Randomization Tests under an Approximate Symmetry Assumption*

Ivan A. Canay†
Department of Economics
Northwestern University
iacanay@northwestern.edu

Joseph P. Romano‡
Departments of Economics and Statistics
Stanford University
romano@stanford.edu

Azeem M. Shaikh§
Department of Economics
University of Chicago
amshaikh@uchicago.edu

November 16, 2016

Abstract

This paper develops a theory of randomization tests under an approximate symmetry assumption. Randomization tests provide a general means of constructing tests that control size in finite samples whenever the distribution of the observed data exhibits symmetry under the null hypothesis. Here, by exhibits symmetry we mean that the distribution remains invariant under a group of transformations. In this paper, we provide conditions under which the same construction can be used to construct tests that asymptotically control the probability of a false rejection whenever the distribution of the observed data exhibits approximate symmetry in the sense that the limiting distribution of a function of the data exhibits symmetry under the null hypothesis. An important application of this idea is in settings where the data may be grouped into a fixed number of "clusters" with a large number of observations within each cluster. In such settings, we show that the distribution of the observed data satisfies our approximate symmetry requirement under weak assumptions. In particular, our results allow for the clusters to be heterogeneous and also have dependence not only within each cluster, but also across clusters. This approach enjoys several advantages over other approaches in these settings.

KEYWORDS: Randomization tests, dependence, heterogeneity, differences-in-differences, clustered data, sign changes, symmetric distribution, weak convergence

JEL classification codes: C12, C14.

*We thank Chris Hansen, Aprajit Mahajan, Ulrich Mueller and Chris Taber for helpful comments. This research was supported in part through the computational resources and staff contributions provided for the Social Sciences Computing cluster (SSCC) at Northwestern University. Sergey Gitlin provided excellent research assistance.

†Research supported by NSF Grant SES-1530534.

‡Research supported by NSF Grant DMS-1307973.

§Research supported by NSF Grants DMS-1308260, SES-1227091, and SES-1530661.

1 Introduction

Suppose the researcher observes data $X^{(n)} \sim P_n \in \mathbf{P}_n$, where $\mathbf{P}_n$ is a set of distributions on a sample space $\mathcal{X}_n$, and is interested in testing

$$H_0 : P_n \in \mathbf{P}_{n,0} \text{ versus } H_1 : P_n \in \mathbf{P}_n \setminus \mathbf{P}_{n,0} \ ,$$

where $\mathbf{P}_{n,0} \subset \mathbf{P}_n$, at level $\alpha \in (0,1)$. The index n here will typically denote sample size. The classical theory of randomization tests provides a general way of constructing tests that control size in finite samples provided that the distribution of the observed data exhibits symmetry under the null hypothesis. Here, by exhibits symmetry we mean that the distribution remains invariant under a group of transformations. In this paper, we develop conditions under which the same construction can be used to construct tests that asymptotically control the probability of a false rejection provided that the distribution of the observed data exhibits approximate symmetry. More precisely, the main requirement we impose is that, for a known function S_n from $\mathcal{X}_n$ to a sample space $\mathcal{S}$,

$$S_n(X^{(n)}) \xrightarrow{d} S \tag{1}$$

as $n \to \infty$ under $P_n \in \mathbf{P}_{n,0}$, where S exhibits symmetry in the sense described above. In this way, our results extend the classical theory of randomization tests. Note that in some cases S_n need not be completely known; see Remark 4.4 below.

While they apply more generally, an important application of our results is in settings where the data may be grouped into q "clusters" with a large number of observations within each cluster. A noteworthy feature of our asymptotic framework is that q is fixed and does not depend on n. In such environments, it is often the case that the distribution of the observed data satisfies our approximate symmetry requirement under weak assumptions. In particular, it typically suffices to consider

$$S_n(X^{(n)}) = (S_{n,1}(X^{(n)}), \ldots, S_{n,q}(X^{(n)}))' \ , \tag{2}$$

where $S_{n,j}(X^{(n)})$ is an appropriately recentered and rescaled estimator of the parameter of interest based on observations from the jth cluster. In this case, the convergence (1) often holds for S that exhibits symmetry in the sense that its distribution remains invariant under the group of sign changes. Importantly, this convergence permits the clusters to be heterogeneous and also have dependence not only within each cluster, but also across clusters. We consider three specific examples of such settings in detail – time series regression, differences-in-differences, and clustered regression.

Our paper is most closely related to the procedure suggested by Ibragimov and Müller (2010). As in our paper, they also consider settings where the data may be grouped into a fixed number of "clusters," q, with a large number of observations within each cluster. In order to apply their results, they further assume that the parameter of interest is scalar and that $S_n(X^{(n)})$ defined

in (2) satisfies the convergence (1) with S satisfying additional restrictions beyond our symmetry assumption. Using a result on robustness of the t-test established in Bakirov and Székely (2006), they propose an approach that leads to a test that asymptotically controls size for certain values of q and α, but may be quite conservative in the sense that its asymptotic rejection probability under the null hypothesis may be much less than α. This same result on the t-test underlies the approach put forward by Bester et al. (2011), which therefore inherits the same qualifications. The methodology proposed in this paper enjoys several advantages over these approaches, including not requiring the parameter of interest to be scalar, being valid for any values of q and α (thereby permitting in particular the computation of p-values), and being asymptotically similar in the sense of having asymptotic rejection probability under the null hypothesis equal to α. As shown in a simulation study, this feature translates into improved power at many alternatives. See Section 2.1.1 and Section S.2 in the Supplemental Material for further details.

The remainder of the paper is organized as follows. Section 2 briefly reviews the classical theory of randomization tests. Here, we pay special attention to an example involving the group of sign changes, which, as mentioned previously, underlies many of our later applications and aids comparisons with the approach suggested by Ibragimov and Müller (2010). Our main results are developed in Section 3. Section 4 contains the application of our results to settings where the data may be grouped into a fixed number of "clusters" with a large number of observations within each cluster, emphasizing in particular differences-in-differences and clustered regression. In Section S.1 of the Supplemental Material to this paper (Canay, Romano and Shaikh, 2015) we also consider an application to time series regression. Simulation results based on the time series regression and differences-in-differences examples are presented in Section S.2. Finally, in Section S.3, we use the clustered regression example to revisit the analysis of Angrist and Lavy (2009), who examine the impact of a cash award on exam performance for low-achievement students in Israel.

2 Review of Randomization Tests

In this section, we briefly review the classical theory of randomization tests. Further discussion can be found, for example, in Chapter 15 of Lehmann and Romano (2005). Since the results in this section are non-asymptotic in nature, we omit the index n.

Suppose the researcher observes data $X \sim P \in \mathbf{P}$, where $\mathbf{P}$ is a set of distributions on a sample space $\mathcal{X}$, and is interested in testing

$$H_0 : P \in \mathbf{P}_0 \text{ versus } H_1 : P \in \mathbf{P} \setminus \mathbf{P}_0 \ , \tag{3}$$

where $\mathbf{P}_0 \subset \mathbf{P}$, at level $\alpha \in (0,1)$. Randomization tests require that the distribution of the data, P, exhibits symmetry whenever $P \in \mathbf{P}_0$. In order to state this requirement more formally, let $\mathbf{G}$

be a finite group of transformations from $\mathcal{X}$ to $\mathcal{X}$ and denote by gx the action of $g \in \mathbf{G}$ on $x \in \mathcal{X}$. Using this notation, the classical condition required for a randomization test is

$$X \overset{d}{=} gX \text{ under } P \text{ for any } P \in \mathbf{P}_0 \text{ and } g \in \mathbf{G} \ . \tag{4}$$

We now describe the construction of the randomization test. Let $T(X)$ be a real-valued test statistic such that large values provide evidence against the null hypothesis. Let $M = |\mathbf{G}|$ and denote by

$$T^{(1)}(X) \leq T^{(2)}(X) \leq \cdots \leq T^{(M)}(X)$$

the ordered values of $\{T(gX) : g \in \mathbf{G}\}$. Let $k = \lceil M(1-\alpha) \rceil$ and define

$$\begin{aligned} M^+(X) &= |\{1 \leq j \leq M : T^{(j)}(X) > T^{(k)}(X)\}| \\ M^0(X) &= |\{1 \leq j \leq M : T^{(j)}(X) = T^{(k)}(X)\}| \ . \end{aligned} \tag{5}$$

Using this notation, the randomization test is given by

$$\phi(X) = \begin{cases} 1 & \text{if } T(X) > T^{(k)}(X) \\ a(X) & \text{if } T(X) = T^{(k)}(X) \\ 0 & \text{if } T(X) < T^{(k)}(X) \end{cases} \ , \tag{6}$$

where

$$a(X) = \frac{M\alpha - M^+(X)}{M^0(X)} \ .$$

The following theorem shows that this construction leads to a test that controls size in finite samples whenever (4) holds. In fact, the test in (6) is similar, i.e., has rejection probability exactly equal to α for any $P \in \mathbf{P}_0$ and $\alpha \in (0,1)$.

