『岡山商大論叢』（岡山商科大学）
第56巻第３号　2021年３月
Journal of OKAYAMA SHOKA UNIVERSITY
Vol.56　No.３　March 2021

《論　　説》

局所的な分布を用いたリスク評価
－実用化に向けた検討－

佐　井　至　道

A Risk Assessment of the Usage of Distributions in the Small Regions
－A Study for Practical Use－

Shido Sai

１　はじめに

個人のプライバシー情報などが含まれる個票データを安全に公開するために，公開前に秘匿措置を施すことがある．秘匿の対象は，例えば年収や病歴のように，本人が第三者に知られたくないセンシティブ変数の場合と，例えば性別や年齢のように，第三者が既に情報を持っていて個体の識別に用いられるキー変数の場合がある．キー変数は準識別子とも呼ばれる．

センシティブ変数に秘匿措置を施せば，第三者に知られたくない情報を隠すという目的は達成できるものの，個体自体の識別を防ぐことはできない．またセンシティブ変数は分析の対象となることが多く，その情報を秘匿してしまうと分析対象としてのデータの価値を下げてしまうことにな

る．そのため，官庁統計などではキー変数に対して秘匿措置を施すことが多く，本稿でも秘匿はキー変数に対して行うことにする．

秘匿措置は，カテゴリの併合やトップコーディング・ボトムコーディングに代表される非攪乱的方法と，ノイズの挿入やスワッピングに代表される攪乱的方法に大別される．国内では，攪乱的方法を施したデータは嘘をついたデータである，という意識が強かったため，長い間非攪乱的方法が主流であった．

秘匿措置が施されていない元の個票データや非攪乱的な秘匿が施された個票データが，どの程度安全かを計るためのリスク評価の指標としては，少なくとも国内では寸法指標の利用がほとんどである．寸法指標は，すべてのキー変数の値の組み合わせについてそれぞれ個体数をカウントし，同じ組み合わせの個体が１個となる頻度，２個となる頻度，…と頻度分布としてまとめたものである．同じ組み合わせが１個しかない，一意またはユニークと呼ばれる頻度が危険性を表す指標として特に重要である．

個票データが母集団から得られている場合には，そのデータ内において安全性を評価するだけでよいが，個票データが標本から得られている場合には，母集団との関係を考える必要がある．そのため，寸法指標をリスク評価の指標として用いる場合には，標本寸法指標を基に母集団寸法指標を推定することになる．その推定には，初期の研究ではBethlehem *et al*.（1990）が提案した実質的に１個のパラメータしか持たないポアソンガンマモデルが用いられることが多かったが，その後，Pitman（1995）が提案した２個のパラメータを持つピットマンモデルのように，実データへの当てはまりの良いモデルが開発され，安定した推定が可能となった．

非攪乱的方法についてはWillenborg and de Waal（1996，2001）を，寸法指標の利用については佐井（2003，2013）を参照していただきたい．

一方，イギリスやアメリカでは，早くから攪乱的な秘匿方法が用いられてきた．攪乱的方法は個票データに含まれる情報の正確性を犠牲にするものの，非攪乱的方法と比較して，リスクが同程度であっても分析対象の多

くの情報を残すことが可能であることが理由であった．国内でも，非攪乱的な秘匿方法を施すだけで安全かつ有用な個票データを作成するには限界があることが認識されてきて，近年では攪乱的方法を用いることが増えてきている．

攪乱的方法の１つであるノイズの挿入が施された個票データのリスク評価として，いくつかの方法が提案されているが，主要な方法の１つが距離計測型リンケージを用いる方法である．これは，対象となる個体の秘匿後のキー変数の値の組から秘匿前のすべての個体のキー変数の値の組までの距離を計算して，対象となる個体について秘匿後と秘匿前の対応づけが可能かどうかを判断する方法である．距離計測型リンケージについては伊藤（2010）を，攪乱的方法で秘匿された個票データに対する，それ以外のリスク評価方法については伊藤他（2014）を参照していただきたい．

攪乱的な秘匿方法が施された個票データのリスク評価はこれまでも行われていたものの，既存の研究では，個票データが母集団から得られている場合しか想定されていなかった．これは個票データが標本から得られていても，未知である母集団の情報を組み入れるのが難しかったからである．

佐井（2015a，2015b，2016，2017a，2017b，2017c，2018a，2018b）の一連の研究では，複数の量的なキー変数に対してノイズを挿入する秘匿措置を行った場合に，距離計測型リンケージによって，秘匿後の個体が秘匿前の同一の個体にリンクされる確率の期待値を真のリンク確率の期待値と呼び，その指標を用いてリスク評価を行う方法を提案した．個票データが標本調査で得られたことを想定し，ノイズを挿入した個体が，標本のみならず母集団でも真のリンクとなる確率について検討を行ったが，その際，キー変数の分布として，一様分布や正規分布のように定義域全体について１つの分布を設定した．

佐井（2018c，2019a，2019b）では，複数の離散型の量的なキー変数にノイズを挿入した場合について，キー変数の局所的な分布を用いてリスク評価を行う方法を提案した．挿入するノイズは小さいことが多いため，真

のリンク確率を考える場合，キー変数の定義域全体を考える必要はなく，ノイズ挿入後のキー変数が存在する可能性のあるその個体の周辺でのみ議論すれば十分であることと，多くの実データでは，キー変数の値はある場所に集中したり，別のある場所にはほとんど存在しなかったりしており，定義域全体の分布では表現しにくいことがこの提案の背景にある．

また局所的な分布を用いることにより，寸法指標を多次元に拡張した多重寸法指標を導入することができて，非攪乱的な秘匿措置が施された個票データのリスク評価に用いられてきた既存の手法との接続も可能となった．

人工的なデータによる検討結果からは，局所的な分布を用いる方法はキー変数の分布に柔軟に対応できていて，キー変数が定義域全体で一様分布に従うことを仮定する推定法よりも，優れている場合のあることが読み取れた．ただ，この検討ではキー変数は３個と少なく，しかも母集団の大きさが２・10^4という小さい母集団を用い，抽出率も1/2で実際の調査と比較して極めて高かった．官庁統計データへの適用を考えると，キー変数の数を増やすとともに，大きな母集団についての検討が不可欠である．また上記の検討では，各キー変数の取り得る値も100と多く，離散型ではあるが連続型に近い設定であった．

本稿では，佐井（2019c）でも検討したように，母集団の大きさを10^6，抽出率を１/100とする．また，キー変数の数を段階的に10まで増やし，キー変数の取り得る値も段階的に20まで減少させ，その影響についても考える．これらの設定により，現実のデータにより近い状況での検討を行うことができるが，それにより解決しなければならない種々の問題が新たに発生するため，それらの点についての解決を図る．

まず第２章では，ノイズを挿入した個票データのリスク評価方法について，佐井（2018c，2019a，2019b）での研究の流れを簡単に述べる．２.１節では，キー変数へのノイズの挿入方法と母集団における個体のリンクについて，そして本稿でもリスクの指標として用いる真のリンク確率の期待

値について説明する．２.２節では，局所的な分布を用いたリスク評価に使われる領域Ｄ，周辺セルＨの定義とともに，多重寸法指標について説明する．また，推定された多重母集団寸法指標を用いた真のリンク確率の期待値の推定方法についての説明も行う．

第３章では，実用化に向けた種々の検討を行う．まず３.１節では，佐井（2019a）で検討した人工的なデータについての結果を簡単に紹介する．３.２節では，本稿で扱うデータについて紹介した後に，実際に母集団から標本を抽出し，標本の各個体のキー変数にノイズを挿入した場合に，それらが母集団において元の個体にリンクされる比率を，以後の推定目標として求める．

３.３節以降では，実際の状況に近い３.２節で扱うようなデータについて，真のリンク確率の期待値を推定する際に生じる問題の提示とともに，その解決策を提案する．まずこの節では，多重標本寸法指標をすべてカウントするための計算時間の問題の解決を図るが，この問題では，多重標本寸法指標のカウント１つずつについて計算時間を要することも原因の１つになっている．３.４節では，その解決を図るための簡易的な方法を提案する．３.５節においては，多重母集団寸法指標の推定に用いるノンパラメトリック法を基礎とした探索法について，計算時間上の問題の解決を図る．

