CATDAP マニュアル

石黒真木夫@統計思考院@統計数理研究所
2016.7.16

ABSTRACT

CATDAP, CATegorical Data Analysis Program, is an AIC-based program published by Katsura,K. and Sakamoto,Y.(1980). It is based on the contingency table analysis method proposed by Prof. Sakamoto(1983) of the Institute of Statistical Mathematics.

This article is a usage manual of an fortified version of CATDAP. This version handles not only the categorical object variable case, but the continuous object variable case either. It is not a CAT anymore, in this sense. We would like to call it TIGERDAP, The Integrated GEneRal Data Analysis Program.

Contents

1 CATDAP とは

CATDAP は坂元の情報量規準 AIC によるクロス表解析法 (坂元, 1983) にもとづく離散データ解析プログラム (Katsura,K. and Sakamoto,Y., 1980) である.

このプログラムはクロス表を解析する道具という以上に, 目的変数が離散変数であるデータをクロス表の形に整理して解析する道具と説明するのが適切であり, 説明変数候補 (群) が連続値を含む場合も扱える. 極めて使いやすい汎用ソフトである.

データの個性に密着した解析にはさらに高度なモデルの開発が必要になる場合は多々あるが, このソフトだけで十分な結果が得られる場合も多々ある. 少なくともデータの挙動が正確に把握できていない段階での予備的な解析の道具として常に手もとに置いておくべき道具であろう.

1.1 クロス表

2つの離散確率変, I ($\in 1,2,...,C_I$) と J ($\in 1,2,...,C_J$) を N 回測定したデータ

$$\text{データ} = \{(I_n, J_n)\,|n = 1,2,\ldots,N\}$$

から, I と J の間の関係を調べたいものとしよう.

$$\delta\,(x,y) = \begin{cases} 1 & (x = y) \\ 0 & (x \neq y) \end{cases} \tag{1}$$

として

$$N_{ji} = \sum_{n=1}^{N} \delta\,(I_n, i) \times \delta\,(J_n, j)$$

を求め, N_{ji} を要素とする行列を表の形に書いたものがクロス表である.

本稿では目的変数と説明変数の関係をみるために作成するクロス表は, 説明変数を行に対応させ, 目的変数を列に対応させるという規約を置く. この規約のもとで N_{ji} は, 説明変数が値 j をとった時に目的変数が値 i をとった回数を示すこととなる.

たとえば

喫煙	脳血管系疾患	虚血性心疾患
吸わない, やめた	7	4
1 日 19 本以下	17	18
1 日 20 本以上	2	10
計	26	32

は喫煙習慣と病気の関係を調べたデータから作られたクロス表である.

(1) のかわりに

$$\delta\,(x,y) = \begin{cases} 1 & (x \text{ が区間 } y \text{ に含まれる}) \\ 0 & (\text{それ以外}) \end{cases}$$

とすれば連続値データ x も離散化してクロス表の形に整理することができる. 上にあげた例では喫煙本数という量が"1 日 20 本以上"のような区間に含まれるか否かといった規準で離散化している.

1.2 クロス表の評価

離散確率変数 I の出現確率が J の値に依存して

$$P_{I|J}\,(I|J) \tag{2}$$

の形で与えられる場合 (相関モデル) と, J に依存しない

$$P_I(I) \tag{3}$$

の形で与えられる場合 (独立モデル) が考えられる.

(2) の場合, (3) の場合それぞれで, データ (I, J) が観測される確率は

$$P_{I|J}(I|J) \times P_J(J)$$

$$P_I(I) \times P_J(J)$$

となる. これから「データ」全体が得られる確率はそれぞれ

$$\prod_{n=1}^{N} P_{I|J}(I_n|J_n) \times P_J(J_n)$$

$$\prod_{n=1}^{N} P_I(I_n) \times P_J(J_n)$$

となり, 対数尤度は

$$\sum_{n=1}^{N} \left\{ \log P_{I|J}(I_n|J_n) + \log P_J(J_n) \right\}$$

$$\sum_{n=1}^{N} \{ \log P_I(I_n) + \log P_J(J_n) \}$$

となる.

$$\text{クロス表} = \{N_{ji} | i = 1, 2, \ldots C_I, j = 1, 2, \ldots, C_J\} \tag{4}$$

の形に集計されたデータがあれば, 対数尤度は

$$\sum_{i=1}^{C_I} \sum_{j=1}^{C_J} N_{ji} \left\{ \log P_{I|J}(i|j) + \log P_J(j) \right\}$$

$$\sum_{i=1}^{C_I} \sum_{j=1}^{C_J} N_{ji} \{ \log P_I(i) + \log P_J(j) \}$$

で計算される.

$$N_{\cdot i} = \sum_{j=1}^{C_J} N_{ji} \tag{5}$$

$$N_{j\cdot} = \sum_{i=1}^{C_I} N_{ji} \tag{6}$$

とすると最尤推定量は

$$
\begin{aligned}
P_I(i) &= N_{\cdot i}/N \\
P_J(j) &= N_{j\cdot}/N \\
P_{I|J}(i|j) &= N_{ji}/N_{j\cdot}
\end{aligned}
$$

で与えられる (証明略). 最大対数尤度は

$$
\sum_{i=1}^{C_I}\sum_{j=1}^{C_J} N_{ji}\left\{\log\frac{N_{ji}}{N_{j\cdot}}+\log\frac{N_{j\cdot}}{N}\right\}
$$

$$
\sum_{i=1}^{C_I}\sum_{j=1}^{C_J} N_{ji}\left\{\log\frac{N_{\cdot i}}{N}+\log\frac{N_{j\cdot}}{N}\right\}
$$

で, AIC は

$$
\begin{aligned}
AIC_1 &= -2\times\left\{\sum_{i=1}^{C_I}\sum_{j=1}^{C_J} N_{ji}\left\{\log\frac{N_{ji}}{N_{j\cdot}}+\log\frac{N_{j\cdot}}{N}\right\}\right\} \\
&\quad +2\times\left\{\left(C_I-1\right)C_J+\left(C_J-1\right)\right\}
\end{aligned}
$$

$$
\begin{aligned}
AIC_0 &= -2\times\left\{\sum_{i=1}^{C_I}\sum_{j=1}^{C_J} N_{ji}\left\{\log\frac{N_{\cdot i}}{N}+\log\frac{N_{j\cdot}}{N}\right\}\right\} \\
&\quad +2\times\left\{\left(C_I-1\right)+\left(C_J-1\right)\right\}
\end{aligned}
$$

