社会システム分析のための統合化プログラム６

― DEA・実験計画法・クラスター分析 ―

福井正康・細川光浩

福山平成大学経営学部経営情報学科

概要

我々は主として教育での利用を目的に、社会システム分析に用いられる手法を統一的に扱うプログラムを作成してきた。今回は事業体等の効率の測定法である包絡分析法（DEA）、実験計画法及び、クラスター分析のプログラムを作成した。この論文ではこれらの分析とプログラムの利用法について説明している。

キーワード

社会システム分析，OR，統計，多変量解析，包絡分析法，DEA，実験計画法，クラスター分析，ソフトウェア，統合化プログラム

URL: http://www.heisei-u.ac.jp/~fukui/

1章　はじめに

我々はこれまで、社会システム分析で利用される手法を統合的に扱うプログラムを MS-Windows 上の Visual Basic によって開発してきたが[1-5]、この論文では新しく追加した 3 つの分析、包絡分析法[6, 7]（DEA）、実験計画法[8, 9]、クラスター分析[9-11]について説明する。

最初は、事業体を対象として、投入と産出から効率性を求める包絡分析法について述べる。効率性は基本的に産出÷投入で与えられるが、産出と投入に複数の要素がある場合、どのような式を用いたらよいのであろうか。例えば線形の式を与えるにしても、そのパラメータはどうすべきか、固定して定数とすれば効率性は決まった要素を重視することになる。これに対して包絡分析法は、ある対象の効率性を最大にするようなパラメータを選択する。即ち、得意分野を評価する分析手法である。我々はこの包絡分析法の中で凸包モデルと呼ばれる基本的なモデルをプログラム化した。具体的には CCR, BCC, IRS, DRS, GRS と呼ばれるモデル及びそれらの出力を対象としたモデルである。

実験計画法は調査や実験に影響を与える要因に関する検定方法であり、大きく分けて 1 つの要因のいくつかの水準間の比較をする 1 元比較の問題と 2 つの要因の水準間の比較をする 2 元比較の問題に分類される。またこれらの検定方法は、正規性と水準間の等分散性によって 2 つに分かれる。ここでは、1 元比較について 1 元配置分散分析と Kruskal-Wallis 検定、2 元比較について 2 元配置分散分析と Friedman 検定及び、順序による差も検出するラテン方格法をプログラム化した。さらに、個々の水準間の比較問題である多重比較の問題についても、結合された不偏分散を用いた t 検定と結合された順位を用いた Wilcoxon の順位和検定を加えている。

クラスター分析は多変量解析と呼ばれる統計分析手法の 1 つで、変数間及び個体間の距離を定義して、類似したもの同士のグループを構成し、データの構造を調べる手法である。この分析の中で我々は階層的方法と呼ばれる手法をプログラム化した。変数間及び個体間の距離には様々な定義があり、また複数の要素からなる 2 つの群を結合して 1 つの群を構成する方法にも多くの種類がある。我々はこれらの中で代表的な距離測定法やクラスター構成法を選んでプログラムに組み込んでいる。

2章　DEA

DEA (Data Envelopment Analysis) は事業体に関して、得意な分野を評価するという姿勢で、その効率性を求める手法である。ここでは効率性を検討する各事業体を DMU (Decision Making Unit) と呼び、効率は r 個の入力変数の線形結合と s 個の出力変数の線形結合の比として表わされる。今、DMU の全数を n として、DMU_i（i 番目の DMU）の入力と出力をそれぞれ、

$$ {}^t\mathbf{x}_i = (x_{i1} \quad x_{i2} \quad \cdots \quad x_{ir}), \quad {}^t\mathbf{y}_i = (y_{i1} \quad y_{i2} \quad \cdots \quad y_{is}), $$

入力と出力に掛かるパラメータをそれぞれ、

$${}^t\mathbf{v}=\begin{pmatrix}v_1 & v_2 & \cdots & v_r\end{pmatrix},\ {}^t\mathbf{u}=\begin{pmatrix}u_1 & u_2 & \cdots & u_s\end{pmatrix},$$

