「クラスター分析（1）：考え方」

■クラスター分析の目的と魅力

クラスター分析（cluster analysis）は、いくつかの変数から構成される多数のケースを類似性の高いグループ（クラスター）にまとめる「分類」のための技法である。人間は、多くの場合、最終的に何らかのグループ枠組みで物事を理解しようとする（例：この人の性格は○○タイプだ、この授業は私と関係のない種類の授業だ、この国の政治制度は独裁制だ、等）。その意味で、クラスター分析の志向はまったく自然なものといえる。

具体的には、いくつかの変数（質問項目）に対してクラスター分析を適用すれば、その回答パターンが総合的に近い人々同士を同じグループにする分類方法を提案してくれる。直感に頼りがちな分類手続きを客観的に行える利点もさることながら、分析者が考えてもいなかった分類枠組みを発見できる可能性が、クラスター分析の魅力である。

（教材）

「学生の恋愛観に関する調査」の問 13

調査対象：関西大学学生（2014 年度 計量社会学 I 受講生＋保田ゼミ 1 期生）＋大阪大学学生（2014 年度 統計学 A-I 受講生）計 188 名
→教材では関大 1 年生 49 名に限定

調査時期：2014 年 6 月

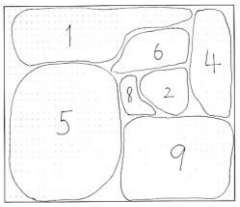
調査方法：講義受講者への集合調査

調査主体：保田ゼミ 2 期生

問 13 理想の恋人を思い浮かべたときに、以下の項目をどれくらい重視しますか。例のように、重視する度合いを面積の大きさで表現してください。すべての項目を記入する必要はありません。

①. 顔の良さ	②. スタイル	③. ファッション
④. 頭の良さ	⑤. 性格の良さ	⑥. ユーモア
⑦. 将来性	⑧. 趣味が合うか	⑨. 価値観が合うか

⑩. その他（具体的に：　　）



(例)
この例の場合、面積の大きな１，５，９は重視する度合いが高く、２，４，６，８は、そこまで重視しないということになる。また、描かれていない３，７，１０に関しては全く重視しないということになる。

回答スペース

■自分で分類してみよう

グループへの「分類」は、日常的におこなわれる作業であるが、我々は何を基準にグループの線引きをしているのだろうか。「学生の恋愛観に関する調査」の結果の一部（n＝49）をもとに、分類の考え方を改めて見直してみよう。話を簡単にするために、2 つの変数だけを取り扱うことにする。問 13 では「理想の恋人に求める条件」として 9 つの項目をどの程度重視するかを得点化しているが、そのうち「顔の良さ」と「性格の良さ」の 2 つで散布図を描くと下のようになる（図 1）。この散布図をもとに、人々を分類してみよう。

（練習）

1. 自分の感覚で 3 つのグループ（クラスター）に分けることを考えて、それぞれのグループを線で囲ってみよう。（メンバーが 1 人しかいないグループがあってもよい）
2. 自分が作ったそれぞれのグループ（クラスター）に名前を付けてみよう。
3. 作業の結果を、周りの人の分類と比べてみよう。分け方が違った場合、線の引き方についてどう考え方が違ったのか、話し合ってみよう。

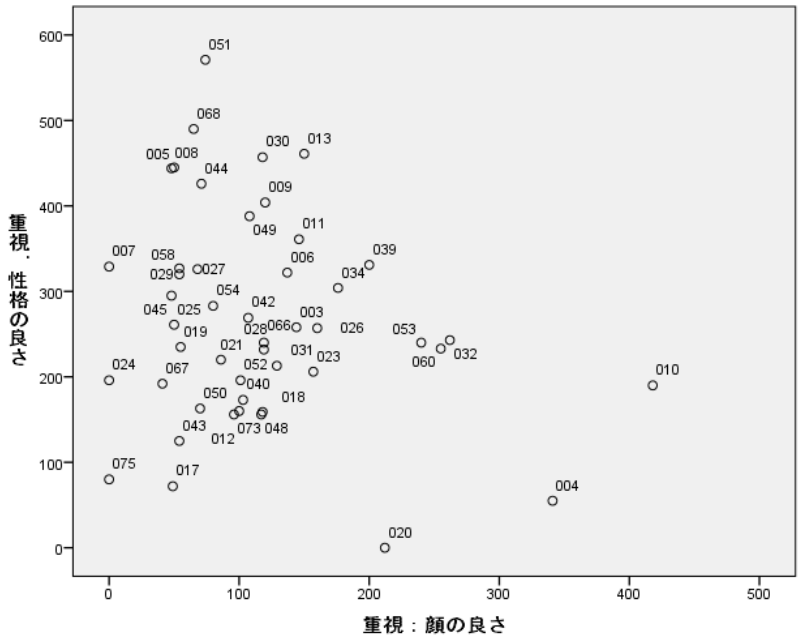


図 1 「顔を重視」と「性格を重視」の散布図

■連結すべきか、しないべきか

グループ（クラスター）の作り方には、当然いろいろな考え方があるが、基本的には、近くの人たちを同じグループに分類しようとしたはずである。クラスター分析もまったく同じ発想で、人々の間の正確な**距離**（distance）を計算して、分類の仕方を考える。

