第16回 匿名データ部会 議事概要

１　日　　時　平成26年10月17日（金）　13:30～15:00

２　場　　所　合同庁舎８号館８階　特別中会議室

３　出 席 者

（部　会　長）北村　行伸

（委　　員）川崎　茂

（専 門 委 員）伊藤　伸介、川口　大司、村田　磨理子

（審議協力者）総務省（政策統括官（統計基準担当））、財務省、文部科学省、厚生労働省、農林水産省、経済産業省、国土交通省、東京都、千葉県

（諮　問　者）総務省統計局：植山克郎調査企画課長、江刺英信労働力統計室長ほか

（事　務　局）内閣府統計委員会担当室：佐々木健一企画官ほか

４　議　　事

（１）社会生活基本調査に係る匿名データの作成について

（２）その他

５　議事概要

（まとめ）

①匿名データ作成に係る諮問案については、おおむね適当であるが、以下の事項は計画の修正が必要と判断

〇地域情報は、調査票Ａと同様、「三大都市圏」と「その他」の２区分にすること

〇これに伴い、調査票Ａと同様、世帯人員数が多い世帯を削除することについては８人以上の世帯にすることが適当

②論点に挙げられている諮問第13号に対する答申における「今後の課題」への対応については、同答申が他の３調査と一緒であるのに対し、今回の検討は本調査についてのみであることから、答申に記述するよりも議事概要に記述して整理

③今後、匿名データの提供時期の短縮化に向けて努力するとともに、利用者の分析結果やニーズを踏まえ、匿名化性の確保を図りつつ有用性の向上を図る検討を深めていくことが必要

（１）前回部会の指摘事項について

前回の部会審議で指摘された、地域区分について全国１区分か三大都市圏、その他の２区分かについて論点を整理すること、年齢について90歳でトップコーディングした場合のサンプルサイズを示すこと、さらに出現頻度の低いデータの削除数を示すことの３点について、総務省統計局からデータ類の提示があり、審議した。主な質疑等は次のとおり。

・分析の用に資する観点からは、年齢のトップコーディングを９０歳に上げることよりは、三大都市圏か否かという地域区分を入れることを優先した方がよいと思う。また、本調査の場合、介護施設等に入居しておらず同居している健常な人の回答が多いことを考慮すると、分析の際にミスリードされる恐れがあるのでトップコーディングは85歳以上でまとめた方がよいのではないか。
・地域区分の表章について、都市圏か否かの情報は有用性が高い。
・有用性の観点からみると、全国１区分から地域２区分にすることによって有用性が高まることは疑いないが、地域２区分の結果表の精度は保証できていないと思われるので、そのことは、利用者に伝えた方がよいのではないか。
・秘匿性の観点からみると、地域を２区分にすることによって新たに標本一意が出現している分布があるが、これについては問題ないと考えているか。
　→標本中では少ないデータだが、母集団では必ずしも一意ではなく、母集団全体の中で問題がなければ削除しないこととしている。
・子供の多い世帯や母子世帯・父子世帯について、住居の種類とクロスした場合の分布において、国勢調査を基にした削除対象の条件に当てはまるレコードは、ほとんど存在しないという理解でよいか。また、全国１区分の場合と地域２区分にした場合で、削除されるレコードはどの程度変わるのか示してほしい。
　→削除されるレコードは、地域を２区分にしても、少ないものと考えている。
・特異なレコードを削除することで匿名性の確保が出来るならば、全国か否かの選択肢であれば、地域２区分の方がよい。
・年齢のトップコーディングの検討に必要と考える国勢調査の90歳以上の分布は、どのようになっているのか。
　→平成17年国勢調査においては、男性の90歳以上は0.4%程度で、女性の場合は1.2%程度である。
・地域区分については、全国１区分から地域２区分にしても削除されるレコード数にほとんど差がないのであれば、有用性が高まることから、２区分にすべきと考える。年齢のトップコーディングについては、高齢化社会を考えると90歳にすべきと考えるが、調査の対象年が平成13年と18年であること、生活時間を把握している調査であることを考慮すると、85歳であっても止むを得ない感がある。ただし、平成23年からは90歳で行うべきと考える。世帯人員が多い世帯を削除することについては、地域を２区分で提供することにより、秘匿性の危険度が高まるのであれば、８人世帯のレコードが削除されても止むを得ないものと考える。
・地域を２区分にした場合、世帯人員数８人以上の世帯のところで、0.5%基準を下回っているから、８人以上の世帯のレコードを削除の対象にするという理解でよいか。
　→0.5%という基準を下回るだけではなく、全国２区分にすることで分布の層が更に薄くなるように思われる。
・年齢のトップコーディングについて、一般論としては、集計表においても匿名データにおいても、85歳以上を90歳以上にする努力は、標本調査であっても標本サイズが許す限りしてほしい。特に国勢調査のような全数のものであれば、90歳以上が望ましいと思う。しかし、この社会生活基本調査においては、今の段階で今後必ずそうした方

がよいと予断を持って言えるかどうかは難しい。生活行動記録のクオリティーがどれだけ正確であるかという問題があるほか、90歳以上になると健康上の理由から施設等世帯に住む高齢者の割合が高く、調査対象である一般世帯に住む高齢者は健康状態の良い人が標本に多く含まれる可能性もある。90歳以上を別に表章するかどうかということは、それぞれの調査ごとにデータを見て判断することが重要ではないか。

（2）総務省統計局の再提案

各委員、専門委員からの意見を踏まえ、統計局が以下の方針を再提案した。

地域区分について、報告書の結果表に使用しているよりも詳細な地域区分を提供することについては、精度の観点も含め、極めて慎重に臨むべきだと考える。しかしながら、匿名データへのニーズとして、例えば回帰分析のように集計表を用いない分析手法をも視野に入れて、大都市圏とそれ以外の地域というダミー変数が加わることによって有用性が高まるという議論があったことを踏まえ、また、地域を２区分にする場合、既存データからの作成が容易であるということから、匿名性の確保に支障がない範囲で地域を２区分にしたいと考える。

