DEPs.xla

Data **E**xploring **P**rocedure**s**

（データ　エクスプローリング　プロシージャーズ）

http://dataexploring.com/

～2016/3/31　Ver.03

0. はじめに

0.0 概要
0.1 DEPsについて
0.2 プログラムおよびサンプルデータ

0.0 概要

- DEPs（Data Exploring Procedures：データエクスプローリングプロシージャーズ）は，DATAEXPLORINGが作成・管理している記述的多変量解析のためのExcelプログラム（xlsmファイル）です．

0.1　DEPsについて

- Microsoft Excel の マクロ有効ワークシート (.xlsm)として配布しています.
- 通常1年間(毎年4月1日(ないし配布時点)～翌年3月末日まで)使用が可能です. 1年ごとにプログラム(xlsmファイル)自体を購入しなおしてください.
- 配布対象は, 企業の管理職者, 大学教員, 各種グループのリーダーなどです(以下,「代表者」). 各代表者が, 自らの判断で行う再配布に関しては, 費用は不要です. 必要に応じ, ファイルをコピーしてください. ただし, 一切のサポートは, 代表者に対してしか行いません. 再配布先からの問い合わせには, 直接／代表者を通じた転送, いずれとも一切応じません.

- 全体を通じての注意
  - 欠測値(無回答)には対応していません.
    - 必要な場合はケースワイズ削除(ひとつでも無回答がある回答者の回答は, その人の回答1人分全部を削除する→すべてに回答している人のデータしか分析に使わない).
  - データ範囲の選択はマウスで行ってください.
    - キーボード操作だと, 想定外の動きをすることがあります.
    - 都度, データシートを選択しなおす必要があります.

0.2 プログラムおよびサンプルデータ

■ 0.2.1 おもなプログラム

1. 重回帰分析
2. 主成分分析
3. 因子分析（探索的，直交回転）
4. 対応分析
5. クラスター分析（階層的）
6. 主座標分析

■ 0.2.2 サンプルデータ

◆ iris

- Edgar Anderson's Iris Data（The R Datasets Package より）※G列（7列目）以右を除く.
- 「フィッシャーのアイリスデータ」として知られるもの．花弁の長さなど4つの測定値と花の種類（計5変数）が，150ケースについて記してある.
- G列（7列）目は新規に追加した「Sepal.Length」（花弁の長さ）に基づく変数.
- H列（8列）～K列（11列）は，B列（2列）～E列（5列）の数値から小数点以下を切り捨てた値.
- L列（12列）は，F列（6列）の「Species」を，「1」～「3」の数値にしたもの.

◆ CT_iris：irisのデータをもとに作ったクロス表（分割表）

- 3つの「花の種類」ごとに，4階級に分けた「花弁の長さ」に基づくカテゴリ変数 を集計したクロス集計表（度数表）

1. 重回帰分析

1.0 概要
1.1 手法について
1.2 プログラムの実行
1.3 出力について

1.0 概要

■ ひとつの従属変数（y）を，複数の独立変数（X）の線形結合で説明します．

- Xが 01型データになれば数量化 I 類．
- Xを直交表を用いて（交互作用を除いた最小限の組み合わせにして）割り付けて，数量化 I 類のような計算をすれば，コンジョイント分析．
- yが0～1の範囲を越えないケースでは，ロジスティック回帰を考えます．

1.1　手法について

- $\boldsymbol{y = b_0 + Xb + \varepsilon}$
  - $\boldsymbol{y}$: 従属変数
  - $\boldsymbol{b_0}$: 切片
  - $\boldsymbol{Xb}$: 独立変数($\boldsymbol{X}$) × 回帰係数($\boldsymbol{b}$)　※$\boldsymbol{X}$は行列
  - $\boldsymbol{\varepsilon}$: 誤差

  ・・・$\boldsymbol{y}$を$\boldsymbol{X}$で当てる．そのための$\boldsymbol{b}$を決める．
- 回帰係数($\boldsymbol{b}$)は，[独立変数+従属変数]の分散共分散行列の逆行列から求まる．
- 実際は$\boldsymbol{Xb}$の部分は，行列とベクトル．

$$\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{p1} \\ x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix}$$

図1.1　$\boldsymbol{Xb}$ の成分

1.2 プログラムの実行

- データは「iris」.
  - ケース数は150.
  - 従属変数(y)には「Species.n 」を指定. 独立変数(X)には「Sepal.Length」「Sepal.Width」「Petal.Length」「Petal.Width」の4つを指定.
- 重回帰分析の指定内容
  - シート「menu」の「btn01」(重回帰分析)をクリックして実行.
  - 表示されるダイアログにて以下指定.
    - 従属変数データ範囲:
      - シート「iris」の1行目H(8)列～151行目H(8)列. (iris!$H$1:$H$151)※ラベル行を含む.
    - 説明変数データ範囲:
      - シート「iris」の1行目B(2)列～151行目E(5)列. (iris!$B$1:$E$151)※ラベル行を含む.

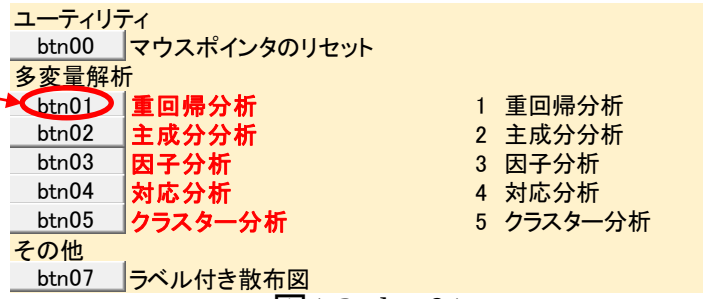


図1.2 btn01

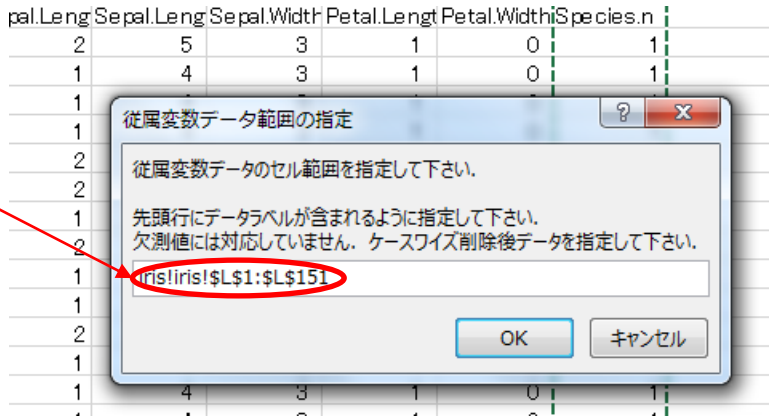


図1.3 従属変数データ範囲の指定

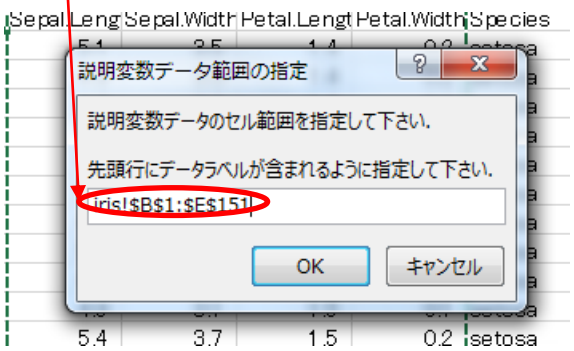


図1.4 説明変数データ範囲の指定

1.3　出力について

- 基本統計量
  - ケース数/平均/標準偏差
- 相関係数行列（下側三角）/分散共分散行列（対角及び上側三角）
- 相関係数の有意性検定結果p-値（下側三角）/データ件数（対角及び上側三角）
- 説明率
  - R・・・重相関係数：yと$\hat{y}$の相関係数
  - R^2・・・決定係数：重相関係数の二乗．yの変動のうち$\hat{y}$の変動で説明できている割合．
  - 調整済R^2・・・自由度調整済み決定係数：R^2はn（ケース数）が小さいと大きくなる．R^2を，ケース数に依存しないように調整したもの．自由度調整済み決定係数（調整済R^2）は，決定係数（R^2）より小さくなる．
  - AIC/extractAIC・・・小さいほうが良い．値に絶対的な意味は無い．
    ※extractAICについては右記参照→https://www.facebook.com/notes/580150388708835
  - 標準誤差・・・分散分析表の「残差」分散の平方根

■ 分散分析

◆ 回帰の分散と残差の分散のF比を見ます.

