『岡山商大論叢』（岡山商科大学）
第52巻第１号　2016年６月
Journal of OKAYAMA SHOKA UNIVERSITY
Vol.52　No.１　June 2016

《論　　説》

ノイズが挿入された個票データのリスク評価

佐　井　至　道

Risk Assessment for Microdata Perturbed by Adding Noise

Shido Sai

１　はじめに

個票データを公開することによってプライバシー侵害の可能性が高いと判断された場合，何らかの秘匿措置が施されることが多いが，国内ではカテゴリの併合のような非攪乱的な秘匿を行うことが多い．そのような個票データに対するリスク評価としては寸法指標を用いるのが一般的であり，データが標本調査で得られた場合には，推定された母集団寸法指標を用いることが多い．

これに対して，海外ではノイズの挿入やスワッピングのような攪乱的な秘匿措置が多用されており，近年，国内においても用いられることが多くなってきた．特に，個体（個人や事業所など）を特定するために用いられるキー変数に対してノイズの挿入を行うと，値が変化してしまうため，寸法指標によるリスク評価は意味を持たなくなる．そのため，秘匿後の個体が元の個体にリンクされる確率やそのような個体数などをリスク評価の指

標とすることが多いが，これまでの議論では，個票データの中でリンクされるか否かについて検討するものがほとんどである．

本報告では，標本から得られた個票データのキー変数に対してノイズを挿入した場合のリスク評価について，母集団も考慮に入れた検討を行う．

まず２節では非攪乱的な秘匿措置と攪乱的な秘匿措置について説明するとともに，それぞれの秘匿措置を施した個票データのリスク評価についても簡単に述べる．

３節ではキー変数が連続型の量的変数の場合について検討を行う．まず3.1 節において，リスク評価方法の一つである個体が真のリンクとなる確率について詳しく述べる．次に 3.2 節では，以後の検討で用いるためにキー変数とノイズに典型的な理論分布を設定し，3.3 節以降では，それぞれの分布の組み合わせに対して，キー変数の数，挿入するノイズの大きさなどについて様々な状況を想定し，標本に含まれる個体が母集団において真のリンクとなる確率を求める．

４節ではキー変数が連続型の量的変数以外の場合について考える．4.1 節ではキー変数が離散型の量的変数の場合について，4.2 節ではキー変数が質的変数の場合について，それぞれ個体が真のリンクとなるケースの検討を行う．

５節では，アメリカの2010年センサスの擬似ミクロデータであるPublic Use Microdata Sampleに対して理論分布の結果を用いてリスク評価を行い，その有用性を検証する．

２　非攪乱的な秘匿措置と攪乱的な秘匿措置

個体の情報がレコードして並べられたデータを個票データやミクロデータと呼ぶ．個票データの公開が危険と判断した場合には，個票データに何らかの変更を加えることが多い．このような操作を秘匿措置（disclosure control），あるいは単に秘匿と呼ぶ．秘匿措置についてはWillenborg and de Waal（1996, 2001）が詳しい．

まず嘘をつかない秘匿とも呼ばれる非攪乱的な秘匿措置について説明する．

最も簡単な秘匿措置は，特定されやすい個体のレコードを削除してしまう方法であるが，個体数が大きく減ってしまうとともに，データに偏りが生じる．また高い年齢や珍しい職業など，個体の特定に結びつきやすい一部の値を削除する方法もある．秘匿としては効果的であるものの，データが虫食い状態となり分析は困難になる．

そこで個票データに含まれるすべての個体に対して同一の方法で秘匿措置を施すことが多い．例えば個体の特定に利用されやすい変数を削除する方法が考えられる．また，年齢を５歳階級にまとめたり都道府県をより広い地域にまとめたりするような，数値の丸めやカテゴリの併合という方法も考えられる．さらに，ある値以上あるいはある値以下をまとめるトップコーディング，ボトムコーディングもある．これらの方法は，すべてカテゴリの併合と考えることも可能でglobal recodingと呼ばれる．

非攪乱的な秘匿措置を施した個票データのリスク評価には，寸法指標を用いることが多い．

例えば，世帯主×男性×35歳×… と組み合わせていくと可能性が絞り込まれていき，場合によっては，そのレコードが誰であるかが分かってしまうかもしれない．このように，第三者にとって一般的に既知で，個体を特定するために使用できる変数をキー変数（key variable）と呼ぶ．そしてキー変数の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．セルがJ個あるとして，各セルに１からJまでの番号をつけ，第jセルに入っている個体数を標本ではf_jと表す．このとき$f_j = l$となるセル，すなわちl個の個体が入っているセル数をs_lと表し，標本寸法指標と呼ぶ．定義関数Ⅱ(・)を用いると

$$s_l = \sum_{j=1}^{J} \mathbb{I}(f_j = l) \tag{1}$$

と書くことができる．定義関数$\mathbb{I}(\cdot)$とは（ ）の中が正しければ1を，正しくなければ0を表す関数である．母集団における定義も同様で，f_j，s_lの代わりにF_j，S_lを用いる．特にs_1，S_1は一意な個体数と呼ばれ，リスク評価の指標としてしばしば利用される．なお，寸法指標はSibuya (1993)が用いたsize indexの訳語である．

官庁が実施する統計調査の多くは標本調査で行われている．個票データが標本から得られている場合には，標本の一意な個体数が多くても必ずしもリスクが高いとは言えないため，標本寸法指標から推定された母集団寸法指標の情報を利用するのが一般的である．リスク評価における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，嘘をつく秘匿とも呼ばれる攪乱的な秘匿措置として，データに誤差を加える攪乱（perturbation）や，複数のレコードの一部の変数の値を入れ替えるスワッピング（swapping）がある．スワッピングについては，地域に関する情報を入れ替えることが効果的と言われている．

攪乱的な秘匿については, 国内では嘘をつくことに対する抵抗感が強く，これまで十分な議論がされてこなかった．しかし欧米では古くから一般的に用いられてきており，国内でも全国消費実態調査のデータに誤差を加えたものが，擬似ミクロデータという名前で提供され始めた．キー変数に相当する質的属性の値の組み合わせごとに集計表を作り，同じ組み合わせの個体が少ない場合には値を変更する秘匿措置を施す．また，値の組み合わせごとに量的属性について平均などの統計量を保存する分布を仮定し，それらの分布から乱数を発生させる方法がとられている．擬似ミクロデータについては統計センター (2011) を，データの作成方法については秋山 他 (2012) を参照していただきたい．

キー変数に対して攪乱的な秘匿措置を施すと，キー変数の値の組み合わ

せが他のすべての個体と異なる一意な個体でも，値そのものが変更されているため危険という判断にはならない．そのため寸法指標による議論は意味を持たなくなる．

そこで，秘匿後の個体のすべてのキー変数の値が秘匿前と一致した場合にリンクされたと判断したり，秘匿前の値が秘匿後の値から一定距離以内となる場合に危険と判断したりする方法があり，その場合にはリンクされる割合などをリスクの指標として考える．このような方法は確定的リンケージと呼ばれる．

また，秘匿後の個体のキー変数の値から秘匿前のすべての個体のキー変数の値までの距離を計算して，元の個体と対応付けが可能かどうかを判断する方法（伊藤（2010））もあり，距離計測型リンケージと呼ばれる．

他にも様々な考え方があるが，それらについては伊藤 他（2014）を参照していただきたい．

本報告ではキー変数に対してノイズを挿入する攪乱的な秘匿措置を施した場合に，距離計測型リンケージによってリンクされるか否かを考えるが，その詳細については次節で説明する．

なお，キー変数以外に対して秘匿措置を施すことも考えられるが，本報告ではそのような方法は扱わない．また，攪乱的な秘匿措置を行うと，措置を施していない値についても正しいという確信が持てなくなるため，第三者に個体の特定を諦めさせてしまう心理的な効果もあるが，この点の議論も本報告では行わないこととする．

