SPSS/PC＋による解析手法　そのII

—多重分類分析と数量化Ｉ類の関連に付いて—

小野寺　孝　義・藤　田　剛　志

問題

本稿の目的はパソコンにおける統計解析パッケージSPSS/PC＋を用いた分散分析について解説すること，及び，そこで得られる多重分類分析(Multiple Classification Analysis：ＭＣＡ)と林の数量化Ｉ類の関連を論じることにある。

SPSS/PC＋は汎用の統計パッケージソフトであり，そこに含まれる多くのプログラムで分散分析を行うことができる。条件間の平均を求めるMEANSプログラム，信頼性係数を求めるRELIABILITYプログラム，回帰分析を行うREGRESSIONプログラムなどでもオプションで分散分析結果を得ることができる。分散分析自体を直接の目的としたプログラムとしては一元配置分散分析を行うONEWAYプログラム，多元配置分散分析を行うANOVAプログラム，そして最も包括的なプログラムとして多変量分散分析を実行するMANOVAプログラムの三種類が用意されている。

このように分散分析といってもSPSS/PC＋には多くのプログラムが用意されているのであるが，全てがまったく同じものというわけではない。利用者はデータの性質や求めたい結果出力によりプログラムを選択すべきなのである。一元配置の分散分析にしてもANOVAプログラムで得られる分析モデルは母数模型(fixed effect model)なのに対して，ONEWAYプログラムでは変量模型(random effect model)が利用できる。また，ONEWAYプログラムでは多重比較のｔ検定を行えるのが大きな特徴である。一方，独立変数が複数ある場合や共分散分析を行いたい場合には，ONEWAYプログラムでは不可能であり，ANOVAプログラムを利用しなくてはならない。さらに，独立変数だけではなく，従属変数も複数ある多変量分散分析や多変量共分散分析を行おうとすればMANOVAプログラムを利用することになる。なお，繰り返しがある測度の分散分析はMANOVAプログラムで行える。ANOVAプログラムでも繰り返しがある測度を分析することもできないことはないが，これには若干の工夫を要する（Woodward & Overall, 1976 参照)。

さて，本稿ではこれらの中でもANOVAプログラムを取り上げ，その出力の多重分類分析（以下ＭＣＡと略記する）について解説していく。そして最後にＭＣＡと林の数量化Ｉ類の関連について説明する。

分散分析

基本的なANOVAプログラムを実行する場合のSPSS/PC＋のコマンド例を以下に示す。

```
DATA LIST FIXED /MENSETU NAISIN
    1-4 TEST 5-7.
BEGIN DATA
```

```
 1 1 72
 1 1 84
 1 2 65
 1 2 61
 2 1 54
 2 1 46
 2 2 48
 2 2 39
 2 2 41
END DATA.
ANOVA VARIABLES=TEST BY
  MENSETU(1,2) NAISIN(1,2)
 /STATISTICS=ALL.
FINISH.
```

この例は入学試験における面接と内申評価が入学試験の成績とどのように関わっているかを調べたものである。面接は変数MENSETU，内申は変数NAISIN，入学試験成績は変数TESTとコーディングされている。BEGIN DATAとEND DATAにはさまれた数値部分がデータであり，各行は被験者一人を，各列は変数を表している。データの第一列目は面接，第二列目は内申，最後の三列目は試験成績である。また，データの行数は九行なので九人の被験者がいることがわかる。ここでは，面接と内申評価を1と2の二段階評定としているが，これは例を簡略化するためであり，別に二値データしか扱えないという意味ではない。それぞれ五段階，七段階としても計算の考え方は同じである。

コマンドを解説すると，DATA LISTでデータの並びと変数名を指定している。はじめに変数名としてMENSETUとNAISINが示されている。次に続く数字1－4はデータの位置が1カラムから4カラムに渡っていることを示している。この場合，面接と内申の二変数なので1－4と書くと自動的に等分幅で，それぞれについて2カラムが割り当てられる。次の変数の入学成績は面接・内申と違い2カラム幅ではなく，3カラム幅なので変数名TESTの後に5－7と書いてデータの位置を新たに示す。

BEGIN DATAに続いてデータがあるが，別なデータファイルからデータを読み込むことも可能である。その場合には次に示すようにDATA LISTを変え，''内にデータファイルの存在するドライブとディレクトリ，ファイル名を書けばよい。次の例ではBドライブのルートディレクトリからBUNSAN. DATというファイル名でデータを読み込むことになる。

```
DATA LIST FILE='B: ¥BUNSAN. DAT'
  FIXED /MENSETU NAISIN 1-4 TEST
  5-7.
```

END DATAでデータの終了を示し，次のANOVAコマンドで分散分析を行うよう指示している。ANOVA VARIABLES=に続けて従属変数名，キーワードBYに続けて独立変数名，括弧内に独立変数の最小値，最大値を書く。独立変数は面接と内申の二つなので二元配置の分散分析が指定されたことになる。次の行のSTATISTICS=ALLは出力可能な分析結果を全て出力するための命令である。最後の行のFINISHコマンドで分析の終了を示している。もし，FINISHの前にその他の分析コマンドを書けば，それらのコマンドも実行されることになり，一度に複数の分析結果を得ることもできる。

