Journal of Modern Applied Statistical Methods

Volume 12 | Issue 1 Article 2

5-1-2013

Conceptual Distinction between the Critical *p* Value and the Type I Error Rate in Permutation Testing

Richard B. Anderson
Bowling Green State University

Conceptual Distinction between the Critical *p* Value and the Type I Error Rate in Permutation Testing

Cover Page Footnote
The author wishes to thank Dr. Michael Doherty, Dr. John Tisak, Dr. Mark Appelbaum, Dr. Joseph Rodgers and Don Zhang for their helpful comments on earlier drafts of the manuscript.

Journal of Modern Applied Statistical Methods
May 2013, Vol. 12, No. 1, 2-8

1538 – 9472/13/$95.00

Invited Debate

Conceptual Distinction between the Critical *p* Value and the Type I Error Rate in Permutation Testing

Richard B. Anderson
Bowling Green State University
Bowling Green, OH

To counter past assertions that permutation testing is not distribution-free, this article clarifies that the critical *p* value (alpha) in permutation testing is not a Type I error rate and that a test's validity is independent of the concept of Type I error.

Key words: Statistics, null hypothesis, non-parametric, permutation test, exchangeability; Type I error; validity.

Introduction

Traditional parametric tests, such as *t* and *F* tests, are said to be robust against violation of the normality assumption (e.g., Keppel & Wickens, 2004), but researchers often hesitate to apply such tests when the extent of the violation is obvious or severe. For example, it is known that most parametric tests are not robust against violations of equal variance and that the situation is exacerbated when sample sizes are unequal (Howell, 2012). Thus, non-parametric tests, including permutation tests (e.g., Edgington & Onghena, 2007; Fisher, 1935; Good, 2011; Ludbrook & Dudley, 1998), have become increasingly popular. Some critiques of permutation tests have questioned whether such tests are genuinely distribution-free in the sense of being valid irrespective of the shapes of the population distributions (Hayes, 1997, 2000; Mewhort, Kelley & Johns, 2009). This article clarifies and demonstrates that the critical *p* values for permutation tests are not estimates of Type I error probability and that the divergence of the two values does not impugn the validity of the permutation test's *p* value.

Richard Anderson is an Associate Professor of Psychology in the Department of Psychology. His research interests include topics in judgment, decision-making, reasoning, memory, statistical cognition and statistical methods. He teaches courses in cognition, statistics and research methods and holds memberships in the Psychonomic Society and the Cognitive Science Society. Email him at: randers@bgsu.edu.

The Logic of the Permutation Test

As described by Edgington & Onghena (2007), Fisher (1935), Good (2011) and others, permutation testing entails the following steps. First, the investigator must formulate the null hypothesis as one that meets the exchangeability requirement. That is, when the null hypothesis is true, the coupling of particular values of the dependent variable with particular values of the independent variable is random. This idea can be further explicated by imagining a failed experiment in which a researcher randomly assigns each of several participants to complete a state-anxiety questionnaire while experiencing either silence (the no noise group) or loud automobile traffic noise (the loud noise group). When the experiment ends, it is discovered that someone forgot to plug-in the machine that plays the recorded noise. Each score in the data set has a label, no noise or loud noise, but the labels are meaningless because of the failed manipulation. Thus, each score's attachment to the no noise versus the loud noise label might as well be random; in other words, when the null hypothesis is true, the scores are exchangeable across the labels. This concept shall be referred to as exchangeability under the null, to emphasize that the exchangeability defines the null hypothesis only. (The scores would not, and could not, be exchangeable when the null hypothesis is false.)

In permutation testing, the null hypothesis is not more specific than the proposition that the coupling of particular values of the dependent variable with particular values of the independent variable is random. For example, whereas a parametric test may assess the null hypothesis that the means are equal across groups, a permutation test is restricted to testing the null hypothesis of random coupling of values to condition labels. (Note: Howell (2013) states that this is also true for the Wilcoxon Rank-Sum test, which is a form of permutation test wherein the data are transformed to ranks prior to permutation.) Having conceptualized the null hypothesis as entailing the exchangeability of scores, the next step is to characterize the sample at hand in terms of one or more test statistics, such as the difference between the means, between the medians or between the variances.