Theorem 2.1. *Suppose $X \sim P \in \mathbf{P}$ and consider the problem of testing (3). Let $\mathbf{G}$ be a group such that (4) holds. Then, for any $\alpha \in (0,1)$, $\phi(X)$ defined in (6) satisfies*

$$E_P[\phi(X)] = \alpha \text{ whenever } P \in \mathbf{P}_0 \ . \tag{7}$$

Remark 2.1. Let $\mathbf{G}^x$ denote the $\mathbf{G}$-orbit of $x \in \mathcal{X}$, i.e., $\mathbf{G}^x = \{gx : g \in \mathbf{G}\}$. The result in Theorem 2.1 exploits that, when $\mathbf{G}$ is such that (4) holds, the conditional distribution of X given $X \in \mathbf{G}^x$ is uniform on $\mathbf{G}^x$ whenever $P \in \mathbf{P}_0$. Since the conditional distribution of X is known for all $P \in \mathbf{P}_0$ (even though P itself is unknown), we can construct a test that is level α conditionally, which leads to a test that is level α unconditionally as well. ■

Remark 2.2. In some cases, M is too large to permit computation of $\phi(X)$ defined in (6). When this is the case, the researcher may use a stochastic approximation to $\phi(X)$ without affecting the finite-sample validity of the test. More formally, let

$$\hat{\mathbf{G}} = \{g_1, \ldots, g_B\} \ , \tag{8}$$

where g_1 = the identity transformation and $g_2, \ldots, g_B$ are i.i.d. Uniform($\mathbf{G}$). Theorem 2.1 remains true if, in the construction of $\phi(X)$, $\mathbf{G}$ is replaced by $\hat{\mathbf{G}}$. ■

Remark 2.3. One can construct a p-value for the test $\phi(X)$ defined in (6) as

$$\hat{p} = \hat{p}(X) = \frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} I\{T(gX) \geq T(X)\} \ . \tag{9}$$

When (4) holds, it follows that $P\{\hat{p} \leq u\} \leq u$ for all $0 \leq u \leq 1$ and $P \in \mathbf{P}_0$. This result remains true when M is large and the researcher uses a stochastic approximation, in which case $\hat{\mathbf{G}}$ as defined in (8) replaces $\mathbf{G}$ in (9). ■

Remark 2.4. The test in (6) is possibly randomized. In case one prefers not to randomize, note that the non-randomized test that rejects if $T(X) > T^{(k)}(X)$ is level α. In our simulations, this test has rejection probability under the null hypothesis only slightly less than α when M is not too small; see Section 2.1.1 below and Sections S.2.1 and S.2.2 in the Supplemental Material for additional discussion. ■

2.1 Symmetric Location Example

In this subsection, we provide an illustration of Theorem 2.1. The example not only makes concrete some of the abstract ideas presented above, but also underlies many of the applications described in Section 4 below.

Suppose $X = (X_1, \ldots, X_q) \sim P \in \mathbf{P}$, where

$$\mathbf{P} = \{\otimes_{j=1}^{q} P_{j,\mu} : P_{j,\mu} \text{ symmetric distribution on } \mathbf{R}^d \text{ about } \mu\} \ .$$

In other words, $X_1, \ldots, X_q$ are independent and each X_j is distributed symmetrically on $\mathbf{R}^d$ about μ, i.e., $X_j - \mu \stackrel{d}{=} \mu - X_j$. The researcher desires to test (3) with

$$\mathbf{P}_0 = \{\otimes_{j=1}^{q} P_{j,\mu} : P_{j,\mu} \text{ a symmetric distribution on } \mathbf{R}^d \text{ about } \mu \text{ with } \mu = 0\} \ .$$

In this case, (4) clearly holds with the group of sign changes $\mathbf{G} = \{-1, 1\}^q$, where the action of $g = (g_1, \ldots, g_q) \in \mathbf{G}$ on $x = (x_1, \ldots, x_q) \in \otimes_{j=1}^{q} \mathbf{R}^d$ is defined by $gx = (g_1 x_1, \ldots, g_q x_q)$. As a result, Theorem 2.1 may be applied with any choice of $T(X)$ to construct a test that satisfies (7).

2.1.1 Comparison with the t-test

Consider the special case of the symmetric location example in which $d = 1$ and $P_{j,\mu} = N(\mu, \sigma_j^2)$, i.e.,

$$\mathbf{P} = \{\otimes_{j=1}^{q} P_{j,\mu} : P_{j,\mu} = N(\mu, \sigma_j^2) \text{ with } \mu \in \mathbf{R} \text{ and } \sigma_j^2 \geq 0\} \tag{10}$$

$$\mathbf{P}_0 = \{\otimes_{j=1}^{q} P_{j,\mu} : P_{j,\mu} = N(\mu, \sigma_j^2) \text{ with } \mu = 0 \text{ and } \sigma_j^2 \geq 0\} \ . \tag{11}$$

For this setting, Bakirov and Székely (2006) show that the usual two-sided t-test remains valid despite heterogeneity in the σ_j^2 for certain values of α and q. More formally, they show that for $\alpha \leq 8.3\%$ and $q \geq 2$ or $\alpha \leq 10\%$ and $2 \leq q \leq 14$,

$$P\{T_{|t\text{-stat}|}(X) > c_{q-1,1-\frac{\alpha}{2}}\} \leq \alpha \text{ for any } P \in \mathbf{P}_0 \ ,$$

where $T_{|t\text{-stat}|}(X)$ is the absolute value of the usual t-statistic computed using the data X and $c_{q-1,1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of the t-distribution with $q-1$ degrees of freedom. Bakirov and Székely (2006) go on to show that this result remains true even if each $P_{j,\mu}$ is allowed to be a mixture of normal distributions as well. This result was further explored by Ibragimov and Müller (2010) and Ibragimov and Müller (2016). Ibragimov and Müller (2016) derived a related result for the two-sample problem, while Ibragimov and Müller (2010) showed that the t-test is "optimal" in the sense that it is the uniformly most powerful scale invariant level α test against the restricted class of alternatives with $\sigma_j^2 = \sigma^2$ for all $1 \leq j \leq q$. In the Appendix, we establish a similar "optimality" result for the randomization test with $T(X) = T_{|t\text{-stat}|}(X)$ and $\mathbf{G} = \{-1,1\}^q$: we show that it is the uniformly most powerful unbiased level α test against the same class of alternatives.

We compare the randomization test with $T(X) = T_{|t\text{-stat}|}(X)$ and $\mathbf{G} = \{-1,1\}^q$ with the t-test. We follow Ibragimov and Müller (2010) and consider the setup in (10)-(11) with $q \in \{8,16\}$ and $\sigma_j^2 = 1$ for $1 \leq j \leq \frac{q}{2}$ and $\sigma_j^2 = a^2$ for $\frac{q}{2} < j \leq q$. Figure 1 shows rejection probabilities under the null hypothesis computed using $100,000$ Monte Carlo repetitions for $\alpha = 5\%$, a ranging over a grid of 50 equally spaced points in $(0.1, 5)$, $q = 8$ (left panel) and $q = 16$ (right panel). As we would expect from Theorem 2.1, the rejection probability of the randomization test equals α for all values of the heterogeneity parameter a (up to simulation error). The rejection probability of the t-test, on the other hand, can be substantially below α when the data are heterogeneous, i.e., $a \neq 1$. Comparing the right and left panels, we see that the performance of the t-test improves as q gets larger, but it is worth emphasizing that the results of Bakirov and Székely (2006) do not ensure the validity of the test for $q > 14$ and $\alpha \geq 8.4\%$.

