『岡山商大論叢』（岡山商科大学）
第42巻第1号　2006年7月
Journal of OKAYAMA SHOKA UNIVERSITY
Vol. 42　No. 1　July 2006

《論　説》

ペナルティ関数を利用した母集団寸法指標のノンパラメトリック推定

佐　井　至　道

Nonparametric Estimation for Population Size Indexes
Using Penalty Functions

Shido SAI

1　はじめに

標本調査で得られた個票データを公開する際のリスク評価には、標本寸法指標から推定された母集団寸法指標を用いるのが一般的である。佐井（2002、2003）において、超母集団モデルを用いない母集団寸法指標の制約付きノンパラメトリック最尤推定法を提案したが、計算時間上の制約から大きな母集団には適用できないという問題点があった。佐井（2005）では、母集団を部分母集団に分割し、個々の部分母集団に対して独立に母集団寸法指標を推定することによる改善策を提案した。

本論文では制約付きノンパラメトリック推定法について、制約条件をペナルティー関数で置き換えるなど、いくつかの改善を図ることによって、計算時間上の問題を克服する可能性を探る。

まず2節において、母集団寸法指標の制約付きノンパラメトリック推定について概略を述べる。

次に3節では、探索する母集団寸法指標を非負整数から非負実数に拡張するとともに、母集団寸法指標に対する制約条件を渋谷（2005）の提案した滑らかなペナルティー関数に置き換え、尤度関数への取り込みを図る。

4節ではアメリカ合衆国で実施されたセンサスの個票データのうち、大きさ1万程度のレコードを母集団と見なし、そこからサンプリングされた標本の寸法指標を基に提案した方法で推定を行い、その有用性を検証するとともに他の推定方法との比較を行う。提案した推定法では事前にペナルティー関数のパラメータや探索方法を決めておく必要がある。5節では、それらの設定を変えながら大きさ25万程度の母集団寸法指標を推定し、推定結果や計算時間の比較を行う。提案した推定方法を官庁統計に適用する場合には、更に大きな母集団の寸法指標を推定する必要があるが、6節では都道県単位の推定を想定し、大きさ500万弱程度の母集団寸法指標の推定を試みる。

これらの検討の結果、ペナルティ関数を導入することにより、母集団寸法指標のノンパラメトリック推定法における計算時間上の問題もほぼ解決され、提案した推定方法の有用性が示されたと考える。

2　制約付きノンパラメトリック推定の概略

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の組み合わせに基づいて K 個のセルに分けられているものとする。そのうちサイズ l のセル数、すなわち l 個の個体が入っているセル数を S_l と表し母集団寸法指標と呼ぶ。また標本（個票データ）の大きさを n、抽出率を $\lambda = n/N$ として、標本寸法指標には $s_{l'}$ という表記を用いる。

佐井（2002、2003）において、標本寸法指標からの母集団寸法指標の制約付きノンパラメトリック推定法を提案した。ここではその要点のみを

簡潔に述べる。

標本寸法指標 $(s_1, s_2, ..., s_L)$ が得られたときに、尤度を最大にするような非負の母集団寸法指標 $(S_1, S_2, ..., S_L)$ を求めるが、その尤度関数は抽出率が十分小さいときに

$$\mathrm{L}(s_1, s_2, ..., s_L \mid S_1, S_2, ..., S_L) = c \cdot \prod_{l'=1}^{L} \frac{e^{-\mu_{l'}} \mu_{l'}^{s_{l'}}}{s_{l'}!} \qquad (1)$$

と近似できる。ただし L はサイズの最大値、c は定数であり、

$$\mu_{l'} = \sum_{l=l'}^{L} S_l \cdot {}_l C_{l'} \lambda^{l'} (1-\lambda)^{l-l'} \qquad (2)$$

である。

しかし、(1)式を最大にする母集団寸法指標は渋谷（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。そこで、佐井（2003）では、推定の際に母集団寸法指標にいくつかの制約（(a)～(e)）を置く場合について、実データを基に比較を行った。ここでは本論文の検討に用いる制約条件のみを列挙する。

(a)　$S_l \geq 0$、$\sum_{l=1}^{L} l \cdot S_l = N$ のみ

(b)　(a)＋母集団寸法指標が単調減少 $(S_1 \geq S_2 \geq \cdots \geq S_L)$

(e)　(a)(b)＋母集団寸法指標の対数が下に凸

$(2 \cdot \log S_l \leq \log S_{l-1} + \log S_{l+1})$

また、標本寸法指標を生成する可能性のあるすべての母集団寸法指標 $(S_1, S_2, ..., S_L)$ について網羅的に尤度を計算することは計算時間上困難であるため、佐井（2003）では数種類の最大尤度の探索法（(A)～(F)）について比較を行った。探索法についても、本論文の検討に関係するものだけを列挙する。なお、これらの探索法では S_l を整数値として扱っている。

(D)　2サイズ探索法

(E) 1サイズ探索法

(F) 打ち切り1サイズ探索法

上記3つの探索法は、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。

(D)では、サイズ1以外の任意の2つのサイズ（1つのサイズは5以下）の頻度を±1の範囲で増減させ、個体数の過不足をサイズ1の頻度で調整する場合を考え、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。以後このステップを繰り返し、どのような変化をさせても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。

(E)の方法はサイズ1以外の任意の1つのサイズの頻度を±1の範囲で増減させる以外は(D)の方法と同様である。探し出せる最大尤度は(D)と比較して若干小さい場合もあるが、要する時間は(D)の1/10程度である。

(F)では、標本寸法指標のサイズの大きい部分を打ち切り、小さいサイズの寸法指標のみから(E)と同様の方法で推定を行う。これは標本寸法指標の大きいサイズについての情報の有無が、母集団寸法指標の重要な推定目標であるS_1などの推定に大きな影響を与えないことを念頭に置き、計算時間の短縮を図ったものである。

