The Journal of Economic Inequality
https://doi.org/10.1007/s10888-023-09600-x

Identification-robust methods for comparing inequality with an application to regional disparities

Jean-Marie Dufour[1,2] · Emmanuel Flachaire[3] · Lynda Khalaf[4] · Abdallah Zalghout[5]

Received: 29 November 2021 / Accepted: 6 October 2023

Abstract
We propose Fieller-type methods for inference on generalized entropy inequality indices in the context of the two-sample problem which covers testing the statistical significance of the difference in indices, and the construction of a confidence set for this difference. In addition to irregularities arising from thick distributional tails, standard inference procedures are prone to identification problems because of the ratio transformation that defines the considered indices. Simulation results show that our proposed method outperforms existing counterparts including simulation-based permutation methods and results are robust to different assumptions about the shape of the null distributions. Improvements are most notable for indices that put more weight on the right tail of the distribution and for sample sizes that match macroeconomic type inequality analysis. While irregularities arising from the right tail have long been documented, we find that left tail irregularities are equally important in explaining the failure of standard inference methods. We apply our proposed method to analyze income per-capita inequality across U.S. states and non-OECD countries. Empirical results illustrate how Fieller-based confidence sets can: (i) differ consequentially from available ones leading to conflicts in test decisions, and (ii) reveal prohibitive estimation uncertainty in the form of unbounded outcomes which serve as proper warning against flawed interpretations of statistical tests.

Keywords Inequality · Generalized entropy · Two samples · Fieller · Identification-robust

1 Introduction

Economic inequality can be broadly defined in terms of the distribution of economic variables, which include income, consumption or health. Various inequality indices have been proposed in the literature, and the Generalized Entropy class has featured prominently in theoretical and empirical studies. Indeed, interest in this family, which includes the Theil

✉ Lynda Khalaf
Lynda_Khalaf@carleton.ca

Extended author information available on the last page of the article

Published online: 14 February 2024

index as a special case, stems largely from its attractive axiomatic properties.[1] This paper proposes improved measures of estimation uncertainty for GE indices, to address a number of statistical irregularities that characterize their sampling distributions.[2] Our framework covers the problem of comparing two indices *i.e.* the so-called two sample problem, which typically involves: (i) testing the statistical significance of the difference in indices, and (ii) the construction of a confidence set for this difference. A confidence set provides much more information than a test, which guards against spurious interpretations of non-rejection when estimation uncertainty is excessively large.

Statistical inference on inequality indices is an enduring challenge that has recaptured the attention of econometricians in the last two decades. One reason behind the poor performance of available methods is that underlying distributions often have thick tails, which contaminate standard asymptotic and bootstrap-based procedures (Davidson and Flachaire 2007; Cowell and Flachaire 2007). Another reason is that two different distributions can yield an identical inequality level, which complicates comparisons (Dufour et al. 2019). While problems may persist in large samples, size distortions are particularly more prominent with samples that match macro-economic data. Resulting spurious statistical decisions are thus more imminent with *e.g.* international inequality analysis, or with dispersion analysis across regions within a country. While within country income inequality has long engaged interest, there is also an abundant literature on country and region level inequality (Deaton 2021; McCann 2020; Young et al. 2008). We provide a constructive solution that is easy to apply, aiming to improve type I error control particularly in such contexts.

The majority of available inference methods for inequality indices focuses on the one-sample problem where the interest is in comparing a measure to a given value (Davidson and Flachaire 2007; Dufour et al. 2018; Cowell and Flachaire 2007). A notable exception is the permutational approach of Dufour et al. (2019) for the two-sample testing case, which is shown to outperform other available asymptotic and bootstrap alternatives unless underlying distributions differ sizably. More broadly, another difficulty we raise here results from definitional discontinuities. Indeed, GE indices can be written as ratios of moments. So by definition, these indices involve transformations that may be ill-defined over some parameter subspace (for example, as the denominator tends to zero). This yields identification problems, where identification refers to our ability to recover objects of interest from available models and data (Dufour and Hsiao 2008). Despite a sizeable literature in econometric theory on the consequences of such problems, these have escaped formal notice in the case of inequality indices.[3] Addressing identification failures is our objective in this paper.

A brief synopsis of the identification problem is helpful at this stage. For a parameter transformation that is not identified over the full parameter space, a valid confidence interval should be *unbounded* with a non-zero probability (Koschat et al. 1987; Gleser and Hwang 1987; Dufour 1997; Dufour and Taamouti 2005, 2007; Bertanha and Moreira 2020). Validity here refers to coverage or type I error control. Standard errors for the GE indices are usually computed using the Delta method which yields a confidence interval with bounded limits (Cowell and Flachaire, 2015, Chapter 6). The Delta method thus violates the above validity

[1] These include scale invariance, the Pigou-Dalton transfer, the symmetry and the Dalton population principle. It is also additively decomposable. See Cowell (2000) for a detailed discussion on these and other properties of indices.

[2] The literature on statistical inference for inequality measures is relatively recent; see Cowell and Flachaire (2015) for a comprehensive survey.

[3] See *e.g.* Dufour (1997), Andrews and Cheng (2013), Kleibergen (2005), Andrews and Mikusheva (2015), Beaulieu et al. (2013), Bertanha and Moreira (2020), and references therein; see also Bahadur and Savage (1956) and Gleser and Hwang (1987).

requirement. Although lesser-known, an alternative method is available that does not suffer from this shortcoming: the procedure proposed by Fieller (1954) - for inference on the ratio of means of two independent normal random variables - can produce a bounded interval, the complement of a bounded interval or even the real line. In this paper, we propose and validate Fieller-type methods for set inference on the GE family of inequality indices in the context of the two-sample problem. Although it seems puzzling for a confidence set to be unbounded, this is provably inevitable for error control. Effectively, an unbounded outcome may serve as a proper warning about factual imprecision.

On the above, this paper has several contributions. *First*, we provide analytical and tractable solutions for proposed confidence sets. *Second*, we show in a simulation study that proposed solutions are more reliable than Delta-method counterparts. *Third*, we find that our approach outperforms most simulation-based alternatives including the permutation test of Dufour et al. (2019). *Fourth*, our solution covers tests for any given value of the difference [i.e. not just zero, in contrast with Dufour et al. (2019)], allowing the construction of confidence sets through test inversion.[4] Because these sets can be unbounded which signals weak identification, they prevent misinterpreting insignificant tests to confirm the no change hypothesis, when failing to reject this hypothesis is due to prohibitive variability. *Fifth*, we provide useful empirical evidence supporting the seemingly counter-intuitive bounds that Fieller-type methods can produce. Taken collectively, our results document the usefulness of the Fieller approach for international or country based inequality analysis, which covers widely popular active and policy-relevant research priorities on inequality.

