# Modern Sampling Methods

Class 1: Introduction

January 10, 2022

# Introduction: Instructors

Keisuke Hirano
Department of Economics, Penn State University
kuh237@psu.edu

Jack Porter
Department of Economics, University of Wisconsin
jrporter@ssc.wisc.edu

This course is about the interaction between:

- data design — creating and collecting data; and
- data analysis — estimation of effects, hypothesis testing, and analysis of welfare properties.

Some key themes:

- designs for identification;
- designs for efficiency/precision;
- designs for intervention/welfare maximization;
- things that can go wrong!

# Course Plan

1. Introduction
2. Randomized Experiments
3. Publication Bias and Preanalysis Plans
4. Treatment and Policy Choice
5. Multi-Wave Experiments
6. Covariate-Adaptive Randomization
7. Bandit Algorithms
8. Applications of Bandits
9. Statistical Inference with Adaptively Generated Data
10. Window Choice in Time Series

- Class sessions will generally match topic list, but we may adjust if some sections take more or less time.
- Syllabus and slides will be posted to: `https://github.com/keihirano/modern-sampling`
- Syllabus contains many additional references and readings for those interested in exploring a topic in greater depth.
- You are invited to post questions to Zoom chat; we will pause occasionally to respond.
- You are also welcome to email us after the class with followup questions.

# Remainder of This Session

- Preview some topics/application areas
- Review the i.i.d. assumption and some standard results
- Identification concepts

# Randomized Controlled Trials in Economics

Randomized controlled trials (RCTs) have become widely used in economics, including field experiments and lab experiments. Some issues we will discuss:

- Understanding the potential benefits of RCT designs for identification;
- Design choices in RCTs: randomization schemes, sample sizes;
- Inference issues that arise from designed RCTs;
- Even with RCTs, bias can arise through reporting and publication decisions.

# A Pre-Analysis Plan

Casey, Glennerster, and Miguel, 2012, "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan," *Quarterly Journal of Economics*.

Large scale randomized experiment designed to study the effect of a community development program in Sierra Leone.

- Provides findings based on pre-analysis plan.
- Provides findings without the constraints of the pre-analysis plan.

Reveals the scope for cherry-picking, hindsight bias, etc.

# Treatment Assignment

Dehejia, 2005, "Program Evaluation as a Decision Problem," *Journal of Econometrics*.

Uses the GAIN experiment data to propose individualized treatment assignment rules (profiling).

- Embeds the problem in a decision theoretic framework.
- Shows how results can depend on welfare objective function.

Draws a distinction between the treatment assignment problem and statistical analysis of the average treatment effect.

# A Multi-Wave RCT

Karlan and Wood (2017)

Experiment to study different ways to solicit donations to a charitable organization in a direct-mail compaign.

- Control: standard appeal (story about a program recipient).
- Treatment: added information about rigorous evaluation studies.

Study was conducted in two waves, with the second wave adjusted based on data from first wave.

# Bandits in E-Commerce

Online platforms make it possible to rapidly adjust prices, advertising strategies, etc. in response to accumulating evidence.

Example: Schwartz, Bradlow, & Fader (2017):

- adaptive online advertising algorithm;
- large decision space (over 500 possible choices);
- field experiment to compare with uniform randomization;
- adaptive rule increased customer acquisition rate by 8%.

We will discuss these “bandit” problems, general heuristics and some properties, and some recent applications.

# Adaptive Randomized experiments

Adaptive economic field experiments, e.g. Caria, Gordon, Kasy, Quinn, Shami, & Teytelboym (2021).

Choice of algorithm depends on goal: in-sample optimization vs. inference about effects (for future policy choice).

Statistical inference for treatment effects using data adaptive experiments is more complicated than with classic RCT designs.

# Window Selection

When analyzing economic times series, common to limit the window of observations (e.g. post-Great Recession, pre-Covid, etc.).

Related issues with cross-sectional data analysis: restriction to certain subpopulations, etc.

There is a rich literature on testing for structural breaks, etc., but what happens to post-test analysis of data?

Issues:

- Robustness to changing parameters/environment;
- Bias/Variance tradeoff in estimation and prediction.

# Background: the i.i.d. Assumption

Most textbook (cross-sectional) econometrics starts with the assumption that data are i.i.d.

This is a mathematical abstraction of a simple survey, where:

- there is a well-defined population;
- individuals are chosen completely randomly and with replacement
- (or the population is infinitely large);
- with no weighting/targeting of specific subpopulations;
- and perfect response.

In reality data collection is more complicated and the design has strong implications for identification, estimation, and inference.

- Many surveys have some degree of stratification with associated survey weights;
- Many surveys and experiments have nonresponse.
- Experimental treatment assignment protocols may break the i.i.d. condition, even in simple one-shot experiments.

