『岡山商大論叢』（岡山商科大学）
第47巻第１号　2011年７月
Journal of OKAYAMA SHOKA UNIVERSITY
Vol.47　No. 1　July 2011

《論　　説》

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

佐　井　至　道

Risk Assessment for Microdata of a Stratified Random Sample

Shido Sai

１　はじめに

個票データの公開を目的とするリスク評価について，これまで様々な方法が考えられてきたが，その個票データが標本調査で得られた場合には，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を用いることが一般的である。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。

官庁統計の標本調査では，層化抽出法，多段抽出法，系統抽出法などを組み合わせてサンプリング方法の設計をすることが多い。特に層化抽出法は推定の精度が大幅に向上することが多いため，広く用いられている。

ところが，層化無作為抽出によって標本が取られ，その寸法指標から母集団寸法指標を推定しようとする場合，仮に個票データの中に層に関する情報が含まれていたとしても，これまでは単純無作為抽出によって標本が

得られたことを前提にして推定を行うことが多かった。これは，層化無作為標本に対するリスク評価方法が構築されていなかったことと，既存のリスク評価方法を用いた場合に，それが結果に与える影響について，十分な検証が行われてこなかったことに原因があると思われる。

本論文では，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，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提案する。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すことが，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う。

まず２節では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。層化無作為抽出は，１つの層の中で見ると，他の層とは独立に非復元無作為抽出が行われているため，標本寸法指標からの母集団寸法指標の推定は容易である。

３節では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，既存の手法では標本寸法指標からの母集団寸法指標の推定は困難であるため，３.１節において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。リスク評価の手順としては，多重標本寸法指標から多重母集団寸法指標を推定し，それを基に母集団寸法指標の推定値を求めることになる。３.２節では，多重標本寸法指標からの多重母集団寸法指標のノンパラメトリック推定について説明する。

４節では，層化無作為抽出された標本を単純無作為抽出されたものと見なすことが，リスク評価に与える影響について考える。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，それぞれ４.１節，４.２節において数値例による検討を行う。

２　層別変数がキー変数の場合のリスク評価

層化無作為抽出は，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ものである。層に分ける際には何らかの変数の値を用いるのが一般的で，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，値と表記することにする。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。

この節では，層別変数がキー変数で，層化無作為抽出された標本から個票データが作成された場合のリスク評価について考える。例えば労働力調査では，北海道，東北などの11地域に分けて，各地域において独立に，異なる抽出率で標本を抽出している。このように，地域の情報は層別にしばしば用いられ，しかも誰もが個体の特定に用いることのできるキー変数と考えられる。

T個のキー変数を$z_1, z_2, \ldots, z_T$とおき，そのうちz_1を層別変数とする。母集団Gが変数z_1の値によってH個の層に分けられており，大きさN_hである母集団の第h層G_hから，大きさをn_hの標本g_hが他の層とは独立に非復元無作為抽出されている。なお，標本全体をgと表し，第h層の抽出率を$\lambda_h = n_h/N_h$とする。

このとき，層別変数を含めた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。値の組み合わせ一つ一つをセルと呼び，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。ただし，$L$はサイズの最大値である。

個票データのリスク評価は，一般に標本寸法指標と共に母集団寸法指標も用いて行われる。これは，例えば標本でサイズ１のセル数が多くても，それらのセルに母集団で多数の個体が入っていれば危険とは言えないためである。標本寸法指標は個票データから求められるが，母集団寸法指標は未知であるため推定する必要がある。

これまで提案されている母集団寸法指標の推定方法では，ポアソンガンマモデルやピットマンモデルなどのモデルを用いる方法でも，ノンパラメトリック最尤推定法でも，非復元単純無作為抽出やベルヌーイ抽出を仮定していた。上記のように，標本が層化無作為抽出されている場合には，既存の手法をそのまま適用することはできないが，各層で見ると非復元無作為抽出されているため，層ごとに既存の推定方法を用いることが可能である。

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。

層がz_1の値と１対１に対応している場合には，第h層のz_1の値はすべて等しくなるため，z_1を除くキー変数の値の組み合わせ$z_2 \times z_3 \times \cdots \times z_T$によってセルが構成されていると考えてもよい。これは，個票データ全体から作られる多次元分割表について，z_1の値ごとにスライスしたものと考えられ，各層のセル数K_hはK/Hとなる。

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$\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_L$を求めることになる。

３　層別変数がキー変数でない場合のリスク評価

3.1　多重寸法指標の導入

この節では，層別変数がキー変数に含まれず，層化無作為抽出された標本から個票データが作成された場合のリスク評価について考える。

ここで，２つの変数「年齢」と「続柄」のみがキー変数で，「地域特性」が層別変数であるもののキー変数でない場合を例として説明する。すなわち，層である「地域特性」ごとに標本が独立に非復元無作為抽出されており，それらを１つにまとめた標本から個票データが作成されている。通常，層別変数である「地域特性」は個票データにも変数として含まれるが，第三者にとって，母集団に含まれる各個体の変数「地域特性」の値は未知であり，個票データの個体を特定するために「地域特性」の情報を用いることができないものとする。

このとき，セルは２つのキー変数「年齢」と「続柄」の値の組み合わせによって構成される。例えば，「年齢」×「続柄」が「35歳」×「世帯主」という組み合わせのセルに５個の個体が入っていれば，その個体がどの層から抽出されたかということとは無関係に，サイズ５のセルが１個とカウントされる点は，非復元単純無作為抽出された標本から個票データが作成された場合と同じである。

ここで，５個のうち２個が「地域特性A」という層から，３個が「地域特性B」という層から取られた個体であるとする。前述したように，その情報は第三者にとって個体の特定を助けるものにはならないが，母集団寸法指標の推定を行う場合には考慮に入れる必要が生じる。

