製表技術参考資料 16

ISSN 1348－7329

教育用擬似ミクロデータの開発とその利用

～平成 16 年全国消費実態調査を例として～

NSTAC

Working Paper No.16

平成 24 年 7 月

独立行政法人 統計センター

製表技術参考資料は、独立行政法人 統計センターの職員がその業務に関連して行った製表技術に関する研究の結果を紹介するためのものである。

ただし、本資料に示された見解は、執筆者の個人的見解である。

目　次

要旨 …… 1
はじめに …… 3
１　背景と目的 …… 3
1.1　統計法の改正 …… 3
1.2　新統計法の概要 …… 4
1.3　新統計法の下でのミクロデータの利用 …… 4
1.4　二次利用の実績 …… 5
1.5　教育用擬似ミクロデータの開発 …… 6
２　擬似ミクロデータの基本的な考え方 …… 7
2.1　調査票情報と匿名データ …… 7
2.2　教育用擬似ミクロデータ …… 8
2.3　高次元の集計表 …… 8
３　教育用擬似ミクロデータの作成方法 …… 9
3.1　量的属性と質的属性の選択 …… 10
3.2　度数１又は２に該当するレコードの処理 …… 12
3.3　高次元の集計表の作成 …… 13
3.4　多変量正規乱数の生成 …… 14
3.5　パターン別集計表を用いた量的属性値０の付与 …… 17
3.6　加法性と収支バランスの調整 …… 18
3.7　集計用乗率の付与 …… 19
４　教育用擬似ミクロデータの分析特性 …… 19
５　教育用擬似ミクロデータの試行提供の現状とアンケートの結果 …… 25
5.1　試行提供の現状 …… 25
5.2　アンケート結果 …… 26
６　おわりに―教育用擬似ミクロデータの作成に関する今後の課題 …… 27
6.1　当面の課題 …… 28
6.2　将来の課題 …… 28
参考１　平成 16 年全国消費実態調査の教育用擬似ミクロデータ　質的属性及び量的属性一覧表…… 29
参考２　教育用擬似ミクロデータを作成するまでの分布イメージ…… 33
参考３　教育用擬似ミクロデータの試行提供利用者アンケートのお願い…… 35
参考４　教育用擬似ミクロデータの試行提供利用者アンケートの回答結果…… 37
付論　ミクロアグリゲーションについて…… 39
参考文献 …… 43

教育用擬似ミクロデータの開発とその利用
～平成 16 年全国消費実態調査を例として～

秋山 裕美*，山口 幸三†，伊藤 伸介‡，星野 なおみ§，後藤 武彦**

要　旨

改正された統計法（平成 19 年法律 53 号）が平成 21 年４月から全面施行されることによって、公的統計のミクロデータの利用が進み、定着しつつあるとみられる。しかしながら、事前の想定よりも利用者が拡大しているとも言えない状況である。そこで、公的統計のミクロデータの利用を推進するために、大学等の教育機関における授業や演習で利用可能な「教育用擬似ミクロデータ」を開発し、統計法令の制約も受けず、ミクロデータを自由に使える環境を整備することを計画した。そうした環境を整備することが、ミクロデータを利用した実証分析ができる人材を育成し、その結果として、ミクロデータの利用を促進し、学術研究水準を向上させることになると考えている。

教育用擬似ミクロデータは、本来の調査票情報（個票データ）から集計した集計表を基に作成した擬似的なミクロデータである。教育用擬似ミクロデータは、個票データとは異なるが、基となったものに近い集計表が復元できるため、個票データの特性を有していると言える。

教育用擬似ミクロデータの作成方法に関する基本的な方法は、個票データから高次元の集計表を作成し、集計表の各セルの量的属性値が多変量（対数）正規分布に従うことを仮定し、多変量（対数）正規乱数を生成することである。具体的には，平成 16 年全国消費実態調査の個票データに基づいて、(1)擬似ミクロデータに含まれる質的属性と量的属性の選択、(2)個票データを用いた高次元の集計表（質的属性を分類事項とした度数、量的属性の平均値及び相関係数行列等）の作成、(3)高次元クロス集計表に含まれる度数１及び度数２のセルの取扱い、(4)高次元の集計表に含まれるセルごとの多変量(対数)正規乱数の生成、(5)擬似ミクロデータにおける量的属性値０の付与、(6)加法性と収支バランスの調整、(7)集計用乗率の付与を行うというものである。これら一連の処理により教育用擬似ミクロデータの作成を行った。この教育用擬似ミクロデータの分布特性については、個票データの分布と比較してより大きな散らばりを持つ場合もあるが、全般的には個票データのそれにほぼ近似していることが実証的に明らかになっている。

* 統計センター情報技術部情報処理課(前統計センター情報技術部研究主幹)
† 統計センター情報技術部
‡ 統計センター情報技術部統計技術研究課非常勤研究員(明海大学経済学部准教授)
§ 統計センター情報技術部統計技術研究課
** 統計センター情報技術部統計技術研究課

このようにして作成された平成 16 年全国消費実態調査の教育用擬似ミクロデータに関する試行提供は、平成 23 年８月 25 日に開始された。平成 24 年６月 30 日現在、試行提供を開始して 10 か月程経過し、提供件数は 51 件(大学教員 20 名、研究員１名、大学院生１名、学部生 29 名)となっている。また、教育用擬似ミクロデータの利用者に利用内容に関するアンケートを実施し、現時点で 17 件の回答が得られている。アンケート結果からの利用者の意見や要望を踏まえ、教育擬似ミクロデータの提供についてのサービスの拡充や教育用擬似ミクロデータの作成に関する研究をさらに進めていく必要があると考えている。

はじめに

新統計法[1]が平成 19 年５月に改正され、平成 21 年４月に全面施行された。新統計法において、公的統計は「国民にとって合理的な意思決定を行うための基盤となる重要な情報である」とされ、言うなれば国民の共有財産と位置付けられている。そうした理念の下に、公的統計の利用促進のために、統計データの二次利用[2]に関する制度が設けられ、学術研究や高等教育の発展に資する場合に、委託による統計表の作成及び匿名データの提供ができることになっている。

この二次利用制度が開始され、匿名データの提供は、定着しつつあるものの、新統計法に規定されている利用目的の制約、利用環境の制約を受けざるを得ない。そのため、多数の学生を対象とした大学等での講義や統計演習などの利用は、現実問題として困難である。このようなことから、統計法に制約されない統計データの開発が、統計委員会等で議論され、大学の研究者からも要望されていた。こうした背景から、自由に利用できる教育用擬似ミクロデータの開発を計画し、平成 16 年全国消費実態調査の教育用擬似ミクロデータを統計的な手法を用いて開発し、実際に試行的に提供できるまでに至っている。

本稿では、まず、教育用擬似ミクロデータを開発するに至った背景と目的について述べ、次に、教育用擬似ミクロデータ作成上の基本的な考え方を示し、その基本的な考え方に基づいた平成 16 年全国消費実態調査を例とする作成方法を述べる。さらに、作成した教育用擬似ミクロデータとその基になっている個票データの分布特性とを比較分析し、試行的に提供した現状と利用者からの意見等を掲げ、最後に今後の課題を提示する。

１　背景と目的

1.1　統計法の改正

統計法（平成十九年法律第五十三号）は、統計に関する基本法として、旧統計法（昭和二十二年法律第十八号）を全部改正し、統計報告調整法（昭和二十七年法律第百四十八号）の廃止とともに、平成 19 年５月 16 日に成立した。この新統計法は、平成 19 年５月 23 日に公布され、戦後、統計制度が再建されて以来 60 年振りの抜本的改革となった。

新統計法のうち、公的統計の整備に関する基本的な計画や統計委員会の設置などに関する一部の規定は平成 19 年 10 月１日に先行施行され、その他の規定も平成 21 年４月１日に

---

1 本稿では、統計法（昭和二十二年法律第十八号）を旧統計法、統計法（平成十九年法律第五十三号）を新統計法と称する。

2 新統計法の第三十二条による調査票情報（ミクロデータ）の利用を二次利用という。すなわち統計調査計画時点での統計表作成以外の作表は、当初の目的外とみなされる。新統計法第三十三条、オーダーメード集計並びに匿名データの作成及び提供を二次的利用としているが、本稿では二次利用で表す。

施行された。

1.2 新統計法の概要

統計法の改正は、「行政のための統計」から「社会の情報基盤としての統計」へ転換し、公的機関が作成する統計が、より体系的・効率的に整備され、国民・事業者にもより使いやすいものとなることを目指しており、①公的統計の体系的かつ効率的な整備及びその有用性の確保を図るため、公的統計の整備に関する基本的な計画の策定、②統計データの利用促進に関する措置、③統計調査の対象者の秘密保護の強化、④統計整備の「司令塔」機能の強化が主な内容となっている。

新統計法の概要は、第一条に「目的」、第二条に「定義」、第三条に「基本理念」が掲げられ、第四条～第三十一条に公的統計の体系的整備と公的統計の整備に関する基本的な計画の策定、第三十二条～第四十三条に統計データの利用促進及び秘密の保護、第四十四条～第五十一条に統計委員会の設置、第五十二条～第六十二条は罰則等となっている。そして、附則の第一条に施行期日が規定されている。

統計委員会は、第四十四条～第五十条及び第五十一条に規定する統計委員会令（平成十九年政令第三百号）に基づき、平成 19 年 10 月 1 日に設置された。統計委員会では、第四条の規定に基づき、「公的統計の整備に関する基本的な計画」（基本計画）を定めた。政府は、この計画を平成 21 年 3 月 13 日に閣議決定し、計画に盛り込まれた内容を着実かつ計画的に推進することとなっている。

1.3 新統計法の下でのミクロデータの利用

ミクロデータの利用については、旧統計法の下でも、第十五条、第十五条の二及び第十五条の四に規定があり、高度の公益性がある場合には、使用が認められていた。これを目的外使用制度という。新統計法の下でも同じような規定が第三十三条に定められている。第三十三条の第一号は公的機関（国の行政機関、地方公共団体、独立行政法人等）が統計の作成等を行う場合の規定であり、第二号は公的機関の統計の作成等と同等の公益性を有する統計の作成等を行う者が統計の作成等を行う場合の規定である。

新統計法の下でのミクロデータの利用については、旧統計制度においても利用されていた方法と同じ利用方法である第三十三条による調査票情報の利用、新統計制度における新たな仕組みである第三十四条による委託による統計の作成等（オーダーメード集計）及び第三十五条・三十六条による匿名データの作成・提供がある。これらのミクロデータの利用を二次利用と称している。

新たな仕組みのうち、オーダーメード集計とは、統計の作成等を希望する者が調査実施者に個別の委託集計を申出、その申出を受けて調査実施者が集計し、委託申出者に集計結果を提供する方式のミクロデータの利用である。委託申出者が直接調査票情報を利用しないので、秘密の保護が確実に保たれる。このように秘密の保護が担保されるので、高度の

公益性を満たさなくても、学術研究の発展に資する場合及び高等教育の発展に資する場合の一定程度の公益性が認められる場合に利用を認められている。公益性については、その利用目的によって判断される。

匿名データの提供は、調査票情報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匿名データを、一般の利用に供する方式でのミクロデータの利用である。

匿名データの作成については、調査実施機関が作成することになっており、基幹統計調査に係る匿名データは統計委員会に諮問し、答申を得なければならない。一般統計調査の匿名データについては、統計委員会に諮問し、答申を得る必要はないが、基幹統計調査と同様の匿名化措置を施さなければならないとされている。

匿名データについては、匿名化されることにより、秘密の保護が図られているので、第三十四条と同様に、高度の公益性を満たさなくても、学術研究の発展に資する場合、高等教育の発展に資する場合並びに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の一定程度の公益性が認められる場合に利用が認められている。

匿名データは匿名化措置を施し、秘密の保護が図られているとしても、調査票情報であることには変りがないので、その管理には十分な注意が必要である。そのため、匿名データの提供におけるガイドライン[3]では、利用者が申出する際には、管理するための条件が掲げられている。

1.4　二次利用の実績

新統計法において、各府省は基本計画の着実な推進が求められることになっており、各府省からの法の施行状況の報告を総務省政策統括官（統計基準担当）において、「統計法施行状況報告」として取りまとめている。

この「統計法施行状況報告」から二次利用の実績をみると、国の行政機関が、第三十三条第二号（第一号は公的機関からのものであるので省いた）に該当するとして、調査票情報を提供した件数は、平成 21 年度 54 件、22 年度 133 件となっている。

平成 21 年度に､国の行政機関がオーダーメード集計のサービスを供している統計調査は、６府省６調査となっている。そして、21 年度に、一般からの申出によって、オーダーメード集計の結果を提供した件数は、４件となっている。これらの申出は、すべて学術研究の発展に資する場合である。22 年度において、サービスを供している統計調査は７府省 20 調査、集計結果を提供した件数は 12 件となっている。これらの申出も、すべて学術研究の発展に資する場合である。

他方、平成 21 年度に、国の行政機関が匿名データの提供のサービスを供している統計調

[3] 総務省政策統括官（統計基準担当）は、第三十三条による調査票情報の提供、第三十四条によるオーダーメード集計、第三十六条による匿名データの提供における各府省共通のガイドラインを設定している。

査は１府省４調査となっている。そして、21 年度に、一般からの申出によって、匿名データを提供した件数は、20 件となっている。これらの申出のうち、学術研究の発展に資する場合は 18 件、高等教育の発展に資する場合は２件である。22 年度においては、サービスを供している統計調査は 21 年度と同様１府省４調査、提供した件数は 38 件、学術研究の発展に資する場合は 36 件、高等教育の発展に資する場合は２件である。

このように平成 21 年度及び 22 年度のオーダーメード集計、匿名データの提供の利用実績をみると、大きな期待をもって開始されたにもかかわらず、申出件数は数少ない状況であった。22 年度は、21 年度に比べて、利用者も、利用件数も着実に増えており、二次利用制度も認知されつつあると思われるが、制度を周知する広報だけでは、利用者の拡大はあまり望めない状況と考えられる。

