『岡山商大論叢』（岡山商科大学）
第58巻第１号　2022年６月
Journal of OKAYAMA SHOKA UNIVERSITY
Vol.58　No.１　June 2022

《論　　説》

ノイズを挿入して撹乱された個票データに対する多重寸法指標を用いた種々のリスク評価方法の比較

佐　井　至　道

Comparison of Several Risk Assessment Methods Using Multi Size Indices for Microdata Perturbed by Adding Noises

SAI Shido

１　はじめに

個体の調査結果（個人や事業所についての変数の値）がそれぞれレコードとして並べられた個票データを秘匿する方法としては，カテゴリの併合のような非撹乱的方法と，ノイズの挿入やスワッピングのような撹乱的方法がある．

このうちノイズの挿入は，個体の識別を防ぐ目的でキー変数に対して行われる場合と，個体の情報そのものを保護するためにセンシティブ変数に対して行われる場合がある．ここでキー変数とは，第三者が個票データの公開前から情報を持っている可能性が高く，個票データの個体を識別する

ために用いられる変数であり，一般的には性別や年齢などが該当する．またセンシティブ変数とは，各個人が他人に知られたくないと思っている情報に関する変数で，病歴や年収などが想定される．

本稿では，標本から得られた個票データについて，個体の識別を防ぐことを目的としてキー変数に対してノイズの挿入を行った場合のリスク評価について考える．なお，第三者は母集団のすべての個体について，すべてのキー変数の値の情報を持っていることを想定してリスク評価を行うが，特に母集団が大きい場合には，実際にそれほどの情報を持っている可能性は小さい．その意味で，本稿での仮定は最悪のシナリオと考えられる．

非攪乱的方法については Willenborg and de Waal（1996, 2001）を，非攪乱的方法で秘匿された個票データに対する寸法指標を利用したリスク評価方法については佐井（2003，2013）を参照していただきたい．本稿では，ノイズの挿入後のリスク評価に距離計測型リンケージを基礎とした真のリンク確率の期待値という指標を用いる．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．なお，距離計測型リンケージを用いるリスク評価と真のリンク確率を用いるリスク評価とでは，リスク評価を個票データ内で行うか母集団内で行うかが異なる．

真のリンク確率をリスクの指標とした一連の研究の概略を述べる．佐井（2015a，2015b，2016，2017a，2017b，2017c，2018a，2018b）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，真のリンク確率の期待値を指標として用いてリスク評価を行う方法を提案した．これはノイズを挿入した個体が，母集団で真のリンクとなる確率の期待値を考えるものだが，その際，キー変数の分布としては，一様分布や正規分布など定義域全体についての連続分布を主に設定した．

一方，佐井（2018c，2019a，2019b，2019c，2020a，2020b，2021a）では，離散型量的変数であるキー変数に対してノイズを挿入した場合について，

本稿で方法［H］と呼ぶ，多重寸法指標を用いるリスク評価方法を提案した．多重寸法指標によって特定のキー変数の値に近い部分の分布，すなわち局所的な分布を表現することができる．

特に佐井（2019c）以降では，検討に用いる人工的なデータについて，母集団の大きさを10^6，抽出率を1/100とするとともに，キー変数の数を段階的に10まで増やし，キー変数のとりうる値の数も段階的に20まで減少させ，より実データに近い状況における提案した方法の有効性について検討した．

本稿は佐井（2021b，2021c）での検討結果をまとめたものであるが，検討は主に２点からなる．１点目は多重寸法指標として従来の方法［H］に加えて，より小さい部分を考える方法［D］を提案し，両者の有効性を比較する．２点目はキー変数の定義域の端付近の処理の問題である．例えば年齢がキー変数の場合，30歳には＋1のノイズを加えて31歳と公表しても，－1のノイズを加えて29歳と公表しても構わないが，０歳には－1を加えて－1歳と公表することはできない．キー変数のとりうる値の数が少なくなると，キー変数の値が端付近になる確率が高くなり，影響は無視できない．この点について特別な対処を行う場合と行わない場合について，有効性の比較を行う．

第２章では，本稿で用いるいくつかの定義について述べるとともに，多重寸法指標の導入方法と真のリンク確率の期待値の推定方法について説明する．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，領域D，和集合Hの定義とともに，これまでの研究でも用いてきた方法［H］の多重寸法指標について説明する．これは和集合Hを周辺セルとするものである．それに加えて，領域D自体を周辺セルとする方法［D］での多重寸法指標の説明を行う．2.3節では，方法［H］，［D］それぞれについて，真のリンク確率の期待値を推定するための具体的な手順を示す．

第３章では，提案した２つの方法［H］，［D］を人工的に作成した様々な分布に対して適用して，その有効性を比較するとともに，特に方法［D］の問題点についても考える．

第４章では，キー変数の値が定義域の端付近にある場合の対処法について検討を行う．まず4.1節では，方法［H］についての対処法［H edge］の手順の概略を述べる．手順は第２章の方法と同じ３つのステップからなるが，１番目と３番目のステップにやや煩雑な調整が入る．4.2節では１番目のステップについて，4.3節では３番目のステップについて，それぞれ詳しく述べることにする．4.4節では，方法［D］についての対処法［D edge］の手順の概略を述べ，それらの対処方法を加えた真のリンク確率の期待値の推定結果を4.5節で示すとともに，［H edge］について若干の改善を図る．

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　リスク評価への２種類の多重寸法指標の導入

2.1　リスクの指標の概要

この節では，キー変数にノイズを挿入した場合の個体のリンクの概念と，本稿でリスクの指標とする真のリンク確率の期待値について述べる．必要最小限の説明にとどめるため，詳細については佐井（2019a，2021a）を参照していただきたい．

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，その標本から個票データが作られているものとする．個票データに含まれるK個のキー変数はすべて離散型量的変数で，第三者は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いると考える．すべてのキー変数をまとめたものをキー変数ベクトルと呼び，キー変数に挿入するノイズについても，すべてをまとめてノイズ変数ベクトルと呼ぶ．

標本の i 番目（$i = 1, 2, ..., n$）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．キー変数ベクトルは

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

と表され，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，キー変数ベクトルそのものではなく，ノイズ変数ベクトルを加えて$\mathbf{x}_i + \mathbf{e}_i$をキー変数ベクトルとする．また，母集団の i 番目（$i = 1, 2, ..., N$）の個体のキー変数ベクトルを$\mathbf{a}_i$とする．ただし，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．

$\mathbf{x}_i$，$\mathbf{e}_i$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と呼ぶが，キー変数が離散型量的変数の場合には格子点の集合になる．本稿では格子点を分割表にならってセルで表し，セルに元のキー変数ベクトルや公開用のキー変数ベクトルが入ることを，「セルに個体が入る」と表現することがある．

次に本稿で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．

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，それに確率変数であるノイズ変数ベクトル$\mathbf{e}_i$を挿入した場合を考える．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$\mathbf{a}_{i''}$が入る確率を $p_{\mathbf{e}_i}(\mathbf{x}_i, \mathbf{a}_{i''})$ と表すとき，$\mathbf{x}_i$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

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

と書ける．$\mathbf{a}_{i''}$ が確率変数でなく確定値の場合は，$p_{\mathbf{e}_i}(\mathbf{x}_i, \mathbf{a}_{i''})$ の値は 0 または 1 であることより，(3) 式も 0 または 1 となる．

$\mathbf{e}_i$ について (3) 式の重み付き平均をとることによって，$\mathbf{x}_i$ が真のリンクとなる確率 $P_t(\mathbf{x}_i)$ が得られる．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，これを真のリンク確率の期待値と呼ぶ．簡単のため $\mathrm{E}(P_t)$ と表すことがある．

これは，第三者が持っている完全な情報を基に，ノイズが挿入された標本の各個体について，キー変数ベクトルの値が最も近い母集団の個体と判断した場合に，平均的にどの程度正しいかを表す値であり，個票データが母集団から作成されたことを想定する距離計測型リンケージの拡張となっている．この指標の値が大きいほどリスクは大きい．

