エビデンスに基づいた匿名化

星野　伸明*

平成 25 年 8 月 1 日

Evidence Based Anonymization

Nobuaki Hoshino*

概 要

匿名データや個人情報は、個体識別が可能か否かで法律上区別される。しかしこの区別の方法は不明確で、改善のための明示的議論の対象になっていない。従って本論文は、個体識別可能性の判定方法を明確化する。このような判定に関する既存研究は、個体識別可能性の定量評価について閾値を定める理論を欠く。この点について本論文では、個体識別が起きていないという観測可能な事実に基づいて閾値を推定する。また部分的にしか観測されず定量評価できない情報も、等しいか否かという判断しやすい方法で利用する。このような観測に基づいて意思決定する態度は、エビデンスに基づいた匿名化と呼ぶのがふさわしい。この立場から、匿名データ審査体制の改善点が指摘できる。

Japan Law discriminates Anonymized Data or personal information by discerning that a specific individual is identifiable. The state of being identifiable, however, is not defined, and thus we can not explicitly improve the evaluation of identifiability. Therefore the present article explicates a method to decide whether given data are identifiable or not. The existing literature on this issue lacks the theory of deciding the critical value of measured identifiability; we estimate it based on a fact that identification has not been observed. Also partially observed factors are compared in our decision, which is called evidence based anonymization. This theory leads to institutional improvements on Anonymized Data.

キーワード: 母集団一意, プライバシー, 統計的開示制限.

*金沢大学経済学類, 〒 920-0927, 石川県金沢市角間町, E-mail: hoshino@kenroku.kanazawa-u.ac.jp

1　はじめに

匿名データは、平成 21 年度に四調査（全国消費実態調査、社会生活基本調査、 就業構造基本調査、 住宅・土地統計調査）から提供が開始された。平成 23 年度末現在、国民生活基礎調査や労働力調査の匿名データ提供も決まっている。新しい制度がこのように実績を重ねてきたことは喜ばしい。ただ今後は実績という経験を活かし、制度を継続的に改善する道筋をつけるべきである。特に利用者からのデータ改善要求に応える必要があろう。

匿名データは元の個票を変換（匿名化）して作られる。例えば全国消費実態調査等の匿名データでは、15 歳から 84 歳までの年齢を 5 歳階級別に変換している。また地域情報は「3 大都市圏」及び「その他の地域」の 2 区分に変換している。このような変換により、各歳別の分析や詳細な地域別分析は不可能となる。データ分析において、匿名化は明らかに望ましくない。故に匿名化の緩和は利用者の典型的な要求である。

しかし全ての匿名化を外せるわけではない。匿名データの定義（統計法第 2 条第 12 項）を引用すると、「一般の利用に供することを目的として調査票情報を特定の個人又は法人その他の団体の識別（他の情報との照合による識別を含む。）ができないように加工したもの」である。元の個票（調査票情報）はこの定義を満たすように匿名化されなければ、匿名データとして提供不可能[1]である。従って匿名化は、個体識別が不可能な範囲で少ない方がよい。つまり匿名データの改善の多くは、個体識別が可能か否かという判断を必要とする。

この判断について、総務省政策統括官（統計基準担当）(2011) による「匿名データの作成・提供に係るガイドライン」（以下、ガイドライン）には、審査用資料として「チェックリスト」を作成することが定められている。そして「チェックリストに記載された内容等を基に」、「匿名化処理の妥当性等に係る審査を実施する」とある。参考として世帯調査のチェックリスト（H23/3/28 改正版）の要約を付録 A に収めた。チェックリストは個体識別に関係する要因を記載しているはずである。しかしその使い方は説明されていない。

結局「一律に匿名化の基準を設定することは困難」なので「一橋大学における匿名標本データの試行的提供の事例[2]及び諸外国の統計機関における同様の提供の事例等を参考に」匿名化せよとガイドラインは書く。同様とはどのような事例で、それをいかに参考にしたらよいか。この点についての判断は審査担当者の見識に委ねられている。個体識別可能と不可能の区別は、不明確である。

この区別の明確化、精密化は匿名データに関してだけの課題ではない。いわゆる個人情報保護法において個人情報は「生存する個人に関する情報であって、当該情報に含まれる氏名、生年月日その他の記述等により特定の個人を識別することができるもの（他の情報と容易に照合することができ、それにより特定の個人を識別することができることとなるものを含む。）」と定義される（第 2 条）。このように個体識別が可能か否かを区分の基準とする例は外国法[3]でも見られる。いか

[1]本論文において統計法改正は手段として除外する。しかし個体識別性をデータ提供の基準とするのは必ずしも望ましくない。例えば個体識別されてもデータが悪用されなければよいという主張は妥当かもしれない。

[2]試行的提供の詳細については山口 (2008) を見よ。

[3]例えば U.S. Privacy Act など。U.S. Office of Federal Statistical Policy and Standards (1978, pp.3–5) の解説を見よ。

に匿名化すれば個体識別が不可能かという問題は普遍的である。

ところがこの基本的な問題がないがしろにされている。匿名化についての多くの研究は個体識別の危険性（開示リスク）の測り方は定める。しかし開示リスクの目標値について、せいぜい危険選好に応じてデータの分析価値（有用性）とバランスさせよ[4] としか言わない。このような態度は要素技術の発展には好都合である。しかし匿名データ等の場合、所与の環境において個体識別が不可能か判定したい。多くの研究はそのような要求に応えていない。

本論文では諸要因と個体識別行為の関係をモデル化し、個体識別が可能という状態を要因と関係づける。このようなモデル化は、開示リスクを具体的に定めることに他ならない。従って、評価される開示リスクの目標値が問題となる。この点について本論文は、個体識別が不可能な状態を過去の事例を基に決める方法を提案する。過去に公開されたデータについて個体識別が観測されていないとすれば、その事実は個体識別不可能ということについて情報を持っている。故に個体識別が観測されることの確率モデルを構成し、既公開の匿名データ等を匿名化の程度判断についての統計的証拠に転ずる。

このような理論なくして、明確な個体識別性の審査はあり得ない。そもそもチェックリストの記載事項は、個体識別と理論的に関係する要因であるべきだ。そしてリストの使い方も理論が定める。また本論文の理論は観測結果と関係を持ち、実証の対象である。いかなる理論も実証を経ることで継続的に改善される。従って本論文は、個体識別に関する統計的証拠—エビデンスに基づいた匿名化 (Evidence Based Anonymization, EBA) を主張する。

本論文の構成は以下の通りである。2 章は全体で、個体識別が可能か否かの判定方式を明らかにする。まず 2.1 節において、個体識別の観測と可能性の関係を確率モデルで表す。次に 2.2 節では、個体識別行為を所与の要因についての確率モデルで表す。2.3 節では、個体識別の要因について計量可能な方式を考察する。ここまでの議論で、情報不足により定性評価で満足せざるをえない個体識別の要因があることになる。2.4 節ではそのような要因を明らかにする。2.5 節では、個体識別の行為と観測の関係について考察する。3 章では 2 章の理論を基に、匿名データ審査体制の改善点を指摘する。

2　個体識別可能性の判定方式

2.1　個体識別の観測モデル

個体識別が可能か否かは、明らかにデータの表現に依存する。ここで匿名化による表現の変化は滑らかだが、個体識別が不可能と可能の差は不連続である。これをモデル化する場合、データ表現の適当な実数特性値が閾値を超えれば個体識別が可能とみなすのが定石であろう。本論文でもこのように考え、個体識別の難易度とみなせる特性値に注目する。この難易度が閾値より高ければ個体識別が不可能とみなすのは自然[5] である。

[4]例えば Duncan et al. (2001) や Domingo-Ferrer and Torra (2001) など。伊藤 (2012) のサーベイを参照せよ。

[5]一般目的汎用ファイル (PUF) の作成においても整合的な考え方である。星野 (2010) の考察の通り、PUF は匿名データよりも強い匿名性を必要とする。しかし個体識別が可能か不可能かという統計法の二元論では、PUF と匿名デー

具体的な難易度測度の設計については後で考察するとして、モデル化をすすめよう。難易度 δ を引数とする関数 f は個体識別が可能なら 1、不可能なら 0 を返すとする。つまり閾値が α として

$$f(\delta) = \begin{cases} 1 & \delta < \alpha \text{ の場合} \\ 0 & \delta \geq \alpha \text{ の場合} \end{cases} \tag{1}$$

ということになる。もし所与のファイルが個体識別可能か否か判定したいなら、その難易度を求めて α と大小を比較すればよい。ただ α が既知となるような難易度測度の設計は難しい。後述されるように、事実上観測可能でない個体識別の要因が残るはずだ。では α が未知の場合にどうするべきか。

まず α が推定可能か考えてみよう。統計的に未知母数 α を推定するには、観測値が必要になる。しかし個体識別が可能か否かは、観測されることではない。観測可能な事実は、個体識別が起きたか否かである。モデルを用いて説明しよう。確率変数 X が 1 なら個体識別が観測され、0 なら観測されないこととする。個体識別が不可能なら必ず $X = 0$ である。個体識別が可能の場合、難易度 δ に依存する確率 $p(\delta)$ で識別が観測されると考えよう。すなわち $\Pr(X = 1; \delta < \alpha) = p(\delta), \Pr(X = 0; \delta < \alpha) = 1 - p(\delta)$ とする。個体識別は決して起きないと考えれば危機管理にならないので、$p(\delta)$ は正と想定するべきだ。

