資料３

総統調第３７２号
平成20年12月22日

統計委員会委員長
竹内　啓　殿

総務大臣
鳩山邦夫

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（諮問）

標記について、別紙のとおり作成するに当たり、統計法（平成19年法律第53号）第35条第２項及び附則第３条の規定に基づき、統計委員会の意見を求める。

別紙

諮 問 の 概 要

1　匿名データの作成の対象とする統計調査

今回、総務省は、以下に掲げる統計調査について、統計法（平成 19 年法律第 53 号）第 35 条第 1 項の規定に基づき匿名データの作成を行う予定である。

匿名データを作成する統計調査名	調査年次
全国消費実態調査	平成元、6、11、16 年
社会生活基本調査	平成 3、8、13 年
就業構造基本調査	平成 4、9、14 年
住宅・土地統計調査	平成 5、10、15 年

（説明）

一般に、世帯・個人は事業所・企業よりも特性のばらつきが小さく、外部情報との照合により特定される可能性が低いため、世帯・個人を対象とする統計調査の方が、事業所・企業を対象とするものよりも、識別可能性のリスクの観点からは匿名データの作成は容易であるとされている。一方、世帯・個人を対象とする統計調査であっても、同一の調査客体を複数回継続的に調査するものや悉皆調査については、匿名データの作成が比較的困難であるとされている。

総務省では、このような点を踏まえ、まず、世帯・個人を対象とする上に掲げた統計調査について匿名データを作成することとした。なお、これらの統計調査については、総務省が一橋大学と共同で行ってきた「学術研究のための政府統計ミクロデータの試行的提供」（平成 16～20 年）の研究において、調査票情報に秘匿措置を講じた場合の当該データの安全性、有用性等について、研究してきたものである。

2　匿名データの作成方法の概要

上記 1 に掲げる統計調査について、匿名化措置を講じ、匿名データを作成することとし、その概要については以下のとおりである。

- 元の統計調査のレコードすべてを匿名データに用いるのではなく、それに間引きを施したものを用いる（レコードのリサンプリング）。
- 識別情報は、レコードから全面的に削除する。また、レコードの配列順が意味をなさないように、無作為に並べ替えを行う（識別情報の削除等）。
- 特徴的な識別情報の値があるレコードは、削除する（裾切りによるレコード削除）。
- 極端に大きな値は、上限値を設けて頭打ちにする（トップコーディング）。
- 分類事項の程度は、詳細なものではなく、粗いものとする（リコーディング）。

匿名データの作成方法の概要

資料３－別添１

（１）情報の削除

ア レコードのリサンプリング　元の統計調査のレコードすべてを匿名データに用いるのではなく、それに間引きを施したものを用いる。

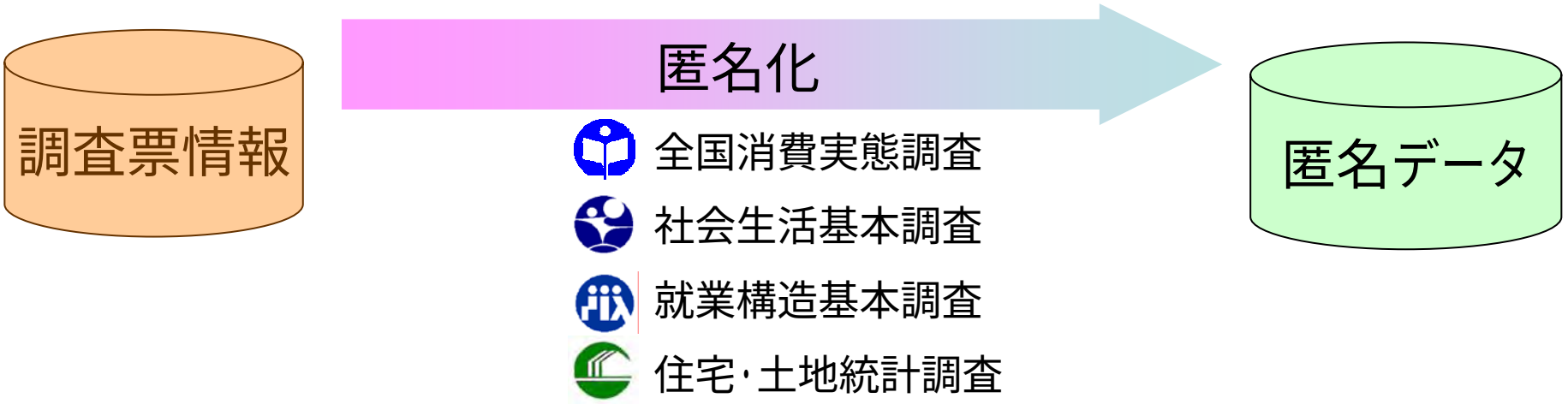
イ 識別情報の削除等　識別情報は、レコードから全面的に削除する。また、レコードの配列順が意味をなさないように、無作為に並べ替えを行う。

