2024/11/19 Tue.

問題解決技法入門

# 3. Data Analysis

# 3. Cluster Analysis using R

堀田 敬介

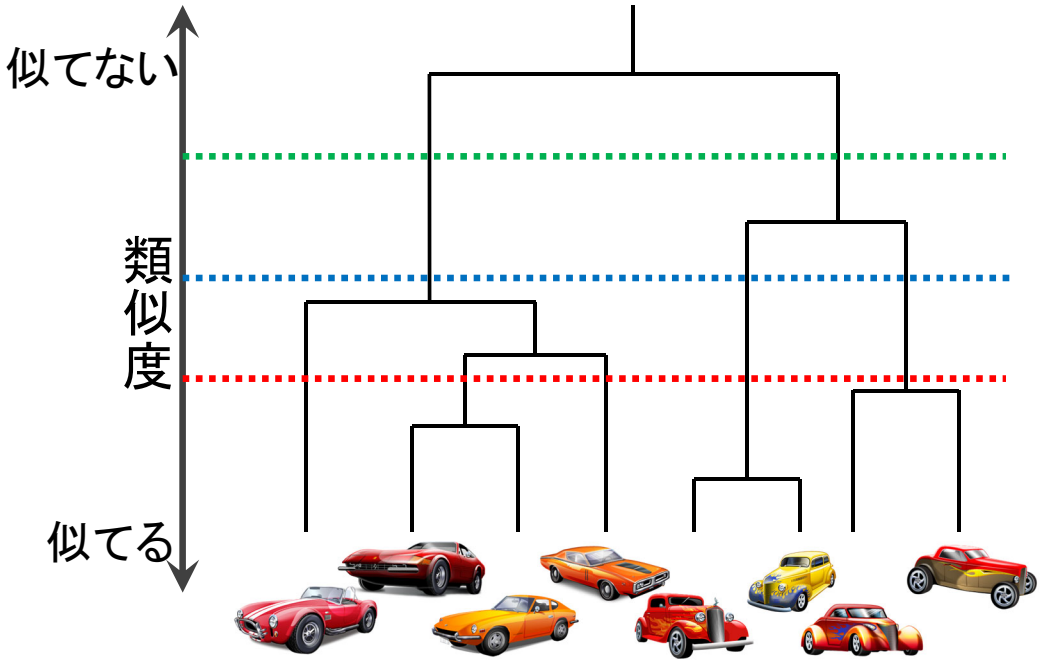
# クラスター分析とは

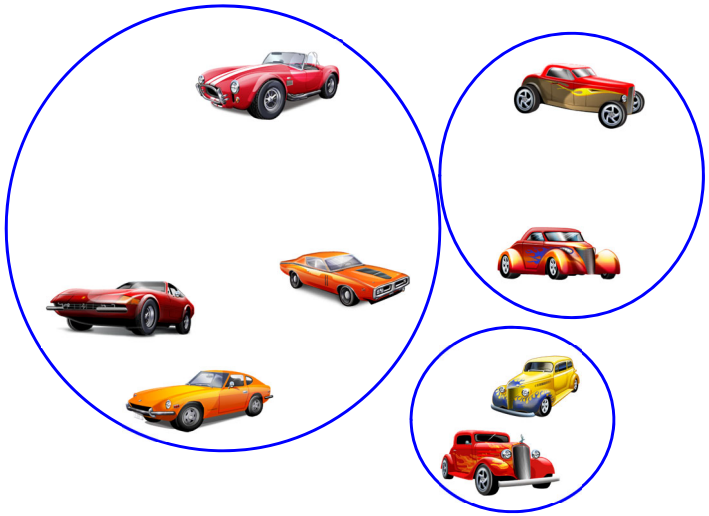
- クラスタ分析とは？
  - 複数の対象（もの，変数など）を，その**属性**によって**類似度（similarity）**をはかり，均質な集団（cluster）に分類する方法の総称

どれとどれが似てる？
（同じクラスター？）

# クラスター分析とは

- クラスタ分析の種類
  - 階層的方法
    - 樹形図（デンドログラム）を作成
    - 目的により高さを決めてクラスタリング
  - 非階層的方法
    - 予めクラスタ数を決めて（or 決まっていて）クラスタリングを行う

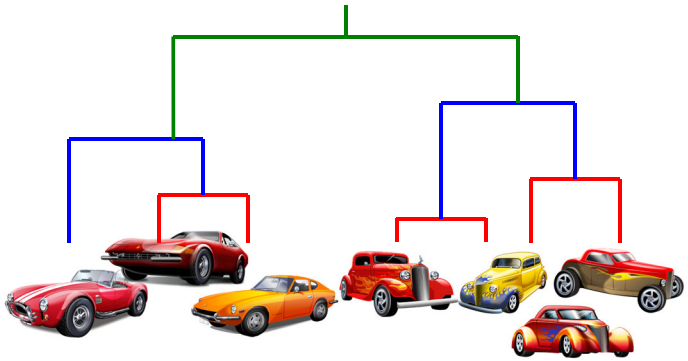


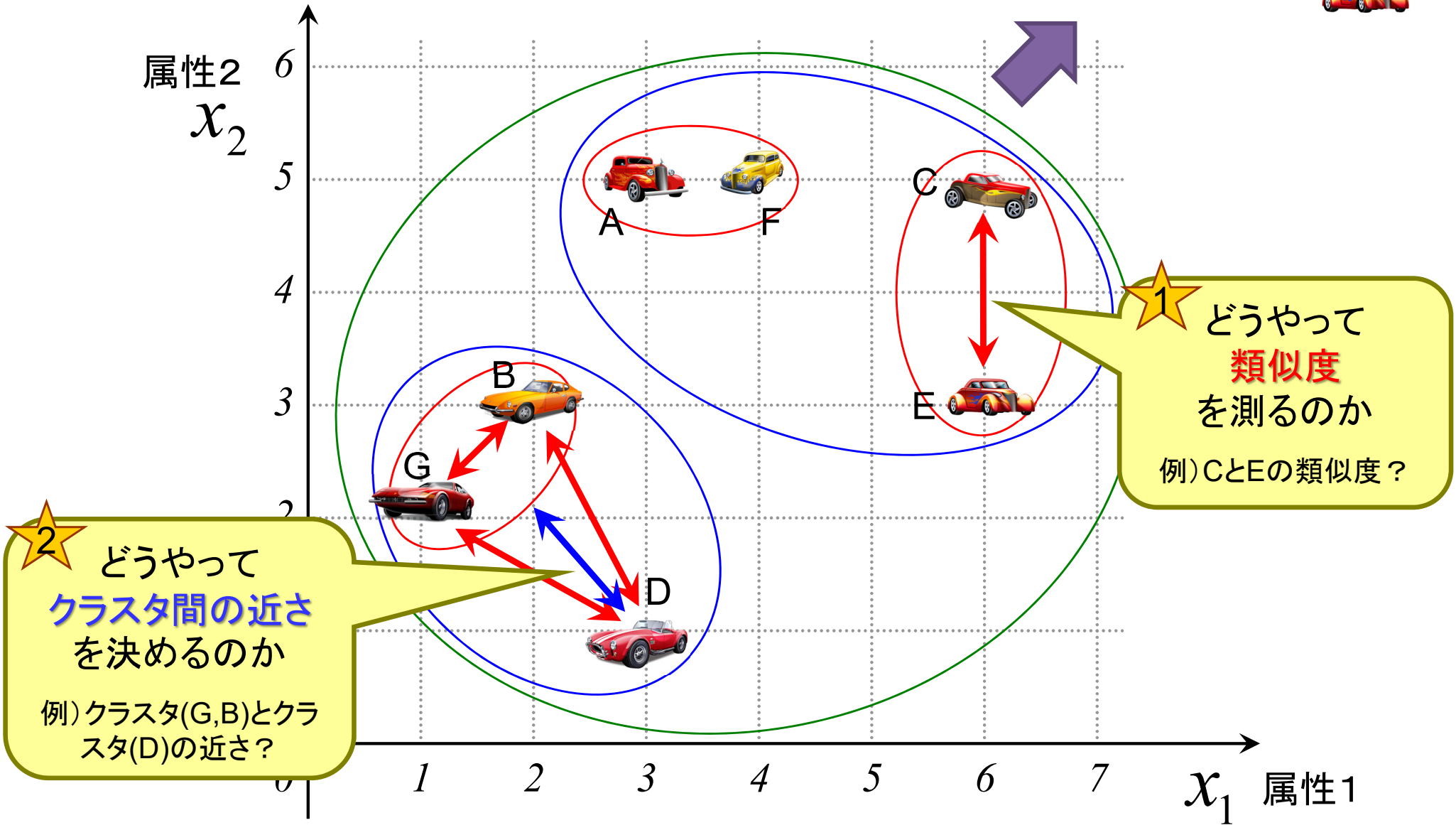



例：3つのクラスタに分類

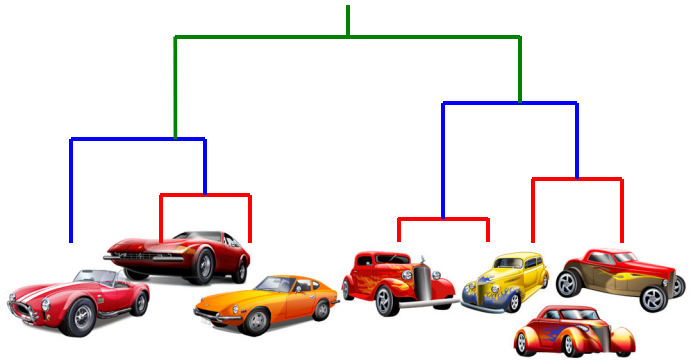
# 1. クラスタ分析概要

– 例：階層的方法（対象の属性が2つの場合）





# 1. クラスタ分析概要



- どうやって類似度を測るか？

| | | |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| | | | 3 | 1 | 2 | 3 | 4 | 6 | 6 |
| | $x_1$ | $x_2$ | 1 | 2 | 3 | 5 | 5 | 5 | 3 |
|  | 3 | 1 | | | | | | | |
|  | 1 | 2 | | | | | | | |
|  | 2 | 3 | | | | | | | |
|  | 3 | 5 | | | | | | | |
|  | 4 | 5 | | | | | | | |
|  | 6 | 5 | | | | | | | |
|  | 6 | 3 | | | | | | | |

# 2. 類似度の測定

類似度は尺度により距離や相関で測る
（距離：近いほうが類似）
（相関：高いほうが類似）

- 距離【間隔尺度】
  - ユークリッド距離
  - ユークリッド平方距離
  - 重み付きユークリッド距離
  - マンハッタン距離
  - ミンコフスキー距離
  - マハラノビス汎距離
- 相関【間隔尺度】
  - Pearsonの積率相関係数
  - ベクトル内積
- 相関【順序尺度】
  - Spearmanの順位相関係数
  - Kendallの順位相関係数
- 距離【名義尺度 [0, 1]】
  - 類似比
  - 一致係数
  - Russel-Rao係数
  - Rogers-Tanimoto係数
  - Hamann係数
  - ファイ係数
- 変量間類似度【名義尺度】
  - 平均平方根一致係数
  - グッドマン・クラスカルのλ

クラスタ分析ノート.pdf

# 2. 類似度の測定

- データと尺度

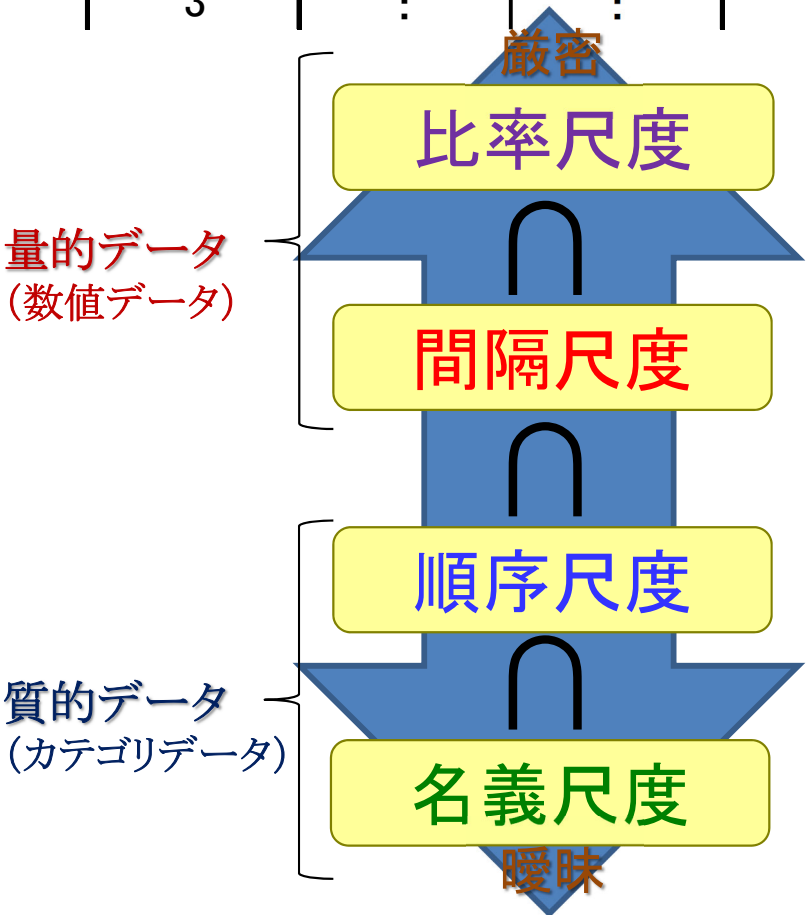
| 学籍番号 | 氏名 | 性別 | 生年月日 | 身長 | 体重 | 問題発見技法成績 | … |
|---|---|---|---|---|---|---|---|
| 1 | 文教太郎 | 男 | 1987.5.6 | 175cm | 69kg | B | … |
| 2 | 湘南花子 | 女 | 1988.1.4 | 163cm | 48kg | AA | … |
| 3 | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | |

名義尺度　名義尺度　名義尺度　名義尺度　名義尺度　名義尺度　順序尺度　名義尺度

順序尺度　順序尺度　順序尺度

間隔尺度　間隔尺度　間隔尺度

比率尺度　比率尺度

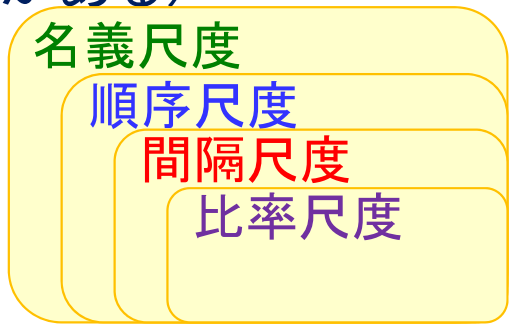



比に意味がある（絶対原点が存在する）
例）身長　180cmのAさんは息子(100cm)の1.8倍背が高い

差に意味がある
例）温度　気温20℃より30℃の方が10℃高い

順序関係がある（順序に意味がある）
例）成績評価　（A > B > C > D）

単なる分類，区別ができる
例）名前，性別



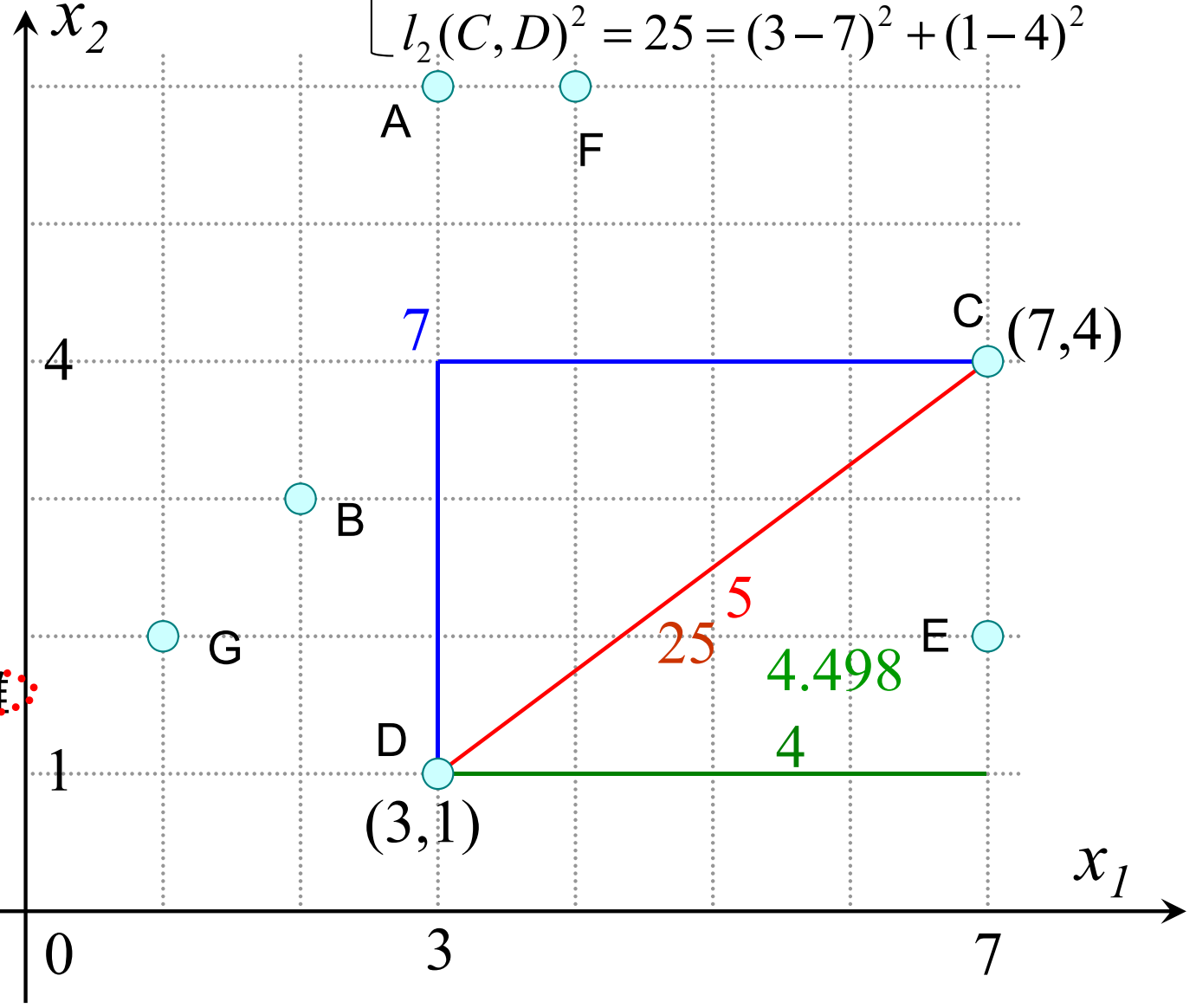
# 2. 類似度の測定

- 個体間類似度
  - ユークリッド距離
    （cf. $l_2$-ノルム）
  - マンハッタン距離
    （cf. $l_1$-ノルム）
  - ミンコフスキー距離
    （cf. $l_p$-ノルム）
    （cf. $l_\infty$-ノルム）
  - マハラノビス汎距離
  - ユークリッド平方距離

クラスター分析でよく使われる

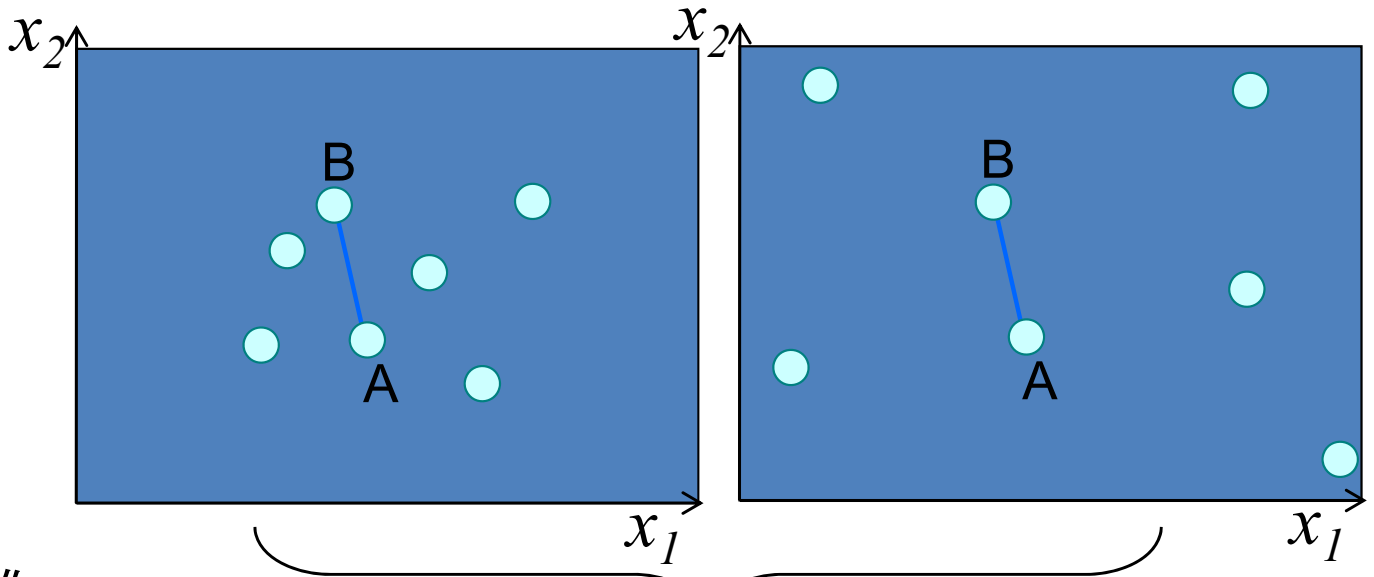
（注：各ノルムとは2変量の差ベクトルに対するノルム）

$$\begin{cases} l_2(C,D) = 5 = \sqrt{(3-7)^2 + (1-4)^2} \\ l_1(C,D) = 7 = |3-7| + |1-4| \\ l_3(C,D) = 4.498 = \sqrt[3]{|3-7|^3 + |1-4|^3} \\ l_\infty(C,D) = 4 = \max\{|3-7|, |1-4|\} \\ l_2(C,D)^2 = 25 = (3-7)^2 + (1-4)^2 \end{cases}$$



# 2. 類似度の測定

- 個体間類似度
  - ユークリッド距離
    （cf. $l_2$-ノルム）
  - マンハッタン距離
    （cf. $l_1$-ノルム）
  - ミンコフスキー距離
    （cf. $l_p$-ノルム）
    （cf. $l_\infty$-ノルム）
  - マハラノビス汎距離




左側の対象内での，A-B間距離と
右側の対象内でのA-B間距離が
異なる！（ユークリッド距離などでは同じ）

2変量版 $\boldsymbol{x}=(x_1, x_2)$

$$D \equiv \sqrt{\frac{u_1^2 + u_2^2 - 2\rho u_1 u_2}{1-\rho^2}}$$

$u_1, u_2$ は $x_1, x_2$ の標準化変量で，

$$u_1 = \frac{x_1 - \mu_1}{\sigma_2}, u_2 = \frac{x_2 - \mu_2}{\sigma_2}$$

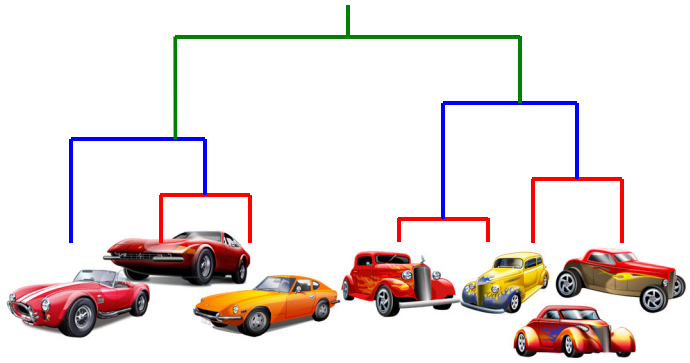
- $\mu_1, \mu_2$ はそれぞれ $x_1, x_2$ の平均
- $\sigma_1, \sigma_2$ はそれぞれ $x_1, x_2$ の標準偏差
- $\rho$ は $x_1, x_2$ の相関係数

多変量版 $\boldsymbol{x}=(x_1, \ldots, x_m)$

$$D \equiv (x_p - x_q)^T \Sigma^{-1} (x_p - x_q)$$

$\Sigma$ は $\boldsymbol{x}_p, \boldsymbol{x}_q$ の分散共分散行列

# 2. 類似度の測定

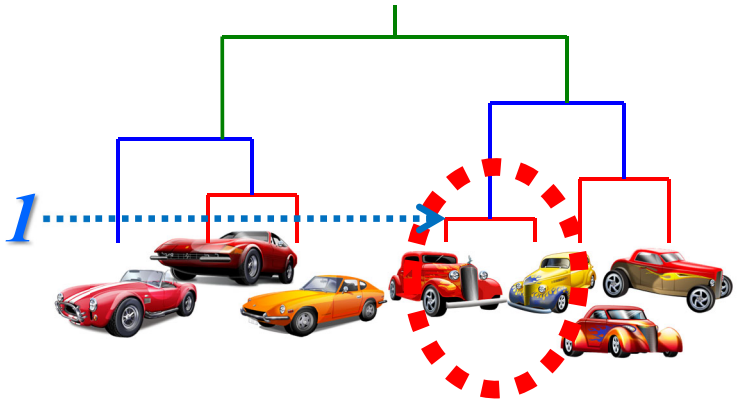


- どうやって類似度を測るか？
  - 例：ユークリッド平方距離

| | | |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| | | | 3 | 1 | 2 | 3 | 4 | 6 | 6 |
| | $x_1$ | $x_2$ | 1 | 2 | 3 | 5 | 5 | 5 | 3 |
|  | 3 | 1 | | 5 | 5 | 16 | 17 | 25 | 13 |
|  | 1 | 2 | | | 2 | 13 | 18 | 34 | 26 |
|  | 2 | 3 | | | | 5 | 8 | 20 | 16 |
|  | 3 | 5 | | | | | 1 | 9 | 13 |
|  | 4 | 5 | | | | | | 4 | 8 |
|  | 6 | 5 | | | | | | | 4 |
|  | 6 | 3 | | | | | | | |

# 2. 類似度の測定

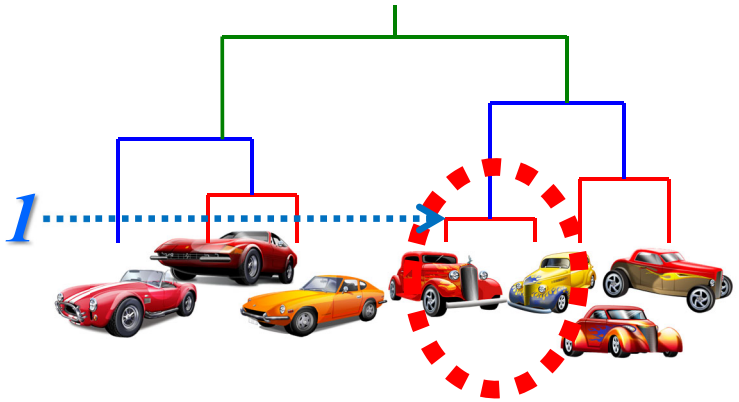
- どうやって類似度を更新するか？




| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 3 | 1 | 2 | 3 | 4 | 6 | 6 |
| | $x_1$ | $x_2$ | 1 | 2 | 3 | 5 | 5 | 5 | 3 |
| | 3 | 1 | | 5 | 5 | 16 | 17 | 25 | 13 |
| | 1 | 2 | | | 2 | 13 | 18 | 34 | 26 |
| | 2 | 3 | | | | 5 | 8 | 20 | 16 |
| | 3 | 5 | | | | | 1 | 9 | 13 |
| | 4 | 5 | | | | | | 4 | 8 |
| | 6 | 5 | | | | | | | 4 |
| | 6 | 3 | | | | | | | |

