IPSS Discussion Paper Series

(No.2025-J01)

二段階調査データの復元推計
生活と支え合いに関する調査における統合ウェイト法・キャリブレーション・分散推定

榊原賢二郎（国立社会保障・人口問題研究所）

2026 年 3 月

国立社会保障・人口問題研究所

〒100-0011　東京都千代田区内幸町 2-2-3
日比谷国際ビル 6F

二段階調査データの復元推計

生活と支え合いに関する調査における統合ウェイト法・キャリブレーション・分散推定

榊原賢二郎[1]

1. はじめに

本稿では、国立社会保障・人口問題研究所が概ね 5 年ごとに実施している「生活と支え合いに関する調査」(以下「支え合い調査」と記載する)を用いた統計的な推計手法について検討する。標本調査は、母集団の値を推計するために行われる。これは広く言えば復元推計と呼ばれるが、復元推計においては通常、標本設計に応じた推計方法が採用されるのはもちろんのこと、無回答などによって生じる真の値からのずれ(バイアス)への対処が行われる。その方法の一つとして、ベースウェイト・無回答補正・事後層化等から最終ウェイトを算出する方法がある。毛塚ほか(2024)は、この一連の方法を上記調査の世帯票に適用している。しかし、これを支え合い調査全体の復元推計に適用するには、まだ課題が残っている。この課題には、他調査にも該当する一般的なものと、支え合い調査の構造から生じるものがある。

一般的な課題としては、第一に、事後層化における小度数セルの回避と、変数の増加がある。事後層化においては、複数の変数をクロスさせたセルの度数分布(同時分布)をそれぞれ何倍かにして、母集団の既知の数値に合わせようとする。この時、度数が 0 や極めて小さいセルについては、倍率が設定できなかったり、極端に大きい倍率が現れたりする。対策の一つは、事後層化の層を統合することであるが、母集団に合わせる調整に使う変数(補助変数)の数を増やすことには困難が伴う。この課題に対応するために、本稿ではキャリブレーションという方法を採用する。キャリブレーションにも各種あるが[2]、ここでは補助変数一つずつの分布(周辺分布)を、既に知られている母集団の分布に合わせるキャリブレーションについて説明する。

第二に、推定値の誤差分散の計算が挙げられる。誤差の計算は、統計の精度を検証するのに不可欠である。しかし、復元推計値の誤差を正確に求める数式は通常存在せず、何らかの近似が必要となる。近似には、ジャックナイフ法や後述の均衡反復複製法(BRR)のような標本再抽出法(一度母集団から選ばれた標本から、さらに一部を取り出すことを繰り返す方法)と、線形化(近似式を求める方法)があるが、本稿では BRR を扱う。

第三に、一部の質問への無回答(項目無回答)への対応も課題である。すべての質問に共通のウェイトは、

---

[1] 国立社会保障・人口問題研究所社会保障応用分析研究部第三室長。なお、本稿は個人の見解を示すものであり、著者の所属組織の見解ではないことに留意されたい。

[2] 母集団における補助変数の周辺分布が既知である場合のほか、標本内の回答者・無回答者の双方で補助変数の情報が得られている場合にも、キャリブレーションが適用できる(Särndal and Lundström, 2005)。また注 10 も参照されたい。

調査項目すべてへの無回答(単位無回答)には対応するが、一部の質問のみへの無回答には対処できない。項目無回答が生じさせる偏りは、そうした無回答を補完する代入法で取り扱う必要がある。特に、分散を過小推計しない観点から、代入して無回答がなくなったデータ(完全データ)を複数作り、そこから推定値や分散を求める多重代入法が推奨されており、ここで取り上げる。

これらに加えて、支え合い調査の構造から生じる課題として、世帯と個人の二段階構造がある[3]。支え合い調査では、調査地区内のすべての世帯と、その世帯に属する 18 歳以上のすべての個人に対して調査協力を依頼する。そのため、世帯に関する情報とその世帯内の個人の情報の両方が得られる。こうした構造を持つ調査においては、世帯へのウェイトと個人へのウェイトを適切に関係付けることが求められる[4]。両者の関係に対して、何らかの制約を課してウェイトを計算することを統合ウェイト法(integrated weighting)と呼ぶ。この統合ウェイト法を用いつつ、世帯票のみならず個人票に対しても適切なウェイトを設定することが本稿の課題となる。

以下では、支え合い調査への適用を前提としたウェイトの計算方法を検討する。次節から順に、ベースウェイト・無回答ウェイト・キャリブレーション・多重代入・分散推定について説明し、その後この方法を 2022 年調査データに適用した結果を、国勢調査等の集計値と比較する。

集計値の改善を評価するにあたり、2022 年支え合い調査を業務利用した。推計には R を用いた。キャリブレーションの計算に sampling パッケージの calib 関数、BRR のウェイトの倍率の計算に as.svrepdesign 関数を用いた。

2. ベースウェイト

調査の標本設計は多様であり、各要素が厳密に同じ確率で選ばれるとは限らない。各要素が標本内に含まれる確率(包含確率)の違いを調整する乗率がベースウェイトであり、包含確率の逆数で求められる。先述の通り、支え合い調査は、全国の各都道府県を層として、各層から調査地区を抽出し、地区内の全世帯・18 歳以上世帯員を調査対象とする層化集落抽出法を採用している。層h地区i世帯jのベースウェイトd_jおよび世帯員(個人)kのベースウェイトd_kは、当該世帯・個人が属する層hにおける、全調査地区数N_h(本稿の文脈では国勢調査区数)を抽出地区数n_hで割った比である(Valliant and Dever, 2018, 22)。層化集落抽出で層内の各地区の包含確率が等しい場合、ある調査地区が選ばれると、その中のすべての世帯・個人が標本となるので、世帯・個人の包含確率は、調査地区のそれと等しい。したがって、その逆数であるベースウェイトも各水準で均しくなる。

$$d_j = d_k = \frac{N_h}{n_h}$$

[3] 抽出方法としては、調査地区を含め三段階あるが、データは二段階で構成され、また統合ウェイトの文脈で二段階抽出と呼ばれるものと同様に考えられるため、ここでは二段階と称している。

[4] 両ウェイトの関係に制約を課さない方法もないわけではない。この場合、世帯と個人それぞれでキャリブレーションを行う。しかし、世帯ウェイトと個人ウェイトに基づく推計値が一致しないという欠点がある(Haziza and Beaumont 2017)。

