資料３

匿名データの作成に関する諮問時における統計委員会委員の意見
(第 41 回統計委員会：平成 22 年 12 月 17 日)

１　リサンプリング率について

○ 先行して提供されている総務省統計局の４調査では、リサンプリング率を８割として提供している調査が多いが、本調査ではリサンプリング率を２割としている。かなり開きがあるが、有用性の観点からリサンプリング率を上げた方が良いのではないか。

○ 統計局４調査の場合、最もサンプルサイズが大きい住宅・土地統計調査(350 万世帯）ではリサンプリング率が約１割であることから、リサンプリング率については、元となる統計調査のサンプルサイズも踏まえながら、統計的に有用と思われるサブサンプルのレコード数を議論すべき。（別添参照）

２　リサンプリング単位について

○ 世帯員単位による８割抽出の匿名データについて、厚労省の検討会では提供が可能であるとの研究成果があると聞いているが、今回の諮問では提供しないこととなっている。将来的には段階的な提供を行うのかどうか、今後の方針を伺いたい。

３　地域情報の提供について

○ 先行して提供されている総務省統計局の調査では、住宅・土地統計調査以外の３調査であっても比較的大きな単位での地域情報は残しているが、本調査では一切の地域情報を提供しないこととしている。本調査は集落抽出（クラスターサンプリング）であることから、十分な秘匿措置が必要なことは理解するが、全ての地域情報を削除することが本当に妥当なのか。

４　介護票の提供について

○ 今回、介護票を一切提供しないが、有用性の観点から提供すべきではないか。

（別添）

総務省４調査の匿名データの作成（リサンプリング率関連）について

総務省諮問資料　参考１（抜粋）

共通事項に関する説明

1. リサンプリング

（リサンプリングの必要性）

一般に、調査票情報の全レコードから構成される匿名データよりも、一部のレコードをリサンプリング（再抽出）した匿名データの方が、調査客体が特定される危険性を抑えられる。なぜならば、匿名データが全レコードから構成されるものであれば、調査対象となった客体のレコードは必ず匿名データに存在することが知られてしまう。一方、匿名データがリサンプリングされたものであるならば、当該客体のレコードが存在するとは限らなくなる。

このため、匿名データは、調査票情報のレコードをリサンプリングすることにより作成することする。

（匿名データの母集団に対する大きさ）

匿名データの大きさは分析に必要とされるデータ量等も勘案して、当面、母集団に対して１％以下にすることを目安とする。

（リサンプリングに関する基本的考え方）

一般に、学術研究における実証分析で用いられる社会調査は、大学や民間によって実施されたものであり、その標本の大きさは数千の規模である。社会調査においては、例えばニートや母子世帯といった比率としてわずかな客体を対象とすることもあるが、そのような客体は無作為抽出によりごく少数しか調査できなかったり、登録モニター制などにより偏った標本しか得られなかったりする。

これに対して、公的統計の統計調査は、全体の標本の大きさは数万又はそれ以上の大きな規模で実施しており、学術研究の社会調査に比べて、比率がわずかな客体であってもかなりの数を無作為に抽出できているという特徴がある。

このような公的統計の統計調査が持つ規模のメリットを生かすためには、リサンプリングを行うにしても、標本のうち相当の割合を確保しておくべきである。

以上のことから、匿名データの作成におけるリサンプリング率は、100%（標本のすべてのレコードを使用）よりも低くしつつ、一方で相当割合を確保するために、基本的には標本に対して 80%を目安とする（就業構造基本調査、社会生活基本調査及び全国消費実態調査に対して適用）。

なお、レコードが個人単位の調査（就業構造基本調査及び社会生活基本調査）の場合、リサンプリングは世帯を単位として行い、その上で、リサンプリング率はレコードを単位として 80%を目安とするように行う。

また、住宅・土地統計調査の匿名データについては、その大きさを母集団の 1%以下の範囲でリサンプリング率を 10%とする。

表１　調査別標本及び匿名データレコード数（概数）

	標本の概数	匿名データレコード数の概数 （リサンプリング率・母集団比率）	（参考） 対応する母集団
全国消費 実態調査	５万（世帯）	４万（世帯） （80%，0.1%）	4957 万世帯 （平成 17 年 国勢調査）
社会生活 基本調査	20 万（個人）	16 万（個人） （80%，0.1%）	１億 2777 万人 （平成 17 年 国勢調査）
就業構造 基本調査	100 万（個人）	80 万（個人） （80%，0.6%）	
住宅・土地 統計調査	350 万（住戸・ 世帯）	35 万（住戸・世帯） （10%，0.6%）	5389 万戸 （平成 15 年住宅 ・土地統計調査）

なお、総務省統計局・一橋大学の共同研究では、上記のリサンプリング率にて作成した匿名データを提供しており、これにより統計調査の二次分析は行われていた。全国消費実態調査、社会生活基本調査及び就業構造基本調査についてリサンプリング率を８割より下げることは、匿名データの有用性を低下させることとなる。また、リサンプリング率を８割より上げることは、安全性に影響が及ぶこととなる。

統計委員会答申（抜粋）

１．計画の適否とその理由等

(2) 理由及び修正点

ア 情報の削除

(ア) レコードのリサンプリング

匿名データの作成に当たっては、作成対象４調査の全ての標本のレコード（調査客体）から、世帯単位により、全国消費実態調査、社会生活基本調査及び就業構造基本調査の３調査については 80%を、また、住宅・土地統計調査については 10%を、無作為または各レコードに付された乗率の大きさに基づく確率比例で再抽出（以下「リサンプリング」という。）したもの（以下「サブサンプル」という。）を用いる計画である。

これについては、次の理由等から適当である。

ａ　リサンプリングは、匿名データの中に特定の調査客体が含まれるか否かの判別を困難とする措置であること

ｂ　特に、今回のリサンプリングにおいては、無作為抽出を基本としつつ、各レコードが持つ集計用乗率に抽出地域との一定の対応関係がある場合、当該乗率から抽出地域が特定されてしまうことを防ぐための措置を採っていること

ｃ　世帯単位による抽出は、匿名データの利用者のニーズが高い世帯収支等世帯に着目した分析が可能となるため、個人単位による抽出よりも当該データの有用性が高まること

ｄ　サブサンプルの抽出率は、各調査の母集団の大きさやそれに含まれる情報の内容等を踏まえ設定しているものであり、当該抽出率によりリサンプリングされたサブサンプルから作成された匿名データによる統計と全レコードから作成された公表統計（以下「公表統計」という。）との間で、代表的な項目の平均値や標準偏差に大きな乖離はなく、当該データの有用性が確保されていること