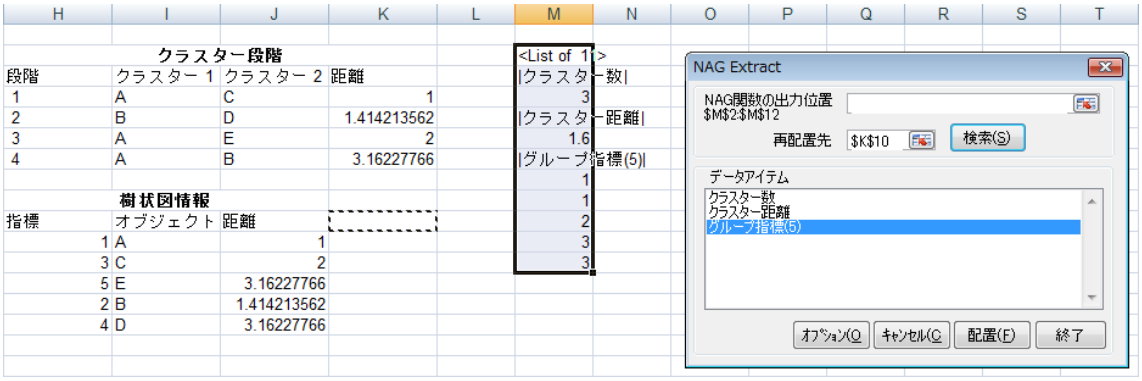
関数を挿入したセル（M2）に <List of 11> などと出力されれば計算は完了です。
引き続き、この選択状態のままで Ctrl + E を押します。

H	I	J	K	L	M	N
	クラスター段階				<List of 11>	
段階	クラスター 1	クラスター 2	距離			
1	A	C	1			
2	B	D	1.414213562			
3	A	E	2			
4	A	B	3.16227766			
	樹状図情報					
指標	オブジェクト	距離				
1	A	1				
3	C	2				
5	E	3.16227766				
2	B	1.414213562				
4	D	3.16227766				

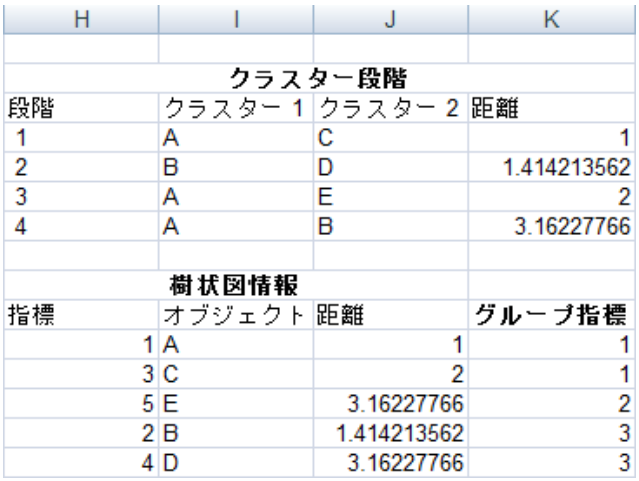
関数を挿入したセル（M2）の下方向に出力結果が展開され、NAG Extract（出力結果再配置ツール）ウィンドウが開きます。



「データアイテム」から「グループ指標」選択し、「再配置先」として樹状図情報の右端（例えば K10）をクリックし、「終了」ボタンを押します。



出力結果（この場合は「グループ指標」）が整理された状態で配置されます。



「グループ指標」は、同じインデックス（指標）の個体（オブジェクト）が同じクラスターに属していることを意味します。
従って、高さ（距離）1.6 で樹形図を切断した場合、{A, C}, {E}, {B, D} の３つのクラスターに分類されることが分かります。

以上で関数 GROUPS_FROM_CLUSTER によるクラスター分けは終了です。

4.4　結果の検討

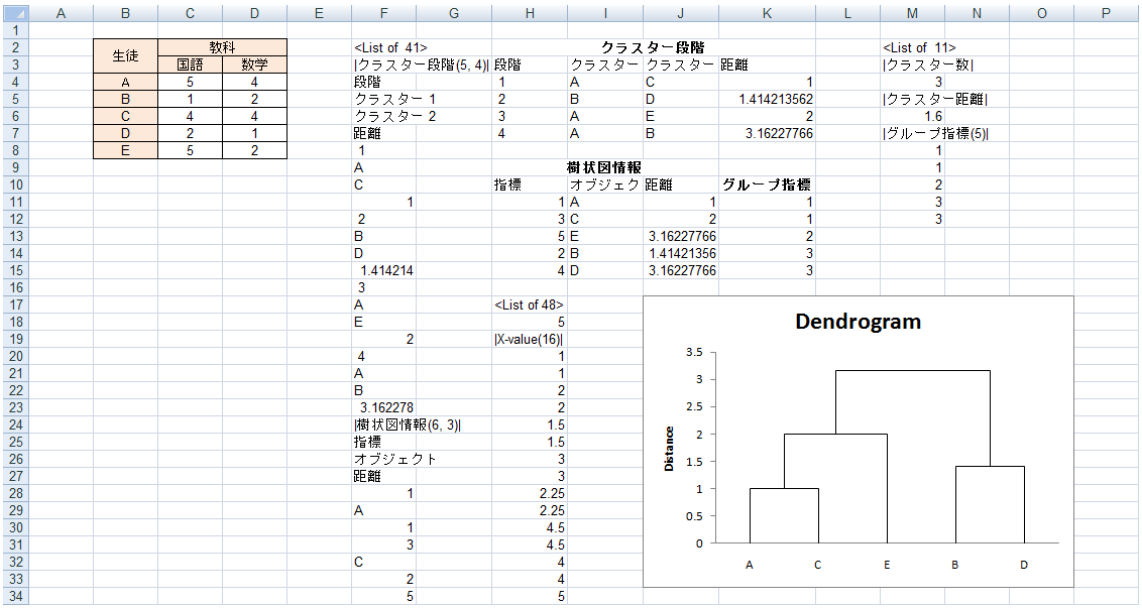


図３　クラスター分析の終了イメージ

最後に、各クラスターの特徴を調べてみます。
表２のデータ（表１の再録）を見ますと、クラスター {A, C} は国語も数学も得意な生徒、クラスター {E} は国語は得意だが数学は苦手な生徒、クラスター {B, D} は国語も数学も苦手な生徒、といった特徴を持っている様です。

表２　５段階評価による成績（表１の再録）

生徒	教科	
	国語	数学
A	5	4
B	1	2
C	4	4
D	2	1
E	5	2

5 参考文献

[1] 田中　豊, 脇本和昌, 「多変量統計解析法」, 現代数学社, 1983.

[2] 永田　靖, 棟近雅彦, 「多変量解析法入門」, サイエンス社, 2001.

[3] Everitt B S, "Cluster Analysis", Heinemann, 1974.

[4] Krzanowski W J, "Principles of Multivariate Analysis",
Oxford University Press, 1990.