しかし、距離を基準にするという方針が決まっていても、誰と誰の連結を優先してグループを作るか、連結の判断をすることは意外と難しい。たとえば、ある 2 人を連結するかどうかは、彼らの距離だけでなく、彼らの周りにもっと距離の近い他人がいるかどうかにもかかってくるだろう（他にもっと近い人がいれば、そちらと連結してグループを作る方がよい）。あるいは、近い人同士をどんどん連結していくと、それぞれの連結には無理がなくても、結果としてできあがったグループが非常に大きな範囲に広がってしまい、端っこの人同士は似ても似つかない、ということになってしまうかもしれない。

クラスター分析では、どの部分を優先的に連結すべきかを 1 つずつ計算して、小グループ同士の連結を徐々に進めていく。しかし、上記のように、その判断は単純ではないので、分析者はその判断基準の設定をさまざまに変更することができる。統計分析ソフトは、それぞれの判断基準に応じた、複数の「正しい」分類結果を出してくれる。

代表的な連結基準は、以下の 3 つである（図 2）。**平均連結**（average linkage）は連結しようとする小グループの間ですべての個体間の組み合わせで距離を調べて、その平均距離が小さいような連結を優先する。近くに見えるものはどんどん連結して、おおらかなグループを作りやすい。これに対して**完全連結［最遠接法］**（complete linkage; furthest neighbor）の考え方は、グループに所属するどの個体を取っても距離が近いという完全性を要求する。すなわち、もっとも遠い個体間でも距離が小さくなるような連結を優先する。こちらは、グループがむやみに大きくなることを避ける判断基準といえる。また、**ウォード法**（Ward's method）は、グループ内での各ペアの距離の合計が大きく膨らまないように連結を進める。やや理解しにくいが、数学的には上の 2 つの基準を両方加味したバランスのよい判断基準である。ただし、かなり保守的な基準なので、あまり意味のない小さなグループをたくさん作ってしまうこともある。

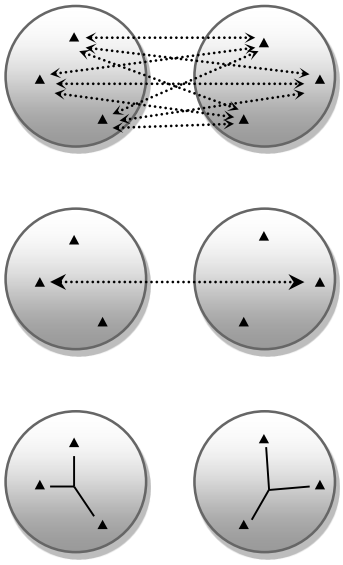


図 2　3 つの連結基準の模式図

これらの連結の基準のうちどれを採用するかは、理念的には分析者の目的に応じて決めるべきだが、実際的には、それぞれ試した上で、有意義な意味が読み取れる分類をしてくれた基準を採用することが多い。分類の価値は、結局、その分類がどれだけ示唆に富んでいるか（他のことを考えるために役立つか）で決まるからである。

（練習）

下の図は、実際にクラスター分析をおこなって３つのクラスターに分けた結果である。一方は、「平均連結」によるクラスターで、もう一方は「完全連結」によるクラスターである。どちらがどちらか予想してみよう。(ウォード法での分類は平均連結と同じ結果になったので省略)