第４章では，３.２節のデータについて示した推定目標について，提案したこれらの解決策を用いて推定を行い，その有効性を示す．

最後に第５章では，残されている課題についても取り上げ，その一部については解決への道筋を示す．

２　局所的な分布を用いたリスク評価の概要

２.１　真のリンク確率の期待値

この章では，複数のキー変数からなるキー変数ベクトルにノイズを挿入

した場合の，個体のリンクと真のリンク確率の期待値について述べた後，局所的な分布を用いた個票データのリスク評価方法について紹介する．詳細については佐井（2019a）を参照していただきたい．

本稿を通じて，個票データは大きさ n の標本から作られたもので，その標本は大きさ N の母集団から非復元単純無作為抽出されたものとする．個票データに含まれるキー変数の数を K として，すべてのキー変数を離散型の量的変数とする．また，第三者が母集団に含まれるすべての個体のキー変数の値について完全な情報を持っており，標本のすべての個体のすべてのキー変数にノイズを挿入する場合を考える．この条件に合わない場合については佐井（2018a）を参照していただきたい．

標本の i 番目（$i = 1, 2, \ldots, n$）の個体のキー変数ベクトルと，そのベクトルに挿入するノイズ変数ベクトルをそれぞれ $\mathbf{x}_i$，$\mathbf{e}_i$ とする．キー変数ベクトルの要素は

$$\mathbf{x}_i \quad = \quad (x_{i,1}, x_{i,2}, \ldots, x_{i,K})' \tag{1}$$

であり，ノイズ変数ベクトルも同様の K 次のベクトルである．公開する個票データでは，キー変数ベクトルの値をノイズを含めて $\mathbf{x}_i + \mathbf{e}_i$ とする．また，母集団の i 番目（$i = 1, 2, \ldots, N$）の個体のキー変数ベクトルを $\mathbf{a}_i$ とする．なお，標本の i 番目の個体が母集団では i' 番目の個体であるとする．

2つの個体のキー変数ベクトル間の距離を $\mathrm{d}(\cdot, \cdot)$ と表す．このとき

$$\mathrm{d}(\mathbf{x}_i + \mathbf{e}_i, \mathbf{a}_{i''}) \leq \mathrm{d}(\mathbf{x}_i + \mathbf{e}_i, \mathbf{x}_i) \tag{2}$$

となる母集団のキー変数ベクトル $\mathbf{a}_{i''}$ $(i'' \neq i')$ が少なくとも 1 個存在すれば，キー変数ベクトルにノイズ変数ベクトルを挿入した標本の i 番目の個体について，母集団内で間違ったリンクが発生し，そのような $\mathbf{a}_{i''}$ が存在しなければ，母集団内で真のリンクが発生したと考える．これが佐井（2016）で提案した真のリンクであり，従来の個票データ内の距離計測型

リンケージを母集団に拡張したものである．従来の距離計測型リンケージについては伊藤（2010，2014）を参照していただきたい．

$\mathbf{x}_i$，$\mathbf{e}_i$が得られたとき，$\mathbf{a}_{i''}$がその領域に入ると間違ったリンクとなる範囲を領域Dと呼ぶことにするが，離散型の量的変数の場合には格子点の集合になる．

図１に，$K=2$で$x_{i,k}$が値１，２，...，10をとりうる場合のイメージを示す．図中の100個のセルは２つのキー変数の値の組み合わせの格子点を表している．以後の図でも格子点の代わりにセルを用いる．またセルにはキー変数ベクトルが入るが，「個体が入る」と表現することがある．

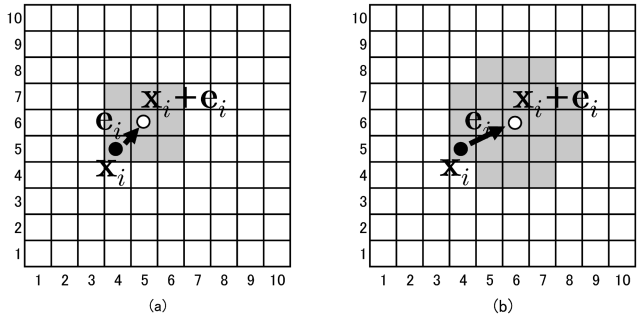

図１　ノイズ変数ベクトルの挿入と領域Dのイメージ（K=2）

図１の（a）では，キー変数ベクトル$\mathbf{x}_i$=（４，５）にノイズ変数ベクトル$\mathbf{e}_i$=（１，１）を加えた場合の領域Dを網掛けで表しており，この場合は３×３の正方形となる．母集団のキー変数ベクトル$\mathbf{a}_{i''}$が正方形の４つの隅のいずれかに存在する場合には（2）式の等号が成立するが，このような場合でも間違ったリンクが発生したと考える．

図１の（b）では，キー変数ベクトル$\mathbf{x}_i$=（４，５）にノイズ変数ベクトルとして$\mathbf{e}_i$=（２，１）を加えた場合の領域Dを網掛けで表している．領域はやや丸みを帯びている．

次にリスクの指標である真のリンク確率の期待値について説明する．

標本の i 番目の個体のキー変数ベクトル $\mathbf{x}_i$ が与えられ，そのキー変数ベクトルに確率変数であるノイズ変数ベクトル $\mathbf{e}_i$ を挿入したとき，キー変数ベクトル $\mathbf{a}_{i''}$ を持つ母集団の特定の個体が間違ったリンクの 1 つとなる確率，すなわち領域 D に特定の 1 個の個体が入る確率を $p_f(\mathbf{x}_i, \mathbf{a}_{i''})$ と表す．ここでの確率変数は $\mathbf{e}_i$ のみである．

このとき $\mathbf{x}_i$ に対応する母集団の $\mathbf{a}_{i'}$ 以外の $N-1$ 個のキー変数ベクトルが間違ったリンクとならない確率，すなわち $\mathbf{x}_i$ が真のリンクとなる確率は

$$P_t(\mathbf{x}_i) = \prod_{i''\,(i''\neq i')} \{1 - p_f(\mathbf{x}_i, \mathbf{a}_{i''})\} \tag{3}$$

と書くことができる．この値が大きいほどリスクは大きいと考えられる．

さらに $\mathbf{x}_i$ について

$$\mathrm{E}\big[P_t(\mathbf{x}_i)\big] = \frac{1}{n}\sum_{i=1}^{n} P_t(\mathbf{x}_i) \tag{4}$$

と期待値をとり，標本の各個体の平均的な真のリンク確率をリスクの指標とする．一部省略して $\mathrm{E}(P_t)$ と表すことがあり，「真のリンク確率の期待値」と呼ぶ．

なお，(3)，(4) 式では母集団の N 個の個体が与えられた下で定義をしているが，母集団のキー変数ベクトル $\mathbf{a}_{i''}$ に分布を入れて確率的な議論をすることがある．

2.2 局所的な分布を用いたリスク評価

この節では，$K = 2$ のときにキー変数ベクトル $\mathbf{x}_i = (4, 5)$ にノイズ変数ベクトルを挿入する場合を例にしながら，局所的な分布を用いたリスク評価方法の概略を説明する．なおノイズ変数 $e_{i,k}$ は，i，k について独立に ± 1 の値を確率 1/2 ずつでとることにする．そのため，

$\mathbf{e}_i = (-1, -1), (-1, 1), (1, -1), (1, 1)$ を確率1/4ずつでとることになる．

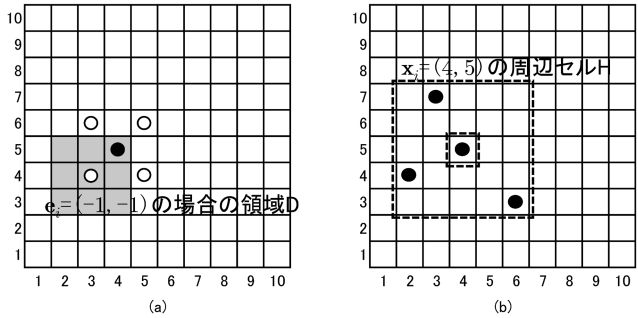

図2　領域Dと周辺セルH

図2の（a）ではキー変数ベクトル $\mathbf{x}_i = (4, 5)$ を黒い点で，ノイズ変数ベクトルを挿入したキー変数ベクトル $\mathbf{x}_i + \mathbf{e}_i$ として可能性のある4つを白抜きの点で示している．$\mathbf{e}_i = (-1, -1)$ を挿入した場合，領域Dは図の網掛けの部分となる．$\mathbf{e}_i = (-1, 1), (1, -1), (1, 1)$ の場合についての領域Dも合わせて，4つの領域Dの和集合を周辺セルHと呼ぶことにする．

