Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2019, Vol. 28(5) 1378–1398

DOI: 10.1177/0962280218756689
journals.sagepub.com/home/smm
SAGE

Randomization-based inference for Bernoulli trial experiments and implications for observational studies

Zach Branson and Marie-Abèle Bind

Abstract

We present a randomization-based inferential framework for experiments characterized by a strongly ignorable assignment mechanism where units have independent probabilities of receiving treatment. Previous works on randomization tests often assume these probabilities are equal within blocks of units. We consider the general case where they differ across units and show how to perform randomization tests and obtain point estimates and confidence intervals. Furthermore, we develop rejection-sampling and importance-sampling approaches for conducting randomization-based inference conditional on any statistic of interest, such as the number of treated units or forms of covariate balance. We establish that our randomization tests are valid tests, and through simulation we demonstrate how the rejection-sampling and importance-sampling approaches can yield powerful randomization tests and thus precise inference. Our work also has implications for observational studies, which commonly assume a strongly ignorable assignment mechanism. Most methodologies for observational studies make additional modeling or asymptotic assumptions, while our framework only assumes the strongly ignorable assignment mechanism, and thus can be considered a minimal-assumption approach.

Keywords

Conditional inference, importance sampling, propensity scores, randomization tests, rejection sampling, strongly ignorable assignment

1 Introduction

Randomization-based inference centers around the idea that the treatment assignment mechanism is the only stochastic element in a randomized experiment and thus acts as the basis for conducting statistical inference.[1] In general, a central tenet of randomization-based inference is that the analysis of any given experiment should reflect its design: the inference for completely randomized experiments, blocked randomized experiments, and other designs should reflect the actual assignment mechanism that was used during the experiment. The idea that the assignment mechanism is the only stochastic element of an experiment is also commonly employed in the potential outcomes framework,[2] which is now regularly used when estimating causal effects in randomized experiments and observational studies.[3,4] While randomization-based inference focuses on estimating causal effects for only the finite sample at hand, it can flexibly incorporate any kind of assignment mechanism without model specifications. Rosenbaum[5] provides a comprehensive review of randomization-based inference.

An essential step to estimating causal effects within the randomization-based inference framework as well as the potential outcomes framework is to state the probability distribution of the assignment mechanism. For simplicity, we focus on treatment versus control experiments, but our discussion can be extended to experiments with multiple treatments. Let the vector $\mathbf{W}$ denote the treatment assignment for N units in an experiment or observational study.

Faculty of Arts and Sciences, Science Center, Harvard University, Cambridge, MA, USA

Corresponding author:
Zach Branson, Faculty of Arts and Sciences, Science Center, Harvard University, One Oxford Street, 7th Floor, Cambridge 02138, MA, USA.
Email: zbranson@g.harvard.edu

It is commonly assumed that the probability distribution of $\mathbf{W}$ can be written as a product of independent Bernoulli trials that may depend on background covariates[6–8]

$$P(\mathbf{W} = \mathbf{w}|\mathbf{X}) = \prod_{i=1}^{N} e(\mathbf{x}_i)^{w_i}[1 - e(\mathbf{x}_i)]^{1-w_i}, \quad \text{where } 0 < e(\mathbf{x}_i) < 1 \quad \forall i = 1, \ldots, N \tag{1}$$

Here, $\mathbf{X}$ is an $N \times p$ covariate matrix with rows $\mathbf{x}_i$, and $e(\mathbf{x}_i)$ denotes the probability that the ith unit receives treatment conditional on pre-treatment covariates $\mathbf{x}_i$; i.e. $e(\mathbf{x}_i) \equiv P(W_i = 1|\mathbf{x}_i)$. The probabilities $e(\mathbf{x}_i)$ are commonly known as propensity scores.[9] An assignment mechanism that can be written as equation (1) is known as an unconfounded, strongly ignorable assignment mechanism.[8] The assumption of an unconfounded, strongly ignorable assignment mechanism is essential to propensity score analyses and other methodologies (e.g. regression-based methods) for analyzing observational studies.[10–13]

In randomized experiments, the propensity scores are defined by the designer(s) of the experiment and are thus known; this knowledge is all that is needed to construct unbiased estimates for average treatment effects.[8] The propensity score $e(\mathbf{x}_i)$ is not necessarily a function of all or any of the covariates: for example, in completely randomized experiments, $e(\mathbf{x}_i) = 0.5$ for all units; and for blocked-randomized and paired experiments, the propensity scores are equal for all units within the same block or pair.

In observational studies, the propensity scores are not known, and instead must be estimated. The $e(\mathbf{x}_i)$ in equation (1) are often estimated using logistic regression, but any model that estimates conditional probabilities for a binary treatment can be used. These estimates, $\hat{e}(\mathbf{x}_i)$, are commonly employed to ''reconstruct'' a hypothetical experiment that yielded the observed data.[8] For example, matching methodologies are used to obtain subsets of treatment and control that are balanced in terms of pre-treatment covariates; then, these subsets of treatment and control are analyzed as if they came from a completely randomized experiment.[8,12,14] Others have suggested regression-based adjustments combined with the propensity score[15,16] as well as Bayesian modeling.[4,17,18] Notably, all of these methodologies implicitly assume the Bernoulli trial assignment mechanism shown in equation (1), but the subsequent analyses reflect a completely randomized, blocked-randomized, or paired assignment mechanism instead. One methodology commonly employed in observational studies that more closely reflects a Bernoulli trial assignment mechanism is inverse propensity score weighting[19–22]; however, the variance of such estimators is unstable, especially when estimated propensity scores are particularly close to 0 or 1, which is an ongoing concern in the literature.[23,24] Furthermore, the validity of such point estimates and uncertainty intervals rely on asymptotic arguments and an infinite-population interpretation.

More importantly, all of the above methodologies—matching, frequentist or Bayesian modeling, inverse propensity score weighting, or any combination of them—assume the strongly ignorable assignment mechanism shown in equation (1), but they also intrinsically make additional modeling or asymptotic assumptions. On the other hand, although randomization-based inference methodologies also make the common assumption of the strongly ignorable assignment mechanism, they do not require any additional model specifications or asymptotic arguments.

However, while there is a wide literature on randomization tests, most have focused on assignment mechanisms where the propensity scores are assumed to be the same across units (i.e. completely randomized experiments) or groups of units (i.e. blocked or paired experiments), instead of the more general case where they may differ across all units, as in equation (1). Imbens and Rubin[25] briefly mention Bernoulli trial experiments, but only discuss inference for purely randomized and block randomized designs. Another example is Basu,[26] who thoroughly discusses Fisherian randomization tests and briefly considers Bernoulli trial experiments, but does not provide a randomization-test framework for such experiments. This trend continues for observational studies: most randomization tests for observational studies utilize permutations of the treatment indicator within covariate strata, and thus reflect a block-randomized assignment mechanism instead of the assumed Bernoulli trial assignment mechanism.[6,27,28] While these tests are valid under certain assumptions, they are not immediately applicable to cases where covariates are not easily stratified (e.g. continuous covariates) or where there is not at least one treated unit and one control unit in each stratum.[5] None of these randomization tests are applicable to cases where the propensity scores (known or unknown) differ across all units.

Most randomization tests that incorporate varying propensity scores focus on the biased-coin design popularized by Efron,[29] where propensity scores are dependent on the order units enter the experiment and possibly pre-treatment covariates as well. Wei[30] and Soares and Wu[31] developed extensions for this experimental design, while Smythe and Wei,[32] Wei,[33] and Mehta et al.[34] developed significance tests for such designs. Good[35] (Section 4.5) provides further discussion on this literature. The biased-coin design is

related to covariate-adaptive randomization schemes in the clinical trial literature, starting with the work of Pocock and Simon.[36] Covariate-adaptive randomization schemes sequentially randomize units such that the treatment and control groups are balanced in terms of pre-treatment covariates,[37–39] and recent works in the statistics literature have explored valid randomization tests for covariate-adaptive randomization schemes.[39,40] Importantly, the randomization test literature for biased-coin and covariate-adaptive designs differs from the randomization test presented here: all of these works focus on sequential designs, and thus depend on the sequential dependence among units inherent in the randomization scheme. In contrast, we assume that all units are simultaneously assigned to treatment according to the strongly ignorable assignment mechanism (equation (1)).

To the best of our knowledge, there is not an explicit randomization-based inference framework for analyzing Bernoulli trial experiments, let alone observational studies. Here we develop such a framework for randomized experiments characterized by Bernoulli trials, with the implication that this framework can be extended to the observational study literature as well. In particular, we develop rejection-sampling and importance-sampling approaches for conducting conditional randomization-based inference for Bernoulli trial experiments, which has not been previously discussed in the literature. These approaches allow one to conduct randomization tests conditional on statistics of interest for more precise inference.

In Section 2, we review randomization-based inference in general, including randomization tests and how these tests can be inverted to yield point estimates and confidence intervals. In Section 3, we develop a randomization-based inference framework for Bernoulli trial experiments, first reviewing the case where propensity scores are equal across units, and then extending this framework to the general case where propensity scores differ across units. Furthermore, we establish that randomization tests under this framework are valid tests, both unconditionally and conditionally on statistics of interest. In Section 4, we demonstrate our framework with a simple example and provide simulation evidence for how our rejection-sampling and importance-sampling approaches can yield statistically powerful conditional randomization tests. In Section 5, we discuss extensions and implications of this work, particularly for observational studies.

2 Review of randomization-based inference

Randomization-based inference focuses on randomization tests for treatment effects, which can be inverted to obtain both point estimates and confidence intervals. Randomization tests were first proposed by Fisher,[1] and foundational theory for these tests was later developed by Pitman[41] and Kempthorne.[42] We follow the notation of Imbens and Rubin[25] in our discussion of randomization tests for treatment versus control experiments.

2.1 Notation

Randomization tests utilize the potential outcomes framework, where the only stochastic element of an experiment is the treatment assignment. Let

$$W_i = \begin{cases} 1 & \text{if the } i\text{th unit receives treatment} \\ 0 & \text{if the } i\text{th unit receives control} \end{cases} \tag{2}$$

denote the treatment assignment, and let $Y_i(W_i)$ denote the ith unit's potential outcome, which only depends on the treatment assignment W_i. Only $Y_i(1)$ or $Y_i(0)$ is ultimately observed at the end of an experiment—never both. Let

$$y_i^{obs} = Y_i(1)W_i + Y_i(0)(1 - W_i) \tag{3}$$

denote the observed outcomes. Finally, let $\mathbb{W} \equiv \{0, 1\}^N$ denote the set of all possible treatment assignments, and let $\mathbb{W}^+ \subset \mathbb{W}$ denote the subset of $\mathbb{W}$ with positive probability, i.e. $\mathbb{W}^+ = \{\mathbf{w} \in \mathbb{W} : P(\mathbf{W} = \mathbf{w}) > 0\}$.

Importantly, the probability distribution of treatment assignments, $P(\mathbf{W})$, fully characterizes the assignment mechanism: because treatment assignment is the only stochastic element in a randomized experiment, the distribution $P(\mathbf{W})$ specifies the randomness in a randomized experiment. Consequently, inference within the randomization-based framework is determined by $P(\mathbf{W})$.

We first review how $P(\mathbf{W})$ is used to perform randomization tests. We then discuss how to invert these tests to obtain point estimates and confidence intervals for the average treatment effect.

2.2 Testing the sharp null hypothesis via randomization tests

The most common use of randomization tests is to test the Sharp Null Hypothesis, which is

$$H_0 : Y_i(1) = Y_i(0) \quad \forall i = 1, \ldots, n \tag{4}$$

i.e. the hypothesis that there is no treatment effect. Under the Sharp Null Hypothesis, the outcomes for *any* randomization from the set of all possible randomizations $\mathbb{W}^+$ are known: regardless of a unit's treatment assignment, its outcome will always be equal to the observed response y_i^{obs} under the Sharp Null Hypothesis. This knowledge allows one to test the Sharp Null Hypothesis.