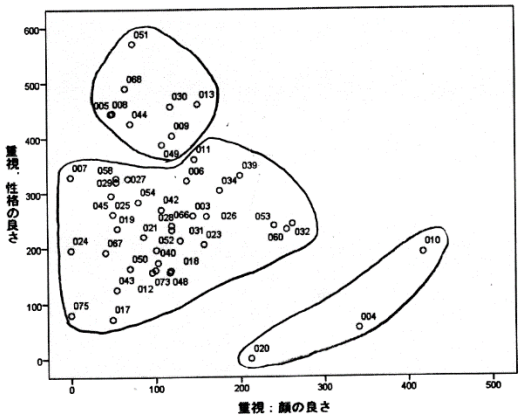


図 3a　（平均連結・完全連結）の結果

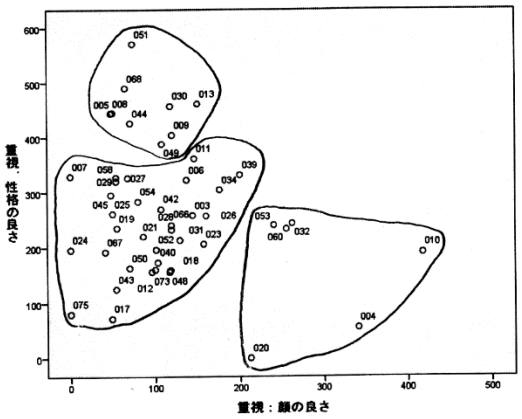


図 3b　（平均連結・完全連結）の結果

※加工したデータや必要なアウトプットの保存用のメディアを持参するように。関大 My ボックスで、大学の Z ドライブに自宅からもアクセスしてもよい。

「関西大学 IT センター 関大 My ボックス」

http://www.itc.kansai-u.ac.jp/services/webstorage.html

※資料やデータは Web 上にも置いているので、必要に応じて復習すること。

「保田時男のページ」

http://www2.itc.kansai-u.ac.jp/~tyasuda/

※自宅の PC への SPSS のインストール方法は下記を参照。

「関西大学 IT センター ダウンロードステーション」

http://www.itc.kansai-u.ac.jp/services/downloadstation.html

「クラスター分析（2）：SPSS で実践」

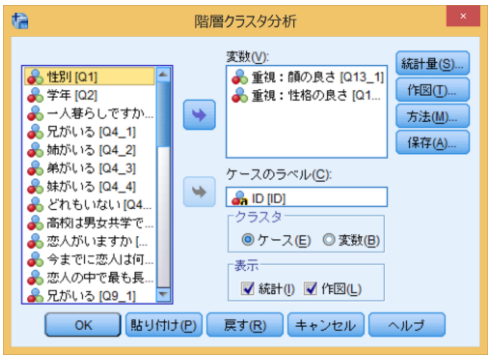
■SPSS でやってみよう

前回は、散布図を見ながら、クラスター分析の考え方について学習した。今回は SPSS を操作して、実際にクラスター分析の結果を出力しながら、一通りの手続きを経験しよう。

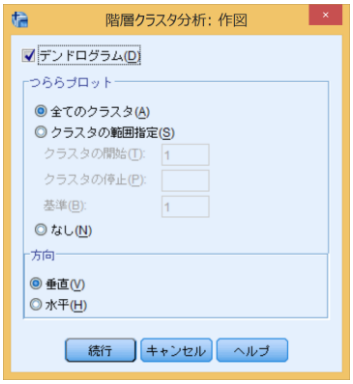
クラスター分析の操作

---

①メニューから、分析➜分類➜階層クラスタ
②分類のために用いる変数群をすべて［変数］欄へ
（②’ 結果をケース番号以外で示す場合は、ラベルの変数を［ケースのラベル］欄へ）



③作図ボタンを押して、一番上の［デンドログラム］にチェックして続行



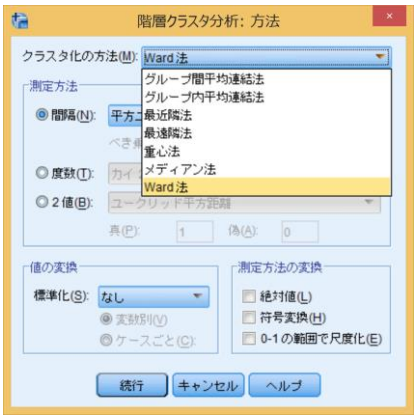
④方法ボタンを押して、［クラスタ化の方法］を選択して続行

平均連結→［グループ間平均連結法］

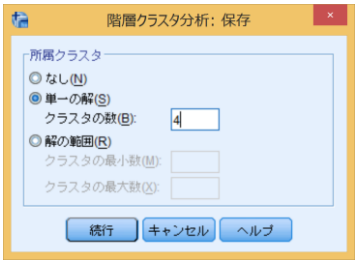
完全連結→［最遠隣法］

ウォード法→［Ward 法］

（④’ すべての変数を標準化してから分析するときは、［標準化］で［Z 得点］を選択）



（⑤<u>各ケースがどのクラスターに分類されたかを、変数で保存するとき</u>は、保存ボタンを押して、［単一の解］を選び、自分が採用する［クラスタの数］を入力）



⑥元の窓でＯＫボタン

■デンドログラムの読み方

クラスター分析の結果は、ふつう**デンドログラム［樹状図］**（dendrogram）と呼ばれるトーナメント表のような図式で表される（図 4）。1 人だけで 1 グループ（全員別グループ）という左端の状態から始めて､右に進むにつれてどことどこを連結していくのが自然かを、順に表現している（最後には全てが連結されて 1 つだけのグループになる）。

グループの個数が決まっているならば、その数になるように区切り線を引いて、誰がどのグループに分類されるのかを確認できる。また、枝が長いことは連結することが困難であることを示しているので、これを参考に適切なグループ数を模索することもできる。枝が長くなっているところは（統計的には）区別を重視しなければならない分類、枝が短いところはひとまとまりにすることに抵抗が少ない分類である。下図の場合、枝の長さだけで考えれば、5～6 個のグループに分類するのがよさそうである。それ以上に分類の数を増やすと急に細かい（枝の短い）分類になってしまうからである。