Fieller's original solution was extended to multivariate normals (Bennett 1959), general exponential (Cox 1967) and linear (Zerbe 1978; Dufour 1997) regression models, dynamic models with possibly persistent covariates (Bernard et al. 2007, 2019) and for simultaneous inference on multiple ratios (Bolduc et al. 2010). For a good review of inference on ratios, see Franz (2007). The Fieller method has now gained popularity in the literature on weak instruments (Andrews et al. 2019), and the macro-economic literature on structural impulse responses (Olea et al. 2021). Our paper brings in novel insights through the inequality case. Indeed, our simulation results illustrate the superiority of Fieller-type methods for sample sizes that are compatible with macro-economic data. Further simulation results can be summarized as follows. (1) size improvements over the Delta method are especially notable for indices that put more weight on the right tail of the distribution *i.e.* as the sensitivity parameter (denoted γ below) increases; (2) size improvements preserve power; (3) results are robust to different assumptions on the shape of the null distributions; (4) tests based on the Fieller-type method outperform available permutation tests when the distributions under the null hypothesis are different. A permutational approach is not available (to date) for the general problem we consider here. Overall, while irregularities arising from the right tail have long been documented, we find that left-tail irregularities are equally important in explaining the failure of standard inference methods for inequality indices.

To demonstrate the practical relevance of these results, we conduct two empirical studies on macroeconomic inequality. Early work in this regard can be traced back to Theil (1979), Ram (1979), and Maasoumi and Jeong (1985). This literature has further developed in the last two decades due to data availability, the increased level of globalization, and the revived interest in international and regional policy circles about economic convergence (Milanovic 2011; Barro 2012). Using per-capita income data for 48 U.S. states, we compare cross-state inequality in the US, between 1946 and 2016. We next compare inequality across non-OECD countries, between 1960 and 2013. We find that inter-state inequality has declined in the US

[4] Inverting a test means collecting the parameter values that are not rejected by this test at a given level.

over the considered period; both Delta and Fieller based confidence sets for the difference in the GE_2 indices (which we formally define below) are bounded, yet while the former covers zero, the latter does not. Further consequential conflicts emerge with non-OECD countries. In this case, the Fieller method for the difference in GE_2 indices returns the real line, whereas the Delta method produces a seemingly precise interval that covers zero. It is tempting to lend credibility to the latter result, which suggests that inequality remained unchanged. Instead, it is precisely the Fieller result that is of interest, since it reveals that sampling variability is too large for us to draw any conclusion from the data.

The rest of the paper is organized as follows. Section 2 derives Fieller-type confidence sets. Section 3 reports the results of the simulation study. Section 4 contains the inter-state convergence application, and Section 5 concludes.

2 Fieller-type confidence sets for generalized entropy inequality indices

An inequality index is a measure of dispersion for the distribution of a random variable. Many such indices, including the GE class, solely depend on the underlying distribution and can typically be written as a functional which maps the space of the cumulative distribution function (CDF) to the nonnegative real line $\mathbb{R}_+^0$. Let X be a positive random variable such that both moments $\mathbb{E}_F(X)$ and $\mathbb{E}_F(X^\gamma)$ are finite, *i.e.*

$$\mathbb{P}[X > 0] > 0\,, \quad 0 < \mu_X := \mathbb{E}_F(X) < \infty\,, \quad 0 < \nu_X(\gamma) := \mathbb{E}_F(X^\gamma) < \infty\,. \tag{2.1}$$

Then the $GE_\gamma(X)$ measure can be expressed as in Shorrocks (1980):

$$\begin{aligned} GE_\gamma(X) &= \tfrac{1}{\gamma(\gamma-1)}\left[\tfrac{\mathbb{E}_F(X^\gamma)}{[\mathbb{E}_F(X)]^\gamma} - 1\right] \quad \text{for} \quad \gamma \neq 0,\ 1\,, \\ GE_0(X) &= \mathbb{E}_F[log(X)] - log[\mathbb{E}_F(X)] \\ GE_1(X) &= \tfrac{\mathbb{E}_F[X\log(X)]}{\mathbb{E}_F(X)} - \log[\mathbb{E}_F(X)]\,. \end{aligned} \tag{2.2}$$

This class of indices includes several common ones, including two well-known indices introduced by Theil (1967): the Mean Logarithmic Deviation (MLD), which is the limiting value of the $GE_\gamma(X)$ as γ approaches zero, and the Theil index, which is the limiting value of the $GE_\gamma(X)$ as γ approaches 1. When $\gamma = 2$, the index is equal to half the squared coefficient of variation and is related to the Hirschman-Herfindahl (HH) index, used in industrial organization (Schluter 2012). The Atkinson index can be obtained from the $GE_\gamma(X)$ index using an appropriate transformation.

Throughout this paper, it will be convenient to focus on income distributions, though our results also apply to other variables relevant to inequality studies, such as wage, health, and consumption distributions. Our aim is to make inference on the GE_γ measure for a given $\gamma \in (0, 2)$. In particular, we wish to build a confidence set for the difference between two indices. For presentation ease, the following discussion sets $\gamma \neq 1, 0$. The Theil index can be treated along the same lines, beginning from the expressions in Eq. 2.2. The MLD measure eschews the statistical irregularities we raise here; for further insights, see Cowell et al. (2018) and the references therein.

To formalize the considered inference problem and maintaining the above notation, let X and Y refer to two random variables with distributions F_X and F_Y that satisfy Eq. 2.1. In this context, our object of interest corresponds to

$$\Delta GE_\gamma := GE_\gamma(X) - GE_\gamma(Y) = \frac{\nu_X(\gamma)\mu_Y^\gamma - \nu_Y(\gamma)\mu_X^\gamma}{\gamma(\gamma-1)\mu_Y^\gamma \mu_X^\gamma}\,, \tag{2.3}$$

where $GE_\gamma(X)$ and $GE_\gamma(Y)$ take the Eq. 2.2 form, and μ_X, ν_X, μ_Y, and ν_Y denote the underlying moments. For further reference, let $\boldsymbol{\lambda}$ denote the vector:

$$\boldsymbol{\lambda} = (\lambda_1,\ \lambda_2,\ \lambda_3,\ \lambda_4)' = (\mu_X,\ \nu_X,\ \mu_Y,\ \nu_Y)'. \tag{2.4}$$

Thus defined, Eq. 2.3 involves a non-linear function of moments that is not continuous throughout its domain, which implies that ΔGE_γ is not identified throughout the parameter space. In this case, the (above cited) econometric literature proves that usual "standard errors" provide a flawed assessment of sampling precision, in the following sense: the confidence interval of the usual form

$$[\text{estimator} \pm \text{critical point} \times \text{standard error of the estimate}]$$

will fail to cover the true parameter value at the hypothesized level. As a matter of fact, any confidence set with bounded limits will suffer from the same distortion. Consequently, there is interest in finding an alternative procedure, whereby a basic requirement is to avoid sets that are necessarily bounded.

Assume we have *i.i.d.* samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$. Our analysis covers two cases defined by the following assumptions.

