個票データにおける個体数とセル数との関係

岡山商科大学 商学部 佐井 至道

Relations between the Number of Individuals and the Number of Cells in Microdata

Shido Sai

要 旨 個票データの開示を行う場合には、個人などのプライバシーの侵害が起こらないようにするため、事前にデータの中で一意である個体がどの程度存在するかを検討しなければならない。本論文では、データに含まれる個体数と、各個体を持っている属性の組み合わせの数であるセル数との比などを考慮に入れて、全数調査における母集団一意である個体数と、標本調査における標本でも母集団でも一意である個体数について、ポアソン分布を用いたモデルとポアソンガンマモデルのもとで検討を行った。また全数調査においてはその期待値、標本調査においてはその推定値の上限を与えるパラメータを求めた。これは例えば標本調査で得られた個票データの開示に向けて、やや厳しくまたは最も厳しく推定した値でもなお安全であることを、官庁統計の個票データの提供を現在検討している研究者や行政関係者、およびそれらの調査に被調査者として協力している人々に示す目的によるものである。本論文の後半では、このような手法を労働力調査の個票の集計データに対して適用し、その安全性について検討を行った。その結果、小地域に関する情報が削除されたものを使用すれば、十分実用に耐えうる、しかも安全なデータの公開が可能であることが示された。

1. はじめに

個票データの開示を考える際、データの中の個体である個人や事業所などのプライバシーに配慮する必要がある。個票データのレコードを構成する数個の変数（項目）のうち、個人の年齢や性別のように第三者が個体の特定に使用できる変数をキー変数 (key variable) と呼ぶが、その変数の組み合わせが他のすべての個体のものと異なる一意 (unique) な個体のレコードをそのまま公表してしまうと、第三者がその個体を識別 (identification) し、プライバシーに関わるセンシティブな変数 (sensitive variable) を知ってしまう可能性がある。

例えば母集団に属するすべての人のレコードからなる個票データがあるとする。もしその中に

Key words: microdata, Poisson-gamma model, population unique, sample unique, statistical disclosure control

37歳の男性が1人しかいなければその人は一意であり、もし第三者がその情報を持っているならば、データの中からその人のレコードを識別することは可能である。その結果同じレコードに含まれる年収や病歴、あるいは失業しているなどのセンシティブな情報を知ってしまう可能性がある。

したがって、母集団についての個票の公開を検討する場合には、上の例のように母集団内で他のすべての個体とキー変数の組み合わせが異なる母集団一意 (population unique) である個体数を十分検討する必要がある。標本についての個票の公開を検討する場合には問題はやや複雑である。標本の中で一意である個体でも母集団で一意でなければ第三者がその個体を識別することはできない。また母集団で一意である個体でも標本として抽出されなければ問題は起きない。したがって標本一意 (sample unique) かつ母集団一意である個体数がどの程度であるかを検討しなければならない。

このような検討の重要性については Willenborg and de Waal (1996) に詳しく述べられており、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 Paass (1988), Blien et al. (1992), Müller et al. (1995) がそれぞれ扱っている。これらの論文では、開示された個票データのレコードを、他のデータベースを利用して特定するマッチング (matching) の問題が主に扱われているが、Marsh et al. (1991) では、ターゲットとなるある1つの個体が個票データの開示によって識別される確率を、その個体が個票データに含まれる確率などの4つの確率の積として計算する方法を提案している。

また Bethlehem et al. (1990) は、ポアソンガンマモデル (Poisson-gamma model) を用いて、標本としてとられた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初めて提案した。Keller and Bethlehem (1992) もほぼ同じ内容を扱っている。また Skinner (1992) もポアソンガンマモデルについて触れるとともに、個票の開示によって個体のセンシティブな情報がある程度の信頼性で予測される予測開示 (prediction disclosure) の問題を扱っている。さらに 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いる。

これらとは別のアプローチとして、Greenberg and Zayatz (1992) はベイズの定理を用いて、母集団一意である個体数のノンパラメトリックな推定法を提案している。また母集団一意であるセル数が、ある決められた個数以下であるという仮説の尤度比検定を 加納 (1997) が扱っている。

このような検討の結果、個票データをそのまま開示することが困難であると考えられた場合、データの各項目について、それが質的変数であればカテゴリーを併合 (global recoding) したり、量的変数であれば丸め (rounding)、トップコーディング、ボトムコーディングを行ったり、さらには項目そのものを削除 (suppression) したりすることがある。これらの方法は分類されるカテゴリーの組み合わせの数 (セル数とよぶ) を減少させる方法としてまとめることができる。これらの方法については、Willenborg and de Waal (1996) が最も詳しい。また別の方法として、元のデータが母集団であれば標本を、標本であれば2次標本を抽出し、個体数そのものを減少させることによって、完全に識別される個体数を減少させる方法がある。

このような検討を行う際、個票データに含まれる個体数とそれらが分類されているセル数との関係を考えることが重要と思われる。渋谷 (1997) は多項分布における一意な個体数の性質について議論する中で、個体数とセル数との関係を重視している。

また一意である個体数が期待値や推定値として得られた場合、個票データの開示に消極的な人

たちの中には、その値について疑問視する者のいることが予想される。開示に向けては、やや厳しい推定を行ったり、場合によってはその上限を求めたりして、それでも安全であることを強調する必要があるだろう。de Waal (1993) は個体数とセル数を考慮に入れて、多項分布を用いたモデルで全数調査における一意な個体数の期待値の上限について議論を行った。

本論文では、ポアソン分布を用いたモデルとポアソンガンマモデルにおいて全数調査と標本調査の各ケースについて同様の問題の検討を行う。

第2章では定義の後に全数調査における問題を扱う。2.2節でポアソン分布を用いたモデルにおいて、母集団一意である個体数の期待値についての性質を求めるとともに、その上限をとるパラメータの組を求める。これはde Waal (1993) の考え方をポアソン分布を用いたモデルに適用するものである。次に2.3節では、より実用性の高いポアソンガンマモデルでの個体数とセル数についての検討を行い、母集団一意である個体数の期待値の上限をとるパラメータについても議論する。

第3章では母集団と標本との関係を考慮に入れて標本調査における問題の検討を行う。まず3.1節ではポアソン分布を用いたモデルで、標本の一意な個体数から、標本でも母集団でも一意な個体数を推定する際のいくつかの性質について考えるとともに、その推定値の上限を示す。次に3.2節において、ポアソンガンマモデルについて同様の議論を行う。このような検討はこれまで行われていなかった。個票データの公開が検討されている調査の多くが標本調査であることを考えると、その結論は全数調査のものよりも利用価値が高いと思われる。

また第4章では、得られた結果を 竹村 (1998) によって求められている労働力調査の個票の集計データに対して適用し、若干の秘匿処理を行うことによって、その公開が可能であることを示す。

2. 母集団一意である個体数について

2.1. 定義

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 ($K \geq 2$) のセルに分けられているものとする。第 i セルに含まれる個体数を $F_i (i=1, 2, \dots, K)$ 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。さらに含まれる個体数が j のセル数、すなわち $F_i=j$ となるセル数を $S_j (j=1, 2, \dots, N)$ とする。このとき、最も問題となるのはサイズが1のセルの数、すなわちユニークセル数 S_1 である。標本では F_i, S_j の代わりに f_i, s_j という表記を用いる。なお、個体自体を呼ぶときには一意という単語を用いていくが、セルに対してはユニークセルという単語を用いることにする。

