『岡山商大論叢』（岡山商科大学）
第40巻第3号　2005年6月
Journal of OKAYAMA SHOKA UNIVERSITY
Vol. 40　No. 3　June 2005

《論　　説》

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

佐　井　至　道

Estimation for Population Size Indexes Utilizing Information of Sub-Populations and Improvement of Nonparametric Estimation Method

Shido Sai

1　はじめに

個票データは調査された項目が個体ごとに並べられたデータで、第三者が個体を識別するために利用できる項目はキー変数、すべてのキー変数の分類の組み合わせはセルと呼ばれる。各セルに含まれる個体数の頻度分布が寸法指標であるが、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。しかし母集団が大きい場合には、使用する超母集団モデルによって推定結果の大きく異なることが明らかになってきた。

母集団寸法指標の推定では母集団の大きさを既知とするが、それ以外に用いられる母集団情報としては、一般に経験的に知られている母集団寸法

指標の形状をモデルに託し、それにより母集団寸法指標の推定に制約が加えられるだけである。

実際の推定では母集団寸法指標は未知であっても、他の調査結果から1個あるいは数個のキー変数の分類の組み合わせについて母集団の大きさは既知であることが多い。例えば、都道府県と性別がキー変数に含まれているとき、都道府県・男女別の人口が既知である場合がそれにあたる。そのような情報を利用することによって、推定を安定させる可能性がある。本論文ではこの点について検討を行う。

超母集団モデルを利用しない制約付きノンパラメトリック推定を佐井(2002,2003)において提案したが、この推定法には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る。母集団を小さく分割することにより、個々の部分母集団についてノンパラメトリック推定が可能となり、推定値を母集団全体について合算することによって母集団全体の寸法指標の推定も可能となる。本論文ではこの点について検討を行うとともに、主に大母集団への適用において問題となるいくつかの点についても併せて検討を行うこととする。

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。

2　大母集団の寸法指標の推定における問題点

本論文では科学研究費補助金研究グループが目的外使用した労働力調査のデータを用いる。目的外使用したのは1989年～2001年の13年間、156か月分の全データであるが、本論文で使用するのは2001年1月分のデータのうち、自衛官と受刑者を除く94003人の個人データである。詳細について

表１：労働力調査の個票データの項目と分類数

	項　　目	秘匿前	秘匿後
(1)	都道府県番号	(47)	47
(2)	県内一連番号	約88万	x
(3)	組符号	8	x
(4)	地域符号	6	6
(5)	抽出単位番号	約550	x
(6)	単位内世帯番号	約100	x
(7)	15歳以上人員計	約10	5
(8)	男計（15歳未満）	約8	4
(9)	女計（15歳未満）	約8	（x）
(10)	続柄	12	7
(11)	性別	2	2
(12)	配偶関係	4	4
(13)	就業状態	9	9
(14)	仕事の主従	3	x
(15)	就業時間	約130	12
(16)	従業上の地位	9	9
(17)	経営組織	4	4
(18)	産業	30	30
(19)	職業	15	15
(20)	従業員数	10	10
(21)	希望意識	6	x
(22)	異動符号	7	7
(23)	年号（出生年月）	(3)	x
(24)	年（出生年月）	約100	17
(25)	月（出生年月）	12	x
(26)	求職理由	6	x

（x：削除を意味する）

は佐井他（2004ａ，2004ｂ）を参照していただきたい。

総務省から提供された個票データの各個人レコードには都道府県番号や性別、就業状態などの26項目が含まれている。表１の「秘匿前」の列には各「項目」の分類数をあげるが、これらすべてをキー変数として考えると全個体が標本一意（標本でサイズ１）である。そこで過去の検討でほぼ安全と判断された形まで秘匿措置を施した。「秘匿後」の列にその分類数を表示しているが、例えば、県内一連番号を削除し就業時間を10時間刻みに

するなど、数個の項目について分類を粗くしている。この点についても、詳細は佐井他（2004 a）を参照していただきたい。

秘匿後に得られた標本寸法指標から母集団寸法指標の推定を行うが、母集団の大きさ（15歳以上人口）としては、総務省から発表されている2001年10月の推計人口を用いて$N = 1.09004 \times 10^8$とする。また実際のサンプリング法には層化無作為二段抽出法が用いられているが、ここでは簡単のため非復元単純無作為抽出を仮定する。

