Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Inference on a New Class of Sample Average Treatment Effects

Jasjeet S. Sekhon & Yotam Shem-Tov

To cite this article: Jasjeet S. Sekhon & Yotam Shem-Tov (2020): Inference on a New Class of Sample Average Treatment Effects, Journal of the American Statistical Association, DOI: 10.1080/01621459.2020.1730854

To link to this article: https://doi.org/10.1080/01621459.2020.1730854

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 00, NO. 0, 1–7: Theory and Methods
https://doi.org/10.1080/01621459.2020.1730854

Inference on a New Class of Sample Average Treatment Effects

Jasjeet S. Sekhon[a] and Yotam Shem-Tov[b]

[a]Department of Statistics & Data Science and Department of Political Science, Yale University, New Haven, CT; [b]Department of Economics, University of California at Los Angeles, Los Angeles, CA

ABSTRACT

We derive new variance formulas for inference on a general class of estimands of causal average treatment effects in a randomized control trial. We generalize the seminal work of Robins and show that when the researcher's objective is inference on sample average treatment effect of the treated (SATT), a consistent variance estimator exists. Although this estimand is equal to the sample average treatment effect (SATE) in expectation, potentially large differences in both accuracy and coverage can occur by the change of estimand, even asymptotically. Inference on SATE, even using a conservative confidence interval, provides incorrect coverage of SATT. We demonstrate the applicability of the new theoretical results using an empirical application with hundreds of online experiments with an average sample size of approximately 100 million observations per experiment. An R package, `estCI`, that implements all the proposed estimation procedures is available. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received November 2017
Accepted February 2020

KEYWORDS
Average treatment effect; Causality; Neyman causal model

1. Introduction

The Neyman variance estimator is the most commonly used variance estimator in randomized experiments (Imbens and Rubin 2015). Under the super-population model, it is a consistent estimator for the variance of the difference-in-means, and this probably accounts for its popularity. However, under Neyman's finite population model, a consistent variance estimator for the difference-in-means does not exist (Neyman 1923/1990), and Neyman's variance estimator is conservative. Sharper, albeit still conservative, variance estimators exist (Aronow, Green, and Lee 2014), but they are not often used.

This article develops new limiting distribution results for inference on sample average treatment effects. To estimate the sample average treatment effect (SATE), researchers use the limiting distribution of the difference-in-means recentered around SATE. We show that changing the estimand and recentering the difference-in-means with respect to sample average treatment effect for the treated (SATT) allows one to obtain a consistent nonconservative variance estimator. The key result of the article is the derivation of valid inference on SATT that generalizes Robins (1988) seminal work.

Inference on SATT yields a prediction interval (PI) that has correct coverage (of SATT) and can potentially be substantially different in length, than a confidence interval (CI) for SATE. It follows that inference on SATE has incorrect coverage and/or is inefficient for the estimation of SATT.[1] The heterogeneity in the response of different units to the treatment is the driving force behind the potential accuracy differences. The change of estimand makes the recentered difference-in-means sensitive to differences in the variance of the treated and control units, and not only to the mean impact of the treatment.

Our possibly surprising results do have an intuitive interpretation. While in the case of the super-population model the Population Average Treatment Effect (PATE) and the Population Average Treatment Effect of the Treated (PATT) are equal.[2] In the case of Neyman's finite population model, the two estimands differ and each recentering choice yields a different variance expression for the recentered difference-in-means. Accuracy differences between inference on SATT relative to inference on SATE come from two channels. First, in the case of SATT, one does not need to use conservative bounds for the variance estimator as it is point-identified. Second, the change of estimand from SATE to SATT changes the variance of the recentered difference-in-means. We discuss and decompose the conditions for accuracy gains from each one of these channels. Robins (1988) studied this phenomenon for the special case of binary outcomes. He emphasized that a CI for SATE does not yield correct coverage of SATT. We extend his result in various ways, including by providing results for nonbinary outcomes and providing conditions under which a PI for SATT yields gains in accuracy.

In general, PIs for SATT are not guaranteed to have correct coverage of SATE. We provide analytical and simulation based evidence for when a PI for SATT has approximately correct coverage of SATE and when it does not. The key factor behind the differences is the variance of the treatment effect distribution. As

CONTACT Jasjeet S. Sekhon sekhon@berkeley.edu Department of Statistics & Data Science and Department of Political Science, 24 Hillhouse Ave, New Haven, CT 06511.

A previous version of this article was entitled "Efficient Estimation of Average Treatment Effects under Effect Heterogeneity."

The R package, `estCI`, that implements the estimation methods described in the article is available at https://github.com/yotamshemtov/estCI.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

[1]This holds even though SATE and SATT are equal in expectation.

[2]The terms SATE, SATT, PATE, and PATT have been first used, to our knowledge, by Imbens (2004).

the impacts of the treatment are more heterogeneous, inference on SATE differs from inference on SATT or sample average treatment effect for the controls (SATC). An example of when SATT can be the estimand of interest is an attributable treatment effect model in which the treatment effect varies across units (Rosenbaum 2001; Feng et al. 2014; Keele, Small, and Grieve 2017).

