『岡山商大論叢』（岡山商科大学）
第48巻第１号　2012年７月
Journal of OKAYAMA SHOKA UNIVERSITY
Vol.48　No.１　July 2012

《論　　説》

事後層化による
個票データのリスク評価の改善

佐　井　至　道

Improvement of the Risk Assessment for Microdata
Using Poststratification

Shido Sai

１　はじめに

個票データを公開する際には，そのデータに含まれる個人の情報が漏洩しないように注意しなければならない。そのため，調査された情報がそのまま並べられた個票データに対して，段階的に秘匿措置を施しながらリスク評価を行い，十分な有用性を保ちながら，個人情報の漏洩が起こらない個票データを探索することになる。

リスク評価には寸法指標が用いられることが多い。寸法指標とは，個体の特定に用いられるキー変数の値の同じ組み合わせに対する個体数をそれぞれ求めた場合の，個体数別の頻度分布である。サイズ１，またはユニークと呼ばれる個体数が１である頻度が，特に重要視される。個票データが標本から得られた場合には，標本寸法指標とともに，それを基にして推定

された母集団寸法指標も用いることが多い。母集団寸法指標の推定にはピットマンモデルなどの超母集団モデルを用いるのが主流であるが，モデルの代わりに母集団寸法指標の推定値に制約条件を置くノンパラメトリック法も考案されている。

これまで母集団寸法指標の推定では，実際のサンプリング法に関わらず非復元単純無作為抽出，またはベールヌーイ抽出を仮定することが多かった。しかし，官庁統計では層化抽出や多段抽出などを組み合わせてサンプリングを行うことが多いため，佐井（2011）では，層化無作為標本から得られる個票データに対するリスク評価方法を，層の構成に用いられる層別変数がキー変数の場合とキー変数でない場合に分けて提案した。また，従来行われてきたように，そのような個票データを単純無作為標本から得られたと見なしてしまうことが，リスク評価の結果にどのような影響を与えるかという点についても検討を行った。

本論文では，非復元単純無作為抽出された標本から得られた個票データに対して，事後層化を行うことによってリスク評価の精度の改善を試みる。佐井（2011）での提案は，層化無作為標本から得られた個票データについて，単純無作為抽出を仮定してリスク評価を行ってしまうと推定値に偏りが生じるため，それを防ぐのが目的であって，リスク評価の精度の改善のために層化無作為抽出法が用いられているわけではなかった。実際，個票データのリスク評価の精度を高めるために標本設計を変更することは，官庁統計の調査でも，それ以外の標本調査でも考えにくい。一方，非復元単純無作為抽出された標本のリスク評価の精度を改善するために事後層化することは，リスク評価を行う当事者にとっても常に可能である。

まず２節では，事後層化に用いる層別変数がキー変数の場合のリスク評価について考える。また層化無作為抽出との違いについても説明する。

3.1節では，労働力調査の個票データに対して実際に事後層化を行うことにより，どの程度リスク評価の精度を改善できるか検討するとともに，3.2節では事後層化による各層の標本の大きさの変動がリスク評価に与え

る影響について検討する。さらに3.3節では，アメリカのセンサスの１％抽出個票データを用いて，どのような事後層化を行うのが有効であるかを検討する。

４節では，事後層化に用いる層別変数がキー変数でない場合のリスク評価について考える。このとき，佐井（2011）と同様に，標本寸法指標からの母集団寸法指標の推定の過程に多重寸法指標を導入する。多重寸法指標は寸法指標におけるサイズを多次元に拡張したもので，佐井（2009a, 2009b）においては，継続調査で得られた複数の時点の個票データに対する同時リスク評価を行うために用いられている。ただし，多重母集団寸法指標の推定は，母集団寸法指標の推定と比較して不安定であるため，その有効性については十分な検討が必要である。

なお，母集団寸法指標の推定の手順は，層別変数がキー変数に含まれるか否かに関わらず，単純無作為抽出された標本を事後層化する場合も，層化無作為抽出された標本に対する場合とほぼ同じものになる。ただ，事後層化では各層の大きさの期待値は比例配分になるものの一定ではない。例えば母平均の推定では，事後層化による推定の精度は，比例配分法を用いた層化無作為標本からの推定の精度よりも一般に劣る。（Cochran（1977），Levy and Lemeshow（2009），Thompson（2012）を参照していただきたい。）リスク評価においても，この点についての検討を行う必要があるが，各層の標本の大きさと母集団寸法指標の推定精度との関係の解析的な解明は容易ではなく，本論文では3.2節における簡単な例による実験にとどめることにする。

２　層別変数がキー変数の場合の事後層化

層化無作為抽出では，母集団をあらかじめいくつかの層（グループ）に分け，各層から決められた大きさの標本を他の層とは独立に非復元無作為抽出する。これに対して事後層化は，母集団から決められた大きさの標本を非復元単純無作為抽出し，その標本を事後的にいくつかの層に分けるも

のである。例えば母平均の推定では，層化無作為標本か，事後層化された単純無作為標本かに関わらず，各層の母集団の重みを利用して，各層の標本平均の重み付き平均によって推定を行うのが一般的である。

事後層化を行う際には通常の層別と同様に何らかの変数の値を用いることが多く，ここではそのような変数を層別変数と呼ぶ。層別変数は量的変数と質的変数の場合があり，量的変数の場合には値によって，質的変数の場合にはカテゴリーによって層別を行うが，ここでは値とカテゴリーを区別せず，主に値と表記する。また，層別変数が複数の場合には値の組み合わせによって層を構成することになるが，本論文では層別変数が 1 個の場合の表現を用いる。

まず標本寸法指標と母集団寸法指標について説明する。

大きさ N の母集団を G とする。母集団 G から大きさ n の標本 g が非復元単純無作為抽出されていて，その標本から個票データが得られているとする。個票データの T 個のキー変数を $z_1, z_2, \ldots, z_T$ とおく。このとき，すべてのキー変数の値の組み合わせ $z_1 \times z_2 \times \cdots \times z_T$ によって多次元分割表を構成する。

キー変数とは，個票データの変数のうち，第三者が既に情報を持っていて，個票データに含まれる個体を特定するのに用いることのできる変数であり，住所や性別などがそれにあたる。

値の組み合わせ一つ一つを分割表の呼び方に倣ってセルと呼ぶ。総セル数を K とするが，その中で，セルに入る個体数が l であるセル数，すなわちサイズ l のセル数を，標本では s_l $(l = 1, 2, \ldots, L)$ と表して標本寸法指標と呼び，母集団では S_l $(l = 1, 2, \ldots, L)$ と表して母集団寸法指標と呼ぶ。L はサイズの最大値であるが，モデルによっては無限大とする場合もある。母集団寸法指標を推定するには，標本寸法指標を基にして母集団寸法指標を直接推定するのが一般的である。

