多変量解析

第13回　クラスター分析

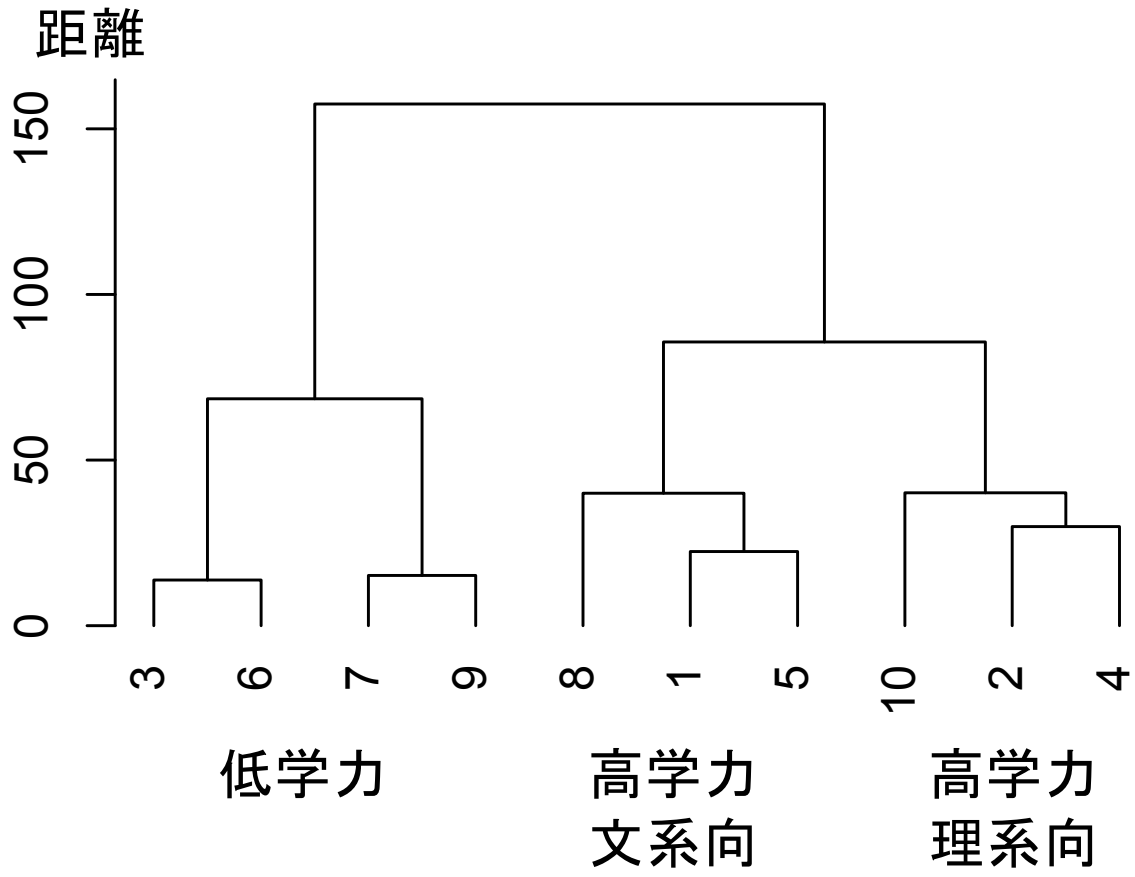
萩原•篠田
情報理工学部

クラスター分析

類似の能力をもつ生徒をグループ化できるか？
それぞれのグループの特徴は何か？

生徒 No.	国語 x_1	英語 x_2	数学 x_3	理科 x_4
1	86	79	67	68
2	71	75	78	84
3	42	43	39	44
4	62	58	98	95
5	96	97	61	63
6	39	33	45	50
7	50	53	64	72
8	78	66	52	47
9	51	44	76	72
10	89	92	93	91