# 2. 類似度の測定

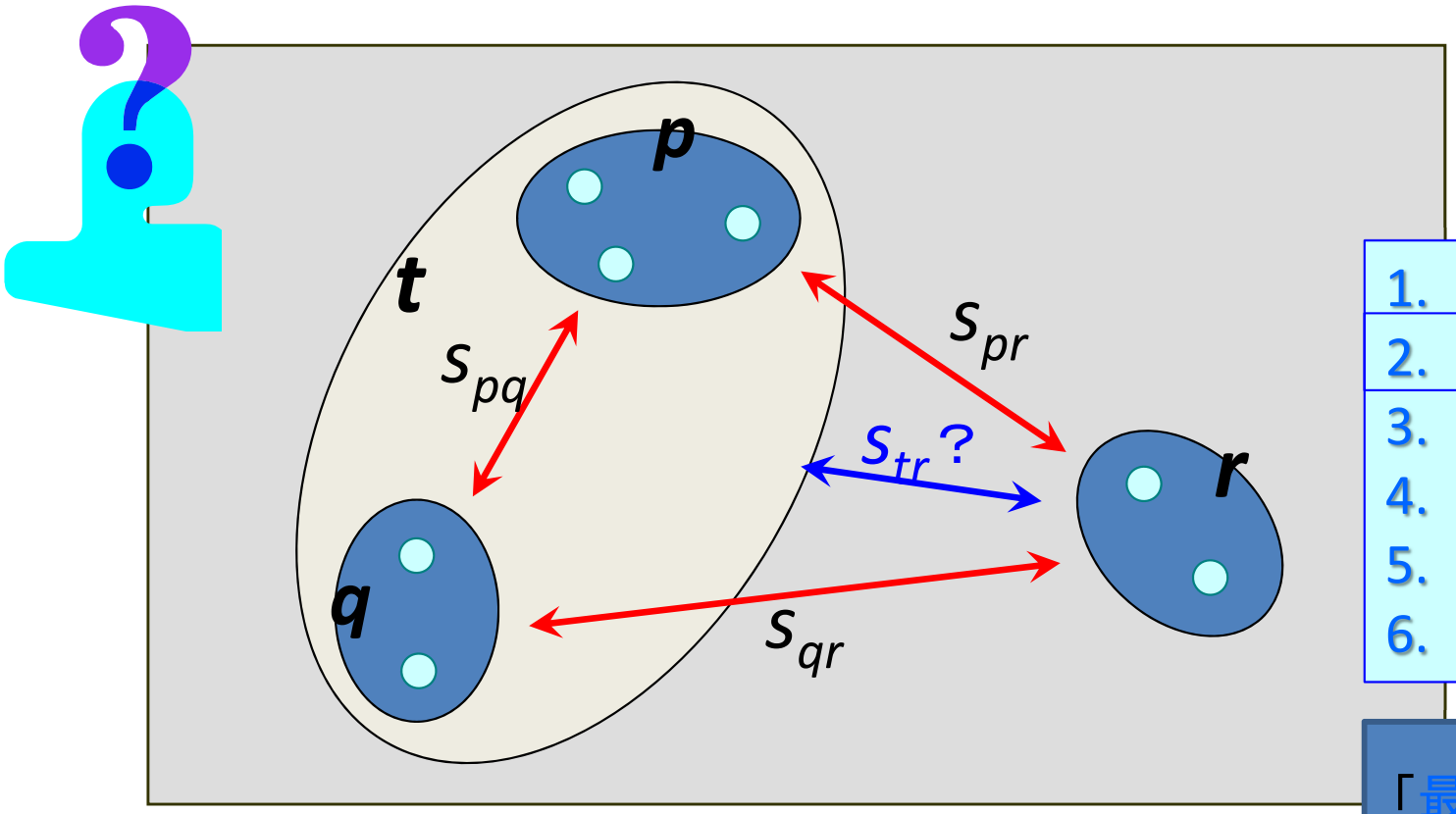
- どうやって類似度を更新するか？




| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |
| | | | 3 | 1 | 2 | 3,4 | 6 | 6 |
| | $x_1$ | $x_2$ | 1 | 2 | 3 | 5,5 | 5 | 3 |
| | 3 | 1 | | 5 | 5 | 16,17 | 25 | 13 |
| | 1 | 2 | | | 2 | 13,18 | 34 | 26 |
| | 2 | 3 | | | | 5,8 | 20 | 16 |
| | 3,4 | 5,5 | | | | 1 | 9,4 | 13,8 |
| | 6 | 5 | | | | | | 4 |
| | 6 | 3 | | | | | | |

# 3. クラスタ化の方法

- 新たなクラスタ生成時の**類似度の更新方法**
  - クラスタ***p***，クラスタ***q*** が一つのクラスタ***t*** になる場合，他のクラスタ***r*** との類似度をどう更新する？



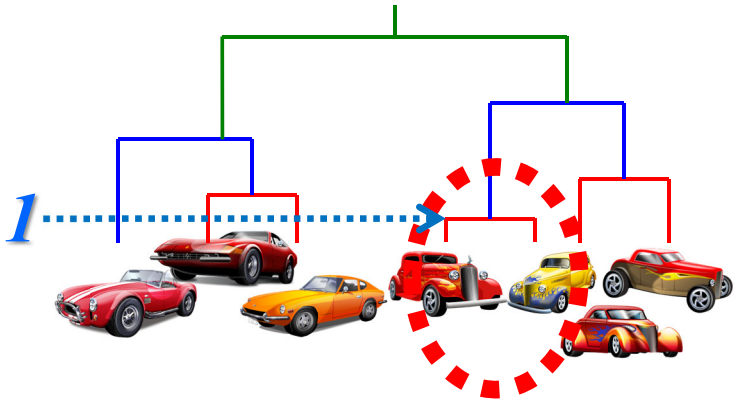

（$s_{pr}$ : クラスタ$p$, $r$の類似度）

1. 最短距離法
2. 最長距離法
3. 群平均法
4. 重心法
5. 中央値法
6. ウォード法

「最短」か「最長」か
何らかの「平均」

# 3. クラスタ化の方法

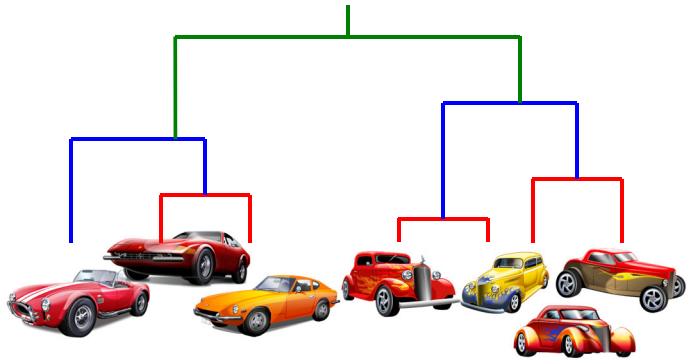
- どうやって類似度を更新するか？




| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |
| | | | 3 | 1 | 2 | 3:4 | 6 | 6 |
| | $x_1$ | $x_2$ | 1 | 2 | 3 | 5:5 | 5 | 3 |
| | 3 | 1 | | 5 | 5 | 16:17 | 25 | 13 |
| | 1 | 2 | | | 2 | 13:18 | 34 | 26 |
| | 2 | 3 | | | | 5:8 | 20 | 16 |
| | 3:4 | 5:5 | | | | 1 | 9:4 | 13:8 |
| | 6 | 5 | | | | | | 4 |
| | 6 | 3 | | | | | | |

# 3. クラスタ化の方法

- どうやって類似度を更新するか？

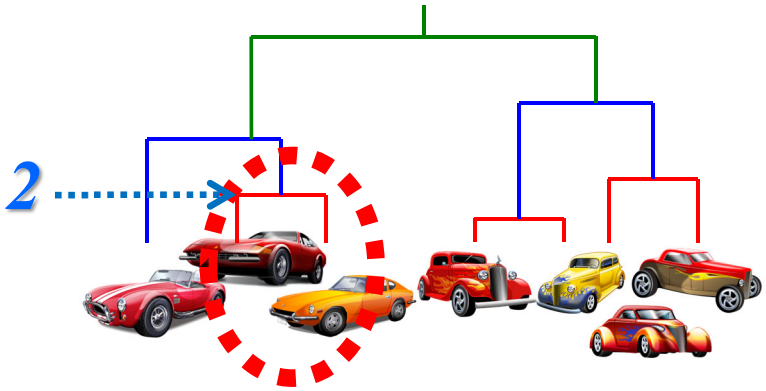


| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 3 | 1 | 2 | 3:4 | $\frac{1+1}{1+1+1}16+\frac{1+1}{1+1+1}17-\frac{1}{1+1+1}1$ | |
| | $x_1$ | $x_2$ | 1 | 2 | 3 | 5:5 | $\frac{1+1}{1+1+1}13+\frac{1+1}{1+1+1}18-\frac{1}{1+1+1}1$ | |
| | 3 | 1 | | 5 | 5 | 21.7 | 25 | 13 |
| | 1 | 2 | | | 2 | 20.3 | 34 | 26 |
| | 2 | 3 | | | | 8.3 | 20 | 16 |
| | 3:4 | 5:5 | $\frac{1+1}{1+1+1}5+\frac{1+1}{1+1+1}8-\frac{1}{1+1+1}1$ | | | 1 | 8.3 | 13.7 |
| | 6 | 5 | | | | | | 4 |
| | 6 | 3 | $\frac{1+1}{1+1+1}9+\frac{1+1}{1+1+1}4-\frac{1}{1+1+1}1$ | | | | | |

$$\frac{1+1}{1+1+1}13+\frac{1+1}{1+1+1}8-\frac{1}{1+1+1}1$$

# 3. クラスタ化の方法

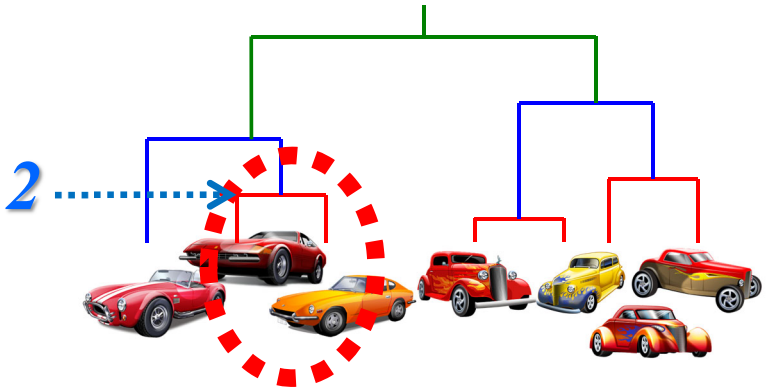
- どうやって類似度を**更新**するか？




| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |
| | | | | | | | | |
| | $x_1$ | $x_2$ | | | | | | |
| | | | | 5 | 5 | 21.7 | 25 | 13 |
| | | | | | 2 | 20.3 | 34 | 26 |
| | | | | | | 8.3 | 20 | 16 |
| | | | | | | | 8.3 | 13.7 |
| | | | | | | | | 4 |
| | | | | | | | | |

# 3. クラスタ化の方法

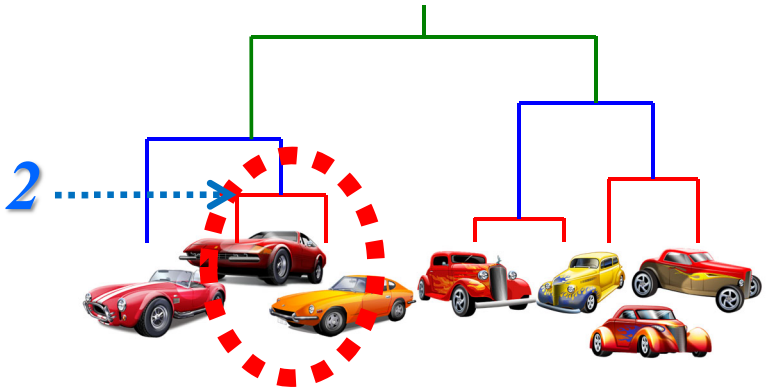
- どうやって類似度を**更新**するか？




| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| | | | | | | | |
| | $x_1$ | $x_2$ | | | | | |
| | | | | 5:5 | 21.7 | 25 | 13 |
| | | | | 2 | 20.3:8.3 | 34:20 | 26:16 |
| | | | | | | 8.3 | 13.7 |
| | | | | | | | 4 |
| | | | | | | | |

# 3. クラスタ化の方法

- どうやって類似度を**更新**するか？

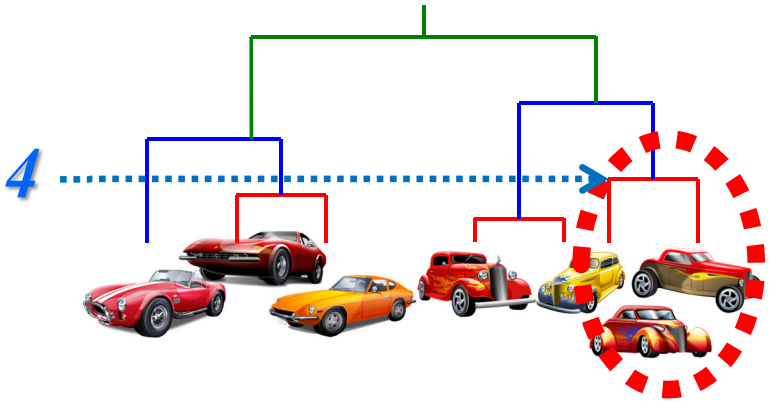



| | $x_1$ | $x_2$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | $\frac{1+2}{1+1+2}20.3+\frac{1+2}{1+1+2}8.3-\frac{2}{1+1+2}2$ | |
| | | | | 6 | 21.7 | 25 | 13 |
| | | | | 2 | 20.5 | 35.3 | 27.3 |
| | | | | | | 8.3 | 13.7 |
| | | | $\frac{1+1}{1+1+1}34+\frac{1+1}{1+1+1}20-\frac{1}{1+1+1}2$ | | | | 4 |
| | | | | | | | |

$$\frac{1+1}{1+1+1}26+\frac{1+1}{1+1+1}16-\frac{1}{1+1+1}2$$

# 3. クラスタ化の方法

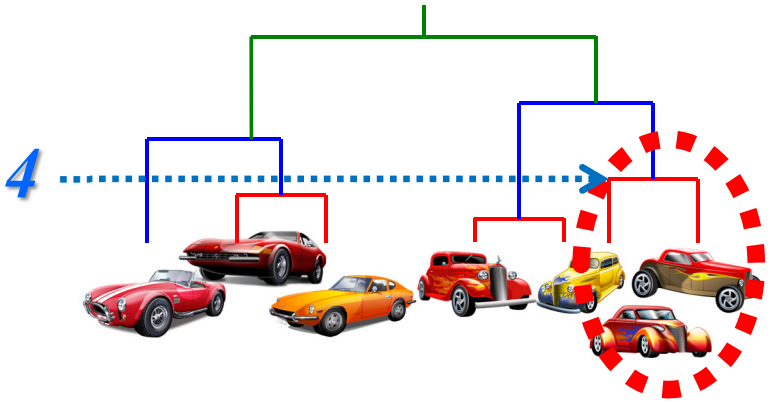
- どうやって類似度を**更新**するか？




| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| | | | | | | | |
| | $x_1$ | $x_2$ | | | | | |
| | | | | 6 | 21.7 | 25 | 13 |
| | | | | | 20.5 | 35.3 | 27.3 |
| | | | | | | 8.3 | 13.7 |
| | | | | | | | 4 |
| | | | | | | | |

