The Journal of Economic Inequality (2023) 21:899–924
https://doi.org/10.1007/s10888-023-09574-w

Assumption-light and computationally cheap inference on inequality measures by sample splitting: the Student *t* approach

Catarina Midões[1,2] · **Denis de Crombrugghe**[3,4]

Received: 18 December 2020 / Accepted: 2 March 2023 / Published online: 24 July 2023

Abstract
Inference on inequality indices remains challenging, even in large samples. Heavy right tails in income and wealth distributions hinder the quality and threaten the validity of asymptotic approximations to finite sample distributions. Attempts to improve on asymptotic approximations by bootstrap techniques or permutation tests are only partial successes. We evaluate a different approach to robust inference, relying on Student t statistics obtained from split samples. This relatively simple 't-based' approach requires no consistent variance estimators, no random sampling of populations, and only mild distributional assumptions. We compare its performance with that of refined bootstrap and permutation techniques. We find that the more complex bootstrap methods still have the edge in one-sample tests, where the t-approach suffers from a negative skew. In two-sample comparisons though, the t-approach offers advantages: it is undersized while bootstrap tests and permutation tests are often oversized. In certain circumstances it is less powerful than permutation tests and bootstrap tests, but for large samples, this difference dissipates. It is also more generally applicable than permutation tests and easily generates confidence intervals. These differences are illustrated with an empirical application using two different sources of household data from the Russian Federation.

Keywords Inference on inequality measures · Difference-in-inequality testing · Bootstrap inference · Permutation tests · Sample splitting

JEL Classification C12 · C46 · D63

✉ Catarina Midões
catarina.midoes@unive.it; mcorreia.catarina@gmail.com

Denis de Crombrugghe
denis.decrombrugghe@nu.edu.kz

1 Institute of Environmental Science and Technology of the Universitat Autònoma de Barcelona (ICTA-UAB), Barcelona, Spain

2 Ca' Foscari University of Venice, Venice, Italy

3 Nazarbayev University, Astana, Kazakhstan

4 Maastricht University, Maastricht, Netherlands

1 Introduction

Practical methods of inference on inequality indices remain in high demand, especially for studying changes and differences in income or wealth distributions. There have been several useful contributions recently, but these are not always versatile and easy to apply. Many reports still choose to offer only point estimates when studying differences in inequality (for example, OECD Reports such as Cohen and Ladaique (2018) or the World Inequality Reports[1]). However, due to the relative frequency of extremely high observations, point estimates behave erratically, even when based on well designed and large samples, thereby undermining confidence in conclusions about changes or differences in inequality. Furthermore, to properly identify the causes and consequences of inequality as well as to evaluate the impact of targeted policies, it is necessary to account for sampling variability. If alleviating poverty and decreasing inequality are indeed fundamental economic, social and political goals, accurate measurement in general and inference in particular must be made a priority.

Unfortunately, the issue is far from straightforward. The sensitivity of inequality measures to extreme observations hinders the construction of confidence intervals and the performance of hypothesis testing. In this context, asymptotic methods are flawed even in large samples, and standard bootstrap approaches bring only modest improvements. Refinements have been developed by Cowell and Flachaire (2007) and Davidson and Flachaire (2007), hereafter C&F and D&F. The *moon* bootstrap and semi-parametric bootstrap they proposed can be substantially more successful in controlling for size in one-tailed hypothesis testing. However, both methods are relatively complex, require additional assumptions, and entail the choice of several parameters.

They also do not seem to yield noteworthy improvements over asymptotic or standard bootstrap tests when testing for inequality differences between two populations, a situation of frequent interest in empirical analyses. The actual size of the tests tends to exceed the nominal level more than in one-sample hypothesis testing, increasing the need for alternative solutions.

More recently, Dufour et al. (2019) have proposed Monte Carlo permutation tests, arguably a bootstrap variant, for inequality comparisons. Observations from the two populations to be compared are mixed and repeatedly rearranged in a random order so as to constitute artificial comparison samples. If the two population distributions are identical, mixing up the observations drawn from them will not affect the sampling distribution of comparison statistics like a studentised difference between two indices. If the two population distributions differ, the inequality measures of interest may still be equal, and permutations tests may still be asymptotically valid, but only under specific conditions. Dufour et al. (2019) analyse those conditions and conduct simulation experiments to study the finite-sample properties of various implementations of permutation tests. Their results indicate good performance, and most notably a significant improvement over standard asymptotic and bootstrap tests, even in unfavourable situations like the comparison of two distributions with very differently shaped upper tails. Still, the main shortcoming of the permutation approach remains that it actually tests the equality of distributions rather than the equality of inequality indices.

Recognising that current methods are not entirely satisfactory, both from a statistical and a practical perspective, we investigate in this paper a simpler and promising testing method, put forward by Ibragimov and Müller (2010) and adapted to difference-in-means testing by Ibragimov and Müller (2016)). It is based on splitting the data in a small number of independent sub-samples or 'groups' and basing Student t statistics on the empirical distri-

[1] https://wir2022.wid.world/

bution of the group estimates, hence the moniker 't-based method'. Unlike its asymptotic and bootstrap contenders, the strategy entirely circumvents the need for consistent variance estimators, offering advantages in dealing with potentially heteroskedastic and correlated data. Unlike permutation tests, it can deal with all sorts of hypotheses and it can quickly generate confidence intervals by test inversion. In the case of group or cluster heteroskedasticity, the method is conservative, in the sense that the effective test size will remain below nominal levels.[2] The method has been applied in the field of inequality measurement by Ibragimov and Ibragimov (2018) and recently in a working paper by Ibragimov et al. (2021) focusing on robustness.

In this paper we offer an assessment of the finite-sample performance and advantages of the Student t-based approach for the purpose of conducting inference about inequality indices. Through Monte Carlo simulations with both random and correlated samples, we confirm previous results, evaluate the size and power of the t-based tests in both one-sample and two-sample problems, and make comparisons with the competing methods.

Beyond confirming results of Ibragimov et al. (2021) under several modifications, we complement those in various ways. For both one-sample and two-sample tests, we expand the range of methods: we examine three different implementations of permutation tests, and also three different bootstrap-based tests. In the case of one-sample tests, we treat left-tailed, right-tailed, and two-tailed tests separately, and we include in our comparisons heavy-tailed distributions relevant in the inequality context, such as the Pareto distribution.

In the case of two-sample tests, we generate new results in a number of directions. First, we expand the range of sample sizes to analyse test size and power differences between methods in realistically large samples. The expectation is that the conservativeness of the t-approach may give it an edge over alternative comparison tests, even when very large surveys are available. By contrast, permutation tests may still over-reject in samples of sizes like 20,000 or even 100,000. In practical work comparing income inequality between, for example, different countries or the same country in different years, such large sample sizes are not exceptional.

Second, we consider how power depends not only on sample size but also on the sample split, and offer guidance as to how many groups to consider when dealing with relatively small or very large samples.

Third, we consider the relative performance of comparison methods under dependence structures between the two samples. To do so, we model positive upper and lower tail dependence through copulas of heavy-tailed and non-heavy-tailed distributions. We conclude that different correlation structures can alter the performance of permutation tests very substantially, while t-based tests are not nearly as sensitive to them.

Fourth, we show how to quickly build confidence intervals, of great interest in practical work, and illustrate their use, alongside the practical feasibility of the method(s), by pairing two sources of household survey data from the Russian Federation. We compare surveys by the official Rosstat federal agency and by the academically managed RLMS-HSE. The academic RLMS-HSE is barely 1/10th the size of the official Rosstat survey, which includes 60,000 households, but in principle they are both representative of the entire population of the same federal territory, so it is interesting to find out how congruent the characteristics of the respective income distributions are.

We conduct inference on key inequality indices using the 2020 data from both sources, and run various comparison tests within and between them. We find contradictions between t-based and permutation tests, where only one of the latter (the most oversized one according to

[2] This property follows from an important result of Bakirov and Székely (2006).

our simulations) rejects the null of equal Theil indices. Assuming the surveys are appropriately built, there is reason to suspect a rejection error. Moreover, when we focus on Russia's main two cities, Moscow and St-Petersburg, where there is much less ground to expect systematic differences between the surveys' coverage than in the Federation's full territory, all the permutation tests reject the null of equality, whereas the t-based tests do not. This further raises the suspicion that rejections by permutation tests might be due to an excessive test size.

The paper is organised as follows. Section 2 briefly outlines our experimental framework: the distributions we simulate, both for one-sample and two-sample problems, and the inequality measures we calculate from them. Section 3 summarises the different hypothesis testing methodologies for one-sample problems, and Section 4 does the same for two-sample (comparison) problems. The methods are briefly referenced, their shortcomings are highlighted, and the t-approach is formally presented. Section 5 reports the main results of our simulation experiments, comparing the size and power of the different tests in both one-sample and two-sample problems. In Section 6, the empirical application pairing two different sources of Russian household income data illustrates the use and flexibility of the proposed methods. Section 7 provides a discussion of the results and Section 8 concludes.

2 Framework

2.1 Distributions

We will consider three different families of distributions in our simulation experiments: the log-normal, the Pareto, and the Singh-Maddala, with the latter functioning as the baseline case. These have been chosen for their prominent role in the inequality literature (Cowell 2009), their good fit to actual income distributions, and their flexibility in generating extreme values (in the case of the Pareto and Singh-Maddala distributions); last but not least, they ensure comparability with the previous literature.

For reference, we specify here the parameterisation of these distributions in terms of their cumulative distribution function (CDF) $F(y;\pi)$ or probability density function (PDF) $f(y;\pi)$, where π stands in for parameters. A random variable obeys our baseline Singh-Maddala distribution if it has the CDF

$$F(y;a,b,c) = 1 - \frac{1}{(1+ay^b)^c} \tag{1}$$

with parameters $a = 100$, $b = 2.8$ and $c = 1.7$. The exact same distribution is used as a baseline by among others C&F, D&F, and Dufour et al. (2019). Like C&F, we also use $c = 1.2$ and $c = 0.7$ to swell the right tail further. Moments of order up to r exist if and only if $bc > r$ (Tadikamalla 1980).