ウ 裾切りによるレコード削除　特徴的な識別情報の値があるレコードは、削除する

（２）識別情報の階級区分統合

ア トップ（ボトム）コーディング　極端に大きな値は、上限値を設けて頭打ちにする

イ リコーディング　分類事項の程度は、詳細なものではなく、粗いものとする



匿名データ作成の対象とする「４調査」の概要

１　全国消費実態調査

（１） 調査の概要

① 調査の目的

国民生活の実態について、家計の収支及び貯蓄・負債、耐久消費財、住宅・宅地などの家計資産を総合的に調査し、世帯の消費・所得・資産に係る水準、構造、分布などを明らかにする。

② 調査の周期

昭和 34 年以降 5 年ごとに実施

（２） 調査の対象

全国すべての世帯のうち、総務大臣の定める方法により選定された二人以上の約５万世帯と約５千単身世帯。

（３） 抽出方法

市は全市を調査対象とし、町村は都道府県ごとに標本設計を行い、一部を抽出する。

① 二人以上の世帯及び一般単身世帯

市部　層化２段抽出法

第１次抽出単位：各市の調査単位区（注)、第２次抽出単位：世帯

郡部　層化３段抽出法

第１次抽出単位：町村、第２次抽出単位：調査単位区（注）

第３次抽出単位：世帯

（注） 調査単位区は近接する二つの国勢調査調査区から構成される。

② 寮・寄宿舎単身世帯

層化２段抽出法

第１次抽出単位：30 人以上の規模の会社等の寮・寄宿舎のある国勢調査調査区

第２次抽出単位：世帯

（４） 調査事項（平成 16 年調査）

① 家計簿Ａ

収入（勤労者世帯と無職世帯)、支出

② 家計簿Ｂ

収入（勤労者世帯と無職世帯)、支出、購入先

③ 世帯票

世帯、世帯員及び住宅・土地に関する事項

④ 耐久財等調査票

主要耐久消費財（40 数品目）に関する事項

⑤ 年収・貯蓄等調査票

年間収入、貯蓄現在高、借入金残高などに関する事項

⑥ 個人収支簿

18 歳以上の世帯員（家計簿記入者を除く。）の個人的な収支

2　社会生活基本調査

(1)　調査の概要

① 調査の目的

国民の社会生活の実態を明らかにするための基礎資料を得る。

② 調査の周期

昭和 51 年以降５年ごとに実施

(2)　調査の対象

指定調査区の中から選定する約 8 万世帯にふだん住んでいる 10 歳以上の世帯員約 20 万人（平成 8 年は約 10 万世帯の 10 歳以上の世帯員、平成 3 年は約 10 万世帯の 15 歳以上の世帯員)。

