岡本安晴　2001.10；2002.8

数量化２類

グループ（群）別に項目の選択データが与えられているとき、それらのグループを区別（判別）するのに最適な項目選択肢の数量化を行うのが数量化２類である。データは、各項目（アイテム）j、$j=1,\cdots,n$、に対する選択肢（カテゴリ）1～N_jから１つk、$k=1,\cdots,N_j$、を選ぶ形式のアイテム・カテゴリ型である。岩坪（1987）では、次の表１の例があげられている。１はその選択肢が選ばれたことを、０は選ばれなかったことを表す。

表１　グループ別に与えられたアイテム（項目）・カテゴリ（選択肢）型データ

		項目１		項目２		
群	個体	選択肢１	選択肢２	選択肢１	選択肢２	選択肢３
A	1	1	0	1	0	0
	2	1	0	1	0	0
	3	1	0	0	1	0
B	1	1	0	0	1	0
	2	0	1	0	1	0
C	1	0	1	0	1	0
	2	0	1	0	0	1
	3	0	1	0	0	1

表１の形式のデータを表す変数$n_i(\nu, jk)$を次のように定義する。

個体iがグループνに属しており、かつアイテムjのカテゴリkが選ばれているとき

$$n_i(\nu, jk)=1$$

それ以外のとき

$$n_i(\nu, jk)=0$$

このとき、アイテムjのカテゴリkに数値$x(jk)$を与えると、グループνのi番目の個体の得点$y_i(\nu)$を次式で与えることができる

$$y_i(\nu)=\sum_j\sum_k x(jk)\cdot n_i(\nu, jk) \qquad (1)$$

$y_i(\nu)$ の値のグループ内での変動は小さく、グループ間の変動は大きくなるように $x(jk)$ の値を定める。すなわち、相関比の２乗

$$\eta^2 = \frac{V_B}{V_T}$$

が最大になるように $x(jk)$ の値を定める。ここで、V_B は $y_i(\nu)$ の群間分散を、V_T は全体の分散を表す。

いま、$n_i(\nu, jk)$ の項目選択肢間の分散共分散行列を A、群間分散共分散行列を B とおくとき、η^2 を極大にする $x(jk)$ は次式を満たすことが導かれる（岩坪、1987）。

$$B\mathbf{x} = \eta^2 A\mathbf{x} \qquad (2)$$

ここで、$\mathbf{x}$ は $x(jk)$ を要素とするベクトルである。

η^2 は（２）式の固有値となっているので、相関比の２乗を最大にするためには（２）式の最大固有値に対応する固有ベクトルを $x(jk)$ の値とする。複数個の解を求めるときは、（２）式の固有値のうち大きさの順に必要な個数をとり、対応する固有ベクトルを用いる。項目選択肢（アイテム・カテゴリ）の値 $x(jk)$ が決まると、（１）式より個体の得点 $y_i(\nu)$ も定まる。

解法

（２）式を満たす $\mathbf{x}$ を岩坪（1987）の式を用いて以下のように求める。

$\mathbf{A}$ の固有値分解を次式で表す。

$$\mathbf{A} = \begin{bmatrix}\mathbf{T} & \mathbf{T}_0\end{bmatrix}\begin{bmatrix}\Lambda_{\mathbf{A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{T}' \\ \mathbf{T}_0'\end{bmatrix}$$

このとき、（２）式は次のように変形できる（岩坪、(3.31）式)。

$$\mathbf{T}'\mathbf{B}\mathbf{T}\mathbf{z} = \lambda\Lambda_{\mathbf{A}}\mathbf{z} \qquad (3)$$

ここで、

$$\mathbf{z} = \mathbf{T}(1k_1, \cdots, nk_n)' \cdot \mathbf{U}(1k_1, \cdots, nk_n)' \cdot \mathbf{x}$$

である（岩坪、1987、(3.30)式)。$\mathbf{T}(1k_1, \cdots, nk_n)$ は、$\mathbf{T}$ の jk_j 行（項目 j のカテゴリ k_j に対応する行）を除いて縦に詰めた行列である。$\mathbf{U}(1k_1, \cdots, nk_n)$ は、$\sum_{j=1}^{n} N_j$ 次単位行列から jk_j 列を取り除いて横に詰めた行列において jk_j 行の項目 j に対応する要素を -1 とおいたものである。項目 j における k_j は適当に選ぶ。