Assumption 2.1 *Samples are of unequal sizes and independent.*

Assumption 2.2 *Samples are of equal sizes and dependent.*

With no further parametric assumptions on F_X and F_Y, there is already a well-developed theory[5] for asymptotic inference on GE measures through the empirical distribution functions (EDFs):

$$\hat{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \le x)\,, \quad \hat{F}_Y(y) = \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}(Y_j \le y)\,, \tag{2.5}$$

where $\mathbf{1}(\cdot)$ is the indicator function that takes the value 1 if the argument is true, and 0 otherwise. Under standard laws of large numbers, we can consistently estimate the index $GE_\gamma(X)$ and $GE_\gamma(Y)$ by

$$\widehat{GE_\gamma}(X) := \frac{1}{\gamma(\gamma-1)}\left[\frac{\hat{\nu}_X(\gamma)}{\hat{\mu}_X^\gamma} - 1\right], \quad \widehat{GE_\gamma}(Y) := \frac{1}{\gamma(\gamma-1)}\left[\frac{\hat{\nu}_Y(\gamma)}{\hat{\mu}_Y^\gamma} - 1\right], \tag{2.6}$$

where

$$\hat{\mu}_X := \int x\, d\hat{F}_X = \frac{1}{n}\sum_{i=1}^{n} X_i\,, \quad \hat{\nu}_X(\gamma) := \int x^\gamma d\hat{F}_X = \frac{1}{n}\sum_{i=1}^{n} X_i^\gamma\,, \tag{2.7}$$

$$\hat{\mu}_Y := \int y\, d\hat{F}_Y = \frac{1}{m}\sum_{j=1}^{m} Y_j\,, \quad \hat{\nu}_Y(\gamma) := \int y^\gamma d\hat{F}_Y = \frac{1}{m}\sum_{j=1}^{m} Y_j^\gamma\,. \tag{2.8}$$

[5] See (Cowell and Flachaire, 2015, Chapter 6) for a review and references.

ΔGE_γ can then be estimated using estimates of the relevant moments [see Eqs. 2.7 - 2.8]:

$$\widehat{\Delta GE_\gamma} := \widehat{GE_\gamma}(X) - \widehat{GE_\gamma}(Y) = \frac{\hat{\nu}_X(\gamma)\hat{\mu}_Y^\gamma - \hat{\nu}_Y(\gamma)\hat{\mu}_X^\gamma}{\gamma(\gamma-1)\hat{\mu}_Y^\gamma \hat{\mu}_X^\gamma}\,. \tag{2.9}$$

Let $\hat{\boldsymbol{\lambda}}$ refer to the estimate of $\boldsymbol{\lambda}$ based on Eqs. 2.7 - 2.8. Assuming that these estimated moments are asymptotically normal, we have:

$$\mathrm{T}^{-1/2}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \xrightarrow{D} N(\mathbf{0},\ \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} := \left[\sigma_{ij}\right]_{i,j=1,\ldots,4} \tag{2.10}$$

where

$$\mathrm{T} := \tau \otimes I_2, \quad \tau := \begin{bmatrix} n & 0 \\ 0 & m \end{bmatrix} \tag{2.11}$$

and I_2 is the 2×2 identity matrix. Using these estimates and limiting distributions, we derive the Wald-type Delta method (DCS) and our proposed Fieller-based alternative confidence set (FCS) for each of Assumptions 2.1 - 2.2. These cases will actually differ only by the expression of the variance of the estimator. Thus to avoid redundancy, we will derive the method in its most general form, and state the restrictions required to obtain the relevant formulae otherwise.

It is natural to begin the discussion with the standard DCS. To pave the way for the introduction of its FCS counterpart, we frame our discussion in terms of test inversion, where it is worth recalling that inverting a test with respect to the parameter tested, means collecting the values of the parameter for which the underlying null hypothesis is not rejected at a given significance level α. Presented from such a perspective, the DCS can be obtained by inverting the square (or the absolute value) of the t-test associated with

$$H_D(\Delta_0) : \Delta GE_\gamma = \Delta_0 \tag{2.12}$$

where Δ_0 is any known admissible value of ΔGE_γ, including possibly $\Delta_0 = 0$, for equality. Given our distributional assumptions, consider the usual statistic of the form $(\widehat{\Delta GE_\gamma} - \Delta_0)/\hat{\mathrm{V}}[\widehat{\Delta GE_\gamma}]^{1/2}$, where $\widehat{\Delta GE_\gamma} = \widehat{GE_\gamma}(X) - \widehat{GE_\gamma}(Y)$ and $\hat{\mathrm{V}}[\widehat{\Delta GE_\gamma}]$ is the estimate of the asymptotic variance. Under Assumption 2.1, the asymptotic variance $V(\widehat{\Delta GE_\gamma})$ in Eq. 2.16 is given by:

$$\mathrm{V}(\widehat{\Delta GE_\gamma}) = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\partial \Delta GE_\gamma}{\partial \lambda_i}\frac{\partial \Delta GE_\gamma}{\partial \lambda_j}\sigma_{ij} + \frac{1}{m}\sum_{i=3}^{4}\sum_{j=3}^{4}\frac{\partial \Delta GE_\gamma}{\partial \lambda_i}\frac{\partial \Delta GE_\gamma}{\partial \lambda_j}\sigma_{ij}\,, \tag{2.13}$$

whereas under Assumption 2.2, we obtain:

$$\mathrm{V}(\widehat{\Delta GE_\gamma}) = \frac{1}{n}\sum_{i=1}^{4}\sum_{j=1}^{4}\frac{\partial \Delta GE_\gamma}{\partial \lambda_i}\frac{\partial \Delta GE_\gamma}{\partial \lambda_j}\sigma_{ij}\,. \tag{2.14}$$

The estimate $\widehat{\mathrm{V}}(\widehat{\Delta GE_\gamma})$ may then be computed by replacing σ_{ij} with $\hat{\sigma}_{ij}$,and λ_i with $\hat{\lambda}_i$.

In practice, inverting a test based on $(\widehat{\Delta GE_\gamma} - \Delta_0)/\hat{\mathrm{V}}[\widehat{\Delta GE_\gamma}]^{1/2}$ can be carried out by solving the following inequality for Δ_0:

$$(\widehat{\Delta GE_\gamma} - \Delta_0)^2 \le z_{\alpha/2}^2 \hat{\mathrm{V}}[\widehat{\Delta GE_\gamma}] \tag{2.15}$$

where $z_{\alpha/2}$ is the asymptotic two-tailed critical value at the significance level α (*i.e.*, $\mathbb{P}[Z \geq z_{\alpha/2}] = \alpha/2$ for $Z \sim N[0,1]$). The solution of Eq. 2.15 yields the familiar Delta-method set

$$\text{DCS}(\Delta GE_\gamma; 1-\alpha) = \left[\widehat{\Delta GE_\gamma} \pm z_{\alpha/2}\,[\hat{\text{V}}(\widehat{\Delta GE_\gamma})]^{1/2}\right], \tag{2.16}$$

which is an interval with bounded limits, and will thus be subject to the above irregularities. We next build on the test inversion representation of Eq. 2.16 to describe our alternative Fieller-type procedure.

