Discussion Paper Series

University of Tokyo
Institute of Social Science
Panel Survey

東京大学社会科学研究所　パネル調査プロジェクト
ディスカッションペーパーシリーズ

標本調査における性別・年齢による層化の効果：100 万人シミュレーション

Effect of Sex-Age-Stratification in Sample Survey:
A Simulation of One Million People

山本耕資
（東京大学社会科学研究所）
Koji YAMAMOTO

April 2007

No.1

東京大学社会科学研究所
INSTITUTE OF SOCIAL SCIENCE　UNIVERSITY OF TOKYO

東京大学社会科学研究所パネル調査プロジェクト
ディスカッションペーパーシリーズ　No.1
2007 年 4 月

標本調査における性別・年齢による層化の効果：100 万人シミュレーション

山本耕資（東京大学社会科学研究所）

要約　標本調査でしばしば用いられるのは地域特性によって層化する 2 段抽出である。この抽出方法に、さらに地点内での性別・年齢による層化を加えた場合に、母集団に対する推定における誤差分散（ないし標準誤差）が減少するのか否かを、本稿ではシミュレーションによって検討する。シミュレーションによる誤差分散の計測の妥当性を理論値との比較によって確認したのちに、上記のような 2 段抽出での性別・年齢による層化は誤差分散を減少させる上で一定の意義があることを、統計的検定によって示す。

付記　本稿での JGSS データの分析にあたり、東京大学社会科学研究所附属日本社会研究情報センターSSJ データアーカイブから個票データの提供を受けた。日本版 General Social Surveys (JGSS)は、大阪商業大学比較地域研究所が、文部科学省から学術フロンティア継続拠点としての指定を受けて（1999-2003 年度)、東京大学社会科学研究所と共同で実施しているプロジェクトである（研究代表：谷岡一郎・仁田道夫、代表幹事：佐藤博樹・岩井紀子、事務局長：大澤美苗)。東京大学社会科学研究所附属日本社会研究情報センターSSJ データアーカイブがデータの作成と配布を行なっている。本稿は 2006 年度日本学術振興会科学研究費補助金（若手研究（スタートアップ））による研究の副次的成果である。本研究にあたり、東京大学社会科学研究所パネル調査実施委員会メンバー各氏に問題関心を共有していただき、とりわけ石田浩教授と三輪哲准教授には有益な助言や支援をいただいた。また、土屋隆裕准教授による標本調査法の授業資料を参考にさせていただいた。石田浩、三輪哲、土屋隆裕の各氏と、東大社研パネル調査実施委員会メンバー各氏、JGSS プロジェクトと SSJ データアーカイブの関係者各位に、深く謝意を表したい。もちろん、本稿のすべての欠陥の責任は筆者にある。

1. はじめに

一般の市民を対象に日本で全国規模の標本調査を行なう場合、多く用いられる手法は層化 2 段無作為抽出法である。この方法では通常、層化は何らかの地域特性を基準に行われる。具体的には、市郡規模と地域ブロック（北海道、東北などの地域区分）がしばしば基準とされる。層化の利点は、観測したい変数の標本誤差を減少させうる点にある。この層化の効果は、層化基準となる変数、すなわち上の例では地域特性が、観察したい変数と関連しているほど大きくなるとされる[1]。

それでは、地域以外の属性を基準に層化するとどうだろうか。社会調査による研究が対象とする変数は､多くの場合で性別や年齢といった回答者の基本的な属性と関連している。他方、抽出台帳として普通用いられる住民基本台帳や選挙人名簿から、対象者の性別と年齢を知ることができる。以上を考え合わせれば、標本調査に性別・年齢による層化という方法を採り入れることが当然に関心の対象となる。実際に、2007 年から実施されている東京大学社会科学研究所パネル調査（社研パネル調査）では、地域特性による層化とともに、地点（2 段抽出の第 1 次抽出単位）内で性別・年齢を基準に層化をしながら対象者を抽出している。本稿は、このような性別・年齢による層化の効果をシミュレーションによって探る試みである。

具体的な本稿の手続きは次のようなものである。まず、コンピュータ内に 100 万人からなる仮想的母集団データを、乱数を用いて生成する。この際、より「現実的」な結果を得るために、既存の標本調査データの分布を参考にする。そののちに、この生成された母集団データに種々の標本抽出法を適用し、4,000 名からなる標本を抽出して比率や平均などの統計量の値を算出する。この操作を、それぞれの抽出方法に関して 50,000 回繰り返し、統計量の誤差分散が抽出方法によってどう違うのかを観察する。このうち、本稿の主たる関心は、地域を基準とする層化 2 段無作為抽出法と、これに加えて地点内で性別・年齢によって層化した方法との比較にある。

2. 母集団の生成

本稿では、より「現実的」な母集団を生成するために、既存の標本調査データに含まれる変数の実際の分布を参考にする。取り上げる（想定される）変数は、「教育程度」、「就労」（職業を有するか否か）、「家庭生活満足度」（以下「家庭満足度」とする）、「家計状態満足度」（以下「家計満足度」とする）である[2]。家庭満足度と家計満足度は間隔尺度と見なす。観察の対象

[1] 例えば盛山（2004: 127）、Groves et al. (2004: 116-117)を参照されたい。

[2] これらの変数を取り上げる理由は以下のとおりである。方法論的関心から、しばしば用

とするのは、各々の 1 変数の分布（比率、平均）に加えて、教育程度と就労の連関、家庭満足度と家計満足度の連関である。これらの連関は、それぞれに実質的意味を持ちうるものではあるが、本稿の関心は直接にはその解釈にはないことを述べておきたい。なお、抽出の際に用いる属性変数として、回答者の居住する「地域」（しばしば「市郡規模」と呼ばれるもの）、「性別」、「年齢」も扱う。

(1) 現実のデータにおける分布

本項では、シミュレーションの母集団生成に際して参考とする実際の標本調査データの分布を示す。用いるデータは JGSS-2003 データである[3]。配布されているデータセットを表 1 のように加工して使用する。欠損の処理は、以下のいずれかの変数に有効な回答がないケースは欠損ケースとしてどの分析からも対象から外す、リストワイズの欠損処理を行なう。

これらの変数の分布を示したのが表 2～表 6 である。まず、回答者の居住する地域の分布が表 2 に示される。表 3 は、それぞれの地域における性別・年齢の分布を表している。地域が 3 タイプに分かれ、性別と年齢はそれぞれ 2 カテゴリに分けられるので、全国のサンプルは 3×2×2=12 の《地域・性別・年齢》カテゴリに分類されることになる。表 4 では、それぞれの《地域・性別・年齢》カテゴリでの、教育程度と就労との 2 変数の分布が示される。これをもとに、同じく《地域・性別・年齢》カテゴリごとに、教育程度が低い場合に職を持つ対数オッズ、教育程度が高い場合に職を持つ対数オッズ、さらに、教育程度と就労との関係を表す対数オッズ比を計算したのが、表 5 である。この対数オッズ比は、教育程度が高い者が職を有する傾向が強いほど大きい値をとるものだが、《地域・性別・年齢》カテゴリによって値が大きく異なっており、連関の向きもしばしば相違している[4]。表

いられる統計量の性質を調べるために、2 値カテゴリ変数の比率、2 値カテゴリ変数同士の対数オッズ比、間隔尺度変数の平均、間隔尺度変数同士の積率相関係数に着目することとした。これらの統計量を算出するために、2 つの 2 値カテゴリ変数と、2 つの間隔尺度変数を取り上げる。2 値カテゴリ変数としては、社会調査において基本的な関心の対象となる教育程度と就労を取り上げた。これらは、本稿で参照する既存調査のデータである JGSS の 2003 年調査データに含まれる。間隔尺度（と見なす）変数としては、一般の関心を引きうるものとして、「生活満足度」の一連の項目（JGSS-2003 面接調査票 A 票 Q7・B 票問 12）を検討した。予備的分析の結果、相関係数が高く、かつ、相関係数の性別・年齢による差が大きい項目のペアであった家庭満足度と家計満足度を取り上げた。性別・年齢による差が大きいペアを選んだのは、のちに性別・年齢によって層化する抽出法の効果をより明瞭に確かめようとする意図による。

[3] 本稿では分析開始時点で入手可能な最新の JGSS データであった 2003 年調査のデータを用いる（JGSS 累積データ 2000-2003 による）。

[4] ただし、これらの連関の実質的意味を解釈しようとするならば、1 つの年齢カテゴリに

6 に示されるのは、家庭満足度と家計満足度の平均、標準偏差と、これら 2 変数の積率相関係数である。

以上で示された変数の分布が、仮想的母集団の生成の際に参照される。

(2) 乱数による母集団の生成

仮想的母集団には 100 万人が存在すると想定する。その生成に際して、母集団における《地域・性別・年齢》カテゴリごとの人数をまず定める。現実データにおける分布は先に示した**表 2**、**表 3** のとおりであり、この分布を参考にしながら、母集団のカテゴリごとの人口を、**表 7**、**表 8** のようにややキリのよい数値に設定した[5]。ここで、**表 3** に示された数

含まれる年齢の幅の広さが問題となりうる。

[5] このようにキリのよい数値に設定する理由の 1 つは、本稿で扱うどの調査方法でも（線形推定量に関して）自動加重標本が得たいというものである。より具体的には、のちに述べる地域層化 2 段抽出法＋性別年齢層化で、200 地点から 20 名ずつ抽出する際に、各地域層・地点・性別年齢層の母集団人口や抽出する個人数は自然数でしかありえないという制約の下で、各個人が抽出される確率をすべて等しくするため、設定に注意が必要となる。具体的には、少なくとも以下の条件が満たされるような設定とする。(a)地域層に地点を割り当てる際、母集団人口に比例させる比例割当を、自然数の人数で偏りなく行なえるようにする。200 地点の配分の問題であるから、各地域の母集団人口相対頻度は 0.5%（=1/200）の自然数倍である必要がある。例えば**表 7** では「大都市」の母集団人口相対頻度は 20%と設定されているため、200 地点のうち 40 地点が「大都市」に割り当てられ、割り当てられる地点数は自然数になる。もし仮にこれが 20%ではなく 20.1%であるならば、割り当てられる地点数は 40.2 となり、自然数にならないので、不適当である。(b)各地点の母集団人口は 20 名の自然数倍とする（後述）ため、各地域の人口も 20 の自然数倍である必要がある。例えば**表 7** では「大都市」の母集団人口は 200,000 と設定されており、これは 20 の自然数倍である。(c)各地点内で、性別・年齢で層化して 20 名を抽出する際に、各性別・年齢カテゴリの母集団人口に対する比例割当を、自然数の人数で偏りなく行なうために、地点内の各性別・年齢カテゴリの母集団人口相対頻度は 5%（=1/20）の自然数倍である必要がある。例えば、「大都市」ではどの地点においても「20-54 歳・女性」は 35%存在する（**表 8** にある相対頻度と等しい）ため、20 名のうちこのカテゴリには 35%にあたる 7 名が割り当てられ、割り当て人数が自然数になる。もし仮にこれが 35%ではなく 36%であるならば、各地点でこのカテゴリから抽出する人数は 7.2 となり、自然数にならないので、不適当である。(d)地点内の各性別・年齢カテゴリの母集団人口が、自然数で、かつ、設定された相対頻度を満たすように、各地点の母集団全体の人口を 20 名の自然数倍とする。これによって、地点内の各性別・年齢カテゴリの母集団人口相対頻度がどのような値に設定されても（5%の整数倍である限りは）カテゴリ別の人口を自然数で表現できる。例えば、「大都市」では「20-54 歳・女性」は 35%存在するので、母集団人口 120 名の地点では 42 名存在することになり、この人数は自然数となる。もし仮に、「大都市」に母集団人口 121 名の地点があるならば、「20-54 歳・女性」の人数は 42.35 となり、自然数にならないので、不適当である。なお、本稿の目的は自動加重性やウェイトの妥当性の検討ではなく誤差分散の評価を行なうことであるが、より「自然」な状況での自動加重性などの問題は別途検討課題となりうることを述べておきたい。

値は非回答による偏りについて修正していない分布であり、例えば女性が男性より明らかに多く含まれている。この分布を参考にしているために、仮想的母集団でも男女比の偏りのある分布が設定されている（表 8）。

