統計数理(2003)
第 51 巻 第 2 号 223–239

特集「個票開示問題の統計理論」
［原著論文］

ミクロデータにおける母集団一意性の事後確率

大森 裕浩†

（受付 2003 年 2 月 6 日；改訂 2003 年 4 月 16 日）

要　　旨

ミクロデータを開示する際には，ある観測値が調査対象となった特定の個人や企業の観測値であると知られるリスクが生じる．このことは官庁統計として得られたミクロデータをそのまま公開することが個人や企業の情報を暴露してプライバシーを侵害する可能性を意味する．実際に調査の結果として得られた標本の観測値は標本一意である(他のどの標本の観測値とも異なる)ことが多いが，それは母集団でも一意であることを必ずしも意味しない．そこで本稿では母集団における一意性の事後確率をミクロデータ開示のリスクとして評価し，これをどのように用いるべきかについて考察する．

キーワード：ミクロデータ，開示リスク，母集団一意性，事後確率，マルコフ連鎖モンテカルロ法．

1. はじめに

近年，官庁統計のミクロデータに対する需要が計量経済分析などの対象として高まりつつある．これまでは経済主体の集計値であるマクロ経済データを用いた分析が主として行われていたが，家計や企業などの経済主体に関するデータが収集されていることから，これら経済主体の行動分析に関心が高まっている．これにともなって欧米ではセンサスデータを始めとしたミクロデータの公開が進んでおり，我が国においてもこれまで厳しく公開を制限されていた官庁統計が少しずつ公開される方向へ進んでいる．公開に際しては調査対象である個人や企業のプライバシーを保護するための配慮が必要であり，そのためのデータ秘匿の方法も数多く提案されている．

一方，ミクロデータ開示にともなうリスクの評価についても提案がなされてきたが，広く用いられている基準はまだ存在しない(Duncan and Lambert (1986, 1989), Paass (1988), Bethlehem et al. (1990), Marsh et al. (1991), Mokken et al. (1992), Willenborg and de Waal (1996))．そこで本稿では比較的解釈のしやすい母集団一意性という概念を用いることにより，ミクロデータの開示によって観測値に含まれる特定の個人に関する情報が暴露されるリスク(識別リスク，identification disclosure risk)の評価を試み，その利用について考察する(識別リスクについてはBethlehem et al. (1990), Lambert (1993), Willenborg and de Waal (1996)を参照)．以下では調査にはカテゴリーデータが多いこと，連続なデータであってもカテゴリーデータに簡単に加工できることを考慮して，カテゴリーデータであるようなミクロデータの開示リスクについて考えていく．

調査の結果得られた標本の観測値が他のどの観測値とも異なる，一意である場合にその観測

† 東京大学大学院 経済学研究科：〒113-0033 東京都文京区本郷 7–3–1; omori@e.u-tokyo.ac.jp

値のことを「標本一意(sample unique)である」というが，この標本一意である観測値が「母集団において一意(population unique)である」かどうかはわからない．もし観測値が母集団一意であれば特定の個人や企業に関する情報が暴露されることになるが，標本一意であることは必ずしも母集団一意である危険を意味するわけではない．従って標本一意であるような観測値が少ない場合には，そのミクロデータを公開しても識別リスクは低いと考えられる．しかし実際のデータでは標本一意である観測値が非常に多く，これを公開すればデータの中に特定の個人や企業と関連づけられるものがないとはいえず，リスクの評価を行うことが必要になる．

そこでもし標本一意である観測値が得られたときにその観測値が母集団一意でもある確率(母集団一意性の事後確率)はどの程度であるかを評価することができれば，それは識別リスクの尺度となりうるはずである．もちろん実際には標本一意であっても観測誤差なども考慮すれば，相異なるカテゴリーデータであっても観測値の類似度は高いと考えられ，また実際に母集団でも似たような個人が多く存在するので母集団一意性の事後確率が示すほど危険ではないと思われる．しかし，標本一意である観測値が，特殊な職業であるとか所得が際だって高いなどという情報をもち，そのような個人や企業が母集団においてもあまり存在しない場合には，母集団一意性の事後確率がそのリスクの目安となるであろう．

以下ではまず第 2 節でカテゴリーの頻度(セル頻度)に関する基本的な確率モデルを紹介し，そのモデルのもとでの母集団一意性の事後確率を試算する．次に第 3 節でカテゴリーの相関構造を考慮したモデルを構成し，これにより事後確率を推定する方法について説明する．最後に第 4 節で結論を述べる．

2. 超母集団モデル

ポアソン=ガンマモデルは数学的に扱いやすいこともあり，カテゴリーデータの確率モデルとしてしばしば用いられている．このモデルでは観測値の個数を N としたとき，第 i 番目のカテゴリー(セルともいう)に属する観測値の頻度は，そのセルの比率 π_i を所与として平均 $N\pi_i$ のポアソン分布に従うとする．ポアソン分布では平均と分散が同じであるが，現実のデータは分散は平均より大きい傾向が見られるため，π_i が互いに独立にガンマ分布に従うと考える．Bethlehem et al. (1990)では，ポアソン=ガンマモデルに基づいて母集団セルの期待頻度を推定し，これを識別リスクの尺度としている．

しかし，このモデルにおいて仮定されるカテゴリー間の独立性や母集団におけるカテゴリー頻度の期待値が等しいことなどは現実のデータに対しては必ずしも適当であるとはいえないため，いくつかの代替的なモデルが提案されている．例えば Skinner and Holmes (1993), Skinner et al. (1994)はポアソン=対数正規モデルを提案し，Takemura (1997)はセルの観測頻度が多項分布でその確率 π_i がディリクレ分布に従うと考えるディリクレ多項分布によってカテゴリー間の相関を考慮したモデルを提案している．この他にも Crescenzi (1993)は母集団におけるセルの頻度が既知であるときに母集団一意性の期待値を推定し，Greenberg and Zayatz (1992)はセルの周辺頻度が超幾何分布に従うとして母集団一意性の事後確率を求めているが，いずれも母集団におけるセル頻度に関する事前情報が存在すると仮定しており，必ずしも現実的であるとはいえない．また Kanoh (1997)では，セル頻度が多変量超幾何分布に従うときに標本一意であるセルのなかに母集団一意であるような観測値が存在するかどうかの尤度比検定を提案している．

この節では Takemura (1997)におけるようにミクロデータの標本は母集団から非復元抽出されるとし，母集団は既知ではない限り確率的に変動すると考えて，ある確率関数で定義された手順により更にその上の母集団(超母集団, superpopulation)から抽出されていると考える．

まず母集団には N 人の個人がおり K 個のセルに属しているとする．また第 i セルの頻度を F_i $(i=1,2,\ldots,K)$ とする．このとき母集団から n 人の標本を非復元抽出すれば，標本において第 i セルに属する人数を f_i $(f_i=0,1,\ldots,F_i)$ として，$f=(f_1,\ldots,f_K)'$ は次のような確率関数を持つ多変量超幾何分布に従う（$(n;f)\sim MH(N;F)$ と表記する）．

$$\Pr(f|F)=\prod_{i=1}^{K}\binom{F_i}{f_i}\Big/\binom{N}{n}\,, \tag{2.1}$$

ただし，

$$N=\sum_{i=1}^{K}F_i\,,\quad n=\sum_{i=1}^{K}f_i\,,\quad F=(F_1,\ldots,F_K)'\,,$$

とする．

2.1 多項分布モデル

もし超母集団における第 i セルの相対頻度 π_i が既知であり，母集団が単純復元抽出によって超母集団から抽出されたと考えれば，F は次のような確率関数をもつ多項分布に従う（$F\sim MN(N;\pi)$ と表記する）．

$$\Pr(F|\pi)=\frac{N!}{F_1!\cdots F_K!}\pi_1^{F_1}\cdots\pi_K^{F_K}\,,\quad \pi=(\pi_1,\ldots,\pi_K)'\,, \tag{2.2}$$

