Hindawi Publishing Corporation
International Journal of Mathematics and Mathematical Sciences
Volume 2010, Article ID 769780, 10 pages
doi:10.1155/2010/769780

Research Article

Invalid Permutation Tests

Mikel Aickin

Department of Family & Community Medicine, University of Arizona, Tucson, AZ 85724, USA

Correspondence should be addressed to Mikel Aickin, maickin@earthlink.net

Received 25 December 2009; Accepted 15 March 2010

Academic Editor: Andrei Volodin

Permutation tests are often presented in a rather casual manner, in both introductory and advanced statistics textbooks. The appeal of the cleverness of the procedure seems to replace the need for a rigorous argument that it produces valid hypothesis tests. The consequence of this educational failing has been a widespread belief in a "permutation principle", which is supposed invariably to give tests that are valid by construction, under an absolute minimum of statistical assumptions. Several lines of argument are presented here to show that the permutation principle itself can be invalid, concentrating on the Fisher-Pitman permutation test for two means. A simple counterfactual example illustrates the general problem, and a slightly more elaborate counterfactual argument is used to explain why the main mathematical proof of the validity of permutation tests is mistaken. Two modifications of the permutation test are suggested to be valid in a very modest simulation. In instances where simulation software is readily available, investigating the validity of a specific permutation test can be done easily, requiring only a minimum understanding of statistical technicalities.

1. Introduction

Permutation tests are frequently recommended on two grounds. First, they require fewer assumptions than corresponding model-based tests, and secondly, their validity (as statistical hypothesis tests) is guaranteed by construction. The purpose of this note is to indicate a way in which the first property undermines the second property. The setting is the usual one in which the Mann-Whitney test is asserted to be appropriate, but going beyond this to the assertion of a more general permutation principle (Lehmann [1]).

Users of statistical methods appear to be of two minds about permutation tests. On one hand, since the "randomization" test in the context of a randomized clinical trial is an example of a permutation test, much of the argument in favor of randomization as an experimental principle has been that there is a guaranteed correct statistical test. This argument has had an enormous impact on the design of biomedical studies, with virtually all researchers agreed that randomization is necessary for a study to be valid. If a randomization test is invalid, however, then the technical argument for randomization becomes rather thin. On the other hand, formal permutation tests occur fairly rarely in statistical practice. In the two-sample situation discussed here, the dominant statistical method is the t-test, and although randomization arguments have been used to justify it, it can also be recommended without mentioning randomization. Perhaps the only other analyses that are used with some frequency are the nonparametric alternatives to the t-tests, and these are often justified by appealing to a "permutation principle", giving the impression that these alternatives minimize untestable assumptions. In fact, the often-used Wilcoxon and Mann-Whitney tests do not follow from a permutation principle, but even if they did I will argue here that this principle is invalid.

First, I will present an exaggerated example that indicates in a practical sense how the logic of the general permutation test can break down. Secondly, using the notion of counterfactuals, I will show in a slightly more theoretical way what the source of the problem is. Thirdly, I will suggest two potential alternatives to the permutation test, and argue in their favor on the basis of a small simulation. Finally, I will indicate how Fisher's exact test fails to follow counterfactual principles, based on an example.

2. Regression Example

The two-sample location problem can be expressed as a model for individual pairs (y, x) of the form $y = \theta x + e$, where x takes the values 1 and −1 (indicating two groups), e is a chance perturbation, having mean 0 and being uncorrelated with x, and θ is the parameter of interest. In this paper, I will only consider the case in which half of the observations have $x = 1$, and so the total sample size n is even. Given data of this form, the conventional estimate of θ is

$$\widehat{\theta} = \frac{\sum xy}{n}. \tag{2.1}$$

The null hypothesis is that $\theta = 0$, and the permutation test of this hypothesis is based on the putative null distribution of the above estimate, obtained in the following way. Let $\tilde{x}$ denote a random permutation of the x-values. For simulation purposes, (y, x) is replaced by $(y, \tilde{x})$, a large number of random permutations $\tilde{x}$ are generated, with y fixed, and for each one the corresponding estimate of θ is computed. Finally, this large sample of θs is taken to estimate the null hypothesis distribution of $\widehat{\theta}$.

