Getting More from Summary Statistics in Online Experiments: Inference on a New Class of Sample Average Treatment Effects

Jasjeet Sekhon

joint work with Yotam Shem-Tov

UC Berkeley

October 26, 2017

Motivation

- With heterogeneity, the average treatment effect is *not* sufficient to evaluate the impacts of an intervention
- Going beyond the mean is difficult:
  - Rank tests and quantile regression have no clear interpretation
    $\Rightarrow$ What is the estimand they identify?
  - Machine learning methods require complex computations
- We can gain efficiency by changing the estimand, even asymptotically

Potential Outcomes Framework

- A fixed population of N units
- A binary treatment is randomly assigned
- Each unit has two potential outcomes:

$$(Y(1), Y(0)) \perp\!\!\!\perp T \in \{0, 1\}$$

- The potential outcomes are fixed (not random variables)
- Let τ_i denote the treatment effect on unit i:

$$\tau_i = Y_i(1) - Y_i(0)$$

- T is the only random component in this data generating process

Potential estimands

1. The sample average treatment effect (**SATE**):

$$\textbf{SATE} = \frac{1}{N} \cdot \sum_{i=1}^{N} \tau_i$$

2. The sample average treatment effect on the treated (**SATT**):

$$\textbf{SATT} = \frac{1}{m} \cdot \sum_{i=1}^{N} \tau_i \cdot T_i, \quad \text{where} \quad m = \sum_{i=1}^{N} T_i$$

3. The sample average treatment effect on the control (**SATC**):

$$\textbf{SATC} = \frac{1}{N-m} \cdot \sum_{i=1}^{N} \tau_i \cdot (1 - T_i)$$

Efficiency gains in actual experiments

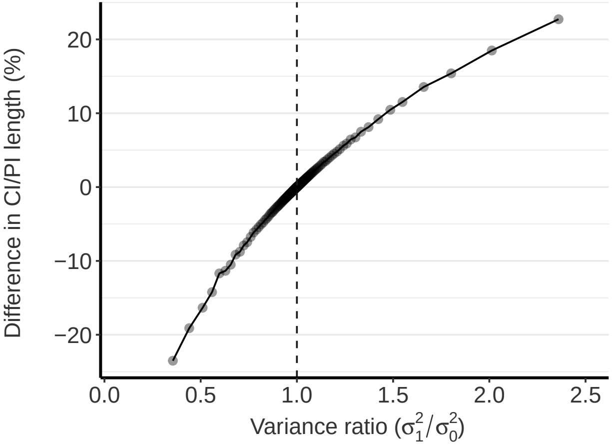


Figure: CI / PI length gains of SATT vs. SATE

What we do

- We generalize Robins (1988) results for non-binary outcomes
- We derive general variance formulas for inference on a new class of estimands:

$$\omega \cdot \textbf{SATT} + (1-\omega) \cdot \textbf{SATC}$$

- Theoretical results (e.g., CLTs) on how to conduct non-parametric inference on a new and general class of estimands
- CI for **SATE** will not have correct coverage of **SATT** or **SATC**
- We provide inference for the estimand that can be estimated most accurately

Outline

1. Inference on **SATE**

2. Inference on **SATT** and a comparison to **SATE**

3. A new class of estimands: The Sample Average Treatment Effect Optimal (**SATO**)

4. Conclusions

Inference on SATE

- The variance of $\left(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_0 - \textbf{SATE}\right)$ is:

$$\underbrace{\frac{\sigma_0^2}{N(1-p)} + \frac{\sigma_1^2}{Np}}_{\text{Neyman's variance estimator}} - \frac{\sigma_\tau^2}{N}$$

  - $p = \Pr(T = 1)$
  - σ_0^2 - variance of $Y(0)$
  - σ_1^2 - variance of $Y(1)$
  - σ_τ^2 - variance of $Y(1) - Y(0) \Rightarrow \sigma_\tau^2 = \sigma_0^2 + \sigma_1^2 - 2\sigma_1\sigma_0\rho$

$$\rho = \text{Corr}\left(Y(1), Y(0)\right)$$

  cannot be identified, and must be bounded

- Inference on SATE is *conservative*

Inference on SATT (and SATC)

- The variance of $\left(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_0 - \textbf{SATT}\right)$ is:

$$\frac{1}{N \cdot (1-p) \cdot p} \cdot \sigma_0^2$$

$\Rightarrow$ Var $\left(\bar{Y}_1 - \bar{Y}_0 - \textbf{SATT}\right)$ is *independent* of ρ

- Inference on **SATT** can be done using a consistent non-conservative variance estimator

Lemma (Decomposition of $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_0$)

