IPSS Discussion Paper Series

(No.2024-J01)

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

毛塚和宏（国立社会保障・人口問題研究所）

2024 年 4 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

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[1]

毛塚和宏（国立社会保障・人口問題研究所）

1. イントロダクション

標本調査は母集団から標本を抽出し，回答を得るプロセスである。通常は推測統計学の技法を使って母集団の特性（e.g. 母平均）を推定する。しかし，調査のプロセスの中で，外的妥当性を失い，適切な推定ができない場合がある。たとえば，抽出確率が不均一であったり，無回答者が特定の属性に集中したり，結果的に得られた標本が一部の層で十分に確保できていなかったりする，といったことによって，サンプルの一般性に懸念が生じることがある。この場合，単純な平均値はバイアスしている可能性がある。

典型的なバイアスの一つは，無回答バイアスである。標本調査における近年の課題は回収率の低下であろう（玉野 2003; 轟ら 2021: 212）。無回答がもし完全にランダムに発生している場合（Missing Completely At Random: MCAR）は，バイアスをもたらさない。多くの標本調査データに基づく分析は，MCAR であるという強い仮定を（暗に）置き，行われている。しかしたいていの場合，そのような仮定が満たされることはめったになく，無回答によるバイアスが生じている可能性が高い（Groves et al. 2004=2011: 196-206）。

このようなバイアスを軽減するために，サンプルに対してウェイトをかけることがある。ウェイトによってあたかも母集団全体を扱っているかのように計算をすることができるからである。実際，Wang et al. (2015)は，Xbox ユーザーというバイアスが著しいサンプルから，ウェイトによって大統領選の予測を適切に行うことができた。標本調査の回収率低下が課題となるなか，ウェイトに対して理解を深めることは重要である。

本論文の目的は，国立社会保障・人口問題研究所で実施される基本調査の一つである，2022 年生活と支え合いに関する調査（以下，支え合い調査と呼ぶ）の世帯票を対象に，適切なウェイトを計算することである。ウェイトの考え方を整理し，実際に計算を行うことで，今後の分析や後続の調査に知見を活用することができる。

---

[1] 本論文で行った分析は，「生活と支え合いに関する調査」に関連する業務の一環として行われた。

2. ウェイト（復元倍率，重み）とその役割

ウェイトとは，標本の観測値に乗じる乗数のことであり，復元倍率や重み，重み付けともよばれている。ウェイトの役割は母集団の推計のプロセスにおいて，サンプルの特性に応じて倍率をかけることで標本のサンプルサイズを母集団サイズに復元[2]し，推計の代表性を担保することである[3]。公的統計の文脈では「復元倍率」が，社会調査の文脈では「重み（付け）」や「ウェイト」と呼ぶことが多い。本論文では，母集団の推計に用いる乗数を以後，ウェイトと呼ぶことにする。

ウェイトによって母集団を復元するメカニズムは，標本を重み付けによって「複製する」[4]ことで，仮想母集団を作成することにある。平均値を例にとって考えよう。いま $x_1, x_2, \ldots, x_n$からなるデータがあるとする。この時，通常の標本平均 $\bar{x}$は次のように表現される。

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_N}{n}.$$

その一方で，ウェイトを用いた標本平均（加重平均） $\bar{x}_w$は次のようにあらわされる。ここに，w_iはi番目のサンプルに対するウェイトである（式(1)参照）。

$$\bar{x}_w = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i x_i = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_n}. \quad (1)$$

加重平均の分子をみると，各サンプルのデータにw_iがかけられていることがわかる。これは，サンプルに一人しかいなかったi番目の人を，w_i人に増やしている，ということを意味する。同時に分母を見ると，ウェイトの総和になっている。よく定義されたウェイトは，ウェイトの総和は母集団の人数Nに一致する[5]。つまり以下が成立する。

---

[2] 社会調査の文脈では「復元」とは，復元抽出（一度選ばれたサンプルを再度母集団に戻して抽出を行う）ことと関連付けられるが，本論文での復元は異なることに注意されたい。

[3] 計量経済学では因果推論のために重みづけが行われることもあるが，本論文で扱う復元倍率と目的が異なる。詳しくは Solon et al.（2015）を参照せよ。

[4] 場合によっては，サンプルとして得られすぎた属性を割り引いて推定する機能を持つこともある。これは議論のセクションで再び触れる。

[5] 厳密には，「サンプルから推定される母集団サイズ」である。なお，後述する支え合い調査におけるベースラインウェイト 1 において，「得られた標本から母集団において調査対象となる世帯数を推定したときの値に一致する。」という説明（p. 5）はこの厳密な意味でウェイトが適切に設定されていることを示している。

$$w_1 + w_2 + \cdots + w_n = N.$$

すなわち，加重平均の分母を書き換えると次のようになる。

$$\bar{x}_w = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{N}.$$

この式から，ウェイトが標本を複製することで，母集団レベルの推計に引き上げていることがわかる。

では，ウェイトはどのように決定すべきなのだろうか。ウェイトは以下の3つの要素に分解される（Biemer and Christ 2012）。具体的には，1）抽出確率への対応（ベースウェイトw^B），2）無回答に対する対応（無回答ウェイトw^{NR}），3）不足・過剰への対応（事後層化ウェイトw^{PS}），である。そして，最終ウェイトw^Fはこの3つのウェイトの積によって与えられる（式(2)参照）。

$$w^F = w^B \times w^{NR} \times w^{PS}. \tag{2}$$

本論文では Biemer and Christ（2012）に依拠しつつ，この 3 つの要素について，それぞれ決定方法を概説しながら，生活と支え合い調査に適用し，ウェイトを計算していく。

3. 支え合い調査の調査プロセスとウェイト計算のデータソース

3.1 支え合い調査の調査プロセス

本論文でウェイトを計算する対象である，2022 年生活と支え合いに関する調査について，調査設計に着目して説明をする。なお，調査項目等の詳細は報告書（国立社会保障・人口問題研究所 2023）や雑誌『厚生の指標』に掲載された記事（黒田ほか 2024）を参照してほしい。

2022 年生活と支え合いに関する調査は集落抽出法を採用している。厚生労働省が実施している令和 4 年国民生活基礎調査が用いた 5,530 地区[6]から無作為に 300 地区を抽出し，その地区内のすべての世帯の世帯主，および 18 歳以上の世帯員が本調査の対象となる。ここに「地区」は令和 2 年国勢調査の調査区から選ばれている。

