科学研究費助成事業　　研究成果報告書

平成　２６　年　　４　月　２１　日現在

機関番号：　１３３０１
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５３０２５１

研究課題名（和文）攪乱・模造データの個票開示リスク評価

研究課題名（英文）The evaluation of the disclosure risk of perturbed or synthetic microdata

研究代表者
星野　伸明（hoshino, nobuaki）

金沢大学・経済学経営学系・教授

研究者番号：００３１３６２７

交付決定額（研究期間全体）：（直接経費）　　3,900,000円、（間接経費）　　1,170,000円

研究成果の概要（和文）：本研究課題の期間中に、攪乱・模造データの個票開示リスク評価について部分的な解を得ることができた。特に受容可能な個票開示リスクの客観的な判定方法の提案が重要な結果と思われる。これにより、匿名化の社会的利用に関するボトルネックが軽減できる。また確率分割の周辺分布について組み合わせ論的評価にも成功し、個票開示リスク測度の厳密な算出に貢献出来た。

研究成果の概要（英文）：Throughout the project a partial progress has been estalished on the evaluation of the disclosure risk of perturbed or synthetic microdata. Especially it is important that an objective method has been proposed on the decision of the threshold value of microdata disclosure risk. It is so because the current practice of that decision is so subjective that anonymization techniques are not widely accepted.
Another progress of this project has been seen in the evaluation of the marginal distribution of a random partitioning distribution. This result employs combinatorial arguments, which lead to an exact derivation of a measure of microdata disclosure risk.

研究分野：　社会科学

科研費の分科・細目：　経済学・経済統計学

キーワード：　官庁統計　匿名化　プライバシー

１．研究開始当初の背景

公的統計などのミクロデータを分析利用に供する場合、調査客体の情報は保護しなければならない。ここでデータの変換による情報保護手法を「匿名化」と呼ぶ。例えば「再符号化」という匿名化手法は、点を（その点を含む）区間に変換する。再符号化されたデータは粗いだけで虚偽ではないと見なされており、再符号化は好んで用いられる。ところが公表される区間中に必ず（点の）観測値が存在するという制約は、特定個体の情報を識別する際の強いヒントとなる。一般に、公表データが虚偽を含まないという制約は極めて強い。従って公表データに虚偽を含める事で匿名性を高めるのは自然な発想であり、そのような手法を「攪乱」と呼ぶ。攪乱は決定論的に行うとバイアスを生ずるので、通常はランダムに行われる。公表データをランダムに生成する手法を「模造」と呼ぶ。攪乱及び模造のサーベイは、星野(2010)にまとめた。

日本では攪乱や模造の実務及び研究の蓄積が薄い。そのような事もあり、現在提供されている匿名データの作成には攪乱も模造も用いられていない。日本の統計調査に関する研究の数少ない例を挙げると、攪乱の一種であるミクロアグリゲーションは伊藤(2008, 統計センター製表技術参考資料 No.10)、伊藤他(2009, 統計センター製表技術参考資料 No.11)が検討している。また一橋大学や統計センターが、現在模造データを試作中である。ただ、いずれも個別、実務的であり、攪乱及び模造の体系的理解を目的としていない。

海外においても、攪乱及び模造については一般性の薄い研究が多い。特に開示リスク評価については、攻撃者の行動シナリオを設けた確率的リンケージの実験で済ませる例が目立つ。確かにこのようなケーススタディの蓄積が有るからこそ、単純な攪乱手法は米国等で普通に使用される。しかし実務的慣行の理論的整理は遅れている。研究計画・方法の項で後述するように、数少ない既存の理論も、匿名化手法が攻撃者にとって既知という非現実的仮定に基づいている。

研究代表者は官庁統計分野では数少ない統計理論家として、匿名化を体系化してきた。主流の匿名化理論では、開示リスクは分割表のセル度数の関数と考えられる。そして再符号化は分割表のセルの併合に他ならない。ここでセル度数を確率変数とみなせば、数理統計の体系が利用可能となり、推定・検定を合理的に行える。これが超母集団モデルアプローチの利点であり、Bethlehem et al. (1990, JASA)以来様々なモデルが提案されている。中でも Hoshino (2001, JOS)で提案した Pitman モデルは、開示リスク評価用モデルでは現時点で最良と思われる。更に Hoshino (2009)ではこれまで提案されてきた様々なモデルを特殊ケースとする分布族（Conditional Compound Poisson, CCP）を提案し、その性質を評価した。また CCP 分布族のメンバーとして擬似多項分布を再発見し、開示リスク評価に必要な結果を導いた。

