J Jpn Ind Manage Assoc 50, 163-170, 1999

Original

A Study of Parametric Estimation in Contingency Tables

Yoshiko KIKUCHI,[3] Masayuki GOTOH[2] and Nobuhiko TAWARA[1]

Abstract

In this paper, we consider the contingency table as a two-way classification with binomial distribution at each of the cells. The contingency table is useful for many practical cases in the field of the quality control, market research, and so on. In the analysis of data of the contingency table, the probabilistic models with parameters are assumed at each cell and the parameters are estimated from observations (data sets). In the case of binomial distribution, the parameters are the probabilities of each cell. The maximum likelihood estimator (MLE) for each cell is one of the parameter estimators. However, the precision of the MLE may not be sufficient when the sample size is small. On the other hand, if we previously assume the prior density on parameter space, we can formulate Bayes optimal estimation of the parameter with the square error loss function. This estimator is Bayes optimal for the finite sample. The another way for precise estimation is to estimate the parameter after the hypothesis, such that the parameter of some cell is equal to that of another cell. If the hypothesis is correct, the estimation error may be reduced rather than parameters being independently estimated at each cell. If we can previously assume a hypothetical set, a hypothesis may be selected from the observations, but in a true hypothesis, the set is unknown. In such a method, although we can regard a hypothesis as a model, this method can be formulated based on the statistical model selection problem, and we can apply conventional information criteria to select the model. However, in Bayes decision theory, the selection of a model is not optimal for prediction. We can formulate Bayes optimal estimator on the condition that the set of candidates of the models is given. In this paper, we propose a method to predict future observations and estimate the parameter values at each cell based on Bayes optimal solution, which uses a mixture model of all candidates. That is, we shall show that the method using the mixture model is Bayes optimal solution with the square error loss function for the parameter estimation similar to the prediction problem. Moreover, we shall show a practical algorithm to calculate Bayes optimal estimator using a mixture model assuming Bata distribution and uniform distribution as the prior distributions for the parameter space and model class, respectively. Through the simulation experiment, we shall show the properties of the proposed method for parameter estimation.

Key words: contingency table, parametric estimation, Bayes decision theory

[1] Musashi Institute of Technology
[2] Waseda University
[3] COPYER Co., Ltd.
Received: April 10, 1998
Accepted: February 26, 1999

論　　文

順序カテゴリカルデータ解析における母数推定に関する研究

菊池淑子[3]，後藤正幸[2]，俵　信彦[1]

本研究では，分割表の各セルの母数を推定する問題について，一つの仮説（モデル）に基づいて推定するのではなく，考えられる複数のモデルの混合モデルを用いて推定する方法を提案する．一般的な分割表の母数推定の方法は，採択された仮説，選ばれたモデルに基づいて推定するものであるが，一つのモデルの下で推定するということは，データ数が少ないときには推定精度が安定していない可能性がある．他方，目的が次に出現するデータの予測である場合には，ベイズ最適な予測はすべてのモデルの混合モデルによって与えられるという研究がなされており，母数の推定においても，モデルを一つに絞るという制約から抜けて推定した方が有効であると考えられる．そこで，まず分割表の各セルに仮定する確率分布が二項分布である場合を取り上げ，ベイズ最適なデータの予測と母数の推定が混合モデルによって与えられることを示す．さらに母数の推定に関して，事前分布にベータ分布を仮定した推定式を提案する．シミュレーション実験の結果，提案した推定式は，推定誤差の点では従来法よりもデータ数の少ないときには有効であるということがわかった．

キーワード：分割表，母数推定，ベイズ決定理論

1. はじめに

データが定量的に得られない場合など，カテゴリカル・データとして分割表の形で示される問題は数多く存在する．実際，品質管理やマーケティングなどの分野では分割表形式でまとめられるデータが多く，順序関係のあるカテゴリカル・データの解析に関する研究はその応用価値からみても興味深いものとなっている〔1〕，〔2〕．

この分割表の解析においては，各セルに確率分布を仮定し，その母数をデータから推定する方法がまず考えられる．しかし各セルの母数を最尤推定法によって求める場合，その最尤推定値は各セルごとに算出されるので，各セルにおけるデータ数が少ないと推定精度が十分でないことがある．また，事前の知識を事前分布として仮定できれば，ベイズ理論によって平均推定誤差最小の母数推定値を導くことができる．分割表においてパラメータに事前分布を仮定したベイズ的アプローチにより安定した推定値が得られるという報告もなされている〔3〕．

一方，分割表における各セル間で母数に差があるかどうかの仮説検定を行い，採択された仮説のもとで母数を推定する方法がある．これを一般的に考えてみると，同じとみなせる母数をもつセル同士まとめたものを一つのモデル（仮説）とし，候補となるモデルの集合の中から適切なモデルを一つ選択（推定）するという問題ととらえることができる．このような立場から，AIC等を用いてモデルを選択し，確率モデルを特定する方法が提案されている〔4〕．モデルを選択する基準はいくつかあるが，これらはどのような目的でモデルを選択するのかによってさまざまであり，この場合の推定値は，選ばれたただ一つのモデルのもとで推定されることになる．しかし，モデルを一つ選択し，そのもとで母数を推定することは，他の候補となるモデルの可能性を無視することになる．一方，ベイズ規準に基づけば，目的が母数の推定ではなく次に出現するデータの予測である場合，（一つのモデルを選択するのではなく）考え得るすべてのモデルの混合モデルを用いて予測することがベイズ最適であることが知られている〔5〕，〔6〕．