図2の（b）では，その周辺セルHを大きく点線で囲んでいるが，元のキー変数ベクトルの値が含まれている点線で小さく囲んだセル(4, 5)は，周辺セルに含めないものとする．周辺セルとの対比で，このセルを「対象となる中央セル」，あるいは単に「中央セル」と呼ぶ．なお，ノイズ変数が±1の値だけをとる場合，領域Dは，$K = 2$ ではこの図にように正方形で，$K = 3$ では立方体となるが，$K \geq 4$ の場合には超立方体を膨らませたような丸みを帯びた形状となる．周辺セルHについても，$K = 2$ では正方形，$K = 3$ では立方体となり，$K \geq 4$ では丸みを帯びた形状の領域Dの和集合となるため複雑な形状となる．

セルに含まれる個体数をサイズと呼ぶ．（4, 5）の中央セルには1個の個体が含まれておりサイズ1で，その周辺セルHには3個の個体が含まれ

サイズ3となっており，そのサイズの組を（1，3）と書くとき，標本におけるそのような組の数を$s_{(1,3)}$と表す．

一般に標本において，対象となる中央セルのサイズがl（$l=0,1,\ldots,n$），その周辺セルHのサイズがh（$h=0,1,\ldots,n$）の組（l，h）の数を$s_{(l,h)}$と表し，多重標本寸法指標と呼ぶ．母集団でも同様に，中央セルのサイズがl（$l=0,1,\ldots,N$），その周辺セルHのサイズがh（$h=0,1,\ldots,N$）の組（l，h）の数を$S_{(l,h)}$と表し，多重母集団寸法指標と呼ぶ．

サイズの組には，中央セルのサイズlが0の場合，その周辺セルHのサイズhが0の場合も含まれる．

対象となる中央セルだけに着目したとき，サイズがlとなるセル数を，標本ではs_lと表して標本寸法指標と呼び，母集団ではS_lと表して母集団寸法指標と呼ぶ．非攪乱的な秘匿方法が施された個票データでは，標本寸法指標を基に母集団寸法指標を推定するのが，リスク評価方法の主流となっている．

同様に，多重標本寸法指標を基に多重母集団寸法指標を推定できれば，対象となる個体が入る中央セルの近くにおける母集団での個体の集散状況が推測できるため，真のリンク確率の期待値$\mathrm{E}(P_t)$を推定することが可能となる．

情報を寸法指標まで縮約すると，同じセルに含まれる母集団と標本の個体数の対応づけができなくなるため，寸法指標を利用したリスク評価には推定された母集団寸法指標のみを用いるのが一般的である．ここでも多重標本寸法指標は多重母集団寸法指標の推定のみに用い，真のリンク確率の期待値の推定は，推定された多重母集団寸法指標の情報のみに基づいて行うことにする．

すべての（l，h）について多重母集団寸法指標の推定値$\hat{S}_{(l,h)}$が得られたとき，母集団で対象となる中央セルのサイズが1，周辺セルHのサイズがhとなる中央セルと周辺セルの組のうちの1つを考える．中央セルのサ

イズが母集団で２以上であれば，そのセルに含まれる個体にどのようなノイズを挿入しても，同じセルの別の個体が間違ったリンクとなるため，サイズ２以上のセルに含まれる個体が真のリンクとなることはない．そのため，真のリンク確率の期待値を求める際には，母集団で対象となる中央セルのサイズが１のものが特に重要である．

ここで周辺セルH内で母集団のキー変数ベクトルが一様に分布すると仮定する．次章以降では，キー変数ベクトルの分布が一様でないデータも扱うものの，周辺セルHは小さいため，その中で一様に分布するとみなすことには大きな問題はないかもしれないが，強い相関構造がある場合などには対処が必要と考えられる．

キー変数ベクトルが $\mathbf{a}_i$ であるサイズ１の中央セルの個体が，母集団で真のリンクとなる確率は

$$P_t(\mathbf{a}_i) = \left(1 - \frac{\#(\mathrm{D}) - 1}{\#(\mathrm{H})}\right)^h \tag{5}$$

と求められる．ただし＃(D)は領域Dのセル数を表し，＃(H)は周辺セルHのセル数を表すため，＃(D)－１は領域Dから中央セルを除いたセル数を意味する．挿入するノイズによって＃(D)が変動するときには，期待値をとることで対処できるが，その場合＃(H)もまた変動するため多重標本寸法指標のカウントや多重母集団寸法指標の推定にも影響する．この点については第５章で簡単に触れる．

前述したように，対象となる中央セルのサイズが２以上の場合は，そのセルに含まれるすべての個体について $P_t(\mathbf{a}_i) = 0$ である．

母集団でサイズが１の中央セルに含まれる個体のうちで，標本として抽出されたものが真のリンクとなる可能性がある．標本のリスクについて議論する場合には，前述のように標本としての観測値を無視して，サイズ１のセルの個体を含めて母集団からの非復元単純無作為抽出についての期待値を考えることになる．抽出される標本における真のリンク確率の期待値

は，母集団における真のリンク確率の期待値そのものを用いて，

$$\hat{\mathrm{E}}\left[P_t(\mathbf{x}_i)\right] = \mathrm{E}\left[P_t(\mathbf{a}_i)\right] = \frac{1}{N}\sum_{i=1}^{N} P_t(\mathbf{a}_i) \tag{6}$$

と推定することができる．

3　実データに近い環境でのリスク評価

3.1　小母集団での検討結果

この章では，提案している個票データのリスク評価方法を官庁統計などの実データに適用することを目的として，それに近い環境を想定して作られた人工的なデータを用いて検討を行い，リスク評価方法の有効性とともに，問題点と解決策について考えるが，その前に佐井（2019a）における検討結果について説明する．

佐井（2019a）では２種類の人工データ（A），（B）を用いて検討を行った．いずれもキー変数の個数は$K = 3$で，各キー変数は１から100までの整数値をとる．これを$M = 100$と表すことにする．母集団の大きさは$N = 2 \cdot 10^4$，標本の大きさは$n = 10^4$で，抽出率１/２であった．（A）は各キー変数が独立な離散型一様分布に従うデータであり，（B）は10毎の周期で確率関数が増減するデータであった．（B）については後に詳しく説明する．

（A），（B）それぞれについて母集団を作成し，それぞれ標本を抽出する．標本の各キー変数に±１のノイズを確率１/２ずつで独立に挿入して，これを公開用の個票データとする．そのデータの各個体について，母集団において真のリンクとなるか間違ったリンクとなるかを１個ずつ確認したところ，真のリンクとなった比率は（A）が0.6033，（B）が0.4890であった．

次に，±１のノイズを確率変数と考えて，佐井（2018a）で提案したキー変数が定義域全体で一様分布に従うことを想定した方法を用いると，真

のリンク確率の期待値は（A）も（B）も0.5828と推定され，（B）においては真のリンクとなった比率との乖離が見られた．これに対して，佐井（2019a）で提案したキー変数の局所的な分布を用いる方法では真のリンク確率の期待値は（A）が0.5862，（B）が0.4710と推定され，キー変数ベクトルの分布によらず安定した推定結果が得られた．

しかし，例えば官庁統計では，母集団の大きさとして日本の全人口や15歳以上人口などを考えることが多く，仮に母集団を都道府県別に分けたとしても母集団の大きさとして$N=10^6$程度を考える必要がある．また上記の検討で用いた抽出率1/2は一般的な調査と比較してかなり高く，より低い抽出率でも安定した推定が可能か検討する必要がある．

各キー変数の取り得る値として上の検討では$M=100$とした．例えば1歳刻みの年齢などが該当するものの，取り得る値がこれほど多い離散型の変数が個票データに多数含まれる可能性は低いため，より小さいMについても検討する必要がある．さらに，実際の個票データではキー変数の個数が10を超えることも珍しくなく，$K=3$より大幅に増やした場合の影響についても考えなければならない．

