解　説

看護研究の統計解析における欠損値の対応方法

谷村　晋

Handling missing values in statistical analysis for nursing research

Susumu Tanimura

Ⅰ．はじめに

欠損値（missing value, missing data）は，欠測値とも訳され，未検査の項目や質問票調査における無回答項目など，ある変数の値が欠落している空欄のことである．論文の表などではNA（not available）[*1]と表現されていることが多い．ほとんどの統計解析手法は完全データを前提としている（Rubin, 1987）ために，欠損値がない完全データであることが少ない看護学の研究データを扱うとき，欠損値から生じるバイアスにより研究結果が歪められる．これを避けるために欠損値への適切な対応が求められる．さらに，研究成果を論文にまとめる際にも，どのように欠損値に対して対応したのかを研究方法に記述し，また，その影響をどのように評価したのかを考察で述べることが求められる（Rue et al., 2008）．

近年，欠損値に対する理論的な整備が行われ，対応方法の開発研究も進んできた．しかしながら，看護学研究における統計解析の教科書（例えば，Grove and Cipher, 2019）には，欠損値の解説が見当たらず，数少ない看護研究における欠損値の解説（例えば，El-Masri and Fox-Wasylyshyn, 2005; Fox-Wasylyshyn and El-Masri, 2005; Patrician, 2002）は網羅性に欠けており，看護研究の初学者は欠損値の扱いに当惑している可能性がある．

そこで，本稿は，欠損値を含む看護研究データへの対応方法について系統的な整理を行った．また，代表的な欠損値対処法である多重代入法の実践的な手順を付録に示した．

Ⅱ．欠損値の生起メカニズム

欠損値への対応を行うためには，欠損値が生じた背景に存在する生起メカニズムやその原因について考察する必要がある．欠損値の生起メカニズムは，完全にランダムな欠損であるMCAR（missing completely at random），他の観測データに依存する欠損であるMAR（missing at random），欠損データに依存する欠損であるMNAR（missing not at random）の3つに大別される．

MCARは，何者にも影響されず完全にランダムな欠損値の発生であり，例えば，偶発的に試料を紛失してデータが欠ける場合や，保健医療情報システムの過度な入力負担から欠損値が生じる業務データなどはMCARに分類される．この場合は対処が最も容易である．MARは，他の変数と連動してランダムに欠損値が発生する場合であり，その変数の値とは無関係な欠損値の発生である．例えば，縦断研究において高齢であるほどフォローアップ調査会場に来ることが困難になりやすい場合は，年齢の影響を受けて欠損値が発生する．学校の体重測定日に女子生徒が仮病で学校を欠席する場合は，性別の影響を受けて体重の欠損値が発生する．MNARは，その変数自体の影響により変数の値に連動してランダムに欠損値が発生する場合である．例えば，自記式調査で年収を尋ねたときに，低収入の者が回答しない傾向にある場合や，未診断の認知症患者が神経心理学的検査を理解できずに欠損値が生じる場合などである．これらの欠損値の生起メカニズムによって，対処方法が異なる．

三重大学大学院医学系研究科

[*1] NAは，not applicable, not assessed, no answerなどの略語でもある． N/AやN.A.と表現されることもある．

III. 欠損値の対処方法

看護研究において，欠損値を含む割合は一般的に10%まで許容されるとされている（Langkamp et al., 2010）．しかし，研究者や研究デザインによって当然ながらさまざまな境界値が提唱されている（Roberts et al., 2017）．

欠損値に対応するためには，まず最初に欠損値を観察し，その割合や欠損が生じた背景と理由などを検証して，欠損値の対処方法を選択する必要がある．現在ではほとんど用いられてない方法も含めて，欠損値への主な対処方法を表1に示す．

リストワイズ法とペアワイズ法は欠損値を無視して削除する単純な方法である（表1）．欠損値の生起メカニズムがMCARであると仮定できる場合は，欠損値を無視できるため，ペアワイズ法やリストワイズ法による削除であっても，解析結果を大きく歪めない（Rue et al., 2008）．また，MARやMNARの場合でも，サンプルサイズに対して欠損値がごく少数であれば，この方法は有効である．しかし，欠損値がごく少数ではない場合にこの方法を適用すると，検出力の低下が問題となり，さらに，特定の傾向を持つ対象者が削除されてしまうため，バイアスにより結果が歪む（Donders et al., 2006; Patrician, 2002; Rubin, 1987; Schafer & Graham, 2002）．

連続量データや計数データなど量的な変数の場合は，平均値代入法，中央値代入法，LOCF法，Hot-deck法，回帰代入法，確率的回帰代入法などの対処方法がある（表1）．しかし，これらの方法により，単一の値に置き換えてしまう方法は，MCARであっても結果が歪むことが知られている（McKnight et al., 2007; Rubin & Schenker, 1991）．

さらに，検査スコアにおいて，DeCrane et al.（2013）は，欠測した場合に0を代入するとリストワイズ法に比べてサンプルサイズが増加し第2種の過誤が減少する理由で，0の代入を提唱している．ランダム化比較試験において共変量が欠測した場合にindicator法が公衆衛生学や疫学の分野で広く使用されてきた（van Buuren, 2018）．これは欠損値を含む変数において欠損値を補完した上で，欠損の有無を意味するもう1つの2値変数を追加する方法であり，単純な代入方法ではない．

