Commentary

フィールド実験

川田恵介
東京大学社会科学研究所 教授

1 初めに

本稿は学術研究のみならず, 実務家による施策決定への応用も注目されている, フィールド実験について紹介する。誌面の都合上, 代表的な実践であるランダム化対照実験（以下, RCT）に基づく因果効果の評価法に絞って紹介する。さらにデータの統計的処理から生じる留意点に焦点を当てる。社会実験の実施には, 実務家を含む実験協力者とチームを構築し, 緊密な意思疎通を前提とした多面的議論が不可欠である。チームの中で,「データ分析の専門家」としての本稿読者に期待される役割は, データ処理についての専門的知見に基づいて助言することであることが多い[1)]。本稿は, その期待に応える出発点となるチュートリアルの提供を目指す。

通常の実験研究では, 研究関心となる変数の一部を実験実施者が人為的に操作できる状況を想定する。フィールド実験は実社会の中で行われる実験であり, 完全な人工的実験環境を整備できない。しかしながら実験計画法の発展によって, 操作できる変数のランダム化を注意深く行えば, 実社会の中で生じる多様な因果効果の推定が可能となっている。

因果効果の解明に向けたアプローチは複数存在する。その中でフィールド実験が注目される理由は, その分析プロセスの透明性を確保しやすい点にある。実験・非実験を問わず, 実証研究への大きな批判は, 分析の結論に至るまでのプロセスの不透明性に向けられている。不透明性の要因の一つは, データ分析方法の選択や結果の解釈を実験実施後に研究者が行う“慣行”にある。このような慣行は, 自覚的・非自覚的な恣意性を分析過程に紛れこませてしまい, 分析の不透明性を拡大させる。これに対し, 正しい手続きを守ったフィールド実験は, 分析プロセスの透明性を大きく前進させる。

2 実験計画書

プロセスの透明性を高める上で, 最も重要なのは,「事前の実験計画ありき」の姿勢である。実際の実験運用では, しばしば予期せぬ事態に直面し, 計画の修正が必要となりうる。しかしながら大量の事後修正は, データの解析結果を大きく損なう可能性がある。このため, 研究チーム内での入念な計画策定が必要となる。

実験計画書が最低限含むべき内容は以下である。(1) 実験参加者の募集方法と人数, (2) 介入変数と結果変数, (3) 介入変数の割り付け方法, (4) 収集する背景変数。これらの項目設定に対し, 後述する研究課題, 識別問題, 統計的推論問題における留意点を踏まえた, 適切な助言が必要となる。

完成した実験計画書は, 可能であれば実験実施前に公開することが望ましい。このような早期公開には, 安易な事後修正を抑制する効果が期待でき, 分析の透明性を向上させるからである。具体的な公開方法としては, 学会が提供するpre-registrationサービス[2)]の活用などの他に, プロジェクトのHPなどで公開することも考えられる。少

なくとも分析チーム内での共有は必須である。

3 研究課題

実験参加者と介入・結果変数は,チームが挑む研究課題と最も密接に関連している。介入変数とは実験実施者が操作する変数であり, 典型的には実験参加者への影響を“評価したい”変数である。結果変数とは, 介入変数の成果を測定する指標である。例えば,「ソロバン教育は, 参加児童の教育到達度をどの程度高めるのか?」という課題においては,ソロバン教育が介入変数, 教育到達度が結果変数, 小学生が実験参加者となる。

研究課題を策定する際には, 原因変数と結果変数のみならず, 実験前に調査する参加者の背景変数も指定すべきである。このような背景変数は後述する, (1) 効果の異質性の探索, (2) 推定精度の改善, (3) 実験結果の実験参加者以外への適用, を達成するために有益である。特に効果の異質性を探索したい変数は, 研究課題設定と密接に関わるので,チーム内で十分議論するべきである。また, 参加者間でのばらつきが大きく, かつ結果変数との相関が強いと予想できる背景変数も極力調査することが望ましい。

