ASYMPTOTICALLY VALID AND EXACT PERMUTATION TESTS BASED ON TWO-SAMPLE U-STATISTICS

By

Eun Yi Chung
Joseph P. Romano

Technical Report No. 2011-09
October 2011

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

ASYMPTOTICALLY VALID AND EXACT PERMUTATION TESTS BASED ON TWO-SAMPLE U-STATISTICS

By

Eun Yi Chung
Joseph P. Romano
Stanford University

Technical Report No. 2011-09
October 2011

This research was supported in part by
National Science Foundation grant DMS 0707085.

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu

Asymptotically Valid and Exact Permutation Tests Based on Two-sample U-Statistics

EunYi Chung and Joseph P. Romano
Departments of Economics and Statistics
Stanford University
Stanford, CA 94305-4065

July 25, 2011

Abstract

The two-sample Wilcoxon test has been widely used in a broad range of scientific research, including economics, due to its good efficiency, robustness against parametric distributional assumptions, and the simplicity with which it can be performed. While the two-sample Wilcoxon test, by virtue of being both a rank and hence a permutation test, controls the exact probability of a Type 1 error under the assumption of identical underlying populations, it *in general* fails to control the probability of a Type 1 error, even asymptotically. Despite this fact, the two-sample Wilcoxon test has been misused in many applications. Through examples of misapplications in academic economics journals, we emphasize the need for clarification regarding both what is being tested and what the implicit underlying assumptions are. We provide a general theory whereby one can construct a permutation test of a parameter $\theta(P, Q) = \theta_0$ which controls the asymptotic probability of a Type 1 error in large samples while retaining the exactness property when the underlying distributions are identical. In addition, the studentized Wilcoxon retains all the benefits of the usual Wilcoxon test, such as its asymptotic power properties and the fact that its critical values can be tabled (which we provide). The results generalize from the two-sample Wilcoxon statistic to general U-statistics. The main ingredient that aids our asymptotic derivations is a useful coupling method. A Monte Carlo simulation study and empirical applications are presented as well.

1 Introduction

The two-sample Wilcoxon test has been widely applied in many areas of academic research, including in the field of economics. For example, the two-sample Wilcoxon test has been used in journals such as *American Economic Review*, *Quarterly Journal of Economics* and *Experimental Economics*. Specifically, in 2009, over one-third of the papers written in *Experimental Economics* utilized the two-sample Wilcoxon test. Furthermore, 30% of the articles across five biomedical journals published in 2004 utilized the two-sample Wilcoxon test (Okeh, 2009). However, we argue that the permutation tests have generally been misused across all disciplines and in this paper, we formally examine this problem in great detail.

To begin, the permutation test is a powerful method that attains the *exact* probability of a Type 1 error in finite samples for *any* test statistic, as long as the assumption of identical distributions holds. Under such an assumption, since all the observations are i.i.d., the distribution of the sample under a permutation is the same as that of the original sample. In this regard, the permutation distribution, which is constructed by recomputing a test statistic over permutations of the data, can serve as a valid null distribution,which enables one to obtain an exact level α test even in finite samples.

To be more precise, assume $X_1, \ldots, X_m$ are i.i.d. observations from a probability distribution P, and independently, $Y_1, \ldots, Y_n$ are i.i.d. from Q. By putting all $N = m+n$ observations together, let the data Z be described as

$$Z \equiv (Z_1, \ldots, Z_N) = (X_1, \ldots, X_m, Y_1, \ldots, Y_n) \ .$$

For now, suppose we are interested in testing the null hypothesis $H_0 : (P, Q) \in \bar{\Omega}$, where $\bar{\Omega} = \{(P, Q) : P = Q\}$. Let $\mathbf{G}_N$ denote the set of all permutations π of $\{1, \ldots, N\}$. Under the null hypothesis $\bar{\Omega}$, the joint distribution of $(Z_{\pi(1)}, \ldots, Z_{\pi(N)})$ is the same as that of $(Z_1, \ldots, Z_N)$ for any permutation $(\pi(1), \ldots, \pi(N))$ in $\mathbf{G}_N$. Given any real-valued test statistic $T_{m,n}(Z)$ for testing H_0, recompute the test statistic $T_{m,n}$ for all $N!$ permutations π, and for given $Z = z$, let

$$T_{m,n}^{(1)}(z) \leq T_{m,n}^{(2)}(z) \leq \cdots \leq T_{m,n}^{(N!)}(z)$$

be the ordered values of $T_{m,n}(Z_{\pi(1)}, \ldots, Z_{\pi(N)})$ as π varies in $\mathbf{G}_N$. Given a nominal level $\alpha, 0 < \alpha < 1$, let k be defined by $k = N! - [N!\alpha]$, where $[N!\alpha]$ denotes the largest integer less than or equal to $N!\alpha$. Let $M^+(z)$ and $M^0(z)$ be the numbers of values $T_{m,n}^{(j)}(z)$ $(j = 1, \ldots, N!)$ that are greater than $T^{(k)}(z)$ and equal to $T^{(k)}(z)$, respectively.

Set

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

Let the permutation test function $\phi(z)$ be defined by

$$\phi(z) = \begin{cases} 1 & \text{if} \quad T_{m,n}(x) > T^{(k)}_{m,n}(z) \ , \\ a(z) & \text{if} \quad T_{m,n}(z) = T^{(k)}_{m,n}(z) \ , \\ 0 & \text{if} \quad T_{m,n}(z) < T^{(k)}_{m,n}(z) \ . \end{cases}$$

Note that, under H_0,

$$\mathrm{E}_{P,Q}[\phi(X_1, \ldots, X_m, Y_1, \ldots, Y_n)] = \alpha \ .$$

In other words, the test ϕ is exact level α under the null hypothesis $H_0 : P = Q$. (As will be seen later, the rejection probability need not be α even asymptotically when $P \neq Q$, but we will be able to achieve this for a general class of studentized test statistics.)

In addition, when the number of elements in $\mathbf{G}_N$ is large, one can instead adopt an approximation to construct the permutation test using simulations. More specifically, first randomly sample permutations π_b for $b = 1, \ldots, B-1$ from $\mathbf{G}_N$ with replacement. Let π_B be $\{1, \ldots, N\}$ so that $Z_{\pi_B} = Z$. Next, calculate the empirical p-value $\tilde{p}$, which is the proportion of permutations under which the statistic value exceeds or equal to the statistic given the original samples, i.e.,

$$\tilde{p} \equiv \frac{1}{B} \sum_{i=1}^{B} \mathrm{I}\{T_{m,n}(Z_{\pi_b(1)}, \ldots, Z_{\pi_b(N)}) \geq T_{m,n}(Z_1, \ldots, Z_N)\} \ ,$$

where $T_{m,n}(Z_{\pi_B}) = T_{m,n}(Z)$. Then, this empirical p-value serves as a good approximation to the true p-value. Thus, the test that rejects when $\tilde{p} \leq \alpha$ controls the probability of a Type 1 error (if the test based on all permutations does).

Also, let $\hat{R}^T_{m,n}(t)$ denote the permutation distribution of $T_{m,n}$ defined by

$$\hat{R}^T_{m,n}(t) = \frac{1}{N!} \sum_{\pi \in \mathbf{G}_N} I\{T_{m,n}(Z_{\pi(1)}, \ldots, Z_{\pi(N)}) \leq t\}, \tag{1}$$

where $\mathbf{G}_N$ denotes a set of all $N!$ permutations of $\{1, \ldots, N\}$. Since the permutation distribution asymptotically approximates the unconditional true sampling distribution of $T_{m,n}$, a statistical inference can be made based on the permutation distribution; roughly speaking, if the test statistic $T_{m,n}$ evaluated at the original sample falls within the 100α % range of the tail of the permutation distribution, the null hypothesis H_0 is rejected.

However, in many applications permutation tests are used to test a null hypothesis Ω_0 which is strictly larger than $\bar{\Omega}$. For example, suppose the null hypothesis of interest Ω_0 specifies a null value θ_0 for some functional $\theta(P, Q)$, i.e., $\Omega_0 = \{(P, Q) : \theta(P, Q) = \theta_0\}$, where $\theta_0 \equiv \theta(P, P)$ so that $\Omega_0 \supset \bar{\Omega}$. For testing Ω_0, unfortunately, one cannot necessarily just apply a permutation test because the argument under which the permutation test is constructed breaks down under Ω_0; observations are no longer i.i.d. and thus, the distribution of the sample under a permutation is no longer the same as that of the original. As a result, the permutation distribution no longer asymptotically approximates the unconditional true sampling distribution of $T_{m,n}$ in general. Problems can arise if one attempts to argue that the rejection of the test implies the rejection of the null hypothesis that the parameter θ is the specified value θ_0. Tests can be rejected not because $\theta(P, Q) = \theta_0$ is not satisfied, but because the two samples are not generated from the same underlying probability law. In fact, there can be a large probability of declaring $\theta > \theta_0$ when in fact $\theta \leq \theta_0$. Therefore, as Romano (1990) points out, the usual permutation construction for the two-sample problems in general is invalid.

For the case of testing equality of survival distributions, however, Neuhaus (1993) discovered that if a survival statistic of interest (log-rank statistic) is appropriately studentized by a consistent standard error, the permutation test based on the studentized statistic achieves asymptotic validity. In other words, it can control the asymptotic probability of a Type 1 error in large samples, even if the censoring distributions are possibly different, but still retains the exact control of the Type 1 error in finite samples when the censoring distributions are identical. This perceptive idea has been applied to other specific applications in Janssen (1997), Neubert and Brunner (2007), and Pauly (2010). Our goal is to synthesize the results of the same phenomenon and apply a general theory to a class of two-sample U-statistics, which includes means, variances and the Wilcoxon statistic as well as many others.

The main purpose of this paper is twofold: (i) to emphasize the need for clarification regarding what is being tested and the implicit underlying assumptions by examples of applications that incorrectly utilize the two-sample Wilcoxon test, and (ii) to provide a general theory whereby one can construct a permutation test of a parameter $\theta(P, Q) = \theta_0$ that can be estimated by its corresponding U-statistic, which controls the asymptotic probability of a Type 1 error in large samples while retaining the exactness property when the underlying distributions are identical. By choosing as test statistic a studentized version of the U-statistic estimator, the correct asymptotic rejection probability can be achieved while maintaining the exact finite-sample rejection probability of α when $P = Q$. Note that this exactness property is what makes the proposed permutation procedure more attractive than other asymptotically valid alternatives, such as bootstrap or subsampling.

Our paper begins by investigating the two-sample Wilcoxon test in Section 2. The two-sample Wilcoxon test has been widely used for its virtues of good efficiency, robustness against parametric assumptions, and simplicity with which it can be performed – since it is a rank statistic, the critical values can be tabled so that the permutation distribution need not be recomputed for a new data set. However, the Wilcoxon test is only valid if the fundamental assumption of identical distributions holds. If the null hypothesis is such that lack of equality of distributions need not be accompanied by the alternative hypothesis, the Wilcoxon test is not a valid approach, even asymptotically. Nevertheless, the Wilcoxon test has been prevalently used for testing equality of means or medians, in which it fails to control the probability of a Type 1 error, even asymptotically unless one is willing to assume a shift model, an assumption that may be unreasonable or unjustifiable. Furthermore, even for the case of testing equality of distributions where the Wilcoxon test achieves exact control of the probability of a Type 1 error, it does not have much power detecting the difference in distributions. The two-sample Wilcoxon test is only appropriate as a test of $H_0 : P(X \leq Y) = \frac{1}{2}$. However, even in this case, the Wilcoxon test fails to control the probability of a Type 1 error, even asymptotically unless the Wilcoxon statistic is properly studentized.

We propose a permutation test based on an appropriately studentized Wilcoxon statistic. We find that the studentized two-sample Wilcoxon test achieves the correct asymptotic rejection probability under the null hypothesis H_0, while controlling the exact probability of a Type 1 error in finite samples when $P = Q$. In addition, the studentized Wilcoxon test has the same asymptotic Pitman efficiency as the standard Wilcoxon test when the underlying distributions are equal. Part of our results on the studentized Wilcoxon test are similar to the results found in Neubert and Brunner (2007), who studied the limiting behavior of the studentized Wilcoxon test. In contrast to their paper, we generalize the result to a class of two-sample U-statistics while also showing that when the underlying distributions are continuous, the studentized Wilcoxon test retains all the benefits of the usual Wilcoxon test. That is, under the continuity assumption of the underlying distributions, the standard error for the Wilcoxon test is indeed a rank statistic, so that the proof of the result becomes quite simple and more importantly, the critical values of the new test can also be tabled. Therefore, the permutation distribution need not be recomputed with a new data set. As such, we provide tables for the critical values of the new test in Table 4.

In Section 3, we provide a general framework where the asymptotic validity of the permutation test holds for testing a parameter $\theta(P, Q)$ that can be estimated by a corresponding U-statistic. Assuming that the kernel of the U-statistic is antisymmetric and there exist consistent estimators for the asymptotic variance, we provide a test procedure that controls the asymptotic probability of a Type 1 error in large samples, but still retains the exactness property if $P = Q$. Another paper by Chung and Romano (2011) also provides a quite general theory for testing $\theta(P) = \theta(Q)$, where the test statistic is based on the difference of the estimators that are asymptotically linear with additive error terms. Although this class of estimators, which include means, medians, and variances, are quite general as well, it does not include cases like the Wilcoxon statistic where the parameter of interest is a function of P and Q together $\theta(P, Q)$ (as opposed to the difference $\theta(P) - \theta(Q)$) and there are no additive error terms involved. Thus, the results derived in Chung and Romano (2011) are not directly applicable in the class of U-statistics and a careful analysis is required. Furthermore, this class of estimators is useful and beneficial because it directly applies to the Wilcoxon test without having to assume continuity of the underlying distributions. And, it also applies to other popular tests, such as tests for equality of means and variances. In deriving our results, we employ a coupling method which allows us to understand the limiting behavior of the permutation distribution under the given samples from P and Q by instead looking at the unconditional sampling distribution of the statistic when all N observations are i.i.d. from the mixture distribution $\bar{P} = pP + qQ$.

Section 4 presents Monte Carlo studies illustrating our results for the Wilcoxon test and lastly, empirical applications of the studentized Wilcoxon test are provided in Section 5.

2 Two-Sample Wilcoxon Test

In this section, we investigate the two-sample Wilcoxon test, a test that is based on a special case of a two-sample U-statistic called the two-sample Wilcoxon statistic. Assume that the observations consist of two independent samples $X_1, \ldots, X_m \sim$ i.i.d. P and $Y_1, \ldots, Y_n \sim$ i.i.d. Q. In this section, both P and Q are assumed continuous. (This assumption will be relaxed in Section 3.) Consider the two-sample Wilcoxon test statistic

$$\hat{\theta} = U_{m,n} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} I(X_i \leq Y_j) \ . \tag{2}$$

The two-sample Wilcoxon test has prominently been applied in the field of economics, primarily in experimental economics. Some examples include Feri, Irlenbusch, and Sutter (2010), Sutter (2009), Charness, Rigotti, and Rustrichini (2007), Plott and Zeiler (2005), Davis (2004), and Sausgruber (2009). The popularity of the two-sample Wilcoxon test can be attributed not only to its robustness against parametric distribution assumptions, but also due to its simplicity with which it can be performed; since it is a rank statistic, the null distribution can be tabled, so that the permutation distribution need not be recomputed with a new data set. Furthermore, as Lehmann (2009) points out, the two-sample Wilcoxon test is fairly efficient under a shift model,

$$Q(y) = P(y - \Delta), \quad \Delta \geq 0 \ . \tag{3}$$

In comparison to the standard two-sample t-test, the two-sample Wilcoxon test has a much higher power under heavy-tailed distributions while barely losing any power under normality. More specifically, the Pitman asymptotic relative efficiency of the two-sample Wilcoxon to the two-sample t-test is 1.5 and 3.0 under double exponential and exponential distribution, respectively, and 0.955 under normality. Such advantageous properties and attributes make the two-sample Wilcoxon test a favored approach in a wide range of research. However, to our surprise, most applications of the two-sample Wilcoxon test in academic journals turn out to be inaccurate; it has been mainly utilized to test equality of means, medians, or distributions. As it will be argued, such applications of the two-sample Wilcoxon test is theoretically invalid or at least incompatible. Our main goal is to understand why such applications can be misleading and ultimately, to construct a test for testing H_0 that controls the asymptotic probability of a Type 1 error in large samples while maintaining the exactness property in finite samples when $P = Q$.