3　キー変数が連続型の量的変数の場合のリスク評価

3.1　個体のリンク

この報告を通して，個票データは標本調査で得られたものとする．また，母集団の大きさを N，標本の大きさを n，キー変数の個数を K とする．

キー変数としては連続型の量的変数，離散型の量的変数，カテゴリに順序のある質的変数，カテゴリに順序のない質的変数などが考えられるが，

この節ではすべてのキー変数が連続型の量的変数の場合について検討する．

個票データのレコードは，通常それぞれの変数に割り当てられる桁数が決まっているため，調査された元の値が連続型の量的変数であっても記録される値は離散型となる．したがって，この節の結果が実際の個票データにそのまま適用されることはないが，連続型の理論分布には多様性があり，様々な分布について比較検討が可能であることと，得られた結果が桁数が多い離散型の量的変数に対して近似的に適用できることから，最初に検討を行うことにする．

標本に含まれる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．本報告のリスク評価は，ノイズを挿入した標本のある個体について行うため，ノイズの挿入が一部の個体の場合にも適用できる．

標本の i 番目（$i = 1, 2, \ldots, n$）の個体の K 個のキー変数からなるベクトルと，そのベクトルに挿入するノイズのベクトルを

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

$$\mathbf{e}_i = (e_{i,1}, e_{i,2}, \ldots, e_{i,K})' \tag{3}$$

として，母集団の i 番目（$i = 1, 2, \ldots, N$）の個体の K 個のキー変数からなるベクトルを

$$\mathbf{a}_i = (a_{i,1}, a_{i,2}, \ldots, a_{i,K})' \tag{4}$$

とする．また，$\mathbf{x}_i$ に対応する母集団のキー変数ベクトルを$\mathbf{a}_{i'}$とする．

２つの個体のキー変数ベクトルまたはノイズを挿入したキー変数ベクトルの距離を $\mathrm{d}(\cdot, \cdot)$ と表すとき，

$$\mathrm{d}(\mathbf{x}_i + \mathbf{e}_i, \mathbf{a}_j) \leq \mathrm{d}(\mathbf{x}_i + \mathbf{e}_i, \mathbf{x}_i) \tag{5}$$

となる $\mathbf{a}_j$（$j \neq i'$）が少なくとも１つ存在すれば，キー変数にノイズを挿入した標本の 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 $\mathbf{a}_j$ 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．なお本報告では，一部の場合を除いて，距離としてユークリッド距離を考える．一般にキー変数ごとに散布度は異なるが，例えば基準化する

などして，ユークリッド距離で個体の距離が測れるものとする．

標本内での議論も同様である．

$$\mathrm{d}(\mathbf{x}_i + \mathbf{e}_i, \mathbf{x}_j) \leq \mathrm{d}(\mathbf{x}_i + \mathbf{e}_i, \mathbf{x}_i) \tag{6}$$

となる $\mathbf{x}_j$（$j \neq i$）が少なくとも 1 つ存在すれば，キー変数にノイズを挿入した標本の i 番目の個体について，標本内で間違ったリンクが発生し，そうでない場合には標本内で真のリンクが発生したと考える．

キー変数が 1 個と 2 個の簡単な例を考える．

キー変数が 1 個で $x_{i,1} = 120$ に挿入したノイズが $e_{i,1} = 4$ とする．このとき母集団に $a_{j,1} = 125$ である個体が存在すると，

$$\mathrm{d}(\mathbf{x}_i + \mathbf{e}_i, \mathbf{a}_j) = \mathrm{d}(120 + 4, 125) = 1,$$
$$\mathrm{d}(\mathbf{x}_i + \mathbf{e}_i, \mathbf{x}_i) = \mathrm{d}(120 + 4, 120) = 4$$

より (5) 式が成り立ち，母集団内で間違ったリンクが発生したと考えられる．なお，(5) 式を満たす個体が複数ある場合，最終的には (5) 式の左辺の値が最も小さい個体にリンクされるが，ここでは (5) 式を満たす個々の個体について間違ったリンクが発生したと表現する。

キー変数が 2 個で，$\mathbf{x}_i = (120, 100)'$ に挿入したノイズが $\mathbf{e}_i = (4, -2)'$ とする．このとき母集団に $\mathbf{a}_j = (121, 101)'$ である個体が存在すると，

$$\mathrm{d}(\mathbf{x}_i + \mathbf{e}_i, \mathbf{a}_j) = \mathrm{d}((120 + 4, 100 - 2)', (121, 101)') = (3^2 + 3^2)^{1/2} = \sqrt{18},$$
$$\mathrm{d}(\mathbf{x}_i + \mathbf{e}_i, \mathbf{x}_i) = \mathrm{d}((120 + 4, 100 - 2)', (120, 100)') = (4^2 + 2^2)^{1/2} = \sqrt{20}$$

より (5) 式が成り立ち，母集団内で間違ったリンクが発生したと考えられる．

母集団内で真のリンクが発生すれば，標本内でも真のリンクが発生することは明らかである．真のリンクが発生するとリスクが高いと考えると，標本内で考えるよりも母集団内で考えた場合の方がリスクは低下する．本報告では，リスク評価の指標として標本の個体が母集団内で真のリンクとなる確率，または標本の個体のうち母集団内で真のリンクとなる個数を考える．

3.2 キー変数とノイズの分布

個体が真のリンクとなる確率が，キー変数が従う連続型分布やキー変数に挿入するノイズが従う連続型分布とどのような関係があるか検討するため，次の6通りの理論分布の組み合わせを考える．キー変数が従う分布をA，ノイズが従う分布をBとしたときA + Bのように表す．

(C1) 一様分布＋一様分布
(C2) 一様分布＋二値分布
(C3) 正規分布＋正規分布
(C4) 正規分布＋一様分布
(C5) 片側三角分布＋一様分布
(C6) 2層一様分布＋一様分布

このうち，(C1) ～ (C3) については，キー変数，ノイズとも多変量分布を考えるが，(C4) ～ (C6) については (C1)，(C3) との比較を目的にしており，1変量分布のみについて考える．

なお二値分布は二点分布などとも呼ばれているが，本報告では二値分布と表記する．

3.3 (C1) 一様分布＋一様分布についての検討

まず，キー変数とそれに挿入するノイズが，ともに独立な一様分布に従う (C1) について考える．

母集団の i 番目の個体の k 番目のキー変数の値 $a_{i,k}$ が他の個体と独立に，また同じ個体の他のキー変数とも独立に $a_{i,k} \sim \mathrm{U}(0,1)$ のように区間 $[0,1]$ の一様分布に従うとする．したがって，標本の i 番目の個体の k 番目のキー変数の値 $x_{i,k}$ も他の個体と独立に，また同じ個体の他のキー変数とも独立に $x_{i,k} \sim \mathrm{U}(0,1)$ のように一様分布に従う．

次に，標本の i 番目の個体の k 番目のキー変数に挿入するノイズ $e_{i,k}$ は，すべてのキー変数と独立に，また他の個体のノイズや同じ個体の他のノイズとも独立に $e_{i,k} \sim \mathrm{U}(-\frac{1}{2}c, \frac{1}{2}c)$ のように一様分布に従うとする．

ここで c は正の実数である．各キー変数の標準偏差は $\mathrm{SD}(x_{i,k}) = \frac{1}{\sqrt{12}}$，各ノイズの標準偏差は $\mathrm{SD}(e_{i,k}) = \frac{1}{\sqrt{12}}c$ となり，c はノイズのキー変数に対する相対的な散布度と考えられる．本報告を通して，c の値を10^{-1}, $10^{-2}, \ldots, 10^{-6}$の 6 通りに設定する．

なお，$x_{i,k}$ が 0 または 1 に近いとき，ノイズを挿入した値 $x_{i,k} + e_{i,k}$ が区間 $[0, 1]$ の外となってしまう可能性がある．これを避けるため次のような端点処理を行う．

$$e_{i,k} \sim \mathrm{U}\big(0, \frac{1}{2}c\big) \quad (0 \leq x_{i,k} < \frac{1}{2}c \text{ のとき}), \tag{7}$$

$$e_{i,k} \sim \mathrm{U}\big(-\frac{1}{2}c, 0\big) \quad (1 - \frac{1}{2}c < x_{i,k} \leq 1 \text{ のとき}). \tag{8}$$