To test this hypothesis, one first chooses a suitable test statistic

$$t(Y(\mathbf{W}), \mathbf{W}) \tag{5}$$

and determines whether the observed test statistic $t^{obs} \equiv t(\mathbf{y}^{obs}, \mathbf{W}^{obs})$ is unlikely to occur according to the randomization distribution of the test statistic (5) under the Sharp Null Hypothesis. For example, one common choice of test statistic is the difference in mean response between treatment and control units, defined as

$$t(Y(\mathbf{W}), \mathbf{W}) = \frac{\sum_{i:W_i=1} Y_i(1)}{\sum_{i=1}^N W_i} - \frac{\sum_{i:W_i=0} Y_i(0)}{\sum_{i=1}^N (1 - W_i)} \tag{6}$$

Such a test statistic will be powerful in detecting a difference in means between the distributions of $Y_i(1)$ and $Y_i(0)$. In general, one should choose a test statistic according to possible differences in the distributions of $Y_i(1)$ and $Y_i(0)$ that one is most interested in. Please see Rosenbaum[5] (Chapter 2) for a discussion on the choice of test statistics for randomization tests.

After a test statistic is chosen, a randomization-test p-value can be computed by comparing the observed test statistic t^{obs} to the set of $t(Y(\mathbf{W}), \mathbf{W})$ that are possible given the set of possible treatment assignments $\mathbb{W}^+$, assuming the Sharp Null Hypothesis is true. The two-sided randomization-test p-value is

$$P(|t(Y(\mathbf{W}), \mathbf{W})| \geq |t^{obs}|) = \sum_{\mathbf{w} \in \mathbb{W}^+} \mathbb{I}(|t(Y(\mathbf{w}), \mathbf{w})| \geq |t^{obs}|) P(\mathbf{W} = \mathbf{w}) \tag{7}$$

where $\mathbb{I}(A) = 1$ if event A occurs and zero otherwise. Importantly, the randomization-test p-value (equation (7)) depends on the set of possible treatment assignments $\mathbb{W}^+$, the probability distribution $P(\mathbf{W})$, and the choice of test statistic $t(Y(\mathbf{W}), \mathbf{W})$.

Thus, testing the Sharp Null Hypothesis is a three-step procedure:

(1) Specify the distribution $P(\mathbf{W})$ (and, consequently, the set of possible treatment assignments $\mathbb{W}^+$).
(2) Choose a test statistic $t(Y(\mathbf{W}), \mathbf{W})$.
(3) Compute or approximate the p-value (equation (7)).

All randomization tests discussed in this paper follow this three-step procedure, with the only difference among them being the choice of $P(\mathbf{W})$, i.e. the first step. The third step notes that exactly computing the randomization-test p-value is often computationally intensive because it requires enumerating all possible $\mathbf{W} \in \mathbb{W}^+$; instead, it can be approximated. A typical approximation is to generate a random sample $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}$ from $P(\mathbf{W})$, and then approximate the p-value (equation (7)) by

$$P(|t(Y(\mathbf{W}), \mathbf{W})| \geq |t^{obs}|) \approx \frac{\sum_{m=1}^M \mathbb{I}(|t(Y(\mathbf{w}^{(m)}), \mathbf{w}^{(m)})| \geq |t^{obs}|)}{M} \tag{8}$$

Importantly, the approximation (8) still depends on the probability distribution of the assignment mechanism, $P(\mathbf{W})$, because the random samples $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}$ are generated using $P(\mathbf{W})$. This distinction will be important in

our discussion of Bernoulli trial experiments, where the probability of receiving treatment—i.e. the propensity scores—may be equal or non-equal across units. In both cases, the set $\mathbb{W}^+$ is the same, but the probability distribution $P(\mathbf{W})$ is different.

Testing the Sharp Null Hypothesis will provide information about the presence of any treatment effect amongst all units in the study. Furthermore, this test can be inverted to obtain point estimates and confidence intervals for the treatment effect.

2.3 Randomization-based point estimates and confidence intervals for the treatment effect

A confidence interval can be constructed by inverting a variation of the Sharp Null Hypothesis that assumes an additive treatment effect. A randomization-based confidence interval for the average treatment effect is the set of $\tau \in \mathbb{R}$ such that one fails to reject the hypothesis

$$H_0^\tau : Y_i(1) = Y_i(0) + \tau \quad \forall i = 1, \ldots, N \tag{9}$$

The above hypothesis is a sharp hypothesis in the sense that, under H_0^τ, every unit's outcome for any treatment assignment is known: under H_0^τ, the missing potential outcome of any treated unit would be $y_i^{obs} - \tau$; likewise, the missing potential outcome of any control unit would be $y_i^{obs} + \tau$. Thus, for any hypothetical treatment assignment $\mathbf{w} \in \mathbb{W}^+$, one can calculate the corresponding potential outcomes $Y(\mathbf{w})$ under H_0^τ in terms of the observed outcomes $\mathbf{y}^{obs}$ and observed treatment assignment $\mathbf{w}^{obs}$

$$Y_i(w_i) = y_i^{obs} + \tau(w_i - w_i^{obs}), \quad \forall i = 1, \ldots, N \tag{10}$$

Therefore, one can obtain a p-value for the hypothesis H_0^τ by drawing many hypothetical randomizations $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}$ from $P(\mathbf{W})$, computing each $Y(\mathbf{w}^{(m)})$ using equation (10), and then using equation (8) to approximate the p-value for any given test statistic $t(Y(\mathbf{W}), \mathbf{W})$.

To construct a 95% confidence interval, one considers many τ (e.g. via a line search), tests the hypothesis H_0^τ for each τ, and defines the confidence interval as the set of τ with corresponding p-values above 0.05.[5,25] Importantly, the confidence interval will depend on the probability distribution $P(\mathbf{W})$ through the draws $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}$ to compute each p-value; thus, the confidence interval will reflect a prespecified assignment mechanism. As we discuss in Section 3.3, this also allows one to flexibly construct confidence intervals that condition on particular statistics of interest.

Testing the hypothesis H_0^τ also yields a natural point estimate: define the point estimate $\hat{\tau}$ as the τ such that the p-value for testing the hypothesis H_0^τ is maximized. For example, given a 95% confidence interval containing τ with corresponding p-values above 0.05, $\hat{\tau}$ is defined as the τ with the highest p-value. The interpretation of such a $\hat{\tau}$ is that this is the "most probable" τ under the assumption of an additive treatment effect. This point estimate is a variant of the Hodges–Lehmann randomization-based point estimate, which equates the test statistic under the hypothesis H_0^τ to its expectation under the randomization distribution.[5,43]

Some have criticized randomization-based confidence intervals constructed by inverting hypotheses such as equation (9) because it assumes a homogeneous treatment effect, which may be an inappropriate assumption. However, in general, confidence intervals can be constructed using any sharp null hypothesis that fully specifies unit-level treatment effects, including sharp null hypotheses that specify heterogeneous treatment effects.[44] Thus, while we focus on homogeneous treatment effects as assumed in (equation (9)), the randomization test framework that we present below can be extended to point estimates and confidence intervals that account for treatment effect heterogeneity to the extent that one can specify sharp null hypotheses that incorporate heterogeneous treatment effects.

3 Randomization-based inference for Bernoulli trial experiments

Here we consider experimental designs that are characterized by Bernoulli trials and develop randomization tests for these designs. First, we review randomization tests for experimental designs where the probability of receiving treatment is the same for all units; this will motivate our development of randomization tests for experimental designs where the probability of receiving treatment differs across units, which is our main contribution. For both cases—first when the propensity scores are equal across units, and then when the propensity scores differ—we will discuss several assignment mechanisms $P(\mathbf{W})$ and sets of possible treatment assignments $\mathbb{W}^+$, which correspond to

different randomization tests. Once $P(\mathbf{W})$ and $\mathbb{W}^+$ are specified, the Sharp Null Hypothesis can be tested by following the three-step procedure in Section 2.2; furthermore, these tests can be inverted to yield point estimates and confidence intervals, as discussed in Section 2.3. For each test, we will state an explicit form for $P(\mathbf{W} = \mathbf{w})$ for any $\mathbf{w} \in \mathbb{W}^+$ to compute the randomization test p-value (equation (7)) exactly, and we will also state how random samples $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}$ can be generated to approximate this p-value using equation (8). In Section 3.3, we introduce rejection-sampling and importance-sampling approaches to perform randomization tests conditional on various statistics of interest, which has not been previously considered for randomization-based inference for Bernoulli trial experiments.

3.1 Case 1: Propensity scores are equal across units

Let $e(\mathbf{x}_i) = P(\mathbf{W}_i = 1|\mathbf{x}_i)$ denote the propensity score, i.e. the probability that the ith unit receives treatment, given a vector of pre-treatment covariates $\mathbf{x}_i$. In this section we assume without loss of generality that $e(\mathbf{x}_i) = 0.5$ for all $i = 1, \ldots, N$; i.e., $P(W_i = 1|\mathbf{x}_i) = P(W_i = 1) = 0.5$ for all units. We consider several sets of possible treatment assignments $\mathbb{W}^+$ and note the corresponding $P(\mathbf{W} = \mathbf{w})$ for each $\mathbf{w} \in \mathbb{W}^+$, which can be used to compute the p-value (7) for testing the Sharp Null Hypothesis.

First consider the set $\mathbb{W}^+ = \mathbb{W} = \{0, 1\}^N$, i.e. experiments that are characterized by independent, unbiased coin flips, where any number of units can receive treatment or control. In this case, $P(\mathbf{W} = \mathbf{w}) = \frac{1}{2^N}$ for all $\mathbf{w} \in \mathbb{W}^+$. To generate random draws $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}$, one simply flips N unbiased coins to generate an N-dimensional vector of 0 s and 1 s.

However, Imbens and Rubin[25] note that when $\mathbb{W}^+ = \{0, 1\}^N$, there is a non-zero probability of $\mathbf{W} = \mathbf{0}_N \equiv (0, \ldots, 0)$ or $\mathbf{W} = \mathbf{1}_N \equiv (1, \ldots, 1)$. In these cases, most test statistics are undefined, and so they do not consider this case further. This concern can be addressed by either defining test statistics for these cases (a common choice being zero) or instead considering the set $\mathbb{W}^+ = \{0, 1\}^N \setminus (\mathbf{0}_N \cup \mathbf{1}_N)$ of possible treatment assignments. In this case, $P(\mathbf{W} = \mathbf{w}) = \frac{1}{2^N - 2}$ for all $\mathbf{w} \in \mathbb{W}^+$. To generate random draws $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}$, one simply flips N unbiased coins and only accepts a random draw $\mathbf{w}^{(m)}$ if it is not $\mathbf{0}_N$ or $\mathbf{1}_N$. This follows the argument of Imbens and Rubin[25] that preventing "unhelpful treatment allocations" will yield more precise inferences for treatment effects.