A difficulty with this approach is illustrated in Figures 1 and 2. Figure 1(a) shows a unit Normal distribution, intended as the distribution of y (or e) under the null hypothesis that $\theta = 0$. Half of the values from this distribution are regarded as being cases when $x = -1$, the remainder when $x = 1$. Figure 1(b) shows the simulated permutation null distribution of the estimate of θ. Figure 2(a) shows a another distribution for y (or e), again under the null hypothesis. Figure 2(b) shows the simulated permutation null distribution of the estimate of θ in this case. Both of these cases are correct, in the sense that if the distribution of e is given on the left, then the null distribution of $\widehat{\theta}$ is validly approximated on the right.

Now let us look at the situation with different eyes. Suppose that Figure 1(a) shows the distribution of e, but that θ is large enough to produce the distribution of y seen in Figure 2(a). The distribution of the parameter estimate under the null hypothesis is still shown in Figure 1(b). But the permutation null distribution that will be computed from data

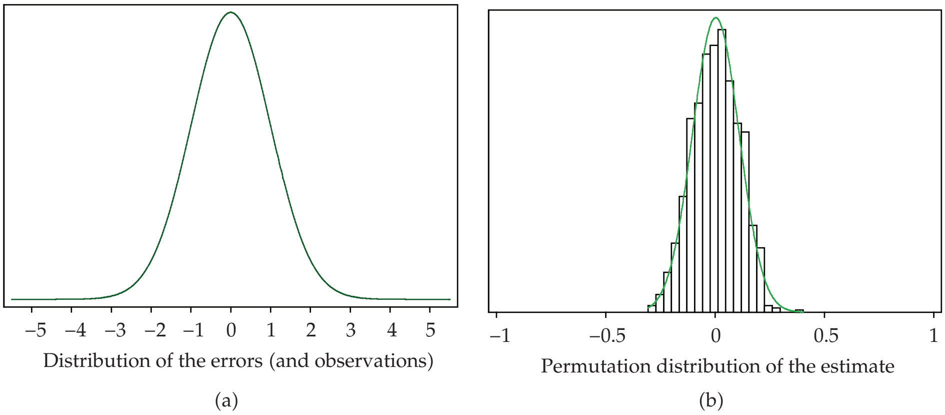

Figure 1: A null situation for generating a permutation distribution.

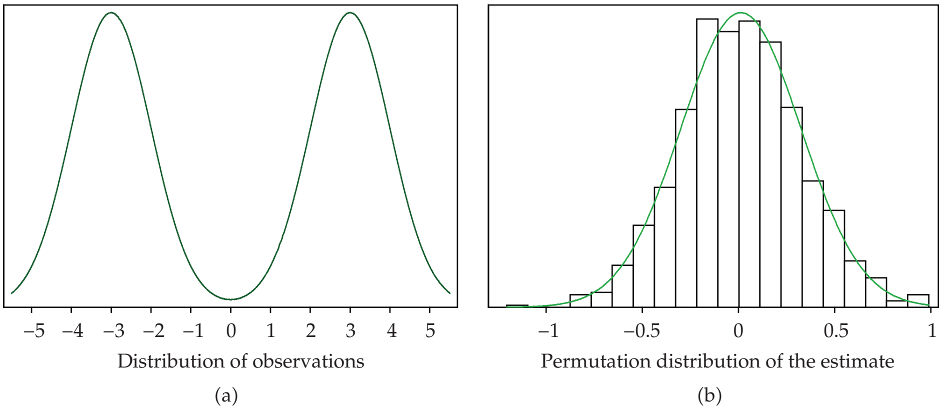

Figure 2: A situation that can be either null or non-null, for simulating a permutation distribution.

coming from the distribution in Figure 2(a) appears in Figure 2(b). This shows an example of the following: in the situation where, as a matter of fact, the null hypothesis is false, the permutation estimate of the null distribution can also be false.