2022 年調査における、調査地区単位のベースウェイトの分布は以下の通りである [5]。

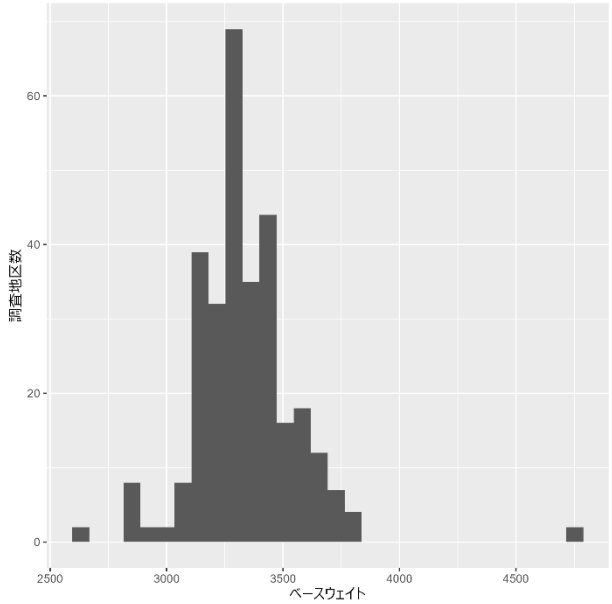

図 1　ベースウェイトの分布 [5]

表 1　ベースウェイトの基本統計量(地区ごと)[5]

平均	中央値	標準偏差	合計	最小値	最大値
3,337.42	3,305.23	227.36	1,001,226	2,659.50	4,779.50

注　鳥取県と島根県は、調査地区が 1 地区のみだったため、分散推定の便宜上統合している。以下同様。

3. 無回答ウェイト

各調査対象が無回答となる傾向には、調査対象者の属性等による差があるため、補正する必要がある。例えば、性別・年齢・市郡区分・地方区分・住居形態などの属性が、調査票の回収不能に関連している(土屋, 2005)。無回答ウェイトは、回答確率の逆数として求められるが、この回答確率は未知である。そこで、回答メカニズムについてのモデルを仮定し、説明変数を選択した上で、回答確率を推定するという形をとる。無回答ウェイトの設定方法は、回答確率の推定方法に応じて複数ある。回答確率の回帰モデルもその一つである。ここでは、より単純なウェイティングクラス(Valliant and Dever, 2018, 34)ないし回答均質集団(response homogeneity group) (Särndal and Lundström, 2005, 53)を用いる。何らかの分類に沿って、標本のベースウェイトの和を、回答者(世帯・個人)のベースウェイトの和で割ると、無回答ウェイトが求められる。ここでは、地域の生活水準の目安として、生活保護級地区分(厚生労働省, 2025)を用いることとする。これは、1 級地-1 から 3 級地-2 の 6 区分から成る市町村の分類であり、生活保護費の算出に用いられる。政令市・その他の市・町・村のような分類と異なる点として、例えば町村同士が合併し

[5] 本稿の図表はすべて、2022 年生活と支え合いに関する調査の個票データを業務利用して、著者が作成した。

て市になっても、元の区分が一定程度維持されることが挙げられる。そのため、この区分は生活水準や都市化の度合いを把握することに資する。生活保護級地区分を無回答補正に用いると、同一地区内の世帯には同一の無回答ウェイトが割り当てられることになる点で、無回答ウェイトとしては特殊である。層h地区iの世帯jは、いずれもベースウェイトd_jが同一であるとし、当該地区の級地区分がcであったとする。当該層の中で級地区分がcである地区の集合をs_{hc}、地区iの標本世帯数をn_{hi}、回答世帯数をm_{hi}とすると、世帯無回答ウェイトは以下の通りである。

$$a_j = \frac{\sum_h \sum_{i \in s_{hc}} n_{hi} d_{hi}}{\sum_h \sum_{i \in s_{hc}} m_{hi} d_{hi}}$$

個人無回答ウェイトa_kは、ここでは各世帯の世帯員の回答割合の逆数とする。層h地区i世帯jにおいて、世帯員数をn_{hij}、回答世帯員数をm_{hij}とすると、個人無回答ウェイトは世帯内で同一となる。

$$a_k = \frac{n_{hij}}{m_{hij}}$$

ただし、これをそのまま適用すると、ウェイトの分布の分散を著しく拡大することが確認されたため、上限(ここでは 3)を設けた。

後で、世帯票・個人票のキャリブレーションを一体的に行う時にも参照するため、世帯の無回答ウェイトに対する個人の無回答ウェイトの比を定義しておく。

$$a_{k|j} = \frac{a_k}{a_j}$$

これは、世帯jの世帯票が存在する時に個人kが回答する条件付き確率の逆数であるが、世帯票がない場合にも定義される。

級地区分ごとの世帯無回答ウェイトは以下の通りであった。大都市部で回収率が低く、その逆数として設定した世帯無回答ウェイトが大きくなっていることが分かる。

表 2　各地区の級地区分ごとの無回答ウェイト

1 級地-1	1 級地-2	2 級地-1	2 級地-2	3 級地-1	3 級地-2
2.43	2.00	1.82	1.57	1.86	1.55

表 3　世帯無回答ウェイトの基本統計量 [5]

平均	中央値	標準偏差	合計	最小値	最大値
1.97	1.86	0.30	17,212.54	1.55	2.43

個人無回答ウェイトは以下の通りである。各世帯内の回収率の逆数を補正倍率としたことから、離散的な分布になっているが、標準偏差は平均に対して著しく大きな値にはなっていないと考えられる。

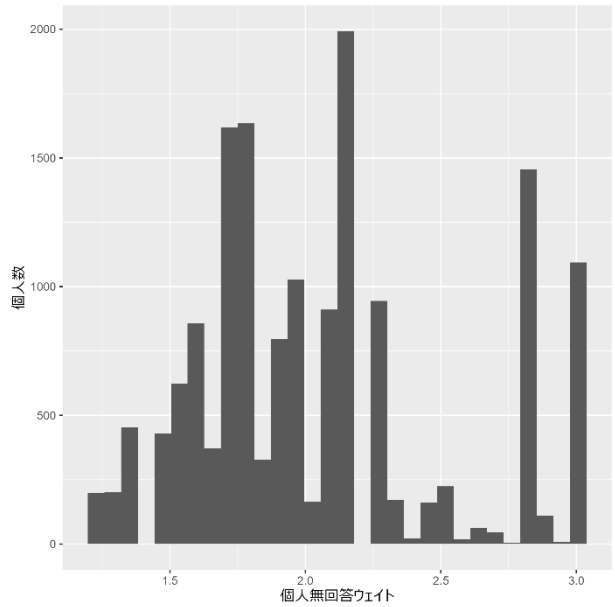

図 2　個人無回答ウェイトの分布

表 4　個人無回答ウェイトの基本統計量 [5]