Indeed, we can even further restrict $\mathbb{W}^+$. It is common to condition on statistics such as the number of units that receive treatment $N_T \equiv \sum_{i=1}^N W_i$. When $\mathbb{W}^+ = \{\mathbf{W} \in \mathbb{W} | \sum_{i=1}^N W_i = N_T\}$ for some prespecified N_T, $P(\mathbf{W} = \mathbf{w}) = 1 \Big/ \binom{N}{N_T}$ for all $\mathbf{w} \in \mathbb{W}^+$. To generate random draws $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}$, one simply flips N unbiased coins and only accepts a random draw $\mathbf{w}^{(m)}$ if $\sum_{i=1}^N w_i^{(m)} = N_T$; equivalently, one can obtain such random draws by randomly permuting the observed treatment assignment $\mathbf{W}^{obs}$. A randomization test that uses such a $\mathbb{W}^+$ and $P(\mathbf{W})$ is the most common randomization test in the literature and corresponds to what is typically referred to as a "completely randomized" experimental design.[25] Because of the equivalence to random permutations of $\mathbf{W}^{obs}$, this randomization test is also often called a permutation test.

3.2 Case 2: Propensity scores differ across units

Now consider the case where $e(\mathbf{x}_i) \neq e(\mathbf{x}_j)$ for some $i \neq j$, i.e. where the propensity scores differ across units. This may be due to differences in the covariate vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ or some other experimental design prespecification. Again we consider several sets of possible treatment assignments $\mathbb{W}^+$, note the corresponding $P(\mathbf{W} = \mathbf{w}|\mathbf{X})$ for each $\mathbf{w} \in \mathbb{W}^+$, and state how to generate random draws $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}$, which can be used to compute or approximate the p-value for testing the Sharp Null Hypothesis.

First consider the set $\mathbb{W}^+ = \mathbb{W} = \{0, 1\}^N$. In this case

$$P(\mathbf{W} = \mathbf{w}|\mathbf{X}) = \prod_{i=1}^N e(\mathbf{x}_i)^{w_i}[1 - e(\mathbf{x}_i)]^{1-w_i} \tag{11}$$

which is identical to the assignment mechanism (1) typically assumed in observational studies. To generate random draws $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}$, one simply flips N *biased* coins with probabilities corresponding to the $e(\mathbf{x}_i)$ to generate an N-dimensional vector of 0 s and 1 s.

However, there is still a chance—though small—that a random draw $\mathbf{w}$ from $\mathbb{W}^+ = \{0,1\}^N$ will be equal to $\mathbf{0}_N$ or $\mathbf{1}_N$, and in this case test statistics will be undefined. Now consider the restricted set $\mathbb{W}^+ = \{0,1\}^N \setminus (\mathbf{0}_N \cup \mathbf{1}_N)$. In this case

$$P(\mathbf{W} = \mathbf{w}|\mathbf{X}) = \frac{\prod_{i=1}^N e(\mathbf{x}_i)^{w_i}[1-e(\mathbf{x}_i)]^{1-w_i}}{1-\prod_{i=1}^N e(\mathbf{x}_i)-\prod_{i=1}^N [1-e(\mathbf{x}_i)]} \tag{12}$$

To arrive at this result, note that when $\mathbb{W}^+ = \{0,1\}^N \setminus (\mathbf{0}_N \cup \mathbf{1}_N)$

$$\sum_{\mathbf{w}\in\mathbb{W}^+} \prod_{i=1}^N e(\mathbf{x}_i)^{w_i}[1-e(\mathbf{x}_i)]^{1-w_i} = 1 - \prod_{i=1}^N e(\mathbf{x}_i) - \prod_{i=1}^N [1-e(\mathbf{x}_i)] \tag{13}$$

Thus, the probabilities equation (12) sum to one. To generate random draws $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}$, one simply flips N *biased* coins and only accepts a random draw $\mathbf{w}^{(m)}$ if it is not $\mathbf{0}_N$ or $\mathbf{1}_N$.

Again, we can further restrict $\mathbb{W}^+$ to incorporate certain statistics of interest, such as the number of units assigned to treatment. Consider the set $\mathbb{W}^+ = \{\mathbf{W} \in \mathbb{W} | \sum_{i=1}^N W_i = N_T\}$ for some prespecified N_T. In this case

$$P(\mathbf{W} = \mathbf{w}|\mathbf{X}) = \frac{\prod_{i=1}^N e(\mathbf{x}_i)^{w_i}[1-e(\mathbf{x}_i)]^{1-w_i}}{P\left(\sum_{i=1}^N W_i = N_T|\mathbf{X}\right)} \tag{14}$$

The denominator, $P(\sum_{i=1}^N W_i = N_T|\mathbf{X}) = \sum_{\mathbf{w}\in\mathbb{W}^+} \prod_{i=1}^N e(\mathbf{x}_i)^{w_i}[1-e(\mathbf{x}_i)]^{1-w_i}$, is seemingly difficult to compute, due to the large number, $\binom{N}{N_T}$, of possible treatment assignments $\mathbf{w} \in \mathbb{W}^+$. Chen and Liu[45] provide an algorithm to compute $P(\sum_{i=1}^N W_i = N_T|\mathbf{X})$ exactly. Alternatively, $P(\sum_{i=1}^N W_i = N_T|\mathbf{X})$ can be estimated, and there are many ways to estimate this quantity. One option is to randomly sample $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}$ from $\mathbb{W}^+$ and use the unbiased estimator

$$\hat{P}\left(\sum_{i=1}^N W_i = N_T|\mathbf{X}\right) = \frac{\binom{N}{N_T}}{M} \sum_{m=1}^M \prod_{i=1}^N e(\mathbf{x}_i)^{w_i^{(m)}}[1-e(\mathbf{x}_i)]^{1-w_i^{(m)}} \tag{15}$$

which is the typical estimator for a population total seen in the survey sampling literature (e.g. Lohr,[46] p.55).

However, computing $P\left(\sum_{i=1}^N W_i = N_T|\mathbf{X}\right)$ is only required when one wants to compute the randomization-test p-value exactly using equation (7). Instead, one can still approximate this p-value using equation (8) by generating random draws $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}$, which is done by flipping N *biased* coins and only accepting a random draw $\mathbf{w}$ if $\sum_{i=1}^N w_i = N_T$.

This introduces straightforward rejection-sampling and importance-sampling procedures for conducting conditional randomization-based inference for Bernoulli trial experiments.

3.3 Rejection-sampling and importance-sampling procedures for conditional randomization tests

As discussed in Section 2, researchers do not typically compute the randomization test p-value equation (7) exactly, but instead generate random draws $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)}$ from the probability distribution $P(\mathbf{W})$ and then approximate the randomization test p-value using equation (8). To conduct conditional randomization-based inference, one generates random draws from conditional probability distributions such as $P(\mathbf{W}|\sum_{i=1}^N W_i = N_T)$ instead of $P(\mathbf{W})$. This is straightforward when the propensity scores are the same across units: for example, as discussed in Section 3.1, samples from $P(\mathbf{W}|\sum_{i=1}^N W_i = N_T)$ correspond to random permutations of the observed treatment assignment $\mathbf{W}^{obs}$ when the propensity scores are equal across units. However, sampling from such conditional distributions when the propensity scores differ across units is less trivial. To the best of our knowledge, a strategy for how to sample from such distributions has not been described in the literature.

Conducting conditional randomization-based inference involves focusing only on "acceptable" treatment assignments $\mathbf{W}$; e.g. $\mathbf{W}$ that are not $\mathbf{0}_N$ or $\mathbf{1}_N$, or $\mathbf{W}$ such that $\sum_{i=1}^N W_i = N_T$ for some prespecified N_T.

To formalize this idea, define an acceptance criterion that is a function of the treatment assignment and pre-treatment covariates

$$\phi(\mathbf{W}, \mathbf{X}) = \begin{cases} 1 & \text{if} \quad \mathbf{W} \text{ is an acceptable treatment assignment} \\ 0 & \text{if} \quad \mathbf{W} \text{ is not an acceptable treatment assignment} \end{cases} \tag{16}$$

The criterion $\phi(\mathbf{W}, \mathbf{X})$ can encapsulate any statistic of interest, such as the number of treated units or forms of covariate balance. The criterion $\phi(\mathbf{W}, \mathbf{X})$ should be defined by statistics that are believed to be related to the outcome, such as the number of treated units with a certain covariate value or the covariate means in the treatment and control groups. See Hennessy et al.[47] for further discussion about the types of statistics that should be conditioned on for conditional randomization-based inference.

Once $\phi(\mathbf{W}, \mathbf{X})$ is defined, one conducts conditional randomization-based inference by performing a randomization test only within the set of randomizations such that the acceptance criterion is satisfied. For example, Sections 3.1 and 3.2 discuss conducting randomization-based inference for the case when $\phi(\mathbf{W}, \mathbf{X}) = 1$ if $\sum_{i=1}^{N} W_i = N_T$ and 0 otherwise. Thus, the true conditional randomization test p-value is

$$p_\phi \equiv \sum_{\mathbf{w} \in \mathbb{W}_\phi^+} \mathbb{I}(|t(Y(\mathbf{w}), \mathbf{w})| \geq |t^{obs}|) P(\mathbf{W} = \mathbf{w}) \tag{17}$$

where $\mathbb{W}_\phi^+ = \{\mathbf{w} : \phi(\mathbf{w}, \mathbf{X}) = 1\}$ is the set of acceptable randomizations. The p-value p_ϕ is nearly identical to the p-value (equation (7)), but using only the set of acceptable randomizations instead of the set of all randomizations. The set of acceptable randomizations is typically large, and thus the p-value p_ϕ cannot always be computed exactly. Instead, it can be unbiasedly estimated using

$$\hat{p}_{RS} = \frac{\sum_{m=1}^{M} \mathbb{I}(|t(Y(\mathbf{w}^{(m)}), \mathbf{w}^{(m)})| \geq |t^{obs}|)}{M}, \quad \text{where } \mathbf{w}^{(m)} \sim P(\mathbf{W} | \phi(\mathbf{W}, \mathbf{X}) = 1) \tag{18}$$

i.e. the approximation presented in equation (8). We propose a rejection-samping procedure for generating random samples $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)} \sim P(\mathbf{W} | \phi(\mathbf{W}, \mathbf{X}) = 1)$: randomly generate draws from $P(\mathbf{W})$, and only accept a draw $\mathbf{w}$ if $\phi(\mathbf{w}, \mathbf{X}) = 1$. For Bernoulli trials, this involves flipping N coins (biased or unbiased, depending on the experimental design), and only accepting a particular assignment $\mathbf{w}$ if $\phi(\mathbf{w}, \mathbf{X}) = 1$.

While the rejection-sampling estimator $\hat{p}_{RS}$ is unbiased for p_ϕ, it may be computationally intensive to generate random samples $\mathbf{w}^{(m)} \sim P(\mathbf{W} | \phi(\mathbf{W}, \mathbf{X}) = 1)$ if $\phi(\mathbf{W}, \mathbf{X})$ is particularly stringent. As an alternative, one can take an importance-sampling approach to biasedly estimate p_ϕ at a much lower computational cost.[48–51] First, define a proposal distribution $P_q(\mathbf{W})$ whose support includes the support of $P(\mathbf{W} | \phi(\mathbf{W}, \mathbf{X}) = 1)$ but is less computationally burdensome to sample from than from $P(\mathbf{W} | \phi(\mathbf{W}, \mathbf{X}) = 1)$. Then, the importance-sampling estimator for p_ϕ is

$$\hat{p}_{IS} = \frac{\sum_{m=1}^{M} \mathbb{I}(|t(Y(\mathbf{w}^{(m)}), \mathbf{w}^{(m)})| \geq |t^{obs}|) \frac{P(\mathbf{W} = \mathbf{w}^{(m)} | \phi(\mathbf{W}, \mathbf{X}) = 1)}{P_q(\mathbf{W} = \mathbf{w}^{(m)})}}{\sum_{m=1}^{M} \frac{P(\mathbf{W} = \mathbf{w}^{(m)} | \phi(\mathbf{W}, \mathbf{X}) = 1)}{P_q(\mathbf{W} = \mathbf{w}^{(m)})}}, \quad \text{where } \mathbf{w}^{(m)} \sim P_q(\mathbf{W}) \tag{19}$$

In other words, the rejection-sampling estimator $\hat{p}_{RS}$ is a simple average based on the random draws $\mathbf{w}^{(m)} \sim P(\mathbf{W} | \phi(\mathbf{W}, \mathbf{X}) = 1)$, whereas the importance-sampling estimator is a weighted average based on the random draws $\mathbf{w}^{(m)} \sim P_q(\mathbf{W})$. Thus, $\hat{p}_{IS}$ will be easier to compute than $\hat{p}_{RS}$ if it is less computationally intensive to sample from the proposal distribution $P_q(\mathbf{W})$ than from the target distribution $P(\mathbf{W} | \phi(\mathbf{W}, \mathbf{X}) = 1)$.

