The Stochastic Inequality Test - a Test for a Directional Treatment Effect*

Karl H. Schlag◇ James Tremewan

University of Vienna

April 13, 2022

Abstract

The *stochastic inequality test* is an exact non-parametric test that can be used to infer whether values in one random sample tend to be higher than in another. In addition it can be used to derive a confidence interval around an intuitive measure of effect size that is readily interpretable for both ordinal and cardinal data, and allows for *ex-ante* power analysis. Unlike other commonly used tests, such as the t-test and the Wilcoxon-Mann-Whitney test, this inference is valid without requiring additional, often unrealistic and untestable, assumptions about the shape of the underlying population distributions.

Keywords: Exact test, distribution free test, stochastic difference

JEL codes: C12, C14, C90

*We are very grateful to Natalia Shestakova for programming the test in Stata.

◇ Department of Economics, University of Vienna. E-mail: karl.schlag@univie.ac.at.

1 Introduction

The purpose of this paper is to encourage the use of the *stochastic inequality test* (Schlag, 2008) by experimental economists. In the following we show why a new test is needed and give a comprehensive explanation of the stochastic inequality test and its advantages over current practice. We aim to provide the working experimentalist with everything they need to know to easily and confidently apply the stochastic inequality test in their own research.

In a typical experiment, the researchers wish to compare a variable across a baseline and a treatment and establish which results in "better" outcomes. This we refer to as a *directional* effect. For example, one might be interested in identifying what mechanism tends to result in greater contributions to a public good or higher revenue in an auction. Current practice is to use the Wilcoxon-Mann-Whitney (Wilcoxon, 1945; Mann and Whitney, 1947) non-parametric test. However, it is well known that without additional, and often unrealistic, assumptions about the distributions of outcomes in the two treatments, the WMW test is simply a test of equality of distributions. Rejection can result from, for example, a difference in variance, skewness, or kurtosis, with no difference in central tendency. These false rejections can occur with large probability and thus make the test invalid for identifying a directional effect. Despite the clear inappropriateness of using WMW as a test of a directional effect, it is ubiquitously used for this purpose in the literature. This means that many claims that one treatment is in any sense better than another are not justified by statistical inference when using this test, potentially contributing to the replication crisis.[1]

The persistence of the WMW test as a directional test may be due to a lack of awareness, exacerbated by false claims in some statistical texts, papers, and wiki entries.[2] Even those aware of the test's limitations may continue with its use due to a view that there is little problem in practice, its power, unfamiliarity with alternatives,

[1]This paper is about between-subject experiments, however we note here that all these criticisms apply equally to the Wilcoxon signed-rank test, which is in general use for within-subject experiments.

[2]See Vargha and Delaney (1998) for a history of errors in the literature. Their paper is about the Kruskal-Wallace, which is a multi-group extension of the WMW test. The same issues apply to both tests.

or simply the convention in the field. We counter these objections by demonstrating that the problems are very real, that SIT is a straightforward alternative for testing directional effects, and that this alternative should prove no hindrance to publication. Furthermore, this paper clears up a common misconception that non-parametric tests are only good for identifying treatment effects, and not magnitudes of those effects.

Rather than looking at differences in means, the stochastic inequality test investigates differences in the probabilities of achieving a higher outcome. Specifically, the treatment effect is defined as the difference between the probability that the treatment yields a higher outcome and that the treatment yields a lower outcome, as compared to the non treated. This measure of effect size is called the *stochastic difference*. It has the advantage of being invariant to monotone transformations of the data.

Invariance to monotone transformations is particularly important when the data is measured on an ordinal scale, such as a Likert scale. It is also valuable for other experimental economics data, where the scale in the laboratory bears little relation to intended applications. Take, for example, a treatment variation that increases donations in a Dictator game by 10c out of a possible \$10. The magnitude of this effect on participants is trivial. However if the finding is externally valid to the population at large, and the donations put to good use, the cumulative impact could save many lives. The relationship between real-life donations and their ultimate benefit is not only likely to be of vastly different magnitude, but also non-linear. In fact, we often do not know how individuals turn outcomes into value or utility. The stochastic difference has the advantage that it only assumes a monotone relationship.

One advantage of the SIT is that it an exact test. Exact tests do not rely on asymptotic theory. Instead, the type I error probabilities apply to the actual sample sizes. Note that if a test is not exact then one has to act as if sample sizes are infinite. SIT is exact for testing any magnitude of the stochastic difference. Similarly, the WMW test is exact for testing identity of the two distributions, but not for testing if their central tendencies are equal. The two sample t-test is exact for comparing two normally distributed random variables that are known to have the same variances. The binomial test is exact for comparing Bernouli distributed random variables. Note that the t-test is an outlier in this list of exact tests as it builds on strong unverifiable assumptions that the data is normally distributed with equal variances. The other

three tests only require that the data within each sample are identically distributed. This is important because data from experimental economics is often clearly non-normal.

