統計数理(2003)
第 51 巻 第 2 号 183–197

特集「個票開示問題の統計理論」
［原著論文］

# 母集団寸法指標のノンパラメトリック推定

佐井 至道†

（受付 2003 年 1 月 20 日；改訂 2003 年 5 月 14 日）

**要　　旨**

標本調査で得られた個票データの開示におけるリスク評価では母集団寸法指標を用いた指標を用いることが多い．標本寸法指標からの母集団寸法指標の推定では，ポアソンガンマモデルやピットマンモデルなどの超母集団モデルを用いるパラメトリックな方法が主流である．

本論文では，これまで提案されているノンパラメトリック推定についてまとめるとともに，ノンパラメトリック最尤推定する方法を新たに提案する．この推定法には，推定が不安定になるというノンパラメトリック推定共通の問題点があり，膨大な時間を要するという計算上の問題もある．

前者の点を解決するために，母集団寸法指標に対して制約を加えることにする．母集団寸法指標の性質が経験的に知られている場合には，単調減少や下に凸のような簡単な制約を置くことによって，典型的な状況では安定した推定が可能になる．

また後者の問題を解決するために，ベルヌーイ抽出による尤度をポアソン確率関数で近似する方法を提案するとともに，すべての可能な母集団寸法指標について網羅的に尤度を計算する代替法として，いくつかの簡便法を提案する．母集団寸法指標ではサイズの小さい部分の推定が特に重要であるため，大きいサイズを打ち切った標本寸法指標を用いることにより，現実的な大きさの母集団に対しても推定が可能になる．

キーワード： 個票データ，キー変数，標本寸法指標，母集団寸法指標.

## 1. はじめに

標本調査で得られた個票データの寸法指標から母集団の寸法指標を推定する方法としては，Bethlehem et al.(1990)がポアソンガンマモデルを用いてから，多項ディリクレモデル，ピットマンモデルなど，少数のパラメータを持つ何らかの超母集団モデルを用いる方法が主流となっている．これらのモデルの詳細については佐井(1998)，Hoshino and Takemura (1998)，Takemura (1999)，Hoshino (2001)を参照されたい．

しかし，既存のどのモデルも寸法指標のすべての形状を表すことができないため，モデルを用いないノンパラメトリックな方法について考えることは重要であり，逆にこのような検討を行うことは適切なモデルの構築の助けになる．ただし，これまでノンパラメトリックな方法を提案あるいは紹介しているものの中で実用的なものは多くない．

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推定法を提案している．ただしその過程において未知パラメータの値が必要となり，それを一般にはかなり異なる個票データからの推定値で置き換えなければならない問題があった．その方法につ

† 岡山商科大学 法経学部：〒700–8601 岡山市津島京町 2–10–1

いて 2.3 節で説明する．また加納(1997)は，母集団のあるセルに含まれる個体数がある決められた値以下であるという仮説の尤度比検定を提案した中で，母集団の 1 つのセルに含まれる個体数の最尤推定量が，標本の対応するセルに含まれる個体数に抽出率の逆数をかけた形で書けることを示した．詳しくは 2.4 節で説明する．更に渋谷(1999)は母集団寸法指標から標本寸法指標の期待値への一次変換行列を用いた Engen (1978)の提案した推定量についての性質を検討したが，数値例による推定結果は非常に悪いとしている．2.5 節において概略を説明する．

なお 2.2 節においては，Zipf (1935)の提案した法則と，それに類似した手法について紹介する．これらの方法はモデルを用いているものの，現在利用されている超母集団モデルに比べて極めて簡素であるため取り上げる．

第 3 章においては標本寸法指標から母集団寸法指標をノンパラメトリック最尤推定する方法を提案する．まず 3.1 節において尤度関数を求め，3.2 節では尤度関数の計算時間上の問題を解決するためにポアソン分布の確率関数による近似を提案する．次に，寸法指標は単調減少や下に凸など共通の特徴を持っていることが経験的に知られているため，推定の際に母集団寸法指標に何らかの制約を置くことを 3.3 節で提案する．これによって安定した推定が可能になる．なおこの方法では，母集団の可能なすべての寸法指標について尤度を計算しなければならない点も計算時間上の障害となるが，これについては 3.4 節で述べる簡便法を用いることによって解決を図りたい．

第 4 章では，アメリカのセンサス個票データなどに対して提案した方法を適用し，その有用性を確かめる．

## 2. これまでのノンパラメトリック推定

### 2.1 定義

母集団の $N$ 個の個体が，個体を特定するために用いられる数種類のキー変数(key variable)の組み合わせに基づいて $K$ 個のセル(cell)に分けられているものとして，第 $i$ 番目のセルのサイズ(含まれる個体数)を $F_i$ $(i = 1, 2, \ldots, K)$ とする．ここで，サイズ $l$ のセル数，すなわち $F_i = l$ となるセル数を $S_l$ $(l = 0, 1, 2, \ldots, L)$ とする．これが母集団寸法指標(population size indexes)である．なお $L$ はセルのサイズの最大値である．

寸法指標は Sibuya (1993)で用いられた size index の訳語であるが，この指標は度数スペクトル(frequency spectrum)，グループ分けされた度数分布(grouped frequency distribution)，度数の度数(frequencies of frequencies)，同一クラス(equivalence classes)など，様々な呼び方をされている．

次に標本(個票データ(microdata))の大きさを $n$ として，抽出率を $\lambda = n/N$ と表す．標本では $F_i$，$S_l$ の代わりに $f_i$，$s_{l'}$ $(l' = 0, 1, 2, \ldots, L)$ という表記を用いるが，後者が標本寸法指標(sample size indexes)である．$s_L$，$s_{L-1}$ などは一般に 0 となる可能性が高い．

### 2.2 Zipf の法則

ここで紹介する Zipf の法則とそれに追随して提案された推定法はモデルを用いる方法ではあるが，現在用いられているパラメトリックな推定法と比較して極めて簡素なものであるため取り上げることにする．なおこの節の手法に関する文献としては Baayen (2001)が詳しいが，そこでの議論は言語統計学の分野への適用を想定して行われたものである．例えば 1 冊の英語の文献を母集団と考えると，個体はその中で使用されている各単語に相当し，$S_l$ は $l$ 回使用されている単語の度数に相当する．