ただし、調査票Ｂの提供は今回が初めてであるということもあり、地域を２区分で提供するとことにより他の匿名化措置において慎重な配慮が必要になってくるのではないかと考える。具体的には、世帯人員と年齢区分については、調査票Ａと同様とし、８人以上の世帯を削除、85歳以上の年齢を一括して示すことにしたい。

なお、出現頻度が低い又は特徴的な値があるレコードの検出方法として、母集団である国勢調査を利用して親の年齢、住宅の所有の関係、子供の数等のクロスから相当する世帯を検出するルールは確立しておきたい。

加えて、地域２区分の結果表に係る精度は保証できないことを利用者に知らしめる措置が肝要ではないかと考えている。

（3）前回部会の指摘事項についての部会長のまとめ

前回の部会から審議してきた「論点」の「１．匿名性及び有用性の確保」については、以下のように取りまとめた。

地域区分については、諮問案は全国としていたが、有用性の観点から「三大都市圏」と「その他」の２区分に変更することが適当と判断した。

情報の削除のうち、出現頻度の低い世帯を削除することについて、９人以上の世帯を削除する案は、地域を２区分で提供することから８人以上の世帯に変更することが適当と判断した。また、母子世帯・父子世帯において子どもの数が多い世帯、三つ子以上がいる世帯など、一定の条件に該当する世帯を削除することも適当と判断した。

分類区分の再編について、年齢のトップコーディングは、一般的には90歳にすることも検討すべきであるが、今回の社会生活基本調査については、85歳とすることは適当と判断した。

（4）前回答申の「今後の課題」

諮問第13号に対する答申における「今後の課題」への対応について、総務省統計局は、「（1）複数の匿名データの作成」については、秘匿性やリソースの観点から対応が難しい、「（2）匿名データの提供時期の短縮化」については、調査から一定の期間を経過することによって個人が特定しづらくなることや、調査報告者の信頼感を考慮すると、最新の調査結果を出すことは不可能、「（3）トップコーディング等が行われた変数」については、本調査において実数による調査項目は存在しないため、該当しないと考える、との説明を行った。

これに対する質疑は、以下のとおり。

・諮問第13号に対する答申は、就業構造基本調査、住宅土地統計調査などを含めた課題であるが、複数の匿名データの作成可能性が考えられる調査もあるのではないか。例えば抽出率を工夫することやパータベーションを適用することなどを考えるべきであると思う。また、複数の匿名データを利用したいニーズはあるので、どういう形であれば複数の匿名データの作成が可能になるのかということも今後、検討がなされればありがたい。

・調査票Ｂについては元々標本が小さく、複数の匿名データを作ることはできないのではないかと思う。

・匿名データの提供時期の短縮化については、引き続き検討してほしい。「前回の調査から５年経過」というのはかなり感覚的なもので、絶対的なものではないと思う。トップコーディング等が行われた変数の扱いは、数量的な変数であれば平均値などを出してもらいたい。

・５年ルールということで、就業構造基本調査は2002年のデータが最新で、2007年のデータは、調査実施から５年経っているが提供されていない。５年と決めたのであれば、５年経ったものに関しては、速やかに公開していただきたい。

・調査票Ｂについては今回初めて出すので、短縮化の話はそぐわないのではないか。

（部会長のまとめ）

・本調査について複数の匿名データを作成することは、現段階では難しいと判断。

・提供時期については、５年以内のものが公開されると調査がしにくいとのことだが、統計法33条を根拠に利用する場合は直近のものを利用する可能性もあるので、提供時期の短縮化に向けて努力してほしい。また、調査実施からデータ提供までのギャップを小さくするためのルールを決めて、委員会で審議しなくても自動的に提供可能となるような仕組みも検討すべきである。

（5）答申の骨子案

答申の骨子案についての主な質疑等は次のとおり。

・特定のレコードを削除しても利用上の特段の不都合を生じさせる集計バイアスはない点について、生活時間の平均値などをレコード削除の前後で比較、検証はしていないが、どう考えるか。

→審議の中で具体的な議論はされなかったが、統計局が80％で仮作成した匿名データを公表結果と比較した結果は、男女別あるいは階層別で大きなかい離がなかったと整理

されている。ただし、特定のレコードを削除して作成されたファイルでは検証されていない。

・レコードがかなりランダムに削られていると考えられるので、それほど平均値は変わらないのではないかというのは容易に想像がつく。

・「削除されるレコードが少数であることから」と追記すればよいのでないか。

・不都合を生じさせるような集計バイアスはないと言い切るよりも、集計バイアスがあったとしても、きわめて無視できる程度のものである可能性が高いという趣旨の表現が適当である。

・平成23年調査の匿名データ作成で改めて検討し直す必要があるというのは、年齢だけではなく情報の削除も条件が変わってくるのではないかと思う。年齢の箇所だけに入れるのは違和感がある。

・諮問第13号に対する答申は４つの統計調査の諮問を受けての対応となっており、今回の答申で記載すると、４調査全てについてと解釈されかねないので、少なくとも答申のどこかに「本調査については」と記載した方がよい。

・諮問第13号に対する答申は４調査まとめて今後の課題を示したが、その対応を逐一記述すると分量が相当増えると思うので、答申に書くよりは議事概要に残すことに止めた方が合理的ではないか。

６　その他

・次回の匿名データ部会は、10月31日（金）13時30分から中央合同庁舎８号館８回特別中会議室で開催することとされた。

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＞