

平成 20 年 6 月 16 日

「統計データの二次利用促進に関する研究会」報告書

(中間取りまとめ案)

(たたき台)

はじめに

当研究会は、統計法（平成 19 年法律第 53 号）において、委託による統計の作成等（第 34 条）及び匿名データの作成・提供（第 35 条、第 36 条）を行うことができるとの規定が設けられたことから、これら統計データの二次的な利用（以下「二次利用」という。）を促進する観点から関連の諸課題について検討するため、総務省政策統括官（統計基準担当）の求めにより開催されたものであり、平成 19 年 10 月 22 日から平成 20 年 月 日まで計 回にわたり、二次利用の運用手続きに関する課題、集計結果表や個別データの秘匿処理に関する技術的な課題を中心に検討を重ね、本報告書を取りまとめるに至った。

基本的な考え方

平成 21 年春の統計法の全部施行に伴い、各府省における二次利用に係る業務が行政活動の一環として適切に実施されることが必要となる。

このため、各府省では、二次利用に係る利用申出手続きや利用目的の審査等の運用に係る事務作業を府省間で整合性を確保しながら適切に実施することが必要である。また、二次利用に係る処理作業を実施するためには、国民の共有財産である行政資源を費やす必要があるとともに、特に匿名データについては、秘匿処理が施されているものの、個別客体の調査結果情報が提供されることから、調査対象者の統計調査に対する信頼を損なわないようにすることが重要である。

このような観点を踏まえつつ、当研究会では、各府省が整合性を確保しながら適切に二次利用の関係業務を実施していくための運用面、技術面における諸課題について検討し、以下のように整理した。

二次利用の運用手続面について

1 利用目的の範囲

統計法第 34 条及び第 36 条では、「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」にのみ、「委託による統計の作成等」（以下「オーダーメイド集計」という。）や「匿名データ」の提供ができると規定している。

ア.「公益性」を求める理由

統計法が、オーダーメイド集計や匿名データの提供に当たって利用目的を制限しているのは、次のような理由によるものである。

(1) オーダーメイド集計

オーダーメイド集計は、依頼者自身が調査票情報を利用することがないことから、提供する統計等に秘匿処理が確実に行われていれば安全性は高いが、その実施に当たっては、相当程度の処理作業を伴い、国民の共有財産である行政資源を費やすこととなり、また、目的を問わずに依頼に応ずることとした場合には、調査対象者の統計調査に対する信頼を損ね、ひいては調査への協力が得られなくなり、また、統計の真実性の確保に支障をきたすおそれがあることから、一定の公益性を求めることとしたものである。

(2) 匿名データの提供

匿名データの提供は、秘匿措置が施されているとしても、集計されていない情報が提供されるものであり、目的を問わずに提供依頼に応ずるとした場合には、調査対象者の統計調査に対する信頼を損ね、ひいては調査への協力が得られなくなり、また、統計の真実性の確保に支障をきたすおそれがあることから、一定の公益性を求めることとしたものである。

イ.「学術研究の発展に資すると認める場合」

(1) 大学・研究機関における場合

典型的な場合として、大学や研究機関において教授、准教授、講師、助教、いわゆるポストドクター（博士研究員）、大学院生等が学術を目的として研究活動を行う場合が想定される。

しかしながら、一定程度の公益性が確保されるためには、このような研究によって社会に対して何らかの貢献が認められることが必要であり、学術論文等の形で研究の成果が公表され、社会に還元されることを要件とするように総務省令で規定すべきである。

(2) 大学・研究機関以外における場合

営利企業に属する者が企業活動の一環として研究を行う場合においても、それが学術的な研究を目的とするもので、研究論文等の形で当該研究の成果が社会に還元される場合であれば、本要件に該当するものと考えられる。