カテゴリカル変数の場合は，中央値や最頻値，あるいはロジスティック回帰モデルや対数線形モデルなどを用いたモデル予測値で欠損値を補完する方法がある．Hot-deck法のように類似した属性の対象者の値で代用する場合もある．また，欠損値を新しい因子水準として追加定義する方法も用いられてきた．例えば，性別の選択肢が「男」と「女」しかない状態で性別を大量に欠測すれば，「無回答」という新しい因子水準を追加し，「男」「女」「無回答」の3群で分析するというアプローチである．

縦断研究など繰り返し測定データの場合は，最も単純なLOCF（last observation carried forward）法（これはlast value carried forward, LVCFともよばれる）やBOCF（baseline observation carried forward）法のほかにも，線形混合モデルの変法であるパターン線形混合モデルによる補完法が用いられる場合もある（Son et al., 2012）．

表1 欠損値の主な対処方法

方法	内容
リストワイズ法	欠損値のあるデータ行を丸ごと除外する方法
ペアワイズ法	相関係数など2変数を扱う場合，2変数のどちらかに欠損値のあるデータのみを除外する方法
平均値代入法	平均値を代入して補完する方法
中央値代入法	中央値を代入して補完する方法
最頻値代入法	（カテゴリカル変数の場合）最頻値を代入して補完する方法
LOCF法	（時系列データの場合）前回の観測値を代入して補完する方法
Hot-deck法	類似した属性の対象者の値を代用して補完する方法（距離関数法やマッチング・パターン法）
回帰代入法	欠損値を除いたデータに基づく回帰モデルで推定して補完する方法
確率的回帰代入法	回帰代入法で推定した値にランダムな誤差を加えて補完する方法
完全情報最尤推定法	欠損値パターンに応じた個別の尤度関数を仮定した最尤推定を行う方法
多重代入法	欠損値に代入したデータセットを複数作成し，各データセットに対して分析を実行し，その結果を統合することにより欠損値を補完する方法

LOCF: last observation carried forward

いずれの補完方法でも，何らかの 1 つの値で代用する場合は，多重代入法（multiple imputation）との対比から，単一代入法（single imputation）とよばれる（Gravesteijn et al., 2021）．単一代入法は，補完値の不確実性を無視した方法であり，過度に標準誤差を縮小し，その使用には問題がある．一方で，予測モデル構築を目的とする場合や大規模データを扱う場合などでは単一代入法が選択肢の 1 つとなる（Gravesteijn et al., 2021）．

補完値を用いるのではなく，観測データに基づく尤度を最大にするモデルパラメータ推定値を直接的に得る方法である完全情報最尤（full information maximum likelihood, FIML）推定法（Newman, 2014）は，多重代入法が普及する前に推奨されていた代表的な方法である．

多重代入法は，量的な変数にも質的な変数にも対応できる．多くの分野において，多重代入法は，最良の方法であると考えられている（van Buuren, 2018）．

以上の方法は，基本的に欠損値の生成メカニズムが MCAR や MAR であると仮定できる場合を想定している．しかし，MNAR であると仮定される場合は，本質的に有効な対処方法はない（Rue et al., 2008）．Joseph et al.（2004）は，そのような場合でも多重代入法でバイアスが軽減されるケースもあると報告しており，Rue et al.（2008）は感度分析による検証を推奨している．

IV. 対処方法の比較

欠損値の対処方法による分析結果の違いを比較するために，実データを用いたデモンストレーションを図 1 に示す．このデータセットは母乳栄養調査の 50 人分のオープンデータ（Dalgaard, 2024）であり，ここでは，データセットに含まれる母親と児の体重を用いて，その関連を検討するために回帰分析を行っている．図 1（a）は，一部を欠損値に置き換える前の完全なデータセットを用いて回帰分析を行った結果である．他の 3 つは，一部を欠損値に置き換えたデータセットを用いて 3 種類の欠損値対処方法をそれぞれ施したのちに，回帰分析を行った結果である．完全なデータセットの結果を正解として参照すると，いずれの対処方法を用いても，回帰係数（β）は，正解から乖離した．最も乖離が大きい