Our results also extend Rigdon and Hudgens (2015), who showed that PIs for SATT can be combined to construct a CI for SATE in the context of binary outcomes. We generalize their result to the case of nonbinary outcomes. In addition, we show that when using the difference-in-means test statistic, combining PIs for SATT and SATC yields the exact same conservative CI for SATE as one would have gotten by using a bound for the variance of the difference-in-means, which assumes that the correlation between potential outcomes is one. More efficient CIs can be constructed by directly using sharper bounds on the unobserved correlation between potential outcomes (Aronow, Green, and Lee 2014). Taken together, our results prove that as long as the test statistic is the difference-in-means, combining separate PIs for SATT and SATC, which have been derived in a nonconservative procedure, does *not* allow one to construct a CI for SATE that is more efficient than existing procedures that use conservative variance estimators and directly conduct inference on SATE.[3]

The remainder of this article is organized as follows. Section 2 describes the theoretical framework, definitions, and notation that are used throughout the article. Section 3 describes the key theoretical results of the article. Section 4 provides Monte Carlo simulations from several data generating processes. Section 5 presents notes and remarks on the theoretical results and possible extensions. Section 6 discusses an empirical data application that consists of hundreds of online experiments with millions of observations. Section 7 concludes.

2. Setting, Definitions, and Notation

We follow Neyman's finite population causal model. Let Y_i be the outcome of interest for unit i and consider a fixed finite population of $N \geq 4$ units and two dimensions, $Y(0)$ and $Y(1)$:

$$\Pi_N = \big\{(Y(0)_{1N}, Y(1)_{1N}),\ (Y(0)_{2N}, Y(1)_{2N}), \ldots, (Y(0)_{NN}, Y(1)_{NN})\big\}. \tag{1}$$

A random sample of m units are assigned to the treatment regime: $\Pi_N^1 = \{Y(1)_{1N}, Y(1)_{2N}, \ldots, Y(1)_{NN}\}$ and the vector of treatment indicators, $\boldsymbol{T} = (T_1, \ldots, T_N)$, denotes the m units allocated to the treatment group.[4] The remaining $N - m$ units are assigned to the control group and they form a random sample of $N - m$ units from the finite population: $\Pi_N^0 = \{Y(0)_{1N}, Y(0)_{2N}, \ldots, Y(0)_{NN}\}$, which is represented by the vector of indicators, $(1 - T_1, \ldots, 1 - T_N)$. Thus, the probability of each unit being assigned to the treatment regime is $p = \frac{m}{N}$. Finally, we also assume that SUTVA (Holland 1986) is satisfied: $Y_i(\boldsymbol{T}) = Y_i(T_i)$.

Let τ_i denote the effect of the treatment on unit i (i.e., $Y_i(1) - Y_i(0)$) and let the vector of treatment effects be denoted by $\boldsymbol{\tau} = \boldsymbol{Y(1)} - \boldsymbol{Y(0)}$. The researcher can be interested in conducting inference on several possible average treatment effect estimands:

$$\text{SATE} \equiv \frac{1}{N} \cdot \sum_{i=1}^{N} \tau_i,\ \text{SATT} \equiv \frac{1}{m} \cdot \sum_{i=1}^{N} \tau_i \cdot T_i,$$
$$\text{SATC} \equiv \frac{1}{N-m} \cdot \sum_{i=1}^{N} \tau_i \cdot (1 - T_i). \tag{2}$$

We do not impose any parametric assumptions on the relationship between $Y(1)$ and $Y(0)$, and allow τ to vary across units. It is important to note that unlike SATE, both SATT and SATC are random variables, even conditional on the sample, as they are a function of $\boldsymbol{T}$. Since SATT is a random variable, a random interval that contains SATT with a probability of $1-\alpha$ is usually referred to as a PI or, in Bayesian terminology, a credible interval, rather than a CI. We are not the first ones to discuss inference on a causal estimand that is a random variable (Robins 1988; Bowers and Hansen 2009; Crump et al. 2009; Li, Morgan, and Zaslavsky 2018).

3. Theory

3.1. Inference on SATT (and SATC) Relative to SATE

The classic estimator for SATE (and SATT) is the difference-in-means between the outcomes of units under the two treatment regimes. For notational convenience, denote the difference-in-means estimator by t_{diff}:

$$t_{\text{diff}} \equiv \frac{1}{m} \sum_{i=1}^{N} Y_i \cdot T_i - \frac{1}{N-m} \sum_{i=1}^{N} Y_i \cdot (1 - T_i). \tag{3}$$

Lemma 3.1 shows that the difference-in-means can be decomposed to three terms of which only two depends on $\boldsymbol{\tau}$. In Equation (4), the first expression is a function of $Y_i(0)$ and T_i and is a random variable, the second is a function of $Y_i(0)$ and is not a random quantity, and the third is SATT. Lemma 3.1 motivates the use of t_{diff} for estimating SATT or SATC by symmetry. When recentering t_{diff} w.r.t. SATE the variance of the recentered estimator does not change; however, recentering w.r.t. SATT does change the variance calculations. This raises the question of whether inference on SATT has correct coverage of SATE and vice versa. Next we address this question both analytically and using Monte Carlo simulations.

Lemma 3.1. The difference-in-means test statistic t_{diff} can be decomposed into three expressions:

$$t_{\text{diff}}(\boldsymbol{Y}, \boldsymbol{T}) = \frac{N}{m \cdot (N-m)} \cdot \sum_{i=1}^{N} Y_i(0) \cdot T_i - \frac{1}{N-m} \cdot \sum_{i=1}^{N} Y_i(0) + \text{SATT}. \tag{4}$$

See Appendix A for the proof.

[3] A related literature discusses the idea of an optimal estimand in terms of covariate balance in observational studies (Crump et al. 2009; Li, Morgan, and Zaslavsky 2018). Crump et al. (2009) suggested a procedure for choosing the optimal estimand in observational studies where there is limited overlap in the covariates. The population overlap issue does not arise in randomized experiments.