3 ペナルティ関数を利用したノンパラメトリック推定

前節で紹介した最大尤度の探索法では、母集団が大きくなるにつれて計算時間上の問題が障害となってくる。大きさn=5000の標本の寸法指標から大きさN=244836の母集団の寸法指標を推定する例（後に検討する）では、“(E) 1サイズ探索法”でもパーソナルコンピュータによる計算に10日以上を要することがあり、官庁統計のように、大きさ10^5～10^8程度の母集団の寸法指標を推定する場合には計算の高速化が不可欠となる。

前節の方法では母集団寸法指標$(S_1, S_2, ..., S_L)$を非負整数に限定してい

た。そのため、例えば“(E) 1 サイズ探索法”を利用する場合、あるステップで S_{100} を 1 だけ増やすためには、個体としては 100 個増加させることになり、総個体数を調整するために S_1 を 100 だけ減少させなければならない。そのため、大きいサイズの頻度の増減が妨げられていることも想定される。そこで母集団寸法指標を非負実数に拡張することによって、よりスムーズな最大尤度の探索の実現を図る。

次に、渋谷（2005）の提案した滑らかなペナルティー関数

$$\mathrm{Pnlt}(x;\varepsilon) = \varepsilon \log\left(e^{\frac{x}{\varepsilon}} + 1\right) \tag{3}$$

の導入を図る。ただし、ε はパラメータである。(3)式の微分は

$$\frac{d\,\mathrm{Pnlt}(x;\varepsilon)}{dx} = \frac{e^{\frac{x}{\varepsilon}}}{e^{\frac{x}{\varepsilon}} + 1} \tag{4}$$

となり、ロジスティック分布の分布関数を横方向に ε 倍だけ拡大したものとなる。

図 1 に $\varepsilon = 0.01, 0.001$ としたペナルティー関数 $y = \mathrm{Pnlt}(x;\varepsilon)$ のグラフを示す。$\varepsilon \to 0$ のとき、グラフは $y = 0\,(x < 0)$, $y = x\,(x \geq 0)$ に近づくため、ε は原点付近での滑らかさを表すパラメータと考えることができる。

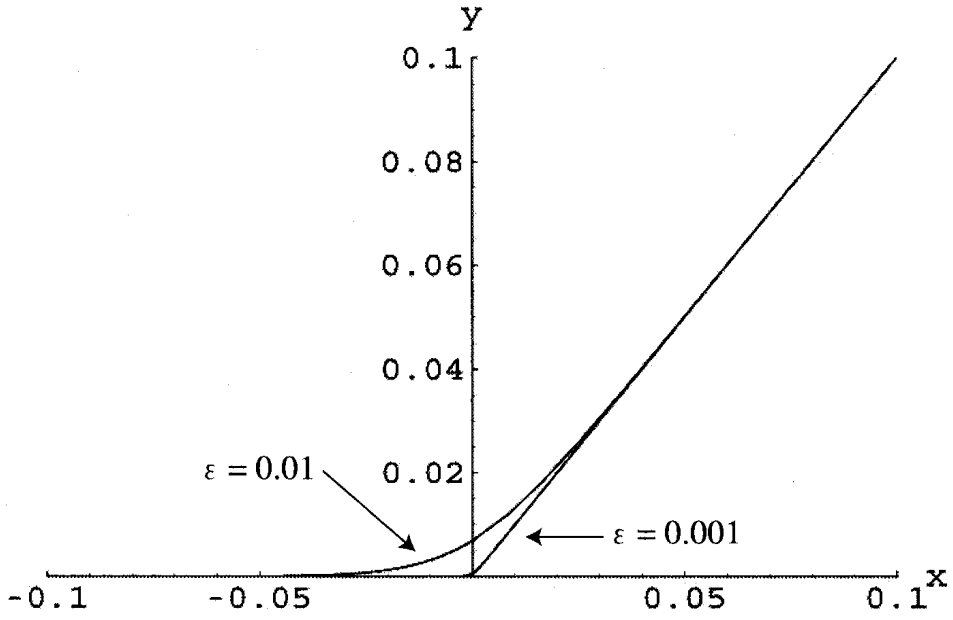

図 1：滑らかなペナルティー関数

ノンパラメトリック推定法では、母集団寸法指標に対して適切な制約条件を置かなければ安定した推定を行うことはできない。前節で述べた探索法では、母集団寸法指標の推定値が制約条件を満たす領域の外部に飛び出さないように、ステップごとにチェックする必要が生じた。ここでは、制約条件をペナルティー関数として表し直すことにより、尤度関数への取り込みを図る。

制約条件として“(e)母集団寸法指標の対数が下に凸”を用いる場合、ペナルティー関数を取り込んだ対数尤度関数を次のように定義する。

$$
\begin{aligned}
&\mathrm{F}(s_1,\ldots,s_L \mid S_1,\ldots,S_L)\\
&=\log \mathrm{L}(s_1,\ldots,s_L \mid S_1,\ldots,S_L)\\
&\quad - c_1\sum_{l=1}^{L}\mathrm{Pnlt}(-S_l;\varepsilon_1) - c_2\sum_{l=2}^{L}\mathrm{Pnlt}(-(S_{l-1}-S_l);\varepsilon_2)\\
&\quad - c_3\sum_{l=2}^{L-1}\mathrm{Pnlt}(-(\log S_{l-1}+\log S_{l+1}-2\log S_l);\varepsilon_3)\\
&=c+\sum_{l'=1}^{L}\{-\mu_{l'}+s_{l'}\log\mu_{l'}-\log(s_{l'}!)\}\\
&\quad - c_1\sum_{l=1}^{L}\varepsilon_1\log\left(e^{-\frac{S_l}{\varepsilon_1}}+1\right) - c_2\sum_{l=2}^{L}\varepsilon_2\log\left(e^{-\frac{S_{l-1}-S_l}{\varepsilon_2}}+1\right)\\
&\quad - c_3\sum_{l=2}^{L-1}\varepsilon_3\log\left(e^{-\frac{\log S_{l-1}+\log S_{l+1}-2\log S_l}{\varepsilon_3}}+1\right) \qquad (5)
\end{aligned}
$$