（３）式を次のように変形する。

$$\Lambda_{\mathbf{A}}^{-1/2}\mathbf{T}'\mathbf{B}\mathbf{T}\Lambda_{\mathbf{A}}^{-1/2}\Lambda_{\mathbf{A}}^{1/2}\mathbf{z}=\lambda\Lambda_{\mathbf{A}}^{1/2}\mathbf{z}$$

いま

$$\mathbf{u}=\Lambda_{\mathbf{A}}^{1/2}\mathbf{z} \qquad (4)$$

とおけば

$$\Lambda_{\mathbf{A}}^{-1/2}\mathbf{T}'\mathbf{B}\mathbf{T}\Lambda_{\mathbf{A}}^{-1/2}\mathbf{u}=\lambda\mathbf{u}$$

と書けるので、λと$\mathbf{u}$は$\Lambda_{\mathbf{A}}^{-1/2}\mathbf{T}'\mathbf{B}\mathbf{T}\Lambda_{\mathbf{A}}^{-1/2}$の固有値と固有ベクトルとして求めることができる。$\mathbf{u}$が求まると、岩坪の（3.35）式

$$\mathbf{w}=\mathbf{T}\mathbf{z} \qquad (5)$$

および（3.39）式

$$x(jk)=w(jk)-w(jk_j) \qquad (6)$$

を用いて、$\mathbf{x}$を求めることができる。すなわち、式（４）と式、（５）から$\mathbf{w}$が

$$\mathbf{w}=\mathbf{T}\Lambda_{\mathbf{A}}^{-1/2}\mathbf{u}$$

と求まる。この$\mathbf{w}$から$\mathbf{x}=\left(x(jk)\right)$が（６）式より算出できる。

$\mathbf{x}$は、項目内のカテゴリに同じ値を加えても（２）式の解であるという性質がある（付録参照)。この性質によって、$x(jk)$の項目内の平均が０になるように基準化する。すなわち、

$$x^*(jk)=w(jk)-\sum_{k=1}^{N_j}\bar{n}(jk)w(jk) \qquad (7)$$

として基準化した最適解$\mathbf{x}^*=\left(x^*(jk)\right)$を求める。（７）式で与えられる解が平均値が０になるように基準化されていることは以下のように確かめられる。

$$\sum_k \bar{n}(jk)x^*(jk)=\sum_k \bar{n}(jk)w(jk)-\sum_k\left[\bar{n}(jk)\left\{\sum_{k'}\bar{n}(jk')w(jk')\right\}\right]$$

$$=\sum_k \bar{n}(jk)w(jk)-\sum_k\left[\bar{n}(jk)\left\{\sum_{k'}\bar{n}(jk')w(jk')\right\}\right]$$

$$=\sum_k \bar{n}(jk)w(jk)-\left\{\sum_{k'}\bar{n}(jk')w(jk')\right\}\sum_k\bar{n}(jk)$$

$$=\sum_k \bar{n}(jk)w(jk)-\left\{\sum_{k'}\bar{n}(jk')w(jk')\right\}\sum_k\left\{\frac{1}{m}\sum_\nu\sum_i n_i(\nu,jk)\right\}$$

$$=\sum_k \bar{n}(jk)w(jk)-\left\{\sum_{k'}\bar{n}(jk')w(jk')\right\}\left[\frac{1}{m}\sum_\nu\sum_i\left\{\sum_k n_i(\nu,jk)\right\}\right]$$

$$=\sum_k \bar{n}(jk)w(jk)-\left\{\sum_{k'}\bar{n}(jk')w(jk')\right\}\left[\frac{1}{m}\sum_\nu\sum_i 1\right]$$

$$= \sum_k \bar{n}(jk)w(jk) - \left\{ \sum_{k'} \bar{n}(jk')w(jk') \right\} \times 1$$
$$= 0$$

プログラム

プログラム PQ2a.dpr は、$x^*(jk)$、$y_i(\nu)$、および $y_i(\nu)$ のグループごとの平均値 $\overline{y(\nu)}$ を求めるものである。このプログラムを実行すると、図１のフォームが表示される。

