第９回評価分科会　議事録

１　日　時　令和３年３月26日（金）10:00～11:26

２　場　所　総務省第二庁舎６階特別会議室

３　出席者

【委　　員】

椿 広計（分科会会長）、岩下 真理（分科会長代理）

【臨時委員】

久我 尚子、山本 涉、美添 泰人

【審議協力者】

厚生労働省政策統括官付参事官付統計企画調整室室長補佐、農林水産省大臣官房統計部企画管理官補佐（統計調整班担当）、経済産業省大臣官房調査統計グループ統計企画室参事官補佐

【事務局（総務省）】

統計委員会担当室：萩野室長、栗原次長、鈴木次長、福田補佐、増成補佐

４　議　事

（１）欠測値の補完に係る主な方法等について

（２）その他

５　議事録

○椿分科会長　おはようございます。定刻となりましたので、ただいまから第９回の評価分科会を開催します。

お集まりの皆様方におかれましては、年度末の大変お忙しい中御出席いただきまして、ありがとうございます。

本日、神林専門委員が御欠席と伺っております。

本日は欠測値の補完に係る主な方向等について審議を行います。

用意されている資料につきまして、事務局から簡単に説明してください。

○福田総務省統計委員会担当室室長補佐

まず１つ、議事次第、その次が、「欠測値の補完に係る主な方法等について（素案）」が資料です。

資料はこれだけですが、席上配布資料として、一枚紙で、「御審議をいただきたい事項について」という紙を１枚、配布しております。

参考資料といたしまして、「欠測値補完に係る主な方法等参考資料」を参考資料１、参考

資料２として、第８回評価分科会議事概要を準備しています。

○椿分科会長　それでは、議事に入ります。

本日は、これまでの各府省統計調査における、欠測値の対応に関しての審議内容を踏まえ、欠測値の補完に係る主な方法につきまして議論をさせていただければと思います。

事務局から資料の説明をお願いいたします。

○栗原総務省統計委員会担当室次長　事務局から説明させていただきます。

これまで個別統計の欠測値への対応を、順次、御審議をいただいてきましたが、調査ごとの状況がそれぞれ異なっており、適切な補完の対応を取ることが必要であるということが一つ分かったのではないかと思います。今回、実務を行う上での参考となるように、補完に用いられる主な方法を全体的に整理し、留意点など関連の情報と合わせてまとめた案を、事務局で用意いたしました。

まず、「はじめに」としまして、この取りまとめの目的を書かせていただいています。

公的統計を作成するための統計調査におきましては、一部調査項目が未回答の場合ですとか、それから回答そのものが得られずに欠測値が発生する未回収などの場合がありますが、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と。

一方で、これらの欠測値につきましては、調査ごとの欠測の発生状況ですとか、補完に利用できるデータなどに違いがありますので、特定の補完方法の適用など、一律の対応は難しい面もある。このため、調査ごとの状況を踏まえた適切な対応が重要となりますが、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、利用上の注意点など、実務上参考となる事項を整理し、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のではないかとしております。

１枚おめくりいただきまして、欠測値の補完に当たっての考え方ですが、まず、欠測による影響です。統計調査においては、調査項目の一部未記入、それから調査回答が得られない場合など欠測が発生した場合においてそのまま集計しますと、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ますので、結果利用上の有用性に影響が生じるということです。

次に補完に関する視点ということですが、まず１つ目で、欠測に対しましては、基本的には、データ収集段階におきまして、できる限り発生しないように対処することが必要ですけれども、最終的に発生した欠測に対しましては、統計的な処理として、可能な限り補完を行うことによりまして、統計表及び平均値などの結果の有用性を確保すると。

それからまた、補完によりまして完全なデータを作成することで、ミクロデータなどによる統計分析など、そういう面からも意義を有することではないかと考えられるところです。

１枚めくっていただきまして、欠測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、一般的には欠測データメカニズムと言われておりますけれども、それらも考慮して、適切な対応を取ることが必要であると考えられます。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には、欠測のメカニズムが完全にランダムな場合以外は結果

に偏りが生じます。

その下で、参考ということで簡単に示しておりますけれども、母集団を項目ｙについて回答する層と回答しない層に分けた場合、欠測の発生が完全にランダムな場合には、それぞれの集団の平均が一致しますので、結果的に偏りはゼロとなるんですが、それ以外の場合にはゼロとはならないということが分かるかと思います。

また１枚おめくりください。

これに対しまして､補完を行う場合には､主にランダムな欠測を仮定できるのであれば、観測されております他の項目、補完の対象となる項目と関連を有し、欠測のしやすさとも関連する項目ですけど、それら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まして、平均値などの結果の偏りを緩和することができるということです。

補完のための方法としましては､幾つかの方法が一般的に用いられておりますけれども、調査ごとに、欠測の発生状況でありますとか、利用可能な補助変数の状況、さらには統計作成の実務上の適用可能性なども考慮して、適切な方法を選択することが必要であるということです。

以下に、補完に関しまして利用される主な方法としまして、項目の欠測の場合とユニット単位での欠測の場合別に示してまいります。

６ページ､ここから幾つか主な方法を紹介するというようなところになってまいります。

まず、項目の欠測、item nonresponse への対応ですが、最初に（１）としまして、層化平均値代入ということです。こちらは欠測値に対しまして、観測された標本の値の平均値を代入すると。層化平均値代入の場合には、全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ます。

