資　料

（素案）

欠測値の補完に係る主な方法等について

令和3年3月26日

総務省政策統括官（統計基準担当）

目次

はじめに

○公的統計を作成するための統計調査において、一部の調査項目が未回答である場合や、回答そのものが得られずに欠測値が発生する場合があるが、結果の有用性を確保するため、これら欠測値への適切な対応が求められる。

○一方で、これらの欠測値については、統計調査ごとに欠測の発生状況や補完に利用できるデータなどに違いがあり、特定の補完方法の適用など一律の対応は困難な面がある。

○このため、統計調査ごとの状況を踏まえた適切な対応が重要となるが、欠測値に関するこれまでの評価分科会における審議や各府省における取組状況等も踏まえて、補完を行うに当たっての主な方法・手順や利用上の注意点など、実務上参考となる事項を整理してまとめて示すことは、公的統計の精度の確保や向上を促進する上で有意義と考えられる。

1　欠測値の補完に当たっての考え方

（1）欠測による影響

○統計調査においては、調査項目の一部について未記入である場合や調査客体から回答が得られない場合などの欠測が発生した場合において、そのままで集計すると母集団としての代表性が損なわれ、平均値などの結果に偏りが発生するおそれがあり、結果利用上の有用性に影響が生じる。

（2）補完に関する視点

○欠測に対しては、基本的にはデータ収集段階においてできる限り発生しないように対処することが必要であるが、最終的に発生した欠測値に対しては、統計的な処理として、可能な限り補完を行うことにより統計表及び平均値などの結果の有用性を確保する。

1　欠測値の補完に当たっての考え方

（2）補完に関する視点（続き）

○欠測への対応を考える上では、欠測が何に依存して発生しているか（欠測データメカニズム）も考慮して適切な対応をとることが必要である。
単純に欠測のない標本のみを用いて母平均を推定する場合、欠測のメカニズムが完全にランダムな場合以外は結果に偏りが生じる。

※参考：欠測のない標本のみで平均値を推定した場合の偏りについて

項目（y）に欠測が発生しているとして、母集団をyについて回答する層と回答しない層に分けると仮定し、それぞれの層の平均値をμ_c、μ_{ic}、対応する構成割合をπ_c、$1-\pi_c$とすると、母集団全体の平均は、

$$\mu = \pi_c \mu_c + (1-\pi_c)\mu_{ic}$$

と書ける。ここで、欠測のない標本のみから平均値を推定した場合はμ_cとなることから、偏りは、

$$\mu_c - \mu = (1-\pi_c)(\mu_c - \mu_{ic}) \quad (*)$$

となる。欠測の発生が完全にランダムな場合には、$\mu_c = \mu_{ic}$となるため、（＊）の偏りは0となるが、それ以外の場合には0とはならない。

1　欠測値の補完に当たっての考え方

（2）補完に関する視点（続き）

○これに対し、補完を行う場合には、主にランダムな欠測を仮定できるならば、観測されている他の項目（補完の対象となる項目と関連を有し、欠測のし易さとも関連する項目）を適切な補助変数として利用して補完を行うことにより、平均値などの結果の偏りを緩和することができる。

○補完のための方法としては、いくつかの方法が一般的に用いられているが、統計調査ごとに、欠測の発生状況や利用可能な補助変数の状況、実務上の適用可能性なども考慮し、適切な方法を選択することが必要である。

○以下に、欠測値の補完に関して利用される主な方法として、項目の欠測の場合と、未回収などユニット単位での欠測の場合別に示す。

2　補完を行うための主な方法

＜項目の欠測（item nonresponse）への対応＞

（1）（層化）平均値代入（Mean Imputation）

〇手法の概要

・欠測値に対し、観測された標本の値の平均値を代入する。層化平均値代入は、すべての標本を適切に層化した上で、その層内の平均値を代入に用いる。

〇手順

・層化平均値代入では、すべての標本について観測されている適切な項目（補助変数）に基づき、欠測のある標本を含めて標本を層化する。

・各層内で、欠測値に対し、観測されている標本の平均値を代入する。

〇利用上の注意点等

・平均値代入は簡易な方法であるが、欠測が完全にランダムに発生している場合以外は、母平均の推定値には偏りが発生する。その改善として、適切な項目により標本を層化した上で代入を行うことにより、偏りを緩和することができる。

