者がいない地区をのぞいてすべての国民生活基礎調査地区を対象とする都道府県がいくつか発生することとなる。都道府県別の対象及び客体見込みの分布を示したものが図3である。

| | | 対象者分布(A) | 中高年縦断調査による分布(B) | 両者の比(B)/(A) |
|---|---|---|---|---|
| 1 | 北海道 | 4.7% | 4.8% | 100.9% |
| 2 | 青森県 | 1.2% | 1.4% | 118.8% |
| 3 | 岩手県 | 1.1% | 1.2% | 108.2% |
| 4 | 宮城県 | 1.8% | 2.3% | 123.4% |
| 5 | 秋田県 | 1.0% | 1.1% | 114.2% |
| 6 | 山形県 | 1.0% | 1.0% | 99.8% |
| 7 | 福島県 | 1.7% | 1.9% | 113.7% |
| 8 | 茨城県 | 2.5% | 2.7% | 108.6% |
| 9 | 栃木県 | 1.7% | 1.9% | 114.0% |
| 10 | 群馬県 | 1.6% | 1.8% | 107.7% |
| 11 | 埼玉県 | 5.7% | 5.4% | 94.6% |
| 12 | 千葉県 | 4.9% | 4.0% | 81.9% |
| 13 | 東京都 | 9.1% | 6.0% | 66.3% |
| 14 | 神奈川県 | 6.6% | 5.9% | 90.3% |
| 15 | 新潟県 | 2.0% | 2.0% | 101.7% |
| 16 | 富山県 | 0.9% | 1.0% | 109.1% |
| 17 | 石川県 | 0.9% | 1.0% | 105.8% |
| 18 | 福井県 | 0.6% | 0.7% | 107.5% |
| 19 | 山梨県 | 0.7% | 0.8% | 114.0% |
| 20 | 長野県 | 1.7% | 1.8% | 110.4% |
| 21 | 岐阜県 | 1.7% | 1.8% | 111.1% |
| 22 | 静岡県 | 3.0% | 3.3% | 109.0% |
| 23 | 愛知県 | 5.4% | 5.4% | 99.9% |
| 24 | 三重県 | 1.4% | 1.6% | 110.1% |
| 25 | 滋賀県 | 1.0% | 1.3% | 121.2% |
| 26 | 京都府 | 2.0% | 2.3% | 111.6% |
| 27 | 大阪府 | 6.8% | 4.9% | 72.0% |
| 28 | 兵庫県 | 4.4% | 4.8% | 109.6% |
| 29 | 奈良県 | 1.2% | 1.3% | 110.6% |
| 30 | 和歌山県 | 0.8% | 1.1% | 135.0% |
| 31 | 鳥取県 | 0.5% | 0.5% | 112.6% |
| 32 | 島根県 | 0.6% | 0.7% | 115.9% |
| 33 | 岡山県 | 1.5% | 1.7% | 109.5% |
| 34 | 広島県 | 2.3% | 2.5% | 109.6% |
| 35 | 山口県 | 1.2% | 1.4% | 112.2% |
| 36 | 徳島県 | 0.7% | 1.0% | 144.4% |
| 37 | 香川県 | 0.8% | 0.9% | 107.9% |
| 38 | 愛媛県 | 1.2% | 1.3% | 111.8% |
| 39 | 高知県 | 0.7% | 0.8% | 121.5% |
| 40 | 福岡県 | 4.0% | 4.5% | 111.2% |
| 41 | 佐賀県 | 0.7% | 0.8% | 119.5% |
| 42 | 長崎県 | 1.2% | 1.2% | 101.5% |
| 43 | 熊本県 | 1.4% | 1.6% | 110.0% |
| 44 | 大分県 | 1.0% | 1.0% | 104.2% |
| 45 | 宮崎県 | 0.9% | 1.1% | 117.5% |
| 46 | 鹿児島県 | 1.4% | 1.6% | 118.5% |
| 47 | 沖縄県 | 1.0% | 1.1% | 116.2% |
| 合計 | | 100.0% | 100.0% | 100.0% |