本研究では簡単のため，分割表における各セルの確率分布が二項分布の場合について取り上げる．二項分布は多くの適用場面をもつ応用上重要なモデル族であり，また本稿の議論を多項分布へ拡張することは容易である〔7〕．モデル選択による定式化と同様に，合わせたセルの組み合わせを一つのモデルと考える．その上でまず，この構成のしかたによって考え得るすべてのモデルの集合に対し，ベイズ最適な予測法を定式化する．さらに，二項分布を仮定した場合には，次のデータを（0なのか1なのか）予測することと各セルの出現確率を推定することは，同じ混合分布を用いた決定がベイズ最適という意味で同等であることを示し，混合モデルによるベイズ最適な母数推定方法を提案す

[1] 武蔵工業大学
[2] 早稲田大学
[3] コピア株式会社
受付：1998年4月10日，再受付（1回）
受理：1999年2月26日

る．最後に，数値実験を通じて，AICでモデルを選択する方法と提案したベイズ最適な推定方法を比較することにより，提案法が推定誤差の点で優れていることを示す．

2. 問題設定

2.1 二項分布を仮定した分割表の基本設定

要因Aと要因Bで構成される分割表において，要因Aの水準を$\{a_1, a_2, \cdots, a_\eta\}$，要因$B$の水準を$\{b_1, b_2, \cdots, b_\zeta\}$とする．水準の組み合わせは分割表の1つのセルを表わし，各セルをa_ib_jとおく$(i=1,2,\cdots,\eta,\ j=1,2,\cdots,\zeta)$．セル$a_ib_j$において，確率$p_{ij}$で1，確率$(1-p_{ij})$で0が出現する二項分布を考える．例えば，ある製造品に対して1が不良品，0が良品とすれば，p_{ij}は不良率を与える．また，薬効問題において各要因の水準を薬の投入量とし，ある効果が現れた場合を1，現れなかった場合を0とすれば，p_{ij}はその効果の現れる確率を表わす．二項分布はこのようにいくつかの重要な応用例をもつ分布である[1]．各セルごとにデータ数n_{ij}が与えられたとき，この二項分布は

$$\frac{n_{ij}!}{(x_{ij})!(n_{ij}-x_{ij})!}p_{ij}^{x_{ij}}(1-p_{ij})^{n_{ij}-x_{ij}} \qquad (1)$$

となる．ここで，x_{ij}はセルa_ib_jにおける1の生起回数であり，$n_{ij}-x_{ij}$は0の生起回数である．

独立にランダムサンプリングされたすべてのデータを$Z^N=\{x_{11}, x_{12}, \cdots, x_{\eta\zeta}\}$，次に出現する（予測しようとしている）データを$y\in\{0,1\}$とする．

いま，要因Aと要因Bの水準には，例えば薬品の投入量などのように，順序があるものとする．このとき，隣り合うセルで母数pが同じであるという仮説を考えることができる．隣り合うセルで母数pが同じ場合には，これらの間の仕切りは意味を持たないので，これらをまとめあわせて新たなセルを作ることを試みる．このセルを本研究では拡大セルと呼び，このようにして作った拡大セルの境界を要因Aについては$\{u_1, u_2, \cdots, u_U\}$，要因$B$については$\{r_1, r_2, \cdots, r_R\}$とする．$u_k, u_k+1, r_l, r_l+1$で囲まれるこの拡大セルを$S_{kl}^{(m)}$，拡大セルでの新たな母数を$\theta_{kl}^{(m)}$と書く（ただし$k=1,2,\cdots,U,\ l=1,2,\cdots,R$）．サフィックスの$(m)$は，拡大セルの作り方は一つのモデル$m$を与えていることを意味し，拡大セルの作り方すべてによって，様々なモデル$m\in\mathcal{M}=\{m_1, m_2, \cdots, m_t\}$が考えられる．本研究では，モデルの集合のうちで真のモデルに近いモデルを選択したり，要因間の関係を見るのではなく，ベイズ的に最適なyの予測と各セルの母数$\theta_{kl}^{(m)}$の推定を考える．