このような状況で閾値が共通する n 件の事例が存在するとしよう。$i, i = 1, 2, \ldots, n,$ 番目について観測されるのは、少なくとも難易度 δ_i と識別の有無 x_i である。単純化のため $\delta_1 < \delta_2 < \cdots < \delta_n$ としよう。この中で個体識別が観測された（$x_i = 1$ となる i が存在する）場合は $\delta_i < \alpha$ と分かる上、実務的に重要でない。故に個体識別がこれまで[6] 起きていない（全ての i について $x_i = 0$）として考察を続ける。この場合モデルの尤度 ℓ は、$\delta_i < \alpha \leq \delta_{i+1}$ の時 $\ell(\alpha) = \prod_{j=1}^{i}(1 - p(\delta_j))$ となる。そして全ての δ について $0 < p(\delta) < 1$ なら、α の最尤推定値 $\hat{\alpha}$ は δ_1 以下である。つまり過去の事例で個体識別が観測されていなければ、その最も低い難易度 δ_1 以下と閾値 α は推定される。このように $\hat{\alpha}$ が一意に定まらないのは、情報が無いので区別出来ないことを意味する。

情報を補うため α の事前分布を用いることは考えられるが、それよりも観測情報の増加を工夫する方が健全だろう。つまり X を $\{0, 1\}$ の二値とするのではなく、α と δ の距離に依存する連続量と出来ればよい。これは治験薬の用量反応関係の推測と考え方が同じである。薬の臨床試験では、人体に決定的な悪影響を及ぼしてはならない。このような制約下では、薬剤の投与を少量から始めて徐々に増やし、危険な兆候が見られれば中止する。生死の二値ではなく、投与量が死亡の閾値と近いことを示す兆候（心拍や呼吸の異常等）を観測するのである。

問題は、α と δ の距離に依存する観測可能な事象として何を用いるかである。例えば個体を識別できたと誰かが誤って主張することなどが、個体識別発生の兆候として考えられる。この場合、個体識別を試す気にさえならない水準よりは難易度が下がっていることがわかる。一般に 1 件の

タを区別できない。ところが個体識別の難易度という概念を用いれば、個体識別が不可能という状態の中で匿名データと PUF を区別出来る。

[6] データの公開直後に識別が起きなくても、ある程度後で識別が起きることはあり得る。閾値の推定をする時点に依存して各 $p(\delta_i)$ は変化するかもしれない。しかし $0 < p(\delta_i) < 1$ なら $\hat{\alpha} \leq \delta_1$ という結論は変わらない。

重大事故の陰には 300 件のヒヤリ・ハットが起きているという。匿名データ提供に対する社会的反応が、警鐘になる可能性はある。また攻撃者の動機を考えることで、ある程度 $p(\delta)$ を定めることができるかもしれない。2.5 節でそのような考察を部分的に行うが、実務への反映は拙速と思われる。

では現時点で、新しく公開するファイルの難易度 δ_{n+1} をどのように決めたらよいか。一つの考え方は $\delta_{n+1} = \delta_1$ とすることだろう。強い仮定を置かずに推測出来るのは、難易度の閾値が δ_1 以下ということまでである。これは δ_1 未満の難易度について個体識別が不可能な証拠がないということだ。個体識別が可能になるという過誤の可能性[7]を考えれば、慎重な判断は正当化されるだろう。また $\delta_{n+1} = \delta_1$ として難易度 δ_1 における観測が蓄積されることは、将来的に意味を持つ。同難易度の複数のファイルについて個体識別が観測されなければ、その難易度が個体識別不可能な確率は高まる。また観測情報が増えないと、閾値との近さについて確かな判断は出来ない。

このように個体識別の難易度という概念を用いれば、個体識別が可能か否かの判断において過去の事例を統合して利用できる。しかし例えば国や公開時期が違う事例において、個体識別が可能となる難易度の閾値 α は同じだろうか。

閾値 α が共通する事例の範囲は δ の具体型に依存する。個体識別に関する要因を全て勘定する理想的な δ を用いる場合、全事例で閾値が共通するとみなしてよい。逆に閾値の変動は、δ が考慮しない要因の変化から生ずる。データ表現の実数特性値として導入した δ だが、その他の要因を算出に用いるべきかもしれない。以下では良い δ の構成を考察しよう。

2.2 個体識別行為の確率モデル

前節では個体識別の難易度に依存して個体識別が確率的に観測されるモデルを考察した。具体的に個体識別の難易度を定めるには、個体識別行為をモデル化する必要がある。本節ではそのようなモデルを構成し、個体識別の難易度の定式化をすすめる。

個体識別行為の確率モデルに関する先行研究としては、英国国勢調査匿名化標本の開示リスクを評価した Marsh et al. (1991)、及びこの論文を再考した Dale and Elliot (2001) が挙げられる。Marsh 等は個体識別が起きる条件を具体的に挙げ、それらが満たされる確率を個別に評価することで、個体識別が起きる確率を計算しようとした。まずこの試みを検討しよう。

最もあり得る個体識別の形態は、識別を試みる者（「攻撃者」）が素性を知る個体を（匿名化された）公開ファイルの中に見つけること[8] だと言われている。既知の個体について攻撃者が知る属性（「キー変数」）を並べたファイルを「攻撃用ファイル」と呼べば、「攻撃」とは公開ファイルと攻撃用ファイルでキー変数が同じレコードを探すことと言える。しかしそのような個体が見つ

[7]個体識別が可能であるにも関わらず識別が起きない ($p(\delta) < 1$) と過誤が生ずる。故に過誤の可能性を減らすには、δ_1 の例だけでも攻撃実験をするなどして識別が可能か確認することも役立つ。2.5 節の議論も参照のこと。

[8]このような行為を竹村 (1997) は「順攻撃」と呼ぶ。一方、公開ファイル中の特定個体を母集団に探す行為は「逆攻撃」と呼ばれる。この区別が意味を持つのは、攻撃者が探して素性を知ることが出来る個体群が、（攻撃用ファイルの）素性を知る個体群と異なる場合である。本稿の想定する攻撃者は統計当局にとって最悪の攻撃用ファイルを持つので、順と逆の区別はしない。

かったとしても、それは母集団に複数存在する属性が同じ個体のうちの一でしかないかもしれない。故に統計当局は、母集団に一しか存在しない個体（「母集団一意[9]」）を公開ファイルの中に攻撃者が発見することを警戒しなければならないと言われている。

このような背景の下、Marsh 等は個体識別が実際に起きる確率を以下のように分解する。

$$\Pr(\text{識別が実際に起きる}) = \Pr(\text{識別が起きる} \mid \text{識別を試みる}) \Pr(\text{識別を試みる}) \tag{2}$$

更に識別を試みた時にそれが成功する事態は、以下の 4 つの条件が成立する場合だという。

(a) 攻撃用ファイルと公開ファイルのキー変数が同じ（時点や分類の）基準で記録されている。

(b) 公開ファイルに個体が含まれている。

(c) 個体が母集団一意である。

(d) 個体が母集団一意と確証出来る。

これらの条件が満たされる事象をそれぞれ a から d と書けば

$$\Pr(\text{識別が起きる} \mid \text{識別を試みる}) = \Pr(a) \Pr(b|a) \Pr(c|a,b) \Pr(d|a,b,c) \tag{3}$$

ということになる。右辺の確率を個別に評価できれば、左辺の確率が求められる。

このような分解により、個体識別という漠然とした行為は直感的に解釈できる事象の積となる。(3) 式の分解で鍵となる母集団一意概念は、この種の議論では珍しく非専門家でも理解が可能であり、よく知られている。

ただし母集団一意は特殊な匿名化[10] において無意味な場合がある。また普通の匿名化手法だけ用いるとしても、個体識別が論理的に可能なのは母集団一意に限らない。母集団二意の個体も、自レコードが分かるなら、もう片方の個体のレコードが識別できる。そして三意以下でも、個体間で結託すれば識別できる。しかし母集団一意数と二意以下の珍しい個体数は経験的に比例する。故に母集団一意数を管理すれば、二意以下の識別可能性も同時におさえられる。母集団一意は実数に意味があるというより、管理対象のリスク測度として分かり易い点が望ましい。

また母集団一意数は、匿名化の程度についてある種の単調性を持つ。主に使われる匿名化手法は「再符号化[11]」と呼ばれ、個体属性を粗く分類する。再符号化で分類を併合してより粗い分類に変換する場合、母集団一意数は非増加である。つまり匿名化の直感的な軽重と母集団一意数の大小は矛盾しない。EBA では匿名化事例に順序をつけるので、矛盾しない順序を得られる方法が重要である。

[9] 公開ファイル中で所与のキー変数の組み合わせの条件を満たす個体数が 1 の場合、そのような個体は「標本一意」と呼ばれる。標本一意でも母集団一意とは限らないが、母集団一意なら標本一意である。

[10] Hoshino (2009, Section 6) で議論したとおり、匿名化された表現が互いに排他的でないと母集団一意は一意にならない。通常用いられる大域的再符号化なら、表現は互いに排他的となる。

[11] トップコーディングや削除 (suppression) も再符号化の特殊ケースである。

このように母集団一意数は使用に異論[12] はあるが、望ましい性質をいくつか持つ。Marsh 等の分解を活かして個体識別の難易度を構成出来ないだろうか。

問題は、Marsh 等が分解した要因 (Pr(識別を試みる), $\Pr(a)$,$\Pr(b|a)$,$\Pr(c|a,b)$,$\Pr(d|a,b,c)$) はそれぞれ評価できるとは限らないことである。まず Pr(識別を試みる) の評価は難しいと Marsh 等も認めており、定性的に議論[13] した上で「識別を試みた例を知らないので識別を試みる確率の最良の推定値は経験からゼロ」と述べている。また $\Pr(d|a,b,c)$ についても分からないので、「非常に多くのキー変数について事前情報が無いはずなのでゼロと信ずるが 0.001 と仮定」している。これでは数値評価が出来ているとは言えない。これらの確率評価は出来るとしても膨大な情報を必要とする。現実的に Marsh 等の方法では、リスク測度 (2) と (3) のいずれも数値評価できない要因を抱える。