分析結果の印刷出力の一部を以下に示す。

```
------------------------------------------------------------------------

                        * * *  C E L L   M E A N S  * * *

                TEST
             BY MENSETU
                NAISIN


TOTAL POPULATION

    56.67
 (     9)


MENSETU

          1           2

    70.50       45.60
 (     4)   (      5)

NAISIN

          1           2

    64.00       50.80
 (     4)   (      5)



            NAISIN

                      1           2

MENSETU

          1       78.00       63.00
             (      2)   (      2)

          2       50.00       42.67
             (      2)   (      3)

------------------------------------------------------------------------
```

図１　セル平均

はじめに平均値が全体，グループ，セルごとに出力される。括弧内の数字は人数である。

* * * ANALYSIS OF VARIANCE * * *

TEST
BY MENSETU
NAISIN

Source of Variation	Sum of Squares	DF	Mean Square	F	Signif of F
Main Effects	1635.273	2	817.636	26.095	.002
MENSETU	1248.073	1	1248.073	39.832	.001
NAISIN	257.473	1	257.473	8.217	.035
2-way Interactions	32.061	1	32.061	1.023	.358
MENSETU NAISIN	32.061	1	32.061	1.023	.358
Explained	1667.333	3	555.778	17.738	.004
Residual	156.667	5	31.333		
Total	1824.000	8	228.000		

図２　分散分析表

図２は分散分析表である。Main Effectsの下に各要因の主効果が，2-way Interactionsの下には面接と内申の交互作用が示されている。最後のResidualのところには誤差平方和が示されている。Sum of Squaresには平方和，DFには自由度，Mean Squareには平方和を自由度で割った平均平方和が示されている。各効果は自由度１なので平方和と平均平方和は一致している。誤差平方和は自由度が５なので，平均平方和は156.667/5で31.333となっている。

Ｆの表示の下にはＦ値が示されている。Ｆ値は効果の自由度と誤差自由度のもとでＦ分布に従い，これをもとに検定が行われる。Ｆ値の計算は（平均平方和／誤差平方和）で求められる。例えば，面接の効果は1248.073/31.333=39.832であるし，面接と内申の交互作用効果は32.061/31.333=1.023となる。あとはＦ分布表で対応する自由度の数値を見て，Ｆ値がその値よりも大きいか否かを検討することで有意差を知ることができる。もちろん，SPSS/PC＋では自動的に有意水準を出力してくれるのでＦ分布表を参照する必要はない。有意水準はSignif of Fの下に示されている。結果は面接と内申の間に交互作用はなく，面接と内申にはそれぞれ１％水準(0.001< 0.01)，５％(0.035< 0.05)水準で主効果が認められている。

多重分類分析（ＭＣＡ）

ＭＣＡはAndrews, Morgan, Sonquist, and Klem(1973)によって提唱された分析手法である。ＭＣＡでは独立変数の各カテゴリーが従属変数に対して及ぼす効果の強さを知ることができる。特徴は名義尺度を含めた質的なデータや非線形データを扱えること，人種や性など実験的にコントロールが困難な変数のために不釣り合い型になっている実験計画を扱えることが挙げられる。

図３は分散分析表に続いて出力された多重分類分析の結果である。

```
---------------------------------------------------------------------
* M U L T I P L E   C L A S S I F I C A T I O N   A N A L Y S I S  * *
                TEST
          By    MENSETU
                NAISIN

Grand Mean =       56.667                                  Adjusted for
                Adjusted for   Independents
                               Unadjusted    Independents    + Covariates
Variable + Category    N       Dev'n  Eta    Dev'n   Beta   Dev'n   Beta

MENSETU
    1                  4       13.83         13.23
    2                  5      -11.07        -10.59
                                    .87             .83

NAISIN
    1                  4        7.33          6.01
    2                  5       -5.87         -4.81
                                    .46             .38

Multiple R Squared                              .897
Multiple R                                      .947
---------------------------------------------------------------------
```

図３　多重分類分析表

Nの後に人数,Unadjusted Dev'nの下には未調整カテゴリー偏差が示されている。この未調整カテゴリー偏差は各カテゴリー平均から全体平均を引いたものである。例えば,面接カテゴリー1の未調整偏差は面接のカテゴリー1にあてはまる人間の成績平均(72+84+65+61)/4=70.5から全体成績平均56.67を引いて求めることができる。同様に面接カテゴリー2の未調整偏差は面接のカテゴリー2に該当した人間の成績平均(55+46+48+39+41)/5=45.6から全体成績平均を引いた数字になる。内申についても各カテゴリー平均(カテゴリー1なら7.33, カテゴリー2なら-5.87)から全体平均を引けばよい。

