A New Method for Constructing Exact Tests without Making any Assumptions[1]

Karl H. Schlag[2]

August 23, 2008

[1]Special thanks to Alexander Schlag Lawrence, Joachim Röhmel, Richard Spady and David Thesmar. Thanks also to Rachel Croson and to Dr. Sabeti-Aschraf for supplying some data.

[2]Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, karl.schlag@upf.edu

Abstract

We present a new method for constructing exact distribution-free tests (and confidence intervals) for variables that can generate more than two possible outcomes. This method separates the search for an exact test from the goal to create a non-randomized test. Randomization is used to extend any exact test relating to means of variables with finitely many outcomes to variables with outcomes belonging to a given bounded set. Tests in terms of variance and covariance are reduced to tests relating to means. Randomness is then eliminated in a separate step.

This method is used to create confidence intervals for the difference between two means (or variances) and tests of stochastic inequality and correlation.*Keywords*: distribution-free, nonparametric, exact hypothesis testing, unavoidable inaccuracy, nonparametric Behrens-Fisher problem, UMPU test, Kendall's tau, Q_n.

JEL classification numbers: C12, C14.

1 Introduction

"Let the data speak!" We are interested in making inference without making assumptions. We wish to make statements in terms of significance that can be deduced directly from the data without having to add additional assumptions. Inference will only be based on knowledge, knowledge for instance about the possible outcomes and on how the data was gathered (e.g. i.i.d.). No distributional assumptions are made, our approach is *distribution-free*. Furthermore we wish to make statements that can be proven mathematically, hence this paper is about *exact* inference. Inference will be concerned with parameters of distributions, such the mean or the variance, our contribution applies to variables that can generate more than two possible outcomes.

Permutation tests have proven to be very useful for comparing distributions, such as for testing identity or independence, when variables can generate many different outcomes. However only few exact tests exist for inference in terms of parameters such as means or variances. A dilemma seems to emerge when the set of underlying data generating processes is rich such as when one allows for any distribution that generates outcomes in a given interval. Formal analysis is required as the entire parameter space cannot be rigorously explored with simulations. Either the mathematical methods for dealing with the complexity of the problem are very crude or the analysis and resulting tests are complex both in terms of derivation and implementation. We explain. Bickel et al. (1989) and Fishman (1991) derive an exact test for the mean of a single sample based on the Hoeffding bounds, the test is simple but the bounds used in the construction are conservative and the test is not very powerful. Romano and Wolf (2000) and Diouf and Dufour (2006) design exact tests for a mean of a single sample and Romano and Wolf (2002) for the variance of a single sample. Their methods are sophisticated, yielding intricate tests, and yet it does not seem feasible to evaluate the performance of these tests in finite samples. Exact distribution-free tests for comparing means given independent samples or for testing for correlation have not been available. We introduce a new method that involves two steps that together allow to construct simple tests for these and other problems of distribution-free inference, exact tests whose performance can be measured for any given sample size.

Given the practical demand for distribution-free tests we find two approaches in

the literature. One has been to simplify inference by ignoring part of the underlying space of distributions. For instance, this is implicitly the case when using the Spearman rank correlation test (Spearman, 1904) to test for correlation. This test is only exact if one ignores distributions that are uncorrelated but not independent. The other approach is more heuristic and builds on insights gained from asymptotic theory, acting as if the sample is sufficiently large. This can be arbitrarily misleading when interested in exact inference as illustrated by Lehmann and Loh (1990) who showed that the size of the t test for testing the mean of a single sample is equal to 1. Exact inference requires that one is able to prove that the stated levels are correct for the given sample size across all possible distributions.

The key to our new method is that it separates the construction of a powerful exact test from the objective to end up with a nonrandomized test. Two separate steps, one for the construction and one for eliminating randomness make the methodology particularly transparent and simple. In fact we can explain the main mechanisms without using formal mathematics. Assume that one is interested in a process that generates outcomes contained in the unit interval $[0,1]$. It is useful to visualize hypothesis testing as a game against nature. Assume that the statistician is able to randomly transform each observed data point into a binary value such that the mean is preserved. As long as the objective of the statistician is formulated in terms of means only, given that this transformation will be undergone anyway, nature might as well choose as true data generating process one that only realizes binary values. The statistician knows this and realizes that after the transformation it is as if he or she faces a much smaller set of possible distributions, namely only those that realize binary values. All the statistician now has to do is to apply some exact test for binary valued data. Note that the random mean preserving transformation is simple, replace observation $y \in [0,1]$ with 1 with probability equal to y and with 0 with probability equal to $(1-y)$. One practically sees without proof that the above method generates an exact test. In fact, a similar random transformation is available to transform the data into one that contains a given finite set of outcomes. This is valuable, as it reduces added variance, provided an exact test is known for this set of outcomes. A caveat is that the transformation and hence also the resulting test is randomized. In the second step we propose a simple way to eliminate this randomness without loosing

the property of being exact. The idea is to reject the null hypothesis if and only if the rejection probability of the above randomized test is above some threshold. The "mistake" that this cutoff strategy introduces is compensated by choosing a smaller size for the underlying exact test for the transformed data.

The power of our method depends on the power of the randomized test used in the first step. In this vein we derive a randomized test that is uniformly most powerful among all unbiased tests for testing the mean of a random variable that has three possible outcomes. This test is a simple extension of the randomized version of McNemar's test (McNemar, 1947, Lehmann, 1959).

The essential value of our construction is that the type II error can be bounded and hence the performance of the resulting test can be measured for any given sample size. One such measure is *inaccuracy* as defined by the maximal expected width of the associated confidence interval. *Relative efficiency* can then be defined as the ratio of the lower bound on inaccuracy to the bound on inaccuracy of the proposed test. For instance, we find in numerical examples that the *relative efficiency* of our nonrandomized test for the mean of a single sample is 68%.

To underline the usefulness of our new method we present many exact tests and confidence intervals. We show how to make inference in terms of means and variance in a single sample, with two independent samples and with matched pairs where we also consider covariance. For this one needs to know a bounded set that will contain all outcomes. We also show how to make inference that relies on ordinal comparisons and hence does not require such known bounds. We treat a stochastic inequality to compare outcomes given independent samples. We test for association that is related to Kendall's tau and show how to investigate a measure of spread that is related to Q_n.

Related literature is mentioned within the text of the main section. We only mention here that the main innovation of this paper, the two step procedure, has been previously used by Gupta and Hande (1992) in a specific problem of statistical decision making.

2 Two Steps to Hypothesis Testing

Our new method for constructing exact nonrandomized distribution-free tests separates the objective of constructing an exact test from the objective to create a test that is nonrandomized. Accordingly it consists of two steps. We first present the second step as it applies to more general settings.

2.1 Eliminating Randomness

In the following we show how to transform an exact randomized test into an exact nonrandomized test.

Consider inference based on a realization z of a random vector Z. Let $\mathcal{Z}$ be the set of possible realizations. Let P_Z be the distribution of z and let Ω be the domain of all possible distributions P_Z. Consider subsets $H_0, H_1 \subset \Omega$ such that $H_0 \cap H_1 = \emptyset$. A test ϕ is described by the probability $\phi(z) \in [0,1]$ of rejecting the null hypothesis H_0 in favor of the alternative hypothesis H_1 when observing z (for each $z \in \mathcal{Z}$). ϕ is called *nonrandomized* if $\phi(z) \in \{0,1\}$ for each $z \in \mathcal{Z}$, otherwise ϕ is called *randomized*. Let $E_Z(\phi) = \int z dP_Z(z)$ be the ex-ante probability of rejecting the null hypothesis before observing the realization of Z. Let $E_Z(1-\phi) = 1 - E_Z(\phi)$. ϕ is a (*exact*) *test with level* α if $\sup_{P_Z \in H_0} E_Z(\phi) \leq \alpha$. ϕ has *type II error* β given H_1 if $\sup_{P_Z \in H_1} E_Z(1-\phi) \leq \beta$. Implicitly it is assumed that these properties refer to statements that have been proven.