The importance-sampling estimator can be reduced to a simple form by first noting that, under the assumption of a strongly ignorable assignment mechanism (1),

$$P(\mathbf{W} = \mathbf{w} | \phi(\mathbf{W}, \mathbf{X}) = 1) = \frac{P(\mathbf{W} = \mathbf{w}, \phi(\mathbf{W}, \mathbf{X}) = 1)}{P(\phi(\mathbf{W}, \mathbf{X}) = 1)} \tag{20}$$

$$= \frac{\prod_{i=1}^{N} e(\mathbf{x}_i)^{w_i} [1 - e(\mathbf{x}_i)]^{1 - w_i}}{P(\phi(\mathbf{W}, \mathbf{X}) = 1)}, \quad \text{where } \mathbf{w} \in \mathbb{W}_\phi^+ \tag{21}$$

$$\propto \prod_{i=1}^{N} e(\mathbf{x}_i)^{w_i}[1-e(\mathbf{x}_i)]^{1-w_i}, \quad \text{where } \mathbf{w} \in \mathbb{W}_{\phi}^{+} \tag{22}$$

where $\mathbb{W}_{\phi}^{+} \equiv \{\mathbf{w} \in \mathbb{W}^{+} : \phi(\mathbf{w}, \mathbf{X}) = 1\}$ is the set of acceptable assignments according to the acceptance criterion. Then, if the proposal distribution is uniform across all acceptable assignments, i.e. if $P_q(\mathbf{W} = \mathbf{w}) = c$ for all $\mathbf{w} \in \mathbb{W}_{\phi}^{+}$, then the importance-sampling p-value approximation reduces to

$$\hat{p}_{IS} = \frac{\sum_{m=1}^{M} \mathbb{I}(|t(Y(\mathbf{w}^{(m)}), \mathbf{w}^{(m)})| \geq |t^{obs}|) \prod_{i=1}^{N} e(\mathbf{x}_i)^{w_i^{(m)}}[1-e(\mathbf{x}_i)]^{1-w_i^{(m)}}}{\sum_{m=1}^{M} \prod_{i=1}^{N} e(\mathbf{x}_i)^{w_i^{(m)}}[1-e(\mathbf{x}_i)]^{1-w_i^{(m)}}}, \quad \text{where } \mathbf{w}^{(m)} \sim P_q(\mathbf{W}) \tag{23}$$

where the quantity $\prod_{i=1}^{N} e(\mathbf{x}_i)^{w_i^{(m)}}[1-e(\mathbf{x}_i)]^{1-w_i^{(m)}}$ is easy to compute because the propensity scores $e(\mathbf{x}_i)$ are known.

For example, sampling from the distribution $P(\mathbf{W}|\sum_{i=1}^{N} W_i = N_T)$ via rejection-sampling may be computationally intensive if the propensity scores differ across units and N is large. One proposal distribution that is uniform across assignments is random permutations of $\mathbf{W}^{obs}$, whose support is equal to the support of $P(\mathbf{W}|\sum_{i=1}^{N} W_i = N_T)$ but is less computational to sample from. Thus, one can still utilize random permutations of $\mathbf{W}^{obs}$ to estimate the conditional randomization test p-value—as in Case 1 in Section 3.1—using the importance-sampling estimator $\hat{p}_{IS}$.

However, as noted earlier, unlike the estimator $\hat{p}_{RS}$, the estimator $\hat{p}_{IS}$ is biased of order M^{-1},[48] which—as we show in Section 4—may break the validity of the conditional randomization test. Thus, we recommend using the rejection-sampling estimator $\hat{p}_{RS}$ to ensure valid inferences from our conditional randomization test if it is not computationally intensive to do so. However, if it is computationally intensive to generate draws $\mathbf{w} \sim P(\mathbf{W}|\phi(\mathbf{W}, \mathbf{X}) = 1)$ but easy to generate draws $\mathbf{w} \sim P_q(\mathbf{W})$ for some proposal distribution, then we recommend using the importance-sampling estimator $\hat{p}_{IS}$ while ensuring that the number of random samples M is large such that the bias of $\hat{p}_{IS}$ is minimal. For an in-depth discussion of rejection-sampling versus importance-sampling, see Robert and Casella (Chapter 3).[49]

The above procedure is closely related to the rerandomization framework developed by Morgan and Rubin,[50] who define an assignment criterion $\phi(\mathbf{W}, \mathbf{X})$ in order to ensure a certain level of covariate balance as part of an experimental design. Recent works on rerandomization have shown how $\phi(\mathbf{W}, \mathbf{X})$ can be flexibly defined: Morgan and Rubin[51] defined $\phi(\mathbf{W}, \mathbf{X})$ such that it incorporates tiers of importance for covariates, and Branson et al.[52] defined $\phi(\mathbf{W}, \mathbf{X})$ such that it incorporates tiers of importance for both covariates and multiple treatment effects of interest.

However, the purpose of the introduction of $\phi(\mathbf{W}, \mathbf{X})$ here is to conduct a conditional randomization test, rather than yield a desirable experimental design. It is similar to the conditional randomization test of Hennessy et al.,[47] who define $\phi(\mathbf{W}, \mathbf{X})$ in terms of categorical covariate balance. However, because Hennessy et al.[47] and other conditional randomization tests (e.g. Rosenbaum[27]) have focused on cases where propensity scores are equal across units or strata, they could sample from $P(\mathbf{W}|\phi(\mathbf{W}, \mathbf{X}) = 1)$ directly via random permutations of $\mathbf{W}^{obs}$. Indeed, both the rerandomization and conditional randomization test literature have focused on cases where the propensity scores are equal across units, whereas our approach addresses the more general case where propensity scores differ across units. Furthermore, if our rejection-sampling approach is computationally intensive, our importance-sampling approach allows one to still utilize random permutations of $\mathbf{W}^{obs}$ to quickly estimate the conditional randomization test p-value at the cost of incurring a small bias.

Now we establish that the unconditional and conditional randomization tests (i.e. the randomization test using p in equation (7) and the randomization test using p_{ϕ} in equation (17), respectively) are valid tests for Bernoulli trial experiments. While these are results for the randomization tests that use the exact p-values p and p_{ϕ}, this also suggests that our rejection-sampling approach for unbiasedly estimating p_{ϕ} yields valid statistical inferences. In Section 4, we empirically confirm the validity of these randomization tests, and we discuss to what extent our importance-sampling approach also yields valid statistical inferences.

3.4 Validity of unconditional and conditional randomization tests for Bernoulli trial experiments

For both theorems presented below, we assume that the treatment is assigned according to the strongly ignorable assignment mechanism (1). First, we establish that the randomization test that uses this assignment mechanism is valid, i.e. that the probability of this α-level randomization test falsely rejecting the Sharp Null Hypothesis is no

greater than α. This result is unsurprising given well-known results about the validity of randomization tests. Then, we establish that the conditional randomization test—i.e., the randomization test that uses the assignment mechanism $P(\mathbf{W}|\phi(\mathbf{W},\mathbf{X})=1)$ for some prespecified criterion $\phi(\mathbf{W},\mathbf{X})$ instead of the assignment mechanism (1)—is also valid. This result is slightly surprising in the sense that the validity of the randomization test holds even if the test uses an assignment mechanism other than the one used to conduct the randomized experiment.

Theorem 3.1 (Validity of Unconditional Randomization Test). *Assume that a randomized experiment is conducted using the strongly ignorable assignment mechanism (1). Define the two-sided randomization-test p-value as*

$$p \equiv \sum_{\mathbf{w}\in\mathbb{W}^+} \mathbb{I}(|t(Y(\mathbf{w}),\mathbf{w})| \geq |t^{obs}|)P(\mathbf{W}=\mathbf{w}) \tag{24}$$

for some test statistic $t(Y(\mathbf{W}),\mathbf{W})$, where $\mathbb{W}^+ = \{0,1\}^N$. Then the randomization test that rejects the Sharp Null Hypothesis when $p \leq \alpha$ is a valid test in the sense that

$$P(p \leq \alpha|H_0) \leq \alpha \tag{25}$$

where H_0 is the Sharp Null Hypothesis defined in (4).

Theorem 3.2 (Validity of Conditional Randomization Test). *Assume that a randomized experiment is conducted using the strongly ignorable assignment mechanism (1). Define the two-sided conditional randomization-test p-value as*

$$p_\phi \equiv \sum_{\mathbf{w}\in\mathbb{W}^+_\phi} \mathbb{I}(|t(Y(\mathbf{w}),\mathbf{w})| \geq |t^{obs}|)P(\mathbf{W}=\mathbf{w}) \tag{26}$$

for some test statistic $t(Y(\mathbf{W}),\mathbf{W})$, where $\mathbb{W}^+_\phi = \{\mathbf{w}\in\mathbb{W}^+ : \phi(\mathbf{w},\mathbf{X})=1\}$ is the set of acceptable randomizations according to some prespecified criterion $\phi(\mathbf{W},\mathbf{X})$. Then the randomization test that rejects the Sharp Null Hypothesis when $p_\phi \leq \alpha$ is a valid test in the sense that

$$P(p_\phi \leq \alpha|H_0) \leq \alpha \tag{27}$$

where H_0 is the Sharp Null Hypothesis defined in (4).

The proofs for Theorems 3.1 and 3.2 are given in Appendix 1.

Now we illustrate our randomization test procedure using a simple example where the randomization test p-value is computed exactly. Then we conduct a simulation study where the randomization test p-value is estimated, and we compare the rejection-sampling and importance-sampling approaches for estimating the p-value. Furthermore, we empirically confirm the validity of our randomization tests as established by Theorems 3.1 and 3.2 above, and we demonstrate how conditioning on various statistics of interest can be used to construct statistically powerful randomization tests for Bernoulli trial experiments.

4 Simulation study of unconditional and conditional randomization tests

4.1 Illustrative example: computing the exact p-value

As discussed in Section 2.2, the randomization-test p-value is typically approximated using equation (8) by drawing many possible treatment assignments $\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(M)}$. However, for small samples, the p-value can be computed exactly using equation (7) by examining each $\mathbf{w}$ in the set of possible treatment assignments $\mathbb{W}^+$. Here we explore a small-sample example to illustrate how to conduct randomization tests and construct confidence intervals when propensity scores vary across units. We also discuss how this procedure differs from the typical case where propensity scores are the same across units.

Consider a randomized experiment with $N=10$ units. The potential outcomes for these units are shown in Table 1, where the true treatment effect is $\tau = 0.5$. Say that a randomized experiment characterized by Bernoulli trials has occurred; the corresponding propensity scores, treatment assignment, and observed outcomes are also shown in Table 1. For now, assume that the task at hand is to conduct randomization-based inference for the average treatment effect given the treatment assignment, observed outcomes, and propensity scores in Table 1.

Table 1. Potential outcomes, treatment assignment, observed outcome, and propensity score for 10 units in a hypothetical randomized experiment.