次に, 事後層化を用いる母集団寸法指標の推定方法について説明するが, この節では層別変数がキー変数の場合について考える。

個票データの T 個のキー変数 $z_1, z_2, \ldots, z_T$ のうち z_1 を層別変数として，標本 g を変数 z_1 の値によって H 個の層に事後層化する。標本の第 h 層 g_h の大きさを n_h とする。また標本の事後層化に対応した母集団の第 h 層を G_h と表し，大きさを N_h として，第 h 層の抽出率を $\lambda_h = n_h/N_h$ とする。なお，事後層化された標本の各層の大きさ $(n_1, n_2, \ldots, n_H)$ は次のように多変量超幾何分布に従うが，この点については3.2節で検討を行う。

$$\mathrm{p}(n_1, n_2, \ldots, n_H) = \frac{\prod_{h=1}^{H} {}_{N_h}C_{n_h}}{{}_{N}C_{n}}. \tag{1}$$

個票データの中で，第 h 層の標本 g_h に対応するレコードについて，すべてのキー変数の値の組み合わせ $z_1 \times z_2 \times \cdots \times z_T$ によって構成されるセルから求められる標本寸法指標を $s_{h,1}, s_{h,2}, \ldots, s_{h,L}$ と表し，第 h 層の標本寸法指標と呼ぶ。また第 h 層の母集団寸法指標も同様に定義して，$S_{h,1}, S_{h,2}, \ldots, S_{h,L}$ と表す。

母集団寸法指標の推定の手順としては，各層ごとの標本寸法指標 $s_{h,1}, s_{h,2}, \ldots, s_{h,L}\,(h = 1, 2, \ldots, H)$ から，各層ごとの母集団寸法指標の推定値 $\hat{S}_{h,1}, \hat{S}_{h,2}, \ldots, \hat{S}_{h,L}\,(h = 1, 2, \ldots, H)$ を独立に求め，それぞれのサイズについて $\hat{S}_l = \sum_{h=1}^{H} \hat{S}_{h,l}\,(l = 1, 2, \ldots, L)$ と合算することによって母集団寸法指標の推定値 $\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_L$ を求めることになる。

3　実データを用いた事後層化の有効性の検討

3.1　労働力調査の個票データに対する事後層化の適用

この節では，労働力調査の個票データに対して，事後層化を用いるリスク評価と事後層化を用いないリスク評価を行い，事後層化の有用性を検証する。

検討に用いるのは2007年12月に実施された労働力調査の基礎票のデータで，これは目的外使用申請が認められたものである。自衛官と受刑者を除

く83709レコードからなる個票データを対象とする。

元の個票データには表１にあげる34変数が含まれるが，比較的緩やかな秘匿措置を施す。表中で変数名の右の列に“−”の記号がある変数はキー変数とは考えず，“×”の記号がある変数は個票データから削除することにする。“○”の記号がある変数はキー変数と考え，元のままの値の公開を想定する。変数名の右の列に数値が書かれた変数は，キー変数であるもののカテゴリー数をその数値まで減らす。例えば変数「15歳以上総数」は，その個人が属する世帯の15歳以上の人数を表すが，ここでは６人以上を１つのカテゴリーにまとめて合計６カテゴリーとする。また，「15歳未満男総数」と「15歳未満女総数」は合計してから５つのカテゴリーに分けている。なお，これらの秘匿措置は，標本寸法指標が実験に耐えうる程度の長さになるように施したもので，この秘匿措置が公開に適しているということを主張をするものではない。

表２の s_l の列に，個票データの標本寸法指標のうち $l \leq 10$ の部分を示す。頻度が０でない最大サイズは175である。以後，表では寸法指標としてサイズ 10以下の部分のみを示すことにする。

また“Pg-srs”，“Pit-srs”の列には，標本寸法指標を基に，それぞれポア

表１：労働力調査の個票データの項目

変数名	
西暦年月	−
年	−
月	−
都道府県番号	○
県内一連番号	×
抽出単位番号	−
単位内世帯番号	−
世帯員番号	×
15歳以上総数	6
15歳未満男総数	5
15歳未満女総数	

変数名	
階級別人員０～３歳	×
階級別人員４～６歳	×
階級別人員７～９歳	×
階級別人員10～12歳	×
階級別人員13～14歳	×
性別	○
続柄	8
出生年月　元年	×
出生年月　年	×
出生年月　月	×
換算年齢	7
配偶関係	○

変数名	
就業状態	○
仕事の主従	−
求職理由	−
就業時間	14
従業上の地位	○
経営組織	○
産業	○
職業	○
従業員数	○
希望意識	−
異動符号	×

ソンガンマモデルと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示す。“srs”は非復元単純無作為抽出（simple random sampling without replacement）された標本をそのまま推定に用いることを意味している。なお, 労働力調査では実際には層化二段抽出によって標本が得られているが, ここでは非復元単純無作為抽出されたと考える。また，母集団の大きさは2007年12月の15歳以上人口の1.10428×10^8を用いている。

ポアソンガンマモデルはBethlehem et al.（1990）が個票データのリスク評価への適用を提案し，この分野の初期の研究でしばしば用いられたモデルである。実質的なパラメータが１個しかないため，実データへの当てはまりが良くないことも多く，母集団寸法指標の小さいサイズの部分，特にS_1を過小評価する傾向があることが知られている。ピットマンモデルはPitman（1995）が提案したモデルで, Ewens-Pitman sampling formula（EPSF）とも呼ばれる。２個のパラメータを持ち，現在考案されているモデルの中では当てはまりの良いことが知られているが，S_1を若干過大評価する傾向があることも指摘されている。この結果でも，ピットマンモデルのS_1の推定値は，ポアソンガンマモデルの推定値の500倍近くになっている。なお，２つのモデルのパラメータの推定には，以後の検討を含めて最尤法を用いている。

表２：労働力調査の個票データの標本寸法指標と母集団寸法指標の推定値

l	s_l	Pg-srs	Pit-srs
1	51424	75679	35028684
2	2057	37814	1609493
3	779	25192	585797
4	442	18881	306356
5	282	15095	189443
6	201	12570	129197
7	171	10767	93979
8	146	9415	71564
9	115	8363	56391
10	82	7521	45631