次のAdjusted for Independents Dev'nの下に示されている数値は調整化カテゴリー偏差である。MCAでは各個人の得点は従属変数の平均値と各独立変数のカテゴリー効果の加算で表わすことができると仮定する。

$$\hat{Y}_{ijk}\ldots=\bar{y}+a_i+b_j+c_k+\cdots+e_{ijk}\ldots \qquad (1)$$

Yは各個人の得点, $\bar{y}$は従属変数の平均値, a_i, b_j, c_k...は各独立変数におけるカテゴリー効果, e_{ijk}...は誤差項である。

このモデルには交互作用効果が含まれていない。言い換えれば, MCAが分析対象とするのは有意な交互作用が見られない場合ということになる。

このとき, a_i, b_j, c_k...の各カテゴリー効果として未調整カテゴリー偏差を利用することはできない。なぜなら, 各カテゴリー効果には重複があるからである。面接で1の評価を得た個人が同時に内申では1の評価を得ているかもしれない。この場合, その個人が面接で1である効果(未調整カテゴリー偏差)には内申が1である効果も含まれていることになる。単純に考えれば, 内申評定が良い学生は面接でも良い印象を与える可能性が高いかもしれないし, 内申評定が低い学生は面接でも低い評定しか受けないかもしれない。そうなると, 面接が良いと評価されたことの中には内申で良いと評価された性質のいくらかが含まれていると考えなくてはならない。この問題を解決するにはダミー変数を利用して正規方程式を解く方法がある。具体的にここで取り上げた例に沿って考えてみよう。面接と内申の二つの独立変数があり, それぞれカテゴリー数が二つあるのでモデルは次のようになる。

$$\hat{Y}_{ij}=\bar{y}+a_i+b_j+e_{ij} \qquad (2)$$
$$i=1,2 \quad j=1,2$$

誤差項e_{ij}が最小化されるようにa_iとb_jを決めるということは, 観測データY_{ij}と(2)式からeの項を除いた$\hat{Y}_{ij}$との差を最小化するということに他ならない。

$$Q=(Y_{ij}-\hat{Y}_{ij})^2 \rightarrow \text{最小化} \qquad (3)$$

(3)式にe_{ij}を除いた(2)式を代入して考えると

$$Q=(Y_{ij}-\bar{y}-a_i-b_j)^2 \qquad (4)$$

この問題を解くにはまず各要素にデータを代入して, 式$f(a_1,a_2,b_1,b_2)$を構成する。次に各変数について偏微分し, 0とおくことで得られた正規方程式を解けばよい。後で触れる多重共線性の問題が生じるので, 方程式を解くには工夫が必要になるはずである。しかし, Andrewら(1973)はその詳細にはふれておらず, 単にこのように逆行列を使って解く方法よりも, より計算時間が短いとして反復手法(iterative procedure)を採用したと述べている。反復手法では未調整カテゴリー偏差を初期値として, 順次各カテゴリー効果の大きさを反復推定していく。ここで挙げたように独立変数が二つの場合なら計算式は

step 1

$a_i^{(1)}$=カテゴリーa_iの平均-全体平均(未調整カテゴリー偏差)

$b_j^{(1)}$=カテゴリーb_jの平均-全体平均(未調整カテゴリー偏差)

step 2

$a_i^{(2)}=a_i^{(1)}-(1/n_i)\Sigma n_{ij}b_j^{(1)}$

$b_j^{(2)}=b_j^{(1)}-(1/n_j)\Sigma n_{ij}a_i^{(2)}$

・

・

・

step k

$a_i^{(k)}=a_i^{(1)}-(1/n_i)\Sigma n_{ij}b_j^{(k-1)}$

$b_j^{(k)}=b_j^{(1)}-(1/n_j)\Sigma n_{ij}a_i^{(k)}$

（nは対応するケース数）

収束条件は様々に設定できるが，一般には規定ステップに達した場合，あるいはステップhとステップh－１で得られた推定効果の差があらかじめ決めた微小定数以下になった場合に計算は収束する。

ここで挙げたデータについて実際に反復手法がどのように行われるかを示そう。面接とカテゴリー各々と内申のカテゴリー各々について調整化カテゴリー偏差が求められることになる。そこで見やすいように，ここでは前出のデータを面接についてカテゴリー毎にソートしたもの（表１）と内申についてカテゴリー毎にソートしたもの（表２）の二つの表示を示す。表１では面接のカテゴリー効果の結果を示し，表２では内申のカテゴリー効果の結果のみを示す。もちろん，これは便宜上の表示にすぎない。面接・

表１　面接でソートしたデータ

（面接）　面接カテゴリー調整偏差

面接	内申	成績	平均	step1	step2	step3	step4	step5
1	1	72	70.50	13.83	13.10	13.23	13.23	13.23
1	1	84	70.50	13.83	13.10	13.23	13.23	13.23
1	2	65	70.50	13.83	13.10	13.23	13.23	13.23
1	2	61	70.50	13.83	13.10	13.23	13.23	13.23
2	1	54	45.60	−11.07	−10.48	−10.58	−10.59	−10.59
2	1	46	45.60	−11.07	−10.48	−10.58	−10.59	−10.59
2	2	48	45.60	−11.07	−10.48	−10.58	−10.59	−10.59
2	2	39	45.60	−11.07	−10.48	−10.58	−10.59	−10.59
2	2	41	45.60	−11.07	−10.48	−10.58	−10.59	−10.59

