統計数理 (2000)
第48巻 第1号 229-251

[原著論文]

予測個体数の期待値に基づく個票データのリスク評価

岡山商科大学* 佐 井 至 道

(受付 2000年1月21日；改訂 2000年3月17日)

要 旨

標本調査によって得られた個票データを公開する際には，プライバシーの侵害の度合いとして標本でも母集団でも一意，すなわちキー変数の組み合わせが他のすべての個体と異なる個体数を指標として用いるのが一般的である．しかし一意以外の個体についてもその程度は低くなるものの危険性をもっていることから，それらを考慮に入れた，より総合的な指標が必要であると考えられる．本論文では，第三者がキー変数を用いて個票データのすべての個体について予測を行うことを仮定した場合に予測される個体数の期待値を指標として提案し，その性質について議論を行う．その際，超母集団モデルとしてポアソンガンマモデルを用いた場合の，期待値とパラメータとの関係などの性質について検討を行う．また，この指標を含むような一般的な指標への拡張も考える．

個票データの開示によってある個体のセンシティブ変数のカテゴリーが第三者にとってある程度予測しやすくなることを予測漏洩と呼ぶが，本論文ではその概念を取り入れ，あるセンシティブ変数のカテゴリーが予測される個体数の期待値の性質についても議論する．センシティブ変数には超母集団モデルとして多項分布を用いたモデルや多項ディリクレモデルを用いることにする．

更に個票データに対して提案した手法を適用して，その有用性をみるとともに，実用上の問題点とその対処方法を探ることにする．

キーワード：個票データ，キー変数，センシティブ変数，予測漏洩，ポアソンガンマモデル，多項ディリクレモデル．

1. はじめに

個票データ (microdata) の項目を構成する変数はキー変数 (key variable)，センシティブ変数 (sensitive variable)，そのどちらでもない変数に大別される．キー変数は性別や年齢のように第三者がその情報を個票データ以外から知りうる項目の変数で，個体 (individual) の識別 (identification) に用いることができるのに対して，センシティブ変数は年間所得や就業状態のように，その項目のカテゴリーが第三者に知れることによってプライバシーの侵害が起こる変数である．なおキー変数かつセンシティブ変数となる項目もあり，場合によっては年間所得のようにキー変数として用いる場合には個票データのカテゴリーよりも粗い分類としてしか用いることのできないものもある．

* 商学部：〒700-8601 岡山市津島京町2-10-1．

個票データを公開する際には，その中に含まれる個体のプライバシーが保護されなければならないが，標本調査の場合にはプライバシーの侵害の度合いとして標本でも母集団でも一意(unique)である個体数，すなわち他のすべての個体とキー変数の組み合わせが異なる個体数を指標として用いるのがこの種の検討では一般的である．第三者が母集団について十分知識を持っているならば直ちにその個体を識別でき，その結果，センシティブ変数に関する情報を知ることが可能というのがその理由である．

一意性の検討の重要性については Willenborg and de Waal (1996) に述べられているが，その他にも文献は多い．例えば一意性を考慮したイギリスのセンサスの例を Marsh et al. (1991, 1994) が，ドイツのセンサスの例をMüller et al. (1995) などがそれぞれ扱っている．また Bethlehem et al. (1990) は，ポアソンガンマモデル (Poisson-gamma model) を用いて，標本である個票データから母集団で一意である個体数を推定する方法を提案し，Keller and Bethlehem (1992)，Skinner (1992)，佐井 (1998) においても同じモデルが用いられている．更に Omori (1997) は多項ディリクレモデル (Dirichlet-multinomial model) を用いて，母集団のセルに含まれる個体数がある決められた値以下となる確率について議論しており，この2つのモデルやエベンスモデル (Ewens model)，対数級数モデルなどの関係を Hoshino and Takemura (1998)，Takemura (1999) が扱っている．これらとは別のアプローチとして，ノンパラメトリックな推定法を Greenberg and Zayatz (1992)，渋谷 (1999) が，検定法を加納 (1997) がそれぞれ扱っている．これらの文献の詳細については佐井 (1997, 1998) を参照していただきたい．

一意性の検討の際には，母集団二意，すなわち母集団において2個の個体のキー変数の組み合わせが同じ組で，そのうち少なくとも1個が標本としてとられているものの数がどの程度あるかを考慮する必要性も指摘され続けていた．2個の個体のうちの一方が他方を識別してしまうことがその理由である．また，母集団二意でしかも標本としてサンプリングされれば，第三者にとっても標本の個々の個体を確率½で予測 (識別ではない) できることになり，その意味でも危険と考えられる．同様に，母集団三意，四意などについても，程度は低くなるものの危険性を持っており，これらを総合的に考えることも必要と思われる．

総合的な指標はこれまでもいくつか提案されている．Bethlehem et al. (1990)，Keller and Bethlehem (1992) ではキー変数の分解 (resolution) が取り上げられている．キー変数の分解は母集団一意，二意などの度数を用いて，母集団から復元抽出された2つの個体のキー変数の組み合わせが等しくなる確率の逆数で定義され，その値が大きいほど危険と考えられる．また Greenberg and Zayatz (1992) はキー変数の分解と同様の性質のエントロピー関数 (entropy function) について議論している．しかしこれらの指標は，標本のサンプリングまでを十分に考慮したものとはいえず，危険性の指標としての意味づけがやや弱いものと思われる．

また予測漏洩 (prediction disclosure) という概念が，Dalenius (1977) などの考えを基に Duncan and Lambert (1986) によって作り上げられ，その後，Skinner (1992)，加納 (1997) などによっても議論が行われている．(予測開示という和訳が一般的だが開示より漏洩の方が適切との意見もあるため，ここでは漏洩という単語を用いる．) その定義はどの文献でも同様であるが，Skinner (1992) では，個票データの開示によってある個体のセンシティブ変数のカテゴリーが第三者にとってある程度予測しやすくなること，としている．ただし実際には，同じキー変数の組み合わせを持つ個体の中で，同じセンシティブ変数のカテゴリーの等しいものがどの程度あるかに限定した議論が多い．例えば10個の個体のキー変数の組み合わせが等しければそれらの個体自体を特定することはできないが，そのうち8個の個体のセンシティブ変数のカテゴリーが等しければ，その情報を高い確率で予測できてしまうというのが基本的な考えである．

Duncan and Lambert (1986) では，個々の個体の漏洩の程度を表す不確実性関数 (uncertainty function) を考え，個票データの公開前後における不確実性関数の値の差を知識の増加

(knowledge gain)，差を公開前の値で割ったものを相対的な知識の増加 (relative knowledge gain) と定義した．不確実性関数の例としてはエントロピー関数などを提案している．また Skinner (1992) では簡単なモデル化を行っているが，未知パラメータの推定が困難であるという問題がある．なお加納 (1997) はセンシティブ変数が連続的である場合についても議論を行っている．

以上のような点を踏まえて，本論文ではキー変数のみを用いて予測される個体数，更にはキー変数とセンシティブ変数を同時に用いてセンシティブ変数のカテゴリーが予測される個体数について，ポアソンガンマモデルや多項ディリクレモデルなどの超母集団モデルを用いて，それぞれ検討を行うことにする．まず第 2 章では，第三者がキー変数を用いて個票データのすべての個体について予測を行うことを仮定した場合に予測される個体数の期待値について議論する．その際，標本のサンプリングと，ポアソンガンマモデルを用いた母集団の超母集団からのサンプリングを段階的に考慮に入れながら，期待値の性質について検討を行う．第 3 章では予測される個体数の期待値を含むような，キー変数のみに基づく危険性の指標を考え，第 2 章と同様の検討を行うことにする．第 4 章では予測漏洩の概念を取り入れ，あるセンシティブ変数のカテゴリーが予測される個体数の期待値の性質について議論する．超母集団モデルとして，キー変数にはポアソンガンマモデル，センシティブ変数には多項分布を用いたモデルや多項ディリクレモデルを用いることにする．更に第 5 章ではアメリカ合衆国で 1990 年に行われたセンサスの 1%抽出データに対して提案した手法を適用して，その有用性をみるとともに，実用上の問題点とその対処方法を探ることにする．

本論文の目的は，これらの指標を用いて個票データ開示の危険性を評価することであるが，個体またはそのセンシティブ変数のカテゴリーが予測される個体数は，識別される個体数の重要性を越える指標とはなり得ない．したがって本論文で提案する手法は，標本でも母集団でも一意である個体数という指標を補う形として用いるのが望ましいと考える．

なお，キー変数，センシティブ変数とも，カテゴリカルなものと連続的な値をとるものがあるが，元々連続的な変数でも個票データ上の桁数の制約によってカテゴリカルなデータに変換されていると考え，すべてカテゴリカルな変数と見なすことにする．

2. キー変数に基づいて予測される個体数の期待値

2.1 定　　義

母集団の N 個の個体が数種類のキー変数の組み合わせに基づいて K 個 ($K \geq 2$) のセル (グループ) に分けられているものとして，第 i 番目のセルに含まれる個体数を F_i ($i = 1, 2, \ldots, K$) とする．各セルの中には個体が全く入っていないものもあるため $F_i = 0$ となるセルも存在する．ここで，含まれる個体数が l のセル数，すなわち $F_i = l$ となるセル数を S_l ($l = 0, 1, 2, \ldots, N$) とする．また標本 (個票データ) の大きさを n として，標本では F_i，S_l の代わりに f_i，s_l ($l = 0, 1, 2, \ldots, n$) という表記を用いることにする．

