Randomization Inference beyond the Sharp Null: Bounded Null Hypotheses and Quantiles of Individual Treatment Effects

Devin Caughey, Allan Dafoe, Xinran Li and Luke Miratrix *

Abstract

Randomization inference (RI) is typically interpreted as testing Fisher's "sharp" null hypothesis that all unit-level effects are exactly zero. This hypothesis is often criticized as restrictive and implausible, making its rejection scientifically uninteresting. We show, however, that many randomization tests are also valid for a "bounded" null hypothesis under which the unit-level effects are all non-positive (or all non-negative) but are otherwise heterogeneous. In addition to being more plausible a priori, bounded nulls are closely related to substantively important concepts such as monotonicity and Pareto efficiency. Reinterpreting RI in this way also dramatically expands the range of inferences possible in this framework. We show that exact confidence intervals for the maximum (or minimum) unit-level effect can be obtained by inverting tests for a sequence of bounded nulls. We also generalize RI to cover inference for quantiles of the individual effect distribution as well as for the proportion of individual effects larger (or smaller) than a given threshold. The proposed confidence intervals for all effect quantiles are simultaneously valid, in the sense that no correction for multiple analyses is required, and are thus a "free lunch" added to conventional RI. In sum, our reinterpretation and generalization provide a broader justification for randomization tests and a basis for exact nonparametric inference for effect quantiles. We illustrate our methods with simulations and applications, finding that Stephenson rank statistics can provide more informative results than the more common Wilcoxon rank or difference-in-means statistics. We also provide an R package RIQITE implementing the proposed approach.

Keywords: causal inference, potential outcome, treatment effect heterogeneity, quantiles of individual treatment effects, randomization test

*Devin Caughey, Department of Political Science, MIT (e-mail: caughey@mit.edu). Allan Dafoe, International Politics of Artificial Intelligence, University of Oxford (e-mail: allandafoe@gmail.com). Xinran Li, Department of Statistics, University of Illinois at Urbana-Champaign (e-mail: xinranli@illinois.edu). Luke Miratrix, Graduate School of Education, Harvard University (e-mail: lmiratrix@g.harvard.edu).

Contents

1. Introduction

1.1. Randomization inference and sharp null hypotheses

Randomization inference (RI), also known as permutation inference, is a general statistical framework for making inferences about treatment effects. RI originated with Fisher (1935), who showed that if the treatment is randomly assigned to units, the hypothesis that no unit was affected by treatment—the "sharp null of no effects"—can be tested exactly, with no further assumptions, by comparing an observed test statistic with its distribution across alternative realizations of treatment assignment. More generally, the Fisher randomization test can be applied to any null hypotheses that is sharp in the sense that under it all potential outcomes are known from the observed data. Furthermore, by testing a sequence of hypotheses, randomization tests can also be used to create exact nonparametric confidence intervals (CIs) for treatment effects (Lehmann 1963). RI thus provides a unified framework of statistical inference that requires neither parametric assumptions about the data-generating distribution nor asymptotic approximations that can be unreliable in small samples (Rosenbaum 2002).

Nevertheless, RI has been subject to trenchant critiques. One line of criticism focuses on the sharp null hypothesis, which has long been dismissed as "uninteresting and academic" (Neyman 1935, 173). Gelman (2011), for example, argues that "the so-called Fisher exact test almost never makes sense, as it's a test of an uninteresting hypothesis of exactly zero effects (or, worse, effects that are nonzero but are identical across all units)." The crux of this critique is that the sharp null "does not accommodate heterogeneous responses to treatment," making it "a very restrictive null hypothesis" (Keele 2015, 330). In this view, rejecting such an implausible hypothesis conveys little information of substantive or scientific interest.

A related line of criticism is that using RI for interval estimation requires assumptions that are arguably as strong as those of its parametric and large-sample competitors. In particular, deriving interpretable CIs for treatment effects typically requires the assumption that effects are constant (i.e., invariant across units) or vary according to some known model (Rosenbaum 2002; Ho and Imai 2006; Bowers et al. 2013). In many settings, constant effects are implausible and the correct model for their variation is unknown. Without some version of the constant-effects assumption, however, RI as currently understood does not justify exact finite-sample inference regarding the *magnitude* of treatment effects.

Defenders of randomization inference have responded to these critiques in various ways. Some, while largely accepting the critiques of the sharp null, argue that RI is nevertheless useful for assessing whether treatment had any effect at all, as a preliminary step to determine whether further analysis is warranted (e.g., Imbens and Rubin 2015). An alternative proposal, advanced by Chung and Romano (2013), is to employ "studentized" test statistics that render permutation tests asymptotically valid under a weak null hypothesis (see, e.g., Ding and Dasgupta 2017; Wu and Ding 2018; Fogarty 2019; Cohen and Fogarty 2020).[1] Some scholars defend the

[1] The term *weak* is typically used to refer to the null hypothesis of no average effect, in contradistinction to the

constant-effects assumption more forthrightly, regarding it as a convenient approximation that is preferable to the shortcomings of parametric methods, such as their sensitivity to assumptions about tail behavior (Rosenbaum 2010a) or inability to account for complex treatment assignments (Ho and Imai 2006).

Though reasonable, all these defenses presume that randomization tests are exactly valid only as tests of a sharp null hypothesis. As such, they do not fully address the concerns of critics who regard sharp nulls as inherently "restrictive," "uninteresting," and "academic." We offer a more fundamental defense. As we show, many randomization tests of the sharp null are exactly valid under a corresponding hypothesis under which unit-level effects are bounded but otherwise hetereogenous. This result in turn provides the basis for CIs for the maximum (or minimum) individual effect and, by extension, for any quantile of the distribution of unit-level effects. These results substantially expand the applicability of RI and permit assessment of the substantive magnitude of treatment effects as well as their statistical significance.

1.2. A motivating example

To motivate our approach, consider Heller et al.'s (2010) experimental study of the effectiveness of professional development for elementary teachers. The study compared 164 teachers assigned to take a professional development course with 69 control subjects (for full details, see Section 8.2). The average gain score, based on tests on content knowledge before and after the courses, was much higher among the teachers who took the courses. A Stephenson rank-sum randomization test (see Section 3.3) yields a p-value near 0, and Lehmann-style test inversion yields a 90% CI of $[16.671, \infty)$ for a constant treatment effect.[2] These results strongly suggest that the professional development improves teacher's content knowledge, but their precise interpretation is less clear. The p-value indicates decisive rejection of the sharp null of no effects, but is rejecting this hypothesis really informative? And if a constant-effects assumption is not plausible, how should we interpret the CI?

Our paper sheds light on both of these questions. First, because the Stephenson rank-sum test is also valid under the null that all unit-level effects are bounded above at 0, the p-value reported above justifies rejection of the hypothesis that no teacher's content knowledge increased as a result of the course. Second, without a constant-effects assumption, the CI reported above can be interpreted as a confidence statement about the *maximum* effect—specifically, that the hypothesis that no effect was as large as 16.670 can be rejected at a significance level of 0.1.

Moreover, we can generalize these results to obtain simultaneously valid CIs for all quantiles of the effect distribution. These inferences are visualized in Figure 1, which reports the simul-

stronger hypothesis of no effect whatsoever (e.g., Freedman et al. 1997, A-32). We use *weak* more generally, to refer to any hypothesis that stipulates the value of some function of the unit-level treatment effects (e.g., their average, maximum or quantiles) but otherwise allows for arbitrary effect heterogeneity. A weak null is a "composite" hypothesis in the sense that it encompasses multiple configurations of potential outcomes rather than a single one as a sharp null does.

[2] The p-value is approximated by Monte Carlo with 10^6 simulated assignments. Consequently, the standard error of this Monte Carlo approximation is at most 5×10^{-4}.

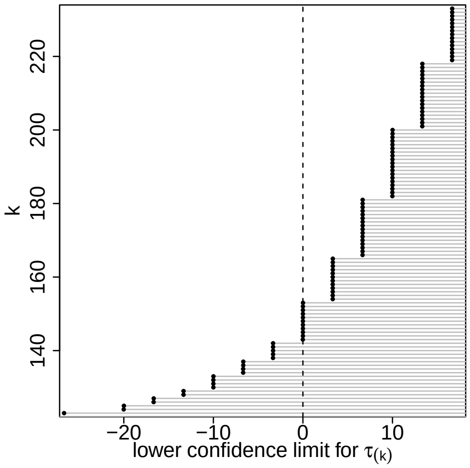


Figure 1: The 90% simultaneous lower confidence limits for all quantiles of individual effects in the study of the effectiveness of professional development for elementary teachers. The uninformative lower confidence limits of $-\infty$ for individual effects at lower ranks are omitted.

taneous 90% one-sided confidence intervals for all informative quantiles of individual effects. Specifically, with $\tau_{(k)}$ denoting kth-ranked effect, the horizontal lines in Figure 1 represents the one-sided confidence intervals for $\tau_{(k)}$'s with $123 \leq k \leq 233$, i.e., the largest 111 individual effects. The 149th and 166th largest individual effects, for example, have 90% CIs of $(0, \infty)$ and $[6.660, \infty)$, respectively. The lower confidence limits for the remaining smallest 122 effects are all negative infinity, which being uninformative are omitted from the figure.

In addition, the lower confidence limit for the number of units with effects greater than c, denoted by $n(c)$, is equivalently the number of quantiles of individual effects $\tau_{(k)}$'s whose confidence intervals do not cover c. We can thus read off these confidence intervals using Figure 1. Consider the dashed vertical line of $c = 0$ as an example. Because the line of $c = 0$ only intersects the confidence intervals for $k \leq 148$, we know we have at least $233 - 148 = 85$ units with significant effects, giving a final interval for $n(0)$ of $[85, 233]$. By the same logic, a 90% confidence interval for the number of units with effects larger than 6, $n(6)$, is $[68, 233]$. Equivalently, we can say we are 90% confident that at least $85/233 = 36.5\%$ of teachers benefited from the professional development, and that for at least $68/233 = 29.2\%$ of them the effect of the course was at least 6.660.

Reversing the direction of the test yields a 90% CI for the minimum effect of $(-\infty, 23.329]$. Since the upper bound of this interval is larger than the lower bound of the CI for the maximum effect, we cannot reject the hypothesis of a constant treatment effect. In sum, even without adding assumptions or changing test statistics, our reinterpretation and generalization of RI justify inferences far richer than mere rejection of the sharp null of no effects.

1.3. Our contribution

The inferences illustrated in Section 1.2 are grounded in a novel set of theoretical results. In particular, in this paper we prove the following:

(i) For any randomization test that employs a test statistic with one of several properties, one-sided rejection of the sharp null hypothesis that the treatment τ_i equals some constant δ_i for each unit i also implies rejection of any null hypothesis under which each τ_i is bounded on one side by its corresponding δ_i.

(ii) Test statistics with the requisite properties include the difference of means, the Wilcoxon rank sum, and many other commonly used statistics.

(iii) For tests in this class, inverting a sequence of tests provides confidence intervals for the maximum (or minimum) individual effect.

(iv) For rank-based members of this class, when treatment assignments are exchangeable across all units, this procedure can be extended to yield simultaneously valid confidence intervals for all quantiles of the treatment-effect distribution (and, by extension, for the proportion of units with effects larger/smaller than a given threshold).

(v) Confidence intervals for the range of unit-level effects can be obtained by combining two one-sided randomization tests, providing an exact test of effect heterogeneity.

These results have important implications for statistical practice.

First, "bounded" hypotheses are often of substantive interest in themselves. Unlike sharp hypotheses, which are concentrated at a particular point in the parameter space, bounded hypothesis cover a range of treatment effects. Consequently, their prior plausibility is greater and their rejection is thus more informative. In addition, there are a number of theoretically or methodologically important special cases of bounded hypotheses. In economics, for example, a change is considered a "Pareto improvement" if it makes at least one person better off while hurting no one (Mishan 1982, 34). Consequently, the claim that an intervention was Pareto improving can be assessed by testing the bounded null hypothesis that all effects were greater than or equal to zero. Similarly, instrumental-variable estimation of causal effects is typically conducted under a monotonicity assumption that the instrument has non-negative or non-positive effects on the treatment (Angrist et al. 1996). In short, there are many situations where testing (and possibly rejecting) a bounded null hypothesis is theoretically or practically important.

Second, these results provide a basis for inferences regarding the distribution of treatment effects across units. Specifically, they permit interval estimation of treatment effect quantiles. Existing nonparametric methods for estimating causal effects focus overwhelmingly on average effects of various kinds. Although averages are often the best summary statistic, quantiles characterize the effects of treatment more completely and robustly, especially when effects are highly heterogeneous. For example, quantiles are more useful in situations where treatment has a positive effect on average but the question of interest is how many subjects are harmed by it.

Although there are existing RI methods for ”quantile treatment effects” (the treated–control difference in *outcome* quantiles; e.g., Cattaneo et al. 2015), to our knowledge ours is the first aimed at quantiles of the distribution of *effects*[3] In addition to being interesting in themselves, quantile CIs from two one-sided tests can be combined to yield tests and CIs for the range of treatment effects.

In sum, our reinterpretation of randomization inference provides the basis for several novel theoretical results as well as practical methods useful to applied statisticians. Significantly, these advances are a ”free lunch” in the sense that they require no additional assumptions. Users of RI can continue to use the same procedures while interpreting them in richer ways, or they can use our open-source R package RIQITE to supplement their analyses with quantile CIs and other extensions. RIQITE is publicly available at `https://github.com/li-xinran/RIQITE`.[4]

The remainder of this paper is organized as follows. Section 2 formally introduces our framework. Section 3 discusses the properties of test statistics to which our results apply and provides examples. Section 4 demonstrates the validity of randomization tests under bounded null hypotheses and explains how this justifies CIs for the maximum individual treatment effect. Section 5 generalizes these results to CIs for treatment effect quantiles. Section 6 shows how CIs for the range of effects can be derived from two one-sided randomization tests. Section 7 conducts simulation studies for the performance of these procedures under various conditions. Section 8 applies the methods to two applications: testing monotonicity of an instrumental variable and evaluating the effectiveness of an experimental professional development program. The paper concludes with a discussion of the broader implications of our reconceptualization and generalization of RI.

2. Framework, notation, and the Fisher randomization test

2.1. Potential outcomes, treatment effects and treatment assignment

We consider a randomized experiment on n units with two treatment arms. Using the potential outcome framework (Neyman 1923; Rubin 1974), we use $Y_i(1)$ and $Y_i(0)$ to denote the potential outcomes under treatment and control, respectively, for units $i = 1, \dots, n$. We use $\boldsymbol{Y}(1) = (Y_1(1), Y_2(1), \dots, Y_n(1))^\top$ and $\boldsymbol{Y}(0) = (Y_1(0), Y_2(0), \dots, Y_n(0))^\top$ to denote the treatment and control potential outcome vectors for all units. Let $\tau_i = Y_i(1) - Y_i(0)$ be the individual treatment effect for unit i, and $\boldsymbol{\tau} = (\tau_1, \tau_2, \dots, \tau_n)^\top$ be the vector of individual treatment effects.[5] Let Z_i be the treatment assignment for unit i, where Z_i equals 1 if the unit receives the active

[3] For the distinction between *differences in outcome distributions* (e.g., average or quantile treatment effects) and *distributions of outcome differences* (e.g., the distribution of individual effects), see Manski (2009, 157–8).

[4] The package's Github page includes installation instructions, detailed explanations of the main functions in R documentation, and a simple illustrating example. In addition, the supplementary materials for this paper contain a replication of the three real data analyses in the paper using RIQITE.

