２次分析研究

専修大学 国里愛彦

2021年3月1日

自己紹介

- 国里愛彦（くにさとよしひこ）

- 所属：専修大学人間科学部心理学科 准教授
- 専門：計算論的アプローチを用いたうつ・不安の維持と認知行動療法の作用メカニズム研究（計算論的臨床心理学）
- 研究の方法論や枠組みを調べるのが好き，オープンサイエンスも好き
- 研究に関することなら気軽にお声掛けください。連絡は，メールフォームからどうぞ！

新型コロナ禍と心理学研究

- 新型コロナ禍により，大学生を対象のデータ収集はかなり困難になった。
- 実証研究ではデータが重要になるため，心理学では新規にデータ収集する事が多い。
- でも，**新規のデータ収集しないと心理学の研究はできないのだろうか？**

良い研究疑問とは？

FINER(Cummings et al., 2013)

1. Feasible = 実行可能である
2. Interesting = おもしろい／興味深い
3. Novel = 新しく独創的である
4. Ethical = 倫理的である
5. Relevant = 切実である（社会的意義）

新規データ収集が必須なわけではない。過去の研究の蓄積の上に少し上乗せできれば良い！

２次分析とは？

２次分析(Secondary Data Analysis)とは，既存データ（preexisting data）に，**既存の研究とは違う目的（視点）をもって分析する**ことである。

２次分析の目的

1. 新しい理論・仮説・モデルの検証，新しい領域への応用
2. 観察データから因果推論（差の差分析，回帰不連続デザイン，操作変数法，マッチング法）
3. 探索的検討による仮説生成（新たに参入する領域のデータの感覚をつかめる）

既存データの種類(Weston et al., 2019)

1. **大規模調査のデータ**（政府統計，パネル調査）
2. **個別の研究データ**（各研究室のデータ）
3. **ビッグデータ**（ウェブスクレイピング，医療報酬の明細書など）

- 既存データの中にはオープンデータも増えている。
- オープンデータとは，**「自由に使えるデータ」**（庄司, 2014）

→どのようなライセンスの下でデータが公開されているのかが重要になる

＊ライセンスは，長谷川他(印刷中)の「実証的研究の事前登録の現状と実践(https://psyarxiv.com/kvgyc/)」の表１が分かりやすい

オープンデータはどこに？

- 心理学関連オープンデータリスト（国里と有志の方がまとめたもの）
- Open Science Framework
- Elsevier DataSearch
- Google Dataset Search
- DataCite Search
- Data Citation Index (Web of Scienceの一部,有料)

どのデータを使う？

- OSFにあるデータは玉石混交（授業での学生プロジェクトのデータなど）
- **データに関連した情報**（著者，論文），**研究手法**（コードがついているか，ビデオも含む論文以上の詳細なマテリアルがある）などが選択の基準となる(池内, 2019a)。**ライセンス**にも注意。
- **問題設定**（データは存在するがまだ検証されてない問題を探すスキル）と**データセットの選定**（問題設定に合ったデータを選べられる，データの質を評価できる）には，高度な専門知識が求められる。

２次分析の良いところ

- 新たにデータ収集を始めるよりもコストが小さくなり，研究資源の節約につながる。
- 入手が難しいデータを利用できる（数十年に渡る縦断データ，大規模なfMRIデータ,コロナ禍の世界中の17万人から収集したCOVIDiSTRESS Global Survey dataset(Yamada et al., 2021)など）
- 理論・仮説，数理モデル（認知モデリング），解析手法が思いついたら，すぐにデータで検証できる。

２次分析の悪いところ

- 関心のある変数が測定されてないor測定の信頼性に疑問がある（大規模調査の場合は，項目数削減のために，過剰に短縮することも）
- 研究パラサイト（research parasites）という不名誉な呼び方があるように，他者の収集したデータで研究することを低く見る文化がある(論文至上主義と同じくデータ収集至上主義がある)。
- データを見た後の仮説検定はα水準のインフレを制御できないが，２次分析の場合データの中身が既知なこともある。*p*-hackingやHARKingに対処しにくい。