4 識別問題

次にデータ処理戦略の策定に移る。データ処理を論じる上で, 有用な論点整理は, 識別問題と統計的推論問題との区別である。識別問題では, 実験参加者数が「無限大」であると仮想的に想定し, 研究課題に回答できる実験デザイン・データ処理法を事前に議論する。

多くの識別問題において重要となるのは,データからでは観察しえない変数の影響を考えることにある[3)]。RCTでは, 因果効果の識別各実験参加者に対して, 介入変数の水準をランダムに決定することで, 観察できない変数の影響をコントロールする。ランダムに決定されている以上, 観察できない背景変数が, 異なる原因変数の間で偏るシステマティックな理由は存在しない。むろん偶然偏ってしまうリスクは存在するが, そのリスクは標準的な方法で定量化が可能である。

標準的なRCTで因果効果を識別するには, 以下の前提条件が必要になる。

Random Sampling：因果効果を推定したい集団は, 実験参加者そのもの, あるいは参加者がランダムサンプリングされたと見做せる集団であること。
No interference：実験参加者間での相互作用は存在しないこと。

上記の条件は, 特に社会科学や実務における応用では満たされないケースも多い。例えばRandom samplingについては, しばしば実験参加者の母集団と真に関心のある集団とが一致しないケースが多いことが指摘され, External validity問題と呼ばれている。この問題を受けて, 実験参加者（あるいはその母集団）以外に実験結果を適用する方法には, 数多くの提案がある。シンプルな方法としては, 実験において収集された背景情報を用いて修正するアプローチである（Stuart et al. 2011）。また, その拡張として, 大量の背景情報の処理法（Cappiello et al. 2021）, 関心のある集団について収集すべき背景情報の同定法（Egami and Hartman 2021）, なども提案されている。さらに, 観察できない背景属性も考慮したsensitivity analysisなども議論されている（Nie, Imbens, and Wager 2021）。

No interferenceについても成り立たないことは多い。また, 特に社会科学においては, 相互作用自体が研究対象となる。相互作用についても含意を得られるような実験計画として, 例えば相互作用を市場均衡として近似する方法（Wager and Xu 2021）, ネットワークとして近似する方法（P. M. Aronow and Samii 2017）

などがある。

さらには効果の大きさだけでなく，効果が生じるメカニズムを明らかにする実験デザイン（Imai, Tingley, and Yamamoto 2013），あるいはRobustな数理モデルを併用することで観察できない厚生指標への因果効果推定（Finkelstein and Hendren 2020）についても数多くの手法研究が存在する。

5 統計的推論問題

次に，予想される実験参加者数を念頭に，識別された因果効果を推定する方法を策定する。現実の実験における実験参加者は，言うまでもなく，有限であり，識別問題をそのまま適用することはできない。そこで，このような有限なサンプルのもとで，どのような結論を得ることができるのかという統計的推論問題を論じる。RCTの入門的な教科書ではしばしば省略されるが，統計的推論においてもRCTは分析結果の頑強性を大幅に高めることができる。

RCTに限らずデータ分析一般において，統計的推論の中心的な論点は，“偶然生じる”データの偏りがもたらす影響の定量的評価である。多くの実践では，ある程度の実験参加者数（例えば150名以上）を前提に，古典的な頻度論的漸近理論に基づく定量評価が行われている。これはランダムサンプリングデータにおける推定誤差を評価する場合にも用いられてきた方法であり，評価結果は標準誤差，及びそこから計算される信頼区間によって報告できる。

誤差評価のための最もシンプルな方法は単純な差の推定法の適用である。この方法は，事後分析の余地がなく，分析の透明性を確保できるという利点がある。しかしながら限定的なサンプルサイズのもとでは，介入の偶然の偏りによる弊害が大きく，推定精度が低い（信頼区間が大きい）場合も多い。よって本稿では，2種類の補正方法を紹介する[4)]。

