Journal of Statistical Planning and Inference 168 (2016) 97–105

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Asymptotically valid and exact permutation tests based on two-sample U-statistics

EunYi Chung[a,*], Joseph P. Romano[b]

[a] Department of Economics, University of Illinois at Urbana–Champaign, United States

[b] Departments of Statistics and Economics, Stanford University, United States

ARTICLE INFO

Article history:
Received 28 June 2015
Received in revised form 14 July 2015
Accepted 15 July 2015
Available online 3 August 2015

Keywords:
Permutation test
U-statistics
Wilcoxon test

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 the Type 1 error under the assumption of identical underlying populations, it in general fails to control the probability of the Type 1 (or Type 3) 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 the Type 1 error in large samples while retaining the exactness property when the underlying distributions are identical. In addition, the new studentized Wilcoxon test 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 in the supplementary appendix). The results are derived for general two-sample U-statistics. A key ingredient that aids our asymptotic derivations is a useful coupling method.

1. Introduction

The two-sample Wilcoxon test has been widely applied in many areas of academic research, including 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 permutation tests have generally been misused in many applications across all disciplines. While the misuse of statistical methods is unfortunately not new, direct application of the classical Wilcoxon test by card carrying statisticians can lead to invalid inference due to lack of Type 1 or Type 3 error control, as explained below. The main goal is less to expose any misuse than to derive a general fix.

To begin, the permutation tests are level α tests even 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

[*] Corresponding author.
E-mail address: eunyi@illinois.edu (E. Chung).

http://dx.doi.org/10.1016/j.jspi.2015.07.004
0378-3758/© 2015 Elsevier B.V. All rights reserved.

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. 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 $T^{(1)}_{m,n}(z) \le T^{(2)}_{m,n}(z) \le \cdots \le T^{(N!)}_{m,n}(z)$ 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 α, $0 < \alpha < 1$, 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 $T^{(j)}_{m,n}(z)$ $(j = 1, \ldots, N!)$ 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

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

Note that, under H_0, $E_{P,Q}[\phi(X_1, \ldots, X_m, Y_1, \ldots, Y_n)] = \alpha$. In other words, the test ϕ has exact size 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 U-statistics.)

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

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

Roughly speaking, if the test statistic $T_{m,n}$ evaluated at the original sample exceeds the $1 - \alpha$ quantile of the permutation distribution, 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 a 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, and Type 1 error is not controlled, even asymptotically. 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 $\theta \le \theta_0$, which is known as a Type 3 or directional error. Indeed, if there are distributions P and Q for which $\theta(P, Q) = \theta_0$ and there is a large probability ($\gg \alpha$) of rejecting the null hypothesis that $\theta = \theta_0$ in favor of $\theta > \theta_0$, then it is easy to "shift" Q by a bit to $\tilde{Q}$ for which $\theta(P, \tilde{Q}) < \theta_0$, but there is still a large probability of rejecting in value of $\theta > \theta_0$. 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, Neuhaus (1993) discovered that if a survival statistic of interest (log-rank statistic) is 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 the Type 1 error in large samples, even if the censoring distributions are 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, 1999, 2005), Janssen and Pauls (2003, 2005), Neubert and Brunner (2007), and Pauly (2010). Our goal is to synthesize the results of the same phenomenon and develop a general theory to a class of two-sample U-statistics, which includes means, variances and the Wilcoxon statistic among 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 the 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, the correct asymptotic rejection probability can be achieved while maintaining the exact finite-sample rejection probability of α when $P = Q$. This exactness property is what makes the proposed permutation procedure more attractive than other asymptotically valid alternatives, such as bootstrap or subsampling (Politis et al., 1999).

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 in which it can be performed—since it is a rank and hence a permutation test, 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. Nevertheless, the Wilcoxon test has been prevalently used for testing equality of means or medians, in which it fails to control the probability of the Type 1 error, even asymptotically.

A general framework is provided 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. We construct a test that controls the asymptotic probability of the Type 1 error in large samples, but still retains the exactness property if $P = Q$. A companion paper by Chung and Romano (2013) 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. Although this class of estimators is quite general, the class of U-statistics not only includes general cases such as comparing means and variances, but also includes 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 $\theta(P) - \theta(Q)$). Thus, the results derived in Chung and Romano (2013) are not directly applicable to 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. However, under the continuity assumption, 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 statistic is indeed a rank statistic, so that the proof of the result becomes quite simple and more importantly, the permutation distribution need not be recomputed with a new data set. As such, we provide the critical values of the new test in the supplementary appendix (see Appendix D). We also prove there that the new test possesses the same Pitman asymptotic relative efficiencies as the classical Wilcoxon test, so that the test is competitive to the classical t-test even under normality in the sense that, while it is slightly less efficient, it is much more robust and can easily outperform the t-test if normality fails.

