The Annals of Statistics
2013, Vol. 41, No. 2, 484–507
DOI: 10.1214/13-AOS1090

EXACT AND ASYMPTOTICALLY ROBUST PERMUTATION TESTS[1]

BY EUNYI CHUNG AND JOSEPH P. ROMANO

Stanford University

Given independent samples from P and Q, two-sample permutation tests allow one to construct exact level tests when the null hypothesis is $P = Q$. On the other hand, when comparing or testing particular parameters θ of P and Q, such as their means or medians, permutation tests need not be level α, or even approximately level α in large samples. Under very weak assumptions for comparing estimators, we provide a general test procedure whereby the asymptotic validity of the permutation test holds while retaining the *exact* rejection probability α in finite samples when the underlying distributions are identical. The ideas are broadly applicable and special attention is given to the k-sample problem of comparing general parameters, whereby a permutation test is constructed which is exact level α under the hypothesis of identical distributions, but has asymptotic rejection probability α under the more general null hypothesis of equality of parameters. A Monte Carlo simulation study is performed as well. A quite general theory is possible based on a coupling construction, as well as a key contiguity argument for the multinomial and multivariate hypergeometric distributions.

1. Introduction. In this article, we consider the behavior of two-sample (and later also k-sample) permutation tests for testing problems when the fundamental assumption of identical distributions need not hold. Assume $X_1, \dots, X_m$ are i.i.d. according to a probability distribution P, and independently, $Y_1, \dots, Y_n$ are i.i.d. Q. The underlying model specifies a family of pairs of distributions (P, Q) in some space Ω. For the problems considered here, Ω specifies a nonparametric model, such as the set of all pairs of distributions. Let $N = m + n$, and write

$$Z = (Z_1, \dots, Z_N) = (X_1, \dots, X_m, Y_1, \dots, Y_n). \tag{1.1}$$

Let $\bar{\Omega} = \{(P, Q) : P = Q\}$. Under the assumption $(P, Q) \in \bar{\Omega}$, the joint distribution of $(Z_1, \dots, Z_N)$ is the same as $(Z_{\pi(1)}, \dots, Z_{\pi(N)})$, where $(\pi(1), \dots, \pi(N))$ is any permutation of $\{1, \dots, N\}$. It follows that, when testing any null hypothesis $H_0 : (P, Q) \in \Omega_0$, where $\Omega_0 \subset \bar{\Omega}$, then an exact level α test can be constructed by a permutation test. To review how, let $\mathbf{G}_N$ denote the set of all permutations π of $\{1, \dots, N\}$. Then, given any test statistic $T_{m,n} = T_{m,n}(Z_1, \dots, Z_N)$, recompute $T_{m,n}$ for all permutations π; that is, compute $T_{m,n}(Z_{\pi(1)}, \dots, Z_{\pi(N)})$ for all

---

Received July 2012; revised December 2012.

[1] Supported by NSF Grant DMS-07-07085.

MSC2010 subject classifications. Primary 62E20; secondary 62G10.

Key words and phrases. Behrens–Fisher problem, coupling, permutation test.

$\pi \in \mathbf{G}_N$, and let their ordered values be

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

Fix a nominal level α, $0 < \alpha < 1$, and let k be defined by $k = N! - [\alpha N!]$, where $[\alpha N!]$ denotes the largest integer less than or equal to $\alpha N!$. Let $M^+(z)$ and $M^0(z)$ be the number of values $T_{m,n}^{(j)}(z)$ $(j = 1, \ldots, N!)$ which are greater than $T^{(k)}(z)$ and equal to $T^{(k)}(z)$, respectively. Set

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

Define the randomization test function $\phi(Z)$ to be equal to 1, $a(Z)$ or 0 according to whether $T_{m,n}(Z) > T_{m,n}^{(k)}(Z)$, $T_{m,n}(X) = T^{(k)}(Z)$ or $T_{m,n}(Z) < T^{(k)}(Z)$, respectively. Then, under any $(P, Q) \in \bar{\Omega}$,

$$E_{P,Q}\bigl[\phi(X_1, \ldots, X_m, Y_1, \ldots, Y_n)\bigr] = \alpha.$$

Also, define the permutation distribution as

$$\hat{R}_{m,n}^T(t) = \frac{1}{N!} \sum_{\pi \in \mathbf{G}_N} I\bigl\{T_{m,n}(Z_{\pi(1)}, \ldots, Z_{\pi(N)}) \le t\bigr\}. \tag{1.2}$$

Roughly speaking (after accounting for discreteness), the permutation test rejects H_0 if the test statistic $T_{m,n}$ exceeds $T_{m,n}^{(k)}$, or a $1 - \alpha$ quantile of this permutation distribution.

It may be helpful to consider an alternative description of the permutation distribution given in (1.2). As a shorthand, for any $\pi \in \mathbf{G}_N$, let $Z_\pi = (Z_{\pi(1)}, \ldots, Z_{\pi(N)})$. Let Π denote a random permutation, uniformly distributed over $\mathbf{G}_N$. Then, $T_{m,n}(Z_\Pi)$ denotes the random variable that evaluates the test statistic, not at the original data Z, but at a randomly permuted data set Z_Π. The permutation distribution $\hat{R}_{m,n}^T(\cdot)$ given in (1.2) is evidently the conditional distribution of $T_{m,n}(Z_\Pi)$ given Z, because conditional on the data Z, $T_{m,n}(Z_\Pi)$ is equally likely to be any of $T_{m,n}(Z_\pi)$ among $\pi \in \mathbf{G}_N$. The asymptotic behavior of this (conditional) distribution $\hat{R}_{m,n}^T(\cdot)$ is the key to establishing properties of the permutation test.

Although the rejection probability of the permutation test is exactly α when $P = Q$, problems arise if Ω_0 is strictly bigger than $\bar{\Omega}$. Since a transformed permuted data set no longer has the same distribution as the original data set, the argument leading to the construction of an α level test fails, and faulty inferences can occur.

To be concrete, if we are interested in testing equality of means, for example, then $\Omega_0 = \{(P, Q) : \mu(P) = \mu(Q)\}$ which, of course, is strictly bigger than $\bar{\Omega}$. So, consider constructing a permutation test based on the difference of sample means

$$T_{m,n} = \sqrt{N}(\bar{X}_m - \bar{Y}_n). \tag{1.3}$$

Note that we are not taking the absolute difference, so that the test is one-sided, as we are rejecting for large positive values of the difference. First of all, we are not concerned about testing $\bar{\Omega} = \{(P, Q) : P = Q\}$, but something bigger than $\bar{\Omega}$. However, we underscore the point that a test statistic (1.3) is not appropriate for testing $\bar{\Omega}$ without further assumptions because the test clearly will not have any power against distributions P and Q whose means are identical but $P \neq Q$.

The permutation test based on the difference of sample means is only appropriate as a test of equality of population means. However, the permutation test no longer controls the level of the test, even in large samples. As is well known (Romano [23]), the permutation test possesses a certain asymptotic robustness as a test of difference in means if $m/n \to 1$ as $n \to \infty$, or the underlying variances of P and Q are equal, in the sense that the rejection probability under the null hypothesis of equal means tends to the nominal level. Without equal variances or comparable sample sizes, the rejection probability can be much larger than the nominal level, which is a concern. Because of the lack of robustness and the increased probability of a type 1 error, rejection of the null may incorrectly be interpreted as rejection of equal means, when in fact it is caused by unequal variances and unequal sample sizes. Even more alarming is the possibility of rejecting a two-sided null hypothesis when observing a positive large difference with the accompanying inference that mean difference is positive when in fact the difference in means is negative, a type 3 error or directional error. Indeed, if for some P and Q with equal means the rejection probability is, say, $\gamma \gg \alpha$, then it follows by continuity that the rejection probability under some P and Q with negative mean difference will be nearly γ as well, where one would conclude that the mean difference is actually positive. Further note that there is also the possibility that the rejection probability can be much less than the nominal level, which by continuity implies the test is biased and has little power of detecting a true difference in means, or large type 2 error.

The situation is even worse when basing a test on a difference in sample medians, in the sense that regardless of sample sizes, the asymptotic rejection probability of the permutation test will be α under very stringent conditions, which essentially means only in the case where the underlying distributions are the same.