・補助変数として利用する項目には、欠測している項目と関連を有し、欠測のし易さとも関連する項目を使うのがよい。

・なお、平均値の補完に伴い、標本分散については過小に評価される。

2　補完を行うための主な方法

（2）回帰代入（Regression Imputation）

○手法の概要

・欠測値に対し、回帰モデルに基づく推定値を代入する。

○手順

・欠測が生じていない標本を用いて、欠測している項目を従属変数とし、観測されている項目を説明変数とする回帰モデルを推定する。

・当該回帰モデルにより推定した値（回帰直線上の理論値）を代入値とする。

○利用上の注意点等

・回帰モデルは、欠測値に対しよい予測値を与える可能性があるが、そのためには適切なモデリングが必要となる。

・説明変数に用いる変数には、連続値のほか、カテゴリカル変数などがある。
なお、説明変数を一定の層への所属を表わすダミー変数とした場合には、層化平均値代入と同じものを表わす。

・線形回帰モデルによる理論値の代入に伴い、標本分散については過小に評価される。欠測値のばらつきを考慮して、予測値に誤差項（乱数）を加える方法は確率的回帰代入と呼ばれる。

2　補完を行うための主な方法

（3）比率補完（Ratio Imputation）

○手法の概要

・欠測が発生している項目と他の項目との比率を利用して、代入値を算出する。

○手順

・欠測が発生していない標本を用いて、補完の対象とする項目（y）と他の項目(x)との比率(r)を計算する。

・欠測が生じている標本において観測されている項目(x)に当該比率(r)を乗じることで得られた値を欠測値への代入値とする。

・比率の算出は、観測されている項目を利用して適切な層区分を設定し、それら層区分ごとに行う。

○利用上の注意点等

・比率を算出する際に利用する項目としては、欠測が生じている項目に対して相関が高い項目を利用するのがよい。

2　補完を行うための主な方法

（4）ホットデック法（Hot Deck Methods）

○手法の概要

・欠測値に対し、同じデータセットの中で、欠測が生じている標本と類似した標本（ドナー）を探し出し、ドナーの観測値を欠測値の代わりとして代入する。

・標本間の距離を定義し、欠測がある標本に近い標本をドナーとする。

○手順

・欠測が生じている標本と欠測が生じていない標本について、共通して観測されている項目（補助変数）の値を基に一定の距離を計算し、最も距離の近い標本の観測値を欠測値に代入する。

○利用上の注意点等

・回帰代入のようなモデルの仮定を要しないが、類似した標本を探し出すための作業が必要となる。

・用いる距離としては、標本に関する補助変数のベクトルに関するユークリッド距離や、マハラノビス距離などがある。マハラノビス距離は、変数間の相関を考慮した距離である。

2　補完を行うための主な方法

- また、標本のすべてについて傾向スコア(注)を推定し、傾向スコアを距離としてその値が最も近い（差の絶対値が最小となる）標本の観測値を代入値とする方法もある。

（注）傾向スコア：標本ごとの補助変数の値に応じて標本が回答する確率を表わし、標本全体を用いてロジットモデルなどにより推定されたモデルを基に推定される。傾向スコアは、欠測の発生がランダムの下、モデルが正確であることが必要となる。

- 距離に基づく以外の方法としては、観測されている項目に基づきすべての標本をセルに分類し、欠測のある標本と同じセル内に存在する欠測のない標本からランダムに選んで、その観測値を代入値とする方法などがある。

- なお、ドナーを同一のデータセットではなく、過去の調査結果など別のデータセットから探す場合はコールドデックと呼ばれる。過去の調査結果を利用する場合、利用するデータが経年で安定的なものであることなどが必要と考えられる。

2　補完を行うための主な方法

（5）LOCF（Last Observation Carried Forward）

○手法の概要

- 同一の客体を複数時点にわたって調査する場合（パネルデータ）において、欠測が発生した以降の各時点の値として、直近の観測値を代入値とする。
- 欠測の発生以降、長期に適用するなどの場合は、経時による変化等を反映させるため、何らかの調整を行うことが考えられる。

○手順

- 欠測が発生している標本について、直近の観測値を欠測値に代入する。
- 経時による調整としては、欠測が生じている項目について、直近の観測値からの伸び率を欠測のない標本を用いて算出し、欠測が発生している標本の直近の観測値に乗じた値を代入値とする。

