統計数理　第36巻　第2号（1988）

公開講演会要旨

カテゴリカルデータのモデル分析

——プログラムCATDAP-01, 02（改訂版）の紹介——

統計数理研究所　坂元慶行・桂　康一

（昭和63年11月4日，統計数理研究所 講堂）

1. はじめに

われわれは，1980年に，カテゴリカルデータにおける変数選択のためのFORTRANプログラムとして，CATDAP-01, 02（Katsura and Sakamoto（1980），坂元（1985））を発表した．CATDAP-01は目的・説明両変数ともカテゴリカルなデータにおいて最適な説明変数を選択するためのプログラムである．しかし，このプログラムは説明変数の最適カテゴリの探索は行なわないから，全ての変数がカテゴリ化されているデータでなければ適用できない．これに対して，CATDAP-02は説明変数の選択だけでなく，カテゴリの最適区分も探索する．連続的な変量も適当に区分すればカテゴリカルな変量に帰着させることができるから，結局，CATDAP-02は目的変数がカテゴリカルでありさえすれば，どんなデータに対しても適用することができることになる．すなわち，このプログラムは，サンプル・サイズ，説明変数の数やその型（連続型，カテゴリカル）の如何にかかわらず適用できるのである．

これらのプログラムは極めて簡単な統計量に基づいてはいるが，種々の効果をもつ説明変数を検出することができ，分析結果の意味も明快である．幸いにしてこのプログラムが多くのユーザを見いだしたのもこのような特長をもっていたからだと思われる．しかし，プログラムCATDAP-02にはこうした長所がある一方で，開発時の大型計算機の環境上の制約から，カード入力・プリンタ出力という想定の下で作成されたため，使いづらいという面もあった．そこで，出力結果をディスプレイで表示するときの見やすさ，入力のしやすさ等を考慮して，入・出力の様式を中心に改訂[注]することにした．現実の統計解析では，所与のデータに対して同種の解析を何回か繰り返して，はじめてデータから有意な情報を抽出できることが多いが，今回の改訂はCATDAPによるこのような分析をより容易にすることを意図したものである（その具体例については2.2節参照）．なお，今回の改訂はパーソナル・コンピュータのためのBASICプログラムCATDAPJ（坂元　他（1988））の考えを受け継いだものである．しかし，通常の実データは極めて多数の連続・離散両変数から成る大量データという性格をもつが，このFORTRAN版はこのような実データに対しても対処可能であり，CATDAPJとは実用性の面で大きく異なる．

本稿の目的は改訂版CATDAP-01, 02の使用法について解説することである．そこで，まず，より適用範囲の広いCATDAP-02（Ver. 1988）の操作例について述べ，次いで，CATDAP-01と02の使用法について略説する．

2. 操 作 例

2.1 処理対象データ

分析例として用いるのは循環器系の集団検診から得られたデータ（駒澤（1978））である．表1にその一部を示したこのデータは，集団検診後1年以内に脳出血（変数番号3の疾患（DIAGNOSIS）のコードが1），脳梗塞（同コード2），心筋梗塞（同コード3），狭心症（同コード4）のいずれかを発症した49～59歳の52人について検査項目の測定値を示したものである．このデータは，上の4つの疾患（DIAGNOSISと表示）のほかに，眼底所見（OPTHALMO.）と心電図所見（ECG）を加えた3個のカテゴリカルな変数と，年齢（AGE），最大血圧（SYS. PRESS），最小血圧（DIA. PRESS），大動脈脈波速度（AORTIC WAVE），血清総コレステロール（CHOLESTEROL）の5個の連続型変数から成っている．

2.2 循環器集団検診データについての操作例

上の4つの疾患に対してどの検診項目が最も多くの情報をもっているかを知りたいものとする．そこで，処理対象データ（表1）を収録したデータセット（仮にJUNKANKI. DATAとする）を用い，3番目の項目DIAGNOSISを目的変数，他の変数を説明変数の候補とみなして，最適な説明変数を探すことにする．