However, in a very insightful paper in the context of random censoring models, Neuhaus [18] realized that by proper studentization of a test statistic, the permutation test can result in asymptotically valid inference even when the underlying distributions are not the same. This result has been extended to other specific problems, such as comparing means by Janssen [9] and certain linear statistics in Janssen [10] (including the Wilcoxon statistic without ties), variances by Pauly [20] and the two-sample Wilcoxon test by Neubert and Brunner [17] (where ties are allowed). Other results on permutation tests are presented in Janssen [11], Janssen and Pauls [12], Janssen and Pauls [13] and Janssen and Pauly [14]. The recent paper by Omelka and Pauly [19] compares correlations by permutation tests, which is a special case of our general results. Note that the importance of studentization when bootstrapping is well known; see Hall and Wilson [7] and Delaigle et al. [3] (though its role for bootstrap is to obtain higher order accuracy while in the context here first order accuracy can fail without studentization).

The goal of this paper is to obtain a quite general result of the same phenomenon. That is, when basing a permutation test using some test statistic as a test of a parameter (usually a difference of parameters associated with marginal distributions), we would like to retain the exactness property when $P = Q$, and also have the asymptotic rejection probability be α for the more general null hypothesis specifying the parameter (such as the difference being zero). Of course, there are many alternatives to getting asymptotic tests, such as the bootstrap or subsampling. However, we do not wish to give up the exactness property under $P = Q$, and resampling methods do not have such finite sample guarantees. The main problem becomes: what is the asymptotic behavior of $\hat{R}^T_{m,n}(\cdot)$ defined in (1.2) for general test statistic sequences $T_{m,n}$ when the underlying distributions differ. Only for suitable test statistics is it possible to achieve both finite sample exactness when the underlying distributions are equal, but also maintain a large sample rejection probability near the nominal level when the underlying distributions need not be equal. In this sense, our results are both exact and asymptotically robust for heterogeneous populations.

This paper provides a framework for testing a parameter that depends on P and Q (and later on k underlying distributions P_i for $i = 1, \dots, k$). We construct a general test procedure where the asymptotic validity of the permutation test holds in a general setting. Assuming that estimators are asymptotically linear and consistent estimators are available for their asymptotic variance, we provide a test that has asymptotic rejection probability equal to the nominal level α, but still retains the exact rejection probability of α in finite samples if $P = Q$ in Section 2. It is not even required that the estimators are based on differentiable functionals, and some methods like the bootstrap would not necessarily be even asymptotically valid under such conditions, let alone retain the finite sample exactness property when $P = Q$. In Section 3, generalizations of the results are discussed with a special attention to the more general k-sample problem of comparing general parameters. Furthermore, Monte Carlo simulation studies illustrating our results are presented in Section 4. The arguments of the paper are quite different from Janssen and previous authors, and hold under great generality. For example, they immediately apply to comparing means, variances or medians. The key idea is to show that the permutation distribution behaves like the unconditional distribution of the test statistic when all N observations are i.i.d. from the mixture distribution $pP + (1 - p)Q$, where p is such that $m/N \to p$. This seems intuitive because the permutation distribution permutes the observations so that a permuted sample is almost like a sample from the mixture distribution. In order to make this idea precise, a coupling argument is given in Section 5.3. Of course, the permutation distribution depends on all permuted samples (for a given original data set). But even for one permuted data set, it cannot exactly be viewed as a sample from $pP + (1 - p)Q$. Indeed, the first m observations from the mixture would include B_m observations from P

and the rest from Q, where B_m has the binomial distribution based on m trials and success probability p. On the other hand, for a permuted sample, if H_m denotes the number of observations from P, then H_m has the hypergeometric distribution with mean mp. The key argument that allows for such a general result concerns the contiguity of the distributions of B_m and H_m. Section 5 highlights the main technical ideas required for the proofs. All proofs are deferred to the supplementary appendix [2].

2. Robust studentized two-sample test.

In this section, we consider the general problem of inference from the permutation distribution when comparing parameters from two populations. Specifically, assume $X_1, \dots, X_m$ are i.i.d. P and, independently, $Y_1, \dots, Y_n$ are i.i.d. Q. Let $\theta(\cdot)$ be a real-valued parameter, defined on some space of distributions $\mathcal{P}$. The problem is to test the null hypothesis

$$H_0 : \theta(P) = \theta(Q). \tag{2.1}$$

Of course, when $P = Q$, one can construct permutation tests with exact level α. Unfortunately, if $P \neq Q$, the test need not be valid in the sense that the probability of a type 1 error need not be α even asymptotically. Thus, our goal is to construct a procedure that has asymptotic rejection probability equal to α quite generally, but also retains the exactness property in finite samples when $P = Q$.

We will assume that estimators are available that are asymptotically linear. Specifically, assume that, under P, there exists an estimator $\hat{\theta}_m = \hat{\theta}_m(X_1, \dots, X_m)$ which satisfies

$$m^{1/2}[\hat{\theta}_m - \theta(P)] = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} f_P(X_i) + o_P(1). \tag{2.2}$$

Similarly, we assume that, based on the Y_j (under Q),

$$n^{1/2}[\hat{\theta}_n - \theta(Q)] = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} f_Q(Y_j) + o_Q(1). \tag{2.3}$$

The functions determining the linear approximation f_P and f_Q can of course depend on the underlying distributions. Different forms of differentiability guarantee such linear expansions in the special case when $\hat{\theta}_m$ takes the form of an empirical estimate $\theta(\hat{P}_m)$, where $\hat{P}_m$ is the empirical measure constructed from $X_1, \dots, X_m$, but we will not need to assume such stronger conditions. We will argue that our assumptions of asymptotic linearity already imply a result about the permutation distribution corresponding to the statistic $N^{1/2}[\hat{\theta}_m(X_1, \dots, X_m) - \hat{\theta}_n(Y_1, \dots, Y_n)]$, without having to impose any differentiability assumptions. However, we will assume the expansion (2.2) holds not just for i.i.d. samples under P, and also under Q, but also when sampling i.i.d. observations from the mixture distribution $\bar{P} = pP + qQ$. This is a weak assumption and replaces having to study the permutation distribution based on variables that are no longer independent nor identically distributed with a simple assumption about the behavior under an i.i.d. sequence. Indeed, we will argue that in all cases, the permutation distribution behaves asymptotically like the unconditional limiting sampling distribution of the studied statistic sequence when sampling i.i.d. observations from $\bar{P}$.

In the next two theorems, the behavior of the permutation distribution is obtained. Note that it is not assumed that the null hypothesis $\theta(P) = \theta(Q)$ necessarily holds. Indeed, the asymptotic behavior of the permutation test under P and Q is the same as when all observations are from the mixture distribution $\bar{P} = pP + (1 - p)Q$, where $p = \lim \frac{m}{N}$. Proofs of all the results in Section 2 are presented along with proofs of the results in Section 5 in the supplementary appendix [2].

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. Consider testing the null hypothesis* (2.1) *based on a test statistic of the form*

$$T_{m,n} = N^{1/2}[\hat{\theta}_m(X_1, \dots, X_m) - \hat{\theta}_n(Y_1, \dots, Y_n)],$$

where the estimators satisfy (2.2) *and* (2.3). *Further assume $E_P f_P(X_i) = 0$ and*

$$0 < \mathrm{E}_P f_P^2(X_i) \equiv \sigma^2(P) < \infty$$

and the same with P replaced by Q. Let $m \to \infty$, $n \to \infty$, with $N = m + n$, $p_m = m/N$, $q_m = n/N$ and $p_m \to p \in (0, 1)$ with

$$p_m - p = O(N^{-1/2}). \tag{2.4}$$

Assume the estimator sequence also satisfies (2.2) *with P replaced by $\bar{P} = pP + qQ$ with $\sigma^2(\bar{P}) < \infty$.*

Then the permutation distribution of $T_{m,n}$ given by (1.2) *satisfies*

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

where

$$\tau^2(\bar{P}) = \frac{1}{p(1-p)}\sigma^2(\bar{P}). \tag{2.5}$$

REMARK 2.1. Under H_0 given by (2.1), the true unconditional sampling distribution of $T_{m,n}$ is asymptotically normal with mean 0 and variance

$$\frac{1}{p}\sigma^2(P) + \frac{1}{1-p}\sigma^2(Q), \tag{2.6}$$

which does not equal $\tau^2(\bar{P})$ defined by (2.5) in general.

EXAMPLE 2.1 (Difference of means). As is well known, even for the case of comparing population means by sample means, under the null hypothesis that $\theta(P) = \theta(Q)$, equality of (2.5) and (2.6) holds if and only if $p = 1/2$ or $\sigma^2(P) = \sigma^2(Q)$.