As one alternative to Singh-Maddala, a Pareto (Type I) distribution has the PDF

$$f(y;\underline{y},\theta) = \frac{\theta \underline{y}^{\theta}}{y^{\theta+1}} \quad , \qquad y \geq \underline{y} \geq 0 \, , \quad \theta \geq 0 \, . \tag{2}$$

Our simulations, in keeping with C&F, set the threshold parameter $\underline{y} = 0.1$ and the shape parameter $\theta \in \{1.5,\ 2.0,\ 2.5\}$. The variance is finite for $\theta > 2$.

A second alternative is the log-normal distribution, with the following PDF:

$$f(y;\mu,\sigma) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log y - \mu)^2}{2\sigma^2}\right) \tag{3}$$

The parameter values used are $\mu = -2$, $\sigma \in \{0.5.\ 0.7,\ 1.0\}$, as in C&F. Log-normal distributions are not heavy-tailed, in the sense that their tails do not decay as power functions but exponentially. Table 1 collects our parameter choices.

Two-sample problems require a pair of distributions which may or may not be the same. We adopt D&F's choice of two Singh-Maddala distributions: distribution A with parameters $a = 100$, $b = 2.8$, $c = 1.7$; and distribution B with parameters $a = 100$, $b = 4.8$, $c = 0.636659$. The latter value of c is chosen so that both distributions have the same *Theil* index, namely 0.140115, even though they have very different *tails*, as visible in Fig. 1. For some tests, we also use a log-normal distribution B with parameters $\mu = -2$ and $\sigma = 0.5293678$, again achieving the same Theil index of 0.140115

2.2 Inequality measures

Generalised Entropy measures are written as follows, for an income or wealth distribution with CDF F and moments $\mu_\alpha(F) \equiv E[Y^\alpha] = \int_0^\infty y^\alpha \, \mathrm{d}F(y)$:

$$GE_\alpha(F) = \begin{cases} \frac{1}{\alpha(\alpha+1)} \left[\frac{\mu_\alpha(F)}{\mu_1(F)} - 1\right], & \text{for } \alpha \notin \{0,1\} \\ -\int_0^\infty \log\left(\frac{y}{\mu_1(F)}\right) \mathrm{d}F\,(y)\ , & \text{for } \alpha = 0 \quad \text{(MLD)} \\ \int_0^\infty \frac{y}{\mu_1(F)} \log\left(\frac{y}{\mu_1(F)}\right) \mathrm{d}F\,(y)\ , & \text{for } \alpha = 1 \quad \text{(Theil)} \end{cases} \tag{4}$$

GE_α is a smooth function of the moments $\mu_1(F)$ and $\mu_\alpha(F)$, and its analog estimator will be asymptotically normal provided the sample moments are. For a Central Limit Theorem (CLT) to apply to the sample moments, these must have finite first and second moments themselves

Table 1 Parameters for simulation study

	Parameters
Singh-Maddala:	$a = 100$, $b =$**2.8**, $c \in \{0.7, 1.2,$ **1.7**$\}$ or
	$a = 100$, $b = 4.8$, $c \approx 0.6367$ (sample B)
Lognormal:	$\mu = -2$, $\sigma \in \{0.5, 0.7, 1\}$ or
	$\mu = -2$ and $\sigma \approx 0.5294$ (sample B)
Pareto:	$\underline{y} = 0.1$, $\theta \in \{1.5, 2, 2.5\}$

In bold, the parameters in the baseline case of the simulations. For two-sample simulations, we resort to the bolded Singh-Maddala distribution as distribution A and to either the Singh-Maddalla sample B or the log-normal sample B. There is no closed expression for the Theil index under a Singh-Maddala and thus neither for c and σ used in the simulations. The exact values used in the simulations are those which minimize the difference between $GE_1(F^A)$ (≈ 0.1401151) and $GE_1(F^B)$ where F^A is the CDF of a Singh-Maddala distribution with the parameters in bold, and F^B is the CDF of a Singh-Maddala with parameters $b = 4.8$ and c or the CDF of a lognormal distribution with parameter σ

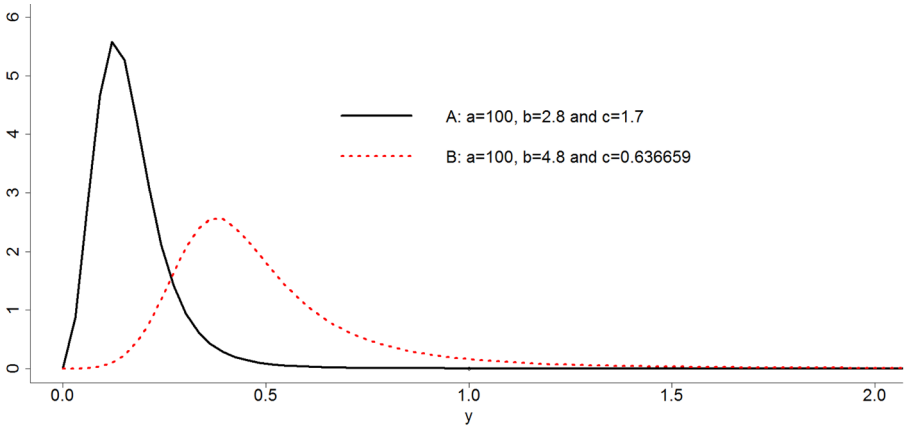

Fig. 1 Singh-Maddala distributions used in our experiments

– requiring finite 2^{nd} and $2\alpha^{th}$ moments of the underlying income or wealth distribution. Attention is mostly focused on the Theil index GE_1 and on the Mean Logarithmic Deviation GE_0 (MLD). GE_2 is a case of extreme sensitivity to inequality at the top of the distribution.

Although lacking some of the desirable properties of the Generalised Entropy class, the Gini index remains the most popular inequality measure. For a distribution with CDF F it is defined as

$$G \equiv \frac{2}{\mu_1(F)} \int_0^\infty yF(y)dF(y) - 1 \tag{5}$$

We use the sample counterpart

$$\widehat{G} = \frac{1}{n-1}\left(n+1-\frac{2}{n\widehat{\mu}_1}\sum_{i=1}^{n}(n+1-i)\,y_{(i)}\right) \tag{6}$$

where $y_{(i)}$ is the i^{th} order statistic and $\widehat{\mu}_1$ is the sample mean. The asymptotic normality of $\widehat{G}$ is discussed by Davidson (2009).[3] In random samples, the Gini index can be written as a transformation of a normalised sum of individual contributions to inequality, which are independently and identically distributed (iid) quantities with expectation zero. Asymptotic normality of the analog estimator follows, provided the first and second moments of the income distribution are finite.

The consistent variance estimator proposed by Davidson (2009) is used:

$$\widehat{Var}(\widehat{G}) = \frac{1}{(n\widehat{\mu})^2}\sum_{i=1}^{n}(\widehat{Z}_i - \bar{Z})^2 \tag{7}$$

where

$$\widehat{Z}_i = -(\widehat{G}+1)y_{(i)} + \frac{2i-1}{n}y_{(i)} - \frac{2}{n}\sum_{j=1}^{i} y_{(j)}\,.$$

The calculation of the $\widehat{Z}_i$ makes the variance estimator computationally expensive. This led C&F to limit their simulation experiments to relatively small samples ($n = 100$, 500, and

[3] Davidson (2009) actually analyses the estimator $\tilde{G} = \frac{(n-1)}{n} \times \widehat{G}$, yet this difference does not affect the asymptotic argument.

1,000, whereas the GE_α measures were studied up to $n = 10{,}000$). We extend the simulations of the Gini to $n = 5{,}000$.

3 Inference on inequality in one-sample problems

3.1 Asymptotic inference

Under the finite moments conditions specified in the preceding section, every inequality index under analysis is asymptotically Gaussian. Take the following statistic W:

$$W = \frac{\hat{I} - I}{S_I} \tag{8}$$

where $\hat{I}$ is an in-sample estimator of the inequality measure I and S_I^2 a consistent variance estimator of $\hat{I}$. If $\hat{I}$ is asymptotically Gaussian, $W \xrightarrow{D} \mathcal{N}(0, 1)$, and valid inference can be based on the standard normal distribution.

However, it has been shown thoroughly by D&F and C&F that asymptotic methods perform poorly, even in samples of considerable size (up to 10,000 observations). The distribution of W is negatively skewed, and its convergence to a normal distribution can be very slow. The fat left tail is a reflection of too frequent inequality underestimation. The immediate consequence for hypothesis testing based on W is that left-sided tests are oversized and, conversely, right-sided tests are undersized.

The more sensitive the measure is to extreme high values, the less reliable the asymptotic approximation becomes. It is worse for GE_2 than for GE_1, and for GE_1 than for GE_0. The Gini index seems to converge faster to normality than all three GE_α measures; see Fig. 2, analogous to Fig. 1 in D&F.

The quality of the approximation deteriorates further for more heavy-tailed distributions. If the shape parameter c is decreased the convergence to the standard normal becomes slower still, for all measures considered; this can be seen by comparing the top and bottom panels of Fig. A1 in the online Appendix. GE_2 is actually no longer asymptotically normal when $c = 1.2$ ($bc = 3.36$), and GE_1 when $c = 0.8$ ($bc = 1.96$).