Given a test ϕ and a threshold $\theta \in (0,1)$ let $\phi|_\theta$ be the nonrandomized test defined by $\phi|_\theta(z) = 1$ if $\phi(z) \geq \theta$ and $\phi|_\theta(z) = 0$ if $\phi(z) < \theta$.

Theorem 1 *Let ϕ be an exact randomized test with level $\theta\alpha$. Then $\phi|_\theta$ is an exact nonrandomized test with level α. The type II error of $\phi|_\theta$ is bounded above by the type II error of the underlying randomized test ϕ divided by $(1-\theta)$, formally:*

$$\sup_{P_Z \in H_1} E_Z(1-\phi|_\theta) \leq \min\left\{1, \frac{1}{1-\theta} \sup_{P_Z \in H_1} E_Z(1-\phi)\right\}. \tag{1}$$

Proof. For Z such that $P_Z \in H_0$ we obtain

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

and hence $\sup_{P_Z \in H_0} E_Z(\phi|_\theta) \leq \alpha$.

Analogously,

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

and hence

$$E_Z(1-\phi|_\theta) \leq \frac{1}{1-\theta} E_Z(1-\phi).$$

Taking the supremum over all $P_Z \in H_1$ on both sides of the above inequality proves (1). ■

Gupta and Hande (1992) previously introduced this method of eliminating randomness for the case of $\theta = 1/2$, it was used to create a specific nonrandomized rule for statistical decision making. The bounds in terms of level and type II error are verified in the proof above very crudely, we do not expect them to be tight. Notice that Dvoretzky et al. (1951) present a formal methodology for eliminating randomness without loosing power, however their tests are not distribution-free, in particular they do not allow (as we do below) for all distributions with support contained in a given set of outcomes.

2.2 Creating Randomized Tests for Variables with Known Bounds

Here we show how to use randomization to extend exact tests for means to richer environments. We start with a test that applies only to data that generates one of finitely many different outcomes and generate one that applies to outcomes that belong to a given bounded set. In later applications we show how to use such tests for means to make inference in terms of variance and covariance. This material only applies if all random variables realize outcomes that belong to a known bounded set.

Processes that generate outcomes with known bounds are wide spread if not typical. Some hypotheses cannot be tested sensibly without such known bounds. For instance, Bahadur and Savage (1956) show that only trivial tests can be exact when

testing for the mean given a single sample without such bounds. Their arguments generalize easily to tests for comparing means given independent samples or matched pairs and to upper confidence bounds on variance.

Let $\mathcal{Y}_i \subset \mathbb{R}$ be a closed and bounded set with $|\mathcal{Y}_i| \geq 2$ for $i = 1, .., m$. Let Y be an m dimensional random vector with underlying distribution P_Y such that $Y_i \in \mathcal{Y}_i$ for $i = 1, .., m$. Let $a_i = \min \{\mathcal{Y}_i\}$ and $b_i = \max \{\mathcal{Y}_i\}$ where $a_i < b_i$. Normalize Y_i with a linear transformation $x \longmapsto (x - a_i) / (b_i - a_i)$ in order to now assume without loss of generality that $\{0, 1\} \subseteq \mathcal{Y}_i \subseteq [0, 1]$ for each i.

Consider inference based on n independent observations of Y that consist either of independent samples of each component or of matched vectors. In the first case there are n_i observations of Y_i for $i = 1, .., m$ where $n_1 + ... + n_m = n$ while in the second case each observation consists of a joint realization of Y. Let f be a random transformation of the data such that $f : (\cup_{i=1}^m \mathcal{Y}_i)^n \to \Delta\Psi^n$ where $\Psi = \{\psi_l\}_{l=0}^c$ and $\psi_0 = 0 < \psi_1 < ... < \psi_c = 1$ for some $c \in \mathbb{N}$. The transformation f is *mean preserving* if $Ef(y) = y$ for all $y \in (\cup_{i=1}^m \mathcal{Y}_i)^n$. In fact, there is no need to apply the same transformation to each variable, this is only done here to simplify presentation. $\phi \circ f$ will be called a *transformed test*.

For the applications we propose the following mean preserving transformations. Consider first the case of independent samples. For each data point y_j, draw a realization z from a uniform distribution on $[0, 1]$. Next determine l such that $\psi_l < z \leq \psi_{l+1}$ (we ignore the case where $z = 0$ as this almost surely does not occur). Independently of other events, replace y_j with either ψ_l or ψ_{l+1} as follows: replace y_j with ψ_{l+1} with probability $(y_j - \psi_l) / (\psi_{l+1} - \psi_l)$ and replace it with ψ_l with probability $(\psi_{l+1} - y_j) / (\psi_{l+1} - \psi_l)$. In the case of matched vectors consider the following two variations of the above transformation. One option is to transform each component of the single observation $y_j = (y_{jk})_{k=1}^m$ separately, thus drawing m realizations from the uniform distribution. The alternative is to transform the m observations jointly, namely to draw a single realization z given y and then to transform each component using the same value of z. It is easily verified that these three specific transformations are mean preserving.

Consider tests that only depend on the underlying means of each component, they are *distribution-free*. Specifically, let A_0 and A_1 will be the vectors of means

that are of interest under the hypotheses H_0 and H_1 respectively, $A_0, A_1 \subset [0,1]^m$ and $A_0 \cap A_1 \neq \emptyset$. We will show how to reduce a possibly *nonparametric* problem to a parametric problem without loosing on inference in terms of means at the expense of adding randomness.[1]

Theorem 2 *Let f be a mean preserving transformation.*

(i) If the test ϕ_g has size α and type II error β for testing

$$H_0 : \{P_Y \in \Delta\Psi^m : EY \in A_0\}, \; H_1 : \{P_Y \in \Delta\Psi^m : EY \in A_1\} \tag{2}$$

then $\phi_g \circ f$ is a test with size α and type II error β for testing

$$H_0' : \{P_Y \in \Delta[0,1]^m : EY \in A_0\}, \; H_1' : \{P_Y \in \Delta[0,1]^m : EY \in A_1\}. \tag{3}$$

(ii) If ϕ_g is a level α test that minimizes the type II error among all level α tests for (2) then $\phi_g \circ f$ is a level α test that minimizes the type II error among all level α tests for (3).

Proof. To prove (i) we use the fact that the transformation is mean preserving. When facing $P_Y \in \Delta[0,1]^m$ and applying f it is as if one is facing $P_{Y^b} \in \Delta\Psi^m$ such that $EY^b = EY$. Formally, $E_Y\left(\phi_g(f)\right) = E_{Y^b}\left(\phi_g\right)$ and hence

$$\sup_{P_Y \in H_0} E_Y\left(\phi_g\right) = \sup_{P_Y \in H_0'} E_Y\left(\phi_g(f)\right) \geq \sup_{P_Y \in H_0} E_Y\left(\phi_g(f)\right) = \sup_{P_Y \in H_0} E_Y\left(\phi_g\right)$$

which shows that both tests have the same size for their respective hypotheses. Similarly it follows that both have the same type II error.

We now prove part (ii). Let ϕ_g^* be a level α test with size β_0 that minimizes the type II error among all level α tests ϕ_g for (2). Following part (i), $\phi_g^* \circ f$ has level α and size β_0 for testing (3). Now note that both hypotheses in (2) are contained in those in (3). Hence, the type II error of some level α test of (3) cannot be strictly below β_0 which shows that the minimal type II error of (3) is equal to that of (2). ■

The random transformation under $c = 1$ and independent samples has appeared independently three times in the literature. Cucconi (1968) used it to create a nonparametric version of the probability ratio test. It was used in statistical decision

[1] 'Nonparametric' means that the set of underlying distributions is infinitely dimensional. Here this is the case if and only if $\mathcal{Y}_j$ is an infinite set for some $j = 1, .., m$.

theory by Gupta and Hande (1992) to design a selection procedure. Schlag (2003) used it to solve decision making under minimax regret when facing a two-armed bandit.

Given a single sample, the random transformation with $c = 1$ maximizes variance among all transformation that leave the mean unchanged. In the worst case, variance increases from 0 to 1/4, that is from the minimal to the maximal possible value of variance as $P_Y \in \Delta[0,1]$. However, once $c > 1$ then by appropriate choice of Ψ the increase in variance will be smaller. For instance, if $c = 2$ and $\psi_1 = 1/2$ then the increase in variance due to the random transformation is at most 1/16. The disadvantage of increasing c is that it then tends to be harder to find an exact test.