実査では調査員が訪問し質問紙を配布する。調査票等配布時に 3 回訪問してなお不在等の場合には，調査票等を投函することで対応している（国立社会保障・人口問題研究所 2022: 277）。以下本論文では，この不在時の投函対応を「ポスティング」と呼ぶことにする。

[6] 国民生活基礎調査の調査地区は，令和 2 年国勢調査の調査区から層化無作為抽出を行って得られている。

3.2 計算に用いるデータソース

ウェイトの計算に際して，2 つのデータソースを用いる。一つは支え合い調査の実査の過程で得られた情報を集約したデータである（以下，このデータソースを実査集約データと呼ぶ）。実査集約データには都道府県，調査区，回答したか否か，調査実施時の世帯名簿に記載されている世帯人員数，調査票をポスティングしたか否かといった情報が含まれている。

もう一つは令和 2 年度国勢調査である。ベースウェイトでは小地域集計「第 1 表　男女別人口及び世帯数－基本単位区」（全都道府県分）を，事後層化ウェイトでは人口等基本集計「第8－1表 世帯員の年齢による世帯の種類，世帯人員の人数別一般世帯数－全国，都道府県，市区町村」を用いる。

4. 抽出確率への対応：ベースウェイト

4.1 ベースウェイトの考え方

無作為抽出の原則は「母集団に属するすべての成員が等しい確率で抽出されること」である（盛山 2004）。この性質は統計学上の様々な推測の妥当性（例：不偏性）を担保するためには重要である。その一方で，標本を母集団レベルに復元するには「抽出されなかった世帯」のことを重みによって「複製する」ことを考えなければならない。これを実現するためには，調査対象世帯が選ばれた確率の逆数をウェイトとしてかけることになる。これをベースウェイト（baseline weight）とよぶ。以後，右上の添え字にBを入れることでベースウェイトw^Bであることを表現する。

標本調査では，母集団と標本以外に，（標本）抽出枠（frame population）と呼ばれる集団が存在する。枠集団は標本抽出する際の分母となる集団である。よって，ベースウェイトは標本を標本抽出枠に拡大するウェイトである，と理解することができる。具体的に数式を用いて説明する。まず，単純無作為抽出の場合を説明し，その後に多段抽出の場合を説明する。その後に，支え合い調査の場合を説明する。

4.1.1 単純無作為抽出の場合

N世帯の母集団に対して，n世帯分の標本を無作為抽出する際，母集団内の世帯kが選ばれる確率p_k^Bはn/Nとしてあらわされる。よって，この場合のウェイトw_kはその逆数で与えられる。

$$w_k^B = \frac{1}{p_k^B} = \frac{N}{n}.$$

4.1.2 多段抽出の場合

標本抽出には，直接世帯を抽出するのではなく，調査対象とする地域を先に抽出し，その地域からさらに世帯を抽出する，といった数段階の抽出プロセスを経る方法がある。これを多段抽出法と呼ぶ[7]。たとえば，まず都道府県内iの調査区jを抽出，その後に世帯kを抽出することを想定する。いま，都道府県iの全調査区数をM_i，抽出した調査区数をm_i，調査区jの総世帯数をN_{ij}，抽出世帯数をn_{ij}とする。このとき，都道府県i，調査区jに住む世帯kが抽出される確率p^B_{ijk}は次で表現される。

$$p^B_{ijk} = \frac{m_i}{M_i} \times \frac{n_{ij}}{N_{ij}}.$$

よって，この逆数を取るとウェイトw_{ijk}を求めることができる。

$$w^B_{ijk} = \frac{1}{p^B_{ijk}} = \frac{M_i}{m_i} \times \frac{N_{ij}}{n_{ij}}. \tag{3}$$

4.2 支え合い調査の場合

支え合い調査は，国民生活基礎調査での調査地区からランダムサンプリングを行って，300 地区を抽出し，選ばれた地区のすべての世帯を調査対象として設定している。ここで，標本抽出枠をどのように設定するかによって，2 通りのベースウェイトを計算することができる。なお，以下都道府県を層として，都道府県i，調査地区j，世帯kのサンプルに対するウェイトw^B_{ijk}を考える。

4.2.1 ベースウェイト1：抽出確率に基づく場合

一つ目の考え方は，抽出確率に忠実にベースウェイトを計算する方法である。国民生活基礎調査は国勢調査における後置番号[8]1と8から層化無作為抽出にサンプルを抽出している。そこで，ベースウェイトの計算にあたり，国勢調査における後置番号1，8の調査地区を抽出枠とし，都道府県を層として，ここから無作為に 300 地区が抽出された，とみなして計算を行う[9]。

---

[7] 詳しくは盛山（2004）を参照せよ。

[8] 後置番号とは調査地区の特性を整理するために付与された番号である。詳細は林（2017）を参照せよ。

[9] 本来は，国民生活基礎調査の調査区抽出プロセス，そしてその調査区から 300 地区を抽出するプロセスを加味すべきであり，この仮定は強い仮定である。本論文では，明快さと実装可能性を考

支え合い調査において世帯が抽出される確率は，ある世帯が属する地区が抽出される確率に等しい。すなわち，式(3)において，$n_{ij} = N_{ij}$とすれば，抽出確率に基づくベースウェイトw_{ijk}^{B1}が得られる。

$$w_{ijk}^{B1} = \frac{1}{p_{ijk}} = \frac{M_i}{m_i}.$$

このとき，ウェイトの総和は，調査対象となった地区の平均世帯数($\sum_{j=!}^{m_i} N_{ij}/m_i$)を復元した世帯数に一致する。

$$\sum_{i=1}^{47}\sum_{j=1}^{m_i}\sum_{k=1}^{N_{ij}} w_{ijk}^{B1} = \sum_{i=1}^{47}\sum_{j=1}^{m_i}\sum_{k=1}^{N_{ij}} \frac{M_i}{m_i}$$
$$= \sum_{i=1}^{47}\frac{M_i}{m_i}\sum_{j=1}^{m_i} N_{ij} = \sum_{i=1}^{47} M_i\left(\frac{1}{m_i}\sum_{j=1}^{m_i} N_{ij}\right).$$