[5] As a side note, we focus on general outcomes with the treatment effect in the difference scale, which may not be most appropriate in some applications; see, e.g., Edwards (1963) and Xie et al. (2008).

treatment and zero otherwise, and $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)^\top$ be the treatment assignment vector for all units. For each unit i, the observed outcome is one of its two potential outcomes, depending on the treatment assignment Z_i. Specifically, $Y_i \equiv Y_i(Z_i) = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$. Let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ be the observed outcome vector for all units. For descriptive convenience, for any treatment assignment vector $\boldsymbol{z} \in \{0,1\}^n$, we define $\boldsymbol{Y}(\boldsymbol{z}) = \boldsymbol{z} \circ \boldsymbol{Y}(1) + (\boldsymbol{1} - \boldsymbol{z}) \circ \boldsymbol{Y}(0)$ to denote the corresponding observed outcome vector, where $\circ$ stands for the element-wise multiplication. We can then write the observed outcome $\boldsymbol{Y}$ as $\boldsymbol{Y} = \boldsymbol{Y}(\boldsymbol{Z}) = \boldsymbol{Z} \circ \boldsymbol{Y}(1) + (1 - \boldsymbol{Z}) \circ \boldsymbol{Y}(0)$.

In this paper, we conduct randomization inference where all potential outcomes are viewed as fixed constants, and use randomization of the treatment assignments as the "reasoned basis" for inference (Fisher 1935). This is equivalent to conducting inference conditional on all the potential outcomes, and thus requires no model or distributional assumptions on the potential outcomes. Consequently, the randomness of the observed data comes solely from the random treatment assignment. Therefore, the distribution of the treatment assignment $\boldsymbol{Z}$, also called the treatment assignment mechanism, plays an important role in governing statistical inference. We use $\mathcal{Z} \subset \{0,1\}^n$ to denote the set of all possible treatment assignments for the n units, and characterize the treatment assignment mechanism by the probability mass function $\Pr(\boldsymbol{Z} = \boldsymbol{z})$ for all $\boldsymbol{z} \in \mathcal{Z}$. Because $\boldsymbol{Z}$ is used to denote the observed treatment assignment, to avoid confusion, we sometimes use $\boldsymbol{A}$ to denote a generic random treatment assignment vector following the same distribution as $\boldsymbol{Z}$, i.e., $\Pr(\boldsymbol{A} = \boldsymbol{z}) = \Pr(\boldsymbol{Z} = \boldsymbol{z})$ for all $\boldsymbol{z} \in \mathcal{Z}$. One class of treatment assignment mechanism that will receive special attention is the *exchangeable* treatment assignment, formally defined as follows.

Definition 1. A treatment assignment is said to be exchangeable if the distribution of treatment assignment vector $\boldsymbol{Z}$ is invariant under permutation of its coordinates, i.e., $(Z_1, Z_2, \ldots, Z_n) \sim (Z_{\pi(1)}, Z_{\pi(2)}, \ldots, Z_{\pi(n)})$ for any permutation $\pi(\cdot)$ of $\{1, 2, \ldots, n\}$.

Popular treatment assignment mechanisms satisfying Definition 1 include *Bernoulli randomized experiment* (BRE) and *completely randomized experiment* (CRE). Specifically, under a BRE, the treatment assignments Z_i's are independent and identically distributed (i.i.d.) Bernoulli random variables with probability p being 1, for some $p \in (0,1)$. Under a CRE, m units will be randomly assigned to treatment, and the remaining $n - m$ units will be assigned to control, where m and $n - m$ are fixed positive integer.

We use Unif$[0,1]$ to denote a uniform random variable on $[0,1]$. We introduce $\preccurlyeq$ to denote element-wise inequality between two vectors: for any two vectors $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ of the same dimension, $\boldsymbol{\xi} \preccurlyeq \boldsymbol{\eta}$ if and only if each coordinate of $\boldsymbol{\xi}$ is less than or equal to the corresponding coordinate of $\boldsymbol{\eta}$.

2.2. Sharp null hypotheses and imputation of potential outcomes

Fisher (1935) proposed to test the null hypothesis that the treatment has no effect for any unit, that is $\boldsymbol{Y}(1) = \boldsymbol{Y}(0)$ or equivalently $\boldsymbol{\tau} = \boldsymbol{0}$. Such a null hypothesis is called a sharp null hypothesis, under which all missing potential outcomes can be imputed from the observed data. Specifically, under Fisher's null of no effect, we have $Y_i(1) = Y_i(0) = Y_i$ for all units $i = 1, \dots, n$. After imputing all the potential outcomes, we can know the null distribution of any test statistic exactly, which further provides an exact p-value. Such a testing procedure is often called the Fisher randomization test, and the resulting p-value is called the randomization p-value. The Fisher randomization test also works for general sharp null hypotheses, as described in detail below.

Consider a general sharp null hypothesis where the individual treatment effect for unit i is δ_i for $i = 1, \dots, n$, where $\boldsymbol{\delta} = (\delta_1, \delta_2, \dots, \delta_n)^\top \in \mathbb{R}^n$ is a predetermined constant vector. That is

$$H_{\boldsymbol{\delta}} : \boldsymbol{\tau} = \boldsymbol{\delta}. \tag{1}$$

Under the null $H_{\boldsymbol{\delta}}$ in (1), we are able to impute all the potential outcomes. Specifically, given the observed data $\boldsymbol{Z}$ and $\boldsymbol{Y}$, for the null $H_{\boldsymbol{\delta}}$, the imputed treatment and control potential outcome vectors for all units are, respectively,

$$\begin{aligned} \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(1) &= \boldsymbol{Y} + (\boldsymbol{1} - \boldsymbol{Z}) \circ \boldsymbol{\delta} = \boldsymbol{Z} \circ \boldsymbol{Y}(1) + (\boldsymbol{1} - \boldsymbol{Z}) \circ \{\boldsymbol{Y}(0) + \boldsymbol{\delta}\}, \\ \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0) &= \boldsymbol{Y} - \boldsymbol{Z} \circ \boldsymbol{\delta} = \boldsymbol{Z} \circ \{\boldsymbol{Y}(1) - \boldsymbol{\delta}\} + (\boldsymbol{1} - \boldsymbol{Z}) \circ \boldsymbol{Y}(0), \end{aligned} \tag{2}$$

where we use the subscripts $\boldsymbol{Z}$ and $\boldsymbol{\delta}$ to emphasize that the imputed potential outcomes are deterministic functions of the observed treatment assignment, the null hypothesis of interest and the true potential outcomes. Importantly, the *imputed* potential outcomes in (2) are generally different from the *true* potential outcomes, and they become the same if and only if the sharp null $H_{\boldsymbol{\delta}}$ in (1) is true.

2.3. Fisher randomization test

There are at least two popular approaches for conducting Fisher randomization tests for general sharp null hypothesis $H_{\boldsymbol{\delta}}$ in (1), represented by the textbooks of Rosenbaum (2002) and Imbens and Rubin (2015), respectively. The main distinction between these approaches lies in the choice of test statistic. For conciseness, we will focus on the approach from Rosenbaum (2002) in the main paper, and relegate the discussion on the other approach from Imbens and Rubin (2015) to the supplementary materials.

Let $t(\cdot, \cdot) : \mathcal{Z} \times \mathbb{R}^n \to \mathbb{R}$ denote a generic function with two arguments: the first a treatment assignment vector $\boldsymbol{z} \in \mathcal{Z}$, and the second an outcome vector $\boldsymbol{y} \in \mathbb{R}^n$. Following Rosenbaum (2002), we consider testing the sharp null $H_{\boldsymbol{\delta}}$ in (1) using test statistic of the form $t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(\boldsymbol{0}))$, which depends on the observed treatment assignment $\boldsymbol{Z}$ and the imputed control potential outcomes $\boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)$ in (2). Often the test statistic $t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0))$ compares the imputed control potential outcomes of treated units to those of control units (e.g., the average difference between the treated outcomes adjusted by corresponding δ_i's and the control outcomes). Under $H_{\boldsymbol{\delta}}$, the imputed control potential outcome $\boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)$ is always the same as the true one $\boldsymbol{Y}(0)$, and thus, for any treatment assignment vector $\boldsymbol{a} \in \mathcal{Z}$, the corresponding value of the test statistic would then be $t(\boldsymbol{a}, \boldsymbol{Y}_{\boldsymbol{a},\boldsymbol{\delta}}(0)) = t(\boldsymbol{a}, \boldsymbol{Y}(0)) = t(\boldsymbol{a}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0))$. This means we can examine what t should be for any possible value of the treatment assignment. Therefore, for the null $H_{\boldsymbol{\delta}}$, the imputed randomization distribution of the test statistic has the following tail probability:

$$G_{\boldsymbol{Z},\boldsymbol{\delta}}(c) \equiv \Pr\left\{t(\boldsymbol{A}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)) \geq c\right\} = \sum_{\boldsymbol{a}\in\mathcal{Z}} \Pr(\boldsymbol{A} = \boldsymbol{a}) \mathbb{1}\left\{t(\boldsymbol{a}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)) \geq c\right\}, \qquad (c \in \mathbb{R}) \tag{3}$$

and the corresponding randomization p-value is the tail probability evaluated at the observed value of the test statistic:

$$p_{\boldsymbol{Z},\boldsymbol{\delta}} \equiv G_{\boldsymbol{Z},\boldsymbol{\delta}}\left\{t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0))\right\} = \sum_{\boldsymbol{a}\in\mathcal{Z}} \Pr(\boldsymbol{A} = \boldsymbol{a}) \mathbb{1}\left\{t(\boldsymbol{a}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)) \geq t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0))\right\}. \tag{4}$$

When $H_{\boldsymbol{\delta}}$ is true (i.e., $\boldsymbol{\tau} = \boldsymbol{\delta}$), the imputed randomization distribution $G_{\boldsymbol{Z},\boldsymbol{\delta}}(\cdot)$ in (3) reduces to $G_{\boldsymbol{Z},\boldsymbol{\delta}}(c) = \Pr\{t(\boldsymbol{A}, \boldsymbol{Y}(0)) \geq c\}$, and the randomization p-value $p_{\boldsymbol{Z},\boldsymbol{\delta}} = p_{\boldsymbol{Z},\boldsymbol{\tau}}$ is stochastically larger than or equal to Unif$[0, 1]$ (the difference is due solely to the discrete nature of the p-value distribution). That is, $p_{\boldsymbol{Z},\boldsymbol{\delta}}$ in (4) is a valid p-value for testing the sharp null $H_{\boldsymbol{\delta}}$.

In contrast to typical descriptions of randomization inference, we do not simplify the imputed potential outcomes in (3) and (4) to the true ones, because later we will investigate the property of the randomization p-value even when the sharp null hypothesis fails. We emphasize that both $G_{\boldsymbol{Z},\boldsymbol{\delta}}(\cdot)$ and $p_{\boldsymbol{Z},\boldsymbol{\delta}}$ are deterministic functions of the treatment assignment $\boldsymbol{Z}$, the null hypothesis of interest $\boldsymbol{\delta}$, the true potential outcomes $(\boldsymbol{Y}(1), \boldsymbol{Y}(0))$, and the treatment assignment mechanism, where the latter two are fixed and the dependence on them is suppressed. Moreover, the randomness in $G_{\boldsymbol{Z},\boldsymbol{\delta}}(\cdot)$ and $p_{\boldsymbol{Z},\boldsymbol{\delta}}$ comes solely from the random treatment assignment $\boldsymbol{Z}$.

Remark 1. Inspired by the previous discussion, it should not be surprising that we can also use test statistics of the form $t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(1))$, which involves the imputed treatment potential outcomes instead of imputed control ones. This can be achieved by switching the labels of treatment and control and changing the signs of the outcomes.

As a side note, the Fisher randomization tests in Imbens and Rubin (2015) use test statistics of form $t(\boldsymbol{Z}, \boldsymbol{Y})$, which often compare the observed outcomes of treated units to those of control units (e.g., the difference in outcome means between treatment and control groups). Analogous to Section 4, such a randomization test can also be valid for testing bounded null hypotheses; see the supplementary materials for details. However, our generalization for inferring quantiles of individual effects (as in Section 5) relies crucially on the use of test statistics of form $t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0))$. This also explains why we focus on the randomization p-value of form (4) in the main paper.

3. Test statistics: properties and examples

3.1. Three properties of test statistics

An important component of the Fisher randomization test described in Section 2.3 is the test statistic. As discussed shortly, test statistics satisfying three certain properties can provide broader justification for the Fisher randomization test, beyond its validity only for sharp null hypotheses.

The first property is called *effect increasing* (a term borrowed from Rosenbaum 2002, who studies the unbiasedness of randomization tests). Intuitively, an effect increasing statistic $t(\boldsymbol{z}, \boldsymbol{y})$, viewed as a function of the outcome vector $\boldsymbol{y}$ with $\boldsymbol{z}$ fixed, is increasing[6] in those y_i's with $z_i = 1$ and decreasing in those y_i's with $z_i = 0$. We formally define it as follows.

Definition 2. A statistic $t(\cdot, \cdot)$ is said to be effect increasing, if $t(\boldsymbol{z}, \boldsymbol{y} + \boldsymbol{z} \circ \boldsymbol{\eta} + (1 - \boldsymbol{z}) \circ \boldsymbol{\xi}) \geq t(\boldsymbol{z}, \boldsymbol{y})$ for any $\boldsymbol{z} \in \mathcal{Z}$ and $\boldsymbol{y}, \boldsymbol{\eta}, \boldsymbol{\xi} \in \mathbb{R}^n$ with $\boldsymbol{\eta} \succcurlyeq 0 \succcurlyeq \boldsymbol{\xi}$.

The second property is called *differential increasing*. Intuitively, a differential increasing statistic $t(\boldsymbol{z}, \boldsymbol{y})$ is one such that, if the outcomes for a subset of units are increased, then the change of the statistic is maximized when it happens that this subset of units receive treatment. We formally define it as follows.

Definition 3. A statistic $t(\cdot, \cdot)$ is said to be differential increasing, if $t(\boldsymbol{z}, \boldsymbol{y} + \boldsymbol{a} \circ \boldsymbol{\eta}) - t(\boldsymbol{z}, \boldsymbol{y}) \leq t(\boldsymbol{a}, \boldsymbol{y} + \boldsymbol{a} \circ \boldsymbol{\eta}) - t(\boldsymbol{a}, \boldsymbol{y})$ for any $\boldsymbol{z}, \boldsymbol{a} \in \mathcal{Z}$ and $\boldsymbol{y}, \boldsymbol{\eta} \in \mathbb{R}^n$ with $\boldsymbol{\eta} \succcurlyeq 0$.

The third property is called *distribution free*; see also Rosenbaum (2007b). Different from the previous two properties, this property depends not only on the test statistic $t(\cdot, \cdot)$, but also on the treatment assignment mechanism. In particular, for a distribution free test statistic $t(\cdot, \cdot)$, the distribution of $t(\boldsymbol{Z}, \boldsymbol{y})$ does not depend on the value of $\boldsymbol{y} \in \mathbb{R}^n$. We formally define it as follows.

Definition 4. A statistic $t(\cdot, \cdot)$ is said to be distribution free if, for any $\boldsymbol{y}, \boldsymbol{y}' \in \mathbb{R}^n$, $t(\boldsymbol{Z}, \boldsymbol{y})$ and $t(\boldsymbol{Z}, \boldsymbol{y}')$ follow the same distribution, where $\boldsymbol{Z}$ follows the treatment assignment mechanism.

For the above three properties, one does not necessarily imply the other; see the supplementary materials for more details. However, many commonly used test statistics are both effect increasing and differential increasing, and many rank-based test statistics (with appropriate handling of ties) are also distribution free when the treatment assignment is exchangeable (e.g., a Bernoulli or completely randomized experiment, as defined in Section 2.1); see the next subsection for more details.

[6] Throughout the paper, an increasing function refers to a nondecreasing function, not a strictly increasing function. Similarly, a decreasing function refers to a nonincreasing function.

3.2. Examples of test statistics

We will discuss two classes of test statistics, which cover many test statistics commonly used for randomization tests. In particular, we will provide sufficient conditions for these test statistics to satisfy the properties introduced in Section 3.1.