2.2 多重寸法指標の設定

この節では，真のリンク確率の期待値を推定するために 2 種類の多重寸法指標を導入する．

本稿では，ノイズ変数は個票データに含まれるすべての個体のすべてのキー変数に対して挿入する．またノイズ変数$e_{i,k}$ 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．例えば $K=2$ の場合は，$\mathbf{e}_i = (-1, -1), (+1, -1), (+1, +1), (-1, +1)$ を確率 $(1/2)^2$ ずつでとることになる．

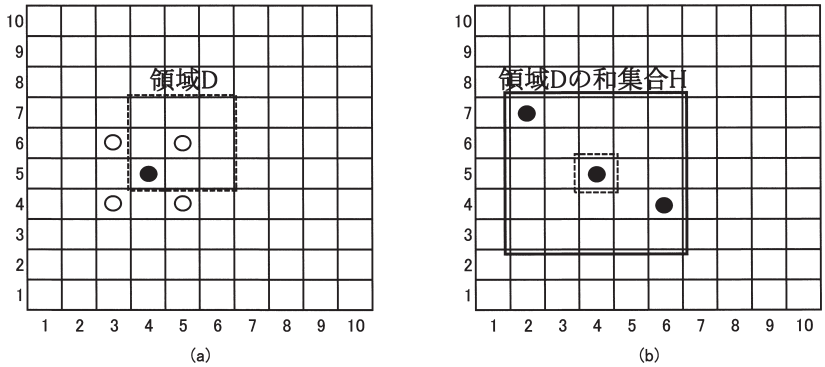


図１　領域Dとその和集合H

図１は$K = 2$で，２個のキー変数のとりうる値がともに１から10の場合の分割表を表している．以後も説明にはこの設定を用いる．(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つのベクトルを白抜きの点で示している．ノイズ変数ベクトル$\mathbf{e}_i = (1,1)$を挿入した場合，領域Dは点線で囲んだ正方形となる．各ノイズ変数が ± 1 をとる場合，領域Dは$K = 2$ではこのように正方形に，$K = 3$では立方体になるが，Kが増加するに従って丸みをおびた超球に近い形状となる．また，ノイズ変数のとる値が大きい場合は，$K = 2$であっても円に近い形状となる．

なお本稿では，キー変数ベクトルはあらかじめ範囲が決められた定義域内の値しかとることができないものと考える．実データにおける年齢の下限は０歳であるが，上限は決まっていない．しかし，個票データでは桁数が定められていることや，極端に高い年齢を表示しないなどの理由で，初めからトップコーディングした値を記録していることが多い．例えば年齢を２桁で表示する場合，99歳以上をまとめて「99」と表示することがある．

このように定義域が定められているとき，キー変数ベクトルが定義域の端にある場合には，ノイズ変数ベクトル挿入後に定義域外にはみ出ないように，ノイズ変数ベクトルの生成に制限をかける必要がある．具体的には，

図１のように各キー変数のとりうる値が１から10の場合，キー変数の値が１であればノイズは確率１で＋1を，値が10であれば確率１で－1をそれぞれ挿入することにする．

ここで多重寸法指標の２種類の設定方法［H］，［D］を考える．まず方法［H］について説明する．

図１（b）では，４通りのノイズ変数ベクトル$\mathbf{e}_i$に対する領域Dの和集合Hを，大きく実線で囲んでいる．この部分を周辺セルと呼ぶことにする．なお，元のキー変数ベクトルの値が含まれている点線で囲んだセル（4, 5）は，周辺セルに含めないものとして，周辺セルとの対比で中央セルと呼ぶ．

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，周辺セルには２個の個体が含まれサイズ２となっており，そのサイズの組を（1, 2）と書くとき，標本におけるそのような組の頻度を$s_{(1,2)}$と表す．

一般に標本において，中央セルのサイズがl（$l = 0, 1, ..., n$），周辺セルのサイズがh（$h = 0, 1, ..., n$）の組（l, h）の頻度を$s_{(l,h)}$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（$l = 0, 1, ..., N$），周辺セルのサイズがh（$h = 0, 1, ..., N$）の組（l, h）の頻度を$S_{(l,h)}$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，周辺セルのサイズhが０の場合も含まれる．

図１（b）のような多重寸法指標の設定方法を［H］と略記する．方法［H］は佐井（2019a，2021a）における設定方法と同じであるが，この後に述べる方法［D］での説明において混乱が生じないように，本稿ではHという記号を，周辺セルではなく領域Dの和集合にあてることにする．

先に述べたように本稿では，ノイズ変数ベクトル挿入後に定義域外にはみ出ないようにノイズを挿入するが，挿入後が定義域内であっても，通常の大きさの領域Dや周辺セルの一部が定義域外にはみ出る可能性がある．具体的には図２（case b）のように元のキー変数ベクトルが（3, 10）の場合には，ノイズ変数ベクトルとして$\mathbf{e}_i$ =（－1, －1），（＋1, －1）を確率 2/4

ずつでとることになるため，図３（case b）のように領域Dは２個しか存在せず，和集合は通常の周辺セルよりも小さくなる．

また図２（case c）のように元のキー変数ベクトルが（9, 9）の場合には，ノイズ変数ベクトルとしては４通りすべてを挿入することが可能であるが，$\mathbf{e}_i = (-1, -1)$ 以外のノイズ変数ベクトルを挿入した場合には，図３（case c）のように，通常の大きさの領域Dであれば一部が定義域外にはみ出てしまい，定義域内の周辺セルは通常の周辺セルよりも小さくなる．

第３章の人工的なデータを用いた検討では，上述したような周辺セルの定義域内の部分が通常の周辺セルよりも小さくなる点については特に考慮しない．多重標本寸法指標をカウントするときには定義域内の個体のみを考えることになるが，真のリンク確率の期待値を推定する場合には，領域Dや周辺セルとしては図４のように定義域外にはみ出た部分も含めた大きさを考える．第４章において，その対処方法を提案するとともに，対処の有無による推定への影響について検討することにする．

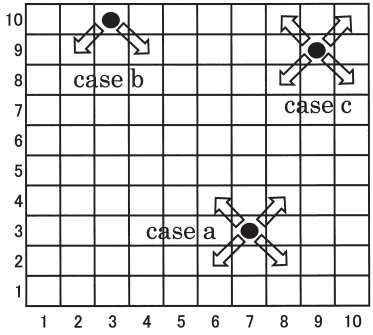


図２　キー変数ベクトルの定義域とノイズ変数ベクトルの挿入

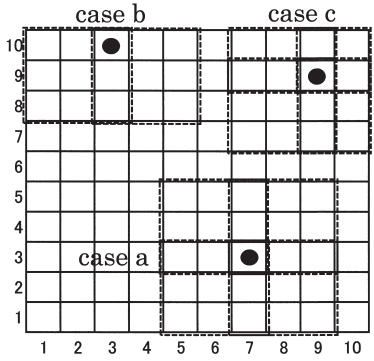


図３　キー変数ベクトルの定義域の端付近の領域D

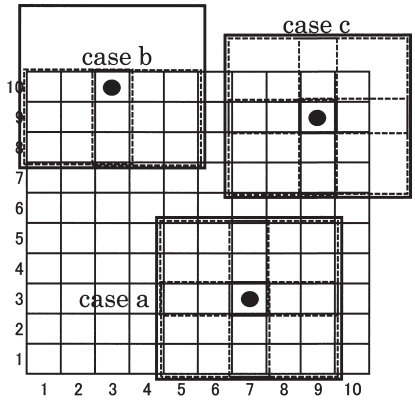


図４　定義域外へのはみ出しを許した場合の領域Dと周辺セル

次に方法［D］について説明する．

図１(a) で，領域Dから中央セルを除いた部分を周辺セルとして，中央セルと周辺セルのサイズの組ごとに頻度をカウントすることによって多重寸法指標を設定する方法を［D］と略記する．