次に事後層化による効果を検証するために，この個票データを用いて次のような実験を行う。83709人分のレコードから抽出率約1/10で8371人分のレコードをサブサンプリングして，得られたサブサンプルの個票データから寸法指標を求めると（6393, 299, 130, 73, 46, 19, 9, 8, 3, 4,...）となる。ここでは，この寸法指標を基に，事後層化を用いる方法と用いない方法で標本寸法指標の推定を行い，実際の標本寸法指標との比較を行う。

変数「就業状態」は９つの値をとるが，そのうち「主に仕事」,「通学のかたわらに仕事」,「家事のかたわらに仕事」,「休業」の者を就業者と呼び,「完全失業者」,「通学」,「家事」,「その他（高齢者など)」に「不詳」を加えて非就業者と呼ぶことにすると，就業者は46441レコード，非就業者は37268レコードとなる。

34変数の中で「就業時間」,「従業上の地位」,「経営組織」,「産業」,「職業」,「従業員数」,「希望意識」は就業者のみについてしか調査されない。このうち「希望意識」を除く６変数はキー変数と見なしているため，非就業者と比較して就業者の方がキー変数の値の組み合わせがかなり細かくなる。実際，標本の個票データのレコードを就業者と非就業者に分けて，それぞれ寸法指標を求めると表３のようになる。就業者の寸法指標の頻度が０でない最大サイズは12で，それより大きいサイズのセルに含まれる個体はすべて非就業者である。

表３：労働力調査の就業者と非就業者の標本寸法指標

l	s_l	就業者	非就業者
1	51424	45266	6158
2	2057	454	1603
3	779	50	729
4	442	15	427
5	282	4	278
6	201	3	198
7	171	1	170
8	146	0	146
9	115	0	115
10	82	0	82

層別では，層内ができるだけ均質に，異なる層の間ではできるだけ異質にすることにより推定の精度を高めることができるとされている。例えば母平均の推定では，目的変数の値が近いものを同じ層にまとめるように層別を行うと推定精度が高くなる。

事後層化による母集団寸法指標の推定では，各層ごとに寸法指標の異なることが推定精度の改善につながると予想されるため，「就業状態」を層別変数として，サブサンプルを就業者と非就業者の２層に事後層化する。サブサンプルの各層の寸法指標から前節の手順で個票データの標本寸法指標を推定するが，サブサンプルにおける各層の大きさの変動の影響を除くため，ここでは比例配分となっているサブサンプルを用いる。就業者の層の寸法指標は（4628, 8, 0, 0, 0, 0, 0, 0, 0, 0,...），非就業者の層の寸法指標は（1765, 291, 130, 73, 46, 19, 9, 8, 3, 4,...）である。各層の大きさが変動することが推定に与える影響については3.2節で検討を行う。

表４の“Pg-ps”，“Pit-ps”の列には，事後層化を用いて，それぞれポアソンガンマモデルとピットマンモデルで推定された標本寸法指標を示す。“ps”は事後層化（poststratification）を意味する。また“Pg-srs”，“Pit-srs”の列には，事後層化を用いずに推定された標本寸法指標を示す。

非復元単純無作為抽出されたサブサンプルを事後層化した場合，各層の大きさは一定でないが，各層のサブサンプルの大きさの条件付きで考えると，層化無作為抽出と同じと見なすことができる。そのため，ここでは各層の標本寸法指標を独立に推定した。なお，サブサンプルの就業者の層では，寸法指標がサイズ２までしかないため，ピットマンモデルの推定はやや不安定である。

ポアソンガンマモデルの推定値は，事後層化を行うことによって大きく改善されており，サイズの小さい部分については，事後層化を行わない場合のピットマンモデルの推定値と比較しても遜色ない結果となっている。一方ピットマンモデルでは，事後層化を行わない場合でも推定値は実際の値に近かったが，事後層化を行うことによりさらに改善されている。

この節の最後に，表３に示した事後層化による各層の標本寸法指標を基に，母集団寸法指標を推定してみる。表５の“Pg-ps”，“Pit-ps”の列には，事後層化を用いて，それぞれポアソンガンマモデルとピットマンモデルで推定された母集団寸法指標を示す。なお，各層には比例配分されていると見なす。事後層化を用いない場合の推定結果も再掲している。

事後層化を用いてもポアソンガンマモデルとピットマンモデルの結果にはまだ大きな開きがある。表４の結果が抽出率約1/10でとられたサブサンプルに基づいて推定されていたのに対して，ここでの推定結果は抽出率1/1000以下でとられた標本に基づいているのが理由の一つと考えられる。

表４：労働力調査の個票データの標本寸法指標の推定値

l	s_l	Pg-srs	Pg-ps	Pit-srs	Pit-ps
1	51424	16170	47462	49029	51882
2	2057	6523	1946	2911	2234
3	779	3509	759	1074	711
4	442	2123	519	563	414
5	282	1370	386	347	278
6	201	921	300	236	201
7	171	637	239	170	153
8	146	450	195	129	120
9	115	323	161	101	97
10	82	234	135	81	80

表５：労働力調査の個票データに基づく母集団寸法指標の推定値

l	Pg-srs	Pg-ps	Pit-srs	Pit-ps
1	75679	1597523	35028684	45480070
2	37814	778175	1609493	1043789
3	25192	505373	585797	371669
4	18881	369223	306356	193039
5	15095	287733	189443	119061
6	12570	233569	129197	81134
7	10767	195018	93979	59027
8	9415	166223	71564	44980
9	8363	143927	56391	35479
10	7521	126181	45631	28744

3.2　各層における標本の大きさの変動が推定結果に与える影響

前節の検討で，労働力調査の個票データに含まれる83709レコードから8371レコードをサブサンプリングしたが，その際，就業者と非就業者の２層に比例配分されているサブサンプルを用いた。サブサンプルの大きさを m ，２層の大きさを (m_1, m_2) で表すことにすると，(m_1, m_2) は（1）式で $H = 2$ とした超幾何分布

$$\mathrm{p}(m_1, m_2) = \frac{{}_{n_1}C_{m_1} \cdot {}_{n_2}C_{m_2}}{{}_{n}C_{m}} \tag{2}$$

