再現性問題で研究がどう変わったか

竹澤正哲

北海道大学文学研究科

社会科学実験研究センター

人間知×脳×AI研究教育センター

日本社会心理学会春の方法論セミナー (2022/3/19)

Perspectives on Psychological Science

Editorial: *Perspectives on Psychological Science*—A Key Journal to Foster the Quality of Research

Perspectives on Psychological Science
2022, Vol. 17(1) 3–5

DOI: 10.1177/17456916211066918
www.psychologicalscience.org/PPS


Scientists must go beyond mere formulaic uncritical compliance with formal guidelines and critically assess what they are doing, what their designs and data imply, and how their work contributes to higher level progress in science.

再現性問題の解決法について思うこと

- 例数設計、プレレジ、追試の奨励、出版バイアスの排除など、概ね解決法は出揃ってきた
- だが「これだけやっていれば大丈夫」「この手法で行われた研究だから信頼できる」と、手法の慣習化が進んだら、元の木阿弥ではないか
- たとえば本来は有益な道具だったはずのp値は、盲目かつ慣習的に利用されたことで問題を引き起こしたのと同様に

なぜ手法によって
結果が正当化されないのか？

理論の重要性①

- QRPが排除された世界について想像してみよう
  - 出版バイアスが抑制され、検定力が高い研究が行われ、プレレジの浸透にり、第一種の過誤の発生率が低い世界
- だがこのような理想的な世界においても、研究者の思いつく仮説が誤ったものばかりであるならば、すべてが台無しとなる可能性が高い（McElreath & Smaldino, 2015)

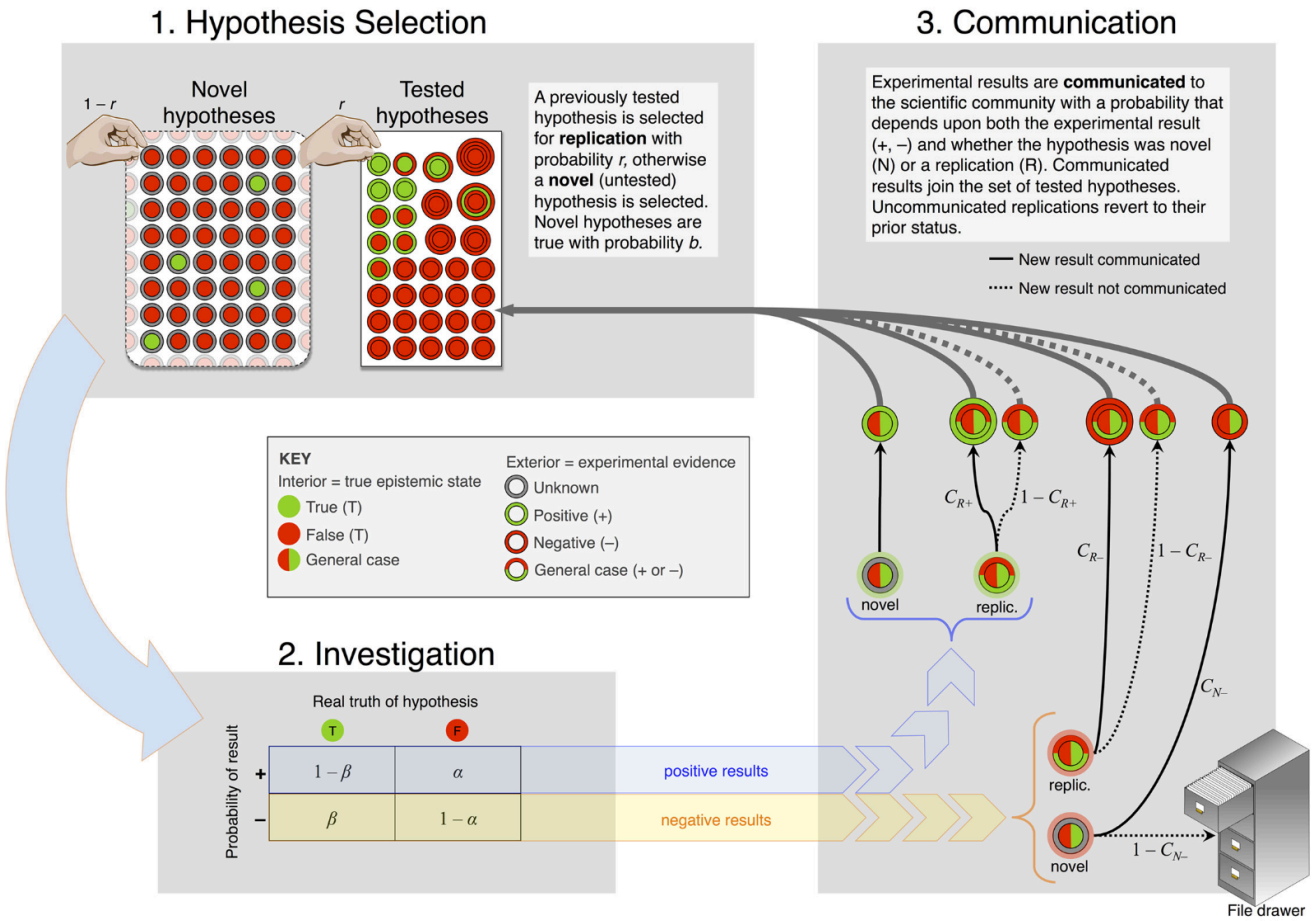


Fig 1. Population dynamics of replication.