○利用上の注意点等

- 欠測が発生した以降、当該項目の値は変化しないとみなすものであるが、補完の対象とする項目によっては長期に固定して用いた場合、妥当な推計とならない可能性がある。

2　補完を行うための主な方法

（6）その他

○演繹的補完（Deductive Imputation）

・欠測が生じている標本において、観測されている項目間の関係から、欠測している項目の値を論理的に定めることができる場合、その値により補完する。

（例）費用合計の回答があり、内訳の一つにのみ欠測が生じていた場合、引き算で算出した欠測値を補完するなど

・補完に際して、一番初めに取り組むべき方法と考えられる。

○他の統計調査の結果、公開情報、行政記録情報等の活用

・欠測が生じている標本について、他の情報（他の統計調査、公開情報、行政記録情報、事業所母集団データベースの情報等）を用いて補完する。

ただし、情報の把握時点の違いや、統計上用いている定義との違いなどに注意する必要がある。

2　補完を行うための主な方法

＜ユニット単位での欠測（unit nonresponse）への対応＞

未回収など調査票の項目のすべてについて回答が得られない場合の対応としては、ウェイトを調整する方法（Weighting Adjustments）がある。

〇手法の概要

・標本設計に基づく通常のウェイトについて、欠測の状況を反映して調整した上で、推定を行う。

〇手順

・標本(i)ごとの回答確率(ϕ_i)を求め、通常のウェイト(w_i：抽出率の逆数に相当)に回答確率の逆数をかけてウェイトを調整し($w_i \times 1/\phi_i$)、当該調整したウェイトを用いて推定を行う。

・回答確率は、標本が回答する確率の推定値であり、未回収を含むすべての標本において観測されている項目※1に基づき標本をクラス（Weighting Class）に分けた場合、当該標本が入るクラスにおける回答標本の割合（回収された標本数／配分された標本数）を用いる。当該標本の区分は、欠測している項目と関連を有し、欠測のし易さとも関連する項目により行うのがよい。

※1 未回収の場合は調査項目の回答は得られないので、ここでは主に標本設計などに用いた変数となる。

2　補完を行うための主な方法

・調整したウェイトを用いて推計を行う場合は、当初配分された標本についてではなく、実際に回答があった標本について合計をとることとなる。

〇利用上の注意点等

・複数の属性などの情報を利用する場合には、標本をクラスに分けるための組み合わせが増えてしまうが、複数の補助変数の情報を回答確率という観点から一つの値に集約するため、傾向スコアを利用する方法もある。その場合、この傾向スコアの値を用いて標本をグループに分ける。

（注）傾向スコアを回答確率として直接推計に用いることもできるが、その場合、傾向スコアの値が非常に小さいと推定結果への影響が過度に大きくなるため、調整後のウェイトが極端になっていないかなどの確認も必要である。

・回答確率により修正したウェイトについて、補助変数に関する母集団総計の情報が別途把握できている場合には、それを利用して更にウェイトを調整することもできる※2 。これは、回答確率を求める区分を事後層と見た事後層化推定となっており、複数の項目によってクラスを構成し、同様の操作を行う場合はレイキングと呼ばれる。

※2 回答確率を求めるクラスごとに、補助変数に関する母集団総計が別途把握できている場合、当該補助変数に関する推計結果が母集団での総計に一致するようにウェイトを調整するものであり、w_i／ϕ_iに更に比率（補助変数の母集団総計／標本から求めた補助変数の母集団総計の推定値）を乗じたものを調整後のウェイトとする。補助変数を1とした場合は、母集団のサイズに関する情報を利用することになる。

2　補完を行うための主な方法

欠測の種類	主な補完方法	補助変数の利用
項目の欠測	層化平均値代入	標本の層化
	回帰代入	説明変数
	比率補完	比率の計算
	ホットデック法	標本間の距離の計算、傾向スコアの算出等
	LOCF	伸び率の計算
ユニット単位での欠測	ウェイト調整	回答確率を計算するクラス、傾向スコアの算出

3　補完の処理の主な手順

①欠測の発生状況の確認

〇欠測が発生しており、補完の対象となる項目の確認

〇欠測値が生じている標本について、欠測の発生状況や、他の項目・特定の属性等との関係が欠測の発生のし易さに影響していないかなどの特徴を把握

※分布を確認することや、全ての標本において観察されている適当な項目で標本を層化し、層ごとの回収率を確認するなど

②補完に利用可能な補助変数等の検討

〇①の確認結果を踏まえ、欠測している変数と関係の強い変数や欠測のし易さに関連していると見られる項目（変数）などの利用可能性を検討

〇その他、欠測の内容に応じて他の情報（当該調査の前回等の結果や外部の関連情報等）の適切な利用可能性についても検討

3　補完の処理の主な手順

③適切な補完方法の検討

○②により利用可能な補助変数等も考慮し、適切な補完の方法について検討

- まず、演繹的な補完や、過去の結果から経年で安定的なものであれば利用を検討
- 項目の欠測に対し、補助変数を基に欠測値を適切に予測できそうな場合は回帰補完や比率補完、ホットデック法等の検討。他には、層化平均値代入、LOCF（時点調整を含む）等の検討
- ユニット単位での欠測の場合はウェイト調整法を検討