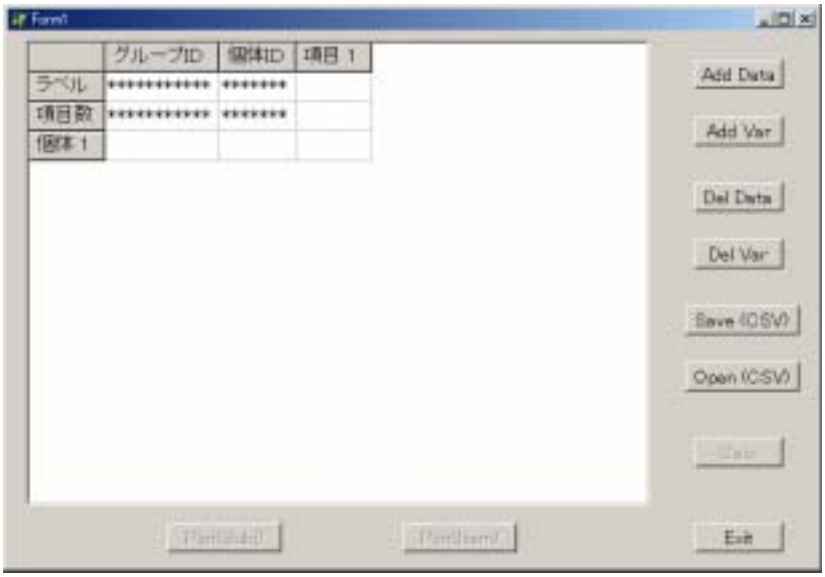


図１　起動時に表示されるフォーム

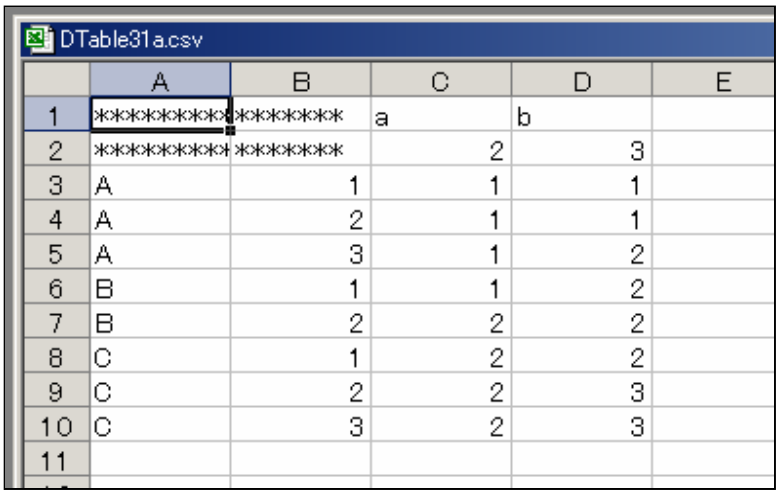
「Add Data」ボタン、および「Add Var」ボタンをクリックして行数および項目（アイテム）数をデータに合わせる。「Add Data」ボタンのクリックでアクティブなセルの下に行が追加・挿入され、「Add Var」ボタンのクリックでアクティブなセルの右側に列（項目欄）が追加・挿入される。セルはそのセルのクリックでアクティブになる。「Del Data」ボタンあるいは「Del Var」ボタンをクリックするとアクティブなセルを含む行あるいは列が削除される。追加・挿入あるいは削除の操作は、データの設定の途中においてもできる。図２はデータの設定例である。

図２　データの設定

第 2 行目（ラベル行）に各項目のラベルを設定する。設定された文字列の１バイト目の文字がマップの描画に用いられる。グループ ID と個体 ID の欄はラベルを設定しない（データとしては無視される)。第 3 行目（項目数の行）に各項目のカテゴリ数を設定する。この行もグループ ID と個体 ID の欄は数値を設定しない（無視される)。第４行目から各個体のデータを設定していく。グループ ID 欄に所属するグループを表す文字（１バイト）を設定する。この文字はマップの描画においてもグループを表すのに用いられる。個体 ID に個体のラベルを設定する。項目欄には各項目における個体の該当するカテゴリ番号を設定する。

項目 j におけるカテゴリ数が N_j であるとき、カテゴリ番号は１から N_j までの数値が当てられているものとする。図２は表１のデータを設定したものである。

図２のように設定したデータは、「Save(CSV)」ボタンのクリックで保存することができる。保存したデータは「Open(CSV)」ボタンのクリックで読み込むことができる。データは CSV 形式のファイルで保存されるので、Excel で開くこともできる。図２のデータを保存したファイルを Excel で開くと図３のように表示される。