ノイズを挿入する前後の平均はともに $\frac{1}{2}$ で，上の処理を行っても平均は $\frac{1}{2}$ に保たれるが，標準偏差などは変化してしまうため，実データへの応用の際にはノイズの挿入方法を含めて注意が必要となる．

標本の i 番目の個体のキー変数ベクトル $\mathbf{x}_i$，ノイズのベクトル $\mathbf{e}_i$ が得られたとき，母集団内のキー変数ベクトル $\mathbf{a}_j$ である個体が間違ったリンクとなる確率 $p(\mathbf{x}_i, \mathbf{a}_j)$ は，$\mathbf{a}_j$ が各辺の長さが 1 である K 次元超立方体内の中心（$x_{i,1} + e_{i,1}, x_{i,2} + e_{i,2}, \ldots, x_{i,K} + e_{i,K}$），半径 $r_i = (\sum_{k=1}^{K} e_{i,k}^2)^{1/2}$ である超球の内部および表面に存在する確率であり，超球の体積に等しい．この超球の内部および表面のような間違ったリンクを生ずる部分を領域 D と呼ぶことにする．

例えば，$K = 1$ のときには超球は線分となり，その長さ $2|e_{i,1}|$ が $p(\mathbf{x}_i, \mathbf{a}_j)$ となる．また $K = 2$ のときには超球は半径 $r_i = (e_{i,1}^2 + e_{i,2}^2)^{1/2}$ の円となり $p(\mathbf{x}_i, \mathbf{a}_j) = \pi(e_{i,1}^2 + e_{i,2}^2)$ となる．なお，$\mathbf{x}_i$，$\mathbf{e}_i$，$\mathbf{a}_j$，$e_{i,k}$ は i, j によらないため，それぞれ $\mathbf{x}_*$，$\mathbf{e}_*$，$\mathbf{a}_*$，$e_{*,k}$ と表すことがある．

$x_{i,k}$ が 0 または 1 に近いとき，(7)，(8) 式のような措置を行っても，超球の一部が超立方体の外となる可能性があるが，この点については無視する．したがって間違ったリンクの確率を若干高めに，真のリンクの確率を

若干低めに見積もることになる.

$\mathbf{x}_i$, $\mathbf{e}_i$ が得られたとき，この個体が母集団内で真のリンクとなる確率 P_t は，$\mathbf{x}_i$ 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を持つ個体が間違ったリンクとならない確率

$$\begin{aligned} P_t &= \prod_{j\,(j\neq i')} \{1-p(\mathbf{x}_i,\mathbf{a}_j)\} \\ &= \{1-p(\mathbf{x}_*,\mathbf{a}_*)\}^{N-1} \end{aligned} \tag{9}$$

である．$\mathbf{x}_i$, $\mathbf{e}_i$ について期待値をとることによって，標本の個体が母集団内で真のリンクとなる確率

$$\mathrm{E}(P_t) = \mathrm{E}\big[\{1-p(\mathbf{x}_*,\mathbf{a}_*)\}^{N-1}\big] \tag{10}$$

が得られるが，各キー変数が一様分布に従うため P_t は $\mathbf{x}_i$ によらず，$\mathbf{e}_i$ についてのみ期待値をとることになる.

例えば $K=1$ のときは，$p\,(\mathbf{x}_*,\mathbf{a}_*) = 2|e_{*,1}| \sim \mathrm{U}(0,c)$ より

$$\begin{aligned} \mathrm{E}(P_t) &= \int_0^c \{1-p(\mathbf{x}_*,\mathbf{a}_*)\}^{N-1}\frac{1}{c}\,dp \\ &= \frac{1}{cN}\big\{1-(1-c)^N\big\} \end{aligned} \tag{11}$$

と解析的に求められる.

$K\geq 2$ のときには計算が困難になるため，乱数を用いて $\{1-p(\mathbf{x}_*,\mathbf{a}_*)\}^{N-1}$ を M 回発生させ，その平均をとるシミュレーション実験を行う．$K=1,2,3,4$ の場合の結果を表1に示す．それぞれの K について，c と N（または n）の組み合わせに対する $\mathrm{E}(P_t)$ の値を示す．母集団内で真のリンクとなることと標本内で真のリンクになることは集団の大きさが異なるだけである．なお，$K=2,3,4$ についてはシミュレーション実験の結果で，$M=2.5\cdot 10^8$ である.

$K=1$, $c=10^{-2}$, $n=10^2$, $N=10^5$ の場合を考えると，標本では $\mathrm{E}(P_t)=0.6340$ より，標本の個体100個のうちで標本内で真のリンクとなる個数の期待値は $n\cdot\mathrm{E}(P_t)=100\cdot 0.6340=63.40$ となるが，母集団では $\mathrm{E}(P_t)=0.0010$ より，標本の個体のうちで母集団内で真のリンクとなる個数の

期待値は $n \cdot \mathrm{E}(P_t) = 100 \cdot 0.0010 = 0.10$ となる．したがって，母集団を考慮に入れてリスク評価すると，標本のみでリスク評価するよりかなり安全と判断されることが分かる．

c が小さくなると $\mathrm{E}(P_t)$ の値は増加し，K が大きくなると $\mathrm{E}(P_t)$ の値は増加する．K が大きいときには N が大きくても $\mathrm{E}(P_t)$ の値が大きくなるため，$\mathrm{E}(P_t)$ の値を小さく抑えるためには c として大きい値を設定する必要がある．実際の個票データでは K は 4 より大きいことが多く，個票データを安全にするためには大きいノイズを入れる必要があることが分かる．

なお，$K = 1$ では $(n, N) = (10^{-1}, 10), (10^{-2}, 10^2), (10^{-3}, 10^3), \ldots$ のよ

表 1：個体が真のリンクとなる確率（(C1) 一様分布＋一様分布）

$K = 1$

$c \setminus N$	10	10^2	10^3	10^4	10^5	10^6	10^7	10^8
10^{-1}	0.6513	0.1000	0.0100	0.0010	0.0001	0.0000	0.0000	0.0000
10^{-2}	0.9562	0.6340	0.1000	0.0100	0.0010	0.0001	0.0000	0.0000
10^{-3}	0.9955	0.9521	0.6323	0.1000	0.0100	0.0010	0.0001	0.0000
10^{-4}	0.9996	0.9951	0.9517	0.6321	0.1000	0.0100	0.0010	0.0001
10^{-5}	1.0000	0.9995	0.9950	0.9516	0.6321	0.1000	0.0100	0.0010
10^{-6}	1.0000	1.0000	0.9995	0.9950	0.9516	0.6321	0.1000	0.0100

$K = 2$

$c \setminus N$	10	10^2	10^3	10^4	10^5	10^6	10^7	10^8
10^{-1}	0.9542	0.6258	0.1000	0.0100	0.0010	0.0001	0.0000	0.0000
10^{-2}	0.9995	0.9948	0.9496	0.6240	0.1000	0.0100	0.0010	0.0001
10^{-3}	1.0000	0.9999	0.9995	0.9948	0.9495	0.6240	0.1000	0.0100
10^{-4}	1.0000	1.0000	1.0000	0.9999	0.9995	0.9948	0.9495	0.6240
10^{-5}	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9948
10^{-6}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999

$K = 3$

$c \setminus N$	10	10^2	10^3	10^4	10^5	10^6	10^7	10^8
10^{-1}	0.9948	0.9453	0.6067	0.0999	0.0100	0.0010	0.0001	0.0000
10^{-2}	1.0000	0.9999	0.9994	0.9943	0.9448	0.6066	0.0999	0.0100
10^{-3}	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9943	0.9448
10^{-4}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
10^{-5}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10^{-6}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

$K = 4$

$c \setminus N$	10	10^2	10^3	10^4	10^5	10^6	10^7	10^8
10^{-1}	0.9994	0.9935	0.9377	0.5828	0.0996	0.0100	0.0010	0.0001
10^{-2}	1.0000	1.0000	1.0000	0.9999	0.9993	0.9935	0.9377	0.5828
10^{-3}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
10^{-4}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10^{-5}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10^{-6}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