# 3. クラスタ化の方法

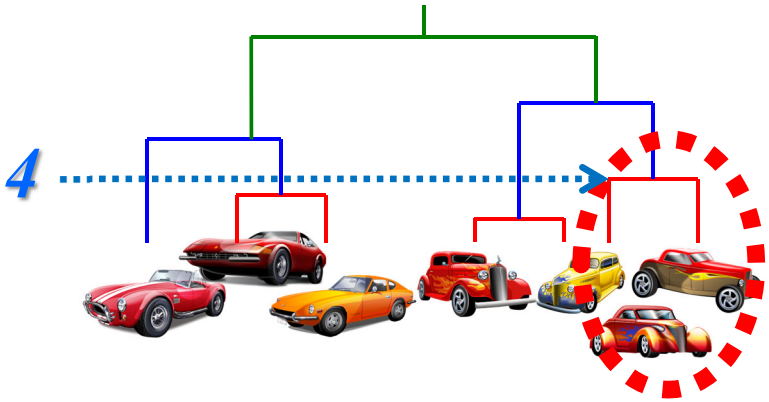
- どうやって類似度を**更新**するか？




| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | |
| | | | | | | |
| | $x_1$ | $x_2$ | | | | |
| | | | | 6 | 21.7 | 25:13 |
| | | | | | 20.5 | 35.3:27.3 |
| | | | | | | 8.3:13.7 |
| | | | | | | 4 |

# 3. クラスタ化の方法

- どうやって類似度を**更新**するか？

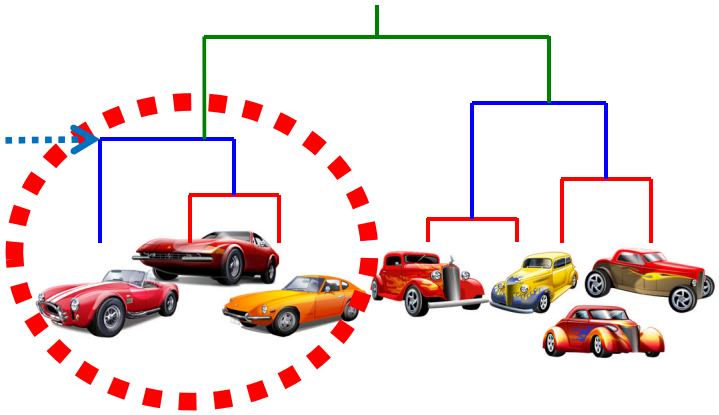



| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | |
| | | | | | | |
| | $x_1$ | $x_2$ | | | | |
| | | | | 6 | 21.7 | 24 |
| | | | | | 20.5 | 45 |
| | | | | | | 14.5 |
| | | | | | | 4 |

$$\frac{1+2}{1+1+2}8.3+\frac{1+2}{1+1+2}13.7-\frac{2}{1+1+2}4$$

# 3. クラスタ化の方法

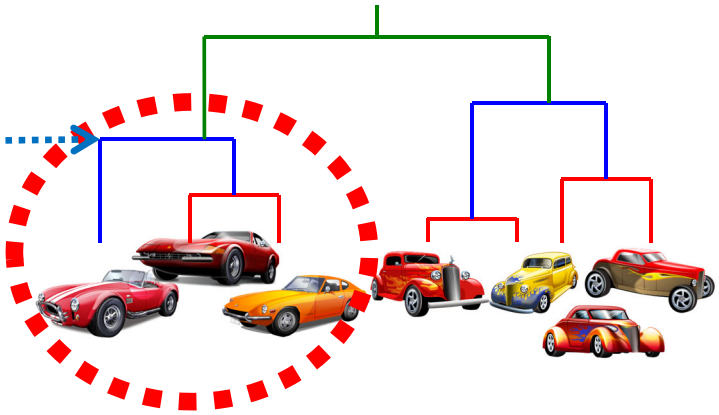
- どうやって類似度を**更新**するか？




| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | |
| | | | | | | |
| | $x_1$ | $x_2$ | | | | |
| | | | | 6 | 21.7 | 24 |
| | | | | | 20.5 | 45 |
| | | | | | | 14.5 |
| | | | | | | |

# 3. クラスタ化の方法

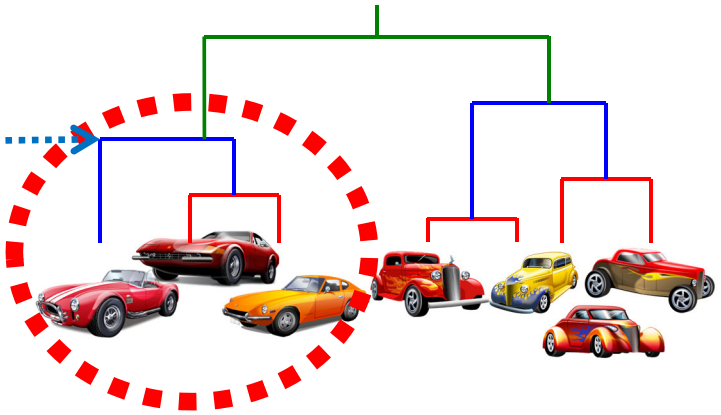
- どうやって類似度を**更新**するか？




| | | | | | |
|---|---|---|---|---|---|
| | | |  |  |  |
| | | | | | |
| | $x_1$ | $x_2$ | | | |
|  | | | 6 | 21.7<br>20.5 | 24<br>45 |
|  | | | | | 14.5 |
|  | | | | | |

# 3. クラスタ化の方法

- どうやって類似度を**更新**するか？



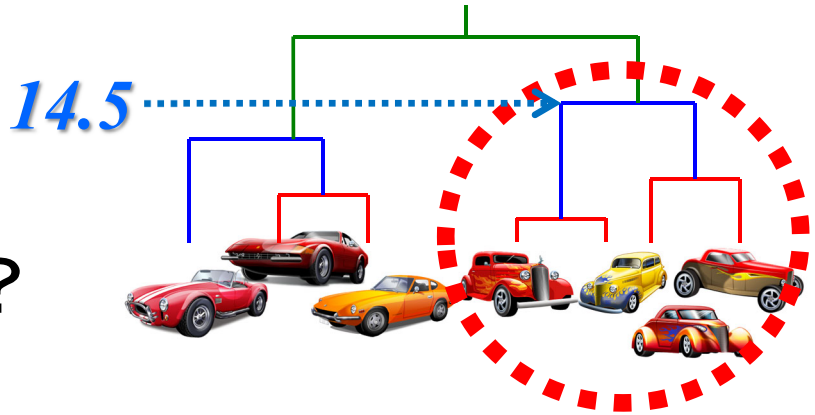

| | | | | | |
|---|---|---|---|---|---|
| | | | | | |
| | | | | | |
| | $x_1$ | $x_2$ | | | |
| | | | 6 | 27 | 48 |
| | | | | | 14.5 |
| | | | | | |

$$\frac{1+2}{1+2+2}21.7+\frac{2+2}{1+2+2}20.5-\frac{2}{1+2+2}6$$

$$\frac{1+2}{1+2+2}24+\frac{2+2}{1+2+2}45-\frac{2}{1+2+2}6$$

# 3. クラスタ化の方法

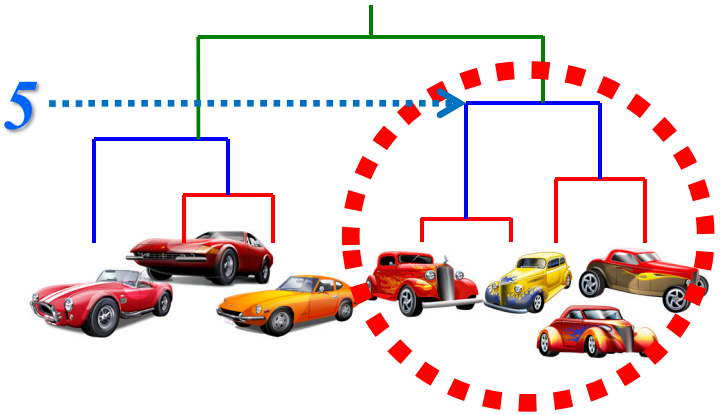
- どうやって類似度を**更新**するか？




| | | |  |  |  |
|---|---|---|---|---|---|
| | | | | | |
| | $x_1$ | $x_2$ | | | |
|  | | | | 27 | 48 |
|  | | | | | 14.5 |
|  | | | | | |

# 3. クラスタ化の方法

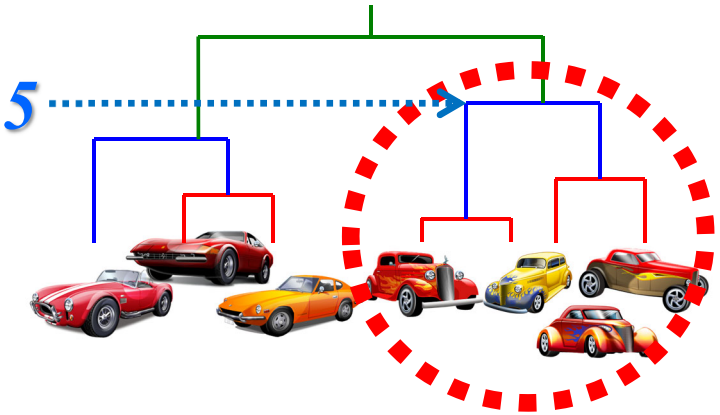
- どうやって類似度を**更新**するか？




| | | |  |  |
|---|---|---|---|---|
| | | | | |
| | $x_1$ | $x_2$ | | |
|  | | | | 27<br>48 |
|  | | | | 14.5 |

# 3. クラスタ化の方法

- どうやって類似度を**更新**するか？

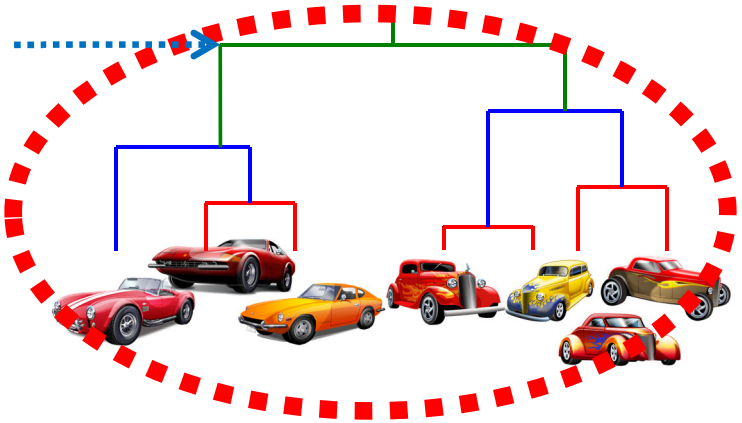



| | | | | |
|---|---|---|---|---|
| | | |  |  |
| | $x_1$ | $x_2$ | | |
|  | | | | 47.4 |
|  | | | | 14.5 |

$$\frac{2+3}{2+2+3}27+\frac{2+3}{2+2+3}48-\frac{3}{2+2+3}14.5$$

# 3. クラスタ化の方法

- どうやって類似度を**更新**するか？




| | | | | |
|---|---|---|---|---|
| | | |  |  |
| | $x_1$ | $x_2$ | | |
|  | | | | 47.4 |
|  | | | | |

# 4. R commanderでクラスター分析

## ① データの準備：csv ファイル

data-seiseki.csv

| | 算数 | 理科 | 国語 | 英語 | 社会 |
|---|---|---|---|---|---|
| 太郎 | 90 | 100 | 70 | 90 | 30 |
| 次郎 | 80 | 60 | 70 | 70 | 20 |
| 三郎 | 100 | 40 | 30 | 70 | 80 |
| 四郎 | 60 | 30 | 40 | 80 | 80 |
| 花子 | 30 | 60 | 80 | 90 | 90 |
| 寒子 | 50 | 60 | 40 | 30 | 60 |
| 湘子 | 90 | 100 | 90 | 80 | 70 |

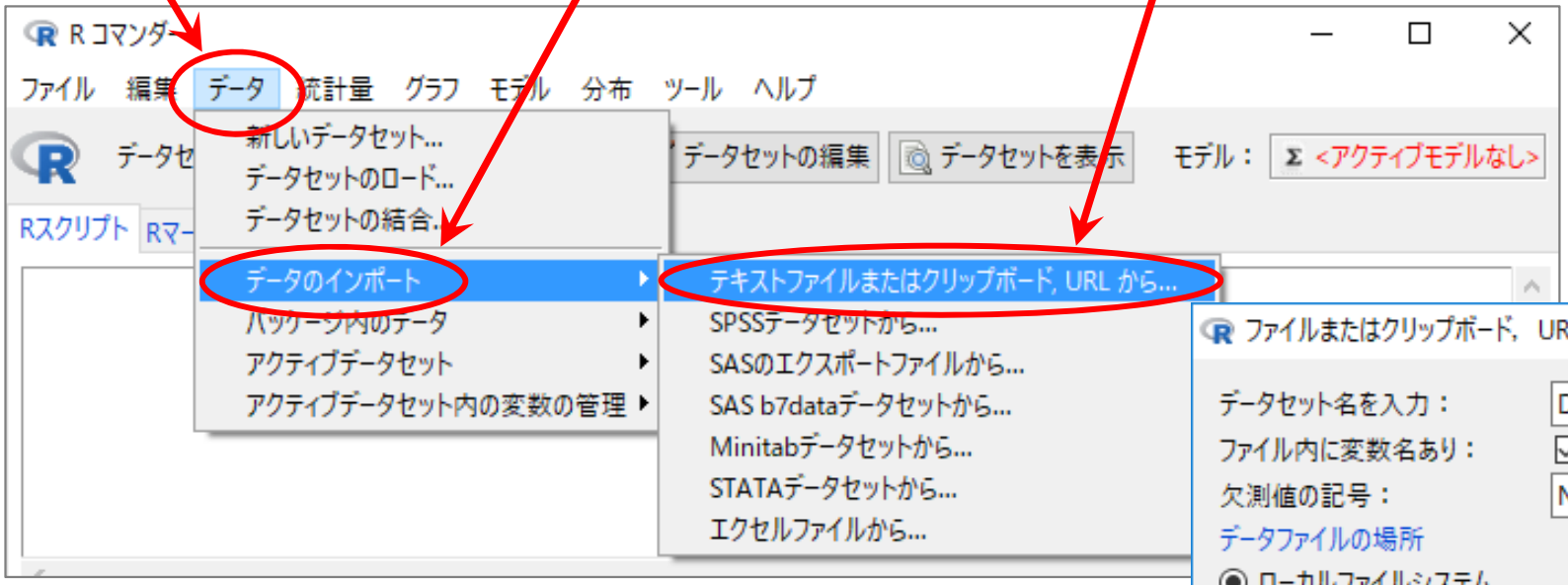
## ② Rの起動：「R x64 X.X.X」を選択

- 注）x64 = 64bit用のプログラム（アプリ），X.X.X = Rのバージョン
- 注）32bit PCの場合は，「R i386 X.X.X」を選択
- 注）起動すると「R Console」が開く．コマンドプロンプト（>）で「library(Rcmdr)」と打ち[Enter] → R commander が起動

