Journal of Business & Economic Statistics

Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: http://amstat.tandfonline.com/loi/ubes20

Permutation Tests for Comparing Inequality Measures

Jean-Marie Dufour, Emmanuel Flachaire & Lynda Khalaf

To cite this article: Jean-Marie Dufour, Emmanuel Flachaire & Lynda Khalaf (2017): Permutation Tests for Comparing Inequality Measures, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2017.1371027

To link to this article: https://doi.org/10.1080/07350015.2017.1371027

Accepted author version posted online: 24 Aug 2017.
Published online: 14 May 2018.

Permutation Tests for Comparing Inequality Measures

Jean-Marie Dufour
Department of Economics, McGill University, Montréal, Québec H3A 2T7, Canada, and Centre Interuniversitaire de Recherche en Analyse des Organisations (CIRANO) and Centre interuniversitaire de recherche en économie quantitative (CIREQ), Montreal, QC, Canada (*jean-marie.dufour@mcgill.ca*)

Emmanuel Flachaire
Aix-Marseille University, AMSE, CNRS, EHESS, Centrale Marseille, 13002, Marseille, France (*emmanuel.flachaire@univ-amu.fr*)

Lynda Khalaf
Economics Department, Carleton University, Ottawa, Ontario K1S 5B6, Canada, and Centre Interuniversitaire de Recherche en Économie Quantitative (CIREQ), and Groupe de Recherche en Économie de l'énergie, de l'environnement et des Ressources Naturelles (GREEN), Université Laval, QC, G1V 0A6, Canada (*Lynda_Khalaf@carleton.ca*)

Asymptotic and bootstrap tests for inequality measures are known to perform poorly in finite samples when the underlying distribution is heavy-tailed. We propose Monte Carlo permutation and bootstrap methods for the problem of testing the equality of inequality measures between two samples. Results cover the Generalized Entropy class, which includes Theil's index, the Atkinson class of indices, and the Gini index. We analyze finite-sample and asymptotic conditions for the validity of the proposed methods, and we introduce a convenient rescaling to improve finite-sample performance. Simulation results show that size correct inference can be obtained with our proposed methods despite heavy tails if the underlying distributions are sufficiently close in the upper tails. Substantial reduction in size distortion is achieved more generally. Studentized rescaled Monte Carlo permutation tests outperform the competing methods we consider in terms of power.

KEY WORDS: Bootstrap; Income distribution; Inequality measures; Permutation test.

1. INTRODUCTION

Income and wealth distributions are typically nonnormal and can take various shapes. In view of this, distribution-free approaches are especially well suited to the task of comparing inequality measures. However, despite a sizable literature, nonparametric methods for inference on such measures perform poorly in finite samples. As the sample size grows, concern shifts from finite-sample distortions to asymptotic problems caused by the failure of the assumptions needed to ensure size control. These problems are often associated with heavy tails, a common situation in applied work. In economics, for example, income inequalities are of primary interest and income distributions are characterized by a prominently heavy right tail. In addition, inequality measures can be equal even if the underlying distributions differ, which also confounds inference.

Consider two variables x and y drawn from two distributions F_x and F_y. We study distribution-free tests of

$$H_0 : \theta(F_x) = \theta(F_y), \tag{1}$$

where $\theta(.)$ is some functional on some subset $\mathscr{F}$ of distributions (further structure is provided below). Inequality indices constitute special cases of $\theta(.)$ and provide the motivation for our work. Formally, we analyze centered and uncentered moments, the Generalized Entropy (GE) class of inequality measures, which includes Theil's index, the Atkinson class of inequality indices (Atkinson 1970), and the Gini index.

While bootstrapping offers a natural alternative to standard asymptotic approximations for this problem, Davidson and Flachaire (2007), Cowell and Flachaire (2007), Schluter and van Garderen (2009), and Davidson (2009, 2012) showed that heavy tails also cause bootstrap failures. A few improvements have been proposed. Davidson and Flachaire (2007) considered a bootstrap data-generating process (DGP), which combines a parametric estimate of the upper tail with a nonparametric estimate of the rest of the distribution. Schluter and van Garderen (2009) proposed normalizing transformations of inequality measures using Edgeworth expansions, to adjust asymptotic Gaussian approximations. Such corrections can be effective in specific instances—for example, when the null

Journal of Business & Economic Statistics
XXXX 2018, Vol. 0, No. 0
DOI: 10.1080/07350015.2017.1371027
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jbes.

hypothesis takes the form $H_1 : \theta(F_x) = \delta_0$ with δ_0 known—but still fail for heavy-tailed distributions.

This article analyzes permutation methods for testing H_0 in (1) and shows that permutational Monte Carlo (MC) test methods (Pitman 1937; Dwass 1957; Dufour 2006) provide substantial improvement. We ask what finite and large-sample assumptions are needed to support reliable permutations, focusing on the specificities of commonly used inequality measures. Our analysis applies and extends the theoretical setups of Romano (1990), Dufour (2006), and Chung and Romano (2013).

We first consider a baseline problem, which restricts (1) to the case where available samples are drawn from the same distribution, that is, when $F_x = F_y$. We show that MC permutation tests provide exact inference in finite samples, even if the common distribution is heavy-tailed. Our result allows for continuous and discrete distributions, and does not require any regularity condition on the form of the functional $\theta(\cdot)$. We also allow for exchangeable (as opposed to iid) observations, hence covering the case of random draws without replacement from a finite population. To the best of our knowledge, although restrictive, this special case provides the only available exact solution for the problem at hand.

The fact remains that (1) does not guarantee that $F_x = F_y$. In this case, the use of permutation tests is not justified from an exact perspective. Romano (1990) showed that, when $F_x \neq F_y$, permutation tests of the hypothesis in (1) are asymptotically valid in specific cases—in the sense that the probability of Type I error tends to the nominal level asymptotically—but they are not generally valid. For instance, permutations work using differences of sample means if the samples are of the same size, but are invalid with differences of medians. We suggest a convenient rescaling that validates permutations for several inequality measures. A bootstrap method for this null hypothesis is also proposed.

More recently, Chung and Romano (2013) showed that permutation tests are asymptotically valid in a more general setting if the underlying statistic is studentized. The importance of studentization is well-known for bootstrapping to achieve asymptotic refinements (Hall 1992). In contrast, with permutation tests, studentization may be required for validity. In particular, when comparing medians, studentized statistics will work while nonstudentized counterparts are invalid. Although Chung and Romano (2013) did not analyze inequality measures, their general statistical setup validates comparing these measures using studentized criteria. The rescaling we introduce may not be necessary for size control with studentized criteria, at least asymptotically. Yet we show that it matters from the power perspective.

Simulation experiments are provided to study the finite-sample properties of the proposed tests when the samples are drawn from similar and different distributions. Some very heavy-tailed distributions are considered to include a worst-case scenario in our design. In terms of Type 1 error or size distortion, our results show that when the samples are drawn from two (strongly) heavy-tailed distributions, which are not too different, permutation tests perform very well in finite samples. When distributions differ dramatically particularly in their tails, while size distortions are not completely eradicated, permutation tests outperform the standard asymptotic and bootstrap tests. In terms of power, our results show that permutation tests based on rescaled samples perform better in small samples than permutation tests based on original samples. We also analyze dependent samples and permutation algorithms with matched (paired) data. Results are broadly similar to the iid case, which confirms the finite-sample superiority of MC permutation methods.

The article is organized as follows. Section 2 describes a general framework and presents the proposed inference methods. In Section 3, we show how exact simulation-based permutation tests for the hypothesis of equal distributions can be obtained using statistics comparing general functionals of empirical distribution functions. In Section 4, we consider the problem of testing the equality of general functionals when the distributions of the two populations can differ. In Section 5, we study specific cases based on moments and commonly used inequality measures. Simulation experiments are reported in Section 6. We conclude in Section 7.

2. FRAMEWORK

In this section, we set notation, define the test statistics, and present alternative permutational and simulation-based p-values for comparing a general functional $\theta(.)$ on two different populations. The specific treatment of inequality measures is deferred to Section 5.

We consider two samples $X = \{X_1, X_2, \ldots, X_n\}$ and $Y = \{Y_1, Y_2, \ldots, Y_m\}$ each of which comprises independent and identically distributed observations with cumulative distribution functions F_x and F_y, respectively. We wish to test general hypotheses of type H_0 as in (1). A natural statistic for such a problem is given by

$$T = \theta(\hat{F}_x) - \theta(\hat{F}_y), \tag{2}$$

where $\hat{F}_x$ and $\hat{F}_y$ are the empirical distribution functions (EDFs) of the samples X and Y, and N is the total number of observations, $N = n + m$. On studentizing T, we get the *studentized* test statistic

$$S = \frac{\theta(\hat{F}_x) - \theta(\hat{F}_y)}{\sqrt{\hat{V}[\theta(\hat{F}_x)] + \hat{V}[\theta(\hat{F}_y)]}}, \tag{3}$$

where $\hat{V}[.]$ denotes an estimate of the variance of the indices in question.

Suppose the asymptotic distribution of S under H_0 is $N[0, 1]$ (as $m, n \to \infty$), and consider a critical region of the form $|S| > c$ where c is a critical value. Then an asymptotic p-value for this test can be obtained as follows:

$$p_a = G_\Phi(|S_0|) = 2\ \min(\Phi(S_0)\,;\ 1 - \Phi(S_0)), \tag{4}$$

where S_0 is the observed value of S,

$$\begin{aligned} G_\Phi(x) &:= \mathbb{P}[|Z| > |x|] \\ &= \Phi(-|x|) + 1 - \Phi(|x|) \\ &= 2 \min(\Phi(x),\ 1 - \Phi(x)), \end{aligned} \tag{5}$$

$Z \sim N[0, 1]$, and $\Phi(\cdot) = \mathbb{P}[Z \leq x]$ is the standard Normal distribution function. Note the identities in (5) depend on the continuity and symmetry of the normal distribution.

We introduce three p-values, described in the following subsections. For clarity, in what follows, p_* refers to the MC-permutation p-value, $p_\flat$ to the bootstrap p-value, and $p_\bullet$ denotes its counterpart that imposes the null hypothesis.

2.1 Monte Carlo Permutational p-Values

The permutational distribution based on the test statistic T in (2), also known as the randomization distribution, is the distribution obtained by permuting in all possible ways the $N = n + m$ observations of the combined sample

$$Z = (X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m)'. \tag{6}$$

We denote $\mathscr{P}(Z)$ the set of all vectors obtained by permuting the components of Z. Two permutations are viewed as distinct as soon as they correspond to different orderings of the components of Z (even if some of the observations are numerically equal), so the total number of different permutations in $\mathscr{P}(Z)$ is $(m + n)!$. Under the assumption that the $m + n$ observations in Z are iid, the $(m + n)!$ permutations in $\mathscr{P}(Z)$ are equally probable, which in turn determines the permutational distribution of T (or S). However, the total number of permutations $(m + n)!$ to consider rapidly becomes prohibitively large as the sample sizes m and n increase.