もちろん Pr(識別が起きる | 識別を試みる) の条件付き確率の積による分解は一意ではない。故にこれを全て評価できる要因の積に書ければ、評価出来ない問題は解決する。しかしそのような分解は不可能であろう。最大の困難は識別が観測されないことである。個票の公開で先行する海外でも識別が起きないように匿名化しているので、ほとんど観測されない事象[14]の確率の推定を強いられる。それにも関わらず、個体識別という事象を細かい要因の積に分解すれば、要因毎に十分な観測数が得られるはずがない。目的の事象の観測が限られる以上、要素の分解に依存したアプローチは、どこかで情報不足の壁に阻まれるであろう。個体識別について数値評価出来ない要因は、存在を前提とするべきである。

実は数値評価する要因 $\boldsymbol{y}_1$ と評価しない要因 $\boldsymbol{y}_2$ が分かれても、ある程度は個体識別の難易度を相対比較できる。数値評価した結果 $g(\boldsymbol{y}_1)$ について、個体識別の難易度 δ が $h(g(\boldsymbol{y}_1),\boldsymbol{y}_2)$ と書けるとしよう。ここで $\boldsymbol{y}_2$ は数値評価できないので、難易度関数 h の具体型は分からない。しかし h は以下のような単調性を持つとする。

$$g(\boldsymbol{y}_1') \geq g(\boldsymbol{y}_1) \Rightarrow h(g(\boldsymbol{y}_1'),\boldsymbol{y}_2) \geq h(g(\boldsymbol{y}_1),\boldsymbol{y}_2) \tag{4}$$

つまり事例 $(\boldsymbol{y}_1',\boldsymbol{y}_2)$ と $(\boldsymbol{y}_1,\boldsymbol{y}_2)$ では、数値評価部が低いほうが難易度が低いということである。単調性 (4) さえ成り立てば、$\boldsymbol{y}_2$ が共通する複数の事例から、最も個体識別の難易度が低いものを選べる。そして 1 節のように $n+1$ 番目の新しい匿名データを公開するとして、数値評価値 g を $\boldsymbol{y}_2$ が共通する過去最低の事例に合わせればよい。このように g の達成目標値は $\boldsymbol{y}_2$ に依存して決まる。

次に Marsh 等の方法と 1 節の個体識別モデルとの関係を整理しよう。まず個体識別が可能ということは、識別を試したときに識別が起きる確率が正ということと同じである。故に個体識別が

[12]例えば同じ母集団一意でも、似た個体が居ない方が目立って識別し易いだろう。故に母集団一意のレコードの中で、似た属性の個体が多いか少ないかで開示リスクを変える考え方を「レコードレベルリスク」と呼ぶ。例えばより低次元の周辺分割表で一意になる個体の方が危険とみなす “Special Unique” は比較的計算しやすい (Elliot et al., 1998)。このような議論は一理あるが、リスク管理の対象として複雑な測度は望ましくない。また匿名化の程度についての単調性が崩れるかもしれない。

[13]曰く Pr(識別が起きる | 識別を試みる) が減少すれば Pr(識別を試みる) も減る。またデータの観測と公開の時点が離れれば Pr(識別を試みる) も減る、等。

[14]開示制限を失敗して個体識別が可能な例は Sweeney (2002) が報告している。

可能かの判断は、(3) 式が正かの判断と同じである。そして識別が実際に起きた場合に必ず観測されるなら、(2) 式の Pr(識別が実際に起きる) は、1 節の $p(\delta)$ と同じ概念となる。しかし攻撃者が識別に成功しても、黙っていれば観測されるか分からない。故に識別が実際に起きることと観測されることは区別した方がよいかもしれない。この議論は 2.5 節へ先送りする。

結局 (3) 式の右辺の要素のどれかが 0 なら、個体識別が不可能と言える。しかし公開される母集団一意が皆無になるのは例外的で、普通は $\Pr(a,b,c)$ は正となる。(3) 式の右辺を書き換えると

$$\Pr(\text{識別が起きる} \mid \text{識別を試みる}) = \Pr(a,b,c)\Pr(d|a,b,c) \tag{5}$$

であり、$\Pr(d|a,b,c)$ が 0 なら個体識別が不可能と考えられよう。つまり Marsh 等の枠組みにおいて通常の場合、個体識別が可能か否かは $\Pr(d|a,b,c)$ が 0 か否かという問題に縮退する。しかし Marsh 等は $\Pr(d|a,b,c)$ の評価に失敗している。

我々の議論に沿って $\Pr(d|a,b,c)$ が 0 か否かの判別方式を構成しよう。これまでの議論では、個体識別の難易度 δ が (1) 式のように閾値 α 未満なら個体識別が可能ということであった。また δ は (4) 式の関数 h で表されると考えていたので、

$$\delta = h(g(\boldsymbol{y}_1), \boldsymbol{y}_2) < \alpha \Rightarrow \Pr(d|a,b,c) > 0 \tag{6}$$

とすればこれまでの議論と整合する。つまり個体識別の難易度 δ が閾値 α を下回れば、個体識別が可能ということである。

このように考えると、関数 h は $\Pr(d|a,b,c)$ が正という判定とできるだけ直接関係するのが望ましい。そして事象 (a,b,c) が条件の確率を判定するなら、h は (a,b,c) を要因とするべきだろう。これを基準化して $-\Pr(a,b,c) = g(\boldsymbol{y}_1)$ とすれば、g が (4) 式の単調性を満たして都合がよい。何故なら確率 $\Pr(a,b,c)$ は、正確に表現されて公開される母集団一意数と比例する。そして正確に表現されて公開される母集団一意数の増加は、母集団一意の確証をより容易にすると考えられる。故にあとは $\Pr(a,b,c)$ が数値評価可能であれば、その評価値に基づいて匿名化を管理できる。

これまでの議論では $\Pr(a,b,c)$ の意味が曖昧だったが、計算方法を定めれば概念は限定される。また関数 g の具体型と必要な情報 $\boldsymbol{y}_1$ も、計算方法に依存して定まる。そして $\boldsymbol{y}_1$ が決まらなければ、$\boldsymbol{y}_2$ も定まらない。これらは理論モデルとは異なる次元の問題なので、節を改めて考察しよう。

2.3 匿名性の計測—実質と下限

EBA において匿名性の評価値を相対比較する際、評価手法のゆれは望ましくない。また出来るだけ広範な経験を利用するには評価が名人芸であってはならず、形式的な手続きでなければならない。そのように匿名性の計算手法は具体的に定めておくべきである。本節では前節のモデルに沿って匿名性の評価値 $g(\boldsymbol{y}_1) = -\Pr(a,b,c)$ の計算を考察する。

匿名性の評価をする際、実質か下限のいずれを求めるのか意識的でなければならない。実質とは実際の攻撃者の能力に合わせた評価という意味であり、下限とは統計当局と同じ情報を持つ「最強」の攻撃者を想定するということである。

両者の違いを母集団一意数を例にとって説明しよう。母集団一意数は、キー変数群の多元分割表における度数 1 のセル数と形容することも出来る。この母集団一意を計算する多元分割表で、各変数の区分（カテゴリー分類）は公開データの区分と一致させるのが常識的である。ただ攻撃用情報の精度が公開データより粗ければ、公開データの区分方法で算出した母集団一意は、攻撃者にとっての母集団一意にならない。例えば攻撃者が五歳階級のデータしかもっていなければ、各歳別でデータが公表されていても階級内で識別できない。故に実質的な母集団一意数の評価では、公開表現と攻撃用情報の粗い方に各変数の区分を合わせる。常識的な方法では公開表現の方が攻撃用情報より常に粗いので、最強の攻撃者が想定されている。

実質的な匿名性評価では、現実の攻撃者の能力を知る必要がある。そして攻撃者の能力を知るための情報収集体制については、Elliot et al. (2010) の重要な議論が存在する。この議論は 2.4 節で紹介するが、そのような情報の完全な収集は資源の制約等から無理であろう。つまり実質的な匿名性評価の問題は、必ずしも評価に必要な情報を得られないことである。

部分的な情報から実際にありそうな攻撃方法（シナリオ）を推定し、匿名性を評価することはできる。このようなシナリオ依存のリスク評価は、例えば Paas (1988) が採用している。しかし想定した攻撃者より強い攻撃者が存在した場合、個体識別の可能性は管理されない。

一方、下限の匿名性は後で確認するように、公開データ表現とその元データから評価する。これらの情報は統計当局にとって常に入手可能であり、シナリオ選択に起因する評価のゆれが起きない。また最強の攻撃者より弱い攻撃者についても、個体識別の可能性は（過剰だが）管理できる。ただ問題は、過去の事例における個体識別の有無が、現実の攻撃者の能力を反映しているということだ。

この問題を一般的に考えよう。匿名性の数値評価値の要因 $\boldsymbol{y}_1 = (\boldsymbol{e}_1, \boldsymbol{e}_2)$ について $\boldsymbol{e}_1$ は公開データ表現とその元データと考える。そして $\boldsymbol{e}_2$ は、必ずしも観測されない攻撃者の能力とする。$\boldsymbol{e}_2$ が観測されるとして、実質的な匿名性の数値評価値が $g(\boldsymbol{e}_1, \boldsymbol{e}_2)$ で表される。一方、最強の攻撃者にとっての匿名性の数値評価値を

$$\inf_{\boldsymbol{e}_2} g(\boldsymbol{e}_1, \boldsymbol{e}_2) =: \underline{g}(\boldsymbol{e}_1)$$

で表そう。ここで $\underline{g}$ を用いて過去の事例で計算した匿名性の最低数値評価値を $\underline{\gamma}_1$ と書く。新規に公開するデータの匿名性を $\underline{g}$ で計算して $\underline{\gamma}_1$ としてよいだろうか。

所与の $\boldsymbol{y}_2$ について、匿名性の数値評価値 g が β 未満なら個体識別が可能としよう。過去最低の実質的な匿名性 γ_1 は $\underline{\gamma}_1$ 以上である。故に閾値 $\beta \leq \gamma_1$ が正しいとしても、$\underline{\gamma}_1 < \beta$ となる場合があり得る。このとき新規に公開するデータの実質的匿名性が例えば $\underline{\gamma}_1$ と等しければ、個体識別は可能となってしまう。