成績平均　56.67

表２　内申でソートしたデータ

（内申）　内申カテゴリー調整偏差

面接	内申	成績	平均	step1'	step2'	step3'	step4'	step5'
1	1	72	64.00	7.33	6.02	6.01	6.01	6.01
1	1	84	64.00	7.33	6.02	6.01	6.01	6.01
2	1	65	64.00	7.33	6.02	6.01	6.01	6.01
2	1	61	64.00	7.33	6.02	6.01	6.01	6.01
1	2	54	50.80	−5.87	−4.82	−4.81	−4.81	−4.81
1	2	46	50.80	−5.87	−4.82	−4.81	−4.81	−4.81
2	2	48	50.80	−5.87	−4.82	−4.81	−4.81	−4.81
2	2	39	50.80	−5.87	−4.82	−4.81	−4.81	−4.81
2	2	41	50.80	−5.87	−4.82	−4.81	−4.81	−4.81

成績平均　56.67

内申のデータの並びが違うだけで元のデータ自体は同一である。

表1の平均は面接についての各カテゴリー毎の成績の平均を示している。ここでは面接評価1の学生の成績平均が70.50で，面接評価2の学生の成績平均が45.60であることがわかる。同様に表2には内申評価が1の学生の成績平均64.00と内申評価2の学生の成績平均50.80が示されている。ＭＣＡの反復手法は表1のstep1にはじまり，表2のstep1'，という具合にstep1→step1'→step2→step2'→step3→step3'→...の順で計算が進められる。はじめに面接の各カテゴリー平均から全体平均56.67を引いた値がstep1の値となる。同様に内申の各カテゴリー平均から全体平均56.67を引いた値がstep1'に示されている。これが反復計算の初期値であり，カテゴリー平均と全体平均の単純な偏差，すなわち未調整カテゴリー偏差である。図3のUnadjusted Dev'nに示されたSPSS/PC＋の分析出力に一致することがわかるだろう。次にstep2ではstep1の初期値から，面接カテゴリー1については（1／面接1のカテゴリー数）×（面接・内申が共に1のカテゴリー数×内申カテゴリー1の効果＋面接が1で内申が2のカテゴリー数×内申カテゴリー2の効果）を引くことになる。面接カテゴリー2については（1／面接2のカテゴリー数）×（面接が2で内申が1のカテゴリー数×内申カテゴリー2の効果＋面接・内申が2のカテゴリー数×内申カテゴリー2の効果）を引くことになる。ここでは計算の最初なのでカテゴリー効果は初期値を用いる。以後の計算では反復計算の結果得られた値をカテゴリー効果の推定値として利用する。こうして徐々に他の変数の効果を除去していくのである。

具体的な数値で示すと

step2　面接カテゴリー1=13.83-(1/4)(2×7.33+2×(-5.87))=13.10
面接カテゴリー2=-11.07-(1/5)(2×7.33+3×(-5.87))=-10.48

step2'　内申カテゴリー1=7.33-(1/4)(2×13.10+2×(-10.48))=6.02
内申カテゴリー2=-5.87-(1/5)(2×13.10+3(-10.48))=-4.82

step3　面接カテゴリー1=13.83-(1/4)(2×6.02+2×(-4.82))=13.23
面接カテゴリー2=-11.07-(1/5)(2×6.02+3×(-4.82))=-10.58

.
.
.

すでにstep4で小数点以下2桁までに関して出力と同じ値が得られていることがわかる。

多重分類分析（ＭＣＡ）におけるその他の統計量

図3のＭＣＡ表のその他の統計量についても簡単に説明しておく。eta統計量は以下の式で表わせる。

$$eta_i = \sqrt{(\sum(\text{未調整カテゴリー偏差の二乗}\times\text{ケース数}/\text{全体平方和})}$$

（iは変数を示す添え字）

例えば，面接カテゴリーの場合なら図2の分散分析表の全体平方和(1824.000)と面接の未調整カテゴリー偏差を用いて次のようになる。

$$eta_1 = \sqrt{((13.83)^2\times 4+(-11.07)^2\times 5)/1824} = 0.87$$

同様に内申カテゴリーのeta係数は以下のようになる。

$$eta_2 = \sqrt{((7.33)^2\times 4+(-5.87)^2\times 5)/1824} = 0.46$$

このeta係数は全体変動に対する各変数の説明力指標の一つと言える。しかし，未調整カテゴリー偏差を用いているために各変数間の効果の重複が除かれておらず，純粋にその変数の全体に占める効果の大きさを常に示しているとは限らない。

