資料１

総務省統計局における匿名データの作成･提供の概要

平成２９年１月３１日
総務省統計局統計調査部調査企画課

１ 匿名データとは

匿名データの作成及び提供は、学術研究の発展や、高等教育の発展に資することを目的に、統計法第 35 条及び第 36 条の規定に基づいて行われる。

総務省統計局では、統計調査を通じて得られた情報を、特定の個人・法人等が識別されないように匿名化処理を行って提供している。

（1）匿名データの定義 （統計法第2条）

匿名データ	一般の利用に供することを目的として調査票情報[注)]を特定の個人又は法人その他の団体の識別（他の情報との照合による識別を含む。）ができないように加工したもの

注）調査票情報 ： 統計調査によって集められた情報のうち、文書、図画又は電磁的記録（電子的方式､磁気的方式その他人の知覚によっては認識することができない方式で作られた記録をいう。）に記録されているもの

（2） 匿名データの作成（統計法第35条）

行政機関の長又は届出独立行政法人等は、その行った統計調査に係る調査票情報を加工して、匿名データを作成することができる。
行政機関の長は、前項の規定により基幹統計調査に係る匿名データを作成しようとするときは、あらかじめ、統計委員会の意見を聴かなければならない。

（3） 匿名データの提供（統計法第36条）

行政機関の長又は届出独立行政法人等は、学術研究の発展に資すると認める場合その他の総務省令で定める場合（※）には、総務省令で定めるところにより、一般からの求めに応じ、前条第一項の規定により作成した匿名データを提供することができる。

※ 匿名データの利用要件（統計法施行規則第１５条）

一 **学術研究の発展に資すると認められる場合**であって、次に掲げる要件のすべてに該当すると認められる場合 イ 匿名データを統計の作成等にのみ用いること。 ロ 匿名データを学術研究の用に供することを直接の目的とすること。 ハ 匿名データを用いて行った学術研究の成果が公表されること。 ニ 匿名データを適正に管理するために必要な措置が講じられていること。
二 **高等教育の発展に資すると認められる場合**であって、次に掲げる要件のすべてに該当すると認められる場合 イ 一のイ及びニに掲げる要件に該当すること。 ロ 匿名データを学校教育法第１条に規定する大学又は高等専門学校における教育の用に供することを直接の目的とすること。 ハ 匿名データを用いて行った教育内容が公表されること。
三 **国際社会における我が国の利益の増進及び国際経済社会の健全な発展に資すると認められる場合**であって、次に掲げる要件のすべてに該当すると認められる場合（以下略）

2 匿名データの作成・提供の流れ

匿名データの作成・提供に係る事務処理については、「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。以下「ガイドライン」という。）に基づいている。

匿名データの作成・提供に係る事務は、統計局から（独）統計センターに委託している。

（1）匿名データ作成の流れ

総務省統計局	○統計局における検討 ・匿名データの作成方法案の検討 ・度数分布表を作成するなど、匿名化処理の妥当性の検証 ・統計局内で有識者会議を開催し、諮問案の確定 ○匿名データ作成に係る諮問（総務大臣→統計委員会）
⇩	
統計委員会	○統計委員会における審議 ○匿名データ作成に係る諮問の答申（統計委員会→総務大臣）
⇩	
総務省統計局	○統計委員会の意見を踏まえ匿名データの作成 ・詳細な仕様書の作成、プログラム開発、作成後のデータの検査、等

＜作成の例＞　平成 19 年就業構造基本調査に係る匿名データ

- 平成26年度から27年度にかけて統計局における検討
- 平成28年3月に諮問、4月に答申
- 平成28年度に匿名データの作成、29年4月頃提供開始予定

（2）匿名データ提供の流れ

利用者	利用の手引、利用規約の確認 利用したい匿名データについて統計センターに事前相談の上、申出
⇩	
統計センター	申出に対して利用目的等の審査、諾否の通知
⇩	
利用者	統計センターの承諾後、統計センターに依頼、手数料[1)]の納付
⇩	
統計センター	匿名データを電磁的媒体に収録して提供（暗号化措置を施す）[2)]
⇩	
利用者	匿名データの利用（集計・分析）[3)、4)]
	利用後、複製した匿名データの消去、電磁的記録媒体の返却、 作成した統計を用いて行った学術研究等の内容の公表、 統計センターに利用実績の報告

注1）　1ファイルにつき約1万円の手数料が必要

注2）　申出から約1か月で提供

注3）　匿名データの利用者は、提供された情報を適正に管理するための措置を講じる

注4）　匿名データについて、提供された目的以外の目的のために利用したり、提供したりすることは禁止されており、自己又は第三者の不正な利益を図る目的で提供したり、盗用した場合、罰せられるほか、利用条件に違反した場合には、提供禁止措置等が課される。