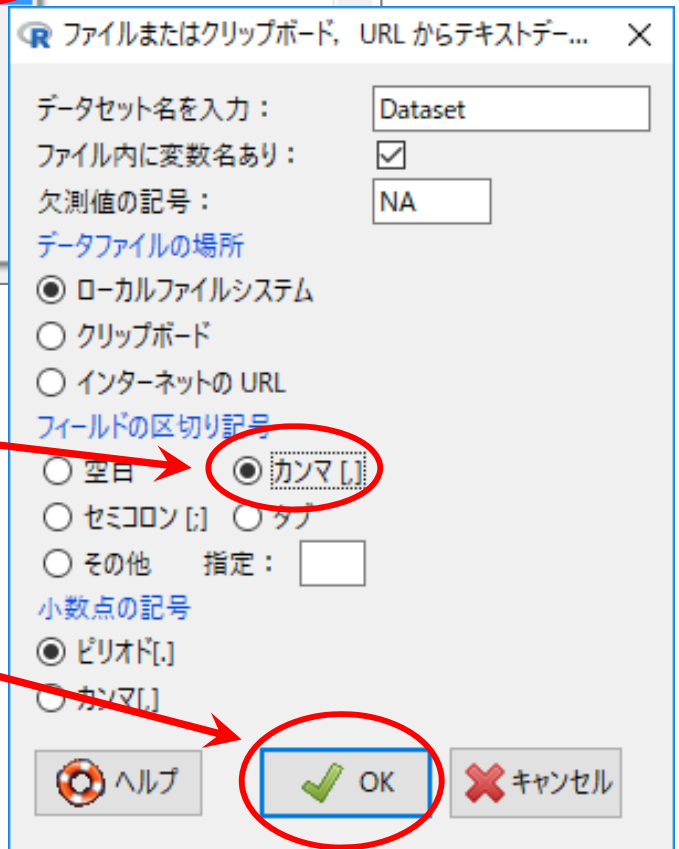
# 4. R commanderでクラスター分析

## ③ データの読込

- 「データ」ー「データのインポート」ー「テキストファイル…」を選択




- [フィールドの区切り記号]で[カンマ[,]]を選び，[OK]クリック



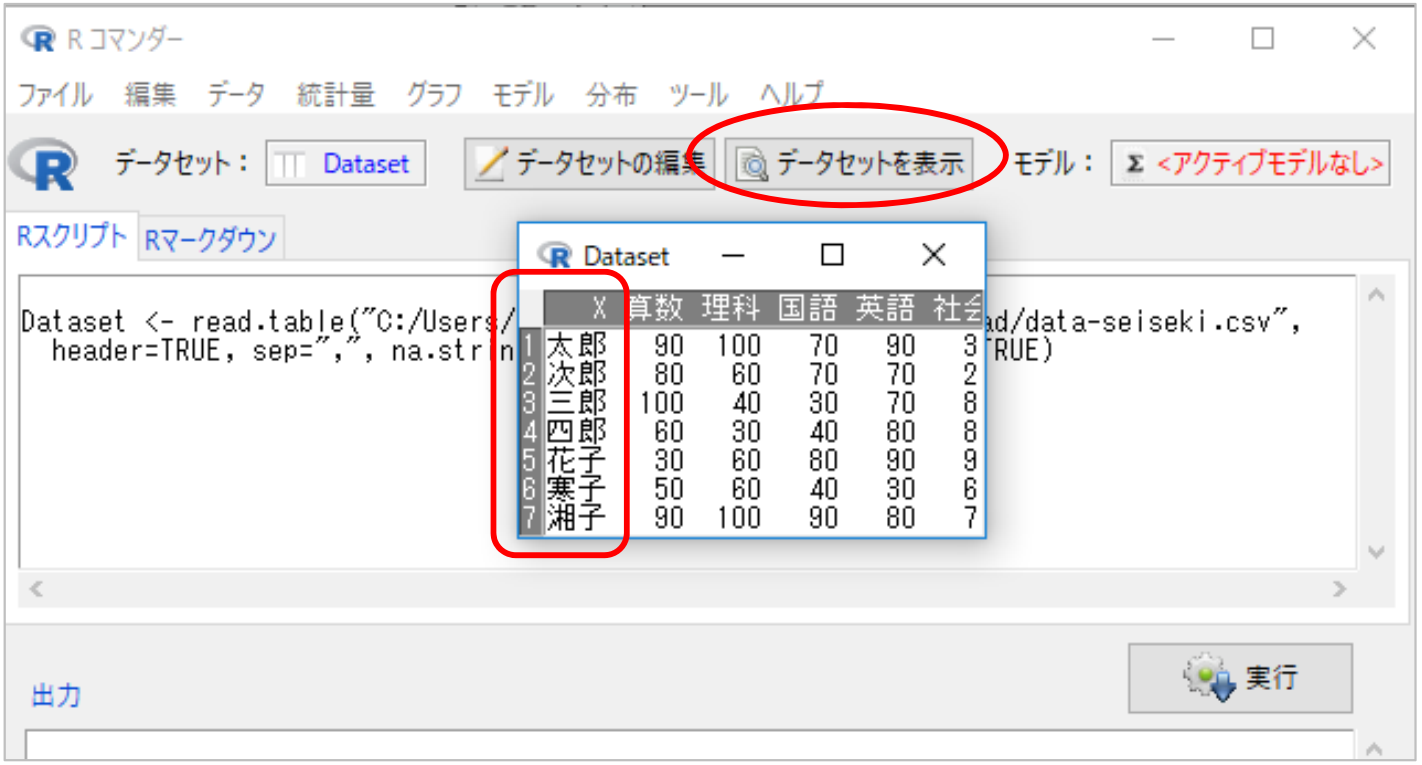

- ①で準備したファイル「*.csv」を選び[開く]

注)「データセットDatasetがすでに存在…上書き…」→[Yes]

# 4. R commanderでクラスター分析

## ③ データの読込（読み込んだファイルの確認）

- [データセットを表示]ボタンをクリックし，内容を確認
  - 注1）氏名の項目名が「X」であることを確認（もとのファイルに項目名がないデータは自動的に「X」となる）
  - 注2）ケース名（左端）が通し番号（1,2,…,7）となっていることを確認

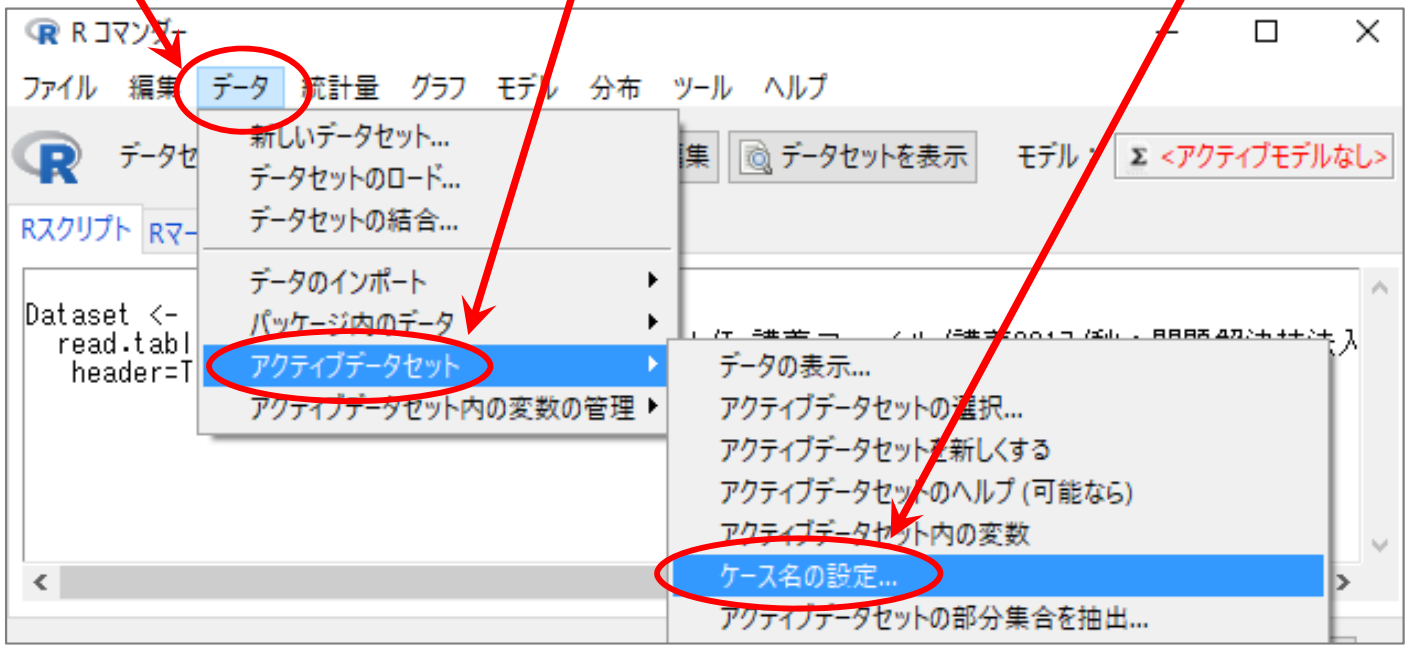



  - 注3）確認後は，必ず「Dataset」の「×」をクリックして「閉じる」こと

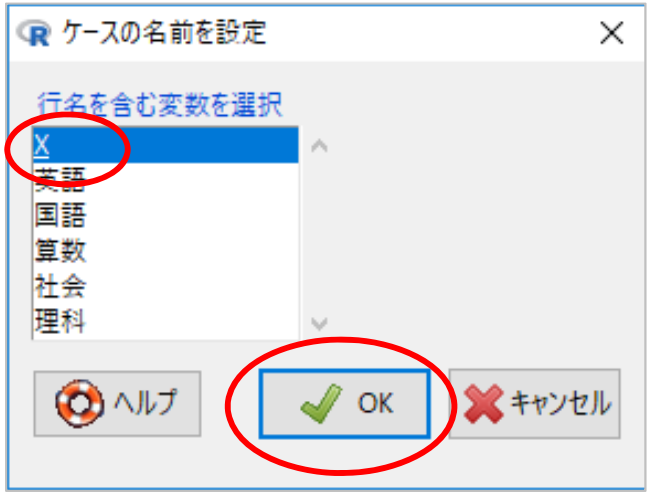
# 4. R commanderでクラスター分析

## ④ データにケース名を設定する

- 「データ」ー「アクティブデータセット」ー「ケース名の設定」選択



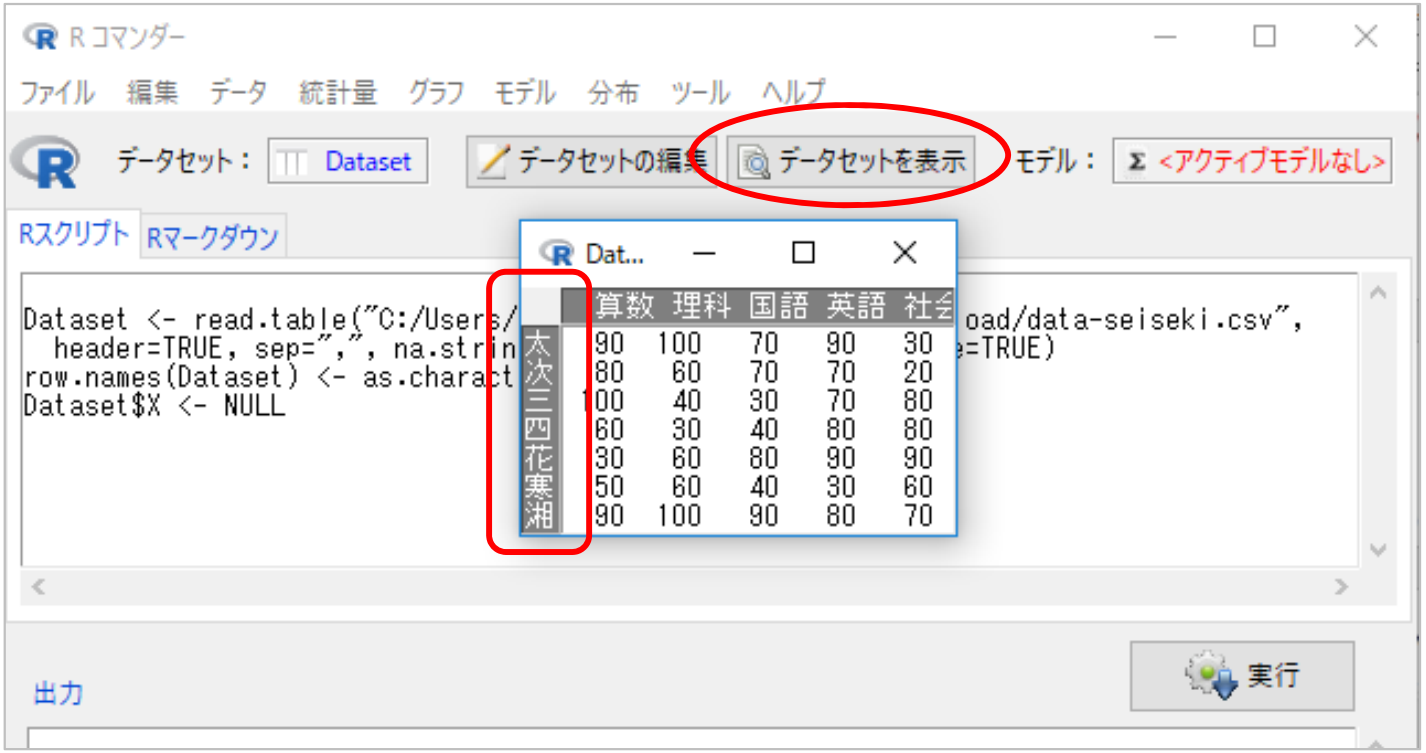

- [行名を含む変数を選択]で[X]を選び[OK]



