Beyond the Sharp Null:
Permutation Tests, Heterogeneous Effects,
and Bounded Null Hypotheses[*]

Devin Caughey   Allan Dafoe   Luke Miratrix

MIT/Princeton   Yale   Harvard

August 24, 2016

Abstract

Permutation tests are often dismissed as tests of an uninteresting and implausible null hypothesis: the sharp null of no effects whatsoever. We dispute this characterization, showing that many one-sided permutation tests are also valid under a much more general "bounded" null hypothesis in which all effects are weakly negative (or positive), thus allowing for heterogenous effects. By inverting such a test we can form one-sided confidence intervals for the maximum (or minimum) treatment effect. These properties hold for all effect-increasing test statistics. These include not only common statistics such as the difference of means, but also Stephenson rank statistics, whose sensitivity to large-but-rare effect permits detection of positive effects even when the average effect is negative. We show that bounded nulls are often of normative or theoretical interest, and illustrate our results' implications of with a re-analysis of a well-known field experiment.

Contents

So long as the *average* yields of any treatments are identical, the question as to whether these treatments affect *separate* yields on *single* plots seems to be uninteresting and academic. . . . [I]t is immaterial whether any two varieties react a little differently to the local differences in the soil. What is important is whether on a larger field they are able to give equal or different yields.

Jerzy Neyman (1935, 173)

1 Motivation

Permutation tests, also known as randomization tests, were developed by R. A. Fisher as general procedures for assessing hypotheses about treatment effects. Fisher (1935) demonstrated that if treatment is randomly assigned to units, the hypothesis that no unit was affected by treatment—the so-called "sharp null of no effects"—can be tested exactly, with no further assumptions, by comparing an observed test statistic with its distribution across alternative permutations of treatment assignment. Thus, unlike likelihood or Bayesian methods, permutation inference does not require parametric assumptions about the data-generating distribution. Nor, unlike other nonparametric methods such as average treatment effect (ATE) estimation, does it rely on asymptotic approximations with uncertain properties in small samples.[1] Rather, the validity of permutation tests depends only on assumptions about how units were assigned to treatments.[2]

---

1. Welch's unequal-variances t test, which is an asymptotically valid (specifically, conservative) nonparametric test of the ATE, is often considered highly robust, even in moderately sized samples of $n = 30$ (on the conservatism of the Welch/Neyman variance estimator, see, e.g., Samii and Aronow 2012). But even the t test can be quite inaccurate if the sample sizes differ between treatment and control and the response distributions are skewed (e.g., Hesterberg 2015, 372). Appendix A illustrates this with the example of a skewed beta distribution and sample sizes of $n_1 = 1000$ and $n_2 = 30$. Under these conditions, the t test with $\alpha = 0.01$ falsely rejects the null of mean equality over 10% of the time. By contrast, the difference-of-means permutation test maintains exact coverage under these conditions.

2. By a *valid* test, we mean one whose true false-rejection rate is no greater than its significance level α.

The guaranteed validity of permutation tests make them an appealing mode of statistical inference, and Fisher's original formulation of them has been extended in many ways. It was soon noted, for example, that permutation tests could be used to assess a null hypothesis of a constant treatment effect of any magnitude, not just Fisher's null of zero effect. Moreover, confidence intervals for a constant treatment effect could be derived by inverting a sequence of such tests (Lehmann 1963). Indeed, Rosenbaum (2002, 2010) has developed a comprehensive statistical framework of testing, estimation, and inference entirely within the Fisherian paradigm.

Nevertheless, permutation tests have also been the target of trenchant critiques. Perhaps the most damning is the charge that permutation tests are not valid tests of so-called "weak" null hypotheses that specify the value of some function of the treatment effects, such as the ATE, rather than the unit-level effects themselves. For example, if two groups differ in spread but not location—as can easily occur if treatment effects are heterogeneous but mean-zero— then interpreting a permutation test as testing the weak null that the ATE is 0 will result in rejection at higher-than-nominal rates. And unlike the t test, which is valid asymptotically, if group sizes are unequal the permutation test of the ATE can remain invalid even as the sample size tends to infinity (Romano 1990).

In the eyes of many political methodologists, permutation tests' invalidity under the weak null is a severe, if not disqualifying, limitation. Like Neyman (1935), many scholars consider sharp nulls to be "uninteresting and academic." 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)." Gelman's concerns are widely shared, even by scholars favorably inclined towards permutation tests in general. Accurately summarizing the current consensus among political methodologists, Keele (2015, 330) describes the sharp null as "a very restrictive null hypothesis" because it does not easily "accommodate heterogeneous responses to treatment." Current pedagogy in social science statistics emphasizes similar

sentiments. "The constant additive treatment effect model is too restrictive," reads one set of lecture notes. "It is difficult to imagine that the treatment effect is constant across units in any social science experiment[,] where treatment effect heterogeneity is a rule rather than an exception."[3]

Defenders of permutation tests have responded to these critiques in various ways. Some, while largely accepting the critiques of the sharp null, argue that permutation tests are 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 the weak null, as well as exactly valid under the sharp null. Other scholars defend the constant-effects assumption more forthrightly, regarding it as a convenient model or approximation that would in any case be required by parametric alternatives.[4] As Rosenbaum (2002) demonstrates, permutation tests can also be used to assess multiplicative, Tobit, quantile, and attributable effects, each of which permits (certain precise kinds of) heterogeneous additive effects (see also Bowers, Fredrickson, and Panagopoulos 2013). Moreover, in principle there is no barrier to assessing any null hypothesis with permutation tests, so long as the hypothesis is sharp in the sense that it fully specifies the unit-level treatment effects. In practice, however, testing arbitrary sharp null hypotheses does not provide informative inferences because the parameter space is too unwieldy (with n units, the space of possible effects is n-dimensional—assuming no spillover!).

---

3. This quotation is from the lecture notes of Kosuke Imai (2013, 7).

4. Unless treatment-effect variation is explicitly modeled, both likelihood and Bayesian estimators (as well as the classical normal-theory regression model) implicitly assume that the treatment effect is a parameter that is constant across units (e.g., Byrk and Raudenbush 1988).

1.1 Our Contribution

While the defenses described above are reasonable, all are predicated on the premise that permutation tests are valid only as tests of a sharp null hypothesis. As such, they do not directly address the concerns of critics who regard sharp nulls as inherently "restrictive," "uninteresting," and "academic." We offer an alternative response to these critics, one founded on the premise that many permutation tests can in fact be interpreted as conservative tests of a weak null hypothesis—that is, one under which unit-level effects are heterogeneous and need not be precisely specified. Specifically, we prove that for a broad class of permutation tests, one-sided rejection of the sharp null that all treatment effects τ_i equal some constant τ^0 also implies rejection of *any* null hypothesis under which the τ_i are bounded on one side by τ^0. This means one can test the *non-superiority* null that all effects are less than or equal to τ^0 ($\tau_i \leq \tau^0 \; \forall i$), or alternatively the *non-inferiority* null $\tau_i \geq \tau^0 \; \forall i$. We refer to these as *bounded* null hypotheses.

In contrast to the sharp null of no effects, bounded null hypotheses are often quite plausible *a priori*. They are also theoretically and normatively important, particularly due to their close connection with the concept of a Pareto improvement. For a treatment or intervention to be Pareto improving, it must make at least one person better off while hurting no one. Thus, rejecting the bounded null hypothesis that all effects are greater than or equal to zero implies rejection of the hypothesis that a treatment is Pareto improving. Bounded null hypotheses are also crucial to instrumental variables estimation, whose standard monotonicity assumption holds that the instrument has non-negative or non-positive effects on treatment.

We show that any permutation test of a given sharp null will be valid under the corresponding bounded null, as long as it employs an *effect-increasing* test statistic. Loosely speaking, an effect-increasing test statistic is one that increases in value as the treated responses increase and the control responses decrease.[5] Although some test statistics, such as

_____

5. We borrow the term *effect increasing* test statistics from Rosenbaum (2002, 37–8), who discusses them in the context of power against particular alternatives under the sharp null. Our definitions of *effect increasing* differ slightly, in that Rosenbaum defines it in terms of observed outcomes and we do so in terms

the studentized statistics described by Chung and Romano (2013), are not effect increasing, many commonly used statistics are, including the difference of means and the Wilcoxon rank sum.

Just as with a t test, where rejecting the weak null $\mu = \mu_0$ against the alternative $\mu > \mu_0$ implies rejection of the composite hypothesis $\mu \leq \mu_0$, rejecting $\tau_i = \tau^0 \ \forall i$ with a one-sided permutation test implies rejection of the bounded hypothesis $\tau_i \leq \tau^0 \ \forall i$. The difference with the t test is that in the permutation case, τ^0 is not the average effect (unless effects are constant) but rather the maximum—that is, the upper bound of the range of unit-level effects. An implication of this is that it is possible to form a one-sided confidence interval for the maximum treatment effect by inverting a sequence of permutation tests for different values of τ^0 (analogous logic applies for the minimum effect). When using permutation tests sensitive to the central tendency, inferences on the upper bound will coincide closely with inferences for the ATE. As we demonstrate in our example application, however, the inferences can diverge substantially if one uses a test statistic sensitive to extreme treatment effects. Indeed, even if the ATE is significantly positive, it is nevertheless possible to conclude that treatment had a negative effect on at least one unit.

In sum, we offer a novel reinterpretation of Fisherian inference. We show that permutation tests do much more than merely assess whether treatment had any effect at all. Not only are they are valid tests of a much less restrictive null hypothesis than is commonly understood, but they also (unlike ATE estimators) can be used to draw inferences about the maximum or minimum treatment effect, a quantity that is often of normative or theoretical interest. All this is possible without asymptotic approximations or any additional assumptions beyond random assignment and SUTVA (Rubin 1980). Thus, scholars analyzing small samples from unknown probability distributions need not resort to dubious assumptions in order to draw scientifically interesting inferences about treatment effects. Rather, they can use permutation tests with statistical confidence and without apology.

---

of potential outcomes.

2 Illustration

At the broadest level, Fisher and Neyman shared the same goal: making inferences about the effects of treatment on a given sample of units, based solely on the assumption that treatment was randomly assigned.[6] In Neyman's (1923) terms, both were interested in the differences in potential outcomes under treatment and control, $\tau_i = Y_i(1) - Y_i(0)$, for n units indexed by i. Under the assumption that the potential outcomes $Y_i(1)$ and $Y_i(0)$ depend only on i's own observed treatment status,[7] these quantities of interest are fully defined by the *potential-outcome schedule* $\mathbb{S}$, which specifies all the potential outcomes in the sample.[8] Drawing inferences about the unobserved elements of the potential-outcome schedule is the core task of causal inference.

