Statistics and Its Interface Volume 16 (2023) 349–361

Testing attributable effects hypotheses with an application to the Oregon Health Insurance Experiment*

Mark M. Fredrickson† and Yuguo Chen

Following a randomized trial, the sum of the differences in the outcomes for the treated units compared to the outcome that would have been observed if the same units had been assigned to the control condition is known as the attributable effect. Most previous methods on testing hypotheses about the attributable effect require the outcome to be binary or ordinal. In this paper, we use a simple approximation to the distribution of a carefully selected test statistic under the hypothesis that the attributable effect is zero to expand attributable effects inference for count and continuous data. The method is efficient and performs well in a variety of simulations. We demonstrate the method using a large medical insurance field experiment.

AMS 2000 subject classifications: Primary 62-08; secondary 62G10.
Keywords and phrases: Attributable effects, Hypothesis testing, Optimization, Randomization inference, Zero-inflated outcomes.

1. INTRODUCTION

In 2008, the state of Oregon engaged in a lottery in which low income residents were selected to be allowed to apply for state funded Medicaid health insurance. Supporters of expanded state sponsored health care argue that offering medical insurance shifts incentives to use expensive emergency room care to less expensive scheduled clinical care, while also lowering the economic burden on low income households. To address these arguments, Finkelstein et al. [9] surveyed those assigned to both the health insurance arm and those who were not selected in the lottery to ascertain the amount spent on out of pocket medical costs. The 11,450 responding households in the control condition reported a total of $4.71 million spent on medical care in the previous six months [9]. On average, this translates to $411.68 per control household, but this averaging obscures the fact that 49% of the control households reported spending zero dollars on out of pocket costs. For nearly half of the control subjects, the average is not very informative about the amount spent on medical care.

Instead of asking about average effects, it may be more useful to ask what portion of the observed costs of the control subjects can be attributed to being prohibited from applying for Oregon's Medicaid program. Rosenbaum [31] calls this quantity "the effect attributable to treatment." Many of the approaches for estimating and testing hypotheses about attributable effects have been focused on binary outcomes [31]. Rosenbaum [32] extends his previous work to matched pair designs. Rigdon and Hudgens [29] allow for attributable effects to both the treatment group and control group in order to get confidence intervals for the average treatment effect. Li and Ding [22] improve the efficiency of these results. Fogarty et al. [12] focus on the particular difficulties in observational studies that attempt to emulate randomized trials and propose numerical solutions that provide tests of effects on binary outcomes along with sensitivity analyses. Also for observational data, Keele et al. [18] develop instrumental variable methods for attributable effects for binary data. Choi [2] provides optimization based techniques for solving attributable effects for binary data and includes methods for improving inference when information is available about interactions between subjects. Some progress has also been made on ordinal outcomes using well defined sequences of alternative hypotheses [23], bounds [24], or introducing nuisance parameters or latent variables [37]. Several papers highlight how physical randomization, binary outcomes, and monotonicity, as assumption that individual treatment effects all have the same sign, combine to generate a multiple hypergeometric likelihood, which can be used for inference for both the attributable effect and the average treatment effect [6, 18].

There are two notable exceptions to the focus on binary data. Several authors suggest using survey sampling methods for estimating attributable effects [30, 15]. While this approach expands the scope of data to include count and continuous outcomes, the method requires large sample approximations to hold, which may be suspect in highly skewed outcomes. Feng et al. [8] provide an exact test for continuous outcomes based on a complex optimization problem.

*This work was supported in part by NSF grants DMS-1406455, DMS-1646108 and DMS-2015561. We appreciate comments and suggestions from Jake Bowers and Ben B. Hansen. We thank the authors of the Oregon Health Insurance Study for publishing their data.
†Corresponding author.

The approach is based on the Mann-Whitney-Wilcoxon sum of ranks test statistic, which degrades in the presence of ties in the values of Y_i. For both methods, large numbers of zeros in the outcomes are difficult to handle.

In this paper, we present a method that can be thought of as a hybrid between the existing exact tests for attributable effects and the survey sampling based estimation approach. We use a normal approximation as part of an optimization routine, but test the resulting hypothesis using exact methods. The method is computationally efficient, and simulations show that it performs well when the data contain large portions of zero values.

The rest of the paper is organized as follows. Section 2 introduces the proposed method, along with notation and assumptions. Section 3 evaluates the accuracy of the key approximation and the statistical properties of the method through a variety of simulations. Section 4 returns to the Oregon Health Insurance Program experiment previously introduced to analyze several outcomes. Section 5 concludes with a discussion.

2. METHODOLOGY

2.1 Setting and notation

Consider N units in a study where n units are assigned to the treatment condition and the remaining $m = N - n$ units are assigned to the control condition, writing $Z_i = 1$ for treatment and $Z_i = 0$ for control. Throughout, we shall notate vectors using boldfaced symbols, so $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_N)'$. We suppose that $\boldsymbol{Z}$ is assigned by *complete random assignment*: $\Pr(\boldsymbol{Z} = \boldsymbol{z}) = n!m!/N!$ if $\sum_{i=1}^{N} z_i = n$ and $\Pr(\boldsymbol{Z} = \boldsymbol{z}) = 0$ otherwise. For all subjects, we hypothesize *potential outcomes* to the different treatment conditions $y_i(1)$ when $Z_i = 1$ and $y_i(0)$ when $Z_i = 0$ [28, 16]. We will require, for reasons discussed later, that both $y_i(1) \geq 0$ and $y_i(0) \geq 0$. Many types of data, such as counts, naturally ensure this assumption. For other data that can take negative values, it is often possible to shift the observations by some constant to ensure that the condition holds.

The observed outcome Y_i is random in that it depends on Z_i: $Y_i = y_i(Z_i)$. The vector $\boldsymbol{Y} = (y_1(Z_1), \ldots, y_N(Z_N)) = \boldsymbol{y}(\boldsymbol{Z})$ defines the outcomes observed after treatment. Implicit in this definition is an assumption that assignment to the treatment or control condition for unit i does not change the outcome of any unit j, often labeled as the Stable Unit Treatment Value Assumption (SUTVA) [34].

Define the vector of individual effects $\boldsymbol{\tau} = \boldsymbol{y}(\mathbf{1}) - \boldsymbol{y}(\mathbf{0})$. Suppose we were able to observe $\boldsymbol{Y}^* = \boldsymbol{y}(\mathbf{1} - \boldsymbol{Z})$. Then $\boldsymbol{\tau}$ could be recovered using $\tau_i = Y_i - Y_i^*$ for $Z_i = 1$ and $\tau_i = Y_i^* - Y_i$ for $Z_i = 0$. Of course, we only observe $\boldsymbol{Y} = \boldsymbol{y}(\boldsymbol{Z})$, so $\boldsymbol{\tau}$ is not identified. While we cannot identify $\boldsymbol{\tau}$ without additional assumptions, we can test a sharp null hypothesis $H_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$. To test the sharp null, we remove the hypothesized treatment effect from the treated units, $\tilde{\boldsymbol{Y}} = \boldsymbol{y}(\boldsymbol{Z}) - \boldsymbol{\tau}_0 \odot \boldsymbol{Z}$, where $\odot$ indicates the element-wise product. Under the null hypothesis, $\boldsymbol{\tau}_0 = \boldsymbol{y}(\mathbf{1}) - \boldsymbol{y}(\mathbf{0})$, so $\tilde{\boldsymbol{Y}} = \boldsymbol{y}(\mathbf{0})$, which is to say a fixed quantity, and we henceforth write $\tilde{\boldsymbol{y}}$ when considering the adjusted outcome under the null. A randomization test can be applied using a suitable test statistic that is increasing in evidence against the null [10, 32, 33]. After selecting a test statistic $T(\boldsymbol{Z}, \tilde{\boldsymbol{y}})$, its distribution under the hypothesis $H_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$ is given by enumerating all possible ways of selecting n out of N units and computing the test statistic applied at each randomization [10]. The p-value of the hypothesis is the proportion of randomizations that lead to a larger test statistic value than the observed value. Indexing all $J = N!/(n!m!)$ possible treatment assignments as $\boldsymbol{z}^{(j)}$, write the p-value as

$$\begin{aligned} p &= \Pr(T(\boldsymbol{Z}, \tilde{\boldsymbol{y}}) \geq T(\boldsymbol{z}, \tilde{\boldsymbol{y}})) \\ &= J^{-1} \sum_{j=1}^{J} I(T(\boldsymbol{z}^{(j)}, \tilde{\boldsymbol{y}}) \geq T(\boldsymbol{z}, \tilde{\boldsymbol{y}})), \end{aligned}$$

where $\boldsymbol{z}$ is the realized treatment assignment in the experiment and $I(\cdot)$ is the indicator function. One of the primary advantages of the Fisherian approach is that it does not rely on large sample approximations or distribution assumptions. The trade-off is that it requires hypothesizing the subject level treatment effects τ_i.