非復元単純無作為抽出では，各個体の抽出率はすべて等しく，抽出は互いに独立ではない。これに対して，上記の例では，異なる層の個体は独立に抽出される上に，一般に層ごとに抽出率も異なるため，仮に標本寸法指標を求めたとしても，それから母集団寸法指標を推定することはできない。

そこで，母集団寸法指標を推定するために，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ている。ただし, 実データへの当てはまりが良くない場合のあることが分かっている。また，佐井（2007, 2008）では馬場，坂口（2006）の紹介した学会所属データに対してノンパラメトリック推定を行っている。このデータは多重寸法指標においてサイズの最大値が１となる特殊なケースである。さらに，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。

前節と同様に，母集団GがあらかじめH個の層に分けられており，大きさN_hである母集団の第h層G_hから，大きさをn_hの標本g_hが他の層とは独立に非復元無作為抽出されているとする。なお, 標本全体をgと表し，第h層の抽出率を $\lambda_h = n_h/N_h$とする。

ここで，すべてのキー変数の値の組み合わせによってセルを構成したとき，各セルに含まれる層ごとの個体数を考える。$(G_1, G_2, \ldots, G_H)$のサイズの組，すなわち母集団で各層におけるサイズの組が$(l_1, l_2, \ldots, l_H)$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。なお，通常の寸法指標でs_0は観測対象から外され，S_0を推定しないのと同様に，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ことにする。

$(0, 0, \ldots, 0)$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

$$\begin{aligned}\boldsymbol{S} &= (S_{(0,0,\ldots,0,1)},\cdots,S_{(0,0,\ldots,0,L_H)},\cdots \\ &\qquad \cdots,S_{(L_1,L_2,\ldots,L_{H-1},0)},\cdots,S_{(L_1,L_2,\ldots,L_{H-1},L_H)}), \quad (1)\\ \boldsymbol{s} &= (s_{(0,0,\ldots,0,1)},\cdots,s_{(0,0,\ldots,0,L_H)},\cdots \\ &\qquad \cdots,s_{(L_1,L_2,\ldots,L_{H-1},0)},\cdots,s_{(L_1,L_2,\ldots,L_{H-1},L_H)}) \quad (2)\end{aligned}$$

と表す。

3.2　多重母集団寸法指標と母集団寸法指標の推定

層化無作為抽出された標本から$\boldsymbol{s}$が得られたときの$\boldsymbol{S}$の尤度関数は，佐井（2009a, 2009b）と同様に

$$\begin{aligned}&\mathrm{L}(\,\boldsymbol{S}\mid\boldsymbol{s}\,)\\ &= \frac{1}{\prod_{h=1}^{H}{}_{N_h}C_{n_h}}\sum_{\mathcal{C}_1}\prod_{(l_1,l_2,\ldots,l_H)}\Bigg\{\frac{S_{(l_1,l_2,\ldots,l_H)}!}{\prod_{(l'_1,l'_2,\ldots,l'_H)}k_{(l_1,l_2,\ldots,l_H),(l'_1,l'_2,\ldots,l'_H)}!}\\ &\qquad\cdot\prod_{(l'_1,l'_2,\ldots,l'_H)}\Big(\prod_{h=1}^{H}{}_{l_h}C_{l'_h}\Big)^{k_{(l_1,l_2,\ldots,l_H),(l'_1,l'_2,\ldots,l'_H)}}\Bigg\} \quad (3)\end{aligned}$$

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，標本において各層のサイズの組が$(l'_1,l'_2,\ldots,l'_H)$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，(3) 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。

層の数H，母集団G_hの大きさN_hと最大サイズL_hが増加するにつれ$\mathcal{C}_1$の組み合わせの数は膨大になり，時間的に (3) 式の計算は困難となるため，次のような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。

$$\mathrm{L_{ap}}(\,\boldsymbol{S}\mid\boldsymbol{s}\,)=\frac{1}{\prod_{h=1}^{H}{}_{N_h}C_{n_h}\lambda_h^{n_h}(1-\lambda_h)^{N_h-n_h}}\cdot\prod_{(l'_1,l'_2,\cdots,l'_H)}\frac{e^{-\mu_{(l'_1,l'_2,\cdots,l'_H)}}\ \mu_{(l'_1,l'_2,\cdots,l'_H)}^{s_{(l'_1,l'_2,\cdots,l'_H)}}}{s_{(l'_1,l'_2,\cdots,l'_H)}!}. \tag{4}$$

ただし，

$$\mu_{(l'_1,l'_2,\cdots,l'_H)}=\sum_{(l_1,l_2,\cdots,l_H)(\geq(l'_1,l'_2,\cdots,l'_H))}\left\{S_{(l_1,l_2,\cdots,l_H)}\cdot\prod_{h=1}^{H}{}_{l_h}C_{l'_h}\lambda_h^{l'_h}(1-\lambda_h)^{l_h-l'_h}\right\} \tag{5}$$

である。また，(4) 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）を参照して頂きたい。

(4) 式を最大とする $\boldsymbol{S}$ が近似的な最尤推定値となるが，抽出率が高い場合を除け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。

母集団寸法指標の推定値は多重母集団寸法指標の推定値を用いて，$\hat{S}_l=\sum_{l_1+l_2+\cdots+l_H=l}\hat{S}_{(l_1,l_2,\cdots,l_H)}$ によって求めることができる。

4　層化抽出を無視した場合のリスク評価への影響

4.1　層別変数がキー変数の場合の影響

これまで，官庁統計の標本から得られた個票データのリスク評価は，実際に行われたサンプリング法に関わらず，非復元単純無作為抽出やベルヌーイ抽出を仮定して行われることが多かった。この節では，層化無作為抽出された標本から個票データが作成されているにも関わらず，その標本を非復元単純無作為抽出されたと見なすことが，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検討する。