3.2　大母集団での検討方法

前節での検討内容を踏まえて，新たに次のような検討を行う．

前節と同様に2種類の人工データ（A），（B）を用いる．いずれの分布でもキー変数の数として$K=1,2,\ldots,10$の10通りを考え，それぞれの場合についてキー変数の取り得る値の数として$M=100,50,20$の3通りを考える．なお，キー変数は1からMまでの整数値をとるものとする．

また母集団の大きさを$N=10^6$，標本の大きさを$n=10^4$として，標本は非復元単純無作為抽出されているものとする．抽出率は$\lambda=1/100$となる．

データ（A）は母集団の各キー変数$a_{i,k}$がi，kについて独立に離散型一様分布に従うデータとする．一様分布の確率関数$f(a_{i,k})=1/M$を用

いて，K個ずつのN組の乱数を独立に発生させ，これを母集団のキー変数ベクトルとする．（A）を簡単に一様分布と呼ぶ．

データ（B）も各キー変数$a_{i,k}$はi，kについて独立で，確率関数は$a_{i,k}$=1,2,3,4,5のとき$f(a_{i,k})$=$a_{i,k}$/(3M)と単調増加で，$a_{i,k}$=6,7,8,9,10のとき$f(a_{i,k})$=$(11-a_{i,k})$/(3M）と単調減少し，以後同様に10毎にM/10回の周期がある分布に従うデータとする．この確率関数を用いてK個ずつのN組の乱数を独立に発生させ，これを母集団のキー変数ベクトルとする．（B）を偏在分布と呼ぶことにする．

２つの母集団から，それぞれ10^4個の個体を実際に非復元単純無作為抽出して標本とする．ここでの標本抽出は１回のみである．次に，標本の各キー変数に±１のノイズを確率１/２ずつで独立に挿入して，これを公開用の個票データと考える．

この公開用のデータの各個体について，母集団で真のリンクとなるか間違ったリンクとなるかを１個ずつ確認したところ，真のリンクとなった比率は表１のようになった．

表１　真のリンクとなった個体の比率

(A) 一様分布

K \ M	100	50	20
1	0.0000	0.0000	0.0000
2	0.0000	0.0000	0.0000
3	0.0000	0.0000	0.0000
4	0.4240	0.0000	0.0000
5	0.9679	0.3555	0.0000
6	0.9985	0.9189	0.0000
7	1.0000	0.9922	0.0348
8	1.0000	0.9995	0.5009
9	1.0000	1.0000	0.8711
10	1.0000	1.0000	0.9738

(B) 偏在分布

K \ M	100	50	20
1	0.0000	0.0000	0.0000
2	0.0000	0.0000	0.0000
3	0.0005	0.0000	0.0000
4	0.3192	0.0043	0.0000
5	0.9326	0.2578	0.0000
6	0.9967	0.8271	0.0034
7	0.9997	0.9803	0.0574
8	1.0000	0.9978	0.2868
9	1.0000	0.9998	0.6518
10	1.0000	1.0000	0.8910

表中で0.0000と表示されているものは標本のすべての個体が母集団で間違ったリンクとなっており，1.0000と表示されているものは標本のすべての個体が母集団で真のリンクとなっていることを意味する．キー変数の個数Kが増加する方向に比率の変化を見ると，しばらく0だった値は急激に1に近づいており，比率が0にも1にも近くないようなKとMの値の組み合わせ，すなわち真のリンクと間違ったリンクの個体が混在するような場合は，それほど多くないことが分かる．その傾向は一様分布の方が強い．

0.02 ～ 0.98の値にアンダーラインを引いているが，これらを本稿での推定目標とする．

佐井（2019a）で提案したリスク評価方法は，次の3つのステップで真のリンク確率の期待値を推定するものであった．

Step 1　多重標本寸法指標$s_{(l,h)}$をカウントする
Step 2　$s_{(l,h)}$を基に多重母集団寸法指標の推定値$\hat{S}_{(l,h)}$を求める
Step 3　$\hat{S}_{(l,h)}$を用いて真のリンク確率の期待値$\mathrm{E}(P_t)$を推定する

本節で設定したような現実のデータに近い状況では，Step 1において次の3つの問題が生じる．

(1) 多重標本寸法指標$s_{(l,h)}$をすべてカウントするのに時間を要する
(2) 多重標本寸法指標$s_{(l,h)}$の1つのカウントに時間を要する
(3) キー変数の定義域の端においてカウントに工夫が必要となる

またStep 2においては次の問題が生じる．

(4) 多重母集団寸法指標の尤度の計算に時間を要する

これら４つの問題のうちで（3）以外について，次節以降で詳しく説明するとともに，その解決方法を探る．具体的には，(1)，(2)，(4) をそれぞれ３.３節，３.４節，３.５節で扱う．なお（3）については，本稿では第５章において問題提起して，対応策の概略を述べるにとどめる．

３.３　多重標本寸法指標のすべてのカウントの高速化

ここでは前節で挙げた（1）の問題について考える．

３.５節で説明する多重母集団寸法指標の推定では，多重標本寸法指標として$s_{(0,0)}$をカウントする必要はないものの，一方のサイズが０のもの，すなわち $s_{(l,0)}$（$l = 1, 2, \ldots, n$），$s_{(0,h)}$（$h = 1, 2, \ldots, n$）をカウントしなければならない．

最も簡単なカウント法は，キー変数ベクトルの定義域内のすべてのセルを中央セルとして設定して，$s_{(0,0)}$を含めてカウントする方法である．M^K通りの中央セルについて，中央セルまたは周辺セルHの{#（H）+ 1}個のセルに，標本のn個の個体が含まれるか否かをチェックすればよい．最後の部分については，{#（H）+ 1}個のセル１個ずつに標本の特定の個体が含まれるかをチェックするよりも，標本の特定の個体が2^K個の領域Dのいずれかに含まれるかどうかを距離で判断した方が計算は速い．図２を例にすると，（b）の{#（H）+ 1} = 25通りに対して，（a）のような領域Dは$M^K = 2^2 = 4$通りである．

しかし，それでも$n \cdot (2M)^K$回の計算が必要になり，Intel Core i7-6700 2.60GHz のパーソナルコンピュータにおいて，$K = 4$，$M = 100$の場合は数時間程度の計算で済むものの，$K = 10$，$M = 20$の場合は理論上，計算に数十年かかることになる．（以後，計算時間としてはこのコンピュータで計算した場合の概算値を示す．）

そこで佐井（2019a）で提案した方法を改善した次のような方法を考える．ここでも，各ノイズ変数$e_{i,k}$は±１の値をとるものとする．図３では$K = 2$の場合を例に説明するが，周辺セルHは５×５の領域から中央

セルを除いたものとなる.

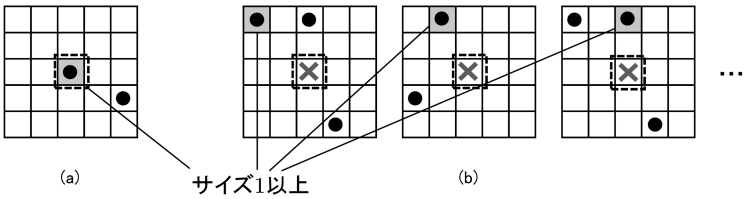

図３　多重標本寸法指標のカウント法

まず図３の（a）のように，対象となる中央セルがサイズ１以上となるものについて多重標本寸法指標$s_{(l,h)}$（$l \geq 1$）をカウントする．標本のn個の個体を順に中央セルに配置したとき，標本の残りの$n-1$個の個体について中央セルまたは周辺セルＨに含まれるか否か2^K回のチェックを行うことになる．$n(n-1)\cdot 2^K$回の計算が必要になるが，$K=10$でも数秒程度でカウントすることができる.

次に対象となる中央セルがサイズ０となるものについて多重標本寸法指標をカウントする．$s_{(0,0)}$をカウントする必要がないため周辺セルの少なくとも１つのセルに個体が含まれるものだけをカウントすればよい.