(3)　抽出方法

標本抽出は、層化 2 段抽出法による。

第 1 次抽出単位：国勢調査調査区（47 都道府県ごとに、人口に基づく確率比例系統抽出）

第 2 次抽出単位：世帯（等確率系統抽出により、各調査区から 12 世帯を抽出）

(4)　調査事項（平成 13 年調査）

① すべての世帯員に関する事項

出生の年月又は年齢、世帯主との続き柄 、在学、卒業等教育又は保育の状況 等

② 10 歳以上の世帯員に関する事項

氏名及び男女の別、配偶者の有無、ふだんの介護の状況、携帯電話やパソコンなどの使用の状況、インターネットの利用の状況、学習・研究活動の状況、スポーツ活動及び趣味・娯楽活動の状況、ボランティア活動の状況、旅行・行楽の状況、1 日の生活時間配分の状況及び天候（調査区ごとに連続する 2 日間を定める。）等

③ 15 歳以上の世帯員に関する事項

ふだんの就業状態、従業上の地位及び雇用形態、仕事の種類、ふだんの 1 週間の就業時間、勤め先・業主などの企業全体の従業者数、ふだんの片道の通勤時間、週休制度 等

④ 60 歳以上の世帯員に関する事項

子どもの住んでいる場所

⑤ 世帯に関する事項

住居の種類、居住室数、自家用車の有無、年間収入、介護支援の利用の状況、不在者の有無等

3　就業構造基本調査

(1) 調査の概要

① 調査の目的

国民の就業及び不就業の状態を調査し、全国及び地域別の就業構造に関する基礎資料を得る。

② 調査の周期

昭和 31 年から 57 年まで概ね 3 年ごと、昭和 57 年以降は 5 年ごとに実施

(2) 調査の対象

抽出単位（世帯が居住することができる建物又は建物の一部）に居住する約 40 万世帯の 15 歳以上の世帯員。

(3) 抽出方法

標本抽出は、層化 2 段抽出法による。

第 1 次抽出単位：国勢調査調査区（以下「調査区」という。）（市区町村ごとに調査区を産業別の人口等によって分類することで層化し、層ごとに 15 歳以上人口をウエイトとして調査区を不等確率系統抽出）

第 2 次抽出単位：住戸（等確率系統抽出により、各調査区の中から住戸を抽出）

(4) 調査事項（平成 14 年調査）

① 15 歳以上の世帯員に関する事項

ア　全員について

氏名、男女の別、配偶者の有無、世帯主との続き柄、出生の年月及びふだんの就業・不就業状態　等

イ　有業者について

従業上の地位、勤め先の事業の内容、仕事の内容、週間就業時間、年間収入、１年前の就業・不就業状態及び前職の有無、（前職ありの場合）離職の時期　等

ウ　無業者について

就業希望の有無、就業希望の理由、希望する仕事の種類、（前職ありの場合）離職の時期、従業上の地位　等

② 世帯に関する事項（世帯主のみ記入）

15 歳未満の年齢別世帯人員、15 歳以上の世帯人員、世帯の収入の種類及び世帯全体の年間収入

4　住宅・土地統計調査

「住宅・土地統計調査」は、昭和 23 年以来 5 年ごとに実施してきた「住宅統計調査」の調査内容等を平成 10 年調査時に調査名を含め変更している。

(1) 調査の概要

① 調査の目的

住宅及び住宅以外で人が居住する建物に関する実態並びに住宅等に居住している世帯に関する実態を調査し、住生活関係諸施策の基礎資料を得る。

② 調査の周期

昭和 23 年以降５年ごとに実施

(2) 調査の対象

抽出した住宅及び住宅以外で人が居住する建物並びにこれらに居住している約 350 万住戸・世帯（平成 15 年調査。平成 10 年以前は約 400 万住戸・世帯）。

(3) 抽出方法

平成 15 年　層化２段抽出法

第１次抽出単位：国勢調査調査区

第２次抽出単位：抽出された標本調査区を基本とする調査単位区内の住戸（抽出された標本調査区、それが大きい場合はそれを分割した調査単位区を設定。一つの単位区はほぼ 50 戸）