表1 2×3分割表

		B		
		b_1	b_2	b_3
A	a_1	x_{11}	x_{12}	x_{13}
	a_2	x_{21}	x_{22}	x_{23}

表2 モデル1

	B		
A	b_1	b_2	b_3
a_1	x_{11}	x_{12}	x_{13}
a_2	x_{21}	x_{22}	x_{23}

表3 モデル2

	B		
A	b_1	b_2	b_3
a_1	x_{11}	x_{12}	x_{13}
u_1			
a_2	x_{21}	x_{22}	x_{23}

表4 モデル3

	B	r_1	
A	b_1	b_2	b_3
a_1	x_{11}	x_{12}	x_{13}
a_2	x_{21}	x_{22}	x_{23}

表5 モデル4

	B	r_1	
A	b_1	b_2	b_3
a_1	x_{11}	x_{12}	x_{13}
a_2	x_{21}	x_{22}	x_{23}

表6 モデル5

	B	r_1	r_2
A	b_1	b_2	b_3
a_1	x_{11}	x_{12}	x_{13}
a_2	x_{21}	x_{22}	x_{23}

表7 モデル6

	B	r_1	
A	b_1	b_2	b_3
a_1	x_{11}	x_{12}	x_{13}
u_1			
a_2	x_{21}	x_{22}	x_{23}

表8 モデル7

	B		r_1
A	b_1	b_2	b_3
a_1	x_{11}	x_{12}	x_{13}
u_1			
a_2	x_{21}	x_{22}	x_{23}

表9 モデル8

	B	r_1	r_2
A	b_1	b_2	b_3
a_1	x_{11}	x_{12}	x_{13}
u_1			
a_2	x_{21}	x_{22}	x_{23}

例えば**表3**の分割表において，要因A, Bについて境界線を引く，引かないの組み合わせを考えると8とおりの分割表モデルを想定することができる（**表2**～**表9**）．**表3**は水準a_1での出現確率とその他すべてのセルの出現確率が等しいモデル，**表9**はすべてのセルの出現確率が等しいと見なせない場合のモデルである．境界線は，**表3**，**表5**のように要因A, Bのみに引いてもよいし，**表7**のように要因で引く数が異なってもよい．混合モデルによる予測および推定では，このように想定できるモデルをすべて用いる．

2.2 適用の際の注意点

本研究では隣り合うセル同士で出現確率が同じとみなせるかどうかで境界線を引き，モデルのパターンを想定しているため，分割表の要因（水準）によっては，本研究での議論を適用できないケースもある．要

[1] 本稿の議論は，そのまま多項分布に拡張することが可能である．

因の水準が順序関係のあるもの，例えば薬品の濃度や投入量などでは適用できるが，色の種類で水準をとる場合では，複雑なセルの区切り方を検討する必要がある．しかし，その場合はそのようなすべてのセルの区切り方を考えてモデルの集合を作り，本稿の結果を用いればよく，その意味で拡張は容易である．

3. ベイズ決定理論に基づく定式化

本研究では，分割表の各セルに次に出現するデータを予測するという問題を予測問題と呼ぶことにする．具体的には，**表 10** のように，各セルに 0 が出現するのか 1 が出現するのかを予測するものである．また，各セルの出現確率を推定する問題を推定問題と呼ぶことにし，これは**表 11** のように，各セル一つ一つにどのような確率が存在しているのかを推定するものである．

3.1 二項分布を仮定した場合のデータの予測

セル a_ib_j の次のデータ $y\in\{0,1\}$ に対する決定関数を $Ay\in\{0,1\}$ とし，損失関数として

$$d_y(y,Ay)=\begin{cases}0 & (y=Ay)\\ 1 & (y\neq Ay)\end{cases} \qquad (2)$$

を定義する．これは，予測値 Ay に対して y が出現した場合の損失である〔8〕．

リスク関数 R は決定関数 $Ay(y|Z^N)$ を用いたときのモデルの期待損失であり，確率モデルの悪さをはかる尺度である．

$$R_{ij}=\sum_y d_y(y,Ay)P_{ij}(y|Z^N,m) \qquad (3)$$

任意の m に対して，一様に R を最小にする決定関数は存在しないので，ベイズ決定理論では事前分布を導入し，リスク関数を平均化したベイズリスク BR の最小化を行う．

$$BR_{ij}=\sum_m\sum_y d_y(y,Ay)P_{ij}(y|Z^N,m)P(m|Z^N)$$
$$=\sum_y d_y(y,Ay)P_{ij}^{mix}(y|Z^N) \qquad (4)$$

ここで，P_{ij}^{mix} は事後混合分布と呼ばれ，

$$P_{ij}^{mix}(y|Z^N)$$
$$=\sum_m P_{ij}(y|Z^N,m)P(m|Z^N)$$
$$=\sum_m\int_{\theta_{kl}^{(m)}} P_{ij}(y|\theta_{kl}^{(m)},m)$$
$$\cdot P(\theta_{kl}^{(m)}|Z^N,m)P(m|Z^N)d\theta_{kl}^{(m)} \qquad (5)$$

ただし，$i=1,2,\cdots,\eta$, $j=1,2,\cdots,\zeta$ に対し，

$$P_{ij}(y|\theta_{kl}^{(m)},m)=\begin{cases}\theta_{kl}, & (y=1)\\ 1-\theta_{kl}, & (y=0)\end{cases}$$
$$\text{かつ}\quad a_ib_j\in S_{kl}^{(m)} \qquad (6)$$

である．

(4)式のベイズリスクを最小にする決定関数 Ay は，

$$Ay=\begin{cases}0, & \left(P_{ij}^{mix}(y=0|Z^N)>\frac{1}{2}\right)\\ 1, & \left(P_{ij}^{mix}(y=1|Z^N)>\frac{1}{2}\right)\end{cases} \qquad (7)$$