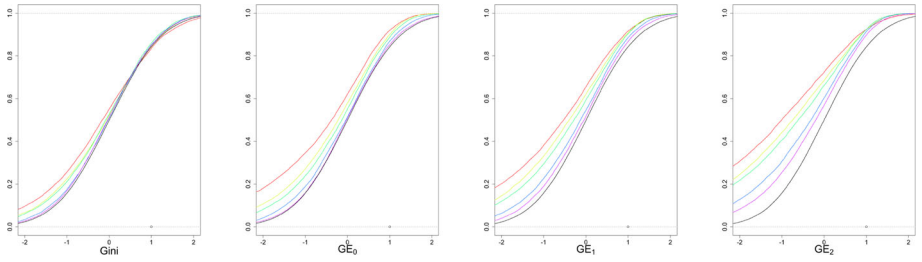

Fig. 2 Distribution of the W statistic for increasing sample sizes. W statistic distribution for the Gini index, the GE_0 (MLD), the GE_1 (Theil) and the GE_2. Approximation based on 10,000 drawings from the Singh-Maddala distribution with $a = 100$, $b = 2.8$, $c = 1.7$. $n = 20$ ___, $n = 50$ ___, $n = 100$ ___, $n = 1000$ ___, $n = 10,000$ ___, Standard Normal distribution $N(0, 1)$ ___

A troublesome conclusion follows: it is for more unequal populations that asymptotic inference poses more severe problems. The inherent difficulty of sampling extremely high values, rare in the population but determinant for inequality, renders inference challenging.

3.2 Bootstrap methods

3.2.1 Bootstrap-*t*

The bootstrap version of W, our quantity of interest, shall be

$$W^* = \frac{\widehat{I}^* - \widehat{I}}{S_I^*},$$

re-centered on the full-sample estimate $\widehat{I}$, to bootstrap under the null hypothesis. As conventional, we use an asterisk to indicate bootstrapped values.

W is, asymptotically, a pivotal quantity, which makes it a suitable candidate for the bootstrap (Beran 1988). This bootstrap based on W - the bootstrap-t - provides higher-order asymptotic approximations to distributions and rejection probabilities of tests (see, for instance, Horowitz (2001)). Whenever asymptotic inference is valid, so are standard bootstrap methods. When asymptotic methods fail, bootstrap methods might fail as well. In the case of inequality, we worry about the underlying income distributions having an infinite variance. In this situation, asymptotic inference fails, and the validity of bootstrap methods is not guaranteed.

3.2.2 *'Moon'* bootstrap

Athreya (1987) demonstrates that the bootstrap-t on the population mean is invalid in some of the specific cases we build, considered illustrative of real income situations. Its distribution does not converge to the same limit as the distribution of the sample mean when the population distribution obeys a power law with stability parameter between 1 and 2 (θ in the Pareto distribution, bc in the Singh-Maddala Distribution).

The m-out-of-n bootstrap can be a remedy for this specific failure of the standard bootstrap. As prescribed by Bickel (1997), if $m/n \to 0$ and $m \to \infty$, the m-out-of-n bootstrap will be consistent for distributions with heavy right tails, and even infinite variances, for which the standard bootstrap methods fail. To bootstrap the W statistic using an m-out-of-n approach, samples of size $m < n$ are taken from the original data with replacement. Still, convergence to the actual sampling distribution can be arbitrarily slow. In the designated '*moon*' bootstrap, following the methods of C&F and D&F, new p-values are constructed, based on the p-values from both the bootstrap-t, $p(n, n)$, and the m-out-of-n bootstrap, $p(m, n)$ with $m = n^{1/2}$. Specifically, for an outcome w,

$$P_{moon}(w) = \Phi(w) + \widehat{a} n^{-1/2} \tag{9}$$

where

$$\widehat{a} = \frac{p(m, n) - p(n, n)}{m^{-1/2} - n^{-1/2}}$$

and $\Phi(w)$ is the standard normal CDF. The p-values $P_{moon}(w)$ are higher than those of asymptotic methods, controlling for size, but converge to them, ensuring validity, since $m = o(n)$ and $m \to \infty$ as $n \to \infty$.

3.2.3 Semi-parametric bootstrap

A semi-parametric bootstrap can also offer an improved performance, as illustrated extensively in C&F. Unlike the *moon* bootstrap, it is a concern of efficiency rather than consistency that motivates its choice. The approach tries to remedy the issue of very high incomes being too irregularly sampled by parameterising the tail of the distribution as a Pareto law governing the occurrence of influential high values.

We adopt the setup of C&F, setting $k = n/10$, where k is the number of sample values considered for the tail parameterisation, and $p_{tail} = 0.04n^{-1/2}$, where p_{tail} defines the probability of needing to draw simulated observations from the Pareto tail. For the parameterised Pareto tail, the threshold value $\underline{y}$ is defined as the sample income with rank $\bar{n} = n(1 - p_{tail})$, rounded to the nearest integer. Given $\underline{y}$ and k, the parameter θ is fitted using the conditional likelihood estimator due to Hill (1975):

$$\widehat{\theta} = k \Big/ \left(\sum_{i=0}^{k-1} \log y_{(n-i)} - k \log y_{(n-k+1)} \right) \qquad \text{for } k > 1 \,.$$

For the constructed semi-parametric bootstrap samples, observations are taken with probability p_{tail} from the fitted Pareto distribution and with probability $1 - p_{tail}$ from the lowest $\bar{n}$ values from the original sample.

3.3 Sample splitting *t*-based approach

In view of the shortcomings of standard and improved inference techniques for inequality, it seems uncomplicated alternatives should be particularly welcome. Ordinary asymptotic inference is unreliable in finite samples; the bootstrap-*t* provides only modest improvements; and the more sophisticated bootstrap methods - *moon* and semi-parametric - though clearly valuable, impose somewhat strong assumptions, require choosing additional parameters, and are computationally expensive.

Ibragimov and Müller (2010) proposed a 'Student t statistic based' approach to robust inference. The principle is to split the original sample in a small number of sub-samples or 'groups' and exploiting the asymptotic normality of the respective group estimators to construct Student t-type statistics. The strategy has been applied, with seemingly promising results, to inequality measures, in Ibragimov and Ibragimov (2018).

The algorithm can be explained in four steps. Let $y = \{y_1, y_2, ..., y_n\}$ be a random sample from a random variable Y with CDF $F(y)$. Suppose we want to test $H_0 : I(F) = I_0$ and consider a consistent estimator $\hat{I}$.

1. Split the sample y randomly into q groups, with $q \geq 2$.
2. Calculate the estimator $\hat{I}$ for each group, giving $\hat{I}_j$ for $j = 1, ..., q$.
3. Construct the t-statistic:

$$t = \sqrt{q} \frac{\left(\bar{I} - I_0\right)}{s_I}, \tag{10}$$

where

$$\bar{I} = \frac{1}{q} \sum_{j=1}^{q} \hat{I}_j \quad \text{and} \quad s_I^2 = \frac{1}{q-1} \sum_{j=1}^{q} \left(\hat{I}_j - \bar{I}\right)^2 .$$

4. Conduct inference by reference to the Student t distribution with $(q - 1)$ degrees of freedom, choosing the appropriate nominal level.

Whenever the population under study has finite second moments, we can invoke a CLT as a rationale for assuming that in each sub-sample our inequality estimators are asymptotically Gaussian, so that $\widehat{I}_j \xrightarrow{D} \mathcal{N}(I, V(\widehat{I}_j)),\ j = 1, ..., q$. In practice, most distributions actually fitting income data do have finite second moments. Of course, any concerns about the asymptotic normality of inequality indices would equally threaten asymptotic and bootstrap inference. Further, in most situations, if the split of the sample is random, it is plausible that $\widehat{I}_j$ and $\widehat{I}_i$ are independent for $i \neq j$. Assuming this to be true, we then have in our hands q independent observations from asymptotically Gaussian random variables, from which we can construct an asymptotically exact Student t test with $q - 1$ degrees of freedom.

The method is simple, intuitive and computationally cheap, particularly in comparison to non-standard bootstrap methods. Moreover, there are no restrictions on the within-group correlations; only independence between groups is necessary, which, provided a split is random, is present by construction.

Thus, spatial correlations, likely in income distributions, or other unknown correlation structures in the population should not harm the performance of the method.

It is worth noting that the test statistic has a symmetric distribution, not having been designed specifically for inequality inference nor for heavy-tailed distributions, but rather, remaining valid in these circumstances. Both the *moon* and the semi-parametric bootstrap, on the contrary, directly tackle the negative skew of the distribution of W (the first by adjusting the p-value and the latter through the simulated extremely high data points).

For simplicity, we keep the significance level at 5% throughout our experiments. Interestingly, in the case of heteroskedasticity across groups, the probability of rejection errors by the t statistic (10) remains below the nominal level, at least if the nominal level does not exceed 8.3%; the same continues to hold for nominal levels up to 10% as long as $q \leq 14$. This important result is due to Bakirov and Székely (2006). Hence, as emphasised by Ibragimov and Müller (2010), the t-approach is conservative, controlling for size better than its contenders which tend to be oversized. In the one-sample problems discussed so far, 'conservativeness' is not played to our advantage. Given a random group split, there is no reason to believe group estimators will have differing (asymptotic) variances, hence smaller than nominal rejection probabilities are not expected. In two-sample problems, on the other hand, this property will come in handy, and we will discuss it further in that context.

For the record, the t-approach easily generates confidence intervals by test inversion, with in large samples coverage probability at least equal to one minus the nominal level:

$$\left[\widehat{I} - |t_{\alpha/2,q-1}| \times \frac{s_I}{\sqrt{q}}\ ,\ \ \widehat{I} + |t_{\alpha/2,q-1}| \times \frac{s_I}{\sqrt{q}}\right] \tag{11}$$

4 Inference on inequality in comparison problems

Two-sample comparisons of inequality indices can in many cases be of greater interest than one-sample inference. It should be clear that equal inequality indices do not imply equal distributions: inequality measures are, like in general moments, characteristics that do not uniquely determine the shape of a distribution. It is much harder, though, to test for the equality of an inequality measure without than with the assumption of identical distributions under the null hypothesis. The problem is analogous to that of comparing means in two (normal) populations with possibly different variances, the vexed Behrens-Fisher problem. Sometimes, the overlap of one-sample confidence intervals has been used as a simple decision

criterion; see e.g. Moran (2006) and Ibragimov et al. (2013). Here we consider more formal tests.