doi:10.1371/journal.pone.0136088.g001

「検証される仮説が間違っている確率」を下げない限り、いくらQRPを排除して追試を繰り返しても、本当は存在しない結果を掲載する論文が蔓延する

McElreath, R., Smaldino, P. (2015). Replication, Communication, and the Population Dynamics of Scientific Discovery. *Plos One* 10, e0136088.

理論の重要性②

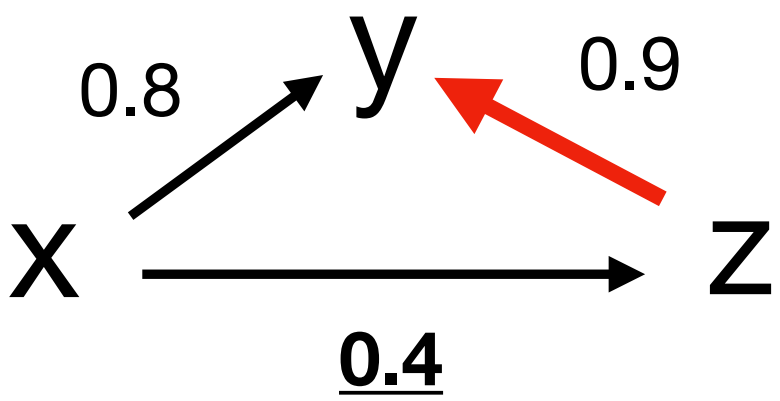
- 変数XがZの原因であることは分かっている
- XやZと関連していそうな変数を統制して、XからZに対する直接効果の強さを推定したい

$$Z = a + b_X X + b_Y Y$$

分析の結果、$b_X = -0.20$

→「XはZに対して負の効果を持っていると結論」

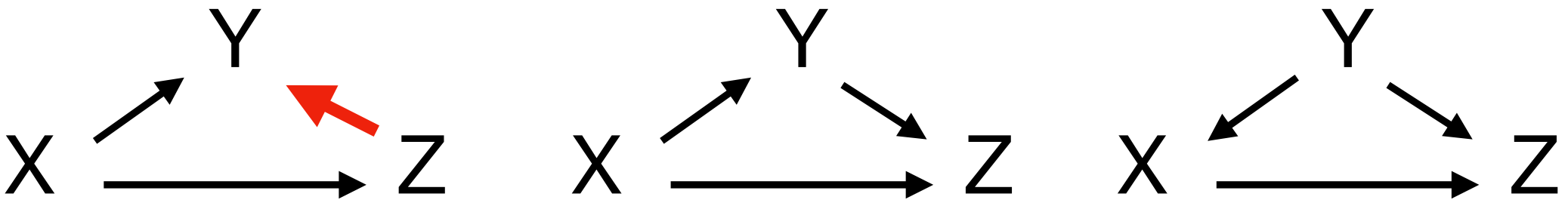
だが真実は異なっていた



$$Z = 0.4X + Normal(0,1)$$
$$Y = 0.8X + 0.9Z + Normal(0,1)$$

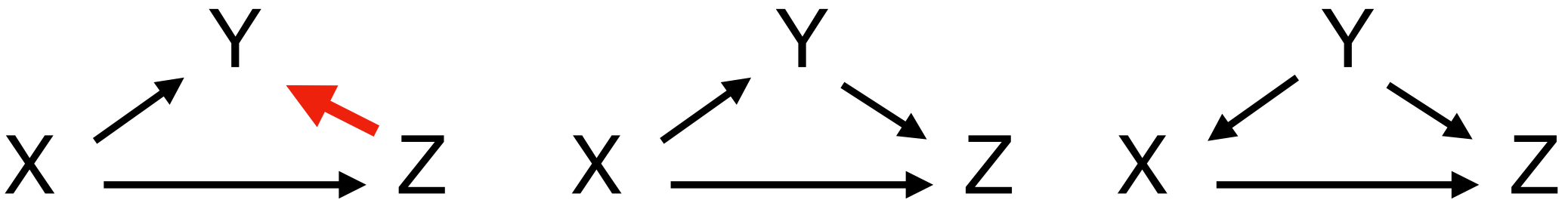
- データはこのような因果プロセスから生成されていた
- XはZに対して本当は正の因果効果を持っていた
- 重回帰分析によって誤った結果が得られてしまった