○補完を行う上で層化を行う場合、適切な層区分の方法についても検討（欠測している項目と関連し、欠測のし易さにも関連する項目（変数）で層化を行うのがよい）

○調査項目ごとに異なる複数の補完方法を用いる場合は、補完の手順等を検討

○適用する補完方法間の比較を行うには、以下の様な方法がある

- 観察されている項目の一部を欠測させるなどのシミュレーションを行い、推定値の真値からの乖離を表わす指標（平均平方誤差（RMSE）など）を利用する方法
- 推定結果をセンサス調査等他の情報源と比較する方法

④補完方法の選択

○実務上の実行性等も勘案し、適切な補完方法を決定

おわりに

○今回、統計調査の実施に伴い発生する欠測値への対応として、統計作成の実務において利用が考えられる主な補完方法について概括的に整理して示した。

○適用する補完の方法は、統計調査ごとの欠測の状況等を踏まえ、適切に選択することが必要であるが、これらの補完方法全体に通じることとして、欠測による結果の偏りを緩和するためには、他の観測されている項目を適切な補助変数として利用して補完を行うことが重要である。

○しかしながら、統計調査ごとに利用できる補助変数の種類や内容などは異なっており、適切な補助変数が常に十分に利用できるとは限らない。統計調査の実施の段階で欠測をできるだけ発生させないようにすることが何より重要である。

（参考1）欠測データメカニズムについて

データの欠測がどのようなメカニズムで発生しているかについて、欠測の発生が何に依存するかという視点から、以下の3つの分類が用いられる。

〇完全にランダムな欠測（MCAR）:

・欠測の発生する確率が当該変数の値及び他の観測されている変数の値に依存しない場合である

・この場合、平均値の推定に偏りは発生しないが、実際上MCARとなる場合は少ないと考えられる

〇ランダムな欠測（MAR）:

・欠測の発生する確率が当該変数の観測された値及び他の観測されている変数の値には依存するが、当該変数の欠測となった値には依存しない場合である

・この場合、適切な補助変数で条件付ければ欠測の発生はランダムとなることから、観測された情報を利用することにより偏りを緩和することが可能である

〇ランダムでない欠測（MNAR）:

・欠測の発生する確率が当該変数自体の値に依存する場合である

・この場合、補完では偏りを緩和できるとは限らず、欠測メカニズムのモデル化など個別に対応を考える必要がある

（参考2）多重代入法について

平均値代入や回帰代入など単一の値を代入する補完法の場合、欠測値に平均的な値が代入されることにより、標本分散については過小に評価されることとなるが、これに対し、補完に関わる不確実性を考慮し、補完による精度の評価を可能とする方法として多重代入法がある。これは、一定の統計的方法により欠測値への補完値を複数発生させ、それらを代入することにより複数の補完されたデータセットを作成し、それらを基に平均値や分散などを推定する方法である。

参考文献

阿部貴行（2016）「欠測データの統計解析」朝倉書店

岩崎学（2015）「統計的因果推論」朝倉書店

高井啓二・星野崇宏・野間久史（2016）「欠測データの統計科学」岩波書店

高橋将宜・渡辺美智子（2017）「欠測データ処理」共立出版

高橋将宜、阿部穂日、野呂竜夫（2015）「公的統計における欠測値補定の研究：多重代入法と単一代入法」、製表技術参考資料

土屋隆裕（2009）「概説　統計調査法」朝倉書店

内閣府経済社会総合研究所（2017）「欠測値補完に関する調査研究報告書」

星野崇宏（2009）「観察データの統計科学ー因果推論・選択バイアス・データ融合」岩波書店

Handbook on Methodology of Modern Business Statistics. (2017), European Commission, Eurostat.

（https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en）

Kalton, G and Kasprzyk, D. (1986), The Treatment of Missing Survey Data, *Survey Methodology* Vol.12,No.1,pp.1-16

Rebecca R.Andridge and Roderick J. A. Little. (2010), A Review of Hot Deck Imputation for Survey Non-response, *International Statistical Review* 78, pp.40-64.

Roderick J. A. Little, and Donald B. Rubin. (2020*), Statistical Analysis with Missing Data*. Third edition, John Wiley ＆ Sons.