As an alternative to specifying the entire $\boldsymbol{\tau}_0$ vector, consider the *attributable effect*

$$A = \boldsymbol{Z}'\boldsymbol{\tau}. \tag{1}$$

Observe that a hypothesis of the form $H_0 : A = A_0$ is a *composite hypothesis* as it contains any $\boldsymbol{\tau}_0$ for which $\boldsymbol{Z}'\boldsymbol{\tau}_0 = A_0$. Theoretically, A_0 could be tested using a randomization test if one could find the $\boldsymbol{\tau}_0$ with the maximum p-value among the set $\{\boldsymbol{\tau}_0 : \boldsymbol{Z}'\boldsymbol{\tau}_0 = A_0\}$, as the p-value of the true $\boldsymbol{\tau}$ must be no greater than the maximum. For count data, enumerating all possible $\boldsymbol{\tau}_0$ is computationally intractable in most circumstances. For continuous data, such an enumeration is not even possible.

2.2 Approximating the adjustment with the largest p-value

Most of the previous approaches to attributable effects relied on "distribution free methods," in which the distribution of the test statistic did not depend the values of the outcomes themselves [27]. In these situations, the problem of finding the largest p-value is equivalent to finding the smallest test statistic value that results from an adjustment $\boldsymbol{\tau}_0$, as any adjustment $\boldsymbol{\tau}_0$ will result in the same null distribution. In this paper, we take the opposite approach: the test statistic remains fixed while we search for a distribution that places the most mass above the test statistic value. While we have wide latitude selecting the test statistic T, a natural

choice is the deviation of the treatment group's mean from the overall mean:

$$T(\boldsymbol{Z},\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{N} Z_i y_i - \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{2}$$

This statistic has been widely studied in both the randomization and permutation literature as an analog of the parametric t-test [21, Chapter 5]. For the present purposes, the primary advantages of this test statistic are the close alignment with the definition of the attributable effect and a convenient normal approximation.

Usefully, the value of the test statistic (2) evaluated at $\boldsymbol{z}$, the observed assignment vector, remains fixed under any possible $\boldsymbol{\tau}_0$ that is compatible with A_0. For observed treatment $\boldsymbol{z}$, observed data $\boldsymbol{y}$, and null hypothesis $\boldsymbol{\tau}_0 = (\tau_{0,1}, \tau_{0,2}, \ldots, \tau_{0,N})^T$ such that $\boldsymbol{z}'\boldsymbol{\tau}_0 = A_0$, define the adjusted data $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{\tau}_0 \odot \boldsymbol{z}$. The value of test statistic when applied to the adjusted data only depends on $\boldsymbol{\tau}_0$ through A_0:

$$\begin{aligned} T(\boldsymbol{z},\tilde{\boldsymbol{y}}) &= \frac{1}{n}\sum_{i=1}^{N} z_i(y_i - z_i\tau_{0,i}) - \frac{1}{N}\sum_{i=1}^{N}(y_i - z_i\tau_{0,i}) \\ &= T(\boldsymbol{z},\boldsymbol{y}) - \frac{(m/n)}{N}A_0. \end{aligned}$$

Therefore any hypothesis compatible with A_0 generates the same value of T.

While the observed test statistic remains unchanged, the distribution of $T(\boldsymbol{Z},\tilde{\boldsymbol{y}})$ depends on the particular $\tau_{0,i}$ values. Since $\tilde{\boldsymbol{y}}$ is a fixed quantity under the null that $A = A_0$, $T(\boldsymbol{Z},\tilde{\boldsymbol{y}})$ can be thought of as a sample average of n items drawn from a finite population of size N, centered on the true population mean $\mu_0 = N^{-1}\left(\sum_{i=1}^{N} y_i - A_0\right)$. The mean and variance of $T(\boldsymbol{Z},\tilde{\boldsymbol{y}})$ follow from standard finite population sampling results [3, Theorems 2.1, 2.2]:

$$\begin{aligned} \mathrm{E}\left(T(\boldsymbol{Z},\tilde{\boldsymbol{y}})\right) &= \frac{1}{n}\sum_{i=1}^{N}\mathrm{E}\left(Z_i\right)\tilde{y}_i - \mu_0 = 0, \\ \mathrm{Var}\left(T(\boldsymbol{Z},\tilde{\boldsymbol{y}})\right) &= \frac{m/n}{N(N-1)}\left[\sum_{i=1}^{N} z_i\left(\tau_{0,i} - y_i + \mu_0\right)^2 \right. \\ &\qquad \left. + \sum_{i=1}^{N}(1-z_i)\left(y_i - \mu_0\right)^2\right]. \end{aligned}$$

While the portion of the sum that depends on the control units is a constant, the portion depending on the treated units is a function of the exact $\tau_{0,i}$ values, even though $\sum_{i=1}^{N} Z_i\tau_{0,i} = A_0$ is fixed.

Under fairly mild conditions, T is approximately normally distributed in large samples ([14]; [20, p. 353]). Consider a set of finite populations indexed by ν. For each population of N_ν subjects, n_ν are assigned to treatment and m_ν are assigned to control. For each population, the null hypothesis $\tau_{0,\nu}$ holds so adjusted values $\tilde{y}_{\nu,i}$ are fixed. The statistic $S_\nu = T_\nu/\mathrm{Var}\left(T_\nu\right)^{1/2}$ converges in distribution to $N(0,1)$ when $N_\nu, n_\nu, m_\nu \to \infty$ and

$$\frac{\max(\tilde{y}_{\nu,i} - \mu_\nu)^2}{\sum_{i=1}^{N_\nu}(\tilde{y}_{\nu,i} - \mu_\nu)^2}\max\left(\frac{n}{m}, \frac{m}{n}\right) \to 0 \quad \text{as } \nu \to \infty,$$

where $\mu_\nu = \frac{1}{N}\tilde{y}_{\nu,i}$. The first term requires that no individual $\tilde{y}$ be so large as to dominate the variance, while the second implies that neither the treated nor control group size become negligible, which seems particularly natural in the context of a series of increasingly larger experiments.

Let $c = \sum_{i=1}^{N}(1-z_i)\left(y_i - \mu_0\right)^2$ be the control subjects' contribution to the variance of T. Under the regularity conditions above, squaring T leads to a scaled χ^2 distribution:

$$\left[T(\boldsymbol{Z},\tilde{\boldsymbol{y}})\right]^2 \sim \frac{m/n}{N(N-1)}\left[\sum_{i=1}^{N} z_i\left(\tau_{0,i} - y_i + \mu_0\right)^2 + c\right]\chi^2_1.$$

Recall that for any fixed value of A_0, the value of T^2 will be the same regardless of the particular values $\tau_{0,i}$. Therefore the vector of adjustments $\boldsymbol{\tau}_0$ that corresponds to the largest possible p-value consistent with A_0 can be found by maximizing the quantity $\sum_{i=1}^{N} z_i(\tau_{0,i} - y_i + \mu_0)^2$. In order to find the $\boldsymbol{\tau}_0$ that maximizes T, we make one of two possible monotonicity assumptions. Either the potential outcome to treatment is at least as large as the potential outcome to control for all units or the potential outcome to control is at least as large as the potential outcome to treatment:

Assumption 1. $0 \le y_i(0) \le y_i(1)$,

or

Assumption 2. $0 \le y_i(1) \le y_i(0)$.

For the purpose of exposition, we focus on the case when Assumption 1 holds, but applying the methods when Assumption 2 holds simply requires substituting $\boldsymbol{W} = \boldsymbol{1} - \boldsymbol{Z}$ for $\boldsymbol{Z}$ throughout.

Without loss of generality, we suppose that the first n units are the treated units (i.e., $z_i = 1$ for $i = 1, \ldots, n$ and $z_i = 0$ for $i = n+1, \ldots, N$). Under Assumption 1, the following optimization problem finds the $\boldsymbol{\tau}_0$ with the largest p-value:

$$\begin{aligned} (P)\ \text{maximize: } & g(\boldsymbol{\tau}_0) = \sum_{i=1}^{n}\left(\tau_{0,i} - y_i + \mu_0\right)^2, \\ \text{subject to: } & \sum_{i=1}^{n}\tau_{0,i} = A_0, \\ & 0 \le \tau_{0,i} \le y_i,\ i = 1, \ldots, n. \end{aligned}$$

This optimization problem comes from the class of "quadratic convex maximization" problems [11]. While maximizing a convex function over a convex set is generally an NP-hard problem, effectively equivalent to enumerating

all possible vertices of the constraint space, the particular form of this problem allows for an efficient solution.

Theorem 1. *Let all $y_i \geq 0$. For $i = 1, \ldots, n$, sort the y_i such that,*

$$y_1 \geq y_2 \geq \cdots \geq y_n.$$

An optimal solution to P is given by:

$$\tau_{0,i} = \begin{cases} 0, & i < s, \\ A_0 - \sum_{i=s+1}^{n} y_i, & i = s, \\ y_i, & i > s, \end{cases}$$

where s is the largest integer such that $\sum_{i=s}^{n} y_i > A_0$.