Figure 2 shows rejection probabilities computed using $100,000$ Monte Carlo repetitions for $\alpha = 5\%$, $\mu \in (0, 1.5)$, $q = 8$, $a = 0.1$ (left panel) and $a = 1$ (right panel). The similarity of the randomization test translates into better power for alternatives close to the null hypothesis. When $a = 0.1$, the rejection probability of the randomization test exceeds that of the t-test for μ less than approximately 0.7; for larger values of μ, the situation is reversed, though the difference in power between the two tests is smaller. When $a = 1$, the t-test slightly outperforms the randomization test, reflecting the previously mentioned optimality property derived in Ibragimov and Müller (2010). It is important to note that this does not contradict the optimality result for the randomization test established in the Appendix, as the t-test is not unbiased. In particular, there are alternatives $P \in \mathbf{P}_1$ under which the t-test has rejection probability $< \alpha$. Moreover, the loss in power of the randomization test relative to the t-test even in this case is arguably negligible. These comparisons

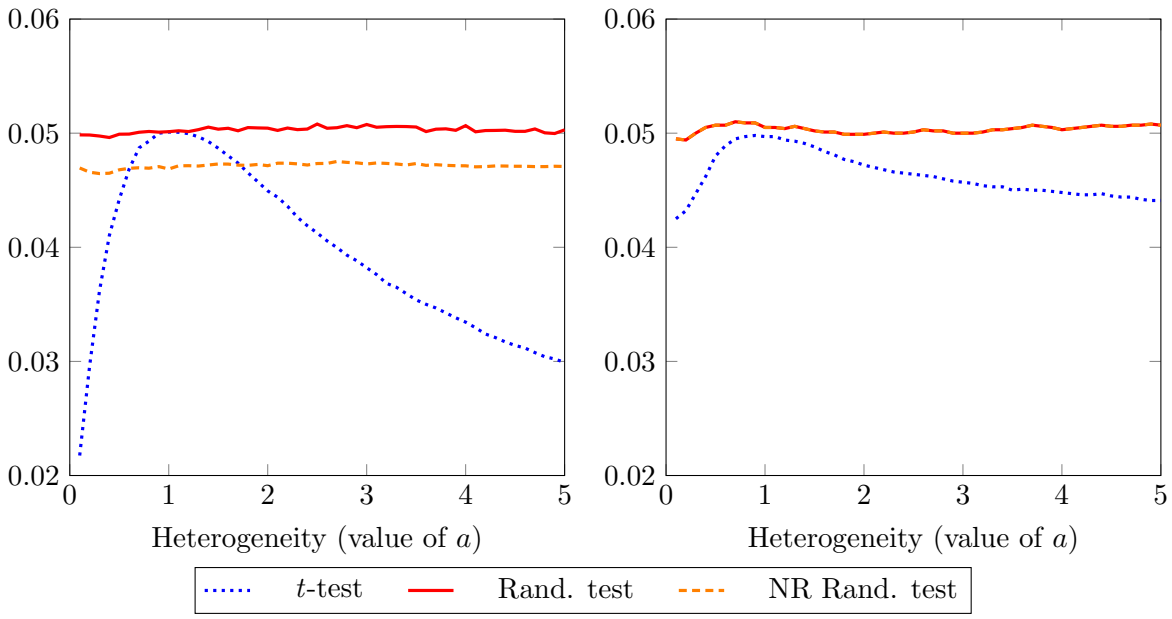

Figure 1: Rejection probabilities under the null hypothesis for different values of a in the symmetric location example. Randomization test (randomized and non-randomized) versus t-test. $q = 8$ (left panel) and $q = 16$ (right panel).

continue to hold even if the randomization test is replaced with its non-randomized version described in Remark 2.4.

In the context of the symmetric location example, the randomization test provides additional advantages over the t-test approach. First, the randomization test works for all levels of $\alpha \in (0, 1)$, which allows for the construction of p-values; see Remark 2.3. Second, the randomization test works for vector-valued random variables, i.e., $d > 1$, while the result in Bakirov and Székely (2006) is restricted to scalar random variables. Third, the construction in Theorem 2.1 works for any choice of test statistic $T(X)$. Finally, the condition in (4) is not limited to mixtures of normal distributions and holds for any symmetric distribution. On the other hand, when q is small the rejection probability of the t-test sometimes exceeds that of the non-randomized version of the randomization test described in Remark 2.4; see Figure 1.

3 Main Result

In this section, we present our theory of randomization tests under an approximate symmetry assumption. Since our results in this section are asymptotic in nature, we re-introduce the index n, which, as mentioned earlier, will typically be used to denote the sample size.

Suppose the researcher observes data $X^{(n)} \sim P_n \in \mathbf{P}_n$, where $\mathbf{P}_n$ is a set of distributions on a

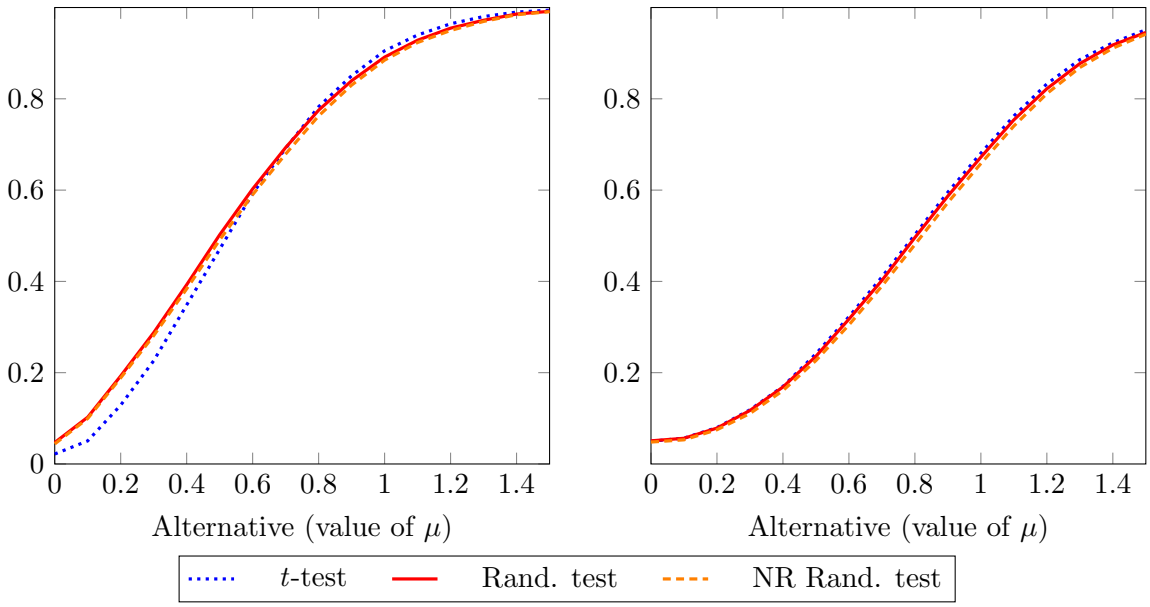

Figure 2: Rejection probabilities for $q = 8$ and different values of μ in the symmetric location example. Randomization test (randomized and non-randomized) versus t-test. $a = 0.1$ (left panel) and $a = 1$ (right panel).

sample space $\mathcal{X}_n$, and is interested in testing

$$H_0 : P_n \in \mathbf{P}_{n,0} \text{ versus } H_1 : P_n \in \mathbf{P}_n \setminus \mathbf{P}_{n,0} \ , \tag{12}$$

where $\mathbf{P}_{n,0} \subset \mathbf{P}_n$, at level $\alpha \in (0,1)$. In contrast to Section 2, we no longer require that the distribution of $X^{(n)}$ exhibits symmetry whenever $P_n \in \mathbf{P}_{n,0}$. Instead, we require that $X^{(n)}$ exhibits approximate symmetry whenever $P_n \in \mathbf{P}_{n,0}$. In order to state this requirement more formally, we require some additional notation. Recall that S_n denotes a function from $\mathcal{X}_n$ to a sample space $\mathcal{S}$. For simplicity, we assume further that $\mathcal{S}$ is a subset of Euclidean space, though this could be generalized to a metric space. As before, let T be a real-valued test statistic such that large values provide evidence against the null hypothesis, but we will assume that T is a function from $\mathcal{S}$ to $\mathbf{R}$ as opposed to from $\mathcal{X}_n$ to $\mathbf{R}$. Finally, let $\mathbf{G}$ be a (finite) group of transformations from $\mathcal{S}$ to $\mathcal{S}$ and denote by gs the action of $g \in \mathbf{G}$ on $s \in \mathcal{S}$. Using this notation, the following assumption is assumed to hold for certain sequences $\{P_n \in \mathbf{P}_{n,0} : n \geq 1\}$:

Assumption 3.1.