平成 10 年以前　層別２段集落抽出法

第１次抽出単位：国勢調査調査区

第２次抽出単位：単位区（抽出された標本調査区を分割して設定。一つの単位区の住戸数はほぼ 25 戸）

(4) 調査事項（平成 15 年調査）

[調査票甲及び乙における共通の調査事項]

① 住宅等に関する事項

居住室の数及び広さ、所有関係に関する事項、敷地面積　等

② 住宅に関する事項

構造、階数、建て方、種類、建築時期、床面積、家賃又は間代に関する事項　等

③ 世帯に関する事項

世帯主又は世帯の代表者の氏名、種類、構成、年間収入

④ 家計を主に支える世帯員又は世帯主に関する事項

従業上の地位、通勤時間、現住居に入居した時期、前住居に関する事項　等

⑤ 住環境に関する事項

敷地に接している道路に関する事項

［調査票乙における調査事項］

⑥ 現住居以外の住宅及び土地に関する事項

所有関係に関する事項、所在地、面積に関する事項　等

資料３－別添３

「４調査」の匿名データ作成方法の共通事項

統計法（平成 19 年法律第 53 号）第 35 条第１項の規定に基づき、総務省統計局が作成する予定の下記１の四つの統計調査の匿名データの作成方法の共通事項は下記２のとおりとする。

記

１　匿名データを作成する統計調査

	標本の概数	備考
全国消費実態調査 （平成元、６、11、16 年）	5 万（世帯）	世帯員に関する情報あり。
社会生活基本調査 （平成３、８、13 年）	20 万（個人） ［８万（世帯）］	個人が属する世帯に関する情報あり。
就業構造基本調査 （平成４、９、14 年）	100 万（個人） ［40 万（世帯）］	個人が属する世帯に関する情報あり。
住宅・土地統計調査 （平成５、10、15 年）	350 万（住戸・世帯）	住戸の土地、住居に居住する世帯及び世帯員に関する情報あり。

2　匿名データ作成の共通事項

（1）　リサンプリング

匿名データは、基本的には、統計調査の標本のレコード（客体）から 80%を目安に無作為にリサンプリング（再抽出）したものとする。また、匿名データの大きさは、当面、母集団の１％以下とする。

（2）　識別情報

ア　地域区分

匿名データの各レコードに付与する地域区分は、基本的には、全国６ブロックとする。ただし、リサンプリング率を 80%よりも低くする場合には、これよりも細かくした区分の付与を行うことも検討する。

イ　個人の年齢

匿名データの各レコードの個人の年齢は、基本的には、５歳階級のグルーピング（階級化）を行い、かつ、一定年齢でトップコーディング（一律の上限値を与える）を行う。

ウ　世帯

①　世帯人員

世帯人員が８人以上の世帯のレコードは、削除する。

②　同一年齢の子どもの数

三つ子以上がいる世帯のレコードは、削除する。

エ　その他

これら以外にも、匿名データの各レコードが識別される危険性を低減するために、レコードの削除やトップ（ボトム）コーディング、リコーディングなど、必要な措置を行う。

トップ（ボトム）コーディング、リコーディングに当たっては、統計調査の本体集計の結果表に用いられる表章、分類を参考にする。

参考１

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。

2. 識別情報

（1）地域区分

一般に、統計分析では、調査客体の属性に着目して分析を行う構造分析と、調査客体の地域に着目して分析を行う地域分析の、二つのアプローチがある。

公的統計の統計表編成においては、全国レベルの集計では属性に関する詳細な分類を用いるのに対して、地域別の集計ではそれよりも粗い分類を用いることがある。これは、集計する統計の精度とともに、地域と属性の両方で詳細に集計し過ぎることにより調査客体が識別される可能性を回避することにも配慮したものである。