Following the suggestion of Dwass (1957), we draw at random B permutations of Z from the set $\mathscr{P}(Z)$. These may be drawn with or without replacement [in $\mathscr{P}(Z)$]. When draws are taken with replacement, the random permutations are iid; when taken without replacement, they are exchangeable. In this article, we focus on the case where the permutations are generated without replacement. Along with the actual data, this yields $B + 1$ random permutations of Z: $Z_{*1}, \ldots, Z_{*B}$. From each permuted sample, the corresponding EDFs $\hat{F}_{x*j}$ and $\hat{F}_{y*j}$ are computed, and the value of the test statistic as defined in (2):

$$T_{*j} = \theta(\hat{F}_{x*j}) - \theta(\hat{F}_{y*j}), \quad j = 1, \ldots, B. \tag{7}$$

Using the above simulated permutational test statistics, we can then compute the following p-value functions:

$$\hat{p}^-_{*B}(x) = \frac{\sum_{j=1}^{B} \mathbf{1}[T_{*j} \le x] + 1}{B + 1}, \quad \hat{p}^+_{*B}(x) = \frac{\sum_{j=1}^{B} \mathbf{1}[T_{*j} \ge x] + 1}{B + 1}, \tag{8}$$

where the indicator function $\mathbf{1}(A)$ is equal to one if A is true, and zero otherwise. We can then obtain one-sided tests of H_0 against $H_1^- : \theta(F_x) < \theta(F_y)$ and $H_1^+ : \theta(F_x) > \theta(F_y)$, by taking the following critical regions, respectively:

$$\hat{p}^-_{*B}(T) \le \alpha, \tag{9}$$

$$\hat{p}^+_{*B}(T) \le \alpha, \tag{10}$$

where α is the level of the test and T is the observed value of the test statistic. To get a two-sided test, we can reject H_0 against $H_1 : \theta(F_x) \neq \theta(F_y)$ when either one of the one-sided tests is significant at level $\alpha/2$:

$$\hat{p}^-_{*B}(T) \le \alpha/2 \quad \text{or} \quad \hat{p}^+_{*B}(T) \le \alpha/2 \tag{11}$$

or equivalently

$$\hat{p}^c_{*B} := 2 \min\{\hat{p}^-_{*B}(T),\ \hat{p}^+_{*B}(T)\} \le \alpha. \tag{12}$$

Another way of building a two-sided test consists in working with the absolute value of the test statistic: setting

$$\hat{p}^a_{*B}(x) = \frac{\sum_{j=1}^{B} \mathbf{1}(|T_{*j}| \ge x) + 1}{B + 1}, \tag{13}$$

H_0 is rejected against H_1 when

$$\hat{p}^a_{*B}(|T|) \le \alpha. \tag{14}$$

The two critical regions in (12) and (14) are not generally equivalent. In this article, we focus on two-sided tests of type (12). In (8)–(14), the statistic T can be replaced by its studentized version S, in which case

$$S_{*j} = \frac{\theta(\hat{F}_{x*j}) - \theta(\hat{F}_{y*j})}{\sqrt{\hat{V}[\theta(\hat{F}_{x*j})] + \hat{V}[\theta(\hat{F}_{y*j})]}}, \quad j = 1, \ldots, B. \tag{15}$$

Of course, tests based on T or S are not generally equivalent.

2.2 Conventional Bootstrap p-Values

A bootstrap test is computed by resampling the original data with replacement. A bootstrap sample, of the same size as the observed sample, is obtained by making n draws with replacement from the n observed realizations $\{X_1, \ldots, X_n\}$, where each X_i has probability $1/n$ of being selected on each draw, and then making, independently, m draws with replacement from the m observed realizations $\{Y_1, \ldots, Y_m\}$, where each Y_i has probability $1/m$ of being selected on each draw. Let $(X_\flat, Y_\flat)$ refer to the bootstrap sample so obtained, and denote by $\hat{F}_{x_\flat}$ and $\hat{F}_{y_\flat}$ the associated EDFs. The bootstrap statistic is computed as was S in (3), except that the null hypothesis tested is that the difference between the two indices is equal to $\theta(\hat{F}_x) - \theta(\hat{F}_y)$ rather than to 0. Formally, the adjusted bootstrap statistic takes the form

$$S_\flat = \frac{\big(\theta(\hat{F}_{x_\flat}) - \theta(\hat{F}_{y_\flat})\big) - \big(\theta(\hat{F}_x) - \theta(\hat{F}_y)\big)}{\sqrt{\hat{V}[\theta(\hat{F}_{x_\flat})] + \hat{V}[\theta(\hat{F}_{y_\flat})]}}. \tag{16}$$

This modification ensures that the hypothesis pertaining to the bootstrap statistics holds true for the population the bootstrap samples are drawn from, that is, the original sample. Let $S_{\flat j}$, $j = 1, \ldots, B$, refer to the series of bootstrap statistics. The bootstrap p-value is the proportion of the bootstrap samples for which the absolute value of the bootstrap statistic is more extreme than the statistic computed from the original data. Thus, for a two-tailed test, the bootstrap p-value is

$$p_\flat = 2 \min \left(\frac{1}{B} \sum_{j=1}^{B} \mathbf{1}(S_{\flat j} \le S_0);\ \frac{1}{B} \sum_{j=1}^{B} \mathbf{1}(S_{\flat j} > S_0) \right). \tag{17}$$

2.3 Bootstrap p-Values Under the Null Hypothesis

The permutation approach does not differ radically from the bootstrap approach. For example, a sample obtained by permuting elements of the combined sample Z defined in (6) is equivalent to resampling *without* replacement N observations from Z. It thus makes sense to resample *with* replacement from Z to form an alternative bootstrap sample that respects the null hypothesis. One can proceed as follows. Draw with replacement n observations in Z to form a sample denoted $X_\bullet$ and then draw with replacement m other observations in Z to form a sample denoted $Y_\bullet$. Again, let $\hat{F}_{x_\bullet}$ and $\hat{F}_{y_\bullet}$ refer to the associated EDFs, respectively. The bootstrap statistic can be computed as

$$S_\bullet = \frac{\theta(\hat{F}_{x_\bullet}) - \theta(\hat{F}_{y_\bullet})}{\sqrt{\hat{V}[\theta(\hat{F}_{x_\bullet})] + \hat{V}[\theta(\hat{F}_{y_\bullet})]}} \tag{18}$$

with no further adjustments since the sampling scheme imposes the null hypothesis. Let $S_{\bullet j}$, $j = 1, \ldots, B$, refer to the series of bootstrap statistics so obtained, leading to the two-tailed bootstrap p-value:

$$p_\bullet = 2 \min \left(\frac{1}{B} \sum_{j=1}^{B} \mathbf{1}(S_{\bullet j} \le S_0);\ \frac{1}{B} \sum_{j=1}^{B} \mathbf{1}(S_{\bullet j} > S_0) \right). \tag{19}$$

3. EXACT MONTE CARLO PERMUTATION TESTS

Before we consider the general problem of testing H_0, it will be of interest to study the problem of testing H_0 when the populations considered have the same distributions, against H_1^-, H_1^+ or H_1. This is equivalent to testing $F_x = F_y$ against alternatives where $\theta(F_x) \neq \theta(F_y)$. This relatively restrictive null hypothesis appears naturally when subsets of a wider population are considered. We will show here that both the level and the size of permutation tests based on general statistics of the form T or S can be controlled, irrespective whether the distribution F_x (or F_y) is discrete or continuous, without any restriction on the form of the functional $\theta(\cdot)$. We also allow for exchangeable (as opposed to iid) observations, hence covering the case of random draws without replacement from a finite population (in addition to iid observations). Thus, the result given here can be viewed as an extension of the basic finding of Dwass (1957) who considered tests that compare arithmetic means of iid random variables with continuous distribution.

Since most estimated inequality measures rely on statistics based on EDFs, which are not continuous, a tie-breaking procedure may be needed to control test size. For this purpose, we propose to use the randomized ordering described in Dufour (2006), which leads to the following procedure. We first draw by simulation $B + 1$ iid random variables $U_0, U_1, \ldots, U_B$, according to a uniform distribution on (0, 1), independently of $(T, T_{*1}, \ldots, T_{*B})$. Then we compute p-value functions similar to those described in Section 2.1, with the difference that the pairs (T_{*j}, U_j), $j = 0, \ldots, B$, are ordered according to the lexicographic order:

$$\begin{aligned}(T_{*i}, U_i) \le (T_{*j}, U_j) \Leftrightarrow [&T_{*i} < T_{*j} \\ \text{or} \quad &(T_{*i} = T_{*j} \text{ and } U_i \le U_j)],\end{aligned} \tag{20}$$

where $T_{*0} = T$ is the statistic obtained from the actual data. More precisely, this yields the following modified (*tie-adjusted*) p-value functions:

$$\tilde{p}^-_{*B}(x) = \frac{\sum_{i=1}^{B} \mathbf{1}[(T_{*i}, U_i) \le (x, U_0)] + 1}{B + 1}, \tag{21}$$

$$\tilde{p}^+_{*B}(x) = \frac{\sum_{i=1}^{B} \mathbf{1}[(x, U_0) \le (T_{*i}, U_i)] + 1}{B + 1}, \tag{22}$$

$$\tilde{p}^a_{*B}(x) = \frac{\sum_{i=1}^{B} \mathbf{1}[(x, U_0) \le (|T_{*i}|, U_i)] + 1}{B + 1}. \tag{23}$$

The tests are performed as before on replacing $\hat{p}$ by $\tilde{p}$. We can then establish the following theorem.

Theorem 1. Suppose the $n + m$ random variables $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$ are exchangeable. Then, for $0 < \alpha < 1$,

$$\mathbb{P}[\hat{p}^-_{*B}(T) \le \alpha] \le \mathbb{P}[\tilde{p}^-_{*B}(T) \le \alpha] = \alpha, \tag{24}$$

$$\mathbb{P}[\hat{p}^+_{*B}(T) \le \alpha] \le \mathbb{P}[\tilde{p}^+_{*B}(T) \le \alpha] = \alpha, \tag{25}$$

$$\mathbb{P}[\hat{p}^a_{*B}(|T|) \le \alpha] \le \mathbb{P}[\tilde{p}^a_{*B}(|T|) \le \alpha] = \alpha, \tag{26}$$

where the p-value functions are defined in (8) and (21)–(23).