一般的な操作の手順やコマンドの意味については3.1.Cおよび3.2.C節で解説するが，表2のようなデータ記述ファイル（JUNKADSC. DATAとする）を用意し，“CATDAP02”とキー・インしてCATDAP-02を起動し，終了を確認した後，“ED”に続いて，たとえば，“FD 2”とキー・インすることにより，ファイル@OUTPUT. DATA上に作成された分析結果のうちの出力〈2〉の単説明変数順位表（表3）をディスプレイで見ることができる．

この表3から，大動脈脈波速度が第1位，続いて血清総コレステロール，最大血圧，最小血圧等の順で効果が強いことが分かる．ところが，“FD 31”ならびに“FD 32”として，出力〈3〉の順位別2次元クロス表の1位と2位の部分（表4）を見ると，第1位の大動脈脈波速度の値の

表1． 循環器系集団検診データ

対象	眼底所見	心電図所見	疾患	年齢	最大血圧	最小血圧	大動脈脈波速度	血清総コレステロール
1	1	1	1	52	154	88	8.4	150
2	2	1	1	55	178	92	6.7	197
3	1	1	1	56	164	94	7.9	173
4	2	2	1	57	178	102	9.3	185
5	2	2	1	53	150	98	8.3	189
6	1	1	1	56	172	106	8.6	218
7	1	1	1	56	168	96	9.1	218
8	1	1	1	54	148	102	8.0	170
9	2	2	1	58	180	98	7.5	197
10	1	1	1	51	186	96	8.6	244
11	1	2	1	52	192	112	7.6	165
12	1	2	1	50	200	118	8.0	197
13	1	1	1	57	130	80	8.4	234
14	2	1	1	54	170	100	7.4	190

［以 下 省 略］

表2.

```
JUNKANKI.DATA
   52    8    0    0    1    0    0    0
(6F4.0,F5.1,F4.0)
  1  OPTHALMO.          1  2  2  0  0  0        0.
  2  ECG                1  2  2  0  0  0        0.
  3  DIAGNOSIS          1  4  2  1  0  0        0.
  4  AGE                0  0  0  0  0  0        1.
  5  SYS. PRESS         0  0  0  0  0  0        1.
  6  DIA. PRESS         0  0  0  0  0  0        1.
  7  AORTIC WAVE        0  0  0  0  0  0      0.1
  8  CHOLESTEROL        0  0  0  0  0  0        1.
```

表3.

```
I LIST OF EXPLANATORY VARIABLES (ARRANGED IN ASCENDING ORDER OF AIC) I
RESPONSE VARIABLE  : (DIAGNOSIS          )
```

NO.	EXPLANATORY VARIABLE	NUMBER OF CATEGORIES OF EXP. VAR.	A I C	DIFFERENCE OF AIC	WEIGHT
1	AORTIC WAVE	2	-7.92	0.0	1.00000
2	CHOLESTEROL	2	-6.84	1.08	0.58252
3	SYS. PRESS	2	-5.13	2.80	0.24713
4	DIA. PRESS	2	-3.87	4.05	0.13182
5	AGE	2	-0.78	7.14	0.02814
6	ECG	2	2.45	10.37	0.00559
7	OPTHALMO.	2	4.57	12.49	0.00194

違いは目的変数のカテゴリ2の脳梗塞の発症確率に与える影響が大きいが，第2位の血清総コレステロールの違いは目的変数のカテゴリ4の狭心症の発症確率に与える影響が大きいというように，説明変数の影響を与える目的変数のカテゴリが説明変数によって異なる．当然のことながら，目的変数のカテゴリが3以上であれば，この現象はつねに起こり得る．

さて，分析者の目的がまさしく「疾患に対してどの検診項目が情報をもつか」にあり，それ以上でも以下でもなければ，このままでよい．しかし，分析者の次の目的が，たとえば，「脳梗塞の発症確率に効果をもつ変数は何か」にあったとすれば，目的変数 DIAGNOSIS のカテゴリ2を1に，その他のカテゴリを2にリコードして CATDAP-02 を適用しなければならないし，目的が「梗塞系疾患の発症確率に影響を与える変数は何か」にあったとすれば，目的変数のカテゴリ2と3を1に，その他のカテゴリを2にリコードして CATDAP-02 を適用しなければならない．