この仮想的母集団において、教育程度と就労の変数は次のように生成した。特定の《地域・性別・年齢》カテゴリに対して、JGSS データでの教育程度と就労の 2 変量の分布が表 4 に示されている。この分布に従う乱数によって、仮想的母集団内の各《地域・性別・年齢》カテゴリの各個体の教育程度と就労の状態を定めた。その結果、仮想的母集団においては表 9 のような分布が見られることとなった。乱数を用いた結果であるため、もとになった表 4 の分布とは厳密には相違するが、極めて近い。仮想的母集団での教育程度と就労との連関を示す対数オッズ比なども算出した（表 10）。この結果も、対応する JGSS データでの表 5 と、一致はしないが、同様の傾向を示している。

家庭満足度と家計満足度は、各《地域・性別・年齢》カテゴリで、JGSS データにおける平均・標準偏差・相関係数（表 6）をパラメータとする 2 変量正規乱数を生成して、各変数の値とした。これらは連続型分布の乱数によっているため、生成されたデータは、5 点尺度であった JGSS データとは異なり、任意の実数をとりうる。結果として、2 つの満足度変数に関して表 11 のような母集団分布が得られた。これらの変数も乱数によって生成しているために、その分布はもととなった分布とはわずかに相違する。

のちに 2 段抽出を行なうために、この仮想的母集団において各個人はそれぞれある 1 つの「地点」に属することとする。各地点はいずれかの地域に属し、100～200 名（20 名刻み）の個人を含む[6]。日本の一般市民対象の全国調査がしばしば地点として用いるのは国勢調査の調査区であり、仮想的母集団の地点あたり人口はこの調査区の母集団人口を模して設定している[7]。地点あたりの母集団人口は一様乱数で決定される[8]。

---

[6] 地点あたりの母集団人数を 20 名刻みとする理由については注 5 を参照されたい。

[7] 2000 年国勢調査では、設定された調査区の数は 939,537 であり、20 歳以上の日本人人口は 99,647,839 名であった（総務省統計局編　2005）。1 調査区あたりの 20 歳以上日本人人口の平均は 106.06 名である。なお、JGSS データのコードブックの調査方法の説明によると、JGSS の調査においては「調査区」ではなく「基本単位区」が地点とされている。

[8] 各地域に対して母集団人口があらかじめ定めてあるため（表 7）、この人口を地点に割り当てるために、乱数を発生させてこれによって地点の母集団人口を定めて割り当て、その人口だけ「残り人口」を引いていくという操作を、「残り人口」がゼロになるまで繰り返した。この際、100 以上 200 以下で 20 刻みの数値を等確率でとる乱数を用いた。ただし、地点母集団人口の合計が、設定された地域の母集団人口に対して過不足がないように、以下のような手順で調整を行なった。「残り人口」が 100 以上 200 以下で、発生させた乱数が「残り人口」より多い場合、「残り人口」すべてを最後の 1 地点に割り当てる。「残り人口」が 200 以下で、発生させた乱数と「残り人口」との差が 100 未満である場合、乱数の値に関わらず「残り人口」すべてを最後の 1 地点に割り当てる。「残り人口」が 200 を超えていて、発生させた乱数と「残り人口」との差が 100 未満である場合、この乱数は放棄し、改めて乱数を発生させる。

仮想的母集団では、特定の地域内においては、地点間で分布の特性に違いは設定されない。これには 2 つの側面がある。第 1 に、同じ地域に含まれるどの地点においても、性別・年齢の構成比は等しい。すなわち、**表 8** の行方向の相対頻度に合致するように、地点の性別・年齢別の母集団人口が定められる。例えば、「大都市」においては「20-54 歳・女性」はどの地点でも 35%存在することになる。2 段抽出で、地点内で性別・年齢によって比例割当で層化する場合には、1 地点から全部で 20 名を選ぶので、35%を占める「20-54 歳・女性」は 7 名選ばれることになる。もし仮に、「大都市」の中でも地点によって「20-54 歳・女性」の構成比が異なるとすれば、地点内層化の際の「20-54 歳・女性」への割り当て数を一律に 7 名とすると、抽出される確率が不均等となり、自動加重標本ではないことになる。地点内層化の際の性別・年齢カテゴリへの割り当て数を同一地域内で等しくしながら自動加重標本を得るために、地点間での性別・年齢の構成を等しくしているのである[9]。第 2 に、データを乱数で生成する際のパラメータは、《地域・性別・年齢》カテゴリの別によってのみ異なっている。すなわち、地点によって乱数生成のパラメータの差はなく、同一地域で同一性別・年齢の個体について、地点間のデータの差は独立なランダムネスの結果として生じるのみである。例えば、「大都市」にいる「20-54 歳・男性」であれば、どの地点にいようと、乱数によって 7.5%の確率で「高等教育を受けた無職」という属性データが与えられる（**表 4** より。ランダムネスによって、結果的には仮想的母集団のこのカテゴリの「高等教育を受けた無職」は 7.4%となっている）。

以上のようにして、100 万人からなる仮想的母集団データをコンピュータ内に生成した。

3. 抽出方法と誤差の算出方法

(1) 抽出方法

本稿のシミュレーションでは、生成された仮想的母集団データから、4,000 名の標本を抽出して統計量を計算するという操作を、8 種類の抽出方法について、それぞれ 50,000 回繰り返す。抽出方法の種類は以下のようなものである。

---

[9] この点は現実の調査実施上の重大な問題を示唆する。地点間で性別・年齢の構成比が異なるという現実的な仮定の下で、2 段抽出で、地点の抽出を地点内母集団全人口に比例させた確率で行ない、地点内で性別・年齢による層化を行なう場合、自動加重標本を得ようとするならば、厳密には、各性別・年齢カテゴリで抽出される個人数を、地点によって個別に検討する必要がある。しかし現実にはこのような方法には困難が伴う可能性がある。全く別の可能性として、地点内で性別・年齢による層化を行なうのではなく、全国の母集団を性別・年齢によって分割し、性別・年齢カテゴリごとに全く独立に地点を選んでさらに個人を選ぶ、という方法もありうるが、これにも現実には困難が伴いうる。

a. 段を設けない方法

まず、段を設けない（2 段抽出ではない）方法として、次の 4 つの方法を用いる。

- 単純無作為抽出法
- 地域層化無作為抽出法
- 性別年齢層化無作為抽出法
- 地域層化＋性別年齢層化無作為抽出法

これらは、単純無作為抽出法をベースとし、地域で層化するか否かと、性別・年齢で層化するか否かによる 4 つのバリエーションとなっている。単純無作為抽出法の場合は母集団全体から、層化を行なう場合は層内から、乱数により無作為に個人を抽出する。いずれの場合も非復元抽出を行なう。層化を行なう際は、各層の標本サイズはその層の母集団サイズに比例させる比例割当で決められる。例えば、母集団において「大都市」に居住する者の割合は 20%であるから、地域層化無作為抽出法で「大都市」に割り当てられる（「大都市」から選ばれる）個人数は、サンプルサイズ 4,000 の 20%にあたる 800 となる。これにより、層化の有無に関わらず、どの個人も選ばれる確率が等しく、平均（・比率）に関する自動加重標本が得られることになる[10]。

b. 2 段抽出に属する方法

さらに、2 段抽出に属する方法として、次の 4 つを用いる。

- 2 段無作為抽出法
- 地域層化 2 段無作為抽出法
- 2 段抽出法＋性別年齢層化
- 地域層化 2 段抽出法＋性別年齢層化

これらの方法では、地点を第 1 次抽出単位（primary sampling unit）とし、地点を選ぶ 1 段目の抽出では、地点の母集団サイズに包含確率を比例させる確率比例抽出[11]を、乱数により行なう。具体的には、Cochran (1977: 265-266)にあるように、系統抽出法を用いて 200 地点を抽出する。1 地点あたりの抽出個人数は 20 名であり、個人を選ぶ 2 段目の抽出は等確率で、乱数を用いて行なわれる[12]。2 段抽出の場合も、1 段目・2 段目ともに、

---

10 ここでは、自動加重標本とは、不偏な線形推定量が、その標本での変数の単純な総和と他の値(例えばサンプルサイズの逆数)との積の形で表現できるような標本を指している。この定義によれば、本稿のシミュレーションの抽出で得られる標本はいずれも、平均について自動加重標本である。比率は 0 と 1 をとる 2 値カテゴリ変数の平均と見なせる。

11 JGSS の調査実施にあたっては、地点の抽出は確率比例抽出では行なわれていない点を注記したい。

12 本稿のシミュレーションでは、地点を選んだ後の 2 段目の抽出では選ばれた地点に属す

非復元抽出が行なわれる。

これらの 4 種類の方法も、地域層化と性別年齢層化の有無での 4 つのバリエーションとなっている。ただし、地域層化と性別年齢層化とでは層化を行なう段階が異なっている。すなわち、地域層化は地点を選ぶ 1 段目でなされる一方、性別年齢層化は地点内で個人を選ぶ 2 段目でなされる。

地域層化を行なう場合、各層の抽出地点数はその層の母集団サイズに比例させる比例割当によって決まる。1 地点あたりの抽出個人数が 20 名で一定であるために、各層の抽出個人総数も層の母集団サイズに比例することになる。

性別年齢層化の場合は、地点内の母集団を性別・年齢で層化して、抽出個人数を各層の人口に比例させる比例割当で決め、抽出する。

これらの 2 段抽出においても、層化の際の比例割当、地点を選ぶ際の確率比例抽出と地点内での等確率抽出によって、どの個人も選ばれる確率が等しく[13]、平均（・比率）に関する自動加重標本が得られることになる。

ここで、これまでの記述と重なるが、抽出の手続きを確認する意味で、地域層化 2 段抽出法＋性別年齢層化での抽出過程の流れを説明したい。まず、各地域に抽出地点数（抽出個人数に比例する）を割り当てる。母集団人口に比例させて、全体で 200 地点を抽出するため、表 7 から、例えば「大都市」には 20%にあたる 40 地点が割り当てられる。次に、「大都市」の地点全体から、40 地点を乱数を用いて選び出す。この際、地点内の母集団人口に比例した確率でその地点が選ばれるようにする。その後、選ばれた地点で個人を選び出す。この際、母集団の各個人を性別・年齢によって「20-54 歳・男性」「55-89 歳・男性」「20-54 歳・女性」「55-89 歳・女性」の 4 カテゴリに分けて層化抽出を行なう。この 4 カテゴリの母集団分布は、表 8 にある行方向の相対頻度のとおりに、地域ごとに一定である。例えば「大都市」に属する地点では、「20-54 歳・女性」は必ず 35%存在する。よって、地点内のこれら 4 カテゴリに抽出個人数を母集団人口に比例させて割り当てる際、各地点から全部で 20 名を選び出すため、20 名の 35%にあたる 7 名を「20-54 歳・女性」から選び出すことになる。地点内の「20-54 歳・女性」の母集団からどの個人を選ぶかは乱数を用いて決定する。以上が地域層化 2 段抽出法＋性別年齢層化での抽出の概観である。

る個人から選び出すが、JGSS の調査では、選んだ地点に属する個人のみを選び出すのではなく、選んだ地点からいわば「はみ出し」て、別の地点の個人を抽出することもありうる方法をとっている。

[13] 2 段抽出でどの個人も等確率で選ばれるようにするための設定上の工夫について、注 5 を参照されたい。

(2) 誤差の計算方法

抽出のシミュレーションで観察対象となる統計量は次のとおりである。教育程度に関して高等教育を受けた者の比率、就労に関して有職者の比率、家庭満足度の平均、家計満足度の平均である。さらに、教育程度と就労の連関を示す対数オッズ比と、2 つの満足度変数の積率相関係数も算出する。

以上の統計量の算出を、各抽出方法につき 50,000 回標本を選ぶごとに行なう。結果として、50,000 の統計量の実現値が得られることになる。この 50,000 の値の分散が誤差分散、その平方根が標準誤差の、それぞれ実現値であると見ることができる[14]。これらの誤差の値が、本稿の主たる関心の対象である。

理論的に定まる誤差分散が存在すると考えると、この理論値とシミュレーションによる実現値との関係は、「母分散」と「標本分散」の関係と見なすことができるはずである。通常、標本分散は不偏に母分散を推定するから、シミュレーションによる誤差分散の実現値を理論値の推定に用いることができる[15]。この場合、誤差分散の理論値に対してシミュレーションによる誤差分散が推定量として振る舞うことになるが、この「誤差分散の推定量」にも分散が存在する。この「誤差分散の推定量の分散」は、シミュレーションで抽出を行なう回数（本稿では 50,000）にほぼ反比例すると考えられる[16]。すなわち、シミュレーションでの抽出の回数を増やすほど、その結果から、理論的に定まっている誤差分散の値をよりうまく推定できるようになる。以上のような考え方をもとに、次節では標準誤差の信頼区間や理論的に予測される分布の区間の算出と、誤差分散の比の検定を行なう。