Proof. By the exchangeability assumption, different permutations of the components of Z are equally probable, so randomly selected permutations (either with or without replacement) are themselves exchangeable. Consequently, the random variables $T, T_{*1}, \ldots, T_{*B}$ are exchangeable. On applying Proposition 2.4 of Dufour (2006), we then get

$$\mathbb{P}[\tilde{p}^-_{*B}(T) \le \alpha] = \mathbb{P}[\tilde{p}^+_{*B}(T) \le \alpha] = \mathbb{P}[\tilde{p}^a_{*B}(|T|) \le \alpha] = \alpha. \tag{27}$$

Finally, the inequalities in (24)–(26) follow on observing that

$$\tilde{p}^-_{*B}(x) \le \hat{p}^-_{*B}(x), \quad \tilde{p}^+_{*B}(x) \le \hat{p}^+_{*B}(x), \quad \tilde{p}^a_{*B}(x) \le \hat{p}^a_{*B}(x). \tag{28}$$

□

Theorem 1 means that the critical regions $\tilde{p}^-_{*B}(T) \le \alpha$, $\tilde{p}^+_{*B}(T) \le \alpha$, and $\tilde{p}^a_{*B}(|T|) \le \alpha$ have size α for testing H_0 when the populations considered have the same distributions, while the critical regions $\hat{p}^-_{*B}(T) \le \alpha$, $\hat{p}^+_{*B}(T) \le \alpha$, and $\hat{p}^a_{*B}(|T|) \le \alpha$ are typically conservative, so they still have level α for testing H_0. Clearly, the same result holds if the test statistic T is replaced by its studentized version S.

Note also the events $\tilde{p}^-_{*B}(T) \le \alpha/2$ and $\tilde{p}^+_{*B}(T) \le \alpha/2$ are mutually exclusive [and similarly for $\hat{p}^-_{*B}(T) \le \alpha/2$ and $\hat{p}^+_{*B}(T) \le \alpha/2$] for $0 < \alpha < 1$, so that

$$\begin{aligned}\mathbb{P}[\hat{p}^-_{*B}(T) \le \alpha/2 \text{ or } \hat{p}^+_{*B}(T) \le \alpha/2] \le \mathbb{P}[&\tilde{p}^-_{*B}(T) \\ &\le \alpha/2 \text{ or } \tilde{p}^+_{*B}(T) \le \alpha/2] = \alpha\end{aligned} \tag{29}$$

under H_0 when the populations considered have the same distributions, for $0 < \alpha < 1$. Thus, on setting

$$\begin{aligned}\hat{p}^c_{*B}(x) &:= 2 \min\{\hat{p}^-_{*B}(x),\ \hat{p}^+_{*B}(x)\}, \\ \tilde{p}^c_{*B}(x) &:= 2 \min\{\tilde{p}^-_{*B}(x),\ \tilde{p}^+_{*B}(x)\},\end{aligned} \tag{30}$$

we have

$$\mathbb{P}[\hat{p}^c_{*B}(T) \le \alpha] \le \mathbb{P}[\tilde{p}^c_{*B}(T) \le \alpha] = \alpha \tag{31}$$

under H_0, so $\hat{p}^c_{*B}(T) \le \alpha$ and $\tilde{p}^c_{*B}(T) \le \alpha$ constitute two-sided critical regions with level α for H_0 $(0 < \alpha < 1)$. The latter are not in general equivalent to $\hat{p}^a_{*B}(|T|) \le \alpha$ or $\tilde{p}^a_{*B}(|T|) \le \alpha$.

4. PERMUTATION TESTS FOR COMPARING LINEAR FUNCTIONALS: DIFFERENT DISTRIBUTIONS

In this section, we identify specific finite-sample permutation test problems and extend the results in Romano (1990) on the asymptotic validity of permutation tests based on T [in (2)] to the case where $\theta(.)$ is a linear functional. This result will be used in the next section to show that permutation tests are asymptotically valid for inequality measures if the samples are previously rescaled adequately. Permutations based on studentized statistics [such as (3) here] have been shown to be valid under rather general regularity conditions by Chung and Romano (2013). The following analysis, which focuses on (2) is nevertheless informative, as the properties we derive are new to this literature. The transformation we introduce for simple contrasts [like (2)] also ends up enhancing power when combined with studentization.

In general, when the assumption $F_x = F_y$ does not hold, permutation tests of H_0 [in (1)] are no longer exact (at level α). However, as pointed out by Romano (1990), such tests can be asymptotically valid in specific cases, in the sense that under the null hypothesis, the rejection frequency of H_0 tends to the nominal level α as the sample size increases.

For the general two-sample problem of testing $\theta(F_x) = \theta(F_y)$, when the statistics $n^{1/2}[\theta(\hat{F}_x) - \theta(F_x)]$ and $m^{1/2}[\theta(\hat{F}_y) - \theta(F_y)]$ converge weakly to Gaussian distributions with mean 0, the corresponding permutation test is asymptotically valid when the asymptotic variances of the statistics $\theta(\hat{F}_x)$ and $\theta(\hat{F}_y)$ are equal; see Romano (1990). This result requires that $\theta(.)$ be appropriately differentiable in the sense of Gill (1988). However, asymptotic validity of the permutation test does not generally hold in such two-sample problems, while it does for one-sample problems.

With two independent samples, $\{X_1, \dots, X_n\}$ and $\{Y_1, \dots, Y_m\}$, drawn from the probability distributions F_x and F_y, the test statistic $n^{1/2}[\theta(\hat{F}_x) - \theta(\hat{F}_y)]$ is asymptotically Gaussian with mean 0 and variance

$$V_{as}\left[\theta(\hat{F}_x)\right] + \frac{1-\lambda}{\lambda} V_{as}\left[\theta(\hat{F}_y)\right] \tag{32}$$

under H_0, where $V_{as}[\theta(\hat{F}_x)]$ and $V_{as}[\theta(\hat{F}_y)]$ are the asymptotic variances of $n^{1/2}[\theta(\hat{F}_x) - \theta(F_x)]$ and $m^{1/2}[\theta(\hat{F}_y) - \theta(F_y)]$, as $n \to \infty$ and $m/(m+n) \to \lambda$, where $0 < \lambda < 1$ (Romano 1990, Theorem 3.3). Further, the permutational distribution of the test statistic is asymptotically Gaussian with mean 0 and variance

$$V_{as}\left[\theta\left((1-\lambda)\hat{F}_x + \lambda\hat{F}_y\right)\right] + \frac{1-\lambda}{\lambda} V_{as}\left[\theta\left((1-\lambda)\hat{F}_x + \lambda\hat{F}_y\right)\right]. \tag{33}$$

This result is useful to consider asymptotic validity of a permutation test when the variances in (32) and (33) are the same, that is, when the following condition holds:

$$V_{as}\left[\theta\left((1-\lambda)\hat{F}_x + \lambda\hat{F}_y\right)\right] = \lambda\, V_{as}\left[\theta(\hat{F}_x)\right] + (1-\lambda)\, V_{as}\left[\theta(\hat{F}_y)\right]. \tag{34}$$

In this case, the permutation test is asymptotically valid, in the sense that the permutational distribution of the test statistic is asymptotically identical to the unconditional sampling distribution of $n^{1/2}[\theta(\hat{F}_x) - \theta(\hat{F}_y)]$. This permutational distribution can then be used to compute a critical value or a p-value. Condition (34) holds in particular when the sample sizes are "equal" (in large samples) or when the distributions have the same asymptotic variances. But it does not hold in general.

In view of applying these general results to compare inequality measures, it will be useful to consider the case where the functionals involved are mixture-linear and the available estimators are asymptotically linear. These concepts are defined in what follows.

Definition 1. Let $\mathscr{D}$ be a set of distribution functions on the real numbers, and $\theta : \mathscr{D} \to \mathbb{R}$ a mapping from $\mathscr{D}$ to the real numbers. We say that θ is a *mixture-linear functional* of order K, where K is a positive integer, if $\mathscr{D}$ is closed under linear mixtures, that is, $F_1, \dots, F_K \in \mathscr{D}$ entails $\sum_{k=1}^{K} \lambda_k F_k \in \mathscr{D}$, and $\theta(\cdot)$ satisfies the condition

$$\theta\left(\sum_{k=1}^{K} \lambda_k F_k\right) = \sum_{k=1}^{K} \lambda_k \theta(F_k), \tag{35}$$

for all $F_1, \dots, F_K \in \mathscr{D}$ and nonnegative scalars $\lambda_1, \dots, \lambda_K \in \mathbb{R}$ such that $\sum_{k=1}^{K} \lambda_k = 1$.

Definition 2. Given a sample $Z = \{Z_1, \dots, Z_n\}$ of independent observations drawn from distribution F_z, an estimator $\theta(\hat{F}_z)$ of the functional $\theta(F_z)$ is *asymptotically linear Gaussian* if

$$n^{1/2}[\theta(\hat{F}_z) - \theta(F_z)] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_z(Z_i) + o_{F_z}(1) \xrightarrow[n\to\infty]{d} \mathrm{N}[0,\, \sigma_\theta(F_z)], \tag{36}$$

where $\sigma_\theta(F_z)$ depends on the underlying distribution.

Theorem 2. Consider the problem of testing $H_0 : \theta(F_x) = \theta(F_y)$ from two independent samples $X = \{X_1, \dots, X_n\}$ and $Y = \{Y_1, \dots, Y_m\}$ of independent observations with distributions F_x and F_y, respectively, where F_x and F_y have finite second moments, and let $\mathscr{D} = \{F_z : F_z = \lambda F_x + (1-\lambda)F_y, 0 < \lambda < 1\}$. Suppose $\theta : \mathscr{D} \to \mathbb{R}$ is a mixture-linear functional of order 2, and the estimator $\theta(\hat{F}_z)$ is asymptotically linear Gaussian for any iid sample $Z_1, \dots, Z_n$ with distribution $F_z \in \mathscr{D}$. If

$$V_{as}[\theta(\hat{F}_x)] = V_{as}[\theta(\hat{F}_y)], \tag{37}$$

where $V_{as}[\theta(\hat{F}_x)]$ and $V_{as}[\theta(\hat{F}_y)]$ are the asymptotic variances of $n^{1/2}[\theta(\hat{F}_x) - \theta(F_x)]$ and $m^{1/2}[\theta(\hat{F}_y) - \theta(F_y)]$, or if

$$m/(m+n) \underset{n\to\infty}{\longrightarrow} \lambda = 1/2, \tag{38}$$

then the permutation test based on T defined in (2) and the MC replications (7) is asymptotically valid.