２次分析の透明性を高める

- 2次分析の透明性を高める方法として，事前登録がある。事前登録を学ぶと2次分析研究の手続きについても分かる!
- Mertens & Krypotos(2019)が，10項目のテンプレートを作成（Psychologica Belgica誌，AsPredictedを意識しているため簡潔，Rパッケージのpssrも用意されている）
- Van den Akker et al.(2019)が，25項目のテンプレートを作成 (https://osf.io/zmua4/)
- 以降は，Mertens & Krypotos(2019)のテンプレートと例を中心に説明する（van den Akker et al.(2019)は補足的に説明しリストは付録につけた）

Q1 調べる仮説は何か？

- 簡潔に理論を説明し，可能な限り**正確に仮説を書く**
- *例：弁別恐怖条件づけが不安関連パーソナリティ特性と正の関連を示すことを検証する。*

Q2 仮説の変数をどのように測定する？

- **仮説の変数をどのように測定するのか正確に書く**。ポジティブな結果が得られた変数だけ報告していると疑われたりしないように，できるだけ明確に書く。
- *例：不安関連パーソナリティ特性は，**STAI(Spielberger, Gorsuch, & Luchene, 1970)**の特性版を用いて測定する。*

Q3 分析するデータの入手元はどこか?

- **どういうデータなのか**という情報を提供し，そのデータが以前に出版されているかどうかも明らかにする。
- *例：上記の仮説は，２つの既に出版されているデータセット＜引用＞を用いて検討される。関連するデータセットは，研究者X,Yそして*Z*から入手する*

Q4 どのようにデータを入手するのか？

- **どのようにデータを請求orアクセスするのかを書く。**
- 事前登録がデータにアクセスするための必要条件の場合は，それを書く。
- データを既に入手している場合は，**データセットについて調べたり，分析したことがあるのか明らかにする**
- *例：恐怖条件づけの研究者X,Yそして*Z*に，恐怖条件づけに関するデータセットを共有するように連絡する。他の２つのデータセットは，私達の研究室で得られたものであり，既に出版されている<引用>。その論文では，今回とは関連しないトピックについて扱っている。*

Q5 データの除外基準があるか？

- **除外する可能性のあるデータセット，観測，時点があるかどうか**明らかにする。
- **除外する理由**についても明記する。
- *例：教示による恐怖条件づけにのみ焦点をあてた研究<引用>は，分析から除外する。それは，教示による恐怖条件づけと教示のない恐怖条件づけとでは，学習メカニズムが異なる可能性があるためである<引用>*

Q6 予定している統計的分析は何か？

- データ分析で使用する**統計モデルを明らかにする**。できるだけ明確にして，曖昧な表現を避ける（分散分析するとだけ書かない）
- 除外や外れ値処理などの**前処理過程も記載する**。
- *例：異なるデータセットを結合する。条件づけフェーズの最後の試行のCS+とCS-の皮膚電位反応差と特性不安得点との間のピアソンの相関係数を算出する。皮膚電位反応は，個人間差を説明するために個人の最大反応で割り，データの分布を標準化するために平方根変換をした。*

Q7 仮説の承認・不承認の基準は何か？

- **仮説の評価のための明確な基準を提供する。**帰無仮説検定においてはα値が使われ，ベイジアン仮説検定の場合は仮説のエビデンスの強さを表すベイズファクターなどが使われる
- 統計モデリングの場合は，**モデルフィッティングに使用される規準を明確にする**（BIC,AIC,RMSEAなど）
- *例：弁別恐怖条件づけと特性不安得点間の正の相関の検証における有意水準は 5%とする*

Q8 分析はデータの一部で既に検証されているか？

- **提案した分析が既にデータの一部やシミュレーションデータで検証されているかどうか示す**（どのくらいのデータ，データセットの%，どの変数）。**第三者が評価できるように，データとシンタックスを提供する。**
- *例：計画された統計モデルは，最初に私達の研究室の利用可能なデータセットで評価した(**data.sav, syntax.sps, output.spv**参照)。これらの結果は，私たちの研究室のデータと研究者**X, Y,Z**によって提供されたデータセットで検証する予定である。*