一方，beta係数はそのような変数間の効果の重複を除いた調整化カテゴリー偏差を用いたものである。従って，純粋に変数の説明力見ることができる。

beta統計量は次の式で表わされる。

$beta_i = \sqrt{(\sum(\text{調整化カテゴリー偏差の二乗} \times \text{ケース数} / \text{全体平方和})}$

（iは変数を示す添え字）

これは未調整カテゴリー偏差のかわりに調整化カテゴリー偏差を用いただけでetaの式と本質的に同じである。

具体的に計算を示すと，面接・内申の効果はそれぞれ次のようになり，図3の出力に一致する。

$$beta_1 = \sqrt{((13.23)^2 \times 4 + (-10.59)^2 \times 5)/1824} = 0.83$$

$$beta_2 = \sqrt{((6.01)^2 \times 4 + (-4.81)^2 \times 5)/1824} = 0.38$$

図3の出力の最後にあるMultiple R SquaredとMultiple Rについて説明しよう。Multiple Rは重相関係数として知られているものである。Multiple R SquaredはMultiple Rを二乗したものである。これは0.947^2=0.897から容易にわかる。相関係数の二乗は決定指数と呼ばれ，全分散に対する説明率を示す。従い，この場合，全分散を100としたとき，89.7%の分散が面接と内申の変数によって説明できることを意味する。このMultiple R Squaredはモデルの適合度を知る上で重要な指標である。ＭＣＡのモデルは式(2)にあるように主効果のみを扱っており，変数間の交互作用（データ例の場合なら，面接と内申の交互作用）は考慮していないことに注意しなくてはならない。逆に言えば，交互作用が見られるデータの場合にはＭＣＡの結果を解釈する意味は薄いということになる。ただし，もし仮に交互作用が見られた場合でも交互作用が見られる変数のカテゴリーを併合すればＭＣＡは意味を持ってくる。面接と内申のデータで交互作用が生じた場合なら，それぞれ二つのカテゴリーだけなので（面接１・内申１）（面接１・内申２）（面接２・内申１）（面接２・内申２）という四つのカテゴリーの組み合わせができる。これを四つのカテゴリーを持つ一つの変数と見なして分析する訳である。

ところで，Multiple Rが重相関係数だとして，ここでの重相関係数とは何を意味しているのであろうか。一般に重相関係数とは重み付けした複数の独立変数を線形結合させた式から得られる値と従属変数の間の相関係数を意味する。独立変数の重み付けはこの従属変数との相関係数を最大化するように決められる。一般に相関係数と呼ばれている二変数間のピアソン相関係数とこの重相関係数には若干の違いがある。二変数間の相関係数は-1から＋1までの値を取るのに対して，重相関係数では0から＋1までの値をとり，マイナスの値は取らない。また，線形式に投入する変数の数が多くなればなるほど重相関係数の値は1に近づいて，大きくなる。従って，重相関係数の値は相関係数ほど直観的に解釈できず，値の大小はあくまで投入変数の数を考慮した上で行われなくてはならない。さて，ここでの重相関係数であるが，式(1)により得られた予測値と実際の観測値，すなわち従属変数との相関係数にあたる。例で挙げたデータに沿って説明すれば，面接と内申の値から合成得点を作り出し，最も入学試験を説明するように重み付けした合成変数得点と試験成績との相関係数になるのである。ただし，一般の重回帰分析では従属変数，独立変数ともに量的データであるのに対し，ここでは独立変数に質的なデータを用いている。

数量化Ｉ類

量的データを従属変数とし，独立変数を質的なデータとする分析には日本で開発された林の数量化Ｉ類と呼ばれるものがある。これは統計数理研究所の林知己夫博士が開発した１群の解析手法の１つである。数量化Ｉ類の他にも有名なものとして数量化II類，数量化III類，数量化IV類などがある。いずれも質的なデータを扱うのが特徴で，それぞれ量的なデータにおける分析に対して，数量化Ｉ類は重回帰分析，数量化II類は判別分析，数量化III類は主成分分析に対応するものと考えられている。

ＭＣＡとの関連を調べる前に，ここで簡単に数量化Ｉ類の考え方を説明してみよう。数量化ではダミー変数を用いる。これまでに用いてきたデータを利用して説明してみよう。例のデータでは面接と内申という二つの変数を考え，そ

表3　ダミー変数への変換

表1'				ダミー変数				
面接	内申	成績		面接1 a_1	面接2 a_2	内申1 b_1	内申2 b_2	成績
1	1	72		1	0	1	0	72
1	1	84		1	0	1	0	84
1	2	65		1	0	0	1	65
1	2	61	→	1	0	0	1	61
2	1	54		0	1	1	0	54
2	1	46		0	1	1	0	46
2	2	48		0	1	0	1	48
2	2	39		0	1	0	1	39
2	2	41		0	1	0	1	41