このような事態は、匿名性の実質と下限の差が変動する場合に起こりうる。新規に公開するケー

スについて匿名性の要因を $(\boldsymbol{e}'_1, \boldsymbol{e}'_2)$ と書く。ただし $\inf_{\boldsymbol{e}'_2} g(\boldsymbol{e}'_1, \boldsymbol{e}'_2) = \underline{\gamma}_1$ となるように匿名化がなされているとしよう。そして $\underline{g}$ を用いて評価した過去最低の匿名性のケースの要因を $(\boldsymbol{e}_1, \boldsymbol{e}_2)$ と書く。つまり $\inf_{\boldsymbol{e}'_2} g(\boldsymbol{e}'_1, \boldsymbol{e}'_2) = \inf_{\boldsymbol{e}_2} g(\boldsymbol{e}_1, \boldsymbol{e}_2)$ が成立している。ここで匿名性の実質と下限の差を過去のケースは $c = g(\boldsymbol{e}_1, \boldsymbol{e}_2) - \inf_{\boldsymbol{e}_2} g(\boldsymbol{e}_1, \boldsymbol{e}_2)$ 、新規のケースは $c' = g(\boldsymbol{e}'_1, \boldsymbol{e}'_2) - \inf_{\boldsymbol{e}'_2} g(\boldsymbol{e}'_1, \boldsymbol{e}'_2)$ で表す。攻撃者の能力が向上して $c > c'$ の時、新規ケースの実質的匿名性は過去最低の実質的匿名性を下回る。そして $\beta = g(\boldsymbol{e}_1, \boldsymbol{e}_2)$ なら、新規ケースの実質的匿名性は $g(\boldsymbol{e}'_1, \boldsymbol{e}'_2) = \underline{\gamma}_1 + c' < \beta$ であり、過去のケースでは不可能だった個体識別が可能となる。

なお上の考察で c は下限評価の歪みを含む。下限評価の真値からのずれは、一定なら問題にならないことは重要だ。つまり実質 g と下限 $\underline{g}$ の差 c がケース毎に変化しなければ、過去最低の匿名性の下限 $\underline{\gamma}_1$ を与えるケースでは実質的な匿名性も過去最低になる。そして $\underline{\gamma}_1 > \beta - c$ なら $\gamma_1 > \beta$ なので、$\underline{g}$ を用いて匿名化の程度を決めれば実質も管理される。言い換えれば、EBA は匿名化を相対比較するので、匿名性の絶対値に意味は無いということである。

匿名性の実質と下限の差 c が一定という重要な条件を満たすには、たとえ歪んでいても同じ方法で測ることが重要である。また $\boldsymbol{e}_2$ は無視できず、変化を確認するべきだ。しかし情報 $\boldsymbol{e}_2$ は入手性に問題があるので、$\boldsymbol{y}_2$ の一部として定性評価するしかないだろう。

このような前提で、匿名性の数値評価は $\underline{g}(\boldsymbol{e}_1)$ を用いるのが望ましい。つまり公開データ表現とその元データから匿名性の下限を求めるということである。そのように $\Pr(a, b, c)$ が計算できるか要素毎に確認しよう。

$\Pr(a)$ の評価　Marsh 等は誤分類や誤記が公開ファイルと攻撃用ファイルのキー変数で起きていない確率を $\Pr(a)$ とした。1981 年の英国センサスの事後調査 (Post Enumeration Survey) で求めた変数の誤分類率を参照して、5 つのキー変数が全て正確に分類されている割合は 0.8 程度と見積もられている。この場合に誤分類が公開ファイルと攻撃用ファイルのキー変数で独立に起きているなら、$\Pr(a) = 0.8^2 = 0.64$ である。なおキー変数が増えれば、全てのキー変数が正確に分類されている確率は減少する。しかし母集団一意数は増えることになる。

実際には、公開ファイルと攻撃用ファイルで調査時点の差や変数の定義の差も存在するだろう。これらの差は $\Pr(a)$ を低下させる。1971 年の英国センサスの 1 年後に再調査した結果、同じ職業だった人の割合が 61%でしかない例を Marsh 等は挙げている。1991 年の英国センサスについては、Dale and Elliot (2001) が各キー変数が経時変化する程度を調べている。ただ Dale and Elliot も述べているように、本気の攻撃者は特定の調査が数年後に公開されることを見込み、同時点に調査した攻撃用ファイルを準備しておくだろう。この場合は、調査時点や変数の定義の差に多くの保護効果を期待出来ない。このように攻撃のシナリオに依存して、$\Pr(a)$ はかなり変化する。

我々は匿名性の下限を評価したいので、最強の攻撃者を想定する。このシナリオでは、匿名化される前のキー変数が全て攻撃者にばれていると考える。この場合キー変数の精度を評価するに

は、匿名化されていない元ファイル[15] と公開ファイルのキー変数を比較する。そして近いレコードが同個体（のペア）と判定し、正しく判定された割合[16] を $\Pr(a)$ と考える。このような手法は開示リスク評価によく用いられるので、研究蓄積が利用可能である。例えば伊藤他 (2009) を見よ。なお我々のシナリオでは、公開ファイルと攻撃用ファイルで調査時点の差は存在しない。それから両者の定義の差は、匿名化によるもののみである。そして元ファイルのキー変数がどれほど誤分類されていたとしても、個体と正しく対応可能である。

このようなシナリオの非現実性は、現実に用意可能な攻撃用データの質と量に依存する。この情報が $\boldsymbol{e}_2$ であり、$\boldsymbol{y}_2$ の一部と考える。一方 $\boldsymbol{y}_1$ は元ファイルと公開キー変数データだが、これらを比較し、正確にマッチされたレコードの割合が $\Pr(a)$ として計算可能であった。注意すべきなのは、毎回同じ方法でマッチさせることである。

$\Pr(b|a)$ の評価　Marsh 等の議論で確率 $\Pr(b|a)$ は、公開個体数が母集団サイズにしめる割合である。例えば 1991 年の英国センサス匿名化標本では 2%となる。全数調査から等確率でサブサンプリングした公開ファイルなら、個体は等確率で公開ファイルに含まれる。その場合に Marsh 等の方法は妥当である。

しかし現実の標本調査では不等確率の複雑な抽出が行われる。また一部の個体について、調査されたか否かを攻撃者が知っているかもしれない。例えば集落抽出を行う調査では、被調査者は隣家も調査されたと推測できる。従って一般に真の $\Pr(b|a)$ は個体毎に異なる。

ただ我々は個体毎（いわゆるレコードレベル）の確率評価をしているのではなく、ファイルレベルの評価が目的である。ファイルレベルでは公開の平均的な可能性を評価すると考えて、Marsh 等の方法を用いることにしよう。

$\Pr(c|a,b)$ の評価　母集団一意数が母集団サイズにしめる割合を求めればよい。なお本節冒頭で議論したように、母集団一意数を計算するための変数の区分は公開表現に従うべきである。それから世帯単位のファイルでは、世帯毎の固まりを「レコード」として母集団一意を計算するのが筋である。具体的には、世帯員のレコードを年齢順に連結したまとまりを一レコードとして扱えば良い。この場合、世帯人数が異なればレコード長も異なる。

ただし全数調査でない限り、母集団一意数は推定しなければならない。そして星野 (2003) で説明したように、母集団一意数の推定は単純ではない。

Marsh 等は英国センサスの全数データが使えなかったため、イタリアのセンサスデータで母集団一意を数えて外挿している。キー変数が 8 つで 10 万人レベルの地域区分を公開するとして、

[15] 攻撃者は補定、エディットのルールを知らないはずなので、補定等を施す前のデータを元ファイルとする方が現実に近いかもしれない。ただそのようなデータが常に利用可能とは限らない。相対比較可能性を考えれば、補定等を施した後のデータを元としてよいだろう。実質と下限の差があるとしても、補定等の割合が小さかったり調査毎に大きく変動しない場合は無視できる。

[16] 何を分母とするかは議論の余地がある。本当に評価したいのは、母集団一意レコードについてのキー変数の精度である。しかし全数調査でないと、母集団一意のレコードを決めるのは難しい。そして評価の歪みより方法の変動を避けたいので、標本一意数を分母とするのが一案である。近さの計算方法によるが、一意にペア相手が見つかるレコード数と標本一意数はほぼ同じである。なお分母が 0 の場合は $\Pr(a) = 0$ とみなして差し支えないだろう。

$\Pr(c|a,b)$ は 2.4%程度とされた。なおこの値は世帯単位ではなく個人単位で評価されている。1991年の英国センサスデータについては、Dale and Elliot (2001) によるとキー変数が 7 つで 12 万人レベルの地域区分を公開する前提で、$\Pr(c|a,b)$ は 4.8%であった。

母集団一意数評価は Marsh 等の時代に比べてかなり進歩しており、（母集団サイズが所与で）公開ファイルの情報だけから推定できる。しかし評価手法による結果の違いが大きいので、同一手法によって評価することの重要性も大きい。

幅広い母集団について一意数の推定精度をルーチンワークとして確保するには、ピットマンモデル（付録 B を参照のこと）の使用を推奨する。この方法においてデータは、無限母集団すなわちピットマン分布からの標本とみなされる。そして母集団[17]も同一無限母集団からの標本とみなすので、データからピットマン分布の母数を最尤推定し、推定値の下で母集団一意数の挙動を求める。より具体的には、付録 C の手順書を参照されたい。

開示リスクを評価するファイルのレコード数は、母集団個体数のせいぜい一割程度であろう。この場合に安定的な母集団一意数の推定量は、全てバイアス[18] を持つ。手順書の推定量も例外でなく、おそらく過大に一意数を推定する。しかし既に考察したように、バイアスは一定であれば問題にならない。

