Permutation Tests at Nonparametric Rates

Marinho Bertanha\textsuperscript{a} and Eunyi Chung\textsuperscript{b}

\textsuperscript{a}Department of Economics, University of Notre Dame, Notre Dame, IN; \textsuperscript{b}Department of Economics, University of Illinois at Urbana-Champaign, Urbana, IL

\textbf{ABSTRACT}

Classical two-sample permutation tests for equality of distributions have exact size in finite samples, but they fail to control size for testing equality of parameters that summarize each distribution. This article proposes permutation tests for equality of parameters that are estimated at root-$n$ or slower rates. Our general framework applies to both parametric and nonparametric models, with two samples or one sample split into two subsamples. Our tests have correct size asymptotically while preserving exact size in finite samples when distributions are equal. They have no loss in local asymptotic power compared to tests that use asymptotic critical values. We propose confidence sets with correct coverage in large samples that also have exact coverage in finite samples if distributions are equal up to a transformation. We apply our theory to four commonly-used hypothesis tests of nonparametric functions evaluated at a point. Lastly, simulations show good finite sample properties, and two empirical examples illustrate our tests in practice. Supplementary materials for this article are available online.

\textbf{1. Introduction}

Applications of permutation tests have gained widespread popularity in empirical analyses in the social and natural sciences. Classical two-sample permutation tests appeal to applied researchers because they are easy to implement and have exact size in finite samples under the so-called “sharp null hypothesis.” The sharp null hypothesis states that two population distributions are equal. However, researchers are often interested in testing equality of parameters that summarize the distributions. For example, one may want to test equality of average outcomes between treatment and control groups while nonparametrically controlling for age and income. Classical permutation tests fail to control size under such nulls, in both finite and large samples.\textsuperscript{1}

This article proposes robust two-sample permutation tests for equality of parameters that are estimated at root-$n$ or slower rates. The tests are robust in the sense that they control size asymptotically while preserving finite-sample exactness under the sharp null. Our general framework covers both parametric and nonparametric models, in cases with two samples from two populations or one sample from a union of two populations. In addition, the article makes three further contributions. First, we derive the asymptotic permutation distribution in our general framework under both null and alternative hypotheses, which requires novel technical arguments. Second, we provide four examples of tests in widely used nonparametric models, and we prove that they satisfy the conditions of our framework. Third, we construct robust confidence sets for differences between the two populations. The confidence sets are robust meaning that they have correct coverage asymptotically and exact coverage in finite samples if the sharp null holds under a class of transformations.

Our framework considers a summary parameter that can be consistently estimated using an asymptotically linear statistic. The influence function depends on the data, the population distribution, and the sample size. There may be two iid samples from two populations, or one iid sample from a union of two populations. In the case of one sample, there is a variable in the data that identifies the population of each observation, and the sample is split into two. The researcher applies the estimator to each of the two samples, computes the difference, and tests whether the two parameters are equal. The classical permutation test compares the estimated difference to critical values from the permutation distribution, that is, the distribution of estimates over all permutations of observations across the two samples.

We derive the asymptotic permutation distribution of the estimated difference, both with and without the null hypothesis, and find it to be generally different from its asymptotic sampling distribution. This leads the classical permutation test to have incorrect size, as shown in a variety of other settings (see related literature below). The derivation has two key technical features. First, we borrow the coupling approximation from Chung and Romano (2013) (henceforth CR) and apply it to our setting. The approximation consists of two steps: (a) couple the original sample with a new sample from a particular mixture of the two populations; and (b) prove that replacing the original sample with the new coupled sample renders no change in the

\textsuperscript{1}The lack of size control of the classical permutation test outside of the sharp null has been studied for a long time (e.g., Romano 1990). Theorem 2.1 confirms the lack of size control in our setting.
asymptotic permutation distribution of the test statistic. We use the coupling approximation because it is more tractable to study the limiting behavior of permutation tests when both samples are drawn iid from the same mixture distribution, rather than different distributions. Our coupling approximation differs from that of CR in that our proof does not assume the null hypothesis. This allows us to study both size and power of the permutation test. Our proof requires a new argument to bound the variance of the approximation error (Section B.2.3 in the supplementary materials). The second key technical feature is that our theory allows for random sample sizes. Sample sizes are random when the researcher splits one sample into two as a function of the data. Thus, the derivation of the limiting distribution must be valid conditional on any sequence of sample splits that occurs with probability one and requires novel arguments: (a) a conditional central limit theorem of weighted sums of triangular arrays, where the weights are random (Lemmas F.3 and F.4 in the supplementary materials); and (b) an asymptotic linear representation for estimators that must hold uniformly over convex combinations of the two populations (Assumption 2.1).

Our proposed permutation test uses a studentized test statistic, which is the estimated difference of parameters divided by a consistent estimator of its standard deviation. We then show that both the asymptotic permutation and sampling distributions are standard normal. It follows that our permutation test has correct size in large samples, and its asymptotic power against local alternatives is identical to that of the test that relies on critical values from a standard normal. Finally, we construct a confidence set by inverting our test, which requires testing null hypotheses that are more general than simple equality of parameters. We propose ways to transform the data in order to test more general hypotheses and preserve finite sample exactness when populations are equal up to the data transformations.

Examples of applications of permutation tests abound. Table 3 in the supplementary materials lists top-publications from the last decade in a variety of disciplines that use permutation tests. This broad applicability motivates our extension of the theory. We illustrate our framework using four nonparametric examples of hypothesis tests that are often used in empirical studies. The first and second examples test equality at a point of nonparametric conditional mean and quantile functions, respectively. The third and fourth examples test continuity at a point of nonparametric conditional mean or probability density functions (PDFs), respectively. We explain how to implement the permutation test in each case and give sufficient conditions to derive the limiting permutation distribution. Implementation requires a sign change and sample splitting in the third and fourth examples. We find that asymptotic size control requires studentization, except in the fourth example.

Related Literature

Randomization inference has recently received keen attention (Canay, Romano, and Shaikh 2017; Shaikh and Toulis 2021). For the case of permutation inference, the insight of robustness through studentization has been proposed before in specific settings that differ from our general framework: Neuhaus (1993), Janssen (1997), Janssen (2005), Neubert and Brunner (2007), Chung and Romano (2013), Pauly, Brunner, and Konietzche (2015), Chung and Romano (2016a), and DiCiccio and Romano (2017). In particular, the framework of Janssen (1997) applies to the problem of testing difference of means, while CR study the more general problem of difference of parameters that are estimable at root-\(n\). It is important to emphasize that this article is not a straightforward generalization of their work, and none of our applications fits CR’s framework. For example, CR handles the difference of parametric quantiles for two independent samples; in contrast, our framework covers the cases of parametric or nonparametric quantiles, for two independent samples or one sample randomly split into two. Moreover, our verification of the coupling approximation differs from that of CR because we do not assume the null hypothesis. All these features make many of our proofs substantially different from theirs.

Studentization is not the only way to achieve robustness. Robustness is also obtained through prepivoting (Chung and Romano 2016b; Fogarty 2021) and the Khmaladze transformation (Chung and Olives 2021). Prepivoting obtains robustness in examples where studentization does not help. For instance, consider a vector of differences in means where one is interested in the maximum absolute difference. The asymptotic distribution cannot be made pivotal by simply studentizing the maximum statistic. Likewise, the Khmaladze transformation is applied to empirical processes to make them asymptotically pivotal.

Previous works have also considered randomization tests for continuity of nonparametric models at a point. Cattaneo, Frandsen, and Titiunik (2015) propose local randomization inference procedures for a sharp null hypothesis, while Canay and Kram (2018) provide permutation tests for continuity of the whole distribution of an outcome variable conditional on a control variable at a point. In contrast, our permutation test applies to testing continuity of summary statistics of the conditional distribution such as mean, quantile, variance, etc. Our fourth example is related to Bugni and Canay (2021), who propose a sign-change test for continuity of PDFs at a point, where critical values come from maximizing a function of a binomial distribution. We show how the same null hypothesis fits into our framework and is simply testable using permutations. The last two papers use the insight that non-iid order statistics converge in distribution to iid variables, which is technically distinct from our coupling approximation. Finally, permutation-based confidence sets have previously been proposed only in specific settings. For example, the confidence sets of Imbens and Rosenbaum (2005) assume that treatment effects divided by treatment doses are constant across individuals, and that the distribution of treatment eligibility is known. To the best of our knowledge, ours is the first paper to provide valid two-sample permutation tests and confidence sets for scalar parameters estimated at nonparametric rates.

The rest of this article is outlined as follows. Section 2 presents the general framework, assumptions, and asymptotic distributions of the classical and robust permutation tests. Section 3 studies how our theory applies to four nonparametric examples. Section 4 explains how to invert permutation tests to build robust confidence sets. Section 5 displays a simulation study that confirms our theory and illustrates good finite sample properties of the robust permutation test. Section 6 illustrates the practical relevance of our procedures to test for continuity.
of conditional mean functions using real-world data. The supplementary materials contains all proofs.

2. Theory

Consider two populations $P_1$ and $P_2$, and a real-valued parameter $\theta(P_k)$ summarizing distribution $P_k$, $k = 1, 2$. The null hypothesis is stated as $H_0 : \theta(P_1) = \theta(P_2)$. Note that our null hypothesis is a bigger set of distributions compared to the set of distributions in the so-called “sharp null hypothesis,” respectively, $\{(P_1, P_2) : \theta(P_1) = \theta(P_2)\}$ versus $\{(P_1, P_2) : P_1 = P_2\}$. For each population $k$, there are iid observations $Z_{k,i} \in \mathbb{R}^d$ from distribution $P_k$, $i = 1, 2, \ldots, n_k$. Observations are independent across $k$, and the total number of observations is $n = n_1 + n_2$. We define $P$ to be the convex hull of $\{P_1, P_2\}$, that is, $P = \{P : P = \eta P_1 + (1 - \eta) P_2, \ \eta \in [0, 1]\}$. Throughout this article, random variables with subscript “$k$” indicate they have distribution $P_k$, for example, $Z_k$. For any other distribution in $P$, the random variable is denoted $V \in \mathbb{R}^d$. Operators such as $\mathbb{P}$ (probability), $\mathbb{E}$ (expectation), or $\mathbb{V}$ (variance) applied to $Z_k$ do not carry the subscript $P_k$, but operators applied to $V$ carry the subscript $P$, for example, $\mathbb{E}[Z_k]$ versus $\mathbb{E}[V]$. The parameter $\theta(P_k)$ is consistently estimated by $\hat{\theta}_k = \theta_{n_1,k}(Z_{k,1}, \ldots, Z_{k,n_k})$, where the functions $\theta_{n_1,k}$ and $\theta_{n_2,k}$ satisfy the following assumption.

