参考２

平成２０年５月１３日
総務省政策統括官（統計基準担当）付
統計企画管理官室

「統計データの二次利用促進に関する研究会」の検討状況について

１　研究会の開催

総務省政策統括官（統計基準担当）では、統計データの二次利用（委託による統計の作成等（以下「オーダーメード集計」という。）及び匿名データの作成･提供）に関するガイドラインの策定に当たり、有識者・利用者から助言等を得るため「統計データの二次利用促進に関する研究会」（以下「研究会」という。）を平成１９年１０月から５回開催しており、これまで運用面の課題（第４ＷＧ第４回会合で報告）及び技術的な課題を検討。

技術的な課題に関する検討状況は以下のとおり。

２　技術的な課題に対する検討状況（議論の要約）

（１）オーダーメード集計における秘匿処理について

① オーダーメード集計においても、客対数が少なく個別の情報の識別につながり易い項目が存在する場合には、当該集計結果表に秘匿処理を行うことが必要。

② 集計結果表における秘匿処理の方法は、調査対象が事業所･企業等か世帯か、全数調査か標本調査かによって異なり、表彰する項目によっても異なる。また、オーダーメード集計の結果表は多岐に渡るため、一律の基準を作成することは困難。ガイドラインでは、各府省が実施している秘匿処理の事例を参考に、目安となるものを整理する方向。

※　各府省が実施している集計結果表の秘匿処理の事例（別紙１）

（２）匿名データにおける秘匿処理について

①　匿名データの作成は、原則として、事業所統計を含めすべての統計調査を対象とする方向。

②　匿名化できない調査は、オーダーメード集計で対応する等何らかの形態で利用できるように求める方向を検討。

③　調査によってデータの安全性、有用性は異なるので、匿名化の方法は、調査ごとに決める方向。

④　各府省が行う秘匿処理の参考とするため、ガイドラインには秘匿処理の技法や匿名化の目安となるものを盛り込む方向。
※　秘匿処理の考え方、技法（別紙２）、匿名化の基準（目安）（別紙３）

（３）秘匿処理の審査のためのチェックリストについて

①　秘匿処理に係るチェックリストは各府省が統計調査ごとに作成する方向。

②　チェックリストの様式は各府省共通とし、内容は調査ごとに異なるものとする方向。
※　匿名データのチェックリストのイメージ（別紙４及び５）

（４）統計委員会における秘匿処理審査への諮問について

統計法第 35 条第 2 項では、基幹統計調査に係る匿名データを作成する場合、行政機関の長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と規定。この場合の審査書類について検討。

○　統計委員会に匿名データについて諮問する際には、匿名データのチェックリストだけではなく、当該統計調査の基本的情報（調査概要、調査票の様式、標本抽出法、提供データの形式等）、匿名データに関する情報（匿名データの作成方針、匿名データのレイアウトフォーム・符号表、匿名化に当たっての留意事項、実施した匿名化の概要）を添付する方向で検討。

（５）秘匿処理の実施機関について

○　秘匿処理にスキルが必要であれば、効率的な実施のために、ある程度処理機関を特定化すべき。

３　今後の検討スケジュール

平成２０年６月頃を目途に研究会報告を取りまとめる予定。

別紙１

「統計データの二次利用促進に関する研究会」第３回会合資料から抜粋

現在、各省が実施している集計結果表の秘匿処理の事例

○　各府省において実施している集計結果表の秘匿処理については、概ね下表のとおりである。

調査対象	標本	全数
事業所等	①客体数が少ない場合、結果を非表示（“ｘ”等に置換え）（賃金引上げ等の実態に関する調査等） ②①の他、合計値からの引き算により秘匿対象が判明する場合は、二次秘匿処理（サービス業基本調査） ③客体数が３未満の場合、客体数は表章するが経営に係る項目は非表示（農林水産関係の統計全般） ④事業所数が一定数以下でかつ従業者数が一定数以下の場合非表示（屋外労働者職種別賃金調査等） ⑤労働者数を 10 人単位で表章（賃金構造基本調査）	○客体数が少ない場合、結果を非表示 （“ｘ”等に置換え） 合計値からの引き算により秘匿対象が判明する場合は、二次秘匿処理 （工業統計調査、商業統計調査、学校教員統計調査等）
世帯	○表章単位の丸め（1000 世帯、万人等） （労働力調査、国民生活基礎調査等）	○表章区分の統合（小地域集計（国勢調査））