Two more properties make SIT and other tests of Schlag (2008) unique among non-parametric tests. First of all, SIT is not only a test to uncover if there is a treatment effect. It can also be used to identify a confidence interval for the magnitude of this treatment effect. Confidence intervals provide valuable information about the uncertainty surrounding the estimate. Their width can give insights as to whether enough data has been gathered, in particular when no significant treatment effect has been detected. Second of all, the test comes together with a bound on the type II error probability. This can be used for ex ante power analysis without having to make any assumptions on the shape of the distributions.

Two other tests of stochastic equality have been proposed (Cliff, 1993; Brunner and Munzel, 2000), but these rely on asymptotic arguments. Simulation studies show these earlier tests are over-sized, especially with small samples and asymmetric distributions (e.g Delaney and Vargha, 2002; Fagerland and Sandvik, 2009). While these tests can be used to construct asymptotically valid confidence intervals, no existing work contains a method for performing ex ante power analysis.

In Section 2 we argue in favour of the need for a new test, discussing the reasonableness or otherwise of the additional assumptions that are required for commonly used tests to perform the same tasks as SIT. In Section 3 we introduce the SIT together with all the information needed to implement it. In particular we present the stochastic difference, the test for identifying a treatment effect, an estimate of the magnitude of the effect together with confidence intervals, ex-ante power analysis and information on the software. In Section 4 we propose an approach for combining inference from the WMW test and the SIT. In Section 5 we conclude.

2 Why do we need a new test?

Data in experimental economics typically has two features that need to be accommodated by statistical tests: small samples and non-normal distributions. To deal with the first issue, we need *exact* tests that do not rely on "sufficiently large" samples. To

deal with the second, we need tests that apply to arbitrary distributions. We discuss each of these features in turn, then discuss the WMW test.

Perhaps the most crucial property of a test is its *size* (or equivalently *type I error probability*), which is the maximal probability of wrongly rejecting the null hypothesis, as calculated before the data is collected. A test has level α if its size is at most α and is called exact if it has the level that it is claimed to have. If the true size of a test is greater than the level it is claimed to have, then conclusions that are made are not based on the underlying methodology. In particular these conclusions are overly positive.

Many tests rely on asymptotic theory to prove their size, and thus can only be known to have the correct level as the sample size approaches infinity. The question then arises as to how many observations we need so that the true size is "close enough" to its asymptotic value. In fact, this depends on the precise distribution of the population, so in general cannot be known, and can in fact be arbitrarily large (Bahadur and Savage, 1956). In any case, the small numbers of observations in most economic experiments make asymptotic arguments of dubious relevance. Therefore, exact tests (that do not rely on asymptotics) are needed.

A commonly used test in the natural sciences is the two sample t-test, which is an exact test when both variables are normally distributed with the same variance. The assumption of normality appears reasonable enough for biological variables such as height, which are known to be very close to normally distributed. However most data in experiments is clearly non-normal, usually multimodal, often with modes at extremes. For example participants in a dictator game typically give half the pie, or nothing, with only a few values in between. Distributions of guesses in beauty contest games have spikes around choices predicted by level-k models. In public goods games subjects typically contribute all, half, or none of their endowment, and similarly in trust games both first and second movers tend to send all, half, or none of the maximum allowable. The fact that our data sets are usually non-normal, and not well described by other distributions used for parametric tests (e.g Poisson distribution), means that we need tests that are valid for arbitrary distributions.

The WMW test is an exact test of the null hypothesis that two distributions are identical. It is applicable to all distributions. So far it satisfies our basic requirements.

The trouble is that all a rejection tells you is that it is likely that the null is incorrect and the distributions differ. Rejection says nothing about *how* the distributions differ. The WMW test is not an exact test for equality of means (or medians). This has been made abundantly clear through simulations. For example, Forsythe et al. (1994) show that the WMW test comparing data drawn from uniform distributions with the same mean (either from $[0, 5]$ or $[2, 3]$) rejects approximately 8% of the time, substantially higher than the advertised 5% level. Fagerland and Sandvik (2009) contains more extreme results, for example estimating that the WMW test can understate the true size by 90 p.p. when two distributions share a mean but differ in skewness and variance.

We consider the above evidence sufficient to conclude that it is not appropriate to use the WMW test for identifying directional treatment effects. However, one could argue that this conclusion only applies to the specific distributions tested in the simulations, and that the WMW test is just fine for *your* data. We now elaborate on the assumptions that must be made about the distributions your data comes from for the WMW test to be valid for directional inference. Remember that you never know the true distributions, and assumptions about them can only be justified by theoretical arguments. For instance, tests for normality, such as the Jarque-Bara test, cannot reveal that the data is normally distributed. They can only provide significant evidence that the data is not normally distributed. As will soon become clear, any test of the assumptions discussed in the remainder of this section would make the WMW test redundant.

The original paper by Mann and Whitney proposed their test as a one-sided test of stochastic dominance[3], i.e. a directional test. Note that stochastic dominance entails an increase in both mean and stochastic difference. In order for the test to move from one of only a difference in distributions to a directional one, the space of alternative hypotheses had to be restricted to distributions that stochastically dominate each other. In other words, for rejection of the null to imply stochastic dominance, one must assume that rejection implies stochastic dominance. Stochastic dominance means that each point in the support of the distribution is either unchanged, or

[3]A cumulative distribution function F first order stochastically dominates G if $F(x) \geq G(x) \forall x$.

moves to the right. This assumption is much stronger than just assuming an increase in mean or stochastic difference because it implies that this is true for each point on the distribution, rather than true in the aggregate.