(i) $S_n = S_n(X^{(n)}) \xrightarrow{d} S$ *under* P_n.

(ii) $gS \overset{d}{=} S$ *for all* $g \in \mathbf{G}$.

(iii) For any two distinct elements $g \in \mathbf{G}$ *and* $g' \in \mathbf{G}$,

$$\text{either } T(gs) = T(g's) \ \forall s \in \mathcal{S} \text{ or } P\{T(gS) \neq T(g'S)\} = 1 \ .$$

Assumption 3.1.(i)-(ii) formalizes what we mean by $X^{(n)}$ exhibiting approximate symmetry. Assumption 3.1.(iii) is a condition that controls the ties among the values of $T(gS)$ as g varies over $\mathbf{G}$. It requires that $T(gS)$ and $T(g'S)$ are distinct with probability one or deterministically equal to each other. For examples of S that often arise in applications and typical choices of T, we verify Assumption 3.1.(iii) (see, in particular, Lemmas S.5.1-S.5.3 in the Supplemental Material).

The construction of the randomization test in this setting parallels the one in Section 2 with S_n replacing X. Let $M = |\mathbf{G}|$ and denote by

$$T^{(1)}(S_n) \leq T^{(2)}(S_n) \leq \cdots \leq T^{(M)}(S_n)$$

the ordered values of $\{T(gS_n) : g \in \mathbf{G}\}$. Let $k = \lceil M(1-\alpha) \rceil$ and define $M^+(S_n)$ and $M^0(S_n)$ as in (5) with S_n replacing X. Using this notation, the proposed test is given by

$$\phi(S_n) = \begin{cases} 1 & T(S_n) > T^{(k)}(S_n) \\ a(S_n) & T(S_n) = T^{(k)}(S_n) \\ 0 & T(S_n) < T^{(k)}(S_n) \end{cases}, \tag{13}$$

where

$$a(S_n) = \frac{M\alpha - M^+(S_n)}{M^0(S_n)} .$$

The following theorem shows that this construction leads to a test that is asymptotically level α whenever $\{P_n \in \mathbf{P}_{n,0} : n \geq 1\}$ is such that Assumption 3.1 holds. In fact, the proposed test is asymptotically similar, i.e., has limiting rejection probability equal to α for all such sequences.

Theorem 3.1. *Suppose $X^{(n)} \sim P_n \in \mathbf{P}_n$ and consider the problem of testing (12). Let $S_n : \mathcal{X}_n \to \mathcal{S}$, $T : \mathcal{S} \to \mathbf{R}$ and $\mathbf{G} : \mathcal{S} \to \mathcal{S}$ be such that $T : \mathcal{S} \to \mathbf{R}$ is continuous and $g : \mathcal{S} \to \mathcal{S}$ is continuous for all $g \in \mathbf{G}$. Then, for any $\alpha \in (0,1)$, $\phi(S_n)$ defined in (13) satisfies*

$$E_{P_n}[\phi(S_n)] \to \alpha \tag{14}$$

as $n \to \infty$ whenever $\{P_n \in \mathbf{P}_{n,0} : n \geq 1\}$ is such that Assumption 3.1 holds.

Remark 3.1. Note that the limiting random variable S that appears in Assumption 3.1 may depend on the sequence $\{P_n \in \mathbf{P}_{n,0} : n \geq 1\}$. ■

Remark 3.2. The assumptions in Theorem 3.1 are stronger than required. The conclusion (14) holds for example, under the following weaker conditions: if T is such that T is only continuous on a set $\mathcal{S}' \subseteq \mathcal{S}$ such that $P\{S \in \mathcal{S}'\} = 1$; if $\mathbf{G}$ is such that g is continuous on a set $\mathcal{S}' \subseteq \mathcal{S}$ such that $P\{S \in \mathcal{S}'\} = 1$ for all $g \in \mathbf{G}$; and whenever $\{P_n \in \mathbf{P}_{n,0} : n \geq 1\}$ is such that for every subsequence $\{P_{n_k} \in \mathbf{P}_{n_k,0} : k \geq 1\}$ there exists a further subsequence $\{P_{n_{k_\ell}} \in \mathbf{P}_{n_{k_\ell},0} : \ell \geq 1\}$ for which Assumption 3.1 is satisfied with $P_{n_{k_\ell}}$ in place of P_n. More generally, as noted by a referee, the assumptions we impose are sufficient to ensure that ϕ is continuous on a set $\mathcal{S}' \subseteq \mathcal{S}$ such that

$P\{S \in \mathcal{S}'\} = 1$. In establishing this, an important observation is that $\phi(s) = \phi(s')$ for any s and s' such that the orderings of $\{T(gs) : g \in \mathbf{G}\}$ and $\{T(gs') : g \in \mathbf{G}\}$ correspond to the same transformations $g_{(1)}, \ldots, g_{(M)}$. This continuity may, of course, be established under alternative sets of assumptions. For example, in the context of a regression discontinuity setting, Canay and Kamat (2015) fruitfully exploit the fact that T is a rank statistic to provide an alternative set of conditions. ■

Remark 3.3. If for every sequence $\{P_n \in \mathbf{P}_{n,0} : n \geq 1\}$ there exists a subsequence $\{P_{n_k} \in \mathbf{P}_{n_k,0} : k \geq 1\}$ for which Assumption 3.1 is satisfied with P_{n_k} in place of P_n, then the conclusion of Theorem 3.1 can be strengthened as follows: for any $\alpha \in (0,1)$, $\phi(S_n)$ defined in (13) satisfies

$$\sup_{P_n \in \mathbf{P}_{n,0}} |E_{P_n}[\phi(S_n)] - \alpha| \to 0$$

as $n \to \infty$. ■

Remark 3.4. As described in Remark 2.1, the validity of the randomization test in finite samples is tightly related to fact that the conditional distribution of X given $X \in \mathbf{G}^x$ is uniform on $\mathbf{G}^x$. While this property holds for the limiting random variable S in our framework, it may not hold even approximately for S_n for large n. ■

Remark 3.5. Earlier work on the asymptotic behavior of randomization tests includes Hoeffding (1952), Romano (1989, 1990), Chung and Romano (2013, 2016a,b). The arguments in these papers involve showing that the "randomization distribution" (see, e.g., Chapter 15 of Lehmann and Romano, 2005) settles down to a fixed distribution as $|\mathbf{G}| \to \infty$. In our framework, $|\mathbf{G}|$ is fixed and the "randomization distribution" will generally not settle down at all. For this reason, the analysis in these papers is not useful in our setting. ■

Remark 3.6. Comments analogous to those made in Remarks 2.2-2.4 after Theorem 2.1 apply to Theorem 3.1. In particular, Theorem 3.1 still holds when $\mathbf{G}$ is replaced by $\hat{\mathbf{G}}$ defined in (8), asymptotically valid p-values can be computed using (9), and the non-randomized test that rejects if $T(S_n) > T^{(k)}(S_n)$ is also asymptotically level α, although possibly conservative. ■

4 Applications

In this section we present two applications of Theorem 3.1 to settings where the data may be grouped into a fixed number of "clusters," q, with a large number of observations within each cluster: differences-in-differences and clustered regression. Before proceeding to these specific examples, we highlight a common structure found in all of the applications.