第６：m_1のパーセント点に対するサブサンプルの各層の大きさと寸法指標
（上：各層の大きさ，中：第１層の寸法指標，下：第２層の寸法指標）

	1％	5％	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
m_1	4544	4573	4589	4607	4621	4633	4644	4655	4666	4680	4699	4715	4744
m_2	3827	3798	3782	3764	3750	3738	3727	3716	3705	3691	3672	3656	3627

l	1％	5％	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
1	4534	4561	4575	4597	4609	4620	4628	4643	4638	4657	4681	4707	4719
2	5	6	7	5	6	5	8	6	14	10	9	4	11
3	0	0	0	0	0	1	0	0	0	1	0	0	1
4	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0

l	1％	5％	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
1	1724	1729	1760	1732	1721	1714	1765	1660	1683	1762	1698	1688	1730
2	338	311	337	371	319	336	291	345	343	332	319	317	321
3	130	134	135	119	121	127	130	137	122	100	135	125	119
4	68	64	63	62	63	71	73	60	79	68	45	63	62
5	37	32	36	35	36	28	46	35	32	50	39	43	31
6	24	32	16	19	21	15	19	21	17	19	22	15	27
7	16	18	18	13	15	15	9	14	13	12	14	13	13
8	8	9	8	4	11	7	8	12	9	7	4	8	6
9	3	5	5	4	4	8	3	3	5	8	6	5	5
10	2	4	4	6	8	4	4	3	8	2	9	4	3

に従うため，各層の大きさが比例配分から大きく離れた場合には，大きさが小さい層の推定結果への影響が懸念される。この節では，この点についての検討を行う。

サブサンプルの大きさ m を前節と同じ8371とする。サブサンプルの第１層の大きさ m_1 の１％から99％までのパーセント点と，それに対応する m_2 の値を表６の上段に示す。例えば m_1<4544，m_2>3827となる確率は１％であることを意味しており，１％と99％を比較しても，各層の大きさはそれぞれ200程度の違いしかない。表６の中段と下段には，サブサンプルの m_1 が各パーセント点となるそれぞれ１つのケースについて，第１層と第２層の寸法指標を示している。

各パーセント点について，表６の２つの層の寸法指標を基にしてポアソンガンマモデルを用いて推定された標本寸法指標を表７に示す。表中の数値は層ごとに独立に推定された各層の標本寸法指標を合算したものであり，50％が表４の“Pg-ps”に示した結果である。

結果には標本誤差が含まれるが，どのパーセント点においても推定結果には大きな違いは認められない。表４の“Pg-srs”の列に示した事後層化を行わない場合の結果と比較すると，比例配分から大きく離れた場合でも推定は十分に改善されている。リスク評価を行う対象となる官庁統計の個票データでは，一般にこの例よりはレコード数が多いため，事後層化における層の数があまり多くない限り，各層の標本の大きさの変動による推定

表７：m_1のパーセント点に対する標本寸法指標の推定値

l	s_l	１％	５％	10％	20％	30％	40％	50％	60％	70％	80％	90％	95％	99％
1	51424	47885	47689	47645	48026	47713	47563	47462	47686	46288	46830	47211	48138	46668
2	2057	1652	1735	1884	1688	1724	1826	1946	1697	2394	2302	1970	1525	2335
3	779	724	724	763	752	724	736	759	713	766	796	737	720	788
4	442	502	500	523	520	500	506	519	493	505	531	503	501	524
5	282	375	374	389	387	374	378	386	369	376	394	375	374	389
6	201	292	291	302	300	291	294	300	288	293	305	292	291	302
7	171	234	233	241	240	233	235	239	231	234	243	234	233	241
8	146	191	190	196	195	190	192	195	189	191	197	191	191	196
9	115	158	158	162	162	158	159	161	157	159	163	158	158	162
10	82	133	133	136	135	133	134	135	132	133	136	133	133	136

への影響については，大きな問題は生じないと考えられる。

なお，サブサンプルの第１層における寸法指標の頻度が０でない最大サイズが非常に小さいことにより，ピットマンモデルモデルを用いた場合には，最尤推定が非常に不安定となるため，ここではポアソンガンマモデルについてのみ検討を行った。

3.3　アメリカのセンサスデータを用いた事後層化の検討

この節では，アメリカにおいて2000年に実施されたセンサスの１％抽出個票データ（U.S. Census Bureau（2003））を用いて，どのような事後層化を行うことがリスク評価の改善につながるかを検討する。このデータはPUMS（Public Use Microdata Sample）と呼ばれる。

ワシントン州に在住している20歳以上の個人レコードを用いるが，個票データに含まれる変数のうち，「続柄」，「性別」，「年齢」，「人種１」，「結婚」，「通勤手段」「通勤時間」「職業」「就業」「労働週数」「週労働時間」「年間収入」の12変数をキー変数として選び，秘匿処置を施さないままの個票データを検討に用いることにする。なお，PUMSは公開前に既に秘匿措置が施されているものの，その詳細については公表されていない。

労働力調査の個票データと同様に，変数として「就業（worked in 1999）」が含まれ，その値が「yes」である就業者については「通勤手段」，「通勤時間」，「職業」，「労働週数」，「週労働時間」，「年間収入」の値が記録されているものの，「no」である非就業者について見ると，０以外の値としては一部の値について記録されているだけである。全41959レコードの標本寸法指標とともに，就業者30234レコード，非就業者11725レコードの標本寸法指標を表８に示す。変数として１ドル単位で記録されている「年間収入」があるため，就業者の頻度が０でない最大サイズは３であり，労働力調査の個票データと比較して，就業者と非就業者の標本寸法指標の違いがさらに顕著になっている。

表８の全レコードの標本寸法指標を基に，ポアソンガンマモデルとピッ

トマンモデルで推定された母集団寸法指標を表９の“Pg-srs”，“Pit-srs”の列に示す。１％抽出個票データであることから，母集団の大きさは簡単に41959×10^2としている。また「就業」による事後層化を用いた母集団寸法指標の推定値を“Pg-ps”，“Pit-ps”の列に示す。なお，事後層化された各層では比例配分されていたと見なし，例えば就業者の母集団の大きさは30234×10^2とした。

実際の母集団寸法指標が未知のため，事後層化による改善の度合いは判断できないが，事後層化によりポアソンガンマモデルによる推定値は大きく変化し，ピットマンモデルによる推定値も，特にサイズ２以上の部分が

表８：センサスの就業者と非就業者の標本寸法指標

l	s_l	就業者	非就業者
1	39327	30099	9228
2	458	66	392
3	73	1	72
4	30	0	30
5	12	0	12
6	7	0	7
7	6	0	6
8	6	0	6
9	4	0	4
10	8	0	8

表９：センサスの個票データに基づく母集団寸法指標の推定値（就業者，非就業者による事後層化）

l	Pg-srs	Pg-ps	Pit-srs	Pit-ps
1	379137	2111011	3619237	3203620
2	172445	337441	32862	135883
3	104576	75929	11150	41736
4	71345	22093	5624	18431
5	51919	8924	3394	9677
6	39356	5103	2272	5646
7	30686	3695	1628	3545
8	24424	3001	1225	2354
9	19748	2561	955	1635
10	16167	2239	765	1179