で与えられ，$P_{ij}^{mix}(y=0|Z^N)=P_{ij}^{mix}(y=1|Z^N)=1/2$ のときはランダム決定とする．よって，ベイズ最適な決定関数 Ay は，混合モデル $P_{ij}^{mix}(y|Z^N)$ を用いれば計算できることになる．

3.2 二項分布を仮定した場合の出現確率の推定

セル a_ib_j における出現確率の推定問題に対する決定関数を $A\theta\in(0,1)$ とおき，真の確率が $\theta\in(0,1)$ であった際の損失関数を二乗誤差損失

$$d_\theta(\theta,A\theta)=\|\theta-A\theta\|^2 \qquad (8)$$

と定義する．このときリスク関数 R は

$$R_{kl}=\int_{\theta_{kl}^{(m)}} d_\theta(\theta_{kl},A\theta_{kl})P(\theta_{kl}^{(m)}|Z^N,m)d\theta_{kl}^{(m)} \qquad (9)$$

となり，ベイズリスク BR_{kl}

$$BR_{kl}=\sum_m\int_{\theta_{kl}^{(m)}} d_\theta(\theta_{kl},A\theta_{kl})$$
$$\cdot P(\theta_{kl}^{(m)}|Z^N,m)P(m|Z^N)d\theta_{kl}^{(m)} \qquad (10)$$

を最小にする決定 $A\theta$ は，

$$A\theta_{ij}^*=\sum_m\int_{\theta_{kl}^{(m)}}\theta_{ij}^{kl}P(\theta_{kl}^{(m)}|Z^N,m)P(m|Z^N)d\theta_{kl}^{(m)} \qquad (11)$$

ただし，$k=1,2,\cdots,U$, $l=1,2,\cdots,R$

$$\theta_{ij}^{kl}=\theta_{kl}\quad\text{かつ}\quad a_ib_j\in S_{kl}^{(m)} \qquad (12)$$

で与えられる．

ここで，(6)の $P_{ij}(y|\theta_{kl}^{(m)},m)$ の $y=1$ のときの値を(5)式に代入すると，(5)式は(11)式と一致することがわかる．このことから，確率分布に二項分布を仮

表 10 予測問題の決定の例

		B			
		b_1	b_2	b_3	b_4
A	a_1	0	1	0	1
	a_2	1	0	0	0
	a_3	1	0	0	1
	a_4	1	1	1	0

表 11 推定問題の決定の例

		B			
		b_1	b_2	b_3	b_4
A	a_1	0.5	0.5	0.4	0.4
	a_2	0.5	0.5	0.4	0.4
	a_3	0.3	0.3	0.7	0.7
	a_4	0.2	0.2	0.1	0.1

定した場合は予測問題と推定問題は，同じ混合モデルを用いて決定できることがわかる．

3.3 本研究の展開

(11)式を用いれば，ベイズ最適な推定を行うことができるが，実際的には，(11)式をより計算しやすい式に変形することが望まれる．そこで，各拡大セルの二項確率に対する事前分布として自然共役事前分布であるベータ分布を仮定し，計算式を導く．ベータ分布は一様分布や Jaffreys prior を含むという意味で事前知識の表現能力も高く，事後分布もベータ分布になるという共役性が計算式の導出を可能にするため，二項分布の事前分布として広く適用される分布である．

(11)式において，$\int_{\theta_{kl}^{(m)}} \theta_{ij}^{kl} P(\theta_{kl}^{(m)}|Z^N, m)d\theta_{kl}^{(m)}$ はベータ分布の期待値であり，

$$P(\theta_{kl}^{(m)}|Z^N, m) = \frac{P(Z^N|m, \theta_{kl}^{(m)})P(\theta_{kl}^{(m)}|m)}{P(Z^N|m)} \quad (13)$$

である．ここで，$P(Z^N|m, \theta_{kl}^{(m)}), P(\theta_{kl}^{(m)}|m)$ は尤度関数であるから

$$\begin{aligned}
&\int_{\theta_{kl}^{(m)}} \theta_{ij}^{kl} P(\theta_{kl}^{(m)}|Z^N, m)d\theta_{kl}^{(m)} \\
&= \int_{\theta_{kl}^{(m)}} \theta_{ij}^{kl} \frac{P(Z^N|m, \theta_{kl}^{(m)})P(\theta_{kl}^{(m)}|m)}{P(Z^N|m)} d\theta_{kl}^{(m)} \\
&= \int_{\theta_{kl}^{(m)}} \prod_{k=1}^{U+1}\prod_{l=1}^{R+1} \theta_{ij}^{kl} \frac{1}{P(Z^N|m)} \\
&\quad \cdot (\theta_{kl}^{(m)})^{\sum_{i=u_{k-1}+1}^{u_k}\sum_{j=r_{l-1}+1}^{r_l} x_{ij}} \\
&\quad \cdot (1-\theta_{kl}^{(m)})^{\sum_{i=u_{k-1}+1}^{u_k}\sum_{j=r_{l-1}+1}^{r_l} (n_{ij}-x_{ij})} \\
&\quad \cdot \frac{1}{B(\alpha, \beta)} \theta_{kl}^{(m)\alpha-1}(1-\theta_{kl}^{(m)})^{\beta-1} d\theta_{kl}^{(m)} \\
&= \prod_{k,l} \frac{1}{P(Z^N|m)} \frac{1}{B(\alpha, \beta)} \\
&\quad \cdot \frac{\Gamma\left(\sum_{i,j} x_{ij}+\alpha+1\right)\Gamma\left(\sum_{i,j}(n_{ij}-x_{ij})+\beta\right)}{\Gamma(\sum_{i,j} n_{ij}+\alpha+\beta+1)} \\
&= \prod_{k,l} \frac{\sum_{i,j} x_{ij}+\alpha}{\sum_{i,j} n_{ij}+\alpha+\beta} \quad (14)
\end{aligned}$$