統計的因果推論と合流点バイアス



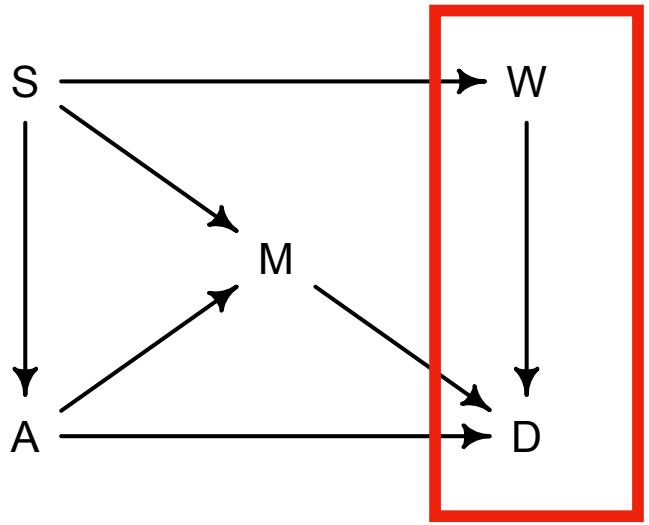
- 変数間の因果関係が中央や右だったら、Z ~ X+Yという重回帰モデルでXの効果を正しく推定できる
- だが、左のような関係だったらYを予測変数に入れてはならない

統計的因果推論と合流点バイアス



- 真の因果は左のはずなのに、「この世界における因果関係は中央 or 右だろう」と誤った理論を研究者が持っていたら、誤った分析をして、誤った結論を導いてしまう（c.f. 吉田・村井, 2021）

統計的因果推論に基づく理想的な重回帰分析



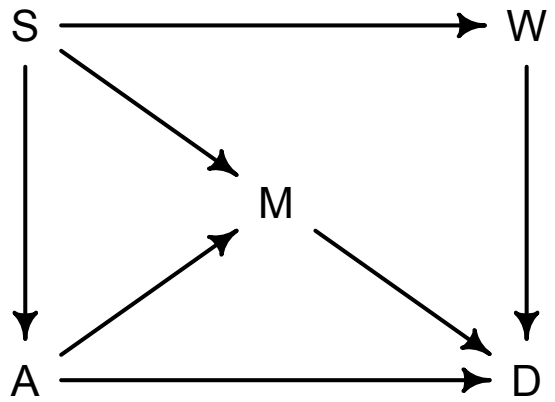
McElreath (2020). Statistical Rethinking 2nd ed., p. 187

```
library(dagitty)
dag_6.2 <- dagitty( "dag {
    A -> D
    A -> M -> D
    A <- S -> M
    S -> W -> D
}")
adjustmentSets( dag_6.2 , exposure="W" , outcome="D" )
```

```
{ A, M } { S }
```

- 研究者が持つ知識に基づいて変数間の因果関係を図示する
- バックドア基準に基づいて、統制すべき変数、統制すべきでない変数を決定する(e.g., Rのdagittyパッケージ)

統計的因果推論に基づく理想的な重回帰分析



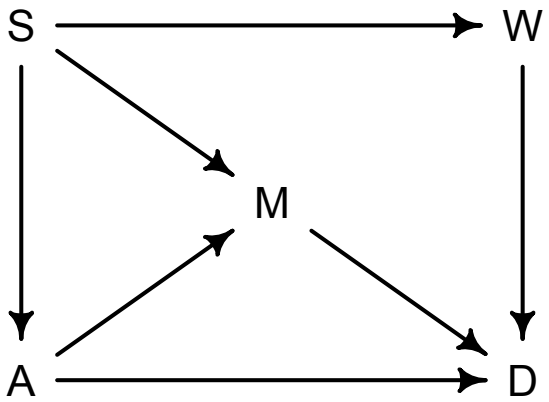
McElreath (2020). Statistical Rethinking 2nd ed., p. 187

```
library(dagitty)
dag_6.2 <- dagitty( "dag {
    A -> D
    A -> M -> D
    A <- S -> M
    S -> W -> D
}")
adjustmentSets( dag_6.2 , exposure="W" , outcome="D" )
```

```
 { A, M } { S }
```

- 当然、得られた結果 ($b_{w \to d}$) の妥当性は、研究者が作る因果ダイアグラムの妥当性に依存する

統計的因果推論に基づく理想的な重回帰分析



McElreath (2020). Statistical Rethinking 2nd ed., p. 187

```
library(dagitty)
dag_6.2 <- dagitty( "dag {
    A -> D
    A -> M -> D
    A <- S -> M
    S -> W -> D
}")
adjustmentSets( dag_6.2 , exposure="W" , outcome="D" )
```

```
 { A, M } { S }
```