4.1 Asymptotic and bootstrap tests

D&F provided simulation results on the size of several difference-in-means tests, allowing for unequal variances between populations, applied to the Theil index and maintaining different underlying Singh-Maddala distributions. They applied the same four approaches as in one-sample hypothesis testing: asymptotic inference, bootstrap-t, *moon* bootstrap and semi-parametric bootstrap. We refer to their paper for detailed descriptions.

By contrast with the results for one-sample hypothesis testing, their Fig. 15 reveals no meaningful differences among the methods: all are oversized, with rejection rates converging to about 10% under the null – corresponding to an Excess (or Error in) Rejection Probability (ERP) of about 5% (though note that only sample sizes up to 3,000 were simulated). Techniques with better size control are desirable.

4.2 Permutation tests

Recently, Dufour et al. (2019) conducted simulation experiments studying the finite-sample properties of various implementations of the permutation tests described in the introduction. By randomly mixing and permuting the observations from the original two samples they construct artificial samples which are exchangeable with the original ones (and with each other) provided the two population distributions are identical. Even if the two population distributions are different, for example if one is more heavy-tailed than the other, they show that the permuted samples may still deliver valid tests of the equality of Gini coefficients or of indices from the Generalised Entropy class, provided specific conditions are met. They propose and test three different implementations of permutation tests, which we describe as follows.

- p_s : the 'standard studentised' permutation test, which is the studentised difference between the indices of the two permuted samples;
- p_s_r : the 'rescaled studentised' permutation test, which is analogous to the preceding one except that the data in the two original samples are rescaled by their respective sample means before they are combined and permuted;
- p_t_r : the 'rescaled but not studentised' permutation test, where the same prior rescaling has taken place but the difference between the indices of the two permuted samples is used directly without being studentised.

Rescaling by the sample mean does not affect the inequality measures (which are scale invariant) but is useful to validate permutation tests asymptotically. The findings are that, compared to standard asymptotic and bootstrap tests, permutation methods improve size control, even in unfavourable situations like comparing two distributions with differently shaped upper tails or samples of different sizes; and that rescaling observations by sample means before permutation improves test power. Clearly, permutation tests are a useful addition to the inequality analysis toolbox; we will see whether the t-approach by sample splitting is able to offer comparable performance. The limitation of the permutation tests is that they do not easily generate confidence intervals or generalise to other problems than two-sample comparison tests.

4.3 Sample splitting *t*-based approach

If one ascribes value to inequality measures as key characteristics of an income or wealth distribution, it seems natural to focus interest on comparisons of an inequality measure of choice across two samples rather than of the entire distribution. In view thereof, the result of Bakirov and Székely (2006) showing how a conservative t-test can be constructed from a scale mixture of normal random variables is particularly helpful for comparing inequality indices. Unless we are willing to assume identical distributions under the null, the variances of the two populations being compared are bound to differ. The result is an extension of the unequal variances t-test of Welch (1947) and ensures that we can still construct correctly sized comparison tests in this situation, an opportunity emphasised by Ibragimov and his co-authors. We summarise the proposed procedure as follows.

Let $y_A = \{y_{1A}, y_{2A}, ..., y_{nA}\}$ be a random sample from a population distribution $F_A(y)$ and $y_B = \{y_{1B}, y_{2B}, ..., y_{mB}\}$ a random sample from a population distribution $F_B(y)$. Suppose we want to conduct inference about the difference between the corresponding two inequality indices, $I(F_A) - I(F_B)$.

1. Split sample y_A randomly into q_A groups and sample y_B randomly into q_B groups, with $q_A, q_B \geq 2$.
2. Calculate the estimator $\hat{I}$ for each group, i.e., $\hat{I}_{iA}$ for $i = 1, ..., q_A$ and $\hat{I}_{jB}$ for $j = 1, ..., q_B$.
3. Construct the following t statistic:

$$t_{ibrag} = \frac{\left(\bar{I}_B - \bar{I}_A\right)}{\sqrt{\frac{s_A^2}{q_A} + \frac{s_B^2}{q_B}}}, \tag{12}$$

where

$$\bar{I}_A = \frac{1}{q_A}\sum_{j=1}^{q_A} I_{jA}, \qquad s_A^2 = \frac{1}{q_A - 1}\sum_{j=1}^{q_A}\left(I_{jA} - \bar{I}_A\right)^2,$$

and $\bar{I}_B$ and s_B^2 are defined in the same way.

4. Conduct inference with the desired nominal level using the critical values of the Student t distribution with $min(q_A, q_B) - 1$ degrees of freedom.

As an illustration, to perform a two-sided test with nominal level α of $H_0 : I_A = I_B$, reject H_0 in favour of the alternative $H_1 : I_A \neq I_B$ if $|t_{ibrag}| \geq t_\alpha$, where $P(|T_{min(q_A,q_B)-1}| \geq t_\alpha) = \alpha$. The test is easily adapted for testing whether the difference between the inequality indices is equal to, greater than, or less than a given threshold Δ_0. To test $H_0 : I_B - I_A \geq \Delta_0$ against $H_1 : I_B - I_A < \Delta_0$, simply subtract Δ_0 from the numerator of t_{ibrag}. Reject the null hypothesis if the (altered) t_{ibrag} is below the α quantile of the Student t distribution with $min(q_A, q_B) - 1$ degrees of freedom.

In cases where the two comparison samples have (very) different sizes, it may be optimal to choose $q_A \neq q_B$. However, in cases where the two comparison samples have (roughly) equal sizes, it is natural to choose $q_A = q_B = q$. In that case, t_{ibrag} coincides with an ordinary two-sample t statistic using the pooled variance estimator. The degrees of freedom assumed differ, however, by a factor 2: here we use $min(q_A, q_B) - 1$ rather than the standard $2(q - 1)$, hence larger critical values. The difference between the two tests dissipates as q increases, and, for $q_A = q_B = q \to \infty$, there is no difference between Ibragimov's test and an ordinary pooled t-test assuming equal variances.

5 Simulation evidence

5.1 One-sample inference

Simulation results on the size of asymptotic and bootstrap-based tests described in Sections 3.1 and 3.2 are presented by C&F. We replicated a small number of simulations using the same four methods and confirmed the authors' findings, eliminating any potential added value from full replication[4].

5.1.1 Size

Table A2 in the online Appendix reports ERPs (Excess Rejection Probabilities) for left-sided t-based tests on the same distributions as those used by C&F and other papers. A detailed comparison across all methods, in terms of size, is possible by reference to C&F's Tables 4 and 5, since our table is designed as a complement to theirs. For the t-approach with $q = 4$, we find high ERPs that are barely better than those of the asymptotic tests; but with $q = 2$, the ERPs are much smaller, less than those of the standard bootstrap, and quite comparable to those of the improved bootstrap methods.

Indeed, for the same nominal level, the t-approach test with $q = 2$ has a smaller size than the bootstrap-t test in virtually all simulations, regardless of the underlying income distribution, its parameter values, the inequality index considered or the sample size.

The issue of inequality underestimation is predictably present in the t-approach, just as in the bootstrap-t. Left-sided tests exhibit large positive ERPs while right-sided tests are undersized. Two-sided tests are likewise oversized, but less so; although hardly noticeable in the baseline case, this tendency comes to the surface in more unequal distributions. Figure 3 plots the ERPs of left, right and two-sided tests. (Numerical values are reported in the online Appendix, Tables A2, A3 and A4.)

Just as the bootstrap-t and the non-standard bootstrap methods, the t-approach is clearly sensitive to the distribution of income, performing worse in the presence of more frequent extreme observations. Even in the log-normal distribution, which is not heavy tailed, merely increasing the variance is enough to raise the test size.

The advantage of the t-approach in comparison with the bootstrap-t is accentuated under heavy tails. In two-sided tests, the ERP of the bootstrap-t for the Theil index exceeds 10 percentage points (10.73, see Table A5 in the online Appendix) even in samples of size 5,000 from Singh-Maddala distributions with very heavy tails ($c = 0.7$ rather than 1.7), as compared to only 2.17 p.p. with the t-approach (see Table A4).

Both methods are of little use in the case of the Pareto Distribution with shape parameter below 2, since then the variance is infinite and neither method is valid (see Table A4 and C&F's Table 4). The *moon* bootstrap remains theoretically valid and performs better, but not much. Only the semi-parametric bootstrap retains its edge, though nevertheless far from the nominal level, with an ERP still as high as 8% for the Theil index (in left-sided tests with $\theta = 1.5$, see C&F's Table 4).

The sensitivity of the different inequality measures to extreme values is also noted, with inference being more challenging for the GE_2 measure, followed by the GE_1 and the GE_0 and Gini Index, the latter two very close in terms of ERP. The extension of the analysis from C&F on the Gini Index to right-sided and two-sided tests and up to $n = 5{,}000$ confirms its behaviour closely mimics the GE_0 (exact results omitted).

[4] Left-sided test sizes comparable to entries in C&F's Table 4 are reported in the online Appendix Table A1.