図3　都道府県別地区数

両者の比を見ると、概ね100%前後となっているものの、東京都で66.3%、大阪府で72.0%など、都市部でやや低めの客体分布の見込みとなっている部分もある。これは、標本抽出時の見込みであるため、実際に調査票の配布や有効回答があったかというものではない。しかしながら、少なくとも標本抽出段階における分布の違いは、標本の代表性の問題に関係するため、記述統計等に対して影響を与えうる可能性が考えられよう。これに関連して、本研究プロジェクトに先行してまとめられた金子(2006)の中においても、委員から、「代表性の問題は、永遠のテーマであるが、ひとつには、何らかの形でウエイトを指

定してかけてみる方法がある。多変量解析の場合は基本的にいいが、全体の descriptive の時にはウエイトを計算することを勧める。」という意見が出されているところである。

これへの対応の一つとして、国民生活基礎調査世帯票で行われているように、各都道府県の抽出率の違いを考慮し、人口を補助変量とする比推定を行う手法が考えられる。国民生活基礎調査世帯票の推計値の作成（大規模年）においては、ある属性を持つ世帯数（あるいは世帯員数）の全国推計値 $\hat{T}$ は、世帯人員を補助変数とする分離比推定により算出される。すなわち、$k$ 県における推計値を $\hat{T}_k$ とおくと、

$$\hat{T}_k = \frac{\sum_j X_{kj}}{\sum_j Y_{kj}} \cdot P_k$$

であり、

$$\hat{T} = \sum_k \hat{T}_k$$

となる。ただし、

$X_{kj}$ : $k$ 県第 $j$ 調査地区のある属性を持つ世帯数（または世帯員数）

$Y_{kj}$ : $k$ 県第 $j$ 調査地区の総世帯員数

$P_k$ : $k$ 県日本人人口

である。

一方、通常、モデルベースの推定では、無限母集団から独立同一分布に従って標本が発生するという前提が置かれていることが多く、デザインエフェクトである有限母集団からの標本抽出が考慮されることはあまり一般的ではない状況にある。しかしながら、このようなデザインエフェクトを考慮した推定や検定等の問題についても、多くの先行研究がなされており、また、様々な議論が行われてきている。金子 (2006) の中においても、委員から、「出生児調査で脱落の検討をされているが、同じようなことを成年者調査での脱落をみると、成年者調査は、サンプルが抽出であり推計の仕方によっては、分子の方だけでなく分母の方も確率変数となるので分子、分母も比推定で普通より分散が大きくなる。その分を考えて行う必要がある。こうした事情もあるので、役所のデータに関しては、パッケージは有効に機能しないことがあるので、注意が必要となる。」といった意見が提言されており、同様の標本抽出を行っている中高年縦断調査においてもこれらについて注意が必要となる。

そこで、本稿では、以下、このような問題に関連し、標本の代表性に関する検定としても用いられるカイ二乗検定に焦点を当て、デザインエフェクトを考慮してカイ二乗検定を行う方法のレビューと簡単なシミュレーションを行うこととする。

## 3. 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

有限母集団からの標本に関するカイ二乗検定に関しては数多くの研究が行われており、代表的なものとして、Holt et al. (1980)、Rao and Scott (1981) などが挙げられる。ここでは、Holt et al. (1980) に沿って有限母集団からの標本に関するカイ二乗検定の一般論をレビューする。

母集団が $k$ 個のカテゴリーに分割されているとし、それぞれの構成割合を

$$p_1, p_2, \ldots, p_k \quad \left(\sum_{j=1}^{k} p_j = 1\right)$$

とする。最後のカテゴリーを省略してベクトル表示を行い、

$$\mathbf{p} = (p_1, p_2, \ldots, p_{k-1})^T$$

と表すこととする。今、帰無仮説

$$H_0 : \mathbf{p} = \mathbf{p_0} = (p_{0,1}, p_{0,2}, \ldots, p_{0,k-1})^T$$

の検定の問題を考える。

$$\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k : p_1, p_2, \ldots, p_k \text{の不偏推定量}$$