3. Counterfactuals

One way of understanding this situation is based on the method of counterfactuals (Rubin, [2]). In the two-sample circumstance, we would write the outcomes as $y(\theta, \psi)$, considered as a stochastic process indexed by the parameter of interest (θ) and the nuisance parameter (ψ). Again, in our situation this would be $y(\theta, \psi) = \theta x + y(0, \psi)$, where now $y(0, \psi)$ plays the role of e above. In the exposition by Lehmann [1], the parameter ψ would represent a probability density, and $y(0, \psi)$ would have density ψ. The value $y(0, \psi)$ is generated by some chance mechanism, and then $y(\theta, \psi)$ is generated by the model equation above. The values we observe are $y(\theta, \psi)$, but for values of θ and ψ that we do not know. Counterfactuality comes in because $y(0, \psi)$ is the value we would have observed if the null hypothesis had been true. As is usually the case in counterfactual arguments, we can only actually observe

$y(0, \psi)$ under special circumstances (when the null hypothesis is true), so that in general it is only a hypothetical quantity.

Here is the argument of Lehmann [1]. The *set* of values $\{y(0, \psi)\}$ (equivalent to the order statistic) is complete and sufficient for the model defined by $\theta = 0$ and $\psi =$ any density. It follows that the null conditional distribution of the *sample* $y(0, \psi)$ given the order statistic $\{y(0, \psi)\}$ is based on the permutation distribution, irrespective of the value of ψ. This is the basis for the permutation test. In practice, however, it is not $\{y(0, \psi)\}$ that is observed, but $\{y(\theta, \psi)\}$. The conventional permutation distribution is based on the observed order statistic, but this is not the order statistic that would have been observed if the null hypothesis had been true. This is the core of the counterfactual objection to the permutation test; it conditions on the observed order statistic $\{y(\theta, \psi)\}$ instead of conditioning on the null order statistic $\{y(0, \psi)\}$, which Lehmann used for his proof. Because it is conditioning on the latter that justifies the test, the argument for conditioning on the former is not obviously correct.

In Lehmann's theorem where the permutation test is derived, he implicitly considers only observations of the form $y(0, \psi)$, and shows that the (unbiased) rejection region R must satisfy $P[R \mid \{y(0, \psi)\}] = \alpha$, the level of the test. In his examples, he implies that R can be found by solving $P[R \mid \{y(\theta, \psi)\}] = \alpha$. But since his theorem implicitly assumed the null hypothesis, he did not show that R must satisfy this latter equation, and as we have seen in the example above, it need not be satisfied.

There is yet another way of seeing the problem, which connects it more meaningfully to the example. The observed order statistic $\{y(\theta, \psi)\}$ has the same distribution as an order statistic $\{y(0, \psi_0)\}$, where ψ_0 is a distribution that depends on θ and ψ. The permutation distribution based on $\{y(0, \psi_0)\}$ is indeed a *possible* null distribution, but it is not the null distribution that would have been obtained if θ had been 0, and ψ had been left unchanged. Thus, it is precisely the fact that the permutation argument permits different values of (θ, ψ) to produce the same order statistic that creates the problem.

Finally, there is a way of seeing how the sufficiency argument actually misleads the inference. The argument of Lehmann [1] is that conditional on the order statistic, the distribution of the sample does not depend on the nuisance parameter ψ, in the submodel where $\theta = 0$. This is perhaps naturally interpreted to mean that the nuisance parameter has been eliminated from the problem. But this is not the correct interpretation of Lehmann's equations. What they really say is that the only influence that the nuisance parameter ψ has on the distribution of the sample is carried by the order statistic. The order statistic completely mediates the effect of ψ on the distribution of the sample. Thus, both θ and ψ jointly influence the permutation distribution of the sample, but the permutation test does not separate their individual influences.