Finally, the principle of exchangeability-under-the-null permits computation of a form of *p* value that is the probability that an effect as large as (or larger than) one computed prior to permutation would occur by chance. This chance process is simulated by repeatedly reassigning the scores to the labels that represent the levels of the independent variable. Following each permutation – that is, each set of reassignments – the test statistic is recomputed. The recomputed statistic is counted as having met the threshold of the original statistic if the former is equal to or more extreme than the latter. (In the case of a two-tailed test, the recomputed and the original statistic are transformed to their absolute values prior to comparison to one another.) An exact *p* value is obtained by calculating the proportion of all possible unique permutations that produce an outcome at least as extreme as the statistic computed for the original, unpermuted data (Fisher's Exact test is an example). Unless the data set is small, it is often more practical to obtain an approximate, Monte Carlo *p* value by calculating the aforementioned proportion for a large number of random permutations (which may occasionally, and by chance, include repetitions of particular permutation patterns) rather than for all possible permutations. In the remainder of this article, permutation test will refer to the Monte Carlo variety unless otherwise specified.

Note that permutation tests can be conceptualized as drawing samples (permutations) from a population, where the so-called population is what a parametric test would regard as a sample (Rodgers, 1999). However, the relationship between the sample and the population in parametric testing is not parallel to the relationship between the samples and the population in permutation testing. In parametric testing the goal is to use a sample statistic to infer a population parameter; in permutation testing the parameters of the so-called population are known and need not be inferred. Permutation-test logic does not use samples to make inferences about population parameters. Rather, a permutation test makes inferences about process. Specifically, it assesses the probability that a random process in which data values are coupled to condition labels would

produce a data set characterized by a given test statistic.

The permutation test is widely used in the form of Fisher's Exact test (and the Monte Carlo variant of Fisher's test) for analyzing relative frequencies in dichotomous data. Other forms of the permutation test have been used infrequently to date but can be implemented with the help of statistical software (e.g., Anderson, 2012; R, version 2.15.1; Stata, version 12.0).

A Permutation Test's Critical *p* Value (α) is not an Estimate of the Type I Error Rate

Some have argued that permutation tests entail highly restrictive assumptions. Hayes (1996), for example, wrote that the "permutation test is not distribution free" (p.1). As an example, he found that in simulated correlational data, under conditions in which the *x*, *y* correlation was zero and in which the variance among *y* observations differed as a function of *x*, both permutation tests and parametric tests rejected the null hypothesis more often than the rate suggested by the tests' *p* values.

Mewhort, et al. (2009) explored a phenomenon in which, when the null hypothesis is true, unequal variance interacts with unequal sample size. In a set of simulations, the researchers created pairs of populations that could have equal or unequal means and that could have equal or unequal variances. They then sampled from the populations to produce a large number of two-group data sets; the two groups could be equal or unequal in size. Thus, some of the data sets consisted of groups that differed both in size and in variance. The researchers conducted a permutation test on each set to assess the rate at which the tests produced Type I errors. With the critical *p* value (α) set at 0.05, the researchers found that when the population variances and the sample sizes were unequal and when the smaller sample had been drawn from the population with the higher variance, the actual Type I error rate was somewhat higher than 0.05. Conversely, when the smaller sample had been drawn from the population with the lower variance, the actual Type I error rate was somewhat lower than 0.05. The authors went on to propose an algorithm to correct the permutation test's apparent bias (for situations which the smaller group is characterized by higher variance).

I argue that there is a conceptual difficulty in using a permutation test's *p* value as a standard to evaluate a test's liberalness or conservativeness. In a traditional, parametric test, *p* refers to a portion of a hypothetical distribution of the values of a test statistic (e.g., a distribution of *t* scores) that would be obtained if one were to generate multiple data sets, via random sampling from a population or set of populations, and then compute the test statistic for each data set. This meaning of *p* allows a researcher to evaluate empirically whether *p* is liberal, conservative or unbiased. This is done by drawing multiple data sets from a population (or set of populations) in which the null hypothesis is known to be true. The researcher then establishes a critical *p* value (α) and computes the obtained Type I error rate as the proportion of simulated data sets in which the *p* value is less than or equal to α. The statistical test is liberal or conservative to the degree that the obtained Type I error rate tends to be higher or lower (respectively) than α.