The difference-in-means can be decomposed to:

$$\frac{N}{m \cdot (N-m)} \cdot \sum_{i=1}^{N} Y_i(0) \cdot T_i - \frac{1}{N-m} \cdot \sum_{i=1}^{N} Y_i(0) + \textbf{\textit{SATT}}$$

Inference on SATT (and SATC)

- The variance of $\left(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_0 - \textbf{SATT}\right)$ is:

$$\frac{1}{N \cdot (1-p) \cdot p} \cdot \sigma_0^2$$

$\Rightarrow$ Var $\left(\bar{Y}_1 - \bar{Y}_0 - \textbf{SATT}\right)$ is *independent* of ρ

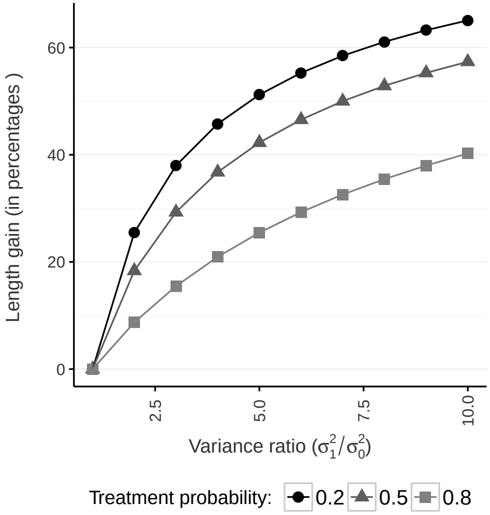
- Inference on **SATT** can be done using a consistent non-conservative variance estimator

Lemma (Decomposition of $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_0$)

The difference-in-means can be decomposed to:

$$\frac{N}{m \cdot (N-m)} \cdot \sum_{i=1}^{N} Y_i(0) \cdot T_i - \frac{1}{N-m} \cdot \sum_{i=1}^{N} Y_i(0) + \textbf{\textit{SATT}}$$

Change of estimand: Efficiency gains relative to the standard benchmark (Neyman's variance estimator)



The estimand that maximizes accuracy (SATO)

- Sample Average Treatment Effect Optimal (**SATO**) is the estimand that maximizes accuracy given the difference-in-means test statistic:

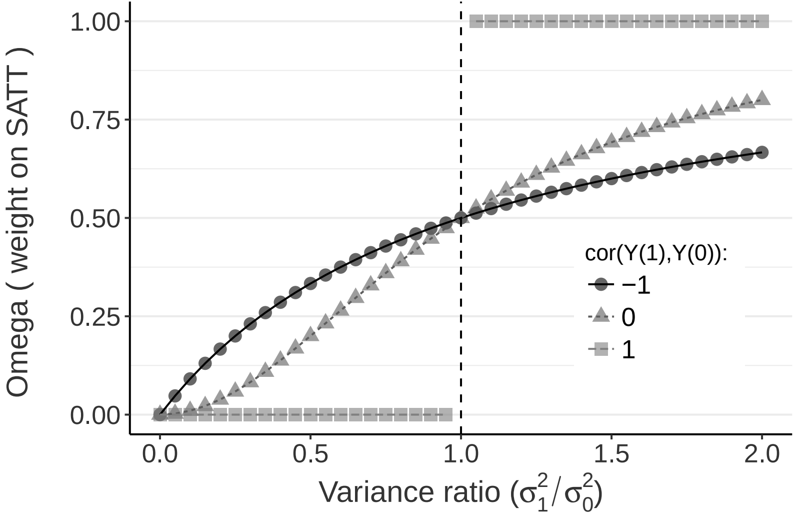
$$\textbf{SATO} \equiv \omega^* \cdot \textbf{SATT} + (1 - \omega^*) \cdot \textbf{SATC}$$
$$s.t$$
$$\omega^* = \underset{\omega}{\text{argmin}} \text{ Var}\left(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_0 - \textbf{SATO}\right)$$

- The optimal ω weight is:

$$\omega^* = \frac{\left(\frac{\sigma_1}{\sigma_0}\right)^2 - \rho \cdot \frac{\sigma_1}{\sigma_0}}{\left(\frac{\sigma_1}{\sigma_0}\right)^2 + 1 - 2\rho\left(\frac{\sigma_1}{\sigma_0}\right)}$$