表2に2001年1月の「秘匿後」の個票データから求められる標本寸法指標（s_lと表記）と、超母集団モデルのポアソンガンマモデル（P–gと表記）、ピットマンモデル（Pitmanと表記）を利用して推定された母集団寸法指標の推定値（$\hat{S}_l$と表記）のサイズ30以下の部分を示す。

ポアソンガンマモデルはBethlehem et al.（1990）が提案したモデルで、母集団寸法指標の期待値は

表2：標本寸法指標と2つの超母集団モデルによる母集団寸法指標の推定値

l	s_l	$\hat{S}_l$	
		P–g	Pitman
1	73000	99039	38758529
2	3672	49475	2175838
3	1262	32953	806639
4	559	24692	425921
5	329	19736	265091
6	196	16432	181670
7	134	14071	132665
8	83	12301	101351
9	56	10925	80085
10	53	9823	64961
11	27	8922	53807
12	27	8171	45338
13	20	7536	38751
14	15	6991	33523
15	8	6519	29301
16	6	6106	25841
17	7	5742	22970
18	5	5418	20559
19	2	5128	18514
20	3	4867	16765
21	1	4631	15252
22	4	4417	13946
23	2	4221	12800
24	2	4041	11792
25	1	3876	10901
26	1	3724	10108
27	1	3582	9401
28	1	3451	8766
29	3	3329	8194
30	0	3215	7678

$$E(S_l) = \frac{K \cdot \prod_{j=1}^{l}\left(\frac{N_0}{K} + (j-1)N_0\beta\right)}{l!(1+N_0\beta)^{l+1/(K\beta)}} \tag{1}$$

と書くことができる。なおβはパラメータである。ポアソンガンマモデルでは母集団の大きさNは期待値としてN_0となるが、ここでは$N_0 = 1.09004 \times 10^8$と考える。

また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

$$E(S_l) = \frac{N}{l}\binom{N-1}{l-1}\frac{(1-\alpha)^{[l-1]}(\theta+\alpha)^{[N-l]}}{(\theta+1)^{[N-1]}} \tag{2}$$

と書くことができる。なおα、θはパラメータである。また、$x^{[k]} = x(x+1)\cdots(x+k-1)$である。2つのモデルともパラメータの推定にはモーメント法を用いた。

2つのモデルによって推定される母集団寸法指標は、母集団が小さい場合にはその差は顕著ではないが、母集団が大きくなるにつれてその差は極めて大きくなる。この結果では、特にリスク評価にしばしば用いられる母集団一意（母集団でサイズ1）の個体数については400倍程度の違いがある。これは竹村（2000）がEwensモデルとピットマンモデルについて指摘したことと同じである。

ポアソンガンマモデルでは、連続する2つのサイズの寸法指標の期待値の比はKが十分大きいとき、(1)式より

$$\lim_{K\to\infty}\frac{E(S_{l+1})}{E(S_l)} = \lim_{K\to\infty}\frac{\frac{N_0}{K} + lN_0\beta}{(l+1)(1+N_0\beta)}$$

$$= \frac{l}{l+1}\cdot\frac{N_0\beta}{1+N_0\beta} \tag{3}$$

と書くことができる。一般に総セル数Kは十分大きいことが多いため、

上式は近似的に成り立つことが多い。ここで更に $N_0\beta \to \infty$ とすると

$$\lim_{N_0\beta\to\infty} \frac{l}{l+1} \cdot \frac{N_0\beta}{1+N_0\beta} = \frac{l}{l+1} \tag{4}$$

となる。すなわち母集団が大きい場合には、すべてのサイズの個体数がほぼ等しいという結果になってしまう。

一方、ピットマンモデルはポアソンガンマモデルと同様の性質を持つモデルも含んだ広いモデルであるため、実データに対して当てはまりが良く、推定された母集団寸法指標の信頼性も高いが、総個体数の35.6%が母集団一意というのは直感的に多すぎると思われる。

3　部分母集団の情報を利用した母集団寸法指標の推定

労働力調査で得られる項目のうち、都道府県、男女、年齢5歳階級（15歳以上については14分類）のクロスの人口が、総務省から10月現在の推計人口として毎年公表されている。これは国勢調査の他、人口動態統計や住民基本台帳人口移動報告などを元に算出しているものである。この節では、各推計人口を部分母集団の大きさと考え、次のような母集団寸法指標の推定法を提案する。

