た出産1年前に有職か無職かの対比の効果が非常に強いため、有職の場合のみに限った分析も行った。

（2）の子育てにかかわるテーマでは、②の育児不安の分析で、まず関連項目を類型化し、就業別に上位項目を調べ、子どもが6ヶ月から4歳半と成長する中で、母親の育児に関する心理はどう変化するのかを母親の就業形態別に調べた。収入別や母親の国籍別による分析も行った。経済的負担感と子育て費用に注目した③では、子育て費用の実態の変化と、母親の就業形態別、年収別の変化、習い事の実態をみた。さらに子育て費用が所得に占める割合が高いのは誰か、負担感が高いのは誰かを見極めるため、第5回のデータを用いて、保育料、父や祖父母との同居の状況、母親の就業状況、収入、習い事の有無、保育料負担感、子どもが病気がちという変数を独立変数にし、従属変数を「子育てで出費がかさむ」（経済的負担感）ならびに「収入に占める子育て費用の割合が30%以上」（子育て費用）としてロジスティック回帰分析を行った。④では子ども観について、関連項目のコレスポンデンス分析を行い、各ケースを数量化得点に従って4象限に分類し、4分類と各ケースの属性の関係を探った。さらに各ケースが4分類それぞれに当てはまるか否かを従属変数としたロジスティック回帰分析を行い、子どもの性別、回答者父のみダミー　回答者父母のみダミー　回答者その他ダミー、きょうだいダミー、祖父母同居ダミー、都市規模、母親の職業、父親の職業、親の学歴、父母の年齢の偏回帰係数を検討した。応用として子ども観の4分類と「しつけ」との関連の分析も行った。

テーマ（3）の子どもの置かれた社会経済状況に関しては、⑤で、脱落と所得階層の関係を確認し、各年度の相対的貧困基準（各年ごとの等価世帯所得の中央値の50%）と絶対的貧困基準（1年目の等価世帯所得の中央値の50%）による貧困率を求め、その推移を確認し、継続状況による4つの類型化を試みた（固定貧困層（4回の調査時点すべてにて貧困基準未満）、慢性的貧困層（2回か3回の時点で貧困基準未満）、一時的貧困層（1回のみ貧困基準未満）、安定層（貧困基準未満なし）。次に、各年の貧困世帯を合わせたケースを用いて、貧困世帯が次の調査時点において貧困であるか否か（貧困を脱出したか否か）を被説明変数とするロジスティック回帰分析を行った。次に、第一回の所得データ（2000年）に所得が貧困基準未満であった世帯を対象とし、貧困脱出までの年数を被説明変数とし、説明変数に世帯類型（二親世帯、母子世帯、父子世帯）、母親年齢、父親年齢、母親学歴、父親学歴を用いたサバイバル分析を行った。

次に、貧困の影響として⑥で0歳から4歳までの貧困経験年数（0～4年）が、子どもの成長や健康に及ぼす影響をOLS分析とロジスティック分析の手法を用いて推計した。4歳時点の子どもの身長・体重（成長）と、4歳時点の通院経験・入院経験の有無を被説明変数とした。「健康ショック」（未熟児で生まれる、低体重で生まれる）の影響が、貧困世帯の子どももそうでない子どもにとって異なるかについてもOLS分析を用いて推計した。

最後の政策評価へデータの活用の検討では、⑦において「子ども・子育て応援プラン」の「仕事と家庭の両立支援と働き方の見直し」「子育ての新たな支え合いと連帯」に関して検討した。政策の成果を測定するひとつの方法として政策手段に対する福祉諸施策の関係や、政策が機能する環境、インプットと成果との結びつき等を示すうえで有用な福祉の生産」(production of welfare)モデルを参考とした。地域間格差の先行研究の検討をふまえ、都道府県の出

生率の変化割合（1995～2005 年）と出生率(2005)を軸に４つのクラスターに分け、設問やサンプル数の関係から、特に女性の就業や保育サービス利用率の部分を中心に、同地域別に数値を出し検討した。

C．研究成果

各サブ分析の結果、以下のような結果が見出された。テーマ（１）(①) の出産後の再就労の分析では出産から４年６ヶ月後までに６割以上の母親が再就労し、出産１年前に職についていた場合は、子どもが１歳６ヶ月の時点で約５割が再就労している。再就労率の違いは出産１年前の従業上の地位の差で大きく表れ、無職だった人は、子が４歳６ヶ月になっても、６割は一度も職に就いていない。常勤だった人は１歳６ヶ月の時点で５割以上が再就労する。都市規模によっても違いが見られ、町村（郡部）の復帰が一番多く、時期も早い。母親が第１子を 30 歳以上になってから持った場合、１年６ヶ月以降の復帰のペースは遅く、全体でも再就労率は最も低い。子どもの祖父母の手助けがある場合は、三世代世帯であることを上回るペースで再就労に移行している。父親の収入については、低位であるほど再就労が多くなる傾向がみられた。

イベントヒストリー分析によると再就労の発生のハザード率が高いのは、出産１年前に有職である場合の諸変数である。祖父母が普段の保育に関わっていることも、次の時点での再就労確率を押し上げている。三世代世帯であることは、有意な効果を示さない、高校卒業と比べ、四大・大学院卒であることは、有意な効果を持たない、父親の収入が高ければ再就労確率は下がり、ひとり親ならば確率は上がる。都市規模は、小さいほど再就労確率が上がっている。都市の規模によって属性の効果が違うかも確かめた結果、全体と同じ傾向のものが多いが、三世代世帯の効果は、都市規模によって全く違う方向に働く。13 大都市では再就労を有意に抑制している。普段の保育に祖父母が関わることは、どの都市規模でも再就労確率を大きく上げている。出産１年前に有職の場合のみに限ってみると、常勤であった場合と比較して、パートの復帰の確率は大きく下がる。父収入は、全体と同じく多いほど有意に再就労確率を下げている。

子育てについてのテーマ（２）における②の負担に思うことや悩みがあるかどうかでは、母親の就業形態にかかわらず８割が感じており、ほとんど差がない。時間不足感・精神的余裕の不足感は常勤層、一時保育の制度不足感は専業主婦、経済的負担感は、パート・アルバイト層で強くなっている。変化の分析結果を一部挙げると、「自分の自由な時間が持てない」は、第１回～第２回　家事(専業)、勤め(常勤)、勤め(パート・アルバイト)それぞれ上昇傾向をみせたが、第４回以降、家事、勤め(パート)は減少、一方、勤め(常勤)は高水準で推移している。「子育てによる身体の疲れが大きい」については、第１回～第３回で、家事が勤め(常勤)、勤め(パート)を上回っている。

③においては、年収に占める子育て費用の割合が 10%をこえる層は、年収 200 万円未満で８割、年収 200～399 万円で６割と大半をしめていた。しかしながら、経済的負担感が最も高いのは、必ずしも年収 200 万円未満ではない。年収 400~599 万円の層が最も高く、続いて 200~399 万円の層が続く。さらに、経済的負担感と実態の費用負担の規定要因を検討した結果、所得に子育て費用が占める割合が 30%以上に有意な正の関連がみられた諸変数のなかで、オッズ比が 1.5 をこえていたのが、ひとり親世帯（別居、死別、離別）、収入 400 万円未満、習い事ありであった。また、経済的負担感（意識面）に有意な正の関連がみ

