How to Generalize?

Noah Dixon* Hanbai Wang†

October 30, 2025

Abstract

KEYWORDS:

JEL Classification:

*Texas A&M University; noah.dixon@tamu.edu
†Texas A&M University; hanbai_wang@tamu.edu

Contents

Introduction

There are many reasons to believe that economics is particularly unique in its "generalization" dilemma. In medicine, for instance, it is widely believed that treatment effects occur on a cellular level. Consequently, it is valid to assume that the results of those studies can be extrapolated to a larger population (although external validity is debated, even in medicine [Ling et al., 2023]). In economics, however, treatment effects are *contagious,* in the sense that they are heavily dependent on regional-level covariates, institutional environments, and interactions among individuals. As a result, causal effects estimated in one environment may not transport cleanly to another environment.

Now, suppose, as a researcher, we are interested in estimating the effects of cash transfers on health outcomes in a developing country. Naturally, we sample a multitude of villages, and likely utilize a staggered treatment approach. The average treatment effect (ATE), in this instance, is identified by experimental design, with standard errors clustered at the village level. Typically, when we cluster, we have this basic conceptualization in mind. However, more generally, there may be additional reasons to cluster, beyond precision. For instance, *ex-ante*, we may care about identifying the population-level effects. Consequently, representative sampling – which occurs by sampling an adequate number of clusters relative to the population – is necessary to identify the population-level effects. Clustering, therefore, is not merely a correction for correlated outcomes – it is also central to

how we conceptualize the population to which our causal estimates generalize.

One issue with representative sampling is that the relevant population is theoretically unbounded. Consider the example above, where we aim to estimate the effects of cash transfers on health outcomes. Naturally, it seems intuitive to target a large, poor city, where a control group and treatment group are easily stratified. Yet, in many countries, the vast majority of the population lives within small, rural villages. Suppose we instead sample a single village (assuming a sufficiently large sample size and no ethical concerns from withholding treatment). In that case, the ATE is internally valid to that specific region. To improve generalizability, we might sample multiple villages throughout a region – but even this may not capture the heterogeneity within a country. And, even if we could hypothetically sample each individual within the country, the question of generalization, across countries, is still unanswered. At each node, we must expand the relevant "population."

Intuitively, we can think of this dilemma as relating to the issue of *scalability.* A practical solution is to "think smaller." In this paper, we investigate exactly how the experimental results from the initial population should be *sampled* to match the distribution of the target population. To do this, we require two things: (i) proper heterogeneity in the initial population, and (ii) sufficiently quantified data in the target population. Consequently, we propose an easy-to-implement algorithm for "precise sampling" from the initial population, to predict treatment effects in the target population. We run a series of

experiments to validate our econometric techniques.

Literature Review

There are four areas of literature that need to examined: (i) optimal experimental design, (ii) policy learning, (iii) transportability, and (iv) covariate resampling. *I will need to expand each of these.*

Transportability

In an implicitly canonical study, Hotz, Imbens, and Mortimer (2004) digress into the econometrics of transportability by evaluating empirical evidence from job training programs. Pearl and Bareinboim (2011; 2014) describe the formal conditions under which the results from one environment can be *transported* into another environment. *Proving that our algorithm satisfies the "do-algebra" proposed in this canonical study is essential to showing its statistical efficiency.* Recently, Wang, Han, and Huang (2024) develop an ongoing-sampling framework that tracks representatives over time to keep online A/B test results externally valid when underlying user characteristics shift.

Covariate Resampling

Historically, propensity scores have been used to balance covariates *within* sample (Rosenbaum & Rubin, 1983). However, an increasingly important question is whether propensity scores can be used to balance covariates *among* samples (Tipton 2013, Stuart et al. 2011, Dahabreh & Hernán 2019). Adjaho and Christensen (2023) describe the conditions under which an individualized treatment policy – like that proposed in Athey and Wager (2021) – from one dataset (via an experiment or observational data) performs well when applied to a different population. Our method is robust to the methodology described in both of these studies.

Optimal Experimental Design

There is an ongoing debate regarding optimal experiment design, with proposed approaches including framing the problem as a dynamic programming problem with Bayesian priors and cost constraints (Higbee, 2024), utilizing stratification trees to improve balance (Tabord-Meehan, 2022), and determining precisely when and how to randomize or rerandomize (Banerjee et al., 2020). Numerous studies have also examined how to optimally choose design parameters – number of clusters, periods, sample sizes per period – to maximize statistical efficiency (Liu & Li, 2024; Watson et al, 2022). There is also the question of *where* to implement a study, to maximize external validity (Gechter et al. 2024). Our method is not

interested in optimal experimental design, but rather optimal data collection. That is to say, it is robust to all the methods listed above.