また、学術研究目的に一部営利目的が含まれている場合であっても、当該研究の成果が社会に還元される場合であれば公益性が認められると考える。

【P：学術研究に一部営利目的が含まれる場合にも二次利用が可能とした場合、統計調査の活動にどのような影響があるか。】

しかしながら、当該研究の成果が企業等の組織内部における業務上の資料として使用される場合や特定の顧客に対するレポート作成の基礎資料とするような場合には公益性があるとは認められないと考える。

ウ．「その他総務省令で定める場合」

大学等の高等教育機関において講義・演習等の教育目的に利用するような場合を規定することが考えられる。この場合、大学の授業において学生がオーダーメイド集計された集計結果表や提供を受けた匿名データを直接利用することが想定され、このような場合には、指導教授等の指導監督の下で、かつ使用場所は教室内や研究室に限るなどの要件を付すことが必要である。

【P：教育目的での利用は高等教育機関に限定すべきか。また、匿名データの利用を学生に認めるべきか。】

2 利用の申出

ア．利用申出書

統計法第 34 条及び第 36 条では、オーダーメイド集計や匿名データの提供を求める場合の申出事項は総務省令で定めることが規定されている。利用者の利便性を考慮すると利用申出書の様式は府省間で統一されることが必要であり、申出書に盛り込むことが必要と考えられる利用事項は以下とおりである。

(1) オーダーメイド集計の場合（統計法第 34 条の総務省令で定める手続き）

- 申出者（所属、氏名、連絡先）
- 使用する統計調査の名称及び年次
- 利用目的
（学術研究目的の場合：研究計画名、具体的な研究内容、研究成果の取扱い（○○学会で発表、○○誌に掲載等））
（教育目的の場合：講義、演習の内容、実施計画、使用場所、使用方法）【P：教育目的による利用の場合の利用目的として何を記載すべきか。】
- 集計等の作業委託内容（集計結果表様式を添付）
- 共同研究者の氏名及び所属
- 作成した集計結果表等の提供を受ける方法

(2) 匿名データの提供の場合（統計法第 36 条の総務省令で定める手続き）

- 申出者（所属、氏名、連絡先）
- 使用する統計調査の名称及び年次
- 利用目的
（学術研究目的の場合：研究計画名、具体的な研究内容、研究成果の取扱い（○○学会で発表、○○誌に掲載等））
（教育目的の場合：講義、演習の内容、実施計画、使用方法）【P：教育目的による利用の場合の利用目的として何を記載すべきか。】
- 使用場所
- データの管理方法（データの保存場所・保存方法、パソコンの使用環境（インターネット非接続等）

- データを取扱う者の範囲
(学術研究目的の場合：共同研究者、外部委託する場合の委託先等)
(教育目的の場合：指導教授、演習履修者等)【P：教育目的による利用の場合のデータを取扱う者の範囲はどのようなものとすべきか。】
- 匿名データの提供を受ける方法

イ．誓約書の添付【P：誓約書を提出させるべきか。また、その記載内容はどのようなものとすべきか。】

誓約書に盛り込むことが必要と考えられる事項は以下とおりである。

- 利用申出書に記載した以外の利用を行わない旨
- 匿名データを適正に管理する旨
- 秘密の保護に遺漏が無いよう厳重に注意する旨
- 不適切利用を行った場合のペナルティに合意する旨
- 利用終了後は提供された匿名データを直ちに返却又は破壊する旨
- 各府省からオーダーメイド集計による集計結果表や匿名データの提供を受けたことを明記する旨

ウ．利用申出者

学術研究を目的としたオーダーメイド集計や匿名データ等の利用申出者は、自分自身が責任を持って研究を行うことができる者である。したがって、大学・研究機関における場合の利用申出者を例示すれば、以下のとおりとなる。

教授等の指導の下で、大学院生や学部学生が研究に携わる場合の匿名データの利用申出者は、指導する教授等である。(注)