[4] For a review of the classic CLT results under the finite population model, see Li and Ding (2016). Note that, the randomization model implies that the number of treated units, m, is a fixed number and not a random variable. The only random component in the model is the treatment indicators, $\boldsymbol{T}$.

The variance of $t_{\text{diff}} - \text{SATE}$ contains the parameter ρ (see Lemma 3.2), which cannot be observed and must be bounded when conducting inference. Lemmas 3.1 and 3.2 illustrate how by changing the estimand from SATE to SATT (or SATC), still using the difference-in-means test statistic, it is possible to conduct inference without needing to either know or bound ρ.

Lemma 3.2. The variance of the difference-in-means when recentered w.r.t. SATE or SATT is

$$\begin{aligned}\text{var}\left(t_{\text{diff}} - \text{SATE}\right) &= \frac{1}{N\cdot(1-p)\cdot p}\cdot\left[p^2\cdot\sigma_0^2 + (1-p)^2\cdot\sigma_1^2 \right.\\ &\qquad \left. + 2p(1-p)\cdot\rho\cdot\sigma_0\cdot\sigma_1\right],\\ \text{var}\left(t_{\text{diff}} - \text{SATT}\right) &= \frac{1}{p\cdot(1-p)\cdot N}\cdot\sigma_0^2,\end{aligned}$$

where $\rho \equiv \frac{\text{cov}(Y(1),Y(0))}{\sigma_0\cdot\sigma_1}$ is the correlation between the potential outcomes under the treatment and control regimes, and σ_j^2 ($j \in \{0,1\}$) is defined as $\sigma_j^2 = \frac{\sum_{i=1}^{N}\left(Y(j)_i - \bar{Y}(j)\right)^2}{N-1}$. See Appendix A for the proof.

The $t_{\text{diff}} - \text{SATT}$ can be more accurately estimated relative to the $t_{\text{diff}} - \text{SATE}$ when $\frac{\sigma_1}{\sigma_0} > 1$, ρ is sufficiently high, and p is not too high (e.g., $p = 1/2$). Theorem 3.1 shows that regardless of the value of $\frac{\sigma_1}{\sigma_0}$ there is a threshold level of ρ that below it $\text{var}\left(t_{\text{diff}} - \text{SATE}\right) \leq \text{var}\left(t_{\text{diff}} - \text{SATT}\right)$ and above which $\text{var}\left(t_{\text{diff}} - \text{SATE}\right) > \text{var}\left(t_{\text{diff}} - \text{SATT}\right)$. Notice that according to Theorem 3.1, it is simple to empirically test whether $\bar{\rho}$ is negative. One can conduct a one-sided hypothesis test of the null hypothesis:

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

and if the null hypothesis is rejected, then we can infer that $\bar{\rho} < 0$.

Theorem 3.1. 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

$$\begin{aligned}\rho \leq \bar{\rho} &\Rightarrow \text{var}\left(t_{\text{diff}} - \text{SATE}\right) \leq \text{var}\left(t_{\text{diff}} - \text{SATT}\right)\\ \rho > \bar{\rho} &\Rightarrow \text{var}\left(t_{\text{diff}} - \text{SATE}\right) > \text{var}\left(t_{\text{diff}} - \text{SATT}\right)\end{aligned}$$

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

See Appendix A for the proof.

In practice, the correlation between potential outcomes is not observed and to estimate $\text{var}\left(t_{\text{diff}} - \text{SATE}\right)$ it must be bounded. The most commonly used estimator was proposed by Neyman, and it ignores the correlation component all together. It can be rewritten as

$$\begin{aligned}\mathbb{V}_{\text{Neyman}} &= \frac{1}{m}\sigma_1^2 + \frac{1}{N-m}\sigma_0^2\\ &= \frac{1}{N\cdot p\cdot(1-p)}\left(\sigma_1^2(1-p) + \sigma_0^2 p\right).\end{aligned} \tag{5}$$

This variance estimator is consistent under the super-population sampling model, and it can be used to conduct inference on the population average treatment effect (PATE).[5] A less conservative estimator for the variance of the difference-in-means bounds ρ at 1:

$$\begin{aligned}\mathbb{V}_{\rho=1} &= \frac{1}{m}\sigma_1^2 + \frac{1}{N-m}\sigma_0^2 - \frac{(\sigma_1-\sigma_0)^2}{n}\\ &= \frac{1}{N\cdot(1-p)\cdot p}\left(p\cdot\sigma_0 + (1-p)\cdot\sigma_1\right)^2.\end{aligned} \tag{6}$$

Theorem 3.2 establishes that the limiting distribution of $t_{\text{diff}} - \text{SATT}$, when standardized using the variance formulas in Lemma 3.2, is standard Normal. An equivalent derivation of the limiting distribution of SATC follows immediately using an analog proof by symmetry.

