WORKING PAPERS

Who gives Direction to Statistical Testing? Best Practice meets Mathematically Correct Tests

Karl H. Schlag

October 2015

Working Paper No: 1512

DEPARTMENT OF ECONOMICS

UNIVERSITY OF VIENNA

All our working papers are available at: http://mailbox.univie.ac.at/papers.econ

Who gives Direction to Statistical Testing? Best Practice meets Mathematically Correct Tests

Karl H. Schlag*

October 8, 2015

Abstract

We are interested in statistical tests that are able to uncover that one method is better than another one. The Wilcoxon-Mann-Whitney rank-sum and the Wilcoxon sign-rank test are the most popular tests for showing that two methods are different. Yet all of the 32 papers in Economics we surveyed misused them to claim evidence that one method is better, without making any additional assumptions. We present eight nonparametric tests that can correctly identify which method is better in terms of a stochastic inequality, median difference and difference in medians or means, without adding any assumptions. We show that they perform very well in the data sets from the surveyed papers. The two tests for comparing medians are novel, constructed in the spirit of Mood's test.

Key words: Nonparametrics, exact, Wilcoxon signed-rank test, Wilcoxon-Mann-Whitney rank-sum test, stochastic inequality, mean, median

JEL codes: C12, C14, C90

1 Introduction

We use statistical tests to uncover findings from data, to see whether what we observe is only random or whether instead it can be attributed to properties of the underlying processes. In this paper we look at methods for comparing two distributions. We may be testing a new voting method or a new drug, or equivalently we may be comparing alternative methods, treatments or populations. Typically we are not primarily interested in showing that a new method is different than the existing one.

*Department of Economics, University of Vienna, karl.schlag@univie.ac.at

We are interested in showing that it is better. This is what we mean by identifying direction, showing that it is better and not only different. Clearly it is easier to establish evidence that the new method is different as one does not have to answer how it is different. To describe how it is different may be done in many ways, for instance by referring to means, medians or variances. In this paper we consider statistical methods for identifying direction when more than two different outcomes are possible.

We observe five different statistical practices for identifying direction. (i) Use a test for establishing evidence of a difference as if it were a test for establishing direction. (ii) Establish statistical evidence of a difference and then look at the data to determine where this difference comes from. (iii) Assume that the sample is sufficiently large and use a statistical test based on an infinitely large sample for identifying direction. (iv) Make additional assumptions on the underlying distributions that cannot be tested and use these to identify direction. (v) Make no additional assumptions and use a test that is designed for the given sample size and that identifies significant evidence for direction correctly.

Clearly the first approach is invalid. Approaches (ii)-(iv) are not reliable in the following sense. There are environments in which for the majority of the data sets the statistician comes to the conclusion that the new method is better even though this is not true. Ad (ii), simply looking at the data can lead to false conclusions, that is why we have statistical tests. Ad (iii), we explain why there is typically no way to determine whether the sample is sufficiently large. Ad (iv), as the assumptions cannot be tested one can easily be making assumptions that are violated in the true data. Only approach (v) is reliable. We answer the title, who gives direction? In (i)-(iv) it is the statistician, she determines what she will find. Only in (v) it is the data itself that reveals direction and allows us to gather statistical evidence for whether the new method is really better. Only this approach is guaranteed to deliver reliable results.

Which approaches are followed in the literature? In this article we focus on the most popular Wilcoxon-Mann-Whitney rank-sum (1945, 1947) and the Wilcoxon signed-rank (1945) test.[1][2] We survey all papers published in a list of leading economics journals since 2014 that use these two tests. We choose the economics journals to get a thorough overview, because this is the area of expertise of the author

[1] It is common to refer to the Wilcoxon-Mann-Whitney rank-sum test instead of to the Mann-Whitney rank-sum test to give credit to Wilcoxon (1945) whose paper also contains a test for two independent balanced samples.

[2] We comment on the t test when discussing mean comparisons.

and because of the special situation in economics. Siegel and Castellan (1988) was an early influential book that contains mistakes that lead the reader to follow (i). Forsythe etal (1994) then show that the Wilcoxon-Mann-Whitney test is not reliable, a paper that is now well cited. Yet none of the 32 papers we surveyed only uses these two tests to establish evidence of a difference. They all follow either (i) or (ii). The two sample t test fits under (iii) or (iv). Either its validity is based on asymptotic theory (fitting into (iii)) or variables are assumed to be normally distributed (fitting into (iv)). Methods belonging to (v) are few and not well known, thus have only been used in a handful of papers. Romano and Wolf (2000) designed a test for the mean of a single sample which can be used to compare means based on matched pairs. The regression method of Dufour and Hallin (1993) can be used to compare data of independent samples if there are no point masses. Unfortunately software is not available online for either method. Schlag (2008) presents tests that are explicitly designed for establishing direction when comparing two samples, the tests are simple and software is available online. In this article we present two further simple tests, these involve comparing medians, again software is available.[3] All of these tests are then applied to a few data sets from each of the surveyed papers (whenever data and sufficient information was available).

We review why the Wilcoxon-Mann-Whitney and the Wilcoxon test are appropriate for testing identity of two distributions but not for testing equality of means or medians. New insights are provided on how severely unreliable these tests can be for uncovering direction. The limitations of the t test are also explained. We then present eight tests that can identify how distributions differ, by looking at an ordinal difference, median difference and differences in medians and means. In particular, two novel tests that show how to correctly compare medians are introduced.

The format of this article is aimed at the broad audience as these are the users of the tests and this is where we hope to change best practice. In particular the aim is to show how simple it is to construct and understand some mathematically correct tests. Once practitioners develop an understanding of reliable testing our hope is that the theorists will design more such tests.

[3]Software in R for the tests in Schlag (2008) and the two new median tests introduced in this article are available at http://homepage.univie.ac.at/karl.schlag/.

2 Statistical Testing

We consider statistical tests for comparing methods when there are more than two possible outcomes. Tests for direction, specifically for comparing means, are well known when only two outcomes are possible, such as the z test (Suissa and Shuster, 1985, 1991). The term *nonparametrics* is used when there are infinitely many possible outcomes and the set of underlying distributions is infinitely dimensional. In practice the set of possible outcomes are large but finite. Many still then talk informally about nonparametrics as typically the set of underlying distributions will never-the-less be too rich for simulations.

Statististical testing is about determining from data whether or not one has evidence in favor of a statement about the process that has generated the data. The statistical test describes, as a function of the data, when the statement can be made. α is called the *level* of the test if the statement made is true with probability at least $1-\alpha$, where the level is calculated before the data is gathered. The smallest possible level of a test is called its *size*. If the level of the test is below 0.05 then one speaks of *significant evidence*. The *p value* is calculated after the data is gathered and is the smallest possible level of the test under which the statement can be made for this data.

To claim that a given test has level α requires a proof. One needs to prove that the probability of claiming to have evidence is at most α when the statement is false. Simulations will not generate a correct, i.e., mathematically correct, statement in a nonparametric framework as given the richness of the set of possible distributions one cannot simulate them all.

In the statistical terminology, the null hypothesis identifies all situations in which the statement is false. To reject the null hypothesis means to make the desired statement. In a literature used to making claims that are only correct if the sample size is sufficiently large, typically without being able to present on a formal bound for what "sufficiently large" means, the term '*exact*' (Yates, 1934) is added to any test that produces results that are mathematically correct. Only very few tests for comparing means or medians receive the attribute 'exact'.[4]

[4]The t test for a single mean and normally distributed random variables, the two sample t test for two means, independent samples and normally distributed random variables with equal variance, and many tests (including Fisher's (1935) exact test and the z test) for comparing means of two binary valued distributions are exact. An exact test for the median of a single random variable is also easily constructed using the binomial test.

3 The Wilcoxon-Mann-Whitney and Wilcoxon Tests

The Wilcoxon-Mann-Whitney and Wilcoxon tests can be used to uncover significant evidence that two underlying distributions are not equal (i.e. not identical), based on a sample of data from each distribution. They are good for establishing significant evidence for the statement "The two distributions are different." The Wilcoxon-Mann-Whitney test applies when comparing two independent samples. The Wilcoxon test is designed for a single sample of matched pairs, a pair consisting of one observation for each treatment. Both tests are permutation tests, they use the fact that under the null hypothesis the probability of realizing any given outcome does not depend on the treatment. They however cannot be used to establish evidence for the statement "The two distributions are different because their two means are different" or for the analogous statement with medians. They reject the null hypothesis too often as they are designed to detect differences in distributions, even if means or medians are equal. We provide the evidence.

Assume that we wish to test if the two means are equal and find that all outcomes in the first sample are greater than those in the second. Then both tests will reject the null hypothesis (if they ever reject the null hypothesis) as for them this is the most extreme evidence in favor of unequal means. However, the means could still be equal if there is a small probability that the second outcome is very large. In fact, the larger it is the smaller the probability is needed to make the two means equal and hence the less likely this large outcome will be observed. So most of the time the two tests falsely reject the null hypothesis. In fact, the size of these tests for claiming that the two means are different is 1. This example is due to Lehmann and Loh (1990).[5]

The two tests are similarly not valid for testing equality of two medians. For matched pairs we obtain that the Wilcoxon test has size 1, using the same intution.[6] Similarly the Wilcoxon-Mann-Whitney test case is oversized for testing equality of medians given independent samples (see Figure 1).[7] For earlier less dramatic examples

[5] Formally, let $\varepsilon > 0$ and $X_1 = 1$ almost surely while $X_2 \in \{1-\varepsilon, 2\}$ such that $EX_2 = 1$. So as ε tends to 0, $P(X_2 = 2)$ tends to 0 and hence for large probability $x_{1i} > x_{2i}$ holds for all $i = 1, .., n$.

[6] Let $\varepsilon > 0$ and assume that $P(X = (0,0)) = \varepsilon$ and $P(X = (-1,0)) = P(X = (0,1)) = (1-\varepsilon)/2$. Then X_1 and X_2 have median 0. If ε is small then with high probability $x_{2i} > x_{1i}$ holds for all $i = 1, .., n$.

[7] The following example is used to derive lower bounds in balanced samples. Let $\varepsilon = 0.0001$. Let X_1 be uniformly distributed on $[0,1]$ while with probability $1/2 + \varepsilon$, let X_2 be uniformly distributed on $\left[1/2 - \varepsilon, 0.5 + 2\varepsilon^2\right]$ and otherwise be uniformly distributed on $[1,2]$. So both r.v.s have median $1/2$.

see Forsythe etal (1994) and Chung and Romano (2013).[8]

The Wilcoxon-Mann-Whitney and the Wilcoxon test are not even approximately valid for establishing evidence in favor of stating "The two means or the two medians are different".

It is of course true that both Wilcoxon-Mann-Whitney and Wilcoxon tests are trivially able to identify differences in means and medians if one assumes that the two distributions are identical whenever they have the same mean or median. Given this assumption, whenever one uncovers that the distributions are not identical, one can immediately claim that the means and the medians are not equal. Such drastic assumptions are made in the location shift model. It is up to the specific application whether it is plausible that these extreme restrictions are viable.[9] One cannot test if these restrictions are true in the data as the property of being a location shift model is degenerate. Arbitrarily small changes in the probabilities, impossible to detect in a finite sample, will ruin the property. We lack mathematical results on how significance changes when the property holds only approximately.