これは式(2.1)における母数 F の事前分布と解釈することができる．式(2.1)(2.2)より，母数 F の事後分布は

$$\Pr(F|\pi,f)=\frac{(N-n)!}{\prod_{j=1}^{K}(F_j-f_j)!}\pi_1^{F_1-f_1}\cdots\pi_K^{F_K-f_K}\,,\quad F_j\geq f_j\,, \tag{2.3}$$

となる．標本のなかに現れなかったセルについては関心がないので，これを除くセルについて考えればよい．従って一般性を失うことなく $f_j\geq 1$ $(j=1,\ldots,K)$ とおくことができる．最初の m 個のセルが標本一意であったとき $(f_1=\cdots=f_m=1,\ m<K)$ 対応する母集団セル頻度 $F_1,\ldots,F_m$ の事後確率は

$$\begin{aligned}&\Pr(F_1,\ldots,F_m|f_1=\cdots=f_m=1)\\&\quad=\frac{(N-n)!}{(N-n-\sum_{j=1}^{m}F_j+m)!\prod_{j=1}^{m}(F_j-1)!}\left(1-\sum_{j=1}^{m}\pi_j\right)^{N-n-\sum_{j=1}^{m}F_j+m}\prod_{j=1}^{m}\pi_j^{F_j-1}\end{aligned}$$

である．実際には π_i は未知であることが多いから，仮に $\pi_i=\pi_0(i=1,\ldots,m)$ とおくと

$$\begin{aligned}&\Pr(F_{i_1},\ldots,F_{i_r}|f_1=\cdots=f_m=1)\\&\quad=\frac{(N-n)!}{(N-n-\sum_{j=1}^{r}F_{i_j}+r)!\prod_{j=1}^{r}(F_{i_j}-1)!}(1-r\pi_0)^{N-n-\sum_{j=1}^{r}F_{i_j}+r}\pi_0^{\sum_{j=1}^{r}F_{i_j}-r}\,,\\&\qquad\qquad\{i_1,\ldots,i_r\}\subseteq\{1,\ldots,m\}\,,\quad 1\leq r\leq m\,,\end{aligned}$$

を得る．従って標本一意であったセルが母集団一意でもある事後確率は

$$\Pr(F_{i_1}=\cdots=F_{i_r}=1|f_1=\cdots=f_m=1)=(1-r\pi_0)^{N-n} \tag{2.4}$$

となる．ここで π_0 の値を $1/n$（つまり標本一意であったセルは母集団から平均して 1 個抽出されたものと考える）としたり，$1/N$（標本一意であったセルの母集団セルは超母集団から平均して 1 個抽出されたものと考える）としたりすることによって式(2.4)を評価することができる．この結果を拡張すると以下の定理を得ることができる（証明は補論を参照）．

定理 2.1. $(n;f)\sim MH(N;F)$, $F\sim MN(N;\pi)$ であるとき，$\pi_1=\cdots=\pi_m=\pi_0$ とする．また $\{i_1,\ldots,i_k\}\subseteq\{1,\ldots,m\}$ とし $\alpha_k=\Pr($ 少なくとも 1 組の $\{i_1,\ldots,i_k\}$ に対して $F_{i_1}=\cdots=F_{i_k}=1|f_1=\cdots=f_m=1)$, $1\le k\le m$ と表記すると

(i)
$$\alpha_k=\sum_{r=k}^{m}\binom{m}{r}\sum_{j=0}^{m-r}(-1)^j\binom{m-r}{j}(1-(r+j)\pi_0)^{N-n}$$

(ii) もし $n(N)$ が十分大きく $\pi_0=1/n(1/N)$ であるならば

$$\alpha_k=1-G(k-1;m,p)\,, \tag{2.5}$$

ただし $G(x;m,p)=\sum_{k=0}^{x}\binom{m}{k}\,p^k(1-p)^{m-k}$ は母数 (m,p) の 2 項分布の分布関数であり，また抽出率を $\theta=n/N$ とすると $p=\exp(1-\theta^{-1})$, $(\exp(\theta-1))$ である．

これを用いれば $k=1$ のときつまり標本一意であるもののなかに母集団一意であるものが少なくとも 1 つある事後確率は次のようになる．

系 2.1. (i) $\alpha_1=\sum_{r=1}^{m}(-1)^{r-1}\,\binom{m}{r}\,(1-r\pi_0)^{N-n}$.

(ii) もし $n(N)$ が十分大きく $\pi_0=1/n(1/N)$ であるならば

$$\alpha_1=1-(1-p)^m\,.$$

この(ii)で $\Pr(F_j=1|f_1=\cdots=f_m=1)=(1-\pi_0)^{N-n}\approx\exp-(N-n)\pi_0=p$ であることに注意すると $\alpha_1=1-\prod_{j=1}^{m}(1-\Pr(F_j=1|f_1=\cdots=f_m=1))$ から $\{F_j=1\}$ $j=1,2,\ldots,m$ が条件付きで独立になっており，m 個の標本一意なセルのうち母集団一意であるセルの期待値は $m(1-\pi_0)^{N-n}$ となる．

さて表 1 は式(2.5)を用いて $m=100$ $(m=10)$ と $\theta=0.1,0.15,0.2$ のときについて α_j を計算した結果である．$\pi_0=1/n$ と考えると 100 個の標本一意な個体が存在したとしても，そのうち少なくともひとつの個体が母集団一意である確率は，標本抽出率が 10%であれば 0.0123 と非常に低い．少なくとも 2 つあるいは 3 つの個体が母集団一意である確率はさらに低くなる．しかし，標本抽出率が 20%になるとその確率は急激に上昇し 0.8425 となる．この傾向は $m=10$ のときでも同様である．式(2.5)で $\alpha_1=0.05$ とおいて m について解いてみると，$\theta=0.1(0.05)$ のとき $m=416(9154945)$ となり 5%抽出はかなり安全である．

一方，$\pi_0=1/N$ と考えるといずれの場合でも母集団一意である事後確率は非常に高い．ただ $\pi_0=1/N$ と考えるのは，標本一意のセルすべてが母集団一意であるという強い事前情報を

表 1. $m=100$ $(m=10)$.

	$\pi_0=1/n$			$\pi_0=1/N$		
θ	α_1	α_2	α_3	α_1	α_2	α_3
0.10	0.0123	0	0	1	1	1
	(0.0012)	(0)	(0)	(0.9898)	(0.9305)	(0.7753)
0.15	0.2929	0.0474	0.0052	1	1	1
	(0.0341)	(0.0005)	(0)	(0.9945)	(0.9575)	(0.8431)
0.20	0.8425	0.5487	0.2774	1	1	1
	(0.1688)	(0.0137)	(0.0007)	(0.9974)	(0.9765)	(0.8997)

用いることになるので，開示リスクの評価としては適切ではない．むしろ母集団に関する情報がない状態では，$\pi_0 = 1/n$ として標本に 1 個あれば平均的には母集団に N/n 個あると考えるのが自然であろう．

2.2 ディリクレ多項分布モデル

これまで π_j は既知であるとした仮定を本節では緩めて，ポアソン=ガンマモデルで独立とされていたカテゴリー間の相関を取り入れるように π_j が多次元分布に従う確率変数であると考える．具体的には以下のように π_j が $\lambda = (\lambda_1, \ldots, \lambda_K)'$ $(\lambda_j > 0)$ を母数とするディリクレ分布に従うとする．

$$f(\pi|\lambda) = \frac{\Gamma(\lambda.)}{\prod_{i=1}^{K}\Gamma(\lambda_j)} \prod_{j=1}^{K} \pi_j^{\lambda_j - 1}, \quad \lambda. = \sum_{i=1}^{K} \lambda_i \,,$$

F と π の同時分布 $\Pr(F|\pi, \lambda) f(\pi|\lambda)$ を π に関して積分すると以下のようなディリクレ多項分布を得る（$F \sim DM(N, \lambda)$ と表記する）．

$$\Pr(F|\lambda) = \frac{N!\Gamma(\lambda.)}{\Gamma(\lambda. + N)} \prod_{i=1}^{K} \frac{\Gamma(\lambda_i + F_i)}{\Gamma(\lambda_i) F_i!} \tag{2.6}$$