また、統計委員会、匿名データ部会等では、新統計法の全面施行の準備をする段階で、いわゆるレプリカデータや統計教育・訓練用データの必要性が議論され、このようなデータが要望されていた。二次利用制度の広報活動を行っている中で、研究者や統計教育の関係者の方からも、このようなデータについての要望が多数あった。政府も二次利用の仕組みを考える際に、課題の一つとしてレプリカデータを取り上げ、「統計データの二次利用促進に関する研究会報告書」に記載しているが、その後、特にレプリカデータに関する具体的な議論はされなかった。

こうしたデータの必要性については、以前からも一部の研究者において指摘されていた。すなわち、日本学術会議学術基盤情報常置委員会（2005）は、リサンプリングとスワッピングによって作成し、研究者に自由に配布できるレプリカデータを提言していた。また、松田（2008）は、①個別の回答者とのリンケージが不可能な、②特別な手続きを必要としない、③大学院生にも自由に使えるレプリカデータの作成を計画していたことに言及している。

このレプリカデータの作成については、日本学術振興会の科学研究費補助金による「ミクロ統計データ活用研究会」（研究代表　松田芳郎、井出満、森博美）において、平成 17 年度に匿名化の度合いを高めた教育用の匿名標本データを作成し、大学院生に利用させることが計画されていた。しかし、計画に対する承認が下りず、レプリカデータの作成については断念している。

1.5　教育用擬似ミクロデータの開発

想定していたよりも低調であったミクロデータの利用を促進していくための１つの方策として、ミクロデータを用いた実証分析ができる人材を拡大または育成することが考えられる。未だミクロデータを利用していない研究者には、まずミクロデータを使って、データ特性等を理解してもらう、若手の研究者や学生には、ミクロデータを用いた実証分析の演習等を行ってもらう、そうしたことができる環境を整備することが肝要である。環境を整備することが、実証分析ができる人材を拡大・育成し、その結果として、ミクロデータ

を用いた実証研究を発展させ、学術研究水準を向上させることになると考えられる。

そのために、未だ利用経験のない研究者、若手の研究者、学生等に実際のミクロデータを自由に利用させることを実現しなければならない。匿名データの利用には、高等教育目的で利用できることになっているものの、統計法令に定められた利用目的や利用環境などの要件を満たさなければならない制約があり、かならずしも自由に使えるというわけではなく、多くの学生を相手にした大学での統計演習などで利用するには現実的に困難な場合がある。美添（2009）は、匿名化措置を強めた匿名データを一般用、教育用として作成し、それぞれ一般社会人、若手の研究者に自由に使わせることを提言しているが、新統計法に基づく統計制度では、そうした利用は想定されていないので、提言を実現できないのが現状である。

そこで、統計法令に縛られない、何の制約も受けない自由に使える教育用擬似ミクロデータの開発を計画することとした。計画した教育用擬似ミクロデータとは、本来の調査票情報である個票データを集計するなどして、個票データとの関連を断ち切った上で、その集計表に基づいて、ミクロデータの形式を持つ擬似的なデータ(以下「擬似ミクロデータ」と呼ぶ)を作成することとした。この擬似ミクロデータは、調査票情報と異なる。そして、このような擬似ミクロデータを、高等教育等の利用が主たるものとして、教育用擬似ミクロデータ[4]として称している。

改めて、実証分析の教育の枠組みを想定すると、次のような段階を経ていくことができると考えられる。まず、①統計調査の調査方法、調査票、標本設計、推定方法などを理解する。次に、②報告書等の既存の統計表データを用いた分析による実態把握を行う。その後、③教育用擬似ミクロデータを用いた演習によってミクロデータの特性、取扱い方、統計的分析手法を習得する。最後に、④匿名データを用いた実証分析する段階に進み、学術研究、学位論文作成を行う。このような教育の枠組みにおいては、教育用擬似ミクロデータは、実証分析のための環境を整備するために必要不可欠なツールであると考えられる。

２　擬似ミクロデータの基本的な考え方

2.1　調査票情報と匿名データ

現在、新統計法で提供されている統計データには、調査票情報、オーダーメード集計、匿名データの３つがある。どのデータも法令に規定されており、利用する上では制約がある。

新統計法における調査票情報とは、「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に記録されているものをいう。」であり、つまり調査客体の調査票ごとのデータであり、「特別の定めがある場合を除き、その行った統計調査の目的以外の目的

[4] 教育用擬似ミクロデータの名称について、開発当初においては擬似データと称していたが、外部への提供を考えた場合に、「擬似」よりも主たる目的である「教育」を冠し、自由に使えるデータと言う意味を込めて、教育用ミクロデータとした。その後、匿名データ等の調査票情報でないことを強調するために、「擬似」を追加し、教育用擬似ミクロデータとした。

のために、統計調査に係る調査票情報を自ら利用し、又は提供してはならない。」と規定されている。「特別の定めがある場合」以外には利用することはできない。匿名データは、調査票情報を加工したもので、調査票情報の一種と考えられている。統計法施行規則（平成十九年総務省令第百十二号）に定められた要件を満たさなければ利用できない。匿名化措置を施されていても、自由には利用できず、どれほど匿名化を強めたとしても、調査票情報には変わりないことを意味する。

したがって、自由に利用できるデータとして教育用のミクロデータを作成するためには、調査票情報から作成することはできないので、別の方法によって作成することが求められた。

2.2 教育用擬似ミクロデータ

調査票情報から作成したものは、調査票情報であることを踏まえて、教育用擬似ミクロデータでは、個票データから高次元の集計表を作成し、その高次元の集計表から個票データに近似したミクロデータを作成するという方法をとっている。集計表から作成するために、個票データでも、匿名データでもない擬似的なミクロデータと言える。それでいて、この教育用ミクロデータは、実証分析に利用した際に、我が国の実態を反映できるように、つまり個票データの分布にできる限り近似するように工夫して作成する方向で考えた。

このように集計表から作成する教育用擬似ミクロデータは、基本的に、①個票データの分布に近づけるなど、元の個票データに近似したデータであること、②量的属性の相関関係を保つなど、量的属性間の関係が整合的であること、③全国消費実態調査で言えば収入総額と支出総額が合致しているなど、調査特有のデータ構造を保持すること、④標本調査における集計用乗率を考慮すること、⑤データ量は元の個票データに合わせること、の考えの下で作成している。作成例としての全国消費実態調査における考慮点として、質的属性[5]については、集計表の作成における分類項目が該当し、その項目数は限られたものになり、量的属性については、分析上必要と思われる収入項目、支出項目を収録する。

このように教育用擬似ミクロデータの作成に当たっての基本的な考え方は、以上のとおりである。これを具体的にどのような統計的な手法を用いて開発したのかは、平成 16 年全国消費実態調査を例として、次節で述べる。

2.3 高次元の集計表

教育用擬似ミクロデータでは、個票データから高次元の集計表を作成し、その高次元の集計表から個票データに近似した擬似ミクロデータを作成するという方法をとっている。この教育用擬似ミクロデータの基になる高次元の集計表の基本的な考え方を述べることとする。

---

[5] 本稿での量的属性は集計表を作成する場合の分類項目、例えば世帯主の性別、年齢階級別であり、質的属性は集計表の表章項目、全国消費実態調査では収入と支出の項目が該当する。

統計作成機関が統計調査を実施し、その調査結果を結果表（又は統計表）として、報告書等で公表する。公表される結果表は、一般的に基本的と考えられる調査項目をクロスさせた集計表であり、低次元の集計表として作成されている。高次元の集計表を作成することは少なく、その公表時期に注目されている問題に対応するために、まれに高次元の集計表を作成することはありえる。

教育用擬似ミクロデータの作成においては、擬似ミクロデータとして収録する分類項目については、その調査項目をすべて使って高次元の集計表の作成をする必要がある。クロスさせなかった調査項目については、データとして収録することはできない。したがって、擬似ミクロデータに収録するべき項目は、基本的な調査項目に限定してもその数は多くならざるを得ない。利用する者にとっては、基本的な調査項目は最低限必要と考えられるからである。平成 16 年全国消費実態調査による擬似ミクロデータでは 14 項目を選定して、収録することにしたが、この 14 項目をクロスさせた集計表を作成し、それが擬似ミクロデータ作成の基になる集計表である。

擬似ミクロデータを作成する方法としては、高次元の集計表から作成する外に、調査票の調査項目ごとに確率分布を作成し、その確率分布から擬似ミクロデータを作成する方法なども考えられる。集計表から作成するという考えは、特別なものではなく、これまでにも１つの方法として指摘されている[6]。集計表は調査票情報から作成されるが、作成された集計表は調査票情報ではないことは統計作成関係者において認識されており、そうでなければ報告書等に掲載される統計表は調査票情報になり、公表できないことになる。今回、集計表から作成することとしたのは、調査票情報と集計表とが切り離され、別のものと理解されているからである。

一方で、ミクロデータに対する匿名化技法の１つであるミクロアグリゲーション[7]を適用して、教育用擬似ミクロデータを作成していると考えることもできる。しかしながら、匿名化されたミクロデータは調査票情報とみなされるため、匿名化技法によって擬似ミクロデータを作成したとしても、それは調査票情報から切り離すことができない。他方、ミクロアグリゲーションの方法論の観点から見れば、高次元の集計表の作成は、ミクロアグリゲーションの枠組みの中に位置付けられることから、教育用擬似ミクロデータの作成においてミクロアグリゲーションの方法論を適用することは可能である。詳細については「付論 ミクロアグリゲーションについて」を参照されたい。

３ 教育用擬似ミクロデータの作成方法

本節では、平成 16 年全国消費実態調査（以下「全消」という）を例に、教育用擬似ミクロデータの作成方法を述べることにしたい。先述のとおり、教育用擬似ミクロデータの基

6 寺崎（2000）、松田（1999）、美添（2009）を参照のこと。
7 ミクロアグリゲーションとは、ミクロデータをｋ個のレコードを有する同質なレコード群にグループ化した上で、そのレコードにおける個々の属性値を平均等の代表値に置き換えることである（伊藤（2008））。

本的な考え方は、集計表から個票データとは異なる擬似ミクロデータを作成することにある。集計表は、一般に表頭及び表側に用いられる分類事項と集計量として表される集計事項(度数等)から構成されるが、それは、擬似ミクロデータにおいてはそれぞれ質的属性と量的属性に対応する。そこで、教育用擬似ミクロデータの作成においては、最初に、擬似ミクロデータに含める質的属性と量的属性を選択する。質的属性と量的属性の選択については、集計表の分類事項によって分割された擬似ミクロデータのレコード群内における度数１又は２の割合が考慮される。次に、集計表の度数１又は２のセルに該当するレコードが出現しないような集計表を作成した上で、作成した集計表のセルごとに多変量正規乱数を発生させることによって量的属性値を作成する。さらに、量的属性については、乱数によって作成した量的属性値から合計値及び内訳値を作成し、収入と支出のバランス調整を行う。最後に、教育用擬似ミクロデータにおける集計用乗率を付与する。以下では、教育用擬似ミクロデータの具体的な作成手順を述べる。

3.1　量的属性と質的属性の選択

教育用擬似ミクロデータ作成における第 1 の手順は、全消に含まれるすべての属性の中から擬似ミクロデータに含める量的属性と質的属性を選び出すことである。教育用擬似ミクロデータは集計表に基づいて作成されることから、量的属性と質的属性の選択は、集計表における分類事項と集計事項の探索的な設定と捉えられる。分類事項の選択によっては、公表されている結果表として存在しない高次元の集計表を作成することができる。このような教育用擬似ミクロデータの基になる集計表は、「超高次元クロス集計表」として位置付けることが可能である(伊藤(2008))[8]。

教育用擬似ミクロデータの作成において検討した質的属性の組合せを一覧したものが表 3-1 である。表 3-1 に示されるように、質的属性は、性別、年齢、就業・非就業の別といった世帯主に関する調査事項と世帯区分、世帯人員階級といった世帯に関する調査事項に区分されている。本研究では、これらの質的属性の中から、12 属性、13 属性、14 属性、16 属性と 18 属性の 5 つのパターンに関する高次元の集計表が想定されている。

ところで、高次元の集計表においては、度数１又は２のセルが出現する可能性がある。度数１に該当するレコードは、個票データと１対１で対応することから、それは個票データとみなされるおそれがある。また、セルに度数１又は２が存在する場合、秘匿性の観点から、公表されている結果表においては、そのセルに「X」等の秘匿処理が施されていることが少なくない。このことから、度数１又は２については、集計表内のセルに出現しないような処理を施す必要がある[9]。

---

[8] 超高次元クロス集計表とは「個別データが有するすべての属性群を集計事項の対象とした上で作成されるｎ次元の多重クロス集計表」である(伊藤(2008, 17 頁))。詳細については、付論「ミクロアグリゲーションについて」を参照。

[9] 後述のとおり、教育用擬似ミクロデータ用の集計表を作成した後に，セルごとに多変量正規乱数が生成される。セルに含まれる度数が１又は２になる場合、多変量正規乱数の生成のために用いられる相関係数

集計表の分類事項の選択によって、集計表に出現する度数１又は２となるセルの数は異なる。表 3-2 は、検討した質的属性の数及び度数１、２と３以上に該当するレコード数及びセル数である。質的属性の数を 12 属性とした場合のセル数は約 9,500 であり、度数 1 に該当するレコードは 4,612 レコードである。それに対して、質的属性の数を 18 属性にした場合のセル数は約 50,000 であって、度数１に該当するレコードは、46,255 レコードに達している。このように、集計表における分類事項として用いられる質的属性が多くなるにつれて、集計表に含まれるセルの総数が増大するから[10]、度数１又は２が出現するセル数も多くなる。そこで、教育用擬似ミクロデータの作成においては、度数１又は２の出現数を考慮した上で、擬似ミクロデータに含める質的属性の数及び種類を選択した。さらに、質的属性の選択においては、(1)世帯主及び世帯に関する基本的な調査項目であること、(2)旧統計法の目的外使用（第 15 条第２項）で提供している属性や結果表で用いられている属性の中で使用頻度が高いことを考慮した。その結果、教育用擬似ミクロデータに含める質的属性として 14 属性を選別した。一方、量的属性については、本稿で用いている作成方法において提供することが可能な 184 属性を選択している。

したがって、教育用擬似ミクロデータで提供する属性は、個票データに含まれる全属性から選び出した、質的属性 14 属性、量的属性 184 属性、及び集計用乗率の全 199 属性である（詳細については、参考１「平成 16 年全国消費実態調査の教育用擬似ミクロデータ　質的属性及び量的属性一覧表」を参照されたい)。