EXAMPLE 2.2 (Difference of medians). Let F and G denote the c.d.f.s corresponding to P and Q. Let $\theta(F)$ denote the median of F, that is, $\theta(F) = \inf\{x : F(x) \geq \frac{1}{2}\}$. Then it is well known (Serfling [24]) that if F is continuously differentiable at $\theta(P)$ with derivative F' (and the same with F replaced by G), then

$$m^{1/2}[\theta(\hat{P}_m) - \theta(P)] = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \frac{1/2 - \mathrm{I}\{X_i \leq \theta(P)\}}{F'(\theta(P))} + o_P(1)$$

and similarly, with P and F replaced by Q and G. Thus, we can apply Theorem 2.1 and conclude that, when $\theta(P) = \theta(Q) = \theta$, the permutation distribution of $T_{m,n}$ is approximately a normal distribution with mean 0 and variance

$$\frac{1}{4p(1-p)[pF'(\theta) + (1-p)G'(\theta)]^2}$$

in large samples. On the other hand, the true sampling distribution is approximately a normal distribution with mean 0 and variance

$$v^2(P, Q) \equiv \frac{1}{p} \frac{1}{4[F'(\theta)]^2} + \frac{1}{1-p} \frac{1}{4[G'(\theta)]^2}. \tag{2.7}$$

Thus the permutation distribution and the true unconditional sampling distribution behave differently asymptotically unless $F'(\theta) = G'(\theta)$ is satisfied. Since we do not assume $P = Q$, this condition is a strong assumption. Hence, the permutation test for testing equality of medians is generally not valid in the sense that the rejection probability tends to a value that is far from the nominal level α.

The main goal now is to show how studentizing the test statistic leads to a general correction.

THEOREM 2.2. *Assume the setup and conditions of Theorem* 2.1. *Further assume that $\hat{\sigma}_m(X_1, \ldots, X_m)$ is a consistent estimator of $\sigma(P)$ when $X_1, \ldots, X_m$ are i.i.d. P. Assume consistency also under Q and $\bar{P}$, so that $\hat{\sigma}_n(V_1, \ldots, V_n) \xrightarrow{P} \sigma(\bar{P})$ as $n \to \infty$ when the V_i are i.i.d. $\bar{P}$. Define the studentized test statistic*

$$S_{m,n} = \frac{T_{m,n}}{V_{m,n}}, \tag{2.8}$$

where

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

and consider the permutation distribution defined in (1.2) *with* T *replaced by* S. *Then*

$$\sup_t \left|\hat{R}^S_{m,n}(t) - \Phi(t)\right| \xrightarrow{P} 0. \tag{2.9}$$

Thus the permutation distribution is asymptotically standard normal, as is the true unconditional limiting distribution of the test statistics $S_{m,n}$. Indeed, as mentioned in Remark 2.1, the true unconditional limiting distribution of $T_{m,n}$ is normal with mean 0 and variance given by (2.6). But, when sampling m observations from P and n from Q, $V^2_{m,n}$ tends in probability to (2.6), and hence the limiting distribution of $T_{m,n}$ is standard normal, the same as that of the permutation distribution.

REMARK 2.2. As previously noted, Theorems 2.1 and 2.2 are true even if $\theta(P) \neq \theta(Q)$. If $\theta(P) = \theta(Q)$, then the true sampling distribution of $S_{m,n}$ and the permutation test become approximately the same. However, if $\theta(P) \neq \theta(Q)$, then we get the power tending to 1. Indeed, the critical value from the permutation distribution asymptotically tends to a finite value $z_{1-\alpha}$ in probability, while the test statistic tends to infinity in probability. Also, see Remark 2.3 for local power.

EXAMPLE 2.1 (Continued). As proved by Janssen [9], even when the underlying distributions may have different variances and different sample sizes, permutation tests based on studentized statistics

$$S_{m,n} = \frac{N^{1/2}(\bar{X}_m - \bar{Y}_n)}{\sqrt{NS^2_X/m + NS^2_Y/n}},$$

where $S^2_X = \frac{1}{m-1}\sum_{i=1}^m (X_i - \bar{X}_m)^2$ and $S^2_Y = \frac{1}{n-1}\sum_{j=1}^n (Y_i - \bar{Y}_m)^2$, can allow one to construct a test that attains asymptotic rejection probability α when $P \neq Q$ while providing an additional advantage of maintaining *exact* level α when $P = Q$.

EXAMPLE 2.2 (Continued). Define the studentized median statistic

$$S_{m,n} = \frac{N^{1/2}[\theta(\hat{P}_m) - \theta(\hat{Q}_n)]}{\hat{v}_{m,n}},$$

where $\hat{v}_{m,n}$ is a consistent estimator of $v(P, Q)$ defined in (2.7). There are several choices for a consistent estimator of $v(P, Q)$. Examples include the usual kernel estimator (Devroye and Wagner [4]), bootstrap estimator (Efron [5]), and the smoothed bootstrap (Hall, DiCiccio, and Romano [6]).

REMARK 2.3. Suppose that the true unconditional distribution of a test $T_{m,n}$ is, under the null hypothesis, asymptotically given by a distribution $R(\cdot)$. Typically a test rejects when $T_{m,n} > r_{m,n}$, where $r_{m,n}$ is nonrandom, as happens in many classical settings. Then, we typically have $r_{m,n} \to r(1-\alpha) \equiv R^{-1}(1-\alpha)$. Assume that $T_{m,n}$ converges to some limit law $R'(\cdot)$ under some sequence of alternatives which are contiguous to some distribution satisfying the null. Then, the power of the test against such a sequence would tend to $1 - R'(r(1-\alpha))$. The point here is that, under the conditions of Theorem 2.2, the permutation test based on a random critical value $\hat{r}_{m,n}$ obtained from the permutation distribution satisfies, under the null, $\hat{r}_{m,n} \xrightarrow{P} r(1-\alpha)$. But then, contiguity implies the same behavior under a sequence of contiguous alternatives. Thus, the permutation test has the same limiting local power as the "classical" test which uses the nonrandom critical value. So, to first order, there is no loss in power in using a permutation critical value. Of course, there are big gains because the permutation test applies much more broadly than for usual parametric models, in that it retains the level exactly across a broad class of distributions and is at least asymptotically justified for a large nonparametric family.

3. Generalizations.

3.1. *Wilcoxon statistic and general U-statistics.* So far, we considered two-sample problems where the statistic is based on the difference of estimators that are asymptotically linear. Although this class of estimators includes many interesting cases such as testing equality of means, medians, and variances, it does not include other important statistics like the Wilcoxon statistic or some rank statistics where the parameter of interest is a function of the joint distribution $\theta(P, Q)$ and not just a simple difference $\theta(P) - \theta(Q)$.

In our companion paper (Chung and Romano [1]), however, we consider these statistics in a more general U-statistic framework. More specifically, assume that $X_1, \dots, X_m$ are i.i.d. P, and independently, $Y_1, \dots, Y_n$ are i.i.d. Q. The problem studied is to test the null hypothesis

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

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

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

where α and β range over the sets of all unordered subsets of r different elements chosen from $\{1, \dots, m\}$ and of r different elements chosen from $\{1, \dots, n\}$, respectively.

This general class of U-statistics covers, for example, Lehmann's two-sample U-statistic to test $H_0 : P(|Y' - Y| > |X' - X|) = 1/2$, the two-sample Wilcoxon statistic to test $H_0 : P(X \leq Y) = P(Y \leq X)$, and some other interesting rank statistics. Under quite weak assumptions, 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 in finite samples when the underlying distributions are identical. The technical arguments involved in this U-statistic problem are different from Section 2, but the mathematics and statistical foundations to be laid out in Section 5 provide fundamental ingredients that aid our asymptotic derivations.

3.2. *Robust k-sample test.* From our general considerations, we are now guided by the principle that the large sample distribution of the test statistic should not depend on the underlying distributions; that is, it should be asymptotically pivotal under the null. Of course, it can be something other than normal, and we next consider the important problem of testing equality of parameters of k-samples (where a limiting Chi-squared distribution is obtained).