Theorem 3.2. The difference-in-means recentered w.r.t. SATT follows a standard Normal distribution under two regularity conditions. When the following two conditions are satisfied:

$$\begin{aligned}&N - m \to \infty, \quad m \to \infty, \quad \text{and}\\ &\frac{\max_{1\leq i\leq N}\left(Y(0)_{Ni} - \bar{Y}(0)_N\right)^2}{\sum_{i=1}^{N}\left(Y(0)_{Ni} - \bar{Y}(0)_N\right)^2}\cdot\max\left(\frac{N-m}{m}, \frac{m}{N-m}\right) \to 0\end{aligned}$$

then

$$\frac{t_{\text{diff}} - \text{SATT}}{\sqrt{\text{var}\left(t_{\text{diff}} - \text{SATT}\right)}} \xrightarrow{d} N(0,1). \tag{7}$$

See Appendix A for the proof.[6]

Therefore, a $1-\alpha$ PI for SATT is

$$\left[t_{\text{diff}} - z_{1-\alpha/2}\cdot\hat{\sigma}_0\cdot\sqrt{k(N,m)},\ t_{\text{diff}} + z_{1-\alpha/2}\cdot\hat{\sigma}_0\cdot\sqrt{k(N,m)}\right], \tag{8}$$

where

$$k(N,m) = \frac{1}{p\cdot(1-p)\cdot N}.$$

Rigdon and Hudgens (2015) showed how to derive a CI for SATE by combining two PIs for SATT and SATC. They focused on binary outcomes for which PIs have been derived using past theoretical results (Robins 1988; Rosenbaum 2001). Theorem 3.3 provides two key results. First, it is a generalization of Theorem 1 in Rigdon and Hudgens (2015), and it shows how a CI for SATE can be constructed in any randomized control trial regardless of whether the outcomes are binary or continuous. This part of the theorem follows directly from the derivations in Rigdon and Hudgens (2015). Second, it shows that combining PIs for SATT and SATC using a Bonferroni-type adjustment, as was done by Rigdon and Hudgens, yields *exactly* the same CI as when one constructs a CI for SATE directly and uses a conservative variance estimator that bounds ρ at 1, $\mathbb{V}_{\rho=1}$. Theorem 3.3 implies that combining PIs for SATT and SATC yield a more conservative CI than using a CI based on a sharper bound for ρ, such as that of Aronow, Green, and Lee (2014).

[5] The estimator in Equation (5) corresponds to Neyman's second variance estimator when the estimand is SATE (Neyman 1923/1990).

[6] These are not the weakest possible conditions.

Theorem 3.3. Let $[L_{\text{SATT}}, U_{\text{SATT}}]$ and $[L_{\text{SATC}}, U_{\text{SATC}}]$ be PIs for SATT and SATC according to the variance formulas in Lemma 3.2; then a CI for SATE is

$$\left[p \cdot L_{\text{SATT}} + (1-p) \cdot L_{\text{SATC}}, \quad p \cdot U_{\text{SATT}} + (1-p) \cdot U_{\text{SATC}}\right]$$

and is equal to

$$\left[t_{\text{diff}} - z_{1-\alpha/2} \cdot \sqrt{\hat{\mathbb{V}}_{\rho=1}},\ t_{\text{diff}} + z_{1-\alpha/2} \cdot \sqrt{\hat{\mathbb{V}}_{\rho=1}}\right].$$

See Appendix A for the proof.

Lemma 3.3. When $\sigma_1 \neq \sigma_0$, a PI for either SATT or SATC is shorter than a CI for SATE that uses either $\mathbb{V}_{\text{Neyman}}$ or $\mathbb{V}_{\rho=1}$.

See Appendix A for the proof. According to Lemma 3.3, whenever $\sigma_0 < \sigma_1$ ($\sigma_0 > \sigma_1$), a PI for SATT (SATC) is shorter than a CI for SATE using Neyman's variance estimator. The gain in terms of interval length (in %) is $1 - \frac{1}{\sqrt{\left(\frac{\sigma_1^2}{\sigma_0^2}(1-p)+p\right)}}$, and it is decreasing with respect to p. For example, in a balanced design, $p = \frac{1}{2}$, with a variance ratio of two, the length of a PI for SATT will be 18.35% shorter relative to a CI for SATE that is based on Neyman's variance estimator.[7]

To obtain intuition about how recentering the difference-in-means w.r.t. SATT can yields a different variance estimator, it is useful to decompose the variance of the difference-in-means. Equation (9) shows that when recentering w.r.t. SATT we cancel two of the elements in the variance expression, var (SATT) and $\text{cov}\left(\text{SATT}, \frac{N}{m(N-m)} \cdot \sum_{i=1}^{N} Y(0)T_i\right)$, and in doing so can potentially reduce/increase the uncertainty.

$$\begin{aligned} &\text{var}\,(t_{\text{diff}} - \text{SATE}) \qquad (9)\\ &= \text{var}\,(\text{SATT}) + \text{var}\left(\frac{N}{m(N-m)} \cdot \sum_{i=1}^{N} Y(0)T_i\right)\\ &\quad + 2 \cdot \text{cov}\left(\text{SATT}, \frac{N}{m(N-m)} \cdot \sum_{i=1}^{N} Y(0)T_i\right) \end{aligned}$$

and the mean squared distance between the two estimands is

$$\text{MSE}(\text{SATE}, \text{SATT}) = \frac{1-p}{m} \cdot \sigma_\tau^2. \qquad (10)$$

See Appendix A for the proof. Thus, the difference between conducting inference on SATE or SATT increases with the heterogeneity of the treatment effect. Given a fixed and bounded value of $\sigma_\tau^2 < \infty$, those difference will converges to zero as $m \to \infty$.