Zipf (1935)は $\log l$ を横軸に，$\log S_l$ を縦軸にとったときに，多くの文献のデータに対して傾

きが負の直線的な関係があることを示した．このような関係は Zipf の法則と呼ばれ，

$$S_l \propto \frac{1}{l^a} \tag{2.1}$$

と表すことができる．ただし $a$ は定数である．

また Zipf は，$\pi_z$ を文献内で $z$ 番目に多く使用されている単語の相対度数とし，これに対しても次のように Zipf の法則を適用した．ただし $b$ は定数である．

$$\pi_z \propto \frac{1}{z^b}\,. \tag{2.2}$$

Zipf はしばしば $b=1$ とおいたが，その場合寸法指標で表現し直すと

$$S_l \propto \frac{1}{l(l+1)} \tag{2.3}$$

となる．Baayen の応用例では (2.1)式より(2.3)式の当てはまりの方が良い．

これらの方法については Church and Gale (1991)，Gale and Sampson (1995)，Naranan and Balasubrahmanyan (1998)が改善を図っているが，Zipf の法則と同様に母集団や標本の大きさによってパラメータの値が変動するなど問題点が多い．

### 2.3 Greenberg and Zayatz の推定方法

この節では Greenberg and Zayatz (1992)が提案した，母集団寸法指標の推定方法を紹介する．

まず母集団のセルの内でサイズ 0 のセルを除いた総セル数を $U=K-S_0$ として，サイズ 1 以上のセルのみを考える．母集団においてサイズ $l$ のセルが標本でサイズ $l'$ となる確率 $P(l' \mid l)$ は，非復元単純無作為抽出を仮定すると超幾何分布を用いて

$$P(l' \mid l) = \frac{{}_lC_{l'} \cdot {}_{N-l}C_{n-l'}}{{}_NC_n} \tag{2.4}$$

で与えられ，標本でサイズ $l'$ となるセル数 $s_{l'}$ の期待値は次のように表される．

$$E(s_{l'}) = \sum_{l=l'}^{L} P(l' \mid l) \cdot S_l\,. \tag{2.5}$$

ここで，あるセルのサイズが $l$ となる確率 $P(l)=S_l/U$ を用いるとベイズの定理より

$$P(l \mid l') = \frac{P(l' \mid l) \cdot \frac{S_l}{U}}{\sum_{l=l'}^{L} P(l' \mid l) \cdot \frac{S_l}{U}} \tag{2.6}$$

となるため，例えば標本で $l'=1$ のセルが観測されたとき，そのうち母集団においても $l=1$ であるセル数の事後分布の期待値は(2.6)式を用いて $s_1 \cdot P(l=1 \mid l'=1)$ と書くことができる．このような値を求めるためには $S_l/U$ の値が必要となるが，標本の対応する値 $s_l/u$ で置き換えるとしている．ただし $u=K-s_0$ である．

しかし加納(1997)が指摘しているように，一般の個票データは小さい抽出率でサンプリングされており，そのような場合，$s_l/u$ での置き換えには大きな問題がある．

### 2.4 加納の推定方法

加納(1997)が尤度比検定を提案した際に，「悪い推定」として紹介した最尤推定について説明する．

非復元単純無作為抽出された標本から母集団の第 $i$ 番目のセルのサイズを推定するために，前節と同様に超幾何分布を考える．母集団の第 $i$ 番目のセルのサイズが $F_i$ のときに標本での同じセルのサイズが $f_i$ である尤度関数は

$$L_i(f_i \mid F_i) = \frac{{}_{F_i}C_{f_i} \cdot {}_{N-F_i}C_{n-f_i}}{{}_{N}C_{n}} \tag{2.7}$$

と書け，もし $1/\lambda$ が整数値であれば，この尤度関数は $F_i = f_i/\lambda$ のときに最大となり，これが最尤推定量となる．$1/\lambda$ が整数値でなければ $f_i/\lambda$ も整数値とならないことがあり，その場合 $F_i$ が $f_i/\lambda$ を挟む整数値のどちらかで最大となるが，議論を簡略化するために，ここでは $1/\lambda$ を整数値として考える．

これを各セルについて行うと，標本の寸法指標 $(s_0, s_1, s_2, \ldots, s_L)$ から最尤推定される母集団の寸法指標は $\hat{S}_0 = s_0,\ \hat{S}_{1/\lambda} = s_1,\ \hat{S}_{2/\lambda} = s_2, \ldots, \hat{S}_{L/\lambda} = s_L$ でそれ以外は 0 となり，標本の寸法指標の柱の間隔を単に $1/\lambda$ 倍にした飛び飛びの形状となる．そのため一般には $\hat{S}_1 = 0$ と推定されてしまう．

個々のセルを区別するノンパラメトリック推定では，後で考えるような制約条件を導入したとしても，推定対象として特に重要な $S_1$ などについてうまく推定が行えない．これを克服するためには，1 つずつのセルを無視して寸法指標のみに基づいた最尤推定を考えなければならないと思われる．

### 2.5 Engen と渋谷の推定方法

次に Engen (1978)が提案し，渋谷(1999)が「素朴な推定」と呼んでその性質を示したノンパラメトリック推定を簡単に紹介する．なお Engen は Goodman (1949)が提案した推定量の不偏性に関する条件を修正したものである．

非復元単純無作為抽出を仮定すると，2.3 節と同様に標本でサイズが $l'$ のセル数の期待値は次のように求められる．

$$\begin{aligned} E(s_{l'}) &= \sum_{l=l'}^{L} \frac{{}_{l}C_{l'} \cdot {}_{N-l}C_{n-l'}}{{}_{N}C_{n}} S_l \\ &= \sum_{l=l'}^{L} {}_{l}C_{l'} \frac{n^{\underline{l'}}(N-n)^{\underline{l-l'}}}{N^{\underline{l}}} S_l \,. \end{aligned} \tag{2.8}$$