よって，(11)式は

$$A\theta_{ij}^{*} = \sum_m \sum_{a_i b_j \in S_{kl}^{(m)}} \frac{x_{ij}+\alpha}{n_{ij}+\alpha+\beta} P(m|Z^N) \quad (15)$$

ここで，モデルの事後分布 $P(m|Z^N)$ について

$$P(m|Z^N) = \frac{P(Z^N|m)P(m)}{P(Z^N)} \quad (16)$$

$$\begin{aligned}
&P(Z^N|m)P(m) \\
&= \int_{\theta_{kl}^{(m)}} P(m)P(Z^N|\theta_{kl}^{(m)}, m)P(\theta_{kl}^{(m)}|m)d\theta_{kl}^{(m)} \\
&= P(m)\prod_{k,l} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \\
&\quad \cdot \frac{\Gamma(\sum_{i,j} x_{ij}+\alpha)\Gamma(\sum_{i,j}(n_{ij}-x_{ij}))+\beta}{\Gamma(\sum_{i,j} n_{ij}+\alpha+\beta)} \quad (17)
\end{aligned}$$

であり，基準化定数 $P(Z^N)$ は

$$\begin{aligned}
P(Z^N) = \sum_m \int_{\theta_{kl}^{(m)}} &P(m) \\
&\cdot P(Z^N|\theta_{kl}^{(m)}, m)P(\theta_{kl}^{(m)}|m)d\theta_{kl}^{(m)} \quad (18)
\end{aligned}$$

となる．

3.4 混合モデルによる出現確率推定に関する考察

混合モデルにおいて，母数 θ は，その積分範囲のために複雑な積分操作を要するが，二項分布に限定し，事前分布をベータ分布と仮定すれば，このような積分操作は排除される．これによって，(15)式のようなまとまった形になり，実務的な使用が可能となる．

(15)式における $\sum_{a_i b_j \in S_{kl}^{(m)}} \frac{x_{ij}+\alpha}{n_{ij}+\alpha+\beta}$ は，あるモデル m において，拡大セルごとにデータ x_{ij}, n_{ij} をそれぞれたし合わせ，さらに α, β を加えた形となっている．これは，ベイズ的アプローチによって出現確率に対して事前分布を考慮したことによるものであり，α, β の値を変化させることによって，出現確率に関する事前情報を組み込ませることが可能である．

また，この項にモデルの事後確率をかけ合わせることは，「真のモデルの可能性」の度合いをそれぞれのモデル m ごとに求め，それをこの項に重み付けして推定することになる．(15)式は，モデルの事前分布 $P(m)$ の値を変えれば，モデルに関する情報をも考慮して推定できることを表している．

4. シミュレーション実験

提案法の有効性を検証するために，AIC によって選ばれたモデルのもとで推定する方法（従来法）〔9〕を取り上げる．従来法と提案法とで推定した値と真の値からそれぞれの平均二乗誤差を求め，これによって評価する．

4.1 シミュレーション条件

各セルに二項分布を仮定し，**表 1** のような 2×3 分割表を取り上げ，想定できるモデルは**表 2**～**表 9** の 8 モデルとし，真のモデルを次のように設定する．

〈シミュレーション条件 1：真のモデルはモデル 6〉

モデル 6 の下で，真の値を次のように変える．

パターン 1：真の値が適当に離れているとき

$p_{11}=0.3, p_{12}=p_{13}=0.7$

$p_{21}=0.5, p_{22}=p_{23}=0.8$

パターン 2：真の値が各セルで近いとき

$p_{11}=0.4, p_{12}=p_{13}=0.6$

$p_{21}=0.5$, $p_{22}=p_{23}=0.7$

パターン3：真の値が離れているとき

$p_{11}=0.1$, $p_{12}=p_{13}=0.7$

$p_{21}=0.2$, $p_{22}=p_{23}=0.8$

各セルごとのデータ数 n_{ij} を等しく $n=n_{ij}=5, 10, 20, 30, \cdots, 200$ と変化させ，シミュレーション回数は1000回とする．また，モデルの事前分布および出現確率の事前分布はともに無情報事前分布とし，一様分布を仮定する〔6〕．すなわち