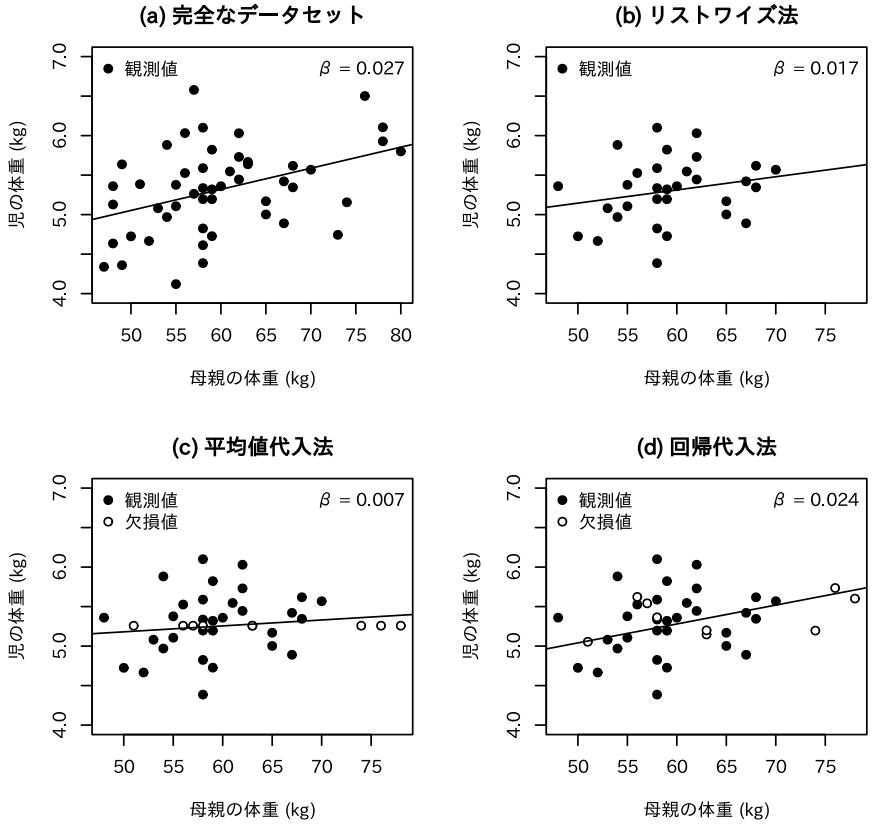

図 1　欠損値の補完結果
完全なデータセットおよび完全なデータセットから 10 個の欠損値を生成したデータセットにおいて回帰分析を実施した．図中の斜線は回帰線であり，β は回帰係数である．
データの出典：Kim Fleischer Michaelsen's data（n = 50）（Dalgaard, 2024）

のは平均値代入法であった．リストワイズ法も，大きな体重の母親が高頻度に欠損し欠損値が全体の 20% である（つまり，MNAR が仮定され欠損割合が大きい）場合に，図 1（b）が示すように，回帰係数は真の値から大きく乖離した．3 つの方法の中では，図 1（d）の回帰代入法が比較的乖離の少ない状態であったが，多重代入法[*2] を用いると β =0.028 になり正解との差異がほぼなくなることを考えると，このデータでは，多重代入法以外を使用すべきではないことは明白である．

V．多重代入法

多重代入法は，基本的にどのような統計解析方法にも適用できる欠損値への対応方法であり，欠損値の予測値を代入した m 組の擬似的な完全データセットを作成するアプローチである（Donders et al., 2006; Rubin, 1987）．

多重代入法の原理を図 2 に示す．単一の値を欠損値に代入するのではなく，欠損値の補完における不確実性を考慮しながら，複数の補完値を準備する方法である．補完値は，補完値を生成する多様なモデル（imputation model）の中から選択されたモデルを使用して生成される．モデル選択について，例えば，臨床経験年数など連続量変数の場合は，線形回帰モデルや予測平均マッチングなどが選択され，性別など 2 値変数の場合は，ロジスティック回帰モデルが選択される．学歴など順序変数の場合は比例オッズモデル，出身地など名義変数の場合は多項ロジスティック回帰モデルが用いられる．

生成された補完値を用いて複数の擬似的完全データセットが用意されたら，次に，それぞれデータセットごとに適切な統計解析を行う（図 2）．解析が終了したら，得られた複数の解析結果を Rubin のルール（Rubin, 1987）を用いて統合化し，全体的な推定値と分散共分散行列を得る（図 2）．最後に，さまざまなシナリオのもとで感度分析を行い，結果が大きく変化しないかを確認する．

擬似的な完全データセット数（m）をどのくらいにすればよいのかという検証が行われてきた．Rubin（1987）は，推定効率の観点から 2 から 10 回でも十分であるとしている．しかし，Bodner（2008）は，$m \leq 10$ の場合に異なる結論が導き出される可能性があることを示し，m 数が大きければ大きいほど，分析の精度が向上することを示した．実際には，必要最低限の m は，欠損割合に依存しており単純な話ではない（Graham et al., 2007）．完全情報最尤推定法との対比により必要な m を見積もった Graham et al.（2007）の報告では，検出力の低下を 1% 未満にしたい場合に，10 から 30% の欠損割合で $m \geq 20$，40% の欠損割合で $m \geq 40$ を推奨している．連鎖方程式による補完を行う場合に White et al.（2011）が提唱するルールは，$m \simeq 100 \times$欠損割合であり，現在，このルールが医学医療の研究分野での標準になっている（van Buuren, 2018）．

多重代入法にはいくつかの限界がある．Rubin（1987）は，他の方法に比べて処理時間が大幅に増加することを欠点としたが，今日の高性能化コンピュータを用いれば問題にならないと考えられる（Graham et al., 2007）．

多重代入法は再現性について限界がある．確率シミュレーションを用いる他の手法と同様に，常に同じ固定された値を得る手法ではないため，再現性について留意しなければならない．しかし，結果の丸め誤差を考慮に入れて疑似的な完全データセット数（m）を慎重に決定すれば，実質的な問題はないと考えられる．

おそらく最も注意を払うべきものは，補完値を生成するモデルの選択に対する妥当性である．この妥当性を担保するために，視覚化による診断などが提唱されて

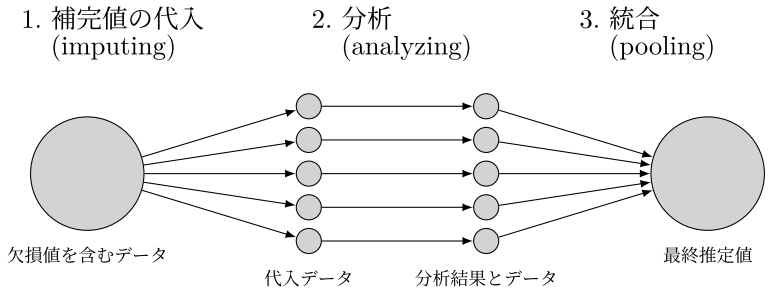

図 2　多重代入法の原理

[*2] 多重代入法は，他の方法と異なり図 1 と同じ方法で示すことができないため，多重代入法を図1に入れていない．

おり（Abayomi et al., 2008; Gelman et al., 2005; van Buuren & Groothuis-Oudshoorn, 2011），評価指標もいくつか提唱されている（van Buuren, 2018）．しなしながら，結局のところ，本当に妥当であるのかを示す手段はない（Gorard, 2020）．さらに，多重代入法は，欠損値の生起メカニズムが MAR であることを前提とした補完値生成モデルを用いる場合が多く，MAR ではない場合に多重代入法を適用するとバイアスが大きくなる（Hughes et al., 2019）ことが報告されている．

