College Analysis レファレンスマニュアル

－ 多変量解析１ －

目次

１．実験計画法

1.1　実験計画法の理論

実験計画法は、異なるいくつかの条件下でデータを求め、その間に差があるかどうか検討する手法の総称である。このプログラムではこれらの分析の関係を図１のようにまとめ、それに基づいて分析メニューが作られている。

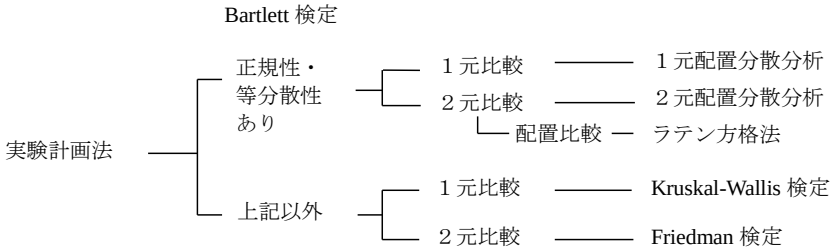


図１ 実験計画法の全体像

1) １元配置分散分析

１元比較の場合、データは表１の形で与えられる。ここに水準数は p、水準 i のデータ数は n_i で与えられ、データは一般に $x_{i\lambda}$ で表わされる。

表１　１元比較のデータ

水準１	水準２	…	水準 p
x_{11}	x_{21}	…	x_{p1}
x_{12}	x_{22}	…	x_{p2}
：	：		：
x_{1n_1}	x_{2n_2}	…	x_{pn_p}

位置母数の比較は正規性と等分散性の有無によって１元配置分散分析か、Kruskal-Wallis 検定かに分かれる。正規性が認められ、多群間の等分散性が認められる場合には、１元配置分散分析が利用できる。この等分散性の検定には Bartlett 検定を利用することができる。

１元配置分散分析のデータ $x_{i\lambda}$ は、水準 i に固有な値 α_i と誤差 $\varepsilon_{i\lambda}$ を用いて以下のように表わされると考える。

$$x_{i\lambda} = \mu + \alpha_i + \varepsilon_{i\lambda}, \qquad \varepsilon_{i\lambda} \sim N(0, \sigma^2)$$ 分布［異なる i, λ について独立］

データの全変動 S は、水準内変動 S_E 及び水準間変動 S_P を用いて以下のように表わされる。

$$S = \sum_{i=1}^{p}\sum_{\lambda=1}^{n_i}(x_{i\lambda} - \bar{x})^2 = \sum_{i=1}^{p}\sum_{\lambda=1}^{n_i}(x_{i\lambda} - \bar{x}_i)^2 + \sum_{i=1}^{p} n_i(\bar{x}_i - \bar{x})^2 = S_E + S_P$$

誤差 $\varepsilon_{i\lambda}$ の正規性から、それぞれの変動は以下の分布に従うことが分かる。

$$S/\sigma^2 \sim \chi^2_{n-1} \text{分布}, \quad S_E/\sigma^2 \sim \chi^2_{n-p} \text{分布}, \quad S_P/\sigma^2 \sim \chi^2_{p-1} \text{分布}$$

１元配置分散分析は、$\alpha_i = 0$ として、以下の性質を利用する。

$$F = \frac{S_P/(p-1)}{S_E/(n-p)} \sim F_{p-1,n-p} \text{分布}$$

2) Kruskal-Wallis の順位検定

Kruskal-Wallis の順位検定は、データの分布型によらず、p 種類の水準の中間値に差があるかどうか判定する手法である。まず、全データの小さい順に順位 $r_{i\lambda}$ を付け、水準ごとの順位和 w_i を求める。但し、同じ大きさのデータにはそれらに順番があるものとした場合の順位の平均値を与える。検定には各水準の中間値が等しいとして以下の性質を利用する。

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{p} n_i \left(\frac{w_i}{n_i} - \frac{n+1}{2} \right)^2 \sim \chi^2_{p-1} \text{分布}$$

3) Bartlett の検定

Bartlett の検定は、各水準の母分散が等しいとして以下の性質を利用する。

$$\chi^2 = \frac{1}{C} \left[(n-p) \log V_E - \sum_{i=1}^{p} (n_i - 1) \log V_i \right] \sim \chi^2_{p-1} \text{分布}$$

ここに、V_E，V_i，C は n を全データ数として以下のように与えられる。

$$V_E = \frac{1}{n-p} \sum_{i=1}^{p} \sum_{\lambda=1}^{n_i} (x_{i\lambda} - \bar{x}_i)^2, \quad V_i = \frac{1}{n_i - 1} \sum_{\lambda=1}^{n_i} (x_{i\lambda} - \bar{x}_i)^2,$$

$$C = 1 + \frac{1}{3(p-1)} \left[\sum_{j=1}^{p} \frac{1}{n_j - 1} - \frac{1}{n-p} \right]$$

4) ２元配置分散分析

２元比較の場合、２つの水準間または水準とブロック間の差を同時に検定する。前者は２つの水準の交点に複数のデータを含んだデータ構造であり、繰り返しのある場合とも言われる。後者は水準とブロックの交点に完備乱塊法によって得た１つのデータが含まれ、繰り返しのない場合とも言われる[8]。２元配置分散分析は、正規性が認められ、各水準やブロック間で分散が等しい場合にのみ有効である。以下２つの場合に分けて分析法について説明する。

表 2　2 元配置分散分析（繰り返しあり）

	水準 Q_1	…	水準 Q_s
水準 P_1	x_{111} : $x_{11n_{11}}$	… … …	x_{1s1} : $x_{1sn_{1s}}$
:	:	:	:
水準 P_2	x_{r11} : $x_{r1n_{r1}}$	… … …	x_{rs1} : $x_{rsn_{rs}}$

まず繰り返しがある場合を考える。データは表 2 の形式で与えられる。各データは水準 P_iに固有の量を α_i、水準 Q_jに固有の量を β_j、水準 P_iと水準 Q_jの相互作用を γ_{ij}、誤差を $\varepsilon_{ij\lambda}$ として、以下のように表わせると考える。

$$x_{ij\lambda} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij\lambda}, \quad \varepsilon_{ij\lambda} \sim N(0, \sigma^2) \text{ 分布［異なる } i, j, \lambda \text{ に対して独立］}$$

但し、各パラメータには以下の条件を付ける。

$$\sum_{i=1}^{r} n_{i\bullet}\alpha_i = 0, \quad \sum_{j=1}^{s} n_{\bullet j}\beta_j = 0, \quad \sum_{i=1}^{r} n_{ij}\gamma_{ij} = 0, \quad \sum_{j=1}^{s} n_{ij}\gamma_{ij} = 0$$

ここにデータ数に関しては以下の記法を用いている。

$$n_{i\bullet} = \sum_{j=1}^{s} n_{ij}, \quad n_{\bullet j} = \sum_{i=1}^{r} n_{ij}, \quad n = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}$$

各水準及び全体のデータ平均を $\bar{x}_{ij}$， $\bar{x}_{i\bullet}$， $\bar{x}_{\bullet j}$， $\bar{x}$ として、全変動 S、水準 P 間の変動 S_P、水準 Q 間の変動 S_Q、相互作用の変動 S_I、水準内変動 S_E を以下で与えると、

$$S = \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{\lambda=1}^{n_{ij}}(x_{ij\lambda} - \bar{x})^2, \quad S_P = \sum_{i=1}^{r} n_{i\bullet}(\bar{x}_{i\bullet} - \bar{x})^2, \quad S_Q = \sum_{j=1}^{s} n_{\bullet j}(\bar{x}_{\bullet j} - \bar{x})^2,$$

$$S_I = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(\bar{x}_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x})^2, \quad S_E = \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{\lambda=1}^{n_{ij}}(x_{ij\lambda} - \bar{x}_{ij})^2,$$

全変動 S はその他の変動を用いて以下のように表わされる。

$$S = S_P + S_Q + S_I + S_E$$

水準間の差や相互作用の有無を検定するためには、以下の性質を利用する。

$\alpha_i = 0$ のとき　　$F_P = \dfrac{S_P/(r-1)}{S_E/(n-rs)} \sim F_{r-1,n-rs}$ 分布　　（水準 P 間の差）

$\beta_j = 0$ のとき　　$F_Q = \dfrac{S_Q/(s-1)}{S_E/(n-rs)} \sim F_{s-1,n-rs}$ 分布　　（水準 Q 間の差）

$\gamma_{ij} = 0$ のとき　　$F_I = \dfrac{S_I/(r-1)(s-1)}{S_E/(n-rs)} \sim F_{(r-1)(s-1),\, n-rs}$ 分布　　（相互作用）

もう１つの２元配置分散分析はブロック毎に無作為化されたデータを用いて、水準やブロック間の差を調べるもので、繰り返しのない場合と呼ばれている。これは対応のある 1 元配置分散分析とも呼ばれ、データは表 3 のようにブロックと水準の交点に１つだけ値が入る。

表 3　２元配置分散分析（繰り返しなし）

	水準 1	水準 2	…	水準 s
ブロック 1	x_{11}	x_{12}	…	x_{1s}
ブロック 2	x_{21}	x_{22}	…	x_{2s}
:	:	:		:
ブロック r	x_{r1}	x_{r2}	…	x_{rs}

水準 j に固有な量を α_j、ブロック i に固有な量を β_i、誤差を ε_{ij} として、データ x_{ij} を以下のように表わす。

$$x_{ij} = \mu + \alpha_j + \beta_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$ 分布［異なる i, j に対して独立］

但し、パラメータ α_j，β_i には以下の条件を付ける。

$$\sum_{j=1}^{s} \alpha_j = 0, \quad \sum_{i=1}^{r} \beta_i = 0$$

水準、ブロック及び全体の平均を、$\bar{x}_{\bullet j}$，$\bar{x}_{i\bullet}$，$\bar{x}$ として、全変動 S、水準間の変動 S_p、ブロック間の変動 S_B、誤差変動 S_E を以下で与えると、

$$S = \sum_{i=1}^{r}\sum_{j=1}^{s}(x_{ij} - \bar{x})^2, \quad S_P = \sum_{i=1}^{r}\sum_{j=1}^{s}(\bar{x}_{\bullet j} - \bar{x})^2, \quad S_B = \sum_{i=1}^{r}\sum_{j=1}^{s}(\bar{x}_{i\bullet} - \bar{x})^2,$$

$$S_E = \sum_{i=1}^{r}\sum_{j=1}^{s}(x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x})^2,$$

全変動 S はその他の変動を用いて以下のように表わされる。

$$S = S_P + S_B + S_E$$

水準間やブロック間の差を検定するためには、以下の性質を利用する。

$\alpha_j = 0$ のとき　　$F_P = \dfrac{S_P/(s-1)}{S_E/(r-1)(s-1)} \sim F_{s-1,(r-1)(s-1)}$ 分布　（水準間の差）

$\beta_i = 0$ のとき　　$F_B = \dfrac{S_B/(r-1)}{S_E/(r-1)(s-1)} \sim F_{r-1,(r-1)(s-1)}$ 分布　（ブロック間の差）

5）Friedman の順位検定

対応のある 1 元比較（繰返しのない 2 元比較）でブロック差が大きい場合や誤差の正規性に問題がある場合は、Friedman の順位検定を用いる。これは各ブロック毎にデータに順位を付け、水準毎の順位和を用いて検定を行なうものである。今、水準 j の順位和を w_j とし、水準間に差がないことを仮定して、以下の性質を用いる。

$$D = \frac{12}{s(s+1)r}\sum_{j=1}^{s} w_j^2 - 3r(s+1) \sim \chi^2_{s-1} \text{ 分布}$$

一般に Friedman 検定は対応のある場合の Wilcoxon の符号付順位和検定の拡張のように考えられがちだが、群間で順位を付ける理論構成から、むしろ McNemar 検定の拡張と言ってもよい。

6）ラテン方格法

実験順序によって結果に影響が出るような場合、それぞれの個体に対する処理（水準と呼ぶ）を順序を変えて 1 回ずつ施す方法がラテン方格法である。表 4 にデータとその処理順序（配置と呼ぶ）の例を示す。

表 4 ラテン方格法のデータと処理順序の例

	水準 1	水準 2	水準 3	水準 4
個体 1	$x_{11(1)}$	$x_{12(2)}$	$x_{13(3)}$	$x_{14(4)}$
個体 2	$x_{21(2)}$	$x_{22(3)}$	$x_{23(4)}$	$x_{24(1)}$
個体 3	$x_{31(3)}$	$x_{32(4)}$	$x_{33(1)}$	$x_{34(2)}$
個体 4	$x_{41(4)}$	$x_{42(1)}$	$x_{43(2)}$	$x_{44(3)}$

配置は、データの添え字に付いた括弧内の数字で表わすが、配置 k は各水準と各個体に一度だけ現れ、水準 j と個体 i による関数とみなすことができる。データ $x_{ij(k)}$ は、水準 j に固有な量を α_j、個体 i に固有な量を β_i、配置差に固有な量を γ_k として、以下のように表わせるものとする。

$x_{ij(k)} = \mu + \alpha_j + \beta_i + \gamma_k + \varepsilon_{ijk}$，　$\varepsilon_{ijk} \sim N(0, \sigma^2)$ 分布［異なる i, j, k に対して独立］

但し、パラメータ α_j，β_i，γ_k には以下の条件を付ける。

$$\sum_{j=1}^{r}\alpha_j = 0,\ \sum_{i=1}^{r}\beta_i = 0,\ \sum_{k=1}^{r}\gamma_k = 0$$

今後の計算のために、水準別合計$T_{\bullet j}$，個体別合計$T_{i\bullet}$，全合計Tを以下のように与える。

$$T_{\bullet j} = \sum_{i=1}^{r} x_{ij(k)},\ T_{i\bullet} = \sum_{j=1}^{r} x_{ij(k)},\ T = \sum_{i=1}^{r}\sum_{j=1}^{r} x_{ij(k)}$$

また、順序 k が付いたデータの合計T_kも求めておく。さて$C = T^2/r^2$とおいて、全変動S、水準間の変動S_P、個体間の変動S_B、配置による変動S_Rを以下で与える。

$$S = \sum_{i=1}^{r}\sum_{j=1}^{r} X_{ij(k)}^2 - C,\ S_P = \frac{1}{r}\sum_{j=1}^{r} T_{\bullet j}^2 - C,\ S_B = \frac{1}{r}\sum_{i=1}^{r} T_{i\bullet}^2 - C,\ S_R = \frac{1}{r}\sum_{k=1}^{r} T_k^2 - C$$

これらの変動から誤差変動S_Eを以下のように定義する。

$$S_E = S - S_P - S_B - S_R$$

水準間の差や個体間の差及び配置による差の検定は、それぞれ以下の性質を利用する。

$\alpha_j = 0$ のとき、 $F_P = \dfrac{S_P/(r-1)}{S_E/(r-1)(r-2)} \sim F_{r-1,(r-1)(r-2)}$ 分布

$\beta_i = 0$ のとき、 $F_B = \dfrac{S_B/(r-1)}{S_E/(r-1)(r-2)} \sim F_{r-1,(r-1)(r-2)}$ 分布

$\gamma_k = 0$ のとき、 $F_R = \dfrac{S_R/(r-1)}{S_E/(r-1)(r-2)} \sim F_{r-1,(r-1)(r-2)}$ 分布

7) 多重比較

1 元比較の場合、1 元配置分散分析も Kruskal-Wallis の順位検定も水準間に差があることは分かってもどこに差があるのか判定することはできない。また、p個の水準から 2 つの水準を選んで 2 群間の差の検定を行なうことはできるが、${}_pC_2$回の検定を行なうことによる有意水準の解釈には問題がある。このような多重比較の場合にどのような検定を行なうかについて、Bonferroni の方法、Tukey の方法、Dunnet の方法等様々な検定方法が考えられてきたが、ここではその中で比較的有効と考えられる結合された（pooled）不偏分散による t 検定及び結合された順位による Wilcoxon の順位和検定をプログラム化した。実際の検定では Fisher の LSD 法を用いて、それぞれ 1 元配置分散分析や Kruskal-Wallis の順位検定と併用する。

結合された不偏分散による t 検定

データは表 1 の形式であり、水準 i のデータ数をn_i、平均を$\bar{x}_i$、不偏分散をs_i^2として、水準 i, j の差について考える。結合された不偏分散s^2は以下のように与えられる。

$$s^2 = \frac{1}{n-p}\sum_{i=1}^{p}(n_i - 1)s_i^2$$

ここに全データ数を n としている。検定には以下の性質を利用する。

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim t_{n-p} \text{ 分布}$$

結合された順位による Wilcoxon の順位和検定

データは上と同様に表 1 の形式であるが、全データの小さい順に順位を付ける。水準 i の順位合計を w_i とし、データ数が十分多いとして以下の性質を利用する。

$$Z_{ij} = \frac{\left|\frac{w_i}{n_i} - \frac{w_j}{n_j}\right| - \frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}{\sqrt{\frac{n(n+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim N(0,1) \text{ 分布}$$

1.2　プログラムの利用法

実験計画法の分析画面を図 2 に示す。

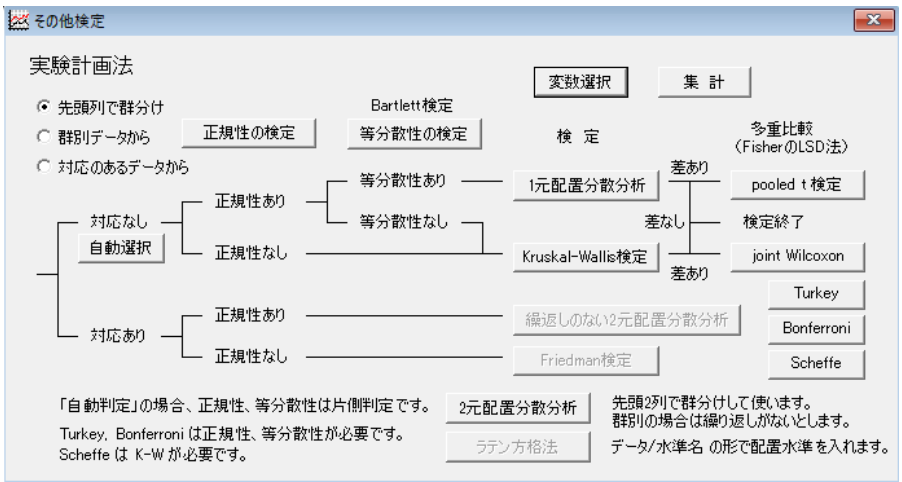


図 2 実験計画法分析画面

画面は基本統計の量的データの検定メニューのように、分析選択手順を図式化したものになっている。データは先頭列で群分けする場合と既に群別になっている場合と 2 通りから選択できる。コマンドボ

タン「集計」は水準毎の基本統計量を出力する。図 3 に「等分散の検定」の出力画面を示す。

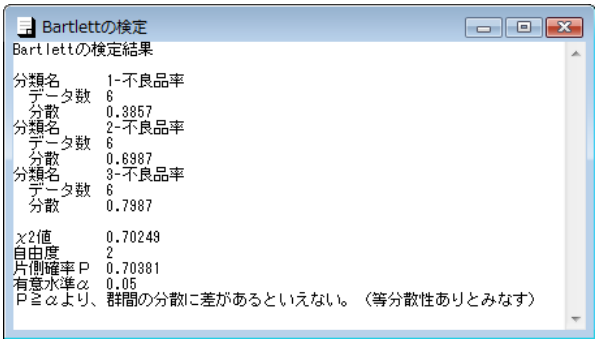


図 3 等分散の検定出力画面

図 4a と図 4b に「１元配置分散分析」の検定結果と分散分析表の出力画面を示す。

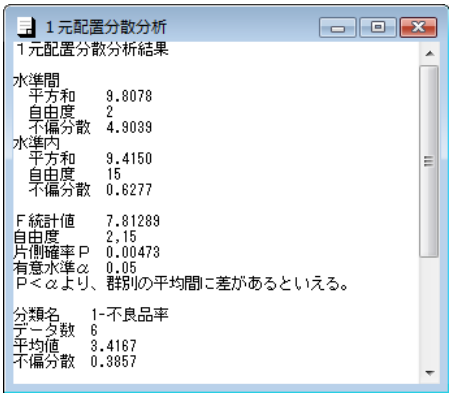


図 4a　１元配置分散分析出力画面

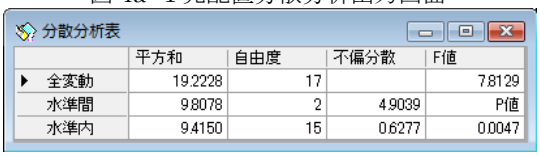


図 4b　１元配置分散分析表

また、図 5 に「Kruskal-Wallis 検定」の検定結果の出力画面を示す。

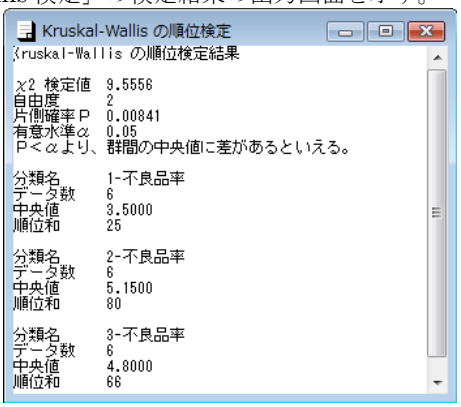


図 5 Kruskal-Wallis 検定出力画面

「繰返しのない 2 元配置分散分析」は、対応のある 1 元配置分散分析とも呼ばれる。「繰り返しのない 2 元配置分散分析」の出力結果と分散分析表をそれぞれ図 6a と図 6b に示す。この場合はブロックと水準の交点に 1 つだけデータがある形式で、群分けされたデータからのみ計算が実行できる。

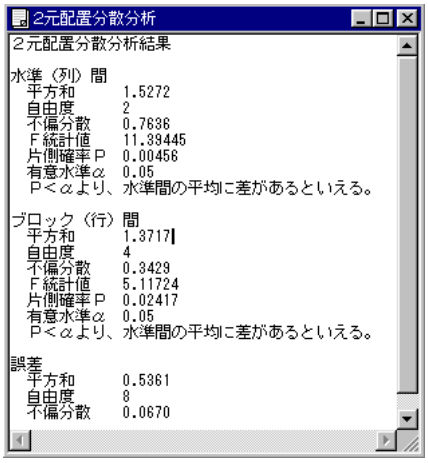
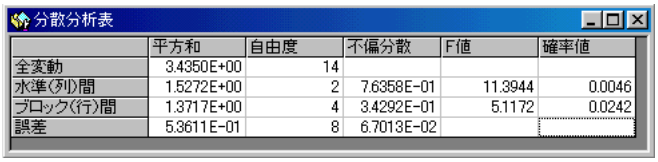


図 6a　２元配置分散分析（繰り返しなし）



	平方和	自由度	不偏分散	F値	確率値
全変動	3.4350E+00	14			
水準(列)間	1.5272E+00	2	7.6358E-01	11.3944	0.0046
ブロック(行)間	1.3717E+00	4	3.4292E-01	5.1172	0.0242
誤差	5.3611E-01	8	6.7013E-02		

図 6b　２元配置分散分析表（繰り返しなし）

対応のある 1 元比較の問題（繰返しのない 2 元比較の問題）で正規性に疑いがある場合やブロック間の平均の差が大きい場合、Friedman 検定を行なう。出力画面を図 7 に示す。

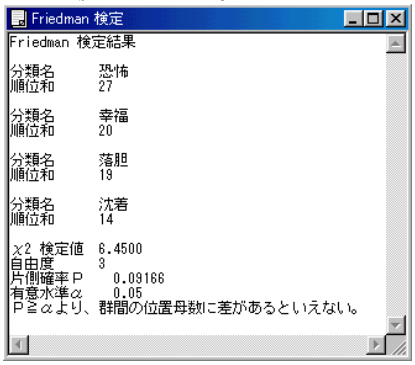


図 7 Friedman 検定出力画面

繰り返しがある場合の「2 元配置分散分析」の出力結果と分散分析表をそれぞれ図 8a と図 8b に示す。この場合、データは先頭 2 列で群分けされたものだけが利用できる。

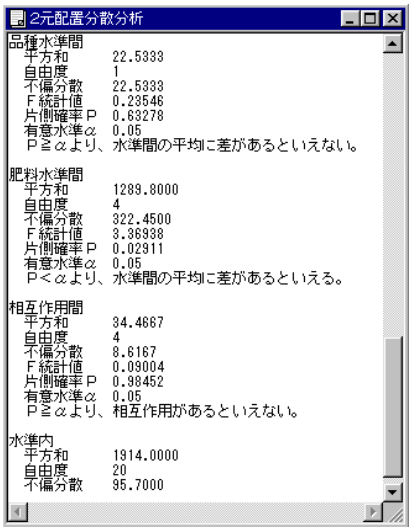
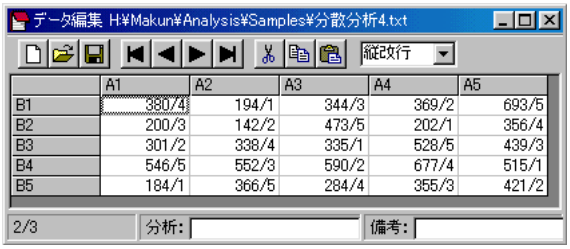


図 8a　2 元配置分散分析（繰り返しあり）

	平方和	自由度	不偏分散	F値	確率値
全変動	3.2608E+03	29			
品種水準間	2.2533E+01	1	2.2533E+01	0.2355	0.6328
肥料水準間	1.2898E+03	4	3.2245E+02	3.3694	0.0291
相互作用間	3.4467E+01	4	8.6167E+00	0.0900	0.9845
水準内	1.9140E+03	20	9.5700E+01		

図 8b　2 元配置分散分析表（繰り返しあり）

データの処理順序の差も検出したい場合、ラテン方格法を利用する。これには処理順序を入力しておく必要があるため、データに加えて順序を「データ/順序」のように / で区切って入力する。このデータ形式の例を図 9 に示す。出力は水準、ブロック、配置間の差を検定した結果を、図 6a と図 6b のようにテキストと分散分析表の 2 種類で表示するが、具体的な画面については省略する。



	A1	A2	A3	A4	A5
B1	380/4	194/1	344/3	369/2	693/5
B2	200/3	142/2	473/5	202/1	356/4
B3	301/2	338/4	335/1	528/5	439/3
B4	546/5	552/3	590/2	677/4	515/1
B5	184/1	366/5	284/4	355/3	421/2

図 9　ラテン方格法データ例

多重比較については、正規性が認められる場合と認められない場合について、結合された不偏分散による t 検定と結合された順位による Wilcoxon の順位和検定の出力結果をそれぞれ図 10 と図 11 に

示す。

Pooled t検定

	工場1	工場2	工場3
データ数	6	6	6
平均	3.4167	5.1333	4.7667
不偏分散	3.8567E-01	6.9867E-01	7.9867E-01
Pooled不偏分散	6.2767E-01		
自由度	15		
確率(両側)			
工場1	1.00000	0.00192	0.00990
工場2	0.00192	1.00000	0.43529
工場3	0.00990	0.43529	1.00000

