参考３

匿名標本データの試行的運用における利用者アンケートの概要について

出典：安田聖、山口幸三、横内宏至(2009)「政府統計ミクロデータの試行的提供」、統計資料シリーズ No.６４、一橋大学経済研究所付属社会科学統計情報研究センター

資料3-5　アンケート結果

匿名標本データ「第Ⅰ期試行的運用」における利用者アンケートの概要

匿名標本データの「第Ⅰ期」利用者に対するアンケート（計30件）の主な結果は、次のとおりである。

1 利用者の資格・利用手続等について

○　大学院生、ポスト・ドクター、民間研究機関の研究者を共同利用者にできるようにしてほしい
○　大学の助手、公的な研究機関の研究員、短期大学の教員など申請者の範囲を広げてほしい
○　６か月では時間不足で、十分な研究ができなかった
○　当初申請していない集計を行えなかったため、分析に支障が生じた
○　２度目の利用のときには、説明会への出席を免除してほしい
○　集計様式の書き方の説明がわかりにくかった
○　個人が複数の調査の申請を出すことを認めてほしい
○　共同研究の場合、どのような申請が可能なのかわかりにくかった
○　試行錯誤を含む集計様式を記載するのは負担なので、省力化してほしい

2 提供データの形式等について

○　地域区分がないので、地域性の分析ができなかった
○　年次により符号の付け方が違っていて不便であった
○　SPSSなどに対応した形式のデータで提供してほしい
○　データが文字型であるため、形式の変換が不便であった
○　人口の多い市、特別区については、地域区分を付与してほしい
○　変数名は日本語のローマ字綴りになっているので読みづらい
○　全国消費実態調査の場合、調査票が全部そろっていない分のデータも提供してほしい

3 リサンプリングと誤差の付加等の比較について

○　下記のいずれのデータを使いたいか
　　抽出率80%のリサンプリング・データ（28）
　　誤差の付加やスワッピングを行った全データ（０）
　　どちらでもよい（２）
○　リサンプリングの率と利用の希望
　　20%でも利用したい（18）
　　50%なら利用したい（４）
　　50%以下では利用しない（６）
　　分からない（２）
○　家計調査などデータ数の少ない調査では、高いリサンプリング率（と誤差の導入）が必要かもしれない
○　サンプルが多くなければできない分析もあるが、20%でも分析可能な研究課題もあると思う

4 試行的提供を知った方法について

学会、研究会で（7）
知り合いの研究者等から（13）
センターのホームページ（10）

5 提供データ（各調査3回分のデータ）以外の希望年次について

就業構造基本調査 …… 過去の年次
全国消費実態調査 …… 過去の年次、最新年次
社会生活基本調査 …… 過去の年次

6 利用したい統計について

国勢調査	国民生活基礎調査
事業所・企業統計調査	賃金構造基本統計調査
住宅・土地統計調査	消費動向調査
サービス業基本調査	法人企業統計調査
家計調査	所得再分配調査
労働力調査	21世紀成年者縦断調査
労働力調査特別調査	パートタイム労働者総合実態調査
高齢者就業実態調査	自動車輸送統計調査

7 ミクロデータ提供のあり方等について

○ 多くの研究者に平等に研究の機会を与えることは非常によい
○ 今後も継続的に提供されることを望む
○ ミクロデータの分析が可能になることは、日本の研究発展に大きな役割を果たす
○ このような提供方式が制度化されることを希望する
○ 研究者であれば自由にアクセスできる仕組みも必要
○ データの種類や年次を拡充してほしい
○ より低い抽出率でもよいので、簡単な手続で使用できるとよい
○ データの開放による調査自体の改善ができるようになる
○ 利用者がいつでも申請できるようにすることや、申請から提供までの期間を短縮することを望む
○ 投稿、査読、再投稿などのプロセスに配慮した形での利用環境を提供してほしい
○ 利用者から提出された論文等のリストを公開するようにできないか
○ 自分の分析結果の妥当性を評価できるよう、ミクロデータによる分析結果の蓄積を期待
○ ミクロデータの優秀論文の表彰などあるとよい