まず初めに，層別変数がキー変数の場合について考える。なお，ここで

は層の数Hを２に限定して検討を行う。

２節で述べたように，層化無作為抽出された標本は，層ごとに見ると非復元無作為抽出されていると見なすことができ，さらに層別変数がキー変数の場合には，各層の寸法指標についてサイズごとに和をとると，全体の寸法指標と一致する。

したがって前述したように，母集団寸法指標を推定するには，第１層の標本寸法指標 $s_{1,1}, s_{1,2}, \ldots, s_{1,L}$ から第１層の母集団寸法指標の推定値 $\hat{S}_{1,1}, \hat{S}_{1,2}, \ldots, \hat{S}_{1,L}$ を求め，第２層の標本寸法指標$s_{2,1}, s_{2,2}, \ldots, s_{2,L}$ から第２層の母集団寸法指標の推定値 $\hat{S}_{2,1}, \hat{S}_{2,2}, \ldots, \hat{S}_{2,L}$ を求め，それぞれのサイズについて $\hat{S}_l = \hat{S}_{1,l} + \hat{S}_{2,l}$ $(l = 1, 2, \ldots, L)$ と合算することによって$\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_L$を求めるのが正しい手順であった。

これに対して，標本が非復元単純無作為抽出されたと見なした場合には，全体の標本寸法指標 $s_1, s_2, \ldots, s_L$ から，母集団寸法指標の推定値 $\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_L$ を直接求めることになる。ここで，$s_l = s_{1,l} + s_{2,l}$ $(l = 1, 2, \ldots, L)$である。

サンプリング法を誤解することが推定に与える影響を考える場合，まず母集団における各セルのサイズを定め，実際に個体をサンプリングすることによって標本の各セルのサイズを求めて標本寸法指標を計算し，その情報を基に母集団寸法指標を推定するシミュレーションを行うことも考えられるが，その場合には推定結果に標本誤差が含まれるため，誤解による影響を議論するためには，これら一連の操作を多数回繰り返す必要が生じる。ここでは標本誤差を排除するために，各セルのサイズに対して超母集団モデルを考え，母集団と標本の寸法指標，すなわち各サイズに対する頻度を，モデルでの期待値をとることによって得ることにする。

超母集団モデルとして，各層で独立なポアソンガンマモデルを用いる。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い。最近はPitman（1995）が提案したピットマンモデル（Ewens-Pitman sampling formula（EPSF）とも呼ばれる）のような２個のパラメータを持つモデルが用いられることが多いが，ここでは，次節で用いる２次元のモデルへの拡張が容易であることから，ポアソンガンマモデルを用いることにした。なおポアソンガンマモデルを用いる場合，サイズの最大値Lは無限大として考える。

ポアソンガンマモデルでは，サンプリング法としてベルヌーイ抽出を仮定している。ベルヌーイ抽出とは，抽出率をλとしたとき，各個体を他の個体とは独立に，確率λで抽出するものである。したがってこの節での検討は，正確に言えば，層化ベルヌーイ抽出された標本を単純なベルヌーイ抽出によって得られたものと見なすことになるが，標本が十分大きい場合には，ベルヌーイ抽出と非復元無作為抽出との違いは無視しても問題は生じない。以後も混乱を避けるため，無作為抽出という表現を用いて説明する。

母集団と標本において，第h層のj番目のセルに含まれる個体数をそれぞれ$F_{h,j}$，$f_{h,j}$ $(h = 1, 2\ ;\ j = 1, 2, \ldots, K_h)$とする。ポアソンガンマモデルでは，標本で各セルに含まれる個体数は

$$f_{h,j} \sim \mathsf{Poisson}(n_h \pi_{h,j}), \tag{6}$$

$$\pi_{h,j} \sim \mathsf{gamma}(\alpha_h, \beta_h) \tag{7}$$

と表現される。ただし，$f_{h,j}$，$\pi_{h,j}$はそれぞれh, jについて互いに独立である。α_h，β_hはパラメータであるが，個体数の期待値が各層の標本の大きさn_hに一致するには$\alpha_h \beta_h = 1/K_h$という制約条件が必要なため，実質的なパラメータはβ_hのみである。ポアソンガンマモデルでの第h層における標本寸法指標の期待値は次のように書ける。

$$\mathrm{E}(s_{h,0}) = \frac{K_h}{(1+n_h\beta_h)^{\alpha_h}}, \tag{8}$$

$$\mathrm{E}(s_{h,l}) = \frac{K_h(n_h\beta_h)^l(\alpha_h+l-1)\cdots\alpha_h}{l!(1+n_h\beta_h)^{\alpha_h+l}} \quad (l \geq 1). \tag{9}$$

母集団で各セルに含まれる個体数$F_{h,j}$，母集団寸法指標の期待値$\mathrm{E}(S_{h,0}), \mathrm{E}(S_{h,l})$は，(6)，(8)，(9) 式において，$n_h$を$N_h$で置き換えるだけである。

層化無作為抽出された標本を単純無作為抽出されたと誤解した場合，佐井（1993a, 1993b）では，比例配分法によって各層への標本の大きさの割り当てが行われた場合には，母平均，母分散の推定への影響が小さいことが示されている。母集団寸法指標の推定においても，標本配分法によって影響の大きさに違いが生じる可能性が大きい。