Q9 仮説検証で用いるデータについて何か知っているか？

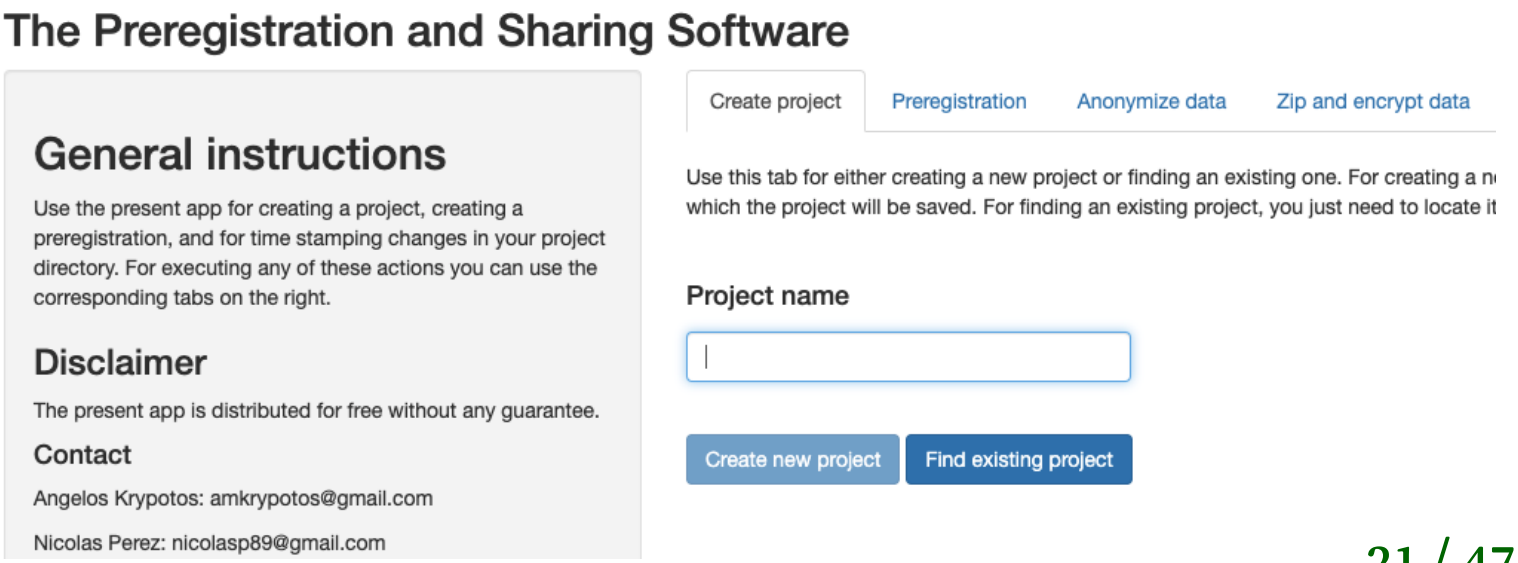
- **研究疑問の検証するデータセットについての事前知識を記載する**（既に出版された論文から変数の平均値を知っているなど）
- 既に知っていることと研究することによって明らかになることを分けられるように書く。
- *例：元の研究で，有意な弁別恐怖条件づけが確立されている。これらのデータセットにおいて特性不安は皮膚電位反応と相関していなかった*

Q10 事前登録における異なるステップの簡潔なタイムラインを示す

- 事前登録のそれぞれのステップの予想される日付(タイムライン)を提供する。
- *例：予定している統計分析のシンタックスは2019年3月15日に完成する。2019年4月1日に研究者X,Y,Zにメールを送り，2019年5月31日まで反応を待つ。2019年7月1日に計画した分析を実行する*

pssrパッケージ

- Krypotos et al.(2019)が開発したRパッケージ。
- (1)研究の事前登録，(2)データの匿名化，(3)データとマテリアル共有のワークフローを提供する。
- 事前登録自体はOSFを使うほうが楽な気もするが，ワークフローや匿名化などが参考になる。



２次分析研究での統計解析

- ２次分析研究は，既に入手可能なデータを扱うこともあり，１次研究よりもハッキングの防止が難しい。
- ２次分析研究でも**事前登録**することで，推論の頑健性が高まる
- ２次分析研究では，**統計解析の頑健性を高める**方法も用いる(Weston et al.,2019)

解析の頑健性を高める(1)