以上のような限界が存在するものの，前節の実データを用いたデモンストレーションでは，多重代入法が最良であった．加えて，看護学研究で頻用される 3 つの測定尺度に対して，リストワイズ法，中央値代入法，hot-deck 法，多重代入法をさまざまなシナリオのもとで比較評価した研究でも，多重代入法が最も優れていた（Xu et al., 2020）．さらに，多重代入法を使用しないと第 2 種の過誤が増大する（van Buuren, 2018）ことが知られている．そのため，欠損値への対応方法について，多重代入法を第一選択にするべきであると考えられる．

VI. おわりに

欠損値を含む看護研究データを分析する際には，その欠損値が生じた背景をよく考察し，生起メカニズムに応じた対応が必要である．欠損値の対応について，さまざまな方法が利用されてきたが，どのような統計解析手法にも適用できる多重代入法が最も広く普及している．しかし，すべての統計解析ソフトウェアが多重代入法に対応しているわけではなく，R など先進的な統計解析ソフトウェアを利用することが求められる．看護研究における統計解析のハードルを下げるためには，看護研究教育における統計解析の教育において，欠損値の対処手段の紹介にとどまらず，より具体的で実践的な手順による演習を行い，また論文への記述についての具体的な例示が必要である．その意味で，本稿では，3 つの統計解析手法について，多重代入法の具体的な手順を付録で紹介している．統計学分野は日進月歩であり，多重代入法が改良され，あるいは多重代入法に置き換わる優れた手法が開発される可能性がある．看護研究者は，看護研究における統計解析手法について，定期的な情報の更新が必要である．

文献

Abayomi, K., Gelman, A., & Levy, M. (2008). Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society Series C: Applied Statistics, 57*(3), 273–291. https://doi.org/10.1111/j.1467-9876.2007.00613.x

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal, 15*(4), 651–675. https://doi.org/10.1080/ 10705510802339072

Bolker, B., & Robinson, D. (2022). *broom.mixed: Tidying methods for mixed models* [R package version 0.2.9.4]. https://CRAN.R-project.org/package=broom.mixed

Dalgaard, P. (2024). *ISwR: Introductory statistics with R* [R package version 2.0-9]. https://CRAN.R-project.org/package=ISwR

DeCrane, S. K., Sands, L. P., Young, K. M., DePalma, G., & Leung, J. M. (2013). Impact of missing data on analysis of postoperative cognitive decline (POCD). *Applied Nursing Research, 26*(2), 71–75.

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology, 59*(10), 1087–1091.

El-Masri, M. M., & Fox-Wasylyshyn, S. M. (2005). Missing data: An introductory conceptual overview for the novice researcher. *Canadian Journal of Nursing Research, 37*(4), 156–171.

Fox-Wasylyshyn, S. M., & El-Masri, M. M. (2005). Handling missing data in self-report measures. *Research in Nursing & Health, 28*(6), 488–495. https://doi.org/https://doi.org/10.1002/nur. 20100

Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D. F., & Meulders, M. (2005). Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics, 61*(1), 74–85. https://doi.org/10.1111/j.0006 -341X.2005.031010.x

Gorard, S. (2020). Handling missing data in numeric analyses. *International Journal of Social Research Methodology, 23*(6), 651–660. https://doi.org/10.1080/13645579.2020.1729974

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science, 8*(3), 206–213. https://doi.org/10.1007/s11121-007-0070-9

Gravesteijn, B. Y., Sewalt, C. A., Venema, E., Nieboer, D., Steyerberg, E. W., & CENTER-TBI Collaborators. (2021). Missing data in prediction research: A five-step approach for multiple imputation, illustrated in the CENTER-TBI study. *Journal of Neurotrauma*, 38(13), 1842–1857.

Grove, S. K., & Cipher, D. J. (2019, October 17). *Statistics for nursing research: A workbook for evidence-based practice* (3rd ed.). Elsevier.

Hughes, R. A., Heron, J., Sterne, J. A. C., & Tilling, K. (2019). Accounting for missing data in statistical analyses: Multiple imputation is not always the answer. *International Journal of Epidemiology, 48*(4), 1294–1304. https://doi.org/10.1093/ ije/

dyz032

Joseph, L., Belisle, P., Tamim, H., & Sampalis, J. (2004). Selection bias found in interpreting analyses with missing data for the prehospital index for trauma. *Journal of Clinical Epidemiology, 57*(2), 147–153.

Langkamp, D. L., Lehman, A., & Lemeshow, S. (2010). Techniques for handling missing data in secondary analyses of large surveys. *Academic Pediatrics*, 10(3), 205–210.

Marshall, A., Altman, D. G., & Holder, R. L. (2010). Comparison of imputation methods for handling missing covariate data when fitting a cox proportional hazards model: A resampling study. *BMC Medical Research Methodology, 10*(1), 112. https://doi.org/10.1186/1471-2288-10-112

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. The Guilford Press.

