カテゴリカルデータにおける変数選択

——プログラム CATDAP を中心に——

統計数理研究所　坂　元　慶　行

(1980年1月　受付)

Variable Selection in Categorical Data

Yosiyuki Sakamoto
(The Institute of Statistical Mathematics)

The purpose of the present paper is to show the construction and use of the Fortran program, CATDAP [5], developed for the variable selection in survey data. The program is applicable to the cases where a response variable is categorical. The practical utility of the program is demonstrated by applying to the analyses of the 1970 survey data of social stratification and status mobility and a set of data of the Japanese National Character and so on.

はじめに

社会学・心理学・医学・疫学等の研究分野では連続変量のみならずカテゴリカルな変量をも同時に含むいわば混合型のデータで，かつ，極めて多くの変数から成るデータが一般的である．このようなデータにおいて，特定のカテゴリカルな目的変数に対して多くの情報を持つ変数を自動的に検出するための方法を文献 [1, 2, 3, 4] で提案した．本稿の目的はこの方法をプログラム化した CATDAP (A Categorical Data Analysis Program Package) [5] の構成と利用法について述べるとともに，方法開発の背景，目的，モデルの構成法等について実例を中心に解説することである．

汎用フォートラン・プログラム CATDAP は目的変数がカテゴリカルでありさえすれば如何なるデータに対しても——説明変数の数，種類等の如何を問わず——適用できる．また，基礎をなす統計量は極めて簡単で，応用範囲が広く，実用的である．

1.　目的とモデルの構成

表1と表2は「社会階層と社会移動に関する調査」(以下，SSM 調査と略称) の 1975 年全国調査の結果である [6]．これら2表に共通してとりあげられている「階層帰属意識」に関する質問はつぎのとおりである．

"かりに現在の日本の社会全体を，この表にかいてあるように5つの層に分けるとすれば，あなた自身は，このどれに入ると思いますか．

1　上　　2　中の上　　3　中の下　　4　下の上　　5　下の下"

表1はこの「階層帰属意識」を「くらしむき」についての自己評価とクロスさせた表であり，表2は同じ項目を「世帯収入」とクロスさせた表である．ただし，3問とも回答に対するカテゴリーは表のようにまとめ直してある．

直井は，文献 [6] において，この「階層帰属意識」と「世帯収入」のクロス表だけでなく，同じ項目と「学歴」，「財産」，「従業上の地位」の各々とのクロス表もつくり，「階層帰属意識」の規定要因を見いだすための検討を行なっている．その結果，従来「階層帰属意識」の規定要

表 1

I_2 \ I_1		階層帰属意識 1. 上，中の上	2. 中の下	3. 下の上，下の下	計
くらしむき	1. 豊か	202 (0.61)	104 (0.31)	25 (0.08)	331 (1.00)
	2. ふつう	418 (0.21)	1183 (0.59)	407 (0.20)	2008 (1.00)
	3. 貧しい	24 (0.08)	113 (0.36)	174 (0.56)	311 (1.00)
計		644	1400	606	2650

表 2

I_3 \ I_1		階層帰属意識 1. 上，中の上	2. 中の下	3. 下の上，下の下	計
世帯収入	1. ～225 万円	213 (0.18)	605 (0.52)	350 (0.30)	1168 (1.00)
	2. 225～375	230 (0.24)	516 (0.55)	196 (0.21)	942 (1.00)
	3. 375～475	80 (0.29)	154 (0.56)	40 (0.15)	274 (1.00)
	4. 475～	121 (0.46)	125 (0.47)	20 (0.07)	266 (1.00)
計		644	1400	606	2650

* 2724 サンプルのうち学生を除外した分について集計した。( ) 内はパーセント.

因として重視されてきた地位変数は階層帰属意識を分化させる決定的要因ではないことを見い出し，さらに分析を重ねた後，「階層帰属意識を分化させる効果が最も大きいのは『くらしむき』変数である」と結論づけている．われわれもまた表1と表2の比較によって，「世帯収入」より「くらしむき」の方が「階層帰属意識」を分化させる効果が大きい，すなわち，「階層帰属意識」に関してより多くの情報を含んでいると判断することができよう．なお，直井が依拠したデータは表1，表2そのものではなく同時に実施されたもう1つの調査の結果も合せたものであるが結論に異なるところはない．

本稿では，この例のようにいくつかの分割表が与えられたときどの項目（変数）がより多くの情報を持っているかを比較するための統計的な方法について述べる．

「階層帰属意識」,「くらしむき」,「世帯収入」を，表1，表2にしめされているように，それぞれ I_1, I_2, I_3 で表わし，各 $I_j(j=1,2,3)$ は値 $1,\cdots,C_j$ をとるものとする．上の例の場合，$C_1=3$, $C_2=3$, $C_3=4$ である．3次元の同時確率を $p(i_1,i_2,i_3)$，対応する観測頻度を $n(i_1,i_2,i_3)$，サンプル・サイズを n $(n=\sum_{i_1,i_2,i_3} n(i_1,i_2,i_3))$ であらわす．必要に応じて

$$p(i_1,i_2)=\sum_{i_3} p(i_1,i_2,i_3)$$
$$n(i_1,i_3)=\sum_{i_2} n(i_1,i_2,i_3)$$

等の記法も用いることにする．

さて，いくつかの表の比較の問題を論じるために最も簡単な例を考えてみよう．いま，項目 I_2(たとえば「くらしむき」) のカテゴリー i_2 の下での項目 I_1 (たとえば「階層帰属意識」) のカテゴリー i_1 の条件付確率を $p(i_1|i_2)$ で表わす．このとき項目 I_1 に関して項目 I_2 が持つ情報の

多さは，i_2 の変化によって $p(i_1|i_2)$ がそれぞれの i_1 に対してとる値が変る様子によって測られる．もし項目 I_2 が何の情報も持たなければ $p(i_1|i_2)$ は i_2 に無関係に一定となり，

$$p(i_1|i_2)=p(i_1) \tag{1.1}$$

が成り立つ．一般に，

$$p(i_1,i_2)=p(i_1|i_2)p(i_2) \tag{1.2}$$

が成り立つから，(1.1) の関係を想定することは

$$p(i_1,i_2)=p(i_1)\,p(i_2) \tag{1.3}$$

を想定することに等しい．これは従来よく用いられてきた独立性のカイ自乗検定における帰無仮説にほかならない [7]．この関係式の意味するところは，あるサンプルが項目 I_2 に関してどのカテゴリーに属するかを知ってもそのサンプルが項目 I_1 に関してどのカテゴリーに属するかの予測に有効な情報とはならないということである．

冒頭の例のような3項目の関係の場合にも，同様の考え方に従って，「階層帰属意識」と「くらしむき」，「世帯収入」との間に種々の形の条件付独立を想定することによって種々の程度に単純化された構造をもつ確率分布の族を得ることができる．これらの分布の族の中から AIC によって現在のデータの特性を最もよく表現するものを選び出せば予測に有効な構造が得られると期待される．

表1は「階層帰属意識」と「くらしむき」とのいわゆる2次元のクロス表である．しかしこれは，「くらしむき」i_2 が与えられれば「世帯収入」i_3 の如何にかかわらず「階層帰属意識」i_1 の条件付確率が定まると想定した3次元のクロス表とみなすことができる．いいかえれば，任意の i_1, i_2, i_3 に対して，

$$p(i_1|i_2,i_3)=p(i_1|i_2,\cdot) \tag{1.4}$$

すなわち確率の定義から，

$$p(i_1,i_2,i_3)=p(i_1,i_2)\,p(i_2,i_3)/p(i_2) \tag{1.5}$$

というモデルを想定したことを意味する．

同様にして，表2に対応するモデルとしてつぎのものを得る．

$$p(i_1,i_2,i_3)=p(i_1,i_3)\,p(i_2,i_3)/p(i_3) \tag{1.6}$$

モデル (1.5)，(1.6) に対する AIC は，

$$\begin{aligned}\mathrm{AIC}^*(I_1;\,I_j)=(-2)\sum_{i_1,i_2,i_3} n(i_1,i_2,i_3)\ln\left[n(i_1,i_j)\,n(i_2,i_3)/\{n\cdot n(i_j)\}\right]\\ +2\{(C_1C_j-1)+(C_2C_3-1)-(C_j-1)\},\\ j=2,3\end{aligned} \tag{1.7}$$

によって与えられる．モデルの比較にとって不要な共通項 $(-2)\sum n(i_2,i_3)\ln[n(i_2,i_3)/n]+2(C_2C_3-1)$ を引き去り，さらに，従来の尤度比統計量（あるいはカイ自乗統計量）との対応を考慮して $2\sum n(i_1)\ln n(i_1)/n-2(C_1-1)$ を共通に加えるとつぎの統計量を得る．

$$\begin{aligned}\mathrm{AIC}(I_1;I_j)=(-2)\sum_{i_1,i_j} n(i_1,i_j)\ln\left[n\cdot n(i_1,i_j)/\{n(i_1)\,n(i_j)\}\right]+2(C_1-1)\,(C_j-1),\\ j=2,3\end{aligned} \tag{1.8}$$

この統計量はモデルの比較に必要な AIC の値を表1，表2だけから計算できることをしめしている．表1，表2に対してはそれぞれ $\mathrm{AIC}(I_1;I_2)=-402.49$，$\mathrm{AIC}(I_1;I_3)=-123.98$ が得られる．AIC の値の小さいモデルほどエントロピー [9] の大きいよいモデルと考えられ，最小の AIC をもつモデルは MAICE と呼ばれる [8], [9]．この例の場合前者が後者より小さいから，表1に対応するモデル (1.5) の方がよいモデルと考えられる．すなわち，「階層帰属意識」に

対しては「世帯収入」は無関係であると想定したモデル (1.5) の方が,「くらしむき」は無関係と想定したモデル (1.6) より真の確率分布に近いと判定されたのである．このことは「階層帰属意識」に対しては「世帯収入」より「くらしむき」の方が多くの情報を含んでいることを意味する．この判定結果は冒頭の直観的な判断と一致する．重要なことは，この判定結果が (1.8) で与えられる統計量の比較という客観的で簡単な操作によって得られたということである．

2. 説明変数の最適な組合せの選択

前節では，説明の便宜上，説明変数の次数をア・プリオリに1次に限定して考察を進めた．この節では多くの変数が説明変数の候補として与えられているときに最適な説明変数の組合せを求める方法について述べよう．