この事前分布を用いると式(2.1)(2.6)より F の事後分布は

$$\Pr(F|f) = \frac{(N-n)!\Gamma(\lambda. + n)}{\Gamma(\lambda. + N)} \prod_{i=1}^{K} \frac{\Gamma(\lambda_i + F_i)}{\Gamma(\lambda_i + f_i)(F_i - f_i)!}$$

となる．前節と同様に $f_1 = \cdots = f_m = 1$ であったとすると $(F_{i_1}, \ldots, F_{i_r})$, $(\{i_1, \ldots, i_r\} \subseteq \{1, \ldots, m\}, 1 \le r \le m)$ の周辺事後分布は

$$\begin{aligned}\Pr(F_{i_1}, \ldots, F_{i_r} | f_1 = \cdots = f_m = 1) &= \frac{(N-n)!\Gamma(\lambda. + n)}{\Gamma(\lambda. + N)} \prod_{j=1}^{r} \frac{\Gamma(\lambda_{i_j} + F_{i_j})}{\Gamma(\lambda_{i_j} + 1)(F_{i_j} - 1)!} \\ &\times \frac{\Gamma(\lambda. - \sum_{j=1}^{r}\lambda_{i_j} + N - \sum_{j=1}^{r} F_{i_j})}{\Gamma(\lambda. - \sum_{j=1}^{r}\lambda_{i_j} + n - r)(N - \sum_{j=1}^{r} F_{i_j} - n + r)!}\end{aligned}$$

となる．このことを用いて以下を導くことができる．

補題 2.1. $(n; f) \sim MH(N; F)$, $F \sim DM(N; \lambda)$ とし，$\{i_1, \ldots, i_r\} \subseteq \{1, \ldots, m\}$ $(1 \le r \le m)$ とすると

(i)
$$\begin{aligned}&\Pr(F_{i_1}, \ldots, F_{i_r} | f_1 = \cdots = f_m = 1) \\ &\quad = \prod_{i=1}^{N-n} \left\{ 1 - \left(w_i \frac{\sum_{j=1}^{r}\lambda_{i_j}}{\lambda.} + (1 - w_i)\frac{r}{N-i} \right) \right\}, \quad w_i = \frac{\lambda.}{\lambda. + N - i} \,.\end{aligned}$$

(ii) もし $\lambda.$ が十分大きく，かつ N が小さいならば重みは $w_i \approx 1$ $(i = 1, \ldots, N-n)$ であるから

$$\Pr(F_{i_1}, \ldots, F_{i_r} | f_1 = \cdots = f_m = 1) = \left(1 - \frac{\sum_{j=1}^{r}\lambda_{i_j}}{\lambda.} \right)^{N-n}$$

さらに $\lambda_1 = \cdots = \lambda_m = \lambda_0$ で $\lambda_0/\lambda.$ が小さいならば $\Pr(F_{i_1} = \cdots = F_{i_r} = 1 | f_1 = \cdots = f_m = 1) = p^r$, $p = \exp(N-n)\lambda_0/\lambda.$ となる．

(iii) もし $\lambda.$ が十分小さいならば重みは $w_i \approx 0$ $(i = 1, \ldots, N-n)$ となり

$$\Pr(F_{i_1}, \ldots, F_{i_r} \,|\, f_1 = \cdots = f_m = 1) = \prod_{i=1}^{r} \frac{n-i}{N-i}\,.$$

さらに $r \ll N$ ならば $\Pr(F_{i_1} = \cdots = F_{i_r} = 1 | f_1 = \cdots = f_m = 1) = \theta^r$ $(\theta = n/N)$ となる.

重み w_i は事前情報の重みであると解釈でき, $\lambda. \to 0$ のとき $w_i \to 0$ で $\lambda. \to \infty$ のとき $w_i \to 1$ となる. もし事前情報がほとんどなく $(\lambda. \approx 0)$, N が十分大きい場合には $\Pr(F_j = 1 | f_1 = \cdots = f_m = 1)$ は標本抽出率 θ となり $\{F_{i_j} = 1\}$ は条件付き独立になる. 従ってこのようなときには Bethlehem et al. (1990) のように F_j に独立性を仮定してポアソン=ガンマモデルをあてはめても同じ結果となる. そして母集団一意であるようなセルの個数の期待値は $m(n-1)/(N-1)$ となる. この補題を用いると母集団一意である事後確率が以下のように求められる.

定理 2.2. $(n; f) \sim MH(N; F)$, $F \sim DM(N; \lambda)$ とし, $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, m\}$ とする.
(i) $\lambda.$ が十分大きく N が小さく, かつ $\lambda_1 = \cdots = \lambda_m = \lambda_0$ であるとすると

$$\alpha_k = \sum_{r=k}^{m} \binom{m}{r} \sum_{j=0}^{m-r} (-1)^j \binom{m-r}{j} \left\{1 - \frac{\lambda_0}{\lambda.}(r+j)\right\}^{N-n}$$

である. さらに $\lambda_0/\lambda.$ が小さければ $\alpha_k = 1 - G(k-1; m, p)$, $p = \exp(N-n)\lambda_0/\lambda.$ である.
(ii) $\lambda.$ が十分小さければ

$$\alpha_k = \sum_{r=k}^{m} \binom{m}{r} \sum_{j=0}^{m-r} (-1)^j \binom{m-r}{j} \prod_{i=1}^{r+j} \frac{n-i}{N-i}\,,$$

であり, さらに $m \ll N$ ならば

$$\alpha_k = 1 - G(k-1; m, \theta)\,, \quad \theta = n/N\,, \tag{2.7}$$

である.

もし(i)で $\lambda_0/\lambda. = \pi_0$ とおけば定理 2.1 と同じ結果を得ることができる. また標本抽出率 θ は $\exp(1 - \theta^{-1}) < \theta < \exp(\theta - 1)$ であることから, 定理 2.1 で $1/N < \pi_0 < 1/n$ である π_0 のなかに(ii)を満たすものがある.

系 2.2. (i) $\lambda.$ が十分大きく N が小さいとする. もし $\lambda_1 = \cdots = \lambda_m = \lambda_0$ ならば $\alpha_1 = \sum_{r=1}^{m} (-1)^{r-1} \binom{m}{r} \left(1 - r\frac{\lambda_0}{\lambda.}\right)^{N-n}$. さらに $\lambda_0/\lambda.$ が小さいならば $\alpha_1 = 1 - (1-p)^m$, $p = -\exp(N-n)\lambda_0/\lambda.$ である.
(ii) $\lambda.$ が十分小さいとき $\alpha_1 = \sum_{r=1}^{m} (-1)^{r-1} \binom{m}{r} \prod_{i=1}^{r} \frac{n-i}{N-i}$ であり, さらに $m \ll N$ ならば $\alpha_1 = 1 - (1-\theta)^m$ である.

表 2 は式(2.7)を用いて α_j を $m = 100$ $(m = 10)$ のときに計算した結果である. この結果は多項分布のモデルで $1/n < \pi_0 < 1/N$ であるような π_0 に対して得られる結果に該当するため, $\pi_0 = 1/n$ の場合に比べると厳しい数値になっている. 例えば $m = 100$ のときで標本抽出率 θ を 0.1 ととると $\alpha_1 \approx 1$ とほぼ確率 1 で母集団一意になり, また標本抽出率をさげてもなかなか

表 2. 母集団一意性の事後確率．$m = 100$ $(m = 10)$.

θ	α_1	α_2	α_3
0.001	0.0952	0.0046	0.0002
	(0.0100)	(0)	(0)
0.01	0.6340	0.2642	0.0794
	(0.0956)	(0.0043)	(0.0001)
0.05	0.9941	0.9629	0.8817
	(0.4013)	(0.0861)	(0.0115)
0.1	1	0.9997	0.9981
	(0.6513)	(0.2639)	(0.0702)

表 3. $\alpha_1 = 0.05$ ($\alpha_1 = 0.01$)であるような標本一意の個数.