表 3-1　教育用擬似ミクロデータの作成において検討した質的属性の一覧表

質　的　属　性		12 属性	16 属性	18 属性	13 属性	14 属性
世帯主事項	性別	○	○	○	○	○
	年齢（各歳）	○	○	○		
	年齢５歳階級					○
	就業･非就業の別	○	○	○	○	○
	企業区分	○	○	○	○	○
	企業規模	○	○	○	○	○
	産業分類	○	○	○	○	○
	職業分類	○	○	○		
	国公立･私立の別	○	○	○		
	学校の種類	○	○	○		
	専修学校	○	○	○		
	各種学校･塾など					

行列が計算できない。このような多変量正規乱数の発生の観点からも、度数１又は２が集計表内のセルに出現しないような処理が必要である。

[10] 厳密には、質的属性の数だけではなく、質的属性の選択肢の数にも影響される。

世帯事項	世帯区分	○	○	○	○	○
	世帯人員階級		○	○	○	○
	就業人員階級		○	○	○	○
	住居の建て方		○	○	○	○
	住宅の所有関係		○	○	○	○
	住居の構造			○	○	○
	建築時期･建築年			○		
	入居時期･入居年				○	○

表 3-2　検討した質的属性の数と度数１、２と３以上に該当するレコード数及びセル数

	12 属性	16 属性	18 属性	13 属性	14 属性
度数１	4,612	26,549	46,255	13,583	22,583
度数２	2,954	6,918	3,526	4,084	5,806
度数３以上	47,490	21,589	5,275	37,387	26,667
セル数	9,505	32,897	49,084	18,538	28,481

3.2　度数１又は２に該当するレコードの処理

教育用擬似ミクロデータの作成における第 2 の手順は、擬似ミクロデータ用の高次元の集計表のセルにおいて度数１又は２に該当するレコードに対して、度数３以上のセルとなるような処理を施すことである。

後述するように、教育用擬似ミクロデータの作成において多変量正規乱数を発生させる必要があることから、高次元の集計表の集計事項として算出されるのは、度数だけでなく、平均、分散・共分散である。なお、平均、分散・共分散は、集計用乗率を乗じることによって計算された重み付きの統計量である。この重み付きの平均、分散・共分散については、度数のように集計表上で作成することができないために、元の個票データから作成する必要がある。そのため、集計表の度数１又は２に該当するレコードにおいて、質的属性の一部の値を不詳に置き換えている。このような度数１又は２に該当するレコードの処理によって、多変量正規乱数の生成で用いる集計表の作成が可能になる。

不詳扱いにする質的属性については、あらかじめ質的属性の有用性を考慮した上で不詳を適用する属性の優先順位を決め、それに従って、該当する質的属性値に対して不詳の付与を行った。なお、世帯事項の「世帯区分」、「世帯人員階級」、「就業人員階級」、世帯主事項の「性別」については、基本的な項目であるため、不詳となる属性から除外した。さらに不詳に関する処理を行っても、度数３以上にならないデータは削除した。不詳を適用する属性の優先順位は次のとおりとなった。

（世帯事項）住居の構造、住居の建て方、住宅の所有関係、入居時期・入居年

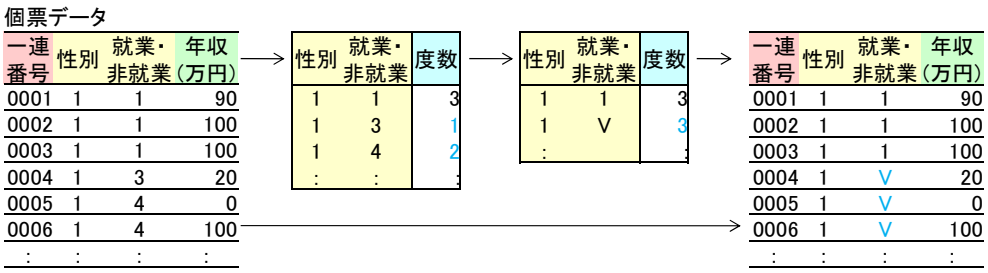
（世帯主事項）年齢５歳階級、就業・非就業の別、企業区分、企業規模、産業符号、
職業符号

度数１又は２の処理に関する具体的な方法は､図 3-1 に示されるような手順で行われる。図 3-1 で用いられている質的属性は、「性別」と「就業・非就業の別」、量的属性は「年間収入」である。

最初に、一連番号、性別、就業・非就業の別、年間収入の４つの属性を持つ個票データを用いて、分類事項として性別と就業・非就業の別を設定し、集計事項については度数を表示する集計表を作成する。作成した集計表をみると、性別が１で就業・非就業の別が３に対応する度数が１、性別が１で就業・非就業の別が４に該当する度数が２であって、度数３を下回っている。

次に、この度数１又は２のセルが度数３以上になるように、集計表の分類事項の「就業・非就業の別」３又は４を不詳（記号としてはＶとする）に置き換える。図 3-1 では、該当するセルが統合されて度数が３になっている。

図 3-1　度数１又は２に該当するデータの処理（イメージ）



3.3　高次元の集計表の作成

第 3 の手順は、度数１又は２を処理した個票データを用いて、分類事項を質的属性、集計事項を度数、基本的な量的属性の平均及び分散・共分散の高次元の集計表を作成することである。本稿では、量的属性の共分散を計算するために、相関係数行列を計測する。なお、相関係数行列の算出においては、選択された 14 の質的属性の中から、量的属性の値が０にならないように質的属性を限定した上で、質的属性によってグループ化されたレコードごとに相関係数行列を算出した。

集計表の作成方法は、以下の①～③の手順で行われる。

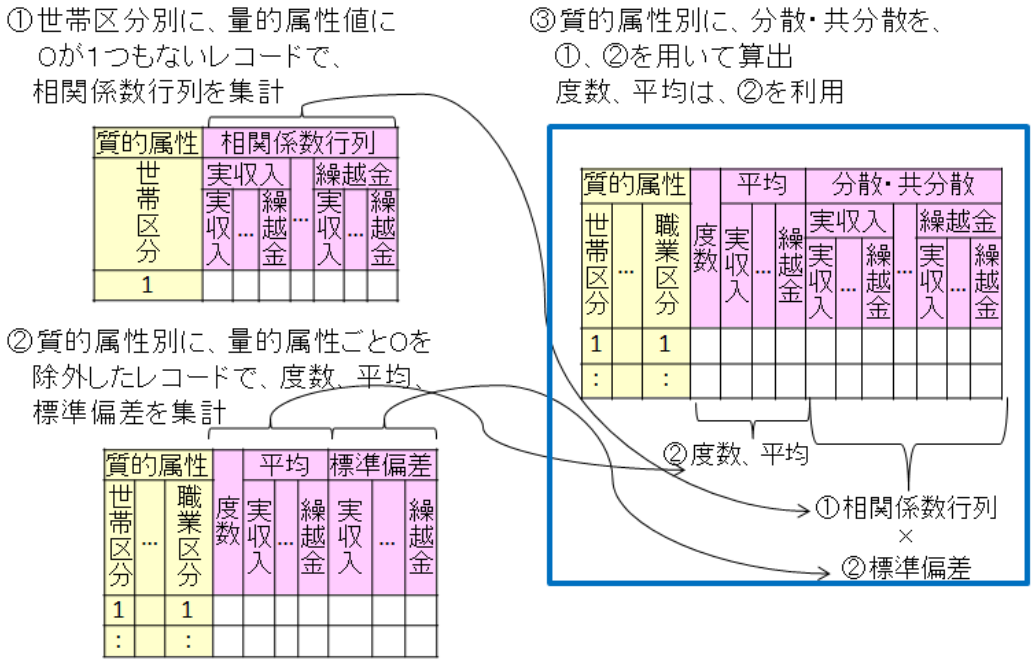
① 世帯区分別にグループ化した上で、量的属性値が０となるレコードが除外されたレコード群について相関係数行列を計算する。

② 質的属性ごとにさらにグループ化した上で度数を計測するだけでなく、量的属性ごとに属性値が０となるレコードを除外したレコード群について、平均、標準偏差を計算する。

③　質的属性ごとにグループ化されたレコード群別の分散・共分散に関しては、②で求めたレコード群別の標準偏差に、①で求めたそのレコード群に対応する相関係数行列を乗じることによって求められる。

なお、全消の量的属性値は、年間収入や消費支出などの収支金額で、必ずしも正規分布に従わないため、量的属性値が対数正規分布に従うことを仮定した。このため、作成する集計表も、度数１又は２を処理した個票データの量的属性値を対数変換し、各統計量を計算している。

図 3-2　教育用擬似ミクロデータを作成するための集計表のイメージ



3.4　多変量正規乱数の生成

第４の手順は､乱数を発生させることによって擬似ミクロデータを生成することである。乱数の発生方法については、最初に、①乱数を発生させずレコード群内の平均値を当てはめる方法(以下「平均法」という)、②単変量正規乱数法、③２変量正規乱数法の３つの方法が、年間収入と消費支出の２つの量的属性を用いた実験によって検討された。

①平均法

集計表の質的属性別の平均値を用いて、ミクロデータ形式でセル内の度数分のデータを作成する。

②単変量正規乱数法

個票データの量的属性について、セル内における各属性の値が属性ごとに正規分布に従うことを仮定し、各属性の平均及び標準偏差を用いて、個票データのばらつきを加味した正規乱数を生成する。

③２変量正規乱数法

個票データの量的属性のなかの２属性が関連性を持ちながら、正規分布（２変量正規分布）に従うことを仮定し、各属性の平均及び２属性間の相関係数を用いて、個票データにおける量的属性の分散を加味した正規乱数を生成する。

上記の３つの方法を用いて作成した擬似ミクロデータにおける年間収入と消費支出の相関係数が、図 3-3 に示されている。また、図 3-4 は年間収入と消費支出の散布図である。図 3-3 と図 3-4 を見ると、平均法における度数分布は個票データと比較して大きく異なることが確認できるが、２変量正規乱数法のように、２つの属性の関連性を考慮した形で乱数を発生させた場合には、擬似ミクロデータにおける相関係数が個票データのそれに近似していくことがわかった。

図 3-3　年間収入と消費支出の相関係数

個票データ

	年間収入	消費支出
年間収入	1.00	
消費支出	0.41	1.00

①平均法

	年間収入	消費支出
年間収入	1.00	
消費支出	0.68	1.00

②単変量正規乱数法

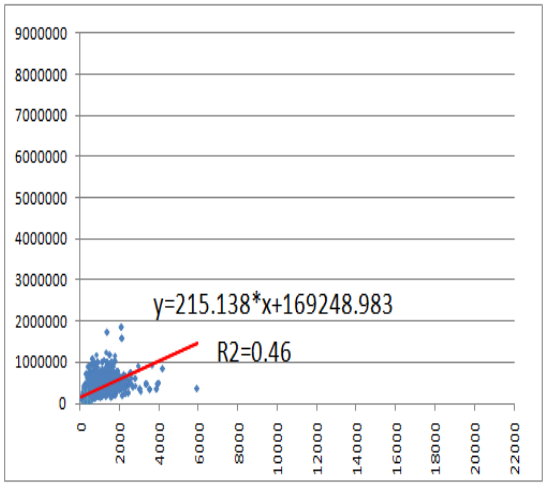
	年間収入	消費支出
年間収入	1.00	
消費支出	0.22	1.00

③２変量正規乱数法

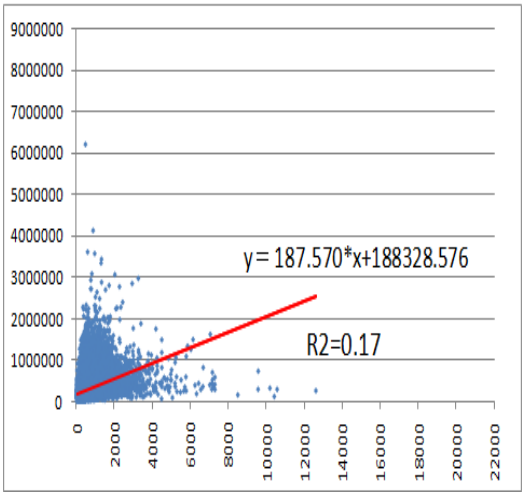
	年間収入	消費支出
年間収入	1.00	
消費支出	0.44	1.00

図 3-4　年間収入と消費支出の散布図および回帰式

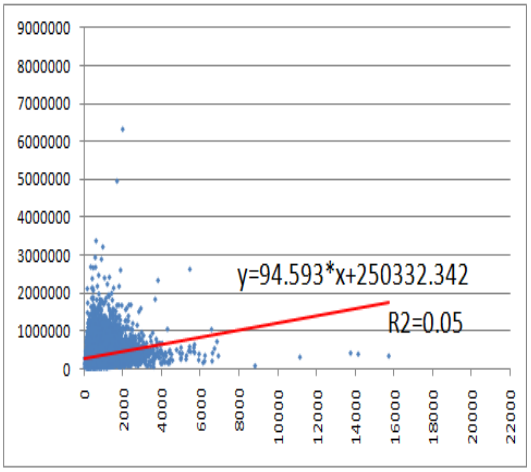
個票データ



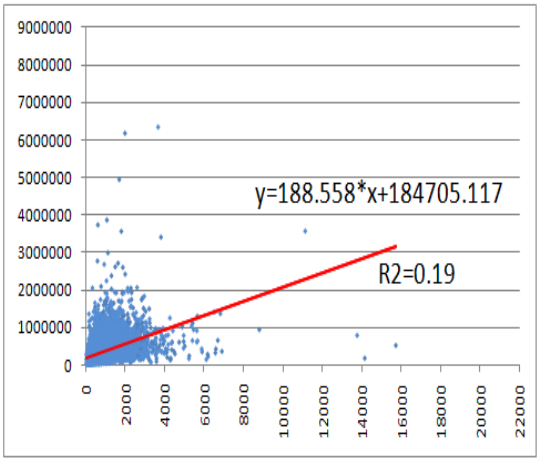
①平均法



②単変量正規乱数法



③２変量正規乱数法



本実験では、量的属性として年間収入と消費支出の２属性で行ってきたが、教育用擬似ミクロデータにおいては、数多くの量的属性を含むデータセットの作成が指向されていることから、本研究では、多変量の属性を用いた作成方法を検討した[11]。

