The Directional Difference Test

Charles A. Holt[1] Daniel Kwiatkowski[2] Sean P. Sullivan[3]

June 2024

When analyzing data from experiments involving more than two treatments that can be ranked in terms predicted effect, it is natural to turn to statistical tests with directional alternative hypotheses. This paper uses Monte Carlo simulations to evaluate the power characteristics of the Directional Difference test, which has a test statistic defined as the sum of treatment differences for all pairs of sample observations in different treatments. When sample data are distributed according to familiar distributions like the normal, uniform, and logistic distributions, we find that the Directional Difference test is better able to detect treatment effects than the commonly used Jonckheere-Terpstra test. The Jonckheere-Terpstra test still performs best for heavy-tailed distributions. Experimenters wishing to exploit the informational content of multiple-treatment samples should consider replacing the Jonckheere-Terpstra test with the Directional Difference test in appropriate circumstances.

[1] University of Virginia, Department of Economics, Charlottesville, VA 22903; cah2k@virginia.edu.

[2] University of Virginia, Department of Economics, Charlottesville, VA 22903; dvk9af@virginia.edu.

[3] University of Iowa College of Law, Iowa City, IA 52241; sean-sullivan@uiowa.edu. Replication material for this work is available at https://github.com/sean-patrick-sullivan/working-paper-2024-ddt.

1 Introduction

Sample size matters for detecting patterns in experimental data but so does the choice of statistical test. At the analysis stage of an experiment, selecting the most powerful test among valid alternatives can mean the difference between reporting significant and insignificant results. At the design stage, selecting an inefficient test can mean wasting laboratory resources as larger samples are collected to compensate for the inefficient exploitation of sample information.

A striking example of how experimenters can obtain statistical sensitivity without expensive sample sizes involves the design of experiments around treatments of varying intensity—especially treatments with monotonic expected effect. In a recent survey of permutation methods for experimental data, Holt and Sullivan (2023) comment on the attractiveness of this design and propose a modification of the familiar Jonckheere-Terpstra test to better exploit the information in interval sample data. The modification replaces ordinal comparisons with interval differences. The authors conjecture that the proposed Directional Difference test would offer power advantages over the Jonckheere-Terpstra test in most situations outside of heavy-tailed distributions. This note evaluates that conjecture.

Using a Monte Carlo framework, we draw the following conclusions. First, the Directional Difference test exhibits superior power characteristics for data that are distributed normally, uniformly, logistically, and in at least one mixture model of normal and uniform distributions. Second, the Jonckheere-Terpstra test exhibits superior power characteristics for the exponential distribution and for heavy-tailed distributions like the Cauchy or normal with outliers. Third, both tests are valid in the sense of controlling type I errors, and both perform similarly for some distributions like the extreme value type I distribution.

In short, while the Directional Difference test does not uniformly outperform the familiar Jonckheere-Terpstra test, it does offer meaningful power gains in interesting and presumably common situations. Experimenters wishing to extract the greatest statistical power from their hard-earned samples should make both tests part of their statistical repertoires.

2 Motivation

For a concrete illustration of the surprising power of ordered treatment effect designs, consider a three-treatment design for an attacker-defender experiment reported by Holt, Sahu, and Smith (2022). Subjects were matched in fixed pairs for 30 rounds. In each round, defenders could decide whether to incur a cost to go on high alert, which would block an attack, attackers could decide whether to attack or not, with an attack decision resulting in a cost ($1, $3, or $5) that was deducted from the subject's earnings, regardless of whether the attack was successful. This design varies the intensity of the treatment change in attack costs, a change that has a somewhat unintuitive feature. Because the attack cost does not directly affect the defender's payoff, the attack rate must be the same in all three treatments for defenders to be indifferent and willing to randomize between going on high alert or not. Hence, the unique (if unintuitive) Nash equilibrium prediction is that the changes in attack cost will have no effect on attack rates. The intuitive directional alternative hypothesis is that a reduction in attack costs will increase attack rates. Table 1 shows average attack rates for 3 pairs in each attack-cost treatment. Note that, on average, attack rates tend to be higher with lower attack costs, but there are some reversals between attack rates in adjacent columns.

Table 1. Average attack rates by attacker-cost treatment.[*]

High ($5) attack cost	Medium ($3) attack cost	Low ($1) attack cost
0.10	0.33	0.57
0.23	0.20	0.43
0.23	0.47	0.40

[*] Selected data from Holt, Sahu, and Smith (2022), Appendix A, first three pairs per treatment.

