Causal Inference using Difference-in-Differences

Lecture 14: Random Treatment Timing

Pedro H. C. Sant'Anna
Emory University

January 2025

Motivation: Application

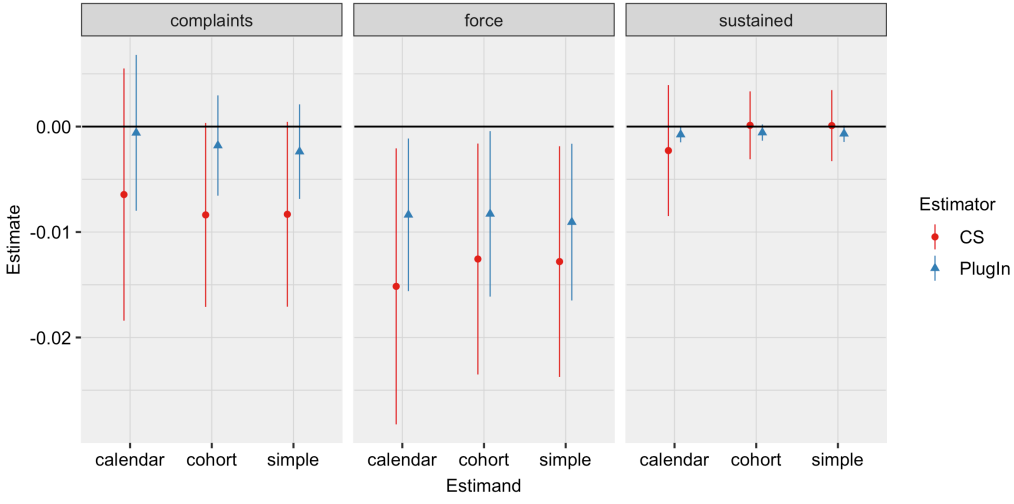
Application - Background

- Reducing police misconduct and the use of force is an important policy objective.
- Wood, Tyler and Papachristos (2020a, PNAS) studied a randomized rollout of a procedural justice training program for police officers.
  - Emphasized respect, neutrality, and transparency in the exercise of authority
- The original study found large & significant reductions in complaints/use of force.
- In Wood, Tyler, Papachristos, Roth and Sant'Anna (2020b), we re-analyzed data using the method of Callaway and Sant'Anna (2021).
  - No significant impacts on complaints; borderline significant effects on force; but CIs for all outcomes were wide

Is the sampling approach to inference here adequate?

Can we do better than DiD when treatment timing is random?

In Roth and Sant'Anna (2023b), "Efficient Estimation for Staggered Rollout Designs", we tackle these problems!



Population: Chicago Police officers that are not on special forces (excluding those selected into the pilot program)

			Plug-In				CS				
Outcome	Estimand	Pre-treat Mean	Estimate	LB	UB	p-val (FRT)	Estimate	LB	UB	p-val (FRT)	CI Ratio
complaints	simple	0.052	–5%	–13%	4%	0.29	–16%	–33%	1%	0.06	2.0
complaints	calendar	0.052	–1%	–15%	13%	0.89	–12%	–35%	11%	0.28	1.6
complaints	cohort	0.052	–3%	–13%	6%	0.47	–16%	–33%	1%	0.06	1.8
sustained	simple	0.004	–15%	–33%	2%	0.09	2%	–75%	79%	0.95	4.3
sustained	calendar	0.004	–17%	–34%	–0%	0.04	–52%	–194%	90%	0.50	8.4
sustained	cohort	0.004	–13%	–31%	5%	0.16	3%	–71%	76%	0.93	4.2
force	simple	0.051	–18%	–32%	–3%	0.03	–25%	–46%	–4%	0.03	1.5
force	calendar	0.051	–16%	–30%	–2%	0.02	–30%	–55%	–4%	0.02	1.8
force	cohort	0.051	–16%	–31%	–1%	0.08	–24%	–46%	–3%	0.04	1.4

Table 1: Estimates and 95% CIs as a Percentage of Pre-treatment Means

Introduction

Introduction

- We are often interested in the causal effect of a treatment that is rolled out to different units at different times.
- Staggered rollouts are often analyzed using methods that extend the simple two-period **difference-in-differences** estimator to the staggered setting.
- This includes two-way fixed effects (TWFE) regressions and recent alternatives that deal better with treatment effect heterogeneity
  (e.g. Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; de Chaisemartin and D'Haultfœuille, 2020).
- The validity of these estimators depends on a **parallel trends** assumption:

$$\mathbb{E}\left[Y_{i,t}(\infty) - Y_{i,t-1}(\infty) | G_i = g\right] = \mathbb{E}\left[Y_{i,t}(\infty) - Y_{i,t-1}(\infty) | G_i = \tilde{g}\right] \text{ for appropriate } \tilde{g}'\text{s.}$$

Introduction

- But the justification for parallel trends is often **(quasi-) random treatment timing**
- In some cases, treatment timing can be randomized by design
  - E.g., our application to the randomized rollout of a police training program in Chicago
- In other cases, treatment timing is argued to be "as good as random"
  - "We exploit the random timing of sudden parental deaths due to car crashes, other accidents, and unexpected heart attacks." (Druedahl and Martinello, 2019)
  - Other examples include quasi-random timing of health shocks (Fadlon and Nielsen, 2021), parental deaths (Nekoei and Seim, 2019), social security office closings (Deshpande and Li, 2019), stimulus payments (Parker, Souleles, Johnson and McClelland, 2013)
- If we have random treatment timing, can we get **more efficient** estimators than DiD?