The first class of test statistics has the following form:

$$t_1(\boldsymbol{z}, \boldsymbol{y}) = \sum_{i=1}^{n} z_i \psi_{1i}(y_i) - \sum_{i=1}^{n} (1 - z_i) \psi_{0i}(y_i), \tag{5}$$

where the $\psi_{1i}(\cdot)$'s and $\psi_{0i}(\cdot)$'s are constant functions from $\mathbb{R}$ to $\mathbb{R}$ but can depend on the treatment assignment mechanism. For example, in a general randomized experiment, if we choose $\phi_{1i}(y) = y/\{n \Pr(Z_i = 1)\}$ and $\phi_{0i}(y) = y/\{n \Pr(Z_i = 0)\}$ for all $1 \leq i \leq n$, then $t_1(\boldsymbol{Z}, \boldsymbol{Y})$ reduces to the usual Horvitz–Thompson estimator for the average treatment effect. In a special case of a CRE with m units receiving treatment, $t_1(\boldsymbol{Z}, \boldsymbol{Y}) = m^{-1} \sum_{i=1}^{n} Z_i Y_i - (n - m)^{-1} \sum_{i=1}^{n} (1 - Z_i) Y_i$ further reduces to the difference-in-means estimator. The following proposition shows that the statistic in (5) satisfies Definitions 2 and 3 as long as $\psi_{1i}(\cdot)$'s and $\psi_{0i}(\cdot)$'s are increasing functions.

Proposition 1. Statistics $t_1(\cdot, \cdot)$ of the form in (5) are both effect increasing and differential increasing if $\psi_{1i}(\cdot)$ and $\psi_{0i}(\cdot)$ are monotone increasing functions for all $1 \leq i \leq n$.

The second class of test statistics depends only on the ranks of the outcomes. For any vector $\boldsymbol{y} \in \mathbb{R}^n$ and for each $1 \leq i \leq n$, we use $\mathrm{r}_i(\boldsymbol{y})$ to denote the rank of the ith coordinate of $\boldsymbol{y}$, where larger rank corresponds to larger outcome value. We briefly defer the exact definition of rank in the presence of ties. The second class of test statistics has the following form:

$$t_2(\boldsymbol{z}, \boldsymbol{y}) = \sum_{i=1}^{n} z_i \phi(\mathrm{r}_i(\boldsymbol{y})), \tag{6}$$

where $\phi(\cdot)$ is a constant function from $\mathbb{R}$ to $\mathbb{R}$. For example, if we choose $\phi(r) = r$ to be the identity function, then $t_2(\boldsymbol{Z}, \boldsymbol{Y})$ reduces to the Wilcoxon rank sum statistic. The classes of statistics satisfying the three properties in Section 3.1 are overlapping but not nested: In general, the statistic $t_1(\cdot, \cdot)$ in (5) is not distribution free, and the statistic $t_2(\cdot, \cdot)$ in (6) is not differential increasing; see the supplementary materials for more details.

Now we discuss the subtle issue of defining ranks when there exist ties. There are multiple ways to define ranks for ties, see, e.g., the R documentation for the function rank (R Core Team 2013). Consider an arbitrary outcome vector $\boldsymbol{y} \in \mathbb{R}^n$. The method "random" puts equal values of y_i's in a random order, the methods "first" and "last" rank equal values of y_i's based on their indices in an increasing or decreasing way[7], and the method "average" replaces their ranks by the corresponding average. The first three methods will always produce ranks from 1 to n,

[7] For example, if we rank the coordinates of $\boldsymbol{y}$ using the "first" method, then $\mathrm{r}_i(\boldsymbol{y}) < \mathrm{r}_j(\boldsymbol{y})$ if and only if (a) $y_i < y_j$ or (b) $y_i = y_j$ and $i < j$.

which is important for the distribution free property in Definition 4, while the ranks from the last generally depend on the value of $\boldsymbol{y}$.

The following proposition gives sufficient conditions for the test statistic in (6) to satisfy Definitions 2 and 4 introduced in Section 3.1.

Proposition 2. (a) The statistic $t_2(\cdot,\cdot)$ in (6) is effect increasing if $\phi(\cdot)$ is a monotone increasing function and the “first” or “last” method is used for ties. (b) The statistic $t_2(\cdot,\cdot)$ is distribution free if the “first” or “last” method is used for ties and the treatment assignment $\boldsymbol{Z}$ is exchangeable as in Definition 1 and independent of the ordering of units.

In practice, we can implement the “random” method for ties by first randomly permuting the coordinates of $\boldsymbol{y}$ and then ranking equal outcomes using the “first” methods. If $\boldsymbol{Z}$ is exchangeable and we use the “random” method for ties, then after we randomly permute the ordering of units, $\boldsymbol{Z}$ must be independent of the ordering, and the ties are ranked based on the “first” method. From Proposition 2, the statistic in (6) will be distribution free. We emphasize that using the “random” method for ties is crucial for the distribution free property. First, a defined method for ties is important even for continuous outcomes, because some outcomes will become the same when we later invert tests for, e.g., a sequence of constant treatment effects. Second, it is important to rank equal outcomes randomly, since in practice a received data set may be ordered, e.g., with all treated units ahead of control ones.

For the usual “average” method for ties, whether the statistic in (6) is effect increasing depends on the values of $\phi(\cdot)$ for tied ranks. If we define the value of $\phi(\cdot)$ for tied ranks as the average value of $\phi(\cdot)$ evaluated at those ranks under the “first” or “last” method, then the statistic (6) is effect increasing. This follows directly from Proposition 2 by noting that the resulting value of the statistic is essentially the average of it under the “first” method of ties over all possible permutations of the ordering. From the above, we can also know that the classical Wilcoxon rank sum statistic with “average” method for ties is effect increasing.

3.3. Stephenson rank sum statistics

Despite the generality of the test statistics discussed before, in this subsection we focus on a special class of test statistics developed by Stephenson and Ghosh (1985), because of its often superior power discussed later. Stephenson rank sum statistics can be defined as the number of subsets of size s in the data in which the largest response is in the treated group. This can be represented as a two-sample statistic of form (6) with a rank score function of

$$\phi(r) = \binom{r-1}{s-1} \quad \text{if } r \geq s, \quad \text{and} \quad \phi(r) = 0 \quad \text{otherwise,}$$

for some fixed integer $s \geq 2$. The Stephenson rank sum statistic with $s = 2$ is almost equivalent to the Wilcoxon rank sum statistic.[8] However, as s increases beyond 2, the Stephenson ranks

[8] The Stephenson ranks with $s = 2$ are each one less than the corresponding Wilcoxon ranks, leading to almost identical behavior (Rosenbaum 2007a, 1168).

place more and more weight on the larger responses.

Rosenbaum (2007a) proposed to use Stephenson ranks to detect uncommon-but-dramatic responses to treatment. Intuitively, this is because as the subset size s increases, it becomes increasingly likely that the largest response in a given subset will be one with an unusually large treatment effect. Thus, compared to the difference-in-means and the Wilcoxon rank sum, whose power is greatest against a constant location shift, the Stephenson rank sum has greater power against alternatives under which effects are heterogeneous and a few are highly positive. This advantage of Stephenson rank sum statistics is particularly relevant for our proposed theory and methods. We defer detailed discussion of this to later simulation studies.

4. Broader justification for Fisher randomization test

4.1. Validity of randomization p-values for bounded nulls

As discussed in Section 2.3, the randomization p-value $p_{\boldsymbol{Z},\boldsymbol{\delta}}$ is always valid for testing the sharp null $H_{\boldsymbol{\delta}}$ in (1) given any test statistic $t(\cdot,\cdot)$. Our question is, can the randomization p-value $p_{\boldsymbol{Z},\boldsymbol{\delta}}$ also be valid for testing a weak null hypothesis that does not fully specify all the individual treatment effects?

We will demonstrate shortly that, under certain intuitive conditions on the test statistic, the randomization p-value $p_{\boldsymbol{Z},\boldsymbol{\delta}}$ for testing the sharp null hypothesis $H_{\boldsymbol{\delta}}$ is also valid for testing a bounded null, which states that each individual treatment effect is less than or equal to the corresponding coordinate of $\boldsymbol{\delta}$. We formally introduce this bounded null hypothesis as follows:

$$H_{\preccurlyeq\boldsymbol{\delta}} : \boldsymbol{\tau} \preccurlyeq \boldsymbol{\delta}, \tag{7}$$

where $\boldsymbol{\delta}$ is a constant vector in $\mathbb{R}^n$. Importantly, the null $H_{\preccurlyeq\boldsymbol{\delta}}$ in (7) is composite, under which the exact distribution of the test statistic is generally unknown. The bounded null hypotheses can often be of interest in practice, and the choice of $\boldsymbol{\delta}$ in (7) depends on the application of interest. For example, we can choose $\boldsymbol{\delta} = \boldsymbol{0}$ if we are interested in whether the treatment has a positive effect for any unit. This is related to pareto efficiency and monotonicity assumption in instrumental variable analysis; see Section 8.1. If we are also interested in the magnitude of the effects, we can choose $\boldsymbol{\delta} = c\boldsymbol{1}$, under which the null hypothesis in (7) assumes that all individual treatment effects are at most c; see Section 4.3.

The following theorem shows that the original Fisher randomization test, designed only for testing sharp null hypotheses, can also be valid for testing certain bounded null hypotheses.

Theorem 1. (a) If the test statistic $t(\cdot,\cdot)$ is either differential increasing or effect increasing, then for any constant $\boldsymbol{\delta} \in \mathbb{R}^n$, the corresponding randomization p-value $p_{\boldsymbol{Z},\boldsymbol{\delta}}$ in (4) for the sharp null $H_{\boldsymbol{\delta}}$ in (1) is also valid for testing the bounded null $H_{\preccurlyeq\boldsymbol{\delta}}$ in (7), i.e., under $H_{\preccurlyeq\boldsymbol{\delta}}$, $\Pr(p_{\boldsymbol{Z},\boldsymbol{\delta}} \le \alpha) \le \alpha$ for any $\alpha \in [0,1]$. (b) If the test statistic $t(\cdot,\cdot)$ is differential increasing, or it is both effect increasing and distribution free, then for any possible assignment $\boldsymbol{z} \in \mathcal{Z}$, the corresponding randomization

p-value $p_{\boldsymbol{z},\boldsymbol{\delta}}$ in (4), viewed as a function of $\boldsymbol{\delta} \in \mathbb{R}^n$, is monotone increasing, i.e., $p_{\boldsymbol{z},\boldsymbol{\delta}} \leq p_{\boldsymbol{z},\overline{\boldsymbol{\delta}}}$ for any $\boldsymbol{\delta} \preccurlyeq \overline{\boldsymbol{\delta}}$.

From Theorem 1(a), when the test statistic satisfies certain properties, the rejection of $H_{\boldsymbol{\delta}}$ also implies that there exists at least one unit i whose treatment effect is larger than δ_i. For the usual null $H_{c\mathbf{1}}$ of additive treatment effect c for some $c \in \mathbb{R}$, the rejection of $H_{c\mathbf{1}}$ then implies that there exists at least one unit whose treatment effect is larger than c. Furthermore, Theorem 1(a) also holds for general assignment mechanisms beyond BRE and CRE, such as blocking (Miratrix et al. 2013) or rerandomization (Morgan and Rubin 2012; Li et al. 2018).

Theorem 1(b) presents a stronger conclusion than (a), but with stronger condition on the test statistic. Specifically, for any $\boldsymbol{\delta} \in \mathbb{R}^n$, if the bounded null $H_{\preccurlyeq\boldsymbol{\delta}}$ holds (i.e., $\boldsymbol{\tau} \preccurlyeq \boldsymbol{\delta}$), then Theorem 1(b) implies that the randomization p-value $p_{\boldsymbol{Z},\boldsymbol{\delta}} \geq p_{\boldsymbol{Z},\boldsymbol{\tau}}$ and is thus stochastically larger than or equal to Unif$[0,1]$. More importantly, Theorem 1(b) is useful for constructing confidence sets for the true treatment effect $\boldsymbol{\tau}$. In principle, one could invert randomization tests for all possible sharp null hypotheses $H_{\boldsymbol{\delta}}$'s to get confidence sets for the true treatment effect $\boldsymbol{\tau}$. However, enumerating all possible sharp null hypotheses is generally computationally intractable, except in the cases of binary or discrete outcomes (see Rigdon and Hudgens 2015). For discrete outcomes where enumeration is possible, Theorem 1(b) can help reduce the number of enumerations (Li and Ding 2016).

For general outcomes, Theorem 1(b) can help provide meaningful confidence sets for $\boldsymbol{\tau}$ under practical computational constraints. For any $\alpha \in (0,1)$, the $1-\alpha$ confidence set for $\boldsymbol{\tau}$ by inverting randomization tests has the following equivalent forms:

$$\{\boldsymbol{\delta} : p_{\boldsymbol{Z},\boldsymbol{\delta}} > \alpha, \boldsymbol{\delta} \in \mathbb{R}^n\} = \{\boldsymbol{\delta} : p_{\boldsymbol{Z},\boldsymbol{\delta}} \leq \alpha, \boldsymbol{\delta} \in \mathbb{R}^n\}^{\complement} = \bigcap_{\boldsymbol{\delta}: p_{\boldsymbol{Z},\boldsymbol{\delta}} \leq \alpha, \boldsymbol{\delta} \in \mathbb{R}^n} \{\boldsymbol{\eta} : \boldsymbol{\eta} \preccurlyeq \boldsymbol{\delta}, \boldsymbol{\eta} \in \mathbb{R}^n\}^{\complement}. \tag{8}$$

Our confidence set consists of all points $\boldsymbol{\eta}$ not in the "shadow" of any given hypothesis that can be rejected. Therefore, for any vector $\boldsymbol{\delta}$ with a corresponding randomization p-value less than or equal to α, the $1-\alpha$ confidence set (8) must be a subset of $\{\boldsymbol{\eta} : \boldsymbol{\eta} \preccurlyeq \boldsymbol{\delta}, \boldsymbol{\eta} \in \mathbb{R}^n\}^{\complement}$, a set in which no element is uniformly bounded by $\boldsymbol{\delta}$.

4.2. Sketch proof of the validity of randomization tests for bounded null

We now give the high level proof of Theorem 1(a)—with some technical details relegated to the supplementary materials—to provide some intuition for this broader justification of Fisher randomization tests. Suppose the bounded null $H_{\preccurlyeq\boldsymbol{\delta}}$ in (7) holds, i.e., the true treatment effect satisfies $\boldsymbol{\tau} \preccurlyeq \boldsymbol{\delta}$. Then, by the definition in (2), the imputed potential outcomes satisfy

$$\boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(1) - \boldsymbol{Y}(1) = (\mathbf{1} - \boldsymbol{Z}) \circ (\boldsymbol{\delta} - \boldsymbol{\tau}) \succcurlyeq \mathbf{0} \quad \text{and} \quad \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0) - \boldsymbol{Y}(0) = \boldsymbol{Z} \circ (\boldsymbol{\tau} - \boldsymbol{\delta}) \preccurlyeq \mathbf{0}. \tag{9}$$

This is because any actual treated value is smaller than or equal to the imputed, and any actual control value is larger than or equal to the imputed.