(注) この場合の大学院生や学部学生は、匿名データの提供を受けた者(教授等)から「業務の委託を受けた者その他の当該委託に係る業務を受託した者」に位置づけられ、統計法第42条第2項(適正管理)及び同法第43条第2項(利用制限)に規定する義務が生じるとともに、同法第61条第3項の罰則規定が適用される。従って、これらの者が匿名データを不正な利益を図る目的で提供や盗用した場合にはその者が直接罰則の対象となり、その際の教授等の責任として、運用上、匿名データの利用取り消し等の措置の対象となることもあり得る。

このため、申出時においては、匿名データの使用人として、大学院生や学部学生を特定することが必要となる。

共同研究として、教授等と大学院生が研究を行う場合の申出者は、その教授等と大学院生である。

教授等の指導監督下でなく、個人として大学院生や学部学生が研究を行う場合の申出者は、その大学院生や学部学生である。

なお、匿名データの提供の場合、秘匿措置が施されているものの、集計さ

れていない情報が提供されるものであり、統計法第42条1項第2号で匿名データの提供を受けた者の適正な管理義務を規定している趣旨から見て、匿名データを提供する場合には、利用申出者の本人確認を十分に行う必要がある。

エ．利用申出の受付窓口

申出に際して、オーダーメイド集計の場合は、オーダーしたい結果表の作成が可能であるか否か等について各府省との個別事前打ち合わせが、匿名データの利用の場合は、匿名データの仕様等について各府省への照会等が必要であることから、利用申出書の受付窓口はいずれの場合も、当該統計調査等の内容を熟知している所管各府省に対して提出することが効率的である。

オ．利用申出の時期

利用申出の受付から提供までの期間を明確にし、年間の受付時期・期間を年数回に限定する等の方法は、行政機関における効率的かつ円滑な運用が可能となり、ひいては利用者としても円滑な提供が期待されることから、検討すべきである。

3 利用目的の審査

ア．利用目的の審査内容

利用目的を審査する場合、研究成果の公表方法が明確になっているとともに、「公益性」の有無がポイントであることから、申出時においては「公益性を示す書類」の添付を求めることが必要である。

(1) 「公益性を示す書類」

公益性を示す書類としては、たとえば次のようなものが想定される。

- 研究計画書
- 所属がある場合は所属を証明する書類
(指導教授や大学、学会からの推薦状、大学に籍があることを証明する書類等)
- 公表の方法に係る書類
(発表予定の学会・大会の名称や活動内容、提出する研究計画書への掲載予定の学術誌、機関紙、専門誌、業界誌等の記述等)
- 教育目的で利用することを証明する書類【P：教育目的での利用申出の場合の公益性を示す書類をどのように考えるべきか。】
(高等教育機関に所属することを証明する書類)
- その他
(研究費を確保するために作成した既存の書類、利用者の著書・論文の一覧等)

(2) 審査内容

学術研究の発展に資すると認める場合の利用目的を審査するに当たっては、依頼者の属する機関が外形的にみて研究を行っている機関であるかどうか、研究の成果が社会に還元(学術論文等の形で研究の成果が公表)されることになるか、提供したデータの管理体制や管理方法に信頼がおけるか、研究内容と利用される統計調査の関係が適切であるか等を審査する必要がある。

教育目的での利用を審査するに当たっては、依頼者の属する機関が外形的にみて高等教育機関であるかどうか、講義等の実施状況がどのように報告されるか、講義で使用した資料や演習で作成した集計結果表を調査実施者に提示するかどうか、提供した集計結果表や匿名データの管理体制や管理方法に信頼がおけるか、講義等の内容と利用される集計結果表や匿名データの関係が適切であるか等を審査する必要がある。【P：教育目的での利用申出の場合の審査内容をどのように考えるべきか。】

(3) 審査の実施主体

利用目的の審査を行政機関が行うことは適当ではなく、第三者専門機関が審査すべきではないかとの提言もなされているが、次の理由から、オーダーメイド集計や秘匿データの提供に関する利用目的の審査は、当該統計調査の実施者が行うことが適当であると考ええる。