- 故に、明確な理論なく、無闇に予測変数を投入した分析の結果は、それがいかなる条件下で妥当なのかすら不明

そんな面倒なことやる必要があるのか？

- 2019年に、マックス・プランク進化人類学研究所に滞在して驚愕した
- フィールドワークに出かける院生に、測定する変数を列挙させ、DAGを描かせ、バックドアが閉ざされているか確認をさせ、さらにダミーデータを生成してモデルリカバリーをさせていた
- フィールドワークなので、全て予定通りに観測することなどできないとしても、研究の前提について徹底的に考えさせていた

なぜ手法の前に理論が重要なのか？

1. 研究者が誤った仮説ばかりを思いつく領域では、検定力を上げ、第一種の過誤を抑制する手法が浸透しても、誤った科学的知識が広まってしまう
2. 本質的に統計分析の結果の妥当性は、研究者が置く前提の妥当性に依拠している

→前提は、必ずしも実データによって検証できるとは限らない

→事前に持っている理論が誤っていたら、正しく分析をしたつもりでも誤った結論が導かれる

APS SPOTLIGHT

Perspectives Editor Klaus Fiedler to Spotlight Pluralism, Theory, "Best Practice" Examplars

APS Fellow Fiedler is APS's first journal editor in chief based at an institution outside North America.

December 29, 2021

As a rule, I believe that statistical methods are subordinate to research design, which is subordinate to logic of science. (…) Likewise, even the most elaborate or compact research design is useless if the theory guiding an investigation is ill-defined, imprecise, or insensitive to prior knowledge.

Japanese Psychological Review
2018, Vol. 61, No. 1, 42–54

心理学におけるモデリングの必要性

竹　澤　正　哲

北海道大学

前半まとめ：再現性問題が教えてくれたこと

- 世界は不確実性に満ち溢れていて「確実」に真実へ辿り着く手法は存在しない
- どれだけ確からしく思える知見も、常に覆る可能性がある
- 当たり前の話だが、この前提を忘れて、「手法によって知見の確からしさ」が担保されると考えてきたことが、再現性問題の根源ではないか
- 「これだけやっていれば大丈夫」を求めるのでなく、ケースごとに自分の頭で考えていくしかない

目指すべき方向性

- ケース・バイ・ケースと言っても、目指すべき方向性ははっきりとしている（村山のトークと共通）

1. データ収集前のアイデア・デザインに時間を費やす（プレレジを含む）
2. 例数設計、大サンプル、汎化性を目指す

- すべての研究で達成できないとしても、できなければできないほど研究の信頼性は下がると考えるべきだろう

だが...

- そう偉そうに言いながらも、集団実験をしていると、思うように目指すべき場所に近づけず、苦戦している
- 相馬先生から打診された時、現在の苦闘の様子を曝け出すことが価値を持つと考え、スピーカーを引き受けた
- お手本となるきれいな話ではないが、七転八倒ぶりから学んでいただけることがあると思っている

集団研究における
ベスト・プラクティス

nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-018-0518-x

(2019)

Social learning strategies regulate the wisdom and madness of interactive crowds

Wataru Toyokawa[1,2,3*], Andrew Whalen[1] and Kevin N. Laland[1]

- 計算論モデルを使ったモデル・リカバリー＋例数設計
- クラウド・ソーシングによって最大27人の集団実験（n=699)を実施

Toyokawa et al. (2019)における研究の手順

1. 個人の意思決定プロセスをモデル化
2. 実験データ（集団レベルで生じる帰結）をシミュレート
3. 架空の実験データにモデルをフィッティングし、リカバリーできることを確認（例数設計）
4. 実験を実施
5. 分析、モデルフィッティング