A proof of Theorem 1 is given in the Appendix. Since this solution finds the maximum possible variance of the statistic (2) under the null $A = A_0$, we label this solution the *variance maximization method* of testing $A = A_0$. As this solution can be implemented using a simple sort of the n treated units, followed by a linear pass through the data, so the complexity of the algorithm is $O(n \log n)$ using typical sorting routines. While Theorem 1 does not assume the data are either real values or integer values, the solution also applies to integer constrained $\boldsymbol{Y}$.

Corollary 1. *When A_0 is an integer and all y_i are integers, the solution to the integer constrained version of P is also given by Theorem 1.*

A proof is given in the appendix.

It is important to note exactly what optimality guarantees Theorem 1 provides. Ultimately, we are seeking the $\boldsymbol{\tau}_0$ vector of adjustments that leads to the maximum p-value over all compatible $\boldsymbol{\tau}_0$ that sum to A_0. Theorem 1, however, finds the $\boldsymbol{\tau}_0$ vector that generates a null distribution for T with the *maximum variance*. When N is large, this distribution will be roughly normal, so the correspondence between maximum variance and maximum p-value will be close. For small samples, or when the normality approximation fails for other reasons such as high skew, this approximation may fail to find the the adjustment with the maximum p-value, despite having the largest variance. In the Appendix, we provide simulations investigating how well the maximum variance approximates the true maximum p-value and find it works well in a variety of samples sizes and data generation processes.

3. SIMULATIONS

We now consider the performance of prediction intervals for A using the method described in Section 2.2, which we refer to as the *variance maximization method*. As a benchmark method for comparison, we use survey sampling based intervals that rely on large sample approximations. Observe that the attributable effect A can be decomposed as

$$\begin{aligned} A &= \sum_{i=1}^{N} Z_i(y_i(1) - y_i(0)) \\ &= \sum_{i=1}^{N} Z_i y_i(1) - \left(\sum_{i=1}^{N} y_i(0) - \sum_{i=1}^{N} (1 - Z_i) y_i(0) \right). \end{aligned}$$

The quantities $\sum_{i=1}^{N} Z_i y_i(1)$ and $\sum_{i=1}^{N}(1-Z_i)y_i(0)$ are completely observed as the totals in the treatment and control groups, respectively. While the total $\sum_{i=1}^{N} y_i(0)$ is not observed, it can be estimated using standard sample survey techniques [30, 15, 35]. This leads to a large sample prediction interval for A:

$$\hat{A} \pm t_{1-\alpha/2} \sqrt{N \frac{n}{m} s_0^2} \tag{3}$$

where $\hat{A} = \sum_{i=1}^{N} Z_i Y_i - \frac{n}{m} \sum_{i=1}^{N}(1 - Z_i)Y_i$, s_0^2 is the sample variance for the control units, and $t_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile from a Student's t-distribution with $m-1$ degrees of freedom (additional details on the derivation of this interval are given in the appendix).

In these simulations, based on 1000 replications, we vary the total experiment size (N), the proportion of the $y_i(0)$ that are zero (p), and the true effect size (e). In each simulation, we compare the coverage rate of a 95% prediction interval as well as the ratio of interval width for the proposed method and that of survey sampling technique. Detailed information on the simulation process is given in the appendix.

In the first simulation, we varied the experimental population size between 10 and 500, treating half of the population in each experiment. Figure 1 shows the results of the simulations. As the figure shows, the variance maximization method is conservative. The survey sampling method under-covers at lower samples sizes but achieves nominal level for the larger sample sizes. For the width of the intervals, the survey sampling method is always smaller, on average, though it is often under-covering for sample sizes less than 200.

In the second simulation, the parameter p, the probability of $y_i(0)$ being zero, varied from 0 to 0.95. Figure 2 again shows the 95% prediction interval coverage and relative widths. With a sample size of 100, the survey sampling based method has modest under-coverage for several values of p. As the proportion of zeros increases, the variance maximizing method becomes less conservative, but never falls below its nominal level. Recall that the solution to the variance maximization optimization problem sets the hypothesized $y_i(0)$ to zero (i.e., $\tilde{y}_i(0) = 0$) for the smallest observed treated units. As the proportion of zeros increases, this solution approaches the true $\boldsymbol{\tau}_0$, whereas in general cases the solution is only guaranteed to generate a p-value larger than

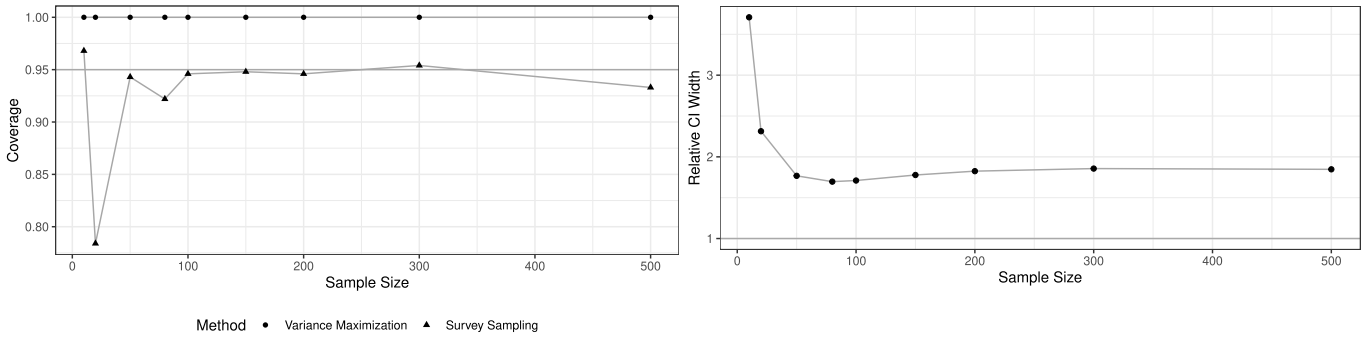

Figure 1. Sample size simulation ($N \in [10, 500]$, $n = N/2$, $p = 0.1$, $e = 1$).

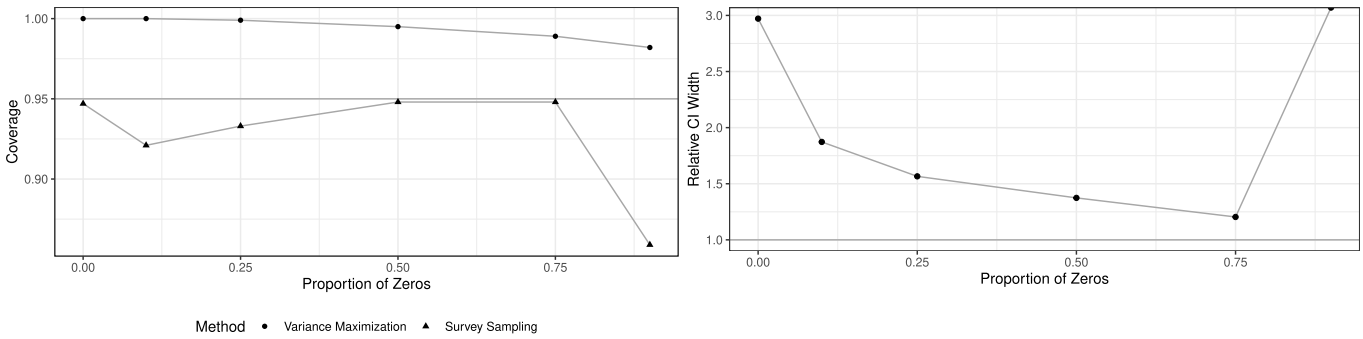

Figure 2. Proportion of zeros in $y(0)$ simulation ($p \in [0, 0.95]$, $e = 1$, $N = 100$, $n = 50$).

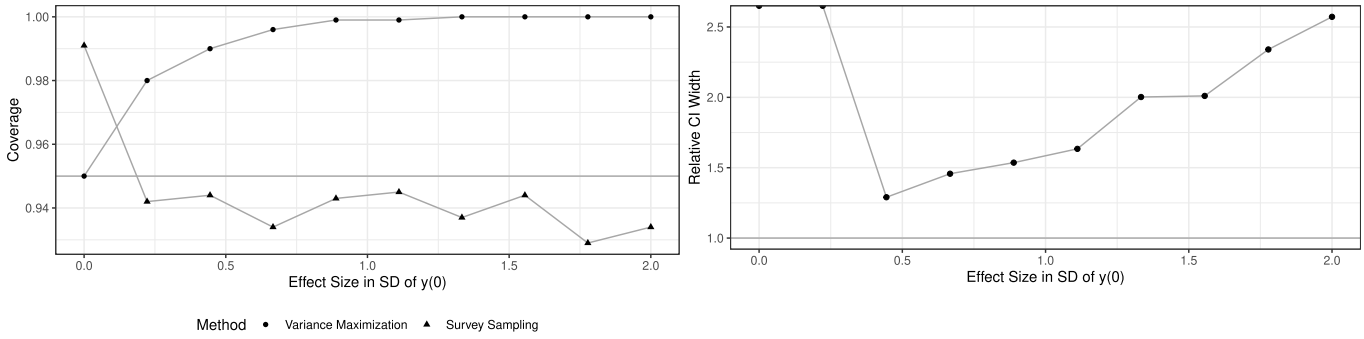

Figure 3. Effect size simulation ($e \in [0, 2]$, $p = 0.1$, $N = 100$, $n = 50$).

that of the true $\boldsymbol{\tau}_0$. Consequently, the method performs particularly well in the case of zero-inflated outcomes. The relative width of the variance maximization interval tends to approach the width of the survey sampling interval; however, the trend reverses for $p = 0.95$, suggesting there may be a limit to the proportion of zeros that this method can efficiently handle.