3 匿名データの利用について

統計局では、６種類の統計調査による匿名データを提供している。（参考１参照）

これらの匿名データは、大学やシンクタンク等における研究や、大学・大学院における演習等に利用されている。匿名データの利用により、公表している結果表以外の集計や多変量解析などの統計的分析が可能となる。

（利用例）

- 社会生活基本調査等によるワーク・ライフ・バランスに関する研究、介護に関する研究
- 全国消費実態調査等による消費行動に関する研究、所得・資産に関する研究
- 就業構造基本調査等による教育と就業に関する研究、労働市場に関する研究
- 労働力調査等による就業継続に関する研究、職業構造に関する研究
- 住宅・土地統計調査等による住宅需要に関する研究、防災対策に関する研究
- 国勢調査等による世帯形成に関する研究

＜参考1＞ 総務省統計局で匿名データを提供している統計調査 （平成 29 年 1 月 1 日現在）

統計調査名	調査年次	レコード数（最新年）	リサンプリング率	地域的情報	備考
国勢調査	平成12年、17年	約 124 万レコード（世帯員）	1%	都道府県、人口 50 万以上市区	全数調査
住宅・土地統計調査	平成5年、10年、15年	約35万レコード（住戸）	10%	都道府県	
全国消費実態調査	平成元年、6年、11年、16年	【二人以上の世帯】約4.4万レコード【単身世帯】約0.4万レコード	80%	全国２区分（３大都市圏、その他）	
労働力調査	平成元年1月～24年12月	約76万レコード（12か月分）（15歳以上世帯員）	80%	全国１区分	
就業構造基本調査	平成 4 年、9 年、14 年	約 75 万レコード（15 歳以上世帯員）	80%	全国２区分（３大都市圏、その他）	
社会生活基本調査		（10 歳以上世帯員）			
調査票Ａ	平成3年、8年、13年、18年	【生活時間編】約27万レコード【生活行動編】約14万レコード	80%	全国２区分（３大都市圏、その他）	
調査票Ｂ	平成13年、18年	約1.5万レコード	80%	全国２区分（３大都市圏、その他）	

4 匿名化処理について

匿名データにおける匿名化処理の考え方等については、ガイドラインに示されている。（参考３参照）

以下では要点のみを示すが、具体例については、参考２を参照されたい。

(1) 匿名化処理の考え方

提供機関は、調査単位及び統計単位（個人、世帯及び事業所等）等が特定又は推定されないよう、各統計調査の特性に応じて、現在、諸外国等で導入されている次の匿名化処理の技法等を組み合わせて匿名化処理を行う。

- 識別情報の削除
- 匿名データの再ソート（配列順の並べ替え）
- 識別情報のトップ（ボトム）・コーディング
- 識別情報のグルーピング（リコーディング）
- リサンプリング
- スワッピング
- 誤差の導入

(2) 匿名化の基準

調査票情報の特性は統計調査ごとに異なることから、各統計調査について一律に匿名化の基準を設定することは困難である。

このため、提供機関は、匿名化する統計調査ごとにその特性を勘案し、一橋大学における匿名標本データの試行的提供の事例及び諸外国の統計機関における同様の提供の事例等を参考に匿名化の基準となる値、例えば、最小値が２件以下とならない等を定める。