- ）利用目的の審査は、上記（2）に示すとおりの形式審査であり、研究内容の有用性や重要度を評価するものではないこと
- ）統計調査の調査対象者の信頼を損なわないため、調査実施者として適正に審査した上でオーダーメイド集計を受託し、匿名データを提供していることを自ら確認しておく必要があること
- ）統計調査の内容や匿名データの内容について最も詳細に把握しているのは調査実施者自身であること

なお、提供状況については、統計法第 55 条に基づいて総務省がとりまとめて公表するとともに統計委員会へも報告することが想定されるが、利用目的の審査の公平性や透明性を確保する観点から、申出に応じなかった場合については、その件数を報告すべきであり、研究者の個人情報保護にも配慮しつつ、申出において示された研究概要や応じなかった理由等を報告することについても検討が必要である。

4 審査結果に対して不服があった場合の対応方法

統計法第 34 条や第 36 条の規定では、それぞれ「統計等の作成を行うことができる」、「匿名データを提供することができる」とされていることから、行政手続法第 2 条第 2 項や行政不服審査法第 2 条第 1 項に規定する「処分」

には該当しないと解されている。このため、審査の結果、申出に応じられないとした場合に、行政手続法や行政不服審査法に基づいた不服申立ては困難ではないかと考えられる。

しかしながら、二次利用という新たな制度を円滑かつ適切に運用する観点から、例えば、二次利用に関する政府統計全体の案内、相談、苦情等の窓口機能を設け、そこで苦情等の受付・斡旋等を行えるような仕組みを検討すべきである。

5 集計結果表等の著作権に関する留意事項

オーダーメイド集計の場合、申出者の創意工夫や集計の事務処理を受託した民間事業者の創意工夫によって付加価値のついた集計結果表が作成されるなど、一連の事務処理の過程で、申出者や委託を受けた民間事業者に著作権が発生する可能性は否定できない。

このため、集計・分析等の業務を民間事業者に委託する場合には、成果物に著作権が発生するような処理を行わないことを契約上明確にしておくことが必要である。また、仮に、特別な分析やグラフの作成等民間事業者の創意工夫の余地があるものを委託せざるを得ないような場合には、その成果物の著作権は調査実施者に帰属する旨を契約書において明確にしておくことが必要である。

さらに、申出者の著作権についても、当該申出による集計結果表の提供後に類似の申出が行われた場合に、先の申出者の著作権があるため応じられないというようなことが生じると運用上支障があることから、これを回避するための手段として、申出者は著作権を主張しないということを申出受理の条件に入れておくことも必要である。

6 その他運用面での留意事項

二次利用の円滑な運用を図る観点から、次のような仕組み等を検討することが必要である。

(1) 提供窓口について

申出書の提出から、オーダーメイド集計による集計結果表や匿名データの提供を受けるまでの期間を最短にする観点からすれば、提供窓口は、匿名データを作成する府省毎に置くことが効果的である。

なお、将来的には、諸外国に見られるように、統計調査票情報の保存と提供を一体化したデータ・アーカイブを構築することにより、利用者の利便性の向上を図るとともに、提供者側の事務処理の効率化を図る必要があると考える。

(2) 提供媒体・提供方法について

- オーダーメイド集計については、利用者が提供を受けた集計結果表を加

工したり、論文等に掲載する場合に、電子的な形態であることが望ましいことから、CD-R等の電子的記録媒体の郵送、直接受渡またはオンラインにより提供することとすべきである。

- 匿名データについては、統計法第42条第1項第2号で適正な管理が規定されている趣旨から見て、集計結果表に比べてより安全性に配慮することが必要であり、当面、電子的記録媒体の書留等による郵送または直接受渡しにより提供すべきである。