2.2. ポアソン分布を用いたモデルについて

この節ではポアソン分布を用いたモデルにおいて、ユニークセル数の期待値 $E(S_1)$ を中心として議論していく。以下の内容は個票データが母集団でも標本でも同様に議論できるが、ここでは母集団の表記を用いる。

本論文で考えているような問題において最も自然なモデルは多項分布であろうが、ポアソン分布から導かれる結果が多項分布の結果とほとんど変わらず、しかも扱いやすいため、ここでは個体がポアソン分布によってセルに落ちるモデルについて考える。なお、多項分布とポアソン分布

を用いた結果の類似性については、佐井 (1998) を参照されたい。また、多項分布とそれを拡張した多項ディリクレモデルについては Omori (1997), Takemura (1997) を参照していただきたい。

母集団の個体数 F_i がそれぞれ独立に次のようにポアソン分布に従うものとする。

$$p(F_i) = \frac{e^{-N_0\pi_i}(N_0\pi_i)^{F_i}}{F_i!}. \quad (1)$$

ただし、 $\pi_i \geq 0$ で $\sum_{i=1}^K \pi_i = 1$, また $N_0 > 0$ とする。 π_i は第 i セルへの個体の入りやすさを表すパラメータであり、 $E(F_i) = N_0\pi_i$ である。したがって全体の個体数 N は一定でないものの、その期待値は N_0 となる。ユニークセル数の期待値は

$$E(S_1) = N_0 \sum_{i=1}^K \pi_i e^{-N_0\pi_i} \quad (2)$$

となる。

特殊な場合としてすべてのパラメータが等しい場合、すなわち、 $\pi_1 = \dots = \pi_K = \frac{1}{K}$ を考えると $E(S_1) = N_0 e^{-\frac{N_0}{K}}$ となる。これがセル数 K に関して単調増加であることは明らかであり、 K を固定すると $N_0 = K$ のときに $E(S_1) = \frac{K}{e}$ で最大となることも容易にわかる。 $\pi_1 = \dots = \pi_K = \frac{1}{K}$ の場合に N_0, K の値を 100 以下で変化させたときのユニークセル数の期待値 $E(S_1)$ の値を図 1 に示す。最大値 ((100, 100) の点) を基準として、その 0.1 倍刻みに 10 段階表示している。濃いほど値は大きい。

なお $\pi_1 = \dots = \pi_K = \frac{1}{K}$ の場合には $E(S_1)$ に関して次のような性質が得られる。 K の値を固定したときに $N_0 = N_1, N_2 (N_1 < K < N_2)$ に対して $E(S_1)$ が等しいとする。すなわち $N_1 e^{-\frac{N_1}{K}} = N_2 e^{-\frac{N_2}{K}}$ とする。ここで、 $N_1 = a_1 K, N_2 = a_2 K (a_1 < 1 < a_2)$ と置き直すと、 a_1, a_2 は次のような K を含まない関係式を満たすことがわかる。

図 1. ポアソン分布 ($\pi_i = \frac{1}{K}$) における $E(S_1)$ の値 ($N_0 \leq 100, K \leq 100$)

$$\frac{\log a_2 - \log a_1}{a_2 - a_1} = 1. \quad (3)$$

例えば $a_1=0.01$ のときに (3) 式から $a_2=6.485$ と得られることから、 $K=1000$ の場合に $N_0=10$ と $N_0=6485$ のユニークセル数の期待値がほぼ等しくなることがわかる。すなわち個体数の期待値を 6485 から小さくしていくと、1000 までは $E(S_1)$ の値は増加しその後減少するものの、大きさが 10 となるまでは元の値より大きいことになる。したがって N_0 が K と比較して非常に大きい場合、個体数を減らすことによってユニークセル数を減らすのは困難であることがわかる。

次に $E(S_1)$ を最大にするパラメータの組 $(\pi_1, \dots, \pi_K)$ を考えると、次のような定理が得られる。

定理 1

$E(S_1)$ を最大にするような π_i の値は高々 2 個からなり

- (i) $N_0 \leq K$ の場合、 $\pi_1 = \dots = \pi_K = \frac{1}{K}$ のときに $E(S_1)$ は最大
- (ii) $K < N_0 \leq 2K$ の場合については不明
- (iii) $2K < N_0$ の場合、 $\pi_1, \dots, \pi_K$ のうち $K-1$ 個が $\pi_{(1)}$ 、1 個が $\pi_{(2)} \left(\pi_{(1)} < \frac{1}{K} < \pi_{(2)} \right)$ をとるときに $E(S_1)$ は最大

証明は付録に示す。(i) の領域では $\pi_1 = \dots = \pi_K = \frac{1}{K}$ のときに、 $E(S_1)$ が最大で $N_0 e^{-\frac{N_0}{K}}$ となるので、個々のパラメータの推定が困難な場合にはこの値を上限として考えることができる。

なお定理 1 は、多項分布を用いたモデルにおける de Waal (1993) の定理と、それを修正した佐井 (1998) の結果にほぼ一致している。佐井 (1998) では、場合分けを (i) $N \leq K+1$, (ii) $K+1 < N \leq 2K$, (iii) $2K < N$ として、多項分布を用いたモデルでも定理 1 と同様の結果が得られている。

2.3. ポアソンガンマモデルについて

上で考えたポアソン分布では、パラメータ $\pi_1, \dots, \pi_K$ の数が一般に極めて多いため、得られたデータからそれらを推定することは困難である。そのためポアソン分布のパラメータがガンマ分布に従う、次式のようなポアソンガンマモデルが Bethlehem et al. (1990) によって提案されており、Keller and Bethlehem (1992), Skinner (1992) においても用いられている。ポアソンガンマモデルと他のモデルとの関係については Takemura (1997), 星野・竹村 (1998) を参照されたい。

$$p(F_i) = \frac{e^{-N_0 \pi_i} (N_0 \pi_i)^{F_i}}{F_i!} \quad (4)$$

かつ

$$\pi_i \sim \text{gamma}(\alpha, \beta). \quad (5)$$

ただし $\alpha\beta = \frac{1}{K}$ である。したがって $\beta \rightarrow 0$ のとき、このモデルはポアソン分布に一致する。また π_i の平均と分散が $E(\pi_i) = \frac{1}{K}$, $V(\pi_i) = \frac{\beta}{K}$ となり、ある数個のセルにおけるパラメータ π_i のみが極端に大きいような場合は表現できないものの、現実のデータにかなり近い状況での議論が可能であろう。

ポアソンガンマモデルにおいて、ユニークセル数の期待値は

$$E(S_1) = \frac{N_0}{(1 + N_0\beta)^{1 + \frac{1}{K\beta}}} \quad (6)$$

と N_0, K, β を用いて書くことができる。 N_0 と β を固定した場合、セル数 K に関して単調増加であることは明らかである。また K, β を固定すると $N_0 = K$ のときに最大となる。

$\beta = 0.001$ の場合に N_0, K の値を 100 以下で変化させたときの、ユニークセル数の期待値 $E(S_1)$ の値を図 2 に示す。

図 2. ポアソンガンマモデルにおける $E(S_1)$ の値 ($\beta = 0.001, N_0 \leq 100, K \leq 100$)

次に N_0, K を固定した場合に $E(S_1)$ を最大にする β の値を求める。これはポアソン分布を用いたモデルで $E(S_1)$ を最大にするパラメータの組 $(\pi_1, \dots, \pi_K)$ を求めたことに対応している。