分布族からのモデル選択という視点を導入する事で、モデルが現実を近似する際のズレによる開示リスク評価の誤差を低減出来る。また CCP 分布族は再符号化について閉じており、匿名性の高低を同一の枠組みで扱える。特に開示リスクの上限が、再符号化の詳細方向への極限として導出可能である。詳細方向への再符号化はセルの分割であり、そのような操作の極限分布は hoshino(2006, 2008)で導出し性質を評価した Limiting CCP(LCCP)分布族である。

このように攪乱及び模造が無い場合の研究は整理が進んでいる。これは公的統計の匿名化が再符号化のみを前提としていた事に対応している。しかし攪乱的模造が日本でも実務的問題として重要性を増す中、本分野に欠ける体系性及び現実性を補う事は喫緊の課題である。従って、方法論的に優れている CCP 分布族の概念に攪乱的模造を統合する事で、研究代表者はこのような要請に応える事とした。

２．研究の目的

本研究プロジェクトは、匿名化手法の詳細が明らかにされないという現実的な仮定の下、攪乱及び模造を含む匿名化の開示リスクを評価する体系を整備する。具体的に述べると、CCP 分布族に攪乱、模造という操作を加えた場合について、確率分布としての挙動を明らかにする。

攻撃者が匿名化手法の詳細を知らないという状況は、攻撃者が匿名化の程度を推定する状況と本研究プロジェクトは定式化する。匿名化の程度はセル度数の分散に比例すると考えられるので、分布論的に分散の挙動を明らかにすれば、推定論の枠組みで問題が整理される。

このような方針の下、疑似多項分布という具体例で攪乱及び模造を分析し、その結果を一般化して CCP 分布族の性質として表現する。

３．研究の方法

研究計画終了時には、CCP 分布族の性質として開示リスクを議論したい。しかし具体例からの抽象化が理論研究の常道である。従って CCP 分布族のメンバーである疑似多項分布を用いて、攪乱的模造のモデル化を開始するのが妥当と考える。なお超母集団からのサンプリングを考えれば、決定論的な匿名化手法も確率的となる。故に研究代表者のアプローチでは模造の有無は重要でなく、攪乱の評価を中心とする。平成２３年度は、代表的な攪乱手法を確率分布への操作として定式化し、疑似多項分布へ適用した場合の挙動を明示

的に表現する事を目標とする。以下では具体的な研究の方針について説明する。

攪乱は分割表上で個体が（セルを）移動する操作と言える。母集団一意など（攪乱が無い場合に）主要なリスク測度はセル度数が小さいものを危険とみなすので、セル度数を大きくするように個体を移動させるのが匿名化の一つの方針である。この移動がランダムとして、小さなセルの度数の期待値を増やすような攪乱手法は、例えばミクロアグリゲーションが挙げられる。しかし、セル度数の期待値を変えない攪乱手法も多い。例えば個体属性にノイズを加える攪乱手法では、セルの度数の期待値が変わらない場合も有る。問題はこのような場合、母集団一意などのリスク測度が平均的に攪乱の安全性を反映しない事である。故に攪乱の開示リスクを評価するには、別の測度を使う必要がある。

例えば計算機科学(Privacy Preserving Data Mining)では、差分プライバシーという概念が存在する。これは一個体の一属性が一単位だけ違う二つの母集団 M_1,M_2 を考え、公表されたデータが M_1 所与で生成される確率と M_2 所与で生成される確率の比を取る。M_1,M_2 を動かした時のこの比の最大値が小さければ、（攪乱されているかもしれない）公表されたデータは真の母集団について情報を持たず、安全と考える。差分プライバシーは情報理論的には自然で、匿名化一般の開示リスク評価に用いる事が出来る。しかし発想としては推測開示（センシティブな情報の区間推定が狭い場合）の抑止が目的であり、データの有用性への配慮を欠く。また匿名化の確率構造が既知としてリスクが評価されるので、匿名化の詳細が公開されない通常の状況と異なる。

統計的開示制限分野での開示リスクの研究は、推測開示ではなく識別開示を主な統制対象としてきた。その中で攪乱の開示リスク評価について、数少ないが理論的研究が存在する。例えば Skinner (2008, in Doming-Ferrer and Saygin (Eds.)“Privacy in Statistical Databases”, LNCS 5262, Springer)は攪乱を含むデータの誤差が前提で、確率的リンケージが成功する可能性を事後確率として表現している。他に Skinner and Shlomo (2007, ISI Invited Paper)は、分割表上での誤分類の確率が既知として同様の考察をした。しかしこれらの研究でも、攪乱手法が既知として開示リスクが評価されている。