4. Better Estimates of the Null Distribution

One obvious strategy to repair the permutation test is to replace the order statistic by $\{y(\theta, \psi) - \widehat{\theta} x\}$. This is an attempt to estimate the true null order statistic $\{y(0, \psi)\}$. The potential objection is that the variability of this estimate might be less than the variability of the true null order statistic, producing an invalid test. The permutation test based on this replacement is called the *adjusted permutation test*.

Another approach is to use a simulation based on a random permutation $\tilde{x}$, but conditioned so that it is orthogonal to x. The corresponding simulation sample of estimates

Table 1: Simulated means of effect estimates.

θ	Null	Conventional	Adjusted	Orthogonal
0	−0.00027	−0.00027	−0.00021	−0.00003
0.1	−0.00007	−0.00006	0.00002	0.00010
0.2	−0.00005	−0.00017	0.00005	−0.00049
0.3	0.00022	0.00022	0.00026	−0.00024
0.4	0.00021	0.00027	0.00021	0.00033
0.5	0.00051	0.00047	0.00048	−0.00009
0.6	−0.00015	−0.00005	−0.00005	−0.00038
0.7	0.00050	0.00053	0.00046	0.00018
0.8	−0.00025	−0.00035	−0.00033	0.00007
0.9	−0.00066	−0.00070	−0.00060	−0.00038
1	−0.00007	−0.00030	−0.00005	0.00004
1.1	−0.00015	0.00032	−0.00007	0.00006
1.2	0.00013	−0.00012	0.00012	−0.00009
1.3	0.00014	−0.00035	0.00011	−0.00019

Means of the sampling distributions of estimates of the null mean, using four procedures: (1) null, the usual method when the null is true, (2) conventional, the usual method when the null is false, (3) adjusted, subtracting off the estimate, (4) orthogonal, using only orthogonal permutations.

are then of the form

$$\frac{\sum \tilde{x}y(\theta,\psi)}{n} = \frac{\sum \tilde{x}y(0,\psi)}{n}. \tag{4.1}$$

The simulated distribution of the estimate is thus based on order statistics that would have counterfactually been seen under the null hypothesis, but with a restriction on the permutations. It is not clear that this procedure is fully justified, because of the orthogonality restriction. Here we call this the *orthogonal permutation test*.

In the simulation, we call the usual permutation test the *conventional permutation test*. Because everything is based on a simulation, we know the true null order statistic, and the permutation test based on it is called the *null permutation test.* This gives us altogether four permutation tests, of which only the null permutation test is (within simulation variation) known to be correct.

5. A Simulation

To compare these four tests, I performed a small simulation in the above regression situation. The simulation used sample size 16, since the orthogonal permutation test is only possible for $\tilde{x}$ taking values −1 and 1 if n is divisible by 4. The distribution of e was Normal (0,1), values of θ were 0(.1)1.3, the number of simulated tests for each value of the parameter was 1000, and the number of permutations in each test was 100. The simulation was carried out in Stata (version 9), using a permutation routine that was written specifically for this research.

The results of the simulation are shown in Tables 1–4. Table 1 shows the average estimate (each based on 1000 replications), verifying that all estimates of the null distribution have the correct mean of 0. For the null case this can be shown theoretically, and so the values here serve as a positive control on the simulations. The standard deviations of the estimates

Table 2: Simulated SDs of effect estimates.