④多変量正規乱数法

個別データの量的属性について、質的属性ごとにグループ化されたレコードにおける属性値が多変量正規分布に従うことを仮定し、各属性の平均値、分散と属性間の共分散を用いて、個票データのばらつきを加味した多変量正規乱数を生成する。

対象となる量的属性については属性値に０が含まれるために、対数変換を行うことができない。そこで、当該量的属性値の０の有無にかかわらず、量的属性値に１を加えてから対数変換し、教育用擬似ミクロデータ用に作成した集計表をもとに生成した乱数を実数に戻してから１を引く処理を行った。

このように、多変量正規乱数法を用いて作成することによって、①平均法、②単変量正規乱数法、及び③２変量正規乱数法と比較して、個票データにより近い分布特性を持つ擬似ミクロデータの作成が期待されるが、量的属性によっては、擬似ミクロデータにおける標準偏差が個票データのそれよりも大きくなる属性が存在する。その理由として、擬似ミクロデータの作成において、個票データの対数をとり、多変量対数正規乱数を生成した後、指数変換によって実数値に戻すような処理を行っているために、右に裾が長い分布となる可能性があることが考えられる。そこで、多変量正規分布の標準偏差に基づいて乱数の生成可能な区間を設定し、その閾値を超えないように乱数を生成する方法を採用した。具体的には、乱数の生成可能な区間として、４σ、３σ、２σ、1.５σ、２σ（住居のみ 1.５σ）の５つのケースを比較・検討した。その結果、２σ（住居のみ 1.５σ）に基づいて作成された擬

11 年間収入又は消費支出を量的属性間における基本的な相関関係として設定し、それ以外の量的属性については年間収入及び消費支出との関係性の中に逐次的に位置づける方法もあるが、年間収入又は消費支出以外の属性間の関連性も考慮するのが望ましいと判断し、多変量で行うこととした。

似ミクロデータが個票データに最も近似していることが明らかになったことから、教育用擬似ミクロデータの作成においては、乱数の生成可能な区間として、2σ（住居のみ 1.5σ）が採用された(表 3-3、表 3-4)。

表 3-3　平均値

平均	個別データ	擬似データ					
		処理なし	4σ	3σ	2σ	2.0σ（住居1.5σ）	1.5σ
年間収入	732.97	728.23	728.23	727.29	720.52	720.73	718.52
実収入	499,195.41	509,730.43	509,730.43	508,786.46	495,737.69	492,935.95	482,167.31
実収入以外の収入	387,220.77	389,486.03	389,486.03	388,257.58	368,413.33	366,435.27	355,314.59
繰入金	76,645.17	81,592.46	81,592.46	81,130.21	74,783.69	74,842.67	67,238.81
食料	72,673.55	72,232.36	72,232.36	72,179.87	71,835.72	71,846.35	71,389.01
住居	18,494.03	29,256.38	29,256.38	29,139.83	21,227.46	16,301.58	16,407.13
光熱・水道	19,491.00	19,392.20	19,392.20	19,377.32	19,312.13	19,337.12	19,202.21
家具・家事用品	9,884.48	9,848.84	9,848.84	9,820.71	9,123.66	9,159.42	8,477.86
被服及び履物	14,298.23	14,723.41	14,723.41	14,687.40	13,739.45	13,967.55	12,706.99
保健医療	11,812.02	12,045.80	12,045.80	11,957.76	11,420.18	11,447.34	10,454.25
交通・通信	50,981.65	49,065.64	49,065.64	48,698.34	47,876.20	47,910.35	45,092.62
教育	21,004.75	22,526.68	22,526.68	22,243.14	20,368.65	20,491.32	18,970.63
教養娯楽	31,499.91	31,643.16	31,643.16	31,598.72	30,012.21	29,940.22	28,933.31
その他の消費支出	86,379.13	86,597.92	86,597.92	86,441.82	82,650.49	82,673.92	79,091.73
非消費支出	75,896.15	77,234.99	77,234.99	77,120.01	74,491.44	74,545.73	73,558.05
実支出以外の支出	471,508.85	472,466.21	472,466.21	471,559.88	460,111.53	459,518.05	451,678.61
繰越金	79,137.60	83,775.36	83,775.36	83,349.45	76,765.58	77,074.92	68,758.33

表 3-4 標準偏差

標準偏差	個別データ	擬似データ					
		処理なし	4σ	3σ	2σ	2.0σ（住居1.5σ）	1.5σ
年間収入	355.49	362.89	362.89	361.16	327.20	329.66	311.86
実収入	313,118.37	457,774.91	457,774.91	455,730.60	405,187.48	321,234.75	225,193.81
実収入以外の収入	334,179.08	436,428.68	436,428.68	433,544.09	295,466.99	258,169.35	229,319.51
繰入金	85,271.65	143,984.99	143,984.99	136,181.10	90,979.43	91,537.30	63,621.91
食料	29,777.53	29,862.35	29,862.35	29,809.28	27,372.66	27,439.17	26,137.28
住居	53,558.37	581,041.68	581,041.68	580,704.40	361,910.19	52,120.42	44,956.43
光熱・水道	8,041.10	8,096.77	8,096.77	8,069.63	7,652.85	7,680.78	7,094.38
家具・家事用品	16,653.53	28,433.90	28,433.90	28,331.69	13,992.71	13,683.43	9,653.25
被服及び履物	19,391.95	27,703.69	27,703.69	27,624.41	20,481.71	38,291.78	14,776.43
保健医療	19,691.49	33,156.22	33,156.22	31,743.35	21,693.65	21,361.73	14,029.66
交通・通信	87,206.49	103,957.31	103,957.31	88,654.91	75,994.82	77,918.22	49,487.79
教育	49,867.14	112,888.83	112,888.83	105,143.03	54,158.48	62,989.77	46,688.69
教養娯楽	31,476.04	52,690.20	52,690.20	52,671.81	29,103.34	28,384.30	24,240.14
その他の消費支出	100,500.49	137,450.39	137,450.39	137,030.53	100,410.46	100,366.66	78,976.20
非消費支出	55,573.93	100,813.17	100,813.17	100,697.20	64,279.44	62,686.33	50,718.26
実支出以外の支出	405,185.34	421,922.06	421,922.06	420,826.23	396,401.89	378,661.31	294,557.16
繰越金	94,957.51	161,278.30	161,278.30	155,061.72	110,699.37	111,416.33	69,925.80

3.5　パターン別集計表を用いた量的属性値０の付与

先述の多変量正規乱数の生成によって作成した擬似ミクロデータには、個票データには含まれる量的属性値の０が存在しない。そこで、教育用擬似ミクロデータ作成における第５の手順として、個票データに近似するように、擬似ミクロデータに量的属性値０を新たに付与した[12]。具体的には、度数１又は２を処理した個票データを用いて、質的属性ごとにグ

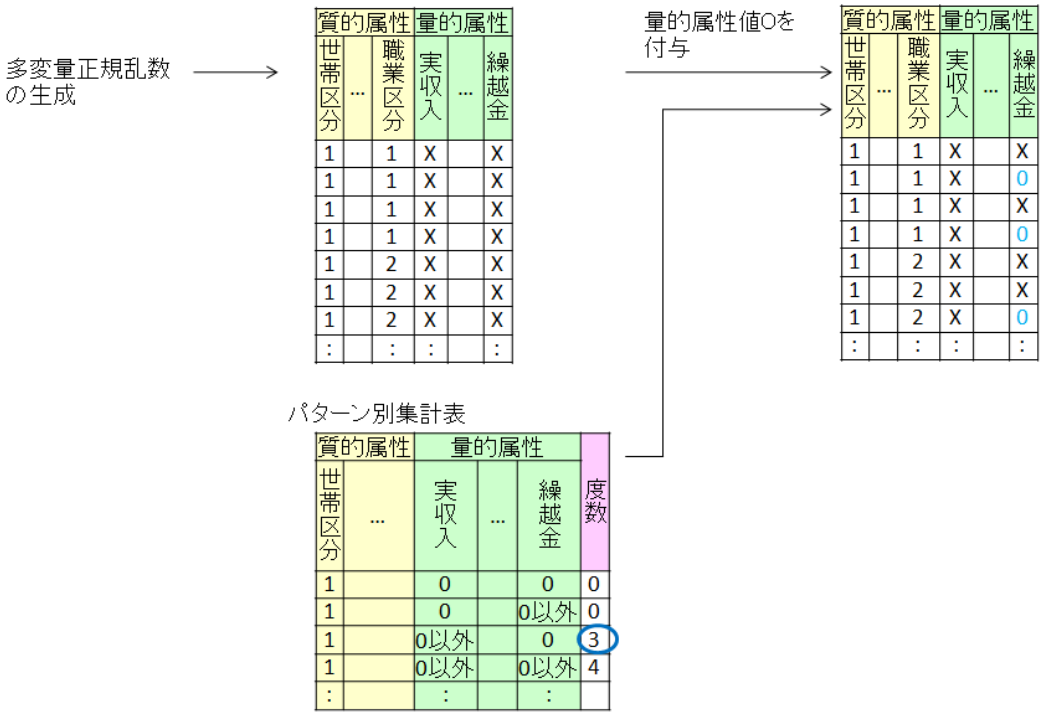
[12]　個票データと教育用擬似ミクロデータにおける分布のイメージについては、「参考２　教育用擬似ミク

ループ化した上で、そのグループ内で量的属性値が０か０以外のパターン別の集計表を作成し、そのパターンに従って、該当する量的属性に０を付与した。具体的な処理の方法は以下のとおりである(図 3-5)。

① 度数１又は２を処理した個票データを用いて、分類事項については世帯区分、実収入及び繰越金(実収入と繰越金については０と０以外の 2 区分)、集計事項については度数とした集計表を作成する。なお、実収入と繰越金の組み合わせに関しては、世帯区分別に、(1) ２属性とも０の場合、(2) ２属性のうちどちらか一方が０の場合、(3) ２属性とも０以外の場合の３つのパターンが考えられる

② 図 3-5 の集計では、世帯区分が１、実収入は０以外(「X」で表示)、繰越金が０に該当する度数は３である。したがって、多変量正規乱数の生成によって作成したレコードの中で、世帯区分が１に該当する７レコードのうち、３レコードの繰越金の値に０が付与される。なお、何番目のレコードの繰越金に０の値が付与されるかについては、乱数を１～７の範囲で３回発生させ（乱数の値が重複した場合、再度発生させる)、乱数の値に基づいて、処理が施される。

図 3-5 量的属性値０の付与



3.6 加法性と収支バランスの調整

第 6 の手順は、レコードレベルで加法性を保ち、収支バランスの調整を行うことである。

ロデータを作成するまでの分布イメージ」を参照されたい。

一般に、統計調査の調査項目には、総計と内訳が存在し、その加法性は保たれている。例えば、全消においては、消費支出の金額と十大費目分類[13]の金額の総計は一致している[14]。このような加法性に基づき、レコードごとに、量的属性の合計について対応する量的属性を合算し、作成した。

一方、全消では、収入（総額）と支出（総額）の値が一致するように収支のバランスが図られている。ゆえに、教育用擬似ミクロデータにおいても、収入と支出のバランスを調整した。具体的には、レコードレベルで支出総額に対する収入総額の割合を算出し、その割合を当該レコードの収入に関する各量的属性値に乗じる。

さらに、量的属性における内訳の値を求めるために、①度数１又は２を処理した個票データを用いて、質的属性別に量的属性値を 1 とした内訳の構成比を集計し、②集計した結果に基づき、収支バランスの調整を行った擬似ミクロデータを用いて、質的属性によってグループ化されたレコード群ごとに、量的属性値にその内訳の構成比を乗じることによって、内訳の値を付与する[15]。

3.7 集計用乗率の付与

第 7 の手順は、レコードに教育用擬似ミクロデータの集計用乗率を付与することである。標本調査では、結果数値の推定を行うため、個票データに集計用乗率が付与されている（推定は当該項目の値に集計用乗率を乗じる）。よって、擬似ミクロデータでも、①度数１又は２を処理した個票データを用いて、質的属性によってグループ化されたレコード群について集計用乗率の平均値を求め、②多変量正規乱数の生成によって作成した擬似ミクロデータの対応するレコード群ごとに集計用乗率の平均値を付与した。

4 教育用擬似ミクロデータの分布特性

本節では、作成された擬似ミクロデータと個票データの分布特性を比較することによって、教育用擬似ミクロデータの特徴を明らかにする。

表 4-1 は、個票データと教育用擬似ミクロデータについて、基本統計量のうち平均、標準偏差を比較したものである。平均値については、「住居」（１世帯当たり１か月間）における差が最も大きかった。個票データが 19,387.99 円であるのに対して、教育用擬似ミクロデータ 17,687.21 円であり、教育用擬似ミクロデータにおける数値のほうが 1,700.78 円（-9%）小さかった。

---

13 十大費目分類は、食料、住居、光熱・水道、家具・家事用品、被服及び履物、保健医療、交通・通信、教育、教養娯楽及びその他の消費支出の 10 分類である。

14 統計調査によっては、内訳の項目としては存在しないが、その値を合計には含める場合もある。

15 当初は、セル内のデータが、全て同じ量的属性値に対する内訳の構成比とならないように、構成比の集計表を消費支出の分位階級別で作成し、分位階級に該当するデータについては、該当分位階級の構成比を用いることとしていた。その後、開発上の都合からこうした処置を施していないため、実際のセル内のデータは画一的な構成比となっている。