The *p* value of a permutation test has a meaning very different from the one previously described, thus the term p_t will be used to refer to the *p* value produced by a traditional, parametric test, and p_{perm} will be used to refer to the kind of *p* value produced by a permutation test.

The value p_{perm}, in contrast to p_t, is about only the data at hand. Therefore, p_{perm} does not pertain to populations, or to multiple samples that could have been draw from a population or to multiple values of a statistic that could have been computed from multiple episodes of random sampling. Instead, p_{perm} is the rate at which the various possible re-assignments of scores to condition labels lead to an effect that matches or exceeds the magnitude of the effect in the un-permuted data. Likewise, p_{perm}'s critical value is conceptually distinct from p_t,'s critical value. This can be observed clearly when considering that a permutation test is valid and useful even when conducted on the entirety of a finite population. Consider the following example: For the uppercase letters in the Modern English alphabet is the central tendency of their ordinal positions significantly different for

symmetrical letters (e.g., "A") than for asymmetrical letters (e.g., "B")?

The question is nonsensical within the framework of parametric testing because the letters in this data set are not sampled (randomly or otherwise) from populations. But the question is eminently sensible from the standpoint of permutation testing. The null hypothesis is that the coupling of ordinal positions with the condition labels symmetrical and asymmetrical has occurred by chance. An alternative hypothesis, that the two conditions differ in the medians of their ordinal positions, is assessed by repeatedly permuting the values of the serial positions across the condition labels and recalculating the medians. For this example, the median ranks for the symmetrical and asymmetrical letters (uppercase) are 20 and 11, respectively (Test Statistic [50,000 permutations] = $Median_A - Median_B$, $p_{perm} \approx .047$). Thus, there is a significant tendency for symmetrical uppercase letters to occur later rather than earlier in the alphabet. Such a finding sets the stage for further scientific inquiry into the genesis of the alphabet, and more importantly, could not have arisen from classical, parametric statistical analysis.

There are two reasons why it would not make sense to ask whether the p_{perm} value, above, is liberal or conservative. First, there is no imaginable population – simulated or otherwise – from which additional Modern English alphabets could be sampled. Consequently, there is no basis for computing a Type I error rate across samples from such a population. Second, and just as importantly, even if one could imagine that the Modern English alphabet is just one random sample among many possible random samples, p_{perm} would still pertain only to permutations of the data at hand and not to a sampling distribution. Therefore, simulated (or otherwise obtained) Type I error rates such as those generated by Hayes (1996, 1997) and Mewhort, et al. (2009) cannot serve as standard to assess bias in p_{perm} because p_{perm} is unrelated to Type I error and p_{perm}'s critical value does not estimate a Type I error rate.

Is Possible to Estimate a Legitimate Type I Error Rate for a Permutation Test?

When a permutation test is conducted on an entire population of values, the concept of a Type I error rate is meaningless because there is no family of alternate samples over which a Type I error rate can be computed, thus in some circumstances, it is impossible to estimate a Type I error rate. But what about situations in which one computes p_{perm} for data that happen to have been sampled randomly from a population? In such a situation, can a Type I error rate be defined? The difficulty in answering yes in this case is that any such error rate must be defined with respect to the permutation test's null hypothesis and not with respect to a parametric null hypothesis. For example, imagine a simulation in which two populations have identical means and in which random samples are repeatedly drawn from the two populations. Computing the Type I error for a parametric test involves simply counting the proportion of samples that lead to the rejection of the null hypothesis of equal population means are equal. However, for a permutation test, such a parametric null hypothesis is not the relevant null hypothesis. Instead, the relevant null is that the arrangement of the data within a fixed set (not multiple, randomly sampled sets) reflects the random coupling of data values to condition labels. Thus, the random sampling procedure described above does not provide a basis on which to assess the rate of incorrect rejection of the particular null hypothesis tested by a permutation test.