ただし $N^{\underline{l}} = N(N-1)\cdots(N-l+1)$ である．

これを元に，母集団寸法指標から標本寸法指標の期待値への一次変換は次のように求められる．

$$\begin{pmatrix} E(s_0) \\ E(s_1) \\ E(s_2) \\ E(s_3) \\ \vdots \\ E(s_L) \end{pmatrix} = \begin{pmatrix} 1 & \frac{N-n}{N} & \frac{(N-n)^{\underline{2}}}{N^{\underline{2}}} & \frac{(N-n)^{\underline{3}}}{N^{\underline{3}}} & \cdots & \frac{(N-n)^{\underline{L}}}{N^{\underline{L}}} \\ 0 & \frac{n}{N} & 2\frac{n(N-n)}{N^{\underline{2}}} & 3\frac{n(N-n)^{\underline{2}}}{N^{\underline{3}}} & \cdots & L\frac{n(N-n)^{\underline{L-1}}}{N^{\underline{L}}} \\ 0 & 0 & \frac{n^{\underline{2}}}{N^{\underline{2}}} & 3\frac{n^{\underline{2}}(N-n)}{N^{\underline{3}}} & \cdots & {}_{L}C_{2}\frac{n^{\underline{2}}(N-n)^{\underline{L-2}}}{N^{\underline{L}}} \\ 0 & 0 & 0 & \frac{n^{\underline{3}}}{N^{\underline{3}}} & \cdots & {}_{L}C_{3}\frac{n^{\underline{3}}(N-n)^{\underline{L-3}}}{N^{\underline{L}}} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \frac{n^{\underline{L}}}{N^{\underline{L}}} \end{pmatrix} \begin{pmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \\ \vdots \\ S_L \end{pmatrix} . \tag{2.9}$$

ここで $E(s_l)$ を実現値 $s_l$ で置き換え，$S_1, S_2, \ldots, S_L$ について解くことによって母集団寸法

指標の推定量として次の式が得られる．

$$\hat{S}_l = \sum_{l'=l}^{L} \left[ \sum_{j=l}^{l'} \frac{{}_N C_j \cdot {}_{l'} C_j \cdot {}_j C_l}{{}_n C_j} (-1)^{j-l} \right] s_{l'} \,. \tag{2.10}$$

Engen は $l \leq n$ であれば，これが母集団寸法指標の唯一の不偏推定量であることを証明した．しかし渋谷は，3 種類の母集団からサンプリングされた 3 種類ずつの標本を元に推定を行ったところ，その結果は非常に悪いと指摘している．

これは(2.9)式の一次変換を連立方程式と見て下から順に解く場合，次のように解釈することもできる．母集団で最大のサイズのセル数 $S_L$ の値は一般には 1 とか 2 のように小さく，$E(s_L) = (n^{\underline{L}}/N^{\underline{L}})S_L$ は一般に期待値として 0 に近い小数値となる．しかし，それを実現値 $s_L$ に置き換えるとその値は 0 とか 1 という整数値になるため，$S_L$ は 0 か，さもなければ極めて大きな値として得られてしまう．その値を連立方程式のその上の式に代入して $S_{L-1}, S_{L-2}, \ldots$ と順々に求めると，つじつまを合わせるために推定値は極めて大きい値や負の値を行き来することになってしまう．

## 3. 標本の寸法指標に基づく制約付き最尤推定

### 3.1 ノンパラメトリック最尤推定

この節では，標本寸法指標から母集団寸法指標をノンパラメトリック最尤推定する方法を考える．つまり標本の寸法指標 $(s_1, s_2, \ldots, s_L)$ が得られたときに，尤度を最大にするような母集団の寸法指標 $(S_1, S_2, \ldots, S_L)$ を求める．なお総セル数 $K$ は未知であることが多いため $s_0$ の観測は困難であるが，$S_0$ は推定の対象とはしないことにする．また $L$ としては $S_L \neq 0$ となる可能性のある最大値を用いるべきだが，実用上はそれより小さな値で置き換えねばならない．

$k_{ll'}$ $(l = 0, 1, 2, \ldots, L\ ;\ l' = 0, 1, 2, \ldots, l\ ;\ l' \leq l)$ を母集団においてサイズ $l$ のセルの内で，標本においてサイズ $l'$ となるものの数とすると，非復元単純無作為抽出の場合に尤度関数は

$$\begin{aligned} &L(s_1, s_2, \ldots, s_L \mid S_1, S_2, \ldots, S_L) \\ &\quad = \frac{1}{{}_N C_n} \sum_{\mathcal{C}_1} \prod_{l=1}^{L} \frac{S_l!}{k_{l0}! k_{l1}! \cdots k_{ll}!} ({}_l C_0)^{k_{l0}} ({}_l C_1)^{k_{l1}} \cdots ({}_l C_l)^{k_{ll}} \end{aligned} \tag{3.1}$$

と書ける．ただし，条件 $\mathcal{C}_1$ は非負整数 $k_{ll'}$ の表 1 に示すような行和と列和に関する条件である．総個体数 $N$ を満たし，得られた標本寸法指標を生み出す可能性のある母集団寸法指標 $(S_1, S_2, \ldots, S_L)$ の中で，尤度関数(3.1)式を最大にするものが求める最尤推定値となる．

表 1. 条件 $\mathcal{C}_1$ （母集団から標本への寸法指標の移動）．

| | | 母集団 | | | | | |
|---|---|---|---|---|---|---|---|
| | | $S_0$ | $S_1$ | $S_2$ | $S_3$ | $\cdots$ | $S_L$ |
| 標本 | $s_0$ | $k_{00}$ | $k_{10}$ | $k_{20}$ | $k_{30}$ | $\cdots$ | $k_{L0}$ |
| | $s_1$ | | $k_{11}$ | $k_{21}$ | $k_{31}$ | $\cdots$ | $k_{L1}$ |
| | $s_2$ | | | $k_{22}$ | $k_{32}$ | $\cdots$ | $k_{L2}$ |
| | $s_3$ | | | | $k_{33}$ | $\cdots$ | $k_{L3}$ |
| | $\vdots$ | | | | | | $\vdots$ |
| | $s_L$ | | | | | | $k_{LL}$ |

### 3.2 尤度のポアソン近似

前節の尤度関数(3.1)式を計算する際には，非負整数 $k_{ll'}$ の膨大な組み合わせについて和を求める必要がある．そのため $N$ が 100 程度でもパーソナルコンピュータによる計算は困難となる．この節ではこの問題を解決するために，尤度関数をポアソン分布の確率関数によって近似する方法を提案する．