DTable31a.csv

	A	B	C	D	E
1	**********	********	a	b	
2	**********	********	2	3	
3	A	1	1	1	
4	A	2	1	1	
5	A	3	1	2	
6	B	1	1	2	
7	B	2	2	2	
8	C	1	2	2	
9	C	2	2	3	
10	C	3	2	3	
11					

図３　Excel で開いたデータ

逆に、Excel で図３の形式によって作成したデータは、CSV 形式のファイルで保存する（拡張子を.csv として保存する）と「Open(CSV)」のクリックで読み込むことができる。

データの設定後、「Calc」ボタンをクリックすると計算が始まる。「Calc」ボタンのクリックで、先ず図４のダイアログボックスが表示される。

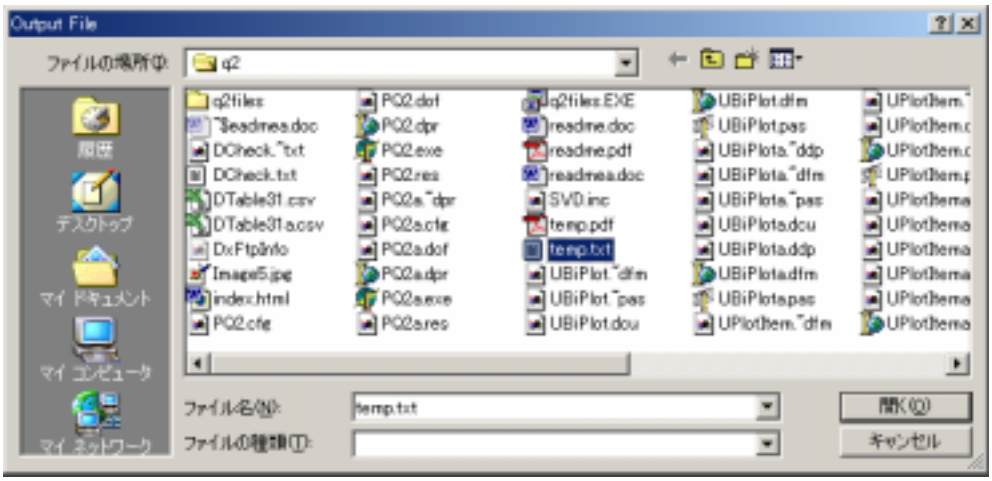


図４　計算結果の書き出し用ファイル名の設定

図４のダイアログボックスで設定した名前のファイルに計算結果がテキストファイルとして書き出される。このファイルは、プログラムの実行終了後エディタなどで開いて見ることができる。