まず母集団全体をいくつかのキー変数を利用して部分母集団に分割し、各部分母集団について、対応する標本から標本寸法指標を求める。次に各標本寸法指標から独立に部分母集団の寸法指標を推定し、その寸法指標を合算することにより全体の母集団寸法指標を推定する。

超母集団モデルではピットマンモデルであっても推定すべきパラメータが2個しかなく、母集団寸法指標の推定に用いる情報としては極めて少ない。部分母集団に分割することによりパラメータ数も増加し、安定した推定が期待できる。また個々の部分母集団の大きさがある程度小さくなるため、ポアソンガンマモデルなどでも妥当な推定結果が得られる可能性がある。

都道府県別の15歳以上人口を利用して、47分割された個々の部分母集団について、対応する標本から推定される母集団寸法指標の合算値を表3の「都道府県別」の列に示す。「全国」の列には表2の推定結果を再掲している。2つのモデルによる推定結果は都道府県の情報を用いない場合よりも接近している。また「都道府県・男女別」の列には、男女別の情報を加え94分割された個々の部分母集団について推定される寸法指標の合算値を示すが、「都道府県別」の結果からの変化はごく僅かである。

4　制約付きノンパラメトリック推定の概略

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を提案したが、ここではその概略を述べる。

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非負の母集団寸法指標 $(S_1, S_2, \ldots, S_L)$ を求める。ただし

$$\sum_{l=1}^{L} l \cdot S_l = N \tag{5}$$

である。

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。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ず、本論文では計算時間などを勘案して主に $L = 120$ としているが、この点については後で検討を行う。

$k_{ll'}(l = 0, 1, 2, \ldots, L;\ l' = 0, 1, 2, \ldots, l;\ l' \leq l)$ を母集団においてサイズ l のセルのうちで、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

表 3 ：部分母集団の情報を用いた母集団寸法指標の推定値

l	$\hat{S}_l$ (全国)		$\hat{S}_l$ (都道府県別)		$\hat{S}_l$ (都道府県·男女別)	
	P–g	Pitman	P–g	Pitman	P–g	Pitman
1	99039	38758529	176896	33017380	195519	32910535
2	49475	2175838	88193	1929374	97380	1759829
3	32953	806639	58626	724380	64669	665642
4	24692	425921	43843	385239	48314	355888
5	19736	265091	34974	240979	38502	223548
6	16432	181670	29061	165787	31962	154325
7	14071	132665	24838	121448	27292	113387
8	12301	101351	21671	93027	23790	87078
9	10925	80085	19208	73676	21066	69124
10	9823	64961	17238	59882	18887	56301
11	8922	53807	15626	49689	17105	46807
12	8171	45338	14283	41935	15621	39574
13	7536	38751	13147	35894	14365	33928
14	6991	33523	12173	31093	13288	29435
15	6519	29301	11329	27210	12357	25796
16	6106	25841	10591	24025	11541	22807
17	5742	22970	9939	21377	10822	20320
18	5418	20559	9360	19152	10183	18227
19	5128	18514	8843	17263	9611	16448
20	4867	16765	8377	15646	9097	14923
21	4631	15252	7955	14249	8632	13606
22	4417	13946	7572	13035	8210	12459
23	4221	12800	7222	11973	7824	11454
24	4041	11792	6902	11037	7471	10569
25	3876	10901	6607	10210	7146	9785
26	3724	10108	6335	9473	6846	9087
27	3582	9401	6083	8815	6569	8463
28	3451	8766	5850	8225	6312	7901
29	3329	8194	5632	7692	6072	7396
30	3215	7678	5429	7211	5849	6939

$$L(s_1, s_2, \ldots, s_L \mid S_1, S_2, \ldots, S_L)$$
$$= \frac{l}{{}_N C_n} \sum_{c_1} \prod_{l=1}^{L} \frac{S_l!}{k_{l0}! k_{l1}! \cdots k_{ll}!} ({}_l C_0)^{k_{l0}} ({}_l C_1)^{k_{l1}} \cdots ({}_l C_l)^{k_{ll}} \qquad (6)$$

表4：条件 c_1（母集団から標本への寸法指標の移動）

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

と書ける。ただし、条件 c_1 は非負整数 $k_{ll'}$ の表4に示すような行和と列和に関する制約条件である。すなわち標本寸法指標 $(s_1, s_2, \ldots, s_L)$ が得られ、1つの母集団寸法指標 $(S_1, S_2, \ldots, S_L)$ を候補として選んだ場合、列和と行和を満たすすべての $k_{ll'}$ の組について和を求める必要がある。得られている標本寸法指標を生成する可能性のある母集団寸法指標のうちで、この尤度関数を最大にするものが最尤推定値となる。