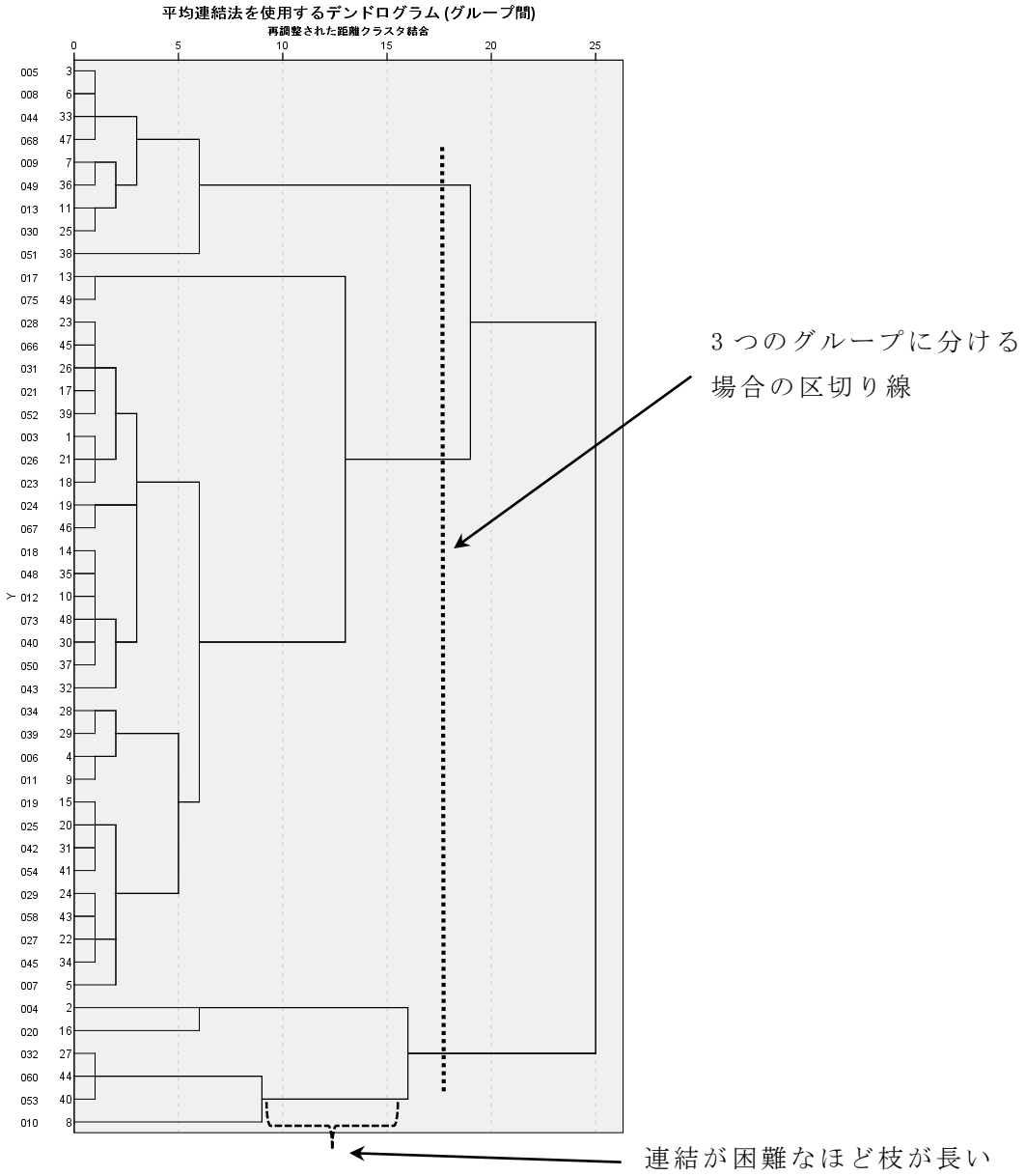
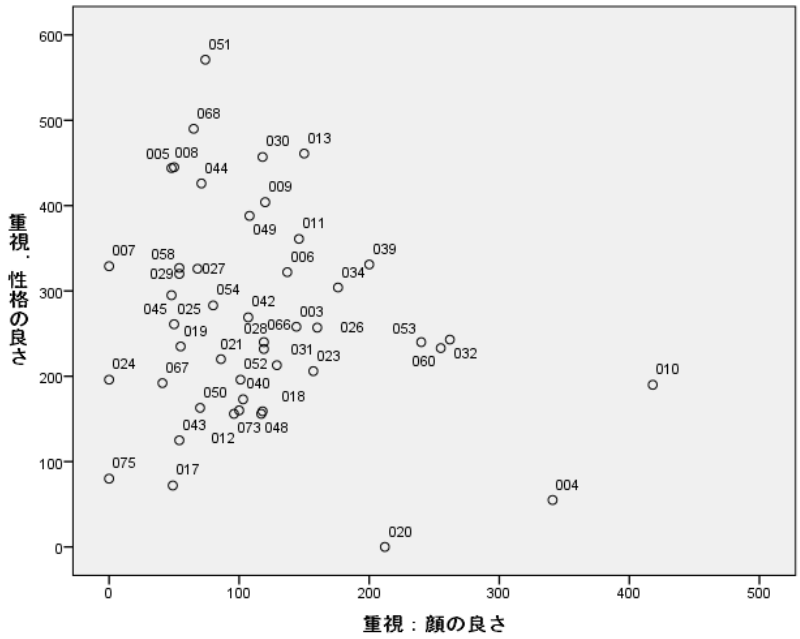