クラスターを樹形図（デンドログラム）で表示

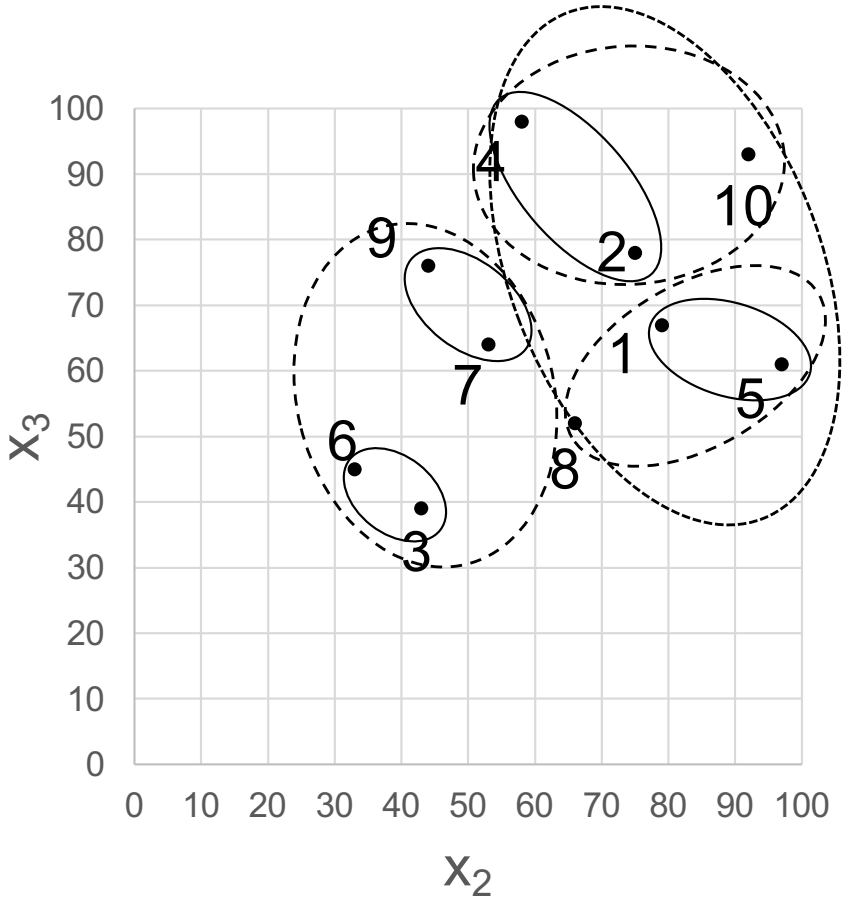


keywords

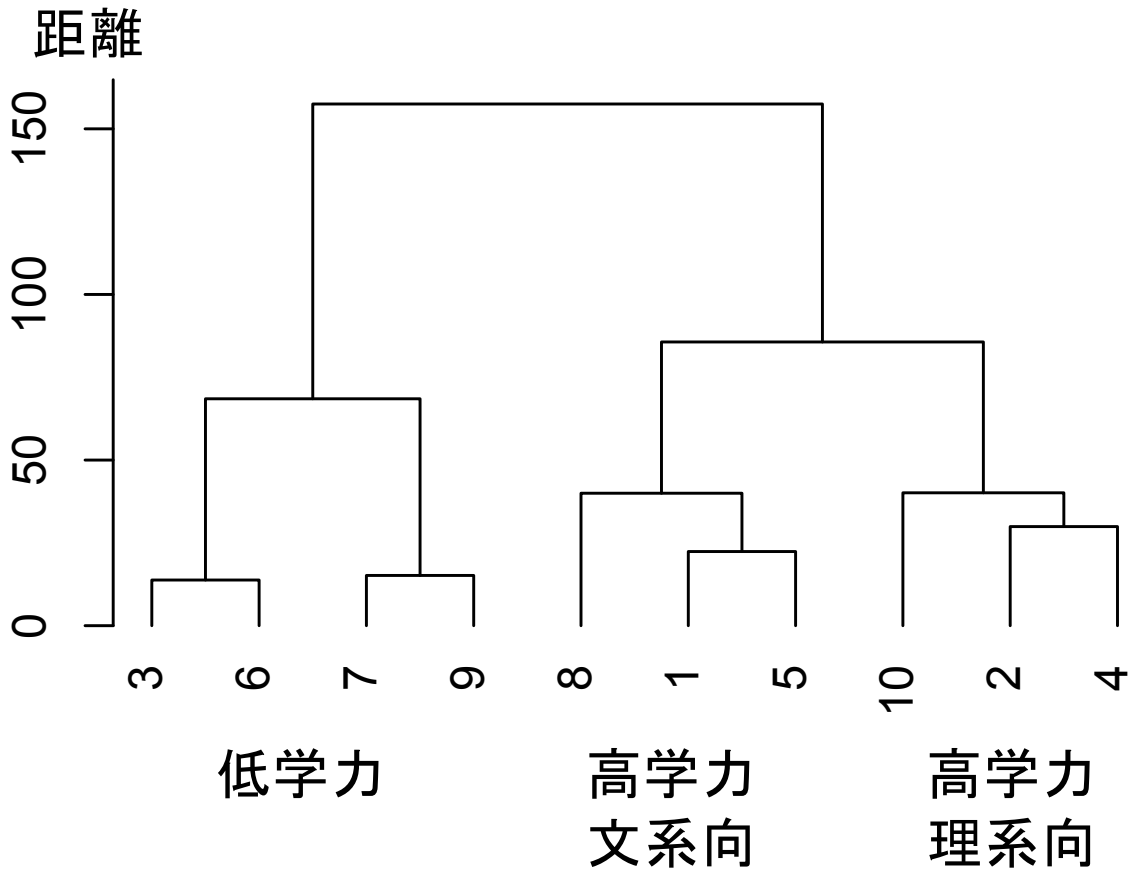
クラスター、距離（ユークリッド、マハラノビス）、デンドログラム

クラスター分析

類似の能力をもつ生徒をグループ化できるか？
それぞれのグループの特徴は何か？



クラスターを樹形図（デンドログラム）で表示



keywords
クラスター、距離（ユークリッド、マハラノビス）、デンドログラム

クラスター分析：

対象物（データの集まり）の中から，互いに似たものを集めて，群れや集団（クラスタ）に分ける手法

「似ている」の定義？，「似ている」程度を測る方法

- ユークリッド距離
- ユークリッド距離の2乗(平方ユークリッド距離)
- マハラノビスの距離
- 相関係数

M 次元空間内の ij 間のユークリッド距離 $d_{ij} = \left\{ \sum_{m=1}^{M} (x_{im} - x_{jm})^2 \right\}^{\frac{1}{2}}$

例えば2次元空間で，(x_{i1}, x_{i2}), (x_{j1}, x_{j2})を i 番目と j 番目の対象データとすると

・ユークリッド距離

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}$$

・ユークリッド距離の2乗(平方ユークリッド距離)

$$d_{ij}{}^2 = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2$$

・マハラノビスの距離

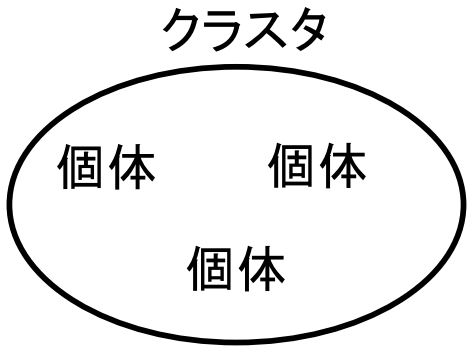
$$D^2 = \frac{(x - \bar{x})^2}{s^2}$$

・相関係数

$$r = \frac{\frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2 \frac{1}{n-1}\sum (y_i - \bar{y})^2}}$$

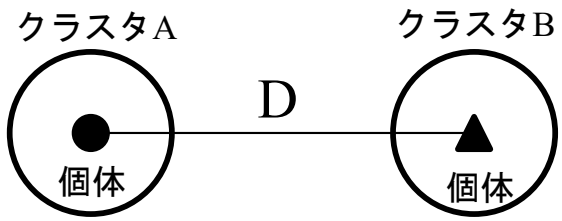
似ている程度を測る方法は、距離の概念の一般化と考えられるので広い意味で距離と呼ぶ