Assume we observe k independent samples of i.i.d. observations. Specifically, assume $X_{i,1},\dots,X_{i,n_i}$ are i.i.d. P_i. Some of our results will hold for fixed $n_1,\dots,n_k$, but we also have asymptotic results as $N \equiv \sum_i n_i \to \infty$. Let $n=(n_1,\dots,n_k)$, and the notation $n\to\infty$ will mean $\min_i n_i \to \infty$. Let $\theta(\cdot)$ be a real-valued parameter, defined on some space of distributions $\mathcal{P}$. The problem of interest is to test the null hypothesis

$$H_0 : \theta(P_1) = \cdots = \theta(P_k) \tag{3.1}$$

against the alternative

$$H_1 : \theta(P_i) \neq \theta(P_j) \qquad \text{for some } i, j.$$

When $P_1 = \cdots = P_k$ holds, one can construct permutation tests with exact level α. However, if $P_i \neq P_j$ for some i, j, then the test may fail to achieve the rejection probability equal to α even asymptotically.

We will assume that asymptotically linear estimators are available, that is, (2.2) holds for i.i.d. samples under P_i for $i=1,\dots,k$, where f_{P_i} can depend on the underlying distribution P_i. Further assume that the expansion also holds for i.i.d. observations $\bar{Z}_{i,1},\dots,\bar{Z}_{i,n_i}$ sampled from the mixture distribution $\bar{P}=\sum_{i=1}^k p_i P_i$, where $n_i/N \to p_i$. Note that the asymptotic linearity conditions need not require any form of differentiability (though of course, some form of differentiability is a sufficient condition). We will argue that the asymptotic linearity conditions under P_i for $i=1,\dots,k$ and $\bar{P}$, are sufficient to derive the asymptotic behavior of the k-sample permutation distribution based on $T_{n,1}$ (defined below), without having to impose any differentiability conditions.

The goal here is to construct a method that retains the exact control of the probability of a type 1 error when the observations are i.i.d., but also asymptotically controls the probability of a type 1 error under very weak assumptions, specifically finite nonzero variances of the influence functions.

LEMMA 3.1. *Consider the above set-up. Assume* (2.2) *holds for* $P_1, \dots, P_k$ *with* $0 < \sigma_i^2 = \sigma_i^2(f_{P_i}) = \mathrm{E}_{P_i} f_{P_i}^2(X_{i,j}) < \infty$. *Assume* $n_i \to \infty$ *with* $n_i/N \to p_i > 0$ *for* $i = 1, \dots, k$. *Let*

$$T_{n,0} = \sum_{i=1}^{k} \frac{n_i}{\sigma_i^2}\left[\hat{\theta}_{n,i} - \frac{\sum_{i=1}^{k} n_i \hat{\theta}_{n,i}/\sigma_i^2}{\sum_{i=1}^{k} n_i/\sigma_i^2}\right]^2, \tag{3.2}$$

where $\hat{\theta}_{n,i} = \hat{\theta}_{n,i}(X_{i,1}, \dots, X_{i,n_i})$ *and* $\sigma_i^2 = \sigma_i^2(f_{P_i}) = \mathrm{E}_{P_i} f_{P_i}^2(X_{i,j})$. *Further assume that* $\hat{\sigma}_{n,i} \equiv \hat{\sigma}_{n,i}(X_{i,1}, \dots, X_{i,n_i})$ *is a consistent estimator of* $\sigma_i = \sigma_i(f_{P_i})$ *when* $X_{i,1}, \dots, X_{i,n_i}$ *are i.i.d.* P_i, *for* $i = 1, \dots, k$. *Define*

$$T_{n,1} = \sum_{i=1}^{k} \frac{n_i}{\hat{\sigma}_{n,i}^2}\left[\hat{\theta}_{n,i} - \frac{\sum_{i=1}^{k} n_i \hat{\theta}_{n,i}/\hat{\sigma}_{n,i}^2}{\sum_{i=1}^{k} n_i/\hat{\sigma}_{n,i}^2}\right]^2. \tag{3.3}$$

Then, under H_0, *both* $T_{n,0}$ *and* $T_{n,1}$ *converge in distribution to the Chi-squared distribution with* $k-1$ *degrees of freedom.*

Let $\hat{R}_{n,1}(\cdot)$ denote the permutation distribution corresponding to $T_{n,1}$. In words, $T_{n,1}$ is recomputed over all permutations of the data. Specifically, if we let

$$(Z_1, \dots, Z_N) = (X_{1,1}, \dots, X_{1,n_1}, X_{2,1}, \dots, X_{2,n_2}, \dots, X_{k,1}, \dots, X_{k,n_k}),$$

then, $\hat{R}_{n,1}(t)$ is formally equal to the right-hand side of (1.2), with $T_{m,n}$ replaced by $T_{n,1}$.

THEOREM 3.1. *Assume the same setup and conditions of Lemma* 3.1 *with* $0 < \sigma_i^2 = \sigma_i^2(f_{P_i}) = \mathrm{E}_{P_i} f_{P_i}^2(X_{i,j}) < \infty$. *Assume* $n_i \to \infty$ *with* $n_i/N \to p_i > 0$. *Further assume that the consistency of* $\hat{\sigma}_{n,i}$ *of* σ_i *under* P_i *also holds under* $\bar{P}$ *as well so that, when the* $\bar{Z}_i$ *are i.i.d.* $\bar{P}$,

$$\hat{\sigma}_{n,i}(\bar{Z}_1, \dots, \bar{Z}_{n_i}) \xrightarrow{P} \sigma(f_{\bar{P}}) \qquad \text{as } n \to \infty$$

with $0 < \sigma^2(f_{\bar{P}}) < \infty$.

Then, under H_0,

$$\hat{R}_{n,1}(t) \xrightarrow{P} G_{k-1}(t), \tag{3.4}$$

where G_d *denotes the Chi-squared distribution with* d *degrees of freedom. Moreover, if* $P_1, \dots, P_k$ *satisfy* H_0, *then the probability that the permutation test rejects* H_0 *tends to the nominal level* α.

EXAMPLE 3.1 (Nonparametric k-sample Behrens–Fisher problem). Consider the special case where $\theta_i(P) = \mu_i(P)$ is the population mean. Also, let $\hat{\theta}_{n,i}$ be the sample mean of the ith sample. When the populations are assumed normal with possibly different unknown variances, this is the classical Behrens–Fisher problem. Here, we do not assume normality and provide a general solution for testing the equality of parameters of several distributions. Indeed, we have exact finite sample type 1 error control when all the populations are the same, and asymptotically type 1 error control when the populations are possibly distinct. (Some relatively recent large sample approaches which do not retain our finite sample exactness property to this specific problem are given in Rice and Gaines [21] and Krishnamoorthy, Lu and Mathew [15].)

4. Simulation results.

Monte Carlo simulation studies illustrating our results are presented in this section. Table 1 tabulates the rejection probabilities of one-sided tests for the studentized permutation median test where the nominal level considered is $\alpha = 0.05$. The simulation results confirm that the studentized permutation median test is valid in the sense that it approximately attains level α in large samples.

In the simulation, odd numbers of sample sizes are selected in the Monte Carlo simulation for simplicity. We consider several pairs of distinct sample distributions that share the same median as listed in the first column of Table 1. For each situation, 10,000 simulations were performed. Within a given simulation, the permutation test was calculated by randomly sampling 999 permutations. Note that neither the exactness properties nor the asymptotic properties are changed at all (as long as the number of permutations sampled tends to infinity). For a discussion on stochastic approximations to the permutation distribution, see the end of Section 15.2.1 in Lehmann and Romano [16] and Section 4 in Romano [22]. As is well known, when the underlying distributions of two distinct independent samples are not identical, the permutation median test is not valid in the sense that the rejection probability is far from the nominal level $\alpha = 0.05$. For example, although a logistic distribution with location parameter 0 and scale parameter 1 and a continuous uniform distribution with the support ranging from -10 to 10 have the same median of 0, the rejection probability for the sample sizes examined is between 0.0991 and

TABLE 1

Monte Carlo simulation results for studentized permutation median test (*one-sided*, $\alpha = 0.05$)

	***m*:**	**5**	**13**	**51**	**101**	**101**	**201**	**401**
Distributions	***n*:**	**5**	**21**	**101**	**101**	**201**	**201**	**401**
$N(0, 1)$	Not studentized	0.1079	0.1524	0.1324	0.2309	0.2266	0.2266	0.2249
$N(0, 5)$	Studentized	0.0802	0.1458	0.095	0.0615	0.0517	0.0517	0.0531
$N(0, 1)$	Not studentized	0.0646	0.1871	0.2411	0.1769	0.1849	0.1849	0.1853
$T(5)$	Studentized	0.0707	0.1556	0.0904	0.0776	0.0661	0.0661	0.0611
Logistic$(0, 1)$	Not studentized	0.0991	0.1413	0.1237	0.2258	0.2233	0.2233	0.2261
$U(-10, 10)$	Studentized	0.0771	0.1249	0.0923	0.0686	0.0574	0.0574	0.0574
Laplace$(\ln 2, 1)$	Not studentized	0.0420	0.0462	0.0477	0.048	0.0493	0.0461	0.0501
exp(1)	Studentized	0.0386	0.0422	0.0444	0.0502	0.0485	0.0505	0.0531

0.2261 and moves further away from the nominal level $\alpha = 0.05$ as sample sizes increase.