なお特定のモデルと決めつけるよりも、モデル集合からデータに良く当てはまるモデルを選択し、そのモデルで一意数を推定する方が正確になろう。しかしモデル集合の空間をうまく張らないと、リスク評価値がぶれる。また経験的に多くの場合、ピットマンモデルが選択[19] される。故に手間や精度及び様々な結果の整合性を勘案すれば、母集団一意数は常にピットマンモデルによって推定するのが最善と思われる。

一点つけ加えておくと、母集団一意数の推定改善にセルの番地情報を使うアプローチはあり得る。しかし大規模かつ疎な分割表では絶対的に情報が不足しているので、うまくいかないであろう。またそのようなアプローチは高度なモデリングが要求され、開示リスク評価の試行錯誤にも向かない。故に実務への採用は難しいはずだ。

このように $\Pr(a,b,c)$ の下限評価に必要なのは、

$$\boldsymbol{e}_1 = (\text{元ファイル}, \text{公開ファイルのキー変数}, \text{母集団サイズ})$$

である。これらの情報が数値評価の対象となり、$\boldsymbol{y}_2$ には含まれないと考えるべきだろう。次節では数値評価しない要因 $\boldsymbol{y}_2$ を確定しよう。

---

[17] 一部が観測されている現実の母集団について推定するのではなく、同サイズの母集団を新たに発生させる場合の挙動が推定される。

[18] 有限母集団から非復元単純無作為抽出する場合、一意数の不偏推定量は一意に存在する。しかしこの不偏推定量は標準誤差が大きく、標本抽出率がかなり高くないと実用に耐えない。そして一意な不偏推定量なので、推定を安定させるためのいかなる工夫もバイアスを生む。

[19] 裾の長いモデルとして代表的な負の二項分布は、統計的開示制限の分野ではポアソン＝ガンマモデルとして知られている。このモデルは基本的に広義のピットマンモデルの特殊ケース ($\alpha \leq 0$ に対応）である。故にピットマンモデルのデータへのあてはまりは、基本的にポアソン＝ガンマモデルを下回らない。そしてポアソン＝ガンマモデルによる母集団一意数の推定値は、必ず Pitman モデルの推定値より（かなり）小さくなると考えて良い。

2.4 定性評価の要因

本節では個体識別について数値評価しない要因 $\boldsymbol{y}_2$ を定める。これまでの議論より、我々は (6) 式に基づいて母集団一意の確証の可能性 $\Pr(d|a,b,c)$ を判断するので、h の引数 $\boldsymbol{y}_2$ は母集団一意の確証にかかる要因である。そして前節では $\boldsymbol{e}_2$、すなわち攻撃用データの質と量が、$\boldsymbol{y}_2$ の一部ということであった。

Marsh 等は母集団一意の確証手法として、全数名簿と公衆の目の利用[20] を検討している。全数名簿の利用とは、職業人名簿等で母集団一意が分かる場合を指す。特定の条件を満たす集団について全数の名簿があれば、その集団内の一意[21] は母集団でも（特定の条件を満たす）一意である。そのような個体について、Marsh 等は特に強い匿名化を求めている。それから公衆の目とは、珍しくて目立つ個体が有名な場合を言う。例えば職業が現職の首相であれば、母集団一意を確証可能である。昨今ではソーシャルネットワークの拡大により、公衆の目は無視できないように思う。

全数名簿が利用出来たり、属性が公衆に知られていたりする個体については、詳しい個体情報が社会に流通しているということだ。母集団一意の確証可能性及び実質的な $\Pr(a,b,c)$ は、そのような個体情報の環境に依存するだろう。個体情報環境を知るため、Elliot et al. (2010) は (i) アクセス制限付きデータベースの調査項目 (ii) 公知の個体データの形態 (iii) ネットショッピング等での web 上データ収集項目 (iv) 商業データベースの情報 (v) 個人情報の収集実験結果 (vi) 情報保有組織における個体情報の取り扱い慣行 (vii) ソーシャルネットワークでの個体データの形態、を調べることを提案している。またそこで現れる様々な変数間の関係を、ツリー構造を用いて記録することとしている。これらの要因は調査できたとしても、定量評価は（識別成功が希なので）難しい。ただこれらについての理解から、現実的な攻撃用データとして

$$\boldsymbol{e}_2 = (\text{外部データに含まれる個体数、変数の種類、精度})$$

を想定するべきだろう。なお世帯データの $\boldsymbol{e}_2$ は、事業所データの $\boldsymbol{e}_2$ と明らかに異なる。従って世帯データの匿名化事例は、事業所データの匿名化のエビデンスとして直接使えないということになる。個体情報環境は個体単位（個人、世帯あるいは事業所等）毎に集約するべきだ。

外部データに含まれる個体数は、実質的に攻撃可能な母集団一意数と比例するだろう。なお外部データに含まれる個体数増加の効果は、サブサンプリングにより $\Pr(b|a)$ を下げれば打ち消すことができる。公開個体率 $\Pr(b|a)$ も攻撃可能な母集団一意数と比例すると考えられるからだ。

外部データの変数の種類は、キー変数の決定に用いる。前節ではキー変数が所与であったが、実際はキー変数を選択しなければリスク評価が出来ない。そしてキー変数の選定基準の揺れは避けた方がよい。これを念頭におき、過去の事例で用いたキー変数の種類は、キー変数の選択で考慮すべきである。キー変数に相当すると判断した根拠の外部データの状況が変わらなければ、同じ変数はキーとして用いなければならない。根拠が変われば、キー変数も変えるべきだろう。なお

[20] 他に母集団一意の確率を統計モデルで求めることを挙げているが、それでは母集団一意の確証にならない。Dale and Elliot (2001) による Marsh 等の議論の再評価でも、統計的推測は母集団一意の確証として扱われていない。

[21] 全数調査において低次元クロス集計の結果の度数が 1 と分かるような場合も該当する。

Elliot et al. (2011) がキー変数の選択基準を考察している。彼らの議論では、変数のアクセス容易性を定性評価した上でキー変数が選択される。

外部データの変数の精度は、$\Pr(a)$ と $\Pr(c|a,b)$ の実質的な値と関係する。なお変数の精度上昇の効果は、匿名化を強く施せば無効化出来る。何故なら匿名化で定まるデータの粗さ以上に変数の精度が上昇しても、開示リスクは変化しない。

$\boldsymbol{e}_2$ 以外の定性評価要因として、匿名化の「曖昧さ」を検討しておこう。ここでは匿名化に用いたデータ変換 m の形を攻撃者が完全には知らない場合を曖昧と呼ぶ。例えば米国センサスミクロデータのように、スワッピングが施されているがその割合やスワップ相手の選択方法などが未公開な状態は曖昧である。他方、労働力調査等の匿名データでは、符号表を読むことで匿名化が施されている変数や程度が完全に分かる。この状態は曖昧ではない。

曖昧さは余り研究されておらず、その効果[22]に定説はない。一つの理由として、計算機科学では曖昧さによる安全性を認めないことが挙げられる。その前提で設計した匿名化は統計当局が隠した情報が漏れても[23] 破られないので、保守的と言える。しかしこのような態度は最強の攻撃者を想定することと同じである。従って下限と実質の差の問題が起きる。

曖昧さが母集団一意の確証に影響する例を挙げよう。年齢と性別の二キー変数について、元ファイルが {(110,M),(120,F)}、公開ファイルが {(120,M),(110,F)} だとする。年齢をスワップしたと考えれば第一レコード同士が同一個体であり、性別をスワップしたと考えれば元ファイルの第一レコードと公開ファイルの第二レコードが同一個体となる。この場合は m について何も知らないと、公開ファイルのレコードが元ファイルのどちらのレコードか分からない。ところが年齢変数に適当なノイズを付加したという情報が有れば、元ファイルと公開ファイルで同一個体のレコードが判明する。そして年齢が 120 歳の母集団一意な個体は、公開ファイルの第二レコードと確証される。

このように母集団一意の実質的確証可能性は、曖昧さの程度に依存するかもしれない。故に曖昧さは匿名化設計の一部として、明示的に考察した方がよい。現実には、攪乱的手法の詳細を公開する程度を $\boldsymbol{y}_2$ の一部として管理するということになろう。なお補定やエディットの母数を明らかにしないことは、曖昧と同じことになる。

ここまでの議論で、既存の情報に基づく個体識別はある程度管理されるだろう。しかし情報が追加できるなら、これまでの枠組みでは管理されない事態が起きる。例えばあるレコードの識別が既存情報から確証できないにせよ、可能性が高いとしよう。この場合に追加の情報を詐取などすれば、確証できるかもしれない。追加情報を想定しての匿名化はあり得るが、詐取の可能性を際限なく考慮すると、有用なファイルの提供は不可能だろう。それよりも追加情報取得の可能性を低く保つ工夫をする方が良い。

[22]特定の曖昧さの効果は、例えば以下のように評価できる。保守的な攻撃者なら、曖昧な部分に自分に不利な事前分布を入れる。このように評価される攻撃の難易度と真の難易度の差が、曖昧さの効果である。

[23]関係者による情報漏洩だけ考慮すれば良いわけではない。攻撃者が攪乱の母数を推定できる可能性がある。例えば匿名データと匿名化されていないデータの分析結果を比較することで、攪乱の率の見当をつけられるかもしれない。攻撃者本人が 33 条申請による目的外使用でデータを手に入れなくても、他人が書いた論文や公の集計表が比較対象になり得る。

重要な追加情報を得るには、当該個体に接触する必要があるのではないか。そして接触するには、広い意味[24]での個体の位置（住所、職場など定期的に訪れる場所、電話番号等）を知らなければならない。故にそのような接触可能性に係る条件で母集団一意な個体は、そうでない母集団一意よりも追加情報を得やすいので、確証の可能性が上がる。従って広い意味での位置情報の精度は、一定以上にならないように管理するべきである。