これは，得られた標本から枠集団の世帯数を推定したときの値に一致する。

4.2.2 ベースウェイト2：世帯数に基づく場合

2 つ目の考え方は世帯数に基づくウェイトである。都道府県i内の総世帯数$N_i(=\sum_{j=1}^{M_i} N_{ij})$から，標本として$n_i(=\sum_{j=1}^{m_i} N_{ij})$世帯だけ無作為抽出された，とみなしてベースウェイトを設定することもできる。この場合のベースウェイトw_{ijk}^{B2}は以下のように定義される。

$$w_{ijk}^{B2} = \frac{N_i}{n_i}.$$

このとき，ウェイトの総和は，調査対象となった総世帯数に一致する。

$$\sum_{i=1}^{47}\sum_{j=1}^{m_i}\sum_{k=1}^{N_{ij}} w_{ijk}^{B2} = \sum_{i=1}^{47}\sum_{j=1}^{m_i}\sum_{k=1}^{N_{ij}} \frac{N_i}{n_i}$$

慮してこのような強い仮定を採用している。

$$= \sum_{i=1}^{47} \frac{N_i}{n_i} \underbrace{\left(\sum_{j=1}^{m_i} N_{ij} \right)}_{=n_i} = \sum_{i=1}^{47} N_i.$$

4.3 ベースウェイトの計算結果

計算の結果として得られたウェイトの概要を表 1 に，ヒストグラムを図 1,2 に示した。全体的にベースウェイト1のほうが2より少ないことがわかる。これは，ウェイトの分子にくるものが，地区数と世帯数と異なる上に，世帯数のほうが明らかに多いためである。

表 1：ベースウェイトの基礎統計（47 都道府県）

	平均	中央値	標準偏差	総和
ベースウェイト 1	506.65	340.33	374.25	4292868.60
ベースウェイト 2	3379.00	3292.94	610.59	28630235.09