6 事前調整

背景変数の偏りを軽減する有力な方法は，介入変数の層化割り付け（Stratified assignment）である。実験参加者が確定したのち，背景情報を収集し，グループ分けを行い，その後グループ内でランダムに介入変数の割り付けを行う方法である。例えば，性別・学年の情報を事前に収集し，同じ性別・学年のグループを作成し，そのグループの中の50%に介入を行う。この方法では，少なくとも性別・学年については，介入を受けた・受けていないグループ間での偏りが生じ得ず，推定の精度が改善する。さらに近年，オンライン上での社会実験が拡大する中で，より豊富な背景情報が活用可能になってきている。このような背景情報の活用法も議論されており（Doudchenko et al. 2021），オフライン実験での応用も期待される。

事前調整の実践上の困難は，単純割り付けに比べて，介入変数の割り付け方に手間がかかる点である。実験参加者の背景情報を収集した後に，それに応じてランダム化を実行する必要があり，実験環境によっては実施が不可能な場合もある。このような場合は，後述する実験後のデータ処理による事後調整の手法を活用することが考えられる。

7 事後調整

事後調整の代表的な方法は回帰分析である。回帰分析については，前提とするモデルの定式化への依存度，その帰結としての分析の不透明化の懸念がある。しかしながら，RCTデータはモデルの誤定式化に対して頑強であり，実験データとしての強みの一つとなっている（P. Aronow et al. 2021）。また，nonparametricな手法としてmatching法や機械学習の応用など，多くの手法が提案されており，すでに容易

に応用できる段階となっている[5]。

8 効果の異質性

以上で紹介した手法を用いて，参加者全体での平均効果（周辺化平均効果）を推定することは比較的容易である。対して，背景属性についてのサブグループ内での平均効果（条件付き平均効果）の推定はよりチャレンジングである。全ての背景変数について条件づけるとサンプルサイズが極めて少数になり，信頼区間の近似計算が信用できない，あるいは推計誤差が大きくなりすぎる可能性が高い。このため，使用する背景変数を適切に“間引く”ことが必要となる。かつてはデータ分析者が実験実施後に，専門・背景知識などに基づいてこの作業を行っていた。しかしながら，分析者主導の事後的アプローチは，しばしば恣意的なサブサンプリングが行われる危険性があり，分析の透明性の低下を招いてしまう。

効果の異質性推定に用いるモデル選択について，機械学習・セミパラメトリック推計の手法を応用し，データ主導で行う手法は過去10年間で急速に発展した[6]。データ主導のアプローチは，分析に用いたコードなどを用いて事後的な検証が可能であり，分析者が事後的にサブサンプル分けを行う場合よりも，分析の透明性を確保しやすい。

9 何を報告すべきか？

統計的推論は推計誤差について有益な情報をもたらすので，それを適切に報告する必要がある。全てのRCTは，常に実験結果についての不確実性をもつ。この不確実性をどのように評価，報告するのかについては，近年改めて活発な議論が行われている[7]。“最大公約数”的に推奨したいのは，「効果＝0を帰無仮説とした統計的検定結果をとりあえず報告する」という慣行を避けることである。このような検定結果がもつ学術的・実務的含意は多くない。RCTによる因果的評価の文脈においても，効果の有無ではなく，どの程度の効果をもつのかという定量的な評価の方が重要な局面は多い。このため，古典的頻度論の枠組み内においても，p値ではなく，標準誤差や信頼区間などの報告が推奨される。

結果変数を事前に絞り込むこともまた重要である。多数の結果変数について，分析結果を示そうとすることは，分析の透明性を阻害する大きな要因となる。このような状況におけるデータ分析法も数多く提案されているが[8]，第2種の過誤の増加など，何らかのコストを伴う。このため，実験計画段階でチーム内で入念に議論し，主たる関心とする結果変数を一つに絞り込むことが推奨される。

10 さらなる取り組み

質・量を問わず，今後の実証分析における大きな課題の一つは，研究プロセスの透明性の確保であり，フィールド実験についてもその例外ではない。本稿で紹介したアプローチは，透明性確保に一定の貢献はなしうるであろうが，当然これだけでは限界がある。

まず必要なのは，Code-firstなデータ分析の徹底である。分析に用いたソフトや追加programのversion，分析用のcodeが保存されていなければ，事後的検証は不可能である。またpythonやRなどのopen-sourseの利用は，外部の事後的検証を容易にするという観点からも推奨される。