モデル比較のためには, 共通部分は不要なのでこれらを消去した形

$$
\begin{aligned}
AIC_{I|J} &= -2\times\left\{\sum_{i=1}^{C_I}\sum_{j=1}^{C_J} N_{ji}\left\{\log\frac{N_{ji}}{N_{j\cdot}}\right\}\right\} \\
&\quad +2\times\left\{\left(C_I-1\right)C_J\right\} \\
&= -2\times\left\{\sum_{i=1}^{C_I}\sum_{j=1}^{C_J}\left(N_{ji}\log N_{ji}-1\right)-\sum_{j=1}^{C_J}\left(N_{j\cdot}\log N_{j\cdot}-1\right)\right\} \\
&= -2\times\left\{\sum_{i=1}^{C_I}\sum_{j=1}^{C_J}\psi_{CATDAP}\left(N_{ji}\right)-\sum_{j=1}^{C_J}\psi_{CATDAP}\left(N_{j\cdot}\right)\right\}
\end{aligned}
\tag{7}
$$

$$
\begin{aligned}
AIC_I &= -2\times\left\{\sum_{i=1}^{C_I}\sum_{j=1}^{C_J} N_{ji}\left\{\log\frac{N_{\cdot i}}{N}\right\}\right\} \\
&\quad +2\times\left\{\left(C_I-1\right)\right\} \\
&= -2\times\left\{\sum_{i=1}^{C_I}\left(N_{\cdot i}\log N_{\cdot i}-1\right)-\left(N\log N-1\right)\right\} \\
&= -2\times\left\{\sum_{i=1}^{C_I}\psi_{CATDAP}\left(N_{\cdot i}\right)-\psi_{CATDAP}\left(N\right)\right\}
\end{aligned}
\tag{8}
$$

と簡明な表現が得られる．ここで

$$\psi_{CATDAP}(x) = x \log x - 1 \tag{9}$$

である．

モデル (2) とモデル (3) を比較して $AIC_{I|J} < AIC_I$ であれば「説明変数」J が「目的変数」I の予測に役立つ情報を持っていることが分る．

1.3 CATDAP の AIC

このことを踏まえて坂元の CATDAP プログラムのオリジナル版では, クロス表を評価する値として

$$\text{クロス表評価}\{N_{ji}\} = AIC_{I|J} - AIC_I$$

を出力するようになっていたがクロス表に空白セルが含まれる場合にこの値による評価が直観に合わない場合があることが分ってきた．

たとえば, 10 円玉 ($j = 1$) と百円玉 ($j = 2$) を投げたときに表が出る ($i = 1$) か裏が出る ($i = 2$) かを調べる実験で１０円玉と１００円玉を一回だけ投げたとき, １０円玉が裏, 百円玉が表だったとする．この実験結果を

i	1	2
N_{1i}	0	1
N_{2i}	1	0
$N_{\cdot i}$	1	1

というクロス表にまとめることができる．「例 1 クロス表」と名付けよう．例 1 クロス表から, 「結果」の数値と使った硬貨の違いに関連があるかどうかを見るのは, データ不足で, 無理であると考えるのが常識的な判断だろう．

ところがこの例 1 クロス表の評価を計算してみると

$$\begin{aligned} AIC_I &= 4.77 \\ AIC_{I|J} &= 4.0 \\ AIC_{I|J} - AIC_I = \text{クロス表評価}\{\text{ 例 1 クロス表 }\} &= -0.77 \end{aligned}$$

と, 「相関モデル」の方が良しとされてしまう．

このような結果が得られるのは, (9) によると空白セルにおいて $\psi_{CATDAP}(0) = -1$ という計算がされるために $AIC_{I|J}$ が過少評価されるためである．

この過少評価は含まれていゼロをあらかじめ $1/e$ で置き換えておいたクロス表を評価することによって避けられる．このような「前処理」の結果得られるクロス表を式の形で

$$\text{空白処理}\{\text{ クロス表 }\}$$

と書くことにしよう．空白処理 { 例 1 クロス表 } は

i	1	2
N_{i1}	1/e	1
N_{i2}	1	1/e
$N_{i\cdot}$	1+1/e	1+1/e

という形になる．$1/e$ というのは $\psi_{CATDAP}(x)$ の最小値を与える x の値である．

{ 空白処理 { 例 1 クロス表 }} は

$$\begin{aligned} AIC_{\text{独立}} &= 5.79 \\ AIC_{\text{相関}} &= 7.18 \\ AIC_{\text{相関}} - AIC_{\text{独立}} &= 1.39 \end{aligned}$$

と, 直観に合う評価となる．

ちょっと「一般化」して

i	1	2
N_{i1}	0	n
N_{i2}	n	0
$N_{i.}$	n	n

という例 n クロス表 で空白処理の効果を調べてみると

クロス表	クロス表評価 { クロス表 }	クロス表評価 { 空白処理 { クロス表 }}
例 1 クロス表	-0.77	1.39
例 1.5 クロス表	-1.82	0.73
例 2 クロス表	-3.55	-0.47
例 3 クロス表	-6.32	-2.69
例 4 クロス表	-9.09	-5.06
例 100 クロス表	-275.26	-266.56

となる. 例 1.5 クロス表 は

i	1	2
N_{i1}	0	1
N_{i2}	2	0
$N_{i.}$	2	1

である.

最新版の CATDAP で AIC として出力される値は「クロス表評価 { 空白処理 { クロス表 }}」である. 空白処理においては一行すべて空白の「空白行」にあっては各セルのゼロを $1/e$ で置き換えるのでなく行全体を削除してしまう.

空白処理後の

コイン	表	裏
10 円玉	0	1
50 円玉	0	0
100 円玉	1	0
$N_{.i}$	1	1

と

コイン	表	裏
10 円玉	0	1
100 円玉	1	0
$N_{.i}$	1	1

は同じ

コイン	1	2
10 円玉	1/e	1
100 円玉	1	1/e
$N_{.i}$	1+1/e	1+1/e

という形になる.

base AIC　CATDAP が出力する AIC の値を見ることによって説明変数 (候補) が目的変数に関する情報を持っているかどうか一目瞭然に分るが, 他のモデルも比較対象に加える場合には この値が $AIC_{I|J}$ の値でないことに注意する必要がある.

CATDAP が当てはめるモデルと, たとえば, ロジスティックモデルを比較しようとすると, ロジスティックモデルの AIC から AIC_I の値を差し引かなくてはならない. 最新の R 版 CATDAP では AIC_I の値が「base AIC」として出力されるようになっている.

たとえば, 付録に置いた「サンプルデータ」で $MODEL_{V1|V2}$, $MODEL_{V1|V3}$, $MODEL_{V1|V4}$ と 4 パラメータのロジスティックモデル

$$P\left(V1=2|V2,V3,V4\right)=\frac{\exp\left\{a0+a1*V2+a2*V3+a3*V4\right\}}{1+\exp\left\{a0+a1*V2+a2*V3+a3*V4\right\}}$$

の比較が可能である. 上のロジスティックモデルの AIC においても, 説明変数 $\{V2, V3, V4\}$ の分布に関しては「不問に付する」のが普通である. CATDAP モデルの AIC (7) や (8) においても説明変数の扱いは同じであり, 比較することに何ら問題はない. データにおける説明変数の分布は, 結果の解釈や利用にあたっては重要なポイントであるが, それは別の問題である.