まず，非復元単純無作為抽出ではなく母集団の各個体が他の個体とは独立に確率 $\lambda$ でサンプリングされる場合の尤度関数を考える．このような抽出法は Särndal et al. (1992)ではベルヌーイ抽出と呼ばれており，ここでもこの用語を用いる．ベルヌーイ抽出では母集団の大きさが固定されても標本の大きさは一定ではない．逆に，大きさ $n$ の標本が得られていても，その母集団の大きさは $n$ 以上のすべての可能性があるが，標本の大きさを $n$，母集団の大きさを $N$ にそれぞれ固定すると(3.1)式に対応する尤度関数は

$$
\begin{aligned}
(3.2)\qquad & L(s_1, s_2, \ldots, s_L \mid S_1, S_2, \ldots, S_L) \\
&= \sum_{\mathcal{C}_1} \prod_{l=1}^{L} \frac{S_l!}{k_{l0}! k_{l1}! \cdots k_{ll}!} \cdot \{(1-\lambda)^l\}^{k_{l0}} \cdot \{{}_lC_1 \lambda^1 (1-\lambda)^{l-1}\}^{k_{l1}} \cdot \cdots \cdot \{{}_lC_l \lambda^l\}^{k_{ll}} \\
&= \lambda^n (1-\lambda)^{N-n} \sum_{\mathcal{C}_1} \prod_{l=1}^{L} \frac{S_l!}{k_{l0}! k_{l1}! \cdots k_{ll}!} ({}_lC_0)^{k_{l0}} ({}_lC_1)^{k_{l1}} \cdots ({}_lC_l)^{k_{ll}}
\end{aligned}
$$

と書けるため，母集団の寸法指標のパターンによらず，ベルヌーイ抽出の尤度関数は単純無作為抽出の尤度関数の ${}_NC_n \cdot \lambda^n (1-\lambda)^{N-n}$ 倍となる．従って，標本寸法指標が単純無作為抽出で得られたとき，最大尤度をとる母集団寸法指標は，ベルヌーイ抽出を想定した場合に大きさが $N$ である母集団寸法指標の中で最大尤度をとるものに一致する．

ベルヌーイ抽出では各 $l$ $(l = 1, 2, \ldots, L)$ について $(k_{l0}, k_{l1}, \ldots, k_{ll})$ が他の $l$ とは独立に多項分布に従い，その確率関数は

$$
\begin{aligned}
(3.3)\qquad & f(k_{l0}, k_{l1}, \ldots, k_{ll}) \\
&= \frac{S_l!}{k_{l0}! k_{l1}! \cdots k_{ll}!} \{{}_lC_0 \lambda^0 (1-\lambda)^l\}^{k_{l0}} \{{}_lC_1 \lambda^1 (1-\lambda)^{l-1}\}^{k_{l1}} \cdots \{{}_lC_l \lambda^l (1-\lambda)^0\}^{k_{ll}}
\end{aligned}
$$

である．ここで抽出率が十分小さいものと仮定する．その場合，(3.3)式において ${}_lC_0 \lambda^0 (1-\lambda)^l$ を除くと ${}_lC_1 \lambda^1 (1-\lambda)^{l-1}, \ldots, {}_lC_l \lambda^l (1-\lambda)^0$ の確率はすべて小さい．また 0 でない異なる $l', l''$ に対して $k_{ll'}$ と $k_{ll''}$ との共分散

$$
(3.4)\qquad \mathrm{Cov}(k_{ll'}, k_{ll''}) = -S_l \cdot \{{}_lC_{l'} \lambda^{l'} (1-\lambda)^{l-l'}\} \cdot \{{}_lC_{l''} \lambda^{l''} (1-\lambda)^{l-l''}\}
$$

も小さい．そこで $k_{ll'}$ $(l' = 1, 2, \ldots, l)$ をそれぞれ独立なポアソン分布で

$$
(3.5)\qquad f(k_{ll'}) = \frac{e^{-S_l \cdot {}_lC_{l'} \lambda^{l'} (1-\lambda)^{l-l'}} \{S_l \cdot {}_lC_{l'} \lambda^{l'} (1-\lambda)^{l-l'}\}^{k_{ll'}}}{k_{ll'}!}
$$

と近似すると，ポアソン分布の再生性から，$s_{l'} = \sum_{l=l'}^{L} k_{ll'}$ もまた他の $l'$ とは独立に次のようにポアソン分布に従う．

$$
(3.6)\qquad f(s_{l'}) = \frac{e^{-\mu_{l'}} {\mu_{l'}}^{s_{l'}}}{s_{l'}!},
$$

ただし

$$
(3.7)\qquad \mu_{l'} = \sum_{l=l'}^{L} S_l \cdot {}_lC_{l'} \lambda^{l'} (1-\lambda)^{l-l'}
$$

である．よって尤度関数 (3.2)式は

$$L(s_1, s_2, \ldots, s_L \mid S_1, S_2, \ldots, S_L) = \prod_{l'=1}^{L} \frac{e^{-\mu_{l'}} {\mu_{l'}}^{s_{l'}}}{s_{l'}!} \tag{3.8}$$

と近似できる．

ここで(3.3)式の(3.5)式を用いた近似について補足する．

$$\begin{aligned} &\frac{S_l!}{k_{l0}!k_{l1}!\cdots k_{ll}!}\{{}_lC_0\lambda^0(1-\lambda)^l\}^{k_{l0}}\{{}_lC_1\lambda^1(1-\lambda)^{l-1}\}^{k_{l1}}\cdots\{{}_lC_l\lambda^l(1-\lambda)^0\}^{k_{ll}} \\ &\quad = \frac{e^{S_l}S_l!}{{S_l}^{S_l}}\cdot\frac{e^{-S_l(1-\lambda)^l}\{S_l(1-\lambda)^l\}^{k_{l0}}}{k_{l0}!} \\ &\qquad \cdot\prod_{l'=1}^{L}\frac{e^{-S_l\cdot{}_lC_{l'}\lambda^{l'}(1-\lambda)^{l-l'}}\{S_l\cdot{}_lC_{l'}\lambda^{l'}(1-\lambda)^{l-l'}\}^{k_{ll'}}}{k_{ll'}!} \end{aligned} \tag{3.9}$$

と書けるが，ここで(3.9)式の右辺の第 1 項と第 2 項の積を $A(\lambda, S_l, k_{l0})$ とおく．

$\lambda$ 以外の値を固定したとき $A(\lambda, S_l, k_{l0})$ は $\lambda = 1$ のとき 0，$\lambda = 1 - (k_{l0}/S_l)^{1/l}$ において極大で 1 以上の値となり，$\lambda = 0$ のとき

$$A(0, S_l, k_{l0}) = \frac{S_l!}{{S_l}^{S_l - k_{l0}}k_{l0}!}$$

で 1 以下である．従って $\lambda$ の値によっては近似によって尤度を過小に評価することもあるが，$\lambda$ が小さい場合には一般に過大に評価する傾向がある．