本節の議論をまとめておこう。定性評価の対象 $\boldsymbol{y}_2$ として過去の事例と比較されるのは、以下の 3 要因である。

1. 同種の母集団についての $\boldsymbol{e}_2$：民間データベース等に含まれる個体数、変数の種類、精度
2. 匿名化の曖昧さ
3. 接触を可能とする情報の精度

本節で考察したように、匿名化手法の変更により攻撃者の能力向上を無効化できる場合がある。故に $\boldsymbol{y}_2$ が過去と同じかどうかの判断は、データ表現にある程度依存してしまう。匿名性の数値評価値 $\underline{g}$ を変えるために匿名化手法を変更すると、定性評価も変わるかもしれないことは注意するべきである。

2.5 識別を試みる確率の決定要因

これまでの議論で後回しにされた、個体識別が実際に起きる確率と観測される確率 $p(\delta)$ の差について本節では議論する。つまり $p(\delta) = \Pr(a,b,c,d) \cdot \Pr(\text{識別を公表するつもりで試みる})$ という関係が (2) 式の関係 $\Pr(\text{識別が実際に起きる}) = \Pr(a,b,c,d)\Pr(\text{識別を試みる})$ と違うかもしれないので、識別を公開すること、公開しないことについて要因の考察を行う。

識別を試みるという意思決定は、識別成功の損得や容易性に依存すると考えられる。Marsh et al. (1991) が指摘するように、$\Pr(\text{識別が起きる} \mid \text{識別を試みる}) = \Pr(a,b,c,d)$ の減少は $\Pr(\text{識別を試みる})$ を減少させるだろう。他に Elliot et al. (2010) は、もっともらしい攻撃シナリオの考察こそが、$\Pr(\text{識別を試みる})$ の妥当なモデル化につながると主張している。

攻撃者が真に識別を成功させた場合、その事実を公表して得られる利益と、識別を隠して得る利益がある。まず識別成功を公表した場合、攻撃者は有名になるだろう。そして識別された個体は情報の漏洩を知ることになり、識別によって入手した情報を用いた詐欺、ストーキング等は難しくなる。そのように識別で得た情報を実用するには、識別成功は公表しない方がよい。また識別成功を公表すれば法的、社会的制裁の対象[25] になるかもしれない。故に例えば商業目的なら識別成功を公表せず、攻撃者は精度の良いマーケティングの利益を享受するだろう。

[24] 狭義の地理情報が強力なキー変数であることは良く知られている。

[25] 匿名データの利用者については、統計法第 43 条第 2 項に「当該匿名データをその提供を受けた目的以外の目的のために自ら利用し、又は提供してはならない」とある。個体識別の成功を公表することは（識別目的でのデータ提供は行われないので）本条に違反するが、直ちに罰則が適用されるわけではない。匿名データの利用者についての罰則は「匿名データを、自己又は第三者の不正な利益を図る目的で提供し、又は盗用した者」に対して「五十万円以下の罰金に処する」(61 条 3 項）とだけ定められている。例えば匿名データの不備を指摘するための個体識別の公表は不正な利

このように考えれば、公開ファイルが含む実用（隠れて悪用）可能な情報が多ければ、Pr(識別を公表するつもりで試みる) を増加させる。また識別を試みるという事象は識別を公表するつもりで試みる事象を包含するので、Pr(識別を試みる) も増加する。

ただファイルが実用可能な情報を含まなくても、識別成功の公表により統計当局の面目を失わせ有名になることを、魅力的に感じる人間が居ないとは言えない。故に Pr(識別を公表するつもりで試みる) は正のはずで Pr(識別を試みる) $= 0$ にはならない。しかし実用の帰結は多様なのに対し、識別成功の公表は帰結が同じである。実用できない情報は公表することによってしか利益を得られないので、攻撃の誘因として全て等価ということになる。

では実用可能な情報とは何か。多様な犯罪を想像して判断するしかないが、匿名化によって実用性は変えられることを指摘しておく。例えば病歴という情報は、削除したり罹患時期を区間表示したりすることで、実用困難にできる。多くの統計調査は適切に匿名化すれば実用可能な情報を含まず、攻撃の誘因は識別成功の公表による利益のみとなる。そしてこの場合 $p(\delta) \doteq$ Pr(識別が実際に起きる) と考えて良いはずだ。

なお調査客体が秘密にしたい調査項目（変数）を「センシティブ変数」と呼ぶ。秘密でない情報は保護に値しないので、実用を妨げる目的での匿名化の対象は、センシティブ変数の一部と考えられる。重要なのは、センシティブか否かは調査客体の主観に依存[26] するということである。故にセンシティブの程度は、攻撃の動機付けの程度と必ずしも一致しない。

3　おわりに—匿名データの審査体制について

最初にこれまでの議論をまとめる。個体識別が不可能かつ有用なデータを統計的根拠に基づいて作成する手順は以下のようになる。

1. $\boldsymbol{y}_2$ が共通する過去の事例をリストアップする。
2. それらの事例について $\Pr(a, b, c)$ の下限 $\underline{g}$ をそれぞれ計算する。
3. その中で最も高い $\underline{g}$ を g^* と書く。
4. データの匿名性の評価値が g^* となるように匿名化する。

なお匿名性の評価値が同じになる複数のデータ表現では、データの有用性が高いものを選べばよい。本稿で有用性の評価は議論しないが、例えば星野 (2010) は基本的な考え方を説明している。以下ではこのような立場から、望ましい匿名データの審査体制を考察する。

益を図る目的と必ずしも言えないので、罰則は適用できないのではないか。なお 33 条申請によって調査票情報を手に入れた者が個人又は法人の秘密を漏らした場合は「二年以下の懲役又は百万円以下の罰金」(57 条 2 項 3 号)、自己又は第三者の不正な利益を図る目的で提供又は盗用した場合は「一年以下の懲役又は五十万円以下の罰金」(59 条 2 項)、と罰に差がつけられている。ところが匿名データの利用者が（個体識別によって入手した）秘密を漏らした場合の罰則規定はなく、そのような事態を統計法は想定していないように思われる。

[26] 全ての調査客体の判断を聞くのは非現実的なので、リスクの評価者がセンシティブ変数を定める際に保守的であれ、ということになる。

まず審査用資料（チェックリスト）は、蓄積して参照するものだということをはっきりさせておきたい。EBA は過去の経験をエビデンスとして用いる。故に過去のチェックリストを、経験の要約として用いたい。この事情は、将来的に α と δ の関係をモデル化する場合も変わらない。

従ってチェックリストの記載事項は、事例の十分統計量であるべきだ。つまり個体識別が可能か否かの判断に用いる情報を（過）不足なく記入するということである。このような観点から、現行のチェックリストは日本の制度にふさわしいだろうか。世帯調査のチェックリスト（H23/3/28 改正版）についてのみ、改善できる点を指摘したい。

チェックリストに記載するべき項目で漏れているのは、まず $\Pr(a,b,c)$ の下限 $\underline{g}$ である。付録の手順書に従えばそれほど計算に手間がかかるとは思えず、匿名化表現の要約として費用対効果が高い情報と考える。またキー変数の情報を持つ部分集団の全数名簿は、母集団一意の確証について重大な影響がある。故に質問項目として特に欄を設けるべきである。そのような名簿が存在するなら、名称、部分集団の種類、個体数、含むキー変数の種類、精度を記述させるとよい。他に狭義の地理情報については記入欄が存在するが、接触を可能とするような広義の地理情報の有無を確認するべきだ。

それからチェックリストに存在する項目で、記入の焦点をしぼるべき箇所がある。まず「ミクロデータを特定できる可能性のある外部ファイル」の存在を記入することになっているが、どのようなファイルが該当するのか明確化するべきである。具体的には、匿名データが含む変数と同じ情報（これがキー変数ということである) をもつ外部ファイルの有無を問うべきだ。そしてそのファイルの名称、含むキー変数の種類、精度、及び個体数を分けて記述してもらう方がよい。また「秘密の情報」（センシティブ変数）のうち、「特に秘匿する必要性の高い調査項目」の有無を聞いているが、必要性の意味を明白にしたほうがよい。個体識別の可能性を制限するための必要性ではなく、実用性を限るための匿名化の必要性を聞かなければならない。またチェックリストには「誤差（ノイズ)」を聞く項目が存在する。誤差の付加は「攪乱」手法の例なのだが、用語の問題は別にして、この項目には匿名化の曖昧さを評価するための情報を記入させるべきだ。具体的には、攪乱手法のパラメータと、その公開方針を分けて書かせるということになる。

チェックリストに記入される情報の使われ方は、説明書を用意するべきであろう。現行のチェックリストも冒頭で匿名化の考え方などが書かれているが、やや説明不足に見える。個体識別可能性の判定方式を明示すれば、焦点がずれたチェックリスト記入の恐れは減る。

Acknowledgements

本研究は科学研究費及び統計数理研究所の共同研究経費の補助を受けている。以下の付録 B,C は星野 (2012) の一部を改訂したものである。

付録

A　世帯調査のチェックリスト（H23/3/28 改正版）要約

1. 地理的情報
   (a) 地理情報のレベル、加工の有無
   (b) 地理情報以外の地理的情報の有無
   (c) 地域分析用の地理情報提供の有無
   (d) 特定の種類の施設の情報の有無
2. 世帯の識別情報
   (a) 世帯のキー変数
   (b) キー変数への匿名化及び分布
   (c) 世帯のまとまりへの匿名化の有無
3. 個人の識別情報
   (a) 個人のキー変数
   (b) キー変数への匿名化及び分布
4. 攪乱の有無
5. サブサンプリングの有無
6. 外部の情報
   (a) 個人・世帯の特定に使える外部情報の存在
   (b) 母集団情報として利用している情報
7. その他
   (a) データの並び順についての匿名化措置
   (b) サンプリング情報により特定の地域や集団であることが明らかになる可能性
   (c) センシティブ変数への匿名化
   (d) 提供時期と調査時点との差
   (e) その他の匿名化処理の有無