Assumption 2.1. Let $V_1, \ldots, V_m$ be an iid sample from a distribution $P \in P$. Let $m$ grow with $n$ such that $m/n \to \gamma$, for some $\gamma \in (0, 1)$. Use these observations to construct the estimator $\hat{\theta} = \theta_{m,n}(V_1, \ldots, V_m)$. Assume the following objects exist: a sequence of functions $\psi_n : \mathbb{R}^d \times \mathcal{P} \to \mathbb{R}$, a function $\xi : \mathcal{P} \to \mathbb{R}$, and a nonincreasing sequence $h_n$ such that $nh_n \to \infty$. Further assume

$$\forall \varepsilon > 0 : \sup_{P \in \mathcal{P}} \mathbb{P}_P \left\{ \left| \sqrt{m} h_n (\hat{\theta} - \theta(P)) \right| > \varepsilon \right\} \to 0, \quad (2.1)$$

$$\mathbb{E}_P[\psi_n(V_i, P)] = 0 \ \forall P \in \mathcal{P}, \quad (2.2)$$

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P [\mathbb{V}_P[\psi_n(V_i, P)]] - \xi^2(P) \to 0, \quad (2.3)$$

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ \psi_n(Z_{k}, P_k) \right] < \infty, \text{ for } k \in \{1, 2\}, \quad (2.4)$$

$$\exists \xi > 0 : n^{-\varepsilon/2} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ \frac{\psi_n(V_i, P)}{\psi_n(V_i, P)} \right]^{2+\xi} \to 0, \text{ and} \quad (2.5)$$

$$\xi^2 \left( \frac{m}{n} P_1 + \frac{n - m}{n} P_2 \right) \to \xi^2 (\gamma P_1 + (1 - \gamma) P_2). \quad (2.6)$$

Assumption 2.1 requires the estimator $\hat{\theta}$ to have an asymptotically linear expansion at rate $\sqrt{nh}$ (Equation (2.1)) with zero-mean influence function $\psi_n$ (Equation (2.2)) and asymptotic variance $\xi^2$ (Equation (2.3)). The expansion must hold for data drawn from any combination of $P_1$ and $P_2$ in $P$ because the limiting permutation distribution of $\hat{\theta}_1 - \hat{\theta}_2$ behaves as if the data were drawn from a sequence of mixtures of $P_1$ and $P_2$ (Theorems 2.1 and 2.2). Assumption 2.1 also imposes bounds on moments of $\psi_n$ (Equations (2.4)–(2.5)) to enable us to apply a central limit theorem and derive the limiting permutation distribution. This requires $\xi^2(P)$ to be smooth with respect to $P$ (Equation (2.6)) in order for the limit of the variance $\xi^2 \left( \frac{m}{n} P_1 + \frac{n - m}{n} P_2 \right)$ to be well-defined.

Situations arise where the number of observations $n_k$ is random rather than deterministic. For example, suppose the researcher desires to compare the female and male subpopulations of a country but only has one iid sample with $n$ individuals from that country. The researcher splits the sample into two subsamples based on the gender of each observation and sample sizes are random. In order to accommodate both deterministic and random sample sizes, we consider a sampling scheme which is dictated by a vector of indicator variables $W_n = (W_1, \ldots, W_n)$, $W_i \in \{1, 2\}$ for $i = 1, \ldots, n$, where $W_n$ has distribution $Q_n$. Conditional on $W_n$, the sample $Z_n = (Z_1, \ldots, Z_n)$ has $Z_i$ drawn from distribution $P_k$ if $W_i = 1$ or from distribution $P_2$ if $W_i = 2$, with observations independent across $i$. This accommodates the standard two-population sampling by making $W_n$ nonrandom with $n_1$ entries equal to 1, $n_2$ entries equal to 2, and $n_1 + n_2 = n$. It also accommodates the example above of male and female subpopulations by making $W_i$ iid and $\text{Pr}(W_i = 1) = \gamma$ equal to the probability of being female. Conditional on $W_n$, there are $n_k$ iid observations $Z_{k,i} \in \mathbb{R}^d$ from distribution $P_k$, $k \equiv 1, 2$, and observations are independent across $k$. As before, $\hat{\theta}_k = \theta_{n_k,k}(Z_{k,1}, \ldots, Z_{k,n_k})$.

Assumption 2.2. There exists $\lambda \in (0, 1)$ such that the sequence of distributions $Q_n$ satisfies $n_1/n \to \lambda$ as $n \to \infty$. Moreover, Assumption 2.1 holds for all sequences of sample sizes $n$ such that $m/n \to \lambda$ or $m/n \to 1 - \lambda$.

The test statistic $T_n$ is a function of the data $(W_n, Z_n)$ as follows:

$$T_n(W_n, Z_n) \doteq \sqrt{nh} (\hat{\theta}_1 - \hat{\theta}_2), \quad (2.7)$$

where we omit the subscript $n$ from the sequence $h_n$ of Assumption 2.1 to simplify notation.

The permutation test is constructed by permuting the order of observations in $Z_n$, while keeping the indicator variables $W_n$ unchanged, and recomputing the test statistic. A permutation is a one-to-one function $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$, where $\pi(i) = j$ says that the $j$th observation becomes the $i$th observation once permutation $\pi$ is applied. Given a permutation $\pi$, the permuted sample becomes $(W_n', Z_n') = (W_{\pi(1)}, Z_{\pi(1)}, \ldots, Z_{\pi(n)})$, and the recomputed value of the test statistic is $T_n' = T_n(W_n', Z_n')$. In other words, permutations swap individuals across the two samples to which they originally belonged according to $W_n$, which remains fixed. The set $\mathsf{G}_n$ is the set of all possible permutations $\pi$. The number of elements in $\mathsf{G}_n$ is $n!$.

The two-sided permutation test with nominal level $\alpha \in (0, 1)$ is conducted as follows. First, recompute the test statistic $T_n(W_n', Z_n')$ for every $\pi \in \mathsf{G}_n$. Rank the values of $T_n'$ across $\pi$: $T_n^{(1)} \leq T_n^{(2)} \leq \ldots \leq T_n^{(n!)}$. Second, fix a nominal level $\alpha \in (0, 1)$ and let $k^- = \lceil n!\alpha/2 \rceil$, that is, the largest integer less than or equal to $n!\alpha/2$, and $k^+ = n! - k^-$. Third, compute the following quantities: (i) $M^+$, the number of values $T_n^{(j)}$, $j = 1, 2, \ldots, k^+$, that
are strictly greater than $T_n^{(k)}$; (ii) $M^-$, the number of values $T_n^{(j)}$ that are strictly smaller than $T_n^{(k)}$; (iii) $M^0$, the number of values $T_n^{(i)}$ that are equal to either $T_n^{(k)}$ or $T_n^{(k)}$; and (iv) $a = (an! - M^+ - M^-)/M^0$. Finally, the outcome of the test is based on the test function $\phi$:

$$\phi(W_n, Z_n) = \begin{cases} 1 & \text{if } T_n > T_n^{(k)} \text{ or } T_n < T_n^{(k)} , \\ a & \text{if } T_n = T_n^{(k)} \text{ or } T_n = T_n^{(k)} , \\ 0 & \text{if } T_n^{(k)} < T_n < T_n^{(k)}. \end{cases} \quad (2.8)$$

For a given sample, if $\phi = 1$, we reject the null hypothesis; if $\phi = a$, we randomly reject the null hypothesis with probability $a$; otherwise, if $\phi = 0$, we fail to reject the null. A classic property of permutation tests is exact size in finite samples under the "sharp null," that is, the null hypothesis stating $P_1 = P_2$.

**Lemma 2.1.** For any $n$, $Q_n$, $P_1$, and $P_2$, if $P_1 = P_2$, then $E(\phi(W_n, Z_n)) = \alpha$.

**Remark 2.1.** Re-computing the test statistic for all $n!$ permutations is costly, even for small $n$. Lemma 2.1 and remaining results in this section are unchanged if, instead of $G_n$, we use a random sample of $G_n$ with or without replacement (Lehmann and Romano 2005, p. 636).

**Remark 2.2.** The randomized outcome in the case of ties is important for exact size in finite samples if $P_1 = P_2$. However, it may be desirable to have a deterministic answer to a hypothesis test after observing a sample of data. An easy way to fix that is to set $\phi = 0$ in the case of ties. This makes the test conservative, that is, the size becomes less than or equal to $\alpha$.

The set of distributions that satisfy the null hypothesis $\theta(P_1) = \theta(P_2)$ is in general larger than the set of distributions that satisfy the sharp null $P_1 = P_2$. Thus, there is no finite sample size control in general. To investigate the asymmetric properties of the test in (2.8), we derive the probability limit of the permutation distribution,

$$\mathbb{R}_{T_n}(t) = \frac{1}{n!} \sum_{\pi \in \Pi_n} I(T_n(W_n, Z_n^\pi) \leq t). \quad (2.9)$$

The hypothesis test (2.8) uses critical values from $\mathbb{R}_{T_n}$. The test has asymptotic size control if, under the null hypothesis, the probability limit of $\mathbb{R}_{T_n}$ equals the cumulative distribution function (CDF) of the limiting distribution of $T_n$. In order to study both size and power, we derive these limiting distributions without imposing the null hypothesis in the following theorems.

**Theorem 2.1.** Under Assumptions 2.1–2.2, the permutation distribution $\mathbb{R}_{T_n}$ converges uniformly in probability to the CDF of a $\mathcal{N}(0, \tau^2)$, that is, $\sup |\mathbb{R}_{T_n}(t) - \Phi(t/\tau)| \to 0$, where $\tau^2 = \xi^2(P_1) / \lambda (1 - \lambda)$ and $\bar{P} = \lambda P_1 + (1 - \lambda) P_2$. Moreover, $T_n - \sqrt{n} \hat{\lambda} (\theta(P_1) - \theta(P_2)) \Rightarrow N(0, \sigma^2)$, where $\sigma^2 = \xi^2(P_1) / \lambda + \xi^2(P_2) / (1 - \lambda)$.

The permutation distribution fails to control size asymptotically because the asymptotic variance of the permutation distribution $\tau^2$ generally differs from $\sigma^2$. To appreciate the implications of Theorem 2.1, we present the parametric example below and four nonparametric examples in Section 3 detailing cases where $\tau^2 = \sigma^2$ and $\tau^2 \neq \sigma^2$.

**Example 2.1** (Parametric Model). For $k = 1, 2$, $Z_k = (X_k, Y_k) \sim P_k$, $X_k \sim U[0, 1]$, $E[Y_k | X_k] = \theta(P_k) + \beta X_k$, and $\mathbb{V}[Y_k | X_k] = \nu_k$. There are two independent samples with $n_1/n \to \lambda$. The ordinary least squares estimator is $\hat{\theta}_k = \frac{\sum_{i=1}^{n_k} X_i Y_i}{\sum_{i=1}^{n_k} X_i^2}$, where $\overline{X_k} = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ki}$ for $s = 1, 2$, and $\overline{X_k Y_k} = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ki} Y_{ki}$. For $\mathbb{V} = (\mathbb{R}, \mathbb{S}) \sim P$, the setting satisfies Assumption 2.1 with $h = 1$; $\psi(\mathbb{V}, P) = \psi(\mathbb{V}, P) = (S - \mathbb{E}P(S|R))(4 - 6R)$; and $\xi^2(P) = \mathbb{E}P[(S - \mathbb{E}P(S|R))^2(4 - 6R)^2]$. Under the null hypothesis, the asymptotic variance of $T_n = \sqrt{n} (\hat{\theta}_1 - \hat{\theta}_2)$ and of the permutation distribution are, respectively, $\sigma^2 = 4 \left[ \frac{\nu_1}{\lambda} + \frac{\nu_2}{1 - \lambda} \right]$ and $\tau^2 = 4 \left[ \frac{\nu_1}{\lambda} + \frac{\nu_2}{1 - \lambda} \right]$. Unless $\lambda = 1/2$ or $\nu_1 = \nu_2$, $\sigma^2$ and $\tau^2$ do not agree in general, distorting the rejection probability under the null for the permutation test. Section C in the supplementary materials displays a graph that illustrates this distortion using simulated data.

Since $\tau^2$ and $\sigma^2$ are generally different, the test statistic $T_n$ must be transformed to become asymptotically pivotal. Thus, we divide $T_n$ by the square root of a consistent estimator for its asymptotic variance. For each population $k \in \{1, 2\}$, let $\hat{\xi}_k^2 = \xi_{n_k}^2(Z_{k,1}, \ldots, Z_{k,n_k})$ be a consistent estimator for $\xi^2(P_k)$ and assume the functions $\hat{\xi}_{n_1,n}^2$ and $\hat{\xi}_{n_2,n}^2$ satisfy the following assumption.