として、その効率を$\theta_i = {}^t\mathbf{u}\mathbf{y}_i \big/ {}^t\mathbf{v}\mathbf{x}_i$で与える。但し、効率性を計算しているDMUを$o$として、$0 \le \theta_i \le 1$の範囲で$\theta_o$を最大化するようにパラメータ$\mathbf{v}$，$\mathbf{u}$を決定する。それ故、効率性を計算するDMU毎にパラメータの値も変わってくる。このパラメータの決定方法が、最初に述べた得意な分野を評価する姿勢を表わしている。さて、ここまで述べたことを分数計画問題として以下のようにまとめておく。

分数計画問題

目的関数　$z = {}^t\mathbf{u}\mathbf{y}_o \big/ {}^t\mathbf{v}\mathbf{x}_o$　最大化

制約式　${}^t\mathbf{u}\mathbf{y}_i \big/ {}^t\mathbf{v}\mathbf{x}_i \le 1$　（$i=1,\cdots,n$），$\mathbf{u} \ge \mathbf{0}$，$\mathbf{v} \ge \mathbf{0}$

この分数計画問題は、以下の線形計画問題として考えることができる。

線形計画問題（主問題）

目的関数　$z={}^t\mathbf{u}\mathbf{y}_o$　最大化

制約式　${}^t\mathbf{v}\mathbf{x}_o = 1$，$-{}^t\mathbf{v}\mathbf{X}+{}^t\mathbf{u}\mathbf{Y} \le \mathbf{0}$，$\mathbf{u} \ge \mathbf{0}$，$\mathbf{v} \ge \mathbf{0}$

但し、$\mathbf{X}=(\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_n)$，$\mathbf{Y}=(\mathbf{y}_1 \quad \mathbf{y}_2 \quad \cdots \quad \mathbf{y}_n)$ である。

この線形計画問題は通常以下の双対問題から解が求められる。

線形計画問題（双対問題）

目的関数　$z' = \theta$　最小化

制約式　$\theta\,\mathbf{x}_o - \mathbf{X}\lambda \ge \mathbf{0}$，$-\mathbf{y}_o + \mathbf{Y}\lambda \ge \mathbf{0}$，$\lambda \ge \mathbf{0}$

ここに、双対問題の変数を、θ と ${}^t\lambda = (\lambda_1 \quad \lambda_2 \quad \cdots \quad \lambda_n)$ で与えた。ところでこの双対問題において、$\theta = 1$，$\lambda_o = 1$，$\lambda_i = 0$（$i \ne o$）は制約式を満たすので解は必ず存在することが分り、このことから必ず$\theta \le 1$となる。

さて、以下のような集合Pを考える。

$$P = \{(\mathbf{x},\mathbf{y}) \mid \mathbf{x} - \mathbf{X}\lambda \ge \mathbf{0}, -\mathbf{y} + \mathbf{Y}\lambda \ge \mathbf{0}, \lambda \ge \mathbf{0}\}$$

今、効率を測定するDMUの入力と出力$\mathbf{x}_o$，$\mathbf{y}_o$について、$(\theta\,\mathbf{x}_o, \mathbf{y}_o) \in P$であれば、双対問題の制約式を満たすことが分かる。$\mathbf{x}_o$を$\theta\,\mathbf{x}_o$として、集合$P$の境界まで縮めたときの倍率$\theta^*$が最小の目的関数値となっている。この集合を生産可能集合と呼ぶ。

θを最小化する最適解でも余剰の自由度は残る。そこで余剰$\mathbf{s}_x = \theta^*\mathbf{x}_o - \mathbf{X}\lambda$及び、$\mathbf{s}_y = -\mathbf{y}_o + \mathbf{Y}\lambda$の成分の合計が最大となるように再度線形計画問題を解く。

線形計画問題（余剰の最大化）

目的関数　$w={}^t\mathbf{e}_x\mathbf{s}_x+{}^t\mathbf{e}_y\mathbf{s}_y$　最大化

制約式　$\mathbf{s}_x = \theta^*\mathbf{x}_o - \mathbf{X}\lambda$，$\mathbf{s}_y = -\mathbf{y}_o + \mathbf{Y}\lambda$，$\lambda \ge \mathbf{0}$，$\mathbf{s}_x \ge \mathbf{0}$，$\mathbf{s}_y \ge \mathbf{0}$

ここに、${}^t\mathbf{e}_x = (1 \quad 1 \quad \cdots \quad 1)$［$r$成分］，${}^t\mathbf{e}_y = (1 \quad 1 \quad \cdots \quad 1)$［$s$成分］である。この