The asymptotic arguments for establishing the asymptotic behavior of the permutation distribution for U-statistics or their studentized versions is not trivial. While the asymptotic behavior of the true unconditional sampling distribution of U-statistics is classical and can be tackled by means of a Hoeffding projection, for example, the analysis of the permutation distribution is harder. Indeed, permutation distributions (which can be viewed as conditional distributions, but only properly so in the case $P = Q$), are random. Moreover, the sampling scheme of permuting the observations (which are treated as fixed when constructing the permutation distribution) are no longer independent, and so the classical methods do not directly apply. However, the use of a coupling argument helps avoid tedious calculations and permits a clean result.

2. Misapplication of the Wilcoxon test

The two-sample Wilcoxon test has prominently been applied in the field of economics, primarily in experimental economics. Some examples include Feri et al. (2010), Sutter (2009), Charness et al. (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 in which it can be performed; since it is a rank and hence a permutation test, 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,

$$f_Q(y) = f_P(y - \Delta), \quad \Delta \geq 0, \tag{2}$$

where $f_Q(\cdot)$ and $f_P(\cdot)$ denote the densities associated with P and Q, respectively. Such advantageous properties 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. 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 the Type 1 error while maintaining the exactness property in finite samples when $P = Q$.

We illustrate examples of misapplication of the two-sample Wilcoxon test 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, assume that the observations consist of two independent samples $X_1, \ldots, X_m$ i.i.d. P and $Y_1, \ldots, Y_n$ i.i.d. Q. Consider the two-sample Wilcoxon test statistic

$$\hat{\theta} = U_{m,n} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} I(X_i \leq Y_j) - \frac{1}{2}, \tag{3}$$

which can be evidently viewed as an estimator of

$$\theta(P, Q) = P_{P,Q}(X \leq Y) - \frac{1}{2} = 0.$$

Although the two-sample Wilcoxon statistic (3) is most suitable for testing $H_0 : \mathrm{P}(X \leq Y) = \frac{1}{2}$, 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 are theoretically invalid or at least incompatible.

First, consider testing equality of medians. 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 (either bigger or smaller than) 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 $\mathrm{P}(X \leq Y) = \frac{1}{2}$ but in the example here, $\mathrm{P}(X \leq Y) = 0.4431$. 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 $\theta(P, Q) = \mathrm{P}(X \leq Y) - \frac{1}{2} = 0$. (Equivalently, one can test that 0 is a median of the distribution of $Y - X$.)

Similarly, using the Wilcoxon test for testing equality of means may cause an analogous problem. One can easily think of situations where two distributions have the same mean but $\mathrm{P}(X \leq Y) \neq \frac{1}{2}$. In such cases, the rejection probability under the null of $\mu(P) = \mu(Q)$ can be very far from the nominal level α. To illustrate how easily things can go wrong, see Section B in the supplementary appendix (see Appendix D).

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, $\mathrm{P}(X \leq Y) = \frac{1}{2}$ is satisfied and thus, the two-sample Wilcoxon test results in exact control of the probability of the 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 $\mathrm{P}(X \leq Y) = \frac{1}{2}$, 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 test to examine the null hypothesis that willingness to pay (WTP) and willingness to accept (WTA) are drawn from the same distribution. The estimated densities of WTP denoted P and of WTA denoted Q are depicted in Fig. S2 in the supplementary appendix (see Appendix D). In their analysis for experiment 3, the Wilcoxon test yields a z value of 1.738 (p-value $=$ 0.0821), resulting in a failure to reject H_0. 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 test yields a p-value of 0.0546.

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 incongruous with what the Wilcoxon test is actually testing ($H_0 : \mathrm{P}(X \leq Y) = \frac{1}{2}$). However, as will be argued in a more general setting in Section 3, even when testing $H_0 : \mathrm{P}(X \leq Y) = \frac{1}{2}$, the standard Wilcoxon test is invalid *unless* it is appropriately studentized. As such, we propose to use the new two-sample studentized Wilcoxon test based on correctly studentized statistic as it attains the rejection probability equal to α asymptotically while still maintaining exact control in finite samples if $P = Q$. The new two-sample studentized Wilcoxon statistic under the continuity assumption warrants special attention and we will briefly examine this case in Section 4 while providing more details with an example of how to implement the new test in the supplementary appendix (see Appendix D).