3 Applications

In the following we show how to use the above to construct distribution-free hypothesis tests. Given Theorem 1 it will be sufficient to present exact randomized tests. We will do this each time for a null hypothesis involving an inequality, analogous tests for the opposite inequality and for two-sided equitailed tests involving an equality are easily constructed. The tests of equality will span the entire parameter space so that one can then construct lower confidence bounds and confidence intervals.

We will mention whenever the proposed randomized test minimizes type II errors. This property is insightful as it means that a nonrandomized test derived from this test cannot be improved in terms of the bound given in (1). Similarly this means that the upper bound on inaccuracy of the associated confidence interval cannot be improved given the presented methodology for constructing nonrandomized tests and bounding type II errors.

In the first three subsections we will consider random variables and random vectors with components that generate outcomes that belong to a known bounded set, normalized to $[0,1]$. Note that this normalization has to be based on the knowledge about the possible outcomes and cannot be based on the largest outcome observed in the data. Of course one may decide to investigate a random variable conditional on it realizing an outcome in some given bounded set. As long as this given bounded set can be justified without referring to the data to be analyzed this approach is valid

too.

3.1 Mean and Higher Moments Given a Single Sample

Consider a random variable Y that can realize outcomes in $\mathcal{Y}$ with $\{0,1\} \subset \mathcal{Y} \subseteq [0,1]$ and $|\mathcal{Y}| > 2$. We wish to make inference in terms of the mean of Y based on a sample of n independent realizations of Y. Specifically, we wish to test

$$H_0 : EY \leq \mu_0,\ H_1 : EY > \mu_0 \tag{4}$$

for some $\mu_0 \in (0,1)$.[2] Following Theorems 2 and (1) we only have to determine c and $\Psi = \{\psi_l\}_{l=0}^{c}$ and then to specify a randomized exact test ϕ_g for outcomes belonging to Ψ. One could choose $c = 1$ so that $\Psi = \{0,1\}$ and then choose the randomized binomial test, denoted by ϕ_{g1}.[3] Since ϕ_{g1} is both unbiased and UMP the combined test $\phi_{g1} \circ f$ is unbiased and minimizes the type II error given $H_1' : EY > \mu_1$ among all tests ϕ for $Y \in \Delta\mathcal{Y}$ for all $\mu_1 > \mu_0$.

However we can do better by choosing $c = 2$ and $\psi_1 = \mu_0$. Note that $EY \leq \mu_0$ holds if and only if either $\Pr(Y = 1 | Y \in \{0,1\}) \leq \mu_0$ or $\Pr(Y = \mu_0) = 1$. Thus it is enough to test $\hat{H}_0 : \Pr(Y = 1 | Y \in \{0,1\}) \leq \mu_0$ which we will do by using the randomized binomial test. The resulting test for $Y \in \Delta\{0, \mu_0, 1\}$ denoted by ϕ_{g2} will be called the *'A' test* (with $\psi_1 = \mu_0$), the combination $\phi_{g2} \circ f$ will be called the *transformed 'A' test.*

Proposition 1 *The 'A' test is uniformly most powerful among all unbiased tests (UMPU) given* $P_Y \in \Delta\{0, \mu_0, 1\}$.[4] *The transformed 'A' test is unbiased for* $P_Y \in$

[2] Note that there is no need to consider the cases $\mu_0 \in \{0,1\}$ as here there are simple nonrandomized exact tests. For instance, if $\mu_0 = 0$ then reject the null hypothesis if at least one realization of Y is strictly greater than 0.

[3] For completeness we specify the randomized binomial test. Assume that $Y = j$ occured a_j times in the sample, $j = 0, 1$. Let

$$f_{a,b} = \sum_{k=a}^{a+b} \binom{n}{k} \mu_0^k (1 - \mu_0)^{n-k}.$$

Then reject the null hypothesis with probability one if $f_{a_1,a_0} \leq \alpha$, reject it with probability $(\alpha - f_{a_1-1,a_0+1}) / (f_{a_1,a_0} - f_{a_1-1,a_0+1})$ if $f_{a_1-1,a_0+1} < \alpha < f_{a_1,a_0}$ and do not reject otherwise.

[4] We avoid the terminology 'UMP unbiased test' as this does not clarify whether the test is UMP and unbiased (such as the randomized binomial test) or whether it is only UMP among the unbiased tests (such as the test of Tocher, 1950).

$\Delta[0,1]$ *and is uniformly more powerful than* $\phi_{g1} \circ f$ *and hence minimizes the type II error for* $H_1 : EY \geq \mu_1$ *among all tests* ϕ *for* $P_Y \in \Delta\mathcal{Y}$ *for all* $\mu_1 > \mu_0$. *Its type II error is equal to that of the randomized binomial test* ϕ_{g1}.

Proof. We show that the 'A' test is UMPU given $Y \in \Delta\{0, \mu_0, 1\}$ by establishing a connection to the randomized version of McNemar's test for matched pairs. In fact the two tests are identical provided one identifies $Y = 1$ with $Y = (0,1)$, 0 with $(1,0)$ and μ_0 with $(0,0)$ and identifies $H_0 : EY \leq \mu_0$ with $H_0 : EY_2 - EY_1 \leq 0$. The fact that the randomized version of McNemar's test is UMPU for $Y \in \Delta\{0,1\}^2$ (Lehmann, 1959) shows that the 'A' test is exact and UMPU for $Y \in \Delta\{0, \mu_0, 1\}$. In particular, this means that ϕ_{g2} is uniformly more powerful than $\phi_{g1} \circ f$ for all $Y \in \Delta\{0, \mu_0, 1\}$. Hence, $\phi_{g2} \circ f$ is uniformly more powerful than $\phi_{g1} \circ f$ for all $Y \in \Delta\mathcal{Y}$. Moreover, the type II error of $\phi_{g2} \circ f$ is equal to that of $\phi_{g1} \circ f$ as $\phi_{g1} \circ f$ attains the minimal type II error. This then means that the type II error of $\phi_{g2} \circ f$ is equal to that of the randomized binomial test ϕ_{g1}. ■

Analogously we can construct tests for $H_0 : EY \geq \mu_0$ and thus also exact two-sided equitailed tests of $H_0 : EY = \mu_0$ and thereby also exact equitailed confidence intervals (CI) for EY.[5]

Consider the nonrandomized test that is based on the 'A' test. We measure its performance in terms of the *inaccuracy* of the associated (family of) equitailed confidence intervals where inaccuracy is defined by the maximal expected width over all $P_Y \in \Delta[0,1]$.[6] The inaccuracy of this test will be compared to the lower bound on inaccuracy across all exact confidence intervals, this lower bound will be called *unavoidable inaccuracy*. Following Pratt (1961), inaccuracy can be derived in terms of an integral over the type II error.[7] Following Proposition 1, unavoidable inaccuracy

[5] A $100(1-\alpha)\%$ confidence region is given by the set of all μ_0 such that $H_0 : EY = \mu_0$ cannot be rejected for the given data. This confidence region is in fact a confidence interval as one can easily verify that the one-sided tests are nested in the following sense: if $H_0 : EY \leq (\geq) \mu_0$ is rejected then $H_0 : EY \leq (\geq) \mu_1$ is rejected for all $\mu_1 < (>) \mu_0$.

[6] Similarly one could measure performance in terms of type II error for a specific alternative hypthesis.

[7] Let $[L, U]$ be the $100 * (1-\alpha)\%$ confidence interval for EY associated to the nonrandomized equitailed test derived from the 'A' test. Let ϕ_{μ_0} be the randomized binomial test with level $\theta\alpha/2$. Let Y_p be such that $P_{Y_p} \in \Delta\{0,1\}$ and $EY_p = p$. Then

is attained by the randomized CI associated to the randomized binomial test. We use (1) to derive an upper bound on the inaccuracy of the nonrandomized test based on the 'A' test (using $\theta = 0.2$), it turns out numerically that this value is attained when $\Pr(Y = 0) = \Pr(Y = 1) = 1/2$. The ratio of unavoidable inaccuracy to this upper bound on inaccuracy is called the *relative efficiency* of our test. Note that the value of relative efficiency not only depends on the properties of the test but also on the tightness of the inequality (1).