れぞれにカテゴリーを二つ持つものとした。ダミー変数とはこのカテゴリーに対してあてはまるか，あてはまらないかで新たに構成された変数のことである。こうして，表1のデータでカテゴリーにあてはまった場合を1，あてはまらない場合を0としてダミー変数を構成すると表3のようになる。新たなダミー変数「面接1」は面接が1のカテゴリーに当てはまったときに1，あてはまらない場合に0となる。ダミー変数「面接2」は面接が2に当てはまっているなら1，そうでなければ0となる。内申についても同様である。こうして，各変数についてカテゴリーの数だけダミー変数が作られる。

このダミー変数をもとに最も良く成績を予測する最小二乗解を求める。ダミー変数面接1をa_1，面接2をa_2，また，内申1をb_1，内申2をb_2とすると最小二乗解を求めるための式は以下のようになる。

$$\begin{aligned} f(a_1,a_2,b_1,b_2)=&(72-a_1-b_1)^2+(84-a_1-b_1)^2\\ &+(65-a_1-b_2)^2+(61-a_1-b_2)^2\\ &+(54-a_2-b_1)^2+(46-a_2-b_1)^2\\ &+(48-a_2-b_2)^2+(39-a_2-b_2)^2\\ &+(41-a_2-b_2)^2 \end{aligned} \tag{5}$$

この式は従属変数と各独立変数の差の二乗からなっているので，式の値が最小化されるようにa_1,a_2,b_1,b_2を求めれば，それが最適予測値になることがわかる。解法としてはa_1,a_2,b_1,b_2それぞれについて偏微分し，0と置く。

$$\begin{aligned} \frac{\partial f}{\partial a_1}=&-2(72-a_1-b_1)-2(84-a_1-b_1)\\ &-2(65-a_1-b_2)-2(61-a_1-b_2)=0 \end{aligned} \tag{6}$$

$$\begin{aligned} \frac{\partial f}{\partial a_2}=&-2(54-a_2-b_1)-2(46-a_2-b_1)-2(48-a_2-b_2)\\ &-2(39-a_2-b_2)-2(41-a_2-b_2)=0 \end{aligned} \tag{7}$$

$$\begin{aligned} \frac{\partial f}{\partial b_1}=&-2(72-a_1-b_1)-2(84-a_1-b_1)\\ &-2(54-a_2-b_1)-2(46-a_2-b_1)=0 \end{aligned} \tag{8}$$

$$\begin{aligned} \frac{\partial f}{\partial b_2}=&-2(65-a_1-b_2)-2(61-a_1-b_2)-2(48-a_2-b_2)\\ &-2(39-a_2-b_2)-2(41-a_2-b_2)=0 \end{aligned} \tag{9}$$

式を整理すると(6)の場合なら

$$\begin{aligned} &-2(72-a_1-b_1)-2(84-a_1-b_1)\\ &-2(65-a_1-b_2)-2(61-a_1-b_2)\\ =&-144+2a_1+2b_1-168+2a_1+2b_1\\ &-130+2a_1+2b_2-122+2a_1+2b_2 \end{aligned}$$

$=-564+8a_1+4b_1+4b_2$

$=-282+4a_1+2b_1+2b_2=0$

定数を右辺に移項すると

$4a_1+2b_1+2b_2=282$

他も同様に計算して，最終的には以下の4つの式が得られる。

$$4a_1+2b_1+2b_2=282 \quad (6)'$$
$$5a_2+2b_1+3b_2=228 \quad (7)'$$
$$2a_1+2a_2+4b_1=256 \quad (8)'$$
$$2a_1+3a_2+5b_2=254 \quad (9)'$$

このようにして得られた(6)'から(9)'は正規方程式と呼ばれる。ここでの例では変数，カテゴリー，データ数いずれも少ないのでこのように偏微分して正規方程式を求める事ができるが，大きなデータではこのような計算は煩雑になる。そこで，一般には行列を用いた計算を行う。表3の右側のデータをそのまま行列Cと見立てて，その転置行列C'を掛ける。

$$\mathbf{C'C} = \begin{bmatrix} 30724 & 282 & 228 & 256 & 254 \\ 282 & 4 & 0 & 2 & 2 \\ 228 & 0 & 5 & 2 & 3 \\ 256 & 2 & 2 & 4 & 0 \\ 254 & 2 & 3 & 0 & 5 \end{bmatrix}$$

結果は第一行目を除いて，それぞれ(6)'から(9)'の正規方程式の係数，及び定数に一致することが確認できる。第一行目を除き，係数だけの行列をD，定数のベクトルをB，求めるべき変数ベクトルをXとすると，正規方程式は以下のようになる。

$$\begin{bmatrix} 4 & 0 & 2 & 2 \\ 0 & 5 & 2 & 3 \\ 2 & 2 & 4 & 0 \\ 2 & 3 & 0 & 5 \end{bmatrix} \times \begin{bmatrix} a_1 \\ a_2 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 282 \\ 228 \\ 256 \\ 254 \end{bmatrix} \quad (10)$$