2.1 Misapplication of the Wilcoxon Test

First, we illustrate examples of misapplication appearing in the academic literature, which makes it clear that it is essential to fully understand the distinction between what one is trying to test and what one is actually testing. To begin, consider testing equality of medians for two independent samples. A suitable test would reject the null at the nominal level α (at least asymptotically). However, the Wilcoxon test may yield rejection probability far from the nominal level α. For example, consider two independent distributions $X \sim N(\ln(2), 1)$ and $Y \sim \exp(1)$. Despite the same median $\ln(2)$, a Monte Carlo simulation study using the Wilcoxon test shows that the rejection probability for a two-sided test turns out to be 0.2843 when α is set to 0.05. The problem is that the Wilcoxon test only picks up divergence from $P(X \leq Y) = \frac{1}{2}$ and in the example considered here, $P(X < Y) = 0.4431 \neq \frac{1}{2}$. Consequently, the Wilcoxon test used for examining equality of medians may lead to inaccurate inferences. Certainly, the Wilcoxon test rejects the null too often, and the conclusion that the population medians differ is wrong. At this point, one may be happy to reject in that it indicates a difference in the underlying distributions. However, if we are truly interested in detecting any difference between P and Q, the Wilcoxon test has no power against P and Q with $\theta(P, Q) = \frac{1}{2}$.

Similarly, using the Wilcoxon test for testing equality of means may cause an analogous problem. One can easily think of situations where two distinct underlying distributions have the same mean but $P(X \leq Y) \neq \frac{1}{2}$. In such cases, the rejection probability under the null of $\mu(P) = \mu(Q)$ can be very far from the nominal level α because $P(X \leq Y)$ may not be $\frac{1}{2}$ even when $\mu(P) = \mu(Q)$. To illustrate how easily things can go wrong, let us consider an example presented in Sutter (2009). Sutter uses the Mann-Whitney U-test to examine the effects of group membership on individual behavior to team decision making. In his analysis, the average investments in PAY-COMM with 18 observations and MESSAGE with 24 observations are compared. The estimated densities using kernel density estimates for PAY-COMM and MESSAGE, denoted P and Q, respectively, are plotted in Figure 1.

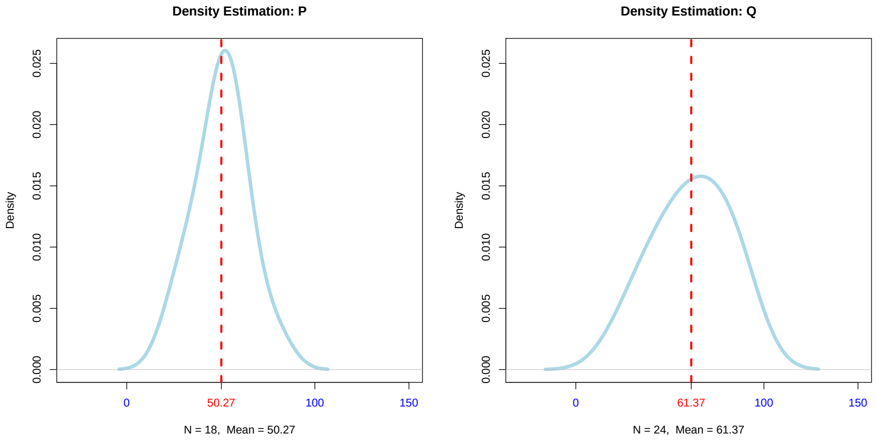

Figure 1: Sutter (2009): Underlying Distributions P & Q

For the two-sample Wilcoxon test to be"valid" in the sense that it controls the probability of a Type 1 error, the two underlying distributions need to satisfy a shift model assumption (3), i.e., these two distributions must be identical under the null hypothesis and the *only* possible difference between the distributions has to be a location (mean). Otherwise, applying the two-sample Wilcoxon test to examine the equality of means may leads to faulty inferences. In his analysis, Sutter rejects the null hypothesis that the average investments in PAY-COMM and MESSAGE are the same at the 10 % significance level (p-value of 0.069). Had it been set to the conventional 5% significance level, however, the Wilcoxon test would have failed to reject the null hypothesis. The studentized permutation t-test, a more suitable test for testing the equality of means (in the sense that it controls the probability of a Type 1 error at least asymptotically), yields a p-value of 0.042 and rejects the null hypothesis even at the 5% significance level.

One of many prevalent applications of the Wilcoxon test is its use in testing the equality of distributions. Of course, when two underlying distributions are identical, $P(X \leq Y) = \frac{1}{2}$ is satisfied and thus, the two-sample Wilcoxon test results in exact control of the probability of a Type 1 error in finite sample cases. However, it does not have much power in detecting distributional differences; since the two-sample Wilcoxon test only picks up divergences from $P(X \leq Y) = \frac{1}{2}$, if the underlying distributions are different but satisfy $P(X \leq Y) = \frac{1}{2}$, the test fails to detect the difference of the two underlying distributions. Despite this fact, the two-sample Wilcoxon test has been prevalently applied to test the equality of distributions. Plott and Zeiler (2005), for example, perform the Wilcoxon-Mann-Whitney rank sum test to examine the null hypothesis that willingness to pay (WTP) and willingness to accept (WTA) are drawn from the same distribution. Figure 2 displays the estimated density of WTP denoted P and that of WTA denoted Q. In their analysis for experiment 3, the Wilcoxon-Mann-Whitney test yields a z value of 1.738 (p-value = 0.0821), resulting in a failure to reject the null hypothesis.

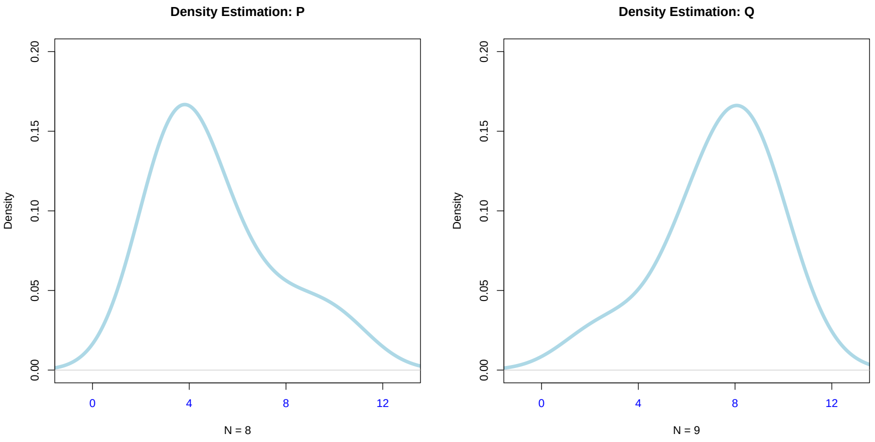

Figure 2: Plott and Zeiler (2005): Underlying Distributions P (WTP) & Q (WTA)

However, when testing equality of distributions, it is more advisable to use a more omnibus statistic, such as the Kolmogorov-Smirnov or the Cramér-von Mises statistic, which captures the differences of the entire distributions as opposed to only assessing a particular aspect of the distributions. In the example of Plott and Zeiler, the Cramér-von

Mises yields a p-value of 0.0546.

In addition, it is worth noting that in testing equality of medians, Plott and Zeiler employed the Mood's median test. Based on a Pearson's chi-square statistic, the Mood's median test examines the null that the probability of an observation being greater than the overall median is the same for all populations. The test yields a Pearson χ^2 test statistic of 1.5159 (p-value = 0.218), leading Plott and Zeiler to fail to reject the null hypothesis that WTP and WTA have the same median. The problem of this median test is that difference of the medians may not be well reflected on the proportion of the observations that are greater than the overall median, especially when the sample sizes are small. As a result, this test has low power in detecting median differences and a more suitable test for testing equality of medians is desirable. The studentized permutation median test (Chung and Romano, 2011) results in a p-value of 0.0290, rejecting the null hypothesis of identical medians. We will revisit this problem in Section 5.

All the cases considered so far exemplify inappropriate applications of the two-sample Wilcoxon test; what the researchers intend to test (testing equality of medians, means, or distributions) is incongruent with what the Wilcoxon test is actually testing ($H_0 : P(X \leq Y) = \frac{1}{2}$). However, as will be argued next, even when testing the null $H_0 : P(X \leq Y) = \frac{1}{2}$, the standard Wilcoxon test is invalid *unless* it is appropriately studentized.

2.2 Studentized Two-Sample Wilcoxon Test

In this section, consider using the Wilcoxon statistic to test the null hypothesis

$$H_0 : P(X \leq Y) = \frac{1}{2}$$

against the alternative

$$H_1 : P(X \leq Y) > \frac{1}{2} \ .$$

Testing H_0 would be desired in many applications, such as testing whether the life span of a particular drug user tends to be longer than that of others as a measure of the treatment effect of drug use, or testing the treatment effects of some policy such as unemployment insurance or a job training program on the unemployment spells.

Under the null hypothesis $H_0 : P(X \leq Y) = \frac{1}{2}$, however, rejecting the null does not necessarily imply $P(X \leq Y) > \frac{1}{2}$ since the null can be rejected not because $P(X \leq Y) > \frac{1}{2}$, but because the underlying distributions P and Q are not identical. This is an inherent common problem of the (usual) permutation test when the fundamental assumption of the identical distributions fails to hold.

Theorem 2.1. *Assume $X_1, \dots, X_m$ are i.i.d. P and, independently, $Y_1, \dots, Y_n$ are i.i.d. Q. Suppose $P(X \leq Y) = \frac{1}{2}$. Let $\min(m, n) \to \infty$ with $\frac{m}{N} \to p \in (0, 1)$. Assume P and Q are continuous. Let $\hat{R}^U_{m,n}$ denote the permutation distribution of $\sqrt{m}(U_{m,n} - \theta)$ defined in (1) with T replaced by U. Then, the permutation distribution $\hat{R}^U_{m,n}$ satisfies*

$$\sup_t |\hat{R}^U_{m,n}(t) - \Phi(t/\tau)| \to 0 \ ,$$

where

$$\tau^2 = \frac{1}{12(1-p)} \ . \tag{4}$$

Under the conditions of Theorem 2.1, the permutation distribution of $\sqrt{m}(U_{m,n} - \theta)$ is approximately normal with mean 0 and variance τ^2, whereas the true unconditional sampling distribution of $\sqrt{m}(U_{m,n} - \theta)$ is asymptotically normal with mean 0 and variance

$$\xi_x + \frac{p}{1-p}\xi_y, \text{ where } \ \xi_x = \text{Var}\left(Q^-_Y(X_i)\right) \text{ and } \xi_y = \text{Var}\left(P_X(Y_j)\right) \ , \tag{5}$$

which does not equal τ^2 unless $(1-p)\xi_x + p\xi_y = \frac{1}{12}$ is satisfied. Since the true sampling distribution can be very far from $\mathcal{N}\big(0, \frac{1}{12(1-p)}\big)$, the permutation test in general fails to control the probability of a Type 1 error. For example, in the case of $P = N(0,1)$ and $Q = N(0,5)$, despite $\theta = P(X \leq Y) = \frac{1}{2}$, the rejection probability of the two-sample Wilcoxon test is 0.0835, which is far from the nominal level $\alpha = 0.05$. (For more examples, see Table 2 in Section 4.) Therefore, one must be careful when interpreting the rejection of the null. Rejecting the null should not be interpreted as rejection of $P(X \leq Y) = \frac{1}{2}$ because it may in fact be caused by unequal distributions.

Remark 2.1. A classical approach to resolve the problem when the underlying distributions are possibly nonidentical is through asymptotics; by appropriate studentization of $U_{m.n}$ by a consistent standard error, an asymptotic rejection probability of α can be obtained in large samples. For example, define the critical value

$$z_{1-\alpha}\sqrt{\hat{\xi}_x + \frac{m}{n}\hat{\xi}_y}$$

where

$$\hat{\xi}_x = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{Q}^-(X_i) - \frac{1}{m}\sum_{i=1}^{m}\hat{Q}^-(X_i)\right)^2 \ , \tag{6}$$

$$\hat{\xi}_y = \frac{1}{n-1}\sum_{j=1}^{n}\left(\hat{P}(Y_j) - \frac{1}{n}\sum_{j=1}^{n}\hat{P}(Y_j)\right)^2 \ , \tag{7}$$

$\hat{Q}^{-}(X_i) = \frac{1}{n}\sum_{j=1}^{n} \mathrm{I}\{Y_j < X_i\}$, and $\hat{P}(Y_j) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{I}\{X_i \leq Y_j\}$. Then, the one-sided test allows one to achieve the rejection probability equal to α in large samples given fixed distributions P and Q, but *only* asymptotically.

For testing the null hypothesis of $\theta = P(X \leq Y) = \frac{1}{2}$ without imposing the equal distribution assumption, we propose to use the two-sample Wilcoxon test based on correctly studentized statistic as it attains the rejection probability equal to α asymptotically while still maintaining exactness property in finite samples if $P = Q$.

Theorem 2.2. *Suppose $P(X \leq Y) = \frac{1}{2}$. Let $\min(m, n) \to \infty$ with $\frac{m}{N} \to p \in (0, 1)$. Assume P and Q are continuous. Define the studentized Wilcoxon test statistic*

$$\tilde{U}_{m,n} = \frac{U_{m,n} - \theta}{\sqrt{\hat{\xi}_x + \frac{m}{n}\hat{\xi}_y}} \ , \tag{8}$$

where $\hat{\xi}_x$ and $\hat{\xi}_y$ are given by (6) and (7), respectively. Then, the permutation distribution of $\sqrt{m}\tilde{U}_{m,n}$ given by (1) with T replaced by $\tilde{U}$ satisfies

$$\sup_t |\hat{R}^{\tilde{U}}_{m,n}(t) - \Phi(t)| \to 0 \ ,$$

and so the critical value $\hat{r}_{m,r}$ satisfies $\hat{r}_{m,n} \to z_{1-\alpha}$.

Under the conditions of Theorem 2.2, the permutation distribution of the test statistic (8) is asymptotically normal with mean 0 and variance 1, which is the same as the limiting sampling distribution. Indeed, the motivation of the studentized two-sample Wilcoxon test stems from the fact that the limiting sampling distribution of $\sqrt{m}\tilde{U}_{m,n}$ no longer depends on the true sample distributions P and Q. By correctly studentizing the standard two-sample Wilcoxon statistic, the true unconditional sampling distribution converges in distribution to standard normal in large samples. This asymptotic "distribution-free" property allows one to achieve asymptotic rejection probability equal to α while maintaining exact rejection probability α in finite samples when $P = Q$.

Remark 2.2. It is crucial to realize that the estimators of ξ_x and ξ_y are themselves rank statistics; $\hat{\xi}_x$, for example, can be calculated from the formula given by

$$\hat{\xi}_x = \frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{1}{n}(S_i - i) - \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n}(S_i - i)\right)\right)^2 \ ,$$

where $S_1 < S_2 < \ldots < S_m$ are the ranks of the Xs in the combined sample. Similarly, ξ_y can be expressed as a function of the ranks of the Ys in the combined sample. The fact that the standard error estimate is a rank statistic allows the studentized two-sample Wilcoxon test to retain all the benefits of the usual two-sample Wilcoxon test as a rank test. In particular, the permutation distribution of the studentized statistic need not be recomputed for a new data set. The critical values of $\tilde{U}_{m,n}$ for the studentized Wilcoxon test are tabulated in Table 4.

2.3 On Asymptotic Power

In this section, we investigate the asymptotic power of the studentized two-sample Wilcoxon test against a sequence of contiguous alternatives and examine its efficiency in comparison to the standard two-sample Wilcoxon test. As will be shown below, there is no efficiency loss in using the studentized Wilcoxon test in comparison to the standard Wilcoxon test.

To begin, recall that the standard Wilcoxon test fails to control the asymptotic probability of a Type 1 error when the null hypothesis is $P(X \leq Y) = \frac{1}{2}$, in which case, the efficiency comparison between the studentized Wilcoxon test and the standard Wilcoxon test is meaningless. Thus, in this section we restrict our attention to the null hypothesis $P = Q$, where both the standard Wilcoxon test and the studentized Wilcoxon test attain exact control of a Type 1 error. For simplicity, assume the usual shift model (3), in which for a fixed P, Δ generates an one-dimensional sub-model.