With three observations per treatment, pairwise comparisons using the Mann-Whitney test could never reach *p*-value smaller than 0.05 in this design, there being nothing more extreme than the case where all three observations in one sample are smaller than all three observations in the other sample, which occurs with probability $1/\binom{6}{3} = 0.05$ under the null hypothesis of no treatment effect. For the resource cost of obtaining one more independent observation per treatment, a minimum *p*-value of about 0.014 could be achieved and with two more observations

per treatment, the minimum *p*-value would be driven down to 0.004. Holt, Sahu, and Smith (2022) use a standard Jonckheere-Terpstra test to assess this ordered conjecture over all three treatments simultaneously. Even with just three observations per treatment, the Jonckheere-Terpstra test based on a directional alternative hypothesis yields a *p*-value of about 0.017, there being 28 out of 1,680 possible permutations of the sample data that result in Jonckheere-Terpstra test statistic values as or more extreme than the observed value.

The ability of tests of ordered alternatives to outperform pairwise comparisons illustrates the potential power gains to be had from designing experiments around treatments of ordered intensity. And the Jonckheere-Terpstra test is not generally the most powerful test of this type. As Holt and Sullivan (2023) note, the Jonckheere-Terpstra test discards considerable sample information when experimental data are measured in interval or continuous terms. The test statistic compares individual observations in ordinal terms, ignoring the extremity of observed differences. To make use of this neglected sample information, Holt and Sullivan (2023) propose the Directional Difference test, based on an interval-difference analogy to the Jonckheere-Terpstra test statistic:

$$D = \sum_{s=1}^{k-1} \sum_{t=s+1}^{k} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} (x_{j,t} - x_{i,s}) \tag{1}$$

where k samples are arranged in order of increasing predicted treatment effect, $x_{i,s}$ is the *i*th observation in the sample data from ordered treatment s, $x_{j,t}$ is the *j*th observation in the sample data from ordered treatment $t > s$, and n_s, and n_t are respective sample sizes. This is essentially the Jonckheere-Terpstra test statistic but with difference terms, $(x_{j,t} - x_{i,s})$, in place of binary indicator functions, $1(x_{j,t} > x_{i,s})$.

Applied to the Holt, Sahu, and Smith (2022) data in Table 1, the Directional Difference test provides a smaller *p*-value of about 0.007, reflecting that only 12 of 1,680 possible permutations yield an observed value of D as or more extreme than the observed value. The intuition for this improvement is that, of the 28 "as or more extreme" permutations for the Jonckheere-Terpstra test statistic, fewer than half involve large enough differences to affect the Directional Difference

test statistic.[4] With fewer "as-or-more-extreme" permutations, the Directional Difference test yields a smaller *p*-value than the Jonckheere-Terpstra. Here, it supports rejection at the 1% level.

As the null distributions of both the Jonckheere-Terpstra test and the Directional Difference test must be constructed by permutation, selection between these tests should turn primarily on considerations of relative power—but, as the Directional Difference test is a recent proposal, no prior study of their comparative power characteristics exists. When collected data support interval interpretation (prices, revenues, efficiencies), Holt and Sullivan (2023) conjecture that the Directional Difference test's greater use of sample information would often provide power gains over the Jonckheere-Terpstra test like those illustrated in the above example. Recognizing that rank-based permutation tests often outperform other tests when dealing with outliers and heavy-tailed distributions, however, Holt and Sullivan (2023) also conjecture that the Jonckheere-Terpstra would have better power characteristics for heavy-tailed distributions. To test these conjectures, and thus to provide an empirical basis for choosing between these tests, we compare the tests in a battery of sampling conditions as summarized below.

3 Procedure

Our approach follows a Monte Carlo methodology illustrated and explained by Moir (1998) in the first issue of this Journal. We compare rejection rates of the two tests over eight different sampling distributions $F = \{F_1, \dots, F_8\}$, where the sample distributions are successively shifted up by a constant *d*, the true size of a treatment effects that varies on a range of differences $R = [0, 100]$. Rejection rates are computed by simulation. In our base case of $T = 3$ samples and $n = 4$ observations per sample, we simulate 40,000 data sets for each $(f, d) \in (F, R)$, with generated samples distributed $s_i \sim f_i + d(i-1)$, for $i \in \{1, 2, 3\}$. That is, our base case involves monotone treatment effects of symmetric size. We conduct both tests on each generated data set, and we

[4] For example, consider switching the 0.10 at the top of the left column with the 0.33, and at the same time switching the 0.23 in the left column with the 0.20 in the middle column. The first of these changes reduces the number of "binary wins" in the predicted direction and the second increases it back to its original level, so this permutation produces a Jonckheere-Terpstra test statistic value "as or more extreme" than the observed value. But, while the first switch reduces the Directional Difference sum by $0.33 - 0.10 = 0.23$, the second switch only increases the directional difference sum by 0.03, so this permutation produces a net 0.2 decrease in the Directional Difference test statistic and is not "as or more extreme" for the Directional Difference test.

record rejection rates in a running tally. For each simulated sample, the p-value (used to determine the incidence of rejections) is calculated as the proportion of permutations of the sample treatment labels that yield a test statistic as great or greater than the test statistic value for that sample.