としたとき、通常のカイ二乗検定量は、

$$\bar{X}^2 = n \sum_{j=1}^{k} \frac{(\hat{p}_j - p_{0,j})^2}{p_{0,j}}$$

で与えられる。このとき、

$$\mathbf{P_0} = \mathrm{diag}(\mathbf{p_0}) - \mathbf{p_0}\mathbf{p_0}^T$$

と定義すると、$\mathbf{P_0}$ 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、

$$\bar{X}^2 = n\,(\hat{\mathbf{p}} - \mathbf{p_0})^T\, \mathbf{P_0}^{-1}\, (\hat{\mathbf{p}} - \mathbf{p_0})$$

とかける。しかしながら、多くの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。

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

もし、

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

がある場合には、Wald Statistic

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

が $\chi_{k-1}^2$ に漸近的に従う事から、これを用いて検定が可能である。Holt et al. (1980) では、複雑な標本設計の下では $\hat{\mathbf{V}}$ の推定が難しいとしているが、オリジナルのミクロデータへのアクセスが可能な状況であれば推定は可能であると考えられる。母集団が既知である場合の数値例に関して次節で示す。

次に、同じく Holt et al. (1980) に基づき、2つの異なる母集団から抽出された標本の構成割合についての場合を示す。今、2つの異なる母集団から大きさ $n_1, n_2$ の標本を抽出することとし、

$$\mathbf{p} = \left(p_{j,1}, p_{j,2}, \cdots, p_{j,k-1}\right)^T$$

を j-th population の構成割合とする。このとき、帰無仮説

$$H_0 : \mathbf{p_1} = \mathbf{p_2}\ (= \mathbf{p})$$

を検定する事を考える。先のケースと同様に、$n_j \to \infty$ に対して、

$$\sqrt{n_j}\left(\hat{\mathbf{p}}_j - \mathbf{p}_j\right) \xrightarrow{L} N\left(\mathbf{0}, \mathbf{V}_j\right) \quad (j = 1,2)$$

を仮定する。このとき、

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

がある場合には、Wald Statistic

$$\bar{X}_{WH}^2 = \left(\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2\right)^T \left[\frac{\hat{\mathbf{V}}_\mathbf{1}}{n_1} + \frac{\hat{\mathbf{V}}_\mathbf{2}}{n_2}\right]^{-1} \left(\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2\right)$$

が $\chi_{k-1}^2$ に漸近的に従う事から、これを用いて検定が可能である。

## 4. 母集団が既知の場合の数値シミュレーション

単純な数値例として、カテゴリーの数 $k = 4$ とし、母集団が二つの層からなる場合を考える。この母集団から層化無作為法により標本抽出を行うこととし、次の3つのケースについて前節において提唱された検定統計量を算定し、理論値と比較する。

CASE1

第1層は母集団の大きさ $N^{(1)} = 5000$、標本の大きさ $n^{(1)} = 50$、第2層は母集団の大きさ $N^{(2)} = 10000$、標本の大きさ $n^{(2)} = 100$ とする。このケースでは第1層と第2層は等質的で、

$$p_1^{(j)} = 0.1,\ p_2^{(j)} = 0.2,\ p_3^{(j)} = 0.3,\ p_4^{(j)} = 0.4 \quad (j = 1,2)$$

であるとする。母集団が既知である事から $\mathbf{V}$ を実際に求める事ができ (石井 2004)、

$$\frac{\mathbf{V}}{n} = \sum_j \frac{N^{(j)2}}{N^2} \cdot \frac{N^{(j)} - n^{(j)}}{N^{(j)} n^{(j)}} \cdot \frac{N^{(j)}}{N^{(j)} - 1} \begin{bmatrix} p_1^{(j)}(1-p_1^{(j)}) & -p_1^{(j)}p_2^{(j)} & -p_1^{(j)}p_3^{(j)} \\ -p_2^{(j)}p_1^{(j)} & p_2^{(j)}(1-p_2^{(j)}) & -p_2^{(j)}p_3^{(j)} \\ -p_3^{(j)}p_1^{(j)} & -p_3^{(j)}p_2^{(j)} & p_3^{(j)}(1-p_3^{(j)}) \end{bmatrix}$$

$$= \begin{bmatrix} 0.0006 & -0.0001 & -0.0002 \\ -0.0001 & 0.0011 & -0.0004 \\ -0.0002 & -0.0004 & 0.0014 \end{bmatrix}$$