解$\boldsymbol{\lambda}^*, \mathbf{s}_x^*, \mathbf{s}_y^*$を最大スラック解と呼ぶ。$\theta^* = 1$，$\mathbf{s}_x^* = \mathbf{0}$，$\mathbf{s}_y^* = \mathbf{0}$のとき、観測している$DMU_o$を効率的であるといい、これ以外のとき非効率的であるという。

DMU_oの改善点を求めるために、入力の過剰量と出力の不足量$\Delta\mathbf{x}$，$\Delta\mathbf{y}$を求める。

$$\Delta\mathbf{x} = \mathbf{x}_o - \mathbf{X}\boldsymbol{\lambda}^* = (1-\theta^*)\mathbf{x}_o + \mathbf{s}_x^*$$
$$\Delta\mathbf{y} = -\mathbf{y}_o + \mathbf{Y}\boldsymbol{\lambda}^* = \mathbf{s}_y^*$$

これによって効率性改善の示唆を得ることができる。

対象となるDMUの特徴と改善点を考える際に、似たDMUで自分より優れたものを知ることは意味がある。DMU_oが非効率的であるとき、以下のE_oをDMU_oに対する優位集合という。

$$E_o = \{ j \mid \lambda_j^* > 0,\ j = 1, \cdots, n \}$$

優位集合に属する活動は効率的であることが知られている。

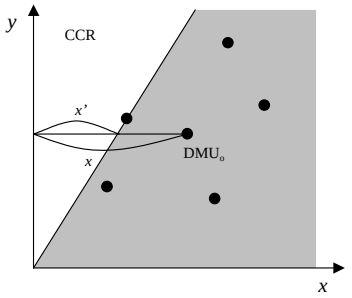

図 2.1 CCR モデルの生産可能集合

生産可能集合を直感的に理解するために1入力、1出力の場合を図で表わしてみる。この場合、生産可能集合は以下となる。

$$P = \{(x, y) \mid x \geq x_1\lambda_1 + \cdots + x_n\lambda_n,\ y \leq y_1\lambda_1 + \cdots + y_n\lambda_n,\ \boldsymbol{\lambda} \geq \mathbf{0}\}$$

この範囲を図で表わすと、図2.1の網掛けの部分になる。また、DMU_oの効率θ^*は図2.1に示したx座標の値を用いて、$\theta^* = x'/x$で与えられる。

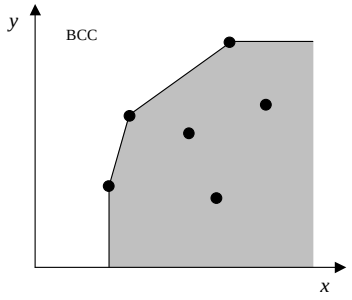

図 2.2 BCC モデルの生産可能集合

様々な基本的なモデルはこの生産可能集合に以下の条件を付けて得られる。

$$L \leq {}^t\mathbf{e}\boldsymbol{\lambda} \leq U \quad (0 \leq L \leq 1, U \geq 1)$$

CCR モデル（$L = 0, U = \infty$）

元々の生産可能集合による効率決定モデルをCCRモデルと呼ぶ。これは生産規模によって効率に優劣が生じない、規模の収穫が一定のモデルである。

BCC モデル（$L = 1, U = 1$）

生産可能集合に${}^t\mathbf{e}\boldsymbol{\lambda} = 1$の条件を付けたものがBCCモデルである。この生産可能集合は図2.2で表わされる。

IRS モデル（$L = 1, U = \infty$）

生産可能集合に${}^t\mathbf{e}\boldsymbol{\lambda} \geq 1$の条件を付けたものがIRS（Increasing Returns to Scale）モデルで、規模の収穫が増

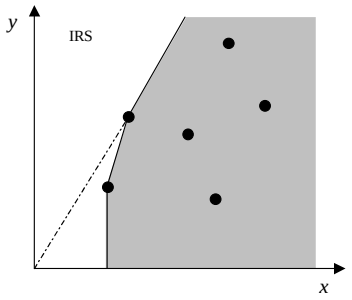

図 2.3 IRS モデルの生産可能集合

加することを想定したモデルである。この生産可能集合は図 2.3 で表わされる。

DRS モデル（$L=0, U=1$）

生産可能集合に ${}^t\mathbf{e}\boldsymbol{\lambda} \le 1$ の条件を付けたものが DRS（Decreasing Returns to Scale）モデルで、規模の収穫が減少することを想定したモデルである。この生産可能集合は図 2.4 で表わされる。