Nassiri, V., Lovik, A., Molenberghs, G., & Verbeke, G. (2018). On using multiple imputation for exploratory factor analysis of incomplete data. *Behavior Research Methods, 50*(2), 501–517.

Nassiri, V., Lovik, A., Molenberghs, G., Verbeke, G., & Busch, T. (2021). *mifa: Multiple imputation for exploratory factor analysis* [R package version 0.2.0]. https://CRAN.R-project.org/package=mifa

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods, 17*(4), 372–411. https://doi.org/10.1177/1094428114548590

Patrician, P. A. (2002). Multiple imputation for missing data. *Research in Nursing & Health*, 25(1), 76–84. https://doi.org/https://doi.org/10.1002/nur.10015

Roberts, M. B., Sullivan, M. C., & Winchester, S. B. (2017). Examining solutions to missing data in longitudinal nursing research. *Journal for Specialists in Pediatric Nursing, 22*(2), e12179. https://doi.org/https://doi.org/10.1111/jspn.12179

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Rubin, D. B., & Schenker, N. (1991). Multiple imputation in health-care databases: An overview and some applications. *Statistics in Medicine, 10*(4), 585–598. https://doi.org/10.1002/sim.4780100410

Rue, T., Thompson, H. J., Rivara, F. P., Mackenzie, E. J., & Jurkovich, G. J. (2008). Managing the common problem of missing data in trauma studies. *Journal of Nursing Scholarship*, 40(4), 373–378. https://doi.org/https://doi.org/10.1111/j.1547-5069. 2008.00252.x

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177.

Son, H., Friedmann, E., & Thomas, S. A. (2012). Application of pattern mixture models to address missing data in longitudinal data analysis using SPSS. *Nursing Research, 61*(3), 195–203.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC Press.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30*(4), 377–399. https://doi.org/10.1002/sim.4067

Xu, X., Xia, L., Zhang, Q., Wu, S., Wu, M., & Liu, H. (2020). The ability of different imputation methods for missing values in mental measurement questionnaires. *BMC Medical Research Methodology, 20*(1), 42. https://doi.org/10.1186/s12874-020-00932-0

要　旨

看護研究データには多くの場合に欠損値が含まれる．しかし，看護研究の統計解析に関する教科書に欠損値への対応方法についての説明が見当たらず，また網羅的に解説した文献も見当たらない．本稿では，看護学研究分野の統計調査データを解析する上で，これまで採用されてきた欠損値の対処方法を網羅的に解説し，加えて，リストワイズ法，平均値代入法，回帰代入法の比較を実データを用いて行った．さらに，近年の代表的な欠損値対処法である多重代入法について，3 種類の統計解析を例示して具体的な手順を詳解した．

キーワード：看護学研究，量的研究，欠損値，多重代入法

付録 A　回帰分析における多重代入法の利用例

架空のデータを用いて重回帰分析における多重代入法の利用例を示す．ここでは重回帰モデル (multiple linear regression) を例に説明するが，他の回帰モデルでも手順はほぼ同じである．まずは，ネットワーク経由で csv ファイルを読み込む

```
fn <- "https://www.medic.mie-u.ac.jp/mnj/data/mnj27-01.csv"
d <- read.csv(fn, stringsAsFactors = TRUE)
```

読み込んだデータフレームを単集計を確認する．

```
> summary(d)
```

```
      ID              age          sex         height          weight           score
Min.   :  1.00   Min.   :19.00   F   :50   Min.   :167.5   Min.   : 66.40   Min.   :1.0
1st Qu.: 25.75   1st Qu.:19.00   M   :45   1st Qu.:177.6   1st Qu.: 76.90   1st Qu.:1.0
Median : 50.50   Median :20.00   NA's: 5   Median :180.6   Median : 81.70   Median :2.5
Mean   : 50.50   Mean   :20.45             Mean   :181.1   Mean   : 82.39   Mean   :2.3
3rd Qu.: 75.25   3rd Qu.:21.25             3rd Qu.:185.3   3rd Qu.: 87.00   3rd Qu.:3.0
Max.   :100.00   Max.   :22.00             Max.   :200.2   Max.   :103.70   Max.   :5.0
                 NA's   :12                NA's   :20      NA's   :7
```

このデータフレームは 6 変数（ID 番号 [ID]，年齢 [age]，性別 [sex]，身長 [height]，体重 [weight]，測定スコア [score]）から構成され，欠損値は年齢が 12 個，性別が 5 個，身長が 20 個，体重が 7 個となっている．このデータを対象に，多重代入法による欠損値の対応を行った上で，目的変数を体重，説明変数を年齢，性別，身長，測定スコアとした重回帰モデルによる分析を行う．まずは最初に，多重代入法の R パッケージである mice パッケージ（van Buuren & Groothuis-Oudshoorn, 2011）を読み込む [*3].

```
> library(mice)
```

次に欠損値を補完した 20 組の擬似的な完全データセットを作成し，dm という名前で保存する．この dm は mids クラスオブジェクトである．

```
> md <- c("", "pmm", "logreg", "pmm", "", "")
> dm <- mice(d, m = 20, method = md, print = FALSE, seed = 1234)
```

補完値の生成方法を指定するには，method オプションを mice() に与える．このオプションで指定する文字列を表 A.1 に示す．1 つの文字列だけを指定する場合は，それが全ての変数に適用される．変数ごとに個別に指定する場合は，例にある通り変数の順番と同じ順番で並んだ文字列ベクトルを指定すればよい．補完しない場合は「""」と空欄文字列を与えればよい．ここでは，method オプションを用いて，年齢は pmm を，性別は logreg を，身長は pmm を指定した（これらの意味については表 A.1 を参照されたい）． method オプションを省略した場合（つまり，補完値の生成方法が無指定の場合）は，生成方法が自動判定される． seed オプションは，デモンストレーションの再現性を担保するために固有の数値を指定しているが，本来は不要なオプションである．