表 4-1　個票データと教育用擬似ミクロデータの比較－基本統計量（平均、標準偏差）

	平均				標準偏差			
	個票データ	教育用擬似ミクロデータ	差		個票データ	教育用擬似ミクロデータ	差
年間収入	740.18	729.81	−0.01	年間収入	358.18	337.69	−0.06
収入総額	971,789.24	946,779.03	−0.03	収入総額	541,290.74	473,480.73	−0.13
実収入	502,133.73	497,655.92	−0.01	実収入	280,695.92	261,558.27	−0.07
実収入以外の収入	391,823.98	372,130.47	−0.05	実収入以外の収入	353,922.37	263,445.65	−0.26
繰入金	77,831.53	76,992.65	−0.01	繰入金	87,036.21	98,947.04	0.14
支出総額	971,789.24	946,779.03	−0.03	支出総額	541,290.74	473,480.73	−0.13
実支出	415,809.39	403,746.63	−0.03	実支出	224,419.69	219,290.60	−0.02
消費支出	339,199.37	328,139.70	−0.03	消費支出	194,501.15	192,447.21	−0.01
食料	73,738.54	72,883.42	−0.01	食料	30,149.02	28,064.49	−0.07
住居	19,387.99	17,687.21	−0.09	住居	52,962.36	60,587.32	0.14
光熱・水道	19,395.36	19,237.81	−0.01	光熱・水道	8,009.23	7,690.12	−0.04
家具・家事用品	9,783.81	9,204.04	−0.06	家具・家事用品	15,977.65	14,933.13	−0.07
被服及び履物	14,649.44	14,137.63	−0.03	被服及び履物	18,837.04	19,823.09	0.05
保健医療	11,936.01	11,366.36	−0.05	保健医療	19,763.39	19,284.07	−0.02
交通・通信	50,740.68	47,960.92	−0.05	交通・通信	85,021.69	84,654.38	0.00
教育	22,332.15	22,269.65	0.00	教育	51,989.72	64,157.45	0.23
教養娯楽	32,472.95	31,389.49	−0.03	教養娯楽	32,161.60	32,723.04	0.02
その他の消費支出	84,762.44	82,003.18	−0.03	その他の消費支出	95,898.83	102,040.97	0.06
非消費支出	76,610.02	75,606.93	−0.01	非消費支出	56,199.75	66,378.49	0.18
実支出以外の支出	475,947.80	464,318.09	−0.02	実支出以外の支出	394,805.29	334,227.09	−0.15
繰越金	80,032.04	78,714.31	−0.02	繰越金	96,421.45	118,055.82	0.22

注 1 教育用擬似ミクロデータの基になる集計表を作成する際に、個票データから一部のデータを削除している（「3.2」参照）ので、その集計表から作成した教育用擬似ミクロデータにも削除したデータが反映されていない。本表は、個票データと教育用擬似ミクロデータの分布特性の比較を目的として作成したものである。そのため、個票データ及び教育用擬似ミクロデータを用いた基本統計量の作成においても、削除したデータは含まれていない。したがって、個票データにおける結果数値が、公表されている結果表のそれと一致しない場合があることに留意されたい（表 4-2～表 4-5 についても同様）。

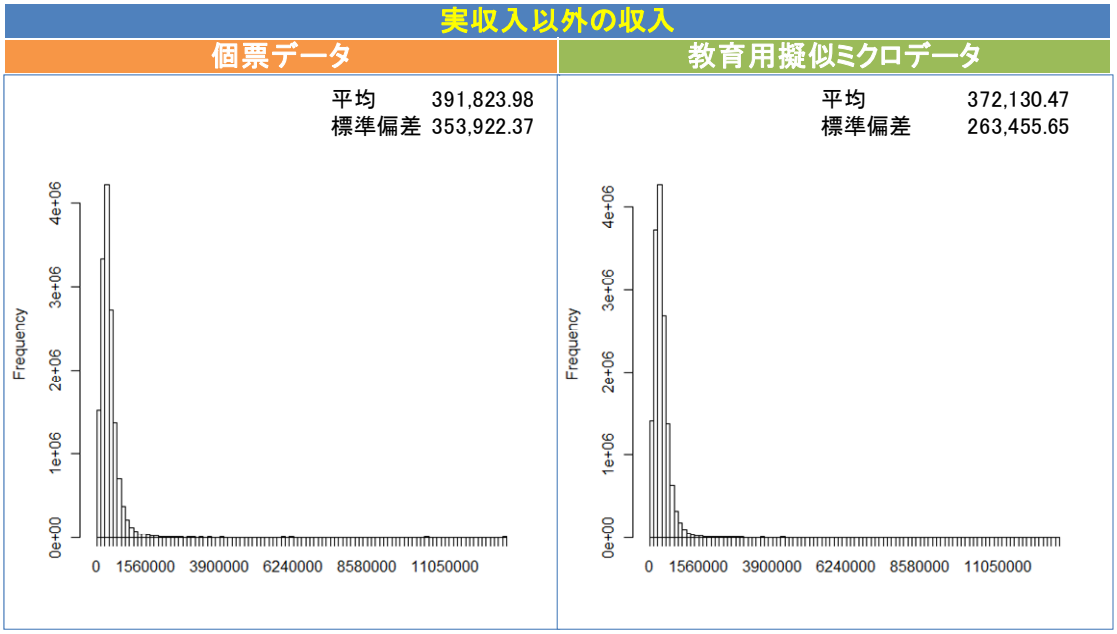
注 2 本表における差の計算方法はつぎのとおりである。

$$差=\frac{教育用擬似ミクロデータ-個票データ}{個票データ}$$

一方、標準偏差を見ると、「教育」においては、個票データが 51,989.72 円であるのに対して、教育用擬似ミクロデータが 64,157.45 円となっており、数値の差が 12,167.73 円（23%）となっている。この原因の一つは、個票データの標準偏差の値が大きいと、ばらつきが大きい乱数が発生する可能性があることから、教育用擬似ミクロデータの場合、基になる個票データの標準偏差と比較して相対的に標準偏差が大きなデータが作成される可能性があることが考えられる。なお、教育用擬似ミクロデータにおいては「その他の消費支出」や「交通・通信」における標準偏差の値そのものが大きいことが確認できるが、個票データと教育用擬似ミクロデータにおける標準偏差の差は小さいことがわかる。

図 4-1 は、個票データと教育用擬似ミクロデータとの標準偏差の差が最も大きな「実収入以外の収入」に関するヒストグラムを示している。図 4-1 を見ると、興味深いことに、２つのヒストグラムの分布は、近似しているように見える。

図 4-1　個票データと教育用擬似ミクロデータの比較－ヒストグラム



つぎに、表 4-2 と表 4-3 はそれぞれ、個票データと教育用擬似ミクロデータに含まれる量的属性間の相関係数行列を示している。個票データと教育用擬似ミクロデータにおいて、相関係数の符号が変化している量的属性があるが(例えば、住居と教養娯楽の相関係数等)、それについては、相関係数が 0 に近いことがその主な理由である。個票データと教育用擬似ミクロデータの相関係数の差が 1 番大きかった「年間収入」と「非消費支出」（1 世帯当たり 1 か月間）は、個票データが 0.70、教育用擬似ミクロデータが 0.50 と、個票データが 0.20 大きかったが、それぞれ正の相関関係である傾向は、変わらなかった。

表 4-2　個票データー相関係数

個票データ

	年間収入	収入総額	実収入	実収入以外の収入	繰入金	支出総額	実支出	消費支出	食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出	非消費支出	実支出以外の支出	繰越金
年間収入	1.00																				
収入総額	0.60	1.00																			
実収入	0.66	0.78	1.00																		
実収入以外の収入	0.35	0.85	0.36	1.00																	
繰入金	0.19	0.26	0.14	0.04	1.00																
支出総額	0.60	1.00	0.78	0.85	0.26	1.00															
実支出	0.60	0.73	0.56	0.63	0.17	0.73	1.00														
消費支出	0.49	0.66	0.45	0.61	0.16	0.66	0.97	1.00													
食料	0.47	0.42	0.37	0.31	0.17	0.42	0.52	0.50	1.00												
住居	-0.02	0.11	0.00	0.16	0.01	0.11	0.24	0.28	-0.03	1.00											
光熱・水道	0.32	0.24	0.22	0.16	0.11	0.24	0.28	0.27	0.44	-0.07	1.00										
家具・家事用品	0.15	0.25	0.12	0.26	0.09	0.25	0.26	0.27	0.17	0.07	0.10	1.00									
被服及び履物	0.30	0.30	0.24	0.24	0.10	0.30	0.39	0.38	0.29	0.02	0.12	0.16	1.00								
保健医療	0.11	0.16	0.10	0.15	0.07	0.16	0.24	0.25	0.15	0.01	0.07	0.08	0.09	1.00							
交通・通信	0.14	0.33	0.15	0.37	0.04	0.33	0.54	0.57	0.12	0.01	0.05	0.05	0.10	0.06	1.00						
教育	0.18	0.23	0.15	0.23	0.03	0.23	0.37	0.39	0.24	-0.03	0.19	0.02	0.09	0.04	0.07	1.00					
教養娯楽	0.32	0.35	0.27	0.30	0.12	0.35	0.44	0.42	0.32	0.02	0.10	0.15	0.26	0.10	0.10	0.09	1.00				
その他の消費支出	0.39	0.46	0.38	0.37	0.12	0.46	0.66	0.66	0.21	0.01	0.13	0.12	0.19	0.11	0.12	0.04	0.16	1.00			
非消費支出	0.70	0.63	0.70	0.38	0.12	0.63	0.62	0.43	0.35	-0.02	0.19	0.12	0.26	0.08	0.17	0.14	0.29	0.34	1.00		
実支出以外の支出	0.44	0.90	0.72	0.79	0.04	0.90	0.40	0.32	0.25	0.01	0.14	0.18	0.17	0.08	0.14	0.11	0.22	0.23	0.49	1.00	
繰越金	0.16	0.24	0.13	0.06	0.86	0.24	0.13	0.12	0.13	0.02	0.10	0.07	0.07	0.05	0.02	0.02	0.08	0.10	0.10	0.01	1.00

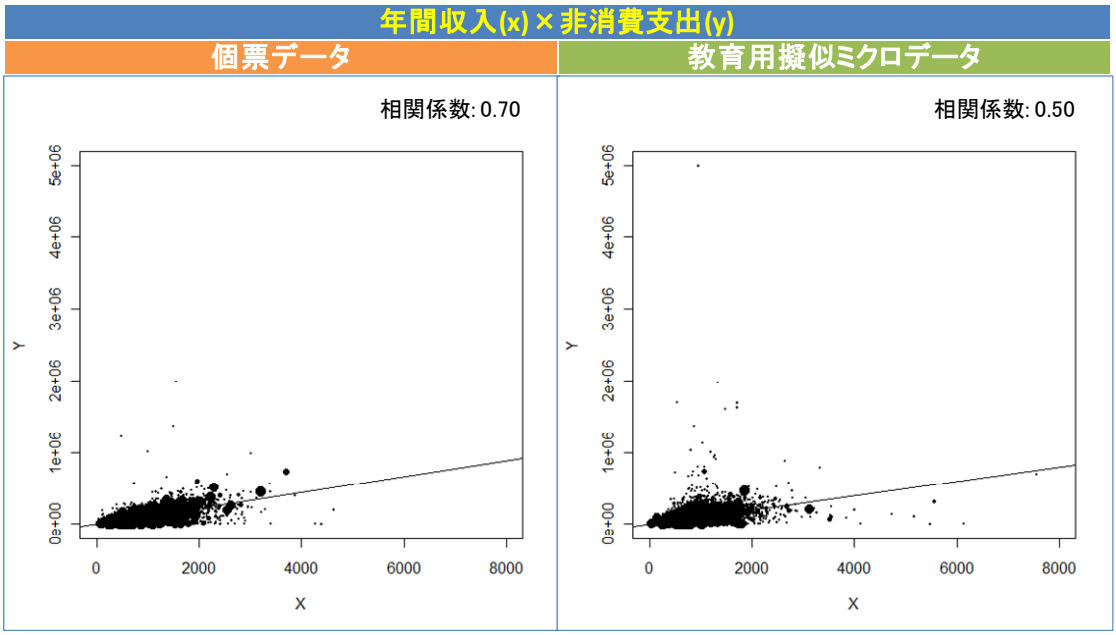
表 4-3　教育用擬似ミクロデーター相関係数

教育用擬似ミクロデータ

	年間収入	収入総額	実収入	実収入以外の収入	繰入金	支出総額	実支出	消費支出	食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出	非消費支出	実支出以外の支出	繰越金
年間収入	1.00																				
収入総額	0.58	1.00																			
実収入	0.63	0.85	1.00																		
実収入以外の収入	0.38	0.83	0.48	1.00																	
繰入金	0.12	0.32	0.15	0.05	1.00																
支出総額	0.58	1.00	0.85	0.83	0.32	1.00															
実支出	0.52	0.71	0.59	0.64	0.14	0.71	1.00														
消費支出	0.42	0.63	0.49	0.60	0.14	0.63	0.96	1.00													
食料	0.46	0.40	0.36	0.32	0.13	0.40	0.45	0.43	1.00												
住居	-0.05	0.08	0.04	0.09	0.03	0.08	0.24	0.28	-0.06	1.00											
光熱・水道	0.32	0.25	0.23	0.18	0.09	0.25	0.26	0.25	0.44	-0.07	1.00										
家具・家事用品	0.12	0.15	0.11	0.14	0.04	0.15	0.19	0.19	0.15	0.00	0.10	1.00									
被服及び履物	0.21	0.23	0.19	0.20	0.06	0.23	0.29	0.28	0.20	0.01	0.08	0.12	1.00								
保健医療	0.07	0.13	0.09	0.13	0.04	0.13	0.19	0.20	0.11	0.00	0.06	0.05	0.05	1.00							
交通・通信	0.12	0.30	0.17	0.35	0.04	0.30	0.50	0.54	0.10	-0.01	0.05	0.03	0.06	0.04	1.00						
教育	0.14	0.24	0.18	0.24	0.02	0.24	0.38	0.41	0.18	-0.02	0.16	0.01	0.04	0.02	0.04	1.00					
教養娯楽	0.26	0.30	0.24	0.28	0.06	0.30	0.35	0.34	0.26	-0.01	0.06	0.12	0.18	0.07	0.07	0.05	1.00				
その他の消費支出	0.33	0.44	0.38	0.37	0.11	0.44	0.63	0.65	0.17	-0.02	0.11	0.07	0.11	0.06	0.09	0.04	0.10	1.00			
非消費支出	0.50	0.50	0.52	0.35	0.07	0.50	0.53	0.26	0.24	-0.04	0.14	0.07	0.14	0.05	0.09	0.07	0.18	0.21	1.00		
実支出以外の支出	0.45	0.85	0.77	0.74	0.07	0.85	0.32	0.25	0.25	-0.05	0.15	0.08	0.13	0.05	0.09	0.09	0.18	0.18	0.35	1.00	
繰越金	0.10	0.28	0.14	0.05	0.82	0.28	0.07	0.07	0.09	0.00	0.08	0.03	0.03	0.02	0.01	0.00	0.02	0.06	0.04	0.00	1.00