Proof. The fact that the permutational distribution (under H_0) of the test statistic $n^{1/2}[\theta(\hat{F}_x) - \theta(\hat{F}_y)]$ is asymptotically Gaussian follows from condition (36); see Chung and Romano (2013). (Alternatively, Romano (1990) relied on differentiability conditions.) Let $\lambda_{1,n}$ and $\lambda_{2,n}$ be fixed positive constants (which may depend on n) such that $\lambda_{1,n} + \lambda_{2,n} = 1$. Since $\theta(.)$ is a linear functional and the two samples are independent, we have

$$V\left[\theta\left(\sum_{k=1}^{2} \lambda_{k,n}\, \hat{F}_k\right)\right] = V\left[\sum_{k=1}^{2} \lambda_{k,n}\, \theta(\hat{F}_k)\right] = \sum_{k=1}^{2} \lambda_{k,n}\, V\left[\theta(\hat{F}_k)\right].$$

For $F_1 = F_x$ and $F_2 = F_y$, we have

$$V\left[\theta\left(\lambda_{1,n}\hat{F}_x + \lambda_{2,n}\hat{F}_y\right)\right] = \lambda_{1,n}\, V[\theta(\hat{F}_x)] + \lambda_{2,n}\, V[\theta(\hat{F}_y)]. \tag{39}$$

Take $\lambda_{2,n} := m/(m+n)$ and $\lambda_{1,n} := n/(m+n) = 1 - \lambda_{2,n}$. Then, $\lambda_{2,n} \to \lambda$ and $\lambda_{1,n} \to 1 - \lambda$ as $n \to \infty$, where $0 < \lambda < 1$, hence

$$V[\theta(\lambda_{1,n}\hat{F}_x + \lambda_{2,n}\hat{F}_y)] \underset{n\to\infty}{\longrightarrow} (1-\lambda)\, V_{as}[\theta(\hat{F}_x)] + \lambda\, V_{as}[\theta(\hat{F}_y)]. \tag{40}$$

The condition (34) is then satisfied if $\lambda = 1/2$ or $V_{as}[\theta(\hat{F}_x)] = V_{as}[\theta(\hat{F}_y)]$. □

Note that the (arithmetic) mean is a linear functional, but the quantile is not a linear functional. Comparing means from samples of similar size with a permutation test is then asymptotically valid, even if the underlying distributions are not identical, while comparing quantiles with a permutation test is no longer valid, in general, if the underlying distributions are not identical.

5. COMPARING INEQUALITY MEASURES

This section focuses on identifying functionals $\theta(.)$ of interest for which permutation tests (7) are valid in the sense of Theorem 2. We study moments and inequality measures for which condition (34) would hold.

5.1 Centered and Uncentered Moments

Consider the functional

$$\theta(F_z) = \int \phi(z)\, dF_z(z), \tag{41}$$

where $\phi(.)$ is any function in $\mathbb{R}$ for which $E[\phi(.)]$ exists. For any random variable w that follows a mixture of K distributions, $w \sim \sum_{k=1}^{K} \lambda_k F_k(w)$, if we consider K random variables $w_1, \ldots, w_K$ from the K component distributions, then

$$\theta(F_w) = E[\phi(w)] = \sum_{k=1}^{K} \lambda_k E[\phi(w_k)] = \sum_{k=1}^{K} \lambda_k \theta(F_{w_k}). \tag{42}$$

The functional (41) is linear and Theorem 2 applies: permutation tests are asymptotically valid if either (37) or (38) holds.

Then, comparing *uncentered* moments between two samples with permutation tests is asymptotically valid, it corresponds to the special case of (41) with $\phi(z) = z^r$, where r is a positive integer. The mean corresponds to the case $\phi(z) = z$. Turning to *centered* moments, consider

$$\theta(F_z) = \int [z - E(z)]^r\, dF_z(z), \tag{43}$$

where r is an integer greater than 1. With w and $w_1, \ldots, w_K$ as in (42), we have

$$\begin{aligned}\theta(F_w) &= E([w - E(w)]^r)\\ &= \sum_{k=1}^{K} \lambda_k E([w_k - E(w_k) + E(w_k) - E(w)]^r).\end{aligned} \tag{44}$$

The two last terms in parenthesis vanish if they are equal leading to

$$\begin{aligned}\theta(F_w) &= \sum_{i=1}^{K} \lambda_k E([w_k - E(w_k)]^r)\\ &= \sum_{i=1}^{K} \lambda_k \theta(F_{w_k}) \quad \text{if} \quad E(w_k) = E(w), \forall k.\end{aligned}$$

This result suggest that (43) is not a linear functional, unless the component distributions share a common mean. From Theorem 2, comparing centered moments from two samples with a permutation test (7) is then invalid, unless the samples come from distributions with the same mean, $\mu(F_x) = \mu(F_y)$, and either (37) or (38) holds.

However, centered moments are translation invariant: calculating centered moments from the original samples or from the *centered* samples,

$$\begin{aligned}&\{X_1 - \mu(F_x), \ldots, X_n - \mu(F_x)\} \quad \text{and}\\ &\{Y_1 - \mu(F_y), \ldots, Y_m - \mu(F_y)\},\end{aligned} \tag{45}$$

gives the same result. The main issue here is that the centered samples have a common mean, which is equal to zero, and the statistic (43) is a linear functional in this particular case. Comparing centered moments from the two *centered* samples rather than from the original samples makes no difference, while it validates (asymptotically) the use of permutation test.

In practice, $\mu(F_x)$ and $\mu(F_y)$ are often unknown. Permutation tests can nevertheless be applied on the combined sample

$$Z_c = \left\{X_1 - \bar{X}, \ldots, X_n - \bar{X}, Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})\right\}, \tag{46}$$

where $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ and $\bar{Y} = m^{-1}\sum_{i=1}^{m} Y_i$ are the sample means of each sample. This procedure is known to perform well in finite samples when testing the equality of variances from two samples, see Lim and Loh (1996) and Boos and Brownie (2004).

5.2 The Generalized Entropy Class

We consider the important GE class of inequality measures, defined by

$$\theta_{\mathrm{GE}}^{\zeta}(F) = \begin{cases} \frac{1}{\zeta^2 - \zeta}\left[\int \left[\frac{y}{\mu(F)}\right]^{\zeta} dF(y) - 1\right], & \text{if } \zeta \neq 0, 1,\\ -\int \log\left[\frac{y}{\mu(F)}\right] dF(y), & \text{if } \zeta = 0,\\ \int \frac{y}{\mu(F)} \log\left[\frac{y}{\mu(F)}\right] dF(y), & \text{if } \zeta = 1.\end{cases} \tag{47}$$

The parameter ζ of the GE class characterizes the sensitivity to differences over different segments of the distribution. The more positive (negative) ζ is, the more sensitive is the inequality measure to differences at the top (bottom) of the distribution. The Mean Logarithmic Deviation (MLD) index, $\theta_{\mathrm{GE}}^{0}(F)$, is the limiting case when $\zeta = 0$. The Theil index, $\theta_{\mathrm{GE}}^{1}(F)$, is the limiting case of the GE when $\zeta = 1$.

The GE class of inequality measures is decomposable, that is, it can be expressed as a simple additive function of within-group and between-group inequality. Let there be K groups and let the proportion of the population falling in group k be λ_k; then the class of GE indices is equal to

$$\theta^{\zeta}_{\mathrm{GE}}(\hat{F}_{\varpi}) = \sum_{k=1}^{K} \lambda_k \left[\frac{\bar{\varpi}_k}{\bar{\varpi}}\right]^{\zeta} \theta^{\zeta}_{\mathrm{GE}}(\hat{F}_{\varpi_k}) - \frac{1}{\zeta^2-\zeta}\left(\sum_{k=1}^{K} \lambda_k \left[\frac{\bar{\varpi}_k}{\bar{\varpi}}\right]^{\zeta} - 1\right), \qquad (48)$$

where $\bar{\varpi}_k$ is the mean income in group k, $\bar{\varpi} = K^{-1}\sum_{k=1}^{K}\lambda_k\bar{\varpi}_k$ is the mean income of the population and $\theta^{\zeta}_{\mathrm{GE}}(\hat{F}_{\varpi_k})$ is the GE index in group k, see Cowell (2011). It is clear that

$$\theta^{\zeta}_{\mathrm{GE}}(\hat{F}_{\varpi}) = \sum_{k=1}^{K} \lambda_k \theta^{\zeta}_{\mathrm{GE},k}(\hat{F}_{\varpi_k}) \qquad \text{if} \qquad \bar{\varpi}_k = \bar{\varpi},\ \forall k. \qquad (49)$$

It follows that $\theta^{\zeta}_{\mathrm{GE}}(\hat{F}_{\varpi})$ is not a linear functional, unless the mean in each group is the same. From Theorem 2, comparing GE inequality measures from two samples with permutation tests (7) is then valid only if the samples come from distributions with the same mean, $\mu(F_x) = \mu(F_y)$, and either (37) or (38) holds.

As is clear from Equation (47), the GE class of inequality measures is scale invariant, which suggests to base a permutation test on the *rescaled* samples, where the observations are divided by their distributional mean,

$$\left\{\frac{X_1}{\mu(F_x)}, \ldots, \frac{X_n}{\mu(F_x)}\right\} \quad \text{and} \quad \left\{\frac{Y_1}{\mu(F_y)}, \ldots, \frac{Y_m}{\mu(F_y)}\right\}. \qquad (50)$$

Comparing Generalized Inequality indices from these rescaled samples rather than from the original samples makes no differences, while it validates (asymptotically) the use of permutation test. In practice, distributional means are often unknown; we thus use sample means $\bar{X}$ and $\bar{Y}$ instead, so the permutation test is based on the following combined sample:

$$Z_s = \left\{\frac{X_1}{\bar{X}}, \ldots, \frac{X_n}{\bar{X}}, \frac{Y_1}{\bar{Y}}, \ldots, \frac{Y_m}{\bar{Y}}\right\}, \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i. \qquad (51)$$

It is worth noting that when we consider the *rescaled* samples (50), the GE inequality measures can be rewritten as a moment $\theta^{\zeta}_{\mathrm{GE}}(F_z) = \int \phi(z)\, dF_z(z)$, as defined in (41), where $\phi(z) = (z^{\zeta}-1)/(\zeta^2-\zeta)$ for $\zeta \neq 0, 1$, $\phi(z) = -\log z$ for $\zeta = 0$, and $\phi(z) = z\log z$ for $\zeta = 1$, which leads us back to the results of Section 5.1.

Clearly the same approach can be applied to the Atkinson class of inequality indices (Atkinson 1970), $\theta^{\zeta}_{\mathrm{Atk}}(F) = 1 - [\int [\frac{y}{\mu(F)}]^{\zeta} dF(y)]^{1/\zeta}$, $\zeta < 1$, which can be rewritten as a function of the GE class of inequality measures

$$\theta^{\zeta}_{\mathrm{Atk}}(F) = \begin{cases} 1 - [(\zeta^2-\zeta)\theta^{\zeta}_{\mathrm{GE}}(F) + 1]^{1/\zeta}, & \zeta \neq 0, \\ 1 - \exp(-\theta^{0}_{\mathrm{GE}}(F)), & \zeta = 0. \end{cases} \qquad (52)$$

5.3 The Gini Coefficient

The Gini index can be expressed in a number of different forms. Let us consider the following expressions,

$$\theta_{\mathrm{Gini}}(F) = \frac{1}{2\mu(F)} \iint |y_1 - y_2|\, dF(y_1)dF(y_2) = \frac{E(|y_1-y_2|)}{2\mu(F)}, \qquad (53)$$

$$\theta_{\mathrm{Gini}}(F) = 1 - 2\int_0^1 L(F; q)\, dq, \qquad (54)$$

where y_1 and y_2 are two random variables independently drawn from F, and $L(F; q)$ is the qth ordinate of the Lorenz curve. Equation (53) presents the Gini as the normalized average absolute difference between all the possible pairs of incomes in the population, while Equation (54) shows that the Gini index is twice the area between the Lorenz curve and the 45° line.