なぜ、手本通りの研究が難しいのか

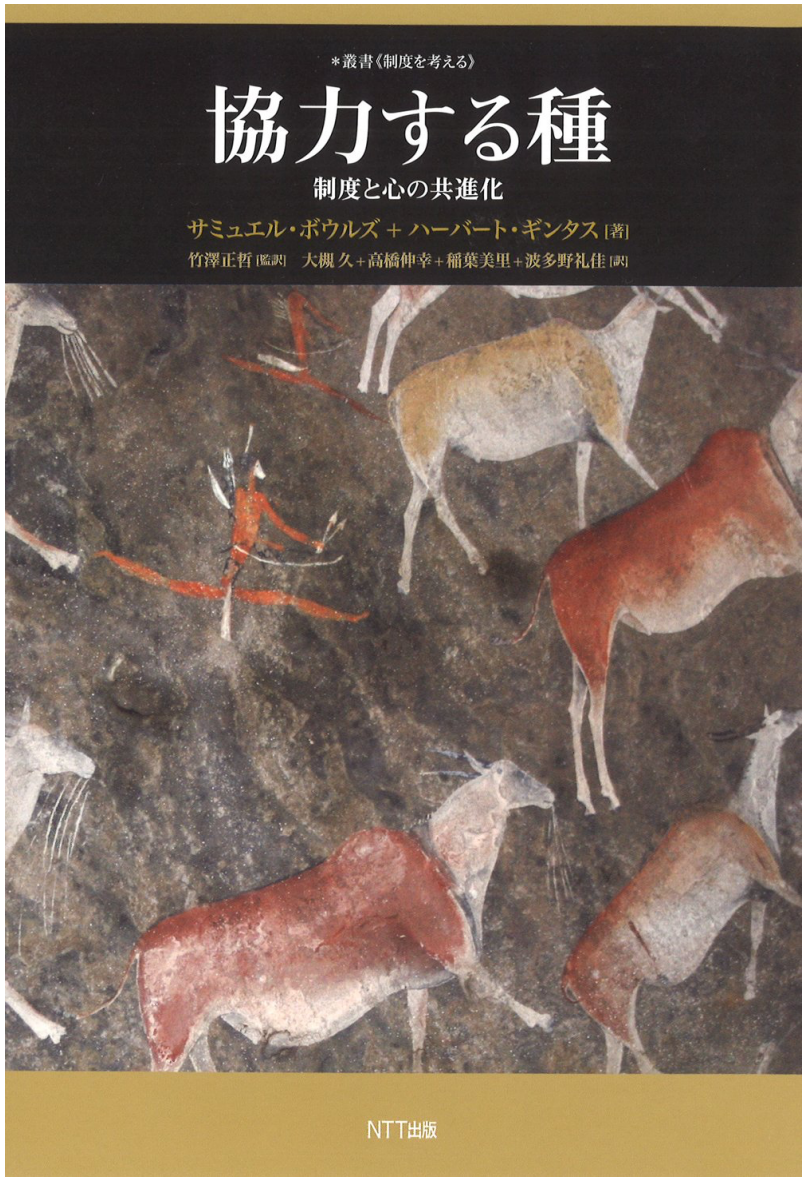
- おそらく多くの心理学者は、教育（学部・大学院）と研究を組み合わせて研究のサイクルを作ってきた
  - 授業とデータ収集の一体化
  - 卒論・修論、学会発表のスケジュールと合わせたデータ収集
- 数ヶ月〜半年を単位とした研究スケジュールにどう組み込んでいくか？

具体的な取り組みの紹介

26

社会に利益をもたらす制度・規範の進化

- 19~20世紀半ばの社会科学における機能主義：人間社会にみられる制度・システムは「集団全体の利益を増加させる」という原理で説明できる
- 記述的には正しく見えるため、大きな影響力を持っていたが、勢力を失う→方法論的個人主義へのシフト（Elster, 1982）
- 文化的集団淘汰理論（Cultural Group Selection Theory）は、人間が持つ社会的学習能力から、機能的な制度・規範が創発することを説明する


＊叢書《制度を考える》
協力する種
制度と心の共進化
サミュエル・ボウルズ＋ハーバート・ギンタス［著］
竹澤正哲［監訳］　大槻久＋高橋伸幸＋稲葉美里＋波多野礼佳［訳］
NTT出版


文化が
ヒトを
進化させた
ジョセフ・ヘンリック
今西康子［訳］
人類の繁栄と
〈文化―遺伝子革命〉
THE SECRET
OF OUR SUCCESS
HOW CULTURE IS DRIVING HUMAN EVOLUTION,
DOMESTICATING OUR SPECIES, AND MAKING US SMARTER
JOSEPH HENRICH
白揚社