Typically it is hard to rule out that that the treatment cannot have an opposite effect on a subject. In fact, there are often reasons to expect heterogenous treatment effects. For example, informing subjects of a descriptive norm may be expected to lead to subjects who would have performed above the norm to perform worse, and those below to perform better, reducing variance with ambiguous theoretical impact on central tendency (the so-called "boomerang effect", Schultz et al., 2007). To assume that the set of alternatives only includes pairs of distributions where one stochastically dominates the other is even more problematic for two-sided hypotheses. This would imply that a treatment either moves every point in the distribution to the left, or every point to the right. Given that the theory underlying the hypothesis is not strong enough to make a one-sided hypothesis, it is hard to justify why any difference should be in the same direction for all points on the distribution.

In more recent times, using the WMW test for directional inference has tended to be justified by assuming the location shift model (Vargha and Delaney, 1998). This model requires even stronger assumptions than stochastic dominance. The location shift model not only requires that all points on the support move in the same direction, but that they move by exactly the same amount. The assumption of shift in location while maintaining identical shape is clearly rejected in much data from experimental economics. As discussed above, in games such as the public goods game and trust game, much of the data falls on either the highest or lowest possible value in all treatments. The fact that any, let alone most, of the data in each distribution is at each end of the support makes it *impossible* that one is a shift of the other. A further disadvantage with the location shift model is that it is not meaningful for ordinal data.

Directional inference is fundamental to experimental economics and there is not yet a test in general use that is adequate. This alone means we need a new test. Yet identifying a treatment effect is merely a first step. Usually we are also interested in the magnitude of any significant treatment effect, and the precision of our estimate. If we fail to find a statistically significant result, a measure of precision is also crucial,

Table 1: What can you do with the tests

	Mann-Whitney			SIT
	No assumption	Assuming SD	Assuming LS	
Difference in distributions	✓	✓	✓	✓
Directional inference	×	✓	✓	✓
Bounded data	✓	✓	×	✓
Ordinal data	✓	✓	×	✓
Power analysis	×	×	×	✓

as it provides information as to whether the failure to reject may be due to small sample size. In other words, we need confidence intervals. When confidence intervals are presented in experimental economics studies, they are almost always based on t-tests (or regressions assuming normal errors), and thus typically invalid for all the reasons given above. There seems to be little awareness that confidence intervals can be constructed using non-parametric tests, although basing them on the WMW test requires assuming the location shift model, which we would not recommend. The SIT represents a test for any magnitude the effect size and reveals a confidence interval, again with no side assumptions on the underlying distributions.

As a final desirable feature of a test we wish to be able to perform an ex ante power analysis. This would enable experimenters to estimate in advance the number of subjects likely necessary to find statistical evidence of a given effect size. The type of information is available for the SIT and presented in this paper. Just like any other test, one can simulate the rejection probability under the SIT for any particular underlying distributions. However the set of underlying distributions is typically too large to get any understanding of its general properties. To perform ex ante power analysis with the WMW test requires assumptions about the precise shapes of the distributions, which entirely defeats the point of using a non-parametric test in the first place. With SIT, there is a simple formula that determines the type II error probability with assumptions on only two parameters: the magnitudes of the stochastic difference under the null and under the alternative hypothesis.

Table 1 summarizes the side assumptions the WMW test requires to perform each of the tasks discussed in this section.

3 The Stochastic Inequality Test

In this section we provide a detailed elaboration of the stochastic inequality test, which was introduced in Schlag (2008). We begin by explaining precisely in what sense the SIT shows that one treatment is better than another, and how the effect size should be interpreted. We then recall the proof that SIT is an exact test for identifying the existence of a directional treatment effect. We go on to explain how to make an inference about the specific size of the effect and use this to derive confidence intervals. Finally we show how *ex ante* power analysis can be performed.

3.1 Stochastic difference

The stochastic inequality test compares outcomes in two different treatments in terms of the *stochastic difference* of the two underlying distributions. This measure quantifies how much more likely it is that a random draw from one distribution is greater than a random draw from another than that it is smaller. Formally, the stochastic difference between two distributions X_1 and X_2 is defined as:

$$d(X_1, X_2) = P(X_2 > X_1) - P(X_2 < X_1)$$

An equally meaningful interpretation of this measure comes from the observation that it is a rescaling of the difference between the expected average rank in the two samples, ranking the data from low to high. This follows easily from Vargha and Delaney (1998). As such, it is the natural analogue to mean comparison for ordinal data.

We say that a distribution is *stochastically greater (less)* than another if the stochastic difference is positive (negative), and *stochastically equal* if the stochastic difference is zero (Vargha and Delaney, 2000). Informally, we say that X_2 tends to be larger than X_1 if X_2 is stochastically greater than X_1.