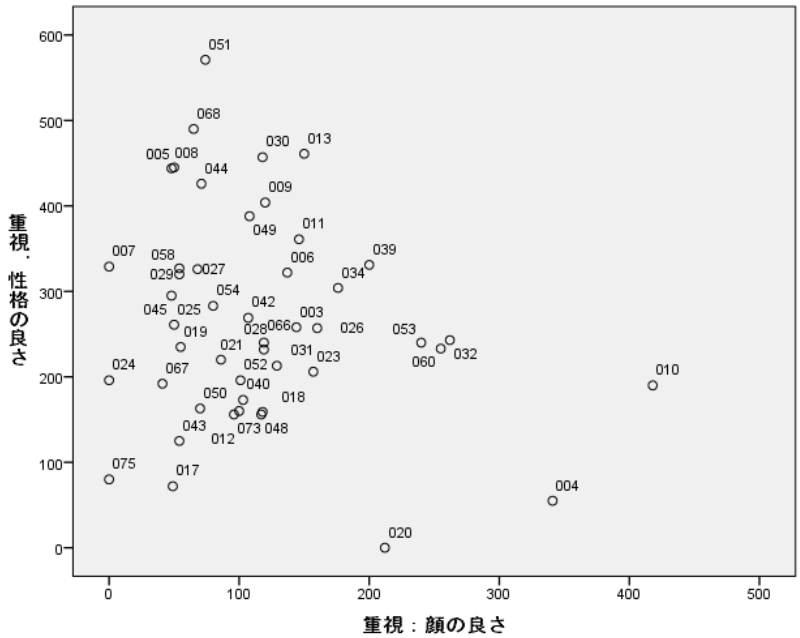
図 4　デンドログラムの例（平均連結）

（練習）

1. 前ページの図 4 は、「性格を重視」「顔を重視」のデータを、平均連結でクラスター分析した結果である。前回の図 3b の散布図での線引きと、このデンドログラムが対応していることを確認しよう。
2. 自分で SPSS を操作して、図 4 と同じデンドログラムを出力しよう。
3. かりに、もっと細かい分類を採用してグループ（クラスター）を **5 個**にするならば、どのような分類になるか。出力したデンドログラムを読み取って、下の散布図にグループ分けの区切り線を描こう。



4. 設定を「ウォード法」に変更してデンドログラムを出力し、その結果から（枝の長さを参考にしつつ）自分が適切だと思うクラスターの数を決めよう。そして、そのクラスターを下の散布図に描こう。



■グループ（クラスター）の特徴を記述する

クラスター分析に使用した変数の数が 2 個しかない場合には、析出されたグループ（クラスター）の特徴は、散布図で簡単に確認できる。しかし、多くの変数を用いた場合は、多次元空間（3 次元、4 次元、……）のの散布図の中でグループ分けをしていることになるので、我々は視覚的にそのグループの特徴を捉えることはできない。

その場合は通常､各変数の平均値などを用いて､それぞれのグループの特徴を記述する。平均値は表にまとめてもよいし、直感的に訴えたければグラフにまとめてもよい。

また､読み取った特徴から各グループには分析者が名前を付けることが多い｡「第 1 クラスターは､……という特徴があるので､顔偏重タイプと名付けました｡第 2 クラスターは､バランスタイプで、……」といったように名前が付いていると、分類の結果が他人にもわかりやすい。ただし、名前に引っ張られて実際の特徴とは異なる意味を感じてしまうこともあるので、命名は慎重におこなおう。

「恋人に重視する条件」の 9 つの項目をすべて活用したクラスター分析で例を示そう。9 項目の変数群なので、9 次元空間での距離からグループ分けを考えていることになる。一番無難なウォード法を用いたところ、そのデンドログラムは図 5 のようになった。枝の長さで考えると、3～7 個程度のクラスターが適切そうだが、今回は 5 つとしてみた。