Unit i	$Y_i(0)$	$Y_i(1)$	W_i^{obs}	y_i^{obs}	$e(\mathbf{x}_i)$
1	−0.56	−0.06	0	−0.56	0.1
2	−0.23	0.27	1	0.26	0.2
3	1.56	2.06	1	2.06	0.3
4	0.07	0.57	0	0.07	0.4
5	0.13	0.63	0	0.13	0.5
6	1.72	2.22	1	2.22	0.5
7	0.46	0.96	1	0.96	0.6
8	−1.27	−0.77	1	−0.77	0.7
9	−0.69	−0.19	0	−0.69	0.8
10	−0.45	0.05	1	0.05	0.9

Note: True treatment effect is $\tau = 0.5$.

With $N = 10$ units, only $2^{10} = 1024$ possible treatment assignments can be considered. Excluding the treatment assignments $\mathbf{0}_N$ and $\mathbf{1}_N$ leaves 1022 possible assignments. Under the Sharp Null Hypothesis, the observed outcomes $\mathbf{y}^{obs}$ will be the same as those in Table 1 for all 1022 of these assignments. We test this hypothesis following the three-step procedure in Section 2.2: First choose $\mathbb{W}^+$ and $P(\mathbf{W})$, then choose a test statistic, and finally compute the randomization test p-value.

We first consider the set $\mathbb{W}^+ = \{0,1\}^N \setminus (\mathbf{0}_N \cup \mathbf{1}_N)$ that was used during randomization, where

$$P(\mathbf{W} = \mathbf{w}|\mathbf{X}) = \frac{\prod_{i=1}^{N} e(\mathbf{x}_i)^{w_i}[1-e(\mathbf{x}_i)]^{1-w_i}}{1-\prod_{i=1}^{N} e(\mathbf{x}_i)-\prod_{i=1}^{N}[1-e(\mathbf{x}_i)]} \tag{28}$$

for each $\mathbf{w} \in \mathbb{W}^+$, as previously shown in equation (12). We choose the mean-difference estimator—given in equation (6)—as the test statistic. We then iterate through each of the 1022 treatment assignments $\mathbf{w} \in \mathbb{W}^+$ and compute the test statistic assuming the Sharp Null Hypothesis is true. Once this is done, the randomization test p-value can be computed exactly using

$$P(|t(Y(\mathbf{W}),\mathbf{W})| \geq |t^{obs}|) = \sum_{\mathbf{w}\in\mathbb{W}^+} \mathbb{I}(|t(Y(\mathbf{w}),\mathbf{w})| \geq |t^{obs}|)P(\mathbf{W} = \mathbf{w}) \tag{29}$$

as previously shown in equation (7). From Table 1, one can calculate the observed test statistic, t^{obs}, which is equal to 1.06.

Figure 1(a) shows the distribution of the absolute value of the test statistic $t(Y(\mathbf{w}),\mathbf{w})$ for each $\mathbf{w} \in \mathbb{W}^+$ assuming the Sharp Null Hypothesis is true. The portion of this distribution that corresponds to test statistics at least as large as the observed one is colored in gray. The randomization test p-value is then the probability of any gray treatment assignment occurring, which we find to be 0.12. If the propensity scores were equal across units—which is typically the case in the randomization test literature—then the randomization test p-value would simply be the number of gray treatment assignments divided by the total number of treatment assignments, which was, in this case, $\frac{164}{1022} \approx 0.16$. Thus, importantly, the p-value reflects the design of the randomized experiment—i.e., it incorporates the propensity scores that were used to randomize the units during the experiment.

Furthermore, we can obtain a confidence interval for the average treatment effect by inverting this randomization test using the procedure outlined in Section 2.3. We did a line search of values $\tau \in \{-3, -2.9, \ldots, 2.9, 3\}$ and defined our 95% confidence interval as the set of τ's for which we obtained p-values greater than 0.05 when testing the hypothesis (9) for each τ. We found the confidence interval to be $(-0.1, 2.4)$. Again, this confidence interval reflects the design of the randomized experiment, because the p-values corresponding to each τ depend on the propensity scores that were used during randomization.

Note that Figure 1(a) displays every possible treatment assignment, including assignments where only one unit is assigned to treatment and the rest to control (and vice versa). However, researchers may want the statistical analysis to only consider treatment assignments similar to the observed one. For example, consider the more

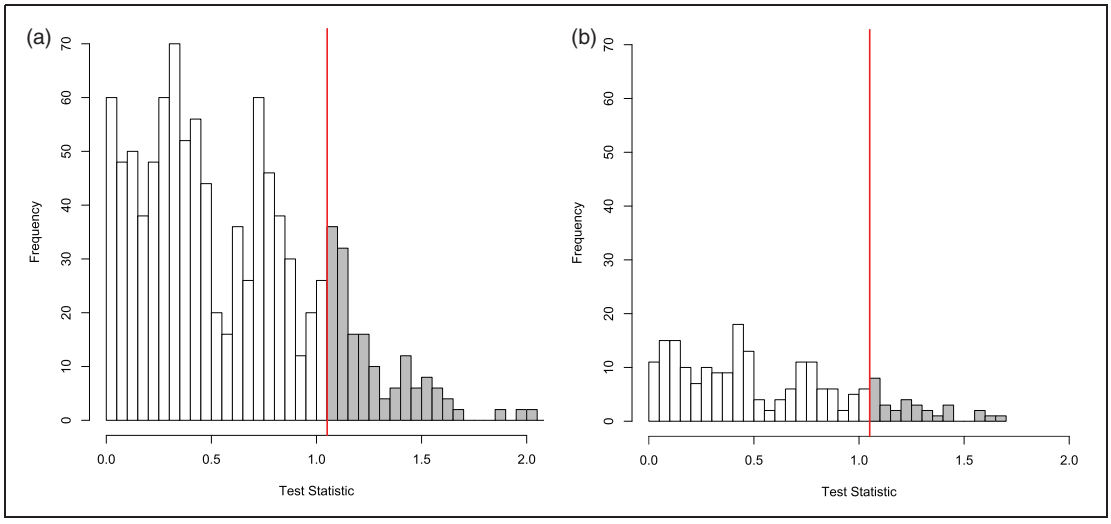

Figure 1. Unconditional and conditional randomization distributions of the test statistic under the Sharp Null Hypothesis. (a) The distribution of $t(Y(\mathbf{w}), \mathbf{w})$ for each $\mathbf{w} \in \mathbb{W}^+$, where $\mathbb{W}^+ = \{0, 1\}^N \setminus (\mathbf{0}_N \cup \mathbf{1}_N)$. The observed test statistic is marked by a red vertical line. Assignments corresponding to test statistics at least as large as the observed one are in gray. (b) The distribution of $t(Y(\mathbf{w}), \mathbf{w})$ for each $\mathbf{w} \in \mathbb{W}^+$, where $\mathbb{W}^+ = \{\mathbf{W} \in \mathbb{W} | \sum_{i=1}^N W_i = N_T\}$.

stringent set of treatment assignments $\mathbb{W}^+ = \{\mathbf{W} \in \mathbb{W} | \sum_{i=1}^N W_i = N_T\}$, where in this example the number of treated units $N_T = 6$, as seen in Table 1. Figure 1(b) shows the distribution of the test statistic for each $\mathbf{w} \in \mathbb{W}^+$ in this case, assuming the Sharp Null Hypothesis is true. Note that there are only $\binom{10}{6} = 210$ treatment assignments, which is a subset of the assignments displayed in Figure 1(a). Again, the randomization test p-value is the probability of any gray treatment assignment occurring, but now the probability of any $\mathbf{w} \in \mathbb{W}^+$ is

$$P(\mathbf{W} = \mathbf{w}|\mathbf{X}) = \frac{\prod_{i=1}^N e(\mathbf{x}_i)^{w_i}[1 - e(\mathbf{x}_i)]^{1-w_i}}{P(\sum_{i=1}^N W_i = N_T|\mathbf{X})} \tag{30}$$

as previously shown in equation (14). Because there are only 210 treatment assignments $\mathbf{w}$ such that $\sum_{i=1}^N w_i = N_T$, we can compute the denominator exactly and thus compute the randomization test p-value exactly as well, which we find to be equal to 0.17. Furthermore, using the same procedure as above, we found the 95% confidence interval to be $(-0.1, 2.4)$. Thus, in addition to reflecting the experimental design, randomization-based inference can also reflect particular experiments of interest, such as ones similar to the observed one.

Now we conduct a simulation study with $N = 100$ units. In this case, it is computationally intensive to compute randomization test p-values exactly, and we instead approximate them. Furthermore, because the propensity scores vary across units, it will be difficult to directly sample from conditional probability distributions such as $P(\mathbf{W} | \sum_{i=1}^N W_i = N_T)$, and thus we will need the rejection-sampling procedure from Section 3.3 to conduct conditional inference.

4.2 Simulation setup

Hennessy et al.[47] conducted a simulation study to show that their randomization test that conditioned on categorical covariate balance was more powerful than unconditional randomization tests when covariates were associated with the outcome. Hennessy et al.[47] consider the case where the propensity scores are the same across units. We modify their simulation study such that units' propensity scores differ. This simulation study serves two purposes:

(1) Confirm the validity of the unconditional and conditional randomization tests discussed in Section 3.2, as established by Theorems 3.1 and 3.2.

Table 2. Contingency table of the number of units assigned to treatment and control (N_T and N_C) and the number of units with covariate values $X = 1$ and $X = 2$ (N_1 and N_2).

		W		
		1	0	
X	1	N_{T1}	N_{C1}	$N_1 = 50$
	2	N_{T2}	N_{C2}	$N_2 = 50$
		N_T	N_C	$N = 100$

Note: The values $N_1 = 50$, and $N_2 = 50$ were fixed across randomizations in the simulation study; the other values varied across randomizations.

(2) Demonstrate how the rejection-sampling and importance-sampling procedures presented in Section 3.3 can be used to construct statistically powerful conditional randomization tests.

Consider $N = 100$ units with a single covariate X, where 50 units have covariate value $X = 1$ and the other 50 units have covariate value $X = 2$. Each unit has two potential outcomes—corresponding to treatment and control—which are generated once from the following

$$\begin{aligned} Y_i(0)|X_i &\sim N(\lambda X_i, 1), \quad i = 1, \ldots, N \\ Y_i(1) &= Y_i(0) + \tau \end{aligned} \tag{31}$$

The parameter λ determines the strength of the association between X and the potential outcomes, while τ is the treatment effect. Similar to Hennessy et al.,[47] we consider the values $\lambda \in \{0, 1.5, 3\}$ and $\tau \in \{0, 0.1, \ldots, 1\}$ in our simulation. The previous example from Table 1 was generated using $\lambda = 0$ and $\tau = 0.5$.

The probability of the ith unit receiving treatment—i.e. its propensity score—was generated once from the following

$$P(W_i = 1|X_i) = P(W_i = 1) \sim \text{Beta}(5, 5), \ i = 1, \ldots, N \tag{32}$$

This generating mechanism resulted in propensity scores being centered but spread around 0.5. In our simulation, propensity scores ranged from 0.22 to 0.87 with a mean of 0.49.

After the potential outcomes and propensity scores were generated, we randomly assigned units to treatment and control according to the probability distribution $P(\mathbf{W})$ defined by the propensity scores. We prevented any single treatment assignment from being $\mathbf{0}_N$ or $\mathbf{1}_N$; in other words, we considered the set of possible treatment assignments $\mathbb{W}^+ = \{0, 1\}^N \setminus (\mathbf{0}_N \cup \mathbf{1}_N)$ during randomization. In this case, there will always be 50 units with $X = 1$ and 50 units with $X = 2$, but the number of units assigned to treatment and control can vary from randomization to randomization. Any randomization of the 100 units to treatment and control can be summarized by Table 2, which includes the number of units assigned to treatment and control (N_T and N_C) and the number of units with covariate values $X = 1$ and $X = 2$ (N_1 and N_2).