◆ Fの大きさが, どれだけ稀なことかは, p−値を見ます.

→回帰式でどれだけyを説明できているかの評価. 式全体の評価です.

■ 回帰係数

◆ 標準化解では切片はゼロです.

◆ トレランス・・・小さい(たとえば0.1以下)時は多重共線性に注意.

■ この場合で言うと. .

◆ 回帰式全体は有意(分散分析).

◆ これら4つの変数の組み合わせで9割以上, 従属変数(あやめの種類)を当てられる(R^2).

◆ 個々の説明変数を見ると, 花弁(Petal)の長さ(Length)／幅(Width)が有意. 萼(がく)片(Sepal)の回帰係数は0でないとは言い切れない(とくに萼片の幅(Sepal.Width)).

◆ ただし, 花弁の長さ／幅は, トレランスが小さい. 実際, Petal.LengthとPetal.Widthの相関は0.963と高いので注意が必要.

2. 主成分分析

2.0 概要
2.1 手法について
2.2 プログラムの実行
2.3 出力について

2.0 概要

- もとの複数の変数を，より少ない数の変数（合成変数）で効率よくあらわします．

2.1　手法について

■ たくさんの顕在変数を少ない合成変数に

※顕在変数：実際に直接測定できる変数.

◆ たとえば・・・身長，体重，胸囲，～などの身体測定値がある．これらを，身体的特徴をしめす，より少数の変数であらわしたい．→ size（大きさ），shape（形（太っているか痩せているか））

◆ 次元縮約のイメージ
・・・分散の大きいところに新しい軸をとおす．

※図1.2参照.

◆ 計算のなかみは・・・

- 変数間の相関行列ないし共分散行列を固有値分解
- 固有ベクトルがそのまま主成分得点.

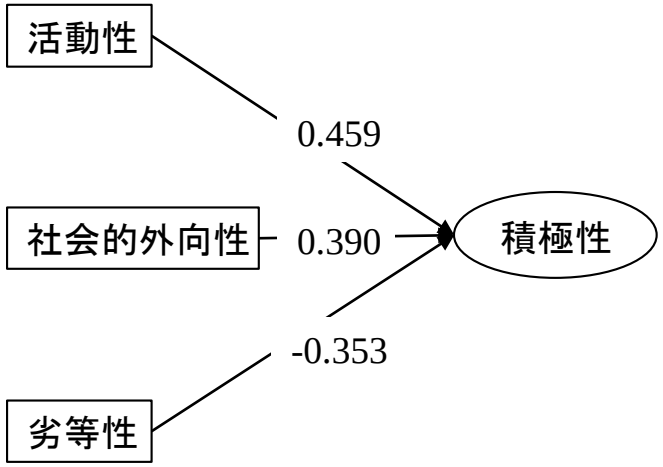


図2.1　主成分へのパス

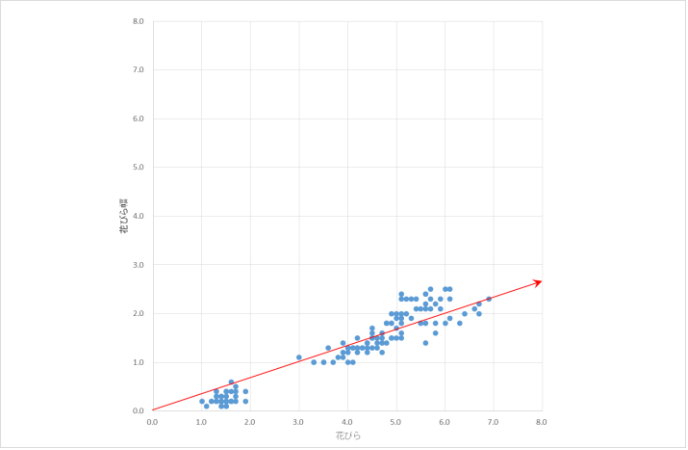


図2.2　次元縮約のイメージ

2.2 プログラムの実行

- データは「iris」.
  - ケース数は150.
  - 変数は「Sepal.Length」～「Petal.Width」の4つを指定.
- 主成分分析の指定内容
  - シート「menu」の「btn02」(主成分分析)をクリックして実行.
  - 表示されるダイアログにて以下指定.
    - 固有値産出行列:「0」(相関係数行列).
    - 主成分得点:「いいえ」.
    - データ範囲:
      シート「iris」の1行目B(2)列～151行目E(5)列.
      (iris!$B$1:$E$151)※ラベル行を含む.

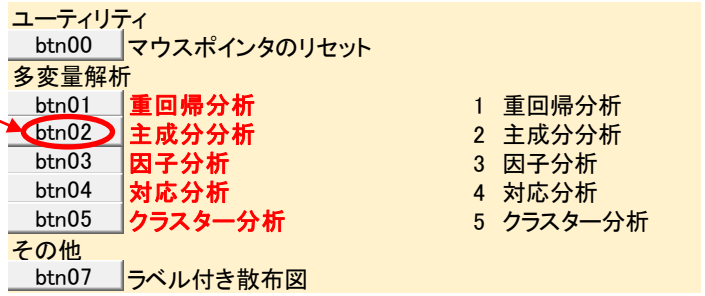


図2.3 btn02

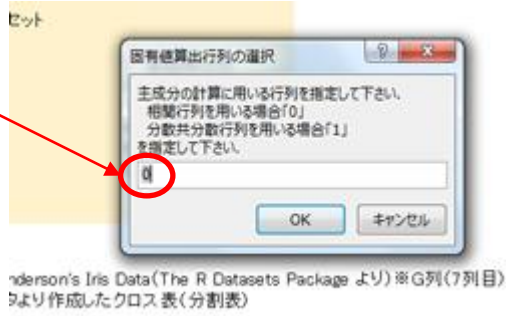


図2.4 固有値産出行列の選択

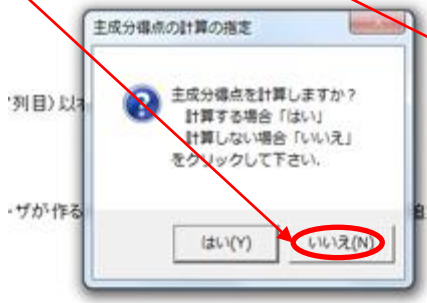


図2.5 主成分得点の指定

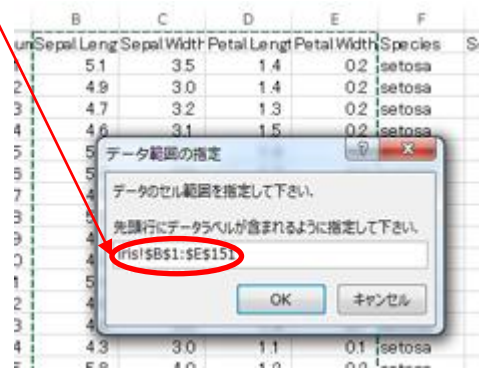


図2.6 データ範囲の指定

■ 散布図の指定内容（主成分分析の実行後に行う）

- ◆ シート「menu」の「btn07」（ラベル付き散布図）をクリックして実行.
  - ● 主成分分析の実行結果，出力されたシート「PCA」の「主成分係数」をデータとして使用.
  - ● 第1次元×第2次元の散布図を描く場合，
    - ‣ x軸に　PCA!$B$40:$B$44　※タイトル行を含める
    - ‣ y軸に　PCA!$C$40:$C$44　※タイトル行を含める
    - ‣ ラベルに　PCA!$A$41:$A$44　※先頭行は含めない