政府統計ミクロデータ提供
平成16年11月～平成18年4月

	H16.11	H17.4	H17.10	H18.4
回収	6	4	11	9
未回収	0	0	1	0

1　利用者の資格、利用手続き等について

申請者及び共同利用者の資格、利用の目的で変更してほしいことはありますか。

	H16.11	H17.4	H17.10	H18.4
1 特にない	2	4	6	4
2 ある	4	0	5	5
未回答	0	0	1	0

○共同利用者の範囲を大学院生、研究機関等の研究員まで広げてほしい(13)
○2回目以降の説明会を省略してほしい。また、申請から利用開始までの期間を短縮してほしい
○共同論文を作成する場合の条件について明示してほしい

今回の利用にあたって、手続や申請書の記入方法などで分かりにくかった点はありましたか。

	H16.11	H17.4	H17.10	H18.4
1 特にない	6	4	9	8
2 ある	0	0	2	1
未回答	0	0	1	0

○集計様式の書き方について、もっと詳しく記載してほしい(2)
○共同研究者と共同利用者の区別を明示してほしい

利用の手続や申請書の書き方などで今後変更してほしいことはありますか。

	H16.11	H17.4	H17.10	H18.4
1 特にない	6	3	10	5
2 ある	0	1	1	4
未回答	0	0	1	0

○2回目以降の説明会を省略してほしい(2)
○申請書の集計表や多変量解析等の変数の記載について簡素化してほしい(2)
○申請から利用開始までの期間を短縮してほしい
○記入様式をダウンロードできるようにしてほしい
○データ提供後でも集計項目等の追加・変更ができるようにしてほしい

「秘匿処理済ミクロデータの使用条件」に、「提供されたミクロデータが万一流出した場合、申請者の意図とは関わりなく、流出元となった申請者が使用条件に違反したものとみなします。」という条件を付け加えた場合、あなたはミクロデータを利用しますか。

	H16.11	H17.4	H17.10	H18.4
1 利用する	4	4	10	7
2 利用しない	1	0	1	1
未回答	0	0	1	1

○違反した場合の罰則内容によるため、回答が難しい
○条件がついても利用はしたいが、罰則等には注意の度合いを反映してほしい
○最大限の努力をしてもリスクが残るため利用しない
○一応利用すると回答したが、他の説明も追加的になされると想定されるので、回答が難しい
○研究上必要なので利用するが、この文言はかなり厳しいのではないか

2　今回提供したデータについて

(1)　提供したデータの形式等

今回はCSV形式のみでデータを提供しましたが、他の形式での提供についての要望はありますか。

	H16.11	H17.4	H17.10	H18.4
1 特にない	5	4	10	7
2 ある	1	0	1	2
未回答	0	0	1	0

○SPSSに対応してほしい(3)
○SAS、SPSS等の形式に対応してほしい
○固定長は利用しにくいため、CSVの方が良い
○固定フィールド長のテキストデータも併せて提供してほしい

試行的提供システムで提供したミクロデータについては、秘密保護のための匿名化処理がされています。このような処理をされたデータであるために分析目的が何か制約を受けましたか。

	H16.11	H17.4	H17.10	H18.4
1 特に制約されなかった	4	4	8	6
2 制約された	2	0	3	3
未回答	0	0	1	0

○地域が2区分のみのため、地域性について分析できなかった(4)
○地域が2区分のため、要因分析(ロジット等)で、変数として利用できるような結果が得られなかった(秘匿のためなのか、そうでないのかわからない)
○2002年の就調の「3大都市圏」が利用できなかった
○地域が2区分のみのため、地方財政との関係について分析を行うことができなかった。人口の多い市・特別区については秘匿しないでほしい
○回帰分析では、地域変数が大きく影響することもあるため、地域区分を、3大都市圏・それ以外だけではなく、人口の多い都道府県については都道府県名で提供してほしい