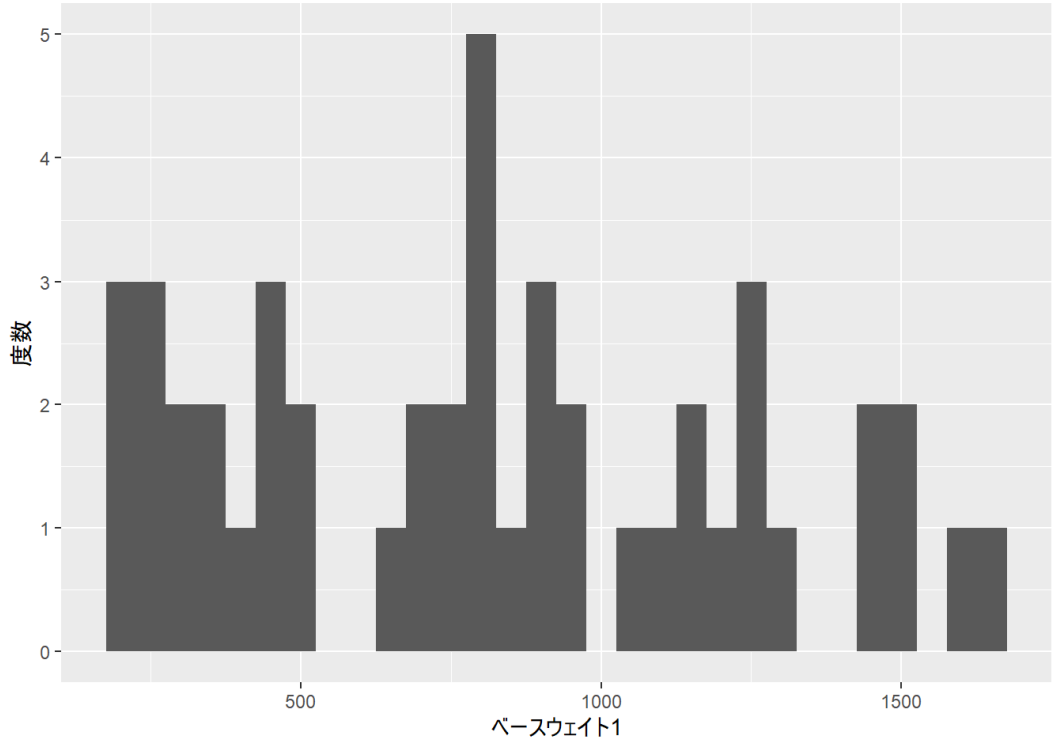

図 1：ベースウェイト1（抽出確率に基づくウェイト）のヒストグラム

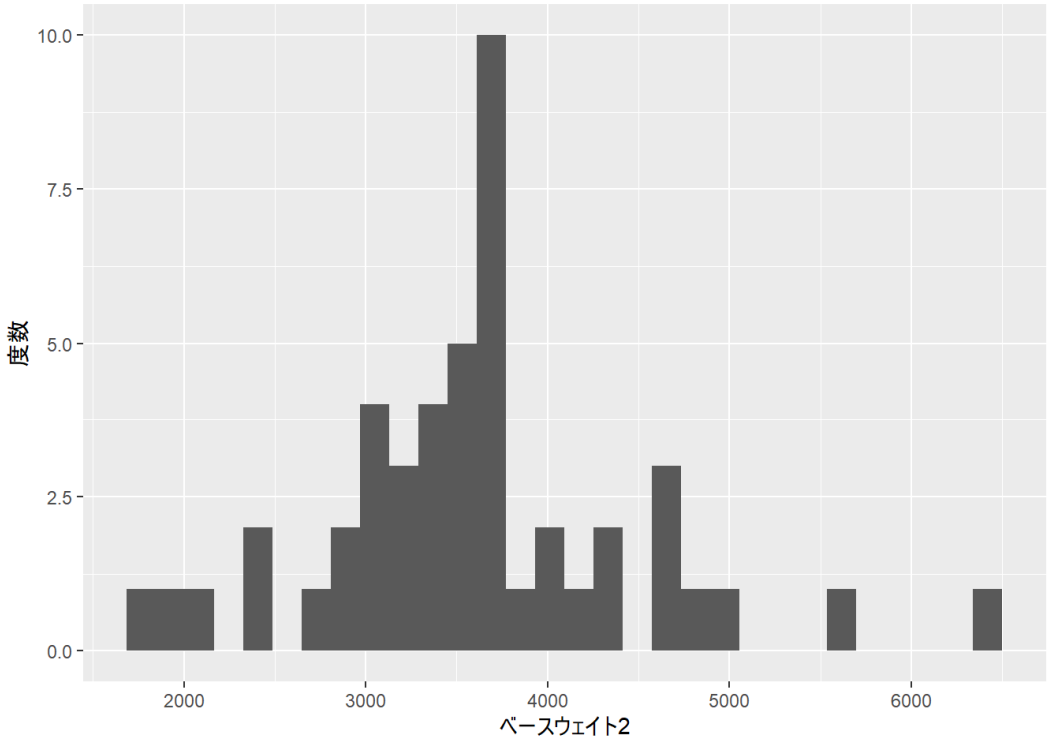

図 2：ベースウェイト2（世帯数に基づくウェイト）のヒストグラム

5. 無回答への対応：無回答ウェイト

5.1 無回答ウェイトの考え方

標本調査では抽出後，実際に世帯に尋ねて回答を得る。しかし，当然ながら様々な形で回答が得られないことがある。調査期間中の訪問した時すべて不在だった，回答を拒否された，調査票を渡したが返送がなかった，などである。この場合，ウェイトの発想としては「無回答になりやすいサンプルほど，手厚く複製する」という考え方に立つ。たとえば，単身世帯の無回答傾向が高い場合は，その少ない分を高いウェイトで補う，ということである。本論文では，この無回答に対するウェイトを無回答ウェイトと呼び，右上の添え字にNRを付け，w^{NR}と表す。世帯kの無回答ウェイトw_k^{NR}は，世帯kに対する推定された回答確率 $\widehat{p_k^{NR}}$の逆数のよって得られる。

$$w_k^{NR} = \frac{1}{\widehat{p_k^{NR}}}.$$

本論文では二項ロジスティック回帰分析を用いて回答確率の推定をする。すなわち，回答ダミーをR，説明変数をXとすると，

$$\widehat{p_k^{NR}} = \Pr(\widehat{R=1}|X_k) = \frac{1}{1+\exp\left(-(\beta_0+\beta X_k)\right)}$$

という形で表現できる。回答確率p^{NR}の推定には，回答・無回答両サンプルともに測定されている変数を用いなければならない。回答するか否かに関連する変数を選んで投入する必要がある。

5.2 支え合い調査の場合

支え合い調査の世帯票の回収率は 50.7%（配布調査票 16,719 票　有効票数 8,473 票）であった。この回収率はさまざまな要因によってちらばっている。まず実査集約データを用いてその異質性を確認する。地区別の回収率をヒストグラムに図示した図 3 を見ると，地区による回収率の散らばりも大きいことを示している。また，表 2 に地域ブロック，調査実施時の世帯人員数，ポスティングによる回収率の違いを示した。地域ブロックからは，地域ごとに回収率の異質性が見られる。また，全体的な傾向として，この両者は回収率と負の相関をもつ傾向にある。すなわち，世帯人員数が多いほど，ポスティングがされているほど，回収率が低い傾向にある。これらの散らばりを踏まえて，回答確率を推定しなければならない。

そこで支え合い調査における回答確率の推定では次の変数を用いて，ロジスティック回帰分析によって回答確率p^{NR}を推定する：ポスティングか否か，調査実施時の世帯人員数，調査区ダミーである。これらの変数を用いたのは，次の 2 つの条件を満たしているためである：1) 回答・無回答世帯双方で得られている，2) 先に確認した通り，回答（＝回収）と関連しうる。

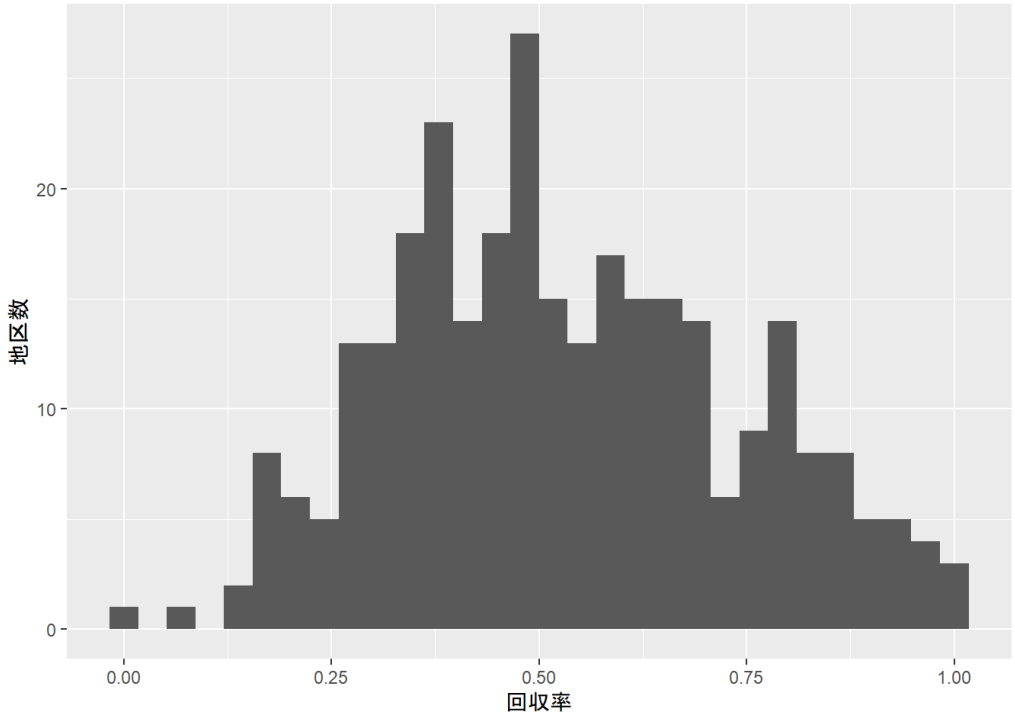

図 3：地区別の回収率のヒストグラム

表 2：地域ブロック，世帯人員数（実施時），ポスティングの有無と回収率

		回収率
地域ブロック	北海道・東北	0.588
	北関東	0.557
	東京圏	0.468
	中部・北陸	0.631
	中京圏	0.481
	大阪圏	0.463
	京阪周辺	0.548
	中国	0.544
	四国	0.483
	九州・沖縄	0.504
世帯人員数（実施時）	1 人	0.477
	2 人	0.662
	3 人	0.611
	4 人	0.552
	5 人	0.518
	6 人	0.548
	7 人以上	0.379
	不明	0.198
ポスティング	あり	0.270
	なし	0.567

表 3 にロジスティック回帰分析の結果を示す。表 3 から，回答率に関するいくつかの示唆を得ることができる。まず，3 回の訪問にもかかわらずポスティングに至った世帯は回答を得にくい傾向にある（オッズ比 0.4）。次に，単身世帯に比べて，2 人・3 人は回答を得やすく，7人以上・不明の場合は回答を得づらい。