Before conducting the full simulation, let us first consider one possible treatment assignment that we may observe during this simulation. We will present four randomization tests one could use to test the Sharp Null Hypothesis.

4.3 Example of one treatment assignment

Consider the case when $\lambda = 3$ and $\tau = 0.5$, i.e. when the covariate is strongly associated with the outcome and the treatment effect is moderate. The potential outcomes were generated using equation (31), the propensity scores were generated using equation (32), and then units were randomized by flipping biased coins corresponding to these propensity scores. Table 3 shows the resulting randomization. Given this randomization and the corresponding dataset, how should we test the Sharp Null Hypothesis?

Any randomization test should involve generating treatment assignments via biased coins corresponding to the prespecified propensity scores, because this is how the randomization observed in Table 3 was generated. However,

Table 3. Example of a possible treatment allocation in our simulation study.

		$\mathbf{W}^{obs}$		
		1	0	
X	1	$N_{T1}^{obs} = 30$	$N_{C1}^{obs} = 20$	$N_1 = 50$
	2	$N_{T2}^{obs} = 24$	$N_{C2}^{obs} = 26$	$N_2 = 50$
		$N_T^{obs} = 54$	$N_C^{obs} = 46$	$N = 100$

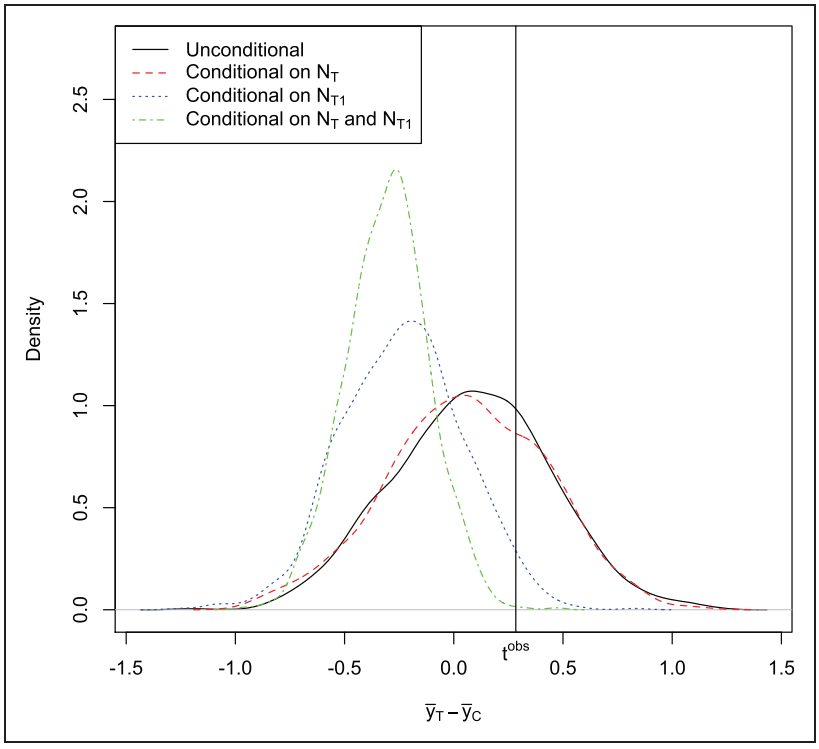

Figure 2. The unconditional and conditional randomization distributions for the mean-difference test statistic under the Sharp Null Hypothesis for the example in Table 3. The observed test statistic for this example dataset is marked by a black vertical line. Each randomization distribution was approximated by drawing $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(1000)}$ from the corresponding $\mathbb{W}^+$ using the rejection-sampling procedure discussed in Section 3.3.

which set of possible treatment assignments $\mathbb{W}^+$ should one consider during the test? We consider four different $\mathbb{W}^+$ and their associated randomization tests:

(1) An unconditional randomization test (as presented in Section 2.2), with $\mathbb{W}^+ = \{0,1\}^N \setminus (\mathbf{0}_N \cup \mathbf{1}_N)$.
(2) A randomization test conditional on the number of units assigned to treatment, with $\mathbb{W}^+ = \{\mathbb{W} | \sum_{i=1}^N W_i = N_T^{obs}\}$.
(3) A randomization test conditional on the number of units with $X=1$ assigned to treatment, with $\mathbb{W}^+ = \{\mathbb{W} | \sum_{i:X_i=1} W_i = N_{T1}^{obs}\}$.
(4) A randomization test conditional on N_T and N_{T1}, with $\mathbb{W}^+ = \{\mathbb{W} | \sum_{i=1}^N W_i = N_T^{obs}$ and $\sum_{i:X_i=1} W_i = N_{T1}^{obs}\}$.

Arguably, the first randomization test is the most natural choice, because it corresponds to the $\mathbb{W}^+$ that was actually used to generate the randomization observed in Table 3; however, because conditional randomization tests can be more powerful than unconditional randomization tests, the other three tests may be options researchers might consider as well.

The above tests are ordered in terms of the restrictiveness of $\mathbb{W}^+$: the first two randomization tests involve flipping biased coins to generate treatment assignments, where the values N_{T1}, N_{C1}, N_{T2}, and N_{C2} in Table 2 can vary across assignments; in the third randomization test, only N_{T2} and N_{C2} can vary; and in the fourth randomization test, none of these values can vary. Because iterating through every possible treatment assignment in $\mathbb{W}^+$ is computationally intensive—for the example in Table 3, $|\mathbb{W}^+| = 2^{100} - 2$ for the first test, and $|\mathbb{W}^+| = \binom{50}{30}$ for the fourth test—we instead generate 1000 treatment assignments $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(1000)}$ using our rejection-sampling procedure discussed in Section 3.3 to approximate the randomization distribution for each test.

The approximate randomization distribution of the mean-difference test statistic $\bar{y}_T - \bar{y}_C$ under the Sharp Null Hypothesis for each of these four tests is shown in Figure 2. The conditional randomization distributions for the third and fourth tests are shifted to the left of the unconditional randomization distribution. This is no coincidence: in Table 3, there are more units with $X = 1$ in the treatment group and more units with $X = 2$ in the control group; as a result, the treatment group will have units with systematically lower potential outcomes, due to the potential outcomes model (31). This is reflected in the conditional randomization distributions but not the unconditional one. Consequently, the conditional and unconditional randomization tests will give different results: two-sided p-values for the four tests are 0.58, 0.57, 0.08, and 0.00, respectively. This suggests that some of these randomization tests may be more powerful at detecting a treatment effect than others, which we further explore below.

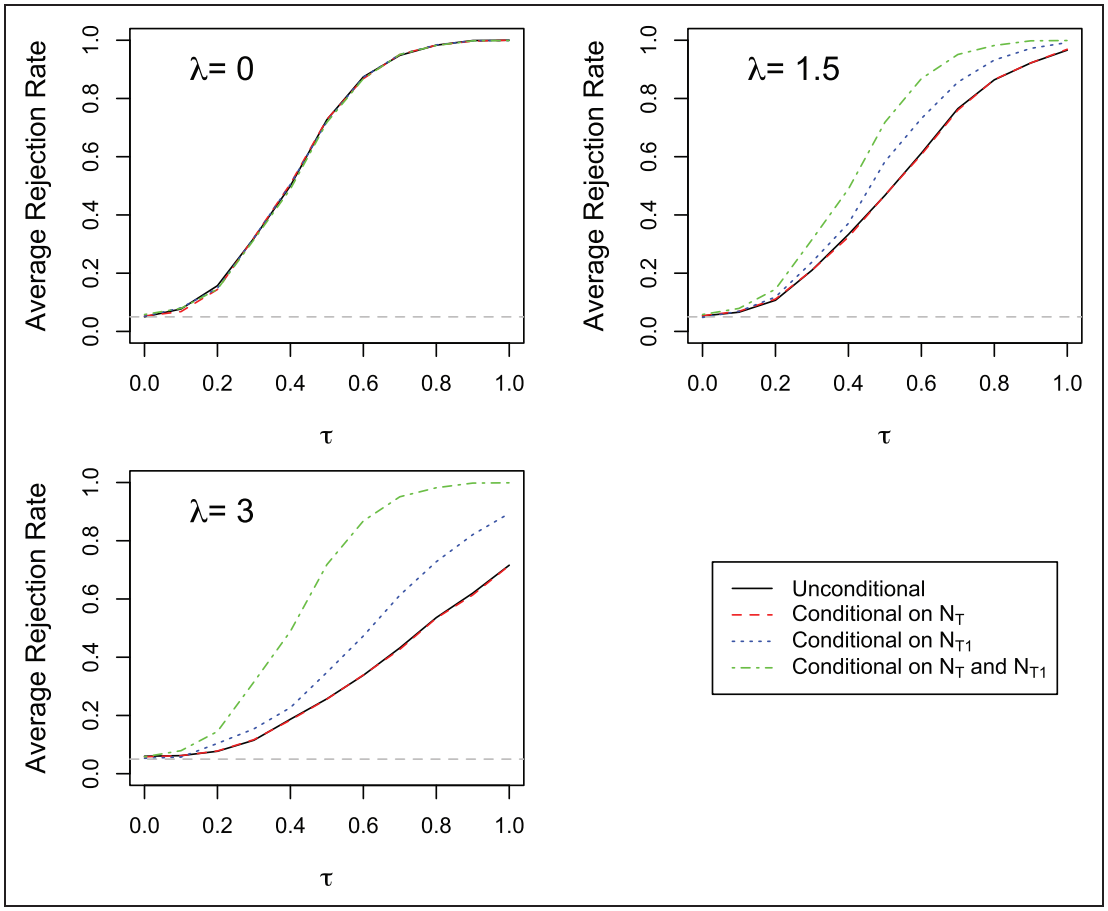

Figure 3. Average rejection rates for the four randomization tests across 1000 randomizations for each value of λ and τ. As λ increases, the covariate becomes more associated with the outcome; as τ increases, the treatment effect should become easier to detect. The gray horizontal line marks 0.05.

4.4 Full simulation study

Now we compare the four randomization tests discussed in Section 4.3 in terms of their power. For each combination of $\lambda \in \{0, 1.5, 3\}$ and $\tau \in \{0, 0.1, \ldots, 1\}$, the potential outcomes were generated using equation (31), the propensity scores were generated using equation (32), and then units were randomized 1000 times by flipping biased coins corresponding to these propensity scores.

For each of the 1000 randomizations, we performed the four randomization tests discussed in Section 4.3 using the rejection-sampling approach to unbiasedly estimate each p-value using $\hat{p}_{RS}$ given in equation (18). For each test, we rejected the Sharp Null Hypothesis if $\hat{p}_{RS} \leq 0.05$. Figure 3 displays the average rejection rate of the Sharp Null Hypothesis—i.e. the power—for each randomization test. When $\tau = 0$, the Sharp Null Hypothesis is true, and all of the randomization tests reject the null 5% of the time. This confirms the validity of our unconditional and conditional randomization tests, as established by Theorems 3.1 and 3.2. When $\lambda = 0$, the covariate is not associated with the outcome, and all of the randomization tests are essentially equivalent. As the covariate becomes more associated with the outcome, the third and fourth conditional randomization tests become more powerful than the unconditional test, while the randomization test that only conditions on N_T remains equivalent to the unconditional randomization test. This is due to the fact that the quantity N_{T1} combined with N_T may be confounded with the treatment effect if there is covariate imbalance between the treatment and control groups, as in the example presented in Table 3 and Figure 2.