θ	m
0.001	51 (10)
0.01	5 (1)
0.05	1 (0)
0.1	0 (0)

この確率は下がらない．これは π_j に関する事前情報がばらついたため標本一意であるようなセルの π_i が非常に小さい場合も含めて考慮してしまうからであると考えられる．

表 3 は母集団一意である事後確率が $\alpha_1 = 0.05(0.01)$ であるような標本一意の個数 m を計算した結果である．近似的に $m = \alpha_1/\theta$ が成り立っており，抽出率が 1%であっても標本一意である個数が 5 個となっている．

以上のことからディリクレ多項分布モデルのように π_i に関する情報をあまり漠然とした形で想定すると，すべての π_j に対して同じような事前のウェイトを与えてしまい，結果として得られるリスク評価の意味が曖昧になりやすいということもできる．ミクロデータの開示リスク評価においては，データの性質によって標本一意であることの危険性が異なるので，むしろ多項分布モデルで π_0 の値をいろいろと変えながらリスクを評価することも必要である．しかし，どの場合においても母集団一意である確率は厳密には 0 にはならないので，確実にリスクを避けたい場合にはデータに対して秘匿のための加工を行う必要がある．

本節の結果は標本一意の観測値の母集団一意性についてであったが，標本におけるセル数が 2 個のカテゴリーが，母集団においても頻度が 2 である事後確率を計算することもできる(Omori (1999))．しかし，それらの確率は標本一意である場合の数値よりも低くなるため，リスク評価の基準としては標本一意である観測値の母集団一意性の事後確率が適当である．

3. カテゴリ間の相関構造のモデル

前節でみたように多項分布モデルのパラメータ π_j に関する事前分布のとりかたによって計算されるリスクは変わってしまう．このためミクロデータの開示に慎重な人々からは母集団一意である危険性は高いという指摘がなされて，開示をすすめる人々からは危険性が直観よりも高すぎるという不満が生じたりしている．実はカテゴリ間の相関を考慮しなくても $\pi_0 = 1/n$ と考えれば比較的安全であるし，$\pi_0 = 1/N$ と考えると危険であることになる．相関構造を考えた場合に π_j にディリクレ分布を仮定すると比較的危険であるという結果になるが，これも

表 4. ミクロデータの形式の例.

変数（項目）	カテゴリ
性別	1 : 女性, 2 : 男性
職業	1 : 会社員, 2 : 公務員, 3 : 自営業, ...
質問 A	1 : そう思わない, 2 : どちらともいえない, 3 : そう思う
質問 B	1 : 大変不満, 2 : やや不満, 3 : ふつう, 4 : やや満足, 5 : 大変満足
年齢	$1 : \leq 19, 2 : 20-29, 3 : 30-39, 4 : 40-49, 5 : 50-59, 6 : \geq 60$
収入	$1 : \leq 300, 2 : 300-500, 3 : 500-700, \ldots$

他の仮定をおけばまた変わってくる.

以下ではカテゴリーの相関構造をモデル化し，開示リスクを評価する代替的な方法について考える．カテゴリーの相関構造はミクロデータごとに異なっているので各カテゴリーの特徴を明らかにしたうえでカテゴリー間の関係を考えることが必要である．通常，ミクロデータは質的な変数の項目と量的な変数の項目からなっているが，量的な変数は正確な値で得られる場合もあるが，区間で打ち切られていることが多いのでここでは区間の両端が既知であるような量的データを考慮する.

すべての変数の項目の合計を J 項目とし，質的な変数のうち，順序のない項目は J_1 個, 順序のある項目は J_2 個あるとする．そして量的な変数の項目は J_3 個とし，合計で $J = J_1 + J_2 + J_3$ 個の項目となる．例えば表 4 のようなミクロデータである.

第 i 個体の第 j 項目についての回答を，順序のない質的な変数，順序のある質的な変数, 区間で打ち切られた量的な変数について，それぞれ X_{ij}, Y_{ij}, Z_{ij} とする．すると

$$
\begin{aligned}
X_{ij} &= k, \quad k = 1, \ldots, K_j, \quad j = 1, \ldots, J_1, \\
Y_{ij} &= l, \quad l = 1, \ldots, L_j, \quad j = 1, \ldots, J_2, \\
Z_{ij} &= m, \quad m = 1, \ldots, M_j, \quad j = 1, \ldots, J_3,
\end{aligned}
$$

$Z_{ij} = m$ ということは，変数 W_{ij} が区間 $[a_{j,m-1}, a_{jm})$ で打ち切られているということを表す（区間の両端 $a_{j0}, \ldots, a_{jM_j}$ は既知）．そこで $X_i = (X_{i1}, \ldots, X_{iJ_1})'$, $Y_i = (Y_{i1}, \ldots, Y_{iJ_2})'$, $Z_i = (Z_{i1}, \ldots, Z_{iJ_3})'$, とすると，第 i 個体の回答は観測値 $(X_i', Y_i', Z_i')'$, で表現される．このとき同時確率密度関数は

$$
f(x_i, y_i, z_i) = f(x_i) g(y_i|x_i) h(z_i|x_i, y_i)
$$

となる.

ミクロデータの開示リスク評価は，これに基づいて得られたパラメータの事後分布から標本一意なサンプルについて行う．まず標本一意なサンプルが (x_0, y_0, z_0) として与えられたときの確率 $f(x_0, y_0, z_0|x, y, z)$ の推定値を求め，母集団の総数を N としたとき標本に含まれなかった観測値が同じセルになる頻度の期待値を $(N-n) f(x_0, y_0, z_0|x, y, z)$ として求める．また標本一意かつ母集団一意でもある事後確率は $\{1 - f(x_0, y_0, z_0|x, y, z)\}^{N-n}$ と考えればよい.

3.1 順序のない質的変数のモデル

順序のない質的変数の項目のモデルとして一般的にすべてのカテゴリの組み合わせを考えるとパラメータ数が非常に大きくなってしまい，推定が困難になる．そこで以下ではパラメータを節約した多変量多項ロジット・プロビットモデルを基礎として考える．多項ロジットモデル

の場合，

$$\log \frac{f(x_{ij})}{1-f(x_{ij})} = \sum_{k=1}^{K_j-1} \alpha_{jk} D_{ijk} = \alpha_j' D_{ij} \,,$$

ただし，D はダミー変数で

$$D_{ijk} = \begin{cases} 1 & x_{ij} = k \text{ のとき}, \\ 0 & \text{それ以外のとき}, \end{cases}$$

であり，$\alpha_j' = (\alpha_{j1}, \ldots, \alpha_{jK_j-1})$, $D_{ij} = (D_{ij1}, \ldots, D_{ijK_j-1})'$, $j = 1, \ldots, J_1$, である．さらに $\alpha = (\alpha_1', \ldots, \alpha_{J_1}')$ とおき，その事前分布として多変量正規分布を仮定し，項目間の相関を共分散を通じて考慮する．たとえば

$$\alpha \sim N(0, \Sigma) \,, \quad \Sigma^{-1} \sim W(\nu_0, H_0)$$

として Σ^{-1} の事前分布には Wishart 分布を仮定する．

$$\pi(\Sigma^{-1}) \propto \left|\Sigma^{-1}\right|^{(\nu_0 - J_1 - 1)/2} \exp -\frac{1}{2} \mathrm{tr} H_0^{-1} \Sigma^{-1} \,.$$

一方, 多項プロビットモデルを仮定する場合には項目ごとに潜在変数 $U_{ij} = (U_{ij1}, \ldots, U_{ij,K_j-1})'$ $j = 1, \ldots, J_1$ を考えて

$$\Pr(X_{ij} = k) = \Pr\left(\max_{k'} U_{ijk'} = U_{ijk} > 0 \right)$$

と定義する．U_{ij} は多変量正規分布に従い，$U_{ij} \sim N(\alpha_j, I)$ とし，さらに α については多項ロジットのように多変量正規分布を想定する．ここでは α が与えられたもとでは各項目は独立であると考えているが，これらに相関構造をさらに考えて拡張することもできる．例えば $U_{ij} \sim N(\alpha_j, \Sigma)$ とすることもできるが(識別のため Σ の(1,1)要素は 1 とおく必要がある)，計算は複雑になる．その場合，逐次プロビットモデルに相関構造を考慮したモデルを利用することも考えられる．