θ	Null	Conventional	Adjusted	Orthogonal
0	0.246	0.246	0.237	0.246
0.1	0.246	0.247	0.237	0.246
0.2	0.246	0.251	0.237	0.246
0.3	0.246	0.257	0.237	0.246
0.4	0.246	0.265	0.237	0.246
0.5	0.246	0.276	0.237	0.246
0.6	0.246	0.288	0.237	0.246
0.7	0.246	0.302	0.237	0.246
0.8	0.247	0.318	0.238	0.246
0.9	0.246	0.335	0.238	0.246
1	0.246	0.353	0.237	0.246
1.1	0.246	0.372	0.237	0.246
1.2	0.247	0.392	0.238	0.246
1.3	0.246	0.412	0.237	0.246

Standard deviations of the estimators in Table 1.

(SDE) are shown in Table 2. Again the null test values are constant, as they should be within the simulation variation. The orthogonal test has essentially the same SDEs as the null test, and the adjusted test is only slightly higher. In contrast, the conventional test estimates an SDE that grows with the size of the parameter, another result that can be verified theoretically (see the Appendix).

The probability of rejecting the null hypothesis is shown in Table 3. Again the results of the orthogonal test are essentially the same as the correct null test, with the adjusted test losing only a small amount of power. Although the conventional test is worse than both the adjusted and orthogonal tests, the difference is rather small. Note that the adjusted and orthogonal tests appear to have been slightly larger than nominal levels, suggesting that possibly some adjustment needs to be made. (In the simulations a value equal to the observed was included in the rejection region, and the number of permutations per test was small, both of which might account for some of the excess level, but more research is warranted.) Table 4 shows the efficiency of the conventional test relative to the others, in terms of sample size. This is an estimation-based rather than a test-based comparison. Clearly the conventional test fares poorly relative to the other tests, and the deficit grows with the size of the effect parameter.

6. Fisher's Exact Test Example

A similar problem affects the exact test for 2 × 2 tables. When the margins of a 2 × 2 table are indeed fixed by the design of the experiment, then the permutation distribution may well make sense. To the contrary, the exact test has been advocated as a general testing procedure that is valid even when the margins are not fixed. The counterfactual approach can be portrayed in these cases by defining indicators as follows: $b_r(u_r, \beta_r, \beta_c, \beta_{rc}) = \text{ind}(u_r < \beta_r), b_c(u_c, \beta_r, \beta_c, \beta_{rc}) = \text{ind}(u_c < \beta_c)$. $b_{rc}(u_{rc}, \beta_r, \beta_c, \beta_{rc}) = \text{ind}(u_{rc} < \beta_{rc})$, where the u's are independent uniform chance variables, and "ind" means "indicator variable of". The row indicator is the larger of b_r and b_{rc}, and the column indicator is the larger of b_c and b_{rc}. The

Table 3: Simulated power of tests.

θ	Null	Conventional	Adjusted	Orthogonal
0	0.053	0.053	0.067	0.061
0.1	0.112	0.1	0.132	0.123
0.2	0.212	0.186	0.221	0.212
0.3	0.337	0.283	0.34	0.316
0.4	0.487	0.426	0.487	0.47
0.5	0.643	0.574	0.645	0.619
0.6	0.759	0.726	0.768	0.756
0.7	0.857	0.835	0.876	0.862
0.8	0.924	0.913	0.94	0.929
0.9	0.963	0.965	0.978	0.974
1	0.984	0.986	0.991	0.991
1.1	0.996	0.995	0.996	0.996
1.2	0.997	0.998	0.999	0.999
1.3	0.999	0.999	1	0.999

Powers of the tests based on the estimators in Table 1.

Table 4: Simulated sample-size efficiency of the conventional test versus other tests.

θ	Null	Adjusted	Orthogonal
0.1	0.992	0.921	0.989
0.2	0.964	0.895	0.961
0.3	0.919	0.853	0.915
0.4	0.861	0.801	0.859
0.5	0.798	0.741	0.793
0.6	0.73	0.678	0.73
0.7	0.664	0.617	0.661
0.8	0.601	0.558	0.597
0.9	0.541	0.503	0.538
1	0.487	0.452	0.485
1.1	0.437	0.406	0.436
1.2	0.396	0.368	0.394
1.3	0.358	0.332	0.356

Efficiency of the estimators in Table 1.

null hypothesis is independent of row and column, which is the same as $\beta_{rc} = 0$. The nuisance parameter is (β_r, β_c).