手順のところですが、層化平均値代入では、全ての標本について観測されている適切な項目、補助変数に基づきまして、欠測のある標本を含めて標本を層化すると。それから、各層内で、欠測内に対し観測されている標本の平均値を代入するということです。

利用上の注意点というところで、実務の担当の方にも参考となるような注意点を情報として書くことにしてございますが、平均値代入は簡易な方法でありますけれども、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と。その改善としまして、適切な項目によって標本を層化した上で代入を行うことによりまして、偏りを緩和することができますと。

補助変数として利用する項目につきましては、欠測している項目と関連を有し、欠測のしやすさとも関連する項目を使うのがよいと。

なお、平均値の補完に伴いまして、補完値がその平均という、ある意味、平均的なところの値を代入するということがありますので、少しばらつきの方が、その分、抑えられるような形になりますので、標本分散については過小に評価されるということに注意が必要です。

７ページ、（２）としまして、回帰代入です。

手法の概要として､欠測値に対し回帰モデルに基づく推定値を代入するというものです。

手順としては、欠測が生じていない標本を用い、欠測している項目を従属変数、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します。当該回帰モデルにより推定した値ということで、回帰直線上の理論値を代入値とするというものです。

ここでの注意点として、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りますが、そのためには適切なモデリングが必要です。

説明変数に用いる変数には、連続値のほかにカテゴリカル変数などもあります。

なお、説明変数を一定の層への従属を表すダミー変数とした場合には、（１）で御説明しました層化平均値代入と同じものを表します。

線形回帰モデルによる理論値の代入に伴いまして、標本分散については過小に評価されます。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ます。

８ページ、（３）比率補完です。こちらの手法につきましては、欠測が発生しています項目と他の項目との比率を利用して、代入値を算出いたします。

手順として、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）を計算します。欠測が生じている標本において観測されています項目（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します。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、それらの層区分ごとに行うということです。

ここでの注意点としましては、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相関が高い項目を利用するのがよいということです。

９ページ、（４）ホットデック法ですが、こちらの手法についは、欠測値に対して、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、ドナーと呼ばれますが、それを探し出して、ドナーの観測値を欠測値に代入するというものです。

標本間の類似性をどう捉えるかということですが、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します。

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の値を基に一定の距離を計算して､最も距離の近い標本の観測値を欠測値に代入します。

ここでの注意点は、回帰代入のようなモデルの仮定は要しませんが、類似した標本を探し出す作業が必要となります。

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ですとか、マハラノビス距離などがあります。マハラノビス距離とは、変数間の相関も考慮した距離です。

10 ページ、利用上の注意点の続きですが、距離の定め方という点で、もう一つ、標本の全てについて、傾向スコアを推定する。傾向スコアとは、注に簡単に書いておりますが、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すもので、標本全体を用いたモデルなどで推定するものですが、モデルを使いまして、その回答する確率を推定して、その傾向スコアを距離とし、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ります。

それから、距離に基づく以外の方法としまして、観測されている項目に基づき、全ての

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び、その観測値を代入値とする方法などもあります。

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ます。過去の調査結果を利用する場合には、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ます。

続きまして、11ページ、（５）のＬＯＣＦ（Last Observation Carried Forward）です。同一の客体を複数時点にわたって調査するような場合、パネルデータのような場合において、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します。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ます。

手順としては､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します。

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します。

ＬＯＣＦでは、欠測発生以降当該項目の値は変化しないとみなしていることになりますが、補完の対象とする項目によっては、カテゴリカル変数の様な時間がたってもあまり変わらないようなものは良いですが、それ以外の数量的な変数などの場合は、長期に固定して用いた場合は、妥当な推計とならない可能性がある。従って、何らかの調整を行うことが考えられるというものです。

12 ページ、（６）その他ということで２点ほど挙げています。１つが演繹的補完で、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、例えば、費用合計については回答があるが、その内訳の一つが欠測しているような場合に、差引きすれば結束している値が一意に定まるというようなケースです。この方法については、補完に際して、一番初めに取り組むべき方法と考えられます。

それから、それ以外の方法で、他の統計調査結果、あるいは公開情報、行政記録情報等のデータを使うということも考えられます。ただし、それぞれのデータ源によって、情報の把握時点の違いや、用いている定義の違いなどがよくありますので、その点は注意する必要があると付言しています。

以上が項目単位の補完の場合の主な方法です。

続きまして、13ページ、ユニット単位での欠測への対応です。

未回収など調査項目の全てについて回答を得られない場合の対応としては、ウエートを調整する方法、Weighting Adjustments という方法があります。

手法の概要は、標本設計に基づく通常のウエートについて、欠測の状況を反映して調整した上で推定を行うというものです。

手順ですが、標本（ｉ）ごとの回答確率（ϕ_i）を求め、通常のウエート（w_i）これは抽出率の逆数に相当するものですが、これに回答確率の逆数を掛けてウエートを調整し、当該調整したウエートを用いて推定します。

回答確率は、標本が回答する確率の推定値でして、未回収を含む全ての標本において観測されている項目、こちらは未回収の場合ですので、調査項目の回答は通常得られませんので、ここでは主に標本設計などに用いた変数になるかと思いますが、それに基づいて標本をクラス、Weighting Class などと呼ばれていますが、それらに分けた場合に、当該欠測している標本が入るクラスにおける回答標本の割合、すなわち配分された標本数分の回収された標本数で、その回答確率を推定するというものです。