いま，われわれの目的が後者であったとすると，“CATDAP02” に続けて “Y” とキー・インしてデータ記述ファイルをディスプレイ上に呼び出し，表5のように変更することによって必要なリコードを行なうことになる（変更箇所は波線部のとおり）．このファイルをセーブして，再度 CATDAP-02 を起動すれば所期の分析が実行される．結果は前と同じ手順で見ることができ，たとえば，“ED” に続けて “FD 2” として単説明変数順位表を見ると，年齢が1位であることが分かる．この結果は，年齢による影響を排除するために，駒澤（1978）によってデータが 49～59 歳の被験者だけに限定されていたにもかかわらず，梗塞系疾患か否かを識別する上では年齢が極めて重要な要因であることを示している．実際，“FD 31” として第1位の2次元クロス表を見ると（表は省略），CATDAP-02 は最適な年齢区分として 51.5 歳を選び，年齢が 51 歳以下だと梗塞系疾患を発症する確率が高いが，52 歳以上になると脳出血か狭心症を発症する

表4.

```
+-----------------------------------------------+
I TWO-WAY TABLES WITH AN OPTIMAL CATEGORIZATION I
+-----------------------------------------------+
<31>
-----------------------------------------------
          (DIAGNOSIS            )
             1     2     3     4
 (AORTIC WAVE          )
-----------------------------------------------
    1     30.3  12.1  27.3  30.3 100.0 (  33)
    2     21.1  57.9  15.8   5.3 100.0 (  19)
-----------------------------------------------
TOTAL     26.9  28.8  23.1  21.2 100.0 (  52)
-----------------------------------------------
<NOTE>
     AORTIC WAVE              CLASS INTERVAL
         1                0.63000D+01 -  0.84500D+01
         2                0.84500D+01 -  0.10200D+02
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
<32>
-----------------------------------------------
          (DIAGNOSIS            )
             1     2     3     4
 (CHOLESTEROL          )
-----------------------------------------------
    1     30.4  39.1  30.4   0.0 100.0 (  23)
    2     24.1  20.7  17.2  37.9 100.0 (  29)
-----------------------------------------------
TOTAL     26.9  28.8  23.1  21.2 100.0 (  52)
-----------------------------------------------
<NOTE>
     CHOLESTEROL              CLASS INTERVAL
         1                0.14600D+03 -  0.19250D+03
         2                0.19250D+03 -  0.27900D+03
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
```

表5.

```
-------------------------------------------------------------------------------
JUNKANKI.DATA
   52    8    1    0    1    0    0    0
(6F4.0,F5.1,F4.0)
  1  OPTHALMO.              1  2  2  0  0  0          0.
  2  ECG                    1  2  2  0  0  0          0.
  3  DIAGNOSIS              1  2  2  1  1  0          0.
  2  1  1  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  4  AGE                    0  0  0  0  0  0          1.
  5  SYS. PRESS             0  0  0  0  0  0          1.
  6  DIA. PRESS             0  0  0  0  0  0          1.
  7  AORTIC WAVE            0  0  0  0  0  0         0.1
  8  CHOLESTEROL            0  0  0  0  0  0          1.
-------------------------------------------------------------------------------
```

確率が高いことを示している.

ここで問題にしている病気はいわゆる成人病であるから, 年齢を有力な説明変数とするこの結果はそれなりに説得力はあるが, 各年齢グループの内部で目的変数が説明変数にどのような形で依存しているかを知ることはできない. そこで, たとえば, 52歳以上のサンプルを除外して, 51歳以下の年齢層における両変数間の関係を見たいものとすると, データ記述ファイルを表6のように変更して(表5からの変更箇所は波線部), 再度CATDAP-02を適用し, 出力〈5〉のサブセット総合順位表の一部として表7を得る. この結果から, この年齢層では, 血清総コレステロールが第1位, 最大血圧×血清総コレステロールが第2位であることが分かる. しか

表6.

```
JUNKANKI.DATA
   52    8    1    1    1    0    0    0
(6F4.0,F5.1,F4.0)
  1  OPTHALMO.              1  2  2  0  0  0          0.
  2  ECG                    1  2  2  0  0  0          0.
  3  DIAGNOSIS              1  2  2  1  1  0          0.
  2  1  1  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  4  AGE                    0  0  0  0  0  1          1.
  1
    51.5     100.0
  5  SYS. PRESS             0  0  0  0  0  0          1.
  6  DIA. PRESS             0  0  0  0  0  0          1.
  7  AORTIC WAVE            0  0  0  0  0  0        0.1
  8  CHOLESTEROL            0  0  0  0  0  0          1.
```

表7.