Suppose the researcher observes data $X^{(n)} \sim P_n \in \mathbf{P}_n$ and considers testing the hypotheses in (12) with

$$\mathbf{P}_{n,0} = \{P_n \in \mathbf{P}_n : \theta_n(P_n) = \theta_0\} ,$$

where $\theta_n(P_n) \in \Theta \subseteq \mathbf{R}^d$ is some parameter of interest. Further suppose that the data $X^{(n)}$ can be grouped into q clusters, $X_1^{(n)}, \ldots, X_q^{(n)}$, where the clusters are allowed to have observations in common. Let $\hat{\theta}_{n,j} = \hat{\theta}_{n,j}(X_j^{(n)})$ be an estimator of $\theta_n(P_n)$ based on observations from the jth cluster such that under weak assumptions on the sequence $\{P_n \in \mathbf{P}_{n,0} : n \geq 1\}$,

$$S_n(X^{(n)}) = \sqrt{n}(\hat{\theta}_{n,1} - \theta_0, \ldots, \hat{\theta}_{n,q} - \theta_0) \xrightarrow{d} N(0, \Sigma) \tag{15}$$

as $n \to \infty$, where $\Sigma = \text{diag}\{\Sigma_1, \ldots, \Sigma_q\}$ and each Σ_j is of dimension $d \times d$. In this setting, the conditions of Theorem 3.1 hold for $\mathbf{G} = \{-1, 1\}^q$ and $T(S_n) = T_{\text{Wald}}(S_n)$, where

$$T_{\text{Wald}}(S_n) = q\bar{S}'_{n,q}\bar{\Sigma}^{-1}_{n,q}\bar{S}_{n,q} \tag{16}$$

with

$$\bar{\Sigma}_{n,q} = \frac{1}{q}\sum_{j=1}^{q} S_{n,j}S'_{n,j} \ , \ \bar{S}_{n,q} = \frac{1}{q}\sum_{j=1}^{q} S_{n,j} \ , \text{ and } S_{n,j} = \sqrt{n}(\hat{\theta}_{n,j} - \theta_0) \ .$$

See Lemma S.5.3 in the Supplemental Material for details. In the special case where $d = 1$, the conditions of Theorem 3.1 also hold for $T(S_n) = T_{|t\text{-stat}|}(S_n)$, where

$$T_{|t\text{-stat}|}(S_n) = \frac{|\bar{S}_{n,q}|}{\sqrt{\frac{1}{q-1}\sum_{j=1}^{q}(S_{n,j} - \bar{S}_{n,q})^2}} \ .$$

See Lemmas S.5.1-S.5.2 in the Supplemental Material for details. Equivalently,

$$T_{|t\text{-stat}|}(S_n) = \frac{|\bar{\hat{\theta}}_{n,q} - \theta_0|}{s_{\hat{\theta}}/\sqrt{q}} \ , \tag{17}$$

with

$$\bar{\hat{\theta}}_{n,q} = \frac{1}{q}\sum_{j=1}^{q}\hat{\theta}_{n,j} \text{ and } s^2_{\hat{\theta}} = \frac{1}{q-1}\sum_{j=1}^{q}(\hat{\theta}_{n,j} - \bar{\hat{\theta}}_{n,q})^2 \ .$$

In each of the applications below, we will therefore simply specify $X_j^{(n)}$ and $\hat{\theta}_{n,j}$ and argue that the convergence (15) holds under weak assumptions on the sequence $\{P_n \in \mathbf{P}_{n,0} : n \geq 1\}$.

Remark 4.1. In the special case where $d = 1$, the idea of grouping the data in this way and constructing estimators satisfying (15) has been previously proposed by Ibragimov and Müller (2010). Using the result on the t-test described in Section 2.1.1, they go on to propose a test that rejects the null hypothesis when $T_{|t\text{-stat}|}(S_n)$ in (17) exceeds the $1 - \frac{\alpha}{2}$ quantile of a t-distribution with $q - 1$ degrees of freedom. Further comparisons with this approach are provided in Section S.2 of the Supplemental Material. ■

Remark 4.2. The convergence (15) permits dependence within each cluster. It also permits some dependence across clusters, but, importantly, not so much that Σ in (15) does not have the required diagonal structure. See, for example, Jenish and Prucha (2009) for some relevant central limit theorems. The convergence (15) further allows for heterogeneity in the distribution of the data across clusters in the sense that Σ_j need not be independent of j in $\Sigma = \text{diag}\{\Sigma_1, \ldots, \Sigma_q\}$. ■

Remark 4.3. The asymptotic normality in (15) arises frequently in applications, but is not necessary for the validity of the test described above. All that is required is that the q estimators (after an appropriate re-centering and scaling) have a limiting distribution that is the product of q distributions that are symmetric about zero. This may even hold in cases where the estimators have infinite variances or are inconsistent. See Remark 4.5 below. ■

Remark 4.4. The test statistics in (16) and (17) are both invariant under scalar multiplication. As a result, the $\sqrt{n}$ in the definition of S_n in (15) may be omitted or replaced with another sequence without changing the results. ■

4.1 Differences-in-Differences

Suppose

$$Y_{j,t} = \theta D_{j,t} + \eta_j + \gamma_t + \epsilon_{j,t} \quad \text{with} \quad E[\epsilon_{j,t}] = 0 \ . \tag{18}$$

Here, the observed data is given by $X^{(n)} = \{(Y_{j,t}, D_{j,t}) : j \in J_0 \cup J_1, t \in T_0 \cup T_1\} \sim P_n$ taking values on a sample space $\mathcal{X}_n = \prod_{j \in J_0 \cup J_1, t \in T_0 \cup T_1} \mathbf{R} \times \{0,1\}$, where $Y_{j,t}$ is the outcome of unit j at time t, $D_{j,t}$ is the (non-random) treatment status of unit j at time t, T_0 is the set of pre-treatment time periods, T_1 is the set of post-treatment time periods, J_0 is the set of controls units, and J_1 is the set of treatment units. The scalar random variables η_j, γ_t and $\epsilon_{j,t}$ are unobserved and $\theta \in \Theta \subseteq \mathbf{R}$ is the parameter of interest.

As before, in order to state the null and alternative hypotheses formally, it is useful to introduce some further notation. Let $W^{(n)} = \{(\epsilon_{j,t}, \eta_j, \gamma_t, D_{j,t}) : j \in J_0 \cup J_1, t \in T_0 \cup T_1\} \sim Q_n \in \mathbf{Q}_n$ taking values on a sample space $\mathcal{W}_n = \prod_{j \in J_0 \cup J_1, t \in T_0 \cup T_1} \mathbf{R} \times \mathbf{R} \times \mathbf{R} \times \{0,1\}$ and $A_{n,\theta} : \mathcal{W}_n \to \mathcal{X}_n$ be the mapping implied by (18). Our assumptions on $\mathbf{Q}_n$ are discussed below. Using this notation, define

$$\mathbf{P}_n = \bigcup_{\theta \in \Theta} \mathbf{P}_n(\theta) \text{ with } \mathbf{P}_n(\theta) = \{Q_n A_{n,\theta}^{-1} : Q_n \in \mathbf{Q}_n\} \ .$$

The null and alternative hypotheses of interest are thus given by (12) with $\mathbf{P}_{n,0} = \mathbf{P}_n(\theta_0)$.

In order to apply our methodology, we must again specify $X_j^{(n)}$ and $\hat{\theta}_{n,j}$ and argue that the convergence (15) holds under weak assumptions on the sequence $\{P_n \in \mathbf{P}_{n,0} : n \geq 1\}$. Different specifications may be appropriate for different asymptotic frameworks. We first consider an asymptotic framework similar to the one in Conley and Taber (2011), where $|J_1| = q$ is fixed, $|J_0| \to \infty$, and $\min\{|T_0|, |T_1|\} \to \infty$ with $\frac{|T_1|}{|T_0|} \to c \in (0, \infty)$. A modification for an alternative asymptotic framework in which $|J_0|$ is also fixed is discussed in Remark 4.10 below. For such an asymptotic framework, for each $j \in J_1$, define

$$X_j^{(n)} = \{(Y_{k,t}, D_{k,t}) : k \in \{j\} \cup J_0, t \in T_0 \cup T_1\}$$

and let $\hat{\theta}_{n,j}$ be the ordinary least squares estimator of θ in (18) using the data $X_j^{(n)}$, including indicator variables appropriately in order to account for η_j and γ_t. Note that in this case the $X_j^{(n)}$ are not disjoint. We may also express $\hat{\theta}_{n,j}$ more simply as

$$\hat{\theta}_{n,j} = \Delta_{n,j} - \frac{1}{|J_0|} \sum_{k \in J_0} \Delta_{n,k} \ , \tag{19}$$

where

$$\Delta_{n,k} = \frac{1}{|T_1|} \sum_{t \in T_1} Y_{k,t} - \frac{1}{|T_0|} \sum_{t \in T_0} Y_{k,t} \ .$$