Theorem 2.3. *Assume $P = Q$. Let $U_{m,n}$ and $\tilde{U}_{m,n}$ denote the two-sample Wilcoxon test and the studentized two-sample Wilcoxon test, defined in (2) and (8), respectively. Let t denote the usual two-sample t-test. Let $\mathrm{ARE}\,(\phi_1, \phi_2)$ denote the asymptotic relative efficiency of ϕ_1 with respect to ϕ_2. Then, $\mathrm{ARE}\left(\tilde{U}_{m,n}, U_{m,n}\right) = 1$ and $\mathrm{ARE}\left(\tilde{U}_{m,n}, t\right) = \mathrm{ARE}\,(U_{m,n}, t)$.*

Remark 2.3. If we make a further assumption that P^m and Q^n satisfy some smoothness condition, then we can obtain the exact form of the limiting distribution of $\sqrt{m}\,(U_{m,n} - \theta)$ under a sequence of contiguous alternatives $\Delta_m = \frac{h}{\sqrt{m}}$. More specifically, assume that the shift model (3) is quadratic mean differentiable. Then, under a sequence of contiguous alternatives $\Delta_m = \frac{h}{\sqrt{m}}$,

$$\sqrt{m}\left(\tilde{U}_{m,n}\right) \xrightarrow{d} N\left(\sigma_{12}, \frac{1}{12(1-p)}\right) ,$$

where $\sigma_{12} = h\mathrm{E}f_P(Y)$ for f_P denoting the density function of P.

3 General Two-Sample U-statistics

In this section, the results regarding the two-sample Wilcoxon statistic considered in Section 2 are extended to a general class of U-statistics. We provide a general framework whereby one can construct a test of parameter $\theta(P, Q) = \theta_0$ based on its corresponding U-statistic, which controls the asymptotic probability of a Type 1 error in large samples while retaining the exact control of a Type 1 error when $P = Q$.

To begin, assume $X_1, \ldots, X_m$ are i.i.d. P and, independently, $Y_1, \ldots, Y_n$ are i.i.d. Q. Let $Z = (Z_1, \ldots, Z_N) = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ with $N = m + n$. The problem is to test the null hypothesis

$$H_0 : \mathrm{E}_{P,Q}\big(\varphi\left(X_1, \ldots, X_r, Y_1, \ldots, Y_r\right)\big) = 0 \ ,$$

which can be estimated by its corresponding two-sample U-statistic of the form

$$U_{m,n}(Z) = \frac{1}{\binom{m}{r}\binom{n}{r}} \sum_{\alpha} \sum_{\beta} \varphi(X_{\alpha_1}, \ldots, X_{\alpha_r}, Y_{\beta_1}, \ldots, Y_{\beta_r}) \ ,$$

where α and β range over the sets of all unordered subsets of r different elements chosen from $\{1, \ldots, m\}$ and of r different elements chosen from $\{1, \ldots, n\}$, respectively. Without loss of generality, assume that φ is symmetric both in its first r arguments and in its last r arguments as a non-symmetric kernel can always be replaced by a symmetric one. Note that the continuity assumption of the underlying distributions in Section 2 is relaxed here and the kernel of U-statistics is also generalized to be of order r.

3.1 Coupling Argument

Our goal is to understand the limiting behavior of the U-statistic under permutations. We first employ what we call the coupling method, which will enable us to reduce the problem concerning the limiting behavior of the permutation distribution under samples from P and Q to the i.i.d. case where all N observations are i.i.d. according to the mixture distribution $\bar{P} = pP + qQ$, where $\frac{m}{N} \to p$ and $\frac{n}{N} \to q$ as $m, n \to \infty$. This reduction of the problem has two main advantages. First, it significantly simplifies calculations involving the limiting behavior of the permutation distribution since the behavior of the permutation distribution based on N i.i.d. observations is typically much easier to analyze than that based on possibly non-i.i.d. observations. Secondly, it provides an intuitive insight as to how the permuted sample asymptotically behave;

the permutation distribution under the original sample behaves approximately like the sampling distribution under N i.i.d. observations from the mixture distribution $\bar{P}$.

To be more specific, let $(\pi(1), \ldots, \pi(N))$ be a permutation of $\{1, \ldots, N\}$. Assume $\bar{Z}_1, \ldots, \bar{Z}_N$ are i.i.d. from the mixture distribution $\bar{P} = pP + qQ$, where $\frac{m}{N} \to p$ and $\frac{n}{N} \to q$ as $m, n \to \infty$ with

$$p - \frac{m}{N} = O(\frac{1}{\sqrt{N}}) \ . \tag{9}$$

We can think of the i.i.d. N observations from $\bar{P}$ as being generated via the following two-stage process: for $i = 1, \ldots, N$, first toss a weighted coin with probability p of coming up heads. If it is heads, sample an observation $\bar{Z}_i$ from P and otherwise from Q. The number of Xs in $\bar{Z}$, denoted B_m, then follows the binomial distribution with parameters N and p, i.e. $B_m \sim B(N, p)$ with mean $Np \approx m$ whereas the number of Xs in Z is exactly m. However, using a certain coupling argument which is described below, we can construct $\bar{Z}$ such that it has "most" of the observations matching those in Z. Then, if we can further show that the difference between the statistic evaluated at Z and also evaluated at $\bar{Z}$ is small in some sense (which we define below), then the limiting permutation distribution based on the original sample Z is the same as that based on the constructed sample $\bar{Z}$.

We shall now illustrate how to construct such a sample $\bar{Z}$ from the mixture distribution $\bar{P}$. First, toss a coin with probability p of coming up heads; if it heads, set $\bar{Z}_1 = X_1$, where X_1 is in Z. Otherwise if it is tails, set $\bar{Z}_1 = Y_1$. Next, if it shows up heads again, set $\bar{Z}_2 = X_2$ from Z; otherwise, if it's different from the first step, i.e., tails, set $\bar{Z}_2 = Y_1$ from Z. Continue constructing $\bar{Z}_i$ for $i = 1, \ldots, N$ from observations in Z according to the outcome of the flipped coin; if heads, use X_i from Z and if tails, use Y_j from Z. However, at some point, we will get stuck since either Xs or Ys have been exhausted from Z. For example, if all the Xs have been matched to $\bar{Z}$ and heads show up again, then sample a new observation from the underlying distribution P. Complete filling up $\bar{Z}$ in this manner. Then, we now have Z and $\bar{Z}$ that share many of common values except for those new observations added to $\bar{Z}$. Let D denote the number of observations that are different between Z and $\bar{Z}$, i.e., $D = |B_m - m|$. Note that the random number D is the number of new observations that are added to fill up $\bar{Z}$.

For example, suppose $m = n = 2$ and $Z = (X_1, X_2, Y_1, Y_2)$. If a coin is flipped four times and three heads are followed by one tail, then $\bar{Z} = (X_1, X_2, Y_1, X_3)$ where X_3 is an additional observation sampled from P. But, Z and $\bar{Z}$ share three elements.

Now, we can reorder the observations in $\bar{Z}$ by a permutation π_0 so that the original sample Z and the constructed sample $\bar{Z}$ will exactly match except for D observations that differ. First, recall that Z has observations that are in order; the first m observations that came from P (Xs) followed by n observations from Q (Ys). Thus, we will shuffle the observations in $\bar{Z}$ such that it is ordered in the (almost) same manner. We first put all the X observations in $\bar{Z}$ up to m slots, i.e., if the number of observations that are Xs in $\bar{Z}$ is greater than or equal to m, then $\bar{Z}_{\pi(i)} = X_i$ for $i = 1, \ldots, m$ and if the number is strictly greater than m, then put all the "left-over" Xs aside for now. On the other hand, if the number of observations that came from P in $\bar{Z}$ is smaller than m, then fill up as many Xs in $\bar{Z}$ as possible and leave the rest D slots in the first m entries empty for now. Next, from the $m+1$ up to Nth slot, fill them up with as many Y observations in $\bar{Z}$ as possible. Lastly, fill up empty slots with the "leftovers." Consequently, $\bar{Z}_{\pi_0}$ is either of the form

$$(\bar{Z}_{\pi_0(1)}, \ldots, \bar{Z}_{\pi_0(N)}) = (X_1, \ldots, X_m, Y_1, \ldots, Y_{n-D}, X_{m+1}, \ldots, X_{m+D})$$

if the number B_m of Xs in $\bar{Z}$ is greater than m; or it is of the form

$$(\bar{Z}_{\pi_0(1)}, \ldots, \bar{Z}_{\pi_0(N)}) = (X_1, \ldots, X_{m-D}, Y_{n+1}, \ldots, Y_{n+D}, Y_1, \ldots, Y_n)$$

if $B_m < m$.

Using this coupling method, if we can show (Lemma A.1) that, for any permutation $\pi = (\pi(1), \ldots, \pi(N))$ of $\{1, \ldots, N\}$,

$$\sqrt{m}U_{m,n}(Z_\pi) - \sqrt{m}U_{m,n}(\bar{Z}_{\pi \cdot \pi_0}) \xrightarrow{P} 0 \ .$$

Then, this condition will enable one to study the permutation distribution based on i.i.d. variables from the mixture distribution $\bar{P} = pP + qQ$ instead of having to dealing with observations that are no longer independent nor identically distributed.

3.2 Main Results

Theorem 3.1. *Consider the above set-up with the kernel φ is assumed antisymmetric across the first r and the last r arguments, i.e.,*

$$\varphi(X_{\alpha_1}, \ldots, X_{\alpha_r}, Y_{\beta_1}, \ldots, Y_{\beta_r}) = -\varphi(Y_{\beta_1}, \ldots, Y_{\beta_r}, X_{\alpha_1}, \ldots, X_{\alpha_r}) \ . \tag{10}$$

Assume $E_{P,Q}\varphi(\cdot) = 0$ and $0 < E_{P,Q}\varphi^2(\cdot) < \infty$ for any permutation of Xs and Ys. Let $m \to \infty$, $n \to \infty$, with $N = m + n$, $m/N \to p > 0$ and $n/N \to q > 0$. Then, the *permutation distribution of $\sqrt{m}U_{m,n}$ given by (1) with T replaced by U satisfies*

$$\sup_t |\hat{R}^U_{m,n}(t) - \Phi(t/\bar{\tau})| \overset{P}{\to} 0 \ ,$$

where

$$\bar{\tau}^2 = r^2 \left(\mathrm{E}\varphi^2_{\cdot\bar{P}^{r-1}\bar{P}^r}(\bar{Z}_i) + \frac{p}{1-p}\mathrm{E}\varphi^2_{\cdot\bar{P}^{r-1}\bar{P}^r}(\bar{Z}_i) \right) = \frac{r^2}{1-p}\mathrm{E}\varphi^2_{\cdot\bar{P}^{r-1}\bar{P}^r}(\bar{Z}_i) \ , \tag{11}$$

for

$$\varphi_{\cdot\bar{P}^{r-1}\bar{P}^r}(a_1) \equiv \int \cdots \int \varphi(a_1, \ldots, a_r, b_1, \ldots, b_r) d\bar{P}(a_2) \cdots d\bar{P}(a_r) d\bar{P}(b_1) \cdots d\bar{P}(b_r) \ .$$

Remark 3.1. Under $H_0 : \mathrm{E}_{P,Q}\varphi = 0$, the true unconditional sampling distribution of $U_{m,n}$ is asymptotically normal with mean 0 and variance

$$r^2 \left(\int \varphi^2_{\cdot P^{r-1}Q^r}(X_i) dP + \frac{p}{1-p} \int \varphi^2_{P^r \cdot Q^{r-1}}(Y_j) dQ \right) \ ,$$

which in general does not equal $\bar{\tau}^2$ defined in (11).

Remark 3.2. The antisymmetry assumption (10) is quite general. Many two-sample U-statistics can be modified such that this condition is satisfied. For example, by modifying the kernel function of the Wilcoxon statistic considered in Section 2, the results regarding the Wilcoxon test can be generalized to the case where the underlying distributions need not be continuous as shown in Example 3.1. However, the antisymmetry assumption is not just one of convenience because, without it, the results do not hold. As an example, consider the following U-statistic which does not satisfy the antisymmetry assumption. Assume $m = n$ and $P = Q$ with mean $\mu = 0$ and variance σ^2. Assume the statistic of interest is, although absurd, of the form $T_{m,m} = \sqrt{m}[\bar{X}_m + \bar{Y}_m]$, i.e., the *sum* of sample means. The true unconditional sampling distribution is asymptotically normal with mean 0 and variance $2\sigma^2$ whereas the permutation distribution is a point mass function at $\sqrt{m}[\bar{X}_m + \bar{Y}_m]$ as the statistic $T_{m,n}$ is invariant under permutations. At the same time, without such a condition the statistic would not enable any kind of comparison between P and Q anyway.

Example 3.1. (Two-sample Wilcoxon test without continuity assumption) The null hypothesis of interest is $H_0 : P(X \leq Y) = P(Y \leq X)$. The corresponding U-statistic that takes into account the possibility of having ties is

$$U_{m,n,1} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\mathrm{I}(X_i < Y_j) + \frac{1}{2}\mathrm{I}(X_i = Y_j) - \frac{1}{2} \right) \ .$$

Note that its kernel

$$\varphi_1 = \mathrm{I}(X_i < Y_j) + \frac{1}{2}\mathrm{I}(X_i = Y_j) - \frac{1}{2}$$

satisfies the antisymmetric assumption. The permutation distribution of $\sqrt{m}U_{m,n,1}$ is approximately a normal distribution with mean 0 and variance $\frac{1}{12(1-p)}$ whereas the true sampling limiting distribution is normal with mean 0 and variance

$$\xi'_x + \frac{p}{1-p}\xi'_y \ ,$$

where $\xi'_x = \mathrm{E}\varphi^2_{\cdot Q} dP = \mathrm{Var}(Q^-_Y(X_i) + \frac{1}{2}f_Q(X_i))$ and $\xi'_y = \mathrm{E}\varphi^2_{\cdot P} dQ = \mathrm{Var}(P^-_X(Y_j) + \frac{1}{2}f_P(Y_j))$ for f_Q and f_P denoting the density function of Q and P, respectively. Hence, the permutation distribution and the true unconditional sampling distribution behave differently asymptotically unless $(1-p)\xi'_x + p\xi'_y = \frac{1}{12}$ is satisfied.

Example 3.2. (Two-sample U-statistic by Lehmann (1951)) The null hypothesis of interest is $H_0 : P(|Y' - Y| > |X' - X|) = \frac{1}{2}$. The corresponding U-statistic is

$$U_{m,n,2} = \frac{1}{\binom{m}{2}\binom{n}{2}} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} \left(\mathrm{I}(|Y_l - Y_k| > |X_j - X_i|) - \frac{1}{2} \right)$$

with its antisymmetric kernel $\varphi_2 = \mathrm{I}(|Y_l - Y_k| > |X_j - X_i|) - \frac{1}{2}$.

Example 3.3. (Two-sample U-statistic by Hollander (1967)) The null hypothesis of interest is $H_0 : P(X + X' < Y + Y') = \frac{1}{2}$. The corresponding U-statistic is

$$U_{m,n,3} = \frac{1}{\binom{m}{2}\binom{n}{2}} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} \left(\mathrm{I}(X_i + X_j < Y_k + Y_l) - \frac{1}{2} \right)$$

with its kernel $\varphi_3 = \mathrm{I}(X_i + X_j < Y_k + Y_l) - \frac{1}{2}$, which is antisymmetric.

Example 3.4. (Comparing Variances) This problem has been addressed by Pauly (2010). A similar argument but in the framework above can also be applied to this problem. The null hypothesis of interest is $H_0 : \sigma_X^2 = \sigma_Y^2 = 0$. The corresponding U-statistic is

$$U_{m,n,4} = \frac{1}{\binom{m}{2}\binom{n}{2}} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} \left(\frac{1}{2}(X_i - X_j)^2 - \frac{1}{2}(Y_k - Y_l)^2 \right) ,$$

where the kernel $\varphi_4 = \frac{1}{2}(X_i - X_j)^2 - \frac{1}{2}(Y_k - Y_l)^2$ satisfies the antisymmetric assumption. The true unconditional sampling distribution is approximately normal with mean 0 and variance

$$\operatorname{Var}\left\{(X - \mu_X)^2\right\} + \frac{p}{1-p} \operatorname{Var}\left\{(Y - \mu_Y)^2\right\}$$

whereas the permutation distribution is asymptotically normal with mean 0 and variance

$$\frac{1}{1-p} \operatorname{Var}\left\{(X - \mu_X)^2\right\} + \operatorname{Var}\left\{(Y - \mu_Y)^2\right\} .$$

The following theorem shows how studentization leads to asymptotic validity.