The third simulation varies the total effect size $\boldsymbol{\mathcal{T}} = \sum_{i=1}^{N} \tau_i$ as a function of the standard deviation of the randomly generated $y_i(0)$. Figure 3 shows the results as the effect size was varied between zero and two standard deviations. As the top panel of Figure 3 shows, the proposed method maintains consistently conservative coverage rates across the different effect sizes, while the survey sampling method under covers somewhat. Interestingly, the relative size of the intervals for the proposed method tended to increase as the effect size increased. This scenario represents the opposite of the zero-inflated situation: as the total effect increases, the true $\boldsymbol{y}(0)$ is less and less like $\tilde{\boldsymbol{y}}(0)$ as the large effect size makes more and more of the τ_i large.

Looking across these simulations, the overall pattern emerges that, at least on these data, the proposed method appears to work well in small samples and when there is a great degree of treatment heterogeneity. The variance max-

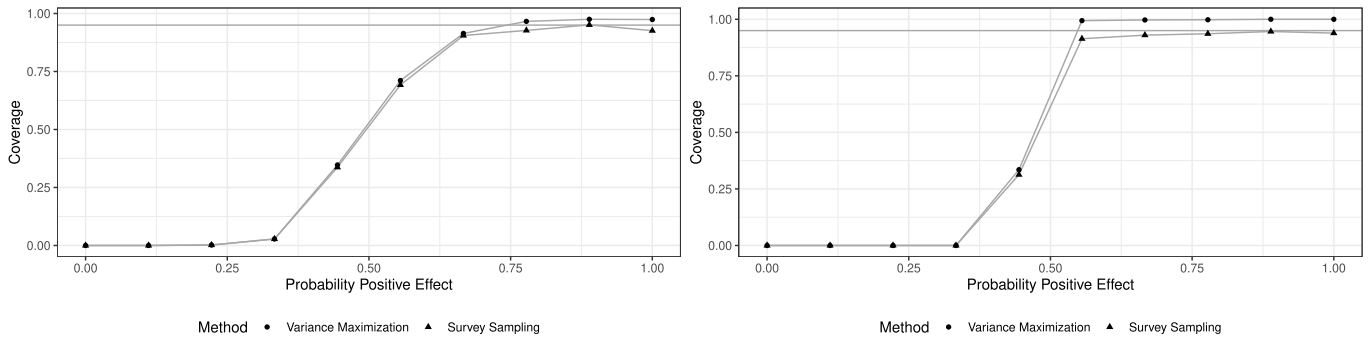

Figure 4. Prediction interval coverage for a simulation in which individual treatment effects may be non-monotonic. The left panel shows a small average individual effect magnitude; the right panel shows larger effects.

imization method is almost always conservative in its coverage rates and has reasonably small interval widths in small samples or when there are many zeros in the data.

We also considered the situation in which the key assumption of monotonicity *does not hold*. In these simulations, we draw $y_i(0)$ and create a magnitude for each individual treatment effects such that $0 \leq \tau_i \leq y_i(0)$. We then vary the probability of a positive treatment effect (θ) from zero to one. Figure 4 shows the coverage rates of 95% prediction intervals as the probability of a positive individual treatment effect is varied from zero to one. The top panel shows a simulation in which the magnitude of the individual effects is small, relative to $y_i(0)$. The bottom panel shows a simulation in which the magnitude of the individual effects is relatively large. In both cases, when most effects are negative ($\theta < 0.5$), the coverage rates of the prediction intervals are quite poor. This is no surprise as $A < 0$ in most cases, and the prediction intervals are constrained to be non-negative. On the other hand, provided $\theta > 0.5$, the method appears relatively robust to having some negative treatment effects. Particularly, when individual effects are relatively large, the coverage of intervals quickly approaches the nominal level. Nevertheless, the possibility remains that when effects are negative the algorithm will not be conservative in finding the maximum possible variance treatment allocation. The true hypothesis can have larger variance than would be possible with strictly positive effects, leading to the under-coverage of intervals demonstrated in Figure 4. Interestingly, the survey sampling method performs similarly, so there is little relative advantage of one method or the other in these cases.

Additional details on the simulations used in this section are given in the appendix. Also included are simulations investigating the fidelity of the null approximation, correct behavior of the optimization routine, and variation in treatment effect distributions.

4. OREGON HEALTH INSURANCE EXPERIMENT

In 2008 the state of Oregon re-opened enrollment for Oregon Healthcare Plan (OHP) Standard, a medical insurance program for low-income households who did not otherwise qualify for health insurance. As enrollment in this program had been closed for several years, officials anticipated a higher demand than could be accommodated under the available budget. To address the issue of over-subscription, state officials implemented a lottery system to allocate opportunities to apply to the program. After an advertising campaign to solicit potential recipients, 74,922 individuals applied for the program. The initial solicitation did not require individuals show eligibility for the program, so being randomly selected into the study provided individuals the opportunity to complete an application, demonstrating eligibility in the program. Of the 74,922 applicants, 29,834 individuals were randomly selected to receive an invitation to apply for OHP. Of these, 8,698 applied and were approved to enroll in OHP. More details on the program and randomization process can be found in Finkelstein et al. [9].

After 12 months, a portion of both the treated individuals (selected to complete an application) and the control individuals (not permitted to apply) were sent a survey requesting self-reported amounts of money spent out-of-pocket for medical care during the previous 6 months. Responses to the question included many zeros and were heavily right skewed. Of the subjects that responded to the questionnaire ($N =$ 22,766), 53% claimed no out of pocket costs in the last 6 months, while 6 individuals reported out of pocket costs in excess of $100,000. Excluding subjects who reported zero out of pocket costs, the median cost reported was $250.

To analyze these questions, we first suppose that for all subjects Assumption 2 holds: $0 \leq y_i(1) \leq y_i(0)$. This assumption supposes that having medical insurance will not raise a subject's out of pocket costs. As the Medicaid program covers nearly all medical costs, this assumption seems plausible. Before applying the methods proposed in this paper, we first create a dichotomous variable indicating whether a subject reported spending more than zero dollars on health care. Applying the method of Rosenbaum [31] to predict the attributable effect yields a 95% prediction interval of $[597, 889]$. This result suggests that had the control subjects had the opportunity to apply for state sponsored

Table 1. 95% prediction intervals for the attributable effect of not having the opportunity to apply for OHP Standard on total out of pocket costs. Numbers in parentheses represent percentage of maximum attributable effect, which is the sum of all control units' outcomes

Method	Lower 95% Prediction Interval	Upper 95% Prediction Interval
Survey Sampling	0 (0)	4,704,329 (100)
Variance Maximization	0 (0)	1,283,000 (27)

health care, between 10.3 and 15.4 percent who had out of pocket costs would have been able to avoid them.

While such savings may be beneficial no matter the overall amount spent, by dichotomizing the cost, this analysis may confuse substantive changes in the amount the control group spent for small, but consistent changes. To answer this question, we apply the proposed method in this paper to predict the attributable effect on the dollar scale and compare it to the survey sampling method. Table 1 shows the results of the different methods. While both methods include an attributable effect of zero in their estimates, the survey sampling method produces an interval that gives no information as it includes every possible value for the attributable effect. The variance maximization method, however, excludes attributable effects greater than 27 percent of the observed total in the control group (at 95% confidence). Had the control group been allowed to apply for health insurance, they would have spent less on health care, but a decrease in spending by more than 27% appears implausible given these results.

Replacing usage of emergency departments with scheduled medical visits is often touted as a justification for expanding government sponsored medical insurance. As emergency departments by law must provide care, regardless of the individual's lack of medical insurance, advocates argue that providing medical insurance can actually *decrease* overall spending as insured individuals can better take advantage of less expensive scheduled care. On the other hand, while emergency departments must treat subjects, they will still bill patients without medical insurance. Having access to Medicaid might incentivize individuals to consume more medical services, in particular emergency department visits, as their own costs will significantly decrease. To answer this controversy, Taubman et al. [36] matched subjects in the Portland, OR area to hospital records to tabulate the number of emergency department visits per subject. Overall, the subset of the experimental population included 24,646 subjects. Again, these data show a large portion of zero values (16,180) and strong right skew.

To address the competing theories of the effect of health insurance on emergency room visits, we analyze the data under both possible assumptions for monotonicity, that either the health insurance does not increase any household's emergency room usage or that insurance does not decrease emergency room usage. We predict the attributable effect for the treated subjects under assumption that none of their usages would decrease as well as the attributable effect for the control subjects under the assumption of that usage would not increase under health insurance. Table 2 provides the results of these tests, again comparing the survey sampling method to the variance maximization method. Both methods tend to favor Assumption 1 — that health care access does not decrease emergency room usage — as the prediction intervals contain a larger portion of the observed data, though all intervals include zero leaving the possibility of no effect or non-monotonic potential outcomes. These results bear some similarities to those reported in Taubman et al. [36], where the authors dichotomized these data at several usage levels and found that treated subjects made more use of emergency facilities.