3つのペナルティー関数は順に、母集団寸法指標が非負、母集団寸法指標が単調減少、母集団寸法指標の対数が下に凸の制約条件を表し、ε_1、ε_2、ε_3 はそれぞれのペナルティー関数の滑らかさを表すパラメータ、c_1、c_2、c_3 はそれぞれのペナルティー関数の関数Fへの影響力を表すパラメータである。

関数Fを各 S_l $(l=1,2,\ldots,L)$ で

$$\frac{\partial \mathrm{F}}{\partial S_l} = -\sum_{l'=1}^{l} {}_lC_{l'}\,\lambda^{l'}(1-\lambda)^{l-l'}\left\{1-\frac{s_{l'}}{\sum_{k=l'}^{L} S_k\cdot {}_kC_{l'}\lambda^{l'}(1-\lambda)^{k-l'}}\right\}$$
$$+c_1\cdot\frac{e^{-\frac{S_l}{\varepsilon_1}}}{e^{-\frac{S_l}{\varepsilon_1}}+1}-c_2\cdot\frac{e^{-\frac{S_{l-1}-S_l}{\varepsilon_2}}}{e^{-\frac{S_{l-1}-S_l}{\varepsilon_2}}+1}+c_2\cdot\frac{e^{-\frac{S_l-S_{l+1}}{\varepsilon_2}}}{e^{-\frac{S_l-S_{l+1}}{\varepsilon_2}}+1}$$
$$-c_3\cdot\frac{2e^{-\frac{\log S_{l-1}+\log S_{l+1}-2\log S_l}{\varepsilon_3}}}{e^{-\frac{\log S_{l-1}+\log S_{l+1}-2\log S_l}{\varepsilon_3}}+1}+c_3\cdot\frac{e^{-\frac{\log S_{l-2}+\log S_l-2\log S_{l-1}}{\varepsilon_3}}}{e^{-\frac{\log S_{l-2}+\log S_l-2\log S_{l-1}}{\varepsilon_3}}+1}$$
$$+c_3\cdot\frac{e^{-\frac{\log S_l+\log S_{l+2}-2\log S_{l+1}}{\varepsilon_3}}}{e^{-\frac{\log S_l+\log S_{l+2}-2\log S_{l+1}}{\varepsilon_3}}+1} \qquad (6)$$

と偏微分し、0と置いた連立方程式を解くことによって関数Fの最大値を得ることが最善であるが、数学的に解くことは困難である。

そこで、ピットマンモデルで推定された母集団寸法指標の推定値を初期値として、ステップごとに、Fが増加する方向に各$\hat{S}_l$を独立に移動させる勾配法によって最大値を探索する。

$\hat{S}_{l(p)}$をp番目のステップにおける母集団寸法指標の推定値とするとき、

$$\hat{S}_{l(p+1)} = \hat{S}_{l(p)} + t_{l(p)}\cdot \mathrm{g}\left(\frac{\partial \mathrm{F}}{\partial S_l}\right) \qquad (7)$$

によって、$p+1$番目のステップの推定値$\hat{S}_{l(p+1)}$を求める。ただし、g(·)は微係数を調整する関数である。また$t_{l(p)}$は移動する距離をコントロールする係数で、サイズl、そのステップでの母集団寸法指標の推定値$\hat{S}_{l(p)}$などを基に各ステップで変化させる。

4　数値例による検討

この節では、提案した推定法と従来の推定法との比較を数値例を用いて行う。

対象とするのは、アメリカ合衆国で1990年に実施されたセンサスの個票データ（Bureau of the Census（1993））の中のワシントン州のデータである。公表されている個票データには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、レコードに含まれる項目から、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 $K = 4.603 \cdot 10^{14}$ 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の標本をサンプリングする。母集団寸法指標は $(S_1, S_2, ..., S_{23}) = (7103, 577, 169, 66, 33, 19, 13, 8, 8, 5, 3, 7, 1, 6, 3, 0, 3, 1, 0, 0, 1, 2, 1)$、標本寸法指標は $(s_1, ..., s_{15}) = (3934, 266, 39, 27, 12, 6, 8, 8, 4, 2, 2, 0, 0, 0, 1)$ である。母集団寸法指標は、サイズの小さい部分において“(e)母集団寸法指標の対数が下に凸”を満たしている。ノンパラメトリック法による推定の際には $L = 30$ とする。

上記の標本寸法指標を基に、ピットマンモデルを用いた推定法（Pitmanと表記）、制約条件として“(e)母集団寸法指標の対数が下に凸”、探索方法として“(D) 2サイズ探索法”を用いた従来の制約付きノンパラメトリック最尤推定法（$(D)(e)_{30}$ と表記）、前節で提案したペナルティー関数を利用した勾配法による推定法（$Pnlt(e)_{30}$ と表記）の比較を行う。

ピットマンモデルはPitman（1995）が提案したモデルで、母集団寸法指標の期待値は

$$\mathrm{E}(S_l) = \frac{N}{l}\binom{N-1}{l-1}\frac{(1-\alpha)^{[l-1]}(\theta+\alpha)^{[N-l]}}{(\theta+1)^{[N-1]}} \tag{8}$$

と書くことができる。なお $x^{[k]} = x(x+1)\cdots(x+k-1)$ である。また α、θ はパラメータで、その推定にはモーメント法を用いる。