I SUMMARY OF SUBSETS OF EXPLANATORY VARIABLES I

RESPONSE VARIABLE : DIAGNOSIS

	EXPLANATORY VARIABLES	NUMBER OF CATEGORIES OF EXP. VAR.	A I C	DIFFERENCE OF AIC	WEIGHT
1	CHOLESTEROL	2	-3.40	0.0	1.00000
2	CHOLESTEROL SYS. PRESS	4	-3.07	0.32	0.85019
3	CHOLESTEROL SYS. PRESS DIA. PRESS	8	-0.67	2.73	0.25584
4	CHOLESTEROL OPTHALMO.	4	-0.53	2.87	0.23794
5	AORTIC WAVE	2	-0.49	2.91	0.23355
6	CHOLESTEROL AORTIC WAVE	4	-0.45	2.95	0.22924
7	SYS. PRESS	2	-0.36	3.04	0.21880
8	- - -	0	0.0	3.40	0.18300
9	CHOLESTEROL DIA. PRESS	4	0.75	4.15	0.12550
10	CHOLESTEROL AGE	4	0.75	4.15	0.12550

し，AICの値の差が0.32であることから分かるように，両者の効果にはほとんど差がなく，最大血圧×血清総コレステロールにも注意を払う必要があると思われる．そこで，表6のデータ記述ファイルの指定のうち，2行8列の0を1に変え，変数の数と変数番号を示す“3 3 5 8”という指定を最下行に1行追加した上で，CATDAP-02を再び稼働し，目的変数と最大血圧×血清総コレステロールとのクロス表を求めてみると表8が得られる．この表から，最大血圧が140 mmHg以下か，血清総コレステロールが192 mg/dl以下であれば必ず梗塞系の疾患を発症し，両者ともこれらの値を超えたとき，初めて非梗塞系疾患の発症確率が高くなるという極めて単純な構造が見いだせる．つまり，この表の説明変数のカテゴリのうちの初めの3行に示されたカテゴリ区分は冗長であり，その3カテゴリを1カテゴリに統合しても情報は減らない．このプーリングを行なえばそのクロス表のAICの値は4だけ小さくなって−7.07となり，先の第1

表8.

```
X(1):DIAGNOSIS
X(2):SYS. PRESS
X(3):CHOLESTEROL

 XX              RESPONSE VARIABLE
 --              -----------------
 23                  1     2
 --              -----------------
 11                  3     0     3
 12                  3     0     3
 21                  6     0     6
 22                  1     4     5
                 -----------------
 TOTAL              13     4    17
                 -----------------

                 ---------------------
                      1      2
                 ---------------------
 11                 100.0    0.0 100.0
 12                 100.0    0.0 100.0
 21                 100.0    0.0 100.0
 22                  20.0   80.0 100.0
                 ---------------------
 TOTAL               76.5   23.5 100.0
                 ---------------------


A I C =      -3.07

 < NOTE >
 X(1):DIAGNOSIS                  1 :      1
                                 2 :      2

 X(2):SYS. PRESS                 1 :  0.10800D+03 -  0.14050D+03
                                 2 :  0.14050D+03 -  0.20000D+03

 X(3):CHOLESTEROL                1 :  0.14600D+03 -  0.19250D+03
                                 2 :  0.19250D+03 -  0.27900D+03
```

位の血清総コレステロールを凌いで，1位に躍り出ることも分かる．

以上に示したように，リコード，一部のデータ，一部の変数の無視等の機能を用いながら，CATDAP-02を介した会話型の処理を行なうことによって効率的にデータの構造を探索することができる．