このモデルをもとに得られた無回答ウェイトの基礎統計を表 4，ヒストグラムを図 4 に示す。無回答ウェイトは必ず 1 を超えるので[10]，平均値も 1.98 と 1 を超えている。無回答ウェイトが大きいほど，回答率が低い属性であることを示唆している。

表 3：ロジスティック回帰分析の結果

	係数		Exp(b)	標準誤差
切片	0.049		1.050	0.339
ポスティングダミー	-0.824	***	0.439	0.058
世帯人員数：実施時（ref. 1人）				
2人	0.506	***	1.659	0.051
3人	0.274	***	1.316	0.058
4人	0.079		1.082	0.066
5人	-0.157		0.855	0.097
6人	-0.115		0.891	0.180
7人以上	-0.877	**	0.416	0.280
不明	-1.157	***	0.314	0.070

$**: p < 0.01$, $***: p < 0.001$

注：調査区ダミーの結果は省略する

表 4：無回答ウェイトの基礎統計（回答した 8473 サンプルに基づく）

	平均	中央値	標準偏差	総和
無回答ウェイト	1.98	1.57	1.35	16806.99

[10] 回答確率p_k^{NR}は，$0 \leq p_k^{NR} \leq 1$なので，その逆数を取ると，$1/p_k^{NR} \geq 1$となるためである。

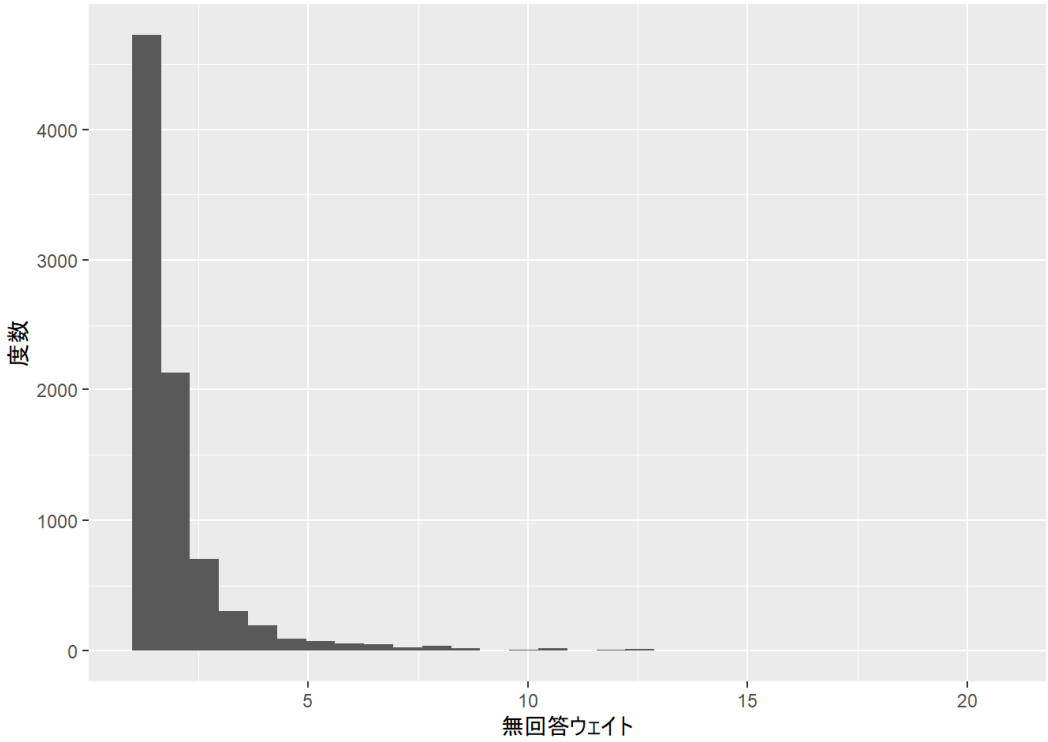

図 4：無回答ウェイトのヒストグラム（回答した 8473 サンプルに基づく）

6. 不足・過剰を補正する：事後層化ウェイト

6.1 事後層化ウェイトの考え方

最後は事後層化ウェイトである。事後層化ウェイトは，ノンカバレッジウェイトとも呼ばれている。事後層化ウェイトを用いることで，特定の属性が不足している場合などに対処することができる。本論文では，事後層化ウェイトを右上にPSの添え字をつけてw^{PS}と表す。

母集団における世帯総数Nを複数の層$h = 1, 2, \ldots, H$に分割し，各層の総世帯数をN_hとする。このとき，層hに属する世帯kに対する事後層化ウェイトw_k^{PS}は次のように計算される。

$$w_k^{PS} = \frac{N_h}{\sum_{t=1}^{n_{rh}} w_t^B w_t^{NR}}.$$

ここに，n_{rh}は層hに属する有効回答世帯数である。このように，事後層化ウェイトは，有効回答サンプルの情報のみで設定可能であることがわかる。また，層hに属する標本における総ウェイトの総和は層hに総世帯数N_hに一致する。

$$\sum_{k=1}^{n_{rh}} w_k^B w_k^{NR} w_k^{PS} = N_h.$$

層の設定を適切に行うことで，特定の属性を持つ標本の不足をカバーすることができる。ある層の事後層化ウェイトが 1 より大きいということは，2 つのウェイトで復元した仮想母集団の総数よりも母集団の総数のほうが大きい，すなわち不足しているのでウェイトによってさらに増やしていることを表す。逆に，1 より小さいということは，母集団の総数のほうが小さいので割り引いていることを示している。

6.2 支え合い調査の場合

本論文では，都道府県ごとの世帯人員を層として設定したい。具体的には，令和 2 年の国勢調査における都道府県別，一般世帯[11]における世帯人員（1 人，2 人，…，6 人，7 人以上）の 47×7＝329 層を用いる[12]。事後層化ウェイトの分母に用いるベースウェイトは 2 種類想定しているので，事後層化ウェイトも 2 種類計算を行い，ベースウェイト1に基づく方を事後層化ウェイト 1，ベースウェイト2を事後層化ウェイト 2 と呼ぶ。

表 5：事後層化ウェイトの基礎統計（回答した 8473 サンプルに基づく）

	平均	中央値	標準偏差	総和
事後層化ウェイト 1	9.73	9.62	5.44	82458.78
事後層化ウェイト 2	0.99	0.94	0.32	8397.00