そのため，まず各層の母集団の大きさ，総セル数，パラメータの値を，N_1=N_2 = 100000，$K_1 = K_2 = 5.0 \times 10^9$，$\beta_1 = \beta_2 = 1.0 \times 10^{-5}$と固定し，標本の大きさの合計も$n$ = 20000 と固定するものの，各層の標本の大きさとして（n_1, n_2）= (10000, 10000) , (11000, 9000) , . . . , (19000, 1000) の10通りの場合を考える。なお，母集団の２つの層の大きさ，セル数，パラメータがすべて等しい設定にしているが，その場合，２つの層を併せた全体のモデルは，N = 200000，$K = 10.0 \times 10^9$，$\beta = 0.5 \times 10^{-5}$のポアソンガンマモデルとなる。

このうち，（n_1, n_2）= (15000, 5000) の場合を例に，検討方法を簡単に説明する。この場合，各層の抽出率はλ_1 = 0.15，λ_2 = 0.05 である。

前述したように，標本誤差を除くために各層における標本の寸法指標として（9）式の期待値を用い，母集団寸法指標として（9）式のn_hをN_hで置き換えた値を用いる。

$$(S_{1,1}, S_{1,2}, S_{1,3}, S_{1,4}, S_{1,5}, \dots) = (49999, 12500, 4167, 1563, 625, \dots), \tag{10}$$

$$(S_{2,1}, S_{2,2}, S_{2,3}, S_{2,4}, S_{2,5}, \dots) = (49999, 12500, 4167, 1563, 625, \dots), \quad (11)$$

$$(s_{1,1}, s_{1,2}, s_{1,3}, s_{1,4}, s_{1,5}, \dots) = (13043, 851, 74, 7, 1, \dots), \quad (12)$$

$$(s_{2,1}, s_{2,2}, s_{2,3}, s_{2,4}, s_{2,5}, \dots) = (4762, 113, 4, 0, 0, \dots) \quad (13)$$

となる。ただし，小数点以下を四捨五入して整数として表示している。

各層の標本寸法指標（12），（13）式からパラメータ β_1, β_2 をそれぞれ推定すると，ともに真の値 1.0×10^{-5} が得られる。その値を（9）式で n_h を N_h で置き換えたものに代入すると，各層の真の母集団寸法指標（10），（11）式が推定され，母集団全体に対する寸法指標の推定値として２つの層の母集団寸法指標の推定値の和

$$(\hat{S}_1, \hat{S}_2, \hat{S}_3, \hat{S}_4, \hat{S}_5, \dots) = (99999, 25000, 8333, 3125, 1250, \dots) \quad (14)$$

が得られる。これが正しい手順による推定である。

単純無作為抽出された標本と誤解した場合には，２つ層の標本寸法指標の和

$$(s_1, s_2, s_3, s_4, s_5, \dots) = (17805, 964, 78, 7, 1, \dots) \quad (15)$$

を基にして，標本が抽出率 λ =（15000 + 5000）/（100000 + 100000）= 0.1 でサンプリングされたと解釈して，母集団寸法指標を推定してしまう。このとき，パラメータの推定値は $\hat{\beta} = 0.625 \times 10^{-5}$ であり，母集団寸法指標の推定値は

$$(\hat{S}_1, \hat{S}_2, \hat{S}_3, \hat{S}_4, \hat{S}_5, \dots) = (88888, 24691, 9145, 3810, 1694, \dots) \quad (16)$$

と得られる。

表１は（n_1, n_2）の各組み合わせについて，単純無作為標本と誤解した方法で推定した $\hat{\beta}$ の値と，母集団全体の寸法指標の推定値のうちでサイズ５以下の部分を示している。

リスク評価において最も重要視される $\hat{S}_1$ に注目すると，（n_1, n_2）=（10000,10000）の場合，すなわち比例配分されている場合には真の値 S_1 に一致するが，２つの層の抽出率が偏るにしたがって値は減少し，（n_1, n_2）=（19000, 1000）では S_1 の71%程度になってしまう。これはリスクを過小評価することにつながるため特に注意が必要である。また，抽出率が偏る

につれ$\hat{S}_2$も単調に減少するが，$\hat{S}_3$以降は単調に増加している。

２層の標本の大きさを$n_1 = n_2 = 10000$と固定して，母集団の大きさの合計も200000に固定するものの，各層の母集団の大きさとして（N_1, N_2）=（100000, 100000），(110000, 90000），. . . ,（190000, 10000）の10通りを考えて同様の検討を行うと，N_1, N_2の値が偏るほどS_1は過大評価され，（N_1, N_2）=（190000, 10000）では$\hat{S}_1$はS_1の約134%になる。

次に２層のポアソンガンマモデルのパラメータβ_1, β_2が異なる場合について考える。

各層の母集団の大きさ，総セル数を，$N_1 = N_2 = 100000$，$K_1 = K_2 = 5.0 \times 10^9$と固定する。パラメータとして，（$\beta_1, \beta_2$）=（$2.0\times10^{-5}, 0.5\times10^{-5}$），（$0.5\times10^{-5}, 2.0 \times 10^{-5}$）の２通りを考え，それぞれについて，（$n_1, n_2$）として表１と同じ10通りの場合を考える。

ポアソンガンマモデルのパラメータβ_1, β_2が大きくなるほど，寸法指標におけるサイズの小さい部分の頻度は小さくなり，寸法指標は裾を長く引くようになる。検討では第１層の抽出率を高くしているが，（β_1, β_2）=（$2.0 \times10^{-5}, 0.5 \times 10^{-5}$）の場合には第１層の方が裾を長く引き，（$\beta_1, \beta_2$）=（$0.5 \times10^{-5}, 2.0 \times 10^{-5}$）の場合には第１層の方が裾が短くなる。