Policy Learning

Athey and Wager (2021) derive an algorithm for selecting the optimal treatment rule, given a sufficiently large quantity of data. While this treatment rule can increase welfare when the cost of prescribing treatment is low (Athey, Keleher & Spiess, 2024), in many instances, there are ethical concerns from withholding treatment from an arbitrary member of the population who could *plausibly* benefit from treatment. Further, there may be institutional barriers limiting the efficacy of policy learning (Wang & Yang, 2025). Additionally, there may be multiple objectives for policy makers to maximize (Rehill & Biddle, 2025), thereby Finally, while policy-learning methods typically fall under reduced-form rule learning, there is an argument that structural counterparts should be viewed as a complements to these algorithmic processes – not a substitute (Todd & Wolpin, 2023).

Methodology

Basic Framework

Let $p, p' \in \mathcal{P}$ represent two arbitrary populations, where $\mathcal{P}$ represents the set of all possible populations we could sample from. Let $X \in \mathbb{R}^{n \times k}$ represent the covariates collected in the experiment conducted in p and let $X' \in \mathbb{R}^{n \times k}$ be covariates collected in p', where n is the number of observations and k is the number of covariates. Assume that $X' \subseteq X$.[1] For each $i \in p, p'$, let $D_i \in \{0, 1\}$ denote treatment status and let Y_i^d denote the potential outcome under treatment state $d \in \{0, 1\}$. Assume that an experiment has already been conducted in p but not p', such that $\tau_p := \mathbb{E}_p[Y^1 - Y^0] = \int (Y^1 - Y^0) dF_p(X, Y^1, Y^0)$ is identified, where proper randomization ensures internal validity. Let $\tau_{p'} := \mathbb{E}_{p'}[Y^1 - Y^0] = \int (Y^1 - Y^0) dF_{p'}(X, Y^1, Y^0)$. While τ_p is observed, $\tau_{p'}$ is not. The goal of this paper is to determine: (i) how to sample from $F_p(X, Y^1, Y^0)$ – using the information collected in X' – to reliably construct $F_{p'}(X, Y^1, Y^0)$, and (ii) the assumptions underlying the process in (i). Consequently, under reasonably assumptions, we can reliably estimate $\hat{\tau}_{p'}$ – using a plethora of tangential methods – without running a separate experiment.

[1] In other words, assume that X is at least as large as X'. While X is collect within the experiment, X' is collected externally (e.g., via census data, survey data, etc.).

Illustrative Example

Table 1: ⋆⋆Replace with Progresa Data⋆⋆

Category	p'	p
Ages 20–40	60%	40%
Ages 40–60	20%	30%
Ages 60–100	20%	30%
Female	30%	50%
Male	50%	50%

Table 1 represents the delicacy of the underlying sampling distributions. If we sample "at random" from p, the resulting sample will be unrepresentative of p' and statistically inefficient.

Algorithmic Outline

Our method includes the following stages:

1. Estimate a target distribution, $G_{p'}$, over the covariates X' in population p'.
2. Draw $b = 1, ..., \mathcal{B}$ bootstrap samples from $G_{p'}$.
3. For each b, we use "nearest neighbor" to match observations to

the nearest covariate profile in X. One issue here is ensuring that there is covariate overlap.

4. Iterate this procedure $s = 1, ..., \mathcal{S}$ times to obtain a stable, Monte-Carlo estimate of $\hat{\tau}_{p'}$. For sufficiently large n (observations constituting X and X'), it is also possible to iterate an algorithmic policy rule (Athey & Wager, 2021). In particular, this method is adventitious because it allows complete research flexibility in the experimental design and policy evaluation phase.

One question is how to accurately sample when the large sample contains clusters. This is often neglected in the machine-learning literature, because variance and bias are equally weighted.

Asymptotic Properties

Data

To test our algorithmic process, we will use the Progresa data set. *I have submitted a request for this data set.* As shown in Skoufias et al. (2001), this data set contains rich household measures on 24,000 households from 506 villages throughout Mexico. Intriguingly, we could use this data set to attempt to predict the CATE in other places, *particularly those in which the effects of cash transfers were negligible.*

Results

Estimation

Robustness

Conclusion

Limitations

Extensions

Sources

Appendix