How to answer this question?

- Roth and Sant'Anna (2023b) introduce a **design-based framework** formalizing the notion of random treatment timing.
- We consider the estimation of a **large class of causal parameters** that aggregate average effects across periods and cohorts.
- We solve for the **efficient estimator** in a class of estimators that nests existing DiD approaches.
  - Efficiency gains can be large: reductions of SEs of 2X (or more) in Monte Carlos/Application!
- We provide both *t*-based and permutation-test-based methods for randomization inference.

Framework

Framework

- Finite population of units: $i = 1, ..., N$
- T periods: $t = 1, ..., T$
- Unit i is first treated at period $G_i \in \mathcal{G} \subset \{1, ..., T\} \cup \{\infty\}$
  - $G_i = \infty$ denotes never treated.
  - Treatment is an "absorbing state" (no switching on-off).
- Potential outcomes: $Y_{i,t}(g)$ = i's outcome in t if first treated at g
- We observe $Y_{i,t} = \sum_g 1[G_i = g] Y_{i,t}(g)$
- Adopt a **design-based framework:** $Y_{i,t}(\cdot)$ and $N_g = \sum_i 1[G_i = g]$ treated as fixed, G is stochastic.

Background on design-based approaches (part 1)

- Two alternative approaches to accounting for uncertainty about causal parameters:
  - **Sampling-based:** Imagine a super-population from which we draw a small fraction of units. The uncertainty reflects that we draw different units in each sampling draw.
  - **Design-based:** Finite population is fixed. The uncertainty reflects that units are assigned to different treatment cohorts in different draws.
- We discuss some of this in Lecture 3 when we talked about inference.
- Design-based approach attractive when it is hard to imagine the "super-population".
  - E.g. if states or counties are units of observation, or one has admin data on the full population (Manski and Pepper, 2018; Abadie, Athey, Imbens and Wooldridge, 2020).
- This has prompted a lot of recent work considering a design-based approach in statistics/econometrics (Imbens and Rubin, 2015; Athey and Imbens, 2022; Abadie et al., 2020; Bojinov, Rambachan and Shephard, 2021; Rambachan and Roth, 2022; Xu, 2021).

Background on design-based approaches (part 2)

- Design-based approaches are also useful for thinking about finite-sample properties of permutation tests (aka Fisher Randomization Tests).
- And as we'll show, there are nice connections between our setting and design-based work on covariate-adjustment in randomized experiments.
- That said, the approaches are *not so different:*
  Even if the sample is drawn from superpopulation, estimators are valid for Sample Average Treatment Effect (SATE).
- But more work to be done connecting efficiency results in design-based and sampling-based frameworks

Table 1: Design-based uncertainty with four time periods and three different treatment cohorts

	Actual Sample				Alternative Sample I				Alternative Sample II				...
Unit	$Y_{i,t}(2)$	$Y_{i,t}(3)$	$Y_{i,t}(4)$	G_i	$Y_{i,t}(2)$	$Y_{i,t}(3)$	$Y_{i,t}(4)$	G_i	$Y_{i,t}(2)$	$Y_{i,t}(3)$	$Y_{i,t}(4)$	G_i	...
1	?	?	✓	4	?	✓	?	3	?	?	✓	4	...
2	✓	?	?	2	?	✓	?	3	?	✓	?	3	...
3	?	?	✓	4	?	?	✓	4	✓	?	?	2	...
4	?	✓	?	3	✓	?	?	2	?	?	✓	4	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	...
N	✓	?	?	2	✓	?	?	2	?	?	✓	3	...

Two Key Assumptions

Assumption (Random treatment timing)

Let $G = (G_1, ..., G_N)$. *Then* $\mathbb{P}(G = \tilde{g}) = (\prod_{g \in \mathcal{G}} N_g!)/N!$ *if* $\sum_i 1[\tilde{g}_i = g] = N_g$ *for all g, and zero otherwise.*

- Any permutation of treatment timing that preserves group size is equally likely

Assumption (No anticipation)

For all i,t and $g, g' > t$, $Y_{i,t}(g) = Y_{i,t}(g')$.

- No Anticipation may fail if treatment timing is announced in advance (Malani and Reif, 2015), as we have discussed this several times already!

Comments on random treatment timing

- Assumption 1 (Random treatment timing) is usually stronger than parallel trends.
  - At least for a fixed set of units.
- Thus, Assumption 1 may not be plausible in all observational settings where parallel trends is used.
- But random timing appears to be the justification for parallel trends in many cases.
- Without random timing, parallel trends will typically be sensitive to functional form as discussed in Extra Lecture 1 (Roth and Sant'Anna, 2023a).
- Roth and Sant'Anna (2023b) propose several pre-tests to assess the plausibility of the Random treatment timing assumptions.
  - See **R** package **staggered**. The function "balance_checks" check for the plausibility of random timing assumption.

Special Case: 2-Period Model

- Suppose $T = 2$ and $\mathcal{G} = \{2, \infty\}$, so some units are treated in period 2 and some are never treated.
- Under Randomization and No Anticipation, this is **analogous to a cross-sectional random experiment** with $Y_{i,t=2}$ the outcome and $Y_{i,t=1}$ playing the role of a fixed covariate.
  - $Y_i = Y_{i,t=2}$;
  - $X_i = Y_{i,t=1} \equiv Y_{i,t=1}(\infty)$
  - $D_i = 1[G_i = 2]$
- We will come back to this example throughout the talk to provide intuition and connect to the literature.