変数 $I_1, \cdots, I_k$ から成る k 次元分割表があり，その同時確率を $p(i_1, \cdots, i_k)$ で，対応する度数を $n(i_1, \cdots, i_k)$ であらわす．ここで，i_j は変数 I_j $(j=1, \cdots, k)$ のとる値 $1, 2, \cdots, C_j$ を表わすことにする．サンプル・サイズを n とすると，

$$\sum_{i_1, \cdots, i_k} n(i_1, \cdots, i_k) = n$$

がなりたつ．また，以下では記号を簡略にするために，

$$p(i_1, \cdots, i_{k-1}) = \sum_{i_k=1}^{C_k} p(i_1, \cdots, i_k),$$

$$n(i_1, \cdots, i_{k-2}) = \sum_{i_{k-1}=1}^{C_{k-1}} n(i_1, \cdots, i_{k-1})$$

などと書くことにする．

さて，変数の集合 $I=\{I_1, \cdots, I_k\}$ の互いに素な任意の部分集合を E, F とし，それらの同時確率を $p(E)$, $p(F)$ であらわす．目的変数 E に対して説明変数 F が持つ情報の量を評価するためにつぎのモデルを考える．

(2.1) $$\text{MODEL } (E;F): \quad p(I) = p(E|F)\, p(E^c)$$

ここで，$p(E|F)$ は F の下での E の条件付確率をしめし，$E^c=I-E$ である．また，$p(E|\phi)$ は $p(E)$ を意味することにする．たとえば目的変数 E として $\{I_1\}$ をとると，(2.1) によってつくりうるモデルは $F=\{I_2, \cdots, I_k\}$ とした最高次のモデル

$$\text{MODEL}(I_1; I_2, \cdots, I_k): \quad p(i_1, \cdots, i_k) = p(i_1|i_2, \cdots, i_k)\, p(i_2, \cdots, i_k)$$

をはじめとして，$F=\phi$ としたモデル

$$\text{MODEL}(I_1; \phi): \quad p(i_1, \cdots, i_k) = p(i_1)\, p(i_2, \cdots, i_k)$$

まで，2^{k-1} 個ある．

また，前節で考えたような2次元クロス表どうしの比較のためのモデルは，つぎの形の $k-1$ 個のモデルに相当する．

$$\text{MODEL}(I_1; I_j): p(i_1, \cdots, i_k) = p(i_1|i_j)\, p(i_2, \cdots, i_k), \quad j=2, \cdots, k$$

なお，モデル (2.1) は，$I=E \cup E^c$ から，

$$p(E|E^c) = p(E|F)$$

とも書くことができ，$E=\{I_1\}$ とすれば，

$$p(i_1|i_2, \cdots, i_k) = p(i_1|F), \quad F \subset \{I_2, \cdots, I_k\}$$

を意味する．したがってこのモデルはサブセット・リグレッションとよく似た考え方に立っていることがわかる．

モデル (2.1) に対する AIC は，

(2.2)　$$\mathrm{AIC}^*(E;F)=(-2)\sum_{E,F} n(E,F)\ln[n(E,F)\,n(E^c)/\{n\cdot n(F)\}] +2\{(C_E C_F-1)+(C_{E^c}-1)-(C_F-1)\}$$

で与えられる．ここで $C_\cdot$ は当該の変数の組のカテゴリー数をしめし，$C_\phi=1$, $n(\phi)=n$ と規約する．前節と同様に共通項を適当に調整して，

(2.3)　$$\mathrm{AIC}(E;F)=(-2)\sum_{E,F} n(E,F)\ln[n\cdot n(E,F)/\{n(E)\,n(F)\}] +2(C_E-1)(C_F-1)$$

としてもモデル比較上問題はない．ここで，$\sum$ は所与の分割表のすべてのセルにわたる総和を意味する．明らかに $\mathrm{AIC}(E;\phi)=0$ である．この統計量は与えられた（低次の）クロス表だけでモデルの比較が可能であることをしめしている．換言すると，与えられた分割表に明示されていない変数間の交互作用の如何を顧慮することなく分割表の比較ができるわけである．また，統計量 (2.3) から，説明変数の次数の不適切な上昇は第2項の値の増大を通じて AIC の増大を招くことがわかる．そこで，重回帰分析における変数増減法と同様に説明変数の組合せを変えながら次数をあげていくことにすれば評価すべき説明変数が数百に及んでもここでの目的を容易に達成することができる．

なお，2次元分割表における従来の独立性の検定は，AIC の立場から見ると [1, 2, 4]，(2.1) において $I=\{I_1,I_2\}$，$E=\{I_1\}$ とし，$F=\{I_2\}$ とおいて得られる無制約モデル

(2.4)　$$p(i_1,i_2)=p(i_1|i_2)\,p(i_2)=p(i_1,i_2)$$

と，$F=\phi$ とおいて得られる独立モデル

(2.5)　$$p(i_1,i_2)=p(i_1|\phi)\,p(i_2)=p(i_1)\,p(i_2)$$

との比較とみなされる．これらのモデルに対応する AIC を $\mathrm{AIC}^*(I_1;I_2)$, $\mathrm{AIC}^*(I_1;\phi)$ とし，従来独立性の検定に用いられてきたカイ2乗統計量を

$$\chi^2=\sum_{i_1,i_2}\{n(i_1,i_2)-n(i_1)\,n(i_2)/n\}^2/\{n(i_1)\,n(i_2)/n\}$$

とすると，(2.3) で $E=\{I_1\}$，$F=\{I_2\}$ とおいて得られる統計量 $\mathrm{AIC}(I_1;I_2)$ とこれらとの間には漸近的につぎの関係が成り立つ．

$$\mathrm{AIC}(I_1;I_2)$$

(2.6)　$$=\mathrm{AIC}^*(I_1;I_2)-\mathrm{AIC}^*(I_1;\phi)$$

(2.7)　$$\cong -(\chi^2-2\times\text{自由度})$$

この関係から，MAICE に従って $\mathrm{AIC}^*(I_1;I_2)$ と $\mathrm{AIC}^*(I_1;\phi)$ のうちで小さい値を与えるモデルを採択することは $\mathrm{AIC}^*(I_1;I_2)-\mathrm{AIC}^*(I_1;\phi)$ の値の正負によって，したがって $\mathrm{AIC}(I_1;I_2)$ の値の正負によって独立か否かを判定することを意味することが，また (2.7) から AIC の立場から見た独立性の意味がわかる．したがって統計量 (2.3) はその符号によって目的変数と説明変数とが独立か否かをしめすとともに（表4，表6の AIC の符号を参照のこと），その値によって説明変数の与える情報の多寡をしめす．

ところで，社会調査などでは調査項目相互間の関連の程度を調査項目全体にわたって比較できると好都合なことが多い．たとえば，2次元クロス表 $\{n(i_1,i_2)\}$ における変数 I_1 と I_2 の関連の強さと $\{n(i_3,i_4)\}$ における I_3 と I_4 の関連の強さとの比較がそうである．このため，つぎのモデルを考える．

(2.8)　$$\mathrm{MODEL}(I_j,I_l)\quad p(i_1,\cdots,i_k)=p(i_j,i_l)\prod_{s=1,s\neq j,l}^{k}p(i_s)$$

(2.9)　$$=\frac{p(i_j,i_l)}{p(i_j)\,p(i_l)}\prod_{s=1}^{k}p(i_s),\qquad j,l=1,\cdots,k,\ j\neq l.$$

モデル (2.8) に対する AIC は,

$$\text{(2.10)}\qquad \mathrm{AIC}^*(I_j, I_l) = (-2)\sum_{i_1,\cdots,i_k} n(i_1,\cdots i_k)\ln\left[\frac{n(i_j,i_l)}{n^{k-1}}\prod_{s=1,s\neq j,l}^{k} n(i_s)\right]$$
$$+2\left[(C_jC_l-1)+\sum_{s=1,s\neq j,l}^{k}(C_s-1)\right]$$

で与えられるが，モデルの比較上不要な共通項を無視して，

$$\text{(2.11)}\qquad \mathrm{AIC}\ (I_j, I_l) = (-2)\sum_{i_j,i_l} n(i_j,i_l)\ln\frac{n\cdot n(i_j,i_l)}{n(i_j)n(i_l)} + 2(C_j-1)(C_l-1)$$

によってもモデル比較が可能である．統計量 (2.11) はモデル (2.9) においてすべてのモデル比較上共通となる $\prod_{s=1}^{k} p(i_s)$ の項を無視したものに対応している．また，これは形式的には統計量 (2.3) において $E=\{I_j\}$, $F=\{I_l\}$ とおいたものに等しい．

3. 最適カテゴリーの設定

表3は表2の「世帯収入」の4つのカテゴリーをまとめなおして3つにしたものと「階層帰属意識」とのクロス表である．これらの表にしめされた2通りのカテゴリーの設定法のうちどちらがデータの特性をよく表現していると言えるだろうか．

このような場合，カテゴリーをプールする前と後の変数の値をそれぞれ別の変数の値とみなせば，前節までのモデルの考え方からその比較法は明らかである．すなわち，カテゴリーの区分法を変えて様々なクロス表をつくりそれぞれに対する統計量 (2.3) の値を比較し，その最小値を与える区分法を採択すればよい [3].