[7] Note that Lemma 3.3 does not cover other estimators for the variance of var $(t_{\text{diff}} - \text{SATE})$, such as that of Aronow, Green, and Lee (2014). CIs based on these other variance estimators may be shorter than our PIs even when $\sigma_1 \neq \sigma_0$. To address these variance estimators, we derive the accuracy gains for a more general case in which ρ can be bounded by ρ^*, $\rho \leq \rho^*$. As the variance of var $(t_{\text{diff}} - \text{SATE})$ is increasing w.r.t. ρ, it follows that substituting ρ^* with ρ yields a conservative variance estimator that is smaller than Neyman's variance estimator. The idea of substituting a bound of ρ instead of the true parameter value was proposed before in the literature (Reichardt and Gallob 1999; Aronow, Green, and Lee 2014). The percentage gain in terms of CI length is: $1 - \frac{1}{\sqrt{p^2+(1-p)^2 \cdot \left(\frac{\sigma_1}{\sigma_0}\right)^2 + 2p(1-p)\cdot\rho^* \cdot \frac{\sigma_1}{\sigma_0}}}$.

4. Analytical Examples and Monte Carlo Simulations

4.1. Additive Random Coefficients Model

We consider a simple additive treatment effect model with heterogeneous impacts across units:

$$Y_i(1) = \tau_i + Y_i(0), \qquad (11)$$

where τ_i is a random variable, which for simplicity is assumed not to be correlated with $Y_i(0)$, nor to be correlated with T_i by construction due to the randomization of treatment assignment. The variance of the treated units is larger by σ_τ^2, and this generates a potential difference between a CI for SATE and a PI for SATT. The variance ratio is: $\frac{\sigma_\tau^2}{\sigma_0^2} + 1$ and it is increasing with the heterogeneity of the treatment effect. Consider the data-generating process in Equation (12) with a sample of $n = 1000$ units. We performed 1000 draws of samples and for each one simulated 1000 different allocations of the treatment according to $p = 1/2$.[8]

$$\begin{aligned} &Y_i(0) \sim N(\mu = 10,\ \sigma_0^2 = 1),\ \tau_i \sim N(\mu = 0,\ \sigma_\tau^2),\\ &Y_i(1) = \tau_i + Y_i(1). \qquad (12) \end{aligned}$$

Figure 1 reports the simulation results. The PIs for SATT are substantially shorter than the CIs for the SATE; this holds both when using Neyman's variance estimator and when using a variance estimator based on a sharp bound for ρ, such as that of Aronow, Green, and Lee (2014). The comparison to the CI based on the true ρ parameter shows that the accuracy differences are mainly due to the change of estimand, rather than a conservative variance estimator. This is contrary to the binary outcome example that is discussed in Appendix C. The simulation is supported by our theoretical results that a change of estimand has larger accuracy impacts as the variance of the treatment effect is higher. The left plot in Figure 1 shows that coverage of the PI for SATT w.r.t. both SATT and SATE. As the treatment effect becomes more heterogeneous the PI for SATT provides worse coverage of SATE and, vice versa, a CI for SATE provides worse coverage of SATT.

4.2. Censored Outcomes (Tobit Model)

Consider a finite population of size N:

$$\begin{aligned} \Pi_N = \big\{&(Y(0)_{1N}, Y(1)_{1N}),\ (Y(0)_{2N}, Y(1)_{2N}), \ldots,\\ &(Y(0)_{NN}, Y(1)_{NN})\big\} \qquad (13) \end{aligned}$$

where $Y(0)$ is a continuous outcome and $Y(1)$ is

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

We used a sample size of 1000 units and performed 1000 draws of samples and for each one simulated 1000 different allocations of the treatment according to $p = 1/2$. For the Monte Carlo simulations we used the following data-generating process:

$$Y_i(0) \sim N(\mu = 0,\ \sigma_0^2 = 1). \qquad (14)$$

[8] In Appendix Figure C.2, we show that the simulation results are not sensitive to our choice of using $N = 1000$ and would have been the same using $N = 100$ or $N = 1300$. This holds for both this data-generating process and the Tobit model that is discussed below.