もしカテゴリーのセル頻度が滑らかに変化する構造が想定されるならば，パラメータ α_j について平滑化する事前分布の重みを強くすることにより，事後確率を平滑化し母集団一意の事後確率を低くすることができる．しかし，平滑化の仮定を正当化できない限りそうした事前分布を機械的に用いることは無意味である．例えば性別というアイテムに属するカテゴリーである男性・女性のセル頻度を考えてみると，母集団の男女比率は必ずしもいつも 50%ずつではない．母集団が化粧品の購入者となれば女性の比率が圧倒的に高くなり，わずかな男性の比率を平滑化によって大きくすることはデータの構造を歪曲することになってしまう．従って母集団一意性が特に問題となるようなセルについては平滑化の仮定がそもそも適当でない可能性が高く，事前分布の仮定には注意が必要である．

推定の方法は潜在変数が多いためマルコフ連鎖モンテカルロ法を用いて行う．多項ロジットモデルの場合，パラメータ α_j の事後分布は正規分布ではないので，対数事後分布をモード $\hat{\alpha}_j$ のまわりでテーラー展開することにより正規分布で近似して Metropolis-Hastings アルゴリズムを適用すればよい．$\alpha_{\backslash j}$ を α_j を除く α の要素とすると

$$\alpha_j | \Sigma, \alpha_{\backslash j} \sim N(m_j, S_j) \,, \quad m_j = \Sigma_{11} \Sigma_{22}^{-1} \alpha_{\backslash j} \,, \quad S_j = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \,,$$

$(\Sigma_{11} = E(\alpha_j \alpha_j'), \Sigma_{12} = \Sigma_{21}' = E(\alpha_j \alpha_{\backslash j}), \Sigma_{22} = E(\alpha_{\backslash j} \alpha_{\backslash j}'))$ であるから，事後分布は

$$f(\alpha_j | x, \Sigma, \alpha_{\backslash j}) \propto \prod_{i=1}^{n} \frac{\exp(\alpha_j' D_{ij})}{1 + \exp(\alpha_j' 1_{K_j - 1})} \times \frac{1}{2} \exp -(\alpha_j - m_j)' S_j^{-1} (\alpha_j - m_j)$$

となり

$$\log f(\alpha_j|x,\Sigma,\alpha_{\setminus j}) = \mathrm{const} + \alpha_j' \sum_{i=1}^{n} D_{ij} - n\log\{1+\exp(\alpha_j' 1_{K_j-1})\} - \frac{1}{2}(\alpha_j - m_j)' S_j^{-1}(\alpha_j - m_j)$$

を 2 次の項までテーラー展開することにより近似分布の提案分布として

$$\mu^* = \Sigma^* \left\{ \sum_{i=1}^{n} D_{ij} - \frac{n\exp(\hat{\alpha}_j' 1_{K_j-1})}{1+\exp(\hat{\alpha}_j' 1_{K_j-1})} 1_{K_j-1} + \frac{n\hat{\alpha}_j' 1_{K_j-1}\exp(\hat{\alpha}_j' 1_{K_j-1})}{\{1+\exp(\hat{\alpha}_j' 1_{K_j-1})\}^2} 1_{K_j-1} + S_j^{-1} m_j \right\},$$

$$\Sigma^* = \left\{ \frac{n\exp(\hat{\alpha}_j' 1_{K_j-1})}{\{1+\exp(\hat{\alpha}_j' 1_{K_j-1})\}^2} 1_{K_j-1} 1_{K_j-1}' + S_j^{-1} \right\}^{-1},$$

を平均，分散共分散行列とする $N(\mu^*, \Sigma^*)$ を用いればよい．Σ^{-1} の事後分布は Wishart 分布で

$$\Sigma^{-1}|x,\alpha \sim W(\nu_0 + 1, \{H_0^{-1} + \alpha\alpha'\}^{-1})$$

となる．

同様に多項プロビットモデルを仮定する場合にも項目ごとに考える．まず空間 R_j を

$$R_j = \begin{cases} \{U_{ij} : u_{ijx_{ij}} = \max_k u_{ijk} > 0\}, & x_{ij} \neq K_j \\ \{U_{ij} : u_{ijx_{ij}} < 0\}, & x_{ij} = K_j \end{cases}$$

とおくと U_{ij} の α_j が与えられたときの条件付き事後分布は切断正規分布 $U_{ij}|\alpha_j, x_{ij} \sim TN_{R_j}(\alpha_j, I)$ である(多変量切断正規分布からのサンプリングは例えば Geweke (1991)参照)．また α_j の U(ただし $U' = (U_1', \ldots, U_n'), U_i' = (U_{i1}, \ldots, U_{iJ})$) が与えられたときの条件付き事後分布は

$$\alpha\,|\,U,\Sigma \sim N(\mu^*, \Sigma^*), \quad \mu^* = \Sigma^* \sum_{i=1}^{n} U_i, \quad \Sigma^* = \left(\Sigma^{-1} + nI\right)^{-1}$$

である．Σ のサンプリングは多項ロジットモデルの場合と同様である．

3.2 順序のある質的変数のモデル

次に，順序のある変数の項目について多変量順序プロビットモデルを用いてカテゴリの構造をモデル化する．ここでは順序のない質的変数に関する回答 x は与えられたものとし，条件つき分布 $g(y|x)$ について考える．回答 y_i は順序尺度であるから，潜在変数 $V_i = (V_{i1}, \ldots, V_{iJ_2})'$ を考えて

$$Y_{ij}|x_i, v_{ij} = l, \quad \text{if } \gamma_{j,l-1} \leq v_{ij} < \gamma_{jl}, \quad j = 1, \ldots, J_2,$$

ただし，$-\infty = \gamma_{j0} \leq \gamma_{j1} \leq \cdots \leq \gamma_{j,L_j-1} \leq \gamma_{j,L_j} = \infty$ とする．パラメータを識別可能とするために $\gamma_{j1} = 0$ とおき，$\gamma = (\gamma_1', \ldots, \gamma_{J_2}'), \gamma_j = (\gamma_{j2}, \ldots, \gamma_{j,L_j-1})'$ と表記する．また潜在変数の分布は x が与えられたとき

$$V_i|\gamma, x_i \sim N(A_i\beta, \Omega),$$

であるとする．$A_i\beta$ は回帰部分で第 i 個体の特徴を表し，$A_i = \mathrm{diag}\{A_{i1}', \ldots, A_{iJ_2}'\}, A_{ij}' = (1, D_{ij}'), \beta' = (\beta_1', \ldots, \beta_{J_2}')$ とする．また識別性のために分散共分散行列 Ω の対角要素は 1 とおく(従って Ω は相関行列)と

$$E(V_{ij}|\gamma, x_i) = A_{ij}'\beta_j = \beta_{j0} + \sum_{h=1}^{J_1} \sum_{k=1}^{K_h-1} \beta_{j,hk} D_{ihk}$$

となる．これを用いれば

$$
\begin{aligned}
&g(y_i|x_i,\beta,\gamma,\Omega)\\
&\quad=\int_{R_i}(2\pi)^{-n/2}|\Omega|^{-n/2}\exp-\frac{1}{2}(v_i-A_i\beta)'\Omega^{-1}(v_i-A_i\beta)dv_i\,,\\
&\quad R_i=(\gamma_{1,y_{i1}-1},\gamma_{1,y_{i1}}]\times\ldots\times(\gamma_{J_2,y_{iJ_2}-1},\gamma_{J_2,y_{iJ_2}}]\,,
\end{aligned}
$$

