新着記事

奥村太一

2017 年度

Depaoli & van de Schoot (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods, 22*(2), 240-261.

ベイズ的手法の利用に関して，

- モデル推定の前
- モデル推定の後，結果解釈の前
- 事前分布の影響の分析
- 結果解釈の後

にチェックすべき項目を 10 個提示したもの。
ウェブサイト[1]より，練習用データとスクリプトなどをダウンロード可能。

van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods, 22*(2), 217-239.

1990 年から 2015 年までに行われた心理学研究へのベイズ的手法の応用（1,579 研究）についてレビューを行ったもの。
ベイズ流の論文は絶対数でも，全体に占める割合でも 2000 年代前半に急激に上昇していた。
対象となった分析では回帰ベースのものが約半数を占めていた。

Hoijtink & Chow (2017). Bayesian hypothesis testing: Editorial to the special issue on Bayesian data analysis. *Psychological Methods, 22*(2), 211-216.

Psychological Methods の 22 巻 2 号は，ベイズ統計学に関する特集の第一集。この第一集では，

1. ベイズ的仮説検定
2. モデル比較
3. 一般的ガイドライン

に関する論文を取り上げる。
なお，第二集では，

1. ベイズ推定
2. ベイズモデリング

[1] http://sarahdepaoli.com/manuscript-files/

を取り上げる。
　このレビューでは，ベイズ的仮説検定の手法として

- ベイズファクター
- 事後予測 p 値

を紹介する。

ベイズファクター（Jeffreys, 1939; 1961）

データは，H_1 を H_2 の何倍支持しているか？
　事前分布は，データにもとづく（情報仮説アプローチ）か主観にもとづく（g-事前分布（の混合））か。
　Kass & Raftery (1995) が入門的。
　ソフトウェア JASP

事後予測 p 値（Rubin, 1984）

H_0 のもとでの事後分布から事後予測分布によって標本統計量を算出し，データから算出された標本統計量の事後予測分布における位置を評価。
　ただし，頻度論における p 値と異なり，H_0 が真であっても，事後予測 p 値は一様分布にならない。

Pek, J., & Flora, D. B. (in press). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods.*

効果量とその信頼区間を報告するよう強く求められていることもあって、心理科学は改革の時にある[2]。
　本研究では、

- 効果量を報告する様々な方法について原理とオススメを提示
- 標準化されていない効果量の解釈と報告について強調

する。
　さらに、

- 一要因分散分析
- カテゴリカルデータ分析
- 線形回帰における交互作用効果
- 単純媒介モデル

について、データをもとにオススメを例示する。

Cohen, B. H. (2017). Why the resistance to statistical innovations?: A comment on Sharpe (2013). *Psychological Methods, 22*(1), 204-210.

量的方法論の研究者が推奨する方法が定着しない理由について、Sharpe (2013) が見逃していること。

- 心理学の中でも諸分野によって研究法は大きく異なる
- 特定の統計改革が受け入れやすいかどうかは分野によって異なる

帰無仮説検定（NHST）を効果量 and/or 信頼区間の報告に置き換える動きについて、

[2] http://dx.doi.org/10.1037/met0000126

- 効果の方向性のみが意味をなす
- 効果量の特定の値を予測する基盤がない（そして大きい標本を取れない）

分野でのインパクトを考える。

結論としては、統計改革論者は NHST に対する批判を一般化しすぎており、異なる研究形態に合わせた提言を行えていない。このことが、NHST を捨て去ることへの抵抗（resistance）を生み出す要因の一部になっている。

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, Published Online

追試の際に、先行研究からサンプルサイズ決定に必要な効果量を求めることについての問題点[3]。

- 不確実性の無視
- 公表バイアスの無視

→ 検定力分析の結果が楽観的になり、実際の検定力は .80 に及ばない。

シミュレーションにより、効果量の不確実性とバイアスを修正する複数の方法で、実際の検定力と意図した検定力の差異を比較。

→ 当初研究の検定力が低い場合は特に、帰無仮説が偽で QRP (questionable research practices) がなくても、当初研究の効果量を用いると再現に失敗しがちであった。

また、効果量の不確実性とバイアスを修正した場合でも、当初研究の検定力が低い場合はあまり改善が見られなかった。

従って、

1. 当初研究の検定力が十分高いこと
2. 追試の検定力が意図した通りのものになるようデザインされていること

が重要であると考えられる。

Tang, Y., Cook, T. D., & Kisbu-Sakarya, Y. (in press). Statistical power for the comparative regression discontinuity design with a nonequivalent comparison group. *Psychological Methods.* [4]

不等価な比較グループにもとづく**比較的回帰分断デザイン** *a comparative regression discontinuity design* (CRD-CG) を提案。