(3) 提供されたデータの利用終了後の措置

- 匿名データについては、秘匿処理が施されているとはいえ調査票情報であることから、統計法第42条第1項第2号で適正な管理が規定されているのであり、調査対象者の信頼を損なわないためには、利用期間終了後は提供したCD-R等が確実に返却又は破壊されたことを提供側で確認できる仕組みを構築すべきである。また、利用期限内に申出した研究等が終了しない場合は、適切な延長手続きを行うことで延長が可能なるようにすべきである。
- オーダーメイド集計により提供された集計結果表についても、その利用目的を公益性のある場合に限定していることから、集計結果表の利用は、利用申出の際に提示された目的の範囲内のみ限定すべきである。

【P：提供された集計結果表を返却又は破壊させる必要があるか。】

(4) 提供された集計結果表や匿名データの利用範囲と成果公表方法について

オーダーメイド集計で提供された集計結果表や匿名データを用いて作成する統計表等の利用は、利用申出書に記載した利用目的の範囲内に限定すべきであり、利用申出書に記載した公表方法でのみ公表するようにすべきである。

(5) 匿名データを作成する統計調査について

どのような統計調査について匿名データを作成するかは、各府省が利用者のニーズ等を勘案して判断することとなるが、匿名化が難しいとされている事業所系の統計調査についても、匿名化の対象とし、匿名化の努力が行われるべきである。なお、匿名化できない調査は、オーダーメイド集計で対応することを検討すべきである。

(6) 二次利用の利用促進方法について【P：どのように利用を促進すべきか】

利用者の利便性を向上させ、二次利用を促進する観点から、各府省は、毎年度当初に、どのような統計調査について、オーダーメイド集計や匿名データの提供を行うかが分かる情報をそのホームページから公表することが必要である。

また、関係学会と連携して、各府省は、どのような統計調査について、

オーダーメイド集計や匿名データの提供を行うかを周知するようにすべきである。

(7) 不適切利用に対するペナルティについて【P：不適切利用に対しどのようなペナルティを課すべきか】

匿名データを自己または第三者の不正な利益を図る目的で提供又は盗用した場合は、統計法第 61 条 3 号に規定する罰則が適用される。また、利用申出書に記載した以外の利用を行った場合や提供された匿名データを適切に管理しなかった場合等、不適切な利用を行った者に対する対処方法としては、不適切利用を行った者は、その後の一定期間、すべての府省において二次利用の申出を一切受け付けないとするようなペナルティを与えることが考えられる。この場合、オーダーメイド集計や匿名データの提供の契約書や利用申出書に添付する誓約書にペナルティの内容を具体的に記載しそのようなペナルティを受けることに利用者が同意する旨を記載するとともに、ペナルティの適用基準等も規定しておくべきである。

(8) 国外の研究者に対する提供について

統計法における罰則規定のうち、同法第 62 条では、匿名データの提供を受けた者あるいは匿名データの取り扱いに関する業務の委託を受けた者等が図利目的で提供又は盗用した場合の罰則（同法第 61 条第 3 号）については、日本国外においてこれらの罪を犯した者に対しても適用すると規定されていることから、制度的な枠組みとしては、匿名データの提供は外国の研究者に対しても可能である。

国外の研究者に提供する場合には、各種手続のための書類や調査情報に関する膨大な書類等を翻訳する必要があり、そのための環境整備を検討することが必要である。

(9) レプリカデータの作成・提供について

レプリカデータは、統計法で規定している匿名データの一形態と考えるか、全くの擬似データと考えるかによりその取り扱いが異なるので、今後さらに定義を明確化した上でその作成・提供について検討する必要がある。

(10) オンサイト利用の検討について

オンサイト利用は、課題の明確化を図り、引続き検討すべきである。

二次利用の技術的な課題について

1 オーダーメイド集計における集計結果表の秘匿処理

オーダーメイド集計においては、客体数が少なく、集計結果表において特定の個人や事業所等が識別されるおそれがある場合には、秘密の保護の観点から、当該集計結果表に対して秘匿処理を行う必要がある。