It follows that for θ as in (18),

$$\begin{aligned} \sqrt{|T_1|}(\hat{\theta}_{n,j} - \theta) \quad &= \quad \sqrt{|T_1|} \left(\frac{1}{|T_1|} \sum_{t \in T_1} \epsilon_{j,t} - \frac{1}{|T_0|} \sum_{t \in T_0} \epsilon_{j,t} \right) \\ &\qquad - \sqrt{|T_1|} \frac{1}{|J_0|} \sum_{k \in J_0} \left(\frac{1}{|T_1|} \sum_{t \in T_1} \epsilon_{k,t} - \frac{1}{|T_0|} \sum_{t \in T_0} \epsilon_{k,t} \right) \ . \end{aligned}$$

For this choice of $X_j^{(n)}$ and $\hat{\theta}_{n,j}$, the convergence (15) (with $|T_1|$ in place of n) therefore holds under $\{P_n \in \mathbf{P}_{n,0} : n \geq 1\}$ with $P_n = Q_n A_{n,\theta_0}^{-1}$ under weak assumptions on $\{Q_n \in \mathbf{Q}_n : n \geq 1\}$. In particular, it suffices to assume that $\epsilon_j = (\epsilon_{j,t} : t \in T_0 \cup T_1)$ are independent across j, that for $1 \leq \ell \leq 2$

$$\frac{1}{|J_0|^2} \sum_{k \in J_0} \left(\frac{1}{|T_\ell|} \sum_{t \in T_\ell} \sum_{s \in T_\ell} E[\epsilon_{k,t} \epsilon_{k,s}] \right) \to 0 \ , \tag{20}$$

and that

$$\left(\frac{1}{\sqrt{|T_1|}} \sum_{t \in T_1} \epsilon_{j,t} \ , \frac{1}{\sqrt{|T_0|}} \sum_{t \in T_0} \epsilon_{j,t} : j \in J_1 \right) \tag{21}$$

satisfies a central limit theorem (see, e.g., Politis et al., 1999, Theorem B.0.1).

Remark 4.5. The construction described above relies on the fact that $\min\{|T_0|, |T_1|\} \to \infty$ in order to apply an appropriate central limit theorem to (21). The construction remains valid, however, even if $|T_0|$ and $|T_1|$ are small provided that

$$\frac{1}{|T_1|} \sum_{t \in T_1} \epsilon_{j,t} \text{ and } \frac{1}{|T_0|} \sum_{t \in T_0} \epsilon_{j,t}$$

are independent and identically distributed. This property will hold, for example, if $|T_0| = |T_1|$ (which may be enforced by ignoring some time periods if necessary) and the distribution of ϵ_j is exchangeable (across t) for all j. While these assumptions may be strong, this discussion illustrates that the estimators $\hat{\theta}_{n,j}$ of θ need not even be consistent in order to apply our methodology. ■

Remark 4.6. The construction described above applies equally well in the case where (18) includes covariates $Z_{j,t}$. The estimators $\hat{\theta}_{n,j}$ of θ can no longer be expressed as in (19), but they may still be obtained using ordinary least squares using the jth cluster of data. Under an appropriate modification of the assumptions to account for the $Z_{j,t}$, the convergence (15) again holds under $\{P_n \in \mathbf{P}_{n,0} : n \geq 1\}$ with $P_n = Q_n A_{n,\theta_0}^{-1}$. ■

Remark 4.7. The requirement that ϵ_j are independent across j can be relaxed using mixing conditions as in Conley and Taber (2011). In order to do so, it must be the case that the ϵ_j can be ordered linearly. ■

Remark 4.8. The construction described above applies equally well in the case where there are multiple observations for each unit j. This situation may arise, for example, when j indexes states and individual-level data within each state is available. ■

Remark 4.9. The construction above may also be used if T_0 and T_1 vary across $j \in J_1$. In this case, we simply define $X_j^{(n)} = \{(Y_{k,t}, D_{k,t}) : k \in J_0 \cup \{j\}, t \in T_{0,j} \cup T_{1,j}\}$. ■

Remark 4.10. The requirement that $|J_0| \to \infty$ can be relaxed by modifying our proposed test in the following way. Suppose $|J_0|$ is fixed and that $|J_1| \leq |J_0|$ (if this is not the case, then simply relabel treatment and control). Denote by $\{\tilde{J}_{0,l} : 1 \leq l \leq q\}$ a partition of J_0. For each $j \in J_1$, define

$$X_j^{(n)} = \{(Y_{k,t}, D_{k,t}) : k \in \tilde{J}_{0,j} \cup \{j\}, t \in T_0 \cup T_1\}$$

and let $\hat{\theta}_{n,j}$ be computed as before using the data $X_j^{(n)}$. For this choice of $X_j^{(n)}$ and $\hat{\theta}_{n,j}$, the convergence (15) continues to hold when $P_n \in \mathbf{P}_{n,0}$ for all $n \geq 1$ under appropriate modifications of the assumptions described above. ■

4.2 Clustered Regression

Suppose

$$Y_{i,j} = \theta D_j + Z_{i,j}'\gamma + \epsilon_{i,j} \quad \text{with } E[\epsilon_{i,j}|D_j, Z_{i,j}] = 0 \ . \tag{22}$$

Here, the observed data is given by $X^{(n)} = \{(Y_{i,j}, Z_{i,j}, D_j) : i \in I_j, j \in J_0 \cup J_1\} \sim P_n$ taking values on a sample space $\mathcal{X}_n = \prod_{i \in I_j, j \in J_0 \cup J_1} \mathbf{R} \times \mathbf{R}^d \times \{0,1\}$, where $Y_{i,j}$ is the outcome of unit i in area j, $Z_{i,j}$ is a vector of covariates of unit i in area j, D_j is the treatment status of area j, I_j is the set of units in area j, J_1 is the set of treated areas, and J_0 is the set of untreated areas. The scalar random variable $\epsilon_{i,j}$ is unobserved, $\gamma \in \Gamma \subseteq \mathbf{R}^d$ is a nuisance parameter, and $\theta \in \Theta \subseteq \mathbf{R}$ is the parameter of interest. The mean independence requirement is stronger than needed; indeed, all that is required is that the $\epsilon_{i,j}$ is uncorrelated with D_j and $Z_{i,j}$. For simplicity, we assume below that $|J_0| = |J_1| = q$, but the arguments are easily adapted to the case where $|J_0| \neq |J_1|$.

As before, in order to state the null and alternative hypotheses formally, it is useful to introduce some further notation. Let $W^{(n)} = \{(\epsilon_{i,j}, D_j, Z_{i,j}) : i \in I_j, j \in J_0 \cup J_1\} \sim Q_n \in \mathbf{Q}_n$ taking values

on a sample space $\mathcal{W}_n = \prod_{i\in I_j, j\in J_0\cup J_1} \mathbf{R} \times \{0,1\} \times \mathbf{R}^d$ and $A_{n,\theta,\gamma} : \mathcal{W}_n \to \mathcal{X}_n$ be the mapping implied by (22). Our assumptions on $\mathbf{Q}_n$ are discussed below. Using this notation, define

$$\mathbf{P}_n = \bigcup_{\theta\in\Theta,\gamma\in\Gamma} \mathbf{P}_n(\theta,\gamma) \text{ with } \mathbf{P}_n(\theta,\gamma) = \{Q_n A_{n,\theta,\gamma}^{-1} : Q_n \in \mathbf{Q}_n\} ,$$

where, as before, $A_{n,\theta,\gamma}^{-1}$ denotes the pre-image of $A_{n,\theta,\gamma}$. The null and alternative hypotheses of interest are thus given by (12) with

$$\mathbf{P}_{n,0} = \bigcup_{\gamma\in\Gamma} \mathbf{P}_n(\theta_0,\gamma) .$$

In order to apply our methodology, we must again specify $X_j^{(n)}$ and $\hat{\theta}_{n,j}$ and argue that the convergence (15) holds under weak assumptions on the sequence $\{P_n \in \mathbf{P}_{n,0} : n \geq 1\}$. Note that the clusters cannot be defined by areas themselves because θ is not identified within a single area. Indeed, D_j is constant within a single area. We therefore define the clusters by forming pairs of treatment and control areas, i.e., by matching each area in J_1 with an area in J_0. In experimental settings, such pairs are often suggested by the way in which treatment status was determined (see, e.g., the empirical application in Section S.3 of the Supplemental Material). More specifically, for each $j \in J_1$, let $k(j) \in J_0$ be the area in J_0 that is matched with j. For each $j \in J_1$, define

$$X_j^{(n)} = \{(Y_{i,l}, Z_{i,l}, D_l) : i \in I_l, l \in \{j, k(j)\}\}$$

and let $\hat{\theta}_{n,j}$ be the ordinary least squares estimator of θ in (22) using the data $X_j^{(n)}$. For this choice of $X_j^{(n)}$ and $\hat{\theta}_{n,j}$, the convergence (15) holds under $\{P_n \in \mathbf{P}_{n,0} : n \geq 1\}$ with $P_n = Q_n A_{n,\theta_0,\gamma}^{-1}$ under weak assumptions on $\gamma \in \Gamma$ and $\{Q_n \in \mathbf{Q}_n : n \geq 1\}$. Some such conditions can be found in Bester et al. (2011, Lemma 1).