モデル	AIC
$CATDAP(Y\|X1)$	155.6
$CATDAP(Y\|X1, X2)$	183.84
$CATDAP(Y\|X1, X2, X3)$	385.21
$LOGISTIC(Y\|X1)$	152.56
$LOGISTIC(Y\|X1, X2)$	154.07
$LOGISTIC(Y\|X1, X2, X3)$	155.58

1.4 決定木分類の評価

目的変数 Y に対して, 説明変数 X1 が測定されており, X1 = 1 の場合にはさらに X2 が測定され, X1 = 2 の場合には X2 でなく X3 が測定されているような場合がある.

```
Y,X1,X2,X3
1, 1, 1,
1, 1, 1,
1, 1, 1,
1, 1, 1,
1, 1, 1,
1, 1, 2,
1, 1, 2,
1, 1, 2,
2, 1, 2,
1, 1, 2,
1, 1, 3,
2, 1, 3,
1, 1, 3,
1, 1, 3,
1, 1, 3,
```

```
Y,X1,X2,X3
1, 2,  , 1
1, 2,  , 2
1, 2,  , 3
1, 2,  , 4
2, 2,  , 5
2, 2,  , 1
2, 2,  , 2
2, 2,  , 3
1, 2,  , 4
2, 2,  , 5
1, 2,  , 1
1, 2,  , 2
1, 2,  , 3
2, 2,  , 4
1, 2,  , 5
```

このデータを分析するにあたっては, 次の決定木に従ってデータを 8 分類して 8 行× 2 列の分割表とみなして AIC を計算すればよい.

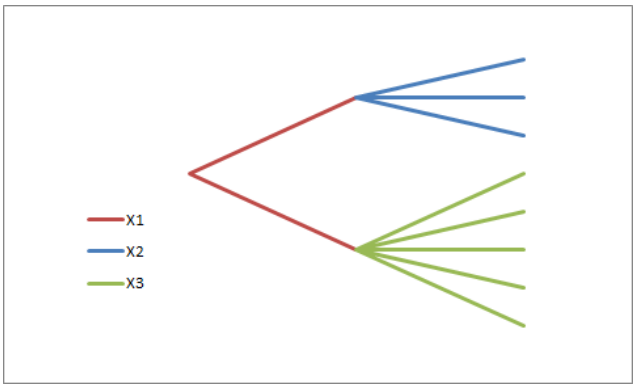

図 1.1 決定木分類

これは, データをつぎのような分割表の形で集計することになる.

			Y=1	Y=2
X1=1	X2=1		5	0
	X2=2		4	1
	X2=3		4	1
X1=2		X3=1	2	1
		X3=2	2	1
		X3=3	2	1
		X3=4	2	1
		X3=5	1	2

図 1.2 決定木分類に基づくデータの集計

すべてのデータについて全ての説明変数, たとえ X1, X2, X3, の値が得られている場合, CATDAP は, 図 1.3 に示される分割表を自動的に生成して AIC を計算する.

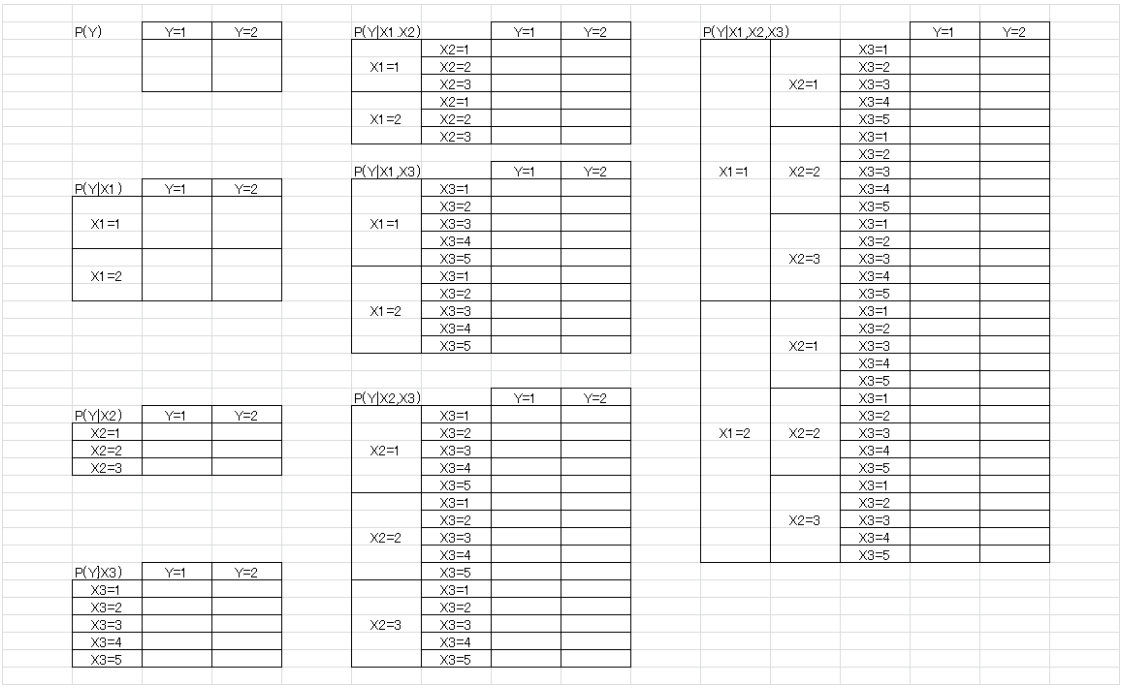

図 1.3 CATDAP が生成する決定木分類

1.5 連続値変数の扱い

CATDAP では説明変数に連続値が含まれている場合に離散化して他の離散値説明変数と同列に扱う. その離散化の方法が 2 通り用意されている.

1.5.1 top-down pooling

連続値変数を一定巾の区間に区切って離散化する方式である. 区間数は AIC 最小化で選ぶ. ただし両端の区間は「一定巾」より短くてよいとする.

1.5.2 bottom-up pooling

連続値変数をまず短い一定巾の区間に区切り, 隣り合う区間における目的変数の分布が似ている場合に合併することを繰り返して最終的な区間分けを得る方式である. 合併することの是非は AIC の値で判定される. クラスタリングに属する手法である.

1.5.3 modified top-down pooling

オリジナルの CATDAP に無かった方式である. topdown pooling と同じであるが, 両端の区間は「一定巾」より長くてよいとする.

1.6 histogram 解析

オリジナル CATDAP では目的変数として離散変数しか受け付けなかったが, 目的変数も離散化する機能を追加することによって, 変数の型によらずに解析できるソフトとなる.

連続値目的変数を離散化して扱うことは連続値変数の histogram を推定することに等しい. その histogram の説明変数への依存の有りかたを調べることは, 全データの histogram を, 説明変数の値で決まる histogram の重みつき混合分布として表現することになる.