この標本を区分するときには、欠測している項目と関連を有し、欠測のしやすさとも関連する項目により行うのがよいというものです。

14 ページ、一番上は、調整したウエートを用いて推定を行う場合は、当初配分された標本、（ｎ）ではなく、実際に回答があった標本（ｒ）についてΣを取るということを書いています。

利用上の注意点としては、複数の属性などの情報を利用する場合は、標本をクラスに分けるための組合せが増えてしまいますが、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という方法もございます。その場合には、この傾向スコアの値を用いて標本をグループに分けます。

なお、注で書いていますが、傾向スコアについては、グループ分けだけではなく、回答確率として直接推計に用いることもできますが、その場合、傾向スコアの値が非常に小さい場合は、その逆数を取りますので、推定結果への影響が過度に大きくなりますので、調整後のウエートが極端になっていないかなどの確認も必要としています。

それから、さらに回答確率により修正したウエートについて、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は、それを利用して、さらにウエートを調整することもできます。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いまして、複数の項目によりクラスを構成し、同様の操作を行う場合にはレイキングと呼ばれます。

15 ページ、こちらで今御説明した手法を少し整理して、一覧形式にしたものです。欠測の種類と、それから主な補完方法、併せて補助変数の利用の仕方ということで整理してみたものです。ここまでが補完方法の説明です。

16 ページは、その補完方法を踏まえ、補完処理の主な手順を整理しました。

まず、①として、欠測の発生状況の確認です。欠測が発生し補完の対象となる項目をまず確認します。欠測が生じている標本について、欠測の発生状況や、その他の項目、あるいは特定の属性等との関係が欠測の発生のしやすさに影響していないかなどの特徴を把握すると。その際、分布の確認や、特定の適当な項目で層化して、層ごとの回収率を確認するとか、そういったことを確認としては使えると。

②としまして、補完に利用可能な補助変数等の検討で、①の確認結果を踏まえまして、欠測している変数と関係の強い変数や、欠測のしやすさに関連していると見られる項目などの利用可能性を検討すると、その他、欠測の内容に応じまして、その他の情報の適切な利用可能性についても検討するとしています。

次のページで、③としまして、適切な補完方法の検討ということです。②によりまして

補助変数等の検討結果も考慮しまして、適切な補完の方法について検討すると。

例えば、ポツで書いてございますけれども、まず、演繹的な補完でありますとか、過去の結果から経年で安定的なものであれば、その利用を検討すると。それから、項目の欠測に対しまして、補助変数を基にして、予測値を適切に予測できそうな場合には、回帰補完や比率補完、あるいはホットデック法等の活用を検討すると。その以外の場合には、層化平均値代入やＬＯＣＦ、時点調整を含めて、そういったものを検討するということが、項目の場合には考えられると。それから、ユニット単位での欠測の場合にはウエートを調整する方法を検討するということを挙げております。

次に、補完を行う上で層化を行う場合には、適切な層区分の方法についても検討するということ、また、項目ごとに異なる複数の補完方法を用いる場合には、その補完の手順等を検討するといったこと、さらに、適用する補完方法間の比較を行う際には、以下のような方法があるということで、観測されております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まして、推定値の真値からの乖離を表す指標として平均平方誤差などで定量的に評価するような方法ですとか、推定結果をより信頼度の高い、より全体的な、例えば、センサス等の情報源と比較するような方法が考えられるとしています。

最後、実務上の実効性等も勘案して適切な補完方法を決定する、このような流れが一つ考えられるのではないかということで整理しております。

18 ページ、「おわりに」ということで、少しまとめ的なコメントを書いております。

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ました。

適用する補完の方法については、統計調査ごとの欠測の状況等を踏まえて、適切に選択することが必要ですが、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と。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こともあり、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としています。

19 ページには、本文中にも出てまいりました欠測データメカニズム、欠測の発生が何に依存しているかという視点で、一般的に使われている定義・用語を挙げています。

それから、本文の方では取り上げませんでしたが、多重代入法と言って、補完に係る不確実性も考慮して、補完による精度の評価を可能とするような方法、少し高度な方法ですが、その様な方法もあると、参考として紹介させていただいております。

最後、20 ページで、今回の整理の参考文献を掲載しています。

資料の方は以上ですが、参考１として、今回取り上げた主な方法について、その数式であるとか、各府省調査の中での、主な活用事例を簡単にまとめています。こちらの方も併せて実務の参考に資していくことを考えております。

また、席上配布資料として、本日御審議いただきたい点の簡単なレジュメを用意しています。資料の説明については以上です。

○椿分科会長　どうもありがとうございました。

席上配布資料にもありますように、全体の構成に係る御意見、取り上げている手法、記述内容に係る御意見・御質問、その他、補完にとって、実務者にとって参考となる情報なに関する御質問や御意見を頂きたいと思います。

○久我臨時委員　こちらの資料はどういう形式で、どのような方々に提供するのでしょうか。各府省の統計実務担当者だとは思うのですが。

○栗原総務省統計委員会担当室次長　最終的には、審議結果報告書の中に、この資料を入れ込んで、各府省に共有を図っていくことを考えています。

○椿分科会長　各府省の実務担当者がこれを読むための予備知識もがどれくらいあるかということも大きな問題ですね。

○久我臨時委員　実務の知識によっては、質問とか相談をしたいというニーズが出てくると思いますが、それは総務省統計局に何か窓口のようなものができるのでしょうか。

○栗原総務省統計委員会担当室次長　今回まとめた内容とか関連の事項であれば、当方でもお答えいたしますし、各府省の技術支援を行う役割として、統計研究研修所も各府省の支援を行うことはできます。