結果を表２，３に示す。これらの場合，２つの層を併せた全体のモデルは標本でも母集団でもポアソンガンマモデルとならないが，パラメータβ^*

表１　$\hat{\beta}$の値と母集団寸法指標の推定値

(n_1, n_2)	$\hat{\beta}$ $(\times10^{-5})$	$\hat{S}_1$	$\hat{S}_2$	$\hat{S}_3$	$\hat{S}_4$	$\hat{S}_5$
(10000, 10000)	0.500	99999	25000	8333	3125	1250
(11000, 9000)	0.505	99501	25000	8375	3156	1269
(12000, 8000)	0.520	98038	24991	8494	3248	1325
(13000, 7000)	0.545	95692	24954	8676	3394	1416
(14000, 6000)	0.580	92591	24863	8902	3585	1540
(15000, 5000)	0.625	88888	24691	9145	3810	1694
(16000, 4000)	0.680	84744	24418	9381	4055	1869
(17000, 3000)	0.745	80320	24032	9587	4303	2060
(18000, 2000)	0.820	75756	23531	9745	4540	2256
(19000, 1000)	0.905	71173	22923	9843	4755	2450
真の値	0.500	99999	25000	8333	3125	1250

のポアソンガンマモデルと見なして推定している。そのため，見かけ上のパラメータの推定値 $\hat{\beta}^*$ が真の値と一致しても，母集団寸法指標の推定値は真の値と一致しない。

表２の（β_1, β_2）＝（$2.0 \times 10^{-5}, 0.5 \times 10^{-5}$）の場合，２つの層の抽出率が偏るにしたがって $\hat{S}_1$ の値は急激に減少し，（n_1, n_2）＝（19000, 1000）では S_1 の43%程度になる。$\hat{S}_2$ も単調に減少しているが，真の値に最も近いのは（n_1, n_2）＝（14000, 6000）のときである。なお，$\hat{S}_3$, $\hat{S}_4$ の値の変化は単調ではない。

表３の（β_1, β_2）＝（$0.5\times10^{-5}, 2.0\times10^{-5}$）の場合には，$\hat{S}_1$ を含め，寸法指標のすべてのサイズについて変化は単調でない。推定値と真の値との

表２　$\hat{\beta}^*$の値と母集団寸法指標の推定値（(β_1, β_2) = ($2.0\times10^{-5}, 0.5\times10^{-5}$)）

(n_1, n_2)	$\hat{\beta}^*$ $(\times10^{-5})$	$\hat{S}_1$	$\hat{S}_2$	$\hat{S}_3$	$\hat{S}_4$	$\hat{S}_5$
(10000, 10000)	0.625	88888	24691	9145	3810	1694
(11000, 9000)	0.706	82901	24269	9473	4160	1948
(12000, 8000)	0.800	76922	23669	9710	4482	2206
(13000, 7000)	0.906	71110	22914	9844	4758	2453
(14000, 6000)	1.025	65573	22037	9875	4978	2677
(15000, 5000)	1.156	60377	21075	9809	5136	2868
(16000, 4000)	1.300	55555	20062	9659	5232	3023
(17000, 3000)	1.456	51118	19026	9442	5272	3139
(18000, 2000)	1.625	47058	17993	9173	5261	3218
(19000, 1000)	1.806	43360	16980	8866	5208	3263
真の値	0.625	99999	22222	7408	3086	1482

表３　$\hat{\beta}^*$の値と母集団寸法指標の推定値（(β_1, β_2) = ($0.5\times10^{-5}, 2.0\times10^{-5}$)）

(n_1, n_2)	$\hat{\beta}^*$ $(\times10^{-5})$	$\hat{S}_1$	$\hat{S}_2$	$\hat{S}_3$	$\hat{S}_4$	$\hat{S}_5$
(10000, 10000)	0.625	88888	24691	9145	3810	1694
(11000, 9000)	0.556	94673	24929	8752	3457	1456
(12000, 8000)	0.500	99999	25000	8333	3125	1250
(13000, 7000)	0.456	104574	24948	7936	2840	1084
(14000, 6000)	0.425	108106	24836	7607	2622	964
(15000, 5000)	0.406	110343	24733	7392	2485	891
(16000, 4000)	0.400	111109	24692	7316	2439	867
(17000, 3000)	0.406	110343	24733	7392	2485	891
(18000, 2000)	0.425	108106	24836	7608	2622	964
(19000, 1000)	0.456	104573	24948	7936	2840	1084
真の値	0.625	99999	22222	7408	3086	1482

差は表２と比較して概ね小さいがその変化は複雑で，$\hat{S}_1$が最大になるのは（n_1, n_2）=（16000, 4000）のときであり，真の値に最も近いのは（n_1, n_2）=（12000, 8000）のときである。

以上のように，推定値のバイアスには様々な要因が絡んでいる。特に，各層のパラメータが異なる場合には，比例配分法を用いて層化無作為抽出が行われていたとしても，バイアスが生じる可能性があることに注意しなければならない。

4.2　層別変数がキー変数でない場合の影響

この節では，層別変数がキー変数でない場合に，層化無作為抽出された標本から個票データが作成されているにも関わらず，その標本を非復元単純無作為抽出されたと見なすことが，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検討する。なお，ここでも層の数Hは２に限定する。

３節で述べたように，層別変数がキー変数でない場合には，層ごとに見ると非復元無作為抽出されていると見なすことができるものの，すべてのキー変数の値の組み合わせによって作られるセルには，複数の層の個体が含まれる。

前述したように，母集団寸法指標を推定するには，まず，標本の各セルに入っている個体を２つの層（d_1, d_2）に分けて（l_1, l_2）とカウントして多重標本寸法指標 $s_{(l_1,l_2)}$の組$\boldsymbol{s}$を求める。そして，それを基に多重母集団寸法指標$S_{(l_1,l_2)}$の組の推定値$\hat{\boldsymbol{S}}$を求める。母集団寸法指標の推定値は多重母集団寸法指標の推定値を用いて，$\hat{S}_l = \sum_{l_1+l_2=l} \hat{S}_{(l_1,l_2)}$ によって求める。これが正しい手順であった。