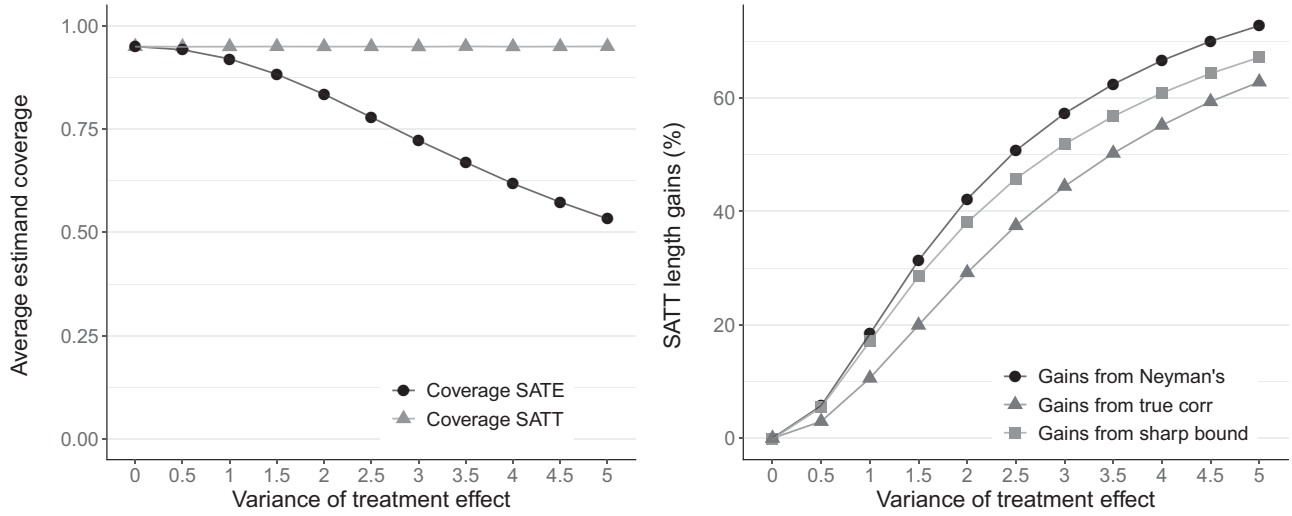


Figure 1. Additive and heterogeneous treatment effect (random coefficient) simulation results: Length and coverage differences. NOTES: The left plot shows the coverage of a PI for SATT w.r.t. SATT and SATE. The coverage of SATT has correct size and there is no evidence of over rejection of the null hypothesis. On the other hand, the coverage of SATE becomes worse as the variance of the treatment effect increases. The right plot illustrates the accuracy gains of using inference on SATT relative to SATE. The PI for SATT is compared to CIs for SATE based on three different values of ρ. The circles represent the accuracy gains (in percentages) relative to Neyman's variance estimator which assumes that $\rho = 0$. The squares dots use a sharp bound for ρ that have been derived by Aronow, Green, and Lee (2014). And the triangles compare the length of a PI for SATT to a CI for SATE when the true value of ρ is known to the researcher.

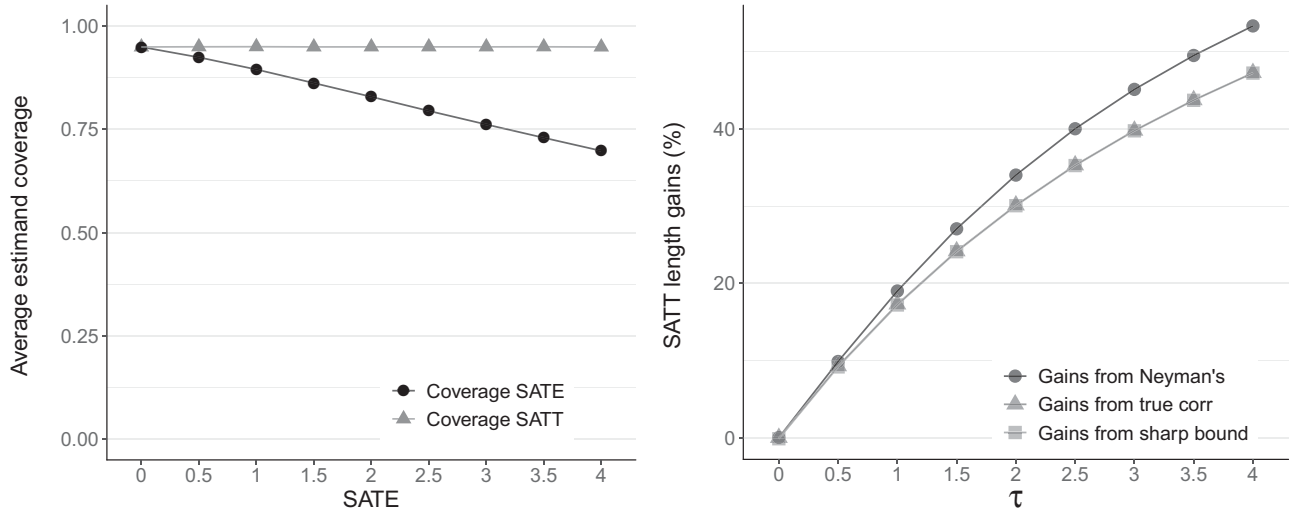


Figure 2. Censored outcome (Tobit) simulation results: Length and coverage differences. NOTES: See the notes in Figure 1.

The above Tobit model (Tobin 1958) implies that the variance of the potential outcomes under treatment is higher than the variance of units under the control regime and the variance ratio is increasing with respect to τ:

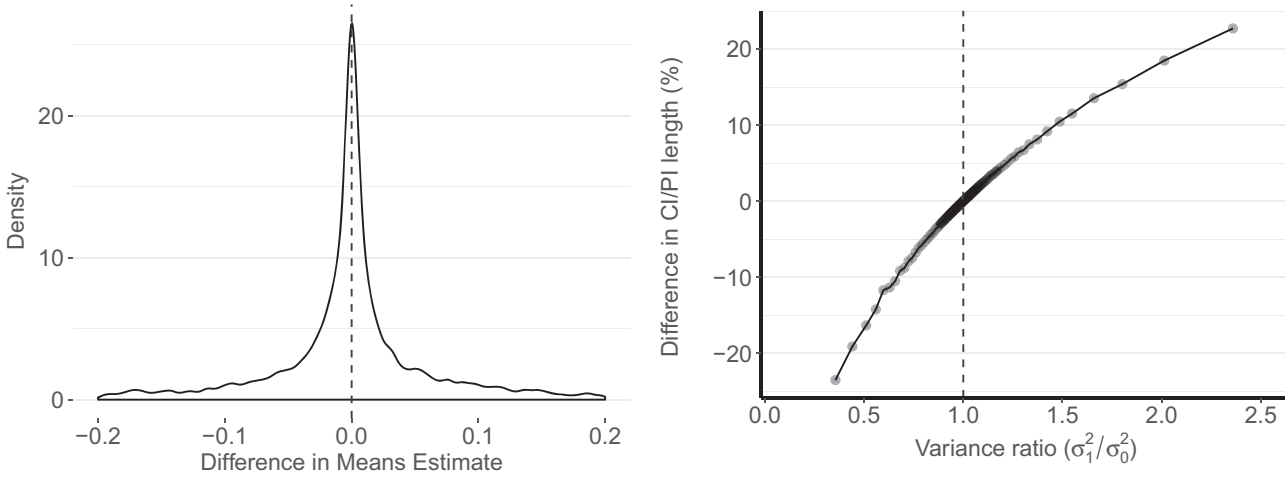
$$\frac{\sigma_1^2}{\sigma_0^2} = 1 + \frac{\begin{array}{c}\Pr(Y(0) > 0) \cdot \tau \cdot [\tau \cdot (1 - \Pr(Y(0) > 0) \\ + \mathbb{E}[Y(0)|Y(0) > 0] - \mathbb{E}[Y(0)]]\end{array}}{\sigma_0^2}$$