ペナルティー関数を利用した勾配法による推定法では、$c_1 = c_2 = 10.0$、$c_3 = 5.0$、$\varepsilon_1 = \varepsilon_2 = 10^{-4}$、$\varepsilon_3 = 10^{-3}$ とする。また(7)式において

$$\mathrm{g}(x) = \mathrm{sign}(x) \cdot |x|^{0.5} \tag{9}$$

$$t_{l(p)} = \mathrm{b}(p) \cdot \hat{S}_{l(p)} / l^{0.7} \qquad (10)$$

として、p 番目のステップにおける母集団寸法指標の推定値が大きいほど大きく変化させ、サイズ l が大きいほど小さく変化させる。また、b(p) は 0.1 を初期値として、3 ステップ連続して F が増加しなかった場合には 0.9 倍し、2 ステップ連続して F が増加した場合に 1.1 倍し、10^{-4} を下回った時点で探索を終了する。

なお、このようにして母集団寸法指標の推定値を変化させた場合、一般に推定値は超平面 $\sum_{l=1}^{L} l \cdot \hat{S}_{l(p)} = N$ 上では得られないが、ここでは推定値と原点とを結ぶ直線とこの超平面との交点へ移動させる補正をステップごとに行う。また、制約条件を大きく逸脱すると F などの計算においてオーバーフローしてしまう恐れがあるため、$\hat{S}_{l(p)}$ が負とならないように、また $\hat{S}_{l(p)} - \hat{S}_{l+1(p)}$ が大きく負とならないようにする調整を行う。

従来のノンパラメトリック推定法では“(e) 母集団寸法指標の対数が下に凸”という制約条件は頻度が 10 以上のサイズにのみ用い、頻度が 10 未満のサイズには“(b) 母集団寸法指標が単調減少”の制約までを課すことにする。提案した推定法では、同じ制約条件を課す場合とともに、頻度が 1 以上まで条件 (e) を課す場合についても推定値を求め、前者を $\mathrm{Pnlt(e)}_{30} - 1$、後者を $\mathrm{Pnlt(e)}_{30} - 2$ と表記する。

表 1 に母集団寸法指標の各サイズ l の頻度の推定値 $\hat{S}_l$ を示す。(S_l の列は実際の母集団寸法指標の値である。) どの結果も大きな差異はなく、特に提案した方法と従来のノンパラメトリック法の推定結果は、母集団寸法指標の実数化や制約条件の違いなどから想定される程度の差しかない。

5　種々の設定についての検討

この節では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 人分を母集団として考える。キー変数としては前節と同じものを用いる。母集団寸法指標は $(S_1, S_2, ..., S_{10}, ...) = (87703, 12340,$

4671, 2408, 1538, 1054, 713, 541, 397, 324, ...) であり、頻度が0でない最大のサイズは520である。この母集団からサンプリングされた大きさ5000の標本の寸法指標は $(s_1, ..., s_{12}) = (3937, 249, 70, 29, 10, 7, 5, 4, 4, 2, 0, 2)$ である。

表1：母集団寸法指標の推定値（N=10000、n=5000）

l	S_l	Pitman	(D) $(e)_{30}$	Pnlt $(e)_{30} - 1$	Pnlt $(e)_{30} - 2$
1	7103	7152.49	6973.00	6973.54	6968.52
2	577	568.99	750.00	795.51	801.41
3	169	186.38	91.00	90.90	92.34
4	66	84.16	57.00	36.96	34.22
5	33	44.27	36.00	26.54	24.85
6	19	25.48	23.00	19.11	18.11
7	13	15.57	15.00	13.80	13.24
8	8	9.92	10.00	10.00	9.72
9	8	6.53	7.00	4.59	8.08
10	5	4.41	3.00	4.59	6.75
11	3	3.03	3.00	4.59	5.65
12	7	2.12	3.00	4.59	4.74
13	1	1.50	3.00	4.59	3.98
14	6	1.08	3.00	4.59	3.34
15	3	0.78	3.00	4.59	2.81
16	0	0.57	3.00	4.59	2.36
17	3	0.42	3.00	4.59	1.98
18	1	0.31	3.00	0.54	1.67
19	0	0.23	1.00	0.15	1.40
20	0	0.17	1.00	0.15	1.18
21	1	0.13	1.00	0.15	1.00
22	2	0.10	1.00	0.15	0.57
23	1	0.08	0.00	0.15	0.28
24		0.06	0.00	0.15	0.21
25		0.05	0.00	0.15	0.16
26		0.04	0.00	0.15	0.13
27		0.03	0.00	0.15	0.10
28		0.02	0.00	0.15	0.07
29		0.02	0.00	0.15	0.05
30		0.01	0.00	0.15	0.03

まず、前節で用いた3つの推定法によって、s_8 で打ち切った標本寸法指標を基に、$L = 160$ と設定して母集団寸法指標を推定する。その際、抽出率 $\lambda = 5000/244836 = 0.02042$ を保つように母集団の大きさも便宜上減少させる。

従来のノンパラメトリック推定法では“(F) 打ち切り1サイズ探索法”を用いるが、最も計算時間が短いこの方法でも計算に要する時間は1日を超える。従って、用いる標本寸法指標と母集団寸法指標のサイズの最大値がこの程度を超えると、従来のノンパラメトリック推定法では推定が困難になる。

提案した方法では $c_1 = c_2 = 10.0$、$c_3 = 5.0$、$\varepsilon_1 = \varepsilon_2 = 10^{-4}$、$\varepsilon_3 = 10^{-3}$ と