出力ファイル名の設定後、ダイアログボックスの「開く」ボタンをクリックすると計算が始まる。計算が終了すると図５のフォームになる。

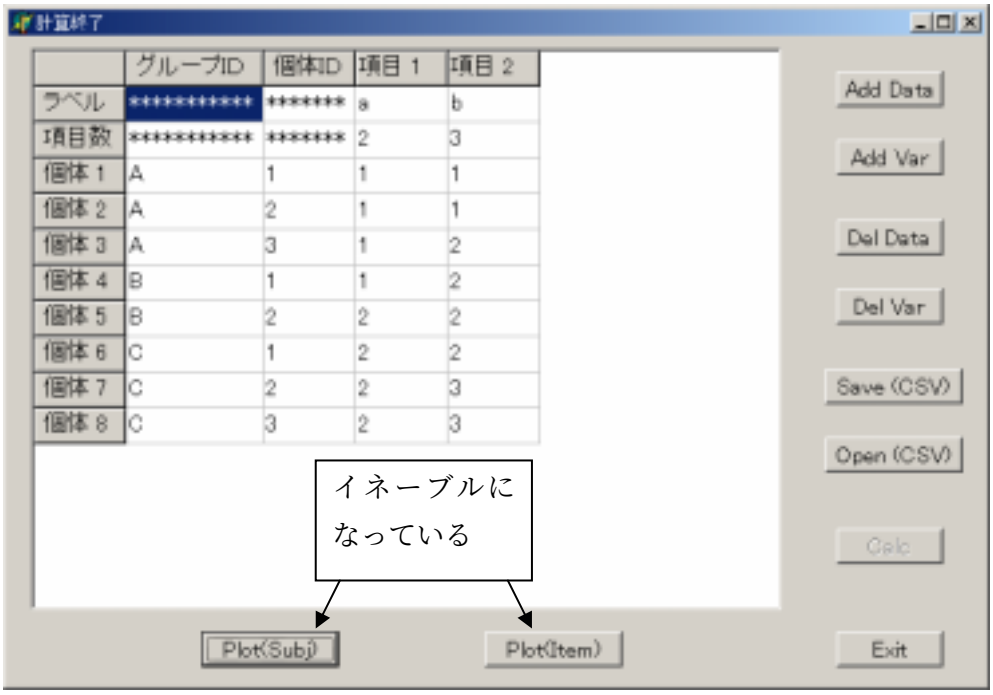


図５　計算終了時のフォーム

「Plot(Subj)」ボタンは個体の値 $y_i(\nu)$ とグループの平均値 $\overline{y(\nu)}$ を図示するものであり、「Plot(Item)」ボタンはアイテム・カテゴリ（項目・選択肢）の値 $x^*(jk)$ とグループの平均値 $\overline{y(\nu)}$ を図示するためのものである。

「Plot(Subj)」ボタンをクリックすると図６のフォームが表示される。

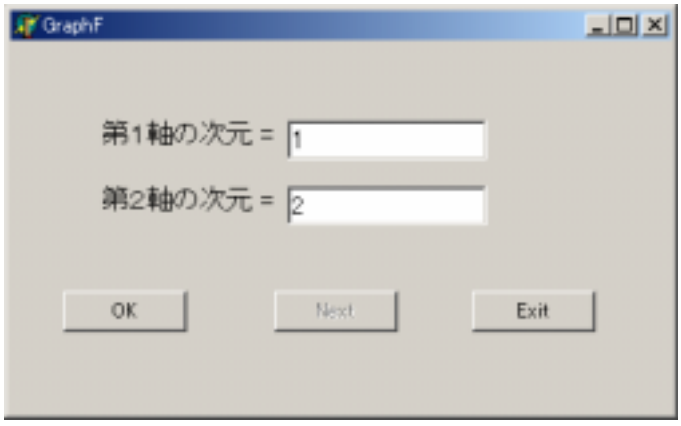


図６　描画次元の選択

個体の値 $y_i(\nu)$ とグループの平均値 $\overline{y(\nu)}$ のプロットにおいて使用する解（相関比の２乗の

大きさの順に対応する解を１軸（第１次元)、２軸（第２次元)、・・・と数える）の軸（次元）を選ぶ。軸の選択後、「ＯＫ」ボタンをクリックすると、選んだ軸の解がプロットされる（図７)。