# 4. R commanderでクラスター分析

## ④ データにケース名を設定する（設定確認）

- [データセットを表示]ボタンをクリックし内容を確認
  - 注1）指定した変数がケース名になっていることを確認

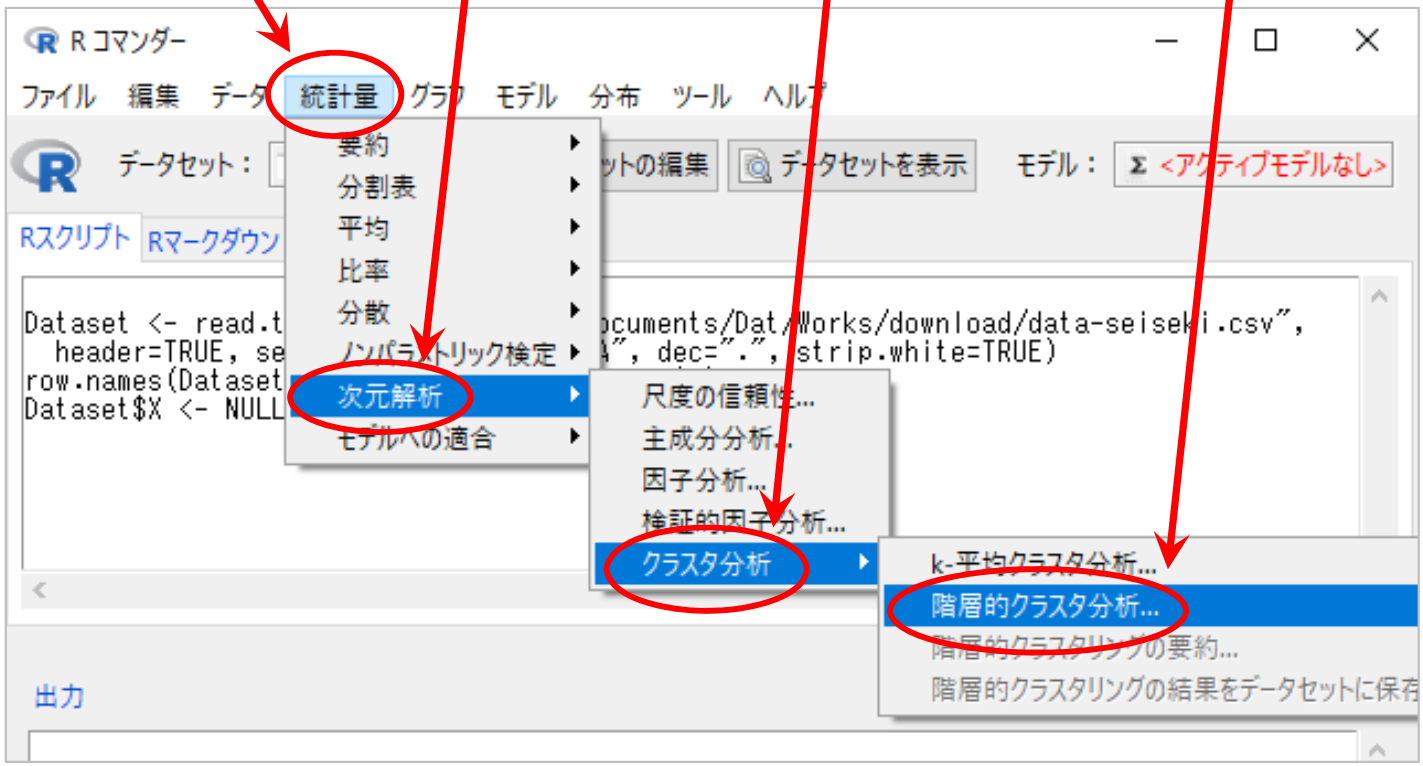



  - 注2）確認後は，必ず「Dataset」の「×」をクリックして「閉じる」こと

# 4. R commanderでクラスター分析

## ⑤ クラスター分析をする

- 「統計量」ー「次元解析」ー「クラスター分析」ー「階層的クラスター分析...」を選択

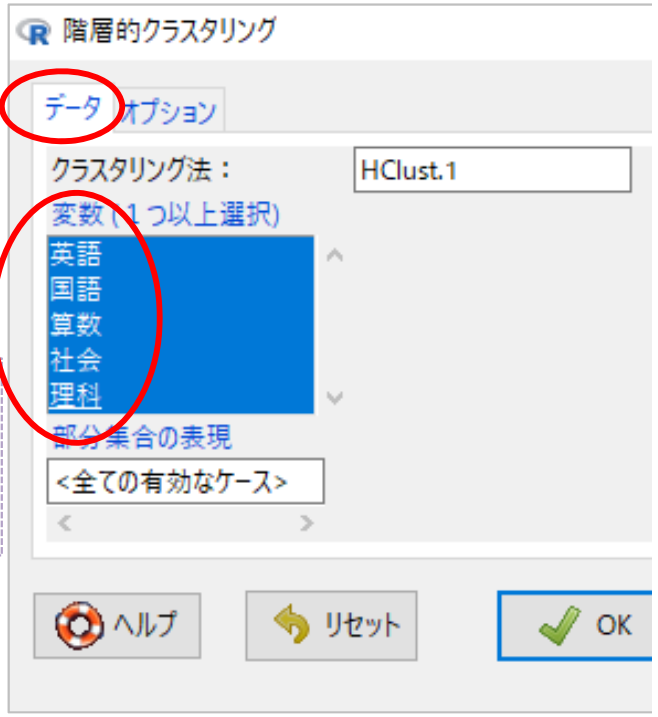


# 4. R commanderでクラスター分析

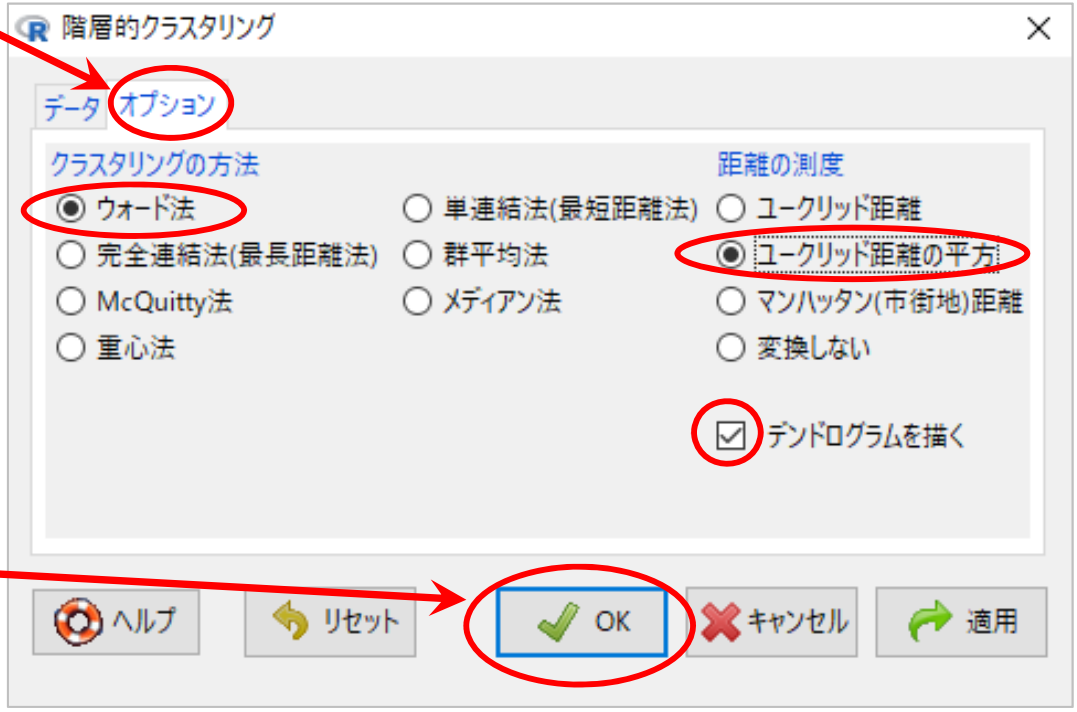
## ⑤ クラスター分析をする

- 『データ』タブで以下を設定
  - [変数（1つ以上選択）]で全科目を選択

注）複数の変数を選択する方法は以下のどちらかを実施
1. [Ctrl]キーを押しっぱなしで，変数を1つずつクリック
2. 1つめをクリック．[Shift]キーを押しながら最後をクリック



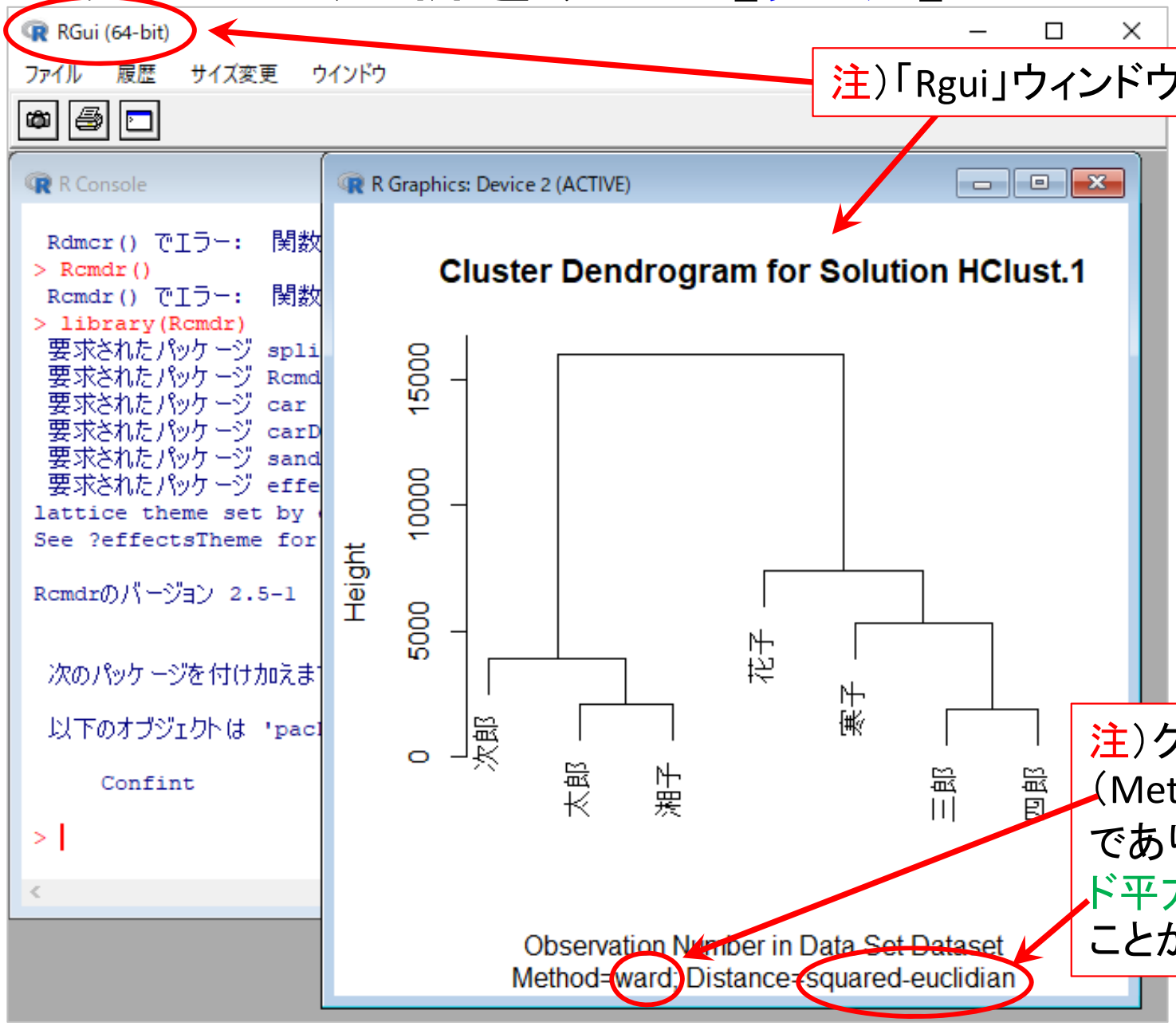

- 『オプション』タブで以下を設定
  - ウォード法
  - ユークリッド距離の平方
  - ☑デンドログラムを描く




- 全て設定後[OK]

# 4. R commanderでクラスター分析

## ⑤ クラスター分析をする 【完成】



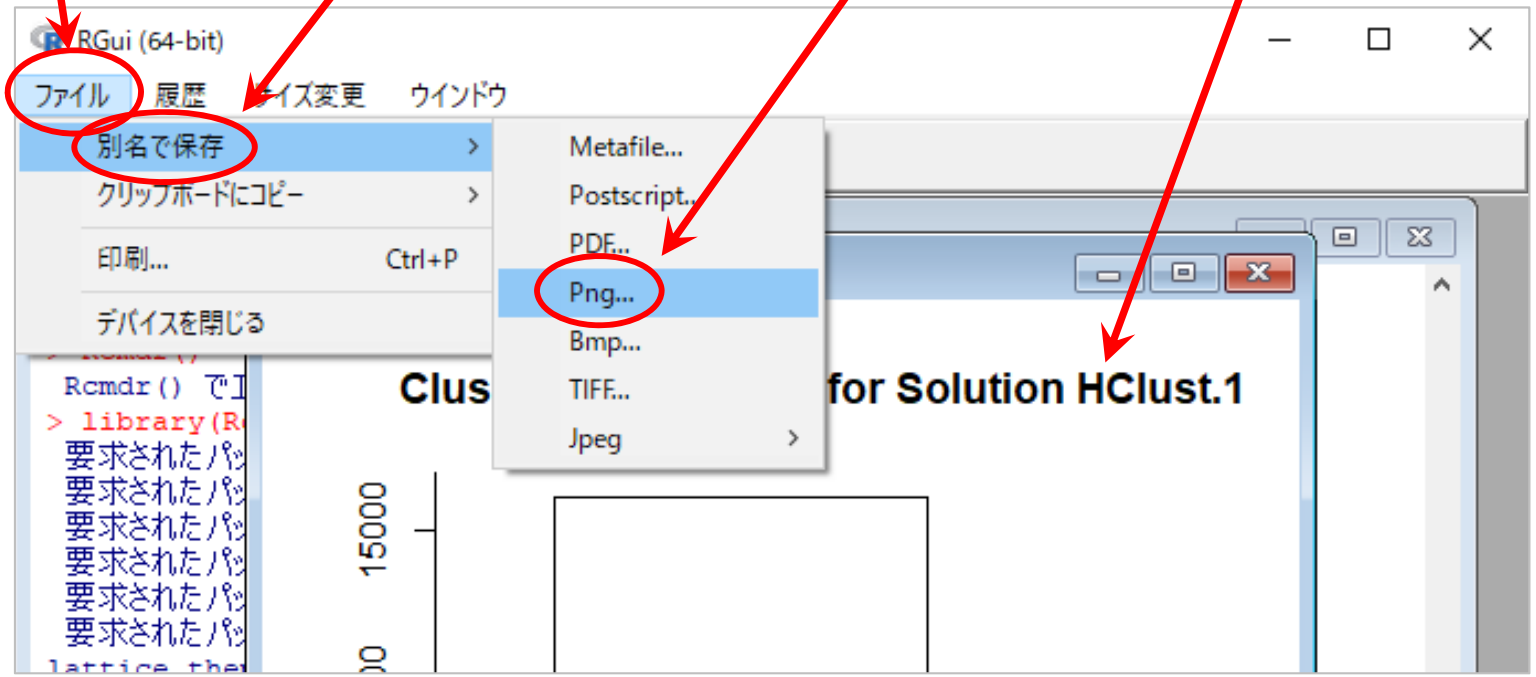

注）「Rgui」ウィンドウの中に表示される

注）クラスタリングの方法（Method）が「ウォード法」であり，距離に「ユークリッド平方距離」を使っていることがわかる

# 4. R commanderでクラスター分析

## ⑥ デンドログラム（樹形図）の保存

- デンドログラムのウィンドウを一番手前に表示した状態で，「ファイル」ー「別名で保存」ー「Png…」を選択




- 名前をつけて保存する

# 5. クラスター分析実施上の注意点

- クラスター分析の長所
  - 探索的手法なので，データ構造を事前に知らなくてよい
  - あらゆる種類のデータに適用可能：数値・カテゴリー
  - 適用が簡単