In contrast, the studentized permutation test results in rejection probability that tends to the nominal level α asymptotically. We apply the bootstrap method (Efron [5]) to estimate the variance for the median $\frac{1}{4f_P^2(\theta)}$ in the simulation given by

$$m\sum_{l=1}^{m}[X_{(l)} - \theta(\hat{P}_m)]^2 \cdot \mathrm{P}(\theta(\hat{P}_m^*) = X_{(l)}),$$

where for an odd number m,

$$\begin{aligned}\mathrm{P}(\theta(\hat{P}_m^*) = X_{(l)}) = \mathrm{P}\Bigl(\text{Binomial}\Bigl(m, \frac{l-1}{m}\Bigr) \le \frac{m-1}{2}\Bigr)\\ - \mathrm{P}\Bigl(\text{Binomial}\Bigl(m, \frac{l}{m}\Bigr) \le \frac{m-1}{2}\Bigr).\end{aligned}$$

As noted earlier, there exist other choices such as the kernel estimator and the smoothed bootstrap estimator. We emphasize, however, that using the bootstrap to obtain an estimate of standard error does *not* destroy the exactness of permutation tests under identical distributions.

5. Four technical ingredients. In this section, we discuss four separate ingredients, from which the main results flow. These results are separated out so they can easily be applied to other problems and so that the main technical arguments are highlighted. The first two apply more generally to randomization tests, not just permutation tests, and are stated as such.

5.1. *Hoeffding's condition.* Suppose data X^n has distribution P_n in $\mathcal{X}_n$, and $\mathbf{G}_n$ is a finite group of transformations g of $\mathcal{X}_n$ onto itself. For a given statistic $T_n = T_n(X^n)$, let $\hat{R}_n^T(\cdot)$ denote the randomization distribution of T_n, defined by

$$\hat{R}_n^T(t) = \frac{1}{|G_n|}\sum_{g \in G_n} I\{T_n(gX^n) \le t\}, \tag{5.1}$$

where $|G_n|$ denotes the cardinality of G_n. Hoeffding [8] gave a sufficient condition to derive the limiting behavior of $\hat{R}_n^T(\cdot)$. This condition is verified repeatedly in the proofs, but we add the result that the condition is also necessary.

THEOREM 5.1. *Let G_n and G_n' be independent and uniformly distributed over $\mathbf{G}_n$ (and independent of X^n). Suppose, under P_n,*

$$(T_n(G_nX^n), T_n(G_n'X^n)) \xrightarrow{d} (T, T'), \tag{5.2}$$

where T and T' are independent, each with common c.d.f. $R^T(\cdot)$. Then, for all continuity points t of $R^T(\cdot)$,

$$\hat{R}_n^T(t) \xrightarrow{P} R^T(t). \tag{5.3}$$

Conversely, if (5.3) holds for some limiting c.d.f. $R^T(\cdot)$ whenever t is a continuity point, then (5.2) holds.

The reason we think it is important to add the necessity part of the result is that our methodology is somewhat different than that of other authors mentioned in the Introduction, who take a more conditional approach to proving limit theorems. After all, the permutation distribution is indeed a distribution conditional on the observed set of observations (without regard to ordering). However, the theorem shows that a sufficient condition is obtained by verifying an unconditional weak convergence property. Nevertheless, simple arguments (see the supplementary appendix [2]) show the condition is indeed necessary and so taking such an approach is not fanciful.

5.2. *Slutsky's theorem for randomization distributions.* Consider the general setup of Section 5.1. The result below describes Slutsky's theorem in the context of randomization distributions. In this context, the randomization distributions are random themselves, and therefore the usual Slutsky's theorem does not quite apply. Because of its utility in the proofs of our main results, we highlight the statement. Given sequences of statistics T_n, A_n and B_n, let $\hat{R}_n^{AT+B}(\cdot)$ denote the randomization distribution corresponding to the statistic sequence $A_nT_n + B_n$; that is, replace T_n in (5.1) by $A_nT_n + B_n$, so

$$\hat{R}_n^{AT+B}(t) \equiv \frac{1}{|\mathbf{G}_n|} \sum_{g \in \mathbf{G}_n} \mathrm{I}\{A_n(gX^n)T_n(gX^n) + B_n(gX^n) \le t\}. \tag{5.4}$$

THEOREM 5.2. *Let G_n and G_n' be independent and uniformly distributed over $\mathbf{G}_n$ (and independent of X^n). Assume T_n satisfies (5.2). Also, assume*

$$A_n(G_nX^n) \xrightarrow{P} a \tag{5.5}$$

and

$$B_n(G_nX^n) \xrightarrow{P} b \tag{5.6}$$

for constants a and b. Let $R^{aT+b}(\cdot)$ denote the distribution of $aT + b$, where T is the limiting random variable assumed in (5.2). Then

$$\hat{R}_n^{AT+B}(t) \xrightarrow{P} R^{aT+b}(t),$$

if the distribution $R^{aT+b}(\cdot)$ of $aT + b$ is continuous at t. [Of course, $R^{aT+b}(t) = R^T(\frac{t-b}{a})$ if $a \neq 0$.]

5.3. *A coupling construction.* Consider the general situation where k samples are observed from possibly different distributions. Specifically, assume for $i = 1, \dots, k$ that $X_{i,1}, \dots, X_{i,n_i}$ is a sample of n_i i.i.d. observations from P_i. All $N \equiv \sum_i n_i$ observations are mutually independent. Put all the observations together in one vector

$$Z = (X_{1,1}, \dots, X_{1,n_1}, X_{2,1}, \dots, X_{2,n_2}, \dots, X_{k,1}, \dots, X_{k,n_k}).$$

The basic intuition driving the results concerning the behavior of the permutation distribution stems from the following. Since the permutation distribution considers the empirical distribution of a statistic evaluated at all permutations of the data, it clearly does not depend on the ordering of the observations. Let n_i/N denote the proportion of observations in the ith sample, and let $p_i = \lim_{n_i \to \infty} n_i/N \in (0, 1)$. Assume that $n_i \to \infty$ in such a way that

$$p_i - \frac{n_i}{N} = O(N^{-1/2}). \tag{5.7}$$

Then the behavior of the permutation distribution based on Z should behave approximately like the behavior of the permutation distribution based on a sample of N i.i.d. observations $\bar{Z} = (\bar{Z}_1, \dots, \bar{Z}_N)$ from the mixture distribution $\bar{P} \equiv p_1 P_1 + \cdots + p_k P_k$. Of course, we can think of the N observations generated from $\bar{P}$ arising out of a two-stage process: for $i = 1, \dots, N$, first draw an index j at random with probability p_j; then, conditional on the outcome being j, sample $\bar{Z}_i$ from P_j. However, aside from the fact that the ordering of the observations in Z is clearly that of n_1 observations from P_1, following by n_2 observations from P_2, etc., the original sampling scheme is still only approximately like that of sampling from $\bar{P}$. For example, the number of observations $\bar{Z}_i$ out of the N which are from P_1 is binomial with parameters N and p_1 (and so has mean equal to $p_1 N \approx n_1$), while the number of observations from P_1 in the original sample Z is exactly n_1.

Along the same lines, let $\pi = (\pi(1), \dots, \pi(N))$ denote a random permutation of $\{1, \dots, N\}$. Then, if we consider a random permutation of both Z and $\bar{Z}$, then the number of observations in the first n_1 coordinates of Z which were X_1's has the hypergeometric distribution, while the number of observations in the first n_1 coordinates of $\bar{Z}$ which were X_1's is still binomial.