理論通りに集団に利益をもたらす戦略が進化するのか？

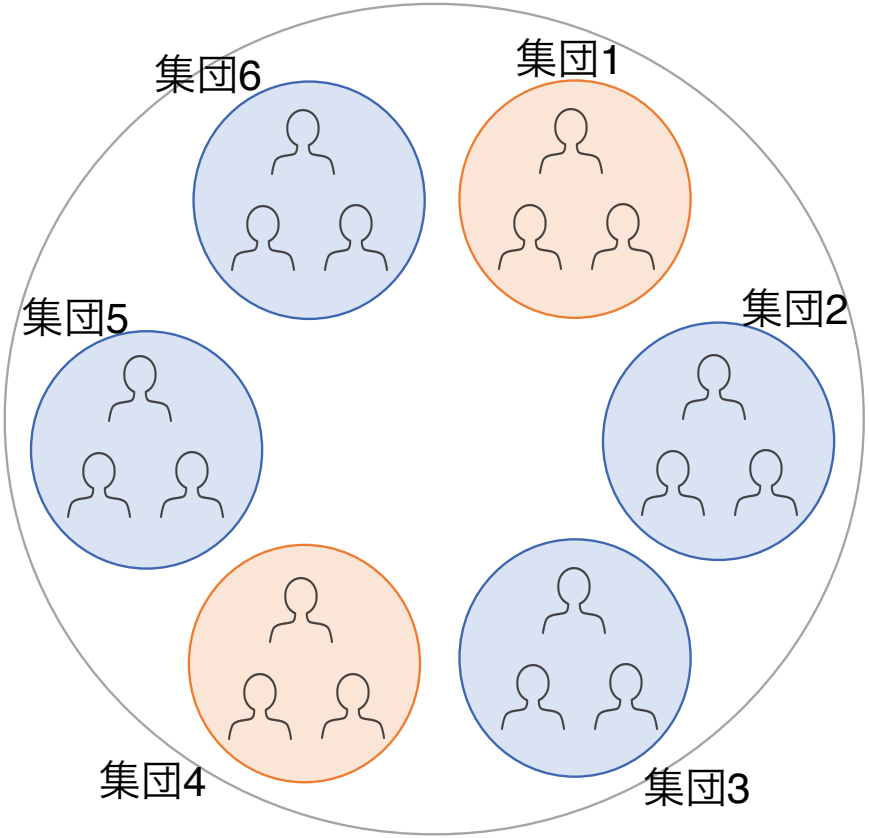
- 文化的集団淘汰理論は、複雑な数理モデルばかりが先行している
- 『集団淘汰』という言葉に惑わされて、食わず嫌いのまま誤解を重ねる研究者も多く、実証研究が少ない
- 理論から予測される通り、成功者模倣バイアスという仕組みによって、集団に利益をもたらす戦略が社会全体に拡散するか検証を試みた

2019年4月
実験デザインの決定

実験の概要

- 18人の参加者が、ランダムに6集団に割り当てられた
- 3人集団でスタグハントゲームが行われた
- ゲームは60試行繰り返され、15試行ごとに移住が発生した

本実験の集団構造



だがそんな綺麗には話は進まない

- 研究がスタートした時点では、集団サイズも、集団の数も決まっていない
  - 実験プログラムはoTree
  - 学部4年生の牧野さんが実験を行う
  - プログラム作成のアドバイスを院生の土田修平が行う
- 牧野さんは就活を進めながら、Python、oTreeを学び、プログラムを作成し、実験準備が終わったのは同年秋だった

2019年秋
実験実施

- 1セッション18人（3人×6集団）という実験デザイン、 2条件×各10セッション（360人） というサンプル数を、実験資源の制約から決定
- 北海道大学社会科学実験研究センターの参加者プール＋実験室で実施
- 実験を開始すると、7セッション終了した時点で、参加者が集まりにくくなったため、実験が終了
- 牧野さんは、卒業論文を執筆して終了