定理 2

- (i) $N_0 \leq 2K$ の場合、 $\beta = 0$ のときに $E(S_1)$ は最大
- (ii) $2K < N_0$ の場合、 $\beta = \beta_c$ ($\beta_c \neq 0$) のときに $E(S_1)$ は最大

ただし β_c は方程式 $N_0\beta(K\beta + 1) - (1 + N_0\beta)\log(1 + N_0\beta) = 0$ の正の解である。

証明は付録に示す。

ポアソン分布を用いたモデルの結果と比較すると、ポアソン分布の (i), (ii) の範囲が上の結果では (i) にまとめられている。これはポアソンガンマモデルが実質的に 1 個のパラメータ β しか持たないため、ポアソン分布のパラメータの組 $(\pi_1, \dots, \pi_K)$ のすべてのパターンを表現できていないことによるものであり、特にある 1 個の π_i のみが極端に大きい場合を表現できないことによるものである。したがってポアソンガンマモデルではポアソン分布と比較してやや狭い範囲で $E(S_1)$ の最大値を探していると言えよう。ただし第 4 章でみるように、実際のデータへの当てはまりはポアソンガンマモデルでも十分なようである。

3. 標本でも母集団でも一意である個体数の推定について

3.1. ポアソン分布を用いたモデルについて

一般に調査は標本調査で行われることが多いが、標本の中で一意である個体でも母集団で一意でなければ第三者がその個体を識別することはできない。また母集団で一意である個体でも標本として抽出されなければ問題は起きない。したがって標本調査で最も問題となるのは、標本で一意でしかも母集団でも一意である個体数である。標本から母集団の一意な個体数を推定、検定する問題は、ポアソングンマモデルについて Bethlehem et al. (1990), Keller and Bethlehem (1992), Skinner (1992) によって検討されており、多項ディリクレモデルについては Omori (1997) によって議論されている。またノンパラメトリックな推定方法を Greenberg and Zayatz (1992) が考案しており、加納 (1997) もノンパラメトリックな尤度比検定を提案している。

この節ではまずポアソン分布を用いたモデルにおいて標本でも母集団でも一意である個体数の推定について考える。

(2) 式を母集団のユニークセル数の期待値とする。ここで標本の抽出は母集団のそれぞれの個体が抽出率 λ で抽出されるベルヌーイ抽出で行われるものとする。(ベルヌーイ抽出については Särndal et al. (1992) の説明が詳しい。) また標本のユニークセル数を s_1 と定義するとその期待値は

$$E(s_1) = \lambda N_0 \sum_{i=1}^K \pi_i e^{-\lambda N_0 \pi_i} \quad (7)$$

となる。したがって、標本でも母集団でも一意である個体数を $\underline{S}_1$ とすると、その推定値 $\hat{\underline{S}}_1$ は (2) 式と (7) 式を用いて

$$\hat{\underline{S}}_1 = s_1 \cdot \frac{E(s_1)}{E(S_1)} \cdot \lambda = s_1 \cdot \frac{\sum_{i=1}^K \pi_i e^{-N_0 \pi_i}}{\sum_{i=1}^K \pi_i e^{-\lambda N_0 \pi_i}} \quad (8)$$

と考えられる。右辺の s_1 の係数は N_0 に関する単調減少関数で、 $N_0 \rightarrow 0$ のとき係数 $\rightarrow 1$, $N_0 \rightarrow \infty$ のとき係数 $\rightarrow 0$ となる性質は認められるものの、パラメータ $\pi_1, \dots, \pi_K$ が未知のままではこれ以上の性質を導き出すのは困難である。

ここで、 $\hat{\underline{S}}_1$ が最大になるようなパラメータの組 $(\pi_1, \dots, \pi_K)$ を求めることにする。ただし (8) 式を最大にするパラメータの組を求めるのは困難であるため、(2) 式を N_0 に関して偏微分した式

$$\frac{\partial E(S_1)}{\partial N_0} = \sum_{i=1}^K \pi_i (1 - N_0 \pi_i) e^{-N_0 \pi_i} \quad (9)$$

の最大値について考える。(8) 式から $E(S_1)/E(s_1)$ が大きいほど $\hat{\underline{S}}_1$ は大きくなるが、この比は一般のパラメータの組 $(\pi_1, \dots, \pi_K)$ について図1のようなグラフを描き、それを等高線と見なしたときに、標本と母集団を示す2点 $(\lambda N_0, K)$, (N_0, K) の高さの比と考えることができる。この2点を結ぶ線分は N_0 軸に平行より、この線分上のすべての点で勾配、すなわち (9) 式を最大にする条件が一定であれば、その条件の下で $E(S_1)/E(s_1)$ も最大となる。なお N_0 は個体数の期待値であるが母集団とは特定しないものとする。このとき次のような定理が得られる。

定理 3

$\frac{\partial E(S_1)}{\partial N_0}$ を最大とするような π_i の値は高々 3 個からなり

(i) $N_0 < t \cdot K$ の場合, $\pi_1 = \dots = \pi_K = \frac{1}{K}$ のときに $\frac{\partial E(S_1)}{\partial N_0}$ は最大

(ii) $t \cdot K \leq N_0$ の場合については不明

ただし t は等式 $t^2 - 3t + 1 = 5e^{t-4}$ の $0 < t$ における解で $t \doteq 0.3225$ である.

証明は付録に示す. (i) の領域に母集団と標本の点があれば, $\pi_1 = \dots = \pi_K = \frac{1}{K}$ のときに (8) 式が最大で $s_1/e^{\frac{N_0}{K}(1-\lambda)}$ となるので, 個々のパラメータの推定が困難な場合にはこの値を上限として考えることができる.

3.2. ポアソンガンマモデルについて

この節ではポアソンガンマモデルについて前節と同様の議論を行う.

(6) 式を母集団のユニークセル数の期待値とし, 抽出率 λ のベルヌーイ抽出でとられた標本のユニークセル数の期待値を

$$E(s_1) = \frac{\lambda N_0}{(1 + \lambda N_0 \beta)^{1 + \frac{1}{K\beta}}} \quad (10)$$

とする. このとき, 標本でも母集団でも一意な個体数の推定値 $\hat{S}_1$ は前節と同様に

$$\hat{S}_1 = s_1 \cdot \left(\frac{1 + \lambda N_0 \beta}{1 + N_0 \beta} \right)^{1 + \frac{1}{K\beta}} \quad (11)$$

と求められる. したがってパラメータ β が既知, または高い精度で推定可能ならば $\hat{S}_1$ を推定できる. (11) 式の s_1 の係数は, $N_0 \rightarrow 0$ のとき係数 $\rightarrow 1$ でポアソン分布の結果に等しくなる.

次に, 前節と同様に

$$\frac{\partial E(S_1)}{\partial N_0} = \frac{1 - \frac{N_0}{K}}{(1 + N_0 \beta)^{2 + \frac{1}{K\beta}}} \quad (12)$$

を最大にするパラメータ β を求めると次のようになる.

定理 4

(i) $N_0 < K$ の場合, $\frac{\partial E(S_1)}{\partial N_0}$ は β に関して単調減少で, $\beta = 0$ のときに最大値 $\left(1 - \frac{N_0}{K}\right)e^{-\frac{N_0}{K}}$ をとる

(ii) $N_0 = K$ の場合, β に関わらず $\frac{\partial E(S_1)}{\partial N_0} = 0$

(iii) $K < N_0 \leq 4K$ の場合, $\frac{\partial E(S_1)}{\partial N_0}$ は β に関して単調増加で, $\beta \rightarrow \infty$ のときに最大値 0 をとる

(iv) $4K < N_0$ の場合, $\frac{\partial E(S_1)}{\partial N_0}$ は $\beta < \beta_d$ で単調減少, $\beta_d < \beta$ で単調増加で, $\beta \rightarrow \infty$ のときに最大値 0 をとる

ただし, β_d は等式 $N_0 \beta (2K\beta + 1) - (1 + N_0 \beta) \log(1 + N_0 \beta) = 0$ の正の解である.