表2：s_8までの標本寸法指標からの母集団寸法指標の推定値

（N=244836、n=5000）

l	S_l	Pitman	(F)$(e)_{160}$	Pnlt$(e)_{160}$
1	87703	102856.43	95152	65893.63
2	12340	8970.27	13231	7895.42
3	4671	3488.02	3034	5597.06
4	2408	1882.34	2041	3997.35
5	1538	1186.15	1373	2873.51
6	1054	819.00	924	2078.10
7	713	600.79	622	1511.48
8	541	460.13	419	1105.46
9	397	363.97	352	812.92
10	324	295.23	299	601.06
151	0	0.63	2	2.76
152	0	0.62	2	2.76
153	1	0.61	2	2.76
154	0	0.60	2	2.76
155	0	0.59	2	2.76
156	0	0.57	2	2.76
157	0	0.56	2	2.76
158	1	0.55	2	2.76
159	0	0.54	2	2.76
160	0	0.53	2	2.76

する。また(7)式において

$$g(x) = \mathrm{sign}(x) \cdot |x|^{0.5} \tag{11}$$

$$t_{l(p)} = b(p) \cdot \hat{S}_{l(p)} / l^{0.6} \tag{12}$$

とする。$b(p)$ については 0.1 を初期値として、6 ステップ連続して F が増加しなかった場合には 0.75 倍し、2 ステップ連続して F が増加した場合に 1.25 倍し、10^{-4}を下回った時点で探索を終了する。

なお、従来のノンパラメトリック推定法では頻度 10 以上のサイズまで、提案した方法では頻度 1 以上のサイズまで制約条件 (e) を課す。

結果を表 2 に示す。パーソナルコンピュータ（CPU：インテル Pentium M プロセッサ 750、1.86GHz）で計算に要する時間は、従来のノンパラメトリック推定法（$(F)(e)_{160}$）では 1 日程度を要するのに対して、提案した推定法（$Pnlt(e)_{160}$）でも 7 時間程度かかり、計算時間の大幅な短縮は実現できていない。

次に、すべての標本寸法指標 $(s_1, ..., s_{12})$ を基に、ピットマンモデルを用いる推定法と提案した推定法で母集団寸法指標を求める。

提案した方法では $c_1 = c_2 = 10.0$、$c_3 = 5.0$、$\varepsilon_1 = \varepsilon_2 = 10^{-4}$、$\varepsilon_3 = 10^{-3}$ とする。また(7)式において

$$g(x) = \mathrm{sign}(x) \cdot |x|^{0.4} \tag{13}$$

$$t_{l(p)} = b(p) \cdot \hat{S}_{l(p)} / l^{0.5} \tag{14}$$

とする。$b(p)$ については 0.1 を初期値として、6 ステップ連続して F が増加しなかった場合には 0.75 倍し、2 ステップ連続して F が増加した場合に 1.25 倍し、10^{-4}を下回った時点で探索を終了する。なお、制約条件 (e) を頻度 1 以上のサイズまで課す。

$L = 200$ とした場合の提案した方法による推定結果を表 3 の $Pnlt(e)_{200}$ の列に示す。計算に要する時間は 155.8 分に減少している。

$\hat{S}_{200} = 1.62$ は十分小さいとは言えず、母集団寸法指標のサイズを 200 で

打ち切っていることが、重要な推定目標である S_1 などの推定に影響を与えていることも懸念される。そこで 200 を超えるサイズについても飛び飛びに頻度が正になることを認める方法についても検討を行う。

ここでは $l = 300, 400, 500, 1000, 2000, 3000, 4000, 5000$ の 8 つのサイズの

表 3：すべての標本寸法指標からの母集団寸法指標の推定値
（N＝244836、n＝5000）

l	S_l	Pitman	Pnlt (e)$_{200}$	Pnlt (e)$_{5000}$	Pnlt (e)$_{5000+}$
1	87703	105625.19	96623.54	100069.25	100236.79
2	12340	8740.11	8162.45	8487.76	8482.46
3	4671	3376.79	3327.78	3351.52	3338.70
4	2408	1817.26	1884.35	1840.36	1826.55
5	1538	1143.50	1247.96	1180.42	1166.89
6	1054	788.98	959.66	829.68	816.65
7	713	578.59	784.42	619.57	605.23
8	541	443.12	645.02	483.05	469.30
9	397	350.58	533.07	388.95	377.70
10	324	284.47	442.55	321.21	312.75
191	0	0.37	1.62	0.52	0.47
192	0	0.36	1.62	0.51	0.46
193	2	0.36	1.62	0.51	0.46
194	0	0.35	1.62	0.50	0.45
195	0	0.35	1.62	0.50	0.45
196	2	0.34	1.62	0.50	0.44
197	1	0.33	1.62	0.49	0.44
198	1	0.33	1.62	0.49	0.43
199	0	0.32	1.62	0.49	0.43
200	1	0.32	1.62	0.48	0.43
300			—	10.93	11.58
400			—	9.02	9.52
500			—	1.74	1.37
1000			—	0.05	0.03
2000			—	0.05	0.03
3000			—	0.04	0.03
4000			—	0.03	0.02
5000			—	0.02	0.02

頻度についてのみ S_l が正になることを認める。例えば S_{500} はサイズ 401～500 を、S_{1000} はサイズ 501～1000 を併合しているものと考え、$S_{500}/(500-400) \geq S_{1000}(1000-500)$ という単調減少の制約条件をペナルティー関数として加える。

なお、S_{5000} まで単調減少という制約条件を課す場合(Pnlt(e)$_{5000}$)と、5000 を超えるサイズも受け持たせるため S_{5000} には単調減少の制約を置かない場合（Pnlt(e)$_{5000+}$）についてそれぞれ推定を行う。また、200 を超えるサイズについては対数が下に凸という制約を置かないことにする。