しかしながら、集計結果表の秘匿処理の方法については、世帯対象か事業所・企業対象か、また全数調査か標本調査か、表章する項目によっても異なるため、一律の基準を設定することは困難である。

このため、「統計データの二次利用に関するガイドライン(仮称)」(以下「ガイドライン」という。)では、現在各府省が統計調査結果を公表する際に実施している秘匿処理の事例(別紙1)を参考に、目安となるものを提示すべきである。

2 匿名データの作成

(1) 匿名データの作成対象

一般的には、事業所・企業を対象とした調査のデータは匿名化が難しいといわれているが、調査によってデータの安全性や有用性は異なるため、匿名化が必要な項目や匿名化の程度も一律ではない。このため、事業所・企業の統計を一律に匿名化の対象外とすることは適当ではなく、匿名データの作成対象は、原則として、事業所・企業関係の統計を含めすべての統計調査とすべきである。

なお、匿名化できない調査は、オーダーメイド集計で対応することを検討すべきである。

(2) 匿名処理の方法

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、調査ごとに決める必要がある。今後、各府省が行う秘匿処理の参考とするため、ガイドラインには別紙2に示す秘匿処理の技法や研究会においてモデルとして示した別紙3等を参考に匿名化の目安となるものを盛り込む必要がある。

3 秘匿処理の審査

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している。

このような統計委員会における審査を効率的、効果的に実施するためには、実施する秘匿処理の内容を明確かつ簡潔に表現するためのチェックリストが必要であり、各府省間で統一化された様式で、各府省がそれぞれの所管の統

計調査ごとに作成することが必要である。

チェックリストの様式は各府省共通とすべきであるが、秘匿処理の内容は調査ごとに異なることから、各府省は所管の統計調査の結果について、ガイドラインに示される秘匿処理の技法や匿名化の目安を参考に必要な秘匿処理の内容を記載することが必要となる。

なお、匿名データのチェックリストのイメージは別紙 4 及び 5 に示すとおりである。

4 秘匿処理の実施機関

基本的には統計調査を所管する府省が秘匿処理を実施する必要があるが、秘匿化という専門的な知識や技術的なスキルを必要とする処理を政府全体として効率的に実施していくためには、ある程度秘匿処理の実施機関を特定化し、秘匿化の業務を委託することも検討すべきである。

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。

調査対象	標本	全数
事業所等	<p>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)</p> <p>その他、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)</p> <p>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)</p> <p>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)</p> <p>労働者数を10人単位で表章(賃金構造基本調査)</p>	<p>○客体数が少ない場合、結果を非表示(“x”等に置換え)</p> <p>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)</p>
世帯	<p>○表章単位の丸め(千世帯、万人等)(労働力調査、国民生活基礎調査等)</p>	<p>○表章区分の統合(小地域集計(国勢調査))</p>

秘匿処理について

(1) 秘匿処理とは

マイ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）とマイクロデータの対応関係を特定し、特定されたマイ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、調査単位とマイ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

提供するマイクロデータには、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、調査単位とマイ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マイ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マイ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。実際のマイクロデータの提供の場合、一部の調査単位のマイ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、マイ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、その調査単位のマイ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とどまる。

(4) 識別情報

調査対象である調査単位とマイ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、提供するマイ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る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。提供されるマイ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、そのリストとマイ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、現実には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、それはマイ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、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」条項によるマイ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マイ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。

□ 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。

□ 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。

□ 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。

□ リサンプリング

マイ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マイ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、提供するマイ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。

□ ミクロデータのソート

マイ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、マイ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、マイ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。

□ 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。

□ 誤差の導入

マイ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、一橋大学におけるマイクロデータの試行的提供では、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ないことがないよう、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、□乱数による誤差の付加（random noise）、□調査単位間の調査情報の交換（swapping）、□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。

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及び年次

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ぬ。

- (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。

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。