- **データブラインド分析:** 天文学や臨床試験では行われている。データの一部を変更して（ノイズ追加，変数ラベルのシャッフル），分析を実施して解析方法を確定。実際のデータに戻して同じ方法で分析。

- **α水準の調整:** α水準のインフレを抑えるために，保守的なα水準を設定する。偽陽性が高くならないように，多重比較補正などを工夫する。

解析の頑健性を高める(2)

- **クロスバリデーション**: モデルがデータに過剰適合しないよう，データを学習セットとテストセットに分けて，学習セットでモデルに学習をさせてから，テストセットで性能評価する。

- **ホールドアウト**: クロスバリデーションの前に，検証用のデータ（ホールドアウトデータ）を確保しておく方法。モデル学習後にホールドアウトデータで性能評価。

解析の頑健性を高める(3)

- **協調分析(Coordinated analysis)**: 複数の独立した大規模調査データがある時，研究知見の一般化可能性を検討することができる。例えば，異なる国でのコホート研究データをまとめて，サンプルが異なる場合にも同一の効果が得られるのか検討したり，効果の推定値の異質性なども検討できる。

- **探索的データ分析**: 探索的に検討しているのに，あたかも事前に仮説を立てていたかのように書くのが悪い。探索的分析であると明示した上で報告するのが良い（仮説検証が前提となる p値や統計的検定も省くことが推奨される）。

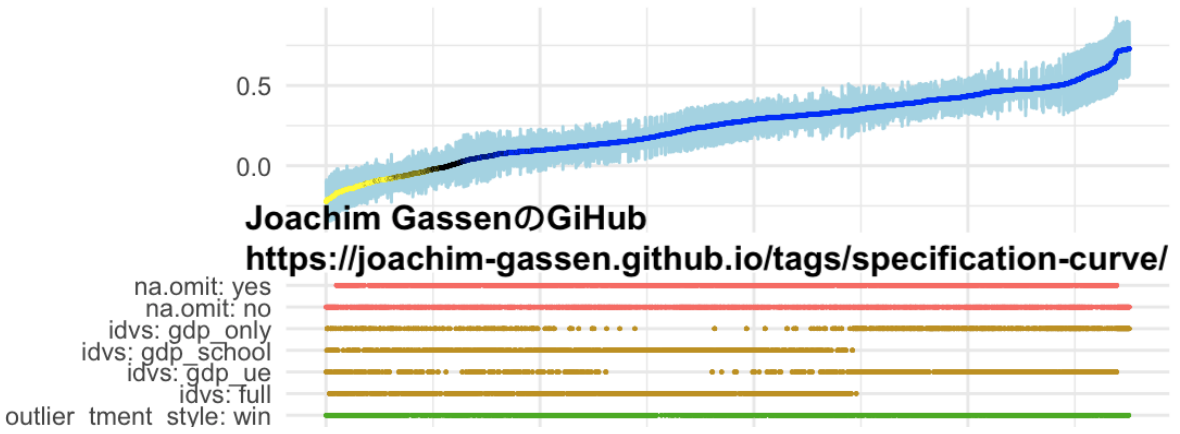
解析の頑健性を高める(4)

- **感度分析**: 分析モデルの一部を変更しても結果に違いがないことを確認する（共変量を追加・除外しても変わらない等）

- **multiverse分析(Steegen et al., 2016)**:　解析データは単一のものではなく，様々な前処理の選択の結果として複数存在しうる多元宇宙(multiverse)と考える。前処理の組み合わせによって作られた多数のデータセットで検定を行って，p値のヒストグラムから検討する（上手に可視化すると前処理において何が結果に影響しているかもわかる）。