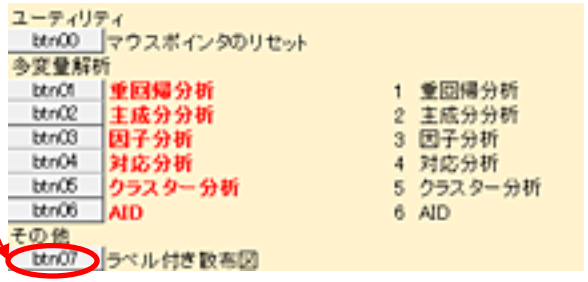


図2.7　btn07

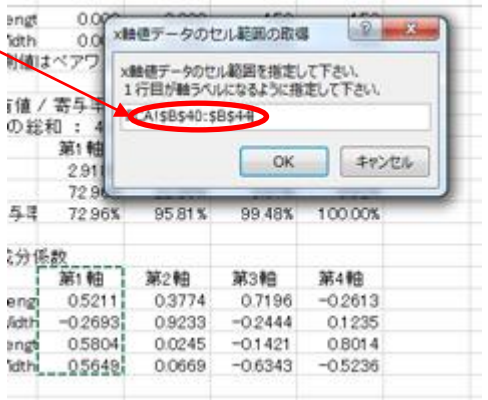


図2.8　x軸値データのセル範囲

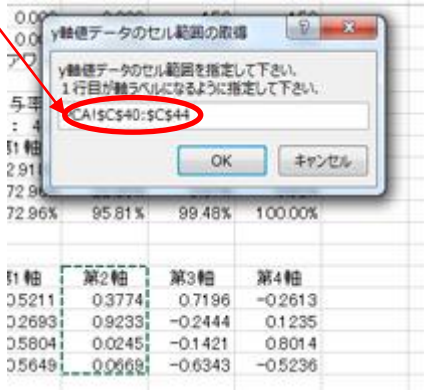


図2.9　y軸値データのセル範囲

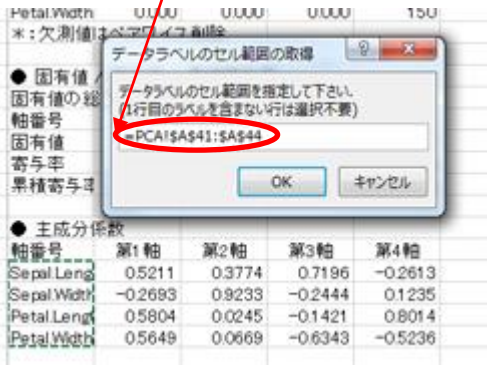


図2.10　データラベルのセル範囲

2.3 出力について

■ 結果

◆ 軸の解釈・・・元のデータをどのくらい表せているか，まず累積寄与率を見る．

- ここでは，第2次元まで（累積寄与95%）の解を見てみることにする．
  - 第1次元：「Sepal.Width」（＋）それ以外 ― Sepal.Width（－）
  - 第2次元：「Sepal/Petal」（＋）Sepal ― Petal（－）

表2.1 固有値/寄与率/累積寄与率

● 固有値 / 寄与率 / 累積寄与率				
固有値の総和 ： 4.000				
軸番号	第1軸	第2軸	第3軸	第4軸
固有値	2.918	0.914	0.147	0.021
寄与率	73.0%	22.9%	3.7%	0.5%
累積寄与率	73.0%	95.8%	99.5%	100.0%

表2.2 主成分係数

● 主成分係数				
軸番号	第1軸	第2軸	第3軸	第4軸
Sepal.Length	0.521	0.377	0.720	−0.261
Sepal.Width	−0.269	0.923	−0.244	0.124
Petal.Length	0.580	0.024	−0.142	0.801
Petal.Width	0.565	0.067	−0.634	−0.524

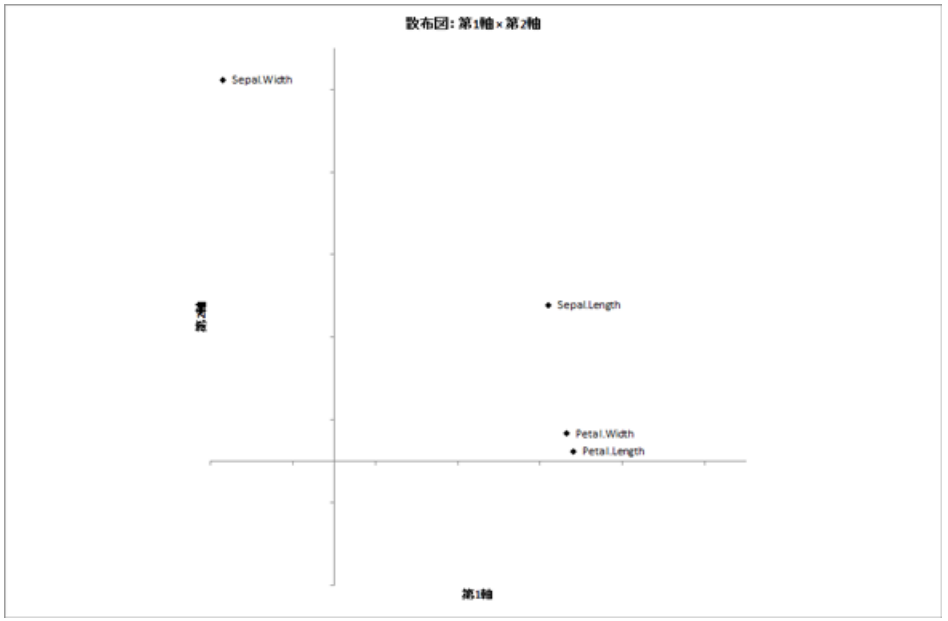


図2.11 第1軸（横軸）×第2軸（縦軸）の散布図

3. 因子分析

3.0 概要
3.1 手法について
3.2 プログラムの実行
3.3 出力について

3.0 概要

■（複数の）変数の背後にある，抽象的な概念（潜在変数）を探し出します．

3.1　手法について

■ 顕在変数の背後にある潜在変数を見つけ出す

◆ 測定できない変数（潜在変数）があって，これが，測定できる変数（顕在変数）に影響を与えている．※図3.1
eg.「文系能力」（潜在変数）が英語の得点や国語の得点（顕在変数）に影響を与えている．

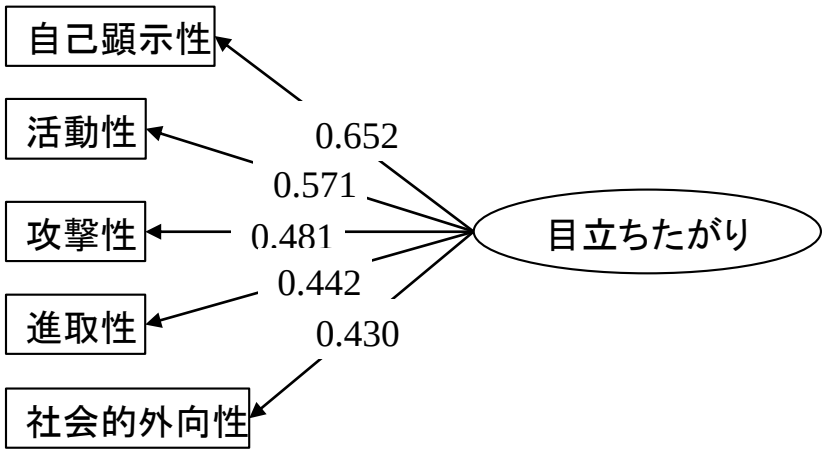


図3.1　潜在変数からのパス

◆ 計算のなかみは・・・

- まず，共通性（潜在因子から影響を受けている部分．潜在因子と関係ない，個々の変数独自の部分は独自性）を推定する．
- 相関係数行列の対角部分（全部「1」）を共通性で置き換えて，固有値分解．
- 求まった固有ベクトルの値に，求まった固有値の平方根（特異値に相当）を乗じて，因子負荷量を求める．
- 求まった因子負荷量を回転する．cf. バリマックス回転→バリアンスがマックスになるように．
- 回転した因子負荷行列を，2次元ソートする（各変数とも最も負荷量の絶対値が大きい次元について，負荷量の絶対値の大きい順に並び替える）．

◆ 潜在因子に名前を付ける．

3.2 プログラムの実行

- データは「iris」.
  - ケース数は150.
  - 変数は「Sepal.Length」～「Petal.Width」の4.
- 因子分析の指定内容
  - シート「menu」の「btn03」（因子分析）をクリックして実行.
  - 「因子得点を計算しますか？」に「いいえ」
    ※主成分分析の場合同様，因子得点で人をクラスタリングも計算はできるが，あるいはSEMを選択することも検討してみる.
  - データ範囲はシート「iris」の1行目B(2)列～151行目E(5)列．(iris!$B$1:$E$151)※ラベル行を含む.