秘匿処理について

(1) 秘匿処理とは

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、秘匿処理とは、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

秘匿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、この問題は秘匿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。

(6) 秘匿処理の技法

対応関係を特定しにくくする秘匿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。

別の概念からの秘匿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。

(7) 秘匿処理の方法の決定

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秘匿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。秘匿処理は、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。

匿名化の基準（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。

別紙４

「統計データの二次利用促進に関する研究会」第５回会合資料から抜粋

匿名データのチェックリスト（案）
―全国消費実態調査を例として―

匿名データを作成する統計データの名称および年次

全国消費実態調査（平成元年、６年、11 年、16 年）

１　地理的情報

(1) 提供するファイルにはどのレベルの地理的情報が含まれていますか。匿名化のために地理的情報を加工していますか。

全国を６地域に区分した地域ブロック。 全国 47 都道府県を６ブロックに集約しています。 ブロックの構成、人口、世帯数は別添１（省略）を参照。

(2) 直接的な地理的情報以外に地理的情報が明らかになるような情報がありますか。

標本データを母集団に復元するための乗率は、都市階級別にそれぞれ固有の値になっているために、地理的情報と組み合わせると、市区町村レベルまで判明するおそれがあります。そのため、匿名化措置として、乗率を階級別に区分し、階級別のその平均値を乗率としています。 乗率の階級、平均値等は別添２（省略）を参照。

(3) 地域分析用に詳細な地理的情報を提供していますか。

特に地域分析用のファイルは作成していません。

(4) ある特定の種類の施設であることが明らかになることはありますか。

特にそのようなことはありません。

２　世帯の識別情報

(1) 世帯の識別情報として考えられるデータ項目を挙げてください。

世帯符号、世帯人員

(2) それぞれの識別情報について、どのような匿名化措置をとっていますか。

世帯符号について、オリジナルの符号は削除し、新たに世帯単位に一連番号を付与しています。
世帯人員が９人以上の世帯は削除しています。

(3) 匿名化措置を施した場合には、その情報を明示してください。

世帯人員９人以上の世帯は母集団全体の約 0.07%を占めています。
世帯人員分布は別添３（省略）を参照。

(4) 世帯単位のデータを提供することに対応して特別な匿名化措置を施していますか。

特別な匿名化措置は施していません。

３　個人の識別情報

(1) 個人の識別情報として考えられるデータ項目を挙げてください。

性別、年齢

(2) それぞれの識別情報について、どのような匿名化措置をとっていますか。

年齢が 80 歳以上のデータについては、すべて 80 歳としています。

(3) 匿名化措置を施した場合には、その情報を明示してください。

年齢が 80 歳以上の人は、母集団全体の約５%を占めています。
年齢分布は別添４（省略）を参照。

４　誤差（ノイズ）

匿名化措置として、誤差を付加する方法を採っていますか。誤差を付加する方法を採っている場合には、その方法を記載してください。

誤差を加える方法は採用していません。

5　リサンプリング

匿名化措置として、リサンプリングをしていますか。リサンプリングをしている場合には、その抽出方法と抽出率を記載してください。

リサンプリングを行っています。 抽出方法は乗率階級別に標本数を比例配分し、乗率階級内は乗率を考慮した確率比例抽出法を採用しています。抽出率は 80%です。

6　外部ファイル

(1) ミクロデータを特定できる可能性のある外部ファイルは存在しますか。

そのような外部ファイルは存在しません。

(2) 母集団情報として利用している情報は何ですか。

母集団情報として利用しているのは国勢調査の調査区情報です。 調査区内の世帯名簿は調査の一環として作成し、その世帯名簿は調査関係者以外見ることはできません。

7　その他

(1) データの一連番号、データの並び順について、何らかの匿名化措置を施していますか。

オリジナルのデータ一連番号は削除しています。 データの並び順は、世帯単位に、乱数によりランダムな並びにしています。 ランダムな並びにしてから、データの一連番号を付与しています。

(2) サンプリング情報によって、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可能性はありますか。

そのような情報はありません。

(3) 秘密の情報のうち、特に秘匿する必要性の高い調査項目がありますか。ある場合には、どのような匿名化措置をとっていますか。

秘密の情報のうち、年間収入について秘匿の必要性を検討したが、年間収入から調査単位が特定される可能性は低いとして、匿名化措置は特に施していません。 また、年間収入は回帰分析などで説明変数としてよく利用され、ジニ係数の計算のため