となる。

そこで、この母集団から標本抽出を10000回行うシミュレーションを実施し、先に述べた Wald Statistic $\bar{X}_w^2$ と通常の $\bar{X}^2$ を算定し、その分布を $\chi_3^2$ と比較した。その結果が図4である。

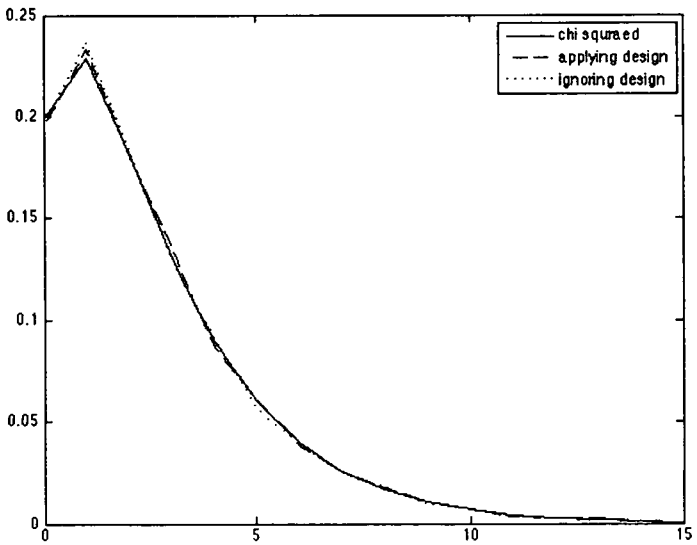


図4　シミュレーション結果 (CASE 1)

これによれば、実線で示した $\chi_3^2$ の分布に対し、破線で示した $\bar{X}_w^2$、点線で示した $\bar{X}^2$ ともにほとんど分布は一致している。このケースでは、各層が等質的かつ抽出率が等しく、抽出率が低いことから無限母集団に近いため、両者がほぼ同じものと考えられることがわかる。

CASE2

CASE2 では、第 1 層と第 2 層が等質的でなく、

$$p_1^{(1)} = 0.1,\ p_2^{(1)} = 0.2,\ p_3^{(1)} = 0.3,\ p_4^{(1)} = 0.4$$

$$p_1^{(2)} = 0.1,\ p_2^{(2)} = 0.6,\ p_3^{(2)} = 0.1,\ p_4^{(2)} = 0.2$$

を仮定した。このときのシミュレーション結果が図 5 である。

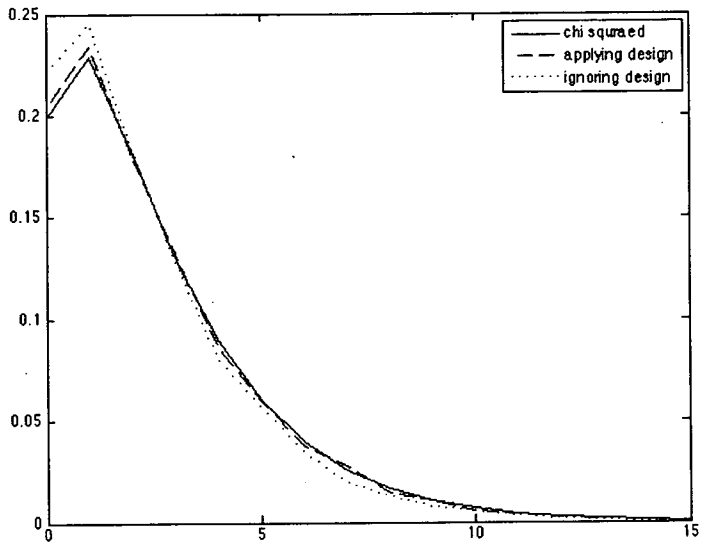


図 5　シミュレーション結果 (CASE 2)