＜参考２＞　ガイドラインに示された匿名化処理と総務省統計局における適用例

ガイドライン別紙１～３に匿名化処理の考え方等が示されており、ここでは、統計局の匿名データにおける具体的な適用例を紹介する。

<table>
<tr><th colspan="2">別紙１　匿名化処理の考え方
別紙２　匿名化処理の技法</th><th>別紙３　匿名化処理の目安</th><th>統計局の匿名データにおける適用例</th></tr>
<tr><td rowspan="5">別紙１</td><td rowspan="5">(3) 特定の可能性
特定の可能性を考えると、地域範囲が狭い場合には、調査対象が絞り込まれるので、識別情報を収集することが容易になり、ミクロデータの地域情報が詳細であれば、特定の可能性が高くなる。また、調査を受けていることが知られていると、その調査単位のミクロデータに必ず存在することが分かるため、対応関係を特定される可能性が高まる。しかし、調査対象のリストは厳格に管理されており、外部の者が調査を受けている調査単位を知る可能性は低く、調査時から数年が経過すれば外部の者が知ることは不可能と言える。
しかし、特殊なデータのときに、特定の可能性は高くなる。例えば、100 歳以上の高齢者がいる世帯や世帯員が 10 人いるというような世帯の数は少ないので、母集団のある個別の世帯に対応するデータ数が少なくなり、そのどれに当たるか決定するのが比較的容易になる。また、複数の属性の特殊な組合せも特定の可能性が高くなる。これに対し、標準的な対象の場合には同じ条件のデータが多数出現することになるので、特定の可能性は比較的低いものにとどまる。</td><td>１　地理的情報について</td><td></td></tr>
<tr><td>(1) 地理的情報としては、地域内に最小でも人口 50 万人以上いなければならない。</td><td>例 1. 国勢調査では都道府県及び人口50万以上市区を提供し、全国消費実態調査等では全国２区分（３大都市圏、その他）を提供している。</td></tr>
<tr><td>(2) 直接的な地理的情報以外で、地理的情報が明らかになる項目（例えば、サンプリング情報など）についても、上記(1)の最小人口 50 万人の基準に適合させなければならない。</td><td></td></tr>
<tr><td>(3) 地域分析用として、人口 50 万人未満の地理的情報を提供するような匿名データを作成する場合には、他の識別情報などの匿名化の程度を高めなければならない。</td><td></td></tr>
<tr><td>(4) 入手可能な外部情報により、ある特定の種類の施設であることが明らかになるようなことがないようにしなければならない。</td><td></td></tr>
</table>

別紙１　匿名化処理の考え方 別紙２　匿名化処理の技法		別紙３　匿名化処理の目安	総務省統計局の匿名データにおける適用例
別紙２	(1) 匿名化処理の技法	２　個人・世帯の識別情報について	
	① 識別情報等の削除 対応関係を特定する危険性の高い識別情報である、世帯や居住地を直接的に特定できるような情報を削除する方法である。	(1) 氏名、住所など個人又は世帯を直接的に識別できる情報は削除されなければならない。	
	② 識別情報のトップ・コーディング 対応関係を特定できる可能性が高くなる特殊な属性を、まとめる方法である。例えば、100歳以上の高齢者がいる世帯や世帯員が10人いる世帯の数は少ないので、対応関係を特定しやすくなるので、特に大きい値や小さい値を「○○以上」、「○○以下」というようにまとめる。海外では、トップ・コーディングされるのが対象全体の0.5%以上としている例などがある。 ③ 識別情報のグルーピング 特定の値をグループ分けして階級区分に変更する方法である。例えば、年齢を例にすると、22歳ではなく、21～25歳とする方法である。また、市町村コードなどの地域情報の場合は、外部の者にも把握しやすい情報であること、対応関係を調べなくてはならないデータの範囲を限定できることなどから特に注意が必要となる。海外では、人口10万人未満の地域区分は提供しないなどの基準が設けられている例などがある。	(2) 間接的に個人又は世帯を識別できる情報、例えば年齢、世帯人員、居住室数などの情報については、年齢の高い個人、世帯員数が多い世帯、居住室数の多い住宅など特定される可能性が高い場合、トップコーディング、グルーピングまたは削除を施す必要がある。トップコーディングにおいては、母集団（個人又は世帯）全体の 0.5%を目安にすることが望ましい。	【削除】 例 2. 世帯人員の多い世帯（8人以上の世帯等）を削除 例 3. 三つ子以上がいる世帯を削除 例 4. 父子世帯を削除 例 5. 年齢差が一定以上大きい夫婦がいる世帯を削除 【トップコーディング】 例 6. 年齢は、5歳階級で85歳以上をトップコーディング 例 7. 居住室数は10室以上をトップコーディング 例 8. 週間就業時間は90時間以上をトップコーディング 【グルーピング】 例 9. 国籍は、「日本人」及び「外国人」の２区分（外国籍内訳を提供しない） 例 10. 世帯の種類は、「一般世帯」及び「施設等の世帯」の２区分（「施設等の世帯」の内訳を提供しない） 例 11. 世帯の家族類型は、22 区分ではなく６区分による提供 例 12. 従業上の地位は、「雇人のある業主」、「雇人のない業主」及び「家庭内職者」を統合 例 13. 産業は、「農業」、「林業」及び「漁業」を統合、「鉱業」及び「建設業」を統合、等 例 14. 職業は、「保安職業従事者」、「農林漁業作業者」及び「運輸・通信従事者」を統合