GRS モデル（$0 \le L \le 1, U \ge 1$）

下限と上限に上記の範囲で任意の値を取ったものを GRS（General Returns to Scale）モデルという。これは一般的なモデルで、利用者が L と U の値を与える。

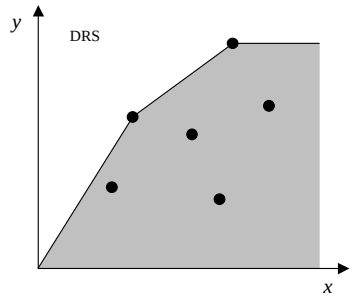

図 2.4 DRS モデルの生産可能集合

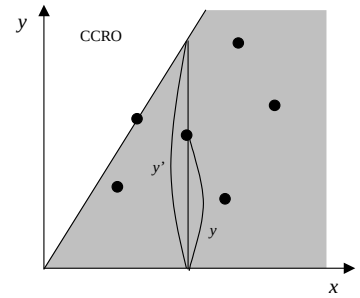

図 2.5 CCRO モデルの生産可能集合

各モデルごとに、効率の測り方として、図 2.5 のように y 軸を用いた方法も考えられる。これは出力型モデルと呼ばれ、それぞれのモデル名の後に O という文字を付け、例えば CCRO モデルのように表わす。CCRO モデルの場合、線形計画問題は以下のように与えられる。

線形計画問題（主問題）

目的関数　$z={}^t\mathbf{v}\mathbf{x}_o$　最小化

制約式　$ {}^t\mathbf{u}\mathbf{y}_o = 1,\ -{}^t\mathbf{v}\mathbf{X}+{}^t\mathbf{u}\mathbf{Y} \le \mathbf{0},\ \mathbf{u} \ge \mathbf{0},\ \mathbf{v} \ge \mathbf{0}$

線形計画問題（双対問題）

目的関数　$z' = \eta$　　最大化

制約式　$\mathbf{x}_o - \mathbf{X}\boldsymbol{\lambda} \ge \mathbf{0},\ -\eta\mathbf{y}_o + \mathbf{Y}\boldsymbol{\lambda} \ge \mathbf{0},\ \boldsymbol{\lambda} \ge \mathbf{0}$

出力型モデルの効率は $1/\eta$ で与えられる。特に CCRO モデルの場合に限り、効率は CCR モデルと一致する。他のモデルでは双対問題に $\boldsymbol{\lambda}$ についての制約が付く。

実際のプログラム実行画面は図 2.6 に示される。利用されるデータは、通常の統計分析のデータと同じ、フィールドとレコードによって表わされる形式のものである。「変数選択」により、どの変数を使用するかを指定し、入力変数の個数を入力する。但し、変数選択

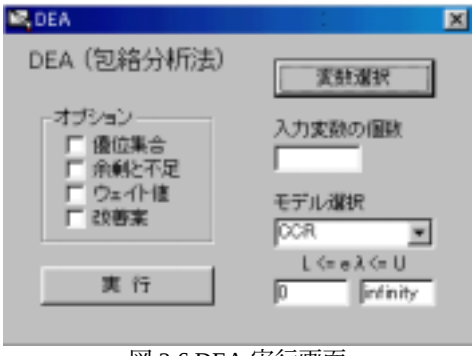

図 2.6 DEA 実行画面

の順番は、入力変数を先に、出力変数を後に選ばなければならない。出力変数の個数は、全部の変数の数から入力変数の数を引いたものとして認識される。

モデルとしては、CCR, BCC, IRS, DRS, GRS モデルとそれぞれの出力モデル CCRO, BCCO, IRSO, DRSO, GRSO モデルが用意されている。GRS モデルの場合、変数 λ についての制約式の上限と下限を入力しなければならない。「実行」ボタンをクリックすると分析結果が表示されるが、表示のオプションとして、優位集合の表示、余剰と不足（スラック変数の値）、ウェイト値（$\mathbf{v}, \mathbf{u}$ の値）、改善案がある。図 2.7 にこれらをすべて指定した場合の出力結果の例を示す。改善案については別のテキスト画面に１つの提言が示される。図 2.8 にその例を示す。

