資料３

匿名データ部会の審議状況について（報告）

第５回匿名データ部会　議事概要

１　日　　時　平成23年２月７日（月）　14：58～17：17

２　場　　所　中央合同庁舎第４号館２階　共用第３特別会議室

３　出　席　者

椿広計部会長、井伊雅子部会長代理、津谷典子委員、廣松毅委員、伊藤伸介専門委員、黒田祥子専門委員、橋本英樹専門委員、安田聖専門委員、石井太氏（国立社会保障・人口問題研究所）、総務省（政策統括官（統計基準担当））、総務省（統計局）、農林水産省、経済産業省、国土交通省、日本銀行

【諮問者（厚生労働省統計情報部）】

中島企画課審査解析室長、上田社会統計課国民生活基礎調査室長、山田企画課審査解析室室長補佐、久住企画課審査解析室匿名データ提供係長

【事務局（内閣府統計委員会担当室）】

若林参事官、谷道参事官補佐

４　議事次第　（１）委員の指名について

（２）国民生活基礎調査に係る匿名データの作成について

（３）その他

５　議事概要

（１）委員の指名について

椿部会長から、資料１に基づき、本部会の委員として津谷委員及び廣松委員が指名された旨の説明があった。

（２）国民生活基礎調査に係る匿名データの作成について

事務局から、前回の部会審議等を踏まえて修正された資料２「『国民生活基礎調査に係る匿名データの作成について』の論点」及び資料３「第42回統計委員会における統計委員会委員の意見」が紹介され、諮問者から論点に沿って説明を受けた後、論点の項目ごとに審議が行われた。

各委員等の主な意見等は次のとおり。

ア　リサンプリング単位と地域区分の関係について

・　世帯レベルでリサンプリングする限り、地域区分を全国一本とすることはやむを得ない。一方、世帯員レベルでのリサンプリングについては、公衆衛生や疫学における有用性だけではなく、地域情報を付加しても特定化されるリスクが十分に低いという意味でも検討すべきではないか。

・　本検討を行った研究会では、世帯員レベルでのリサンプリングについて、乗率を用いない分析に利用する前提であれば、世帯が特定できないように世帯構造が再構成されない形で世帯員単位に全部ばらした上で乗率を考慮しないでリサンプリングし、全国推計を行わないといった制限も掛けれ

ば、８割程度のリサンプリング率で匿名データを作成できるという結論が得られ、地域情報についても提供できる可能性があるということだった。

・　例えば、生活習慣病などで、親がたばこを吸っている場合に子供はたばこを吸いやすいかとか、ソーシャルリレーションシップなどが行動や健康に与える影響について分析する時には、世帯レベルでリサンプリングされたデータを世帯員にばらした方が有用であるものの、一般的な関係を分析する時には、世帯員同士の影響を受ける可能性があるので、世帯員レベルでリサンプリングされた匿名データの方がバイアスの少ないデータの提供が受けられるという点で望ましい。

≪部会長のまとめ≫

・　二段抽出の方法は妥当であり、今回、結果として約２割のリサンプリング率となったことはやむを得ない。今回は、世帯員単位での匿名データ化が諮問されていないが、①公衆衛生・疫学の分野における強い提供ニーズ、②世帯員単位で提供することで地域情報を付加できる可能性など、作成するメリットが大きいので、今後の課題として検討いただきたい。

イ　地域区分の提供について

・　「三大都市圏とそれ以外の地域」などの区分ならば有用性のある情報だと思うが、平成の大合併の途中である平成 16 年調査において、厚生労働省から提案があった「15 万人以上の市及びそれ未満の市及び郡部」という情報を付加しても、他の年次と比較することが困難であり、有用性は期待できない。一方で、地域情報を付加することによる個体識別リスクが高まることを考えると、現段階では地域区分を全国一本とすることはやむを得ない。

≪部会長のまとめ≫

・　「15 万人以上の市及びそれ未満の市及び郡部」という地域区分では、ユーザーにとって有用性が低いと考えられるので、今回は地域区分を全国一本とすることはやむを得ないこととし、今後、「三大都市圏」などといった有用性が見込まれる地域区分の提供について、開示リスクを踏まえながら検討していただきたい。

ウ　世帯員の年齢について

・　年齢について、85 歳以上でトップコーディングしているが、我が国の人口構成において一番急激に増加しているのが「85～89 歳」の階級。人口構成が変化し、この階級の割合が増加することは間違いないので、新しい年次の匿名データを作成するときに「85～89 歳」階級が１％を超えた場合には、トップコーディングする年齢階級を見直しする必要があるのではないか。