**Assumption 2.3.** Let $V_1, \ldots, V_m$ be an iid sample from a distribution $P \in \mathbb{P}$. Use these observations to construct the estimator $\hat{\xi}_k^2 = \hat{\xi}_{n_k}^2(V_1, \ldots, V_m)$. Assume that, for any sequence of sample sizes $m$ such that $m/n \to \lambda$ or $m/n \to 1 - \lambda$, $\hat{\xi}_k^2(P)$ converges in probability to 0 uniformly over $P \in \mathbb{P}$.

Then, the studentized test statistic $S_n$ is

$$S_n(W_n, Z_n) = \frac{T_n(W_n, Z_n)}{\sigma_n} \quad (2.10)$$

where $\sigma_n$ is the square root of the consistent estimator for the asymptotic variance of $T_n$, that is, $\sigma_n^2 = \frac{n \xi_{n_1}^2}{n_1} + \frac{n \xi_{n_2}^2}{n_2}$.

**Theorem 2.2.** Let $\mathbb{R}_{S_n}$ be the permutation CDF defined in (2.9) with $T_n$ replaced by $S_n$. Under Assumptions 2.1–2.3, $\mathbb{R}_{S_n}$ converges uniformly in probability to the CDF of a $\mathcal{N}(0, 1)$, that is, $\sup |\mathbb{R}_{S_n}(t) - \Phi(t)| \to 0$. Moreover, $S_n - \sqrt{n} \hat{\lambda} (\theta(P_1) - \theta(P_2)) \Rightarrow N(0, 1)$.

**Example 2.1 (Continued).** If the test statistic $T_n$ is appropriately studentized by the usual ordinary least squares formula for standard errors, both permutation and sampling distributions of $S_n$ coincide asymptotically. Section C in the supplementary materials simulates data for this example and presents these two distributions graphically.

Note that the standard deviation $\sigma_n$ that divides $T_n$ must be consistent for $\sigma$, as opposed to $\tau$. However, $\sigma_n$ is evaluated
using permuted samples, it converges in probability to \( p \). Under the null hypothesis, both the permutation distribution and the test statistic \( S_n \) are asymptotically standard normal. Therefore, our robust permutation test in (2.8) with \( T_n \) replaced by \( S_n \) has asymptotic size equal to the nominal level \( \alpha \), even if \( P_1 \neq P_2 \). In case \( P_1 = P_2 \), this test has exact size in finite samples. Since Theorems 2.1 and 2.2 are true regardless of whether the null hypothesis holds or not, we can now study the power properties of the permutation test.

**Corollary 2.1.** Let \( \phi(W_n, Z_n) \) be the permutation test in (2.8) with \( T_n \) replaced by \( S_n \), and suppose Assumptions 2.1–2.3 hold. If the null hypothesis holds, then \( \mathbb{E}(\phi(W_n, Z_n)) \rightarrow \alpha \); otherwise, \( \mathbb{E}(\phi(W_n, Z_n)) \rightarrow 1 \). Moreover, assume \( S_n \) has a limiting distribution under a sequence of local alternatives contiguous to the null. Then, the asymptotic power of the robust permutation test against local alternatives is the same as that of the test that uses critical values from the limiting null distribution of \( S_n \).

---

### 3. Applications

In this section, we apply our theory to four different non-parametric problems: testing for equality of conditional mean and quantile functions evaluated at a point, and testing for continuity of conditional mean and PDF at a point. For a simple and intuitive presentation, we use the Nadaraya–Watson (NW) type of kernel estimators throughout, but proofs in this section generalize to other types of estimators, for example, local-polynomial regression (LPR), sieves, etc. In particular, we demonstrate the generalization of our third application in Section 3.3 to the LPR estimator in Section D.5 of the supplementary materials. We obtain the usual rate restrictions on \( h \), the bandwidth tuning parameter, and intuitive presentation, we use the Nadaraya–Watson (NW) type of kernel estimators throughout, but proofs in this case. A conventional solution is to subtract a first-order bias term \( h^2 \beta_k \) from \( B_k \), where \( B_k \) is nonparametrically estimated by \( B_k = B_{0,n}(Z_{k,1}, \ldots , Z_{k,n}) \). We give the analytical formulas for \( B(P_k) \) and \( B_k \) in Section D.1 (Equation D.6) of the supplementary materials. Our permutation tests use the bias-corrected NW estimator \( \hat{\theta}_k = \theta_{0,n}(Z_{k,1}, \ldots , Z_{k,n}) - h^2 B_{0,n}(Z_{k,1}, \ldots , Z_{k,n}) = \hat{\theta}_k - h^2 B_k \). Note that no bias correction is needed if \( h = o(n^{-1/3}) \) because \( \sqrt{n} h^2 B_k \). An alternative solution to the bias issue consists of replacing the NW estimator \( \hat{\theta}_k \), which is the LPR estimator of order zero, with the LPR estimator of order two. For LPR estimation at an interior point \( x \), if \( \hat{\theta}_k \) is LPR of order \( \rho \), the asymptotic bias of \( \sqrt{n} h^2 B_k - \theta(P_k) \) is \( O \left( \sqrt{n} h^2 \rho + 2 \right) \) if \( \rho \) is odd, or \( O \left( \sqrt{n} h^2 \rho^{1} \right) \) if \( \rho \) is odd (Theorem 3.1 by Fan and Gijbels 1996). Thus, if an LPR of order \( \rho \) has asymptotic bias, that bias vanishes if we increase the order of the polynomial by two if \( \rho \) is even, or by one if \( \rho \) is odd. We keep the NW estimator for a simple and intuitive presentation of our theory in this section, but our permutation tests also apply to LPR estimators. We demonstrate this in the context of our third application (Section 3.3) in the supplementary materials (Section D.5). In the same spirit, our simulations in Section 5 and empirical examples in Section 6 use LPR of order one with MSE-optimal bandwidth and bias-correct it using LPR of order two. Other options of bias correction include higher-order kernels (Li and Racine 2007, sect. 1.11) and the bootstrap (Racine 2001).

When the distribution of \( (X_i, Y_i) \) equals that of \( (X_2, Y_2) \), the permutation test in (2.8) has exact size in finite samples. For other cases, we rely on asymptotic size control, which depends on Assumptions 2.1 and 2.2. Below, we describe regularity

---

2 Alternative resampling methods such as the bootstrap and subsampling also share the same asymptotic local power as the permutation test because they produce critical values that are consistent for standard Gaussian critical values under the null hypothesis.

3 Suppose one uses an estimator \( \hat{\theta}_k^2 \) that assumes the null hypothesis is true; that is, \( \hat{\theta}_k^2 \) is consistent for \( \rho^2 \) under the null hypothesis but has a different probability limit under the alternative. Regardless of whether the null is true, such an estimator applied to a random permutation of the data is generally consistent for \( \rho^2 \), and Corollary 2.1 remains true. Consistency for \( \rho^2 \) comes from the fact that an estimator applied to a random permutation behaves as if it were applied to data from a mixture distribution, where the null is always true (Section B.3.2 in the supplementary materials).
conditions such as continuous differentiability of conditional moments, and Proposition 3.1 proves the asymptotic linear representation of the bias-corrected NW estimator.

**Assumption 3.1.** As \( n \to \infty, n_1/n \to \lambda \in (0, 1), h \to 0, nh \to \infty, \) and \( \sqrt{nh}/c \to c \in [0, \infty). \)

**Assumption 3.2.** For \( k = 1, 2, \) the distribution of \( X_k \) has PDF \( f_{X_k}(x_k) \) that is bounded, that is, there exist \( m \) and \( n \) that are twice differentiable with bounded derivatives.

**Assumption 3.3.** For \( k = 1, 2, m_{Y_k|X_k}(x_k) \equiv \mathbb{E}[Y_k|X_k = x_k] \) is bounded, that is, \( \beta \neq 0 \) and its derivatives are bounded; there exists \( \zeta > 0 \) such that \( \mathbb{E}[|Y_k|^{2+\zeta}|X_k] \) is almost surely bounded.

**Assumption 3.4.** For \( k = 1, 2, \) \( \psi_{Y_k|X_k}(x_k) = \mathbb{P}[Y_k|X_k = x_k] \) is bounded, that is, \( \psi_{Y_k|X_k}(x_k) > 0 \).

**Assumption 3.5.** Let \( V_1, \ldots, V_m \) be an iid sample from a distribution \( P \in \mathcal{P} \), where \( m \) grows with \( n \). Let \( B = B_{m,n}(V_1, \ldots, V_m) \) be a consistent estimator for the first-order bias term \( B(P) \). Assume that \( \sqrt{B - B(P)} \to 0 \) uniformly over \( P \in \mathcal{P} \) for any sequence \( m \) such that \( m/n \to \lambda \) or \( m \to 1 - \lambda \).

**Proposition 3.1.** Suppose Assumptions 3.1–3.5 hold. Let \( V_1 = (R_1, S_1), \ldots, V_m = (R_m, S_m) \) be an iid sample from a distribution \( P \in \mathcal{P} \), where \( m \) grows with \( n \), and consider the bias-corrected estimator \( \theta = \theta_{m,n}(V_1, \ldots, V_m) \) as described in the text. Then, Propositions 2.1 and 2.2 hold for \( \theta \) with \( \psi_n(V, P) = K(x_n) \left( \frac{\psi_{m,n}(R, P)}{\sqrt{m,n}(x_n)} \right) \) and \( \mathbb{E}[\psi^2(V, P)] = \psi_{m,n}(R, P) \), where \( m_{S|R}(x; P) \) is the conditional mean of \( S \) given \( R = x \), \( \psi_{S|R}(x; P) \) is the conditional variance of \( S \) given \( R = x \), and \( \bar{f}_R(x; P) \) is the PDF of \( R \) at \( x \), for all assuming \( V = (R, S) \sim P \in \mathcal{P} \).

The proof of Proposition 3.1 adapts conventional arguments for nonparametric asymptotics (e.g., Theorem 2.2 by Li and Racine 2007) and is found in Section D.1 of the supplementary materials. Under the null hypothesis, the asymptotic variance of \( T_n \) and of the permutation distribution are, respectively, \( \sigma^2 = \kappa_0 \left( \frac{\psi_{Y|X}(x)}{f_X(x)} + \frac{\psi_{Y|X}(x)}{(1-\lambda)f_X(x)} \right) \) and \( \tau^2 = \kappa_0 \left( \frac{f_X(x)f_{Y|X}(x)}{(1-\lambda)f_X(x)} + \frac{f_X(x)f_{Y|X}(x)}{f_Y(x)} \right) \), where \( f_X(x) \) is the square of the PDF of \( R \) evaluated at \( x \) under \( P = \lambda P_1 + (1-\lambda)P_2 \) \( P \in \mathcal{P} \). These variances are generally different, except in special cases, for example, when \( f_X(x) = f_X(x) \) and \( \lambda = 1/2 \), then \( f_Y(x) \) and \( f_{Y|x}(x) \). Thus, in general, the researcher must use the studentized test statistic for the permutation test to have asymptotic size control.

### 3.2. Controlled Quantiles

In this section, we examine equality of conditional quantile functions for two populations. For example, a researcher may wish to compare not only averages (Section 3.1) but also other features of a conditional distribution between \( P_1 \) and \( P_2 \), such as the median, tails, interquantile range, etc. The goal is to test the difference of the \( \chi \)th quantile of the outcomes \( Y \) between the two populations, after controlling for a given value of the variable \( X \). For instance, the immune response \( Y \) of a certain treatment conditional on age \( X \) may differ for individuals at the bottom, median, or top of the immunity distribution.