○椿分科会長　統計研究研修所に相談や報告が気楽に行けるような雰囲気になってくるといいですね。

○山本臨時委員　各府省自身のところでやっている補完方法はこれだなと分かると思うのですが、それ以外を見たときに用語全体が専門的、統計的なので、例えばこの統計でこれが用いられているといった、相談する際に各府省の統計の例示があるとよいように思います。

○栗原総務省統計委員会担当室次長　補足ですが、参考１によりそれぞれの手法についてまとめたものも併せて各省の方にお配りしたいと思っています。

○山本臨時委員　参考１、拝見してこれはありがたいと思います。調査名を例示して頂くだけでも、各府省で調査の更新作業をするときに参考にできるといいと思います。

○美添臨時委員　私が見ている限りは、なかなかバランスが取れた表現だと思います。

先程、久我先生、山本先生からも質問や指摘があったように、この結果をどう位置付けるかを考えると、参考文献を見て、統計委員会担当はよく勉強しているというアピールにはなっている。何か所か語句の手直しをお願いしたいところがありますが、不正確なところを直せば、今回の目的には立派な答えだと思います。

さらに言えば､アメリカで 25 年前に｢標本調査における不完全データ｣(Incomplete Data in Sample Surveys)という、統計局、センサス局と大学の先生方、参考文献に出てくるルービンさんの先生のテンプスター先生、モステラー先生、コクラン先生も入っていた研究グループがまとめた本で３巻もありますよね。１巻が理論編だったと思う。２巻と３巻目には実例も含めている。あれは、センサス局がやったわけですよ。日本で数年前に内閣府が数人の先生方を巻き込んで調査研究報告書を作ったと思いますが、一番大事なのはアメリカのセンサス局が行った様に総務省統計局が行うことでしょう。今回はこれでいいですが、今後もアメリカの不完全データに関する膨大な報告書の様に、理論があり、実例、各省で何をしているかも書いてあるものを、外部の先生方を巻き込んで研究会を数年行って

作るぐらいの意気込みがあるとすばらしい。将来的に、それを是非やっていただきたい。

２ページ目「はじめに」で、「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」とあり、これについて、６ページにアイテム・ノンレスポンスが、13 ページにユニット・ノンレスポンスがある。５ページで、欠測値の補完に当たっての考え方の一番下に、説明無しに「未回収などユニット単位での欠測」という表現が出てくるので、例えば、ここに「13 ページで解説する」とか、「13 ページ以下参照」とか入れていただくと親切か思います。

２ページに戻ると、「一部の調査項目が未回答である場合」とある。ここに何ページ以下の項目欠測、アイテム・ノンレスポンスとかユニット・ノンレスポンスを参照とか書くといい。

２ページ目は、なかなか格好いいこと言っているなと思います。

６ページ目、平均値補完で平均を推定するとあるが、統計検定の世界でも用語を統一しようという人たちがいて、「平均」「平均値」の用語を、統一すべきことかどうかも含めて考えていただけますか。

最後 18 ページ、これが一番大事なので、これを冒頭にも同じことを書いてほしい。欠測値の補完に当たっての考え方。そもそも欠測値は難しいので、発生させないような努力、減らすような努力が求められるところであるとした上で、実際に欠測値が出たら、ここに以下に書いてあるような手法を考えてくださいというのが、姿勢としては正しい。

○栗原総務省統計委員会担当室次長　その点につきましては、３ページにまずできるだけ発生しないように対処することが必要とした上で補完の方法を説明し、最後にもう一度改めて、入念に強調している形にしています。

○美添臨時委員　分かりました。「基本・・必要であるが」とありますが、「基本的には必要である」で、１回打ち切った方がいいのではないか。書いてありますが、これは強調した方がいいかなと思いました。「基本的には・・対処することが必要である。」で終わりにする。２番目の白丸を設けて、欠測値が発生してしまったら、そのままにしておくのではなくて、何らかの補正をしましょうということで、「最終的に発生した欠測値に対して・・」云々というところをまとめる。

可能な限り補完を行うことにより統計及び平均値などの結果有用性を確保するとありますが、口頭では説明があったのですが、ミクロデータとして提供する際の結果の有効性を書いてもいいのではないか。この頃、各府省はミクロデータを公開する方向なので、その際にも有用な方法を考えるということで。ミクロデータを公開しないなら平均値代入でいいのではないかという考え方もありますので。

４ページはおおむねすっと読めました。

５ページ、ちょっと引っかかったのは最初の○「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」という記述。19 ページに欠測の定義が書いてあって、そこを見れば、ここは MCAR ではない方、MAR を指しているのではと思われるのですが、このランダムな欠測が MCAR だとすると、別に観測されている他の項目、補助変数がなくてもいいのですよね。サンプルが減っただけなのだというように思われないような、何か工夫をした方

が良いのでは。「ランダムな欠測を仮定できるならば」という記述は、削ってしまってもいいのではないかという気がします。

６ページ目は、利用上の注意点等で１行目は平均値代入とあり、２行目は母平均の推定値。母平均値とは言わないだろうな、母集団平均値なのか。何か気持ち悪いなと思って、微妙な違和感がありました。それは文章点検のときにお任せします。