更に $\lambda = 0$ の場合について考える．$A(0, S_l, k_{l0})$ は $S_l$ に関して単調減少，$k_{l0}$ に関して単調増加で，$k_{l0} = S_l$ のときに最大で 1 となる．$\lambda$ が十分小さいときには一般に $k_{l0}$ も $S_l$ に近い値をとることが多く(3.5)式は良い近似となるが，$k_{l0}$ が $S_l$ より小さくなるにつれ $A(0, S_l, k_{l0})$ も小さくなり，近似により尤度を過大に評価することになる．

ただし母集団と標本の個体数が固定されているため，すべての $l$ について $S_l$ と比較して $k_{l0}$ が小さいことはなく調整機能が働く．また $k_{l0}$ が $S_l$ より極端に小さい場合には(3.9)式の右辺の第 3 項も一般には小さい．同様のことは $\lambda > 0$ の場合にも言える．

4.1 節の数値例で扱うようないくつかの小さな母集団からサンプリングされた標本の寸法指標について，各母集団寸法指標のすべてのパターンに対する(3.2)式と(3.8)式の値を比較したところ，尤度のずれは生じていたものの，尤度の順番に母集団寸法指標を並べて比較した場合には，尤度の大きい部分での問題となるような逆転現象は起こらなかった．

### 3.3 母集団寸法指標に対する制約条件

これまで述べた制約付きノンパラメトリック最尤推定には，まだ 2 つの問題が残っている．

その一つは得られた標本寸法指標を生み出す可能性のある母集団寸法指標 $(S_1, S_2, \ldots, S_L)$ のすべてのパターンについて網羅的に尤度を計算しなければならないという計算時間上の問題であるが，この解決策については次節で考えることにして，この節ではもう一つの問題について考える．

その問題は，ノンパラメトリック推定される母集団寸法指標が渋谷(1999)の結果と同様に大きく増減を繰り返し，このままでは実用性に乏しい点である．一方，実際のデータの寸法指標はほとんどの場合ほぼ単調減少や下に凸など，共通の特徴を持っている．そこで推定を安定させるために，母集団寸法指標に対して次のような数種類の制約条件を置くことを考える．

(a) 制約条件を用いない
(b) 母集団寸法指標が単調減少　$(S_1 \geq S_2 \geq \cdots \geq S_L)$

(c) 母集団で各サイズの個体数そのものが単調減少 $(S_1 \geq 2 \cdot S_2 \geq \cdots \geq L \cdot S_L)$
(d) 条件(b)に加えて，母集団寸法指標が下に凸 $(2 \cdot S_l \leq S_{l-1} + S_{l+1})$
(e) 条件(b)に加えて，母集団寸法指標の対数が下に凸 $(2 \cdot \log S_l \leq \log S_{l-1} + \log S_{l+1})$

ただし(c), (d), (e)を用いる場合にも度数がある程度大きいサイズにのみ適用することとして，度数の小さいサイズについては条件(b)を適用する．後の適用例では度数が 10 以上のサイズのみに適用することとする．

なお超母集団モデルの多くは寸法指標が単調減少であることを想定して考案されているが，その点については佐井(2002)を参照されたい．

### 3.4 計算時間短縮のための簡便法

前節で述べたように，実際に推定対象となるような大きさの母集団では，ポアソン分布による近似を用いたとしても，可能性のあるすべての母集団寸法指標 $(S_1, S_2, \ldots, S_L)$ について網羅的に尤度を計算することは計算時間上不可能である．そのため計算時間を短縮するための簡便法を用いる必要がある．簡便法については，推定の精度と要する時間とを勘案して選択する必要があるが，次のような数種類の方法を考えることにする．

(A) (3.1)式を用いて網羅的に尤度を求める方法
(B) (3.8)式のポアソン近似を用いて網羅的に尤度を求める方法

(A), (B)は可能なすべての母集団寸法指標 $(S_1, S_2, \ldots, S_L)$ のパターンについて網羅的に尤度を計算するもので，(B)と以下の方法では(3.8)式のポアソン近似を用いる．

(C) 個体追加法

まず観測された標本寸法指標に個体を 1 個追加したものを大きさ $n+1$ の母集団の寸法指標と見なし，そのような母集団寸法指標の内で尤度を最大にするものを見つける．次にその寸法指標に個体を 1 個追加したものを大きさ $n+2$ の母集団の寸法指標と見なし，その中で尤度を最大にするものを見つける．以後この操作を続け，母集団の大きさが $N$ となった時点で最大尤度をとる寸法指標を最尤推定値と考える．なお個体の追加法として 2 種類を考える．一つはサイズ 1 のセル数を 1 だけ増加させる方法であり，もう一つは，変化させるサイズを任意の 2 つとして，その内の小さいサイズのセル数を 1 だけ減少させ，大きいサイズのセル数を 1 だけ増加させ，個体の不足が生ずればサイズ 1 のセル数の減少で補う方法である．その中で最も尤度が大きくなるものを選択することにする．

(D) 2 サイズ探索法

ピットマンモデルで推定された母集団寸法指標を初期値とする．次にサイズ 1 以外の任意の 2 つのサイズの度数を $\pm 1$ の範囲で増減させ，個体の過不足をサイズ 1 のセルで調整する．そのすべてのパターンの中で尤度が最も増加するもので初期値を置き換える．以後これを繰り返し，もはや尤度が増加しなくなった場合に，その寸法指標を最尤推定値と考える．

なおピットマンモデルは Pitman (1995)によって提案されたモデルで，2 つのパラメータを持つ．現在この分野で用いられている超母集団モデルの中では，最も良い結果が得られているモデルの一つである．ピットマンモデルについては，Pitman (1996)，Yamato and Sibuya (2000)，Hoshino (2001)を参照されたい．