5. DISCUSSION

In this paper, we presented a novel method for testing hypotheses for the effect attributable to treatment, the sum of the individual effects of the subjects within the treatment group. This method expands the scope of attributable effects to count and continuous data, provided the researchers are able to assume that effects are non-negative and that responses under the treatment condition are no less than responses under the control condition (Assumption 1). Alternatively, this method can be applied to recover the sum of treatment effects for the control group when control responses are assumed to be greater than treatment responses (Assumption 2). This method is computationally efficient, and simulations show that using a normal approximation to the true null distribution adds little error compared to the true solution. From a statistical perspective, the method appears to perform well in small samples or when there is a high degree of treatment effect heterogeneity. This method might be most useful when combined with a prescreening method used to detect heterogeneity [e.g., 5] and employed only if the constant treatment effect assumption seems to be a poor approximation to the true treatment effect distribution.

To evaluate the new method, we compared it to the survey sampling based method that estimates A [30, 15] and found favorable results, particularly in small samples or with many zeros. It should be noted that the survey sampling based estimator made use of neither the assumption of monotonicity nor the constraint that $0 \leq A \leq \sum_{i=1}^{N} Z_i Y_i$. While it lies outside the scope of the current paper to amend the estimator to take advantage of these assumptions, there is a lengthy literature on both using monotonic-

Table 2. 95% prediction intervals for the attributable effect of number of emergency department visits. Under the assumption that $0 \leq y_i(0) \leq y_i(1)$, the effect of having the opportunity to apply for Medicaid is identified. Under the assumption that $0 \leq y_i(1) \leq y_i(0)$, the effect for the control group is identified. Numbers in parentheses indicate the percentage of the observed total attributed to the treatment

Method	Lower 95% Prediction Interval	Upper 95% Prediction Interval
	Assuming $0 \leq y_i(0) \leq y_i(1)$	
Survey Sampling	0 (0%)	1068 (10.7%)
Variance Maximization	0 (0%)	1077 (10.8%)
	Assuming $0 \leq y_i(1) \leq y_i(0)$	
Survey Sampling	0 (0%)	213 (1.5%)
Variance Maximization	0 (0%)	160 (1.1%)

ity assumptions to derive bounds for average treatment effects [26, 19, 4, 13, 17] as well as estimating means under boundedness assumptions [1, 25, 7], which could be combined to provide a more efficient estimator under the assumptions invoked for the proposed method.

Another possible extension lies in bounding all potential outcomes from above, as well as below. Assuming an *a priori* upper bound c such that $0 \leq y_i(0) \leq y_i(1) \leq c$ for all units would allow simultaneously testing hypotheses about the attributable effect in both the treated and control groups. The average treatment effect is then the sum of the attributable effects in each group divided by the number of units in the study. This is the approach taken in Rigdon and Hudgens [29] for binary data where $c = 1$.

We conclude with a discussion of a seeming contradiction between the proposed method and well known methods for estimating average treatment effects. The statistic used in the optimization routine at the heart of the proposed approach is very similar to the difference of means statistic frequently used to estimate average treatment effects. The variance of this statistic is given by $S_1^2/n + S_0^2/m - S_\tau^2/N$, where S_1^2, S_0^2 and S_τ^2 represent the sample variance of $\boldsymbol{y}(1)$, $\boldsymbol{y}(0)$, and $\boldsymbol{\tau}$ respectively. A conservative estimator of this variance omits the unidentifiable S_τ^2 term, which is equivalent to having a constant treatment effect. In the proposed approach, however, the solution to the optimization problem suggests $\boldsymbol{\tau}$ with maximum variance.

The resolution to this contradiction emerges when we realize that $\boldsymbol{y}(1) = \boldsymbol{y}(0) + \boldsymbol{\tau}$, so there are really only two degrees of freedom in the variance expression, which can be written only in terms of control potential outcomes and treatment effects,

$$\frac{S_1^2}{n} + \frac{S_0^2}{m} - \frac{S_\tau^2}{N} = \left(\frac{1}{n} + \frac{1}{m}\right) S_0^2 + \left(\frac{1}{n} - \frac{1}{N}\right) S_\tau^2 - \frac{2}{n} S_{0\tau}.$$

This expression shows that a conservative estimator would assume large treatment effect variation and a negative covariance between treatment effects and control potential outcomes ($S_{0\tau}$), which is what the solution to the optimization problem establishes when it implies that the control potential outcomes for units with the largest observed $y_i(1)$ values are zero.

APPENDIX A. LARGE SAMPLE PREDICTION INTERVALS FOR A

By definition, for any given $\boldsymbol{Z}$, the attributable effect of treatment can be decomposed as

$$A = \sum_{i=1}^{N} Z_i y_i(1) - \left(\sum_{i=1}^{N} y_i(0) - \sum_{i=1}^{N} (1 - Z_i) y_i(0) \right).$$

As $\sum_{i=1}^N Z_i y_i(1) = \sum_{i=1}^N Z_i Y_i$ and $\sum_{i=1}^N (1 - Z_i) y_i(0) = \sum_{i=1}^N (1 - Z_i) Y_i$ are observed quantities, we need only estimate $\sum_{i=1}^N y_i(0)$ using

$$\hat{Y}_0 = \frac{N}{m} \sum_{i=1}^{N} (1 - Z_i) Y_i.$$

Plugging this estimator into the decomposition of A yields

$$\begin{aligned} \hat{A} &= \sum_{i=1}^{N} Z_i Y_i - \left(\frac{N}{m} \sum_{i=1}^{N} (1 - Z_i) Y_i - \sum_{i=1}^{N} (1 - Z_i) Y_i \right) \\ &= \sum_{i=1}^{N} Z_i Y_i - \frac{n}{m} \sum_{i=1}^{N} (1 - Z_i) Y_i. \end{aligned}$$

Under conditions stated in Section 2.2, when N is large, $\hat{Y}_0$ is approximately normal with mean $\sum_{i=1}^N y_i(0)$ and variance $N \frac{n}{m} \sigma_0^2$ [3, Theorem 2.2], where σ_0^2 is the finite population variance of $y_i(0)$. By estimating σ_0^2 with s_0^2, the sample variance of the control units, a $100 \times (1 - \alpha)\%$ prediction interval for A has the form:

$$\sum_{i=1}^{N} Z_i Y_i - \frac{n}{m} \sum_{i=1}^{N} (1 - Z_i) Y_i \pm t_{1-\alpha/2} \sqrt{N \frac{n}{m} s_0^2},$$

where $t_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of a t-distribution with $m - 1$ degrees of freedom.

APPENDIX B. PROOF OF THEOREM 1

Proof. Recall that we wish to maximize:

$$\begin{aligned} g(\boldsymbol{\tau}_0) &= \sum_{i=1}^{n} \left(\tau_{0,i} - y_i + \mu_0)\right)^2 \\ &= \sum_{i=1}^{n} \tau_{0,i}^2 + \sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n} \mu_0^2 - \\ &\quad 2\sum_{i=1}^{n} \tau_{0,i} y_i + 2\mu_0 \sum_{i=1}^{n} \tau_{0,i} - 2\mu_0 \sum_{i=1}^{n} y_i \\ &= \sum_{i=1}^{n} (y_i - \tau_{0,i})^2 + \sum_{i=1}^{n} \mu_0^2 + 2\mu_0 A_0 - 2\mu_0 \sum_{i=1}^{n} y_i. \end{aligned}$$

As the term $\sum_{i=1}^{n} \mu_0^2 + 2\mu_0 A_0 - 2\mu_0 \sum_{i=1}^{n} y_i$ does not depend on τ_0, maximizing $g(\boldsymbol{\tau}_0)$ is equivalent to maximizing

$$h(\boldsymbol{\tau}_0) = \sum_{i=1}^{n} (y_i - \tau_{0,i})^2.$$

In other words, we can equivalently maximize the sum of squared remainders left after removing $\tau_{0,i}$. Writing $r_i = y_i - \tau_{0,i}$, rewrite the maximization problem as

$$\begin{aligned} (P') \text{ maximize: } & h(\boldsymbol{r}) = \sum_{i=1}^{n} r_i^2 \\ \text{subject to: } & \sum_{i=1}^{n} r_i = \sum_{i=1}^{n} y_i - A_0 = R_0 \\ & 0 \le r_i \le y_i \end{aligned}$$

A simple greedy algorithm provides an optimal solution to P'. Sort the observations so that $y_1 \ge y_2 \ge \cdots \ge y_n$. Initialize $R_0^{(1)} = R_0$. For $i = 1, \ldots, n$, do:

1. If $y_i \ge R_0^{(i)}$, set $x_i = R_0^{(i)}$. For all $j > i$, set $r_j = 0$ and stop.
2. Otherwise, set $r_i = y_i$ and $R_0^{(i+1)} = R_0^{(i)} - y_i$.
3. If $i = n$, stop. Otherwise, update $i = i + 1$ and repeat the loop.