Table 1: Inaccuracy of 95% Equitailed Confidence Intervals

n	20	30	40	50	60
Unavoidable Inaccuracy	0.41	0.35	0.3	0.27	0.25
Upper Bound on Inaccuracy of the Nonrandomized CI on the 'A' Test	0.59	0.5	0.44	0.4	0.37
Relative Efficiency	69%	70%	68%	68%	68%

Given the properties of the type II error of the 'A' test, smaller values of relative efficiency are not possible using our methodology (basing calculations on (1 and using $\theta = 0.2$).

We can similarly construct tests relating to higher moments. For instance to test the null hypothesis that $E\left(Y^k\right) \leq \gamma$ replace observation y_i by $(y_i)^k$ for all $i = 1, .., n$ and then proceed as above when testing means.

Exact tests for the mean of a single sample have been previously created by Bickel et al. (1989), Fishman (1991), Romano and Wolf (2000) and by Diouf and Dufour (2006), none has found much attention in applications. The test by Bickel et al. (1989) and Fishman (1991) is simple as it relies on the Hoeffding bound (Hoeffding, 1963). An upper bound on its type II error is easily derived using the Hoeffding bound but it is substantially worse than that of our test.[8] Neither of the other two papers

$$E_Y(U - L) \leq \int_0^1 \min\left\{1, \frac{1}{1-\theta} E_{Y_x}(1 - \phi_x)\right\} dx$$

which can then be used to derive inaccuracy $\sup_Y E_Y(U - L)$.

[8] The relative efficiency of the test of Bickel et al. (1989) and Fishman (1991) for the values of n given in Table 1 is not more than 56%.

provide (finite sample) bounds on the type II error, possibly due to the intricate nature of the underlying tests.

3.2 Comparing Means

Consider two random variables Y_1 and Y_2 with respective outcome spaces $\mathcal{Y}_i$ that satisfy $\{0,1\} \subset \mathcal{Y}_i \subseteq [0,1]$, $|\mathcal{Y}_i| \geq 2$ and $|\mathcal{Y}_1| + |\mathcal{Y}_2| > 4$. We wish to test

$$H_0 : EY_2 - EY_1 \leq d,\ H_1 : EY_2 - EY_1 > d. \tag{5}$$

If $d > 0$ then this is called a *test of superiority* of Y_2 over Y_1. If $d < 0$ then it is called a *test of noninferiority* of Y_2 over Y_1. Tests for the above can then be combined with their counterparts for $H_0 : EY_2 - EY_1 \geq d$ to yield nonrandomized two-sided tests of

$$H_0 : EY_2 - EY_1 = d,\ H_1 : EY_2 - EY_1 \neq d$$

as well as *tests of equivalence* for

$$H_0 : |EY_2 - EY_1| \geq d,\ H_1 : |EY_2 - EY_1| < d.$$

3.2.1 Matched Pairs

Consider inference based on a sample of matched pairs. To minimize variance, transform the two observations within the pair jointly. We propose a randomized exact test to be inserted in the first step. Consider $d = 0$. The obvious choice is to set $c = 1$ and to use the UMPU test for $Y \in \Delta\{0,1\}^2$ which is the randomized version of McNemar's (1947) test due to Lehmann (1959). In fact, Schlag (2008) has shown that it minimizes the type II error for $H_1 : EY_2 - EY_1 = d_1$ among all tests of level α. In this sense the transformed UMPU test is most powerful for $d = 0$ whenever the difference between the two means is the only parameter of interest.

Consider now the case where $d \neq 0$. Here we apply our results for testing the mean of a single sample and let $Z = (1 + Y_2 - Y_1)/2$. Then the null hypothesis in (5) is identical to $H_0 : EZ \leq \frac{1}{2}(1 + d)$ and our proposal is to apply the transformed 'A' test to test this null hypothesis. We find that there is no loss in terms of inference of treating matched pairs as a single sample whenever the difference between the two means is the only parameter of interest.

Proposition 2 *Interpreting the matched sample as a single sample of the random variable* $Z = (1 + Y_2 - Y_1)/2$ *and then testing* $H_0 : EZ \leq \frac{1}{2}(1 + d)$ *using the transformed 'A' test generates a randomized test that is unbiased and minimizes the type II error given* $H_1 : EY_2 - EY_1 > d_1 > d$. *When* $d = 0$ *then this test is identical to the transformed UMPU test.*

Proof. The fact that the 'A' test is unbiased for (4) implies that the proposed test is unbiased for (5). Consider $H_1 : EY_2 - EY_1 = d_1$ for some $d_1 > d$. Following Proposition 1 the type II error of the transformed 'A' test is attained when $Z \in \Delta\{0, 1\}$ which means that $Y \in \Delta\{(1, 0), (0, 1)\}$. Within this class of distributions the transformed 'A' test is uniformly most powerful. Hence it attains the minimal type II error. ■

Given Proposition 2 the relative efficiency of the CI derived from our proposed test is equal to that of the CI proposed for a single sample. This is because the factor of 2 that enters the integration when moving from matched pairs to a single sample appears both when deriving the lower bound on inaccuracy and the upper bound on the inaccuracy of our specific test.

3.2.2 Two Independent Samples

Now consider inference based on two independent samples of possibly different sizes. Consider first the case where $d = 0$. The obvious choice is to set $c = 1$ and to use as randomized test for the binary valued data the randomized version of Fisher's (1935b) exact test that is due to Tocher (1950). Note that this test is UMPU. Of course one can apply any other test for comparing means of Bernoulli distributions. When the samples are approximately balanced the Z test with pooled variance (see Suissa and Shuster, 1985) is a valuable alternative. Note that the selection of which test to use can be made based on the sample sizes but not on the data itself. In Section 5 we give some help for selecting the test used in step 1.

For the case where $d \neq 0$ one can again set $c = 1$ and use any exact test for the binary valued case. However as the tests used in the literature for this shifted null hypothesis are designed to be nonrandomized and hence intricate (see Röhmel, 2005 for an overview) we suggest a simpler but randomized test for the binary valued case. The idea is to reverse the roles of the null and of the alternative hypothesis in the

test of Tocher (1950) and to adjust the size appropriately. Specifically, let $\phi^u_{\alpha'}$ be the UMPU test with size α' for testing $H'_0 : EY_2 \geq EY_1$ and choose α'' such the maximal probability of not rejecting H'_0 using $\phi^u_{\alpha''}$ among all $Y \in H_0 = \{EY_2 - EY_1 \leq d\}$ is equal to α. Our proposal is then to use $\phi^d_\alpha := 1 - \phi^u_{\alpha''}$ as the randomized test in step 1. We call ϕ^d_α the *reversion* of $\phi^u_{\alpha'}$. By construction, ϕ^d_α is a randomized test with size α for testing (5) when Y_1, Y_2 are Bernoulli distributed.

We provide some more details and intuition. Assume $d = 0$. Then it is easily verified that ϕ^d_α is identical to the UMPU test ϕ'^u_α for testing $H''_0 : EY_2 \leq EY_1$. In other words, $\phi'^u_\alpha \equiv 1 - \phi^u_{1-\alpha}$. Assume $d < 0$. Here we are looking for α'' such that the type II error of $\phi^u_{\alpha''}$ given $\{EY_2 - EY_1 \leq d\}$ is equal to α, or equivalently that the minimal probability of rejection is equal to $1 - \alpha$. Compared to the case of $d = 0$ this can only be achieved by lowering α'', hence $\alpha'' < 1 - \alpha$. In all numerical examples we have found that the maximal type II error is attained on the off-diagonal where $EY_1 + EY_2 = 1$. Now consider $d > 0$. In this case we have to move the size in the opposite direction and find $\alpha'' > 1 - \alpha$. Here we have found in the numerical examples that the maximal type II is attained on the border where $EY_2 = 0$ and $EY_1 = -d$. It namely turns out that the level sets of $\phi^u_{\alpha'}$ are bent away from the straight-line defined by $\{\phi^u_{\alpha'} = \alpha'\}$.