平均	中央値	標準偏差	合計	最小値	最大値
2.05	1.97	0.47	32,722.88	1.22	3.00

4. キャリブレーションと統合ウェイト法

続いて、ベースウェイトと無回答ウェイトの積をさらに補正することで、ウェイトをかけた合計値が、母集団の値と見なしうる数値(国勢調査の集計値等)の合計値と一致するように調整する。調査間の一致の他に、無回答バイアスやカバレッジエラー[6]の低減、推計値の誤差分散の縮小にも寄与する(Haziza and Beaumont 2017)。

ここでは、国勢調査における以下の各世帯数・個人数に合わせて、支え合い調査のキャリブレーションを行う。

- 都道府県別世帯数 [7]
- 世帯人員数別世帯数(世帯規模 1～5 人のそれぞれと 6 人以上)
- 住宅所有形態(持ち家・民営賃貸・その他)別世帯数 [8]

[6] 推定を行う対象である母集団と、名簿のように、そこから標本を抽出してくるところの集団(標本抽出枠)とのずれを指す。

[7] ここでも島根県と鳥取県は統合した。

[8] なお、この変数には少数の項目無回答が存在する。通常キャリブレーションは、項目無回答がない変

■　個人票(18 歳以上)による性・年齢(原則 5 歳階級)ごとの人数[9]

これらを、令和 2 年国勢調査(総務省統計局 2022)の一般世帯における集計値に合わせる。支え合い調査の集計値は一般調査区(寮を含む)の世帯に基づき、他方ここで参照する国勢調査の集計結果はすべての調査区の一般世帯に準拠するので、わずかなずれが生じる。しかし、これはカバレッジエラーに相当するので、キャリブレーションで対応できる。

先述したように、この補正には、事後層化という方法もある。複数の変数があった時に、それをすべてクロスさせたセル度数を合わせるものである[10]。ただ、先述の通り、度数 0 ないし小度数のセルが問題となる。例えば、各都道府県ごとに世帯人員数を国勢調査のものと合わせようとすると、世帯数が少ない大規模世帯が問題となる。国勢調査より規模が小さい支え合い調査の側で、いくつかの都道府県の大規模世帯数が 0 となり、ウェイトの合計が母数(ここでは国勢調査一般世帯数)を下回るという、望ましくない状況が生じうる。対応方法の一つは、セルを統合することであるが、支え合い調査側のセル度数が大きくない状況には変わりなく、潜在的な誤差が大きい。また、小度数セルの問題があるので、多変数への拡張が難しい。

ここでは、周辺度数(一変数ごとの各カテゴリーの度数)を母集団に合わせるため、キャリブレーションを採用する(Deville and Särndal 1992; Deville et al. 1993)。世帯および個人に関する各変数の母集団合計を並べたベクトルをそれぞれ$\boldsymbol{t}_{\boldsymbol{x}_J}$および$\boldsymbol{t}_{\boldsymbol{x}_K}$とする。$\boldsymbol{t}_{\boldsymbol{x}_J}$および$\boldsymbol{t}_{\boldsymbol{x}_K}$は、世帯・個人について補正に使う変数の数と同じ要素数を持つベクトルである。以下のキャリブレーション方程式を満たす世帯最終ウェイトw_jと個人最終ウェイトw_kを求める。

$$\sum_{h,i}\sum_{j\in r_{hi}} w_j \boldsymbol{x_j} = \boldsymbol{t}_{\boldsymbol{x}_J}$$

$$\sum_{h,i}\sum_{j\in r_{hi}}\sum_{k\in r_{hij}} w_k \boldsymbol{x_k} = \boldsymbol{t}_{\boldsymbol{x}_K}$$

ただし、r_{hi}は、層h地区i内で回答した世帯の集合、r_{hij}は層h地区i世帯j内で回答した個人の集合である。

数で行うので、これは問題となる。しかし、世帯番号を割り振る際、集合住宅内の住戸や戸建て地域の隣家は前後に並ぶことが多いと想定されるため、居住形態の項目無回答を前後の回答を用いて補完した。前後が相違する場合は、世帯員数や世帯番号が近い方で補完した。当然ながら、補完しない元変数は、このキャリブレーションによっても厳密には国勢調査と一致しないが、項目無回答がそこまで多くなく、補完がある程度の妥当性を持っている場合、おおよそ国勢調査の数値を近似できると考えられる。国勢調査側にも少数の無回答(不詳)が存在するが、居住形態については不詳を按分しなかった。

[9] 国勢調査の性年齢別個人数については、無回答者の性年齢分布を推定した不詳補完結果が公表されているが、一般世帯限定の集計表がないので、不詳を補完していない性年齢別一般世帯の表を用い、不詳を按分した。

[10] キャリブレーションにおいても、同時分布(各変数をクロスさせたセル度数)を調整することもできる。ただし、本稿で実施しているように、キャリブレーションは複数の周辺分布を同時に調整することができる点で、事後層化と異なる。

ここで、w_jとw_kの関係に以下の制約を加える(Estevao and Särndal 2006)[11]。

$$w_k = d_{k|j} a_{k|j} w_j$$

ここで$d_{k|j}$は、世帯jが標本として選ばれた時に個人kが標本として選ばれる条件付き確率の逆数である。支え合い調査においては、標本として選ばれた世帯の全員を個人票の対象とするため、$d_{k|j} = 1$であり、$w_k = a_{k|j} w_j$である。このように、世帯と個人という二つの段階のウェイトの関係が制約されるようなウェイトの定め方を、統合ウェイト法と呼ぶ(Lemaître and Dufour 1987)。

この関係をキャリブレーション方程式に代入すると以下のようになる(Estevao and Särndal 2006)。

$$\sum_{h,i} \sum_{j \in r'_{hi}} w_j \left(\sum_{k \in r'_{hij}} a_{k|j} \boldsymbol{x'_k} \right) = \begin{pmatrix} \boldsymbol{t}_{x_J} \\ \boldsymbol{t}_{x_K} \end{pmatrix} = \boldsymbol{t_x}$$

ここで、支え合い調査では、対応する世帯票がない個人票やその逆が存在することに留意する必要がある。そこで、世帯票回答世帯の集合に、個人票のみ回答世帯(いずれか一名以上の世帯員が回答している世帯)を加えてr'_{hi}とし、この世帯のうち、世帯jの回答世帯員の集合をr'_{hij}とする。そして、$\boldsymbol{x'_j} = \boldsymbol{x_j}\ (j \in r_{hi}), \boldsymbol{x'_j} = \boldsymbol{0}\ (j \notin r_{hi})$と置く。個人も同様。こうすると、回答していない世帯・個人は合計に影響を与えず、かつ一括してキャリブレーションを行うことができる。