≪部会長のまとめ≫

・　トップ（ボトム）コーディングについて、少し余裕を持って１％基準としたことに関しては、経年的な分析の観点から理解できるものの、将来的に人口構成が変化した場合などには、トップ（ボトム）コーディングの閾値について適宜見直すことが必要。

エ　トップコーディングの閾値について

・　今回、匿名データＡ及びＢのトップコーディングの閾値を同じにすることについては、実際の分布を確認したところ大きな差が認められないので了解。ただし、今回のように同一の調査票から複数の匿名データを提供するに当たって、基となる対象サンプルの分布が著しく異なる場合には、トップコーディングの閾値を変えることが、有用性の観点だけでなく、秘匿性の観点からも必要と思われる。

・　トップコーディングの閾値について、本来は母集団による概念であるが、多くの項目では母集団情報がないことから、代替策としてサンプル全体の上位１％でトップコーディングをかけている。匿名データＡ及びＢの母集団はほぼ同一であるため、基本的に閾値は一つで良いと考えている。

≪部会長のまとめ≫

・　今回、匿名データＡとＢの閾値を同一にすることについては妥当であるが、同一の調査票から複数の匿名データを作成する場合にトップコーディングの閾値を母集団から計れない時には、それぞれの対象サンプルの分布に違いが無いかを確認する必要があるのではないか。

オ　外観識別可能性の低い項目の秘匿措置の緩和について

・　総務省統計局の４調査のような試行的提供の実績がないことから、安全性を優先していることは理解できるが、秘匿措置の見直しに関する十分な検証期間とはどのくらいなのか。

・　利用実績の蓄積や利用者の意見を基に、秘匿性と有用性を考慮して検討を行うことになる。総務省統計局の４調査の匿名データについては、提供開始から約２年になろうとしており統計センターのＷＥＢページ上でその成果が公表されているが、研究が終了したものはまだ多くなく、十分な利用実績は蓄積されていないと思われる。それを勘案すると１～２年ということではなく、中期的な時間は必要と考える。

・　「希望する仕事の形」に関しては、秘匿措置をせず提供することに変更するとのことだが、これ以外の項目についてどうなっているのか。健康票にも外観識別が不可能と思われる項目に秘匿措置がかけられていると思われるので確認をお願いしたい。

≪部会長のまとめ≫

・　まだ精査が必要とのことなので、次回の部会で厚労省から回答をお願いしたい。

カ　トップコーディングした階級の平均値などのメタデータの提供について

・　トップコーディングした階級の平均値以外にも、研究者から様々な要望が挙がる可能性があり、そのような対応が多くなると収拾がつかなくなり、匿名データの提供の趣旨から外れてしまうのではないか。

・　現段階では本調査の利用実績が無いことから判断しかねるので、今挙げられているような要望が多くなった場合には、匿名データの提供に併せて平均値等も提供すればよいのではないか。

・　ジニ係数算出のためにはこれらの情報が重要との意見であるが、白書等では、他部局が現物給付等を含めて把握している「所得再分配調査」による再配分後等諸種のジニ係数が公表されているのに対し、本調査のジニ係数は、参考値として年間所得金額ベースで算出しているものであり、両者は異なるものであることをご承知おきいただきたい。

≪部会長のまとめ≫

・　トップコーディングした階級の平均値などの提供は、なかなか難しい問題であり、匿名データの提供を進める中で、研究者のニーズを見ながら判断していく必要がある。

キ　所得票及び貯蓄票に関する事項について

・　所得の内訳は、社会保障や所得格差を分析する者としては非常に重要な情報であり、その内訳が分からないのは研究者にとって非常に重要な問題。今回は安全性を優先するということで必要な措置と理解するが、所得の内訳を提供することについては、総務省統計局の４調査でも提供していないので、各省が連携して提供の方法を検討していただきたい。

・　所得等の量的項目に関する匿名化技法については、トップコーディングだけではなく、ラウンディング、カテゴリー化、ミクロアグリゲーション等幾つかの方法がある。トップコーディングだけでは秘匿が難しい項目についても、他の匿名化技法も組み合わせることによって提供が可能となることも考えられる。例えば、特定の内訳項目に対してカテゴリー化を適用することによって、本調査の匿名データにおいて量的項目の内訳の提供が可能になるかもしれない。
・　一橋大学で匿名データを試行的に提供していたときには、出来る限り調査票に近い形で提供してほしいという意見が多かったので、最低限の秘匿措置としてトップコーディングなどに限定し、スワッピングやパータベーションをかけることは避けてきた。現時点では、加工してほしくないという意見の方が多いのではないか。
・　どのようにデータを加工したらよいかは、研究者や研究目的によっても違いがある。カテゴリー化に関しても、どこで区切るかということに恣意的な側面があるので、年齢の５歳階級であれば、比較的多くの者が一般的であるとの共通認識があるが、そうでなければ加工は最小限にすべきではないか。