Note that the broad approach of test inversion requires a test that can serve to assess a non-zero Δ_0. Any method that can only assess equality of indices (including the permutation test of Dufour et al. (2019)) cannot be used for our purposes. Yet another fundamental concern stems from the discontinuities in the definition of ΔGE_γ. One way of dealing with this is to test a linearized counterpart of $H_D(\Delta_0)$ rather than Eq. 2.12. Conformably, we thus reformulate the null hypothesis into the following linear form (without the ratio transformation)

$$H_F(\Delta_0) : \Theta(\Delta_0) = 0 \qquad \text{where} \qquad \Theta(\Delta_0) := \theta_1 - \theta_2\Delta_0 \tag{2.17}$$

where θ_1 and θ_2 are the numerator and the denominator in Eq. 2.3

$$\theta_1 = \nu_X(\gamma)\mu_Y^\gamma - \nu_Y(\gamma)\mu_X^\gamma, \quad \theta_2 = \gamma(\gamma-1)\mu_Y^\gamma\mu_X^\gamma\,. \tag{2.18}$$

We then consider the t-test of $H_F(\Delta_0)$, with acceptance region:

$$\hat{\Theta}(\Delta_0)^2 \leq z_{\alpha/2}^2\,\hat{\text{V}}[\hat{\Theta}(\Delta_0)] \tag{2.19}$$

where we use the moment-type estimators based on Eqs. 2.7 - 2.8, *i.e.*

$$\hat{\Theta}(\Delta_0) := \hat{\theta}_1 - \hat{\theta}_2\Delta_0, \quad \hat{\theta}_1 := \hat{\nu}_X(\gamma)\hat{\mu}_Y^\gamma - \hat{\nu}_Y(\gamma)\hat{\mu}_X^\gamma, \quad \theta_2 := \gamma(\gamma-1)\hat{\mu}_Y^\gamma\hat{\mu}_X^\gamma \tag{2.20}$$

and $\hat{\text{V}}[\hat{\Theta}(\Delta_0)]$ is a consistent estimator of $\text{V}[\hat{\Theta}(\Delta_0)$, the asymptotic variance of $\hat{\Theta}(\Delta_0)$ under $H_F(\Delta_0)$. Note that the latter consistency needs to hold only under the null hypothesis $H_F(\Delta_0)$. On using the asymptotic normality assumption Eq. 2.10, the acceptance region Eq. 2.19 yields a confidence set for ΔGE_γ with level $1-\alpha$ (asymptotically):

$$\text{FCS}[\Delta GE_\gamma; 1-\alpha] = \{\Delta_0 : \hat{\Theta}(\Delta_0)^2 \leq z_{\alpha/2}^2\hat{\text{V}}[\hat{\Theta}(\Delta_0)]\}. \tag{2.21}$$

We call $\text{FCS}[\Delta GE_\gamma; 1-\alpha]$ the level-$(1-\alpha)$ *Fieller-type confidence set* for ΔGE_γ. Estimating $\text{V}[\hat{\Theta}(\Delta_0)]$ will require estimating the asymptotic covariance of $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)'$. For future reference, we denote the latter and the corresponding estimator as follows:

$$\text{V}(\hat{\theta}) = \begin{bmatrix} \text{V}(\hat{\theta}_1) & \text{C}(\hat{\theta}_1, \hat{\theta}_2) \\ \text{C}(\hat{\theta}_1, \hat{\theta}_2) & \text{V}(\hat{\theta}_2) \end{bmatrix}, \quad \hat{\text{V}}(\hat{\theta}) = \begin{bmatrix} \hat{\text{V}}(\hat{\theta}_1) & \hat{\text{C}}(\hat{\theta}_1, \hat{\theta}_2) \\ \hat{\text{C}}(\hat{\theta}_1, \hat{\theta}_2) & \text{V}(\hat{\theta}_2) \end{bmatrix}. \tag{2.22}$$

The form of the Fieller-type confidence set may not be clear from Eq. 2.21. The following theorem characterizes $\text{FCS}[\Delta GE_\gamma; 1-\alpha]$ in an explicit way.

Theorem 2.1 *Let* $\hat{\text{V}}(\hat{\theta})$ *be an estimate of* $\text{V}(\hat{\theta})$ *in Eq. 2.22. Then the confidence set* $\text{FCS}[\Delta GE_\gamma; 1-\alpha]$ *defined in Eq. 2.21 can be computed as follows*:

$$\text{FCS}[\Delta GE_\gamma; 1-\alpha] = \{\Delta_0 : A\Delta_0^2 + B\Delta_0 + C \leq 0\}$$

$$= \begin{cases} \left[\frac{-B-\sqrt{D}}{2A}, \frac{-B+\sqrt{D}}{2A}\right] & \text{if } D \geq 0 \text{ and } A > 0 \\ \left]-\infty, \frac{-B+\sqrt{D}}{2A}\right] \cup \left[\frac{-B-\sqrt{D}}{2A}, +\infty\right[& \text{if } D \geq 0 \text{ and } A < 0 \\ \left]-\infty, -\frac{C}{B}\right] & \text{if } A = 0 \text{ and } B > 0 \\ \left[-\frac{C}{B}, \infty\right[& \text{if } A = 0 \text{ and } B < 0 \\ \mathbb{R} & \text{if } [A = B = 0 \text{ and } C \leq 0] \text{ or } [D < 0 \text{ and } A \leq 0] \\ \emptyset & \text{if } [A = B = 0 \text{ and } C > 0] \text{ or } [D < 0 \text{ and } A > 0] \end{cases} \tag{2.23}$$

where

$$A := \hat{\theta}_2^2 - z_{\alpha/2}^2 \hat{\mathrm{V}}(\hat{\theta}_2), \quad B := -2[\hat{\theta}_1\hat{\theta}_2 - z_{\alpha/2}^2 \hat{\mathrm{C}}(\hat{\theta}_1, \hat{\theta}_2)], \quad C := \hat{\theta}_1^2 - z_{\alpha/2}^2 \hat{\mathrm{V}}(\hat{\theta}_1)\,, \tag{2.24}$$

$$\begin{aligned} D := B^2 - 4AC = 4z_{\alpha/2}^2 \{ & [\hat{\theta}_1^2 \hat{\mathrm{V}}(\hat{\theta}_2) + \hat{\theta}_2^2 \hat{\mathrm{V}}(\hat{\theta}_1) - 2\hat{\theta}_1\hat{\theta}_2 \hat{\mathrm{C}}(\hat{\theta}_1, \hat{\theta}_2)] \\ & + z_{\alpha/2}^2 [\hat{\mathrm{C}}(\hat{\theta}_1, \hat{\theta}_2)^2 - \hat{\mathrm{V}}(\hat{\theta}_1)\hat{\mathrm{V}}(\hat{\theta}_2)] \}\,. \end{aligned} \tag{2.25}$$

If furthermore $\hat{\mathrm{V}}(\hat{\theta})$ *is positive definite, then*

$$D < 0 \implies [A < 0 \text{ and } C < 0]\,. \tag{2.26}$$

The proof is available in Appendix A. Theorem 2.1 allows for non-positive definite matrix $\hat{\mathrm{V}}(\hat{\theta})$ [at least, for the specific sample considered]. When $\hat{\mathrm{V}}(\hat{\theta})$ is positive definite, Eq. 2.26 implies that FCS$[\Delta GE_\gamma; 1-\alpha]$ may be empty only when $A = B = 0$ and $C > 0$, *i.e.* $A\Delta_0^2 + B\Delta_0 + C = C > 0$ [an event with zero probability when $(\hat{\theta}_1, \hat{\theta}_2)'$ has a Gaussian distribution]. Note that the condition $A > 0$ means that θ_2 is significantly different from zero [according to the criterion $\hat{\theta}_2^2/\hat{\mathrm{V}}(\hat{\theta}_2) > z_{\alpha/2}^2$], while $C > 0$ means that θ_1 is significantly different from zero [according to the criterion $\hat{\theta}_1^2/\hat{\mathrm{V}}(\hat{\theta}_1) > z_{\alpha/2}^2$].