The Gini index is also closely related to a measure of dispersion of a distribution. The most popular measure of dispersion is the standard deviation, which is the square root of the variance that can be rewritten as follows:

$$V(y) = E[(y-\mu(F))^2] = E\left[\frac{1}{2}(y_1-y_2)^2\right]. \qquad (55)$$

Another well-known measure of dispersion is the Gini's mean difference,

$$\Delta(F) = E(|y_1 - y_2|). \qquad (56)$$

Both measures of dispersion are translation invariant. In Section 5.1, we prove that testing the equality of variances from two samples with different means can be done with permutation tests based on the combined sample of the recentered individual samples. Boos, Janssen, and Veraverbeke (1989) proved that this procedure is asymptotically correct for a large class of U-statistics, from which the Gini's mean difference is a special case. We can then use the relationship between the Gini (inequality) index and the Gini's mean difference to justify asymptotically the use of permutation test with the Gini index. Indeed, we have

$$\Delta(F) = 2\mu(F)\theta_{\mathrm{Gini}}(F). \qquad (57)$$

With $\mu(F) = 1$, the Gini's mean difference is twice the Gini index. Comparing the Gini's mean difference or the Gini index from two samples is then equivalent if the underlying distributions share a common mean equal to one.

The last condition does not hold in general. However, the Gini index is scale invariant. Then, calculating Gini index from the original samples or from the *rescaled* samples, where the observations are divided by their distributional mean,

$$\left\{\frac{X_1}{\mu(F_x)}, \ldots, \frac{X_n}{\mu(F_x)}\right\} \quad \text{and} \quad \left\{\frac{Y_1}{\mu(F_y)}, \ldots, \frac{Y_m}{\mu(F_y)}\right\}, \qquad (58)$$

gives the same results. The main issue here is that these rescaled samples share a common mean, equals to one. Comparing Gini inequality measures from the two *rescaled* samples in (58) rather than from the original samples makes no difference for scale invariant statistic, while it validates asymptotically the use of permutation test. In practice, distributional means are replaced by sample means and permutation tests are based on the combined sample of empirically rescaled individual data Z_s as defined in (51).

6. SIMULATION STUDY

Overall, we focus our simulation study to extreme cases of (very) heavy-tailed distributions in small samples to stress-test the methods employed in testing. The heavy-tailed distribution used as a benchmark in previous studies is a more favorable case here, and we use much more heavy-tailed distributions with a very small number observations in each samples. In our experiments, we test the equality of Gini and Theil inequality measures between two samples.

6.1 Model Design

We make use of simulated datasets drawn from the Singh–Maddala distribution, which can quite successfully mimic observed income distributions in various countries (McDonald 1984; Kleiber and Kotz 2003). The CDF of the Singh–Maddala distribution, $\mathrm{SM}(x; a, b, q)$, can be written as $F(x) = 1 - [1 + (\frac{x}{b})^a]^{-q}$, where a, b, q are positive, b is a scale parameter and a, q are shape parameters; q only affects the right tail, whereas a affects both tails. The kth moment exists for $-a < k < aq$. The upper-tail of the Singh–Maddala distribution behaves like a Pareto distribution with a tail index equal to $\xi = aq$ (Schluter and Trede 2002). Smaller is ξ, heavier is the upper tail of the distribution.

As a benchmark, we use the parameter values $a = 2.8$, $b = 100^{-\frac{1}{2.8}}$, $q = 1.7$. This distribution is used in Davidson and Flachaire (2007) and Cowell and Flachaire (2007) to show poor finite-sample performance of asymptotic and bootstrap inference. Its tail index is equal to $\xi = aq = 4.76$. We will depart from this distribution using heaviest-tailed distributions (Singh–Maddala distributions with smaller tail parameters ξ) for which we know that bootstrap inference is poorest.

We compute the Theil and the Gini indices as follows:

$$\theta^1_{\mathrm{GE}}(\hat{F}_y) = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{\hat{\mu}} \log\left(\frac{y_i}{\hat{\mu}}\right) \quad \text{and}$$

$$\theta_{\mathrm{Gini}}(\hat{F}_y) = \frac{2\sum_{i=1}^{n} i y_{(i)}}{\hat{\mu} n(n-1)} - \frac{n+1}{n-1}, \tag{59}$$

where $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and, the $y_{(i)}$, $i = 1, \dots, n$ are the order statistics ($y_{(1)} \le \cdots \le y_{(n)}$). (The variance of the Theil index is computed as $\widehat{\mathrm{var}}(\theta^1_{\mathrm{GE}}(\hat{F}_y)) = \frac{1}{n^2}\sum_{i=1}^{n}(Z_i - \bar{Z})^2$, where $Z_i = \frac{y_i}{\hat{\mu}}[\log(\frac{y_i}{\hat{\mu}}) - \theta^1_{\mathrm{GE}}(\hat{F}_y) - 1]$, and $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$. The variance of the Gini index is computed as $\widehat{\mathrm{var}}(\theta_{\mathrm{Gini}}(\hat{F}_y)) = \frac{1}{(n\hat{\mu})^2}\sum_{i=1}^{n}(Z_i - \bar{Z})^2$, where $Z_i = -(\theta_{\mathrm{Gini}}(\hat{F}_y) + 1)y_{(i)} + \frac{2i-1}{n} y_{(i)} - \frac{2}{n}\sum_{j=1}^{i} y_{(j)}$ and $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$, see Davidson 2009; Cowell and Flachaire 2015.)

Our results are presented with figures, with the following legend:

- `asymptotic`: asymptotic test.
- `bootstrap`: standard bootstrap test S_b (defined in (16))
- `Perm T, rescaled`: permutation test T_* based on Z_s (defined in (7) and (51))
- `Perm S, rescaled`: permutation test S_* based on Z_s (defined in (15) and (51))
- `Perm S, standard`: permutation test S_* based on Z (defined in (15) and (6))
- `Boot S, rescaled`: bootstrap test $S_\bullet$ based on Z_s (defined in (18) and (51))
- `Boot S, standard`: bootstrap test $S_\bullet$ based on Z (defined (18) and (6))

The number of replications is equal to 10,000. The number of bootstrap and permutation samples are similar, $B = 999$. The permutation and bootstrap p-values are obtained as described above. We compute the rejection probability, or rejection frequency, as the proportion of p-value less than a nominal level equals to 0.05.

6.2 Size

In the experiments, we consider several Singh–Maddala distributions for which the Theil inequality measure index is the same and the tail index varies, $\xi \in [2.9, 6.26]$. (Singh–Maddala distributions with parameters (a, q) equal to $(2.5, 2.502199)$, $(2.6, 2.149747)$, $(2.7, 1.894309)$, $(2.8, 1.7)$, $(3.0, 1.4223847)$, $(3.2, 1.2320215)$, $(3.4, 1.0922125)$, $(3.8, 0.8984488)$, $(4.8, 0.6366578)$, and $(5.8, 0.4996163)$, share the same (scale-invariant) Theil index, equal to 0.1401151. The tail parameters are, respectively, equal to $\xi = 6.26, 5.59, 5.11, 4.76, 4.27, 3.94, 3.71, 3.41, 3.06, 2.9$.) The Singh–Maddala distribution with $\xi = 2.9$ is then the heaviest-tailed distribution considered here. Similar experiments are conducted for the Gini index, with slightly different tail parameters, $\xi \in [2.59, 6.6]$. (Singh–Maddala distributions with parameters (a, q) equal to $(2.5, 2.640350)$, $(2.6, 2.218091)$, $(2.7, 1.920967)$, $(2.8, 1.7)$, $(3.0, 1.3921126)$, $(3.2, 1.1866026)$, $(3.4, 1.0388049)$, $(3.8, 0.8387663)$, $(4.8, 0.5784599)$, and $(5.8, 0.4473111)$, share the same (scale-invariant) Gini index, equals to 0.2887138. The tail parameters are, respectively, equal to $\xi = 6.6, 5.77, 5.19, 4.76, 4.18, 3.80, 3.53, 3.19, 2.78, 2.59$.) Inference is exact if the rejection probability is equal to 0.05.

6.2.1. Identical Distributions. Figure 1 shows empirical rejection frequencies for asymptotic, bootstrap, and permutation tests for the Theil index, when $F_x = F_y$, as the upper tail becomes heavier. The sample size is very small $n = m = 50$. Figure 2 shows similar results for the Gini index. When the upper-tail of the distribution becomes heavier (as ξ_y decreases), asymptotic and standard bootstrap tests perform very poorly, while permutation and bootstrap under the null tests based on the studentized statistic (perm S rescaled, perm S standard, boot S rescaled, and boot S standard) provide empirical frequencies almost equal to 0.05. Note that studentized permutation tests based on the combined original samples (perm S standard) provides exact inference—as shown by Chung and Romano (2013), not permutation tests based on the combined rescaled samples (perm S rescaled): it is because samples are previously divided by sample means rather than by distributional means.

6.2.2. Different Distributions. We then generate samples from different distributions, $F_x \neq F_y$, with the same value of the inequality index. Figure 3 shows rejection frequencies for asymptotic, bootstrap, and permutation tests for the Theil index,

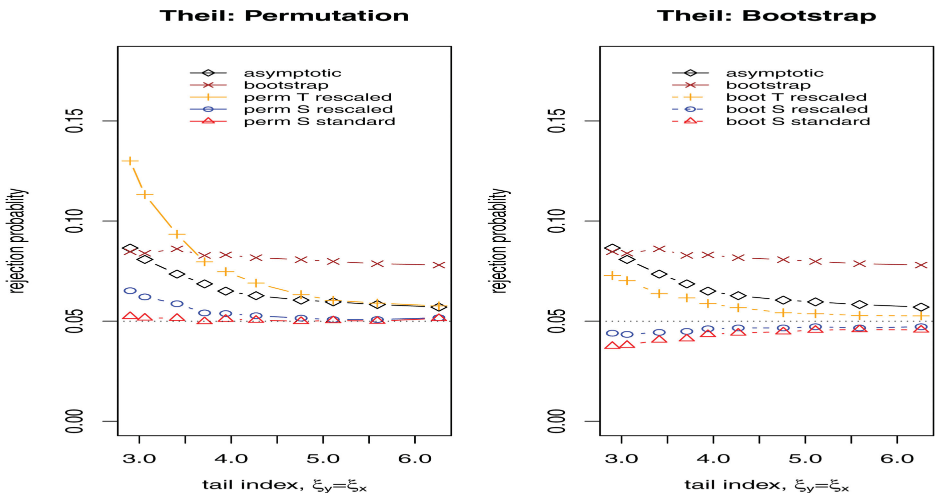

Figure 1. Size: Rejection frequencies of asymptotic, permutation, and bootstrap tests for the problem of testing the equality of Theil inequality measures between two samples. The two distributions are identical, $\mathbf{F_x} = \mathbf{F_y}$. The upper tail is heavier as ξ_y decreases, with $\xi_y = \xi_x \in [2.9; 6.26]$ and $n = m = 50$.