We first consider the case in which the test statistic is effect increasing. Similar to (2), for any $\boldsymbol{a} \in \mathcal{Z}$, the imputed control potential outcome if the observed treatment assignment was $\boldsymbol{a}$ would be $\boldsymbol{Y}_{\boldsymbol{a},\boldsymbol{\delta}}(0) = \boldsymbol{Y}(\boldsymbol{a}) - \boldsymbol{a} \circ \boldsymbol{\delta} = \boldsymbol{a} \circ \{\boldsymbol{Y}(1) - \boldsymbol{\delta}\} + (\boldsymbol{1} - \boldsymbol{a}) \circ \boldsymbol{Y}(0)$, and the difference between $\boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)$ and $\boldsymbol{Y}_{\boldsymbol{a},\boldsymbol{\delta}}(0)$ has the following equivalent forms:

$$\begin{aligned}\boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0) - \boldsymbol{Y}_{\boldsymbol{a},\boldsymbol{\delta}}(0) &= \boldsymbol{a} \circ \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0) + (\boldsymbol{1} - \boldsymbol{a}) \circ \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0) - \boldsymbol{a} \circ \{\boldsymbol{Y}(1) - \boldsymbol{\delta}\} - (\boldsymbol{1} - \boldsymbol{a}) \circ \boldsymbol{Y}(0) \\ &= \boldsymbol{a} \circ \{\boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(1) - \boldsymbol{Y}(1)\} + (\boldsymbol{1} - \boldsymbol{a}) \circ \{\boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0) - \boldsymbol{Y}(0)\}\end{aligned}$$

From (9) and Definition 2, $t(\boldsymbol{a}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)) \geq t(\boldsymbol{a}, \boldsymbol{Y}_{\boldsymbol{a},\boldsymbol{\delta}}(0))$. Thus, for any $c \in \mathbb{R}$, the imputed tail probability $G_{\boldsymbol{Z},\boldsymbol{\delta}}(c)$ in (3) of the test statistic can be bounded by

$$G_{\boldsymbol{Z},\boldsymbol{\delta}}(c) = \sum_{\boldsymbol{a} \in \mathcal{Z}} \Pr(\boldsymbol{A} = \boldsymbol{a}) \mathbb{1}\{t(\boldsymbol{a}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)) \geq c\} \geq \sum_{\boldsymbol{a} \in \mathcal{Z}} \Pr(\boldsymbol{A} = \boldsymbol{a}) \mathbb{1}\{t(\boldsymbol{a}, \boldsymbol{Y}_{\boldsymbol{a},\boldsymbol{\delta}}(0)) \geq c\},$$

where the right hand side is actually the tail probability of the true randomization distribution of the test statistic $t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0))$. Therefore, we can derive that $p_{\boldsymbol{Z},\boldsymbol{\delta}} \equiv G_{\boldsymbol{Z},\boldsymbol{\delta}}(t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)))$ is stochastically larger than or equal to Unif$[0,1]$, i.e., it is a valid p-value for testing $H_{\preccurlyeq\boldsymbol{\delta}}$.

We then consider the case in which the test statistic is differential increasing. From (9),

$$t(\boldsymbol{a}, \boldsymbol{Y}(0)) - t(\boldsymbol{a}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)) = t(\boldsymbol{a}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0) + \boldsymbol{Z} \circ (\boldsymbol{\delta} - \boldsymbol{\tau})) - t(\boldsymbol{a}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)). \tag{10}$$

From Definition 3, the change of the statistic in (10) is maximized at $\boldsymbol{a} = \boldsymbol{Z}$, which immediately implies that $t(\boldsymbol{Z}, \boldsymbol{Y}(0)) - t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)) \geq t(\boldsymbol{a}, \boldsymbol{Y}(0)) - t(\boldsymbol{a}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0))$ for any $\boldsymbol{a} \in \mathcal{Z}$. Thus, $t(\boldsymbol{a}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)) - t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)) \geq t(\boldsymbol{a}, \boldsymbol{Y}(0)) - t(\boldsymbol{Z}, \boldsymbol{Y}(0))$ for any $\boldsymbol{a} \in \mathcal{Z}$, and the randomization p-value $p_{\boldsymbol{Z},\boldsymbol{\delta}}$ in (4) is bounded by

$$\begin{aligned}p_{\boldsymbol{Z},\boldsymbol{\delta}} &= \sum_{\boldsymbol{a} \in \mathcal{Z}} \Pr(\boldsymbol{A} = \boldsymbol{a}) \mathbb{1}\{t(\boldsymbol{a}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)) \geq t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0))\} \\ &\geq \sum_{\boldsymbol{a} \in \mathcal{Z}} \Pr(\boldsymbol{A} = \boldsymbol{a}) \mathbb{1}\{t(\boldsymbol{a}, \boldsymbol{Y}(0)) \geq t(\boldsymbol{Z}, \boldsymbol{Y}(0))\},\end{aligned}$$

which is actually the tail probability of the true randomization distribution of $t(\boldsymbol{Z}, \boldsymbol{Y}(0))$ evaluated at its realized value. Therefore, we can derive that $p_{\boldsymbol{Z},\boldsymbol{\delta}}$ is stochastically larger than or equal to Unif$[0,1]$, i.e., it is a valid p-value for testing $H_{\preccurlyeq\boldsymbol{\delta}}$.

Although both effect increasing and differential increasing test statistics can lead to valid randomization tests for bounded null hypotheses, their proofs are rather different, as shown above. Specifically, the randomization p-value $p_{\boldsymbol{Z},\boldsymbol{\delta}}$ using effect increasing test statistic is bounded by the tail probability of $t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0))$ evaluated at its realized value, while that using differential increasing test statistic is bounded by the tail probability of $t(\boldsymbol{Z}, \boldsymbol{Y}(0))$ evaluated at its realized value. The two bounds are generally different, although they are both stochastically larger than or equal to Unif$[0,1]$.

4.3. Inference for the maximum individual treatment effect

Consider a sharp null hypothesis $H_{c\mathbf{1}}$ of constant treatment effect c for some $c \in \mathbb{R}$. From Theorem 1(a), under our conditions on the test statistic, the randomization p-value for the sharp null $H_{c\mathbf{1}}$ can also be valid for the bounded null $H_{\preccurlyeq c\mathbf{1}}$, which, letting $\tau_{\max} \equiv \max_{1\leq i\leq n} \tau_i$, is equivalent to the null hypothesis that $\tau_{\max} \leq c$. This immediately implies that inverting randomization tests for a sequence of constant treatment effects can provide confidence sets for the maximum individual effect $\tau_{\max}$. Moreover, from Theorem 1(b), the resulting confidence sets can be intervals of forms $(\underline{c}, \infty)$ or $[\underline{c}, \infty)$. We summarize the results in the following corollary.

Corollary 1. (a) If the test statistic $t(\cdot, \cdot)$ is either differential increasing or effect increasing, then for any $\alpha \in (0, 1)$, the set $\{c : p_{\boldsymbol{Z},c\mathbf{1}} > \alpha, c \in \mathbb{R}\}$ is a $1 - \alpha$ confidence set for the maximum individual effect $\tau_{\max}$. (b) If the test statistic is differential increasing, or it is both effect increasing and distribution free, then the confidence set must have the form of $(\underline{c}, \infty)$ or $[\underline{c}, \infty)$ with $\underline{c} = \inf\{c : p_{\boldsymbol{Z},c\mathbf{1}} > \alpha, c \in \mathbb{R}\}$.

The confidence intervals for the maximum individual effect in Corollary 1 can also be thought of as intervals stating where at least some of the individual treatment effects lie, and the more homogeneous the effects are, the more individual effects these intervals will contain.

5. Randomization test for quantiles of individual treatment effects

5.1. Randomization test for general null hypotheses

For a general null hypothesis of interest, e.g., $\boldsymbol{\tau} \in \mathcal{H} \subset \mathbb{R}^n$, we can always obtain a valid p-value by maximizing the randomization p-value $p_{\boldsymbol{Z},\boldsymbol{\delta}}$ in (4) over $\boldsymbol{\delta} \in \mathcal{H}$. That is, $\sup_{\boldsymbol{\delta}\in\mathcal{H}} p_{\boldsymbol{Z},\boldsymbol{\delta}}$ is a valid p-value for testing the null hypothesis of $\boldsymbol{\tau} \in \mathcal{H}$. Unfortunately, such an optimization can be quite challenging, due to the complicated dependence of the imputed null distribution $G_{\boldsymbol{Z},\boldsymbol{\delta}}(\cdot)$ in (3) on the hypothesized effects $\boldsymbol{\delta}$. Specifically, the exact calculation of $G_{\boldsymbol{Z},\boldsymbol{\delta}}(\cdot)$ involves enumerating all possible treatment assignments, which is generally infeasible even with moderate sample size; in practice, we often use the Monte Carlo approximation. Thus, $G_{\boldsymbol{Z},\boldsymbol{\delta}}(\cdot)$ generally has no simple expression. Moreover, the p-value $p_{\boldsymbol{Z},\boldsymbol{\delta}}$ is the value of $G_{\boldsymbol{Z},\boldsymbol{\delta}}(\cdot)$ evaluated at the realized value of the test statistic, $t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0))$, which itself also depends on $\boldsymbol{\delta}$. These indicate the difficulty of the optimization of $p_{\boldsymbol{Z},\boldsymbol{\delta}}$ over $\boldsymbol{\delta}$. To ease computation, we consider using a distribution free test statistic. As discussed below, the optimization for $\sup_{\boldsymbol{\delta}\in\mathcal{H}} p_{\boldsymbol{Z},\boldsymbol{\delta}}$ will reduce to one for the realized value of the test statistic.

From Definition 4, the imputed randomization distribution (3) of a distribution free test statistic $t(\cdot, \cdot)$ under the null hypothesis $H_{\boldsymbol{\delta}}$ in (1) has the following equivalent forms:

$$G_{\boldsymbol{Z},\boldsymbol{\delta}}(c) = \sum_{\boldsymbol{a}\in\mathcal{Z}} \Pr(\boldsymbol{A} = \boldsymbol{a}) \mathbb{1}\left\{t(\boldsymbol{a}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)) \geq c\right\} = \sum_{\boldsymbol{a}\in\mathcal{Z}} \Pr(\boldsymbol{A} = \boldsymbol{a}) \mathbb{1}\left\{t(\boldsymbol{a}, \boldsymbol{y}) \geq c\right\} = G_0(c), \quad (11)$$

where $\boldsymbol{y} \in \mathbb{R}^n$ can be any fixed vector and $G_0(c)$ is a tail probability function that does not depend on the observed assignment $\boldsymbol{Z}$ or the null hypothesis of interest $\boldsymbol{\delta}$. By the definition in (4) and the fact that $G_0(c)$ is decreasing in c, (11) implies that

$$\sup_{\boldsymbol{\delta}\in\mathcal{H}} p_{\boldsymbol{Z},\boldsymbol{\delta}} = \sup_{\boldsymbol{\delta}\in\mathcal{H}} G_{\boldsymbol{Z},\boldsymbol{\delta}} \left\{t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0))\right\} = \sup_{\boldsymbol{\delta}\in\mathcal{H}} G_0 \left\{t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0))\right\} \leq G_0 \left\{ \inf_{\boldsymbol{\delta}\in\mathcal{H}} t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)) \right\}, \quad (12)$$

where the last inequality becomes equality if $G_0(\cdot)$ is continuous at $\inf_{\boldsymbol{\delta}\in\mathcal{H}} t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0))$ or $t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0))$ can achieve its infimum at some $\boldsymbol{\delta} \in \mathcal{H}$. Therefore, the right hand side of (12) is also a valid p-value for testing the null of $\boldsymbol{\tau} \in \mathcal{H}$, and more importantly, its optimization becomes much simpler, because it now involves only minimization over a known and generally closed-form function of $\boldsymbol{\delta}$. For completeness, we summarize the results below.

Theorem 2. For any distribution free test statistic and any constant region $\mathcal{H} \subset \mathbb{R}^n$, the supremum of the randomization p-value $\sup_{\boldsymbol{\delta}\in\mathcal{H}} p_{\boldsymbol{Z},\boldsymbol{\delta}}$, as well as its upper bound on the right hand side of (12), is valid for testing the null hypothesis of $\boldsymbol{\tau} \in \mathcal{H}$.

Theorem 2 avoids the intractability of $G_{\boldsymbol{Z},\boldsymbol{\delta}}(\cdot)$ and reduces the optimization to cases where the target $t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0))$ can have a closed-form expression as a function of $\boldsymbol{\delta}$. Furthermore, as demonstrated in the next subsection, when considering null hypotheses on quantiles of individual effects and using rank sum statistics, such an optimization can have a closed-form solution.

Below we give two additional remarks. First, the distribution free property for test statistics is also utilized by Rosenbaum (2007b) for analyzing the magnitude of treatment effects in the presence of interference. In particular, it helps derive the distributions of certain statistics under a uniformity trial. By contrast, the distribution free property here mainly helps ease the computation of the valid p-value in (12). Second, as suggested by a reviewer, another way to overcome the complex dependence of $G_{\boldsymbol{Z},\boldsymbol{\delta}}(\cdot)$ on $\boldsymbol{\delta}$ is through asymptotic approximation, under which we can approximate $G_{\boldsymbol{Z},\boldsymbol{\delta}}(\cdot)$ by a closed-form expression of $\boldsymbol{\delta}$. However, as discussed shortly, when considering null hypotheses on quantiles of individual effects, some units are allowed to have infinitely large individual effects. Consequently, we will consider cases allowing elements of $\boldsymbol{\delta}$ to take extreme and even infinite values, which may destroy the asymptotic approximation for $G_{\boldsymbol{Z},\boldsymbol{\delta}}(\cdot)$. On the contrary, using Theorem 2 with rank-based distribution free test statistics, the inference will be not only finite-sample valid but also robust to extreme values in $\boldsymbol{\delta}$; see also the related discussion in Remark 2.

5.2. Inference for quantiles of individual treatment effects

We sort the true individual treatment effects in an increasing order: $\tau_{(1)} \leq \tau_{(2)} \leq \ldots \leq \tau_{(n)}$, where $\tau_{(n)}$ is equivalently the maximum individual effect $\tau_{\max}$ studied in Section 4.3. In this subsection, instead of only the maximum individual effect, we intend to infer general quantiles of the individual treatment effects $\tau_{(k)}$'s for $1 \leq k \leq n$, where $k = n$ corresponds to the maximum or largest individual effect, $k = n - 1$ corresponds to the second largest individual effect, and so on. Specifically, for any $1 \le k \le n$ and any constant $c \in \mathbb{R}$, we consider the following null hypothesis that the individual effect of rank k is at most c:

$$H_{k,c} : \tau_{(k)} \le c. \tag{13}$$

In the special case of $k = n$, $H_{n,c}$ reduces to the bounded null $H_{\preccurlyeq c\mathbf{1}}$ as in (7), which focuses on the maximum individual effect $\tau_{(n)} \equiv \tau_{\max}$. Define $\mathcal{H}_{k,c} = \{\boldsymbol{\delta} \in \mathbb{R}^n : \delta_{(k)} \le c\} \subset \mathbb{R}^n$ as the set of vectors whose elements of rank k are smaller than or equal to c. Then the null hypothesis $H_{k,c}$ in (13) can be equivalently represented as $\boldsymbol{\tau} \in \mathcal{H}_{k,c}$.

We consider testing the null hypothesis $H_{k,c}$ in (13) using the randomization p-value $p_{\boldsymbol{Z},\boldsymbol{\delta}}$ in (4) with an effect increasing and distribution free statistic $t(\cdot,\cdot)$. Specifically, we focus on randomized experiments with exchangeable treatment assignment including BRE and CRE, and use the test statistic in (6) with a monotone increasing function $\phi(\cdot)$ and the "random" method for ties. We assume that the ordering of the units has been randomly permuted and is independent of the treatment assignment. Under this new ordering, we simply rank the units using the "first" method for ties. For descriptive convenience, we call such a statistic a rank score statistic, formally defined as follows.

Definition 5. A statistic $t(\cdot,\cdot)$ is a rank score statistic, if it can be written as $t(\boldsymbol{z},\boldsymbol{y}) = \sum_{i=1}^{n} z_i \phi(\mathrm{r}_i(\boldsymbol{y}))$, where the score function $\phi(\cdot)$ is monotone increasing and the rank function $\mathrm{r}(\cdot)$ uses the "random" method, or equivalently the "first" method assuming the ordering of units has been randomly permuted, for ties.