られた諸変数は子どもが病気がち、母親が仕事を探している、保育料が負担の場合であった。さらに、母親が専業主婦の場合は負の関連が見られた。

④の子ども観の分析結果では、第一次元では「物を大切にする子ども(該当)」など、他者や環境に従順で協調性のある子ども像を示すものが高い正の得点、「好奇心が旺盛な子ども(該当)」、「ねばりづよい子ども(該当)」などの、積極的で自発的な子ども像が高い負の得点を示した。第一次元は積極的・自発的か調整的・協調的かという子ども観の差異を表すと思われる。二次元は、「人の話をよく聞く子ども(該当)」などの得点が正に高く(「じょうぶなからだの子ども(非該当)」などの得点が負に高い。知性か感性かという子ども観の差異を表していると考えられる。次に、出生児調査の各ケースを積極的か調整的か、知性重視か感性重視かの2軸を交差させ、4つの子ども観に分類し、各タイプについて、ロジスティック回帰分析を行った結果、子どもが女子、祖父母と同居、母親が主婦であることが、「知性×調整」または「感性×調整」の調整型と正の関係性を持ち、「知性×積極」または「感性×積極」の積極型と負の関係性を持つ。回答者が父、母親が常勤、父親が大卒であることなどは、これと逆の関係性を示した。しつけとの関連では「感性」型、中でも「感性×積極」型は児童中心主義的で対話的なしつけ方法をとりやすいという傾向が明らかになった。

子どもの社会経済的背景については、⑤で本調査でも他調査と同様に脱落がより低所得層に偏っていることを明らかにした。0歳から4歳までの貧困率は1年目(出生前年)から2年目(出生前+0歳)にかけて、またその後も、(絶対的)貧困率は改善する傾向にある。貧困の類型別の割合を他の研究と比較すると本調査では比較的に安定層が多く(4回を通じて79%)、固定貧困層は少ない(同、1.9%)。安定層・貧困固定層は、分析対象の年数が多くなると共に減少し、年間のうち貧困を2回以上経験した割合は9.3%で、約1割の子どもである。母親や父親の年齢をコントロールした上でも、2001年から2002年にかけての貧困脱出の確率に比べ、2004年から2005年にかけての確率が有意に低くなっており、貧困が慢性化していることがうかがえる。1年目に貧困であった世帯が貧困から脱出するのに何年かかるかというサバイバル分析の結果、父親の学歴や母子世帯であることが有意に影響していることがわかった。次いで⑥の分析では、身長・体重双方で貧困経験回数は負で有意で、身長は貧困年数が1年増えると約0.08cm、体重では約30g低くなっている。その他の変数では未熟児ダミーと出生時体重が大きく影響していた。貧困経験の回数は通院経験には負で有意、入院経験については正で有意である。「健康ショック」の影響が、貧困世帯の子どもとそうでない子どもにとって異なるかについては、貧困経験は概ね負で有意であり、0歳からの貧困経験が体重に悪影響を及ぼしている。

最後の⑦政策評価における研究については、項目の細かい分析結果は省略するが、その分析作業を通じ、子育て行動・意識は、個人の属性、家庭環境、労働環境、地域環境など、様々な要因が絡み合ったものであり、自治体の行う次世代育成支援事業と子育て行動・意識との直接的な因果関係を導くことは難しいことがみられた。

D. 考察

テーマ(1)については、出産1年前に職をもっていると、出産後の再就労確率は大きく上がること、女性が出産後に再就労するために、祖父母(女性の親・義親)の

サポートが大きな効果を持つという結果が大きな知見である。2000年以降の出産は、対象児が第1子の場合と考えられ、母親の年齢層は多様であるため、母親のライフコースの中での出産時期の位置を分かりやすく指標化するために、変数を工夫する必要がある。

テーマ（2）においては、「専業主婦か、働いているか」という視点からの育児不安論を超え、所得や国籍も含め、母親の多様な側面、多様な状況に即した育児不安の議論が必要であること、また、子育て中の女性の子育て支援ニーズも、「専業主婦か、働いているか」という二分法より就業別、所得別、国籍別といった側面から、そのニーズを把握し、政策対応を見直していくことが重要だということがいえる。

また、子ども観の分類については、さまざまな分析に応用できることが指摘できる。たとえば、第6回調査では父母の子どもへの接し方や子どもの側の父母への接し方を尋ねているので、これらと子ども観の関係性を分析すれば、保育者の子ども観と子どもへの接し方の関係性や、それに対する子どもの側のフィードバックなどを検討できる。また小学校入学以後の調査では、さらに教育方針等との関係を探ることもできる。

テーマ（3）については、通院と入院の符号は逆であるが、貧困経験の蓄積は、身長・体重といった子どもの成長を有意に抑制し、0歳から4歳といった幼児期において、多かれ少なかれ子どもは病気をするものなので、通院といった日常における健康のケアについては、貧困経験がない子どものほうがより多く受け、貧困経験が多くなるほど、通院しない子が増える。入院を伴うような大きな疾病は、貧困経験が多い子どもの方が多くなると解釈できる。

テーマ（4）においては昨年度のヒアリング調査でもあったように、出生児調査は地域の次世代育成支援政策の立案や実行のために基礎的かつ重要なデータの宝庫である。都道府県ごとの分析結果の発信や共有化など、自治体の少子化対策担当者も強く期待しているので、第1～5回目のデータが蓄積された現在、地域別に即した資料提供など、より具体的な動きへとつなげていくこともできる。政策と行動・意識との関連の検討を課題とする必要がある。今回掲げた指標は限られたものであり、改善の余地もあるが、平成20年度に本格化するであろう、次世代育成支援行動計画のニーズ再調査、政策評価の議論へ向けて、次の課題を検討することができる。

E．結論

本研究では、厚生労働省の「21世紀縦断調査」の分析システムの開発を目指し、「出生児縦断調査」の1回～5回の個票データを用いて、行政上重要である、出産後の就労のタイミング、子育ての意識や実態、子どもの社会経済環境のテーマに絞って、事例分析を行った。また、本調査データを政策評価にいかに活用するかの検討も行った。

当然であるが、ここで扱ったテーマは本調査データから導き出すことのできるもののごく一部である。ここでは別々に扱っているテーマについても、たとえば子育てに関する分析で経済的負担感と子育て負担感と子ども観の関連性をみたり、子育て観と出産後の再就労のタイミングを検討したりするなど、無限の可能性を秘めている。分析手法も、これらのテーマについて有用なものをすべて試みるところまでには至っていない。しかしここで示した主たるテーマに関する分析は、パネル・データが整備されたからこそ可能となったものであり、本調査の重要性が再認識される。また、他で行われているパネル調査において得られた

結果との比較をすることで、調査をパネル分析という大きな枠組みの中で検討することもできた。ここで示した分析結果全体を考慮すると、父親の収入や、貧困か否かなどを含む子どもの社会経済的な背景、母親の就業状況、都市規模などがキーとなっていることがうかがえ、今後どのテーマの分析をする際も、注意を払う必要のある要素であるといえる。なお、政策効果の検証については、「２１世紀出生児縦断調査」が今後も積み重ねられていくことにより、現在講じられているさまざまな政策の長期的な効果の分析が可能となるため、継続的なデータの蓄積が望まれる。