しかし、条件を満たす非負整数 $k_{ll'}$ の組み合わせは一般に膨大であり、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。そこで佐井（2002, 2003）では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した。その導出過程は省略するが、抽出率が十分小さいときには、尤度関数(6)式は

$$L(s_1, s_2, \ldots, s_L \mid S_1, S_2, \ldots, S_L) = c \cdot \prod_{l'=1}^{L} \frac{e^{-\mu_{l'}} \mu_{l'}^{s_{l'}}}{s_{l'}!} \tag{7}$$

と近似でき、尤度関数の計算時間は大幅に短縮される。ただし c は定数であり、

$$\mu_{l'} = \sum_{l=l'}^{L} S_l \cdot {}_l C_{l'} \lambda^{l'} (1-\lambda)^{l-l'} \tag{8}$$

である。

このようなノンパラメトリック法で推定され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次のような制約を置く場合について、実データを元に比較を行った。

(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

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

これまで経験上知られている寸法指標のほとんどは、最も強い制約である(e)を満たしており、実データに対する推定結果でも(e)から最も安定した推定値が得られた。

本論文の例でも制約条件として主に(e)を用いるが、その妥当性については後に検討を行う。なお、対数が下に凸という条件は頻度が10以上のサイズにのみ適用することとして、頻度の小さいサイズについては条件(b)のみを適用することとする。

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$(S_1, S_2, \ldots, S_L)$について網羅的に尤度を計算することは計算時間上困難であるため、佐井（2003）では次のような数種類の探索方法についての比較を行った。(C)と(F)については説明を省略する。

(A) (6)式を用いて網羅的に尤度を求める方法

(B) (7)式を用いて網羅的に尤度を求める方法

(C) 個体追加法

(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

(D)、(E)の方法では、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のセル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。以後これを繰り返し、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。

後の例では、まず(E)を用いて最大尤度を探し、その結果を初期値として更に(D)によって最大尤度を探している。

5　超母集団モデルとノンパラメトリック法との結果の比較

前節で述べた制約付きノンパラメトリック推定法では、ピットマンモデルで得られた初期値にもよるが、母集団の大きさが10^5程度でもパーソナルコンピュータでの計算には数日程度を要し、10^6を超えると計算が困難となる。

この節では、94分割された部分母集団の1つである秋田県の男性を母集団と見なし、超母集団モデルとともに前節で述べた制約付きノンパラメトリック推定の結果も加えて比較する。その際、新たに　15歳～39歳、40歳～59歳、60歳～という年齢3階級の部分母集団に分割して、個々の年齢階級についても母集団寸法指標を推定する。制約付きノンパラメトリック推定では、制約条件として“(e)　母集団寸法指標が単調減少でその対数が下に凸”を用い、“(E)　1サイズ探索法”の結果を初期値として“(D)　2サイズ探索法”を用いて最大尤度を探索する。このような手法の組み合わせ

を(E)(D)−(e)と表記することにする。

表5に、「秋田県、男性、15歳～39歳」($n = 119$、$N = 1.6 \cdot 10^5$) の標本寸法指標と、2つの超母集団モデルと制約付きノンパラメトリック推定による母集団寸法指標の推定結果を示す。ポアソンガンマモデルとピットマンモデルの推定値には依然大きな隔たりがある。しかし、ノンパラメトリック法の推定値はそのどちらからもかけ離れ、特に母集団一意の推定値は2つのモデルの推定値の間に位置している。

図1に、「秋田県、男性」の年齢3階級について、各母集団寸法指標の推定値を図示している。横軸はサイズ l、縦軸は $\log \hat{S}_l$ を表す。「40歳～59歳」のみ、ノンパラメトリック法 (Nonpara と表示) で推定された母集団寸法指標はピットマンモデルの結果に近い。労働力調査では就業者についてのみ調査される項目があり、就業者についてはキー変数の組み合わせによる分類が細かくなり、一意である可能性が高くなる。「40歳～59歳」には就業者が多いこともあり $(s_1, s_2, s_3, s_4, s_5, s_6) = (154, 1, 0, 1, 0, 1)$ という標本寸法指標であるが、そのような点が影響しているかもしれない。