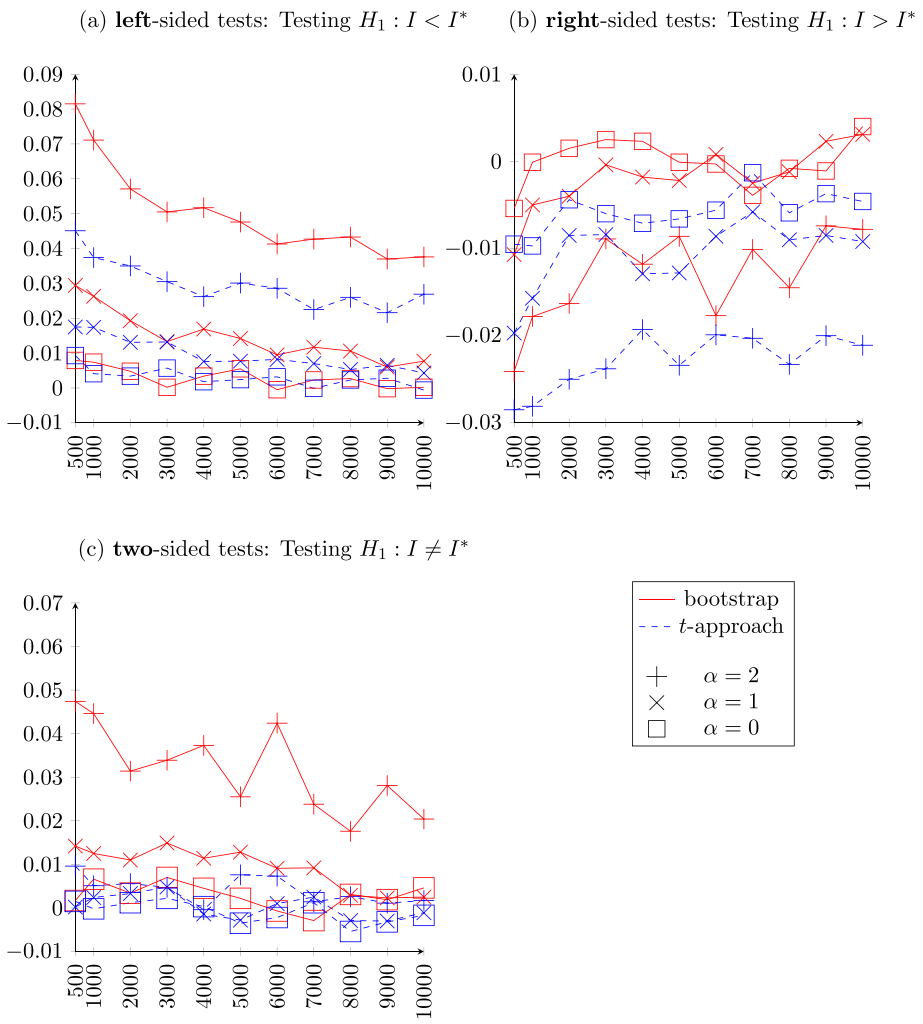

Fig. 3 ERP of tests on Generalised Entropy Measures using bootstrap methods and t-approach, $H_0 : I = I^*$. The true value of I is $I^* = 0.140115$. Testing $H_0 : I = I^*$ vs H_1 defined in panel title, with I standing for GE_α. Singh-Maddala distribution with $a = 100$, $b = 2.8$, $c = 1.7$. Nominal level of 5%, ERP = Proportion of rejections - 0.05, t-approach uses $q = 2$

The choice of q - the number of groups into which the sample is split - has a clear effect on the size of the method. Increasing q from 2 to 4 is enough to double the ERP for left-sided tests, with further increases as q is augmented.

5.1.2 Power

Improvements of the t-approach in size control compared to the bootstrap-t are obtained at the expense of efficiency. The method is substantially less powerful, with differences hardly subsiding with increased sample size. This is illustrated for left-sided tests in Fig. 4, but

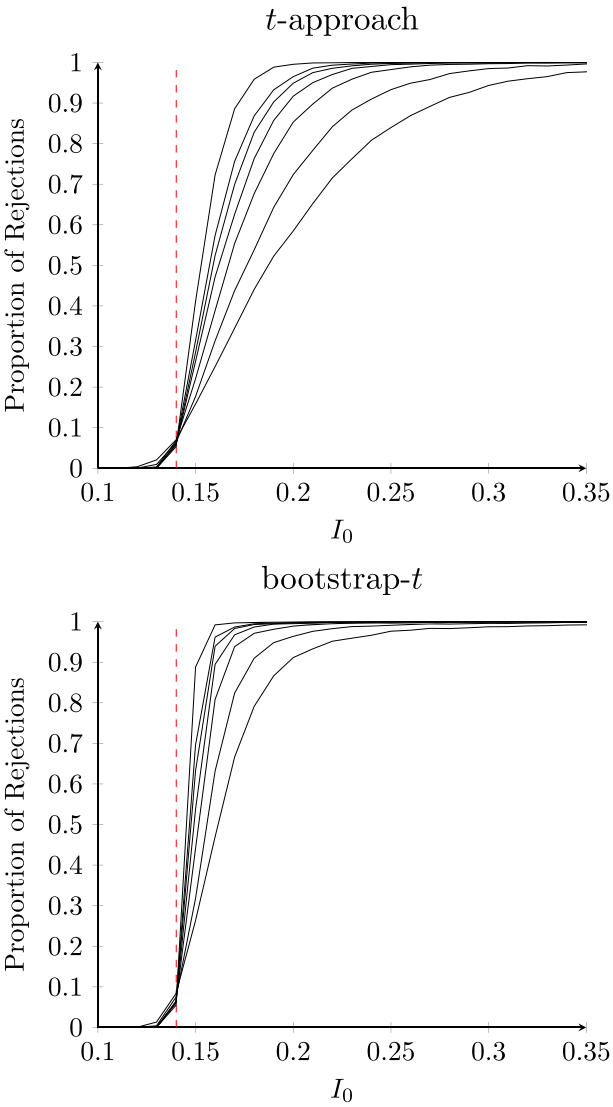

Fig. 4 Power of **left**-sided tests on the Theil index. Testing $H_0 : I = I_0$ vs $H_1 : I < I_0$, with I standing for GE_1 (Theil index). Singh-Maddala distribution with $a = 100$, $b = 2.8$, $c = 1.7$; The true value of I is $I^* = 0.140115$. Nominal size of 5%; $n \in \{500, 1{,}000, 2{,}000, 3{,}000, 4{,}000, 5{,}000, 10{,}000\}$

is also very visible in two-sided and right-sided tests (See Figs. A2 and A3 in the online Appendix).

More unequal distributions again harm method performance, with both methods presenting noticeably lower power when parameter c of the Singh-Maddala distribution is reduced. (For a graphical comparison, contrast panels (a) and (b) of Fig. A4 in the online Appendix).

We now look into the comparison of the t-approach with the remaining bootstrap methods in terms of power. The t-approach is not only noticeably less powerful than the bootstrap-t, but also than the non-standard bootstrap methods. The difference in power between the *moon*

bootstrap, the semi-parametric bootstrap and the bootstrap-t is small. Though in extremely small samples, simple asymptotic inference is still noticeably more efficient, the difference dissipates with $n = 5{,}000$, as visible in Fig. 5.

5.2 Two-sample comparisons

5.2.1 Size

In our simulations all methods seem to perform slightly better than in D&F, but we reach the same conclusion of virtually no differences among them. In comparison to these, the added value of the t-approach becomes clear for difference in means testing. The t-approach controls for size remarkably well, and even more so when compared to the bootstrap and asymptotic

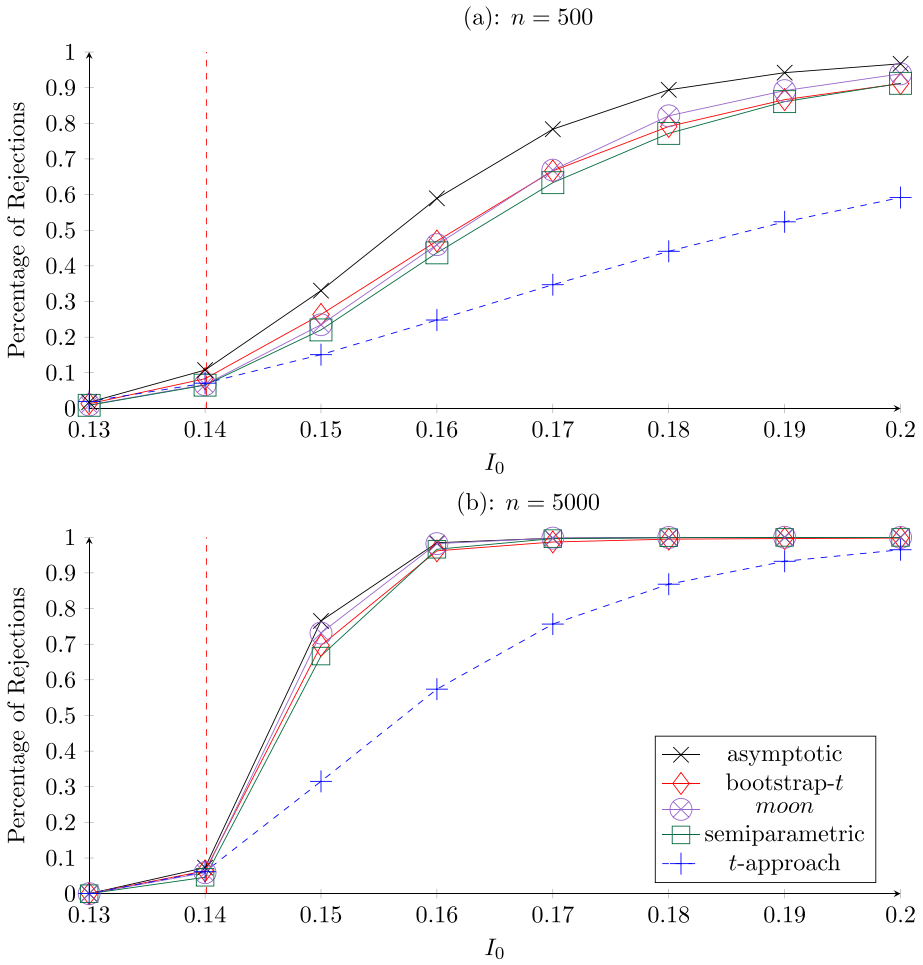

Fig. 5 Power of left-sided tests on the Theil index using all bootstrap methods and the t-approach. Testing $H_0 : I = I_0$ vs $H_1 : I < I_0$, with I standing fr GE_1 (Theil index). Singh-Maddala distribution with $a = 100$, $b = 2.8$ and $c = 1.7$; The true value of I is $I^* = 0.140115$. Nominal size of 5%

contenders (see Fig. 6). Conservativeness from heteroskedasticity allows for an effective test size below nominal levels for $q = 2$ and $q = 4$. Higher q leads to size increases, yet $q = 8$ is still preferable to asymptotic and bootstrap methods, approaching nominal levels faster. A q of 16 makes size comparable to that of the *moon* bootstrap (not shown). Conservativeness is present both with small ($n = 500$) and large samples. For example, when $n = 50{,}000$, we find a test size of 5.4% for $q = 16$ and below 5% for $q = 8$ (see Table A6 in the online Appendix).