クラスター分析ではデータのことを個体と呼び、個体と個体が集まってクラスタを構成することになる



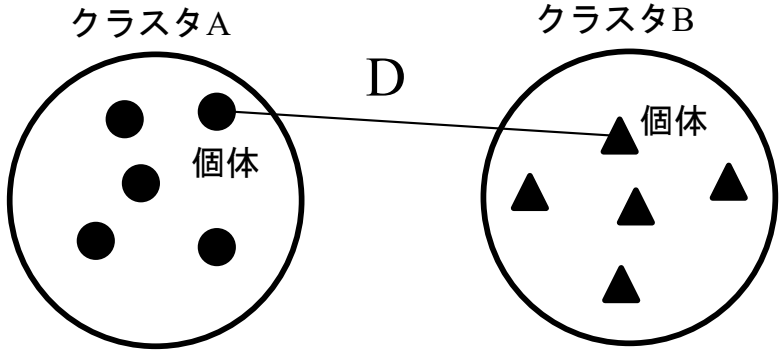
クラスタ間の距離の決め方

・クラスタの成分が1個だけからなる場合



個体と個体との距離 = クラスタ間の距離D

・クラスタの成分が2個以上からなる場合



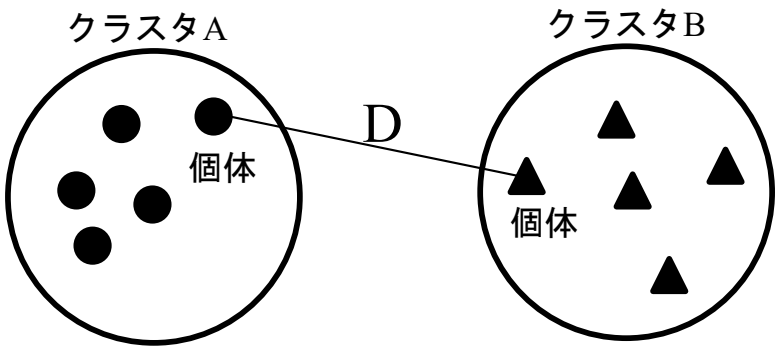
Aのどの個体とBのどの個体の間を測ればいいのか。

主なものとして、以下の方法がある

1. 最短距離法
2. 最長距離法
3. 群平均法
4. メディアン法
5. 重心法
6. ウォード法

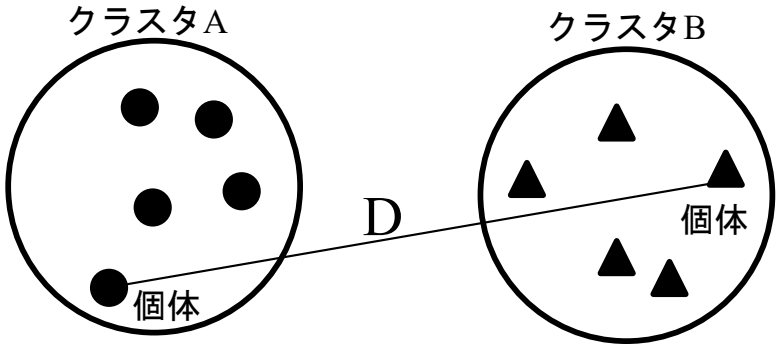
1. 最短距離法

クラスタAの個体とクラスタBの個体とのすべての
組み合わせについて距離を求めてその中で
最も短い距離＝2つのクラスタA,B間の距離D



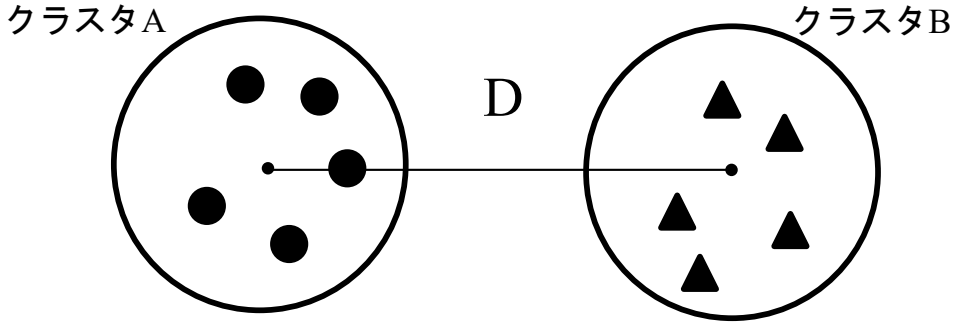
2. 最長距離法

クラスタAの個体とクラスタBの個体とのすべての
組み合わせについて距離を求めてその中で
最も長い距離＝2つのクラスタA,B間の距離D



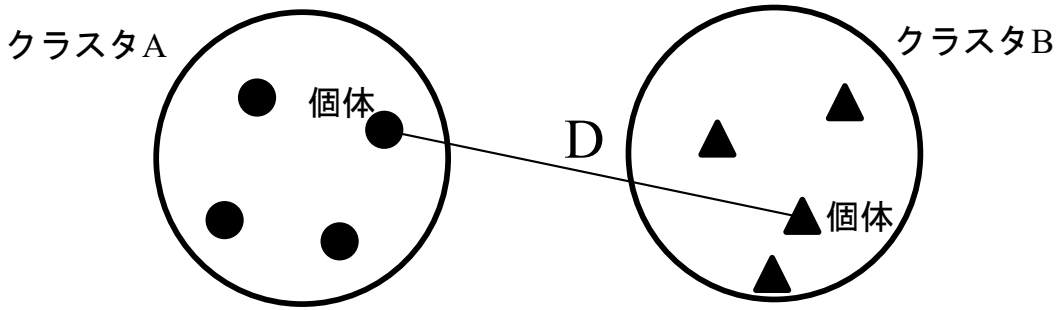
3. 群平均法

クラスタAの個体とクラスタBの個体とのすべての
組み合わせについて距離を求めて
その距離の平均値＝2つのクラスタA,B間の距離D



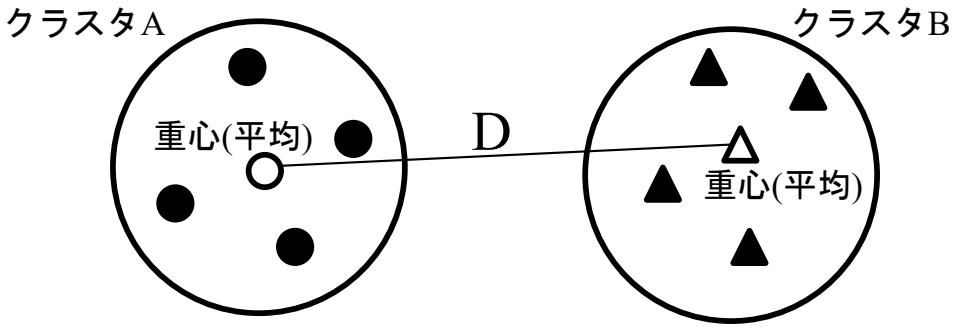
4. メディアン法

クラスタAの個体とクラスタBの個体とのすべての
組み合わせについて距離を求めて
その距離を順番に並べたときの中央値
＝2つのクラスタA,B間の距離D



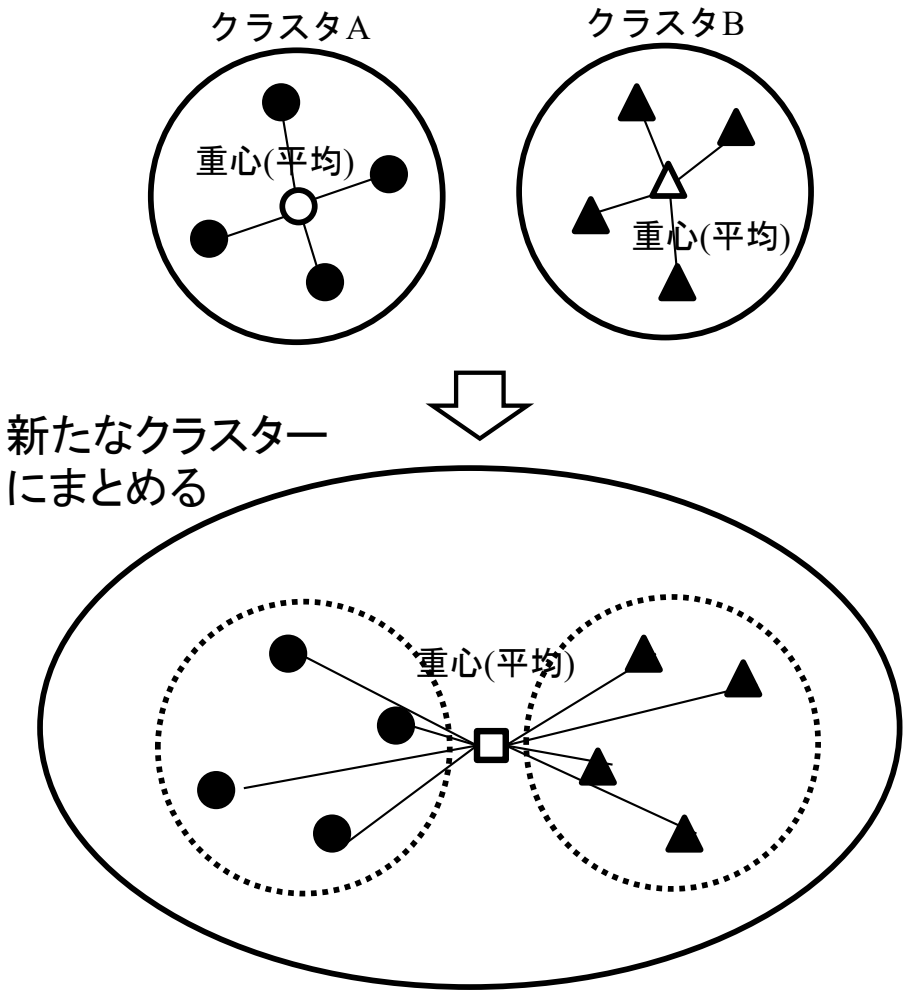
5. 重心法

クラスタAの重心とクラスタBの重心との距離
＝2つのクラスタA,B間の距離D



6. ウォード法

新たに統合されるクラスター内の平方和を最も小さくする
という基準でクラスターを形成していく方法



２つのクラスターA,Bを統合したと仮定したとき、

S_{AB}：新たなクラスターの重心とクラスター内の
各サンプルとの距離の平方和

S_A，S_B：元々の２つのクラスター内での重心と
それぞれのサンプルとの距離の平方和

としたとき

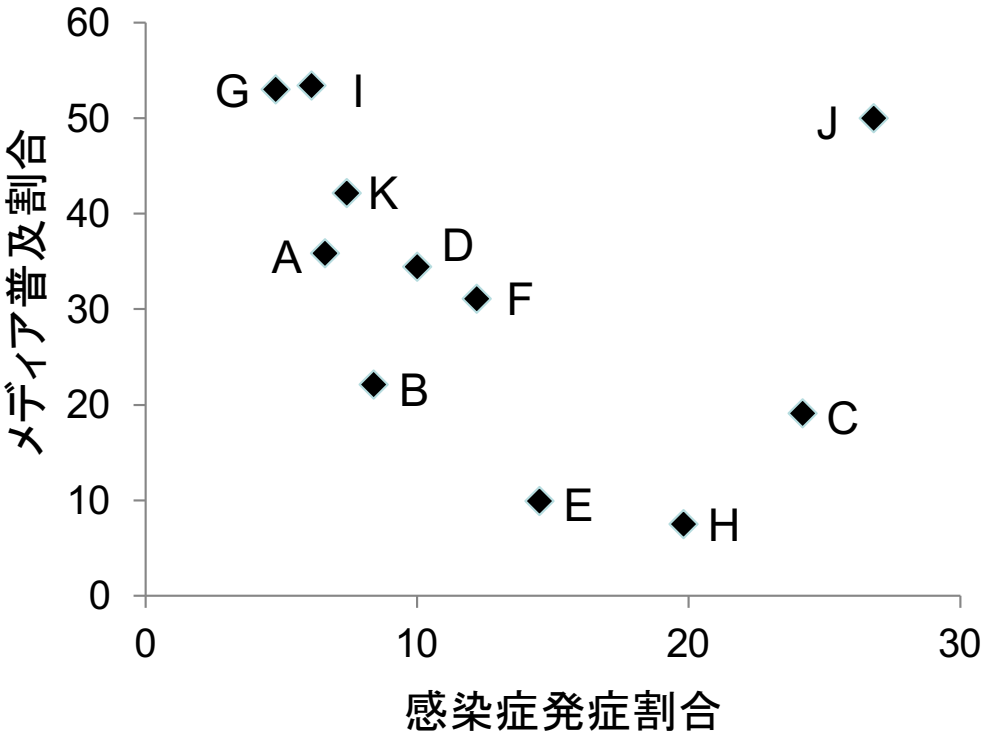
$$\Delta S_{AB} = S_{AB} - S_A - S_B$$

が最小となるようにクラスター同士を統合する。
この平方和の増加分がウォード法における距離と
なる

クラスター分析の手順(1)

国名	感染症発症割合	メディア普及割合
A	6.6	35.8
B	8.4	22.1
C	24.2	19.1
D	10	34.4
E	14.5	9.9
F	12.2	31.1
G	4.8	53
H	19.8	7.5
I	6.1	53.4
J	26.8	50
K	7.4	42.1

→



散布図を見ると、{G,I}、{A,B,D,F,K}、{C,E,H}、{J}のような4つのクラスタになりそう

→デンドログラム(樹形図)というグラフで表現

クラスター分析の手順(2)