変化している。事後層化を用いたポアソンガンマモデルとピットマンモデルの推定値は，表４ほどは近くないものの，表５よりは近い。これは1/ 100という抽出率が，２つのケースの抽出率の間に位置していることが理由であると推測される。

次に，上で用いた個票データのレコードから就業者の30234レコードをとり出し，これを標本と見なす。また，標本寸法指標が実験に耐えうる程度の長さになるように，キー変数から「年間収入」を削除した。この標本について，男女別に事後層化した場合と，年齢階級別に事後層化した場合の各層の標本寸法指標を，それぞれ表10，11に示し，それらを基に推定された母集団寸法指標の推定値を，それぞれ表12，13に示す。なお，事後層

表10：センサスの就業者の男女別の標本寸法指標

l	就業者	男性	女性
1	25815	13433	12382
2	1045	625	420
3	281	179	102
4	117	81	36
5	67	46	21
6	30	18	12
7	20	14	6
8	16	14	2
9	9	8	1
10	8	7	1

表11：センサスの就業者の年齢階級別の標本寸法指標

l	就業者	20歳代	30歳代	40歳代	50歳代	60歳以上
1	25815	5860	6520	6793	4534	2108
2	1045	180	270	328	216	51
3	281	40	72	105	55	9
4	117	13	44	38	19	3
5	67	4	17	26	20	0
6	30	1	9	12	8	0
7	20	1	6	8	5	0
8	16	0	7	6	3	0
9	9	0	3	4	1	1
10	8	0	2	4	2	0

化された各層では比例配分されていたと見なし，各層の母集団の大きさは標本の大きさに10^2を乗じたものとした。

表10からは男性よりも女性の方が標本寸法指標が短く，表11からは60歳以上，20歳代の標本寸法指標が他の年齢階級よりも短いことが分かる。これは，就業者に限った場合，男性の30歳代から50歳代ではフルタイムで働く者が多く，例えば「週労働時間」は40時間程度の者が多くなる。そのためキー変数の値の組み合わせも同じ者が多く，それによってサイズの大きいセルが多くなるものと推測される。このように，標本寸法指標の長さは層ごとに異なるものの，表12，表13を見ると，ポアソンガンマモデルでも

表12：センサスの個票データに基づく母集団寸法指標の推定値（男女による事後層化）

l	Pg-srs	Pg-ps	Pit-srs	Pit-ps
1	135843	145793	1661719	1647279
2	64871	69100	83460	84666
3	41304	43676	30542	31060
4	29586	31062	15998	16288
5	22606	23569	9895	10082
6	17992	18633	6745	6875
7	14729	15155	4902	4998
8	12308	12585	3728	3803
9	10449	10619	2934	2992
10	8982	9074	2370	2418

表13：センサスの個票データに基づく母集団寸法指標の推定値（年齢階級による事後層化）

l	Pg-srs	Pg-ps	Pit-srs	Pit-ps
1	135843	158811	1661719	1620782
2	64871	74261	83460	87225
3	41304	46344	30542	32122
4	29586	32568	15998	16873
5	22606	24437	9895	10452
6	17992	19117	6745	7131
7	14729	15396	4902	5185
8	12308	12670	3728	3944
9	10449	10601	2934	3105
10	8982	8988	2370	2508

ピットマンモデルでも，事後層化を行うことによる推定値の変化はほとんどない。これは各層の寸法指標の長さの異なることが，必ずしも事後層化による改善につながらないことを意味している。

図１と図２には，表８と表10の各層の寸法指標について対数をとったものを図示している。横軸はサイズの20以下の部分を，縦軸はそれぞれのサイズの頻度の対数を表す。図１では，就業者の標本寸法指標の対数がほぼ直線的な形状となることが読み取れる。特にポアソンガンマモデルはこのような形状の寸法指標への当てはまりの良いことが分かっており，事後層化を行うことにより，就業者の母集団寸法指標の推定が改善されたことが，結果の変化につながっていると考えられる。これに対して，図２では男性と女性の寸法指標の長さや減少の度合いは異なるものの，その形状自体には大きな違いはない。ポアソンガンマモデルでもピットマンモデルでも，事後層化によって標本寸法指標の当てはまりに大きな改善がないため，母集団寸法指標の結果にも違いが現れないものと考えられる。