キー変数による分類が細かく、標本で s_1, s_2 などの比率が高い場合、過度に小さな部分母集団に分けすぎると、ほとんどの個体が標本一意になり母集団寸法指標の推定が不安定になる可能性がある点にも注意しなければならない。

「秋田県、男性」の標本寸法指標から推定された母集団寸法指標を表6の「全年齢」の3列に、年齢階級別に標本寸法指標から推定された母集団寸法指標を合算したものを「年齢階級別」の3列にそれぞれ示す。

母集団一意の値に注目すると、「15歳～39歳」、「60歳～」の2つの年齢階級ではノンパラメトリック法の推定値はピットマンモデルの推定値よりかなり小さいものの、「40歳～59歳」の推定結果が近いため、「年齢階級別」で合算したノンパラメトリック法の推定値はピットマンモデルのものに近づいている。いくつかの都道府県の男女について同様の検討を行ったところ、「40歳～59歳」のように標本一意の比率が極めて高い集団が存在

表5：母集団寸法指標の推定値（秋田県、男性、15歳～39歳）

l	s_l	$\hat{S}_l$		
		P–g	Pitman	(E)(D)−(e)
1	106	588	87976	9262
2	3	293	2643	6342
3	1	195	934	4344
4	1	145	481	2976
5	0	116	294	2040
6	0	96	199	1399
7	0	82	144	960
8	0	72	109	660
9	0	63	86	455
10	0	57	69	315
11	0	52	57	219
12	0	47	48	153
13	0	43	40	108
14	0	40	35	77
15	0	37	30	56
16	0	35	27	42
17	0	33	24	33
18	0	31	21	27
19	0	29	19	23
20	0	27	17	21
21	0	26	16	20
22	0	25	14	18
23	0	24	13	17
24	0	23	12	16
25	0	22	11	15
26	0	21	10	14
27	0	20	9	13
28	0	19	9	13
29	0	18	8	12
30	0	18	8	12