Before explaining the virtues of the stochastic difference as a measure of how two random variables X_1 and X_2 differ, we first summarize other related measures. Two similar measures have been proposed in the past. The common language statistic $P(X_2 > X_1)$ was suggested by McGraw and Wong (1992) who identified its favourable

features. It is invariant to monotone transformations of the data and has a straightforward interpretation readily understood by non-statisticians. In contrast, the magnitude of a shift in measures such as mean or median clearly depends on scale. Statistics which normalize the effect size in some way, such as Cohen's d, are only invariant to uniform scaling, require an understanding of means and standard deviations, and differ in interpretation depending on the precise shapes of the distributions.

However the common language statistic is less useful when data is discrete and $P(X_2 = X_1) \neq 0$. To see this, suppose $P(X_2 > X_1) = 0.4$. Does this suggest that treatment 1 is better treatment 2? In fact, we cannot tell: if $P(X_2 = X_1) < 0.2$, then $P(X_1 > X_2) > 0.4 = P(X_2 > X_1)$ and treatment 1 is to be preferred to treatment 2, however if $P(X_2 = X_1) > 0.2$, the reverse is true. To address this concern, Vargha and Delaney (2000) introduced the measure of *stochastic superiority* defined by $A_{12} = P(X_1 > X_2) + \frac{1}{2}P(X_1 = X_2)$.

The stochastic difference shares the advantages of Vargha and Delaney (2000)'s measure in being invariant to monotone transformations, readily interpretable for arbitrary distributions, and appropriate for discrete data. An added attraction is its symmetry, in the sense that $d(X_2, X_1) = -d(X_1, X_2)$.

The stochastic difference involves comparing those better and those worse off. On the other hand, mean comparisons can produce a positive effect even if benefits are only realized by a minority. Therefore, the stochastic difference captures the impact of a treatment on each individual, without trading off differential effects across subjects. It is invariant to monotone transformations of the data and hence, unlike mean comparisons, is not influenced by the scale with which outcomes are measured. Interestingly, the stochastic difference is equal to the difference between expected average ranks in the two samples. As such, the stochastic difference can be interpreted as a natural form of scale free mean comparison. Last but not least, the stochastic difference can be tested using very small sample sizes. In contrast, means tests that are exact require substantially larger sample sizes or mandate an assumption of normality and equal variances.

These nice features make the stochastic difference an ideal concept for comparing treatments. It is particularly relevant when an agent must make a single choice from a distribution, and must select which distribution to choose from. Examples include

someone choosing whether to invest their retirement savings in an actively or passively managed fund or evaluating whether the extra productivity of hiring an MBA is likely to be worth the wage premium.

Sometimes means are the primary statistic of interest, for example when investigating programs where the ultimate goal is to increase total productivity in the country. Nevertheless, due to the difficulty of testing differences in means, stochastic difference remains useful for checking robustness. The stochastic difference and its associated test can help check if a discovered effect in means can be statistically undermined albeit under a slightly different concept of difference. Additional checks are generally welcome when evaluating interventions.

Stochastic difference also has direct relevance when interested in group averages. The stochastic difference can answer the question "which treatment is most likely to result in a better average outcome?" This has a clear connection to the common experimental design where subjects interact repeatedly in matching groups, and matching group averages are used as independent observations.

The user should note that stochastic difference is not a transitive relationship. It is possible that A is stochastically greater than B, B stochastically greater than C, and C stochastically greater than A (see Vargha and Delaney (1998) for an example). This intransitivity is to be expected under such ordinal inference, similar to the appearance of the Condorcet cycle under majority voting. This is not a weakness, but rather a reminder that a single statistic of a distribution will not always be sufficient. Given a non-transitive relationship between three distributions one must look closer at where the differences lie, and make a decision based on the features of the distributions that are most important for the application at hand. For example, if poverty is a concern one should focus on the lower bound, or for promoting innovation it may be the upper tail that matters most.

3.2 Identifying a Treatment Effect Using SIT

Here we present the SIT, a test that can be used to infer how two treatments differ in terms of the stochastic difference. Given the discussion in the previous section, a rejection by this test has two natural interpretations. First, it is evidence that a

random draw from one treatment is likely to be greater than a random draw from the other. Second, it is evidence that the expected average rank of that treatment is higher than the other.

Conceptually, the test consists of two parts. The first building block is a *randomized test*. This test is essentially a sign test but, as a randomized test, rejection is probabilistic. While randomized tests are perfectly valid in theory, in practice they are generally not accepted because the way in which data is treated will depend on the outcome of the randomization. To resolve this issue, the SIT removes randomness. More precisely, it yields a rejection whenever the randomized test, evaluated at a lower significance level, yields a rejection probability that lies above a given threshold.

We now present the two parts of the SIT in more detail. First, the data in the two sets of observations are randomly matched in pairs. So if the sample sizes are unequal then some observations are unmatched. Hence it is most efficient when samples are balanced. Next one checks whether there is evidence that $X_2 > X_1$ occurs more often than $X_2 < X_1$ among these matched pairs using the sign test. Note that the sign test is an exact test for its original setting where data is given as matched pairs. Embedding the sign test in this construction we obtain an exact randomized test. This is because the sign test is applied to a sample of matched pairs that is the result of a random pairing of the original data. Because the sign test is a uniformly most powerful unbiased test, so is this test. If one is happy with a randomized test we are done.[4] In particular, the randomized test inherits the size of the sign test used therein.