$c_1 = c_2 = 10.0$、$c_3 = 5.0$ とするが、200 を超えるサイズについては $c_2 = 1.0$ として単調減少に関するペナルティー関数の影響をやや弱める。また $\varepsilon_1 = \varepsilon_2 = 10^{-4}$、$\varepsilon_3 = 10^{-3}$ とする。(7)式において

$$\mathrm{g}(x) = \mathrm{sign}(x) \cdot |\, x \,|^{0.4} \tag{15}$$

$$t_{l(p)} = \mathrm{b}(p) \cdot \hat{S}_{l(p)} / l^{0.4} \tag{16}$$

として、b(p) は 0.1 を初期値とし、関数 F が 8 ステップ連続して F が増加しなかった場合には 0.75 倍し、2 ステップ連続して F が増加した場合に 1.25 倍し、10^{-4}を下回った時点で探索を終了する。なお、制約条件(e)を頻度 1 以上のサイズまで課す。

結果を表 3 の右の 2 列に示す。どちらの推定法でもサイズ 500 程度まで無視できない頻度が存在しており、サイズ 200 付近の推定結果の改善に伴い、Pnlt(e)$_{200}$ と比較して S_1 の推定値も多少変化している。S_{5000} に対する単調減少という制約条件の有無は、この場合、結果的に推定にはほとんど影響を与えていないと判断される。なお、計算に要する時間はそれぞれ 21.9 分、36.9 分で、$L = 200$ とした場合より 1 ステップでの計算量を増やす設定をしているにも関わらず、計算時間は逆に大幅に減少している。

次に、244836 人分の母集団から大きさ 5000 の標本を無作為に 10 回サンプリングし、各標本の寸法指標から母集団寸法指標を推定した結果を図 2 に示す。推定方法は、ピットマンモデルを用いた推定法（Pitman）、s_8

で打ち切った標本寸法指標を用いた従来のノンパラメトリック推定法（$(F)(e)_{160}$）、$L = 5000$ まで飛び飛びに頻度が正になることを許す提案した方法（$\mathrm{Pnlt}(e)_{5000+}$）である。

グラフにおいて横軸は母集団寸法指標のサイズ l、縦軸は推定された頻度の対数 $\log(\hat{S}_l)$ を表し、太線が実際の母集団寸法指標、細線が推定された10個の母集団寸法指標を示す。$\mathrm{Pnlt}(e)_{5000+}$ での平均計算時間は62.4分である。

ピットマンモデルを用いた推定は観測された標本寸法指標からあまり影響を受けず、極めて安定している。しかし、このグラフでは分かりにくいが、例えば S_1 の推定値にはバイアスが存在している。それに対し $L = 5000$ まで飛び飛びに頻度が正になることを許す提案した方法による推定値は、散らばりは大きいもののバイアスが小さいという特徴がある。

図3に、ピットマンモデルを用いた推定法と $\mathrm{Pnlt}(e)_{5000+}$ について、サイズ20以下における母集団寸法指標の推定値と実際の値の差 $\hat{S}_l - S_l$ を示す。ただしサイズの大きい部分では提案した推定法のバイアスが大きくなる場合もあることには注意を要する。

6　大母集団に対する適用の可能性

前節では244836人分の個票データを母集団とし、そこから取り出した5000人を標本とした。個票データはワシントン州の5%抽出データであったため、州の人口はおよそ $244836 \cdot 20 \doteqdot 4.897 \cdot 10^6$ と考えられる。この節では前節までと同様に5000人を標本とし、その標本寸法指標 $(s_1, \ldots, s_{12}) = (3937, 249, 70, 29, 10, 7, 5, 4, 4, 2, 0, 2)$ から大きさ $4.897 \cdot 10^6$ の母集団の寸法指標を推定する。この場合の抽出率は $5000/(4.897 \cdot 10^6) = 0.00103$ で、例えば日本の労働力調査における抽出率にほぼ等しく、都道府県単位での推定と同様の状況と見なすことができる。

推定には、ピットマンモデルを用いた推定法、$L = 5000$ まで飛び飛びに頻度が正になることを許す提案した方法（$\mathrm{Pnlt}(e)_{5000+}$）を用いる。