F．研究発表

1．論文発表

なし

2．学会発表

なし

G．知的所有件の取得状況

なし

Ⅱ. 個別研究報告

（分析システム・分析手法）

1　分析総合システムの開発と実装

金子 隆一
三田 房美

はじめに

本事業では、縦断調査によって毎年継続して蓄積されて行く統計データに対し、複数の調査票に対応するデータセットを同時に対象として、統合的操作で集計、分析法をも考慮した有効で具体的なデータ管理・分析システムを検討し、開発することを目指している。パネル調査(縦断調査)は同一対象(個人)を追跡しながら継続的に調査を実施するものであり、その有効性を十分に引き出すためには横断調査とは異なるデータ管理、および分析が要求される。最も異なる点は、逐次累積される複数の調査回データセットを個人をキーとしながら連係させて集計・分析を行わなくてはならない点であろう。成年者縦断調査においては、男性票・女性票データセットを用いて夫婦単位の集計分析なども行われるため、データ操作はさらに複雑となる。

本事業におけるこれまでの研究では、縦断調査のデータ管理・分析システムとしての基本的課題とその対処するためのシステム用件について検討を行いエクセル・コード表を中心としたデータ管理・分析システムを開発した。ここではこれらを当初の構想にしたがってさらに発展させ、現在広く用いられている標準的リレーショナル・データベース・システムの枠組みを応用することにより、各回および各調査票に対応する多数のデータセットを連携させるシステムを開発した。これによれば、特定のテーマに対する集計ならびに統計分析の際に、必要な変数値を複数のデータセットから自動的に集約し、集計ならびに統計分析用のソフトウェアに受け渡すことができるので、分析者において毎回多数の指定を行う負担を軽減し、また多様かつ多重な変数の扱いにおける混乱や誤りを防ぐことができると期待される。

本システムは、データ管理のために開発された 21 世紀パネル（縦断）調査データベースシステム（PDB21）を基として開発し、統計分析パッケージと連携させて総合的な分析システムとなることから、21 世紀パネル（縦断）調査データ分析システム（PDA21）と称する。これは縦断調査の調査票単位の複数のデータセットから種々の統計分析パッケージ・ソフトウェアに対して、分析対象の変数からなるデータセットを生成して供給するシステムである。これによれば、分析者は分析に必要な項目（変数）を複数の調査回から選択するだけで、パッケージ独自形式のデータセットを得る事ができる。統計パッケージとして、現在、SAS、SPSS に対応しており、その他集計ソフト ADAM、データ管理用の言語として Perl にも対応している。対応する統計パッケージは、現在 STAT、S-Plus への拡張を行っている。

システムは、リレーショナル・データベース・システム PostgreSQL(Ver.8.0.4)をデータベースエンジンとして使用し、DICS-Ⅳ(または ADAM)、SPSS、SAS、Perl その他の一般の集

計または統計分析ソフトウェアと連携する。ただし、基本的ユーザインターフェースは、すべてEXCELを用いることでユーザに対する操作習熟に対する要求は最低限度となるように配慮した。ここではシステムの成り立つについての概要について記す。

システムの概要

通常、調査データは、調査票単位のデータセットとして保管されるが、その際、変数コード記述表（略称コード表）と呼ばれるデータ内容を変数ごとに記述する表が付帯する（図1）。21 世紀縦断調査においても、調査データセットの情報は厚生労働省の統計情報部が独自の形式で「調査変数定義表」として維持管理している。この表は、集計や統計分析にも必要なデータに関する情報をすべて含んでいるので、PDA21 では、データ管理、集計・分析にこれを直接活用することとしている。まず、PDA21 では、統計情報部独自形式の調査変数定義表から集計分析に必要な情報のみ抽出してプログラム等からの操作が可能なように定式化した「分析用コード表」と呼ぶ二次的なコード表を作成する。これらはどちらもExcel シートを媒体としており、この作成自体もプログラムによって自動化されている。そして、PDA21 ではこの分析用コード表を中心として、新たなデータセットを融合、生成したり、統計分析パッケージ用のデータセットやプログラムの生成を行っている。

図1 調査票～データセット・ユニット

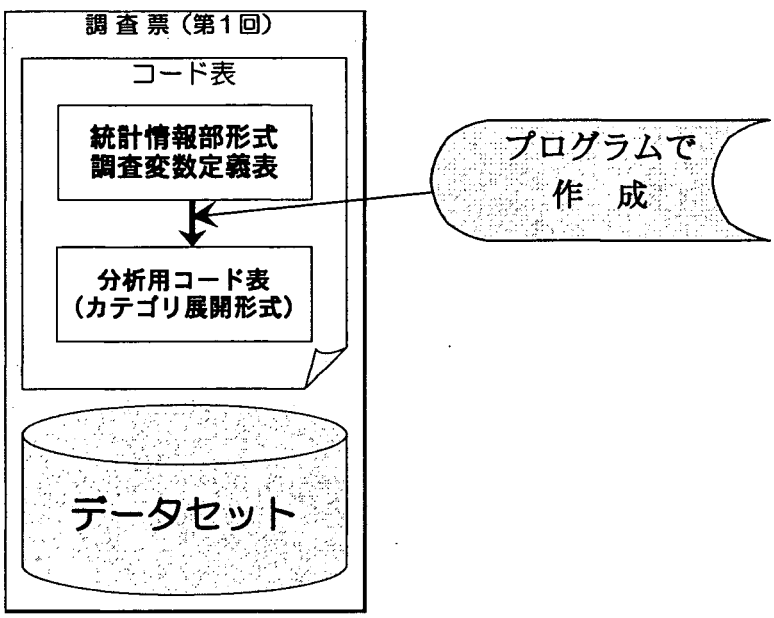
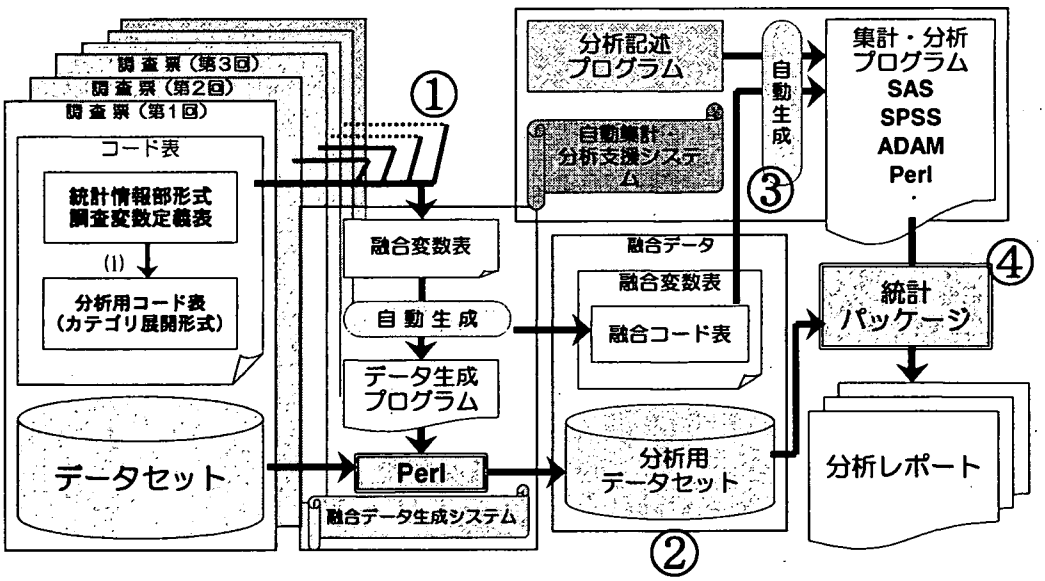