ここから先は、個人票の変数を$\sum_{r'_{hij}} a_{k|j} \boldsymbol{x'_k}$のように、各世帯内で総和にして新たな説明変数とすることで、通常のキャリブレーションに持ち込める。改めて世帯jの補助変数ベクトルを$\boldsymbol{x_j}$とする。ここには個人単位の補助変数の世帯単位の合計値が含まれている。キャリブレーション方程式は、以下の通りである。

$$\sum_j w_j \boldsymbol{x_j} = \boldsymbol{t_x}$$

この式を満たすウェイトは、無数に存在する(通常、変数の数は回答世帯数より少ないため)。そこで、キャリブレーション前のウェイトになるべく近い最終ウェイトを求めることとする。この近さを精確に表すと、何らかの距離の関数Gを使って、

$$\sum_j d_j G\left(\frac{w_j}{d_j}\right)$$

を最小化し、かつキャリブレーション方程式を満たすようにw_jを定める(Deville and Särndal 1992; Deville et al. 1993)。ここにGは$G(1) = G'(1) = 0, G''(1) = 1$を満たす 2 階微分可能な狭義凸関数である。ラグランジュの未定乗数法により、$\boldsymbol{x_j}$と同じ要素数のベクトル$\boldsymbol{\lambda}$を使い、

$$\sum_j d_j G\left(\frac{w_j}{d_j}\right) - \boldsymbol{\lambda}^t \left(\sum_j w_j \boldsymbol{x_j} - \boldsymbol{t_x} \right)$$

---

[11] 無回答ウェイト$a_{k|j}$は著者が追加した。なお、ウェイト間の関係には以下のような定め方もある(Estevao and Särndal 2006)。ただし無回答補正は省略し、世帯規模をn_{hij}、世帯員集合をs_{hij}とする。

$$\sum_{k \in s_{hij}} w_k = n_{hij} w_j$$

を最小にするw_jと$\boldsymbol{\lambda}$を求める。ここでGの導関数をgとする。w_jで偏微分して 0 とおく[12]。

$$g\left(\frac{w_j}{d_j}\right) - \boldsymbol{\lambda}^t \boldsymbol{x_j} = g\left(\frac{w_j}{d_j}\right) - \boldsymbol{x_j^t}\lambda = 0$$

gの逆関数が存在するものとし、それをFとすると、

$$w_j = d_j F(\boldsymbol{x_j^t \lambda})$$

となる。これをキャリブレーション方程式に代入すると、

$$\sum_j d_j F(\boldsymbol{x_j^t \lambda}) \boldsymbol{x_j} = \boldsymbol{t_x}$$

これを$\boldsymbol{\lambda}$について解くと、キャリブレーションされた最終ウェイトが求まる。

距離関数の取り方には何種類かある(Deville and Särndal 1992; Haziza and Beaumont 2017)。

- 線形関数—GREG(一般化回帰)推定量: $G(u) = \frac{1}{2}(u-1)^2, F(u) = 1 + u$
- 乗法関数—レイキング比推定量: $G(u) = u\log u - u + 1, F(u) = e^u$
- 最尤レイキング: $G(u) = u - 1 - \log u, F(u) = \frac{1}{1-u}$
- 切断線形関数: $G(u) = \frac{1}{2}(u-1)^2\ (L \le u \le U), G(u) = \infty (u < L, u > U),$
  $F(u) = 1 + u\ (u \in [L-1, U-1]), F(u) = L\ (u < L-1), F(u) = U\ (u > U-1)$
- ロジット関数: $G(u) = \frac{1}{A}[(x-L)\log\frac{x-L}{1-L} + (U-x)\log\frac{U-x}{U-1}]\ (L < u < U), G(u) = \infty\ (u \le L, u \ge U), F(u) = \frac{L(U-1)+U(1-L)\exp(Au)}{U-1+(1-L)\exp(Au)}, A = \frac{U-L}{(1-L)(U-1)}$

ここに$L < 1 < U$である。

GREG 推定量は負のウェイトを、レイキング比推定量は極端に大きなウェイトをもたらす可能性がある。切断線形関数およびロジット関数は、ウェイトの補正倍率に下限と上限を設けて、極端なウェイトが生じないようにする方法である。ここではロジット関数によるキャリブレーション関数を採用する。ここでは下限 0、上限 10 で計算したが、より狭く、下限 1/5、上限 5 等もありうる。

キャリブレーション方程式を$\boldsymbol{\lambda}$についいて解く必要があるが、$\boldsymbol{\lambda}$を解析的に求める数式が存在するのは GREG 推定量のみであり、それ以外は近似計算で数値解を求める必要がある。数値解の計算法として有名なのが Newton-Raphson 法であるが、ここでは多変数に適用する。キャリブレーション方程式の左辺を$\boldsymbol{\lambda}$の関数

$$\boldsymbol{\phi}(\boldsymbol{\lambda}) = \sum_j d_j F(\boldsymbol{x_j^t \lambda}) \boldsymbol{x_j}$$

と置き、その$\boldsymbol{\lambda}$による偏微分を

$$\boldsymbol{\phi}'(\boldsymbol{\lambda}) = \sum_j d_j F'(\boldsymbol{x_j^t \lambda}) \boldsymbol{x_j x_j^t}$$

[12] $g\left(\frac{w_j}{d_j}\right)$に代えて、$g\left(\frac{w_j}{d_j}\right) / q_k$という定式化もあるが(Deville and Särndal 1992)、ここでは。$q_k = 1$として割愛する。このとき関数Fは、$F(0) = 1, F'(0) = q_k = 1$という条件を満たす。

その逆行列を$\phi'^{-1}(\boldsymbol{\lambda})$と表す。初期値を与えて以下の式に基づいて近似解を求める。

$$\hat{\boldsymbol{\lambda}}_{\boldsymbol{n+1}} = \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}} + \phi'^{-1}(\hat{\boldsymbol{\lambda}}_{\boldsymbol{n}})\left(\boldsymbol{t_x} - \phi(\hat{\boldsymbol{\lambda}}_{\boldsymbol{n}})\right) \ (n = 0, 1, 2, \dots)$$

$\hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}$ないし $\phi(\hat{\boldsymbol{\lambda}}_{\boldsymbol{n}})$の変化量がある閾値を下回ったら反復計算を終了する。