と表現できる．潜在変数を導入することによりパラメータのサンプリングが容易になるが，相関行列 Ω のサンプリングが難しくなるため，以下のように Nandram and Chen (1996)の変数変換を行う．Ω の (i,j) 要素を ω_{ij} とおき

$$
\xi_j=1/\gamma_{jL_j-1}\,,\quad \gamma^*_{jk}=\xi_j\gamma_{jk}\,,\quad \beta^*_j=\xi_j\beta_j\,,\quad v^*_{ij}=\xi_jv_{ij}\,,\quad \omega^*_{ij}=\xi_i\xi_j\omega_{ij}\,,\\
j=1,\ldots,J_2,\quad k=2,\ldots,L_j,
$$

(従って $\xi_j=\sqrt{\omega^*_{jj}}$)と変換すると

$$
\begin{aligned}
&Y_{ij}\,|\,x_i,v^*_{ij}=l\,,\quad \text{if}\quad \gamma^*_{j,l-1}\le v^*_{ij}<\gamma^*_{jl}\,,\quad j=1,\ldots,J_2\,,\\
&V^*_i\,|\,\gamma^*,x_i\sim N(A_i\beta^*,\Omega^*)\,,\\
&\qquad -\infty=\gamma^*_{j0}\le 0\le\gamma^*_{j2}\le\cdots\le\gamma^*_{jL_j-2}\le\gamma^*_{jL_j-1}=1\le\gamma^*_{jL_j}=\infty\,,
\end{aligned}
$$

となる．これにより Ω^* の事前分布に $\Omega^{*-1}\sim W(\nu_1,H_1)$ とおくと事後分布も Wishart 分布から簡単にサンプリングできるようになる．特に $\gamma^*=(\gamma_1^{*'},\ldots,\gamma_{J_2}^{*'})'$, $\gamma^*_j=(\gamma^*_{j2},\ldots,\gamma^*_{jL_j-2})'$ が与えられたとき，β^* の事前分布を $N(\beta^*_0,\Omega^*_0)$ とするならば

- $\Omega^{*-1}|x,v^*,\beta^*,\gamma^*\sim W\left(\nu_1+n,\left\{H_1^{-1}+\sum_{i=1}^n(v^*_i-A_i\beta^*)(v^*_i-A_i\beta^*)'\right\}^{-1}\right)$
- $\beta^*|x,v^*,\Omega^*,\gamma^*\sim N(\hat\beta^*,\Omega^*_1)$ ただし

$$
\hat\beta^*=\Omega^*_1\left(\Omega^{*-1}_0\beta^*_0+\sum_{i=1}^nA_i'\Omega^{*-1}v^*_i\right)\,,\quad \Omega^*_1=\left(\Omega^{*-1}_0+\sum_{i=1}^nA_i'\Omega^{*-1}A_i\right)^{-1}\,.
$$

- $v^*_i|x,\beta^*,\Omega^*,\gamma^*\sim TN_{R_i}(A_i\beta^*,\Omega^*)$ ただし $R_i=\left(\gamma^*_{1,y_{i1}-1},\gamma^*_{1,y_{i1}}\right]\times\ldots\times\left(\gamma^*_{J_2,y_{iJ_2}-1},\gamma^*_{J_2,y_{iJ_2}}\right]$.

である．

γ^* のサンプリングには Albert and Chib (1993), Nandram and Chen (1996), Chen and Dey (1996)による方法があるが，Albert and Chib (1993)は v^*_i と $\gamma^*_{\setminus jt}$ で両端が決まる一様分布を用いるので得られたサンプルの相関が高くなり収束が遅くなりやすい(特に $n\ge 50$ のとき)．Nandram and Chen (1996)はディリクレ分布を近似分布の提案分布として Metropolis-Hastings アルゴリズムを行うが，あるセルの度数が 0 に近くなるような場合にはよい提案分布として機能しない(Chen et al. (2000))．このため Chen and Dey (1996)に従って次のような変数変換を行う．

$$
\gamma^*_{jl}=\frac{\gamma^*_{j,l-1}+\exp(\zeta_{jl})}{1+\exp(\zeta_{jl})}\,,\quad l=2,\ldots,L_j-2\,,
$$

$j=1,\ldots,J_2$. そして v^*_i を上のように発生させるのではなく

(1) まず $\zeta_{j2},\ldots,\zeta_{jL_j-2}|v^*_{ik},\gamma^*_k(k\ne j),x,\beta^*,\Omega^*$ を発生させる．
(2) 次に $v^*_{ij}|v^*_{ik}(k\ne j),\gamma^*,x,\beta^*,\Omega^*$ $(i=1,\ldots,n)$ を発生させる．

というように γ_j^* と v_{ij}^* $(i=1,\ldots,n)$ を同時にサンプリングする．v_{ik}^* $(k\neq j)$ が与えられたときの v_{ij}^* の分布を $N(m_{ij},s_{ij}^2)$ $(i=1,\ldots,n)$ とする (各個人の v_i^* の同時分布における第 j 変数の条件付き分布)と条件つき周辺尤度は

$$\prod_{i=1}^{n}\prod_{t=1}^{L_j-2}\left\{\Phi\left(\frac{\gamma_{jt}^*-m_{ij}}{s_{ij}}\right)-\Phi\left(\frac{\gamma_{j,t-1}^*-m_{ij}}{s_{ij}}\right)\right\}$$

であるから,(1)ではこれをモード $\hat{\zeta}_j$ のまわりでテーラー展開して正規分布で近似し, Metropolis-Hastings アルゴリズムを用いて ζ_j をサンプリングする．ζ_j (従って γ_j^*)が得られたら(2)では v_{ij}^* を切断正規分布 $TN_{(\gamma_{j,y_{ij}-1}^*,\gamma_{j,y_{ij}}^*]}(m_j,s_{ij}^2)$ からサンプリングする．

3.3 区間で打ち切られた量的変数のモデル

最後に，区間で打ち切られた量的変数の項目に関するモデルを考える．ここでは打ち切られる前の変数 w_{ij} に対して一般化線形モデルを仮定し指数型分布族の密度関数をもつとする．区間の両端 a_{jm}'s の値は既知であり

$$\begin{aligned}
h(z_{ij}=m\,|\,x,y,\eta_{ij},\phi_j)&=\int_{a_{j,m-1}}^{a_{jm}}p(w_{ij}|x,y,\eta_{ij},\phi_j)dw_{ij}\,,\\
p(w_{ij}\,|\,x,y,\eta_{ij},\phi_j)&=\exp\{[w_{ij}\theta_{ij}-a(\theta_{ij})+b(w_{ij})]/\phi_j\}\\
\mu_{ij}&=E(w_{ij}\,|\,x,y,\eta_{ij},\phi_j)=a'(\theta_{ij})\,,\\
\tau_{ij}&=\mathrm{var}(w_{ij}\,|\,x,y,\eta_{ij},\phi_j)=a''(\theta_{ij})\,,\\
q(\mu_{ij})&=\eta_{ij}\,,\quad \tau_{ij}=r(\mu_{ij})\phi_j\,,\\
&\qquad j=1,\ldots,J_3\,,
\end{aligned}$$

となる．ここで q,r はそれぞれ既知のリンク及び分散関数である．また連続変数の項目間には相関関係があるので，$w_i=(w_{i1},\ldots,w_{iJ_3})'$ の多変量分布を考える必要があるが，ここでは母数 η_i が与えられた時には w_{ij} は条件付きで独立とし，η_i を通じて階層的に相関構造を考えていく．そこで

$$\eta_i=B_i\delta\,,\quad \delta\sim\ N(\delta_0,Q)\,,\quad \eta_i=(\eta_{i1},\ldots,\eta_{iJ_3})'\,,\quad i=1,\ldots,n\,,$$