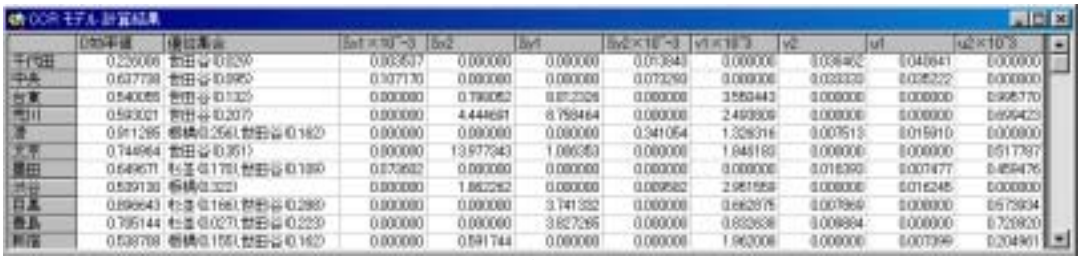

図 2.7 結果表示

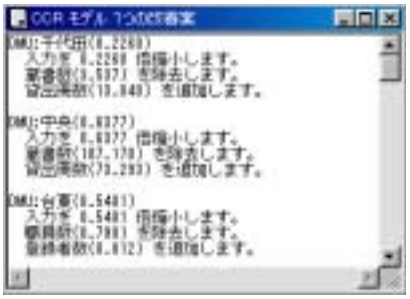

図 2.8 改善案の提示

３章　実験計画法

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 3.1 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。

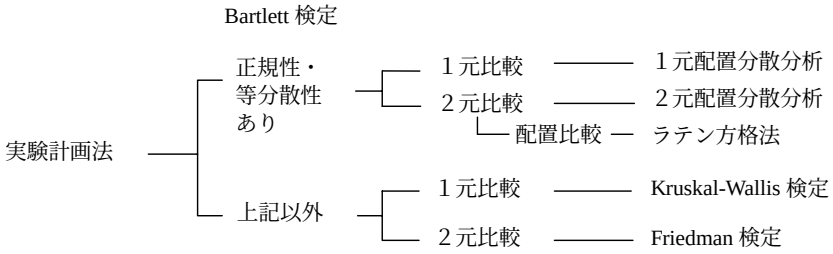

図 3.1 実験計画法の全体像

１元配置分散分析

１元比較の場合、データは表 3.1 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。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。

表 3.1　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}

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。

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

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。

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

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。

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

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。

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

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

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[9]。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 3.2　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_r	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 3.2 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。

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

ここにデータ数に関しては以下の記法を用いている。

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

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、

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

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

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。データは表 3.3 のようにブロックと水準の交点に 1 つだけ値が入る。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。

表 3.3　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}

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

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）

Friedman の順位検定

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。

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

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 3.4 にデータとその処理順序（配置と呼ぶ）の例を示す。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。

表 3.4 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	$x_{41(4)}$	$x_{42(1)}$	$x_{43(2)}$	$x_{44(3)}$

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

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

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 [8]。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 3.1 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。

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

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 3.1 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。

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

実験計画法の分析画面を図 3.2 に示す。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 3.3 に「等分散の検定」の出力画面を示す。

図 3.2 実験計画法分析画面

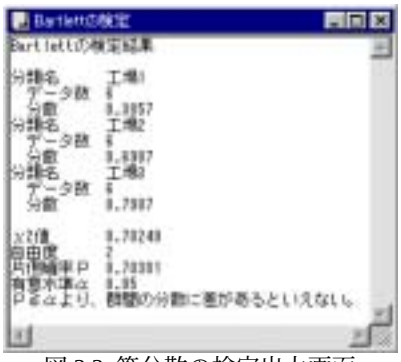

図 3.3 等分散の検定出力画面

図 3.4a と図 3.4b 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。また、図 3.5 に「Kruskal-Wallis 検定」の検定結果の出力画面を示す。