- Inference on **SATO** is generally *more* efficient than **SATE**

Optimal ω for different $\frac{\sigma_1}{\sigma_0}$ and ρ



Efficiency gains in actual experiments

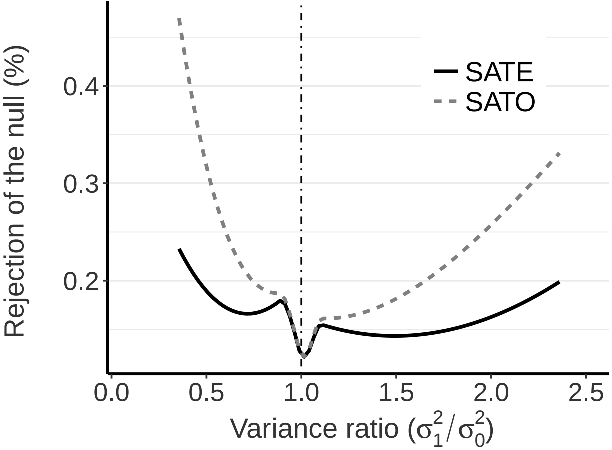


Figure: Rejection rate

Conclusions

- We derive a unified framework for identifying and estimating average treatment effects
- Implementation requires *only* aggregate data (e.g., $\hat{\bar{Y}}_1$, $\hat{\sigma}_0^2$)
  $\Rightarrow$ Ideal for online platforms that run thousands of experiments
- Combining these results with sequential testing
- ω^* is *independent* of p unlike **SATE** ($\omega = p$)
- **SATE** is equal to **SATO** under a constant treatment effect model

Additional slides

Monte Carlo simulations

1. Random coefficient data generating process:

$$
\begin{aligned}
Y_i(0) &\sim N(\mu = 10,\ \sigma_0^2 = 1) \\
\tau_i &\sim N(\mu = 0,\ \sigma_\tau^2) \\
Y_i(1) &= \tau_i + Y_i(1)
\end{aligned}
$$

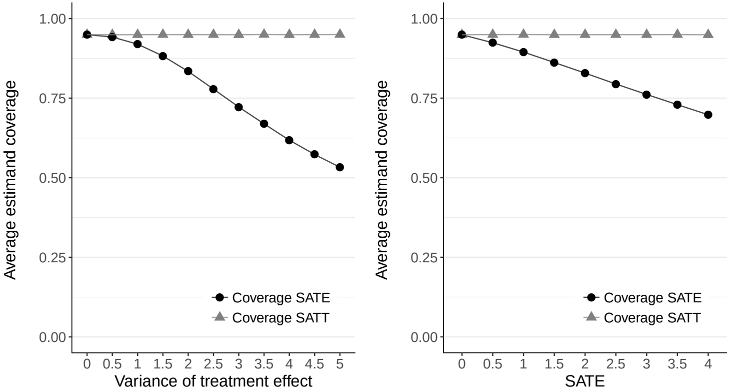
2. Tobit data generating process:

$$
Y(1) = \begin{cases} Y(0) + \tau, & Y(0) \geq 0 \\ Y(0), & Y(0) < 0 \end{cases} \qquad \text{and} \quad \tau > 0
$$

Inference on **SATE** relative to **SATT**:

- CI/PI length (efficiency)
- Coverage (Type-I error)

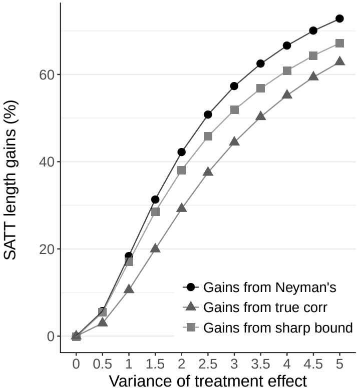
Figure: Coverage (Type-I error rate) of SATE and SATT when using a PI for SATT



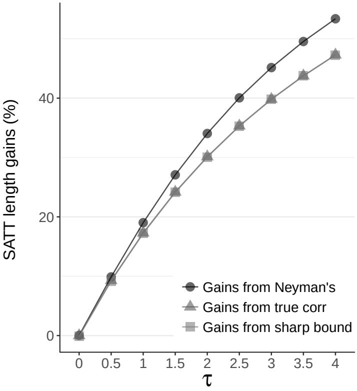
: Random coefficient

: Tobit

Figure: Confidence Interval/Prediction Interval length



: Random coefficient



: Tobit

Heterogeneity in estimated treatment effects

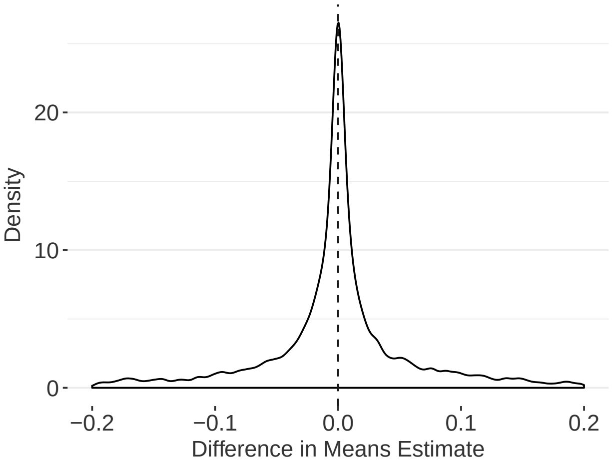


Figure: Confidence intervals for average treatment effects (using data from Tunca and Egeli, 1996)