ここで $F_i \neq 0$ である第 i 番目のセルに注目するとき，もし第三者が標本の f_i 個の個体それぞれを確率 $1/F_i$ ずつで母集団の個々の個体と予測することを考えると，このセル内では期待値として $f_i \cdot (1/F_i)$ 個の個体が予測されることになる．したがって，全セルについて同様の予測を行ったときに予測される個体数を I_1 とすると，その期待値は

$$E_{pr}(I_1) = \sum_{\substack{i=1 \\ (F_i \neq 0)}}^{K} \frac{f_i}{F_i} \tag{2.1}$$

と書くことができる．ただし $E_{pr}(\cdot)$ は予測 (prediction) に関する期待値を意味している．

2.2 母集団が固定されたときの期待値

この節では母集団の各個体のセルへの配置を固定した場合の，予測される個体数の期待値について検討を行うが，まず始めに期待値の上限と下限をとる母集団の各個体のセルへの配置を考えると，次のことが簡単にわかる．なおサンプリング法は，非復元単純無作為抽出またはベルヌーイ抽出を仮定して，抽出率を $\lambda = n/N$ とおく．ベルヌーイ抽出法は母集団の個体1個ずつが同じ抽出率で独立にサンプリングされる方法で，Särndal et al. (1992) の説明が詳しい．

ここで $N \leq K$ と仮定すると，上限は母集団の N 個のセルに1個ずつの個体が入っている場合，例えば $F_1 = F_2 = \cdots = F_N = 1$，$F_{N+1} = F_{N+2} = \cdots = F_K = 0$ のときで $E_{samp}(E_{pr}(I_1)) = n$ となる．ここで $E_{samp}(\cdot)$ は標本 (sample) のサンプリングに関する期待値を意味している．また下限は母集団の1個のセルにすべての個体が入っている場合，例えば $F_1 = N$，$F_2 = F_3 = \cdots = F_K = 0$ のときで $E_{samp}(E_{pr}(I_1)) = n/N = \lambda$ となる．すなわち $\lambda \leq E_{samp}(E_{pr}(I_1)) \leq n$ である．なお $K \leq N$ の場合にも下限についての結果は変わらないが，上限については下の議論から λK となることがわかる．

母集団での個体の各セルへの任意の配置 $F_1, F_2, \ldots, F_K$ に対して，予測される個体数の期待値を考えると

$$
\begin{aligned}
E_{samp}(E_{pr}(I_1)) &= E_{samp}\left(\sum_{\substack{i=1 \\ (F_i \neq 0)}}^{K} \frac{f_i}{F_i}\right) \\
&= \sum_{\substack{i=1 \\ (F_i \neq 0)}}^{K} \lambda \\
&= \lambda(K - S_0)
\end{aligned} \tag{2.2}
$$

と求められ，抽出率 λ と，含まれる個体数が0でないセル数のみで表されることがわかる．

2.3 超母集団モデルを仮定した予測

この節では，母集団が超母集団からサンプリングされたものとみなす超母集団モデルを導入した場合の予測される個体数の期待値について検討する．その場合，第三者の予測と標本のサンプリングに関する期待値に加えて，次のように母集団の超母集団からのサンプリングに関する期待値も考慮することになる．

$$
\begin{aligned}
E_{pop}(E_{samp}(E_{pr}(I_1))) &= E_{pop}(\lambda(K - S_0)) \\
&= \lambda\{K - E_{pop}(S_0)\}.
\end{aligned} \tag{2.3}
$$

ここで $E_{pop}(\cdot)$ は母集団 (population) の超母集団からのサンプリングを表しているが，この部分を実際に計算するためには超母集団モデルを決定する必要がある．ここでは Bethlehem et al. (1990) によって提案されたポアソンガンマモデルを用いることにする．F_i が他のセルと独立にポアソン分布

$$
p(F_i) = \frac{(N_0\pi_i)^{F_i} e^{-N_0\pi_i}}{F_i!} \tag{2.4}
$$

に従い，各セルへの入りやすさを表すパラメータ π_i が他のパラメータと独立にガンマ分布

$$
f(\pi_i) = \frac{\pi_i^{\alpha-1} e^{-\frac{\pi_i}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)} \tag{2.5}
$$

に従うとするのがポアソンガンマモデルである．α，β は正のパラメータで $\alpha\beta = 1/K$ という制約をおいているため，N_0，K の他には π_i の散らばりを表す1個のパラメータ β のみで表される

モデルである．なお $\beta \to 0$ のとき $\pi_i \equiv 1/K$ のポアソン分布のみのモデルに退化する．ポアソンガンマモデルでは超母集団からサンプリングされる母集団の個体数Nは期待値としては N_0 であるものの一定ではなく，しかも標本のサンプリングとしてベルヌーイ抽出を仮定しなければならない．ただし，母集団，標本とも十分大きく，個々のセルが興味の対象でなければ，大きな問題は起きないものと考えられる．ポアソンガンマモデルの性質については佐井 (1998) を参照されたい．

このモデルの下では，母集団において，含まれる個体数が 0 のセル数の期待値が

$$
\begin{aligned}
(2.6) \qquad E_{pop}(S_0) &= \sum_{i=1}^{K}\left\{\int_0^{\infty}\frac{(N_0\pi_i)^0 e^{-N_0\pi_i}}{0!}\cdot\frac{\pi_i^{\alpha-1}e^{-\frac{\pi_i}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}\,d\pi_i\right\} \\
&= \sum_{i=1}^{K}\left\{\frac{1}{(1+N_0\beta)^{\alpha}}\int_0^{\infty}\frac{\pi_i^{\alpha-1}e^{-\frac{1+N_0\beta}{\beta}\pi_i}}{\left(\frac{\beta}{1+N_0\beta}\right)^{\alpha}\Gamma(\alpha)}\,d\pi_i\right\} \\
&= \frac{K}{(1+N_0\beta)^{\frac{1}{K\beta}}}
\end{aligned}
$$

と表される．したがって予測される個体数の期待値は

$$
(2.7) \qquad E_{pop}(E_{samp}(E_{pr}(I_1))) = \lambda K\left\{1-\frac{1}{(1+N_0\beta)^{\frac{1}{K\beta}}}\right\}
$$

と書き直され，次のような性質が得られる．

定理 1. N_0, K を固定した場合 $E_{pop}(E_{samp}(E_{pr}(I_1)))$ は β に関する単調減少関数で，

$$
\begin{aligned}
&\lim_{\beta\to 0} E_{pop}(E_{samp}(E_{pr}(I_1))) = \lambda K\left(1-e^{-\frac{N_0}{K}}\right), \\
&\lim_{\beta\to\infty} E_{pop}(E_{samp}(E_{pr}(I_1))) = 0
\end{aligned}
$$

となる．

証明.

$$
\log E_{pop}(S_0) = \log K - \frac{1}{K\beta}\log(1+N_0\beta)
$$

を β で偏微分すると

$$
\frac{\partial \log E_{pop}(S_0)}{\partial\beta} = \frac{(1+N_0\beta)\log(1+N_0\beta) - N_0\beta}{K\beta^2(1+N_0\beta)}
$$

となるが，右辺の分子を $g_1(\beta)$ とおくと

$$
g_1'(\beta) = N_0\log(1+N_0\beta) > 0 \qquad (\beta > 0)
$$

となることがわかる．したがって $g_1(\beta)$ は $\beta > 0$ で単調増加で $g_1(0) = 0$ より $g_1(\beta)$ も同じ範囲で正となり，$\beta > 0$ で $E_{pop}(S_0)$ が単調増加，すなわち $E_{pop}(E_{samp}(E_{pr}(I_1)))$ が単調減少となることが言える．

また

$$E_{pop}(E_{samp}(E_{pr}(I_1))) = \lambda K\left\{1 - \frac{1}{\left\{(1+N_0\beta)^{\frac{1}{N_0\beta}}\right\}^{\frac{N_0}{K}}}\right\}$$

より $\lim_{\beta\to 0} E_{pop}(E_{samp}(E_{pr}(I_1))) = \lambda K(1-e^{-\frac{N_0}{K}})$, $\lim_{\beta\to\infty} E_{pop}(E_{samp}(E_{pr}(I_1))) = 0$ となることがわかる. □

$E_{pop}(E_{samp}(E_{pr}(I_1)))$ は，他の2つのパラメータを固定した場合 N_0 に関しても K に関しても単調増加であることも上と同様にして確かめられるが，β に関する性質を特に重視するのは，N_0 は一般に既知または高い精度で推定可能であり，K は既知または自由に設定可能であるのに対して，β は個票データから推定する必要があるためである．β の推定値そのものを用いると，期待値を実際の値よりも小さめに評価してしまう可能性もある．そこで β をやや小さめに推定したり，場合によっては0とすることによって，やや厳しく，またはもっとも厳しく評価することが可能となると考えられる．

3. キー変数に基づく危険性の指標

3.1 母集団が固定されたときの指標

前章では予測される個体数の期待値の性質について検討したが，これは第三者が個票データに含まれる個体すべてについて，母集団の同じキー変数の組み合わせを持つものの中から，いわば当てずっぽうで予測を行った場合に平均的に的中する個数である．しかし母集団内で同じキー変数の組み合わせを持つ個体があまりに多く，的中する確率が低いと思われれば，第三者は予測することをあきらめてしまうかもしれない．また母集団一意かつ標本一意であればどの第三者でも識別することが可能であり，母集団二意かつ標本一意または二意であれば，その2つの個体間でのみ識別することが可能であるが，母集団三意以上であれば単独での識別は不可能となるため危険性は大きく減少すると考えることもできる．

このような立場で危険性の検討を行うためには，母集団の第 i 番目のセルにおいて，前章で考えた予測される確率 $1/F_i$ に代えて，その関数 $v(1/F_i)$ を用いることで対応できる．ただし $v(x)$ は1以上の個体数に対応する重みを表す $x = 1, 1/2, 1/3, \ldots$ で定義される関数とする．