周辺セルＨの＃(H)個のセルに，中央セルとの相対的位置で１から＃(H)の番号をつける．図３の（b）で網掛けしているセルが左の図から相対的位置が１番目，２番目，３番目とする．まず１番目のセルのサイズが１以上のものをカウントする．標本のn個の個体を順にこのセルに配置したとき，標本の残りの$n-1$個の個体が周辺セルＨに含まれるかをチェックする．これには$n(n-1)\cdot 2^K$回の計算が必要になる．同様に２番目のセルのサイズが１以上のものをカウントし，以後＃(H)番目のセルのサイズが１以上のものまでカウントを行う.

図３の（b）で左から１番目と３番目の図は中央セルとの相対的位置で同じセルに個体が入っているが，中央セルが同じ場合と異なる場合がある．このうち同じ中央セルの場合を考えると，例えば$h=3$であれば，全く同

じ個体の組み合わせによる多重標本寸法指標 $s_{(0,3)}$ が重複して３回カウントされることになる．これらについては仮に求めた多重標本寸法指標を最終的に hで割って調整する．佐井（2019a）では，例えば３番目のセルのサイズが１以上のものをカウントする場合には，１番目，２番目のサイズが０であるものだけを考えることにより重複を避けていたが，そのために標本のすべての個体について，周辺セルＨの特定のセルに含まれるか否かをチェックする必要が生じるという短所があった．

ただし本稿で改善した方法でも，＃(Ｈ)・$n(n-1)\cdot 2^{K}$回の計算が必要となり$K=4$，$M=100$では28分程度で済むものの，$K=10$，$M=20$では理論上，計算に1387日程度かかる．

この問題を解決するために，多重標本寸法指標のうち$s_{(0,h)}$（$h \geq 1$）のカウントにおいて，中央セルとの相対的位置の１～＃(Ｈ)について，抽出を行うことにする．

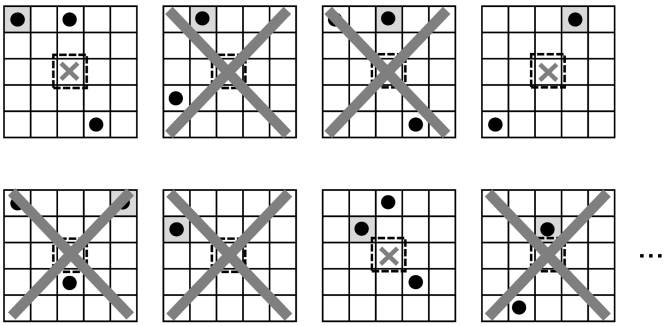

図４　中央セルとの相対的位置に関する系統抽出のイメージ

図４には，相対的位置が１番目，４番目，７番目と３毎に抽出する系統抽出のイメージを示す．×印がつけられたものについては多重標本寸法指標をカウントせず，３毎の抽出であれば，得られた多重標本寸法指標の頻度を最終的に３倍して$s_{(0,h)}$とする．後の例では計算時間を数時間程度に抑えるために11毎，101毎，1001毎など様々な抽出間隔を用いているが，

例えば$K = 5$，$M = 50$の一様分布の推定に系統抽出を用いた場合の結果は表２のようになる．

表２：中央セルとの相対的位置に関する系統抽出の影響

（一様分布，$K = 5$，$M = 50$）

	1 毎	11 毎	101 毎	真の比率
$\hat{\mathrm{E}}(P_t)$	0.3783	0.3779	0.3775	0.3555

「真の比率」は標本の個体のうち母集団で実際に真のリンクとなった比率で，表１でも示した推定の目標値である．抽出間隔として「１毎」，「11毎」，「101毎」の場合の推定値を示すが，「１毎」はすべての相対的位置についてカウントした場合の結果である．系統抽出の間隔を大きくしても推定値は大きくは変わらず，抽出による影響はそれほど大きくないと考えられる．

例えば「3001毎」の系統抽出をすることによって，多少の誤差はあるものの，$K = 10$でも計算を11時間程度に縮めることができ，さらに大きなKについても推定が可能となる．

３.４　多重標本寸法指標の１つずつのカウントの高速化

前節では，多重標本寸法指標のカウントを高速化する方法について考えたが，カウントに時間がかかる要因の１つは，多重標本寸法指標の１つずつについて計算に時間を要する点である．

前述したように，±１のノイズを挿入する場合，周辺セルHは$K = 2$では正方形，$K = 3$では立方体であるが，$K \geq 4$では丸みを帯びた形状の領域Dの和集合となり，複雑な形状となる．そのため，標本の特定の個体が周辺セルHに含まれるか否かを判断する際には，領域Dの１つずつに含まれるか否かをチェックするため2^K回の計算が必要になった．

もし領域Dが超立方体であれば周辺セルHも超立方体となり，標本の特

定の個体のK個のキー変数の値それぞれについて，超立方体の端を表す上限と下限の間に入っているか否かをチェックすればよい．計算は$2K$回で済み，N=10の場合には1/50以下の回数となる．

この節では，周辺セルHに含まれる超立方体Hcを周辺セルHの代わりに用いる方法を提案し，計算時間の短縮を図る．

表3：超立方体Hcと周辺セルHのセル数

K	#(Dc)	#(D)
1	3	3
2	9	9
3	27	27
4	81	89
5	243	333
6	729	1,341
7	2,187	5,449
8	6,561	21,697
9	19,683	84,663
10	59,049	327,829

K	#(Hc)	#(H)
1	4	4
2	24	24
3	124	124
4	624	688
5	3,124	4,244
6	15,624	27,528
7	78,124	177,804
8	390,624	1,122,912
9	1,953,124	6,983,332
10	9,765,624	43,424,504

表3は，キー変数ベクトルのK個の変数に±1のノイズを挿入する場合の，領域Dに含まれる超立方体Dc，領域D，超立方体Hc，周辺セルHのセル数を示している．＃(Dc)と＃(Hc)はそれぞれの超立方体のセル数を表すが，超立方体Dcは1辺が3で3^K個のセルからなり，超立方体Hcは1辺が5で中央セルを除いて5^K-1個のセルからなる．なお周辺セルHに完全に含まれる超立方体としてはこのHcが最大である．$K \leq 3$では，HcとHのセル数は等しいが，$K=4$では少しの差が生じ，K=10では周辺セルHのセル数が4倍以上となっている．

真のリンク確率の期待値の推定では，3.2節で述べた3つのステップすべてに変更を加えることになる．

Step 1では，多重標本寸法使用をカウントする際に，周辺セルHに代えて超立方体Hcを用いるのが唯一の変更である．前述した計算の高速化が図られ，$K=4$，M=100では22分程度で大きな変化はないものの，K=10，M=20では理論上，計算は217日程度まで短縮され，1387日程

度かかる周辺セルHを用いる方法の1/6以下になる．計算回数が1/50以下になっていることと比較して減少率が小さいのは，周辺セルHを用いる方法でも，まず標本の特定の個体が超立方体Hcに含まれているか否かを事前にチェックして，含まれていない場合には，領域Dの1つずつに含まれるか否かをチェックするなどの計算時間を短縮する工夫をしていたためである．

なお，周辺セルHを用いる場合，標本の各個体が周辺セルHに含まれるか否かは2^K回まで計算回数を抑えられたが，最初の個体が周辺セルHのどのセルに含まれるかを決める必要があり，中央セルとどのような相対的位置にあるセルが周辺セルHのセルであるかを事前に調べておく必要がある．K=10では，工夫を施してもこの計算に数時間程度を要し，さらにKが大きくなった場合にはすべて求めることが困難になる．その場合には，相対的位置の事前調査の段階で，前節で述べた系統抽出を予め行うなどの対応が必要となろう．

表4にはK=5，M=50の一様分布について，周辺セルHと超立方体Hcを用いてそれぞれ求めた多重標本寸法指標を示す．超立方体を用いるとセル数が少なくなるため，hについて縮小傾向が認められる．