うに左上から右下へ斜めに並ぶ $\mathrm{E}(P_t)$ の値がほぼ等しい．これは，領域 D である線分の長さが c の値に比例するためで，c が 0.1 倍されたとき，母集団の個体数が10倍となれば，領域 D の中に個体が存在する確率がほぼ一定に保たれるためである．

一般の K では，c が 0.1 倍されたとき領域 D の体積は 10^{-K}倍となるため，母集団の個体数が 10^K倍となれば，領域 D の中に個体が存在する確率がほぼ一定に保たれ，$\mathrm{E}(P_t)$ の値もほぼ一定に保たれる．この性質を利用すれば，様々な設定について $\mathrm{E}(P_t)$ の値を容易に試算することができる．

3.4 （C2）一様分布＋二値分布についての検討

この節では，キー変数が独立な一様分布に従い，それに挿入するノイズが独立な二値分布に従う（C2）について考える．

（C1）と同様に，母集団と標本のキー変数ベクトル $a_{i,k}$，$x_{i,k}$ が i,k について独立に一様分布 $\mathrm{U}(0,1)$ に従うとする．またノイズ $e_{i,k}$ は，すべてのキー変数と独立に，また i,k について独立に２つの値 $-\frac{1}{2}m_1c$ と $\frac{1}{2}m_1c$ を確率$\frac{1}{2}$ずつでとる二値分布に従うとする．

ここで $m_1=\frac{1}{\sqrt{3}}$ は，ノイズのキー変数に対する相対的な散布度を（C1）に合わせるための係数である．実際，$\mathrm{SD}(e_{i,k})=\frac{1}{2}m_1c=\frac{1}{\sqrt{12}}c=c\cdot\mathrm{SD}(x_{i,k})$ となるため，c がノイズとキー変数の標準偏差の比となる．

なお，$x_{i,k}$ が 0 または 1 に近いときの端点処理は，

$$e_{i,k}=\frac{1}{2}m_1c\ \ (0\le x_{i,k}<\frac{1}{2}m_1c\ \text{のとき}),$$
$$e_{i,k}=-\frac{1}{2}m_1c\ \ (1-\frac{1}{2}m_1c<x_{i,k}\le 1\ \text{のとき})$$

とする．

領域 D は（C1）と同様に超球の内部および表面となるが，$|e_{i,k}|\equiv\frac{1}{2}m_1c$ であるため超球の半径は常に

$$r_i = \Big(\sum_{k=1}^{K} e_{i,k}^2\Big)^{\frac{1}{2}} = \frac{1}{2}\sqrt{K} m_1 c$$

となり，$p(\mathbf{x}_i, \mathbf{a}_j)$，$P_t$ の値は $\mathbf{x}_i$，$\mathbf{e}_i$ の値によらず一定である．

$K = 1, 2, 3, 4$ の場合の $\mathrm{E}(P_t) = P_t$ の値を表２に示す．また，$c = 10^{-1}$ の場合の $K = 1, 2, \ldots, 10$ の結果を表３に示す．

K，c の値にかかわらず，$\mathrm{E}(P_t)$ の値が大きい場合には（C1）との差は小さいが，$\mathrm{E}(P_t)$ の値が小さい部分では（C1）の値よりも極めて小さくなる．例えば，$K = 2$，$c = 10^{-3}$，$N = 10^7$ のとき，（C1）では $\mathrm{E}(P_t) = 0.1000$ であるが（C2）では $\mathrm{E}(P_t) = 0.0053$ である．

表２：個体が真のリンクとなる確率（（C2）一様分布＋二値分布）

	$c \setminus N$	10	10^2	10^3	10^4	10^5	10^6	10^7	10^8
$K=1$	10^{-1}	0.5855	0.0028	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	10^{-2}	0.9492	0.5637	0.0031	0.0000	0.0000	0.0000	0.0000	0.0000
	10^{-3}	0.9948	0.9444	0.5616	0.0031	0.0000	0.0000	0.0000	0.0000
	10^{-4}	0.9995	0.9943	0.9440	0.5614	0.0031	0.0000	0.0000	0.0000
	10^{-5}	0.9999	0.9994	0.9942	0.9439	0.5614	0.0031	0.0000	0.0000
	10^{-6}	1.0000	0.9999	0.9994	0.9942	0.9439	0.5614	0.0031	0.0000
$K=2$	$c \setminus N$	10	10^2	10^3	10^4	10^5	10^6	10^7	10^8
	10^{-1}	0.9539	0.5947	0.0053	0.0000	0.0000	0.0000	0.0000	0.0000
	10^{-2}	0.9995	0.9948	0.9490	0.5924	0.0053	0.0000	0.0000	0.0000
	10^{-3}	1.0000	0.9999	0.9995	0.9948	0.9490	0.5924	0.0053	0.0000
	10^{-4}	1.0000	1.0000	1.0000	0.9999	0.9995	0.9948	0.9490	0.5924
	10^{-5}	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9948
	10^{-6}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
$K=3$	$c \setminus N$	10	10^2	10^3	10^4	10^5	10^6	10^7	10^8
	10^{-1}	0.9953	0.9495	0.5926	0.0053	0.0000	0.0000	0.0000	0.0000
	10^{-2}	1.0000	0.9999	0.9995	0.9948	0.9490	0.5924	0.0053	0.0000
	10^{-3}	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9948	0.9490
	10^{-4}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	10^{-5}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	10^{-6}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$K=4$	$c \setminus N$	10	10^2	10^3	10^4	10^5	10^6	10^7	10^8
	10^{-1}	0.9995	0.9946	0.9467	0.5779	0.0042	0.0000	0.0000	0.0000
	10^{-2}	1.0000	1.0000	1.0000	0.9999	0.9995	0.9945	0.9466	0.5779
	10^{-3}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	10^{-4}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	10^{-5}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	10^{-6}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

表３：個体が真のリンクとなる確率（(C2) 一様分布＋二値分布）

$c=10^{-1}$ $K \setminus N$	10	10^2	10^3	10^4	10^5	10^6	10^7	10^8
1	0.5855	0.0028	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.9539	0.5947	0.0053	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.9953	0.9495	0.5926	0.0053	0.0000	0.0000	0.0000	0.0000
4	0.9995	0.9946	0.9467	0.5779	0.0042	0.0000	0.0000	0.0000
5	0.9999	0.9994	0.9941	0.9427	0.5544	0.0027	0.0000	0.0000
6	1.0000	0.9999	0.9994	0.9936	0.9374	0.5242	0.0016	0.0000
7	1.0000	1.0000	0.9999	0.9993	0.9929	0.9309	0.4886	0.0008
8	1.0000	1.0000	1.0000	0.9999	0.9992	0.9920	0.9230	0.4486
9	1.0000	1.0000	1.0000	1.0000	0.9999	0.9991	0.9910	0.9136
10	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9990	0.9898

これは，(C1) では超球の半径 r_i が変動するため，半径が極端に小さいときには，N が大きくても超球の内部に個体が存在しない確率が高いが，(C2) では超球の半径が一定であるため，N がある程度大きくなると，超球の内部に個体が存在する可能性が高くなることが理由である．

3.5 (C3) 正規分布＋正規分布についての検討

この節では初めに，キー変数とそれに挿入するノイズが，ともに独立な正規分布に従う (C3) について考える．

母集団と標本のキー変数ベクトル $a_{i,k}$，$x_{i,k}$ が i,k について独立に $a_{i,k}\sim \mathrm{N}(0,1)$, $x_{i,k}\sim \mathrm{N}(0,1)$ のように標準正規分布に従うとする．またノイズ $e_{i,k}$ はすべてのキー変数と独立に，また i,k について独立に $e_{i,k}\sim \mathrm{N}(0,c^2)$ のように正規分布に従うとする．

領域 D は (C1) と同様に，中心 $(x_{i,1}+e_{i,1}, x_{i,2}+e_{i,2}, \ldots, x_{i,K}+e_{i,K})$，半径 $r_i=(\sum_{k=1}^{K} e_{i,k}^2)^{1/2}$ の超球の内部および表面となり，この領域に多変量正規分布に従う母集団の $\mathbf{a}_j$ が存在する確率が $p(\mathbf{x}_i,\mathbf{a}_j)$，母集団の $N-1$ 個の個体が領域 D 内に存在しない確率が P_t となる．