それから、７ページ目回帰代入では、手順には回帰モデルと書いてあるだけで線形とは特定していないが、下から３行目は線形と特定している。下から６行目にはカテゴリカル変数がありますし、「線形」という記載は要らないのではないか。カテゴリカル変数の、クロスタームを考えるときには掛け算するのを線形というのでしょうか。

それから、８ページの一番下ですが、比率補完って、Ratio。Ratio だったら相関係数ではなくて比例関係ですよ。比推定の関係と同じで、原点から離れたところで線形関係があっても困るわけで。

○椿分科会長　比例関係が強いにしないといけません。

○美添臨時委員　13 ページ、ほかのところは何とか分かるにしても、このウエート調整、は式がないと具体的に何をするのか分からないと思います。参考資料に入っていますかね。次の 14 ページもそうですね。レイキングも実例がどこかで参考資料にあると、何を言っているのかがよく分かると思います。

○栗原総務省統計委員会担当室次長　今のウエート調整については、参考資料の 19 ページに式があります。

○美添臨時委員　ほかのところも含めて、参考１に実例や細かい説明が書いてあるということを書いていただけると、安心して読めると思います。

最後のページに参考文献を付けているのは、よく勉強しているなという感じが分かり、各府省にもこんなに勉強しなきゃいけないのかというメッセージが与えられて、なかなかすばらしいと思いました。もっと書いてもいいのではないか。統計局とか関係者の方がいるわけですよね。全員統数研と統計研修所、統計センター絡みの人ばかりですよね。できたら、これを外国語で出して、頑張って宣伝したらいいなと思いました。

参考の１は、各府省でやっている報告をまとめていますが、この中に、例えば、財務省は法人企業統計等について随分丁寧にやっていますので、情報を出してくれると思うのですが。

○栗原総務省統計委員会担当室次長　ありがとうございます。法人企業は８ページの方に記載があります。

○美添臨時委員　分かりました。活用事例は手法ごとに書いてあるのですね。細かいことで気になった点を除けば、全体のバランスもいいと思います。

○岩下分科会長代理　いろいろな指標ごとにやることを１回決めましたと。これが時代の変遷に伴って、多少、統計の内容も見直していく段階で、どういうタイミングで、このままでいいとかを決めていかれているのか教えてください。

○椿分科会長　欠測値の補完は、欠測されているものの偏りのない推定値とか予測値というものをうまく作って、精度よく作って入れればよくなるというような話を入れておけば、

時代にあまり流されないで、方法は少しずつ変わるということかと思いますが、事務局からは何かありますか。

○栗原総務省統計委員会担当室次長　今回、欠測対応が十分に行われていないものなどをこの分科会で取り上げていただいて、対応がかなり進むようになってきたというところがあります。それから、統計局のように、欠測対応をしっかりやっているようなところでは、５年ごとの周期調査の見直しのときなどに、欠測方法もブラッシュアップするとか、そういう形でやられているところもあるようですので、それが一つ理想的な形、時代の社会の変化も踏まえて見直していくということかとは思います。

○岩下分科会長代理　まさしく５年だと、大体基準年の改正ですよね。

○椿分科会長　５年に１度ぐらいというのはいいタイミングかもしれませんね。今、割と世の中すごく、人工知能とかが目立ってきましたから。

○岩下分科会長代理　そうですサイクルが速くなっている気がしますね。

○山本臨時委員　参考資料の１の 11 ページ、この（４）という式はミニマックスと書いてあるのですが、マックスしか取っていないので、ミニマックスなのでしょうか、距離関数と書いてありますが。

○椿分科会長　Ｌ無限大というものではないでしょうか。

○栗原総務省統計委員会担当室次長　マックスが一番小さくなるようなものをミニマックスと呼んでいるようです。

○椿分科会長　この距離関数のミニマムを取るという話ではないですか。最近隣法ですから。

○山本臨時委員　距離関数はマックスなのですが、これはおそらく最大距離といいまして、それをミニマムにするようにマッチングをするのでマックスなので、距離関数としては、多分マックス距離とか、多分最大距離のようなものだったと思うのですが。それで最近隣を取るとミニマックスになると思います。

○美添臨時委員　山本先生の、（２）式がミニマックスでと言いたいのでしょうからね。

○山本臨時委員　はい、そうですね。

○美添臨時委員　11 ページの式４のところは、「無限大に近づけると、以下の距離関数が得られる」だけで、「ミニマックスと称される」という記述を削るのが正しいのでは。

○椿分科会長　その方が無難かもしれない。確かにそうかもしれない。

○山本臨時委員　美添先生おっしゃるとおり。「ミニマックスと称される」は除くと良いと思います。

○岩下分科会長代理　「実際のところ」からですね。

○山本臨時委員　「実際のところ・・」を「式３において、「Ｚを無限大に近づけると、以下の距離関数が得られる」にしてしまって、名称だけ取るという美添先生の案でどうでしょうか。