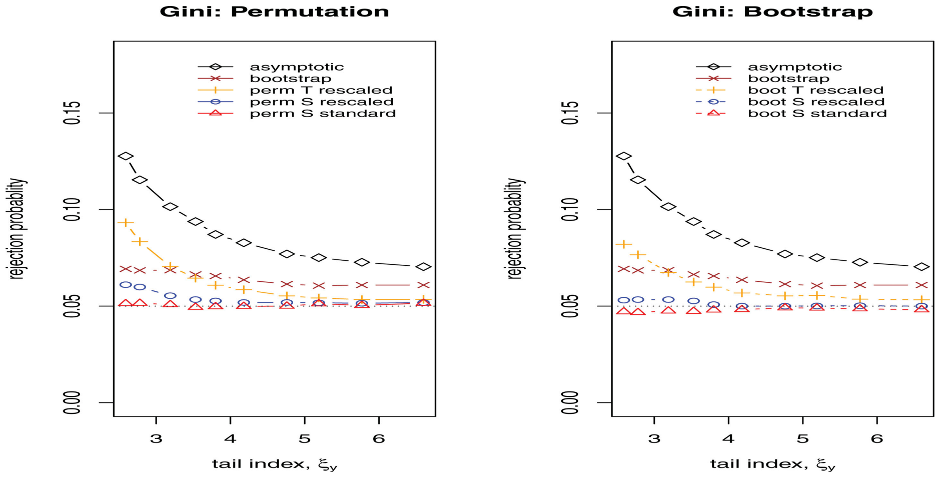

Figure 2. Size: Rejection frequencies of asymptotic, permutation, and bootstrap tests for the problem of testing the equality of Gini inequality measures between two samples. The two distributions are identical, $\mathbf{F_x} = \mathbf{F_y}$. The upper tail is heavier as ξ_y decreases, with $\xi_y = \xi_x \in [2.59; 6.6]$ and $n = m = 50$.

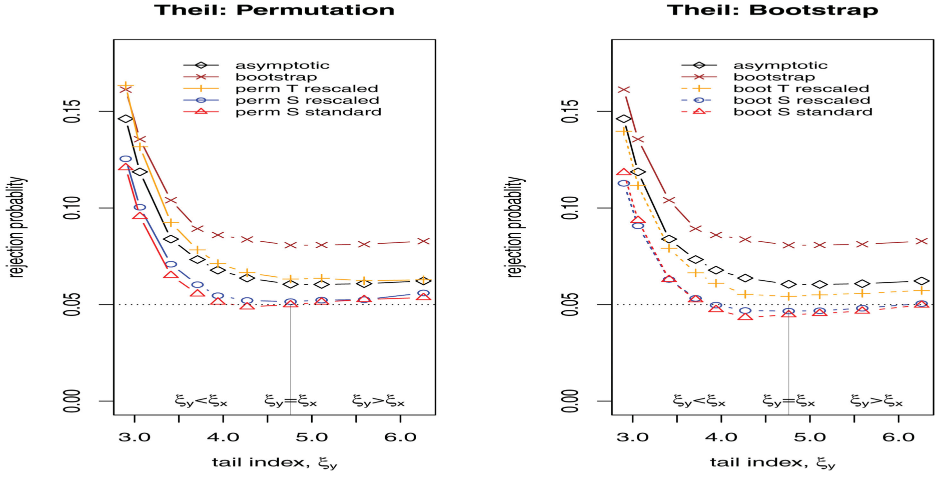

Figure 3. Size: Rejection frequencies of asymptotic, permutation, and bootstrap tests for the problem of testing the equality of Theil inequality measures between two samples. The distribution F_x is fixed ($\xi_x = 4.76$) and $\mathbf{F_x} \neq \mathbf{F_y}$. The distribution F_y goes away from F_x, being heavier tailed as ξ_y decreases, with $\xi_y \in [2.9; 6.26]$ and $n = m = 50$.

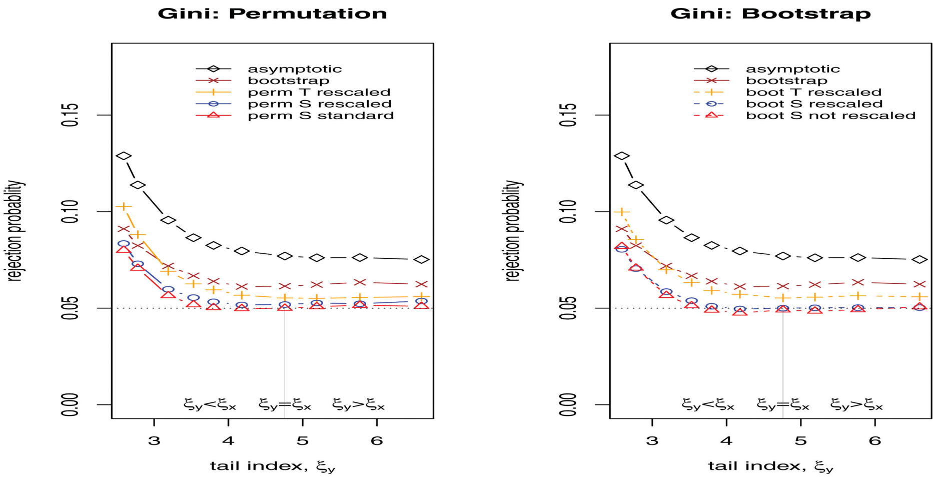

Figure 4. Size: Rejection frequencies of asymptotic, permutation, and bootstrap tests for the problem of testing the equality of Gini inequality measures between two samples. The distribution F_x is fixed ($\xi_x = 4.76$) and $\mathbf{F_x} \neq \mathbf{F_y}$. The distribution F_y goes away from F_x, being heavier tailed as ξ_y decreases, with $\xi_y \in [2.59; 6.6]$ and $n = m = 50$.

as the distribution F_y moves away from F_x. The distribution F_x is fixed, with a tail index $\xi_x = 4.76$, while F_y has varying tail indices. When the tail index of F_y is smaller (higher) than that of F_x, that is, when $\xi_y < \xi_x$, F_y is more (less) heavy-tailed than F_x. Figure 4 shows similar results for the Gini index. From these figures, we can see that the results deteriorate when F_y tends to be much more heavy-tailed than F_x, that is, when $\xi_y < 3.5$. Overall, permutation and bootstrap under the null tests based on the studentized statistic (perm S rescaled, perm S standard, boot S rescaled, and boot S standard) perform similarly and they outperform other methods. They perform very well when $\xi_y > 3.5$, that is, when F_y is not much more heavy-tailed than F_x.

6.2.3. Sample Size. Figure 5 shows rejection frequencies for the Theil measure, as the sample size increases ($n = m = 50, \ldots, 10{,}000$), with identical distributions ($F_x = F_y$), in the worst case previously studied ($\xi_x = \xi_y = 2.9$). We can see that the rejection frequencies decrease slowly as the sample size increases with asymptotic tests, and, even more slowly with standard bootstrap tests. In contrast, permutation tests and bootstrap under the null perform very well in all cases when they are based on studentized statistic: rejection frequencies are always almost equal to 0.05 for perm S rescaled, perm S standard, boot S rescaled, and boot S standard. (We obtain similar results for the Gini index (results not reported).)

Figure 6 shows rejection frequencies for the Theil measure, as the sample size increases ($n = m = 50, \ldots, 10{,}000$) with different distributions ($F_x \neq F_y$), in the worst cases previously studied ($\xi_x = 4.76$, $\xi_y = 2.9$). We can see that, for each method, the rejection frequencies decrease very slowly as the sample size increases. Moreover, permutation and bootstrap under the null tests based on a studentized statistic outperform other methods. (We obtain similar results for the Gini index (results not reported).)

Figure 7 shows rejection frequencies for the Theil measure, with unequal sample sizes ($n = 100, 110, \ldots, 190$

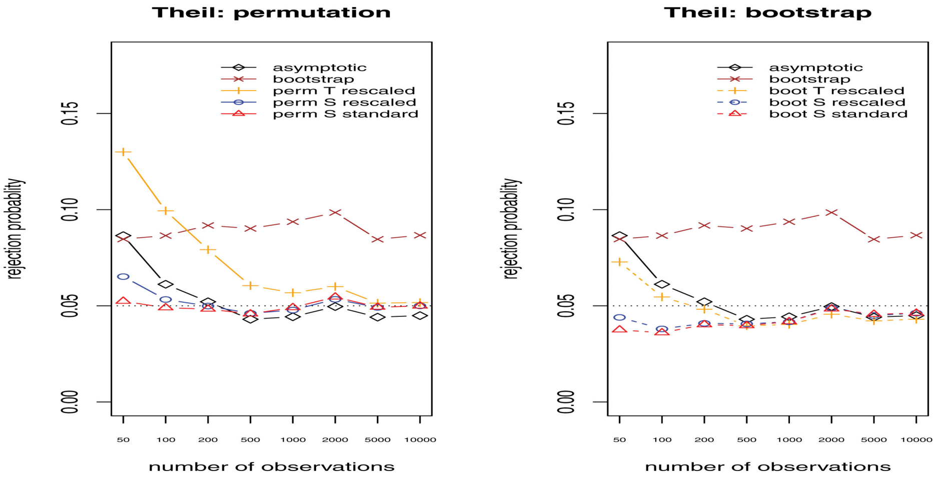

Figure 5. Size: Rejection frequencies of asymptotic, permutation, and bootstrap tests for the problem of testing the equality of Theil inequality measures between two samples, as the sample size increases. The two distributions are identical, $\mathbf{F_x} = \mathbf{F_y}$, and very heavy-tailed, $\xi_y = \xi_x = 2.9$.