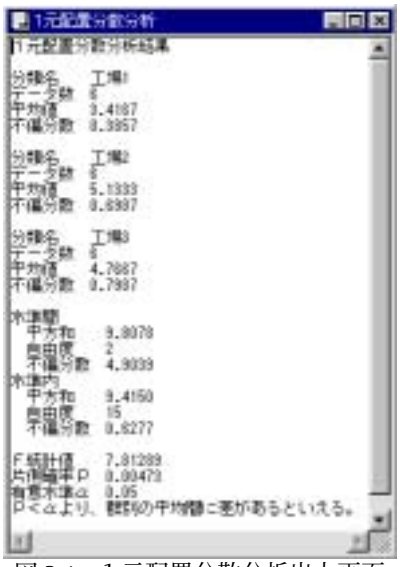

図 3.4a １元配置分散分析出力画面

図 3.4b １元配置分散分析表

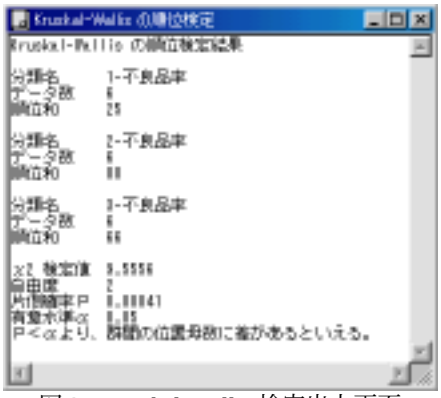

図 3.5 Kruskal-Wallis 検定出力画面

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 3.6a と図 3.6b 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。また、繰り返しがない場合の「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 3.7a と図 3.7b 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

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。

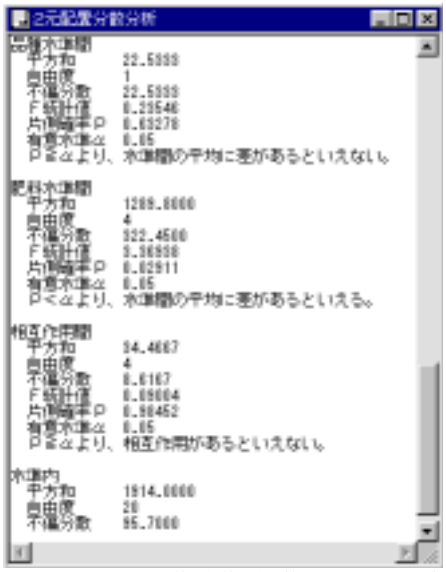

図 3.6a　２元配置分散分析（繰り返しあり）

図 3.6b　２元配置分散分析表（繰り返しあり）

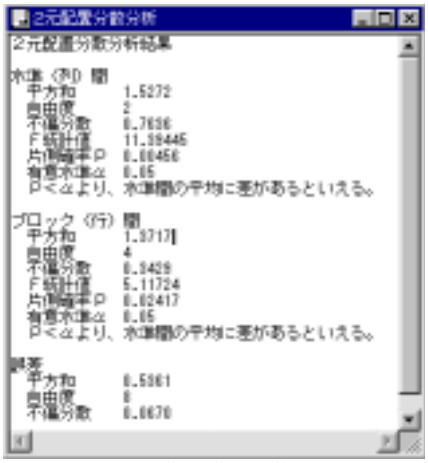

図 3.7a　２元配置分散分析（繰り返しなし）

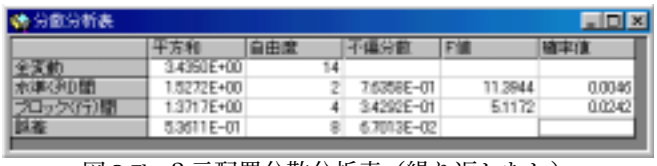

図 3.7b　２元配置分散分析表（繰り返しなし）

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 3.8 に示す。

さらに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 3.9 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 3.7a と図 3.7b 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。