To transform the randomized test into a nonrandomized test, we run the randomized test on *all possible matchings*, and reject if "enough" of these tests result in rejection. The question now is how to choose the size of the randomized tests and the proportion of rejections such that the overall test can be guaranteed to have the desired size. Suppose that our sign test has size $\hat{\alpha}$ and we reject overall if the

[4]Similarly we are done if there is some exogenous order that can be used to match the pairs. This order has to be precommitted to prior to gathering the data. For instance, the order can be the order in which the data was created. The downside that makes such tests not acceptable is that not observations are treated equally and that it is very hard to precommit to such an order.

proportion of rejections in our subtests is above θ. It can easily be shown that the resulting test has size $\alpha = \frac{\hat{\alpha}}{\theta}$ (see Appendix A). Therefore, for any value $\theta \in (0, 1)$ we can create a test of level α which rejects if a fraction of θ randomized tests reject at the $\theta\alpha$ level. Note that while it is easy to prove that this test has level alpha, the size of this test is unknown apart from the fact that it is at most equal to alpha.

So far we have a family of tests, indexed by the parameter theta. It remains to select an appropriate value of θ. It is crucial that this is done before the data is collected, and for greater credibility a method for selecting θ should be agreed on as a standard convention. The following method for choosing theta has been proposed by Schlag (2008) and has been used ever since. One chooses the value of theta according to the power. However, the power functions for the different values of theta cannot be ordered according to FOSD. Typically, one value of theta does not lead to a uniformly more powerful test when comparing to a different value. So we compare the different tests according to their ability to guarantee a rejection probability that lies above 0.5. One selects the value of theta that leads to the largest set of alternative hypotheses under which rejection probability is at least 0.5. The value of 0.5 was chosen as a salient value that identifies values of the stochastic difference that are neither too close or too far from the null hypothesis. In particular, the value of θ is chosen based on the sample sizes but not on the observations.

Note that the stochastic inequality test is not constructed in the same way as most other tests. In particular, there is no test statistic to be reported. The common procedure is to define a test statistic, to compute the distribution of this test statistic under the null hypothesis and then to reject the null if the test statistic of the observed data lies in the a tail of this distribution. However this method has not been successfully applied to create an exact test. As the SIT is created under a different methodology, there is no test statistic that one can report. If one wishes to demonstrate how easy it was to reject the null hypothesis, one can report how close the fraction of rejections was to the threshold θ. However, a more advisable approach is simply to compute the p value. This is obtained by varying α until the null hypothesis is just barely rejected.

The intuition behind the SIT is relatively simple. A typical statistical test rejects the null hypothesis in favour of identifying a treatment effect if the estimated effect

is sufficiently large. The SIT does something a bit more general, rejecting the null hypothesis if sufficiently many matchings yield a sufficiently large estimate. Note that there are two cut-offs in this criterion: "Sufficiently many matchings" refers to there being a higher proportion than theta, while "sufficiently large estimate" refers to the cut-off determined by the binomial test.

3.3 Magnitudes and confidence intervals

An unbiased estimate of the stochastic difference between the nontreated and treated can be computed as follows. It is the difference between the proportion of possible matchings in which the treated outcome is greater and the proportion in which the nontreated outcome is greater. Formally, the estimate $\hat{d}(X_1, X_2)$ of the stochastic difference between X_1 and X_2 is given by:

$$\hat{d}(X_1, X_2) = \frac{1}{n_1 n_2} \sum_{k=1}^{n_2} \sum_{j=1}^{n_1} \left(\mathbb{I}\{x_{2,k} > x_{1,j}\} - \mathbb{I}\{x_{2,k} < x_{1,j}\} \right)$$

To build a confidence interval for the stochastic difference, the procedure for identifying a treatment effect needs to be modified slightly to test the hypothesis

$$d(X_1, X_2) = P(X_1 > X_2) - P(X_1 < X_2) = d_0$$

for values of $d_0 \neq 0$. Roughly speaking, the test is as described in the previous section, but the sign test is replaced by a binomial test which asks if the proportion of times an observation from the treated group is greater than an observation from the untreated group is sufficiently large. More precisely, code the pairs where $X_1 > X_2$ with 1 and those with $X_1 < X_2$ with 0. Drop the pairs with $X1 = X2$ with probability $\frac{1}{1+|d_0|}$, otherwise code them with 0 if $d_0 > 0$ and with 1 if $d_0 < 0$. Then use the binomial test to test if the proportion of 1 in the resulting sample is significantly larger than $\frac{1+d_0}{2}$. Intuitively, the number $\frac{1+d_0}{2}$ comes from the following: after removing ties as described above, $P(X_1 < X_2) = 1 - P(X_1 > X_2)$, so our hypothesis can be rewritten $2P(X_1 > X_2) - 1 = d_0$, which means that $P(X_1 > X_2) = \frac{1+d_0}{2}$.