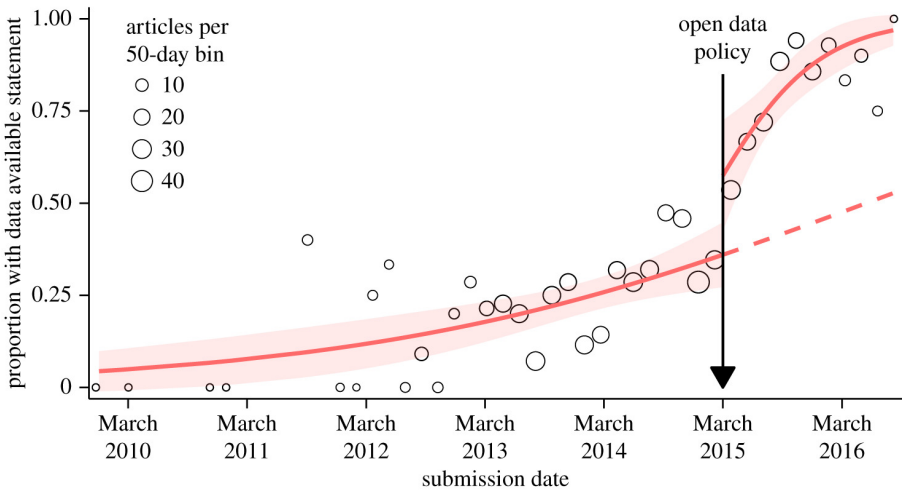
解析の頑健性を高める(5)

- **specification-curve分析（Simonsohn et al., 2020）**：分析する上で決める仕様を網羅的に検討し，解析を実施する。その多数の解析結果を効果の順番で並べ替えて検討する(descriptive specification curvve)。また，その要約統計量も検討する(inferential specification curve:効果の中央値，有意な結果の数など)



オープンデータの増加

- Hardwicke et al.(2018)は，Cognition誌がオープンデータ方針を導入した前後で，論文内でのデータの利用可能性の記載の変化を回帰不連続デザインで検討。明らかに増加。
- 今後，オープンデータはさらに増加すると予想される。



オープンデータのRでの利用

- 国里のウェブサイトにて詳しい解説をしています（https://kunisatolab.github.io/main/how-to-open-data.html ）
- openPsychData: 心理学関連オープンデータをダウンロードするパッケージ（国里自家製野良パッケージ，現状はOpen-Source Psychometrics Projectのデータのみ対応）

```
remotes::install_github("ykunisato/openPsychData")
library(openPsychData)
data <- load_openPsyData(dataset_name = "16PF", codebook = TRUE)
```

オープンデータのRでの利用

- osfr: Open Science Frameworkにアップロードされているデータやコードを**R**に読み込むのに使えるパッケージ

```
library(tidyverse)
library(osfr)
# Yamada et al.(2020)のJapan PVD 2018リポジトリのデータを確認
osf_retrieve_node("qw2af") %>%
  osf_ls_files()
```

- GitHubにアップされているデータやコードをリンクアドレスから**R**に読み込むこともできる。

オープンデータを作る側になる

- データ中心科学においては，**研究成果=データ生成**という側面があり，**研究することがデータのサプライチェーン(三浦, 2018)を回すこと**になる。
- データ共有はまだ一般的でなく，面倒に感じたり，どうやればよいのか分からないという意見も多いが(Houtkoop et al., 2018)，自身の研究データを公開し，データ提供者になることも研究の循環において非常に重要

オープンデータ化については，FAIR(Findability, Accessibility, Interoperability, Reusability; Wilkinson et al.,2016)やAdvances in Methods and Practices in Psychological ScienceのChallenges in Making Data Available(2018,1(1))が参考になる。

オープンデータを作る心理学へ

- 科学技術政策として，2013年のG8サミットから論文とデータのオープン化が推進されている(日本学術会議,2020)。
- 今後は，**オープンデータを作ることを意識して新規データを収集する（参加同意の取得に工夫）**。
- 現状では，データとマテリアルのオープン化は個人の創意工夫で行っている。その実施方法について研究領域レベルでのコンセンサスが必要（機械が読みやすく，ヒトも理解しやすいデータのフォーマットなど）
- **日本人を対象とした研究のマテリアルとデータについては，日本の心理学関連学会が協働して，どのような枠組みで蓄積すべきか検討・提言していく必要があるのでは？**

まとめ

- 既存データを用いた２次分析でも意義のある心理学研究はできる! データ収集の制約がなくなるぶん，自由なテーマ設定もできる。
- ２次分析は，pハッキングなどの不適切な研究実践に繋がってしまう可能性がある。そのため，事前登録や解析の頑健性の検証などが必要になる。
- 未来の２次分析のためにオープンデータの提供者にもなろう！