ここで予測される個体数の期待値 $E_{pr}(I_1)$ に対応する危険性の指標を

$$I_2 = \sum_{\substack{i=1\\(F_i\neq 0)}}^{K} f_i \cdot v\left(\frac{1}{F_i}\right) \tag{3.1}$$

と表すことにすると，まず母集団の各個体のセルへの配置を固定したときの標本のサンプリングに関する期待値は次のように書ける．

$$\begin{aligned} E_{samp}(I_2) &= \sum_{\substack{i=1\\(F_i\neq 0)}}^{K} E_{samp}(f_i)\cdot v\left(\frac{1}{F_i}\right) \\ &= \lambda \sum_{\substack{i=1\\(F_i\neq 0)}}^{K} F_i \cdot v\left(\frac{1}{F_i}\right). \end{aligned} \tag{3.2}$$

3.2 超母集団モデルを仮定した指標

この節では (3.2) 式に，次のように超母集団からの母集団のサンプリングに関する期待値も

加えて検討を行う．

$$E_{pop}(E_{samp}(I_2)) = \lambda E_{pop}\left(\sum_{\substack{i=1\\(F_i\neq 0)}}^{K} F_i\cdot v\left(\frac{1}{F_i}\right)\right). \tag{3.3}$$

ここでも超母集団モデルとしてポアソンガンマモデルを用いることにする．なお具体的に関数 $v(x)$ を決める必要があるが，すべての個体数 l ($l=1, 2, \ldots$) に対して正で定義される関数の中では実際に計算可能なものは多くない．ここでは F_i が大きくなるほど危険性をより小さく評価するような例として $v(1/l)=1/a^l$ ($a>1$) の場合について考える．

例 1. $v\left(\frac{1}{l}\right)=\frac{1}{a^l}$ ($a>1$) の場合．

母集団の超母集団からのサンプリングに関する期待値を，ポアソン分布に関する期待値（添え字 *pop*1）とガンマ分布に関する期待値（添え字 *pop*2）に分けて考える．便宜上 $l=0$ のとき $v(1/l)=0$ と定義しておくと，ポアソン分布に関して

$$\begin{aligned} E_{pop1}\left(F_i\cdot v\left(\frac{1}{F_i}\right)\right) &= E_{pop1}\left(\frac{F_i}{a^{F_i}}\right) \\ &= \sum_{F_i=0}^{\infty}\frac{F_i}{a^{F_i}}\cdot\frac{e^{-N_0\pi_i}(N_0\pi_i)^{F_i}}{F_i!} \\ &= \frac{N_0\pi_i}{a}e^{-N_0\pi_i}\sum_{F_i=1}^{\infty}\frac{\left(\frac{N_0\pi_i}{a}\right)^{F_i-1}}{(F_i-1)!} \\ &= \frac{N_0\pi_i}{a}\cdot e^{-\frac{(a-1)N_0\pi_i}{a}} \end{aligned} \tag{3.4}$$

と期待値が求められ，更にガンマ分布について期待値が

$$\begin{aligned} &E_{pop2}\left(\frac{N_0\pi_i}{a}\cdot e^{-\frac{(a-1)N_0\pi_i}{a}}\right) \\ &= \int_0^1\frac{N_0\pi_i}{a}\cdot e^{-\frac{(a-1)N_0\pi_i}{a}}\cdot\frac{\pi_i^{\alpha-1}e^{-\frac{\pi_i}{\beta}}}{\beta^\alpha\Gamma(\alpha)}\,d\pi_i \\ &= \frac{N_0\Gamma(\alpha+1)\left\{\frac{a\beta}{a+(a-1)N_0\beta}\right\}^{\alpha+1}}{a\Gamma(\alpha)\beta^\alpha}\int_0^1\frac{\pi_i^\alpha e^{-\left\{\frac{1}{\beta}+\frac{(a-1)N_0}{a}\right\}\pi_i}}{\left\{\frac{a\beta}{a+(a-1)N_0\beta}\right\}^{\alpha+1}\Gamma(\alpha+1)}\,d\pi_i \\ &= \frac{N_0}{aK\left(1+\frac{a-1}{a}N_0\beta\right)^{1+\frac{1}{K\beta}}} \end{aligned} \tag{3.5}$$

と求められるため，

$$\begin{aligned} E_{pop}(E_{samp}(I_2)) &= \lambda\sum_{i=1}^{K}\frac{N_0}{aK\left(1+\frac{a-1}{a}N_0\beta\right)^{1+\frac{1}{K\beta}}} \\ &= \lambda\cdot\frac{N_0}{a\left(1+\frac{a-1}{a}N_0\beta\right)^{1+\frac{1}{K\beta}}} \end{aligned} \tag{3.6}$$

が得られる.

この期待値は明らかにKに関して単調増加で,N_0に関しては$N_0=(a/(a-1))K$において極大値をとることが偏微分することによって容易に求められる.またβに関して次の性質が得られる.

定理 2. $v\left(\frac{1}{l}\right)=\frac{1}{a^l}$ $(a>1)$ の場合,次のことがいえる.

(i) $N_0\leq(2a/(a-1))K$ のとき,$E_{pop}(E_{samp}(I_2))$ は β に関して単調減少である.

(ii) $(2a/(a-1))K<N_0$ のとき,$E_{pop}(E_{samp}(I_2))$ は $0<\beta<\beta_1$ において単調増加,$\beta_1<\beta$ において単調減少となる.

ただし β_1 は

$$\log\left(1+\frac{a-1}{a}N_0\beta\right)-(K\beta^2+\beta)\frac{\frac{a-1}{a}N_0}{1+\frac{a-1}{a}N_0\beta}=0$$

の正の解 β である.

証明.

$$\log E_{pop}(E_{samp}(I_2))=\log\frac{\lambda N_0}{a}-\left(1+\frac{1}{K\beta}\right)\log\left(1+\frac{a-1}{a}N_0\beta\right)$$

を β に関して偏微分すると

$$\frac{\partial\log E_{pop}(E_{samp}(I_2))}{\partial\beta}=\frac{1}{K\beta^2}\left\{\log\left(1+\frac{a-1}{a}N_0\beta\right)-(K\beta^2+\beta)\frac{\frac{a-1}{a}N_0}{1+\frac{a-1}{a}N_0\beta}\right\}$$

となるが,右辺の括弧の中を $g_2(\beta)$ とおくと

$$g_2'(\beta)=\frac{\frac{a-1}{a}N_0\beta}{1+\frac{a-1}{a}N_0\beta}\cdot\frac{-\frac{a-1}{a}KN_0\beta+\left(\frac{a-1}{a}N_0-2K\right)}{1+\frac{a-1}{a}N_0\beta}$$

となる.

したがって $N_0\leq(2a/(a-1))K$ のとき,上式の右辺後半の分子が $\beta>0$ において常に負となり $g_2'(\beta)<0$ となる.ここで $g_2(0)=0$ より $g_2(\beta)$ も同じ範囲で負となり,$\beta>0$ において $E_{pop}(E_{samp}(I_2))$ が単調減少であることがわかる.

また $(2a/(a-1))K<N_0$ のとき $\beta_2=(1/K)-2/(((a-1)/a)N_0)$ とおくと,右辺後半の分子は $\beta<\beta_2$ において正,$\beta_2<\beta$ において負となる.ここで $g_2(0)=0$,また $\lim_{\beta\to\infty}g_2(\beta)=-\infty$ より $g_2(\beta_1)=0$ となる正の β_1 が存在し,$0<\beta<\beta_1$ において $g_2(\beta)>0$ すなわち $E_{pop}(E_{samp}(I_2))$ は単調増加で,$\beta_1<\beta$ では $g_2(\beta)<0$ すなわち $E_{pop}(E_{samp}(I_2))$ は単調減少となる. □

よって(i)の場合には β を小さめに，(ii)の場合には β を β_1 に近づけて推定することによって，やや厳しい推定をすることが可能となる．

これに対して，有限個の l の値以外では $v(1/l)=0$ と定義する関数については一般に計算は可能で，実用的なものも多いと思われる．例えば，次のように母集団二意までを危険，母集団三意以上を安全と評価するなど，多くの危険性の指標が考えられる．

例 2.

$$v\left(\frac{1}{l}\right)=\begin{cases}1 & (l=1)\\ \dfrac{1}{2} & (l=2)\\ 0 & (l\geq 3)\end{cases}\quad \text{の場合.}$$

このときも例1と同様に，便宜上 $l=0$ のとき $v(1/l)=0$ と定義しておくと，ポアソン分布に関して期待値が

$$E_{pop1}\left(F_i\cdot v\left(\frac{1}{F_i}\right)\right)=e^{-N_0\pi_i}N_0\pi_i+\frac{1}{2}e^{-N_0\pi_i}(N_0\pi_i)^2$$

と得られるため，

$$E_{pop}(E_{samp}(I_2))=\lambda\sum_{i=1}^{K}E_{pop2}\left(e^{-N_0\pi_i}N_0\pi_i+\frac{1}{2}e^{-N_0\pi_i}(N_0\pi_i)^2\right)$$
$$=\lambda\left\{\frac{N_0}{(1+N_0\beta)^{1+\frac{1}{K\beta}}}+\frac{N_0^2\left(\frac{1}{K}+\beta\right)}{2(1+N_0\beta)^{2+\frac{1}{K\beta}}}\right\}\tag{3.7}$$

と求めることができる．すなわち $E_{pop}(E_{samp}(I_2))=\lambda\{E_{pop}(S_1)+E_{pop}(S_2)\}$ である．