今回提供したのは、全体から80%のデータを無作為抽出(リサンプリング)したものです。このようなリサンプリングを行わず、その代わりにランダムな誤差を加えたり、(レコード間で一部のデータをランダムに入れ替える)スワッピングを行った全データが提供されるとしたら、あなたはどちらのデータを使いたいですか。

	H16.11	H17.4	H17.10	H18.4
1 リサンプリング・データ	5	4	10	9
2 誤差の付加やスワッピングを行った全データ	1	0	0	0
未回答	0	0	2	0

前問で1または2のデータの利用を希望されるのはなぜですか。その理由を具体的に記入してください。

	H16.11	H17.4	H17.10	H18.4
記入有	5	4	11	9
未回答	1	0	1	0

○原データそのものの情報を活かしたいから
○説明変数に誤差が入っていると修正に手間がかかるから
○人為的に加工したデータではなく、生の(リサンプリングであっても)データの方を使いたいから
○抽出率が80%でも、かなり十分なデータ数であり、誤差の付加やスワッピングよりも結果の解釈はしやすいと感じるため
○分析によるとは思うが、一般に、より問題が少ないと思われるから
○計量分析のデータとして論文で引用しやすい。出生力の分析などサブ・サンプルを利用する場合、誤差の推定が困難
○誤差の付加やスワッピングによるデータの変化について、知識がないため
○加工処理は単純な方がよいと思うため

○抽出率80%で十分であるため
○真のデータにより近いデータセットの方が望ましいため
○ミクロデータそのものがサンプリングデータなので、80%抽出されていても特に問題はないが、データそのものが加工されるのは抵抗があるため
○リサンプリング・データの方が、もしランダムサンプリングであるならば、後者に比べて統計学的性質がはるかに良く、後者はバイアスの問題などあり、統計学的に扱いがやっかいであるため
○現行の80%抽出データは多重クロス集計にも有効なデータと思われるため
○いずれでも構わない
○リサンプリング・データの方が、方法が単純で明快であるから
○データの整合性が心配であるから
○誤差の付加やスワッピングされたデータはOLSで推計値の偏りを生むから
○誤差の付加やスワッピングによる効果が明確でないため
○原データの情報をより活用して分析できると思うから
○特に変更の必要を感じないため
○分析上、ほとんど問題ないと思われるから
○現在のやり方が、秘匿処理をした上で、データの安定的な利用につながる最もわかりやすい方法であると思うから
○変数間の関係を調べるのが目的なので、誤差が入るよりはサンプルサイズが小さいほうが許容できるから
○プライバシーの問題が無ければ全レコード使えることが望ましい。次善の策として現状と同等と思うが、誤差の付加やスワッピングと比較してのメリット・デメリットがよくわからない
○スワッピングを行った結果、たとえば世帯人員を実際に属している世帯とは異なる世帯に入れ替えるようなことがなされることになるとすると、世帯の属性と個人の属性を関連付けることが正確に行えなくなるのではないかと予想される
○誤差の付加やスワッピングよりリサンプリング・データの方が全データの性格をよりはっきり反映しており、比較した時、理解しやすい
○リサンプリング・データでも十分な標本数が得られているので、報告されたとおりのデータに基づく分析をしたいから
○スワップデータの利用可能性にかんする検証が十分になされてきたとは言えないから
○誤差やスワッピングにより、調査されたデータに忠実ではなくなると思われ、抽出率80%の方が正確性を一定程度確保できると思うから

今回は全体の80%のデータをリサンプリングして提供しましたが、リサンプリングの率が20%または50%の場合でも利用したいとお考えでしょうか。

	H16.11	H17.4	H17.10	H18.4
1 20%でも利用したい	5	2	4	7
2 20%では利用したいと思わないが、50%なら利用したい	0	1	2	1
3 50%以下では利用したいと思わない	0	1	4	1
4 分からない	1	0	1	0
未回答	0	0	1	0