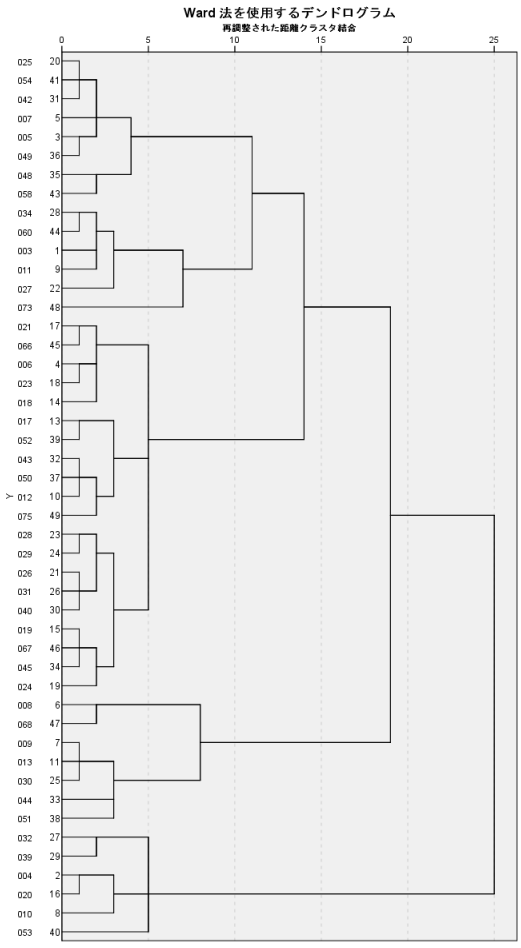


図 5　9 変数のクラスター分析でのデンドログラム（ウォード法）

5つのクラスターごとに各変数の平均点を算出すると、表1のようになった。図6はこれをレーダーチャートに表わしたものである。直感的には、表よりもこちらの方がわかりやすい。平均点をよく見て、それぞれのグループに「趣味共有型」「ルックス型」「享楽型」「バランス型」「性格オンリー型」という名前を付けた。一番人数が多いのはバランス型（20人）で、他は横並びである。性格を第一にあげながらも他もそれぞれ大切というバランス重視の学生が、やはり一番多いようである。ルックス型は顔やスタイルを重視する傾向が突出している。趣味共有型と享楽型は価値観と性格を重視する点で似ているが、前者は趣味を通じて、後者はユーモアを通しての交流を重視している。性格オンリー型はとにかく性格重視であるが、もしかすると具体的な理想のイメージがわいていないのかもしれない。

表1　クラスター別の各変数の平均点

	クラスター1（趣味共有型）	クラスター2（ルックス型）	クラスター3（享楽型）	クラスター4（バランス型）	クラスター5（性格ｵﾝﾘｰ型）
顔の良さ	148	279	71	85	92
スタイル	92	191	35	66	84
ファッション	46	89	45	64	42
頭の良さ	27	48	52	74	76
性格の良さ	273	177	307	203	465
ユーモア	37	31	200	96	80
将来性	26	55	19	77	78
趣味が合うか	300	30	145	97	49
価値観が合うか	207	34	265	154	136
分類ケース数	6	6	8	20	7

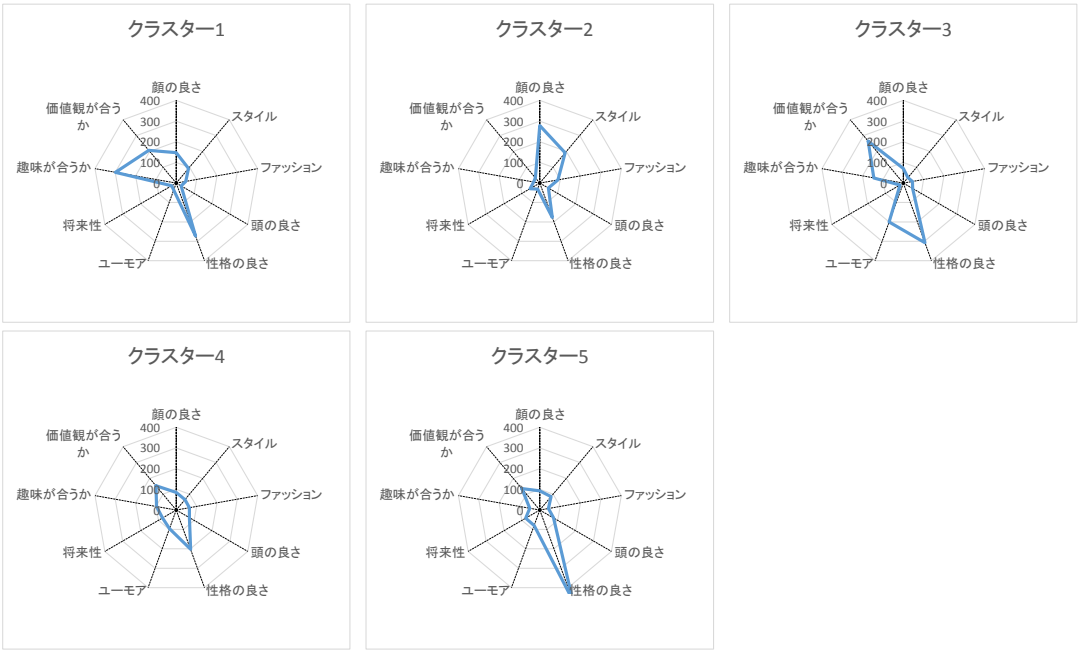


図6　クラスター別のレーダーチャート

（練習）

表 1 と同じ数値が読み取れるまでの操作を SPSS で再現してみよう。

(1) 9 個の変数でウォード法のクラスター分析を実行

(2) デンドログラムから 5 個のクラスターが適切と読み取る

(3) 各ケースがどのクラスターに属するのかを示す変数を作成

(4) 作成されたグループ分けの変数を使って、グループ別の平均値を算出

■（参考）関連する SPSS の操作

グループ別に平均点を出力

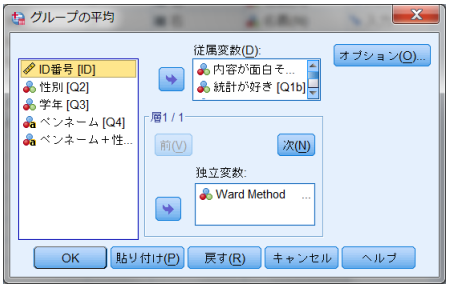
クラスター分析で、グループごとに平均点を算出するためには、まずそれぞれのケースが何個目のグループ（クラスター）に属するかを示す変数を作成しなければならない。クラスター分析を実行する際にオプション指定をすれば、この変数は「CLU...」といった変数名で自動的に作成される（p.6 の⑤）。