○栗原総務省統計委員会担当室次長　申し訳ありません。ここはＥＵの関連のマニュアルを和訳したものとなっておりまして、この表現が実際に使われているというものです。

○椿分科会長　Ｌ無限大距離だというふうに数学で言っていることはわかりますがね。

○美添臨時委員　ここは間違いですとコメント付けるのと、黙って削除するのと、どっちが親切かという話でしょう。

○山本臨時委員　消し線でもいい気はするのですが。

○美添臨時委員　黙って修正していいのではないですか。ＥＵにここをこう直しましたと連絡しておけば。

○山本臨時委員　もしかすると、訳が以下の距離関数を用いたミニマックスに一致する、ミニマックスになるというようなことかもしれない、おそらく文法だけかもしれないですが。

○椿分科会長　後で原文確認してみますかね。たまにそういうことをやっておくといいかもしれない。距離としてミニマックス的な意味を持っていることも理解できますが。

○山本臨時委員　資料の方に戻りますが、３ページについては私も美添先生と同意見で、補完技術を紹介するときには、まず可能な限り回収率を上げる努力をすべきであるとか、項目無回答に関しては、客体に確認するなどの努力をした上でやむを得ず生じた欠測については補完するという様な枕詞がメッセージとして最初にあった方が良いのではないでしょうか。内容としては再度ですが、一つ目の○の「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・・」という記述箇所で、「まず回収率の向上の努力や、不完全な回答の客体の確認の努力を重ねた上でも」といった様な記述を追加しておいた方がよいように感じました。技術としては整理すべき情報ではあるのですが、私も美添先生と同じく、活用しないのがもちろん望ましい、最初から前提としないのが望ましいという意見です。

また、何か所かに傾向スコアについての記載があります。10 ページのホットデックのところにあるのはおそらく傾向スコアマッチングだと思います。

○椿分科会長　「傾向スコアを距離として」という記載のある部分ですね。

○山本臨時委員　はい。もう一つ、推定に「傾向スコアを利用する方法もある」という記載が 14 ページにあります。

○椿分科会長　ユニット単位での欠測への対応についての箇所ですね。

○山本臨時委員　はい。10 ページのホットデックのところには、「傾向スコアマッチング」という名称を括弧書きで付け、14 ページには、「傾向スコアによる推定」の様な付記をした方が良いのでは。同じ傾向スコアでいろいろな事ができると思われるといけないので。

○椿分科会長　前半はマッチングですが、後半はその確率自体の推定が意味がある。

○山本臨時委員　そうですので、ただし書とか括弧書でよいので記して分けた方が親切かと思いました。

それから、従来ほかの先生方が御指摘されていたことですけど、用語の揺らぎが今回もあります。「調査客体」と「標本」と「ユニット」と「単位」という用語が使われていますが、同じものです。もう一つ、「アイテム」と「項目」と「変数」とが一緒です。ですから、調査票に関しては項目だと思いますが、データに関しては変数といった様な使い分けがあってもいいかなと思いました。

多分、そうですね。「ユニットノンレスポンス」と、英語を片仮名で訳したところがある

ので、「単位」って出てくると思うんですが、「標本」という用語にはデータ全体というか、サンプル全体としての「標本」を指す場合と、レコードとしての「標本」を指す場合と、２つの使い方がありますので、そこは少なくとも使い分けをした方がよいと思います。「客体」という言葉が途中で出てくるので、レコードは「客体」としてもいいかもしれないです。

あとは、先ほどやっぱり美添先生がおっしゃっていたことですが、４ページや、６ページの平均値代入の利用上の注意点に「欠測が完全にランダムに発生している場合以外は」と書いてありまして、この時点で「完全にランダム」が意味することを理解していることを前提とするのは難しいので、小さくてよいので、19 ページに参考で付けてある「完全にランダムな欠測」についての説明を、４ページの下に書くと良い。次に５ページに、「主にランダムな欠測を仮定できるなら」とありますが、「主に」はなくていいような気がします。ランダムな欠測は、５ページの下にスペースありますので、ここに引用していただいて、ミッシング・ノンアットランダムはこの資料には出てこないので、19 ページにある参考１は参考資料にしていただいてもいいような気がします。

17 ページになります。これも今、議論がありましたところですが、16 ページで、ほかの他府省調査や、ミクロデータの活用のようなことが視野に入っていると思うのですが、17 ページの③の１つ目のポツの次ですね。「まず、演繹的な補完や過去の結果から」と書いてありますが、ここに過去の結果だけでなく、他府省のデータ・調査結果も入るのではないかと思いますので、重複かもしれませんが、明記された方が良いと思います。

○椿分科会長　演繹的な補完の中には行政情報からの補完とか、問合せによる補完とか、いろいろなものがあり得ると思いますね。

○椿分科会長　非常によくまとめていただいたということを前提の上で、先ほど美添先生のご意見と、事務局の説明もありましたが、この前文の有用性というのは、どちらかというときちっとした欠測値補完をすると集計の推計精度が上がるという有用性がほとんどだったのですが、この評価分科会で時々ふれていますが、ミクロデータが公開される場合には、欠測値補完は、無から有を作ったという誤解を受ける可能性が極めて高い。ＬＯＣＦなどは、過去のデータをずっと使い続けるもので、それを続けてマスコミにねつ造と言われたこともあります。ですから、欠測値補完は集計精度を上げるために必要だということと、公的統計の世界には、残念ながら未回答や欠測があり、その適切な処理を事前に定めておく必要があるということを徹底的に言っておかないとミクロデータが公開された時に怖い。研究者に公表するミクロデータの中に欠測情報を入れるか入れないかということも極めて重要な問題です。研究者が使ってみて、実はこのデータは、本来の調査データじゃなくて、欠測を補完したデータだったといった場合に、研究者側の利用者の反応は非常に心配なものがあります。