なお、初期値$\hat{\boldsymbol{\lambda}}_{\boldsymbol{0}} = \boldsymbol{0}$とすると、$\hat{\boldsymbol{\lambda}}_{\boldsymbol{1}}$は GREG 推定量に対応する(Deville and Särndal 1992, 380)。収束した時の$\boldsymbol{\lambda} = \hat{\lambda}$を使って、最終ウェイトを求める。

キャリブレーションによる推定値は、母集団の各要素について、以下の条件を満たす**$\boldsymbol{\lambda}$と同じサイズの**ベクトル$\boldsymbol{\alpha}$が存在する時には、真の値に収束する(一致推定量となる)(D'Arrigo and Skinner 2010, 183)。

$$F(x_i^t \boldsymbol{\alpha}) = q_i^{-1}$$

ただし、q_i^{-1}は、各要素の回答確率の逆数である。事後層化におけるように、補助変数x_iにより分類された各集団内で回答確率が均一であれば、こうした$\boldsymbol{\alpha}$は必ず存在する(Särndal and Lundström, 2005)。そのためこの前提は特異なものではない。

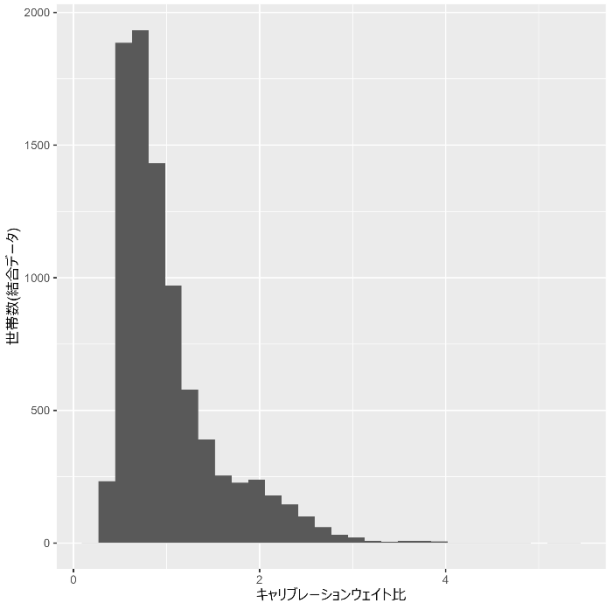

図 3　キャリブレーション倍率の分布

表 5　キャリブレーション倍率の基本統計量 [5]

平均	中央値	標準偏差	合計	最小値	最大値
1.01	0.84	0.55	8,808.18	0.25	5.43

図 3・表 5 は、ベースウェイト・無回答ウェイトに対する補正倍率の分布である。世帯無回答ウェイトで回収率の相違に一定程度対応しているため、キャリブレーション倍率の平均はほぼ 1 となっている。右裾が長い分布であるが、中心性が見て取れる。

以上の手続きを経て得られた最終ウェイトを図 4 および図 5 に示す。ウェイトの性能を示す値として、不等加重効果(unequal weighting effect)がある(Valliant and Dever, 2018, 71)。ウェイトの変動係数

(標準偏差/平均)をcvと表すと、不等加重効果は$1+cv^2$で定義される。この値が大きいほど、ウェイトがばらついており、そのことが推計値の分散をより大きくすることを意味する。一律の判断基準が存在するわけではないものの、世帯最終ウェイト・個人最終ウェイトのいずれも、許容可能であると思われる。なお、小数点計算の精度の問題で、個人最終ウェイトの合計値については、一の位の正確性は保証されていない。

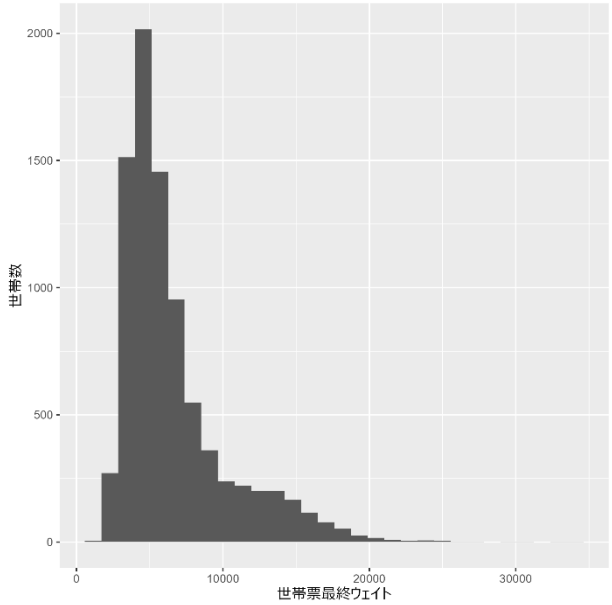

図 4　世帯最終ウェイトの分布

表 6　世帯最終ウェイトの基本統計量[5]

平均	中央値	標準偏差	合計	最小値	最大値
6,574.41	5,413.37	3,662.91	55,704,949	1,446.32	34,382.05
不等加重効果		1.31			

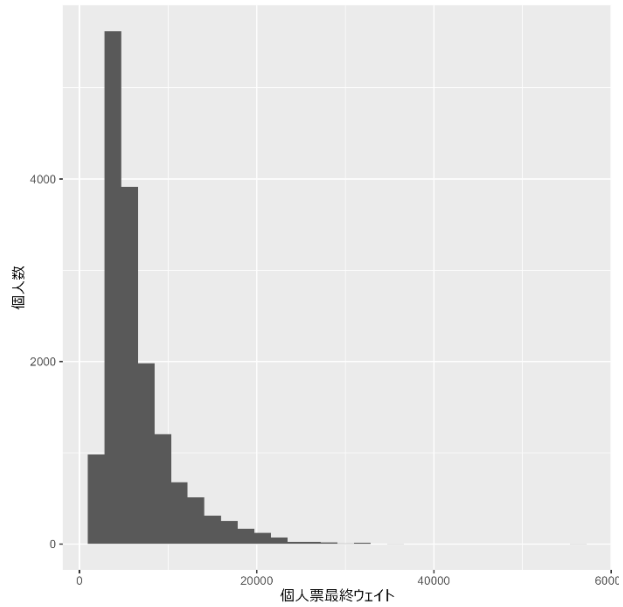

図 5　個人最終ウェイトの分布

表 7　個人最終ウェイトの基本統計量 [5]

平均	中央値	標準偏差	合計	最小値	最大値
6,564.87	5,171.96	4,345.54	104,571,826	1,105.51	55,528.16
不等加重効果		1.44			