4. シミュレーションの結果

(1) 結果の概観

以上の設定と方法のもとで、抽出のシミュレーションを行なった結果が表 12 に示され

---

[14] 正確に言えば、本稿で着目するのは推定量（統計量）の分散（とその平方根）であって、平均 2 乗誤差（mean square error; MSE）ではない。ただし不偏推定量ではこれらは一致する。本稿では推定量（統計量）の分散を誤差分散、その平方根を標準誤差と呼んでいる。

[15] シミュレーションで得られた誤差分散の実現値が理論値の不偏推定量であるためには、当該統計量について標本分布からランダムに実現値が得られるという条件が必要であるように思われるが、本稿のシミュレーションではこの条件は基本的に満たされると考える。

[16] ある統計量についてシミュレーションで得られた誤差分散を s^2、理論的に定まる誤差分散を σ^2、シミュレーションでの抽出回数を N とし、その統計量の標本分布を正規分布と仮定すると、$(N-1)s^2/\sigma^2$ は自由度(N-1)の χ^2 分布に従う。自由度(N-1)の χ^2 分布の分散は $2(N-1)$であることから、s^2 の分散は $2(\sigma^2)^2/(N-1)$となる。

ている。各統計量について、母集団における値（対応する母数）と、各抽出方法で得られた実現値の平均・標準誤差を掲げた[17]。

この結果からまず、各統計量に関して、母集団での値（母数）と各抽出方法での実現値の平均とを比較したい。教育程度と就労の比率に注目すると、いずれの抽出方法においても統計量の実現値の平均と母集団での値との差、すなわち抽出による偏りが、ほとんどないことがわかる。例えば、有職者の比率は、母集団において 0.5757 である一方、各抽出方法での平均を見ると、1 つの方法を除いてすべて 0.5757 という数値になっており、例外である地域層化 2 段抽出法＋性別年齢層化でも 0.5756 という値であり母数との差は極めて小さい。家庭満足度と家計満足度の平均についても、抽出による偏りはほとんどない。ここでは比率と平均は不偏推定量となっているはずであり、上記の結果はこれを裏付けている。家庭満足度と家計満足度の相関係数でも、母集団における相関係数と標本での相関係数の平均値の差はほとんどない。相関係数は偏りのある推定量だが、少なくとも単純無作為抽出法のもとでは一致性があり、4,000 というサンプルサイズはここでは十分な程度の収束をもたらしていると考えられる。標本における教育程度と就労との対数オッズ比は、他の統計量と比べると、母数との差が大きくなっているが、対数オッズ比も一般には一致性はあるが不偏性はないという点と、標準誤差も大きいという点に注意すべきである。

表 12 に示された標準誤差をグラフで表現したのが図 1 である。ここには、標準誤差の 95%信頼区間の上限・下限も示した[18]。理論的に定まる「真の」標準誤差が存在するならば、それはこの信頼区間の中に 95%の確率で入るはずである。このように幅のある区間で「真の」標準誤差を推定するのは、前述のようにシミュレーションで得られる標準誤差にも「誤差」が存在するからである。

これらの図から、まず、高等教育を受けた者・有職者の比率については層化の効果が認められる。地域層化の効果が見られる部分もあるが、性別年齢層化の効果の方がより大きい。教育程度と就労との対数オッズ比では、2 段抽出において性別年齢層化の効果が見られる一方、地域層化をするとむしろ誤差が大きくなるという結果となっている[19]。また、

---

[17] 表 12 にある各抽出法で、50,000 回の抽出・統計量の算出を行なうのに要した時間を参考までに計測した。もちろん計算時間はアルゴリズムや計算環境に依存するが、段を設けない抽出法（4 種類）では計算時間がそれぞれ 700～1000 秒台であった一方、2 段抽出の 4 つのバリエーションでは、それぞれ 60 秒台で計算が終了した。2 段抽出と段を設けない抽出法とで計算時間に大きな差がある。

[18] ここでの信頼区間の算出に際しては、統計量の標本分布が正規分布であると仮定している。

[19] 任意の 2 つの抽出方法に関して、各統計量の誤差分散が、等しいのか、有意に異なるのかを調べるため、本稿 4.(3)（表 15）にあるような F 検定を行なった（ただし、ここでは両側検定とした）。その結果、教育程度と就労との対数オッズ比の誤差が 2 段抽出において地域層化をするとむしろ大きくなるという効果は 5%水準で有意に認められる。

対数オッズ比では、2 段抽出の方が段を設けない抽出よりも誤差が小さくなる[20]。家庭満足度・家計満足度の平均に関しては、層化の効果は明瞭には認められない。地域層化によって誤差が大きくなる部分があるようにも見えるが、信頼区間の広さを考えると決定的ではないと思われる[21]。満足度変数同士の相関係数でも、地域層化の効果は明確ではない一方、性別年齢層化の効果が確認できる部分もある。

以上をまとめれば、地域層化より性別年齢層化に効果が認められる場合が多く、その効果は教育程度と就労において顕著であると言える。2 段抽出が段を設けない場合と比べて標準誤差を増す傾向はほとんどないが、これは仮想的母集団の設定上、同一地域内であれば地点間で同質的な個人データが生成されてしまうことに起因していると思われる。

(2) 理論値との比較

本項では、シミュレーションによって得られる統計量の誤差の値と、理論的に算出される誤差の理論値との比較を行なう。ここでは、標準誤差を比較的容易に理論式から計算できる抽出方法・統計量を対象とする。

まず、単純無作為抽出法について、標準誤差の理論値を計算する。各統計量に関する理論式は次のものを用いる[22]。標本サイズを n、母集団サイズを N とする。比率の標準誤差は、母比率を p として、

$$\sqrt{\frac{p(1-p)}{n-1}\cdot\left(1-\frac{n}{N}\right)}$$

で計算される。対数オッズ比については、対象となる 2×2 クロス表の各セルに該当する 4 カテゴリの、母集団での比率をそれぞれ π_1、π_2、π_3、π_4 とすると、漸近的な標準誤差は、

$$\sqrt{\frac{1}{n\pi_1}+\frac{1}{n\pi_2}+\frac{1}{n\pi_3}+\frac{1}{n\pi_4}}$$

となるとされる。平均の標準誤差は、母集団における i 番目の個体のデータを y_i として、

$$\sqrt{\frac{S^2}{n}\cdot\left(1-\frac{n}{N}\right)} \quad \text{ただし} \quad S^2=\frac{\sum_{i=1}^{N}(y_i-\bar{y})^2}{N-1},\ \bar{y}=\frac{\sum_{i=1}^{N}y_i}{N}$$

である。積率相関係数の標準誤差は、母相関係数を ρ として、近似的に、

---

[20] 注 19 と同様の検定の結果、この効果も有意に認められる。

[21] 注 19 と同様の検定の結果、家庭満足度・家計満足度の平均の標準誤差が地域層化によって有意に大きくなる組み合わせは認められない。

[22] これらの理論式に関しては、Groves et al. (2004: 100)、Agresti (2002: 76)、Cochran (1977: 23)、南風原（2002: 115）を参照されたい。

$$\frac{1-\rho^2}{\sqrt{N}}$$

となるとされる。

これらの理論式によって算出した理論値と、シミュレーションの結果とを比較したのが、表 13 である。シミュレーションで得られた標準誤差は理論値にほぼ等しいことが確認できる[23]。

次に、段を設けない層化無作為抽出法について、比率・平均の標準誤差の理論値を、Cochran (1977: 89-93)に依って計算した。この結果は表 14 に示されている。ここには単純無作為抽出法での結果も載せてある。さらにこれらを図示したのが図 2 である。この図には、理論値を前提として、シミュレーションによる標準誤差の実現値が 95%の確率で入る範囲の上限・下限も示した[24]。この範囲は、いわば「標準誤差の誤差」の範囲の目安となる。ここで明らかなように、シミュレーションの結果として得られた標準誤差は理論値から予想される誤差の範囲内にほぼ入っている。

以上のように、シミュレーション結果は理論式によって得られる理論値と実質的に変わらない。これは本稿でなされたシミュレーションの妥当性を確認している。単純無作為抽出法や層化無作為抽出法であれば、比率・平均の正確な標準誤差を母集団分布から比較的容易に計算可能だが、非復元抽出による 2 段抽出においては正確な計算は煩雑となる。よって、次項での、2 段抽出における層化の効果の比較は、シミュレーションの強みを生かしたものであると言えよう。

ただし、ここで次の点に注意すべきである。図 2 の予測上限・下限の幅が示すとおり、シミュレーションで得られた標準誤差の値は、理論値から離れた値をとることもありうる。このため、異なる抽出方法の間でシミュレーションによる標準誤差を単純に比較した場合、理論的には正しくない結論を導く可能性がある。例えば、表 14 にある、家計満足度の平均の標準誤差に着目しよう。性別年齢層化無作為抽出法の場合と、これに地域層化も加えた場合とで、この標準誤差を比較すると、理論値では地域層化を加えた方が標準誤差が減る（0.01715 から 0.01714 に）のに対し、シミュレーションでは逆に増える（0.01714 から 0.01717 に）、という結果となっている。しかし、図 2 から明らかなように、シミュレーションで得られた家庭満足度平均の標準誤差は、理論的に十分ありうる範囲、いわばありうる「標準誤差の誤差」の範囲に入っている。よって、シミュレーションで得られる標

[23] 単純無作為抽出法での相関係数の標準誤差の理論値は、シミュレーションの結果から得られた 95%信頼区間（図 1 参照）に入っておらず、理論値とシミュレーション結果との差が顕著であると言える。ただし、相関係数の標準誤差の理論値は近似的なものである点、および、シミュレーション結果から信頼区間を計算するに際して標本分布の正規性を仮定している点に注意する必要があろう。

[24] この範囲の算出に際しては、統計量の標本分布が正規分布であると仮定している。

準誤差を異なる抽出方法の間で比較する場合には、この「標準誤差の誤差」に注意する必要がある。ある抽出方法と別の抽出方法とで、理論的に標準誤差が異なると言えるのかをシミュレーションから判断するには、「標準誤差の誤差」を考慮した統計的検定が有力な手段となるであろう。このような観点から、次項では誤差分散の比の検定も行なう。

(3) 地域層化 2 段抽出法における性別年齢層化の有無の比較

それでは、本稿のそもそもの問題意識に立ち戻り、より実践的なトピックを検討したい。地域層化をした上での 2 段抽出は広く標本調査において用いられる手法であるが、その際に、さらに地点内で性別年齢層化を施すことの意義を調べるのである。すなわち、地域層化 2 段無作為抽出法と、地域層化 2 段抽出法＋性別年齢層化との比較をより詳細に行なう。

まず、これら 2 つの抽出方法に関して、得られた統計量の標本分布をヒストグラムで示したのが図 3 である。地域層化 2 段抽出法によるものを点線で、これに性別年齢層化を加えた抽出方法によるものを実線で、それぞれ表している。2 つの抽出方法によるヒストグラムは、かなりの程度重なっているものの、高等教育を受けた者の比率と有職者の比率においては、性別年齢層化をする方がモード付近により多くの実現値が集中していることが明確にわかる。

次に、統計的検定によって、果たして性別年齢層化が誤差を減少させるのか否かを判断しよう。表 15 には、シミュレーションで得られた、性別年齢層化がない場合とある場合の標準誤差が記されている。どの統計量についても、性別年齢層化を行なった場合の方が標準誤差が小さい。ただし、前項で述べたように、この違いは「標準誤差の誤差」の範囲内にある可能性もある。そこで、この違いを統計的に検定した。これは片側の F 検定による。誤差分散の比と、検定結果を判断する p 値が表 15 に示されている。5%水準で判断すると、家庭満足度・家計満足度の平均の誤差分散は、性別年齢層化によって有意には減少していないものの、これら以外の 4 つの統計量、すなわち、教育程度、就労の比率と、これらの対数オッズ比、それに満足度変数同士の相関係数の誤差分散は、性別年齢層化をした方が有意に小さいことが明らかとなった。

5. まとめ

本稿では、乱数を用いたシミュレーションによって、標本抽出法における誤差の検討を行なった。シミュレーションで得られる標準誤差の値に妥当性があることを、比較的簡便に算出できる理論値との比較によって確認し、そののちに、地域層化を伴う 2 段抽出に地点内での性別年齢層化を加えることに、誤差分散を減少させる観点から一定の意義がある