So the Wilcoxon-Mann-Whitney and Wilcoxon tests should be used to understand if there is any difference in the treatments. But they should not be used for more, in particular they cannot identify whether either means or medians differ between the two treatments.[10]

How could best practice focus on making mathematically incorrect claims? Early influential books (e.g. Siegel and Castellan, Example 5.3b and Section 6.7) suggest this malpractice. Alternative statistical tests have only been recently available. The largest obstacle that interferes with the use of correct methods for comparing means or medians is the fact that these will typically be less powerful than either the Wilcoxon-Mann-Whitney or the Wilcoxon test. This is because their null hypothesis is much larger, and hence it is much more difficult to establish sufficient evidence against it.

Note that the Wilcoxon test has also been used to compare a single distribution to a single value. This is a valid procedure if one wishes to uncover that the distribution is not symmetrically distributed around this value. However, typically there is no empirical interest in uncovering asymmetry. In fact, in most applications this sym-

[8] Forsythe etal (1994, Table V, distributions 2-5) find rejection probabilities up to 0.12, Chung and Romano (2013, Table 1) up to 0.23.

[9] The author knows of no context in which this restriction arises naturally.

[10] There are one sided versions of these two tests, which can be used to claim that the one distribution does not first order stochastically dominate the other. However, hardly ever do these really fit the application, they are mainly used because of their apparent power, their p value is half of the one of the two sided test.

metry assumption is not even mentioned. Note that statistical statements about the median can be derived using the binomial test, the uniformly most powerful test for this setting. A simple and powerful test for investigating the mean of the underlying distribution has been constructed by Schlag (2008).

4 Tests for Comparing Means

The two sample t test (Student, 1908) cannot be used to compare means without imposing drastic assumptions. The two sample t test is a parametric test, it is exact when the two underlying distributions are normally distributed and have equal variances. One can apply the two sample t test if one explicitly assumes that the two distributions are normally distributed with equal variances. Note that the property underlying this assumption is knife-edge. Arbitrarily small changes in some probabilities can destroy this property. Thus one cannot claim that this property follows from previous investigations. Moreover, it is not possible to test this property. In fact, data cannot be normally distributed if all possible outcomes are bounded or discrete. Tests can only reveal that the data is approximately normally distributed but we do not know how to adjust the size of the t test for this case. In particular, it is not good practice (but see FDA, 1996 VI.A and FDA 2007, IV.B.4.a.iii.) to assume normality if no evidence against normality can be gathered. Using the two-sample t test is like testing the claim "Either the two means are different or the data is not normally distributed with equal variances".

Typically one is interested in establishing evidence for the claim "The two means are different". Indeed the t test is approximately valid if the there are sufficiently many observations in each sample. However, the number of observations needed depends on the underlying distributions which are unknown. The paradox is that, for any given number of observations, provided there are at least 3 observations in each sample, the two sample t test and Welch's t test (1947) are arbitrarily bad for establishing this claim. Their size is 1. This result follows from the arguments of Lehmann and Loh (1990) for the single sample t test, we used the corresponding example above (see Footnote 5). To use the two sample t test is as if one is testing the claim "Either the two means are different or not enough data has been gathered." Similarly the permuation test proposed by Chung and Romano (2013) is asymptotically valid but by the same arguments, in any finite sample, either it has size 1 or it never rejects the null hypothesis.

To establish significant evidence for the statement "The two means are different"

one needs some additional information about the underlying process. Otherwise, following Bahadur and Savage (1956), non trivial tests do not exist. This is because of fat tails. The values of very large and rare outcomes can influence whether or not the two means are equal. But if these outcomes are so rare, they will most likely not be observed in the given finite sample, hence there is no way to tell whether or not the two means are different.

In an earlier paper (Schlag, 2008) we have constructed valid tests for the case where the statistician knows some bounds on any possible outcome. Such bounds emerge naturally when outcomes are measured on a bounded scale. Their construction is simple. First the data is randomly transformed into binary valued data by a mean preserving transformation. Then significant evidence of differences is investigated using an exact test for binary valued data for the relevant sampling context. In a final step the randomized element that was inserted by the random transformation is eliminated.

5 Old and New Tests for Comparing Medians

Mood's test (Westenberg, 1948, Mood, 1950) is a candidate for comparing medians of independent continuous distributions.[11] In a first step one estimates the median of the combined sample. In the second step one investigates whether there is evidence that this is not the common median. To do this, count in each sample how many observations are above the estimated median. Then test whether these two proportions are different, for instance it has been suggested to use Fisher's exact test (1935). Mood's test relies on having a good estimate of the common sample median. However, even for large sample sizes, the estimated median need not be close to the true median. Hence, sometimes Mood's test is in fact comparing quantiles and these need not be equal even when the medians are, which can lead to overrejection.

The robust rank order test of Fligner and Policello (1981) is another test for comparing medians, this test is drastically oversized. We show lower bounds in Figure 1 in the appendix.[12] More recently, Chung and Romano (2013) suggest a test for equality of medians that is asymptotically valid, however their Table 1 shows that it is often oversized in finite samples, for instance its rejection probability is 0.09 when

[11] The median of a random variable X is any value x that satisfies $P(X \geq x) \geq \frac{1}{2}$ and $P(X \leq x) \geq \frac{1}{2}$. We let $med(X)$ denote the set of values with this property and write $med(X) > 0$ or $med(X) = 0$ if all elements in the set have this property.

[12] In both cases we used the example from Footnote 7.

$\alpha = 0.05$, $n_1 = 51$ and $n_2 = 101$.

In this article we demonstrate how easy it is to adjust Mood's test to obtain a valid and powerful test. We construct a confidence set for the pair of medians. We rule out a candidate pair of medians if too many or too few observations are above these values. We also rule it out if, as in Mood's test, there are sufficient differences between the two samples. If by this procedure all potential values are ruled out, one rejects that the medians are equal. The test is invariant to a common monotone transformation of all outcomes.

As this is the first place where this test appears in the literature we provide a few more details how to construct the test for $H_0 : med\,(X_1) \leq med\,(X_2)$ (an extended description of the test can be found in the online material to this article). The construction is based on combining two separate tests, one for the extremes and one for the intermediate values. First it needs to be decided how much of the overall level α one assigns to each of these parts. Following power analyses we suggest to allocate $0.1 \cdot \alpha$ to the first test and the remaining to the second. For each potential value m of the median of X_2 one proceeds roughly as follows. The value m is ruled out if, using the binomial test at level $\alpha/10$, the total number of outcomes above these values in each sample is significantly different from 1/2. It is also ruled out if there are significantly more observations of X_1 above this value than there are of X_2 above this value, evaluated using a test for comparing proportions at level $9\alpha/10$. A natural candidate test for this comparison is the z test (Suissa and Shuster, 1985), it is more powerful than Fisher's exact test in balanced samples. In fact, power analyses show that our new test is even more powerful if one replaces the exponent 1/2 in the denominator of the z test statistic by 3/2. The details of the test as described in the supplementary material are chosen so that the above test remains valid when the distributions have point masses.

The same approach can be used to construct tests for comparing medians based on matched pairs, using the z test (Suissa and Shuster, 1991) and allocating 1/10 of the level to choosing the set of potential medians.[13]

One can use the above to derive a confidence interval for the difference between the two medians. This is particularly useful to understand what is going on when the null hypothesis of equality of the two medians cannot be rejected. Is there really evidence that two medians are close or is the sample just too small?

[13] Unlike in the case of independent samples we here do not suggest to adjust the exponent in the denominator of the z statistic.

6 Tests for Stochastic Equality and Median Difference

Median comparison tests are natural ordinal tests when the median of each distribution is of central of interest. However, if the main interest is to compare two samples, it makes sense to test for an ordinal property that pertains to the comparison. The aim is to generate more rejections as we are directly comparing the two distributions.

One idea is to compare two random observations, one from each distribution, and compare the likelihood that the one is larger than the other, ignoring matches where they are equal. This is the idea behind stochastic inequality. We say that a random outcome X_1 tends to be higher than a random outcome X_2 if X_1 is more likely to generate a higher outcome than X_2 than vice versa. In other words, if in the majority of cases in which the two outcomes are different we expect that X_1 is higher than X_2. Formally, we test the null hypothesis $H_0 : P(X_1 > X_2) \leq P(X_1 < X_2)$. To reject this means to have evidence for the statement "X_1 tends to be larger than X_2". The two-sided test for $P(X_1 > X_2) = P(X_1 < X_2)$ is called a test of *stochastic equality* (Brunner and Munzel, 2000).

It is easy to construct an exact test of stochastic equality when there are matched pairs. Use the binomial test to test if there are significantly more observations (x_1, x_2) with $x_1 > x_2$ than there are with $x_1 < x_2$ (this is also called the sign test).

Consider now the case of independent samples. The test of Brunner and Munzel (2000) is designed using asymptotic theory and assumes that neither X_1 nor X_2 puts all mass on a single point. Simple examples show that it is moderately oversized in many samples (see also Medina etal, 2010). The Wilcoxon-Mann-Whitney test is also oversized for testing this null hypothesis. Lower bounds on the rejection probabilities of these two tests are shown in Figure 2 in the appendix.[14]

An exact test for independent samples is developed in (Schlag, 2008). Its construction relies on the following three steps. First the data is matched in pairs and it is recorded which observation is larger or smaller (matches with equal observations are dropped). Then the binomial test is applied to identify if there is a significant difference between these two events. Finally, the random component introduced by the data matching is eliminated. Confidence intervals for the so-called *stochastic difference* $P(X_1 > X_2) - P(X_1 < X_2)$ help understand the degree of the difference in

[14] In both cases we use the following example with balanced samples. Let X be uniformly distributed on $[0.9, 1.1]$ and Y be equally likely equal to 0 and 2.

tendency.

An alternative idea is to look at the median difference between two random observations, one from each distribution. This is in analogy to the means test where we are investigating the mean difference between two random observations, one from each distribution. We establish evidence for the statement "The median difference between X_1 and X_2 is strictly positive" when we are able to reject $H_0 : med(X_1 - X_2) \leq 0$. We would thus establish evidence that X_1 tends to be larger than X_2 in the sense of the stochastic inequality. The converse is true if $P(X_1 = X_2) = 0$. This is because "$med(X_1 - X_2) > 0$" $\iff$ "$P(X_1 - X_2 \leq 0) < \frac{1}{2}$" $\iff$ "$P(X_1 > X_2) > P(X_1 \leq X_2)$" $\implies$ "$P(X_1 > X_2) > P(X_1 < X_2)$". Thus it is weakly harder to establish evidence against median difference equal to 0 than against stochastic equality. A test for $H_0 : med(X_1 - X_2) \leq 0$ can be constructed like the test of stochastic inequality, only now recording whether $X_1 \geq X_2$ in the matched pair and not dropping matches with equal observations.