When the second distribution is log-normal (not heavy-tailed), the size of the t-approach becomes smaller, for all q. The test seems promising even for very different underlying distributions. Asymptotic and bootstrap methods also have a smaller size when the second distribution is not heavy-tailed (Table A7 in the online Appendix shows more details). Under those circumstances, over-sized tests are not as big a concern, regardless of the method. We focus instead on more heavy-tailed distributions, for which the t-approach reduces over-rejection vis-à-vis bootstrap and asymptotic methods, and investigate whether this advantage holds against permutation tests too.

The fundamental difference between permutation tests and the t-approach (and the bootstrap related methods here considered) is the null hypothesis under consideration. Indeed, in permutation tests, the *natural* null hypothesis is equality of distributions, as opposed to equality of the inequality indices. When the underlying distributions are different, but inequality indices the same, our simulation results demonstrate that all three versions of the permutation tests as suggested by Dufour et al. (2019) are over-sized, even at sample sizes of 20,000 - see Table 2. This is comparable to the results of Dufour et al. (2019) in their Fig. 6.

When the distributions are the same, the permutation tests instead have the correct size of 5%, while the t-approach remains conservative (results omitted).

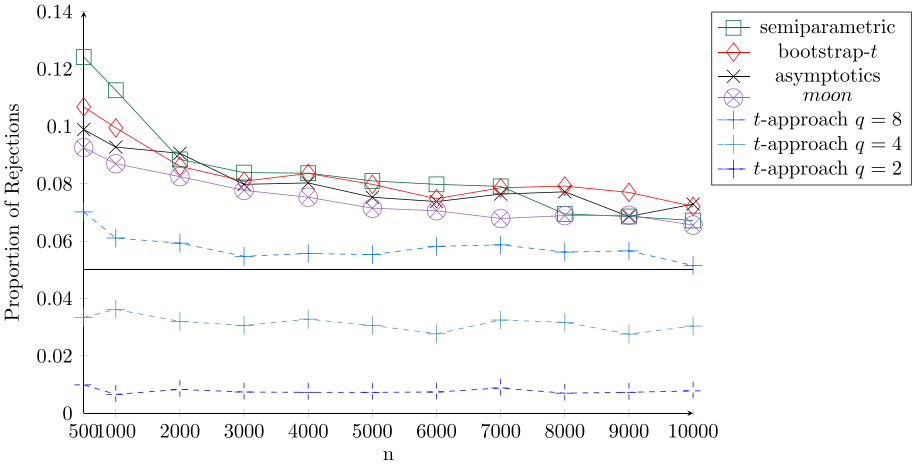

Fig. 6 **Size** of test on equality of Theil index between two distributions. Testing $H_0 : I_A = I_B$ vs $H_1 : I_A \neq I_B$, with I standing for GE_1 (Theil index). Singh-Maddala distribution A with $a = 100$, $b = 2.8$, $c = 1.7$; $I_A = 0.140115$. Singh-Maddala distribution B with $a = 100$, $b = 4.8$, $c = 0.636659$; $I_B = I_A$. Nominal level of 5%, size = proportion of rejections

Table 2 Size of tests of equality of Theil index between two Singh-Maddala distributions: t-approach versus permutation methods

n	p_s	p_s_r	p_t_r	$t_{q=4}$	$t_{q=6}$	$t_{q=8}$	$t_{q=10}$	$t_{q=12}$
50	0.096	0.1	0.126	0.034	0.051	0.063	0.081	0.091
100	0.101	0.109	0.131	0.031	0.053	0.075	0.079	0.101
500	0.102	0.105	0.112	0.043	0.058	0.066	0.085	0.085
1,000	0.081	0.076	0.08	0.038	0.045	0.054	0.065	0.068
50,00	0.084	0.088	0.09	0.035	0.058	0.064	0.071	0.07
10,000	0.073	0.077	0.078	0.032	0.049	0.054	0.057	0.061
20,000	0.055	0.058	0.058	0.02	0.027	0.033	0.045	0.043
100,000	0.072	0.068	0.069	0.032	0.047	0.054	0.061	0.062

Size of tests of $H_0 : I_A = I_B$ vs $H_1 : I_A \neq I_B$ when $I_A^* = I_B^* = 0.140115$, with I standing for GE_1 (Theil index), using the t-approach and 1,000 simulations. Singh-Maddala distribution A with $a = 100$, $b = 2.8$, $c = 1.7$; Singh-Maddala distribution B with $a = 100$, $b = 4.8$, $c = 0.636659$. Nominal level $\alpha = 0.05$

5.2.2 Power

The t-approach becomes more powerful as q increases. At the lowest number of splits ($q = 2$), power is substantially below asymptotic inference (see Fig. 7). A q of 4 is enough for a powerful test, yet, it is with a q of 8 and 16 that efficiency closely approaches that of the asymptotic test. Bootstrap methods closely follow asymptotic inference in terms of power (results omitted).

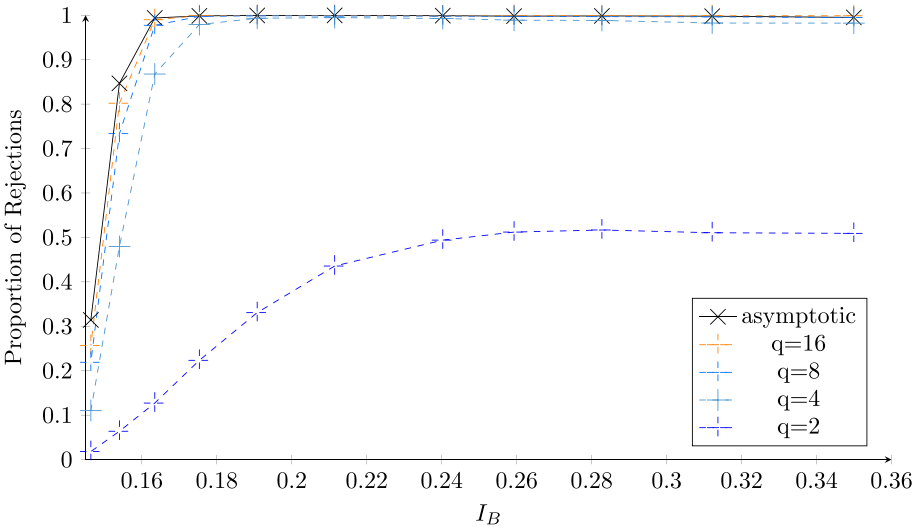

Fig. 7 Power of test of equality of the Theil index based on two Singh-Maddala distributions. Testing $H_0 : I_A = I_B$ vs $H_1 : I_A \neq I_B$, with I standing for GE_1 (Theil index). Distribution A: Singh-Maddala distribution with $a = 100$, $b = 2.8$, $c = 1.7$; $I_A = 0.140115$. Distribution B: Singh-Maddala with a=100, b=2.8 and $c \in \{1.6, 1.5, 1.4, 1.3, 1.2, 1.1, 1, 0.95, 0.9, 0.85, 0.8\}$ to match the values of the Theil index I_B on the horizontal axis: $I_B \in \{0.146, 0.154, 0.163, 0.175, 0.191, 0.211, 0.240, 0.259, 0.283, 0.312, 0.350\}$. Nominal level of 5%, $n = 10{,}000$

Permutation tests are also more powerful than the t-approach, particularly at small sample sizes. As the sample sizes increase, the difference dissipates. Table A8 in the online Appendix presents detailed results on the power comparison between permutation tests and the t-approach.

5.2.3 Correlated samples

The previous simulation analysis considered only independent variables. Here, instead, we resort to simulations from a bivariate Gumbel Copula, which allows us to model upper tail dependence, and briefly from a Clayton Copula, which allows us to model lower tail dependence. We firstly consider a bivariate Gumbel copula composed of two Singh-Maddala distributions with a Kendall's rank correlation coefficient of 0.4 (Gumbel copula parameter of 5/3).

The t-approach is less powerful than permutation contenders, but maintains a lower size, particularly for smaller sample sizes (online Appendix Tables A9 and A10).

The permutation tests which are, in independent samples, oversized, are yet much less so in this application, since the positive correlation specifically on the upper tail makes the samples more similar in terms of Theil index (more sensitive to extremely high than extremely low observations), overriding the tendency the tests have to find statistically significant differences too frequently.

This lower overrejection compared to independently distributed random variables is not generalizable to all types of dependence. For instance, under a positive correlation of the lower tail, based on a Clayton copula with a Kendall's rank correlation coefficient of 0.45, the level of overrejection of the permutation tests is similar to that for independently distributed random variables, i.e., considerably above nominal levels and above the t-based approaches (see the lower panel of Table A9).

6 Empirical application

In this section we present an empirical application comparing income distributions of households in the Russian Federation obtained from two different sources: Rosstat and RLMS-HSE. Rosstat, the governmental statistics agency of the Russian Federation, conducts sampling surveys of Russian household budgets continuously over the calendar year. We use the data for 2020, the most recent available to us, with records from 60,000 households. The microdata were obtained online from a dedicated website.[5]

RLMS-HSE, on the other hand, is the result of an academic collaboration between the Higher School of Economics in Moscow and the Carolina Population Center of the University of North-Carolina at Chapel Hill. The project has been going on since the 1990s, and the surveys have a useful panel component. The sample size is much smaller than that of Rosstat: we have 6,365 useful observations for the year 2020 (again the most recent year available), just over one tenth the size of the Rosstat sample. The RLMS-HSE maintains two samples: one which follows all households previously included whenever feasible, for the sake of the continuity of the survey's panel component; and another designed to be cross-sectionally representative, which drops those panel households that would affect the current cross-sectional representativity of the sample. Strictly speaking it is the latter,

[5] The dedicated website has been closed but, as of May 2023, the official Rosstat site (https://rosstat.gov.ru) provides a Russian-language link to 'Censuses and Surveys', through which the microdata are made accessible.