B　ピットマンモデルについて

自然数 $n \in \mathbb{N} := \{1, 2, 3, \ldots\}$ を自然数の和で表す事を分割と呼ぶ。この和の中で自然数 i が足される回数を s_i で表せば、$\boldsymbol{s}_n := (s_1, s_2, \ldots, s_n)$ は（順序無しの）分割を表す。非負整数の集合を $\mathbb{N}_0$ で表すと、n の全ての分割の集合は $\mathcal{S}_{|n} := \{\boldsymbol{s}_n : s_i \in \mathbb{N}_0, i \in \{1, 2, \ldots, n\}, \sum_{i=1}^n i s_i = n\}$ で表される。この集合上の分布が自然数の確率分割である。以下では $u := \sum_{i=1}^n s_i$ とする。

Pitman 分布 (Pitman, 1995) は自然数の確率分割であり、母数 $0 \leq \alpha < 1, \theta > -\alpha$ について確率関数は以下のように書ける。

$$p(s_1, s_2, \ldots, s_n) = n! \frac{\theta^{[u:\alpha]}}{\theta^{[n]}} \prod_{j=1}^n \left(\frac{(1-\alpha)^{[j-1]}}{j!} \right)^{s_j} \frac{1}{s_j!}, \quad \boldsymbol{s}_n \in \mathcal{S}_{|n}, \tag{7}$$

ただし $\theta^{[u:\alpha]} = \theta(\theta+\alpha)\cdots(\theta+(u-1)\alpha)$, $\theta^{[n]} = \theta(\theta+1)\cdots(\theta+n-1)$ である。

個票データとの対応を述べておこう。匿名化の程度を定めることで、キー変数に関する分割表が出来る。分割表の情報のうち度数を、第 j セルについて f_j と書く。ただしセル総数が J として $j \in \{1, 2, \ldots, J\}$ である。ここで $i = 1, 2, \ldots, n$ について度数 i のセルの数を s_i と表す。つまり指示関数 $1(\cdot)$ を使えば、$s_i = \sum_{j=1}^J 1(f_j = i)$ である。例えば s_1 は標本で一意なレコード数となる。このように作られる $\boldsymbol{s}_n$ を「寸法指標」と呼び、$n \ll J$ なら Pitman 分布の標本とみなせる。母集団サイズ $\tilde{n}$ も J よりかなり小さいなら、確率変数 $\boldsymbol{S}_{\tilde{n}} := (S_1, S_2, \ldots, S_{\tilde{n}})$ が Pitman 分布に従う場合、S_1 で母集団一意数の挙動が表せる。

経験ベイズ的に母集団一意数推定の論理を説明すると、以下の通りになる。$\boldsymbol{S}_{\tilde{n}}$ の事前分布が Pitman 分布であり、その母数 (α, θ) は超母数である。超母数はデータ $\boldsymbol{s}_n$ により（最尤）推定される。推定したい母数は S_1 であり、$(S_2, S_3, \ldots, S_{\tilde{n}})$ は局外母数である。母数 S_1 の周辺分布については Hoshino (2012, Theorem 3) を見よ。

Pitman 分布に従う $\boldsymbol{S}_n$ の任意の周辺階乗モメントは、Yamato and Sibuya (2000) が与えている。特に

$$\mathrm{E}(S_i) = \frac{(1-\alpha)^{[i-1]} n^{(i)}}{i!} \theta \left(\frac{(\theta+\alpha)^{[n-i]}}{\theta^{[n]}} \right), \tag{8}$$

である。

超母数の最尤推定量を $(\hat{\alpha}, \hat{\theta})$ と書けば、(8) 式に $\alpha = \hat{\alpha}, \theta = \hat{\theta}, n = \tilde{n}, i = 1$ を代入して母集団一意数の点推定量が得られる。すなわち

$$\hat{S}_1 = \tilde{n} \frac{(\hat{\theta}+\hat{\alpha})(\hat{\theta}+\hat{\alpha}+1)\cdots(\hat{\theta}+\hat{\alpha}+\tilde{n}-2)}{(\hat{\theta}+1)(\hat{\theta}+2)\cdots(\hat{\theta}+\tilde{n}-1)}. \tag{9}$$

なお Hoshino (2001, Proposition 3) によれば、$\alpha \geq 0$ について $\lim_{n\to\infty} \mathrm{E}(S_1)/\mathrm{E}(U_n) = \alpha$ であ

る。ただし $U_n := \sum_{i=1}^{n} S_i$ は度数が 0 でないセルの総数なので、母集団で空でないセルのうち一意のセル数の割合は α と解釈出来る。

次にフィッシャー情報量を確認しておこう。まず対数尤度関数を

$$L(\alpha,\theta) = \sum_{i=1}^{u-1} \log(\theta + i\alpha) - \sum_{i=1}^{n-1} \log(\theta + i) + s_1 + \sum_{i=2}^{n} s_i \sum_{j=1}^{i-1} \log(j - \alpha) + \text{Const.} \tag{10}$$

で表す。二次の微分係数は

$$\frac{\partial^2 L(\alpha,\theta)}{\partial \theta^2} = -\sum_{i=1}^{u-1} \frac{1}{(\theta + i\alpha)^2} + \sum_{i=1}^{n-1} \frac{1}{(\theta + i)^2}, \tag{11}$$

$$\frac{\partial^2 L(\alpha,\theta)}{\partial \alpha^2} = -\sum_{i=1}^{u-1} \frac{i^2}{(\theta + i\alpha)^2} - \sum_{i=2}^{n} s_i \sum_{j=1}^{i-1} \frac{1}{(j - \alpha)^2} < 0, \tag{12}$$

$$\frac{\partial^2 L(\alpha,\theta)}{\partial \theta \partial \alpha} = -\sum_{i=1}^{u-1} \frac{i}{(i\alpha + \theta)^2} < 0 \tag{13}$$

である。(13) 式より $\hat{\alpha}$ と $\hat{\theta}$ は負の相関を持つ。情報量はこれらの式について u を U_n に、s_i を S_i に置き換えて期待値をとる。$\mathrm{E}(S_i)$ は (8) 式で与えられているので、あとは

$$\mathrm{P}(U_n = u) = \frac{\theta^{[u:\alpha]}}{\theta^{[n]}} (-1)^{n-u} C(n,u,\alpha) \alpha^{-u}, \quad u \in \{1,2,\ldots,n\}. \tag{14}$$

を利用して数値的に評価できる。$C(\cdot,\cdot,\cdot)$ は C-ナンバーと呼ばれ、一般化されたスターリング数である。C-ナンバーについては Charalambides and Sing (1988) を参照のこと。Sibuya and Yamato (2001, Proposition 5) がフィッシャー情報量行列のオーダーを評価しており、$n \to \infty$ の時 $I_{\theta\theta} = O(1), I_{\theta\alpha} = O(\log n), I_{\alpha\alpha} = O(n^{\alpha})$ である。特に θ の推定精度は悪い。

C　母集団一意数の推定手順

以下では標本サイズを n、母集団サイズを $\tilde{n}$ と記す。

1. 評価するキー変数とその精度を決める。
2. 決められたキー変数全てについてクロス集計する。つまり（高次元の）分割表を作り、各セルに所属するレコード数（度数）を数える。
    - セル総数 J は、全てのキー変数のカテゴリー数の積である。連続変数でも現実には有限個の表現しかとらず、その表現の数をカテゴリー数と考える。

- 第 j セルの度数を $f_j, j = 1, 2, \ldots, J,$ と書く。以下の結果はインデクス j の付け方に依存しない。

3. 空でないセルの度数の度数（寸法指標）を数える。

   - $i = 1, 2, \ldots, n$ について度数 i のセルの数を s_i と表す。つまり指示関数 $1(\cdot)$ を使えば、$s_i = \sum_{j=1}^{J} 1(f_j = i)$ である。
   - 最大のセルの度数が m ならば、$m < i$ について $s_i = 0$ である。

4. データを生成した構造（確率分布）を推定する。

   - 現実の母集団を無限母集団（超母集団）からの標本とみなす。この場合、手元の標本から超母集団の分布を推定すれば、母集団の挙動も推定される。
   - 超母集団の分布として広義の Pitman モデルを仮定し、その母数を最尤推定する。
   - Pitman モデルは 2 母数 (α, θ) を持ち、α が負の場合と正の場合で分けて考えた方が良い。どちらの場合も $u = \sum_{i=1}^{n} s_i$, $n = \sum_{i=1}^{n} i s_i$ である。
     (a) $0 \leq \alpha < 1, \theta > -\alpha$ について Pitman モデルの確率関数は (7) 式で表される。
     (b) $\alpha < 0$ の場合は (7) 式で $\theta = -J\alpha$ とおき、さらに $-\alpha = \gamma$ とおく。すると一母数の確率関数を得る：

$$p(s_1, s_2, \ldots, s_n) = \frac{n! J! \Gamma(J\gamma)}{\Gamma(J\gamma + n)} \prod_{i=0}^{n} \left(\frac{\Gamma(\gamma + i)}{\Gamma(\gamma) i!} \right)^{s_i} \frac{1}{s_i!}. \tag{15}$$

     ここで $\gamma > 0$ であり、$s_0 = J - u$ である。

   - モデル (7) を「(狭義の)Pitman モデル」と呼ぶ。モデル (15) を「多項ディリクレモデル」と呼ぶ。本来は AIC 等によりデータ依存でいずれかをモデル選択するのが良いが、ここでは簡易的な選択基準を示す：
     (a) 母集団サイズ $\tilde{n}$ が総セル数 J より大の場合、多項ディリクレモデルを用いる。
     (b) その他の場合は Pitman モデルを用いるが、尤度の最大化に失敗する（繰り返し計算が収束しない）場合、多項ディリクレモデルを用いる。
   - 狭義の Pitman モデルの最尤推定は以下のように行えば良い。
     (a) 対数尤度関数は (10) 式で与えられている。
     (b) 最尤推定量は以下の同時方程式の解である.