(7) は目的変数が多項分布に従うとした場合の確率関数の尤度に基づいた式で

$$\begin{aligned} AIC_{I|J} &= -2 \times \left\{ \sum_{i=1}^{C_I} \sum_{j=1}^{C_J} N_{ji} \left\{ \log \frac{N_{ji}}{N_{j.}} \right\} \right\} \\ &\quad +2 \times \{(C_I - 1)\, C_J\} \end{aligned}$$

という形になっていたが, これを目的変数が histogram モデルに従うとした場合尤度に基づいた式とするには, 確率のを離散化の区間巾で割った確率密度にすればよい. i 番目の区間の巾を s_i とすると

$$\begin{aligned} AIC^H_{I|J} &= -2 \times \left\{ \sum_{i=1}^{C_I} \sum_{j=1}^{C_J} N_{ji} \left\{ \log \frac{N_{ji}}{N_{j.} s_i} \right\} \right\} \\ &\quad +2 \times \{(C_I - 1)\, C_J\} \\ &= -2 \times \left\{ \sum_{i=1}^{C_I} \sum_{j=1}^{C_J} N_{ji} \left\{ \log \frac{N_{ji}}{N_{j.}} \right\} \right\} \\ &\quad +2 \times \{(C_I - 1)\, C_J\} \\ &\quad +2 \times \left\{ \sum_{i=1}^{C_I} N_{.i} \left\{ \log s_i \right\} \right\} \end{aligned}$$

s_i が一定値 s に等しい場合には

$$AIC^H_{I|J} = AIC_{I|J} + 2 \times N \times \log s$$

となる. AIC^H_I についても同様である. この AIC は他の, たとえば重回帰モデルや分散分析モデルの AIC と比較可能である.

1.7 欠測値の扱い

データには欠測がつきものである. しかし, 離散データの場合には欠測も一種の観測値として扱うことが可能であり, 説明変数の欠測は本質的な問題を起こさない.

連続値目的変数における欠測も, ヒストグラムモデルの拡張で扱うことができる. 具体的には N 個のデータのうち, M 個が欠測である場合, $(N - M)$ 個のデータを, $(C_I - 1)$ 個の巾 s の区間に落し, M 個のデータを C_I 番目のカテゴリーに落して, $AIC_{I|J}$ を求め、

$$AIC^M_{I|J} = AIC_{I|J} + 2 \times (N - M) \times \log s$$

で欠測を含むモデルの AIC と定義すればよい.

$$AIC^M_{I|J} = AIC_{I|J} + 2 \times (N - M) \times \log s + 2 \times M \times \log 1$$

であるから, C_I 番目のカテゴリーだけ「巾」を s でなく 1 とする事と等価であることに注意しておく. また, データによっては数種の欠測を区別して扱いたい場合にも, 同様な方法で扱うことができることも注意しておく.

また, 区間巾を一定としない hisitogram を描くとき, 端の「区間」の「巾」をどうすべきか悩むが, このような「区間」を欠測として扱うことでこの問題から逃げることが可能かもしれない.

なお, 連続値データ X のモデルとしてヒストグラムでない確率密度関数 $f(x)$ を考えることが可能てある. データが欠測となる確率を p, $f(x)$ に従う確率を $(1-p)$ とするモデルを考えることが出来る. データ M 個が欠測で,$\{x_1, x_2, ..., x_N\}$ が観測されている場合のこのモデルの対数尤度は

$$\sum_{i=1}^{N} \log(1-p) f(x_i) + M \log p$$

で与えられる. このモデルのパラメータ数はモデル $f(x)$ のパラメータ数より一つ多い.

2 CATDAP を用いたデータ解析

2.1 疾患を分ける要因の発見

坂元 (2001) は循環器系疾患に関する集団検診の後で, 脳血管系疾患あるいは虚血性心疾患を発症した５８人について, 疾患を分ける要因を見付けるのに CATDAP を用いてみせた.

疾患と喫煙の関係をまとめたクロス表１と疾患と最小血圧の関係をまとめた分割表 2.1 と 2.2 では, クロス表 2,2 の AIC の値が低く, 最小血圧の方がより有効であることが (医学的知識のない筆者にも) 分るというのが坂元の結論である.

表 2.1

喫煙	脳血管系疾患	虚血性心疾患
吸わない, やめた	7	4
1 日 19 本以下	17	18
1 日 20 本以上	2	10
計	26	32

表 2.2

最小血圧	脳血管系疾患	虚血性心疾患
〜79mmHg	2	9
80〜89mmHg	4	12
90〜99mmHg	7	6
100〜 mmHg	13	5
計	26	32

n	CATDAP	CATDAPmod
表 2.1	-2.06	-2.06
表 2.2	-6.14	-6.14

このような空白セルのないクロス表の評価は CATDAP と CATDAPmod で同じである. なお, エストレーラの記事の後半で, 表２のもとになった「個票データ」を CATDAPmod で解析することによって, 表２におけるよりさらに有効な最小血圧の区分が得られることが示されている.

また, Kogiso,T. et al.(2009) では「非アルコール性脂肪肝 (?)」の診断と関係が深い因子の発見に CATDAP が使われている.

2.2 統計指標の構成

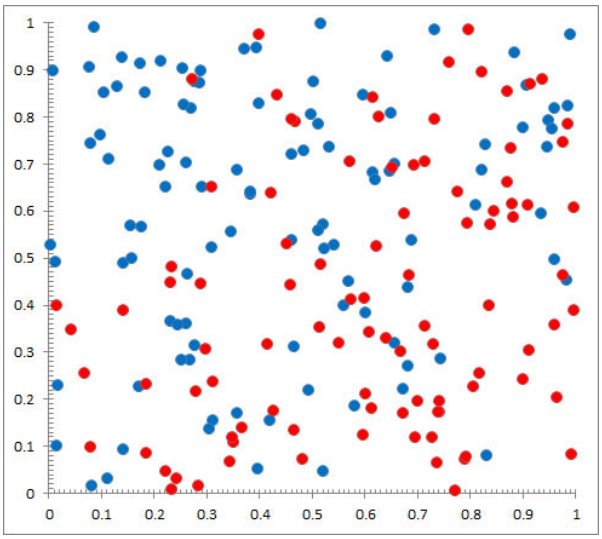

図 2.1 X-Y 平面における YES(青) と NO(赤) の分布

図 2.1 は, 平面上にばらまいた 200 個の点において, その位置 (X,Y) と YES/NO という値を観測して得られたデータである. たとえば, 200 人の人それぞれのある事柄に関する賛否とその人の身長・体重の関係を調べようとしてとったデータを図示すればこんなものになるだろう.

この X, Y と YES/NO の関係を CATDAP で調べると「ある事柄」に賛成する確率が次のようになっていることが分る.