	B	C	D	E	F	G	H	I	J	K
A	190.9	588.7	13.5	733.2	53.5	299.1	975.1	310.0	609.7	40.3
B		258.6	153.9	186.1	95.4	967.8	343.1	985.0	1117.0	401.0
C			435.7	178.7	288.0	1525.6	153.9	1504.1	961.6	811.2
D				620.5	15.7	373.0	819.7	376.2	525.6	66.1
E					454.7	1951.7	33.9	1962.8	1759.3	1087.3
F						534.4	614.7	534.5	570.4	144.0
G							2295.3	1.9	493.0	125.6
H								2294.5	1855.3	1350.9
I									440.1	129.4
J										438.8

- この組み合わせの中で“距離”が最小なのは、GとIの組み合わせ{G,I}を構成する

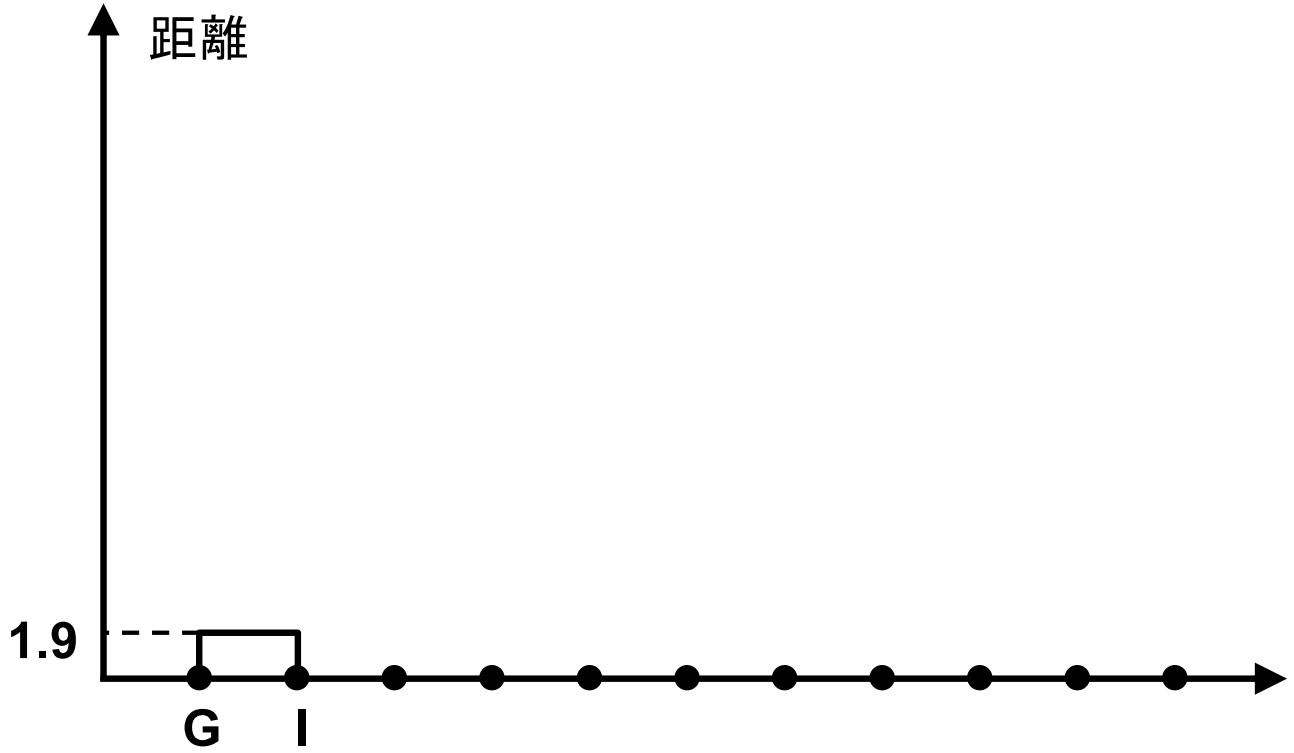
  ユークリッド距離の2乗(平方ユークリッド距離)

  $d_{ij}{}^2 = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2$

- 式は $(4.8-6.1)^2 + (53.0-53.4)^2 = 1.85$ となる

国名	感染症発症割合	メディア普及割合
G	4.8	53
I	6.1	53.4

クラスター分析の手順(3)



デンドログラム第1段階

- 階層クラスター分析
  → 重心法(平均)
  → 平方ユーグリッド距離

GとIが一つのクラスターになったので

	B	C	D	E	F	G･I	H	J	K
A	190.9	588.7	13.5	733.2	53.5	304.1	975.1	609.7	40.3
B		258.6	153.9	186.1	95.4	975.9	343.1	1117.0	401.0
C			435.7	178.7	288.0	1514.4	153.9	961.6	811.2
D				620.5	15.7	374.1	819.7	525.6	66.1
E					454.7	1956.8	33.9	1759.3	1087.3
F						534.0	614.7	570.4	144.0
G･I							2294.4	466.1	127.0
H								1855.3	1350.9
J									438.8

- GとIの重心(平均)は

$(4.8+6.1)/2=5.45,\ (53+53.4)/2=53.2$

AとG･Iのユークリッド距離は

$(6.6-5.45)^2+(35.8-53.2)^2=304.1$

国名	感染症発症割合	メディア普及割合
A	6.6	35.8
G	4.8	53
I	6.1	53.4
G･I	5.45	53.2

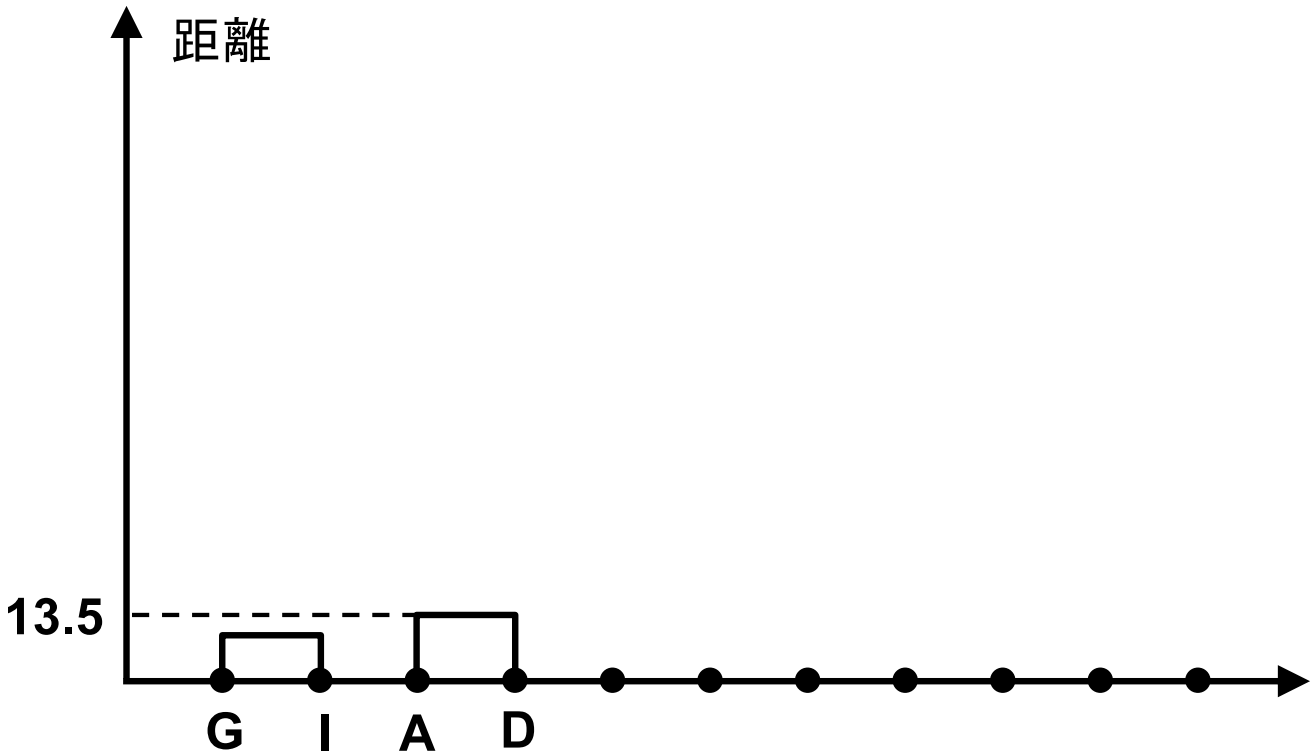
クラスター分析の手順(4)

	B	C	D	E	F	G･I	H	J	K
A	190.9	588.7	13.5	733.2	53.5	304.1	975.1	609.7	40.3
B		258.6	153.9	186.1	95.4	975.9	343.1	1117.0	401.0
C			435.7	178.7	288.0	1514.4	153.9	961.6	811.2
D				620.5	15.7	374.1	819.7	525.6	66.1
E					454.7	1956.8	33.9	1759.3	1087.3
F						534.0	614.7	570.4	144.0
G･I							2294.4	466.1	127.0
H								1855.3	1350.9
J									438.8