このため、今回の匿名データの作成に当たっては、初回ということもあり、以下のような地域区分を採用する。

（就業構造基本調査、社会生活基本調査、全国消費実態調査）

80％リサンプリングのこれら３調査についての匿名データでは、相当の割合のレコード（調査客体）が抽出されていることに配慮する必要がある。これらの調査は、個人・世帯に着目したものであり、個人に関して職業分類や産業分類といった属性情報も付されることから、レコードに付与する地域区分は都道府県のレベルを避けることとし、基本的には、全国をいくつかのブロックに分割したものとする。

具体的には、①都道府県レベルでの特定を避けるため北海道や沖縄を単独で表章しないこと、②編成するブロックの間で人口規模に過大な差が生じないようにすることを条件として、地域区分を６ブロックとする（ブロック編成は表２参照）。

（住宅・土地統計調査）

10％リサンプリングとする住宅・土地統計調査の匿名データでは、他の３調査に比べてかなりのレコードを元の標本から間引くこととなる。すなわち、標本のレコードが匿名データに含まれる可能性は１割しかないことから、匿名データ中のレコードが調査客体として識別されるリスクは他に講じる匿名化措置と併せることで相当抑えられる。

このため、10％リサンプリングの住宅・土地統計調査の匿名データでは、そのレコードに付与する地域区分は都道府県のレベルとする。

なお、住宅・土地統計調査と就業構造基本調査とで比較すると、匿名データのレコード数（概数）がそれぞれ 35 万、80 万である一方、その地域区分はそれぞれ 47 都道府県、6 ブロックとなり、住宅・土地統計調査は、就業構造基本調査に比べてレコード数が少ないにもかかわらず地域区分が詳細、というように、一見して逆転状態となっている。しかし、これは、住宅・土地統計調査のリサンプリング率を 80％ではなく 10％に抑えていることから、住宅・土地統計調査はその分だけ匿名データ

のレコードが識別されるリスクを低下させていることを踏まえたものである。

なお、総務省統計局・一橋大学のミクロデータ試行的提供において、開始当初（平成 16 年）の匿名データ（就業構造基本調査・社会生活基本調査・全国消費実態調査）に付与していた地域区分は、慎重を期して、大都市圏又は大都市圏以外の 2 分法を取っていた。その後、平成 18 年に 6 ブロック（就業構造基本調査・社会生活基本調査・全国消費実態調査）とした。なお、同年に新たに提供を始めた住宅・土地統計調査の匿名データの地域区分は、47 都道府県である。