図 10 pooled t 検定出力結果

Pooled Wilcoxon順位和検定

	工場1	工場2	工場3
データ数	6	6	6
順位和	25.000	80.000	66.000
確率(両側)			
工場1	1.00000	0.00350	0.03055
工場2	0.00350	1.00000	0.48208
工場3	0.03055	0.48208	1.00000

図 11 pooled Wilcoxon 検定出力結果

さらに多重比較について、単独で検定する手法として、比較的よく用いられる Turkey の方法、適用範囲が広い Bonferroni の方法が含まれている。これらの検定には、正規性と等分散性が必要である。またこれらの条件が満たされない場合に用いられる手法として Scheffe の方法が含まれている。これには最初に Kruskal-Wallis の検定が必要である。これらの検定の実行結果を分散分析 1.txt のデータを用いて図 12 から図 14 に与えておく。

Tukey-Kramerの方法

	1-不良品率	2-不良品率	3-不良品率
データ数	6	6	6
平均	3.417	5.133	4.767
不偏分散	0.386	0.699	0.799
Pooled不偏分散	0.628		
自由度	15		
確率範囲(両側)			
1-不良品率	1.0000	p<0.01	p<0.01
2-不良品率	p<0.01	1.0000	n.s.
3-不良品率	p<0.01	n.s.	1.0000

図 12　Turkey の方法検定結果

pooled Bonferroniの方法

	1-不良品率	2-不良品率	3-不良品率
データ数	6	6	6
平均	3.417	5.133	4.767
不偏分散	0.386	0.699	0.799
Pooled不偏分散	0.628		
自由度	15		
確率*回数(両側)			
1-不良品率	1.0000	0.0058	0.0297
2-不良品率	0.0058	1.0000	1.0000
3-不良品率	0.0297	1.0000	1.0000

図 13　Bonferroni の方法検定結果

Scheffeの方法

	1-不良品率	2-不良品率	3-不良品率
データ数	6	6	6
平均	3.417	5.133	4.767
不偏分散	0.386	0.699	0.799
Pooled不偏分散	0.628		
自由度	15		
確率(片側)			
1-不良品率	1.0000	0.0070	0.0323
2-不良品率	0.0070	1.0000	0.7301
3-不良品率	0.0323	0.7301	1.0000

図 14　Scheffe の方法

２．重回帰分析

2.2　重回帰分析の理論

重回帰分析は、目的変数を複数の説明変数の線形回帰式で予測する手法である。データは以下の表1の形式で与えられる。

表 1 重回帰分析のデータ

目的変数	説明変数 1	…	説明変数 p
y_1	x_{11}	…	x_{p1}
y_2	x_{12}	…	x_{p2}
:	:		:
y_n	x_{1n}	…	x_{pn}

実測値は以下のような1次式と正規分布する誤差 ε_λ で与えられるものと考える。

$$y_\lambda = \sum_{i=1}^{p} b_i x_{i\lambda} + b_0 + \varepsilon_\lambda , \qquad \varepsilon_\lambda \sim N(0,\sigma^2)$$ 分布［異なる λ について独立］

線形回帰式は偏回帰係数 b_i，b_0 を用いて、以下の形で与えられる。

$$Y_\lambda = \sum_{i=1}^{p} b_i x_{i\lambda} + b_0$$

これらの偏回帰係数は実測値と予測値のずれの 2 乗和 EV が最小になるように決定される。

$$EV = \sum_{\lambda=1}^{n} (y_\lambda - Y_\lambda)^2 \qquad \text{最小化}$$

即ち、b_i と b_0 についての EV の微係数を 0 とおいて以下の式を得る。

$$b_i = (\mathbf{S}^{-1}\mathbf{S}_y)_i , \quad b_0 = \bar{y} - \sum_{i=1}^{p} b_i \bar{x}_i$$

ここに、$\mathbf{S}^{-1}$ は説明変数の共分散行列 $\mathbf{S}$ の逆行列、$\mathbf{S}_y$ は目的変数と説明変数の分散共分散ベクトルである。

$$(\mathbf{S})_{ij} = \frac{1}{n-1}\sum_{\lambda=1}^{n}(x_{i\lambda} - \bar{x}_i)(x_{j\lambda} - \bar{x}_j) , \quad (\mathbf{S}_y)_i = \frac{1}{n-1}\sum_{\lambda=1}^{n}(y_\lambda - \bar{y})(x_{i\lambda} - \bar{x}_i)$$

偏回帰係数は変数の平均や分散によって影響を受け、係数の重要性が分かりにくいが、データを以下のように標準化して重回帰分析を行なうと変数の影響力の強さがはっきりと示される。ここに s_y^2，s_i^2 は目的変数及び説明変数 i の不偏分散である。

$$\tilde{y}_\lambda = \frac{y_\lambda - \bar{y}}{s_y} , \quad \tilde{x}_{i\lambda} = \frac{x_{i\lambda} - \bar{x}_i}{s_i}$$

これらの新しいデータ $\tilde{y}_\lambda$ と $\tilde{x}_{i\lambda}$ で作った重回帰式の偏回帰係数 $\tilde{b}_i$ を標準化偏回帰係数と言い、回帰

式は以下のように表わされる。

$$\tilde{Y}_\lambda = \sum_{i=1}^{p} \tilde{b}_i \tilde{x}_{i\lambda}$$

標準化偏回帰係数と偏回帰係数との関係は $\tilde{b}_i = b_i \, s_i / s_y$ で与えられる。

重相関係数 R は実測値と予測値の相関係数であり、以下のように与えられる。

$$R = s_{yY} / (s_y s_Y)$$

ここに、s_{yY} は実測値 y と予測値 Y の共分散、s_y^2 と s_Y^2 は実測値と予測値の不偏分散である。

$$s_{yY} = \frac{1}{n-1}\sum_{\lambda=1}^{n}(y_\lambda - \bar{y})(Y_\lambda - \bar{Y}),\ s_y^2 = \frac{1}{n-1}\sum_{\lambda=1}^{n}(y_\lambda - \bar{y})^2,\ s_Y^2 = \frac{1}{n-1}\sum_{\lambda=1}^{n}(Y_\lambda - \bar{Y})^2$$

実測値の全変動 SV は回帰変動 RV と残差変動 EV の和として表わされる。

$$SV = \sum_{\lambda=1}^{n}(y_\lambda - \bar{y})^2 = \sum_{\lambda=1}^{n}(y_\lambda - Y_\lambda)^2 + \sum_{\lambda=1}^{n}(Y_\lambda - \bar{Y})^2 = EV + RV$$

全変動に占める回帰変動の割合は、予測値が実測値を説明する割合を表わしていると考えられ、その値を寄与率という。寄与率は重相関係数の 2 乗に等しいことが示されるので、記号 R^2 で表わすことにする。

$$R^2 = RV / SV$$

寄与率や重相関係数の値は説明変数の数が増えれば大きくなることが知られており、これを緩和するために以下のような自由度調整済み重相関係数 $\bar{R}$ が考えられている。

$$\bar{R} = \sqrt{1 - \frac{EV/(n-p-1)}{SV/(n-1)}}$$

重回帰式の有効性は回帰変動と残差変動を比べて、回帰変動が十分大きいことが重要で、この検定には、以下の性質が利用される。

$$F = \frac{RV/p}{EV/(n-p-1)} \sim F_{p,n-p-1} \text{分布}$$

重回帰式全体の有効性とは別に、それぞれの偏回帰係数の有効性も検討される。これらは偏回帰係数が 0 と異なることを示して確かめられる。この検定には以下の性質が利用される。

$b_i = 0$ の検定　　$t_i = \dfrac{b_i}{\sqrt{a^{ii} EV/(n-p-1)}} \sim t_{n-p-1}$ 分布

$b_0 = 0$ の検定　　$t_0 = \dfrac{b_0}{\sqrt{\left(\dfrac{1}{n} + \sum_{i=1}^{p}\sum_{j=1}^{p} \bar{x}_i \bar{x}_j a^{ij}\right) EV \Big/ (n-p-1)}} \sim t_{n-p-1}$ 分布

ここに a^{ij} は $\mathbf{A} = (n-1)\mathbf{S}$ としたときの行列 $\mathbf{A}$ の逆行列 $\mathbf{A}^{-1}$ の i, j 成分である。

説明変数 i を除く他の説明変数で作った $x_{i\lambda}$ の予測回帰式を以下のように書く。

$$X_{i\lambda} = b_1^{(i)}x_{1\lambda} + \cdots + b_{i-1}^{(i)}x_{i-1\lambda} + b_{i+1}^{(i)}x_{i+1\lambda} + \cdots + b_p^{(i)}x_{p\lambda} + b_0^{(i)}$$

また、説明変数 i を除く他の説明変数で作った目的変数の予測回帰式を以下のように書く。

$$Y_{i\lambda} = b_1'^{(i)}x_{1\lambda} + \cdots + b_{i-1}'^{(i)}x_{i-1\lambda} + b_{i+1}'^{(i)}x_{i+1\lambda} + \cdots + b_p'^{(i)}x_{p\lambda} + b_0'^{(i)}$$

実測値からこれらの予測値を引いた値をそれぞれ $x'_{i\lambda}$， $y'_{i\lambda}$ として、

$$x'_{i\lambda} = x_{i\lambda} - X_{i\lambda}\,,\quad y'_{i\lambda} = y_\lambda - Y_{i\lambda}\,,$$

この $x'_{i\lambda}$ と $y'_{i\lambda}$ の相関係数を偏相関係数と呼び、$\tilde{r}_{iy}$ で表わす。偏相関係数は他の変数の影響を除いた相関係数と見ることができ、以下のように表わすこともできる。

$$\tilde{r}_{iy} = -r^{iy}\Big/\sqrt{r^{ii}r^{yy}}$$

ここに r^{iy}，r^{ii}，r^{yy} は、目的変数と説明変数を合せた相関行列 $\mathbf{R}$ の逆行列 $\mathbf{R}^{-1}$ の成分である。

$$\mathbf{R} = \begin{pmatrix} 1 & r_{y1} & \cdots & r_{yp} \\ r_{1y} & 1 & \cdots & r_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{py} & r_{p1} & \cdots & 1 \end{pmatrix},\quad \mathbf{R}^{-1} = \begin{pmatrix} r^{yy} & r^{y1} & \cdots & r^{yp} \\ r^{1y} & r^{11} & \cdots & r^{1p} \\ \vdots & \vdots & \ddots & \vdots \\ r^{py} & r^{p1} & \cdots & r^{pp} \end{pmatrix}$$

また、モデルの適合度を表すのに、AIC の値が利用されることがあるが、これは以下のように定義される。

$$AIC = n\left(\log(2\pi)+1\right) + n\log\left(EV/n\right) + 2p$$

2.2　プログラムの利用法

具体的な分析画面を図 1、データを図 2 に示す。変数選択で、全てのデータを選択する。

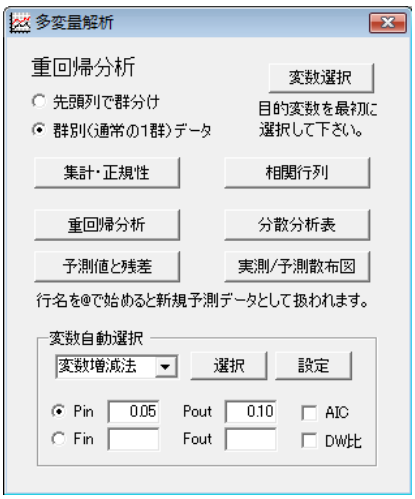


図 1 重回帰分析メニュー画面

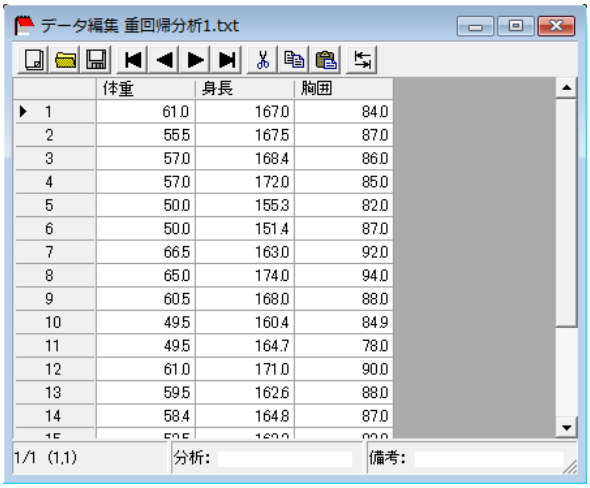


図 2 重回帰分析データ

「相関行列」ボタンでは目的変数と説明変数を含んだ相関行列 **R** が表示される。その際、相関係数を 0 と比較する検定の確率値も表示される。「重回帰分析」ボタンでは、テキスト画面とグリッド画面の２つのウィンドウが開き、図 3a と図 3b の分析結果が表示される。

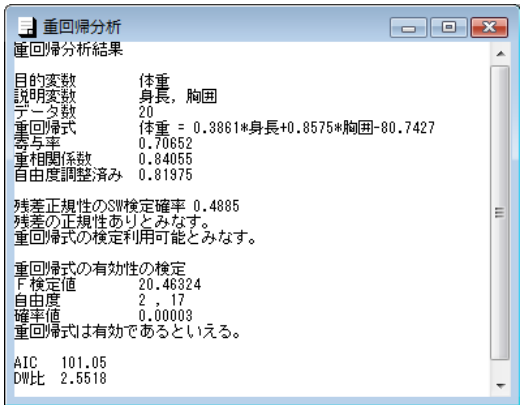


図 3a 重回帰分析出力画面 1

図 3b 重回帰分析出力画面 2

次に、「分散分析表」ボタンをクリックすると、図 4 に示す結果が表示される。

	平方和	自由度	不偏分散	F検定値
全変動	462.4055	19		20.4632
回帰変動	326.7009	2	163.3504	確率値
残差変動	135.7046	17	7.9826	0.0000

図 4 分散分析表画面

「予測値と残差」ボタンでは、図 5 のように各レコード毎の実測値、予測値、残差が示される。

	実測値	予測値	残差
1	61.0	55.762	5.238
2	55.5	58.528	-3.028
3	57.0	58.018	-1.018
4	57.0	58.550	-1.550
5	50.0	49.530	0.470
6	50.0	52.312	-2.312
7	66.5	61.078	5.422
8	65.0	67.040	-2.040
9	60.5	59.579	0.921
10	49.5	53.986	-4.486

図 5 予測値と残差

また、「実測／予測値の散布図」ボタンでは、図 6 のように実測値と予測値の散布図が描かれる。

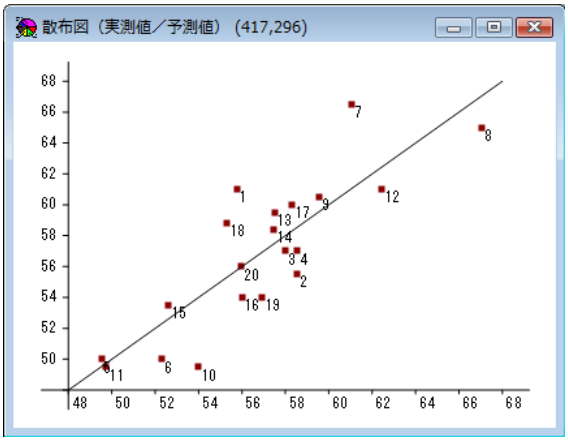
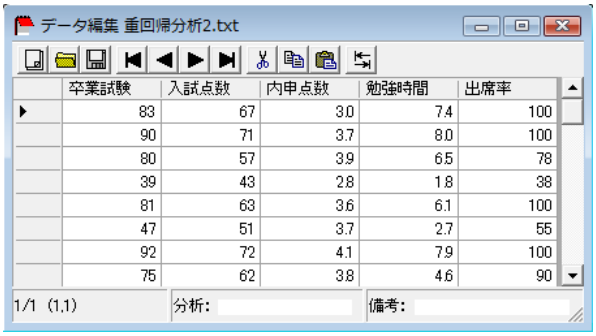


図 6 実測値と予測値の散布図

次に変数の自動選択について、図 7 のデータを用いて説明する。

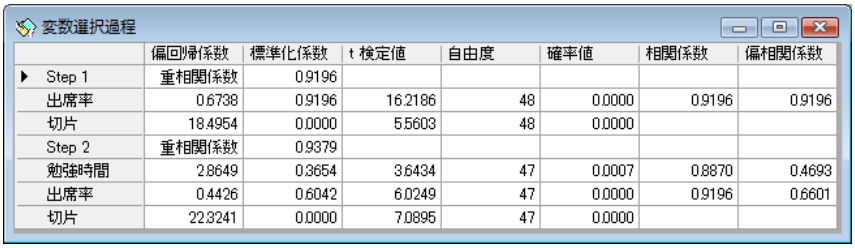


	卒業試験	入試点数	内申点数	勉強時間	出席率
▶	83	67	3.0	7.4	100
	90	71	3.7	8.0	100
	80	57	3.9	6.5	78
	39	43	2.8	1.8	38
	81	63	3.6	6.1	100
	47	51	3.7	2.7	55
	92	72	4.1	7.9	100
	75	62	3.8	4.6	90

図 7 変数自動選択のデータ

最初に全ての変数を選択して分析を実行する。変数の追加と削除の基準は、追加と削除の変数の係数についての検定確率またはＦ検定値のどちらかで与えられる。「Pin」左側のラジオボックスをチェックすると検定確率で指定し「Fin」左側のラジオボックスをチェックするとＦ検定値で指定することになる。デフォルトは検定確率になっている。

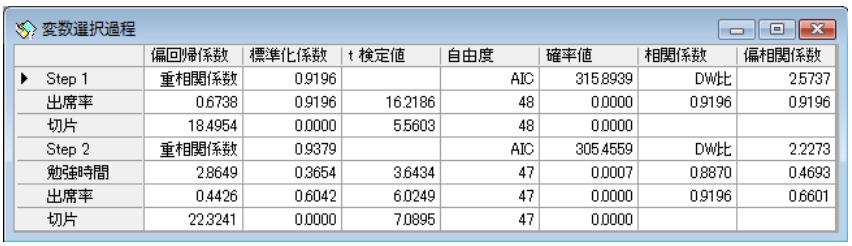
変数の選択法として、変数増加法、変数減少法、変数増減法のどれかを選び、「選択」ボタンをクリックすると図８のように選択過程での種々の統計量が表示される。



	偏回帰係数	標準化係数	t 検定値	自由度	確率値	相関係数	偏相関係数
▶ Step 1	重相関係数	0.9196					
出席率	0.6738	0.9196	16.2186	48	0.0000	0.9196	0.9196
切片	18.4954	0.0000	5.5603	48	0.0000		
Step 2	重相関係数	0.9379					
勉強時間	2.8649	0.3654	3.6434	47	0.0007	0.8870	0.4693
出席率	0.4426	0.6042	6.0249	47	0.0000	0.9196	0.6601
切片	22.3241	0.0000	7.0895	47	0.0000		

図 8 変数選択過程表示画面

この場合は、２段階で変数が２つ選択されている。図１で「AIC」チェックボックスや「DW 比」チェックボックスにチェックを入れると、各過程での AIC の値やダービン・ワトソン比が図８の画面上に図９のように追加して表示される。



	偏回帰係数	標準化係数	t 検定値	自由度	確率値	相関係数	偏相関係数
▶ Step 1	重相関係数	0.9196		AIC	315.8939	DW比	2.5737
出席率	0.6738	0.9196	16.2186	48	0.0000	0.9196	0.9196
切片	18.4954	0.0000	5.5603	48	0.0000		
Step 2	重相関係数	0.9379		AIC	305.4559	DW比	2.2273
勉強時間	2.8649	0.3654	3.6434	47	0.0007	0.8870	0.4693
出席率	0.4426	0.6042	6.0249	47	0.0000	0.9196	0.6601
切片	22.3241	0.0000	7.0895	47	0.0000		

図 9 AIC と DW 比を加えた変数選択過程表示画面

重回帰分析は１つの目的変数を複数の説明変数の線形結合で予測するモデルであるが、データによっては、１つの線形結合として表すのではなく、複数の線形結合の混じり合ったものとして表す方が良い予測結果を与える場合がある。我々はこの問題について、１変数の回帰分析では分類別に回帰分析を行うプログラムを開発していたが、多変数の重回帰分析では今回新たに機能を追加した。ここではこの機能について図 10 の例を用いて説明する。変数選択では、最初に群分け用変数、次に目的変数、続けて説明変数を選択する。ここで群による違いを明確にするために、故意に説明変数は両群同じ値にしている。

図 10 群分けした重回帰分析のデータ

データの形式は図 1 の分析メニューで、「先頭列で群分け」ラジオボタンを選択する。

「相関行列」ボタンをクリックすると、図 11 のように、「群」変数で群分けしたデータ毎の相関行列が表示される。

図 11 群分けした相関行列

また、「重回帰分析」ボタンをクリックすると、図 12a と図 12b のような群分けした結果が表示される。

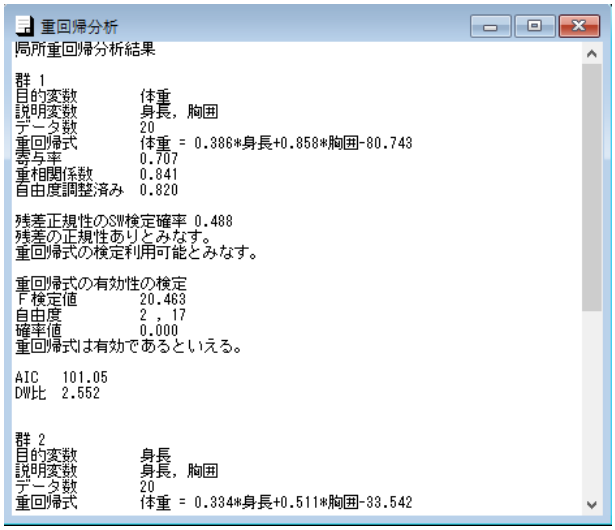
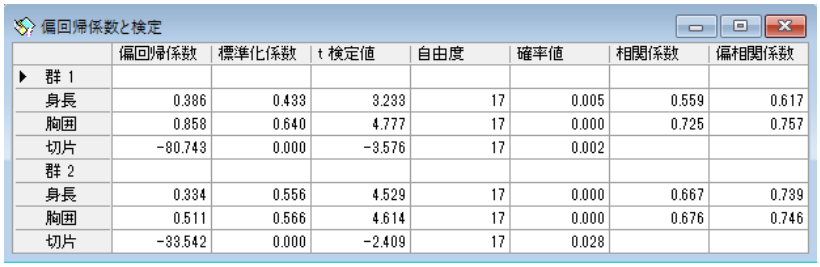


図 12a 群分けした重回帰分析結果１



	偏回帰係数	標準化係数	t 検定値	自由度	確率値	相関係数	偏相関係数
▶ 群 1							
身長	0.386	0.433	3.233	17	0.005	0.559	0.617
胸囲	0.858	0.640	4.777	17	0.000	0.725	0.757
切片	-80.743	0.000	-3.576	17	0.002		
群 2							
身長	0.334	0.556	4.529	17	0.000	0.667	0.739
胸囲	0.511	0.566	4.614	17	0.000	0.676	0.746
切片	-33.542	0.000	-2.409	17	0.028		

図 12b 群分けした重回帰分析結果２

ここで、図 12a の画面下方には、群分けした結果の他に、図 12c のような、全体的な指標も表示される。

図 12c 群分けした重回帰分析結果３

これは、群分けした結果から、予測値を求め、それを元にして全体的な予測の程度を与えたものである。重回帰分析では、実測値と予測値の相関係数（重相関係数）の 2 乗と回帰変動／全変動（寄与率）の結果が一致するが、ここの定義だと異なっている。

「分散分析表」ボタンをクリックすると、図 13 のように、群別に計算された分散分析表が表示される。

分散分析表

	平方和	自由度	不偏分散	F検定値	F確率値
▶ 群 1					
全変動	462.405	19		20.463	0.000
回帰変動	326.701	2	163.350		
残差変動	135.705	17	7.983		
群 2					
全変動	209.598	19		26.015	0.000
回帰変動	157.980	2	78.990		
残差変動	51.618	17	3.036		

図 13 群分けされた分散分析表

「予測値と残差」ボタンをクリックすると、レコード順に、群別に計算された予測値と残差を図 14 のように表示する。

予測値と残差

	群	実測値	予測値	残差
17	1	60.000	58.327	1.673
18	1	58.800	55.291	3.509
19	1	54.000	56.946	-2.946
20	1	56.000	55.978	0.022
21	2	63.300	65.067	-1.767
22	2	67.500	66.766	0.734
23	2	68.300	66.556	1.744
24	2	67.200	67.245	-0.045

図 14 群分けされた予測値と残差結果

「実測／予測散布図」ボタンをクリックすると、図 15 のように、上の予測値を用いたグラフが表示されるが、このグラフの回帰直線は一致しており、重なって表示されている。

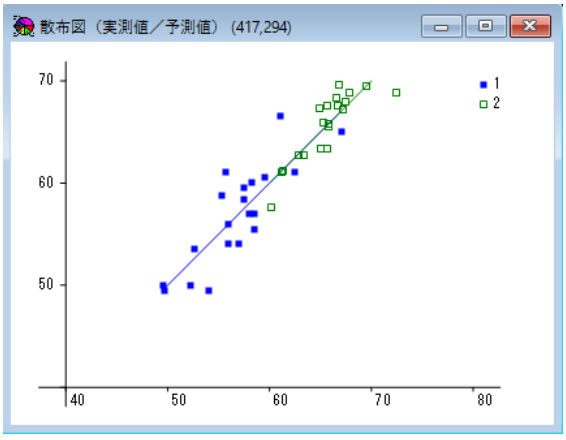


図 15 群分けされた実測値／予測値散布図

３．判別分析

3.1　判別分析の理論

判別分析は外的基準によって群別に分類されたデータから、群を判別するための線形関数を見出すことを目的としている。データは例えば 2 群の場合、表 1 のような形式で与えられる。

表 1 判別分析のデータ（2 群の場合）

群 1			群 2		
変数 1	…	変数 p	変数 1	…	変数 p
x_{11}^1	…	x_{p1}^1	x_{11}^2	…	x_{p1}^2
x_{12}^1	…	x_{p2}^1	x_{12}^2	…	x_{p2}^2
：		：	：		：
$x_{1n_1}^1$	…	$x_{pn_1}^1$	$x_{1n_2}^2$	…	$x_{pn_2}^2$

変数の一般的な表式 $x_{i\lambda}^{\alpha}$ において、α は群、i は変数、λ はレコード番号を表わす。

1）マハラノビス距離を用いた方法

ここでは、最初に 2 群の場合の理論について考える。2 つの群 G_1 と G_2 について、群 $G_1 \cup G_2$ から、G_α（$\alpha = 1,2$）の要素を取り出す確率を P_α とし、G_α の要素を G_β（$\alpha \neq \beta$）と誤判別する損失を $C_{\beta\alpha}$ とする。また、群 α の確率密度関数を $f_\alpha(\mathbf{x})$ とすると、G_α の要素を G_β と誤判別する確率 $Q_{\beta\alpha}$ は以下となる。

$$Q_{\beta\alpha} = \int_{R_\beta} f_\alpha(\mathbf{x}) d\mathbf{x}$$

ここに領域 R_β は、R_β 内の要素を G_β の要素と判別する領域である。これから、誤判別による損失 L は以下のように与えられる。

$$\begin{aligned} L &= C_{21}P_1Q_{21} + C_{12}P_2Q_{12} \\ &= C_{21}P_1\int_{R_2} f_1(\mathbf{x})d\mathbf{x} + C_{12}P_2\int_{R_1} f_2(\mathbf{x})d\mathbf{x} \\ &= C_{21}P_1\int_{R_1\cup R_2} f_1(\mathbf{x})d\mathbf{x} + \int_{R_1}[C_{12}P_2 f_2(\mathbf{x}) - C_{21}P_1 f_1(\mathbf{x})]d\mathbf{x} \end{aligned}$$