ことを示した。

最後に、本稿の方法の限界を指摘しておきたい。本稿では、現実のデータを参照してより現実的なシミュレーションを目指したものの、「不自然」な設定も多い。例えば、第 1 に、本稿では同一地域内であれば、地点間では分布の差異を設けずにデータを生成している。具体的には、同一地域内であれば地点内の性別・年齢カテゴリ別の分布は全く同一であり、性別・年齢カテゴリ別の各変数の生成は同一の分布からの乱数による。このような地点間での同質性が現実に存在するとは考えられない。しかし、この設定は、本稿で 2 段抽出と段を設けない抽出とで顕著な差が見られない理由であると思われ、この点は抽出方法を議論する際に重要となるはずである。第 2 に、地域・性別・年齢別の人口の設定は、現実の分布を参考にしつつも、それを何らかの明確な基準で近似したというものではなく、ある種の恣意性をもって単純化したものである。単純化した理由の 1 つは、本稿で扱うどの抽出方法においても自動加重標本を得たいというものであったが[25]、抽出されるべき層別または地点別の個人数が自然数であるという制約のもとでは、厳密な意味で自動加重標本にならないのはむしろ通常の標本の性質である。第 3 に、非回答に伴う推定の偏りを考慮していない。本稿の主眼は性別年齢層化による標本誤差の減少を確かめることであったが、実際の調査では非回答による推定の偏りが深刻な問題となる。本稿は自動加重標本が得られる前提で推定の誤差を問題にしているが、実際には非回答の問題によってそもそも自動加重標本が得られないのである。しかも、本稿で仮想的母集団の生成に際して参考にしている JGSS データの分布自体が非回答による偏りにさらされていると考えられる。第 4 に、層化の際のカテゴリが少なすぎると思われる。地域層化の際に市郡規模による 3 カテゴリにしか分けないのは現実になされる方法にそぐわないし、年齢で層化する場合に 2 カテゴリにしか分けないのも全く不十分であろう。

このような「不自然」な設定を改め、現実にある、より複雑な状況をシミュレーションによって再現してみることは、実践的な知見を得る上で意味を有するであろう[26]。さらに、この再現の過程で、研究者は標本調査を「疑似体験」することができると言っても過言ではない。本稿のようなシミュレーションは、標本誤差の検討のみならず、調査方法そのものへの理解を深め共有するという意味でも有益であると考える。

---

[25] 注 5、注 10 を参照されたい。

[26] より現実的なシミュレーションを目指すにあたって、以下のような方法が検討に値すると思われる。第 1 に、仮想的母集団の生成にあたり、JGSS などの既存の標本調査データから、リサンプリング手法によって母集団を構成するという方法である。その際、既存データの非回答による偏りを補正しておくことが望ましいであろう。第 2 に、仮想的母集団の生成にあたり、国勢調査のデータを参照するという方法が挙げられる。国勢調査は全数調査であり、このデータは母集団データとして用いるのに極めて適当であろう。第 3 に、本稿では計算時間の短縮のために母集団サイズを 100 万としたが、より現実に近い数値の設定が検討に値する。

引用文献

Agresti, Alan, 2002, *Categorical Data Analysis (Second Edition)*, John Wiley & Sons, Inc.

Cochran, William G., 1977, *Sampling Techniques (Third Edition)*, John Wiley & Sons, Inc.

Groves, Robert M., Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau, 2004, *Survey Methodology*, John Wiley & Sons, Inc.

南風原朝和，2002,『心理統計学の基礎　統合的理解のために』有斐閣.

盛山和夫，2004,『社会調査法入門』有斐閣.

総務省統計局編，2005,『平成 12 年国勢調査最終報告書　日本の人口（資料編)』総務省統計局.

表1. JGSSデータの分析で使用する変数

変数	JGSS変数名	内容･加工方法
地域	SIZE	SIZEをそのまま用いる。「大都市」「その他の市」「郡部」の3カテゴリに分かれている。
性別	SEXA	SEXAをそのまま用いる。「男性」「女性」の2カテゴリに分かれている。
年齢	AGEB	「20-54歳」と「55-89歳」の2カテゴリに分ける。
教育程度	XXLSTSCH	最後に通った（中退を含む）、または在学中の学校の程度で、「高等教育」と「中等教育（以下）」とに分ける。旧制学校に関しては、尋常小･高等小･中学･高等女学校･実業学校を「中等教育（以下）」とする。
就労	XJOB1WK	「先週仕事をした」「先週仕事をすることになっていた」という場合に「有職」、「仕事をしていない」という場合に「無職」とする。
家庭満足度	ST5LIFEY	ST5LIFEYをそのまま用いる。1から5までの順序尺度を間隔尺度と見なす。
家計満足度	ST5ECNY	ST5ECNYをそのまま用いる。1から5までの順序尺度を間隔尺度と見なす。

表2. JGSSデータにおける対象者の居住地域（市郡規模）の分布

居住地域（市郡規模）	n	(%)
大都市	660	(18.5)
その他の市	2,043	(57.1)
郡部	873	(24.4)
計	3,576	(100.0)

Note: 「大都市」とは、札幌市、仙台市、さいたま市、千葉市、東京都区部、横浜市、川崎市、名古屋市、京都市、大阪市、神戸市、広島市、北九州市、福岡市を指す。JGSSデータセットのうち、本稿が分析対象とする変数についてリストワイズで欠損ケースを定義し、有効なケースのみを扱う。この欠損処理はJGSSデータを用いる以下の表すべてで同様である。

Source: JGSS-2003.

表3. JGSSデータにおける地域別の性別・年齢の分布

居住地域	20-54歳		55-89歳		計
	男性	女性	男性	女性	
大都市	147 (22.3)	215 (32.6)	136 (20.6)	162 (24.6)	660 (100.0)
その他の市	423 (20.7)	618 (30.3)	445 (21.8)	557 (27.3)	2,043 (100.0)
郡部	178 (20.4)	217 (24.9)	223 (25.5)	255 (29.2)	873 (100.0)
計	748 (20.9)	1,050 (29.4)	804 (22.5)	974 (27.2)	3,576 (100.0)

Note: 括弧なしの数値は実数、括弧付きの数値は行方向のパーセントを示す。

Source: JGSS-2003.

表4. JGSSデータにおける《地域・性別・年齢》カテゴリ別の教育程度と就労の分布

居住地域	年齢 性別	中等教育		高等教育		計
		有職	無職	有職	無職	
大都市	20-54 男性	36.1	1.4	55.1	7.5	100.0
	20-54 女性	30.2	16.7	34.0	19.1	100.0
	55-89 男性	36.0	34.6	15.4	14.0	100.0
	55-89 女性	24.7	58.6	7.4	9.3	100.0
その他の市	20-54 男性	52.2	3.8	40.7	3.3	100.0
	20-54 女性	38.2	16.8	27.0	18.0	100.0
	55-89 男性	35.3	44.3	11.2	9.2	100.0
	55-89 女性	27.5	62.1	3.9	6.5	100.0
郡部	20-54 男性	64.6	3.4	29.8	2.2	100.0
	20-54 女性	44.7	18.9	26.7	9.7	100.0
	55-89 男性	47.1	42.2	6.7	4.0	100.0
	55-89 女性	25.5	68.6	1.6	4.3	100.0
計		37.9	32.4	20.4	9.3	100.0

Note: 数値は行方向のパーセントを示す。

Source: JGSS-2003.

表5. JGSSデータにおける《地域・性別・年齢》カテゴリ別の教育程度と就労との連関

居住地域	年齢	性別	有職の対数オッズ		対数オッズ比（高等教育 / 中等教育）
			中等教育	高等教育	
大都市	20-54	男性	3.277	1.997	-1.281
	20-54	女性	0.591	0.577	-0.014
	55-89	男性	0.042	0.100	0.058
	55-89	女性	-0.865	-0.223	0.642
その他の市	20-54	男性	2.626	2.508	-0.117
	20-54	女性	0.819	0.408	-0.411
	55-89	男性	-0.227	0.198	0.425
	55-89	女性	-0.816	-0.492	0.324
郡部	20-54	男性	2.953	2.584	-0.369
	20-54	女性	0.861	1.016	0.155
	55-89	男性	0.111	0.511	0.400
	55-89	女性	-0.990	-1.012	-0.021
全体			0.157	0.782	0.625

Note: 表中の対数オッズ比は、高等教育を受けた者が中等教育を受けた者よりも相対的に有職となりやすい度合いを示す。

Source: JGSS-2003.

表6. JGSSデータにおける《地域・性別・年齢》カテゴリ別の家庭満足度と家計満足度の分布

居住地域	年齢 性別	家庭満足度		家計満足度		相関係数
		平均	標準偏差	平均	標準偏差	
大都市	20-54 男性	2.333	1.075	3.109	1.080	0.488
	20-54 女性	2.447	1.053	3.121	1.201	0.415
	55-89 男性	2.331	1.047	2.941	0.987	0.535
	55-89 女性	2.340	1.004	2.914	1.066	0.556
その他の市	20-54 男性	2.383	1.012	3.116	1.081	0.393
	20-54 女性	2.405	0.988	3.199	1.082	0.438
	55-89 男性	2.407	0.999	2.987	1.029	0.525
	55-89 女性	2.370	1.012	2.932	1.145	0.585
郡部	20-54 男性	2.534	1.126	3.303	1.035	0.365
	20-54 女性	2.424	1.002	3.171	1.047	0.434
	55-89 男性	2.296	0.960	3.036	1.056	0.363
	55-89 女性	2.278	1.045	2.882	1.091	0.616
全体		2.383	1.017	3.061	1.089	0.481

Note: 表中の「標準偏差」は偏差平方和を(n-1)で割ったものの平方根で定義される。

Source: JGSS-2003.

表7. 仮想的母集団における居住地域（市郡規模）の分布

居住地域（市郡規模）	N	(%)
大都市	200,000	(20.0)
その他の市	550,000	(55.0)
郡部	250,000	(25.0)
計	1,000,000	(100.0)

Note: これらの数値は仮想的母集団の生成の際に設定される数値である。乱数によって生成された結果ではない。

表8. 仮想的母集団における地域別の性別・年齢の分布

居住地域	20-54歳		55-89歳		計
	男性	女性	男性	女性	
大都市	40,000 (20.0)	70,000 (35.0)	40,000 (20.0)	50,000 (25.0)	200,000 (100.0)
その他の市	110,000 (20.0)	165,000 (30.0)	110,000 (20.0)	165,000 (30.0)	550,000 (100.0)
郡部	50,000 (20.0)	62,500 (25.0)	62,500 (25.0)	75,000 (30.0)	250,000 (100.0)
計	200,000 (20.0)	297,500 (29.8)	212,500 (21.3)	290,000 (29.0)	1,000,000 (100.0)

Note: 括弧なしの数値は実数、括弧付きの数値は行方向のパーセントを示す。これらの数値は仮想的母集団の生成の際に設定される数値である。乱数によって生成された結果ではない。

表9. 仮想的母集団における《地域・性別・年齢》カテゴリ別の教育程度と就労の分布

居住地域	年齢	性別	中等教育		高等教育		計
			有職	無職	有職	無職	
大都市	20-54	男性	36.1	1.4	55.1	7.4	100.0
	20-54	女性	30.4	16.7	33.7	19.2	100.0
	55-89	男性	36.1	34.7	15.1	14.1	100.0
	55-89	女性	24.6	58.8	7.5	9.1	100.0
その他の市	20-54	男性	52.1	3.8	40.8	3.3	100.0
	20-54	女性	38.1	16.9	27.1	17.9	100.0
	55-89	男性	35.2	44.4	11.1	9.3	100.0
	55-89	女性	27.4	62.1	3.9	6.6	100.0
郡部	20-54	男性	64.9	3.4	29.5	2.2	100.0
	20-54	女性	44.5	18.7	27.1	9.7	100.0
	55-89	男性	46.6	42.5	6.8	4.0	100.0
	55-89	女性	25.4	68.7	1.6	4.3	100.0
計			37.5	33.0	20.1	9.4	100.0

Note: 数値は行方向のパーセントを示す。これらの数値は、教育程度と就労の2変量の分布を各《地域・性別・年齢》カテゴリで表4にあるような確率で設定して、乱数によって生成したデータを集計したものである。