図１で見た調査データとコード表のユニットは、調査票ごとに管理されている（図２）。パネル調査では、これら複数のユニットから変数を抽出して集計分析を実施することになる。パネル調査でも調査回数が少ない場合には、これらのデータセットをすべて一つのデータセットとして接合し、分析することが可能であり、その場合には横断調査による集計分析手法をそのまま用いることも可能である。しかし、調査回数が多くなってきた場合、接合したデータは膨大なものとなる上、パネル分析で特徴とする調査回間の変数変化などを新たなデータとして加えるなどすると、変数の数も膨大で複雑なものとなるため、統計プログラムを実施する時間が無駄に長くなる上、その管理はたいへん煩雑になると考えられる。この場合、保管された調査票ごとのデータセットから、集計や分析の目的に応じて自動的に必要なデータセットが生成されれば、毎回そのような冗長な作業を繰り返す必要がなくなる。

このように複数のユニットから変数を抽出して集計分析用のデータセットを自動生成するのが、PDA21 である（図２）。そして、このシステムは単に分析用データセットを生成するだけでなく、これを使用する統計パッケージ独自のデータ形式としたり、統計分析のためのプログラムの作成を支援したりすることができる。図２によって、この作業の流れを見よう。

図２　データ分析システム（PDA21）の開発

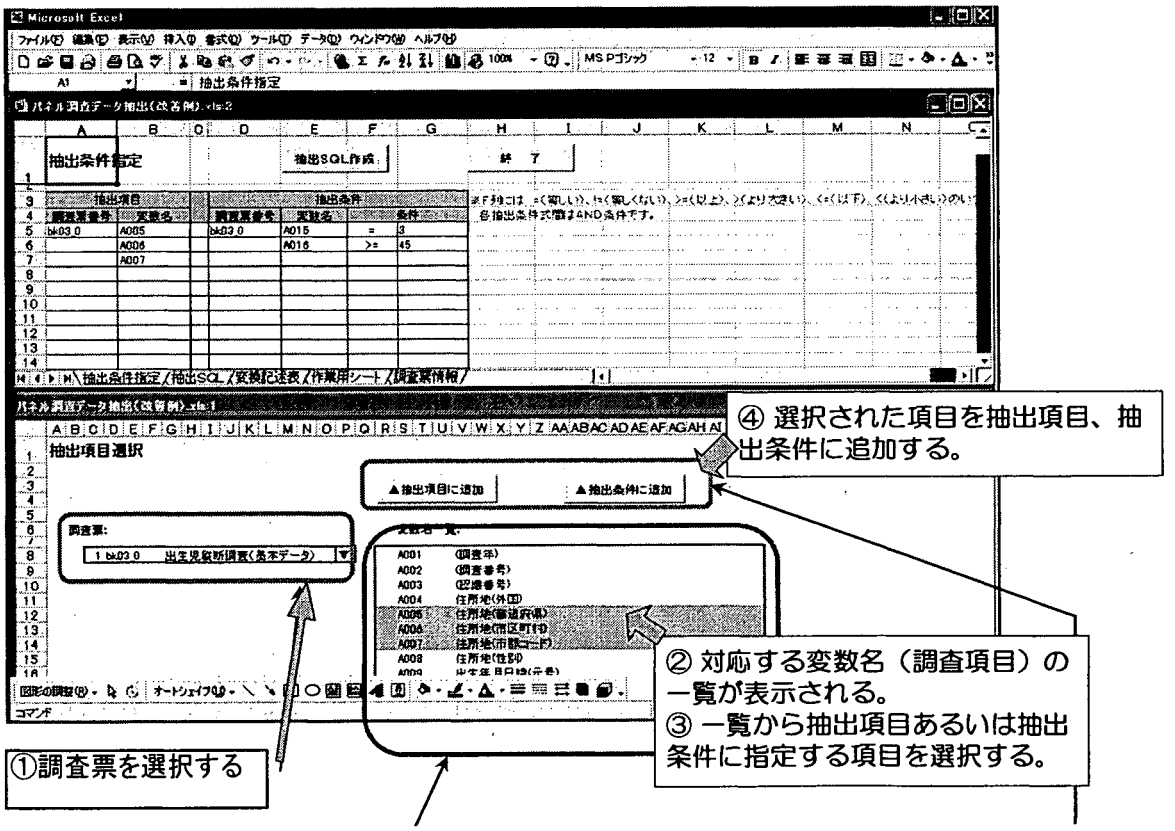
― パネルデータ集計分析プログラム自動生成 ―



まず、上述のように複数のデータセットのユニットから、分析に必要な変数データを抽出する（図２－①）。この際に、実際にはまず分析用コード表から変数を抽出し、そのデー

タを用いて分析用データセットを作成するプログラムが生成される。コード表からの変数の抽出には、独自のインターフェースが用意されており（図３）、ユーザによる選択作業を支援する。ユーザは調査票を選択した後、表示される変数名（調査項目）の一覧の中から変数を選択する。次に必要に応じて各変数の抽出条件を選択する。抽出条件は、たとえば対象者の「出生年を昭和 50 年以降に限定する」「有職者のみとする」「所得 200 万円未満とする」など、この段階で分析対象を絞り込むことができる。このようにしてコード表から抽出された変数データに関する情報は、そのまま新たなデータセットのコード表（図２融合変数表）となる。このコード表からデータ生成のプログラム（現在は Perl 言語による）を自動生成し、実施すると、分析用データセットが生成し、融合変数表と合わせて、新たなデータユニットとなる（図２－②）。