As illustration, consider a sample of 16 units, 8 of which have been randomly assigned to treatment ($W_i = 1$) and 8 to control ($W_i = 0$). Table 1 presents what we know about the sample. For each unit, only one potential outcome is observed; the other potential outcome is missing, and so is each unit's treatment effect. Suppose that we are interested in assessing the alternative hypothesis that units were positively affected by the treatment. Based on the observed outcomes, we calculate a treated–control difference of means of $t^{\text{obs}} = \bar{Y}_1 - \bar{Y}_0 = +1.13$. How unlikely is a difference of means this large, relative to what would be expected by chance?

Answering this question requires comparing t^{obs} to its reference distribution under some null hypothesis H_0. In the Fisherian paradigm, H_0 consists of an n-vector $\boldsymbol{\tau}^0$ of treatment effects, based on which we can create a null potential-outcome schedule $\tilde{\mathbb{S}}_{\boldsymbol{\tau}^0}$ with the missing potential outcomes filled in. Using the imputed $\tilde{\mathbb{S}}_{\boldsymbol{\tau}^0}$, we can "re-run" our experiment on the same units and calculate the test statistics that would have been observed under alternative permutations of treatment assignment. The collection of these values across permutations

---

6. Since we focus on finite-sample inference, which regards the potential outcomes as fixed rather than random, we set aside Neyman's additional interest in the population (as opposed to sample) ATE.

7. This is known as the stable unit treatment value assumption, or SUTVA (Rubin 1980).

8. This is what Rubin (2005, 323–4) calls the "science" table. We use "potential-outcome schedule" to echo Freedman's (2009) term "response schedule."

i	W_i	Y_i	$Y_i(0)$	$Y_i(1)$	τ_i
1	0	−0.90	−0.90	?	?
2	0	0.18	0.18	?	?
3	0	1.59	1.59	?	?
4	0	−1.13	−1.13	?	?
5	0	−0.08	−0.08	?	?
6	0	0.13	0.13	?	?
7	0	0.71	0.71	?	?
8	0	−0.24	−0.24	?	?
9	1	2.98	?	2.98	?
10	1	0.86	?	0.86	?
11	1	1.42	?	1.42	?
12	1	1.98	?	1.98	?
13	1	0.61	?	0.61	?
14	1	−0.04	?	−0.04	?
15	1	2.78	?	2.78	?
16	1	−1.31	?	−1.31	?

Table 1. The potential-outcome schedule $\mathbb{S}$ for our 16-unit illustration.

constitutes the statistic's reference distribution under the null hypothesis, conditional on the observed data. The proportion of permutations with a value of the test statistic at least as large as t^{obs} is the p-value under H_0.

This procedure is most straightforward for the sharp null of no effect. Under this hypothesis, $\tau_i = 0\ \forall i$, so the missing potential outcomes imputed under this hypothesis simply equal the observed outcomes (see Table 2, columns 4–5). However, the same procedure may be used for any arbitrary vector $\boldsymbol{\tau}^0$. If $W_i = 1$, we simply impute the missing $Y_i(0)$ as $\tilde{Y}_i(0)Y_i - \tau_i^0$; likewise, if $W_i = 0$ we impute $\tilde{Y}_i(1) = Y_i + \tau_i^0$. Columns 7–9 of Table 2 illustrate this procedure for a constant-effect null of $\tau_i = -1\ \forall i$, and columns 10–12 do so for a bounded null under which most effects are 0 but two are negative. For any such sharp null, we can generate the reference distribution by repeatedly permuting the treatment variable W, determining potential outcomes that would have been observed under that treatment assignment, and calculating the value of the test statistic in each permutation.[9]

9. See, for example, Ding, Feller, and Miratrix (2016, 660). It is worth noting that this procedure differs slightly from that described by Rosenbaum (e.g., 2002, 44), who instead proposes testing whether the imputed potential outcomes under control, $\tilde{Y}_i(0) = Y_i - W_i\tau_i^0$, satisfy the sharp null of no effects. A disadvantage of the Rosenbaum procedure is its arbitrary choice of $\tilde{Y}_i(0)$ rather than $\tilde{Y}_i(1)$ as a baseline.

Observed Data			H_0: No Effects			H_0: Constant Effect			H_0: Non-Superiority		
i	W_i	Y_i	$\tilde{Y}_i(0)$	$\tilde{Y}_i(1)$	τ_i^0	$\tilde{Y}_i(0)$	$\tilde{Y}_i(1)$	τ_i^0	$\tilde{Y}_i(0)$	$\tilde{Y}_i(1)$	τ_i^0
1	0	-0.90	-0.90	-0.90	0	-0.90	-1.90	-1	-0.90	-0.90	0
2	0	0.18	0.18	0.18	0	0.18	-0.82	-1	0.18	-1.82	-2
3	0	1.59	1.59	1.59	0	1.59	0.59	-1	1.59	1.59	0
4	0	-1.13	-1.13	-1.13	0	-1.13	-2.13	-1	-1.13	-1.13	0
5	0	-0.08	-0.08	-0.08	0	-0.08	-1.08	-1	-0.08	-0.08	0
6	0	0.13	0.13	0.13	0	0.13	-0.87	-1	0.13	0.13	0
7	0	0.71	0.71	0.71	0	0.71	-0.29	-1	0.71	0.71	0
8	0	-0.24	-0.24	-0.24	0	-0.24	-1.24	-1	-0.24	-0.24	0
9	1	2.98	2.98	2.98	0	3.98	2.98	-1	2.98	2.98	0
10	1	0.86	0.86	0.86	0	1.86	0.86	-1	0.86	0.86	0
11	1	1.42	1.42	1.42	0	2.42	1.42	-1	1.42	1.42	0
12	1	1.98	1.98	1.98	0	2.98	1.98	-1	2.98	1.98	-1
13	1	0.61	0.61	0.61	0	1.61	0.61	-1	0.61	0.61	0
14	1	-0.04	-0.04	-0.04	0	0.96	-0.04	-1	-0.04	-0.04	0
15	1	2.78	2.78	2.78	0	3.78	2.78	-1	2.78	2.78	0
16	1	-1.31	-1.31	-1.31	0	-0.31	-1.31	-1	-1.31	-1.31	0
$t^{\mathrm{obs}} = +1.13$			$p = 0.040$			$p = 0.002$			$p = 0.027$		

Table 2. Potential-outcome schedules imputed under the sharp null hypotheses of no effects (columns 4–6), a constant effect of -1 (columns 7–9), and non-superiority (columns 10–12). The bottom row lists the observed difference of means (t^{obs}) and its one-sided permutation p-values under the three null hypotheses.

The bottom row of Table 2 first gives t^{obs}, the observed difference of means, and then lists the p-values of this statistic under each sharp null hypothesis. Notice that the p-values under the constant-effect null and the bounded null are both smaller than the p-value under the null of no effects whatsoever. This is no coincidence. Rather, as we later prove, the p-value under *any* sharp null that satisfies $\tau_i \leq 0 \; \forall i$ is guaranteed to be no larger than the p-value under the sharp null of no effects ($\tau_i = 0 \; \forall i$). This result follows from the fact that the reference distribution generated under the null of no effects weakly stochastically dominates the distribution under any non-superiority null. This fact is illustrated visually in Figure 1, which plots the reference distribution for the no-effects null (solid black line), the constant-


This choice can affect the results of the test if, for example, the difference of means is the test statistic and null stipulates heterogeneous additive effects. For intuition on this point, observe that the vector $\tilde{\mathbf{Y}}(1)$ imputed under the non-superiority null (Table 2, column 11), because it incorporates the -2 treatment effect for unit 2, deviates more strongly from the sharp null than does $\tilde{\mathbf{Y}}(0)$, which incorporates only the smaller -1 effect for unit 12. The exact p-values under the no-effects null are 0.025 for $\tilde{\mathbf{Y}}(1)$ and 0.035 for $\tilde{\mathbf{Y}}(0)$. In expectation, either test is valid, but for any realized treatment assignment their results can differ.
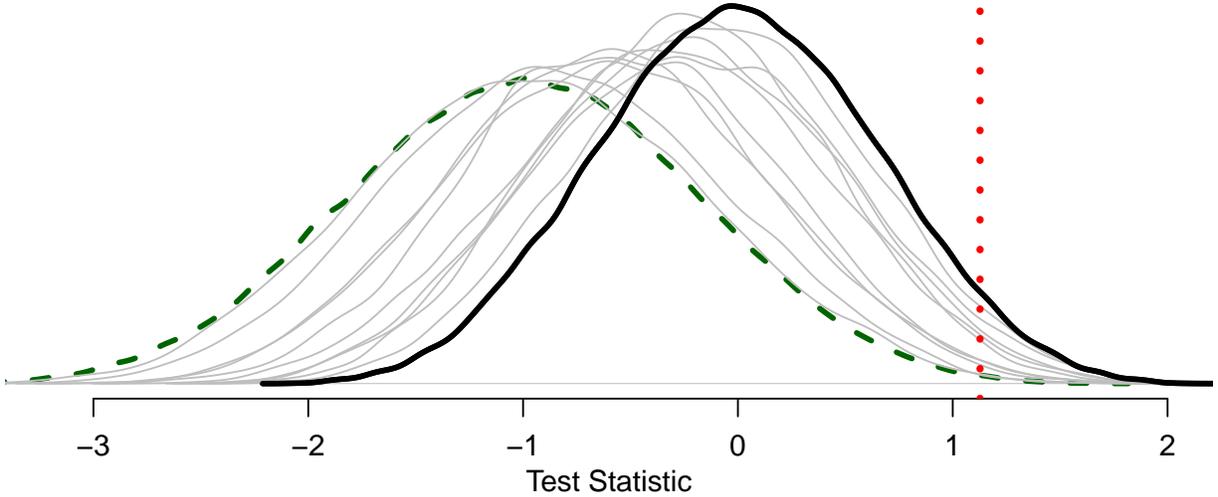
Figure 1. Reference distributions under different null hypotheses. The dark black line is the reference distribution of $t = \bar{Y}_1 - \bar{Y}_0$ under the sharp null hypothesis of no effects. The red dotted line denotes the observed t^{obs}. The p-value is the area under the curve to the right of this, and as it is small we would reject the null. The dashed green line is the reference distribution under the null hypothesis of a constant treatment effect of $\tau^0 = -1$. The grey lines are a sample of 10 possible reference distributions for ten different non-superiority nulls of no positive effects. They are all stochastically lower than the sharp zero null, and thus have lower p-values.

effects null of -1, and ten randomly generated nulls with heterogeneous treatment effects between -1 and 0. Note that the density lines of the heterogeneous-effect nulls are all to the left of the solid no-effects density line and to the right of the dashed constant-effect line. Consequently, the cumulative density greater than t^{obs} (vertical dotted line)—that is, the p-value—is largest for the no-effect null, smallest for the constant-effect null, and somewhere in the middle for each of the heterogeneous nulls.