From Proposition 2, under experiments with exchangeable treatment assignment, the rank score statistic in Definition 5 is both effect increasing and distribution free. Consequently, from Theorem 2, to test the null hypothesis $H_{k,c}$ in (13), it suffices to minimize the value of the test statistic $t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0))$ over $\boldsymbol{\delta} \in \mathcal{H}_{k,c}$. As we demonstrate below, this minimization has a closed-form solution. Let $m = \sum_{i=1}^{n} Z_i$ be the number of treated units, and $\mathcal{I}_k$ be the set of indices of treated units with the largest $\min(n-k, m)$ observed outcomes for $1 \le k \le n$; when $k = n$, $\mathcal{I}_n$ is an empty set. We then define a column vector as follows:

$$\boldsymbol{\xi}_{k,c} = (\xi_{1k,c}, \xi_{2k,c}, \dots, \xi_{nk,c}) \in \mathbb{R}^n, \quad \text{where } \xi_{ik,c} = \begin{cases} \infty, & \text{if } i \in \mathcal{I}_k, \\ c, & \text{otherwise,} \end{cases} \quad (1 \le i \le n). \tag{14}$$

Theorem 3. Under a randomized experiment with exchangeable treatment assignment as in Definition 1, for any rank score statistic $t(\cdot,\cdot)$ in Definition 5, any $1 \le k \le n$ and any constant $c \in \mathbb{R}$,

$$p_{\boldsymbol{Z},k,c} \equiv \sup_{\boldsymbol{\delta} \in \mathcal{H}_{k,c}} p_{\boldsymbol{Z},\boldsymbol{\delta}} = G_0 \left\{ \inf_{\boldsymbol{\delta} \in \mathcal{H}_{k,c}} t(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)) \right\} = G_0 \left\{ t(\boldsymbol{Z}, \boldsymbol{Y} - \boldsymbol{Z} \circ \boldsymbol{\xi}_{k,c}) \right\} \tag{15}$$

is a valid p-value for testing the null hypothesis $H_{k,c}$ in (13), where G_0 and $\boldsymbol{\xi}_{k,c}$ are defined in (11) and (14), respectively. Specifically, under $H_{k,c}$, $\Pr(p_{\boldsymbol{Z},k,c} \leq \alpha) \leq \alpha$ for any $\alpha \in [0,1]$.

From Theorem 3, we are able to test whether any quantile of the individual treatment effects is bounded above by any constant. When $k = n$, the null hypothesis $H_{n,c}$ in (13) reduces to $H_{c\mathbf{1}}$ in (7), the vector $\boldsymbol{\xi}_{n,c}$ reduces to $c\mathbf{1}$, the p-value $p_{\boldsymbol{Z},k,c}$ in (15) reduces to $p_{\boldsymbol{Z},c\mathbf{1}}$ in (4), and Theorem 3 reduces to a special case of Theorem 1.

In Theorem 3, intuitively, when testing null hypothesis $H_{k,c}$ of $\tau_{(k)} \leq c$, we allow the $\tau_{(j)}$'s with $j > k$ to be arbitrarily large. Moreover, we assign these infinity values to the treated units with largest outcomes to minimize the value of the test statistic, or equivalently to maximize the randomization p-value. As a result, the calculation of $p_{\boldsymbol{Z},k,c}$ in (15) involves ranking vectors with infinite elements. In practice, we can replace those infinite elements of $\boldsymbol{\xi}_{k,c}$ by any constant larger than the difference between the maximum treated observed outcome and the minimum control observed outcome, and the value of $p_{\boldsymbol{Z},k,c}$ will remain the same. For simplicity, we use infinity, and view two negative infinite elements as equal in ranking. This is also compatible with the R software.

Remark 2. The ranking aspect of rank statistics plays an important role in making the test statistic distribution free and thus eases the computation. From Theorem 3 and the discussion before, the rank statistic also has the advantage that it is robust to extreme outcome values. Specifically, although we allow some individual treatment effects to be infinity when maximizing the p-value over $\boldsymbol{\delta} \in \mathcal{H}_{k,c}$, the rank statistic is still able to provide significant p-values against the null; see, e.g., the simulation in Section 7.2 and the application in Section 8.2.

5.3. Confidence intervals for quantiles of individual treatment effects

Similar to Corollary 1, we are now able to construct confidence sets for quantiles of the individual treatment effects based on Theorem 3. Moreover, the p-value $p_{\boldsymbol{Z},k,c}$ in (15) enjoys a certain monotonicity property that helps simplify the confidence sets. We summarize the results in the following theorem.

Theorem 4. Under a randomized experiment with exchangeable treatment assignment as in Definition 1, for any rank score statistic in Definition 5, any $1 \leq k \leq n$ and any $\alpha \in (0,1)$, (a) for any fixed $\boldsymbol{z}$ and k, $p_{\boldsymbol{z},k,c}$, defined as in (15), is increasing in c; (b) a $1-\alpha$ confidence set for $\tau_{(k)}$ is $\{c : p_{\boldsymbol{Z},k,c} > \alpha, c \in \mathbb{R}\}$, which must have the form of $(\underline{c}, \infty)$ or $[\underline{c}, \infty)$ with $\underline{c} = \inf\{c : p_{\boldsymbol{Z},k,c} > \alpha, c \in \mathbb{R}\}$.

In the special case of $k = n$, Theorem 4 is directly implied by Theorem 1 and Corollary 1. However, Theorem 4 generalizes the previous results for only the maximum individual effect to all quantiles of the individual effects. The intervals from Theorem 4 also give a sense of the sizes of effects across all units, and help understand the effect heterogeneity. For a specific k, the interval for $\tau_{(k)}$ states where the largest $n-k+1$ individual effects lie with certain confidence, and it can cover all individual effects under assumptions weaker than constant effects: if the smallest k individual treatment effects are homogeneous, then the interval for $\tau_{(k)}$ covers all the individual treatment effects.

Importantly, the inference in Theorem 4 on quantiles of individual effects can sometimes be more appropriate than Neyman (1923)'s inference on the average treatment effect. Specifically, when the outcomes have heavy tails and outliers, the average effect may be sensitive to these outliers, and the finite population asymptotic approximation (Li and Ding 2017) may work poorly. However, the quantiles are more robust to outliers than the average. Moreover, the inference in Theorem 4 is exactly valid in finite samples and does not require any large-sample approximation. See Section 7.3 for a simulation study.

Remark 3. When $k \leq n - m$, $\mathcal{I}_k$ in (14) contains the indices of all treated units, whose treatment effects are all hypothesized to be arbitrarily large. The resulting confidence interval for $\tau_{(k)}$ is usually the uninformative $(-\infty, \infty)$. Therefore, m, the size of the treatment group, can affect the performance of the method in Theorems 3 and 4. Moreover, generally larger m can lead to more quantiles of effects with informative confidence intervals. This asymmetric role of the treatment and control group sizes comes from the fact that the randomization p-value $p_{\boldsymbol{Z},\boldsymbol{\delta}}$ uses only the imputed control potential outcomes. When the treatment group size is expected to be small, we may want to use randomization p-value that uses the imputed treatment potential outcomes. As discussed in Remark 1, this can be achieved by switching the labels for treatment and control and changing the signs of the outcomes.

5.4. Inference for the number of units with effects larger than a threshold

We now consider an equivalent form of the null hypothesis $H_{k,c}$ in (13), which relates to the proportion of units with effects larger than a certain threshold. Because such a quantity can often be of interest in practice, we give a detailed discussion on its statistical inference below.

For any constant $c \in \mathbb{R}$, define

$$n(c) = \sum_{i=1}^{n} \mathbb{1}(\tau_i > c) \tag{16}$$

as the number of units whose treatment effects are larger than c. We can verify that, for any $1 \leq k \leq n$ and $c \in \mathbb{R}$, $\tau_{(k)} \leq c$ if and only if $n(c) \leq n - k$. Therefore, the null hypothesis $H_{k,c}$ in (13) has the following equivalent forms:

$$H_{k,c} : \tau_{(k)} \leq c \iff \boldsymbol{\tau} \in \mathcal{H}_{k,c} \iff n(c) \leq n - k, \qquad (1 \leq k \leq n, c \in \mathbb{R}). \tag{17}$$

Theorem 3 immediately implies that we are able to test null hypotheses about the number of units with effects larger than any threshold, as shown in the following theorem. For descriptive convenience, we define $p_{\boldsymbol{z},0,c} = 1$ for any $\boldsymbol{z}$ and c, due to the fact that $H_{0,c}$ is true by definition.

Corollary 2. Under a randomized experiment with exchangeable treatment assignment as in Definition 1, for any rank score statistic in Definition 5, (a) the p-value $p_{\boldsymbol{Z},k,c}$ in (15) is valid for testing the null hypothesis $H_{k,c}$ as given in (13) and (17) for any $1 \le k \le n$ and $c \in \mathbb{R}$; (b) for any fixed $\boldsymbol{z}$ and c, $p_{\boldsymbol{z},k,c}$, defined as in (15), is decreasing in k; (c) a $1-\alpha$ confidence interval for $n(c)$ in (16), i.e, the number of units with effects larger than c, is $\{n-k : p_{\boldsymbol{Z},k,c} > \alpha, 0 \le k \le n\}$, which must have the form of $\{j : n-\overline{k} \le j \le n\}$ with $\overline{k} = \sup\{k : p_{\boldsymbol{Z},k,c} > \alpha, 0 \le k \le n\}$.

From Theorem 4 and Corollary 2, we can know that, by construction, the $1-\alpha$ lower confidence limit of $n(c)$ is equivalently the number of quantiles of individual effects $\tau_{(k)}$'s whose $1-\alpha$ confidence intervals do not cover c.

Theorem 4 and Corollary 2 provide confidence intervals for quantiles of individual effects $\tau_{(k)}$'s as well as number (or equivalently proportion) of units with effects larger than any threshold. However, both theorems do not provide point estimation for these quantities. Indeed, these quantities are generally not identifiable due to no joint observation of the treatment and control potential outcomes for any unit. Consequently, consistent estimators for them generally do not exist. Recently, for binary or ordinal outcomes, Lu et al. (2018) and Huang et al. (2019) studied sharp bounds and constructed confidence intervals for the proportions of units with positive effects, i.e, $n(0)/n$. Importantly, our confidence intervals in Corollary 2 work for general outcomes.

5.5. Simultaneous inference for quantiles $\tau_{(k)}$'s and numbers $n(c)$'s

From Theorem 4, we are able to construct $1-\alpha$ confidence intervals for all quantiles of individual treatment effects $\tau_{(k)}$'s. Similarly, from Corollary 2, we are able to construct $1-\alpha$ confidence intervals for the numbers $n(c)$'s of units with effects larger than the thresholds c's. As demonstrated shortly, these confidence intervals will cover their corresponding truth simultaneously with probability at least $1-\alpha$, in the sense that there is no need for any correction due to multiple analyses.

Recall that the set $\mathcal{H}_{k,c}$ introduced in Section 5.2 has the following equivalent forms:

$$\mathcal{H}_{k,c} = \Big\{\boldsymbol{\delta} \in \mathbb{R}^n : \delta_{(k)} \le c\Big\} = \Big\{\boldsymbol{\delta} \in \mathbb{R}^n : \sum_{i=1}^{n} \mathbb{1}(\delta_i > c) \le n-k\Big\} \subset \mathbb{R}^n, \tag{18}$$

in parallel with the equivalence relationship in (17). Using (18), we can equivalently represent the confidence intervals for the quantiles $\tau_{(k)}$'s and the numbers $n(c)$'s as confidence sets for the treatment effect vector $\boldsymbol{\tau}$. Specifically, for any $1 \le k \le n$, the $1-\alpha$ confidence interval for $\tau_{(k)}$ in Theorem 4 has the following equivalent form as a $1-\alpha$ confidence set for $\boldsymbol{\tau}$:

$$\tau_{(k)} \in \{c : p_{\boldsymbol{Z},k,c} > \alpha, c \in \mathbb{R}\} \Longleftrightarrow \boldsymbol{\tau} \in \bigcap_{c : p_{\boldsymbol{Z},k,c} \le \alpha} \mathcal{H}_{k,c}^{\complement}, \tag{19}$$

and for any $c \in \mathbb{R}$, the $1-\alpha$ confidence interval for $n(c)$ in Corollary 2 has the following equivalent form:

$$n(c) \in \{n-k : p_{\boldsymbol{Z},k,c} > \alpha, 0 \le k \le n\} \Longleftrightarrow \boldsymbol{\tau} \in \bigcap_{k : p_{\boldsymbol{Z},k,c} \le \alpha} \mathcal{H}_{k,c}^{\complement}. \tag{20}$$

Therefore, the combination of all the confidence intervals for the $\tau_{(k)}$'s can be viewed as a confidence set for all the individual treatment effects $\boldsymbol{\tau}$, which is the intersection of the sets in (19) over $1 \leq k \leq n$. Similarly, the combination of all the confidence intervals for the $n(c)$'s can be viewed as a confidence set for $\boldsymbol{\tau}$, which is the intersection of the sets in (20) over all $c \in \mathbb{R}$. As shown in the following theorem, these two confidence sets for $\boldsymbol{\tau}$ are the same, and more importantly, they are indeed confidence sets with at least $1-\alpha$ coverage probability.

Theorem 5. Under a randomized experiment with exchangeable treatment assignment as in Definition 1 and using the p-value $p_{\boldsymbol{Z},k,c}$ in (15) with any rank score statistic $t(\cdot,\cdot)$ in Definition 5, for any $\alpha \in (0,1)$, the intersection of $1-\alpha$ confidence intervals for all $\tau_{(k)}$'s, viewed as a confidence set for the individual treatment effect vector $\boldsymbol{\tau}$, is the same as that for all $n(c)$'s. In particular, it has the following equivalent forms:

$$\bigcap_{k=1}^{n} \bigcap_{c: p_{\boldsymbol{Z},k,c} \leq \alpha} \mathcal{H}_{k,c}^{\complement} = \bigcap_{c \in \mathbb{R}} \bigcap_{k: p_{\boldsymbol{Z},k,c} \leq \alpha} \mathcal{H}_{k,c}^{\complement} = \bigcap_{k,c:\ p_{\boldsymbol{Z},k,c} \leq \alpha} \mathcal{H}_{k,c}^{\complement}$$

Moreover, it has at least $1-\alpha$ probability to cover the true individual treatment effects $\boldsymbol{\tau}$, i.e.,

$$\Pr\left(\boldsymbol{\tau} \in \bigcap_{k,c: p_{\boldsymbol{Z},k,c} \leq \alpha} \mathcal{H}_{k,c}^{\complement}\right) \geq 1-\alpha.$$

From Theorem 5, in practice, we can simultaneously construct confidence intervals for all quantiles of individual effects, or equivalently numbers of units with effects larger than any threshold. Moreover, these confidence intervals can be conveniently visualized, as illustrated in Section 1.2 using Figure 1. Note that the confidence interval for the maximum individual effect is the same as that under usual randomization inference with additive treatment effects assumption. By the simultaneous validity in Theorem 5, we can get confidence intervals on all quantiles of individual effects as free lunches, because these additional intervals will not reduce our confidence levels.

In principle, we can construct confidence sets for the n-dimensional individual effect vector $\boldsymbol{\tau}$ by inverting tests for all sharp null hypotheses. In general, however, testing arbitrary sharp null hypotheses is computationally infeasible, and it does not provide informative inferences because the parameter space is typically too unwieldy (with n units, the space of possible effects is n-dimensional). As noted by Rosenbaum (2010b), such a confidence set would not be intelligible, since it would be a subset of an n-dimensional space. Some special forms of outcomes or test statistics can lead to efficient computation and intuitive confidence sets. For example, Rosenbaum (2001) used carefully designed test statistics involving the attributable effects, and Rigdon and Hudgens (2015) considered binary outcomes. A key property utilized by these approaches is that many sharp null hypotheses are equally likely in the sense of producing the same randomization p-value, which not only avoids enumeration over all possible values of $\boldsymbol{\tau}$ but also provides a convenient form for the resulting confidence sets. By contrast, our confidence sets constructed in Theorem 5 works for general outcomes and relies mainly on the valid p-value (15) for testing null hypotheses about quantiles of individual effects, which involves efficient optimization of randomization p-value in Theorem 3. Moreover, as illustrated in Section 1.2, Theorem 5 provides nice confidence sets in $\mathbb{R}^n$ that are easy to understand, interpret and visualize.

In fact, our confidence set in Theorem 5 essentially provides a confidence band for the quantile (or equivalently distribution) function of the individual treatment effects. Relatedly, Lei and Candès (2021) constructed prediction intervals for the (random) individual treatment effect, assuming random sampling of units from some superpopulation. One main difference is that we focus on the fixed population distribution of individual effects, while they focus on a random draw from the superpopulation distribution of individual effects.