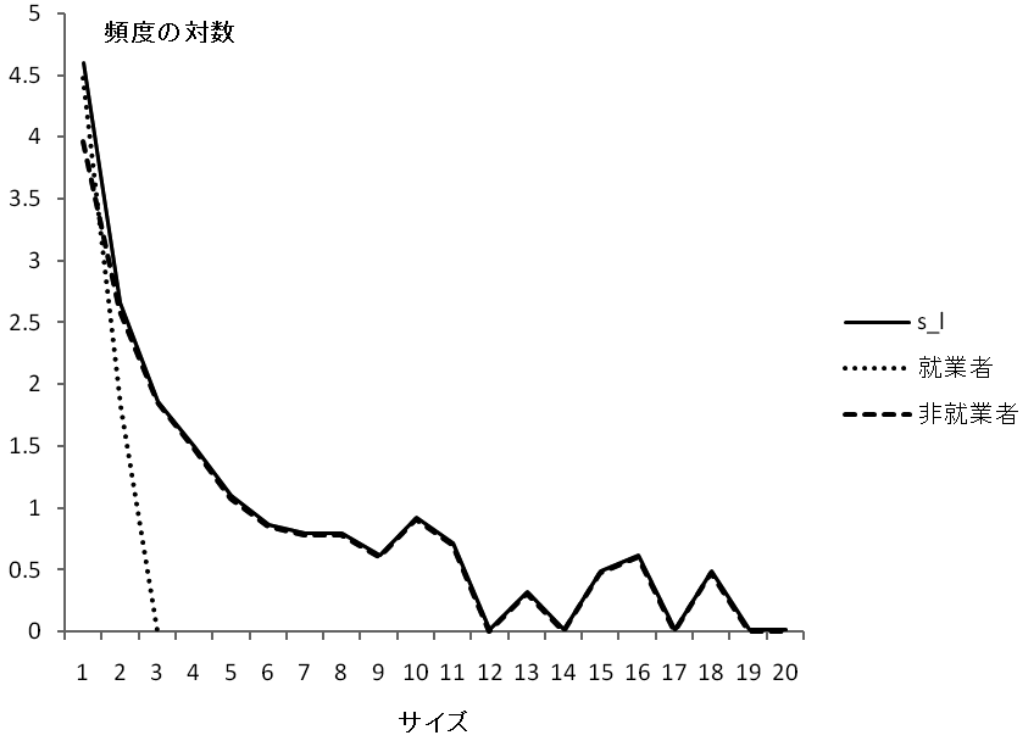

図１：センサスの就業者と非就業者の標本寸法指標

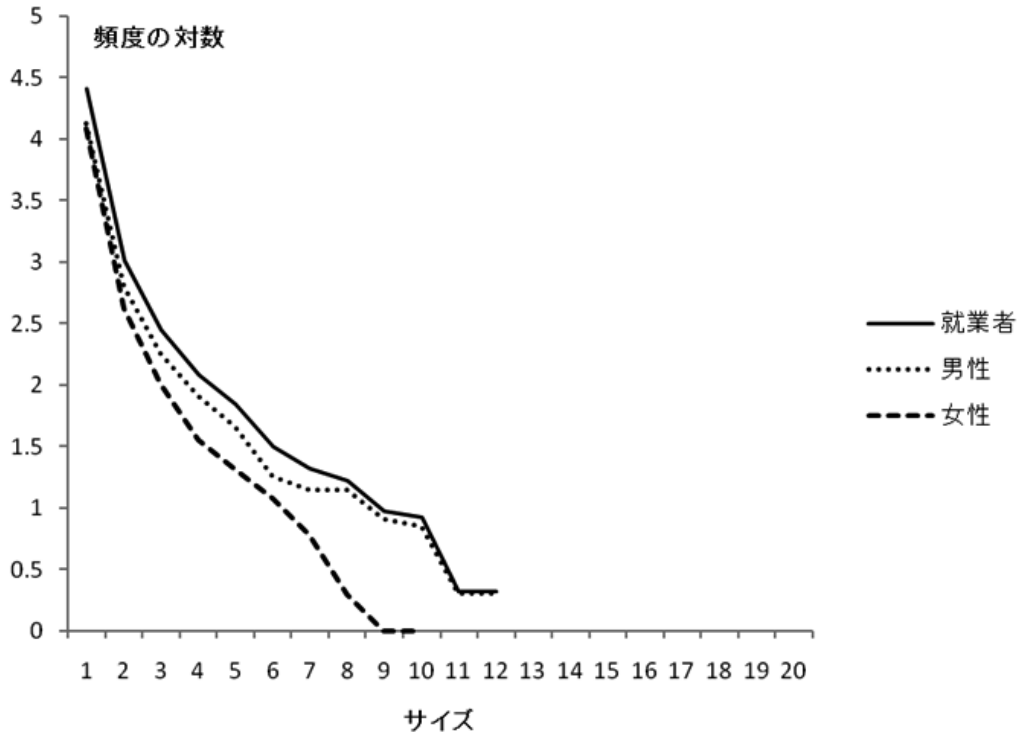

図２：センサスの就業者の男女別の標本寸法指標

４　層別変数がキー変数でない場合の事後層化

4.1　多重寸法指標を用いたリスク評価方法

この節では，標本が非復元単純無作為抽出によって得られているとき，キー変数以外の情報によって事後層化を行い，それを利用して母集団寸法指標の推定を行う方法について考える。

3.1節で行った労働力調査の個票データに対する検討では，表１に示した「都道府県番号」など14個の変数をキー変数と見なしたが，それ以外の変数の情報も事後層化に利用することができる。例えば，「県内一連番号」という変数は，国勢調査調査区と呼ばれる各都道府県内の小さなエリアを表しており，秘匿前の個票データではキー変数と見なされるが，そのまま公開すると個体の特定される可能性が著しく高くなるため，3.1節での検討では削除した。しかし，たとえ公開される個票データから削除されたとしても，事後層化を行う際にその値そのものや調査区の特性などを利用することは可能である。また，標本設計などに利用した内部的な情報など，

個票データに含まれない情報を基にして事後層化を行うことも可能である。後者も含めて，以後，キー変数以外の変数による事後層化と表現する。

キー変数以外の変数で事後層化を行った場合，すべてのキー変数の値の組み合わせによって作られる個々のセルには，複数の層の個体が含まれることになる。例えば個票データから削除された「県内一連番号」別の就業者構成を利用して，「地域特性１」，「地域特性２」，「地域特性３」という３つの層に事後層化したとする。キー変数の値の組み合わせで作られるセルのうちの１つに注目したとき，そのセルに５個の個体が入っていれば，サイズ５のセルが１個とカウントされる。ここで，５個のうち２個が「地域特性１」という層に，２個が「地域特性２」という層に，１個が「地域特性３」という層に分けられたとする。その情報は第三者には知らされないが，母集団寸法指標の推定を行う者にとっては利用が可能である。

このような事後層化に対応するために多重寸法指標を導入する。

多重寸法指標については渋谷，佐井（2007），Sibuya and Sai（2008）が紹介して，多変量に拡張したピットマンモデルの適用を検討した。ただし，実データへの当てはまりの良くない場合があることが分かっている。佐井（2009a, 2009b）では，継続調査によって得られた複数の個票データに対する同時リスク評価のために多重寸法指標を導入し，多重標本寸法指標からの多重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を提案している。適用例での推定結果は概ね満足できるものであったが，抽出率が低い場合には推定が不安定となることが懸念される。また佐井（2011）では，層化無作為抽出された標本から個票データが得られているとき，層別変数がキー変数でない場合に多重寸法指標の導入を提案している。

サンプリング法は２節と同様である。大きさ N の母集団を G とする。母集団 G から大きさ n の標本 g が非復元単純無作為抽出されていて，その標本から個票データが得られているとする。個票データの T 個のキー変数を $z_1, z_2, \cdots, z_T$ とおく。このとき，すべてのキー変数の値の組み合わせ $z_1 \times z_2 \times \cdots \times z_T$ によって多次元分割表を構成する。

ここで，キー変数以外の変数で標本を H 個の層に事後層化する。標本の第 h 層 g_h の大きさを n_h とする。また標本の事後層化に対応した母集団の第 h 層を G_h と表し，大きさを N_h として，第 h 層の抽出率を $\lambda_h = n_h/N_h$ とする。

各セルに含まれる層ごとの個体数を考えたとき，母集団での各層 $(G_1, G_2, \ldots, G_H)$ のサイズの組が $(l_1, l_2, \ldots, l_H)$ となるセル数を $S_{(l_1,l_2,\ldots,l_H)}$ と表し，多重母集団寸法指標と呼ぶ。通常のサイズ，すなわち層を無視した場合のサイズを l とすると，$l = \sum_{h=1}^{H} l_h$ である。標本でも同様に，$(d_1, d_2, \ldots, d_H)$ のサイズの組が $(l_1, l_2, \ldots, l_H)$ となるセル数を $s_{(l_1,l_2,\ldots,l_H)}$ と表し，多重標本寸法指標と呼ぶ。