This result has several implications. First, it means that if the sharp null hypothesis of no effect $\tau_i = 0 \ \forall i$ can be rejected at level α, so can any null hypothesis such that $\tau_i \leq 0 \ \forall i$. In other words, tests of the no-effect null are conservative tests of the more general bounded null. (See Figure 2 for a visual representation of the relationship between the Fisher's no-effect null and the bounded null.) Moreover, this result extends to any constant-effect null: Rejection of $H_0 : \tau_i = \tau^0 \ \forall i$ implies rejection of $H_0 : \tau_i \leq \tau^0 \ \forall i$. This observation is particularly important for confidence intervals (CIs), which in the permutation framework are defined as the collection of sharp null hypotheses not rejected at a given significance level.
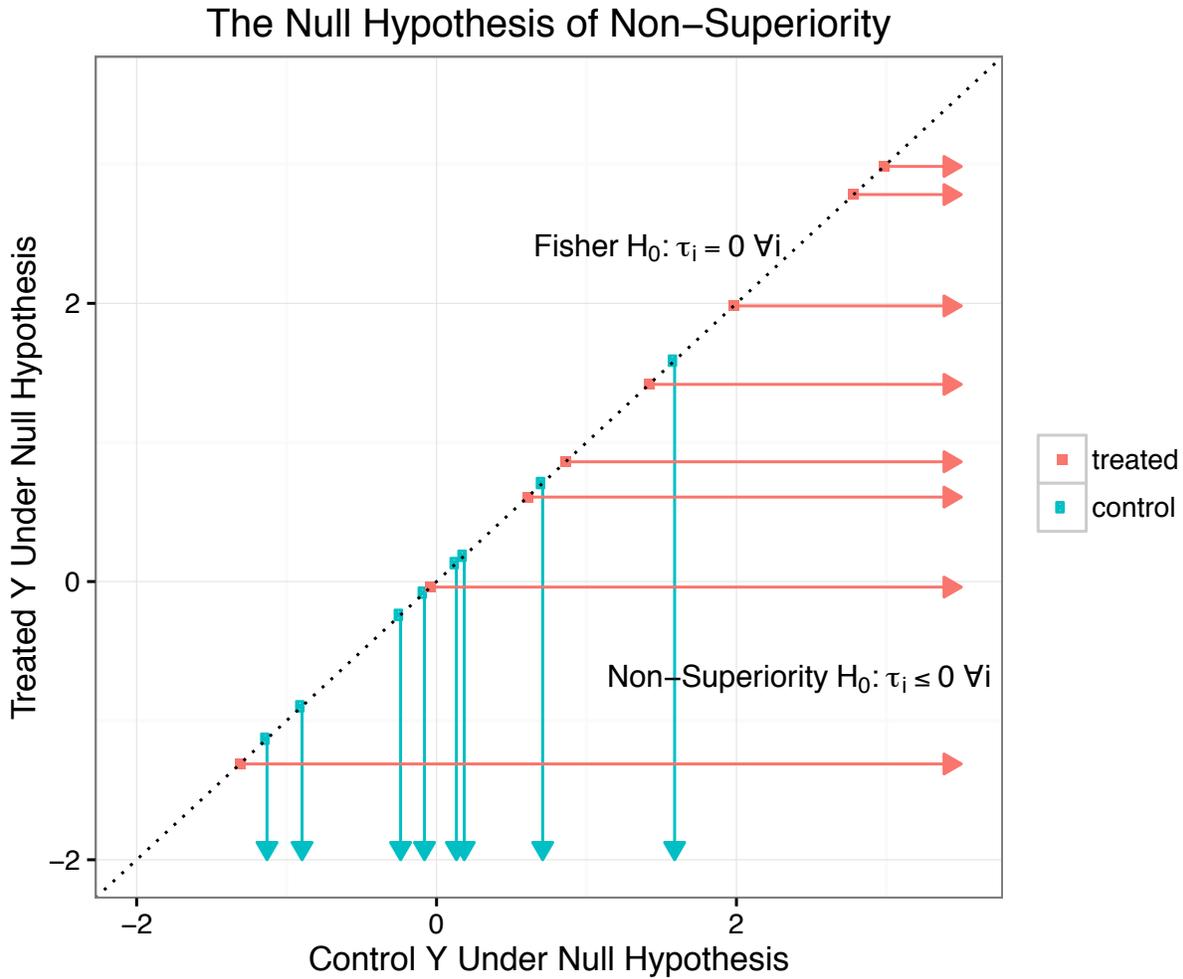
Figure 2. The null hypothesis of non-superiority. The horizontal and vertical axis indicate, respectively, $\tilde{Y}_i(1)$ and $\tilde{Y}_i(0)$: the treated and control potential outcomes imputed under the null hypothesis. The dotted $45°$ line represents Fisher's sharp null of no effects, under which $\tilde{Y}_i(1) = \tilde{Y}_i(0)$. The red squares indicate the observed treated outcomes $Y_i(1)$, and the horizontal red lines indicate their possible values of $\tilde{Y}_i(0)$ under the bounded null. The blue circles and vertical blue lines indicate the $Y_i(0)$ and possible $\tilde{Y}_i(1)$ for units actually assigned to control.

Typically, permutation CIs are calculated under a constant-effect assumption, but the above result suggests an alternative interpretation that does not require this assumption. Under this re-interpretation, a one-sided permutation CI for a constant effect is also a valid CI for the lower bound on the maximal unit-level effect, τ^{max}. Thus, an α-level CI of $[\tau_\ell, \infty)$ will miss the true τ^{max} with probability α, and we can conclude with $100 \times (1 - \alpha)\%$ confidence that at least some units had a treatment effect as large as τ_ℓ.

For intuition, consider the example of a newly developed drug, whose side effects on pain we wish to compare to those of an existing drug. In particular, we wish to assess whether the new drug increases any subject's pain level Y_i relative to the existing drug. Given random assignment to the existing drug ($W_i = 0$) and the new ($W_i = 1$), we can do so using an appropriate one-sided permutation test of the null hypothesis $\mathrm{H}_0 : Y_i(1) - Y_i(0) \equiv \tau_i = \tau^0 \; \forall i$ for a sequence of τ^0 values. If τ_ℓ, the smallest value of τ^0 that cannot be rejected at $\alpha = 0.05$, is greater than 0, then we can conclude with 95% confidence that the new drug caused at least one subject at least τ_ℓ more pain than the existing drug would have. In other words, we can conclude not only that the sharp null of no effects is implausible, but that at least one subject was adversely affected by the new drug.

Having illustrated the intuition behind our argument, we now turn to formal exposition. We first prove that an effect-increasing permutation test of a given sharp null $\boldsymbol{\tau} = \boldsymbol{\tau}^0$ is also a valid test of the general bounded null that the unit-level treatment effects are bounded on one side by the vector $\boldsymbol{\tau}^0$. We then show that without further assumptions this result can be exploited to derive confidence intervals for the maximum (or minimum) effect. Finally, we prove that several commonly used test statistics are effect increasing, explain why some others are not, and discuss power considerations in the selection of test statistics. Following this formal exposition, we turn to a discussion of the theoretical and normative status of bounded null hypotheses, and finally illustrate with an empirical application.

3 Proof of Validity under the Bounded Null

As noted above, permutation tests are typically conducted under a sharp null hypothesis that precisely specifies the n-vector of unit-level treatment effects $\boldsymbol{\tau}$. Let $\mathrm{H}_{\boldsymbol{\tau}^0}$ denote the sharp null hypothesis $\boldsymbol{\tau} = \boldsymbol{\tau}^0$, where $\boldsymbol{\tau}^0$ is a vector of hypothesized treatment effects, not necessarily equal to 0 or constant. Testing such a sharp null hypothesis entails first choosing a test statistic $T(W, Y)$, which is a function of the treatment assignment vector W and the observed outcomes Y.[10] Since Y is itself a function of W and the potential outcomes $Y(0)$ and $Y(1)$,[11] we can also write the test statistic as $T(W, Y(0), Y(1)) = T(W, \mathbb{S})$. $\mathbb{S}$ being fixed, the randomness in $T(W, \mathbb{S})$ comes only from the randomness in W. Examples of test statistics include the treated-control difference of means,

$$T(W, \mathbb{S}) = \sum W_i Y_i(1) / \sum W_i - \sum (1 - W_i) Y_i(0) / \sum (1 - W_i), \qquad (1)$$

but there are many other options, several of which we discuss later in the paper.