A test for $H_0 : med(X_1 - X_2) = d$ can be constructed for each $d \in \mathbb{R}$, using the fact that $med(X_1 - X_2) - d = med(X_1 - (X_2 + d))$. In particular, if $x_{1i} - x_{2j} \neq d$ for all i and j then the test of $H_0 : med(X_1 - X_2) = d$ is equivalent to the test of $H_0 : P(X_1 > X_2 + d) = P(X_1 < X_2 + d)$. Confidence intervals for the median difference help understand how X_1 and X_2 differ.

Note that the test for stochastic equality and the one for median differences are invariant to a common monotone transformation of all outcomes.

7 Data Examples

To establish current practice and to show how the tests introduced above perform we collected all papers in which the Wilcoxon-Mann-Whitney (WMW) or the Wilcoxon (W) test has been used and that have been recently published in top economics journals (from January 2014 to June 2015). We found a total of 32 papers using either the WMW or W test.

Our first step is to investigate current practice. The result is very disappointing. In none of these papers the WMW and W tests are only used for what they are designed for. In (Masella etal, 2014) the data is not independent, in (Cohn etal, 2015) the test is used only to show high p values. In all other cases they are used to reveal insights about direction. Note that the location shift model is not mentioned in any of the articles.

Our second step is to use the data in these papers to show how the tests for

direction presented above perform. We were able to obtain data and the necessary information for 22 of these papers.[15] For each of them we select a few salient results derived using either W or WMW test, where the observations within the sample are independent by construction. We choose data sets in which authors establish significant differences, whenever possible we choose ones with strongly significant treatment effects and ones with only significant treatment effects. This yields 42 comparisons with corresponding data sets, 18 with matched pairs and 23 with independent samples.

For these 42 cases we first run the WMW and W tests. Often our results are different from the ones in the paper as we use the exact versions and we do not find one-sided tests to be appropriate. We then investigate the four nonparametric tests with direction, comparing medians, means and investigating stochastic inequality.

We summarize our findings. In most cases the WMW and W tests generate much smaller p values than any of the other four tests. However in 2 cases the test for comparing medians and in 6 cases the test for stochastic equality yielded a smaller or equal p value. Generally, and as expected, significant evidence for one of the four tests with direction can only be reached if there is very strong evidence of non identical distributions.

In 20 (28) out of the 42 data sets at least one significant (marginally significant) result is obtained with one of the four directional tests. As anticipated, the test for stochastic equality performs best. In fact, when running the one-sided tests whenever appropriate, it establishes at least a marginally significant result in 34 out of the 42 tests. A detailed account of how many significant and marginally significant results each test produced is shown in Table 1.

[15] The papers that are not part of this subset of 22 papers and that are not mentioned above are (Becchetti etal, 2014, Brocas etal, 2014, Cabral etal, 2014, Filiz-Ozbay etal, 2015, Ockenfels and Selten, 2015), the data sets of the following papers arrived too late to be considered in this first draft (Kamijo etal, 2014, Khalmetski etal, 2015, Regner, 2015).

p value	≤ 0.05	$\in (0.05, 0.1]$	> 0.1
WMW or W	33	6	3
stochastic equality	18 (20)	10 (14)	14 (8)
median difference	10 (11)	4 (8)	28 (23)
equality of medians	10 (11)	3 (4)	29 (27)
equality of means	6 (6)	1 (2)	35 (34)
at least one test with direction	20 (22)	8 (12)	14 (8)

Table 1: Significance levels for our 42 data sets for the various tests. Numbers in brackets refer to results obtained by running the one-sided test whenever we found that the paper indicates this as being the relevant one.

We add a few more observations (see table with all results in online material). In the extremely small data sets, identified by having between 5 and 7 observations in each sample, the test for stochastic equality identified a significant finding in 3 out of the 6 cases. Note that our novel test for testing equality of medians performs quite well in data sets in which the minimal sample size is at least 15. In 9 out of 22 cases it establishes a significant result. The equality of means test requires larger samples, the 6 cases with significance were obtained among the 17 cases with at least 25 observations in each sample.

For a subset of the data sets we also derive confidence sets for the difference in medians, for the median difference and for the stochastic difference. These nicely complement our basic tests. In 4 out of the 17 data sets the confidence set for the difference in medians contains a single point. This means that one can significantly identify the difference in medians. One of these nicely highlights the difference between tests for identity of distributions and for identity of medians. In this data set with 36 matched pairs there is significant evidence that the two distributions are different yet there is also significant evidence that both have the same median.

For the same set of journals and same time period we found 7 papers that use the Wilcoxon test to compare a single variable to specific value or sequence of values. Symmetry is mentioned only twice but not motivated. We obtained data sets and the necessary information to replicate the results for 6 of these papers. We choose 9 salient data sets from them and find that both the median and the mean analysis produce significant results in 7 cases, in particular without assuming symmetry.

8 Conclusion

The Wilcoxon-Mann-Whitney and Wilcoxon tests are the most prominent tests for identifying that two treatments generate different outcome distributions. These two tests are extremely useful for getting a first understanding when comparing treatments. In addition, the WMW test stands out due to its unique properties.[16] Conventions are important in statistical testing.[17] Hence, alternative tests such as the Kolmogorov-Smirnov test (Kolmogorov, 1933, Smirnov, 1939) should only be used if they can be justified based on the design of the treatments, irrespective of the specific data gathered. However, neither the WMW nor the W test should be used to make claims about which method is better. Nor should any other permutation or different test for identity of distributions be used for this cause.[18]

We find no justification for using the two sample t test and its robust counterpart (Welch, 1947) in finite samples. It is known, yet generally overlooked (see the low number of citations), that tests based on asymptotic theory can be very inappropriate in finite samples (eg see Lehmann and Loh, 1990, Dufour, 2003, Romano, 2004, Medina, 2010).

In this article, we present eight correct and valid tests and show how useful they are for understanding real data. We find that it is useful to explain evidence in terms of ordinal differences as one can then apply the test for stochastic inequality. If one wishes to identify differences in terms of means then the following has to be taken into account. It is not possible to claim significant evidence of a difference in means if the underyling random variables have no natural bounds. The samples should not be too small as in most experimental investigations if one wishes to compare means (Note that we only obtained signficant results in 6 of our data sets, see Table 1).

Statistical testing is a mathematical field that naturally requires to make correct claims. Of course it will not be possible to construct exact tests for each statistical model of the underlying data generating process. Asymptotic theory remains to be helpful to guide the design of tests. However, we think that it is only fair to report on extensive simulations whenever using a test whose finite sample properties are not mathematically founded. Oversizedness in these simulations should lead to an

[16] The WMW test is uniformly most powerful among all unbiased tests that are invariant to monotone transformations of the data (Lehmann and Romano, 2005).

[17] Otherwise the user may be tempted to try several different tests and only report the results of the one that is most favorable.

[18] Note that the test used in Cohen and Dupas (2010) is a permutation test that is misused to identify differences in means.

upwards adjustment in the reported p values.

Statistical analysis can have a drastic impact on life, such as when used to design development policies for the third world or to analyze the effectiveness of new drugs. As such we think that we owe it to us to do this analysis mathematically correct.

References

[1] Asparouhova, Elena, Peter Bossaerts, Jon Eguia, and William Zame, "Asset Pricing and Asymmetric Reasoning," *Journal of Political Economy*, 123 (2015), 66-122.

[2] Bahadur, R. R., and Leonard J. Savage, "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems," *The Annals of Mathematical Statistics*, 27 (1956), 1115–1122.

[3] Becchetti, Leonardo, Maurizio Fiaschetti, and Giancarlo Marini, "Card Games and Economic Behavior," *Games and Economic Behavior,* 88 (2014), 112–129.

[4] Bhattacharya, Sourav John Duffy, and Sun-Tak Kim, "Compulsory Versus Voluntary Voting: An Experimental Study," *Games and Economic Behavior,* 84 (2014), 111–131.

[5] Brocas, Isabelle, Juan D. Carillo, Stephie W. Wang, and Colin F. Camerer, "Imperfect Choice or Imperfect Attention? Understanding Strategic Thinking in Private Information Games,'" *Review of Economic Studies* (2014), 81, 944–970.

[6] Brunner, Edgar and Ullrich Munzel (2000), "The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation," *Biometrical Journal*, 42, 17-25.

[7] Cabral, Luis, Erkut Y. Ozbay, and Andrew Schotter, "Intrinsic and Instrumental Reciprocity: An Experimental Study," *Games and Economic Behavior*, 87 (2014), 100–121.

[8] Cason, Timothy N., Daniel Friedman, and Ed Hopkins, "Cycles and Instability in a Rock–Paper–Scissors Population Game: A Continuous Time Experiment," *Review of Economic Studies* (2014), 81, 112–136.

[9] Cason, Timothy N., and Charles R. Plott, "Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing," *Journal of Political Economy*, 122 (2014), 1235-1270.

[10] Charness, Gary, Ramón Cobo-Reyes, and Natalia Jiménez, "Identities, Selection, and Contributions in a Public-Goods Game," *Games and Economic Behavior*, 87 (2014), 322–338.

[11] Chen, Yan, Sherry Xin Li, Tracy Xiao Liu, and Margaret Shih, "Which Hat to Wear? Impact of Natural Identities on Coordination and Cooperation," *Games and Economic Behavior,* 84 (2014), 58–86.

[12] Cheung, Stephen L., "Comment on "Risk Preferences Are Not Time Preferences": On the Elicitation of Time Preference under Conditions of Risk," *American Economic Review*, 105 (2015), 2242–2260.

[13] Chung, EunYi, Romano, Joseph P., "Exact and Asymptotically Robust Permutation Tests," *The Annals of Statistics*, 41 (2015), 484–507.

[14] Cohen, Jessica, and Pascaline Dupas, "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment," *The Quarterly Journal of Economics*, 75 (2010), 1-45.

[15] Cohn, Alain, Jan Engelmann, Ernst Fehr, and Michel André Maréchal, "Evidence for Countercyclical Risk Aversion: An Experiment with Financial Professionals," *American Economic Review*, 105 (2015), 860–885.

[16] Dittmann, Ingolf, Dorothea Kübler, Ernst Maug, and Lydia Mechtenberg, "Why Votes have Value: Instrumental Voting with Overconfidence and Overestimation of Others' Errors," *Games and Economic Behavior*, 84 (2014), 17–38.

[17] Duffy, John, and Daniela Puzzello, "Gift Exchange versus Monetary Exchange: Theory and Evidence," *American Economic Review*, 104 (2014), 1735–1776.

[18] Dufour, Jean-Marie, "Identification, Weak Instruments, and Statistical Inference in Econometrics," *The Canadian Journal of Economics / Revue canadienne d'Economique*, 36 (2003), 767-808.