6. Extension: two-sided alternatives and effect range

In the previous discussion, we mainly focused on one-sided testing for the treatment effect $\boldsymbol{\tau}$, where the alternative hypotheses favor larger treatment effects. In fact, these results immediately imply that we are able to test the other side for $\boldsymbol{\tau}$, where the alternative hypotheses favor smaller treatment effects. We can achieve this simply by multiplying the outcomes by -1 or by switching the labels for treatment and control. By Bonferroni correction, we can also construct confidence intervals for all quantiles of individual effects using both sides of alternatives.

It is also possible to combine the confidence intervals for the maximum and minimum individual effects into a single confidence statement about the range of treatment effects. Suppose $\hat{\tau}^L_{\max}$ is a $1-\alpha/2$ lower confidence limit for the maximum individual effect $\tau_{\max}$, and $\hat{\tau}^U_{\min}$ is a $1-\alpha/2$ upper confidence limit for the minimum individual effect $\tau_{\min}$. Using Bonferroni correction, we are $1-\alpha$ confident that the effect range $\tau_{\max}-\tau_{\min}$ is at least $\hat{\tau}^L_{\max}-\hat{\tau}^U_{\min}$, based on which we are able to test whether the treatment effect is constant, an issue discussed in detail in Ding et al. (2016). For completeness, we summarize the results in the following theorem.

Theorem 6. Suppose that $[\hat{\tau}^L_{\max},\infty)$ is a $1-\alpha/2$ confidence interval for $\tau_{\max}$, and $(-\infty,\hat{\tau}^U_{\min}]$ is a $1-\alpha/2$ confidence interval for $\tau_{\min}$. Then (a) $[\max\{\hat{\tau}^L_{\max}-\hat{\tau}^U_{\min},0\},\infty)$ is a $1-\alpha$ confidence interval for the effect range $\tau_{\max}-\tau_{\min}$; (b) for the null hypothesis of constant treatment effect, i.e., $H_{c\mathbf{1}}$ holds for some $c\in\mathbb{R}$, rejecting the null if and only if $\hat{\tau}^L_{\max}-\hat{\tau}^U_{\min}>0$ leads to a valid test at significance level α.

7. Simulation studies

7.1. A simulation study for inferring the maximum individual effect

We first conduct a simulation study to investigate the power of randomization tests with different test statistics for detecting positive maximum individual treatment effect $\tau_{\max}$, including the settings where the average treatment effect is close to zero or even negative. In particular, we investigate difference-in-means, Wilcoxon rank sum and Stephenson rank sum as test statistics in a completely randomized experiment. Using Propositions 1 and 2, we know that the difference-in-means statistic is differential increasing, and the Wilcoxon rank sum and Stephenson rank sum statistics are both effect increasing and distribution free under the CRE. Therefore, from Theorem 1, the randomization p-values $p_{\boldsymbol{Z},\boldsymbol{\delta}}$ with these test statistics are all valid for testing the bounded null $H_{\preccurlyeq\boldsymbol{\delta}}$. In the following, we will investigate the power of various test statistics averaging over randomly generated potential outcomes.

We generate the potential outcomes as i.i.d. $(Y_i(0), \tau_i)$ pairs from the following model and randomize half of the units to treatment group and the remaining to control group:

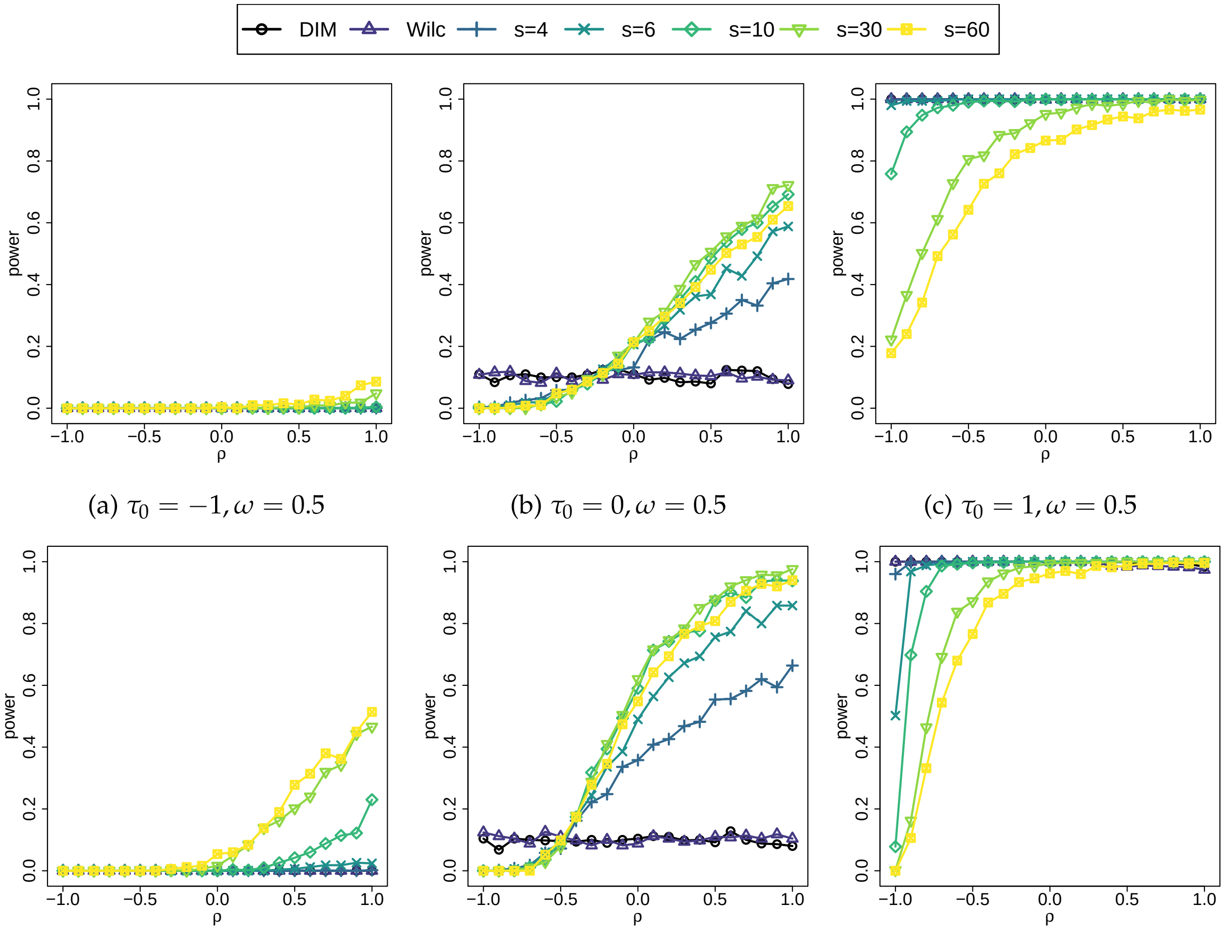
$$\begin{pmatrix} Y_i(0) \\ \tau_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \tau_0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\omega \\ \rho\omega & \omega^2 \end{pmatrix} \right), \quad Y_i(1) = Y_i(0) + \tau_i, \tag{21}$$

where τ_0 characterizes the magnitude of the average treatment effect, ρ reflects the correlation between the individual treatment effect and the control potential outcome, and ω characterizes the variability of the individual treatment effect. If ρ takes a positive value, then units with larger control potential outcomes tend to have larger individual treatment effects.

We test the bounded null $H_{\preccurlyeq\boldsymbol{0}}$ that all individual treatment effects are non-positive (or equivalently $\tau_{\max} \leq 0$) based on the randomization p-values $p_{\boldsymbol{Z},\boldsymbol{0}}$. Figures 2 shows the power of the test using different test statistics with sample size $n = 120$ and significance level 0.1, under different parameter values of (τ_0, ω, ρ). The columns correspond to different average treatment effects ranging from quite negative ($\tau_0 = -1$), to 0, to quite positive ($\tau_0 = 1$). The top row has less variation in individual treatment impacts ($\omega = 0.5$), and the bottom row has more variation ($\omega = 1$).

We first see that the performance of the difference-in-means and Wilcoxon rank sum (equivalent to Stephenson rank sum with $s = 2$ under the CRE) statistics are very similar. When the average treatment effect is non-positive, both of them have almost no power to detect positive maximum effect. However, when s increases the Stephenson rank sum statistic is able to detect the presence of positive maximum effects, even when the average treatment effect is non-positive (see bottom-left).

The choice of s for the Stephenson rank sum statistic is a researcher choice. From Section 3.3, as the value of s increases, the Stephenson rank places greater weights on larger outcomes, making the test statistic more sensitive to larger outcome. Therefore, intuitively, the "optimal" choice of s will depend on the right tails of the distributions of treatment and control potential outcomes. First, from Figures 2(a), (b), (d) and (e), when the average treatment effect is non-positive in expectation and the individual effects are not very negatively correlated with the control potential outcomes (i.e., ρ is not very small), the power of the Stephenson rank sum test generally increases with s. This is intuitive, since in this case the treated group will tend to have larger outcomes than the control group. Second, from Figures 2(c) and (f), when the average treatment effect is positive, the power of the Stephenson rank test can decrease with



(a) $\tau_0 = -1, \omega = 0.5$

(b) $\tau_0 = 0, \omega = 0.5$

(c) $\tau_0 = 1, \omega = 0.5$

(d) $\tau_0 = -1, \omega = 1$

(e) $\tau_0 = 0, \omega = 1$

(f) $\tau_0 = 1, \omega = 1$

Figure 2: Power of randomization tests using $p_{\boldsymbol{Z},\mathbf{0}}$ with different test statistics for the null hypothesis $H_{\preccurlyeq \mathbf{0}}$ or equivalently $\tau_{\max} \le 0$ at significance level equals 0.1. The potential outcomes are generated from (21) with sample size $n = 120$ and different values of (τ_0, ω, ρ).

s, especially for small or negative ρ. This is also intuitive, since in this case the control group is more likely to have larger outcomes, which will reduce the observed Stephenson rank sum statistic, and thus reduce power. Overall, we suggest a moderately large s for randomization tests to infer maximum treatment effects. As a side note, Conover and Salsburg (1988) examined the asymptotic relative efficiency of a closely related class of test statistics, and found that when only a small fraction of treated respond, the optimal subset size s is between 5 and 6.

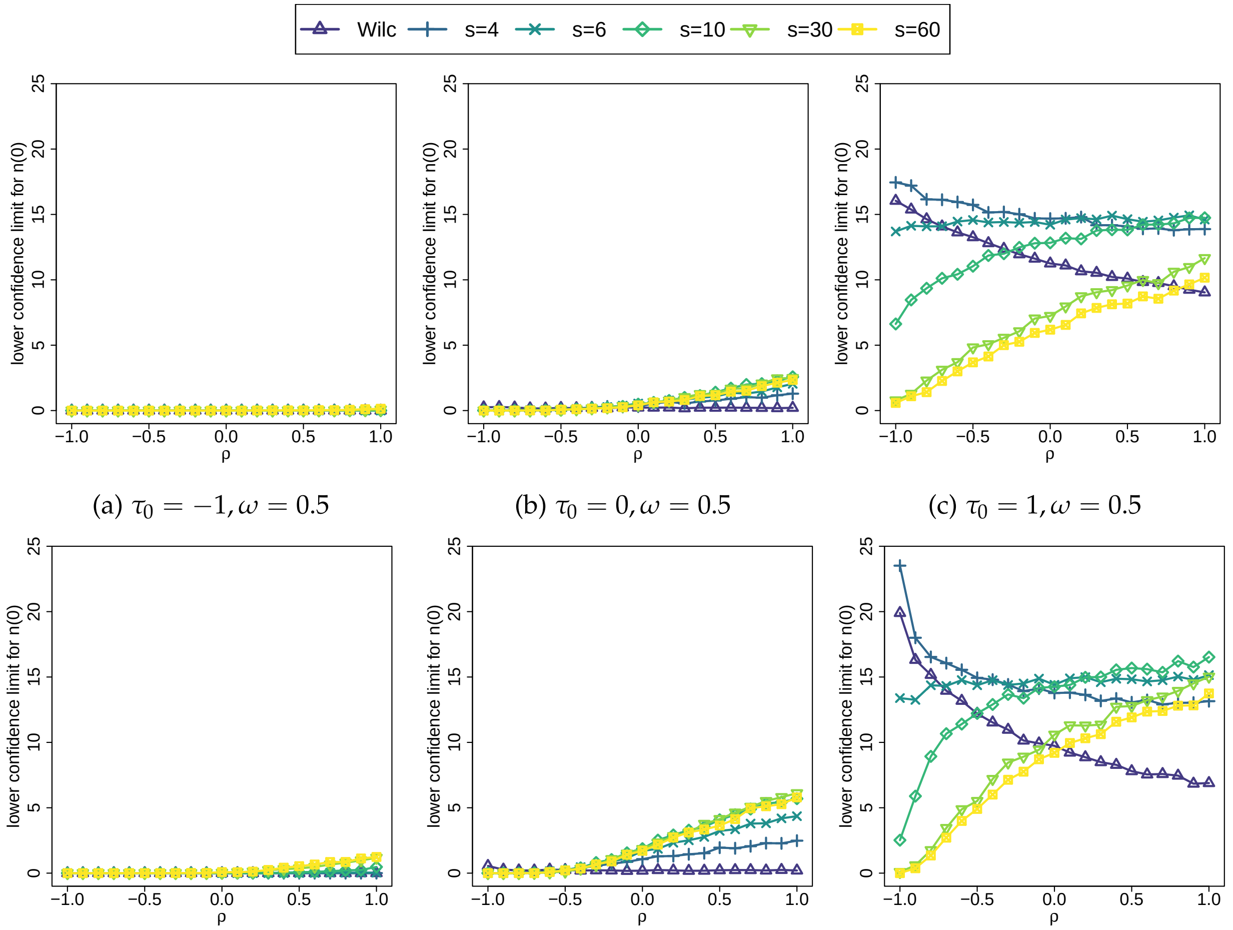
In sum, although the Wilcoxon rank sum statistic is commonly used in practice and has greatest relative power when treatment effects are close to constant, the Stephenson rank sum statistic can be preferred due to its sensitivity to extreme treatment effects, which can lead to tighter confidence intervals for the maximum effect when the maximum differs greatly from the mean or median. It is even possible for a Stephenson rank sum test to reject the bounded null $H_{c\mathbf{1}}$ for positive values of c when the average treatment effect estimate is negative, if some individual treatment effects are sufficiently positive. Thus, when treatment effects are heterogeneous, the behavior of the Stephenson rank sum test can differ markedly from the Wilcoxon rank sum or difference-in-means, while, like them, still providing a valid test for the bounded null hypothesis. This also means that it can have greater power when using permutation testing in the classic sense of testing whether there is any violation of the sharp null of no treatment effects whatsoever.

7.2. A simulation study for inferring quantiles of individual effects

We next conduct a simulation study to investigate the power of different rank score statistics for detecting positive quantiles of individual treatment effects. We generate the potential outcomes and treatment assignment in the same way as that in Section 7.1, and focus on the inference of the number of units with effects larger than zero, i.e, $n(0)$ as defined in (16). Recall that the lower confidence limit of $n(0)$ is equivalently the number of $\tau_{(k)}$'s whose confidence intervals do not cover zero.

Figure 3 shows the average lower bounds of the 90% confidence intervals for $n(0)$ using the Stephenson rank sum statistics with parameter s ranging from 2 to 60, where $s = 2$ corresponds to the Wilcoxon rank sum statistic. From Figures 3(a), (b), (d) and (e), when the average treatment effect is less than or equal to zero, the Wilcoxon rank sum statistic has almost no power to provide informative confidence intervals for the number of units with positive effects. However, the Stephenson rank sum statistics are able to detect significant amount of positive treatment effects, where larger s tends to give larger lower confidence limits. From Figures 3(c) and (f), when the average treatment effect is positive, the power of the Stephenson rank sum statistic becomes non-monotone in s. In particular, very large value of s can lead to deteriorated confidence limits for $n(0)$, especially when the individual treatment effect is negatively correlated with the control potential outcome. The intuition is similar to that discussed in Section 7.1: the control group is likely to have larger outcomes and the Stephenson rank with large s places greater weights on these larger outcomes, making the test less powerful to detect positive individual effects.