This approach allows us to assess both the size and the power of these tests for different distributions and treatment-effect sizes. The size of a test, for a given distribution, is the rejection rate when the treatment effect is zero, $d = 0$. The power of a test, again for a given distribution, is the rejection rate as a function of the size of the non-zero treatment effect, $d > 0$. For larger values of d (bigger treatment effects), we expect the power of the tests to converge, but for smaller values of d (smaller treatment effects), we expect the power of the tests to differ meaningfully.

Table 2 summarizes the distributions from which we draw the simulated sample data in this study. For consistency with prior research in the experimental literature, we use roughly the same distributions as Moir (1998). For each distribution, we choose parameter values to ensure that errors are mean-zero and that both the Jonckheere-Terpstra test and the Directional Difference test reach approximately 100% power at treatment value $d = 100$.

Table 2. Distributions used in Monte Carlo simulation.

Distribution		Parameterization
1	Normal	Mean 0 and variance of 60
2	Uniform	Range [-110,110]
3	Normal with outliers	Combination of 95% chance of normal with mean 0 and variance 45 and 5% chance of normal with mean 0 and variance 130.
4	Normal plus uniform	Normal distribution with mean 0 and variance 35 added to an independent uniform distribution with range [-80,80]
5	Logistic	Location 0 and scale 30
6	Cauchy	Location 0 and scale 7
7	Gumbel	Location -23 and scale 40
8	Exponential	Scale 60, normalized to mean zero by subtracting 60

4 Results

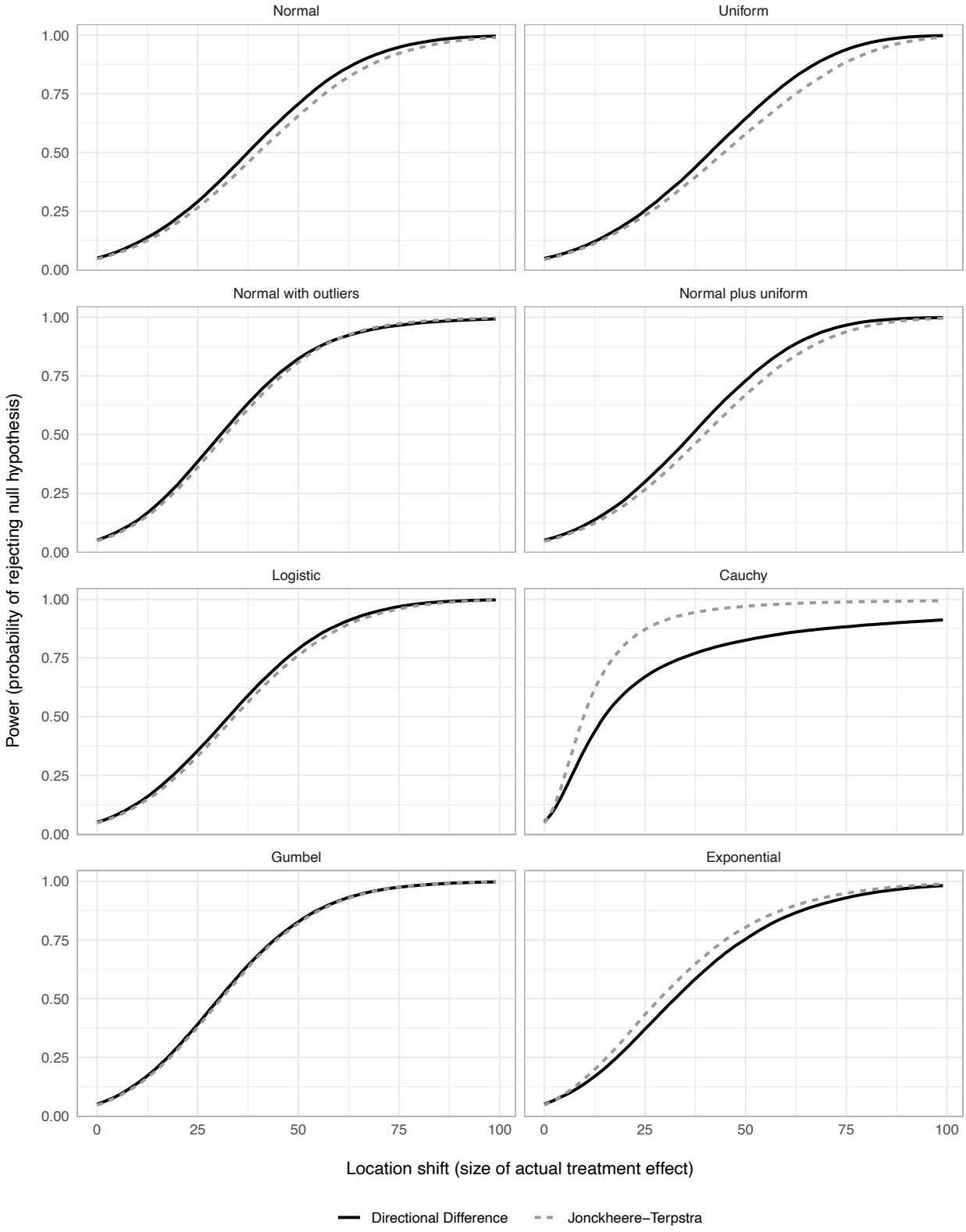
Power plots for the two tests are shown in Figure 1 for each of the 8 sampling distributions in Table 2. Additional power plots are provided in an appendix for alternative specifications such as asymmetric treatment-effects (different d values for samples 1 and 2 than for samples 2 and 3) and for different sample shapes ($T = 4$ samples and $n = 3$ observations per sample). Results from the alternative specifications support the same conclusions as Figure 1.