Causal parameters of interest

Estimands

- With staggered treatment timing, there are many possible causal estimands. Let's consider a flexible class of possible aggregations.
- **Building block:** let $\tau_{t,gg'}$ be average effect on outcome in period t of switching treatment from g' to g:

$$\tau_{t,gg'} \equiv ATE(g', g, t) = \frac{1}{N} \sum_i Y_{i,t}(g) - Y_{i,t}(g').$$

- This is why, in Lecture 1, we spent some time discussing this type of parameter!
- Consider a (scalar) estimand that aggregates this building blocks:

$$\theta = \sum_{t,gg'} a_{t,gg'} \tau_{t,gg'}$$

Special Case: 2-Period Model

- **Set-up:** Two periods ($T = 2$). Units treated in period 2 or never ($\mathcal{G} = \{2, \infty\}$)
- **Target parameter:** Average treatment effect (ATE) in period 2:

$$\theta = \tau_{2,2\infty} = \frac{1}{N} \sum_i Y_{i,t=2}(2) - Y_{i,t=2}(\infty)$$

Estimands in the Staggered Case

- In the staggered case, there are many possible ways of aggregating effects across cohorts and time periods.
- One useful parameter is $ATE(g,t)$, the average effect at time t of being treated at g relative to being never treated:

$$ATE(g,t) = \frac{1}{N}\sum_i Y_{i,t}(g) - Y_{i,t}(\infty).$$

- Following Callaway and Sant'Anna (2021), one might also be interested in summary parameters that are weighted averages of $ATE(g,t)$ along different dimensions.
- In Roth and Sant'Anna (2023b), we change the order of g and t; don't ask me why!

- The **simple weighted average** averages $ATE(g,t)$ across all (g,t) pairs with $t \geq g$:

$$\theta^{simple} = \frac{1}{\sum_t \sum_{g:g \leq t} N_g} \sum_t \sum_{g:g \leq t} N_g ATE(g,t).$$

- Let's define the **time-specific and cohort-specific weighted averages** as

$$\theta_t = \frac{1}{\sum_{g:g \leq t} N_g} \sum_{g:g \leq t} N_g ATE(g,t) \text{ and } \theta_g = \frac{1}{T-g+1} \sum_{t:t \geq g} ATE(g,t),$$

  and introduce the **calendar and cohort summary** parameters

$$\theta^{calendar} = \frac{1}{T} \sum_t \theta_t \text{ and } \theta^{cohort} = \frac{1}{\sum_{g:g \neq \infty} N_g} \sum_{g:g \neq \infty} N_g \theta_g.$$

- Finally, let's introduce **event-study parameters** that aggregate the treatment effects at a given lag l since treatment

$$\theta_l^{ES} = \frac{1}{\sum_{g:g+l \leq T} N_g} \sum_{g:g+l \leq T} N_g ATE(g+l,g).$$

Class of estimators we consider

Estimators

- Define $\widehat{\theta}_0$ to be the sample plug-in estimator for θ:

$$\widehat{\theta}_0 = \sum_{t,gg'} a_{t,gg'} \widehat{\tau}_{t,gg'},$$

  where $\widehat{\tau}_{t,gg'} = \overline{Y}_{tg} - \overline{Y}_{tg'}$ and $\overline{Y}_{tg}$ is the sample mean of $Y_{i,t}$ for cohort g.

- Let's will consider the class of estimators of the form

$$\widehat{\theta}_\beta = \widehat{\theta}_0 - \widehat{X}'\beta,$$

  where $\widehat{X}$ is a vector guaranteed to be mean-zero by No Anticipation.

- Formally, each element of $\widehat{X}$ aggregates differences between groups before either is treated

$$\widehat{X}_j = \sum_{(t,g,g'):g,g'>t} b^j_{t,gg'} \widehat{\tau}_{t,gg'}.$$

Example: 2 period Example

- Our proposed estimator is of the form

$$\begin{aligned}\widehat{\theta}_\beta &= \widehat{\theta}_0 - \beta\widehat{X} \\ &= \underbrace{(\overline{Y}_{22} - \overline{Y}_{2\infty})}_{\text{Post-treatment diff}} - \beta \underbrace{(\overline{Y}_{12} - \overline{Y}_{1\infty})}_{\text{Pre-treatment diff}}\end{aligned}$$

- The difference-in-differences estimator corresponds with $\beta = 1$.
- The simple difference-in-means estimator corresponds with $\beta = 0$
- For $\beta \in (0, 1)$, $\widehat{\theta}_\beta$ is a weighted average of DiD and DiM
- More generally, $\widehat{\theta}_\beta = \widehat{\theta}_0 - \widehat{X}'\beta$ is isomorphic to class of *regression adjusted estimators* in Freedman (2008b,a); Lin (2013), treating $Y_{i,t=1}$ as a fixed covariate.

Estimators in this class in the staggered case

- Several previously proposed estimators correspond with $\widehat{\theta}_1 = \widehat{\theta}_0 - \widehat{X}$ for an appropriately specificied $\widehat{\theta}_0$ and $\widehat{X}$.