$$\mathbf{DX} = \mathbf{B}$$

あとはXを求めるだけなので，Dの逆行列を左から掛ければよいことになる。ところが，Dの逆行列は求める事ができない。それはDがランク落ちしているからであり，実際Dの行列式の値は0になる。これは多重共線性の問題として知られているが，簡単に言えば，連立方程式において求める未知数に対して十分な数の式がない状況である。(10)式の場合，求める未知数はa_1からb_2の四つであり，方程式も四つあるので一見問題が無いように思える。しかし，一行目の式と二行目の式を足すと三行目の式と四行目の式を足した式に等しくなる。これは線形従属関係があるということで，式は四つあるが実質的には同じ情報が重複して含まれているため，未知数を求めるに十分な情報量がないことを意味する。この多重共線性の問題はこのように質的なデータをダミー変数を用いて分析する場合には独立変数が一つ以上の場合を除き，常に生じる。

数量化Ⅰ類ではこの問題を解決するために任意の未知数を0とおく。仮に$b_2=0$と置けば，以下の式を解いて$a_1=64.825$, $a_2=41.455$, $b_1=10.86$, $b_2=0$と値を求めることができる。

$$\begin{bmatrix} 4 & 0 & 2 \\ 0 & 5 & 2 \\ 2 & 2 & 4 \end{bmatrix} \times \begin{bmatrix} a_1 \\ a_2 \\ b_1 \end{bmatrix} = \begin{bmatrix} 282 \\ 228 \\ 256 \end{bmatrix}$$

このように任意に未知数を0と置くのはおかしなことのよう思えるかも知れない。そこで，他の未知数を0と置いた場合の結果を表4に示す。

表4からあきらかなように，どの未知数を0と置くかで各値は異なってくるが，変数のカテゴリー間の差は一定になっている。どの場合にも(a_1-a_2)の値は23.818であるし，(b_1-b_2)は10.818である。すなわち，各カテゴリー間の相対的な差が一定になることが保証されるのである。

数量化Ⅰ類の考え方は以上であり，こうして求められた値が数量化の値である。このままでもよいのであるが，各変数ごとに付与する数値

表4　各ダミー変数を0と置いた場合の数量化結果

	a_1=0の場合	a_2=0の場合	b_1=0の場合	b_2=0の場合
a_1	0.000	23.818	75.909	65.091
a_2	−23.818	0.000	52.091	41.273
b_1	75.909	52.091	0.000	10.818
b_2	65.091	41.273	−10.818	0.000

の平均が0になるように規準化することが一般に行われる。b_2=0の場合で考えてみる。a_1=65.091，a_2=41.273，b_1=10.818，b_2=0を表3に当てはめてみる。すると面接に関しては面接1に該当するのが四人に，面接2に該当するのが五人，内申も同様なので合計，平均は次のようになる。

面接　合計(4×65.091)+(5×41.273)=466.729
　　平均466.729/9=51.859
内申　合計(4×10.818)+(5× 0)　= 43.272
　　平均 43.272/9= 4.808

求めた各カテゴリーの数量値からこの平均を引く。

	b_2=0の場合
a_1	65.091-51.859= 13.232
a_2	41.273-51.859=-10.586
b_1	10.818- 4.808= 6.010
b_2	0.000- 4.808=- 4.808

こうして得られた値，すなわちa_1=13.232，a_2=-10.586，b_1=6.010，b_2=-4.808が一般に統計量として示される数量化Ⅰ類の結果である。ここで用いた面接と内申の平均点の合計(51.859+4.808=56.667)が成績の全体平均に一致することもわかる。

MCAと数量化Ⅰ類の関連に付いて

数量化Ⅰ類の分析で得られた値はよく注意してみると，MCAで得られた図3の調整化カテゴリー偏差と一致している。これは偶然なのだろうか。

数量化Ⅰ類のモデルに付いて検討してみると，ダミー変数xの線形結合式が最も予測変数Yを説明できるように重み付けtを考えている。例のデータの場合なら式(11)のようになる。そしてここでのtこそが数量化なのである。

$$\hat{Y}=t_{11}x_1+t_{12}x_1+t_{21}x_2+t_{22}x_2+e \qquad (11)$$

$(\hat{Y}-Y)^2$→最小化

式(11)の場合，ダミー変数$t_{11}x_1$と$t_{12}x_1$は実は面接という一つの変数であったものであるし，$t_{21}x_2$と$t_{22}x_2$も内申という一つの変数であった。そして，被験者はカテゴリーのいずれか一つに反応していたわけで，反応していないダミー変数は0であることを考えれば，ある一つの変数についてダミー変数が同時に複数存在することはない。そこでダミー変数をもとの変数に戻して，それぞれαとβと表してみると式(12)のように表すことができる。

$$Y_{ij}=\alpha_i+\beta_j+e_{ij} \qquad (12)$$
$$i=1,2 \quad j=1,2$$

式(12)はMCAの式(2)と極めて似ていることに気が付く。実際，この式でα，βについて求めた解は表4で得られた数値になる。さらに数量化Ⅰ類ではカテゴリー平均が0になるように調整を行った。具体的には各カテゴリー平均を引いたのである。同じ操作を式(2)に加え，αとβの平均値を引くと式(13)のようになる。

$$Y_{ij}-\bar{y}=\alpha_i+\beta_j-(\bar{\alpha}_i+\bar{\beta}_j)+e_{ij} \qquad (13)$$
$$i=1,2 \quad j=1,2$$