As in Section 3.1, there are two independent samples, \( Z_{i,j} = (X_{i,j}, Y_{i,j}), i = 1, \ldots, n_1, \) and \( Z_{k,l} = (X_{k,l}, Y_{k,l}), i = 1, \ldots, n_2, \) and the vector \( W_{n} \) is nonrandom. For a given interior point \( x \), the parameter of interest is the \( \chi \)th conditional quantile, that is, \( \theta(X_k) = \mathbb{Q}_x(Y_k|x) = x = \arg\min_x \mathbb{E}[\rho_x(Y_k - a)|X_k = x] \), where \( \rho_x(Y_k - a|X_k = x) \).

For a bandwidth \( h > 0 \) and a kernel density function \( K \), a consistent estimator of the NW style is \( \hat{\theta}^b = \hat{\theta}^b_{m,n}(Z_{1,1}, \ldots, Z_{k,m}) = \arg\min_x \sum_{i,j=1}^{n_1} \rho_x(Y_{i,j} - a)K_h(x - x_i) \), for \( k = 1, 2 \).

The superscript \( b \) indicates that there is bias in the asymptotic distribution of \( \sqrt{m,n}(\hat{\theta}^b_{m,n} - \theta(P_k)) \) whenever the bandwidth choice converges to zero at the slowest possible rate, that is, \( h = O(n^{-1/5}) \) (Proposition 3.2). This is the case of mean squared error (MSE) optimal bandwidths, and inference requires bias correction in this case. A conventional solution is to subtract a first-order bias term \( h^2B(P_k) \) from \( \hat{\theta}^b \), where \( B(P_k) \) is nonparametrically estimated by \( \hat{B}_{m,n}(Z_{k,1}, \ldots, Z_{k,m}) \). We give the analytical formulas for \( B(P_k) \) and \( \hat{B}_{m,n} \). We give the analytical formulas for \( B(P_k) \) and \( \hat{B}_{m,n}(Z_{k,1}, \ldots, Z_{k,m}) \). Note that no bias correction is needed if \( h = o(n^{-1/5}) \) because \( \sqrt{m,n}(\hat{\theta}^b - \theta(P_k)) \). An alternative solution to the bias issue consists of replacing \( \hat{\theta}^b \) with the LPR estimator, while keeping the MSE-optimal bandwidth choice for the NW estimator. See Sections 3.1 and 3.3 for details.

Proposition 3.2 demonstrates validity of Assumptions 2.1–2.2 for the NW quantile regression estimator. It relies on some of the same assumptions as Section 3.1 plus a couple of assumptions on the distribution of \( Y_k \) conditional on \( X_k \).

**Assumption 3.3'.** For \( k = 1, 2, \) the distribution of \( Y_k \) conditional on \( X_k \) has PDF \( f_{Y|X}(y_k|x_k) \) that is bounded and differentiable function of \( (x_k, y_k) \), has bounded partial derivatives with respect to \( x_k \) and \( y_k \), and is bounded away from zero over \( y_k \) for \( x_k = x \).

**Assumption 3.4'.** For \( k = 1, 2, \) the distribution of \( Y_k \) conditional on \( X_k \) has CDF \( F_{Y|X}(y_k|x_k) \) that is three times partially differentiable with respect to \( x_k \) with bounded partial derivatives.

**Proposition 3.2.** Suppose Assumptions 3.1, 3.2, 3.3', 3.4', and 3.5 hold. Let \( V_k = (R_1, S_1), \ldots, V_m = (R_m, S_m) \) be an iid sample from a distribution \( P \in \mathcal{P} \), where \( m \) grows with \( n \), and consider the bias-corrected estimator \( \hat{\theta} \equiv \theta_{m,n}(V_1, \ldots, V_m) \) as described in the text. Then, \( \hat{\theta} \) satisfies Assumptions 2.1 and 2.2 with \( \psi_n(V, P) = -K(x_n) \left( \frac{|S-S_{\theta}|}{f_{\theta}(\theta(P))} \right) \left( \frac{f_{R|\theta}(P)|\theta(P)|}{f_{R|\theta}(P)|P_\theta(P)} \right) \) and \( \mathbb{E}[\psi^2(V, P)] = \mathbb{E}[\psi_{m,n}(R, P)] \).
is the conditional CDF of $S$ given $R = r$, all three assuming $V = (R, S) \sim P \in \mathcal{P}$.

The proof of Proposition 3.2 adapts arguments by Pollard (1991), Chaudhuri (1991), and Fan, Hu, and Truong (1994) and is found in Section D.2 of the supplementary materials. The asymptotic variance of $\tilde{T}_n$ and of the permutation distribution are, respectively, $\sigma^2 = \frac{k_0 \omega_{1}X(1 - \chi)}{(1 - \lambda) f_{x_0}^{(1)}(\theta(P_1)x) f_{x_1}^{(1)}(x)} + \frac{k_0 \omega_{2}X(1 - \chi)}{(1 - \lambda) f_{x_0}^{(1)}(\theta(P_2)x) f_{x_1}^{(1)}(x)}$ and $\tau^2 = \frac{k_0 \omega_{2}X(1 - \chi)}{(1 - \lambda) f_{x_0}^{(1)}(\theta(P_1)x) f_{x_1}^{(1)}(x)}$. These variances are generally different except in special cases. For example, the null hypothesis implies $\theta(P_1) = \theta(P_2) = \theta(\hat{P})$. If $f_{x_0}^{(1)}(\theta(P_1)x) = f_{x_0}^{(1)}(\theta(P_2)x)$ and $f_{x_1}^{(1)}(x) = f_{x_1}^{(1)}(x)$, then $\sigma^2 = \tau^2$. However, in general, the researcher must use the studentized test statistic for the permutation test to have asymptotic size control.

### 3.3. Discontinuity of Conditional Mean

Numerous empirical studies in the social sciences have relied on estimation and inference on the size of a discontinuity in a conditional mean function at a certain point. In the so-called regression discontinuity design (RDD), an individual $i$ receives treatment if and only if a running variable $X_i$ is above a fixed threshold. If individuals do not know the threshold or do not have perfect manipulation over $X$, untreated individuals who barely missed the cutoff serve as a control group for treated individuals who barely made it across the cutoff. Assume the threshold for treatment is 0 without loss of generality. The difference in side limits $E[Y|X = 0^+] - E[Y|X = 0^-]$ identifies the causal effect of treatment on an outcome variable $Y$. Thus, the null hypothesis of zero causal effect is equivalent to continuity of the conditional mean function $E[Y|X = x]$ at $x = 0$.

This idea first appeared in psychology (Thistlethwaite and Campbell 1960), made its way to economics (Hahn, Todd, and Van der Klaauw 2001), and has a growing number of applications in the social sciences. Examples include Agarwal et al. (2017) in economics, Valentine, Konstanopoulos, and Goldrick-Rab (2017) in education, Abou-Chadi and Krause (2020) in political science, and Zoorob (2020) in sociology. We focus on the conditional mean function in this section, but the whole argument goes through if one desires to use the conditional quantile function.

The researcher has a sample of $n$ iid observations $(X_i, Y_i)$ and the NW estimator for the discontinuity is
\[
\frac{\sum_{i=1}^{n} I\left(\frac{X_i}{\Delta} \right) Y_i}{\sum_{i=1}^{n} I\left(\frac{X_i}{\Delta} \right)} - \frac{\sum_{i=1}^{n} I\left(\frac{X_i + \Delta}{\Delta} \right) Y_i}{\sum_{i=1}^{n} I\left(\frac{X_i + \Delta}{\Delta} \right)}.
\]

RDD fits in our two-population framework by splitting the sample based on $X$ being above or below the cutoff. Construct $W_i = 2 - I[X_i \geq 0]$ so that $n_1/n \rightarrow \lambda = P(W_i = 1)$. Record the sample such that the $n_1$ observations with $W_i = 1$ come first, and the $n_2$ observations with $W_i = 2$ come second. Define $Z_{1,i} = (X_{1,i}, Y_{1,i}) = (X_i, Y_i)$ for $i = 1, \ldots, n_1$ and $Z_{2,i} = (X_{2,i}, Y_{2,i}) = (-X_{n_1+1+i}, Y_{n_1+1+i})$ for $i = 1, \ldots, n_2$. We have $Z_{n} = (Z_{1,1}, \ldots, Z_{1,n_1}, Z_{2,1}, \ldots, Z_{2,n_2})$. Conditional on $W_i$, the distribution of $Z_{1,i}$ is $P$, which equals the distribution of $(X, Y)$ conditional on $X \geq 0$. Likewise for $Z_{2,i}, P$ is the distribution of $(-X, Y)$ conditional on $X < 0$. Permutations reorder observations in $Z_n$ but keep $W_i$ unchanged. The RDD parameter becomes $\theta(P_1) - \theta(P_2)$, where $\theta(P_i) = E[Y_k|X_k = 0^+]$ for $k = 1, 2$.

The NW estimator for the RDD parameter is $\hat{\theta}_1 - \hat{\theta}_2$, where $\hat{\theta}_k$ is defined in Equation (3.1) for $k = 1, 2$ and the evaluation point $x$ is set to zero. The superscript $b$ indicates that there is bias in the asymptotic distribution of $\sqrt{n h_b} (\hat{\theta}_k - \theta(P_k))$ whenever the bandwidth choice converges to zero at the slowest possible rate, that is, $h \rightarrow O(n^{-1/3})$ (Proposition 3.3). This is the case of MSE-optimal bandwidths, and inference requires bias correction in this case. A conventional solution is to subtract a first-order bias term $hB(P_k)$ from $\hat{\theta}_k$, where $B(P_k)$ is nonparametrically estimated by $\hat{B}_k = B_{n,n}(Z_{k,1}, \ldots, Z_{k,n_k})$. We give the analytical formulas for $B(P_k)$ and $\hat{B}_k$ in Section D.3 (Equation D.21) of the supplementary materials. Our permutation tests use the bias-corrected NW estimator $\hat{\theta}_k = \hat{\theta}_k + hB_{n,n}(Z_{k,1}, \ldots, Z_{k,n_k}) - hB_{n,n}(Z_{k,1}, \ldots, Z_{k,n_k}) = \hat{\theta}_k - h\hat{B}_k$.

Note that no bias correction is needed if $h = o(n^{-1/3})$ because $\sqrt{n h_b} (\hat{\theta}_k - \theta(P_k)) - hB(P_k) = \sqrt{n h_b} (\hat{\theta}_k - \theta(P_k)) + o(1)$. An alternative solution to the bias issue consists of replacing the NW estimator $\hat{\theta}_k$, which is the LPR estimator of order zero, with the LPR estimator of order one, while keeping the MSE-optimal bandwidth choice for the NW estimator. For LPR estimation at a boundary point, if $\hat{\theta}_k$ is LPR of order $\rho$, the asymptotic bias of $\sqrt{n h_b} (\hat{\theta}_k - \theta(P_k))$ is $O(\sqrt{n h^\rho})$ (Theorem 3.2 by Fan and Gijbels 1996). Thus, if an LPR of order $\rho$ has asymptotic bias, that bias vanishes if we increase the order of the polynomial by one. We keep the NW estimator for a simple and intuitive presentation of our theory in this section, but we demonstrate that our permutation tests also apply to LPR estimators of any order in Section D.5 of the supplementary materials. Section 6 illustrates our procedures with two empirical examples of RDD using LPR estimators and MSE-optimal bandwidths, which is standard practice in applied literature. A third option for bias correction is the method proposed by Armstrong and Kolesár (2018).

In case the distribution of $(X, Y)$ equals that of $(-X, Y)$, then $P_1 = P_2$ and the permutation test in (2.8) has exact size in finite samples. Note that this X-symmetry restriction eliminates the impossibility problem in RDD tests (Kamat 2018; Bertanha and Moreira 2020) because there is no bias in estimation. For other cases, we rely on asymptotic size control, which depends on Assumptions 2.1 and 2.2. Proposition 3.3 builds on assumptions similar to those of Section 3.1, which we rewrite below in terms of the originally sampled variables $(X, Y)$, that is, before the variables are transformed into $(X_k, Y_k), k = 1, 2$.