注: 合計の数値は、一の位の精度を保証するものではない。

5. 多重代入

以上で得られた最終ウェイトは、すべての質問に無回答である場合(単位無回答)に対処するが、個別の質問に回答しない項目無回答には対応していない。項目無回答への対応には、代入法が用いられる。データが得られていない(欠測している、ここでは無回答)項目について、単に除去するのではなく、予測値で置き換えることで、妥当な統計的推測を可能にしようとするものである。単位無回答にウェイト、項目無回答に代入を適用する方法は、複合法(combined approach)とも呼ばれる(Särndal and Lundström, 2005, 155)。

代入に当たっては、条件付きで無作為な欠測(Missing At Random = MAR)の仮定が置かれる。これは、データが得られる条件付き確率について、欠測データを無視できるというものである。データ行列を$\boldsymbol{D}$、そのうち観測データを$\boldsymbol{D_{obs}}$、観測確率の行列を$\boldsymbol{K}$とする時、$\Pr(K|D) = \Pr(K|D_{obs})$が満たされることである。これを前提として、観測データから欠測データの分布を予測し、そこから欠測のない完全データを構築する。

代入法には、単一代入と多重代入がある。前者は、説明変数によって欠損値を予測し、予測値一つで置

き換えるものである。この方法では、分散を過小評価するという問題がある。そこで、複数の組の代入値で元のデータを置き換えた、複数のデータセットを生成し、そこから推定値と分散を計算する。複数のデータから得られる推定値のばらつきを活かすことで、分散に欠測データの不確実性を反映し、分散を過小評価することを防ぐことができる。

代入済みデータセット数をM、m番目の完全データセットに基づく推定値を$\hat{\theta}_m$、その分散の推定値を$\hat{V}(\hat{\theta}_m)$とする。このM組の推定値を統合して得られる推定値$\bar{\theta}$および分散の推定値Tは以下の通りである(Rubin 1987)。

$$\bar{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$$

$$T = W + \left(1 + \frac{1}{M}\right)B$$

ここで、

$$W = \frac{1}{M}\sum_{m=1}^{M}\hat{V}(\hat{\theta}_m), B = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m - \bar{\theta}\right)^2$$

多重代入のアルゴリズムは複数ある。ここではカテゴリカルデータにも適用可能な MICE (van Buuren 2018)を使う。標準では、ある変数を他のすべての変数で予測するようになっているが、計算負荷が大きい上に、潜在的に観測 0 や小度数セルを多く含むクロス表を扱うことになり、推定上の問題が生じる。そこで、カテゴリカル変数を前提として、各変数と他の変数の組み合わせから、観測 0 を含む組み合わせを排除した上で、関連の強さを示すクラメールのVが最も大きい変数最大 2 個を予測変数として投入することとした。

6. 分散推定

上記で得られた代入データとキャリブレーションウェイトを用いることで、補助変数以外の変数$\boldsymbol{y}$の合計値は以下で推計できる。

$$\hat{t}_{yCAL} = \sum_j w_j \boldsymbol{y_j}$$

先述の通り、この推計値の誤差分散を正確に計算する数式は、一般に存在しない。ここでは BRR (Balanced Repeated Replication) (Valliant and Dever, 2018, 89-94; Wolter, 2007, 107-150)という方法を用いる。元々この方法は、各層から 2 件ずつ抽出する標本設計を前提とした誤差の推定方法である。L個の層のそれぞれから、2 つの要素のうちいずれかを取り出す方法は2^L通り存在する。取り出した要素のウェイトを 2 倍した上で、推計値$\hat{\theta}_\alpha$を計算する。全標本による推計値(または$\hat{\theta}_\alpha$の平均)を$\hat{\theta}^*$とする。この時、推計値の誤差分散は以下のように推定される。

$$\hat{V}(\hat{\theta}) = \frac{1}{2^L}\sum_{\alpha=1}^{2^L}\left(\hat{\theta}_\alpha - \hat{\theta}^*\right)^2$$

実際には、層が多いと、2^L回の推計は容易ではない。そこで Hadamard 行列という、$+1$と-1を要素に

持ち、各列(または各行)が直交する行列を活用する。この行列の要素の値を参照しながら、各層からどの要素を選択するかを決定する。これにより、層の数より大きい 4 の倍数組の推計値を計算すれば済み、計算量が軽減される[13]。この 4 の倍数をAとする。A行A列の Hadmard 行列を$\boldsymbol{H}$とし、そのi行j列の要素を$\delta_{ij} \in \{1, -1\}$とする。各列の直交性は、任意の列$j, j'(j \neq j')$が以下を満たすことを意味する。

$$\sum_{i=1}^{A} \delta_{ij}\delta_{ij'} = 0$$

この行列を、各層 2 要素からいずれを再抽出するかと結びつける。$\boldsymbol{H}$のうちL列分を使用する。各列を層、各行を半標本(標本から半分再抽出したもの)と対応させる。α番目の半標本において、$\delta_{\alpha h} = 1$であれば層hから 1 番目の要素を選択し、$\delta_{\alpha h} = -1$であれば層hから 2 番目の要素を選択する。選択された要素のウェイトを 2 倍し、選択されなかった要素のウェイトを 0 とする。これにより得られる推計値を$\hat{\theta}_\alpha$とすると、推計値$\hat{\theta}$の分散の不偏推定量は、以下の式で得られる。

$$\hat{V}(\hat{\theta}) = \frac{1}{A}\sum_{\alpha=1}^{A}\left(\hat{\theta}_\alpha - \hat{\theta}^*\right)^2$$

以上は標準的な BRR の方法であり、要素が選択されたかどうかに応じて、元のウェイトの 0 倍または 2 倍にウェイトを変更することに均しい。これに対して、ウェイトの変化幅として別の倍率を使うこともできる。例えば、元のウェイトのρ倍と$2 - \rho$倍$(0 \leq \rho < 1)$を使うことができる(Fay の BRR)。この時、誤差分散の推定値は以下のようになる。

$$\hat{V}(\hat{\theta}) = \frac{1}{A(1-\rho)^2}\sum_{\alpha=1}^{A}\left(\hat{\theta}_\alpha - \hat{\theta}^*\right)^2$$

Fay の BRR には、母集団の一部(ドメイン)についての推計を行う際、ドメインが小規模であっても、誤差分散を安定的に計算できるという長所がある(Valliant and Dever, 2018, 90)。本稿では$\rho = 0.5$とした Fay の BRR を用いる。

各層の要素数が 2 を超える場合には、いくつかの要素をまとめて各層 2 グループとする方法や、2 要素ずつ擬似的な層に分割する方法がある(Wolter, 2007, 128-138)。本稿ではグループ BRR を採用した。

無回答ウェイトは、厳密に言えば各組の推計値を計算する時に再計算する必要があり(Wolter, 2007, 138)、本稿でもこれに従った。これにより、無回答ウェイトがもたらす分散の増大をより良く反映することができる。