Theorem 3.2. *Assume the same setup and conditions of Theorem 3.1. Further assume that $\hat{\sigma}_m^2(X_1, \ldots, X_m)$ is a consistent estimator of $\int \varphi^2_{.P^{r-1}Q^r} dP$ when $X_1, \ldots, X_m$ are i.i.d. P and that $\hat{\sigma}_n^2(Y_1, \ldots, Y_n)$ is a consistent estimator of $\int \varphi^2_{P^r.Q^{r-1}} dQ$ when $Y_1, \ldots, Y_n$ are i.i.d. Q. Assume consistency also under the mixture distribution $\bar{P}$, i.e., $\hat{\sigma}_m^2(\bar{Z}_1, \ldots, \bar{Z}_m)$ is a consistent estimator of $\int \varphi^2_{.\bar{P}^{r-1}\bar{P}^r} d\bar{P}$ when $\bar{Z}_1, \ldots, \bar{Z}_m$ are i.i.d. $\bar{P}$. Define the studentized U-statistic*

$$S_{m,n} = \frac{U_{m,n}}{V_{m,n}} ,$$

where

$$V_{m,n} = r\sqrt{\hat{\sigma}_m^2(X_1, \ldots, X_m) + \frac{m}{n}\hat{\sigma}_n^2(Y_1, \ldots, Y_n)} .$$

Then, the permutation distribution $\hat{R}^S_{m,n}(\cdot)$ of $\sqrt{m}S_{m,n}$ given by (1) with T replaced by S satisfies

$$\sup_t |\hat{R}^S_{m,n}(t) - \Phi(t)| \xrightarrow{P} 0 . \tag{12}$$

Example 3.1. (continued) Define the studentized Wilcoxon statistic

$$\tilde{S}_{m,n,1} = \frac{U_{m,n,1}}{\sqrt{\hat{\xi}'_x + \frac{m}{n}\hat{\xi}'_y}} ,$$

where

$$\hat{\xi}'_x = \frac{1}{m-1}\sum_{i=1}^{m}\left\{\hat{\zeta}_{x,1}(X_i) - \frac{1}{m}\sum_{i=1}^{m}\hat{\zeta}_{x,1}(X_i)\right\}^2 \text{ and } \hat{\xi}'_y = \frac{1}{n-1}\sum_{j=1}^{n}\left\{\hat{\zeta}_{y,1}(Y_j) - \frac{1}{n}\sum_{j=1}^{n}\hat{\zeta}_{y,1}(Y_j)\right\}^2,$$

for

$$\hat{\zeta}_{x,1}(X_i) \equiv \frac{1}{n}\sum_{j=1}^{n}\left(\mathrm{I}\{Y_j < X_i\} + \frac{1}{2}\mathrm{I}\{Y_j = X_i\}\right)$$

and

$$\hat{\zeta}_{y,1}(Y_j) \equiv \frac{1}{m}\sum_{i=1}^{m}\left(\mathrm{I}\{X_i < Y_j\} + \frac{1}{2}\mathrm{I}\{X_i = Y_j\}\right) .$$

Then, by Theorem 3.2, both the permutation distribution and the true unconditional sampling distribution are approximately standard normal.

Example 3.2. (continued) The variance of the true sampling distribution of $\sqrt{m}U_{m,n,2}$ can be estimated by

$$V^2_{m,n,2} \equiv 4\left[\frac{1}{m-1}\sum_{i=1}^{m-1}\left\{\hat{\zeta}_{x,2}(X_i) - \frac{1}{m-1}\sum_{i=1}^{m-1}\hat{\zeta}_{x,2}(X_i)\right\}^2 + \frac{m}{n}\frac{1}{n-1}\sum_{k=1}^{n-1}\left\{\hat{\zeta}_{y,2}(Y_k) - \frac{1}{n-1}\sum_{k=1}^{n-1}\hat{\zeta}_{y,2}(Y_k)\right\}^2\right],$$

where

$$\hat{\zeta}_{x,2}(X_i) = \sum_{j=i+1}^{m}\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}\mathrm{I}(|Y_k - Y_l| > |X_i - X_j|)$$

and

$$\hat{\zeta}_{y,2}(Y_k) = \sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\sum_{l=k+1}^{n}\mathrm{I}(|Y_k - Y_l| > |X_i - X_j|) .$$

Example 3.3. (continued) The variance of the true sampling distribution of $\sqrt{m}U_{m,n,3}$ can be estimated by

$$V^2_{m,n,3} \equiv 4\left[\frac{1}{m-1}\sum_{i=1}^{m-1}\left\{\hat{\zeta}_{x,3}(X_i) - \frac{1}{m-1}\sum_{i=1}^{m-1}\hat{\zeta}_{x,3}(X_i)\right\}^2 + \frac{m}{n}\frac{1}{n-1}\sum_{k=1}^{n-1}\left\{\hat{\zeta}_{y,3}(Y_k) - \frac{1}{n-1}\sum_{k=1}^{n-1}\hat{\zeta}_{y,3}(Y_k)\right\}^2\right],$$

where

$$\hat{\zeta}_{x,3}(X_i) = \sum_{j=i+1}^{m}\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}\mathrm{I}(Y_k + Y_l - X_j \le X_i)$$

and

$$\hat{\zeta}_{y,3}(Y_k) = \sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\sum_{l=k+1}^{n}\mathrm{I}(X_i + X_j - Y_l < Y_k) .$$

Example 3.4. (continued) The corresponding studentized U-statistic can be defined by

$$\tilde{S}_{m,n,4} = \frac{U_{m,n,4}}{\sqrt{V_{m,n,4}}} \ ,$$

where

$$V^2_{m,n,4} \equiv \frac{1}{m-1}\sum_{i=1}^{m}\left\{(X_i-\bar{X})^2 - \frac{1}{m}\sum_{i=1}^{m}(X_i-\bar{X})^2\right\}^2 + \frac{m}{n}\frac{1}{n-1}\sum_{j=1}^{n}\left\{(Y_j-\bar{Y})^2 - \frac{1}{n}\sum_{j=1}^{n}(Y_j-\bar{Y})^2\right\}^2 \ .$$

4 Simulation Results

Monte Carlo simulation studies illustrating our results regarding the two-sample Wilcoxon test are presented in this section. Summarized in Table 1 are the rejection probabilities of two-sided tests for the studentized Wilcoxon test under the null hypothesis where the nominal level is $\alpha = 0.05$. The simulation results confirm that the studentized two-sample Wilcoxon test is valid in the sense that it controls the asymptotic probability of a Type 1 error in large samples.

The underlying distributions that are considered in the Monte Carlo simulations are presented in the first column of Table 1. All the distributions studied are chosen in such a way that $\theta(P, Q) = P_{P,Q}(X \leq Y) = \frac{1}{2}$ is satisfied despite $P \neq Q$. As displayed in Table 1 below, when the underlying distributions of the two independent samples are not identical, the standard Wilcoxon test fails to control the asymptotic probability of a Type 1 error. For example, when P is a normal distribution with mean 0 and variance 5 and Q is a Student's t-distribution with 3 degrees of freedom, despite $P(X \leq Y) = \frac{1}{2}$, the rejection probabilities of the usual Wilcoxon test for the sample sizes considered range between 0.0723 and 0.1213, which are far larger than the nominal level $\alpha = 0.05$. With the increased rejection probability, rejection of the test may be erroneously construed as rejection of $P(X \leq Y) = \frac{1}{2}$, when the rejection is actually caused by the inequality of distributions. Furthermore, in the case P = gamma(1,2) and Q = gamma(0.63093, 4), where gamma(α, β) is defined as in Casella and Berger (2001), the rejection probability can be much less than the nominal level $\alpha = 0.05$. This implies that by continuity, the probability of rejection under some alternatives can be less than the level of the test. Thus, the test is biased and its power of detecting the true probability of $P(X \leq Y)$ can be very small. As displayed in the 'Not Studentized' sections, the rejection probabilities under some P and Q satisfying $\theta(P, Q) = \frac{1}{2}$ do not tend to the nominal level $\alpha = 0.05$ asymptotically.

Distributions	m	4	12	50	100
	n	4	18	100	100

N(0,1)
N(0,5)
	Not Studentized	0.0864	0.0496	0.0367	0.0835

	Asymptotic	0.0678	0.0644	0.0516	0.0527
	Studentized	0.0678	0.0477	0.0512	0.0505

N(0,1)
T(3)
	Not Studentized	0.0799	0.0992	0.1213	0.0723

	Asymptotic	0.0674	0.0767	0.0567	0.0501
	Studentized	0.0674	0.0599	0.0570	0.0486

T(3)
T(10)
	Not Studentized	0.0493	0.0473	0.0524	0.0506

	Asymptotic	0.0878	0.0637	0.0477	0.0524
	Studentized	0.0878	0.0477	0.0480	0.0500

N(0.993875, 5)
Exp(0.993875)
	Not Studentized	0.0971	0.0513	0.0421	0.0804

	Asymptotic	0.0593	0.0654	0.0534	0.0509
	Studentized	0.0593	0.0475	0.0538	0.0486

Beta(10,10)
U(0,1)
	Not Studentized	0.0810	0.0474	0.0353	0.0719

	Asymptotic	0.0730	0.0660	0.0515	0.0514
	Studentized	0.0890	0.0499	0.0516	0.0480

Gamma(1,2)
Gamma(0.63093, 4)
	Not Studentized	0.0495	0.0400	0.0432	0.0548

	Asymptotic	0.0825	0.0618	0.0532	0.0550
	Studentized	0.0825	0.0467	0.0535	0.0517

Cauchy(0,1)
N(0,5)
	Not Studentized	0.0594	0.0436	0.0406	0.0627

	Asymptotic	0.0708	0.0625	0.0516	0.0534
	Studentized	0.0708	0.0488	0.0505	0.0514

Table 1: Monte-Carlo Simulation Results for Studentized Two-Sample Wilcoxon Test (Two-sided, $\alpha = 0.05$)

As explained earlier, the failure of the standard Wilcoxon test to control the asymptotic probability of a Type 1 error is due to the fact that the true sampling distribution and the permutation distribution behave differently. For each pair of sample distributions considered in this study, the limiting variance (4) of the permutation distribution (with p replaced by $\frac{m}{N}$) is tabled below and is compared with the limiting variance (5) of the unconditional sampling distribution (with p replaced by $\frac{m}{N}$). As shown in Table 2, the limiting variances of (4) and (5) are generally not the same. For instance, the limiting variance of the permutation distribution with samples of equal size under $P = N(0,1)$ and $Q = N(0,5)$ is $\tau^2 = 0.3333$ whereas the limiting variance of the unconditional true sampling distribution is $\sigma^2 = 0.4237$. Hence, for testing the null hypothesis $\theta = P(X \leq Y) = \frac{1}{2}$, the critical values of the permutation test, which converges to $z_{1-\alpha}\tau$ in probability, are not valid because what is required instead is that the critical values tend to $z_{1-\alpha}\sigma$ in probability.

Distributions	m=4 & n=4		m=12 & n=18		m=50 & n=100	
	Perm.(τ^2)	Sample(σ^2)	Perm.(τ^2)	Sample(σ^2)	Perm.(τ^2)	Sample(σ^2)
N(0,1) & N(0,5)	0.3333	0.4237	0.3472	0.3582	0.375	0.3269
N(0,5) & T(3)	0.3333	0.4040	0.3472	0.4943	0.375	0.5867
T(3) & T(10)	0.3333	0.3342	0.3472	0.3577	0.375	0.3933
N(0.993875, 5) & Exp(0.993875)	0.3333	0.4336	0.3472	0.5371	0.375	0.6416
Beta(10,10)& U(0,1)	0.3333	0.3999	0.3472	0.3433	0.375	0.3179
Gamma(1,2) & Gamma(0.63093, 4)	0.3333	0.3389	0.3472	0.3301	0.375	0.3400
Cauchy(0,1) & N(0,5)	0.3333	0.3686	0.3472	0.3397	0.375	0.3351

Table 2: Variance Discrepancy Between the Limiting Permutation Distribution (23) and the Limiting True Sampling Distribution (22)

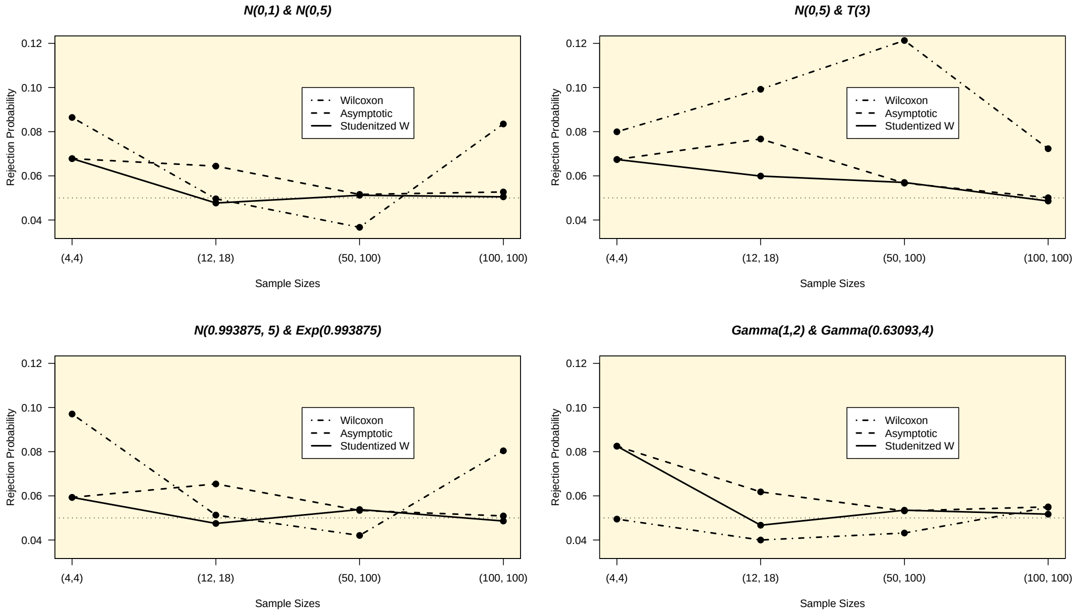

Figure 3: Probability Rejection Comparison (Two-sided, $\alpha = 0.05$)

Results of the usual asymptotic approach using the normal approximation are also presented in the 'Asymptotic' sections of Table 1. As the sample sizes increase, the rejection probability better approximates $\alpha = 0.05$. The Monte Carlo simulation studies also confirm that the studentized two-sample Wilcoxon test attains asymptotic rejection probability α in large samples. In the 'Studentized' sections of Table 1, the rejection probability under some P and Q satisfying the null hypothesis $P(X \leq Y) = \frac{1}{2}$ tends to the nominal level $\alpha = 0.05$ asymptotically in large samples. Note that both the asymptotic approach and the studentized version of the Wilcoxon test are asymptotic results.

When the sample sizes are small such as m = n = 4, the rejection probability is not close to α, as expected. However, as displayed in Figure 3, the rejection probability seems to converge to the nominal level α much more rapidly with the studentized Wilcoxon test than with the asymptotic approach. Under the studentized two-sample Wilcoxon test, the rejection probability quickly tends to the nominal level α even with relatively small sample sizes, like 12 and 18. Unlike the asymptotic approach, the studentized Wilcoxon test retains the exact rejection probability of α in the case $P = Q$.

5 Empirical Applications

In this section, we present empirical applications of the studentized Wilcoxon test, employing the experimental data used in Dohmen and Falk (2011) and Plott and Zeiler (2005). Dohmen and Falk conduct a laboratory experiment to study how individual characteristics affect an individual's self-selection decision between fixed- and variable-payment schemes. They consider three variable-payment schemes (piece rate, tournament, and revenue sharing) and find that individuals who self-select into the variable-payment schemes get more answers correct than those who self-select into the fixed-payment scheme (see Dohmen and Falk (2011) for more details).