**Assumption 3.1**. As $n \rightarrow \infty, h \rightarrow 0, n h \rightarrow \infty$, and $\sqrt{n h} \rightarrow c \in [0, \infty)$.

**Assumption 3.2**. The distribution of $X$ has PDF $f_X(x)$ that is bounded, bounded away from zero, twice differentiable except at $x = 0$, and has bounded derivatives.

**Assumption 3.3**. $E[Y|X = x]$ is bounded, twice differentiable except at $x = 0$, and has bounded derivatives. There exists $\zeta > 0$ such that $\mathbb{E}[|Y|^{\zeta}] < \infty$.
Assumption 3.4*. \( 
abla [Y | X = x] \) is bounded, differentiable except at \( x = 0 \), has bounded derivative, \( \nabla [Y | X = 0^+] > 0 \), and \( \nabla [Y | X = 0^-] > 0 \), where \( 0^+ \) and \( 0^- \) denote side limits as \( x \to 0 \).

**Proposition 3.3.** Suppose Assumptions 3.1*, 3.2*, 3.3*, 3.4*, and 3.5 hold. Let \( V_1 = (R_1, S_1), \ldots, V_m = (R_m, S_m) \) be an iid sample from a distribution \( P \in \mathcal{P} \), where \( m \) grows with \( n \), and consider the bias-corrected estimator \( \hat{\theta} = \hat{\theta}_{m,n}(V_1, \ldots, V_m) \) as described in the text. Then, \( \hat{\theta} \) satisfies Assumptions 2.1 and 2.2 with \( \psi(V, P) = K \left( \frac{b}{n} \right) \left( \frac{S - m \rho_R(R,P)}{\sqrt{m \rho_R(R,P)/2}} \right) \) and \( \hat{\xi}^2(P) = \frac{\rho_{SR}(0^+)}{\rho_{SR}(0^-)}, \) where \( m \rho_R(r; P) \) is the conditional mean of \( S \) given \( R = r, \rho_{SR}(r; P) \) is the conditional variance of \( S \) given \( R = r \), and \( f_R(r; P) \) is the PDF of \( R \) at \( r \), all three assuming \( V = (R, S) \sim P \in \mathcal{P} \).

The proof is in Section D.3 of the supplementary materials. The conditions and proof of Proposition 3.3 follow along the lines of the conditional mean case (Proposition 3.1). Unlike Section 3.1, the evaluation point \( x \) lies at the boundary of the support of \( X_k \). As a result, \( h \) must be chosen to zero to bound the asymptotic bias of \( T_n \) (i.e., \( \sqrt{n} h \to c \) in Assumption 3.1* versus \( \sqrt{n} h \to c \) in Assumption 3.1). Proposition 3.3 is extended to LPR estimators of any order \( p \) in Section D.5 of the supplementary materials. Common practice in RDD uses local-linear regression (\( p = 1 \)) with a MSE-optimal bandwidth, and the resulting asymptotic bias is often corrected by using a local-quadratic regression (\( p = 2 \)).

If agents do not manipulate \( X \) to change their treatment status, which is a key assumption in RDD, then the PDF of \( X \) should be continuous at the cutoff. This implies that \( f_X(0^+) = f_X^*(0^+) = f_X^*(0^-) = f_X(0^-) = f_X(0) \). In this case, under the null hypothesis, the asymptotic variance of \( T_n \) and of the permutation distribution are, respectively, \( \sigma^2 = \frac{\kappa_0^2}{f_X(0)^2} \left( \sum_{i=1}^n f_Y | X_i(0^+) + v_i \right) \) and \( \tau^2 = \frac{\kappa_0^2}{f_X(0)^2} \left( \sum_{i=1}^n f_Y | X_i(0^-) + v_i \right) \). These are generally different, except when \( \lambda = 1/2 \). Thus, in general, the researcher must use the studentized test statistic for the permutation test to have asymptotic size control.

### 3.4. Discontinuity of Density

In many settings, the distribution of a random variable may exhibit a discontinuity at a given point if a certain phenomenon of interest occurs. For example, estimating agents' responses to incentives is a central objective in the social sciences. A continuous distribution of agents who face a discontinuous schedule of incentives results in a distribution of responses with a discontinuity at a known point. For example, Saez (2010) looks for evidence of a mass point in the distribution of reported income at tax brackets as evidence of agents' responses to tax rates. Caetano (2015) proposes an exogeneity test in nonparametric regression models, where the distribution of the potentially endogenous regressor may have a mass point. Identification of causal effects with RDD depends heavily on continuity assumptions, and these imply that the PDF of the control variable is continuous at the cutoff.

Consider a scalar random variable \( X \) with PDF \( f \) that is continuous, except for point \( x = 0 \). We want to test the null hypothesis of continuity of the PDF at \( x = 0 \). For a sample with \( n \) iid observations \( X_n \), a kernel density estimator for the size of the discontinuity is

\[
\frac{2}{nh} \sum_{i=1}^n K \left( \frac{X_i}{n} \right) \left[ I[X_i \geq 0] - I[X_i < 0] \right].
\]

The problem fits in our two-population framework by randomly splitting the sample as follows. Make \( n_1 = \lfloor n/2 \rfloor \) and \( n_2 = n - n_1 \). For observations \( 1 \leq i \leq n_1 \), set \( W_i = 1 \) and let \( Z_{i,1} = X_{i,1} = X_i \) for observations \( n_1 < i \leq n \), set \( W_i = 2 \) and let \( Z_{i,2} = X_{i,2} = X_i \). This implies that \( n_1/n \to 1/2 \). Permutations reorder observations in \( Z_n = (X_{1,1}, \ldots, X_{1,n_1}, X_{1,2}, \ldots, X_{n,n_2}) \) but keep \( W_n \) unchanged. Conditional on \( W_n \), the distribution of \( X_{i,1} \) is \( P_1 \), which equals the distribution of \( X \). Likewise for \( X_{i,2} \).

Let \( V \) be a scalar variable \( R \sim P \in \mathcal{P} \). The parameter of interest is \( \theta(P) = 1/2(f_0^*(0^+; P) - f_0^*(0^-; P)) \), where \( f_0^*(0^+; P) \) and \( f_0^*(0^-; P) \) are the side-limits at zero of the PDF of \( R \) under the \( P \) distribution. The discontinuity parameter is \( \theta(P_1) - \theta(P_2) \), which equals \( f_0^*(0^+) - f_0^*(0^-) \) in terms of the PDF of \( X \). The kernel estimator for the density discontinuity becomes \( \hat{\theta}^b_1 - \hat{\theta}^b_2 \), where \( \hat{\theta}^b_1 \) is defined as \( \hat{\theta}^b_1 = \frac{1}{nh} \sum_{i=1}^n K \left( \frac{X_i}{n} \right) \left[ I[X_i \geq 0] - I[X_i < 0] \right] \), for \( k = 1, 2 \).

The b superscript denotes asymptotic bias in the limiting distribution of \( \sqrt{n} h (\hat{\theta}^b_1 - \theta(P_1)) \) whenever the bandwidth choice converges to zero at the slowest possible rate, that is, \( h = O(n^{-1/3}) \) (Proposition 3.4). This is the case of MSE-optimal bandwidths, and inference requires bias correction in this case. A conventional solution is to subtract a first-order bias term \( h B(P_k) \) from \( \hat{\theta}^b_k \), where \( B(P) \) is nonparametrically estimated by \( \hat{B}_k = B_{n,l,n}(Z_{k,1}, \ldots, Z_{k,n}) \). We give the analytical formulas for \( B(P) \) and \( \hat{B}_k \) in Section D.4 (Equation D.24) of the supplementary materials. Our permutation tests utilize the bias-corrected NW estimator \( \hat{\theta}_k = \hat{\theta}_{n,l,n}(Z_{k,1}, \ldots, Z_{k,n}) \approx \theta_{n,l,n}(Z_{k,1}, \ldots, Z_{k,n}) - h B_{n,l,n}(Z_{k,1}, \ldots, Z_{k,n}) = \hat{\theta}_k - h \hat{B}_k \).

Note that no bias correction is needed if \( h = o(n^{-1/3}) \) because \( \sqrt{n} h (\hat{\theta}^b_1 - \theta(P_1)) = O(1) \). Alternative solutions to the bias issue include using transformed-kernel estimators (Marron and Ruppert 1994) or LPR density estimators (Cattaneo, Jansson, and Ma 2020).

It is important to emphasize that if the sample size \( n \) is even and split in half, then the test statistic \( T_n = \sqrt{nh} (\hat{\theta}_1 - \hat{\theta}_2) \) is invariant to the way the original sample is split. In case the distribution of \( X \) is symmetric at 0, then \( P_1 = P_2 \) and the permutation test in (2.8) has exact size in finite samples. For other cases, we rely on asymptotic size control, and thus need to verify Assumptions 2.1 and 2.2.

**Proposition 3.4.** Suppose Assumptions 3.1*, 3.2*, and 3.5 hold. Let \( V_1 = (R_1, S_1), \ldots, V_m = (R_m, S_m) \) be an iid sample from a distribution \( P \in \mathcal{P} \), where \( m \) grows with \( n \), and consider the bias-corrected estimator \( \hat{\theta} = \hat{\theta}_{m,n}(V_1, \ldots, V_m) \).
as described in the text. Then, Assumptions 2.1 and 2.2 hold for \( \theta \) with \( \psi_n(V, P) = h_n^{1/2} \left( \frac{K}{\sqrt{n}} \right) \left( I[R \geq 0] - I[R < 0] \right) - h_n^{1/2} E_P[K \left( \frac{K}{\sqrt{n}} \right) \left( I[R \geq 0] - I[R < 0] \right)] \) and \( \xi^2(P) = \kappa_{0,2}^{-1} \left( f_R(0^+; P) + f_R(0^-; P) \right) \).

The proof of Proposition 3.4 is in Section D.4 in the supplementary materials. The asymptotic variance of \( T_n \) and of the permutation distribution are, respectively,

\[
\sigma^2 = 2 \kappa_{0,2}^{-1} \left[ f_X(0^+) + f_X(0^-) + f_X(0^+) + f_X(0^-) \right] = 4 \kappa_{0,2}^{-1} \left[ f_R(0^+; P) + f_R(0^-; P) \right] = \tau^2.
\]

These are the same regardless if the null hypothesis is true or not. Thus, unlike the previous examples, we do not need to studentize the test statistic for asymptotic validity of the permutation test.

4. Confidence Sets

This section constructs robust confidence sets for a discrepancy measure \( \Psi(P_1, P_2) \) between the two populations by “inverting” the permutation test for the null hypothesis \( \Psi(P_1, P_2) = \delta, \delta \in \mathbb{R} \). The discrepancy measure satisfies two requirements. First, there exists a unique \( \delta_0 \in \mathbb{R} \) such that \( \Psi(P_1, P_2) = \delta_0 \) is equivalent to \( \theta(P_1) = \theta(P_2) \). Second, for every \( \delta \neq \delta_0 \), there exists a known data transformation \( \psi_{\delta} \) that applies to observations from the first sample such that the distribution \( \hat{P}_1 \) of \( \psi_{\delta}(Z_{1,i}) \) satisfies \( \theta(\hat{P}_1) = 0 \).