謝辞

本発表は，JSPS科研費JP20K20870の助成を受けたものです。

引用文献

- 池内有為 (2019a). 研究データの信頼性―データの選択方法と質の向上. 情報の科学と技術, 69(9), 435–437.
- 池内有為 (2019b). 研究データの検索ツール. 情報の科学と技術, 69(6), 256–258.
- Cummings, S. R., Browner, W. S., & Hulley, S. B. (2013). Conceiving the Research Question and Developing the Study Plan. In S. B. Hulley, S. R. Cummings MD, W. S. Browner MD MPH, D. G. Grady MD MPH, & T. B. Newman MD MPH (Eds.), *Designing Clinical Research* (Fourth Edition, pp. 14--22). Lippincott Williams & Wilkins.
- Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Henry Tessler, M., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal Cognition. Royal Society Open Science, 5(8), 180448.
- Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V. M., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data Sharing in Psychology: A Survey on Barriers and Preconditions. Advances in Methods and Practices in Psychological Science, 1(1), 70–85.
- Krypotos, A.-M., Klugkist, I., Mertens, G., & Engelhard, I. M. (2019). A step-by-step guide on preregistration and effective data sharing for psychopathology research. Journal of Abnormal Psychology, 128(6), 517–527.
- Mertens, G., & Krypotos, A.-M. (2019). Preregistration of Analyses of Preexisting Data. Psychologica Belgica, 59(1), 338–352.
- オープンサイエンスの深化と推進に関する検討委員会 (2020). オープンサイエンスの深化と推進に向けて　日本学術会議

引用文献

- 庄司昌彦(2014)オープンデータの定義・目的・最新の課題　智場119号特集号オープンデータ　国際大学GLOCOM
- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. Nature Human Behaviour, 1–7.
- Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 11(5), 702–712.
- Stewart, D. W. (2012). Secondary analysis and archival research: Using data collected by others. In APA handbook of research methods in psychology, Vol 3: Data analysis and research publication, pp. 473–484.
- Van den Akker, O., Weston, S. J., Campbell, L., Chopik, W. J., Damian, R. I., Davis-Kean, P., Hall, A. N., Kosie, J. E., Kruse, E. T., Olsen, J., Ritchie, S. J., Valentine, K. D., van 't Veer, A. E., & Bakker, M. (2019). Preregistration of secondary data analysis: A template and tutorial. https://doi.org/10.31234/osf.io/hvfmr
- Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2019). Recommendations for Increasing the Transparency of Analysis of Preexisting Data Sets. Advances in Methods and Practices in Psychological Science, 2(3), 214–227.

引用文献

- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ⋯ Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 160018.
- Yamada Y, Xu H and Sasaki K. A dataset for the perceived vulnerability to disease scale in Japan before the spread of COVID-19 [version 1; peer review: awaiting peer review]. F1000Research 2020, 9:334 (https://doi.org/10.12688/f1000research.23713.1)
- Yamada, Y., Ćepulić, DB., Coll-Martín, T. et al. COVIDiSTRESS Global Survey dataset on psychological and behavioural consequences of the COVID-19 outbreak. Sci Data 8, 3 (2021). https://doi.org/10.1038/s41597-020-00784-9

付録

- Mertens & Krypotos(2019)のプレレジ案
- van den Akker et al.(2019)のプレレジ案
- Krypotos et al.(2019)の6ステップ・チェックリスト

Mertens & Krypotos(2019)のプレレジ案

- Q1 調べる仮説は何か？
- Q2 必須な変数をどのように測定するのか？
- Q3 分析するデータの入手元はどこか?
- Q4 どのようにこのデータを入手するのか？
- Q5 データの除外基準があるか？
- Q6 予定している統計的分析は何か？
- Q7 仮説の承認・不承認の基準は何か？
- Q8 データの一部において既に検証された分析があるか？
- Q9 検証する仮説に関連づけられるデータについて何かしっているか？
- Q10 事前登録における異なるステップの簡潔なタイムラインを示す