[19] Dufour, Jean-Marie, and Marc Hallin, "Improved Eaton Bounds for Linear Combinations of Bounded Random Variables, with Statistical Applications," *Journal of the American Statistical Association*, 88 (1993), 1026–1033.

[20] Eckel, Catherine C., and Sascha C. Füllbrunn, "Thar SHE Blows? Gender, Competition, and Bubbles in Experimental Asset Markets," *American Economic Review*, 105 (2015), 906–920.

[21] FDA, *Regulatory Information, Statistical Guidance for Clinical Trials of Non Diagnostic Medical Devices*, The Division of Biostatistics, U.S. Food and Drug Administration, 1996.

[22] FDA, Redbook 2000, Guidance for Industry and Other Stakeholders Toxicological Principles for the Safety Assessment of Food Ingredients, U.S. Department of Health and Human Services, Food and Drug Administration, Center for Food Safety and Applied Nutrition, 2007.

[23] Feltovich, Nick, "Critical values for the robust rank-order test", *Communications in Statistics - Simulation and Computation*, 34 (2005), 525-547.

[24] Filiz-Ozbay, Emel, Kristian Lopez-Vargas, and Erkut Y.Ozbay, "Multi-Object Auctions with Resale: Theory and Experiment," *Games and Economic Behavior,* 89 (2015) 1–16.

[25] Fisher, R. A. (1935), "The Logic of Inductive Inference," *J. Roy. Stat. Soc.* **98**, 39–54.

[26] Fligner, Michael A., and George E. Policello II, "Robust Rank Procedures for the Behrens-Fisher Problem," Journal of the American Statistical Association, 76 (1981), 162-168.

[27] Forsythe, Robert, Joel L. Horowitz, N. E. Savin, and Martin Sefton, "Fairness in Simple Bargaining Experiments," *Games and Economic Behavior,* 6 (1994), 347–369.

[28] Friedman, Daniel, Steffen Huck, Ryan Oprea, and Simon Weidenholzer, "From Imitation to Collusion: Long-run Learning in a Low-Information Environment," *Journal of EconomicTheory*, 155 (2015), 185–205.

[29] Hugh-Jones, David, Morimitsu Kurino, and Christoph Vanberg, "An Experimental Study on the Incentives of the Probabilistic Serial Mechanism," *Games and Economic Behavior,* 87 (2014), 367–380.

[30] Isoni, Andrea, Anders Poulsen, Robert Sugden, and Kei Tsutsui, "Efficiency, Equality, and Labeling: An Experimental Investigation of Focal Points in Explicit Bargaining," *American Economic Review*, 104 (2014), 3256–3287.

[31] Kamijo, Y., T. Nihonsugi, A. Takeuchi, and Y. Funaki, "Sustaining Cooperation in Social Dilemmas: Comparison of Centralized Punishment Institutions," *Games and Economic Behavior*, 84 (2014), 180–195.

[32] Khalmetski, Kiryl, Axel Ockenfels, and Peter Werner, "Surprising Gifts:Theory and Laboratory Evidence," *Journal of Economic Theory,* 159 (2015), 163–208.

[33] Kolmogorov A., "Sulla Determinazione Empirica di una Legge di Distribuzione," *Giornale dell'Istituto Italiano degli Attuari*, 4 (1933), 83–91.

[34] Kosfeld, Michael and Devesh Rustagi, "Leader Punishment and Cooperation in Groups: Experimental Field Evidence from Commons Management in Ethiopia," *American Economic Review,* 105 (2015), 747–783.

[35] Lai, Ernest K., Wooyoung Lim, and Joseph Tao-yi Wang, "An Experimental Analysis of Multidimensional Cheap Talk," *Games and Economic Behavior*, 91 (2015), 114–144.

[36] Lehmann, E. L., and Wei-Yin Loh (1990), "Pointwise verses Uniform Robustness in some Large-Sample Tests and Confidence Intervals," *Scandinavian Journal of Statistics,* 17, 177–187.

[37] Lehmann, Erich L., and Joseph P. Romano, *Testing Statistical Hypotheses*, New York: Springer, 2005.

[38] Mann, H. B., and D. R. Whitney, "On a Test Whether One of Two Random Variables is Stochastically Larger Than the Other," *Annals of Mathematical Statistics*, 18 (1947), 50–60.

[39] Markussen, Thomas, Louis Putterman, and Jean-Robert Tyran, "Self-Organization for Collective Action: An Experimental Study of Voting on Sanction Regimes," *Review of Economic Studies*, 81 (2014) , 301–324.

[40] Masella, Paolo, Stephan Meier, and Philipp Zahn, "Incentives and Group Identity," *Games and Economic Behavior*, 86 (2014), 12–25.

[41] McNemar, Q., "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages," *Psychometrika,* 12 (1947), 153–157.

[42] Medina, Jared, Daniel Y. Kimberg, Anjan Chatterjee, and H. Branch Coslett, "Inappropriate Usage of the Brunner-Munzel Test in Recent Voxelbased Lesion-Symptom Mapping Studies," *Neuropsychologia,* 48 (2010), 341–343.

[43] Miao, Bin and Songfa Zhong, "Comment on "Risk Preferences Are Not Time Preferences": Separating Risk and Time Preference," *American Economic Review,* 105 (2015), 2272–2286.

[44] Mood, A.M., *Introduction to the Theory of Statistics*, New York: McGraw-Hill, 1950.

[45] Noussair, Charles N., Stefan T. Trautmann, and Gijs van de Kuilen, "Higher Order Risk Attitudes, Demographics, and Financial Decisions," *Review of Economic Studies,* 81 (2014), 325–355.

[46] Ockenfels, Axel, and Reinhard Selten, "Impulse Balance in the News Vendor Game," *Games and Economic Behavior,* 86 (2014), 237–247.

[47] Oprea, Ryan, "Survival Versus Profit Maximization in a Dynamic Stochastic Experiment," *Econometrica*, 82 (2014), 2225–2255.

[48] Romano, Joseph P., "On Non-Parametric Testing, the Uniform Behaviour of the t-Test, and Related Problems," *Scandinavian Journal of Statistics*, 31 (2004), 567-584.

[49] Petersen, Luba, and Abel Winn, "Does Money Illusion Matter?: Comment," *American Economic Review,* 104 (2014), 1047–1062.

[50] Pratt, J. W., "Remarks on Zeros and Ties in the Wilcoxon Signed Rank Procedures," Journal of the American Statistical Association 54 (1959), 655-667.

[51] Regner, Tobias, "Social Preferences? Google Answers!," *Games and Economic Behavior,* 85 (2014), 188–209.

[52] Romano, Joseph P., and Michael Wolf, "Finite Sample Non-Parametric Inference and Large Sample Efficiency," *Annals of Statistics,* 28 (2000), 756–778.

[53] Schlag, Karl H., "A New Method for Constructing Exact Tests without Making any Assumptions," Department of Economics and Business Working Paper, 1109 (2008), Universitat Pompeu Fabra.

[54] Schlag, Karl H., *Exact Hypothesis Testing without Assumptions - New and Old Results not only for Experimental Game Theory*, mimeo, http://homepage.univie.ac.at/karl.schlag/research/statistics/exacthypothesistesting.pdf, 2011.

[55] Shapiro, Dmitry, Xianwen Shi, and Artie Zillante, "Level-k Reasoning in a Generalized Beauty Contest," *Games and Economic Behavior*, 86 (2014), 308–329.

[56] Siegel, Sidney and N. John Castellan Jr., *Nonparametric Statistics for The Behavioral Sciences,* 2nd ed. (McGraw-Hill), 1988.

[57] Smirnov, N. V., "Estimate of Deviation Between Empirical Distribution Functions in Two Independent Samples," *Bulletin Moscow University* 2 (1939), 3–16.

[58] Spearman, C., "The Proof and Measurement of Association Between Two Things," *The American Journal of Psychology*, 15 (1904), 72–101.

[59] Student, "The Probable Error of a Mean," *Biometrika* 6 (1908), 1–25.

[60] Suissa, Samy, and Jonathan J. Shuster, "Exact Unconditional Sample Sizes for the 2×2 Binomial Trial," *Journal of the Royal Statistitical Society, Series A*, 148 (1985), 317–327.

[61] Suissa, Samy, and Jonathan J. Shuster, "The 2 x 2 Matched-Pairs Trial: Exact Unconditional Design and Analysis," *Biometrics,* 47 (1991), 361–372.

[62] Tocher, K. D., "Extension of the Neyman-Pearson Theory of Tests to Discontinuous Variates," *Biometrika*, 37 (1950), 130–144.

[63] Wantchekon, Leonard, Marko Klasnja, and Natalija Novta, "Education and Human Capital Externalities: Evidence from Colonial Benin," *The Quarterly Journal of Economics,* 130 (2015), 703–757.

[64] Welch, B. L., "The Generalization of 'Student's' Problem when Several Different Population Variances are Involved," *Biometrika*, 34 (1947), 28-35.

[65] Westenberg, J., "Significance Test for Median and Interquartile Range in Samples from Continuous Populations of Any Form," form," Proceedings of the Koninklijke Nederlandse Akademie van Wetenschapp, 51 (1948), 252-261.

[66] Wilcoxon, F., "Individual Comparisons by Ranking Methods," *Biometrics Bulletin,* 1 (1945), 80–83.

[67] Yates, F., "Contingency Tables Involving Small Numbers and the χ^2 Test," *Supplement to the Journal of the Royal Statistical Society*, 1 (1934), 217–235.

A Figures on Oversizedness

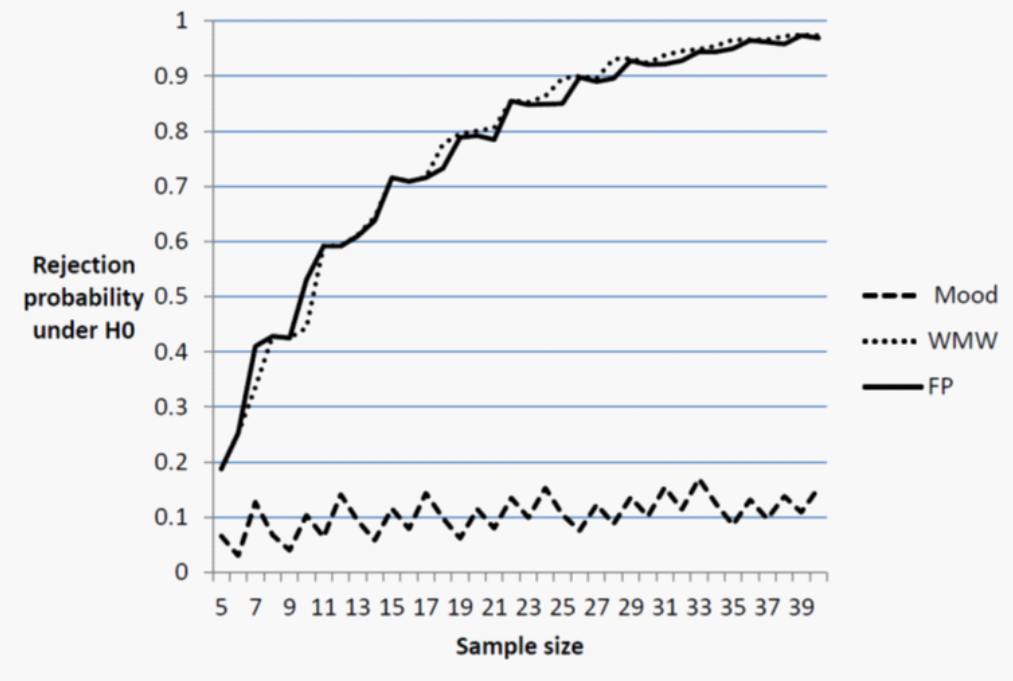

Figure 1: Lower bound on rejection probability under the null hypothesis of test for comparing medians with balanced independent samples and nominal size $\alpha = 0.05$.