ここでは $\mathrm{E}(P_t)$ の値をシミュレーション実験によって求める．まず，$\mathbf{x}_*$ と $\mathbf{e}_*$ のベクトルを多変量正規乱数によって作成する．次に，多変量正規乱数 $\mathbf{a}_*$ を $N-1$ 個発生させ，領域 D に 1 個も存在しないかどうかを観測する．この手順を M 回繰り返し，領域 D 内に 1 個も存在しないことが

表４：個体が真のリンクとなる確率（(C3) 正規分布＋正規分布）

$K=1$

$c \setminus N$	10	10^2	10^3	10^4	10^5
10^{-1}	0.7004	0.1793	0.0251	0.0030	0.0004
10^{-2}	0.9610	0.6848	0.1785	0.0250	0.0030
10^{-3}	0.9960	0.9571	0.6834	0.1785	0.0250
10^{-4}	0.9996	0.9956	0.9568	0.6833	0.1785
10^{-5}	1.0000	0.9996	0.9955	0.9568	0.6833
10^{-6}	1.0000	1.0000	0.9996	0.9955	0.9567

$K=2$

$c \setminus N$	10	10^2	10^3	10^4	10^5
10^{-1}	0.9578	0.6966	0.2410	0.0464	0.0069
10^{-2}	0.9995	0.9950	0.9531	0.6932	0.2397
10^{-3}	1.0000	0.9999	0.9995	0.9950	0.9531
10^{-4}	1.0000	1.0000	1.0000	0.9999	0.9995
10^{-5}	1.0000	1.0000	1.0000	1.0000	1.0000
10^{-6}	1.0000	1.0000	1.0000	1.0000	1.0000

$K=3$

$c \setminus N$	10	10^2	10^3	10^4	10^5
10^{-1}	0.9948	0.9476	0.6967	0.2851	0.0697
10^{-2}	1.0000	0.9999	0.9994	0.9941	0.9461
10^{-3}	1.0000	1.0000	1.0000	1.0000	0.9999
10^{-4}	1.0000	1.0000	1.0000	1.0000	1.0000
10^{-5}	1.0000	1.0000	1.0000	1.0000	1.0000
10^{-6}	1.0000	1.0000	1.0000	1.0000	1.0000

$K=4$

$c \setminus N$	10	10^2	10^3	10^4	10^5
10^{-1}	0.9993	0.9930	0.9389	0.6951	0.3176
10^{-2}	1.0000	1.0000	1.0000	0.9999	0.9993
10^{-3}	1.0000	1.0000	1.0000	1.0000	1.0000
10^{-4}	1.0000	1.0000	1.0000	1.0000	1.0000
10^{-5}	1.0000	1.0000	1.0000	1.0000	1.0000
10^{-6}	1.0000	1.0000	1.0000	1.0000	1.0000

起こる比率を求め，$\mathrm{E}(P_t)$ の近似値とする．

$K=1,2,3,4$ の場合の結果を表４に示す．なお，通算で発生する正規乱数の個数を一定とするため，繰り返しの回数 M を$1.250 \cdot 10^3 \sim 5 \cdot 10^5$と変化させている．また，$N$ の最大値は10^5とする．

傾向はキー変数，ノイズとも一様分布に従う（C1）と同じであるが，特に $\mathrm{E}(P_t)$ が小さい部分では，値は（C3）の方が若干大きい．

次に，分布の変量間に相関がある場合（C3’とする）について検討する．

$K=2$ として，母集団と標本のキー変数ベクトル $(a_{i,1}, a_{i,2})'$，$(x_{i,1}, x_{i,2})'$ が i について独立に $(a_{i,1}, a_{i,2})' \sim \mathrm{N}(\mathbf{0}, \Sigma_1)$，$(x_{i,1}, x_{i,2})' \sim \mathrm{N}(\mathbf{0}, \Sigma_1)$ のように，相関係数 r の２変量正規分布に従うとする．ただし，

$$\Sigma_1 = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \tag{12}$$

である．またノイズベクトル $(e_{i,1}, e_{i,2})'$ はすべてのキー変数と独立に，また i について独立に $(e_{i,1}, e_{i,2})' \sim \mathrm{N}(\mathbf{0}, c^2\Sigma_1)$ のように，相関係数 r の2変量正規分布に従うとする．

ここで領域 D として（C3）と同様に，中心 $(x_{i,1} + e_{i,1}, x_{i,2} + e_{i,2})$，半径 $r_i = (e_{i,1}^2 + e_{i,2}^2)^{1/2}$ の円の内部および表面とした場合の，$\mathrm{E}(P_t)$ の近似値をシミュレーション実験によって求める．$c = 10^{-1}$ のときのいくつかの r の値に対する結果を表5に示す．繰り返しの回数 M を $2.5 \cdot 10^3 \sim 2.5 \cdot 10^5$ と変化させている．

（C3）のように相関がない場合，2変量 $(a_{i,1}, a_{i,2})'$，$(x_{i,1}, x_{i,2})'$，$(e_{i,1}, e_{i,2})'$ の確率密度が一定の曲線はそれぞれ真円となるが，（C3'）のように相関がある場合，確率密度が一定の曲線は楕円となる．例えば $(x_{i,1}, x_{i,2})'$ の確率密度が一定の曲線は $x_{i,1}^2 - 2rx_{i,1}x_{i,2} + x_{i,2}^2 = \text{const.}$ と書ける．領域 D として同様の楕円の領域を設定すると $\mathrm{E}(P_t)$ の値は（C3）の結果と等しくなる．これは，相関のない真円をすべて楕円にしたことにより，結局，求める値が等しくなるためである．表5の結果は，第三者が相関構造を無視してリンクを行った場合とも言える．

相関係数が高くなるにつれて $\mathrm{E}(P_t)$ の値は小さくなり，$r = 1.0$ に近づくと $K = 1$ の相関がない場合の結果に近づいている．

表5：個体が真のリンクとなる確率（（C3'）正規分布＋正規分布）

$K = 2$
$c = 10^{-1}$

$r \setminus N$	10	10^2	10^3	10^4	10^5
0.0	0.9578	0.6966	0.2410	0.0464	0.0069
0.5	0.9520	0.6781	0.2347	0.0456	0.0069
0.9	0.9142	0.5786	0.2004	0.0413	0.0064
0.999	0.7313	0.2668	0.0842	0.0228	0.0044
0.99999	0.7010	0.1837	0.0343	0.0090	0.0015
一変量 (0.0)	0.7004	0.1793	0.0251	0.0031	0.0004

3.6　(C4) 正規分布＋一様分布についての検討

この節では (C3) との比較のために，キー変数が独立に一変量の正規分布に従い，それに挿入するノイズが独立に一変量の一様分布に従う (C4) について考える．

母集団と標本のキー変数ベクトル $a_{i,1}$，$x_{i,1}$ が i について独立に標準正規分布 $\mathrm{N}(0,1)$ に従うとする．またノイズ $e_{i,1}$ はすべてのキー変数と独立に，また i について独立に $e_{i,1} \sim \mathrm{U}(-\frac{1}{2}m_2 c, \frac{1}{2}m_2 c)$ のように一様分布に従うとする．ここで $m_2 = \sqrt{12}$ は，ノイズのキー変数に対する相対的な散布度を (C1) に合わせる係数である．

シミュレーション実験で得られた $\mathrm{E}(P_l)$ の値を表 6 に示す．繰り返しの回数 M を $2.5 \cdot 10^3 \sim 2.5 \cdot 10^5$ と変化させている．また，N の最大値を 10^5 とする．ノイズが正規分布に従う (C3) の結果と大きな違いはない．