これより、損失を最小にするためには R_1 として第 2 項の被積分関数が負になる領域を選べばよい。即ち各群の領域として、以下のような領域を考えれば良いことが分かる。

$$\begin{aligned} R_1 &= \{\mathbf{x} \mid C_{12}P_2 f_2(\mathbf{x}) - C_{21}P_1 f_1(\mathbf{x}) \le 0\}, \\ R_2 &= \{\mathbf{x} \mid C_{12}P_2 f_2(\mathbf{x}) - C_{21}P_1 f_1(\mathbf{x}) > 0\} \end{aligned}$$

これを $h = C_{12}P_2 / C_{21}P_1$ として書き換えて、以下のような条件を得る。

$$R_1 = \{\mathbf{x} \mid \log f_1(\mathbf{x}) / f_2(\mathbf{x}) - \log h \geq 0\},$$
$$R_2 = \{\mathbf{x} \mid \log f_1(\mathbf{x}) / f_2(\mathbf{x}) - \log h < 0\}$$

ここに、判別の分点は 0 である。

今、群 α の変数 i の平均 $\bar{x}_i^{\alpha}$ と各群共通な共分散 s_{ij} をそれぞれ以下のように求め、

$$\bar{x}_i^{\alpha} = \frac{1}{n_\alpha}\sum_{\lambda=1}^{n_\alpha} x_{i\lambda}, \quad s_{ij} = \frac{1}{n_1 + n_2 - 2}\sum_{\alpha=1}^{2}\sum_{\lambda=1}^{n_\alpha}(x_{i\lambda}^{\alpha} - \bar{x}_i^{\alpha})(x_{j\lambda}^{\alpha} - \bar{x}_j^{\alpha}),$$

これらを成分とする平均ベクトル $\bar{\mathbf{x}}^{\alpha}$ と共分散行列 $\mathbf{S}$ を用いて、以下の多変量正規分布の確率密度関数を考える。

$$f_\alpha(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k \mid \mathbf{S} \mid}} \exp\left[-\frac{1}{2}{}^t(\mathbf{x} - \bar{\mathbf{x}}^{\alpha})\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}^{\alpha})\right]$$

これを判別関数に代入して以下の線形判別関数を得る。

$$z = \log f_1(\mathbf{x}) / f_2(\mathbf{x}) - \log h$$
$$= {}^t\mathbf{x}\mathbf{S}^{-1}(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2) - \frac{1}{2}{}^t(\bar{\mathbf{x}}^1 + \bar{\mathbf{x}}^2)\mathbf{S}^{-1}(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2) - \log h$$

$\mathbf{a} = \mathbf{S}^{-1}(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2)$ とすると、判別関数は以下のように書くことができる。

$$z = {}^t\mathbf{x}\mathbf{a} - \frac{1}{2}{}^t(\bar{\mathbf{x}}^1 + \bar{\mathbf{x}}^2)\mathbf{a} - \log h \tag{1}$$

判別関数は、変数 x_i の標準化値 u_i と不偏分散 s_i を用いて以下のように書くこともできる。

$$z = {}^t\mathbf{u}\mathbf{c} + {}^t\bar{\mathbf{x}}\mathbf{a} - \frac{1}{2}{}^t(\bar{\mathbf{x}}^1 + \bar{\mathbf{x}}^2)\mathbf{a} - \log h, \quad c_i = a_i s_i \tag{2}$$

この係数 $\mathbf{c}$ を標準化係数と呼ぶ。標準化係数は変数の重要性をみるときに利用される。

判別関数 (1) は各群の平均 $\bar{\mathbf{x}}^{\alpha}$ から、$\mathbf{x}$ までのマハラノビスの平方距離 $D^{2(\alpha)}$ の差として以下のように定義することもできる。

$$z = \frac{1}{2}(D^{2(2)} - D^{2(1)}) - \log h \ , \ D^{2(\alpha)} = {}^t(\mathbf{x} - \bar{\mathbf{x}}^{\alpha})\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}^{\alpha})$$

この z は $\log h$ が 0 の場合、$\mathbf{x}$ が 2 つの群別平均の中央である $(\bar{\mathbf{x}}^1 + \bar{\mathbf{x}}^2)/2$ のとき、0 になっている。

変数 z の確率分布は、個体 $\mathbf{x}$ が群 1 に属するか、群 2 に属するかに応じて、以下のような正規分布に従うことが知られている。

$z \sim N(D^2/2, D^2)$ $\quad$ $\mathbf{x} \in G_1$ の場合

$z \sim N(-D^2/2, D^2)$ $\quad$ $\mathbf{x} \in G_2$ の場合

ここに、D^2 は群平均 $\bar{\mathbf{x}}^1$ と $\bar{\mathbf{x}}^2$ のマハラノビスの平方距離で、以下のように定義される。

$$D^2 = {}^t(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2)\mathbf{S}^{-1}(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2)$$

この性質から誤判別の理論確率は以下で与えられることが分かる

$$Q_{21} = \int_{-\infty}^{\log h} \frac{1}{\sqrt{2\pi D^2}} \exp\left[-\frac{(z - D^2/2)^2}{2D^2} \right] dz = Z\left(\frac{\log h - D^2/2}{D} \right)$$

$$Q_{12} = \int_{\log h}^{\infty} \frac{1}{\sqrt{2\pi D^2}} \exp\left[-\frac{(z + D^2/2)^2}{2D^2} \right] dz = 1 - Z\left(\frac{\log h + D^2/2}{D} \right)$$

これは判別分析の有効性を示している。

判別分析では、判別関数の係数についてもその有効性を検定できる。変数iの係数が 0 であるかどうかの検定は、以下の性質を利用する。

$$F_i = \frac{(n_1 + n_2 - p - 1) n_1 n_2 (D^2 - D_i^2)}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 D_i^2} \sim F_{1, n_1 + n_2 - p - 1} \text{分布}$$

ここに、D_i^2は両群の変数iを除いたマハラノビスの平方距離である。

以上のような理論では、線形判別関数で表わされる判別分析がうまく利用できる条件は、分布が多変量正規分布に従うことに加えて 2 群の共分散が等しいことである。この検定には以下の性質が利用される。

$$\chi^2 = \left[1 - \left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} - \frac{1}{n_1 + n_2 - 2} \right) \frac{2p^2 + 3p - 1}{6(p+1)} \right] \log \frac{|\mathbf{S}|^{n_1 + n_2 - 2}}{|\mathbf{S}^1|^{n_1 - 1} |\mathbf{S}^2|^{n_2 - 1}} \sim \chi^2_{p(p+1)/2} \text{分布}$$

ここに、$\mathbf{S}^\alpha$は群αの共分散行列である。しかし、後に述べるような正準形式では、2 群の場合、分布の形を仮定することなく同等な結論を導く。

3 群以上（群の数をm）の判別には以下の判別関数を考え、z^αが最大になる群αに属するものと判定する。

$$z^\alpha = {}^t\mathbf{x}\mathbf{S}^{-1}\bar{\mathbf{x}}^\alpha - \frac{1}{2}{}^t\bar{\mathbf{x}}^\alpha \mathbf{S}^{-1}\bar{\mathbf{x}}^\alpha + \log C_\alpha P_\alpha m$$

但し、C_αは群αを他の群と間違えた場合の損失である。定数項に含まれるmは、各群の生起確率が同じで誤判別損失が 1 の場合、これらを考えない理論と繋がるように、定数項を 0 にするための定数である。

$\mathbf{a}^\alpha = \mathbf{S}^{-1}\bar{\mathbf{x}}^\alpha$として、この判別関数は以下のように書くこともできる。

$$z^\alpha = {}^t\mathbf{x}\mathbf{a}^\alpha - \frac{1}{2}{}^t\bar{\mathbf{x}}^\alpha \mathbf{a}^\alpha + \log C_\alpha P_\alpha m \qquad (3)$$

2 群の場合と同様に、判別関数は変数x_iの標準化値u_iと不偏分散s_iを用いて以下のように書くこともできる。

$$z^{\alpha} = {}^{t}\mathbf{u}\mathbf{c}^{\alpha} + {}^{t}\bar{\mathbf{x}}\mathbf{a}^{\alpha} - \frac{1}{2}{}^{t}\bar{\mathbf{x}}^{\alpha}\mathbf{a}^{\alpha} + \log C_{\alpha}P_{\alpha}m, \qquad c_i^{\alpha} = a_i^{\alpha}s_i \tag{4}$$

この係数 $\mathbf{c}^{\alpha}$ を標準化係数と呼ぶ。

上で与えた 2 群の場合の判別関数は、この判別関数を用いて $z = z^1 - z^2$ として求めることができる。

2) 正準形式を用いた方法

正準形式の判別分析（正準判別分析と呼ばれる）は、判別関数の拡がりを最大化するように係数を求めるもので、特に 3 群以上の場合は、判別得点を複数次元の空間上に配置し、判別をより分かり易く表現する手法である。これまでのプログラムでは、数量化Ⅱ類でその中の主要な 1 次元を取り出して判別する方法を導入している。以下に正準判別分析の理論を示す。

正準判別分析は、判別群で分けられたデータについて、「群間分散／群内分散」を最大化するように線形判別関数の係数を決定する手法である。判別関数を以下のように表す。ここに z_0 は後に決める定数項である。

$$z = \sum_{i=1}^{p} a_i x_i + z_0$$

判別群を α，群別のデータの番号を λ，変数の番号を i，としてデータを $x_{i\lambda}^{\alpha}$（$\alpha = 1,\cdots,m$，$\lambda = 1,\cdots,n_{\alpha}$，$i = 1,\cdots,p$）と表す。このデータを用いて、群 α の λ 番目の判別関数の値 z_{λ}^{α} は以下のようになる。

$$z_{\lambda}^{\alpha} = \sum_{i=1}^{p} a_i x_{i\lambda}^{\alpha} + z_0$$

この z_{λ}^{α} による群間分散 s_B^2，群内分散 s^2 を以下のように定義する。

$$s_B^2 = \frac{1}{n-m}\sum_{\alpha=1}^{m} n_{\alpha}\left(\bar{z}^{\alpha} - \bar{z}\right)^2, \quad s^2 = \frac{1}{n-m}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_{\alpha}}\left(z_{\lambda}^{\alpha} - \bar{z}^{\alpha}\right)^2$$

ここに、$\bar{z}^{\alpha} = \frac{1}{n_{\alpha}}\sum_{\lambda=1}^{n_{\alpha}} z_{\lambda}^{\alpha}$，$\bar{z} = \frac{1}{n}\sum_{\alpha=1}^{m} n_{\alpha}\bar{z}^{\alpha}$，$n = \sum_{\alpha=1}^{m} n_{\alpha}$ である。

これより、$\bar{x}_i^{\alpha} = \frac{1}{n_{\alpha}}\sum_{\lambda=1}^{n_{\alpha}} x_{i\lambda}^{\alpha}$，$\bar{x}_i = \frac{1}{n}\sum_{\alpha=1}^{m} n_{\alpha}\bar{x}_i^{\alpha}$ として、s_B^2 と s^2 は以下のようになる。

$$s_B^2 = \frac{1}{n-m}\sum_{\alpha=1}^{m} n_{\alpha}\left[\sum_{i=1}^{p} a_i\left(\bar{x}_i^{\alpha} - \bar{x}_i\right)\right]^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} a_i b_{ij} a_j$$

$$s^2 = \frac{1}{n-m}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_{\alpha}}\left[\sum_{i=1}^{p} a_i\left(x_{i\lambda}^{\alpha} - \bar{x}^{\alpha}\right)\right]^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} a_i s_{ij} a_j$$

ここに、

$$b_{ij} = \frac{1}{n-m}\sum_{\alpha=1}^{m} n_\alpha \left(\overline{x}_i^\alpha - \overline{x}_i\right)\left(\overline{x}_j^\alpha - \overline{x}_j\right)$$

$$s_{ij} = \frac{1}{n-m}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha}\left(x_{i\lambda}^\alpha - \overline{x}_i^\alpha\right)\left(x_{j\lambda}^\alpha - \overline{x}_j^\alpha\right)$$

である。行列の成分として、$\left(\mathbf{B}\right)_{ij} = b_{ij}$ ， $\left(\mathbf{S}\right)_{ij} = s_{ij}$ ， $\left(\mathbf{a}\right)_i = a_i$ とすると、s_B^2 と s^2 はこれらの行列を用いて次のように書ける。

$$s_B^2 = {}^t\mathbf{aBa} \quad , \quad s^2 = {}^t\mathbf{aSa}$$

ここに、$n \geq m$ の場合、一般に $rank\left(\mathbf{B}\right) = m-1$，$rank\left(\mathbf{S}\right) = n-m$ である。

群間分散を群内分散で割った分散比 ρ は以下のようになる。

$$\rho = s_B^2 / s^2 = {}^t\mathbf{aBa} / {}^t\mathbf{aSa}$$

この分散比を最大化するには、以下の解を求める。

$$\partial\rho / \partial\mathbf{a} = \frac{1}{\left(s^2\right)^2}\left[\partial s_B^2 / \partial\mathbf{a}\, s^2 - s_B^2\, \partial s^2 / \partial\mathbf{a}\right] = \mathbf{0}$$

$\partial s_B^2 / \partial\mathbf{a} = 2\mathbf{Ba}$ ， $\partial s_W^2 / \partial\mathbf{a} = 2\mathbf{Sa}$ であるので、上の式は以下となる。

$$\mathbf{Ba} = \rho\mathbf{Sa} \tag{5}$$

これを対称行列の固有方程式にするために、適当な下三角行列 $\mathbf{F}$ を用いて対称行列 $\mathbf{S}$ を $\mathbf{S} = \mathbf{F}\,{}^t\mathbf{F}$ のように書いて、上式を以下のようにする。

$$\mathbf{F}^{-1}\mathbf{B}\,{}^t\mathbf{F}^{-1}\,{}^t\mathbf{Fa} = \rho\,{}^t\mathbf{Fa}$$

ここで $\mathbf{A} = \mathbf{F}^{-1}\mathbf{B}\,{}^t\mathbf{F}^{-1}$ ， $\mathbf{u} = {}^t\mathbf{Fa}$ （$\mathbf{a} = {}^t\mathbf{F}^{-1}\mathbf{u}$）とすると、上式は以下のような対称行列の固有方程式となる。

$$\mathbf{Au} = \rho\mathbf{u} \tag{6}$$

${}^t\mathbf{uu} = 1$ の規格化条件を付けて r 番目の固有値 $\rho^{(r)}$ について方程式を解いた答えを、$\mathbf{u}^{(r)}$ とすると、正準判別関数の係数は以下で与えられる。

$$\mathbf{a}^{(r)} = {}^t\mathbf{F}^{-1}\mathbf{u}^{(r)}$$

以上より、第 r 番目の固有値に対応する判別関数 $z^{(r)}$ は以下のようになる。

$$z^{(r)} = {}^t\mathbf{xa}^{(r)} - {}^t\tilde{\mathbf{x}}^\alpha\mathbf{a}^{(r)} \tag{7}$$

ここに $\tilde{\mathbf{x}}^\alpha = \frac{1}{m}\sum_{\alpha=1}^{m}\overline{\mathbf{x}}^\alpha$ である。定数項については、後に述べる 2 群の場合のマハラノビス形式と正準形式の同一性から、各固有ベクトルに対応する判別関数の群別平均の単純平均が 0 になるように決めた。

マハラノビス形式と同様、変数 x_i の標準化値 u_i と不偏分散 s_i を用いて判別関数は以下のように書くこともできる。

$$z^{(r)} = {}^t\mathbf{u}\mathbf{c}^{(r)} + {}^t\bar{\mathbf{x}}\mathbf{a}^{(r)} - {}^t\tilde{\mathbf{x}}\mathbf{a}^{(r)}, \qquad c_i^{(r)} = a_i^{(r)} s_i \tag{8}$$

この係数 $\mathbf{c}^{(r)}$ を標準化係数と呼ぶ。

(6) 式から、

$$\rho^{(r)} = {}^t\mathbf{u}^{(r)}\mathbf{A}\mathbf{u}^{(r)} = {}^t\mathbf{u}^{(r)}\mathbf{F}^{-1}\mathbf{B}\,{}^t\mathbf{F}^{-1}\mathbf{u}^{(r)} = {}^t\mathbf{a}^{(r)}\mathbf{B}\mathbf{a}^{(r)} = s_B^{(r)2}$$

となり、r 番目の固有値は群間分散の第 r 成分に等しくなる。この性質を用いて、r 番目の固有値に対する変動の寄与率 $P^{(r)}$ を以下で与える。

$$P^{(r)} = \rho^{(r)} \Big/ \sum_{k=1}^{m-1} \rho^{(k)}$$

3) 2 群におけるマハラノビスの形式と正準形式の同等性

さて、ここで述べてきた従来の理論とマハラノビスの距離を用いた判別分析とはどのような関係にあるのだろうか。(5)式について再考する。ここに方程式を再度挙げておく。

$$\mathbf{B}\mathbf{a} = \rho\mathbf{S}\mathbf{a}$$

行列 $\mathbf{B}$ は成分を用いて書くと以下のように表される。

$$\begin{aligned} b_{ij} &= \frac{1}{n-m}\sum_{\alpha=1}^{m} n_\alpha \left(\bar{x}_i^\alpha - \bar{x}_i\right)\left(\bar{x}_j^\alpha - \bar{x}_j\right) \\ &= \frac{1}{n(n-m)}\sum_{\alpha=1}^{m}\sum_{\beta=1}^{m} n_\alpha n_\beta \left(\bar{x}_i^\alpha \bar{x}_j^\alpha - \bar{x}_i^\alpha \bar{x}_j^\beta\right) \\ &= \frac{1}{2n(n-m)}\sum_{\alpha=1}^{m}\sum_{\beta=1}^{m} n_\alpha n_\beta \left(\bar{x}_i^\alpha - \bar{x}_i^\beta\right)\left(\bar{x}_j^\alpha - \bar{x}_j^\beta\right) \end{aligned}$$

これより、$(\mathbf{S}_B\mathbf{a})_{ij}$ は以下のように書ける。

$$\begin{aligned} (\mathbf{S}_B\mathbf{a})_i &= \frac{1}{2n(n-m)}\sum_{\alpha=1}^{m}\sum_{\beta=1}^{m}\sum_{j=1}^{p} n_\alpha n_\beta \left(\bar{x}_i^\alpha - \bar{x}_i^\beta\right)\left(\bar{x}_j^\alpha - \bar{x}_j^\beta\right) a_j \\ &= \sum_{\alpha=1}^{m}\sum_{\beta=1}^{m} c_{\alpha\beta}\left(\bar{x}_i^\alpha - \bar{x}_i^\beta\right) \\ c_{\alpha\beta} &= \frac{n_\alpha n_\beta}{2n(n-m)}\sum_{j=1}^{p}\left(\bar{x}_j^\alpha - \bar{x}_j^\beta\right) a_j \end{aligned}$$

特に 2 群の判別の場合、方程式(5)は以下となる。

$$\rho\mathbf{S}\mathbf{a} = \mathbf{S}_B\mathbf{a} = c(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2)$$

$$c = 2c_{12} = -2c_{21} = \frac{n_1 n_2}{n(n-2)} \sum_{j=1}^{p} \left(\bar{x}_j^1 - \bar{x}_j^2 \right) a_j$$

これより、解 $\mathbf{a}$ を求めると以下となる。

$$\mathbf{a} = \frac{c}{\rho} \mathbf{S}^{-1} (\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2)$$

これは、(1)式で与えられたマハラノビス形式の判別関数の係数の定数倍である。よって、判別の分点を 0 にするような判別関数は以下となる。

$$z = \frac{c}{\rho} {}^t\mathbf{x} \mathbf{S}^{-1} (\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2) - \frac{c}{2\rho} {}^t(\bar{\mathbf{x}}^1 + \bar{\mathbf{x}}^2) \mathbf{S}^{-1} (\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2)$$

これは、判別関数全体が定数倍となっただけで、判別結果は $-\log h$ の項を除いて同等である。

3.2　プログラムの利用法

メニュー［分析－多変量解析等－判別分析］をクリックすると、図 1 のような判別分析実行画面が表示される。

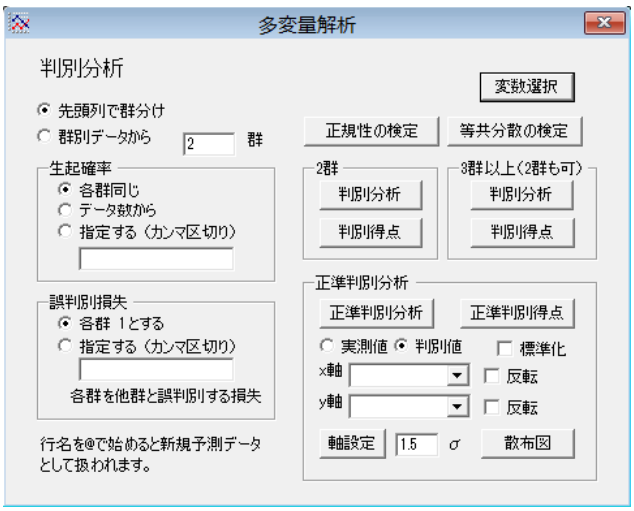


図　1 判別分析画面

データの形式は、先頭列で群分けする場合と最初から群分けされている場合が扱える。但し、後者の場合、予め群の数を入力しておかなければならない。各群の生起確率や誤判別損失の値は、オプションボタンの「指定する」を選び、テキストボックス内に値をカンマ区切りで入力することによって、自由に設定することができる。但し、確率の値は合計が 1 になることが必要であるので、無限小数の場合は 1/3 のように、分数で入力する。これらのデフォルト値は生起確率が「各群同じ」、誤判別損

失が「各群 1 とする」である。

2 群の判別の場合、「等共分散の検定」ボタンで等共分散性を調べることができる。図 2 に「等共分散の検定」の出力結果を示す。図 3 と図 4 に 2 群の判別分析と判別得点の出力結果を示す。判定は判別得点を判別の分点 0 と比較して決定される。

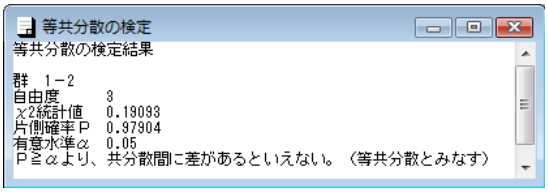


図 2　等共分散の検定

	勉強時間	平均点	定数項
判別関数	2.2461	0.2007	-23.0187
標準化係数	2.6210	2.2787	-0.3788
F検定値	19.8822	15.0274	
自由度	1,27	1,27	
確率	0.0001	0.0006	
マハラノビスの距離	5.6823		
誤判別確率	1群を2群と	2群を1群と	
理論から	0.1167	0.1167	
実測から	0.0769	0.0588	
判別数（実＼予）	1群	2群	
1群	12	1	
2群	1	16	
判別確率(実＼予)	1群	2群	
1群	0.9231	0.0769	
2群	0.0588	0.9412	

図 3　判別分析実行画面（2 群の形式）

標準化係数の定数項は、重回帰分析などでは 0 になるが、判別分析では、判別の分点を 2 つの群の群別平均のデータ数による加重平均ではなく、単純平均にしていることから、2 つの群のデータ数が異なる場合、一般に 0 にならない。

	所属群	判別得点	判別群
6	1	0.3280	1
7	1	-0.7743	2
8	1	4.9054	1
9	1	1.3153	1
10	1	1.8934	1
11	1	1.0704	1
12	1	4.0450	1
13	1	2.2301	1
14	2	-4.8682	2
15	2	-0.0469	2
16	2	-0.9540	2
17	2	-2.1784	2

図 4　判別得点（2 群の形式）

比較のために同じデータを用いて 3 群以上の判別のプログラムを実行した出力結果を図 5 と図 6 に示す。本来は 3 群以上で利用すべきであるが、2 群の判別で用いても問題はない。

判別分析

	勉強時間	平均点	定数項
1群判別関数	8.7369	1.0833	-61.8513
2群判別関数	6.4908	0.8826	-38.8327
1群標準化係数	10.1951	12.2975	47.1974
2群標準化係数	7.5741	10.0189	47.5762
マハラノビスの距離	1群	2群	
1群	0.0000	5.6823	
2群	5.6823	0.0000	
誤判別確率	1群を他群と	2群を他群と	
実測から	0.0769	0.0588	
判別数 (実＼予)	1群	2群	
1群	12	1	
2群	1	16	
判別確率(実＼予)	1群	2群	
1群	0.9231	0.0769	
2群	0.0588	0.9412	

図 5　判別分析実行画面（3 群以上の形式）

判別得点：既存データの判別

	所属群	1群	2群	判別群
6	1	53.3136	52.9857	1
7	1	43.6412	44.4156	2
8	1	72.6009	67.6956	1
9	1	58.3039	56.9886	1
10	1	54.6531	52.7597	1
11	1	50.2115	49.1412	1
12	1	65.9266	61.8816	1
13	1	62.2250	59.9949	1
14	2	23.2397	28.1079	2
15	2	48.9213	48.9682	2
16	2	44.3643	45.3183	2
17	2	37.7561	39.9345	2

図 6　判別得点（3 群以上の形式）

次に我々は正準形式に基づく判別の結果を示す。これは正準判別分析とも呼ばれている。正準相関分析における判別関数は、変数の数≧分割数、の場合は、分割数－1 個作られる。同じデータを用いた結果を図 7 に示す。

正準判別分析

	勉強時間	平均点	定数項
判別1	0.9423	0.0842	-9.6565
標準化1	1.0995	0.9559	-0.1589
	固有値	寄与率	累積寄与率
判別1	1.4950	1.0000	1.0000
判別の分点	0		
	1群を他群と	2群を他群と	
誤判別確率	0.0769	0.0588	

図 7　正準相関分析

生起確率が同じで誤判別損失が 1 の場合、2 群のハラノビス形式と正準形式の同等性から、判別関数の係数は比例している。また、判別の分点は 2 つの形式とも 0 に設定している。

正準判別分析の判別得点では、図 8 のように最後に群別得点平均が付く。これは 3 群以上の場合でも同様である。

判別得点

	所属群	判別得点1	判別群
25	2	-1.8352	2
26	2	-2.3991	2
27	2	-2.4203	2
28	2	-1.8778	2
29	2	-0.4873	2
30	2	-2.0510	2
群別得点平均	1	1.1919	
	2	-1.1919	