Under these settings, one can also test whether the median of the differences between fixed- and variable-payment schemes equal zero. To be more precise, let X denote the number of correct answers under the fixed-payment scheme and let Y represent the number of correct answers under the variable-payment schemes. The null hypothesis of interest is

$$P(X \leq Y) = P(Y \leq X) \ ,$$

which is the same as testing $P(X \leq Y) = 1/2$ in the case of continuous distributions. The studentized Wilcoxon test for testing $P(X \leq Y) = P(Y \leq X)$ yields studentized Wilcoxon statistics of 9.874385, 3.380504, and 3.326384 (all three p-values < 0.0001 against one-sided alternatives) in the treatment of piece rates, tournament, and revenue sharing, respectively, indicating that the median of the differences between fixed- and variable-payment schemes is not zero.

A similar analysis can be applied to the experimental data used in Plott and Zeiler (2005). Plott and Zeiler study whether the observed WTP-WTA gap, a tendency for an individual to state that a minimum amount of values the individual is willing to accept in order to give up an item (WTA) is greater than the maximum amount of values that the same individual would pay in exchange for the same item (WTP), can be attributed to loss aversion, the notion of a fundamental feature of human preferences in which gains are valued less than losses. We do not attempt to provide an alternative solution to answer their primary question, but rather to present what else can be answered and how it should be tested.

In their experiment, subjects were divided into two groups - buyers and sellers. Each subject was given a mug, and WTP for buyers and WTA for sellers were reported (see Plott and Zeiler (2005) for more details). WTP in experiment 3 consists of $\{2.50, 5.85, 6, 7.50, 8, 8.50, 8.50, 8.78, 10\}$ with sample size 9 and median 8 and WTA is composed of $\{3, 3, 3.50, 3.50, 5, 5, 7.50, 10\}$ with sample size 8 and median 4.25. The estimated densities using kernel density estimates for WTP and WTA, denoted P and Q, respectively, are plotted in Figure 4.

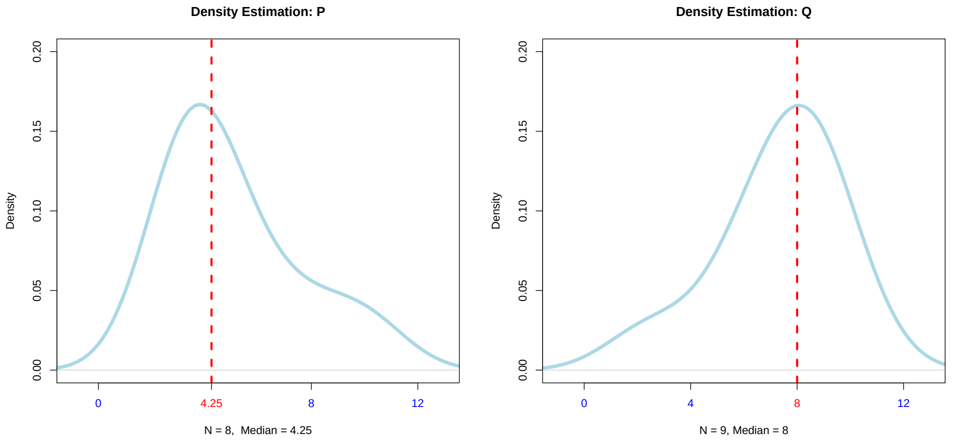

Figure 4: Experiment 3: WTP (P) & WTA (Q) (Plott and Zeiler, 2005)

One valid and interesting question is whether the probability of WTA tends to be greater than WTP is the same as that of WTP tends to be greater than WTA, i.e., $\mathrm{P}(WTP \leq WTA) = \mathrm{P}(WTA \leq WTP)$, which is equivalent to testing $\mathrm{P}(WTP \leq WTA) = \frac{1}{2}$ in the case of continuous distributions. Testing this parameter is in fact equivalent to testing whether the median of the difference between WTA and WTP is zero. The studentized Wilcoxon test yields a studentized Wilcoxon statistic of 1.5435 (p-value of 0.0588 against one-sided alternatives), which barely fails to reject the null hypothesis that the probability of WTA tends to be greater than WTP is one half at level $\alpha = 0.05$.

As pointed out earlier, the Wilcoxon test has been often misused for testing equality of medians. We discuss other tests that are often presented as alternatives for testing equality of medians and compare them with the studentized permutation test based on the difference in sample medians (Chung and Romano, 2011). We compare the performance and the assumptions required for Mood's median test, Wilcoxon test, robust rank-order test, and studentized median permutation test using the experimental data from Plott and Zeiler (2005). We provide evidence that for testing equality of medians, using the studentized median permutation test is most advisable especially when imposing additional assumptions on the underlying distributions is undesirable.

In their paper, Plott and Zeiler used the Mood's median test for testing equality of medians. The median test is based on a Pearson's chi-square statistic and tests the hypothesis that the probability of an observation being greater than the overall median is the same for all populations. The test yields a Pearson χ^2 test statistic of 1.5159 (p-value = 0.218), leading Plott and Zeiler to fail to reject the null hypothesis that WTP and WTA share the same median. However, as is well-known, the Mood's median test is limited in its power to detect median differences. The reason for lower efficiency is that the difference of the medians may not be reflected by the proportion of observations that are greater (smaller) than the overall median, especially when the sample sizes are small, causing the researchers to fail to reject the null hypothesis. Thus, a more suitable test is desired for testing equality of medians especially when the sample sizes are small, like 9 and 8 as our example.

As an alternative, the Wilcoxon test is often used for testing equality of medians. Feri, Irlenbusch, and Sutter (2010) and Waldfogel (2005) are some examples among many that use the Wilcoxon test for testing equality of medians. The Wilcoxon test yields a p-value of 0.0821, again resulting in a failure to reject the null. However, the Wilcoxon test is only valid for testing equality of medians when the two underlying distributions are identical under the null. If such an assumption does not hold, however, it fails to control the probability of a Type I error, even asymptotically. Furthermore, it can severely affect the power of the test. The problem is that the Wilcoxon test only picks up divergence from $P(WTP \leq WTA) = \frac{1}{2}$. Thus, the Wilcoxon test can fail to reject the null just because $P(WTP \leq WTA) = \frac{1}{2}$ holds, not because medians are identical.

Fligner and Policello (1981) proposed the robust rank-order test as an another alternative for testing equality of medians. As Feltovich (2003) pointed out, the robust rank-order test tends to outperform the Wilcoxon test under more general settings as the robust rank-order test only requires the underlying distributions to be symmetric. Of course, this assumption is weaker than having identical distributions. However, as Fligner and Policello pointed out, if the underlying distributions are not symmetric, then unfortunately, the performance of the robust-rank-order test may not be satisfactory. In our example, the robust rank-order test fails to reject the null at the 5% level (see Table 3), but this result may be caused by the asymmetry of the underlying distributions P and Q as seen in Figure 4 rather than not having significantly different medians.

	Median Test (χ^2)	Wilcoxon Test (z) †	Robust Rank-Order Test†	Studentized Median Permutation Test
Statistic Value	1.5159	1.7385	1.7066	3.1499
p-value	0.2182	0.0821	> 0.05	0.0290

† Adjusted for having ties

Table 3: Tests Used For testing Equality of Medians

To test equality of medians, we propose using the studentized median permutation test. Unlike the Wilcoxon test and the robust rank-order test, the studentized median permutation test does not require any assumptions on the underlying distributions. As long as there exists a consistent estimator for the standard error, it controls the asymptotic probability of Type I error in large samples while retaining the exact control of the probability of Type I error when the underlying distributions are identical. A consistent estimator can be obtained by the usual kernel estimator (Devroye and Wagner, 1980), bootstrap estimator (Efron, 1979), or the smoothed bootstrap (Hall, DiCiccio, and Romano, 1989). The permutation distribution for the example considered using the kernel estimator for the standard errors is plotted in Figure 5. Since the sample statistic of the studentized median difference (3.1499) falls within the 5% range of the tail of the permutation distribution (p-value of 0.0290), we *reject* the null that the underlying distributions have the same median, which is contrary to all the other tests considered earlier.

6 Conclusion

Although permutation tests are useful tools in obtaining exact level α in finite samples under the fundamental assumption of identical underlying distributions, it lacks robustness of validity against inequality of distributions. If the underlying distributions of the two independent samples are not identical, the usual permutation test can fail to control the probability of a Type 1 error, even asymptotically. Thus, a careful interpretation of a rejection of the permutation test is necessary; rejection of the test does not necessarily imply the rejection of the null hypothesis that some real-valued parameter $\theta(F, G)$ is some specified value θ_0. Thus, one needs to clarify both what is being tested and what the implicit underlying assumptions are. We provide a general theory whereby one can construct a test of a parameter $\theta(P, Q) = \theta_0$ based on its corresponding U-statistic, which controls the Type 1 error in large samples. Moreover, it also retains the exactness property of the permutation test when the underlying distributions are identical, a

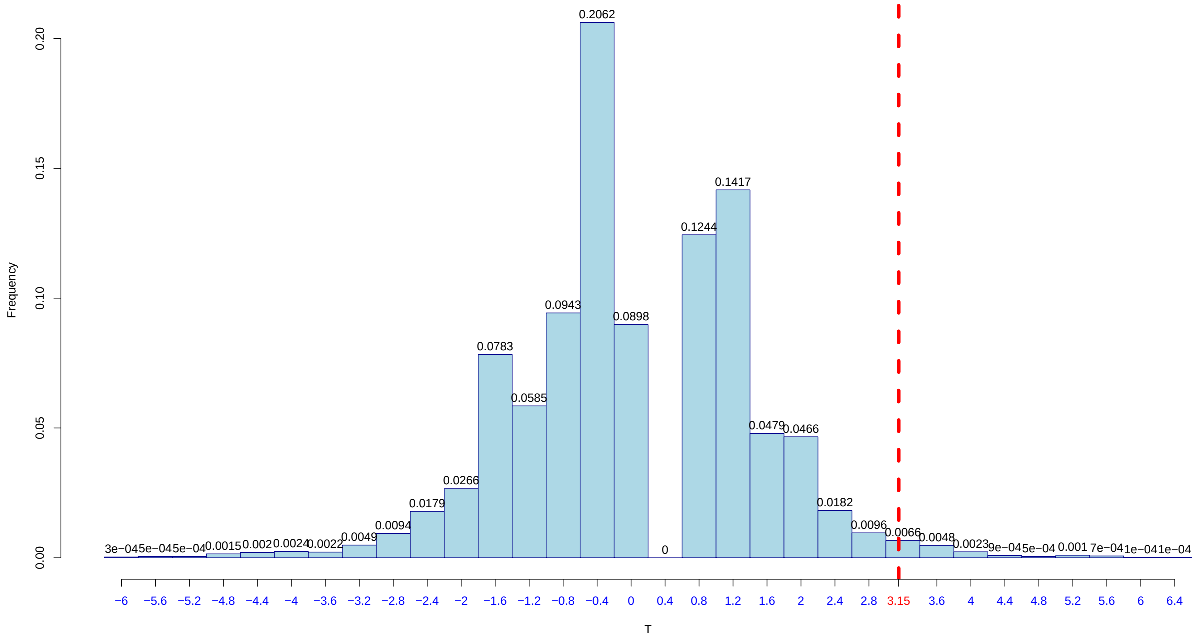

Figure 5: Permutation Distribution of the Studentized Median Statistic

desirable property that other resampling methods do not possess.

For example, in the case of the Wilcoxon test, we have constructed a new test that retains the exact control of the probability of a Type 1 error when the underlying distributions are identical while also achieving asymptotic validity of the test for testing $\mathrm{P}(X < Y) = \frac{1}{2}$. Moreover, when the underlying distributions are assumed continuous, the new test is also a rank test, so that the critical values of the new table can tabled as displayed in Table 4. Also, it achieves the same asymptotic relative efficiency compared to the two-sample t-test as the usual Wilcoxon test.

When testing $\theta(P, Q) = \theta_0$, as long as its corresponding U-statistic is studentized by a consistent standard error, the permutation test based on the studentized U-statistic controls the asymptotic probability of a Type 1 error in large samples while enjoying the exact control of the rejection probability when the underlying distributions are identical. This result is applicable for any test that is based on a two-sample U-statistic that satisfies the antisymmetry condition.

Table 4: Critical Values of the Studentized Wilcoxon Test, c ($P(\tilde{U}_{m,n} \leq c)$

m								
3	4	5	6	7	8	9	10	n
2.4749 (.0500)	3.5355 (.02857)	4.5962 (.01786)	5.6568 (.01190)	6.7175 (.00833)	7.7782 (.00606)	5.1430 (.00912)	5.8138 (.00701)	3
1.1180 (.1500)	1.7889 (.05714)	2.4597 (.03571)	2.2361 (.04762)	3.8013 (.01667)	4.4721 (.01212)	4.7958 (.01368)	5.4083 (.01054)	
0.6124 (.2000)	1.7321 (.08571)	2.3452 (.05357)	2.000 (.05952)	2.2613 (.05000)	2.4279 (.04848)	2.4042 (.4549)	2.3590 (.04897)	
0.5000 (.3500)	1.1619 (.11429)	1.5000 (.08929)	1.5504 (.09524)	1.5321 (.10000)	2.3518 (.05455)	2.3351 (.05005)	2.3570 (.05247)	
	0.6547 (.22857)	1.3887 (.10714)	1.5076 (.10714)	0.6124 (.25000)	1.5275 (.09697)	1.5765 (.09999)	1.6000 (.09782)	
	0.6433 (.25714)	0.6350 (.25000)	0.6202 (.25000)		1.4314 (.10909)	1.5428 (.10452)	1.5757 (.10132)	
					0.6831 (.24242)	0.6822 (.24549)	0.6566 (.24840)	
					0.6481 (.25455)	0.6585 (.25004)	0.6498 (.25192)	
	4.9497 (.01429)	6.3640 (.00794)	4.3301 (.00952)	5.0410 (.00909)	4.5691 (.00809)	4.0556 (.00983)	4.0000 (.00996)	4
	2.5981 (.04286)	3.4641 (.01587)	4.2258 (.01429)	3.8891 (.01212)	4.1110 (.01011)	3.9865 (.01123)	3.8307 (.01096)	
	1.8464 (.05714)	2.1602 (.04762)	2.0000 (.04762)	2.1019 (.04848)	2.1114 (.04839)	2.0448 (.04899)	2.0871 (.04995)	
	1.5811 (.08571)	1.7955 (.05556)	1.9069 (.05238)	1.9403 (.00515)	1.9797 (.05040)	2.0272 (.05041)	2.0706 (.05094)	
	1.2247 (.12857)	1.3525 (.09524)	1.3728 (.10000)	1.3555 (.10000)	1.3830 (.09889)	1.4347 (.09929)	1.4049 (.09989)	
	0.5477 (.24286)	1.3443 (.10317)	0.6283 (.24762)	0.6325 (.24848)	1.3728 (.10092)	1.4275 (.10070)	1.3887 (.10089)	
	0.5000 (.34286)	0.6626 (.24603)	0.6246 (.25238)	0.6298 (.25152)	0.6565 (.24844)	0.6690 (.24605)	0.6478 (.24986)	
		0.6547 (.25497)		0.6547 (.25046)	0.6576 (.25021)	0.6462 (.25085)		
		8.1317 (.00397)	4.1906 (.00866)	3.8184 (.00882)	3.7170 (.00926)	3.5380 (.00999)	3.5841 (.00996)	5
		4.4772 (.01190)	3.7033 (.01082)	3.7932 (.01009)	3.6927 (.01004)	3.5000 (.01048)	3.5660 (.01029)	
		2.0616 (.04365)	1.9781 (.04978)	1.8906 (.04926)	1.9322 (.04966)	1.8842 (.04988)	1.9160 (.04961)	
		1.9365 (.05556)	1.9734 (.05195)	1.8572 (.05053)	1.9211 (.05043)	1.8779 (.05038)	1.9156 (.05028)	
		1.4720 (.09921)	1.3522 (.09957)	1.3122 (.09968)	1.3840 (.09939)	1.3571 (.09990)	1.3612 (.09971)	

Continued on Next Page. . .