Let s be the largest integer such that $\sum_{i=1}^{s-1} y_i < R_0$. The result of the algorithm $\boldsymbol{r}$ has the form:

$$r_i = \begin{cases} y_i, & i < s, \\ R_0 - \sum_{i=1}^{s-1} y_i, & i = s, \\ 0, & i > s. \end{cases}$$

To show this is optimal, we show that we can transform any optimal solution into the greedy solution. Let $\boldsymbol{r}$ be the solution found by the greedy algorithm and $\tilde{\boldsymbol{r}}$ be any optimal solution. At each stage of the following algorithm, transform $\tilde{r}_i$ into r_i while maintaining the objective function value $h(\tilde{\boldsymbol{r}})$. At each state the proposed optimal solution has $\tilde{r}_j = r_j$ for $j < i$. Starting from $i = 1$,

1. If $\tilde{r}_i = r_i$, continue to $i + 1$.
2. Otherwise, consider the two possible values of r_i:
   (a) $r_i = R_0^{(i)}$: Observe that in this case $r_j = 0$ for $j > i$. As the solution $\tilde{\boldsymbol{r}}$ is feasible, it must be the case that $\sum_{j=i}^{n} \tilde{r}_j = R_0^{(i)} = r_i$. Since the $\tilde{r}_j$ are non-negative, this implies a contradiction that $h(\tilde{\boldsymbol{r}})$ is maximal:

$$\begin{aligned} h(\boldsymbol{r}) - h(\tilde{\boldsymbol{r}}) &= (R_0^{(i)})^2 - \sum_{j=i}^{n} \tilde{r}_j^2 \\ &= \left(\sum_{j=1}^{n} \tilde{r}_j\right)^2 - \sum_{j=i}^{n} \tilde{r}_j^2 \\ &= \sum_{j=i}^{n} \sum_{j'=i}^{n} \tilde{r}_{j'} > 0. \end{aligned}$$

   Therefore, when $y_i \ge R_0^{(i)}$ the only optimal solution is the greedy one. At this point, we can stop, having found that the greedy solution is optimal.

   (b) $y_i < R_0^{(i)}$ and $r_i = y_i$. Since $\tilde{r}_i$ is also bounded by y_i, it must be the case that $\tilde{r}_i < r_i$. Again, since $\sum_{j=i}^{n} \tilde{r}_j = R_0^{(i)}$ and $\tilde{r}_j < r_i < R_0^{(i)}$, there must exist at least one $j > i$ such that $\tilde{r}_j > 0$. Let $\delta = \min(y_i - \tilde{r}_i, \tilde{r}_j)$. Then the solution $\hat{\boldsymbol{r}} = \tilde{r}_1, \ldots, \tilde{r}_i + \delta, \ldots, \tilde{r}_j - \delta, \ldots, \tilde{r}_n$ is also feasible. Comparing the difference of objective functions, we see:

$$\begin{aligned} h(\hat{\boldsymbol{r}}) - h(\tilde{\boldsymbol{r}}) &= (\tilde{r}_i + \delta)^2 - \tilde{r}_i^2 + (\tilde{r}_j - \delta)^2 - \tilde{r}_j^2 \\ &= 2\delta^2 + 2\tilde{r}_i\delta - 2\tilde{r}_j\delta. \end{aligned}$$

   As δ is the lesser of $y_i - \tilde{r}_i$ or $\tilde{r}_j$, consider both cases:

   i. $\delta = y_i - \tilde{r}_i$: Then

$$\begin{aligned} \delta^2 + \tilde{r}_i\delta - \tilde{r}_j\delta &= y_i^2 - y_i\tilde{r}_i - \tilde{r}_j y_i + \tilde{r}_i\tilde{r}_j \\ &= (y_i - \tilde{r}_i)(y_i - \tilde{r}_j) \end{aligned}$$

   We already know that $y_i > \tilde{r}_i$. By the ordering of units, since $i > j$, we know that $y_i \ge y_j \ge \tilde{r}_j$). Therefore $(y_i - \tilde{r}_i)(y_i - \tilde{r}_j) \ge 0$ so the solution $\hat{\boldsymbol{r}}$ is also optimal. Since $\delta = y_i - \tilde{r}_i$, then $\hat{r}_i = y_i = r_i$.

   ii. $\delta = \tilde{r}_j$: Then

$$\delta^2 + \tilde{r}_i\delta - \tilde{r}_j\delta = \tilde{r}_j^2 + \tilde{r}_i\tilde{r}_j - \tilde{r}_j^2 = \tilde{r}_i\tilde{r}_j$$

   As both $\tilde{r}_i \ge 0$ and $\tilde{r}_j \ge 0$, the solution $\hat{\boldsymbol{r}}$ is also optimal. As $\hat{r}_i = \tilde{r}_i + \tilde{r}_j < r_i$, it must be the case that some other unit j' is also non-zero and can be used to create $\delta' = \min(y_i - \hat{r}_i, \tilde{r}_{j'})$ and another optimal solution. This logic can be repeated until an optimal solution can be found that includes $\hat{r}_i = y_i$.

3. Update $\tilde{r}_i = \hat{r}_i = r_i$. At this point, $\tilde{r}_j = r_j$ for all $j \leq i$.
4. Continue for $i+1$ and $R_0^{(i+1)} = R_0^{(i)} - r_i$.

At the end of this algorithm, $\tilde{\boldsymbol{r}} = \boldsymbol{r}$, the greedy solution, showing that any optimal solution can be transformed into the greedy solution while maintaining $h(\boldsymbol{r}) \geq h(\tilde{\boldsymbol{r}})$ at each step.

With a solution $\boldsymbol{r}$ to P', we can then translate back to P using the relationship $\boldsymbol{\tau}_0 = \boldsymbol{r} - \boldsymbol{y}$. Consequently, $\boldsymbol{\tau}_0$ has the form:

$$\tau_{0,i} = \begin{cases} 0, & i < s, \\ A_0 - \sum_{i=s+1}^{n} y_i, & i = s, \\ y_i, & i > s, \end{cases}$$

where s is the largest integer such that $\sum_{i=s}^{n} y_i > A_0$. □

APPENDIX C. PROOF OF COROLLARY 1

Proof. Observe that P is the continuous relation of the version of the problem for integer y_i. Let τ^* be the solution to P. For $i > s$, $\tau_i^* = y_i$, which are integer values. For $i = s$, $\tau_s^* = A_0 - \sum_{i=s+1}^{n} y_i$ is also an integer, as A_0 is an integer and the sum of any y_i values must also be an integer. For $i < s$, $\tau_i^* = 0$. Thus τ^* is an integer solution and must be the optimal solution to the integer constrained version of P. □

APPENDIX D. ADDITIONAL SIMULATION DETAILS

D.1 Testing the normal approximation

For each hypothesized attributable effect A_0, there may be many compatible unit level sharp hypotheses $\boldsymbol{\tau}_0$ such that $\sum_{i=1}^{N} Z_i \boldsymbol{\tau}_i = A_0$. Rejecting A_0 at the α level implies that all compatible hypotheses must also be rejected at the α level. In the suggested methodology of Section 2.2, we propose using a normal approximation to the null distribution to find $\boldsymbol{\tau}_0$ with the largest p-value. We now present several simulations to assess how well the approximation works.

For n treated units and a hypothesis A_0, there are at most $\binom{n+A_0-1}{n}$ ways to allocate the A_0 to the n treated units when the potential outcomes $y_i(1)$ and $y_i(0)$ are integer values. For small experiments, these allocations can be explicitly enumerated to find the $\boldsymbol{\tau}_0$ vector with the largest p-value. This presents a way to compare how well the approximation holds in finding the largest p-value, at least for a sufficiently small experiments and effect sizes for which all $\binom{n+A_0-1}{n}$ possible allocations can be enumerated and checked.

For a small experiment ($N = 10$, $n = 5$), we generated $\boldsymbol{y}(\boldsymbol{0})$ and then allocated A to the different units, with $A \in \{1, \ldots, 6\}$. The true A was either spread out or clustered it on only a few units. Figure 5 shows the relative error of the p-value from the normal approximation comapred to p-value from complete enumeration. On the whole, the approximation works quite well, even for this small experiment. The approximation performed least well in these examples where the true treatment effect was larger and evenly distributed. Recalling that the solution to the approximation concentrates the adjustments to the smallest values, it makes sense that the approximation does not perform well in this situation.