表 2.3

	X ≦ 0.39	0.39 < X
0.49 < Y	95 %	52 %
Y ≦ 0.49	43 %	29 %

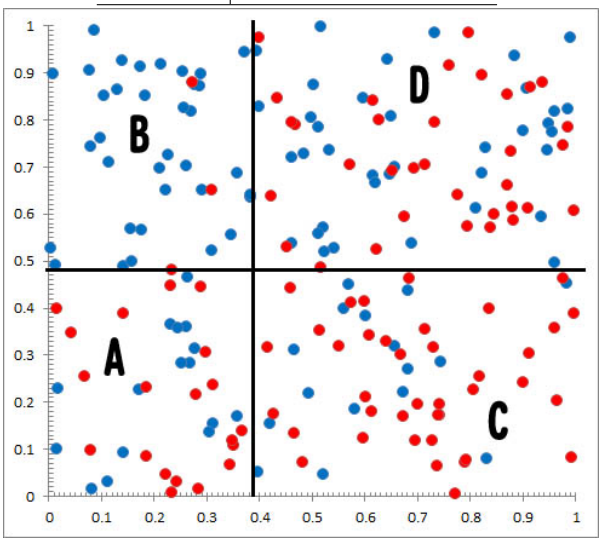

図 2.2 X-Y 平面の分割と YES(青) と NO(赤) の分布

実は, 図 2.1 のデータは X と Y の差 $(X-Y)$ の値がマイナスの時は賛成比率が高く, $(X-Y)$ の値がプラスの時は賛成比率が低くなるように作ったデータである. CATDAP は説明変数候補の組合せで目的変数の分布を説明しようとするが, 用意されているモデルは説明変数の空間を格子状に分割する形のものだけである. X と Y が $(X-Y)$, あるいは一般に X, Y の関数 $f(X,Y)$, の値を経由して目的変数の分布に影響している場合にも格子分割を細かくできればよいが, データ数の制約によってあまり細かい分割はできない. 細かく分割したモデルの AIC は大きくなり, そのようなモデルが採用されることがないのである.

ただし, CATDAP の結果を検討すると, $f(X,Y)$ の形がおぼろげに分る場合がある. 表 2.3 はそのような場合と言っていいだろう. このような場合には, X, Y に加えて $Z = f(X,Y)$ を定義して説明変数候補に加えて解析しなおすのがよい.

表 2.4 $Z = f(X,Y) = X - Y$ を説明変数候補に加えた場合の最適モデル

$X - Y \leqq -0.36$	$-0.36 < X - Y \leqq 0.21$	$0.21 < X - Y$
89 %	57 %	24 %

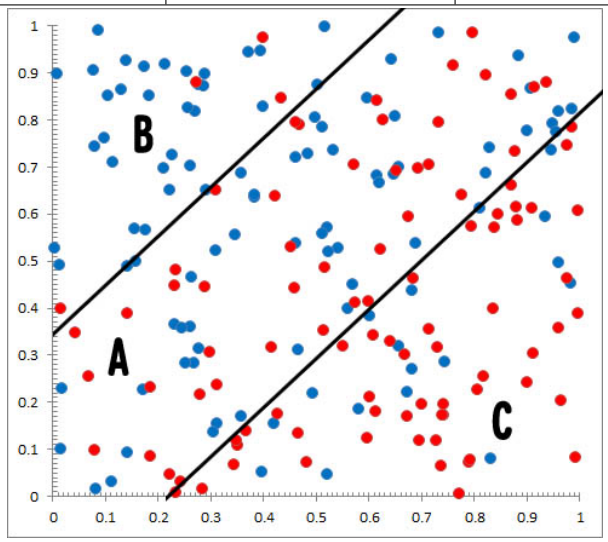

図 2.3 「X マイナス Y」という「指標」と YES(青) と NO(赤) の分布

なんらかの $Z = f(X,Y)$ の値で賛否が精度よく予測できるようになったら, この Z を賛否を判断する「指標」として利用することができる. SVM などはこのような指標を探す手法と考えることができる.

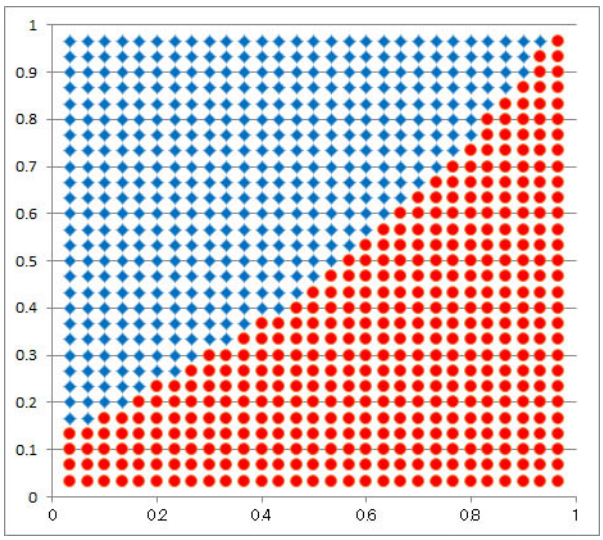

図 2.4 SVM ソフトに図 2.1 のデータをトレーニングデータとして与えた場合の学習結果の例

2.3 ヒストグラム解析の例

分割表 vs. ヒストグラム

X1=1	9	29	38	15	10	2	AIC = −0.8
X1=2	3	21	35	32	10	2	
total	12	50	73	47	20	4	AIC = 0.0

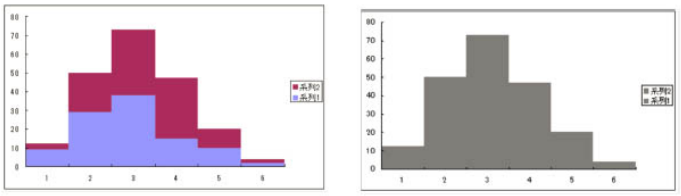

図 2.5 ヒストグラム解析

2.4 「関係」解析

変量 X と変量 Y の関係は X と Y の同時分布に現れる．この同時分布の他の変量への依存を見ることによって, X と Y の関係の有り方を研究することができる．

X と Y がともに離散変数であれば, X と Y を組合せて新しい離散変数を定義するという簡単な方法で, 上記の目的が達成される．たとえば 2 つの 2 値変数 X と Y から W という 4 値変数

$$W = \begin{cases} 1 & X = 0, Y = 0 \\ 2 & X = 0, Y = 1 \\ 3 & X = 1, Y = 0 \\ 4 & X = 1, Y = 1 \end{cases}$$

を構成して, W を目的変数として解析すればよい．

これに対して, 連続値変数 x と y の間の関係の第 3 の変数 z への依存を調べるには工夫がいる．たとえば,

$$y = a0\,(z) + a1\,(z) \times x + a2\,(z) \times x^2$$

のようなモデルを使う必要があるはずだ．

3 R 版 CATDAP

CATDAP を R の関数の形として使えるようにし，グラフ出力機能を強化したパッケージ catdap が CRAN(*) から入手可能である．

https://cran.r-project.org/web/packages/catdap/index.html

には，ソースファイルの他に Windows 用バイナリ，OS X 用バイナリ，PDF マニュアルが置かれている．インストール方法等については RjpWiki (www.okadajp.org/RWiki/) を参考されたい．