この場合，キー変数ベクトル$\mathbf{x}_i$を固定しても，ノイズ変数ベクトル$\mathbf{e}_i$が確率変数ベクトルであるため，例えば $K = 2$ の場合には各サイズの組を確率 $(1/2)^2$ でとる，一般の K では確率 $(1/2)^K$ でとる，と考える必要が

ある．

方法［D］においても，前述のように，元のキー変数ベクトルが図２（case b）の（3, 10）の場合には，領域Dである周辺セルは図３（case b）のように４個のうち２個しか存在しない．図２（case c）の（9, 9）の場合には，ノイズ変数ベクトル挿入後のキー変数ベクトルはすべて定義域内にあるものの，図３（case c）のように通常の大きさの領域Dでは，３個についてその一部が定義域外にはみ出てしまい，定義域内の周辺セルは通常のものよりも小さくなるが，これについても方法［H］と同様に，第３章の人工的なデータを用いた検討では特に考慮せず，第４章においてその対処方法を考える．

2.3　真のリンク確率の期待値の推定方法

この節では２種類の多重寸法指標の設定方法［H］，［D］について，真のリンク確率の期待値 $\mathrm{E}(P_t)$ の推定方法について説明する．

方法［H］については既に佐井（2019a, 2021a）で述べているが，［H］，［D］のどちらの方法においても，次の３つのステップで推定を行うことになる．

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

なお，l と h の分布が等しい場合には，Pitman（1995）が考案した Pitman モデルを拡張した多重 Pitman モデルを用いて多重母集団寸法指標を推定する方法があるが，ここでは l と h が非対称であるため，制約付きノンパラメトリック最尤推定法を用いる．実際には，最尤推定値を勾配法と呼ばれる探索的な方法で見つけ出すことになる．Step 1，2 については佐井（2021a）を参照していただきたい．また多重Pitmanモデルについては渋谷，佐井（2007），Sibuya and Sai（2008）を参照のこと．

ここでは Step 3 について詳細を説明する．

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．そのため，非攪乱的な秘匿措置が施された個票データに対して寸法指標を用いたリスク評価を行う際には，推定された母集団寸法指標の情報のみを用いるのが一般的である．多重寸法指標についても同様で，多重標本寸法指標における特定のサイズの組を，多重母集団寸法指標における特定のサイズの組に対応づけすることはできない．そのため，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．

まず，方法［H］の場合の推定法について説明するが，詳細は佐井（2021a）に譲り，ここでは概略のみを述べる．

$(0,0)$ を除くすべてのサイズの組 (l,h) について，多重母集団寸法指標の推定値 $\hat{S}_{(l,h)}$ が得られたとする．サイズの組 $(1,h)$ $(h=0,1,...,N)$ のうちの１つを考える．

ここで，周辺セル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．このとき，キー変数ベクトルが$\mathbf{a}_i$である母集団でサイズ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 $\#(\mathrm{D})$ は領域Dのセル数を表し，$\#(\mathrm{H})$ は領域Dの和集合Hである周辺セルのセル数を表している．$\#(\mathrm{D})-1$ は領域Dから中央セルを除いたセル数を意味している．

中央セルのサイズ l 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，母集団で中央セルのサイズが２以上の場合，そのセルに含まれるキー変数ベクトル$\mathbf{a}_i$の個体については $P_t(\mathbf{a}_i)=0$ となる．

方法［H］，［D］とも，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，

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

と推定することができる．したがって，方法［H］では

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

と推定される．

次に方法［D］の場合の推定法について説明する．

$(0,0)$ を除くすべてのサイズの組 (l,h) について多重母集団寸法指標の推定値 $\hat{S}_{(l,h)}$ が得られたとする．ここで各サイズの組の確率は $w = (1/2)^K$ である．

方法［H］と同様に，母集団で中央セルのサイズ l が２以上であれば，そのセルに含まれるキー変数ベクトル $\mathbf{a}_i$ の個体については $P_t(\mathbf{a}_i) = 0$ となる．また $l = 1$ であっても $h \geq 1$ であれば，ノイズ挿入後の個体が周辺セルに含まれる他の個体にリンクされるため，間違ったリンクとなる．そのため真のリンクとなるのは母集団で $(l,h) = (1,0)$ という組み合わせのみであり，その場合には確率１で真のリンクとなる．なお方法［D］では，方法［H］のように周辺セル内の分布を仮定する必要はない．

したがって，抽出される標本における真のリンク確率の期待値は

$$\hat{\mathrm{E}}\big[P_t(\mathbf{x}_i)\big] = \frac{1}{N} w \hat{S}_{(1,0)} \tag{8}$$

と推定される．

３　人工的なデータによる推定結果の比較

この章では，２種類の多重寸法指標の設定方法［H］，［D］を用いた個票

データのリスク評価方法を，人工的に作成した数値データに対して適用することにより，それぞれの方法の推定精度と特徴を比較する．

佐井（2021a）と同じ人工的な数値データ（A），（B）を用いる．どちらのデータでも，キー変数の個数として $K = 1, 2, ..., 10$ の10通りを考える．それぞれのキー変数は1から始まる連続的な自然数をとるものとして，その最大値としては $M = 100, 50, 20$ の3通りを考える．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である．確率関数 $f(a_{i,k}) = 1/M$ を用いて，K個ずつ$N = 10^6$ 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．データ（A）を簡単に一様分布と呼ぶ．実際の個票データにおけるキー変数には，年齢，誕生月，性別のように一様分布で近似できるものも多い．

データ（B）も母集団の各キー変数 $a_{i,k}$は i，k について独立で，確率関数は $1 \leq a_{i,k} \leq 5$ のとき $f(a_{i,k}) = a_{i,k}/(3M)$ と単調増加で，$6 \leq a_{i,k} \leq 10$ のとき $f(a_{i,k}) = (11 - a_{i,k})/(3M)$ と単調減少し，以後同様に10毎の周期がM/10回ある．この確率関数を用いてK個ずつの $N = 10^6$ 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．データ（B）を偏在分布と呼ぶ．実際の個票データでは，キー変数の定義域の中で個体が集まる部分と個体がほとんど存在しない部分が存在することが多い．（B）はそのようなデータを人工的に再現しているものである．

それぞれの母集団から $n = 10^4$ 個の個体に対応するキー変数ベクトルを，乱数を用いて非復元単純無作為抽出して，これを標本のキー変数ベクトルとする．次に，乱数を用いて確率 1/2 ずつで ± 1 をとるノイズを作り，各キー変数に独立に挿入することにより公開用のデータを作成する．標本のキー変数ベクトルのデータ，公開用のキー変数ベクトルのデータとも，作成したのは各分布についてそれぞれ1個ずつである．

佐井（2021a）において，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，真のリンクとなる比率は表１のようになった．表中で下線を引いている 0.02 ～ 0.98 の範囲の値を本稿での推定目標とする．

表２には方法［H］を用いた推定結果を，表３には方法［H］の Step 1 の計算時間を短縮するために，領域Dの和集合Hの代わりに，Hに含まれる最大の超立方体Hcを周辺セルとして用いた場合の推定結果を，表４には方法［D］を用いた推定結果を，それぞれ示す．

方法［H］で用いる周辺セルは，領域Dを含むより広い領域内での個体の密度を測るために利用されている．２番目の方法では，それよりもやや小さい超立方体を領域として用いることになるが，中央セルの近くの密度が極端に高い場合などを除き，両者における周辺セル内の個体の密度が大きく異なることはないため，推定結果にも大きな影響は与えないと思われる．この方法を［Hc］と表す．詳細については佐井（2021a）を参照していただきたい．

また図５，６には，KとMの８通りの組み合わせを横軸として，表２，表３，表４で示した推定値の，表１の実際に真のリンクとなる個体の比率からの差を分布別に示す．例えば横軸の4 −100は$K = 4$，$M = 100$を，ま

表１　真のリンクとなる個体の比率

（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

た縦軸の－0.10は差が－10%であることを意味する.