Step 2では，表4の（b）のように求められた多重標本寸指標を用いて

表4：多重標本寸法指標（一様分布，K=5，M=50）

(a) 周辺セル H を用いた場合

		h					
		0	1	2	3	4	5
l	0	-	33005478	2009511	78310	2030	14
	1	8849	1100	51	0	0	0

(b) 超立方体 Hc を用いた場合

		h					
		0	1	2	3	4	5
l	0	-	25268219	1131678	32158	558	1
	1	9167	802	31	0	0	0

多重母集団寸法指標を推定するが，その方法については，周辺セルHを用いる場合も含めて次節で説明する．

Step 3では推定された多重母集団寸法指標を基にして，真のリンク確率の期待値を推定することになる．

まず周辺セルHを用いる場合には，2.2節で説明した方法を基礎として，

$$\hat{\mathrm{E}}\big[P_t(\mathbf{x}_i)\big] = \mathrm{E}\big[P_t(\mathbf{a}_i)\big] = \frac{1}{N}\sum_{h=0}^{N} \hat{S}_{(1,h)}\Big(1 - \frac{\#(\mathrm{D})-1}{\#(\mathrm{H})}\Big)^h \quad (7)$$

と推定することになる．

超立方体Hcを用いた場合に（7）式の右辺の和をとる部分を

$$\hat{S}_{(1,h)\mathrm{c}}\Big(1 - \frac{\#(\mathrm{Dc})-1}{\#(\mathrm{Hc})}\Big)^h \quad (8)$$

のように，Dc，Hcのセル数を用いた式で置き換えてしまうと過大推定となってしまうため，

$$\hat{\mathrm{E}}\big[P_t(\mathbf{x}_i)\big] = \mathrm{E}\big[P_t(\mathbf{a}_i)\big] = \frac{1}{N}\sum_{h=0}^{N} \hat{S}_{(1,h)\mathrm{c}}\Big(1 - \frac{\#(\mathrm{D})-1}{\#(\mathrm{Hc})}\Big)^h \quad (9)$$

とする．ここで$\hat{S}_{(1,h)\mathrm{c}}$は超立方体Hcを用いて推定された多重母集団寸法指標を表す．すなわち超立方体Hcのセル数は用いるものの，超立方体Dcのセル数ではなく領域Dのセル数を最後の段階で用いることにより補正を行うことになる．

図5は補正のイメージである．$K = 2$では，前述したように実際は領域D，周辺セルHともに正方形になるが，ここでは$K \geq 4$の状況を2次元的に表現しているものと考えていただきたい．

（a）は領域Dと周辺セルHを，（b）は超立方体Dcと超立方体Hcを，（c）は領域Dと超立方体Hcをそれぞれ表現している．ここで提案している

方法は，Step 1，Step 2では（b）のように超立方体で代替するものの，Step 3では（c）のように領域Dだけを元に戻すことになる．（c）では（a）と比べて周辺セルが小さいが，その分hも縮小しているため個体の密度は調整されている．また領域Dの一部が超立方体Hcの外にはみ出ることがあるが，キー変数ベクトルが周辺セルHの付近で一様に分布している場合，推定には大きな影響は与えないと考えられる．

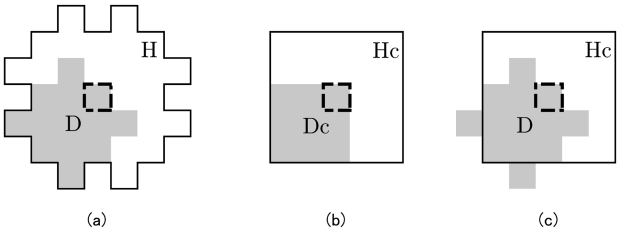

図５　超立方体Dc，Hcを用いた推定のイメージ

表５には，$K=5$，$M=50$の一様分布について，上記の補正を行った場合の真のリンク確率の期待値の推定値を，他のいくつかの値とともに示している．「H利用」は周辺セルHを用いた推定値を，「Hc利用」は超立方体Hcを用いた推定値を，「Hc利用＋補正」はこの節で提案した方法による推定値をそれぞれ表している．超立方体Hcを用いるとかなりの過大推定となるが，その後に補正を行った推定値は周辺セルHを利用した推定値と大きな差はない．

他のケースについての結果は第４章で説明する．

表５：超立方体Hcを用いた後に補正した場合の推定値

（一様分布，$K=5$，$M=50$）

	H 利用	Hc 利用	Hc 利用+補正	真の比率
$\hat{\mathrm{E}}(P_t)$	0.3783	0.5024	0.3850	0.3555

3.5　多重母集団寸法指標の尤度の計算の改善

多重標本寸法指標$s_{(l,h)}$を基にして多重母集団寸法指標$S_{(l,h)}$を推定するための方法として，パラメトリックな方法とノンパラメトリックな方法がそれぞれ提案されている．

パラメトリックな推定方法としては，渋谷，佐井（2007），Sibuya and Sai（2008）が提案した，母集団寸法指標の推定に用いられているピットマンモデルを多重寸法指標に拡張した多重ピットマンモデルを用いる方法がある．しかしこのモデルは，lとhについて対称，すなわち$S_{(l,h)} = S_{(h,l)}$を想定したものであり，本稿で議論している非対称なデータへの適用は難しい．

そのため本稿では佐井（2019a）で用いたノンパラメトリック最尤推定法を用いることにする．ここでは理論の詳細は省略して，近似された尤度関数を示すのにとどめる．詳細については上記論文の他，佐井（2008, 2009）も参照していただきたい．

（l, h）=（0, 0）を除くサイズの組について，多重標本寸法指標を並べたベクトルと多重母集団寸法指標を並べたベクトルを

$$\boldsymbol{s} = (s_{(0,1)}, s_{(0,2)}, \ldots, s_{(1,0)}, s_{(1,1)}, \ldots), \tag{10}$$

$$\boldsymbol{S} = (S_{(0,1)}, S_{(0,2)}, \ldots, S_{(1,0)}, S_{(1,1)}, \ldots) \tag{11}$$

と表す．

大きさNの母集団から大きさnの標本が非復元単純無作為抽出されたとき，抽出率$\lambda = n/N$が十分小さい場合，多重標本寸法指標の組sが得られたときの多重母集団寸法指標の組Sの尤度関数は次のようにポアソン分布の確率関数で近似される．

$$\mathrm{L_{ap}}(\ \boldsymbol{S} \mid \boldsymbol{s}\ ) = \frac{1}{\binom{N}{n}\lambda^n(1-\lambda)^{N-n} \cdot \binom{m/\lambda}{m}\lambda^m(1-\lambda)^{m/\lambda-m}} \cdot \prod_{(l',h')} \frac{e^{-\mu_{(l',h')}}\mu_{(l',h')}^{s_{(l',h')}}}{s_{(l',h')}!}. \tag{12}$$

ただしm＝＃（H）・nは標本における周辺セルHのサイズの合計で，

$$\mu_{(l',h')} = \sum_{(l,h)(\geq(l',h'))} \left\{ S_{(l,h)} \cdot \binom{l}{l'} \lambda^{l'} (1-\lambda)^{l-l'} \cdot \binom{h}{h'} \lambda^{h'} (1-\lambda)^{h-h'} \right\} \quad (13)$$

である．なお，（12）式の積において，標本におけるサイズの組み合わせ（l', h'）は（0，0）を含まない．

この尤度関数を用いた推定は非常に不安定で負の推定値が含まれることが多いため，多重母集団寸法指標の推定値に次のような制約条件を課し，探索法によって近似的な解を求める．

（1）　$\hat{S}_{(l,h)} \geq 0 \quad (l = 0, 1, \ldots, N,\ h = 0, 1, \ldots, N)$

（2）　$\hat{S}_{(l,h)} \leq \hat{S}_{(l-1,h)} \quad (l = 1, \ldots, N,\ h = 0, 1, \ldots, N)$

（1）は多重母集団寸法指標の推定値が非負であることを，（2）はhを固定したとき，多重母集団寸法指標の推定値がlに関して単調減少であることを意味する．

探索法としては，$S_{(l,h)}$の適当な初期値から始め，各ステップで（12）式を各$S_{(l,h)}$で偏微分して，それらの値を基に（12）式の値が増加する方向に移動していく勾配法を用いる．各ステップでの移動距離については，短めから始めて，尤度が順調に増加する場合には距離を長く，尤度が増加しなくなった場合には短く設定して，移動距離が下限を下回った時点で探索を終了して，そのときの値を整数値に四捨五入して推定値とする．