これによれば、両者ともこのケースにおいても $\chi_3^2$ に概ね近いが、やや $\bar{X}^2$ の方が分布が低い方に集中する傾向が見られる。

CASE 3

CASE3 では CASE2 に加え、さらに抽出率を変え、第 1 層は母集団の大きさ $N^{(1)} = 5000$ に対して標本の大きさ $n^{(1)} = 60$、第 2 層は母集団の大きさ $N^{(2)} = 10000$ に対して標本の大きさ $n^{(2)} = 90$ とした。このときのシミュレーション結果が図 6 である。

このケースでは、$\bar{X}^2$ の分布が高い方に寄っていることが見られる。中高年縦断調査においても層間における抽出率の違いが存在していたことから、このような検定統計量への影響も考えられる。

CASE4

次に、有限母集団という特性をより考慮しなければならない状況として、第 1 層は母集団の大きさ $N^{(1)} = 5000$ に対し標本の大きさ $n^{(1)} = 500$、第 2 層は母集団の大きさ $N^{(2)} = 10000$ に対し標本の大きさ $n^{(2)} = 1000$ とする。このケースでは再び第 1 層と第

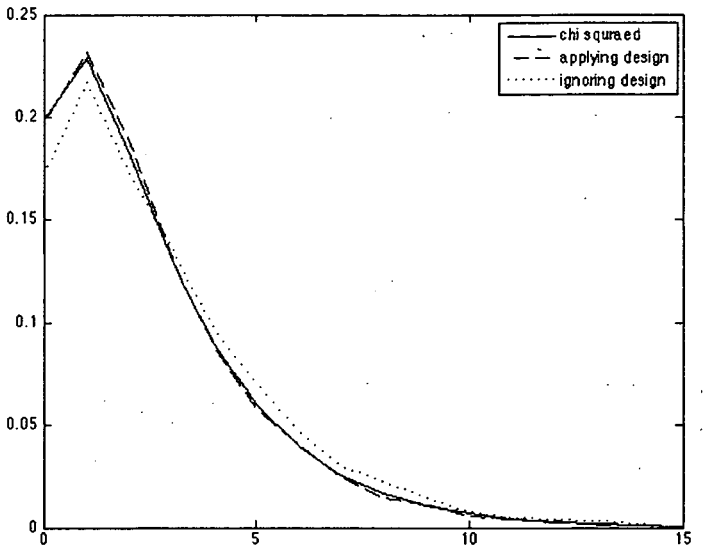


図6　シミュレーション結果 (CASE 3)

2層は等質的で、

$$p_1^{(j)} = 0.1,\ p_2^{(j)} = 0.2,\ p_3^{(j)} = 0.3,\ p_4^{(j)} = 0.4 \quad (j = 1,2)$$

であるとする。このときのシミュレーション結果が図7である。

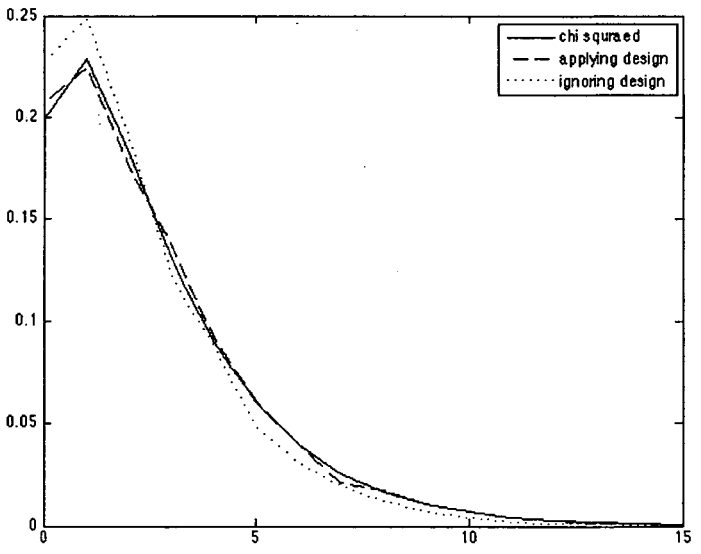


図7　シミュレーション結果 (CASE 4)

このケースでは、$\tilde{X}^2$ の分布が低い方に寄っていることが見られる。

## 5. おわりに

本稿では、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ることとし、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。なお、今回の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、$\mathbf{V}$ を求めることができたが、実際の標本調査では、標本からの推定値 $\hat{\mathbf{V}}$ に基づいて検定を行うこととなる。また、今回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考慮も必要となることから、今後、より実際の標本抽出に近い形でのシミュレーションの実行とともに、実データによる検証も必要となると考えられる。