前問で1～3と答えられた場合、その理由を具体的に記入してください。

	H16.11	H17.4	H17.10	H18.4
記入有	5	4	10	9
未回答	1	0	2	0

○サンプル数が多い調査では20%でも良いが、例えば家計調査ではサンプル数が少ないため、高いリサンプリング率が必要
○継続的な利用が可能であれば、20%でも有用と考えるから
○多いにこしたことはないが、20%であっても分析はおそらく可能であるため
○恒常的に提供される仕組みが作られれば、再抽出率が低くても、使用したい
○20%であっても、個票データを扱うメリットが大きいと思えるから
○サンプル抽出率に応じて異なる分析が可能、但し、20%の場合、当然のことながら、かなり制限を受ける

○50%データでも研究の工夫に有効で、ある程度試行した上で抽出率の大きいデータの申請すれば効率的
○50%以下では、原データと乖離が大きいと思うから
○リサンプリング率が低くても、研究を行うのに十分な量であるため
○標本数が大きく、無作為抽出である場合、リサンプリング率20%でも、バイアスが少ないため
○多重クロス集計を行った場合、現状でも1セルのサンプル数が100人以下になる場合がある。20%では信頼性のある結果が出せない
○標本サイズが大きければ、20%のリサンプリング率でも、ランダムなリサンプリングが行われれば、統計学的には充分に良い性質を保持するであろうから
○多重クロス集計の有効性が低下するため
○サンプリング率は大きいほうが望ましいが、仮にそれが小さくとも利用目的によっては有用であるから
○50%以下では、100%との間で、特性などの乖離が大きくなると思うから
○50%以下では、マイクロ・データとしてサンプル数が不足するから
○多いほうがよいが、20%でもかなりのサンプル数になるから
○分析方法にもよるが、一般的に50%でも有意な結果が得られると考えられるため
○リサンプリング率が大きい方が、母集団の情報をより小さい誤差で利用できると思うから
○利用する調査のサンプルの数による
○分析の内容ならびに目的次第では、20%抽出でも統計的に意味がある分析は可能だと思われるため
○秘匿処理の程度にもよるが、リサンプリング率が低くても使用できるデータがそれしかない場合、使用せざるを得ないから
○ジャーナルによってはデータの質として不十分だと判断されると思われるから
○リサンプリング率は高いほうが当然望ましい。ただ、20%でも分析可能な研究課題もあると思う
○標本数が多ければ、20%でも、ランダム・サンプリングであれば、全データを利用するのと違わない
○まったく利用する機会が得られないよりは、限定的でも利用できたほうが良いから。ただし、分析内容によってはかなり制約を受けると思うのでできるだけ100%に近いデータを提供してほしい
○「ミクロ統計データ活用研究会」において、20%のリサンプリングデータに関するユーザビリティーが検証されているから
○リサンプリング・データの利用は研究上必要不可欠なので、80%を維持してほしいが、統計法規やプライバシーなどの高まりを考慮して譲歩して50%とした

(2) **就業構造基本調査**

変数名の付け方等で分かりにくいところ、変更してほしいところはありますか。

	H16.11	H17.4	H17.10	H18.4
1 特にない	1	2	2	4
2 ある	0	0	2	3
未回答	5	2	8	2

○項目の内容やコードに変化がない場合には、調査年を通じて同一の変数名としてほしい。例えば、1992年の"Shugyoldo"と、1997年、2002年の"Shuugyouldou"はどちらも「就業異動」なので、同じにしてほしい
○他国のミクロデータでは、「ラベリング」がなされているので、変数にラベリング処理を付したデータの提供をしてほしい
○日本語をローマ字つづりにしたものは読みづらく、英語の方がわかりやすい
○どの調査年度でも同様と思われるにもかかわらず、変数名が微妙に異なっていることがあった。質問項目の様式が調査年毎に異なるのであれば利用者に注意を喚起するために変数名を変更することも大変有意義と思うが、そうでない場合については、統一したほうが利用者の便宜をはかれると思われる
○年度によって、変数名が異なっていることがあるので、統一してほしい