本稿で扱っているデータで，標本における周辺セルHのサイズhの最大サイズを見てみると，例えばK＝7，8，9，M＝20の偏在分布では，いずれもl＝0のときが最も大きく，それぞれ27，17，10である．ここで問題となるのは（13）式における母集団のl，hの最大サイズ$l_{\max}$，$h_{\max}$の設定，特に後者の設定である．定義では$h_{\max}$＝Nとしているが，探索法の各ステップにおいて（13）式を更新しなければならず，その都度，各（l', h'）のすべての組み合わせに対して，それ以上の値のすべての組

(l, h)についての計算が必要となるため，$h_{\max}$をある程度小さく設定しないと計算時間上の問題が生じる．

佐井(2019a)では抽出率が1/2と小さかったこともあり$l_{\max}=h_{\max}=50$と設定していたが，例えば$K=8$，$M=20$の偏在分布について同じ設定で多重母集団寸法指標を推定すると表6のような結果になる．表には周辺セルHのサイズhが大きい部分のみを表示しているが，おそらく$h>50$として推定されるはずだった頻度が$h=50$の値として推定されていることが推測される．

表6：多重母集団寸法指標の推定値（h_{max}＝50）

（偏在分布，K＝8，M＝20）

		h							
		⋯	43	44	45	46	47	48	49	50
	0	⋯	0	0	0	0	0	0	0	17396101825
	1	⋯	0	0	0	0	0	0	0	1087322
l	2	⋯	0	0	0	0	0	0	0	2128
	⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

この問題を回避するために，$l_{\max}=50$，$h_{\max}=500$とするが，そのままでは計算時間が大幅に増えて実用には耐えられない．そのため，周辺セルHのサイズで50を超えるものについては，$h=100$，150，200，250，300，350，400，450，500と50飛びのサイズだけを設定する．推定結果では周辺セルのサイズについて誤差が生じることになるが，サイズが大きい部分であるため，真のリンク確率の期待値の推定への影響はそれほど大きくはないと考えられる．

表7に上記の方法で推定した結果を示す．

$h_{\max}$を500まで増やしても，$\hat{S}_{(0,500)}$に若干の頻度が残っている．$K=9$では，推定値が正の値となるhの最大値は150であるが，$K=7$では$\hat{S}_{(0,500)}=62062461$となり，さらに大きな$h_{\max}$を設定する必要性が見てとれる．ただ，計算時間上の判断で，次章での検討ではすべてのケースについて$h_{\max}=500$の設定を用いる．

表７：多重母集団寸法指標の推定値（h_{max}＝500）

（偏在分布，K＝８，M＝20）

		h					
		⋯	46	47	48	49	50
	0	⋯	472474457	491852897	512385488	534090302	556983297
	1	⋯	25350	28553	32081	35953	40188
l	2	⋯	0	0	0	0	0
	⋮		⋮	⋮	⋮	⋮	⋮

		h					
		100	150	200	250	300	350
	0	2059974196	624577432	125881921	55893742	43417752	12825695
	1	395234	73928	20531	37833	75204	8477
l	2	0	0	1	27	279	36
	⋮	⋮	⋮	⋮	⋮	⋮	⋮

		h		
		400	450	500
	0	808490	75851	1451692
	1	12	0	0
l	2	0	0	0
	⋮	⋮	⋮	⋮

４　改善した方法を用いた推定結果

この章では，３.２節で設定した分布，母集団，標本について，予め求めておいた真のリンクとなった個体の比率を推定目標として，３.３～３.５節で提案した解決策を用いた推定結果を示すが，３.４節で提案した超立方体Hcを用いる方法だけでなく，周辺セルＨを用いる方法の結果も合わせて示す.

探索法の各ステップで，尤度が増加する方向に多重母集団寸法指標の暫定的な推定値を移動していく．その移動距離については，尤度が順調に増加する場合には距離を1.1倍して大きな増加を目指すが，荒い動きになるのを防ぐため距離に上限を設ける．逆に，２回続けて尤度の増加が認められない場合には，移動距離を0.9倍して，基準以下となった時点で探索を

終了して，その時点での暫定的な多重母集団寸法指標の推定値を最終的な推定値とする．距離の上限として１つ目の設定の１/10を２つ目に，さらにその１/10を３つ目の設定として，それぞれの設定の下で推定を行い，そのうちで尤度が最大のものを推定値とするが，２つ目の設定でステップ数が十分多かったものについては，３つ目の設定で推定しなかったものもある．

なお，探索法では多重母集団寸法指標の推定値に対する制約条件を渋谷（2005）が提案したペナルティー関数で表現して，対数尤度関数に取り込んでいるが，その詳細についても佐井（2019a）を参照していただきたい．

表８には周辺セルＨを用いた方法，表９には超立方体Hcを用いて多重母集団寸法指標を推定した後に超立方体Dcを領域Ｄで置き換えることによって補正する方法で，それぞれ推定された真のリンク確率の期待値を示す．

また図６，図７には，表８，表９で示した推定値の，真のリンクとなった個体の比率（表１）からの差を分布別に示す．図には，超立方体Hcを用いて多重母集団寸法指標を推定して，そのまま補正を行わなかった場合の結果も加えている．

周辺セルＨを用いて推定した場合の結果を見ると，どちらの分布とも良い推定結果が得られている．これに対して，超立方体Hcを用いて補正を

表８：真のリンク確率の期待値の推定値（周辺セルＨ利用）

(A) 一様分布

		M		
		100	50	20
K	4	0.4281		
	5	0.9657	0.3783	
	6		0.9321	
	7			0.0543
	8			0.6261
	9			0.9133
	10			0.9835

(B) 偏在分布

		M		
		100	50	20
K	4	0.3114		
	5	0.9422	0.1936	
	6		0.8454	
	7			0.0349
	8			0.2577
	9			0.6915
	10			0.9195

表９：真のリンク確率の期待値の推定値（超立方体Hc利用＋補正）

(A) 一様分布

		M		
		100	50	20
K	4	0.3365		
	5	0.9487	0.3850	
	6		0.9298	
	7			0.0847
	8			0.6131
	9			0.8975
	10			0.9581

(B) 偏在分布

		M		
		100	50	20
K	4	0.2706		
	5	0.9385	0.1957	
	6		0.8346	
	7			0.0436
	8			0.2214
	9			0.6429
	10			0.6307

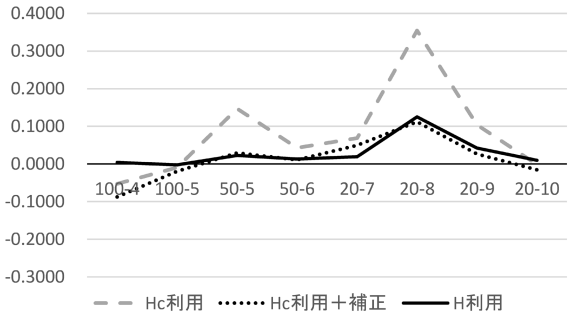

図６　推定値の実際のリンク比率との差（一様分布）

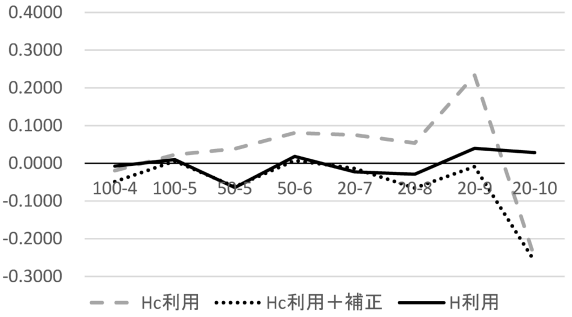

図７　推定値の実際のリンク比率との差（偏在分布）

行わなかった場合には$M=50$，20では過大に推定されているものの，最後に領域Dのセル数を用いた簡単な補正を行うだけで，周辺セルHを用いる方法とほぼ等しい推定結果が得られていることが分かる．

$K=10$，$M=20$の偏在分布では，超立方体Hcを用いた場合，最後の補正の有無にかかわらずかなりの過小推定となっている．このケースでは，多重母集団寸法指標を探索的に推定する際に，移動距離の上限を替えた３通りの場合とも10000ステップに満たずに探索が終了してしまい，３通りの中でステップ数が7568と最も多かった結果を採用した．そのため十分高い尤度に到達していない可能性がある．この点の解明は今後の課題である．