Table 4 Critical Values of the Studentized Wilcoxon Test – Continued

m								
3	4	5	6	7	8	9	10	n
		1.3132 (.10714)	1.3422 (.10173)	1.3089 (.10220)	1.3579 (.10015)	1.3537 (.10040)	1.3587 (.10037)	5
		0.7071 (.22222)	0.6840 (.24892)	0.6758 (.24886)	0.6598 (.24953)	0.6437 (.24909)	0.6470 (.24988)	
		0.6736 (.25794)	0.6612 (.25108)	0.6576 (.25138)	0.6586 (.25031)	0.6426 (.25059)	0.6455 (.25088)	
			3.9131 (.00971)	3.3597 (.00993)	3.3712 (.01000)	3.2466 (.00998)	3.2099 (.00990)	6
			3.3371 (.01189)	3.1884 (.01052)	1.8067 (.05000)	3.2395 (.01018)	3.2066 (.01016)	
			1.9174 (.04977)	1.8561 (.04962)	1.3363 (.09989)	1.8391 (.04974)	1.8353 (.04990)	
			1.8436 (.05191)	1.8216 (.05020)	1.3356 (.10022)	1.8380 (.05014)	1.8341 (.05015)	
			1.4142 (.08980)	1.3251 (.09975)	0.6304 (.24964)	1.3444 (.09982)	1.3253 (.09995)	
			1.3640 (.10278)	1.3241 (.10152)	0.6298 (.25032)	1.3429 (.10002)	1.3241 (.10008)	
			0.6202 (.24656)	0.6611 (.24999)		0.6636 (.24988)	0.6521 (.24966)	
			0.6030 (.27258)	0.6588 (.25059)		0.6630 (.25028)	0.6516 (.25042)	
				3.3072 (.00987)	3.1399 (.00993)	3.0645 (.00993)	3.0735 (.00997)	7
				3.1774 (.01103)	3.1196 (.01009)	3.0616 (.01002)	3.0674 (.01003)	
				1.8544 (.04920)	1.8212 (.04989)	1.8183 (.04996)	1.7897 (.04997)	
				1.8148 (.05005)	1.8161 (.05004)	1.8179 (.05005)	1.7889 (.05002)	
				1.3481 (.09903)	1.3192 (.09992)	1.3134 (.09992)	1.3253 (.09986)	
				1.3229 (.10251)	1.3188 (.10054)	1.3128 (.10001)	1.3248 (.10002)	
				0.6753 (.24353)	0.6563 (.24991)	0.6594 (.24994)	0.6598 (.24999)	
				0.6637 (.25022)	0.6561 (.25052)	0.6591 (.25002)	0.6595 (.25004)	
					3.0678 (.00952)	2.9916 (.01000)	2.95097 (.00998)	8
					3.0402 (.01044)	1.7966 (.04993)	2.95084 (.01001)	
					1.7889 (.04978)	1.7962 (.05001)	1.77702 (.04999)	

Continued on Next Page...

Table 4 Critical Values of the Studentized Wilcoxon Test – Continued

m								
3	4	5	6	7	8	9	10	n
					1.7856 (.05096)	1.3091 (.09991)	1.77655 (.05001)	8
					1.3229 (.09975)	1.3089 (.10012)	1.30922 (.09996)	
					1.3132 (.10068)	0.6518 (.24997)	1.30921 (.10002)	
					0.6692 (.24980)	0.6517 (.25001)	0.66652 (.24996)	
					0.6623 (.25121)		0.66623 (.25001)	
						2.9354 (.00985)	2.8642 (.01000)	9
						2.9155 (.01006)	1.7705 (.04999)	
						1.7761 (.04996)	1.7703 (.05004)	
						1.7579 (.05161)	1.3146 (.09999)	
						1.3171 (.09834)	1.3145 (.10003)	
						1.3168 (.10140)	0.6415 (.25000)	
						0.6407 (.24967)		
						0.6402 (.25409)		
							2.8180 (.00997)	10
							2.8108 (.01013)	
							1.7564 (.04980)	
							1.7487 (.05037)	
							1.3093 (.09985)	
							1.3073 (.10137)	
							0.6631 (.24668)	
							0.6591 (.25016)	

A Two Useful Lemmas

The following two lemmas hold for general two-sample U-statistics with a kernel of orders r and s. Note that the first lemma relies on the coupling arguments in Section 3.1.

Lemma A.1. *Assume $X_1, \ldots, X_m$ are i.i.d. P and, independently, $Y_1, \ldots, Y_n$ are i.i.d. Q. Consider a U-statistic of the form*

$$U_{m,n}(Z) = \frac{1}{\binom{m}{r}\binom{n}{s}} \sum_{\alpha} \sum_{\beta} \varphi(X_{\alpha_1}, \ldots, X_{\alpha_r}, Y_{\beta_1}, \ldots, Y_{\beta_s}) \ ,$$

where α and β range over the sets of all unordered subsets of r different elements chosen from $\{1, \ldots, m\}$ and of s different elements chosen from $\{1, \ldots, n\}$, respectively. Assume $E_{P,Q}\varphi(\cdot) = 0$ and $0 < E_{P,Q}\varphi^2(\cdot) < \infty$ for any permutation of $r + s$ Xs and Ys. Let $m \to \infty$, $n \to \infty$, with $N = m + n$, $m/N \to p > 0$ and $n/N \to q > 0$. Further assume that (9) holds. Also, let $\bar{Z}$ and π_0 be constructed by the coupling method. Then, for any random permutation $\pi = (\pi(1), \ldots, \pi(N))$ of $\{1, \ldots, N\}$,

$$\sqrt{m}U_{m,n}(Z_\pi) - \sqrt{m}U_{m,n}(\bar{Z}_{\pi \cdot \pi_0}) \overset{P}{\to} 0 \ . \tag{13}$$

Proof of Lemma A.1 First, consider the size of the number of observations D where Z and $\bar{Z}_{\pi_0}$ differ. By the central limit theorem, $|B_m - Np| = O_P(\sqrt{N})$, where B_m denotes the number of Xs in $\bar{Z}$, which follows a binomial distribution with parameters N and p. This fact together with (9) implies

$$D = |B_m - m| \leq |B_m - Np| + |Np - m| = O_P(\sqrt{N}) \ .$$

Furthermore,

$$\begin{aligned} \mathrm{E}(D) = \mathrm{E}(|B_m - m|) &\leq \mathrm{E}(|B_m - Np|) + |Np - m| \\ \leq \sqrt{\mathrm{E}((B_m - Np)^2)} + O(\sqrt{N}) &= \sqrt{Np(1-p)} + O(\sqrt{N}) = O(\sqrt{N}) \ . \end{aligned}$$

Now, to show (13), since the mean is assumed to be zero under any permutation of Xs and Ys, it suffices to show,

$$\mathrm{Var}\left(\sqrt{m}U_{m,n}(Z_\pi) - \sqrt{m}U_{m,n}(\bar{Z}_{\pi \cdot \pi_0})\right) \to 0 \ .$$

Note that given the sample Z, π_0 is a uniquely determined fixed permutation of $\{1, \ldots, N\}$ constructed by the coupling method. Everything stated below is implicitly conditioned on π_0 but the notation will be suppressed for simplicity. For a given π, let $C_{m,1}$ and $C_{n,2}$

be the number of observations where Z_π and $\bar{Z}_{\pi\cdot\pi_0}$ differ in the first m coordinates and in the second n coordinates, accordingly. Note that $C_{m,1}+C_{n,2}=|B_m-m|=D$. Then, by the law of total variance,

$$\begin{aligned}\mathrm{Var}\left(\sqrt{m}U_{m,n}(Z_\pi)-\sqrt{m}U_{m,n}(\bar{Z}_{\pi\cdot\pi_0})\right) &= \mathrm{Var}\left(\mathrm{E}\left(\sqrt{m}(U_{m,n}(Z_\pi)-U_{m,n}(\bar{Z}_{\pi\cdot\pi_0}))|C_{m,1},C_{n,2}\right)\right)\\ &\quad+\mathrm{E}\Big(\mathrm{Var}\left(\sqrt{m}(U_{m,n}(Z_\pi)-U_{m,n}(\bar{Z}_{\pi\cdot\pi_0}))|C_{m,1},C_{n,2}\right)\Big)\ ,\end{aligned}$$

where the first term on the right side is zero since $\mathrm{E}\left(\sqrt{m}(U_{m,n}(Z_\pi)-U_{m,n}(\bar{Z}_{\pi\cdot\pi_0}))|C_{m,1},C_{n,2}\right)=0$ by assumption. Thus, it suffices to show

$$\mathrm{E}\Big(\mathrm{Var}\left(\sqrt{m}(U_{m,n}(Z_\pi)-U_{m,n}(\bar{Z}_{\pi\cdot\pi_0}))|C_{m,1},C_{n,2}\right)\Big)\to 0\ . \tag{14}$$

To begin, note that conditioning on π and hence $C_{m,1}$ and $C_{n,2}$, $\sqrt{m}\left(U_{m,n}(Z_\pi)-U_{m,n}(\bar{Z}_{\pi\cdot\pi_0})\right)$ can be divided into $(r+1)(s+1)$ cases given by

$$\sqrt{m}\left(U_{m,n}(Z_\pi)-U_{m,n}(\bar{Z}_{\pi\cdot\pi_0})|C_{m,1},C_{n,2}\right)=\frac{\sqrt{m}}{\binom{m}{r}\binom{n}{s}}\sum_{r_0=0}^{r}\sum_{s_0=0}^{s}S_{r_0,s_0}\ ,$$

where S_{r_0,s_0} each denotes the linear sum of terms in the case where r_0 of the first r coordinates and s_0 of the last s coordinates in $\varphi(Z_{\pi(\alpha_1)},\ldots,Z_{\pi(\alpha_r)},Z_{\pi(\beta_1)},\ldots,Z_{\pi(\beta_s)})$ and $\varphi(\bar{Z}_{\pi_0\cdot\pi(\alpha_1)},\ldots,\bar{Z}_{\pi_0\cdot\pi(\alpha_r)},\bar{Z}_{\pi_0\cdot\pi(\beta_1)},\ldots,\bar{Z}_{\pi_0\cdot\pi(\beta_s)})$ are the same. For instance, in the case $r=s=1$, $S_{0,0},S_{0,1},S_{1,0}$, and $S_{1,1}$ each denotes the sum of all the $\varphi(Z_{\pi(i)},Z_{\pi(j)})-\varphi(\bar{Z}_{\pi_0\cdot\pi(i)},\bar{Z}_{\pi_0\cdot\pi(j)})$ terms where

(i) $Z_{\pi(i)}\neq\bar{Z}_{\pi_0\cdot\pi(i)}$ and $Z_{\pi(m+j)}\neq\bar{Z}_{\pi_0\cdot\pi(m+j)}$ ($C_{m,1}C_{n,2}$ terms),

(ii) $Z_{\pi(i)}\neq\bar{Z}_{\pi_0\cdot\pi(i)}$ and $Z_{\pi(m+j)}=\bar{Z}_{\pi_0\cdot\pi(m+j)}$ ($C_{m,1}(n-C_{n,2})$ terms),

(iii) $Z_{\pi(i)}=\bar{Z}_{\pi_0\cdot\pi(i)}$ and $Z_{\pi(m+j)}\neq\bar{Z}_{\pi_0\cdot\pi(m+j)}$ ($(m-C_{m,1})C_{n,2}$ terms), and

(iv) $Z_{\pi(i)}=\bar{Z}_{\pi_0\cdot\pi(i)}$ and $Z_{\pi(m+j)}=\bar{Z}_{\pi_0\cdot\pi(m+j)}$ ($(m-C_{m,1})(n-C_{n,2})$ terms), respectively.

Thus, the conditional variance given $C_{m,1}$ and $C_{n,2}$ becomes

$$\begin{aligned}\mathrm{Var}\left(\sqrt{m}(U_{m,n}(Z_\pi)-U_{m,n}(\bar{Z}_{\pi\cdot\pi_0}))|\pi_0,C_{m,1},C_{n,2}\right) &= \frac{m}{\binom{m}{r}^2\binom{n}{s}^2}\mathrm{Var}\left(\sum_{r_0=0}^{r}\sum_{s_0=0}^{s}S_{r_0,s_0}\Big|C_{m,1},C_{n,2}\right)\\ &\le\frac{m}{\binom{m}{r}^2\binom{n}{s}^2}\left(\sum_{r_0=0}^{r}\sum_{s_0=0}^{s}\sqrt{\mathrm{Var}\left(S_{r_0,s_0}|C_{m,1},C_{n,2}\right)}\right)^2\ .\end{aligned}\tag{15}$$

Note that when $r_0=r$ and $s_0=s$, the conditional variance of the difference given π_0, $C_{m,1}$, and $C_{n,2}$ is obviously zero, i.e., $\mathrm{Var}(S_{r,s}|C_{m,1},C_{n,2})=0$. Now consider the case where at least one of them is strictly smaller than its order, i.e., when $r_0 < r$ or $s_0 < s$. First, let $V_{a,b}$ be the second moment of φ with a numbers of Xs in the first r arguments (so $(r-a)$ Ys in the first r arguments) and b numbers of Ys in the last s arguments (so $(s-b)$ Xs in the last s arguments) , i.e., (up to symmetry of φ)

$$V_{a,b} \equiv \mathrm{E}\Big[\varphi(X_1, \dots, X_a, Y_{b+1}, \dots, Y_{b+r-a}, Y_1, \dots, Y_b, X_{a+1}, \dots, X_{a+s-b})^2\Big] \ .$$

Let $B_{a,b}$ be the bound of $V_{a,b}$. Then, the by the Cauchy-Schwarz inequality, all the covariance terms as well as variances are bounded by B, where $B = \max_{a=0,\dots,r\&b=0,\dots,s}\{B_{a,b}\}$. Note that when both $\mathrm{Var}(A)$ and $\mathrm{Var}(A')$ are bounded by B,

$$\mathrm{Var}(A - A') \le (\sqrt{\mathrm{Var}(A)} + \sqrt{\mathrm{Var}(A')})^2 \le (\sqrt{B} + \sqrt{B})^2 = 4 \cdot B \ . \tag{16}$$

Thus, it follows from (16) and the Cauchy-Schwarz inequality that

$$\frac{m}{\binom{m}{r}^2\binom{n}{s}^2} \mathrm{Var}\left(S_{r_0,s_0} | C_{m,1}, C_{n,2}\right) \le \frac{m}{\binom{m}{r}^2\binom{n}{s}^2} A_{r_0,s_0}(C_{m,1}, C_{n,2}) \cdot 4 \cdot B \ , \tag{17}$$

where $A_{r_0,s_0}(C_{m,1}, C_{n,2})$ is a constant depending on $C_{m,1}$ and $C_{n,2}$; more precisely, $A_{r_0,s_0}(C_{m,1}, C_{n,2}) =$

$$\binom{m - C_{m,1}}{r_0}\binom{C_{m,1}}{r - r_0}\binom{n - C_{n,2}}{s_0}\binom{C_{n,2}}{s - s_0}\left[\binom{m - C_{m,1}}{r_0}\binom{C_{m,1}}{r - r_0}\binom{n - C_{n,2}}{s_0}\binom{C_{n,2}}{s - s_0}\right.$$
$$\left. - \binom{m - C_{m,1} - r_0}{r_0}\binom{C_{m,1} - r + r_0}{r - r_0}\binom{n - C_{n,2} - s_0}{s_0}\binom{C_{n,2} - s + s_0}{s - s_0}\right] .$$

The constant $A_{r_0,s_0}(C_{m,1}, C_{n,2})$ is the number of nonzero covariance terms in $\mathrm{Var}(S_{r_0,s_0} | C_{m,1}, C_{n,1})$. For example, in the case $r = s = 1$ and $r_0 = 1$ and $s_0 = 0$,

$$\begin{aligned}
\frac{1}{mn^2}\mathrm{Var}(S_{1,0} | C_{m,1}, C_{n,2}) &= \frac{1}{mn^2}\mathrm{Var}\left(\sum_{i=1}^{m - C_{m,1}} \sum_{j = n - C_{n,2}}^{n} \left(\varphi\left(Z_i, X_{m+j}\right) - \varphi\left(Z_i, Y_{m+j}\right)\right) | C_{m,1}, C_{n,2}\right) \\
&= \frac{1}{mn^2}\Bigg[(m - C_{m,1})C_{n,2}\mathrm{E}\left(\varphi\left(Z_1, X_2\right) - \varphi\left(Z_1, Y_2\right)\right)^2 \\
&+ (m - C_{m,1})C_{n,2}\left(C_{n,2} - 1\right)\mathrm{E}\left(\varphi\left(Z_1, X_2\right) - \varphi\left(Z_1, Y_2\right)\right)\left(\varphi\left(Z_1, X_3\right) - \varphi\left(Z_1, Y_3\right)\right) \\
&+ (m - C_{m,1})\left(m - C_{m,1} - 1\right)C_{n,2}\mathrm{E}\left(\varphi\left(Z_1, X_3\right) - \varphi\left(Z_1, Y_3\right)\right)\left(\varphi\left(Z_2, X_3\right) - \varphi\left(Z_2, Y_3\right)\right)\Bigg] \\
&\le \frac{(m - C_{m,1})C_{n,2}(m + C_{n,2} - C_{m,1} - 1)}{mn^2} \cdot 4 \cdot \tilde{B} = \frac{1}{mn^2}A_{1,0}(C_{m,1}, C_{n,2}) \cdot 4 \cdot \tilde{B} \ ,
\end{aligned}$$

where $\tilde{B} = \max\{B_{0,0}, B_{0,1}, B_{1,0}, B_{1,1}\}$.