(E) 1 サイズ探索法

サイズ 1 以外の任意の 1 つのサイズの度数を $\pm 1$ の範囲で増減させる以外は(D)の方法と同様である．

（F） 打ち切り 1 サイズ探索法

観測された標本寸法指標の内で，小さいサイズのみを用いて(E)と同様に母集団寸法指標を推定する方法であり，後の適用例ではサイズ 8 以下のみを用いることにする．母集団寸法指標で通常興味の対象となるのは，開示における危険性の指標としてしばしば用いられる $S_1$ や $S_2$ であるため，これらにあまり影響を与えない標本寸法指標の大きいサイズを打ち切ったものである．

制約条件と簡便法との様々な組み合わせについて母集団寸法指標の推定を行い，その精度について比較検討を行うが，例えば(A)と(a)との組み合わせを(A)-(a)のように表記する．

また比較のためモデルを用いる 2 種類のパラメトリック推定法も考える．パラメータの推定は最尤法を用いる．

（G） ピットマンモデルを用いた推定法

（H） ポアソンガンマモデルを用いた推定法

ポアソンガンマモデルは Bethlehem et al. (1990)によって提案され，この分野における超母集団モデルの中で最も早くから用いられていたモデルである．しかし実質的なパラメータを 1 個しか持たず，寸法指標への当てはまりに問題のあることが指摘されている．

## 4. 数値例による検討

### 4.1 制約条件の必要性とパラメトリック推定法との比較

まず制約条件の必要性の認識とパラメトリック法との比較のために，(A)-(a), (A)-(b), (B)-(b), (C)-(b), (G), (H)を用いて母集団寸法指標の推定を行う．母集団の大きさは計算時間上の制約から $N = 50$ として，母集団寸法指標を $(S_1, S_2, S_3, S_4) = (14, 6, 4, 3)$ とする．またポアソンガンマモデルで必要になる総セル数は，便宜上 $K = 10^4$ とする．

この母集団から大きさ 25 の標本をサンプリングし，標本寸法指標から母集団寸法指標を推定する．この一連の流れを 10 回繰り返し，実際の母集団寸法指標とその推定値との近さをみることにする．なお推定の際に，母集団のセルのサイズの最大値 $L$ は未知として十分大きくとるべきだが，ここでは $L = 5$ とした．以後の検討でも $L$ としては実際の母集団におけるサイズの最大値よりも若干大きい値を用いている．

結果を図 1 に示す．各グラフにおいて，横軸はセルのサイズを，縦軸はその度数を示す．また太い線は母集団の実際の寸法指標を，細い線は標本の寸法指標から推定された 10 個の寸法指標を表している．

(3.1)式そのものを用いて網羅的に尤度を求める(A)-(a)では，推定された寸法指標は大きく増減を繰り返しており，渋谷の「素朴な推定」と同様の結果が得られている．これに対して(A)-(b)の制約条件を用いる場合では，推定ごとにばらつきはあるものの概ね実際の寸法指標に近い結果が得られている．これは制約条件によって(a)での増減を押さえ込んでいると考えることができる．ポアソン近似を用いた(B)-(b)，個体追加法を用いた(C)-(b)の結果は(A)-(b)とほぼ同じである．これに対して(G), (H)では推定は安定しているものの，サイズによっては過大あるいは過小に推定されている．

### 4.2 制約条件についての比較

ここでは制約条件の比較を行うために，簡便法を(C)の個体追加法に固定し，(C)-(b), (C)-(c), (C)-(d), (C)-(e)によって 4 つの制約条件の比較を行う．また，比較のために(G)ピットマンモデルでも推定を行う．

用いる母集団はアメリカ合衆国で 1990 年に実施されたセンサスの個票データ(Bureau of the

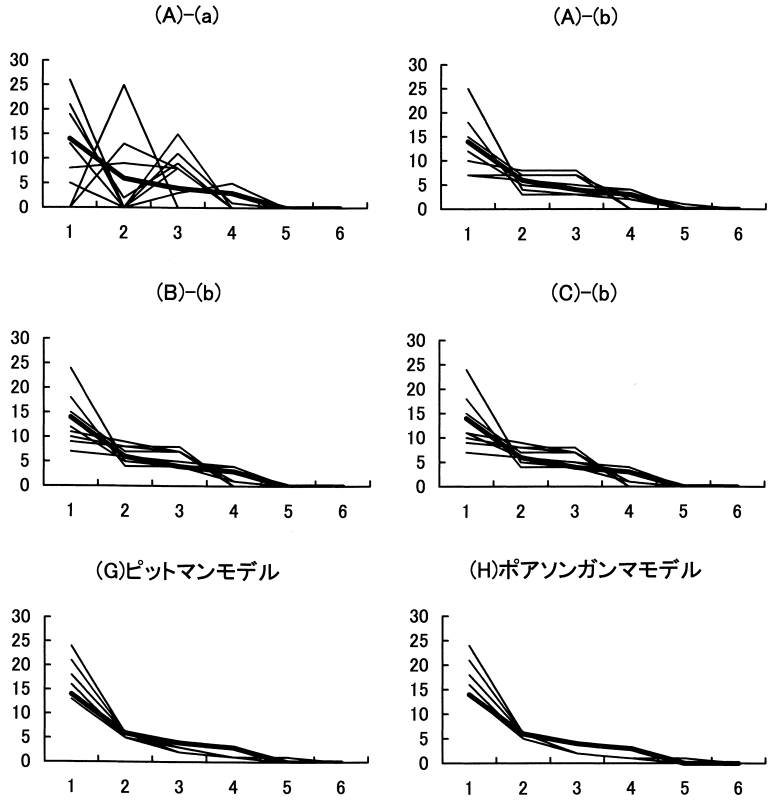


図 1. 制約条件の効果（$N = 50$, $n = 25$）.