今回、データを利用していて何か誤りを発見されましたか。

	H16.11	H17.4	H17.10	H18.4
1 特にない	1	1	4	4
2 ある	0	1	0	3
未回答	5	2	8	2

○2002年の地域変数が利用できない(2)
○文字型でデータが入っていて、形式の変換に手間どった
○一部変数で不詳等でアルファベットが含まれるものがあった。また、都道府県コードがないのは難点だと感じた
○年齢と継続就業期間をクロス集計したとき、論理的に矛盾していると思われる年齢階層が見られた

今回、データを使用する上で何か直面した問題はありましたか。

	H16.11	H17.4	H17.10	H18.4
1 特にない	0	1	3	4
2 ある	1	1	1	3
未回答	5	2	8	2

○空白も含めて読み込んだため、データが桁ずれを起した
○調査を行うとき、時系的分析(比較)が行えるよう、慎重に検討して欲しい。特に「就調」は他の調査と比較して、変更が多い気がする
○初歩的なことだが、SPSSデータに変換する際、スペースの処理を誤って、最初はとまどった
○世帯項目には、収入の種類を尋ねているのに、個人に関する事項では、その項目がない。どうして、個人に関してはないのか。また、既存の研究結果から、年金、あるいは年金の所得代替率が就業ないし就業率と関係していると言われているので、そうした項目を個人に関する項目に追加すべきであると思われる
○秘匿処理のため地域情報(市町村コード)を利用できなかったこと。財政との関連で分析をすすめる場合は、観測値の居住地が重要になることが多く、市町村コードが使用できることが望ましい。市レベルでは秘匿の問題はないと思われるので、せめて市コードまでは利用したい
○データについて、不詳や産業・職業分類にアルファベットが使われていたため、文字型変数として読み込むしかなく、数値型に変換するのが大変であった(SPSS使用)。不詳を9などすべて数字が割り当てられたデータだったら、分析に使う状態までデータセットを整えるのに大幅に時間が節約できたと思われる。また、都道府県コードがないので政策効果を分析する際に、データマッチングができないことは難点だと感じた。この点は改善してほしい

(3) 全国消費実態調査

今回は購入先、曜日別データを提供していませんが、そのために分析目的が何か制約を受けましたか。

	H16.11	H17.4	H17.10	H18.4
1 特に制約を受けなかった	3	3	8	3
2 制約された	1	0	0	1
未回答	2	1	4	5

○供給者側の変化、曜日別の消費者行動の分析が難しくなる
○単身個人に関するサンプルが少なすぎて、性別の年令別分析に支障をきたしている

今回はすべての調査票がそろった世帯から提供用のデータを抽出しました。今回の方式と、すべての調査票がそろっていない世帯も含めて提供用のデータを抽出する方式のどちらが使用する上でいいですか。

	H16.11	H17.4	H17.10	H18.4
1 現在の方式でよい	3	3	7	2
2 調査票がそろっていない世帯も含める方式がよい	1	0	1	2
未回答	2	1	4	5

○調査への協力の度合いにより、調査票への記入の正確さが変わると思われる。すべての調査票がそろっていない世帯についての集計によりその効果を推測することができるため
現在の方法で構わないが、調査票がそろっていないことの理由が経済変数と相関をもつかどうかは気になる
○属性により、調査票のそろっている世帯の率が異なると考えられるので
○ランダム・サンプリングにしてほしい。全データとの比較のため。資産データがない場合でも含めてほしい
○80%の抽出率があれば、どちらでもよいと思う。調査票がそろっていない世帯も含める場合、研究上必要な変数のデータが一定程度揃えば、使えるケースが増えると思うので、メリットもあるかと思われるが