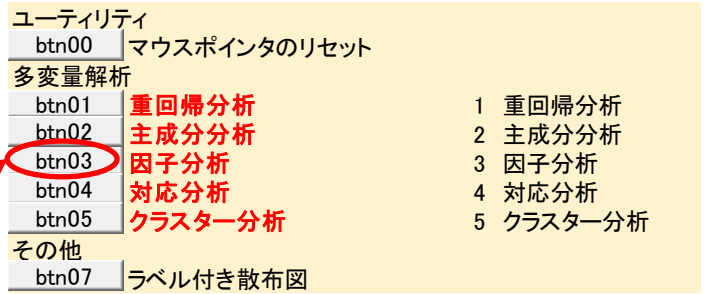


図3.2 btn03

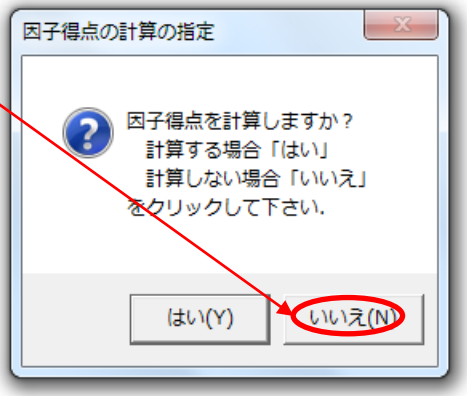


図3.3 因子得点の計算の指定

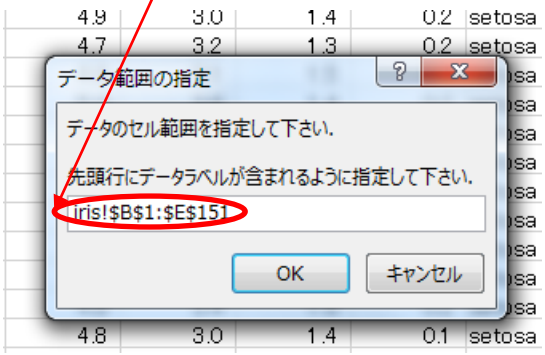


図3.4 データ範囲の指定

- 計算する因子数は，固有値が1を越える＋肘の基準で「2」を指定．
  ※スクリープロット参照（図3.6）．固有値の表から作成．→因子（共通性推定後）では無い．
- 因子抽出は主因子法，回転はバリマックス．
  ※最近は，因子抽出には最尤法を使うことも多い．ある程度傾向がはっきりしている場合は，どちらでも解は同様となる．
- データ範囲の指定はそのまま「OK」．
  ※因子負荷量の範囲の指定→2次元ソート，が行われる．
  ※固有値の総和は，共通性を推定すると独自性の分が減る．

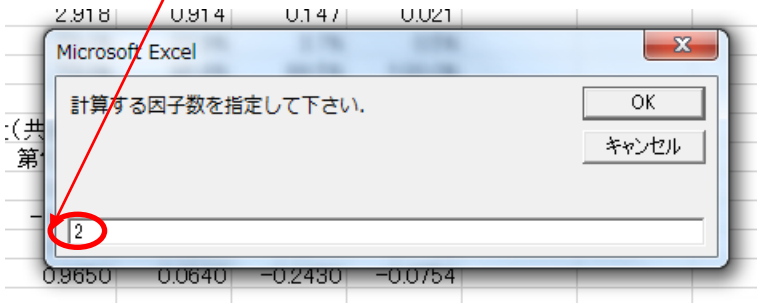


図3.5　クロス表（ウラウズ×NPI）

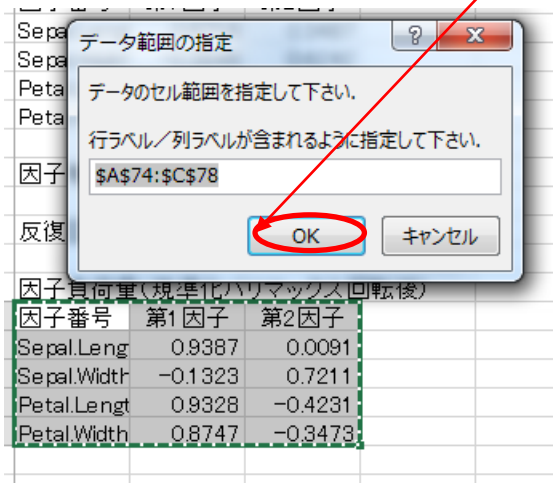


図3.7　（負荷行列の）データ範囲

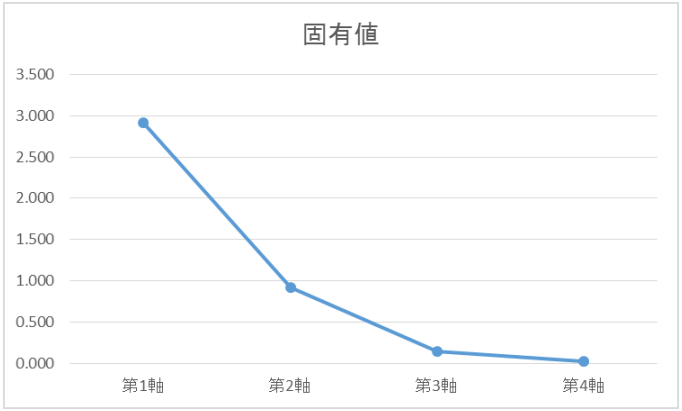


図3.6　スクリープロット

3.3 出力について

■ 結果

◆ 第2因子まで，元のデータのおよそ96%（共通性を推定する前なので，目安）を使って計算．

- **因子の解釈**（※因子負荷量：回転後（表3.1）を見る）
  - 因子寄与は通常，あまり気にする必要はありません（回転するとある程度なれるため）．
  - 因子負荷量の絶対値が大きい項目に着目（因子負荷量の絶対値が小さい場合あまり気にしにしない．）
  - 共通性が大きい項目に着目（独自性が大きい場合あまり気にしない…とえば，「Sepal.width」）

‣因子に名前を付ける．たとえば・・・

- 第1因子：「Sepal.width」以外
- 第2因子：「Sepal.width」

※そもそも潜在因子を想定できるケースでもなく．ちょっと例が良くないです．．

表3.1 因子負荷量

因子負荷量：回転後 — 2次元ソート				
因子番号	第1因子	第2因子	共通性	独自性
Sepal.Length	0.939	0.009	0.881	0.119
Petal.Length	0.933	−0.423	1.049	−0.049
Petal.Width	0.875	−0.347	0.886	0.114
Sepal.Width	−0.132	0.721	0.537	0.463
寄与	2.534	0.820	3.354	0.646
累積寄与	2.534	3.354		
寄与率	76%	24%		

4. 対応分析

4.0 概要
4.1 手法について
4.2 プログラムの実行
4.3 出力について

4.0 概要

■ データ行列の，（基本的には）行要素同士／列要素同士の距離関係を図示します．

※数量化Ⅲ類（パタン分類の数量化）と，数学的には同等だが，標準的には同じデータを分析したら同じ数字が出てくるのではない．同じように使えるが，必ずしも同じ目的で作られたものでは無い．（最終的には同等の手法として使えるが，各手法の生み出される筋道は林知己夫（パタン分類の数量化）とベンゼクリ（コレスポンデンスアナリシス）とで異なる．）

※行要素と列要素の同時布置は解釈に注意が必要．

※「数量化〇類」という呼称は，飽戸弘氏によるもの．林知己夫氏本人も使っていたが，林知己夫氏本人の命名ではない．

4.1　手法について

■ 分割表（クロス表，度数表）を図示する．※数表を眺めるより，絵のほうが直感的にわかる

数式は右記参照．→http://insightxinside.com/correspondence_analysis_correspondence-analysis/

クロス表（表4.1）の各セルの値は，それぞれ行和，列和が違う．→各セルの値を，行和，列和を使って標準化する．（ポアソン分布する値の標準化）

◆ 標準化した行列を特異値分解する．

表4.1　分割表（クロス表，度数表）

Sepal.Length.c				
表側：Spec	1	2	3	4
setosa	20	30	0	0
versicolor	1	29	20	0
virginica	1	8	29	12

■ 対応分析では，しばしば，行と列の同時布置を見る．

◆ ただし，行と列の相関（特異値）は常に1を下回っているので，行と列は，本来次元を共有していない．そのため，行と列の同時布置を解釈することの是非については議論がある（経験的には有用でも理論的には正しくない）．eg.西里静彦（2010）行動科学のためのデータ解析．培風館．

■ その他．．※そもそも固有値分解は（特異値分解も），第1次元から順次もっともよく分散を説明するように次元を決めていくので，最適な次元数が分かっている場合は，現代的MDS推奨．（https://www.facebook.com/notes/551449451578929）

※非対称の行列は，自分自身の転置行列を掛ければ，対象行列になる．

4.2 プログラムの実行

- 「btn04」(対応分析)を実行.
- データ範囲には, シート「CT_iris」の「CT_iris!$A$2:$E$5」を指定.
- 「OK」をクリックして実行.
- シート「CA」に計算結果が出力される.