証明は付録に示す。

したがって、ポアソンガンマモデルを適用することには問題がないが、標本が小さいなどの理由で β を十分高い精度で推定することが困難な場合には、 β を (i) の領域では十分小さめに、(iii) の領域では十分大きめに推定することによって、 $\hat{S}_1$ をやや厳しく求めることができる。なお (iv) の領域では、 β を β_a から遠ざける方向にずらして推定することになる。仮に N_0 と λN_0 が複数の領域にまたがってもそれぞれの領域ごとに分けて考えることが可能である。

また、何らかの理由でパラメータ β を推定することが困難な場合には、上記の各範囲における最大値をとる β を用いることによって $\hat{S}_1$ の上限を求めることができる。

4. 労働力調査についての考察

4.1. 個票データについての考察

この節では労働力調査の個票の集計データについて、主にポアソンガンマモデルについての結果を用いながら考察を行う。

労働力調査は層化2段抽出法によって行われ、調査対象は15歳以上の者で、毎月約4万世帯の約10万人が調査されている。ここでは、竹村(1998)によって求められている関東1都6県と山梨県、長野県の合計9都県における1995年1月調査分の個票の集計データについて検討を行う。

表1. 各項目の分類数とセル数

項 目	検討1	検討2	検討3
(A) 都道府県番号	(9)	(9)	(9)
(B) 県内一連番号	824	824	824
(C) 15歳以上人員計	8	<u>3</u>	<u>3</u>
(D) 15歳未満人員 (男計)	6	<u>3</u>	<u>2</u>
(E) 15歳未満人員 (女計)	5	<u>3</u>	上に併合
(F) 続柄	12	<u>5</u>	<u>5</u>
(G) 性別	2	2	2
(H) 換算年齢	100	<u>20</u>	<u>10</u>
(I) 配偶関係	4	4	4
セル数	1898496000	17798400	1977600

項 目	検討4	検討5	検討6
(A) 都道府県番号	9	9	9
(B) 県内一連番号	<u>x</u>	<u>x</u>	<u>x</u>
(C) 15歳以上人員計	8	8	8
(D) 15歳未満人員 (男計)	6	6	<u>6</u>
(E) 15歳未満人員 (女計)	5	5	上に併合
(F) 続柄	12	12	12
(G) 性別	2	2	2
(H) 換算年齢	100	<u>20</u>	<u>20</u>
(I) 配偶関係	4	4	4
セル数	20736000	4147200	829440

集計データは表 1, 2, 3 のすべてと表 4~9 中の標本の一意な個体数の部分である。労働力調査年報 (1996) によると, 1995 年 1 月~3 月のこの地域の 15 歳以上人口を 3,585 万人としているので, ポアソン分布の母集団の個体数の期待値を $N_0=35850000$ とする。なおこの個票の集計データは欠損値のあるレコードを除く 27,158 人分のレコードを用いたものである。したがって $\lambda N_0=27158$ より $\lambda=0.0007575$ と考えられる。労働力調査では非復元の層化 2 段抽出法によってサンプリングが行われているが, ここではベルヌーイ抽出されたものとする。母集団, 標本とも十分大きいので, 非復元抽出をベルヌーイ抽出と解釈してもほとんど問題はないと考えられるが, 層化抽出の影響を無視することにはやや問題があると思われる。また個人レコードを世帯単位で利用する点についても今後の課題となろう。

各個人レコードは 200 桁程度からなるが, これから議論の対象となる主要部分は定期調査項目から世帯符号などを除いた 32 桁, 22 項目である。この中で, 第三者が個人を識別する際に用いる可能性があるキー変数として, 表 1 にあげる 9 項目を考えている。

検討 1 の分類数 $K=1898496000$ はこの標本におけるものであり, 母集団では例えば (C) などの分類数がさらに多くなる可能性がある。(なお (B) は (A) を含んで 824 分類である。)

一意性の問題を扱う上ではセンシティブな変数の項目を指定する必要はないが, ここでは特に就業状態 ((J) とする) に注目する。これは他の項目をこの項目と比較するとセンシティブな度合いがかなり低いためである。(J) は表 2 の 9 カテゴリーに分類されているが, その中で (d) 休業と (e) 完全失業であることが第三者に知られてしまうことが特に問題であり, 仮に (a) 仕事主であることがわかって大きな問題とはならない。つまりある個人がキー変数の組み合わせから特定された場合, その人が失業中や休業中であることが知られることは一般にプライバシーの侵害と考えられるが, 仕事をしているのは通常の状態であり, それを知られたくないということは考えにくい。すなわちセンシティブなのは項目ではなく項目の分類であると考えられることもできよう。このような考え方も含めて検討を行っていく。

検討 1

まずポアソンガンマモデルの当てはめを考える。ガンマ関数のパラメータ β の推定方法としては Bethlehem et al. (1990) の提案しているモーメント法を用いる。パラメータ β の推定量は次式である。

$$\hat{\beta} = \frac{1}{\lambda N_0} \left(\frac{K}{\lambda N_0} s_f^2 - 1 \right), \quad (13)$$

表 2. 就業状態の度数分布

(J)	就業状態	度 数
(a)	仕事主	14034
(b)	通学主	343
(c)	家事主	2116
(d)	休業	294
(e)	完全失業	503
(f)	通学	2510
(g)	家事	4390
(h)	その他	2949
(i)	不詳	19
合 計		27158

ただし

$$s_f^2 = \frac{1}{K-1} \sum_{i=1}^K (f_i - \bar{f})^2, \quad (14)$$

$$\bar{f} = \frac{1}{K} \sum_{i=1}^K f_i \quad (15)$$

である。なお f_i は標本において第 i セルに含まれる個体数である。したがって β の推定には標本の度数のみが必要で, 個票データそのものは必要でない。

セルに含まれる個体数で分けたセルの度数は竹村 (1998) によって表 3 のように求められている。個体数 1 の度数 25046 が標本のユニークセル数 s_1 である。

表 3. セルに含まれる個体数とその度数

個体数	度 数	個体数	度 数
0	1898470077		
1	25046	11	2
2	753	12	1
3	66	13	1
4	20	14	0
5	11	15	0
6	9	16	0
7	4	17	0
8	2	18	1
9	2	19	1
10	2	20	0
		25	1
		28	1

この結果を用いてパラメータ β を推定すると $\hat{\beta}=7.7928 \times 10^{-6}$ となる。そこで (11) 式を用いて、標本でも母集団でも一意である個体数は $\hat{S}_1=108.20$ と推定される。なお、以後の検討を含めて、ポアソンガンマモデルのデータへの当てはまりはよいようである。

ところで推定量 $\hat{\beta}$ の分散の推定量が

$$\hat{V}(\hat{\beta}) = \frac{K^2}{(\lambda N_0)^4} \hat{V}(s_f^2), \quad (16)$$