図 8　正準判別分析の判別得点

次に 3 群以上の正準判別分析の結果を図 9 に示す。

正準判別分析

	がくの長さ	がくの幅	花弁の長さ	花弁の幅	定数項
判別1	0.8294	1.5345	-2.2012	-2.8105	2.1051
判別2	0.0241	2.1645	-0.9319	2.8392	-6.6615
標準化1	0.6868	0.6688	-3.8858	-2.1422	0.0000
標準化2	0.0200	0.9434	-1.6451	2.1641	0.0000
	固有値	寄与率	累積寄与率		
判別1	32.1919	0.9912	0.9912		
判別2	0.2854	0.0088	1.0000		

図 9　正準判別分析結果

ここでは標準化係数が 0 になっているが、これは 3 つの群のデータ数がすべて同じであることによる偶然で、一般には 0 と異なる。3 群の判別得点は 2 つの固有値に対応して図 10 のように 2 種類出力される。

判別得点

	所属群	判別得点1	判別得点2
1	1	8.0618	0.3004
2	1	7.1287	-0.7867
3	1	7.4898	-0.2654
4	1	6.8132	-0.6706
5	1	8.1323	0.5145
6	1	7.7019	1.4617
7	1	7.2126	0.3558
8	1	7.6053	-0.0116
9	1	6.5606	-1.0152
10	1	7.3431	-0.9473

図 10　正準判別分析の判別得点

これは 2 次元上の点であるので、「軸設定」を行い、「散布図」ボタンをクリックすることにより、図 11 のような散布図が表示される。

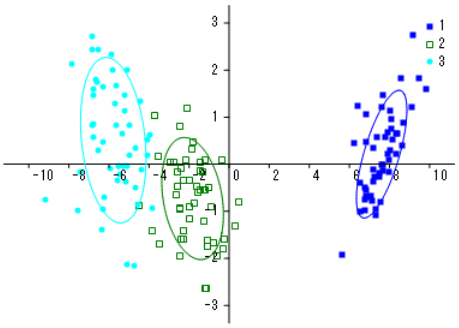


図 11　判別得点散布図

ここには、各群の分布を 2 変量正規分布とみなした場合の、1.5σ の確率楕円が示されている。確率楕円の大きさ、軸の向き等はメニューで変更できる。

この 2 変量正規分布の密度関数式は、グラフメニュー「設定－正規楕円半径－密度関数数式」で図 12 のように表示される。

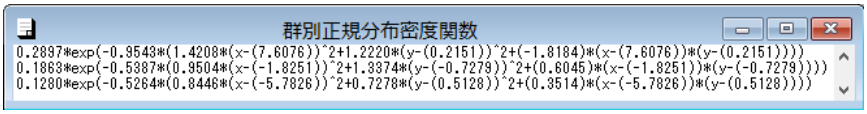


図 12　2 変量正規分布密度関数式

この式をコピーし、分析メニュー「数学－2 変量関数グラフ」のテキストボックスに貼り付けて（[Shift+Ins] または [Ctrl+v]）、（範囲を設定、分割数を増加、色を指定に）表示させると、図 13 のように 3 つの密度関数グラフを重ね合わせて視覚化することもできる。これによってどの程度分離ができているのか直感的に見ることもできる。

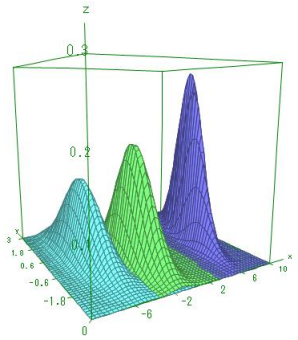


図 13　確率密度関数の視覚化

謝辞

正準判別分析とその表示方法については、岩村忠昭氏にいろいろと助言をいただきました。有難うございました。

４．主成分分析

4.1　主成分分析の理論

主成分分析は、変数の１次結合により、新しい意味付けのできる特徴的な変数を作り出すことを目的としている。この新しい変数を主成分と呼ぶ。主成分分析のデータ形式は表１で与えられる。

表１ 主成分分析のデータ

変数１	変数２	…	変数 p
x_{11}	x_{21}	…	x_{p1}
x_{12}	x_{22}	…	x_{p2}
:	:	…	:
x_{1n}	x_{2n}	…	x_{pn}

我々は新しい変数として以下の１次式を考える。

$$y_\lambda = \sum_{i=1}^{p} u_i x_{i\lambda}$$

特徴的な変数とは、データの変化に最も敏感であることと考え、係数 u_i は変数 y の不偏分散 s^2 が最大になるように求める。但し、スケールの自由度を無くすため係数に ${}^t\mathbf{uu}=1$ の制約を付ける。ここに $\mathbf{u}$ は成分が u_i の縦ベクトルである。

不偏分散 s^2 は係数ベクトル $\mathbf{u}$ と共分散行列 $\mathbf{S}$ を用いて以下のように与えられる。

$$s^2 = \frac{1}{n-1}\sum_{\lambda=1}^{n}(y_\lambda - \bar{y})^2 = {}^t\mathbf{uSu}\,, \qquad (\mathbf{S})_{ij} = \frac{1}{n-1}\sum_{\lambda=1}^{n}(x_{i\lambda} - \bar{x}_i)(x_{j\lambda} - \bar{x}_j)$$

この制約付き最大化問題は、Lagrange の未定定数法を用いて以下の量 L の極値問題となり、解は行列 $\mathbf{S}$ の固有方程式で与えられる。

$$L = {}^t\mathbf{uSu} - \lambda({}^t\mathbf{uu} - 1) \qquad \rightarrow \qquad \mathbf{Su} = \lambda\mathbf{u}$$

この最大固有値に対する固有ベクトル $\mathbf{u}$ を用いて作られた変数 y を第 1 主成分といい、順次固有値の大きい方から第 2 主成分、第 3 主成分と呼ぶ。一般に p 変数の場合、第 p 主成分まで選ぶことができる。

係数 u_i は変数の平均や分散から影響を受けるので、変数を標準化して分析を実行する場合も多い。この場合固有方程式は相関行列 $\mathbf{R}$ を用いて上と同様に与えられる。

$$\mathbf{Ru} = \lambda\mathbf{u}$$

正規化された固有ベクトルを求めることは、線形変換における座標回転の角度を決めることを意味する。即ち、主成分分析は、座標回転によって最も分散の大きな主軸を選び、さらにその主軸に直交し、分散が最大になるような軸を次々と定めてゆく方法である。

これらの固有方程式の第 a 固有値 λ_a に対する固有ベクトル $\mathbf{u}^a$ の成分を以下のように表わす。

$$ {}^t\mathbf{u}^a = (u_1^a \quad u_2^a \quad \cdots \quad u_p^a) $$

固有値 λ_a は第 a 主成分の分散を表わすことが知られている。このことから、全分散 s^2 に対する第 a 主成分の分散の割合 c_a は以下で与えられ、寄与率と呼ばれる。

$$ c_a = \lambda_a \Big/ \sum_{i=1}^{p} \lambda_i $$

因子負荷量 r_{ai} は第 a 主成分と変数 i の相関係数として与えられるが、これは共分散行列と相関行列を元にした場合に分けて、それぞれ以下のような形に表わされる。

$$ r_{ai} = \frac{\sqrt{\lambda_a}\, u_i^a}{s_i} \quad \text{（共分散行列から）}, \qquad r_{ai} = \sqrt{\lambda_a}\, u_i^a \quad \text{（相関行列から）} $$

ここで s_i^2 は変数 i の不偏分散である。

主成分得点 y_λ^a は個体毎の第 a 主成分の値として以下のように定義される。

$$ y_\lambda^a = \sum_{i=1}^{p} u_i^a x_{i\lambda} $$

主成分分析において主成分を区別するためには、その固有値の大きさに差がなければならない。そこで固有値を $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$ とした場合、大きいほうから r 個だけ値が異なり、残りは $\lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_p$ となるかどうかの Anderson による sphericity の検定を行なう。この検定には以下の性質が利用される。

$$ \chi^2 = -n \sum_{a=r+1}^{p} \log \lambda_a + n(p-r) \log \left(\sum_{a=r+1}^{p} \lambda_a \Big/ (p-r) \right) \sim \chi^2_{(p-r-1)(p-r+2)/2} \text{ 分布} $$

4.2 プログラムの利用法

実際の主成分分析のメニュー画面を図 1 に与える。主成分分析は、表 1 に与えたデータの形から実行する場合に加え、それを集計した共分散行列や相関行列から実行する場合も想定される。それ故データの形式としてこれら 3 つの場合が含まれている。等固有値の検定にはデータ数も必要になることから、集計結果からの計算ではデータ数を入力する必要もある。計算を実行するモデルには、通常のデータから計算する「共分散行列から」と標準化されたデータから計算する「相関行列から」の 2 種類がある。勿論、データ形式で相関行列を選んだ場合は共分散行列からの計算はできない。

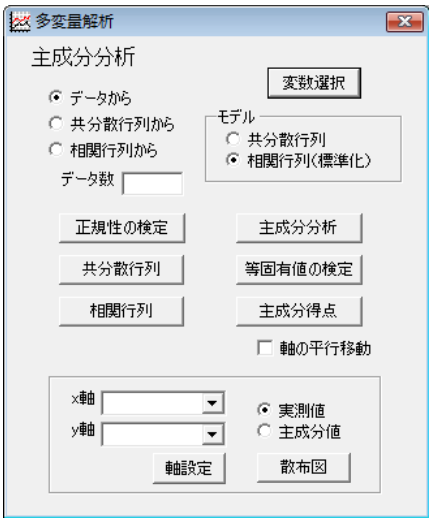


図 1 主成分分析のメニュー

計算結果の表示としては「共分散行列」や「相関行列」も必要と思われるので加えてある。主成分分析は「主成分分析」ボタンで実行され、出力例は、図 2 に示される。

主成分分析

	主成分1	主成分2	主成分3	主成分4
固有値	3.5411	0.3134	0.0794	0.0661
寄与率	0.8853	0.0783	0.0199	0.0165
累積寄与率	0.8853	0.9636	0.9835	1.0000
固有ベクトル				
身長	0.4970	-0.5432	0.4496	0.5057
体重	0.5146	0.2102	0.4623	-0.6908
胸囲	0.4809	0.7246	-0.1752	0.4615
座高	0.5069	-0.3683	-0.7439	-0.2323
因子負荷量				
身長	0.9352	-0.3041	0.1267	0.1300
体重	0.9683	0.1177	0.1303	-0.1776
胸囲	0.9049	0.4056	-0.0494	0.1187
座高	0.9539	-0.2062	-0.2096	-0.0597

図 2 主成分分析出力結果

等固有値の検定結果は図 3 に示される。

等固有値の検定 (by Anderson)

利用主成分	第1主成分	第2主成分	第3主成分
χ2値	67.0395	10.1275	0.1093
自由度	9	5	2
等固有値確率	0.00000	0.07170	0.94683
利用可能性	可	不可	不可

図 3 等固有値の検定結果

ここに表示された第 i 主成分の χ^2 値は、固有値を大きさの順番に並べた場合、第 i 主成分以降の固有値がすべて等しいとみなせるかどうかの検定値であり、等固有値確率はその確率値を表わす。それゆえ等固有確率が有意水準より大きい主成分以降が利用に適さないことを示している。極端な例として、第 1 主成分の等固有値確率が有意水準より小さい場合、主成分分析自体があまり意味を持たない。

「主成分得点」の出力は各主成分毎に図 4 に与えられ、2 つの主成分に関する主成分得点の散布図は図 5 に与えられる。これによって主成分で見た場合の個体の類似度を把握することが容易となる。

主成分得点

	主成分1	主成分2	主成分3	主成分4
1	-0.0687	0.2341	0.3491	-0.2616
2	2.8001	-0.3830	0.0957	-0.2748
3	2.6936	-0.0169	-0.3541	0.3526
4	1.3972	0.0595	-0.2074	-0.0435
5	0.9189	0.5749	0.0867	0.1780
6	-2.7897	-0.3429	-0.0325	-0.0306
7	2.4015	0.1649	0.4613	-0.1602
8	-2.7662	0.3126	0.0324	-0.2183
9	1.5295	1.6757	0.3257	0.0074
10	2.4794	-0.9564	-0.1196	-0.3841
11	0.7829	-0.1603	-0.1257	-0.2892

図 4 主成分得点出力結果

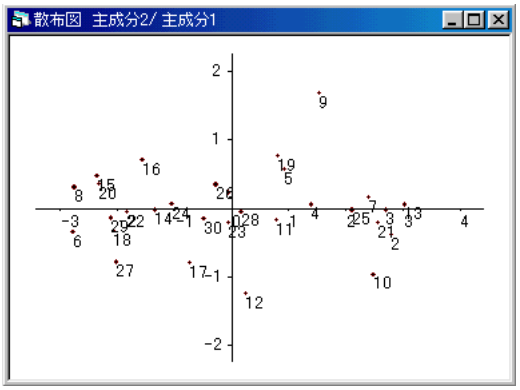


図 5 主成分得点散布図

５．因子分析

5.1　因子分析の理論

因子分析が取り扱うデータは主成分分析等と同様に p 変数、n 個体（レコード）の変量 $x_{i\lambda}$（$i=1,2,\cdots,p,\ \lambda=1,2,\cdots,n$）である。これらのデータから各変数 x_i に内在すると思われる因子を抽出することが因子分析のねらいである。

因子分析では変数 x_i を標準化した変数 $t_i=(x_i-\overline{x}_i)/u_i$ を用いることが多いので、今後はこの変数 t_i を用いて議論を進める。ここで $\overline{x}_i$ は変数 x_i の標本平均、u_i は不偏分散から求めた標準偏差である。

因子分析では各データに内在すると考えられる共通因子 $f_{\alpha\lambda}$（$\alpha=1,2,\cdots,q\le p$）の線形結合によって、変数 t_i が以下のように表わされるものと考える。

$$t_i=\sum_{\alpha=1}^{q}a_{i\alpha}f_\alpha+\varepsilon_i \tag{1}$$

係数 $a_{i\alpha}$ は α 因子の因子負荷量と呼ばれている。ここで ε_i は誤差であり、共通因子 f_α との相関や互いの相関はないものと考える。

$$E\left(f_\alpha\varepsilon_i\right)=0,\quad E\left(\varepsilon_i\varepsilon_j\right)=0\ （i\neq j）$$

また共通因子 f_α についても互いの相関はなく、平均 0、分散 1 に標準化されているものとする。

$$E\left(f_\alpha f_\beta\right)=\delta_{\alpha\beta},\quad E\left(f_\alpha\right)=0,\quad V\left(f_\alpha\right)=1$$

これらを利用すると変数 x_i と x_j との相関係数 r_{ij} は以下のように表わせる。

$$r_{ij}=E\left(t_it_j\right)=\sum_{\alpha=1}^{q}a_{i\alpha}a_{j\alpha}\ （i\neq j）,\quad r_{ii}=V\left(t_i\right)=\sum_{\alpha=1}^{q}a_{i\alpha}^2+V\left(\varepsilon_i\right)=1$$

ここで、$h_i=\sum_{\alpha=1}^{q}a_{i\alpha}^2=1-V\left(\varepsilon_i\right)$ と置くと、上式は以下のように表わされる。

$$\mathbf{A}^t\mathbf{A}=\mathbf{R},$$

$$\mathbf{A}=\begin{pmatrix}a_{11}&a_{12}&\cdots&a_{1q}\\a_{21}&a_{22}&\cdots&a_{2q}\\\vdots&\vdots&\ddots&\vdots\\a_{p1}&a_{p1}&\cdots&a_{pq}\end{pmatrix},\quad \mathbf{R}=\begin{pmatrix}h_1&r_{12}&\cdots&r_{1p}\\r_{12}&h_2&\cdots&r_{2p}\\\vdots&\vdots&\ddots&\vdots\\r_{1p}&r_{2p}&\cdots&h_p\end{pmatrix} \tag{2}$$

この中で特に h_i は共通性と呼ばれる。

共通性の和を取ると、

$$\sum_{i=1}^{p} h_i = \sum_{i=1}^{p}\left(\sum_{\alpha=1}^{q} a_{i\alpha}^2\right) = \sum_{\alpha=1}^{q}\left(\sum_{i=1}^{p} a_{i\alpha}^2\right)$$

$$= \sum_{i=1}^{p} V(t_i) - \sum_{i=1}^{p} V(\varepsilon_i) = p - \sum_{i=1}^{p} V(\varepsilon_i)$$

となるが、この関係式を利用し、誤差 $\varepsilon_{i\lambda}$ が 0 に近づけば左辺は p に近づくことを考えて、因子 α の寄与率を以下のように定義する。

$$P_\alpha = \sum_{i=1}^{p} a_{i\alpha}^2 \Big/ p$$

我々は (2) 式を解いて因子負荷量 $a_{i\alpha}$ を求めようとするが、その求め方にもセントロイド法、主因子法、主成分分析法、最尤法、最小 2 乗法等種々の方法があるが [9]、ここでは歴史的に有名なセントロイド法と広く知られている主因子法、主成分分析法を取り上げる。

セントロイド法と主因子法では、最初に適当な推定値 h_i を用いて、因子負荷量 $a_{i\alpha}$ を計算し、その値を使って再度 $h_i = \sum_{\alpha=1}^{q} a_{i\alpha}^2$ で共通性 h_i を計算し、それをまた推定値として再び因子負荷量を計算する。これを共通性 h_i が収束するまで（このプログラムでは前回との差が 0.001 以下になるまで）繰り返すという方法で近似値を求める。その際最初の共通性 h_i の推定値には変数 x_i と他の変数の重相関係数や他との相関係数の中で最大のものなどが利用される。主成分分析法では、相関行列の固有ベクトルをそのまま推定値として利用し、必要な次元までを採用する。以後詳しく見て行く。

セントロイド法は第 1 因子から逐次因子負荷量を求めていく手法で、

$$a_{i1} = \sum_{j=1}^{p} r_{ji} \Big/ \sqrt{\sum_{j=1}^{p}\sum_{k=1}^{p} r_{jk}} \qquad (r_{ii} = h_i)$$

の形で第 1 因子の因子負荷量を与える。次に $r_{ij}^{(1)} = r_{ij} - a_{i1}a_{j1}$ として新たな相関行列を定義するが、その際対角要素は各行の非対角要素の絶対値の最大値を用い、負の相関係数をできるだけ少なくするために、参考文献 8) のアルゴリズムに従い座標反転を行なう。この相関行列を利用して新たに第 2 因子の因子負荷量を同様の方法で計算する。

$$a_{i2} = \sum_{j=1}^{p} r_{ji}^{(1)} \Big/ \sqrt{\sum_{j=1}^{p}\sum_{k=1}^{p} r_{jk}^{(1)}}$$

さらに $r_{ij}^{(2)} = r_{ij}^{(1)} - a_{i2}a_{j2}$ を用いて新たな相関行列を作り、上に述べた方法で対角要素と負の相関についての処理を行ない、次の因子の因子負荷量を計算して行く。

次に主因子法は対角成分を共通性 h_i で置き換えた相関行列 $\mathbf{R}$ の固有値と固有ベクトルによって因

子負荷量 $a_{i\alpha}$ が計算される。即ち、第 α 因子の因子負荷量 $a_{i\alpha}$ は、行列 $\mathbf{R}$ の固有値 λ_α と規格化された固有ベクトル $u_{i\alpha}$ を使って、

$$a_{i\alpha} = \sqrt{\lambda_\alpha} u_{i\alpha}$$

のように与えられる。

主成分分析法は、相関行列 $\mathbf{R}$ をそのまま使い、固有値と固有ベクトルによって因子負荷量 $a_{i\alpha}$ を計算する。第 α 因子の因子負荷量 $a_{i\alpha}$ は、相関行列 $\mathbf{R}$ の固有値 λ_α と規格化された固有ベクトル $u_{i\alpha}$ を使って、

$$a_{i\alpha} = \sqrt{\lambda_\alpha} u_{i\alpha}$$

のように与える。共通性は $h_i = \sum_{\alpha=1}^{p} a_{i\alpha}^2$ のように因子負荷量から計算する。

次に各因子、各個体毎の因子得点 $f_{\alpha\lambda}$ の値について考える。前にも述べたとおり、誤差項が特定できない限り、一般に観測値 $x_{i\lambda}$ から因子得点 $f_{\alpha\lambda}$ を決定することはできない。そこで我々は分散で重み付けされた誤差の 2 乗項

$$\sum_{\lambda=1}^{n}\sum_{i=1}^{p} \varepsilon_{i\lambda}^2 / u_i^2 = \sum_{\lambda=1}^{n}\sum_{i=1}^{p} (t_{i\lambda} - \sum_{\alpha=1}^{q} a_{i\alpha} f_{\alpha\lambda})^2 \Big/ u_i^2$$

が最小になるように仮定して、因子得点 $f_{\alpha\lambda}$ を推定する。この解は成分が

$$(\mathbf{F})_{\lambda\alpha} = f_{\alpha\lambda}, \quad (\mathbf{T})_{\lambda i} = t_{i\lambda}, \quad (\mathbf{A})_{i\alpha} = a_{i\alpha}, \quad (\mathbf{D})_{ij} = u_i^2 \delta_{ij},$$

のように与えられる行列 $\mathbf{F}, \mathbf{T}, \mathbf{A}, \mathbf{D}$ を用いて以下のように求められる。

$$\mathbf{F} = \mathbf{T}\mathbf{D}^{-1}\mathbf{A}({}^t\mathbf{A}\mathbf{D}^{-1}\mathbf{A})^{-1}$$

この推定法は Bartlett の重みつき最小 2 乗推定法と呼ばれる。

この他にも回帰推定法と呼ばれるものがある。(1)式から、共通因子の推定値と変数は以下のような関係にあると考える。

$$t_{i\lambda} = \sum_{\alpha=1}^{q} a_{i\alpha} \hat{f}_{\alpha\lambda}$$

これから、

$$\sum_{i=1}^{p} a_{i\alpha} t_{i\lambda} = \sum_{i=1}^{p}\sum_{\beta=1}^{q} a_{i\alpha} a_{i\beta} \hat{f}_{\beta\lambda} = \sum_{\beta=1}^{q} \lambda_\alpha \delta_{\alpha\beta} \hat{f}_{\beta\lambda} = \lambda_\alpha \hat{f}_{\alpha\lambda}$$

となり、以下を得る。

$$\hat{f}_{\alpha\lambda} = \frac{1}{\lambda_\alpha}\sum_{i=1}^{p} a_{i\alpha} t_{i\lambda} = \sum_{i=1}^{p}\sum_{j=1}^{p} r^{ij} a_{j\alpha} t_{i\lambda} \equiv \sum_{i=1}^{p} b_{\alpha i} t_{i\lambda} \qquad (3)$$

ここで、$\mathbf{R}\mathbf{a}_\alpha = \lambda \mathbf{a}_\alpha \Leftrightarrow \mathbf{R}^{-1}\mathbf{a}_\alpha = (1/\lambda_\alpha)\mathbf{a}_\alpha$ の関係を用いた。これにより、因子得点を求める係

数$b_{\alpha i}$は以下のように与えられる。

$$b_{\alpha i} = \sum_{j=1}^{p} r^{ij} a_{j\alpha}$$

この関係は、(3)式が $\hat{f}_{\alpha}$ を推定する重回帰分析の式（目的変数には実測値がないが）であると考えることによっても導かれる。重回帰式の標準化係数は$b_i = \sum_{j=1}^{p} r^{ij} r_{jy}$ であり、r_{jy} は変数 j と目的変数の相関係数である。この場合目的変数は因子 α なので、相関係数は因子負荷量 $a_{j\alpha}$ である。

ここで求めた因子負荷量 $a_{i\alpha}$ には、$a^{*}_{i\alpha} = \sum_{\beta=1}^{q} o_{\alpha\beta} a_{i\beta}$ ，$\sum_{\gamma=1}^{q} o_{\alpha\gamma} o_{\beta\gamma} = \delta_{\alpha\beta}$ のような回転の自由度が存在する。この変換により、(1)式は以下のように変わり、因子も回転を受ける。

$$t_i = \sum_{\alpha=1}^{q} a^{*}_{i\alpha} f^{*}_{\alpha} + \varepsilon_i \quad , \quad f^{*}_{\alpha} = \sum_{\beta=1}^{q} o_{\alpha\beta} f_{\beta}$$

しかし、(2)式、寄与率、因子の平均と分散や直交性は不変である。この性質を利用して、因子負荷量の各因子の分散を最大化するように回転させると因子の解釈が容易になる。この直交回転をバリマックス回転という。

最後に、このようにして推測された共通因子からデータはどの程度推測できるのであろうか。実際に以下の式によってデータを推測し、観測値との相関係数を調べてみるとモデルの良さが実感できる。

$$\hat{t}_{i\lambda} = \sum_{\alpha=1}^{q} a_{i\alpha} \hat{f}_{\alpha\lambda}$$

その後、参考文献 [1]を用いて、プロマックス回転についてプログラムを作成したので、追加しておく。

斜交回転の軸に用いられる用語として、プライマリ因子軸とは、斜交回転をした場合の斜交軸のことであり、参考因子軸とは、プライマリ因子軸と直交する斜交軸のことである。

ステップ１

直交回転後の因子負荷行列 $\mathbf{A}$ から始める。

$\mathbf{A}$ の各要素を、各行の２乗和が１となるように共通性を用いて基準化する。

絶対値最大の２乗和が±1（我々の場合はバリマックス回転ですでに正）となるように定数倍する。

$\mathbf{A}$ の各要素を k 乗したものを目標行列 $\mathbf{A}^{*}$ とする。ここで k が奇数の場合はそのまま、偶数の場合は要素の符号をかけておく。通常 k は３か４を指定するが、我々の場合は４にしている。

これによって、絶対値が１に近いものを除き、他の要素は０に近づく。

ステップ２

回転後の $\mathbf{A}$ が $\mathbf{A}^*$ と最小２乗法の意味で最も近くなるような変換（プロクラステス変換）行列 $\mathbf{T}_r$ は以下の式で与えられる。

$$\mathbf{T}_r = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{A}^*$$

ステップ３

プライマリ因子の因子構造を計算するための回転行列 $\mathbf{T}_p$ は、$\mathbf{T}_p = (\mathbf{T}_r')^{-1}$ の列ノルムを１に基準化した行列である。

ステップ４

プライマリ因子軸と直交する（参考因子軸に沿う）成分である因子構造行列 $\mathbf{S}_p$、参考因子軸と直交する（プライマリ因子軸に沿う）成分である因子パターン行列 $\mathbf{P}_p$、因子間の相関行列 Φ_p を以下より求める。ここで、結果には因子構造行列 $\mathbf{S}_p$ と因子パターン行列 $\mathbf{P}_p$ を用いる。

$$\mathbf{S}_p = \mathbf{A}\mathbf{T}_p, \quad \mathbf{P}_p = \mathbf{A}(\mathbf{T}_p')^{-1}, \quad \Phi_p = \mathbf{T}_p'\mathbf{T}_p$$