# Useful Results for i.i.d. Data

Suppose

$$Y_1, \ldots, Y_n \overset{iid}{\sim} F,$$

where $F$ is some distribution.

Law of Large Numbers: if $\mu = E[Y_i]$ exists, then

$$\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i \xrightarrow{p} \mu.$$

Central Limit Theorem: if $\mu = E[Y_i]$ and $\sigma^2 = V[Y_i]$ exists, then

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

In fact, we can estimate the entire distribution (CDF) of $Y_i$: for any $y$, let

$$F(y) = \Pr(Y_i \leq y).$$

Then the empirical CDF

$$\hat{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} 1(Y_i \leq y) \xrightarrow{p} F(y).$$

Similar results hold for vectors $(X_i, Y_i)$, etc.

The upshot is that under i.i.d. sampling we can "learn" (estimate) the joint distributions of <u>observable</u> variables.

We can also estimate objects like conditional expectations

$$m(x) = E[Y_i | X_i = x].$$

- If we specify linear conditional mean $E[Y_i | X_i = x] = x'\beta$, then we have a classic linear regression problem.
- Can estimate $m(x)$ flexibly, using series, or kernel regression methods.
- "Machine learning" methods can be useful when $X_i$ is high-dimensional.
- Other objects like conditional quantiles can also be estimated.

# Identification Concepts

Models in economics often involve *latent* variables, e.g.

- Individual preferences/valuations in demand and auction models
- Potential outcomes in causal inference models

Marschak (1953): want to evaluate some possible *interventions*; a "structural" model is invariant to those interventions.

Let $S$ denote full set of (latent) model variables, and let $G$ denote their distribution:

$$S \sim G, \quad G \in \mathcal{G}.$$

We are interested in some parameter (e.g. treatment effect) $\theta(G)$.

There is some mapping from the latent variables to observable variables: $S \mapsto Z$, inducing a distribution

$$Z \sim F, \quad F = \mathrm{Obs}(G).$$

However, some information may be lost in the transition from $G$ to $F$.

The identification problem: assume we "know" (can estimate) $F$. Can we recover $G$ or $\theta(G)$?

For any (feasible) $F$, let the set of possible $G$ that generated $F$ be:

$$\tilde{\mathcal{G}}(F) = \{G \in \mathcal{G} : F = \mathrm{Obs}(G)\}.$$

and let

$$\Theta(F) = \left\{\theta(G) : G \in \tilde{\mathcal{G}}(F)\right\}.$$

If $\Theta(F)$ is a singleton, we say that $\theta$ is (point-) identified at $F$. In principle, since we can estimate $F$, we should be able to estimate $\theta$.

More generally, if $\Theta(F)$ is a (strict) subset of $\Theta$, the parameter is partially identified and we may be able to estimate the identified set.

# Example: Recovering Bidder Valuations

There are multiple auctions $j = 1, \ldots, J$ with bidders $m = 1, \ldots, M$ in each.

Suppose individuals have independent private values $V_i \stackrel{iid}{\sim} G$ (same across auctions).

In a second-price sealed-bid auction, dominant strategy is to bid your valuation:

$$B_i = V_i \quad \Rightarrow \quad F = G.$$

If we observe bids, then we can estimate

$$\hat{F}_n(v) = \frac{1}{n} \sum_{i=1}^{n} 1(B_i \leq v) \xrightarrow{p} F(v) = G(v).$$

In a first-price sealed-bid auction, equilibrium strategy is to shade your bid:

$$B_i \leq V_i \quad \Rightarrow \quad F \neq G.$$

However, there is a one-to-one mapping between $F$ and $G$ (that depends on $M$); see e.g. Athey & Haile (2002).

We can still consistently estimate $F$, and then use the mapping to obtain the implied $G$.

# Example: Nonresponse in Surveys

Suppose $Z_i \overset{iid}{\sim} G$ and we are interested in $G$ or

$$\theta(G) = E[Z_i] = \int z dG(z).$$

We sample individuals but not all respond:

- $T_i = 1$ if individual responds
- Observe $T_i$ and $T_i \cdot Z_i$.

Then we can learn $F =$ distribution of $(T_i, T_i \cdot Z_i)$.

Whether we can recover $G$ from $F$ depends on the nature of $T_i$. For example, if $T_i$ is independent of $Z_i$, then

$$E\left[Z_i \mid T_i = 1\right] = E[Z_i] = \theta,$$

and $\theta$ is identified provided $\Pr(T_i = 1) > 0$.

Similarly we can learn the entire distribution $G$ through the distribution of $Z_i \mid T_i = 1$.

However, if $T_i$ is dependent with $Z_i$, what we can learn may be more limited, and depends on the restrictions we place on the nonresponse model (and the model for $Z$). See e.g. Manski (1995).