2020年春

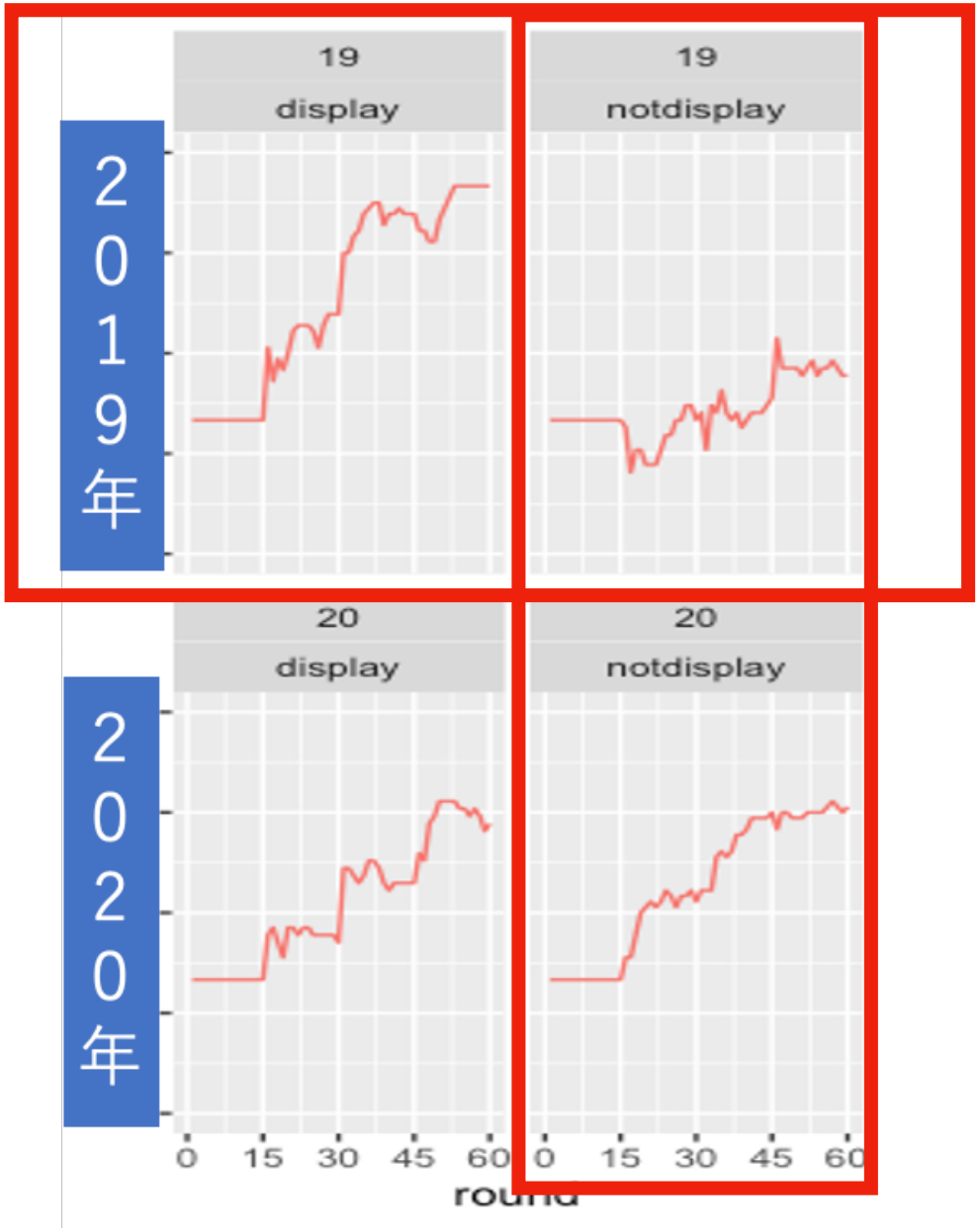
- 昨年は、合計20セッション（n=340）を目標としていたが、7セッション（n=126)しかデータが集まらなかった。
- この時点で有意な効果が得られていたが、当初の予定通りのサンプル数に達するまで実験続行を決定
- 学部4年生の阿部紗采さんが実験を実施することに
- だが北海道ではCOVID-19が蔓延しており、大学への立ち入り制限が続く…

2020年秋

- 埒が明かないので、参加者に自宅からオンライン実験に参加してもらって実験実施
- 社会科学実験研究センターのプールを利用して参加者をリクルート＋アマゾンギフトカードで謝礼
- 1人で参加していることを確認するために、zoomで参加者の様子を見ながら実験実施
- 実験プログラムはherokuにアップロードし、urlを個別にメール送信→自宅PCからアクセス

- 7セッション実施した時点で、参加者の集まりが悪くなったため終了
- 阿部さんは計14セッション（n=252）のデータを分析して、卒論執筆
- だが分析において、予期せぬ現象が見つかった

- 2019年度（上段）は予測通りの結果だったが、2020年度は統制条件（下段・右）で予期せぬ増加が観察された
- 例数設計もなく、実験状況も異なるので、このままだったら『良く分からない』で終わってしまう…