L_h を第 h 層でのサイズの最大値とするとき $l_h = 0, 1, \ldots, L_h$ として，サイズ 0 を含める。ただし，通常の寸法指標で s_0 は観測対象から外され，S_0 を推定しないため，ここでは $s_{(0,0,\ldots,0)}$ を求めず，$S_{(0,0,\ldots,0)}$ は推定対象にしない。

$(0, 0, \ldots, 0)$ を除く，各層におけるすべてのサイズの組み合わせに対する多重母集団寸法指標の組と多重標本寸法指標の組を，それぞれ

$$\begin{aligned}\boldsymbol{S} \;=\; (&S_{(0,0,\ldots,0,1)}, \cdots, S_{(0,0,\ldots,0,L_H)}, \cdots \\ &\cdots, S_{(L_1,L_2,\ldots,L_{H-1},0)}, \cdots, S_{(L_1,L_2,\ldots,L_{H-1},L_H)}),\end{aligned} \tag{3}$$

$$\begin{aligned}\boldsymbol{s} \;=\; (&s_{(0,0,\ldots,0,1)}, \cdots, s_{(0,0,\ldots,0,L_H)}, \cdots \\ &\cdots, s_{(L_1,L_2,\ldots,L_{H-1},0)}, \cdots, s_{(L_1,L_2,\ldots,L_{H-1},L_H)})\end{aligned} \tag{4}$$

と表す。

母集団寸法指標の推定の手順としては，多重標本寸法指標を基に多重母集団寸法指標を推定し，$\hat{S}_l = \sum_{l_1+l_2+\cdots+l_H=l} \hat{S}_{(l_1,l_2,\cdots,l_H)}$ によって母集団寸法指標の推定値を求めることになる。

4.2　ノンパラメトリック法による推定

前述したように，多重標本寸法指標に基づいて多重母集団寸法指標を推定する方法としては，多変量のピットマンモデルを用いる方法とノンパラメトリック法が提案されている。ここでは，層化無作為標本から得られた個票データのリスク評価のために，佐井（2011）において多重寸法指標への適用を行ったノンパラメトリック推定法に，多少の変更を加えることにより事後層化に対応させる。

非復元単純無作為抽出された標本をキー変数以外の変数で事後層化することによって，多重標本寸法指標 $\boldsymbol{s}$ が得られたときの多重母集団寸法指標 $\boldsymbol{S}$ の尤度関数は

$$
\begin{aligned}
&\mathrm{L}(\ \boldsymbol{S} \mid \boldsymbol{s}\ ) \\
=\ & \frac{1}{{}_N C_n} \sum_{\mathcal{C}_1} \prod_{(l_1,l_2,\ldots,l_H)} \left\{ \frac{S_{(l_1,l_2,\ldots,l_H)}!}{\prod_{(l'_1,l'_2,\ldots,l'_H)} k_{(l_1,l_2,\ldots,l_H),(l'_1,l'_2,\ldots,l'_H)}!} \right. \\
& \left. \cdot \prod_{(l'_1,l'_2,\ldots,l'_H)} \left(\prod_{h=1}^{H} {}_{l_h} C_{l'_h} \right)^{k_{(l_1,l_2,\ldots,l_H),(l'_1,l'_2,\ldots,l'_H)}} \right\} \qquad (5)
\end{aligned}
$$

と表される。ただし，$k_{(l_1,l_2,\ldots,l_H),(l'_1,l'_2,\ldots,l'_H)}$ は母集団において各層のサイズの組が $(l_1, l_2, \ldots, l_H)$ のセルのうち，標本において各層サイズの組が $(l'_1, l'_2, \ldots, l'_H)$ として観測されるセル数を表し，$\mathcal{C}_1$ は $\boldsymbol{S}$ から $\boldsymbol{s}$ が生成されるような $k_{(l_1,l_2,\ldots,l_H),(l'_1,l'_2,\ldots,l'_H)}$ の組についてのすべての組み合わせを表す。なお，(5) 式の積の記号において母集団における各層のサイズの組み合わせ $(l_1, l_2, \ldots, l_H)$ は $(0, 0, \ldots, 0)$ を含まず，標本における各層のサイズの組み合わせ $(l'_1, l'_2, \ldots, l'_H)$ は $(0, 0, \ldots, 0)$ を含む。

(5) 式の右辺の $\mathcal{C}_1$ に関する和を A と置き，右辺を

$$
\frac{1}{{}_N C_n} \cdot A = \frac{\prod_{h=1}^{H} {}_{N_h} C_{n_h}}{{}_N C_n} \cdot \frac{1}{\prod_{h=1}^{H} {}_{N_h} C_{n_h}} \cdot A \qquad (6)
$$

と変形すると，(6) 式の右辺の積から第 1 項を除いたものは，層化無作為標本から多重標本寸法指標が得られたときの多重母集団寸法指標の尤度関数に等しくなる。2 節で述べたように，事後層化では各層の標本の大きさ n_h が一定でないが，各層の標本の大きさの条件付きで考えると，層化無作為抽出と同じと見なすことができることから, これは自然な結果である。

一般に (5) 式の計算は計算時間的に困難であるため，佐井 (2011) と同様にポアソン分布の確率関数の積で近似を行い，計算量の軽減を図る。

$$
\begin{aligned}
\mathrm{L_{ap}}(\ \boldsymbol{S} \mid \boldsymbol{s}\ ) &= \frac{\prod_{h=1}^{H} {}_{N_h}C_{n_h}}{{}_{N}C_{n}} \cdot \frac{1}{\prod_{h=1}^{H} {}_{N_h}C_{n_h} \lambda_h^{n_h} (1-\lambda_h)^{N_h - n_h}} \\
&\qquad \cdot \prod_{(l'_1, l'_2, \ldots, l'_H)} \frac{e^{-\mu_{(l'_1, l'_2, \ldots, l'_H)}}\ \mu_{(l'_1, l'_2, \ldots, l'_H)}^{s_{(l'_1, l'_2, \ldots, l'_H)}}}{s_{(l'_1, l'_2, \ldots, l'_H)}!} \\
&= \frac{1}{{}_{N}C_{n}} \cdot \frac{1}{\prod_{h=1}^{H} \lambda_h^{n_h} (1-\lambda_h)^{N_h - n_h}} \\
&\qquad \cdot \prod_{(l'_1, l'_2, \ldots, l'_H)} \frac{e^{-\mu_{(l'_1, l'_2, \ldots, l'_H)}}\ \mu_{(l'_1, l'_2, \ldots, l'_H)}^{s_{(l'_1, l'_2, \ldots, l'_H)}}}{s_{(l'_1, l'_2, \ldots, l'_H)}!} \cdot \quad (7)
\end{aligned}
$$