We can make a more precise statement by constructing a certain coupling of Z and $\bar{Z}$. That is, except for ordering, we can construct $\bar{Z}$ to include almost the same set of observations as in Z. The simple idea goes as follows. Given Z, we will construct observations $\bar{Z}_1, \dots, \bar{Z}_N$ via the two-stage process as above, using the observations drawn to make up the Z_i as much as possible. First, draw an index j among $\{1, \dots, k\}$ at random with probability p_j; then, conditionally on the outcome being j, set $\bar{Z}_1 = X_{j,1}$. Next, if the next index i drawn among $\{1, \dots, k\}$ at random with probability p_i is different from j from which $\bar{Z}_1$ was sampled, then $\bar{Z}_2 = X_{i,1}$; otherwise, if $i = j$ as in the first step, set $\bar{Z}_2 = X_{j,2}$. In other words, we are going to continue to use the Z_i to fill in the observations $\bar{Z}_i$. However, after a certain point, we will get stuck because we will have already exhausted all the n_j observations from the jth population governed by P_j. If this happens and an index j was drawn again, then just sample a new observation X_{j,n_j+1} from P_j. Continue in this manner so that as many as possible of the original Z_i observations are used in the construction of $\bar{Z}$. Now, we have both $\bar{Z}$ and Z. At this point, $\bar{Z}$ and Z have many of the same observations in common. The number of observations which differ, say D, is the (random) number of added observations required to fill up $\bar{Z}$. (Note that we are obviously using the word "differ" here to mean the observations are generated from different mechanisms, though in fact there may be a positive probability that the observations still are equal if the underlying distributions have atoms. Still, we count such observations as differing.)

Moreover, we can reorder the observations in $\bar{Z}$ by a permutation π_0 so that Z_i and $\bar{Z}_{\pi_0(i)}$ agree for all i except for some hopefully small (random) number D. To do this, recall that Z has the observations in order, that is, the first n_1 observations arose from P_1 and the next set of n_2 observations came from P_2, etc. Thus, to couple Z and $\bar{Z}$, simply put all the observations in $\bar{Z}$ which came from P_1 first up to n_1. That is, if the number of observations in $\bar{Z}$ from P_1 is greater than or equal to n_1, then $\bar{Z}_{\pi(i)}$ for $i = 1, \dots, n_1$ are filled with the observations in $\bar{Z}$ which came from P_1, and if the number was strictly greater than n_1, put them aside for now. On the other hand, if the number of observations in $\bar{Z}$ which came from P_1 is less than n_1, fill up as many of $\bar{Z}$ from P_1 as possible, and leave the rest of the slots among the first n_1 spots blank for now. Next, move onto the observations in $\bar{Z}$ which came from P_2 and repeat the above procedure for $n_1 + 1, \dots, n_1 + n_2$ spots; that is, we start filling up the spots from $n_1 + 1$ as many of $\bar{Z}$ which came from P_2 as possible up to n_2 of them. After going though all the distributions P_i from which each of observations in $\bar{Z}$ came, one must then complete the observations in $\bar{Z}_{\pi_0}$; simply "fill up" the empty spots with the remaining observations that have been put aside. (At this point, it does not matter where each of the remaining observations gets inserted; but, to be concrete, fill the empty slots by inserting the observations which came from the index P_i in chronological order from when constructed.) This permuting of observations in $\bar{Z}$ corresponds to a permutation π_0 and satisfies $Z_i = \bar{Z}_{\pi_0(i)}$ for indices i except for D of them.

For example, suppose there are $k = 2$ populations. Suppose that N_1 of the $\bar{Z}$ observations came from P_1 and so $N - N_1$ from P_2. Of course, N_1 is random and has the binomial distribution with parameters N and p_1. If $N_1 \geq n_1$, then the above construction yields the first n_1 observations in Z and $\bar{Z}_{\pi_0}$ completely agree. Furthermore, if $N_1 > n_1$, then the number of observations in $\bar{Z}$ from P_2 is $N - N_1 < N - n_1 = n_2$, and $N - N_1$ of the last n_2 indices in Z match those of $\bar{Z}_{\pi_0}$, with the remaining differ. In this situation, we have

$$Z = (X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2})$$

and

$$\bar{Z}_{\pi_0} = (X_1, \dots, X_{n_1}, Y_1, \dots, Y_{N-N_1}, X_{n_1+1}, \dots, X_{N_1}),$$

so that Z and $\bar{Z}_{\pi_0}$ differ only in the last $N_1 - n_1$ places. In the opposite situation where $N_1 < n_1$, Z and $\bar{Z}_{\pi}$ are equal in the first N_1 and last n_2 places, only differing in spots $N_1 + 1, \dots, n_1$.

The number of observations D where Z and $\bar{Z}_{\pi_0}$ differ is random and it can be shown that

$$E(D/N) \le N^{-1/2}; \tag{5.8}$$

see supplementary appendix [2]. In summary, the coupling construction shows that only a fraction of the N observations in Z and $\bar{Z}_{\pi_0}$ differ with high probability. Therefore, if the randomization distribution is based on a statistic $T_N(Z)$ such that the difference $T_N(Z) - T_N(\bar{Z}_{\pi_0})$ is small in some sense whenever Z and $\bar{Z}_{\pi_0}$ mostly agree, then one should be able to deduce the behavior of the permutation distribution under samples from $P_1, \dots, P_k$ from the behavior of the permutation distribution when all N observations come from the same distribution $\bar{P}$. Whether or not this can be done requires some knowledge of the form of the statistic, but intuitively it should hold if the statistic cannot strongly be affected by a change in a small proportion of the observations; its validity though must be established on a case by case basis. Although the assessment of the validity needs to be taken on a case by case basis, it readily extends to a broader class of statistics such as "mean-like" statistics. (However, this coupling argument and the contiguity results in Section 5.4 together allow us to prove quite general results.) The point is that it is a worthwhile and beneficial route to pursue because the behavior of the permutation distribution under N i.i.d. observations is typically much easier to analyze than under the more general setting when observations have possibly different distributions. Furthermore, the behavior under i.i.d. observations seems fundamental as this is the requirement for the "randomization hypothesis" to hold, that is, the requirement to yield exact finite sample inference.

To be more specific, suppose π and π' are independent random permutations, and independent of the Z_i and $\bar{Z}_i$. Suppose we can show that

$$\bigl(T_N(\bar{Z}_{\pi}), T_N(\bar{Z}_{\pi'})\bigr) \xrightarrow{d} (T, T'), \tag{5.9}$$

where T and T are independent with common c.d.f. $R(\cdot)$. Then, by Theorem 5.1, the randomization distribution based on T_N converges in probability to $R(\cdot)$ when all observations are i.i.d. according to $\bar{P}$. But since $\pi\pi_0$ (meaning π composed with π_0 so π_0 is applied first) and $\pi'\pi_0$ are also independent random permutations, (5.9) also implies

$$\bigl(T_N(\bar{Z}_{\pi\pi_0}), T_N(\bar{Z}_{\pi'\pi_0})\bigr) \xrightarrow{d} (T, T').$$

Using the coupling construction to construct Z, suppose it can be shown that

$$T_N(\bar{Z}_{\pi\pi_0}) - T_N(Z_{\pi}) \xrightarrow{P} 0. \tag{5.10}$$

Then, it also follows that

$$T_N(\bar{Z}_{\pi'\pi_0}) - T_N(Z_{\pi'}) \xrightarrow{P} 0,$$

and so by Slutsky's theorem, it follows that

$$\big(T_N(Z_\pi), T_N(Z_{\pi'})\big) \xrightarrow{d} (T, T'). \tag{5.11}$$

Therefore, again by Theorem 5.1, the randomization distribution also converges in probability to $R(\cdot)$ under the original model of k samples from possibly different distributions. In summary, the coupling construction of $\bar{Z}$, Z and π_0 and the one added requirement (5.10) allow us to reduce the study of the permutation distribution under possibly k different distributions to the i.i.d. case when all N observations are i.i.d. according to $\bar{P}$. We summarize this as follows.

LEMMA 5.1. *Assume* (5.9) *and* (5.10). *Then* (5.11) *holds, and so the permutation distribution based on k samples from possibly different distributions behaves asymptotically as if all observations are i.i.d. from the mixture distribution $\bar{P}$ and satisfies*

$$\hat{R}^T_{m,n}(t) \xrightarrow{P} R(t),$$

if t is a continuity point of the distribution R of T in (5.9).