---

[13] 各列の直交性だけであれば、層の数以上の 4 の倍数で良い。これに加えて各列の総和が 0 という条件も満たすようにするためには、層の数より大きい 4 の倍数が必要となる。すぐ後で言及する記法に基づくと列の総和が 0 という性質は以下のように表現される(jは任意の列)。

$$\sum_{i=1}^{A} \delta_{ij} = 0$$

この条件が満たされると、半標本の推計値の平均が標本全体からの推計値と一致する(Wolter, 2007, 112)。

7. 2022 年調査に基づく復元推計

以上の手続きで得られた推計値の性能を評価するため、国勢調査等の推計値と比較したのが表 8 である。比較可能な項目として、生活保護世帯数と就業者数を選んだ。生活保護世帯数については、被保護者調査月次確定値の結果の概要(厚生労働省社会・援護局保護課, 2022)を参照した。就業者数については、国勢調査においても無回答が存在するため、統計局が公表している不詳補完結果(総務省統計局, 2022, 令和 2 年国勢調査に関する不詳補完結果→労働力状態・産業・職業・従業上の地位の不詳補完→表 1)を用いた。

更なる比較対象として、支え合い調査に基づき、単純無作為抽出および層化集落抽出を前提とした場合の推計値と誤差を求めた。すなわち、多重代入で無回答を補完した完全データのそれぞれにおいて、無回答補正・キャリブレーションはかけず、単純無作為抽出および層化集落抽出の推定値・分散を求める。各完全データで得られた推定値・分散は、多重代入の項で言及した方法に従い統合する。単純無作為抽出を仮定する場合、カテゴリカル変数の特定のカテゴリーの選択比率 $\hat{p}$、標本規模n、母集団規模Nとすると、母集団の推定値は$N\hat{p}$、分散の推定値は$N\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$である。

層化集落抽出に関しては、ここでは複合比推定(厚生統計協会 2004, 214)を用いる。変数x, yを考える。ただしyは各調査地区の世帯数または個人数を調整するための変数であり、$y = 1$である。xがカテゴリカル変数の場合、各カテゴリーが選択されている場合に 1、それ以外の場合は 0 のダミー変数とする。層(都道府県)hの集落(調査地区)iにおけるxの総和を$t_{x_{hi}}$、層(都道府県)hの総和をt_{x_h}とし、yも同様とする。各調査地区の母集団・標本における集落数をそれぞれM_h, m_hとし、すべての層のM_hの合計をMとする。xの 1 集落あたり総和の標本推定値 $\bar{t}_x$を以下の式で定める。yについても同様とする。

$$\bar{t}_x = \frac{1}{M}\sum_h \frac{M_h}{m_h} t_{x_h}$$

母集団における総世帯ないし個人数をt_yとすると、複合比推定の推定値は$\hat{T}_x^C = t_y \frac{\bar{t}_x}{\bar{t}_y}$で与えられる。また、層$h$における$t_{x_{hi}}, t_{y_{hi}}$の標本分散をそれぞれ$S_{x_h}, S_{y_h}$、$t_{x_{hi}}, t_{y_{hi}}$の標本共分散を$S_{x_h y_h}$とすると、複合比推定の推定値の標準誤差は以下の式で近似される。

$$\sqrt{\hat{V}(\hat{T}_x^C)} \approx M\bar{t}_x \sqrt{\sum_h \left(\frac{M_h}{M}\right)^2 \left(\frac{1}{m_h} - \frac{1}{M_h}\right)\left(\frac{S_{x_h}}{\bar{t}_x^2} - 2\frac{S_{x_h y_h}}{\bar{t}_x \bar{t}_y} + \frac{S_{y_h}}{\bar{t}_y^2}\right)}$$

なお、推定値$\hat{T}_x^C$は、層化集落抽出の単純推定に比べると分散が小さい(効率的である)が、偏りがある。偏りの推定値は以下の式で与えられる。

$$\widehat{Bias}\left(\hat{T}_x^C\right) \approx M\bar{t}_x \sum_h \left(\frac{M_h}{M}\right)^2 \left(\frac{1}{m_h} - \frac{1}{M_h}\right)\left(\frac{S_{y_h}}{\bar{t}_y^2} - \frac{S_{x_h y_h}}{\bar{t}_x \bar{t}_y}\right)$$

Mが大きい場合や、$t_{y_{hi}}$の層内の変動が少ない場合には、偏りは小さくなる。上式により求めた偏りは、生活保護世帯数について 321.78、就業者個人数について-182.59 であった。これは推計値に比べると十分小さく、無視可能である。

表 8　復元推計結果と令和 2 年国勢調査の比較 [5]

	生活保護世帯	(誤差)	就業者	(誤差)
指標値	被保護者調査		国勢調査	
	1,636,959		65,468,436	
生活と支え合いに関する調査				
単純無作為抽出	1,245,193	89,975	61,120,259	409,030
層化集落抽出	1,253,475	90,492	61,037,737	395,173
復元推計	1,726,385	210,428	66,015,414	1,358,856

注: 生活保護世帯は世帯数、就業者数は個人数。数値は小数点以下を四捨五入した。令和 2 年国勢調査の就業者数については、不詳補完結果の総数を参照したため、15-19 歳が含まれる。このうち 15-17 歳の人数が不明であるため、総数のままとした。被保護者調査の数値は、令和 2 年度確定値の月次平均を用いた。生活と支え合いに関する調査の推計値については、いずれも項目無回答に対して多重代入を行っている。単位無回答・カバレッジエラーの補正は、「復元推計」の行のみ行っている。「単純無作為抽出」の行は、本来層化集落抽出で得られた標本に、単純無作為抽出を当てはめて、推計値と分散を求めた数値を示している。

各推計値の 95%信頼区間は、推計値±誤差×1.96 で求まる。信頼区間で見ると、単純無作為抽出を仮定した推計値の 95%信頼区間は、生活保護世帯数、就業者数のいずれでも、指標値を含んでいない。復元推計の場合、生活保護世帯数、就業者数のいずれでも、国勢調査の集計値に近付いており、指標値は 95%信頼区間に含まれている [14]。生活保護世帯・就業者数が指標値に接近したということは、無回答補正やキャリブレーションが、要支援層や若年層の把握に貢献しうることを示唆している。このことは、生活上の諸困難や、幅広い世代の社会経済状況を扱う本調査の目的をより良く達成することに資する。無回答補正やキャリブレーションが推計値に与える影響に比べると、単純無作為抽出による近似と層化集落抽出法に基づいた推計の相違はわずかである。すなわち、推計値のバイアスの主要因は、単位無回答にあることが示唆されている。