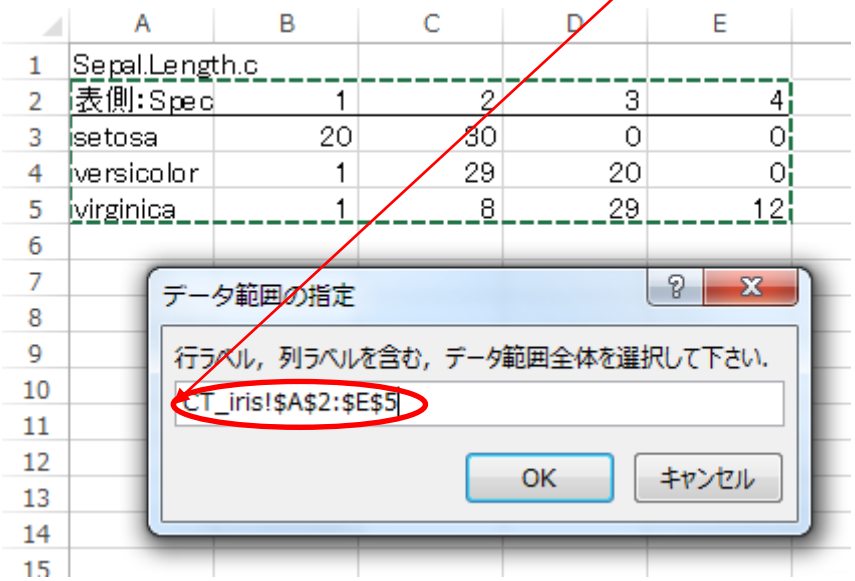


図4.2 btn04

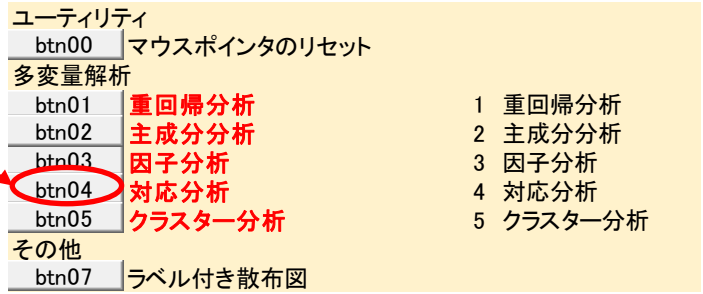


図4.1 btn03

4.3 出力について

■ 出力内容

◆ 固有値 / 寄与率 / 累積寄与率 / 相関係数

- 算出した軸の数だけ出ます.
- 「固有値の総和」は, CA(対応分析)のときだけ意味があります. MCA(多重対応分析)の場合は無視してください. なお, 度数表にしか対応していません(相対度数(%)表の際も無視してください). (読み込んだ分割表の(周辺度数を除く)すべてのセルの値を足しあげたものが総度数(N)に等しい場合のみ有効)

◆ スコア

- 座標値. この値を使って散布図を作成します.

◆ 絶対寄与

- 各軸ごと, すべての行要素／列要素を足すと(縦に足すと), 100%になります.
- ある軸については, どの要素が重要なのかを見ます.

◆ 平方相関(cos^2θ)

- 各要素ごと, すべての軸を足すと(横に足すと), 1になります.
- ある要素については, どの軸が重要なのか見ます.

4.3 出力について

■ 出力の解釈

- ◆ 累積寄与は第1次元だけで81%.
- ◆ 平方相関を見ると，1軸は，Sepal（がく片）のLength（長さ）が「1」か「3」の場合，2軸は「4」あるいは「1」が引っ張っている.
- ◆ 絶対寄与を見ると，Sepal（がく片）のLength（長さ）が「1」のときは，1軸も2軸も同等，「2」では2軸，「3」では1軸，「4」では2軸に対する寄与が大きい.

→累積寄与を見ると，基本的に1軸でおおよその概況をあらわす．他方，2軸への絶対寄与が大きい要素もある．なので，2軸までの結果（散布図）を見る.

※相対寄与，平方相関・・・https://www.facebook.com/notes/293428184047725

■ 散布図（次頁）

- ◆ 行と列の同時布置の場合，基本は角度を見る．度数が他より小さいセル（たとえばSepal.Length.c=4）は，原点から大きく離れる場合があるので注意．平方相関の小さい軸では無視．行要素だけ，列要素だけの場合は，単純に距離を解釈すればよい。

・・・ https://www.facebook.com/notes/404801642910378

◆ 第1次元 × 第2次元

- 曲線状に，列については「1」→「2」→「3」→「4」，行については「setosa」（セトサ）→「versicolor」（バーシカラー）→「virginica」（バージニカ），が並んでいる．

※軸に無理に名前を付けることはオススメしません．

※すべて解釈しようとすることはオススメしません．（絶対寄与，平方相関が小さい場合は無理に解釈しないで無視する）

※ちなみに：**散布図はタテ・ヨコのスケールを揃える！**

※この例では「馬蹄形問題」が起こっている可能性も（本来1次元で足りる情報に2次元が使われている）．

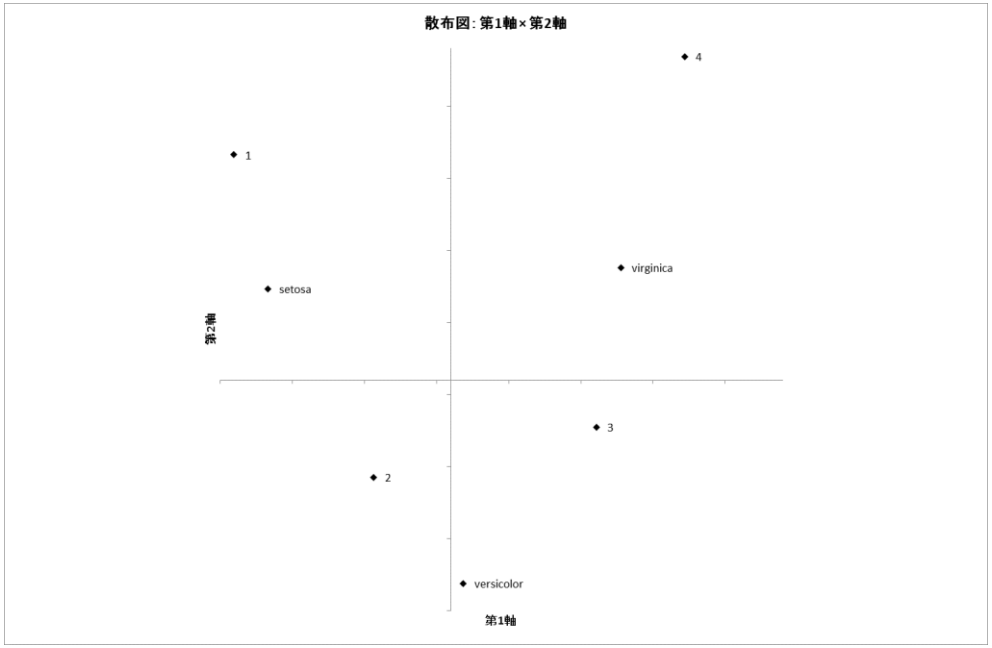


図4.3　対応分析の結果の散布図（Species × Sepal.Length.c）

5. クラスター分析

5.0 概要
5.1 手法について
5.2 プログラムの実行
5.3 出力について

5.0 概要

- クラスター分析
  - 多くの対象を，少ないグループ（クラスター）にまとめる.