表3に対する AIC を $\mathrm{AIC}(I_1; I_3')$ とすれば $\mathrm{AIC}(I_1; I_3') = -109.86$ を得る．表2に対する AIC の値は −123.98 であったから，表3より表2のカテゴリゼーションの方が良いと判定される．なお，SSM 調査では「世帯収入」はもともと50万円キザミの97個のカテゴリーとして回答が得られている．この場合に上の方法を適用して得られた最適なカテゴリーは，4節で触れるように，25万未満，25〜75万円，75〜225万円，225〜375万円，375〜475万円，475万円以上の6カテゴリーであった．

ところで，このような現実のデータを処理する際にはプールの仕方は説明変数の性格（名義尺度，序数尺度等）や分析目的に応じて変えなければならない．プールの方式としてはおおよそつぎのような仕方が考えられよう．

i) 可能なすべての組み合せによるプール
ii) 相隣るカテゴリーの不等間隔なプール
iii) 等間隔のプール

なお，プログラム CATDAP-02 では変数ごとに上の ii), iii) のプールの形式を指定できる．

表 3

I_3' \ I_1		階層帰属意識 1. 上，中の上	2. 中の下	3. 下の上，下の下	計
世帯収入	1. 〜225万円	213	605	350	1168
	2. 225〜375	230	516	196	942
	3. 375〜	201	279	60	540
計		644	1400	606	2650

さらに，説明変数が連続型であっても，その観測値の観測精度を級間隔としてカテゴリー化し，上の統計量を用いて最適なカテゴリゼーションを見い出すことにすれば全く同様に処理できる．たとえば，観測値が18.3, 19.1, ･･･等と与えられていれば，0.1の級間隔で区分（カテゴリー化）すれば連続変量も分割表に帰着させることができる．

こうして，ここで提案した方法は，概念的に言えば，所与の説明変数によってつくりうる最大限のセルをもつ多重クロス表を基礎におくが，説明変数のとる値を条件とした目的変数の条件付確率分布が同一とみなしうる限りで説明変数のカテゴリーをプールし，パラメータの減少を図り，データの無意味な統計的変動の影響を押え，有意な関係を抽き出す方法である．この立場から見れば変数選択の問題も最適カテゴリーゼーションの問題も同一の問題とみなすことができる．こうして，ある目的変数に対する最適な説明変数としてたとえば〔性×年令〕が合計4カテゴリーで採択されても〔年令〕だけが4カテゴリーで採択されても，実体的な意味はもちろん異なるが，この方法の上では何ら区別されるところはない．とはいえ，現実の調査データにこの統計量を適用した経験から言えば，説明変数を変えることによるAICの値の変化は大きいが，カテゴリー数に極端な差がない限りカテゴリーゼーションの違いによる値の変化はさほど大きくない．

なお，目的変数のカテゴリー数が2のときの最適なカテゴリー化については文献[10]でも論じられている．

4. 処理の実例

まず，これまでの表でもとりあげた1975年SSM調査における「階層帰属意識」を例にとろう．1975年SSM全国調査は「社会階層と階層移動に関する調査（A調査）」と「職業威信調査（B調査）」の2本立てで行なわれ，前者は職歴，家庭環境，学歴，収入，社会的地位，財産所有状況，満足感等に関する100以上の質問から成っている．ここでは，A調査の回答者のうち学生74人を除いた2650人の結果に基づいて，「階層帰属意識」に対してどのような項目が多くの情報をもっているかを検討してみよう．このため，前節までに述べた方法をプログラム化したCATDAP-02を用いると表4のようなアウト・プットが得られる．ここで，目的変数の「階層帰属意識」は表1〜3にしめした3カテゴリーとし，各説明変数はそのカテゴリーが順序をもっている限り前節ii）の方式で自動的なカテゴリーゼーションを行なうよう入力時に指定した．また，「無答」等はあらかじめ中間的なカテゴリーに含めてしまった場合もある．なお，分析でとりあげた項目数は73項目である．

表4の右側から2列目の項目名の配列を見ればわかるように，「階層帰属意識」に対しては表1でもとりあげた「くらしむき」の評価が1番多くの情報をもち，以下，「階級帰属意識」，「所有財産の種類の数」，「生活に対する満足度」となっており，表2の「世帯収入」は5番目にでてくる．また，これらの順位に対応するクロス表は表5のように得られる．表5の右端の数字は説明変数のプールする前のカテゴリーをしめしており，たとえば1行目の'2'は「くらしむき」の'1'というカテゴリーは当初のカテゴリーの1から2まで（'非常に豊か'と'やや豊か'に相当）をプールしたものであることを意味する．このクロス表から，「くらしむき」が豊かな人は「階層帰属意識」も高いことが見られる．なお，「所有財産の種類の数」は表3の9番目に出てくる応接セット，10番目の電子レンジ等合計20項目にわたる財産のうち何種類所有しているかを説明変数としてとりあげたもので，種類数の級境界値はやはり表4の最右列にしめされている．このようにこのプログラムでは当初の説明変数を何らかの形で加工したものも当初の変数と同時に説明変数の候補とみなして処理することができる．

表 4

LIST OF EXPLANATORY VARIABLES ARRANGED IN ASCENDING ORDER OF AIC
RESPONSE VARIABLE NAME: (Q18 KAISO)

NO.	EXPLANATORY VARIABLE NAME	NUMBER OF CATEGORIES OF EXPLANATORY VARIABLE	AIC	DIFFERENCE OF AIC	〔階層帰属意識〕	〔階級帰属意識〕
1	Q24 KURAS	4	−406.36	0.0	(くらしむき)	(所有財産の種類数)
2	Q18B KAIKY	3	−213.65	192.71	(階級帰属意識)	(階層帰属意識)
3	*ZAISAN SU	6	−170.49	43.16	(所有財産の種類数)	(本人の従業上の地位)
4	Q9-5 SEIKA	5	−150.09	20.41	(生活に対する満足度)	(くらしむき)
5	Q26A SYU-S	6	−129.65	20.44	(世帯の収入)	(本人の収入)
6	Q2 15KUR	4	−118.30	11.35	(15才時のくらしむき)	(世帯の収入)
7	Q9-3 SYUNU	3	−101.84	16.46	(収入に対する満足度)	(企業の経営者とのつきあいの程度)
8	Q26B SYU-H	5	−88.08	13.76	(本人の収入)	(株券あるか)
9	25-14 OSETU	2	−74.70	13.37	(応接セットあるか)	(電子レンジあるか)
10	25-7 RANGE	2	−72.24	2.46	(電子レンジあるか)	(スポーツ会員権あるか)
11	Q19B YOKAM	3	−63.45	8.80	(余暇に対する満足度)	(応接セットあるか)
12	25-6 CAR	2	−63.28	0.17	(乗用車あるか)	(エア・コンディショナーあるか)
13	Q19-6 TRIP	3	−58.58	4.70	(国内旅行の頻度)	(本人の学歴)
14	Q20-3 SYACH	3	−58.22	0.36	(企業経営者とのつきあいの程度)	(大学の先生とのつきあいの程度)
15	Q21-1 SYOKU	3	−57.70	0.52	(職場での影響力の程度)	(ゴルフの頻度)
16	Q3A GAKU	4	−55.20	2.49	(本人の学歴)	(支持政党)
17	Q9-1 SIGOT	3	−55.00	0.20	(仕事に対する満足度)	(15才時のくらしむき)
18	Q19-2 PARTY	3	−54.01	0.99	(友人との会食の頻度)	(地域社会での影響力の程度)
19	Q20B TUKIM	3	−47.78	6.23	(つきあいに対する満足度)	(職場での影響力の程度)
20	25-18 RADIO	2	−45.17	2.61	(携帯用ラジオあるか)	(ピアノあるか)
21	Q9-4 GAKU	4	−43.86	1.31	(学歴に対する満足度)	(収入に対する満足度)
22	Q9-2 TUTOM	4	−43.29	0.57	(勤め先に対する満足度)	(地方議会議員とのつきあいの程度)
23	Q21-4 CHIIK	3	−41.59	1.70	(地域社会での影響力の程度)	(父親の従業上の地位——経営者, 役員)
24	Q20-1 GIIN	3	−41.43	0.16	(地方議会議員とのつきあいの程度)	(国内旅行の頻度)
25	25-5 CAMER	2	−40.75	0.69	(カメラあるか)	(貸付信託あるか)
26	25-4 YUWAK	2	−40.04	0.71	(ガス瞬間湯沸器あるか)	(本人の職場の雇用者数)
27	Q20-4 PROF	2	−39.07	0.96	(大学の先生とのつきあいの程度)	(セントラル・ヒーティングあるか)
28	Q8-2 GANBA	2	−39.00	0.07	(性格——目標たてて頑張る)	(年令)
29	Q4A HCHII	6	−37.14	1.86	(本人の従業上の地位)	(自家風呂あるか)
30	Q16 NOZOM	8	−37.00	0.14	(望ましい仕事の第一条件)	(父の従業上の地位——一般従業者)
31	Q19-8 SIBAI	3	−36.03	0.97	(芝居見物・コンサートなどの頻度)	(電話あるか)
32	25-15 STOCK	2	−34.65	1.38	(株券あるか)	(ガス瞬間湯沸器あるか)
33	Q21-3 DANTA	2	−32.29	2.36	(青年会, サークルでの影響力の程度)	(芝居見物・コンサートなどの頻度)
34	Q21B HATUM	4	−28.96	3.33	(仲間, 会, 団体の影響力に対する満足度)	(つきあいに対する満足度)
35	25-12 TOCHI	2	−26.61	2.35	(宅地あるか)	(学歴に対する満足度)
36	Q25-1 FURO	2	−26.25	0.36	(自家風呂あるか)	(持家あるか)
37	25-19 TEL.	2	−25.48	0.77	(電話あるか)	(友人との会食の頻度)
38	25-11 CENTR	2	−24.09	1.39	(セントラル・ヒーティングあるか)	(性格——世話好き)
39	Q21-2 JICHI	3	−23.44	0.65	(町内会や自治会での影響力の程度)	(宅地あるか)
40	25-17 COOL	2	−23.31	0.13	(エア・コンディショナーあるか)	(ステレオあるか)
41	Q8-3 NONKI	2	−22.63	0.68	(性格——のんきに人に従う)	(仕事に対する満足度)
42	25-8 PIANO	2	−22.34	0.30	(ピアノあるか)	(読書の頻度)
43	Q8-1 NECHU	2	−22.23	0.11	(性格——物事に熱中し, ものにする)	(乗用車あるか)
44	25-10 KAIIN	2	−21.91	0.31	(スポーツ会員権あるか)	(妻の父の従業上の地位)
45	Q22 SEITO	4	−21.25	0.66	(支持政党)	(性格——目標たてて頑張る)
46	25-13 IE	2	−21.16	0.09	(持家あるか)	(生活に対する満足度)
47	25-9 TV.	2	−21.09	0.06	(カラーテレビあるか)	(性格——お山の大将好き)
48	Q3C NAME	10	−19.86	1.23	(本人の学歴——どこの大学)	(望ましい仕事の第一条件)
49	Q15 TENSY	3	−18.58	1.28	(転職に賛否)	(カメラあるか)
50	Q19-3 GOLF	2	−17.40	1.17	(ゴルフ・テニスなどの頻度)	(本人の学歴——どこの大学)
51	Q12 HAHAG	4	−16.13	1.27	(母の学歴)	(青年会・サークルでの影響力の程度)
52	Q8-6 OYAMA	2	−15.13	1.00	(性格——お山の大将好き)	(町内会などの役員とのつきあいの程度)
53	Q1 NENRE	6	−14.11	1.02	(年令)	(性格——のんきに人に従う)
54	25-20 SINTA	2	−13.70	0.42	(貸付信託あるか)	(海外旅行の頻度)
55	Q8-4 KIMAM	2	−13.68	0.01	(性格——きままにのんびりやる主義)	(全収に対する満足度)
56	Q20-2 JICHI	2	−12.62	1.06	(町内会などの役員とのつきあいの程度)	(町内会や自治会での影響力の程度)
57	25-16 STERE	2	−12.44	0.19	(ステレオあるか)	(勤め先に対する満足度)
58	Q19-7 OVERS	2	−10.70	1.74	(海外旅行の頻度)	(携帯用ラジオあるか)
59	Q19-5 READ	2	−10.64	0.06	(読書の頻度)	(父の学歴)
60	Q8-5 MENDO	2	−10.49	0.15	(性格——世話好き)	(仲間, 会, 団体の影響力に対する満足度)
61	Q19-4 TOZAN	3	−9.77	0.72	(登山, ハイキングなどの頻度)	(母の学歴)
62	Q17 SYUSE	11	−8.00	1.77	(立身出世の第一条件)	(転職に賛否)
63	Q19-9 KEIKO	2	−7.16	0.83	(楽器演奏, おけいこごとの頻度)	(カラーテレビあるか)
64	Q11 CHII1	8	−6.53	0.63	(父の従業上の地位——経営者, 役員)	(既婚か未婚か)
65	Q14 H-P-C	8	−4.53	2.00	(妻の父の従業上の地位)	(楽器演奏, おけいこごとの頻度)
66	25-3 REIZO	2	−1.82	2.71	(電気冷蔵庫あるか)	(性格——物事に熱中し, ものにする)
67	Q13 KEKON	5	−0.33	1.49	(既婚か未婚か)	(電気冷蔵庫あるか)
68	Q10 PAPAG	4	−0.14	0.19	(父の学歴)	(立身出世の第一条件)
69	Q25-2 BESSO	2	1.78	1.92	(別荘あるか)	(性格——きままにのんびりやる主義)
70	Q19-1 EIGA	2	2.57	0.78	(映画の頻度)	(別荘あるか)
71	Q4F INZU	2	3.38	0.82	(本人の職場の雇用者数)	(登山・ハイキングなどの頻度)
72	Q11 CHII2	8	8.65	5.27	(父の従業上の地位——一般従業者)	(映画の頻度)