3. 入・出力について

3.1 CATDAP-01

3.1.A. 入力（データ記述ファイルの書式）

〈1行目〉 処理対象データ・セットのタイトル．

〈2行目〉

1列目　サンプル・サイズ

2列目　変数の総数

3列目　目的変数の総数

4列目　リコードする変数の総数

5列目　出力したい2次元クロス表の数　（−1は全表，0は10とみなされる）

6列目　一部の変数だけを説明変数の候補とするときは1，他は0
7列目　6列目が1のときの変数の数　(6列目が0なら0)
8列目　2次元クロス表の比較だけが必要なら1，多次元分析も必要なら0
9列目　たとえば,「性」という変数の「男」という属性をもつデータは除外して分析する場合のように，特定の属性をもつデータを除外して分析する場合の除外する変数の総数
10列目　特に指定した変数組によるクロス表が必要な場合の変数の組数

〈3行目〉 処理対象データのリード・フォーマット(必ずIフォーマット・コードで指定．1行は72桁.)

〈4行目以降〉 この指定のみ固定フォーマット（カッコ内の桁数）で指定する.

1列目　(1～3桁)　変数の一連番号
2列目　(6～25桁)　変数のタイトル
3列目　(26～28桁)　変数の（リコード後の）コードの最小値
4列目　(29～31桁)　変数の（リコード後の）コードの最大値
5列目　(32～34桁)　マイナス値ならその変数は無視され，0なら無視されない
6列目　(35～37桁)　その変数を目的変数とみなせば1，他は0
7列目　(38～40桁)　リコードするなら1,不要なら0.(1なら,原コードの1から19まで（必ず19まで指定）に対応する新コードをこの順に次行に指定)
8列目　(41～43桁)　その変数が一部のデータの無視に関与すれば1,他は0.(この指定が1なら次行に無視するコードの数を（同じ行の）その後に無視するコードを指定.なお,7列目も8列目も1で2行追加する場合は7列目の対応する行を先に置く．これは他の場合も同様)
9列目　(44～46桁)　2行6列目が1で,この指定が1なら,この変数は説明変数の候補とみなされる（2行6列目が0なら指定は不要)

〈最終行〉 特に指定した変数によるクロス表の分析の場合(入力書式の2行10列が0でない場合)の，変数の数と変数番号.(1組の変数につき1行を用い，変数番号の先頭には目的変数の変数番号を指定.)

3.1.B. 出力

以下の記述で出力番号中のkは目的変数の変数番号を表わす．したがって，出力番号にkが含まれているものは目的変数ごとに出力され，出力番号の先頭部に%が含まれるものは全体で一つ出力されることを示す.

〈%.1〉 データ記述ファイル　データ記述ファイルの内容,各パラメータの意味,目的変数の一覧.

〈k.2〉 単説明変数順位表　変数番号kの変数を目的変数としたときの結果で，AICに関する昇順に，順位，説明変数のタイトル(20字以下)，説明変数のカテゴリ数，AIC，各モデルとMAICEとのAICの差，MAICEモデルを1としたときの各モデルの（尤度に関する）ウェイト等を表示する．ここで，モデルjのウェイト$w(j)$は，MAICEモデルのAICをAIC(1)，モデルjのAICをAIC(j)とするとき,

$$w(j)=\exp\{-1/2\cdot(\mathrm{AIC}(j)-\mathrm{AIC}(1))\}$$

である．なお，ウェイト $w(j)$ のグラフも，その値が 0.01 以上のモデルに限って，付記される．

〈k.3〉 順位別2次元クロス表 上の〈k.2〉に対応するクロス表で，実数，パーセンテージで表示．なお，〈$k.3m$〉は変数番号 k を目的変数としたときの第 m 位の2次元クロス表を表わす．

〈k.4〉 パーセンテージのグラフ 目的変数のカテゴリごとに，上の〈k.3〉のパーセンテージを，〈k.2〉の説明変数の順に，図示．なお，〈$k.4m$〉は変数番号 k を目的変数としたときの第 m 位の説明変数に関するグラフを表わす．

〈%.5〉 2次元クロス表の AIC の総括表 全ての目的変数と説明変数とで作られ得る2次元クロス表に対する AIC の値を総括．目的変数ごとに，AIC の値の平均と分散を付記．

〈%.6〉 2次元クロス表に対する AIC の総括表の図化 上の〈%.5〉の5段階図示．

〈k.7〉 次数別サブセット順位表 このプログラムでチェックされた説明変数のサブセットをその次数ごとに整理し，サブセットを構成する説明変数のタイトル，サブセットの総カテゴリ数，AIC の値，その MAICE との差，各モデルのウェイト等を出力．なお，各次数とも上位 100 位のサブセットまで表示可能．