ただし

$$\hat{V}(s_f^2) = \left(\frac{K}{K-1} \right)^2 \frac{1}{K} \left[\frac{1}{K} \sum_{i=1}^K (f_i - \bar{f})^4 - \left\{ \frac{1}{K} \sum_{i=1}^K (f_i - \bar{f})^2 \right\}^2 \right] \quad (17)$$

と求められるので、この式を用いて推定量 $\hat{\beta}$ の標準誤差を表 3 から推定すると、 $\widehat{SD}(\hat{\beta})=1.6248 \times 10^{-6}$ となる。 $N_0 < K$ の範囲では β の値が小さいほど $\hat{S}_1$ が増加することから、例えば $\hat{\beta}$ からその標準誤差の 3 倍を引いた値を $\hat{\beta}$ の代わりに用いて、やや厳しい推定を行うこともできる。このとき $\hat{\beta} - 3 \cdot \widehat{SD}(\hat{\beta}) = 2.9183 \times 10^{-6}$ で、 $\hat{S}_1 = 255.71$ となり推定値は 2 倍以上に増加する。

さらに厳しく、上記の上限の取り方では不十分と考えるならば、 $N_0 < K$ におけるいわば最悪のモデルである $\beta=0$ の場合を考えなければならない。このとき標本でも母集団でも一意である個体数は $\hat{S}_1=24577.84$ と推定されるが、これはポアソン分布を用いたモデルでの $N_0 \leq 0.3225K$ における $\hat{S}_1$ が最大となるケースでもある。

以上の議論は個票データ全体に対して行ったものであるが、前に述べたように (J) 就業状態が (d) 休業と (e) 完全失業である場合のみがセンシティブと考え、その中で標本でも母集団でも一意である個体数を推定することもできる。上で検討した、 $\hat{\beta}$ をそのまま用いる方法、標準誤差の 3 倍を引いた値を用いる方法、 $\beta=0$ とする方法について、竹村 (1998) によって集計されている標本で一意な個体数 278,480 を用いて $\hat{S}_1$ を求めると次のようになる。(個票データ全体についての結果も再掲する。)

この 2 つの分類に絞ると、標準誤差の 3 倍を引いた値を用いる場合でも期待値は比較的小さいと思われる。

表 4. 検討 1 における標本でも母集団でも一意である個体数の推定値

標本の一意な個体数	全 体	(d) 休業	(e) 完全失業
	25046	278	480
$\hat{\beta}$	108.20	1.20	2.07
$\hat{\beta}-3 \cdot \text{標準誤差}$	255.71	2.84	4.90
$\beta=0$	24577.84	272.80	471.03

4.2. 安全な個票データの開示法について

検討 1 で見たように、個票データ全体を考えた場合には標本でも母集団でも一意である個体数は決して少なくない。そこで、この節では個票データとしての利用価値を失わない範囲でセルの併合を検討する。図 1, 2 からわかるように、個体数がセル数よりも十分小さい場合には、セルの併合よりも標本から 2 次標本をとる方がユニークセルを減らす効果は大きい、各個体から少量の情報を失う方が個体そのものを失うよりも良いと判断した。

セルの併合を行う際、特に問題となるのが地域変数の扱いである。キー変数とした 9 項目のうち (A) 都道府県番号と (B) 県内一連番号 が地域変数で、他の 7 項目は個人（または世帯）に関する変数である。(B) は市町村コードとは異なるものの、県内の地域を特定する重要な地域変数である。個票データの利用者の中に、小地域に関する情報に対して興味を持つ者がいることは事前に予想されるが、その分類が非常に細かいため、前節の検討 1 における一意な個体数を増加させる要因となっている。

そこで表 1 のように、検討 2, 3 では地域に関する情報を重視する立場で、検討 4, 5, 6 では個人に関する情報を重視する立場で竹村(1998)によってそれぞれ段階的にセルが併合された集計データを用いて、標本でも母集団でも一意な個体数の推定を行っていく。なお分類数にアンダーラインが引かれた項目は検討 1 の状態から分類数が減少していることを示す。また $\times$ は項目が削除されたことを表す。なお検討 2~6 において、検討 1 の表 3 に対応する度数分布については省略する。

検討 2

項目 (C), (D), (E) についてはトップコーディングが行われ、(F) 続柄については近い性質のカテゴリーが併合された。また (H) 換算年齢は 5 歳刻みにまとめられた。カテゴリーの併合によってセル数は $K=1898496000$ から $K=17798400$ に減少した。併合後の検討を行う前に、2.3 節の (6) 式から標本のユニークセル数を予測してみると 22383.30 である。大きな効果は期待できそうにないが、前節と同様の検討を行うことにする。

変則的なセルの併合であるため、 β の値は改めて推定した。また K の値が小さくなったことによって、 $\lambda N_0 < K < N_0 \leq 4K$ となり、 $(\lambda N_0, K)$ と (N_0, K) を結ぶ線分は定理 4 の場合分けの (i) と (iii) の領域にまたがるため、やや厳しい推定を行う場合には、(i) の範囲では $\hat{\beta}$ からこの標準誤差の 3 倍を引いた値を、(iii) の範囲では $\hat{\beta}$ にこの標準誤差の 3 倍を足した値をそれぞれ用いた。下表では $\hat{\beta} \pm 3 \cdot \text{標準誤差}$ と表示している。最も厳しい推定の場合も同様で、(i) の範囲では $\beta=0$ 、(iii) の範囲では $\beta=\infty$ として $\hat{S}_1$ の推定を行った。結果は次の通りである。

カテゴリーの併合前と比較すると、最も厳しい推定においてかなりの改善が見られるが、データの開示に消極的な者からは危険性が指摘されうる数値にとどまっている。

表 5. 検討 2 における標本でも母集団でも一意である個体数の推定値

標本の一意な個体数	全 体	(d) 休業	(e) 完全失業
	18275	222	356
$\hat{\beta}$	33.14	0.40	0.65
$\hat{\beta} \pm 3 \cdot \text{標準誤差}$	38.61	0.47	0.75
$\beta = 0, \infty$	3342.85	40.61	65.12

検討 3

次に、個人情報に関する項目について可能な限りのセルの併合について考える。(D) と (E) の 2 つの項目が併合され、しかも分類が「いる」と「いない」の 2 つにまで減らされた。また (H) 換算年齢は 10 歳刻みとされた。地域情報に関する項目を残す場合、これが個票データの利用価値を維持しうるセルの併合の限界と思われる。上と同様の検討を行って次の結果を得た。なおこの場合、 $\lambda N_0 < K$, $4K < N_0$ となり、定理 4 の場合分けの (i), (iii), (iv) の 3 つの領域にまたがるが、(iv) についても (iii) と同様に、やや厳しい推定を行う際には $\hat{\beta}$ にその標準誤差の 3 倍を足したものをうい、最も厳しい推定を行う際には $\beta = \infty$ とした。 $\hat{\beta} = 4.87299 \times 10^{-5}$ が (iv) の範囲のすべての N_0 に対する β_d の値より大きくなることが確かめられるため、 $\hat{\beta}$ を大きくすればよいことがわかる。

表 6. 検討 3 における標本でも母集団でも一意である個体数の推定値

標本の一意な個体数	全 体	(d) 休業	(e) 完全失業
	12919	162	281
$\hat{\beta}$	16.03	0.20	0.35
$\hat{\beta} \pm 3 \cdot \text{標準誤差}$	17.85	0.22	0.39
$\beta = 0, \infty$	265.80	3.33	5.78