表 6：個体が真のリンクとなる確率（(C4) 正規分布＋一様分布）

$K = 1$

$c \setminus N$	10	10^2	10^3	10^4	10^5
10^{-1}	0.6697	0.1441	0.0191	0.0023	0.0003
10^{-2}	0.9573	0.6530	0.1434	0.0190	0.0023
10^{-3}	0.9956	0.9533	0.6515	0.1434	0.0190
10^{-4}	0.9996	0.9952	0.9530	0.6514	0.1434
10^{-5}	1.0000	0.9995	0.9951	0.9529	0.6514
10^{-6}	1.0000	0.9999	0.9995	0.9951	0.9529

3.7　(C5) 片側三角分布＋一様分布についての検討

この節では (C1) との比較のために，キー変数が独立に一変量の片側三角分布に従い，それに挿入するノイズが独立に一変量の一様分布に従う (C5) について考える．

母集団と標本のキー変数ベクトル $a_{i,1}$，$x_{i,1}$ が i について独立に，密度関数 $f(x) = -2x + 2$ $(0 \leq x \leq 1)$ をもつ片側三角分布に従うとする．標準偏差は $1/\sqrt{18}$ である．またノイズ $e_{i,1}$ はすべてのキー変数と独立に，また

i について独立に $e_{i,1} \sim \mathrm{U}(-\frac{1}{2}m_3 c, \frac{1}{2}m_3 c)$ のように一様分布に従うとする．ここで $m_3 = \sqrt{6}\,/\,3$ は，ノイズのキー変数に対する相対的な散布度を (C1) に合わせる係数である．端点処理は (C1) と同様である．

シミュレーションで得られた $\mathrm{E}(P_t)$ の値を表 7 に示す．繰り返しの回数は $M = 5 \cdot 10^9$ である．キー変数が一様分布に従う (C1) の結果と大きな違いはない．

表 7：個体が真のリンクとなる確率（(C5) 片側三角分布＋一様分布）

$K=1$	$c \setminus N$	10	10^2	10^3	10^4	10^5	10^6	10^7	10^8
	10^{-1}	0.6396	0.1154	0.0121	0.0012	0.0001	0.0000	0.0000	0.0000
	10^{-2}	0.9526	0.6230	0.1150	0.0122	0.0012	0.0001	0.0000	0.0000
	10^{-3}	0.9951	0.9482	0.6214	0.1150	0.0122	0.0012	0.0001	0.0000
	10^{-4}	0.9995	0.9946	0.9478	0.6213	0.1150	0.0122	0.0012	0.0001
	10^{-5}	1.0000	0.9995	0.9946	0.9477	0.6213	0.1150	0.0122	0.0012
	10^{-6}	1.0000	0.9999	0.9995	0.9946	0.9477	0.6213	0.1150	0.0122

3.8 （C6）２層一様分布＋一様分布についての検討

この節では (C1) との比較のために，キー変数が独立に一変量の２層一様分布に従い，それに挿入するノイズが独立に一変量の一様分布に従う (C6) について考える．

母集団と標本のキー変数ベクトル $a_{i,1}$，$x_{i,1}$ が i について独立に，密度関数

$$f(x) = \begin{cases} t & (0 \le x \le s) \\ \frac{1-ts}{1-s} & (s < x \le 1) \end{cases}$$

を持つ２層一様分布に従うとする．ただし s は $0 < s < 1$ の実数，t は正の実数で，後の例では $s = 0.1$，$t = 2.0$ とする．このとき標準偏差は $\sqrt{1.13/12}$ である．またノイズ $e_{i,1}$はすべてのキー変数と独立に，また i について独立に $e_{i,1} \sim \mathrm{U}(-\frac{1}{2}m_4 c, \frac{1}{2}m_4 c)$ のように一様分布に従うとする．ここで $m_4 = \sqrt{1.13} \doteqdot 1.063015$ は，ノイズのキー変数に対する相対的な散布度を (C1) に合わせる係数である．端点処理は (C1) と同様である．また，領域 D が２つの層にまたがる点については無視をする．

解析的に求めた $\mathrm{E}(P_t)$ の値を表 8 に示す．(C1)，(C5) の結果と大きな

違いはない.

表８：個体が真のリンクとなる確率（(C6) ２層一様分布＋一様分布）

$K=1$	$c \setminus N$	10	10^2	10^3	10^4	10^5	10^6	10^7	10^8
	10^{-1}	0.6350	0.0941	0.0094	0.0009	0.0001	0.0000	0.0000	0.0000
	10^{-2}	0.9535	0.6176	0.0941	0.0094	0.0009	0.0001	0.0000	0.0000
	10^{-3}	0.9952	0.9492	0.6160	0.0941	0.0094	0.0009	0.0001	0.0000
	10^{-4}	0.9995	0.9948	0.9487	0.6158	0.0941	0.0094	0.0009	0.0001
	10^{-5}	1.0000	0.9995	0.9947	0.9487	0.6158	0.0941	0.0094	0.0009
	10^{-6}	1.0000	0.9999	0.9995	0.9947	0.9487	0.6158	0.0941	0.0094

3.9　連続型分布の結果の比較

これまで検討した (C1) ～ (C6) の結果を比較する．表 9 に，$K=1$, $c=10^{-1}$，10^{-2}，10^{-3} の場合の$\mathrm{E}(P_t)$ の値を再掲する．

(C1) を基準に考えると, $\mathrm{E}(P_t)$ の値が比較的小さい部分では (C3), (C4) の値の方が大きい傾向がある．これは正規分布では裾に密度が低い部分が存在するため，キー変数の絶対値が大きい場合に間違ったリンクが起きにくいことと，極端に小さいノイズが多く存在することによるものと推測されるが，この点についてはさらなる検討が必要である．

また前述したように，$\mathrm{E}(P_t)$ の値が比較的小さい部分では (C2) の値は他の分布の組み合わせより小さい傾向がある．これは小さいノイズが存在しないためである．リスクを減少させるためには，二値分布などの離散型のノイズの挿入や，絶対値がある程度以上のノイズの挿入などの方法が効果的と考えられる．

表９：個体が真のリンクとなる確率（(C1) ～ (C6) の比較，$K = 1$）

$c = 10^{-1}$ 分布 \ N	10	10^2	10^3	10^4	10^5
(C1) 一様＋一様	0.6513	0.1000	0.0100	0.0010	0.0001
(C2) 一様＋二値	0.5855	0.0028	0.0000	0.0000	0.0000
(C3) 正規＋正規	0.7004	0.1793	0.0251	0.0030	0.0004
(C4) 正規＋一様	0.6997	0.1441	0.0191	0.0023	0.0003
(C5) 三角＋一様	0.6396	0.1154	0.0121	0.0012	0.0001
(C6) 2 層＋一様	0.6350	0.0941	0.0094	0.0009	0.0001

$c = 10^{-2}$ 分布 \ N	10	10^2	10^3	10^4	10^5
(C1) 一様＋一様	0.9562	0.6340	0.1000	0.0100	0.0010
(C2) 一様＋二値	0.9492	0.5637	0.0031	0.0000	0.0000
(C3) 正規＋正規	0.9610	0.6848	0.1785	0.0250	0.0030
(C4) 正規＋一様	0.9573	0.6530	0.1434	0.0190	0.0023
(C5) 三角＋一様	0.9526	0.6230	0.1150	0.0122	0.0012
(C6) 2 層＋一様	0.9535	0.6176	0.0941	0.0094	0.0009

$c = 10^{-3}$ 分布 \ N	10	10^2	10^3	10^4	10^5
(C1) 一様＋一様	0.9955	0.9521	0.6323	0.1000	0.0100
(C2) 一様＋二値	0.9948	0.9444	0.5616	0.0031	0.0000
(C3) 正規＋正規	0.9960	0.9571	0.6834	0.1785	0.0250
(C4) 正規＋一様	0.9956	0.9533	0.6515	0.1434	0.0190
(C5) 三角＋一様	0.9951	0.9482	0.6214	0.1150	0.0122
(C6) 2 層＋一様	0.9952	0.9492	0.6160	0.0941	0.0094

