資料２

試行的提供における秘匿処理

一橋大学経済研究所社会科学
統計情報研究センター　山口幸三

１　秘匿処理の考え方

(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。

2 試行的提供における秘匿処理

試行的提供は我が国で初めて国が実施するもので、開始段階では秘匿処理の実際上の問題に関する経験は十分には蓄積されていなかった。このため、当初の段階では強度な秘匿処理を施して安全性に疑問がないようにしておき、その後、利用の実態、利用者及び関係者の意見等に基づき、順次改善を図っていくことが適当であると考えた。

このような考え方に基づき、具体的には、次のような秘匿処理を行った。

(1) 地域区分

地域符号を、「３大都市圏」と「その他の地域」の２区分とする。海外の例と比べると強度な秘匿処理で、これ以外の秘匿処理は不要といえるほどである。なお、今回提供する調査の場合、地域をこのように統合しても十分に利用価値があるものと考えられる。それと同時に、集計用乗率についても集約する。これは、市町村等によって抽出率に差があるので、地域を秘匿しても集計用乗率から市区町村等が判明する場合があるからである。

(2) 世帯の基礎的属性

外部から把握しやすく、誰もが秘匿の安全性に不安を持つような情報である世帯の基礎的属性について秘匿処理を行う。具体的には、年齢と世帯員数について、それぞれでデータを区分したとき、出現するデータの個数が少数になった場合、他の区分と統合する。

ア 年齢

80 歳以上はすべて 80 歳に変換する。なお、調査対象特定の可能性を低下させるため、出生年月の情報は提供せず年齢のみを提供する。

【80 歳とする理由】

○　年齢別の人口分布は下記のとおりであり、80 歳代でもかなりの人口がある。これは人口の割合なので、世帯にいる高齢者の割合で考えれば割合は更に高くなる。したがって、80～84 歳についてそのまま年齢を提供しても世帯数は多く、世帯特定の危険性は高いものではない。

平成 12 年国勢調査

総人口	100.000%
75 歳	0.794
76 歳	0.705
77 歳	0.645
78 歳	0.591
79 歳	0.540
80 歳	0.538
81 歳	0.411
82 歳	0.398
83 歳	0.371
84 歳	0.346
85 歳以上	1.763

○　しかし、利用の面から考えれば、75 歳以上の後期老年人口は通常就業していることも稀であるので、今回提供する調査の場合、一括して分析することにしてもほとんど問題ない。

○　すなわち、80～84 歳については秘匿処理の必要性は低く、利用面からは 75 歳以上をまとめても問題ない。一方、年齢は外部から把握しやすく誰もが秘匿の安全性に不安を持つような情報であることから、何らかの秘匿処理を行っておくべきである。これらのことを総合的に勘案して 80 歳以上をまとめることが適当と考える。

イ　世帯員数

世帯員が９人以上の世帯は、その世帯全体を削除する。

【９人以上とする理由】

○　世帯員数の分布と推計されるデータ数は、下記のとおりである。８人世帯までは、全国消費実態調査、社会生活基本調査とも約 200 以上のデータがあり特定の可能性は低いが、９人以上は少なくなる。

平成 12 年国勢調査		社会調	全消
一般世帯	100.0%	約７万世帯	約６万世帯
６人	3.1	2,170	1,860
７人	1.3	910	780
８人	0.31	217	186

９人	0.06	42	36
10 人以上	0.02	14	12

○　世帯員数は外部から把握しやすく誰もが秘匿の安全性に不安を持つような情報であるが、その不安は、どのような条件のときに特殊な世帯であると人々が想定するかということにもよる。親夫婦に子夫婦とその子供の三世代世帯などは世帯人員数も多く、どこにでもある世帯であることから、世帯員数８人までのデータはそのまま提供しても不安を与えるおそれはない。

○　一方、利用の面から考えれば、今回提供するような調査で９人以上の世帯をターゲットとして分析することは考えにくく、削除しても分析への影響は少ない。

(3) 調査事項のうち年間収入

年間収入については、秘匿の必要性が高いことから、全国消費実態調査において 3,000 万円以上の場合をすべて 3,000 万円に変換すべきであるとの考え方があったが、以下の理由によりそのような加工は行わず、原データのまま提供することとする。

○　年間収入階級別の分布は下記のとおりである。報告書からはこれ以上の年間収入の分布は分からないが、1,000 万円台が 20.62%、2,000 万円以上が 1.64%なので、1,000 万円幅で 13 分の１くらいになっている。そのことに基づいて推計すれば、3,000 万円以上で 0.13%程度となる（通常、裾を引く分布になるので過少推計と思われる。）。これは、集計世帯数で 80 世帯程度となり、外部から分かる情報の場合であれば特定の危険性の観点から、3,000 万円台、4,000 万円台の数値をそのまま提供せず 3,000 万円以上として一括するのが適当であるということになる。

【平成 11 年全国消費実態調査　二人以上の一般世帯】

	100.00%	集計世帯数
1,000～1,250	10.98	5,710
1,250～1,500	5.23	2,766
1,500～2,000	4.41	2,245
2,000 以上	1.64	825

○　しかし、年間収入は外部から分かる識別情報ではなく調査事項であり、世帯を特定する危険性を高めるわけではない。また、年間収入は各種の分析でよく利用される。もし 3,000 万円以上を一括するような加工を行うと、ジニ係数の計算などは不可能となってしまう。

○　年間収入については、その内訳（勤め先収入、事業収入、社会保障給付など）もよく分析に利用される。もし年間収入の総額についてトップコーディングを行うと、内訳の値を不詳としなくてはならなくなる。また、不詳にならないよう何らかの推計を行ったとすると、誤った分析結果を導いてしまうおそれがある。

(4) リサンプリング

今回は我が国で初めて国が行うデータ提供であるので、秘匿の安全性を担保するため

に十分すぎる対策を施すこととした。ここで、リサンプリングを行っておけば、調査を受けていたという情報が存在する条件（時間的な情報の劣化を考えれば、現実的な想定ではないと思われる。）の下であっても、対象が特定できたとの主張に対し、全データを提供しているわけではないのでその主張には無理があると説明できることになる。このため、全体から８割の世帯をリサンプリングしたデータを提供することとした。

【リサンプリングの方法】

リサンプリングとして、①単純任意抽出を行う方法、②集計用乗率で確率比例抽出を行う方法、の２つの方法が考えられる。単純任意抽出の場合、集計用乗率の大きさに関係なく抽出されるので、標本調査の標本と同じ構成の 80%標本が得られる。したがって、集計用乗率を用いない集計結果は母集団を代表するものにはならない。確率比例抽出の場合は、集計用乗率の大きさに比例して抽出されるので、標本調査の母集団と同じ構成の 80%標本が得られる。したがって、集計用乗率を考慮しなくても、集計結果は母集団を代表するものになり、データとしては扱いやすい。しかし、集計用乗率の大きいデータが 80%標本の中に複数抽出されるという問題が生じる。

以上のことを考えて、基本的には単純任意抽出を採用している。しかし、全国消費実態調査では、集計用乗率によって市区町村等が明らかになる可能性があるので、次のような処置をとった。

集計用乗率の大きさ別に階級区分し、その階級ごとに確率比例抽出を行うことにし、それぞれのデータには、集計用乗率として、その階級区分の集計用乗率の平均値を付す。結果的には、単純任意抽出と確率比例抽出を組み合わせたリサンプリングの方法となっている。この場合、集計用乗率を考慮しなくても集計結果が母集団を代表するものとなるという確率比例抽出の長所はなくなっている。