〈k.8〉 サブセット総合順位表 説明変数のサブセットの総合順位．

〈k.9〉 最適なサブセットによるクロス表 上の出力〈k.8〉の第1位に対応する多次元クロス表で，説明変数の最適なサブセットをもつクロス表を実数とパーセンテージで表示．

〈k.10〉 指定変数のみによるクロス表 特に指定された変数だけで作られるクロス表，その AIC，パーセンテージ等を表示．

なお，CATDAP-01 の出力ファイル @OUTPUT. DATA には，実際には，上に示した出力の順ではなく，各目的変数に対する2次元クロス表分析の結果が全目的変数分収録された（つまり，目的変数の総数を M とすると，〈k.2〉～〈k.4〉 $k=1, \cdots, M$）後，それに続いて，各目的変数に対する多次元分析の結果が2次元クロス表分析の場合と同じ順で収録されている．また，次節 3.1.C で解説するように，この @OUTPUT. DATA の内容は簡単なキー操作で見られるようにしておくと分析上便利である．

3.1.C. 操作法

あらかじめ，処理対象データの他に，データ記述ファイルを，3.1.A 節の書式に従って，用意しておく．“CATDAP01”（プログラム名は CATDAP-01 だがプログラムの稼働にはハイフンをつけない．CATDAP-02 の場合も同様）とキー・インして CATDAP-01 を起動すると，

```
DATA DESCRIPTIVE FILE=(直前に用いたデータ記述ファイルの名称)
MAIN DATA=(直前に用いた処理対象データのファイル名)
  "GO" OR "EDIT"? (TO CHANGE MAIN DATA=N ; TO EDIT DATA DESCRIP-
TIVE FILE=Y)
```

と表示される．

① そのファイルでよければリターンのみでCATDAP-01の実行に移る.

ファイルの変更には2通りのケースがある. まず,

② ファイル名を変更する場合は, "N"とキー・インすると, データ記述ファイル名と処理対象データ名を聞いてくるから, 適当に応答すると, CATDAP-01の実行に移る.

③ ディスプレイ上のデータ記述ファイル内の指定を訂正する場合は, "Y"とキー・インして, 適当に編集, セーブして, いったんジョブを終了させた後, 再び"CATDAP01"として起動する.

CATDAP-02の場合について, 以上の操作を支援するコマンド・プロシジャーの例を付録に示した(CATDAP-01の場合もほとんど同様である).

ところで, 出力結果はファイル @OUTPUT. DATA 上に作成されるから, 少ないキー操作で出力結果をディスプレイで見られるようにしておくのが便宜である. たとえば, 筆者の場合, 付録に示すように, EDとFDという二つのコマンドを用意し, "ED"とキー・インすることによって @OUTPUT. DATA の編集モードに入り, 引き続き"FD〈3.1.B節の出力番号〉"とキー・インすることによってその番号の出力結果が見られるようにしている. なお, CATDAP-01には, 目的変数とみなされている変数の変数番号を知るために目的変数の一覧表が用意されており, "FD RV"とキー・インすることによってそれを見ることができる.

3.2 CATDAP-02

3.2.A. 入力(データ記述ファイルの書式)

〈1行目〉 処理対象データ・セットのタイトル.

〈2行目〉

1列目　サンプル・サイズ
2列目　変数の総数
3列目　リコードする変数の総数
4列目　(特定の変数の)特定の属性をもつデータを除外して分析する場合の除外する変数の数
5列目　処理対象データが整数タイプなら0, 他は1
6列目　2次元クロス表の比較だけが必要なら1, 多次元分析も必要なら0
7列目　出力〈2〉の上位の変数だけを最適多次元クロス表探索の候補とする場合, 候補の数に1を加えた数
8列目　特に指定した変数組によるクロス表が必要な場合の変数の組数

〈3行目〉 処理対象データのリード・フォーマット
(上記2行5列目が1なら必ずFフォーマット・コードで指定. (Iフォーマット・コードとFフォーマット・コードの混在は不可.)また, 指定は2行(1行は72桁).)