'representative' sample that should be used for comparisons with Rosstat, which reduces our sample size to 4,831 households.

The Rosstat data are reported on a yearly basis, and we scale them to monthly amounts for comparability with RLMS-HSE. Both data sets contain some very small values of income, and the RLMS-HSE incomes, which are recorded on a monthly basis, even include a few zero or negative values. Rather than dropping the affected households, we winsorise the monthly incomes to a minimum of 1,000 rubles (approximately 12 USD).[6] This intervention concerns 32 RLMS-HSE households and 16 Rosstat households. The resulting datasets are available from the authors upon request.

6.1 Comparisons within surveys

We begin with comparisons of the Theil index within each survey. The main outcomes are reported in Table 3. Our first application is to test whether, within the Rosstat data, there is a difference between the inequality among total incomes ($ytot$) and among disposable incomes ($ydisp$). Despite a rather small difference in magnitude between the two Theil indices, all tests - both versions of the t-approach, with $q = 4$ and $q = 8$, and the three versions of the permutation test proposed by Dufour et al. (2019) - reject the null of equality. The Theil index estimated from total incomes is higher than that estimated from disposable incomes, as would be expected as a result of redistribution policies. In the third column of the table, a t-based 95% confidence interval (with $q = 4$) puts the fiscal reduction in Theil inequality between 0.007 and 0.030.

Rows b) and c) of Table 3 report comparisons within the RLMS-HSE data. In row b), we see that there is a large and according to all tests significant difference between the inequalities calculated from total incomes ($ytot$) and from total expenditures ($xtot$). Surprisingly, expenditures are much more unequal than incomes. This intriguing finding might be a consequence of not just differential spending patterns but also income under-reporting behaviour. In row c), by contrast, we are unable to detect any rejection of the null of equal income inequality between the full RLMS-HSE sample ($ytot$) and its 'representative' sub-sample ($ytotrep$).

6.2 Comparisons across surveys

The motivation for the next empirical application is that the Rosstat and RLMS-HSE surveys are both designed to represent the same federal population and both report a total income variable. Assuming they are appropriately built, the inequality indices of total income from the two surveys should not be statistically different. Results are presented in Table 4. Comparing all regions, the t-based tests indeed do not reject the null of equal Theil indices, whereas one version of the permutation test does, namely p_t_r; however, both our simulations and those of Dufour et al. (2019) show that the p_t_r version is the most oversized one. The results are practically identical whether we use the RLMS-HSE full sample or the smaller one designated as representative (row e). There is no trace of a rejection any more if we draw a Rosstat sub-sample of the same size as the RLMS sample ($n = 6365$, row f).

Finally, we focus the comparison on the Moscow & St. Petersburg city areas, both together and separately, on the premise that this leaves less room for sampling differences between

[6] Zero or negative reported values indicate very low incomes, but our inequality measures cannot deal with nonpositive values. Furthermore, extremely low values risk being influential, whereas their precise value hardly matters in practice.

Table 3 Comparison of Theil indices within each survey: p-values and confidence intervals

Distribution A vs B	I_A	I_B	CI $t_{q=4}$	$t_{q=4}$	$t_{q=8}$	p_s	p_s_r	p_t_r
a) Rosstat: $ytot$ vs $ydisp$	0.233	0.215	[0.007,0.030]	0.10%	1.52%	0.00%	0.00%	0.00%
b) RLMS: $ytot$ vs $xtot$	0.261	0.489	[-0.301,-0.152]	0.02%	0.24%	0.00%	0.10%	0.10%
c) RLMS: $ytot$ vs $ytotrep$	0.261	0.262	[-0.207,0.174]	96.2%	96.8%	86.3%	89.6%	89.8%

I_A and I_B are the GE_1 (Theil) indices of distributions A and B under comparison.
CI $t_{q=4}$ are t-based 95% Confidence Intervals for their difference.
a) Rosstat total income ($ytot$) vs disposable income ($ydisp$)
b) RLMS-HSE total income ($ytot$) vs total expenditures ($xtot$)
c) RLMS-HSE total income in complete sample ($ytot$) vs total income in cross-sectional 'representative sample' ($ytotrep$), which drops households that are followed for the purposes of the panel but affect the cross-sectional representativity of the sample

surveys. Unexpectedly, though, all permutation tests reject the null in these regional comparisons, whereas the t-based tests continue passing it. Note that the table also presents 95% confidence intervals for the various contrasts we have tested. The straightforward delivery of confidence intervals is a practical advantage of the sample-splitting t-based approach compared to permutation tests.

Table 4 Comparison of Theil indices between surveys: p-values and confidence intervals

Distrib. A vs B	I_A	I_B	CI $t_{q=4}$	$t_{q=4}$	$t_{q=8}$	p_s	p_s_r	p_t_r
All regions								
d) Rosstat $ytot$ vs RLMS $ytot$	0.233	0.261	[-0.087,0.032]	22.8%	23.6%	8.81%	8.11%	1.10%
e) Rosstat $ytot$ vs RLMS $ytotrep$	0.233	0.262	[-0.081,0.023]	16.4%	17.6%	6.0%	7.51%	0.90%
f) Rosstat S $ytot$ vs RLMS $ytot$	0.230	0.261	[-0.127,0.067]	33.1%	40.2%	19.6%	18.5%	18.5%
Moscow & St. Petersburg								
g) Rosstat $ytot$ vs RLMS $ytot$	0.195	0.313	[-0.286,0.063]	5.1%	13.6%	0.00%	0.10%	0.10%
h) Rosstat $ytot$ vs RLMS $ytotrep$	0.195	0.335	[-0.406,0.154]	11.6%	24.8%	0.20%	0.10%	0.10%
Moscow								
i) Rosstat $ytot$ vs RLMS $ytot$	0.206	0.328	[-0.328,0.102]	11.1%	19.4%	0.00%	0.10%	0.10%
St. Petersburg								
j) Rosstat $ytot$ vs RLMS $ytot$	0.144	0.236	[-0.263,0.068]	8.9%	15.8%	0.60%	0.30%	0.10%

I_A and I_B are the GE_1 (Theil) indices of distributions A and B under comparison. CI $t_{q=4}$ are t-based 95% Confidence Intervals for their differences. RLMS is short for RLMS-HSE. $ytot$ is total income, $ydisp$ is disposable income, $ytotrep$ is total income in the RLMS-HSE 'representative sample', which drops households that are followed for the purposes of the panel but affect the cross-sectional representativity of the sample. 'Rosstat S' is a random sample from Rosstat of the same size as the RLMS sample ($n = 6365$)

7 Discussion

7.1 One-sample inference

The t-approach presents ERPs always lower than the bootstrap-t, comparable to the *moon* bootstrap, and in most cases higher than the semi-parametric bootstrap. Our conclusions mostly echo the statement in Ibragimov et al. (2013) and Ibragimov et al. (2021) that "(the size properties of t-based...) are comparable and in many cases dominate the size properties of computationally expensive alternatives". However, the ERP alone is not enough to motivate the choice of an inference procedure. The use of a low-size but possibly low-power test begs the question of whether our hypothesis is not rejected because it is true, or because the evidence against it is not strong enough to be detected by our test. Thus, the possible trade-off between size and power was investigated.

The t-approach controls better for size than the bootstrap-t but in clear detriment of power. The other bootstrap methods - *moon* and semi-parametric - on the contrary, achieve size control while only slightly reducing power, making them better candidates for inference. These results give strength to the *moon* and particularly to the semi-parametric bootstrap as inference techniques, supporting the conclusions of C&F.

Furthermore, since the semi-parametric bootstrap tackles the problem of inequality inference at its core (absence of representative high incomes) the asymmetry issue disappears: there are no longer substantial differences between right-sided and left-sided tests (results not shown).

Some caveats are in order. The baseline simulations, based on a Singh-Maddala distribution, are well-suited for a Pareto parameterisation, given the power decay of both distributions. Indeed, when one takes the log-normal distribution instead, where the tail decays exponentially, the ERP of the semi-parametric bootstrap increases considerably, actually surpassing that of the t-approach. Moreover, our simulations regarding one-sample inference were limited to i.i.d. samples. The unbiasedness of the simple full-sample variance estimators, present in all bootstrap methods, is ensured. In that sense, there is hardly a situation more tailored to improved performance of the methods. The question remains whether in a different scenario, the t-approach, avoiding possibly biased variance calculations, could be beneficial.

Some uncertainty exists surrounding the ideal parameters of a semi-parametric bootstrap. D&F advise further analysis on the choice of p_{tail}. The value chosen in our experiments (following C&F) is reasonable; p_{tail} still tends to 0 as $n \to \infty$ but not too fast. Hence, the need to sample from the fitted distribution diminishes and ultimately vanishes as our sample grows, but simultaneously, because convergence to zero is not too fast, the probability is still relevant in considerably large samples. To the extent of our knowledge no further exploration of the issue has been attempted. A more thorough answer to these questions could both improve the performance of the method and broaden its use.

It is worth highlighting that this paper is not meant to provide a fully comprehensive review of inference in inequality approaches. We have not looked into other bias-corrected methods, such as those developed by Schluter and van Garderen (2009), which are more demanding in terms of moments of the underlying distribution (the GE_2 measure would not converge in our baseline case). A direct comparison of results is not available since simulation conditions differ.

Some attempts at improving the t-approach were undertaken. A semi-parametric adaptation, where for each split the semi-parametric inequality measure is computed, satisfies

the conditions for validity. However, in most of our experiments the parameterisation led to higher variances, overriding the lower bias in the mean group Theil index.