これに対して，標本が非復元単純無作為抽出されたと見なした場合には，各セルに入っている個体数の合計から得られる標本寸法指標$s_1, s_2, \ldots, s_L$を基に，母集団寸法指標の推定値 $\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_L$を直接

求めることになる。ここで，$s_l = \sum_{l_1+l_2=l} s_{(l_1,l_2)}$ である。

前節と同様の検討を行うために，次のような２次元ポアソンガンマモデルを用いる。

母集団と標本において，j 番目のセルに含まれる第 h 層の個体数をそれぞれ $F_{h,j}, f_{h,j}$（$h = 1, 2$；$j = 1, 2, \ldots, K$）とする。前節と同じ記号を用いているが，層別変数がキー変数の場合には $f_{1,j}$ と $f_{2,j}$ は異なるセルに入っている個体数を表しているのに対して，ここでは $f_{1,j}$ と $f_{2,j}$ は同じセルに入っている個体を，層ごとに分けてカウントした数を表している。標本で各セルに含まれる個体数は

$$
\begin{aligned}
f_{1,j} &\sim \mathsf{Poisson}(n_1\pi_{1,j}), \quad \pi_{1,j} = \pi_{0,j} + \pi'_{1,j}, && (17)\\
f_{2,j} &\sim \mathsf{Poisson}(n_2\pi_{2,j}), \quad \pi_{2,j} = \pi_{0,j} + \pi'_{2,j}, && (18)\\
\pi_{0,j} &\sim \mathsf{gamma}(r\alpha_0, \beta_0), && (19)\\
\pi'_{1,j} &\sim \mathsf{gamma}((1-r)\alpha_0, \beta_0), && (20)\\
\pi'_{2,j} &\sim \mathsf{gamma}((1-r)\alpha_0, \beta_0) && (21)
\end{aligned}
$$

と表現される。ただし，$\pi_{0,j}$，$\pi'_{1,j}$，$\pi'_{2,j}$ は互いに，そしてすべての j についても独立である。

２つの層のパラメータ α_0，β_0 を共通にしているのは，ガンマ分布の再生性を利用して，$\pi_{1,j} \sim \mathsf{gamma}(\alpha_0, \beta_0)$，$\pi_{2,j} \sim \mathsf{gamma}(\alpha_0, \beta_0)$ とするためである。このとき，$\pi_{1,j}$ と $\pi_{2,j}$ との相関係数は r となるが，仮に r = １としても $f_{1,j}$ と $f_{2,j}$ との相関係数は１とはならないため，モデルとしての表現力は不十分である。

この点を補うため，３つの独立なポアソンガンマモデルの確率変数 $f'_{0,j}$, $f'_{1,j}$ と $f'_{2,j}$ を用いて，$f_{1,j} = f'_{0,j} + f'_{1,j}$，$f_{2,j} = f'_{0,j} + f'_{2,j}$ のようにそれぞれ２つの確率変数の和とすること，または $f'_{1,j}$ と $f'_{2,j}$ を独立とせず，上述の２次元ポアソンガンマモデルとすることも可能であるが，多重寸法指標の期待値が，実データから得られるものとはかけ離れたものとなるこ

とから，ここでは使用を見送った。

２次元ポアソンガンマモデルでの多重標本寸法指標の期待値は，次のように書ける。

$$\begin{aligned}\mathrm{E}(s_{(l_1,l_2)}) \;=\; & \frac{K(n_1\beta_0)^{l_1}(n_2\beta_0)^{l_2}}{l_1!\;l_2!} \\ & \cdot\sum_{i_1=0}^{l_1}\sum_{i_2=0}^{l_2} {}_{l_1}C_{i_1}\cdot{}_{l_2}C_{i_2}\cdot\frac{\{r\alpha_0+(i_1+i_2)-1\}\cdots(r\alpha_0)}{\{1+(n_1+n_2)\beta_0\}^{r\alpha_0+(i_1+i_2)}} \\ & \cdot\frac{\{(1-r)\alpha_0+(l_1-i_1)-1\}\cdots\{(1-r)\alpha_0\}}{(1+n_1\beta_0)^{(1-r)\alpha_0+(l_1-i_1)}} \\ & \cdot\frac{\{(1-r)\alpha_0+(l_2-i_2)-1\}\cdots\{(1-r)\alpha_0\}}{(1+n_2\beta_0)^{(1-r)\alpha_0+(l_2-i_2)}}. \quad (22)\end{aligned}$$

ただし，(22) 式の右辺の和における３つの分数の分子は，それぞれ $i_1+i_2=0$，$i_1=l_1$，$i_2=l_2$ のときには１と考えるものとする。

母集団で各セルに含まれる個体数$F_{1,j}$，$F_{2,j}$，多重母集団寸法指標の期待値$\mathrm{E}(S_{(l_1,l_2)})$は，(17)，(18)，(22) 式において，$n_h$を$N_h$で置き換えるだけである。

検討では，各層の母集団の大きさ，総セル数，パラメータの値を，$N_1=N_2=100000$，$K=1.0\times10^{10}$，$\beta_0=1.0\times10^{-5}$ と固定し，相関係数として $r=0.0, 0.4, 0.8$ の３通りを考える。標本の大きさの合計を$n=20000$と固定するものの，各層の標本の大きさとして$(n_1,n_2)=(10000,10000),(11000,9000),\ldots,(19000,1000)$の10通りの場合を考える。なお，各セルに含まれる個体数の合計 $f_j=f_{1,j}+f_{2,j}$ は，$r=0.0$ の場合にはポアソンガンマモデルに従うが，$r\neq0.0$ の場合にはポアソンガンマモデルとはならない。