- この組み合わせで、13.5が最小の距離なので、
  AとDが2つ目のクラスタ{A,D}を構成

式は $(10.0-6.6)^2+(34.4-35.8)^2=13.52$ となる

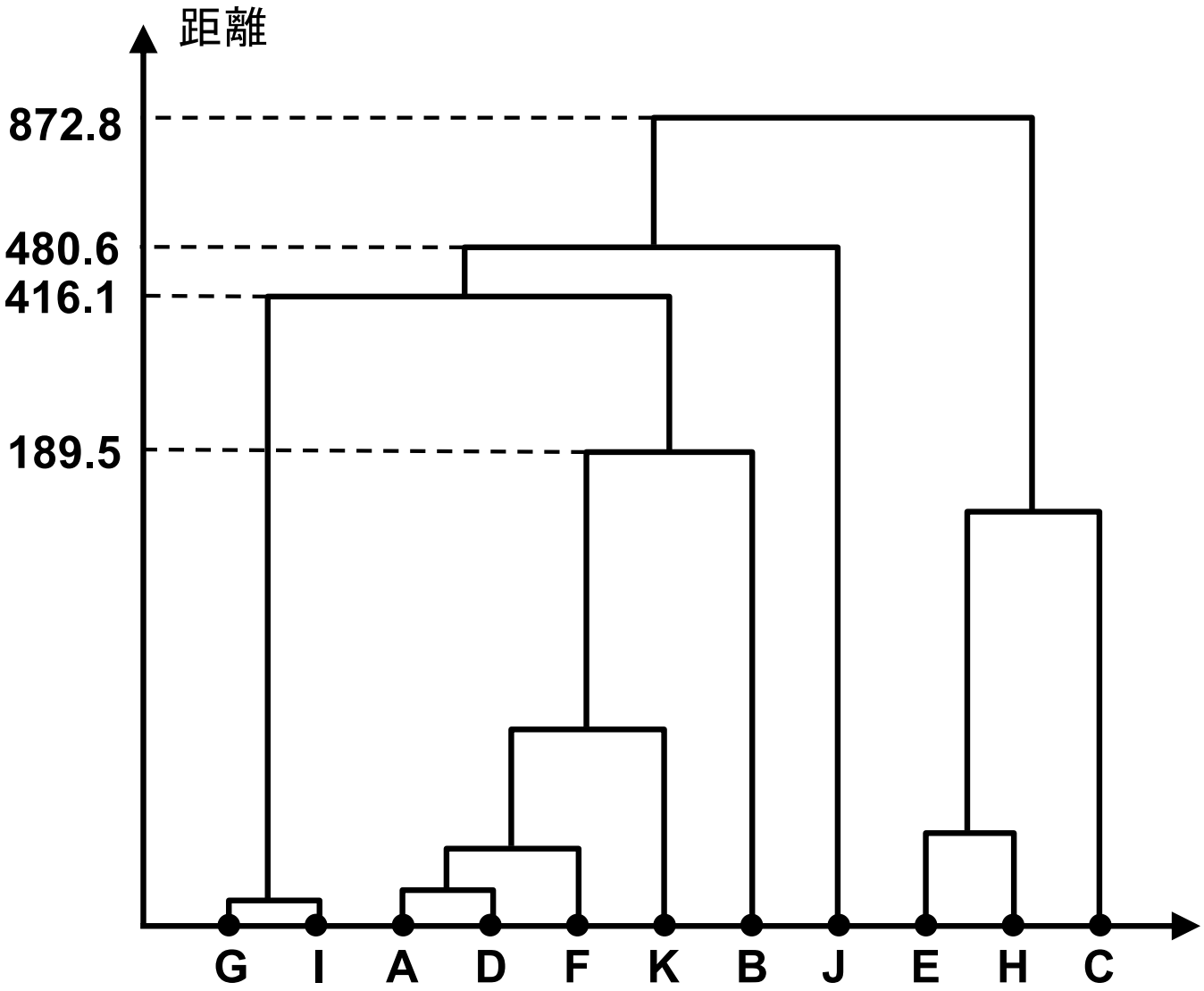
クラスター分析の手順(5)