5.1　手法について

■ クラスター分析（階層的方法）

◆ たくさんの変数を順次結合していく

- ひとつひとつ（一人ひとり）から始まって，順次，「近い」ものをまとめていく．
- 「近い」をどう定義するかでいろいろな方法があるが，良く使うのは ward法．
  - ward法では，ある一つを，ほかに統合する（同じクラスタに含める）際，もっとも群内平方和の増加が小さいクラスタに統合する．
  - いろいろな方法は，いくつかのパラメタの値を変えることで，ひととおりの式で表現できる（組合せ的方法：conbinational method）（Lance and Williams，1967）．
  - ward法は，鎖効果（一つのクラスタに，順次，ひとつの対象が追加されていく…図2.1のA）が起こりにくい．
- この「近い」ものを統合していく過程をあらわしたものが樹形図（デンドログラム）．

例）同じデータを，それぞれ

A）標準化データ／最近隣法／平方距離

B）非標準化データ／ward法／平方距離

で計算した際の樹形図．Aでは鎖効果が起こっている．※図2.1

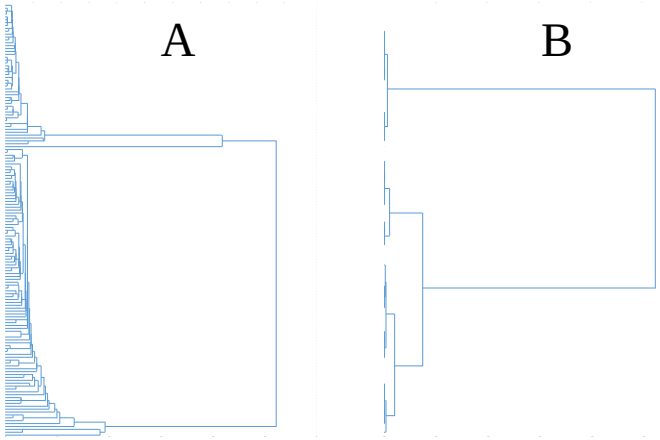


図5.1　デンドログラムの例（A：最近隣法／B：ward法）

5.2 プログラムの実行

- データは「iris」.
  - ケース数は150.
  - 変数は「Sepal.Length」～「Petal.Width」の4.
- クラスター分析の指定内容
  - シート「menu」の「btn05」（クラスター分析）をクリックして実行.
  - データ範囲は，シート「iris」の「iris!$B$2:$E$151」（※ラベルは含まない）.
  - ここでは，データは正規化しない（これはデータによる）
  - 手法は ward法（ウォード法：最小分散法）.
    - 群内の平方和の増分がもっとも小さいクラスターに併合する.