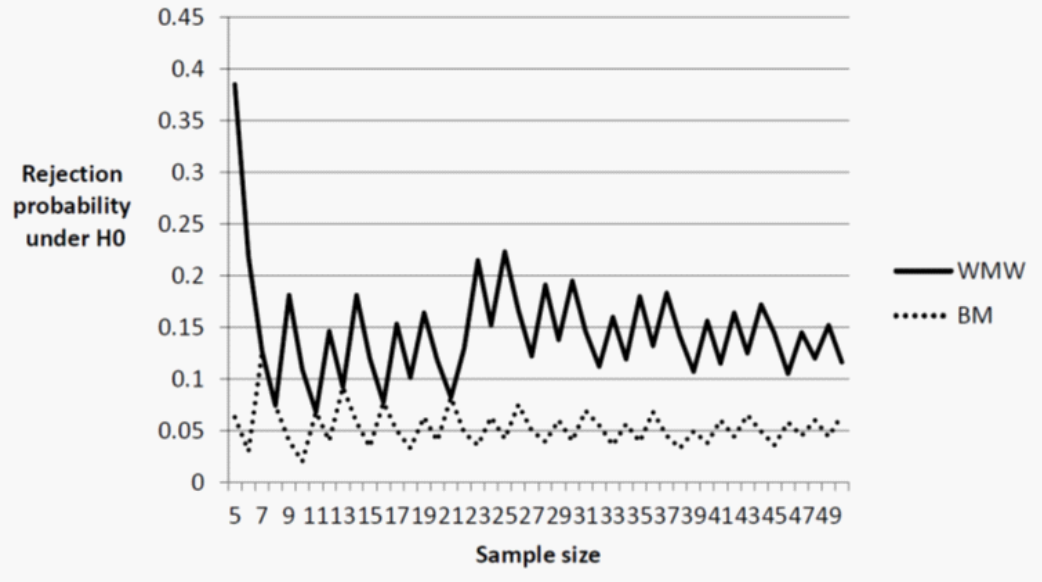

Figure 2: Lower bound on rejection probability under the null hypothesis of tests of stochastic inequality with balanced independent samples and nominal size $\alpha = 0.05$.

B Comparing Tests in Published Data Sets

We considered the following journals: American Economic Review (not Papers and Proceedings), Games and Economic Behavior, Econometrica, Journal of Economic Theory, Journal of Political Economy and Review of Economic Studies. We found 32

papers that either use the Wilcoxon-Mann-Whitney or the Wilcoxon test. We were able to obtain the data and the necessary information to replicate the findings for 22 of these papers, either from online sources or directly from the authors. For each of these papers we selected a few salient hypotheses involving comparing two variables in which the WMW and W tests can be formally applied, so where observations are independent. Whenever possible we selected data both where p values of these two tests were below 0.01 and where they were between 0.01 and 0.05.

We summarize our findings in a series of tables followed by a "key for abbreviations" table in which more information about the data source is given. In particular, we identify for each of the hypotheses the authors of the article, enough information to identify which two variables are being considered and where in the published paper the relevant statement for WMW or W can be found. In the column "x" we mention whether we found that the authors are interested in providing evidence of inequality, highlighted by "$\neq$" or instead where a directional hypothesis can be identified from the formulation of the problem as being of central interest.

For each hypothesis we reran the Wilcoxon-Mann-Whitney and Wilcoxon tests, using the exact versions. Then we ran the four tests for identifying direction as described in the main article. For each test, in the column "r", we highlight the direction of the evidence whenever it is significant at 5%, put it in brackets if only significant at 10%. While the cases where one-sided tests are of importance is highlighted in column "x", all p values in these two tables refer to the two-sided tests.

The equality of medians test is introduced for the first time in this article. It tests $H_0 : med(X_1) = med(X_2)$. In our tables we include an ad-hoc measure $\hat{m}$ of the evidence that the two medians are different. It is given by the difference between the number of observations in each sample that are above the common sample median, normalized by the number of observations in each sample. Formally, $\hat{m}$ is defined by

$$\hat{m} = \frac{|\{k : x_{1k} \geq med(\{x_{ij}\})\}|}{n_1} - \frac{|\{k : x_{2k} \geq med(\{x_{ij}\})\}|}{n_2},$$

where x_{ij} is the jth observation in the sample of X_i.

The median difference test is concerned with $H_0 : med(X_1 - X_2) = 0$. The test for independent samples follows the lines of the test for stochastic inequality introduced in (Schlag, 2008). The test for matched pairs is easily designed using the binomial test. The estimated median difference is denoted by $m(x_1 - x_2)$. For matched pairs $m(x_1 - x_2)$ is given by the median of the sample differences. For independent samples the estimate is given by the expected value that is obtained by first forming $\min\{n_1, n_2\}$ random pairs, one observation from each sample, and then proceeding

as in the case of matched pairs.

The test for stochastic equality tests $H_0 : P(X_1 > X_2) = P(X_1 > X_2)$, the sample estimate is denoted by SI. The test for independent samples is introduced by (Schlag, 2008), the test for matched pairs is easily designed using the binomial test.

The test for the equality of means (Schlag, 2008) is concerned with $H_0 : EX_1 = EX_2$, the estimate Δ is the difference between the sample means divided by the range.

Information on the software implementation is provided in a separate section below.

	2015			WMW or W		stochastic equality			median difference		
	statement	x	n_1, n_2	p value	r	p value	r	SI	p value	r	$m\,(x_1 - x_2)$
A	G2 $\neq$ G3	$\neq$	12 m	0.06	$(\neq)$	0.04 (!!!)	$>$	0.67	0.04	$>$	2.13
	G2 $\neq$ G4	$\neq$	12 m	0.005	$\neq$	0.04	$>$	0.67	0.04	$>$	3.86
Cg	Cert $\neq$ Ind	$\neq$	63 m	10^{-9}	$\neq$	10^{-4}	$>$	0.56	0.31		20
	Ind $\neq$ Corr	$\neq$	63 m	0.015	$\neq$	0.08	$(>)$	0.21	1		0
	Cert $\neq$ Corr	$\neq$	63 m	0.04	$\neq$	0.08	$(<)$	-0.17	1		0
	r5: Ind $\neq$ Corr	$\neq$	63 m	0.001	$\neq$	0.001 (!!!)	$>$	0.38	0.21		9
E	Bias: m $\neq$ f	$\neq$	6, 6	0.004	$\neq$	0.04	$>$	0.94	0.04	$>$	95.5
	Turn: m $\neq$ f	$\neq$	6, 6	0.03	$\neq$	0.09	$(<)$	-0.75	0.09	$(<)$	-2.9
F	Ds: p1 $\neq$ p25	$>$	18 m	0.02	$\neq$	0.1 ^{w}	$(<)$	-0.44	0.1 ^{w}	$(<)$	-1.87
	Tl: b1 $\neq$ b3	$\neq$	6 m	0.03	$\neq$	0.04	$>$	1	0.04	$>$	0.95
L	F2$\neq$F1	$>$	4, 4	0.03	$\neq$	0.13 ^{z}		1	0.13 ^{z}		0.09
M	Cer$\neq$Pos	$\neq$	111 m	10^{-5}	$\neq$	10^{-4}	$>$	0.2	1		0
	Pos$\neq$Neg	$\neq$	111 m	10^{-4}	$\neq$	10^{-4} (!!!)	$>$	0.32	1		20
	Cer$\neq$Pos, 16	$\neq$	111 m	0.04	$\neq$	0.06	$(>)$	0.1	1		0
O	I-HS $\neq$ I-LS	$>$	28, 26	10^{-8}	$\neq$	10^{-5}	$>$	0.82	10^{-5}	$>$	0.35
W	K: t $\neq$ n w/o	$>$	89, 151	10^{-6}	$\neq$	10^{-4}	$>$	0.39	0.004	$>$	2
	D: t $\neq$ n w	$>$	89, 154	0.005	$\neq$	0.05	$>$	0.22	0.64		1

Table 2: Summary of findings for 2015 papers: p values for WMW or W, stochastic inequality and median difference.

	2015			WMW or W		equality of medians			equality of means			
	statement	x	n_1, n_2	p value	r	p value	r	$\hat{m}$	p value	r	range	Δ
A	G2 $\neq$ G3	$\neq$	12 m	0.06	$(\neq)$	0.56		0.33	1		$[0, 18]^y$	0.096
	G2 $\neq$ G4	$\neq$	12 m	0.005	$\neq$	0.56		0.33	1		$[0, 18]^y$	0.16
Cg	Cert $\neq$ Ind	$\neq$	63 m	10^{-9}	$\neq$	10^{-6}	$>$	0.43	0.001	$>$	$[0, 100]$	0.23
	Ind $\neq$ Corr	$\neq$	63 m	0.015	$\neq$	0.25		0.11	0.59		$[0, 100]$	0.09
	Cert $\neq$ Corr	$\neq$	63 m	0.04	$\neq$	0.05	$<$	0^{e*}	0.4		$[0, 100]$	-0.09
	r5: Ind $\neq$ Corr	$\neq$	63 m	0.001	$\neq$	0.57		0.25	0.27		$[0, 100]$	0.13
E	Bias: m $\neq$ f	$\neq$	6, 6	0.004	$\neq$	0.08	$(>)$	0.67	0.21		$[-80, 140]^y$	0.45
	Turn: m $\neq$ f	$\neq$	6, 6	0.03	$\neq$	0.3		-0.67	0.95		$[5, 25]^y$	-0.23
F	Ds: p1 $\neq$ p25	$>$	18 m	0.02	$\neq$	0.19 ^{z}		-0.33	0.94		$[0, 8]^y$	-0.071
	Tl: b1 $\neq$ b3	$\neq$	6 m	0.03	$\neq$	too small			1		$[0, 5]^y$	0.15
L	F2$\neq$F1	$>$	4, 4	0.03	$\neq$	0.25		1	1		$[0, 1]$	0.096
M	Cer$\neq$Pos	$\neq$	111 m	10^{-5}	$\neq$	0.001	$>$	0.18	0.003	$>$	$[0, 100]$	0.13
	Pos$\neq$Neg	$\neq$	111 m	10^{-4}	$\neq$	10^{-6} (!!)	$<$	-0.13^e	0.02	$>$	$[0, 100]$	0.16
	Cer$\neq$Pos, 16	$\neq$	111 m	0.04	$\neq$	0.006 (!!)	$>$	0.099	0.55		$[0, 100]$	0.056
O	I-HS $\neq$ I-LS	$>$	28, 26	10^{-8}	$\neq$	0.001	$>$	0.63	0.002	$>$	$[0, 1]$	0.34
W	K: t $\neq$ n w/o	$>$	89, 151	10^{-6}	$\neq$	0.03	$>$	0.27	0.02	$>$	$[0, 20]$	0.12
	D: t $\neq$ n w	$>$	89, 154	0.005	$\neq$	0.37		0.13 ^{z}	0.5		$[0, 41]$	0.045

Table 3: Summary of findings for 2015 papers: p values for WMW or W, equality of medians and means.