さらに、図 4-2 は、個票データと教育用擬似ミクロデータで相関係数の差が最も大きかった「年間収入」と「非消費支出」の散布図を示したものである。教育用擬似ミクロデータの散布図には、左上及び右下に、外れ値のようなデータがみられる。これは、教育用擬似ミクロデータにおいては、個票データに存在する外れ値(特異値)よりもさらに外側に位置する外れ値が存在することを示している。各セルのデータによっては、対数変換した後の標準偏差の数値が大きくなり、乱数を発生したときに、個票データよりも大きく離れた値が出現したものと考えられる。しかしながら、全体的な分布は、個票データと教育用擬似ミクロデータは、近似しているように見える。

図 4-2　個票データと教育用擬似ミクロデータの比較－散布図



つぎに、表 4-4 は、個票データと教育用擬似ミクロデータについてクロス集計表で比較したものである。具体的には、表 4-4 は、「平成 16 年全国消費実態調査報告」の「第 13 表 世帯人員、年間収入階級別 1 世帯当たり 1 か月間の支出金額」を参考に、世帯人員、年間収入階級別 1 世帯当たり 1 か月間の支出の表を作成し、このうち、個票データと教育用擬似ミクロデータの平均値の差が 1 番大きかった世帯人員 2 人を表示している。「教育」について、年間収入階級 1,250～1,500 万円の階級については、個票データの教育費が 887 円であるのに対して、教育用擬似ミクロデータにおけるそれが 0 円となっている。また、年間収入階級 1,500 万円以上では、個票データにおける教育費が 1,142 円であるのに対して、教育用擬似ミクロデータにおける教育費が 10,052 円となっている。このように年間収入が高い階級において消費支出の内訳について大きな差がみられる。これは、擬似ミクロデータが高次元の集計表を基に作成されるから、高次元の集計表の中で度数の小さいセルにつ

いては、セルに含まれるレコードにおける属性値の標準偏差が相対的に大きくなり、それが、個票データと擬似ミクロデータにおけるクロス集計表の分布の違いに反映されていると思われる。

表 4-4　個票データと教育用擬似ミクロデータの比較－クロス集計表

世帯人員2人、年間収入階級別1世帯当たり1か月間の支出

		平均	年間収入階級（万円）									
			200 未満	200 ～ 300	300 ～ 400	400 ～ 500	500 ～ 600	600 ～ 800	800 ～1000	1000 ～1250	1250 ～1500	1500 以上
個票データ	食料	60,984	38,439	46,341	51,449	55,430	59,147	64,313	68,487	71,419	81,540	92,303
	住居	23,905	25,703	22,623	23,331	26,316	24,298	22,386	24,143	27,417	22,171	15,331
	光熱・水道	15,096	12,304	13,709	14,515	14,544	14,654	15,349	15,822	16,155	16,640	20,022
	家具・家事用品	9,286	4,398	5,549	7,483	7,368	8,698	10,255	12,056	10,972	13,492	14,796
	被服及び履物	13,654	5,904	6,825	7,955	8,821	11,513	14,737	17,283	22,484	26,940	37,843
	保健医療	11,134	5,501	8,175	9,444	10,006	11,713	11,314	13,163	12,805	14,661	16,623
	交通・通信	45,703	20,996	25,966	32,570	38,339	46,288	46,286	55,709	63,407	78,109	81,492
	教育	1,213	3,641	2,530	1,260	836	1,153	785	697	1,940	887	1,142
	教養娯楽	30,863	11,051	16,564	18,288	23,541	26,479	34,068	37,126	51,438	56,794	67,421
	その他の消費支出	89,539	27,955	38,910	51,071	60,584	76,484	97,967	120,000	153,095	163,964	208,602
		平均	年間収入階級（万円）									
			200 未満	200 ～ 300	300 ～ 400	400 ～ 500	500 ～ 600	600 ～ 800	800 ～1000	1000 ～1250	1250 ～1500	1500 以上
教育用擬似ミクロデータ	食料	60,342	40,970	44,783	50,789	55,577	59,450	62,807	68,987	72,671	77,010	87,092
	住居	22,241	23,467	25,408	22,322	24,755	24,034	19,762	25,761	15,321	15,644	9,864
	光熱・水道	15,062	12,565	13,113	14,135	14,410	15,090	15,190	16,060	16,440	17,481	19,122
	家具・家事用品	8,550	4,413	6,072	7,048	7,472	8,272	9,069	9,775	10,675	13,353	13,103
	被服及び履物	13,215	7,535	6,748	7,746	9,586	11,361	14,166	18,306	20,846	24,587	29,413
	保健医療	10,386	7,824	7,702	8,618	9,970	9,575	10,846	12,150	12,488	13,335	13,873
	交通・通信	43,772	23,271	24,048	34,239	35,225	44,714	46,207	53,821	61,657	59,127	79,587
	教育	1,337	4,640	4,141	1,196	1,489	830	744	398	634	0	10,052
	教養娯楽	30,194	11,654	15,620	19,395	23,153	28,587	32,610	40,740	43,390	50,752	60,672
	その他の消費支出	85,264	26,578	38,004	52,043	59,657	77,723	90,088	116,778	142,395	160,150	216,446

最後に、個票データと教育用擬似ミクロデータについて、回帰分析を行った結果を紹介したい。表 4-5 は、「平成 16 年全国消費実態調査報告　第９巻」の「ＸⅢ　弾力係数について（第 35 表～第 38 表)」を参考に、以下の回帰式により、計算した。

用途項目の１か月当たり支出金額(円)＝f（消費支出(円)）

回帰分析の結果から、個票データと教育用擬似ミクロデータが近似的な傾向にあることがわかった。

これまでの分析結果から、作成した全消の教育用擬似ミクロデータは、属性によっては、個票データと比べて散らばりが大きい場合もあるが、全般的には個票データにほぼ近似しているとみなすことができる。ゆえに、実際に教育用擬似ミクロデータを用いて、実証的な分析を行った場合でも、実態に即した結果が得られると考えている。

表 4-5　個票データと教育用擬似ミクロデータの比較－回帰

用途項目の1か月当たり支出金額(円)=f(消費支出(円))										
		消費支出			定数			決定係数	調整済決定係数	F値
		係数	標準誤差	P値	係数	標準誤差	P値			
個票データ	食料	0.07681	0.00075	0.000	47,683.0	294.2	0.000	0.2456	0.2456	10,425
	住居	0.07649	0.00146	0.000	-6,558.3	571.0	0.000	0.0789	0.0789	2,744
	光熱・水道	0.01123	0.00022	0.000	15,585.8	86.6	0.000	0.0744	0.0744	2,574
	家具・家事用品	0.02205	0.00044	0.000	2,305.1	172.9	0.000	0.0720	0.0720	2,486
	被服及び履物	0.03669	0.00050	0.000	2,202.6	195.8	0.000	0.1436	0.1435	5,368
	保健医療	0.02551	0.00055	0.000	3,284.6	214.9	0.000	0.0630	0.0630	2,153
	交通・通信	0.25071	0.00200	0.000	-34,299.9	782.4	0.000	0.3289	0.3289	15,699
	教育	0.10323	0.00138	0.000	-12,682.5	538.7	0.000	0.1491	0.1491	5,613
	教養娯楽	0.06986	0.00084	0.000	8,775.9	327.5	0.000	0.1785	0.1785	6,959
	その他の消費支出	0.32741	0.00206	0.000	-26,296.4	805.5	0.000	0.4410	0.4410	25,262
		消費支出			定数			決定係数	調整済決定係数	F値
		係数	標準誤差	P値	係数	標準誤差	P値			
教育用擬似ミクロデータ	食料	0.06232	0.00074	0.000	52,433.5	280.3	0.000	0.1826	0.1826	7,156
	住居	0.08949	0.00169	0.000	-11,679.6	641.7	0.000	0.0808	0.0808	2,815
	光熱・水道	0.01002	0.00022	0.000	15,951.0	82.2	0.000	0.0628	0.0628	2,147
	家具・家事用品	0.01509	0.00043	0.000	4,251.4	161.8	0.000	0.0378	0.0378	1,259
	被服及び履物	0.02908	0.00055	0.000	4,594.6	210.1	0.000	0.0797	0.0797	2,774
	保健医療	0.01958	0.00055	0.000	4,940.7	208.9	0.000	0.0382	0.0382	1,271
	交通・通信	0.23652	0.00207	0.000	-29,652.4	788.4	0.000	0.2891	0.2891	13,023
	教育	0.13728	0.00170	0.000	-22,777.2	645.8	0.000	0.1696	0.1695	6,539
	教養娯楽	0.05808	0.00089	0.000	12,331.1	339.7	0.000	0.1167	0.1166	4,230
	その他の消費支出	0.34253	0.00226	0.000	-30,392.9	860.4	0.000	0.4173	0.4173	22,934

５　教育用擬似ミクロデータの試行提供の現状とアンケートの結果

5.1　試行提供の現状

独立行政法人統計センターでは、本研究によって作成した教育用擬似ミクロデータの試行提供(以下「試行提供」という。)を平成 23 年８月 25 日に開始した。提供している教育用擬似ミクロデータは、以下のとおりである。

・平成 16 年全国消費実態調査　二人以上の勤労者世帯

・レコード数：約 32,000 レコード

・データ形式：CSV 形式

（教育用擬似ミクロデータのイメージ）

世帯情報		世帯員情報			集計用乗率	年間収入	収入		支出			
		世帯主										
世帯区分	…	…	年齢	…			収入総額	…	支出総額	…	消費支出	…
3	…	…	50	…	27	5000	790000	…	790000	…	280000	…

試行提供を開始してから 10 か月が経った平成 24 年６月末現在の実績は、提供件数が 51 件であり、その内訳は、大学教員 20 名、研究員１名、大学院生１名、学部生 29 名となっている。大学教員が授業等を行った際の教育用擬似ミクロデータの利用者数は、大学院生約 30 人、学部生約 800 人であった。提供した件数や利用した学生数の評価については、様々な見方ができると思われるが、ミクロデータを用いた講義や演習は、通常の講義と違って、パーソナルコンピュータを利用する場合が多く、一度に多人数の学生を相手に指導することが困難なことを考えれば、必ずしも少ない人数ではないと考えている。また、教える側の教員にとっても、講義や演習のために新たな教材を用意し、指導内容も検討しなければならないので、一気に利用が広がる性格のものではないと思われる。従って、ミクロデータを用いた講義や演習は徐々に拡大し、提供件数もそれに伴って増えていくものと考えている。

試行提供は、大学研究者、教育者等の関係者に教育用擬似ミクロデータの試行的な提供を行うことによって、利用に関する意見や要望等の情報を収集するだけでなく、作成方法の改善や別の統計調査への適用、さらには今後における開発・提供の可能性について検討することを目指している。そのため、教育用擬似ミクロデータの更なる検討を行うための参考資料とするために、教育用擬似ミクロデータを利用した大学教員等に対してアンケートを実施した(アンケートの内容については、「参考 3 教育用擬似ミクロデータの試行提供利用者アンケートのお願い」を参照されたい)。試行提供を開始後、10 か月が経過した平成 24 年 6 月末までに提供した 51 件について、アンケートを実施した。アンケートの回答件数は 17 件であった。回答された 17 件のすべてが大学等の高等教育での利用である。

5.2　アンケート結果

教育用擬似ミクロデータの試行提供の利用者アンケートに関する結果は、参考 4 に示されている。アンケートの結果を見ると、利用者のほとんどが、大学等の高等教育用に教育用擬似ミクロデータを用いていることがわかる。また、後述のように、教育用擬似ミクロデータの試行提供全般についての様々な要望があることから、全般的に教育用の教育用擬似ミクロデータの提供に対する期待の高さがうかがえる。その一方で、まだ、高等教育用として提供して時間も経っていないこともあることから、教育用擬似ミクロデータのデータ構造については特段の意見はなく、ファイル形式や符号表等のデータの提供方法や提供内容に関する意見が少なくなかった。それらの意見等を要約すると、次のようであった。

①ファイル形式について

提供した CSV 形式で問題がなかったとする意見が多いものの、SPSS や R といった統計解析用ソフトウェアに対応した提供や Excel ファイル形式での提供に関する要望も見られた。

②符号表及びレイアウトについて

符号表やレイアウトについては特に問題なしと回答した利用者は多いものの、より使

いやすい符号表を提供してほしいという要望もあった。具体的には、全消においては「収支項目分類表」に関する情報と符号表を併せて提供してほしいといった意見が見られた。符号表・レイアウトの利便性については、提供の仕方を工夫することによって、利便性をさらに高める余地があると考えられる。

③教育用擬似ミクロデータに含まれるレコード数や属性について

教育用擬似ミクロデータに含まれるレコード数や属性については、利用者から様々な意見が出された。具体的には、レコード数が少ない簡易的な擬似ミクロデータの提供や、地域属性等のより詳細な属性を含む擬似ミクロデータの提供、勤労者世帯以外のデータ、単身世帯データの提供等の要望があった。このことから、教育用擬似ミクロデータの提供形態についてさらなる検討を行う必要があると考える。

④その他

上記の意見の他にも、つぎのような様々な意見・要望が見られた。

- 全国消費実態調査以外の世帯・人口系の統計調査に関する教育用擬似ミクロデータを作成してほしい。
- 事業所・企業系のデータに関する教育用擬似ミクロデータが提供されれば、大きなニーズが見込まれると思われる。
- 教育用擬似ミクロデータの作成方法について、わかりやすい資料を提供してほしい。
- 基本統計量などの算出方法に関する詳細な情報を提供してほしい。
- 教育用擬似ミクロデータの利用例をホームページで紹介してほしい。
- 教育用擬似ミクロデータによる分析結果の検証システム（送付プログラムによる実行結果の返送）の提供（具体的にはオンラインよる提供）してほしい。
- 教育用擬似ミクロデータを活用した教材開発のプロジェクト及びそれに関する事業を立ち上げるだけでなく、そのプロジェクトの成果としての教材を刊行してほしい。