$$P(m)=\frac{1}{|\mathcal{M}|},\quad \alpha=\beta=1 \tag{19}$$

とする．

4.2 シミュレーションの結果および考察

どのパターンについても同じような結果が得られたが，ここでは代表として条件1パターン2の結果を示す．

各パターンの結果をまとめると

（I） 平均二乗誤差は，データ数 n が大きくなっていくにしたがってAICでも混合法でも小さくなっていく．

（II） 平均二乗誤差は，すべてのパターンで混合による推定の方がAICよりも小さく，その差は n が少ないときほど大きい．

（III） 真のモデルの事後確率は，n が大きくなるとそれに伴って上昇する．

AIC規準でモデルを一つ選択して推定する従来法よりも，すべてのモデルの混合モデルを用いて推定する提案法の方が平均二乗誤差が小さいのは，提案法はモデルの事後確率で平均化した推定量を用いているためと考えられる．**図3**をみると，データ数が少ないところでは事後確率が各モデルに残っており，一つのモデルの事後確率が高くなっていない．これは定性的には，データ数が少ないために，まだ真のモデルを特定できていないことに起因する．このようなデータ数が少なく，まだ曖昧性が残っている段階で，従来法ではモデルを一つ選定して推定しているため，データにオーバーフィッティングしていると考えられる．一方，提案法では各モデルに平均的に重み付けし，すべてのモデルを混合する．このため，従来法はデータ数の小さいところで平均二乗誤差が大きくなるのに対し，提案法は平均二乗誤差をおさえることができたと思われる．

また，混合によるものの平均二乗誤差が n の増加に伴って減少する理由としては，結果(III)に見られたように，真の事後確率がデータ数と共に1に近づくことがあげられる．

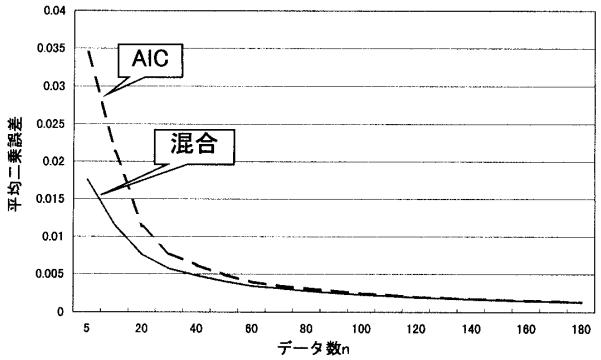

図1 条件1パターン2 平均二乗誤差

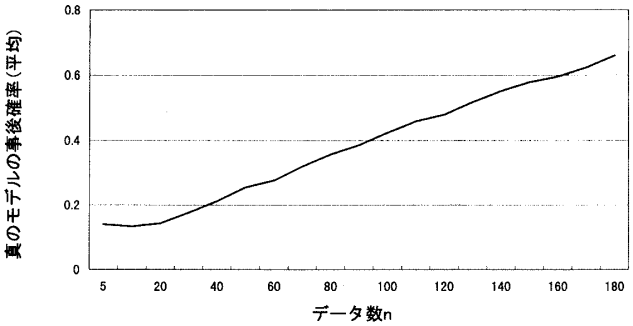

図2 条件1パターン2 真のモデルの事後確率

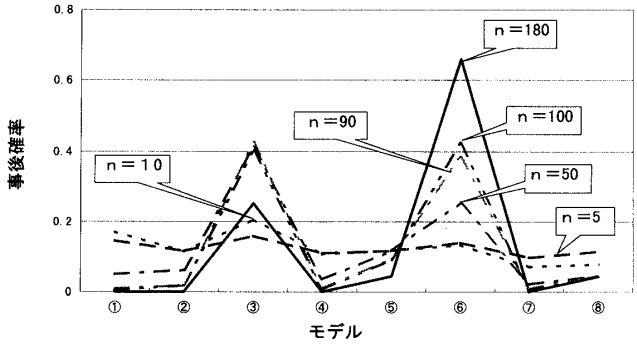

図3 条件1パターン2 データ数別モデルの事後確率

次に，提案法の特性をさらに深く検討するため，現実的かどうかは別として，真のモデルが極端な構造を持っている場合を想定し，次の条件を設定した．

追加シミュレーション実験

〈シミュレーション条件2：真のモデルはモデル1〉

これは，すべてのセルで母数が同じであるという最も単純なモデルが真である場合である．

パターン1：真の値が小さいとき

$p_{11}=p_{12}=p_{13}=p_{21}=p_{22}=p_{23}=0.1$

パターン2：1と0の真の出現確率が同じくらいのとき

$p_{11}=p_{12}=p_{13}=p_{21}=p_{22}=p_{23}=0.4$

パターン3：真の値が大きいとき

$p_1=p_{12}=p_{13}=p_{21}=p_{22}=p_{23}=0.8$

〈シミュレーション条件３：真のモデルはモデル８〉

これは，すべてのセルの母数が異なる値を持つという最も複雑なモデルが真の場合である．

パターン１

$p_{11}=0.2, p_{12}=0.3, p_{13}=0.4$

$p_{21}=0.6, p_{22}=0.7, p_{23}=0.8$

ここでは条件２パターン１，条件３パターン１の結果を示す．

これらの結果で条件１の結果と異なったものをあげると

(IV)　平均二乗誤差は，条件２のパターン１は $n=5$ で，条件３のパターン１は $n=60$ 以上で AIC による推定の方が小さくなるが，その他のパターンではすべて混合による推定の方が小さく，n が少ないときほど顕著な差がみられる．