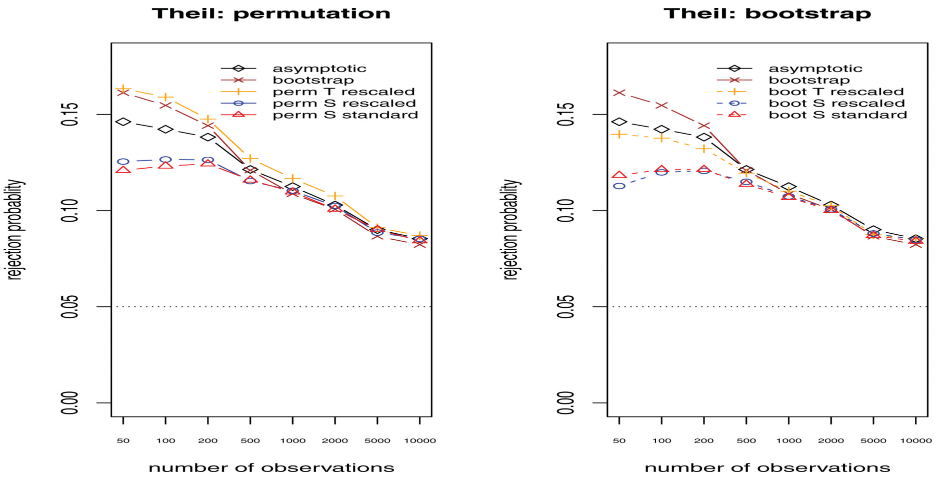

Figure 6. Size: Rejection frequencies of asymptotic, permutation, and bootstrap tests for the problem of testing the equality of Theil inequality measures between two samples, as the sample size increases. The two distributions are very different in their upper tails, $\mathbf{F_x} \neq \mathbf{F_y}$, with tail parameters equal to $\xi_x = 4.76$ and $\xi_y = 2.9$.

and $m = 200 - n$), in the worst case previously studied ($\xi_x = \xi_y = 2.9$). The x-axis goes from equal sizes ($n = m = 100$) to extremely unequal sizes ($n = 190$, $m = 10$). With identical distributions (left panel: $\mathbf{F_x} = \mathbf{F_y}$), we can see that over-rejections increase quickly with asymptotic and bootstrap tests, as the sample sizes are more unequal. By contrast, permutation tests perform very well in all cases when they are based on studentized statistic. With different distributions (right panel: $\mathbf{F_x} \neq \mathbf{F_y}$), over-rejections increase quickly with asymptotic and bootstrap tests, as the sample sizes are more unequal. They increase slowly with permutation tests. Overall, we can see that permutation tests outperform asymptotic and standard bootstrap tests, with unequal sample sizes. They perform very well when the two distributions are similar.

6.3 Power

To study the power, we test the equality of an inequality measure between two samples, when the samples come from two distributions with different values of the inequality measure. From the study on the size, studentized permutation and bootstrap under the null tests outperform other methods. They also perform similarly when the null hypothesis is true, we can thus compare power between these methods.

In our experiments, the distribution F_x is fixed and the distribution F_y varies:

$$F_x = \mathrm{SM}(x; 2.8, 100^{-\frac{1}{2.8}}, 1.7) \text{ and}$$

$$F_y = \mathrm{SM}(y; 2.8, 100^{-\frac{1}{2.8}}, q), \quad q \in [0.7; 31.7]. \qquad (60)$$

As q increases, the tail index and the inequality measure increase. (We take $q = 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3, 1.4, 1.5, 1.7, 1.9, 2.2, 2.7, 3.7, 5.7, 31.7$ from which we have the tail indices $\xi = 2.8\,q \in [1.96; 88.76]$. The true null hypothesis is, respectively, equal to $H_0 : \theta(F_y) - \theta(F_x) = 0.328, 0.21, 0.143, 0.1, 0.071, 0.051, 0.035, 0.023, 0.014, 0, -0.01, -0.02, -0.03, -0.041, -0.049, -0.06$ for the Theil index.) Figure 8 shows rejection frequencies for

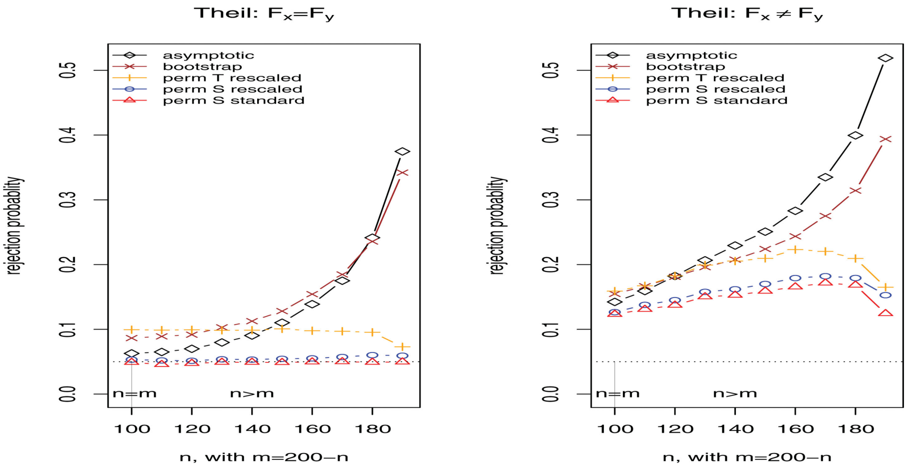

Figure 7. Size: Rejection frequencies of asymptotic, permutation, and bootstrap tests for the problem of testing the equality of Theil inequality measures between two samples of unequal sizes, $m = 200 - n$.

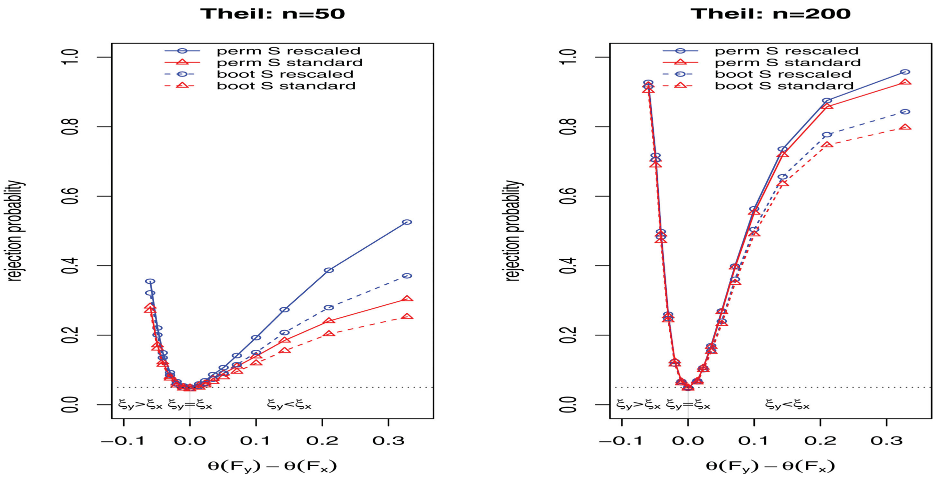

Figure 8. Power: Rejection frequencies of permutation and bootstrap tests for the problem of testing the equality of Theil inequality measures between two samples, when the true null hypothesis is equal to $\theta(F_y) - \theta(F_x)$. The distribution F_x is fixed and the distribution F_y is heavier tailed as $\theta(F_y) - \theta(F_x)$ increases.

testing the equality of the Theil measure between two samples, when the true null hypothesis, $\theta(F_y) - \theta(F_x)$, goes away from 0. We consider a small sample ($n = m = 50$, on the left) and a moderate sample ($n = m = 200$, on the right). We consider different cases:

- $\theta(F_y) - \theta(F_x) = 0$: the two distributions are identical and H_0 is true (size);
- $\theta(F_y) - \theta(F_x) \neq 0$: the two distributions are different and H_0 is not true (power);
- $\theta(F_y) - \theta(F_x) < 0$: F_y is less heavy-tailed than F_x ($\xi_y > \xi_x$);
- $\theta(F_y) - \theta(F_x) > 0$: F_y is more heavy-tailed than F_x ($\xi_y < \xi_x$).

Power comparison of the considered permutation and bootstrap methods are valid since rejection probabilities under the null hypothesis $\theta(F_y) - \theta(F_x) = 0$ are close to the nominal level (here 0.05), which in Figure 8 is represented via the dashed horizontal line.

From Figure 8, we can see that the curves are asymmetric around 0. When F_y is less heavy-tailed than F_x ($\theta(F_y) - \theta(F_x) < 0$), the null is quickly rejected as the true null hypothesis moves away from 0. On the other side, when F_y is more heavy-tailed than F_x ($\theta(F_y) - \theta(F_x) > 0$), the null is slowly rejected as the true null hypothesis moves away from 0. Overall, we can see that the permutation approach (perm S rescaled and standard) is more powerful than the bootstrap under the null approach (boot S rescaled and standard), the difference between the two approaches being resampling without replacement rather than with replacement. In addition, the studentized permutation tests based on the combined rescaled samples (perm S rescaled) outperform other methods. It rejects the null much more faster than other methods, especially when F_y is heavier-tailed than F_x ($\theta(F_y) - \theta(F_x) > 0$).[6.2.3]

6.4 Dependent Samples

It is often of great interest to compare inequality levels between pretax and post-tax income distributions, for instance to measure the impact of a specific taxation scheme on inequality. In that case, the two samples are correlated and of equal size, with the same individuals in both samples. Bootstrapping and permuting should be modified to take into account such dependencies:

- A bootstrap sample is obtained by making n draws by pairs with replacement from the n observed incomes, where each pairs (X_i, Y_i) has probability $1/n$ of being selected on each draw.
- A permuted sample is obtained by permuting elements within pairs, that is, by permuting X_i and Y_i (or $X_i/\bar{X}$ and $Y_i/\bar{Y}$ for the case of rescaling), with probability 1/2, for $i = 1, \ldots, n$.

The denominator of the studentized statistic, in (3), (15), and (18), needs also to be modified to take into account the dependence (matched-pair) between the samples. (The denominator in (3) becomes $\{\hat{V}[\theta(\hat{F}_x)] + \hat{V}[\theta(\hat{F}_y)] - 2\epsilon_{xy}/n\}^{1/2}$. For the Theil index, we have: $\epsilon_{xy} = \frac{1}{\hat{\mu}_x\hat{\mu}_y}\{(\frac{\hat{\nu}_x}{\hat{\mu}_x} + 1)(\frac{\hat{\nu}_y}{\hat{\mu}_y} + 1)\hat{\mu}_x\hat{\mu}_y - (\frac{\hat{\nu}_x}{\hat{\mu}_x} + 1)\hat{\mu}_x\hat{\nu}_y - (\frac{\hat{\nu}_y}{\hat{\mu}_y} + 1)\hat{\nu}_x\hat{\mu}_y + \hat{\nu}_x\hat{\nu}_y\}$ and, $\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^n x_i$, $\hat{\nu}_x = \frac{1}{n}\sum_{i=1}^n x_i \log x_i$, $\hat{\mu}_y = \frac{1}{n}\sum_{i=1}^n y_i$, $\hat{\nu}_y = \frac{1}{n}\sum_{i=1}^n y_i \log y_i$, see Zheng and Cushing (2001).)