Remark 4.11. In the application described in this section as well as the one described in the previous section when both $|J_0|$ and $|J_1|$ are small (see Remark 4.10), our methodology requires the researcher to match treated units and control units. While there may be a natural way of doing so in some empirical settings (see, e.g., Section 4.1), this may not be the case in all empirical settings. The test proposed by Ibragimov and Müller (2016), which can be used in these applications, may therefore sometimes be an attractive alternative in that it does not require the researcher to match treated units and control units in this way. However, unlike the approach proposed in this paper, their test, which relies on a generalization of the result by Bakirov and Székely (2006) described in Section 2.1 to two-sample problems, may be quite conservative even under restrictive homogeneity assumptions. To illustrate this point, consider the application described in this section with $|J_0| = |J_1| = 3$ and suppose that the data is i.i.d. across both $i \in I_j$ and $j \in J_0 \cup J_1$. Even under such strong assumptions, the limiting rejection probability of their test with a nominal level of 5% when the null hypothesis is true is approximately 1%. This same probability when $|J_0| = |J_1| = 8$ is 3%. This conservativeness stems from the rule they use for choosing the degrees of freedom for the quantile of the t-distribution with which they compare their test statistic. ■

A Optimality of Randomization Test

Define

$$\begin{aligned}\mathbf{P} &= \{\otimes_{j=1}^{q} P_{j,\mu} : P_{j,\mu} = N(\mu, \sigma_j^2) \text{ with } \mu \geq 0 \text{ and } \sigma_j^2 \geq 0\} \\ \mathbf{P}_0 &= \{\otimes_{j=1}^{q} P_{j,\mu} : P_{j,\mu} = N(\mu, \sigma_j^2) \text{ with } \mu = 0 \text{ and } \sigma_j^2 \geq 0\} \ .\end{aligned}$$

Let $X = (X_1, \ldots, X_q) \sim P \in \mathbf{P}$ consider testing (3) at level $\alpha \in (0,1)$. Below we argue that the randomization test with $T(X) = T_{t-\text{stat}}(X)$ and $\mathbf{G} = \{-1,1\}^q$ is the uniformly most powerful unbiased level α test against the restricted class of of alternatives with $\sigma_j^2 = \sigma^2 > 0$ for all $1 \leq j \leq q$. A similar argument can be used to establish the corresponding two-sided result for the randomization test with $T(X) = T_{|t-\text{stat}|}(X)$ and $\mathbf{G} = \{-1,1\}^q$ when $\mathbf{P}$ and $\mathbf{P}_0$ according to (10)-(11). Related results have been obtained previously in Lehmann and Stein (1949).

Consider a test $\tilde{\phi}(X) = \tilde{\phi}(X_1, \ldots, X_q)$. Since the test is unbiased, it must be the case that $E_P[\tilde{\phi}(X)] \leq \alpha$ for all $P \in \mathbf{P}_0$ and $E_P[\tilde{\phi}(X)] \geq \alpha$ for all $P \in \mathbf{P}_1$. Using the dominated convergence theorem, it is straightforward to show that the requirement of unbiasedness therefore implies that the test is similar, i.e., $E_P[\tilde{\phi}(X)] = \alpha$ for all $P \in \mathbf{P}_0$.

Next, note that $U = (|X_1|, \ldots, |X_n|)$ is sufficient for $\mathbf{P}_0$. Indeed, the distribution of $X|U$ under any $P \in \mathbf{P}_0$ is uniform over the 2^n points of the form $(\pm|X_1|, \ldots, \pm|X_n|)$. Furthermore, $\mathbf{P}_0^U$, the family of distributions for U under P as P varies over $\mathbf{P}_0$, is complete. To see this, for $\gamma \in \mathbf{R}^n$, define P_γ to be the distribution with density

$$C(\gamma) \exp\left(-\sum_{j=1}^{n} \gamma_j x_j^2\right) \ ,$$

where $C(\gamma)$ is an appropriate constant. By construction, $P_\gamma \in \mathbf{P}_0$, so the desired result follows from Theorem 4.3.1 in Lehmann and Romano (2005). Therefore, by Theorem 4.3.2 in Lehmann and Romano (2005), we see that all similar tests have Neyman structure, i.e., $E_P[\tilde{\phi}(X)|U = u] = \alpha$ for all $P \in \mathbf{P}_0$ and all u except those in a set N such that $\sup_{P \in \mathbf{P}_0} P\{U \in N\} = 0$.

To find an optimal test, we therefore maximize the power of the test under $P = \otimes_{j=1}^{q} N(\mu, \sigma^2)$ where $\mu > 0$ and $\sigma^2 > 0$. Under the null, the distribution of $X|U$ is uniform, as described above. Under this alternative, the conditional probability mass function is proportional to

$$\prod_{1 \leq i \leq n} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) = \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{1 \leq i \leq n} x_i^2 - 2\mu \sum_{1 \leq i \leq n} x_i + \mu^2\right)\right) \ .$$

Since $\sum_{1 \leq i \leq n} X_i^2$ is constant conditional on $U = u$, the Neyman-Pearson Lemma implies that the optimal (conditional) test rejects when $\sum_{1 \leq i \leq n} X_i > c(u)$ and rejects with probability $\gamma(u)$ when $\sum_{1 \leq i \leq n} X_i = c(u)$, where the constants $c(u)$ and $\gamma(u)$ are chosen so that the test has (conditional)

rejection probability equal to α. Such tests are, of course, randomization tests with underlying choice of test statistic equal to $\sum_{1\leq i\leq n} X_i$, and this test is identical to the randomization test with underlying choice of test statistic equal to $T_{t-\text{stat}}(X)$ (see Example 15.2.4 in Lehmann and Romano (2005) for details). Denote this test by $\phi(X)$.

It remains to show that $\phi(X)$ is indeed unbiased. By construction, it is similar and therefore has rejection probability $= \alpha$ for all $P \in \mathbf{P}_0$. To see that the rejection probability is $\geq \alpha$ under any $P \in \mathbf{P}_1$, note that $\phi(X)$ is weakly increasing in each of its arguments. We therefore have that $E_P[\phi(X_1+\mu,\ldots,X_n+\mu)] \geq \alpha$ for all $\mu > 0$ and any $P \in \mathbf{P}_0$, from which the desired result follows.

Remark A.1. It is important to emphasize that this optimality result, like the one in Ibragimov and Müller (2010), is only for a restricted class of alternatives. On the other hand, it can readily be shown that the specified randomization test is in fact admissible whenever the set of alternatives contains this class and α is a multiple of $\frac{1}{2^q}$. The argument hinges on the fact that the above argument using the Neyman-Pearson lemma together with Lemma S.5.1 in the Supplemental Material guarantees that the optimal test is non-randomized for these values of α. ■

Remark A.2. The argument presented above in fact shows that the specified randomization test remains uniformly most powerful unbiased against the same class of alternatives even if $\mathbf{P}_0$ is enlarged so that each $P_{j,\mu}$ is only required to be symmetric about zero. ■

B Proof of Theorem 3.1

Let $\{P_n \in \mathbf{P}_{n,0} : n \geq 1\}$ satisfying Assumption 3.1 be given and define $M = |\mathbf{G}|$. By Assumption 3.1(i) and the Almost Sure Representation Theorem (c.f van der Vaart, 1998, Theorem 2.19), there exists $\tilde{S}_n$, $\tilde{S}$, and $U \sim U(0,1)$, defined on a common probability space $(\Omega, \mathcal{A}, P)$, such that

$$\tilde{S}_n \to \tilde{S} \text{ w.p.1 } ,$$

$\tilde{S}_n \overset{d}{=} S_n$, $\tilde{S} \overset{d}{=} S$, and $U \perp (\tilde{S}_n, \tilde{S})$. Consider the randomization test based on $\tilde{S}_n$, this is,

$$\tilde{\phi}(\tilde{S}_n, U) \equiv \begin{cases} 1 & T(\tilde{S}_n) > T^{(k)}(\tilde{S}_n) \text{ or } T(\tilde{S}_n) = T^{(k)}(\tilde{S}_n) \text{ and } U < a(\tilde{S}_n) \\ 0 & T(\tilde{S}_n) < T^{(k)}(\tilde{S}_n) \end{cases} .$$

Denote the randomization test based on $\tilde{S}$ by $\tilde{\phi}(\tilde{S}, U)$, where the same uniform variable U is used in $\tilde{\phi}(\tilde{S}_n, U)$ and $\tilde{\phi}(\tilde{S}, U)$.