しない場合には、部分母集団に分けても両者の結果にはなお開きがあった。

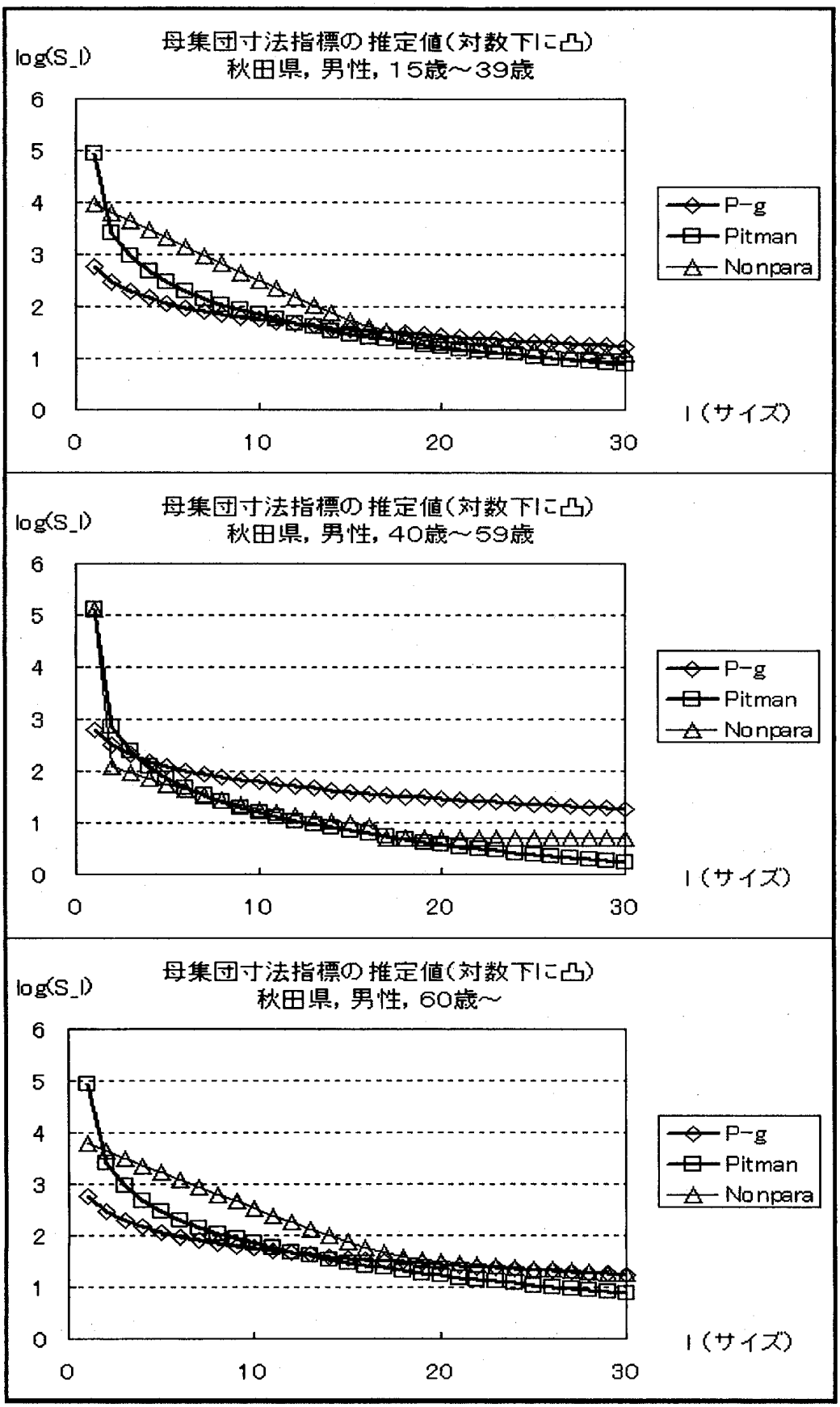

図１：年齢階級別の母集団寸法指標の推定値（秋田県、男性）

表6：母集団寸法指標の推定値（秋田県、男性）

l	$\hat{S}_l$（全年齢）			$\hat{S}_l$（年齢階級別）		
	P–g	Pitman	(E)(D)−(e)	P–g	Pitman	(E)(D)−(e)
1	1489	266895	22930	1606	271231	144660
2	742	8100	17250	800	6532	10925
3	493	2864	12977	532	2368	7656
4	369	1475	9763	397	1237	5371
5	294	903	7347	317	764	3773
6	244	611	5530	263	521	2655
7	209	442	4163	225	379	1872
8	182	335	3135	196	288	1324
9	161	263	2362	174	227	940
10	145	212	1781	156	184	671
11	131	174	1344	141	152	482
12	120	146	1015	129	128	349
13	110	124	768	119	109	255
14	102	107	582	110	94	189
15	95	93	442	102	82	143
16	89	82	337	95	72	111
17	83	73	258	89	64	86
18	78	65	199	84	57	72
19	74	58	155	79	52	63
20	70	52	122	75	47	58
21	67	48	97	71	42	55
22	63	43	78	68	39	51
23	60	40	64	65	36	48
24	58	37	54	62	33	45
25	55	34	47	59	30	43
26	53	31	42	57	28	41
27	51	29	39	54	26	39
28	49	27	37	52	24	38
29	47	25	35	50	23	36
30	45	24	33	48	21	35

6　母集団寸法指標に対する制約条件の妥当性

これまで個票開示問題で扱ってきた寸法指標は、そのほとんどが対数をとってもなお下に凸という(e)の条件を満たすものであった。しかしこれらの多くは標本の寸法指標であり、$N = 10^5$ を超えるような母集団寸法指標

についての知識の十分な蓄積はない。

図1においてノンパラメトリック法で推定された母集団寸法指標を観察すると、例えば「15歳～39歳」のサイズ15以下の部分のように寸法指標が直線的な形状である部分があり、対数下に凸という条件が厳しすぎることも懸念される。ここでは、「秋田県、男性、15歳～39歳」の3つの年齢階級について、条件(e)を用いた結果を、4節で紹介した“(d)　母集団寸法指標が単調減少で下に凸”、“(b)　母集団寸法指標が単調減少”という条件の下でそれぞれノンパラメトリック推定された結果と比較する。

表7：制約条件別の推定結果の比較（秋田県、男性、15歳～39歳）

l	$\hat{S}_l$		
	(E)(D)−(e)	(E)(D)−(d)	(E)(D)−(b)
1	9262	8332	5341
2	6342	6338	5341
3	4344	4344	5341
4	2976	2981	2981
5	2040	2040	2039
6	1399	1399	1400
7	960	960	960
8	660	660	660
9	455	455	455
10	315	315	315
11	219	219	219
12	153	153	153
13	108	108	108
14	77	77	77
15	56	61	61
16	42	52	52
17	33	44	44
18	27	37	37
19	23	31	31
20	21	26	26