これらの意見については、今後教育用擬似ミクロデータの提供を議論する上で貴重な参考資料と考えている。その一方で、教育用擬似ミクロデータの利用においては、教育用擬似ミクロデータが個票データではないために、教育用擬似ミクロデータの分布特性が、個票データの分布に対して大きく異なるケースがあることを利用者に理解させることも必要かと考える。

６ おわりに―教育用擬似ミクロデータの作成に関する今後の課題

本稿では、最初に、擬似ミクロデータの作成に関する基本的な考え方を論じた上で、全消の個票データを用いた教育用擬似ミクロデータの作成を行った。つぎに、教育用擬似ミクロデータと個票データの分布特性を比較することによって、高等教育における授業・演習用として教育用擬似ミクロデータを用いることが有用であることを明らかにした。さらに、試行提供の現状とアンケート結果から明らかになった課題についても紹介した。

教育用擬似ミクロデータについては、これまでは試行提供という形で、既に研究者や学

生に提供してきた。アンケートによって利用実態が把握できたことによって、擬似ミクロデータの作成についてさらに検討すべき課題が存在することがわかった。また、教育用擬似ミクロデータのさらなる利用拡大を図るための方策も併せて検討する必要があると思われる。最後に、今後の課題について述べることにしたい。

6.1 当面の課題

当面の課題としては、擬似ミクロデータの作成において明らかになった問題点として、①多変量対数正規乱数を発生させた場合に、実際の個票データでは出現していない範囲を超えた数値が作成されること、②高次元の集計表の中で度数が少ないセルのデータについては、個票データの分布との乖離が大きくなること、③セル内の度数に対応するレコードの属性値群のすべてが、基本的な量的属性値の内訳に関して同一の構成比とならないように、レコードによって構成比を変えること等がある。これらについては今後の検討課題だと考える。一方、提供する全消のデータセットについては、①勤労者世帯のみで勤労者世帯以外の世帯が含まれていないこと、②二人以上の世帯のみで、単身者世帯が含まれていないことが問題点として指摘できることから、これらの課題に関する検討を進める必要があろう。また、利用者の裾野を広げるためにも、教育用擬似ミクロデータの利便性を高める必要があることから、①複数のファイル形式による提供、②簡易的なミクロデータを含む様々なタイプのミクロデータの提供、③データセットにおける変数名の付与等については、対処したいと考えている。

6.2 将来の課題

将来の課題としては、就業構造基本調査の教育用擬似ミクロデータの開発を考えている。全国消費実態調査と異なり、就業構造基本調査では、レコードに含まれる量的属性は少なく、その多くが質的属性である、作成方法についても全消の教育用擬似ミクロデータの作成方法をそのまま適用するのは難しいことから、作成方法を再検討する必要がある。一方、匿名データが世帯系の調査のみで事業所・企業系の調査では匿名データが提供されていないので、事業所・企業系の調査のミクロデータを利用する機会が乏しい。そうしたことを考えると、擬似データといえども、教育用として擬似ミクロデータを開発する意義は高いと考えるので、その次は、事業所・企業系の調査に関する教育用擬似ミクロデータの開発を目指したい。

参考１　平成 16 年全国消費実態調査の教育用擬似ミクロデータ　質的属性及び量的属性一覧表

質的属性

【世帯事項】
世帯区分（勤労，勤労以外，無職）
世帯人員階級（2 人～）
就業人員階級（1 人～、不詳）
住居の構造（木造，防火木造，鉄骨・鉄筋コンクリート造，その他，不詳）
住居の建て方（一戸建，長屋建，共同住宅（4 区分），その他，不詳）
住居の所有関係（持家（2 区分)），民営賃貸住宅（2 区分），県市区町村営賃貸住宅，都市再生機構・公社等賃貸住宅，社宅・公務員住宅，貸間，寮・寄宿舎，不詳）
入居時期・入居年（持家以外：昭和 63 年以前，平成元～16 年，不詳）
【世帯主事項】
性別（男，女）
年齢５歳階級（24 歳未満～75 歳以上，不詳）
就業・非就業の別（就業，非就業，不詳）
企業区分（就業のみ：民営，自営，官公，不詳）
企業規模（非就業・官公除く：1-4 人～1000 人以上，不詳）
産業分類（農業～その他，不詳）
職業分類（常用労務作業者～無職，不詳）

量的属性

【収入事項】		他の経常収入
年間収入		財産収入
収入総額		社会保障給付
実収入		公的年金給付
経常収入		他の社会保障給付
勤め先収入		仕送り金
事業・内職収入		特別収入
農林漁業収入		受贈金
家賃収入		その他
他の事業収入		実収入以外の収入
内職収入		預貯金引出
本業以外の勤め先・事業・内職収入		保険取金

個人・企業年金保険取金		乾物・海藻
他の保険取金		大豆加工品
有価証券売却		他の野菜・海藻加工品
株式売却		果物
他の有価証券売却		生鮮果物
土地家屋借入金		果物加工品
他の借入金		油脂・調味料
分割払・一括払購入借入金		油脂
財産売却		調味料
その他		菓子類
繰入金		調理食品
【支出事項】		主食的調理食品
支出総額		他の調理食品
実支出		飲料
消費支出		茶類
食料		コーヒー・ココア
穀類		他の飲料
米		酒類
パン		外食
めん類		一般外食
他の穀類		学校給食
魚介類		賄い費
生鮮魚介		住居
塩干魚介		家賃地代
魚肉練製品		設備修繕・維持
他の魚介加工品		設備材料
肉類		工事その他のサービス
生鮮肉		光熱・水道
加工肉		電気代
乳卵類		ガス代
牛乳		他の光熱
乳製品		上下水道料
卵		家具・家事用品
野菜・海藻		家庭用耐久財
生鮮野菜		家事用耐久財

冷暖房器具		自動車等維持
一般家具		通信
室内設備・装飾品		教育
寝具類		授業料等
家事雑貨		教科書・学習参考教材
家事用消耗品		補習教育
家事サービス		教養娯楽
被服及び履物		教養娯楽用耐久財
和服		教養娯楽用品
洋服		書籍・他の印刷物
男子用洋服		教養娯楽サービス
女子用洋服		宿泊料
子供用洋服		パック旅行費
シャツ・セーター類		月謝類
男子用シャツ・セーター類		他の教養娯楽サービス
女子用シャツ・セーター類		その他の消費支出
子供用シャツ・セーター類		諸雑費
下着類		理美容サービス
男子用下着類		理美容用品
女子用下着類		身の回り用品
子供用下着類		たばこ
生地・糸類		その他の諸雑費
他の被服		こづかい（使途不明）
履物類		交際費
被服関連サービス		食料
保健医療		家具・家事用品
医薬品		被服及び履物
健康保持用摂取品		教養娯楽
保険医療用品・器具		他の物品サービス
保険医療サービス		贈与金
交通・通信		他の交際費
交通		仕送り金
自動車等関係費		非消費支出
自動車購入費		直接税
自転車購入費		勤労所得税

個人住民税		個人・企業保険掛金
他の税		他の保険掛金
社会保険料		有価証券購入
公的年金保険料		株式購入
健康保険料		他の有価証券購入
介護保険料		土地家屋借金返済
他の社会保険料		他の借金返済
他の非消費支出		分割払・一括払購入借入金返済
実支出以外の支出		財産購入
預貯金		その他
保険掛金		繰越金

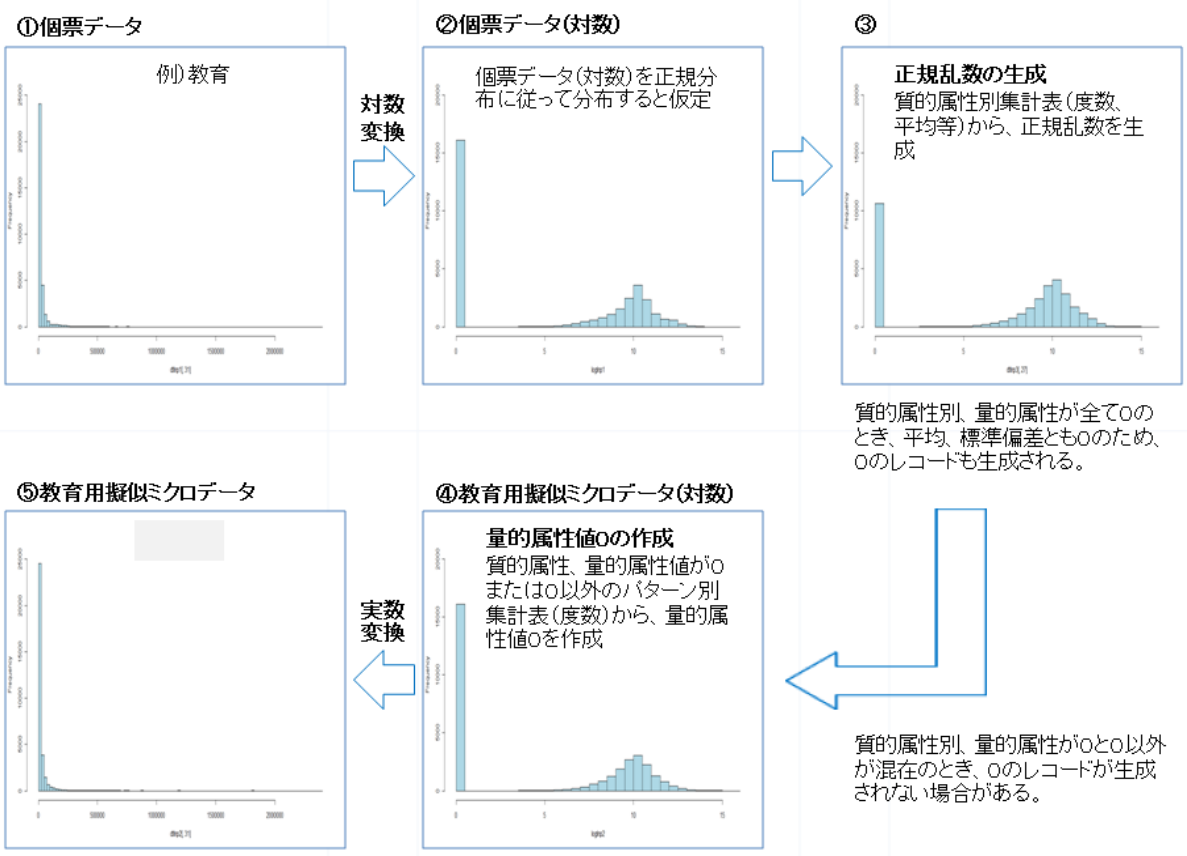
集計用乗率

集計用乗率

参考２　教育用擬似ミクロデータを作成するまでの分布イメージ

以下の参考図は、個票データと教育用擬似ミクロデータの分布特性に関するイメージ図を用いて、教育用擬似ミクロデータの作成過程を図示したものである。説明を単純化するために、全消における「教育」を例に、教育用擬似ミクロデータの作成に至るプロセスを示している。

参考図　個票データと教育用擬似ミクロデータの分布特性に関するイメージ



イメージ図の解説

(1) 個票データにおける教育費のヒストグラムである（図①）。

(2) 個票データを対数変換した場合のヒストグラムである。分布の全体を見ると正規分布とは言いがたい。そこで、値が０であるレコードと０以外のレコードに類別すると、０以外のレコードを対象にした分布は、正規分布に近い形のようにみえることから、対数変換した分布は、正規分布に従っていると仮定する（図②）。

(3) 対数変換した属性値について求めた平均値等に基づいて正規乱数を生成すると、図③のような分布となる。値が０となるレコードが生成されている。その理由としては、同一の質的属性値をもつレコードにおいて、量的属性値が全て０である場合には、平均と

標準偏差のいずれも０であることから、０のレコードが生成されることが指摘される。一方で、同一の質的属性値をもつレコードにおいて、量的属性値が０又は０以外のレコードが混在している場合は、次の(4)の方法で処理する。

(4) 質的属性別、量的属性０又は０以外のパターン別の度数を集計し、その度数に従って量的属性値０が該当するレコードに追加的に付与される。これらの新たに付与された量的属性値０を含めた度数分布は、図④のようになる。

(5) 図④の分布を実数に変換すると図⑤のような分布となる。図⑤の分布は、図①の個票データの分布に近似していることがわかる。このような方法で教育用擬似ミクロデータは作成されている。

参考３　教育用擬似ミクロデータの試行提供利用者アンケートのお願い

このたびは、『教育用擬似ミクロデータ』試行提供について、申出いただき、ありがとうございました。今後の教育用擬似ミクロデータについて、さらなる検討を行うため、「教育用擬似ミクロデータ利用者アンケート」を実施しております。ご多忙のところ、誠に申し訳ありませんが、アンケートの趣旨をご理解いただき、各質問事項に該当する回答を選択し、統計センターに返送くださいますよう、よろしくお願いいたします。

なお、各質問事項への回答は、□を黒く塗りつぶしてください。（例：■　問題なかった）

１　教育用擬似ミクロデータのファイル形式について

□　CSV 形式に問題なかった

□　不都合があった

（理由：　　）

□　どちらともいえない

２　符号表及びレイアウトについて

□　問題なかった

□　不都合があった

（理由：　　）

□　どちらともいえない

３　統計センターホームページ及びパンフレットについて

□　わかりやすかった

□　わかりにくかった

（理由：　　）

□　どちらともいえない

４　教育用擬似ミクロデータの構成について

(1)　レコード数について

□　ちょうどよい

□　多い　（希望するレコード数：　　）

□　少ない（希望するレコード数：　　）

(2)　属性（世帯人員、世帯主の性別、年間収入、消費支出など）について

□　ちょうどよい

□　多い　（掲載を希望しない属性：　　）

□　少ない（掲載を希望する属性　：　　）

(3)　その他、教育用擬似ミクロデータの構成についての意見・要望等がありましたらご記入ください。

（　　）

５　教育用擬似ミクロデータの利用目的について

(1)　どのような目的で利用されたかご記入ください。

□　大学等の高等教育に用いた（利用人数：　学部生　　人、大学院生　　人　）

□　匿名データの申出前の分析に用いた

□　その他

具体的な利用の仕方（教育や分析等の内容）をご記入ください。

（　　）

(2)　今後、上記以外の目的で利用の可能性があればご記入ください。

（　　）

申出書に記載された利用目的以外でご利用される場合は、再度申出書の提出をお願いします。

６　集計用ソフトについて

何を利用されましたか。具体的にご記入ください。（例：R、SPSS）

（　　）

７　教育用擬似ミクロデータの利用手続や利用方法について

意見・要望等がありましたらご記入ください。

（　　）

８　教育用擬似ミクロデータの作成方法について

意見・要望等がありましたらご記入ください。

（　　）

９　教育用擬似ミクロデータの使用料について

今回、モニターとして、利用後のアンケートにご協力いただき、無償でご利用いただきましたが、有償でもご利用いただけますか。

□　有償でも利用したい。（許容できる額：　　）

□　有償では利用しない。（理由：　　）

10　教育用擬似ミクロデータの試行提供全般について

意見・要望等がありましたらご記入ください。

（　　）

教育用擬似ミクロデータを用いた教育事例について、今後の教育用擬似ミクロデータのさらなる検討のために参考にさせていただければと考えています。後日、教育事例をご報告頂く際には、ご協力よろしくお願いいたします。