In terms of the examples of Sections 3.1–3.3, we set \( \Psi(P_1, P_2) = \theta(P_1) - \theta(P_2), \) and it follows that \( \delta_0 = 0 \) and \( \psi_{\delta}(Z_{1,i}) = (X_{1,i}, Y_{1,i} - \delta) \) satisfy the two requirements for the discrepancy measure. Note that the null hypothesis \( \psi_{\delta}(P_1, P_2) = \delta \) is equivalent to \( \mathbb{E}[Y_1|X_1 = x] = \mathbb{E}[Y_2|X_2 = x] = \delta \) in the conditional mean case, to \( \mathbb{Q}_x(Y_1|X_1 = x) = \mathbb{Q}_x(Y_2|X_2 = x) = \delta \) in the conditional quantile case, and to \( \mathbb{E}[Y|X = 0^+] = \mathbb{E}[Y|X = 0^-] = \delta \) in the discontinuity of conditional mean case. For the discontinuity of PDF example of Section 3.4, we make \( \Psi(P_1, P_2) = f_X(0^+)/f_X(0^-) \), and we have that \( \delta_0 = 1 \) and \( \psi_{\delta}(Z_{1,i}) = X_{1,i} \{ \|X_{1,i} \geq 0 \} + (1/\delta) \{ \|X_{1,i} < 0 \} \) satisfy the two requirements. In this example, \( \Psi(P_1, P_2) = \delta \) is equivalent to \( f_X(0^+)/f_X(0^-) = \delta \).

Define \( \phi_{\delta}(W_n, Z_n) \) to be the test described in Equation (2.8) with the studentized test statistic \( S_n \) of Equation (2.10) replacing \( T_n \). This test applies to the null hypothesis \( \Psi(P_1, P_2) = \delta_0 \). Next, for \( \delta \neq \delta_0 \), we first transform the data \( \tilde{Z}_n \) to \( \tilde{Z}_n = (Z_{1,1}, \ldots, Z_{1,n_1}, Z_{2,1}, \ldots, Z_{2,n_2}) \) to \( \tilde{Z}_n \). The robust permutation test for the null hypothesis \( \Psi(P_1, P_2) = \delta \) is defined as \( \phi_{\delta}(W_n, Z_n) \) is \( \phi_{\delta}(W_n, Z_n) \).

Let \( U \) be a uniform random variable in \([0, 1]\) and independent of the data. The confidence set with \( 1 - \alpha \) nominal coverage is \( C(W_n, Z_n) = \{ \delta : U > \phi_{\delta}(W_n, Z_n) \} \). The set almost-surely includes all values of \( \delta \) for which the test fails to reject, and it excludes the ones the test rejects. For those values of \( \delta \) for which the test outcome is randomized with rejection probability \( a \), the inclusion in the confidence set occurs with probability \( 1 - a \). The purpose of a randomized confidence set is to guarantee exact coverage whenever the test \( \phi_{\delta} \) has exact size. A nonrandomized confidence set is \( \hat{C}(W_n, Z_n) = \{ \delta \in \mathbb{R} : \phi_{\delta}(W_n, Z_n) < 1 \} \), and its coverage is conservative, especially in small samples.

Lemma 2.1 implies that \( \phi_{\delta}(W_n, Z_n) \) has exact size \( \alpha \) in finite samples for any \( P_1 \) and \( P_2 \) such that \( \Psi(P_1, P_2) = \delta \) and \( \hat{P}_1 = \psi_{\delta}P_1 = P_2 \). This implies that the confidence set \( C(W_n, Z_n) \) has exact coverage in finite samples if distributions are equal up to a transformation \( \psi_{\delta} \). In the examples of Sections 3.1–3.2, exactness occurs when the distributions of \( P_1 \) and \( P_2 \) are such that there exists \( \delta \in \mathbb{R} \) for which the distribution of \( (Y_{1,i} - \delta, X_{1,i}) \) equals that of \( (Y_{2,i} - \delta, X_{2,i}) \) in Section 3.3, when there exists \( \delta \in \mathbb{R} \) for which the distribution of \( (Y - \delta, X) \) equals that of \( (Y_i, X_i) \) \( X_i \) is zero and in Section 3.4, when there exists \( \delta \in (0, \infty) \) such that \( \mathbb{E}[x|X = 0] = \delta \) is symmetric around \( x = 0 \). If these restrictions do not apply, then the confidence set has correct coverage asymptotically.

Corollary 4.1. Consider the discrepancy measure \( \Psi \) and class of data transformations \( \psi_{\delta} \) discussed above. For any \( n, q_n, P_1, \) and \( P_2, \) if \( \hat{P}_1 = \psi_{\delta}P_1 = P_2 \) for \( \delta = \Psi(P_1, P_2) \), then \( P_1 | \Psi(P_1, P_2) \in C(W_n, Z_n) \} = 1 - \alpha \). Assume instead that \( \hat{P}_1 = \psi_{\delta}P_1 \neq P_2 \) and Assumptions 2.1–2.3 hold. Then, as \( n \to \infty \), \( \mathbb{P}[\Psi(P_1, P_2) \in C(W_n, Z_n)] \to 1 - \alpha \).

5. Monte Carlo Simulations

We conduct Monte Carlo simulations to compare the finite sample performance of our permutation test to the conventional \( t \)-test, that is, the test that rejects the null if \( |S_n| > 1 - \Phi(1-\alpha/2) \). The goal is not to show that the permutation test dominates the \( t \)-test in all cases; instead, the goal is to verify the theoretical predictions of Section 2 and explore DGP variations that illustrate pros and cons of our methods. The exercise confirms the theoretical findings of size control in large samples and in finite samples under the sharp null; it also shows similar power curves between permutation and \( t \)-tests in large samples. Moreover, we find several cases where the permutation test performs significantly better than the \( t \)-test, both in power and in size control outside of the sharp null. We also compare our permutation test to two other popular resampling procedures, namely, the bootstrap and subsample, and the permutation test compares favorably to them.

The basic setup of our DGP is as follows. For \( k = 1, 2, X_k \sim U[0,1], \varepsilon_k \sim N(0, \sigma^2_k) \), where \( X_k \) is independent of \( \varepsilon_k, Y_k = m_{\psi_k}(X_k) + \varepsilon_k \), and the variances are \( \sigma^2_k, \sigma^2_k \in \{ (1, 1), (1, 5) \} \). The experiments simulate iid samples from the following designs:

**Design 1.** \( m_{\psi_k}(x) = g_k(x), \) where \( g_k(x) = 5(x - 0.2)(x - 0.8)\{\|x - 0.5\| > 0.3\} \) and \( g_k(x) = -15(x - 0.2)(x - 0.8)\{\|x - 0.5\| > 0.3\} \); the sample sizes are \( (n_1, n_2) \in \{ (100, 1900), (250, 4750), (500, 9500), (40, 166), (100, 4900), (200, 9800) \} \).
and the null hypothesis is \( H_0 : \theta(P_1) = m_{y,1}(0.5) = m_{y,2}(0.5) = \theta(P_2) \);

**Design 2.** \( m_{y,\delta}(x) = 1 + g_{\delta}(x) \) for \( g_{\delta}(x) \) of Design 1; \( D_{\delta} = m_{d,\delta}(X_k) + \delta g_{\delta} \) for mutually independent \( (X_k, \delta g_{\delta}) \) and \( \delta g_{\delta} \sim N(0, 1) \); \( m_{y,\delta}(x) = \mu \in (1, 10) \); the sample sizes are \( (n_1, n_2) \in \{(75, 75), (150, 150), (1000, 1000)\} \); and the null hypothesis is \( H_0 : \theta(P_1) = m_{y,1}(0.5)/m_{d,1}(0.5) = m_{y,2}(0.5)/m_{d,2}(0.5) = \theta(P_2) \).

**Design 1** is an example of the controlled means case studied in Section 3.1, and **Design 2** is a variation of that case. **Design 2** has controlled means of two different variables with the ratio of means being of interest. The asymptotic behavior of the ratio estimator can be obtained via the Delta method, and Assumptions 2.1 and 2.2 can be verified using arguments similar to those in Section 3.1.

These designs represent practical situations in which the \( t \)-test is known to perform poorly in small samples. **Design 1** corresponds to cases of sample imbalance, that is, cases where the sample sizes are very different. For example, a researcher has a much larger sample of women than men and is interested in comparing the average outcome from a professional training between men and women, after conditioning on a test score. **Design 2** encompasses scenarios where the ratio of conditional mean functions is of interest, and the denominator may be small. An example is estimating the efficacy rate of a vaccine conditional on blood pressure, and comparing this rate between men and women. The efficacy rate in vaccine trials is the difference in proportions of infected individuals between treatment and control groups, divided by the proportion in the control group.

Both designs have conditional mean functions for \( Y \) given \( X \) that are equal and flat for \( |X - 0.5| \leq 0.3 \), but different and nonlinear otherwise (Figure 4 in Section E of the supplemental materials). This shape of conditional mean function allows us to experiment with scenarios with or without estimation bias, and inside or outside the sharp null, depending on the choice of bandwidth. Both designs fall under the null hypothesis; they fall under the sharp null if we further set \( \sigma_1^2 = \sigma_2^2 \) and restrict the sample to \( |X - 0.5| \leq 0.3 \).

The test statistic \( T_n \) is the difference of consistent estimators \( \hat{\theta}_1 - \hat{\theta}_2 \) multiplied by \( \sqrt{n} h \) (Equation (2.7)). The studentized statistic \( S_n \) equals the difference of consistent estimators divided by the standard error of the difference (Equation (2.10)). The conditional mean functions at point 0.5 are consistently estimated by local linear regressions with triangular kernel and a bandwidth choice \( h \) that shrinks to zero as \( n \) increases. A practical choice for \( h \) is the estimated MSE-optimal bandwidth for local-linear regression (LLR), which decreases at rate \( n^{-1/5} \). In particular, we adapt the algorithm of Imbens and Kalyanaraman (2012), denoted IK bandwidth, to our setting. This choice of bandwidth implies that \( T_n \) and \( S_n \) have asymptotically distributions not centered at zero. Thus, we employ local quadratic regressions, using the same kernel and bandwidth as before, to construct the test statistics and avoid the asymptotic

| \( \sigma_1^2 \) | \( \sigma_2^2 \) | \( \mu \) | \( n_1 = n_2 \) | \( \hat{\eta}_{n,\infty} \) |
|---|---|---|---|---|
| **Design 1** | | | | |
| \( h = 0.1 \) | \( h = 0.3 \) | \( h = 0.5 \) | \( \hat{T}_n \) | \( \hat{S}_n \) |
| \( \hat{\eta}_{n,\infty} \) | \( \eta_{n,\infty} \) | \( \hat{\eta}_{n,\infty} \) | \( \eta_{n,\infty} \) | \( \hat{\eta}_{n,\infty} \) |