Consistent estimation of these depends on the assumptions made on the observations $[X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m]$. For the assumptions 2.1 and 2.2, we get (using the Delta method):

$$\text{under Assumption 2.1}: \quad \mathrm{V}(\hat{\theta}_1) = \frac{1}{n} S_{11}\,, \quad \mathrm{V}(\hat{\theta}_2) = \frac{1}{m} S_{22}\,, \quad \mathrm{C}(\hat{\theta}_1, \hat{\theta}_2) = \frac{1}{n} S_{12} + \frac{1}{m} S_{21}\,, \tag{2.27}$$

$$\text{under Assumption 2.2}: \quad \mathrm{C}(\hat{\theta}_k, \hat{\theta}_l) = \frac{1}{n} \sum_{i=1}^{4} \sum_{j=1}^{4} \frac{\partial \theta_k}{\partial \lambda_i} \frac{\partial \theta_l}{\partial \lambda_j} \sigma_{ij} \quad \text{for } k = 1, 2,\ l = 1, 2, \tag{2.28}$$

where

$$S_{11} := \sum_{i=1}^{4} \sum_{j=1}^{4} \frac{\partial \theta_1}{\partial \lambda_i} \frac{\partial \theta_1}{\partial \lambda_j} \sigma_{ij}\,, \quad S_{22} := \sum_{i=1}^{4} \sum_{j=1}^{4} \frac{\partial \theta_2}{\partial \lambda_i} \frac{\partial \theta_2}{\partial \lambda_j} \sigma_{ij}\,, \tag{2.29}$$

$$S_{12} := \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{\partial \theta_1}{\partial \lambda_i} \frac{\partial \theta_2}{\partial \lambda_j} \sigma_{ij}\,, \quad S_{21} := \sum_{i=3}^{4} \sum_{j=3}^{4} \frac{\partial \theta_1}{\partial \lambda_i} \frac{\partial \theta_2}{\partial \lambda_j} \sigma_{ij}\,. \tag{2.30}$$

The above presumes asymptotic normality of the underlying criteria. In fact, the considered indices are known transformations of two moments the estimators of which are asymptotically normal under standard regularity assumptions; see Davidson and Flachaire (2007) and Cowell

and Flachaire (2007). These typically require that the first two moments exist and are finite. Asymptotic normality of the statistics in Eqs. 2.15 and 2.19 thus follows straightforwardly. Nevertheless, convergence in this context is known to be slow, especially when the distribution of the data is heavy-tailed and with indices that are sensitive to the upper tail. Our simulations confirm these issues, yet the Fieller-based criteria perform better than the Delta method-based counterparts in finite samples because the former eschews problems arising from the ratio.

Extending the above results to other measures is worth considering, and in particular to the Gini coefficient which also takes a ratio form. However, in this case an estimator for the covariance term in $\hat{\mathrm{V}}[\hat{\Theta}(\Delta_0)]$ in Eq. 2.19 is not readily available. The GE class is based on moments, which allowed us to build on the influence function approach in (Cowell and Flachaire, 2015, Chapter 6) to derive the underlying covariance terms. Because of its popularity, extensions of the Fieller method to the Gini index is a worthy objective, however, it is beyond the scope of this paper.

3 Simulation evidence

This section describes a simulation study designed to compare the finite-sample properties of FCS to the standard DCS. This will be done for the two popular inequality indices from the general entropy family: the Theil Index (GE_1), and half of the coefficient of variation squared (GE_2) which is related to the Hirschman-Herfindahl (HH) index. The tables and figures are in the appendix.

We report the *rejection frequencies* of the tests underlying the proposed confidence sets, under both the null hypothesis (level control) and the alternative (power). Under the null hypothesis, these can also be interpreted as 1 minus the corresponding *coverage probability* for the associated confidence set. So we are studying here both the operating characteristics of tests used and the coverage probabilities of the confidence sets defined above. For further insight on confidence set properties, we also study the width of the bounded sets.

Since available inference methods perform poorly when the underlying distributions are heavy-tailed, we designed our simulation experiments to cover such distributions by simulating the data from the Singh-Maddala distribution, which was found to successfully mimic observed income distributions for developed countries such as Germany (Brachmann et al. 1995). Another reason to use the Singh-Maddala distribution is that it was widely used in the literature which makes our results directly comparable to previously proposed inference methods. The CDF of the Singh-Maddala distribution can be written as

$$F_X(x) = 1 - \left[1 + \left(\frac{x}{b_X}\right)^{a_X}\right]^{-q_X} \tag{3.1}$$

where a_X, q_X and b_X are the three parameters defining the distribution. a_X influences both tails, while q_X only affects the right tail. b_X is a scale parameter to which we give little attention as the inequality indices considered in this paper are scale-invariant. This distribution is a member of the five-parameter generalized beta distribution and its upper tail behaves like a Pareto distribution with a tail index equal to the product of the two shape parameters a_X and q_X ($\xi_X = a_X q_X$). The k-th moment exists for $-a_X < k < \xi_X$ which implies that a sufficient condition for the mean and the variance to exist is $-a_X < 2 < \xi_X$.

The moment of order γ of Singh-Maddala distribution have the following closed form:

$$\mathbb{E}(X^\gamma) = \frac{b_X^\gamma \, \Gamma\left(\gamma a_X^{-1} + 1\right) \Gamma\left(q_X - \gamma a_X^{-1}\right)}{\Gamma\left(q_X\right)} \tag{3.2}$$

where $\Gamma(\cdot)$ is the gamma function. For $\gamma = 1$, this yields the mean of X [$\mu_X = \mathbb{E}(X)$] and, for $\gamma = 2$, the second moment of X [$\nu_X = \mathbb{E}(X^2)$]. Similarly, replacing X by Y in the above expressions, we can compute μ_Y and ν_Y. Using the values of these moments, we compute analytical expressions for $GE_\gamma(X)$ and $GE_\gamma(Y)$. Each experiment involves 10000 replications. The nominal level α is set at 5%.

The hypotheses of interest take the form $H_0(\gamma) : GE_\gamma(X) - GE_\gamma(Y) = \Delta_0$, for $\gamma = 1$ or 2. Even though we emphasize the important problem of testing equality ($\Delta_0 = 0$), we also consider the problem of testing nonzero differences ($\Delta_0 \neq 0$) to document the coverage of the confidence sets. Our simulation experiments rely on the following designs.