この期待値の β に関する増減について次のような性質が得られる．

定理 3. $v(1/l)$ が例2のように決められる場合，次のことがいえる．

(i) $N_0\leq cK$ のとき $E_{pop}(E_{samp}(I_2))$ は β に関して単調減少である．

(ii) $cK<N_0$ のとき $E_{pop}(E_{samp}(I_2))$ は $0<\beta<\beta_3$ において単調増加，$\beta_3<\beta$ において単調減少である．

ただし c は方程式 $c^3-6c-4=0$ の正の解で $c\doteqdot 2.7321$，β_3 は方程式

$$\frac{3K^2N_0\beta^2}{2K+N_0+3KN_0\beta}+\log(1+N_0\beta)-(2K\beta^2+\beta)\frac{N_0}{1+N_0\beta}=0$$

の正の解である．

証明.

$$E_{pop}(E_{samp}(I_2))=\lambda\cdot\frac{N_0}{2}\cdot\frac{2+\frac{N_0}{K}+3N_0\beta}{(1+N_0\beta)^{2+\frac{1}{K\beta}}}$$

の対数をとり，β に関して偏微分すると

$$\frac{\partial \log E_{pop}(E_{samp}(I_2))}{\partial \beta}=\frac{1}{K\beta^2}\left\{\frac{3K^2N_0\beta^2}{2K+N_0+3KN_0\beta}+\log(1+N_0\beta)-(2K\beta^2+\beta)\frac{N_0}{1+N_0\beta}\right\}$$

となる．ここで右辺の{ }内を $g_3(\beta)$ とおき，再び β で微分すると

$$\begin{aligned}g_3'(\beta)=&\frac{N_0\beta}{(2K+N_0+3KN_0\beta)^2(1+N_0\beta)^2}\\&\cdot\{-9K^3N_0^3\beta^3+(3K^2N_0^3-30K^3N_0^2)\beta^2+(4KN_0^3-8K^2N_0^2-23K^3N_0)\beta\\&+(N_0^3-6K^2N_0-4K^3)\}\end{aligned}$$

となる．ここで右辺後半の β の3次式を $g_4(\beta)$ とおき，その符号を得るために $y=g_4(\beta)$ の極値の数，変曲点をとる β の符号，$g_4(0)$ の符号，極値をとる β の符号について考える．

(I) 極値の数

$$g_4'(\beta)=-KN_0\{27K^2N_0^2\beta^2-6KN_0(N_0-10K)\beta-(4N_0^2-8KN_0-23K^2)\}$$

より，{ }内の β の2次式を0とする方程式の判別式を計算することにより，$1\leq N_0/K\leq 31/13$ のとき極値を持たず，$0<N_0/K<1$ または $31/13<N_0/K$ のとき2つの極値を持つことがわかる．

(II) 変曲点をとる β の符号

$g_4''(\beta)=-6K^2N_0^2\{9KN_0\beta-(N_0-10K)\}$ より，変曲点の β の値は $0<N_0/K\leq 10$ のとき0以下で，$N_0/K>10$ のとき正となる．

(III) $g_4(0)$ の符号

$$g_4(0)=K^3\left\{\left(\frac{N_0}{K}\right)^3-6\frac{N_0}{K}-4\right\}$$

となるが，ここで

$$y=x^3-6x-4$$

のグラフを考えることにより，$c^3-6c-4=0$ を満たす正の数 $c\doteqdot 2.7321$ が存在して，$0<N_0/K\leq c$ のとき $g_4(0)\leq 0$ で，$c<N_0/K$ のとき $g_4(0)>0$ となることがわかる．

(IV) 極値をとる β の符号

$y=g_4(\beta)$ が2つの極値を持つとき，すなわち $0<N_0/K<1$ または $31/13<N_0/K$ のとき，2つの極値をとる β の値の積は

$$\frac{-4N_0^2+8KN_0+23K^2}{27K^2N_0^2}=\frac{1}{27N_0^2}\left\{-4\left(\frac{N_0}{K}\right)^2+8\frac{N_0}{K}+23\right\}$$

となるので，$0<N_0/K\leq(2+3\sqrt{3})/2$ のとき積は0以上で極値をとる β の値は同符号または1つが0となり，$(2+3\sqrt{3})/2<N_0/K$ のとき積は負で極値をとる β の値は異符号となる．

ここで N_0/K の値によって次のように場合分けして考える．

(i) $0<N_0/K<1$, $31/13<N_0/K\leq c$ のとき，$y=g_4(\beta)$ は(I)より極値は2個，(II)より変曲点の β の値は負，(III)より $g_4(0)$ は0以下，(IV)より極値をとる β の値は同符号であることより $g_4(\beta)<0$ $(\beta>0)$ となることがわかる．したがって $g_3(\beta)$ は同じ範囲で単調減少で $g_3(0)=0$ より $g_3(\beta)<0$ $(\beta>0)$ となり $E_{pop}(E_{samp}(I_2))$ は $\beta>0$ において単調減少である．

(ii) $1\leq N_0/K\leq 31/13$ のとき，(I), (III)より(i)と同様に $E_{pop}(E_{samp}(I_2))$ は $\beta>0$ において単調減少である．

(iii)　$c < N_0/K \leq (2+3\sqrt{3})/2$ のとき，$y = g_4(\beta)$ は(I)より極値は2個，(II)より変曲点の β の値は負，(III)より $g_4(0)$ は正である．したがって $g_4(\beta_4) = 0$ となる正の数 β_4 が存在して $0 < \beta < \beta_4$ のとき $g_4(\beta) > 0$ より $g_3(\beta)$ は単調増加，$\beta_4 < \beta$ のとき $g_4(\beta) < 0$ より $g_3(\beta)$ は単調減少となる．ここで $g_3(0) = 0$ と $\lim_{\beta\to\infty} g_3(\beta) = -\infty$ より $g_3(\beta_3) = 0$ となる正の数 β_3 が存在して $0 < \beta < \beta_3$ のとき $g_3(\beta_3) > 0$ より $E_{pop}(E_{samp}(I_2))$ は単調増加，$\beta_3 < \beta$ のとき $g_3(\beta_3) < 0$ より $E_{pop}(E_{samp}(I_2))$ は単調減少である．

(iv)　$(2+3\sqrt{3})/2 < N_0/K$ のとき，(I)，(III)，(IV)より(ii)と同様に $E_{pop}(E_{samp}(I_2))$ は $0 < \beta < \beta_3$ において単調増加，$\beta_3 < \beta$ において単調減少である．

(i)～(iv)より証明された．□

なお $N_1 = (2K^2\beta + \sqrt{(2K\beta+1)(2K\beta+2)})/(3K\beta+1)$ とおくとき，$E_{pop}(E_{samp}(I_2))$ は $N = N_1$ において極大となり，K に関しては単調増加となることも，それぞれ偏微分することによって求めることができる．

ここでは $v(1/1) = 1$，$v(1/2) = 1/2$ としたが，一意である個体は任意の第三者から特定される危険性があるのに対して，母集団二意の個体は相互にしか特定できないため危険性をより低く評価して，$v(1/2) = (1/2)w$ $(0 < w \leq 1)$ としても同様の議論が可能となるが，計算はかなり複雑である．ただし $N_0 < K$ の場合に $E_{pop}(E_{samp}(I_2))$ が単調減少となることは容易に求められる．現実の個票データでは秘匿措置を十分施した後でもセル数は個体数よりも大きいのが一般的であるため，β をやや小さめに推定することによって，より厳しい評価が可能となると思われる．

また例2を拡張して，任意の自然数 m に対して $v(1/l) = 1/l$ $(l \leq m)$，$v(1/l) = 0$ $(l > m)$ とした場合に $E_{pop}(E_{samp}(I_2)) = \lambda \sum_{l=1}^{m} E_{pop}(S_l)$ となることも容易に求められる．ただし m が増加するにしたがって，定理3のような性質を得るのは難しくなる．

4. センシティブ変数も考慮した場合の予測される個体数の期待値

4.1 定　　義

第2章ではキー変数の組み合わせによって予測される個体数の期待値について議論したが，この章では更にセンシティブ変数による予測も考慮に入れることにする．センシティブ変数はキー変数に含まれる場合と含まれない場合がある．また例えば個票データに1週間の従業時間という項目が1時間刻みで記載されていて，しかもその項目がセンシティブ変数と考えられるとき，キー変数として用いる場合にはせいぜい5時間刻みとしてしか利用できないことも考えられる．この章ではセンシティブ変数として，キー変数に含まれないもの，またはキー変数よりも更に細かいカテゴリーに分けられる部分を考えていくことにする．

次のような例を考える．キー変数の組み合わせによってできるセルの1つにおいて，母集団で5個，標本で3個の個体が含まれているとする．キー変数による予測だけを考えれば，2.1節で議論したように $3\cdot(1/5) = 3/5$ 個がこのセル内で平均的に予測される．ここで，あるセンシティブ変数に関して，母集団の5個の個体のうち3個，1個，1個が別々のカテゴリーに属し，そのうちの2個，1個，0個が標本としてサンプリングされているものとすると，例えば標本の最初の2個それぞれが母集団のどの個体であるかを予測される確率は依然として1/5であるものの，そのセンシティブ変数のカテゴリーを予測される確率は3/5と高まり，結果的にこのセル内でセンシティブ変数の項目が予測される個体数の期待値は $2\cdot(3/5) + 1\cdot(1/5) = 7/5$ と高まることになる．

センシティブ変数については，いくつかのセンシティブ変数の組み合わせがプライバシーの侵害を生むような場合も考えられるし，個々のセンシティブ変数，またはそのいくつかのカテゴリーだけが問題となる場合もある．ここでは個々のセンシティブ変数についての定義で考えていくが，組み合わせの場合についても同様である．まず次のように定義を修正，追加する．