- Callaway and Sant'Anna (2021) consider estimators that aggregate 2x2 diff-in-diff estimators:

$$\widehat{\tau}_W^{CS} = \sum_{t,g} w_{t,g} \left[\underbrace{(\overline{Y}_{t,g} - \overline{Y}_{t,\infty})}_{\text{Diff in period } t} - \underbrace{(\overline{Y}_{g-1,g} - \overline{Y}_{g-1,\infty})}_{\text{Diff in period } g-1} \right].$$

- This can be viewed as an estimator of the form $\widehat{\theta}_0 - \widehat{X}$, where

$$\widehat{\theta}_0 = \sum_{t,g} w_{t,g} \widehat{\tau}_{t,g\infty} \quad \text{and} \quad \widehat{X} = \sum_{t,g} w_{t,g} \widehat{\tau}_{g-1,g\infty}$$

Related Staggered Estimators

- Several variants to the Callaway and Sant'Anna (2021) estimator have been proposed that can likewise be cast into this class
- Callaway and Sant'Anna (2021) propose an alternative estimator using *not-yet-treated* instead of *never-treated* as the comparison
- Sun and Abraham (2021) propose a similar estimator using *last-to-be-treated* as the comparison
- de Chaisemartin and D'Haultfœuille (2020)'s estimator equivalent to Callaway and Sant'Anna (2021) estimator for a particular choice of weights, corresponding with event-study at lag 0

TWFE Estimators

- It is common to estimate the parameter $\widehat{\theta}^{TWFE}$ from the OLS regression:

$$Y_{i,t} = \alpha_i + \lambda_t + D_{i,t}\theta^{TWFE} + \epsilon_{i,t}$$

  where $D_{i,t} = 1[G_t \leq t]$ is an indicator for having received treatment.

- Athey and Imbens (2022) show that $\widehat{\theta}^{TWFE}$ can be written in the form:

$$\widehat{\theta}^{TWFE} = \sum_t \sum_{\substack{(g,g'):\\ min(g,g')\leq t}} \gamma_{t,gg'}\widehat{\tau}_{t,gg'} - \sum_t \sum_{\substack{(g,g'):\\ min(g,g')>t}} \gamma_{t,gg'}\widehat{\tau}_{t,gg'}$$

- Hence, the TWFE estimator can able be written in the form $\widehat{\theta}_0 - \widehat{X}$.

- However, the weights $\gamma_{t,gg'}$ may be negative, and so it is not obvious that the estimand θ is interesting under staggered treatment.

The efficient estimator

Unbiasedness

Proposition (Unbiasedness)

The estimator $\widehat{\theta}_\beta = \widehat{\theta}_0 - \widehat{X}'\beta$ is unbiased over the randomization distribution for any β,

$$\mathbb{E}\left[\widehat{\theta}_\beta\right] = \theta \text{ for all } \beta.$$

Proof sketch:

- By usual arguments, under randomization
$$\mathbb{E}\left[\overline{Y}_{t,g}\right] = \mathbb{E}_f\left[Y_{i,t}(g)\right].$$
- Since $\widehat{\theta}_0$ is a linear combination of sample means, it follows that $\mathbb{E}\left[\widehat{\theta}_0\right] = \theta$
- Likewise, $\mathbb{E}\left[\widehat{X}\right] = 0$ under no anticipation

Efficient Estimator

Proposition

The variance of $\widehat{\theta}_\beta = \widehat{\theta}_0 - \widehat{X}'\beta$ is uniquely minimized at

$$\beta^* = \underbrace{(\mathbb{V}ar\left[\widehat{X}\right])^{-1}}_{=V_{\widehat{X}}} \underbrace{Cov\left[\widehat{X}, \widehat{\theta}_0\right]}_{=V_{\widehat{X},\widehat{\theta}_0}}$$

if $V_{\widehat{X}}$ is positive definite.

Proof sketch: By definition,

$$\underset{\beta}{\text{argmin}}\, \mathbb{V}ar\left[\widehat{\theta}_0 - \widehat{X}'\beta\right] = \underset{\beta}{\text{argmin}}\, \mathbb{E}\left[\left((\widehat{\theta}_0 - \mathbb{E}\left[\widehat{\theta}_0\right]) - (\widehat{X} - \mathbb{E}\left[\widehat{X}\right])'\beta\right)^2\right].$$

This is just OLS! The solution follows immediately from the usual OLS formula.

Solving for the Variance

Recall that $\widehat{\theta}_0$ and $\widehat{X}$ are both linear functions of cohort sample means $\overline{Y}_g$.
Can write them as:

$$\widehat{\theta}_0 = \sum_g A_{\theta,g}\overline{Y}_g \text{ and } \widehat{X} = \sum_g A_{0,g}\overline{Y}_g.$$

Proposition

$$\mathbb{V}ar\left[\begin{pmatrix}\widehat{\theta}_0\\ \widehat{X}\end{pmatrix}\right] = \begin{pmatrix}\sum_g N_g{}^{-1} A_{\theta,g} S_g A'_{\theta,g} - N^{-1} S_\theta, & \sum_g N_g{}^{-1} A_{\theta,g} S_g A'_{0,g}\\ \sum_g N_g{}^{-1} A_{0,g} S_g A'_{\theta,g}, & \sum_g N_g{}^{-1} A_{0,g} S_g A'_{0,g}\end{pmatrix},$$

where $S_g = \mathbb{V}ar_f[Y_i(g)]$, $S_\theta = \mathbb{V}ar_f\left[\sum_g A_{\theta,g} Y_i(g)\right]$.

- Depends on estimable variances of potential outcomes (S_g), and non-estimable variances of treatment effects S_θ.
- But β^* depends only on estimable quantities, not on heterogeneous treatment effects.