今回のデータでは世帯ごとに1レコードとしてありますが、そのことで使う上で何か問題はありましたか。

	H16.11	H17.4	H17.10	H18.4
1 特にない	4	3	7	3
2 ある	0	0	1	0
未回答	2	1	4	6

○世帯人員の情報を集計するときに、具体的説明がないので、集計作業に時間がかかった
○個人レベルに変換できるので、特に無いとしましたが、世帯ごとだけでなく、個人ごとのデータも提供されれば、世帯内の個人の関係を研究する場合では、面倒な作業を減らせる。ちなみに、イギリスでは、世帯調査(SARsかGHS)が、世帯レベルと個人レベルのデータで提供されていた

○今回のデータでは、秘匿のために乗率を集約して付け替えていますが、そのことで分析を行う上で何か支障がありましたか。

	H16.11	H17.4	H17.10	H18.4
1 特に支障はなかった	4	3	8	4
2 支障があった	0	0	0	0
未回答	2	1	4	5

変数名の付け方等で分かりにくいところ、変更してほしいところはありますか。

	H16.11	H17.4	H17.10	H18.4
1 特にない	4	3	7	2
2 ある	0	0	1	2
未回答	2	1	4	5

○数値型にして、ラベリング処理を施したデータを提供してほしい
○年収の項目のコーディングが調査年によりまちまちなようなので、統一してもらえると処理が簡単になると思われる
○符号だけではなく、変数名やラベル名をつけてほしい

今回、データを利用していて何か誤りを発見されましたか。

	H16.11	H17.4	H17.10	H18.4
1 特にない	3	3	6	4
2 ある	1	0	2	0
未回答	2	1	4	5

○1989年の集計用乗率が、他年度と比較して、1桁異なっていた
○平成11年における二人以上の一般世帯続き柄を示すS1_Tsuzukiの一部に「1」以外の数値があり、S2_Tsuzuki～S6_Tsuzukiの一部に「1」が、それぞれ入っていた

今回、データを使用する上で何か直面した問題はありましたか。

	H16.11	H17.4	H17.10	H18.4
1 特にない	1	2	8	2
2 ある	3	1	0	1
未回答	2	1	4	6

○当初申請した以外のデータの集計により研究が進展しそうだったが、途中で集計表の追加ができないため、試行錯誤ができない
○分析時間が足りない
○用途別分類と品目別支出で、いくつかの家計で菓子への支出額が異なり、その結果、食品支出また消費支出が異なった
○1994年の年収項目の区分が1989年、1999年とは異なる
○学歴の項目が無い

(4) 社会生活基本調査

今回はアフターコード調査票のデータは提供していませんが、そのために分析目的が何か制約を受けましたか。

	H16.11	H17.4	H17.10	H18.4
1 特に支障はなかった	1	0	1	0
2 支障があった	0	0	0	0
未回答	5	4	11	9

変数名の付け方等で分かりにくいところ、変更してほしいところはありますか。

	H16.11	H17.4	H17.10	H18.4
1 特にない	0	0	1	0
2 ある	1	0	0	0
未回答	5	4	11	9

○ボランティアで1996年はVolunteer1～8となっていて、できれば2001年や他の行動と同様に01～08の方が良かった

今回、データを利用していて何か誤りを発見されましたか。

	H16.11	H17.4	H17.10	H18.4
1 特にない	1	0	1	0
2 ある	0	0	0	0
未回答	5	4	11	9

今回、データを使用する上で何か直面した問題はありましたか。

	H16.11	H17.4	H17.10	H18.4
1 特にない	0	0	1	0
2 ある	1	0	0	0
未回答	5	4	11	9

○都道府県別、地域ブロック別または都市階級別があるとよい
生活時間と生活行動を結びつけた分析ができなかった