図３　パネルデータ自動集計・分析支援システム



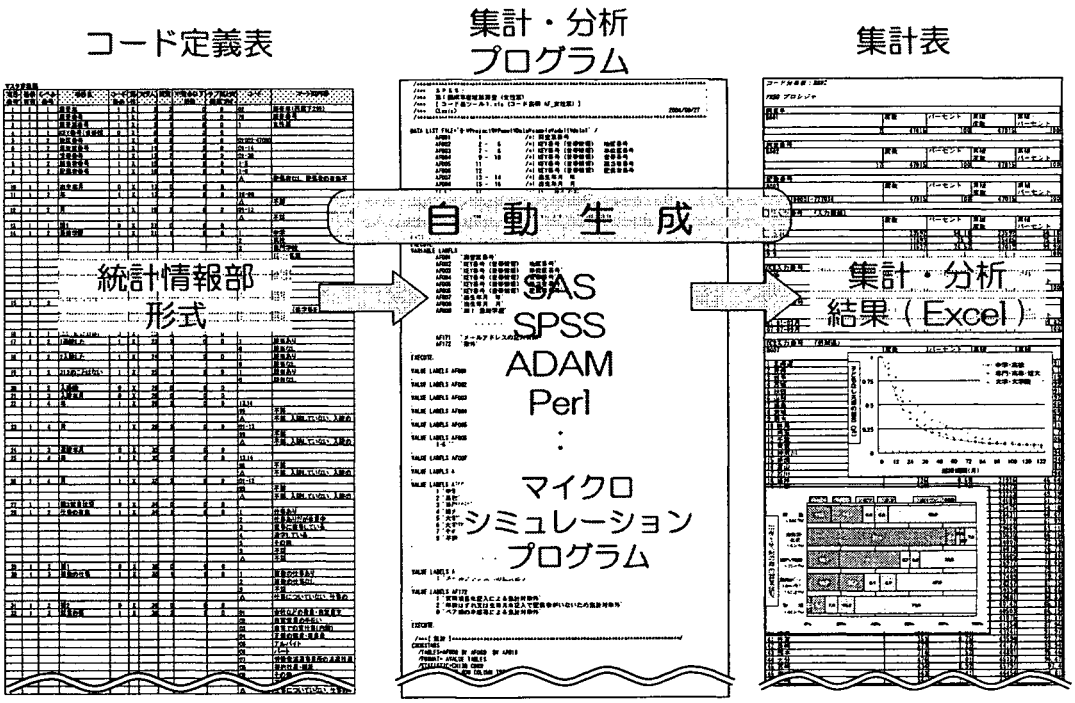
次に、PDA21 は、分析用データセットのコード表（融合変数表）に対して、各種統計パッケージ、集計ソフトウェア等の専用データセットを生成するプログラムや、分析用のプログラムを、ユーザの指定に従って自動生成する（図２－③）。ただし、現在の PDA21 で

は、単純集計等の基礎的な作業以外の集計分析以外の各パッケージ用の分析プログラムは、ユーザが用意しなくてはならない。最後に、これらを統計パッケージ・ソフトウェアなどに投入することによって、分析結果が得られることとなる（図２－④）。

現状では、ユーザがこれらの工程の管理を順に進めて行く形式となっているが、システムのインターフェースの改良により、最低限の指示作業と十分な作業速度によって行うことができれば、ユーザは変数の選択と統計分析結果の検討のみを行いながら試行錯誤的な分析サイクルを繰り返すことができるようになる。現在そうした改良にとりくんでいるところである。

図 4 には、システムの元となるコード表が、パッケージプログラムに変化され、集計分析結果として出力される流れを具体的記述イメージで示した。

図４　パネルデータ自動集計・分析支援システム



PDA21 は、統計パッケージ用のプログラムだけでなく、21 世紀縦断調査用のマイクロシミュレーションシステムのプログラムも生成するよう設計されている。マイクロシミュレーションとは、各種属性を持った個人の集団をコンピュータ上に構成して、おのおのの行動や状態変化を発生させることにより、集団の変化を再現するシミュレーション手法であり、縦断調査においても、対象集団の将来予測、行政制度・施策の効果の予見をはじめ、行動メカニズムの解明や統計手法の精度評価などに有効であるため、本事業においてもこれを実施するシステムを開発しているものである。

今後の開発

PDA21 は、さまざまな制約から開発が遅れた。システム開発には、データベース技術を始めとした高度の IT 技術を要するため、専門業者との連携が必須である。しかし、現在においては、ユーザインターフェースの改善等の周辺的課題を残しているものの、基本機能の実現には成功しており、直ちに実用化が可能であると考えている。したがって、今後の本事業において応用的に利用されるとともに、行政機関における調査実務においてもできるだけ活用され、実用の中で改善されて行くことが望まれる。

システムの拡張について、まず新たな統計パッケージへの対応については、システムにおけるパッケージ対応部分を独立させていることから、類似のプログラム構造を持つものであれば比較的対応が容易であり、現在 STATA、S-Plus への対応を進めている。また、上述のシミュレーションシステムとの連携を進めており、これが完成すると、21 世紀縦断調査データを直接用いたシミュレーション分析が用意となり、諸外国の例に見るような、政策効果の分析等への応用に道が開かれると考える。

21 世紀縦断調査のようなシリーズとして大きな価値を発揮する実地調査に関しては、これに即した高度分析のためのインフラが必要であり、PDA21 はその中で、データ管理および統計分析の総合的インフラの中核として機能するべきシステムであり、今後も充実を図って行くものとする。

2　中高年縦断調査における標本設計と分析方法の検討 (2)

石井　太

1.　はじめに

厚生労働省統計情報部において現在実施されている「中高年縦断調査」は、平成 17 年 10 月末時点で 50～59 歳であった男女を対象とし、健康・就業・社会活動について、意識面・事実面の変化の過程を継続的に把握することを目的とした統計調査である。そもそも本調査は、「統計行政の新たな展開方向」において、「中高年齢者についても、既存の調査と併せ、その行動の変化や事象間の関連性などについて把握することにより、より詳細な分析が可能となるよう、データの整備・充実を図ることが求められている」という問題意識が提示されたことを踏まえて企画・立案されたものであり、本調査の結果分析手法の充実は、調査企画本来の主旨に沿うとともに、調査体系においても重要な位置を占めるものであるといえる。

昨年度の本研究では、中高年縦断調査の標本設計について整理を行うとともに、標本の代表性の問題に関連してサンプリングデザインの考慮が与える影響等に関する問題点の例に関して考察を行った。特に、有限母集団から層化抽出された標本に関するカイ二乗検定に焦点を当て、Holt et al. (1980) による方法に関してレビューするとともに、数値シミュレーションによる評価を行った。

昨年度の研究において行った数値シミュレーションでは、母集団が既知であることから母集団の分散・共分散行列を直接求めて検定を行ったが、実際の標本調査では、標本からの推定値に基づいて検定を行う必要がある。また、昨年度のシミュレーションでは集落抽出の考慮は行わなかったが、実際の中高年縦断調査においてはこのような抽出法上の考慮も必要となる。そこで、本年度においては、昨年度の研究成果を基礎としつつ、実際の中高年縦断調査のデータを用い、集落抽出法など実際の標本抽出法を考慮し、標本からの分散・共分散行列の推定を行った上で、カイ二乗検定を行うとともに、統計量の分布について考察を行うこととする。

2.　中高年縦断調査の標本設計に基づく各種推定量の理論的整理

本研究では、以下、年齢分布の適合度に関する仮説検定として、第一回中高年縦断調査実施時における調査対象者の年齢分布（51～60 歳）の推定値を、平成 17 年国勢調査結果による年齢分布と比較する例に基づき議論を進める。本節においては、各種推定量と検定統計量について、通常行われるように無限母集団から単純無作為抽出を行ったと考えた場合の推定量と、有限母集団から標本設計を考慮して抽出を行ったと考えた場合の推定量２

