匿名データ有識者会議（第２回）　議事概要

１　日　時　令和元年６月 14 日（金）　16：00～17：45

２　場　所　総務省第２庁舎６階　特別会議室

３　出席者

【構成員】

情報セキュリティ大学院大学	
情報セキュリティ研究科客員教授	廣松 毅（座長）
中央大学経済学部教授	伊藤 伸介
統計数理研究所長	椿 広計
(一社)新情報センター会長	美添 泰人

【オブザーバー】

一橋大学経済研究所教授	北村 行伸

【事務局】

統計研究研修所長
統計研究研修所次長
統計研究研修所新規情報活用技術研究官
統計局総務課長
政策統括官（統計基準担当）付統計企画管理官

４　議題

（１）統計委員会の審議状況等
（２）匿名データの作成及び検討のスケジュールについて
（３）有用性の指標の検討方法及び複数の匿名データの作成・提供について
（４）その他

５　議事概要

（１）統計委員会の審議状況等

事務局より匿名データの作成・提供に係る手続の変更に関する統計委員会での審議状況及びガイドラインの改正について報告を行った。

以下、主な意見。

・　多人数で匿名データを利用した場合の手数料はどうなるのか。高額となることを懸念している。

→　資料１－１に例示したとおり、従来は利用者１人につき１ファイル

を原則としていたため手数料が高額となり、利用が進まない一因になっていた。このため、統計法の改正が施行された５月１日より、一定のセキュリティ確保を条件に、申出者側での複数回の複製を認めることとし、手数料の額そのものの見直しと合わせて利用しやすい手数料の価格設定となった。

・　セキュリティの観点から、適正管理の要件は何か。
→　申出時に省令に規定されたセキュリティに係る適正管理措置の内容を確認する。適正管理がなされると判断できれば提供を認める運用である。

・　匿名化処理基準を変更する際、統計委員会への諮問は必要か。
→　必要である。どのような変更に対して諮問審議が必要になるかは、まさに議論すべき点である。まずは匿名データの年次の追加について、現行の匿名化処理基準を用いて早急に提供を行い、その後、審議をどのように進めていくかを、本会議を含む関係各所と相談していきたい。

・　学生の利用など、利用目的の拡大に伴い利用方法が変わる可能性があり、付随するセキュリティ上のリスク対策を検討すべきである。法による規制と攪乱手法など技術による抑止が考えられるが、どちらの方法を採るかによって、匿名データの作成方法が変わる。特に、後者は作成方法に強く影響するため、慎重な議論を要する。

（２）匿名データの作成及び検討のスケジュールについて

事務局より匿名データの作成における課題の提示と、その検討スケジュールについて説明があった。また、当該課題の検討のため、匿名有識者会議の下に「匿名データ作成方法ワーキンググループ」及び「共通課題検討ワーキンググループ」の２つのワーキンググループ（以下、「ＷＧ」という。）を開催する旨が報告された。

以下、主な意見。

・　資料に挙げた事項を全部やるとなれば、相当の作業量が予想される。

・　有用性の指標は、利用者が評価するためではなく、提供者が利用者に対して品質を担保するための指標だと考えている。そのため、資料２－１の「利用者が評価する」旨の記載については考慮願いたい。
→　利用者の意見を聞いて、改善点を受け付ける窓口は必要であろう。

・　作成方法は共通課題と密接な関わりがある。議論や研究の成果がまとまったときに、それらを反映した匿名データを作ることとなるが、資料中の作成スケジュールは前後し得るか。

→　作成スケジュールの変更はあり得る。まずは、匿名データの年次の追加について優先して対応し、共通課題の検討状況を踏まえ、その後の作成スケジュールを改めて検討する。

→　現行の匿名化処理基準を用いて作成する匿名データについても、複数系統のファイル作成を念頭に、リサンプリング率などの基準を考慮する必要がある。

→　リサンプリング率も議論が必要と考えている。しかし、匿名データの提供時期の早期化の趣旨を踏まえ、匿名データの年次の追加を最優先とし、これらの作成・提供が終わった後、複数ファイルとの整合性について整理を行う予定である。

→　現行の匿名化処理基準で作成するものについては、できるだけ早い提供を目指す。既に提供した匿名データについての複数系統のファイルの作成の要否については、ＷＧでの論点の一つとする。