別紙１　匿名化処理の考え方 別紙２　匿名化処理の技法		別紙３　匿名化処理の目安	統計局の匿名データにおける適用例
別紙２			例 15. 住宅の所有の関係は、「公営の借家」及び「都市機構・公社の借家」を統合、「給与住宅」及び「間借り」を統合 例 16. 建物の階数は、実数ではなく階級で提供し、トップコーディングも併用 例 17. 住宅の床面積は、実数ではなく階級で提供し、トップコーディングも併用
		(3) 少数の特定の集団を対象とする場合、トップコーディングの基準を３～５％にすることを考慮すべである。	
		(4) トップコーディングするデータ項目については、その情報（平均値や中央値など）を明らかにすることが望ましい。	例 18. トップコーディング対象のレコード数、平均値、標準偏差を提供
		(5) 世帯単位のデータを提供する場合、調査単位が特定されることがないよう、必要があれば、匿名化を考慮する必要がある。	
別紙２	別の概念からの匿名化処理の技法としては、ミクロデータから正確な対応関係を知ることができないようにする方法がある。具体的には、ミクロデータを加工して正しくないものにしてしまう方法である。 ① スワッピング 任意の２つの調査単位の間で、一部の調査事項の値を入れ替える方法である。 ② 誤差の導入 ミクロデータの一部の調査事項（識別情報又は秘密の情報自体）に誤差を導入する方法である。	３　誤差（ノイズ）	
		(1) ミクロデータに誤差を加えることによって、調査データと外部情報との対応関係を特定する可能性を低めることができる。他に適当な匿名化の技法がない場合には、研究・分析上の有用性を損なわない範囲で誤差を付加することを考慮すべきである。	
		(2) 誤差を加える方法としては、①乱数による誤差の付加（random noise）、②調査単位間の調査情報の交換（swapping）、③ブランク（blank）への置換え又は補定（imputation）がある。	例 19. 一部世帯を他の地域の類似世帯と入れ替えるスワッピング（国勢調査）

別紙１　匿名化処理の考え方 別紙２　匿名化処理の技法		別紙３　匿名化処理の目安	統計局の匿名データにおける適用例
別紙２	④ リサンプリング ミクロデータをすべて提供するのではなく、そこから抽出した一部のミクロデータだけを提供する方法である。この方法によれば、提供するミクロデータが少なくなるので、対応関係を特定できる可能性を低下させることができる。 また、特定できたとの主張に対し、特定できたと考えることが適当ではないと主張する方法でもある。	4　リサンプリング ミクロデータを全て提供する場合は、その一部を提供する場合に比べて、調査単位の特定の可能が高くなる。例えば、ある人が調査を受けたことがわかっている場合には、ミクロデータの中に必ずその人のデータがあるはずとの前提で探すことができる。したがって、必要に応じて、ミクロデータの全てではなく、一部のデータだけを提供することを考慮すべきである。	例 20. すべての匿名データでリサンプリングを行っている。
別紙１	(4) 識別情報 調査対象である調査単位とミクロデータの対応関係を特定しようとするときに用いる識別情報とは、提供するミクロデータに含まれていて、かつ、統計調査以外からも知ることができる情報である 個人又は世帯を対象とした統計の場合、比較的容易に入手できる識別情報としては、外観からでも把握できるような基本的な属性が考えられ、例えば、県、市町村などの地域情報や、世帯員数、世帯員の性別、住宅の大きさなどが挙げられる。このほか、自宅で営業している世帯であればその産業・職業を知ることができるし、子供の年齢は通学している学年で分かると思われる。ただし、これらの情報だけでは、一般には対応関係を特定することはできない。また、これらの情報の収集は比較的簡単ではあるが、多数の調査単位について情報を収集しようとすれば大きな作業量を必要とする。	5　外部ファイルとのマッチングの可能性	
		(1) ミクロデータと外部の既存ファイルのデータを突き合わせることにより調査単位が識別されるような可能性があれば、それを回避するための措置をとらなければならない。	例 21. 公表統計により母集団一意又は二意であることが判明しているレコードを含む世帯を削除（国勢調査）
		(2) 調査のための標本フレームが、国勢調査の母集団情報以外の情報によって提供されている場合には、調査データと標本フレームの元の情報とを一致させることが可能となるおそれがあるので、事前に回避する措置をとらなければならない。	