キー変数の組み合わせによって分けられているセル数を K_1 個 ($K_1 \geq 2$) と定義し直して，その第 i 番目のセルにおいて，キー変数に含まれていない目的変数であるセンシティブ変数のカテゴリーに応じて，更に K_2 個 ($K_2 \geq 2$) の2次セルに分けられているものとする．各セルごとに2次セル数が異なることもあり得るが，そのときには最大数を K_2 と決めることにする．第 i 番目のセルの第 j 番目の2次セルに含まれる母集団の個体数を $F_{i,j}$，標本の個体数を $f_{i,j}$ とする．$F_i = \sum_{j=1}^{K_2} F_{i,j}$，$f_i = \sum_{j=1}^{K_2} f_{i,j}$ である．このときセンシティブ変数のカテゴリーが予測される個体数を I_3 とすると，前述の例のように考えて

$$E_{pr}(I_3) = \sum_{\substack{i=1 \\ (F_i \neq 0)}}^{K_1} \sum_{j=1}^{K_2} f_{i,j} \frac{F_{i,j}}{F_i} \tag{4.1}$$

と期待値が求められる．

4.2 母集団が固定されたときの予測

ここでは2.2節と同様に，母集団の各個体のセルへの配置 $F_{i,j}$ ($i = 1, 2, \ldots, K_1$; $j = 1, 2, \ldots, K_2$) を固定した場合の予測される個体数の期待値について検討を行う．標本のサンプリングに関する $E_{pr}(I_3)$ の期待値は

$$\begin{aligned} E_{samp}(E_{pr}(I_3)) &= \sum_{\substack{i=1 \\ (F_i \neq 0)}}^{K_1} \sum_{j=1}^{K_2} E_{samp}(f_{i,j}) \cdot \frac{F_{i,j}}{F_i} \\ &= \lambda \sum_{\substack{i=1 \\ (F_i \neq 0)}}^{K_1} \sum_{j=1}^{K_2} \frac{F_{i,j}^2}{F_i} \end{aligned} \tag{4.2}$$

となり，センシティブ変数を考慮したことによる，予測される個体数の期待値の増加分は (4.2) 式から (2.2) 式を引くことによって

$$E_{samp}(E_{pr}(I_3)) - E_{samp}(E_{pr}(I_1)) = \lambda \sum_{\substack{i=1 \\ (F_i \neq 0)}}^{K_1} \sum_{j=1}^{K_2} \frac{F_{i,j}^2 - F_{i,j}}{F_i} \tag{4.3}$$

と求められる．

4.3 ポアソンガンマ-多項ディリクレモデル

この節と次節では，超母集団からの母集団のサンプリングも含めた評価を行う．キー変数によるセルとセンシティブ変数による2次セルを同時に考えて K_1K_2 個のセルに対して1つのモデルを仮定することも考えられるが，やや柔軟性に欠けるため，ここではセルと2次セルに対して段階的に2つのモデルを当てはめることにする．

キー変数の組み合わせによってできるセルに対しては2.3節で用いたポアソンガンマモデルをそのまま利用することにすると，2次セルに対してはベルヌーイ抽出を仮定できるような，ポアソン分布を用いたモデルやポアソンガンマモデルが便利であるが，両者とも第 i 番目のセルに含まれる個体数 F_i よりも，同じセル内の2次セルに含まれる個体数の和 $\sum_{j=1}^{K_2} F_{i,j}$ の方が大きくなる可能性があり，特に F_i が小さい場合には影響が大きいと思われる．そのため2次セルに対しては，多項分布を用いたモデルや多項ディリクレモデルのように総個体数が固定され

ているモデルの方が適当であろう．前者は2次セル数が小さい場合か，すべてのセルにおいて2次セルへの個体の入り方が等しいと考えられる場合などに有効であるが，後者の方が利用範囲は広いと思われるため，この節ではセンシティブ変数に対して多項ディリクレモデルを適用することにする．なお，ポアソンガンマモデルにおいて個体数の和を固定した条件付きの分布が多項ディリクレモデルとなることなど，この2つのモデルを含む数種類のモデルの関係については Hoshino and Takemura (1998)，Takemura (1999) を参照されたい．

センシティブ変数による2次セルに対して，次のように多項ディリクレモデルを用いる．ポアソンガンマモデルについては 2.3 節での定義をそのまま用い，2次セル内の個体数 $F_{i,j}$ ($j=1, 2, \ldots, K_2$) については

$$p(F_{i,1}, F_{i,2}, \ldots, F_{i,K_2}) = \frac{F_i!}{\prod_{j=1}^{K_2} F_{i,j}!} \prod_{j=1}^{K_2} \pi_j'^{F_{ij}}, \tag{4.4}$$

$$f(\pi_1', \pi_2', \ldots, \pi_{K_2}') = \frac{\Gamma(K_2\gamma)}{\{\Gamma(\gamma)\}^{K_2}} \prod_{j=1}^{K_2} \pi_j'^{\gamma-1} \tag{4.5}$$

とする．(4.5) 式は通常のディリクレ分布の K_2 個のパラメータをすべて γ と等しくしたものである．なお $F_{i,j}$ は i について独立とする．また π_j' も i について独立に決定される方が適用範囲は広くなるものの，パラメータの推定の観点からは π_j' がすべてのセルについて共通とする方が望まれる．ただしこの独立性は以下の議論に影響を与えないため，ここではこの点について特に限定しないことにする．

このモデルをポアソンガンマ-多項ディリクレモデルと呼ぶことにすると，次のような結果が得られる．

定理 4. ポアソンガンマ-多項ディリクレモデルを用いた場合，$E_{samp}(E_{pr}(I_3))$ の母集団の超母集団からのサンプリングに関する期待値は

$$E_{pop}(E_{samp}(E_{pr}(I_3))) = \lambda\left\{\frac{\gamma+1}{K_2\gamma+1} N_0 + \frac{(K_2-1)\gamma}{K_2\gamma+1}\left\{K_1 - \frac{K_1}{(1+N_0\beta)^{\frac{1}{K_1\beta}}}\right\}\right\} \tag{4.6}$$

$$= \lambda\left\{K_1 - \frac{K_1}{(1+N_0\beta)^{\frac{1}{K_1\beta}}}\right\} + \lambda\cdot\frac{\gamma+1}{K_2\gamma+1}\left\{N_0 - \left\{K_1 - \frac{K_1}{(1+N_0\beta)^{\frac{1}{K_1\beta}}}\right\}\right\} \tag{4.7}$$

である．

証明. センシティブ変数に関する期待値と分散をそれぞれ $E_{sens}(\cdot)$，$V_{sens}(\cdot)$，更に多項分布とディリクレ分布に関しては添え字 “1”，“2” をその後に付けることにすると

$$E_{sens2}(\pi_j') = \frac{1}{K_2},$$

$$V_{sens2}(\pi_j') = \frac{K_2-1}{K_2^2(K_2\gamma+1)}$$

であるから

$$\begin{aligned} E_{sens}(F_{i,j}^2) &= V_{sens}(F_{i,j}) + \{E_{sens}(F_{i,j})\}^2 \\ &= V_{sens2}(E_{sens1}(F_{i,j})) + E_{sens2}(V_{sens1}(F_{i,j})) + \{E_{sens2}(E_{sens1}(F_{i,j}))\}^2 \\ &= V_{sens2}(F_i\pi_j') + E_{sens2}(F_i\pi_j'(1-\pi_j')) + \{E_{sens2}(F_i\pi_j')\}^2 \end{aligned}$$

$$= \frac{\gamma+1}{K_2(K_2\gamma+1)} F_i^2 + \frac{(K_2-1)\gamma}{K_2(K_2\gamma+1)} F_i$$

となるので，キー変数に関する期待値を $E_{key}(\cdot)$ と書くことにすると，$E_{samp}(E_{pr}(I_3))$ の母集団の超母集団からのサンプリングに関する期待値は

$$\begin{aligned}
&E_{pop}(E_{samp}(E_{pr}(I_3))) \\
&= \lambda \cdot E_{key}\left[\sum_{\substack{i=1 \\ (F_i \neq 0)}}^{K_1} \sum_{j=1}^{K_2} \frac{1}{F_i} \left\{ \frac{\gamma+1}{K_2(K_2\gamma+1)} F_i^2 + \frac{(K_2-1)\gamma}{K_2(K_2\gamma+1)} F_i \right\} \right] \\
&= \lambda \cdot E_{key}\left[\sum_{\substack{i=1 \\ (F_i \neq 0)}}^{K_1} \left\{ \frac{\gamma+1}{K_2\gamma+1} F_i + \frac{(K_2-1)\gamma}{K_2\gamma+1} \right\} \right] \\
&= \lambda \left\{ \frac{\gamma+1}{K_2\gamma+1} N_0 + \frac{(K_2-1)\gamma}{K_2\gamma+1} \left\{ K_1 - \frac{K_1}{(1+N_0\beta)^{\frac{1}{K_1\beta}}} \right\} \right\} \\
&= \lambda \left\{ K_1 - \frac{K_1}{(1+N_0\beta)^{\frac{1}{K_1\beta}}} \right\} + \lambda \cdot \frac{\gamma+1}{K_2\gamma+1} \left\{ N_0 - \left\{ K_1 - \frac{K_1}{(1+N_0\beta)^{\frac{1}{K_1\beta}}} \right\} \right\}
\end{aligned}$$