表10. 仮想的母集団における《地域・性別・年齢》カテゴリ別の教育程度と就労との連関

居住地域	年齢	性別	有職の対数オッズ		対数オッズ比（高等教育 / 中等教育）
			中等教育	高等教育	
大都市	20-54	男性	3.260	2.002	-1.258
	20-54	女性	0.598	0.565	-0.033
	55-89	男性	0.040	0.070	0.030
	55-89	女性	-0.871	-0.185	0.687
その他の市	20-54	男性	2.628	2.508	-0.119
	20-54	女性	0.811	0.413	-0.398
	55-89	男性	-0.232	0.179	0.411
	55-89	女性	-0.819	-0.524	0.296
郡部	20-54	男性	2.944	2.576	-0.368
	20-54	女性	0.865	1.023	0.159
	55-89	男性	0.091	0.527	0.436
	55-89	女性	-0.996	-1.017	-0.021
全体			0.126	0.761	0.634

Note: 表中の対数オッズ比は、高等教育を受けた者が中等教育を受けた者よりも相対的に有職となりやすい度合いを示す。

表11. 仮想的母集団における《地域・性別・年齢》カテゴリ別の家庭満足度と家計満足度の分布

居住地域	年齢 性別	家庭満足度		家計満足度		相関係数
		平均	標準偏差	平均	標準偏差	
大都市	20-54 男性	2.337	1.078	3.116	1.087	0.489
	20-54 女性	2.444	1.051	3.116	1.206	0.413
	55-89 男性	2.332	1.046	2.942	0.989	0.535
	55-89 女性	2.332	1.001	2.910	1.067	0.559
その他の市	20-54 男性	2.381	1.013	3.111	1.081	0.393
	20-54 女性	2.405	0.990	3.205	1.082	0.440
	55-89 男性	2.401	0.998	2.984	1.026	0.521
	55-89 女性	2.368	1.011	2.933	1.147	0.585
郡部	20-54 男性	2.532	1.124	3.307	1.031	0.361
	20-54 女性	2.425	1.001	3.171	1.046	0.435
	55-89 男性	2.295	0.960	3.042	1.056	0.361
	55-89 女性	2.278	1.044	2.882	1.088	0.617
全体		2.380	1.019	3.059	1.093	0.482

Note: 表中の「標準偏差」は偏差平方和を(n-1)で割ったものの平方根で定義される。これらの数値は、2つの満足度の2変量分布を各《地域・性別・年齢》カテゴリで表6にあるような平均・標準偏差・相関係数で設定して、乱数によって生成したデータを集計したものである。

表12. 仮想的母集団からの種々の方法による抽出シミュレーションの結果

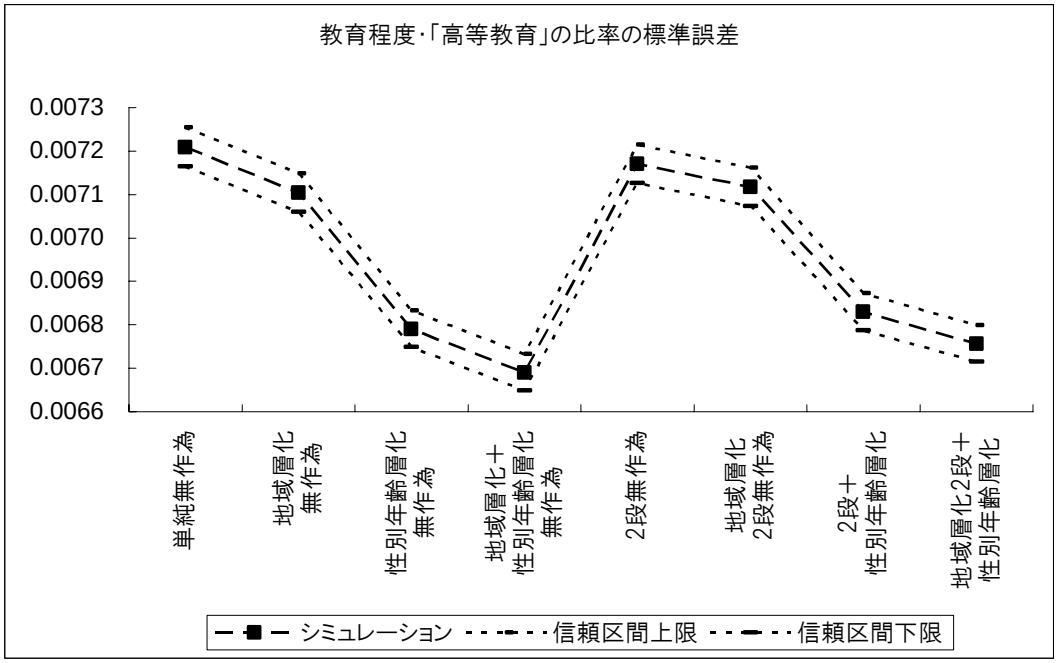
変数 / 統計量		母集団（対応する母数）	単純無作為抽出法		地域層化無作為抽出法		性別年齢層化無作為抽出法		地域層化＋性別年齢層化無作為抽出法	
			平均	（標準誤差）	平均	（標準誤差）	平均	（標準誤差）	平均	（標準誤差）
教育程度	「高等教育」の比率	0.2946	0.2947	(0.0072)	0.2947	(0.0071)	0.2946	(0.0068)	0.2947	(0.0067)
就労	「有職」の比率	0.5757	0.5757	(0.0078)	0.5757	(0.0078)	0.5757	(0.0069)	0.5757	(0.0069)
教育程度・就労	対数オッズ比	0.6345	0.6355	(0.0728)	0.6352	(0.0730)	0.6353	(0.0723)	0.6355	(0.0725)
家庭満足度	平均	2.3805	2.3806	(0.0161)	2.3805	(0.0161)	2.3803	(0.0161)	2.3804	(0.0160)
家計満足度	平均	3.0589	3.0589	(0.0173)	3.0588	(0.0172)	3.0588	(0.0171)	3.0588	(0.0172)
家庭満足度・家計満足度	相関係数	0.4821	0.4821	(0.0123)	0.4821	(0.0122)	0.4821	(0.0121)	0.4820	(0.0121)

Note: 各方法で、4,000名からなる標本を50,000回抽出した。表中の対数オッズ比は、高等教育を受けた者が相対的に有職となりやすい傾向を示す。表中の標準誤差は、抽出を繰り返した際の統計量の偏差平方和を(N_{iter}-1)で割ったものの平方根で定義される。ただしN_{iter}は抽出回数（ここでは50,000）を表す。

表12. *(Continued)*

変数 / 統計量		母集団（対応する母数）	2段無作為抽出法		地域層化2段無作為抽出法		2段抽出法＋性別年齢層化		地域層化2段抽出法＋性別年齢層化	
			平均	（標準誤差）	平均	（標準誤差）	平均	（標準誤差）	平均	（標準誤差）
教育程度	「高等教育」の比率	0.2946	0.2946	(0.0072)	0.2946	(0.0071)	0.2946	(0.0068)	0.2946	(0.0068)
就労	「有職」の比率	0.5757	0.5757	(0.0078)	0.5757	(0.0078)	0.5757	(0.0071)	0.5756	(0.0069)
教育程度・就労	対数オッズ比	0.6345	0.6354	(0.0712)	0.6348	(0.0723)	0.6352	(0.0704)	0.6353	(0.0714)
家庭満足度	平均	2.3805	2.3804	(0.0159)	2.3805	(0.0160)	2.3806	(0.0159)	2.3805	(0.0160)
家計満足度	平均	3.0589	3.0588	(0.0173)	3.0588	(0.0173)	3.0590	(0.0171)	3.0590	(0.0172)
家庭満足度・家計満足度	相関係数	0.4821	0.4821	(0.0123)	0.4821	(0.0123)	0.4821	(0.0122)	0.4821	(0.0122)

図1. 仮想的母集団からの種々の方法による抽出シミュレーションの結果



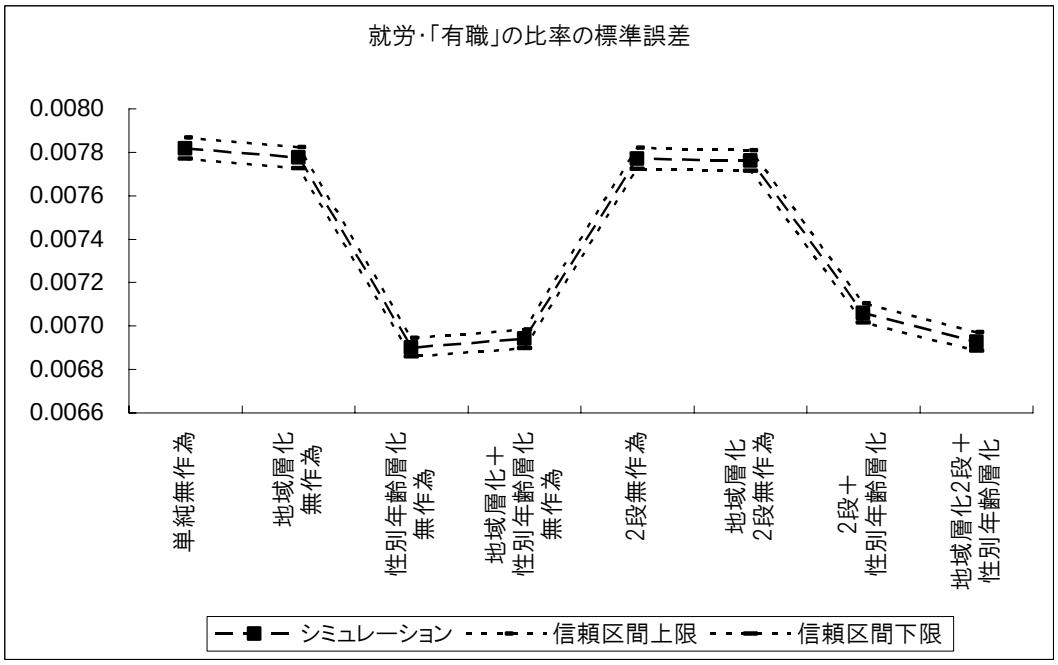
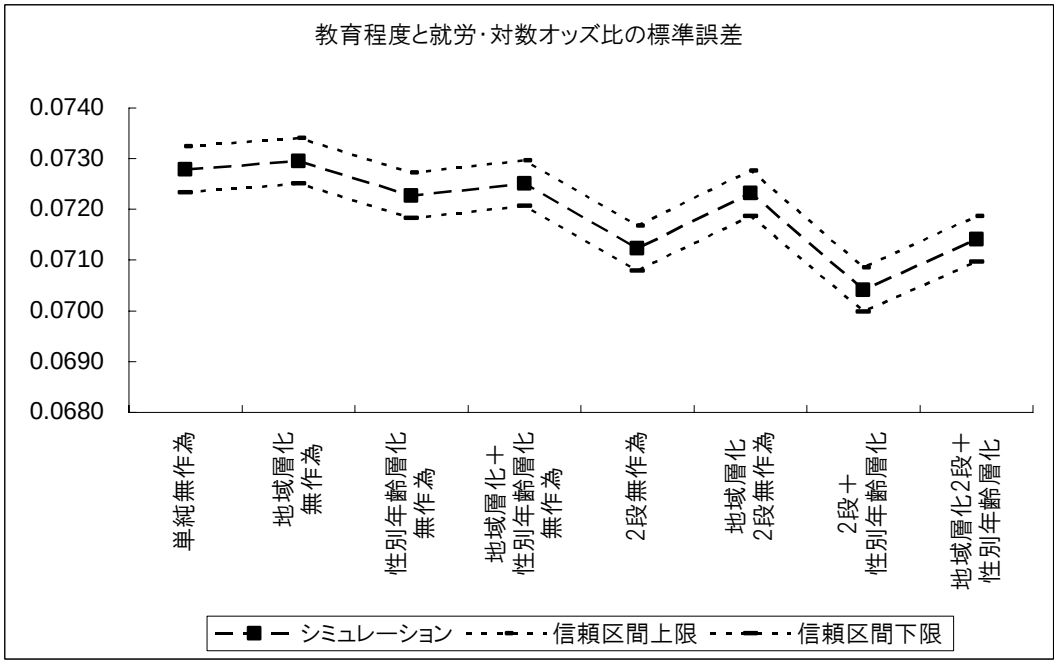