Thus, the conditional variance given $C_{m,1}$ and $C_{n,2}$ given in (15) is bounded by

$$\mathrm{Var}\left(\sqrt{m}(U_{m,n}(Z_\pi) - U_{m,n}(\bar{Z}_{\pi\cdot\pi_0}))|\pi_0, C_{m,1}, C_{n,2}\right) \leq \frac{m}{\binom{m}{r}^2\binom{n}{s}^2}\left(\sum_{r_0=0}^{r}\sum_{s_0=0}^{s}\sqrt{A_{r_0,s_0}(C_{m,1}, C_{n,2}) \cdot 4 \cdot B}\right)^2 . \tag{18}$$

Since both $C_{m,1}$ and $C_{n,2}$ are $O_P(\sqrt{N})$, the bound in (18) is conditionally at most $O_P(\frac{1}{\sqrt{N}})$. Moreover, since both $\mathrm{E}(C_{m,1})$ and $\mathrm{E}(C_{n,2})$ are $O(\sqrt{N})$,

$$\mathrm{E}\left(\frac{m}{\binom{m}{r}^2\binom{n}{s}^2}\left(\sum_{r_0=0}^{r}\sum_{s_0=0}^{s}\sqrt{A_{r_0,s_0}(C_{m,1}, C_{n,2}) \cdot 4 \cdot B}\right)^2\right)$$

is at most $O(\frac{1}{\sqrt{N}})$. Hence, (14) follows. ■

We shall now show that the two-sample U-statistic $U_{m,n}(\bar{Z})$ can be approximated by its Hàjek projection.

Lemma A.2. *Assume the same setup and conditions of Lemma A.1. Define the Hàjek projection of $U_{m,n}$ as*

$$\tilde{U}_{m,n}(\bar{Z}) \equiv \sum_{i=1}^{N} \mathrm{E}[U_{m,n}|\bar{Z}_i] .$$

Then,

$$\sqrt{m}U_{m,n}(\bar{Z}) - \sqrt{m}\tilde{U}_{m,n}(\bar{Z}) \xrightarrow{P} 0 .$$

Proof of Lemma A.2 For the conditional expectation of $U_{m,n}$ given $\bar{Z}_i$,

$$\mathrm{E}[U_{m,n}|\bar{Z}_i] = \frac{1}{\binom{m}{r}\binom{n}{s}}\sum_{\alpha}\sum_{\beta}\mathrm{E}\big(\varphi(\bar{Z}_{\alpha_1}, \ldots, \bar{Z}_{\alpha_r}, \bar{Z}_{\beta_1}, \ldots, \bar{Z}_{\beta_s})|\bar{Z}_i\big) ,$$

since the kernel φ is symmetric in the first r and last s arguments, if $i \leq m$,

$$\mathrm{E}[U_{m,n}|\bar{Z}_i] = \frac{1}{\binom{m}{r}}\left[\binom{m-1}{r-1}\varphi_{\cdot\bar{P}^{r-1}\bar{P}^s}(\bar{Z}_i)\right]$$

where

$$\varphi_{\cdot\bar{P}^{r-1}\bar{P}^s}(a_1) \equiv \int\cdots\int \varphi(a_1, \ldots, a_r, b_1, \ldots, b_s)d\bar{P}(a_2)\cdots d\bar{P}(b_s) .$$

On the other hand, when $i > m$, then

$$\mathrm{E}[U_{m,n}|\bar{Z}_i] = \frac{1}{\binom{n}{s}}\left[\binom{n-1}{s-1}\varphi_{\bar{P}^r.\bar{P}^{s-1}}(\bar{Z}_i)\right]$$

where

$$\varphi_{\bar{P}^r.\bar{P}^{s-1}}(\bar{Z}_i)(b_1) \equiv \int\cdots\int \varphi(a_1,\ldots,a_r,b_1,\ldots,b_s)d\bar{P}(a_1)\cdots d\bar{P}(a_r)d\bar{P}(b_2)\cdots d\bar{P}(b_s)\ .$$

Hence, the Hájek projection $\tilde{U}_{m,n}$ becomes

$$\begin{aligned}\sum_{i=1}^{N}\mathrm{E}[U_{m,n}|\bar{Z}_i] =& \frac{1}{\binom{m}{r}}\sum_{i=1}^{m}\binom{m-1}{r-1}\varphi_{.\bar{P}^{r-1}\bar{P}^s}(\bar{Z}_i) + \frac{1}{\binom{n}{s}}\sum_{i=m+1}^{N}\binom{n-1}{s-1}\varphi_{\bar{P}^r.\bar{P}^{s-1}}(\bar{Z}_i)\\ =& \frac{r}{m}\sum_{i=1}^{m}\varphi_{.\bar{P}^{r-1}\bar{P}^s}(\bar{Z}_i) + \frac{s}{n}\sum_{i=m+1}^{N}\varphi_{\bar{P}^r.\bar{P}^{s-1}}(\bar{Z}_i)\ .\end{aligned}$$

Assuming $\frac{m}{N} \to p$, the limiting variance of $\sqrt{m}\tilde{U}_{m,n}(\bar{Z})$ then can be obtained by

$$\begin{aligned}m\,\mathrm{Var}(\tilde{U}_{m,n}(\bar{Z})) &= r^2\mathrm{E}\varphi^2_{.\bar{P}^{r-1}\bar{P}^s}(\bar{Z}_i) + \frac{m}{n}s^2\mathrm{E}\varphi^2_{\bar{P}^r.\bar{P}^{s-1}}(\bar{Z}_i)\\ &\to r^2\mathrm{E}\varphi^2_{.\bar{P}^{r-1}\bar{P}^s}(\bar{Z}_i) + \frac{p}{1-p}s^2\mathrm{E}\varphi^2_{\bar{P}^r.\bar{P}^{s-1}}(\bar{Z}_i) \equiv \tau^2\ .\end{aligned} \tag{19}$$

By the central limit theorem, we now have

$$\mathrm{P}\{\sqrt{m}\tilde{U}_{m,n} \le t\} \to \Phi(\frac{t}{\tau})\ , \tag{20}$$

where τ^2 is defined in (19).

Now, for the exact variance $\mathrm{Var}\left(\sqrt{m}U_{m,n}(\bar{Z})\right)$

$$= \frac{m}{\binom{m}{r}^2\binom{n}{s}^2}\sum_{\alpha}\sum_{\alpha'}\sum_{\beta}\sum_{\beta'}\mathrm{Cov}\left(\varphi(\bar{Z}_{\alpha_1},\ldots,\bar{Z}_{\alpha_r},\bar{Z}_{\beta_1},\ldots,\bar{Z}_{\beta_s}),\varphi(\bar{Z}_{\alpha'_1},\ldots,\bar{Z}_{\alpha'_r},\bar{Z}_{\beta'_1},\ldots,\bar{Z}_{\beta'_s})\right)\ ,$$

let a_0 and b_0 be the number of $\bar{Z}_i$s that differ between $\bar{Z}_\alpha$ and $\bar{Z}_{\alpha'}$ and between $\bar{Z}_\beta$ and $\bar{Z}_{\beta'}$ in the covariance term, respectively. Then, the exact variance of $\sqrt{m}U_{m,n}(\bar{Z})$ can be expressed as

$$\frac{m}{\binom{m}{r}\binom{n}{s}}\left[\sum_{a_0=0}^{r}\sum_{b_0=0}^{s}\binom{r}{a_0}\binom{s}{b_0}\binom{m-r}{a_0}\binom{n-s}{b_0}\mathrm{E}\varphi^2_{.r-a_0\,\bar{P}^{a_0}.s-b_0\,\bar{P}^{b_0}}\right]\ ,$$

where

$$\mathrm{E}\varphi^2_{.r-a_0\,\bar{P}^{a_0}.s-b_0\,\bar{P}^{b_0}} = \int\cdots\int \varphi(a_1,\cdots,a_{r-a_0},a_{r-a_0+1},\cdots,a_r,b_1,\cdots,b_{s-b_0},\bar{b}_{s-b_0+1},\cdots,b_s)\cdot$$

$$\varphi(a_1,\cdots,a_{r-a_0},a'_{r-a_0+1},\cdots,a'_r,b_1,\cdots,b_{s-b_0},b'_{s-b_0+1},\cdots,b'_s)d\bar{P}(a_1)\cdots d\bar{P}(a_{r-a_0})d\bar{P}(b_1)\cdots d\bar{P}(b_{s-b_0})\,.$$

Note that either when $a_0 = r-1$ and $b_0 = s$ or when $a_0 = r$ and $b_0 = s-1$, the expression above has the biggest order and the rest term is $O(1/m)$. Thus, it can be easily shown that $m\,\mathrm{Var}\left(U_{m,n}(\bar{Z})\right)$

$$= \frac{m}{\binom{m}{r}\binom{n}{s}}\left[r\binom{m-r}{r-1}\binom{n-s}{s}\mathrm{E}\varphi^2_{.\bar{P}^{r-1}\bar{P}^s}+s\binom{m-r}{r}\binom{n-s}{s-1}\mathrm{E}\varphi^2_{\bar{P}^r.\bar{P}^{s-1}}\right]+O(\frac{1}{m})\,. \quad (21)$$

Thus, from (19) and (21), one can readily show that the difference between the exact variance of $\sqrt{m}U_{m,n}$ and the projection variance of $\sqrt{m}\tilde{U}_{m,n}$ becomes

$$\begin{aligned}\mathrm{E}(\sqrt{m}\tilde{U}_{m,n}-\sqrt{m}U_{m,n})^2 &= m\,\mathrm{Var}(U_{m,n}(\bar{Z}_\pi)) - m\,\mathrm{Var}(\tilde{U}_{m,n}(\bar{Z}_\pi))\\ &= O(\frac{1}{m})\,.\end{aligned}$$

Hence, by Chebychev's Inequality,

$$\mathrm{P}\{|\sqrt{m}U_{m,n}-\sqrt{m}\tilde{U}_{m,n}|>\varepsilon\} \le \frac{\mathrm{EE}[m(\tilde{U}_{m,n}-U_{m,n})^2]}{\varepsilon^2}\to 0\,,$$

which implies

$$\sqrt{m}U_{m,n}(\bar{Z})-\sqrt{m}\tilde{U}_{m,n}(\bar{Z})\xrightarrow{P}0\,.\ \blacksquare$$

Remark A.1. Note that since $\bar{Z}$s are i.i.d., Lemma A.2 implies that for any random permutation π of $\{1,\ldots,N\}$, $\sqrt{m}U_{m,n}(\bar{Z}_\pi)-\sqrt{m}\tilde{U}_{m,n}(\bar{Z}_\pi)\xrightarrow{P}0$.

B Proofs of Theorems

Proof of Theorem 2.1 Assuming $\frac{m}{N}\to p\in(0,1)$, it is well-known (van der Vaart, 1998) that the limiting sampling distribution of $\sqrt{m}(U_{m,n}-\theta)$ satisfies

$$\sqrt{m}\,(U_{m,n}-\theta)\xrightarrow{d}\mathcal{N}\left(0,\xi_x+\frac{p}{1-p}\xi_y\right)\,, \quad (22)$$

where $\xi_x = \text{Var}\left(Q_Y^-(X_i)\right)$ and $\xi_y = \text{Var}\left(P_X(Y_j)\right)$. Note that if the two samples share the same continuous underlying distribution $P = Q$, then the sampling distribution of $\sqrt{m}(U_{n,m} - \theta)$ satisfies

$$\sqrt{m}\,(U_{m,n} - \theta) \overset{d}{\to} \mathcal{N}\left(0, \frac{1}{12(1-p)}\right) \ . \tag{23}$$

To analyze the permutation distribution $\hat{R}^U_{m,n}$ of $\sqrt{m}(U_{m,n} - \theta)$, recall the important property of the Wilcoxon statistic: it is a rank statistic. Since the statistic is solely based on ranks, the permutation distribution, which is the conditional distribution given the order statistics, is a fixed distribution that depends only on sample sizes. Thus, its behavior would be the same, regardless of what the true underlying distributions P and Q are. This property naturally allows the randomization distribution to be understood by the sampling distribution in the case of $P = Q$. Thus, the result follows. ■

Proof of Theorem 2.2 First, notice that both the numerator, the standard Wilcoxon statistic, and the denominator, a consistent estimator of the variance, of $\tilde{U}_{m,n}$ are rank statistics. Since they are based solely on ranks, the permutation distribution is some fixed distribution which depends on the sample sizes m and n. Thus, for the given (sequence of) sample sizes, the limiting behavior of the permutation distribution of $\sqrt{m}\tilde{U}_{m,n}$ is the same regardless of the true underlying distribution P and Q. Hence, we can conclude that the permutation distribution of the studentized Wilcoxon statistic has the same limiting distribution as the sampling distribution in the case of $P = Q$. But by Lemma B.1 below, the unconditional sampling distribution, regardless of what P and Q are, converges in distribution to a standard normal. Consequently, the permutation distribution of $\sqrt{m}\tilde{U}_{m,n}$ and the true unconditional sampling distribution have the same limiting distribution $N(0, 1)$. This result confirms that the critical value $\hat{r}_{m,n}(1 - \alpha)$ of the studentized two-sample Wilcoxon test converges to the $(1 - \alpha)$ quantile of the true sampling distribution $z_{1-\alpha}$. ■

Lemma B.1. *Assume the setup and the conditions of Theorem 2.2. Let $\hat{\xi}_x$ and $\hat{\xi}_y$ be defined by (6) and (7), respectively. Then, the sampling distribution of $\sqrt{m}\tilde{U}_{m,n}$ converges weakly to a standard normal distribution.*

Proof of Lemma B.1 Since $\sqrt{m}(U_{m,n} - \theta)$ is asymptotically normal with mean 0 and variance $\xi_x + \frac{p}{1-p}\xi_y$, it suffices to show that

$$\hat{\sigma}_{m,n} \equiv \hat{\xi}_x + \frac{m}{n}\hat{\xi}_y \tag{24}$$

is a consistent estimator of $\xi_x + \frac{p}{1-p}\xi_y$ since then by Slutsky's Theorem, the limiting distribution of $\sqrt{m}\tilde{U}_{m,n}$ converges weakly to a standard normal distribution. First, let $\hat{Q}_n(y) = \frac{1}{n}\sum_{j=1}^{n} \mathrm{I}\{Y_j \le y\}$ be the empirical cdf of the Ys. It follows that

$$\left|\frac{1}{m}\sum_{i=1}^{m} \hat{Q}_n(X_i) - \frac{1}{m}\sum_{i=1}^{m} Q_Y(X_i)\right| \le \sup_y \left|\hat{Q}_n - Q_Y(y)\right| \to 0 \quad \text{a.s.}$$

by Glivenko-Cantelli Theorem. By the strong law of large numbers,

$$\frac{1}{m}\sum_{i=1}^{m} Q_Y(X_i) \to \mathrm{E}Q_Y(X_i) \quad \text{a.s.},$$

which implies

$$\frac{1}{m}\sum_{i=1}^{m} \hat{Q}_n(X_i) \to \mathrm{E}Q_Y(X_i) \;\; \text{a.s.} \tag{25}$$

Furthermore, $\hat{\xi}_x$ can be expressed by

$$\begin{aligned}
\hat{\xi}_x &= \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{Q}_n(X_i) - \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_n(X_i)\right)^2 \\
&= \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{Q}_n(X_i) - \mathrm{E}Q_Y(X_i)\right)^2 + \frac{1}{m-1}\sum_{i=1}^{m}\left(\mathrm{E}Q_Y(X_i) - \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_n(X_i)\right)^2 \\
&+ \frac{2}{m-1}\sum_{i=1}^{m}\left(\hat{Q}_n(X_i) - \mathrm{E}Q_Y(X_i)\right)\left(\mathrm{E}Q_Y(X_i) - \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_n(X_i)\right) .
\end{aligned}$$