なお、一般論として、成年者縦断調査など国民生活基礎調査を親標本とした調査のサンプリングデザインでは、有限母集団修正が小さいことなどから、サンプルを無限母集団からの単純無作為標本とみなすことにそれほど大きい問題はないと考えられ、また、出生児縦断調査や成年者縦断調査ではウエイトの問題もほとんど影響しないことから、本研究で対象としたカイ二乗検定などの分析に与える影響は限定的なものであると考えられる。

しかしながら、本研究において見た通り、中高年縦断調査ではややウエイトによる影響が想定されること、また、分析レベルが細かくなってきた場合にはサンプリングデザインが分析結果に与える影響を考慮しなければならない可能性もあり、これらの定量的な影響を予め把握しておくことは基礎的な研究として重要なものである。いずれにせよ、このような問題についてはわが国での実務的な研究蓄積が多くないことから、他の分析手法についての影響評価も視野に入れつつ、今後研究を深めていく必要があると考えられる。

## 参考文献


Holt, D., A. Scott, and P. Ewings (1980) "Chi-squared Test with Survey Data", *Journal of Royal Statistical Society A*, Vol. 143, No. 3, pp. 303–320.

石井太 (2004) 『よくわかる標本調査法 第2部標本設計理論編』, (財) 厚生統計協会.

金子隆一 (編) (2006) 『『パネル調査 (縦断調査) のデータマネジメント方策及び分析に関する総合的システムの開発研究』厚生労働科学研究統計情報高度利用総合研究事業, 総合研究報告書』.

Rao, J. and A. Scott (1981) "The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables", *Journal of the American Statistical Association*, Vol. 76, No. 374, pp. 221–230.

# （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 $\chi^2$ 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、

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

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 $\chi^2_9$ 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 $\chi^2_{k-1}$ 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$\bar{X}^2_{ws}$ 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。

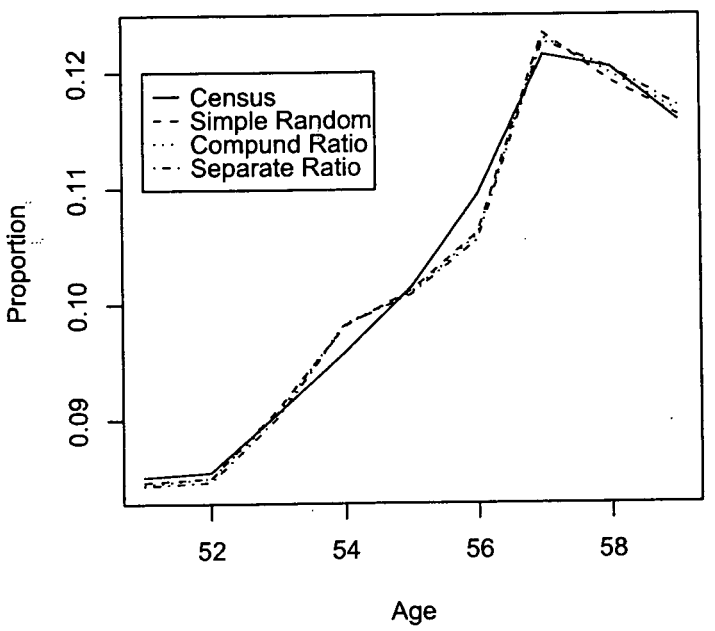


図1　年齢分布の推定量の比較

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 $\chi^2$ 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 $\chi^2$ 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)、