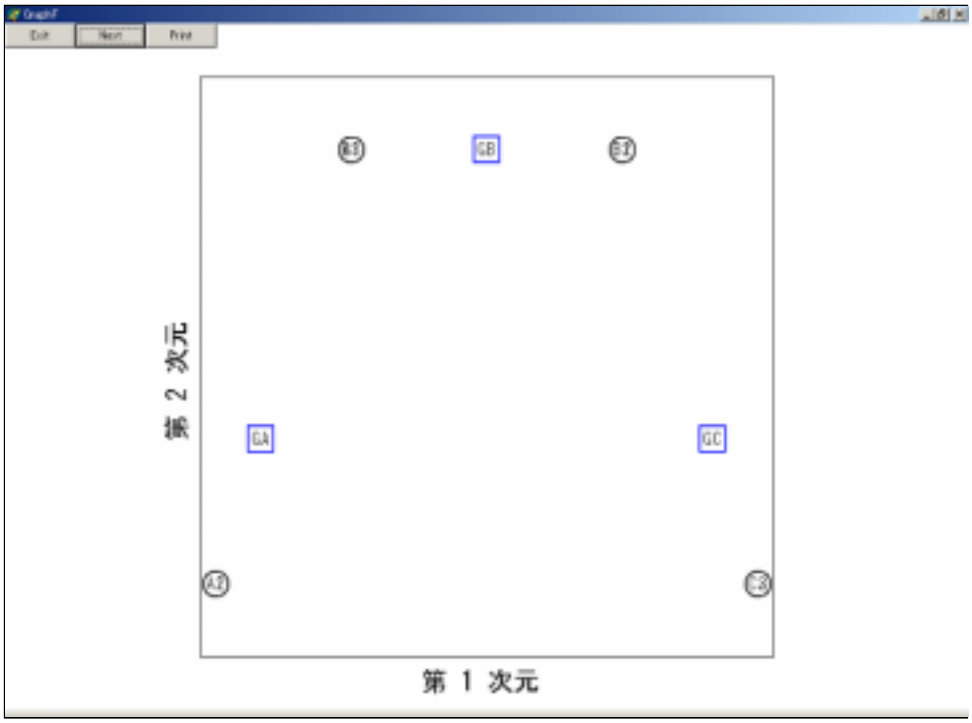


図７　個体の値とグループ平均値の描画

個体の値は小円によって表されている。小円内の先頭の文字は、グループのラベルとして入力されたものが表示されている。２番目の数値はグループ内の個人の番号である。入力データファイルに書かれている順番を表す。グループ内の個人の数が９人を超えるときは、１０人目以降は？で表される。図７の表示では、グループＡに属する個人は左側、グループＢに属する個人は上方、グループＣに属する個人は右側に位置している。グループの平均値は、小正方形で表されている。小正方形内の２文字目がそのグループを表すラベルである。それぞれのグループ内の個人を表す位置の平均位置にグループを表す小正方形が位置していることがわかる。ディスプレイに描画されている図は、図７の画面の左上のボタン「Print」をクリックするとプリンタに出力される。「Next」ボタンをクリックすると図６の描画次元の選択に戻る。描画したい次元を再設定した後、「OK」ボタンをクリックすると再設定した次元での描画が行われる。「Exit」ボタンのクリックで図５の計算終了時のフォームに戻る。

アイテム・カテゴリの位置を表示するときは、図５のフォームにおいて「Plot(Item)」ボタンをクリックする。「Plot(Item)」ボタンのクリックで図８のフォームが表示される。

図８　描画次元の選択

図６の場合と同様に、描画する軸（次元）を選び、「ＯＫ」ボタンをクリックすると選択した軸のアイテム・カテゴリ値 $x^*(jk)$ とグループの平均値 $\overline{y(\nu)}$ がプロットされる（図９）。

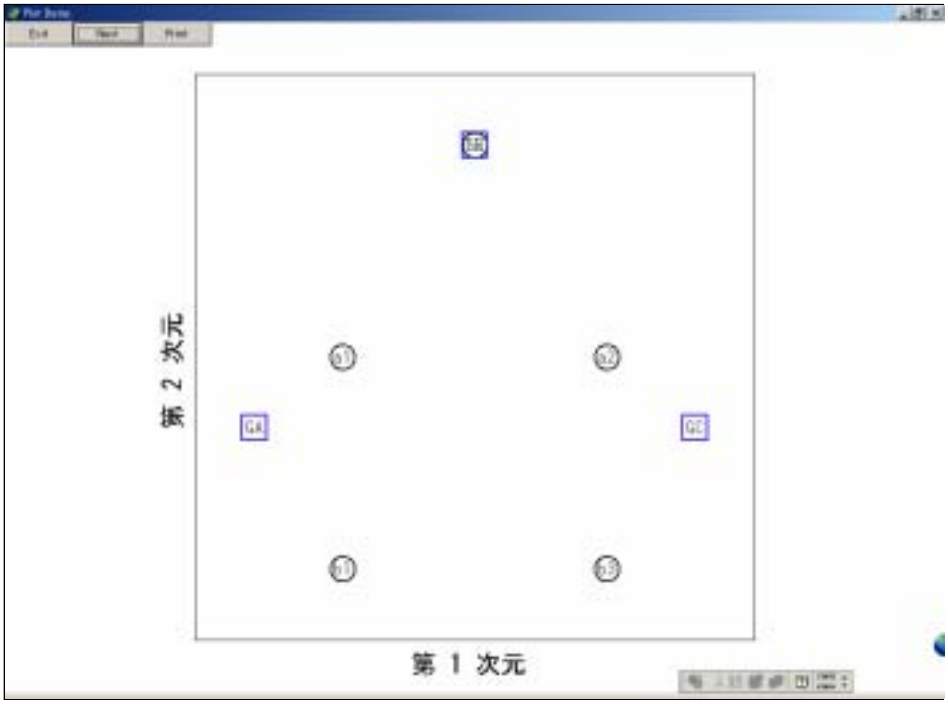


図９　アイテム・カテゴリ値とグループ平均値のプロット

グループ平均値は図７の個人の値のプロットのときと同じように描かれている。アイテム・カテゴリの値は小円によって描かれ、小円内の最初の文字がそのアイテム（項目）を表す。アイテムは、図２のデータ設定おいて項目のラベルとして設定された文字列の先頭