The simulation results in Figure 2 show that similarly to the random coefficient model the change of estimand is the main cause of the differences in length/accuracy. Similarly to the previous simulations, as σ_τ^2 increases the differences between estimands become more stark and inference for SATT (SATE) has bad coverage w.r.t. SATE (SATT).

5. Comments

Several remarks on the previous results and possible extensions are in order.

Remark 1. The previous results can be extended to include covariate adjustment of pretreatment characteristics using the procedure proposed by Rosenbaum (2002). Denote by X_i a $1 \times p$ dimensional vector of the pretreatment characteristics of unit i. The matrix $\boldsymbol{X}$ has dimensions $n \times p$ and each row i contains the pretreatment characteristics of unit i. It is common to adjust Y_i for X_i for efficiency purposes. Define $Y_i^{\text{adjusted}} = Y_i - X_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ as the adjusted/residualized responses. All the results for inference on Y_i also apply for inference on Y_i^{adjusted}.



(a) Distribution of the difference-in-means (b) Differences in variance: SATT vs. SATE

Figure 3. Inference accuracy: A comparison of SATT to SATE across online experiments. NOTES: The left plot (a) presents the distribution of the difference-in-means test statistic across the hundreds of online experiments. The right plot (b) describes the relationship between the variance ratio in the experiment and the gain in accuracy, difference in PI length relative to CI length, from conducting inference on SATT relative to SATE.

Remark 2. The inference results for SATT can also be extended to additional treatment assignment models that differ from the classic complete randomization mechanism. For example, Theorem E.1, in Appendix E, provides variance and limiting distribution results for inference on SATT when the treatment assignment is done by random independent Bernoulli trials.

Remark 3. The regularity conditions of the above theoretical results will usually be satisfied in applied research, yet it is important to understand when they will not. For example, imagine a control regime in which all the individuals die ($Y_i(0) = 0 \quad \forall_i$), while under the treatment regime units have a strictly positive survival probability, $\mathbb{E}[Y(1)] = p_1$ and $\text{var}(Y(1)) = p_1(1-p_1)$. As the variance of the control units is strictly lower (zero) than that of the treated units, there are potential accuracy gains from estimating SATT instead of SATE. In this scenario as $\sigma_0^2 = 0$ the PI for SATT contains only one point, the difference-in-means estimate, which is clearly wrong. The example above does not stand in contradiction to Theorem 3.2, as in the above case the regularity condition of the theorem is not satisfied: $\frac{\max_{1\leq i\leq N}\left(Y(0)_{Ni}-\bar{Y}(0)_N\right)^2}{\sum_{i=1}^{N}\left(Y(0)_{Ni}-\bar{Y}(0)_N\right)^2} = \frac{0}{0}$ and $\frac{0}{0}$ is not a well-defined expression.

Remark 4. The variance formulas in Lemma 3.2 can be extended to a super-population model in which the sampling procedure has two steps. First a sample of $1, \ldots, N$ units is randomly drawn from the population. Second, m units within the sample are randomly allocated to the treatment regime and the remaining $N - m$ units to the control regime. Appendix B discusses how to conduct inference on SATT and SATE in this setting. Lemma B.1 provides variance formulas that extend Lemma 3.2 to the super-population sampling model.

Remark 5. The variance and inference results for SATT can be extended to linear combinations of SATT and SATC. In Appendix D, we derive inference results for ωSATT + $(1-\omega)$SATC when ω is chosen to minimize the variance of the recentered difference-in-means, $t_{\text{diff}} - (\omega\text{SATT} + (1-\omega)\text{SATC})$. We denote the weighting of SATT and SATC that minimizes MSE as Sample Average Treatment Effect Optimal (SATO). Inference on SATO can be used, for example, to reject the sharp null of no treatment effect and will be more efficient for this purposes than conducting inference on SATE.

6. Real Data Application: Online Experiments

To better understand the trade-offs in conducting inference on different average treatment effect estimands, we analyze a sample of online field experiments that have been conducted by a large internet firm as product improvement tests. Our sample consists of 278 experiments with an average sample size of approximately 100 million units per experiment. Many different outcome metrics are analyzed for various subgroups. The average subgroup consists of 1.1 million observations, and there are 826 unique outcome metrics across all of the experiments. In total, twenty-five thousand different treatment effects are estimated. The data analyzed were aggregated and de-identified.