Figure 9 shows size results, that is, empirical rejection frequencies for asymptotic, standard bootstrap and permutation tests for the Theil index in small samples ($n = 50$), when the two distributions are identical (left panel: $\mathbf{F_x} = \mathbf{F_y}$) and different (right panel: $\mathbf{F_x} \neq \mathbf{F_y}$). The correlation between the two samples is generated using a Gumbel copula, and is very strong: the Kendall correlation coefficient is high, $\rho = 0.8$. We can see from this figure that the permutation t-tests outperform other methods and that they perform very well when F_x is not too far from F_y. Overall, the simulation results are quite similar to those obtained previously in case of independent samples, in left panels of Figures 1 and 3, with less distortions in the case of identical distributions and more distortions in the case of different distributions.

Figure 10 shows power results, that is, rejection frequencies for testing the equality of the Theil index between two dependent samples, when the true null hypothesis, $\theta(F_y) - \theta(F_x)$, goes

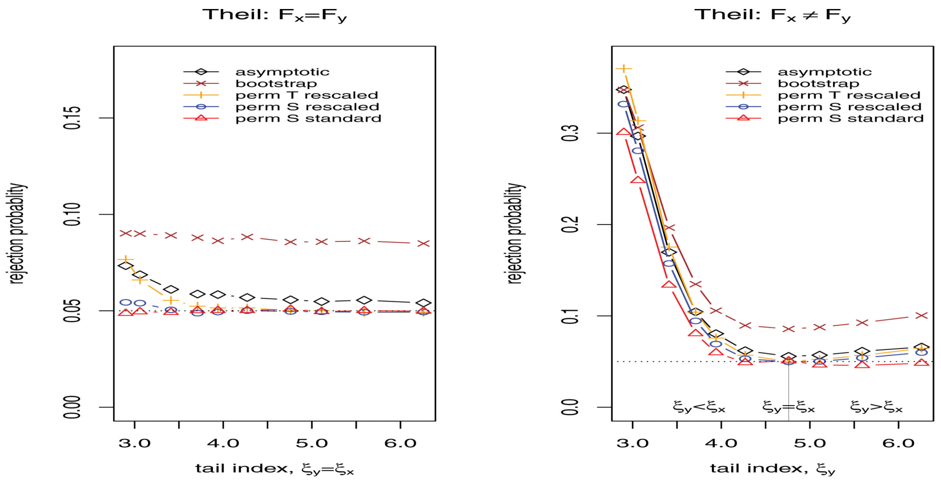

Figure 9. Size: Rejection frequencies of asymptotic, permutation, and bootstrap tests for the problem of testing the equality of Theil inequality measures between two dependent samples, $n = 50$.

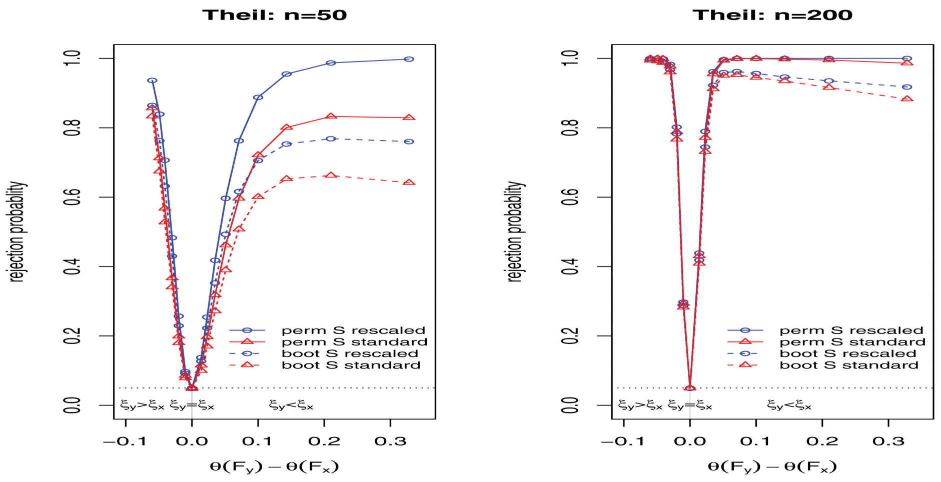

Figure 10. Power: Rejection frequencies of permutation and bootstrap tests for the problem of testing the equality of Theil inequality measures between two dependent samples, when the true null hypothesis is equal to $\theta(F_x) - \theta(F_y)$. The distribution F_x is fixed and the distribution F_y is heavier tailed as $\theta(F_y) - \theta(F_x)$ increases.

away from 0. We consider a small sample (left panel: $n = 50$) and a moderate sample (right panel: $n = 200$). Overall, the simulation results are quite similar to those obtained previously in case of independent samples, in Figure 8. The studentized permutation tests based on the combined rescaled samples (perm S rescaled) outperform other methods in small samples. It rejects the null much more faster than other methods, especially when F_y is heavier-tailed than F_x ($\theta(F_y) - \theta(F_x) > 0$).

It is interesting to note that, if a taxation scheme is proportionate (the amount of the tax is a fixed proportion of incomes), the pretax and post-tax distributions are identical for the Theil and MLD indices (because of the scale independence property). In such cases, inference should then be (nearly) exact with studentized permutation tests, even if the income distributions are heavy-tailed.

The above results suggest that further extensions relaxing the iid assumption hold credible promise. One may consider, in particular, non-iid observations within each sample as occurs with stratified or clustered survey-based data; see Deaton (1997) and Bhattacharya (2005, 2007). Although beyond the scope of our article, permutation and bootstrap schemes that account for such dependencies are a useful research direction.

7. CONCLUSION

We study Monte Carlo permutation and bootstrap methods for the problem of testing the equality of inequality measures between two samples. For scale-independent measures, as the Gini, Theil, Generalized Entropy, and Atkinson indices, we introduce a convenient rescaling to validate and enhance performance. Our simulation results show that permutation tests control size regardless of tail thickness, when underlying distributions are not too distant (with respect to scale). When underlying distributions differ substantially in their upper tails, proposed permutation methods still provide significant improvement over standard asymptotic and bootstrap tests. In addition, results suggest that rescaling observations by sample means before permutation improves power in finite samples.

ACKNOWLEDGMENTS

The authors thank an anonymous referee and the editor for several useful comments.

FUNDING

This work was supported by the William Dow Chair in Political Economy (McGill University), the Bank of Canada (Research Fellowship), a Guggenheim Fellowship, a Konrad-Adenauer Fellowship (Alexander-von-Humboldt Foundation, Germany), the Institut de finance mathématique de Montréal (IFM2), the Social Sciences and Humanities Research Council of Canada, the Fonds de recherche sur la société et la culture (Government of Québec), the Institut Universitaire de France and the projects ANR-11-IDEX-0001-02 and ANR-16-CE41-0005, managed by the French National Research Agency (ANR).

[Received February 2016. Revised January 2017.]

REFERENCES

Atkinson, A. B. (1970), "On the Measurement of Inequality," *Journal of Economic Theory*, 2, 244–263. [1,7]

Bhattacharya, D. (2005), "Asymptotic Inference From Multi-Stage Samples," *Journal of Econometrics*, 126, 145–171. [13]

——— (2007), "Inference on Inequality From Household Survey Data," *Journal of Econometrics*, 137, 674–707. [13]

Boos, D. D., and Brownie, C. (2004), "Comparing Variances and Other Measures of Dispersion," *Statistical Science*, 19, 571–578. [6]

Boos, D. D., Janssen, P., and Veraverbeke, N. (1989), "Resampling From Centered Data in the Two-Sample Problem," *Journal of Statistical Planning and Inference*, 21, 327–345. [7]

Chung, E., and Romano, J. P. (2013), "Exact and Asymptotically Robust Permutation Tests," *Annals of Statistics*, 41, 484–507. [2,5,8]

Cowell, F. A. (2011), *Measuring Inequality*, Oxford: Oxford University Press. [7]

Cowell, F. A., and Flachaire, E. (2007), "Income Distribution and Inequality Measurement: The Problem of Extreme Values," *Journal of Econometrics*, 141, 1044–1072. [1,8]

——— (2015), "Statistical Methods for Distributional Analysis," in *Handbook of Income Distribution* (Vol. 2), eds. A. B. Atkinson, and F. Bourguignon, New York: Elsevier, pp. 359–465. [8]

Davidson, R. (2009), "Reliable Inference for the Gini Index," *Journal of Econometrics*, 150, 30–40. [1,8]

——— (2012), "Statistical Inference in the Presence of Heavy-Tails," *Econometrics Journal*, 15, C31–C53. [1]

Davidson, R., and Flachaire, E. (2007), "Asymptotic and Bootstrap Inference for Inequality and Poverty Measures," *Journal of Econometrics*, 141, 141–166. [1,8]

Deaton, A. S. (1997), *The Analysis of Household Surveys*, Baltimore, Maryland: Johns Hopkins Press for the World Bank. [13]

Dufour, J.-M. (2006), "Monte Carlo Tests With Nuisance Parameters: A General Approach to Finite Sample Inference and Nonstandard Asymptotics," *Journal of Econometrics*, 133, 443–477. [2,4]

Dwass, M. (1957), "Modified Randomization Tests for Nonparametric Hypotheses," *Annals of Mathematical Statistics*, 28, 181–187. [2,3,4]

Gill, R. (1988), "Non- and Semiparametric Maximum Likelihood Estimators and the Von-Mises Method (part I)," *Scandinavian Journal of Statistics*, 16, 97–128. [5]

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion* (*Springer Series in Statistics*), New York: Springer Verlag. [2]

Kleiber, C., and Kotz, S. (2003), *Statistical Size Distributions in Economics and Actuarial Sciences* (Wiley Series in Probability and Statistics), New York: Wiley. [8]

Lim, T.-S., and Loh, W.-Y. (1996), "A Comparison of Tests of Equality of Variances," *Computational Statistics and Data Analysis*, 22, 287–301. [6]

McDonald, J. B. (1984), "Some Generalized Functions for the Size Distribution Income," *Econometrica*, 52, 647–663. [8]

Pitman, E. J. G. (1937), "Significance Tests Which May be Applied to Samples From any Populations," *Journal of the Royal Statistical Society*, Series A, 4, 119–130. [2]

Romano, J. P. (1990), "On the Behavior of Randomized Tests Without a Group Invariance Assumption," *Journal of the American Statistical Association*, 85, 686–692. [2,5]

Schluter, C., and Trede, M. (2002), "Tails of Lorenz Curves," *Journal of Econometrics*, 109, 151–166. [8]

Schluter, C., and van Garderen, K. J. (2009), "Edgeworth Expansions and Normalizing Transforms for Inequality Measures," *Journal of Econometrics*, 150, 16–29. [1]

Zheng, B., and Cushing, B. C. (2001), "Statistical Inference for Testing Inequality Indices With Dependent Samples," *Journal of Econometrics*, 101, 315–335. [12]