Under this reversion technique confidence bounds and intervals are particularly simple to derive. For instance, assume that we wish to find a $100\,(1-\alpha)\,\%$ lower confidence bound L on $EY_2 - EY_1$. So we are searching for the largest value of L such that $\phi^d_\alpha|_\theta$ recommends a rejection of the data for all $d < L$. This can be done directly as follows. First find α'' such that the transformed UMPU test $\phi^u_{\alpha''} \circ f$ rejects $H'_0 : EY_2 \geq EY_1$ with probability $1 - \theta$ and then use the power function of $\phi^u_{\alpha''}$ to find L such that the maximal probability of not rejecting H'_0 using $\phi^u_{\alpha''}$ is at most $\theta\alpha$ when $EY_2 - EY_1 \leq L$. Then L is the desired lower confidence bound.

In the following we evaluate the inaccuracy of this test in balanced samples and compare it to the two approximations of the unavoidable inaccuracy. The upper bound on unavoidable inaccuracy is given by evaluating the inaccuracy of the transformed UMPU test. This is not equal to the lower bound due to the additional constraints imposed by unbiasedness. The lower bound on unavoidable inaccuracy is derived in Schlag (2008). When $n_1 = n_2 = n$ then it is attained by the most powerful

test when facing $\Delta\{(1,0),(0,1)\}$ and hence given Proposition 1 by the transformed 'A' test based on $2n$ matched pairs. The fact that this lower bound comes very close to the inaccuracy of the transformed UMPU test gives us a good understanding of unavoidable inaccuracy. Schlag (2008) shows more, namely that this lower bound on inaccuracy is also valid within the larger set of all sequential tests that sample at most $2n$ observations. The relative efficiency of our test presented below hence also refers to the class of all sequential tests.

Table 2: Inaccuracy of 95% Equitailed Confidence Intervals

$n_1 = n_2 =$	20	30	40
Lower Bound on Inaccuracy	0.606	0.5	0.436
Inaccuracy of Transformed Reversed UMPU Test	0.612	0.502	0.442
Upper Bound on Inaccuracy of the Nonrandomized CI Based on the UMPU Test	0.89	0.74	0.65
Relative Efficiency	68%	68%	67%

We do not know of any other exact test for comparing means when there are more than two possible outcomes. Recall that both the Mann-Whitney-Wilcoxon test (Wilcoxon, 1945, Mann and Whitney, 1947) and the Wilcoxon rank sum test (Wilcoxon, 1945) are permutation tests and hence exact for the null hypothesis $H_0 : P_{Y_1} \equiv P_{Y_2}$. However it is well known and easily verified by example that they are not exact for testing equality of means.[9]

3.3 Variance and Covariance

A transformation of Walsh (1962) allows us to reduce tests in terms of variance and covariance to tests in terms of means. The idea is to pair data to create a new sample that has mean equal to the variance or covariance of the original sample. We have

[9] Consider

	state A	state B
Y_1	0	1
Y_2	$1-\varepsilon$	$1-\varepsilon$
prob.	ε	$1-\varepsilon$

for $\varepsilon > 0$ with ε small.

not seen this technique used in practice. Apart from the test of Romano and Wolf (2002) for variance we are not aware of any other exact tests for these hypotheses.

3.3.1 Variance

Consider a random variable Y that can realize outcomes in $\mathcal{Y}$ with $\{0,1\} \subset \mathcal{Y} \subseteq [0,1]$ and $|\mathcal{Y}| > 2$. We wish to make inference in terms of the variance of Y based on a sample of n independent realizations of Y. Specifically we wish to test

$$H_0 : VarY \leq \sigma_0^2,\ H_1 : VarY > \sigma_0^2$$

for some $\sigma_0^2 > 0$. The algorithm leading to an exact randomized test is as follows. Randomly combine the sample $(y_j)_{j=1}^n$ into $\lfloor n/2 \rfloor$ pairs. For the ith pair $\{y_k, y_l\}$ compute $z_i = 1/2 + (y_k - y_l)^2/2$. It is then as if we have $\lfloor n/2 \rfloor$ independent observations $(z_i)_{i=1}^{\lfloor n/2 \rfloor}$ of a random variable Z that generates outcomes in $[0,1]$ and where $EZ = 1/2 + VarY$. Then apply the transformed 'A' test with $\psi_1 = 1/2 + \sigma_0^2$ to test the null hypothesis that $EZ \leq 1/2 + \sigma_0^2$.

Note that the sample size is cut in half when combining the observations into pairs. This reduces the performance of the test. Its type II error can be bounded using (1), inserting the power of the randomized binomial test given $\lfloor n/2 \rfloor$ independent observations. It should come at no surprise that inference in terms of variance is more difficult than when concerned with means as variance incorporates a difference between second and first moments.

Romano and Wolf (2002) present an exact test for the above pair of hypotheses. In contrast to our tests, they do not require known bounds for $\mathcal{Y}$ when determining the lower confidence bound. However they do not provide a finite sample bound on the type II error of their test.

3.3.2 Covariance and Correlation

Next consider a random vector (Y_1, Y_2) with respective outcome spaces $\mathcal{Y}_i$ that satisfy $\{0,1\} \subset \mathcal{Y}_i \subseteq [0,1]$ and $|\mathcal{Y}_1| + |\mathcal{Y}_2| > 4$. We wish to test

$$H_0 : Cov(Y_1, Y_2) \leq \gamma,\ H_1 : Cov(Y_1, Y_2) > \gamma \tag{6}$$

based on a sample of n matched pairs realized from (Y_1, Y_2). We propose a test using a very similar algorithm to the one above. Randomly combine the sample into $\lfloor n/2 \rfloor$ pairs. For the ith pair $\{(y_{1k}, y_{2k}), (y_{1l}, y_{2l})\}$ compute $z_i = 1/2 + (y_{1k} - y_{1l})(y_{2k} - y_{2l})/2$. Then apply the transformed 'A' test with $\psi_1 = 1/2 + \gamma$ to test $H_0 : EZ \leq 1/2 + \gamma$ based on $(z_i)_{i=1}^{\lfloor n/2 \rfloor}$.

One may also be interested in testing whether Y_1 and Y_2 are correlated, specifically we wish to test

$$H_0 : \rho(Y_1, Y_2) \leq 0, \; H_1 : \rho(Y_1, Y_2) > 0 \tag{7}$$

where $\rho(Y_1, Y_2) = Cov(Y_1, Y_2) / \sqrt{VarY_1 \cdot VarY_2}$ if $VarY_1 \cdot VarY_2 > 0$ and $\rho(Y_1, Y_2) = 0$ otherwise. Since $\{\rho(Y_1, Y_2) \leq 0\} \subset \{Cov(Y_1, Y_2) = 0\}$ we can directly apply our test of (6).

Note that the classic distribution-free test used for analyzing correlation is the Spearman rank correlation test (Spearman, 1904). It is a permutation test and as such is an exact test for the null hypothesis that Y_1 and Y_2 are independent. However it is easy to see by example that it is generally not an exact test for the null hypothesis that Y_1 and Y_2 are uncorrelated or negatively correlated.[10]

3.3.3 Comparing Variances

Consider the same setting as above but now we wish to test

$$H_0 : VarY_2 - VarY_1 \leq d, \; H_1 : VarY_2 - VarY_1 > d.$$

Here we suggest to create exact randomized tests by first applying the transformation of Walsh and then continuing with the algorithms for comparing means. More specifically, for the case of independent samples, match within each sample the n_j observations of Y_j into pairs (y_k, y_l) and compute $1/2 + (y_l - y_k)^2/2$. This results in two independent sample of size $\lfloor n_1/2 \rfloor$ and $\lfloor n_2/2 \rfloor$, the values belong to $[0, 1]$ and their expectations are equal to $VarY_j$. Then can continue with the algorithm for

[10] Consider

$Y_1 \backslash Y_2$	$1/2 - \varepsilon$	$1/2$	1
0	$(1 - \lambda)/2$	0	$\lambda/2$
1	0	$1/2$	0

where $\lambda = \varepsilon / (1/2 + \varepsilon)$ and $\varepsilon > 0$ with ε small.

comparing means. For the case of matched pairs, first pair these n matched pairs, replace the i'th pair $((y_{1k}, y_{2k}), (y_{1l}, y_{2l}))$ by $z_{i'}$ where $z_{ji'} = 1/2 + (y_{jl} - y_{jk})^2/2$ and then continue to analyze the mean of the matched pairs $(z_{1i'}, z_{2i'})_{i=1}^{\lfloor n/2 \rfloor}$.