表 5 に事後層化ウェイトの基礎統計，図 5，6 にヒストグラムを示した。ベースウェイト1に基づく事後層化ウェイトは全体的に 1 より大きく，ベースウェイト 2 に基づくウェイトは 1 より小さいことが示されている。これは，基準が地区か世帯数かの点が影響していると考えられる。

[11] 一般世帯に関するデータを用いるということは，復元する母集団として，国勢調査の一般世帯を想定することを意味する。この部分を変えることで，復元する母集団を指定することができる。
[12] なお，今回用いた世帯票において，世帯人員数が不明なサンプルは存在せず，すべての世帯が必ずいずれかの層に属することとなった。

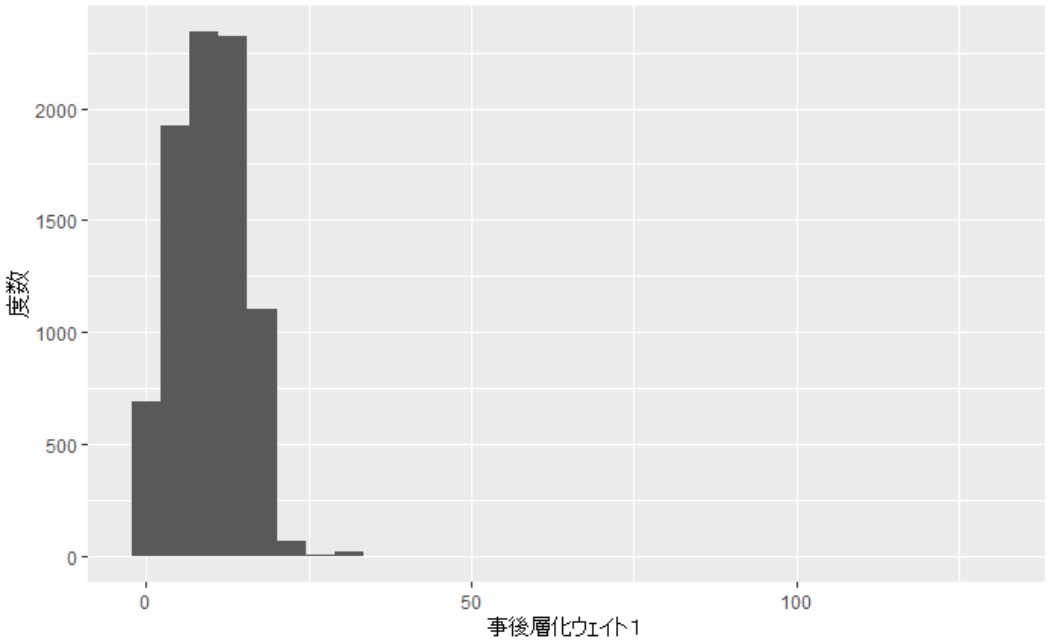

図 5:事後層化ウェイト 1 のヒストグラム(回答した 8473 サンプルに基づく)

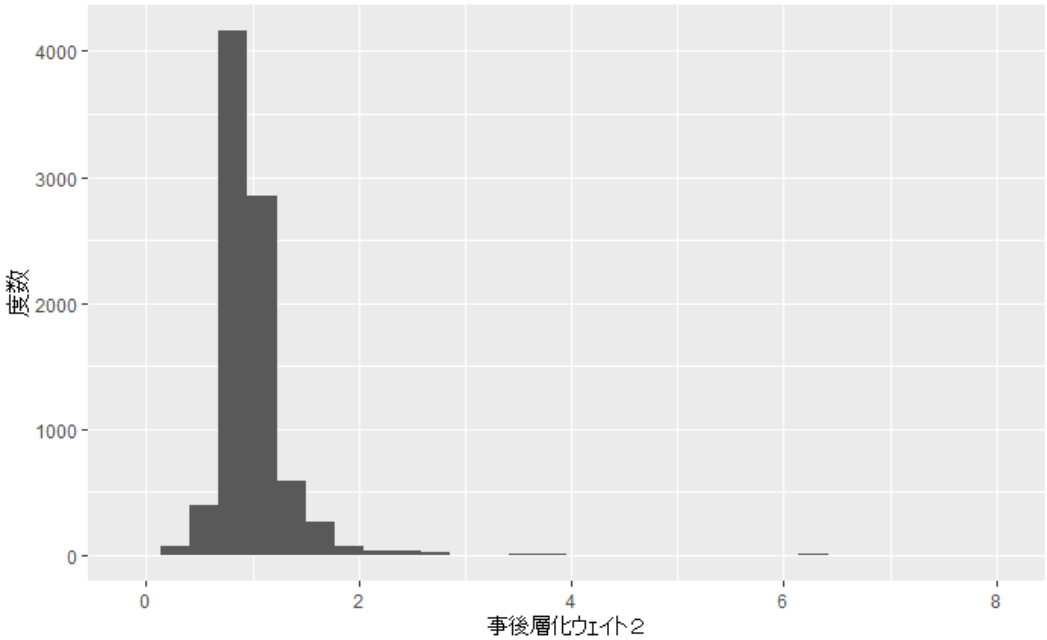

図 6:事後層化ウェイト 2 のヒストグラム(回答した 8473 サンプルに基づく)

7.最終ウェイトと各ウェイトへの評価

7.1 支え合い調査の最終ウェイト

第 2 節で示したようにベースウェイト，無回答ウェイト，事後層化ウェイトのすべての積（式(2)）によって最終的なウェイトが与えられる。本論文ではこれを最終ウェイトと呼ぶことにする。今回の分析では，どちらのベースウェイトを用いても同じ最終ウェイトw_kを与える（証明と含意については補遺を参照せよ）。

$$w_k = w_k^{B1} \times w_k^{NR} \times w_k^{PS1} (= w_k^{B2} \times w_k^{NR} \times w_k^{PS2})$$