3. General two-sample *U*-statistics

In this section, the problem regarding the two-sample Wilcoxon statistic considered in Section 2 is 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 the Type 1 error in large samples while retaining the exact control of the 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

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

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.

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

$$\varphi(X_{\alpha_1}, \ldots, X_{\alpha_r}, Y_{\beta_1}, \ldots, Y_{\beta_r}) = -\varphi(Y_{\beta_1}, \ldots, Y_{\beta_r}, X_{\alpha_1}, \ldots, X_{\alpha_r}). \tag{4}$$

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

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

where $\Phi(\cdot)$ is the standard normal cdf and

$$\bar{\tau}^2 = r^2 \left(\mathrm{E}\varphi^2_{.\bar{P}^{r-1}\bar{P}^r}(\bar{Z}_i) + \frac{p}{1-p}\mathrm{E}\varphi^2_{.\bar{P}^{r-1}\bar{P}^r}(\bar{Z}_i) \right) = \frac{r^2}{1-p}\mathrm{E}\varphi^2_{.\bar{P}^{r-1}\bar{P}^r}(\bar{Z}_i), \tag{5}$$

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

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

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

where $\varphi_{.P^{r-1}Q^r(a_1)} \equiv \int \cdots \int \varphi(a_1, \ldots, a_r, b_1, \ldots, b_r) dP(a_2) \cdots dP(a_r) dQ(b_1) \cdots dQ(b_r)$ and $\varphi_{.P^r Q^{r-1}(b_1)} \equiv \int \cdots \int \varphi(a_1, \ldots, a_r, b_1, \ldots, b_r) dP(a_1) \cdots dP(a_r) dQ(b_2) \cdots dQ(b_r)$. Note that this asymptotic variance in general does not equal $\bar{\tau}^2$ defined in (5). See the supplementary appendix for examples of U-statistics.

Remark 3.2. The antisymmetry assumption (4) 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 C.1. However, the antisymmetry assumption is not just one of conveniences because, without it, the results do not hold. See Remark C.1 in the supplementary appendix.

The following theorem shows how studentization leads to asymptotic validity while still maintaining exactness property in finite samples if $P = Q$.

Theorem 3.2. *Assume the same setup and conditions of Theorem 3.1. Further assume that $\hat{\sigma}^2_m(X_1, \ldots, X_m)$ 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 $\hat{\sigma}^2_n(Y_1, \ldots, Y_n)$ 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., $\hat{\sigma}^2_m(\bar{Z}_1, \ldots, \bar{Z}_m)$ 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

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

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

$$\sup_t |\hat{R}^S_{m,n}(t) - \Phi(t)| \xrightarrow{P} 0. \tag{6}$$

Under the conditions of Theorem 3.2, the permutation distribution of $S_{m,n}$ is asymptotically standard normal, which is the same as the limiting sampling distribution. 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 $P = Q$. Examples of the studentized U-statistics are provided in the supplementary appendix (see Appendix D).

4. New two-sample studentized Wilcoxon test

As suggested in Theorem 3.2, for testing $H_0 : \mathrm{P}(X \le Y) = \frac{1}{2}$ without imposing the equal distribution assumption under the null, the most suitable test statistic to be considered when the underlying distributions are continuous is

$$\tilde{U}_{m,n} = \frac{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \mathrm{I}\{X_i \le Y_j\} - \frac{1}{2}}{\sqrt{\frac{1}{m}\hat{\xi}_x + \frac{1}{n}\hat{\xi}_y}},$$

where

$$\hat{\xi}_x = \frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n}\mathrm{I}\{Y_j \le X_i\} - \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n}\mathrm{I}\{Y_j \le X_i\}\right)\right)^2$$

and

$$\hat{\xi}_y = \frac{1}{n-1}\sum_{j=1}^{n}\left(\frac{1}{m}\sum_{i=1}^{m}\mathrm{I}\{X_i < Y_j\} - \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{m}\sum_{i=1}^{m}\mathrm{I}\{X_i < Y_j\}\right)\right)^2.$$