表２　真のリンク確率の期待値の推定値（方法［H］）

（A）一様分布

K \ M	100	50	20
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

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

表３　真のリンク確率の期待値の推定値（方法［Hc］）

（A）一様分布

K \ M	100	50	20
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

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

表４　真のリンク確率の期待値の推定値（方法［D］）

（A）一様分布

K \ M	100	50	20
4	0.3486		
5	0.9616	0.1036	
6		0.9115	
7			0.0437
8			0.4113
9			0.8627
10			0.9722

（B）偏在分布

K \ M	100	50	20
4	0.2739		
5	0.9145	0.3497	
6		0.7997	
7			0.1117
8			0.5261
9			0.7952
10			0.9154

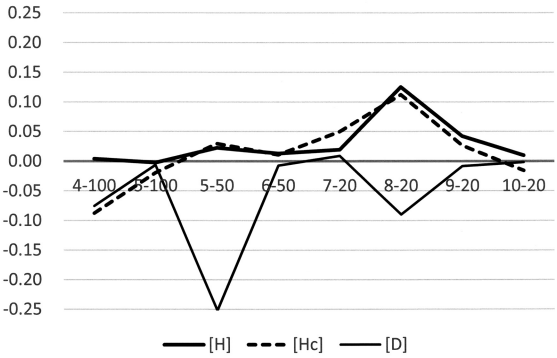


図５　推定値の実際の真のリンク比率からの差（(A) 一様分布）

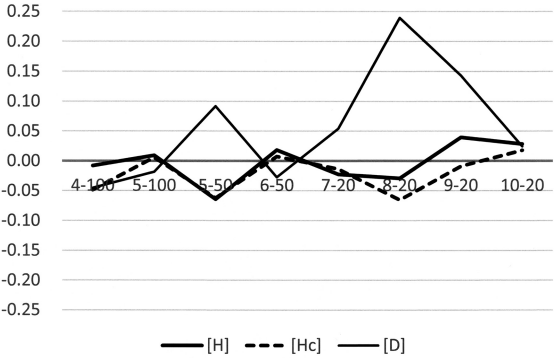


図６　推定値の実際の真のリンク比率からの差（(B) 偏在分布）

方法［H］の推定結果を見ると，どちらの分布でも実際に真のリンクとなる個体の比率に近い値が得られており，方法［Hc］でも遜色のない結果が得られている．それと比較して方法［D］では推定が不安定で，過大推定と過小推定が混在する．±25%程度の差も見られ，このままでは実用にはほど遠い．

(A) 一様分布，$K = 10$，$M = 20$の場合の，方法［H］の観測された多重標本寸法指標を表５(a) に，推定された多重母集団寸法指標を (b) に，それぞれ示す．10^6 以上の値は概数を指数表示している．ここでは $h \leq 5$ の

部分のみを表示しているが，(a) において正の値をとる h の最大値は5，(b) においては22である．KとMの組み合わせによっては，推定された多重母集団寸法指標で正の値をとる h の最大値が100を超えることもある．なお，すべての場合について，多重母集団寸法指標の最尤推定値を勾配法で探索する際には，各 l，h に対して $S_{(l,h)} \geq 0$ という非負条件と，各 h について l に関する単調減少，すなわち $S_{(l,h)} \geq S_{(l+1,h)}$ という条件を制約条件として課した．

表6には，(A) 一様分布，$K = 10$，$M = 20$ の場合の，方法［D］の観測された多重標本寸法指標と推定された多重母集団寸法指標を示す．(b) において正の値をとる h の最大値は5である．方法［D］についても多重母集団寸法指標に課す制約条件は方法［H］と同じである．

表5　観測された多重標本寸法指標と推定された多重母集団寸法指標
（方法［H］，(A) 一様分布，$K = 10$，$M = 20$）

(a) 多重標本寸法指標（$h \leq 5$ の部分）

		h					
		0	1	2	3	4	5
l	0	-	$2.112 \cdot 10^{11}$	$2.661 \cdot 10^{9}$	$2.607 \cdot 10^{7}$	210820	1200
	1	9784	214	2	0	0	0
	2	0	0	0	0	0	0

(b) 多重母集団寸法指標の推定値（$h \leq 5$ の部分）

		h					
		0	1	2	3	4	5
l	0	-	$2.984 \cdot 10^{12}$	$2.213 \cdot 10^{12}$	$1.505 \cdot 10^{12}$	$9.370 \cdot 10^{11}$	$5.338 \cdot 10^{11}$
	1	109409	183333	335027	226810	99403	33581
	2	0	0	0	0	0	0

方法［H］では（7）式から分かるように，多重母集団寸法指標のうち，サイズの組み合わせ $(1, h)$（$h = 0, 1, ..., N$）のすべてについて真のリンクとなる可能性がある．例えば $S_{(1,4)}$ が過小推定されたとしても，$S_{(1,3)}$ や $S_{(1,5)}$ がその分過大推定されれば過不足は相殺される．このことが推定を

表6　観測された多重標本寸法指標と推定された多重母集団寸法指標
（方法［D］，（A）一様分布，$K = 10$，$M = 20$）

（a）多重標本寸法指標（$h \leq 5$ の部分）

		h					
		0	1	2	3	4	5
l	0	-	$2.627 \cdot 10^{12}$	$3.732 \cdot 10^{8}$	0	0	0
	1	$1.024 \cdot 10^{7}$	2843	0	0	0	0
	2	0	0	0	0	0	0

（b）多重母集団寸法指標の推定値（$h \leq 5$ の部分）

		h					
		0	1	2	3	4	5
l	0	-	$2.553 \cdot 10^{14}$	$3.731 \cdot 10^{12}$	$1.979 \cdot 10^{8}$	12263	11
	1	$9.956 \cdot 10^{8}$	$2.842 \cdot 10^{7}$	3159	0	0	0
	2	0	0	0	0	0	0

安定させていると考えられる．

これに対して方法［D］では，多重母集団寸法指標のうちで真のリンクとなるサイズの組み合わせは，(8) 式のように $(1, 0)$ のみである．その推定結果がそのまま真のリンク確率の期待値の推定に影響を及ぼすため，推定が不安定になっていると予想される．この点については第5章で再考する．

なお，方法［H］において，（A）一様分布，$K = 8$，$M = 20$ では10%を超える過大推定となっている．Mが小さい場合には，キー変数ベクトルが定義域の端付近となる可能性が高い．前述したように，ここまでの検討では通常の大きさの周辺セルが定義域外にはみ出しても，補正などは行わずに定義域内の個体を多重標本寸法指標としてカウントした．一方，真のリンク確率の期待値の推定では，定義域外の部分も存在すると仮定して通常の大きさの領域D，周辺セルのセル数を用いた．そのため周辺セルの大きさと比較して周辺サイズ h が小さく観測される一方で，領域Dを実態よりも大きく設定することになる．推定への影響は相殺されている可能性があるが，補正をしない場合には，様々な大きさの周辺セルが混在した多重標

本寸法指標を用いることになり，多重母集団寸法指標の推定に影響を与えている可能性がある．この点について次章で検討を行う．

4　キー変数の定義域の端についての対処

4.1　方法［H］における対処方法［H edge］の概要

前述したように，キー変数ベクトルが定義域の端付近にある場合，通常の大きさの周辺セルが定義域外にはみ出すことがある．例えば図7の中央セルの位置にキー変数ベクトルがあるとき，方法［H］における通常の大きさの周辺セルのうちで定義域内となる比率は $(15 - 1)/(25 - 1) = 0.583$ となる．

第3章で用いた人工データのうちで，（A）一様分布，$K = 4$，$M = 100$ と，（A）一様分布，$K = 10$，$M = 20$ について，$s_{(0,0)}$ 以外の多重標本寸法指標としてカウントされる個々の場合について，方法［H］における通常の大きさの周辺セルのうちで定義域内となる比率をそれぞれ計算し，度数分布としてまとめた結果を図8に表す．ただし計算時間の短縮のため，周辺セルとしては，領域Dの和集合Hの代わりに方法［Hc］の超立方体を用いている．