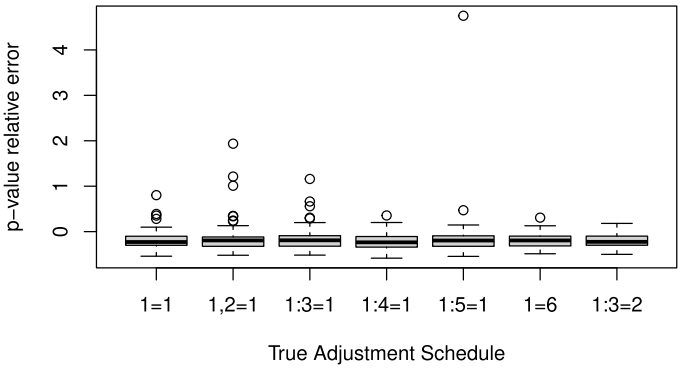

Figure 5. Boxplot comparing of normal approximation maximum p-value ($\hat{p}$) to true maximum p-value (p) using relative error $(p - \hat{p})/p$. The x-axis labels indicate the units that had positive τ_i values. For example, "1:3=2" indicates that $\tau_1 = \tau_2 = \tau_3 = 2$ and $\tau_i = 0$ for $i > 3$.

As an additional check on the performance of the algorithm, the variance of T generated by the adjustment schedule found by the proposed algorithm was compared to the variances of T for all possible adjustments via enumeration. In all simulations, the adjustment selected had the largest variance of any possible solution. While this does not always imply the largest p-value, as seen in Figure 5, the algorithm is performing its job properly. As the sample size increases and the normal approximation improves, the accuracy with respect to finding the true maximum p-value should increase.

To test the suitability using the variance of the null distribution to approximate the p-values of the sharp null hypotheses, we simulated a small experiment with 10 units from which 5 were assigned to treatment. First, the potential responses under control were simulated as:

$$y_i(0) = P + 20B,\ P \sim \text{Poisson}(7),\ B \sim \text{Binomial}(0.01, 2).$$

Next, the set of true treatment effects were added to the treated units' scores based on Table 3. The columns represent the treatment unit, and each row shows the individual effect of the treatment $\tau_{0,i}$. The true attributable effect for each row is the sum of row values. The first experiment adds one to the first treated unit, the second adds one to both the first and second, and so on. We also consider placing a much larger effect of six on the first unit and adding two to the first three units. For each allocation, the true attributable effect $A = \sum_{i=1}^{5} \tau_i$ was used to generate $\boldsymbol{y}(\boldsymbol{1})$ from $\boldsymbol{y}(\boldsymbol{0})$ and a hypothesis test of $A_0 = A$ was performed

Table 3. Strategies for allocating treatment effects used in small sample size simulations. Columns represent the true effect of treatment for each of treated units. The attributable effect A is the sum of the row values

	1	2	3	4	5
1	1	0	0	0	0
2	1	1	0	0	0
3	1	1	1	0	0
4	1	1	1	1	0
5	1	1	1	1	1
6	6	0	0	0	0
7	2	2	2	0	0

using the normal approximation strategy. Recall that the normal approximation is guaranteed to find the adjustment that leads to the largest variance of the null distribution of the test statistic T, but this may not correspond to the adjustment with the largest p-value, which is the true target. By enumerating all compatible allocations $\boldsymbol{\tau}_0$ and performing an exact randomization test, we can find the adjustment with maximum p-value and compare this p-value found by the normal approximation by computing the relative error $|p - \hat{p}|/p$, where p is the largest p-value and $\hat{p}$ is found from the method given in Section 2.2. In both cases, p-values were generated by completely enumerating all $\binom{10}{5}$ possible treatment allocations, generating the null distribution of the test statistic T^2, and comparing the observed test statistic to the null distribution. The simulations were repeated 100 times, each with a new $\boldsymbol{y}(\mathbf{0})$, for each true allocation.

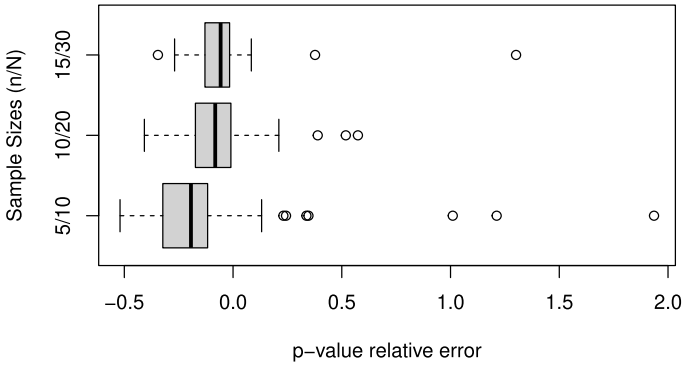

Figure 6. Boxplot of relative error when finding the largest p-value using the normal approximation method compared to complete enumeration. For $N = 10, n = 5$, p-values are computed exactly. For the simulations with 10 out of 20 and 15 out of 30 assigned to treatment, p-values are computed using 10,000 Monte Carlo samples.

In order to completely enumerate all possible treatment allocations compatible with a given A_0 as well as perform exact hypothesis tests, the simulations so far have been kept fairly small. To consider the effect of sample size on the performance of the variance maximization method, we repeated the simluations for larger experiments using 10 out of 20 treated and 15 out of 30 treated. For each experiment, the true treatment effect was 1 for 2 of the treated units and zero for the remainder. These experiments start to push the boundaries of convenient computation when completely enumerating the entire randomization distribution, so a sample of 10,000 treatment assignments was used instead. If the method is working well, the distribution of p-values under the null should be approximately uniform when the null hypothesis is true. Figure 6 shows that the method performs reasonable well by this metric.

D.2 Coverage and Predicition interval widths

The main paper reports three simulations comparing the proposed methods to the survey sampling based method. Here we provide additional details on the simulation process.

For each simulation, the $\boldsymbol{y}(\mathbf{0})$ data were generated for N experimental subjects using power law type distribution:

$$y_i(0) = \lfloor 2^{10B} \rfloor - 1, B \sim \text{Beta}(2, 5).$$

This model was chosen to mimic several of the features of the observed data in the Oregon Health Insurance experiment, with many zeros and a very long right tail.

In order to create a full experiment, we must also generate $\boldsymbol{y}(\mathbf{1})$. To get the individual treatment effects τ_i, the population-level standard deviation σ_0 for the $\boldsymbol{y}(\mathbf{0})$ values are measured and a total effect computed as $\boldsymbol{\mathcal{T}} = \lfloor eN\sigma_0 \rfloor$, where e is the effect size multiplier. As $\boldsymbol{y}(\mathbf{0})$ were discrete, the total treatment effect must be applied in integer amounts. There are $\binom{N+\boldsymbol{\mathcal{T}}-1}{N-1}$ possible ways to distribute the total effect $\boldsymbol{\mathcal{T}}$ to the N units. One was chosen uniformly at random and used to generate $\boldsymbol{y}(\mathbf{1})$.

For 5000 replications, a treatment assignment was generated and the observed data were created using the $y_i(1)$ for the treated units and the $y_i(0)$ for the control units. The true value of A was computed by subtracting the true $\boldsymbol{y}(\mathbf{0})$ from the observed data. For each replication, 95% prediction intervals were generated using the proposed method and the survey sampling method. The interval widths were recorded as well as whether the intervals covered the true A value. To compute the p-value for the proposed method, 1000 Monte Carlo samples from the assignment mechanism were used.

D.3 Variation in treatment effect distributions

The simulations reported in the paper, and detailed in the previous section, randomly assigned treatment effects to individuals independently of $y_i(0)$. Whether treatment effects are correlated with $y_i(0)$ can also influence the power of the test, particularly for the variance maximization method. Instead of randomly assigning treatment effects, Figure 7 shows the cumulative distribution functions for p-value of