The $100(1-\alpha)\%$ confidence interval is then simply the set of values of d_0 such that the above hypothesis cannot be rejected at level α.

3.4 Ex Ante Power Analysis

There is a simple formula for determining how powerful the SIT is for a given sample size. This formula determines the type II error probability as a function of only two parameters. These are the magnitudes of the stochastic difference under the null and under the alternative hypothesis.

Given estimates of the magnitudes of the stochastic difference under the null and under the alternative hypothesis and significance level, it is possible to compute the number of observations necessary for the stochastic inequality test to reject with the power desired by the experimentalist. Thus, when intending to use the SIT, one can preselect an appropriate sample size. Details of the computation can be found in Appendix C.

3.5 Implementation and Software

The stochastic inequality test has been implemented in both Stata and R. [5] The package allows the user to test for treatment differences, calculate effect size, p-values, and confidence intervals, and perform ex ante power analysis. To reduce computing time, rather than using all possible pairings of the data, a Monte Carlo procedure is used which, crucially, retains the exact nature of the test (see package documentation for details).

4 A Proposal for Two-Step Testing

The SIT can, of course, be used as a stand-alone test to investigate tretment differences. However, we propose to use it in conjunction with the WMW test in a two-step procedure. We first describe this approach, then discuss its benefits.

Step one is to run a WMW test to decide whether or not there is a difference in distributions. If the null hypothesis is not rejected, we are done, as there is no evidence of any kind of treatment effect. If the null is rejected, we conclude that there is some kind of treatment effect, but not necessarily a directional one. We therefore

[5] Available at https://homepage.univie.ac.at/karl.schlag/statistics.php.

move on to step two, which is to run a stochastic inequality test which, if the null is rejected, gives us evidence of a directional effect.

There are two advantages of keeping the WMW test test as a first pass. The less scientific reason is that as long as statistical significance is an advantage or requirement for publication in many journals (something we strongly disagree with), making the less powerful SIT an additional test rather than a replacement should encourage uptake of a proper directional test: authors are no less likely to identify a statistically significant treatment difference (although directional claims may need to be tempered). Potentially more interesting is in the possibility that the WMW test rejects but SIT does not. Of course it is possible that a directional difference exists, but there are too few observations for the requisite power. On the other hand, it may prompt the researcher to seriously consider heterogenous treatment effects, leading to further insights and ideas for future research.

Note that it may be possible that the WMW test does not reject the null hypothesis of equal distributions but that the SIT detects a treatment effect. We conjecture that this is not possible, but have no proof. This possibility does not change the value of the approach described in this section.

With respect to identifying directional treatment differences in within-subject experimental designs, the commonly used Wilcoxon rank-sum (WRS) test has the same shortcomings as the WMW test. A two-step approach can be easily followed in this case without need for a new test: first perform a WRS test for differences in distributions, then a sign test for a directional effect.

Finally, we would also like to point out to those nervous of using a novel approach, this methodology has already been accepted by many referees in a variety of economics journals. For published examples, see Galbiati et al. (2013), Montag and Tremewan (2018), Rizzolli and Tremewan (2018), Krysowski and Tremewan (2020), and Lippert and Tremewan (2021).

5 Conclusion

In this paper we have argued that experimentalists need a new test. The most crucial reason is that the current gold standard test for directional inference, the

Wilcoxon-Mann-Whitney test, is over-sized. This means that there are likely more false-positives in the literature than their should be, potentially contributing to the "replication crisis". Furthermore, confidence intervals in most experimental economics publications are typically based on an assumption of normally distributed errors, an assumption which is clearly inappropriate for much of our data. As a solution to these problems, we propose the adoption of the stochastic inequality test. As an exact test for arbitrary distributions, not only does it solve these problems but, in addition, allows for ex ante power analysis. We hope that our readers are persuaded by our arguments, and the two-step procedure outlined in Section 4 eases the entry of the stochastic inequality test into our field.

References

Bahadur, R. R. and L. J. Savage (1956). The nonexistence of certain statistical procedures in nonparametric problems. The Annals of Mathematical Statistics 27(4), 1115–1122.

Brunner, E. and U. Munzel (2000). The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation. Biometrical Journal: Journal of Mathematical Methods in Biosciences 42(1), 17–25.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. Psychological Bulletin 114(3), 494.

Delaney, H. D. and A. Vargha (2002). Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. Psychological Methods 7(4), 485.

Fagerland, M. W. and L. Sandvik (2009). The Wilcoxon–Mann–Whitney test under scrutiny. Statistics in Medicine 28(10), 1487–1497.

Forsythe, R., J. L. Horowitz, N. E. Savin, and M. Sefton (1994). Fairness in simple bargaining experiments. Games and Economic Behavior 6(3), 347–369.

Galbiati, R., K. H. Schlag, and J. J. Van Der Weele (2013). Sanctions that signal: An experiment. Journal of Economic Behavior & Organization 94, 34–51.

Krysowski, E. and J. Tremewan (2020). Why does anonymity make us misbehave: Different norm or less compliance? Economic Inquiry.

Lippert, S. and J. Tremewan (2021). Pledge-and-review in the laboratory. Games and Economic Behavior 130, 179–195.