度数分布では定義域内となる比率について，0以上0.1未満を第1階級，0.1以上0.2未満を第2階級，…，0.9以上1.0未満を第10階級，1.0のみを第11階級として，11階級に分けて表示している．図では階級の範囲の代わりに，階級値を％表示している．Mが小さいほど，またKが大きいほど，定義域内の比率の小さい場合が増加する傾向がある．

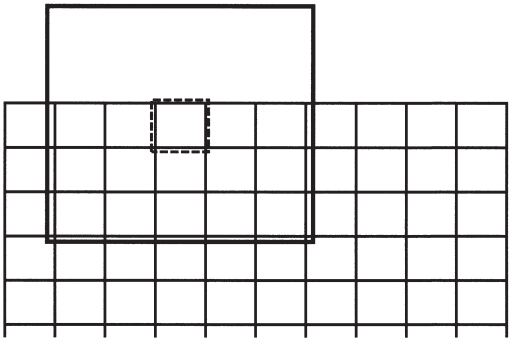
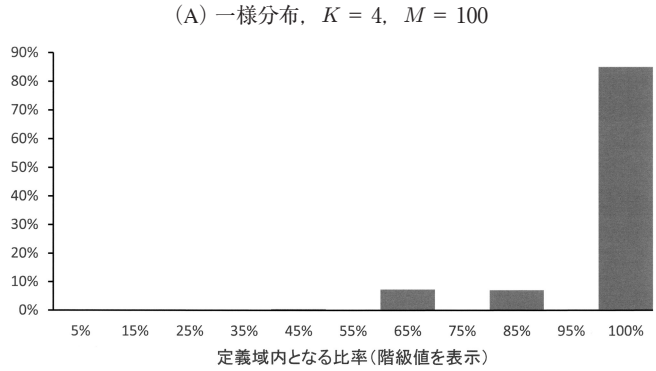


図7　定義域と周辺セルとの関係の例



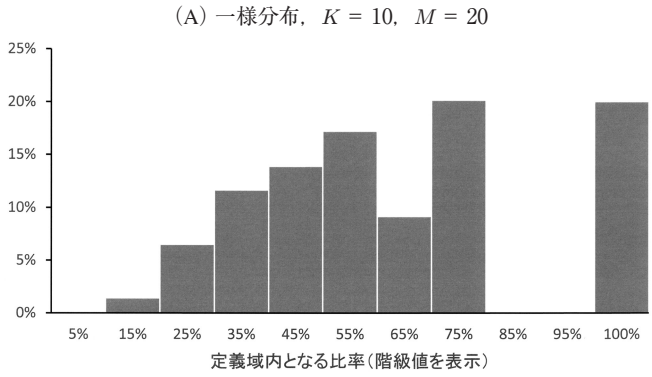


図8　周辺セルのうち定義域内となる比率の度数分布（サイズ調整に使用）

この章では，このように通常の大きさの領域Dや周辺セルの一部が定義域外にはみ出す場合の対処法を考える．まず方法［H］の対処方法［H edge］について検討する．

第３章における方法［H］での真のリンク確率の期待値の推定では，図４に示すような通常の大きさに統一した領域Dと周辺セルを用いた．しかし実際には，領域Dに母集団の個体が入るのは定義域内に限られる．例えば図３（case c）では，領域Dの定義域内となるセル数は，左下から反時計回りに，９セル，６セル，４セル，６セルとなり，その和集合を周辺セルとすべきであろう．

ただ，多重標本寸法指標を基にした多重母集団寸法指標の推定では，（case c）からカウントされる特定の多重標本寸法指標が，推定された多重母集団寸法指標のどれに対応しているかという関連付けはできない．そのため，領域Dや周辺セルの大きさの情報を，真のリンク確率の期待値の推定まで持ち続けても，それを適用する多重母集団寸法指標は全体から無作為に選ぶような方法をとらざるを得ない．

ここでは周辺セルの大きさについては通常の大きさに統一した多重寸法指標を設定するものの，周辺サイズ h については補正し，真のリンク確率の期待値の推定においては，定義域内の領域Dの大きさについての情報を用いることにする．

より正確には次のような手順となる．

Step 1’　図９のように，定義域外へのはみ出しを許して，周辺セルを通常の大きさに統一する．定義域内でのみ観測した標本における周辺サイズ h を，拡大した周辺セルに合わせて $\underline{h}$ に増やす調整を行い，調整された多重標本寸法指標 $\underline{s}_{(l,\underline{h})}$ を求める．この２つの周辺セルを，定義域内周辺セル，拡大周辺セルと呼ぶことにする．なお中央セルの大きさは変わらないため，l についての調整は行わない．

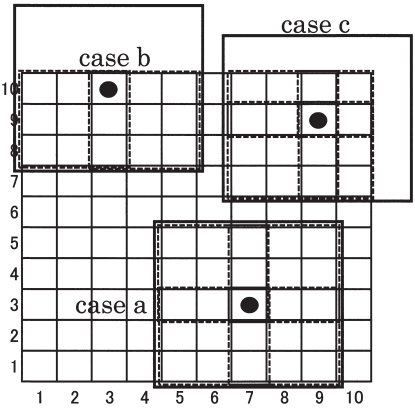


図９　［H edge］において通常の大きさに統一した周辺セル

Step 2’　調整された多重標本寸法指標を基に多重母集団寸法指標 $\underline{S}_{(l,\underline{h})}$ を推定する．

Step 3’　多重母集団寸法指標の推定値 $\hat{\underline{S}}_{(l,\underline{h})}$ を基に真のリンク確率の期待値を推定する．その際，領域Dの大きさとしては定義域内に入る比率についての度数分布の情報を用いる．

上記のステップのうち，Step 1’と Step 3’については，それぞれ4.2節と4.3節で詳細を述べる．

4.2　周辺サイズの調整方法

Step 1’で行う周辺サイズの調整では，通常の大きさの周辺セルのうち定義域内となる比率が例えば0.2のときは，その逆数を用いて，$\underline{h} = 5h$ とするのが最も簡単な方法である．しかし，そのようにして調整した多重標本寸法指標では，$\underline{h}$ の値について飛び飛びに正の頻度が出現してしまう．また，頻度が正となるような $\underline{h}$ の最大値 $\underline{h}_{\max}$ も，h の最大値 $h_{\max}$ と比較して極めて大きくなる．そのような多重標本寸法指標を用いると，Step 2’の多重母集団寸法指標の最尤推定値の探索が早期に終了するなど，推定は不安定となった．$\underline{h}_{\max}$ を小さく抑える補正を行っても状況は変わらなかっ

た．そのため次のような方法を用いる．

4.1節のように，$s_{(0,0)}$ 以外の多重標本寸法指標としてカウントされる個々の場合について，通常の大きさの周辺セルのうち定義域内となる比率をそれぞれ求め，それらを11階級に分ける．第 j 階級の階級値をm_jとして，第 j 階級に含まれるすべての周辺セルについて，定義域内となる比率をm_jとみなす．

中央セルと定義域内周辺セルに含まれるサイズの組を (l, h)，中央セルと拡大周辺セルに含まれるサイズの組を $(l, \underline{h})$ とする．ここで拡大周辺セル内で個体が一様分布すると仮定すると，拡大周辺セルを定義域内周辺セルに限定することによりサイズの組が $(l, \underline{h})$ から (l, h) となる推移確率は，第 j 階級に属する周辺セルについては

$$p^j_{\underline{h},h} = \binom{\underline{h}}{h} m_j^h (1 - m_j)^{\underline{h}-h} \tag{9}$$

と書くことができる．この確率には l は関係していない．図10は中央セルと定義域内周辺セルのサイズの組 $(l, 2)$ に推移する，中央セルと拡大周辺セルのサイズの組のイメージである．図では $l = 1$ としている．