別紙１　匿名化処理の考え方 別紙２　匿名化処理の技法		別紙３　匿名化処理の目安	統計局の匿名データにおける適用例
別紙２		6　その他の問題	
	⑤ ミクロデータのソート 　ミクロデータの配列順を並べ替えることでランダムにし、対応関係を探り出すことができないようにする方法である。	(1) データの一連番号、データの並び順によって、およその地域範囲が推測されるおそれがあるので、削除、付替え又は並べ替えをするべきである。	例 22. 世帯順をランダムに並び替え
		(2) サンプリングに関する情報によっては、地理的情報以外に特定の地域や集団であることが明らかになるおそれがあるので、そのような情報は削除すべきである。	
		(3) 秘密の情報のうち秘匿の必要性の高い調査項目については、その調査項目自体についてグルーピング、削除等の匿名化を施す必要がある。	例 23. ５年前の住居の所在地、常住地による従業地・通学地について、都道府県名等を提供しない。
別紙１	(4)の後半 　実際の問題としては、時間が経つとともに識別情報を正確に知ることは難しくなる。提供されるミクロデータは数年前の調査の結果であり、そのときに個々の調査対象がどのような属性を有していたか知ることは、たとえ世帯の基本的な属性であっても難しい。既存のリストのようなものの場合も、そのリストとミクロデータの時点が一致していないと対応関係の特定には多くの誤りが生じることになる。	(4) 時間の経過とともに、調査データを外部情報と照合することは困難になる。提供時期は調査時点から最低限２年間以上は離すべきである。	

＜参考３＞　匿名データの作成・提供に係るガイドライン（抄）

制 定　　平 成　２１　年　２　月　１７　日
（中略）
改 正　　平 成　２８　年　１　月　２２　日
総 務 省 政 策 統 括 官（統 計 基 準 担 当）決 定

目　次

第１　ガイドラインの目的

匿名データの作成・提供に係るガイドライン（以下「本ガイドライン」という。）は、統計法（平成19年法律第53号。以下「法」という。）第35条及び第36条の規定に基づいて行う匿名データの作成及び提供に係る事務処理の明確化及び標準化を図ることにより、行政機関又は届出独立行政法人等及び法第37条に基づき事務の全部を受託する独立行政法人等が、これらの事務を適切かつ円滑に実施できるようにすることを目的とするものである。

（中略）

第５　匿名データの作成

１　匿名データを作成する統計調査の範囲

（中略）

２　匿名データの匿名化処理の方法

(1) 匿名処理の考え方（別紙１参照）

提供機関は、調査単位及び統計単位（個人、世帯及び事業所等）等が特定又は推定されないよう、各統計調査の特性に応じて、現在、諸外国等で導入されている次の匿名化処理の技法（別紙２参照）等を組み合わせて匿名化処理を行う。

・　識別情報の削除
・　匿名データの再ソート（配列順の並べ替え）
・　識別情報のトップ（ボトム）・コーディング
・　識別情報のグルーピング（リコーディング）
・　リサンプリング
・　スワッピング
・　誤差の導入　　等

なお、基幹統計調査の場合は個別具体的に用いた匿名化の方法について取りまとめた資料を、統計委員会に対する諮問において提出するほか、必要に応じて第６の３で掲げる情報提供事項とともに公開又は、匿名データ提供の際に利用者に提供する。

(2) 匿名化の基準

調査票情報の特性は統計調査ごとに異なることから、各統計調査について一律に匿名化の基準を設定することは困難である。

このため、提供機関は、匿名化する統計調査ごとにその特性を勘案し、一橋大学における匿名標本データの試行的提供の事例及び諸外国の統計機関における同様の提供の事例等を参考に匿名化の基準となる値、例えば、最小値が２件以下とならない等を定める。

なお、個人・世帯を対象とする統計調査の匿名化について、一橋大学で行われた試行的な取組で用いた基準は別紙３「匿名化処理の目安」のとおり。

第６　匿名データの匿名化処理の実施手順

１　匿名化処理の審査

(1) 審査表の作成

提供機関及び統計委員会における匿名化処理の審査を効率的、効果的に実施するため、提供機関は作成する匿名データごとに、母集団情報や識別情報などその実施する匿名化処理の方法等を記述した審査表を作成する（別紙様式第１号参照）。

(2) 提供機関内における審査

提供機関はその組織内に匿名化処理等に関する審査体制等を設けるとともに、(1)により作成した審査表に記載された内容等を基に、実際に統計表を作成して得られた分布を確認するなどにより、匿名化処理の妥当性等に係る審査を実施する。

２　統計委員会への諮問

行政機関が基幹統計調査に係る匿名データを作成する場合、法第35条第２項に基づきあらかじめ統計委員会に諮問する必要がある。

諮問に当たり、行政機関は提供開始の時期等を勘案して事前に統計委員会事務局と審議日程等について調整を図るほか、次のとおり対応する。

(1) 初めて匿名データを作成する統計調査の場合

次に掲げる資料を準備する。

＜統計委員会の諮問資料＞

○　審査表
○　当該統計調査の基本情報
　・　調査概要
　・　調査票様式
　・　標本抽出法　　等
○　匿名データに関する資料
　・　匿名データの作成方針
　・　匿名化に当たって留意すべき事項　　等