Here is an illustrative example. Table 5 shows the result of a single simulation with $\beta_r = 0.6$, $\beta_c = 0.6$, and $\beta_{rc} = 0.3$. The one-sided P-value from Fisher's exact test is .517 (according to Stata). If we eliminate b_{rc} (counterfactually imposing the null hypothesis), then the results are as in Table 6. If we take it that there were 10 observations in the lower-right cell (as in Table 5), but with the margins of Table 6, the one-sided P-value is .0124 (again from Stata). This example makes it clear that the failure of the null hypothesis has an effect on the conditioning statistic, which in this case consists of the marginal frequencies in the table. The conditioning in Fisher's exact test is not on the frequencies that would have been seen under the null hypothesis with $\beta_{rc} = 0$ but β_r and β_c unchanged, but rather under a different version of the

Table 5: Single simulation (dependent).

1	4	**5**
5	10	**15**
6	**14**	**20**

Simulated data based on b_r, b_c, and b_{rc}, exhibiting dependence of row and column, but with $P = .517$

Table 6: Single simulation (independent).

2	5	**7**
7	6	**13**
9	**11**	**20**

Data from the simulation in Table 5, but omitting the b_{rc} factor, thereby correctly estimating the null margins. 10 observations in the lower right cell would result in a P-value of .0124.

null hypothesis in which the actual nonnull value of β_{rc} implicitly gives different values to β_r and β_c. From the counterfactual standpoint, Fisher's exact test uses the wrong margins.

7. Discussion

It has become increasingly common in the statistical and biomedical literature to see assertions that amount to a general "permutation principle". The issue is whether two (or more) variables are related, and a test is performed based on an estimated null distribution, which is produced by permuting one (or more) of the variables while leaving the remainder fixed. In randomized trials, this is called the "randomization test", and more recently "re-randomization" has also been used for it, even in situations where no original randomization has been performed. The impression is given that either no or very few assumptions are necessary for the correctness of this procedure.

For example, one now sees the justification for the Wilcoxon one-sample test as the application of this general permutation principle to the signed ranks of the observations, as if no further assumptions were required. The Wilcoxon test is, to the contrary, the consequence of a careful argument (Pratt and Gibbon [3]) that requires the symmetry of the underlying distribution, the disregard of which has been noted for some time (Kruskal [4]). It is known that the Wilcoxon test can be invalid in the presence of asymmetry; that is, the test detects the asymmetry rather than a departure from the null hypothesis, when the null hypothesis regarding the mean is true. Thus it is a specific argument based on symmetry and not a general permutation principle that justifies the Wilcoxon test. Similar comments apply to the Mann-Whitney test. There have been some cautionary articles about the general validity of the permutation principle (Romano [5], Hayes [6], Zimmerman [7], Lu et al. [8], Zimmerman [9], Modarres et al. [10]), but the dominant statistical thinking has been to ignore the cautions.

Based on the counterfactual argument and simulation of a regression case presented here, it seems warranted to say that permutation tests need to be revisited. Despite the fact that permutation tests are not widely used in practice, there is a very large literature on them, and several books that explain in detail how they can be used in a wide variety of situations. This literature is attractive because it seems to offer valid statistical procedures, even in complex cases, and in cases where there are technical barriers to obtaining theoretical results. Indeed, when a permutation principle does apply to a specific situation, the argument in favor of using it seems considerable, due to the reduction of untestable assumptions. But when the permutation principle fails, then there is a risk of raising false confidence in an unreliable procedure, with obvious negative consequences. The conclusion is that a substantial amount of new research is required to distinguish between valid and invalid permutation tests, and potentially also to devise modifications of the generally recommended tests that would be appropriate in practice. In any case, it should be recognized that the general assertion of a permutation principle that automatically produces valid tests appears itself to be invalid.