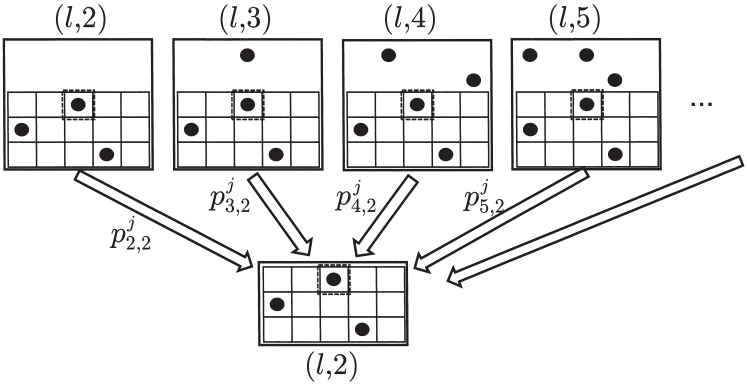


図10　周辺サイズの推移のイメージ

定義域内周辺セルを用いた多重標本寸法指標のうちで第 j 階級に含まれるものを $s^j_{(l,h)}$，拡大周辺セルを用いた調整後の多重標本寸法指標のうちで第 j 階級に含まれるものを $\underline{s}^j_{(l,\underline{h})}$ とすると，中央セルのサイズが l であるものについて推移確率を用いて次のように書くことができる．

$$\begin{pmatrix} s^j_{(l,0)} \\ s^j_{(l,1)} \\ s^j_{(l,2)} \\ \vdots \\ s^j_{(l,h_{\max})} \end{pmatrix} = \begin{pmatrix} p^j_{0,0} & p^j_{1,0} & p^j_{2,0} & \cdots & p^j_{\underline{h}_{\max},0} \\ & p^j_{1,1} & p^j_{2,1} & \cdots & p^j_{\underline{h}_{\max},1} \\ & & p^j_{2,2} & \cdots & p^j_{\underline{h}_{\max},2} \\ & & & & \vdots \\ & & & & p^j_{\underline{h}_{\max},h_{\max}} \end{pmatrix} \begin{pmatrix} \underline{s}^j_{(l,0)} \\ \underline{s}^j_{(l,1)} \\ \underline{s}^j_{(l,2)} \\ \vdots \\ \underline{s}^j_{(l,\underline{h}_{\max})} \end{pmatrix}. \quad (10)$$

ただし

$$s_{(l,h)} = \sum_{j=1}^{11} s^j_{(l,h)} \quad (11)$$

である．

ここで $\underline{h}_{\max} = h_{\max}$ とすると，(10) 式の両辺に推移確率を要素とする行列の逆行列を乗ずることにより，調整後の第 j 階級の多重標本寸法指標 $\underline{s}^j_{(l,\underline{h})}$ を求めることができ，最終的には

$$\underline{s}_{(l,\underline{h})} = \sum_{j=1}^{11} \underline{s}^j_{(l,\underline{h})} \quad (12)$$

により，すべての階級を合わせた調整後の多重標本寸法指標が得られる．

4.3　真のリンク確率の計算方法

Step 3’では，推定された多重母集団寸法指標を用いて，真のリンク確率の期待値を推定する．拡大周辺セルを用いて周辺サイズも調整された多重標本寸法指標を基に推定されているため，多重母集団寸法指標についても周辺セルは拡大周辺セルで，周辺サイズについても調整されたものが得られていると考えられる．

ここで図 9 の（case a），（case b），（case c）について，（5）式に対応する真のリンク確率について考える．なおそれぞれについて，中央セルには個体が 1 個のみ含まれている場合を考える．

（case a）では，確率 1/4 ずつでとる 4 個の領域Dすべてが定義域内に含まれ，真のリンク確率は

$$P_t(\mathbf{a}_i) = 4 \cdot \frac{1}{4}\Big(1 - \frac{9-1}{25-1}\Big)^{\underline{h}} = \Big(1 - \frac{9-1}{25-1}\Big)^{\underline{h}} \tag{13}$$

と計算される．この場合，拡大周辺セルと定義域内周辺セルは等しく，$\underline{h}$ を h と置き換えても結果は変わらないため，(13) 式は方法［H］で用いた (5) 式と等しい．

（case b）では，挿入できるノイズ変数ベクトルが 2 個しかない．そのため領域Dも 2 個しか存在しないが，どちらも定義域内に含まれている．この場合，2 個の領域Dにそれぞれ確率 2/4 を割り振り，真のリンク確率は

$$P_t(\mathbf{a}_i) = 2 \cdot \frac{2}{4}\Big(1 - \frac{9-1}{25-1}\Big)^{\underline{h}} = \Big(1 - \frac{9-1}{25-1}\Big)^{\underline{h}} \tag{14}$$

と計算される．

方法［H］では (14) 式の $\underline{h}$ を h に置き換えた式を用いて真のリンク確率を計算していた．$h \leq \underline{h}$ より，上記のような対処を行わないと真のリンク確率を大きく計算することになり，過大推定の要因になると考えられる．

（case c）では挿入できるノイズ変数ベクトルは 4 個あるものの，前述したように領域Dの定義域内となるセル数は，左下から反時計回りに，9 セル，6 セル，4 セル，6 セルとなる．そのため真のリンク確率は

$$\begin{aligned} P_t(\mathbf{a}_i) \;=\; & \frac{1}{4}\Big(1 - \frac{9-1}{25-1}\Big)^{\underline{h}} + \frac{1}{4}\Big(1 - \frac{6-1}{25-1}\Big)^{\underline{h}} \\ + \; & \frac{1}{4}\Big(1 - \frac{4-1}{25-1}\Big)^{\underline{h}} + \frac{1}{4}\Big(1 - \frac{6-1}{25-1}\Big)^{\underline{h}} \end{aligned} \tag{15}$$

と計算すべきである．ただ，例えば $K = 10$ の場合には領域Dの個数は 2^{10} 個あり，真のリンク確率の期待値の推定まで，多重標本寸法指標の頻度

に対応する値の組を持ち続けなければならない．ここでは $s_{(0,0)}$ 以外の多重標本寸法指標を対象とするものの，それでも（A）一様分布，$K = 10$, $M = 20$ の場合，およそ $2.139 \cdot 10^{11}$ 組にもなる．

そこで，拡大周辺セルのうちで定義域内となる比率を，その周辺セルを構成する通常の大きさの領域Dのうち定義域内となる平均的な比率と見なし，その値で代用することにする．例えば図９（case c）では，中央セルも含めると，定義域内周辺セルは拡大周辺セルの $(4/5)^2 = 16/25$ の大きさであることから，近似的な真のリンク確率を

$$P_t(\mathbf{a}_i) = \left(1 - \frac{16}{25} \cdot \frac{9-1}{25-1}\right)^{\underline{h}} \tag{16}$$

と簡単に求める．例えば $\underline{h} = 2$ の場合，（15）式の値は 0.6159，（16）式の値は0.6188 である．

$s_{(0,0)}$ 以外の多重標本寸法指標としてカウントされる個々の場合について上記の平均的な比率を計算し，Step 1’と同じ区分の11階級の度数分布として情報を記録する．図11は図８と同じ２種類の分布について，上記の度数分布を表したものである．この中には（case a），（case b）のような場合もすべて含まれている．（case b）については，図８の度数分布の情報は h を $\underline{h}$ に増やす調整のために用いられるため，見かけ上の定義域内の比率として記録されていたが，領域Dの大きさに関しては（case b）は（case a）と同様に調整不要のため，図11では比率は1.0として記録されている．そのため両者の度数分布は異なっている．