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 3.10 と図 3.11 に示す。

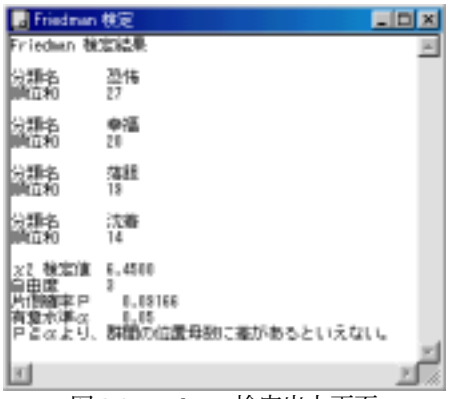

図 3.8 Friedman 検定出力画面

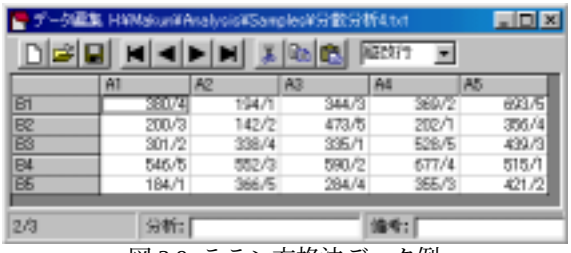

図 3.9 ラテン方格法データ例

図 3.10 pooled t 検定出力結果

図 3.11 pooled Wilcoxon 検定出力結果

4章　クラスター分析

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 4.1 のように与えられる。

表 4.1 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。

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

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

ここに、a,b,c,d は以下のように与えられる。

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

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。

実際の分析画面を図 4.1 に、「クラスター構成と距離」の出力結果を図 4.2 に、「デンドログラム」の出力結果を図 4.3 に与える。図 4.2 のような 2 つのクラスターの結合では、結合後左側のクラスター名になるものとする。即ち最初の行は、クラスターr3 とクラスターr7 が結合され、クラスターr3 になる、と読む。

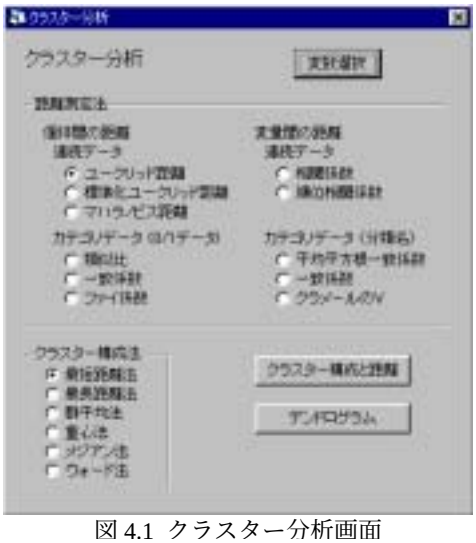

図 4.1 クラスター分析画面

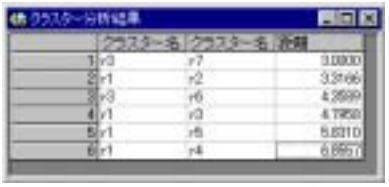
図 4.2 クラスターの構成

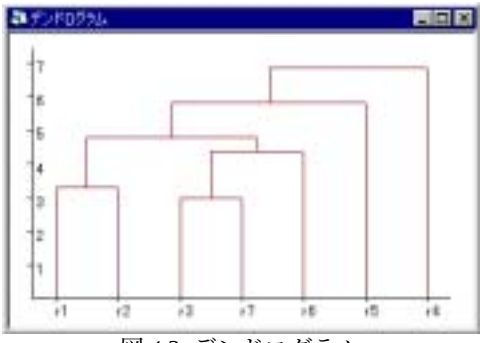
図 4.3 デンドログラム

5章　その他の変更

ここでは、これまでに作られたプログラムの拡張された機能について説明する。主な変更点

は以下の４つである。

1) 変数選択の画面に簡単な並べ替え機能を追加した。
2) 散布図とヒストグラムについて独自のグラフ表示を追加した。
3) 基本統計量，度数分布表，ヒストグラム，正規性の検定等で群分けして集計する機能を追加した。
4) AHP の構造図を簡易的に表示する機能を追加した。

1) については、図 5.1 のように選んだ変数の順序を変えるコマンド[Top],[Up],[Down]を追加し、簡単に並べ替えができるようにした。また、変数の削除も連続的にできるようにプログラムに細かい修正を加えた。また、2) については、これまで棒グラフを援用したり、散布図の表示が見にくい等の欠点があったが、もう少し標準的で見易いように、簡単なグラフ作成プログラムを標準的な MSChart とは別に作成した。これにより今後のグラフィック表示の可能性が拡がった。3) については、学生実習から得られた問題点を元に、選択変数の先頭列で群分けを行い、それぞれの処理を行う機能を追加した。これにより、データを加工することなく直接分析結果を得ることができるようになった。最後に 4) については、これまで AHP の構造図を 0, 1 行列で表わしており、初心者にはその形が分りにくいという意見があった。そのため新たに階層図を描く機能を追加した。しかし、専用ソフトのように階層図をエディターとして用いて構造を決める機能はなく、今後の課題になっている。