**Table 1.** Simulated rejection rates—5% nominal size.

**Table 2.** Simulated rejection rates—5% nominal size.

**Table 3.** Simulated rejection rates—5% nominal size.

**Table 4.** Simulated rejection rates—5% nominal size.

**Table 5.** Simulated rejection rates—5% nominal size.

**Table 6.** Simulated rejection rates—5% nominal size.

**Table 7.** Simulated rejection rates—5% nominal size.

**Table 8.** Simulated rejection rates—5% nominal size.

**Table 9.** Simulated rejection rates—5% nominal size.

**Table 10.** Simulated rejection rates—5% nominal size.

**Table 11.** Simulated rejection rates—5% nominal size.

**Table 12.** Simulated rejection rates—5% nominal size.

**Table 13.** Simulated rejection rates—5% nominal size.

**Table 14.** Simulated rejection rates—5% nominal size.

**Table 15.** Simulated rejection rates—5% nominal size.

**Table 16.** Simulated rejection rates—5% nominal size.

**Table 17.** Simulated rejection rates—5% nominal size.

**Table 18.** Simulated rejection rates—5% nominal size.

**Table 19.** Simulated rejection rates—5% nominal size.

**Table 20.** Simulated rejection rates—5% nominal size.

**Table 21.** Simulated rejection rates—5% nominal size.

**Table 22.** Simulated rejection rates—5% nominal size.

**Table 23.** Simulated rejection rates—5% nominal size.

**Table 24.** Simulated rejection rates—5% nominal size.

**Table 25.** Simulated rejection rates—5% nominal size.

**Table 26.** Simulated rejection rates—5% nominal size.

**Table 27.** Simulated rejection rates—5% nominal size.

**Table 28.** Simulated rejection rates—5% nominal size.

**Table 29.** Simulated rejection rates—5% nominal size.

**Table 30.** Simulated rejection rates—5% nominal size.

**Table 31.** Simulated rejection rates—5% nominal size.

**Table 32.** Simulated rejection rates—5% nominal size.

**Table 33.** Simulated rejection rates—5% nominal size.

**Table 34.** Simulated rejection rates—5% nominal size.

**Table 35.** Simulated rejection rates—5% nominal size.

**Table 36.** Simulated rejection rates—5% nominal size.

**Table 37.** Simulated rejection rates—5% nominal size.

**Table 38.** Simulated rejection rates—5% nominal size.

**Table 39.** Simulated rejection rates—5% nominal size.

**Table 40.** Simulated rejection rates—5% nominal size.

**Table 41.** Simulated rejection rates—5% nominal size.

**Table 42.** Simulated rejection rates—5% nominal size.

**Table 43.** Simulated rejection rates—5% nominal size.

**Table 44.** Simulated rejection rates—5% nominal size.

**Table 45.** Simulated rejection rates—5% nominal size.

**Table 46.** Simulated rejection rates—5% nominal size.

**Table 47.** Simulated rejection rates—5% nominal size.

**Table 48.** Simulated rejection rates—5% nominal size.

**Table 49.** Simulated rejection rates—5% nominal size.

**Table 50.** Simulated rejection rates—5% nominal size.

**Table 51.** Simulated rejection rates—5% nominal size.

**Table 52.** Simulated rejection rates—5% nominal size.

**Table 53.** Simulated rejection rates—5% nominal size.

**Table 54.** Simulated rejection rates—5% nominal size.

**Table 55.** Simulated rejection rates—5% nominal size.

**Table 56.** Simulated rejection rates—5% nominal size.

**Table 57.** Simulated rejection rates—5% nominal size.

**Table 58.** Simulated rejection rates—5% nominal size.

**Table 59.** Simulated rejection rates—5% nominal size.

**Table 60.** Simulated rejection rates—5% nominal size.

**Table 61.** Simulated rejection rates—5% nominal size.

**Table 62.** Simulated rejection rates—5% nominal size.

**Table 63.** Simulated rejection rates—5% nominal size.

**Table 64.** Simulated rejection rates—5% nominal size.

**Table 65.** Simulated rejection rates—5% nominal size.

**Table 66.** Simulated rejection rates—5% nominal size.

**Table 67.** Simulated rejection rates—5% nominal size.

**Table 68.** Simulated rejection rates—5% nominal size.

**Table 69.** Simulated rejection rates—5% nominal size.

**Table 70.** Simulated rejection rates—5% nominal size.

**Table 71.** Simulated rejection rates—5% nominal size.

**Table 72.** Simulated rejection rates—5% nominal size.

**Table 73.** Simulated rejection rates—5% nominal size.

**Table 74.** Simulated rejection rates—5% nominal size.

**Table 75.** Simulated rejection rates—5% nominal size.

**Table 76.** Simulated rejection rates—5% nominal size.

**Table 77.** Simulated rejection rates—5% nominal size.

**Table 78.** Simulated rejection rates—5% nominal size.

**Table 79.** Simulated rejection rates—5% nominal size.

**Table 80.** Simulated rejection rates—5% nominal size.
Figure 1. Design 1—Simulated power curves.

NOTE: Panels (a)–(b) compare power of the studentized permutation test (SP) with the t-test (t). Panels (c)–(d) compare SP with the studentized bootstrap (SB) and studentized subsample (SS). Darker lines correspond to bigger sample sizes, that is, Panels (a) and (c): black \((n_1, n_2) = (500, 9500)\), dark gray \((n_1, n_2) = (250, 4750)\), and light gray \((n_1, n_2) = (100, 1900)\); Panels (b) and (d): black \((n_1, n_2) = (200, 9800)\), dark gray \((n_1, n_2) = (100, 4900)\), and light gray \((n_1, n_2) = (40, 1960)\). The x-axis represents the difference \(\theta(P_1 - \theta(P_2)\) and the y-axis, the simulated probability of rejection. Sizes of all tests are artificially adjusted such that the simulated rejection under the null is always equal to 5%. All estimates use the IK MSE-optimal bandwidth \(\hat{h}_{\text{mse}}\).

![Graph](image)

We consider 10,000 simulated samples and 1000 random permutations for each variation of the two designs. We compare the null rejection probability of three different tests at 5% nominal size: the non-studentized permutation test (NSP), the studentized permutation test (SP), and the t-test (t). All tests use the same bandwidth choice, and we experiment with four possibilities: three fixed choices of \(h\), 0.1, 0.3, and 0.5, and the IK data-driven MSE-optimal bandwidth \(\hat{h}_{\text{mse}}\).

Table 1 displays simulated rejection rates under the null hypothesis. DGPs with \(\sigma_1^2 = \sigma_2^2\) and \(h = 0.1\) or 0.3 fall under the sharp null hypothesis. As the theory predicts, both NSP and SP control size in these cases, but t fails to do so, most notably in Design 1 with \(n_1 = 40\) and Design 2 with \(\mu = 1\). Models with \(\sigma_1^2 = \sigma_2^2\) and \(h = 0.5\) fall outside the sharp null, and the local quadratic estimators are biased. Bias makes the mean of \(T_n\) diverge to infinity as the sample size increases, which explains the increasing size distortion of all tests. In these cases, all tests fail to control size, although the distortions are smaller for SP than for t.

The rows of Table 1 with \(\sigma_1^2 \neq \sigma_2^2\) fall outside of the sharp null. Cases with \(\sigma_1^2 \neq \sigma_2^2\) and \(h \leq 0.3\) violate the sharp null, but the distribution of \(T_n\) does not diverge as in the case of \(h = 0.5\). Design 1 with \(h = 0.1\) or 0.3 has a large size distortion of NSP and a small size distortion of SP that decreases with \(n\), as predicted by our theory. The size distortion of SP is much smaller than that of t, especially for smaller samples. For Design 2 with \(h = 0.1\) or 0.3, the size distortions of the permutation test are again smaller than t and decrease with \(n\). Finally, the IK MSE-optimal bandwidth \(\hat{h}_{\text{mse}}\) balances bias and variance, and does a good job keeping low size distortions of SP in all cases.

Since Hall and Hart (1990), many authors have proposed resampling procedures to test nonparametric hypotheses. It is natural to ask how some of these procedures compare to our robust permutation test. We implement the tests of Designs 1 and 2 using critical values for \(S_n\) generated by the wild bootstrap of Cao-Abad (1991) and the subsample of Politis, Romano, and Wolf (1999). The last two columns of Table 1 report the simulated rejection rates of studentized bootstrap (SB) and studentized subsample (SS) using the IK MSE-optimal bandwidth (Table 4 in Section E of the supplementary materials reports the rates for fixed \(h\)). The performance of SB and SS are generally

---

6 We use White’s robust formula for local quadratic regressions to compute standard errors, where the squared residuals are obtained by the nearest-neighbor matching estimator using three neighbors (Abadie and Imbens 2006).

---

6 For more details on bias correction see discussion in Section 3.1. Section D.5 of the supplementary materials demonstrates validity of our permutation tests with the LPR estimator.
similar to that of the t-test, with SS having worse size control in some cases.\footnote{It is worth noting that SB takes longer than SP to compute because it requires re-estimation of $h_k(X_k)$ for two different bandwidths and multiple observations \cite[p. 2227]{Cao-Abad1991}. We compare the computation time of all tests with two empirical illustrations in the next section (Table 2).} We shift the conditional mean function of $Y_2$ in both designs so that $\theta(P_1) \neq \theta(P_2)$. We then examine the power of SP and compare it to that of the $t$, SB, and SS tests. A direct comparison is difficult because not all tests control size in all cases. Thus, we artificially adjust the size of the tests to make sure they have a simulated rejection rate of 5% under the null hypothesis.\footnote{For the t-test, we obtain critical values from the simulated distribution under the null hypothesis, and keep those critical values to examine the simulated rejection rates under the alternative hypotheses. For the permutation, bootstrap, and subsampling tests, we numerically search for an artificial level $\alpha$ that gives us the simulated rejection rate of 5% under the null hypothesis. Once that artificial nominal level is found, we fix that nominal level and compute the simulated rejection probabilities under the various alternative hypotheses. Rejection may be randomized in case of ties (Equation (2.8)) in order for the numerical search to find a solution.}

Figures 1 and 2 display the simulated power curves for Designs 1 and 2, respectively, for cases with $\sigma_1^2 = \sigma_2^2 = 1$. Cases with $\sigma_1^2 > \sigma_2^2 = 1$ are found in Figures 5–6 of Section E in the supplementary materials. Panels (a)–(b) compare SP (solid line) to $t$ (dashed line), and Panels (c)–(d) compare SP (solid line) to SB (dashed line) and SS (dotted line). Each panel displays curves associated with three different sample sizes, with darker colors representing larger samples. The x-axis shows $\theta(P_1) - \theta(P_2)$, and the y-axis plots the simulated probability of rejection. All estimates in our power analysis use the IK MSE-optimal bandwidth $h_{\text{mse}}$.

In Design 1 (Figure 1), SP outperforms $t$, SB, and SS, particularly in cases with $n_1/(n_1 + n_2) = 0.02$; the dominance over SS is more pronounced for smaller samples. In Design 2 (Figure 2), there is no clear pattern of dominance between SP and $t$, however, SP dominates SB and SS in all cases. The discrepancies between SP and other tests converge to zero as $n$ increases, as predicted by our theory. Overall, we conclude that SP has size control superior to other tests without substantial costs, if any, in terms of power.

6. Empirical Examples

In this section, we revisit two classical examples in the RDD literature: Lee (2008) on U.S. House elections and Ludwig and Miller (2007) on the Head Start (HS) funding program. We illustrate the performance of our permutation test in practice and compare it to that of the $t$, bootstrap, and subsample tests.

Lee (2008) studies the electoral advantage of incumbent parties, using data on U.S. House of Representatives elections from 1946 to 1998. Since districts where a party’s candidate narrowly won an election are comparable to districts where the party’s candidate lost by a small margin, the difference in the electoral outcomes between these two groups in the subsequent election identifies the causal effect of party incumbency. Lee (2008) finds...
that an incumbent party has a significant causal advantage of a 0.08 vote share increase in the next election (Table 2 of Lee 2008).

Ludwig and Miller (2007) study the effects of HS on students’ health and schooling. The HS program was established in 1965 to provide preschool, health, and social services to poor children, aged three to five, and their families. The program provided technical assistance to the 300 poorest counties in the United States, based on the 1960 poverty rate. This created a discontinuity in program funding between the 300th and 301st poorest counties. Ludwig and Miller compare child mortality rates above and below this cutoff, and estimate that HS reduces mortality by 1.198 per 100,000 children (see Table 3 of Ludwig and Miller 2007, ages 5–9, HS-related causes, 1973–1983).

We test the null hypothesis of zero discontinuity using our robust permutation test on both datasets.9 We estimate the discontinuities with local-quadratic regressions and MSE-optimal bandwidths \( \hat{h}_{\text{mse}} \) for local-linear regressions, as explained in Section 5. The number of observations is 6558 for Lee (2008) and 3103 for Ludwig and Miller (2007), respectively.