Census (1993))の中のワシントン州のデータである．個票データには 1%抽出データと 5%抽出データがあり，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 1%抽出データの個人レコードのみを用いることにする．また，含まれる項目から年齢，性別，世帯主との続柄，配偶関係，出身地など 16 項目のみをキー変数として考えることにする．総セル数は項目のカテゴリー数の積として $K = 4.603 \times 10^{14}$ である．

ワシントン州の個票データには 49045 人分の個人レコードが含まれるが，ここではその中の 10000 人分を母集団と考え，そこから大きさ 5000 の標本を 10 回繰り返しサンプリングする．母集団の寸法指標は $(S_1, S_2, \ldots, S_{23}) = (7103, 577, 169, 66, 33, 19, 13, 8, 8, 5, 3, 7, 1, 6, 3, 0, 3, 1, 0, 0, 1, 2, 1)$ である．この母集団寸法指標は，サイズの小さい部分ではすべての制約条件を満たしている．

結果を図 2 に示す．縦軸はセル数の自然対数を表す．厳しい制約条件を用いるほど推定が安定していくことが読みとれる．

ところでピットマンモデルを用いた推定は提案した方法と比較して安定した推定が行われている．しかし前節の例と同様に，サイズごとに過大あるいは過小に推定されている．推定された $S_1$ の 10 個の推定値の平均を図 3 に示す．ノンパラメトリック推定法は標準誤差は大きいものの，比較的偏りの小さい推定が行われている．

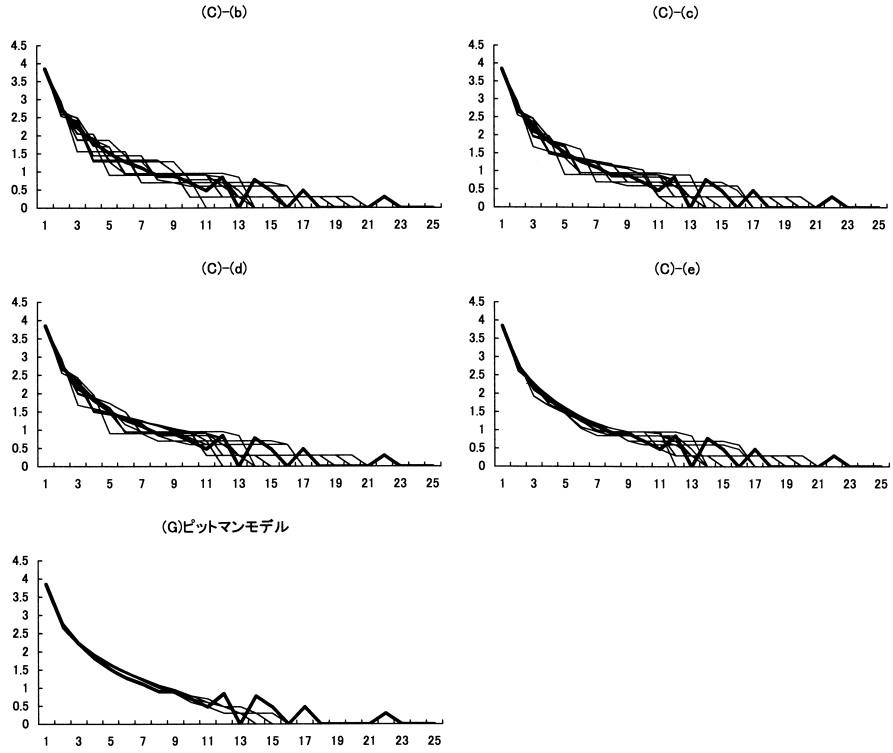


図 2. 制約条件の比較（$N = 10000,\ n = 5000$）.

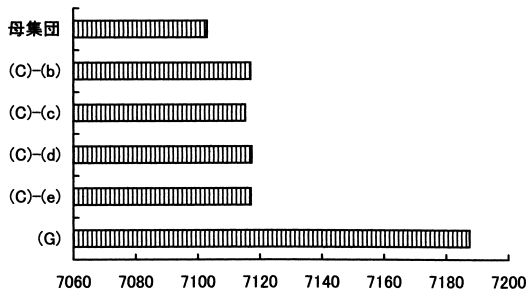


図 3. $S_1$ の推定値の平均（$N = 10000,\ n = 5000$）.

### 4.3 簡便法についての検討

次に，推定精度の観点でどの程度の簡便法を用いることが可能であるか検討を行う．ここでは制約条件を(e)の寸法指標の対数が下に凸という条件に固定し，簡便法として(C)-(e), (D)-(e), (E)-(e), (F)-(e)の比較を行う．なお，用いる母集団，標本は前節と同じである．結果を図 4 に示す.

どの簡便法を用いても特にサイズの小さい部分については十分良い推定が行われている．な

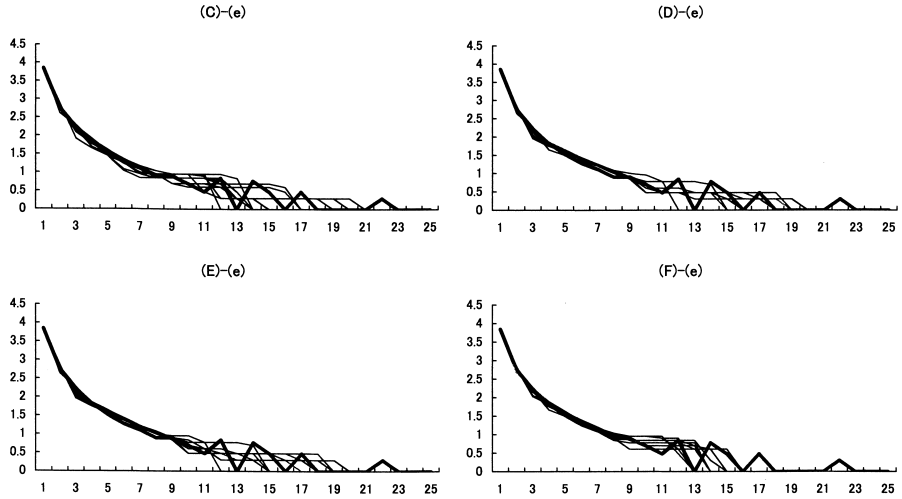


図 4. 簡便法の比較（$N = 10000,\ n = 5000$）.

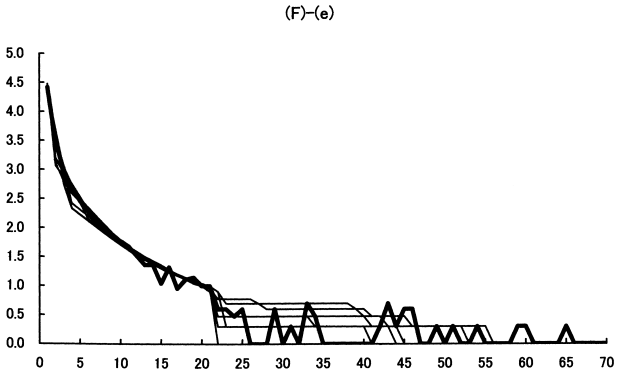


図 5. 大母集団の寸法指標の推定（$N = 244836,\ n = 5000$）.