以下に，医学データを R 関数 catdap2( ) を使って解析した結果の一部を例として示す．

クロス表の評価などの出力の後に，目的変数（診断）に対し最も有効な説明変数の組（血清レベル，最大血圧）の多次元分割表と，その度数を可視化した帯グラフが表示される．帯の幅はその部分の比率の推定に寄与したデータ数に比例し, 比率の推定精度を反映するものとなっている．

```
 X   X                 response variable X(1)
(2) (3)            1                  2                  3                  4          Total
-------------------------------------------------------------------------------------------------
  1   1        4 (  25.0 )        6 (  37.5 )        6 (  37.5 )        0 (   0.0 )      16 ( 100.0 )
  1   2        3 (  42.9 )        3 (  42.9 )        1 (  14.3 )        0 (   0.0 )       7 ( 100.0 )
  2   1        1 (   5.6 )        3 (  16.7 )        4 (  22.2 )       10 (  55.6 )      18 ( 100.0 )
  2   2        6 (  54.5 )        3 (  27.3 )        1 (   9.1 )        1 (   9.1 )      11 ( 100.0 )
-------------------------------------------------------------------------------------------------
   Total      14 (  26.9 )       15 (  28.8 )       12 (  23.1 )       11 (  21.2 )      52 ( 100.0 )

<Note>
 X(1) : Diagnosis
  1     1
  2     2
  3     3
  4     4
 X(2) : Lev.Serum
  1       146.00   -     192.50
  2       192.50   -     279.00
 X(3) : Sys.press.
  1        98.00   -     165.50
  2       165.50   -     216.00

 AIC =    -7.54

> |
```

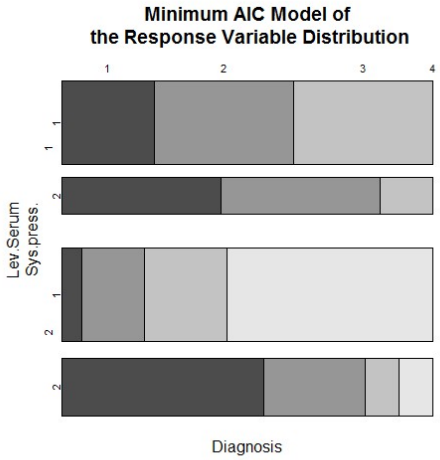

図 3.1 R 版 CATDAP の出力例

(*) CRAN (The Comprehensive R Archive Network) とは, フリーソフト R とその基本パッケージ及びユーザが開発した拡張（貢献）パッケージが公開されているアーカイブネットワークである.

4 参考文献

- Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle", 2nd Inter.Symp.on Information Theory, Petrov, B.N. and Csaki, F. eds., Akademiai Kiado, Budapest, 267-281. (Reproduced in Breakthroughs in Statistics, Vol.1, Foundations and Basic Theory, Kotz, S. and Johnson, N.J., eds., Springer-Verlag, New York, (1992) 610-624.)
- Katsura,K. and Sakamoto,Y.(1980);CATDAP, A categorical data analysis program; Computer Science Monographs, 14, The Institute of Statistical Mathematics, Tokyo.
- 坂元・石黒・北川 (1983). 情報量統計学. 共立出版.
- 坂元　慶行 (1985). カテゴリカルデータのモデル分析、共立出版. (英語版、Y.Sakamoto(1991). Categorical data analysis by AIC, Kluwer.)
- 坂元 (2001). 質的データのデータマイニング— 最適なクロス表の自動探索 CATDAP(1)(2),ESTRELA, No.91(pp.82-85) No.92(pp.84-87).

- Kogiso,T. Moriyoshi,Y. and Shimizu,S, et al.(2009). High-sensitivity C-reactiveprotein as a serum predictor of nonalcoholic fatty liver disease based on the Akaike Information Criterion scoring system in the general Japanese population, J. Castroenterol, 44:313-321, DOI 10:1007/s00535-009-0002-5.

5 付録

5.1 サンプルデータ

```
Y,X1,X2,X3
1, 1, 1, 1
1, 1, 1, 2
1, 1, 1, 3
1, 1, 1, 4
1, 1, 1, 5
1, 1, 2, 1
1, 1, 2, 2
1, 1, 2, 3
2, 1, 2, 4
1, 1, 2, 5
1, 1, 3, 1
2, 1, 3, 2
1, 1, 3, 3
1, 1, 3, 4
1, 1, 3, 5
1, 1, 4, 1
1, 1, 4, 2
2, 1, 4, 3
1, 1, 4, 4
1, 1, 4, 5
1, 1, 5, 1
1, 1, 5, 2
1, 1, 5, 3
1, 1, 5, 4
1, 1, 5, 5
1, 2, 1, 1
1, 2, 1, 2
1, 2, 1, 3
1, 2, 1, 4
2, 2, 1, 5
2, 2, 2, 1
2, 2, 2, 2
2, 2, 2, 3
1, 2, 2, 4
2, 2, 2, 5
1, 2, 3, 1
1, 2, 3, 2
1, 2, 3, 3
2, 2, 3, 4
1, 2, 3, 5
2, 2, 4, 1
```

```
1, 2, 4, 2
1, 2, 4, 3
2, 2, 4, 4
2, 2, 4, 5
1, 2, 5, 1
1, 2, 5, 2
2, 2, 5, 3
1, 2, 5, 4
2, 2, 5, 5
1, 3, 1, 1
2, 3, 1, 2
1, 3, 1, 3
1, 3, 1, 4
1, 3, 1, 5
1, 3, 2, 1
1, 3, 2, 2
2, 3, 2, 3
2, 3, 2, 4
1, 3, 2, 5
1, 3, 3, 1
2, 3, 3, 2
2, 3, 3, 3
2, 3, 3, 4
1, 3, 3, 5
2, 3, 4, 1
1, 3, 4, 2
2, 3, 4, 3
2, 3, 4, 4
1, 3, 4, 5
2, 3, 5, 1
1, 3, 5, 2
2, 3, 5, 3
1, 3, 5, 4
1, 3, 5, 5
2, 4, 1, 1
2, 4, 1, 2
1, 4, 1, 3
2, 4, 1, 4
2, 4, 1, 5
1, 4, 2, 1
2, 4, 2, 2
2, 4, 2, 3
1, 4, 2, 4
1, 4, 2, 5
1, 4, 3, 1
1, 4, 3, 2
2, 4, 3, 3
1, 4, 3, 4
2, 4, 3, 5
2, 4, 4, 1
1, 4, 4, 2
2, 4, 4, 3
2, 4, 4, 4
2, 4, 4, 5
```

```
1, 4, 5, 1
2, 4, 5, 2
2, 4, 5, 3
2, 4, 5, 4
1, 4, 5, 5
2, 5, 1, 1
1, 5, 1, 2
2, 5, 1, 3
2, 5, 1, 4
2, 5, 1, 5
2, 5, 2, 1
2, 5, 2, 2
2, 5, 2, 3
2, 5, 2, 4
1, 5, 2, 5
1, 5, 3, 1
2, 5, 3, 2
1, 5, 3, 3
2, 5, 3, 4
2, 5, 3, 5
2, 5, 4, 1
2, 5, 4, 2
2, 5, 4, 3
2, 5, 4, 4
2, 5, 4, 5
2, 5, 5, 1
2, 5, 5, 2
1, 5, 5, 3
2, 5, 5, 4
2, 5, 5, 5
```

5.2 離散予測の誤差表示

離散的な確率事象の生起確率を表現するのにドーナツグラフを使うことが出来る. たとえば, 生起確率 3/5 を次の図の青の部分で表現できる.