最尤法と最小２乗法による因子分析（追加）

標準化された観測変数を以下のように p 個の因子（内生変数）で表すものとする。

$$x_{i\lambda} = \sum_{j=1}^{p} a_{ij} f_{p\lambda} + b_i e_{i\lambda}$$

これを用いると、相関係数 s_{ij} は以下のように書ける。

$$\begin{aligned} s_{ij} &= \frac{1}{N}\sum_{\lambda=1}^{N} x_{i\lambda}x_{j\lambda} = \frac{1}{N}\sum_{\lambda=1}^{N}(\sum_{k=1}^{p} a_{ik} f_{k\lambda} + b_i e_{i\lambda})(\sum_{l=1}^{p} a_{jl} f_{l\lambda} + b_j e_{j\lambda}) \\ &= \sum_{k=1}^{p}\sum_{l=1}^{p} a_{ik}a_{jl}\frac{1}{N}\sum_{\lambda=1}^{N} f_{k\lambda} f_{l\lambda} + b_i b_j \frac{1}{N}\sum_{\lambda=1}^{N} e_{i\lambda}e_{j\lambda} \\ &= \sum_{k=1}^{p} a_{ik}a_{jk} + b_i b_j \delta_{ij} \equiv \mathbf{\Sigma}_{ij} \end{aligned}$$

ここに、因子と誤差について、以下の関係があるものとする。

$$\frac{1}{N}\sum_{\lambda=1}^{N} f_{k\lambda} f_{l\lambda} = \delta_{kl}, \quad \frac{1}{N}\sum_{\lambda=1}^{N} e_{i\lambda}e_{j\lambda} = \delta_{ij}, \quad \sum_{\lambda=1}^{N} f_{k\lambda}e_{i\lambda} = 0$$

最尤法では、以下の尤度を考える。

$$L = \prod_{\lambda=1}^{N} \frac{1}{\sqrt{(2\pi)^p \mid \mathbf{\Sigma} \mid}} \exp\left[-\frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} (\mathbf{\Sigma}^{-1})_{ij} x_{i\lambda} x_{j\lambda} \right]$$

$$= (2\pi)^{-pN/2} \mid \mathbf{\Sigma} \mid^{-N/2} \exp\left[-\frac{N}{2} tr(\mathbf{\Sigma}^{-1}\mathbf{S}) \right]$$

これを最大化するために、符号を反対にした対数尤度の最小化を考える。

$$-\log L = \frac{pN}{2} \log(2\pi) + \frac{N}{2} \log \mid \mathbf{\Sigma} \mid + \frac{N}{2} tr(\mathbf{\Sigma}^{-1}\mathbf{S})$$

$$= \frac{N}{2} \left[tr(\mathbf{\Sigma}^{-1}\mathbf{S}) - \log \mid \mathbf{\Sigma}^{-1} \mid \right] + const.$$

この形から、$\mathbf{\Sigma}$ と $\mathbf{S}$ が完全に一致する場合に 0 になるように、評価関数 $f_{ML}(\mathbf{a},\mathbf{b})$ を以下のように定義してこれを最小化する。

$$f_{ML}(\mathbf{a},\mathbf{b}) = tr(\mathbf{\Sigma}^{-1}\mathbf{S}) - \log \mid \mathbf{\Sigma}^{-1}\mathbf{S} \mid - p$$

最小 2 乗法では、評価関数 $f_{MS}(\mathbf{a},\mathbf{b})$ を以下のように定義してこれを最小化する。

$$f_{MS}(\mathbf{a},\mathbf{b}) = \sum_{i=1}^{p} \sum_{j=1}^{p} (\Sigma_{ij} - s_{ij})^2$$

評価関数が複雑なため、パラメータの初期値の与え方が重要であるが、我々のプログラムでは、確実に値の求まる最小 2 乗法の結果を初期値として用いる。

この手法の追加に伴い、セントロイド法は選択肢から外している。

5.2　プログラムの利用法

因子分析の実際の実行画面を図 1 に示す。データとしては主成分分析と同じように個体毎の元データ、共分散行列、相関行列が選択できる。因子負荷量を求める方法では、主因子法、主成分分析、最小 2 乗法、最尤法が利用できる。歴史的なセントロイド法については、最小 2 乗法と最尤法を導入した際に削除した（必要があれば復活も可能である）。

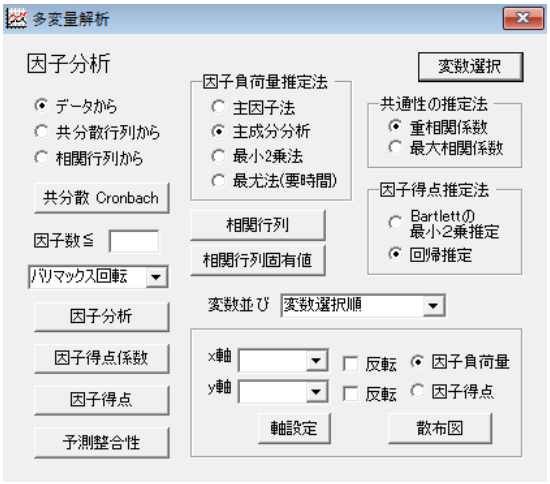


図 1 因子分析画面

図 2 に因子数を 2 としてバリマックス回転にチェックを入れ、「因子分析」のボタンをクリックした場合の出力画面を示す。

	因子1	因子2	共通性
国語	0.0745	-0.7869	0.6248
英語	0.2205	-0.7367	0.5914
社会	0.1566	-0.7625	0.6059
数学 I	0.8511	-0.2033	0.7657
数学 II	0.8451	-0.2964	0.8021
理科	0.8890	-0.0397	0.7919
寄与率	0.3846	0.3124	
累積寄与率	0.3846	0.6970	
符号調整済α	0.6691	0.5381	

図 2 因子分析出力画面

因子数で指定した数だけ因子負荷量と寄与率、累積寄与率が表示されている。但し、因子数を指定しない場合は、セントロイド法で累積寄与率が 0.9 を超えたところで、主因子法では固有ベクトルの値が 0.5 未満になったところで因子の出力を停止する。また、因子数を指定した場合でも、主因子法で固有値が 0 に近い負の値を取ることも見つかっており、指定した個数より少なく表示される場合もある。この原因は現在考察中である。符号調整済みαは、因子負荷量の符号が同じになるように、変数の符号を調整して因子負荷量の大きさで組み分けした場合の Cronbach のα係数である。これは、一般には 0.8 程度以上が良いとされている。

「因子得点」ボタンをクリックすると図 3 のように個体毎の因子得点が表示される。ここでは因子得点の推定に、Bartlett の重みつき最小 2 乗推定法を用いている。「散布図」ボタンをクリックすると図 4 のように因子得点 1 を横軸に因子得点 2 を縦軸にした散布図を作成する。

	因子1	因子2
37	1.0797	1.3139
38	0.7078	-1.6708
39	-1.0450	0.0860
40	2.5198	0.5770
41	0.5220	-0.1223
42	-0.4754	-0.9796
43	0.2715	-1.4933
44	-0.4913	-1.1754
45	-0.3260	-0.0892
46	0.0810	-0.1131
47	-0.4149	-0.6058
48	-0.6418	0.9986
49	-1.0654	0.4957
50	0.9558	1.8029
平均	0.0000	0.0000
標準偏差	0.9899	0.9899

図 3 因子得点出力画面

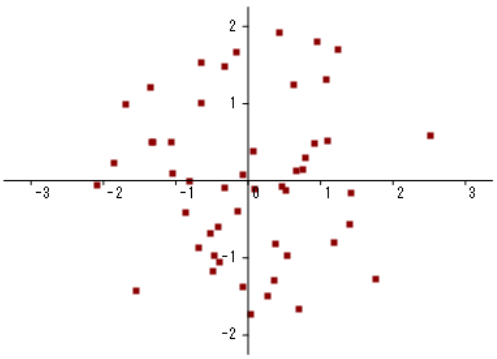


図 4 因子得点散布図

新しくバリマックス回転の機能を追加したが、それ以外に因子負荷量推定法に主成分分析を、因子得点の推定法として回帰推定も追加した。これらはよく利用されているのでデフォルトで、使うように設定している。

「因子得点係数」ボタンをクリックすると、因子得点を求めるための係数が、図 5 のように表示される。実データから求める場合と標準化されたデータ（不偏分散による）から求める場合の 2 種類の係数が示されている。


因子得点の係数

	国語	英語	社会	数学 I	数学 II	理科	定数項
因子1(標準化データ)	-0.13276	-0.04763	-0.08572	0.38865	0.36543	0.44342	
因子2(標準化データ)	-0.47650	-0.41340	-0.44341	0.05726	-0.00231	0.16796	
因子1(実データ)	-0.00944	-0.00278	-0.00586	0.02523	0.02669	0.02156	-3.07657
因子2(実データ)	-0.03388	-0.02416	-0.03033	0.00372	-0.00017	0.00817	4.87724

図 5 因子得点を求める場合の係数

「予測整合性」というボタンは、因子得点を計算して、逆に元のデータを予測し、実データと比較して、因子分析の効果を実感してもらうためのものである。その実行画面を図 6 に示す。


予測整合性（標準化変数値と予測値）

	国語	予測値	英語	予測値	社会	予測値	数学 I	予測値	数
45	0.492	0.046	0.394	-0.006	-0.709	0.017	-0.491	-0.259	
46	0.919	0.095	-1.009	0.101	0.386	0.099	0.288	0.092	
47	0.492	0.446	-0.892	0.355	1.207	0.397	-0.621	-0.230	
48	-0.219	-0.834	-1.652	-0.877	-0.845	-0.862	-0.815	-0.749	
49	-0.717	-0.469	-0.191	-0.600	-0.709	-0.545	-1.010	-1.008	
50	-1.286	-1.348	0.219	-1.117	-2.350	-1.225	0.223	0.447	
平均値	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
標準偏差	1.000	0.782	1.000	0.761	1.000	0.771	1.000	0.866	
相関係数		0.782		0.761		0.771		0.866	

図 3.3 実測値と予測値の比較画面

因子分析のバージョンアップで、因子負荷量推定法に主成分分析を加えたことは前に述べたが。これによって因子数を変数の数まで任意に選ぶことができるようになり、主成分分析の同じ主成分数の場合と累積寄与率が等しくなる。また、他の推定法に比べても累積寄与率の値は向上する。その他に、出力変数の並びをこれまでの変数選択順の他に、因子負荷量の大きさで 2 通りに並べ替える方法を加えた。これによって因子ごとに因子負荷量の大きい変数同士を並べて表示できるようになり、因子の解釈がより容易になる。

新たに最小 2 乗法と最尤法も加えたが、最尤法は変数が多くなると時間を要する。

参考文献

[1] 田中豊・垂水共之編，Windows 版統計解析ハンドブック多変量解析，共立出版，1995.

６．クラスター分析

6.1　クラスター分析の理論

クラスター分析は個体や変数間の様々に定義された距離に基づき、これらを分類する手法である。その中でもここで取り扱うのはクラスターを 1 つずつまとめてゆく階層的方法と呼ばれるものである。クラスター分析のデータは変数と個体のシート形式で、表１のように与えられる。

表１ クラスター分析のデータ

	変数 1	変数 2	…	変数 p
個体 1	x_{11}	x_{21}	…	x_{p1}
個体 2	x_{12}	x_{22}	…	x_{p2}
:	:	:	:	:
個体 n	x_{1n}	x_{2n}	…	x_{pn}

クラスター分析には距離の測定方法やクラスターの構成法にさまざまな種類があるが、ここでは利用者の理解し易い代表的な数種のものについて取り上げている。距離の測定は２つの個体または変数の間で定義される。これらが複数個集まったクラスター間の距離の定義にはクラスター構成法を利用する。

ここではまず、距離の測定方法を個体間のものと変数間のものに分けて説明する。個体 μ と個体 ν との距離には以下のようなものがある。最初に量的なデータに対してその定義を示す。

ユークリッド距離　　$d_{\mu\nu}^2 = \sum_{i=1}^{p} (x_{i\mu} - x_{i\nu})^2$

標準化ユークリッド距離　　$d_{\mu\nu}^2 = \sum_{i=1}^{p} \frac{1}{s_i^2} (x_{i\mu} - x_{i\nu})^2$

マハラノビス距離　　$d_{\mu\nu}^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} (x_{i\mu} - x_{i\nu}) s^{ij} (x_{j\mu} - x_{j\nu})$

ここに s_i^2 は変数 i の不偏分散、添え字の上に付いた s^{ij} は共分散行列 $\mathbf{S}$ の逆行列 $\mathbf{S}^{-1}$ の i, j 成分である。

$$s_i^2 = \frac{1}{n-1} \sum_{\lambda=1}^{n} (x_{i\lambda} - \bar{x}_i)^2 , \quad (\mathbf{S})_{ij} = s_{ij} = \frac{1}{n-1} \sum_{\lambda=1}^{n} (x_{i\lambda} - \bar{x}_i)(x_{j\lambda} - \bar{x}_j)$$

次に、0/1 の値で与えられるカテゴリデータに対しては、以下の統計量を距離として用いる。

類似比　　$d_{\mu\nu} = a/(a+b+c)$

一致係数　　$d_{\mu\nu} = (a+d)/(a+b+c+d)$

ファイ係数　　$d_{\mu\nu} = (ad - bc) \Big/ \sqrt{(a+b)(c+d)(a+c)(b+d)}$

ここに、a, b, c, d は以下のように与えられる。

$$a = \sum_{i=1}^{p} x_{i\mu} x_{i\nu}\ ,\quad b = \sum_{i=1}^{p} x_{i\mu}(1 - x_{i\nu})\ ,\quad c = \sum_{i=1}^{p} (1 - x_{i\mu}) x_{i\nu}\ ,\quad d = \sum_{i=1}^{p} (1 - x_{i\mu})(1 - x_{i\nu})$$

次に、変数 *i, j* 間の距離について述べる。数値データに対しては、以下の統計量を距離として用いる。

相関　　$d_{ij} = 1 - s_{ij} / s_i s_j$　（1−相関係数）

順位相関　　$d_{ij} = 1 - \tilde{s}_{ij} / \tilde{s}_i \tilde{s}_j$　（1−順位相関係数）

ここに、$\tilde{s}_i$ 及び $\tilde{s}_{ij}$ は、データの代わりに変数別に付与された順位データを用いて求めた、標準偏差と共分散である。

カテゴリデータに対しては、まず以下のような変数 *i, j* に対する統計量 χ_{ij}^2 を求める。

$$\chi_{ij}^2 = \sum_{k=1}^{r_i} \sum_{l=1}^{r_j} \frac{\left(\left| n_{kl} - n_{k\bullet} n_{\bullet l} / n \right| - 1/2 \right)^2}{n_{k\bullet} n_{\bullet l} / n}$$

ここに、r_i は変数 *i* の分類数、n_{kl} は変数 *i* の *k* 番目の分類と変数 *j* の *l* 番目の分類に含まれるデータ数及び、$n_{k\bullet}$ と $n_{\bullet l}$ はそれぞれ n_{kl} の *l* についての和と *k* についての和である。

これを用いて以下のように距離を定義する。

平均平方根一致係数　　$d_{ij} = \sqrt{\chi_{ij}^2 / n}$

一致係数　　$d_{ij} = \sqrt{\chi_{ij}^2 / (\chi_{ij}^2 + n)}$

クラメールの V　　$d_{ij} = \sqrt{(\chi_{ij}^2 / n) / \min(r_i - 1, r_j - 1)}$

次にクラスター構成法について述べる。ここではクラスター *f* とクラスター *g* を結合してクラスター *h* を作り、他のクラスター *l* との距離を求める場合を考える。クラスター *h* とクラスター *l* の距離を D_{hl} で表わすと、これらの関係は以下のように与えられる。

最短距離法　　$D_{hl} = \frac{1}{2} D_{fl} + \frac{1}{2} D_{gl} - \frac{1}{2} \left| D_{fl} - D_{gl} \right|$

最長距離法　　$D_{hl} = \frac{1}{2} D_{fl} + \frac{1}{2} D_{gl} + \frac{1}{2} \left| D_{fl} - D_{gl} \right|$

メジアン法　　$D_{hl} = \frac{1}{2} D_{fl} + \frac{1}{2} D_{gl} - \frac{1}{4} D_{fg}$

重心法　　$D_{hl}^2 = \frac{n_f}{n_h} D_{fl}^2 + \frac{n_g}{n_h} D_{gl}^2 - \frac{n_f n_g}{n_h^2} D_{fg}^2$

群平均法　　$D_{hl}^2 = \frac{n_f}{n_h} D_{fl}^2 + \frac{n_g}{n_h} D_{gl}^2$

ウォード法　　$$D_{hl}^2 = \frac{1}{n_h + n_l}\left[(n_f + n_l)D_{fl}^2 + (n_g + n_l)D_{gl}^2 - n_l D_{fg}^2\right]$$

但し、重心法、群平均法、ウォード法について、距離はユークリッド距離をとるものとする。

6.2　プログラムの利用法

メニュー［分析－多変量解析－クラスター分析］を選択して表示される、クラスター分析の分析画面を図 1 に示す。

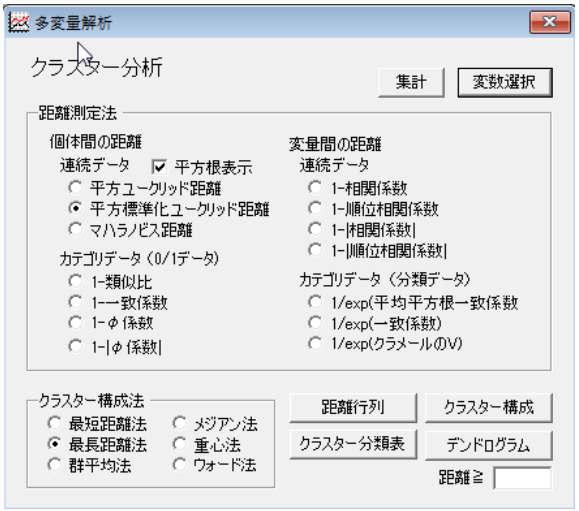
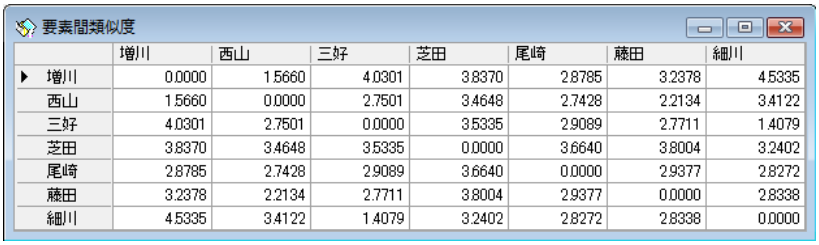


図 2 クラスター分析メニュー画面

変数を選択して「距離行列」ボタンをクリックした場合の出力結果を図 2 に示す。これは各要素の類似度（距離）を表示したものである。


要素間類似度

	増川	西山	三好	芝田	尾崎	藤田	細川
▶ 増川	0.0000	1.5660	4.0301	3.8370	2.8785	3.2378	4.5335
西山	1.5660	0.0000	2.7501	3.4648	2.7428	2.2134	3.4122
三好	4.0301	2.7501	0.0000	3.5335	2.9089	2.7711	1.4079
芝田	3.8370	3.4648	3.5335	0.0000	3.6640	3.8004	3.2402
尾崎	2.8785	2.7428	2.9089	3.6640	0.0000	2.9377	2.8272
藤田	3.2378	2.2134	2.7711	3.8004	2.9377	0.0000	2.8338
細川	4.5335	3.4122	1.4079	3.2402	2.8272	2.8338	0.0000

図 2 類似度行列

クラスター分析で最も利用する「デンドログラム」の出力結果を図 3 に与える。

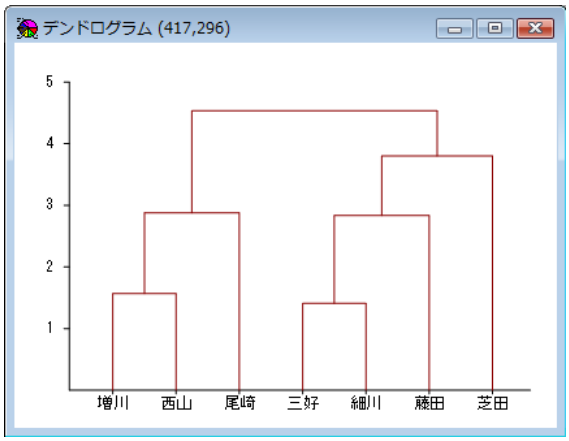


図 3 デンドログラム

デンドログラムでは構成の際の類似度が読みづらいので構成順を表にして示す。「クラスター構成」ボタンをクリックすると図 4 に示される結果が表示される。

	クラスター名	クラスター名	類似度
▶ 1	E:三好	E:細川	1.4079
2	E:増川	E:西山	1.5660
3	C:三好	E:藤田	2.8338
4	C:増川	E:尾崎	2.8785
5	C:三好	E:芝田	3.8004
6	C:増川	C:三好	4.5335

図 4 クラスターの構成

クラスター名の先頭に E の付いたものは要素名、C の付いたものはクラスターである。クラスター名はデンドログラムで表示される左端の要素名で代表される。例えば、最初の行は、要素「三好」と要素「増川」が結合され、クラスター「三好」になる、と読む。また、3 番目の行は、クラスター「三好」と要素「藤田」が結合され、クラスター「三好」になる、と読む。

「クラスター分類表」ボタンをクリックすると、例えば、図 3 のデンドログラムを表形式で表した図 5 のクラスター分類表が表示される。これはクラスター構成の各段階での分類を表示している。これによって例えば全体を 2 分割するときに各個体がどちらのクラスターに属するか簡単に知ることができる。また、これを利用して 2 つのクラスター間での有意差検定などを行いたい場合、この表の列をコピーして元データに加え、簡単に群分けすることができるようになる。


クラスター分類表

	並び	7	6	5	4	3	2	1
▶	増川	1	1	1	1	1	1	1
	西山	2	2	2	1	1	1	1
	三好	4	3	3	3	3	3	1
	芝田	7	4	4	4	4	4	1
	尾崎	3	5	5	5	5	1	1
	藤田	6	6	6	6	3	3	1
	細川	5	7	3	3	3	3	1

図 5 クラスター分類表

他の分析でも同様であるが、これまで予測値は欠損値データを除いて表示していたが、新しいデータを作成することを考えると欠損値を加えたままで表示し、元のデータに簡単に追加できるようにする方が賢明である。例えばこのクラスター分類表で、芝田のデータに欠損がある場合、図 6 の形式で表示すべきである。


クラスター分類表

	並び	6	5	4	3	2	1
▶	増川	1	1	1	1	1	1
	西山	2	2	2	1	1	1
	三好	3	3	3	3	3	1
	芝田						
	尾崎	6	4	4	4	3	1
	藤田	5	5	5	5	3	1
	細川	4	6	3	3	3	1

図 6 欠損値のある場合の分類表の表示

この考えをすべての多変量解析に適用し、予測値には欠損値も加えて表示するように変更した。特に予測値の並びが変わった分析は、判別分析と数量化Ⅱ類である。これらは今まで群ごとに予測値を表示していたが、新たにデータ並びの順に表示するように作り変えた。

７．正準相関分析

7.1　正準相関分析の理論

正準相関分析は変数 $x_1, x_2, \cdots, x_r$ と変数 $y_1, y_2, \cdots, y_s$ を含む２群間の相関係数を、これらの変数を用いた１次関数間の相関係数と定義し、この相関係数が最大となるように係数を決める手法である。

まず、以下のような線形結合により、新しい変数 u，v を考える。

$$u = {}^t\mathbf{ax}, \quad v = {}^t\mathbf{by},$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_r \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_r \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_s \end{pmatrix}$$

ここに、$\mathbf{a}$，$\mathbf{b}$ は係数ベクトルである。

変数 $x_1, x_2, \cdots, x_r$ と変数 $y_1, y_2, \cdots, y_s$ の分散共分散行列をそれぞれ $\mathbf{S}_{xx}$，$\mathbf{S}_{yy}$ とし、２組の変数間の分散共分散行列を $\mathbf{S}_{xy}$（$\mathbf{S}_{yx} = {}^t\mathbf{S}_{xy}$）とすると、$u$ と v の相関係数 r_{uv} は以下となる。

$$r_{uv} = {}^t\mathbf{aS}_{xy}\mathbf{b}$$

但し係数ベクトルは u，v の分散が１になるように ${}^t\mathbf{aS}_{xx}\mathbf{a} = 1$，${}^t\mathbf{bS}_{yy}\mathbf{b} = 1$ と規格化している。

制約条件 ${}^t\mathbf{aS}_{xx}\mathbf{a} = 1$，${}^t\mathbf{bS}_{yy}\mathbf{b} = 1$ を入れ、Lagrange の未定定数法を用いて r_{uv} が最大となるように係数を求めると、以下の固有値問題に帰着する。

$$\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{a} = \rho^2\mathbf{a}, \quad {}^t\mathbf{aS}_{xx}\mathbf{a} = 1,$$

$$\mathbf{b} = \frac{1}{\rho}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{a}$$

ここに ρ は未定定数であるが、r_{uv} に等しいことが上の計算過程から分かっており、最大の相関係数の２乗は最大の固有値に等しい。この固有値に対応する固有ベクトル $\mathbf{a}$，$\mathbf{b}$ で決まる変量 u，v を（第１）正準変量、その時の相関係数を（第１）正準相関係数という。これに倣って α 番めに大きい固有値に対応する固有ベクトルから同様に求まるものをそれぞれ第 α 正準変量、第 α 正準相関係数という。

個体（レコード）λ について、変数 x_i のデータを $x_{i\lambda}$，変数 y_j のデータを $y_{j\lambda}$ とするとこの個体の正準変量 u_λ，v_λ は以下のように与えられる。

$$u_\lambda = \sum_{i=1}^{r} a_i x_{i\lambda}, \quad v_\lambda = \sum_{j=1}^{s} b_j y_{j\lambda}$$

ここでは元のデータから分散共分散行列を用いて求める方法を示したが、変数の大きさ（ばらつき）に極端な差があるときは、各変数を標準化して相関行列から同様の計算を進める。

正準変数 u と変数 x_i との相関係数 r_{ui}、正準変数 v と変数 y_j との相関係数 r_{vj} を正準負荷量という。正準負荷量を使った以下の定義を寄与率 P_u, P_v という。

$$P_u = \sum_{i=1}^{r} r_{ui}^2 / r \,, \quad P_v = \sum_{j=1}^{s} r_{vj}^2 / s$$

正準変数 u と変数 y_j との相関係数 r_{uj} 、正準変数 v と変数 x_i との相関係数 r_{vi} を交差負荷量という。公差負荷量を使った以下の定義を冗長性係数 Q_u, Q_v という。

$$Q_u = \sum_{j=1}^{s} r_{uj}^2 / s \,, \quad Q_v = \sum_{i=1}^{r} r_{vi}^2 / r$$

7.2　プログラムの利用法

正準相関分析の実行画面を図 1 に示す。

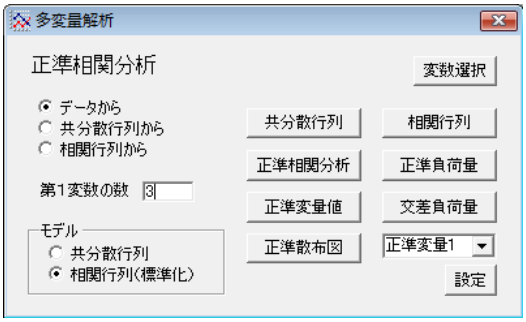


図 1 正準相関分析画面

分析は、主成分分析等と同様、元データ、分散共分散行列、相関行列から実行できるが、正準変量の値と正準変量の散布図については、当然元データがないと求められない。計算のモデルは、データをそのまま利用する場合と、標準化して相関行列を用いて計算する場合のどちらかを選ぶようになっている。直感的に分り易いのはそのままの値を利用するものであるが、変数の大きさが相当違う場合や係数から重要性を読み取ろうとする場合には標準化した方がよい。図 2 は 5 つの変数を、3 つと 2 つに分け、「正準相関分析」ボタンをクリックした実行結果である。