先に指摘したように、攪乱手法は通常明らかにされない。そのような不確実性は安全要因として働くので、未知の攪乱の逆変換可能性をリスク測度に反映させる方が妥当である。例えば、攪乱が無い場合に確率的リンケージが成功する可能性と逆変換可能性の積が自然なリスク測度となる。これは Hoshino (2009)で一般的に考察した攪乱の開示リスク評価方法の一例となる。ただしそこでは、逆変換の可能性を具体的に評価する手法まで検討しなかった。

攪乱の逆変換可能性を測る一つのアイデアは、攪乱によりセル度数の分散は増加するので、不自然な分散増加を検出する事である。不自然な増加分は攪乱の程度に対応すると考えられ、これが大きければファイルとして逆変換可能性は低くなるはずだ。なおセル度数の分散の増加は、過分散の程度の増加を意味する。疑似多項分布は過分散の程度を母数として持つので、攪乱の効果を明示的にモデルに入れやすい。ただし攪乱の手法により周辺度数を固定するなどの制約が異なるので、場合分けをして考察を進めるべきだろう。加法的ノイズなど代表的な攪乱手法について、匿名化の程度を疑似多項分布の母数変化として表す事を当初の目的とする。そして疑似多項分布モデルについて数値実験を行い、次の段階へ移行して良いか検証する。

過分散のモデル化が成功すれば、セル度数の分散増加をデータから検出する方法を検討しなければならない。攪乱による分散増加は誤差の一部として現れる。通常は真の構造に誤差が加わったものが観測値と考え、誤差は剰余として定まる。この誤差は、誤記やエディティングに由来する通常の意味での誤差の他、攪乱を含むと考えられる。通常の意味での誤差も攪乱と同様に情報保護効果を持つので、あえて区別する必要はないだろう。

上記の意味での誤差は、データから真のモデル（分布）を推定した後、剰余として計測可能と思われる。しかし剰余の大きさは真のモデルの自由度に依存し、CCP 分布族では自由度を定める自由があるので、制約を加えないと剰余が一意に定まらない。要するに、何を真のモデルとするか、限定的に考える必要がある。

最も限定的なのは、全てのセル確率が等しいセル可換な場合である。この場合はこれまでの研究蓄積で、剰余としての誤差評価は問題ないと考える。特に可換な疑似多項分布の推定論は Hoshino (2009)で整理済みなので、後は数値実験により検証すれば良い。

ただ、出来ればセル確率のバリエーションを許す方向で理論を詰めたい。一つのアイデアは観測度数がゼロのセルについて、セル確率を非零の母数とする事であろう。このようにすれば通常の多項分布（は観測度数ゼロのセル確率はゼロと推定される）の場合とセル確率の推定値が変わる。結果として、標本数を増やした際に新しいセルから個体が現れないという多項分布の応用上の問題が解消される。

以上のような方針で、疑似多項分布の推定論と開示リスク評価の統合を目指す。またこの議論を一般化し、CCP 分布族の性質として記述したい。

４．研究成果

本研究プロジェクトでは攪乱を伴う匿名化の個票開示リスクについて、現実的な評価方法を提案することが出来た。また個票開示リスクについて、小標本の厳密な分布を求めることに成功した。

まず個票開示リスクの現実的な評価方法についてだが、研究の方法の項で述べたとおり、模造の有無は重要でなく、攪乱の評価が本質的である。従って、攪乱のあるデータの個票開示リスク評価方法を定めればよい。

このようなリスク評価は、公表ファイルと元ファイルで最も近いレコードを探すような方法（リンケージ）で行われるのが普通である。しかしこの方法は母集団一意か否かは考慮せず、せいぜい標本一意の検出しか出来ない。おそらくは、母集団一意やリンケージに関する研究蓄積を活かした、より妥当な個票開示リスク評価手法が望ましい。

このような既存研究の統合は、Marsh et al. (1991, JRSS, A)の方法の改良で可能になるというのが、本研究プロジェクトの着眼点である。Marsh らの方法は、攪乱の効果と再符号化による一意数の管理効果を確率評価に反映する。しかし確率評価の方法に妥当性を欠き、実質的に個票開示リスクの定量評価に失敗している。我々は、この失敗している部分を改善するということである。