The observed value of the test statistic is $t^{\mathrm{obs}} = T(\boldsymbol{w}^{\mathrm{obs}}, \boldsymbol{y}^{\mathrm{obs}})$, where $\boldsymbol{w}^{\mathrm{obs}}$ and $\boldsymbol{y}^{\mathrm{obs}}$ are, respectively, the observed treatment and outcome vectors. To evaluate whether a test statistic value as large as t^{obs} would be unusual under the null hypothesis, we compare it to its permutation distribution under the null. Being sharp, $\mathrm{H}_{\boldsymbol{\tau}^0}$ allows us to impute the potential-outcomes schedule under the null, $\tilde{\mathbb{S}}_{\boldsymbol{\tau}^0} = \tilde{\mathbb{S}}(\boldsymbol{w}^{\mathrm{obs}}, \boldsymbol{y}^{\mathrm{obs}}, \mathrm{H}_{\boldsymbol{\tau}^0})$, through the relations

$$\tilde{Y}_i(0) = \begin{cases} w_i = 0 & y_i^{\mathrm{obs}} \\ w_i = 1 & y_i^{\mathrm{obs}} - \tau_i^0 \end{cases}$$

---

10. We are simplifying the exposition by considering univariate outcomes and by ignoring covariates. More generally, however, a test statistic may be a function of multiple outcome variables as well as of covariates.
11. Specifically, $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$.

and

$$\tilde{Y}_i(1) = \begin{cases} w_i = 0 & y_i^{\text{obs}} + \tau_i^0 \\ \\ w_i = 1 & y_i^{\text{obs}}. \end{cases}$$

Using $\tilde{\mathbb{S}}_{\boldsymbol{\tau}^0}$, we can then compute the test statistic value that would have been observed under any alternative realization of W. Let $\boldsymbol{w}^*$ denote a random draw from the space of potential treatment assignments, and let $t^* = T(\boldsymbol{w}^*, \tilde{\mathbb{S}}_{\boldsymbol{\tau}^0})$ be the value of the test statistic given $\boldsymbol{w}^*$ and $\tilde{\mathbb{S}}_{\boldsymbol{\tau}^0}$.[12] The p-value of t^{obs} under $\mathrm{H}_{\boldsymbol{\tau}^0}$ is thus the probability across permutations of observing a test statistic value at least as large as t^{obs}:

$$p_{\boldsymbol{\tau}^0} \equiv \mathbf{P}\left\{T(\boldsymbol{w}^*, \tilde{\mathbb{S}}_{\boldsymbol{\tau}^0}) \geq t^{\text{obs}}\right\}. \tag{2}$$

This exact p-value can be estimated to arbitrary precision by sampling J random permutations of treatment $\boldsymbol{w}_j^*$ and calculating

$$\hat{p}_{\boldsymbol{\tau}^0} = \frac{1}{J}\sum_{j=1}^{J} \mathbf{1}\left\{T(\boldsymbol{w}_j^*, \tilde{\mathbb{S}}_{\boldsymbol{\tau}^0}) \geq t^{\text{obs}}\right\} \approx p_{\boldsymbol{\tau}^0}. \tag{3}$$

We next show that, for a class of *effect-increasing* test statistics, a test of the sharp null is actually a valid test of the much more general bounded null hypothesis, under which the treatment effects are bounded on one side by the sharp null but otherwise may be arbitrarily heterogeneous. When the test statistic is an increasing function of the treatment effects, such as the difference of means, this bounded null is one of *non-superiority*—that is, one bounded above by the sharp null:

$$\mathrm{H}_{\boldsymbol{\tau}^\vee} : \tau_i \leq \tau_i^\vee \equiv \tau_i^0 \quad \forall i \in 1\ldots n. \qquad \text{(Non-Superiority)}$$

---

12. The assignment vector W is random according to an assignment mechanism which returns a specific $\boldsymbol{w}$ with given probability $\mathbf{P}\{W = \boldsymbol{w}\}$. For instance, in the completely randomized design, $\mathbf{P}\{W = \boldsymbol{w}\} = \binom{N}{N_T}^{-1}$ for any $\boldsymbol{w}$ such that $\sum w_i = N_T$ for some pre-specified N_T. In most typical experiments, all treatment assignments that have non-zero probability are equiprobable.

Analogously, when the test statistic is decreasing in treatment effects, the weak null is one of *non-inferiority*, on which the sharp null is a lower bound:

$$H_{\boldsymbol{\tau}^\wedge} : \tau_i \geq \tau_i^\wedge = \tau_i^0 \quad \forall i \in 1 \dots n. \qquad \text{(Non-Inferiority)}$$

For ease of exposition we focus in this section on the non-superiority null $H_{\boldsymbol{\tau}^\vee}$, which we simply refer to as the bounded null, but all our results can be extended to $H_{\boldsymbol{\tau}^\wedge}$ by multiplying Y by -1.

Unlike the sharp null $H_{\boldsymbol{\tau}^0}$, which corresponds to a single null potential-outcomes schedule $\tilde{\mathbb{S}}_{\boldsymbol{\tau}^0}$, the bounded null $H_{\boldsymbol{\tau}^\vee}$ corresponds to an infinitely large set of such schedules that are consistent with the observed data and the treatment-effect bound (e.g., every set of points on the lines in Figure 2). Thus, if we reject this null, we are rejecting a set of null schedules (and associated treatment-effect vectors) rather than a single point null. Because each weak null permits infinitely many possible potential-outcome schedules, each of which implies a different null distribution for the test statistic, no single p-value will be exact (i.e., have a false-positive rate of exactly α) for all possible schedules. We can show, however, that for a class of test statistics the p-value associated with any of these null distributions will be bounded above by the p-value under $H_{\boldsymbol{\tau}^0}$, making the sharp null p-value valid but possibly conservative for the bounded null. (A hypothesis test is *conservative* if, for any nominal significance level α, the true probability of incorrectly rejecting the null hypothesis is no greater than α.)

The above property holds for permutation tests that employ an effect-increasing test statistic. To define this class of statistics, we must first introduce the notion of ordering potential-outcome schedules:

Definition: Ordering of Potential-Outcome Schedules. Two potential-outcome schedules $\mathbb{S}$ and $\mathbb{S}'$ are ordered as $\mathbb{S} \preceq \mathbb{S}'$ if $Y_i(1) \leq Y_i'(1)$ and $Y_i(0) \geq Y_i'(0)$ $\forall i \in 1 \dots n$. That is, $\mathbb{S} \preceq \mathbb{S}'$ if and only if no unit's potential outcome under treatment is smaller in $\mathbb{S}'$ than

in $\mathbb{S}$ and no unit's potential outcome under control is larger in $\mathbb{S}'$ than in $\mathbb{S}$. An immediate consequence of such ordering is that the individual treatment effects are also ordered: $\tau_i \leq \tau_i' \; \forall i$.

Our class of statistics is then defined as those that satisfy the following:

Definition: Effect Increasing (EI). A test statistic T is *effect increasing* if, for two potential-outcome schedules $\mathbb{S}$ and $\mathbb{S}'$, $\mathbb{S} \preceq \mathbb{S}'$ implies $T(\boldsymbol{w}, \mathbb{S}) \leq T(\boldsymbol{w}, \mathbb{S}')$ for all allowed realizations $\boldsymbol{w}$ of the treatment variable W. In other words, a test statistic T is effect increasing if it is weakly increasing in the potential outcomes under treatment and weakly decreasing in the potential outcomes under control (cf. Rosenbaum 2002, 37–8). Since $\mathbb{S} \preceq \mathbb{S}'$ implies $\tau_i \leq \tau_i' \; \forall i$, an EI statistic is also increasing in the individual treatment effects (hence the label "effect increasing").

For effect-increasing statistics we have the following proposition:

Proposition 1. If $\boldsymbol{\tau}_0 = \boldsymbol{\tau}^\vee$, a permutation test of $\mathrm{H}_{\boldsymbol{\tau}^0}$ is a conservative (and thus valid) test of $\mathrm{H}_{\boldsymbol{\tau}^\vee}$ in that for all $\mathbb{S}_h \in \mathrm{H}_{\boldsymbol{\tau}^\vee}$

$$\mathbf{P}\{\text{Reject } \mathrm{H}_{\boldsymbol{\tau}^\vee} | \mathbb{S}_h\} = \mathbf{P}\Big\{ T(\boldsymbol{w}^*, \tilde{\mathbb{S}}_{\boldsymbol{\tau}^0}) \geq T(\boldsymbol{w}, \mathbb{S}_h) \Big\} \leq \alpha,$$

with the probabilities taken across both $\boldsymbol{w}$ and $\boldsymbol{w}^*$, each random draws from the assignment mechanism.

Proof: Let $\mathbb{S}_h \in \mathrm{H}_{\boldsymbol{\tau}^\vee}$ be any potential-outcomes schedule satisfying the non-superiority null hypothesis $\tau_i \leq \tau_i^\vee \equiv \tau_i^0 \; \forall i \in 1 \dots n$. Suppose $\mathbb{S}_h$ holds. We randomize the units and obtain $\mathbf{w}^{\text{obs}}$, $\mathbf{y}^{\text{obs}}$, and $t^{\text{obs}} = T(\mathbf{w}^{\text{obs}}, \mathbb{S}_h)$. We then impute $\tilde{\mathbb{S}}_{\boldsymbol{\tau}^0} = \tilde{\mathbb{S}}(\mathbf{w}^{\text{obs}}, \mathbf{y}^{\text{obs}}, \mathrm{H}_{\boldsymbol{\tau}^0})$ and obtain $\tilde{p}$, our nominal p-value. Even though $\mathrm{H}_{\boldsymbol{\tau}^\vee}$ holds, $\mathrm{H}_{\boldsymbol{\tau}^0}$ might not, and so it is possible that $\tilde{\mathbb{S}}_{\boldsymbol{\tau}^0} \neq \mathbb{S}_h$ and thus $\tilde{p} \neq p$. However, if the test statistic is effect increasing, then we can place

17

a bound on $\tilde{p}$. In particular, note that

$$\tilde{Y}_i(1) = \begin{cases} Y_i(1) & w_i^{\text{obs}} = 1 \\ Y_i(0) + \tau_i^0 & \text{otherwise} \end{cases}$$

for all $i \in 1 \ldots n$. Under $\mathrm{H}_{\boldsymbol{\tau}^\vee}$

$$\tau_i^0 \geq \tau_i = Y_i(1) - Y_i(0),$$

so

$$Y_i(1) \leq Y_i(0) + \tau_i^0$$

and thus

$$Y_i(1) \leq \tilde{Y}_i(1).$$

In other words, every true potential outcome under treatment is no larger than its imputed equivalent. By analogous logic $Y_i(0) \geq \tilde{Y}_i(0)$. These inequalities imply $\mathbb{S}_h \preceq \tilde{\mathbb{S}}_{\boldsymbol{\tau}^0}$. Since $T(W, \mathbb{S})$ is effect increasing, $T(\boldsymbol{w}, \mathbb{S}_h) \leq T(\boldsymbol{w}, \tilde{\mathbb{S}}_{\boldsymbol{\tau}^0})$ for any realization $W = \boldsymbol{w}$. In other words, because the potential-outcome schedules are ordered $\mathbb{S}_h \preceq \tilde{\mathbb{S}}_{\boldsymbol{\tau}^0}$, the values of T simulated from $\tilde{\mathbb{S}}_{\boldsymbol{\tau}^0}$ will be pointwise weakly larger than T's true reference distribution. As a consequence,

$$\tilde{p} = \mathbf{P}\Big\{ T(\boldsymbol{w}^*, \tilde{\mathbb{S}}_{\boldsymbol{\tau}^0}) \geq t^{\text{obs}} \Big\} \geq \mathbf{P}\big\{ T(\boldsymbol{w}^*, \mathbb{S}_h) \geq t^{\text{obs}} \big\} = p,$$

i.e., the estimated p-value will be at least as large as the true one. This gives a valid (though potentially conservative) test:

$$\mathbf{P}\{\tilde{p} \leq \alpha\} \leq \mathbf{P}\{p \leq \alpha\} \leq \alpha.$$

$\square$

This result means that if we consider a test of a given sharp null as a test of the associated bounded null, we still have a valid test. In particular, rejecting $\mathrm{H}_{\boldsymbol{\tau}^\vee}$ when the nominal p-value

for a permutation test of $H_{\boldsymbol{\tau}^0}$ is less than α is a valid testing procedure for $H_{\boldsymbol{\tau}^\vee}$.