	正準変量 1	正準変量 2
正準相関係数	0.9560	0.3004
1群係数		
英語	1.1926	2.5235
国語	-0.0813	-2.3912
社会	-0.1494	-0.4650
2群係数		
数学	0.7392	-1.3634
理科	0.3141	1.5188

図 2 正準相関分析出力画面

この場合正準変量 $\boldsymbol{u}$ に含まれる変数の数として 3 を指定する。また、変数は同じ組の変数が並ぶように、選択順を調整する。結果は２つの正準変量の値と２つの正準相関係数の値を表示する。

次に図３に「正準変量の値」ボタンをクリックした場合の実行結果を示す。

	正準値1-1	正準値1-2	正準値2-1	正準値2-2
1	-0.4094	-0.2969	2.1992	1.2314
2	0.5306	0.4012	-0.4186	1.5912
3	1.0370	0.9764	-1.3427	-1.2820
4	-0.0323	0.3452	1.1427	-0.9415
5	1.0365	1.4806	-0.3974	-0.7291
6	1.0236	0.6838	-0.7758	-1.5660
7	0.5674	0.6696	-1.8596	-0.8808
8	-1.0215	-1.2806	-1.5344	0.2449
9	-1.2187	-0.7940	-0.1201	0.3359
10	-1.2154	-1.0970	-0.3800	-0.7526

図３ 正準変量の値画面

各個体毎に正準変量の値を計算して表示している。ここでは標準化されたデータから計算を進めたので、結果は標準化された値となる。これらのデータから第１正準変量について散布図を作ったものが、図４である。正準変量の選択は「設定」ボタンでできる。

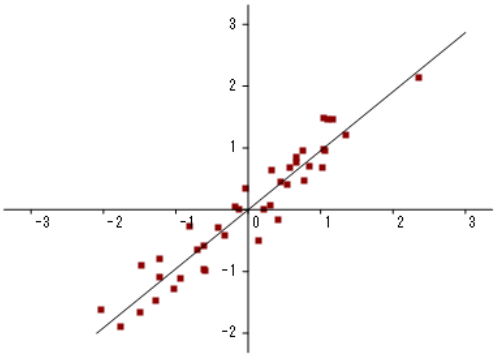


図４ 正準変量の散布図

第１正準変量のうちの一方を横軸に、もう一方を縦軸にとっているが、相当高い正準相関係数になることが見て取れる。

正準変数と、それと同じ組の変数との間の相関係数を正準負荷量という。「正準負荷量」ボタンをクリックすると、正準負荷量と各正準変量の寄与率が図５のように表示される。

正準負荷量

	正準負荷量1	正準負荷量2
1群		
寄与率	0.7944	0.0973
英語	0.9953	-0.0694
国語	0.9025	-0.4291
社会	0.7604	-0.3207
2群		
寄与率	0.8659	0.1341
数学	0.9793	-0.2025
理科	0.8791	0.4766

図 5 正準負荷量

正準変数と、それと違う組の変数との間の相関係数を交差負荷量という。「交差負荷量」ボタンをクリックすると、交差負荷量の値が図 6 のように表示される。

交差負荷量

	交差負荷量	交差負荷量
正準変数1		
冗長性係数	0.7914	0.0121
数学	0.9362	-0.0608
理科	0.8404	0.1432
正準変数2		
冗長性係数	0.7261	0.0088
英語	0.9515	-0.0208
国語	0.8628	-0.1289
社会	0.7270	-0.0963

図 6 交差負荷量

８．数量化Ｉ類

8.1　数量化Ｉ類の理論

数量化Ｉ類は、目的変数をカテゴリデータから推測する手法で、量的データの重回帰分析に相当する。数量化Ｉ類の変数は目的変数とアイテム毎に複数個含まれるカテゴリ変数からなる。データの基本的な形は表 1.1 に示される。カテゴリデータは各アイテム中の１つのカテゴリを選択するようになっており、選択された値が 1 で、他の値が 0 であるように定められている。これはデータの一般的な書式 $x_{ij\lambda}$ を用いて以下のように表わすこともできる。

$$x_{ij\lambda} \in \{0,1\},\ \sum_{j=1}^{r_i} x_{ij\lambda} = 1$$

表 1.1 数量化Ｉ類のデータ

目的変数	アイテム 1				アイテム p		
	カテゴリ 1	…	カテゴリ r_1	…	カテゴリ 1	…	カテゴリ r_p
y_1	x_{111}	…	x_{1r_11}	…	x_{p11}	…	x_{pr_p1}
y_2	x_{112}	…	x_{1r_12}	…	x_{p12}	…	x_{pr_p2}
:	:		:		:		:
y_n	x_{11n}	…	x_{1r_1n}	…	x_{p1n}	…	x_{pr_pn}

これより全カテゴリ数 r_c は以下で与えられる。

$$r_c = \sum_{i=1}^{p} r_i$$

目的変数は第 2 アイテム以降の第 1 カテゴリを除いた、以下の式で予測される。

$$Y_\lambda = \sum_{j=1}^{r_1} \hat{a}_{1j} x_{1j\lambda} + \sum_{i=2}^{p} \sum_{j=2}^{r_i} \hat{a}_{ij} x_{ij\lambda}$$

ここに、係数 $\hat{a}_{ij}$ は以下の残差変動 EV を最小化するように求める。後に述べるが係数はすべて独立ではない。このうちの１つは他の係数で求めることができる。それにより係数の数 r_d は以下で与えられる。

$$r_d = r_c - p$$

残差変動 EV の係数 $\hat{a}_{ij}$ についての微係数を 0 として、以下の解を得る。

$$EV = \sum_{\lambda=1}^{n} (y_\lambda - Y_\lambda)^2 \qquad \rightarrow \qquad \hat{\mathbf{a}} = ({}^t\mathbf{XX})^{-1}{}^t\mathbf{Xy}$$

ここに、各行列やベクトルは以下のように定義されるが、第 2 アイテム以降の第 1 カテゴリを外しているのは、行列 ${}^t\mathbf{XX}$ の正則性を失わせないためである。

$$
{}^t\hat{\mathbf{a}} = (\hat{a}_{11} \quad \cdots \quad \hat{a}_{1r_1} \quad \hat{a}_{22} \quad \cdots \quad \hat{a}_{2r_2} \quad \cdots \quad \hat{a}_{p2} \quad \cdots \quad \hat{a}_{pr_p})
$$

$$
{}^t\mathbf{y} = (y_1 \quad y_2 \quad \cdots \quad y_n)
$$

$$
\mathbf{X} = \begin{pmatrix}
x_{111} & \cdots & x_{1r_1 1} & x_{221} & \cdots & x_{2r_2 1} & \cdots & x_{p21} & \cdots & x_{pr_p 1} \\
x_{112} & \cdots & x_{1r_1 2} & x_{222} & \cdots & x_{2r_2 2} & \cdots & x_{p22} & \cdots & x_{pr_p 2} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{11n} & \cdots & x_{1r_1 n} & x_{22n} & \cdots & x_{2r_2 n} & \cdots & x_{p2n} & \vdots & x_{pr_p n}
\end{pmatrix}
$$

また、この係数は、

$$
Y_\lambda = \sum_{i=1}^{p} \sum_{j=2}^{r_i} \hat{a}_{ij} x_{ij\lambda} + \hat{a}_0
$$

として通常の重回帰分析の手法で求めることもできる。もちろん値は前のものと異なる。

ここで係数の自由度について考えてみる。

アイテム数をp個、第iのアイテムのカテゴリ数をr_i個とし、第iアイテムの第kカテゴリ、レコードλのデータを$x_{i(k)\lambda} = \{0,1\}$とし、数量化Ⅰ類の予測式が以下で与えられたとする。

$$
y_\lambda = \sum_{i=1}^{p} \sum_{k=1}^{r_i} b_{i(k)} x_{i(k)\lambda} + b_0 \ , \quad \sum_{k=1}^{r_i} x_{i(k)\lambda} = 1
$$

この式から、以下の関係も与えられる。

$$
\bar{y} = \sum_{i=1}^{p} \sum_{k=1}^{r_i} b_{i(k)} \bar{x}_{i(k)} + b_0
$$

この係数（カテゴリウェイト）には以下の自由度が存在する。

$$
b'_{i(k)} = b_{i(k)} - c_i \ , \quad b'_0 = b_0 + \sum_{i=1}^{p} c_i
$$

なぜなら、

$$
\sum_{i=1}^{p} \sum_{k=1}^{r_i} b'_{i(k)} x_{i(k)\lambda} + b'_0 = \sum_{i=1}^{p} \sum_{k=1}^{r_i} (b_{i(k)} - c_i) x_{i(k)\lambda} + b_0 + \sum_{i=1}^{p} c_i = \sum_{i=1}^{p} \sum_{k=1}^{r_i} b_{i(k)} x_{i(k)\lambda} + b_0
$$

この解に対して代表的なカテゴリウェイトを作ってみる。

重回帰ウェイト

$$
c_i = b_{i(0)}
$$

これにより、$b'_{i(0)} = 0$となる。

通常のカテゴリウェイト

$$
c_1 = -b_0 - \sum_{i=2}^{p} c_i \ , \quad c_i = b_{i(0)} \ (i \neq 1)
$$

これにより、$b'_0 = 0,\ b'_{i(0)} = 0\ \ (i \neq 1)$となる。

基準化ウェイト（これが最も重要である）

$$c_i = \sum_{k=1}^{r_i} b_{i(k)} \bar{x}_{i(k)}$$

これにより、

$$\sum_{i=1}^{p} c_i = \sum_{i=1}^{p} \sum_{k=1}^{r_i} b_{i(k)} \bar{x}_{i(k)} = \bar{y}$$

となり、予測式は以下となる。

$$y_\lambda = \sum_{i=1}^{p} \sum_{k=1}^{r_i} b'_{i(k)} x_{i(k)\lambda} + \bar{y}$$

これは$b'_{i(k)}$が目的変数を平均より上げるか下げるか分かるようになる。

分析の寄与率R^2（重相関係数R）、自由度調整済み寄与率R^{*2}（自由度調整済み重相関係数R^*）は、以下のように全変動SV、回帰変動RV、残差変動EVを用いて与えられる。。

$$SV = \sum_{\lambda=1}^{n} (y_\lambda - \bar{y})^2 = \sum_{\lambda=1}^{n} (y_\lambda - Y_\lambda)^2 + \sum_{\lambda=1}^{n} (Y_\lambda - \bar{y})^2 = EV + RV$$

$$R^2 = RV/SV = 1 - EV/SV\ ,\quad R^{*2} = 1 - \frac{EV/(n - r_d - 1)}{SV/(n-1)}$$

各アイテムと目的変数の共分散行列s_{ij}, s_{iy}, s_{yy}を以下で定義する。

$$s_{ij} = \frac{1}{n-1} \sum_{\lambda=1}^{n} (X_{i\lambda} - \overline{X}_i)(X_{j\lambda} - \overline{X}_j)\ ,\quad s_{iy} = \frac{1}{n-1} \sum_{\lambda=1}^{n} (X_{i\lambda} - \overline{X}_i)(y_\lambda - \bar{y})\ ,$$

$$s_{yy} = \frac{1}{n-1} \sum_{\lambda=1}^{n} (y_\lambda - \bar{y})^2$$

ここに、アイテムiの予測値$X_{i\lambda}$及びその平均$\overline{X}_i$は以下で与えられる。

$$X_{i\lambda} = \sum_{j=1}^{r_i} \tilde{a}_{ij} x_{ij\lambda}\ ,\quad \overline{X}_i = \frac{1}{n} \sum_{\lambda=1}^{n} X_{i\lambda}$$

上で定義した共分散行列を用いた相関行列$\mathbf{R}$の逆行列$\mathbf{R}^{-1}$の成分r^{ij}, r^{iy}, r^{yy}から、アイテムiと目的変数との偏相関係数$\tilde{r}_{iy}$は以下のように求められる。

$$\tilde{r}_{iy} = -r^{iy} \Big/ \sqrt{r^{ii} r^{yy}}$$

アイテムの重要性を調べるために、p個のアイテムに１つ付け加える場合を考える。全変動SV、p個のアイテムの回帰変動RV、p個のアイテムの残差変動EV、係数の数r_d、$p+1$個のアイテムの回帰変動RV'、残差変動EV'、係数の数r'_dを用いて、付け加えるアイテムの重要性のＦ値

は以下となる。

$$F = \frac{(EV - EV')/(r'_d - r_d)}{EV'/(n - r'_d - 1)} \qquad \text{自由度 } r'_d - r_d,\ n - r'_d - 1$$

また、p 個のアイテムの数量化 I 類による式の有効性の F 値は以下となる。

$$F = \frac{RV/r_d}{EV/(n - r_d - 1)} \qquad \text{自由度 } r_d,\ n - r_d - 1$$

8.2　プログラムの利用法

実際の分析メニュー画面は図 1 に与える。

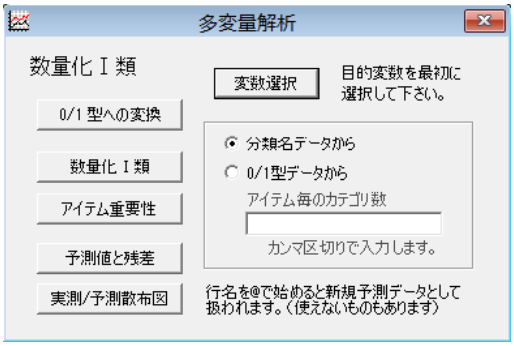


図 1 数量化 I 類メニュー画面

入力にはアイテム毎にカテゴリ名が記されているものとアイテム内をカテゴリ数に分け 0/1 で回答を表わしたものの 2 種類のデータが利用できる。もちろん 0/1 で表わされたデータには、アイテム毎のカテゴリ数を与える必要があり、テキストボックス内にカンマ区切りで入力する。コマンドボタン「0/1 型への変換」ではカテゴリ名データからもう 1 つの入力型である 0/1 型データに変換する。出力結果を図 2 に示す。

データ基本型

	販売率	地域:1	地域:2	気候:1	気候:2	気候:3
1	30	1	0	0	1	0
2	18	0	1	1	0	0
3	15	0	1	0	1	0
4	33	1	0	0	1	0
5	22	1	0	0	0	1
6	20	1	0	0	0	1
7	35	1	0	1	0	0
8	20	0	1	1	0	0
9	17	1	0	0	0	1
10	23	1	0	0	0	1

図 2 0/1 型データへの変換

カテゴリウェイトと基準化されたカテゴリウェイトの値はコマンドボタン「カテゴリウェイト」を

クリックすることによって得られる。また、これらの値による予測値から得られる重相関係数と寄与率も与えられる。出力画面は図 3 に示す。

カテゴリウェイト

	地域:1	地域:2	気候:1	気候:2	気候:3	定数項
▶ カテゴリウェイト	3.5167	1.8917	0.0000	-0.3750	-1.4667	0.0000
重回帰 ウェイト	0.0000	-1.6250	0.0000	-0.3750	-1.4667	3.5167
基準化 ウェイト	0.4875	-1.1375	0.6992	0.3242	-0.7675	2.3300
重相関係数	0.9679	調整済	0.9514			
寄与率	0.9367	調整済	0.9051			
有効性F値	29.6205	自由度	3,6			
参考p値	0.0005					

図 3 カテゴリウェイト

ここでは定数項を 0 としたカテゴリウェイトの他に、各アイテムのカテゴリの影響の正負がはっきり分かる基準化カテゴリウェイトや、各アイテムの第 1 カテゴリを 0 とした重回帰ウェイトが求められる。重回帰ウェイトは 0/1 データから、第 1 カテゴリを 0 として、重回帰分析を実行した場合と同じ結果となる。有効性 F 値は、残差に正規性があるとは考えられないので、F 分布にはならず、p 値を求めることはできないが、参考のため F 分布の際の上側確率を与えている。

目的変数とアイテム間の相関行列、目的変数とアイテム間の偏相関係数、ウェイト範囲、変数の重要性の F 値等は「アイテム重要性」ボタンをクリックすることにより図 4 のように表示される。重要性 F 値についても参考のため F 分布の際の上側確率を与えている。

アイテム重要性

	販売率	地域	気候
▶ 販売率	1.0000	0.5584	0.3152
地域	0.5584	1.0000	-0.5843
気候	0.3152	-0.5843	1.0000
ウェイト範囲		1.6250	1.4667
偏相関係数		0.9642	0.9529
重要性F値		76.5861	29.6388
自由度		1,6	2,6
参考p値		0.0001	0.0008

図 4 アイテム重要性

各アイテムが目的変数をどのように予測するかを個体毎に示すアイテムの予測値は「アイテム予測値」ボタンで図 5 のように示される。変更：この結果はカテゴリウェイトに依存するので、ボタンを削除した。

アイテム予測値

	観測値	地域	気候
1	3.0	3.517	-0.375
2	1.8	1.892	0.000
3	1.5	1.892	-0.375
4	3.3	3.517	-0.375
5	2.2	3.517	-1.467
6	2.0	3.517	-1.467
7	3.5	3.517	0.000
8	2.0	1.892	0.000
9	1.7	3.517	-1.467
10	2.3	3.517	-1.467

図 5 アイテム予測値

目的変数に対する予測値と残差は「予測値と残差」ボタンで図 5 のように与えられ、その「散布図」を図 6 に示す。

予測値と残差

	観測値	予測値	残差
1	3.0	3.142	-0.142
2	1.8	1.892	-0.092
3	1.5	1.517	-0.017
4	3.3	3.142	0.158
5	2.2	2.050	0.150
6	2.0	2.050	-0.050
7	3.5	3.517	-0.017
8	2.0	1.892	0.108
9	1.7	2.050	-0.350
10	2.3	2.050	0.250

図 6 予測値と残差

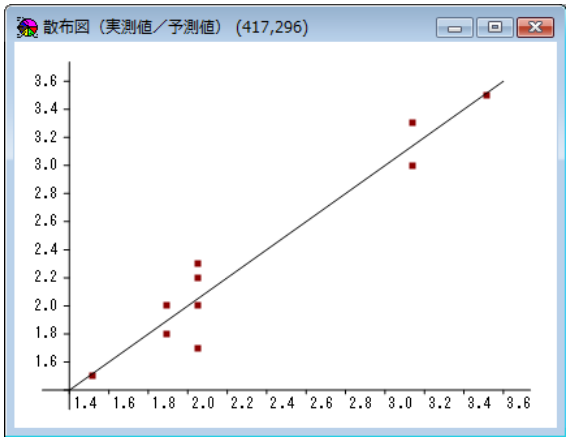


図 6 予測値と実測値の散布図

参考文献

1) 河口至商, 多変量解析入門Ⅱ, 森北出版, 1978.

2) 永田靖・棟近雅彦, サイエンス社, 2001.

９．数量化Ⅱ類

9.1　数量化Ⅱ類の理論

数量化Ⅱ類はカテゴリデータに関する線形判別関数を定義し、個体を分類することが狙いであり、判別分析に相当する。カテゴリデータで群分類を行なう数量化Ⅱ類は、群の数を m 、群 α のデータ数を n_α、アイテム数を p、アイテム i のカテゴリ数を r_i として、表１のデータ形式を元にする。

表 1 数量化Ⅱ類のデータ

	アイテム 1				アイテム p		
	カテゴリ 1	…	カテゴリ r_1	…	カテゴリ 1	…	カテゴリ r_p
群 1	x^1_{111} : $x^1_{11n_1}$	… …	$x^1_{1r_11}$: $x^1_{1r_1n_1}$	…	x^1_{p11} : $x^1_{p1n_1}$	… …	$x^1_{pr_p1}$: $x^1_{pr_pn_1}$
:	:		:		:		:
群 m	x^m_{111} : $x^m_{11n_m}$	… …	$x^m_{1r_11}$: $x^m_{1r_1n_m}$	…	x^m_{p11} : $x^m_{p1n_m}$	… …	$x^m_{pr_p1}$: $x^m_{pr_pn_m}$

一般にデータを $x^\alpha_{ij\lambda} \in \{0,1\}$ の形で表わすと、$\alpha\,(1,2,\cdots,m)$ は群、$\lambda\ (1,2,\cdots,n_\alpha)$ は個体、$i\ (1,2,\cdots,p)$ はアイテム、$j\,(1,2,\cdots,r_i)$ はアイテム毎のカテゴリである。各変数には次の関係がある。

$$\sum_{j=1}^{r_i} x^\alpha_{ij\lambda} = 1 \tag{1}$$

このため、アイテムごとに独立なカテゴリの数は１つ少なくなる。通常は第１カテゴリを除いた変数を用いて分析を実行する。

ここで、$x^\alpha_{ij\lambda}$ の表式を判別分析と類似のものとするため、新しい表記として $x^\alpha_{I\lambda}$ を導入する。この大文字の I はアイテム i 、その中のカテゴリ $j\,(=2,\cdots,r_i)$ について、順番にアイテム 1 から並べた数で、$I \equiv \sum_{k=1}^{i-1}(r_k-1)+(j-1)$ で定義される。変数 I の範囲は $I=1,2,\cdots,P \equiv \sum_{k=1}^{p}(r_k-1)$ である。この変数表記法を用いると第１カテゴリを除いた数量化Ⅱ類は判別分析と同等であることが理解し易い。以後は

$$\sum_{I=1}^{P} f_I \quad \Leftrightarrow \quad \sum_{i=1}^{p}\sum_{j=1}^{r_i} f_{ij}$$

と置き換えることによって、両者の書式を使い分けることにする。

1）マハラノビスの距離に基づく方法

新しい変数表記法 $x_{I\lambda}^{\alpha}$ でデータを見ると 0,1 型のデータであっても、判別分析と同等に扱うことができる。よってデータの判別はマハラノビスの距離に基づく方法を用いて、判別分析と同じように行うことができる。但し、データの分布は正規分布でないので、判別分析の最初のところで述べた分布関数による判別の理由付けはできない。しかし、3.3 節で述べたように、２群の場合は正準形式と同等であるので、判別関数による群間分散の最大化の方法による理由付けは説得力がある。３群以上の場合は、群間の１対比較によって判別を行うものと解釈すると、判別の問題は判別分析と全く同等に考えることができる。

２群の場合、判別分析と同じように作られた係数を用いて判別関数は以下のように与えられる。ここでは判別関数との類似性を強調するため、新しい変数表示法を用いている。

$$z=\sum_{I=1}^{P}a_I x_I-\frac{1}{2}\sum_{I=1}^{P}(\bar{x}_I^1+\bar{x}_I^2)a_I \quad , \quad a_I=\sum_{J=1}^{P}(\mathbf{S}^{-1})_{IJ}(\bar{x}_J^1-\bar{x}_J^2) \tag{2}$$

また、３群以上の場合、群 α の判別関数は以下のように与えられる。

$$z^{\alpha}=\sum_{I=1}^{P}a_I^{\alpha}x_I-\frac{1}{2}\sum_{I=1}^{P}\bar{x}_I^{\alpha}a_I^{\alpha} \quad , \quad a_I^{\alpha}=\sum_{J=1}^{P}(\mathbf{S}^{-1})_{IJ}\bar{x}_J^{\alpha} \tag{3}$$

２群の場合も３群以上の場合も、係数ベクトル a_{ij} は各アイテムの第１カテゴリを除いたものであるので、以下のような基準化された係数 d_{ij} $(i=1,\cdots,p,\ j=1,2,\cdots,r_i)$ も計算しておく。

２群の場合　$d_{ij}=\hat{a}_{ij}-\sum_{k=1}^{r_i}\tilde{x}_{ik}\hat{a}_{ik}$,　$\hat{a}_{ij}=\begin{cases}0 & j=1\\ a_{ij} & j\neq 1\end{cases}$

３群以上の場合　$d_{ij}^{\alpha}=\hat{a}_{ij}^{\alpha}-\sum_{k=1}^{r_i}\tilde{x}_{ik}\hat{a}_{ik}^{\alpha}$,　$\hat{a}_{ij}^{\alpha}=\begin{cases}0 & j=1\\ a_{ij}^{\alpha} & j\neq 1\end{cases}$

ここに基準化ウェイトの意味がカテゴリの影響が判別に正に働くか負に働くかを見ることであると考えて、以下のように、$\tilde{x}_{ik}$ はアイテム i カテゴリ k における群平均の単純平均とした。

$$\tilde{x}_{ik}=\frac{1}{m}\sum_{\alpha=1}^{m}\bar{x}_{ik}^{\alpha}$$

基準化されたカテゴリウェイトを用いると、判別関数値は以下のように与えられる。

２群の場合　$z=\sum_{i=1}^{p}\sum_{j=1}^{r_p}d_{ij}x_{ij}$ (4)

３群以上の場合　$z^{\alpha}=\sum_{i=1}^{p}\sum_{j=1}^{r_i}d_{ij}^{\alpha}x_{ij}+\sum_{i=1}^{p}\sum_{j=1}^{r_i}\tilde{x}_{ij}\hat{a}_{ij}^{\alpha}-\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{r_i}\bar{x}_{ij}^{\alpha}\hat{a}_{ij}^{\alpha}$ (5)

判別分析は変数一つひとつが独立であったが、数量化Ⅱ類の場合は、１つのアイテムが判別分析の１つの変数に対応する。その中にはいくつかのカテゴリが含まれているために、アイテムの重要性は複数のカテゴリをまとめた重要性と解釈される。そのため、アイテムの重要性をみるには、カテゴリによる判別関数値の変化幅であるウェイト範囲や以下に述べるアイテムと判別関数値との相関係数、アイテムと判別関数値との偏相関係数の値などが参照される。

アイテムと判別関数間の相関係数を次のように与える。

$$r_{ij} = s_{ij} \Big/ \sqrt{s_{ii} s_{jj}}\ ,\quad r_{iz} = s_{iz} \Big/ \sqrt{s_{ii} s_{zz}}$$

ここに、アイテムと判別関数間の共分散 s_{ij} ， s_{iz} ， s_{zz} は以下のように定義される。

$$s_{ij} = \frac{1}{n-1} \sum_{\alpha=1}^{m} \sum_{\lambda=1}^{n_\alpha} (x_{i\lambda}^{\alpha} - \overline{x}_i)(x_{j\lambda}^{\alpha} - \overline{x}_j)\ ,\quad s_{iz} = \frac{1}{n-1} \sum_{\alpha=1}^{m} \sum_{\lambda=1}^{n_\alpha} (x_{i\lambda}^{\alpha} - \overline{x}_i)(z_{\lambda}^{\alpha} - \overline{z})\ ,$$

$$s_{zz} = \frac{1}{n-1} \sum_{\alpha=1}^{m} \sum_{\lambda=1}^{n_\alpha} (z_{\lambda}^{\alpha} - \overline{z})^2$$