Table 2 reports the \( p \)-values of our studentized permutation test (SP) and compares them to those of the \( t \), studentized bootstrap (SB), and studentized subsample (SS) tests, as implemented in Section 5. The party incumbency effect of Lee (2008) is strongly significant and robust across different tests for both choices of \( \hat{h}_{\text{mse}} \), that is, the \( \hat{h}_{\text{mse}} \) of Calonico, Cattaneo, and Titunik (2014) in the first row, and that of Imbens and Kalyanaraman (2012) in the second row. On the other hand, Ludwig and Miller (2007) find the effect of HS on child mortality to be marginally significant, and we find the significance level ranges between 1% and 12%, depending on the test and choice of \( \hat{h}_{\text{mse}} \). Finally, both examples demonstrate that our permutation test is feasible to compute in mere seconds, comparing favorably with the bootstrap in terms of computation time.

7. Conclusion

Classical two-sample permutation tests for the sharp null hypothesis of equal distributions are easy to implement and have exact size in finite samples. However, for testing equality of parameters that summarize distributions, classical permutation tests fail to control size. To fix this problem, we propose robust permutation tests based on studentized test statistics. Our framework is general enough to cover both parametric and nonparametric models with two samples or one sample split into two subsamples. We also propose confidence sets with correct asymptotic coverage that have exact coverage in finite samples if population distributions are the same up to a class of transformations. In a simulation study, our permutation test has good size control and power curves in finite samples, outperforming the conventional \( t \)-test and other resampling methods. Finally, we illustrate our permutation test with two empirical examples and show that its computation time is feasible, comparing favorably to the bootstrap.

Supplementary Materials

The online supplementary materials contain a PDF document with all the proofs, tables, and graphs that are referred to in the article. It also includes replication code for the simulations and empirical examples.

Acknowledgments

We thank Roger Koenker, Marcelo Moreira, Joe Romano, Azeem Shaikh, Xiaofeng Shao, the editor, associate editor, and two anonymous referees for many insightful comments. The article benefited from feedback given by seminar participants at the U. of Sao Paulo (FEARP), U. of Michigan, Boston U., U. of Essex, IAAE-Rotterdam, Bristol-ESG, KER Int’l Conference, UIUC, and Penn State. Bertanha acknowledges financial support received while visiting the Kenneth C. Griffin Department of Economics at the University of Chicago, where part of this work was conducted.

References

Abadie, A., and Imbens, G. W. (2006), “Large Sample Properties of Matching Estimators for Average Treatment Effects,” *Econometrica*, 74, 235–267. [2839]
Abou-Chadi, T., and Krause, W. (2020), “The Causal Effect of Radical Right Success on Mainstream Parties’ Policy Positions: A Regression Discontinuity Approach,” *British Journal of Political Science*, 50, 829–847. [2839]
Agarwal, S., Chomisengphet, S., Mahoney, N., and Stroebel, J. (2017), “Do Banks Pass Through Credit Expansions to Consumers Who Want to Borrow?” *Quarterly Journal of Economics*, 133, 129–190. [2839]
Armstrong, T. B., and Kolesár, M. (2018), “Optimal Inference in a Class of Regression Models,” *Econometrica*, 86, 655–683. [2839]
Bertanha, M., and Moreira, M. J. (2020), “Impossible Inference in Econometrics: Theory and Applications,” *Journal of Econometrics*, 218, 247–270. [2839]
Bugini, F. A., and Canay, I. A. (2021), “Testing Continuity of a Density via g-Order Statistics in the Regression Discontinuity Design,” *Journal of Econometrics*, 221, 138–159. [2834]
Caetano, C. (2015), “A Test of Exogeneity without Instrumental Variables in Models with Bunching,” *Econometrica*, 83, 1581–1600. [2840]
Calonico, S., Cattaneo, M. D., and Titunik, R. (2014), “Robust Nonparametric Confidence Intervals for Regression-discontinuity Designs,” *Econometrica*, 82, 2295–2326. [2845]
Canay, I. A., and Kamat, V. (2018), "Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design," Review of Economic Studies, 85, 1577–1608. [2834]
Canay, I. A., Romano, J. P., and Shaikh, A. M. (2017), "Randomization Tests under an Approximate Symmetry Assumption," Econometrica, 85, 1013–1030. [2834]
Cao-Abad, R. (1991), "Rate of Convergence for the Wild Bootstrap in Nonparametric Regression," The Annals of Statistics, 19, 2226–2231. [2843,2844]
Cattaneo, M. D., Frandsen, B. R., and Titiunik, R. (2015), "Randomization Tests in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate," Journal of Causal Inference, 3, 1–24. [2834]
Cattaneo, M. D., Jansson, M., and Ma, X. (2020), "Simple Local Polynomial Regression," The Annals of Statistics, 115, 1449–1455. [2840]
Chaudhuri, P. (1991), "Nonparametric Estimates of Regression Quantiles and Their Local Bahadur Representation," Journal of the American Statistical Association, 86, 760–777. [2839]
Chung, E., and Romano, J. P. (2013), "Exact and Asymptotically Robust Permutation Tests for The Regression Discontinuity Estimator," Review of Economic Studies, 80, 176–183. [2839]
—— (2016a), "Asymptotically Valid and Exact Permutation Tests Based on Two-sample U-statistics," Journal of Statistical Planning and Inference, 168, 97–105. [2834]
—— (2016b) "Multivariate and Multiple Permutation Tests," Journal of Econometrics, 193, 76–91. [2834]
Chung, E., and Olivares, M. (2021), "Permutation Test for Heterogeneous Treatment Effects with a Nuisance Parameter," Journal of Econometrics, 225, 148–174. [2834]
Chung, E., and Romano, J. P. (2013), "Exact and Asymptotically Robust Permutation Tests," The Annals of Statistics, 41, 484–507. [2833,2834]
Cao-Abad, R. (1991), "Rate of Convergence for the Wild Bootstrap in Nonparametric Regression," The Annals of Statistics, 19, 2226–2231. [2843,2844]
Cattaneo, M. D., Frandsen, B. R., and Titiunik, R. (2015), "Randomization Tests in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate," Journal of Causal Inference, 3, 1–24. [2834]
Cattaneo, M. D., Jansson, M., and Ma, X. (2020), "Simple Local Polynomial Density Estimators," Journal of the American Statistical Association, 115, 1449–1455. [2840]
Chaudhuri, P. (1991), "Nonparametric Estimates of Regression Quantiles and Their Local Bahadur Representation," Journal of the American Statistical Association, 86, 760–777. [2839]
Chung, E., and Olivares, M. (2021), "Permutation Test for Heterogeneous Treatment Effects with a Nuisance Parameter," Journal of Econometrics, 225, 148–174. [2834]
Chung, E., and Romano, J. P. (2013), "Exact and Asymptotically Robust Permutation Tests," The Annals of Statistics, 41, 484–507. [2833,2834]
—— (2016a), "Asymptotically Valid and Exact Permutation Tests Based on Two-sample U-statistics," Journal of Statistical Planning and Inference, 168, 97–105. [2834]
—— (2016b) "Multivariate and Multiple Permutation Tests," Journal of Econometrics, 193, 76–91. [2834]
DiCiccio, C. J., and Romano, J. P. (2017), "Robust Permutation Tests for Correlation and Regression Coefficients," Journal of the American Statistical Association, 112, 1211–1220. [2834]
Fan, J., and Gijbels, I. (1996), Local Polynomial Modelling and its Applications, 66 of Monographs on Statistics and Applied Probability, Boca Raton, FL: CRC Press. [2837,2839]
Fan, J., Hu, T.-C., and Truong, Y. K. (1994) “Robust Non-parametric Function Estimation,” Scandinavian Journal of Statistics, 21, 433–446. [2839]
Fogarty, C. B. (2021), "Prepivoted Permutation Tests," arXiv preprint arXiv:2102.04423. [2834]
Hahn, J., Todd, P., and Van der Klaauw, W. (2001), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," Econometrica, 69, 201–209. [2839]
Hall, P., and Hart, J. D. (1990) "Bootstrap Test for Difference Between Means in Nonparametric Regression," Journal of the American Statistical Association, 85, 1039–1049. [2843]
Imbens, G. W., and Kalyanaraman, K. (2012) “Optimal Bandwidth Choice for The Regression Discontinuity Estimator,” Review of Economic Studies, 79, 933–959. [2842,2845]
Imbens, G. W., and Rosenbaum, P. R. (2005), "Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education," Journal of the Royal Statistical Society, Series A, 168, 109–126. [2834]
Janssen, A. (1997), "Studentized Permutation Tests for Non-Li.d. Hypotheses and the Generalized Behrens-Fisher Problem," Statistics & Probability Letters, 36, 9–21. [2834]
—— (2005) "Resampling Student's t-type Statistics," Annals of the Institute of Statistical Mathematics, 57, 507–529. [2834]
Kamat, V. (2018), "On Nonparametric Inference in the Regression Discontinuity Design," Econometric Theory, 34, 694–703. [2839]
Lee, D. S. (2008) "Randomized Experiments from Non-random Selection in U.S. House Elections," Journal of Econometrics, 142, 675–697. [2844,2845]
Lehmann, E. L., and Romano, J. P. (2005) Testing Statistical Hypotheses, New York: Springer. [2836]
Li, Q., and Racine, J. S. (2007), Nonparametric Econometrics: Theory and Practice, Princeton: Princeton University Press. [2837,2838]
Ludwig, J., and Miller, D. (2007) "Does Head Start Improve Children's Life Chances? Evidence From a Regression Discontinuity Design," Quarterly Journal of Economics, 122, 159–208. [2844,2845]
Marron, J. S., and Ruppert, D. (1994), "Transformations to Reduce Boundary Bias in Kernel Density Estimation," Journal of the Royal Statistical Society, Series B, 56, 653–671. [2840]
Neubert, K., and Brunner, E. (2007), "A Studentized Permutation Test for the Non-parametric Behrens-Fisher Problem," Computational Statistics & Data Analysis, 51, 5192–5204. [2834]
Neuhaus, G. (1993), "Conditional Rank Tests for the Two-Sample Problem Under Random Censorship," The Annals of Statistics, 21, 1760–1779. [2834]
Pauly, M., Brunner, E., and Konietschke, F. (2015) "Asymptotic Permutation Tests in General Factorial Designs," Journal of the Royal Statistical Society, Series B, 77, 461–473. [2834]
Politis, D. N., Romano, J. P., and Wolf, M. (1999), Subsampling, New York: Springer. [2843]
Pollard, D. (1991), "Asymptotics for Least Absolute Deviation Regression Estimators," Econometric Theory, 7, 186–199. [2839]
Racine, J. (2001), "Bias-Corrected Kernel Regression," Journal of Quantitative Economics, 17, 25–42. [2837]
Romano, J. P. (1990) "On the Behavior of Randomization Tests Without a Group Invariance Assumption," Journal of the American Statistical Association, 85, 686–692. [2833]
Saez, E. (2010) "Do Taxpayers Bunch at Kink Points?" American Economic Journal: Economic Policy, 2, 180–212. [2840]
Shaikh, A. M., and Toulis, P. (2021) "Randomization Tests in Observational Studies with Staggered Adoption of Treatment," Journal of the American Statistical Association, 116, 1835–1848. [2834]
Thistlreithwaite, D. L., and Campbell, D. T. (1960), "Regression-discontinuity Analysis: An Alternative to the Ex Post Facto Experiment," Journal of Educational Psychology, 51, 309–317. [2839]
Valentine, J. C., Konstantopoulos, S., and Goldrick-Rab, S. (2017) "What Happens to Students Placed into Developmental Education? A Meta-analysis of Regression Discontinuity Studies," Review of Educational Research, 87, 806–833. [2839]
Zoorob, M. (2020), "Do Police Brutality Stories Reduce 911 Calls? Reassessing an Important Criminological Finding," American Sociological Review, 85, 176–183. [2839]