おそれぞれの簡便法で 1 回の推定に要する計算時間は CPU : Intel Celeron プロセッサ 500 MHz のパーソナルコンピュータを用いて, (C)-(e)：約 67.3 分, (D)-(e)：約 46.4 分, (E)-(e)：約 3.5 分, (F)-(e)：約 2.2 分 である.

### 4.4 大きな母集団に対する適用の可能性

ここまで検討した例の抽出率はすべて 1/2 で，母集団の大きさも最大で 10000 だった．ここでは現実的な大きさの母集団として，アメリカのワシントン州におけるセンサスの 5%抽出の全個人レコード 244836 人分を考える．キー変数としては前節までと同じものを考える．母集団の寸法指標は $(S_1, S_2, \ldots, S_{10}, \ldots) = (87703, 12340, 4671, 2408, 1538, 1054, 713, 541, 397, 324, \ldots)$ となり，度数が 0 でない最大のサイズは 520 である．

母集団から約 1/50 の抽出率で大きさ 5000 の標本を 10 回サンプリングし，母集団寸法指標の対数が下に凸という制約条件の下，打ち切り 1 サイズ探索法を用いて母集団寸法指標の推定

を行う．結果を図 5 に示す．1 回の推定に約 1 日を要するものの，サイズの小さい部分では概ね満足できる推定結果が得られている．

## 5. おわりに

本論文では，標本寸法指標から母集団寸法指標をノンパラメトリック推定する方法についてまとめるとともに，ノンパラメトリック最尤推定する方法を新たに提案した．単純なこの推定法がこれまで提案されなかった理由として，モデルを用いる方法と比較して推定が不安定になることが予測されたことと，計算時間上の問題があげられる．

前者の点を解決するために，3.3 節において母集団の寸法指標に対して制約を加えた．これは非常に簡単な工夫ではあるが，母集団の寸法指標の性質が経験的に知られている場合には，単調減少や，対数が下に凸のような簡単な制約を置くことによって安定した推定が可能になる．

また後者の問題を解決するために，3.2 節でベルヌーイ抽出による尤度をポアソン分布の確率関数で近似する方法を提案するとともに，3.4 節ではすべての可能な母集団寸法指標について網羅的に尤度を計算する代わりに，個体追加法と探索法を提案した．特に，標本寸法指標の大きいサイズを打ち切ったものを用いることにより，現実的な大きさの母集団に対しても推定が可能になった．

## 謝　辞

本研究を進めるにあたり，東京大学の竹村彰通先生，高千穂大学の渋谷政昭先生を初めとして，共同研究のメンバーからは数々の助言を戴いた．ここに感謝いたします．

また査読者の有益な助言に対して感謝いたします．

本論文は統計数理研究所共同研究プログラム(14-共研-2024)「個票データの開示におけるリスクの評価と官庁統計データの公開への応用」，日本学術振興会科学研究費補助金・課題番号 13553001「官庁統計におけるサンプリング法の改善と個票データとしての開示に関する諸問題の研究」の研究成果に基づくものであり，日本学術振興会科学研究費補助金・課題番号 14208023 の援助も受けている．

## 参 考 文 献

Baayen, R. H. (2001). *Word Frequency Distributions*, Kluwer Academic Publishers, Dordrecht.

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.

Church, K. and Gale, W. (1991). A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams, *Computer Speech and Language*, **5**, 19–54.

Engen, S. (1978). *Stochastic Abundance Models*, Chapman and Hall, London.

Gale, W. A. and Sampson, G. (1995). Good-Turing frequency estimation without tears, *Journal of Quantitative Linguistics*, **2**, 217–237.

Goodman, L. A. (1949). On the estimation of the number of classes in a population, *Ann. Math. Statist.*, **20**, 572–579.

Greenberg, B. V. and Zayatz, L. V. (1992). Strategies for measuring risk in public use microdata file,

*Statist. Neerlandica*, **46**, 33–48.

Hoshino, H. (2001). Applying Pitman's sampling formula to microdata disclosure risk assessment, *Journal of Official Statistics*, **17**, 499–520.

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 年度日本学術振興会科学研究費補助金報告書(課題番号 08209107), 1–22.

Naranan, S. and Balasubrahmanyan, V. (1998). Models for power law relations in linguistics and information science, *Journal of Quantitative Linguistics*, **5**, 35–61.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions, *Probab. Theory Related Fields*, **102**, 145–148.

Pitman, J. (1996). Random discrete distributions invariant under size-biased permutation, *Advances in Applied Probability*, **28**, 525–539.

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.

佐井至道(2002). サイズインデックスの制約付き最尤推定, 岡山商大論叢, **37**(3), 61–79.

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

Sibuya, M. (1993). A random clustering process, *Ann. Inst. Statist. Math.*, **45**, 459–465.

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.

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.

Yamato, H. and Sibuya, M. (2000). Moments of some statistics of Pitman sampling formula, *Bull. Inform. Cybernet.*, **32**, 1–10.

Zipf, G. K. (1935). *The Psycho-biology of Language*, Houghton Mifflin, Boston.

Proceedings of the Institute of Statistical Mathematics Vol. 51, No. 2, 183–197 (2003) 197

# Simple Methods for Nonparametric Estimation of Population Size Indexes

Shido Sai

(Okayama Shoka University)

Observed sample size indexes and the estimated population size indexes are often used to assess the disclosure risk of microdata sampled from a population. The parametric method with superpopulation models such as a Poisson gamma model and a Pitman model is presently the main method used for estimation of population size indexes based on sample size indexes.

This article introduces some already proposed nonparametric estimation methods, and proposes the nonparametric maximum likelihood estimation method. There are two problems with the proposed method: the estimation result is unstable and enormous computing time is necessary.

In order to resolve the first problem, we set some simple restrictions for population size indexes, for example, monotone decreasing and downwards convex. These restrictions enable us to carry out stable estimation.

To remove the second problem, we approximate the likelihood function under Bernoulli sampling with the product of Poisson probability functions, and propose some convenient computational methods that do not need exhaustive computation. The target of estimation is almost always indexes of lower sizes. Thus, estimation of actual population may be feasible if we use sample size indexes of only lower sizes.

Key words: Microdata, key variable, population size indexes, sample size indexes.