Example: 2 period case

- Recall that

$$\begin{aligned}\widehat{\theta}_{\beta} &= \widehat{\theta}_0 - \beta\widehat{X} \\ &= \underbrace{(\overline{Y}_{22} - \overline{Y}_{2\infty})}_{\text{Post-treatment diff}} - \beta \underbrace{(\overline{Y}_{12} - \overline{Y}_{1\infty})}_{\text{Pre-treatment diff}}\end{aligned}$$

- Thus $\beta = 1$ is DiD and $\beta = 0$ is DiM
- Our results imply that the efficient β^* is equal to $\frac{N_\infty}{N}\beta_2 + \frac{N_2}{N}\beta_\infty$, where β_g is the coefficient from a regression of $Y_{i2}(g)$ on Y_{i1} (and a constant); see Lin (2013) as well.
- Intuitively, put more weight on lagged outcomes if they are more predictive of current outcomes:
  - If $Y_{i2}(g)$ and Y_{i1} are uncorrelated, then DiM ($\beta^* = 0$) is optimal;
  - If correlation is such that $\beta_2 = \beta_\infty = 1$, then DiD ($\beta^* = 1$) is optimal.

Properties of the Plug-In Efficient Estimator

The Plug-In Estimator

- So far we have solved for the efficient β^*, but it depends on the variances of potential outcomes S_g, which are typically **not known ex-ante**.
- Consider the feasible **plug-in efficient estimator** based on $\widehat{\beta}^*$, which replaces S_g with a sample analog $\widehat{S}_g$ in the expression for β^*.
  - $\widehat{S}_g = 1/(N_g - 1) \sum_i 1[G_i = g](Y_i - \overline{Y}_g)(Y_i - \overline{Y}_g)'$.
- Will show that in large populations the plug-in estimator $\widehat{\theta}_{\widehat{\beta}^*}$ has similar properties to the "oracle" estimator $\widehat{\theta}_{\beta^*}$.

Large population asymptotics

- Consider a sequence of populations in which N_g grows large for all g, satisfying certain regularity conditions

Assumption

(i) *Cohort shares converge to a constant:*

- *For all* $g \in \mathcal{G}$, $N_g/N \to p_g \in (0,1)$.

(ii) *Variances of potential outcomes converge to a constant:*

- *For all* g, g', S_g *and* $S_{gg'}$ *have limiting values denoted* S_g^* *and* $S_{gg'}^*$, *respectively, with* S_g^* *positive definite.*

(iii) *No individual dominates the variance of potential outcomes (Lindeberg-type condition):*

- $\max_{i,g} ||Y_i(g) - \overline{Y}(g)||^2/N \to 0$.

Asymptotic Properties of the Plug-In Estimator

- Under the given asymptotic conditions, the plug-in efficient estimator is asymptotically normally distributed with the same variance as the "oracle" efficient estimator.

Proposition

Under the given asymptotic conditions,

$$\sqrt{N}(\widehat{\theta}_{\widehat{\beta}^*} - \theta) \to_d \mathcal{N}\left(0, \sigma_*^2\right),$$

where

$$\sigma_*^2 = \lim_{N\to\infty} N\mathbb{V}ar\left[\widehat{\theta}_{\beta^*}\right].$$

Proof sketch: Apply CLT from Li and Ding (2017) to show that $(\widehat{\theta}_0, \widehat{X})$ asymptotically normal, then apply Slutsky's lemma plus convergence of $\widehat{\beta}^* \to_p \beta^*$

Inference

Covariance Estimation

- As is common in finite-population settings, the variance of $\widehat{\theta}_{\widehat{\beta}^*}$ can only be **estimated conservatively.**
- The issue is that the variance of $\widehat{\theta}_{\widehat{\beta}^*}$ contains the term $-S_\theta = -\mathbb{V}\text{ar}_f\left[\sum_g A_{\theta,g} Y_i(g)\right]$. This is not consistently estimable since it depends on covariances of potential outcomes that are never observed together.
- A natural conservative approach is the **Neyman-style variance estimate**, which ignores S_θ and replaces S_g with $\widehat{S}_g$ in the variance formula.
- Roth and Sant'Anna (2023b) shows that a **less conservative** variance estimator can be obtained by estimating the part of S_θ explained by $\widehat{X}$.

What about Fisher Randomization Tests

Fisher Randomization Tests

- An alternative approach to inference uses **Fisher Randomization Tests (FRTs)**
- We show that an FRT using a **studentized version of the efficient estimator** has dual advantages :
  1. has exact size under the sharp null of no treatment effects for all units;
  2. is asymptotically valid for the weak null that $\theta = 0$.
- Studentization is key!
  - In general, (un-studentized) FRTs may not have the correct size for such weak null hypotheses even asymptotically (Wu and Ding, 2020).
  - Roth and Sant'Anna (2023b) builds on Wu and Ding (2020) and Zhao and Ding (2020) to show that studentization bypasses this problem: FRT is asy. equiv. to testing that 0 falls within the t-based confidence interval CI_{**}

Fisher Randomization Tests

The following regularity condition imposes that the means of the potential outcomes have limits and that their fourth moment is bounded.

Assumption

Suppose that for all g, $\lim_{N\to\infty} \mathbb{E}_f[Y_i(g)] = \mu_g < \infty$*, and there exists* $L < \infty$ *such that* $N^{-1}\sum_i ||Y_i(g) - \mathbb{E}_f[Y_i(g)]||^4 < L$ *for all N.*

With this assumption in hand, we can make precise the sense in which the FRT is asymptotically valid under the weak null.