4　キー変数が連続型の量的変数以外の場合のリスク評価

4.1　キー変数が離散型の量的変数の場合のリスク評価

この節では，すべてのキー変数が離散型の量的変数の場合について考える．以下の考え方は，キー変数が質的変数であっても，そのカテゴリに順序がある場合には流用できる．

キー変数ベクトル，ノイズベクトル，距離の定義については 3.1 節と同じである．連続型の量的変数との違いは (5)，(6) 式において等号成立があり得る点である．なお，キー変数ごとに散布度は異なり，２つのキー変数の数値の差の意味は異なるが，ここでは例えば各変数ごとに基準化するなどして，ユークリッド距離で個体の距離が測れるものとする．

図１に $K = 2$ で，$x_{i,1}$，$x_{i,2}$ がそれぞれ 1 ～ 10 の10個ずつの整数値を取りうる離散型の量的変数の場合のイメージを示す．横軸が１番目のキー変数，縦軸が２番目のキー変数で，セル一つ一つが整数値の組み合わせを表

すものとする．まず左図の黒い点はキー変数ベクトル $(x_{i,1}, x_{i,2})' = (4,5)'$ である．またノイズベクトルを挿入した $(x_{i,1} + e_{i,1}, x_{i,2} + e_{i,2}) = (6,6)$ が白い点とする．

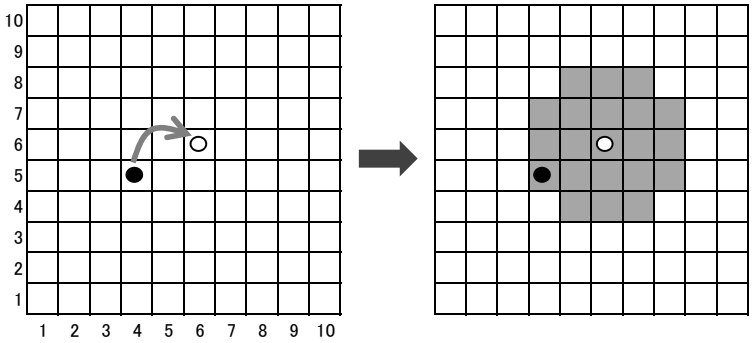

図１：領域Dのイメージ（離散型の量的変数の場合）

このとき右図の円状の網掛け部分が領域 D となり，この領域に元の個体以外のキー変数ベクトルが存在すると間違ったリンクが発生することになる．

例えば $(a_{j,1}, a_{j,2}) = (8,5)'$ の場合には (5) 式の等号が成立して，母集団内で間違ったリンクが発生する．確率1/2で真のリンクが発生するとも考えられるが，第三者は確信を持ってリンクできないため，ここでは間違ったリンクと考えることにする．また，$(a_{j,1}, a_{j,2}) = (4,5)'$ のようにキー変数ベクトルの値が同じ個体が元々複数存在すれば，どのようなノイズを挿入するかに関わらず等号が成立するため，間違ったリンクが発生すると考えられる．

このように離散型の量的変数の場合，特に各変数の取り得る値が少ない場合には，連続型の量的変数の場合と比較して間違ったリンクとなる確率がやや高くなる．

次に図２には，図１と同様に $K = 2$，$(x_{i,1}, x_{i,2})' = (4,5)'$ のとき，横軸

である１番目のキー変数のみにノイズを挿入し，縦軸である２番目のキー変数には挿入しない場合のイメージを示す．

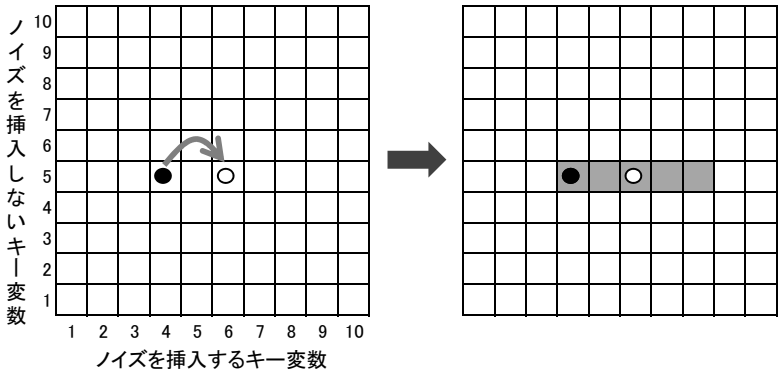

図２：領域Dのイメージ（ノイズ挿入の有無が混在する場合）

このとき第三者がノイズ挿入の有無について正確な情報を持っていれば，２番目のキー変数の値が同じ個体についてリンクを行うため，右図の横に伸びる線状の網掛け部分が領域 D となり，この領域に元の個体以外のキー変数ベクトルが存在すると間違ったリンクが発生すると考えられる．第三者が，２番目のキー変数にノイズを挿入していないという情報を持たなければ，図１と同様の円状の部分が領域 D となる．

なお，キー変数が連続型の量的変数の場合には，あるキー変数にノイズを挿入していないという情報を第三者が持つと，そのキー変数の値のみでリンクできるため，すべての個体が真のリンクとなってしまう．

4.2 キー変数が質的変数の場合のリスク評価

この節では，キー変数に離散型の量的変数とカテゴリに順序のない質的変数が混在する場合について考える．

図３に $K = 2$ で，横軸である１番目のキー変数がカテゴリに順序のない質的変数，縦軸である２番目のキー変数が離散型の量的変数の場合のイ

メージを示す．１番目のキー変数についてはダミー変数を用い，$(x_{i,1}, x_{i,2})'$ $=(4,5)'$として，ノイズベクトルを挿入した$(x_{i,1}+e_{i,1}, x_{i,2}+e_{i,2})=(6,6)$が白い点とする．

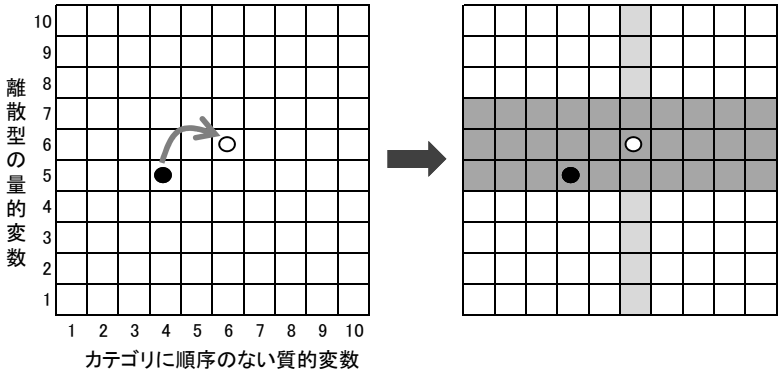

図３：領域Dのイメージ（カテゴリに順序のない質的変数が存在する場合）

このとき，１番目のキー変数である質的変数の値の差に関する距離を考えることができず，質的変数の値が同じか異なるかだけの判断となる．そのため，ノイズを挿入する前後のキー変数ベクトルを比較すると，１番目のキー変数の値は異なり，２番目のキー変数の値は１だけ異なる，という判断となる．したがって，右図の帯状の濃い網掛け部分が領域 D となり，この領域に元の個体以外のキー変数ベクトルが存在すると間違ったリンクが発生することになる．これ以外に，１番目のキー変数の値が６の縦に並ぶ薄い網掛けのセルにキー変数ベクトルが存在する場合には，ノイズ挿入後のベクトルからみて，元のベクトルとどちらが近いかという判断が必要となる．

なお，すべてのキー変数がカテゴリに順序のない質的変数の場合には，各変数が図３の１番目のキー変数のような扱いになるため，０ベクトルでないノイズベクトルが挿入されたすべての個体は間違ったリンクとなってしまう．

5　個票データに対する適用例

この節では，実際の擬似ミクロデータを用いて実験を行う．

データは2010年にアメリカ合衆国で実施されたセンサスの結果から作成されたPublic Use Microdata Sample（U.S. Census Bureau（2014））で，2000年のセンサスから作成されたPublic Use Microdata Sample（U.S. Census Bureau（2003））とは異なり，秘匿措置が施されたフェイスシートの一部のみについて，10%抽出データがダウンロード可能である．ここでは，ワシントン州の672957人分のレコードのデータを用いる．なお，この擬似ミクロデータには既に秘匿措置が施されているため，ここで行う実験はデータが実際に危険かどうかを判断することを目的とするものではない．