念の為に補完値が生成された方法を確認にする．

```
> dm$method

   ID       age       sex    height    weight     score
   ""     "pmm" "logreg"     "pmm"        ""        ""
```

ここで年齢と身長は予測平均マッチングによる補完値の生成を行った．この方法は，回帰モデルを用いた補完値の生成方法をさらに改良した方法であり，相対的に優れている（Marshall et al., 2010）ことが明らかになっており，広く使用されている．

なお，mice() は補完値の生成を行う際に，対象の変数以外の変数をすべて使用して生成を行うため，変数が数百以上あるような場合は処理時間が問題になり，自由記述の回答が含まれている場合や欠損値が NA ではなくて 9999 などの数値が入っている場合は，正常に補完値の生成ができない（van Buuren, 2018）ことに注意する．

この 20 組の擬似的な完全データセットが入った dm を用いて，重回帰分析を行う．

表 A.1 mice() で利用できる補完値の生成方法 [a]

文字列表記 [b]	変数の型	内容
pmm	問わず	予測平均マッチング
midastouch	問わず	重み付き予測平均マッチング
sample	問わず	観測値から無作為抽出
cart	問わず	CART（決定木分析）
rf	問わず	ランダムフォレスト
mean	連続量	条件づけなしの平均値
norm	連続量	ベイズ線型回帰モデル
lasso.norm	連続量	Lasso 回帰モデル
logreg	二値	ロジスティック回帰モデル
polr	順序	比例オッズモデル
polyreg	名義	多項ロジスティック回帰モデル
lda	名義	線形判別分析

[a] 主な補完値の生成方法を mice() のヘルプから抜粋した
[b] mice() の method オプションに指定する文字列

[*3] 事前に mice パッケージをインストールしておく必要がある．インストールする際には，install.packages("mics") を実行する．

```
> f1 <- with(dm, lm(weight ~ height + age + sex + score))
```

20 組の擬似的な完全データセットをモデルに当てはめた結果（20 回分の重回帰分析の結果）を `f1` に保存した．この 20 組の結果を `pool()` で統合した上で `summary()` で結果を観察する．

```
> summary(pool(f1))
```

```
         term     estimate   std.error  statistic       df       p.value
1 (Intercept) -104.5719622 15.88318197 -6.5838169 66.44571  8.620986e-09
2      height    0.9202228  0.09387192  9.8029611 68.33887  1.162756e-14
3         age    0.9353491  0.54503968  1.7161119 44.23465  9.313829e-02
4        sexM    2.6573845  1.03128373  2.5767734 64.81957  1.225852e-02
5       score    0.0539069  0.38607999  0.1396263 71.76974  8.893464e-01
```

この結果から作表したものを表 A.2 に示す．対比のためにリストワイズ法で欠損値の対応を行った重回帰分析の結果を併記した．

リストワイズ法と多重代入法の結果において，計算方法が異なるため回帰係数が異なるのは当然であるが，検出力の違いから有意水準を 5% としたときに統計的有意性に違いが生じている．このデモデータにおいて，性別は 2 つの方法で異なる結論が導かれる．これまでリストワイズ法で対応してきた過去の研究データにおいても，多重代入法で再解析すれば，この事例にように異なる結論になる可能性がある．

付録 B　因子分析における多重代入法の利用例

多重代入法は主に回帰モデルなど推定値を得る統計解析を想定して開発されてきた．分析結果に質的な情報が含まれる探索的因子分析やクラスター分析の場合は，他の解析方法のように擬似的な完全データセットごとの推定値を集約して 1 つにまとめるアプローチを取ることができない．そこで，Nassiri et al.（2018）は，分析前に擬似的な完全データを統合して探索的因子分析を行う方法を提案している．この方法は，`mifa` パッケージ（Nassiri et al., 2021）に Nassiri 自身が実装している．なお，確証的因子分析（CFA）の場合は，モデルパラメータを推定する方法であるため，回帰モデルと同様に多重代入法を適用できる．

まず最初に，デモンストレーション用のデータをネットワーク経由で取り込み，`d` という名前のデータフレームとして保存する．

表 A.2　リストワイズ法および多重代入法による欠損値の対応を行った重回帰分析の結果

	リストワイズ法		多重代入法	
	回帰係数	p 値	回帰係数	p 値
身長	0.96	< 0.001	0.92	< 0.001
年齢	0.60	0.284	0.90	0.093
性別女性	0.00	参照	0.00	参照
男性	2.06	0.080	2.63	0.013
測定スコア	0.14	0.741	0.01	0.987

```
> url <- "https://www.medic.mie-u.ac.jp/mnj/data/mnj27-02.csv"
> d <- read.csv(url)
> head(d)

  a1 a2 a3 a4 a5 b1 b2 b3 b4 c1 c2 c3
1  4  3  3  3  3  4  3  3  4  5  4  4
2  2  2  1  2  1  5  5  5  5  5  5  5
3  4  3  4 NA  3  4  4  4  3  5  4  4
4  2  2  1  1  2  3  2  3  2  3  2  3
5  5  5  4  5  4  1  0  1  0  2  2  1
6  5  4  5  4  5  5  5  5  5  5  4  4
```