Tests on Simulated Data

I now present what I believe to be a conceptually coherent assessment of Type I errors in permutation testing. Within a given simulation, the following procedure is employed:

(1) Decide on a set of numbers (i.e., a seed set) that will be constant throughout the simulation.

(2) Prior to conducting any permutation tests, perform many permutations of the data set. Each result of this initial set of permutations

is not part of a permutation test, but instead constitutes a data set to be analyzed via a permutation test. Such permutation-prior-to-testing is necessary because it produces data sets in which the null hypothesis (which is the true hypothesis in the present simulations) is of the permutation-test variety rather than the parametric-test variety.

It should be noted that while it would be possible to arrange things so that the two groups would differ systematically on some dimension (in their variances, for example), this would make the null hypothesis false rather than true. Such a situation would allow a permutation-test assessment of Type II errors (i.e., erroneous acceptance of the null hypothesis of random coupling of data values to condition labels), but it would miss the point of the present simulations, which is to assess Type I rather than Type II errors.

(3) Conduct the permutation tests in the usual manner. However, unlike typical simulations (e.g., Hayes, 1996, 1997; Mewhort, et al., 2009), each test is conducted on a data set generated by permuting a set that has not been sampled from a population and whose membership is the same for all tests.

(4) Compute the Type I error rate as the proportion of tests that reject the null hypothesis. But do not subsequently compare the Type I error rate to the conceptually distinct, critical p_{perm}.

Three simulations were conducted. In Simulation 1, the seed data set was the set of consecutive, non-repeating integers 1 through 30. Thus, the seed data were uniformly distributed. In Simulation 2, the seed data consisted of 30 values that were the squares of the 30 values in Simulation 1. Thus in Simulation 2, the seed data were exponentially distributed. In Simulation 3 the seed data consisted of the cubes of values in Simulation 1.

In all simulations, the procedure was as follows. The 30 seed values were randomly permuted 1,000 times across two equal-sized groups, yielding 1,000 data sets, with each set composed of 30 unique values divided among two groups (n = 15 per group). Each of the 1,000 two-group data sets was then submitted to a permutation test of the difference between means (Test Statistic = $mean_1 - mean_2$; 1,000 random permutations) and a permutation test of the difference between variances (*Test Statistic* = $variance_1 - variance_2$; 1,000 random permutations). Overall, each simulation included 1,000 times 1,000 permutations of data. For each permutation test, the critical p_{perm} was set at .05. The results are shown on Table 1.

Perhaps the most important aspect of Table 1 is that is does not permit assessment of the validity of p_{perm}. Because neither p_{perm} nor its critical value pertain to the Type 1 error rate, the table does not permit comparison of the Type I error rates to a standard (Note: The critical p_{perm} is not such a standard.) Another result is that the computed Type I error rates happen to correspond (roughly) to the value of p_{perm}, irrespective of whether the test assesses differences in means or differences in variances and irrespective of which seed data set served as the basis for generating the tested data. Finally, the simulations show that Type I error rates in permutation testing can meaningful (though such error rates are not essential), despite the absence of random sampling from populations.

Implications

The present arguments show that the critical p_{perm} for a permutation test is not a Type I error rate, and that consequently, a permutation test's validity does not depend on whether the numerical value of p_{perm} matches the Type I error rate that would be relevant to a parametric test. It is also clear that the true Type I error rate for a family of permutation tests will likely depend on how one chooses to define the family tests. In my view, the absence of a well-defined Type I error rate in no way impugns the validity or usefulness of permutation testing. The test computes the probability that a random process could have produced the observed assignment of data to condition labels, and this is a sufficient basis for deciding whether to reject the hypothesis of a random process. Thus, unlike other statistical approaches, permutation testing entails no parametric assumptions, and does not require the to-be-analyzed data to be randomly sampled from a population.