Appendix

The purpose of this appendix is to derive the standard deviation of the permutation estimate of a regression parameter. Let $y_i (i = 1, 2, \ldots, n = 2m)$ be any collection of numbers. Let vector x be chosen at random from among the ordered lists of n values +1 and −1, with half being +1 and half being −1. It is obvious that

$$E\left(\sum_i x_i y_i\right) = 0, \tag{A.1}$$

where the expectation is taken with respect to the distribution of the x_i's. By algebra,

$$\left(\sum_i x_i y_i\right)^2 = \sum_i y_i^2 + \sum_{i \neq j} x_i x_j y_i y_j. \tag{A.2}$$

Employing an elementary combinatorial argument

$$pr(x_i x_j = 1) = pr(x_i = x_j) = 2\frac{\binom{2m-2}{m-2}}{\binom{2m}{m}} = \frac{m-1}{2m-1} \tag{A.3}$$

from which it follows immediately that

$$pr(x_i x_j = -1) = \frac{m}{2m-1}. \tag{A.4}$$

Thus

$$E\left[\sum_{i \neq j} x_i x_j y_i y_j\right] = \left(\frac{m-1}{2m-1} - \frac{m}{2m-1}\right)\sum_{i \neq j} y_i y_j = -\frac{1}{2m-1}\sum_{i \neq j} y_i y_j \tag{A.5}$$

which after a few additional manipulations gives

$$E\left[\left(\frac{\sum_i x_i y_i}{n}\right)^2\right] = \frac{1}{n}\frac{\sum_i (y_i - \overline{y})^2}{n-1}. \tag{A.6}$$

The import of this result for permutation inference for a regression parameter is the following. If the model $y_i = \theta x_i + e_i$ holds, then the last equation above shows that the variance of the permutation distribution of the regression estimator depends on θ. Consequently, for the observed values of y_i the permutation distribution of the regression estimator has the correct variance if and only if $\theta = 0$, that is, if and only if the null hypothesis is in fact true.

Acknowledgment

This research was supported by Grant AT001906 from the National Institutes of Health.

References

[1] E. L. Lehmann, *Testing Statistical Hypotheses*, John Wiley & Sons, New York, NY, USA, 1959.

[2] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974.

[3] J. W. Pratt and J. D. Gibbons, *Concepts of Nonparametric Theory*, Springer Series in Statistics, Springer, New York, NY, USA, 1981.

[4] W. H. Kruskal, "Review of nonparametric and shortcut statistics," *Journal of the American Statistical Association*, vol. 53, pp. 595–598, 1958.

[5] J. P. Romano, "On the behavior of randomization tests without a group invariance assumption," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 686–692, 1990.

[6] A. F. Hayes, "Randomization test and the equality of variance assumption when comparing group means," *Animal Behaviour*, vol. 59, no. 3, pp. 653–656, 2000.

[7] D. W. Zimmerman, "Statistical significance levels of nonparametric tests biased by heterogeneous variances of treatment groups," *Journal of General Psychology*, vol. 127, no. 4, pp. 354–364, 2000.

[8] M. Lu, G. Chase, and S. Li, "Permutation tests and other test statistics for ill-behaved data: experience of the NINDS t-PA stroke trial," *Communications in Statistics—Theory and Methods*, vol. 30, no. 7, pp. 1481–1496, 2001.

[9] D. W. Zimmerman, "A warning about the large-sample Wilcoxon-Mann-Whitney test," *Understanding Statistics*, vol. 2, no. 4, pp. 267–280, 2003.

[10] R. Modarres, J. L. Gastwirth, and W. Ewens, "A cautionary note on the use of non-parametric tests in the analysis of environmental data," *Environmetrics*, vol. 16, no. 4, pp. 319–326, 2005.