- クラスター分析の短所
  - どんな属性値を選んだらいいのか？
  - どの類似度（距離）測定法を選んだらいいのか？
  - どのクラスタ化更新法を選んだらいいのか？
  - データのスケーリング
  - 結果の解釈が困難な可能性がある

迷ったらとりあえず
「**ユークリッド平方距離**」
で

迷ったらとりあえず
「**ウォード法**」
で

# 参考文献


- 田中豊・脇本和昌『多変量統計解析法』現代数学社（1983）
- 河口至商『多変量解析入門Ⅱ』森北出版（1978,2005）
- 青木繁伸『Rによる統計解析』オーム社（2009）
- 荒木孝治『RとRコマンダーではじめる多変量解析』日科技連（2007）
- 金明哲『Rによるデータサイエンス』森北出版（2007）
- 新納浩幸『Rで学ぶクラスタ解析』オーム社（2007）


# もっと知りたい人へ

- 関連する経営学科の授業
  - 「**基礎統計**」（1/2セメ）
  - 「**基礎統計演習**」（3/4セメ）
  - 「**データ処理応用**」（2/3セメ）
  - 「**統計モデル分析**」（5セメ）
  - 「**ビッグデータ・AI演習**」（6セメ）

# 6. Rでクラスター分析

- Rを起動，csv ファイルをデータとして読込み
  - 「マイドキュメント（Y:）」の「R」フォルダに保存

data-seiseki.csv

| | 算数 | 理科 | 国語 | 英語 | 社会 |
|---|---|---|---|---|---|
| 太郎 | 90 | 100 | 70 | 90 | 30 |
| 次郎 | 80 | 60 | 70 | 70 | 20 |
| 三郎 | 100 | 40 | 30 | 70 | 80 |
| 四郎 | 60 | 30 | 40 | 80 | 80 |
| 花子 | 30 | 60 | 80 | 90 | 90 |
| 寒子 | 50 | 60 | 40 | 30 | 60 |
| 湘子 | 90 | 100 | 90 | 80 | 70 |

- csvファイルを読み込み，変数seiseki に代入

```
> seiseki <- read.csv("Y:/R/data-seiseki.csv", header=T, row.names=1)
```

※読み込むファイル名　※1行目にheaderあり　※各行1列目は名前

# 6. Rでクラスター分析

- 関数 dist() で距離を計算し，seiseki.dに代入

```
> seiseki.d <- dist(seiseki, “manhattan”)
```

※マンハッタン距離（”manhattan”）を用いて距離を計算している
他の距離を使いたいときは”manhattan”を以下に変更

“euclidean” =ユークリッド距離
“minkowski”, p=3 = p=3のミンコフスキー距離
“maximum” =$l_\infty$ノルム（える むげんだい のるむ）

- 階層クラスター分析をし，結果をseiseki.hcに代入

```
> seiseki.hc <- hclust(seiseki.d, “ward.D2”)
```

※ウォード法（”ward.D2”）を用いてクラスター分析を実施している
他の方法を使いたいときは、”ward.D2”を以下に変更

“single”=最短距離法, “complete”=最長距離法
“average”=群平均法, “centroid”=重心法, “median”=中央値法

# 6. Rでクラスター分析

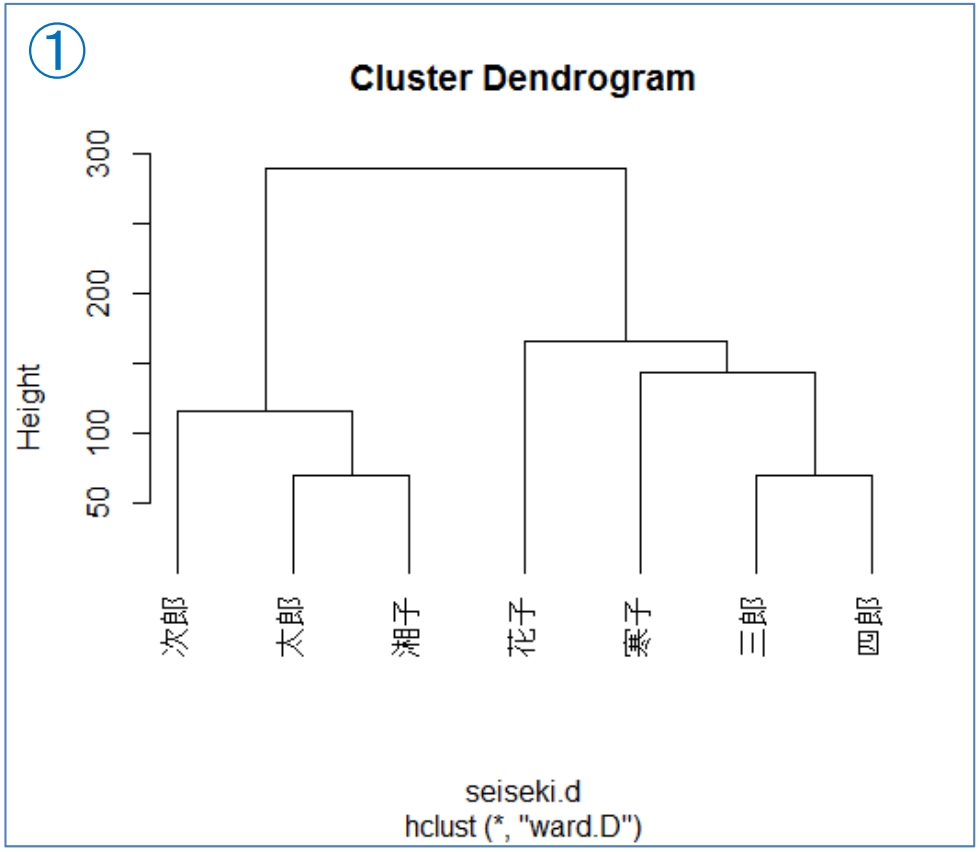
- 結果をデンドログラム（樹形図）で描画①

```
> plot(seiseki.hc, hang=-1)
```

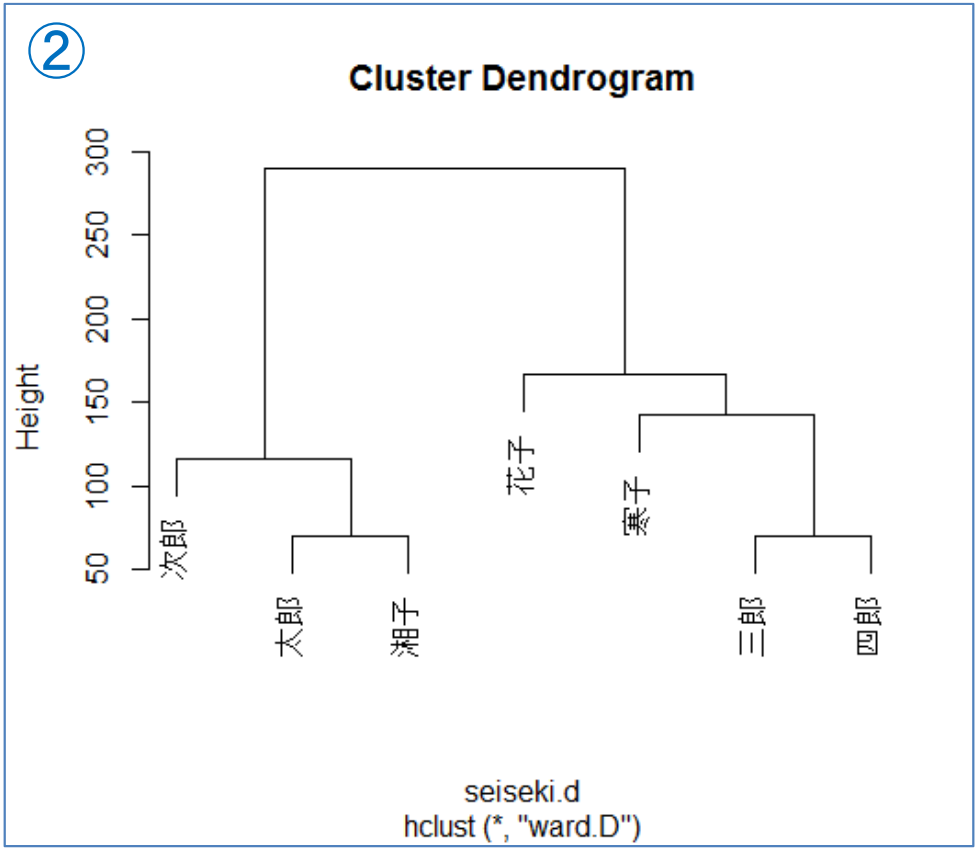
- 結果をデンドログラム（樹形図）で描画②

```
> plot(seiseki.hc)
```

①




②



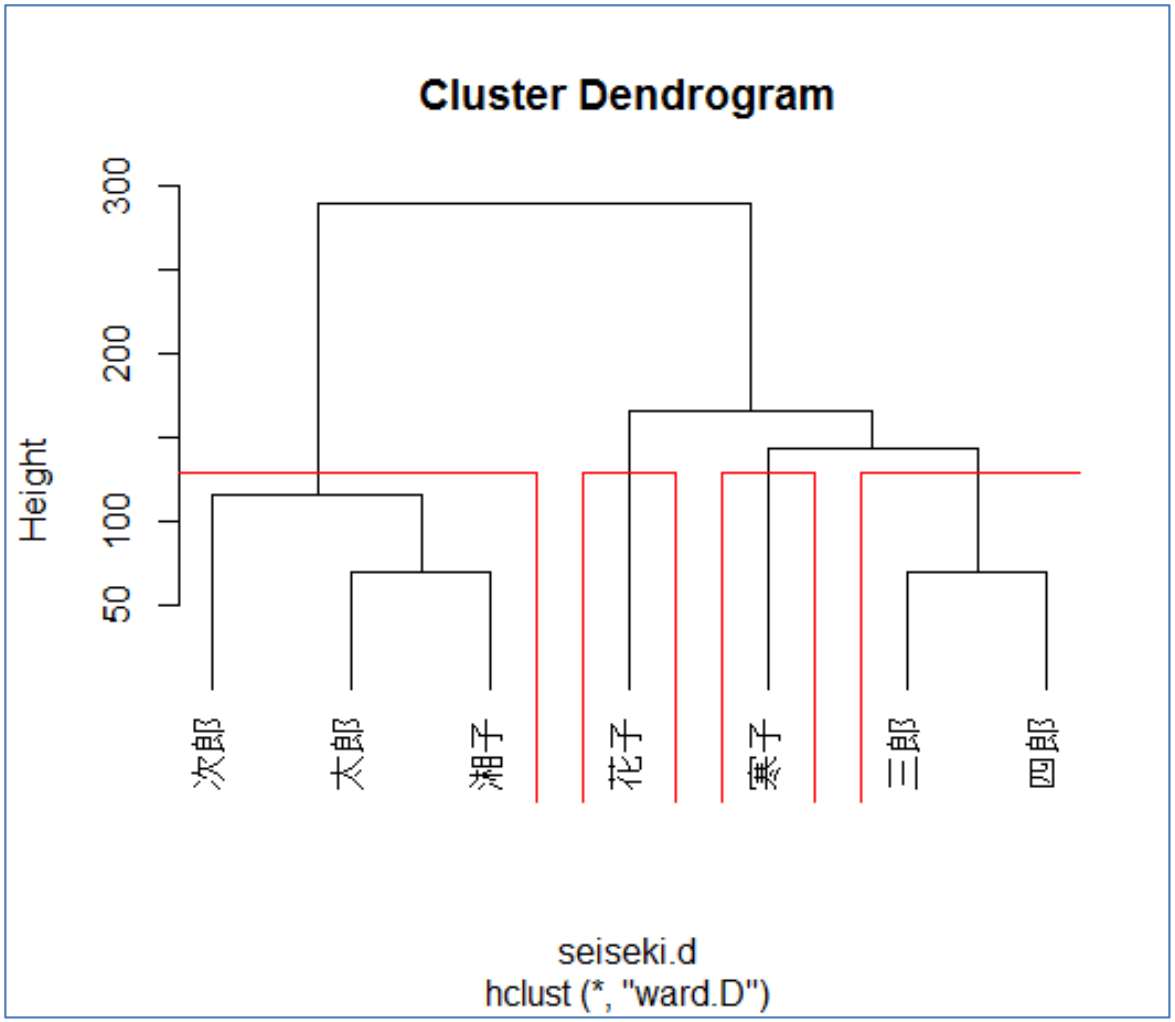
# 6. Rでクラスター分析

- デンドログラム（樹形図）を4つに分割

```
> plot(seiseki.hc, hang=-1)
> rect.hclust(seiseki.hc, k=4, border="red")
```