５　おわりに

本稿では，個票データに含まれる個体のキー変数ベクトルにノイズを挿入した場合のリスクとして，距離計測型リンケージを基にした真のリンク確率の期待値を考えた．その推定では，対象となる個体が含まれるセルのサイズとともにその周辺セルのサイズも合わせて考え，その表現として多重寸法指標を用いた．キー変数としては離散型の量的変数を想定したが，これまで検討したケースよりも官庁統計などの調査の実態に近づけるために，キー変数の数を多く，しかも取り得る値を少なく，母集団を大きく，抽出率を小さくした人工的なデータを用いた．そのような状況の下で検討を行ったところ，特に計算時間に関するいくつかの問題点が浮かび上がってきたものの，本稿で提案した対応策で，多くの部分については解決できたか，あるいはその道筋が見えた．

本稿では，挿入するノイズとしては最小の±１のみを考えたが，リスクを効率的に軽減するためには，より大きなノイズの挿入も必要であろう．その場合，例えば±２というノイズだけを挿入するよりは，±１，±２のノイズを確率的に挿入する方が自然である．そうすると，領域Dとして複数のパターンが生じ，何を周辺セルHに設定するか，そして多重寸法指標

をどのようにカウントするか，という問題が新たに生じる．

これについては3.4節の議論が解決策となる．例えば周辺セルHとしては，ノイズの大きさに応じて変動する中での平均的なものを設定しておいて，多重母集団寸法指標を推定した後に領域Dの正確なセル数で補正する際，(9) 式において複数の＃(D)とそれに対応する確率を用いて期待値を求めることにより解決は可能と考えられる．

一方で，本稿では解決できなかった課題もある．

3.5節で提案した，ノンパラメトリック推定法の改良によって，本稿で扱った程度のデータについては，真のリンク確率の期待値を推定することができたが，母集団の大きさが日本の人口程度の場合や，多重母集団寸法指標において周辺セルHのサイズhの最大値$h_{\max}$をさらに大きくしなければならない場合を考えると，ノンパラメトリック法では限界があり，パラメトリックな方法の構築が必要である．

第4章の図6と図7を見ると，特に図6の一様分布の結果において，$M=20$の場合に，周辺セルHを用いる方法，超立方体Hcを用いて最後に補正する方法とも，やや過大推定となっている．各キー変数の取り得る値が少なくなると，キー変数ベクトルが定義域の端になる確率が増えるため，その処理の仕方が結果により強く影響している可能性がある．

図8で$K=2$を例に説明すると，(a) のように対象となる中央セルが（4，1）のとき，2番目のキー変数に－1のノイズを挿入すると定義域外に移動してしまうため，3.2節で標本の個体を1個ずつ確かめて真のリンクとなる比率を計算した際には＋1のみを挿入していた．一方で，多重寸法指標を考える場合には，(b) のように周辺セルHの一部がキー変数の定義域外にはみ出ることを許してカウントを行っていた．そのため，周辺セルのサイズを過小にカウントしてしまい，最終的には真のリンク確率の期待値を過大に推定してしまっていることが理由として考えられる．

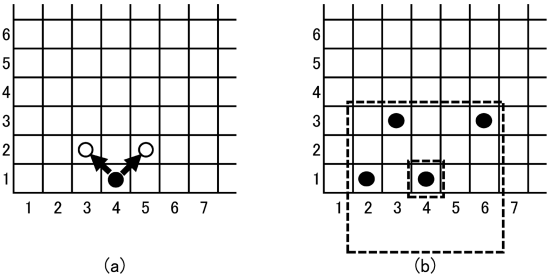

図８　キー変数の定義域の端の処理

対応策としては，周辺セルHを実態に合わせて小さく設定することも考えられるが，様々な大きさの周辺セルが混在してしまうため，多重母集団寸法指標の推定が困難となる．そのため，周辺セルの大きさは通常の大きさに固定して，周辺セルのサイズを水増しする方が容易である．図８の(b)の場合，＃(H)＝24であるが，そのうち定義域内には14セルが含まれている．この図のように標本で３個の個体が周辺セルに含まれている場合，セル数の比で3･24/14≒5.14と水増しして，$h=5$とすることになる．実際にこの方法を用いて，$K=8,9$，$M=20$の一様分布，偏在分布について多重標本寸法指標を求め，それを基に真のリンク確率の期待値を推定してみたが，推定値の十分な減少が確認できないばかりか，逆に値が増加するものもあり，期待された結果は得られなかった．

この方法では，例えば定義域内に含まれる部分が１/100であれば，hの値は100倍に水増しされるため，0, 100, 200, ...のようなサイズしかとれなくなってしまい，不自然な多重標本寸法指標が作られることが原因かもしれない．別の補正方法が必要と思われ，この点についても今後の課題となろう．

謝辞

本稿は，日本学術振興会科学研究費補助金（18H00835），統計数理研究

所共同利用研究（2019-ISMCRP-2042，2019-ISMCRP-4305，2020-ISMCRP-2002，2020-ISMCRP-4501）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 Association*, **85**, 38–45.
［2］伊藤伸介（2010）．ミクロデータにおける秘匿性の評価方法に関する一考察，明海大学経済学論集，**22**（2），1–17.
［3］伊藤伸介，村田磨理子，高野正博（2014）．ミクロデータにおける匿名化技法の適用の可能性－全国消費実態調査と家計調査を用いて－，統計研究彙報，**71**，83–124.
［4］Pitman, J.（1995）. Exchangeable and partially exchangeable random partitions, *Probability Theory and Related Fields*, **102**, 145–148.
［5］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，**51**（2），183–197.
［6］佐井至道（2008）．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，**43**（3），1–18.
［7］佐井至道（2009）．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.
［8］佐井至道（2013）. 官庁統計におけるセンシティブデータ，システム/制御/情報，**57**（4），147–152.
［9］佐井至道（2015a）．ノイズが加えられた個票データのリスク評価，2015年度統計関連学会連合大会予稿集，87.
［10］佐井至道（2015b）．ノイズが加えられた個票データのリスク評価，研究集会「官庁統計データの公開における諸問題の研究と他分野への応用」報告資料，1–5.
［11］佐井至道（2016）．ノイズが挿入された個票データのリスク評価，岡山商大論叢，**52**（1），23–50.
［12］佐井至道（2017a）．曖昧な母集団情報を考慮に入れたノイズを含む個票データのリスク評価，岡山商大論叢，**53**（1），27–57.
［13］佐井至道（2017b）．ノイズが挿入された個票データの変数の型によるリスクの差について，2017年度統計関連学会連合大会予稿集，81.
［14］佐井至道（2017c）．個票データに挿入するノイズの型によるリスクの差について，研究集会「公的大規模データの利用におけるプライバシー保護の理論と応用」報告資料，1–4.
［15］佐井至道（2018a）．個票データのキー変数の型と開示リスクとの関係，岡山商大論叢，**54**（1），1–28.
［16］佐井至道（2018b）．リスクの高いデータの秘匿について，2018年度統計関連学会連合大会予稿集，18.
［17］佐井至道（2018c）．秘匿措置が施されたデータのリスク評価の試み，研究集会「公的大規模データの利用におけるプライバシー保護の理論と応用」報告資料，1–5.
［18］佐井至道（2019a）．局所的な分布を用いた個票データのリスク評価，岡山商大論叢，**55**（2），45–70.
［19］佐井至道（2019b）．局所的な分布を用いた個票データのリスク評価，2019年度統

計関連学会連合大会予稿集，60.
［20］佐井至道（2019c）．局所的な分布を利用した個票データのリスク評価の実用化に向けて，研究集会「公的データの利用とプライバシー保護の理論」報告資料，1–10.
［21］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.
［22］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.
［23］Sibuya, M. and Sai, S.（2008）. Analysis of a Dataset for Statistical Disclosure Control by Random Partition of a Multi-index, *Cherry Bud Workshop 2008*, 1–13.
［24］Willenborg, L. and de Waal, T.（1996）. *Statistical Disclosure Control in Practice*, Springer.
［25］Willenborg, L. and de Waal, T.（2001）. *Elements of Statistical Disclosure Control,* Springer.