Proposition

Suppose Assumptions 1-4 hold. Let $t^\pi = (\widehat{\theta}^* / \widehat{se})^\pi$ *be the studentized statistic under permutation* π*. Then* $t^\pi \to_d \mathcal{N}(0, 1)$*,* P_G*-almost surely. Hence, if* p_{FRT} *is the p-value from the FRT associated with* $|t^\pi|$*, then under* $H_0 : \theta = 0$*,*

$$\lim_{N \to \infty} P(p_{FRT} \leq \alpha) \leq \alpha,$$

P_G*-almost surely, with equality if and only if* $S^*_\theta = 0$*.*

Fisher Randomization Tests for Callaway and Sant'Anna (2021)

- Nothing prevents us to adopt this randomization-based approach to inference for the Callaway and Sant'Anna (2021) estimators.
- Recall that it is a special case of Roth and Sant'Anna (2023b) when $\beta = 1$.
- Thus, one can use this approach to conduce "design-based" inference for Callaway and Sant'Anna (2021).
- Although, technically speaking, one (implicitly) imposes random treatment timing, Rambachan and Roth (2022)'s results suggest that the results are still valid (but conservative) without that assumption.

Simulations

Simulations

Monte Carlo 1

Monte Carlo 2

- Based on our application to the procedural justice training program for police officers studied in Wood et al. (2020a)
  - $Y_{i,t}$ is complaints against officer i in period t
  - G_i is the date on which officer i is first-trained
- Run simulations under sharp null in which $Y_{i,t}(g)$ is equal to the observed outcome in the data for all g
  - In paper, also consider specifications with heterogeneous treatment effects.
- Simulate G_i so that the number treated N_g match the data
  - We have 72 periods, 48 cohorts, 7785 officers. Cohort sizes range from 6 to 642.
- Compare plug-in efficient estimator to Callaway and Sant'Anna (2021) and Sun and Abraham (2021) estimators for several aggregation schemes (simple, calendar, cohort, and instantaneous event-study)

Results

Estimator	Estimand	Bias	Coverage	FRT Size	Mean SE	SD
PlugIn	calendar	0.00	0.93	0.06	0.27	0.29
PlugIn	cohort	0.00	0.92	0.06	0.24	0.24
PlugIn	ES0	0.01	0.94	0.05	0.26	0.27
PlugIn	simple	0.00	0.92	0.06	0.22	0.22
CS	calendar	0.00	0.94	0.05	0.55	0.55
CS	cohort	-0.01	0.95	0.05	0.41	0.41
CS/dCDH	ES0	0.01	0.94	0.07	0.36	0.36
CS	simple	-0.01	0.96	0.05	0.41	0.40
SA	calendar	0.06	0.93	0.04	1.30	1.30
SA	cohort	0.05	0.92	0.05	1.34	1.38
SA	ES0	0.03	0.94	0.06	0.83	0.89
SA	simple	0.06	0.92	0.04	1.46	1.49

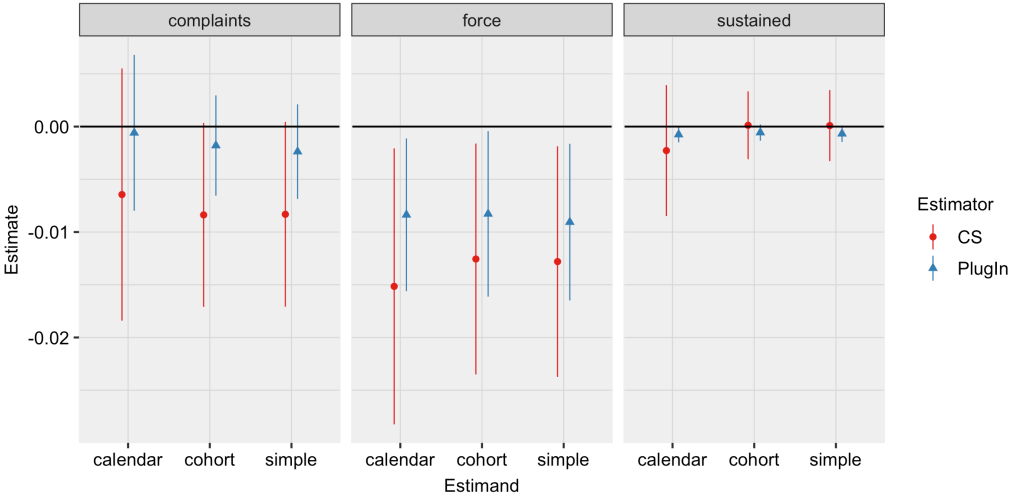
Results

	Ratio of SD to Plug-In	
Estimand	CS	SA
calendar	1.92	4.57
cohort	1.67	5.68
ES0	1.36	3.33
simple	1.82	6.76

Application

Application - Background

- Reducing police misconduct and use of force is an important policy objective.
- Wood et al. (2020a, PNAS) studied a randomized roll out of a procedural justice training program for police officers
  - Emphasized respect, neutrality, and transparency in the exercise of authority
- In Wood et al. (2020b), we re-analyzed data using the method of Callaway and Sant'Anna (2021)
  - No significant impacts on complaints; borderline significant effects on force; but CIs for all outcomes were wide



Population: Chicago Police officers that are not on special forces (excluding those selected into the pilot program)

			Plug-In				CS				
Outcome	Estimand	Pre-treat Mean	Estimate	LB	UB	p-val (FRT)	Estimate	LB	UB	p-val (FRT)	CI Ratio
complaints	simple	0.052	–5%	–13%	4%	0.29	–16%	–33%	1%	0.06	2.0
complaints	calendar	0.052	–1%	–15%	13%	0.89	–12%	–35%	11%	0.28	1.6
complaints	cohort	0.052	–3%	–13%	6%	0.47	–16%	–33%	1%	0.06	1.8
sustained	simple	0.004	–15%	–33%	2%	0.09	2%	–75%	79%	0.95	4.3
sustained	calendar	0.004	–17%	–34%	–0%	0.04	–52%	–194%	90%	0.50	8.4
sustained	cohort	0.004	–13%	–31%	5%	0.16	3%	–71%	76%	0.93	4.2
force	simple	0.051	–18%	–32%	–3%	0.03	–25%	–46%	–4%	0.03	1.5
force	calendar	0.051	–16%	–30%	–2%	0.02	–30%	–55%	–4%	0.02	1.8
force	cohort	0.051	–16%	–31%	–1%	0.08	–24%	–46%	–3%	0.04	1.4

Table 4: Estimates and 95% CIs as a Percentage of Pre-treatment Means

Can we assess the validity of the assumptions?