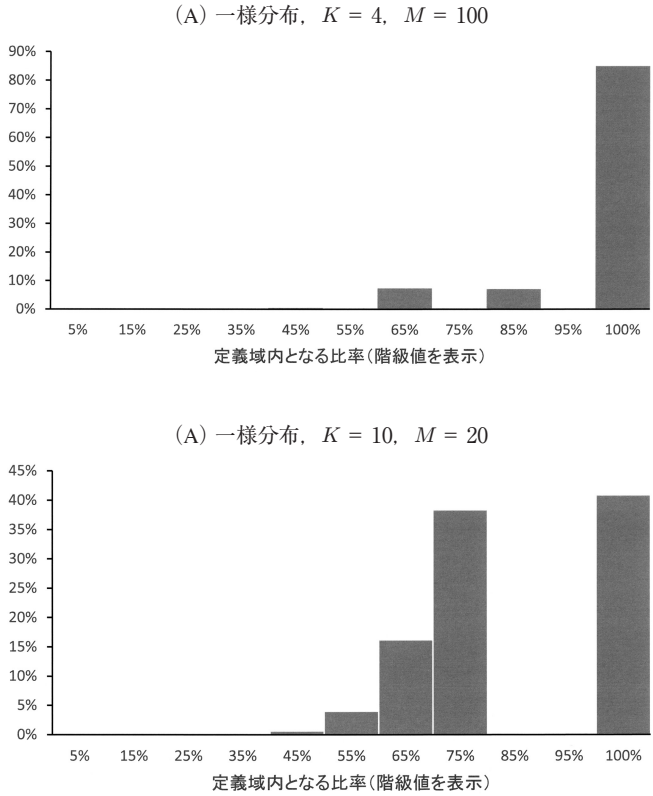


図11　周辺セルのうち定義域内となる比率の度数分布（推定に使用）

前述したように，多重標本寸法指標と多重母集団寸法指標の対応はつかないため，推定された多重母集団寸法指標の個々のサイズの組み合わせ $(l, \underline{h})$ に対して，度数分布の各度数に比例して比率を配分して，近似的な真のリンク確率を重み付け平均として求める．サイズの組み合わせによって配分を変えることはしない．

なお，例えば領域Dが（case b）のように２個のみで，その２個の領域Dの一部が（case c）のように定義域外にはみ出す場合もあるが，その場合には上記の（case b）と（case c）の両方の対処を行うことになる．

4.4　方法［D］における対処方法［D edge］の概要

この節では, キー変数ベクトルが定義域の端付近にある場合の, 方法［D］における対処方法［D edge］について述べる．方法［H edge］とは周辺セルの設定方法が異なるものの手順は同様であるため，概略を述べるにとどめる．

Step 1”　図12のように，定義域外へのはみ出しを許して，領域Dである周辺セルを通常の大きさの拡大周辺セルに統一する．定義域内周辺セルのサイズ h を，拡大周辺セルのサイズ $\underline{h}$ に増やす調整を行い，調整された多重標本寸法指標 $\underline{s}_{(l,\underline{h})}$ を求める．調整には，周辺セルが定義域内となるセルの比率で分けられた11階級別に，（10）式と同様の推移行列を用いる．

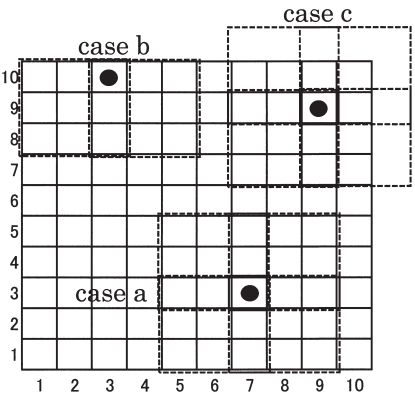


図12　［D edge］において大きさを統一した周辺セル

Step 2”　調整された多重標本寸法指標を基に多重母集団寸法指標 $\underline{S}_{(l,\underline{h})}$ を推定する．

Step 3”　多重母集団寸法指標の推定値 $\hat{\underline{S}}_{(l,\underline{h})}$ を基に真のリンク確率の期待値を推定する．方法［D］では推定された多重母集団寸法指標のうちで $\hat{S}_{(1,0)}$ のみが真のリンク確率の期待値の推定に用いられていたが，［D edge］では中央セルのサイズが $l = 1$ であるすべての場合

に真のリンクの可能性がある．実際，拡大周辺セルを用いたサイズの組が $(1, \underline{h})$ と推定されたとき，それが第 j 階級に属している場合，真のリンクとなる確率は第 j 階級の階級値 m_j を用いて $(1 - m_j)^{\underline{h}}$ と求められる．したがって，真のリンク確率の期待値は，方法［D］の (8) 式に代えて

$$\hat{\mathrm{E}}\big[P_t(\mathbf{x}_i)\big] = \frac{1}{N} w \sum_{\underline{h}=0}^{\underline{h}_{\max}} \hat{S}_{(1,\underline{h})} \sum_{j=1}^{11} f_j (1 - m_j)^{\underline{h}} \tag{17}$$

と推定することができる．ただし f_j は第 j 階級の標本における相対度数である．

なお方法［D edge］では，Step 3”で用いる度数分布は Step 1”で用いるものと同じである．

4.5　対処による結果への影響について

図13，14には，真のリンク確率の期待値の推定値の，真のリンクとなる個体の比率からの差について，前節までの対処を行った場合の結果を加えて示す．［H edge］，［D edge］がそれぞれ［H］，［D］について対処を行った場合の結果である．

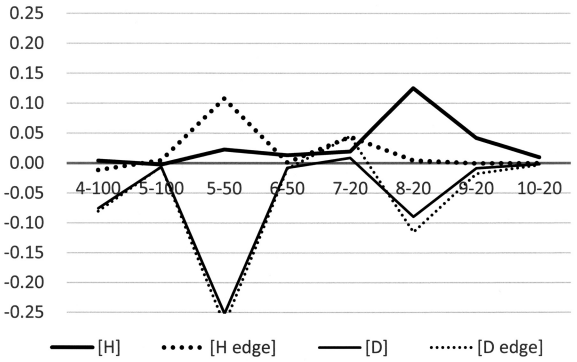


図13　推定値の実際の真のリンク比率からの差（(A) 一様分布）

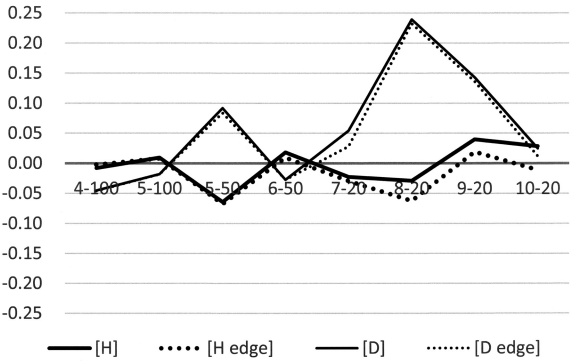


図14　推定値の実際の真のリンク比率からの差（(B) 偏在分布）

一様分布，偏在分布とも，方法［D edge］の推定結果は不安定で方法［D］との違いもほとんど見られない．

これに対して方法［H edge］では（A）一様分布，$K = 8, 9$，$M = 20$では方法［H］で見られた過大推定が解消され改善が図られたように見える．ただ，（A）一様分布，$K = 5$，$M = 50$では推定結果は悪化しており，この結果だけで判断するのは難しい．（B）偏在分布，$K = 8, 9, 10$，$M = 20$でも推定値は方法［H］よりも小さくなり，$M = 20$ の場合は全体的にやや過小推定となっているように見える．偏在分布にもかかわらず，周辺セル内で一様に分布するという仮定を置いているため，真のリンク確率の期待を過小推定してしまうことと，端の処理をしないために過大推定してしまうことが相殺され，方法［H］においては見かけ上偏りの小さい推定結果となっていることが原因と考えられる．この点については今後の検討課題である．

ここで図 4，9 の（case c）について考える．方法［H］では，真のリンク確率は（case a）などと同様に

$$P_t(\mathbf{a}_i) = \left(1 - \frac{9-1}{25-1}\right)^h \tag{18}$$

と求められていた．ここで (16) 式と (18) 式を展開すると，前者は

$$P_t(\mathbf{a}_i) = 1 - \underline{h}\,\frac{16}{25}\cdot\frac{9-1}{25-1} + \frac{\underline{h}(\underline{h}-1)}{2}\Big(\frac{16}{25}\cdot\frac{9-1}{25-1}\Big)^2 - \cdots, \quad (19)$$

後者は

$$P_t(\mathbf{a}_i) = 1 - h\,\frac{9-1}{25-1} + \frac{h(h-1)}{2}\Big(\frac{9-1}{25-1}\Big)^2 - \cdots \quad (20)$$