全体に対する最も厳しい推定の結果を除くと十分小さい値が得られている。この個票の集計データからは、ポアソンガンマモデルにおいて最も厳しい推定で用いられている β の値をとる可能性がほとんどないことを考えると、十分実用に耐えうると判断できる。しかし、それでも開示した場合のリスクを懸念する者が多いとすると、地域情報を重視して開示を行うことは不可能であろう。

検討 4

次に個人に関する情報を重視する立場で、元の個票データから (B) 県内一連番号をまず初めに削除し、その後、他の個人情報に関する項目について、どの程度セルの併合を行えば $\hat{S}_1$ の値を許容限度に抑えることができるのか検討することにする。(B) が削除された結果は次のようである。

(d) 休業、(e) 完全失業の分類の $\hat{\beta}$ をそのまま用いる方法、標準誤差の 3 倍を引いた値を用いる方法では値は十分小さいものの、その他の値からはまだ危険性が感じられる。

次に個人情報に関する項目について段階的に行われたセルの併合について考える。これは、個票データの都道府県単位の利用が一般的に多いため、(A) 都道府県番号については削除あるいは併合が難しいことによるものである。

表 7. 検討 4 における標本でも母集団でも一意である個体数の推定値

	全 体	(d) 休業	(e) 完全失業
標本の一意な個体数	8049	99	147
$\hat{\beta}$	7.32	0.09	0.13
$\hat{\beta} \pm 3 \cdot \text{標準誤差}$	7.54	0.09	0.14
$\beta = 0, \infty$	1714.95	21.09	31.32

検討 5

個人情報に関する項目で特に分類数の多い (H) 換算年齢が 5 歳刻みとされている。その結果は次の通りである。

表 8. 検討 5 における標本でも母集団でも一意である個体数の推定値

	全 体	(d) 休業	(e) 完全失業
標本の一意な個体数	3813	54	65
$\hat{\beta}$	3.00	0.04	0.05
$\hat{\beta} \pm 3 \cdot \text{標準誤差}$	3.05	0.04	0.05
$\beta = 0, \infty$	163.34	2.31	2.78

検討 6

次に (D) と (E) の項目の併合が行われ、新たに 15 歳未満人員 (0~5 の 6 分類) という項目が作られた。その結果は次の通りである。

標本でも母集団でも一意な個体数の期待値は小さく、実際の母集団がこの 3 倍程度であることを考慮しても十分安全と思われる。しかも、情報も地域変数以外は元のデータからほとんど失われておらず、個票データとしても利用価値は十分保たれていると思われる。

表 9. 検討 6 における標本でも母集団でも一意である個体数の推定値

	全 体	(d) 休業	(e) 完全失業
標本の一意な個体数	2974	42	58
$\hat{\beta}$	2.32	0.03	0.05
$\hat{\beta} \pm 3 \cdot \text{標準誤差}$	2.35	0.03	0.05
$\beta = 0, \infty$	26.16	0.37	0.51

5. おわりに

本論文では 2 つのモデルのもとで、一意な個体数の期待値を指標として検討を行った。

まず 2.2 節では、ポアソン分布を用いたモデルについて、母集団一意である個体数の期待値についての性質を求めるとともに、その上限をとるパラメータの組を求めた。2.3 節では実用性のあるポアソングンマモデルでの個体数とセル数についての検討を行った。また一意な個体数の期待値の上限をとるスケールパラメータについても議論し、例えば個体数の期待値がセル数の 2 倍以下であればスケールパラメータが 0、すなわちパラメータがすべて等しいポアソン分布を用いたモ

デルに退化した場合が上限であることを示した。

第3章では母集団と標本との関係を考慮に入れて検討を行った。まず3.1節ではポアソン分布を用いたモデルで、標本の一意な個体数から、標本でも母集団でも一意な個体数を推定する際のいくつかの性質について考えるとともに、その推定値の上限を示した。次に3.2節において、ポアソンガンマモデルについて同様の議論を行った。

ここで注意すべきことは、一般に第3章で考えた上限が標本のデータからはほとんどあり得ない値として得られる点である。これは、例えばポアソンガンマモデルでは、上限をとるスケールパラメータが標本からの推定値からはかけ離れた値をとると言い換えることができる。そこで第4章の応用例中に示したように、パラメータの値を推定値が増加する方向にある程度ずらす方法が現実的であろう。もしこの考え方をさらに厳密に適用するのであれば、標本でも母集団でも一意である個体数の信頼区間を得る必要があるが、そのためにはブートストラップ法の利用が有効であろう。

第4章で労働力調査の個票の集計データに対して提案した方法を基に6つの検討を行ったが、地域変数の県内一連番号を削除しなければ個票データは完全に安全とは言えないと思われる。したがって、地域に関する情報を重視する立場の者が満足するような開示はやや難しいが、個人に関する情報を重視すると開示は比較的容易で、検討6のセルの併合を行う程度で十分安全な個票データの開示が可能と思われる。

6. 付 録

定理1の証明

(2) 式を $\sum_{i=1}^K \pi_i = 1$ の条件の下に π_i について

$$\frac{\partial \{N_0 \sum_{i=1}^K \pi_i e^{-N_0 \pi_i} - l(\sum_{i=1}^K \pi_i - 1)\}}{\partial \pi_i} = N_0(1 - N_0 \pi_i) e^{-N_0 \pi_i} - l \quad (18)$$

と偏微分し0とおくことによって、 $E(S_i)$ が最大となる必要条件が $(1 - N_0 \pi_i) e^{-N_0 \pi_i} = \frac{l}{N_0}$ ($i=1, 2, \dots, K$)であることが得られる。ここで左辺の $N_0 \pi_i$ を x とおいた式を $g_1(x) = (1-x)e^{-x}$ ($0 \leq x \leq N_0$)と定義して $y = g_1(x)$ のグラフを考えると、 $0 < x < 2$ において単調減少で $g_1(0) = 1$, $g_1(1) = 0$, $g_1(2) = -\frac{1}{e^2}$, $2 < x$ で単調増加で $g_1(N_0) = (1 - N_0)e^{-N_0} < 0$ である。($N_0 \leq 2$ の場合には $x \leq N_0$ の部分のみを考える。)したがって $g_1(x) = \frac{l}{N_0}$ は高々2個の実数解しかもたず、 π_i も高々2個の値しかとらないことがわかる。

(i) $N_0 \leq K$ のとき

すべての i について $1 < N_0 \pi_i$ であれば $K < N_0$ となり矛盾する。したがって $N_0 \pi_i$ のうち少なくとも1つは1以下であるが、 $g_1(N_0 \pi_i) \geq 0$ であるので $g_1(N_0 \pi_i) = g(x)$ となる x は $y = g_1(x)$ のグラフから $N_0 \pi_i$ 以外に存在しないことがわかり、 $\pi_1 = \dots = \pi_K = \frac{1}{K}$ のときに $E(S_i)$ は最大となる。

(ii) $2K < N_0$ の場合

(2) 式の右辺の第 i 項の $N_0\pi_i$ を x とおいた式を $g_2(x) = xe^{-x}$ ($0 \leq x \leq N_0$) と定義してそのグラフを考えると, $0 < x < 1$ において単調増加かつ上に凸, $1 < x < 2$ において単調減少かつ上に凸, $2 < x < N_0$ において単調減少かつ下に凸である. $E(S_1)$ の最大値を求めるためには, $\sum_{i=1}^K x_i = N_0$ となる K 個の点 $(x_i, g_2(x_i))$ ($i=1, 2, \dots, K$) をグラフ上にとり, $\sum_{i=1}^K g_2(x_i)$ を最大にするを考えればよい.