		Main Estimation Sample					Including pilot + special				
Outcome	Estimand	Xhat	t-stat	p-val	p-val (FRT)	Joint p-val (FRT)	Xhat	t-stat	p-val	p-val (FRT)	Joint p-val (FRT)
complaints	Simple	0.007	1.55	0.12	0.12	0.15	0.005	1.22	0.22	0.21	0.44
complaints	Cohort	0.008	1.76	0.08	0.08	0.15	0.004	1.04	0.30	0.33	0.44
complaints	Calendar	0.006	1.22	0.22	0.22	0.15	0.010	1.30	0.19	0.24	0.44
complaints	ES0	0.004	1.27	0.21	0.18	0.15	0.003	1.03	0.30	0.31	0.44
sustained	Simple	−0.001	0.46	0.64	0.64	0.89	−0.001	0.97	0.33	0.34	0.55
sustained	Cohort	−0.001	0.43	0.67	0.67	0.89	−0.002	1.03	0.30	0.31	0.55
sustained	Calendar	0.002	0.48	0.63	0.68	0.89	−0.001	0.42	0.68	0.73	0.55
sustained	ES0	0.000	0.23	0.82	0.82	0.89	0.000	0.32	0.75	0.76	0.55
force	Simple	0.005	0.91	0.36	0.37	0.36	0.005	1.04	0.30	0.33	0.02
force	Cohort	0.006	1.10	0.27	0.27	0.36	0.004	0.91	0.36	0.39	0.02
force	Calendar	0.008	1.22	0.22	0.21	0.36	0.013	1.59	0.11	0.15	0.02
force	ES0	0.005	1.28	0.20	0.21	0.36	0.008	2.91	0.00	0.00	0.02

Take-way message

Conclusion

- When treatment timing is random, classical DiD estimators "leave too much money on the table".
- Roth and Sant'Anna (2023b) show how you use additional information to "collect" the money!
- Estimators and inference procedures can easily used in **R**, via the **staggered** package.
- I recommend this approach when treatment timing is (quasi-) random.
  - But note other procedures are valid under the weaker assumption of parallel trends!
  - Gains are coming from additional rationalization/structure of the problem!

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge, "Sampling-Based versus Design-Based Uncertainty in Regression Analysis," *Econometrica*, 2020, *88* (1), 265–296.

Athey, Susan and Guido Imbens, "Design-Based Analysis in Difference-In-Differences Settings with Staggered Adoption," *Journal of Econometrics*, 2022, *226* (1), 62–79.

Bojinov, Iavor, Ashesh Rambachan, and Neil Shephard, "Panel experiments and dynamic causal effects: A finite population perspective," *Quantitative Economics*, 2021, *12* (4), 1171–1196.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

de Chaisemartin, Clément and Xavier D'Haultfœuille, "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 2020, *110* (9), 2964–2996.

Deshpande, Manasi and Yue Li, "Who Is Screened Out? Application Costs and the Targeting of Disability Programs," *American Economic Journal: Economic Policy*, November 2019, *11* (4), 213–248.

Druedahl, Jeppe and Alessandro Martinello, "Long-Run Saving Dynamics: Evidence from Unexpected Inheritances," 2019, p. 79.

Fadlon, Itzik and Torben Heien Nielsen, "Family Labor Supply Responses to Severe Health Shocks: Evidence from Danish Administrative Records," *American Economic Journal: Applied Economics*, July 2021, *13* (3), 1–30.

Freedman, David A., "On Regression Adjustments in Experiments with Several Treatments," *The Annals of Applied Statistics*, 2008, *2* (1), 176–196.

_ , "On regression adjustments to experimental data," *Advances in Applied Mathematics*, 2008, *40* (2), 180–193.

Imbens, Guido W. and Donald B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, 1 edition ed., New York: Cambridge University Press, April 2015.

Li, Xinran and Peng Ding, "General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference," *Journal of the American Statistical Association*, October 2017, *112* (520), 1759–1769.

Lin, Winston, "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique," *Annals of Applied Statistics*, March 2013, *7* (1), 295–318.

Malani, Anup and Julian Reif, "Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform," *Journal of Public Economics*, 2015, *124*, 1–17.

Manski, Charles F. and John V. Pepper, "How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions," *Review of Economics*

and Statistics, 2018, *100* (2), 232–244.

Nekoei, Arash and David Seim, "How do Inheritances Shape Wealth Inequality? Theory and Evidence from Sweden," SSRN Scholarly Paper ID 3192778, Social Science Research Network, Rochester, NY March 2019.

Parker, Jonathan A, Nicholas S Souleles, David S Johnson, and Robert McClelland, "Consumer Spending and the Economic Stimulus Payments of 2008," *American Economic Review*, October 2013, *103* (6), 2530–2553.

Rambachan, Ashesh and Jonathan Roth, "Design-Based Uncertainty for Quasi-Experiments," *arXiv:2008.00602*, 2022.