表２　地域６ブロック、都道府県別人口及び世帯数（平成 17 年国勢調査）

	人口	構成比（%）	世帯数	構成比（%）
全　国	**127,767,994**	**100.0%**	**49,566,305**	**100.0%**
北海道・東北	**15,262,654**	**11.9%**	**5,729,566**	**11.6%**
01　北海道	5,627,737	4.4%	2,380,251	4.8%
02　青森県	1,436,657	1.1%	510,779	1.0%
03　岩手県	1,385,041	1.1%	483,926	1.0%
04　宮城県	2,360,218	1.8%	865,200	1.7%
05　秋田県	1,145,501	0.9%	393,038	0.8%
06　山形県	1,216,181	1.0%	386,728	0.8%
07　福島県	2,091,319	1.6%	709,644	1.4%
関東	**44,575,465**	**34.9%**	**18,027,536**	**36.4%**
08　茨城県	2,975,167	2.3%	1,032,476	2.1%
09　栃木県	2,016,631	1.6%	709,346	1.4%
10　群馬県	2,024,135	1.6%	726,203	1.5%
11　埼玉県	7,054,243	5.5%	2,650,115	5.3%
12　千葉県	6,056,462	4.7%	2,325,232	4.7%
13　東京都	12,576,601	9.8%	5,890,792	11.9%
14　神奈川県	8,791,597	6.9%	3,591,866	7.2%
19　山梨県	884,515	0.7%	321,261	0.6%
20　長野県	2,196,114	1.7%	780,245	1.6%
北陸・東海	**20,560,076**	**16.1%**	**7,386,655**	**14.9%**
15　新潟県	2,431,459	1.9%	819,552	1.7%
16　富山県	1,111,729	0.9%	371,815	0.8%
17　石川県	1,174,026	0.9%	424,585	0.9%
18　福井県	821,592	0.6%	269,577	0.5%
21　岐阜県	2,107,226	1.6%	713,452	1.4%
22　静岡県	3,792,377	3.0%	1,353,578	2.7%
23　愛知県	7,254,704	5.7%	2,758,637	5.6%
24　三重県	1,866,963	1.5%	675,459	1.4%
近畿	**20,893,067**	**16.4%**	**8,246,987**	**16.6%**
25　滋賀県	1,380,361	1.1%	479,217	1.0%
26　京都府	2,647,660	2.1%	1,079,041	2.2%
27　大阪府	8,817,166	6.9%	3,654,293	7.4%
28　兵庫県	5,590,601	4.4%	2,146,488	4.3%
29　奈良県	1,421,310	1.1%	503,068	1.0%
30　和歌山県	1,035,969	0.8%	384,880	0.8%
中国・四国	**11,762,204**	**9.2%**	**4,523,175**	**9.1%**
31　鳥取県	607,012	0.5%	209,541	0.4%
32　島根県	742,223	0.6%	260,864	0.5%
33　岡山県	1,957,264	1.5%	732,346	1.5%
34　広島県	2,876,642	2.3%	1,145,551	2.3%
35　山口県	1,492,606	1.2%	591,460	1.2%
36　徳島県	809,950	0.6%	298,480	0.6%
37　香川県	1,012,400	0.8%	377,691	0.8%
38　愛媛県	1,467,815	1.1%	582,803	1.2%
39　高知県	796,292	0.6%	324,439	0.7%
九州・沖縄	**14,714,528**	**11.5%**	**5,652,386**	**11.4%**
40　福岡県	5,049,908	4.0%	2,009,911	4.1%
41　佐賀県	866,369	0.7%	287,431	0.6%
42　長崎県	1,478,632	1.2%	553,620	1.1%
43　熊本県	1,842,233	1.4%	667,533	1.3%
44　大分県	1,209,571	0.9%	469,270	0.9%
45　宮崎県	1,153,042	0.9%	451,208	0.9%
46　鹿児島県	1,753,179	1.4%	725,045	1.5%
47　沖縄県	1,361,594	1.1%	488,368	1.0%

（２）個人の年齢

年齢は、個人を識別する重要な情報であり、年齢を各歳別（１歳刻み）で指定することで対象となる人口はおおよそ 100～200 万人に絞り込むことができる（人口全体の約 1～2%）。さらに、世帯を構成する世帯員それぞれについて年齢を特定することにより世帯の絞り込みも可能である。世帯の世帯人員が多ければ、それだけ年齢を指定する次元が増し、世帯や世帯員が識別される危険性は高まる。

このように、世帯に属する世帯員の年齢を各歳別で示すことは、世帯が識別されるリスクを高める可能性がある。

一方で、統計調査の個票データでは、各レコードに世帯に関する事項は必ず収録され、そこで年齢という情報が重要な役割を持つことは論を待たない。

このことから、統計調査の匿名データの作成に当たり個人及び世帯の識別リスクを低減するために、年齢は１歳刻みではなく、５歳階級にグルーピングすることとする。また、一定値を上回る高齢者の年齢については、トップコーディングを施す。

なお、子どもに関しては、１歳刻みで所属する世帯やその個人の経済社会活動に与える影響は大きく、育児問題に関連した分析においてとりわけ重要な意義を持つ。また、子どもは非就労であることから、職業分類といった情報までは付与されない。このため、子どもの年齢は５歳階級グルーピングを必ずしも適用しないこととする。