この検定力について、

- 通常の回帰分断デザイン RD よりも高い
- カットオフや割り当て変数の分布の影響を受けにくい
- 処置ユニットが少なくて済む

National Head Start Impact のデータを用いて、RD、CRD-CG、RCT の効率性を数値的に予測。
→ RD よりも CRD-CG の方が予測されたパラメータに近い結果を得た。

[3] http://dx.doi.org/10.1080/00273171.2017.1289361
[4] http://psycnet.apa.org/doi/10.1037/met0000118

Usami, S. (2017). Generalized sample size determination formulas for investigating contextual effects by a three-level random intercept model. *Psychometrika, 82*(1), 133-157. [5]

文脈効果 ***contextual effect*** 例えば、レベル 1 の説明変数を統制した場合にレベル 2 の説明変数が従属変数に及ぼす効果、など[6][7]。本文では、**個人レベル、集団レベルの説明変数の組み合わせが個人レベルの結果変数に及ぼす効果**と定義。

文脈効果を検証するための

- 検定力分析
- 信頼区間幅

にもとづくサンプルサイズ決定の公式を提示。

- 3 レベルのランダム切片モデル
- 各レベルに興味のある説明変数が 1 つずつある

公式に含まれる指標の値が文脈効果の SE に及ぼす影響を検証。
シミュレーションにより明らかになったこと:

1. 公式に則って算出されたサンプルサイズは正負双方のバイアスを持ちうる
2. バイアスは、文脈効果の信頼性、多重共線性、分散の仮定などに左右される

Höhn, H., & Chiu, C. (2017). A procedure for assessing the completeness of the Q-matrix of cognitively diagnostic tests. *Psychometrika, 82*(1), 112-132. [8]

受験者におけるすべての熟達度を特定できていれば、Q 行列は完璧であるというが、多特性を多項目で計るテストでは完璧な Q 行列を得るのは難しい上に、どのような認知診断モデル（cognitive diagnosis model: CDM）で検証するかによって「完璧度合い」は変化しうる（完璧さは行列そのものの性質ではない）。

ある CDM のもとで、Q 行列が完璧か査定する方法を提示する。

McClelland, G. H., Irwin, J. R., Disatnik, D., & Sivan, L. (2017). Multicollinearity is a red herring in the search for moderator variables: A guide to interpreting moderated multiple regression models and a critique of Iacobucci, Schneider, Popovich, and Bakamitsos (2016). *Behavior Research Methods, 49*(1), 394-402. [9]

多重共線性は、調整関係（moderator relationship）を検討する上で厄介なものである。本研究では、Iacobucci et al. の誤りを指摘するとともに、

- 2 段階検定
- 平均センタリング
- spotlighting
- 直交化
- floodlighting

などの調整関係を検証するために用いられる手法について述べる。

多くの先行研究をレビューした上で、調整された回帰分析と結果の報告について推奨される点を提供する。

[5] https://link.springer.com/article/10.1007/s11336-016-9532-y
[6] http://mumu.jpn.ph/forest/computer/2016/03/20/2811/
[7] http://koumurayama.com/koujapanese/mediation.pdf
[8] http://link.springer.com/article/10.1007/s11336-016-9536-7
[9] http://link.springer.com/article/10.3758/s13428-016-0785-2

Bishara, A. J., & Hittner, J. B. (2017). Confidence intervals for correlations when data are not normal. *Behavior Research Methods, 49*(1), 294-309. [10]

正規分布しないデータについて、種々の CI の算出法の頑健性をシミュレーションによって検討。

- フィッシャーの Z 変換では、95% CI は実測信頼水準が 68% 程度に落ちることもある
- フィッシャーの Z 変換は、尖度が 2、歪度の絶対値が 1 あれば不正確になる
- 頑健な方法は、スピアマンの順位相関、順位に基づく逆正規変換（rank-based inverse normal: RIN）、一部のブートストラップ法のみ。

Rights, J. D., & Sterba, S. K. (in press). A framework of R-squared measures for single-level and multilevel regression. *Psychological Methods.* [11]

シングルレベル、マルチレベルの**回帰混合モデル** *regression mixture models* において決定係数 R^2 を報告するための新たな枠組みを提示。

以下の観点から、11 個の R^2 指標を区別し、関連づける。

- 結果の分散は何と研究者がか考えているか
- 分散の予測に貢献するものとして研究者は何を考えているか

説明された分散を、どのようにして本質的に意味のある要因に分割するか例示

Lai, M. H. C., Kwok, O., & Hsiao, Y. (in press). Finite population correction for two-level hierarchical linear models. *Psychological Methods.* [12]