The new studentized Wilcoxon test has the same asymptotic Pitman efficiency (Noether, 1995) as the standard Wilcoxon test when the underlying distributions are equal (see the supplementary appendix for more detail). Moreover, it is crucial to realize that the estimators $\hat{\xi}_x$ and $\hat{\xi}_y$ 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 $S_1 < S_2 < \cdots < S_m$ are the ordered ranks of the Xs in the combined sample. Similarly, ξ_y can be expressed as a function of the ordered ranks of the Ys in the combined sample. The fact that the standard error estimate is a rank statistic renders much simpler proofs (as provided in the supplementary appendix, see Appendix D) and more importantly, it allows the new studentized 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 S3 in the supplementary appendix (see Appendix D). We also provide an example of how to carry out the new studentized Wilcoxon test in the appendix.

5. Conclusion

Although permutation tests are useful tools in obtaining exact level α in finite samples under the fundamental assumption of identical underlying distributions, they lack robustness of validity against inequality of distributions. If the underlying distributions are not identical, the usual permutation test can fail to control the probability of the Type 1 error, even asymptotically. Thus, a careful interpretation of a rejection of the permutation test is necessary; rejection 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 $P = Q$, a 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 the 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 be tabled as displayed in Table S3. Also, it achieves the same asymptotic relative efficiency compared to the two-sample t-test as the usual Wilcoxon test. As such, it loses very little efficiency compared to the t-test if the underlying distributions are normal, but is much more efficient in other scenarios; see Lehmann (2009). Now that a test with all the simple and desirable properties of the classical Wilcoxon test has been constructed, but for which $P = Q$ is not naively assumed, inference can proceed in a safe and valid way.

More generally when testing $\theta(P, Q) = \theta_0$, as long as a U-statistic is studentized by a consistent standard error, the permutation test based on the studentized U-statistic controls the asymptotic probability of the 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.

Appendix A. 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. Second, 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

$$p - \frac{m}{N} = O\left(\frac{1}{\sqrt{N}}\right). \tag{7}$$

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 is 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 $\bar{Z}$.

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

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

or it is of the form

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

Using this coupling method, if we can show (8) in Lemma B.1, 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.

Appendix B. 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 Appendix A.

Lemma B.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 $\{1, \dots, m\}$ and of s different elements chosen from $\{1, \dots, n\}$, respectively. Assume $E_{P,Q}\varphi(\cdot) = 0$ and $0 < E_{P,Q}\varphi^2(\cdot) < \infty$ for any permutation of $r + s$ X s and Y s. 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 (7) holds. Also, let $\bar{Z}$ and π_0 be constructed by the coupling method. Then, for any fixed permutation $\pi = (\pi(1), \dots, \pi(N))$ of $\{1, \dots, N\}$,

$$\sqrt{m}U_{m,n}(Z_\pi) - \sqrt{m}U_{m,n}(\bar{Z}_{\pi\cdot\pi_0}) \xrightarrow{P} 0. \tag{8}$$

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

Lemma B.2. *Assume the same setup and conditions of Lemma B.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,

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

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

Appendix C. Proofs of theorems

Proof of Theorem 3.1. Independent of the Zs, let π and π' be independent permutations of $\{1, \dots, N\}$. 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 (5). However, if we can show

$$\left(\sqrt{m}\tilde{U}_{m,n}(\bar{Z}_\pi), \sqrt{m}\tilde{U}_{m,n}(\bar{Z}_{\pi'})\right) \xrightarrow{d} (T, T'), \tag{9}$$

then, by Lemma B.2 and Remark B.1, (9) implies

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

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

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

which also implies by Lemma B.1 that

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

To show (9), note that the antisymmetry assumption of φ defined by (4) 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

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

where $W_i = 1$ if $\pi(i) \le m$ and $-\frac{m}{n}$ otherwise. That is, this problem is reduced to the general mean case based on observations $\varphi_{.\bar{P}^{r-1}\bar{P}^r}(\bar{Z}_1), \dots, \varphi_{.\bar{P}^{r-1}\bar{P}^r}(\bar{Z}_N)$ instead of $Z_1, \dots, Z_N$. Thus, it follows by Theorem 15.2.5 of Lehmann and Romano (2005) that $\left(\sqrt{m}\tilde{U}_{m,n}(\bar{Z}_\pi), \sqrt{m}\tilde{U}_{m,n}(\bar{Z}_{\pi'})\right)$ converges in distribution to a bivariate normal distribution with independent, identically distributed marginals having mean 0 and variances given by (5). ■