と求められる．□

$E_{pop}(E_{samp}(E_{pr}(I_3)))$ が β に関して単調減少であることは (4.6) 式と定理1から示され，γ に関して単調減少であることは (4.7) 式から明らかである．すなわち，$V(\pi_i)$ が小さく，$V(\pi_j')$ が大きいほど期待値は大きくなる．また (4.7) 式の前半の項はキー変数のみに基づく予測される個体数で，後半の項がセンシティブ変数を考慮したことによってカテゴリーが予測される個体数の期待値の増加分となるが，この項は β に関する増加関数であるから，$V(\pi_i)$ が大きいほど増加分の大きくなることがわかる．

ところで (4.4)，(4.5) 式に代えて

$$p(F_{i,j}) = \frac{(F_i\pi_j')^{F_{i,j}} e^{-F_i\pi_j'}}{F_{i,j}!}, \tag{4.8}$$

$$f(\pi_j') = \frac{\pi_j'^{\alpha'-1} e^{-\frac{\pi_j'}{\beta'}}}{\beta'^{\alpha'}\Gamma(\alpha')}, \tag{4.9}$$

ただし $\alpha'\beta' = K_2$，また $F_{i,j}$，π_j' は i, j について独立として，センシティブ変数に対してもポアソンガンマモデルを用いるとき，(4.7) 式に対応するカテゴリーが予測される個体数の期待値は

$$E_{pop}(E_{samp}(E_{pr}(I_3))) = \lambda \left\{ K_1 - \frac{K_1}{(1+N_0\beta)^{\frac{1}{K_1\beta}}} \right\} + \lambda N_0 \left(\beta' + \frac{1}{K_2} \right) \tag{4.10}$$

となる．この式も (4.7) 式と同様に β に関して単調減少で β' に関して単調増加，すなわち $V(\pi_i)$ が小さく，$V(\pi_j')$ が大きいほど期待値は大きくなる．しかしセンシティブ変数に関する総個体数が変動する影響から (4.10) 式右辺の後半の項に β が含まれず，センシティブ変数を考慮したことによるカテゴリーの予測される個体数の期待値の増加分が，キー変数に関する個体の入り方とは独立になってしまう．

なお (4.7) 式で，キー変数に関するモデルをポアソンガンマモデルから多項ディリクレモデルに代えても，$K_1/(1+N_0\beta)^{1/(K_1\beta)}$ の部分が

$$K_1\cdot\frac{\Gamma(K_1\gamma')\Gamma((K_1-1)\gamma'+N)}{\Gamma(K_1\gamma'+N)\Gamma((K_1-1)\gamma')}=K_1\cdot\prod_{l=0}^{N-1}\left(1-\frac{\gamma'}{K_1\gamma'+l}\right)$$

に変わるだけであるが，ポアソンガンマモデルと比較してその性質を議論するのはやや難しい．ただし γ' はディリクレ分布共通のパラメータである．多項ディリクレモデルの性質については Takemura (1999) を参照されたい．

4.4 ポアソンガンマ-多項分布モデル

前節ではセンシティブ変数によって分けられる2次セルへの個体の入り方について多項ディリクレモデルを用いたが，2次セル数がそれほど多くなく，しかもすべてのセルにおいて2次セルへの入り方が等しいと考えられる場合には，多項分布を用いることも可能である．また多項分布を用いたモデルでは2次セルを特定した議論ができるため，例えば就業状態という項目の完全失業というカテゴリーのように，あるセンシティブ変数で特にプライバシーの侵害につながるカテゴリーに属する個体数のみに限定した議論も可能となる．

多項分布をセンシティブ変数に対して適用する場合，(4.4)，(4.5) 式のうちで前者のみを用いることになる．すなわち π'_j はすべてのセルに共通である．これをポアソンガンマ-多項分布モデルと呼ぶことにすると，次の結果が得られる．

定理 5. ポアソンガンマ-多項分布モデルを用いた場合，$E_{samp}(E_{pr}(I_3))$ の母集団の超母集団からのサンプリングに関する期待値は

$$E_{pop}(E_{samp}(E_{pr}(I_3)))=\lambda\left\{K_1-\frac{K_1}{(1+N_0\beta)^{\frac{1}{K_1\beta}}}\right\}+\lambda\left(K_2S^2_{\pi'}+\frac{1}{K_2}\right)\left\{N_0-\left\{K_1-\frac{K_1}{(1+N_0\beta)^{\frac{1}{K_1\beta}}}\right\}\right\},\tag{4.11}$$

ただし

$$S^2_{\pi'}=\frac{1}{K_2}\sum_{j=1}^{K_2}\left(\pi'_j-\frac{1}{K_2}\right)^2\tag{4.12}$$

である．

証明. 多項ディリクレモデルの場合と同様に

$$\begin{aligned}E_{sens}(F^2_{i,j})&=V_{sens}(F_{i,j})+\{E_{sens}(F_{i,j})\}^2\\&=F_i\pi'_j(1-\pi'_j)+(F_i\pi'_j)^2\end{aligned}$$

を用いると，カテゴリーが予測される個体数の期待値は

$$\begin{aligned}&E_{pop}(E_{samp}(E_{pr}(I_3)))\\&=\lambda\cdot E_{key}\left[\sum_{\substack{i=1\\(F_i\neq0)}}^{K_1}\sum_{j=1}^{K_2}\frac{1}{F_i}\{F_i\pi'_j(1-\pi'_j)+(F_i\pi'_j)^2\}\right]\\&=\lambda\cdot E_{key}\left[\sum_{\substack{i=1\\(F_i\neq0)}}^{K_1}\left\{1+(F_i-1)\sum_{j=1}^{K_2}\pi'^2_j\right\}\right]\end{aligned}$$

$$= \lambda \cdot E_{key}\left[\sum_{\substack{i=1 \\ (F_i \neq 0)}}^{K_1}\left\{1+(F_i-1)\left(K_2 S_{\pi'}^2+\frac{1}{K_2}\right)\right\}\right]$$

$$= \lambda\left\{K_1-\frac{K_1}{(1+N_0\beta)^{\frac{1}{K_1\beta}}}\right\}+\lambda\left(K_2 S_{\pi'}^2+\frac{1}{K_2}\right)\left\{N_0-\left\{K_1-\frac{K_1}{(1+N_0\beta)^{\frac{1}{K_1\beta}}}\right\}\right\}$$

と求められる．□

この期待値の性質も多項ディリクレモデルを用いた場合と同様である．

また，センシティブ変数の特定のカテゴリーに含まれることがプライバシーの侵害につながるような場合を想定して，第 j 番目の2次セルに含まれる個体のうちでそのカテゴリーが予測される個体数 $I_{3,j}$ の期待値も

(4.13)
$$E_{pop}(E_{samp}(E_{pr}(I_{3,j})))$$
$$= \lambda \cdot E_{key}\left[\sum_{\substack{i=1 \\ (F_i \neq 0)}}^{K_1}\{\pi_j'(1-\pi_j')+F_i\pi_j'^2\}\right]$$
$$= \lambda\pi_j'(1-\pi_j')\left\{K_1-\frac{K_1}{(1+N_0\beta)^{\frac{1}{K_1\beta}}}\right\}+\lambda N_0\pi_j'^2$$

と求められる．

5. 適 用 例

5.1 個票データを標本として州全体の個体を予測する場合

この章ではアメリカ合衆国のセンサスからの抽出データに対して前章までの手法を適用し，その有用性をみるとともに，実際のデータへの応用における問題点やその対処方法などを考えていくことにする．

ここで使用するのはアメリカ合衆国で1990年に実施されたセンサスの1%抽出の個票データ (Bureau of the Census (1993)) で，これは5%抽出データとともに一般ユーザー向けに販売されているものである．個票データは各州ごとに分類され，それぞれには世帯と個人のレコードが含まれているが，ここでは個人のレコードのみを用いることにする．個人レコードは231桁で，年齢，性別，世帯主との続柄などの項目の他，人種，市民権や使用する言語に関する項目，兵役に関する項目，労働に関する項目，収入に関する項目などから構成されている．以下の検討では，年齢 (91カテゴリー)，性別 (2カテゴリー)，世帯主との続柄 (14カテゴリー)，配偶関係 (5カテゴリー)，出身地 (14カテゴリー) など15項目のみをキー変数として考えることにするが，そのカテゴリーの組み合わせの数は $K_1 = 4.603 \times 10^{12}$ である．またキー変数に含まれないセンシティブ変数として最近1週間の従業時間を考える．従業時間は1時間刻みで0時間から98時間と99時間以上のカテゴリーに分けられ $K_2 = 100$ となる．なおこの個票データはトップコーディングなどの目に見える秘匿措置の他にも，明らかにされていない秘匿措置が施されている可能性があるため，以下の検討では一つの題材としてこの個票データを用いるのに止め，個票データの実際の危険性などについて深い議論はしないことにする．

50州のうち比較的人口の多いメリーランド州 (人口 4.781×10^6)，ミズーリ州 (5.117×10^6)，ワシントン州 (4.867×10^6) の3州のデータを用いることにする．キー変数のみを用いた指標としては，標本でも母集団でも一意である個体数 ($\underline{S_1}$ と表記)，予測される個体数の期待値 ($E(I_1)$)，3.2節の例1の危険性の指標の期待値 ($E(I_{2,1})$)，例2の危険性の指標の期待値

($E(I_{2,2})$) について検討する．なお $E(I_{2,1})$ については $a=2$ とする．またセンシティブ変数も考慮した指標として，ポアソンガンマ-多項ディリクレモデルを用いた場合に，従業時間のカテゴリーが予測される個体数の期待値 ($E(I_3)$)，そのうち従業時間が 0 時間以外であるカテゴリーが予測される個体数の期待値 ($E(I_{3,1+})$)，更にポアソンガンマ-多項分布モデルを用いた場合に従業時間が 80 時間以上であることが予測される個体数の期待値 ($E(I_{3,80+})$) を考える．