1. Independent samples of unequal sizes ($m \neq n$):
   (a) $\Delta_0 = 0$ with $F_X = F_Y$; (b) $\Delta_0 = 0$ with $F_X \neq F_Y$; and (c) $\Delta_0 \neq 0$ (hence $F_X \neq F_Y$).
2. Dependent samples of equal sizes ($m = n$):
   (a) $\Delta_0 = 0$ with $F_X = F_Y$; (b) $\Delta_0 = 0$ with $F_X \neq F_Y$; and (c) $\Delta_0 \neq 0$ (hence $F_X \neq F_Y$).

We consider sample sizes ranging from 50 to 2000 observations which include sizes that match macroeconomic type inequality analysis. We aim to document the usefulness of our proposed methodology in such problems, for which corrections are most needed because of data limitations. We also consider samples as large as 200000 for some designs as reported below. For most experiments, the simulation results are presented graphically through plotting the rejection frequencies against the number of observations. When the number of observations is different between the two samples, we plot the rejection frequencies against the number of observations of the smallest sample.

For the Delta method, we use the critical region $[\widehat{\Delta GE_\gamma} - \Delta_0]^2 > z^2_{\alpha/2}\hat{\mathrm{V}}[\widehat{\Delta GE_\gamma}]$, based on Eq. 2.15; for the Fieller method, we use the critical region $\hat{\Theta}(\Delta_0)^2 > z^2_{\alpha/2}\hat{\mathrm{V}}[\hat{\Theta}(\Delta_0)]$, as described in Eq. 2.19. Power is investigated by assuming distributions with heavier left and right tails to draw the first sample, and distributions with less heavy left and right tails to draw the second sample. We do so by considering DGPs with a lower value of the shape parameter a_X and a higher value of the shape parameter a_Y. The rejection frequencies under the alternative are not size-controlled, yet we compare power when both methods have similar sizes.

Our extensive simulation study reveals several important results. First, the Fieller-type method outperforms the Delta method under most specifications, and when it does not, it performs as well as the Delta method. Put differently, the Fieller-type method was never dominated by Delta method. Improvements are most useful with sample sizes that match macro-economic data. Second, the Fieller-type method is more robust to irregularities arising from both the left and right tails. Third, the Fieller-type method gains become more sizeable as the sensitivity parameter γ increases. Fourth, the performance of the Fieller-type method matches, and for some cases exceeds, the permutation method which is considered one of the best performing methods proposed in the literature so far for the two-sample problem. In the remainder of this section we take a closer look at the simulation evidence supporting the above findings.

Independent samples of unequal sizes Empirically, when comparing inequality levels spatially or over time, it is unlikely one encounters samples with the same size. Thus, it is useful to assess the performance of our proposed method when the sample sizes are unequal. To do so, we set the number of observations of the second sample to be as twice as large as the first sample. If we denote the size of the first sample by n and that of the second by m, then

$n = 2m$.[6] The left panels of Figs. 1 and 2 depict the rejection frequencies against the sample size for GE_1 and GE_2 respectively. Here the distributions are assumed identical [$F_X = F_Y$]. Comparing the two panels, we notice that better size control with the Fieller-type method is more noticeable for GE_2: the size gains are larger when the index used is more sensitive to the changes in the right tail of the underlying distributions. As the sample size increases the rejection probabilities of the two methods converge to the same level.

In the second specification, the indices are identical, but the underlying distributions are not [$\Delta_0 = 0$ with $F_X \neq F_Y$]. The left panel of Figs. 3 and 4 plots the FCS and DCS rejection frequencies for this scenario. Again, the results suggest that the Fieller-type method outperforms the Delta method in small samples in terms of size, and the gains are most prominent for GE_2. The gains are smaller in this scenario compared to the previous one. As we will show later, the Fieller-type method will not solve the over-rejection problem under all scenarios, but it will reduce size distortions in many cases, particularly with small samples, and when it does not, it performs as well as the Delta method.

We now move to the third scenario, where we consider different distributions under the null hypothesis and unequal inequality indices [$\Delta_0 \neq 0$]. In this scenario, the difference under the null hypothesis can take any admissible value (possibly different from zero). Testing a zero value, although informative, does not always translate into a confidence interval. Hence, one of our contributions lies in considering the non-zero null hypothesis which allows us to rely for inference on the more-informative confidence sets approach rather than testing the equality of the difference between the two indices to one specific value.

The results, as shown in the left panels of Figs. 5 and 6, suggest important improvements. In both panels, the Fieller-type method leads to size gains and almost achieves correct size. The improvements are more pronounced for the GE_2 index. The right panels of Figs. 1 to 6 illustrate the power of FCS and DCS for both GE_1 and GE_2 under the three scenarios considered: [$\Delta_0 = 0$ with $F_X = F_Y$], [$\Delta_0 = 0$ with $F_X \neq F_Y$] and [$\Delta_0 \neq 0$] respectively. The results show that the Fieller-type method is as powerful as the Delta method when compared at sample sizes where both FCS and DCS have similar empirical rejection frequencies[7].

Dependent samples of equal sizes Another interesting case is the one where the samples are dependent. This occurs mostly when comparing inequality levels before and after a policy change, such as comparing pre-tax and post-tax income inequality levels, or comparing the distributional impact of a macroeconomic shock. To accommodate for such dependencies, we modify the simulation design as follows: the samples are drawn in pairs from the joint distribution, which we denote F_{XY}, where the correlation between the two marginal distributions is generated using a Gumbel copula with a high Kendall's correlation coefficient of 0.8. For this case, results are in line with the independent cases, in small samples and when larger γ is used. Size and power plots under this scenario can be found in the online appendix.

Comparing the Fieller-type method with the permutation method As outlined in the introduction, the permutation-based Monte-Carlo test approach proposed in Dufour et al. (2019) stands out as one of the best performing nonparametric inference method for testing the equality of two inequality indices. The authors focus on the GE_1 and the Gini indices. The permutation testing approach provides exact inference when the null distributions are identical ($F_X = F_Y$) and it leads to a sizeable size distortion reduction when the null distributions are

[6] The results presented here are not sensitive to choice of the ratio between n and m

[7] The simulation results for the scenario of independent samples with equal size are similar to those obtained in the first experiment and can be found in the online appendix.

sufficiently close ($F_X \approx F_Y$). However, as the null distributions differ, the performance of the method deteriorates.

Figures 7 and 8 plot size and power of the permutation Fieller-type methods against the tail index of F_Y. As in Dufour et al. (2019), we fix the tail index of the null distribution F_X to 4.76. When the distributions under the null hypothesis are identical, the permutation method is exact and thus it is important to compare methods when exactness does not hold. For the GE_1 index, as shown in Fig. 7, the Fieller-type method leads to considerable size and power gains especially when the right tail of the second distribution is much heavier than that of the first distribution ($\xi_X < \xi_Y$). These gains are more pronounced when considering the GE_2 index. The attraction of the Fieller-type method with respect to the permutation approach goes beyond the superior performance highlighted above. Unlike the Fieller-type method, the permutation method applicability is restricted to the null hypothesis of equality ($\Delta_0 = 0$), and further theoretical developments would be needed to test more general hypotheses. Building confidence intervals using a permutation-based or another simulation-based method (such as the bootstrap) would also require a computationally intensive numerical inversion (*e.g.*, through a grid search). So another appealing feature of the Fieller-type approach comes from the fact that it is computationally easy to implement.