表7に推定結果を示す。

制約条件(b)では$\hat{S}_1 = \hat{S}_2 = \hat{S}_3$と推定され、制約条件(d)でも寸法指標の対数が多少上に凸となる部分があることが分かる。もちろん母集団寸法指標がこのような形状をしている保証はないのに加えて、後で述べるサイズ打ち切りの影響も考えられるが、特に重要な推定目標であるサイズの小さい部分の寸法指標の推定が制約条件によって左右されており、この点については今後十分な検討が必要である。

7　母集団寸法指標の整数頻度についての検討

尤度関数を最大にする母集団寸法指標を探索する際には、これまで各サ

イズの頻度を整数として扱ってきた。例えば“(E)　1サイズ探索法”による最大尤度探索の過程において、サイズ120の頻度を1だけ増やすためには、サイズ1の頻度を120だけ減らすことが必要であった。各サイズの頻度を整数に限定することが、特に大きなサイズの頻度の増減を困難にしている可能性がある。

この節では、増減の単位を1頻度から1個体に変更することによって、推定される母集団寸法指標の変化をみることにする。(D)、(E)に対応した1個体増減の方法を(D')、(E')と表記する。このような変更によって、例えばサイズ120の頻度を1/120の単位で増減させることが可能となり、より柔軟な推定が行われることも考えられる。

表8に、“(e)　母集団寸法指標が単調減少でその対数が下に凸”、“(d)　母集団寸法指標が単調減少で下に凸”の2つの制約条件の下での、1頻度増減法と1個体増減法を用いた母集団寸法指標の推定結果を示す。

制約条件(d)の結果では、1個体増減法の大きいサイズの部分の頻度が若干増加しているため$\hat{S}_1$などが減少しているが、1頻度増減法と1個体増減法との推定結果の差はさほど大きくない。

8　母集団寸法指標の最大サイズについての検討

これまで$N = 10^5$程度までの比較的小さな母集団の寸法指標の推定では、最大のサイズLとして$\hat{S}_L = 0$と推定されるような十分大きな値（$L = 50, 80, 120, 160$など）を試行錯誤の後に設定する場合が多かったが、本論文の例では計算時間上の理由から$L = 120$としてきた。表9の(D)(E)−(e)の列に、ノンパラメトリック法によるサイズ1～10と111～120の母集団寸法指標の推定値を再掲しているが、$L = 120$はサイズの最大値として十分大きな値とは言えず、寸法指標の大きいサイズの部分を打ち切っていることが小さいサイズの寸法指標の推定値に影響を与えていることも考えられる。

ここで、この問題を克服するために2つの代替法を考える。

表8：1頻度増減法と1個体増減法との推定結果の比較（秋田県、男性、15歳～39歳）

l	$\hat{S}_l$			
	(E)(D)−(e)	(D')(E')−(e)	(E)(D)−(d)	(D')(E')−(d)
1	9262	9254.00	8332	7070.00
2	6342	6340.00	6338	5707.00
3	4344	4343.67	4344	4344.00
4	2976	2976.00	2981	2981.00
5	2040	2039.00	2040	1838.00
6	1399	1399.00	1399	1399.00
7	960	960.00	960	960.00
8	660	660.00	660	660.00
9	455	455.00	455	455.00
10	315	315.00	315	315.00
111	9	9.00	9	9.35
112	9	9.00	9	9.35
113	9	9.00	9	9.35
114	9	9.00	9	9.34
115	9	9.00	9	9.34
116	9	9.00	9	9.34
117	9	9.00	9	9.33
118	9	9.00	9	9.33
119	9	9.00	9	9.33
120	9	9.00	9	9.33

1つ目の方法は“(e)　母集団寸法指標が単調減少でその対数が下に凸”という条件をサイズ119以下の寸法指標にのみ適用し、S_{120} については全く制約を置かない方法である。この制約条件を(e')と表記する。これは、打ち切っている120より大きなサイズの母集団寸法指標の値を S_{120} に受け持たせることにより、サイズ119以下の推定結果へのサイズ打ち切りによる影響を軽減するのが目的である。