○　その他諮問に当たって必要とされる資料及び統計委員会が法第50条に基づき要求する資料

※　将来的な作成年次の追加を予定している場合は、その旨を明示。

なお、行政機関は、統計委員会の意見を踏まえ匿名データを作成するとともに、匿名化処理が適切に行われていることを検証する。

⑵ 匿名データの作成年次を追加する場合

①　提供機関は次に掲げる資料を作成する。

○　審査表

○　提供機関における検討経緯や直近の統計委員会答申における「今後の課題」への対応に関する資料

②　提供機関は統計委員会事務局と連携し、匿名化手法に関する資料を基に次の匿名化手法を確認する。

○　追加・変更された調査事項の匿名化手法

○　識別情報の匿名化手法

○　しきい値基準によるトップコーディング・ボトムコーディングの匿名化手法

③　匿名化手法について上記①及び②により、次のⅰ）～ⅲ）の全てが確認できた場合は、前回統計委員会答申からの変更がないものと判断できるため、統計委員会への諮問を要さないものとし、それ以外の場合は統計委員会に諮問する。その判断に当たっては、統計委員会事務局と連携し、必要に応じて統計委員会の意見を聴きつつ判断する。

ⅰ）母集団情報に変更がないこと

ⅱ）調査事項別の匿名化手法に変更がないこと

ⅲ）調査事項の変更が形式的（技術的な名称変更や選択肢の統合等）であること

④　諮問時の資料は、上記①に掲げる資料のほか、「初めて匿名データを作成する統計調査の場合」を参考に必要な資料を追加する。

⑶ その他

提供機関は法 55 条に基づく総務大臣からの要請に基づき匿名データの作成に関する検討・実施状況（統計委員会答申における「今後の課題」の検討状況も含む。）について、総務省に報告を行う。

総務省は、提供機関から報告を受けた匿名データの作成に関する検討・実施状況を取りまとめ、統計委員会に報告する。

３　匿名データ提供の周知

提供機関等は、提供が可能となった匿名データについて、次の内容をホームページ等に掲載することにより情報提供を行う。（関連：第４、第７）

○　統計調査の名称及び年次

○　匿名データの名称

○　提供の条件

・　セキュリティ要件、利用環境要件

・　その他法令等により定められた要件　　等

○　提供する項目及び符号表（必要に応じてデータレイアウトフォーム）

○　匿名化処理の方法（項目ごとの匿名化処理方法、リサンプリング率等）

○　受付窓口、受付期間等

○　提供依頼申出方法

○　必要となる費用の概算

○　提供可能な方法（媒体）

○　提供予定時期

（後略）

別紙１

匿名化処理の考え方

(1) 匿名化処理とは

ミクロデータから世帯や個人の秘密の情報を知るということは、調査対象である調査単位（世帯や個人）とミクロデータの対応関係を特定し、特定されたミクロデータから調査単位の秘密に属する事項を知るということを意味する。どの調査事項が、秘密の情報に当たるかは一概には決めることができないし、時代とともに変化し、普遍的ではないと思われるので、匿名化処理とは、基本的には、調査単位とミクロデータの対応関係を特定されないようにするということである。

(2) 対応関係

提供するミクロデータには、氏名、住所などの直接的に世帯や個人が特定できる情報は付与されていないので、調査単位とミクロデータの対応関係は、性別や年齢などの属性（識別情報）が同じかどうかで判断することになる。

全国の全調査単位のミクロデータが提供されていて、かつ、全調査単位について識別情報が分かる場合、識別情報が一致する調査単位とミクロデータがそれぞれ１つしかない場合には同じ世帯や個人と判断でき、それぞれ複数ある場合はそのうちのいずれかと判断できる。実際のミクロデータの提供の場合、一部の調査単位のミクロデータが提供されていて、かつ、一部の調査単位の識別情報がわかるに過ぎず、このような状況では、対応関係を特定するのは現実的ではないと考えられる。

(3) 特定の可能性

特定の可能性を考えると、地域範囲が狭い場合には、調査対象が絞り込まれるので、識別情報を収集することが容易になり、ミクロデータの地域情報が詳細であれば、特定の可能性が高くなる。また、調査を受けていることが知られていると、その調査単位のミクロデータに必ず存在することが分かるため、対応関係を特定される可能性が高まる。しかし、調査対象のリストは厳格に管理されており、外部の者が調査を受けている調査単位を知る可能性は低く、調査時から数年が経過すれば外部の者が知ることは不可能と言える。