ところで，直井がこのSSM調査データの，経験をまじえた吟味に基づいて，「階層帰属意識を分化させる効果が最も大きいのは『くらしむき』変数である」と結論していることは既に述べた．従って，われわれは何らの検討・予備知識なしに客観的な方法で瞬時にして同一の結論に達したことになる．しかし，ここで注意すべき点はわれわれの結論が直井が行なった7変数のみによる検討からではなく，目の子による検討では厖大な時間を要する72変数との分析から得られたということである．

表6は最適な変数の組合せを探索する過程で検討された変数の組合せをAICの値の小さい順に配列しなおしたものである．この表から，「くらしむき×階級帰属意識×収入に対する満足度」という3変数の組が最適な組み合せであることがわかる．2次元クロス表の分析では7

表 5

TWO-WAY TABLES WITH AN OPTIMAL CATEGORIZATION OF EACH SINGLE EXPLANATORY VARIABLE

RESPONSE VARIABLE NAME : Q18 KAISO

(Q24 KURAS) \ (Q18 KAISO)	1	2	3		CLASS INTERVAL
1	202	104	25	331	2
2	4181183	4072008			3
3	22	105	143	270	4
4	2	8	31	41	5
TOTAL	6441400	6062650			

(Q24 KURAS) \ (Q18 KAISO)	1	2	3		
1	61.0	31.4	7.6	100.0	2
2	20.8	58.9	20.3	100.0	3
3	8.1	38.9	53.0	100.0	4
4	4.9	19.5	75.6	100.0	5
TOTAL	24.3	52.8	22.9	100.0	

(Q18B KAIKY) \ (Q18 KAISO)	1	2	3		CLASS INTERVAL
1	3361051	5341921			1
2	250	291	48	589	2
3	58	58	24	140	3
TOTAL	6441400	6062650			

(Q18B KAIKY) \ (Q18 KAISO)	1	2	3		
1	17.5	54.7	27.8	100.0	1
2	42.4	49.4	8.1	100.0	2
3	41.4	41.4	17.1	100.0	3
TOTAL	24.3	52.8	22.9	100.0	

(HZAISAN SU) \ (Q18 KAISO)	1	2	3		CLASS INTERVAL
1	14	78	63	155	4
2	72	249	163	484	7
3	103	344	141	588	9
4	176	347	136	659	11
5	232	344	99	675	15
6	47	38	4	89	20
TOTAL	6441400	6062650			

(HZAISAN SU) \ (Q18 KAISO)	1	2	3		
1	9.0	50.3	40.6	100.0	4
2	14.9	51.4	33.7	100.0	7
3	17.5	58.5	24.0	100.0	9
4	26.7	52.7	20.6	100.0	11
5	34.4	51.0	14.7	100.0	15
6	52.8	42.7	4.5	100.0	20
TOTAL	24.3	52.8	22.9	100.0	

(Q9-5 SEIKA) \ (Q18 KAISO)	1	2	3		CLASS INTERVAL
1	116	116	44	276	1
2	340	664	2031207		2
3	140	433	205	778	3
4	35	152	107	294	4
5	13	35	47	95	5
TOTAL	6441400	6062650			

(Q9-5 SEIKA) \ (Q18 KAISO)	1	2	3		
1	42.0	42.0	15.9	100.0	1
2	28.2	55.0	16.8	100.0	2
3	18.0	55.7	26.3	100.0	3
4	11.9	51.7	36.4	100.0	4
5	13.7	36.8	49.5	100.0	5
TOTAL	24.3	52.8	22.9	100.0	

(Q26A SYU-S) \ (Q18 KAISO)	1	2	3		CLASS INTERVAL
1	32	67	31	130	1
2	7	13	21	41	2
3	174	525	298	997	5
4	230	516	196	942	8
5	80	154	40	274	10
6	121	125	20	266	97
TOTAL	6441400	6062650			

(Q26A SYU-S) \ (Q18 KAISO)	1	2	3		
1	24.6	51.5	23.8	100.0	1
2	17.1	31.7	51.2	100.0	2
3	17.5	52.7	29.9	100.0	5
4	24.4	54.8	20.8	100.0	8
5	29.2	56.2	14.6	100.0	10
6	45.5	47.0	7.5	100.0	97
TOTAL	24.3	52.8	22.9	100.0	

(Q2 15KUR) \ (Q18 KAISO)	1	2	3	CLASS INTERVAL
1	111	107	38	256
2	395	910	3001605	
3	114	294	184	592
4	24	89	84	197
TOTAL	6441400	6062650		

(Q2 15KUR) \ (Q18 KAISO)	1	2	

位であった「収入に対する満足度」が最適な組合せの中にくい込んできた点が興味深い．表7はこの組合せに対応するクロス表を実数とパーセンテージでしめしたものである．CATDAP-02によって最適とされた各変数のカテゴリゼーションは表7の下部にしめされている．この表から，「くらしむき」の回答がカテゴリー'1'か'2'に属し，「階級帰属意識」が'1'で，「収入に対する満足度」が'2'以下の人は89人中43人が「中の上」より高い「階層帰属意識」をもっていること，等がわかる．こうして，あるサンプルの上記3項目に対する回答が与えられればそのサンプルの「階層帰属意識」を確率的に予測することができる．なお，表4，表6は当該の組合せの下での最適なカテゴリゼーションに対応するAICの値のみをしめしている．