	2014			WMW or W		stochastic equality			median difference		
	statement	x	n_1, n_2	p value	r	p value	r	SI	p value	r	$m\,(x_1 - x_2)$
B	R: C $\neq$ VC	$<$	$4, 4$	0.03	$\neq$	0.13 ^{z}		-1	0.13 ^{z}		-0.06
	Red $\neq$ Blue	$<$	4 m	0.13		0.13 ^{z} (!!!)		-1	0.13 ^{z}		-0.095
C	$U_a \neq S$	$\neq$	$6, 6$	0.002	$\neq$	0.04	$>$	1	0.04	$>$	2.2
CP	r 4 $\neq$ r 5	$<$	$45, 39$	0.02	$\neq$	0.06 ^{w}	$(<)$	-0.31	0.86		-0.5
	r 4 $\neq$ r 6	$<$	$45, 39$	0.001	$\neq$	0.005	$<$	-0.43	0.19 ^{z}		-0.5
Ch	Gr $\neq$ No	$>$	$24, 24$	0.0001	$\neq$	0.001	$>$	0.68	0.001	$>$	4.5
Cn	t $\neq$ c	$>$	$11, 10$	0.13		0.29		0.4	1		1
Dt	T $\neq$ B	$\neq$	$10, 10$	0.02	$\neq$	0.07	$(<)$	-0.64	1		2.4
Du	M6 $\neq$ M14	$\neq$	$4, 4$	0.03	$\neq$	0.13		1	0.13		0.09
H	AH3 $\neq$ SH1	$\neq$	10 m	0.005	$\neq$	0.07	$(>)$	0.64	0.07	$(>)$	0.14
	AH1: PS$\neq$RSD	$\neq$	$11, 6$	0.05	$\neq$	0.22		-0.61	0.24		-1
	AH2: PS$\neq$RSD	$\neq$	$11, 6$	0.001	$\neq$	0.04	$>$	1	0.04	$>$	12

Table 4A: Summary of findings for 2014 papers, authors A-H: p values for WMW or W, stochastic equality and median difference.

	2014			WMW or W		stochastic equality			median difference		
	statement	x	n_1, n_2	p value	r	p value	r	SI	p value	r	$m\,(x_1 - x_2)$
I	G9: Fav ≠ Un	>	36 m	0.04	≠	0.11 ^{z}		0.17	1		0
	G17: Fav ≠ Un	>	36 m	0.003	≠	0.002 (!!!)	>	0.44	0.62		1
	G26: Fav ≠ Un	>	42 m	0.06	(≠)	0.12 ^{z}		0.21	1		0
K	pct ≠ pct2	>	16 m	0.003	≠	0.02	>	0.63	0.03	>	25.3
M	IS end ≠ IS ex	>	6, 4	0.02	≠	0.15 ^{z}		0.92	0.15 ^{z}		4.5
	FS end ≠ FS ex	>	9, 7	0.07	(≠)	0.27		0.56	0.27		2.35
N	high ≠ hypo	≠	1065, 994	10^{-10} ^{a}	≠	10^{-10} (!!!)	<	−0.21	1		0
	real ≠ lab	≠	1395, 109	0.09	(≠)	0.24		−0.095	1		0
P	NS ≠ NS+	≠	19, 19	0.06	(≠)	0.19		0.37	0.23		2.25
	NS ≠ RS	≠	19, 15	0.02	≠	0.09	(>)	0.48	0.29		3
	price NS ≠ NS+	≠	7, 5	0.03	≠	0.16		−0.71	0.72		−1
S	Action≠NE	≠	5 m	0.07	(≠)	0.07	(>)	1	0.07	(>)	15.78

Table 4B: Summary of findings for 2014 papers authors I-Z: p values for WMW or W, stochastic equality and median difference.

	2014			WMW or W		equality of medians			equality of means			
	statement	x	n_1, n_2	p value	r	p value	r	$\hat{m}$	p value	r	range	Δ
B	R: C $\neq$ VC	$<$	$4, 4$	0.03	$\neq$	0.25		-1	0.43		$[0.85, 1]^y$	-0.4
	Red $\neq$ Blue	$<$	4 m	0.13		too small			0.81		$[0.85, 1]^y$	-0.58
C	$U_a \neq S$	$\neq$	$6, 6$	0.002	$\neq$	0.07	(>)	1	0.06	(>)	$[47, 51]^y$	0.56
CP	r 4 $\neq$ r 5	$<$	$45, 39$	0.02	$\neq$	0.43		-0.18	0.99		$[0, 5]$	-0.074
	r 4 $\neq$ r 6	$<$	$45, 39$	0.001	$\neq$	0.06 ^{w}	(<)	-0.21	0.85		$[0, 6]$	0.086
Ch	Gr $\neq$ No	$>$	$24, 24$	0.0001	$\neq$	0.02	$>$	0.5	0.12 ^{z}		$[0, 25]$	0.19
Cn	t $\neq$ c	$>$	$11, 10$	0.13		1		0.31	0.76		$[1, 7]$	0.15
Dt	T $\neq$ B	$\neq$	$10, 10$	0.02	$\neq$	0.44		-0.4	0.25		$[-1, 8]$	-0.28
Du	M6 $\neq$ M14	$\neq$	$4, 4$	0.03	$\neq$	0.25		1	1		$[0, 1]$	0.095
H	AH3 $\neq$ SH1	$\neq$	10 m	0.005	$\neq$	0.31		0.46	1		$[0, 1]$	0.14
	AH1: PS$\neq$RSD	$\neq$	$11, 6$	0.05	$\neq$	0.56		-0.47	0.94		$[91, 97]$	-0.19
	AH2: PS$\neq$RSD	$\neq$	$11, 6$	0.001	$\neq$	0.05	$>$	0.818	0.11		$[70, 97]$	0.45

Table 5A: Summary of findings for 2014 papers, authors A-H: p values for WMW or W, equality of medians and means.

	2014			WMW or W		equality of medians			equality of means			
	statement	x	n_1, n_2	p value	r	p value	r	$\hat{m}$	p value	r	range	Δ
I	G9: Fav $\neq$ Un	$>$	36 m	0.04	$\neq$	1		0.17	1		$[0, 10]^y$	0.025
	G17: Fav $\neq$ Un	$>$	36 m	0.003	$\neq$	0.58		0.47	1		$[0, 10]^y$	0.053
	G26: Fav $\neq$ Un	$>$	42 m	0.06	$(\neq)$	1		0.14	1		$[0, 10]^y$	0.064
K	pct $\neq$ pct2	$>$	16 m	0.003	$\neq$	0.19 ^{z}		0.5	1		$[0, 200]^y$	0.11
M	IS end $\neq$ IS ex	$>$	$6, 4$	0.02	$\neq$	0.21		0.83	1		$[0, 20]$	0.2
	FS end $\neq$ FS ex	$>$	$9, 7$	0.07	$(\neq)$	0.43		0.38	1		$[0, 20]$	0.11
N	high $\neq$ hypo	$\neq$	$1065, 994$	10^{-10} ^{a}	$\neq$	0.002	$<$	-0.18	10^{-8} ^{b}	$<$	$[0, 5]$	-0.115
	real $\neq$ lab	$\neq$	$1395, 109$	0.09	$(\neq)$	1		-0.11	0.17 ^{b}		$[0, 5]$	-0.079
P	NS $\neq$ NS+	$\neq$	$19, 19$	0.06	$(\neq)$	0.48		0.26	0.86		$[-3, 25]^y$	0.098
	NS $\neq$ RS	$\neq$	$19, 15$	0.02	$\neq$	0.14		0.35	0.84		$[-2, 25]^y$	0.11
	price NS $\neq$ NS+	$\neq$	$7, 5$	0.03	$\neq$	0.78		-0.71	0.98		$[-4, 1]^y$	-0.2
S	Action$\neq$NE	$\neq$	5 m	0.07	$(\neq)$	too small			0.88		$[44, 80]^y$	0.46

Table 5B: Summary of findings for 2014 papers authors I-Z: p values for WMW or W, equality of medians and means.

	statement	2015	statement	where
A	G2 ≠ G3	Asparouhova etal	Mispricing: G2 ≠ G3	p90, pr-1
	G2 ≠ G4		Mispricing: G2 ≠ G4	p90, pr-1
Cg	Cert ≠ Ind	Cheung	Cert ≠ Ind	T2, row 1
	Ind ≠ Corr		Ind ≠ Corr	T2, row 3
	Cert ≠ Corr		Cert ≠ Corr	T2, row 4
	r5: Ind ≠ Corr		Ind ≠ Corr	T2, row 5
E	Bias: m ≠ f	Eckel etal	Avgbias: all m ≠ all f	p911, pr-1
	Turn: m ≠ f		Turnover: all m ≠ all f	p911, pr-1
F	Ds: p1 ≠ p25	Friedman etal	Duo short: per 1 ≠ per 25	p192, pr1
	Tl: b1 ≠ b3		Tri long: block 1 ≠ block 3	p193 pr-3
L	F2≠F1	Lai etal	F(Game 2)≠F(Game 1)	T3
M	Cer≠Pos	Miao & Zhong	Cer≠Pos, g1 1-5	TA1
	Pos≠Neg		Pos≠Neg, g1 1-5	TA1
	Cer≠Pos, 16		Cer≠Pos, g1 5-16	TA1
O	I-HS ≠ I-LS	Oprea	Invest: I-HS ≠ I-LS	p2248, pr1
W	K: t ≠ n w/o	Wantchekon etal	Kids: treat ≠ no treat w/o sch	TXII
	D: t ≠ n w		Desc.: treat ≠ no treat w sch	TXII

Table 6: Key to abbreviations used in Tables 2 and 3.