※分割数を4に指定

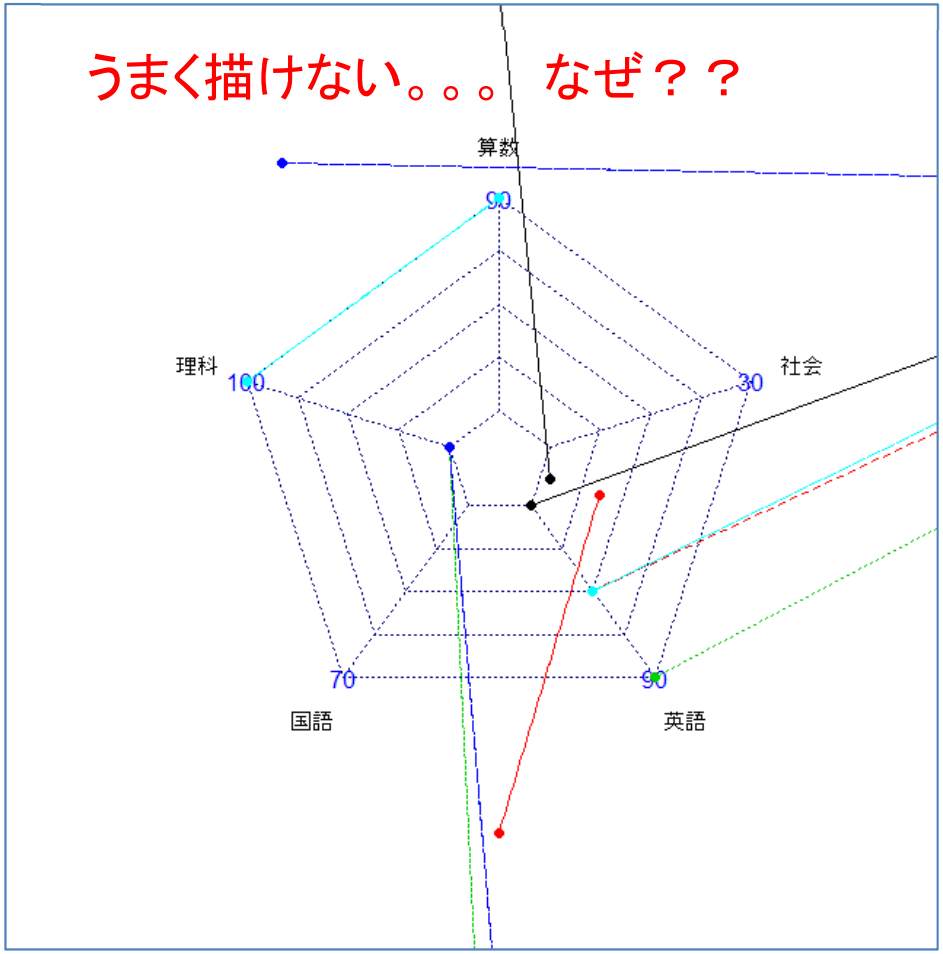
※分割線の色を赤に指定

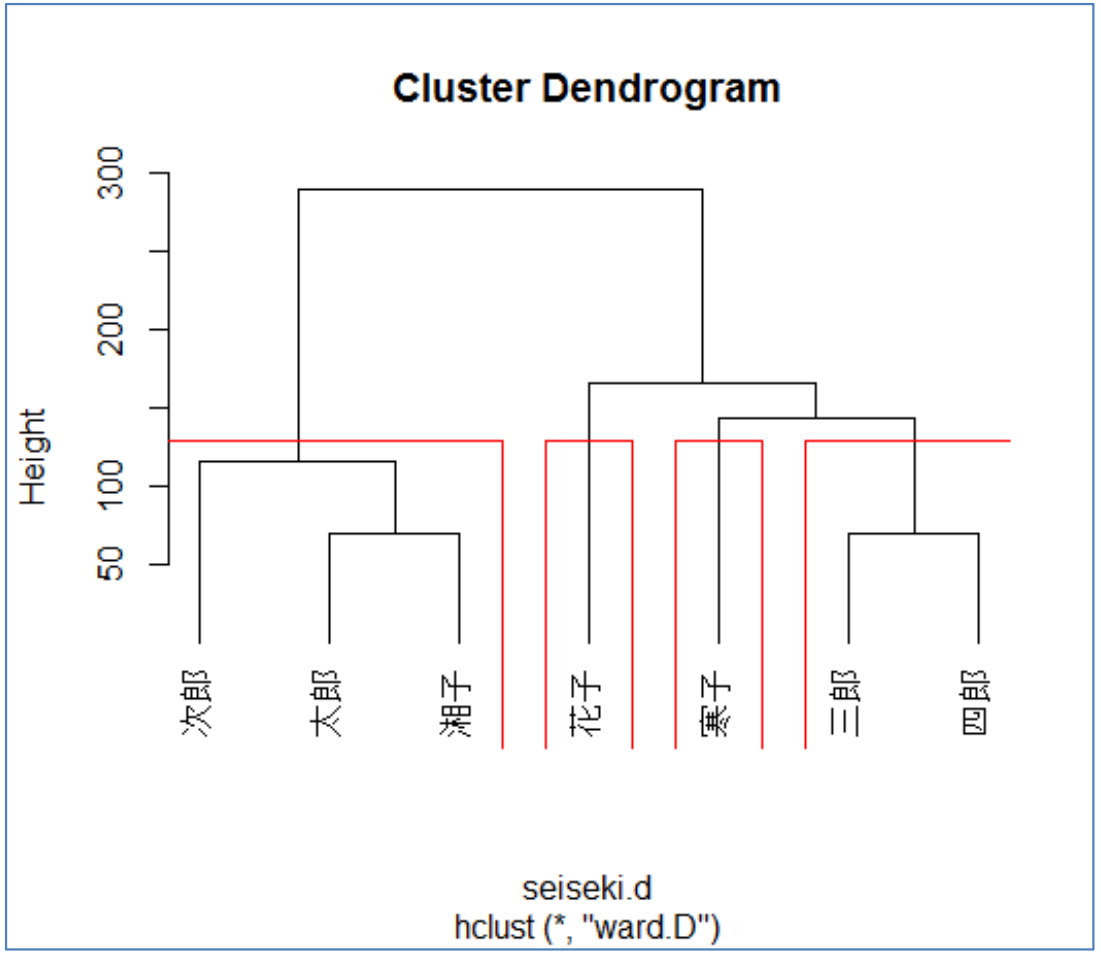


# 6. Rでクラスター分析

## Tips!　結果（樹形図）をレーダーチャートと比較

```
> install.packages("fmsb")
> library(fmsb)
> radarchart(seiseki, axistype=2, オプション指定勉強せよ)
```



# 6. Rでクラスター分析

## 【練習】距離とクラスター化の方法，分割数を以下の設定に従ってクラスター分析をし，樹形図を描き，比較せよ

| | 距離 | クラスター化の方法 | 分割数 |
|---|---|---|---|
| ① | ユークリッド距離（euclidean） | 最短距離法（single） | 4 |
| ② | ユークリッド距離（euclidean） | 最長距離法（complete） | 4 |
| ③ | ユークリッド距離（euclidean） | 群平均法（average） | 4 |
| ④ | ユークリッド距離（euclidean） | 重心法（centroid） | 4 |
| ⑤ | ユークリッド距離（euclidean） | 中央値法（median） | 4 |
| ⑥ | ユークリッド距離（euclidean） | ウォード法（ward.D2） | 4 |

| | 距離 | クラスター化の方法 | 分割数 |
|---|---|---|---|
| ① | マンハッタン距離（manhattan） | 最短距離法（single） | 4 |
| ② | マンハッタン距離（manhattan） | 最長距離法（complete） | 4 |
| ③ | マンハッタン距離（manhattan） | 群平均法（average） | 4 |
| ④ | マンハッタン距離（manhattan） | 重心法（centroid） | 4 |
| ⑤ | マンハッタン距離（manhattan） | 中央値法（median） | 4 |
| ⑥ | マンハッタン距離（manhattan） | ウォード法（ward.D2） | 4 |

※たくさん計算するので変数は整理して使う
例えば，距離は
seiseki.man <- dist(...)
seiseki.euc <- dist(...)
などとし，ユークリッド距離とマンハッタン距離を計算した結果を，わかり易い名前の別変数で区別し，クラスター化も[距離_方法]付加で区別等
seiseki.e_si <-hclust(...)
seiseki.e_cp <-hclust(...)
seiseki.e_av <-hclust(...)
seiseki.e_ce <-hclust(...)
seiseki.e_m <-hclust(...)
seiseki.m_si <-hclust(...)
seiseki.m_cp <-hclust(...)

# 6. Rでクラスター分析

## Tips!　画面を分割して，複数の図を比較する

```
> par(mfrow=c(2,3))
```

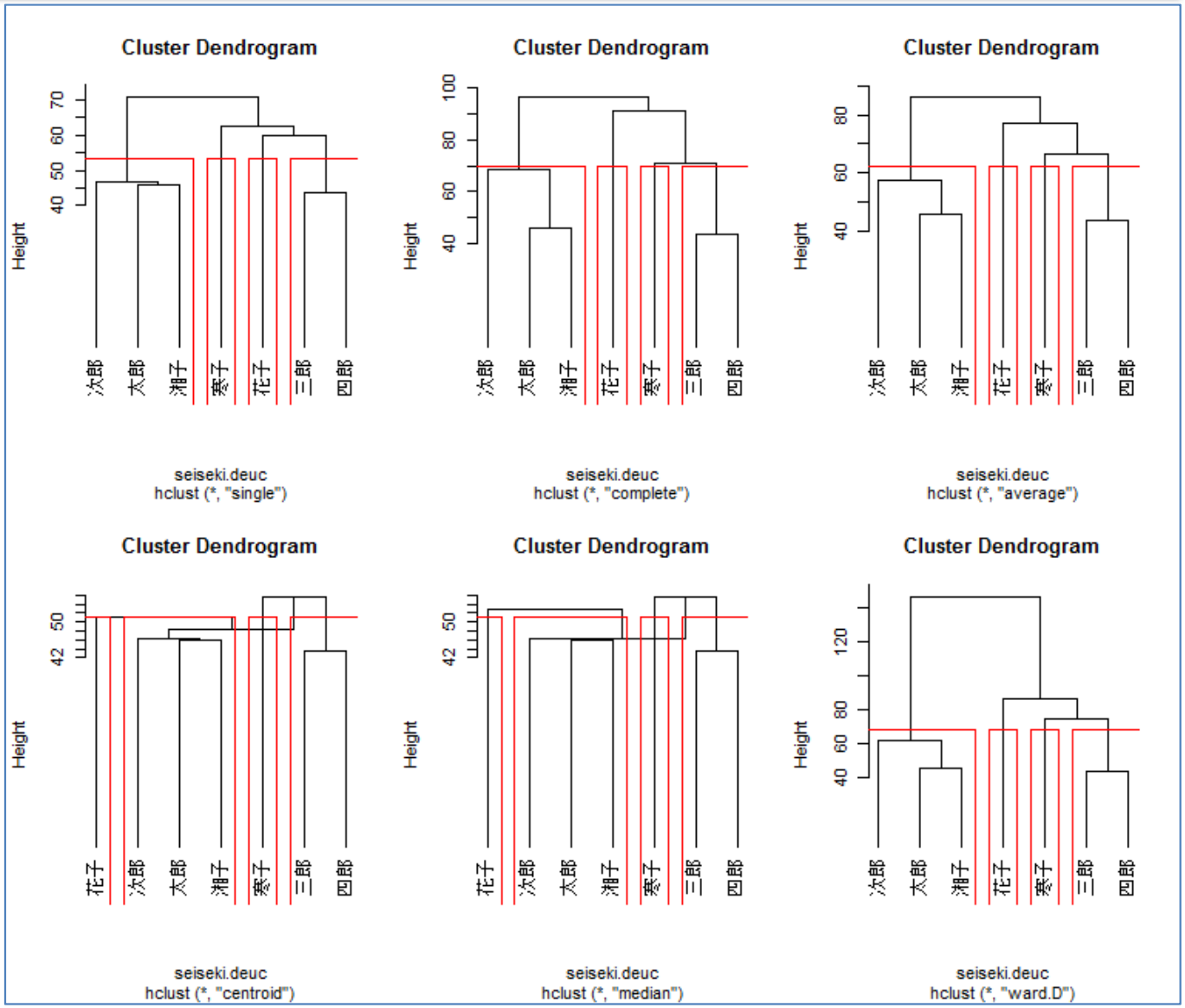
←　※1画面を2×3に分割

※c(x,y) の x, y に分割したい行数(x)と列数(y)を指定する

※この命令の後，plotなどで図を描画すると，左上から順に描画されていく

※6個描かれた後，7個目を描くと，画面がクリアされてまた左上から順に描画される

※別の分割に変えたい場合は，変えたい設定でもう一度実行すれば良い（何度でも変更可能）



# 6. Rでクラスター分析

## Tips! 画面分割，複数図描画（前ページの場合の実行例）

```
> seiseki.euc <- dist(seiseki, "euclidean")

> seiseki.e_si <- hclust(seiseki.euc, "single")
> seiseki.e_cp <- hclust(seiseki.euc, "complete")
> seiseki.e_av <- hclust(seiseki.euc, "average")
> seiseki.e_ce <- hclust(seiseki.euc, "centroid")
> seiseki.e_m  <- hclust(seiseki.euc,"median")
> seiseki.e_wa <- hclust(seiseki.euc,"ward.D2")

> par(mfrow=c(2,3))

> plot(seiseki.e_si,hang=-1)
> rect.hclust(seiseki.e_si,k=4,border="red")
> plot(seiseki.e_cp,hang=-1)
> rect.hclust(seiseki.e_cp,k=4,border="red")
> plot(seiseki.e_av,hang=-1)
> rect.hclust(seiseki.e_av,k=4,border="red")
> plot(seiseki.e_ce,hang=-1)
> rect.hclust(seiseki.e_ce,k=4,border="red")
> plot(seiseki.e_m,hang=-1)
> rect.hclust(seiseki.e_m,k=4,border="red")
> plot(seiseki.e_wa,hang=-1)
> rect.hclust(seiseki.e_wa,k=4,border="red")
```

ユークリッド距離を計算

クラスター分析実施
上から順に，6つの方法でそれぞれ計算し結果を保存

6つの結果を描画したいので画面を2x3の6分割

6つの結果を順に描画
それぞれ2行で1つの画面を作っており，
plot(...) が樹形図描画
rect.hclust(...) が分割線描画をしている

# 6. Rでクラスター分析

## Tips!　たくさんの命令を打つのは大変だし間違えちゃう！

## 一度にまとめて命令したい！

① まとめて実行したい命令（右）を1つのファイルに書く．制作には「TeraPad」や「メモ帳」「秀丸」などのテキストエディタを使う

② ファイルの種類を「全てのファイル」にし，「ファイル名.R」で保存．このとき，ファイル名は半角アルファベットが良い（例：ファイル名「euc_clust.R」とし「Y:/R/」フォルダに保存）

③ R（R Studio）で以下を実行

```
> source("Y:/R/euc_clust.R")
```

※ソースコード「euc_clust.R」内に間違いがなければ全て順に実行される．間違いがある場合は，その場所でエラーが出て止まる

【演習】manhattan 距離で同様のファイル「man_clust.R」をつくり実行しよう

```
seiseki.euc <- dist(seiseki, "euclidean")

seiseki.e_si <- hclust(seiseki.euc, "single")
seiseki.e_cp <- hclust(seiseki.euc, "complete")
seiseki.e_av <- hclust(seiseki.euc, "average")
seiseki.e_ce <- hclust(seiseki.euc, "centroid")
seiseki.e_m  <- hclust(seiseki.euc,"median")
seiseki.e_wa <- hclust(seiseki.euc,"ward.D2")

par(mfrow=c(2,3))

plot(seiseki.e_si,hang=-1)
rect.hclust(seiseki.e_si,k=4,border="red")
plot(seiseki.e_cp,hang=-1)
rect.hclust(seiseki.e_cp,k=4,border="red")
plot(seiseki.e_av,hang=-1)
rect.hclust(seiseki.e_av,k=4,border="red")
plot(seiseki.e_ce,hang=-1)
rect.hclust(seiseki.e_ce,k=4,border="red")
plot(seiseki.e_m,hang=-1)
rect.hclust(seiseki.e_m,k=4,border="red")
plot(seiseki.e_wa,hang=-1)
rect.hclust(seiseki.e_wa,k=4,border="red")
```