4   Confidence Intervals for Maximum/Minimum Effects

For sharp null inference, confidence intervals are generally obtained by inverting a sequence of sharp-null level-α tests of hypotheses $H_{\boldsymbol{\tau}^0}$. For example, we might consider a sequence of constant shift hypotheses $H_{\tau^0} : \tau_i = \tau^0 \; \forall i$. For each candidate value $\tau_h^0 = \tau^0$ we calculate

$$p(\tau_h^0) = \mathbf{P}\Big\{ T(\boldsymbol{w}^*, \tilde{\mathbb{S}}_{\tau_h^0}) \geq t^{\mathrm{obs}} \Big\},$$

and if $p(\tau_h^0) \leq \alpha$, we conclude that τ_h^0 is implausible. This gives a confidence set of plausible τ^0 values of

$$\mathrm{CI} \equiv \big\{ \tau_h^0 : p(\tau_h^0) \geq \alpha \big\}.$$

The confidence sets are random depending on the randomization. They are valid in the sense that if the treatment effect is in fact constant ($\mathbb{S} \in H_{\tau_h^0}$ for some τ_h^0), then the confidence set CI will contain τ with probability at least $1 - \alpha$. Unfortunately, if the treatment effect is not constant, there is no immediate reason for CI to contain any particular summary of the treatment effects (e.g., the ATE). This is one of the primary complaints against permutation inference. By viewing these as tests of a bounded null, however, the associated confidence interval does in fact have a general interpretation that does not depend on the implausible assumption of constant effects.

To show this, we first need a small lemma:

Lemma 1. *For our one-sided testing case, and regardless of the character of the true $\mathbb{S}$, the one-sided CI for an effect-increasing statistic will be a half-interval $[L, \infty)$, indicating that the constant-shift treatment effect is no smaller than L.*

Proof: Say our CI is not a half-interval. Then there exists $\tau_1 < \tau_2$ such that τ_1 is not in CI and τ_2 is. But the proof of our main result shows that if we are testing $H_{\tau_2^0}$ then the p-value

will be lower for any $\mathbb{S}' \preceq \tilde{\mathbb{S}}$, including the one corresponding to a constant treatment effect shift of $\tau_1 < \tau_2$. But this means τ_1 would not be in CI, which is a contradiction. Therefore the CI is a half-interval. □

These confidence intervals can easily generalize to bounded nulls. Let CI be the above confidence set generated by inverting a sequence of constant-effect nulls. Now consider the true potential-outcomes schedule $\mathbb{S}$. Assuming all potential outcomes are defined, let

$$\tau^* \equiv \arg\max_i \tau_i$$

be the largest treatment effect in schedule $\mathbb{S}$. Then $H_{\boldsymbol{\tau}^\vee} : \tau_i \leq \tau^* \ \forall i$ is true, and so testing the associated $H_{\boldsymbol{\tau}^0} : \tau_i = \tau^* \ \forall i$ will reject with probability no greater than α. We therefore will include $\tau^* \in$ CI with probability no less than $1 - \alpha$, giving a valid CI for the maximum effect.

The typical permutation CI for a constant effect can thus be interpreted as a confidence set on the maximum treatment effect in the sample (or minimum, in the case of a non-inferiority hypothesis). In other words, given a set $[L, \infty)$ we can say that we are at least $1 - \alpha$ confident that some units have a treatment effect of at least L. This statement does not depend on any specific structure on the individual effects; we can have arbitrary heteroskedasticity. That being said, the less heterogenous the effects, the more individual effects we would expect to be in the CI. Of course, in the limiting case of no effect heterogeneity (i.e., a constant effect), $[L, \infty)$ will, as discussed above, contain all the individual effects with probability $1 - \alpha$.

It should be emphasized that the permutation CI for the maximum effect will have correct coverage regardless of the test statistic used, as long as that statistic is effect increasing. The CI itself will, however, vary depending on the test statistic's power against different alternatives. In particular, unless the treatment effects are close to constant, using a statistic sensitive to the central tendency may result in relatively uninformative confidence bounds

for the maximum. Thus, if heterogenous effects are expected, it may be preferable to use a statistic that is sensitive to the largest effects, such as the Stephenson rank sum (see below).

5   Effect-Increasing Test Statistics

We have shown that any effect-increasing test statistic is valid under the bounded null. We next show that the difference of means, the Wilcoxon rank sum, and other common test statistics are effect increasing. We also note that others, in particular the studentized difference of means, are not effect increasing. We then briefly touch on the issue of statistical power against different alternatives.

As discussed above, an effect-increasing statistic should grow as the treatment potential outcomes get larger or the control potential outcomes get smaller. The difference in means, $T(Y, W) = \bar{Y}_1 - \bar{Y}_0$, is an intuitive example of such a statistic: if we increase any $Y_i(1)$ then the difference in means will either be unchanged if $W_i = 0$, or go up if $W_i = 1$. Similarly, if we lower any $Y_i(0)$ the control side mean will either shrink, which will raise t, or stay the same. This general argument applies for several broad classes of statistics, as we present below in two lemmas. Proofs are in Appendix B.

First, statistics that look like differences in means, or sums of rescaled treatment outcomes, are all effect increasing:

Lemma 2. *Let $Q(\cdot)$ be a non-decreasing function of outcomes and $a, b \geq 0$ be two constants. Then the class of statistics defined as*

$$T(Y, W) = a \sum_{i=1}^{n} W_i Q(Y_i) - b \sum_{i=1}^{n} (1 - W_i) Q(Y_i)$$

are effect increasing. The difference-in-means is a special case with $a = 1/n_T$, $b = 1/n_C$, and $Q(Y) = Y$.

Furthermore, many common rank statistics are also effect increasing:

Lemma 3. *For continuous outcomes any statistic of the form*

$$T(W,Y) = \frac{1}{n_T}\sum_i W_i a(R_i) - \frac{1}{n_C}\sum_i (1-W_i)a(R_i),$$

where R_i is the rank of Y_i and $a(R_i)$ is some non-decreasing function of the ranks, is effect-increasing.

For many statistics this result extends to outcomes with ties. In particular, see Appendix B for derivations for the Wilcoxon rank sum statistic and Stephenson rank sum statistics (see below).

The EI property does *not* hold, however, if the difference of means is "studentized" by a consistent estimate of its standard error:

$$t = \frac{\bar{Y}_T - \bar{Y}_C}{\sqrt{s_T^2/N_T + s_C^2/N_C}}.$$

The t statistic is not EI because a large increase in one unit's treated outcome can have such a large effect on the standard deviation s_T that it outweighs the effect on the mean $\bar{Y}_T$, thus decreasing the statistic overall. More obviously, statistics not sensitive to one-sided location shifts—such as the absolute difference of means, the difference of variances, and the two-sided Kolmogorov-Smirnov statistic—are also not effect-increasing.

Any effect-increasing statistic will give a valid test of the general bounded null. They are not all the same, however. In particular, a given statistic can be more or less sensitive to different alternatives, which can influence power. In particular one class of statistics potentially well-suited to this circumstance are Stephenson rank statistics (Stephenson 1981; Stephenson and Ghosh 1985). Stephenson rank statistics are two-sample statistics with a score function of

$$a(R_i) = \begin{cases} \binom{R_i-1}{s-1} & R_i \geq s \\ 0 & \text{otherwise} \end{cases}$$

for some fixed integer $s \geq 2$. Stephenson rank statistics are equivalent to summing the number of subsets of size s in which the largest response is in the treated group. The Stephenson rank statistic with $s = 2$ is almost identical to the Wilcoxon rank sum. However, as s increases beyond 2, the Stephenson ranks place more and more weight on the largest responses.

Stephenson rank statistics are particularly interesting in the context of this paper due to their power to detect uncommon-but-dramatic responses to treatment (Rosenbaum 2007). 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.[13] Thus, compared to the difference of means and the Wilcoxon rank sum, whose power is greatest against a constant location shift, the Stephenson ranks have greater power against alternatives under which effects are heterogeneous and a few are highly positive (relative to the null). This sensitivity to extreme treatment effects leads 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 test to reject the bounded null when the ATE estimate is *negative*, if some treatment effects are sufficiently positive. Thus, when treatment effects are heterogeneous, the behavior of the Stephenson test can differ markedly from the rank sum or difference of means, while like them still providing a valid test of the bounded null hypothesis.

6 Relevance of Bounded Null Hypotheses

So far, we have shown that effect-increasing permutation tests are valid under the general bounded null; that inverting a sequence of such tests produces valid confidence intervals for the maximum effect; and that many familiar test statistics are effect increasing. We now

---

13. Examining the asymptotic relative efficiency of a closely related class of test statistics, Conover and Salsburg (1988, 196) find that when a only small fraction of treated respond, the optimal subset size s is between 5 and 6.

consider the substantive relevance of bounded null hypotheses. Such hypotheses are sometimes evoked as assumptions of statistical models, perhaps most saliently in instrumental variables estimation, which relies on the "monotonicity" assumption that the instrument's effects on treatment are non-negative (or non-positive). This rarely tested assumption could be assessed using a permutation test with a statistic sensitive to rare negative (or positive) effects, the existence of which would violate monotonicity.