3.4 Ordinal Comparisons

In the next three subsections we present tests that do not require that outcomes belong to a known bounded set. In fact, they can also be applied to *ordinal data*.[11]

3.4.1 Stochastic Inequality for Two Independent Samples

Consider two random variables Y_1 and Y_2. Inference will be based on two independent samples consisting of n_j realizations of Y_j, $n_j > 0$, $j = 1, 2$. A measure for comparing outcomes realized of by these two random variables is the *stochastic difference* (Vargha and Delaney, 2000) between Y_2 and Y_1 defined by

$$\delta = \Pr(Y_2 > Y_1) - \Pr(Y_2 < Y_1)$$

where $\delta \in [-1, 1]$. While $\delta = 0$ holds if Y_1 and Y_2 are identically distributed, the converse is generally not true. However these two statements are equivalent if Y_1 and Y_2 are binary valued, in which case δ is equal to the difference between the two means.

We wish to test the following pair of hypotheses, also known as a *stochastic inequality*:

$$H_0 : \Pr(Y_2 > Y_1) \leq \Pr(Y_2 < Y_1), \; H_1 : \Pr(Y_2 > Y_1) > \Pr(Y_2 < Y_1) \tag{8}$$

which is equivalent to testing $H_0 : \delta \leq 0$ against $H_1 : \delta > 0$. Upon rejection one may claim to have significant evidence that "Y_2 tends to larger values than Y_1" (Brunner and Munzel, 2000). More generally we present tests of

$$H_0 : \delta \leq \delta_0, \; H_1 : \delta > \delta_0 \tag{9}$$

for $\delta_0 \in (-1, 1)$.[12] These tests can then be used to construct tests of $H_0 : \delta = \delta_0$

[11] For the case of ordinal data, let Y_1 and Y_2 be associated to distributions across a set of outcomes. Let $\succsim$ be a complete and transitive ordering within this set of outcome. For this alternative, replace $\geq, >, =$ with $\succsim, \succ, \sim$ respectively below.

[12] Simple exact nonrandomized tests can be constructed directly for the extreme cases where $\delta_0 \in \{-1, 1\}$.

against $H_1 : \delta \neq \delta_0$. The special case where $\delta_0 = 0$ is referred to as a *stochastic equality* (Vargha and Delaney, 1998).

We describe a randomized test for (9) in terms of the following algorithm. Let $n = \min\{n_1, n_2\}$. First randomly match n observations of Y_1 to n observations of Y_2. Ignore the $|n_1 - n_2|$ unmatched observations of the larger sample. For each $i = 1, .., n$ replace the ith matched pair $\left(y_1^k, y_2^l\right)$ with 0, 1/2 or 1 depending on whether $y_2^l - y_1^k < 0$, $= 0$ or > 0. This yields n independent observations of a random variable $Z \in \Delta\{0, 1/2, 1\}$ where $\Pr(Z = 0) = \Pr(Y_2 < Y_1)$, $\Pr(Z = 1/2) = \Pr(Y_2 = Y_1)$, $\Pr(Z = 1) = \Pr(Y_2 > Y_1)$ and $EZ = (1 + \delta)/2$. Apply the transformed 'A' test with $\psi_1 = (1 + \delta_0)/2$ to test the null hypothesis that $EZ \leq (1 + \delta_0)/2$.

The above test is randomized and unbiased and minimizes the type II error among all tests will level α for $H_0 : \delta \leq \delta_0$ against $H_1 : \delta \geq \delta_1$ for each $\delta_1 > \delta_0$. This follows from the properties of the transformed 'A' test.

To connect to the literature, note that any exact test of (8) will also be an exact test of

$$H_0 : EY_2 \leq EY_1, \; H_1 : EY_2 > EY_1 \tag{10}$$

provided Y_1 and Y_2 are distributed symmetrically (e.g. normally distributed). The objective to find a test for (10) when Y_1 and Y_2 are normally distributed is called the *Behrens-Fisher problem* (Behrens, 1929, Fisher, 1935a), for an overview of solutions see Weerahandi (1994). Given this connection Brunner and Munzel (2000) describe (8) as a nonparametric Behrens-Fisher problem. So our test of (8) also solves of the Behrens-Fisher problem.

The search for tests of the stochastic inequality seems to originate in Cliff (1993). For recent treatments including tests of stochastic equality see Vargha and Delaney (1998, 2000), Brunner and Munzel (2000), Borges del Rosal et al. (2003) and Reizigel et al. (2005). Our test seems to be the first exact test given independent samples. Note that one can easily adapt the Z test to construct tests for the simpler case of matched pairs.

3.4.2 Association Related to Kendall's Tau

Consider a random vector $(Y, Z) \in \mathbb{R}^2$ and an independent sample of n matched pairs drawn from this random vector. We wish to make inference about the association

between the two random variables Y and Z. Two pairs of observations $\left(y^k, z^k\right)$ and $\left(y^l, z^l\right)$ are called a *concordant (discordant) pair* if $\left(y^l - y^k\right)\left(z^l - z^k\right) > 0$ (< 0). We measure association by the *concordant difference* τ', defined by the difference between the probability that two random observations are concordant and discordant. So

$$\tau'(Y, Z) = \Pr\left(\left(Y_2 - Y_1\right)\left(Z_2 - Z_1\right) > 0\right) - \Pr\left(\left(Y_2 - Y_1\right)\left(Z_2 - Z_1\right) < 0\right),$$

where Y_1, Y_2 and Z_1, Z_2 are independent copies of Y and Z. Then $\tau' \in [-1, 1]$. While $\tau' = 0$ if Y and Z are independent the converse is generally not true. However the converse is true if Y and Z can only realize two different outcomes. In this special case τ' is equal to the covariance between Y and Z.

An unbiased and symmetric estimate of τ' involves recording the difference between the percentage of concordant and discordant pairs among all possible pairings. This estimator is called *Kendall's tau* (Kendall, 1938).[13]

We wish to design an exact test for

$$H_0 : \tau' \leq \tau'_0,\ H_1 : \tau' > \tau'_0$$

when $\tau'_0 \in (-1, 1)$.[14] When rejecting the case where $\tau'_0 = 0$ one can claim to have significant evidence that "(Y, Z) tends to generate concordant pairs."

The algorithm defining the test starts by randomly matching the vectors $\{(y^i, z^i), i = 1, .., n\}$ into pairs, ignoring the unmatched observation when n is odd. Then replace any matched pair of vectors $\left(y^k, z^k\right)$ and $\left(y^l, z^l\right)$ with the outcome $\left(1 + sign\left(\left(y^l - y^k\right)\left(z^l - z^k\right)\right)\right)/2$. Finally apply the 'A' test with $\psi_1 = 1/2$ to test the null that $EX \leq \tau'_0$ where X is the random vector underlying the transformed sample.

It seems that Fechner (1897) first investigated the concordant difference (for more on the origins see Kruskal, 1958). For recent tests that are however not exact see Kochar and Gupta (1987) and Samara and Randles (1988).

3.4.3 Spread Related to Q_n

Consider a random variable Y and a single sample of n independent realizations of Y. We wish to make inference in terms of the spread of Y. Let m_q be the qth quantile of

[13] To avoid confusion, we hence denote the concordant difference by τ' and not by τ.