デンドログラム第2段階

以下同様の手順を繰り返す

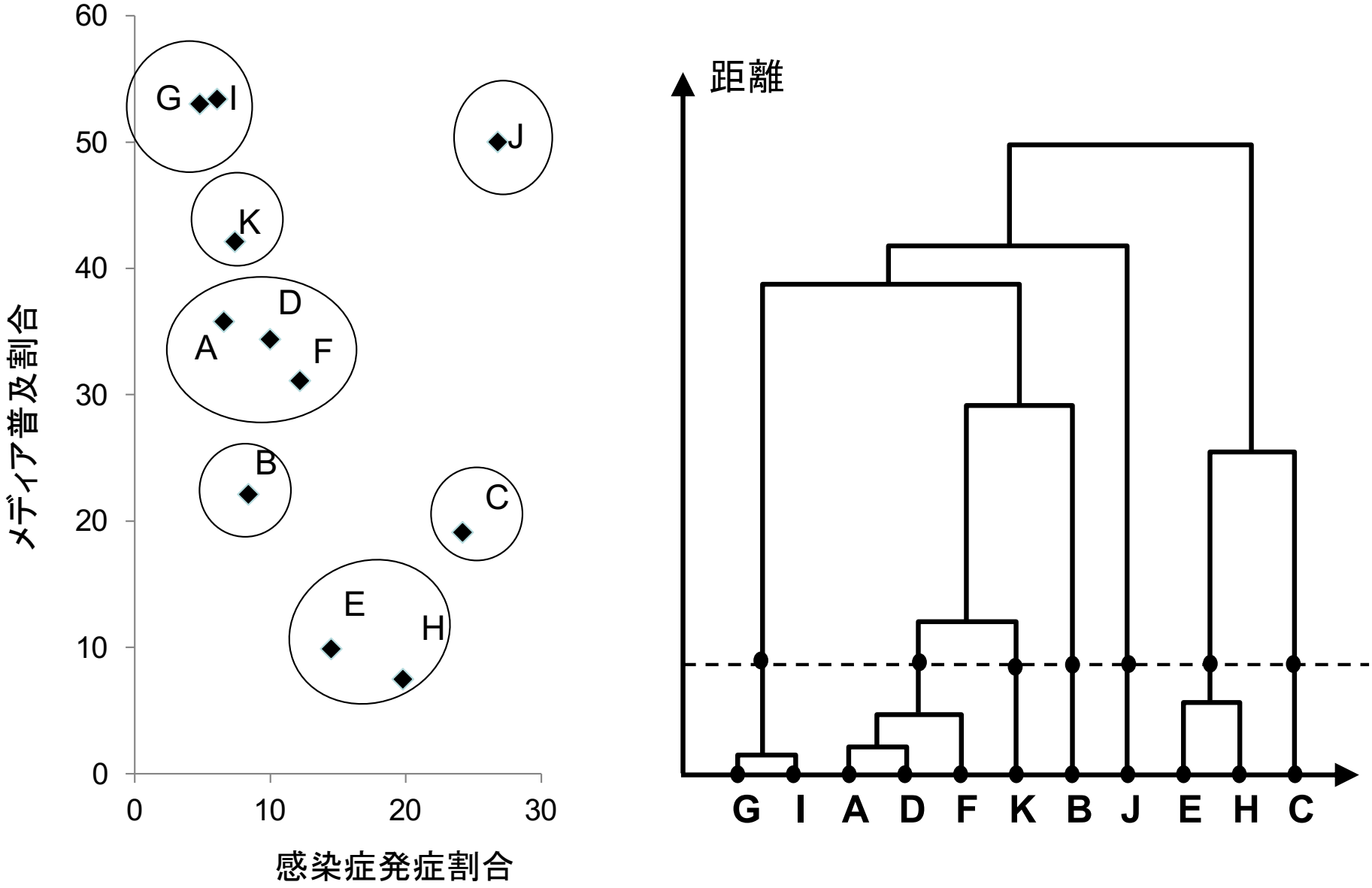
クラスター分析の手順（最終）



デンドログラム完成

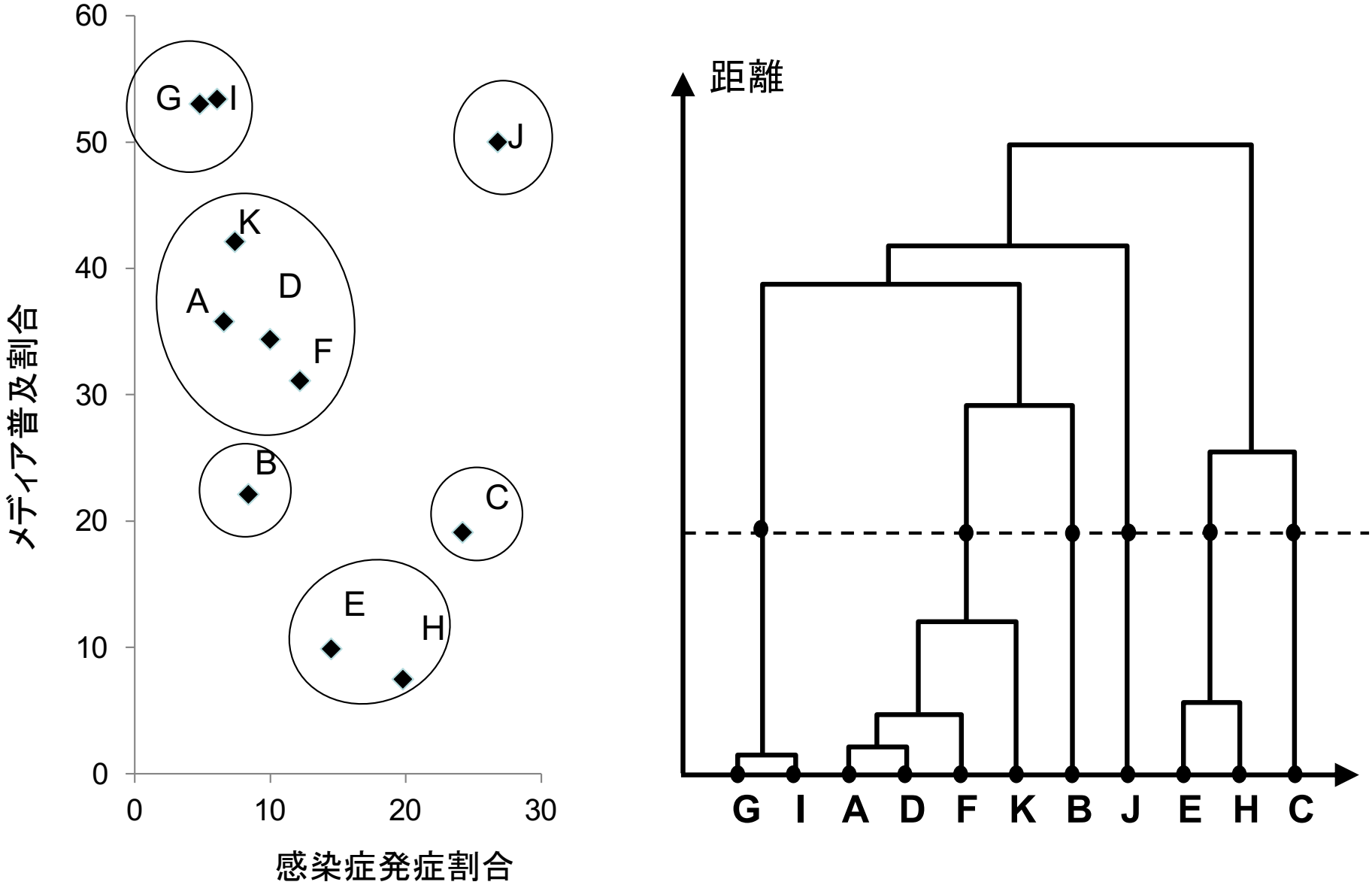
デンドログラムの使い方

- 縦軸が類似度を示す距離なので、横軸に平行に切って、
  デンドログラムの縦線とぶつかった個数がクラスタの個数になる



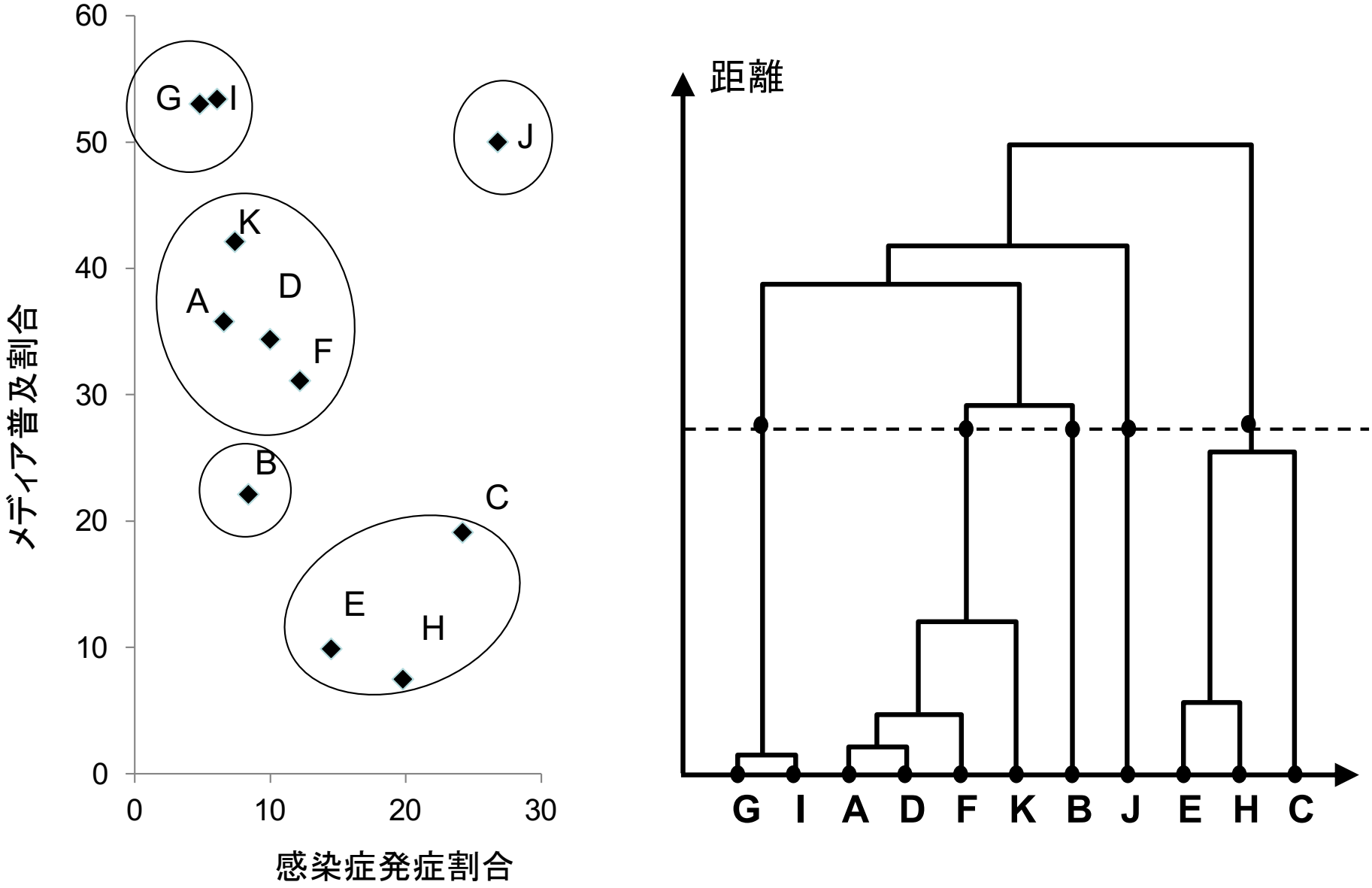
デンドログラムの使い方

- 縦軸が類似度を示す距離なので、横軸に平行に切って、
  デンドログラムの縦線とぶつかった個数がクラスタの個数になる



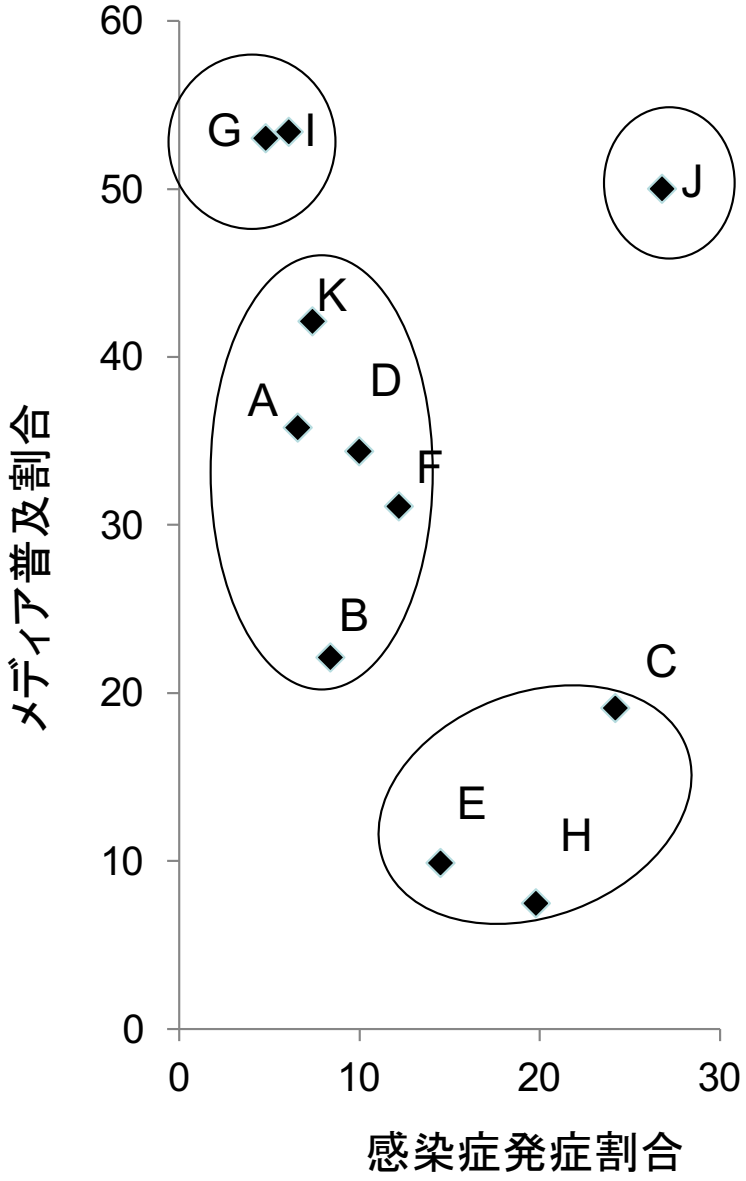
デンドログラムの使い方

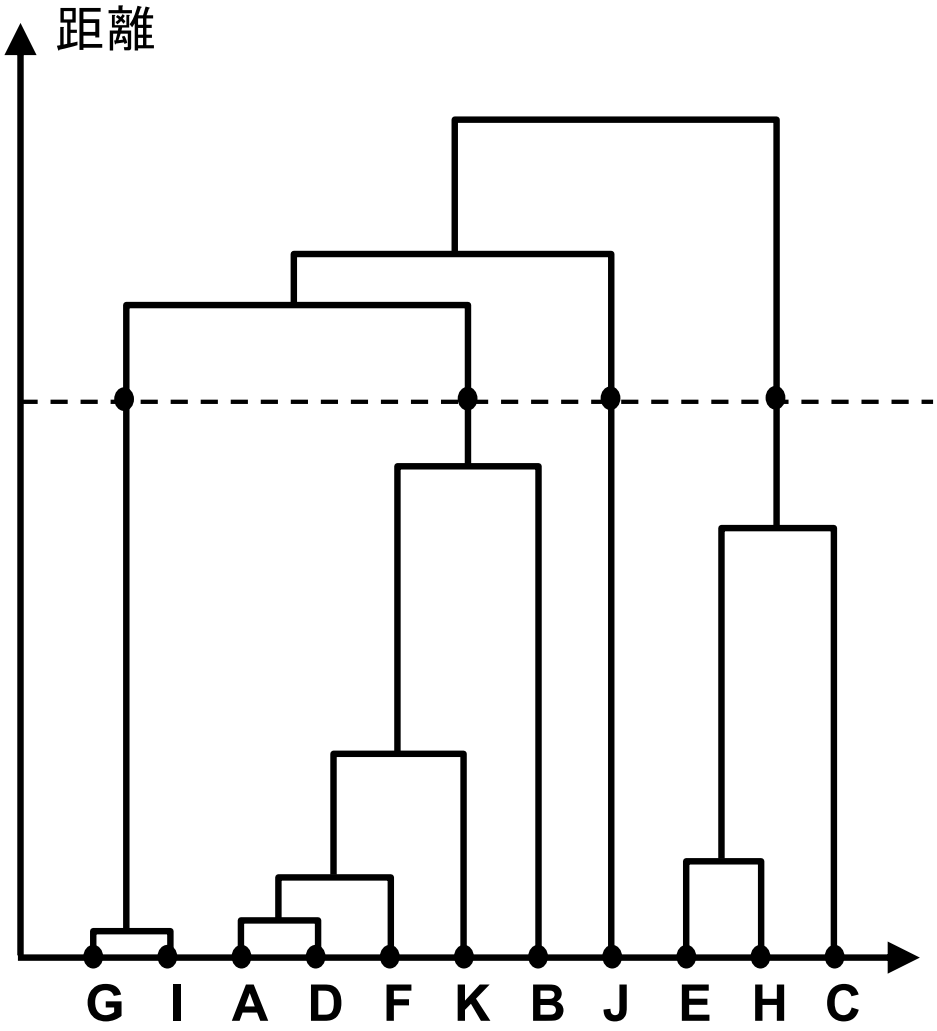
- 縦軸が類似度を示す距離なので、横軸に平行に切って、
  デンドログラムの縦線とぶつかった個数がクラスタの個数になる