表 6：最終ウェイトの基礎統計（回答した 8473 サンプルに基づく）

	平均	中央値	標準偏差	総和
最終ウェイト	6532.63	5102.30	4816.70	55350944

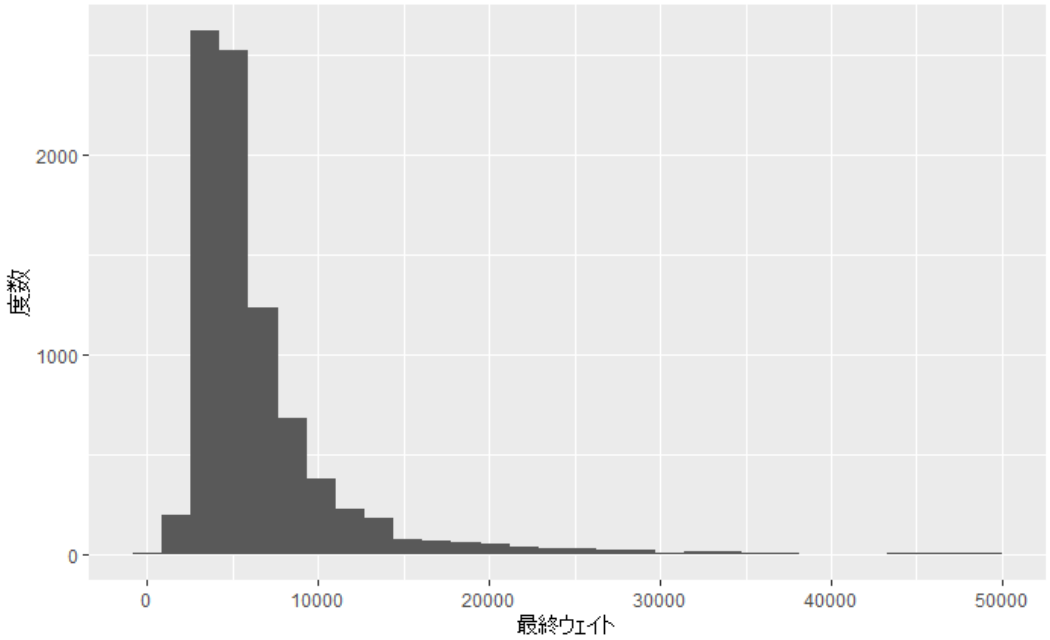

図 7：最終ウェイトのヒストグラム（回答した 8473 サンプルに基づく）

表 6 にウェイトの基礎統計，図 7 にヒストグラムを示した。表 5 から平均的には，サンプル 1 世帯が約 6532 世帯を代表させていることがわかる。最終的なウェイトの総和は令和 2 年度の国勢調査の一般世帯の総数 55,704,949 世帯には届いていない。これは，サンプルが事後層化ウェイトに用

いた 329 層のすべてを網羅していないので，その分が少なくなっているためである[13]。

7.2 平均世帯人員数の評価

では，このウェイトを用いて世帯人員数の平均値，標準偏差，95%信頼区間を計算する。なお，加重平均$\bar{x}_w$は式(1)を，標準偏差σ_wと 95%信頼区間$95\%CI_w$は以下の式を用いた[14]。ここにnはサンプルサイズを表す。

$$\sigma_w = \sqrt{\frac{1}{\left(\sum_{k=1}^{n} w_k\right) - 1} \sum_{k=1}^{n} w_k (x_k - \bar{x}_w)^2},$$

$$95\%CI_w: \left(\bar{x}_w - 1.96\frac{\sigma_w}{\sqrt{n}}, \bar{x}_w + 1.96\frac{\sigma_w}{\sqrt{n}}\right).$$

表 7:各ウェイトに基づく平均世帯人員数

			95%信頼区間	
ウェイト	平均	標準偏差	下限	上限
なし	2.343	1.205	2.317	2.369
ベースウェイト 1 w^{B1}	2.404	1.249	2.378	2.431
ベースウェイト 2 w^{B2}	2.345	1.211	2.319	2.371
無回答ウェイトw^{NR}	2.244	1.242	2.218	2.271
ベースウェイト 1×無回答ウェイト$w^{B1}w^{NR}$	2.331	1.299	2.303	2.358
ベースウェイト 2×無回答ウェイト$w^{B2}w^{NR}$	2.249	1.249	2.222	2.275
最終ウェイト$w^{B}w^{NR}w^{PS}$	2.186	1.245	2.159	2.212
参考：2022 年 国民生活基礎調査の平均人員数	2.25 人			

表 7 に世帯人員数の分析結果を示した。2022 年国民生活基礎調査の平均世帯人員数は 2.25 人であった（厚生労働省 2023）。これを踏まえると，ウェイトを用いない単純平均値 2.343 は上方にバイアスしている可能性がある，ということを示している。

[13] このような状況に対しては，レイキングという手法が存在する。詳細は Valliant and Dever（2018）を参照せよ.

[14] ここに示した標準偏差は最もシンプルなものである。サンプリングも考慮した分散の計算は Horvitz and Thompson（1952）を参照せよ。また，ブートストラップ法による推定も考えられるが，今後の課題としたい。

ウェイトを用いたなかで，ベースウェイト1，2，ベースウェイト1×無回答ウェイトは過大推定，最終ウェイトは過小推定をしている。その中で，ベースウェイト 2×無回答ウェイトが点推定として最も近い値を，次いで無回答ウェイトが近い値を示している。ここから，世帯数に基づくベースウェイトと無回答ウェイトを組み合わせたウェイトを用いることが最適なウェイトとして機能しうることが示唆される。

8. 議論

本論文では，支え合い調査世帯票に対するウェイトを検討した。具体的には 3 つのウェイト（ベースウェイト，無回答ウェイト，事後層化ウェイト）を計算した。平均世帯人員数を用いて複数のウェイトがもたらす補正性能を確認した結果，世帯数に基づくベースウェイトと無回答ウェイトを組み合わせたウェイト（$w^{B2}w^{NR}$）が最も高いことが確認された[15]。

ただし課題もある。本論文では平均世帯人員数で評価を行ったが，ほかの変数への適用可能性など，最終的な判断はさらなる評価を待たねばならない。また，過去の支え合い調査に対して，本論文のスキームがどの程度効果的か，今後検討も必要だろう。

どのウェイトを採用するか，事前に決定することは難しい。たとえば，今回はベースラインウェイト，無回答ウェイト，事後層化ウェイトの組み合わせによって，6 つのウェイトの可能性が存在した。これらの性能は，今回の世帯人員数のように事前に明らかになっている指標との比較によって初めて明らかになる。この意味で，ウェイトの決定は事後的にならざるを得ない部分もあるのは事実であろう。

適切なウェイトは，柔軟な標本設計を可能にする。その一例が，オーバーサンプリングである。たとえば，単身若年世帯の回収率が低いことを見越して，事前に多めにサンプリングしたとする。このとき，単身若年世帯の抽出確率はほかの世帯に比べて相対的に高いことになる。そのように得られた標本に対して，なにもウェイトを施さずに推定を行うとバイアスした推定値になる。よって，抽出確率に応じてベースウェイトを変化させることで，オーバーサンプリングを補正することで対応する。サンプルサイズを確保しつつ，不偏性に配慮した分析を，ウェイトは可能にするのである。

社会調査データを用いて二次分析をするとき，ウェイトはあまり顧みられない。しかし，今後，回収率の観点からみて，社会調査をめぐる環境が改善する見込みは薄い。適切にウェイトを設定することは分析精度を高めうる。ウェイトを用いることは今後の分析において必須となりつつあるのだ。本論文を通して，ウェイトの重要性が顧みられることを願いたい。

---

[15] ただし，ベースウェイト 2 を用いる場合は不偏性を持たない比推定の形になっており，新たなバイアスに対する懸念もある。詳しくは，Goodman and Hartly（1958）を参照せよ。

文献

Biemer, P. P. and S. L. Christ, 2012, “Weighting survey data,” *International handbook of survey methodology,* pp. 317-341, Routledge.