Note that both the second term and the third term on the right hand side converge to 0 with probability 1 due to (25). In order to investigate the first term, let us further expand the first term as follows,

$$\begin{aligned}
\frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{Q}_n(X_i) - \mathrm{E}Q_Y(X_i)\right)^2 &= \frac{1}{m-1}\sum_{i=1}^{m}\hat{Q}_n(X_i)^2 - \frac{2}{m-1}\sum_{i=1}^{m}\hat{Q}_n(X_i)\mathrm{E}Q_Y(X_i) \\
&+ \frac{1}{m-1}\sum_{i=1}^{m}\left(\mathrm{E}Q_Y(X_i)\right)^2 .
\end{aligned}$$

The second term $-\frac{2}{m-1}\sum_{i=1}^{m}\hat{Q}_n(X_i)\mathrm{E}Q_Y(X_i)$ converges a.s. to $-2\left(\mathrm{E}Q_Y(X_i)\right)^2$ by Slutsky's Theorem and (25). To examine the first term $\frac{1}{m-1}\sum_{i=1}^{m}\hat{Q}_n(X_i)^2$, first note that

$$\begin{aligned}\sup_y\left|\hat{Q}_n(y)^2-Q_Y(y)^2\right| &= \sup_y\left|\hat{Q}_n(y)+Q_Y(y)\right|\left|\hat{Q}_n(y)-Q_Y(y)\right|\\ &\le 2\sup_y\left|\hat{Q}_n(y)-Q_Y(y)\right| \to 0 \quad \text{a.s. },\end{aligned}$$

since we know both $\hat{Q}_n(y)$ and $Q_Y(y)$ are bounded by 1 and $\sup_y|\hat{Q}_n(y)-Q_Y(y)|\to 0$ with probability 1. Thus, with a similar argument as before, we can conclude that

$$\frac{1}{m-1}\sum_{i=1}^{m}\hat{Q}_n(X_i)^2 \to \mathrm{E}\left[Q_Y(X_i)^2\right] \quad \text{a.s.}$$

and therefore

$$\begin{aligned}\frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{Q}_n(X_i)-\mathrm{E}Q_Y(X_i)\right)^2 &\xrightarrow{P} \mathrm{E}\left[Q_Y(X_i)^2\right]-\left(\mathrm{E}Q_Y(X_i)\right)^2\\ &= \mathrm{Var}(Q_Y(X_i)) = \xi_x \ .\end{aligned}$$

Similarly, it can be shown that $\hat{\xi}_y$ converges to ξ_y in probability. Hence,

$$\hat{\xi}_x+\frac{m}{n}\hat{\xi}_y \xrightarrow{P} \xi_x+\frac{p}{1-p}\xi_y \ .$$

■

Proof of Theorem 2.3 Consider a sequence of situations with $X=X^m, Y=Y^n, P=P^m, Q=Q^n, \mathcal{X}=\mathcal{X}_{m,n}, \mathbf{G}=\mathbf{G}_{m,n}, T=T_{m,n}$, etc., defined for an infinite sequence of positive integers m and n. As shown in (23), the limiting permutation distribution of $\sqrt{N}(U_{m,n}-\theta)$ under $\Delta=0$ satisfies

$$\sqrt{m}(U_{m,n}-\theta) \xrightarrow{d} N\left(0,\tau^2\right) \ ,$$

where $\tau^2=\frac{1}{12(1-p)}$ and the critical value $\hat{r}_{m,n}(1-\alpha)$ of the permutation test converges to $\tau z_{1-\alpha}$ under $\Delta=0$. Hence, under a sequence of contiguous alternatives Δ_m converging to the null $\Delta=0$ at rate $O(\frac{1}{\sqrt{N}})$, the critical value also tends to $\tau z_{1-\alpha}$ in probability. Since $\sqrt{m}(U_{m,n}-\theta)$ is tight under $\Delta=0$, by Theorem 12.3.2 (ii) of Lehmann and Romano (2005), so is $\sqrt{m}(U_{m,n}-\theta)$ under a sequence of contiguous alternatives $\Delta_m=\frac{h}{\sqrt{m}}$. Then, by Prohorov's Theorem, there exists a subsequence along which

$$\sqrt{m}\left(U_{m_i,n_j}-\theta\right) \xrightarrow{d} \mathcal{L}_h \ , \tag{26}$$

for some $\mathcal{L}_h$ under P_{Δ_m}. Thus, the limiting power of the two-sample Wilcoxon test against local alternatives along this subsequence is

$$\mathrm{P}_{\Delta_m}\left(\sqrt{m}\left(U_{m_i,n_j}-\theta\right) > \tau z_{1-\alpha}\right) \to 1-\mathcal{L}_h(\tau z_{1-\alpha}) \ . \tag{27}$$

Now, we want to compare (27) with the limiting power of the studentized Wilcoxon test along the same subsequence. For $\tilde{U}_{m,n}$ given by (2), it is proved in Theorem 2.2 that the limiting permutation distribution of $\sqrt{m}\tilde{U}_{m,n}$ is a standard normal distribution. Thus, the critical value $\hat{r}_{m,n}(1-\alpha)$ of the permutation test converges to $z_{1-\alpha}$ under $\Delta = 0$. Hence, under a sequence of contiguous alternatives Δ_m converging to the null $\Delta = 0$ at rate $O(\frac{1}{\sqrt{N}})$, the critical value also tends to $z_{1-\alpha}$ in probability. Furthermore, we know by contiguity that

$$\hat{\sigma}_{m,n} \xrightarrow{P} \tau, \quad \text{under} \quad \Delta_m = \frac{h}{\sqrt{m}} \ , \tag{28}$$

where $\hat{\sigma}_{m,n}$ is defined in (24). Combining (26) and (28), we can conclude by Slutsky's Theorem that along the same subsequence,

$$\mathrm{P}_{\Delta_m}\left(\sqrt{m}\left(\frac{U_{m_i,n_j}-\theta}{\hat{\sigma}_{m,n}}\right) > z_{1-\alpha}\right) \to 1-\mathcal{L}_h(\tau z_{1-\alpha}) \ . \tag{29}$$

By (27) and (29), we achieve that the limiting power against local alternatives of tests based on $\tilde{U}_{m,n}$ is the same as that of the standard Wilcoxon test under the null of $\Delta = 0$. Thus, the studentized Wilcoxon test has the same asymptotic Pitman efficiency as the standard Wilcoxon test. In particular, the studentized Wilcoxon test has the same asymptotic relative efficiency with respect to the usual two-sample t-test as the standard Wilcoxon test. ■

Proof of Theorem 3.1 Independent of the Zs, let π and π' be independent permutations of $\{1,\ldots,N\}$. By Theorem 15.2.3 of Lehmann and Romano (2005), it suffices to show that the joint limiting behavior satisfies

$$\left(\sqrt{m}U_{m,n}(Z_\pi), \sqrt{m}U_{m,n}(Z_{\pi'})\right) \xrightarrow{d} (T, T') \ ,$$

where T and T' are independent, each with common c.d.f. $\Phi(t/\bar{\tau})$ for $\bar{\tau}^2$ defined in (11). However, if we can show

$$\left(\sqrt{m}\tilde{U}_{m,n}(\bar{Z}_\pi), \sqrt{m}\tilde{U}_{m,n}(\bar{Z}_{\pi'})\right) \xrightarrow{d} (T, T') \ , \tag{30}$$

then, by Lemma A.2 and Remark A.1, (30) implies

$$\big(\sqrt{m}U_{m,n}(\bar{Z}_{\pi}), \sqrt{m}U_{m,n}(\bar{Z}_{\pi'})\big) \xrightarrow{d} (T, T').$$

But since $\pi \cdot \pi_0$ and $\pi' \cdot \pi_0$ are also independent permutations, it follows

$$\big(\sqrt{m}U_{m,n}(\bar{Z}_{\pi\cdot\pi_0}), \sqrt{m}U_{m,n}(\bar{Z}_{\pi'\cdot\pi'})\big) \xrightarrow{d} (T, T') \ ,$$

which also implies by Lemma A.1 that

$$\big(\sqrt{m}U_{m,n}(Z_{\pi}), \sqrt{m}U_{m,n}(Z_{\pi'})\big) \xrightarrow{d} (T, T').$$

To show (30), note that the antisymmetry assumption of φ defined by (10) implies $\varphi_{.\bar{P}^{r-1}\bar{P}^r}(\bar{Z}_i) = -\varphi_{\bar{P}^r.\bar{P}^{r-1}}(\bar{Z}_i)$. As a result, the Hàjek projection of $\sqrt{m}U_{m,n}$ becomes

$$\begin{aligned}\sqrt{m}\tilde{U}_{m,n}(\bar{Z}) &= \frac{r}{\sqrt{m}}\left(\sum_{i=1}^{m}\varphi_{.\bar{P}^{r-1}\bar{P}^r}(\bar{Z}_i) - \frac{m}{n}\sum_{i=m+1}^{N}\varphi_{.\bar{P}^{r-1}\bar{P}^r}(\bar{Z}_i)\right)\\ &= \frac{r}{\sqrt{m}}\sum_{i=1}^{N}W_i\varphi_{.\bar{P}^{r-1}\bar{P}^r}(\bar{Z}_i) \ ,\end{aligned}$$

where

$$W_i = \begin{cases} 1 & \text{if} \quad \pi(i) \leq m \\ -\frac{m}{n} & \text{if} \quad \pi(i) > m \ . \end{cases}$$

That is, this problem is reduced to the general mean case based on observations $\varphi_{.\bar{P}^{r-1}\bar{P}^r}(\bar{Z}_1)$, $\ldots$, $\varphi_{.\bar{P}^{r-1}\bar{P}^r}(\bar{Z}_N)$ instead of $Z_1, \ldots, Z_N$. Thus, it follows by Theorem 15.2.5 of Lehmann and Romano (2005) that

$$\Big(\sqrt{m}\tilde{U}_{m,n}(\bar{Z}_{\pi}), \sqrt{m}\tilde{U}_{m,n}(\bar{Z}_{\pi'})\Big)$$

converges in distribution to a bivariate normal distribution with independent, identically distributed marginals having mean 0 and variances given by (11). ■

Proof of Theorem 3.2 We first shall show that $V^2_{m,n}(Z_{\pi(1)}, \ldots, Z_{\pi(N)})$ is a consistent estimator for $\bar{\tau}^2$, i.e.,

$$V^2_{m,n}(Z_{\pi(1)}, \ldots, Z_{\pi(N)}) \xrightarrow{P} \bar{\tau}^2 \ ,$$

where $\bar{\tau}^2$ is defined in (11). To do so, it suffices to show that

$$\hat{\sigma}^2_m(Z_{\pi(1)}, \ldots, Z_{\pi(m)}) \xrightarrow{P} \int \varphi^2_{.\bar{P}^{r-1}\bar{P}^r}d\bar{P} \qquad (31)$$

and

$$\hat{\sigma}_n^2(Z_{\pi(m+1)}, \ldots, Z_{\pi(N)}) \xrightarrow{P} \int \varphi^2_{\cdot \bar{P}^{r-1} \bar{P}^{r}} d\bar{P} \ . \tag{32}$$

However, (31) and (32) follow from a key contiguity argument for the binomial and hypergeometric distributions shown in Lemma 3.2 and Lemma 3.3 of Chung and Romano (2011). Now let $R^V_{m,n}(\cdot)$ denote the permutation distribution corresponding to the statistic $V_{m,n}$, as defined in (1) with T replaced by V. By Slutsky's Theorem for stochastic randomization distribution (Theorem 3.2 of Chung and Romano (2011)), $\hat{R}^V_{m,n}(t)$ converges to $\delta_{\bar{\tau}^2}(t)$ for all $t \neq \bar{\tau}^2$, where $\delta_c(\cdot)$ denotes the c.d.f. of the distribution placing mass one at the constant c. Thus, we can apply Slutsky's Theorem for stochastic randomization distribution again together with Theorem 3.1 to conclude that the permutation distribution of $\sqrt{m} S_{m,n}$ satisfies (12). ■

References

Casella, G. and Berger, R. (2001). *Statistical Inference*. 2nd edition, Duxbury Press, Pacific Grove, California.

Charness, G., Rigotti, L., and Rustichini, A. (2007). Individual Behavior and Group Membership. *American Economic Review* **97**, 1340–1352.

Chung, E. and Romano, J. (2011) Exact and Asymptotically Robust Permutation Tests.

Davis, J (2004). An Annual Index of U. S. Industrial Production, 1790-1915. *Quarterly Journal of Economics* **119**, 1177–1215.

Devroye, L. and Wagner, T.J. (1980). The strong uniform consistency of kernel density estimates. *Multivariate Analysis V* (P.R. Krishnaiah, *ed.*). North Holland, 59–77.

Dohmen, T. and Falk, A. (2011). Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender. *American Economic Review* **101**, 556–590.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics* **7**, 1–26.

Feltovich, N. (2003). Nonparametric Tests of Differences in Medians: Comparison of the Wilcoxon-Mann-Whitney and Robust Rank-Order Tests. *Experimental Economics* **6**, 273–297.

Feri, F., Irlenbusch, B. and Sutter, M. (2010). Efficiency Gains from Team-Based CoordinationLarge-Scale Experimental Evidence. *American Economic Review* **100**, 1892–1912.

Fligner, M. A. and Policello, G. E. II. (1981). Robust Rank Procedures for the Behrens-Fisher Problem. *Journal of the American Statistical Association* **76**, 162–168.

Hall, P., DiCiccio, T., and Romano, J. (1989). On Smoothing and the Bootstrap. *Annals of Statistics* **17**, 692–704.

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

Hollander, M. (1967). Asymptotic efficiency of two nonparametric competitors of Wilcoxon's two sample test. *Journal of the American Statistical Association* **62**, 939–949.

Janssen, A. (1997). Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens-Fisher problem. *Statistics and Probability Letters* **36**, 9–21.

Janssen, A. (2005). Resampling student's t-type statistics. *Annals of the Institute of Statistical Mathematics* **57**, 507–529.

Janssen, A. and Pauls, T. (2003). How do bootstrap and permutation tests work? *Annals of Statistics* **31**, 768–806.

Janssen, A. and Pauls, T. (2005). A Monte Carlo comparison of studentized bootstrap and permutation tests for heteroscedastic two-sample problems. *Computational Statistics* **20**, 369–383.

Kowalski, J. and Tu, X. (2008). *Modern Applied U-Statistics.* Wiley, New Jersey.

Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *Annals of Mathematical Statistics* **22**, 165–179.

Lehmann, E. L. (1998). *Nonparametrics: Statistical Methods Based on Ranks.* revised first edition, Prentice Hall, New Jersey.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory.* Springer-Verlag, New York.

Lehmann, E. L. (2009). Parametric versus nonparametrics: two alternative methodologies. *Journal of Nonparametric Statistics* **21**, 397–405.

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

Neubert, K. and Brunner, E. (2007). A Studentized permutation test for the nonparametric Behrens-Fisher problem. *Computational Statistics & Data Analysis* **51**, 5192–5204.

Neuhaus, G. (1993) Conditional Rank Test for the Two-sample Problem under Random Censorship. *Annals of Statistics* **21**, 1760–1779.

Noether, G. E. (1995). On a Theorem of Pitman. *Annals of Mathematical Statistics* **26**, 64–68.

Okeh, U. M. (2009), Statistical analysis of the application of Wilcoxon and Mann-Whitney U test in medical research studies. *Biotechnology and Molecular Biology Reviews* **4**, 128–131.

Pauly, M. (2010). Discussion about the quality of F-ratio resampling tests for comparing variances. *TEST*, 1–17.

Plott, C. and Zeiler, K. (2005), The Willingness to Pay-Willingness to Accept Gap, the "Endowment Effect," Subject Misconceptions, and Experimental Procedures for Eliciting Valuations. *American Economic Review* **95**, 530–545.

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

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

Romano, J. (2009). Discussion of "parametric versus nonparametrics: Two alternative methodologies".

Sausgruber, R. (2009). A note on peer effects between teams. *Experimental Economics* **12**, 193-201.

Sutter, M. (2009). Individual Behavior and Group Membership: Comment. *American Economic Review* **99**, 2247–2257.

van der Vaart, A. W. (1998). *Asymptotic statistics.* Cambridge University Press, New York.

Waldfogel, J. (2005). Does Consumer Irrationality Trump Consumer Sovereignty?, *Review of Economics and Statistics* **87**, 691–696.