の１バイトが用いられる。アイテムを表す文字の右側の数値はカテゴリを表す。カテゴリはアイテム内の通し番号で１、２、・・・、N_jと数値で表している。図９から、アイテム a のカテゴリ１およびアイテム b のカテゴリ１を選ぶ者はグループＡに、アイテム a のカテゴリ２およびアイテム b のカテゴリ３を選ぶ者はグループＣに属する傾向のあることがわかる。アイテム b のカテゴリ２を選ぶ者はグループＢに属する傾向にある。これらの傾向は表１のデータにおいて確認できる。

「Exit」,「Print」、「Next」ボタンの使い方は、図７の場合と同じである。

図５のフォームにおいて「Exit」ボタンをクリックすると、プログラムの実行終了となる。プログラムの実行終了後、図４で設定した名前の計算結果の出力ファイルを開いて見ることができる。２図のデータ設定の場合、出力ファイルの内容はリスト１のようになっている。

リスト１　計算結果の出力例

```
  ***********  *******  a  b
  ***********  *******  2  3
  A  1  1  1
  A  2  1  1
  A  3  1  2
  B  1  1  2
  B  2  2  2
  C  1  2  2
  C  2  2  3
  C  3  2  3


入力データ...

グループ: A
        1    10 100
        2    10 100
        3    10 010

グループ: B
        1    10 010
        2    01 010

グループ: C
        1    01 010
        2    01 001
        3    01 001

個体総数 = 8

固有値（相関比の２乗） =
                              0.833333333333333
                              0.333333333333333

基準化された最適スコア
  a1 ->        -0.63246         0.00000
  a2 ->         0.63246        -0.00000
  b1 ->        -0.63246        -1.00000
  b2 ->         0.00000         1.00000
  b3 ->         0.63246        -1.00000
```

```
個体得点
グループ A
  1  <1>        -1.265    -1.000
  2  <2>        -1.265    -1.000
  3  <3>        -0.632     1.000
グループ B
  1  <1>        -0.632     1.000
  2  <2>         0.632     1.000
グループ C
  1  <1>         0.632     1.000
  2  <2>         1.265    -1.000
  3  <3>         1.265    -1.000

グループ平均得点
グループ A
    -1.054    -0.333
グループ B
     0.000     1.000
グループ C
     1.054    -0.333
```

リスト１を見ると、まず、グリッドのセルに設定されていたデータがそのまま書き出されている。続いて、入力データがグループ別に書き出されるが、このときは項目のカテゴリの選択は０－１のパターンで書き出されている。選択されたカテゴリの値が１、それ以外のカテゴリの値は０である。グループごとのデータの出力後、個体の総数が出力される。固有値（相関比の２乗）の書き出しに続いて、図９で表示されているアイテム・カテゴリの値が基準化された最適スコアとして書き出され、図７で表示されている個体の値が個体得点として書き出されている。最後にグループごとの平均得点がグループ平均得点として書き出されている。

付　録

$\sum_{j=1}^{n} N_j$ 次列ベクトルにおいて項目 j に対応する要素を c、それ以外の要素を 0 とおいたものを次式のように $\mathbf{d}$ とおく。

$$\mathbf{d}' = \left[0 \quad \cdots \quad 0 \quad \underbrace{c \quad \cdots \quad c}_{\text{項目}j\text{のカテゴリ}} \quad 0 \quad \cdots \quad 0 \right]$$