ただし,

$$
\begin{aligned}
&\mu_{(l'_1, l'_2, \ldots, l'_H)} \\
&= \sum_{(l_1, l_2, \ldots, l_H)(\geq (l'_1, l'_2, \ldots, l'_H))} \left\{ S_{(l_1, l_2, \ldots, l_H)} \cdot \prod_{h=1}^{H} {}_{l_h}C_{l'_h} \lambda_h^{l'_h} (1-\lambda_h)^{l_h - l'_h} \right\} \quad (8)
\end{aligned}
$$

である。また，(7) 式の積の記号において，標本における各層のサイズの組み合わせ $(l'_1, l'_2, \ldots, l'_H)$ は $(0, 0, \ldots, 0)$ を含まない。近似の詳細については佐井 (2009a，2009b) も参照して頂きたい。

(7) 式を最大とする S が近似的な最尤推定値であるが，抽出率が十分高くなければ推定は不安定であるため，経験的な知識に基づいて，多重母集団寸法指標の推定値に対して制約条件を課す必要がある。

ノンパラメトリック法を用いた母集団寸法指標の推定でも，母集団寸法指標の推定値に対して制約条件を課すことになるが，非負，サイズの増加に関して頻度が単調減少，連続する３つのサイズについて頻度の対数が下に凸という比較的簡単な制約条件で，推定をある程度安定させることが可能であった。それに対して，多重母集団寸法指標は多次元のため推定はさらに不安定で，２層への事後層化に用いる２次元の多重母集団寸法指標でさえも，多くの制約条件を課さなければ安定した推定を行うことのできないことが，これまでの検討でも分かっている。

キー変数でない層別変数を用いて標本が層化無作為抽出されてしまった場合には，多重標本寸法指標を用いなければ正しいリスク評価ができないが，事後層化では，リスク評価を行う当事者が自由に層別変数を選択することが可能である。

また層化無作為抽出では，各層において独立に標本がとられ，しかも標本配分法によっては各層における抽出率が大きく異なることも想定され，それが寸法指標を多重にしなければならない大きな要因となっていた。もし多重でない寸法指標を用いてしまうと，抽出率の異なる個体を同じサンプリングで得られたと見なすことになってしまうからである。これに対して，事後層化では標本は単純無作為抽出されているため，寸法指標を多重にする理由はだいぶ薄れており，推定精度を加味すると，多重にするメリットがあるかどうか見極めるには本論文の検討だけでは不十分である。

事後層化によって母集団寸法指標の推定に大きな改善が望めそうなキー変数以外の層別変数の候補と，それよりは改善の度合いは若干小さいと予想されるキー変数の層別変数の候補がある場合，後者を選択することもあり得るであろう。

５　おわりに

本論文では，非復元単純無作為抽出された標本から得られた個票データに対して，事後層化を行うことによってリスク評価の精度の改善，より正

確には母集団寸法指標の推定の改善を試みた。事後層化に用いる変数がキー変数か否かによって，２節で提案した層ごとの寸法指標を用いる方法と，４節で提案した多重寸法指標を用いる方法を使い分けなければならない。これらの使い分けは，標本が層化無作為抽出されたときに，その標本から得られた個票データのリスク評価を行う場合に類似している。

しかし一般に層化無作為抽出では，興味の対象となる値の平均や合計の推定精度を上げるように層別変数が選ばれるため，層別の際にリスク評価の精度の改善を考慮に入れることは考えにくい。また，層化無作為抽出の場合にサンプリング法を考慮に入れたリスク評価を行うのは，そうしなければ結果に偏りが生じるためであって，推定精度の改善を目的としたものではない。

それに対して本論文で提案した事後層化は，リスク評価の精度の改善に目的を絞っており，非復元単純無作為抽出された標本に対して，調査後，リスク評価を行う際に層別変数を自由に選択して層別を行えるため，標本設計に影響を与えない。

３節の実データに対する検討で示されたように，モデルの標本寸法指標への当てはまりが良くなく，事後層化による各層の標本寸法指標への当てはまりが良い状況が起これば，提案した方法は有効である。ただ，労働力調査やアメリカのセンサスデータにおいては有効な事後層化の方法は限られていた。事後層化によって各層で形状が異なる標本寸法指標が得られても，モデルの当てはまりが良くなければ推定は改善されない。本論文の検討だけではその点の解明は不十分で，今後，有効な事後層化の方法を考える必要があるだろう。

謝　　辞

本論文で扱った多重寸法指標に関して，慶應義塾大学名誉教授の渋谷政昭先生を初め，共同研究をしている諸先生方からは数々の有益な助言をいただきました。ここに感謝いたします。

本論文は科学研究費補助金（課題番号 22300097），統計数理研究所共同利用研究プログラム（22-共研-2031，23-共研-2028，24-共研-2030）の研究成果に基づくものである。

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］ Cochran, W. G.（1977）. *Sampling Techniques, 3rd edition*, Wiley, New York.
［3］ Levy, P. S. and Lemeshow, S.（2009）. *Sampling of Populations: Methods and Applications, 4th edition*, Wiley, New York.
［4］ Pitman, J.（1995）. Exchangeable and partially exchangeable random partitions, *Probability Theory and Related Fields*, **102**, 145–148.
［5］ 佐井至道（2009a）. 母集団多重寸法指標のノンパラメトリック推定，岡山商大論叢，**45**（1），1–22.
［6］ 佐井至道（2009b）．多重母集団寸法指標のノンパラメトリック最尤推定－２時点の個票データへの適用－，統計数理，**57**（2），425–442.
［7］ 佐井至道（2011）．層化無作為標本から得られる個票データに対するリスク評価，岡山商大論叢，**47**（1），1–22.
［8］ 渋谷政昭，佐井至道（2007）．多重指標の確率分割，応用統計学会 2007年度年会講演予稿集，13–18.
［9］ Sibuya, M. and Sai, S.（2008）. Analysis of a Dataset for Statistical Disclosure Control by Random Partition of a Multi-index, *Cherry Bud Worlshop 2008*, 1–13.
［10］ Thompson, S. K.（2012）. *Sampling, 3rd edition*, Wiley, New York.
［11］ U. S. Census Bureau（2003）. Census 2000 Public Use Microdata Sample（microdata）, Washington, D. C.