未知のパラメータのうち，ポアソンガンマモデルの N_0 は母集団の大きさで代用するが，多項ディリクレモデルの γ，多項分布の π_j' の推定にはモーメント法を用いることにする．

多項ディリクレモデルでは π_j' がすべてのセルにおいて共通に決定されるモデルを考える．この場合，第 j 番目の 2 次セルに含まれる個体数の総和 $f_{\cdot,j}=\sum_{i=1}^{K_2} f_{i,j}$ が共通の多項ディリクレモデルから発生したと考えることができる．このとき $f_{\cdot,j}$ の平均と分散が

$$E(f_{\cdot,j})=\frac{n}{K_2} \tag{5.1}$$

$$V(f_{\cdot,j})=n\cdot\frac{1}{K_2}\left(1-\frac{1}{K_2}\right)\frac{K_2\gamma+n}{K_2\gamma+1} \tag{5.2}$$

と書かれるので，(5.2) 式の左辺を

$$\hat{V}(f_{\cdot,j})=\frac{1}{K_2-1}\sum_{j=1}^{K_2}\left(f_{\cdot,j}-\frac{n}{K_2}\right)^2 \tag{5.3}$$

で置き換えた式を γ について解いた式をその推定量とする．また (5.3) 式の標準誤差を推定し，(5.3) 式がその標準誤差の 1 倍または 2 倍程度大きく推定された場合，すなわち $\hat{V}(f_{\cdot,j})+\hat{SD}(\hat{V}(f_{\cdot,j}))$ または $\hat{V}(f_{\cdot,j})+2\cdot\hat{SD}(\hat{V}(f_{\cdot,j}))$ を用いて γ を推定した場合も合わせて考えることにする．4.3 節で述べたように γ が小さいほど $E(I_3)$ が大きくなるため，安全性を考慮して推定量を上方修正するのが目的である．

またセンシティブ変数に多項分布モデルを用いた場合に従業時間が 80 時間以上であることが予測される個体数の期待値では，80 時間以上のパラメータの和をその相対度数の実現値で置き換えることになる．この場合にも，多項ディリクレモデルと同様の上方修正を考えることにする．

ポアソンガンマモデルでも上記のモーメント法や最尤法によってパラメータ β を推定することは可能であるが，この個票データから得られる $s_0, s_1, \ldots$ の度数分布の s_1 から s_2 への減少の度合いが大きいため，モデルの当てはまりが良くない．そこで $\sum_{l=1}^{10}(s_l-\hat{s}_l)^2/\hat{s}_l$ を最小にするようにパラメータを推定し，含まれる個体数が小さいために危険性が大きいセル数の部分にモデルが適合するようにする．ただし s_l は実現値で $\hat{s}_l$ はポアソンガンマモデルによる推定値である．この場合にも便宜上，モーメント法において $V(f_i)$ の推定量からその標準誤差の推定量の 1 倍または 2 倍を引いたものを用いて β を推定した場合のこのパラメータの減少分を，上で推定した β から減じることによって上方修正を行うことにする．(他の超母集団モデルと合わせて，後の表ではこれらを「修正 1」，「修正 2」と略記する．) なお，どの州でも N_0 は K_1 と比較して十分小さく，$E(I_{2,1})$，$E(I_{2,2})$ も β が小さいほど大きくなる．

超母集団モデルを仮定した場合の $E(I_{3,1+})$ を除くポアソンガンマモデルと多項ディリクレモデルのパラメータの推定値と各指標の期待値の推定値を表 1 に示す．$E(I_{3,1+})$ については従業時間が 0 時間のデータを除いた個票データに対してモデルを当てはめるため，そのパラメータは他の指標に用いるものとは異なる．

「推定値」の列は推定値そのものをモデルに代入した場合の結果で，「修正 1」と「修正 2」の列は上述したように危険性を段階的に高めに評価した結果である．この個票データは実際に

表1. 各指標の期待値の推定値(個票データを標本として州全体の個体を予測する場合).

州	メリーランド			ミズーリ			ワシントン		
母集団の大きさ	4.781×10^6			5.117×10^6			4.867×10^6		
標本の大きさ	46424			50978			49045		
	推定値	修正1	修正2	推定値	修正1	修正2	推定値	修正1	修正2
$\beta \times 10^5$	4.140	3.207	2.273	5.830	4.269	2.709	4.340	2.980	1.621
$\gamma \times 10^2$	2.530	1.075	0.835	1.908	0.758	0.519	2.027	0.818	0.578
$\underline{S_1}$	233	301	423	170	232	365	231	336	614
$E(I_1)$	1241	1526	2006	974	1258	1816	1244	1685	2724
$E(I_{2,1})$	232	299	419	170	231	363	230	334	606
$E(I_{2,2})$	349	450	633	255	348	546	346	503	917
$E(I_3)$	14366	23400	26420	18495	29752	34342	17356	27955	32242
$E(I_{3,1+})$	5865	9541	10450	5006	8345	9564	5091	8304	9048
$E(I_{3,80+})$	8.18	10.43	11.14	7.31	9.57	10.20	10.04	13.71	14.56

表2. 母集団の値を用いた各指標の期待値(標本から元の個票データを予測する場合).

州	メリーランド		ミズーリ		ワシントン	
母集団の大きさ	46424		50978		49045	
標本の大きさ	9285		10196		9809	
	$E_{pr}(\cdot)$	$E_{samp}(E_{pr}(\cdot))$	$E_{pr}(\cdot)$	$E_{samp}(E_{pr}(\cdot))$	$E_{pr}(\cdot)$	$E_{samp}(E_{pr}(\cdot))$
$\underline{S_1}$	3435	3350	2743	2746	3297	3280
$E(I_1)$	4730	4682	4046	4052	4664	4659
$E(I_{2,1})$	2155	2129	1789	1790	2112	2107
$E(I_{2,2})$	4017	3960	3284	3290	3929	3917
$E(I_3)$	7097	7059	7369	7364	7499	7461
$E(I_{3,1+})$	2991	3019	2306	2340	2646	2659
$E(I_{3,80+})$	28.95	28.12	18.71	23.33	35.37	31.63

は母集団から層化無作為抽出されたもので，しかも前述したような明らかにされていない秘匿措置が施されている可能性があるが，それを無視して考えると，キー変数のみを考えた指標ではミズーリ州が最も安全と判断されるのに対して，センシティブ変数を考慮した$E(I_3)$では逆に最も危険と考えられるのは興味深い．ここで$E(I_3)$が非常に大きいのは従業時間が0，すなわち子供，主婦，失業者などが全体の半数程度を占めていることによるものである．ただし従業時間が0時間以外のカテゴリーが予測される個体数の期待値$E(I_{3,1+})$も$E(I_1)$に比べて値は非常に大きく，センシティブ変数を考慮することの重要性は認識させられる．なお$E(I_3)$, $E(I_{3,1+})$についてポアソンガンマ-多項分布モデルを用いても，例えばメリーランド州ではそれぞれ14240, 5817と推定され，多項ディリクレモデルを用いた値に近く，この程度の2次セル数であればどちらのモデルを用いても結果に大きな違いはないものと推測される．

5.2 個票データを母集団とするサンプリング実験

この節ではモデルの当てはまりや手法の修正方法などを検討するために各州の個票データを未知の母集団と考え，抽出率1/5で非復元無作為抽出した大きさ約1万の標本について，各指標の期待値を超母集団モデルを用いて推定することにする．この場合，実際には未知である母集団の値を用いることによって，例えばI_1については母集団と標本が決定したときの期待値

表3．サンプリング実験による各指標の期待値の推定値（標本から元の個票データを予測する場合）．

州	メリーランド			ミズーリ			ワシントン		
母集団の大きさ	46424			50978			49045		
標本の大きさ	9285			10196			9809		
	推定値	修正1	修正2	推定値	修正1	修正2	推定値	修正1	修正2
$\beta \times 10^5$	6.800	5.594	4.389	9.730	7.686	5.641	7.350	5.636	3.921
$\gamma \times 10^2$	2.497	1.057	0.817	1.884	0.746	0.507	2.013	0.811	0.571
$\underline{S_1}$	2234	2581	3057	1711	2073	2631	2130	2606	3356
$E(I_1)$	4191	4577	5063	3669	4145	4803	4155	4704	5471
$E(I_{2,1})$	1801	2020	2300	1465	1723	2091	1750	2059	2500
$E(I_{2,2})$	3082	3513	4082	2423	2899	3607	2964	3563	4459
$E(I_3)$	5684	6889	7405	5975	7637	8399	6069	7540	8242
$E(I_{3,1+})$	2861	3381	3639	2392	3025	3566	2602	3048	3241
$E(I_{3,80+})$	24.54	30.43	34.10	19.25	24.71	27.72	32.10	40.47	44.69

表4．サンプリング実験による各指標の期待値とその推定値（メリーランド州：副標本から標本を予測する場合）．

母集団の大きさ	9285				
標本の大きさ	1875				
	標本の値を利用		超母集団モデルを利用		
	$E_{pr}(\cdot)$	$E_{samp}(E_{pr}(\cdot))$	推定値	修正1	修正2
$\beta \times 10^5$			12.270	9.850	7.429
$\gamma \times 10^2$			2.590	1.104	0.864
$\underline{S_1}$	1153	1150	876	979	1110
$E(I_1)$	1394	1389	1252	1332	1426
$E(I_{2,1})$	674	671	597	643	697
$E(I_{2,2})$	1296	1287	1110	1213	1336
$E(I_3)$	1651	1657	1430	1593	1669
$E(I_{3,1+})$	785	787	763	843	899
$E(I_{3,80+})$	12.70	8.88	10.05	13.45	16.21