Mann, H. B. and D. R. Whitney (1947). On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics, 50–60.

McGraw, K. O. and S. P. Wong (1992). A common language effect size statistic. Psychological Bulletin 111(2), 361.

Montag, J. and J. Tremewan (2018). Let the punishment fit the criminal: an experimental study. Journal of Economic Behavior & Organization.

Rizzolli, M. and J. Tremewan (2018). Hard labor in the lab: Deterrence, non-monetary sanctions, and severe procedures. Journal of Behavioral and Experimental Economics 77, 107–121.

Schlag, K. H. (2008). A new method for constructing exact tests without making any assumptions. Department of Economics and Business Working Paper 1109, Universitat Pompeu Fabra.

Schultz, P. W., J. M. Nolan, R. B. Cialdini, N. J. Goldstein, and V. Griskevicius (2007). The constructive, destructive, and reconstructive power of social norms. Psychological Science 18(5), 429–434.

Vargha, A. and H. D. Delaney (1998). The kruskal-wallis test and stochastic homogeneity. Journal of Educational and Behavioral Statistics 23(2), 170–192.

Vargha, A. and H. D. Delaney (2000). A critique and improvement of the cl common language effect size statistics of mcgraw and wong. Journal of Educational and Behavioral Statistics 25(2), 101–132.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin 1(6), 80–83.

Appendix A The Derandomization Trick

Here we recall the so-called derandomization trick of Schlag (2008) that shows how to create an exact nonrandomized test from an exact randomized test. This is how it works. Evaluate the randomized test at a more conservative level and then reject with certainty whenever the randomized test produces a rejection probability above a given threshold. The following result shows the details. Specifically, if the threshold is given by θ then evaluate the randomized test at level $\theta\alpha$ to generate a nonrandomized test with level α.

Proposition 1 *Let ϕ be an exact randomized test with level $\bar{\alpha}$, let $\theta \in (0,1)$ and let ϕ_θ be the test which rejects the null hypothesis if and only if the rejection probability under ϕ is above θ. Then ϕ_θ has level $\bar{\alpha}/\theta$. So if $\bar{\alpha} = \theta\alpha$ then ϕ_θ has level α.*

Proof. Let z be the data sample that is distributed according to P_Z. The fact that ϕ has level $\bar{\alpha}$ means that

$$\int \phi(z) dP_Z(z) \leq \bar{\alpha}.$$

Let $\phi_\theta(z) = 1$ if $\phi(z) \geq \theta$ and $\phi_\theta(z) = 0$ if $\phi(z) < \theta$. Then

$$\begin{aligned}
\bar{\alpha} &\geq \int \phi(z) dP_Z(z) \\
&= \int_{z:\phi(z)\geq\theta} \phi(z) dP_Z(z) + \int_{z:\phi(z)<\theta} \phi(z) dP_Z(z) \\
&\geq \int_{z:\phi(z)\geq\theta} \phi(z) dP_Z(z) \\
&\geq \int_{z:\phi(z)\geq\theta} \theta dP_Z(z) \\
&= \theta \int_{z:\phi(z)\geq\theta} 1 dP_Z(z) \\
&= \theta \int_{z:\phi(z)\geq\theta} 1 dP_Z(z) + \theta \int_{z:\phi(z)<\theta} 0 dP_Z(z) \\
&= \theta \int \phi_\theta(z) dP_Z(z)
\end{aligned}$$

and hence ϕ_θ has level $\bar{\alpha}/\theta$. ■

The same type of arguments can be used to identify the type II error probability of the test ϕ_θ.

Proposition 2 *Let ϕ be an exact randomized test with level $\bar{\alpha}$, let $\theta \in (0,1)$ and let $\phi_\theta \in \{0,1\}$ be such that $\phi_\theta(z) = 1$ if and only if $\phi(z) \geq \theta$. Then the type II error probability of ϕ_θ is at most $\frac{1}{1-\theta}$ times the type II error probability of ϕ.*

Proof. Let z be the data sample that is distributed according to P_Z. We compute

$$\begin{aligned}
\int \phi dP_Z(z) &= \int_{z:\phi(z)\geq\theta} \phi dP_Z(z) + \int_{z:\phi(z)<\theta} \phi dP_Z(z) \\
&\leq \int_{z:\phi(z)\geq\theta} \phi dP_Z(z) + \theta \int_{z:\phi(z)<\theta} 1 dP_Z(z) \\
&= \int_{z:\phi(z)\geq\theta} \phi dP_Z(z) + \theta \left(1 - \int_{z:\phi(z)\geq\theta} 1 dP_Z(z)\right) \\
&= \int \phi_\theta dP_Z(z) + \theta \left(1 - \int \phi_\theta dP_Z(z)\right) \\
&= 1 - (1-\theta) \int (1 - \phi_\theta) dP_Z(z)
\end{aligned}$$

and hence obtain

$$\int (1-\phi_\theta)\, dP_Z(z) \leq \frac{1}{1-\theta}\left(1-\int \phi dP_Z(z)\right) = \frac{1}{1-\theta}\left(\int (1-\phi)\, dP_Z(z)\right)$$

where $\int (1-\phi')\, dP_Z(z)$ is the type II error probability of the test ϕ'. ■

Appendix B Choosing θ

The following methodology has been proposed by Schlag (2008) for selecting the threshold θ, here applied to the test of stochastic inequality. We first present the formalities and then provide some intuition. We do this here for the family of one sided tests ϕ_θ for $H_0 : d(X) \leq d_0$ with level α. Two-sided tests are then constructed by combining two one sided tests that each have half the level.

