参考２

第１回匿名データ部会　議事概要

１　日　　時　平成 21 年 1 月 26 日（月）16：30～18：30

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

廣松部会長、井伊部会長代理、宇賀臨時委員、椿臨時委員、津谷臨時委員、西郷専門委員、永瀬専門委員、星野専門委員、安田専門委員

総務省（政策統括官室（統計基準担当））、厚生労働省、経済産業省、国土交通省、東京都、千葉県、日本銀行

【諮問者】

杉山総務省統計局統計調査部調査企画課長、栗原総務省統計局統計調査部調査企画課調査官

【事務局】

中島内閣府大臣官房統計委員会担当室長、高木内閣府大臣官房統計委員会担当室参事官

４　議事次第　（１）匿名データ部会の公開について

（２）全国消費実態調査、社会生活基本調査、就業構造基本調査及び住宅・土地統計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

冒頭、部会長、委員、臨時委員、専門委員及び中島内閣府大臣官房統計委員会担当室長のあいさつに引き続き、井伊委員が部会長代理に指名された後、以下の議事が進められた。

（１）匿名データ部会の公開について

事務局から、資料１、２に基づき、匿名データ部会の公開方法等について説明があり、調査客体の特定リスクの防止の観点から、会議及び議事録は非公開に、また、議事概要及び配布資料は公開にすることで了承された。

（２）全国消費実態調査、社会生活基本調査、就業構造基本調査及び住宅・土地統計調査に係る匿名データの作成について

杉山総務省統計局統計調査部調査企画課長から、資料４に基づき、諮問第 13 号「全国消費実態調査、社会生活基本調査、就業構造基本調査及び住宅・土地統計調査に係る匿名データの作成について」の内容説明の後、部会長から当該作成に関する論点（案）が示され、個別の論点に沿って審議が行われた。各委員等の主な意見は次のとおり。

ア　リサンプリングの方法

・　リサンプリングの方法として、世帯ごとにサブサンプルを抽出すると、同じファイルの中に世帯員全員に係る情報が必ず含まれることになることから特定し易くなる危険性が高まるため、個人単

位でサブサンプルを抽出する方法を採り、より特定可能性を低くすることを検討する必要があるのではないか。

・　匿名データを使って世帯の収支等に係る経済分析を行う際などでは、リサンプリングの方法が個人単位でサブサンプルを抽出する方法を採ると十分な分析ができないため、有用性の観点から、世帯単位での抽出方法とすべき。

・　標本調査の中には、結果精度を向上させる目的で人口の少ない地域について標本抽出率を高く設定しているものがあり、そのような調査の場合、当該地域における調査客体の特定リスクが生じることのないよう、標本設計の内容を踏まえて、リサンプリングの方法を検討する必要がある。

・　リサンプリングデータについて、基のデータの解析結果との比較等により、その有効性を確認する必要があるのではないか。

・　80%に設定している住宅・土地統計調査以外の３調査について、一橋大学の試行提供においては、提供データから作成した統計と公表されている統計の整合性を確認したところ大きな乖離がなかったことから、データ利用に支障は無いと判断される。

イ　裾切り及びトップコーディング・ボトムコーディングの基準

・　年齢などの変数において、全体に占める構成比が 0.5%未満と極めて小さい変数はレコード削除又はコーディングを行うとの考え方を採用している。これは、諸外国の基準を参考として設定したとのことだが、そうであれば、就業者の少ない職業など構成比 0.5%未満の変数については全てこのルールを適用すべきではないか。

・　ごく一部の都道府県のデータが構成比 0.5%未満である場合、全ての都道府県のデータを一つ上又は下の階級に統合してコーディングするのではなく、該当する都道府県のデータのみをコーディングすれば良いのではないか。

・　住宅・土地統計調査の「家賃・間代」のトップコーディングについて、各都道府県の上限値を揃える必然性は少なく、工夫の余地がある。

・　就業構造基本調査に係る「親の年齢」については、80 歳以上でトップコーディングすることとしているが、近年 80 歳以上は急激に増加しており、当該項目の非常に多くが 80 歳以上に該当してしまい、項目の意味がなくなる可能性もある。親子関係の分析の観点から、80 歳台も５歳階級別に提供することとし、トップコーディングの上限値を 90 歳以上に引き上げるべきではないか。

ウ　世帯員の年齢のリコーディング

・　就業構造基本調査については、その主な目的が就業行動の把握であるが、その分析に当たり、年齢は大変重要な情報であり、年齢を使った分析は大変多い。したがって、少なくとも主要な世帯員については、各歳別に提供すべきではないか。その場合、特定リスクがあるのであれば、代わりに利用度が低いと考えられる職業分類を大括りにすることも考えられる。

エ　世帯人員による裾切り

・　世帯人員が８人以上の世帯については、全体に占める構成比が小さいことから、そのレコードを削除することとなっているが、当該構成比は全国で見ると小さいとしても、地域別に見ると必ずしも小さくない場合もあるため、有用性が損なわれるおそれがある。

・　世帯人員が８人以上の世帯のレコードについては、一橋大学の試行提供の中では、当該レコードを使って特別な分析が行われたことはなく、また、全体に占める構成比が 0.5 未満と小さくデータ利用に大きなバイアスを与えるものではないことを勘案すると、当該レコードを削除しても問題はないと考える。

オ　その他

（ア）訓練・教育用の匿名データの必要性

・　初めて匿名データを利用する研究者の訓練や大学生の教育に利用するための簡易な匿名データを作成・提供することを考えたらどうか。5,000 サンプル程度で卒業論文にも利用できるようなデータが広く提供されれば、利用者の裾野も大きく広がると思う。

・　簡易なデータで訓練しても、その水準での利用ができるようにしかならず、適当でない。むしろ、ある程度難しい匿名データが扱えるよう訓練しなければならない。

・　匿名データの利用訓練のためであれば、インターネットで公開されている米国のデータが多数あるので、それを利用すれば十分であり、また、大学の研究者自身が訓練用のデータを作成することも可能である。むしろ、大学院生等には研究者と同レベルの匿名データが利用できる機会を与えた方が良い。

（イ）その他

・　全国消費実態調査については、世帯の収支、貯蓄額、資産等に関する詳細な情報を把握しているものであり、かつ、収支等を階級値ではなく実額で把握していることから、秘匿の必要性は極めて高い。これに対し、就業構造基本調査や社会生活基本調査については、漠然としたデータも多いため、調査客体が特定されるリスクは比較的低いのではないか。

・　今回の計画ではそうなっているが、リサンプリングデータセットは、結果の再現性の担保及び匿名性の確保の観点から１調査１セットに固定して提供することが望ましい。

・　どのような匿名データであっても全ての研究者のニーズに対応することは不可能である。したがって、匿名性の確保に関する議論に当たっては、安全性の確保と共にどのような匿名化措置であれば利用者をより一層増やすことができるかという観点から審議すべきである。

・　匿名データの作成に関する審議に当たっては、パブリックコメントを求めること等により、研究者のニーズを広く把握する必要があるのではないか。

（３）その他

次回の匿名データ部会は 2 月 13 日（金）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＞