ご協力ありがとうございました。

参考４　教育用擬似ミクロデータの試行提供利用者アンケートの回答結果

教育用擬似ミクロデータの試行提供利用者アンケートにおける主な回答結果を紹介する(回答総数は 17 件である。以下では、主な質問項目についての回答数を表示している)。

1 教育用擬似ミクロデータのファイル形式について

①	CSV形式に問題なかった	14
②	不都合があった	3
③	どちらともいえない	0

２ 符号表及びレイアウトについて

①	問題なかった	11
②	不都合があった	5
③	どちらともいえない	1

３ 統計センターホームページ及びパンフレットについて

①	わかりやすかった	16
②	わかりにくかった	0
③	どちらともいえない	1

４ 教育用擬似ミクロデータの構成について

(1) レコード数

①	ちょうどよい	12
②	多い	3
③	少ない	1
	無回答	1

(2) 属性

①	ちょうどよい	10
②	多い	1
③	少ない	5
	無回答	1

５ 教育用擬似ミクロデータの利用目的について

(1) どのような目的で利用されたかご記入ください。

①	大学等の高等教育に用いた	15
②	匿名データの申出前の分析に用いた	0
③	その他	1
	無回答	1

大学等の高等教育に用いた利用人数

学部生	270
大学院生	33
計	303

６ 集計用ソフトについて※複数回答

Excel	8
R	5
SPSS	5
その他	4
無回答	1

７ 教育用擬似ミクロデータの使用料について

①	有償でも利用したい。	6
②	有償では利用しない。	10
	無回答	1

付論　ミクロアグリゲーションについて

教育用擬似ミクロデータの作成方法に関しては、ミクロデータに対する匿名化技法の 1 つであるミクロアグリゲーション(Microaggregation)を方法的に展開したものと考えることができる(Bethlehem *et al.*(1990),Höhne(2003))。ミクロアグリゲーションとは、ミクロデータ(個別データ)をk個(kは閾値(threshold))のレコードを有する同質的なレコード群にグループ化した上で、そのレコードにおける個々の属性値を平均値等の代表値に置き換えることであって(伊藤(2008, 6 頁))、ヨーロッパ諸国を中心に、事業所・企業系のミクロデータにおける匿名化技法として、ミクロアグリゲーションに関する調査研究が進められてきた[16]。

ミクロアグリゲーションは、一般に、ミクロデータに含まれる量的属性に対して適用される。それに対して、質的属性については、対象となる質的属性のおのおのにおいて同一の属性値を有するレコードをグループ化した場合、それらの属性値をグループの代表値への置き換えとみなせば、質的属性値に関するレコードのグループ化もミクロアグリゲーションの一形態として位置付けることが可能である。その場合、ミクロアグリゲートデータは、付図１の例(性別、雇用形態、週間就業時間と年間収入のみを含む個別データによる例)で示されるように、特定のグループ内で同一の質的属性値群とそれに対応する量的属性の平均値を含むレコード群と考えられる(付図１における「ミクロアグリゲートデータ」を参照)。このようなミクロアグリゲートデータは、質的属性値群と量的属性の平均値群から構成された個別データに準じたデータとみなすことができるが、各レコードが持つ属性値群は、あくまで集計値として位置付けられている。

一方、グループ化の対象となる質的属性を分類事項としたクロス集計表を作成することが可能であるが、このクロス集計表におけるある特定のセルの度数とミクロアグリゲートデータにおいて対応するグループ内のレコード数は一致する。このことは、クロス集計表が高次元になるにしたがって、グループ化において使用する質的属性の数が増えることを意味している。こうした議論を展開することによって、「個別データが有するすべての属性群を集計事項の対象とした上で作成されるｎ次元の多重クロス集計表」である「超高次元クロス集計表」(付図２)を考えることができ(伊藤(2008, 19 頁))、その集計表に含まれるセルと対応関係を持ったレコード群から構成されるミクロアグリゲートデータが理論的に設定可能となる。なお、超高次元クロス集計表は、付図２に示すように、一次元からｎ次元までのあらゆる次元のクロス集計表を包含している。このことは、超高次元クロス集計

16 Thorogood(1999)によれば、ヨーロッパ諸国の企業におけるイノベーションの活動状況を調査した Community Innovation Survey(1994)においては、匿名化技法の１つとしてミクロアグリゲーションが適用されていることが知られている。

表の枠組において、擬似ミクロデータ作成のもとになる高次元の集計表を設定するための様々な次元のクロス集計表が作成可能であることを意味している。

他方、ミクロアグリゲーションでは、個別データに含まれるレコードを閾値ｋのレコード群にグループ化した上で、グループ内のレコードにおける個々の属性値を平均値等の代表値に置き換える。この場合、対象となる属性群について同一の属性値を有するレコード群(以下「同質属性値レコード群」と呼ぶ。)に存在するレコード数は、同じ属性群を分類事項として作成された超高次元クロス集計表におけるセルの度数と対応関係にある。したがって、同質属性値レコード群内に含まれるレコード数の下限が決まれば、超高次元クロス集計表に含まれるセルの度数に関する閾値が確定する。閾値ｋを設定した場合、超高次元クロス集計表の分類事項となる属性群から、属性の組合せを適当に選択することによって、超高次元クロス集計表に含まれるすべてセルが０以外のｋ未満の数にならないように集計表を構成することができる。本稿で作成した全国消費実態調査の擬似ミクロデータの場合も、閾値を３に設定し、不詳による処理等を行うことによって、３以上のレコードを持つ同質属性値レコード群へのグループ化が行われている。

このようにして、ミクロアグリゲーションにおいて超高次元クロス集計を方法的に位置付けることが可能になる。こうした超高次元クロス集計表に基づく個別データに準じたデータの作成・提供は、公的統計の二次利用における新たな可能性を提示するように思われる。なぜなら、個別データに準じたデータは集計表をもとに作成されることから、擬似的なデータの側面を持つものの、個別データと同様の属性群を有していると考えられるからである。ゆえに、超高次元クロス集計に基づいた擬似ミクロデータの作成は、公的統計の二次利用促進の観点から見ても新たな方向性を示すことができると考える。

付図１　個別データとミクロアグリゲートデータとの関係

(1)個別データ(属性群として性別，雇用形態，週間就業時間と年間収入のみが配列されていると想定)

一連番号	性別	雇用形態	週間就業時間	年間収入(千円)
1	1	1	4	2300
2	1	1	2	1500
3	1	1	4	2100
4	1	3	1	1500
5	1	3	3	2700
6	1	3	2	1800
7	2	2	3	3600
8	2	4	4	2800
9	2	2	3	4000

性別　1: 男　2:女
雇用形態　1:正規の職員・従業員　2:パート　3:アルバイト　4:派遣・契約社員
週間就業時間　1:35時間未満　2:35～48時間　3:49～59時間　4:60時間以上

(2)性別，雇用形態と週間就業時間に関する同質属性値レコード群

一連番号	性別	雇用形態	週間就業時間	年間収入(千円)
2	1	1	2	2300
1	1	1	4	1500
3	1	1	4	2100
4	1	3	1	1500
5	1	3	2	1800
6	1	3	3	2700
7	2	2	3	3600
9	2	2	3	2800
8	2	4	4	4000

対応関係

(3)性別，雇用形態，週間就業時間別クロス集計表

性別	男				女				計
雇用形態	正規の職員・従業員	パート	アルバイト	派遣社員	正規の職員・従業員	パート	アルバイト	派遣社員	
週間就業時間									
35時間未満	0	0	1	0	0	0	0	0	1
35～48時間	1	0	1	0	0	0	0	0	2
49～59時間	0	0	1	0	0	2	0	0	3
60時間以上	2	0	0	0	0	0	0	1	3
計	3	0	3	0	0	2	0	1	9

性別	雇用形態	週間就業時間	総数(N)	年間収入の総計
1	1	2	1	2300
1	1	4	2	3600
1	3	1	1	1500
1	3	3	1	2700
1	3	2	1	1800
2	2	3	2	6400
2	4	4	1	4000

(4)ミクロアグリゲートデータ

ミクロアグリゲーション後の一連番号	性別	雇用形態	週間就業時間	年間収入
1	1	1	2	2300
2	1	1	4	1800
3	1	1	4	1800
4	1	3	1	1500
5	1	3	3	2700
6	1	3	2	1800
7	2	2	3	3200
8	2	2	3	3200
9	2	4	4	4000

出所　伊藤(2008，18 頁)

付図 2　超高次元クロス集計のイメージ―全国消費実態調査を例に―

すべての属性に関するクロス集計表

(世帯主の)性別	就業・非就業の別	企業規模	(世帯主の)職業符号	・・・	総数(N)
1	1	1	1		1
1	1	1	1	・・・	1
1	1	1	2		2
・		・	・		・
・		・	・		・
・		・	・		・
1	4	5	12		2
1	4	5	12	・・・	1
・		・	・		・
・		・	・		・
・		・	・		・
2	1	1	1		1
2	1	1	1		1
2	1	1	2	・・・	2
2	1	1	3		3
・		・	・		・
・		・	・		・
・		・	・		・

性別を選択

1つの属性を選択

就業・非就業の別を選択

(世帯主の)性別	就業・非就業の別	企業規模	(世帯主の)職業符号	・・・	総数(N)
1	―	―	―		50664
2	―	―	―	・・・	4392
―	1	―	―		40783
―	2	―	―		1895
―	3	―	―		11721
―	4	―	―		657
・	・	・	・	・・・	・
・	・	・	・		・

2つの属性を選択

性別と就業・非就業の別を選択

(世帯主の)性別	就業・非就業の別	企業規模	(世帯主の)職業符号	・・・	総数(N)
1	1	―	―		38578
1	2	―	―	・・・	908
1	3	―	―		10644
1	4	―	―		534
2	1	―	―		2205
2	2	―	―		987
2	3	―	―		1077
2	4	―	―		123
・	・	・	・	・・・	・
・	・	・	・		・

すべての属性を選択

(世帯主の)性別	就業・非就業の別	企業規模	(世帯主の)職業符号	・・・	総数(N)
1	1	1	1		1
1	1	1	1	・・・	1
1	1	1	2		2
・	・	・	・		・
・	・	・	・		・
・	・	・	・		・
1	4	5	12		2
1	4	5	12	・・・	1
・	・	・	・		・
・	・	・	・		・
・	・	・	・		・
2	1	1	1		1
2	1	1	1		1
2	1	1	2	・・・	2
2	1	1	3		3
・	・	・	・		・
・	・	・	・		・
・	・	・	・		・

出所　伊藤(2008, 20 頁)

参考文献

伊藤伸介(2008)「ミクロアグリゲーションに関する研究動向」『製表技術参考資料』No.10, pp.3-31.

総務省政策統括官（統計基準担当）(2008)「統計データの二次利用促進に関する研究会報告書」

総務省政策統括官（統計基準担当）(2010)「平成21年度　統計法施行状況報告」

総務省政策統括官（統計基準担当）(2011)「平成22年度　統計法施行状況報告」

総務省統計局（2009)「平成16年全国消費実態調査報告書」

寺崎康博（2000)「リスト形式による集計表とパターン化変数」松田芳郎・伴金美・美添泰人(編著)『講座ミクロ統計分析－ミクロ統計の集計解析と技法』日本評論社,pp.111-122.

統計委員会（2009)「第20回統計委員会議事録」

統計委員会匿名データ部会（2009)「第1回匿名データ部会議事概要」

日本学術会議学術基盤情報常置委員会（2005)「政府統計・世帯調査等の一次データ（含む個票データ）の体系的保存と活用・公開方策について」

星野伸明（2010)「公的統計ミクロデータ提供制度の課題」『日本統計学会誌』第40巻第1号, pp.23-45.

山口幸三(2008)「政府統計の個票利用と統計法改正」『経済研究』第59巻第2号,pp.139-152.

松田芳郎（1999)『ミクロ統計が描く社会経済像』日本評論社

松田芳郎（2008)「日本におけるミクロデータ政府統計活用の新しい夜明け」『統計』第59巻第12号, pp.2-9.

美添泰人(2009)「統計の有効活用に関する展望と課題」『ESTRELA』2009年4月号(No.181), pp.9-17.

Bethlehem, J. G., Keller, W. J. and Pannekoek, J.(1990) “Disclosure Control of Microdata”, *Journal of the American Statistical Association*, Vol.85, No.409, pp.38-45.

Höhne, J.(2003) “SAFE- A Method for Statistical Disclosure Limitation of Microdata”, Joint ECE/Eurostat Work Session on Statistical Data Confidentiality (Luxenbourg, 7-9 April 2003)

http://www.unece.org/fileadmin/DAM/stats/documents/2003/04/confidentiality/wp.37.s.e.pdf

Thorogood D.(1999) “Protecting the Confidentiality of Eurostat Statistical Outputs”, *Netherlands Official Statistics*, Volume 14, Spring, pp.30-33.

製 表 技 術 参 考 資 料 16

平成 24 年 7 月発行

編集・発行　　独立行政法人　統計センター

〒162－8668

東京都新宿区若松町 19－1

電　話　代表　03（ 5273 ）1200

掲載論文を引用する場合は、事前に下記まで連絡してください

情報技術部統計技術研究課　　TEL ： 03－5273－1368

E-mail ： research@nstac.go.jp