まずこのデータから，抽出率1/10，1/100，1/1000で大きさ67296，6730，673のサブサンプルを抽出し，元のデータを合わせて4種類のデータを作る．母集団の大きさNと標本の大きさnとして，例えば$(N, n) = (67296, 6730)$など，4種類のデータの様々な組み合わせを想定して，標本の個体が母集団で真のリンクとなる実際の比率を調べるとともに，分布を用いた推定も行う．

続柄，性別，人種，年齢の4変数をキー変数と考え，標本に相当するデータに含まれる個体の年齢に-1～1のノイズを挿入する．すなわち，各個体の年齢に，-1，0，1のノイズを確率1/3ずつで挿入する．

この擬似ミクロデータでは，年齢が0歳から89歳まではそのままの年齢が記録されており，90歳以上の場合にはトップコーディングが行われ，90歳以上の平均である93歳が記録されている．90歳以上の4120人についてはノイズを挿入しない．また，0歳には確率2/3で0，確率1/3で1のノイズを，89歳には確率1/3で-1，確率2/3で0のノイズを，それぞれ挿入する端点処理を行う．

このとき，標本に含まれる各個体について，母集団内でリンクされるかどうかを判定し，母集団内で真のリンクとなる比率を求める．具体的には，続柄，性別，人種の組み合わせごとに，ノイズ挿入後の年齢からみて，元

の年齢より近いあるいは同じ距離の母集団の年齢がなければ真のリンクと判定する．

結果を表10に示す．例えば，$(N, n) = (67296, 6730)$ のとき，2.348％に相当する158の個体が母集団では真のリンクとなっている．真のリンクとなるか否かの判定は標本の 1 個ずつの個体について行われるため，各 N についての結果は n の値によらずほぼ一定である．

なお，ノイズを挿入する前の一意な個体数は，大きさ673の集団では311（46.21％），大きさ672957の集団では3282人（0.49％）である．

表10：個体が真のリンクとなる比率（PUMS2010）

$n \setminus N$	673	6730	67296	672957
673	0.35067	0.10104	0.02823	0.00446
6730		0.10713	0.02348	0.00342
67296			0.02355	0.00282
672957				0.00279

次に，年齢に離散型一様分布を当てはめて，真のリンクとなる確率を推定する．まず続柄の17のカテゴリごとの年齢分布を見ると，0 歳から高年齢まで広く分布しているものもあるが，偏りがある分布も多い．例えば続柄カテゴリ01は世帯主で，262007人おり全員15歳以上である．また続柄カテゴリ08は（世帯主の）孫で，10757人おり 0 ～ 57歳の間で分布している．

そこで，続柄のカテゴリの特性に応じて，90カテゴリ，75カテゴリ，60カテゴリの離散型一様分布を当てはめる．例えば続柄カテゴリ01の世帯主には 0 ～ 89歳の90カテゴリを，続柄カテゴリ08の孫には 0 ～ 59歳の60カテゴリを設定する．

続柄，性別，人種の各組み合わせについて，標本に含まれる個体数 n' を抽出率で割り，母集団に含まれる同じ組み合わせの個体数を推定する．その値を N' とする．標本の各個体の年齢に確率1/3ずつで -1，0，1 のノイズを挿入することから，例えば続柄に90カテゴリを設定した場合には，標本で続柄，性別，人種のある組み合わせの個体のうち母集団で真のリン

クとなる個体数は，ノイズが 0 と ±1の場合に分けて

$$n'\{\frac{1}{3}(\frac{89}{90})^{N'-1}+\frac{2}{3}(\frac{87}{90})^{N'-1}\} \tag{13}$$

と求められる．なお，90歳以上については，すべての個体が間違ったリンクとなると考える．この手法を用いて推定した母集団で真のリンクとなる確率を表11に示す．表10と同様に母集団と標本の大きさの組み合わせについて推定を行っている．

表11：個体が真のリンクとなる確率の推定値（PUMS2010）

$n \setminus N$	673	6730	67296	672957	6729570
673	0.41941	0.13865	0.01018	0.00000	0.00000
6730		0.15668	0.02867	0.00140	0.00000
67296			0.03329	0.00343	0.00011
672957				0.00345	0.00020

標本のサイズが小さい場合には，続柄，性別，人種の各組み合わせとなる個体数が少なくなるため，推定は不安定である．ここでは年齢に離散型一様分布を当てはめたが，実際の分布は高齢者の度数が少なかったり，特定の年齢層が多かったりしている．そのことが影響しているのかもしれないが，表10の比率より多少大きい値が得られる傾向にある．しかし，比較的近い値が得られていると判断される．

このデータが10％抽出であることから，ワシントン州の人口を $N = 672957{\cdot}10$として推定した母集団で真のリンクとなる比率を，表11の $N = 6729570$の列に示す．例えば $n = 672957$ の場合には，標本の個体が母集団で真のリンクとなる確率は 0.00020，真のリンクとなる個体数の期待値は $672957{\cdot}0.00020 = 134.59$ となる．

6　おわりに

本報告では，個票データに対する秘匿措置のうち，これまで国内で用い

られることが少なかったノイズの挿入を行った場合に，標本のみならず母集団においても真のリンクとなる確率をリスクの指標として検討を行った．

キー変数とノイズが従う分布とも一様分布や正規分布などの典型的な連続型分布についての検討がほとんどであったが，特にキー変数が従う分布による結果の違いはそれほど大きくないこと，ノイズの分布としてあまり小さいノイズを発生しない二値分布のようなものを用いると真のリンクとなる確率が小さくなることがわかった．

また，母集団まで考慮に入れると，標本のみで考えた場合と比較して真のリンクとなる確率は大きく減少することもわかった．

これらの結果は直感的には明らかで，本報告以前も考慮されてきたことであるが，具合的なリスクについては検討されておらず，新たな貢献と考えられる．

キー変数の数が増加すると真のリンクとなる確率は急激に増加することも示されたが，これは第三者が母集団に関する完全な情報を保持している場合を想定しているものである．実際には第三者が保持している情報は限定的で，持っている情報もあやふやな場合が多いであろう．その意味では母集団の個体のキー変数にもノイズを入れるような検討をした方が現実的とも考えられる．このような考え方はMarsh et al.（1991）のように，欧米では初期の研究から用いられている．

キー変数が離散型の量的変数の場合や質的変数の場合については，本報告では概念の紹介や，実例に対する簡単な検討にとどまっている．連続型の量的変数の結果が近似的に用いることのできる場合もあるが，今後のさらなる検討が必要と考えられる．

謝　辞

本論文は，科学研究費補助金（16H02791），統計数理研究所共同利用研究プログラム（27-共研-2030，27-共研-2064，28-共研-2027），国立がん研

究センター研究開発費（25-A-21）の研究成果に基づくものである.

参考文献

［1］秋山裕美，山口幸三，伊藤伸介，星野なおみ，後藤武彦（2012）. 教育用擬似ミクロデータの開発とその利用～平成16年全国消費実態調査を例として～，製表技術参考資料，16，1-42.

［2］独立行政法人統計センター（2011）. 平成16年全国消費実態調査擬似ミクロデータ（microdata）.

［3］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.

［4］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.

［5］Marsh, C., Skinner, C., Arber, S. et al.（1991）. The case for samples of anonymized records from the 1991 census, *Journal of the Royal Statistical Society*, Ser. A, 154（2）, 305-340.

［6］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.

［7］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.

［8］Sibuya, M.（1993）. A random clustering process, *Annals of Institute of Statistical Mathematics,* 45, 459-465.

［9］U.S. Census Bureau（2003）. Census 2000 Public Use Microdata Sample（microdata）, Washington, D. C.

［10］U.S. Census Bureau（2014）. Census 2010 Public Use Microdata Sample（microdata）, Washington, D. C.

［11］Willenborg, L. and de Waal, T.（1996）. *Statistical Disclosure Control in Practice*, Springer.

［12］Willenborg, L. and de Waal, T.（2001）. *Elements of Statistical Disclosure Control*, Springer.