≪部会長のまとめ≫

・　利用者にとって所得等の内訳に対する提供ニーズが高いことは理解するものの、開示リスクを考えるとトップコーディング等のみで十分な匿名性を確保することは困難。ノイズの付加などのこれまで採用してこなかった匿名化技法を用いてでも内訳項目を提供すべきか、それともこれまでどおりトップコーディング等のみで対応し、出来るだけ個票に近い形で提供していくのか、いずれの立場をとるかを含めて今後の大きな問題である。

ク　外観識別可能な項目について

・　年齢差の大きい夫婦を削除することに加え、年齢差の大きい親子についても確認すべきではないか。他にもさまざまなパターンはあるように思われる。

≪部会長のまとめ≫

・　世帯人員が８人以上の世帯、３つ子以上の世帯、年齢差の大きい夫婦を削除することは適当。年齢差の大きい親子についても確認するとともに、他にも外観識別可能と思われるパターンがあれば、次回の部会までに指摘いただきたい。

ケ　統計調査以外の情報との関係について、本人が識別できる場合について

・　現状では、匿名データが学術目的または高等教育目的に限定され、利用者を制限した中で制度が運用されている。また、開示請求に対しても匿名データは開示される情報ではない。これらを踏まえると、現時点では、外部情報との照合の可能性については一般人基準で運用することが適当であると思われる。しかし、今後、例えば利用目的等が拡大し、利用者の範囲も拡大することとなった場合には、開示リスクも高まるので、特定人基準で外部情報との照合可能性を考える必要が出てくる可能性はある。
・　行政記録情報とは違い、匿名データは開示請求に対して不開示。また、少なくとも現時点では、利用目的が厳しく制限されている。その意味では、他の外部情報との照合可能性について特定人基準とまでしてしまったり、他人が識別できなくとも本人には識別できるからと言って制限してしまうと、せっかくの匿名データが全く使い物にならなくなる恐れがある。
・　申請手続きを厳しくして、万一悪用された場合の処罰規定を整備する方が建設的ではないか。そのあたりのコスト＆ベネフィットを考えた方がいいのではないか。

・　匿名データに関しては、統計法上の根拠があって提供するものなので、もし自分の利益を得るために利用した場合は、当然統計法上の罰則が適用されることになる。
・　現状の匿名データは、誰でもダウンロードできてしまうパブリックユースではなく、きちんと申請した研究者が研究目的に限定して利用するという条件が付いているので、厳格に特定人からも絶対に特定されないようにする必要はない。今後、利用者に関する条件が緩和されたり変更されたりしたときには、また議論が必要になるというようにまとめればよいのではないか。

≪部会長のまとめ≫

・　一般人基準と特定人基準については、開示リスクをできる限り低くすることを目指すものの、現状、研究者がそれなりの誓約を行った上で使う状況においては、特定人基準として厳格に運用する必要はないであろう。また、「他人からは識別できなくても本人が識別できる場合」についても、利用者や利用目的の範囲が限定されていること、リサンプリング等の秘匿措置を行っており完全には特定できないことなどから、調査へ悪影響が及ばないよう配慮されている。ただし、利用者の範囲を拡大した場合には、基準を厳しくする可能性もあるだろう。

コ　提供時期について

・　総務省の４調査はたまたま５年周期だったので５年経過してから提供することとしているが、本調査は３年周期なので、５年とすると直近の結果が出ないということになり、利用上の不都合が発生する。調査周期はそれぞれ違っているので、今後、提供時期についてどのようにしていくか議論すべき。

≪部会長のまとめ≫

・　本調査は総務省の４調査に比べかなり大きな開示リスクを持っているので、現時点では５年経過後で妥当とするが、有用性の観点から提供期間を短くすることについても検討すべき。

サ　その他の意見について

○　介護票の匿名データ化について

・　介護票の匿名データ作成に当たっては、十分な秘匿措置を図るためにはリサンプリングの必要があるが、リサンプリングによりサンプルサイズが小さくなり有用性が低くなること、小規模標本調査の匿名化に関する研究実績がないことから今回、作成しないとした判断は妥当。匿名化技法としては、ノイズの付加等もあり得るが、本調査に限らず我が国では研究蓄積がないことから、今後研究していく必要があるのではないか。

○　後続調査との関係について

・　本調査を親標本とした後続調査の客体が特定されるリスクについては、①地域情報を削除していること、②出現率の低い項目を削除している等、十分に危険性が低くなっていると考えられ適当。

（３）その他

次回の匿名データ部会は３月８日（火）に開催することとなった。

以上

＜文責　内閣府大臣官房統計委員会担当室　速報のため事後修正の可能性あり＞