EXAMPLE 5.1 (Difference of sample means). To appreciate what is involved in the verification of (5.10), consider the two-sample problem considered in Theorem 2.1, in the special case of testing equality of means. The unknown variances may differ and are assumed finite. Consider the test statistic $T_{m,n} = N^{1/2}[\bar{X}_m - \bar{Y}_n]$. By the coupling construction, $\bar{Z}_{\pi\pi_0}$ and Z_π have the same components except for at most D places. Now,

$$\begin{aligned} T_{m,n}(\bar{Z}_{\pi\pi_0}) - T_{m,n}(Z_\pi) = {}& N^{1/2}\left[\frac{1}{m}\sum_{i=1}^{m}(\bar{Z}_{\pi\pi_0(i)} - Z_{\pi(i)})\right] \\ & - N^{1/2}\left[\frac{1}{n}\sum_{j=m+1}^{N}(\bar{Z}_{\pi\pi_0(j)} - Z_{\pi(j)})\right]. \end{aligned}$$

All of the terms in the above two sums are zero except for at most D of them. But any nonzero term like $\bar{Z}_{\pi\pi_0(i)} - Z_{\pi(i)}$ has variance bounded above by

$$2\max\big(\mathrm{Var}(X_1), \mathrm{Var}(Y_1)\big) < \infty.$$

Note the above random variable has mean zero under the null hypothesis that $E(X_i) = E(Y_j)$. To bound its variance, condition on D and π, and note it has conditional mean 0 and conditional variance bounded above by

$$N\frac{1}{\min(m^2, n^2)}2\max\big(\mathrm{Var}(X_1), \mathrm{Var}(Y_1)\big)D$$

and hence unconditional variance bounded above by

$$N\frac{1}{\min(m^2, n^2)}2\max(\mathrm{Var}(X_1), \mathrm{Var}(Y_1))O(N^{1/2}) = O(N^{-1/2}) = o(1),$$

implying (5.10). In words, we have shown that the behavior of the permutation distribution can be deduced from the behavior of the permutation distribution when all observations are i.i.d. with mixture distribution $\bar{P}$.

Two final points are relevant. First, the limiting distribution R is typically the same as the limiting distribution of the true unconditional distribution of T_N under $\bar{P}$. This is intuitively the case because the permutation distribution is invariant under any permutation of the combined data, and so the set of N observations with exactly n_i observations sampled from P_i and then randomly permuting them behaves very nearly the same as a sample of N observations from $\bar{P}$. On the other hand, the true limiting distribution of the test statistic under $(P_1, \ldots, P_k)$ need not be the same as under $\bar{P}$ as it will in general depend on the underlying distributions $P_1, \ldots, P_k$. However, suppose the choice of test statistic T_N is such that it is an asymptotic pivot in the sense that its limiting distribution does not depend on the underlying probability distributions. Then, the limiting distribution of the test statistic will be the same whether sampling from $(P_1, \ldots, P_k)$ or $(\bar{P}, \ldots, \bar{P})$. In such cases, the randomization or permutation distribution under $(P_1, \ldots, P_k)$ will asymptotically reflect the true unconditional distribution of T_N, resulting in asymptotically valid inference. Indeed, the general results in Section 2 yield many examples of this phenomenon. However, that these statements need qualification is made clear by the following two (somewhat contrived) examples.

EXAMPLE 5.2. Here, we illustrate a situation where coupling works, but the true sampling distribution does not behave like the permutation distribution under the mixture model $\bar{P}$. In the two-sample setup with $m = n$, suppose $X_1, \ldots, X_n$ are i.i.d. according to uniformity on the set of x where $|x| < 1$, and $Y_1, \ldots, Y_n$ are i.i.d. uniform on the set of y with $2 < |y| < 3$. So, $E(X_i) = E(Y_j) = 0$. Consider a test statistic $T_{n,n}$ defined as

$$T_{n,n}(X_1, \ldots, X_n, Y_1, \ldots, Y_n) = N^{-1/2}\left[\sum_{i=1}^{n} I\{|Y_i| > 2\} - I\{|X_i| < 2\}\right].$$

Under the true sampling scheme, $T_{n,n}$ is zero with probability one. However, if all $2n$ observations are sampled from the mixture model, it is easy to see that $T_{n,n}$ is asymptotically normal $N(0, 1/4)$, which is the same limit for the permutation distribution (in probability). So here, the permutation distribution under the given distributions is the same as under $\bar{P}$, though it does not reflect the actual true unconditional sampling distribution.

EXAMPLE 5.3. Here, we consider a situation where both populations are indeed identical, so there is no need for a coupling argument. However, the point is that the permutation distribution does not behave like the true unconditional sampling distribution. Assume $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ are all i.i.d. $N(0, 1)$ and consider the test statistic

$$T_{n,n}(X_1, \dots, X_n, Y_1, \dots, Y_n) = N^{-1/2} \sum_{i=1}^{n} (X_i + Y_i).$$

Unconditionally, $T_{n,n}$ converges in distribution to $N(0, 1)$. However, the permutation distribution places mass one at $\frac{n}{\sqrt{N}}(\bar{X}_n + \bar{Y}_n)$ because the statistic $T_{n,n}$ is permutation invariant.

Examples 5.2 and 5.3 show that the intuition provided in the paragraph before Example 5.2 does not always work. However, in the two examples, the test statistic does not reflect an actual comparison between P and Q. Of course, our theorems apply to tests of equality of parameters, and therefore the test statistics are based on appropriate differences.

5.4. *An auxiliary contiguity result.* Consider the general situation involving k (possibly distinct) populations for $i = 1, \dots, k$ with n_i observations from population i. Set $N = \sum_{i=1}^{k} n_i$ and $n = (n_1, \dots, n_k)'$, where the notation $n \to \infty$ means $\min_i n_i \to \infty$. Assume all N observations are mutually independent. Define $p_{n,i} = n_i/N \to p_i \in (0, 1)$ as $n_i \to \infty$ for $i = 1, \dots, k$. Let P_n be the multinomial distribution based on parameters $s = s(n)$ and $p_n = (p_{n,1}, \dots, p_{n,k})$. So, under P_n, let $M_{n,j}$ be the number of observations of type i when s observations are taken with replacement from a population with n_i observations of type i. So, $M_n \equiv (M_{n,1}, \dots, M_{n,k}) \sim P_n$. Also, let Q_n be the multivariate hypergeometric distribution. Under Q_n, let $H_{n,i}$ be the number of observations of type i when s observations are taken without replacement. So, $H_n \equiv (H_{n,1}, \dots, H_{n,k}) \sim Q_n$.

We shall show that the multinomial distribution P_n and the multivariate hypergeometric distribution Q_m are mutually contiguous, which will allow us to obtain the limiting behavior of a statistic under the given samples from k probability distributions P_i for $i = 1, \dots, k$, by instead calculating the limiting behavior of the statistic when all N observations are i.i.d. from the mixture distribution $\bar{P} = \sum_{i=1}^{k} p_i P_i$, which is relatively easier to obtain. For basic details on contiguity, see Section 12.3 in Lehmann and Romano [16].

LEMMA 5.2. *Assume the above setup with* $s/N \to \theta \in [0, 1)$ *as* $n \to \infty$. *Consider the likelihood ratio* $L_n(x) = dQ_n(x)/dP_n(x)$.

(i) *The limiting distribution of* $L_n(M_n)$ *satisfies*

$$L_n(M_n) \xrightarrow{L} (1 - \theta)^{-(k-1)/2} \exp\left\{-\frac{\theta}{2(1-\theta)} \chi_{k-1}^2\right\}, \tag{5.12}$$

where χ^2_{k-1} *denotes the Chi-squared distribution with* $k-1$ *degrees of freedom.*

(ii) Q_n *and* P_n *are mutually contiguous.*

REMARK 5.1. With $M_n \equiv (M_{n,1}, \dots, M_{n,k})$ having the multinomial distribution with parameters s and $p_n = (p_{n,1}, \dots, p_{n,k})$ as in Lemma 5.2, also let $\bar{M}_n \equiv (\bar{M}_{n,1}, \dots, \bar{M}_{n,k})$ have the multinomial distribution with parameters s and $p = (p_1, \dots, p_k)$. Then, the distributions of M_n and $\bar{M}_n$ are contiguous if and only if $p_{n,i} - p_i = O(n_i^{-1/2})$, not just $p_{n,i} \to p_i$ for all $i = 1, \dots, k$.