デンドログラムの使い方

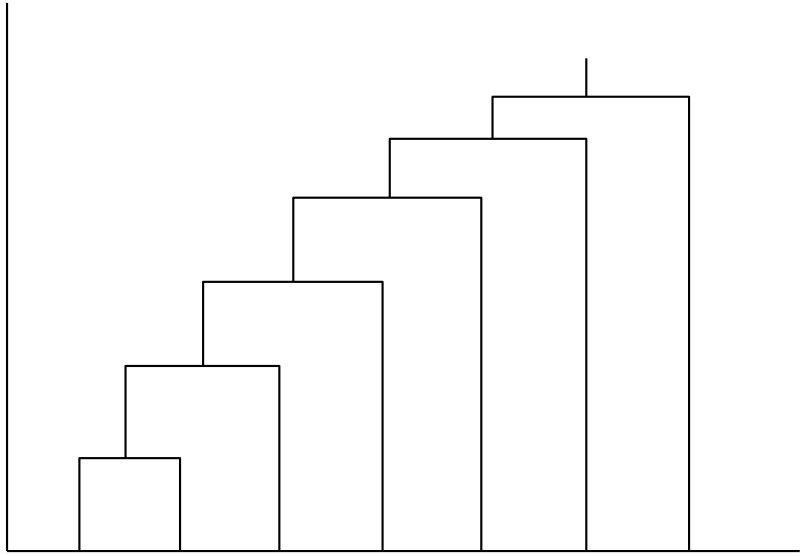
- 縦軸が類似度を示す距離なので、横軸に平行に切って、
  デンドログラムの縦線とぶつかった個数がクラスタの個数になる



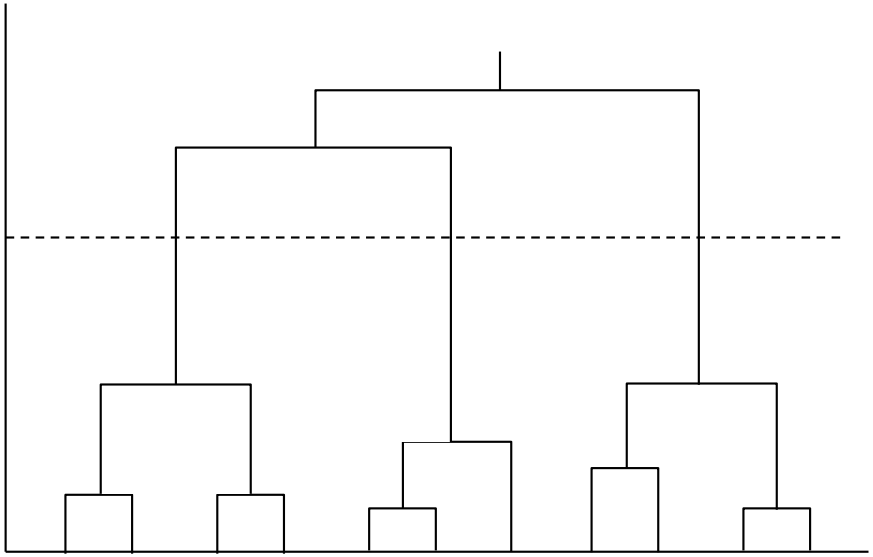


クラスター分析の目的

対象全体をいくつかのグループに分けて特徴を把握すること



鎖効果を示すデンドログラム



“よい”クラスター分析の結果を示すデンドログラム

鎖効果;
ある1つのクラスターに対象が1つずつ吸収されてクラスターが形成されていく現象。
従って、どの距離で切っても、あるクラスターとその他の対象」1つずつで構成され、グループに分けたことにならない

最短距離法では鎖効果が起こりやすく、
ウォード法では鎖効果が起こりにくいことが知られている

ウォード法

新たに統合されるクラスター内の平方和が最小となるようにクラスターをまとめる方法

変数が2個の場合のウォード法

国語と英語の成績(5段階評価)