In addition to technical applications such as instrumental variables, bounded null hypothesis also have more general theoretical and normative relevance. First, bounded null hypotheses can be appropriate for one of the most common questions motivating social scientists: whether some treatment has a (policy relevant sized) effect on the outcome. Typically scientists wishing to use randomization inference would limit themselves to testing the knife-edge null of absolutely zero effect for every unit ($\tau_i = 0 \; \forall i$). This null, however, is only relevant for questions in which strong plausible theory suggests there is no possibility of effects. This kind of strong theory is found in physics, for example in the hypothesis that neutrinos cannot go faster than the speed of light. By contrast for social phenomena we rarely have strong beliefs of zero causal effects for all units, except when considering treatments affected outcomes in the past (the basis for placebo tests of design assumptions; Rosenbaum 2002).

However, in many social contexts it is plausible that a treatment has at most small effects—so small that they fall within what might be called a "policy-irrelevant" interval (τ_L, τ_U). In such cases, we may wish to assess whether treatment had any "policy-relevant" effect outside this window. This can be done by testing either the bounded null $H_{\boldsymbol{\tau}^\vee} : \tau_i \leq \tau_U \; \forall i$ or $H_{\boldsymbol{\tau}^\wedge} : \tau_i \geq \tau_L \; \forall i$, or both (possibly correcting for multiple testing; see, e.g., Caughey, Dafoe, and Seawright, Forthcoming). Equivalently, one could calculate permutation CIs for the maximum and minimum effects, $[\delta_g, \infty)$ and $(-\infty, \delta_l]$, to see whether $\delta_g > \tau_U$ or $\delta_l < \tau_L$. If the same measure of central tendency, such as the difference of means, is used as a test statistic for both $H_{\boldsymbol{\tau}^\vee}$ and $H_{\boldsymbol{\tau}^\wedge}$, then the CIs will overlap (i.e., $\delta_g < \delta_l$), meaning that at most

one hypothesis can be rejected. If, however, a statistic sensitive to the tails of the treatment effect distribution, such as the Stephenson rank sum, is used, then it is possible for the CIs to be disjoint, as they are in Figure 3. In such a case, it can be concluded that treatment had at least one negative *and* at least one positive policy-relevant effect.
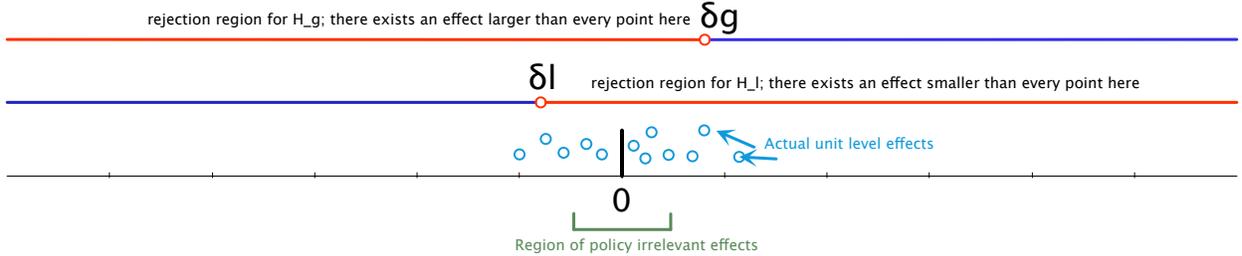
Figure 3. Two one-sided confidence intervals for maximum and minimum effects (blue lines). The confidence intervals $[\delta_g, \infty)$ and $(-\infty, \delta_l]$ do not overlap in this illustration, which is possible if they are based on a test statistic sensitive to the tails of the distribution of treatment effects (e.g., the Stephenson rank sum).

Another domain where the bounded null hypothesis is especially applicable involves testing for the existence of harm. It is a widely held ethical and moral principle that one should avoid doing harm: It is not enough for one's good and bad actions to "average out" as positive, but one must systematically avoid committing bad actions, even if that leads to less good "overall." Many people would oppose a policy intervention or new drug if it were shown to inflict harm on individuals, whether or not on average it had beneficial effects. Experimental moral philosophy, as exemplified by the "Fat Man" variant of the Trolley problem, has shown that people may perceive an action that "on average" is beneficial (saves the most lives) to be morally repugnant if it involves causing harm to a single individual. Amongst medical doctors the ancient maxim to "at least do no harm" can be interpreted as reflecting this moral aversion to committing any injury even if in the service of an expected overall benefit. Economists and others often strive for Pareto-improving policies, which by definition do not harm anyone. In tort law people are held liable for various forms of injuries that they cause to other individuals, largely irrespective of whether there are net benefits of the

action to society.

Thus, in many circumstances people and society perceive the relevant question to be whether any one individual was harmed by an action, i.e., whether an action was not Pareto-improving. This is readily expressed as a bounded null that treatment had weakly positive effects for all individuals ($\tau_i \geq 0\ \forall i$) against the alternative that some individuals were harmed ($\exists i\ \tau_i < 0$). As always with randomization inference, in implementing such a test one should use the test statistic that will be most sensitive to the alternative. If the alternative is that everyone may have been somewhat injured (say an additive effect), then a difference in means is a good choice for a powerful test statistic, even if what one is interested in detecting is the existence of any harm. However, if the most plausible alternative is that most people were benefited from the treatment, but some people suffered great harm, then one wants a test statistic most likely to detect this great harm, such as the Stephenson rank statistic.[14]

7 Application: Campaign Effects in Benin

To illustrate how these insights can improve our understanding of a real study, we re-analyze data from Wantchekon's (2003) well-known field experiment in Benin.[15] Wantchekon convinced Beninese presidential candidates to randomly vary the content of their campaign appeals in different villages, stratified by electoral district. The original experiment involved three treatments—a *clientelist* campaign, a *policy* campaign, and a *control* group exposed to both campaigns—with 8 villages per treatment group. Wantchekon hypothesized that the clientelist campaign would be more effective than the control campaign, and that the policy campaign would be less effective than control. Consistent with these hypotheses, candidates earned an average vote share of 84% in villages where they ran clientelist campaigns, 74%

---

14. To be clear, we can reject the null that an intervention is Pareto improving in favor of the alternative that it is not Pareto improving. We cannot use this framework for the reverse inference to reject the null of some harm against the alternative of Pareto improvement. In fact, without strong assumptions no method will have statistical power to detect whether *every* unit has not be harmed.

15. There is no publicly available replication dataset for this study, but the village-level data are reported in Table 2 on page 412 of Wantchekon (2003).
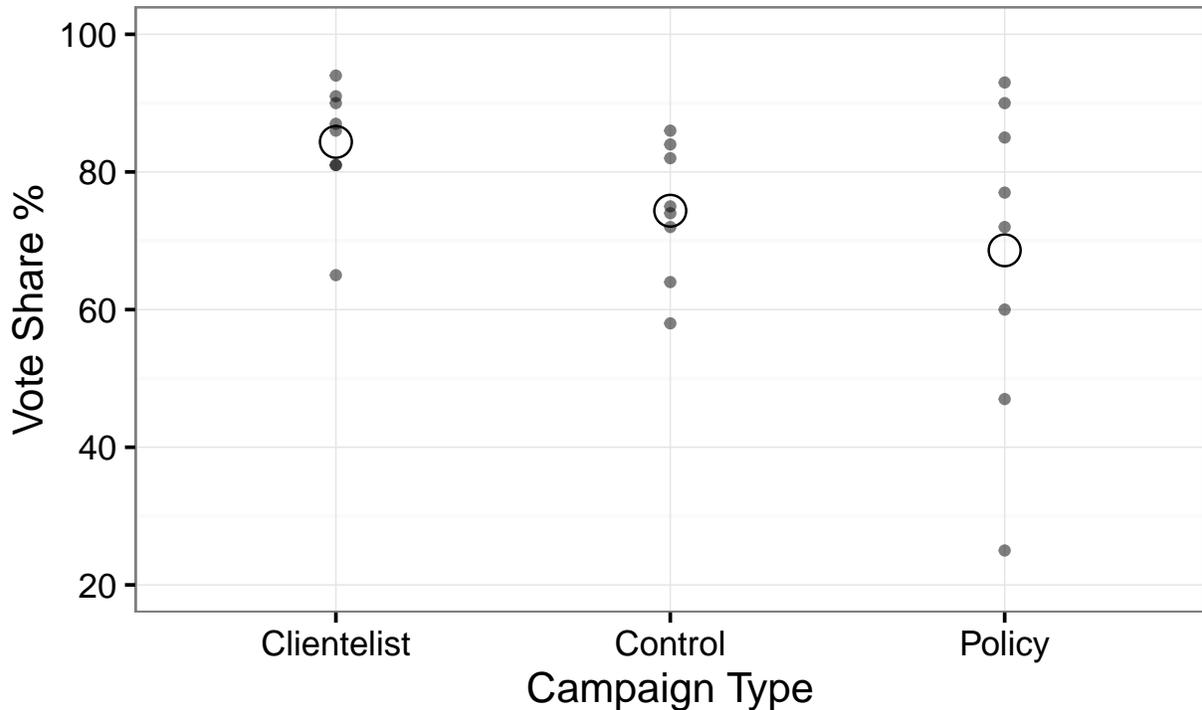
Figure 4. Data from Wantchekon (2003). Hollow circles indicate group means. Note that two villages in the clientelist condition are over-plotted because they returned the same vote share (81%).

in control villages, and 69% where they ran policy campaigns (see Figure 4). Wantchekon notes, however, that the policy campaign may have been *more* effective than the control in a few villages (2003, 413). According to the original paper all the mean differences between treatment conditions are highly statistically significant, but, as Green and Vavreck (2008) note, Wantchekon's analysis ignored village-level clustering and thus vastly overstated the precision of his estimates.

Here we use permutation inference to re-analyze Wantchekon's data at the village level.[16] We begin with a comparison of the clientelist and control conditions. As Figure 4 indicates, the distribution of outcomes in the clientelist and control conditions have about the same variance, but the control distribution is shifted 10 percentage points lower. As Table 3 indicates, a paired t test yields a p-value of 0.039 and a one-sided 90% CI for the ATE of $[3.1\%, \infty)$. A difference-of-means permutation test (permuting only within district strata)

---

16. For an earlier re-analysis of these data, see Caughey, Dafoe, and Seawright (Forthcoming).

yields almost precisely the same p-value and a slightly tighter CI of $[3.9\%, \infty)$, which is consistent with the two tests' asymptotic equivalence when group sizes are equal (Romano 1990).[17] The p-values of Wilcoxon and Stephenson rank tests (0.031 and 0.019, respectively) are slightly smaller and their CI ($[5.1\%, \infty)$ for both tests) is somewhat tighter, probably because the rank-scores discount the negative outlier in the clientelist group.