あとはこの変数を使って、ふつうにグループ別に平均値を出力する命令を出せばよい。

①メニューから、分析 ➜ 平均の比較 ➜ グループの平均

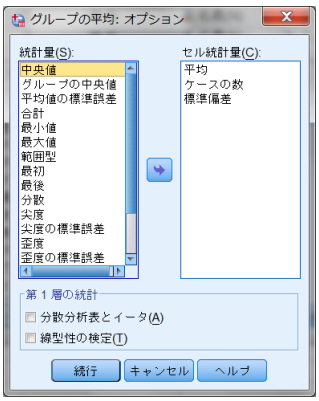
②グループ分けに使う変数（CLU...）を［独立変数］へ。

平均値を出す変数群（クラスター分析に使った一連の変数）を［従属変数］へ。



③オプションボタンを押して、集計したい統計量を［セル統計量］に選択して続行

（必要ならば、平均以外にもいろいろな統計量を、グループ別に集計できる）



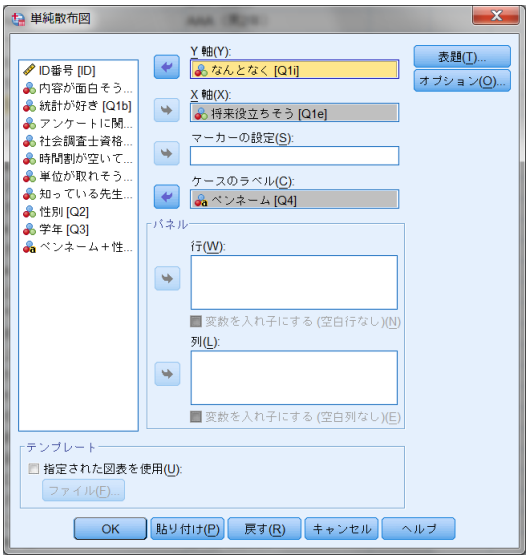
④元の窓でＯＫボタン

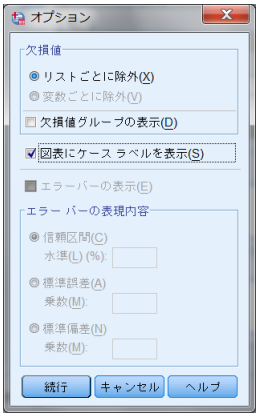
散布図を描く

2 変数だけでクラスター分析をする場合には、あらかじめ散布図で各ケースの位置関係を把握しておくと、結果が理解しやすい。

①メニューから、グラフ ➜ レガシーダイアログ ➜ 散布図/ドット
②［単純な散布］を選択して、定義ボタン

③［Y 軸］欄と［X 軸］欄にそれぞれの変数を移動





（④散布図の印を個人識別する場合、オプションボタンを押して、［図表にケースラベルを表示］にチェックして続行）
（個人識別をケース番号以外で示したい場合、ラベルの変数を［ケースのラベル］欄へ）
⑤元の窓でＯＫボタン

「クラスター分析（3）：発展」

■非階層的クラスター分析

以上の説明でクラスター分析の一連の手続きが使える状態になった（はず）。ただ、クラスター分析は「分類」の技法であり、分類の仕方は多種多様であるから、クラスター分析にはもっといろいろな側面から発展の可能性が残されている。比較的よく知られる側面をいくつか紹介しておこう。

この講義で扱ったクラスター分析は、全ケースがばらばらのクラスターという状況から出発して、距離の近いものを順次連結していくデンドログラムを用いた方法であった。このようなクラスター分析を、とくに**階層的クラスター分析**（hierarchical clustering）と呼ぶ。異なるアプローチとして、クラスターの個数を自分で指定してやり、適当な初期分類から 1 ケースずつクラスター間を移動させながらより適切な分類に近づいていく**非階層的クラスター分析**（non-hierarchical clustering）がある。階層的クラスター分析は、ケース数が少なく、個別のケースの所属先やまとまり方を記述したい場合に向いている。逆に、非階層的クラスター分析は、ケース数が多く、個別のケースよりも全体としてどのような特徴のクラスターが形成されたかを記述したい場合に向いている。

非階層的クラスター分析では、通常、小さな数のクラスターから順に徐々にクラスターの数の設定を大きくしていき、分類が過剰になったところで止めて、直前のクラスター数を採択する。

SPSS では「分析 ➔ 分類 ➔ 大規模ファイルのクラスタ」のメニューで非階層的クラスター分析が実行できる。基本的に、「クラスターの個数」を自分で指定してやる以外は、階層的クラスター分析と同様である。デンドログラムは作ることができないので、クラスター別の各変数の平均値が自動的に出力され、各クラスターの内容を記述してくれる。

■変数の標準化

クラスター分析では、ケース間の距離を基準にして分類を考えるので、分析する変数群の分散（ばらつき具合）には気を遣う必要がある。かりに使用した変数のうち 1 つだけが分散（ばらつき具合）が大きい変数だったりすると、その変数の値だけが距離に大きく影響してしまうからである。このため、まったく内容の異なる（分散の異なる）変数群で分析を行う際には、各変数を平均値 0、標準偏差 1 の得点（z スコア）に標準化してから分析に使うことが多い。たとえば、勤続年数と年収でクラスター分析をすると、勤続年数は±10 年くらいのばらつきなのに対して、年収は±1,000,000 円くらいのばらつきがあるため、年収の方が分類の仕方に大きな影響をおよぼしたりすることがある。この授業では、同じ尺度の変数だけを使ったので、標準化は考えなかった。