欠測データに対するミクロデータレベルの話をここでする必要はないかもしれませんけど、将来、研究者が使えるミクロデータについて、欠測値であったかどうかという情報を付与するかどうかという話について、研究者側からいったら間違いなく付与すべきだと私も思いますが、それが一朝一夕にできるかどうかというと大問題なので、それを論点とし

て残しておいていただきたい。この手続き、よく書けているのですが、これも今のような問題が無く、集計精度を上げるという問題だけだったら、とにかくそれにベストを尽くしてくださいという話です。しかし、結果を見て方法を決めているのではという話になったとき、つまり、この推計精度が、こっちの方にぶれてくれたからよかったとかという形で欠測値補完技術が選択されるとなると、やはり世の中に非常に大きな誤解を与えるので、どういう欠測値補完をやるかは、前回の調査結果を見ながらこういうふうにしましたというような手続き的な透明性みたいなものというのも本当はあった方がいいかと思う。特にミクロデータに限らず、集計の有用性においても何かそのような視点があった方がいいかと思います。統計の集計側だと今回のデータの推計精度を上げようとベストを尽くすので、あまり首を絞めるべきではないと思うのです。そういう点は少し感じられる。

手法はこれから非常に進化していって、美添先生がおっしゃられたように、線形モデルに限らず、いろんな補完技術ができてくるだろと思うのです。それはそれとして、推計精度を上げるためには欠測になったデータをできるだけ偏りなく、精度よく予測推定することが必要だという原理原則を入れておいていただければ、それでいいのではないかということです、先ほどの線形モデルの話を伺って強く思いました。

また、美添先生、山本先生おっしゃられたように、欠測をそもそもなくす努力の話が前文に入ることと、その他で書いてありますが、事前の処理として、演繹的な処理とか、そういうものをきちっとやった上で、最終的に統計的処理を行う、統計的な補完以外に重要な補完技術があるということですね。

あとは、最初に先ほど美添先生がおっしゃられた、アメリカセンサス局の研究報告書ですが、シンフォニカか何かから報告書で出版された記憶がありますが。

○美添臨時委員　しましたね、３年ぐらい、４冊ぐらい出したかな。あれ、科研費研究の報告書でしたね。

○椿分科会長　そういうものも含めて、私としては、今回、非常にいいものを政策統括官の方でまとめていただいたので、これを統計研究研修所などの、データアナリストやデータアナリスト補の研修に生かしていくようなこともやっていただくと非常に良いのではないかと考えています。

第１段階としては大変いいものをまとめていただいたのですが、世の中が欠測値というものに対して、どのように考えているかと、統計のプロフェッショナルが考えている印象と、一般の方が考えている印象とは大分違っていると思いますので、周知を入念にしていただいて、この技術がバイアスなどのない統計を作るために重要なことを周知徹底していただけるといいと思います。

○美添臨時委員　各府省で、欠測があった、未回収があったにもかかわらず、全く無視して集計していて偏りがあることを指摘された、それが大事ですという簡単な話を入れてもいいかもしれないですね。

○椿分科会長　過去の失敗に学んでいるということですね。

○美添臨時委員　そもそも、それは決定的にまずいのだと。30 年ぐらい前に当時の通商産業省が企業活動基本調査について、最初は 90％ぐらいの回収率だったのが回収率が 70％

台に下がったときに、欠測値対応せずに集計すると景気がよくなっているのに数字が減っているというおかしなことになって、中で研究会を実施してそこは直したのです。そのような苦い経験もあって学んできたわけですよね。そんなこともやって、成果を上げたという事を、どこかに書いてあるといいかなという気がします。

もう一つ、ミクロデータ公開はずっと気になっているのですが、研究者に対して、どこまで情報を提供するかという点については、一橋の松田先生が開催されたミクロデータの研究会に、椿先生、山本先生と一緒に参加させてもらいましたが、基本的には欠測値処理をしたらフラグを立てろということでしたね。しかし、欠測値処理の方法にも、椿先生の指摘のように、統計的な方法だけではなくて、行政記録情報の活用や、演繹的な方法があるので、それは意味が違うのです。行政記録情報の正確な数字を活用して欠測を補完しましたというのは、欠測値補完というかなという気がします。ミクロデータで、行政記録情報の活用や演繹的補完で合計が合わないところを埋めましたというのと、統計的に補完をしましたというのとはちょっと違うかなという気がします。

○椿分科会長　もともと研究者に渡すデータのレベルというのがあって、私が言っているのは昔の環境関係の衛星データですけど、ノイズで欠測のあるそのままのものを渡すというのがレベルゼロで、次に、それを使える人は滅多にいないのでという理由で、補完やノイズ処理を行って渡すというレベル１というデータがあるのです。ただ、それを言うのは簡単ですが、それを２つ作れと今すぐに言うのは無理という感覚もありますね。フラグが付いていれば、レベル１のデータをレベルゼロに戻すことは容易なので。

○美添臨時委員　そうですね。

○椿分科会長　一応、そういうことも含めて課題はあるだろうと思うのです。ミクロデータという問題が出てきたために、新たな課題が出てきたことは間違いない。

○久我臨時委員　全体の構成のお話ですが、私も、先生方御指摘のとおり、もう少し枕詞で、回収率向上だとか、欠測値のこれまでの話を入れた方がいいなと思ったのと、まず方法についてずらっと述べられて、３番目に補完の主な手順というのが来るのですが、手順があった上での方法かと思うので、２と３を逆にした方がよいと思いました。