	statement	2014	statement	where
B	R: C $\neq$ VC	Bhattacharya etal	Red all: C $\neq$ VC	T5
	Red $\neq$ Blue		C all: Red $\neq$ Blue	T6
C	$U_a \neq S$	Cason etal	Payoff cont. instant.: $U_a \neq S$	T2
CP	r 4 $\neq$ r 5	Cason & Plott	Offer: range 4 $\neq$ range 5	T2
	r 4 $\neq$ r 6		Offer: range 4 $\neq$ range 6	T2
Ch	Gr $\neq$ No	Charness etal	Per 1: GrNoTy $\neq$ NoGrNoTy	footn 25
Cn	t $\neq$ c	Chen etal	Comp: treat$\neq$control	p72, pr-1
Dt	T $\neq$ B	Dittmann etal	Premium: Treat$\neq$Base	T5
Du	M6 $\neq$ M14	Duffy & Puzello	Efficiency M6$\neq$M14	p1754, top
H	AH3 $\neq$ SH1	Hugh-Jones etal	AH3 $\neq$ SH1	p376, top
	AH1: PS$\neq$RSD		AH1: PS$\neq$RSD	T3
	AH2: PS$\neq$RSD		AH2: PS$\neq$RSD	T3
I	G9: Fav $\neq$ Un	Isoni etal	Fav Earn $\neq$ Unfav Earn	T2A G9
	G17: Fav $\neq$ Un		Fav Earn $\neq$ Unfav Earn	T2A G17
	G26: Fav $\neq$ Un		Fav Earn $\neq$ Unfav Earn	T2B G26
K	pct $\neq$ pct2	Kosfeld & Rustagi	L_{NP} pct $\neq$ pct2	p775, pr -1
M	IS end $\neq$ IS ex	Markussen etal	Contr: IS end $\neq$ IS ex	p318, pr 2
	FS end $\neq$ FS ex		Contr: FS end $\neq$ FS ex	p318, pr 3
N	high $\neq$ hypo	Noussair etal	Riskaversion: high $\neq$ hypohigh	T3
	real $\neq$ lab		Riskaversion: real $\neq$ lab	T3
P	NS $\neq$ NS+	Petersen & Winn	dev NS $\neq$ NS+	p1057, pr -3
	NS $\neq$ RS		dev NS $\neq$ RS	p1057, pr -3
	price NS $\neq$ NS+		Type y firms: price NS $\neq$ NS+	p1058, pr 3
S	Action$\neq$NE	Shapiro etal	r=0.15,type5: Action$\neq$NE	footn 6

Table 7: Key to abbreviations used in Tables 4AB and 5AB.

Explaining symbols used in Tables: 2,3,$AB,5AB

- all tests are two-sided
- all p values rounded up, reported on the grid $10^{-10}, .., 10^{-4}, 0.001, .., 0.009$, and above to two decimals behind the comma
- WMW and W use exact distribution, W deals with ties as in Pratt (1959) to be comparable to STATA although STATA uses asymptotic p values.
- 'T' stands for table
- 'pr' stands for paragraph
- 'x' indicates desired alternative hypothesis in paper
- 'm' refers to matched pairs
- 'r' refers to significant evidence at level 0.05 of the two-sided test, in brackets if marginally significant at level 0.1
- '$\hat{m}$' adhoc measure of difference in medians, defined as difference of proportion of observations above the common median
- '*': instance where $\hat{m} = 0$ but medians are different, given by 0 and 20.
- 'e': one of the sample medians is at the extreme points of data range
- 'range' indicates range of outcomes used for running test for comparing means
- 'y' indicates that range chosen more or less arbitrarily based on data, so inference is conditional on outcomes belonging to that interval
- 'SI': estimate of the stochastic inequality, given by the difference in the proportions of matches in which observation in first sample greater and smaller
- 'Δ': difference in means normalized by the range
- '(!!)' highlights that p value of median test less or equal to that of W or WMW test
- '(!!!)' highlights that p value of median test less or equal to that of W or WMW test
- 'b': in means test with very large data sets we exogenously choose $\theta = 0.3$

- ‘too small’: indicates that sample size too small for sensible results, for these sample sizes no data can lead to p value below 0.2 with our median test
- ‘z‘: the authors are interested in a one-sided test, for this the result is marginally significant (p value at most 0.1)
- ‘w‘: the authors are interested in a one-sided test, for this the result is significant (p value at most 0.05)

C Confidence Intervals

For a subset of the above data we compute confidence intervals for the three tests that involve ordinal comparisons.

Confidence Intervals				stochastic difference $P(X_1 > X_2) - P(X_1 < X_2)$		median difference $med(X_1 - X_2)$	
	statement	where	n_1, n_2	95% CI	SI	95% CI	$m(x_1 - x_2)$
Friedman etal (2015)	Duo short: per 1 $\neq$ per 25	p192, pr1	18 m	$[-0.81, 0.06]$	-0.44	$[-2.6, 0.08]$	-1.87
Isoni (2014)	Fav Earn $\neq$ Unfav Earn	T2A G9	36 m	$[-0.11, 0.94]$	0.167	$[0, 0]$	0
	Fav Earn $\neq$ Unfav Earn	T2B G26	42 m	$[-0.07, 0.66]$	0.21	$[0, 1.5]$	0
Cheung (2015)	Cert $\neq$ Ind	T2, row 1	63 m	$[0.8, 1]$	0.56	$[0, 45]$	20
	Ind $\neq$ Corr	T2, row 3	63 m	$[-0.02, 0.54]$	0.21	$[0, 10]$	0
Miao & Zhong (2015)	Cer$\neq$Pos, g1 1-5	TA1	111 m	$[0.5, 0.98]$	0.2	$[0, 0]$	0
Eckel etal (2015)	Avgbias: all m $\neq$ all f	p911, pr-1	$6, 6$	$[0.05, 0.99]$	0.94	$[13, 178]$	95
Chen etal (2014)	Comp: treat$\neq$control	p72, pr-1	$11, 10$	$[-0.35, 0.93]$	0.4	$[0, 3]$	1
Petersen & Winn (2014)	dev NS $\neq$ NS+	p1057, pr -3	$19, 19$	$[-0.12, 0.74]$	0.37	$[-0.7, 5.7]$	2.2
	dev NS $\neq$ RS	p1057, pr -3	$19, 15$	$[-0.06, 0.88]$	0.48	$[0, 6]$	3
Charness etal (2014)	Per 1: GrNoTy $\neq$ NoGrNoTy	footn 25	$24, 24$	$[0.28, 0.9]$	0.68	$[1.2, 8.9]$	4.5
Oprea (2015)	Invest: I-HS $\neq$ I-LS	p2248, pr1	$28, 26$	$[0.5, 0.95]$	0.82	$[0.14, 0.54]$	0.35
Cason & Plott (2014)	Offer: range 4 $\neq$ range 5	T2	$45, 39$	$[-0.7, 0.01]$	-0.31	$[-0.86, 0]$	-0.5
Wantchekon etal (2015)	Kids: treat $\neq$ no treat w/o sch	TXII	$89, 151$	$[0.18, 0.62]$	0.39	$[1, 3]$	2
	Desc.: treat $\neq$ no treat w sch	TXII	$89, 154$	$[0.01, 0.47]$	0.22	$[0, 3]$	1
Noussair etal (2014)	Riskaversion: high $\neq$ hypohigh	T3	$1065, 994$	$[-0.4, -0.15]$	-0.21	$[-1, 0]$	0
	Riskaversion: real $\neq$ lab	T3	$1395, 109$	$[-0.37, 0.12]$	-0.095	$[-1, 0]$	0

Table 8A: Summary of findings: confidence intervals for stochastic and median difference

	Confidence Intervals				difference in medians $med(X_1) - med(X_2)$	
	statement	where	n_1, n_2	95% CI	$m(x_1) - m(x_2)$	dr
Friedman etal (2015)	Duo short: per 1 $\neq$ per 25	p192, pr1	18 m	$[-3, 0.9]$	-1.41	$[-3.53, 1.02]$
Isoni (2014)	Fav Earn $\neq$ Unfav Earn	T2A G9	36 m	$[0, 0]$ (!!)	0	$[0, 0.5]$
	Fav Earn $\neq$ Unfav Earn	T2B G26	42 m	$[-1.5, 3]$	1.5	$[-1.5, 1.5]$ (!)
Cheung (2015)	Cert $\neq$ Ind	T2, row 1	63 m	$[50, 50]$	50	$[50, 50]$
	Ind $\neq$ Corr	T2, row 3	63 m	$[-2, 39]$	5	$[-5, 40]$
Miao & Zhong (2015)	Cer$\neq$Pos, g1 1-5	TA1	111 m	$[50, 50]$	0	$[0, 50]$
Eckel etal (2015)	Avgbias: all m $\neq$ all f	p911, pr-1	$6, 6$	too small	106	-
Chen etal (2014)	Comp: treat$\neq$control	p72, pr-1	$11, 10$	$[0, 2]$	0	$[-3, 4]$
Petersen & Winn (2014)	dev NS $\neq$ NS+	p1057, pr -3	$19, 19$	$[-2.7, 6.7]$	3.75	$[-5.75, 9]$
	dev NS $\neq$ RS	p1057, pr -3	$19, 15$	$[-2, 7.4]$	4.5	$[-2.75, 8.5]$
Charness etal (2014)	Per 1: GrNoTy $\neq$ NoGrNoTy	footn 25	$24, 24$	$[0.3, 10.4]$	5	$[-1.5, 10.5]$
Oprea (2015)	Invest: I-HS $\neq$ I-LS	p2248, pr1	$28, 26$	$[0.2, 0.6]$	0.4	$[0.1, 0.6]$
Cason & Plott (2014)	Offer: range 4 $\neq$ range 5	T2	$45, 39$	$[-1, 0]$	-0.5	$[-1, 0.13]$
Wantchekon etal (2015)	Kids: treat $\neq$ no treat w/o sch	TXII	$89, 151$	$[1, 4]$	3	$[0, 4]$
	Desc.: treat $\neq$ no treat w sch	TXII	$89, 154$	$[0, 4]$	2	$[-1, 5]$
Noussair etal (2014)	Riskaversion: high $\neq$ hypohigh	T3	$1065, 994$	$[-2, -2]$	-2	$[-2, -2]$
	Riskaversion: real $\neq$ lab	T3	$1395, 109$	$[-1, 1]$	-1	$[-1, 1]$

Table 8B: Summary of findings: confidenceintervals for difference in medians.

Additional information for the table with confidence intervals:

- $m\,(x_i)$ denotes the sample median of sample from X_i
- 'dr': naive estimate of range of difference in medians, computed by looking at range of differences among the median values that belong to 95% CI in each sample
- (!): this is the only case where the naive measure of range is smaller than the exact 95% CI
- (!!): In this data set we find significant evidence that the two distributions are different (see previous table) and significant evidence that the two medians are equal.

D Single Sample

For the same set of journals and same time period we found 7 papers that use the Wilcoxon test to compare observations in a single sample to a value. We obtained the data and necessary information to replicate their results for 6 of them from which we selected 9 hypotheses. The results are summarized below. The median test considers $H_0 : med(X) = m_0$ and is constructed using the binomial test, observing that if $med(X) \leq m_0$ then $P(X \leq m_0) \geq \frac{1}{2}$. The mean test is explained in (Schlag, 2008) and investigates $H_0 : EX = m_1$, the sample average is denoted by $\bar{x}$.