この復元推計では、住宅所有形態をキャリブレーション変数に用いている。支え合い調査の回答者と国勢調査の住宅所有形態(総務省統計局, 2022, 人口等基本集計　表 18-1)の乖離は大きい。令和 2 年国勢調査における持ち家世帯数は、33,729,416 世帯であった。不詳は 56 世帯に止まるため、大きな影響はない。これに対して、支え合い調査に基づき、単純無作為抽出を仮定した推計値は 42,819,112 世帯(標準誤差 256,130 世帯)、層化集落抽出を前提とした推計値は 42,889,972 世帯(標準誤差 449,710 世帯)であった。これらの推計値を国勢調査と比較すると、持ち家世帯の過大推計になっている。そこで、注 8 の要領で補完した住宅所有形態変数をキャリブレーションに用いた場合、元の変数の復元推計値は 33,960,505 世帯となった。元々住宅所有形態の項目無回答が少なかったこともあり、キャリブレーションの合計値として参照する国勢調査の値に近いのは当然であるが、この項目に関しては国勢調査との乖離を縮めることができたことになる。

[14] 被保護者調査の場合、一般世帯という限定がないので、厳密な比較ではなく若干のずれがある。

しかし、以上のことは、全ての側面で推計値が改善したことを意味しない。特に注意を要するのは最終学歴である。支え合い調査における大学・大学院卒業者の復元推計値は 32,475,646 人(標準誤差 993,639 人)であった。これは単純無作為抽出を仮定した 28,709,492 人(標準誤差 379,726 人)、層化集落抽出に基づく 28,476,867 人(標準誤差 426,550 人)より大きい。国勢調査の最終卒業学校の集計表では、大学卒業者 19,839,068 人、大学院卒業者 2,060,874 人、卒業学校不詳者 15,059,305 人、在学か否か不詳の者 2,551 人であった。不詳の規模が大きく、単純な比較は難しいが、以上の手続きによる復元推計値が過大推計である可能性も大いにある。単純無作為抽出や層化集落抽出による推計値と比較して、復元推計により大学卒業者の推計値が増加したのは、無回答補正やキャリブレーションにおいて、大都市居住者の低回収率に対処してウェイトを大きくしたことによると考えられる。最終学歴による無回答ウェイトを用いることで改善を図ることができるかどうかについては今後の課題とする。

8. 結論

以上の手続きを経て、世帯および個人を調査対象とする支え合い調査のウェイトを設定することが可能となった。これにより、無回答バイアス、すなわち回答者と無回答者の違いによる推計値の偏りを、完全ではないにせよ補正しつつ、母集団の推定値を求められるようになった。昨今、社会調査全般において、回収率が低下しており、無回答バイアスへの対応は喫緊の課題である。回収率向上のための対策はもちろんであるが、集計において無回答に対処する方法の重要性も増してきていると考えられる。回収率が低い集団、例えば大都市部居住者や若年層などの層のデータに、より大きな重みを与えることで、推計の偏りを低減させられる可能性があり、実際 2022 年調査データの復元推計では、一部項目の比較ではあるが、バイアスが緩和された。このことは、無回答補正・キャリブレーションを伴う復元推計を導入すれば必ずバイアスが低減することを意味するものではないが、無回答バイアス等を放置することによる問題ももはや看過できず、適切なウェイトのあり方の検討が求められている。回収率の向上と推計方法の精緻化を両輪として、統計の精度確保をより一層図っていく必要がある。

引用文献

毛塚和宏・三輪哲・榊原賢二郎(2024)「2022 年生活と支え合いに関する調査世帯票のウェイトについての考え方と方法」IPSS Discussion Paper Series 2024-J01、2026 年 3 月 4 日取得、https://www.ipss.go.jp/publication/j/DP/dp2024_J01.pdf。

厚生統計協会、2004、『よくわかる標本調査法――厚生統計で学ぶ標本設計の理論と実践』厚生統計協会。

厚生労働省(2025)「お住まいの地域の級地を確認」2026 年 1 月 26 日取得、https://www.mhlw.go.jp/content/kyuchi.3010.pdf

厚生労働省社会・援護局保護課(2022)「令和 2 年度被保護者調査　月次調査(確定値)結果の概要」2026 年 2 月 26 日取得、https://www.e-stat.go.jp/stat-search/file-download?statInfId=000032172574&fileKind=2。

総務省統計局、2022、「令和 2 年国勢調査」2025 年 5 月 2 日取得、https://www.e-stat.go.jp/stat-search/files?page=1&toukei=00200521&tstat=000001136464。

土屋隆裕(2005)「調査不能者の特性に関する一考察 －「日本人の国民性 第 11 次全国調査」への協力理由に関する事後調査から－」『統計数理』53(1): 35-56。

D'Arrigo, Julia and Skinner, Chris J. 2010. Linearization Variance Estimation for Generalized Raking Estimators in the Presence of Nonresponse. *Survey Methodology*, 36(2): pp. 181-192.

Deville, Jean-Claude and Särndal, Carl-Erik. 1992. Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association* 87(418): 376-382.

Deville, Jean-Claude and Särndal, Carl-Erik, and Sautory, Olivier. 1993. *Journal of the American Statistical Association* 88(423): 1013-1020.

Estevao, Victor M. and Särndal, Carl-Erik. 2006. Survey Estimates by Calibration on Complex Auxiliary Information. *International Statistical Review* 74(2): 127-147.

Eurostat. 2023. Methodological Guidelines and Description of EU-SILC Target Variables: 2023 Operation. Retrieved May 19, 2025 from https://circabc.europa.eu/ui/group/853b48e6-a00f-4d22-87db-c40bafd0161d/library/334d943f-6f71-4f4b-9c7e-a6767a3fe164?p=1&n=-1&sort=name_DESC .

Haziza, David and Jean-François Beaumont. 2017. Construction of Weights in Surveys: A Review. *Statistical Science* 32(2): 206-226.

Lemaître, George and Dufour, Johane. 1987. An Integrated Method for Weighting Persons and Families. *Survey Methodology* 13(2): 199-207.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Särndal, Carl-Erik and Lundström, Sixten. 2005. *Estimation in Surveys with Nonresponse*. John Wiley and Sons.

Valliant, Richard and Dever, Jill A. 2018. *Survey Weights: A Step-by-Step Guide to Calculation*. Stata Press.

Van Buuren, Stef. 2018. *Flexible Imputation of Missing Data*. CRC Press.

Wolter, Kirk M. 2007 *Introduction to Variance Estimation*. 2nd ed. Springer.