Behavior with respect to the tails To better understand under what circumstances does the Fieller-type method improves level control, we assess the performance of the proposed method to different tail shapes. The literature has focused on the role of heavy right tails in the deterioration of the Delta method confidence sets. However, as our results indicate, heavy left tails also contribute to the under-performance of the standard inference procedures. The Fieller-type method is less prone to such irregularities arising from both ends of the distributions and thus it reduces size distortions whether the cause of the under-performance is arising from the left tail or the right tail. This is supported by the results reported below in Tables 3 and 4. Table 3 reports the percentage difference of the rejection frequencies as the right tails of the two distributions become thicker. The right-tail shape is determined by the tail index ($\xi_X = a_X q_X$). The smaller the tail index, the thicker is the right tail of the distribution under consideration. The reliability advantage of the Fieller-type method (over the Delta method) increases as the right tail of the distributions gets thicker.

To study the impact of the left tail, the parameters of the first distribution are fixed at $a_X = 2.8$ and $q_X = 1.7$, while a_Y and q_Y are varied such that the left tail becomes thicker and the right tail is left unchanged. This is done by decreasing a_Y, and increasing q_Y enough to keep the tail index fixed ($\xi_X = \xi_Y = 4.76$). The last column of Table 4 shows the percentage difference of the rejection frequencies between the Fieller-type and Delta methods. As the left tail thickens, the performance of the Delta method deteriorates relative to the Fieller-type method, and thus the Fieller method better captures irregularities in the left tail. This conclusion holds regardless of whether the left tail of the second distribution is lighter or thicker than the left tail of the first distribution.

Fieller-type method and the sensitivity parameter γ A consistent conclusion from our results is that the Fieller's-induced size gains are more prominent for GE_2 compared to GE_1, that is, when the sensitivity parameter γ increases from 1 to 2. This might suggest that as γ increases, size gains from the Fieller-type method increase. Such generalization is indeed supported by simulation evidence illustrated by Fig. 9. The left panel plots rejection frequencies of DCS and FCS for $\gamma \in [0, 3.5]$ for independent samples. The right panel considers dependent samples. As γ becomes larger, FCS outperforms DCS at an increasing rate. The superiority of the Fieller-type method in this context is unaffected by the independence

assumption as shown in the right panel where the rejection frequencies are plotted against γ for dependent samples with Kendall's correlation of 0.8.

Recall that the parameter γ characterizes the sensitivity of the index to changes at the tails of the distribution. For instance, the index becomes more sensitive to changes at the upper tails as γ increases (assuming positive γ). Thus, relative to the Delta method, the performance of the Fieller-type method in the two-sample problem improves as the right tail of the underlying distributions becomes heavier. This conclusion, as we saw from the results above, is robust to the assumptions about the independence of the samples and to the distance between the two null distributions.

The identical performance of the Fieller-type method and Delta method at $\gamma = 0$ is expected as the underlying t-tests inverted in the process of building FCS and DCS are identical since the null hypothesis is no longer a ratio. To see that, recall that the limiting solution for $GE_\gamma(\cdot)$ at $\gamma = 0$ is equal to $\mathbb{E}_F[\log(X)] - \log[\mathbb{E}_F(X)]$. Graphically, we can see that both methods start off at the same rejection frequencies when $\gamma = 0$, and then diverge as γ increases.

Robustness to the shapes of the null distributions So far, our simulation experiments have focused on comparing the finite-sample performance of FCS and DCS by studying their behavior as the number of observation increases, holding the parameters of the two underlying null distributions constant. Here we try to check the robustness of our results by fixing the number of observations at 50 and allowing the parameters (a_X,q_X, a_Y and q_Y) to vary. This type of analysis highlights the (in)sensitivity of our conclusions regarding the Fieller-type method to the shape of the null distributions. In left panel of Fig. 10, we plot the rejection frequencies of both methods against the sensitivity parameter ξ_X for the GE_1 index. We set ξ_X equal to 4.76 and allow ξ_Y to vary between 2.89 and 6.357. In the right panel, we focus on the GE_2 index. Here ξ_X is fixed at 4.76 again and the parameter ξ_Y ranges between 2.89 and 6.357.

For small samples, the gains of the Fieller-type method are maintained regardless the shape of the distribution. The gains are more pronounced for GE_2 compared to GE_1. These two graphs show that the gains attained by the Fieller-type method are not arbitrary and that they hold for various parametric assumptions of the underlying distributions.

Slow convergence Inequality estimates are characterized by slow convergence when underlying distributions are heavy-tailed. This problem has in fact motivated most of the proposed asymptotic refinements in this literature [see Davidson and Flachaire (2007); Cowell and Flachaire (2007)]. Our results in Table 5 corroborate this fact, as over-rejections remain even with samples as large as 200000, particularly with the GE_2 which puts more weight on the upper tail of the distribution. On balance, our main finding is the superiority of the Fieller method in samples of sizes compatible with macroeconomic data.

Widths of the confidence sets The last two columns of Table 5 show the average widths of the FCS and the DCS for the two sample problem. Since the Fieller's method can produce unbounded confidence sets, we take the average of the widths based on the bounded confidence sets. In general, compared to the FCS widths, the DCS widths are shorter with small samples, *i.e.* they are shorter when the Delta method rejection frequencies are higher than those of Fieller. This suggest that the DCS are too short and thus they tend to undercover the true difference between the indices. As the sample size increases, the two methods exhibit similar performance and the widths coincides.

4 Application: income inequality across US states and non-OECD countries

In this section, we present empirical evidence on the relevance of our theoretical results to applied economic work. In view of our simulation results, we focus on a macro-type problem, specifically the analysis of: (i) income per-capita inequality across U.S. states between 1946 and 2016, and (ii) per-capita inequality across non-OECD countries between 1960 and 2013. Empirically, policy-makers are interested in learning about the dynamics of income dispersion across regions/states to plan or assess redistributive policies. Following the terminology proposed by Milanovic (2011), one can differentiate between three approaches to international or regional inequality. The first relies on income per-capita; the second also uses income per-capita but accounts for population differences; the third is a microeconomic approach based on cross-country household surveys. In this paper, we follow the first popular approach. In particular, we aim to illustrate our methodology with standard and publicly available data.