ところで，表4の最右列は，資本家，中産，労働者のうちどの階級に属するかという「階級

表 6

SUMMARY OF SUBSETS OF EXPLANATORY VARIABLES

RESPONSE VARIABLE NAME : Q18 KAISU

	EXPLANATORY VARIABLE NAME	NUMBER OF CATEGORIES OF EXPLANATORY VARIABLE(S)	A I C	DIFFERENCE OF AIC
1	Q24 KURAS Q18B KAIKY Q9-3 SYUNU	18	-517.70	0.0
2	Q24 KURAS Q18B KAIKY	9	-514.87	2.83
3	Q24 KURAS Q18B KAIKY 25-6 CAR	18	-513.86	1.01
4	Q24 KURAS Q18B KAIKY 25-7 RANGE	18	-511.27	2.59
5	Q24 KURAS Q18B KAIKY 25-14OSETU	18	-499.62	11.65
6	Q24 KURAS Q18B KAIKY Q19-2PARTY	27	-494.54	5.08
7	Q24 KURAS Q18B KAIKY Q9-3 SYUNU 25-7 RANGE	36	-487.41	7.13
8	Q24 KURAS Q18B KAIKY Q9-3 SYUNU 25-6 CAR	36	-487.31	0.11
9	Q24 KURAS Q18B KAIKY Q21-1SYOKU	27	-484.74	2.57
10	Q24 KURAS Q18B KAIKY Q9-1 SIGOT	27	-484.60	0.14
11	Q24 KURAS Q18B KAIKY Q20B TUKIM	27	-484.60	0.00
12	Q24 KURAS Q18B KAIKY Q19-6 TRIP	27	-481.97	2.63
13	Q24 KURAS Q18B KAIKY Q19B YOKAM	27	-480.17	1.80
14	Q24 KURAS Q18B KAIKY Q9-5 SEIKA	45	-479.82	0.34
15	Q24 KURAS Q18B KAIKY Q9-3 SYUNU 25-14OSETU	36	-475.77	4.06
16	Q24 KURAS Q18B KAIKY Q2 15KUR	36	-474.88	0.88
17	Q24 KURAS Q18B KAIKY Q20-3SYACH	27	-471.95	2.93
18	Q24 KURAS Q18B KAIKY Q3A GAKU	36	-457.05	14.90
19	Q24 KURAS Q9-3 SYUNU 25-7 RANGE	12	-452.18	4.87
20	Q24 KURAS Q9-5 SEIKA	15	-451.34	0.84
21	Q24 KURAS Q18B KAIKY Q9-3 SYUNU Q19-2PARTY	54	-450.02	1.33
22	Q24 KURAS Q18B KAIKY #ZAISAN SU	45	-445.08	4.93
23	Q24 KURAS #ZAISAN SU	15	-437.62	7.46
24	Q24 KURAS 25-7 RANGE	6	-437.24	0.38
25	Q24 KURAS Q2 15KUR	12	-433.91	3.33
26	Q24 KURAS Q18B KAIKY Q9-3 SYUNU Q21-1SYOKU	54	-431.09	2.82
27	Q24 KURAS Q18B KAIKY Q9-3 SYUNU Q20B TUKIM	54	-428.32	2.76
28	Q24 KURAS 25-6 CAR	6	-426.88	1.44
29	Q24 KURAS Q18B KAIKY Q9-3 SYUNU Q19B YOKAM	54	-426.51	0.38
30	Q24 KURAS Q9-3 SYUNU	6	-425.15	1.35
31	Q24 KURAS Q18B KAIKY Q9-3 SYUNU Q19-6 TRIP	54	-424.59	0.57
32	Q24 KURAS Q18B KAIKY Q9-3 SYUNU Q9-1 SIGOT	54	-423.23	1.36
33	Q24 KURAS Q19-2PARTY	9	-421.33	1.90
34	Q24 KURAS Q9-1 SIGOT	9	-419.66	1.67
35	Q24 KURAS Q18B KAIKY Q9-3 SYUNU Q20-3SYACH	54	-417.27	2.39
36	Q24 KURAS Q18B KAIKY Q26B SYU-H	45	-416.56	0.71
37	Q24 KURAS Q3A GAKU	12	-416.50	0.06
38	Q24 KURAS Q20-3SYACH	9	-415.48	1.03
39	Q24 KURAS 25-14OSETU	6	-415.24	0.23
40	Q24 KURAS Q21-1SYOKU	9	-414.97	0.27
41	Q24 KURAS Q26A SYU-S	18	-413.16	1.81
42	Q24 KURAS Q20B TUKIM	9	-412.73	0.43
43	Q24 KURAS Q19-6 TRIP	9	-412.39	0.34
44	Q24 KURAS Q26B SYU-H	15	-411.51	0.88
45	Q24 KURAS Q19B YOKAM	9	-407.85	3.66
46	Q24 KURAS	3	-402.49	5.36
47	Q24 KURAS Q18B KAIKY Q26A SYU-S	54	-402.46	0.03
48	Q24 KURAS Q18B KAIKY Q9-3 SYUNU Q2 15KUR	72	-400.14	2.32
49	Q24 KURAS Q18B KAIKY Q9-3 SYUNU Q3A GAKU	72	-356.38	43.76
50	Q24 KURAS Q18B KAIKY Q9-3 SYUNU Q9-5 SEIKA	90	-341.26	15.12
51	Q24 KURAS Q18B KAIKY Q9-3 SYUNU #ZAISAN SU	90	-322.11	19.15
52	Q18B KAIKY Q9-3 SYUNU 25-7 RANGE	12	-298.44	23.67
53	Q24 KURAS Q18B KAIKY Q9-3 SYUNU Q26B SYU-H	90	-295.98	2.46
54	Q18B KAIKY Q9-3 SYUNU	6	-271.26	24.72
55	Q24 KURAS Q18B KAIKY Q9-3 SYUNU Q26A SYU-S	108	-256.16	15.09
56	Q18B KAIKY	3	-213.65	42.51
57	#ZAISAN SU	5	-169.65	44.00
58	Q9-5 SEIKA	5	-150.09	19.56
59	Q26A SYU-S	6	-129.65	20.44
60	Q2 15KUR	4	-118.30	11.35
61	Q26B SYU-H	5	-88.08	30.23
62	Q9-3 SYUNU	2	-79.19	8.89
63	25-14OSETU	2	-74.70	4.49
64	25-7 RANGE	2	-72.24	2.46
65	Q19B YOKAM	3	-63.45	8.80
66	25-6 CAR	2	-63.28	0.17
67	Q19-6 TRIP	3	-58.58	4.70
68	Q20-3SYACH	3	-58.22	0.36
69	Q21-1SYOKU	3	-57.70	0.52
70	Q3A GAKU	4	-55.20	2.49
71	Q9-1 SIGOT	3	-55.00	0.20
72	Q19-2PARTY	3	-54.01	0.99
73	Q20B TUKIM	3	-47.78	6.23
74	- - -	0	0.0	47.78

帰属意識」を目的変数とみなして同様の処理をした結果で，情報の多い順に項目の略称だけをしめしたものである．この配列を左側の「階層帰属意識」の配列と対比すると，「階級帰属意識」の場合には財産，地位，収入等のいわば客観的・地位的な項目が上位を占めるのに対して，「階層帰属意識」の場合にはくらしむき，満足度等の主観的・情緒的な項目が多く並んでいる．したがって回答者に受け取られた質問内容はこれら2問の間で相当異なっていると考えられる．CATDAP を適用して容易に実現されるこのような吟味が最近マスコミをにぎわしたいわゆる「新中間層」論争の進展に寄与するところは大きいと考えられる．

上のように，調査票に盛られた質問項目をつぎつぎに目的変数とみなして統計量 (2.11) を適用すると，全ての――質問数が k のとき ${}_kC_2$ 通りの――質問間の関連を見ることができる．図1は1973年の「日本人の国民性調査（K型調査票）」[11] の全質問についてこの値を計算し，その値の小大を記号の重複度（濃淡）で表現し一目で質問間の関連の程度を見られるようにし

表 7

CONTINGENCY TABLE WITH THE OPTIMAL COMBINATION AND CATEGORIZATION OF EXPLANATORY VARIABLES

X(1):Q18 KAISO X(2):Q24 KURAS X(3):Q18B KAIKY X(4):Q9-3 SYUNU

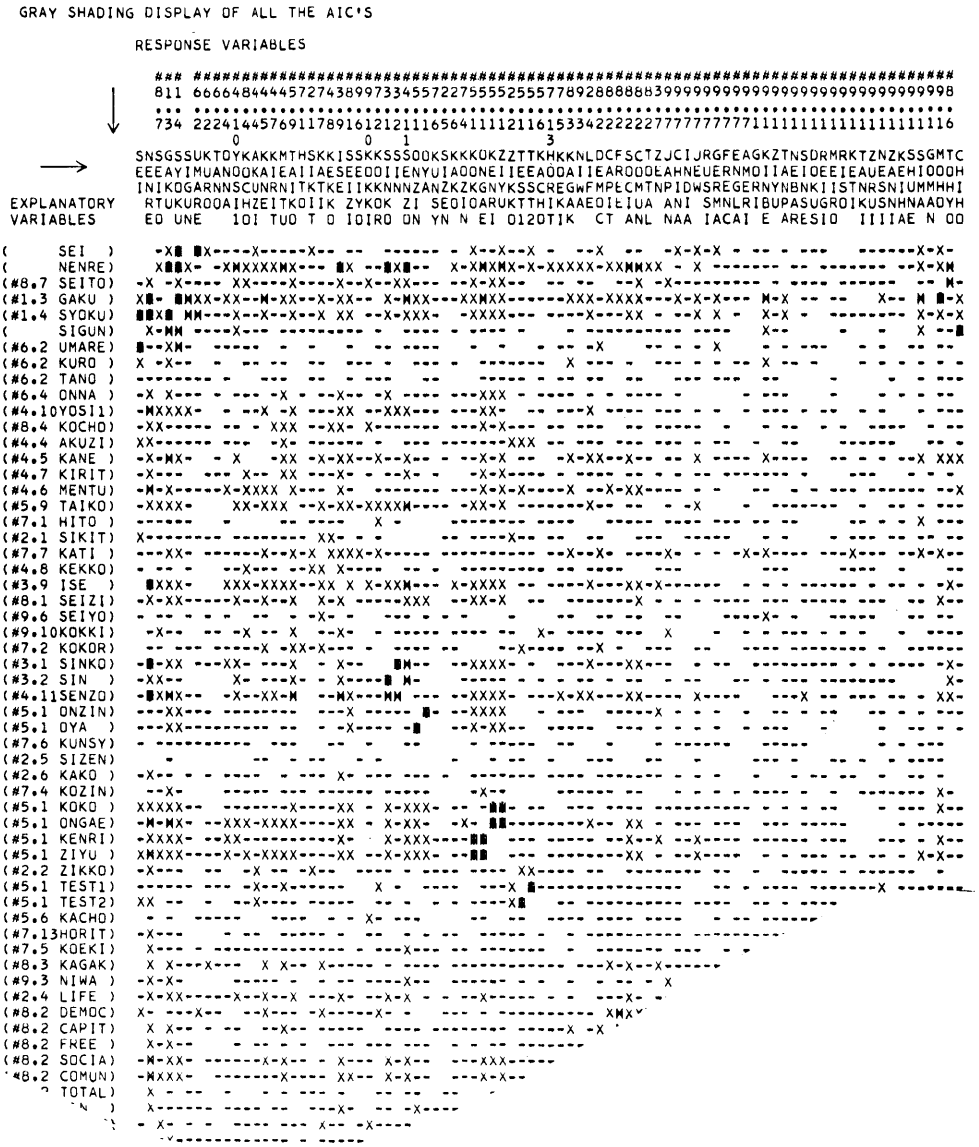
X X X - - - 2 3 4	RESPONSE VARIABLE			
X(1)	1	2	3	
1 1 1	43	36	10	89
1 1 2	33	27	10	70
1 2 1	68	25	1	94
1 2 2	35	9	4	48
1 3 1	19	4	0	23
1 3 2	4	3	0	7
2 1 1	98	299	109	506
2 1 2	144	592	239	975
2 2 1	81	93	14	188
2 2 2	63	150	23	236
2 3 1	21	13	6	40
2 3 2	11	36	16	63
3 1 1	4	15	30	49
3 1 2	14	82	136	232
3 2 1	0	0	1	1
3 2 2	3	14	5	22
3 3 1	1	0	1	2
3 3 2	2	2	1	5
TOTAL	644	1400	606	2650

X(1)	1	2	3	
1 1 1	48.3	40.4	11.2	100.0
1 1 2	47.1	38.6	14.3	100.0
1 2 1	72.3	26.6	1.1	100.0
1 2 2	72.9	18.8	8.3	100.0
1 3 1	82.6	17.4	0.0	100.0
1 3 2	57.1	42.9	0.0	100.0
2 1 1	19.4	59.1	21.5	100.0
2 1 2	14.8	60.7	24.5	100.0
2 2 1	43.1	49.5	7.4	100.0
2 2 2	26.7	63.6	9.7	100.0
2 3 1	52.5	32.5	15.0	100.0
2 3 2	17.5	57.1	25.4	100.0
3 1 1	8.2	30.6	61.2	100.0
3 1 2	6.0	35.3	58.6	100.0
3 2 1	0.0	0.0	100.0	100.0
3 2 2	13.6	63.6	22.7	100.0
3 3 1	50.0	0.0	50.0	100.0
3 3 2	40.0	40.0	20.0	100.0
TOTAL	24.3	52.8	22.9	100.0