Let Z_X be the two independent random samples generated from joint distribution X. Let $\phi^{(\alpha')}$ be the randomized test with size α' for testing $H_0 : d(X) \leq d_0$. Let $B(\theta', d')$ be the bound on the type II error probability of the stochastic inequality test with parameter θ' when the data is generated by the joint distribution X with $d(X) = d'$. Following Proposition 2,

$$B(\theta', d') = \frac{1}{1-\theta} \sup_{X: d(X)=d'} \left(\int \left(1-\phi^{(\theta'\alpha)}\right) dP_{Z_X}(z)\right).$$

Now we incorporate the property that $B(\theta', d')$ is weakly decreasing in d'. Let

$$d_{1/2}(\theta') = \min\left\{d' : B(\theta', d') \leq \frac{1}{2}\right\}.$$

So $B(\theta', d') \leq \frac{1}{2}$ if and only if $d' \geq d_{1/2}(\theta')$. The proposal is to choose

$$\theta \in \arg\min_{\theta'} d_{1/2}(\theta').$$

Here is some intuition. ϕ_θ is an exact test with a parameter θ. The parameter θ is part of the test. Hence θ has to be chosen prior to observing the data. Of course the choice of θ may depend on the two sample sizes. We are interested in detecting a treatment effect when there is one. The best test would be one that has the smallest type II error probability given the true treatment effect. However we do not know the true treatment effect and different values of θ will differ in their power for detecting small or large treatment effects. The idea is to be good at detecting treatment effects

that can be detected. One might consider a treatment effect detectable if the type II error probability of the test lies below 1/2. Then $d_{1/2}(\theta')$ is the smallest treatment effect that is detectable when choosing $\theta = \theta'$ where all treatment effects above $d_{1/2}(\theta')$ are also detectable. So it is natural to choose the value of θ that makes the most treatment effects detectable. This means that we choose the value θ that yields the smallest value of $d_{1/2}(\theta')$ across all θ'. Of course, using a different threshold in the conceptualization of detectability will lead to a different choice of θ. The value of 1/2 has been chosen as it is salient and as it turns out to identify the region in the power curve where changes in θ have the most impact.

Appendix C Power analysis

Here we provide the formula of the type II error probability of the test of stochastic inequality of $H_0 : d(X_1, X_2) \leq 0$.

Let $\mathbb{I}$ be the indicator function, so $\mathbb{I}(x) = 1$ if $x \geq 0$ and $\mathbb{I}(x) = 0$ if $x < 0$. Let p denote the probability of success of a Bernoulli random variable.

Let g_1 be the randomized (UMP) binomial test of $H_0 : p \leq z$ with size α when observing k successes in a random sample of n observations. So

$$\begin{aligned} g_1(k, n, z, \alpha) &= \mathbb{I}\left(\alpha - \sum_{j=k}^{n} \binom{n}{j} z^j (1-z)^{n-j}\right) \\ &\quad + \left(1 - \mathbb{I}\left(\alpha - \sum_{j=k}^{n} \binom{n}{j} z^j (1-z)^{n-j}\right)\right) \cdot \mathbb{I}\left(\alpha - \sum_{j=k+1}^{n} \binom{n}{j} z^j (1-z)^{n-j}\right) \\ &\quad \cdot \frac{\alpha - \sum_{j=k+1}^{n} \binom{n}{j} z^j (1-z)^{n-j}}{\binom{n}{k} z^k (1-z)^{n-k}}. \end{aligned}$$

Let g_2 be probability of rejection under the randomized binomial test of $H_0 : p \leq z$ when $p = \mu$. So

$$g_2(\alpha, \mu, n, z) = \sum_{k=0}^{n-1} \binom{n}{k} \mu^k (1-\mu)^{n-k} g_1(k, n, z, \alpha) + \mu^n \left(w(\alpha - z^n) + (1 - w(\alpha - z^n)) \frac{\alpha}{z^n}\right).$$

Following Proposition 2, the type II error probability of the SIT of $H_0 : d(X_1, X_2) \leq 0$, when $d(X_1, X_2) = d'$, is equal to $\frac{1}{1-\theta}$ times the type II error probability of the randomized binomial test with size $\alpha\theta$ for testing $p \leq \frac{1}{2}$ when $p = \frac{1+d'}{2}$ and sample size

is $n = \min\{n_1, n_2\}$. So the type II error of the SIT equals

$$\frac{1 - g_2\left(\theta\alpha, \left(1 + d'\right)/2, n, 1/2\right)}{1 - \theta}.$$

Note that this is an exact bound on the type II error probability. However, this formula is not tight, in practice performance is much better.