図 5.1 変数設定画面

６章　おわりに

今回は元々の懸案であった DEA と多変量解析を中心にプログラムを作成したが（本シリーズの７も同時に書き上げている）、これにより統計処理のソフトウェアとしてもある程度の機能を持つことができるようになった。しかし、緻密さにはまだまだ問題が残る。ここでは分析毎に不足している機能や心残りな機能を説明する。

まず DEA に関しては、このプログラムに組み込まれた分析は基本的な凸包モデルと呼ばれるものである。これに対して、制御不能変数を含む凸包モデルや階層的カテゴリデータを扱うモデル等、複雑化には様々なバリエーションが存在する。しかし、初心者向けにはここでプログラム化した分析程度が混乱を招かなくて良いのではないかと考える。その他の複雑化された分析は専門のソフトウェアの領域であろう。

実験計画法は当初の予定より分量が相当多くなった。最初は１元比較の問題のみを想定し、

１元配置分散分析、Kruskal-Wallis 検定、多重比較問題だけが分析の対象であった。しかし、実験計画法の全体像を明らかにするために２元比較の問題も含めることにした。そのためラテン方格法等データの形式に少し統一性に欠ける部分が生じている。また、多重比較の問題でも比率に関するものが取り入れられていない点等改善すべき問題もある。

クラスター分析では似た形式の様々な距離測定法があり、実際に何を用いるかはデータの種類や実務者の経験等によって選択が分かれる。また、クラスター構成法の選択方法によっても結果が大きく左右されるので、手法はできるだけ多く含めておく必要があろう。しかし、メニュー画面の大きさや見易さを最優先に考えているので、どこまで含めるかは難しい問題である。今後、必要に応じてこれらの選択肢が増えてゆくかも知れない。また、２つ以上の等距離のクラスターが存在する場合、どれを最初に結合させるかによって結果が異なってくる。この問題をどのように扱うべきか今後の課題である。

参考文献

1) 福井正康・田口賢士, 社会システム分析のための統合化プログラム, 福山平成大学経営情報研究, 3 号, 109-127, 1998.
2) 福井正康・田口賢士, 社会システム分析のための統合化プログラム２ －産業連関分析・KSIM・AHP－, 福山平成大学経営情報研究, 3 号, 129-144, 1998.
3) 福井正康・増川純一, 社会システム分析のための統合化プログラム３ －線形計画法・待ち行列シミュレーション－, 福山平成大学経営情報研究, 4 号, 99-115, 1999.
4) 福井正康, 社会システム分析のための統合化プログラム４ －基本統計－, 福山平成大学経営情報研究, 5 号, 89-100, 2000.
5) 福井正康, 社会システム分析のための統合化プログラム５ －システムの改良・ISM－, 福山平成大学経営情報学研究, 6 号, 91-104, 2001.
6) A.Charnes, W.W.Cooper and E.Rhodes, “Measuring the Efficiency of Decision Making Units”, European Journal of Operational Research, 2, 429-444, 1978.
7) 刀根薫, 経営効率性の測定と改善 －包絡分析法 DEA による－, 日科技連出版社, 1993.
8) 丹後俊郎, 新版医学への統計学, 朝倉書店, 1993.
9) 河口至商, 多変量解析入門Ⅱ, 森北出版, 1978.
10) 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, 共立出版社, 1995.
11) 田中豊・脇本和昌, 多変量統計解析法, 現代数学社, 1983.

Multi-purpose Program for Social System Analysis 6

- DEA, Method of Experimental Design, Cluster Analysis -

Masayasu FUKUI and Mitsuhiro HOSOKAWA

Department of Management Information, Faculty of Management,
Fukuyama Heisei University

Abstract

We have been constructing a unified program on social system analysis for the purpose of education. Now we added some programs on data envelopment analysis (DEA) that is used in measurement of the efficiency, method of experimental design and cluster analysis, to our system. The purpose of this paper is to explain these analyses, reformed part of our system and operation of our program.

Keywords

social system analysis, statistics, multivariate analysis, data envelopment analysis, DEA, method of experimental design, cluster analysis, software, unified program

URL: http://www.heisei-u.ac.jp/~fukui/