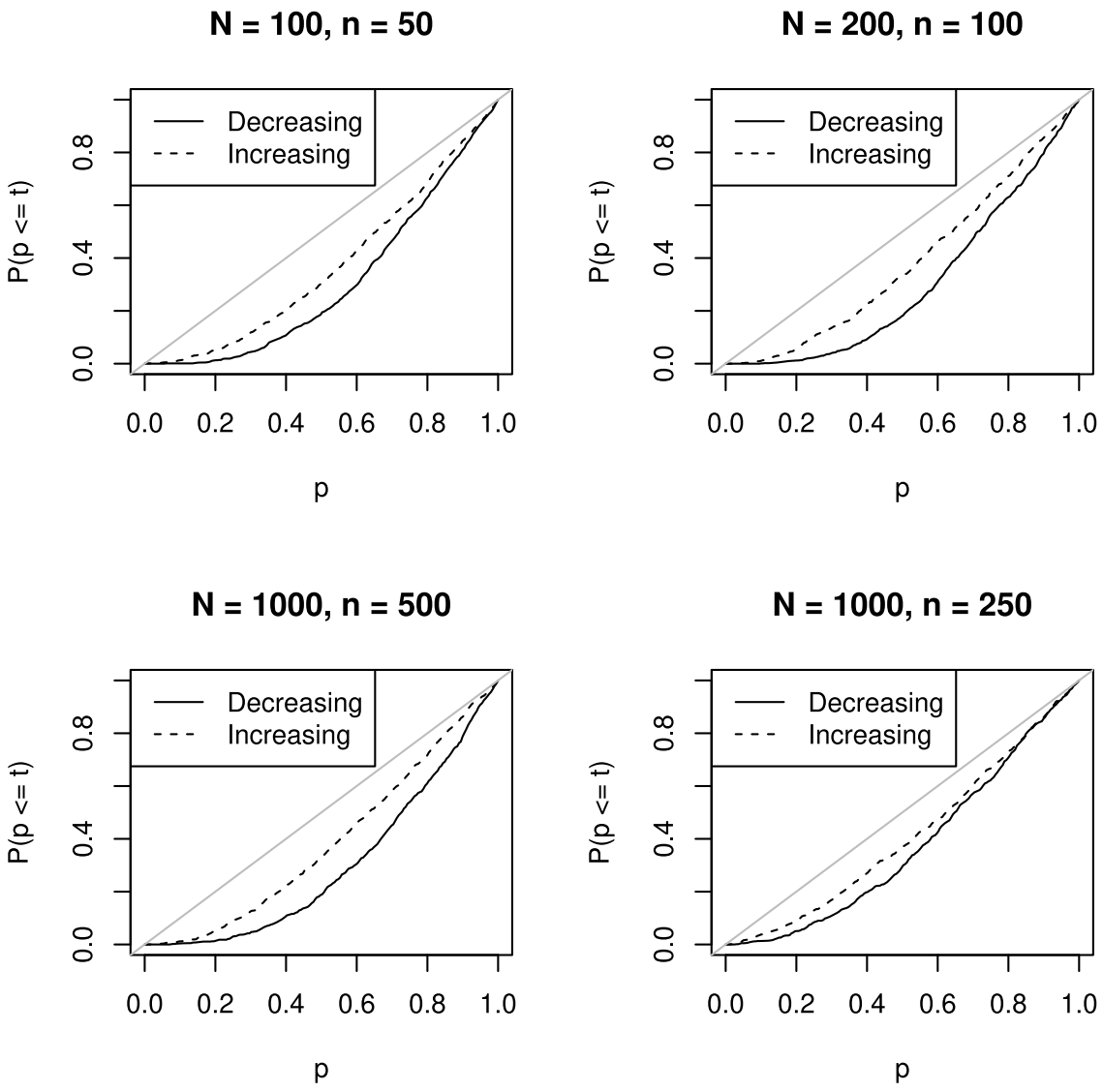

Figure 7. Cumulative distribution function of p-values when testing a true null hypothesis about A. All simulation parameters held at defaults. Potential outcomes to control are sorted such that $y_1(0) \geq y_2(0) \geq \cdots \geq y_N(0)$ and treatment effects are sorted in either increasing ($\tau_1 \leq \tau_2 \leq \cdots \leq \tau_N$) or decreasing ($\tau_1 \geq \tau_2 \geq \cdots \geq \tau_N$) order.

the test when the largest treatment effects are allocated to either the subjects with the largest $y_i(0)$ or smallest $y_i(0)$. To perform this simulation, $y_i(0)$ and treatment effects are generated using the simulation default settings. The $y_i(0)$ are sorted from largest to smallest and treatment effects are sorted in either increasing or decreasing order. When the effects are decreasing, the treatment helps the subjects that already have large $y_i(0)$ values; when the effects are increasing, the effects help those with the lowest $y_i(0)$. While the test is conservative for both sorting methods, it is less conservative when the largest effects are given to those with the smallest $y_i(0)$. As the optimization routine tests a hypothesis in which the treatment effects are concentrated on subjects with $y_i(0)$, it is unsurprising that the test is most powerful when the true treatment effect allocation is similar to the result of the optimization routine.

Received 9 July 2021

REFERENCES

[1] Casella, G. and Strawderman, W. E. (1981). Estimating a bounded normal mean. *The Annals of Statistics*, 9(4):870–878. MR0619290

[2] Choi, D. S. (2017). Estimation of monotone treatment effects in network experiments. *Journal of the American Statistical Association*, 112(519):1147–1155. MR3735366

[3] Cochran, W. (1999). *Sampling Techniques*. John Wiley & Sons, third edition. MR0474575

[4] Demuynck, T. (2015). Bounding average treatment effects: A linear programming approach. *Economics Letters*, 137(Supplement C):75–77. MR3432014
[5] Ding, P., Feller, A., and Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B*, 78(3):655–671. MR3506797
[6] Ding, P. and Miratrix, L. W. (2019). Model-free causal inference of binary experimental data. *Scandinavian Journal of Statistics*, 46(1):200–214. MR3915272
[7] Evans, S. N., Hansen, B. B., and Stark, P. B. (2005). Minimax expected measure confidence sets for restricted location parameters. *Bernoulli*, 11(4):571–590. MR2158252
[8] 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(2):773–797. MR3235398
[9] Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., and Baicker, K. (2012). The Oregon Health Insurance Experiment: Evidence from the first year. *The Quarterly Journal of Economics*, 127(3):1057–1106.
[10] Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.
[11] Floudas, C. A. and Visweswaran, V. (1995). Quadratic optimization. In Horst, R. and Pardalos, P. M., editors, *Handbook of Global Optimization*, pages 217–269. Springer US, Boston, MA. MR1377086
[12] Fogarty, C. B., Shi, P., Mikkelsen, M. E., and Small, D. S. (2017). Randomization inference and sensitivity analysis for composite null hypotheses with binary outcomes in matched observational studies. *Journal of the American Statistical Association*, 112(517):321–331. MR3646574
[13] Frandsen, B. R. and Lefgren, L. J. (2021). Partial identification of the distribution of treatment effects with an application to the Knowledge is Power Program (KIPP). *Quantitative Economics*, 12:143–171. MR4220376
[14] Hájek, J. (1961). Some extensions of the Wald-Wolfowitz-Noether theorem. *The Annals of Mathematical Statistics*, 32(2):506–523. MR0130707
[15] Hansen, B. B. and Bowers, J. (2009). Attributing effects to a cluster-randomized get-out-the-vote campaign. *Journal of the American Statistical Association*, 104(487):873–885. MR2562000
[16] Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960. MR0867618
[17] Huang, E. J., Fang, E. X., Hanley, D. F., and Rosenblum, M. (2017). Inequality in treatment benefits: Can we determine if a new treatment benefits the many or the few? *Biostatistics*, 18(2):308–324. MR3825122
[18] 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 (Statistics in Society)*, 180(2):569–586. MR3601784
[19] Kim, J. H. (2014). Identifying the distribution of treatment effects under support restrictions. Unpublished manuscript.
[20] Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, Inc., San Francisco. MR0395032
[21] Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, New York, third edition. MR2135927
[22] Li, X. and Ding, P. (2016). Exact confidence intervals for the average causal effect on a binary outcome. *Statistics in Medicine*, 35(6):957–960. MR3457618
[23] Lu, J., Ding, P., and Dasgupta, T. (2015). Construction of alternative hypotheses for randomization tests with ordinal outcomes. *Statistics & Probability Letters*, 107:348–355. MR3412795
[24] Lu, J., Ding, P., and Dasgupta, T. (2018). Treatment effects on ordinal outcomes: Causal estimands and sharp bounds. *Journal of Educational and Behavioral Statistics*, 43(5):540–567.
[25] Mandelkern, M. (2002). Setting confidence intervals for bounded parameters. *Statist. Sci.*, 17(2):149–172. MR1939335
[26] Manski, C. F. (1997). Monotone treatment response. *Econometrica*, 65(6):1311–1334. MR1604297
[27] Maritz, J. S. (1981). *Distribution-Free Statistical Methods*. Chapman and Hall, London. MR0644802
[28] Neyman, J. S. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–480. (Originally in Roczniki Nauk Tom X (1923) 1–51 (Annals of Agricultural Sciences). Translated from original Polish by Dambrowska and Speed.). MR1092986
[29] Rigdon, J. and Hudgens, M. G. (2015). Randomization inference for treatment effects on a binary outcome. *Statistics in Medicine*, 34(6):924–935. MR3310672
[30] Robins, J. M. (1988). Confidence intervals for causal parameters. *Statistics in Medicine*, 7(7):773–785.
[31] Rosenbaum, P. R. (2001). Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika*, 88(2):219–231. MR1841270
[32] Rosenbaum, P. R. (2002). Attributing effects to treatment in matched observational studies. *Journal of the American Statistical Association*, 97(457):183–192. MR1963391
[33] Rosenbaum, P. R. (2020). *Design of Observational Studies*. Springer, New York, second edition. MR4225301
[34] Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.
[35] Sekhon, J. S. and Shem-Tov, Y. (2021). Inference on a new class of sample average treatment effects. *Journal of the American Statistical Association*, 116(534):798–804. MR4270025
[36] Taubman, S. L., Allen, H. L., Wright, B. J., Baicker, K., and Finkelstein, A. N. (2014). Medicaid increases emergency-department use: Evidence from Oregon's Health Insurance Experiment. *Science*, 343(6168):263–268.
[37] Volfovsky, A., Airoldi, E. M., and Rubin, D. B. (2015). Causal inference for ordinal outcomes. ArXiv preprint arXiv:1501.01234.

Mark M. Fredrickson
Department of Statistics
University of Michigan
323 West Hall
1085 S. University Ave.
Ann Arbor, MI 48109
USA
E-mail address: mfredric@umich.edu

Yuguo Chen
Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL 61820
USA
E-mail address: yuguo@illinois.edu