と仮定する．$B_i\delta$ は回帰部分であり $B_i=\mathrm{diag}\{B_{i1}',\ldots,B_{iJ_3}'\}$, $B_{ij}'=(1,D_{ij}',D_{ij}^{*'})$, $j=1,\ldots,J_2$, $\delta'=(\delta_1',\ldots,\delta_{J_3}')$ とする．ここでは順序のある質的変数の潜在変数 v を説明変数とするとサンプリングが難しくなるため，順序のない質的変数と同様にして定義されるダミー変数 D_{ij}^* を用いることとする．従って

$$\eta_{ij}=B_{ij}'\delta_j=\delta_{j0}+\sum_{h=1}^{J_1}\sum_{k=1}^{K_h-1}\delta_{j,hk}D_{ihk}+\sum_{h=1}^{J_2}\sum_{k=1}^{K_h-1}\psi_{j,kh}D_{ihk}^*\,,$$

と仮定する．分散のパラメータ ϕ_j が未知の場合は事前分布としてガンマ分布を考える．するとパラメータのサンプリングは次のようになる．

- $w_{ij}\,|\,\delta,Q,\phi$ は，区間 $(a_{j,m-1},a_{jm}]$ で切断された指数型分布族に従って発生させる．
- $Q|\delta,w,\phi.$　Q の事前分布を $Q^{-1}\sim W(\nu_2,H_2)$ とすれば事後分布は

$$Q^{-1}\,|\,\delta,w,\phi\sim W(\nu_2+1,\{H_2^{-1}+(\delta-\delta_0)(\delta-\delta_0)'\}^{-1})\,.$$

- $\phi_j|\delta, Q, w_{ij}$. ϕ_j^{-1} の事前分布を $\mathrm{Gamma}(\varphi_{0j}, \varphi_{1j})$ とすれば

$$\phi_j^{-1} \sim \mathrm{Gamma}\left(\varphi_{0j}+n, \left\{\varphi_{1j}^{-1}+\sum_{i=1}^{n} w_{ij}\theta_{ij}-a(\theta_{ij})+b(w_{ij})\right\}^{-1}\right).$$

また回帰係数 δ の事後分布 $\delta|Q, w, \phi$ では 尤度関数をテーラー展開することにより得られる正規分布を近似する提案分布として Metropolis-Hastings アルゴリズムによりサンプリングする．Q, w, ϕ が与えられたときの対数尤度関数 $l(\delta)$ は定数項部分と δ_0 の事前分布に関する部分を除いて

$$\begin{aligned} l(\delta) = & \sum_{i=1}^{n}\sum_{j=1}^{J_3} \phi_j^{-1}\{w_{ij}\theta_{ij}-a(\theta_{ij})\} \\ \approx & \, l(\hat{\delta}) + \sum_{i=1}^{n}\sum_{j=1}^{J_3}(\delta_j-\hat{\delta}_j)' l'_{ij}(\hat{\delta}_j) + \frac{1}{2}(\delta_j-\hat{\delta}_j)' C_{ij}(\hat{\delta}_j)(\delta_j-\hat{\delta}_j) \\ = & \, l(\hat{\delta}) + \sum_{i=1}^{n}\sum_{j=1}^{J_3}\{B'_{ij}(\delta_j-\hat{\delta}_j)\}\frac{w_{ij}-\hat{\mu}_{ij}}{\phi_j a''(\hat{\theta}_{ij})q'(\hat{\mu}_{ij})} - \frac{\{B'_{ij}(\delta_j-\hat{\delta}_j)\}^2}{2\phi_j a''(\hat{\theta}_{ij})\{q'(\hat{\mu}_{ij})\}^2} \end{aligned}$$

ただし

$$\begin{aligned} l'_{ij}(\delta_j) &= \frac{\partial l_i(\delta)}{\partial \delta_j} = \frac{w_{ij}-\mu_{ij}}{\phi_j a''(\theta_{ij})q'(\mu_{ij})} B_{ij}\,, \\ C_{ij}(\delta_j) &= E\left[\frac{\partial^2 l_i(\delta)}{\partial \delta_j \partial \delta'_j}\right] = -\frac{1}{\phi_j a''(\theta_{ij})\{q'(\mu_{ij})\}^2} B_{ij}B'_{ij}\,, \end{aligned}$$

であるから，

$$w^*_{ij} \equiv B'_{ij}\hat{\delta}_j + (w_{ij}-\hat{\mu}_{ij})q'(\hat{\mu}_{ij})$$

と定義し $P_{ij} = \phi_j a''(\hat{\theta}_{ij})(q'(\hat{\mu}_{ij}))^2$ として w の分布を以下のような w^* の正規分布で近似する．δ_j の事前分布の条件付き分布を $\delta_j \sim N(m_j, Q_j)$ とおくと

$$w^*_{ij} = B'_{ij}\delta_j + \epsilon_{ij}\,, \quad \epsilon_{ij} \sim N(0, P_{ij})\,,$$

すると近似する分布のモードは

$$\hat{\delta}_j = \left\{\sum_{i=1}^{n} B_{ij}P_{ij}^{-1}B'_{ij} + Q_j^{-1}\right\}^{-1}\left\{\sum_{i=1}^{n} B_{ij}P_{ij}^{-1}w^*_{ij} + Q_j^{-1}m_j\right\},$$

となり，その分散共分散行列は

$$\Sigma(\hat{\delta}_j) = \left\{\sum_{i=1}^{n} B_{ij}P_{ij}^{-1}B'_{ij} + Q_j^{-1}\right\}^{-1}$$

となるので δ_j の近似分布として $N(\hat{\delta}_j, \Sigma(\hat{\delta}_j))$ を用いればよい．実際に $\hat{\delta}$ を求めるには，適当な初期値から始めて P, w^* を評価して $\hat{\delta}$ を求め，その値を利用して P, w^* を評価して $\hat{\delta}$ を求めるということを数回繰り返せばよい．なぜならそのプロセスは最尤推定量を求めるためのスコアリング法になっているからである．また近似分布では厳密なモードが必要とされているわけではないので，近似分布を得る目的としては数回程度の繰り返しで十分である．

4. 結論

本稿ではミクロデータにおいて数多く観察される標本一意である観測値がミクロデータの開示リスクとなるかどうかについて，母集団一意性の事後確率を用いて評価する方法について提案した．まずカテゴリ間に詳しい相関構造を考えない場合，多項分布やディリクレ多項分布を用いて事後確率を試算した．特に事前分布に構造を考えることができない場合には，標本一意であるようなカテゴリーのセルの頻度は母集団では N/n 程度(標本抽出率の逆数)であるという事前分布を用いると母集団一意である事後確率は比較的低いことがわかった．しかし，標本一意であるようなセルは母集団一意であるとする事前分布を用いたり，母集団のセル頻度が大きくばらつく事前分布を用いたりするとその事後確率は必要以上に高くなってしまう傾向があり，事前分布の選択はミクロデータの性質によって慎重に行わなければならない．

一方，ミクロデータの構造を詳細に反映したい場合には相関構造をモデル化する必要があり，多項ロジット・プロビットモデル，多変量順序プロビットモデル，切断正規分布モデルなどを組み合わせて推定する必要がある．推定の方法は潜在変数が多いためにマルコフ連鎖モンテカルロ法によるが，計算の負荷は大きくならざるを得ない．またいろいろな現実のデータに適用し，その妥当性について検討を重ねていくことも今後の課題である．

本稿ではどのセルについても同じく重要であると考えて議論を進めたが，実際の応用においてその仮定は常に適当であるとは限らない．例えば職業が俳優で住所がある地域であるという特定のセルについてのリスクが実際には問題になると考えられるため，単純に標本一意であるようなセルの個数を考えるのではなく，プライバシーの侵害につながるような標本一意のセルの個数を用いて本稿で提案した開示リスクの評価を行うといった柔軟な適用も必要である．

補 論

定理 2.1 の証明.