まず攪乱の効果の計量については、近年のリンケージ研究の成果を用いればよい。また母集団一意数の推定については、研究代表者等が蓄積した超母集団モデルの技術が使える。そして確率の計量が難しい部分（母集団一意の確証が可能となる確率など）は、法的に必要な判断だけなら、実は正か 0 かだけ分かればよい。この判断を、本研究プロジェクトでは統計的推定として定式化した。

個票の公開において法的に問題となる判断は、個体識別が可能か否か、である。この事情は統計法だけでなく、個人情報保護法や外国法でも変わらない。そして現実の攻撃者によって個体識別がなされるかどうかは、個体識別が（今のところ）起きていないという観測可能な事実が統計的証拠になっている。故にこれを用いて推定する枠組みは、現実に広範囲で主観的に行われている判断を客観化出来る。結果として、匿名化が社会に普及する上でのハードルを大きく下げることが可能である。

既存研究では、リスク測度のしきい値は主観的に決定するものとして扱われてきた。本研究プロジェクトの成果は、この重大だが無視されてきた問題を改善する無二の提案である。

個体識別が可能か否かを統計的推定問題として定式化する基本的なアイデアは、個体識別の難易度概念を潜在変数として用いることである。そして難易度が母数（しきい値）以下なら個体識別が可能と考える。これを確率モデルとして記述できるため、個票開示リスクの適当な測度を「難易度」として観測し、識別が起きていない（または起きた）ことを観測すれば、母数は最尤法で推定できる。

このような観測値に基づいての判断は、攻撃者が匿名化の程度を知らないかもしれないという問題意識に対応するものでもある。研究計画で述べたように、現実的な攻撃者を想定しての個票開示リスク評価を考察した結果、データに聞くしかないという結論を得た。そのため、上記の統計的推定論の枠組みでの意思決定が望ましいと判断した。

なお方法の詳細はディスカッションペーパーにまとめた。この論文は投稿済みであり、現在査読者の意向に沿って改訂中である。

このように現実的な個票開示リスク評価手法を定めるという計画は達成されたが、攪乱を CCP 分布族の性質として数理的に記述するという当初の想定アプローチはとらなかった。

しかし分布論的な考察は進み、個票開示リスクの小標本分布の厳密評価に関する成果を得た。CCP 分布に極限操作を施すと、LCCP 分布という確率分割族を得る。この族を含むコルチンモデルという族について、少数の変数の周辺分布の評価方法を、二通り明らかに出来た。

一番目の方法は、周辺モメントの逆転による。これまで逆転公式は明示されていなかったが、階乗モメントの逆転公式を証明することが出来た。

二番目の方法は、コルチンモデルの混合分布としての構造を利用する。正の整数上の独立同一分布の数を、それらとは独立な正の整数上の分布で混合し、さらに総度数を条件付けしてコルチンモデルは得られる。このような構造を見ると、（操作しやすい）独立同一分布について周辺分布を求め、混合と条件付けを経て、確率分割の周辺分布が得られる。この方法での明示的な解を算出した。

これらの分布論的成果は、本研究プロジェクトへの助成金で開催した研究集会の予稿集に書いたが、未だ広く公表していない。今後は数値計算等を補い、公刊論文にする予定である。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　1　件）

(1) Nobuaki Hoshino, Random partitioning over a sparse contingency table, Annals of the Institute of Statistical Mathematics, Vol. 64, pp. 457-474, 2012, DOI:10.1007/s10463-011-0327-8. 査読有.

〔学会発表〕（計　6　件）

(1) Nobuaki Hoshino, Learning from the experience of publishing useful data, The 8th International Workshop on Security **(招待講演)**, 2013/11/19, **那覇市**.

(2) **星野伸明**, **自然数の確率分割における周辺分布・続**, **統計関連学会連合大会**, 2013/9/10, **豊中市**.

(3) **星野伸明**, **官庁統計の情報保護基準**, **人工知能学会全国大会 (招待講演)**, 2013/6/6, **富山市**.

(4) **星野伸明**, **エビデンスに基づいた匿名化**, **統計関連学会連合大会**, 2012/9/10, **札幌市**.

(5) Nobuaki Hoshino, Invitation to mathematical statistical disclosure control, The 2nd IMS APRM, 2012/7/3, **つくば市**.

(6) **星野伸明**, **自然数の確率分割における周辺分布**, **統計関連学会連合大会**, 2011/9/7, **九州大学伊都キャンパス**.

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

http://stat.w3.kanazawa-u.ac.jp/owner/papers.html

６．研究組織

(1)研究代表者
星野　伸明（HOSHINO, Nobuaki）
金沢大学・経済学経営学系・教授
研究者番号：00313627

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：