但し、 $x_{i\lambda}^{\alpha} = \sum_{j=1}^{r_i} \hat{a}_{ij} x_{ij\lambda}^{\alpha}$ ， $\overline{x}_i = \frac{1}{n} \sum_{\alpha=1}^{m} \sum_{\lambda=1}^{n_\alpha} x_{i\lambda}^{\alpha}$ ， $\overline{z} = \frac{1}{n} \sum_{\alpha=1}^{m} n_\alpha z^{\alpha}$ である。

変更点を明らかにするために、プログラム変更以前の定義も与えておく。

$$s_{iz} = \frac{1}{n-1} \sum_{\alpha=1}^{m} \sum_{\lambda=1}^{n_\alpha} (x_{i\lambda}^{\alpha} - \overline{x}_i)(\overline{z}^{\alpha} - \overline{z})\ ,\quad s_{zz} = \frac{1}{n-1} \sum_{\alpha=1}^{m} n_\alpha (\overline{z}^{\alpha} - \overline{z})^2\ ,\quad \overline{z}^{\alpha} = \frac{1}{n_\alpha} \sum_{\lambda=1}^{n_\alpha} z_{\lambda}^{\alpha}$$

アイテム i と判別関数との偏相関係数 $\tilde{r}_{iy}$ は、上の相関係数を用いた相関行列 $\mathbf{R}$ の逆行列 $\mathbf{R}^{-1}$ の成分 r^{ij}, r^{iz}, r^{zz} を用いて、以下のように与えられる。

$$\tilde{r}_{iz} = -r^{iz} \Big/ \sqrt{r^{ii} r^{zz}}$$

数量化Ⅱ類では２群の判別の場合、各アイテムについて判別分析と同様にその有効性の F 値を求めることができる。アイテム i の有効性の F 値は以下となる。最後の分布形は仮に変数の正規性が成り立つ場合の性質であるが、当然数量化Ⅱ類のデータでは成り立たない。参考までの仮の表示である。

$$F_i = \frac{(n_1 + n_2 - p - 1) n_1 n_2 (D^2 - D_i^2)}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 D_i^2} \sim F_{r_i - 1,\, n_1 + n_2 - p - 1} \text{ 分布}$$

ここに、 D_i^2 は両群のカテゴリ i を除いたマハラノビス距離である。

2) 正準形式に基づく方法

マハラノビス形式と同様に、判別関数は係数 $a_{ij}\ (i = 1, \cdots, p,\ j = 2, \cdots, r_i)$ と定数 z_0 を用いて以下のように与える。

$$z_\lambda = \sum_{i=1}^{p}\sum_{j=2}^{r_i} a_{ij}x_{ij\lambda} + z_0$$

この判別関数は新しい変数表記法では以下となる。

$$z_\lambda = \sum_{I=1}^{P} a_I x_I + z_0$$

この表記法では、第 1 カテゴリを除いた数量化Ⅱ類と判別分析が同等である。

我々は z_λ^α の群間の変動 s_B^2 と群別変動の合計 s^2 を以下のように定義し、群間の変動を際立たせるために、これらの分散比 $\rho = s_B^2 / s^2$ を最大化することを考える。

$$s_B^2 = \sum_{\alpha=1}^{m} n_\alpha (\overline{z}^\alpha - \overline{z})^2 \ , \quad s^2 = \sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha} (z_\lambda^\alpha - \overline{z}^\alpha)^2$$

ここに、$\overline{z}^\alpha = \dfrac{1}{n_\alpha}\sum_{\lambda=1}^{n_\alpha} z_\lambda^\alpha$ ， $\overline{z} = \dfrac{1}{n}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha} z_\lambda^\alpha$ ， $n = \sum_{\alpha=1}^{m} n_\alpha$ である。

この分散比を係数で微分することにより、判別分析と同様に以下の方程式が得られる。

$$\mathbf{Ba} = \rho\mathbf{Sa} \tag{6}$$

この方程式はデータを以下のようにまとめ、

$$\mathbf{X} = \begin{pmatrix}
x_{121}^1 & \cdots & x_{1r_1 1}^1 & \cdots & x_{p21}^1 & \cdots & x_{pr_p 1}^1 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{12n_1}^1 & \cdots & x_{1r_1 1}^1 & \cdots & x_{p2n_1}^1 & \cdots & x_{pr_p n_1}^1 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{121}^m & \cdots & x_{121}^m & \cdots & x_{p21}^m & \cdots & x_{pr_p 1}^m \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{12n_m}^m & \cdots & x_{12n_m}^m & \cdots & x_{p2n_m}^m & \cdots & x_{pr_p n_m}^m
\end{pmatrix}$$

$$\overline{\mathbf{X}}_B = \begin{pmatrix}
\overline{x}_{12}^1 & \cdots & \overline{x}_{1r_1}^1 & \cdots & \overline{x}_{p2}^1 & \cdots & \overline{x}_{pr_p}^1 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\overline{x}_{12}^1 & \cdots & \overline{x}_{1r_1}^1 & \cdots & \overline{x}_{p2}^1 & \cdots & \overline{x}_{pr_p}^1 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\overline{x}_{12}^m & \cdots & \overline{x}_{12}^m & \cdots & \overline{x}_{p2}^m & \cdots & \overline{x}_{pr_p}^m \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\overline{x}_{12}^m & \cdots & \overline{x}_{12}^m & \cdots & \overline{x}_{p2}^m & \cdots & \overline{x}_{pr_p}^m
\end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} n_1 \\ \vdots \\ \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} n_m \end{matrix}$$

$$\bar{\mathbf{X}}=\begin{pmatrix} \bar{x}_{12} & \cdots & \bar{x}_{1r_1} & \cdots & \bar{x}_{p2} & \cdots & \bar{x}_{pr_p} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_{12} & \cdots & \bar{x}_{1r_1} & \cdots & \bar{x}_{p2} & \cdots & \bar{x}_{pr_p} \end{pmatrix} \Bigg\} n$$

方程式中の行列を以下のように定義することによって得られる。

$${}^t\mathbf{a}=(a_{12} \quad \cdots \quad a_{1r_1} \quad \cdots \quad a_{p2} \quad \cdots \quad a_{pr_p})$$

$$\mathbf{S}=\frac{1}{n-m}{}^t(\mathbf{X}-\bar{\mathbf{X}}_B)(\mathbf{X}-\bar{\mathbf{X}}_B),\quad \mathbf{B}=\frac{1}{n-m}{}^t(\bar{\mathbf{X}}_B-\bar{\mathbf{X}})(\bar{\mathbf{X}}_B-\bar{\mathbf{X}})$$

ここに n はすべての群のデータ数の合計、m は群の数である。

方程式 (6) は正準判別分析と同様の方法で変形され、以下となる。

$$\mathbf{Au}=\rho\mathbf{u} \tag{7}$$

ここに、$\mathbf{A}=\mathbf{F}^{-1}\mathbf{B}\,{}^t\mathbf{F}^{-1}$，$\mathbf{u}={}^t\mathbf{Fa}$、また $\mathbf{F}$ は $\mathbf{S}=\mathbf{F}\,{}^t\mathbf{F}$ となる下三角行列である。

(7) 式の第 r 固有値に対する規格化された固有ベクトル $\mathbf{u}^{(r)}$ を使って、係数は $\mathbf{a}^{(r)}={}^t\mathbf{F}^{-1}\mathbf{u}^{(r)}$ となり、これにより判別関数は以下となる。

$$z^{(r)}=\sum_{I=1}^{P}a_I^{(r)}x_I-\sum_{I=1}^{P}a_I^{(r)}\tilde{x}_I \tag{8}$$

ここで定数項については、正準判別分析と同様に、各固有値に対応する判別関数の群別平均の単純平均が 0 になるようにしている。

係数 $a_{ij}^{(r)}$ は各アイテムの第1カテゴリを除いたものであるので､以下のような基準化した係数 $d_{ij}^{(r)}$ $(i=1,\cdots,p,\ j=1,2,\cdots,r_i)$ も計算しておく。

$$d_{ij}^{(r)}=\hat{a}_{ij}^{(r)}-\sum_{k=1}^{r_i}\hat{a}_{ik}^{(r)}\tilde{x}_{ik},\qquad \hat{a}_{ij}^{(r)}=\begin{cases}0 & j=1\\ a_{ij}^{(r)} & j\neq 1\end{cases}$$

ここに基準化ウェイトの意味を考えて、$\tilde{x}_{ik}$ は判別関数のときと同様に、アイテム i カテゴリ k における群平均の単純平均とした。

$$\tilde{x}_{ik}=\frac{1}{m}\sum_{\alpha=1}^{m}\bar{x}_{ik}^{\alpha}$$

基準化されたカテゴリウェイトを用いると、判別関数は以下のように与えられる。

$$z^{(r)}=\sum_{i=1}^{p}\sum_{j=1}^{r_i}d_{ij}^{(r)}x_{ij} \tag{9}$$

9.3　プログラムの利用法

メニュー［分析－多変量解析等－数量化Ⅱ類］を選択すると、数量化Ⅱ類のメニュー画面が図 1 のように表示される。

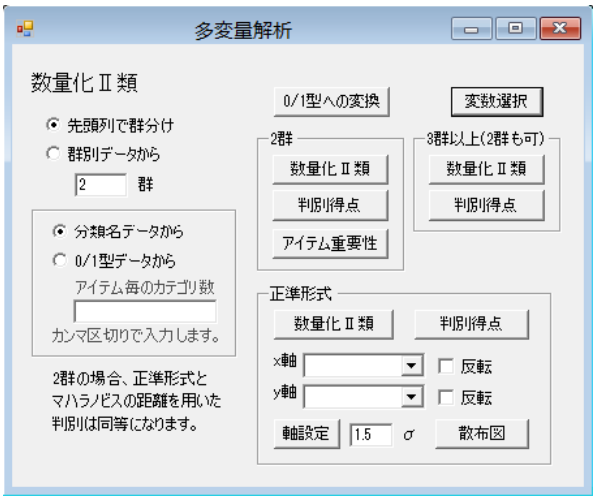
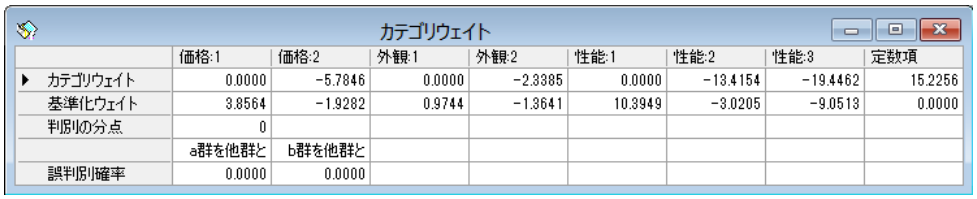


図 1 数量化Ⅱ類分析画面

データは先頭列で群分けを行なう場合と既に群別になっている場合が取り扱えるが、群別データからの場合は群の数を入力する必要がある。データの形式は各アイテムについてカテゴリ名を与える場合とカテゴリが既に 0/1 データとして分けられている場合があるが、0/1 データの場合、各アイテムのカテゴリ数をカンマ区切りで入力しなければならない。また、計算方式としては、上部に示された、参考文献 3）で与えられるマハラノビス形式と下部に示された、参考文献 4）で与えられる正準形式のどちらかを選択できる。正準形式は、これまでの計算結果を踏襲するものであるが、定義の違いから、係数について定数倍の違いがある。しかし、判別結果については同じである。マハラノビス形式は、2 群の場合、判別分析のところで示したように、正準形式と定数倍の違いを除いて同じである。しかし、3 群以上の場合では大きく異なり、判別分析と同様の結果を出力する。マハラノビス形式の結果は、各カテゴリの第 1 アイテムを除いた変数で判別分析を行った結果と一致する。我々はまず、2 群の場合の結果を比較して、3 群の場合の違いを見ることにする。

「数量化Ⅱ類」コマンドボタンをクリックした結果を比較する。マハラノビス形式の結果を図 2a に、正準形式の結果を図 2b に与える。



カテゴリウェイト

	価格:1	価格:2	外観:1	外観:2	性能:1	性能:2	性能:3	定数項
▶ カテゴリウェイト	0.0000	-5.7846	0.0000	-2.3385	0.0000	-13.4154	-19.4462	15.2256
基準化ウェイト	3.8564	-1.9282	0.9744	-1.3641	10.3949	-3.0205	-9.0513	0.0000
判別の分点	0							
	a群を他群と	b群を他群と						
誤判別確率	0.0000	0.0000						

図 2a　マハラノビス形式のカテゴリウェイト

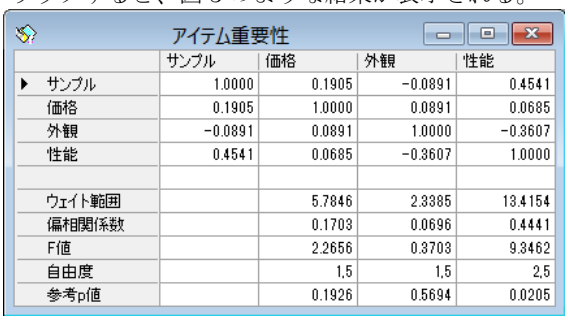
カテゴリウェイト

	価格:1	価格:2	外観:1	外観:2	性能:1	性能:2	性能:3	定数項
▶ 判別1	0.0000	-1.4636	0.0000	-0.5917	0.0000	-3.3943	-4.9202	3.8524
基準化1	0.9757	-0.4879	0.2465	-0.3451	2.6301	-0.7642	-2.2901	0.0000
	固有値	寄与率	累積寄与率					
判別1	4.6862	1.0000	1.0000					
判別の分点	0							
	a群を他群と	b群を他群と						
誤判別確率	0.0000	0.0000						

図 2b　正準形式のカテゴリウェイト

ここではカテゴリウェイト、基準化されたカテゴリウェイト、判別の分点、誤判別確率が表示される。2 群の判別の場合、判別の分点は 0 にしている。2 つのカテゴリウェイトはそれぞれ比例している。正準形式の場合は、固有値と寄与率、累積寄与率が表示されるが、2 群の場合、寄与率と累積寄与率は定義より 1 になる。

2 群の場合、2 つの方法は同等であるので、以後はマハラノビス形式の結果のみを表示する。「アイテム重要性」ボタンをクリックすると、図 3 のような結果が表示される。



アイテム重要性

	サンプル	価格	外観	性能
▶ サンプル	1.0000	0.1905	-0.0891	0.4541
価格	0.1905	1.0000	0.0891	0.0685
外観	-0.0891	0.0891	1.0000	-0.3607
性能	0.4541	0.0685	-0.3607	1.0000
ウェイト範囲		5.7846	2.3385	13.4154
偏相関係数		0.1703	0.0696	0.4441
F値		2.2656	0.3703	9.3462
自由度		1,5	1,5	2,5
参考p値		0.1926	0.5694	0.0205

図 3 アイテム重要性

ここでは、相関行列とそれを元に計算される偏相関係数及びアイテム毎のカテゴリウェイトの最大と最小の差であるウェイト範囲が表示される。ウェイト範囲は各アイテムの重要性を見るのに用いられる。またアイテムの重要性を示す F 値等も表示される。データに正規性がないために、F 値の確率は参考 p 値として表示してある。

図 4 は「判別得点」をクリックした場合の結果を表わしている。各個体が元々所属する群とその個体の数量化された値が示される。判別の助けとなるように各群の判別得点の平均や 2 群の場合は判別の分点も示されている。

判別得点

	所属群	判別得点	予測群
1	a	1.8103	a
2	a	9.4410	a
3	a	12.8872	a
4	a	7.1026	a
5	b	-4.2205	b
6	b	-12.3436	b
7	b	-6.3128	b
8	b	-10.0051	b
9	b	-3.9744	b
10	b	-10.0051	b
群別得点平均	a	7.8103	
	b	-7.8103	
判別の分点		0	

図 4 判別得点

以後は 3 群以上の場合を扱う。3 群の場合、正準形式とマハラノビス形式ではかなり異なる。マハラノビス形式では群別の判別関数が出力されるのに対して、正準形式では固有値に対応する判別関数が出力される。前者はどの判別関数の値が大きいかによって判別結果を決めるが、後者は判別結果を多次元上に表示するためのものである。結果を比較して示しておく。それぞれ、図 5a と図 5b のように結果が表示される。

カテゴリウェイト

	吐き気:0	吐き気:1	吐き気:2	頭痛:0	頭痛:1	頭痛:2	定数項
1群判別関数	0.0000	3.2656	3.7813	0.0000	2.0625	1.7188	-0.5328
2群判別関数	0.0000	15.9844	20.5759	0.0000	8.9375	10.0670	-12.7114
3群判別関数	0.0000	16.3281	22.0491	0.0000	10.3125	13.3080	-15.8739
1群基準化関数	-2.5495	0.7161	1.2318	-1.2432	0.8193	0.4755	3.2599
2群基準化関数	-13.0993	2.8850	7.4766	-6.3913	2.5462	3.6757	6.7792
3群基準化関数	-13.6903	2.6379	8.3589	-8.0233	2.2892	5.2847	5.8397
	1群を他群と	2群を他群と	3群を他群と				
誤判別確率	0.2000	0.4000	0.2500				

図 5a マハラノビス距離を用いたカテゴリウェイト

カテゴリウェイト

	吐き気:0	吐き気:1	吐き気:2	頭痛:0	頭痛:1	頭痛:2	定数項
判別1	0.0000	-2.9339	-4.0012	0.0000	-1.7342	-2.3017	3.8454
判別2	0.0000	-2.0006	-1.4483	0.0000	0.3109	2.2173	0.3633
基準化1	2.4717	-0.4622	-1.5295	1.3737	-0.3605	-0.9280	0.0000
基準化2	1.3014	-0.6992	-0.1469	-0.9381	-0.6272	1.2793	0.0000
	固有値	寄与率	累積寄与率				
判別1	5.5682	0.9778	0.9778				
判別2	0.1263	0.0222	1.0000				

図 5b　正準形式を用いたカテゴリウェイト

それぞれの方法の「判別得点」をクリックした結果を図 6a と図 6b に示す。

判別得点

	所属群	1群判別得	2群判別得	3群判別得	予測群
1	1	1.5297	−3.7739	−5.5614	1
2	1	2.7328	3.2730	0.4542	2
3	1	−0.5328	−12.7114	−15.8739	1
4	1	−0.5328	−12.7114	−15.8739	1
5	1	−0.5328	−12.7114	−15.8739	1
6	2	4.4516	13.3400	13.7623	3
7	2	3.2484	7.8645	6.1752	2
8	2	4.7953	12.2105	10.7667	2
9	2	4.4516	13.3400	13.7623	3
10	2	5.3109	16.8020	16.4877	2
11	3	4.4516	13.3400	13.7623	3
12	3	4.9672	17.9315	19.4833	3
13	3	4.7953	12.2105	10.7667	2
14	3	4.9672	17.9315	19.4833	3

図 6a　マハラノビス距離を用いた判別得点

判別得点

	所属群	判別得点1	判別得点2
1	1	2.1112	0.6742
2	1	0.9115	−1.6372
3	1	3.8454	0.3633
4	1	3.8454	0.3633
5	1	3.8454	0.3633
6	2	−1.3902	0.5801
7	2	−0.1558	−1.0850
8	2	−0.8227	−1.3263
9	2	−1.3902	0.5801
10	2	−1.8900	−0.7741
11	3	−1.3902	0.5801
12	3	−2.4575	1.1324
13	3	−0.8227	−1.3263
14	3	−2.4575	1.1324
群別得点平均	1	2.9118	0.0254
	2	−1.1298	−0.4050
	3	−1.7820	0.3796

図 6b　従来の方法による判別得点

マハラノビス形式では、判別関数の値の最も大きい群に判別されることが示されているが、正準形式では判別結果は明確に示されていない。正準形式では複数の次元の判別点を見て判断を下すため。2 次元上に散布図を描画する機能が付けられている。メニューの「軸設定」で表示する次元を設定し、「散布図」ボタンにより、図 7 のように判別得点を平面上に表示する。図中の楕円は1.5σ を表す楕円である。重なった点が多いため、散布図はあまり見易いとは言えない。

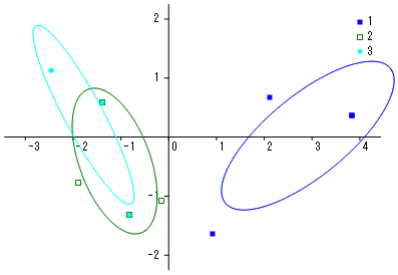


図 7　判別得点による散布図

１０．数量化Ⅲ類

10.1　数量化Ⅲ類の理論

数量化Ⅲ類はカテゴリと個体にそれぞれ数値を与えて、データの持つ類似性を解明しようとするものである。個々のデータはカテゴリに反応した場合 1、反応しない場合は 0 で与えられる。

$$x_{i\lambda} \in \{0,1\}$$

ここに、i はカテゴリ、λ は個体を表わす。また、カテゴリ数を p、データ数を n（$p \le n$）とする。

この分析では、カテゴリと個体に対してカテゴリウェイトと個体ウェイトと呼ばれる特徴的な点数 u_i と v_λ を与える。そのようにすると λ 番目の個体の i 番目のカテゴリの回答に対して、数値の組 $(u_i x_{i\lambda}, v_\lambda x_{i\lambda})$ が割り当てられる。即ち、各回答の反応した位置には数値の組 (u_i, v_λ) が割り当てられる。この反応した点を１つのデータ点と考えると、カテゴリと個体に割り当てられた数値間の散布図が得られる。各カテゴリや個体への数値の与え方によって散布図の形状は変わってくる。与えられた数値の順にカテゴリや個体を並べ替えると考えると、並べ替えによって大まかに散布図の形状を変えていると考えてもよい。似た回答をされたカテゴリや個体に属するデータ点を近くにまとめ、それと異なる回答をしたカテゴリや個体に属するデータ点を遠く離すには、この散布図の相関係数が最大になるように（データ点が直線状に並ぶように）点数を与えるとよい。数量化Ⅲ類では、このような考え方に基づき議論を進めて行く。

まず、各点の平均について考え、これが 0 になるように変数の原点を決める。即ち、以下とする。

$$\bar{u} = \frac{1}{T}\sum_{i=1}^{p}\sum_{\lambda=1}^{n} x_{i\lambda}u_i = \frac{1}{T}\sum_{i=1}^{p} c_i u_i = 0 \quad ,$$

$$\bar{v} = \frac{1}{T}\sum_{i=0}^{p}\sum_{\lambda=1}^{n} x_{i\lambda}v_\lambda = \frac{1}{T}\sum_{\lambda=1}^{n} d_\lambda v_\lambda = 0$$

$$c_i = \sum_{\lambda=1}^{n} x_{i\lambda} \quad , \qquad d_\lambda = \sum_{i=1}^{p} x_{i\lambda} \quad , \qquad T = \sum_{i=1}^{p}\sum_{\lambda=1}^{n} x_{i\lambda} \, ,$$

これによって、２変量 (u_i, v_λ) の分散、共分散は以下で与えられる。

$$S_u^2 = \frac{1}{T}\sum_{i=1}^{p}\sum_{\lambda=1}^{n} x_{i\lambda}^2 u_i^2 = \frac{1}{T}\sum_{i=1}^{p} c_i u_i^2 \quad ,$$

$$S_v^2 = \frac{1}{T}\sum_{i=1}^{p}\sum_{\lambda=1}^{n} x_{i\lambda}^2 v_\lambda^2 = \frac{1}{T}\sum_{\lambda=1}^{n} d_\lambda v_\lambda^2$$

$$S_{uv} = \frac{1}{T}\sum_{i=1}^{p}\sum_{\lambda=1}^{n} x_{i\lambda}u_i x_{i\lambda}v_\lambda = \frac{1}{T}\sum_{i=1}^{p}\sum_{\lambda=1}^{n} x_{i\lambda}u_i v_\lambda$$

これからカテゴリと個体の相関係数を $\rho = S_{uv} / S_u S_v$ と表わす。点数の分散を 1 とする制約条件

を付けて、この相関係数 ρ を最大にする点数を求めるために、Lagrange の未定乗数法を用いる。

$$L = S_{uv} - \eta\left(S_u^2 - 1\right) - \mu\left(S_v^2 - 1\right)$$

ここに η と μ は未定乗数である。これを u_i と v_λ で微分して、以下の方程式を得る。

$$\sum_{\lambda=1}^{n} x_{i\lambda} v_\lambda - 2\eta c_i u_i = 0\,,\quad \sum_{i=1}^{p} x_{i\lambda} u_i - 2\mu d_\lambda v_\lambda = 0$$

これらの式を行列で表示すると以下のようになる。

$$\mathbf{Xv} = 2\eta\mathbf{Cu}\,,\quad \mathbf{X}'\mathbf{u} = 2\mu\mathbf{Dv} \tag{1}$$

ここに

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pn} \end{pmatrix},\quad \mathbf{C} = \begin{pmatrix} c_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & c_p \end{pmatrix},\quad \mathbf{D} = \begin{pmatrix} d_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_n \end{pmatrix},$$

$$\mathbf{u}' = \begin{pmatrix} u_1 & \cdots & u_p \end{pmatrix},\quad \mathbf{v}' = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix}$$

これらの行列を用いると、以下の関係も示される。

$$\mathbf{u}'\mathbf{Cu} = TS_u^2 = T\,,\quad \mathbf{v}'\mathbf{Dv} = TS_v^2 = T\,,\quad \mathbf{u}'\mathbf{Xv} = \mathbf{v}'\mathbf{X}'\mathbf{u} = TS_{uv} = T\rho$$

(1) の方程式で、左式に左から $\mathbf{u}'$ を掛けると上の関係から、$\rho = 2\eta$、同様に右式に左から $\mathbf{v}'$ を掛けると $\rho = 2\mu$ を得る。右式を $\mathbf{v}$ について解いて左式に代入すると以下となる。

$$\mathbf{C}^{-1}\mathbf{X}\mathbf{D}^{-1}\mathbf{X}'\mathbf{u} = \rho^2\mathbf{u}\,,\quad \text{また、}\ \mathbf{v} = \rho^{-1}\mathbf{D}^{-1}\mathbf{X}'\mathbf{u} \tag{2}$$

また $\mathbf{v}$ についても対等に同様の関係が示されるが、ここでは省略する。

さて、ここで $S_u^2 = 1$ としたことから、$\mathbf{u}$ の規格化条件が $\dfrac{1}{T}\mathbf{u}'\mathbf{Cu} = 1$ となるので、新たに以下のベクトル $\mathbf{z}$ を考える。

$$\mathbf{z} = \frac{1}{\sqrt{T}}\mathbf{C}^{1/2}\mathbf{u}\,,\quad \text{ここに}\ \mathbf{C}^{1/2} = \begin{pmatrix} \sqrt{c_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{c_p} \end{pmatrix}$$

これを用いて最終的に方程式 (2) は以下となる。

$$\mathbf{Az} = \rho^2\mathbf{z}\,,\quad \mathbf{A} = \mathbf{C}^{-1/2}\mathbf{X}\mathbf{D}^{-1}\mathbf{X}'\mathbf{C}^{-1/2}\,,\ \text{規格化条件}\quad \mathbf{z}'\mathbf{z} = 1 \tag{3}$$