$r=0.8$，$(n_1,n_2)=(15000,5000)$の場合を例に，検討方法を簡単に説明する。この場合，各層の抽出率は$\lambda_1=0.15$，$\lambda_2=0.05$である。

多重母集団寸法指標と多重標本寸法指標としては，前節と同様に，次のようにモデルでの期待値をとったものを用いる。ただし，小数点以下を四

捨五入して整数として表示している。また$S_{(0,0)}$，$s_{(0,0)}$は数値の表示を省いて，“＊”と表示している。

$$\begin{pmatrix} S_{(0,0)} & S_{(0,1)} & S_{(0,2)} & \cdots \\ S_{(1,0)} & S_{(1,1)} & S_{(1,2)} & \cdots \\ S_{(2,0)} & S_{(2,1)} & S_{(2,2)} & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix} = \begin{pmatrix} * & 36666 & 6944 & \cdots \\ 36666 & 8889 & 2963 & \cdots \\ 6944 & 2963 & 1481 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}, \tag{23}$$

$$\begin{pmatrix} s_{(0,0)} & s_{(0,1)} & s_{(0,2)} & \cdots \\ s_{(1,0)} & s_{(1,1)} & s_{(1,2)} & \cdots \\ s_{(2,0)} & s_{(2,1)} & s_{(2,2)} & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix} = \begin{pmatrix} * & 4286 & 92 & \cdots \\ 12608 & 417 & 17 & \cdots \\ 795 & 52 & 3 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}. \tag{24}$$

多重標本寸法指標からパラメータβ_0と相関係数rを推定すると，それぞれ真の値が得られ，それらの値を（22）式でn_hをN_hで置き換えたものに代入すると，真の多重母集団寸法指標（23）式が推定される。したがって，母集団寸法指標の推定値は，

$$\hat{S}_1 = \hat{S}_{(0,1)} + \hat{S}_{(1,0)} = 36666 + 36666 = 73332, \tag{25}$$

$$\hat{S}_2 = \hat{S}_{(0,2)} + \hat{S}_{(1,1)} + \hat{S}_{(2,0)} = 6944 + 8889 + 6944 = 22777 \tag{26}$$

のように多重母集団寸法指標の和をとることによって，

$$(\hat{S}_1, \hat{S}_2, \hat{S}_3, \hat{S}_4, \hat{S}_5, \dots) = (73332, 22778, 9568, 4576, 2357, \dots) \tag{27}$$

と得られる。これが正しい手順による推定である。

単純無作為抽出された標本と誤解した場合には，各セルに入っている個体数の合計から得られる標本寸法指標から，直接，母集団寸法指標を推定してしまう。このとき標本寸法指標は

$$s_1 = s_{(0,1)} + s_{(1,0)} = 4286 + 12608 = 16894, \tag{28}$$

$$s_2 = s_{(0,2)} + s_{(1,1)} + s_{(2,0)} = 92 + 417 + 795 = 1304 \tag{29}$$

のように多重標本寸法指標の和をとることによって，

$$(s_1, s_2, s_3, s_4, s_5, \dots) = (16894, 1304, 139, 17, 2, \dots) \quad (30)$$

と得られる。これをパラメータ β^* のポアソンガンマモデルから得られたものとみなし，抽出率 $\lambda = (15000 + 5000) / (100000 + 100000) = 0.1$ でサンプリングされたと解釈してパラメータを推定すると $\hat{\beta}^* = 0.925 \times 10^{-5}$ となり，それを基に母集団寸法指標の推定値を求めると

$$(\hat{S}_1, \hat{S}_2, \hat{S}_3, \hat{S}_4, \hat{S}_5, \dots) = (70175, 22776, 9856, 4799, 2492, \dots) \quad (31)$$

となる。

表４，５，６はそれぞれ相関係数が $r = 0.0, 0.4, 0.8$ の場合に，(n_1, n_2) の各組み合わせについて同様の方法で推定した $\hat{\beta}$ または $\hat{\beta}^*$ の値と，母集団寸法指標の推定値のうちでサイズ５以下の部分を示している。前述したように，$r \neq 0.0$ の場合にはポアソンガンマモデルとはならないため，β^* は見かけ上のパラメータであり，推定値 $\hat{\beta}^*$ が真の値と一致しても，母集団寸法指標の推定値 $\hat{S}_l$ は真の値と一致しない。$r = 0.0$ の場合にはポアソンガンマモデルとなるため，パラメータの推定値は $\hat{\beta}$ と表記している。

相関係数が大きくなるにつれて真の値との差は一般に小さくなり，$r = 0.8$では，$(n_1, n_2) = (19000, 1000)$ の場合でも，$\hat{S}_1$ の値は真の値の92%である。また，$r = 0.0$ の場合には，$N = 200000$，$K = 1.0 \times 10^{10}$，$\beta = 0.5 \times 10^{-5}$ のポアソンガンマモデルとなり，推定結果は層別変数がキー変数の場合にほぼ等しくなる。

表４　$\hat{\beta}$ の値と母集団寸法指標の推定値（$r = 0.0$）

(n_1, n_2)	$\hat{\beta}$ $(\times 10^{-5})$	$\hat{S}_1$	$\hat{S}_2$	$\hat{S}_3$	$\hat{S}_4$	$\hat{S}_5$
(10000, 10000)	0.500	99999	25000	8333	3125	1250
(11000, 9000)	0.505	99501	25000	8375	3156	1269
(12000, 8000)	0.520	98038	24991	8494	3248	1325
(13000, 7000)	0.545	95692	24954	8676	3394	1416
(14000, 6000)	0.580	92591	24863	8902	3585	1540
(15000, 5000)	0.625	88888	24691	9145	3810	1694
(16000, 4000)	0.680	84745	24418	9381	4055	1869
(17000, 3000)	0.745	80320	24032	9587	4303	2060
(18000, 2000)	0.820	75757	23531	9745	4540	2256
(19000, 1000)	0.905	71174	22923	9843	4755	2450
真の値	0.500	99999	25000	8333	3125	1250