(V)　真のモデルが複雑であると，真のモデルの事後確率は，n が増えても，他のモデルの事後確率より大きくなりにくい．逆に，簡単なモデルであるほど他のモデルよりも真のモデルの事後確率が大きくなる n は小さいようである．

まず(IV)の結果について考察すると，条件III パターン１では，AIC が提案法より優れている場合が存在するが，これは AIC が複雑なモデルを選択しやすい性質をもつことがそのまま結果に表れたと考えられる．すなわち，真のモデルが一番複雑なモデルであるとき，AIC ではそのモデルを選択する傾向があるた

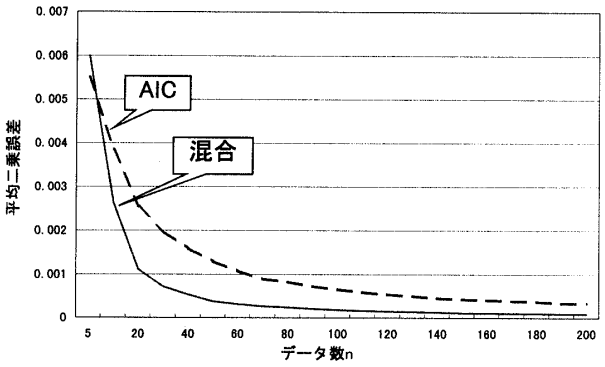

図４　条件２パターン１　平均二乗誤差

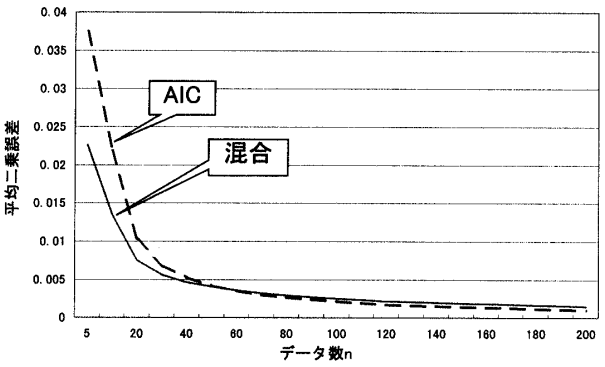

図７　条件３パターン１　平均二乗誤差

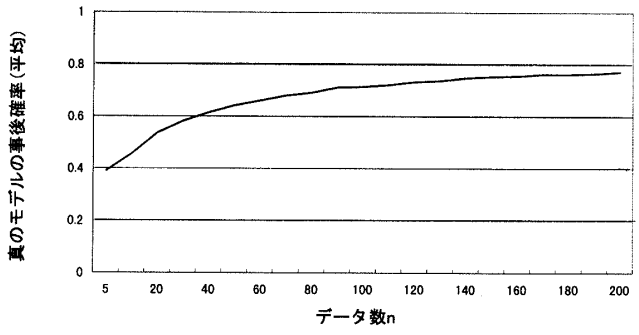

図５　条件２パターン１　真のモデルの事後確率

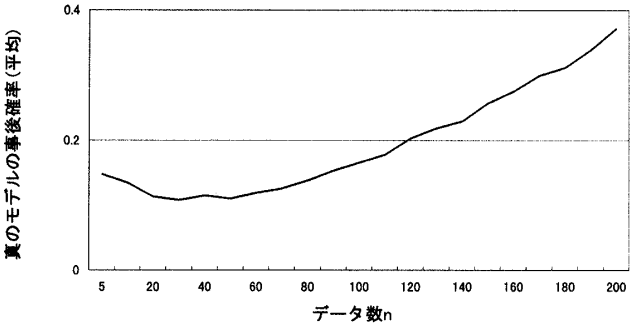

図８　条件３パターン１　真のモデルの事後確率

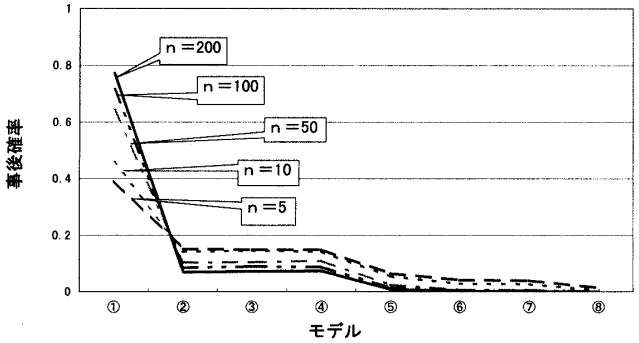

図６　条件２パターン１　データ数別モデルの事後確率

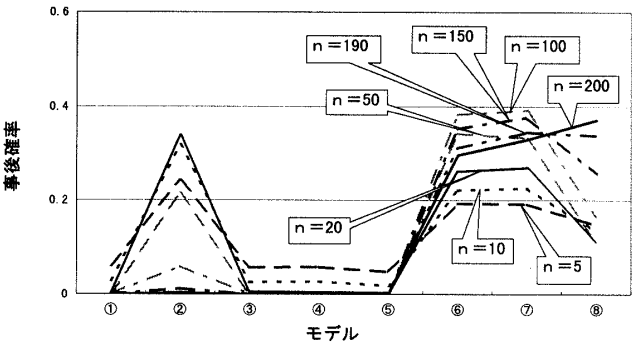

図９　条件３パターン１　データ数別モデルの事後確率

めである．

一方，(V)の結果について考察すると，図5および図8からわかるように，真のモデルが最も次数の低いモデルのときには，データ数が少ない時点ですでに真のモデルの事後確率が大きな割合を占めているのに対し，真のモデルが複雑なときはnに対する真のモデルの事後確率の上昇度が大きくないことから，このような結果が得られたと考えられる．