2つ目の代替法は、120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で、ここでは $l = 200, 300, 400, 500, 1000, 2000, 3000, 4000, 5000$ の9つのサイズの頻度についてのみ S_l が正になることを認めている。ここで、例えば S_{200} はサイズ121～200を、S_{300} はサイズ

201～300を受け持つと考え、$S_{200}/(200-120) \geq S_{300}/(300-200)$ を満たす母集団寸法指標のみを探索することにより、120を超えるサイズの寸法指標にも単調減少の条件を入れるものとする。なお S_{5000} は5000を超えるサイズも受け持たせるため、上記の代替法の S_{120} と同様に制約を置かない。また、120を超えるサイズの頻度は十分小さくなるため、この部分については母集団寸法指標の対数が下に凸という条件を考慮しないこととする。制約条件以外の手法も含むこのような方法全体を(e")と表す。

この方法では、$\mu_{121} = \mu_{122} = \cdots = \mu_{200}$ となることなどを利用することにより、(8)式、(7)式の計算において簡便化が図られる。また最大尤度の探索時には $l = 1, 2, \ldots, 120$ に加えて9つのサイズのみの変化を考えるだけでよいため、最大サイズが120の場合と比較しても計算時間はさほど増加しない。

表9の(E)(D)－(e')、(E)(D)－(e")の列に、上記の方法で推定された母集団寸法指標を示す。(E)(D)－(e")では、サイズ3000程度まで母集団寸法指標は伸びているが、サイズの小さい部分の寸法指標は(E)(D)－(e)とさほど違わない。これに対して(E)(D)－(e')では $\hat{S}_1$ の値が極端に減少してしまっており、最大サイズの調整は慎重に扱わなければならない問題であることが分かる。今後の検討が必要である。

9　おわりに

本論文では母集団を部分母集団に分割することによって、超母集団モデルによる推定を安定させるとともに、複数の超母集団モデルによる推定値の接近を図った。実際、ピットマンモデルとポアソンガンマモデルで推定される母集団寸法指標は部分母集団の情報を用いることにより接近したものの、その差は依然として大きく、更なる検討が必要である。

また母集団を分割することにより制約付きノンパラメトリック推定が可能となったが、推定された母集団寸法指標は上記2つの超母集団モデルによって得られたいずれのものとも異なる形状であった。この点に関する詳

しい検討のためにはノンパラメトリック推定の高速化が不可欠であり、計算方法の根本的な見直しも必要となろう。本論文の後半で行ったノンパラメトリック推定の種々の検討結果は、その際にも極めて重要と思われる。

謝辞

本論文をまとめるにあたり、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、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（16－共研－2024）の研究成果に基づくものである。

表9：サイズ打ち切りを回避する代替法による推定値（秋田県、男性、15歳～39歳）

l	$\hat{S}_l$		
	(E)(D)−(e)	(E)(D)−(e′)	(E)(D)−(e″)
1	9262	1225	10233
2	6342	904	6945
3	4344	670	4714
4	2976	497	3201
5	2040	369	2190
6	1399	274	1506
7	960	181	1037
8	660	120	715
9	455	82	494
10	315	68	342
111	9	1	1
112	9	1	1
113	9	1	1
114	9	1	1
115	9	1	1
116	9	1	1
117	9	1	1
118	9	1	1
119	9	1	1
120	9	923	1
200	—	—	1
300	—	—	1
400	—	—	1
500	—	—	1
1000	—	—	3
2000	—	—	5
3000	—	—	2
4000	—	—	0
5000	—	—	0

参考文献

[1] Bethlehem, J. G., Keller, W. J. and Pannekoek, J. (1990). Disclosure control of microdata, *Journal of the American Statistical Asociation*, **85**, 38–45.

[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、瀧敦弘、加納悟、福重元嗣 (2004a). 労働力調査、家計調査、全国消費実態調査の個票データに対するリスク評価、科学研究費補助金（基盤研究(B)(1)）研究成果報告書（課題番号13553001、研究代表者　佐井至道)、85－114.
[6] 佐井至道、瀧敦弘、加納悟、福重元嗣 (2004b). 労働力調査、家計調査、全国消費実態調査の標本寸法指標 (CD-ROM).
[7] 佐井至道 (2004). 部分母集団の情報を用いた母集団寸法指標の推定、2004年統計関連学会連合大会予稿集、241－242.
[8] 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.
[9] 竹村彰通 (2000). 母集団一意数推定に関するいくつかの論点、科学研究費補助金研究会資料、1－10.