〈4行目以降〉 この指定のみ固定フォーマット(カッコ内の桁数)で指定する.

1列目　(1～3桁)　変数の一連番号
2列目　(6～25桁)　変数のタイトル
3列目　(26～28桁)　変数の(リコード後の)コードの最小値. 実数なら0と指定
4列目　(29～31桁)　変数の(リコード後の)コードの最大値. 実数なら0と指定
5列目　(32～34桁)　カテゴリのプーリングの形式. 等間隔のプーリングが必要なら0, 不

等間隔のプーリングが必要なら1，プーリング不要なら2．なお，マイナス値ならその変数そのものが無視される

6列目 (35～37桁) その変数を目的変数とみなせば1，他は0

7列目 (38～40桁) リコードするなら1，不要なら0．(1なら，原コードの1から19まで（必ず19まで指定）に対応する新コードをこの順に次行に指定)

8列目 (41～43桁) その変数が一部のデータの無視に関与すれば1，他は0．(この指定が1で，次の9列目の指定が0.0でない場合に限り，次行に無視する区間数を，そのまた次の行に各区間の左端と右端を対で指定する．この指定が1でも，上の場合以外は，次行に無視するコード数を（同じ行の）その後に無視するコードを指定)

9列目 (44～53桁) その変数の観測精度（上記2行5列目が0ならこの指定は不要)

〈最終行〉 特に指定した変数によるクロス表の分析の場合（入力書式の2行8列が0でない場合）の，変数の数と変数番号．

3.2.B. 出力

〈1〉 データ記述ファイル CATDAP-01の〈%.1〉と同じ（表2，表5，表6参照)．

〈2〉 単説明変数順位表 CATDAP-01と同じ．なお，各説明変数のAICは最適カテゴリ区分時のAICである（表3参照)．

〈3〉 順位別2次元クロス表 説明変数の最適カテゴリ区分も表示されること以外はCATDAP-01に同じ（表4参照)．

〈4〉 最適カテゴリ区分下でのパーセンテージのグラフ 基本的にCATDAP-01に同じ．

〈5〉 次数別サブセット順位表とサブセット総合順位表 〈5〉は総合順位表，〈5k〉は$k+1$次元クロス表の順位表（表7参照)．

〈6〉 最適なサブセットによるクロス表 上の出力〈5〉の第1位に対応する多次元クロス表(説明変数の最適なサブセットで，かつ，最適なカテゴリ区分をもつクロス表)を実数とパーセンテージで表示．各説明変数のカテゴリの境界値も注記．

〈7〉 指定変数のみによるクロス表 特に指定された変数だけで作られるクロス表，そのAIC，最適カテゴリ区分を表示（表8参照)．

3.2.C. 操作法

原則としてCATDAP-01と同じである．操作の具体例については2.2節を参照されたい．

4. おわりに

様々な分野における優れた科学的発見と同様，優れた統計分析においても，直観に基づく仮説の想定からデータによる検証へ，という一方通行的な認識過程だけでなく，データ構造の抽出から新しい仮説の提示へ，という逆方向の認識過程が不可欠である．これら双方向の認識を繰り返し，徐々に仮説を純化することによって，はじめてわれわれは対象の特質に迫ることができる．統計的分析の場合，この後者の局面においては，大量の統計的情報の中から，観測ノイズを除去して，有意な情報を自動的に抽出する手法が不可欠である．すなわち，仮説検証的

なデータ解析法ではなく，本来の意味での構造探索的なデータ解析法である．たとえば，筆者が研究対象としている日本人の意識や世論の動向の場合にも，第一次オイル・ショック以降，その把握が困難になっており，実用性に富んだ構造探索的な解析法がますます必要になっている．

このような構造探索的なデータ解析を実現するために，現実のカテゴリカルなデータにおいて，特定の目的変数に対して有効な情報をもつ説明変数を探索するには，次のような条件を満たす手法を開発しなければならない．

① サンプリング誤差の処理が考慮されていること
② 説明変数の次数やカテゴリ数の影響を評価できること
③ 説明変数がどんなに多くても対処できること
④ 高次の交互作用の評価ができること
⑤ 説明変数の候補に連続変量が混在していても対処できること