Roth, Jonathan and Pedro H. C. Sant'Anna, "When Is Parallel Trends Sensitive to Functional Form?," *Econometrica*, 2023, *91* (2), 737–747.

_ and Pedro H.C. Sant'Anna, "Efficient Estimation for Staggered Rollout Designs," *Journal of Political Economy: Microeconomics*, 2023, *1* (4), 669––709.

Sun, Liyan and Sarah Abraham, "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 2021, *225* (2).

Wood, George, Tom R. Tyler, and Andrew V. Papachristos, "Procedural justice training reduces police use of force and complaints against officers," *Proceedings of the National Academy of Sciences*, May 2020, *117* (18), 9815–9821.

_ , _ , _ , Jonathan Roth, and Pedro H.C. Sant'Anna, "Revised Findings for "Procedural justice training reduces police use of force and complaints against officers"," *Working Paper*, 2020.

Wu, Jason and Peng Ding, "Randomization Tests for Weak Null Hypotheses in Randomized Experiments," *Journal of the American Statistical Association*, 2020, pp. 1–16.

Xu, Ruonan, "Potential outcomes and finite-population inference for M-estimators," *Econometrics Journal*, 2021, (Forthcoming).

Zhao, Anqi and Peng Ding, "Covariate-adjusted Fisher randomization tests for the average treatment effect," *arXiv:2010.14555 [math, stat]*, November 2020. arXiv: 2010.14555.

Monte Carlo Experiment 1

- Based on a two-period running example.
- Draw $Y_i(\infty) \sim \mathcal{N}\left(0, \Sigma_\rho\right)$, where Σ_ρ has 1's on diagonal and ρ's off diagonal.
  - ρ governs the serial correlation of POs
- Set $Y_{i,t=2}(2) = Y_{i,t=2}(\infty) + \gamma(Y_{i,t=2}(\infty) - \mathbb{E}_f\left[Y_{i,t=2}(\infty)\right])$
  - γ governs the heterogeneity of treatment effects
- We hold the POs fixed across sims, and re-draw treatment assignment across simulation draws so that N_2 are treated in period 2 and N_∞ are never treated.
- Compare performance of plug-in efficient estimator ($\widehat{\theta}_{\widehat{\beta}^*}$), simple difference in means ($\widehat{\theta}_0$), and difference-in-differences ($\widehat{\theta}_1$)
- Use $\rho \in \{0, .5, .99\}$, $\gamma \in \{0, .5\}$ and $N_2 = N_\infty \in \{25, 1000\}$. Back

Results

				Bias			SD			Coverage			FRT Size	
N_1	N_0	rho	gamma	PlugIn	DiD	DiM	PlugIn	DiD	DiM	PlugIn	DiD	DiM	PlugIn	DiD
1000	1000	0.99	0.0	0.00	0.00	−0.00	0.01	0.01	0.04	0.95	0.95	0.95	0.05	0.05
1000	1000	0.99	0.5	0.00	0.00	−0.00	0.01	0.01	0.06	0.95	0.95	0.95	0.04	0.06
1000	1000	0.50	0.0	0.00	0.00	0.00	0.04	0.04	0.05	0.94	0.95	0.94	0.06	0.05
1000	1000	0.50	0.5	0.00	0.00	0.00	0.05	0.05	0.06	0.95	0.95	0.95	0.06	0.05
1000	1000	0.00	0.0	−0.00	0.00	−0.00	0.04	0.07	0.04	0.95	0.94	0.95	0.05	0.06
1000	1000	0.00	0.5	−0.00	0.00	−0.00	0.06	0.07	0.06	0.95	0.95	0.95	0.04	0.05
25	25	0.99	0.0	0.00	0.00	−0.03	0.04	0.04	0.27	0.94	0.94	0.94	0.04	0.05
25	25	0.99	0.5	0.00	−0.01	−0.04	0.05	0.08	0.34	0.92	0.93	0.93	0.06	0.06
25	25	0.50	0.0	−0.01	0.02	−0.02	0.24	0.29	0.26	0.94	0.95	0.94	0.04	0.04
25	25	0.50	0.5	−0.01	0.01	−0.03	0.30	0.32	0.33	0.94	0.95	0.94	0.04	0.04
25	25	0.00	0.0	−0.03	−0.02	−0.03	0.28	0.38	0.27	0.93	0.95	0.93	0.06	0.04
25	25	0.00	0.5	−0.04	−0.02	−0.04	0.35	0.42	0.34	0.93	0.94	0.94	0.06	0.05

					SD Relative to Plug-In		
N_1	N_0	rho	gamma	betastar	PlugIn	DiD	DiM
1000	1000	0.99	0.0	0.99	1.00	1.00	7.09
1000	1000	0.99	0.5	1.24	1.00	1.71	7.07
1000	1000	0.50	0.0	0.52	1.00	1.13	1.15
1000	1000	0.50	0.5	0.65	1.00	1.04	1.15
1000	1000	0.00	0.0	−0.03	1.00	1.45	1.00
1000	1000	0.00	0.5	−0.03	1.00	1.31	1.00
25	25	0.99	0.0	0.97	1.00	0.99	6.58
25	25	0.99	0.5	1.22	1.00	1.47	6.31
25	25	0.50	0.0	0.41	1.00	1.21	1.10
25	25	0.50	0.5	0.51	1.00	1.08	1.10
25	25	0.00	0.0	0.10	1.00	1.35	0.98
25	25	0.00	0.5	0.13	1.00	1.22	0.98

- Can be as much as 1.7 or 7 times more efficient!