Although the large-sample t test and exact permutation tests yield similar results, the interpretations of the two sets of tests are different. The permutation test signals the implausibility not only of the sharp null of no effects, but also of the bounded null that no treatment effects were positive. In fact, from the rank-test CI of $[5.1\%, \infty)$ we can conclude with 90% confidence that in at least one village, running a clientelist-only campaign increased candidate vote share by at least 5.1 percentage points. The inferences justified by the t test are different. It provides evidence that the null of mean equality is implausible, and that we can be approximately 90% confident that clientelism increased candidates' average vote share by at least 3.1 points.

The comparison of the policy and control conditions is more nuanced. Consistent with Wantchekon's expectations, average vote share was about 6 points lower in the policy condition. This difference, however, is not statistically significant under either a t test or a difference-of-means permutation test (in both cases, $p \approx 0.22$). This not the result of outliers reducing power: a Wilcoxon-Mann-Whitney test yields a one-sided p-value of 0.48. It would thus appear that little can be inferred about the treatment effects of policy campaigns relative to control.

We can, however, say more than this by focusing on extreme effects rather than typical (e.g., average or median) ones. If treatment effects are heterogeneous, as is suggested by the difference in spread between the control and policy distributions, then using a test statistic sensitive to the most negative or positive effects can provide greater power.[18] Indeed, as Fig-

---

17. The permutation difference of means is also asymptotically equivalent to the t when the treatment and control variances are equal, which also appears to be true in these data.

18. A permutation test of the difference of variances (which, it should be noted, is not an effect-increasing statistic) indicates that of the null of distributional equality (i.e., the sharp null of no effects) can be rejected.

ure 4 indicates, the policy condition contains both the smallest *and* the largest observations in the two groups. This pattern is consistent with the possibility, suggested by Wantchekon, that in some villages policy campaigns were *more* effective than the control. We can assess this hypothesis formally using the Stephenson rank test, which is sensitive to large-but-rare effects. Testing the null of no effects against the alternative of a few large positive effects, we obtain a p-value of 0.086, providing suggestive evidence that the policy campaign treatment was indeed more effective in at least some villages.[19] The 90% CI for the largest effect of the policy treatment is $[1.1\%, \infty)$. If we instead test for large-but-rare *negative* effects, we find weaker evidence against the bounded null ($p = 0.125$). Thus, despite the fact that the estimated ATE of the policy treatment is negative (though far from statistically significant), there is actually stronger evidence that the policy campaigns increased vote share in at least one village than that it decreased any village's share. This inference is made possible by combining our new interpretation of permutation tests with a test statistic specifically suited to maximize our power to detect extreme effects.

Alternative Hypothesis	Statistical Test	p	90% CI
Control < Clientelist	Paired t (Asymptotic)	0.04	$[+3.1, +\infty)$
Control < Clientelist	Difference-of-Means (Exact)	0.04	$[+3.9, +\infty)$
Control < Clientelist	Wilcoxon Rank (Exact)	0.03	$[+5.1, +\infty)$
Control < Clientelist	Stephenson Rank (Exact)	0.02	$[+5.1, +\infty)$
Policy < Control	Paired t (Asymptotic)	0.23	$[-4.4, +\infty)$
Policy < Control	Difference-of-Means (Exact)	0.22	$[-3.6, +\infty)$
Policy < Control	Wilcoxon Rank (Exact)	0.48	$[-7.9, +\infty)$
Policy < Control	Stephenson Rank (Exact)	0.13	$[-4.0, +\infty)$
Control < Policy	Paired t (Asymptotic)	0.77	$[-15.9, +\infty)$
Control < Policy	Difference-of-Means (Exact)	0.79	$[-14.5, +\infty)$
Control < Policy	Wilcoxon Rank (Exact)	0.54	$[-13.9, +\infty)$
Control < Policy	Stephenson Rank (Exact)	0.09	$[+1.1, +\infty)$

Table 3. Results of Different Analyses of the Benin Dataset.

---

19. This is for the Stephenson ranks with subset size $m = 6$, but p-values for $m = 7, 8, 9$ and 10 are also around 0.09.

8　Conclusion

The rise of nonparametric causal inference in the tradition of Neyman (1923) and Rubin (1974) has been one of the most important recent developments in quantitative social science. This perspective, with its focus on average treatment effects and its acceptance of effect heterogeneity as the rule rather than the exception, has rightly prompted greater skepticism of statistical methods that heavily rely on parametric models or assumptions. It is then perhaps no surprise that permutation inference, which has traditionally been motivated in terms of shift hypotheses or other highly structured models of treatment effects (e.g., Lehmann 1975; Rosenbaum 2002), has shared in this skepticism despite its freedom from many distributional assumptions.

What we have sought to demonstrate in this paper is that the view of permutation tests now dominant among political methodologists—that despite these tests' virtues, they are 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 permutation tests using effect-increasing test statistics are valid under a more general bounded null, that this fact can be exploited to derive confidence intervals for the maximum effect, and that many familiar test statistics are effect increasing. 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 of bounded null hypotheses. Finally, we have re-analyzed a well-known experiment to demonstrate that, when coupled with the new interpretation we have advanced, permutation tests can yield substantively interesting inferences about treatment effects that are not possible based on ATE estimation alone.

In sum, we have developed a novel perspective on permutation tests that we hope tempers the skepticism that many political methodologists hold towards this otherwise-appealing mode of statistical inference. Permutation tests are by no means a cure-all; nor are they

a substitute for ATE estimation when that is the goal of the analysis. But in many cases, particularly when samples are small, treatment groups unequal, or treatment assignment complex, they are the most reliable form of statistical inference. Moreover, even when this is not the case, they often make possible inferences about treatment effects that other methods cannot. For these reasons, permutation tests deserve a secure place in the quantitative social scientists' toolbox.

References

Bowers, Jake, Mark M. Fredrickson, and Costas Panagopoulos. 2013. "Reasoning about Interference Between Units: A General Framework." *Political Analysis* 21 (1): 97–124.

Byrk, Anthony S., and Stephen W. Raudenbush. 1988. "Heterogeneity of Variance in Experimental Studies: A Challenge to Conventional Interpretations." *Psychological Bulletin* 104 (3): 396–404.

Caughey, Devin, Allan Dafoe, and Jason Seawright. Forthcoming. "Nonparametric Combination (NPC): A Framework for Testing Elaborate Theories." *Journal of Politics.*

Chung, EunYi, and Joseph P. Romano. 2013. "Exact and Asymptotically Robust Permutation Tests." *Annals of Statistics* 41 (2): 484–507.

Conover, William J., and David S. Salsburg. 1988. "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.

Ding, Peng, Avi Feller, and Luke Miratrix. 2016. "Randomization inference for treatment effect variation." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78 (3): 655–671.

Fisher, Ronald A. 1935. *Design of Experiments.* Edinburgh: Oliver / Boyd.

Freedman, David A. 2009. *Statistical Models: Theory and Practice.* Revised. New York: Cambridge UP.

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

Green, Donald P., and Lynn Vavreck. 2008. "Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches." *Political Analysis* 16 (2): 138–152.

Hesterberg, Tim C. 2015. "What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum." *The American Statistician* 69 (4): 371–386.

Imai, Kosuke. 2013. "Statistical Hypothesis Tests." Lecture note. Accessed May 18, 2016. Department of Politics, Princeton University, March 24. `http://imai.princeton.edu/teaching/files/tests.pdf`.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction.* New York: Cambridge UP.

Keele, Luke. 2015. "The Statistics of Causal Inference: A View from Political Methodology." *Political Analysis* 23 (3): 313–335.

Lehmann, E. L. 1963. "Nonparametric Confidence Intervals for a Shift Parameter." *Annals of Mathematical Statistics* 34 (4): 1507–1512.

———. 1975. *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day.

Neyman, Jerzy. 1923. "On the application of probability theory to agricultural experiments. Essay on principles. Section 9." *Roczniki Nauk Roiniczych, Tom X:* 1–51. Reprinted in *Statistical Science*, 5 (4): 465–472, 1990. Translated from Polish by D. M. Dabrowska and T. P. Speed.

———. 1935. "Statistical Problems in Agricultural Experimentation." *Supplement to the Journal of the Royal Statistical Society* 2 (2): 107–180.

Romano, Joseph P. 1990. "On the Behavior of Randomization Tests Without a Group Invariance Assumption." *Journal of the American Statistical Association* 85 (411): 686–692.

Rosenbaum, Paul R. 2002. *Observational Studies.* 2nd. New York: Springer.

———. 2007. "Confidence Intervals for Uncommon but Dramatic Responses to Treatment." *Biometrics* 63 (4): 1164–1171.

———. 2010. *Design of Observational Studies.* New York: Springer.

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

———. 1980. "Randomization Analysis of Experimental Data: The Fisher Randomization Test: Comment." *Journal of the American Statistical Association* 75 (371): 591–593.

———. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association* 100 (469): 322–331.

Samii, Cyrus, and Peter M. Aronow. 2012. "On Equivalencies Between Design-based and Regression-based Variance Estimators for Randomized Experiments." *Statistics and Probability Letters* 82 (2): 365–370.

Stephenson, Robert W., and Malay Ghosh. 1985. "Two Sample Nonparametric Tests Based on Subsamples." *Communications in Statistics: Theory and Methods* 14 (7): 1669–1684.

Stephenson, W. Robert. 1981. "A General Class of One-Sample Nonparametric Test Statistics Based on Subsamples." *Journal of the American Statistical Association* 76 (376): 960–966.

Wantchekon, Leonard. 2003. "Clientelism and Voting Behavior: Evidence from a Field Experiment in Benin." *World Politics* 55 (3): 399–422.