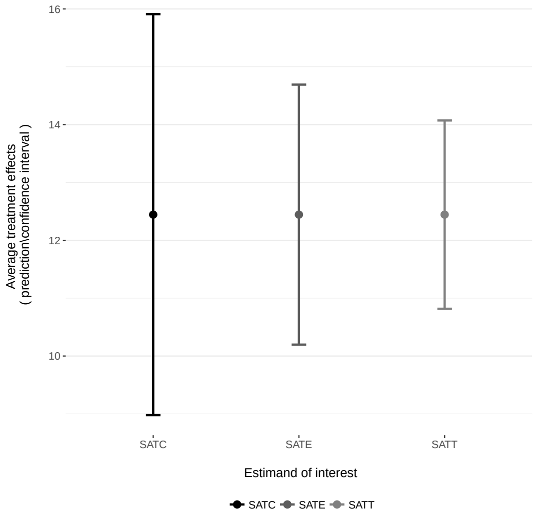


Table: **Example for when the SATT can substantially differ from the SATE**

Unit	$Y(1)$	$Y(0)$
1	1	0
2	-1	0
3	-100	0
4	100	0

A comparison of SATE and SATT when ρ is *known*

Theorem

For all σ_0 and σ_1 such that $\sigma_0 < \sigma_1$:

1. *There exists a threshold level of ρ, $\bar{\rho}$ such that:*

$$\rho \leq \bar{\rho} \Rightarrow \mathit{Var}\left(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_0 - \textbf{\textit{SATE}}\right) \leq \mathit{Var}\left(\bar{Y}_1 - \bar{Y}_0 - \textbf{\textit{SATT}}\right)$$

$$\rho > \bar{\rho} \Rightarrow \mathit{Var}\left(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_0 - \textbf{\textit{SATE}}\right) > \mathit{Var}\left(\bar{Y}_1 - \bar{Y}_0 - \textbf{\textit{SATT}}\right)$$

2. *When $\frac{\sigma_1}{\sigma_0} > \sqrt{\frac{1-p^2}{(1-p)^2}}$ then, $\bar{\rho} < 0$.*

We can empirically test whether $\bar{\rho}$ is negative:

$$H_0: \frac{\sigma_1}{\sigma_0} \leq \sqrt{\frac{1-p^2}{(1-p)^2}},$$

$\Rightarrow$ if the null is rejected, then $\bar{\rho} < 0$

Estimating SATE vs. SATT when ρ is *unknown*

- The classic variance estimator for Var $\left(\bar{Y}_1 - \bar{Y}_0 - \textbf{SATE}\right)$ is:

$$\mathbb{V}_{\text{Neyman}} = \frac{1}{m}\sigma_1^2 + \frac{1}{N-m}\sigma_0^2$$

- More efficient estimators exist by bounding ρ (e.g., Arronow et al., 2014)

Theorem

When $\sigma_1 \neq \sigma_0$, a prediction interval for either ***SATT*** *or* ***SATC*** *will be shorter than a confidence interval for the* ***SATE*** *using $\mathbb{V}_{Neyman}$.*

- The gain in terms of interval length (in %) is:

$$1 - \frac{1}{\sqrt{\left(\frac{\sigma_1^2}{\sigma_0^2}(1-p) + p\right)}}$$

Estimating SATE vs. SATT when ρ is *unknown*

- The classic variance estimator for Var $\left(\bar{Y}_1 - \bar{Y}_0 - \textbf{SATE}\right)$ is:

$$\mathbb{V}_{\text{Neyman}} = \frac{1}{m}\sigma_1^2 + \frac{1}{N-m}\sigma_0^2$$

- More efficient estimators exist by bounding ρ (e.g., Arronow et al., 2014)

Theorem

*When $\sigma_1 \neq \sigma_0$, a prediction interval for either **SATT**or **SATC**will be shorter than a confidence interval for the **SATE**using $\mathbb{V}_{\textit{Neyman}}$.*

- The gain in terms of interval length (in %) is:

$$1 - \frac{1}{\sqrt{\left(\frac{\sigma_1^2}{\sigma_0^2}(1-p) + p\right)}}$$