Different Stephenson rank statistics can be preferred for different estimands of interest, even



(a) $\tau_0 = -1, \omega = 0.5$

(b) $\tau_0 = 0, \omega = 0.5$

(c) $\tau_0 = 1, \omega = 0.5$

(d) $\tau_0 = -1, \omega = 1$

(e) $\tau_0 = 0, \omega = 1$

(f) $\tau_0 = 1, \omega = 1$

Figure 3: Average lower limits of 90% confidence intervals for the number of units with positive effects $n(0)$. The potential outcomes are generated from (21) with sample size $n = 120$ and different values of (τ_0, ω, ρ).

under the same data generating process. For example, Figure 4 shows the power of various Stephenson rank statistics for testing whether each quantile of individual effect is bounded by zero, i.e., $H_{k,0} : \tau_{(k)} \leq 0$ over all $1 \leq k \leq n$, under the data generating model with $n = 120$ and (τ_0, ω, ρ) equal to $(1, 1, -0.9)$. (We omit those values of k for which all tests under consideration have zero power.) From Figure 4, we can see that $s = 2$ is preferred for larger quantiles of individual effects, while $s = 4$ is preferred for smaller quantiles.

Selecting s is a decision the researcher has to make when planning their data analysis. To aid with this, we provide functions in our developed package that, under a user-specified data generating model, compare the performance of various Stephenson rank sum statistics in terms of either power for testing the null hypotheses for given quantiles of individual effects or the average magnitude of confidence limits for the number of units with effects passing any given threshold. Users can then specify a range of possible distributions of control-side potential out-

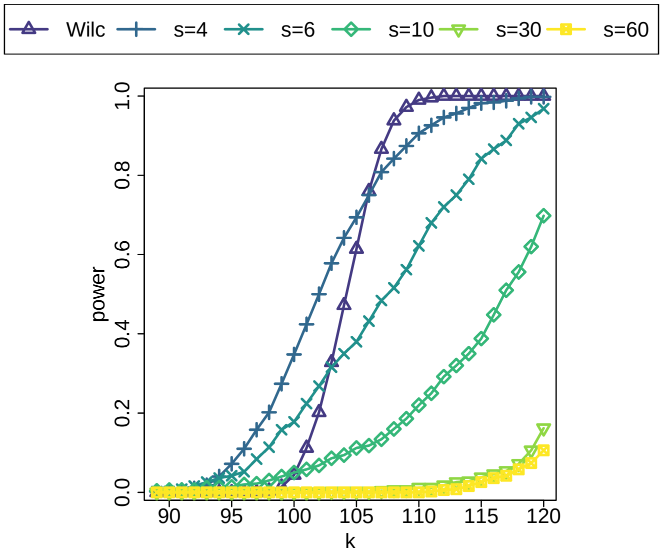


Figure 4: Power of various Stephenson rank statistics for testing the null hypothesis of $H_{k,0}$: $\tau_{(k)} \leq 0$ over all k, under model (21) with $n = 120$ and (τ_0, ω, ρ) equal to $(1, 1, -0.9)$.

comes, treatment impact models, and potential correlations of the two, along with their primary estimand of interest (such as $\tau_{(k)}$ for some k or $n(c)$ for some c), and compare the performances of different s values for these hypothetical scenarios at a given sample size n. They would then select s based on which value generally had superior performance for the targeted estimands of interest across the explored scenarios. The distributions of outcomes and effects, and their correlation, would be obtained from the empirical data and prior knowledge (such as from pilot studies).

In our applications, we mainly consider the Stephenson rank sum statistic with $s = 6$ or 10, based on this process and the other simulations in this paper. Generally $s = 6$ appears a reasonable starting point for a versatile choice in the absence of empirical information.

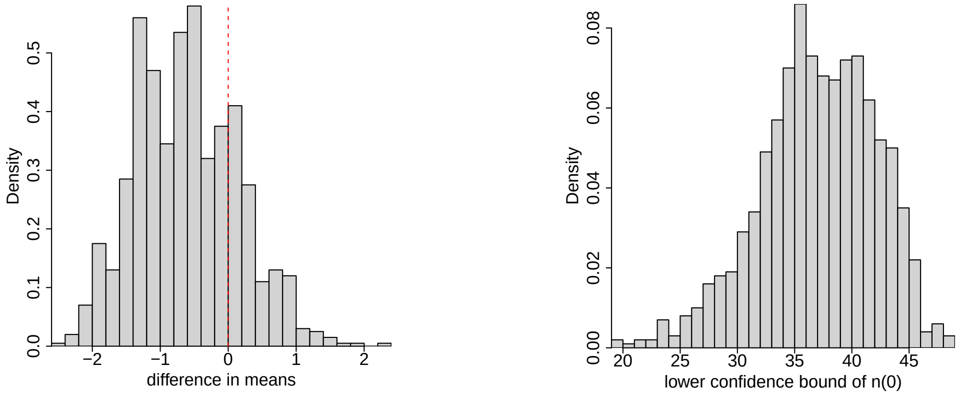
7.3. Heavy-tailed outcomes and individual effects

We next conduct a simulation with heavy-tailed outcomes and individual treatment effects. We will demonstrate that, compared to inference on average treatment effects (e.g., Neyman 1923), the proposed inference on quantiles does not require any large-sample approximation and can be more robust to outliers. In particular, we consider the case of generally positive constant individual effects with a few extreme negative outliers. In the supplement, we also consider a scenario with heavy-tailed distributions of the potential outcomes under Fisher's null of no effect.

We simulate an experiment with 120 units, among which two thirds will be randomly assigned to treatment and the remaining to control. We assume the treatment increases a certain risk factor by 2 for 95% of the units, and decreases it by 50 for the remaining 5% of units. We simulate the control potential outcomes from the standard normal distribution. Once generated, all the potential outcomes are fixed. For the final sample, about 5% of the units receive large benefit, but the remainder would incur a negative effect. The true average treatment effect is negative, indicating an average benefit, even though the majority of units are harmed.

Figure 5(a) shows the histogram of the sampling distribution of the usual difference-in-means estimator. Over all the simulated assignments, the difference-in-means estimator is negative about 77% of the time, and its average is close to the true average effect -0.6. In general, the difference in means estimator would indicate that the treatment is reducing overall risk level and is apparently beneficial. However, such a conclusion would potentially be misleading, because the treatment is harmful for most units.

Figure 5(b) shows the histogram of the 90% lower confidence limit of the number of units with higher risk level under treatment than control (i.e., $n(0)$), based on the Stephenson rank statistic with $s = 6$. Over all simulated assignments, the lower confidence limit of $n(0)$ has an average value of 37.5 (about 31% of the units), and can sometimes reach 49 (about 41% of the units). Our method reliably detects that the treatment harms a significant amount of units; such a finding would signal to researchers that the treatment would need further careful investigation despite an apparent average benefit.



(a)

(b)

Figure 5: Simulation when individual treatment effects have extreme values or outliers. (a) shows the histogram of the usual difference-in-means estimator. (b) shows the histogram of the 90% lower confidence limit of the number of units with positive effects using Theorem 4.

8. Applications

8.1. Testing monotonicity of an instrumental variable

The assumption that the instrument has monotonic effects on the treatment, though conventionally invoked for identification of instrumental variable (IV) estimates (Angrist et al. 1996), is rarely evaluated in empirical applications. Recently, however, the issue of non-monotonicity has received attention in the active literature on school-entry age, in which numerous studies

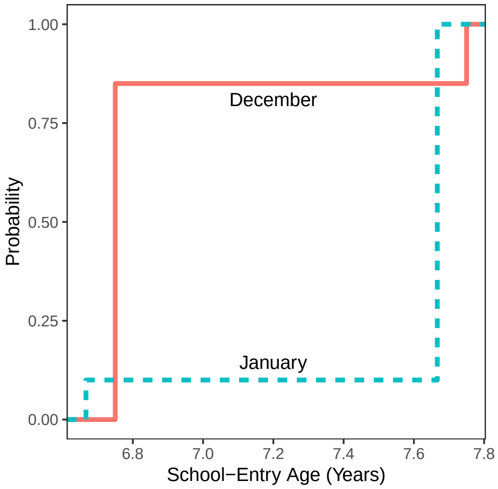


Figure 6: Empirical cumulative distribution functions of school-entry age for units born in December and January, respectively (Black et al. 2011).

involve instruments based on laws regulating entry age (Aliprantis 2012; Barua and Lang 2016; for an overview, see Fiorini and Stevens 2014). Typically, these laws select an arbitrary date of birth before which children are allowed to enter school in given calendar year. If the cutoff date is January 1, for example, most children born in December will be about 11 months younger when they enter school than children born in January. Due to imperfect compliance with the instrument, however, some fraction of December children may "redshirt" and start the following school year, at which time they will be one month *older* than than January children who started on time. Unless the December children who redshirt would also have redshirted had they been born in January, monotonicity is violated. That is, the effect of December birth on school-entry age is typically negative, but for a few children it is positive (an analogous logic holds for January children who start early).

For a simple illustration of how randomization inference can be used to evaluate monotonicity, we re-analyze data on 104,000 children born in December or January from Black et al. (2011)'s IV study of school-entry age, which have previously been analyzed from a sampling-based perspective by Fiorini and Stevens (2014)[9]. The latter authors note that the cumulative distribution functions of December- and January-born children in these data cross each other, suggesting a violation of monotonicity. Figure 6 indicates this clearly. Most children started school at an older age if they were born in January rather than December. Indeed, this first-stage relationship is incredibly strong, with an average effect of -0.667 years and an F statistic of 106,256. This would conventionally be considered persuasive evidence of a valid instrument. Note, however, that at the tails of the distribution the relationship between month of birth and entry age reverses:

[9] Table 1 of Black et al. (2011) cross-tabulates month of birth and school entry age (early/on-time/late) in terms of proportions, and Table 3 there reports the total number of subjects born in January or December (the "discontinuity subsample"): 104,023. Similar to Fiorini and Stevens (2014), we assume equal numbers of subjects born in January and December, and round the total number of subjects to 104,000, to ensure integer number of subjects in each subgroup.

January birthdays predominate among the youngest starters, and December does so among the oldest.

Table 1: Randomization tests for the bounded null $H_{\preccurlyeq \mathbf{0}}$, which is equivalent to the monotonicity assumption. The intervals are 90% confidence intervals for the maximum individual effect, constructed by inverting the corresponding randomization tests. Columns 2–4 show results from randomization inference using different test statistics. The last column shows results from the classical Student's t-test.

	Difference-in-means	Wilcoxon	Stephenson	Student's t-test
p-value	1	1	0	1
90% CI	$[-0.669, \infty)$	$[-0.916, \infty)$	$[0.084, \infty)$	$[-0.670, \infty)$

If we designate December birth as the assigned-to-treatment condition and January birth as control, then the monotonicity assumption is equivalent to the null hypothesis $H_{\preccurlyeq \mathbf{0}}$: being born in December did not cause any child to go to school at an older age than they would have if born in January. We assume that birth month is as-if randomly assigned, and conduct randomization inference for the effect of December birth on the school-entry age. We test the bounded null $H_{\preccurlyeq \mathbf{0}}$ using the randomization p-value $p_{\boldsymbol{Z},\mathbf{0}}$ with various test statistics satisfying the conditions in Theorem 1(a), including the difference-in-means, Wilcoxon rank sum and Stephenson rank sum with $s = 10$. Table 1 lists the results from randomization tests using these three test statistics, supplemented by the classical Student's t-test. We emphasize that, although both the randomization p-values and the corresponding intervals are numerically the same as that under the usual constant treatment effect assumption, they are also valid p-values for the bounded null $H_{\preccurlyeq \mathbf{0}}$ and valid confidence intervals for the maximum individual effect $\tau_{\max}$, as demonstrated in Section 4. Because the difference-in-means estimator for the average effect is negative, from the simulation results in Section 7.1, it is not surprising that neither difference-in-means or Wilcoxon rank sum give significant p-values. However, the Stephenson rank sum gives an almost zero p-value, strong evidence of the existence of units violating the monotonicity assumption. Intuitively, the significant p-value is driven by the Stephenson rank placing more weight on larger outcomes coupled with 15% of December-born children having a school-entry age of 7.75, which is larger than the maximum school-entry age 7.67 for the January-born children. This means the treatment group has a large share of the most extreme observations. The corresponding 90% lower confidence limit is 0.084 year, or equivalently about 1 month, suggesting some children would first enter the school one-month older if born in December than in January. We can therefore confidently conclude that being born in December increased school-entry age for at least some students, i.e., the IV monotonicity assumption is violated in this application.

We now apply Theorem 6 to study the effect range. Using the Stephenson rank sum statistic with $s = 10$, the 95% upper confidence limit for the minimum effect of December birth is -0.917 years (-11 months) while the 95% lower confidence limit for the maximum effect is 0.083 years (1 month). We are therefore 90% confident that the range of the effect of birth month on school-entry age is at least 1 year, indicating significant individual effect heterogeneity. This of course is hardly surprising given Figure 6, which shows that despite the negative average effect of December birth, the children with the oldest school-entry age were born in this month.

8.2. Evaluating the effectiveness of professional development

Heller et al. (2010) studied the effectiveness of professional development on elementary teachers, classrooms and students using a randomized experiment conducted at eight national research sites. A sample of fourth grade teachers were randomly assigned to treatment and control, where treated teachers would participate in a professional development course encompassing eight three-hour sessions focusing on the teaching of electric circuits. We are interested in the effect of professional development on the teachers' electric circuits content knowledge, as measured by the gain scores based on tests before and after the professional development courses. The actual experiment was randomized within site and school, and the active treatment had three versions with regard to additional activities for the development of pedagogical content knowledge. For simplicity and to illustrate our approach, we analyze it as a completely randomized treatment-control experiment and exclude teachers with missing outcomes[10], resulting in 164 treated teachers and 69 control teachers. Figure 8(a) shows the histograms of the observed gain scores in the treatment and control groups. The treated teachers tend to have larger gain scores, and the corresponding histogram is close to a positive shift of that for the control teachers, with a magnitude of 15 to 20. Therefore, intuitively, we expect our approach to infer a significant proportion of teachers with positive effects as there is little sign of treatment effect heterogeneity and a large share of treated teachers' outcomes are larger than nearly all control teacher outcomes.

Before proceeding with our analysis, we need to select the tuning parameter s for our Stephenson statistic. We do this via a power simulation, generating a series of datasets with an empirical control-side distribution bootstrapped from the control units in our data, the same estimated average treatment effect, different hypothesized distributions of treatment effects (constant, normal, exponential) with different levels of impact variation, and different correlations of treatment impact with baseline outcome. We calibrate an assumed treatment variation by assuming zero correlation and comparing the variances of the treated and control groups, but explore other values as well. We additionally explore both positive and negative correlations as a sensitivity check. For each dataset we calculate simultaneous confidence intervals for the largest 100 individual effects, using the method described in Section 5. For each of these top 100, we then calculate the median lower bound of their confidence intervals across simulation runs along with power (the probability of the CI excluding zero).

Figure 7 shows the median lower bounds as a function of s averaged over three different groups of quantiles (the top 10, next 20, and next 70). We generally see high informativeness of the intervals for low s. Averaged across scenarios, we find optimal s of 4, 5, and 6 for the high, middle, and low groups. Our supplementary replication file gives further details of the above

[10] Since the outcomes of all control teachers are missing in one of the eight sites, we exclude that site in our analysis.

procedure and aggregation.

We also examine the performance on the number of significant units by examining the median of the lower confidence bound across the simulation runs for the different scenarios; here we find that $s = 5$ generally gives the largest number of significant units, unless the pattern of impacts is exponential, in which case a larger $s = 6$ or $s = 8$ is warranted.

Given the above, especially considering the constant shift suggesting possible constant treatment impact, we select $s = 5$ for our analysis.