にも実数でないと困ることから、利用の面も考慮してそのまま提供しています。

(4) 提供時期と調査時点とはどの程度の期間が開いていますか。

調査による結果がすべて公表されてから、匿名データを提供しています。したがって、最短の期間でも調査時点から２年以上は開いています。

(5) そのほか、データを匿名化するに当たり、措置していることがありますか。

特にありません。

別紙５

「統計データの二次利用促進に関する研究会」第５回会合資料から抜粋

匿名データのチェックリスト（案）
―就業構造基本調査を例として―

匿名データを作成する統計データの名称および年次

就業構造基本調査（平成４年、９年、14 年）

１　地理的情報

(1) 提供するファイルにはどのレベルの地理的情報が含まれていますか。匿名化のために地理的情報を加工していますか。

全国を６地域に区分した地域ブロック。 全国 47 都道府県を６ブロックに集約しています。 ブロックの構成、人口、世帯数は別添１（省略）を参照。

(2) 直接的な地理的情報以外に地理的情報が明らかになるような情報がありますか。

特にありません。

(3) 地域分析用に詳細な地理的情報を提供していますか。

特に地域分析用のファイルは作成していません。

(4) ある特定の種類の施設であることが明らかになることはありますか。

特にそのようなことはありません。

２　世帯の識別情報

(1) 世帯の識別情報として考えられるデータ項目を挙げてください。

世帯符号、世帯人員

(2) それぞれの識別情報について、どのような匿名化措置をとっていますか。

世帯符号について、オリジナルの符号は削除し、新たに世帯単位に一連番号を付与しています。 世帯人員が９人以上の世帯は削除しています。

(3) 匿名化措置を施した場合には、その情報を明示してください。

世帯人員９人以上の世帯は母集団全体の約 0.07%を占めています。
世帯人員の分布は別添２（省略）を参照。

(4) 世帯単位のデータを提供することに対応して特別な匿名化措置を施していますか。

特別な匿名化措置は施していません。

３　個人の識別情報

(1) 個人の識別情報として考えられるデータ項目を挙げてください。

性別、年齢

(2) それぞれの識別情報について、どのような匿名化措置をとっていますか。

年齢が 80 歳以上のデータについては、すべて 80 歳としています。

(3) 匿名化措置を施した場合には、その情報を明示してください。

年齢が 80 歳以上の人は、母集団全体の約５%を占めています。
年齢分布は別添３（省略）を参照。

４　誤差（ノイズ）

匿名化措置として、誤差を付加する方法を採っていますか。誤差を付加する方法を採っている場合には、その方法を記載してください。

誤差を加える方法は採用していません。

５　リサンプリング

匿名化措置として、リサンプリングをしていますか。リサンプリングをしている場合には、その抽出方法と抽出率を記載してください。

リサンプリングを行っている。
抽出方法は単純任意抽出法を採用し、抽出率は 80%です。

6　外部ファイル

(1) ミクロデータを特定できる可能性のある外部ファイルは存在しますか。

そのような外部ファイルは存在しません。

(2) 母集団情報として利用している情報は何ですか。

母集団情報として利用しているのは国勢調査の調査区情報です。
調査区内の世帯名簿は調査の一環として作成し、その世帯名簿は調査関係者以外見ることはできません。

7　その他

(1) データの一連番号、データの並び順について、何らかの匿名化措置を施していますか。

オリジナルのデータ一連番号は削除しています。
データの並び順は、世帯単位に、乱数によりランダムな並びにしています。
ランダムな並びにしてから、データの一連番号を付与しています。

(2) サンプリング情報によって、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可能性はありますか。

そのような情報はありません。

(3) 秘密の情報のうち、特に秘匿する必要性の高い調査項目がありますか。ある場合には、どのような匿名化措置をとっていますか。

特に秘匿する必要性の高い調査項目はありません。

(4) 提供時期と調査時点とはどの程度の期間が開いていますか。

調査による結果がすべて公表されてから、匿名データを提供しています。したがって、最短の期間でも調査時点から２年以上は開いています。

(5) そのほか、データを匿名化するに当たり、措置していることがありますか。

特にありません。