[14] Simple exact nonrandomized tests are constructed directly for the extreme cases where $\tau'_0 \in \{-1, 1\}$.

the distribution underlying $|Y_2 - Y_1|$ where Y_1 and Y_2 are two independent copies of Y. Then $m_q(|Y_2 - Y_1|)$ is a measure of the spread of Y. An estimator for $m_q(|Y_2 - Y_1|)$ is the qth quantile among all possible pairs of observations. While it is most intuitive to set $q = 1/2$ (Shamos, 1976, Bickel and Lehmann, 1979), it turns out that the choice of $q = 1/4$ makes the estimator, referred to as Q_n, maximally "robust" (Rousseeuw and Croux, 1993).

We wish to design a randomized test for

$$H_0 : m_q(|Y_2 - Y_1|) \leq d,\ H_1 : m_q(|Y_2 - Y_1|) > d.$$

The test calls to randomly pair the data and then to use the randomized binomial test to test the null that the proportion of pairs in which the absolute difference is below d is at least q.

3.5 Other Applications

There are many other possible applications. Our method only requires knowledge of an exact test for binary valued data. One can construct tests of Markov dependence by building on the results of Klotz (1973). Another obvious application is to testing simultaneous equality of m means, $H_0 : EY_1 = .. = EY_m$ for $Y_j \in [0, 1]$. Here one simply needs to insert an exact test of homogeneity for $2 \times m$ tables (e.g. see Mehta and Patel, 1980) into our two step construction.

4 Illustrating Examples

We illustrate some of our tests using three data examples.

4.1 Anti-Self-Dealing

The 'anti-self-dealing' indices in Djankov et al. (2008, Table III) were gathered for 72 countries to measure minority shareholder protection against self-dealing (i.e. investor expropriation) of a controlling shareholder. In Table 3 we present the average anti-self-dealing indices across different regions characterized by the origin of their law system together with the number of countries in each of these regions. Thereby, countries with civil law are subdivided into the French, German and Scandinavian

region. Indices belong to $[0, 1]$ by construction. For our analysis we assume that the indices were constructed independently across countries. 95% confidence intervals are provided in Table 3 for the means using the test in Section 3.1 and in Table 4 for the expected mean difference (test in Section 3.2.2) and stochastic difference (test in Section 3.4.1) between common law and civil law and its subregions.

Table 3: Mean Anti-Self-Dealing Indices across Regions

	Common law	Civil law	French	German	Scand.
n_i	21	51	32	14	5
sample mean	0.66	0.35	0.33	0.38	0.39
95% CI	$[0.51, 0.78]$	$[0.28, 0.42]$	$[0.24, 0.43]$	$[0.26, 0.56]$	$[0.15, 0.72]$

Table 4: Anti-Self Dealing Indices: Comparing Common Law to Other Regions

	Civil law	French	German	Scand.
sample difference	0.31	0.33	0.28	0.27
95% CI	$[0.08, 0.52]$	$[0.07, 0.56]$	$[-0.04, 0.55]$	$[-0.19, 0.65]$
estimated stochastic difference δ	0.67	0.69	0.66	0.58
95% CI of δ	$[0.27, 0.89]$	$[0.29, 0.91]$	$[0.14, 0.92]$	$[-0.27, 0.94]$

4.2 Pain Diagnosis

Consider the randomized medical experiment of Sabeti-Aschraf et al. (2005) designed to evaluate the outcome of two alternative diagnosis methods. 50 patients suffering from shoulder tendinitis received shock wave therapy. Prior to the intervention patients were randomly assigned to one of two treatments. For 25 of the patients the location of the therapy was determined manually. For the other 25 patients a computer assisted in determining the location. Before and after the intervention the level of pain of each patient was measured on the visual analog scale (VAS), a scale that takes values in $[0, 100]$. In Table 5 we present the descriptive statistics together with the 95% confidence interval for the change in VAS derived using our test for matched pairs presented in Section 3.2.1.

Table 5: Shock-Wave Therapy

Assistance	Manual	Computer
n	25	25
Average VAS before intervention	68	66
Average VAS after intervention	33	18
Estimated change	-35	-48
95% CI	$[-59, -12]$	$[-71, -21]$

We also verify that the seemingly better performance of the computer assistance (-13 in terms of mean, 0.25 in terms of stochastic difference) is not statistically significant at the 10% level using our tests, either when comparing means or in terms of stochastic inequality.

4.3 A Laboratory Experiment

Croson and Buchan (1999) conducted the following randomized double-blind laboratory experiment. Subjects were matched via computers in pairs. Both subjects received an endowment in terms of tokens, here normalized to a total of 1 unit. One of the two subjects was selected as sender to be allowed to transfer all or part of his or her endowment to the other subject. The amount transferred was tripled by the experimenter. The recipient then had to decide how many tokens to return to the sender. Thereafter the experiment ended. Notice that while the recipient did not have to return any tokens, one may expect him or her to do so in order to reward the sender for making the "investment".

To clarify the outcome let $S \in [0, 1]$ be the amount sent and let R be the amount returned. Then $R \in [0, 1 + 3 * S]$. After the experiment the sender has $1 - S + R$ tokens and the recipient has $1 + 3 * S - R$ tokens.

We wish to investigate $Cov(S, R)$ where $Cov(S, R) \in [-1, 1]$. We find marginally significant evidence (level 10%) that the covariance between amount sent and amount returned is strictly positive and report the 95% confidence interval for covariance in Table 6. We also derive Kendall's tau and present a 95% CI for the concordant difference τ'. Here we find strong statistically significant evidence (at level 1%) that sending more tokens tends to be rewarded with more being returned.

Table 6: A Laboratory Experiment

n	94
mean of S	0.67
$Cov\,(S,R)$	0.13
95% CI of $Cov\,(S,R)$	$[-0.055, 0.32]$
$\tau\,(S,R)$	0.53
95% CI of $\tau'\,(S,R)$	$[0.29, 0.72]$

5 Fine-Tuning the Threshold θ

It is natural to include the choice of the threshold θ in the design of a test or of confidence intervals, anticipating its impact on inference as measured using (1). The choice of θ is particularly transparent when interested in confidence intervals. Inaccuracy is a popular measure of performance, one then chooses θ in order to minimize its upper bound as derived in Footnote 7. For the case of a single sample we find that this fine-tuning only yields marginal improvements over choice of $\theta = 0.2$. For instance, for the values of n given in Table 1, the choice of θ lies between 0.21 and 0.23. The resulting reduction in the upper bound on inaccuracy is smaller than the rounding error.

In fact, the choice of θ is simplest when concerned with a specific pair of hypotheses. θ can then be chosen to minimize the bound given in (1). One example would be noninferiority tests where particular attention is on testing $H_0 : EY_2 - EY_1 \leq d$ against $H_1 : EY_2 \geq EY_1$ for a given value of $d < 0$.

However, in many applications, such as when testing equality of two means, one is interested in testing a particular null hypothesis without being focused on a specific alternative hypothesis. One has to then determine how to choose θ as the choice of θ will typically depend on the specific alternative hypothesis. Smaller θ tend to improve inference for alternatives that lie closer to the null hypothesis. The statistical decision theory approach would be to assign a loss to each recommendation, naturally assigning a greater loss to false negatives when the true data generating process is "further away" from the null hypothesis.

6 Conclusion and Outlook

"The race has begun?" We have demonstrated that it is possible to construct exact nonrandomized tests in rich environments and to measure their power of inference for the given sample size. The next step is to work on improving these tests. Two types of improvements immediately come to mind. One could improve the first step in our construction. For instance one could consider a finer grid for the random transformation and then try to find a test that is uniformly more powerful. The downside of a finer grid is that it is then more difficult to design an exact test. Alternatively one could try to improve the second step. The bounds used to evaluate the loss of inference when eliminating randomness are admittingly very crude. Their advantage is that the underlying proof is extremely simple. More insights are needed on how much inference can be improved by choosing $\theta \neq 0.2$.

An alternative line of future research is to consider other environments where the case of binary valued data is well understood to then extend the exact tests for binary valued data to nonparametric settings.

A downside of our method is that we have not (yet) been able to use it to construct a test for comparing medians. Of course, tests for the median, or more generally for any quantile, given a single sample are easily designed using the binomial test. The only exact test we know for comparing medians (or quantiles) of two independent samples then involves looking at the intersection of these confidence intervals, adjusting their coverage appropriately.