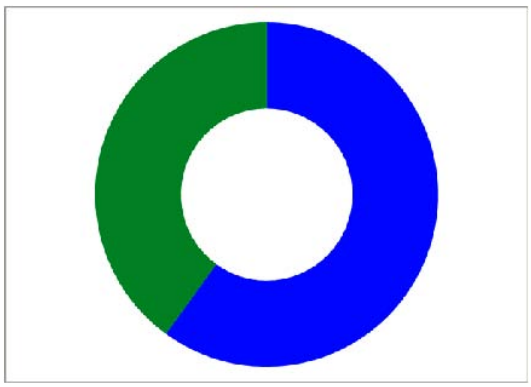

図 7.1 青 3/5 緑 2/5

3/5 という確率の表示としてはこれで十分であるが, 3/5 という確率がデータに基づく推定値である場合には, 誤差を含まないということはあり得ない. この誤差の程度は 3/5 が 5 回の観測中 3 回起きた事象の生起確率の推定値であるばあいと, 500 回の観測中 300 回起きたというデータに基づく場合でかなり違うはずである.

以下で離散事象の生起確率の推定値の誤差の程度を表現する方法を考える.

事象 $i(=1,2,...,C)$ がそれぞれ N_i 回起きたというデータにもとづく事象 i の生起確率の最尤推定量は

$$N=\sum_{i=1}^{C}N_i$$

として

$$P\left(i\right)=N_i/N$$

で与えられ, その推定誤差の標準偏差は

$$\sigma\left(i\right)=\sqrt{\frac{P\left(i\right)\left(1-P\left(i\right)\right)}{N}}$$

と見積もられる. 図 7.1 は

$$N_1=3$$
$$N_2=2$$

というデータから求めた $\{P(1),P(2)\}$ を表示している. 青の部分が $P(1)$ を表し, 緑の部分が $P(2)$ を表している.

この図の中に $\{\ \sigma\ (1),\sigma\ (2)\}$ の情報は反映されていない. $P(1)+P(2)=1$ であるために, 「ドーナツ」の中に $\{\ \sigma\ (1),\sigma\ (2)\}$ を置く余地がないのである.

この情報を表示する余地を作るために

$$\tilde{P}\left(i\right)=P\left(i\right)-\ \sigma\ \left(i\right)$$

を定義する. そして

$$\left\{\sigma\ \left(i\right)/2,\tilde{P}\left(i\right),\sigma\ \left(i\right)/2\right\}\quad\left(i=1,2,\ldots,C\right)$$

に比例するコンポーネントをドーナツの円環部分に配置した結果が 図 7.2 である. 図の青に部分が $\tilde{P}(1)$ に対応, 緑の部分が $\tilde{P}(2)$ にあたる. それ以外の部分がグレーにしてある.

$$\sum_{i=1}^{C}\sigma\ \left(i\right)/2+\tilde{P}\left(i\right)+\ \sigma\ \left(i\right)/2=1$$

であるから, きっちりと円環の中に納まるのは当然である. この形の表示を $DOUGHNUTwithERROR\{N_1,N_2,...\}$ と「名付けることとする.

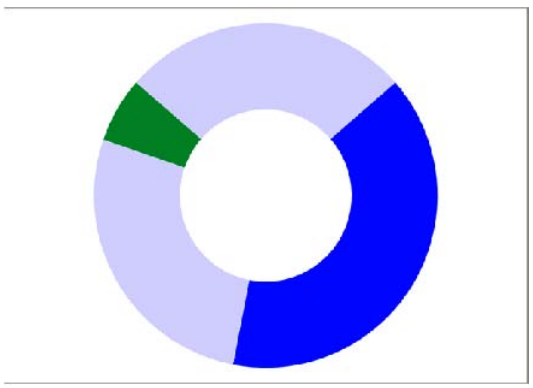

図 7.2 $DOUGHNUTwithERROR\{2,1\}$

P(1) という推定値は誤差を含んでいるが, その値が青で示されている値より小さくなる確率は 15 %程度であるから, 実用的には「P(1) の値は誤差を含んでいるが青で示されているよりは大きい」と考えることが許されるだろう. もちろん P(1) が過小推定で, 真値はグレーの部分に大きく食い込んでいる可能性もある. いずれにせよ「真値」にはグレーの部分はない.

5.3 分割表モデルの選択と誤差表示

(2) モデルでデータ J と I の関係を記述するということはデータ $\{N_{ij}|i=1,...,C_I, j=1,2,...,C_J\}$ を C_J 個の部分データ群 $\{N_{i1}|i=1,...,C_I\}$, $\{N_{i2}|i=1,...,C_I\}$,..., $\{N_{ij}|i=1,...,C_I\}$, ... に分割してそれぞれの確率構造を $DOUGHNUTwithERROR\{N_{11},N_{21},...\}$, $DOUGHNUTwithERROR\{N_{12},N_{22},...\}$, ..., $DOUGHNUTwithERROR\{N_{1j},N_{2j},...\}$, ... で表現するということである.

$$AIC\left\{N_1,N_2,\ldots=-2\times\left\{\sum_{i=1}^{C}\left(N_i\log N_i-1\right)-\left(N\log N-1\right)\right\}\right.$$

と書くことにすると (8) 式は

$$AIC_I=AIC\left\{N_{1.},N_{2.},\ldots\right\}$$

と表され, (7) 式は

$$AIC_{I|J}=\sum_{j=1}^{C_J}AIC\left\{N_{1j},N_{2j},\ldots\right\}$$

と書ける. ここで $AIC\{N_{1j},N_{2j},...\}$ が $DOUGHNUTwithERROR\{N_{1j},N_{2j},...\}$ を評価する量になっていることに注意.