Table 1: The Mean of Group Means, Mean of Group Variances, Central Tendencies of p_{perm} and Type I Error Rates for Three Different Seed Data Sets

	Descriptives		Test of Means			Test of Variances		
Seed	*M* of M_1 (*M* of S_1^2)	*M* of M_2 (*M* of S_2^2)	*M* of p_{perm}	*Mdn* of p_{perm}	Type 1 Error Rate	*M* of p_{perm}	*Mdn* of p_{perm}	Type 1 Error Rate
Uniform	15 (72)	16 (73)	.509	.501	.047	.504	.512	.049
Squared	316 (73623)	315 (73797)	.493	.487	.058	.486	.474	.056
Cubed	7251 (62408539)	7164 (62667777)	.500	.500	.046	.500	.488	.045

Notes: For each permutation test, the critical value of p_{perm} was 0.05. This critical value does not refer to and is conceptually distinct from the Type I error rate. For each tested data set, the null hypothesis is true in that the coupling of data values to group labels is random.

Conclusion

Although this article demonstrates that permutation testing's validity is independent of the idea of Type I error, it does not resolve all of the outstanding questions concerning permutation tests. For example, there is some uncertainty as to the formal relationship between the null hypothesis and the test statistic used to reject that null hypothesis. Consider the following data set: Group A (10, 8, 77, 2, 40, 92, 88), Group B (7, 4, 2, 5, 5, 3, 3). With a critical p_{perm} of 0.05 and with 50,000 permutations, the null hypothesis of random coupling of values to group labels is rejected, whether the test statistic is the difference between means ($p_{perm} \approx 0.01$), between medians ($p_{perm} \approx 0.03$), or between variances ($p_{perm} \approx 0.02$). Yet, in all three cases, the rejected null hypothesis – that is, that the coupling of values to condition labels is random – is precisely the same hypothesis. Thus, there is unresolved ambiguity concerning the degree to which each of the three tests, above, provides unique information about the null hypothesis.

Limitations

There is also a question about the possible definitions of a type one error rate, within permutation testing. The simulations in this study utilized just a few of an infinite number of possible seed data sets. In principle each seed set could yield a different Type I error rate, therefore, it is not known whether this type of simulation typically yields Type I error rates that approximate the critical p_{perm}. Moreover, there are likely other methods (besides the permutation of seed data) for generating a family of data sets to serve as the basis for computing Type I error. This article does not resolve these questions. Nevertheless, if future research on permutation testing is to provide such answers, the Type I error rates must be established either empirically or by some means that does not interpret p_{perm} to refer to a Type I error rate. It should be reiterated, however, that the logic of permutation testing does not require establishing a Type I error rate.

Acknowledgements

The author wishes to thank Dr. Michael Doherty, Dr. John Tisak, Dr. Mark Appelbaum, Dr. Joseph Rodgers and Don Zhang for their helpful comments on earlier drafts of the manuscript.

References

Anderson, R. B. (2012). *Anderson's permutation tester*. Available at: https://sites.google.com/site/complexcognitionbgsu/calculators.

Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (*4th Ed.*). Boca Raton, FL: Taylor & Francis Group.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver & Boyd.

Good, P. I. (2011). *A practitioner's guide to resampling for data analysis, data mining, and modeling*. Boca Raton, FL: Taylor & Francis Group.

Hayes, A. F. (1996). Permutation test is not distribution-free: Testing H_0: $\rho = 0$. *Psychological Methods*, *1*, 184-198.

Hayes, A. F. (1997). Cautions in testing variance equality with randomization tests. *Journal of Statistical Computation and Simulation*, *59*, 25-31.

Hayes, A. F. (2000). Randomization tests and the equality of variance assumption when comparing group means. *Animal Behaviour*, *59*, 653-656.

Howell, D. C. (2012). *Statistical methods for psychology* (*8th Ed*.). Belmont, CA: Wadsworth Publishing.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (*4th Ed.*). Upper Saddle River, NJ: Pearson Education.

Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to *t* and *F* tests in biomedical research. *American Statistician*, *52*, 127-132.

Mewhort, D. J. K., Kelly, M., & Johns, B. T. (2009). Randomization tests and the unequal-N/unequal-variance problem. *Behavior Research Methods*, *41*, 664-667.

Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research*, *34*, 441-456.