図1. *(continued)*



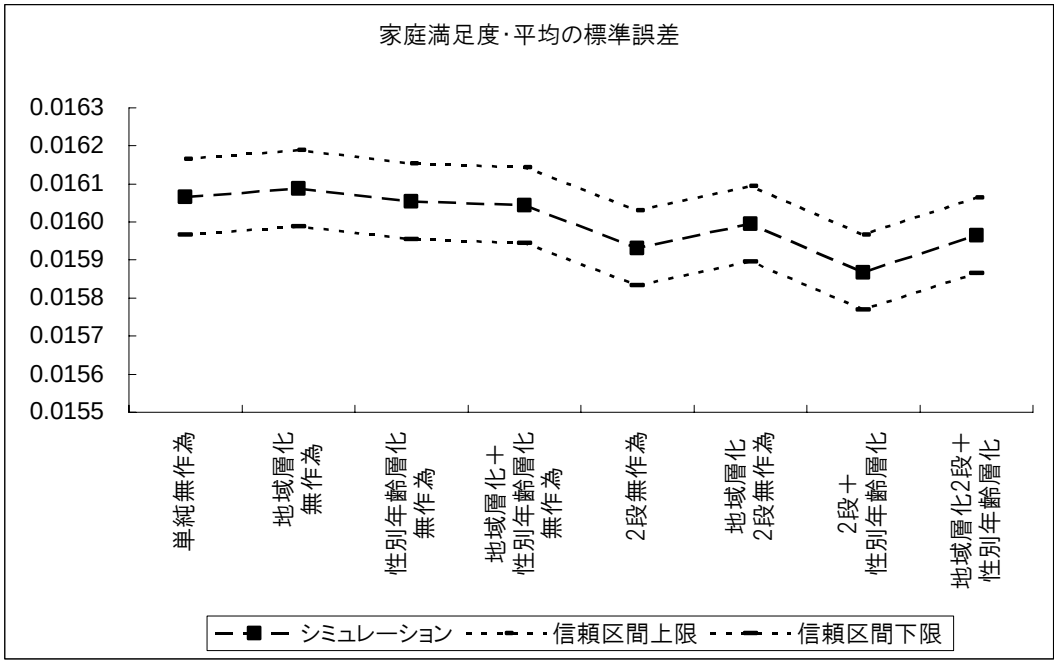
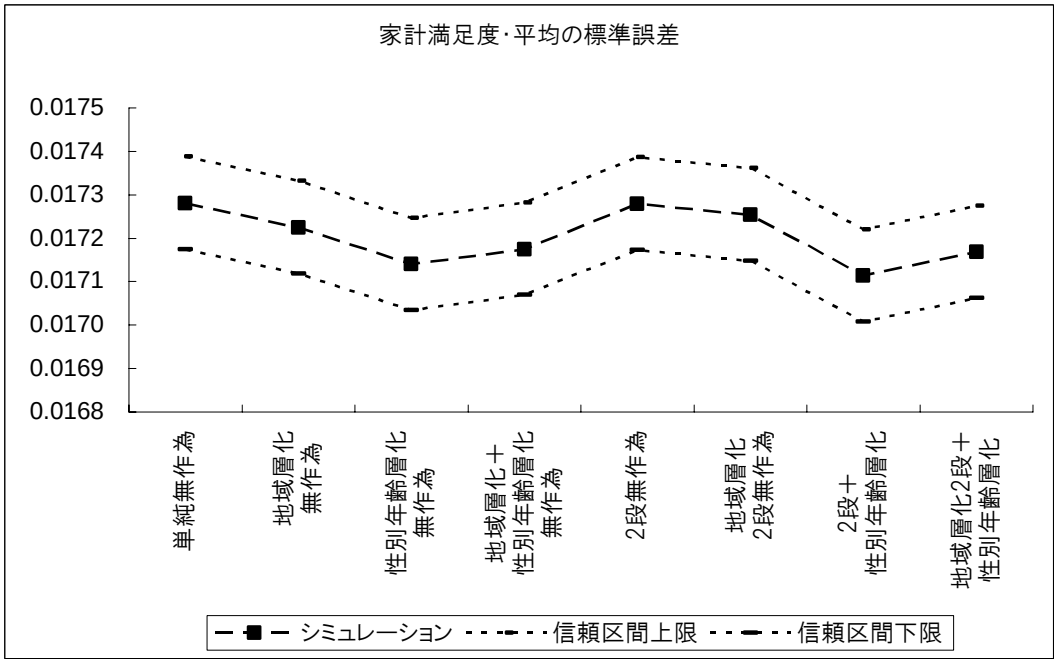
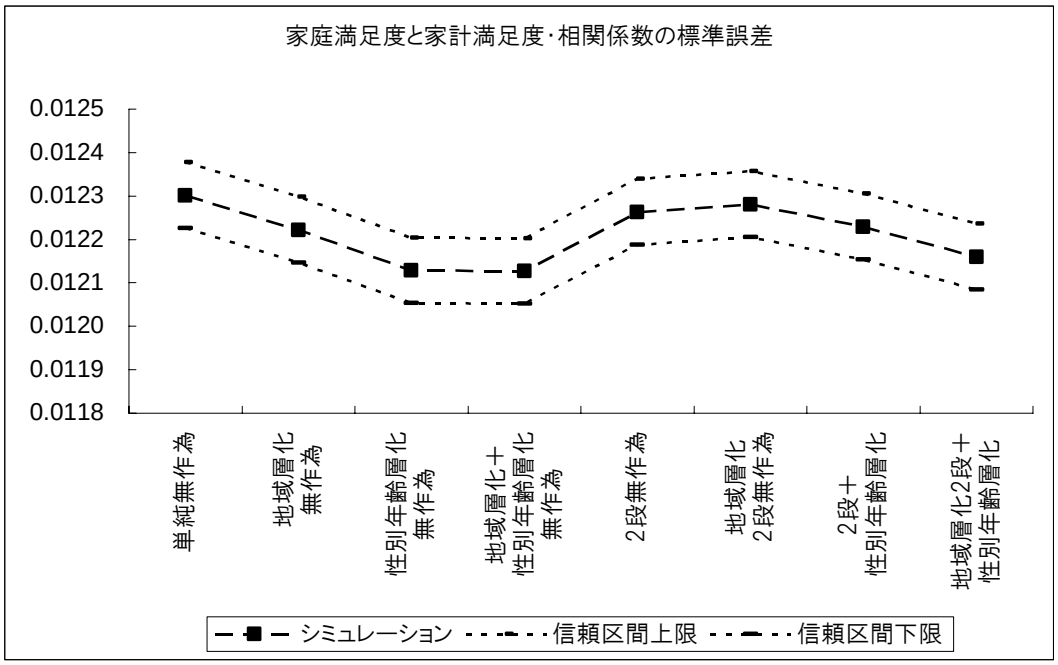


図1. *(continued)*





Note: 各抽出方法で、サイズ4,000の標本を50,000回抽出して統計量を計算した結果が示されている。図示された標準誤差の信頼区間の上下限は、「真の」誤差分散を母分散と考えてこれをシミュレーション結果から推定する際の、95%信頼区間の上下限値の平方根の値を示す。ここでは統計量の標本分布が正規分布であると仮定している。この算出には自由度49,999のχ^2分布を用いた。

表13. 単純無作為抽出法における理論値とシミュレーション結果

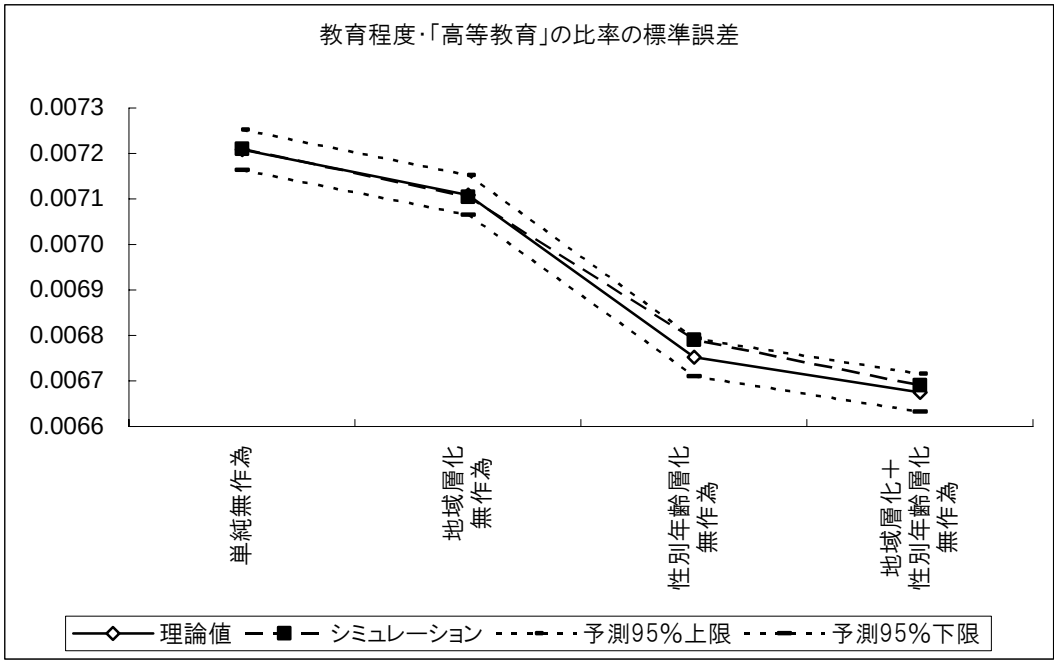
変数 / 統計量		母集団		単純無作為抽出法		
		対応する母数	（標準偏差）	シミュレーション結果 平均	（標準誤差）	理論値（標準誤差）
教育程度	「高等教育」の比率	0.2946	-	0.2947	(0.0072)	(0.0072)
就労	「有職」の比率	0.5757	-	0.5757	(0.0078)	(0.0078)
教育程度・就労	対数オッズ比	0.6345	-	0.6355	(0.0728)	(0.0730)
家庭満足度	平均	2.3805	(1.0191)	2.3806	(0.0161)	(0.0161)
家計満足度	平均	3.0589	(1.0931)	3.0589	(0.0173)	(0.0172)
家庭満足度・家計満足度	相関係数	0.4821	-	0.4821	(0.0123)	(0.0121)

表14. 層化無作為抽出法における理論値と抽出シミュレーション結果

変数 / 統計量		単純無作為抽出法		地域層化 無作為抽出法		性別年齢層化 無作為抽出法		地域層化 ＋性別年齢層化 無作為抽出法	
		標準誤差		標準誤差		標準誤差		標準誤差	
		シミュレーション	理論値	シミュレーション	理論値	シミュレーション	理論値	シミュレーション	理論値
教育程度	「高等教育」の比率	0.00721	0.00721	0.00710	0.00711	0.00679	0.00675	0.00669	0.00667
就労	「有職」の比率	0.00782	0.00781	0.00777	0.00780	0.00690	0.00695	0.00694	0.00694
家庭満足度	平均	0.01607	0.01608	0.01609	0.01608	0.01605	0.01607	0.01604	0.01606
家計満足度	平均	0.01728	0.01725	0.01722	0.01725	0.01714	0.01715	0.01717	0.01714

Note: 標準誤差の理論値の算出はCochran (1977: 89-93)による。「シミュレーション」の標準誤差はサイズ4,000の標本を50,000回抽出して得たものである。