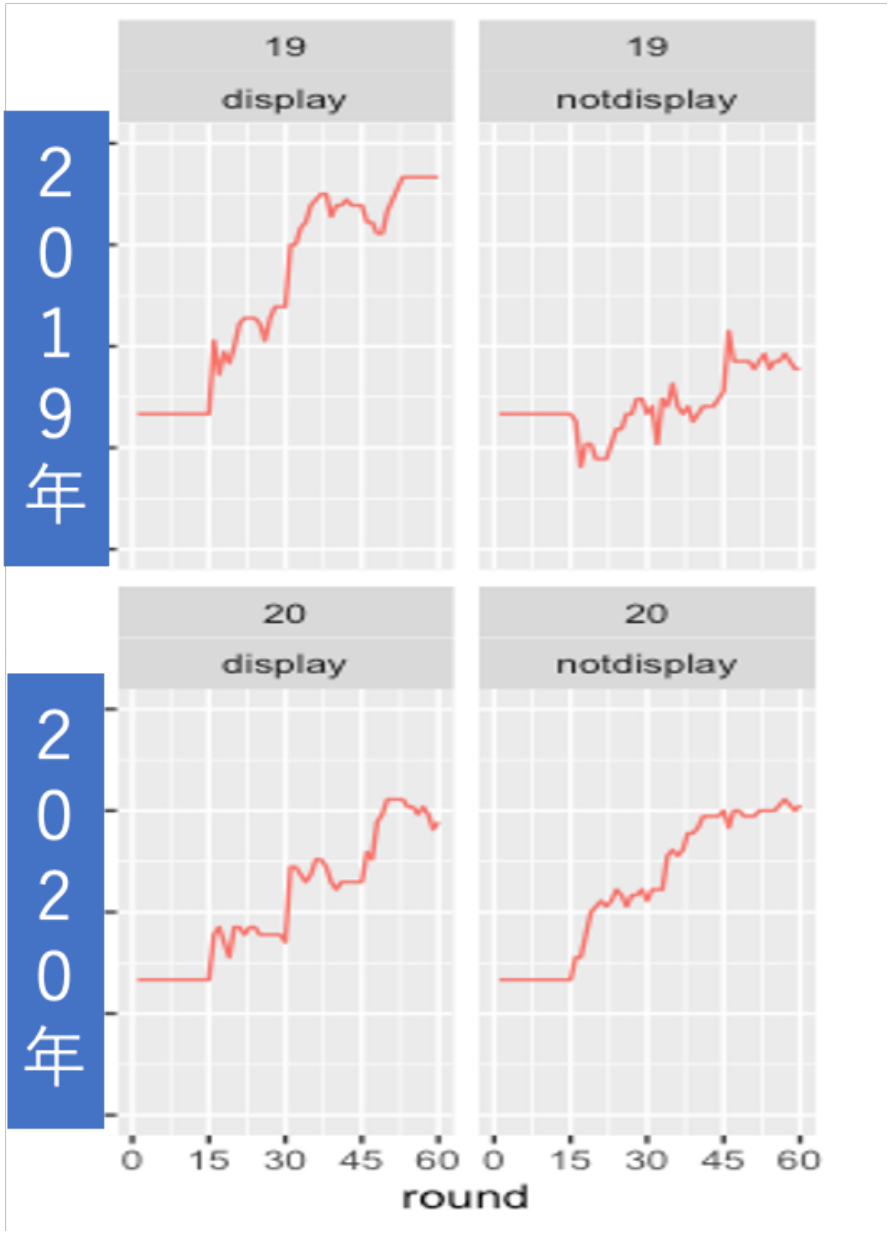


2021年

- 理論的には成功者模倣という学習バイアスが原因となって、戦略が拡散すると予測されていた
- 個人の意思決定プロセスを数理的にモデル化して、シミュレーション、データ分析することとした
- 修士課程に進学した阿部さんが実施：
  - 実験と同じ状況下では、目的となる戦略が拡散するために成功者模倣バイアスが必要であることを、シミュレーションで確認
  - データに対してモデルフィッティング→成功者模倣バイアスを持つモデル群が、そうでないモデル群よりも、WAICや事後予測プロットの当てはまりが良いことなどを確認

- 以上の分析を通して、2020年の統制条件で観察された予期せぬ結果は、偶然の産物である可能性が示唆された

→事前に例数設計を行い、大サンプルで実験を実施できていたら回避できたかも…



ここまでの流れをまとめると...

2019〜2020年（2年間）

1. 実験を実施

2021年（1年間）

2. 個人の意思決定プロセスをモデル化
3. 実験データ（集団レベルで生じる帰結）をシミュレート
4. ~~架空の実験データにモデルをフィッティングし、リカバリーできることを確認（例数設計）~~
5. 分析、モデルフィッティング

2022年（1年間）

6. 論文化、投稿

ここから何を伝えたいのか？

研究のモジュール化

小サンプルの実験を数多く重ねるのではなく、
一つの大きな研究を独立したモジュールに分割し実施

1つのモジュールに時間をかけて取り組む

- 3年生でゼミ配属されるまで、研究を担当した学生たちは、Python, R, Stan, 統計モデリング、計算論モデルについて授業を受けたことはない
- そのため、1つのモジュールが始まる都度、Pythonの入門書、oTreeのドキュメント、緑本、アヒル本、片平本を渡して、学習させるところから始まるので、時間はかかる

- ただし、モジュール化できず、従来のような小さな実験を１年で計画・立案する卒論もまだ多い（むしろ多数）

最後に：
学生やECRの研究評価の仕組み

「面白さ」から「手法の堅実さ」へ

1. データ収集前のアイデア・デザインに時間を費やす（プレレジを含む）
2. 例数設計、大サンプル、汎化性を目指す

- 1つの卒業論文／修士論文ですべて行うのではなく、これらを分割して実施するという提案をした
- そのためには、卒論レベルでは『結果の面白さ』ではなく『手法の堅実さ』を評価するよう頭を切り替える

経済学におけるjob market paperモデル

- 博士号を取得して、就職するにあたり、job market paperという論文を一本だけ執筆し、C.V.と論文一本で公募の審査が行われる
- 手堅く、大きく、質の高い研究に対する圧力となりうる
- ECRの公募においては「最近の論文5篇」を求めるのではなく、このような考え方の転換を目指すべきではないだろうか