$$\frac{\partial L(\alpha, \theta)}{\partial \theta} = \sum_{i=1}^{u-1} \frac{1}{\theta + i\alpha} - \sum_{i=1}^{n-1} \frac{1}{\theta + i} = 0,$$

$$\frac{\partial L(\alpha,\theta)}{\partial \alpha} = \sum_{i=1}^{u-1} \frac{i}{\theta + i\alpha} - \sum_{i=2}^{n} s_i \sum_{j=1}^{i-1} \frac{1}{j-\alpha} = 0.$$

(c) $L(\alpha,\theta)$ の最大化は汎用最大化ルーチン (R の optim() 関数等) に任せても良いだろう。

(d) 最尤推定値を自前で評価するなら、二次の微分係数 (11),(12),(13) 式を用いたニュートン＝ラフソン法が適当である。

(e) $c = s_1(s_1-1)/s_2$ として、以下の近似的なモメント推定量を得る。これらをニュートン＝ラフソン法の初期値として使うことが考えられる。

$$\hat{\theta} = \frac{nuc - s_1(n-1)(2u+c)}{2s_1u + s_1c - nc}, \quad \hat{\alpha} = \frac{\hat{\theta}(s_1-n) + (n-1)s_1}{nu},$$

(f) θ の推定は不安定なので、初期値をランダムに変えて繰り返し計算が同じ値に収束するか確認するのが望ましい。

- 多項ディリクレモデルの最尤推定は以下のように行えば良い。

(a) 対数尤度関数は定数を除いて

$$L(\gamma) = -\sum_{i=0}^{n-1} \log(J\gamma + i) + \sum_{i=1}^{n} s_i \sum_{j=0}^{i-1} \log(\gamma + j).$$

(b) 最尤推定値は尤度方程式

$$\frac{dL(\gamma)}{d\gamma} = -\sum_{i=0}^{n-1} \frac{J}{J\gamma + i} + \sum_{i=1}^{n} s_i \sum_{j=0}^{i-1} \frac{1}{\gamma + j} = 0$$

の解である。

(c) $L(\gamma)$ の最大化は汎用最大化ルーチン (R の optimize() 関数等) に任せても良いだろう。

(d) 最尤推定値を自前で評価するなら、二次の微分係数

$$\frac{d^2L(\gamma)}{d\gamma^2} = \sum_{i=0}^{n-1} \frac{J^2}{(J\gamma + i)^2} - \sum_{i=1}^{n} s_i \sum_{j=0}^{i-1} \frac{1}{(\gamma + j)^2}$$

を用いたニュートン＝ラフソン法が適当である。

(e) 尤度関数は単峰であり、それほど初期値に依存せず最大化が可能である。ただ最尤推定値が無限大に発散する事はあり得て、それは確率関数が等確率 J 項分布である事を意味する。また最尤推定値が 0 の場合、狭義の Pitman モデルの方が適切と思われる。

- 狭義の Pitman モデルと多項ディリクレモデルの境界 ($\alpha = 0$) のモデルを Ewens モデルという。Ewens モデルの確率関数は以下の通り：

$$p(s_1, s_2, \ldots, s_n) = n! \frac{\theta^u}{\theta^{[n]}} \prod_{j=1}^{n} \left(\frac{1}{j}\right)^{s_j} \frac{1}{s_j!}. \tag{16}$$

  - 同じデータについて Ewens モデルの最尤推定値を $\hat{\theta}_E$ と書き、Pitman モデルの最尤推定値を $(\hat{\alpha}, \hat{\theta}_P)$ と書く。もし $\hat{\alpha} > 0$ ならば $\hat{\theta}_E > \hat{\theta}_P$.
  - 上の結果は Pitman モデルのチェックに使える。また最尤推定の繰り返し計算の範囲を限定できる。
  - Ewens モデルの尤度関数は単峰であり、最大化は容易である。

5. 同定されたデータ構造の下で母集団一意数の推定値 $\hat{S}_1$ を求める。

   (a) 狭義の Pitman モデルの場合、母数の最尤推定値を $\hat{\alpha}, \hat{\theta}$ と書けば (9) で推定される。

   (b) 多項ディリクレモデルの場合、母数の最尤推定値を $\hat{\gamma}$ と書けば

$$\hat{S}_1 = \tilde{n}(J-1)\hat{\gamma} \frac{((J-1)\hat{\gamma}+1)((J-1)\hat{\gamma}+2)\cdots((J-1)\hat{\gamma}+\tilde{n}-2)}{(J\hat{\gamma}+1)(J\hat{\gamma}+2)\cdots(J\hat{\gamma}+\tilde{n}-1)}.$$

   - これらの推定値はモデルの下での度数 1 のセル数の期待値である。
   - 注 1）Ewens モデルの母集団一意数推定式は、Pitman モデルの推定式に $\hat{\alpha} = 0$ を代入して得られる。
   - 注 2) 等確率 J 項分布の母集団一意数推定値は $\tilde{n}(1 - 1/J)^{\tilde{n}-1}$ である。

参考文献

[1] Charalambides, C.A. and Singh, J. (1988) A Review of the Stirling Numbers, Their Generalizations and Statistical Applications. *Communications in Statistics*, Thoer. Meth., **17**, 2533–2595.

[2] Dale, A. and Elliot, M. (2001) Proposal for 2001 Samples of Anonymized Records: An Assessment of Disclosure Risk. *Journal of the Royal Statistical Society*, Series A, **164**, 427–447.

[3] Domingo-Ferrer, J. and Torra, V. (2001) A Quantitative Comparison of Disclosure Control Methods for Microdata. *Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, Doyle et al. (Eds.), Elsevier, Amsterdam, 111-133.

[4] Duncan, G., Keller-McNulty, S.A. and Stokes, S.L. (2001) Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Technical Report 121, National Institute of Statistical Sciences, Durham, North Carolina.

[5] Elliot, M. J., Skinner, C. J., and Dale, A. (1998) Special Uniques, Random Uniques, and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk. *Research in Official Statistics*, **1**, 53–67.

[6] Elliot, M., Lomax, S., Mackey, E. and Purdam, K. (2010) Data Environment Analysis and the Key Variable Mapping System. *Privacy in Statistical Databases*, Domingo-Ferrer, J. and Magkos, E. (Eds.), LNCS 6344, 138–147, Springer-Verlag, Berlin Heidelberg.

[7] Elliot, M., Mackey, E. and Purdam, K. (2011) Formalizing the Selection of Key Variables in Disclosure Risk. *Int. Statistical Inst.: Proceedings of the 58th World Statistical Congress*, 2777–2784.

[8] Hoshino, N. (2001) Applying Pitman's Sampling Formula to Microdata Disclosure Risk Assessment, *Journal of Official Statistics*, **17**, 499–520.

[9] 星野伸明 (2003) 「超母集団モデルによる個票開示リスク評価」, 統計数理, **51**, 297–319.

[10] Hoshino, N. (2009) The Quasi-multinomial Distribution as a Tool for Disclosure Risk Assessment, *Journal of Official Statistics*, **25**, 269–291.

[11] 星野伸明 (2010)「公的統計ミクロデータ提供制度の課題」, 日本統計学会誌, **40**, 23–45.

[12] 星野伸明 (2012)「公的統計の開示リスク評価—労働力調査の論点」, 『経済統計・政府統計の数理的基礎と応用-I 』, 国友直人・山本拓共編, CIRJE 研究報告書シリーズ, CIRJE-R-10, 40–56.

[13] Hoshino, N. (2012) On the Marginals of a Random Partitioning Distribution. 研究集会「数理統計学の沃野」予稿集, 78–86.

[14] 伊藤伸介 (2012) 「政府統計ミクロデータの提供における匿名化措置—イギリス統計法における法制度的措置と攪乱的手法の適用可能性を中心に—」, 明海大学経済学論集, **24**, 1–14.

[15] 伊藤伸介・磯部祥子・秋山裕美 (2009) 「秘匿性の評価方法に関する実証研究—全国消費実態調査のミクロアグリゲートデータを用いて—」, 統計センター製表技術参考資料, **11**, 12–14.

[16] Marsh, C., Skinner, C., Arber, S., Penhale, P., Openshaw, S., Hobcraft, J., Lievesley, D. and Walford, N. (1991) The Case for a Sample of Anonymized Records from the 1991 Census. *Journal of the Royal Statistical Society*, Series A, **154**, 305–340.

[17] Paass, G. (1988) Disclosure Risk and Disclosure Avoidance for Microdata. *Journal of Business and Economic Statistics*, **6**, 487–500.

[18] Pitman, J. (1995) Exchangeable and Partially Exchangeable Random Partitions. *Probability Theory and Related Fields*, **102**, 145–158.

[19] Sibuya, M. and Yamato, H. (2001) Pitman's Model of Random Partitions. 数理解析研究所講究録, **1240**, 64–73.

[20] 総務省政策統括官（統計基準担当）(2011). 「匿名データの作成・提供に係るガイドライン（平成 23 年 3 月 28 日改正版）」

[21] Sweeney, L. (2002) k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, **10**, 557–570.

[22] 竹村彰通 (1997) 「個票データ開示の理論」, 科学研究費補助金（課題番号 08209102）報告書, 2–25.

[23] U.S. Office of Federal Statistical Policy and Standards (1978). *Report on Statistical Disclosure and Disclosure Avoidance Techniques.* Statistical Policy Working Paper 2, U.S. Department of Commerce, Washington DC.

[24] Yamato, H. and Sibuya, M. (2000). Moments of Some Statistics of Pitman Sampling Formula. *Bulletin of Informatics and Cybernetics*, **32**, 1–10.