レベル 1 とレベル 2 の固定効果の SE を、有限母集団に合わせて補正する方法を提案。

もし有限母集団であることを考慮しないと、SE が過大評価され、検定力の低下、CI 幅の増大が生じる。

シミュレーションによると、バイアスは

- レベル 2 のサンプルサイズがレベル 2 の母集団サイズの 10%を上回ると顕著
- 旧内相関とともに増加
- 集団の数とともに増加
- 平均的な集団の大きさとともに増加

する。

提案された方法では、クラスター数が 30、クラスターの大きさが 10 あれば不偏な SE となる。

Scherer, E. A., Huang, L., & Shrier, L. A. (2017). Application of correlated time-to-event models to ecological momentary assessment data. *Psychometrika, 82*(1), 223-244. [13]

EMA を通じて、繰り返し測定される構成概念と散発的に起こるイベントとの関連を見たいとき、データは観測された時間、打ち切られた時間が相関したものから構成される。

→ **時間事象分析** *time-to-event analysis*（生存時間分析）[14]を使用する必要がある。

感情状態に関する研究、うつ状態の青年における性行動に関する研究に、proportional hazards, accelerated failure time modeling を適用。

[10] http://link.springer.com/article/10.3758/s13428-016-0702-8
[11] http://psycnet.apa.org/doi/10.1037/met0000139
[12] http://psycnet.apa.org/psycinfo/2017-12025-001/
[13] http://link.springer.com/article/10.1007/s11336-016-9495-z
[14] https://www.slideshare.net/okumurayasuyuki/ss-29212460

Michiels, B., Heyvaert, M., Meulders, A., & Onghena, P. (2017). Confidence intervals for single-case effect size measures based on randomization test inversion. *Behavior Research Methods, 49*(1), 363-381.

帰無仮説が棄却されない範囲を CI とする方法（hypothesis test inversion: HTI）を単一事例実験のランダマイゼーション検定に応用（randomization test inversion: RTI）。

- 完全無作為単一事例実験デザインでの平均値差（非標準化、標準化）
- 他の単一事例実験デザイン

への適用を紹介し、R コードを提供。

Manolov, R., & Patrick, O. (in press). Analyzing data from single-case alternating treatment designs. *Psychological Methods.* [15]

一事例実験では、**処置交替デザイン** *alternating treatment designs (ATDs)* はあまり相対的に重要視されてこなかった。

1. ATDs の方法論的に望ましい特徴を先行研究とあわせてレビュー
2. ATDs データの分析法のオプションを紹介
3. 新たな 2 つの分析方法を提示
4. 従来の方法と提示した方法の応用
5.

Finnigan, K. M., & Vazire, S. (in press). The incremental validity of average state self-reports over global self-reports of personality. *Journal of Personality and Social Psychology*

考え方や感じ方、行動の一般的パターンを自己報告させる、というパーソナリティの測定法 *global self-reports* が本当に良いのかという疑問がある[16]。

Whole Trait Theory（全体特性理論？ Fleeson & Jayawckreme, 2015）によると、**自己報告の反復測定**（多数状態の平均）が、特性レベルを測る上で代わりとなる、そして潜在的に優れた方法。

→ 平均的な状態 *average states* が、パーソナリティの個人差を測る上で妥当なのだろうか?（global self-reports に対する**増分妥当性** *incremental validity* の検証）

結果は、以下の通り。

- average state self-reports とインフォーマント報告との相関は、global self-reports とインフォーマント報告との相関より低い
- global self-reports を統制すると、average state self-reports はインフォーマント報告を有意に予測しない

以上より、以下のことが示唆される。

1. average state self-reports は、パーソナリティの個人差について global self-reports 以上の情報を有していない
2. average state self-reports は、一般に信じられているよりも多くの自己バイアスを含んでいる

→ 日々のパーソナリティの表れや自己報告の正確さに関する研究に対する示唆

[15] http://psycnet.apa.org/doi/10.1037/met0000133
[16] http://dx.doi.org/10.1037

はてな？

EMA のような方法でパーソナリティを把握しようとしても、決して（典型的なパーソナリティの個人差を検出する上での）妥当性は高くならないということか？

はてな？ここまで

Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple Imputation of Missing Data in Multilevel Designs: A Comparison of Different Strategies. *Psychological Methods, 22*(1), 141-165.

不完全なマルチレベルデータに対する多重代入法を比較。

- 代入時にマルチレベル構造を考慮しないと、級内相関係数の推定値にネガティブなバイアスをもたらす
- マルチレベル構造を表すために代入モデルに事後的にダミー変数を含めるのは問題

多変量線形混合モデルにもとづく代入が、シミュレーションで扱ったほとんどの条件下で妥当な結果をもたらす唯一の手法。