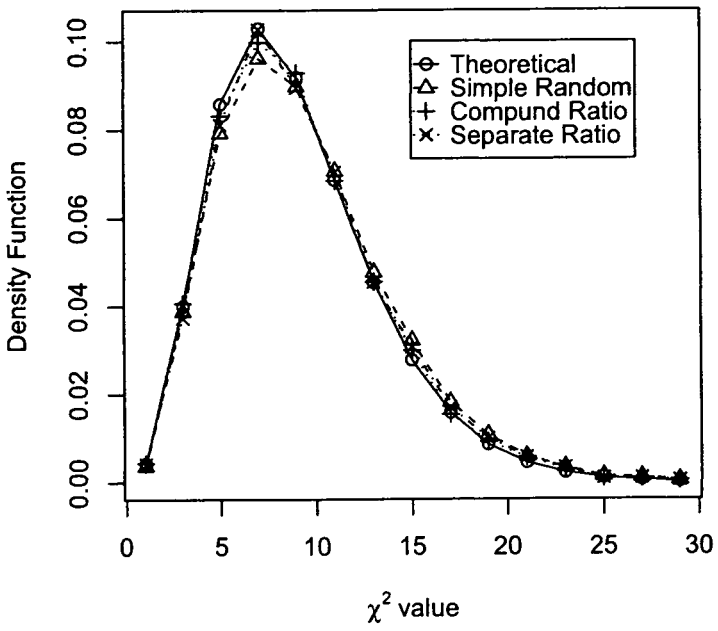


図2　検定統計量の密度関数の比較

分離比推定形式による推定量 (Separate Ratio) においても、理論値とのずれはややあるものの、無限母集団からの単純無作為標本とみなした場合 (Simple Random) のずれほどには大きくなく、標本設計を考慮せずに検定統計量を作成したことが影響を与えていることがわかる。表2はいくつかの p-value に対応する検定統計量の点を示したものであるが、ここからも Simple Random と Theoretical の乖離が最も大きいことがわかる。

| | 75% | 50% | 25% | 10% | 5% |
|---|---|---|---|---|---|
| Theoretical | 5.899 | 8.343 | 11.389 | 14.684 | 16.919 |
| Simple Random | 6.082 | 8.708 | 11.900 | 15.391 | 17.787 |
| Compound Ratio | 5.940 | 8.411 | 11.471 | 14.936 | 17.178 |
| Separate Ratio | 6.047 | 8.508 | 11.657 | 15.216 | 17.537 |

表2　検定統計量の p-value に対応する点

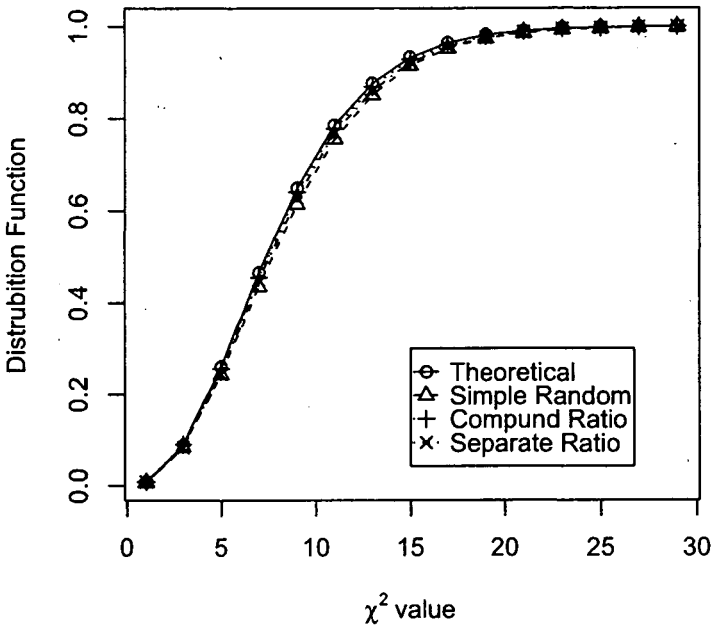


図3　検定統計量の分布関数の比較

## 4. おわりに

本研究においては、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った。

昨年度の考察において、一般論として、成年者縦断調査など国民生活基礎調査を親標本とした調査のサンプリングデザインでは、有限母集団修正が小さいことなどから、サンプルを無限母集団からの単純無作為標本とみなすことにそれほど大きい問題はないと考えられ、また、出生児縦断調査や成年者縦断調査ではウエイトの問題もほとんど影響しないことから、カイ二乗検定などの分析に与える影響は限定的なものであると考えられる一方で、中高年縦断調査ではややウエイトによる影響が想定されること、また、分析レベルが細かくなってきた場合にはサンプリングデザインが分析結果に与える影響を考慮しなければならなくなる可能性もあり、これらの定量的な影響を予め把握しておくことは基礎的な研究として継続していく必要性があることを指摘した。