種類に関して、理論的側面から整理を行う。

昨年度の本研究においては、中高年縦断調査の標本設計のレビューを行った。このうち、本年度の研究に関連する部分について、以下に簡単にまとめる。

- 中高年縦断調査の標本抽出にあたっては、まず、平成16年国民生活基礎調査世帯票として、国勢調査の調査地区約90万地区から、都道府県別に定められた地区数に基づき、5,280地区が層化抽出される。さらに、この世帯票の地区から、中高年縦断調査として、都道府県別に定められた地区数に基づいて2,515地区が標本抽出される。
- 一般に国民生活基礎調査の後続調査においては、集計を簡便にする観点から、国勢調査の地区数に比例するように地区数を設定し、各都道府県別での抽出率が一定となるように標本抽出を行うことが多い。中高年縦断調査でも基本的にはこのような考え方に基づき地区数が設定されているが、本調査では約4万程度の客体に対して調査を行う必要性から、一定の地区数を確保することが求められ、都市部でやや低めの客体分布の見込みとなっているなど、必ずしも一定となっていない部分も存在する。このことに対しては、地区数によるウエイトを考慮した推定を行う、人口などの補助変量を利用した比推定などの手法を用いるという対応法が考えられる。

理論的整理に当たり、以下のように記号を定義しておく。標本設計における層として都道府県を考慮し、これを $h(=1,\cdots,47)$ で表す。

N : 国勢調査地区総数
N_h : 都道府県別国勢調査地区数
n_h : 都道府県別中高年縦断調査調査地区数
P : 平成17年国勢調査に基づく51～59歳人口（平成17年10月末現在）
$P^{(j)}$: 平成17年国勢調査に基づ年齢区分 j の人口 ($j=1,\cdots,10$, 年齢 $50+j$ 歳)（平成17年10月末現在）
P_h : 平成17年国勢調査に基づく都道府県別51～59歳人口（平成17年10月末現在）
$X_{hi}^{(j)}$: 母集団、層 h, 地区 i における年齢区分 j の人数 ($j=1,\cdots,10$, 年齢 $50+j$ 歳)
Y_{hi} : 母集団、層 h, 地区 i における人数 ($=\sum_j X_{hi}^{(j)}$)
$x_{hi}^{(j)}$: 標本、層 h, 地区 i における年齢区分 j の人数 ($j=1,\cdots,10$, 年齢 $50+j$ 歳)
y_{hi} : 標本、層 h, 地区 i における人数 ($=\sum_j X_{hi}^{(j)}$)

このとき、母集団における年齢区分 j の人口割合 $p^{(j)}$ は、

$$p^{(j)} = \frac{\sum_h \sum_i X_i^{(j)}}{\sum_h \sum_i Y_i}$$

となる。これらを $j=10$ のカテゴリーを省略して、

$$\mathbf{p} = (p^{(1)}, p^{(2)}, \ldots, p^{(9)})^T$$

と表す。

さて、ここで

$$\mathbf{p_0} = (p_0^{(1)}, p_0^{(2)}, \ldots, p_0^{(9)})^T$$

としたとき、我々の問題は、母集団に関する帰無仮説

$$H_0 : \mathbf{p} = \mathbf{p_0}$$

を、母集団から得た標本に基づいて検定することとなる。

しかしながら、この問題は、どのような標本設計に基づいてどのような推定量を用いるかにより検定統計量が異なるものとなる。そこで、本研究では、

- 無限母集団から単純無作為抽出が行われたとした場合
- 有限母集団から中高年縦断調査の標本設計に従って抽出が行われ、地区数のウエイトを考慮して推定を行う場合
- 有限母集団から中高年縦断調査の標本設計に従って抽出が行われ、補助変量を考慮して比推定を行う場合

の３通りについて、それぞれの推定量と検定法を比較することとする。

2.1 無限母集団から単純無作為抽出が行われたとした場合

通常行われる χ^2 検定では、得られた標本が、無限母集団から独立同一分布に従って単純無作為抽出されたと考えて検定統計量が構成される。この条件の下では、母集団年齢分布 $\mathbf{p}$ の推定量 $\hat{\mathbf{p}}$ を、

$$\hat{p}^{(j)} = \frac{\sum_h \sum_i x_{hi}^{(j)}}{\sum_h \sum_i y_{hi}}$$

とし、検定統計量は、

$$\bar{X}^2 = n \sum_{j=1}^{k} \frac{\left(\hat{p}^{(j)} - p_0^{(j)}\right)^2}{p_0^{(j)}}$$

で与えられる。このとき、

$$\mathbf{V_n} = \mathrm{diag}(\mathbf{p_0}) - \mathbf{p_0}\mathbf{p_0}^T$$

と定義すると、$\mathbf{V_n}$ は H_0 が真であるとき、独立に単純無作為抽出を行ったとした場合の分散共分散行列であり、このとき、$\bar{X}^2$ は、

$$\bar{X}^2 = n\left(\hat{\mathbf{p}} - \mathbf{p_0}\right)^T \mathbf{V_n}^{-1}\left(\hat{\mathbf{p}} - \mathbf{p_0}\right)$$

とかける。今回の例ではカテゴリー数は 10 であり、この $\bar{X}^2$ が χ^2_9 に従うことに基づき、検定を行うことになる。

2.2 有限母集団から中高年縦断調査の標本設計に従って抽出が行われ、地区数のウエイトを考慮して推定を行う場合

さて、昨年度の研究において見た通り、一般的に標本が有限母集団から層化抽出されているようなケースにおいては、必ずしも前節のような議論が成立しない場合があった。そこで、ここでは、得られた標本が中高年縦断調査の標本設計に従って抽出され、地区数による抽出率の違いをウエイトによって考慮して推定を行った場合に、どのような検定等計量を考えればよいか考察することとする。

昨年度のレビューにある通り、Holt et al. (1980) によれば、標本調査においては標本設計が層化抽出・多段抽出などによっており、独立に単純無作為抽出を行うとの仮定は現実的でなく、その代わりとして、標本の大きさ $n \to \infty$ のとき、ある正定値行列 $\mathbf{V}$ に対して、以下が成立する事を仮定する。

$$\sqrt{n}\left(\hat{\mathbf{p}} - \mathbf{p}\right) \xrightarrow{L} N\left(\mathbf{0}, \mathbf{V}\right)$$

もし、

$\hat{\mathbf{V}}$: $\mathbf{V}$ の一致推定量

がある場合には、Wald Statistic

$$\bar{X}^2_w = n\left(\hat{\mathbf{p}} - \mathbf{p_0}\right)^T \hat{\mathbf{V}}^{-1}\left(\hat{\mathbf{p}} - \mathbf{p_0}\right)$$

が χ^2_{k-1} に漸近的に従う事から、これを用いて検定が可能であった。

昨年度の本研究では、母集団分布が既知、すなわち、$\mathbf{V}$ についても既知であるという前提の下に数値シミュレーションを行った。しかし、実際の標本調査においては $\mathbf{V}$ の値を事前に知ることはできないため、$\hat{\mathbf{V}}$ の推定が必要となる。そこで、以下、中高年縦断調査の標本設計に基づき、その推定をどのように行うかついて述べる。