以上の結果から，真のモデルの構造が極端に単純であったり極端に複雑であるような特殊な場合には，従来法による推定が優れる場合があるといえる．これを定性的に考えると，ベイズ規準は事前分布に対して平均的に最適な推定を行っているので，その分布の裾(端)の部分に対しては推定量を分布の中心方向へ引き戻す性質があるためと思われる．しかし，これらの極端なモデルは現実的には考えにくいケースでもあるので，モデルを一つ選択して推定するという制約から抜けて，提案法のように推定する方法は現実問題で有効となると思われる．また，複雑なモデルが事前に想定できるときには，複雑なモデルの事前分布を高く設定することで，提案法では良いパフォーマンスを示すことは容易に想像できる．

5. ま と め

本研究では隣同士のセルで出現確率が同じとみなせるものをまとめて拡大セルを作り，考え得るすべてのモデルの混合モデルを用いた分割表における予測および出現確率の推定について，ともにベイズ最適な解を示した．さらに，事前分布にベータ分布を仮定した場合の推定に関する計算式を導出し，シミュレーション実験では，提案式は従来法（AIC）と比べて平均二乗誤差の点で優れていることを検証できた．これによって，提案式を用いれば，実務面で有用な情報を与えることができると思われる．

参 考 文 献

〔1〕 児玉寛典, 上坂浩之, 後藤昌司：“順序カテゴリーデータにおける2標本検定”, 品質, Vol. 8, No. 3, pp. 31-39 (1978)

〔2〕 辻谷将明：“順序カテゴリカル・データ解析における数量化法と連関モデル”, 品質, Vol. 21, No. 2, pp. 16-22 (1991)

〔3〕 繁桝算男, 高瀬慎也：“ベイズ的アプローチによる分割表の平滑化”, 応用統計学会誌, Vol. 23, No. 2, pp. 81-93 (1994)

〔4〕 坂元慶行：「カテゴリカルデータのモデル分析」, 共立出版 (1985)

〔5〕 後藤正幸, 松嶋敏泰, 平澤茂一：“ベイズ決定理論に基づく統計的モデル選択について”, 信学技報, IT97-21, pp. 37-42 (1997)

〔6〕 繁桝算男：「ベイズ統計入門」, 東京大学出版会 (1985)

〔7〕 柳川 堯：「離散多変量データの解析」, 共立出版 (1986)

〔8〕 松嶋敏泰：“統計モデル選択の概要”, オペレーションズ・リサーチ, 7月号, pp. 369-374 (1996)

〔9〕 赤池弘次：“AICとMDLとBIC”, オペレーションズ・リサーチ, 7月号, pp. 375-378 (1996)

付 録

(7)式の証明

予測と決定が一致しない誤り確率を$e(y)$とおくと

$$e(y)=\begin{cases}P_{ij}^{mix}(y=0|Z^N), & (Ay=1)\\ P_{ij}^{mix}(y=1|Z^N), & (Ay=0)\end{cases} \tag{20}$$

$e(y)$を最小にする決定Ayは

$$\begin{cases}P_{ij}^{mix}(y=0|Z^N)>P_{ij}^{mix}(y=1|Z^N)\\ \qquad\qquad \text{のとき } Ay=0\\ P_{ij}^{mix}(y=0|Z^N)<P_{ij}^{mix}(y=1|Z^N)\\ \qquad\qquad \text{のとき } Ay=1\end{cases}$$

$P_{ij}^{mix}(y=0|Z^N)=P_{ij}^{mix}(y=1|Z^N)$のときは任意であるが，誤り$e(y)$を等確率にしたいときは0, 1のランダム決定となる．よって

$$P_{ij}^{mix}(y=0|Z^N)=1-P_{ij}^{mix}(y=1|Z^N) \tag{21}$$

から，証明される．

(11)式の証明

BR_{kl}を最小にする決定$A\theta$は(10)式を$A\theta$について偏微分して0とおくことによって求められる．

$$\begin{aligned}0&=\frac{\partial}{\partial A\theta}\left[\sum_m\int_\theta(\theta^2-2\theta(A\theta)+(A\theta)^2)\right.\\&\qquad\left.\cdot P(\theta|Z^N,m)P(m,Z^N)d\theta\right]\\&=-2\sum_m\int_\theta\theta P(\theta|Z^N,m)P(m,Z^N)d\theta\\&\qquad+2(A\theta)\sum_m\int_\theta P(\theta|Z^N,m)P(m,Z^N)d\theta\end{aligned} \tag{22}$$

よって

$$A\theta=\sum_m\int_\theta\theta P(\theta|Z^N,m)P(m,Z^N)d\theta \tag{23}$$