プログラム CATDAP-01, 02 はそのための一つの方法であるが，今回の改訂によって会話型の処理が可能になったので，実データに対する構造探索的な解析がより容易になったと思われる．

謝　辞

本稿の執筆に当って有益なコメントをいただいた統計数理研究所の中村 隆助教授に心から感謝します．

注

今回の改訂では，入・出力の様式以外に，次の訂正も行なった．目的変数や説明変数に関する周辺度数が 0 のセルは，AIC の計算に当って，自由パラメータ数に算入しないように修正した．また，AIC の計算に当って，度数 0 のセルは e^{-1} とみなして計算するように修正した．

参 考 文 献

Katsura, K. and Sakamoto, Y. (1980). CATDAP, A categorical data analysis program package, *Comput. Sci. Monographs*, No. 14.

駒澤　勉（1978）．多次元データ分析の基礎，朝倉書店．

坂元慶行（1985）．カテゴリカルデータのモデル分析，共立出版．

坂元慶行，中村　隆，桂　康一，原　　聡（1988）．CATDAPJ：『パソコンによるデータ解析（村上，田村編）』別売ソフト，朝倉書店．

付録　コマンド・プロシジャーのリスト

CATDAP02

```
00100 PROC 0
00200 SET DSN1=JUNKADSC.DATA
00300 SET DSN2=JUNKANKI.DATA
00400 WRITENR DATA DESCRIPTIVE FILE =
00500 WRITE &DSN1
00600 WRITENR MAIN DATA =
00700 WRITE &DSN2
00800 WRITE "GO" OR "EDIT"? (TO CHANGE MAIN DATA = N:+
```

```
00900        TO EDIT DATA DESCRIPTIVE FILE = Y)
01000 READ &Y
01100 IF &Y¬=N && &Y¬=Y THEN GOTO A
01200 IF &Y=Y THEN GOTO B
01300 WRITENR DATA DESCRIPTIVE FILE?
01400 READ &DSN1
01500 WRITENR MAIN DATA?
01600 READ &DSN2
01700 E MCATPROC.CLIST(CATDAP02)
01800 200 SET &STR(DSN1)=&DSN1
01900 300 SET &STR(DSN2)=&DSN2
02000 END S
02100 GOTO A
02200 B:E &DSN1
02300 EXIT
02400 A:ALLOC DD(FT10F001) DS(&DSN2) REU
02500 ALLOC DD(FT05F001) DS(&DSN1) REU
02600 ALLOC DD(FT01F001) TE(WORK) SP(1 1) CY REU
02700 ALLOC DD(FT04F001) DS(@OUTPUT.DATA) REU
02800 ALLOC DD(FT06F001) DUMMY REU
02900 CALL CATDAPTS(CATDAP2)
03000 ALLOC DD(FT05F001) DS(*) REU
03100 ALLOC DD(FT06F001) DS(*) REU
```

なお，CATDAP-01に対しても同様に用意できるが，ここでは省略する．

CATDAP-01，02共通

ED

```
00100 E @OUTPUT.DATA
00200 V
```

FD

```
00100 PROC 1,NUM(2)
00200 V OFF
00300 TOP
00400 FIND /<&NUM>/
00500 LD
00600 V ON
```

ここで，LDならびにSCROLLは中村　隆によって作成されたもので，リストは以下のとおりである．

LD

```
00100 PROC 2 NUM() NL()
00200 /*****************************/
00300 /*       MACRO.CLIST:LD      */
00400 /*****************************/
00500 SET &NULL=
00600 IF &NL=&NULL THEN SET &NL=&SCROLL
00700 ELSE SCROLL &NL
00800 /*-------------------------*/
00900 IF &NUM=&NULL THEN -
01000  DO
01100   V ON
01200   L * &EVAL(&NL+1)
01300  END
01400 ELSE -
01500  DO
01600   L &NUM
01700   V OFF
01800   D
01900   V ON
02000   L * &NL
02100  END
```

SCROLL

```
00010 PROC 1 SCRL(20)
00020 /*****************************/
00030 /*     MACRO.CLIST:SCROLL    */
00040 /*****************************/
00050 X SETCS &&SCROLL VALUE(&SCRL)
```