However, our rejection-sampling approach can be computationally expensive. Generating 1000 samples for the unconditional randomization test, the randomization test conditional on N_T, the randomization test conditional on N_{T1}, and the randomization test conditional on N_T and N_{T1} took on an average of 0.25 s, 1.22 s, 2.14 s, and

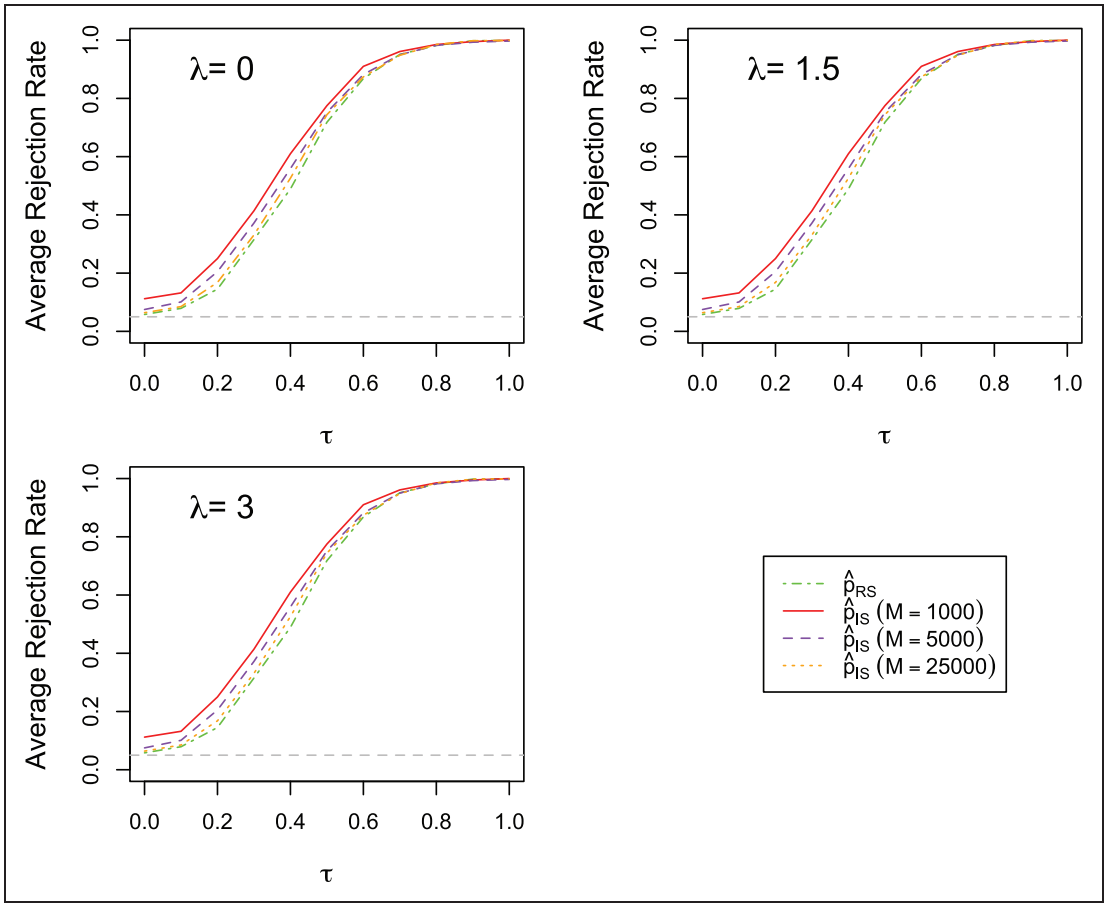

Figure 4. Average rejection rates for the rejection-sampling and importance-sampling approaches conditional on N_T and N_{T1}. For importance-sampling, we tried various numbers of proposals M. The line for $\hat{p}_{RS}$ (i.e. the rejection-sampling approach) is the same as the line for "Conditional on N_T and N_{T1}" in Figure 3.

34.75 s, respectively. As an alternative to the rejection-sampling approach for computing the randomization test p-value $\hat{p}_{RS}$ conditional on N_T and N_{T1}, we can take our importance-sampling approach discussed in Section 3.3. Instead of sampling directly from $P(\mathbf{W}|\phi(\mathbf{W},\mathbf{X})=1)$ via rejection-sampling, we generate M proposals $\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(M)}$ uniformly from the set of acceptable randomizations $\{\mathbf{w}:\sum_{i=1}^{N} w_i = N_T$ and $\sum_{i:w_i=1}\mathbb{I}(X_i=1)=N_{T1}\}$; this corresponds to random permutations of $\mathbf{W}^{obs}$ within the $X=1$ and $X=2$ strata. Then, we compute $\hat{p}_{IS}$ given in equation (23) and reject if $\hat{p}_{IS}\leq 0.05$.

Figure 4 compares the rejection-sampling approach (i.e. rejecting the Sharp Null Hypothesis if $\hat{p}_{RS}\leq 0.05$) with the importance-sampling approach (i.e. rejecting the Sharp Null Hypothesis if $\hat{p}_{IS}\leq 0.05$) for different values of M. The importance-sampling approach is computationally less intensive than the rejection-sampling approach: the importance-sampling approach using $M=1000$, $M=5000$, and $M=25000$ took on an average of 0.68 s, 3.30 s, and 16.31 s, respectively. Note that even the $M=25000$ case required less than half the time as the rejection-sampling approach. However, as noted in Section 3.3, $\hat{p}_{IS}$ has a bias of order M^{-1}, and thus the p-value for the importance-sampling approach may be notably biased for low M. This can be seen in Figure 4: for $M=1000$, the importance-sampling approach falsely rejects the Sharp Null Hypothesis when $\tau=0$ at a substantially higher rate than 0.05; this suggests that the importance-sampling approach has a negative bias in this case. However, as M increases, this bias is less substantial, and results using $\hat{p}_{IS}$ approach those using $\hat{p}_{RS}$. Thus, the bias of importance-sampling can break the validity of our randomization test, but this can be alleviated by increasing the number of proposals M at a minimal computational cost.

In summary, these results reinforce the idea of Hennessy et al.[47] that conditional randomization tests are more powerful than unconditional randomization tests when the acceptance criterion $\phi(\mathbf{W},\mathbf{X})$ incorporates statistics that are associated with the outcome. Furthermore, this demonstrates how our rejection-sampling procedure can be used to condition on several combinations of statistics of interest, thus yielding statistically powerful randomization tests for Bernoulli trial experiments. Finally, when this rejection-sampling procedure is computationally intensive, our importance-sampling approach is a viable alternative; however, we recommend generating a large number of proposals M such that the bias of the importance-sampling approach is negligible and thus still yields valid statistical inferences.

5 Discussion and conclusion

Here we presented a randomization-based inference framework for experiments whose assignment mechanism is characterized by independent Bernoulli trials. Our framework and corresponding randomization tests encapsulate all strongly ignorable assignment mechanisms, including experiments based on complete, blocked, and paired randomization, as well as the general case where propensity scores differ across all units. In particular, we introduced rejection-sampling and importance-sampling approaches for obtaining randomization-based point estimates and confidence intervals conditional on any statistics of interest for Bernoulli trial experiments, which has not been previously studied in the literature. We also established that our randomization test is a valid test, and the power of this test can be improved by conditioning on various statistics of interest without sacrificing the validity of the test.

While our discussion of point estimates and confidence intervals are based on a sharp hypothesis that assumes a constant additive treatment effect, our framework can be extended to any sharp hypothesis, including those that incorporate heterogeneous treatment effects. Recent works in the randomization-based inference literature have begun to address treatment effect heterogeneity (e.g. Ding et al.[53] and Caughey et al.[44]), and our framework can be extended to these discussions.

Throughout, we assumed that the propensity scores are known, as in randomized experiments. In observational studies, the propensity scores are estimated, typically with model-based methodologies like logistic regression. Nonetheless, propensity score methodologies still assume a strongly ignorable assignment mechanism as in equation (1), with the assumption that the estimated propensity scores $\hat{e}(\mathbf{x})$ are "close" to the true $e(\mathbf{x})$, i.e. the propensity score model is well-specified. An implication of our randomization-based inference framework is that it can still be applied to observational studies, where estimates $\hat{e}(\mathbf{x})$ are used instead of known $e(\mathbf{x})$. Such a test is valid to the extent that the $\hat{e}(\mathbf{x})$ are "close" to the true $e(\mathbf{x})$; this is not a limitation of our framework specifically but of propensity score methodologies in general. Determining when our randomization test is valid for observational studies is future work.

However, our randomization test would seem to be the most natural randomization test to use for observational studies, because it directly reflects the strongly ignorable assignment mechanism (1) that is assumed in most of the observational study literature. Other proposed randomization tests for observational

studies reflect other assignment mechanisms, such as blocked and paired assignment mechanisms; these randomization tests are not immediately applicable to cases where the propensity score varies across all units.

There are many other methodologies for analyzing randomized experiments and observational studies, such as regression with or without inverse probability weighting, matching, and Bayesian modeling. Importantly, all of these methodologies assume the strongly ignorable assignment mechanism (1) in addition to other assumptions about model specification, asymptotics, or units' propensity scores within covariate strata. Our framework only makes the strongly ignorable assignment mechanism assumption, and thus is a minimal-assumption approach while still yielding point estimates and confidence intervals that directly reflect the assignment mechanism. Furthermore, we established the validity of our randomization test and demonstrated how conditioning on relevant statistics of interest can yield powerful randomization tests for Bernoulli trial experiments.

Acknowledgements

We would like to thank Donald Rubin for enlightening discussions about this work, as well as two anonymous referees for their thoughtful comments, which substantially improved this work.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1144152, and by the Office of the Director, National Institutes of Health under Award Number DP5OD021412. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.

ORCID iDs

Zach Branson http://orcid.org/0000-0001-9496-9979
Marie-Abèle http://orcid.org/0000-0002-0422-6651

References

1. Fisher RA. *The design of experiments*. Edinburgh, UK: Oliver and Boyd, 1935.
2. Neyman J. On the application of probability theory to agricultural experiments. Essay on principles (with discussion). Section 9 (translated). *Stat Sci* 1923; **5**: 465–472.
3. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Education Psychol* 1974; **66**: 688.
4. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. *J Am Stat Assoc* 2005; **100**: 322–331.
5. Rosenbaum PR. *Observational studies*. New York, NY: Springer, 2002.
6. Rosenbaum PR. Covariance adjustment in randomized experiments and observational studies. *Stat Sci* 2002; **17**: 286–327.
7. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med* 2007; **26**: 20–36.
8. Rubin DB. For objective causal inference, design trumps analysis. *Ann Appl Stat* 2008; **2**: 808–840.
9. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
10. Dehejia RH and Wahba S. Propensity score-matching methods for nonexperimental causal studies. *Rev Econ Stat* 2002; **84**: 151–161.
11. Sekhon JS. Opiates for the matches: matching methods for causal inference. *Ann Rev Political Sci* 2009; **12**: 487–508.
12. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci: Rev J Inst Math Stat* 2010; **25**: 1.
13. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 2011; **46**: 399–424.
14. Ho DE, Imai K, King G, et al. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analys* 2007; **15**: 199–236.