地区数によるウエイトを考慮した場合、母集団年齢分布 $\mathbf{p}$ の推定量 $\hat{\mathbf{p}}$ は、複合比推定により、

$$\hat{p}^{(j)} = \frac{\sum_h \frac{N_h}{N} \quad \frac{1}{n_h} \quad \sum_i x^{(j)}_{hi}}{\sum_h \frac{N_h}{N} \quad \frac{1}{n_h} \quad \sum_i y_{hi}}$$

と推定される。このとき、この推定量の分散及び共分散は母集団における変量を用いて以下のように表される（これらの導出については、石井 (2004) を参照。以下同様)。

$$\mathbf{V}(\hat{p}^{(j)})$$
$$=p^{(j)2}\sum_h \left(\frac{N_h}{N}\right)^2 \left(\frac{1}{n_h}-\frac{1}{N_h}\right)\left(\frac{\sigma^2_{X_h^{(j)}}}{\mu^2_{X^{(j)}}}-2\frac{\sigma_{X_h^{(j)}Y_h}}{\mu_{X^{(j)}}\mu_Y}+\frac{\sigma^2_{Y_h}}{\mu^2_Y}\right)$$

$$\mathbf{COV}(\hat{p}^{(j)},\hat{p}^{(k)})$$
$$=p^{(j)}p^{(k)}\sum_h \left(\frac{N_h}{N}\right)^2 \left(\frac{1}{n_h}-\frac{1}{N_h}\right)\left(\frac{\sigma_{X_h^{(j)}X_h^{(k)}}}{\mu_{X^{(j)}}\mu_{X^{(k)}}}-\frac{\sigma_{X_h^{(j)}Y_h}}{\mu_{X^{(j)}}\mu_Y}-\frac{\sigma_{X_h^{(k)}Y_h}}{\mu_{X^{(k)}}\mu_Y}+\frac{\sigma^2_{Y_h}}{\mu^2_Y}\right)$$

さらに、これらは、標本を用いて以下のように推定することができる。

$$\hat{\mathbf{V}}(\hat{p}^{(j)})$$
$$=(\hat{p}^{(j)})^2\sum_h \left(\frac{N_h}{N}\right)^2 \left(\frac{1}{n_h}-\frac{1}{N_h}\right)\left(\frac{Var(x_h^{(j)})}{(\bar{x}^{(j)})^2}-2\frac{Cov(x_h^{(j)},y_h)}{(\bar{x}^{(j)})\bar{y}}+\frac{Var(y_h)}{\bar{y}^2}\right)$$

$$\hat{\mathbf{COV}}(\hat{p}^{(j)},\hat{p}^{(k)})$$
$$=(\hat{p}^{(j)})(\hat{p}^{(k)})\sum_h \left(\frac{N_h}{N}\right)^2 \left(\frac{1}{n_h}-\frac{1}{N_h}\right)$$
$$\times\left(\frac{Cov(x_h^{(j)},x_h^{(k)})}{(\bar{x}^{(j)})(\bar{x}^{(k)})}-\frac{Cov(x_h^{(j)},y_h)}{(\bar{x}^{(j)})\bar{y}}-\frac{Cov(x_k^{(k)},y_h)}{(\bar{x}^{(k)})\bar{y}}+\frac{Var(y_h)}{\bar{y}^2}\right)$$

よって、これらから分散共分散行列 $\hat{\mathbf{V}}_\mathbf{c}$ を

$$\frac{\hat{\mathbf{V}}_\mathbf{c}}{n}=\begin{bmatrix} \hat{\mathbf{V}}(\hat{p}^{(1)}) & \hat{\mathbf{COV}}(\hat{p}^{(1)},\hat{p}^{(2)}) & \cdots & \hat{\mathbf{COV}}(\hat{p}^{(1)},\hat{p}^{(9)}) \\ \hat{\mathbf{COV}}(\hat{p}^{(2)},\hat{p}^{(1)}) & \hat{\mathbf{V}}(\hat{p}^{(2)}) & \cdots & \hat{\mathbf{COV}}(\hat{p}^{(2)},\hat{p}^{(9)}) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{COV}}(\hat{p}^{(9)},\hat{p}^{(1)}) & \hat{\mathbf{COV}}(\hat{p}^{(9)},\hat{p}^{(2)}) & \cdots & \hat{\mathbf{V}}(\hat{p}^{(9)}) \end{bmatrix}$$

として推定し、これを用いて Wald Statistic$\bar{X}^2_{wc}$ を求め、カイ二乗検定を行うことができる。

2.3 有限母集団から中高年縦断調査の標本設計に従って抽出が行われ、補助変量を考慮して比推定を行う場合

有限母集団から中高年縦断調査の標本設計に従って抽出が行われるとした場合、国民生活基礎調査世帯票の推定で行われているように、人口を補助変量とする比推定を行った推

定量を考えることができる。そこで、ここでは、各都道府県の 51～60 歳人口が補助変量として既知であることを仮定し、これを利用した分離比推定による推定量を考察することとする。

人口を補助変量とする比推定による場合、母集団年齢分布 $\mathbf{p}$ の推定量 $\hat{\mathbf{p}}$ は、分離比推定により以下の式により推定される。

$$\hat{p}^{(j)} = \sum_h \left(\frac{P_h}{P}\right) \left(\frac{\sum_i x_{hi}^{(j)}}{\sum_i y_{hi}}\right)$$

このとき、この推定量の分散及び共分散は母集団における変量を用いて以下のように表される。

$$\mathbf{V}(\hat{p}^{(j)})$$
$$= \sum_h \left(\frac{P_h}{P}\right)^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \left(\frac{\sigma^2_{X_h^{(j)}}}{\mu^2_{X_h^{(j)}}} - 2\frac{\sigma_{X_h^{(j)}Y_h}}{\mu_{X_h^{(j)}}\mu_{Y_h}} + \frac{\sigma^2_{Y_h}}{\mu^2_{Y_h}}\right) {p_h}^{(j)2}$$

$$\mathbf{COV}(\hat{p}^{(j)}, \hat{p}^{(k)})$$
$$= \sum_h \left(\frac{P_h}{P}\right)^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \left(\frac{\sigma_{X_h^{(j)}X_h^{(k)}}}{\mu_{X_h^{(j)}}\mu_{X_h^{(k)}}} - \frac{\sigma_{X_h^{(j)}Y_h}}{\mu_{X_h^{(j)}}\mu_{Y_h}} - \frac{\sigma_{X_h^{(k)}Y_h}}{\mu_{X_h^{(k)}}\mu_{Y_h}} + \frac{\sigma^2_{Y_h}}{\mu^2_{Y_h}}\right) {p_h}^{(j)} {p_h}^{(k)}$$