さらには個々のチームを超えた，“コミュニティ”としての取り組みも重要である。例えばPre-registrationの仕組みにどのように実効性をもたすのか，論文として成果を発表する場合はどのような査読を行うのか，これらの課題については学会や協会レベルでの取り組みが必要不可欠である。

注

1) 詳細なデータ処理法を含むより包括的な議論は，Duflo and Banerjee (2017) などを参照のこと。

2) 例えばAmerican Economic Associationは，AEA RCT Registration (https://www.socialscienceregistry.org/) を提供している。

3) 代表的な理論的枠組みは潜在結果モデル(Imbens and Rubin 2015) や有向非循環グラフ (Pearl and Mackenzie 2018) がある。

4) 両補正方法ともに，Declaredeisgn project (https://declaredesign.org/) にて提供されているR packages, estimatorやrandomizrで容易に実装できる。

5) Athey and Imbens (2017)

6) Rでの簡便な実装としては，grfパッケージ(https://github.com/grf-labs/grf) などがある。

7) 直近のまとめとしては，Imbens (2021) などを参考のこと。

8) 多重検定問題と呼ばれる。

文献

Aronow, Peter M. and Cyrus Samii, 2017, "Estimating Average Causal Effects Under General Interference, with Application to a Social Network Experiment", *The Annals of Applied Statistics* 11 (4): 1912-1947.

Aronow, P. M., James M. Robins, Theo Saarinen, Fredrik Sävje and Jasjeet Sekhon, 2021, "Nonparametric Identification Is Not Enough, but Randomized Controlled Trials Are", *arXiv Preprint arXiv*:2108.11342.

Athey, Susan and Guido W. Imbens, 2017, "The State of Applied Econometrics: Causality and Policy Evaluation", *Journal of Economic Perspectives*, 31 (2): 3-32.

Cappiello, Lauren, Zhiwei Zhang, Changyu Shen, Neel M. Butala, Xinping Cui and Robert W. Yeh, 2021, "Adjusting for Population Differences Using Machine Learning Methods", *Journal of the Royal Statistical Society: Series C (Applied Statistics).*

Doudchenko, Nick, Khashayar Khosravi, Jean Pouget-Abadie, Sebastien Lahaie, Miles Lubin, Vahab Mirrokni, Jann Spiess and others, 2021, "Synthetic Design: An Optimization Approach to Experimental Design with Synthetic Controls", *Advances in Neural Information Processing Systems* 34.

Duflo, Esther and Abhijit Banerjee, 2017, *Handbook of Field Experiments*: Elsevier.

Egami, Naoki and Erin Hartman, 2021, "Covariate Selection for Generalizing Experimental Results: Application to a Large-Scale Development Program in Uganda", *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Finkelstein, Amy and Nathaniel Hendren, 2020, "Welfare Analysis Meets Causal Inference", *Journal of Economic Perspectives* 34 (4): 146-167.

Imai, Kosuke, Dustin Tingley and Teppei Yamamoto, 2013, "Experimental Designs for Identifying Causal Mechanisms", *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176 (1): 5-51.

Imbens, Guido W., 2021, "Statistical Significance, *p*-Values, and the Reporting of Uncertainty", *Journal of Economic Perspectives* 35 (3): 157-174.

Imbens, Guido W. and Donald B. Rubin, 2015, *Causal Inference in Statistics, Social, and Biomedical Sciences*: Cambridge University Press.

Nie, Xinkun, Guido Imbens and Stefan Wager, 2021, "Covariate Balancing Sensitivity Analysis for Extrapolating Randomized Trials Across Locations", *arXiv Preprint arXiv*:2112.04723.

Pearl, Judea and Dana Mackenzie, 2018, *The Book of Why: The New Science of Cause and Effect*: Basic books.

Stuart, Elizabeth A., Stephen R. Cole, Catherine P. Bradshaw and Philip J. Leaf, 2011, "The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials", *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174 (2): 369-386.

Wager, Stefan and Kuang Xu, 2021, "Experimenting in Equilibrium", *Management Science*.