15. Robins JM and Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *J Am Stat Assoc* 1995; **90**: 122–129.
16. Rubin DB and Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. *J Am Stat Assoc* 2000; **95**: 573–585.
17. Rubin DB. Bayesian inference for causal effects: the role of randomization. *Ann Stat* 1978; **6**: 34–58.
18. Zigler CM and Dominici F. Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *J Am Stat Assoc* 2014; **109**: 95–107.
19. Hirano K and Imbens GW. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Services Outcomes Res Methodol* 2001; **2**: 259–278.
20. Hirano K, Imbens GW and Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 2003; **71**: 1161–1189.
21. Lunceford JK and Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med* 2004; **23**: 2937–2960.
22. Hernán MA and Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Commun Health* 2006; **60**: 578–586.
23. Cole SR and Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* 2008; **168**: 656–664.
24. Austin PC and Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med* 2015; **34**: 3661–3679.
25. Imbens GW and Rubin DB. *Causal inference in statistics, social, and biomedical sciences*. Cambridge: Cambridge University Press, 2015.
26. Basu D. Randomization analysis of experimental data: the Fisher randomization test. *J Am Stat Assoc* 1980; **75**: 575–582.
27. Rosenbaum PR. Conditional permutation tests and the propensity score in observational studies. *J Am Stat Assoc* 1984; **79**: 565–574.
28. Rosenbaum PR. Permutation tests for matched pairs with adjustments for covariates. *Appl Stat* 1988; **37**: 401–411.
29. Efron B. Forcing a sequential experiment to be balanced. *Biometrika* 1971; **58**: 403–417.
30. Wei L. An application of an urn model to the design of sequential controlled clinical trials. *J Am Stat Assoc* 1978; **73**: 559–563.
31. Soares JF and Wu C. Some restricted randomization rules in sequential designs. *Commun Stat-Theory Meth* 1983; **12**: 2017–2034.
32. Smythe R and Wei L. Significance tests with restricted randomization design. *Biometrika* 1983; 496–500.
33. Wei L. Exact two-sample permutation tests based on the randomized play-the-winner rule. *Biometrika* 1988; **75**: 603–606.
34. Mehta CR, Patel NR and Wei L. Constructing exact significance tests with restricted randomization rules. *Biometrika* 1988; **75**: 295–302.
35. Good P. *Permutation tests: a practical guide to resampling methods for testing hypotheses*. New York, NY: Springer Science & Business Media, 2013.
36. Pocock SJ and Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 1975; 103–115.
37. Loux TM. A simple, flexible, and effective covariate-adaptive treatment allocation procedure. *Stat Med* 2013; **32**: 3775–3787.
38. Lin Y, Zhu M and Su Z. The pursuit of balance: an overview of covariate-adaptive randomization techniques in clinical trials. *Contemporary Clin Trial* 2015; **45**: 21–25.
39. Simon R and Simon NR. Using randomization tests to preserve type I error with response adaptive and covariate adaptive randomization. *Stat Probabil Lett* 2011; **81**: 767–772.
40. Shao J and Yu X. Validity of tests under covariate-adaptive biased coin randomization and generalized linear models. *Biometrics* 2013; **69**: 960–969.
41. Pitman EJG. Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika* 1938; **29**: 322–335.
42. Kempthorne O. The design and analysis of experiments. New York: John Wiley & Sons, Inc., 1952.
43. Hodges JL and Lehmann EL. Estimates of location based on rank tests. *Ann Math Stat* 1963; **34**: 598–611.
44. Caughey D, Dafoe A and Miratix L. Beyond the sharp null: permutation tests actually test heterogeneous effects. In: *Summer meeting of the Society for Political Methodology*, Rice University, 21-23 July 2016. vol. 22.
45. Chen SX and Liu JS. Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Stat Sinica* 1997; **7**: 875–892.
46. Lohr S. *Sampling: Design and Analysis*. 2nd ed. Pacific Grove, CA: Duxbury Press, 2009.
47. Hennessy J, Dasgupta T, Miratrix L, et al. A conditional randomization test to account for covariate imbalance in randomized experiments. *J Causal Inference* 2016; **4**: 61–80.
48. Kong A. *A note on importance sampling using standardized weights*. Department of Statistics, University of Chicago, Tech Report, 1992, pp.348.
49. Robert CP and Casella G. *Monte Carlo statistical methods*. New York, NY: Springer, 1999.

50. Morgan KL and Rubin DB. Rerandomization to improve covariate balance in experiments. *Ann Stat* 2012; **40**: 1263–1282.
51. Morgan KL and Rubin DB. Rerandomization to balance tiers of covariates. *J Am Stat Assoc* 2015; **110**: 1412–1421.
52. Branson Z, Dasgupta T, Rubin DB, et al. Improving covariate balance in 2K factorial designs via rerandomization with an application to a New York City Department of Education High School Study. *Ann Appl Stat* 2016; **10**: 1958–1976.
53. Ding P, Feller A and Miratrix L. Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2016; **78**: 655–671.

Appendix 1. Proof of Theorem 3.1

This proof closely follows the proof provided in Hennessey et al.[47] (p. 64), but with a focus on Bernoulli trial experiments instead of completely randomized experiments.

Define T_W as a random variable whose distribution is the same as $|t(Y(\mathbf{W}), \mathbf{W})|$, for some test statistic $t(Y(\mathbf{W}), \mathbf{W})$, where $\mathbf{W} \sim P(\mathbf{W}|\mathbf{X})$ is specified by the strongly ignorable assignment mechanism (1). Furthermore, let $F_{T_W}(\cdot)$ be the CDF of T_W. Note that T_W must be defined for all $\mathbf{W} \in \mathbb{W}^+$, including $\mathbf{W} = \mathbf{1}_N$ or $\mathbf{W} = \mathbf{0}_N$; without loss of generality, one can define $T_W = 0$ for these two cases.

Under the Sharp Null Hypothesis H_0 defined in equation (4), $Y(\mathbf{W}) = \mathbf{y}^{obs}$ for all $\mathbf{W} \in \mathbb{W}^+$. Thus, under H_0

$$|t(\mathbf{y}^{obs}, \mathbf{W})| \sim T_W \tag{33}$$

i.e. the distribution of the observed test statistic $|t^{obs}| \equiv |t(\mathbf{y}^{obs}, \mathbf{W}^{obs})|$ across randomizations is the same as the distribution of T_W.

Now note that the randomization test p-value defined in equation (24) of Theorem 3.1 is such that, under H_0

$$p = 1 - F_{T_W}(|t^{obs}|) \tag{34}$$

Furthermore, given equation (33), we have that the distribution of p across randomizations is

$$p \sim 1 - F_{T_W}(T_W) \tag{35}$$

under H_0.

If T_W were continuous, then $(1 - F_{T_W}(T_W)) \sim \text{Unif}(0, 1)$ by the probability integral transform; however, T_W is discrete due to the discreteness of $\mathbb{W}^+$. Nonetheless, $(1 - T_W)$ stochastically dominates $U \sim \text{Unif}(0, 1)$, and thus

$$P(p \leq \alpha | H_0) \leq P(U \leq \alpha | H_0) \tag{36}$$

$$\leq \alpha \tag{37}$$

where equation (36) follows from the definition of stochastic dominance, and equation (37) follows from properties of the standard uniform distribution. This concludes the proof of Theorem 3.1.

Proof of Theorem 3.2

Define a set of partitions $\mathbb{W}_1^+, \ldots, \mathbb{W}_B^+$, where $\mathbb{W}_b^+ \cap \mathbb{W}_{b'}^+ = \emptyset$ for all $b \neq b'$ and $\cup_{b=1}^B \mathbb{W}_b^+ = \mathbb{W}^+ = \{0, 1\}^N$. In other words, the $\mathbb{W}_1^+, \ldots, \mathbb{W}_B^+$ partition the set of possible randomizations under the strongly ignorable assignment mechanism (1) into non-overlapping sets. Consider a randomization test that is conducted only within a particular one of these partitions; the associated randomization test p-value is

$$p_b \equiv \sum_{\mathbf{w} \in \mathbb{W}_b^+} \mathbb{I}(|t(Y(\mathbf{w}), \mathbf{w})| \geq |t^{obs}|) P(\mathbf{W} = \mathbf{w}) \tag{38}$$

Importantly, by Theorem 3.1, for randomizations $\mathbf{W} \in \mathbb{W}_b^+$, the randomization test that rejects the Sharp Null Hypothesis when $p_b \leq \alpha$ is a valid test, i.e.

$$P(p_b \leq \alpha | H_0, \mathbf{W} \in \mathbb{W}_b^+) \leq \alpha \quad \text{for all } b = 1, \ldots, B \tag{39}$$

The acceptance criterion $\phi(\mathbf{W}, \mathbf{X})$ determines the particular partition in which the conditional randomization test is conducted. Without loss of generality, say that $\phi(\mathbf{W}, \mathbf{X})$ is defined such that

$$\phi(\mathbf{W}, \mathbf{X}) \equiv \begin{cases} 1 & \text{if} \quad \mathbf{W} \in \mathbb{W}_b^+ \quad \text{for some } b = 1, \ldots, B \\ 0 & \text{otherwise.} \end{cases} \tag{40}$$

Defined this way, $\phi(\mathbf{W}, \mathbf{X})$ varies across randomizations $\mathbf{W} \in \mathbb{W}^+$; as a result, the set of acceptable randomizations $\mathbb{W}_\phi^+ \equiv \{\mathbf{w} \in \mathbb{W}^+ : \phi(\mathbf{w}, \mathbf{X}) = 1\}$ varies across $\mathbf{W} \in \mathbb{W}^+$ as well. As an example, consider the criterion $\phi(\mathbf{W}, \mathbf{X})$ defined as equal to 1 if $\sum_{i=1}^N W_i = N_T$ and equal to 0 otherwise. The number of treated units N_T can vary across $\mathbf{W} \in \mathbb{W}^+$, and thus $\mathbb{W}_\phi^+$ will vary across $\mathbf{W} \in \mathbb{W}^+$ as well, based on the realization of N_T. In this case, the partitions $\mathbb{W}_1^+, \ldots, \mathbb{W}_B^+$ are defined as the sets of treatment assignments corresponding to the unique values of N_T. In general, the criterion $\phi(\mathbf{W}, \mathbf{X})$ will be a function of statistics, and the partitions $\mathbb{W}_1^+, \ldots, \mathbb{W}_B^+$ can be defined by the unique values of these statistics. This setup is a generalization of the covariate balance function discussed in Hennessy et al.[47] (p. 67).

Thus, for each $b = 1, \ldots, B$, there is an associated probability $P(\mathbf{W} \in \mathbb{W}_b^+) = P(\mathbb{W}_\phi^+ = \mathbb{W}_b^+)$, and this probability is determined by the strongly ignorable assignment mechanism (1). Once it is determined which partition the set of acceptable randomizations is equal to, the randomization test is conducted within this partition, i.e. the p-value p_b is used for the b such that $\mathbb{W}_\phi^+ = \mathbb{W}_b^+$.

Thus, for the conditional randomization test p-value p_ϕ defined in Theorem 3.2, we have that

$$P(p_\phi \leq \alpha | H_0) = \sum_{b=1}^B P(p_\phi \leq \alpha | H_0, \mathbb{W}_\phi^+ = \mathbb{W}_b^+) P(\mathbb{W}_\phi^+ = \mathbb{W}_b^+) \text{ (by law of total probability)} \tag{41}$$

$$= \sum_{b=1}^B P(p_b \leq \alpha | H_0, \mathbb{W}_\phi^+ = \mathbb{W}_b^+) P(\mathbb{W}_\phi^+ = \mathbb{W}_b^+) \tag{42}$$

$$\leq \sum_{b=1}^B \alpha P(\mathbb{W}_\phi^+ = \mathbb{W}_b^+) \quad \text{(by Theorem 3.1)} \tag{43}$$

$$= \alpha \left(\text{because } \sum_{b=1}^B P(\mathbb{W}_\phi^+ = \mathbb{W}_b^+) = 1 \right) \tag{44}$$

which is our desired result.