まず $x_i \equiv \frac{N_0}{K}$ ($i=1, 2, \dots, K$) のとき $\sum_{i=1}^K g_2(x_i)$ が最大でないことを証明する. $x_i \equiv \frac{N_0}{K}$ ($i=1, 2, \dots, K$) の場合 $2 < \frac{N_0}{K}$ で, $g_2(x)$ は $2 < x$ で下に凸である. したがって $\sum_{i=1}^K x_i = N_0$ を満たしつつ $x_i = x_a, x_b$ ($x_a < x_b$) なる 2 つの値に分けた方が $\sum_{i=1}^K g_2(x_i)$ は大きくなるので, $x_i \equiv \frac{N_0}{K}$ のとき最大とはならない. よって, π_i が 2 個の値をとるとき $\sum_{i=1}^K g_2(x_i)$ は最大になることがいえる. また明らかに $x_b > 2$ で, $y = g_1(x)$ のグラフの形より $1 < x_a < 2$ であることもわかる.

次に x_a が $K-1$ 個, x_b が 1 個のとき $\sum_{i=1}^K g_2(x_i)$ が最大になることを, x_a が k_a 個, x_b が k_b 個 ($k_b \geq 2$) ある場合には最大とならないことを示すことによって証明する. ただし $k_a + k_b = K$, $k_a x_a + k_b x_b = N_0$ である.

ここで 2 点 $(x_a, g_2(x_a)), (x_b, g_2(x_b))$ を結ぶ直線を $y = l(x)$ とすると, $y = g_2(x)$ のグラフの形状より $x_b < x < N_0$ において $l(x) < g_2(x)$ となることがわかる. ここで仮に $(x_a, g_2(x_a))$ を 1 個増やし $k_a + 1$ 個にし, 点 $(x_b, g_2(x_b))$ を直線 $y = l(x)$ に沿って動かし, $k_b - 1$ 個を $(x_c, l(x_c))$ ($((k_a + 1)x_a + (k_b - 1)x_c = N_0)$) とすると, $(k_a + 1)g_2(x_a) + (k_b - 1)l(x_c)$ の値は $k_a g_2(x_a) + k_b g_2(x_b)$ と変わらない. しかし実際には $(x_a, g_2(x_a))$ を 1 個増やし $k_a + 1$ 個にすると, $k_b - 1$ 個が $(x_c, g_2(x_c))$ となり, $l(x_c) < g_2(x_c)$ であることより $\sum_{i=1}^K g_2(x_i)$ の値はさらに大きくなる. したがって $k_b \geq 2$ の場合には最大とはならず, x_a が $K-1$ 個, x_b が 1 個のとき $\sum_{i=1}^K g_2(x_i)$ は最大になることがいえる.

定理 2 の証明

(6) 式の右辺の分母を

$$g_3(\beta) = (1 + N_0\beta)^{1 + \frac{1}{K\beta}} \quad (19)$$

とおく. ここで

$$\log g_3(\beta) = \left(1 + \frac{1}{K\beta}\right) \log(1 + N_0\beta) \quad (20)$$

の両辺を β で微分することによって

$$g_3'(\beta) = \frac{(1 + N_0\beta)^{1 + \frac{1}{K\beta}}}{K\beta^2(1 + N_0\beta)} \{N_0\beta(K\beta + 1) - (1 + N_0\beta)\log(1 + N_0\beta)\} \quad (21)$$

が得られる。この右辺の後半を

$$g_4(\beta) = N_0\beta(K\beta + 1) - (1 + N_0\beta)\log(1 + N_0\beta) \quad (22)$$

と置き直し、微分すると

$$g_4'(\beta) = N_0\{2K\beta - \log(1 + N_0\beta)\} \quad (23)$$

となる。\$y=2K\beta\$ と \$y=\log(1+N_0\beta)\$ のグラフを考えることによって次の結果が得られる。

(i) \$N_0 \le 2K\$ のとき

\$\beta > 0\$ で \$g_4'(\beta) > 0\$。したがって \$g_4(\beta)\$ は単調増加で \$g_4(0)=0\$ より \$\beta > 0\$ で \$g_4(\beta) > 0\$。ここで (21) 式の右辺の前半が \$\beta > 0\$ で正より、\$g_3'(\beta)\$ も \$\beta > 0\$ で正となる。したがって \$g_3(\beta)\$ は単調増加で \$E(S_1)\$ が単調減少より \$\beta=0\$ のとき \$E(S_1)\$ が最大で

$$\begin{aligned} \lim_{\beta \rightarrow 0} E(S_1) &= \lim_{\beta \rightarrow 0} \frac{N_0}{(1 + N_0\beta) \cdot \{(1 + N_0\beta)^{\frac{1}{N_0\beta}}\}^{\frac{N_0}{K}}} \\ &= N_0 e^{-\frac{N_0}{K}} \end{aligned} \quad (24)$$

となる。

(ii) \$2K < N_0\$ のとき

\$g_4'(\beta_c)=0\$ (\$0 < \beta_c\$) となる \$\beta_c\$ が 1 つ存在し、\$0 < \beta < \beta_c\$ で \$g_4'(\beta) < 0\$、\$\beta_c < \beta\$ で \$g_4'(\beta) > 0\$。したがって \$g_4(\beta)\$ は \$0 < \beta < \beta_c\$ で単調減少で \$\beta_c < \beta\$ で単調増加である。また \$g_4(0)=0\$、\$\lim_{\beta \rightarrow \infty} g_4(\beta) = \infty\$ より、\$g_4(\beta_c)=0\$ (\$\beta_c < \beta_c\$) となる \$\beta_c\$ が 1 つ存在する。したがって、\$0 < \beta < \beta_c\$ で \$g_3'(\beta) < 0\$、すなわち \$g_3(\beta)\$ は単調減少で、\$\beta_c < \beta\$ で \$g_3'(\beta) > 0\$、すなわち \$g_3(\beta)\$ は単調増加である。よって \$E(S_1)\$ は \$\beta = \beta_c\$ で最大となる。

定理 3 の証明

(9) 式を \$\sum_{i=1}^K \pi_i = 1\$ の条件の下に \$\pi_i\$ について

$$\begin{aligned} & \frac{\partial \{ \sum_{i=1}^K \pi_i (1 - N_0 \pi_i) e^{-N_0 \pi_i} - l (\sum_{i=1}^K \pi_i - 1) \}}{\partial \pi_i} \\ &= \{ (N_0 \pi_i)^2 - 3N_0 \pi_i + 1 \} e^{-N_0 \pi_i} - l \end{aligned} \quad (25)$$

と偏微分し 0 とおくことによって、(9) 式が最大となる必要条件が \$\{(N_0 \pi_i)^2 - 3N_0 \pi_i + 1\} e^{-N_0 \pi_i} = l\$ (\$i=1, 2, \dots, K\$) であることが得られる。ここで \$N_0 \pi_i\$ を \$x\$ とおいた式を \$g_5(x) = (x^2 - 3x + 1)e^{-x}\$ (\$0 \le x \le N_0\$) と定義して \$y = g_5(x)\$ のグラフを考えると、\$0 < x < 1\$ において単調減少で \$g_5(0) = 1\$、\$g_5(1) = -\frac{1}{e}\$、\$1 < x < 4\$ で単調増加で \$g_5(4) = \frac{5}{e^4}\$、\$4 < x\$ で単調減少で \$g_5(N_0) = (N_0^2 - 3N_0 + 1)e^{-N_0}\$ である。(\$N_0 \le 4\$ の場合には \$x \le N_0\$ の部分のみを考える。) したがって \$g_5(x) = l\$ は高々 3 個の実数解しか持たず、\$\pi_i\$ も高々 3 個の値しかとらないことが証明できる。