References

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

[2] Behrens,W. V. (1929), "Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen," *Landwirtschaftliche Jahrbuecher* **68**, 807–837.

[3] Bickel, P., Godfrey, J., Neter, J. and Clayton, H. (1989), "Hoeffding Bounds for Monetary Unit Sampling in Auditing," *International Statistical Institute, Contributed Paper, Paris Meeting.*

[4] Bickel, P. J. and Lehmann, E. L. (1979), "Descriptive Statistics for Nonparametric Models IV: Spread," in *Contributions to Statistics, Hájek Memorial Volume*, ed. J. Jurečková, Prague: Academia, 33–40.

[5] Borges del Rosal, A., San Luis, C. and Sánchez-Bruno, A. (2003), "Dominance Statistics: A Simulation Study on the d Statistic," *Quality and Quantity* **37**, 303–316.

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

[7] Cliff, N. (1993), "Dominance Statistics: Ordinal Analyses to Answer Ordinal Questions," *Psychological Bulletin* **114**, 494–509.

[8] Croson, R. and Buchan, N. (1999), "Gender and Culture: International Experimental Evidence from Trust Games," *American Economic Review Papers and Proceedings* **89**, 386–391.

[9] Cucconi, O. (1968), "Contributi all'Analisi Sequenziale nel Controllo di Accettazione per Variabili" (in Italian), *Atti dell' Ass. Italiana per il Controllo della Qualità* **6**, 171–186.

[10] Diouf, M. A. and Dufour, J.-M. (2006), "Exact Nonparametric Inference for the Mean of a Bounded Random Variable," in *American Statistical Association, Proceedings of the Business and Economic Statistics Section.*

[11] Djankov, S., La Porta, R., Lopez-de-Silanes, F. and Shleifer, A. (2008), "The Law and Economics of Self-Dealing," *Journal of Financial Economics* **88**, 430–465.

[12] Dvoretzky, A., Wald, A., and Wolfowitz, J. (1951), "Elimination of Randomization in Certain Statistical Decision Procedures and Zero-Sum Two-Person Games," *The Annals of Mathematical Statistics* **22**, 1–21.

[13] Fechner, G. T. (1897), *Kollektivmasslehre*, Leipzig: Wilhem Engelmann (published posthumously, completed and edited by G. F. Lipps).

[14] Fisher, R. A. (1935a), "The Fiducial Argument in Statistical Inference," *Ann. Eugenics* **6**, 391–398

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

[16] Fishman, G. S. (1991), "Confidence Intervals for the Mean in the Bounded Case," *Statistics and Probability Letters* **12**, 223–227.

[17] Gupta, S. S. and Hande, S. N. (1992), "On Some Nonparametric Selection Procedures," *Nonparametric Statistics and Related Topics*, A.K.Md.E. Saleh (Editor), Amsterdam: Elsevier, 33–49.

[18] Hoeffding, W. (1963), "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association* **58**, 13–30.

[19] Kendall, M. G. (1938), "A New Measure of Rank Correlation," *Biometrika* **30**, 81–93.

[20] Klotz, J. (1973), "Statistical Inference in Bernoulli Trials with Dependence," *The Annals of Statistics* **1**, 373-379

[21] Kochar, S. C, and Gupta, R. P. (1987), "Competitors of the Kendall-Tau Test for Testing Independence Against Positive Quadrant Dependence," *Biometrika* **74**, 664–666.

[22] Kruskal, W. H. (1958), "Ordinal Measures of Association," *JASA* **53**, 814–861.

[23] Lehmann, E. L. (1959), *Testing Statistical Hypotheses*. New York: Wiley.

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

[25] Lehmann, E. L. and Romano, J. P. (2005), *Testing Statistical Hypotheses*. New York: Springer.

[26] Mann, H. B., and Whitney, D. R. (1947), "On a Test whether one of two random variables is stochastically larger than the other," *Annal. Math. Statistics* **18**, 50–60.

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

[28] Mehta, C. R. and Patel, N. R. (1980), "A Network Algorithm for the Exact Treatment of the $2 \times k$ Contingency Table," Communications in Statistics - Simulation and Computation **9**, 649-664.

[29] Pratt, J. W. (1961), "Length of Confidence Intervals," *Journal of the American Statistical Association* **56**, 549–567.

[30] Reiczigel, J, Zakariás, I. and Rózsa, L. (2005), "A Bootstrap Test of Stochastic Equality of Two Populations," *The American Statistician* **59**, 156–161.

[31] Röhmel, J. (2005), "Problems with Existing Procedures to Calculate Exact Unconditional P-Values for Non-Inferiority/Superiority and Confidence Intervals for Two Binomials and How to Resolve Them," *Biometrical Journal* **47**, 37–47.

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

[33] Romano. J. P. and Wolf, M. (2002) "Explicit Nonparametric Confidence Intervals for the Variance with Guaranteed Coverage," *Communication in Statistics - Theory and Methods* **31**, 1231–1250.

[34] Rousseeuw, P. T. and Croux, C. (1993), "Alternatives to the Median Absolute Deviation," *JASA* **88**, 1273–1283.

[35] Sabeti-Aschraf, M., Dorotka, R., Goll, A. and K. Trieb, (2005), "Extracorporeal Shock Wave Therapy in the Treatment of Calcific Tendinitis of the Rotator Cuff," *Amer. J. Sports. Med.* **33(9)**, 1–4.

[36] Samara B, Randles R. H. (1988), "A Test for Correlation Based on Kendall's tau," *Communications in Statistics - Theory and Methods* **17**, 3191–3205.

[37] Schlag, K. H. (2003). *How to Minimize Maximum Regret in Repeated Decision-Making*, Unpublished Manuscript, European University Institute.

[38] Schlag, K. H. (2008), *Bringing Game Theory to Hypothesis Testing: Establishing Finite Sample Bounds on Inference*, Universitat Pompeu Fabra Working Paper No. 1099.

[39] Shamos, M. I. (1976), "Geometry and Statistics: Problems at the Interface," in *New Directions and Recent Results in Algorithms and Complexity*, ed. J.F. Traub, New York: Academic Press, 251–280.

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

[41] Suissa, S. and Shuster, J. J. (1985), "Exact Unconditional Sample Sizes for the 2×2 Binomial Trial," *J. Roy. Stat. Soc. (A)* **148**, 317–327.

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

[43] Vargha, A., and Delaney, H. D. (1998), "The Kruskal-Wallis Test and Stochastic Homogeneity," *Journal of Educational and Behavioral Statistics* **23**, 170–192.

[44] Vargha, A. and Delaney, H. (2000). "A Critique and Improvement of the CL Common Language Effect Size Statistics of McGraw and Wong," Journal of Educational and Behavioral Statistics **25**, 101–132.

[45] Walsh, J. E. (1962), *Handbook of Nonparametric Statistics, Investigation of Randomness, Moments, Percentiles, and Distributions*, Princeton: D. van Nostrand Company Inc..

[46] Weerahandi, S. (1994) *Exact Statistical Methods for Data Analysis*, New York: Springer.

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

7 Appendix: Summary Table of Novel Exact Tests and Confidence Intervals

Table 7

Data	DGP	Parameter or Test	Section
Single sample	more than 2 outcomes	$m_q\left(\lvert Y_1 - Y_2\rvert\right)$	3.4.3
"	bounded DGP*	CI for mean and higher moments	3.1
"	bounded DGP*	CI for variance	3.3.1
Two independent samples	more than 2 outcomes	test for stochastic inequality (with CI)	3.4.1
"	bounded DGP*	CI for difference in means	3.2.2
"	bounded DGP*	CI for difference in variances	3.3.3
Matched pairs	more than 2 outcomes	test for association (with CI)	3.4.2
"	bounded DGP*	CI for difference in means and higher moments	3.2.1
"	bounded DGP*	CI for difference in variances	3.3.3
"	bounded DGP*	CI for covariance	3.3.2
"	bounded DGP*	Test for correlation	3.3.2

* There is some known bounded set that will contain any outcome that can be generated.