Figure 1 suggests several guidelines for pursuing statistical power when evaluating sample data in ordered treatment effect designs. First, when observed treatment effects are extreme, both tests are likely to reject the null hypothesis with high probability. When collected samples exhibit no overlaps or departures from the predicted order, for example, both tests will support the same conclusions. In the interesting case where observed treatment effects are less extreme, however, the choice of test has important power implications.

Second, when sample data are not affected by outliers, and do not exhibit heavy tails, the Directional Difference test should be preferred. In our simulations, the Directional Difference test outperforms the Jonckheere-Terpstra test for the normal distribution, the uniform distribution, the logistic distribution, and for a mixture of the uniform and normal distributions. Intuitively, when observed treatment effects are not too extreme, the Directional Difference test's sensitivity to differences between sample observations permits correct rejection of the null hypotheses in some situations where the Jonckheere-Terpstra test does not.

Third, for heavy tailed distributions, sensitivity to differences in collected observations becomes a liability, and the Jonckheere-Terpstra test offers superior power characteristics. In our simulations, the Jonckheere-Terpstra test outperforms the Directional Difference test for the Cauchy and the exponential distributions. This finding is consistent with other comparisons between rank-based and non-rank-based tests in the statistics literature. Miller (1997) notes the Wilcoxon test's superiority to the *t*-test for heavy-tailed distributions, and the sign test's superiority to the two-sample *t*-test for the Cauchy distribution and its asymptotic optimality for the two-tailed exponential (Laplace) distribution. Heavy-tailed distributions can produce samples containing extreme observations; these observations act like outliers and substantially influence the Directional Difference test while exerting little effect on a test based on ordinal comparisons, like the Jonckheere-Terpstra test.

Figure 1. Power plots for Jonckheere-Terpstra and Directional Difference tests.



Fourth, both tests perform similarly in important respects. Both tests control type I errors for all distributions in our study. When the treatment effect is zero, both tests support rejection of the null hypothesis with a probability of 5% or less. Power differences are also close enough to be ignored for some distributions, like the Gumbel (type I extreme value) distribution.

Conclusion

To make the most efficient use of laboratory resources, experimenters sometimes must look beyond familiar designs and statistical tests. One way to achieve impressive statistical sensitivity with small sample sizes is to design experiments around tests of monotone alternatives. When conducting statistical tests in this design, additional power can be gained from relying on the Directional Difference test instead of the Jonckheere-Terpstra test. The Directional Difference test is not uniformly more powerful for all distributions, but its more complete exploitation of interval-data sample information provides meaningful power gains over the Jonckheere-Terpstra test when sampling distributions are approximately normal, uniform, or similar. Both tests should be a part of every experimenter's statistical toolkit.

5 References

Holt, C. A., Sahu, R., & Smith, A. M. (2022). An experimental analysis of risk effects in attacker-defender games. *Southern Economic Journal*, 89(1), 185–215.

Holt, C. A. & Sullivan, S. P. (2023). Permutation tests for experimental data. *Experimental Economics*. 26, 775–812.

Jonckheere, A. R. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika*, 41, 133–145.

Miller, R. G. (1997). Beyond ANOVA: Basics of Applied Statistics. Boca Raton: Chapman & Hall/CRC.

Moir, R. (1998). A Monte Carlo analysis of the fisher randomization technique: Reviving randomization for experimental economists. *Experimental Economics*, 1(1), 87–100.

Terpstra, T. J. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae*, 14, 327–333.