このデータは，質問が 12 項目の 100 人分の欠損値を含むデータである．まずは，リストワイズ法でどの程度が削除されるのが確認する．

```
> nrow(d) - nrow(na.omit(d))

[1] 30
```

このデータを用いて因子分析をおこなう．因子分析の手順として必要な因子数の決定，Bartlett の球面性検定などは省略する．因子数を 3 因子として探索的因子分析を行う．

最初に因子分析に必要な psych パッケージを読み込む．psych パッケージの利用には GPArotation パッケージが必要であるため，psych パッケージと GPArotation パッケージを事前にインストールしておく必要がある．さらに，因子分析で多重代入法を行うパッケージである mifa パッケージも読み込む．

```
> library(mifa)
> library(psych)
```

次に擬似的な完全データセットを 20 組作成し，mi という名前で保存する．

```
> mi <- mifa(data = d, print = FALSE, m = 20)
```

ここで m = 20 を省略すると，mice パッケージの既定値である m = 5 が使用される．

```
> res.im <- fa(mi$cov_combined, n.obs = 100, nfactors = 3,
+              fm = "ml", rotate = "oblimin",
+              scores = "regression")
```

結果を表示する．

```
> res.im

> Factor Analysis using method =  ml
Call: fa(r = mi$cov_combined, nfactors = 3, n.obs = 100,
    rotate = "oblimin", scores = "regression", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix
```

```
     ML1   ML2   ML3   h2   u2 com
a1  0.70 -0.09  0.03 0.48 0.52 1.0
a2  0.67 -0.16  0.00 0.44 0.56 1.1
```

[後略]

記載を省略しているが，因子分析の結果として必要な情報はほぼ表示されている．

APA スタイルで作表した結果を表 B.1 に示す．以上がデータの読み込みから作表までの手順の流れである．ここで示したように多重代入法を導入することで生じる煩雑な操作はほとんどなく，通常の因子分析の手順とほぼ同じように分析を実行できる．

最後に，リストワイズ法の結果と比較するために，リストワイズ法による結果と多重代入法の結果を並べたものを表 B.2 に示す．

2 つの方法を比較した結果，結論に影響するほどの因子負荷量の違いは認められなかったが，最大で 0.12 の差が生じていた．

付録 C　介入研究データにおける多重代入法の利用例

患者に対する看護介入方法や学生に対する看護教育方法の効果を比較する介入研究デザインは看護学研究分野で広く用いられている．介入研究デザインは，アウトカムを繰り返し測定することが多く，この場合のデータを繰り返し測定データという．性別や年齢など対象者の属性の影響を取り除いた介入効果の比較を行う場合は，線形混合モデルが適用される．ここでは，介入 2 群においてアウトカムである尺度を介入前後で 2 回測定したデータおよび属性のデータから構成されるデモンストレーション用のデータを用いて解説する．

最初にネットワーク経由で csv ファイルを読み込む．

表 B.1　多重代入法を用いた因子分析

Variable	ML1	ML2	ML3	h^2	u^2	com
a1	**0.70**	− 0.09	0.03	0.48	0.52	1.04
a2	**0.67**	− 0.16	0.00	0.44	0.56	1.11
a3	**0.75**	0.05	0.12	0.59	0.41	1.06
a4	**0.71**	0.07	− 0.06	0.53	0.47	1.03
a5	**0.70**	0.14	− 0.12	0.56	0.44	1.14
b1	0.07	**0.67**	0.02	0.47	0.53	1.02
b2	− 0.01	**0.70**	− 0.09	0.47	0.53	1.03
b3	0.18	**0.60**	0.05	0.44	0.56	1.19
b4	− 0.19	**0.61**	0.11	0.40	0.60	1.26
c1	− 0.14	0.07	**0.64**	0.45	0.55	1.12
c2	0.07	0.12	**0.52**	0.31	0.69	1.15
c3	0.07	− 0.06	**0.74**	0.54	0.46	1.03

	ML1	ML2	ML3
SS loadings	2.61	1.76	1.3

	ML1	ML2	ML3
ML1	1.00	0.15	− 0.06
ML2	0.15	1.00	0.12
ML3	− 0.06	0.12	1.00

表 B.2 リストワイズ法と多重代入法における因子負荷量の違い

	リストワイズ法			多重代入法		
	ML1	ML2	ML3	ML1	ML2	ML3
a1	0.74	− 0.12	0.09	0.70	− 0.09	0.03
a2	0.63	− 0.04	− 0.00	0.67	− 0.16	0.00
a3	0.81	0.02	0.09	0.75	0.05	0.12
a4	0.76	0.10	− 0.11	0.71	0.07	− 0.06
a5	0.68	0.09	− 0.14	0.70	0.14	− 0.12
b1	0.10	0.68	0.06	0.07	0.67	0.02
b2	− 0.03	0.80	− 0.12	− 0.01	0.70	− 0.09
b3	0.16	0.63	0.10	0.18	0.60	0.05
b4	− 0.16	0.58	0.16	− 0.19	0.61	0.11
c1	− 0.15	0.13	0.58	− 0.14	0.07	0.64
c2	0.09	0.03	0.58	0.07	0.12	0.52
c3	0.01	− 0.06	0.68	0.07	− 0.06	0.74

```
> url <- "https://www.medic.mie-u.ac.jp/mnj/data/mnj27-03.csv"
> d <- read.csv(url, stringsAsFactors = TRUE)
```