生徒No.	国語x_1	英語x_2
1	5	1
2	4	2
3	1	5
4	5	4
5	5	5

対象間のウォード法における距離(1)

生徒No.	1	2	3	4
1				
2	1.00			
3	16.00	9.00		
4	4.50	2.50	8.50	
5	8.00	5.00	8.00	0.50

No.1とNo.2の生徒を統合した時のクラスター内での平方和S_{12}

$$S_{12} = \sum_{i=1}^{2}\sum_{k=1}^{2}(x_{ik} - \bar{x}_{\cdot k})^2$$
$$= (5-4.5)^2 + (4-4.5)^2 + (1-1.5)^2 + (2-1.5)^2$$
$$= 1.00$$

対象間のウォード法における距離(1)

生徒No.	1	2	3	4
1				
2	1.00			
3	16.00	9.00		
4	4.50	2.50	8.50	
5	8.00	5.00	8.00	0.50

国語と英語の成績(5段階評価)

生徒No.	国語x_1	英語x_2
1	5	1
2	4	2
3	1	5
4	5	4
5	5	5

対象間のウォード法における距離(1)で、
次に統合した時のクラスター内平方和の増加分が最小のものを統合
→　No.4とNo.5の統合が増加分が最小

C1(4,5)とNo.1~No.3の各対象を統合し、その時の平方和の増加分を計算

C1(4,5)とNo.1では

$$S_{145} = (5-5.00)^2 + (5-5.00)^2 + (5-5.00)^2 + (1-3.33)^2 + (4-3.33)^2 + (5-3.33)^2 = 8.67$$

平方和の増加分ΔS_{145}は
統合前の平方和が$S_1 = 0,\ S_{45} = 0.5$であるので

$$\Delta S_{145} = S_{145} - S_1 - S_{45} = 8.67 - 0 - 0.50 = 8.17$$

従って、クラスターC1(4,5)とNo.1の距離は8.17
同様に$\Delta S_{245} = 4.83$, $\Delta S_{345} = 10.83$ を求める

対象間のウォード法における距離(2)

生徒No.	1	2	3
1			
2	1.00		
3	16.00	9.00	
C1(4,5)	8.17	4.83	10.83

対象間のウォード法における距離(2)

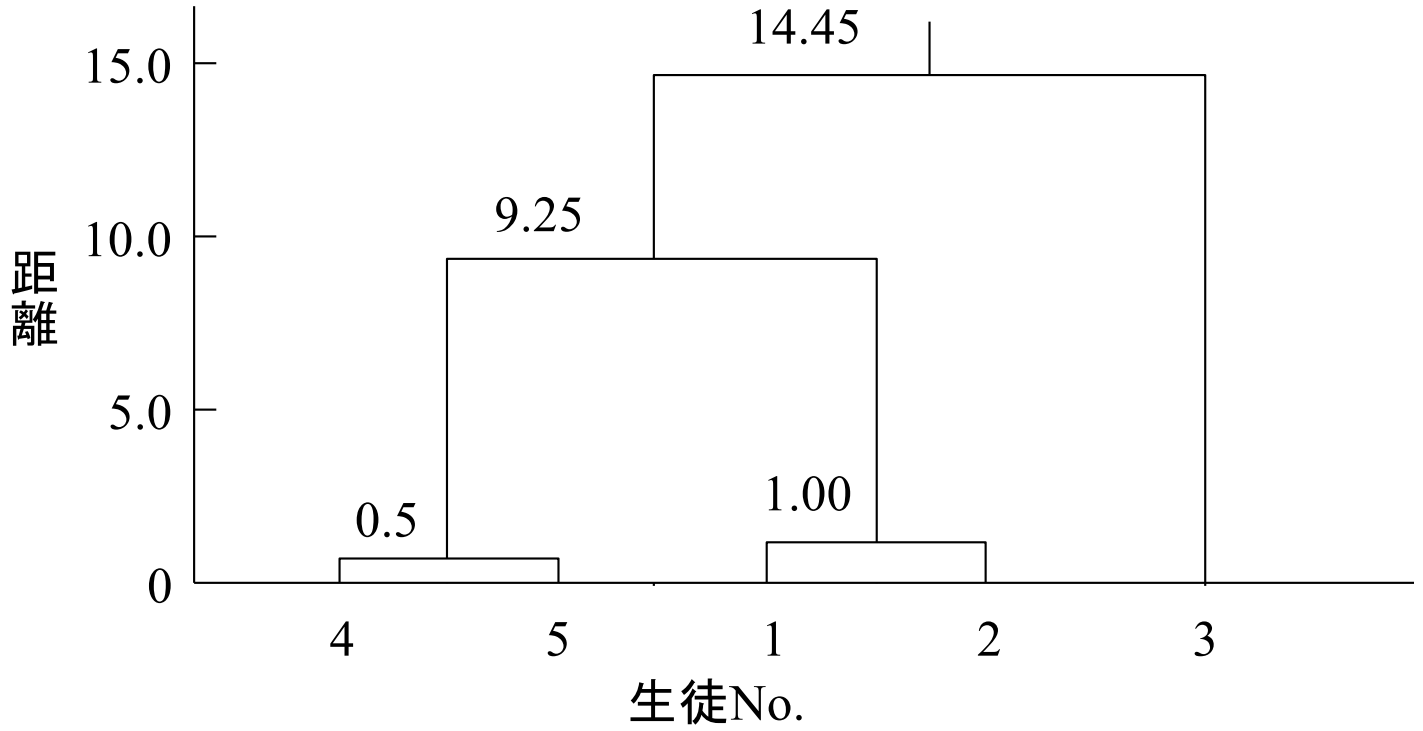
生徒No.	1	2	3
1			
2	1.00		
3	16.00	9.00	
C1(4,5)	8.17	4.83	10.83

対象間のウォード法における距離(3)

生徒No.	C2(1,2)	3
C2(1,2)		
3	16.33	
C1(4,5)	9.25	10.83

対象間のウォード法における距離(3)

生徒No.	C3(1,2,4,5)
C3(1,2,4,5)	
3	14.45



国語と英語の成績データのデンドログラム

変数がp個の場合のウォード法

変数が3個以上の場合も考え方は前述の2個の場合と同様

平方和の増加分の一般式

クラスタlとクラスタmを統合してクラスタlmを作成する場合
以下の関係が成り立つ

x_{lik}, x_{mik}；クラスタlとクラスタmに属する第k変数のi番目のデータ

n_l, n_m；サンプルサイズ

$$S_l = \sum_{i=1}^{n_l}\sum_{k=1}^{p}(x_{lik} - \bar{x}_{l\cdot k})^2$$

$$S_m = \sum_{i=1}^{n_m}\sum_{k=1}^{p}(x_{mik} - \bar{x}_{m\cdot k})^2$$

$$S_{lm} = S_l + S_m + \Delta S_{lm}$$

$$\Delta S_{lm} = \frac{n_l n_m}{n_l + n_m}\sum_{k=1}^{p}(\bar{x}_{l\cdot k} - \bar{x}_{m\cdot k})^2$$

$$S_{lm} = \sum_{i=1}^{n_l}\sum_{k=1}^{p}(x_{lik} - \bar{x}_k)^2 + \sum_{i=1}^{n_m}\sum_{k=1}^{p}(x_{mik} - \bar{x}_k)^2$$

$$\bar{x}_{l\cdot k} = \frac{1}{n_l}\sum_{i=1}^{n_l} x_{lik} \qquad \bar{x}_{m\cdot k} = \frac{1}{n_m}\sum_{i=1}^{n_m} x_{mik}$$

$$\bar{x}_k = \frac{n_l \bar{x}_{l\cdot k} + n_m \bar{x}_{m\cdot k}}{n_l + n_m}$$

デンドログラムの問題点

- 最適のクラスタの個数は何個か？
  → はっきりとした基準がない

- 「何個のクラスタに分類するか」、
  「それらの特徴は何か」は
  そのデータを研究している人に任されている
  (解析者の意図が入る)

クラスター分析

①クラスター分析とは何か。語句で説明せよ。

②高齢者の転倒事故が多く見られる住宅空間についての調査データを因子分析した結果、左下に示す表の結果を得た。この結果をクラスター分析しデンドログラムに描いたものが右下の図である。因子分析の結果を散布図に示し4つのクラスターに分け表示せよ。

	因子	
	水まわり	段差のある所
浴室	.858	-.098
食堂	.844	.039
トイレ	.744	.023
廊下	.634	.168
庭	.602	.029
玄関	.063	.818
ベランダ	.292	.459
階段	-.101	.372
居間	.161	-.091
寝室	-.311	.211