さらに、これらは、標本を用いて以下のように推定することができる。

$$\hat{\mathbf{V}}(\hat{p}^{(j)})$$
$$= \sum_h \left(\frac{P_h}{P}\right)^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \left(\frac{Var(x_h^{(j)})}{(\bar{x_h}^{(j)})^2} - 2\frac{Cov(x_h^{(j)}, y_h)}{(\bar{x_h}^{(j)})\bar{y_h}} + \frac{Var(y_h)}{{\bar{y_h}}^2}\right) (\hat{p_h}^{(j)})^2$$

$$\hat{\mathbf{COV}}(\hat{p}^{(j)}, \hat{p}^{(k)})$$
$$= \sum_h \left(\frac{P_h}{P}\right)^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right)$$
$$\times \left(\frac{Cov(x_h^{(j)}, x_h^{(k)})}{(\bar{x_h}^{(j)})(\bar{x_h}^{(k)})} - \frac{Cov(x_h^{(j)}, y_h)}{(\bar{x_h}^{(j)})\bar{y_h}} - \frac{Cov(x_k^{(k)}, y_h)}{(\bar{x_h}^{(k)})\bar{y_h}} + \frac{Var(y_h)}{{\bar{y_h}}^2}\right) (\hat{p_h}^{(j)})(\hat{p_h}^{(k)})$$

よって、これらから分散共分散行列 $\hat{\mathbf{V}}_\mathbf{s}$ を

$$\frac{\hat{\mathbf{V}}_\mathbf{s}}{n} = \begin{bmatrix} \hat{\mathbf{V}}(\hat{p}^{(1)}) & \hat{\mathbf{COV}}(\hat{p}^{(1)}, \hat{p}^{(2)}) & \cdots & \hat{\mathbf{COV}}(\hat{p}^{(1)}, \hat{p}^{(9)}) \\ \hat{\mathbf{COV}}(\hat{p}^{(2)}, \hat{p}^{(1)}) & \hat{\mathbf{V}}(\hat{p}^{(2)}) & \cdots & \hat{\mathbf{COV}}(\hat{p}^{(2)}, \hat{p}^{(9)}) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{COV}}(\hat{p}^{(9)}, \hat{p}^{(1)}) & \hat{\mathbf{COV}}(\hat{p}^{(9)}, \hat{p}^{(2)}) & \cdots & \hat{\mathbf{V}}(\hat{p}^{(9)}) \end{bmatrix}$$

として推定し、これを用いて Wald Statistic$\tilde{X}^2_{ws}$ を求め、カイ二乗検定を行うことができる。

3. 中高年縦断調査データによる推定量の検討

前節においては、理論的側面から推定量と対応する検定統計量について整理を行った。そこで、本節ではこれを実際の中高年縦断調査のデータに当てはめ、カイ二乗検定を実行するとともに、その検定統計量の分布について数値シミュレーションを行うことにより、標本設計を考慮することによる影響を評価することとする。

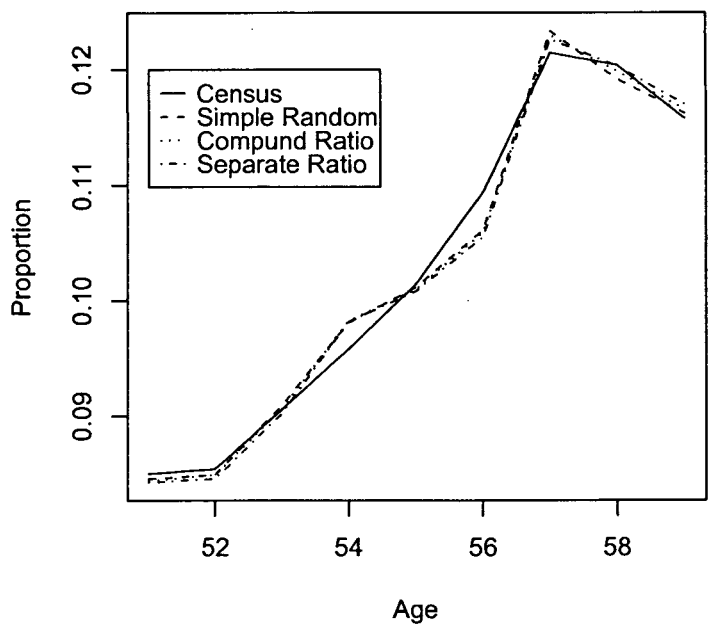


図1　年齢分布の推定量の比較

図1は、平成17年国勢調査に基づいて推計した51～59歳の年齢分布 (Census) を、前節において述べた3種類の推定量、すなわち、無限母集団から単純無作為抽出が行われたとした場合の推定量 (Simple Random)、地区数のウエイトを考慮した複合比推定形式による推定量 (Compound Ratio)、人口を補助変量とした分離比推定形式による推定量 (Separate Ratio) と比較したものである。これによれば、この3つの推定量は若干の違いはあるものの、概ね同様の推定結果となっていることがわかる。すなわち、本調査のデータによれば、年齢分布の推定に関しては、単純無作為抽出とみなして推定を行ったとしても、ウエイトや補助変量を考慮した推定量と大きな違いはないと考えてよいことがわ

かる。

次に、前節において述べた、カイ二乗検定に用いる検定統計量 $\bar{X}, \bar{X}^2_{wc}, \bar{X}^2_{ws}$ を検討する。推定結果を表したものが表1である。

	$\chi^2 - value$	df	p-value
Simple Random	7.044	9	0.6325
Compound Ratio	7.716	9	0.5630
Separate Ratio	9.356	9	0.4050

表1　カイ二乗検定の結果

この結果によれば、検定統計量はどれも帰無仮説を棄却するほどの大きさではなく、どの推定量についても国勢調査結果と大きく異なる年齢分布という帰結にはならない点では同様である。しかしながら、各検定統計量の値は選定する推定量によって幅のある結果となっており、場合によっては結論に影響を与えうる可能性があるといえる。

ところで、この検定統計量における検定結果が有効なものとなるためにはこれらが実際に自由度9のカイ二乗分布に従うことが前提である。しかしながら、昨年度の数値シミュレーションにおいて明らかとなったように、有限母集団から標本抽出された標本を無限母集団からの単純無作為標本とみなした場合には、その検定統計量が必ずしもカイ二乗分布に従わないことがあった。そこで、次に、これらの検定統計量がどのように分布するのか、数値シミュレーションを用いて考察してみることとする。

数値シミュレーションはブートストラップ法に基づいて行った。具体的には以下のような方法による。各都道府県毎に、その都道府県に割り当てられている地区の数だけ、当該都道府県の地区を復元抽出する。この手続きによって得られた一組のブートストラップ標本に基づき、各種推定量及び検定統計量を構成する。今回は、10000組のブートストラップ標本に基づいて検定統計量の分布を作成し、比較・検討を行うこととした。

図2は、検定統計量の密度関数の比較を行ったものである。これによれば、どの推定量についても分布は概ね理論的な χ^2 分布 (Theoretical) と一致していることが観察される。しかしながら、より詳細に観察すると、標本設計を考慮した複合比推定形式による推定量 (Compound Ratio)、分離比推定形式による推定量 (Separate Ratio) では、理論的な χ^2 分布 (Theoretical) と分布がよく一致しているものの、無限母集団からの単純無作為標本とみなした場合 (Simple Random) には分布の形状に影響がやや出ていることがわかる。

図3は、検定統計量の分布関数の比較を行ったものである。これからも同様の傾向を見ることができるが、これをより詳細に観察するため、分布関数について、各種検定統計量の理論値に対する比を見てみたものが図4である。

これによれば、標本設計を考慮した複合比推定形式による推定量 (Compound Ratio)、