しかし、特殊なデータのときに、特定の可能性は高くなる。例えば、100 歳以上の高齢者がいる世帯や世帯員が 10 人いるというような世帯の数は少ないので、母集団のある個別の世帯に対応するデータ数が少なくなり、そのどれに当たるか決定するのが比較的容易になる。また、複数の属性の特殊な組合せも特定の可能性が高くなる。これに対し、標準的な対象の場合には同じ条件のデータが多数出現することになるので、特定の可能性は比較的低いものにとどまる。

(4) 識別情報

調査対象である調査単位とミクロデータの対応関係を特定しようとするときに用いる識別情報とは、提供するミクロデータに含まれていて、かつ、統計調査以外からも知ることができる情報である

個人又は世帯を対象とした統計の場合、比較的容易に入手できる識別情報としては、外観からでも把握できるような基本的な属性が考えられ、例えば、県、市町村などの地域情報や、世帯員数、世帯員の性別、住宅の大きさなどが挙げられる。このほか、自宅で営業している世帯であればその産業・職業を知ることができるし、子供の年齢は通学している学年で分かると思われる。ただし、これらの情報だけでは、一般には対応関係を特定することはできない。また、これらの情報の収集は比較的簡単ではあるが、多数の調査単位について情報を収集しようとすれば大きな作業量を必要とする。

実際の問題としては、時間が経つとともに識別情報を正確に知ることは難しくなる。提供されるミクロデータは数年前の調査の結果であり、そのときに個々の調査対象がどのような属性を有していたか知ることは、たとえ世帯の基本的な属性であっても難しい。既存のリストのようなものの場合も、そのリストとミクロデータの時点が一致していないと対応関係の特定には多くの誤りが生じることになる。

(5) 特定の試み

匿名化処理の方法を決めるときには、現実にどのような危険があるかについても考えておく必要がある。最近、個人情報の流出がよく問題となるが、そのような例では、住所（メールのアドレス等も含む。）、氏名などが流出しており、それは、商業目的などにそのまま利用できる。しかし、統計情報の場合、住所、氏名が流出することはあり得ない。また、前述のとおり、特殊な対象の場合には特定の可能性が比較的高くなるが、多くの標準的な対象の場合には特定の可能性は比較的低いものにとどまる。一部の対象についてだけ特定できたとしても、商業目的での利用価値は少ないであろう。したがって、対象を特定しようとするような試みが、最近問題になっているような商業目的で行われる可能性は低いものと考えられる。そもそも、数年前の統計情報では利用する価値もないであろう。

しかし、もし対象を特定するような試みが実際に行われたら、それはミクロデータ提供の危険性、ひいては統計調査の危険性を指摘するものとして利用されてしまうであろう。ところが、絶対的な匿名性を担保しようとすると、ドイツでの経験のように提供できる情報が極めて限られてしまう。したがって、この問題は匿名化処理だけで対策を考えるべきものではなく、そのような試みを行うこと自体を制限しておくことが必要となる。このため、データを提供するときには、利用目的を限定し、データの管理を適正に行わせることを義務付けておかなくてはならない。

注：ドイツは、1980 年の連邦統計法で「絶対的な匿名化」条項によるミクロデータの提供を行ってきたが、多くの情報が失われることになり、科学研究の要求に応じられず、ほとんど利用されなかった。そのため、1987 年の連邦統計法ではミクロデータが莫大な時間や経費をかけない限り識別できないという「事実上の匿名性」の概念に法規定を改正している。

別紙２

匿名化処理の技法

(1) 匿名化処理の技法

対応関係を特定しにくくする匿名化処理の方法としては、下記のような方法がある。

① 識別情報等の削除

対応関係を特定する危険性の高い識別情報である、世帯や居住地を直接的に特定できるような情報を削除する方法である。

② 識別情報のトップ・コーディング

対応関係を特定できる可能性が高くなる特殊な属性を、まとめる方法である。例えば、100歳以上の高齢者がいる世帯や世帯員が 10 人いる世帯の数は少ないので、対応関係を特定しやすくなるので、特に大きい値や小さい値を「○○以上」、「○○以下」というようにまとめる。海外では、トップ・コーディングされるのが対象全体の 0.5%以上としている例などがある。

③ 識別情報のグルーピング

特定の値をグループ分けして階級区分に変更する方法である。例えば、年齢を例にすると、22 歳ではなく、21～25 歳とする方法である。また、市町村コードなどの地域情報の場合は、外部の者にも把握しやすい情報であること、対応関係を調べなくてはならないデータの範囲を限定できることなどから特に注意が必要となる。海外では、人口 10 万人未満の地域区分は提供しないなどの基準が設けられている例などがある。