異なる固有値 ρ_α^2（$\alpha = 1, \cdots, p$）に対する固有ベクトルを $\mathbf{z}^\alpha$ とすると、各点数は以下のように表される。

$$\mathbf{u}^\alpha = \sqrt{T}\mathbf{C}^{-1/2}\mathbf{z}^\alpha\,,\quad \mathbf{v}^\alpha = \rho_\alpha^{-1}\sqrt{T}\mathbf{D}^{-1}\mathbf{X}'\mathbf{C}^{-1/2}\mathbf{z}^\alpha \tag{4}$$

ここでもう一度 (2) 式について考える。この方程式を成分表示すると以下となる。

$$\sum_{\lambda=1}^{n}\sum_{j=1}^{p}\frac{1}{c_i}x_{i\lambda}\frac{1}{d_\lambda}x_{j\lambda}u_j=\rho^2 u_i$$

ここで、$u_j=1$ とすると。上式は以下となる。

$$\rho^2=\sum_{\lambda=1}^{n}\sum_{j=1}^{p}\frac{1}{c_i}x_{i\lambda}\frac{1}{d_\lambda}x_{j\lambda}=\frac{1}{c_i}\sum_{\lambda=1}^{n}x_{i\lambda}\frac{1}{d_\lambda}\sum_{j=1}^{p}x_{j\lambda}=\frac{1}{c_i}\sum_{\lambda=1}^{n}x_{i\lambda}=1$$

$$v_\lambda=\sum_{j=1}^{p}\frac{1}{d_\lambda}x_{j\lambda}=1$$

即ち(2) 式には $\rho^2=1$, $\mathbf{u}=\mathbf{1}$, $\mathbf{v}=\mathbf{1}$ の自明な解が存在するが、この解は

$$\bar{u}=\frac{1}{T}\sum_{i=1}^{p}c_i u_i=1\neq 0\ ,\quad \bar{v}=\frac{1}{T}\sum_{\lambda=1}^{n}d_\lambda=1\neq 0$$

であるから、除外する。

点数 $\mathbf{u}$, $\mathbf{v}$ の与え方には、以下のように相関係数を掛ける方法もある。

$$\tilde{\mathbf{u}}^\alpha=\rho_\alpha\mathbf{u}^\alpha,\quad \tilde{\mathbf{v}}^\alpha=\rho_\alpha\mathbf{v}^\alpha$$

ここで $p\le n$ を仮定してきたが、$p>n$ の場合、先に $\mathbf{v}$ について求め、後で $\mathbf{u}$ について求めるが、方法は同様であるので省略する。

このカテゴリウェイト $\mathbf{u}^\alpha$ と個体ウェイト $\mathbf{v}^\alpha$ を用いてカテゴリ得点 $\mathbf{y}^\alpha$ と個体得点 $\mathbf{w}^\alpha$ をそれぞれ以下のように定義する場合もあるが、ここでは省略する。

$$\mathbf{y}^\alpha=\mathbf{X}\mathbf{v}^\alpha,\ \mathbf{w}^\alpha=\mathbf{X}'\mathbf{u}^\alpha$$

各成分の重要性を表すために、自明な解に対する固有値を ρ_p^2 として、これを除いて寄与率 λ_α を以下のように定義する。

$$\lambda_\alpha=\rho_\alpha^2/\sum_{\beta=1}^{p-1}\rho_\beta^2\quad(\alpha\neq p)$$

10.2　プログラムの利用法

メインメニューの中の「分析-多変量解析-数量化Ⅲ類」メニューを選択すると図 1 に示される分析メニューが表示される。

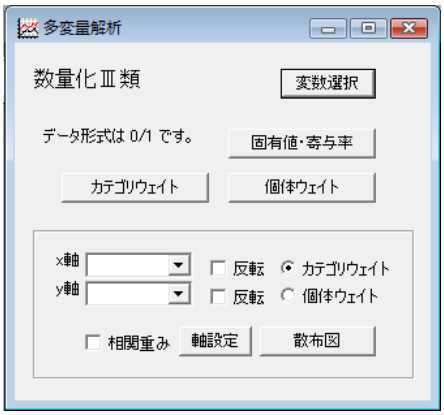
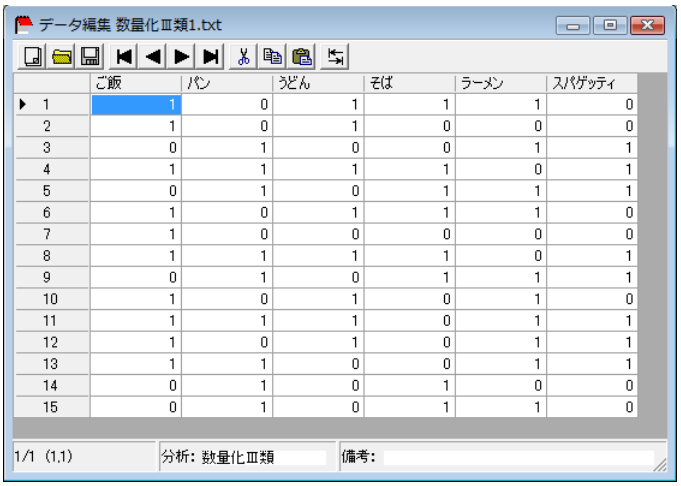


図 1 分析メニュー

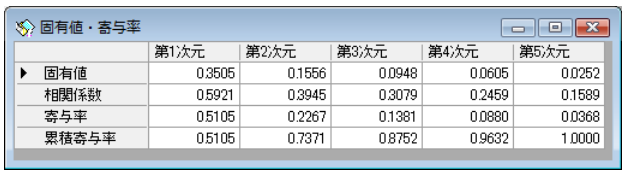
分析は図 2 のような {0, 1} の値を持つデータから実行される。



	ご飯	パン	うどん	そば	ラーメン	スパゲッティ
1	1	0	1	1	1	0
2	1	0	1	0	0	0
3	0	1	0	0	1	1
4	1	1	1	1	0	1
5	0	1	0	1	1	1
6	1	0	1	1	1	0
7	1	0	0	0	0	0
8	1	1	1	1	0	1
9	0	1	0	1	1	1
10	1	0	1	0	1	0
11	1	1	1	0	1	1
12	1	0	1	0	1	1
13	1	1	0	0	1	1
14	0	1	0	1	0	0
15	0	1	0	1	1	0

図 2 分割表データ

変数を選択して、「固有値・寄与率」ボタンをクリックすると図 3 のような結果が表示される。



	第1次元	第2次元	第3次元	第4次元	第5次元
固有値	0.3505	0.1556	0.0948	0.0605	0.0252
相関係数	0.5921	0.3945	0.3079	0.2459	0.1589
寄与率	0.5105	0.2267	0.1381	0.0880	0.0368
累積寄与率	0.5105	0.7371	0.8752	0.9632	1.0000

図 3 固有値・寄与率画面

ここで表示される固有値は、(3.2) 式のρ^2、相関係数は同じくρである。

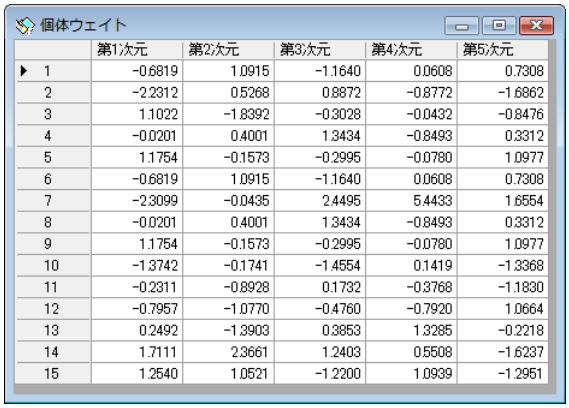
分析メニューで「カテゴリウェイト」ボタンをクリックすると図 4 のような結果が表示される。

	第1次元	第2次元	第3次元	第4次元	第5次元
ご飯	-1.3676	-0.0171	0.7543	1.3384	0.2630
パン	1.2003	-0.0615	0.8530	0.3157	-1.6177
うどん	-1.2744	0.4328	-0.2079	-1.7697	-0.7989
そば	0.8258	1.9285	-0.0892	-0.0449	1.1017
ラーメン	0.2012	-0.6217	-1.8909	0.5360	-0.1014
スパゲッティ	0.5563	-1.4935	0.7582	-0.8836	1.3151

図 4 カテゴリウェイト画面

ここでは自明な解に対応する結果は表示されていない。

分析メニューの「個体ウェイト」ボタンをクリックすると、図 5 の個体ウェイト画面が表示される。



	第1次元	第2次元	第3次元	第4次元	第5次元
1	-0.6819	1.0915	-1.1640	0.0608	0.7308
2	-2.2312	0.5268	0.8872	-0.8772	-1.6862
3	1.1022	-1.8392	-0.3028	-0.0432	-0.8476
4	-0.0201	0.4001	1.3434	-0.8493	0.3312
5	1.1754	-0.1573	-0.2995	-0.0780	1.0977
6	-0.6819	1.0915	-1.1640	0.0608	0.7308
7	-2.3099	-0.0435	2.4495	5.4433	1.6554
8	-0.0201	0.4001	1.3434	-0.8493	0.3312
9	1.1754	-0.1573	-0.2995	-0.0780	1.0977
10	-1.3742	-0.1741	-1.4554	0.1419	-1.3368
11	-0.2311	-0.8928	0.1732	-0.3768	-1.1830
12	-0.7957	-1.0770	-0.4760	-0.7920	1.0664
13	0.2492	-1.3903	0.3853	1.3285	-0.2218
14	1.7111	2.3661	1.2403	0.5508	-1.6237
15	1.2540	1.0521	-1.2200	1.0939	-1.2951

図 5 個体ウェイト画面

カテゴリウェイトや個体ウェイトを図で表示するには、まずどちらを表示するかをラジオボタンで選択し、「軸設定」ボタンをクリックして x 軸と y 軸の成分を選択する。その後、「散布図」ボタンをクリックすると図 6 や図 7 のような散布図が表示される。

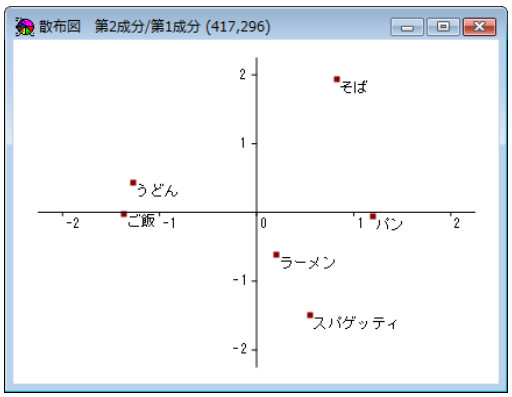


図 6 カテゴリウェイトの散布図

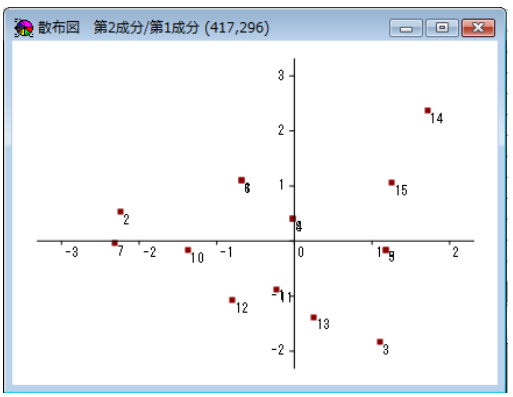


図 7 個体ウェイトの散布図

散布図の各成分には相関係数をかけて表示する場合があるが、その時には図１の「相関重み」チェックボックスにチェックを入れて散布図を表示する。また、成分を反転させて表示する場合は、反転チェックボックスにチェックを入れる。

１１．コレスポンデンス分析

11.1　コレスポンデンス分析の理論

今 2 つの質的な変数、変数 1 と変数 2 があるとする。変数 1 のカテゴリ数を p 、変数 2 のカテゴリ数を q （一般性を失わず $p \le q$ ）とする。この 2 つの変数に対して p 行 q 列の 2 次元分割表を考え、変数 1 のカテゴリ i 、変数 2 のカテゴリ j に属するデータ数を n_{ij} とする。またデータ数の合計を以下のように定義する。

$$n_{i\bullet} \equiv \sum_{j=1}^{q} n_{ij}, \quad n_{\bullet j} \equiv \sum_{i=1}^{p} n_{ij}, \quad n \equiv \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}$$

次に変数 1 のカテゴリ i のデータに点数 u_i 、変数 2 のカテゴリ j のデータに点数 v_j を与え、これらの点数の値によって各カテゴリ間の特徴的な関係を考えることとする。但し、これらの関係は変数 1 の点数と変数 2 の点数との相関係数を最大にするものとして与える。

これらの点数に対して、２つの変数の相関係数 ρ は以下のように与えられる。

$$\rho = \frac{S_{uv}}{S_u S_v} \ ,$$

$$S_{uv} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} u_i v_j, \quad S_u^2 = \frac{1}{n}\sum_{i=1}^{p} n_{i\bullet} u_i^2, \quad S_v^2 = \frac{1}{n}\sum_{j=1}^{q} n_{\bullet j} v_j^2$$

ここに、S_{uv} は共分散、S_u^2 と S_v^2 は分散であり、2 つの変数の点数について平均は 0 としている。

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{p} n_{i\bullet} u_i = 0, \quad \bar{v} = \frac{1}{n}\sum_{j=1}^{q} n_{\bullet j} v_j = 0$$

この相関係数 ρ について、点数の分散を 1 とする制約条件を付けて最大値を求めるために Lagrange の未定乗数法を用いる。

$$L = S_{uv} - \lambda\left(S_u^2 - 1\right) - \mu\left(S_v^2 - 1\right)$$

ここに λ と μ は未定乗数である。これを u_i と v_j で微分して、以下の方程式を得る。

$$\sum_{k=1}^{q} n_{ik} v_k - 2\lambda n_{i\bullet} u_i = 0, \qquad \sum_{k=1}^{p} n_{kj} u_k - 2\mu n_{\bullet j} v_j = 0$$

これらの式を行列で表示すると上式は以下のようになる。

$$\boldsymbol{N}\boldsymbol{v} = 2\lambda \boldsymbol{D}_r \boldsymbol{u}, \quad \boldsymbol{N}'\boldsymbol{u} = 2\mu \boldsymbol{D}_c \boldsymbol{v}$$

ここに

$$\boldsymbol{N} = \begin{pmatrix} n_{11} & \cdots & n_{1q} \\ \vdots & \ddots & \vdots \\ n_{p1} & \cdots & n_{pq} \end{pmatrix}, \quad \boldsymbol{D}_r = \begin{pmatrix} n_{1\bullet} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & n_{p\bullet} \end{pmatrix}, \quad \boldsymbol{D}_c = \begin{pmatrix} n_{\bullet 1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & n_{\bullet q} \end{pmatrix},$$

$$\boldsymbol{u}' = \begin{pmatrix} u_1 & \cdots & u_p \end{pmatrix}, \quad \boldsymbol{v}' = \begin{pmatrix} v_1 & \cdots & v_q \end{pmatrix}$$

上の方程式で、左式に左から $\boldsymbol{u}'$ を掛けると $\rho = 2\lambda$、同様に右式に左から $\boldsymbol{v}'$ を掛けると $\rho = 2\mu$ を得る。右式を $\boldsymbol{v}$ について解いて左式に代入すると以下となる。

$$\boldsymbol{D}_r^{-1}\boldsymbol{N}\boldsymbol{D}_c^{-1}\boldsymbol{N}'\boldsymbol{u} = \rho^2\boldsymbol{u}, \quad \text{また、} \boldsymbol{v} = \rho^{-1}\boldsymbol{D}_c^{-1}\boldsymbol{N}'\boldsymbol{u} \tag{1}$$

また $\boldsymbol{v}$ についても同様の関係が示されるが、ここでは省略する。

ここで $S_u^2 = 1$ としたことから、$\boldsymbol{u}$ の規格化条件を $\frac{1}{n}\boldsymbol{u}'\boldsymbol{D}_r\boldsymbol{u} = 1$ として、新たに以下のベクトル $\boldsymbol{z}$ を考える。

$$\mathbf{z} \equiv \frac{1}{\sqrt{n}}\boldsymbol{D}_r^{1/2}\boldsymbol{u}, \quad \text{ここに} \quad \boldsymbol{D}_r^{1/2} = \begin{pmatrix} \sqrt{n_{1\cdot}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{n_{p\cdot}} \end{pmatrix}$$

これを用いて(1)式は最終的に以下となる。

$$\boldsymbol{A}\mathbf{z} = \rho^2\mathbf{z}, \quad \mathbf{z}'\mathbf{z} = 1, \quad \boldsymbol{A} \equiv \boldsymbol{D}_r^{-1/2}\boldsymbol{N}\boldsymbol{D}_c^{-1}\boldsymbol{N}'\boldsymbol{D}_r^{-1/2} \tag{2}$$

異なる固有値 ρ_α^2（$\alpha = 1, \cdots, p$）に対する固有ベクトルを $\mathbf{z}^\alpha$ とすると、各点数は以下のように表される。

$$\boldsymbol{u}^\alpha = \sqrt{n}\boldsymbol{D}_r^{-1/2}\mathbf{z}^\alpha, \quad \boldsymbol{v}^\alpha = \rho_\alpha^{-1}\sqrt{n}\boldsymbol{D}_c^{-1}\boldsymbol{N}'\boldsymbol{D}_r^{-1/2}\mathbf{z}^\alpha$$

ところで、(1) 式には $\rho^2 = 1$, $\boldsymbol{u} = 1$ の自明な解が存在し、それに基づく固有値と固有ベクトルが得られるが、この解は除外される。

その他、点数 $\boldsymbol{u}$, $\boldsymbol{v}$ の与え方には、以下のように相関係数を掛ける方法もある。

$$\tilde{\boldsymbol{u}}^\alpha = \rho_\alpha\boldsymbol{u}^\alpha, \ \tilde{\boldsymbol{v}}^\alpha = \rho_\alpha\boldsymbol{v}^\alpha$$

各成分の重要性を表すために、自明な解に対する固有値を ρ_p^2 として、以下で与えられる寄与率 λ_α を考える場合もある。

$$\lambda_\alpha = \rho_\alpha^2 / \sum_{\beta=1}^{p-1}\rho_\beta^2 \quad （\alpha \neq p）$$

11.2 プログラムの利用法

メニュー「分析−多変量解析−コレスポンデンス分析」を選択すると図 1 に示される分析メニューが表示される。

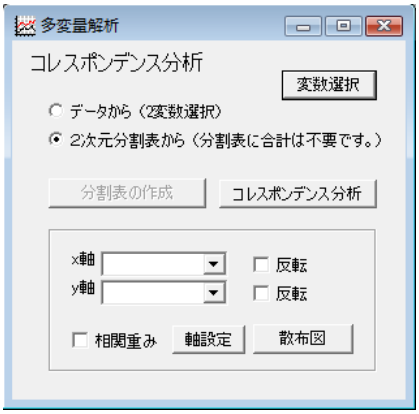


図 1 分析メニュー

分析は通常の質的データと図 2 のような分割表の 2 通りから選択できる。

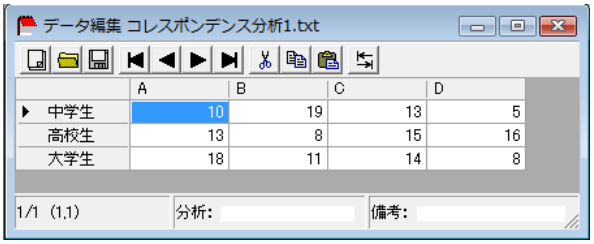


図 2 分割表データ

変数を選択して、「コレスポンデンス分析」ボタンをクリックすると図 3 のような分析結果が表示される。

	群	第1成分	第2成分	重み1成分	重み2成分
固有値		0.0763	0.0183		
相関係数		0.2762	0.1352		
中学生	1	-1.3287	-0.6528	-0.3670	-0.0883
高校生	1	1.1333	-0.7748	0.3130	-0.1048
大学生	1	0.0690	1.3916	0.0190	0.1882
A	2	0.2373	1.5238	0.0655	0.2060
B	2	-1.4691	-0.6411	-0.4058	-0.0867
C	2	0.0596	-0.1102	0.0165	-0.0149
D	2	1.5032	-1.1547	0.4152	-0.1561

図 3 コレスポンデンス分析実行結果

出力される成分数は 2 つの変数のカテゴリ数の小さい方から自明な固有値の数の 1 を引いた数であり、この例の場合 2 である。重み成分はそれぞれの成分に相関係数をかけたものである。

この結果を図の上で表示するには、まず「軸設定」ボタンをクリックし、図 11.4 のように x 軸と y 軸に表示される成分の中で適切なものを選択する。通常は x 軸に第 1 成分、y 軸に第 2 成分を表示する。「散布図」ボタンをクリックすると図 11.5 のような結果が表示される。

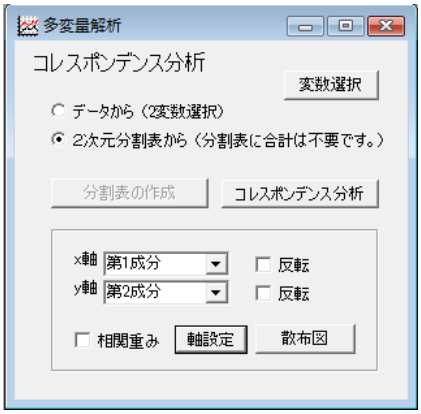


図 4 軸設定された分析メニュー

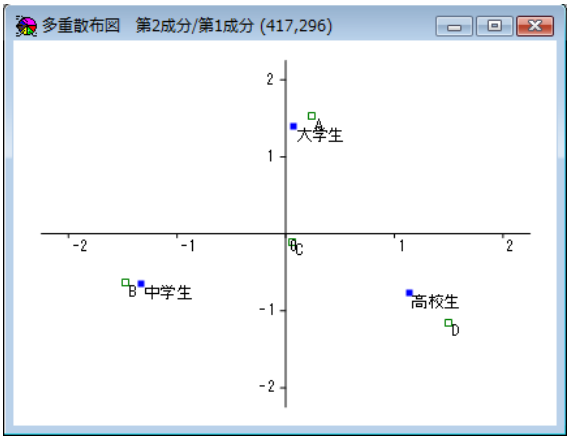
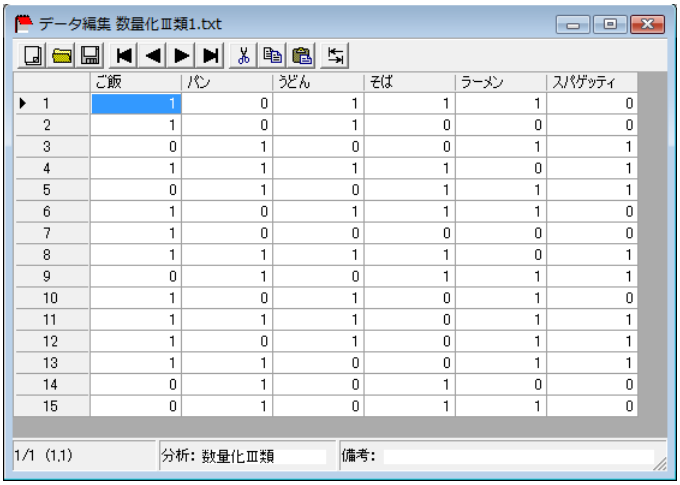


図 5 散布図画面

相関係数の重みを付ける場合は、「相関重み」チェックボックスにチェックを入れ、軸を反転させて表示したい場合は、それぞれの軸の「反転」チェックボックスにチェックを入れて散布図を表示する。

最後に、コレスポンデンス分析と数量化Ⅲ類の関係を述べておく。例えば数量化Ⅲ類の図 6 のデータを用いて 2 つの分析でカテゴリウェイトと個体ウェイトについての散布図を示しておく。コレスポンデンス分析ではこのデータを分割表として扱うが、数量化Ⅲ類の呼び名に合わせることにする。



	ご飯	パン	うどん	そば	ラーメン	スパゲッティ
1	1	0	1	1	1	0
2	1	0	1	0	0	0
3	0	1	0	0	1	1
4	1	1	1	1	0	1
5	0	1	0	1	1	1
6	1	0	1	1	1	0
7	1	0	0	0	0	0
8	1	1	1	1	0	1
9	0	1	0	1	1	1
10	1	0	1	0	1	0
11	1	1	1	0	1	1
12	1	0	1	0	1	1
13	1	1	0	0	1	1
14	0	1	0	1	0	0
15	0	1	0	1	1	0

図 7 はカテゴリウェイトについて、左が数量化Ⅲ類の結果、右がコレスポンデンス分析の結果である。同様に図 8 は個体ウェイトについての同じ結果である。

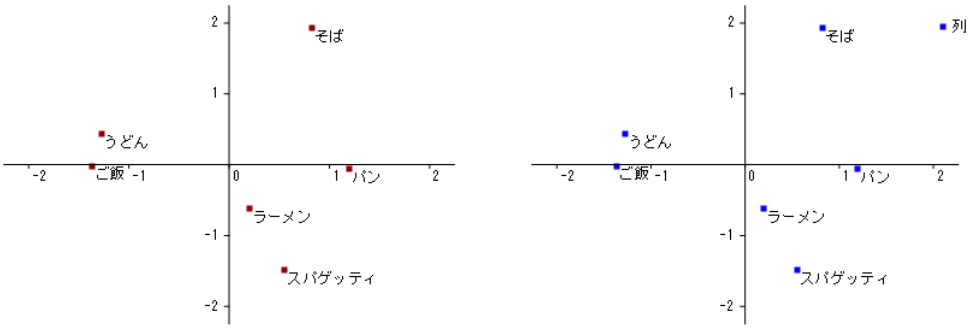


図 7　数量化Ⅲ類とコレスポンデンス分析のカテゴリウェイト

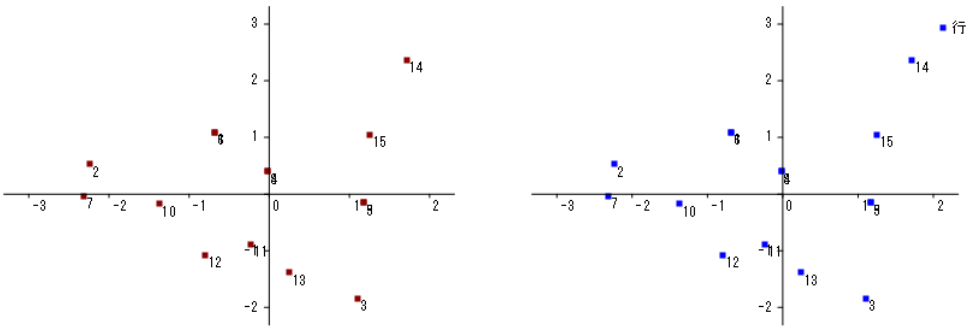


図 8　数量化Ⅲ類とコレスポンデンス分析の個体ウェイト

これらは明らかに同じものであり、0/1 データを用いる限り両者は同じである。コレスポンデンス分析ではこれ以外の分割表データも扱えるので、コレスポンデンス分析は数量化Ⅲ類の拡張になっている。