このとき $\mathbf{Ad}$ の $(j_1 k_1)$ 行目の要素は以下のようになる。

$$\sum_{k=1}^{N_j} a(j_1 k_1) c = \frac{c}{m} \sum_{k} \left[\sum_{\nu=1}^{g} \sum_{i=1}^{m_\nu} \{n_i(\nu, j_1 k_1) - \bar{n}(j_1 k_1)\}\{n_i(\nu, jk) - \bar{n}(jk)\} \right]$$

$$= \frac{c}{m} \sum_{\nu} \sum_{i} \left[\{n_i(\nu, j_1 k_1) - \bar{n}(j_1 k_1)\} \left\{ \sum_{k} n_i(\nu, jk) - \sum_{k} \bar{n}(jk) \right\} \right] \qquad \text{(A1)}$$

ここで、

$$\sum_{k} n_i(\nu, jk) = 1$$

および

$$\begin{aligned} \sum_{k} \bar{n}(jk) &= \sum_{k} \left\{ \frac{1}{m} \sum_{\nu'} \sum_{i'} n_{i'}(\nu', jk) \right\} \\ &= \frac{1}{m} \sum_{\nu'} \sum_{i'} \sum_{k} n_{i'}(\nu', jk) \\ &= \frac{1}{m} \sum_{\nu'} \sum_{i'} 1 \\ &= 1 \end{aligned}$$

であることに注意すると、（A1）式は

$$\frac{c}{m} \sum_{\nu} \sum_{i} \left[\{n_i(\nu, j_1 k_1) - \bar{n}(j_1 k_1)\}\{1 - 1\} \right] = 0$$

となる。すなわち、次式が成り立つ。

$$\sum_{k=1}^{N_j} a(j_1 k_1) c = 0$$

上式は $\mathbf{Ad}$ の任意の $(j_1 k_1)$ 行について成り立つ。したがって、

$$\mathbf{Ad} = \mathbf{0} \qquad \text{(A2)}$$

である。

同様に、**Bd** の (j_1k_1) 行目の要素は次のように計算できる。

$$\sum_k b(j_1k_1, jk)c = \frac{c}{m}\sum_k\left[\sum_\nu m_\nu\{\bar{n}(\nu, j_1k_1)-\bar{n}(j_1k_1)\}\{\bar{n}(\nu, jk)-\bar{n}(jk)\}\right]$$

$$= \frac{c}{m}\sum_\nu\left[m_\nu\{\bar{n}(\nu, j_1k_1)-\bar{n}(j_1k_1)\}\left[\sum_k\{\bar{n}(\nu, jk)-\bar{n}(jk)\}\right]\right] \qquad \text{（A3）}$$

ここで、次式が成り立っている。

$$\sum_k\{\bar{n}(\nu, jk)-\bar{n}(jk)\} = \sum_k\left\{\frac{1}{m_\nu}\sum_{i'}n_{i'}(\nu, jk)\right\}-\sum_k\left\{\frac{1}{m}\sum_{\nu'}\sum_{i'}n_{i'}(\nu', jk)\right\}$$

$$= \frac{1}{m_\nu}\sum_{i'}\sum_k n_{i'}(\nu, jk)-\frac{1}{m}\sum_{\nu'}\sum_{i'}\sum_k n_{i'}(\nu', jk)$$

$$= \frac{1}{m_\nu}\sum_{i'}1-\frac{1}{m}\sum_{\nu'}\sum_{i'}1$$

$$= 1-1$$

$$= 0$$

故に（A3）式は次のようになる。

$$\frac{c}{m}\sum_\nu\left[m_\nu\{\bar{n}(\nu, j_1k_1)-\bar{n}(j_1k_1)\}\times 0\right] = 0$$

すなわち、

$$\sum_k b(j_1k_1, jk)c = 0$$

が成り立つ。

上式は **Bd** の任意の (j_1k_1) 行について成り立つ。したがって、

$$\mathbf{Bd} = \mathbf{0} \qquad \text{（A4）}$$

が成り立つ。

（A2）式と（A4）式より、**x** が

$$\mathbf{Bx} = \lambda\mathbf{Ax} \qquad \text{（A5）}$$

を満たすとき、

$$\mathbf{B(x+d)} = \mathbf{Bx}+\mathbf{Bd} = \mathbf{Bx} = \lambda\mathbf{Ax} = \lambda\mathbf{Ax}+\lambda\mathbf{Ad} = \lambda\mathbf{A(x+d)}$$

が成り立つ。すなわち、$\mathbf{x}+\mathbf{d}$ も（A5）式を満たす。

参考文献

岩坪秀一「数量化法の基礎」朝倉書店、1987