ここで \$g_5(x) = g_5(4)\$ (\$x \neq 4\$) の解を \$x = t\$ とすると、\$t \doteq 0.3225\$ で、\$0 < x < t\$ において \$g_5(x) = l\$ は 1 個の実数解しか持たず、したがって \$N_0 < t \cdot K\$ のとき定理 1 の証明の (i) と同様に \$\pi_1 = \dots = \pi_K = \frac{1}{K}\$ のときに \$\frac{\partial E(S_1)}{\partial N_0}\$ が最大となることが証明できる。

定理 4 の証明

(12) 式の右辺の分母を

$$g_6(\beta) = (1 + N_0\beta)^{2 + \frac{1}{K\beta}} \quad (26)$$

とおく。ここで

$$\log g_6(\beta) = \left(2 + \frac{1}{K\beta}\right) \log(1 + N_0\beta) \quad (27)$$

の両辺を β で微分することによって

$$g'_6(\beta) = \frac{(1 + N_0\beta)^{2 + \frac{1}{K\beta}}}{K\beta^2(1 + N_0\beta)} \{N_0\beta(2K\beta + 1) - (1 + N_0\beta)\log(1 + N_0\beta)\} \quad (28)$$

が得られる。この右辺の後半を

$$g_7(\beta) = N_0\beta(2K\beta + 1) - (1 + N_0\beta)\log(1 + N_0\beta) \quad (29)$$

と置き直し、微分すると

$$g'_7(\beta) = N_0\{4K\beta - \log(1 + N_0\beta)\} \quad (30)$$

となり、 $y = 4K\beta$ と $y = \log(1 + N_0\beta)$ のグラフを考えることによって次の結果が得られる。

$N_0 \leq 4K$ のとき

$\beta > 0$ で $g'_7(\beta) > 0$ 。したがって定理 2 の証明の (i) と同様にして $g_6(\beta)$ は単調増加であることがいえる。ここで (12) 式の右辺の分子 $1 - \frac{N_0}{K}$ の符号を考えると次のような結論が得られる。

(i) $N_0 < K$ のとき

$1 - \frac{N_0}{K} > 0$ より $\frac{\partial E(S_1)}{\partial N_0}$ は単調減少で $\beta = 0$ のときに最大値 $\left(1 - \frac{N_0}{K}\right)e^{-\frac{N_0}{K}}$ をとる

(ii) $N_0 = K$ のとき

$1 - \frac{N_0}{K} = 0$ より $\frac{\partial E(S_1)}{\partial N_0} = 0$

(iii) $K < N_0 \leq 4K$ のとき

$1 - \frac{N_0}{K} < 0$ より $\frac{\partial E(S_1)}{\partial N_0}$ は単調増加で $\beta \rightarrow \infty$ のときに最大で $\lim_{\beta \rightarrow \infty} \frac{\partial E(S_1)}{\partial N_0} = 0$

また

(iv) $4K < N_0$ のとき

$g'_7(\beta_a) = 0$ ($0 < \beta_a$) となる β_a が 1 つ存在し、 $0 < \beta < \beta_a$ で $g'_7(\beta) < 0$ 、 $\beta_a < \beta$ で $g'_7(\beta) > 0$ 。したがって $g_7(\beta)$ は $0 < \beta < \beta_a$ で単調減少で $\beta_a < \beta$ で単調増加である。また $g_7(0) = 0$ 、 $\lim_{\beta \rightarrow \infty} g_7(\beta) = \infty$ より、 $g_7(\beta_a) = 0$ ($\beta_a < \beta_a$) となる β_a が 1 つ存在する。したがって、 $0 < \beta < \beta_a$ で $g'_6(\beta) < 0$ 、すなわち $g_6(\beta)$ は単調減少で、 $\beta_a < \beta$ で $g'_6(\beta) > 0$ 、すなわち $g_6(\beta)$ は単調増加である。また (12) 式の右辺の分子 $1 - \frac{N_0}{K} < 0$ であることと合わせると $\frac{\partial E(S_1)}{\partial N_0}$ は $\beta < \beta_a$ で単調減少、 $\beta_a < \beta$ で単調増加で、 $\beta \rightarrow \infty$ のときに最大値 0 をとる。

謝 辞 本論文の作成にあたり，東京大学経済学部の竹村彰通先生と横浜国立大学の加納悟先生を初めとして，文部省科学研究費補助金の竹村班，加納班のメンバーからは多くのアドバイスを頂き，ここに感謝いたします。また論文の訂正にあたって，査読者と編集理事の方々から適切なコメントを頂いたことに感謝いたします。本研究を進める上で，文部省科学研究費補助金（統計情報活用フロンティアの拡大—マイクロデータによる社会構造解析—）の援助を受けている。

参 考 文 献

- Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990) : Disclosure Control of Microdata, *Journal of the American Statistical Association* **85**, 38-45.
- Blien, U., Wirth, H. and Müller, M. (1992) : Disclosure Risk for Microdata Stemming from Official Statistics, *Statistica Neerlandica* **46**, 69-82.
- Greenberg, B.V. and Zayatz, L.V. (1992) : Strategies for Measuring Risk in Public Use Microdata File, *Statistica Neerlandica* **46**, 33-48.
- 星野伸明, 竹村彰通 (1998) : 個票開示リスクの対数級数モデルによる推定, 応用統計学会・日本計量生物学会 1998 年度合同年次大会講演予稿集, 7-12.
- 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, *Statistica Neerlandica* **46**, 5-19.
- Marsh, C., Skinner, C. et al. (1991) : The Case for Samples of Anonymized Records from the 1991 Census, *Journal of the Royal Statistical Society, 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.
- Paass, G. (1988) : Disclosure Risk and Disclosure Avoidance for Microdata, *Journal of Business and Economic Statistics* **6**, 487-500.
- 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について, 平成 9 年度科学研究費補助金報告書(竹村班), 69-82.
- Särndal, C.E., Swensson, B. and Wretman, J. (1992) : *Model Assisted Survey Sampling*, Springer.
- 渋谷政昭 (1997) : 多項分布における度数 0, 1 セルの数—漏洩管理のための基礎事実—, 応用統計学 **26**, 161-170.
- Skinner, C.J. (1992) : On Identification Disclosure and Prediction Disclosure for Microdata, *Statistica Neerlandica* **46**, 21-32.
- 総務庁統計局 (1996) : 労働力調査年報 平成 7 年.
- Takemura, A. (1997) : Some Superpopulation Models for Estimating the Number of Population Uniques, Discussion Paper 97-f-29, Faculty of Economics, University of Tokyo.
- 竹村彰通 (1998) : 労働力調査に見られるサイズインデックス, 平成 9 年度科学研究費補助金報告書(竹村班), 96-104.
- de Waal, T. (1993) : The Maximum Expected Number of Unique Individuals in a Population, Internal Report, Netherlands Central Bureau of Statistics.
- Willenborg, L. and de Waal, T. (1996) : *Statistical Disclosure Control in Practice*, Springer.

(1998 年 6 月 12 日受付 10 月 29 日最終修正)

著者連絡先: 〒 700-8601 岡山市津島京町 2-10-1 岡山商科大学商学部