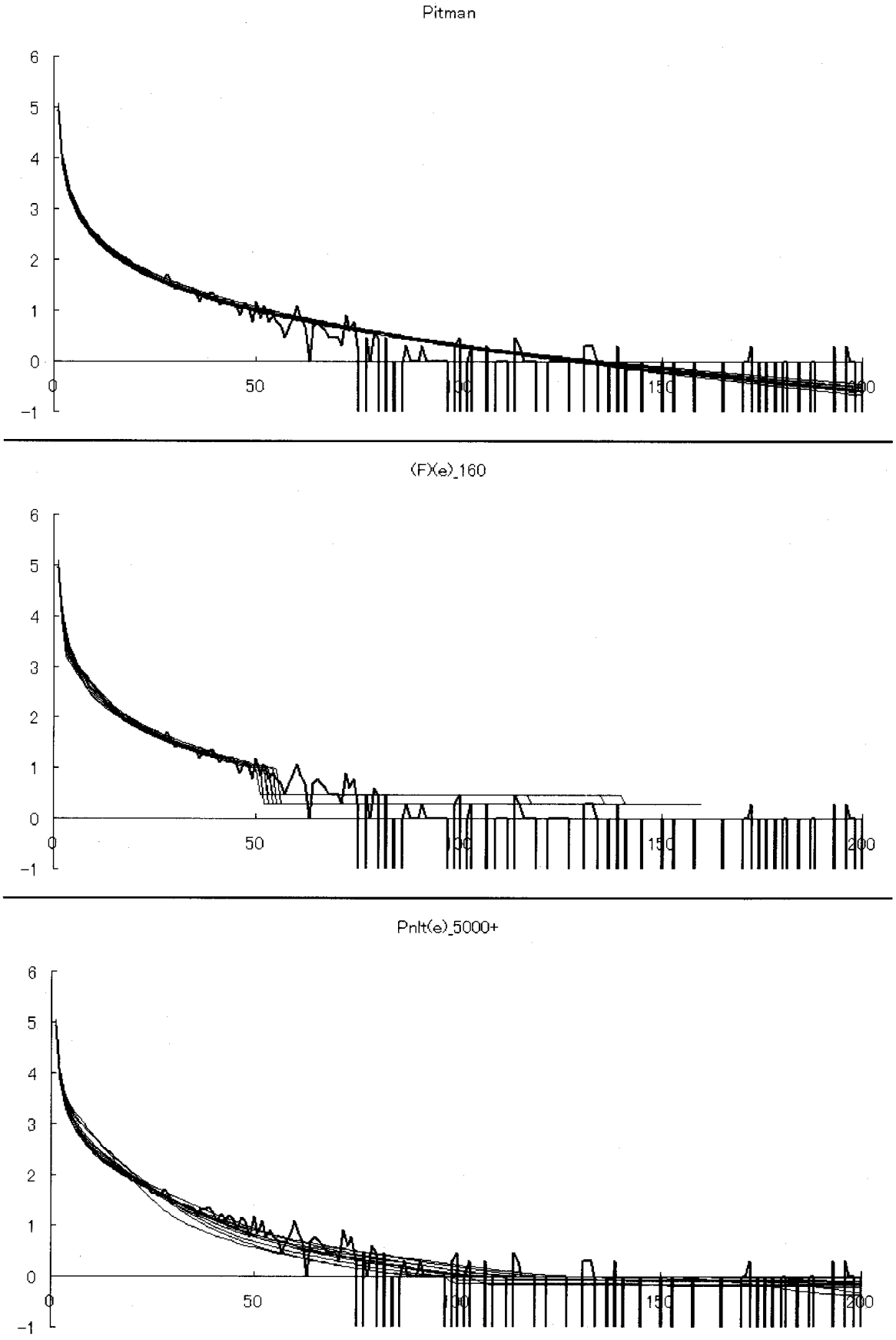

図 2：母集団寸法指標の 10 回の推定結果

後者では $c_1 = c_2 = 10.0$ (200を超えるサイズについては $c_2 = 1.0$)、$c_3 = 5.0$、$\varepsilon_1 = \varepsilon_2 = 10^{-4}$、$\varepsilon_3 = 10^{-3}$ とする。また(7)式において