Proof of Theorem 3.2. We shall first show that $V^2_{m,n}(Z_{\pi(1)}, \dots, Z_{\pi(N)})$ is a consistent estimator for $\bar{\tau}^2$, i.e., $V^2_{m,n}(Z_{\pi(1)}, \dots, Z_{\pi(N)}) \xrightarrow{P} \bar{\tau}^2$, where $\bar{\tau}^2$ is defined in (5). To do so, it suffices to show that

$$\hat{\sigma}^2_m(Z_{\pi(1)}, \dots, Z_{\pi(m)}) \xrightarrow{P} \int \varphi^2_{.\bar{P}^{r-1}\bar{P}^r}\, d\bar{P} \tag{10}$$

and

$$\hat{\sigma}_n^2(Z_{\pi(m+1)}, \ldots, Z_{\pi(N)}) \xrightarrow{P} \int \varphi^2_{.\bar{P}r-1\bar{P}r} \, d\bar{P}. \tag{11}$$

However, (10) and (11) follow from a key contiguity argument for the binomial and hypergeometric distributions shown in Lemmas 3.2 and 3.3 of Chung and Romano (2013). 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, 2013), $\hat{R}^V_{m,n}(t)$ 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 Theorem 3.2 of Chung and Romano (2013) again together with Theorem 3.1 to conclude that the permutation distribution of $\sqrt{m}S_{m,n}$ satisfies (6). ■

Appendix D. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jspi.2015.07.004.

References

Charness, G., Rigotti, L., Rustichini, A., 2007. Individual behavior and group membership. Amer. Econ. Rev. 97, 1340–1352.

Chung, E., Romano, J., 2013. Exact and asymptotically robust permutation tests. Ann. Statist. 41, 484–507.

Davis, J., 2004. An annual index of US industrial production, 1790–1915. Quart. J. Econ. 119, 1177–1215.

Feri, F., Irlenbusch, B., Sutter, M., 2010. Efficiency gains from team-based coordination—large-scale experimental evidence. Amer. Econ. Rev. 100, 1892–1912.

Janssen, A., 1997. Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens–Fisher problem. Statist. Probab. Lett. 36, 9–21.

Janssen, A., 1999. Testing nonparametric statistical functionals with application to rank tests. J. Statist. Plann. Inference 81, 71–93. Erratum 92 (2001) 297.

Janssen, A., 2005. Resampling student's t-type statistics. Ann. Inst. Statist. Math. 57, 507–529.

Janssen, A., Pauls, T., 2003. How do bootstrap and permutation tests work? Ann. Statist. 31, 768–806.

Janssen, A., Pauls, T., 2005. A Monte Carlo comparison of studentized bootstrap and permutation tests for heteroscedastic two-sample problems. Comput. Statist. 20, 369–383.

Lehmann, E.L., 2009. Parametric versus nonparametrics: two alternative methodologies. J. Nonparametr. Stat. 21, 397–405.

Lehmann, E.L., Romano, J., 2005. Testing Statistical Hypotheses, third ed. Springer-Verlag, New York.

Neubert, K., Brunner, E., 2007. A studentized permutation test for the non-parametric Behrens–Fisher problem. Comput. Statist. Data Anal. 51, 5192–5204.

Neuhaus, G., 1993. Conditional rank test for the two-sample problem under random censorship. Ann. Statist. 21, 1760–1779.

Noether, G.E., 1995. On a theorem of pitman. Ann. Math. Statist. 26, 64–68.

Okeh, U.M., 2009. Statistical analysis of the application of Wilcoxon and Mann–Whitney U test in medical research studies. Biotechnol. Mol. Biol. Rev. 4, 128–131.

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

Plott, C., Zeiler, K., 2005. The willingness to pay-willingness to accept gap, the "endowment effect," subject misconceptions, and experimental procedures for eliciting valuations. Amer. Econ. Rev. 95, 530–545.

Politis, D., Romano, J., Wolf, M., 1999. Subsampling. Springer-Verlag, New York.

Romano, J., 1990. On the behavior of randomization tests without a group invariance assumption. J. Amer. Statist. Assoc. 85, 686–692.

Sausgruber, R., 2009. A note on peer effects between teams. Exp. Econ. 12, 193–201.

Sutter, M., 2009. Individual behavior and group membership: Comment. Amer. Econ. Rev. 99, 2247–2257.