((2.1) 式)，母集団のみを固定した標本のサンプリングに関する期待値 ((2.2) 式) と比較することが可能となる．

表2の $E_{pr}(\cdot)$, $E_{samp}(E_{pr}(\cdot))$ の列にこの2種類の期待値を示している．抽出率が大きいために，2種類の期待値の差は $E(I_{3,80+})$ を除くとどれもかなり小さい．$E(I_{3,80+})$ の値の差は，標本において従業時間が80時間以上の2次セルに含まれる個体数が各州において15, 53, 75（個票データでは54, 295, 317）とやや少ないためと思われる．

次に，表3に標本の値を元に超母集団モデルによって推定された各指標の期待値の推定値を示す．この場合セル内の個体数が全体的に少なくなるため，ポアソンガンマモデルでは $\sum_{l=1}^{5}(s_l - \hat{s}_l)^2/\hat{s}_l$ を最小とするようにパラメータ β を推定することにする．

前述したようにこのデータへのポアソンガンマモデルの当てはまりがやや悪いため，β の値が個票データから推定される値よりも大きく推定され，推定値をそのまま用いてしまうと表2の値と比較して危険性を過小に評価する傾向のあることがわかる．しかし $\underline{S_1}$ を除くと表2の値は表3の「修正1」の値の付近にあり，この程度の修正を行うとかなり近い値が推定値として得られることがわかる．なお $\underline{S_1}$ は一意である個体数のみを対象にするため，モデルの部分的な

当てはまりの悪さが結果に顕著に現れることがあると考えられる．ここでは$\lambda \cdot N_0/(1+N_0\beta)^{1+1/(K_1\beta)}$を用いて推定しているが，佐井 (1998) で用いたように，実現値s_1にポアソンガンマモデルで推定した値の比$\hat{S}_1/\hat{s}_1$を乗じてS_1を推定する方法も効果的である．しかしこの例では著しい改善はない．

ここでメリーランド州について，この標本から更に抽出率1/5で副標本を非復元無作為抽出する．標本の値を用いた危険性の各指標の期待値と，超母集団モデルを用いたその推定値を表4に示す．

この場合の期待値とその推定値とのずれは，表2，3のずれとほぼ同じ動きをしており，その意味で個票データと標本との関係がほぼ維持されることがわかる．従って個票データを未知の母集団と考え，標本からこの母集団の予測を行う際の危険性を評価する場合にも，標本と副標本との関係と同様の上方修正を行うべきであることが推測できる．

なお元の個票データから州全体の母集団の予測を行う際の危険性の評価については，上の検討と抽出率が異なるために個票データと標本との関係がそのまま維持される保証はないが，「修正2」の程度，指標によってはそれ以上の十分な上方修正を行えば安全と思われる．

このように，実用可能性の有無やパラメータ推定の性質は，標本である個票データそのものから，または個票データからサブサンプリングを行うことによってある程度導き出すことが可能であるため，個別のデータに対応した対処が必要と考えられる．

6. おわりに

本論文では，キー変数のみを用いて予測される個体数，更にはキー変数とセンシティブ変数を同時に用いてセンシティブ変数のカテゴリーが予測される個体数について，ポアソンガンマモデルや多項ディリクレモデルなどの超母集団モデルを用いて検討を行った．提案した危険性の指標は大きく3つに分けることができる．

まず第2章では危険性の指標として，第三者がキー変数を用いて個票データのすべての個体について予測を行うことを仮定した場合に予測される個体数を提案し，標本のサンプリングを考慮するとその値は標本の抽出率と個体が入っているセル数のみで表せることを示すとともに，超母集団モデルとしてポアソンガンマモデルを仮定した場合のパラメータと期待値との関係を求めた．次に第3章では上記の指標を拡張して，キー変数のみに基づく危険性の一般的な指標を考案するとともに，実用性のある2つの例を考え，その性質について議論を行った．例えば表1と表3で，これらの期待値や標本でも母集団でも一意である個体数などの比を求めてもそれぞれ若干の違いがあり，危険性の定義の仕方によって個票データの評価の異なることが認識される．更に第4章では，あるセンシティブ変数のカテゴリーが予測される個体数を指標として提案し，センシティブ変数に多項ディリクレモデルと多項分布のモデルを用いた場合のパラメータとその期待値との関係について議論した．第5章の応用例のように，キー変数のみを用いた検討では危険と思われない場合でも，センシティブ変数を考慮することによって危険と判断されるケースが考えられ，この点を勘案することの重要性が認識される．

第5章でみたように，提案した手法は概ね実用可能であるもの，特にポアソンガンマモデルのパラメータの推定に対しては課題が残る．その推定には標本の性質や，場合によってはサブサンプリングを行って標本と母集団との関係を推測する必要もあろう．例えば個体数が小さくなるほどパラメータとして大きい値が推定されるような場合には，副標本と標本から推定される値を比較することによって，母集団の推定値を修正することも可能であろう．またモデルの当てはまりの悪さやパラメータの推定誤差を考慮に入れて危険性を十分大きく評価することも重要である．当然，汎用性のより高いモデルを仮定することも対処法の一つである．Hoshino

(2000) は従来用いられているモデルよりもエベンスモデルを拡張したピットマンモデル (Pitman model) の方が当てはまりがよい例を示している．ただし他のモデルを用いる場合にも，本論文で議論した性質はほぼ維持されると思われる．

謝　　辞

本論文の訂正にあたって査読者の方々からは適切なコメントを頂き，感謝致します．本論文は統計数理研究所共同研究プログラム (11-共研-2033) の研究成果に基づくものであり，文部省科学研究費補助金・課題番号 11480055「調査データ公有化における理論的課題」の援助も受けている．

参 考 文 献

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

Bureau of the Census (1993). 1990 Census of Population and Housing, Public Use Microdata Samples (microdata), Washington, D. C.

Dalenius, T. (1977). Towards a methodology for statistical disclosure control, *Statistisk Tidskrift*, **5**, 429-444.

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

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

Hoshino, H. (2000). On application of Pitman's sampling formula to microdata disclosure risk assessment,「調査データ公有化における諸問題」シンポジウム資料，1-14.

Hoshino, H. and Takemura, A. (1998). Relationship between logarithmic series model and other superpopulation models useful for microdata disclosure risk assessment, *J. Japan Statist. Soc.*, **28**, 125-134.

加納　悟 (1997)．ミクロ個票データの公開におけるリスク評価の理論，平成 8 年度科学研究費補助金報告書（加納班），1-22.

Keller, W. J. and Bethlehem, J. G. (1992). Disclosure protection of microdata: Problems and solutions, *Statist. Neerlandica*, **46**,.5-19.

Marsh, C., Skinner, C., 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.

Marsh, C., Dale, A. and Skinner, C. (1994). Safe data versus safe settings : Access to microdata from the British Census, *International Statistical Review*, **62**, 35-53.

Müller, W., Blien, U. and Wirth, H. (1995). Identification risks of microdata, evidence from experimental studies, *Sociological Methods and Research*, **24**, 131-157.

Omori, Y. (1997). Measuring identification disclosure risk for categorical microdata by posterior population uniqueness, Research Paper Series No. 6, Faculty of Economics, Tokyo Metropolitan University.

佐井至道 (1997)．データの統計的開示制限に関連した文献とその分類，平成 8 年度科学研究費補助金報告書（重点領域，竹村班），106-136.

佐井至道 (1998)．個票データにおける個体数とセル数との関係，応用統計学，**27**，127-145.

Särndal, C. E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer, New York.

渋谷政昭 (1999)．size index の推測，日本計量生物学会・応用統計学会合同年次大会予稿集，11-14.

Skinner, C. J. (1992). On identification disclosure and prediction disclosure for microdata, *Statist. Neerlandica*, **46**, 21-32.

Takemura, A. (1999). Some superpopulation models for estimating the number of population uniques, *Statistical Data Protection — Proceedings of the Conference, Lisbon*, 25 to 27 March 1998-1999

edition, 45-58, Office for Official Publications of the European Communities, Luxembourg.
Willenborg, L. and de Waal, T. (1996). *Statistical Disclosure Control in Practice*, Springer, New York.

Microdata Disclosure Risk Assessment Based on the Expected Number of Predicted Individuals

Shido Sai

(Okayama Shoka University)

In the microdata, if the combination of categories on the key variables of some individual is different from all other individuals', it is called unique. When the microdata which is sampled from the population is disclosed, the number of unique individuals both in the microdata and the population is usually used to assess the risk of disclosure. However, most individuals other than unique ones are not safe, so it may be necessary to consider the comprehensive index including uniqueness. In this article, new index, the expected number of the individuals which is predicted when the investigator tries to connect all individuals in the microdata to ones in the population using key variables, is proposed, and some properties about it are considered. The Poisson-gamma model is used as a superpopulation model for the key variables. Furthermore, the extension to the general index is carried out.

It is called prediction disclosure that microdata enables the investigator to predict the categories of the sensitive variables for individuals with some degree of confidence. In this article we investigate the expected number of individuals in which the categories of a sensitive variable is predicted, when the investigator tries to estimate them for all individuals in the microdata. To add to the Poisson-gamma model for the key variables, the multinomial distribution and the Dirichlet-multinomial models are adopted for the sensitive variable as a superpopulation model.

Several types of indexes proposed in this article are applied to the sample microdata set of Census in U.S. to make sure of usefulness of them and to find the problems to be solved.

Key words: Microdata, key variable, sensitive variable, prediction disclosure, Poisson-gamma model, Dirichlet-multinomial model.