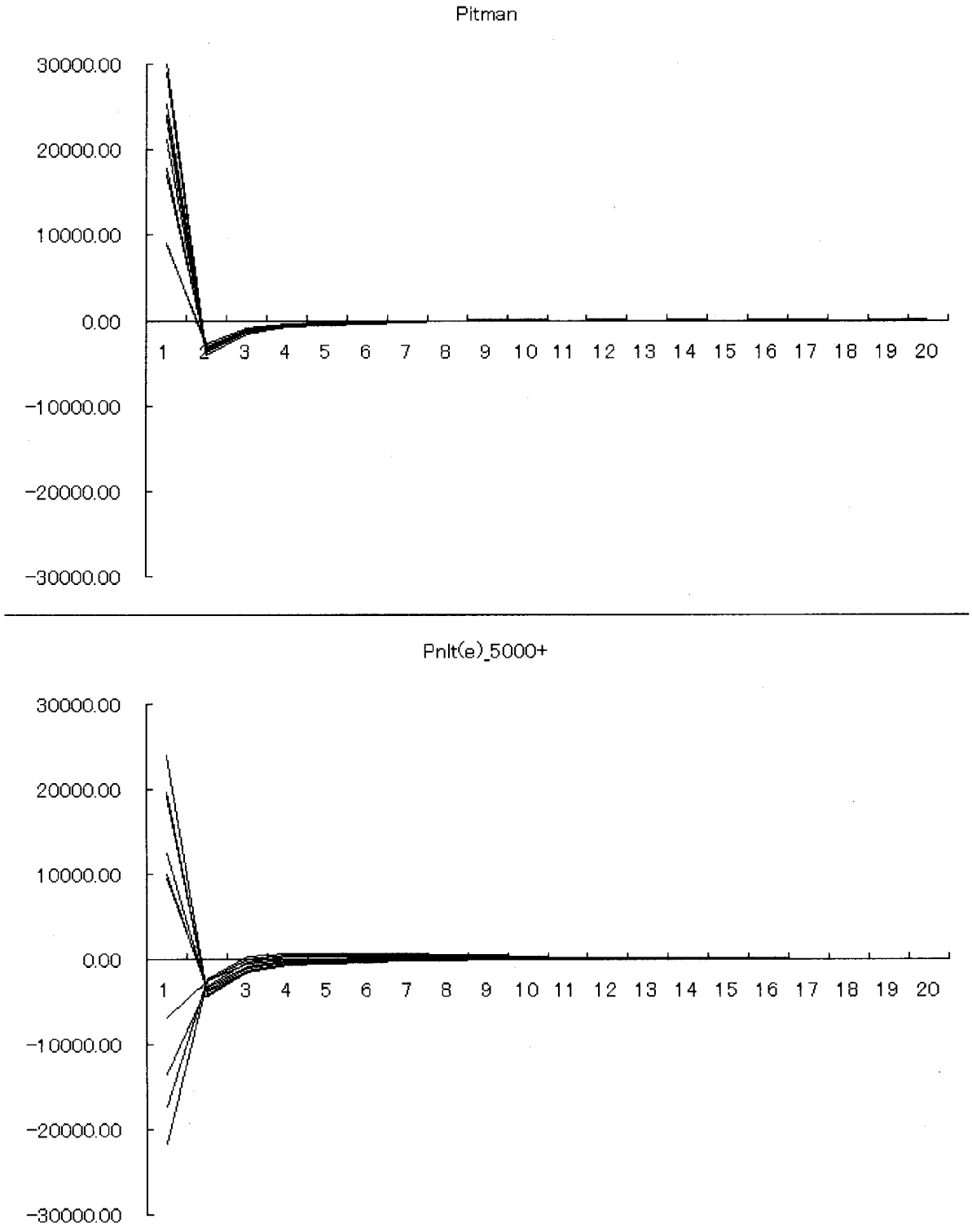

図3：母集団寸法指標の推定値と実際の値の差 $\hat{S}_l - S_l$

$$g(x) = \mathrm{sign}(x) \cdot |\, x \,|^{0.4} \tag{17}$$

$$t_{l(p)} = b(p) \cdot \hat{S}_{l(p)} / l^{0.2} \tag{18}$$

とする。$b(p)$ については 0.1 を初期値として、8 ステップ連続して F が増加しなかった場合には 0.75 倍し、1 ステップでも F が増加した場合には 1.25 倍し、10^{-4}を下回った時点で探索を終了する。なお、制約条件 (e) を頻度 1 以上のサイズまで課す。

結果を表 4 に示す。$\mathrm{Pnlt}(e)_{5000+}$ での計算時間は 166.0 分である。母集団寸法指標が未知であるため推定結果を評価することは難しいが、S_{5000} に単調減少の制約を課さない手法が機能していることが読み取れる。また計算時間から見ても十分に実用可能であると考えられる。

7 おわりに

本論文では、母集団寸法指標の制約付きノンパラメトリック最尤推定について、母集団寸法指標の推定値の頻度を実数に拡張するとともに、制約条件を渋谷（2005）の提案した滑らかなペナルティー関数に置き換えることにより計算時間の短縮を図った。

推定にはピットマンモデルを用いて推定された母集団寸法指標を初期値として用いるため、提案した推定法での結果がピットマンモデルでの結果と乖離するほど計算時間を要する。見方を変えると、そのような場合にはピットマンモデルが母集団寸法指標の推定に対して、強すぎる制約を置いていると考えることもできる。なお、初期値を変えて探索した場合、必ずしも同じ推定値に到達しないことにも注意を要する。

ペナルティー関数には事前に決めるべき数種類のパラメータがあり、ステップ間の推定値の移動方法についても事前に定める必要がある。その選択によっては、早い段階で探索が打ち切られてしまったり、逆に探索がなかなか終了しないことがあり、本論文では、尤度がほとんど増加しなくなった頃に探索が終了するようなパラメータや関数を設定している。

表 4：母集団寸法指標の推定値
($N=4.897\cdot10^6$、$n=5000$)

l	Pitman	Pnlt (e)$_{5000+}$
1	1283916.97	1469694.70
2	106891.90	110057.39
3	41551.94	39933.67
4	22498.95	20522.72
5	14244.26	12452.88
6	9888.41	8339.14
7	7296.08	5962.91
8	5622.15	4468.59
9	4475.38	3468.88
10	3653.67	2767.90
191	14.31	28.25
192	14.17	28.12
193	14.03	28.01
194	13.89	27.92
195	13.76	27.84
196	13.62	27.78
197	13.49	27.73
198	13.36	27.70
199	13.24	27.69
200	13.11	27.68
300		267.05
400		218.81
500		203.14
1000		721.50
2000		42.76
3000		1.17
4000		0.42
5000		53.44

本論文で用いた標本は比較的小さく、標本寸法指標の最大サイズも 12～20 であるため、提案した方法でも推定はやや不安定であった。本論文で提案した方法では頻度 1 以上のサイズまで制約条件 (e) を課すことが多かったが、頻度 10 以上のサイズまでしか課さない場合についても同様に推定を行ったところ、推定値の散布度は一般に大きくなった。まれなケー

スではあるが S_1 の推定値が実際の値の 1/3 程度となってしまうこともあった。従って、標本が小さい場合や抽出率が小さい場合には、強めの制約条件を課す必要があると考えられる。

ペナルティー関数の導入によって計算時間は大幅に短縮されたが、母集団の大きさ N や母集団寸法指標のサイズの最大値 L が増加するにつれて（特に後者が増加するにつれて）計算時間は一般的に増加する。その場合には本論文の後半で提案した、サイズの大きい部分について飛び飛びに頻度が正になることを許す方法が効果的と考えられる。本論文の例ではサイズ 200 を超える部分についてそのような方法を採用することによって計算時間が短縮されたが、これはサイズ 200 以下の部分がより自由に変化できることによるものではないかと思われる。

謝　辞

ペナルティー関数を導入するというアイディアに関して、丁寧に説明して頂いた高千穂大学の渋谷政昭先生に感謝致します。また本論文をまとめるにあたり、東京大学の竹村彰通先生、一橋大学の加納悟先生を初めとして、個票開示問題関連の研究グループのメンバーからは数々の有益な助言をいただきました。ここに感謝致します。

本論文は科学研究費補助金（課題番号　14208023、16203014）、統計数理研究所共同利用研究プログラム（17－共研－2028）の研究成果に基づくものである。

参考文献

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

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

［3］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、第37巻、第3号、61－79.

［4］佐井至道（2003）．母集団寸法指標のノンパラメトリック推定、統計数理、第51巻、第2号、183－198.

［5］佐井至道（2005）．部分母集団の情報を用いた母集団寸法指標の推定とノンパラメトリック推定の改良、岡山商大論叢、第40巻、第3号、33－35.

［6］佐井至道（2005）．母集団寸法指標の制約付きノンパラメトリック推定におけるいくつかの改善について、2005年統計関連学会連合大会予稿集、205－206.

［7］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.

［8］渋谷政昭（2005）．滑らかな罰金関数、科学研究費補助金研究会資料、1－3.