Since $\tilde{S}_n \overset{d}{=} S_n$, it follows immediately that $E_{P_n}[\phi(S_n)] = E_P[\tilde{\phi}(\tilde{S}_n, U)]$. In addition, since $\tilde{S} \overset{d}{=} S$, Assumption 3.1(ii) and Theorem 2.1 imply that $E_P[\tilde{\phi}(\tilde{S}, U)] = \alpha$. It therefore suffices to show

$$E_P[\tilde{\phi}(\tilde{S}_n, U)] \to E_P[\tilde{\phi}(\tilde{S}, U)] \ . \tag{23}$$

In order to show (23), let E_n be the event where the orderings of $\{T(g\tilde{S}) : g \in \mathbf{G}\}$ and $\{T(g\tilde{S}_n) : g \in \mathbf{G}\}$ correspond to the same transformations $g_{(1)}, \ldots, g_{(M)}$. We first claim that $I\{E_n\} \to 1$ w.p.1. To see this, note that by Assumption 3.1(iii) and $\tilde{S} \stackrel{d}{=} S$, any two $g, g' \in \mathbf{G}$ are such that either

$$T(gs) = T(g's) \ \forall s \in \mathcal{S} \ , \tag{24}$$

or

$$T(g\tilde{S}) \neq T(g'\tilde{S}) \text{ w.p.1 under } P \ . \tag{25}$$

It follows that there exists a set with probability one under P such that for all $\omega \in \Omega$ in this set, $\tilde{S}_n(\omega) \to \tilde{S}(\omega)$ and $T(g\tilde{S}(\omega)) \neq T(g'\tilde{S}(\omega))$ for any two $g, g' \in \mathbf{G}$ not satisfying (24). For any ω in this set, let $g_{(1)}(\omega), \ldots, g_{(M)}(\omega)$ be the transformations such that

$$T(g_{(1)}(\omega)\tilde{S}(\omega)) \leq T(g_{(2)}(\omega)\tilde{S}(\omega)) \leq \cdots \leq T(g_{(M)}(\omega)\tilde{S}(\omega)) \ .$$

For any two consecutive elements $g_{(j)}(\omega)$ and $g_{(j+1)}(\omega)$ with $1 \leq j \leq M-1$, there are only two possible cases: either $T(g_{(j)}(\omega)\tilde{S}(\omega)) = T(g_{(j+1)}(\omega)\tilde{S}(\omega))$ or $T(g_{(j)}(\omega)\tilde{S}(\omega)) < T(g_{(j+1)}(\omega)\tilde{S}(\omega))$. If $T(g_{(j)}(\omega)\tilde{S}(\omega)) = T(g_{(j+1)}(\omega)\tilde{S}(\omega))$ then by (24) it follows that

$$T(g_{(j)}(\omega)\tilde{S}_n(\omega)) = T(g_{(j+1)}(\omega)\tilde{S}_n(\omega)) \ \forall n \geq 1 \ .$$

If $T(g_{(j)}(\omega)\tilde{S}(\omega)) < T(g_{(j+1)}(\omega)\tilde{S}(\omega))$, then

$$T(g_{(j)}(\omega)\tilde{S}_n(\omega)) < T(g_{(j+1)}(\omega)\tilde{S}_n(\omega)) \ \text{ for } n \text{ sufficiently large} \ ,$$

as $\tilde{S}_n(\omega) \to \tilde{S}(\omega)$ and the continuity of $T : \mathcal{S} \to \mathbf{R}$ and $g : \mathcal{S} \to \mathcal{S}$ imply that $T(g_{(j)}(\omega)\tilde{S}_n(\omega)) \to T(g_{(j)}(\omega)\tilde{S}(\omega))$ and $T(g_{(j+1)}(\omega)\tilde{S}_n(\omega)) \to T(g_{(j+1)}(\omega)\tilde{S}(\omega))$. We can therefore conclude that

$$I\{E_n\} \to 1 \ w.p.1 \ ,$$

which proves the first claim.

We now prove (23) in two steps. First, we note that

$$E_P[\tilde{\phi}(\tilde{S}_n, U)I\{E_n\}] = E_P[\tilde{\phi}(\tilde{S}, U)I\{E_n\}] \ . \tag{26}$$

This is true because, on the event E_n, if the transformation $g = g_{(m)}$ corresponds to the mth largest value of $\{T(g\tilde{S}) : g \in \mathbf{G}\}$, then this same transformation corresponds to the mth largest value of $\{T(g\tilde{S}_n) : g \in \mathbf{G}\}$. In other words, $\tilde{\phi}(\tilde{S}_n, U) = \tilde{\phi}(\tilde{S}, U)$ on E_n. Second, since $I\{E_n\} \to 1$ $w.p.1$ it follows that $\tilde{\phi}(\tilde{S}, U)I\{E_n\} \to \tilde{\phi}(\tilde{S}, U)$ w.p.1 and $\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\} \to 0$ w.p.1. We can therefore use (26) and invoke the dominated convergence theorem to conclude that,

$$\begin{aligned} E_P[\tilde{\phi}(\tilde{S}_n, U)] &= E_P[\tilde{\phi}(\tilde{S}_n, U)I\{E_n\}] + E_P[\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\}] \\ &= E_P[\tilde{\phi}(\tilde{S}, U)I\{E_n\}] + E_P[\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\}] \\ &\to E_P[\tilde{\phi}(\tilde{S}, U)] \ . \end{aligned}$$

This completes the proof. ■

References

Angrist, J. D. and Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American Economic Review* 1384–1414.

Bakirov, N. K. and Székely, G. (2006). *Journal of Mathematical Sciences*, **139** 6497–6505.

Bester, C. A., Conley, T. G. and Hansen, C. B. (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, **165** 137–151.

Canay, I. A. and Kamat, V. (2015). Approximate permutation tests and induced order statistics in the regression discontinuity design. Tech. rep., CeMMAP working paper CWP27/15.

Canay, I. A., Romano, J. P. and Shaikh, A. M. (2015). Supplement to "Randomization tests under an approximate symmetry assumption". Manuscript.

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

Chung, E. and Romano, J. P. (2016a). Asymptotically valid and exact permutation tests based on two-sample U-statistics. *Journal of Statistical Planning and Inference*, **168** 97–105.

Chung, E. and Romano, J. P. (2016b). Multivariate and multiple permutation tests. Working Paper.

Conley, T. G. and Taber, C. R. (2011). Inference with "difference in differences" with a small number of policy changes. *The Review of Economics and Statistics*, **93** 113–125.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics*, **23** 169–192.

Ibragimov, R. and Müller, U. K. (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics*, **28** 453–468.

Ibragimov, R. and Müller, U. K. (2016). Inference with few heterogenous clusters. *The Review of Economics and Statistics*, **98** 83–96.

Jenish, N. and Prucha, I. R. (2009). Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of econometrics*, **150** 86–98.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. 3rd ed. Springer, New York.

Lehmann, E. L. and Stein, C. (1949). On the theory of some non-parametric hypotheses. *The Annals of Mathematical Statistics* 28–45.

Politis, D. N., Romano, J. P. and Wolf, M. (1999). *Subsampling*. Springer, New York.

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *The Annals of Statistics*, **17** 141–159.

Romano, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, **85** 686–692.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.