表３　年齢５歳階級別人口比率（時系列）（国勢調査）

	構成比%			
	平成2年	7	12	17
0歳～	5.3	4.8	4.7	4.4
5～	6.0	5.2	4.7	4.6
10～	6.9	6.0	5.2	4.7
15～	8.1	6.8	5.9	5.1
20～	7.1	7.9	6.6	5.8
25～	6.5	7.0	7.7	6.5
30～	6.3	6.5	6.9	7.6
35～	7.3	6.2	6.4	6.8
40～	8.6	7.2	6.1	6.3
45～	7.3	8.5	7.0	6.0
50～	6.5	7.1	8.2	6.9
55～	6.2	6.3	6.9	8.0
60～	5.5	6.0	6.1	6.7
65～	4.1	5.1	5.6	5.8
70～	3.1	3.7	4.6	5.2
75～	2.4	2.6	3.3	4.1
80～	1.5	1.8	2.1	2.7
85～	0.7	0.9	1.2	1.4
90～	0.2	0.3	0.4	0.7
95～	0.0	0.1	0.1	0.2
100歳～	0.0	0.0	0.0	0.0

（3）世帯

世帯を構成する世帯員の人数、それぞれに関する年齢、続柄などの情報は、世帯の外部から比較的容易に把握可能な属性である。特定の世帯の属性を指定したとしても、同じパターンが多数存在するのであれば、そのことによって世帯を識別することは不可能である。

しかし、以下に掲げる極端なケースについては、世帯を識別し得る可能性が生じるため、匿名化を施しておく必要がある。

ア　世帯人員８人以上の世帯について、レコードを削除

世帯人員は外見上分かりやすい情報であり、８人の世帯は0.5%以下と希少である。このため、世帯人員が８人を超える場合は、その世帯に属するレコードを匿名データからは削除することとする。

表４ 世帯人員別世帯数（一般世帯）（国勢調査）

年次		一般世帯 総数	1人	2	3	4	5	6	7	8	9	10人以上
平成2年	1990	100.0	23.1	20.6	18.1	21.6	9.4	4.7	2.0	0.5	0.1	0.0
7	1995	100.0	25.6	23.0	18.5	18.9	8.0	3.9	1.7	0.4	0.1	0.0
12	2000	100.0	27.6	25.1	18.8	16.9	6.8	3.1	1.3	0.3	0.1	0.0
17	2005	100.0	29.5	26.5	18.7	15.7	5.8	2.5	1.0	0.2	0.1	0.0

イ　三つ子以上のいる世帯について、レコードを削除

人口動態統計によると、分娩１万件当たりに対して出生児数が３以上となるのは約２件（0.02%）となっている。すなわち、世帯内に三つ子、四つ子（又はそれ以上）がいる場合は、かなりまれである。

このため、世帯に同一年齢の子どもが３人以上いる場合は、その世帯に属するレコードを匿名データからは削除することとする。

表５ 出生児数別分娩件数 （人口動態統計を基に加工）

	平成７年 1995	12 2000	17 2005	18 2006
総数（不詳含む）	1,215,174	1,216,168	1,081,393	1,110,448
1出生児	1,168,268	1,167,787	1,036,816	1,064,878
2出生児	9,608	11,220	11,628	11,806
3出生児	280	298	206	219
4出生児	24	4	3	0
5出生児	1	0	0	0

参考２

匿名データの作成に関する基本的考え方及び匿名化の技法

1．　匿名データの作成に関する基本的考え方

匿名データの作成・提供は海外で広く行われている。

「匿名データの作成・提供に係るガイドライン」（総務省政策統括官（統計基準担当）決定予定）では、「別紙２　匿名処理の技法」において

「実際に海外で行われている匿名化処理の方法をみるとかなり詳細なデータをそのまま提供しているのが普通である。匿名化処理は、論理的に可能性だけを考えると極めて厳しく行わなくてはならないことになるが、実際には、秘匿の必要性や利用面も考慮して現実的な判断の下で決定している。」