左辺に$\bar{y}$があるのは右辺で引いた$\bar{\alpha}_i$と$\bar{\beta}_j$の合計は先に触れたように成績の全体平均に等しくなるからである。式(13)で左辺の全体平均を右辺に移項して整理すると式(14)のようになる。

$$Y_{ij}=\bar{y}+(\bar{\alpha}_i-\alpha_i)+(\bar{\beta}_j-\beta_j)+e_{ij} \qquad (14)$$
$$i=1,2 \quad j=1,2$$

ここでの$(\alpha_i-\bar{\alpha})$, $(\beta_j-\bar{\beta})$が実はＭＣＡのモデルのa_i, b_jに他ならず，それぞれを置き換えると式(2)に一致することがわかる。つまり，ＭＣＡと数量化Ｉ類は解法にこそ違いがあるものの，数学的にはまったく同じものと考えられるのである。従い，両分析の解釈も同様なものとなる。たとえば，予測値は従属変数の平均点に各カテゴリーへの反応得点を加算して求められる。面接が１で内申も１の学生の予測成績はそれぞれのカテゴリーに関して求められた得点の合計（13.23＋6.01=19.24）に成績平均56.67を足した75.91であるし，面接が１で内申が２の学生の予測成績はそれぞれのカテゴリー得点合計(13.23-4.81=8.42）に成績平均56.67を足して65.09となる。他も同様である。全体としての予測の正確さ，言い替えれば加算モデルの適合度は重相関係数により知る事ができる。また，カテゴリーの数値のレンジが大きな変数はそれだけ従属変数の予測に寄与していることがわかるし，正負の符号によってその寄与の方向も知る事ができるのである。

まとめ

SPSS/PC＋における簡単な分散分析のやり方とその結果の解釈について解説した。そして，結果出力オプションとして得られる多重分類分析（ＭＣＡ）が数学的には林の数量化Ｉ類に他ならず，SPSS/PC＋において数量化Ｉ類の分析結果を得る事ができることを示した。

文献

Andrews, F., J. Morgan, J. Sonquist, and L. Klem 1973 *Multiple classification analysis.* 2nd ed. Ann Arbor: University of Michigan.

藤沢偉作　1985　楽しく学べる多変量解析法　現代数学社.

本田正久・島田一明　1977　経営のための多変量解析法　産業能率大学出版部.

岩坪秀一　1987　数量化法の基礎　朝倉書店.

河口至商　1973　多変量解析入門Ｉ　森北出版.

小林竜一　1981　数量化理論入門　日科技連.

駒澤勉　1992　数量化理論　放送大学教育振興会.

古谷野亘　1988　数学が苦手の人のための多変量解析ガイド　川島書店.

三宅一郎・中野嘉弘・水野欽司・山本嘉一郎　1977　SPSS統計パッケージ　解析編　東洋経済新報社.

三宅一郎・山本嘉一郎・垂水共之・白倉幸男・小野寺孝義　1991　新版SPSS^X III 解析編２，東洋経済新報社.

SPSS Inc. 1991 *SPSS Statistical Algorithms,* 2nd ed. Chicago, SPSS inc.

SPSS JAPAN Inc. SPSS/PC＋ Base Manual V3.0J (SPSS/PC＋ NEC PC-9800対応版マニュアル).

竹内啓編集代表　1989　統計学辞典　東洋経済新報社

田中豊・脇本和昌　1983　多変量統計解析　現代数学社.

田中豊・垂水共之　脇本和昌　1984　パソコン統計解析ハンドブックII　多変量解析編　共立出版.

垂水共之・西脇二一・石田千代子・小野寺孝義　1990　新版SPSS^X II 解析編１東洋経済新報社.

山本嘉一郎・吉村英・竹村和久　1991　パソコンSPSS　基礎編　東洋経済新報社.

渡部洋　編著　1988　心理・教育のための多変量解析法入門　基礎編　福村出版.

Woodward, J and J. Overall 1976 Nonorthogonal analysis of variance in repeated measures experimental designs, *Educational and psychological measurement,* 36, 855-859.

ABSTRACT

Data analysis using SPSS/PC+: II
- The relationship between Multiple classification analysis and Hayasi's first method of quantification -

In this article, we offer a simple example of analysis of variance (ANOVA) and multiple classification analysis (MCA) using SPSS/PC +. Furthermore, we examine the relationship between MCA and Hayasi's first method of quantification. It is indicated that, although both analyses use different algorithms for solution, that is, iterative technique for MCA and dummy variables for the Hayasi's first method of quantification, each analysis finally yields same results.

Mathematical implications of both analysis models are discussed.