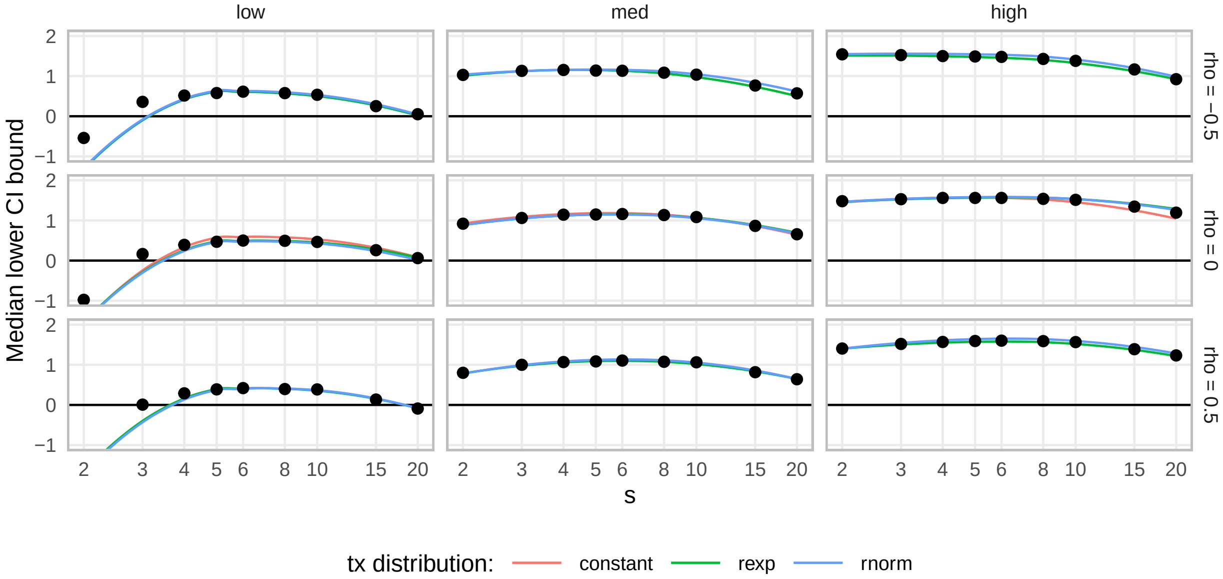
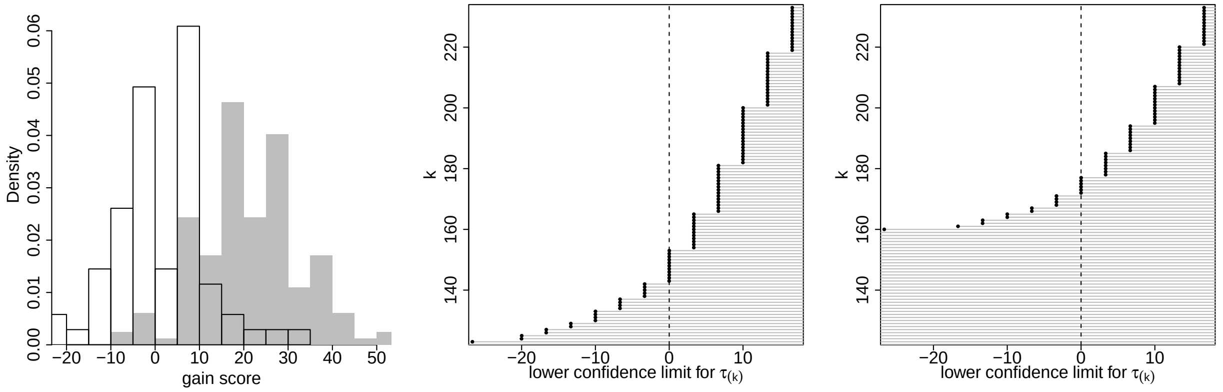


Figure 7: Median CI lower bounds over different groups of quantiles and simulation scenarios as a function of s for an exploratory simulation calibrated to the teacher professional development data. From left to right we have top 13 quantiles, next 20, and next 30. Rows correspond to different correlations of $Y_i(0)$ and τ_i. Black dots give empirical averages across all scenarios considered.

With our selected s we then analyze the original data, calculating confidence intervals for all quantiles. Figure 8(b) shows the 90% lower confidence limits for all the $\tau_{(k)}$'s using the Stephenson rank sum statistic with $s = 5$ that have finite lower limits (the lower confidence limits for $\tau_{(k)}$'s with $k \leq 122$ are all negative infinity). From Figure 8(b), the lower confidence limits of $\tau_{(k)}$'s with $149 \leq k \leq 233$ are all larger than zero, implying that a 90% confidence interval for $n(0)$ is $[85, 233]$. Equivalently, we are 90% confident that at least $85/233 = 36.5\%$ units would benefit from the professional development courses. Similarly, a 90% confidence interval for $n(6)$ is $[68, 233]$. That is, we are 90% confident that at least $68/233 = 29.2\%$ teachers would have gained six more points in the test if they had participated in the professional development.

To show the benefit of using Stephenson ranks, we also plot the 90% lower confidence limits for all quantiles $\tau_{(k)}$'s using the usual Wilcoxon rank sum statistic, as shown in Figure 8(b). We have noninformative limits for all $\tau_{(k)}$'s with $k \leq 159$, rather than 122, and the lower confidence



(a)

(b) $s = 5$

(c) $s = 2$

Figure 8: Histograms of observed gain scores and 90% confidence intervals for quantiles of individual effects. (a) shows the histograms of the observed gain scores in treatment (grey) and control (white) groups, respectively. (b) and (c) shows the 90% lower confidence limits for all quantiles of individual effects using the Stephenson rank statistics with s equal to 10 and 2, respectively. The uninformative confidence intervals of $(-\infty, \infty)$ for individual effects at lower ranks are omitted from (b) and (c).

limits for $n(0)$ and $n(6)$ are respectively 59 and 48, corresponding to 25.3% and 20.6% of the teachers. Obviously, the inference results using Stephenson ranks are far more informative.

Finally, we apply Theorem 6 to study the effect range. The confidence intervals for the minimum and maximum effects overlap substantially, yielding a completely uninformative confidence interval for the effect range and precluding rejection of the hypothesis of a constant effect. This is consistent with the graphical evidence in Figure 8(a), where the distribution of the observed outcome for treated units is close to a shift of that for control units.

9. Conclusion and Discussion

The rise of nonparametric causal inference in the Neyman-Rubin tradition has been one of the most important statistical developments in the social and biomedical sciences. This perspective, with its focus on the average effects and its acceptance of effect heterogeneity as the rule rather than the exception, has rightly prompted greater skepticism of statistical methods that rely on parametric assumptions. It is then perhaps no surprise that RI, which has traditionally been motivated in terms of shift hypotheses or other highly structured models of treatment effects (e.g., Lehmann 1963; Rosenbaum 2002), has also been regarded with skepticism, despite its freedom from other distributional assumptions.

We have argued that the view of RI prevalent among statisticians and applied researchers — that RI is useful only for assessing the typically uninteresting and implausible sharp hypothesis that treatment had no effect at all—is too limited. We have proved that randomization tests can be valid under a more general bounded null, that this fact can be exploited to derive confidence intervals for the maximum or minimum effect, and that many familiar test statistics can lead to tests with this property. We then extended the RI for the maximum (or minimum) effect to general quantiles of the individual effects, which in turn provides confidence intervals for the number (or equivalently proportion) of units with effects larger than (or smaller than) any threshold. Moreover, the confidence intervals for all quantiles of individual effects are simultaneously valid.

We have also highlighted the value of less familiar statistics such as the Stephenson rank sum, which is sensitive to the extremes of the treatment effect distribution, and explained the normative and theoretical relevance to infer quantiles of individual effects. There are many interesting directions worth exploring in the future. First, as shown in Section 7.2, different Stephenson rank sum statistics can be preferred under different data generating models and for different causal estimands. Thus, it will be preferable to choose the test statistics adaptively based on the observed data. Second, Ghosh et al. (2021) recently studied asymptotic properties of randomization test under sharp null hypotheses with constant individual effects. It will be interesting to extend their theory to bounded null hypotheses and quantiles of individual effects. Third, in this paper we focus on inferring quantiles of effects in randomized experiments with exchangeable treatment assignments. It is our future work to extend it to more general randomized experiments, such as stratified/blocked randomized experiments.

In sum, we have offered a novel perspective on RI that we hope tempers the skepticism that many applied scientists hold towards this otherwise-appealing mode of statistical inference, and have developed a novel method to infer all individual effects simultaneously, in particular confidence intervals for quantiles of the effects and proportions of units with effects larger (or smaller) than any threshold. RI is by no means a cure-all; nor is it a substitute for average treatment effect estimation when that is the goal of the analysis. But in many cases, particularly when sample size is small or outcomes are heavy-tailed, they are the most reliable form of statistical inference. Moreover, even when this is not the case, they can provide fruitful results that may help understand the pattern of treatment effects across the whole population. For these reasons, RI deserves a secure place in the quantitative applied scientists' toolbox.

Supplementary Materials

The supplementary materials contain two files. The first file includes (i) a discussion on the validity of randomization tests using test statistics of form $t(\boldsymbol{Z}, \boldsymbol{Y})$ for bounded null hypotheses, (ii) the proofs of all theorems, corollaries and propositions, (iii) additional simulation studies, and (iv) an application for evaluating the effects of six-month nutrition therapy using a randomized controlled trial. The second file includes (i) installation of the developed R package, (ii) replication for the three data analyses in the paper and the supplementary materials, and (iii) illustration for the R functions to compare the power of different Stephenson rank sum statistics. The R package RIQITE implementing the proposed methods is available at `https://github.com/li-xinran/RIQITE`.

Acknowledgments

We thank the Editor, the Associate Editor and all the reviewers for insightful and constructive comments.

REFERENCES

D. Aliprantis. Redshirting, compulsory schooling laws, and educational attainment. *Journal of Educational and Behavioral Statistics*, 37:316–338, 2012.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–455, 1996.

R. Barua and K. Lang. School entry, educational attainment, and quarter of birth: A cautionary tale of a local average treatment effect. *Journal of Human Capital*, 10:347–376, 2016.

S. E. Black, P. J. Devereaux, and K. G. Salvanes. Too young to leave the nest? the effects of school starting age. *Review of Economics and Statistics*, 93:455–467, 2011.

J. Bowers, M. M. Fredrickson, and C. Panagopoulos. Reasoning about interference between units: A general framework. *Political Analysis*, 21(1):97–124, 2013.

M. D. Cattaneo, B. R. Frandsen, and R. Titiunik. Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate. *Journal of Causal Inference*, 3(1):1–24, 2015.

E. Chung and J. P. Romano. Exact and asymptotically robust permutation tests. *Annals of Statistics*, 41:484–507, 2013.

P. L. Cohen and C. B. Fogarty. Gaussian prepivoting for finite population causal inference. *arXiv preprint arXiv:2002.06654*, 2020.

W. J. Conover and D. S. Salsburg. Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to 'respond' to treatment. *Biometrics*, 44(1): 189–196, 1988.

P. Ding and T. Dasgupta. A randomization-based perspective on analysis of variance: a test statistic robust to treatment effect heterogeneity. *Biometrika*, 105:45–56, 7/9/2019 2017. doi: 10.1093/biomet/asx059. URL `https://doi.org/10.1093/biomet/asx059`.

P. Ding, A. Feller, and L. Miratrix. Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78:655–671, 2016.

A. W. F. Edwards. The measure of association in a 2 × 2 table. *Journal of the Royal Statistical Society: Series A (General)*, 126:109–114, 1963.

M. Fiorini and K. Stevens. Assessing the Monotonicity Assumption in IV and Fuzzy RD designs. Working papers, University of Sydney, School of Economics, 2014. URL `http://EconPapers.repec.org/RePEc:syd:wpaper:2014-13`.

R. A. Fisher. *Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

C. B. Fogarty. Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. *Journal of the American Statistical Association*, (just-accepted):1–35, 2019.

D. Freedman, Robert P., and Roger P. *Statistics*. Norton, New York, 3 edition, 1997.

A. Gelman. Why it doesn't make sense in general to form confidence intervals by inverting hypothesis tests. *Statistical Modeling, Causal Inference, and Social Science*, 2011. URL `http://andrewgelman.com/2011/08/25/why_it_doesnt_m/`.

A. Ghosh, N. Deb, B. Karmakar, and B. Sen. Efficiency of regression (un)-adjusted rosenbaum's rank-based estimator in randomized experiments. *arXiv preprint arXiv:2111.15524*, 2021.

J. L. Heller, M. Shinohara, L. Miratrix, S. R. Hesketh, and K. R. Daehler. Learning science for teaching: Effects of professional development on elementary teachers, classrooms, and students. *Proceedings from Society for Research on Educational Effectiveness.*, 2010.

D. E. Ho and K. Imai. Randomization inference with natural experiments: An analysis of ballot effects in the 2003 california recall election. *Journal of the American Statistical Association*, 101 (475):888–900, 2006.

E. J. Huang, E. X. Fang, D. F. Hanley, and M. Rosenblum. Constructing a confidence interval for the fraction who benefit from treatment, using randomized trial data. *Biometrics*, 75:1228–1239, 2019.

G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, 2015.

L. Keele. The statistics of causal inference: A view from political methodology. *Political Analysis*, 23(3):313–335, 2015.

E. L. Lehmann. Nonparametric confidence intervals for a shift parameter. *Annals of Mathematical Statistics*, 34(4):1507–1512, 1963.

L. Lei and E. J. Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83:911–938, 2021.

X. Li and P. Ding. Exact confidence intervals for the average causal effect on a binary outcome. *Statistics in Medicine*, 35:957–960, 2016.

X. Li and P. Ding. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American statistical Association*, 112:1759–1769, 2017.

X. Li, P. Ding, and D. B. Rubin. Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 115: 9157–9162, 2018.

J. Lu, P. Ding, and T. Dasgupta. Treatment effects on ordinal outcomes: Causal estimands and sharp bounds. *Journal of Educational and Behavioral Statistics*, 43:540–567, 2018.

C. F Manski. *Identification for prediction and decision*. Harvard University Press, 2009.

L. W. Miratrix, J. S. Sekhon, and B. Yu. Adjusting treatment effect estimates by post-stratification in randomized experiments adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75: 369–396, 2013.

Edward J. Mishan. *Introduction to Political Economy*. Hutchinson, 1982.

K. L. Morgan and D. B. Rubin. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40:1263–1282, 2012.

J. Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Roczniki Nauk Roiniczych, Tom X*, pages 1–51, 1923.

J. Neyman. Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society*, 2(2):107–180, 1935.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`.

J. Rigdon and M. G. Hudgens. Randomization inference for treatment effects on a binary outcome. *Statistics in Medicine*, 34(6):924–935, 2015. doi: 10.1002/sim.6384. URL `https://doi.org/10.1002/sim.6384`.

P. R. Rosenbaum. Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika*, 88(1):219–231, 2001.

P. R. Rosenbaum. *Observational Studies*. Springer, New York, 2 edition, 2002.

P. R. Rosenbaum. Confidence intervals for uncommon but dramatic responses to treatment. *Biometrics*, 63(4):1164–1171, 2007a.

P. R. Rosenbaum. Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102:191–200, 2007b.

P. R. Rosenbaum. Design sensitivity and efficiency in observational studies. *Journal of the American Statistical Association*, 105(490):692–702, 2010a.

Paul R. Rosenbaum. *Design of Observational Studies*. Springer, New York, 2010b.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

R. W. Stephenson and M. Ghosh. Two sample nonparametric tests based on subsamples. *Communications in Statistics: Theory and Methods*, 14(7):1669–1684, 1985.

J. Wu and P. Ding. Randomization tests for weak null hypotheses. *arXiv preprint arXiv:1809.07419*, 2018.

X. Xie, Z. Ma, and Z. Geng. Some association measures and their collapsibility. *Statistica Sinica*, 18:1165–1183, 2008. ISSN 10170405, 19968507. URL `http://www.jstor.org/stable/24308536`.