「そのような現実的な判断を行うために、海外では権威ある委員会などが処理の方法を最終承認する方式をとっている。」

とされている。、

総務省統計局では、上記のガイドラインの「別紙３　匿名処理の目安」を参考にしつつ、以下の考え方により、匿名データの作成方法を検討することとしている。また、これらの作成方法については、統計法（平成 19 年法律第 53 号。以下「新統計法」という。）に基づき、統計委員会に諮問することとしている。

（１）匿名データの安全性

匿名データの作成に当たっては、統計調査の秘密を守り、調査客体が識別できないようにする。

さらに、安全性に加えて、調査客体の安心感についても留意し、統計調査に対する不信感を招くことのないように配慮する。

（２）匿名データの有用性

匿名データの作成に当たっては、学術研究や高等教育における具体的なニーズを勘案することとする。

（３）匿名データの試行的提供との関連

一橋大学との共同研究「学術研究のための政府統計ミクロデータの試行的提供」（平成 16～20 年）では、就業構造基本調査、社会生活基本調査、全国消費実態調査及び住宅・土地統計調査の４周期調査の匿名データが提供された。この試行的提供

では、現行の統計法（昭和 22 年法律第 18 号）に基づく目的外使用の枠組みに沿って高度な公益性に関する審査や官報への告示を要し、その利用者の範囲は大学の専任講師以上であった。これまで、延べ 100 件超の匿名データが提供されており、それを利用した分析により種々の論文が学会誌や紀要へ投稿されるなどしている。この試行的提供では、匿名データが外部に漏洩したり、調査客体が特定されたりするといった事案は起きていない。

一方、新統計法に基づく匿名データの作成・提供においては、学術研究又は高等教育を目的とする利用が認められており、利用者は民間企業の研究者や大学等の学生にまで広がる予定である。

新統計法に基づく匿名データの作成・提供においては、試行的提供に係る経緯と実績を踏まえつつ、匿名データの提供先が広がることに留意して、特に秘密の漏洩等を防止するための措置については万全を期すこととする。

2. 匿名化の技法（別紙参照）

調査票情報に対して適用する匿名化技法には、識別情報のうち直接的な識別子（氏名、住所など）を削除すること以外に、一般に、(1) 情報の削除、(2) 識別情報の階級区分統合、(3) 識別情報に対する真実と異なる修正・加工がある。

匿名データの作成に当たっては、これらの技法が複合的に適用されることにより行う。したがって、匿名データの作成方法に関する議論の際には、個別の技法を単独で取り上げて行うべきものではなく、適用される技法全体を総体で認識して行うべきものである。

（別紙）匿名化の技法

（１）情報の削除

ア レコードのリサンプリング

元の統計調査のレコードすべてを匿名データに用いるのではなく、それに間引き（リサンプリング）を施したものを用いる（レコードの一部の削除、とも言える。）

イ 識別情報の削除等

調査客体が直接的に特定できるような識別情報は、レコードから全面的に削除すること。また、レコードの配列順が意味をなさないように、無作為に並べ替えを行う。

ウ 裾切りによるレコード削除

レコードに、極端に大きな数など特徴的な識別情報の値がある場合、客体が識別される可能性が高くなるため、当該レコードを作為的に削除すること

（２）識別情報の階級区分統合

ア トップ（ボトム）コーディング

極端に大きな（小さな）値は、上限（下限）値を設けて頭打ちにすること（上限（下限）値に置き換える）

イ リコーディング

分類事項の程度を詳細なものではなく粗いものとすること、又は、連続値を階級区間で表章し直すこと

（３）識別情報に対する真実と異なる修正・加工

ア 誤差の導入

資産額といった開示リスクの高い識別情報をそのまま使用せず、誤差（ノイズ）による加算、乗算等を行うこと

イ データスワッピング

異なる地域のデータセットから、あらかじめ指定した識別情報と完全一致するレコードを抽出し、当該レコード間の他の識別情報の一部（又は全部）をスワッピング（交換）すること