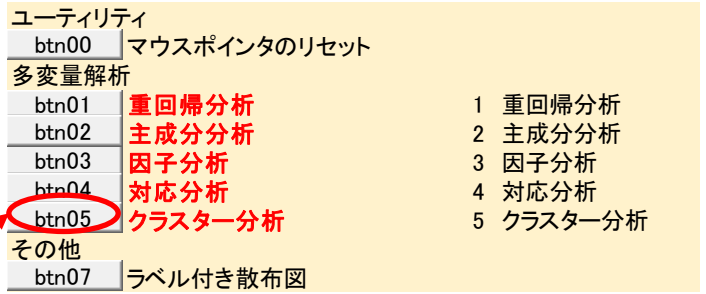


図5.2 btn05

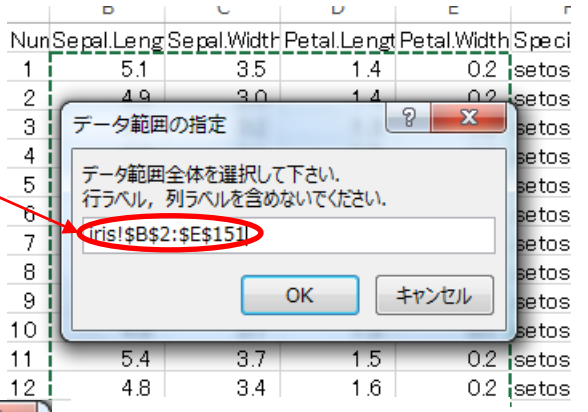


図5.3 データ範囲の指定

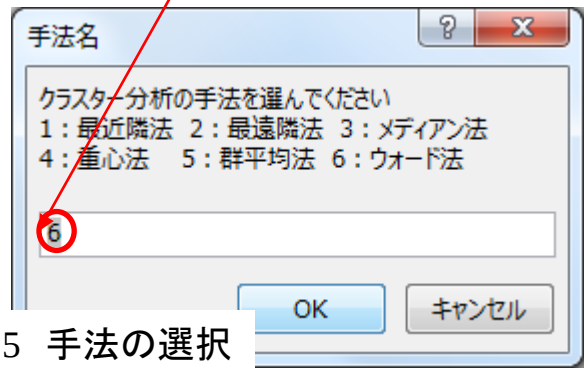


図5.5 手法の選択

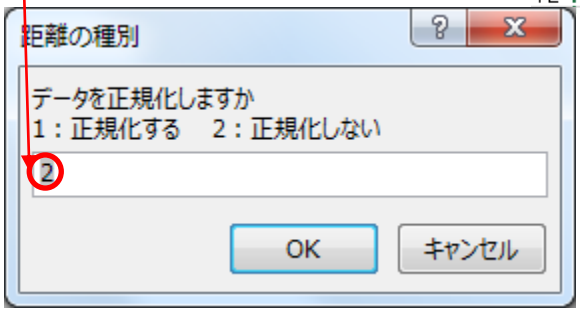


図5.4 正規化の有無

5.3　出力について

■ 結果（デンドログラム（図5.6）参照）

◆ 2あるいは3クラスタ？（赤線は3クラスタとなるように引いたもの）

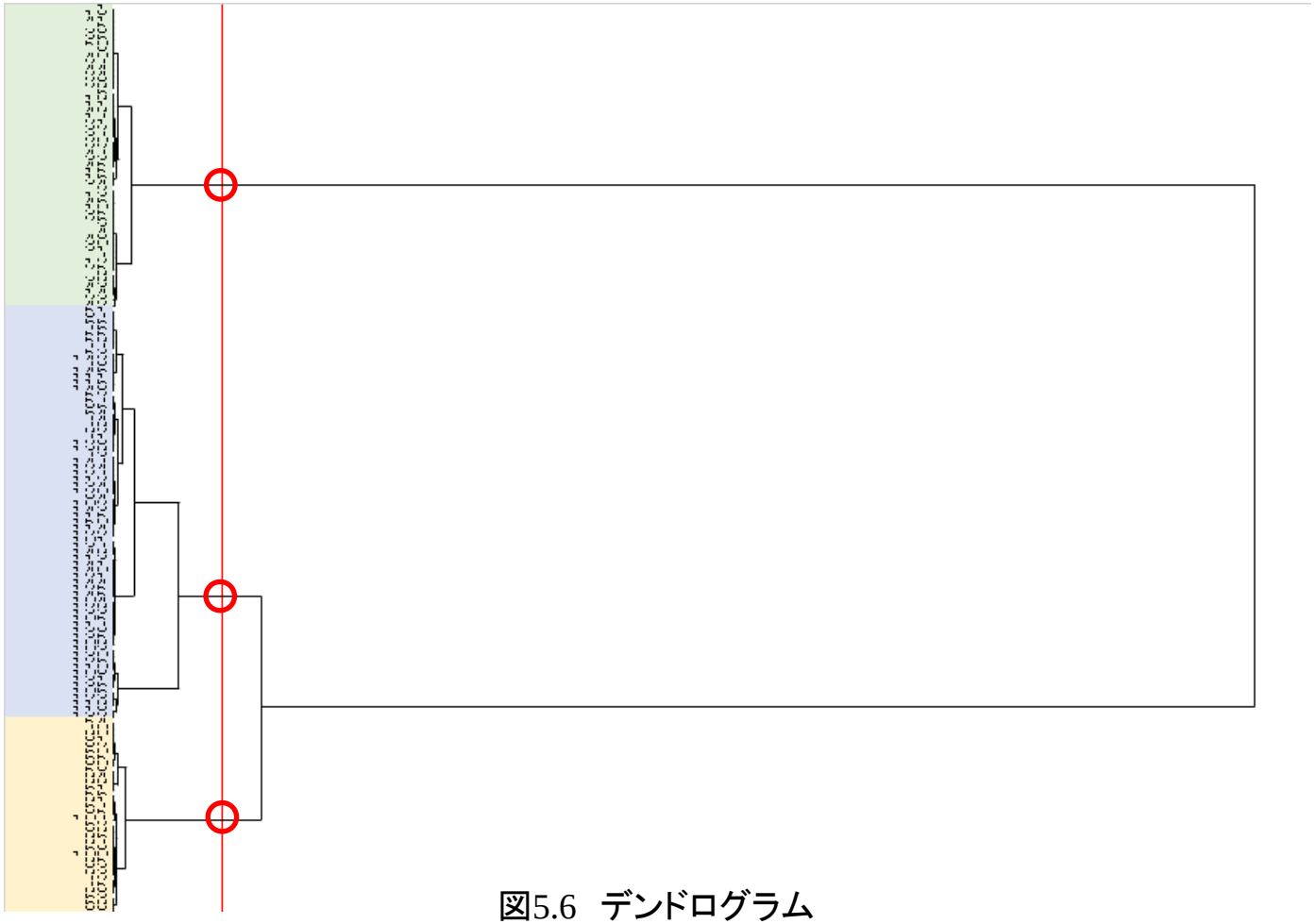


図5.6　デンドログラム

■ どの人がどのクラスタに割り振られたか．．？
（※ここは全くソフトウェアに依存します）

- ◆ デンドログラム部分の行の高さを標準の高さに広げます（ここでは13.5に）．
- ◆ 各クラスタに割り振られたサンプルNO．を確認し易くします（ここでは，セルに色を付けています）．
- ◆ 列を挿入してスラスター分析の結果に基づくクラスタ番号を割り振ります．
- ◆ サンプルNO．でソートすると，もとのデータに合併する（くっつける）ことができます．

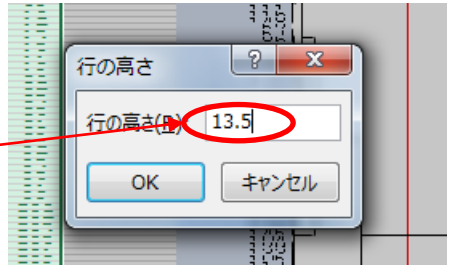


図5.7　デンドログラム

Species.n	クラスター番号	サンプルNO
1	1	1
1	1	2
1	1	3
1	1	4
1	1	5
1	1	6
1	1	7
1	1	8
1	1	9
1	1	10
1	1	11
⋮	⋮	⋮
3	2	139
3	2	140
3	2	141
3	2	142
3	2	143
3	2	144
3	2	145
3	2	146
3	2	147
3	2	148
3	2	149
3	2	150

図5.9　デンドログラム

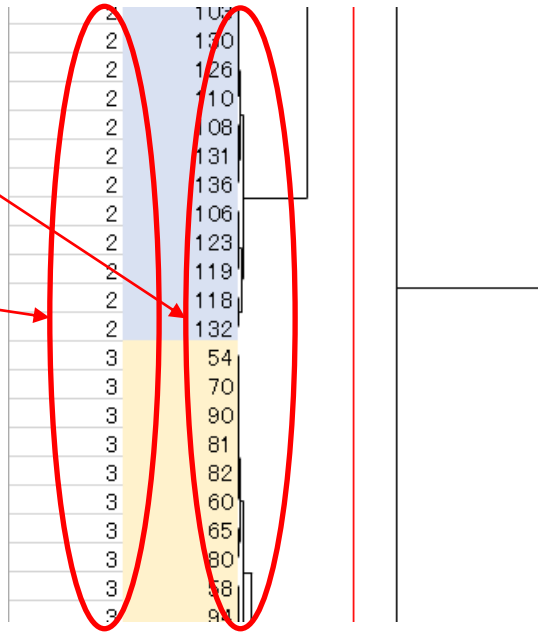


図5.8　デンドログラム

■ クラスタリングの結果

- ◆ ここで使用した4つの変数を使って，ウォード法でクラスタリングした結果に基づいて，全150ケースを3つのクラスタに分けた．このクラスタと，もとの種類をクロス集計すると，結果は表5.1のとおり．がく片の長さ，幅，花弁の長さ，幅の4つの変数では，全体として150ケースのうち130ケース（87%）を正しくクラスタリングすることができた．

表5.1　もとの種類／データに基づくクラスタ

クラスタ				
表側: Species	全体	1	2	3
全体	150	50	68	32
	(100.0)	(33.3)	(45.3)	(21.3)
setosa	50	50	−	−
	(100.0)	(100.0)	(−)	(−)
versicolor	50	−	19	31
	(100.0)	(−)	(38.0)	(62.0)
virginica	50	−	49	1
	(100.0)	(−)	(98.0)	(2.0)

5.4 発展

■ クラスタクロス

- ◆ 各クラスタがどういう性質か示すため，各クラスタを表側，クラスタリングに使った各変数を表頭にとったクロス集計表（平均値表）を作る場合が多い．※表5.2
- ◆ レーダー図を描くことも多い．※図5.10
- ◆ さらに，各クラスターごとに，いろいろな項目について集計してみると，より，各クラスターの性質がわかる．

表5.2　平均値表

表側：クラスタ	データ件数	Sepal.Leng	Sepal.Width	Petal.Lengt	Petal.Width
全体	150	5.843	3.057	3.758	1.199
clus1	50	5.006	3.428	1.462	0.246
clus2	68	6.562	2.987	5.326	1.879
clus3	32	5.625	2.628	4.013	1.244

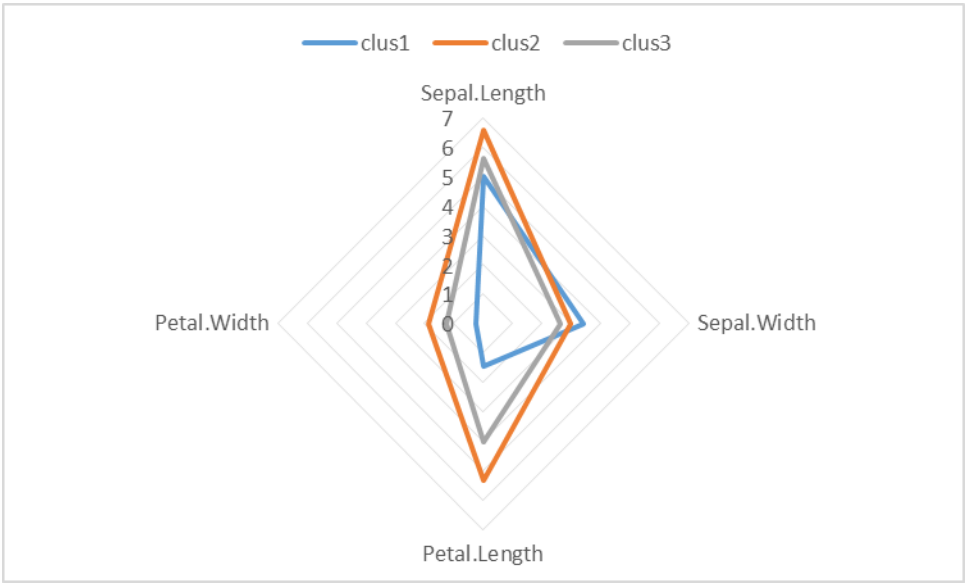


図5.10　レーダー図

6. 主座標分析

6.0 概要
6.1 手法について
6.2 プログラムの実行
6.3 出力について

6.0 概要

■ 単相2元の距離行列から，図（散布図）を得ます．数表が絵になり，見て分かるものになります．

※実行時，エラーメッセージらしい数字のみのダイアログが出ます．気にせず，「OK」をクリックしてください．

6.1　手法について

- 読み込む行列は距離行列
  - 単相2元の距離行列を読み込みます．すでに各要素間は距離を成しているという前提でデータを読み込みます．ですので，加算定数の計算はしません（非類似度を距離にするために，個々の要素ごと，最低限必要な数値をプラスする計算）．
- ２重中心化
  - 読み込んだ行列を2重中心化（ヤング・ハウスホルダーの変換）します．式で書くと，

$$P = -\frac{1}{2}\left(d_{ij}^2 - \bar{d}_{i.}^2 - \bar{d}_{.j}^2 + \bar{d}_{..}^2\right) \quad / \quad P = -\frac{1}{2} J_n D^{(2)} J_n$$