■距離の測り方

日常で「距離」と言えば、ふつう定規で測れる長さのことであるが、数学的にはこれをユークリッド距離と呼ぶ。クラスター分析で距離を観察するのは、ケース間の類似性、非

類似性を調べるためであるから、必ずしもユークリッド距離にこだわる必要はない。実際、標準的には、平方ユークリッド距離と呼ばれる測り方が用いられることが多い。これはユークリッド距離の 2 乗である。こちらを採用した方が、平均距離を算出したりする際に、その離れ具合を加速度的に評価することになって、ふつうは望ましい。その他にも多種多様な距離概念が存在する。

この講義では、標準の平方ユークリッド距離を用いて分析をおこなうことにするので、あまり深入りする必要はないが、そういう別の考え方もあるということを知っておこう。連結の基準の設定と同じように、理念的には、分析者の考えにマッチする距離概念を選択して採用すべきであるが、実際的には、分析結果とにらみ合って、よく意味のわからない分類が示されたときに、距離概念の設定を変更して、結果が変わらないか試してみたりすることが多い。

■使用した変数以外でのクラスターの記述

この講義では、それぞれのクラスターの特徴を記述する際に、使用した変数の平均値を参照する方法を紹介した。同じように、分析に「使用していない」変数の平均値も、当然算出することができる。たとえば、平均学年とか、女子学生の割合（男子を 0、女子を 1 としたダミー変数の平均値）で、各クラスターの特徴を記述できる。

このような記述は、分類のために使った変数（この科目への関心理由）が他の変数とどう関係しているのかを読み取るために、非常に便利である。手軽な方法なので、有効に活用してほしい。

■よい分類とは？

クラスター分析では、いろいろな分類の可能性が示されるので、結局、どの分類の仕方がよいのか迷ってしまう場合も多い。そのときには、改めて、なぜ分類がしたいのかを思い出してみよう。クラスター分析をおこなうべき状況は、大きく分けて 2 つある。

1 つは、だいたい分類の仕方は想像がつくのだけれども、具体的にどのケースがどのグループに属するのかを知りたいという場合である。匿名の回答者データの場合、このような関心は起こらないはずであるが、たとえば記名式調査であるとか、あるいは都道府県がケース単位になっているデータのように、何らかの匿名でない意味のある個体が 1 ケースを表わしている場合、このような関心がありえる。

もう 1 つは、そもそもどんな分類の仕方が考えられるのか、想像がつかない、あるいは自己流の分類ではしっくり来ない、というような場合である。このようなとき、クラスター分析は、統計という客観的基準によって、主観的には把握しにくいような分類の仕方を見つけ出してくれる。だから、新しい、そして納得のいく分類の仕方に出会えれば、それはよい分類であり、そのクラスター分析は成功である。デンドログラムを眺めているだけではよい分類かどうかは判断できない。使用した変数やそれ以外の変数で、クラスターの特徴を記述する（平均値を出す）ことで、何か訴えてくるものがある分類方法がないか、よく探索、吟味しよう。

■クラスター分析の手順のまとめ

（階層的）クラスター分析の一般的な手順を、改めてまとめておく

（1）クラスターの連結法の選択

まず、クラスターの連結法をそれぞれ選択する（必要があれば、上記の「距離の測り方」も選択する）。とくに理由がなければ、ウォード法から始めるのが堅実なことが多い。実際的には、いろいろと試してみて納得できる結果を事後的に採用する。

（2）クラスター数の選択

クラスターの連結過程は、デンドログラムに表される。どの個数でクラスターを切れば、どのようなクラスターが形成されるかは一目瞭然なので、納得のいく個数でクラスターを切って、クラスターの数を決める。距離が離れているクラスターを連結する際には、デンドログラムの枝も長くなるように表現されているので、基本的には枝の長い分類は残すように考えればよい。

（3）読み取るべきポイント

・各クラスターの特徴（使用した変数の平均値の比較）
・各ケースがどのクラスターに所属しているか
・より適切な分類方法はないか模索
・析出されたクラスターと他の変数の関係（他の変数の平均値の記述など）

＜参考文献＞

Landau, Sabine and Brian S. Everitt, 2004, *A Handbook of Statistical Analyses Using SPSS*, Chapman & Hall/CRC.

村瀬洋一・高田洋・廣瀬毅士, 2007, 『SPSS による多変量解析』 オーム社.

小田利勝, 2007, 『ウルトラ・ビギナーのための SPSS による統計解析入門』 プレアデス出版.（←SPSS でのデータ分析を学習するために、絶対おすすめ！）

＊＊＊　小課題　＊＊＊

Web ページ（http://www2.itc.kansai-u.ac.jp/~tyasuda/）から「小課題用データ」をダウンロードして、大学生の幸福感の軌跡について、クラスター分析をおこない、「小課題提出ファイル」のレポートを完成させなさい。（**10 月 28 日提出**、1 回遅れまでは減点で受け取ります）