・　参考３に掲げた匿名データ作成方法及び共通課題検討の両ＷＧの開催について合意した。本日の議論を踏まえ、検討作業についてはＷＧで行うこととする。

（３）有用性の指標の検討方法及び複数の匿名データの作成・提供について事務局より共通課題検討ＷＧにおける検証の方針が提案された。

以下、主な意見。

・　利用者が匿名データの品質に納得できるような指標があればよいと考えている。匿名化していない元のデータにも誤差は相当程度あるため、統計的分析結果に関する情報の公開の仕方は相当に難しく、不適切な公開方針を採った結果、使い物にならないデータであると利用者が判断してしまうおそれもある。

・　まずは、海外の事例を日本のデータにあてはめ、有用性の尺度の事例として紹介する程度が順当ではないか。

→　資料の２．(1)①からまず着手する予定。どのように公開すべきかについても重要な問題であり、考えていく必要がある。海外の事例を参考に、まず実現可能な事項から試算し、提示していきたい。

・　実際に、解析に大きく影響を与えるのは外れ値である。リサンプリングした標本に、代表値から大きくかい離した外れ値が入っていると、分析結果は多大な影響を受ける。匿名データの利用者が、外れ値や統計的分布の歪みなどの高度な検討事項に対して、どう対処出来るのかを考える必要がある。外れ値の発生の頻度を確認できるデータを提供すべきか、逆に、ソフトな立場で外れ値の多くを除外し、綺麗な結果が出るような匿名データを提供すべきかといった論点がある。

→　一般の分析者にとって外れ値の扱いは難しい。これまでの匿名データは（匿名化処理基準を適用した項目については）匿名化の過程で外れ値が削除されており、ある程度安定した結果が得られると思っている。一方で、元となったデータと比べると、外れ値の影響により全く結果が違うということは起こり得る。ロバストな分析(注1)であれば匿名データと元となったデータとを比較できるが、ロバストでない分析では匿名データと元となったデータとを比較することは不適切である。このため、レンジ、平均又は分散などのロバストではない指標を提供すると、利用者に誤解を与えるため控えたほうが良い。

注１：ロバスト（頑健）な分析とは、少数のはずれ値の存在によって、結果が大きく変化しない統計的分析手法を指す。例えば、中央値は極めて頑健だが平均値は頑健性を持たない。

・　質的尺度及び離散データの匿名化は理論的にも相当難しく、検討すべき点である。

・　膨大なパターンを想定する必要があるため、公表し得る指標の開発については、幾分時間を要することが見込まれる。

・　複数系統のファイルを同時に提供すると、差分攻撃(注2)を受けたときにどのような結果となるか予想し難しい。各個のファイルに関しては匿名化処理基準によって作成されるためリスクが低減されているが、同一の利用者が同時に複数系統のファイルを分析することは制約すべきである。また、地域情報は極めて有用だが、地域情報を提供する場合にはこれまでと違ったノイズを付加するなどの方法を採るべきである。

注２：異なるデータ同士の重複している情報を元に、調査客体を特定してデータを結合することで、元の各個のデータに含まれていない情報を新たに得る情報セキュリティ上の攻撃手法。

→　誓約書を提出しているとはいえ使い方は自由なので、適切な秘匿処理を施さなければ安全性に懸念がある。

→　地域のマクロ情報を用いて層別化し、適切に地域分析できるよう情報を与えるという方法論は検討に値する。

・　ＰＵＦとしての利用者、匿名データの分析から研究のヒントを導き出す利用者、本格的にミクロデータを扱う利用者などを階層化するような視点が必要である。

・　複数系統ファイルを、どのような切り口で作るかという論点がある。経常調査についても複数系統のファイルを作れるのか、周期調査を個人又は世帯単位で抽出するのか、地域情報又は世帯情報の詳細化をするかなど、データの特性をみて判断するものだと思う。海外の状況を見据えながら検討することとなる。

・　諸外国と比較すると、現状、匿名データに関する議論は教育目的で利用する方向に進んでいるように思う。匿名データの利活用の促進をどのような方向に進めるかによって匿名データの作成に求められる要件は異なってくるため、検討を進める上でニーズの把握は必要である。

・　共通課題の解決に向けて様々な可能性があると思うが、そこには当然制約がある。統計研究研修所及び統計局等の人的リソースを考慮すると、どこまで実現可能であるのか。更に、複数系統のファイルを提供するにしても、ニーズを把握しておかなければ、作成しても利用されず、目的が果たされないこととなってしまう。今後、制約とニーズも検討の論点とする。

（４）その他

・　次回は令和元年７月頃を予定。

・　本日の会議資料は、全て統計研究研修所のホームページに掲載。

以上