行列の要素で書いた場合／行列で書いた場合

※高根芳雄（1980）．多次元尺度法．東京大学出版会．pp.41-45.
（行に関しても列に関しても中心化している）

- 固有値を求めます
  - ２重中心化した距離行列を固有値分解します．
- 主座標
  - 固有ベクトル×$\sqrt{\text{固有値}}$で主座標を求めます．

6.2 プログラムの実行

- 使用するデータ
  - シート「CT_iris」にある相関係数行列を使います（図6.1）.
- データの指定
  - データ指定の範囲」ダイアログで，データ範囲を指定します．シート「CT_iris」の「CT_iris!$A$8:$E$12」を指定します．行ラベル・列ラベルを含む範囲を指定してください．（図6.2）
- 実行
  - [OK]をクリックすると，計算結果が，シート「Pcoa」に書き出されます．固有値などと，座標が，出力されます（図6.3）

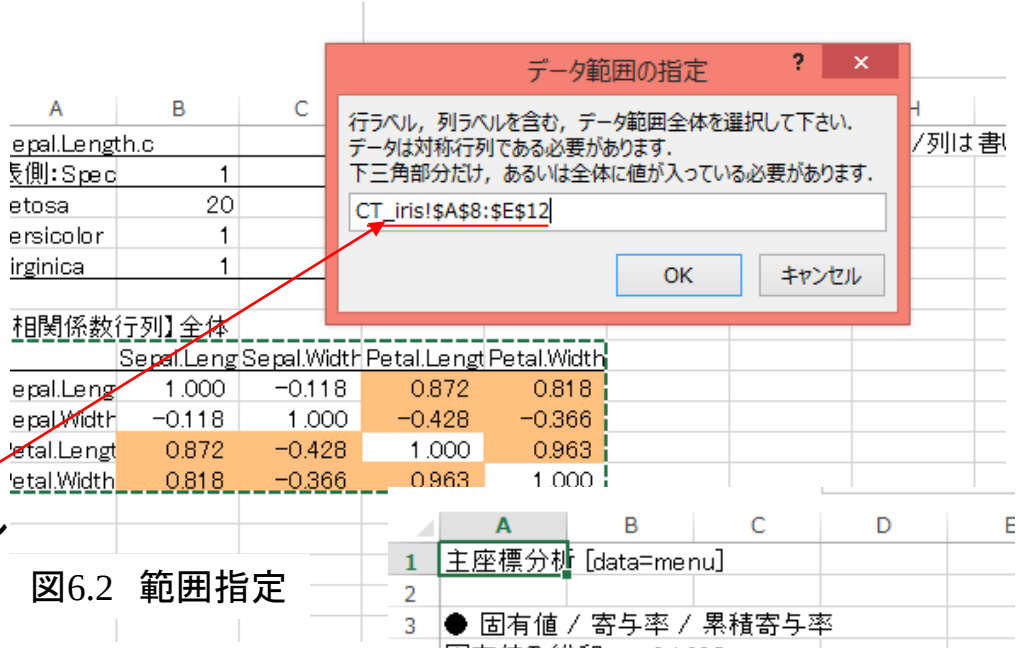


図6.2 範囲指定

【相関係数行列】全体

	Sepal.Leng	Sepal.Width	Petal.Lengt	Petal.Width
Sepal.Leng	1.000	−0.118	0.872	0.818
Sepal.Width	−0.118	1.000	−0.428	−0.366
Petal.Lengt	0.872	−0.428	1.000	0.963
Petal.Width	0.818	−0.366	0.963	1.000

図6.1 相関係数行列

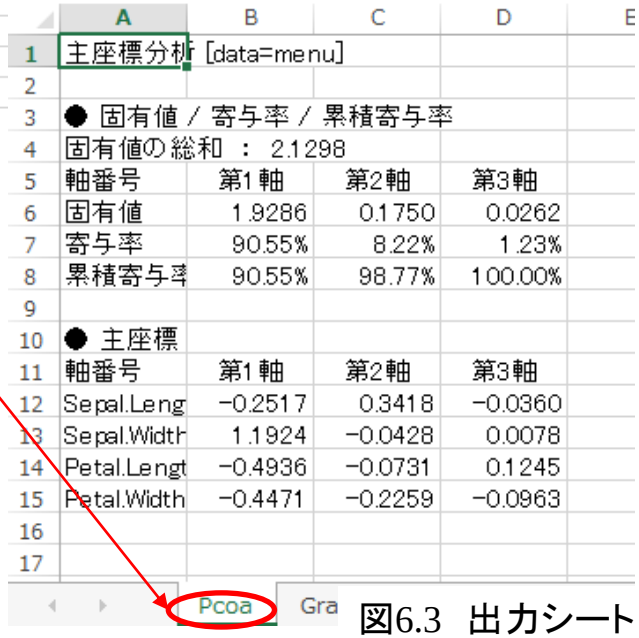
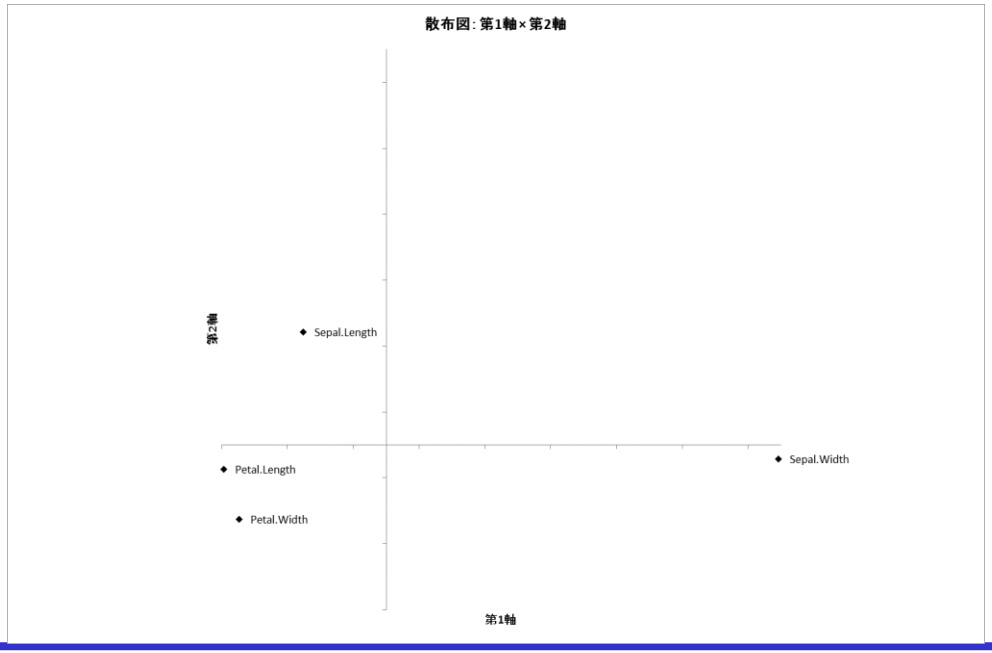


図6.3 出力シート

6.3　出力について

- 固有値と主座標
  - シンプルに上記のみ（右記図参照）
  - 上段が固有値など
  - 下段が主座標
- 散布図（map）
  - 主座標を使って散布図を描くと，図のとおり．



	A	B	C	D
1	主座標分析 [data=menu]			
2				
3	● 固有値 / 寄与率 / 累積寄与率			
4	固有値の総和 ： 2.1298			
5	軸番号	第1軸	第2軸	第3軸
6	固有値	1.9286	0.1750	0.0262
7	寄与率	90.55%	8.22%	1.23%
8	累積寄与率	90.55%	98.77%	100.00%
9				
10	● 主座標			
11	軸番号	第1軸	第2軸	第3軸
12	Sepal.Leng	−0.2517	0.3418	−0.0360
13	Sepal.Width	1.1924	−0.0428	0.0078
14	Petal.Lengt	−0.4936	−0.0731	0.1245
15	Petal.Width	−0.4471	−0.2259	−0.0963
16				
17				

Pcoa　Graph1　menu　iris

図6.4　出力シート（再掲）

横軸（第１軸）でsepal.widthとそれ以外が大きく分かれる．

縦軸（第2軸）でSepal.lengthが分かれる

図6.5　散布図（map）