④ リサンプリング

ミクロデータをすべて提供するのではなく、そこから抽出した一部のミクロデータだけを提供する方法である。この方法によれば、提供するミクロデータが少なくなるので、対応関係を特定できる可能性を低下させることができる。

また、特定できたとの主張に対し、特定できたと考えることが適当ではないと主張する方法でもある。

⑤ ミクロデータのソート

ミクロデータの配列順を並べ替えることでランダムにし、対応関係を探り出すことができないようにする方法である。

別の概念からの匿名化処理の技法としては、ミクロデータから正確な対応関係を知ることができないようにする方法がある。具体的には、ミクロデータを加工して正しくないものにしてしまう方法である。

① スワッピング

任意の２つの調査単位の間で、一部の調査事項の値を入れ替える方法である。

② 誤差の導入

ミクロデータの一部の調査事項（識別情報又は秘密の情報自体）に誤差を導入する方法である。

(2) 匿名化処理の方法の決定

上記のような問題があるものの、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。我が国においても同様の手続きを踏むべきであり、試行的提供では、統計局の「匿名標本データ作成・利用研究会」の承認を得ている。

別紙３

匿名化処理の目安

1　地理的情報について

(1) 地理的情報としては、地域内に最小でも人口 50 万人以上いなければならない。

(2) 直接的な地理的情報以外で、地理的情報が明らかになる項目（例えば、サンプリング情報など）についても、上記(1)の最小人口 50 万人の基準に適合させなければならない。

(3) 地域分析用として、人口 50 万人未満の地理的情報を提供するような匿名データを作成する場合には、他の識別情報などの匿名化の程度を高めなければならない。

(4) 入手可能な外部情報により、ある特定の種類の施設であることが明らかになるようなことがないようにしなければならない。

2　個人・世帯の識別情報について

(1) 氏名、住所など個人又は世帯を直接的に識別できる情報は削除されなければならない。

(2) 間接的に個人又は世帯を識別できる情報、例えば年齢、世帯人員、居住室数などの情報については、年齢の高い個人、世帯員数が多い世帯、居住室数の多い住宅など特定される可能性が高い場合、トップコーディング、グルーピングまたは削除を施す必要がある。トップコーディングにおいては、母集団（個人又は世帯）全体の 0.5%を目安にすることが望ましい。

(3) 少数の特定の集団を対象とする場合、トップコーディングの基準を３～５%にすることを考慮すべである。

(4) トップコーディングするデータ項目については、その情報（平均値や中央値など）を明らかにすることが望ましい。

(5) 世帯単位のデータを提供する場合、調査単位が特定されることがないよう、必要があれば、匿名化を考慮する必要がある。

3　誤差（ノイズ）

(1) ミクロデータに誤差を加えることによって、調査データと外部情報との対応関係を特定する可能性を低めることができる。他に適当な匿名化の技法がない場合には、研究・分析上の有用性を損なわない範囲で誤差を付加することを考慮すべきである。

(2) 誤差を加える方法としては、①乱数による誤差の付加（random noise）、②調査単位間の調査情報の交換（swapping）、③ブランク（blank）への置換え又は補定（imputation）がある。

4　リサンプリング

ミクロデータを全て提供する場合は、その一部を提供する場合に比べて、調査単位の特定の可能が高くなる。例えば、ある人が調査を受けたことがわかっている場合には、ミクロデータの中に必ずその人のデータがあるはずとの前提で探すことができる。したがって、必要に応じて、ミクロデータの全てではなく、一部のデータだけを提供することを考慮すべきである。

5　外部ファイルとのマッチングの可能性

(1) ミクロデータと外部の既存ファイルのデータを突き合わせることにより調査単位が識別されるような可能性があれば、それを回避するための措置をとらなければならない。

(2) 調査のための標本フレームが、国勢調査の母集団情報以外の情報によって提供されている場合には、調査データと標本フレームの元の情報とを一致させることが可能となるおそれがあるので、事前に回避する措置をとらなければならない。

6　その他の問題

(1) データの一連番号、データの並び順によって、およその地域範囲が推測されるおそれがあるので、削除、付替え又は並べ替えをするべきである。

(2) サンプリングに関する情報によっては、地理的情報以外に特定の地域や集団であることが明らかになるおそれがあるので、そのような情報は削除すべきである。

(3) 秘密の情報のうち秘匿の必要性の高い調査項目については、その調査項目自体についてグルーピング、削除等の匿名化を施す必要がある。
(4) 時間の経過とともに、調査データを外部情報と照合することは困難になる。提供時期は調査時点から最低限２年間以上は離すべきである。