A larger number of splits q, for fixed n, decreases the size of each group, harming the quality of the asymptotic approximation. On the other hand, a higher q gives us additional, though more biased, observations of the inequality measure. A possible improved technique considering these two competing forces would look into re-sampling the splits, i.e., from the original sample, creating two groups and again, resorting to the original sample, creating two new groups, and repeating the process several times. Preliminary results suggest power improvements but size control seems to require bias adjustments.

7.2 Two-sample comparisons

The t-approach appears to be well-suited for difference-in-means tests. Its conservativeness allows for effective size below nominal level and not in serious detriment of power. Unlike for one-sample hypothesis testing, increases in q do not severely harm size control, since conservativeness with respect to heteroskedasticity serves as a counterweight.

Permutation tests proposed by Dufour et al. (2019; 2020), are, even for very large sample sizes ($n = 100,000$), oversized, as are the bootstrap-based methods. The t-approach retains its edge here: it minimises type I error in comparison with the recently suggested permutation tests and the previous bootstrap contenders.

Permutation tests are shown to be, as were the bootstrap contenders, more powerful than the t-approaches. Yet, importantly, once the sample size is large enough (in our examples, we consider samples of size 20,000), the t-approaches become almost as powerful. Indeed, they are able in approximately 99% of cases to correctly identify a difference between a Theil index of 0.14 and a Theil index of 0.163. When dealing with large samples, thus, there can be an argument for choosing a t-approach, as one is able to minimize type I error with little negative impact on power.

The case for considering the t-approaches is made stronger when taking together results by Ibragimov et al. (2021) which show how in certain cases, depending on the different sample sizes of each of the two samples and the heaviness of the tails, the t-based approaches can actually dominate the permutation tests in terms of power while keeping their edge in terms of reduced size.

Moreover, we show the performance of the permutation tests is sensitive to the exact correlation structure of the samples (lower or upper tail dependence), substantially more so than the t-approaches.

With smaller samples, a permutation test remains a strong option. It should be kept in mind that the possibility of over-rejection is always present. A t-approach can be used as an additional test to find out whether a difference found empirically is robust. Such confirmation may be important especially if one is assessing policy impacts.

Another advantage of the t-approaches is the easiness to construct confidence intervals around the difference between inequality indices. The difference follows the Student t distribution, with which practitioners are very familiar. In empirical analyses considering, for example, drivers of inequality, or structural breaks in inequality, having an easily computable test with known distributional properties can be of great help.

Recommendations in other scenarios advised a q of 8 or 16 - Ibragimov et al. (2021) specifically state "The numerical results presented in Ibragimov and Müller (2010) demonstrate that, for many (...) heterogeneity settings considered in the literature and typically encountered in practice (...), the choice $q = 8$ or $q = 16$ leads to robust tests with attractive

finite sample performance." A q of 16 however, in this case, does not control for size better than bootstrap contenders.

Our simulation exercises look at q ranging from 4 to 12 for two-sample testing. Higher q makes the test less conservative but, on the other hand, more powerful. For large sample sizes - n =10,000 or 20,000 -, the test is already very powerful with $q = 4$; so such a choice is advisable, if one prioritises type I error control. For sample sizes below 10,000, $q = 8$ appears, from the simulation results, to still guarantee ERPs below any of the three permutation-based tests, and obtain a test more able to find meaningful differences where these exist.

An interesting research question left for further work is whether the division into several subgroups can reveal additional information about inequality structures in a population. This might be possible by exploiting the decomposability of inequality indices into within-group and between-group inequality. It would be of high interest as a practical application to test not only whether, for example, the Theil index is different between two populations, but also whether inequality between (within) regions in one population is different from inequality between (within) regions for another population.[7]

8 Conclusion

The t-approach is a simple, intuitive and computationally cheap inference method. Not having been developed specifically for inference in inequality (nor for heavy-tailed distributions), it falls prey to some of the issues that hinder asymptotic and standard bootstrap methods, namely asymmetric rejections areas. Furthermore, though in terms of size control it appears superior to these approaches, a power analysis reveals strikingly lower efficiency.

Tackling the problem of inequality underestimation from sensitivity to extreme values at its core, the *moon* and especially the semi-parametric bootstrap are clearly preferable, minimising size (in most cases, lower than the t-approach) without losing power. Though more unequal distributions remain challenging for the semi-parametric bootstrap - a disappointing result from an empirical perspective - its edge is clear and arguably justifies the added complexity.

In the difference-in-means framework, conservativeness comes into play for the t-approach, creating a very interesting test. Size is maintained below nominal levels, substantially below those of the contenders, both in very small and in large samples, and quite high power - not substantially lower than asymptotic, bootstrap and permutation contenders - is achieved when using 4 and 8 groups (q), especially in large samples which are routinely found when studying income distributions. In certain situations, as Ibragimov et al. (2021) show, the t-based approaches might not even be less powerful, while keeping its edge on size reduction. The need for such a test is clear, given the economic interest in comparing two populations and the type I error rate of current methods in doing so. The t-based approaches also allow for quick construction of confidence intervals for the difference between inequality indices, resorting to a Student t distribution with which practitioners are very familiar.

Other scenarios besides heterogeneity in which the t-approach showed promise dealt with different forms of dependence between distributions in two-sample tests. Whether the dependence exacerbates or decreases the relative edge of the t-based approaches versus permutation tests depends on the form of dependence. While positive upper tail dependence

[7] We thank one of the reviewers for this interesting point.

reduces the risk of overrejection from permutation tests, the same does not happen with positive lower tail dependence. This is because the degree of overrejection of permutation tests is sensitive to the exact (and in practice, often unknown) correlation structures.

Realistic income distribution situations involving, say, spatial correlations, could reveal new successful applications of the t-approach, as Ibragimov et al. (2021) explore in the one-sample test case. If unbiased estimators of the full-sample variance are unavailable, methods other than the t-approach and permutation tests are at a disadvantage.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10888-023-09574-w.

Acknowledgements Work on this paper was initiated while both authors were at Maastricht University, The Netherlands. We thank Zsolt Darvas for crucial and timely support and the referees for insightful suggestions.

Funding This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 956107, "Economic Policy in Complex Environments (EPOC)". Open Access Funding provided by Universitat Autonoma de Barcelona.

Data Availibility The raw data for the empirical application is publicly available and the few transformations undertaken are described in Section 6. The exact data used for the empirical application following these minor transformations is also directly available from the corresponding author upon request.
The simulation analysis simulates data and is fully reproducible from the R scripts built. The R scripts are directly available from the corresponding author upon request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Athreya, K.B.: Bootstrap of the mean in the infinite variance case. Ann. Stat. **15**(2), 724–731 (1987)
Bakirov, N.K., Székely, G.J.: Student's t-test for Gaussian scale mixtures. J. Math. Sci. **139**(3), 6497–6505 (2006)
Beran, R.: Prepivoting Test Statistics: A Bootstrap View of Asymptotic Refinements. J. Am. Stat. Assoc. **83**(403), 687–697 (1988)
Bickel, P.J., Götze, F., van Zwet, W.R.: Resampling fewer than n observations: gains, losses, and remedies for losses. Statist. Sin. **7**, 1–32 (1997)
Cohen, G., Ladaique, M.: Drivers of Growing Income Inequalities in OECD and European Countries. In: Carmo, R.M., Rio, C., Medgyesi, M. (eds.) Reducing Inequalities: A Challenge for the European Union?, pp. 31–43. Springer International Publishing, Cham (2018)
Cowell, F.A.: Measuring Inequality. London School of Economics Perspectives in Economic Analysis (2009)
Cowell, F.A., Flachaire, E.: Income distribution and inequality measurement: The problem of extreme values. J. Econ. **141**(2), 1044–1072 (2007)
Davidson, R.: Reliable inference for the Gini index. J. Econ. **150**(1), 30–40 (2009)
Davidson, R., Flachaire, E.: Asymptotic and bootstrap inference for inequality and poverty measures. J. Econ. **141**(1), 141–166 (2007)

Dufour, J.M. Flachaire, E., Khalaf, L., Zalghout, A.: Identification-Robust Inequality Analysis. Working Paper, Series Cahiers de recherche (2020)
Dufour, J.M., Flachaire, E., Khalaf, L.: Permutation Tests for Comparing Inequality Measures. J. Bus. Econ. Stat. **37**(3), 457–470 (2019)
Hill, B.M.: A simple general approach to the inference about the tail of a distribution. Ann. Stat. **3**(5), 1163–1174 (1975)
Horowitz, J.: The bootstrap. In: Handbook of econometrics, vol. 5, chap. 52, pp. 3159–3228. Elsevier (2001)
Ibragimov, M., Ibragimov, R., Karimov, J., Yuldasheva, G.: Robust Analysis of Income Inequality Dynamics in Russia: t-Statistic Based Approaches. wiiw Balkan Observatory Working Papers No. 105 (2013)
Ibragimov, R., Kattuman, P., Skrobotov, A.: Robust Inference on Income Inequality: t-Statistic Based Approaches. Working Paper, SSRN Electronic Journal. (2021)
Ibragimov, M., Ibragimov, R.: Heavy tails and upper-tail inequality: The case of Russia. Empir. Econ. **54**(2), 823–837 (2018)
Ibragimov, R., Müller, U.K.: t-Statistic Based Correlation and Heterogeneity Robust Inference. J. Bus. Econ. Stat. **28**(4), 453–468 (2010)
Ibragimov, R., Müller, U.K.: Inference with few heterogeneous clusters. Rev. Econ. Stat. **98**(1), 83–96 (2016)
Moran, T.P.: Statistical Inference for Measures of Inequality With a Cross-National Bootstrap Application. Sociol. Methods Res. **34**(3), 296–333 (2006)
Schluter, C., van Garderen, K.J.: Edgeworth expansions and normalizing transforms for inequality measures. J. Econ. **150**(1), 16–29 (2009)
Tadikamalla, P.R.: A Look at the Burr and Related Distributions. Int. Stat. Rev. Rev. Int. Stat. **48**(3), 337 (1980)
Welch, B.L.: The generalisation of student's problems when several different population variances are involved. Biometrika **34**(1–2), 28–35 (1947)