図2. 層化無作為抽出法における理論値と抽出シミュレーション結果



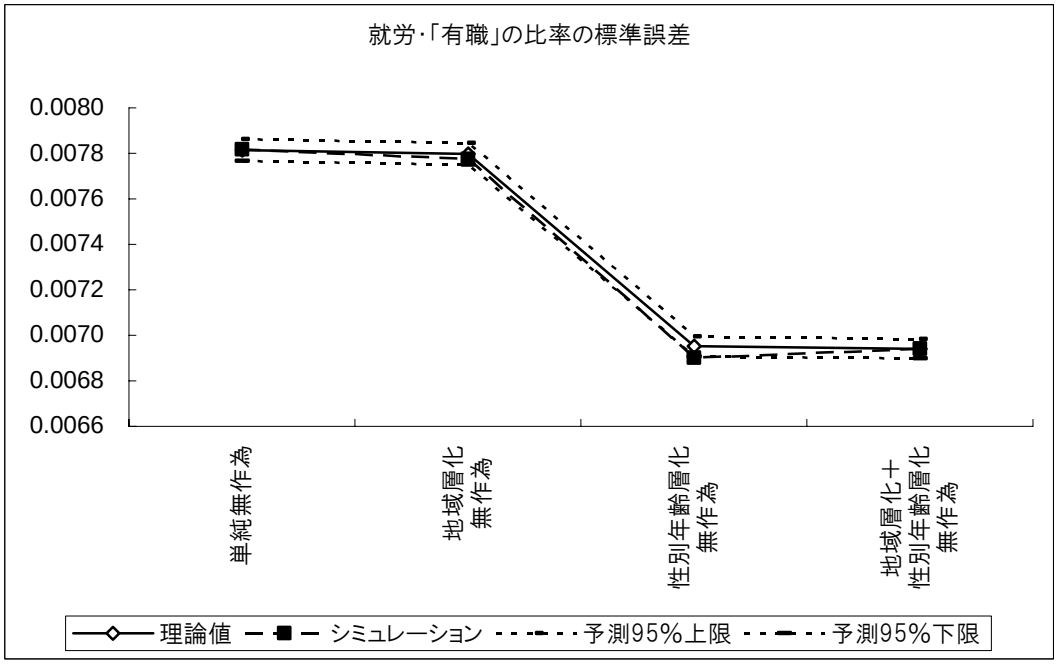
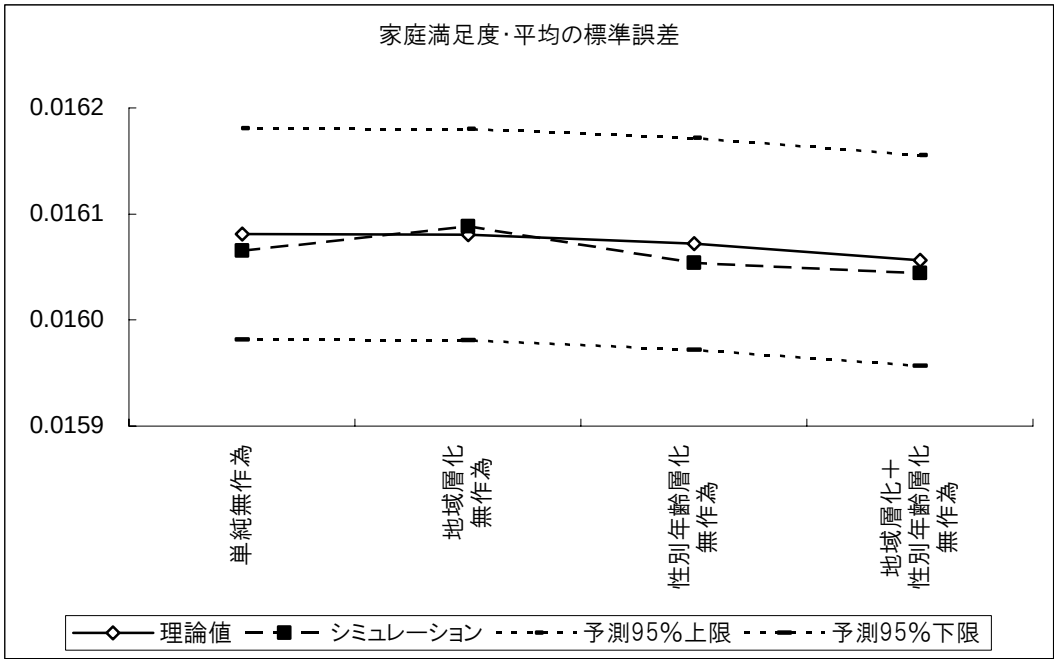
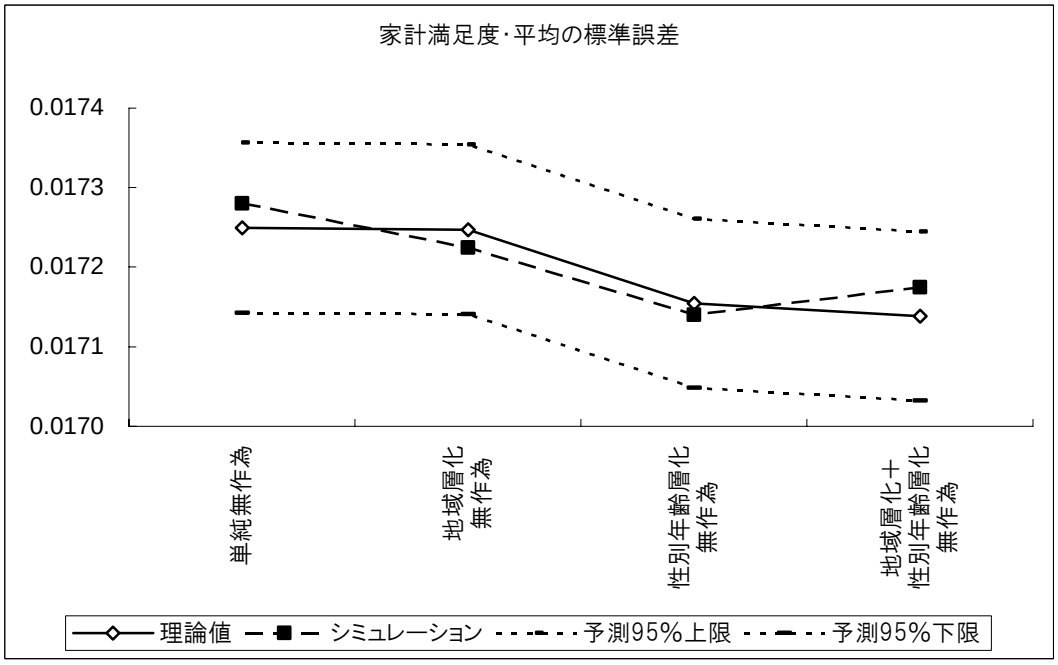


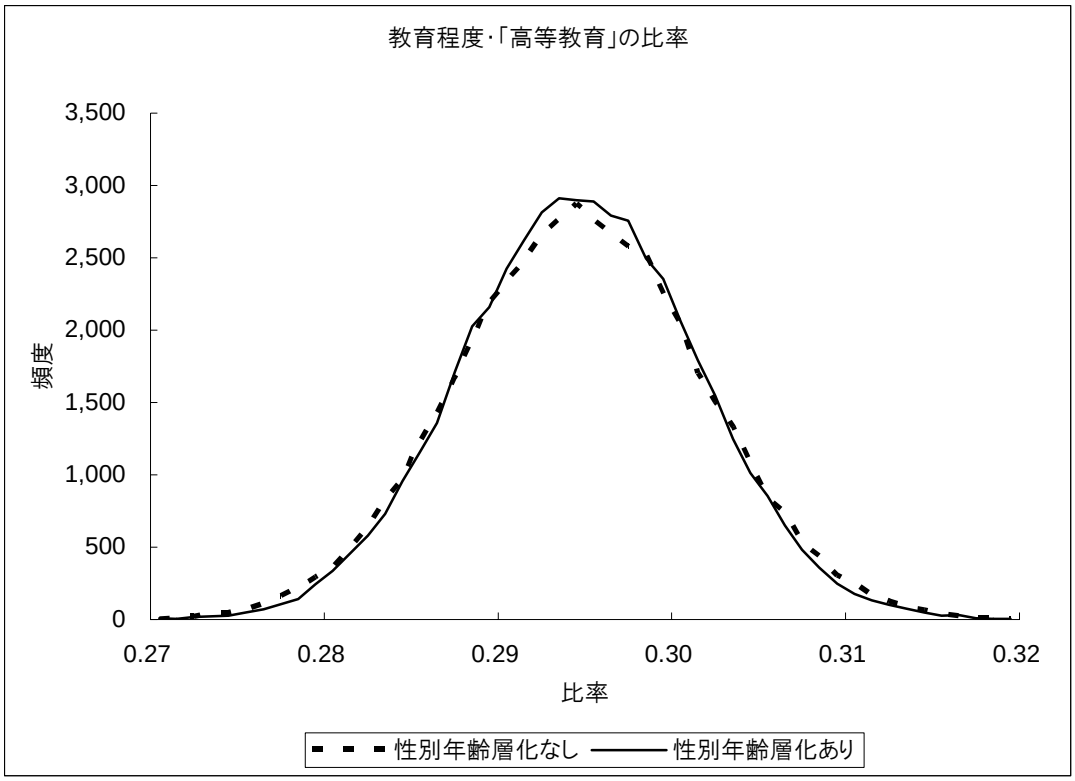
図2. *(continued)*





Note: 各抽出方法で、サイズ4,000の標本を50,000回抽出して統計量を計算した結果が示されている。図示された予測95%上下限は、誤差分散の理論値を母分散と考えた場合の、50,000抽出するシミュレーションでの誤差分散の標本分布の、上側97.5%点と上側2.5%点の平方根を示す。ここでは統計量の標本分布が正規分布であると仮定している。この算出には自由度49,999のχ^2分布を用いた。

図3. 地域層化2段抽出法における地点内性別年齢層化の有無による比較:ヒストグラム



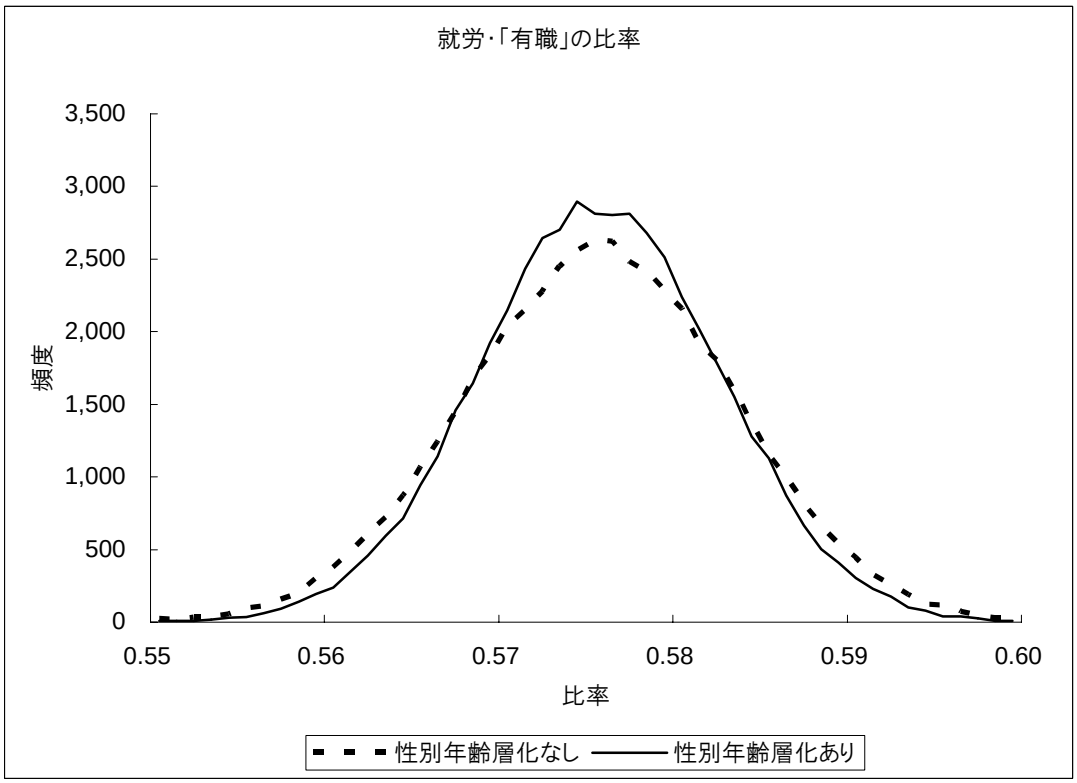
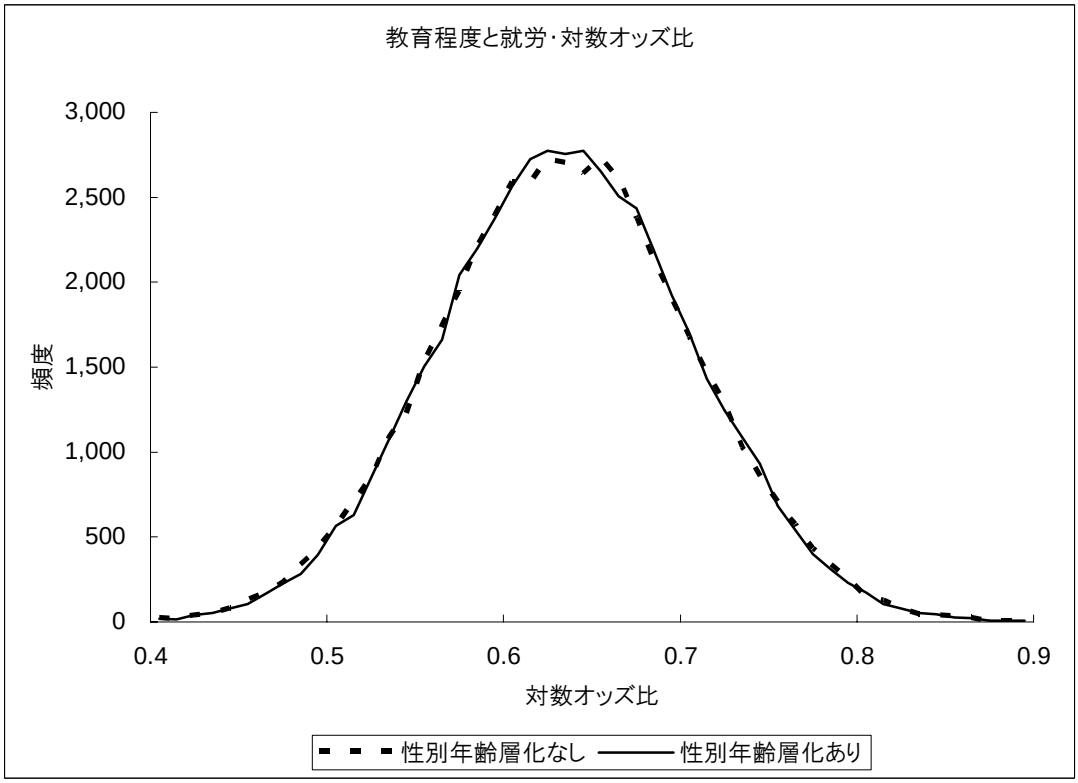