CLASS INTERVAL

X (1)	: Q18 KAISO	1 :	1
X (1)	:	2 :	2
X (1)	:	3 :	3
X (2)	: Q24 KURAS	1 :	2
X (2)	:	2 :	3
X (2)	:	3 :	5
X (3)	: Q18B KAIKY	1 :	1
X (3)	:	2 :	2
X (3)	:	3 :	3
X (4)	: Q9-3 SYUNU	1 :	2
X (4)	:	2 :	5

図 1



たものの一部で，CATDAP-01 のアウト・プットの1つとして得られる．この図では濃い色の箇所に対応する変数どうしほど関連が強く，白地のところはその欄に対応する変数どおしが独立——AIC の立場から見た独立——であることを意味している．したがって，たとえば図1の4行2列（あるいは2行4列）は1番濃い色だからこの欄に対応する学歴と年令は強い関連があり，2行1列の年令と性は白地だから独立，… 等というように変数間の関連に一目で見当がつけられる．この図から，ある程度のサンプル・サイズをもつ調査データにおいては，独立とみなされる項目がいかに少ないかも読みとることができる．従来のカイ自乗検定が実用性をもち得なかった理由の一端はここにもあるのである．また，年令の列を縦（あるいは横）に見ると他の列にくらべて濃い色の欄が多く，国民性調査のデータでは年令と関連の強い項目が多

いことも分る．ところで，この図の各項目は調査票の質問順にならべられているから対角線に近いほど質問番号が近いことになる．従って，この図で対角線の近くに濃い色の欄が多いということは，接近した質問どうしは強い関連をもつことをしめす．これは質問形式に起因するみかけ上の関連をしめしている場合もあるが，内容の似た質問は近くにまとめられることが多いという調査票の質問の配列法や前の質問の回答が次の質問に影響を与えるという世論調査に特有の事情を反映している場合もある．いずれにせよ図1によって大量のデータにおける変数間の相互関係を1枚の図表に縮約することの効用は極めて大きい．

なお，この図では5段階に分けたが，AICの値のランクづけの仕方を工夫して，他の統計調査によって事前にその関連度に見当のつけられる項目――年令と学歴，学歴と性等――をいくつか選び，それらの項目の値によって段階づけをし，新たな項目間の関連度を直観的・経験的に把握できるようにする方法なども考えられる．

つぎに，CATDAP-02の出入力の例として文献[5]でとりあげられているのはいわゆる連続型変数とカテゴリカルな変数が混在しているデータである．このデータ[12, 171頁]は循環器系の集団検診例で，眼底所見，心電図所見，疾患（脳出血，脳梗塞，心筋梗塞，狭心症）の3つのカテゴリカルな変数と，年令（1才きざみ），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の5つの連続型変数から成る52例のデータである．このデータの解析については文献[4]を参照されたい．

5. CATDAPと他の方法との関係

統計調査において2組の変数をとりあげ相互の関係を見る場合，目的を異にする2通りの見方がある．1つは，たとえば「各収入階層の何%の人が中流意識をもっているか」という問題のように，目的変数も説明変数も，それぞれのカテゴリーの仕方もあらかじめ指定されていて，両変数の関係を表現する数値を推定ないしは確定することが目的とされる場合である．他の1つは，「階層帰属意識の規定要因は何か」というような研究課題に応えるために，目的変数だけが指定されていてそれに対して多くの情報を持っている説明変数を探すことが目的になっている場合である．社会調査などでは同一の意図をもってしても幾通りもの質問文が考えられるが，特定の質問文が調査者の意図どおりの内容・文脈で被調査者に理解されているかどうかがわからない場合も多い．そこで，着目した質問がどういう質問項目と強い関連を持っているかを見ることによって質問文の妥当性の吟味の一助とすることが考えられる．説明変数の検索はこのような状況でも必要になる．

変数間の関連性の程度を評価し重要な変数を探すという試みはカテゴリカルデータの分野においてもことさら新しい問題という訳ではない．クラメールの係数のようなカイ自乗統計量に修正を加えた様々な統計量や相関比等は従来のデータ解析でしばしば見られる例である．ただ，これらはいわば記述統計学的な立場からの統計量であって，所与のデータが全数調査によるものであろうとランダム・サンプリングによるものであろうと何ら区別せず，与えられたクロス表における関連の程度を数値的に表現することをねらったものである．いうまでもなく，これらの統計量によるならば，説明変数の次数やカテゴリー数の増加はどのような尺度を用いても常に「関連度」の増大をもたらす．したがって，この立場にあっては経験的な判断を加味しない限り説明変数の選択はできない．表1～3の場合でもこれらの統計量によっては変数の効果の3表間の比較は不可能であり，したがってもちろん，もともと50万円きざみの97個のカテゴリーとして回答が得られている「世帯収入」の最適なカテゴリー化の決定も不可能である．

これに対して本稿の方法は各クロス表に対応するモデルの尤度の比較によって構造（確率分

布）の推定をめざす方法である．この立場から見れば，説明変数の選択とは前述のモデルの中で最大の尤度をもつモデルを採択することを意味する．

検定や推定という統計学的な概念を基礎に置いてカテゴリカルな変数間の関係を評価する方法としては対数線型モデルを用いる方法が最も一般的である．しかしながら従来の方法はモデルの構成の不適切さとモデルの悪さの統一的な評価基準の欠如という2つの理由から現実のデータを処理しきれなかった．統計調査は非常に多くの質問項目から成る調査票を用いて行われるのが常態で，100問を越えることもことさらめずらしいことではない．また既に実例を見たように，調査データはカテゴリカルな変数ばかりでなく連続型の変数も同時に含んでいるのが普通である．ところが，従来の考え方・方法に従う限り，質問が100項目あれば100次元分割表をつくりそこに現われる全変数相互間のありとあらゆるタイプの交互作用について考察せざるをえず，事実上分析不能に陥る．いうまでもなく，クロス表の次数の増大は検討さるべきモデル数の指数関数的増加を招くばかりでなく，有限なサンプルを尨大な数のセルにバラまく不合理からほとんどのセルの観測度数が0になってしまうという理由に因る．しかしながら分析の目的によっては説明変数の数に等しい次数をもつクロス表をつくる必要はない．目的変数に対する説明変数の効果の比較という目的に即応したモデルを構成し，すべてのモデルの比較を可能にする統一的な評価基準を導入すれば説明変数の候補となる変数がいくら多くても低次のクロス表だけの比較によって所期の目的を達成することができるのである．

われわれの日常の経験をふり返ってみると，たとえデータの項目数が何百あっても質問数とは無関係に，問1と問2は関連が強く，問1と問3は無関係で，…というような直観的な評価を行なっている．ここで重要なことは，問1という特定の質問に対する問2や問3の関係を考察する際には問2と問3の関係の如何は無視できるということである．この着想を基礎に，説明変数相互間の関係には何の制約も加えず，目的変数と説明変数との間だけに条件付確率という制約を種々の形で課すことによって得られたのが本稿のモデルである．このモデルにAICを適用し，モデル比較上無用な項を無視することによって，当面の分割表だけに基づきそこに明示された説明変数の目的変数に対する効果を評価しうる統計量 (1.8) もしくは (2.3) を得る．

クロス表の分析に当って特定の目的変数（反応変数）に対する他の説明変数の効果に注目する場合と，所与の変数すべてによって作られる様々な相互関係に注目する場合とを明確に区別すべきであるという論点は文献 [13] において初めて指摘された．本稿の方法はこの観点の延長線上に位置する．なお，対応するモデルの解説は [14] においても見られ，本稿のモデルと対数線型モデルとの対比については [15] で触れられている．表8は [14] の記法を模して3変数の場合について本稿のモデルと対数線型モデルとの対応関係をしめしたものである．この表では，

$$F_{i_1 i_2 i_3} = n \cdot p(i_1, i_2, i_3)$$

であるから，表8の1行目はモデル

(5.1) $$\log_e F_{i_1 i_2 i_3} = u + u_{i_1} + u_{i_2} + u_{i_3} + u_{i_1 i_2} + u_{i_1 i_3} + u_{i_2 i_3} + u_{i_1 i_2 i_3}$$

表 8

No.	本稿のモデル $p(i_1, i_2, i_3) =$	対数線型モデル $\log_e F_{i_1, i_2, i_3} =$	
1	$p(i_1, i_2, i_3)$	$u_{i_1 i_2} + u_{i_1 i_3} + u_{i_1 i_2 i_3}$	$+u + u_{i_1} + u_{i_2} + u_{i_3} + u_{i_2 i_3}$
2	$p(i_1, i_2)\, p(i_2, i_3) / p(i_2)$	$u_{i_1 i_2}$	$+u + u_{i_1} + u_{i_2} + u_{i_3} + u_{i_2 i_3}$
3	$p(i_1, i_3)\, p(i_2, i_3) / p(i_3)$	$u_{i_1 i_3}$	$+u + u_{i_1} + u_{i_2} + u_{i_3} + u_{i_2 i_3}$
4	$p(i_1)\, p(i_2, i_3)$		$u + u_{i_1} + u_{i_2} + u_{i_3} + u_{i_2 i_3}$

を意味する．各モデルに対する AIC の値を全てのモデルにわたって比較したとすれば，表中の点線枠でしめされた全てのモデルに共通する項は各モデルの AIC の差に寄与しないから，本稿のモデルの比較は点線枠外の項の効果を比較していることがわかる．これは本稿の統計量が当該のクロス表に明示された説明変数の目的変数に対する効果のみを比較していることをしめしている．また，この表から，本稿のモデルは (5.1) の飽和モデルと，$u_{i_1i_2i_3}=0$ だけを想定したモデル

(5.2)　　$$\log F_{i_1i_2i_3}=u+u_{i_1}+u_{i_2}+u_{i_3}+u_{i_1i_2}+u_{i_1i_3}+u_{i_2i_3}$$

とを識別できないことがわかる．すなわち，本稿のモデルは各交互作用項の効果の評価に関しては相対的に粗いが，対数線型モデルによるならば変数の増加とともに検討すべきモデル数が急激に増加し，またしても分析困難に陥る．一方，変数の選択という観点に立てば，(5.1)，(5.2) のいずれのモデルが選択されても i_2, i_3 の 2 変数とも説明変数として採択するという結論に変りはなく，対数線型モデルはムダが多いことになる．本稿のモデルは変数の選択という目的に対して必要最小限のモデルということである．分析に当っていずれのモデルを適用するか，あるいは両者をどう組合せて使うかは分析の目的，データの精度・性質，計算時間の制約等によって決定されるべき問題である．