The left plot of Figure 3(a) shows the distribution of the difference-in-means across all of the online experiments. It is clear that on average, across experiments, the treatment had a zero effect. The distribution of difference-in-means is tightly centered around zero. However, this does not imply that the treatment had no effect. Treatment effect heterogeneity can generate positive effects for some units and negative effects for others that cancel each other on average. However, elaborate tests and computations are difficult to carried out with millions of observations. Next we compare inference on SATE and SATT in this setting and to what degree do they differ.

The right plot of Figure 3(b) demonstrates that variance differences across experiments exist. As the variance ratio σ_1^2/σ_0^2 increases, the PI for SATT becomes shorter, and in some cases there can be large variance gains from changing the estimand to SATT. As the variance ratio departs from one there can be substantial differences between conducting inference on SATT relative to SATE. The differences in PI/CI length, and hence also

in rejection rates, between inference on SATE and SATT are evidence of treatment effect heterogeneity.

7. Discussion

Making inferences about SATT (or analogously on SATC) using a CI for SATE relies on variance estimators that are not consistent and are not guaranteed to have correct coverage or be efficient. We derive efficient variance formulas for inference on a new and general class of estimands derived from any mixing between SATT and SATC. The variance formulas are used to construct PIs that are non-parametrically guaranteed to have correct coverage and to be nonconservative—unlike inference on SATE. All inference procedures discussed in the article use the difference-in-means as the test statistic, and therefore have the same point estimates as existing methods. Note that all three estimands, SATE, SATT, and SATC, are equal in expectation. The key difference is in the variance calculations.

Taken together, the Monte Carlo simulations demonstrate that: (i) the choice of estimand has a direct implication on the accuracy of the inference that can be conducted; and (ii) using variance formulas that have correct size for SATE will *not* have correct coverage of other sample average treatment effect estimands such as SATT. The application to online experiments provides a real data empirical example that even in experiments with millions of observations there can still be meaningful differences between valid inference on SATT relative to SATE.

The large potential differences in coverage (and efficiency) emphasizes that researchers should think carefully about which causal estimand they want to conduct inference on. If the answer is SATE, then they should not use any of the results in this article; however, if SATT (or SATC) is also of interest, then the variance formulas and inference results presented in this article will be of direct use.

Supplementary Materials

The supplemental material includes replication code, proofs, and additional results that are mentioned in the text only briefly.

Acknowledgments

We thank Peter Aronow, Max Balandat, Eytan Bakshy, Avi Feller, Johann Gagnon-Bartsch, Peng Ding, Ben Hansen, Guido Imbens, James Robins, Sören Kunzel, Winston Lin, Juliana Londoño Vélez, Fredrik Sävje, and John Myles White for helpful comments and discussions. In addition, we thank Max Balandat, Eytan Bakshy, and John Myles White for help with the data application. We also thank the participants of the Berkeley Statistics Annual Research Symposium 2017 and the Atlantic Causal Inference Conference 2017.

Funding

Sekhon wishes to acknowledge Office of Naval Research (ONR) grants N00014-15-1-2367 and N00014-17-1-2176.

References

Aronow, P. M., Green, D. P., and Lee, D. K. K. (2014), "Sharp Bounds on the Variance in Randomized Experiments," *The Annals of Statistics*, 42, 850–871. [1,2,3,4,5]

Bowers, J., and Hansen, B. B. (2009), "Attributing Effects to a Cluster Randomized Get-Out-the-Vote Campaign," *Journal of the American Statistical Association*, 104, 873–885. [2]

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009), "Dealing With Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96, 187–199. [2]

Feng, X., Feng, Y., Chen, Y., and Small, D. S. (2014), "Randomization Inference for the Trimmed Mean of Effects Attributable to Treatment," *Statistica Sinica*, 24, 773–797. [2]

Holland, P. (1986), "Statistics and Causal Inference" (with discussion), *Journal of the American Statistical Association*, 86, 945–970. [2]

Imbens, G. W. (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29. [1]

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, New York: Cambridge University Press. [1]

Keele, L., Small, D., and Grieve, R. (2017), "Randomization-Based Instrumental Variables Methods for Binary Outcomes With an Application to the 'IMPROVE' Trial," *Journal of the Royal Statistical Society*, Series A, 180, 569–586. [2]

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018), "Balancing Covariates via Propensity Score Weighting," *Journal of the American Statistical Association*, 113, 390–400. [2]

Li, X., and Ding, P. (2016), "General Forms of Finite Population Central Limit Theorems With Applications to Causal Inference," Technical Report. [2]

Neyman, J. (1923/1990), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9" (Trans. Dorota M. Dabrowska and Terence P. Speed), *Statistical Science*, 5, 465–472. [1,3]

Reichardt, C., and Gallob, H. (1999), "Justifying the Use and Increasing the Power of a *t* Test for a Randomized Experiment With a Convenience Sample," *Psychological Methods*, 4, 117–128. [4]

Rigdon, J., and Hudgens, M. G. (2015), "Randomization Inference for Treatment Effects on a Binary Outcome," *Statistics in Medicine*, 34, 924–935. [2,3]

Robins, J. M. (1988), "Confidence Intervals for Causal Parameters," *Statistics in Medicine*, 7, 773–785. [1,2,3]

Rosenbaum, P. (2001), "Effects Attributable to Treatment: Inference in Experiments and Observational Studies With a Discrete Pivot," *Biometrika*, 88, 219–231. [2,3]

——— (2002), "Covariance Adjustment in Randomized Experiments and Observational Studies," *Statistical Science*, 17, 286–304. [5]

Tobin, J. (1958), "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 26, 24–36. [5]