	Single Sample			Wilcoxon				median				mean		
	statement	x	n	p value	r	p value	r	95% CI	$m\,(x)$	p value	r	range	95% CI	$\bar{x}$
C	S Cont Inst: payoff$\neq$ 48	$\neq$	6	0.06	($\neq$)	0.22		$[47.2, 48.2]$ [k]	47.53	0.96		$[47, 51]^y$	$[47.3, 49.1]$	47.6
D	Voting A share $\neq 0$	$>$	10	0.002	$\neq$	0.002	$>$	$[15.47, 22.27]$	18.27	0.02	$>$	$[-10, 27]$	$[4.4, 23.1]$	18.52
E	Averagebias all male$\neq$ 0	$>$	6	0.03	$\neq$	0.04	$>$	$[15.2, 131]$ [k]	74.72	0.05	$>$	$[-80, 140]^y$	$[1, 121]$	74.1
N	Lab: riskaversion$\neq$ 2.5	$>$	109	10^{-10}	$\neq$	10^{-9}	$>$	$[4, 4]$	4	$10^{\cdot-10}$	$>$	$[0, 5]$	$[3.3, 3.9]$	3.6
	Hypohigh: riskaversion$\neq$ 2.5	$>$	994	10^{-10}	$\neq$	10^{-10}	$>$	$[4, 4]$	5 (!)	10^{-10}	$>$	$[0, 5]$	$[3.65, 3.89]$	3.78
	Riskaversion$\neq$ 2.5	$>$	3563	10^{-10} [a]	$\neq$	10^{-10}	$>$	$[4, 4]$	4	10^{-10}	$>$	$[0, 5]$	$[3.32, 3.45]$	3.38
O	Inv I-HS $\neq 0.5$	$\neq$	28	10^{-7}	$\neq$	0.0001	$>$	$[0.95, 1]$	1	10^{-6}	$>$	$[0, 1]$	$[0.78, 0.97]$	0.92
P	Average Dev NH$\neq$ 0 in $T+1$	$\neq$	10	0.002	$\neq$	0.002	$>$	$[2.25, 8.75]$	5.625	0.04	$>$	$[-10, 10]^y$	$[0.4, 8.2]$	5.63
	Average Dev NH$\neq$ 0 in $T+5$	$\neq$	10	0.02	$\neq$	0.75		$[0, 6.5]$	0.875	0.27		$[-10, 10]^y$	$[-1.9, 5.9]$	2.6

	Key to Abbreviations	where
C	Cason etal (2014)	T2
D	Dittmann etal (2015)	T3A
E	Eckel etal (2015)	p910, pr-1
N	Noussair etal (2014)	T3
		T3
		T3
O	Oprea (2015)	footn 34
P	Petersen Winn (2014)	T5
		T5

Table 9: Summary of findings: p values for W, p values and confidence intervals for median and mean.

Additional information for the table with data on single samples:

- 'a': software could not compute exact distribution with this sample size
- '(!)' highlights that the sample median lies outside the confidence interval of the median of the underlying distribution
- 'k' this is the entire range of data

E Tests for Comparing Medians

In this article we introduce two novel exact tests for comparing medians, one for two independent samples and one for matched pairs. The only other known test for comparing medians is for two independent samples and involves looking at the overlap of the confidence intervals of each sample (see Schlag, 2011). In all of the data sets investigated above this alternative test is less powerful. In the following we explain the details behind these two tests. It is enough to construct exact tests of the one-sided null hypothesis $H_0 : med(X_1) \leq med(X_2)$.

Consider inference based on two independent samples. Choose r such that based on the binomial test there is evidence at level αw that $med(X_1) = med(X_2) = m$ is not true whenever $s_1 + s_2 < r$ or $s_1 + s_2 > n_1 + n_2 - r$. This means we choose the largest value of r such that the probability of obtaining strictly less then r successes among $n_1 + n_2$ independent draws with success probability $1/2$ is at most $w\alpha/2$. We set $w = 1/10$.

Consider some $m \in \mathbb{R}$. Let $s_1 = |\{j : x_{1j} > m\}|$ and $s_2 = |\{j : x_{2j} \geq m\}|$. Assume that $med(X_1) \leq med(X_2) = m$. Let $\phi_m(s_1, s_2) = \infty$ if $s_1 + s_2 < r$ or $s_1 + s_2 > n_1 + n_2 - r$ and let

$$\phi_m(s_1, s_2) = \frac{s_1 - s_2}{\left(\frac{s_1+s_2}{n_1+n_2}\left(1 - \frac{s_1+s_2}{n_1+n_2}\right)\right)^{3/2}}$$

if $r \leq s_1 + s_2 \leq n_1 + n_2 - r$. We now find z such that the maximal probability of obtaining $\phi_m \geq z$ is at most α when H_0 is true. The test recommends to reject H_0 if $\phi_m \geq z$ holds for all m. Let $Z_i \in \{0, 1\}$ be such that $Z_1 = 1$ if $X_1 > m$ and $Z_2 = 1$ if $X_2 \geq m$. Under H_0, we have $P(Z_2 = 1) \geq \frac{1}{2}$ and $P(Z_1 = 1) = P(X_1 > m) \leq P(X_1 > med(X_1)) \leq \frac{1}{2}$ as $med(X_1) \leq m$. Clearly ϕ_m is maximized when $P(Z_1 = 1) = \frac{1}{2}$. We choose the smallest value of z that satisfies

$$\max_{q \in [0,1]} \frac{1}{2^{n_1+n_2}} \sum_{s_1=0}^{n_1} \sum_{s_2=0}^{n_2} \binom{n_1}{s_1} \binom{n_2}{s_2} 1_{\{\phi_m(s_1,s_2) \geq z\}} \leq \alpha$$

where $1_{\{.\}}$ is the indicator function.

We now prove that the rejection probability is bounded above by α when H_0 is true. Clearly this is the case when $P(X_2 \geq med(X_2)) = \frac{1}{2}$. Assume that $P(X_2 \geq med(X_2)) > \frac{1}{2}$. Then there exists an independent random variable $Q \in \{0,1\}$ such that $P(X_2 \geq med(X_2), Q = 1) = \frac{1}{2}$. It is as if each of the realizations of $X_2 = med(X_2)$ have a label Q. Assume that Q is observable each time X_2 is observable. Replace Z_2 above by $Z_2' \in \{0,1\}$ defined by $Z_2' = 1$ if $X_2 \geq m$ and $Q = 1$ and replace s_2 by $s_2' = |\{j : x_{2j} \geq m, q_j = 1\}|$. Then $P(Z_2' = 1) = 1/2$ under H_0 and the rejection probability of the above test is bounded above by α. Note that $s_2' \leq s_2$. If we then replace back s_2' by s_2 then the test statistic gets smaller and we reject less often, in particular the rejection probability is bounded above by α. But if we replace s_2' by s_2 then we do not need to observe Q and the proof is complete.

The test for matched pairs is very similar to the one for independent samples. Choose r such that the probability of obtaining strictly less than r successes among n independent draws with success probability $\frac{1}{2}$ is at most $w\alpha/2$ where $w = 1/10$.

Consider again some $m \in \mathbb{R}$. Let $b = |\{j : x_{1j} > m > x_{2j}\}|$, $c = |\{j : x_{1j} \leq m \leq x_{2j}\}|$ and $d = |\{j : x_{1j} > m, x_{2j} \geq m\}|$. Assume that $med(X_1) \leq med(X_2) = m$. Let $\phi_m(b,c) = \infty$ if $b + c + 2d < r$ or if $b + c + 2d > 2n - r$, for $r \leq b + c + 2d \leq 2n - r$ let $\phi_m(b,c) = \frac{b-c}{\sqrt{b+c}}$ if $b + c > 0$ and $\phi_m(0,0) = 0$. We choose z such that the maximal probability of obtaining $\phi_m \geq z$ is at most α when H_0 is true. Let $Z \in \{0,1\}^2$ such that $Z_i = 1$ if and only if $X_i \geq m$. Under H_0, we have $P(Z_2 = 1) \geq \frac{1}{2}$ and $P(Z_1 = 1) = P(X_1 > m) \leq P(X_1 > med(X_1)) \leq \frac{1}{2}$ as $med(X_1) \leq m$. Clearly ϕ_m is maximized when $P(Z_1 = 1) = \frac{1}{2}$. Let $q = P(Z = (1,0))$. Then $P(Z = (0,1)) = q$ and $P(Z = (1,1)) = \frac{1}{2} - q$ when $P(Z_2 = 1) = \frac{1}{2}$. We choose the smallest z such that

$$\max_{q \in [0,1]} \sum_{b=0}^{n} \sum_{c=0}^{n-b} \sum_{d=0}^{n-b-c} \frac{n!}{b!c!(n-b-c-d)!} q^{b+c} \left(\frac{1}{2} - q\right)^{n-b-c} 1_{\{\phi_m(b,c) \geq z\}} \leq \alpha.$$

It follows, using analogous arguments as in the case of independent samples, that the rejection probability of the test is bounded above by α when H_0 is true.

The above tests for two independent samples and matched pairs are easily generalized to be able to consider $H_0 : med(X_1) = m_1, med(X_2) = m_2$ for all $(m_1, m_2) \in \mathbb{R}^2$ and thus to construct confidence intervals for $med(X_1) - med(X_2)$.

F Software

All software is available in R. We used the implementations of the Wilcoxon-Mann-Whitney and Wilcoxon tests available in the R package 'coin'. For the robust rank order test we took the test statistic from the R package 'NSM3' and critical values from Feltovich (2005) which are only available for $n_i \leq 40$. For the test of stochastic inequality of Brunner and Munzel (2000) we used the R package 'lawstat'. Software for the tests for stochastic inequality, equality of means (for small and moderate sample sizes) and for the mean of a single sample are contained in the R package 'npExact' that can be found at https://github.com/zauster/npExact.git. The equality of means test for large sample sizes and the two tests for comparing medians are available at https://homepage.univie.ac.at/karl.schlag/. The median difference test is derived from the test for stochastic inequality as follows. We wish to test $H_0 : med\,(X_1 - X_2) \leq d$, which is equivalent to $H_0 : P\,(X_1 \leq X_2 + d) \geq \frac{1}{2}$ which is equivalent to $H_0 : P\,(X_1 > X_2 + d) \leq P\,(X_1 \leq X_2 + d)$. Consider some $\varepsilon > 0$ with $\varepsilon < \min_{i,j} \{x_{1i} - x_{2j} : x_{1i} \neq x_{2j}\}$ and test $H_0 : P\,(X_1' > X_2') \leq P\,(X_1' < X_2')$, setting $X_1' = X_1 - \varepsilon$ and $X_2' = X_2 + d$. So we decreased X_1 slightly so that there are no more ties and so that $x_{1i} \leq x_{2j} + d$ holds if and only if $(x_{1i} - \varepsilon) < x_{2j} + d$. To test $H_0 : med\,(X_1 - X_2) \geq d$ use the fact that this is equivalent to testing $H_0 : med\,(X_2 - X_1) \leq -d$.