van den Akker et al.(2019)のプレレジ案

- Q1 研究タイトル
- Q2 著者名
- Q3 研究疑問（「AとBは関係するか？」など，シンプルかつ具体的な疑問にする。根拠も必要）
- Q4 研究疑問に対応した検証可能な仮説（方向性のある仮説かそうじゃないかを明示）
- Q5 データ名と内容（研究デザイン，測定項目，対象者など）
- Q6 公開されているデータかどうか（オープンデータか，許可を得てアクセスするものか）
- Q7 どのようにデータにアクセスできるか（DOI, URL, 入手方法など）
- Q8 ダウンロードやアクセスした日（データが追加されていくものなら，いつDLしたかが重要）
- Q9 データ収集方法（サンプルの特徴，どうやって収集？）
- Q10 コードブック（あるならURL。codebookなどのRパッケージ）

van den Akker et al.(2019)のプレレジ案

- Q11 実験操作（介入群など，操作方法を詳細に）
- Q12 使用する変数（測定や得点化など詳細に）
- Q13 適格・除外基準とサンプルサイズ（事前登録段階で分からない場合は保守的な推定値）
- Q14 欠測値とその対処，欠測を除外したサンプルサイズ（分からない場合は推定値）
- Q15 外れ値とその対処，対処後のサンプルサイズ
- Q16 サンプリング・ウェイト（データから重み付けをしてサンプリングする場合）
- Q17 論文に関わる著者全員が過去に当該データを用いて行った発表をリストにする（過去にデータセットを使用した履歴の情報になるため。サラミ出版を防ぐことにもつながる）
- Q18 各著者のデータについての予備知識（要約統計，分布，相関係数，データ操作経験，当該データを使った論文などを読んだ経験など。）

van den Akker et al.(2019)のプレレジ案

- Q19 仮説の検討で使う統計モデル（ここに記載していない分析は探索的分析に。コードも含める）
- Q20 予想した効果量（予備研究やメタ分析などを参考に）
- Q21 検定力分析（予想した効果量，欠測・外れ値除外後のサンプルサイズを使用する）
- Q22 統計的推論で用いる基準（効果量，信頼区間，p 値，ベイズファクター，適合度指標など。片側or両側，多重比較補正）
- Q23 統計モデルの仮定の違反，モデルが収束しないなどの解析の問題が生じた時の対処（天井効果があった場合，正規分布しなかった場合などの工夫を書いておく。恣意的な運用をしないように自由度を下げておく）
- Q24 統計的検定の強さ・信頼性・頑健性の報告（研究内追試，共変量追加による感度分析，クロスバリデーション，ウェイトの使用，SEMの場合に制約の追加，過学習を防ぐ方法，シミュレーション・サンプリング・ブートストラップなど）

van den Akker et al.(2019)のプレレジ案

- Q25 探索的検討を予定している場合に記載する

Krypotos et al.(2019)のチェックリスト

ステップ１:研究疑問と予測を決める

- 確証的な仮説
- 予測
- 探索的な仮説
- 予測

ステップ2:データ収集前に方法と解析プランを決める

- 方法
- 刺激
- 手続き
- プロトコル

Krypotos et al.(2019)のチェックリスト

ステップ2:データ収集前に方法と解析プランを決める(続き)

- 従属変数
- 独立変数
- 統計分析
- 従属変数名
- 独立変数名
- 用いる統計的検定のタイプ
- データ削減
- 頻度論分析：α水準，検定力，予想される効果量を決める
- ベイジアン分析：事前分布，予想されるエビデンスの強さを決める
- モデル選択：モデルパラメータ，比較規準，必要なら事前分布を決める
- 相関分析：予想される相関係数，α水準，検定力を決める
- 分析スクリプトの作成

Krypotos et al.(2019)のチェックリスト

ステップ3:マテリアルの収集

- 情報を載せたパンフレット，インフォームド・コンセント
- 研究プロトコル
- 実験課題と質問紙
- すべてのマテリアルのライセンス

ステップ4:パイロット研究

- パイロット研究の実施
- 現在の研究プロトコルの修正

Krypotos et al.(2019)のチェックリスト

ステップ5:事前登録

- ウェブサイトでの事前登録（例えば，osf.io, aspredicted.org）
- 事前登録プロジェクトのタイムスタンプ

ステップ6:データのアップロードと結果の報告

- データの匿名化