データフレーム d の先頭 3 行と末尾 3 行を表示してデータを確認する．

```
> head(d, n = 3)

  id sex age x1 x2 scale type   time
1  1   M  58  2 31    28    A before
2  2   M  58  2 31    22    A before
3  3   M  52  2 25    25    A before

> tail(d, n = 3)

   id sex age x1 x2 scale type  time
38 18   F  46  3 NA    20    B after
39 19   M  58  1 31    25    B after
40 20   M  59  2 32    26    B after
```

このデータは，A 群 10 名，B 群 10 名を 2 回測定した 40 行から構成され，ID 番号 (id)，性別 (sex)，年齢 (age)，属性 1 (x1)，属性 2 (x2)，アウトカムである尺度 (scale)，介入種別 (type)，測定時期 (time) が含まれる．`read.csv()` に `stringsAsFactors = TRUE` を追加しているため，文字列データである `sex` 変数，`type` 変数，`time` 変数は因子型に変換されている．それぞれの因子水準は，M と F，A と B，before と after である．繰り返し測定のデータ構造として long 型と wide 型があるが，データ入力をする際には人間がわかりやすい wide 型が便利であるが，データ分析する場合はコンピュータがわかりやすい long 型にする．long 型のデータとは，測定値を縦に積む並びのデータ構造であり，例えば，今回のデータでは，20 人を 2 回測定しているので，20 行ではなく 40 行のデータになる．

欠損値の状態を確認するために，`summary()` を用いて各変数の単集計を行う．

```
> summary(d)

       id          sex          age              x1
Min.   : 1.00   F   :10   Min.   :46.00   Min.   :1.000
1st Qu.: 5.75   M   :24   1st Qu.:48.00   1st Qu.:2.000
Median :10.50   NA's: 6   Median :52.00   Median :2.000
Mean   :10.50             Mean   :52.89   Mean   :2.056
3rd Qu.:15.25             3rd Qu.:58.00   3rd Qu.:2.000
Max.   :20.00             Max.   :59.00   Max.   :3.000
                          NA's   :4       NA's   :4
      x2            scale       type       time
Min.   :11.00   Min.   :20.00   A:20   after :20
1st Qu.:20.00   1st Qu.:23.00   B:20   before:20
Median :25.00   Median :26.00
Mean   :23.94   Mean   :25.62
3rd Qu.:31.00   3rd Qu.:28.00
Max.   :32.00   Max.   :31.00
NA's   :4
```

sex 変数の欠損値は 6 個，age 変数の欠損値は 4 個，x1 変数の欠損値は 4 個，x2 変数の欠損値は 4 個であり，その他の変数には欠損値がない状態である．

さらなる記述統計やクロス集計などは省略する．最初に必要なパッケージを読み込む [*4]．

```
> library(lmerTest)
> library(mice)
```

次に，mice パッケージの mice() を用いて，20 組の擬似的な完全データセットを作成する．

```
> md <- c("", "logreg", "pmm", "pmm", "pmm", "", "", "")
> dm <- mice(d, m = 20, method = md, print = FALSE, seed = 1234)
```

mice() のオプションについての説明は付録 A を参照されたい．次に，線形混合モデルにデータを当てはめる．

```
> co <- list(type = "contr.sum", time = "contr.sum")
> m1 <- with(dm,
+            lmer(scale ~ type * time + sex + age + x1 + x2 + (1|id),
+                 contrasts = co))
```

介入研究では，群と測定時期の交互作用項が統計的に有意であると，2 群において介入効果に差が認められたと判定する．最後に，pool() を用いて結果を統合し，その最終的な結果を表示する．この際に，mice パッケージ version 3.16.0 では，事前に broom.mixed パッケージ（Bolker & Robinson, 2022）を読み込んでおかないとエラーになるので注意を要する．

[*4] 事前にインストールしておく

```
> library(broom.mixed)
> summary(pool(m1))
```

```
          term    estimate   std.error    statistic       df       p.value
1 (Intercept) 19.63872278 11.75867597  1.670147458 22.16907  1.089508e-01
2       type1  1.62874460  0.67298954  2.420163334 26.79208  2.257473e-02
3       time1  0.67022026  0.09768640  6.860937012 25.69344  2.958757e-07
4        sexM -0.03453519  3.65547469 -0.009447526 19.11218  9.925600e-01
5         age  0.22813472  0.28276370  0.806803415 20.26348  4.291482e-01
6          x1 -0.99906616  1.66636969 -0.599546526 24.21716  5.543770e-01
7          x2 -0.17237243  0.29732679 -0.579740662 19.27018  5.688064e-01
8 type1:time1  0.49333689  0.09642297  5.116383607 26.21528  2.416594e-05
```

多重代入法を適用した上で，交互作用項 type1:time1 が $p < 0.001$ であったことから，性別，年齢，属性 1，属性 2 の影響を取り除いた上で，A 群と B 群の介入効果に有意な差があったと言える．最後に，リストワイズ法と多重代入法の結果を表 C.1 に示す．

リストワイズ法では欠損値によるサンプルサイズ不足のため，x2 変数の計算ができていない．サンプルサイズを確保する意味でも多重代入法は有用である．

表 C.1 線形混合モデルにおけるリストワイズ法と多重代入法の分析結果

	リストワイズ法		多重代入法	
	回帰係数	p 値	回帰係数	p 値
type1	0.31	0.773	1.63	0.023
time1	0.68	< 0.001	0.6	< 0.001
sexM	− 2.77	0.527	− 0.04	0.993
age	0.11	0.638	0.23	0.429
x1	− 0.67	0.803	− 1.00	0.554
x2	—	—	− 0.17	0.569
type1:time1	0.58	<0.001	0.49	< 0.001