さらに、その２番ですが、内容はすごく分かりやすくて、頭にすっと入ってくる文章で、よく勉強されておられるなと思ったのですが、15 ページの一覧表が先に来た方がまず全体論が見えて分かりやすいのかと思います。それから、一つ一つ説明していただいた方が分かりやすいかと思いました。

○椿分科会長　補完方法について初めて目にするユーザーを対象とすると、俯瞰があって次に具体的な方法に入っていく方が分かりやすい、それからあと、手続きがあって、その中で使う方法があった方が分かりやすいのではないかということですね。

○久我臨時委員　そうですね。実務の御担当者の方には伝わりやすいのかなと思いました。

○椿分科会長　事務局の方、そのような御意見があったということを参考にして、構成を考えていただけますでしょうか。私も普通のユーザーのことを考えたら、確かに久我先生のおっしゃる方が読みやすいだろうなと思います。また、そうしていただかないと、その他のところで書いてある補完が実は重要な処理であるということがあまり伝わらないかも

しれないですね。

○久我臨時委員　その上で、相談先というか問合せ先が一番後ろにあった方が良いです。

○椿分科会長　この件に関して御相談するときは統計研究研修所にというようなことで良いでしょうか。

○久我臨時委員　良いかもしれないですが、各府省の御担当者の方が、その統計の担当者の方に御連絡を取るというのも、なかなか難しいのではないでしょうか。

○椿分科会長　窓口が必要ということですね。

○久我臨時委員　はい。

○栗原総務省統計委員会担当室次長　以前、この分科会でもご報告があったかと思いますが、統計研究研修でも相談窓口というのを既に作られて周知を図られていますし、私どもの方に一旦つなぎ、ここを経由してつなぐのでも、そこはもちろんかまいません。

○椿分科会長　こういう文書を作って周知するだけでなく、いざとなったら相談できる窓口というのは非常に重要な気がします。

今回の本当にこういう補完の方法って、先ほどあったように、確かに過去からいろいろな問題があって、実は統計の数値がおかしくなったと、出す値がおかしくなったということ、いろいろあったわけですね。そういうことを踏まえて、これができているというような位置付けを明確にすること。統計的な欠測値補完というのは、あくまで最後の手段で、その前にやることがあるというのが、きちっと伝わるようにしておくこと、それを明確にした上で、技術的な問題についても、いろいろご意見頂戴しましたので直していく。それから、重要なこととして、構成上、誰がユーザーだから、どのような順番で書くのが便利かということについて配慮していただくということで、これは事務局と私も含めて、最終構成をこのようにしますということを、評価分科会の委員、専門委員の先生方にも少しチェックしていただくといった形で取りまとめをしていきたいと思います。

今日頂いた意見を反映したものを事務局に整理してもらって、それを確認いただくという手続きをした上で、最終的に審議結果、報告書にしていくという進め方になると思いますが、それでよろしいでしょうか。

（「異議なし」の声あり）

○美添臨時委員　参考１、この場で初めて見て、勉強になりそうだなと思って見ているのですが、各府省から頂いた資料で、引用が書いてあります。例えば、6ページ目のDe Waal et (2001)、でも詳細情報がないので、出典を、どこかに書いていただけませんか。少なくとも引用しているところは、ここに載せていただけると使いやすいかなと思いました。Farrell and Barrera(2007)、Rao(1996)も。

○椿分科会長　それは調べていただいて、正確な文献、引用文献一覧を作っておいてください。審議報告書に入れますので。

○山本臨時委員　目次１ページ目に、今、大項目しか挙がってないんですが、（１）、（２）も明記していただくと、どういう手法が書いてあるとか、考え方がどういう観点があるのかみたいなことが分かるので、プレゼン資料としては、この目次がよいとは思うのですが、配布する資料としましては、（１）、（２）までを羅列で、縦ではなくて、（１）何とか、（２）

と、改行しないタイプでよいので、明記していただいた方が良いです。あと、参考１も同様に目次を、（１）、（２）までつけていただくと、資料として、かなり有効なのではないかと思います。

○椿分科会長　昔だと審議報告書という紙媒体で、冊子媒体にしているから目次が付いている。今はこのような形にして、場合によっては、Ｚｏｏｍで開催して、そのまま流すと、簡単な研修になります。

しかし、いずれにせよ、目次に付けていただく。事務局と調整の上、させていただければと思います。

それでは、先ほどのような形で、事務局とともにまとめて、先生方にまたチェックしていただくということで、大変恐縮ですが、よろしくお願いいたします。

それでは、予定されている議事は以上で終了いたしましたので、ここまでとさせていただきます。

最後に事務局から、次回の日程について連絡をお願いできればと思います。

○福田総務省統計委員会担当室室長補佐　次回は４月 15 日木曜日の午後４時に開催する予定です。４月 15 日木曜日の午後４時から開催いたします。場所は今回と同じ第６特別会議室で行います。よろしくお願いいたします。

○椿分科会長　本日は大変お忙しいところ、年度末にもかかわらず、御審議に御協力いただいたこと、本当に感謝申し上げます。どうもありがとうございました。これで閉会したいと思います。