(i) $\Pr(\text{少なくとも一組の } \{i_1,\ldots,i_k\} \text{ について } F_{i_1}=\cdots=F_{i_k}=1 | f_1=\cdots=f_m=1)$

$$
\begin{aligned}
&= \sum_{r=k}^{m} \sum_{i_1<\cdots<i_r} \Pr(F_{i_1}=\cdots=F_{i_r}=1, F_i\neq 1, i\notin\{i_1,\ldots,i_r\} | f_1=\cdots=f_m=1)\\
&= \sum_{r=k}^{m} \binom{m}{r} \Pr(F_1=\cdots=F_r=1, F_i\neq 1, i=r+1,\ldots,m | f_1=\cdots=f_m=1)
\end{aligned}
$$

であることから

$$
\begin{aligned}
&\Pr(F_1=\cdots=F_r=1, F_i\neq 1, i=r+1,\ldots,m | f_1=\cdots=f_m=1)\\
&\quad= \Pr(F_1=\cdots=F_r=1 | f_1=\cdots=f_m=1)\\
&\qquad- \sum_{i=r+1}^{m} \Pr(F_1=\cdots=F_r=F_i=1 | f_1=\cdots=f_m=1)\\
&\qquad+\cdots+(-1)^{m-r}\Pr(F_1=\cdots=F_m=1 | f_1=\cdots=f_m=1)\\
&\quad= \sum_{j=0}^{m-r}(-1)^j \binom{m-r}{j} (1-(r+j)\pi_0)^{N-n}
\end{aligned}
$$

より結果を得る．

(ii) n が十分に大きく $\pi_0=1/n$ であるとき $\Pr(F_1=\cdots=F_s=1 | f_1=\cdots=f_m=1)=\left(1-\frac{s}{n}\right)^{n(\theta^{-1}-1)}\approx p^s$, $p=\exp(1-\theta^{-1}), \theta=n/N$ である．同様に N が十分に大きく

$\pi_0 = 1/N$ であるとき，この確率は $p^s, p = \exp(\theta - 1)$ となる．従って $(1-(r+j)\pi)^{N-n} = p^{r+j}$ を用いると $\Pr(F_1 = \cdots = F_r = 1, F_i \neq 1, i = r+1, \ldots, m | f_1 = \cdots = f_m = 1) = \sum_{j=0}^{m-r}(-1)^j \binom{m-r}{j} p^{r+j} = p^r(1-p)^{m-r}$ となるから結果を得る． □

定理 2.2 の証明. (i) $\Pr(F_1 = \cdots = F_r = 1, F_i \neq 1, i = r+1, \ldots, m | f_1 = \cdots = f_m = 1) = \sum_{j=0}^{m-r}(-1)^j \binom{m-r}{j} \left(1 - \frac{\lambda_0}{\lambda_\bullet}(r+j)\right)^{N-n}$ より最初の等号を得る．もし $\lambda_0/\lambda_\bullet$ が小さいならば補題 2.1 の(ii)によりその確率は $\sum_{j=0}^{m-r}(-1)^j \binom{m-r}{j} p^{r+j} = p^r(1-p)^{m-r}$, $p = \exp -\frac{\lambda_0}{\lambda_\bullet}(N-n)$ であるから結果を得る．

(ii) $\Pr(F_1 = \cdots = F_r = 1, F_i \neq 1, i = r+1, \ldots, m | f_1 = \cdots = f_m = 1) = \sum_{j=0}^{m-r}(-1)^j \binom{m-r}{j} \prod_{i=1}^{r+j} \frac{n-i}{N-i}$, より最初の等号を得る．$N$ が十分大きければ $\prod_{i=1}^{r+j} \frac{n-i}{N-i} = \theta^{r+j}$ となり同様に結果を得ることができる． □

系 2.1 の証明． $\Pr(F_1 = \cdots = F_r = 1 | f_1 = \cdots = f_m = 1) = (1 - r\pi_0)^{N-n}$ および $\Pr(\text{ある } j \text{ について } F_j = 1 | f_1 = \cdots = f_m = 1) = \sum_{i=1}^{m} \Pr(F_i = 1 | f_1 = \cdots = f_m = 1) - \sum_{i<j} \Pr(F_i = 1, F_j = 1 | f_1 = \cdots = f_m = 1) + \cdots + (-1)^{m-1} \Pr(F_1 = \cdots = F_m = 1 | f_1 = \cdots = f_m = 1)$ より結果を得る． □

参 考 文 献

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotmous response data, *J. Amer. Statist. Assoc.*, **88**, 669–679.

Bethlehem, J. G., Keller, W. J. and Pannekoek, J. (1990). Disclosure control of microdata, *J. Amer. Statist. Assoc.*, **85**, 38–45.

Chen, M.-H. and Dey, D. K. (1996). Bayesian analysis and computation for correlated ordinal data, Tech. Report, Department of Statistics, University of Connecticut, Storrs, Connecticut.

Chen, M.-H., Shao, Q.-M. and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*, Springer-Verlag, New York.

Crescenzi, F. (1993). On estimating population uniques: Methodological proposals and applications on Italian census data, *Proceedings of the International Seminar on Statistical Confidentiality*, Dublin.

Duncan, G. T. and Lambert, D. (1986). Disclosure-limited data dissemination, *J. Amer. Statist. Assoc.*, **81**, 10–28.

Duncan, G. T. and Lambert, D. (1989). The risk of disclosure for microdata, *J. Bus. Econom. Statist.*, **7**, 207–217.

Geweke, J. (1991). Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints, *Computing Science and Statistics*: *Proceedings of the Twenty-Third Symposium on the Interface*, 571–578, Interface Foundation of North America Inc., Fairfax Station, Virginia.

Greenberg, B. V. and Zayatz, L. V. (1992). Strategies for measuring risk in public use microdata files, *Statist. Neerlandica*, **46**, 33–48.

Kanoh, S. (1997). Statistical disclosure control of microdata for public use, *Bulletin of the International Statistical Institute*, 51st session, Contributed papers TOME LVII, 425–426.

Lambert, D. (1993). Measure of disclosure risk and harm, *Journal of Official Statistics*, **9**, 313–331.

Marsh, C., Skinner, C. J., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D. and Walford,

N. (1991). The case for samples of anonymized records from the 1991 census, *J. Roy. Statist. Soc. Ser. A*, **154**, 305–340.

Mokken, R. J., Kooiman, P., Pannekoek, J. and Willenborg, L. (1992). Disclosure risks for microdata, *Statist. Neerlandica*, **46**, 49–67.

Nandram, B. and Chen, M.-H. (1996). Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence, *J. Statist. Comput. Simulation*, **545**, 129–144.

Omori, Y. (1999). Measuring identification disclosure risk for categorical microdata by posterior population uniqueness, *Statistical Data Protection-Proceedings of the Conference, Lisbon, 25 to 27 March 1998–1999 Edition*, 59–76, Office for Official Publications of the European Communities, Luxembourg.

Paass, G. (1988). Disclosure risk and disclosure avoidance for microdata, *J. Bus. Econom. Studies*, **6**, 487–500.

Skinner, C. J. and Holmes, D. J. (1993). Modeling population uniques, *Proceedings of the International Seminar on Statistical Confidentiality*, Dublin.

Skinner, C. J., Marsh, C., Openshaw, S. and Wymer, C. (1994). Disclosure control for census microdata, *Journal of Official Statistics*, **10**, 31–51.

Takemura, A. (1997). Some superpopulation models for estimating the number of population uniques, Discussion Paper 97-F-29, Faculty of Economics, University of Tokyo, Tokyo.

Willenborg, L. and de Waal, T. (1996). *Statistical Disclosure and Control in Practice*, Lecture Notes in Statist., **111**, Springer, New York.

Proceedings of the Institute of Statistical Mathematics Vol. 51, No. 2, 223–239 (2003) 239

Posterior Probability of Population Uniqueness in Microdata

Yasuhiro Omori

(Faculty of Economics, University of Tokyo)

There is always an identification disclosure risk when we release microdata. Some individuals or companies might be identified by sample unique observations in the dataset and then their privacy would be violated. Thus, the identification disclosure risk needs to be evaluated when we disclose microdata such as official statistics. We usually have many sample unique observations in the microdata, but most of them are not considered to be population unique. This article evaluates a probability of identification disclosure of any individual using the posterior probability of population uniqueness, and discusses how we should use it as a criterion of disclosure risk.

Key words: Identification disclosure risk, Markov chain Monte Carlo, microdata, population uniqueness, posterior probability.