図3. *(continued)*



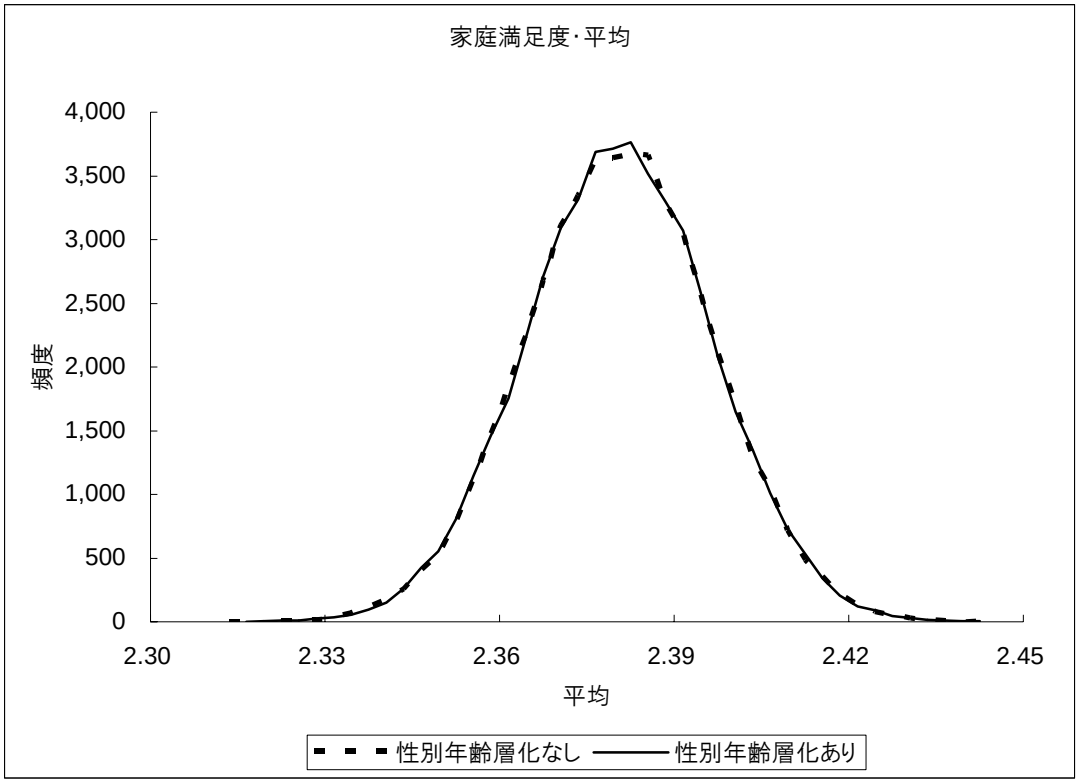
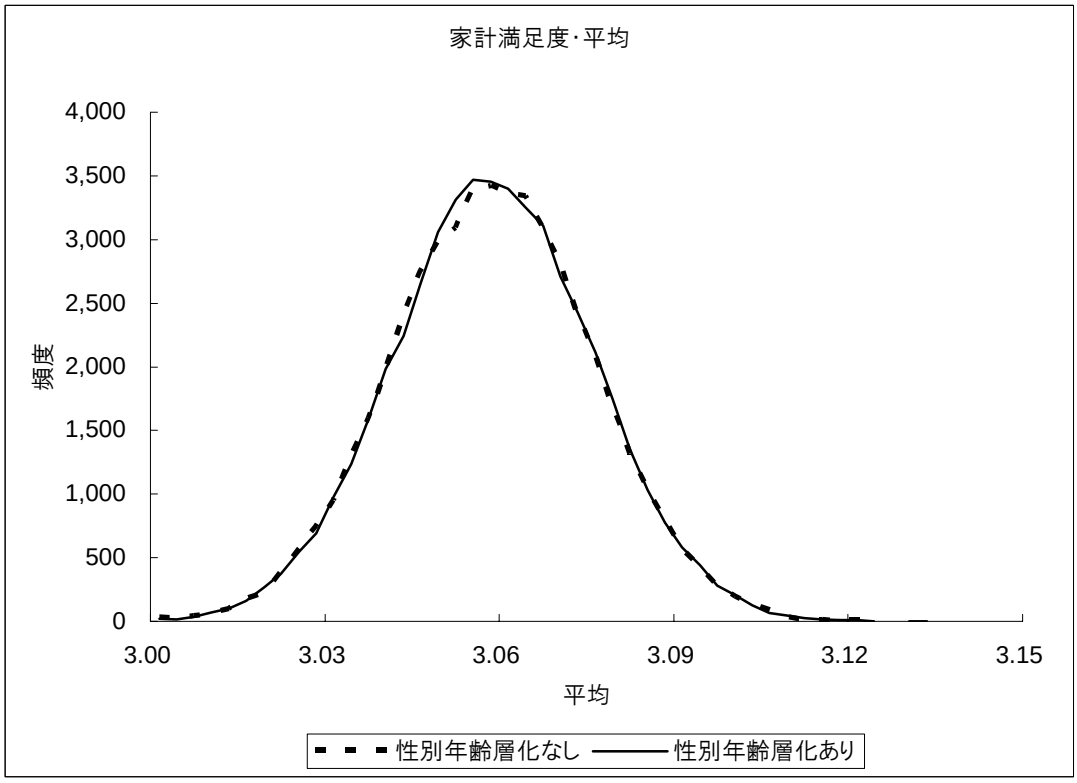
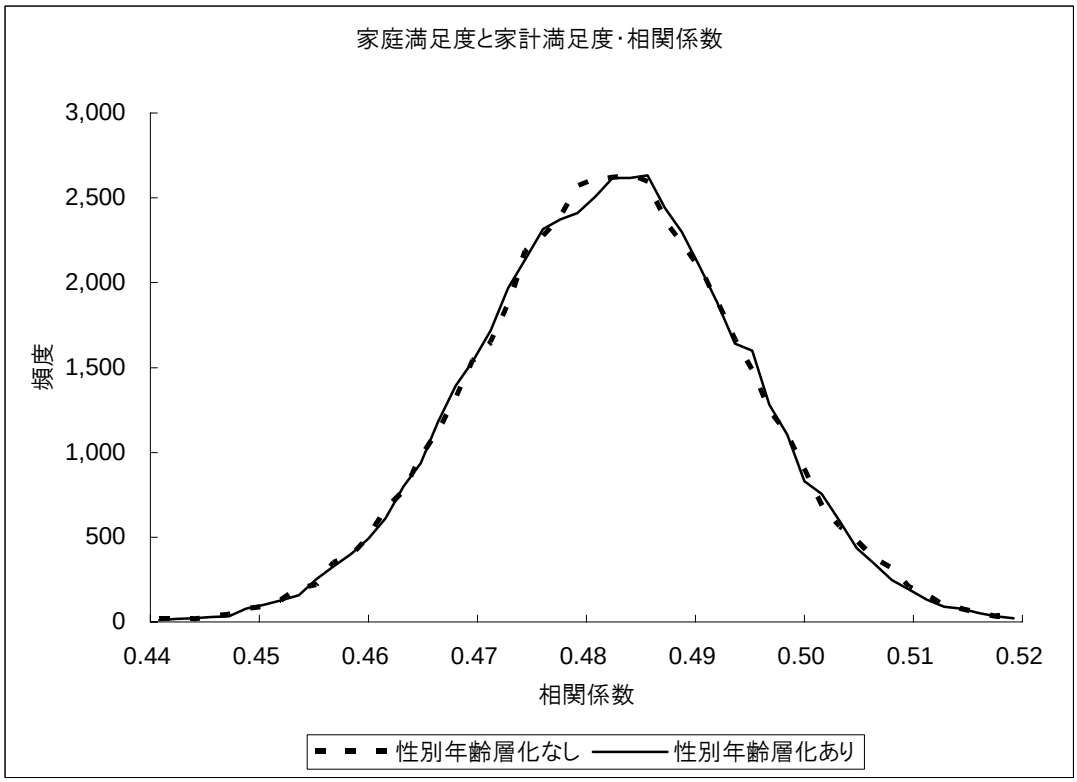


図3. *(continued)*





Note: 各抽出方法で、サイズ4,000の標本を50,000回抽出して統計量を計算した結果が示されている。ヒストグラムの作成に際して、図の横軸に示されている範囲を50階級に分割して各階級の頻度をプロットした。

表15. 地域層化2段抽出法における地点内性別年齢層化の有無による比較：F検定

変数 / 統計量		標準誤差		誤差分散の比	p値（片側）
		性別年齢層化なし	性別年齢層化あり		
教育程度	「高等教育」の比率	0.00712	0.00676	1.110	0.000
就労	「有職」の比率	0.00776	0.00693	1.255	0.000
教育程度・就労	対数オッズ比	0.07232	0.07141	1.025	0.003
家庭満足度	平均	0.01599	0.01596	1.004	0.336
家計満足度	平均	0.01725	0.01717	1.010	0.130
家庭満足度・家計満足度	相関係数	0.01228	0.01216	1.020	0.014

Note: 各抽出方法で、サイズ4,000の標本を50,000回抽出して統計量を計算し、標準誤差を算出した。検定に際しては統計量の標本分布が正規分布であると仮定している。F検定では自由度（49,999 , 49,999）のF分布を用いた。

東京大学社会科学研究所パネル調査プロジェクトについて

労働市場の構造変動、急激な少子高齢化、グローバル化の進展などにともない、日本社会における就業、結婚、家族、教育、意識、ライフスタイルのあり方は大きく変化を遂げようとしている。これからの日本社会がどのような方向に進むのかを考える上で、現在生じている変化がどのような原因によるものなのか、あるいはどこが変化してどこが変化していないのかを明確にすることはきわめて重要である。

本プロジェクトは、こうした問題をパネル調査の手法を用いることによって、実証的に解明することを研究課題とするものである。このため社会科学研究所では、若年パネル調査、壮年パネル調査、高卒パネル調査の３つのパネル調査を実施している。

本プロジェクトの推進にあたり、以下の資金提供を受けた。記して感謝したい。

文部科学省・独立行政法人日本学術振興会科学研究費補助金
基盤研究 S：平成 18 年度～平成 22 年度

厚生労働科学研究費補助金
政策科学推進研究：平成 16 年度～平成 18 年度

奨学寄付金
株式会社アウトソーシング（代表取締役社長・土井春彦、本社・静岡市）：2006 年～

東京大学社会科学研究所パネル調査プロジェクト
ディスカッションペーパーシリーズについて

東京大学社会科学研究所パネル調査プロジェクトディスカッションペーパーシリーズは、東京大学社会科学研究所におけるパネル調査プロジェクト関連の研究成果を、速報性を重視し暫定的にまとめたものである。

東京大学社会科学研究所パネル調査プロジェクト　ディスカッションペーパーシリーズ

No.1　　山本耕資　　標本調査における性別・年齢による層化の効果：100 万人シミュレーション（2007 年 4 月発行）

東京大学社会科学研究所 パネル調査プロジェクト
http://ssjda.iss.u-tokyo.ac.jp/panel/