と書ける．

ここで，拡大周辺セルと定義域内周辺セルの大きさの比，すなわち (16) 式の近似に用いた値の逆数を用いて $\underline{h} = (5/4)^2 h$ とすると，(19) 式と (20) 式について第 2 項までは一致し，第 3 項以降についてもその差は小さいことが分かる．例えば $h = 2$ では 2 つの式の値はそれぞれ 0.4724, 0.4624，$h = 10$ では 2 つの式の値はそれぞれ 0.0235，0.0211 である．実際には，推移行列を用いて $\underline{h}$ を求めたため，上の結果がそのまま当てはまるわけではないが，(case c) のような場合については，方法 [H edge] のような対処を行わないことによる結果への影響は大きくないと考えられる．

そのため，方法 [H edge] を用いる必要性としては (case b) の対処に限られる．なお方法 [D] では (case b) の領域Dは定義域内に収まっているため，この場合についても方法 [D edge] を用いる必要性は認められない．

図15，16には，真のリンク確率の期待値の推定値の，真のリンクとなる個体の比率からの差について，4.3節の対処のうち (case b) のみを行った場合の結果を加えて示す．方法 [H edge 2] が新たに追加した結果で，方法 [D]，[D edge] については省略している．

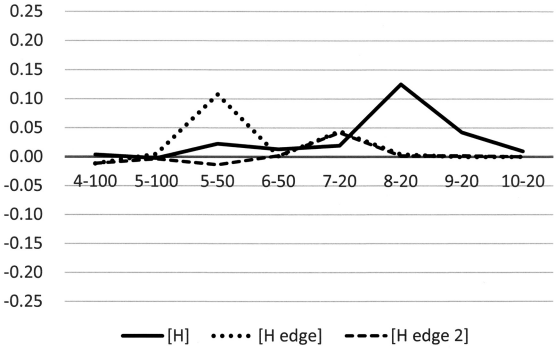


図15　推定値の実際の真のリンク比率からの差（(A) 一様分布）

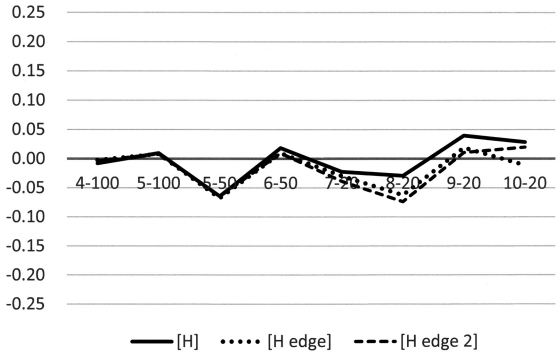


図16　推定値の実際の真のリンク比率からの差（(B) 偏在分布）

一様分布を見ると，方法［H edge］で過大推定となっていた$K = 5$, $M = 50$では方法［H edge 2］の推定値の真の値との差はわずかで，改善が図られたように見える．

一様分布のKとMのそれ以外の組み合わせと偏在分布では，方法［H edge 2］の結果は方法［H edge］の結果とほとんど変わらない．総合的に考えると（case b）のみの対処を行う方法が最も優れているが，今後，理論的な検証を行うか，適用するデータを増やすなどして，さらなる検討を行う必要がある．

5　おわりに

本稿では, 標本調査で得られた個票データを秘匿するために, 個体のキー変数ベクトルにノイズ変数ベクトルを挿入した場合に，リスク評価の指標として真のリンク確率の期待値を用い，その推定のために多重寸法指標を用いる方法について検討を行った．多重寸法指標としては，これまで用いてきた領域Dの和集合Hを周辺セルとする方法［H］に加えて，領域Dそのものを周辺セルとする方法［D］を追加し，人工的ないくつかのデータに適用することによって有効性を比較した.

その結果については既に述べたところだが，検討の過程でいくつかの課題も見つかった.

方法［D］については推定が非常に不安定であった．第２章，第３章でも述べたように，方法［D］では，真のリンク確率の期待値の推定には，推定された多重母集団寸法指標のうちで $\hat{S}_{(1,0)}$ のみが用いられる．表６を見ると，(a) 多重標本寸法指標では $s_{(1,0)} = 1.024 \cdot 10^7$，(b) 多重母集団寸法指標の推定値では $\hat{S}_{(1,0)} = 9.956 \cdot 10^8$ となっている．抽出率は0.01であることから，$s_{(1,0)}$ の 97.25％程度が $\hat{S}_{(1,0)}$ から生み出されており，残りは $\hat{S}_{(1,1)}$, $\hat{S}_{(1,2)}$ から生み出されていると見なすことができる．なお，表６で用いた（A）一様分布，$K = 10$，$M = 20$ では方法［D］でも偏りの小さい推定が行われていた.

一方，方法［D］で過小推定となっている（A）一様分布，$K = 5$, $M = 50$ においては，多重標本寸法指標では $s_{(1,0)} = 316722$，多重母集団寸法指標の推定値では $\hat{S}_{(1,0)} = 3.317 \cdot 10^6$ となっており，10.47％程度しか $\hat{S}_{(1,0)}$ から生み出されていない．一方 $\hat{S}_{(1,1)} = 2.455 \cdot 10^7$ が $\hat{S}_{(1,0)}$ を大きく上回り，$\hat{S}_{(1,1)}$ から生み出されたものが 76.74％程度と見なされ，$\hat{S}_{(1,0)}$ の過小推定が真のリンク確率の期待値の過小推定に直結してしまっていると推測される．$\hat{S}_{(1,0)}$ の推定精度を高めることも重要であるが，$\hat{S}_{(1,0)}$ 以外の多重母集団寸法指標を利用して推定を安定させることについても，今後の

検討課題である.

方法［H］については概ね良好な推定が行われているものの，前述したように,周辺セル内で一様分布と仮定した点については改善の余地がある.周辺セル内の分布は厳密である必要ではなく，個々の個体が，領域Dの内部にあるか外部にあるかを確率的に表現できれば十分である．そのような分布の導入により,偏在分布に対する推定の改善が図られる可能性もあり,この点についても今後の課題となる.

実用化につなげるためには，より大きなK，より小さなMについての検討も進める必要がある．特にMについては，現在の方法で推定が可能な最小値を見極めるとともに，それよりも小さい場合に有効な方法についても別途検討する必要があろう.

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（2021-ISMCRP-2002，2022-ISMCRP-2003）の研究成果に基づくものである.

参考文献

［1］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.
［2］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.
［3］Pitman, J.（1995）．Exchangeable and partially exchangeable random partitions, *Probability Theory and Related Fields*, **102**, 145-148.
［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，**51**(2)，183-197.
［5］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.
［6］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.
［7］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.
［8］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.
［9］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.
［10］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.
［11］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.
［12］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.
［13］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.
［14］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.
［15］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.
［16］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.
［17］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.
［18］佐井至道（2020a）．局所的な分布を用いたリスク評価の実用化に向けた検討，2020年度統計関連学会連合大会予稿集，256.
［19］佐井至道（2020b）．ノイズを挿入した個票データの多重寸法指標を用いたリスク評価－とりうる値の少ないキー変数への対処について－，研究集会「大規模データの公開におけるプライバシー保護の理論と応用」報告資料，1–42.
［20］佐井至道（2021a）．局所的な分布を用いたリスク評価－実用化に向けた検討－，岡山商大論叢，**56**(3)，43–76.
［21］佐井至道（2021b）．ノイズが挿入された個票データに対する多重寸法指標を用いたいくつかのリスク評価方法の比較，2021年度統計関連学会連合大会予稿集，270.
［22］佐井至道（2021c）．個票データに対するいくつかの多重寸法指標を用いたリスク評価方法の比較とその改善，研究集会「大規模データの公開におけるプライバシー保護の理論と応用」報告資料，1–33.
［23］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.
［24］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.
［25］Willenborg, L. and de Waal, T.（1996）．*Statistical Disclosure Control in Practice*, Springer.
［26］Willenborg, L. and de Waal, T.（2001）．*Elements of Statistical Disclosure Control*, Springer.