A The t Test with Skewness and Unequal Sample Sizes

The following code gives rejection rates of 17% against the null with a nominal α of 5%, demonstrating how the t test can fail even in moderate-sized samples if there is a strong skew.

```r
### DGP with one small group and one large, and skew.
n1 <- 30
n2 <- 1000
alpha <- .2
beta <- 20
## Plot density
plot(seq(0, 1, 0.01), dbeta(seq(0,1,0.01), alpha, beta), type="l")
## Simulations
sims <- 10000
ps <- rep(NA, sims)
set.seed(1)
for (i in 1:sims){
  ys <- rbeta((n1 + n2), 0.1, 5) ## beta
  zs <- sample(c(rep(1, n1), rep(0, n2)), replace=FALSE)
  ps[i] <- as.numeric(t.test(ys ~ zs, var.equal=FALSE)[3]) ## Welch/Neyman
}
ps <- ps[!is.na(ps)]
length(ps)
mean(ps < 0.01)
mean(ps < 0.05)
```

B Effect-Increasing Statistics

Many statistics are effect increasing. In this appendix we first prove that differences in rescaled outcomes are effect increasing. We then show that a general class of rank-based statistics is EI in the absence of ties, and then extend this results to the case of ties for, the rank-sum and Stephenson rank-sum statistics, two primary statistics discussed in the paper.

Proof (Lemma 2): Let $Q_i = Q(Y_i)$ denote the observed scores, and let $Q_i(1) = Q(Y_i(1))$ and $Q_i(0) = Q(Y_i(0))$ indicate the scores of the potential outcomes. Denote as $S(W, \mathbb{S})$ any statistic with the form $a \sum_i W_i Q(1)_i = a \sum_i W_i Q_i$ with $a \geq 0$, i.e., the scaled sum of the scores of the treated observations.[20] We first show that for any $S(W, \mathbb{S})$ and any pair of potential-outcome schedules $\mathbb{S} \preceq \mathbb{S}'$, $S(W, \mathbb{S}) \leq S(W, \mathbb{S}')$:

$$Y_i(1) \leq Y_i'(1) \ \forall i \implies Q_i(1) \leq Q_i'(1) \ \forall i \qquad \text{b/c } Q(Y_i) \text{ is non-decreasing in } Y_i$$

$$\implies Q_i(1) - Q_i'(1) \leq 0 \ \forall i$$

$$\implies \sum W_i [Q_i(1) - Q_i'(1)] \leq 0 \qquad \text{b/c every element is } \leq 0$$

$$\implies \sum [W_i Q_i(1)] - \sum [W_i Q_i'(1)] \leq 0$$

$$\implies a \sum [W_i Q_i(1)] \leq a \sum [W_i Q_i'(1)]$$

$$\implies S(W, \mathbb{S}) \leq S(W, \mathbb{S}').$$

The above includes the special case of the sum of the treated responses, $\sum_i W_i Y(1)_i$, for which $Q(Y_i) = Y_i$. The intuition behind the above is that raising any individual potential outcome on the treatment side will either (if the unit was treated) increase $\sum_i W_i Y(1)_i$ or (if the unit was not treated) not affect the test statistic at all.

As is well known, the sum of the treated responses is permutationally equivalent to the treated-control difference of means, so the latter is also an effect-increasing statistic. So too is the difference of mean scores. This follows from the fact that the average of the treated scores, $\bar{Q}_T = N_T^{-1} \sum [W_i Q_i(1)]$, being the sum multiplied by a constant, is effect increasing, and so is the additive inverse of the average of the control scores, $-\bar{Q}_C = -N_C^{-1} \sum [W_i Q_i(0)]$. Because $\mathbb{S} \preceq \mathbb{S}' \implies \bar{Q}_T \leq \bar{Q}_T'$ and $\mathbb{S} \preceq \mathbb{S}' \implies -\bar{Q}_C \leq -\bar{Q}_C'$, we can conclude that $\mathbb{S} \preceq \mathbb{S}' \implies \bar{Q}_T - \bar{Q}_C \leq \bar{Q}_T' - \bar{Q}_C'$, that is, the difference of mean scores is an effect-

---

20. Note that as $Q(Y_i)$ depends only on unit i, not on the entire vector Y. This is a crucial difference from the class of sum statistics defined by Rosenbaum (2002, 35). As a result of this distinction, the class of statistics $S(W, \mathbb{S})$ defined here excludes rank statistics because the rank of Y_i depends on the values of other observations. We treat ranks statistics separately below.

increasing statistic as well. This again includes the difference of means as a special case where $Q(Y_i) = Y_i$. Finally, the above easily extends to any rescaling a and b of the two sums of the rescaled outcomes. □

Proof (Lemma 3): Let the R_i be the ranks of the (adjusted) observed outcomes and $a(\cdot)$ a monotonic mapping of these ranks to the real numbers.

Consider two schedules $\mathbb{S}' \preceq \mathbb{S}$ that are identical except that for some specific unit k with $Y'_k(1) \leq Y_k(1)$. Conceptually consider making $\mathbb{S}'$ by reducing potential outcome $Y_k(1)$ for some specific k, and leaving the other potential outcomes alone.

Now given any assignment vector W, we have $Y_i^{obs}, i = 1, \ldots, n$, and $R_i, i = 1, \ldots, n$ the associated ranks. Assume no ties in ranks. By lowering $Y_k(1)$ to $Y'_k(1)$ we potentially could change some ranks. In particular, the rank of unit k could go down and if it does the ranks of some other units would increase. Let $G = \{j : R'_j > R_j \text{ and } W_j = 1\}$ be the set of units in the treatment group with increased ranks. Let m denote the size of this set. Let $j_1, \ldots, j_m$ be the indices of the units in G arranged in increasing order by rank, so $R_{j_a} < R_{j_{a+1}}$. Importantly, for any unit in G, we have a change of at most 1 rank, so $R_{j_{a+1}} \geq R_{j_a} + 1 \geq R'_{j_a}$, giving $a(R'_{j_a}) \leq a(R_{j_{a+1}})$. Because no unit can increase its rank to above R_k by reducing unit k we have $R'_{j_m} \leq R_k$. Similarly, the reduced R'_k must be less than the rank of any unit in G, giving $R'_k \leq R_{j_1}$. This gives:

$$a(R'_k) + a(R'_{j_1}) + \ldots + a(R'_{j_{m-1}}) + a(R'_{j_m}) \leq a(R_{j_1}) + a(R_{j_2}) + \ldots + a(R_{j_m}) + a(R_k)$$

due to a pairwise comparison (the first elements of the two sums are ordered, the second, etc., up to the mth).

This means that $T' \leq T$ because the treatment average decreases, and the control average can only go up (those units impacted in the control group all have ranks that are larger).

A similar argument shows that $T' \leq T$ if we consider a pair of potential-outcome schedules where only a single control potential outcome is increased from $\mathbb{S}$ to $\mathbb{S}'$.

Finally, take any two potential-outcome schedules $\mathbb{S}' \preceq \mathbb{S}$. Generate a chain of potential-outcome schedules from $\mathbb{S}'$ to $\mathbb{S}$ by changing one potential outcome at a time. For example, the first step in the chain would be to modify $\mathbb{S}'$ to $\mathbb{S}''$ so $Y_1(1)'' = Y_1(1)$ and all other $Y_i''(z) = Y_i'(z)$. By transitivity along this sequence we finally have $T' \leq T$ for any W. Therefore, T is potential outcomes monotonic. $\qquad\square$

Examples. If $a(r) = r$ we have the classic rank sum test. Similarly, if

$$a(r) = \binom{r-1}{s-1} \text{ for } r \geq s \text{ and } a(r) = 0 \text{ otherwise}$$

for some fixed s (representing how many subsets of size s can be formed where our unit with rank r is biggest), we have the Stephenson Rank test.

Ties. Unfortunately, this general proof does not go through if there are ties. This is because by lowering a potential outcome, an entire group of mid-ranks can shift. Consider the case where the original ranks of treated units are $2, 2, 2, 8$ and we lower the rank-8 so much that it becomes rank 1. The other units then will have ranks $3, 3, 3$ giving final ranks of $1, 3, 3, 3$. Now, if $a(r) = 0$ if $r < 3$ and 1 otherwise, the sum of the four goes from 1 to 3. The control units are unaffected. This violates the monotonicity. Many specific rank based statistics are, however, EI even in the presence of ties. This can be shown by direct proof. We next do this for the rank-sum and the Stephenson rank-sum statistic.

The rank-sum test. Let

$$T(W, \mathbb{S}, H_{s\delta}) = \sum W_i \text{rank}(Y_i^{\text{obs}} - \delta W_i^{\text{obs}})$$

This statistic is equivalent to the Mann-Whitney statistic summing all pairs of treatment-control observations with the treatment beating the control

$$T_{MW}(W, \mathbb{S}) = \sum_{i,j} W_i(1 - W_j)\mathbf{1}\{Y_i^{\text{obs}} - \delta \geq Y_j^{\text{obs}}\} = \sum_{i,j} W_i(1 - W_j)\mathbf{1}\{Y_i(1) - \delta \geq Y_i(0)\}$$

with $\mathbf{1}\{a \leq b\}$ equalling $1/2$ if $a = b$.

Then, for any two potential-outcome schedules with $\mathbb{S} \preceq \mathbb{S}'$ we have

$$\mathbf{1}\{Y_i(1) - \delta \geq Y_i(0)\} \leq \mathbf{1}\{Y_i'(1) - \delta \geq Y_i'(0)\}$$

since we are moving the left side up and the right side down, which only increases the chance of having the left side be higher. Plugging this in to our sum of pairwise comparisons easily obtains our result of $T_{MW}(W, \mathbb{S}) \leq T_{MW}(W, \mathbb{S}')$.

The Stephenson rank test. Represent this statistic as a sum of indicators across all subsets where the indicator is 1 if a treatment unit is (tied for) the largest. We have, letting G indicate a size-s subset of unit indices and $\mathcal{G}$ the collection of all such G,

$$T_S(W, \mathbb{S}) = \sum_{G \in \mathcal{G}} H_G$$

where

$$H_G = \max_{i \in G} W_i \text{ s.t. } \tilde{Y}_i \geq \max_{j \in G} \tilde{Y}_j$$

with $\tilde{Y}_i$ being an adjusted outcome (i.e., imputed control outcome under the null). The above simply says that H_G is 1 if there is a treated unit that is (tied for) largest value in the set G. Alternatively, substitute Y_i^{obs} for $\tilde{Y}_i$.

Then, for any two potential-outcome schedules with $\mathbb{S} \preceq \mathbb{S}'$ and a given G we have H_G and H_G' with

$$H_G \leq H_G'$$

since for any unit under treatment, Y_i^{obs} can only be larger, and for control, smaller. Therefore, for each subset where a treatment unit was largest for $\mathbb{S}$, we will still see one being largest for $\mathbb{S}'$. These inequalities sum, giving $T_S \leq T_S'$ for any W, which implies monotonicity.