ところで，統計量 (1.7) から (1.8) への変形，あるいは (2.2) から (2.3) への変形が可能であるのは AIC がエントロピーの漸近的な不偏推定量であり，エントロピーはモデルの悪さを，想定したモデルすべてにわたって比較するために導入された統一的な尺度 [8, 9] であるという事情に基づく．これは AIC による方法が適合度のカイ自乗検定と際立って異なる点である．適合度検定は最高次の分割表にしめされた観測値に対する個々のモデルの適合度を単独に評価するものであって，自由度の異なるモデルどおしの比較を可能にするものではないからである．適合度検定によっては表 1 と表 2 の比較という形式の問題は解決できないのである．実際のデータ処理に当ってサンプル・サイズや変数の数を一切考慮する必要がなくなったのはこの形式でクロス表の比較が可能になったことに根ざすのであるから，その実用上の意義は極めて大きい．

こうして，データさえ与えられれば，仮に他の情報が何も与えられなくても，エントロピー最大化原理 [9] に導かれて最適なカテゴリゼーションをもつ説明変数の最適な組合せを検出することができる．とはいえ実際にデータを処理する場合には，データの性質や分析目的によって様々な形の事前情報や制約条件が与えられていることが多い．事前情報の多寡によって状況を分類すると，本節の冒頭の状況のように分布の型が指定されていてパラメータの値の推定のみが問題になる場合が一方の極端をなし，上のように事前情報が全くなく変数の最適な組合せとカテゴリゼーションを検索する場合が他方の極端をなす．しかし，一般には，カテゴリーのプールの方式や説明変数の次数等に関して種々の事前情報や制約が与えられていることが多い．たとえば，「ある目的変数に対してどの項目がより多くの情報を持っているか」という問題意識には既に説明変数の次数は 1 次であるという指定が暗黙のうちに含まれている．従来の記述統計的な尺度によって関連性の程度を把握する場合には実は暗黙のうちにカテゴリゼーションや次数について条件がつけられていたのである．本稿の立場に立つと，これらは種々の制約に応じたモデルの部分集合の中でモデルの選択を通じて確率分布の推定をめざす問題であるとみなすことができる．分布の推定，予測こそが本来の目標であって，「関連性の強さ」という日常的な概念はその一助としてたまたま問題とされる概念にすぎないのである．既に 1 節の最後部で例示したように，変数間の関連の強さは，次数とカテゴリゼーションに関する制約をもったモデルを設定し，それらの中で採択されたモデルの意味を解釈しなおすことによって結果的に比較することができるのである．

なお，連続変量をカテゴリー化することには不自然さがないわけではない．しかし，ヒスト

グラムは連続変量をカテゴリー化することによって母集団分布の型を推測するための方法である[3]．本稿の方法の不自然さはヒストグラムの不自然さと同程度であるといえよう．

さらに，既に明らかなように，本稿における目的変数とは着目した変数という程の意味であって，この意味では着目変数もしくは被説明変数とでも呼ぶ方がよりふさわしい．したがって目的変数であったものを次の分析では説明変数とみなしても構わないし，いったん着目したら所与の変数のうちのどの変数でも，その実質的な意味とは無関係に，目的変数とみなすこともできる。蛇足ではあるが，いくつかの地点で得られた連続変量だけから成るデータに基づいて地点間比較をするような場合にも，地点はカテゴリカルであるから，その変数を目的変数とみなしてCATDAPを適用することができる．

6. CATDAPの構成と利用法

6.1 概　　要

CATDAPは2つのメイン・プログラムから成っている．すなわち，統計的な手法を目的変数と説明変数のタイプによって下のように分類すると，この2つのプログラムの適用領域はつぎのようになる．

	説明変数		
	カテゴリカル型	連続型	混合型
目的変数：カテゴリカル型	CATDAP-01, 02	CATDAP-02	CATDAP-02

CATDAP-01は目的・説明両変数ともカテゴリカルであって，かつ，説明変数のカテゴリーのプールは一切行なわず単に説明変数の選択だけが問題となる場合のために作られたものである．CATDAPは，さらに，PART-1とPART-2とに細分される．PART-1は，実用上2次元のクロス表が分析に用いられることが多いことを配慮して，可能な2次元のクロス表を全て作成し指定された目的変数に対して情報の多い順（AICの値の小さい順）に説明変数を順序づける．一方，PART-2は，重回帰分析等における場合と同様の変数増減法を用いて，与えられた説明変数の候補の中から最適な組合せを探す．

CATDAP-02は説明変数の選択だけでなく，カテゴリーの自動的なプールによって最適なカテゴリゼーションをも併せて求める場合に適用されるべきプログラムである．したがって，説明変数の候補が連続型変数，もしくは混合型であればもちろんのこと，カテゴリカルであっても自動的なカテゴリーのプールが必要な場合にはCATDAP-02を使わなければならない．したがって，概念的にはCATDAP-01はCATDAP-02に含まれるが，CATDAP-01では1回のランで目的変数として何項目でも指定しうるのに対し，CATDAP-02では1回につき1項目しか指定しえず，この点でプログラムとしての機能が異なる．なお，人為的なコードの変換――単なるリコードを意味することもあり，カテゴリーのプールを意味することもある――はいずれのプログラムでも可能である．

出入力の詳細についてはプログラムごとに節を分け，CATDAP-01については6.2節で，CATDAP-02については6.3節で述べる．

なお，CATDAPは，ヒストグラムの自動描画のためのプログラム(CATDAP-11)も含めて，磁気テープの貸し出しが可能である．

6.2 CATDAP-01

6.2.1 出力について

CATDAP-01のPART-1のアウトプットはつぎの4種であるが，前半の2種は目的変数ご

とに出力され，後半の2種は全体で1枚だけ出力される．なお，まとまった出力例については[5, 21-34 頁]を参照されたい．

1) 指定された目的変数と各説明変数とによってつくられる2次元クロス表すべての間で比較をし，目的変数に対する情報の多いの順に配列しなおした説明変数の一覧表（順位，説明変数名，説明変数のカテゴリー数，AIC の値，1つだけ上位との AIC の値の差）．この AIC の値の差は情報の量のちがいを比較するのに有効である．

2) 2次元クロス表

上の1) の順位に従って実数，パーセンテージが打ち出され，何位まで打ち出すかは指定できる．

3) AIC の値の総括表

指定された目的変数の各々と説明変数の各々とによって作られる2次元クロス表に対する AIC の値を総括したもので，特定の変数間の従属度が考えうる（2次元の）組合せの中でどの程度であるかを知るために用いる．これらの数値の小大を符号の濃淡で表現し，2変数間の従属度（関連の強さ）を視覚的に捉えやすくしたものがつぎの4) である．

4) AIC の値の大きさを表わす濃淡図（図1参照）

この図で，符号の濃度が大きいほど当該2変数間の従属度が強いこと（AIC の値が小さいこと）を表わす．ただし，AIC の値の区切り方（各符号の境界値の与え方）は恣意的である．なお白地（スペース）は当該2変数が AIC の立場から見て独立であることをしめす．

PART-2 のアウトプットはつぎの3種であり，いずれも目的変数ごとに出力される．

5) 説明変数の次数ごとの AIC の値

変数増減法によってチェックした説明変数の組合せを次数ごとに整理し，組合せた説明変数の名称，説明変数の組の総カテゴリー数，AIC の値，上位との差が，各次数の中で AIC の小さい順にプリントされる．各次数とも上位100位までで印刷は打ち切られる．

6) AIC の値の総括表

上記の5) で得られた AIC の値を全ての組合せにわたって小さい順に整理し，打ち出すもので，形式は5) と同じ．印刷は上位100位までで打ち切られる．

7) 説明変数の最適な組合せをもつ多重クロス表

実数とパーセンテージの2様でしめされる．なお，どんな説明変数の組合せをとっても AIC の値が負にならない場合（独立ではない変数の組がない場合）には目的変数に関する周辺度数分布（いわゆる単純集計）だけが印刷される．

なお，出力の最後に作業領域がバイトでプリントされる．この作業領域は引数 'IALIM' と 'ALIM' によって指定できる．

6.2.2 入力について

このプログラムには最大限9種類のカード・イメージ・データのインプットが必要である．なお，インプットの書式例については[5, 19〜20頁]を参照されたい．これは100人のデータのうち 'AGE' がコード '4' の人を除外して集計・分析した例である．

1) データ1

NSAMP（カラム 1〜4）：サンプル・サイズ

N（カラム 5〜8）：変数（あるいは項目）の総数

L（カラム 9〜12）：目的変数としてとりあげる変数の数．L個の変数が順番に1個ずつ目的変数とみなして分析される．

RECODE（カラム 13〜16）：人為的なリコードをする変数の総数．

ICROSS（カラム 17〜20）：アウトプット 2）で印刷する 2 次元クロス表の数．特に指定がない場合には 10 番までしか印刷されない．

IEXP（カラム 21〜24）：所与の変数の 1 部だけを説明変数の候補とみなす場合には '1' とおき，他の場合には '0' とおくこと．たとえばフェース・シートだけとの関連度を問題にする場合が前者である．

N1（カラム 25〜28）：'IEXP' が 1 の場合の説明変数の数．'IEXP' が 0 ならこれも 0 とおくこと．

JSAMP（カラム 29〜32）：分析に必要なメイン・データ（処理対象データ）までの読み飛ばすデータ数．

IN（カラム 33〜36）：メインデータの入力装置番号．特に指定されない場合は '5' とみなされる．

IPART（カラム 37〜40）：PART-1 だけの分析が必要なら 1，PART-2 だけなら 2，両方 (CATDAP-01 全体）必要なら 0 とおく．

2 次元のクロス表だけで分析におおよその見当をつけたい時などこの命令を活用することができる．

ISKIP1（カラム 41〜44）：ある属性をもつサンプルを除外して分析する場合の除外する属性の種類（＝変数）の数．所与のデータを全て使用する場合は 0 とおけばよい．

これは男のデータを除外して分析するとか，20 才台以外のデータをのぞいて分析する場合に使用する．また，世論調査のデータにおいてはある質問で 'D. K.'（わからない）と答えた人は別の質問でも 'D. K.' と答えがちであるため，みかけ上 2 問間に強い関連があるとみるとみなされることがある．これを防ぐためには目的変数に関して 'D. K.' という回答をしたサンプルを除いておけばよい．

2) データ 2

ITEM1(I)，ITEM2(I)，I=1, N：所与の各変数の（リコード後の）コードの最小値と最大値．1 変数につき 4 カラムを 2 つずつ使って N 個の変数分続けて指定する．