林玲子, 2017,「国勢調査における後置番号別人口」『IPSS Working Paper Series』15:1-5.

Horvitz, D. G. and D. J. Thompson, 1952, “A Generalization of Sampling Without Replacement From a Finite Universe,” *Journal of the American Statistical Association*, 47(260), 663–685.

厚生労働省, 2023,『2022(令和4)年 国民生活基礎調査の概況』厚生労働省,(2024 年 2 月 22 日取得, https://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa22/dl/14.pdf).

国立社会保障・人口問題研究所, 2022,『2022 年社会保障・人口問題基本調査 生活と支え合いに関する調査 報告書』国立社会保障・人口問題研究所,(2024 年 2 月 22 日取得, https://www.ipss.go.jp/ss-seikatsu/j/2022/SSPL2022_houkokusho/SSPL2022_houkokusho.pdf).

黒田有志弥, 毛塚和宏, 河西奈緒, 佐々木織恵, 榊原賢二郎, 盖若琰, 泉田信行, 2024,「「生活と支え合いに関する調査」結果の概要について」『厚生の指標』71(2):30-37.

Goodman, L. A. and H. O. Hartley, 1958, "The Precision of Unbiased Ratio-Type Estimators," *Journal of the American Statistical Association,* 282:491-508.

Groves, R. M. F. J. Fowler, Jr., M. P. Couper, J. M. Lepkowski, E. Singer and R. Tourangeau, 2004, *Survey Methodology*, John Wiley & Sons.(=大隅昇監訳, 氏家豊, 大隅昇, 松本渉, 村田磨理子, 鳰真紀子訳, 2011,『調査法ハンドブック』朝倉書店).

盛山和夫, 2004,『社会調査法入門』有斐閣.

Solon, G., S. J. Haider and J. M. Wooldridge, 2015, “What are we weighting for?” *The Journal of Human Resources*, 50(2): 301-316.

玉野和志, 2003,「サーベイ調査の困難と社会学の課題」『社会学評論』53(4): 537-551.

轟亮, 杉野勇, 平沢和司編, 2021,『入門・社会調査法〔第4版〕――2ステップで基礎から学ぶ』法律文化社.

Valliant, R. and J. A. Dever, 2018, *Survey Weights: A Step-by-Step Guide to Calculation*, College Station, Texas: Stata Press.

Wang, W., D. Rothschild, S. Goel and A. Gelman, 2015, “Forecasting elections with non-representative polls,” *International Journal of Forecasting*, 31(3): 980-991.

補遺 最終的なウェイトが一致することの証明

本論文で作成した最終ウェイトw_kは，ベースウェイトのいずれを用いても一致した。以下，この点について以下に証明を与える。

命題 本論文で示したウェイトの作成の仕方に基づき，世帯kに注目し，ベースウェイト1を用いた場合の最終ウェイトをw_k^1，ベースウェイト2を用いた場合をw_k^2とする。このとき，すべての世帯kについて次が成立する。

$$w_k^1 = w_k^2.$$

証明） ウェイトの決め方から，事後層化ウェイトの同じ層h（同じ都道府県，同じ世帯人員数）に属している世帯k, lについて，この両者のベースウェイトは一致する。すなわち，ある値w^{hB1}, w^{hB2}が存在して，以下が成立する。

$$w_k^{B1} = w_l^{B1} = w^{hB1}, w_k^{B2} = w_l^{B2} = w^{hB2}.$$

これは，ベースウェイトが都道府県別に設定されているためである。

いま，世帯kがある層hに属していると考える。最終ウェイトw_k^1を計算すると，次のようにベースウェイトに依存しない形に変形できる。

$$\begin{aligned} w_k^1 &= w_k^{B1} w_k^{NR} w_k^{PS1} \\ &= \frac{w^{hB1} w_k^{NR} N_h}{\sum_{t=1}^{n_{rh}} w^{hB1} w_t^{NR}} = \frac{w^{hB1} w_k^{NR} N_h}{w^{hB1} \sum_{t=1}^{n_{rh}} w_t^{NR}} \\ &= \frac{w_k^{NR} N_h}{\sum_{t=1}^{n_{rh}} w_t^{NR}}. \end{aligned}$$

同様に，w_k^2についても計算することで，$w_k^1 = w_k^2$が示される。

$$\begin{aligned} w_k^2 &= w_k^{B2} w_k^{NR} w_k^{PS2} \\ &= \frac{w^{hB2} w_k^{NR} N_h}{\sum_{t=1}^{n_{rh}} w^{hB2} w_t^{NR}} = \frac{w^{hB2} w_k^{NR} N_h}{w^{hB2} \sum_{t=1}^{n_{rh}} w_t^{NR}} \\ &= \frac{w_k^{NR} N_h}{\sum_{t=1}^{n_{rh}} w_t^{NR}} = w_k^1. \end{aligned}$$

これがすべての層hについて言え，すべてのサンプルはかならずどこかの層に重複なく属しているので，命題が示された。

■

この命題は，事後層化ウェイトの層（＝都道府県，世帯人員）がベースウェイトで用いた層（＝調査地区）と包含関係にある際に当てはまる。同時に，この命題が成立する場合は，無回答ウェイトと層の規模のみによって最終的なウェイト（$w_k = w_k^{NR} N_h / \sum_{t=1}^{n_{rh}} w_t^{NR}$）を決定することができることを示している。

仮に，調査地区内の全世帯でなく，系統抽出などによって一部を抽出した場合，ベースウェイトは都道府県よりも細かい調査地区ごとに決めることができる。このような場合は包含関係にないので，命題が示すように無回答ウェイトのみでウェイトが決定する事態は生じない。