これらのことから，層化無作為抽出された標本を単純無作為抽出されたと誤解した場合のリスク評価への影響は，層別変数がキー変数の場合の方が大きい可能性があり，層別変数がキー変数でない場合には，相関係数が小さいほど影響が大きい可能性があることが分かる。

表５　$\hat{\beta}^*$の値と母集団寸法指標の推定値（$r=0.4$）

(n_1, n_2)	$\hat{\beta}^*$ $(\times 10^{-5})$	$\hat{S}_1$	$\hat{S}_2$	$\hat{S}_3$	$\hat{S}_4$	$\hat{S}_5$
(10000, 10000)	0.700	83332	24306	9452	4135	1930
(11000, 9000)	0.703	83124	24288	9462	4147	1939
(12000, 8000)	0.712	82507	24235	9491	4182	1965
(13000, 7000)	0.727	81499	24144	9537	4238	2009
(14000, 6000)	0.748	80127	24013	9595	4313	2068
(15000, 5000)	0.775	78430	23837	9660	4404	2141
(16000, 4000)	0.808	76452	23614	9725	4506	2227
(17000, 3000)	0.847	74238	23341	9785	4614	2321
(18000, 2000)	0.892	71838	23017	9833	4726	2423
(19000, 1000)	0.943	69299	22644	9865	4835	2528
真の値	0.700	86666	23889	8951	3850	1804

表６　$\hat{\beta}^*$の値と母集団寸法指標の推定値（$r=0.8$）

(n_1, n_2)	$\hat{\beta}^*$ $(\times 10^{-5})$	$\hat{S}_1$	$\hat{S}_2$	$\hat{S}_3$	$\hat{S}_4$	$\hat{S}_5$
(10000, 10000)	0.900	71428	22959	9840	4744	2440
(11000, 9000)	0.901	71377	22952	9840	4746	2442
(12000, 8000)	0.904	71224	22930	9843	4753	2448
(13000, 7000)	0.909	70972	22893	9846	4764	2459
(14000, 6000)	0.916	70621	22842	9851	4779	2473
(15000, 5000)	0.925	70175	22776	9856	4799	2492
(16000, 4000)	0.936	69637	22695	9862	4821	2514
(17000, 3000)	0.949	69012	22600	9868	4847	2540
(18000, 2000)	0.964	68305	22489	9872	4875	2568
(19000, 1000)	0.981	67521	22363	9875	4906	2600
真の値	0.900	73332	22778	9868	4576	2357

５　おわりに

本論文では，層を構成する際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とに分けて，層化無作為抽出された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提案した。また，層化無作為標本から得られた個票データを，非復元単純無作為標本から得られたと見なすことが，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。

官庁統計では多くの場合，キー変数と考えられる変数によって層別が行われている。層別変数がキー変数の場合には，わずかな手順の追加によって母集団寸法指標の推定が可能である。一方で，非復元単純無作為標本と見なすことがリスク評価に与える影響，特に母集団寸法指標のサイズの小さい部分の推定に与える影響が大きい可能性がある。また，層ごとに性質が異なる場合には，比例配分されていたとしても安心できない。これらの結果は，手順の追加の労力を惜しむことが危険であることを意味している。

層別変数がキー変数でない場合は，官庁統計ではそれほど多くはないと思われるが，リスク評価を正確に行うためには多重寸法指標の導入が不可欠である。例えば労働力調査では，全国を11の地域に分けた後に，それぞれの地域に含まれる調査区の産業・従業上の地位別の就業者構成を基にしてさらに層別を行っている。これらの情報の取得は可能であるものの，11の地域と比較すると個体の特定に利用するには労力を要する。必要とされる労力がさらに増してあるレベルを超えると，層別変数が実質的にキー変数でないと考えるべき場合もあるだろう。

本論文では層の数が２の場合について例示したが，官庁統計の標本調査では一般に層の数が多く，その場合，安定した多重母集団寸法指標の推定を行うのは簡単ではない。非復元単純無作為標本と見なした場合のリスク評価への影響は，層別変数がキー変数の場合と比較して小さい傾向にあるが，相関が十分高い場合を除くとその影響は無視することはできない。この板挟みの状態を解決するための新たな手法の検討が今後の課題と考え

る。

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19300098，22300097），統計数理研究所共同利用研究プログラム（21−共研−2033，22−共研−2031，23−共研−2028）の研究成果に基づくものである。

参考文献

［1］馬場康維，坂口尚文(2006). 複数名簿のマッチングによる共通集合の推定, 2006年度統計関連学会連合大会予稿集, 117−117.

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

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

［4］佐井至道(1993a). サンプリング法の誤解による推定への影響について－第１部母平均推定の場合－，行動計量学，**20** (2)，24−34.

［5］佐井至道(1993b). サンプリング法の誤解による推定への影響について－第２部母分散推定の場合－，行動計量学，**20** (2)，35−47.

［6］佐井至道(2007). 多重寸法指標のノンパラメトリック推定，2007年度統計関連学会連合大会予稿集, 186−186.

［7］佐井至道(2008). 多重寸法指標のノンパラメトリック推定，岡山商大論叢，**43** (3)，1−18.

［8］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.

［9］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.

［10］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.

［11］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.