3) データ 3

ICONV(I, J)，I=1, RECODE, J=1, 20：人為的にリコードする場合の変数番号と新しいコードの指定．リコードを要する変数ごとにカードを 1 枚ずつ用いて，カラム 1〜4 に変数番号，カラム 8 から 80 まで 4 カラムごとに事前のコード 1, 2, ···, 19 に対応するリコード後のコードをパンチすること．リコードが不要な場合には，入力（カード）も不要．

4) データ 4

FACE(I)，I=1, L：目的変数の変数番号を 4 カラムごとにパンチすること．たとえば，N 個の変数でつくりうる ${}_NC_2$ 通りの 2 変数間の関連を比較する場合には '1, 2, ···, N' とパンチする．

5) データ 5

FACE2(I)，I=1, N1：変数の一部のみを説明変数の候補とみなす場合（'IEXP=1' の場合）の変数番号．全てを説明変数の候補とみなす場合には入力も不要．

6) データ 6

FMT(I)，I=1, 20：メイン・データ（処理対象データ）のリード・フォーマット．

7) データ 7

TITLE(I, K)，I=1, 10, K=1, N：全変数の変数名．10 文字で指定．

8) データ 8

ISKIP(I, J)，I=1, ISKIP1，J=1, 20：ある属性をもつサンプルを除外して分析する場合に必要なカードで，当該の変数ごとにカードを1枚ずつ使い，カラム4に変数番号，カラム8にコードの個数，カラム12以降にその個数だけのリコード前のコードをパンチする．指定された変数に関してあるサンプルの回答がここで指定されたコードのどれかに一致すると，そのサンプルは集計・分析から除外される．

9) データ9

メイン・データ（処理対象データ）

6.3 CATDAP-02

6.3.1 出力について

CATDAP-02 のアウト・プットはつぎの5種である．なお，出力例については文献 [5, 71-73 頁] を参照されたい．

1) 従属度の強さの順に配列された説明変数の一覧表（形式は6.2.1節の1)と同じ，表4参照）CATDAP-02 では説明変数を最適にカテゴライズしたときの AIC の値，クロス表等のみをプリントアウトし，他の種々のカテゴリゼーションに対応する計算結果は印刷されない．

2) 最適にカテゴライズされたときの2次元クロス表（表5参照）

1) に対応する表で実数，パーセンテージの他に，各説明変数の各カテゴリーの境界値も右端にしめされる．

3) 説明変数の次数ごとの AIC の値

出力形式は6.2.1節の5) と同じである．上位100位まで印刷される．

4) AIC の値の総括表（表6参照）

6.2.1節の6) と同一の出力形式．

5) 最適な説明変数をもつクロス表（表7参照）

上記4) の第1位に相当する多重クロス表で，最適な変数の組合せ・最適なカテゴリーゼーションをもつクロス表が実数とパーセンテージ両様でしめされると同時に各変数の各カテゴリーの境界値もしめされる．

6.3.2 入力について

このプログラムには最大限10種類のカード・イメージ・データのインプットが必要である．インプットの多くが CATDAP-01 と同じ形式である．なお，インプットの書式例については文献 [5, 70 頁] を参照されたい．

1) データ1

NSAMP（カラム 1〜4)：サンプル・サイズ

N（カラム 5〜8)：変数の総数

RECODE（カラム 9〜12)：人為的なリコードをする変数の総数

IJP（カラム 13〜16)：分析に必要なメイン・データ（処理対象データ）までの読み飛ばすデータ数

IN（カラム 17〜20)：メイン・データの入力装置番号

ISKIP1（カラム (21〜24)：ある属性をもつサンプルを除外して分析する場合の除外する属性の種類（＝変数）の数．

IT（カラム 25〜28)：全変数が整数タイプとして読み込める場合は0とおき，それ以外は1とおく．

IPART（カラム 29〜32)：2次元のクロス表による分析（各説明変数の最適カテゴリーの探索と関連度の比較）だけが必要なら1，それ以外の分析も必要なら0とおく．この指

定は，2次元のクロス表の検討によって説明変数の候補を絞り，以後の多重クロス表による解析の処理時間短縮を図る場合にも利用できる．

NOV（カラム 33〜36）：6.3.1節の1)の変数のうちの上位何位かまでだけを最適な多重クロス表を求める際の説明変数の候補とみなす場合の候補の数に1を加えた数．たとえば上位10位までの変数を候補とみなす場合には '11' とかくこと．従って，'IPART' が1のときはこの指定は不要で，該当するカラムはブランクにしておく．

2) データ2

ITEM1(I)，ITEM2(I)，I=1，N：各変数の（リコード後の）コードの最小値と最大値．1変数につき4カラムを2つずつ使ってN個の変数分続けてパンチする．連続型変数があれば該当箇所は (0, 0) とおくこと．

3) データ3

XX(I)，I=1, N：連続型変数の観測精度．例えば，ある変数が 18.3，19.2，… と与えられていれば該当箇所は0.1と10カラムごとに書く．ただしカテゴリカルな変数の該当箇所は0.0と書く．従って，データ2で '0, 0' となっている変数はデータ3では0.0以外の数値がパンチされ，データ3で0.0となっている変数はデータ2では '0, 0' 以外の数値がパンチされているはずである．'IT=1' で，ある観測値が整数値の場合には，データ2を '0, 0' データ3を '1.0' としてもよいし，データ3を '0.0' としてデータ2で最小値，最大値を指定してもよい．なお，全変数が整数（'IT=0'）ならこのカードは不要．

4) データ4

ITY(I)，I=1, N：各変数のカテゴリーのプールの形式の指定．

等間隔のプールが必要なら0，相隣るカテゴリーの（不等間隔の）プールが必要なら1，一切プールしないなら2と指定する．なお，このプログラムでは，コード1とコード3をプールするというようなコードを飛び越したプールは行なえない．

5) データ5

ICONV(I, J)，I=1，RECODE，J=1, 20：6.2.2節の3)の入力形式と同じ．

6) データ6

FACE：目的変数の変数番号

7) データ7

FMT(I)，I=1, 20：メイン・データ（処理対象データ）のリード・フォーマット．1つでもFタイプで読み込まれる変数があれば他もすべてFタイプで指定すること．たとえば 10I1, F2.1）は許されず，(10F1.0, F2.1）と書くこと．

8) データ8

TITLE(I, J)，I=1, 10，J=1，N：全変数の変数名．10文字で指定．

9) データ9

特定の属性（実数型の変数を含む）をもつサンプルを除外して分析する場合に必要なカードで，次の3つのケースがある．

i) 全てのデータを使用する場合はカード不要．

ii) 全変数が整数タイプで読み入まれ，かつ，除外する属性がある場合（'IT=0' かつ 'ISKIP1'>0 の場合）には CATDAP-01 と同一形式（6.2.2節の8)）で指定．

iii) 実数タイプの変数が含まれ，かつ，除外する属性がある場合（'IT=1' かつ 'ISKIP1'>0 の場合）には，当該の変数ごとに2枚のカードを使って指定し，1枚目のカードにはその変数の変数番号と除外される区間の数を書き，2枚目のカードには指定する区間の最小値と最大値を10カラムごとに書く．

10)　データ10
　　メイン・データ（処理対象データ）

むすびにかえて

プログラム CATDAP は目的変数がカテゴリカルでありさえすれば，サンプル・サイズ，変数の数，種類の如何にかかわらず，どんなデータにも適用することができる．

この方法はどんなデータも分割表に帰着させ，その分割表を目的変数と説明変数の2組に分けて統計量 (2.3) を適用するだけの簡単な方法である．それにもかかわらず，CATDAP はいかなる形の効果をもつ説明変数も検出することができ，分割表の形で与えられる分析結果の意味は明快で，解釈が恣意的になる危険性も少ない．

CATDAP はデータさえ与えられれば大量のデータに含まれている重要な情報を瞬時にして検出することができる．CATDAP は実質科学的に意味のない変数を拾いあげることはあっても重要なものを見逃がすことはない．したがって，社会学，心理学，医学，疫学等カテゴリカルデータをとりあつかうことの多い様々の分野で各種の現象の因果関係の究明に大いに寄与しうるものと期待される．

謝　　辞

この研究を進めるに当たっては統計数理研究所の林知己夫，西平重喜，赤池弘次，駒沢　勉田辺国士，石黒真木夫，尾形良彦，仁木直人，北川源四郎，桂　康一，大阪大学の稲垣宣生の諸氏に多くの激励，教示を受けました。殊に，赤池，石黒，北川の三氏には日常的な討論の中で種々の教示・協力を得ました。記して心から感謝します。

参　考　文　献

[1] Sakamoto, Y. and Akaike, H. (1978). Analysis of cross-classified data by AIC, *Ann. Inst. Statist. Math.*, **30**, B, 185–197.
[2] Sakamoto, Y. and Akaike, H. (1978). Robot data screening of cross-classified data by an information criterion, *Proc. International Confernce on Cybernetics and Society*, 398–403.
[3] Sakamoto Y. (1977). A model for the optimal pooling of categories of the predictor in a contingency table, *Research Memorandum* No. 119, Institute of Statistical Mathematics, Tokyo.
[4] 坂元慶行 (1981). カテゴリカルデータの解析,「数理科学」, サイエンス社, 1981年3月号掲載予定.
[5] Katsura, K. and Sakamoto, Y. (1980). CATDAP, *Computer Science Monographs* No. 14, Institute of Statistical Mathematics, Tokyo.
[6] 直井道子 (1979). 階層意識と階級意識,「日本の階層構造」(富永健一編), 365-388, 東大出版会.
[7] 赤池弘次 (1976). 情報量基準 AIC とは何か――その意味と将来への展望,「数理科学」, 153号, 5-11.
[8] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory* (eds. B.N. Petrov and F. Csaki), Akademiai Kiado, Budapest, 267–281.
[9] Akaike, H. (1976). On entropy maximization principle, *Application of Statistics* (ed. P.R. Krishnaiah), North-Holland, Amsterdam, 27–41.
[10] Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections, *Commun. Statist.* A7(1), 13–26.
[11] 統計数理研究所国民性調査委員会 (1975).「第3日本人の国民性」, 至誠堂.
[12] 駒沢　勉 (1978).「多元的データ分析の基礎」, 朝倉書店.
[13] Bhapker, V.P. and Koch, Gary G. (1968). Hypothesis of 'no interaction' in multi-dimensional contingency tables, *Technometrics*, **10**, 107–123.
[14] Everitt, B.S. (1980).「質的データの解析」(山内光哉監訳), 新曜社.
[15] 大隅　昇 (1978). 多重クロス表による社会調査データのモデル解析,「京大数理解析研究所講究録」, 345.