5.3.1 データ数 8 の場合

V2	V1=1	V1=2
1	2	1
2	3	2
計	5	3

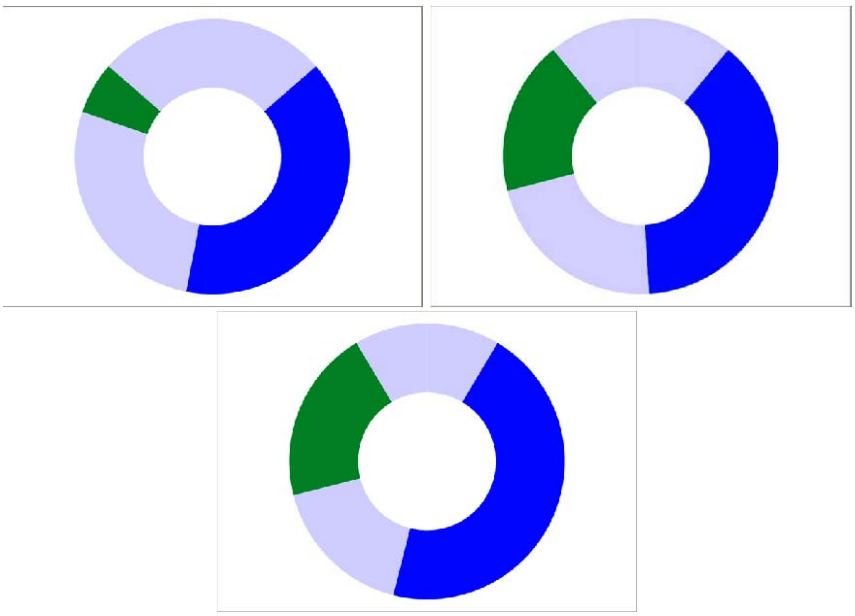

図 7.3 $AIC\{2,1\} + AIC\{3,2\} = 5.82 + 8.73 > 12.59 = AIC\{5,3\}$

5.3.2 データ数 80 の場合

V2	V1=1	V1=2
1	20	10
2	30	20
計	50	30

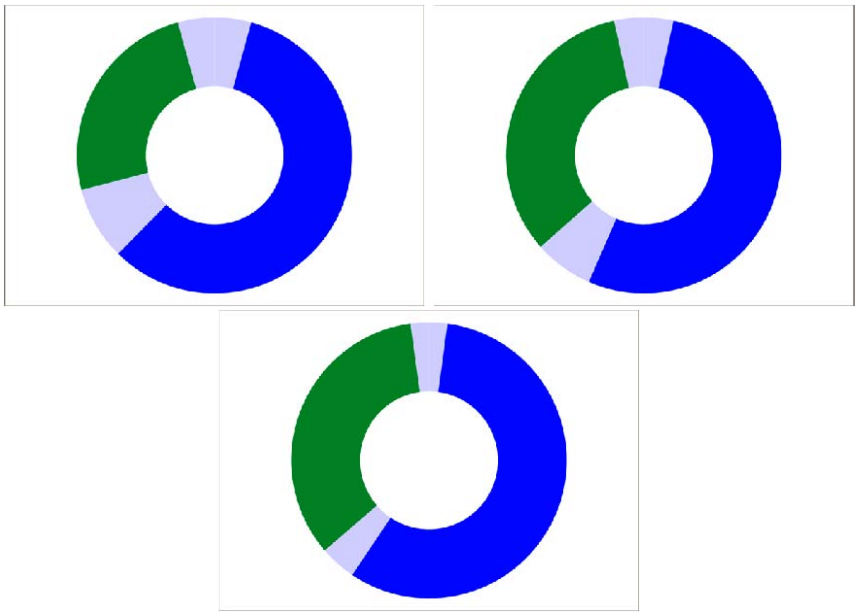

図 7.4 $AIC\{20,10\} + AIC\{30,20\} = 40.19 + 69.30 > 107.85 = AIC\{50,30\}$

5.3.3 データ数 800 の場合

V2	V1=1	V1=2
1	200	100
2	300	200
計	500	300

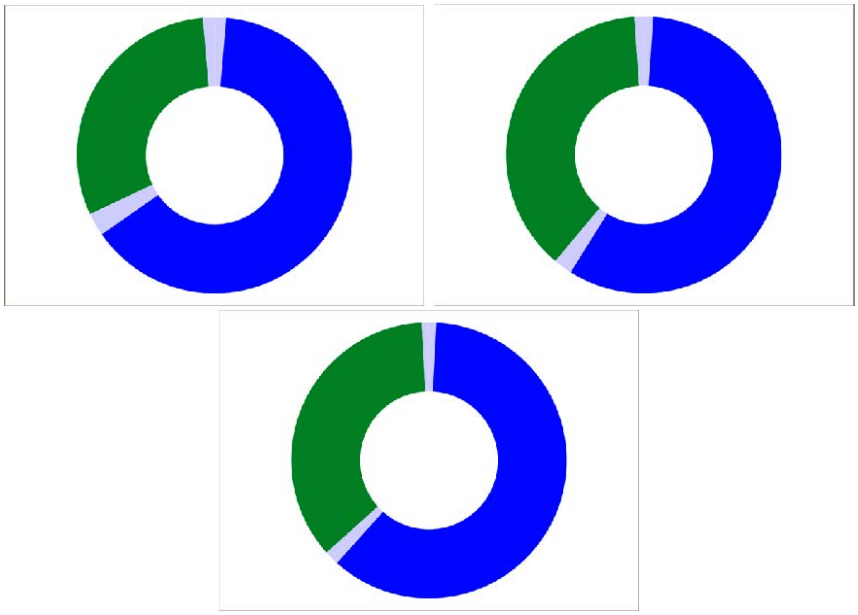

図 7.5 $AIC\{200, 100\} + AIC\{300, 200\} = 383.91 + 675.01 < 1060.50 = AIC\{500, 300\}$

5.3.4 データ数 8000 の場合

V2	V1=1	V1=2
1	2000	1000
2	3000	2000
計	5000	3000

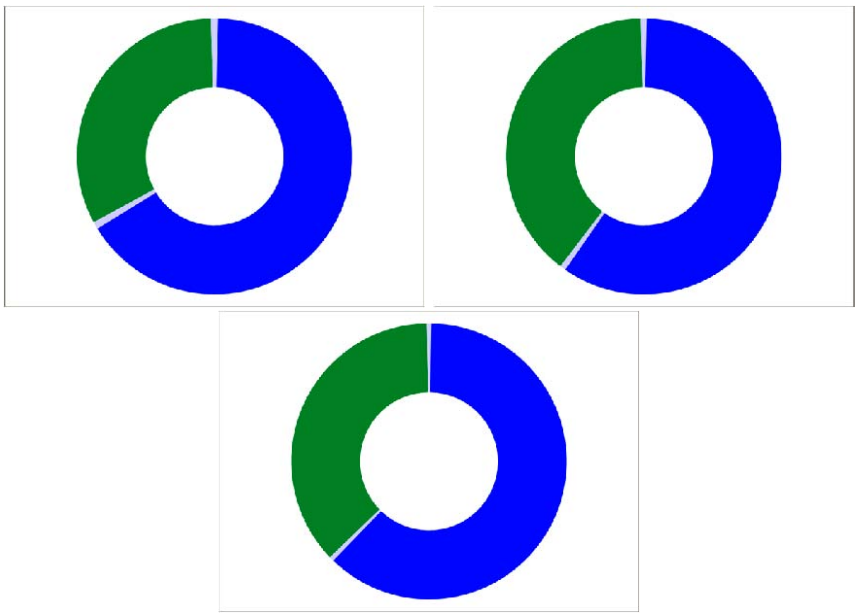

図 7.6 $AIC\{2000, 1000\} + AIC\{3000, 2000\} = 3821.09 + 6732.12 < 10587.01 = AIC\{5000, 3000\}$