本年度の研究結果は、この点について、一定の定量的分析を与えることができたものと考える。すなわち、年齢分布の適合度に関するカイ二乗検定については、中高年縦断調査

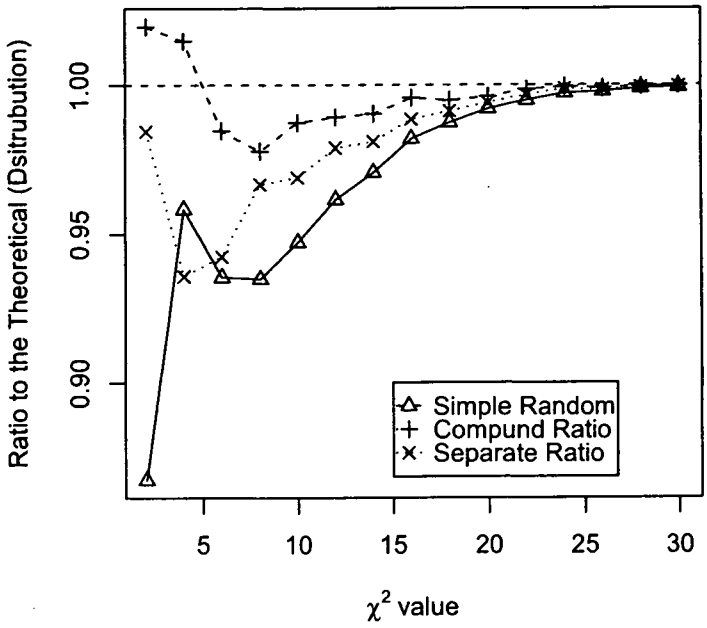


図4　検定統計量の分布関数の理論値に対する割合

においてサンプルを無限母集団からの単純無作為標本とみなした場合においても、一般的な検定統計量の分布と理論的なカイ二乗分布との乖離は、それほど大きなものではないということが観察された。しかしながら、その一方で、その影響は完全に無視できるというわけではない。サンプルを無限母集団からの単純無作為標本とみなした場合の検定統計量の分布は、有限母集団からの標本抽出と考え標本設計を考慮した検定統計量の分布に比べ、理論的なカイ二乗分布との乖離はより大きいものであることが観察され、細かい分析を行うなど、検定統計量の小さな違いを重視しなければならないケースにおいては、その影響が無視できなくなることもあり得ると考えられる。このような場合については、本研究で検討を行った、有限母集団からの標本抽出と考え標本設計を考慮した検定統計量を用いて検討を行うことが望ましいと言えよう。また、その場合、本研究において見た通り、推定方法によってもその検定統計量が異なることから、推定方法に見合った検定統計量を用いることが求められる。

中高年縦断調査は他の縦断調査に比べてその歴史がまだ浅く、十分なデータ蓄積には未だ至っていない状況ではあるものの、今後データが蓄積されてきた状態を想定し、高齢者の状況等に関して縦断調査というメリットを活かした分析手法を予め研究しておくことは、本調査結果を政策に利用していく上でも重要な課題である。このような観点から、本

研究においては、中高年縦断調査におけるカイ二乗検定に焦点を当てて分析手法の検討を行ってきたが、これ以外の分析手法について、同様の考察を行っていくことも今後必要であると考えられる。また、本研究で対象とした、年齢分布以外の分析対象についても、実データに基づく定量的評価を行い、さらなる検討を続けていくことも重要であろう。これらについては、今後の研究課題として引き続き取り組む必要があるものと考えられる。

## 参考文献

Holt, D., A. Scott, and P. Ewings (1980) "Chi-squared Test with Survey Data", *Journal of Royal Statistical Society A*, Vol. 143, No. 3, pp. 303–320.

石井太 (2004) 『よくわかる標本調査法 第2部標本設計理論編』, (財) 厚生統計協会.