LEMMA 5.3. *Suppose* $V_1, \dots, V_s$ *are i.i.d. according to the mixture distribution*

$$\bar{P} \equiv \sum_{i=1}^{k} p_i P_i,$$

where $p_i \in (0, 1)$, $\sum_{i=1}^k p_i = 1$ *and* P_i*'s are probability distributions on some general space. Assume, for some sequence* W_n *of statistics,*

$$W_n(V_1, \dots, V_s) \xrightarrow{P} t \tag{5.13}$$

for some constant t (*which can depend on the* P_i*'s and* p_i*'s*). *Let* $n_i \to \infty$, $s(n) \to \infty$, *with* $s/N \to \theta \in [0, 1)$, $N = \sum_{i=1}^k n_i$, $p_{n,i} = n_i/N$, *and* $p_{n,i} \to p_i \in (0, 1)$ *with*

$$p_{n,i} - p_i = O\bigl(n_i^{-1/2}\bigr). \tag{5.14}$$

Further, let $X_{i,1}, \dots, X_{i,n_i}$ *be i.i.d.* P_i *for* $i = 1, \dots, k$. *Let*

$$(Z_1, \dots, Z_N) = (X_{1,1}, \dots, X_{1,n_1}, \dots, X_{k,1}, \dots, X_{k,n_k}).$$

Let $(\pi(1), \dots, \pi(N))$ *denote a random permutation of* $\{1, \dots, N\}$ (*and independent of all other variables*). *Then,*

$$W_n(Z_{\pi(1)}, \dots, Z_{\pi(s)}) \xrightarrow{P} t. \tag{5.15}$$

REMARK 5.2. The importance of Lemma 5.3 is that is allows us to deduce the behavior of the statistic W_n under the randomization or permutation distribution from the basic assumption of how W_n behaves under i.i.d. observations from the mixture distribution $\bar{P}$. Note that in (5.13), the convergence in probability assumption is required when the V_i are $\bar{P}$ (so the P over the arrow is just a generic symbol for convergence in probability).

6. Conclusion. When the fundamental assumption of identical distributions need not hold, two-sample permutation tests are invalid unless quite stringent conditions are satisfied depending on the precise nature of the problem. For example, the two-sample permutation test based on the difference of sample means is asymptotically valid only when either the distributions have the same variance or they are comparable in sample size. Thus, a careful interpretation of rejecting the null is necessary; rejecting the null based on the permutation tests does not necessarily imply a valid rejection of the null that some real-valued parameter $\theta(F, G)$ is some specified value θ_0. We provide a framework that allows one to obtain asymptotic rejection probability α in two-sample permutation tests. One great advantage of utilizing the proposed test is that it retains the exactness property in finite samples when $P = Q$, a desirable property that bootstrap and subsampling methods fail to possess.

To summarize, if the true goal is to test whether the parameter of interest θ is some specified value θ_0, permutation tests based on correctly studentized statistic is an attractive choice. When testing the equality of means, for example, the permutation t-test based on a studentized statistic obtains asymptotic rejection probability α in general while attaining exact rejection probability equal to α when $P = Q$. In the case of testing the equality of medians, the studentized permutation median test yields the same desirable property. Moreover, the results extend to quite general settings based on asymptotically linear estimators. The results extend to k-sample problems as well, and analogous results hold in k-sample problem of comparing general parameters, which includes the nonparametric k-sample Behrens–Fisher problem. The guiding principle is to use a test statistic that is asymptotically distribution-free or pivotal. Then, the technical arguments developed in this paper can be shown that the permutation test behaves asymptotically the same as when all observations share a common distribution. Consequently, if the permutation distribution reflects the true underlying sampling distribution, asymptotic justification is achieved.

As mentioned in the Introduction, proper implementation of a permutation test is vital if one cares about confirmatory inference through hypothesis testing; indeed, proper error control of types 1, 2 and 3 errors can be obtained for test of parameters by basing inference on test statistics which are asymptotically pivotal. Thus, the foundations are laid for considering more complex problems in modern data analysis, such as two-sample microarray genomics problems, where a very large number of tests are performed simultaneously. (Indeed, there are many microarray analyses which have begun by performing a permutation test for each gene, without proper studentization.) The role of permutations in multiple testing cannot be properly understood without a firm basis for single testing. Thus, future work will further develop the ideas presented here so that permutation tests can be applied to other measures of error control in multiple testing such as the false discovery rate.

SUPPLEMENTARY MATERIAL

Supplement to "Exact and asymptotically robust permutation tests" (DOI: 10.1214/13-AOS1090SUPP; .pdf). Contains proofs of all the results in the paper.

REFERENCES

[1] CHUNG, E. and ROMANO, P. J. (2011). Asymptotically valid and exact permutation tests based on two-sample U-statistics. Technical Report 2011-09, Dept. Statistics, Stanford Univ.

[2] CHUNG, E. and ROMANO, P. J. (2013). Supplement to "Exact and asymptotically robust permutation tests." DOI:10.1214/13-AOS1090SUPP.

[3] DELAIGLE, A., HALL, P. and JIN, J. (2011). Robustness and accuracy of methods for high dimensional data analysis based on Student's t-statistic. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73** 283–301. MR2815777

[4] DEVROYE, L. P. and WAGNER, T. J. (1980). The strong uniform consistency of kernel density estimates. In *Multivariate Analysis, V* (*Proc. Fifth Internat. Sympos., Univ. Pittsburgh, Pittsburgh, Pa.*, 1978) 59–77. North-Holland, Amsterdam. MR0566330

[5] EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681

[6] HALL, P., DICICCIO, T. J. and ROMANO, J. P. (1989). On smoothing and the bootstrap. *Ann. Statist.* **17** 692–704. MR0994260

[7] HALL, P. and WILSON, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics* **47** 757–762. MR1132543

[8] HOEFFDING, W. (1952). The large-sample power of tests based on permutations of observations. *Ann. Math. Statist.* **23** 169–192. MR0057521

[9] JANSSEN, A. (1997). Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens–Fisher problem. *Statist. Probab. Lett.* **36** 9–21. MR1491070

[10] JANSSEN, A. (1999). Testing nonparametric statistical functionals with applications to rank tests. *J. Statist. Plann. Inference* **81** 71–93. MR1718393

[11] JANSSEN, A. (2005). Resampling Student's t-type statistics. *Ann. Inst. Statist. Math.* **57** 507–529. MR2206536

[12] JANSSEN, A. and PAULS, T. (2003). How do bootstrap and permutation tests work? *Ann. Statist.* **31** 768–806. MR1994730

[13] JANSSEN, A. and PAULS, T. (2005). A Monte Carlo comparison of studentized bootstrap and permutation tests for heteroscedastic two-sample problems. *Comput. Statist.* **20** 369–383. MR2242115

[14] JANSSEN, A. and PAULY, M. (2009). Asymptotics and effectiveness of conditional tests with applications to randomization tests. Technical report, Univ. Duesseldorf.

[15] KRISHNAMOORTHY, K., LU, F. and MATHEW, T. (2007). A parametric bootstrap approach for ANOVA with unequal variances: Fixed and random models. *Comput. Statist. Data Anal.* **51** 5731–5742. MR2407673

[16] LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York. MR2135927

[17] NEUBERT, K. and BRUNNER, E. (2007). A studentized permutation test for the non-parametric Behrens–Fisher problem. *Comput. Statist. Data Anal.* **51** 5192–5204. MR2370717

[18] NEUHAUS, G. (1993). Conditional rank tests for the two-sample problem under random censorship. *Ann. Statist.* **21** 1760–1779. MR1245767

[19] OMELKA, M. and PAULY, M. (2012). Testing equality of correlation coefficients in a potentially unbalanced two-sample problem via permutation methods. *J. Statist. Plann. Inference* **142** 1396–1406.

[20] PAULY, M. (2011). Discussion about the quality of F-ratio resampling tests for comparing variances. *TEST* **20** 163–179. MR2806315

[21] RICE, W. R. and GAINES, S. D. (1989). One-way analysis of variance with unequal variances. *Proc. Natl. Acad. Sci. USA* **86** 8183–8184. MR1021453

[22] ROMANO, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Statist.* **17** 141–159. MR0981441

[23] ROMANO, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *J. Amer. Statist. Assoc.* **85** 686–692. MR1138350

[24] SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165

DEPARTMENT OF ECONOMICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-6072
USA
E-MAIL: eunyi@stanford.edu

DEPARTMENTS OF STATISTICS AND ECONOMICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4068
USA
E-MAIL: romano@stanford.edu