Macro-inequality is typically analysed within the neoclassical growth model, which predicts that income per-capita of less developed countries/regions is expected catch-up with the developed ones in the long run. In contrast, endogenous growth models postulate that once knowledge differentials are factored in, incomes may diverge and the gap might widen in the long run (Romer 1994; Rebelo 1991). Among the various measures of convergence/dispersion provided in the literature, two definitions appear to dominate the work on this topic: β-convergence and σ-convergence (Barro 2012; Barro and Sala-i Martin 1992; Quah 1996; Sala-i Martin 1996; Higgins et al. 2006). Although related, these two measures might lead to different conclusions as they capture different dimensions of economic convergence. For an analytical treatment of the relationship between the two measures, see Higgins et al. (2006). The σ-convergence concept focuses on the dispersion of the income distribution which is typically measured in this literature by the variance of the logs. The variance of logs is scale-independent and thus multiplying the per-capita incomes by a scale k has no impact on the dispersion level. Alternative scale-independent measures of dispersion such as inequality indices have generally not been utilized in convergence analysis. Exceptions include Young et al. (2008) and Evans(1996) who used the Gini coefficient and the variance of logs respectively.

Inequality indices respect the Pigou-Dalton principle, which states that a rank preserving transfer from a richer individual/state to a poorer individual/state should make the distribution at least as equitable. In the context of economic convergence, this principle is particularly relevant. For instance, if the US government makes a transfer from a richer state to a poorer one, one would expect dispersion between states to decline. The Gini and GE indices would capture this decline, whereas the variance of logs might indicate no change or even an increase in dispersion. The fact that the variance of logs violates the Pigou-Dalton principle is usually neglected in the literature on the grounds that the problem occurs only at the extreme right tail of the distribution. However, Foster and Ok (1999) shows that disagreement between the variance of logs and inequality indices can result from changes in incomes in other parts of the distribution including the left tail. The following example (Foster and Ok 1999) underscores the importance of the Pigou-Dalton principle and its implications for convergence. Consider two income distributions defined by the following incomes $(2, 5, 10, 28, 40)$ and $(2, 5, 10, 34, 34)$ where the latter is associated with a transfer from the richest [40 to 34] incomes to poorer ones [28 to 34]. The resulting change in the variance of logs, from 1.5125

Table 1 Estimates and confidence intervals of the change in inequality across U.S. states between 1946 and 2016

	GE_1	GE_2
First sample - 1946	0.02743	0.02679
Second sample -2016	0.0144	0.01516
$GE_\gamma(2016) - GE_\gamma(1946)$	-0.01303	-0.01163
Delta C.I.	$[-0.02486, -0.001204]$ Inequality decreases	$[-0.02349, 0.00024]$ No change in Inequality
Fieller′s C.I.	$[-0.02531, -0.00155]$ Inequality decreases	$[-0.02456, -0.00043]$ Inequality decreases
Permutation test p − Value	0.014 Inequality decreases	0.014 Inequality decreases
Number of states	48	48

to 1.5154, suggests an increase of inequality. In contrast, the GE_2 index declines from 0.3696 to 0.3446, thereby capturing the expected distributional impact of such a transfer.

Our empirical analysis of per-capita income dispersion across the US is motivated by comparably peculiar statistics. Consider the publicly available per-capita income at the state level for 48 out of the 50 states (the data for Alaska and Hawaii is not available). The variance of logs between the years 2000 and 2016 indicates a 3% increase in dispersion, whereas GE_2 indicates a decline in dispersion by 0.3%. This provides a compelling basis for the more comprehensive inferential analysis reported next.

Using the same data source, we first compute the GE_1 index for the per-capita income distributions of 1946 and 2016. Then we construct the Delta and Fieller confidence sets for the difference between the two indices. A standard interpretation of differences between the two confidence intervals (at the considered level) implies that one will reject the null hypothesis $\Delta GE_\gamma = \Delta_0$ for a given Δ_0 while the other fails to reject it. Special attention should be paid to the $\Delta_0 = 0$ case, as decisions might reverse the conclusion on whether convergence holds or not. In what follows tests and confidence sets are at the 5% level.

Based on the GE_1 index, our results in the first column of Table 1 indicate that per-capita income inequality across states has declined between 1946 and 2016. The decline in inequality implies convergence. This is compatible with the general convergence trend reported in the literature (Barro and Sala-i Martin 1992; Bernat Jr 2001; Higgins et al. 2006). Although the Fieller and Delta-method confidence sets are not identical, they still lead to the same conclusion which is that the decline of inequality is statistically different from zero at the level used.

In the second column of Table 1, we consider the same problem using GE_2 index which puts more weight on the right tail of the distribution. In this case, the results also indicate a decline of inequality across states. Inequality in 1946 was 0.02679 and declined by 0.01163 in 2016. The confidence sets based on the Delta and Fieller-type methods lead to opposite conclusions about the statistical significance of the decline in inequality:the former suggests that the decline is insignificant whereas the latter indicates it is significant.

Table 2 Estimates and confidence intervals of the change in inequality across non-OECD countries

	GE_1	GE_2
First sample - 1960	0.717621	1.46631
Second sample -2013	0.78726	1.45076
$GE_\gamma(2013) - GE_\gamma(1960)$	0.06964	−0.01554
Delta C.I.	$[-0.35694, 0.49623]$	$[-1.15143, 1.120337]$
Fieller′s C.I.	$[-0.40436, 0.63075]$	$\mathbb{R}$
Permutation test p − value	0.886	0.992
Number of countries	72	72

In addition to the DCS and FCS, we report the permutational p-values. For the GE_2, the p-value is less than 5% which entails the rejection of the null hypothesis of no change in inequality contradicting the conclusion based on the Delta method. This constitutes an empirical evidence supporting the findings of Dufour et al. (2019).

Two conclusions can be drawn from our findings. First, the Fieller-type and the Delta methods can lead to different confidence sets in practice which documents the empirical relevance of our theoretical findings. Second, disparities between both sets can lead to spurious conclusions about inequality changes if one set includes zero while the other does not. From a policy point of view, this disparity is crucial, especially if important policy actions are motivated by the underlying analysis.

We next turn to non-OECD countries between 1960 and 2013. Table 2 presents estimates and confidence sets for the difference of inequality indices between the two periods. The main finding here is that the Fieller-type confidence set based on the GE_2 index is the whole real line $\mathbb{R}$. These results confirm that decisions based on Delta-method are spurious, and that a no-change conclusion is flawed: data and indices are, instead, uninformative.

The permutational method leads to results similar to Delta and the Fieller-type methods for non-OECD countries. Available permutation tests although preferable size-wise to their standard counterparts, are difficult to invert to build confidence sets. Instead, the confidence sets proposed here can be unbounded and thus avoid misleading statistical inferences and policy decisions, in particular from seemingly insignificant tests. The econometric literature on inequality has long emphasized the need to avoid over-sized tests. Rightfully, spurious rejections are misleading. Our results document a different, although related, problem: even with adequately sized no-change tests, weak identification can undercut the reliability of policy advice resulting form insignificant no-change test outcomes. Far more attention needs to be paid to confidence sets. Moreover, sets that can be unbounded, although might seems counter-intuitive at first, make empirical and policy work far more credible than it can be using bounded alternatives or no-change tests that cannot be inverted.

5 Conclusion

This paper introduces a Fieller-type method for two-sample inference problem on the GE class of inequality indices. Simulation results confirm that the proposed method outperforms standard counterparts including the permutation test. Improvements are most notable for indices that put more weight on the right tail of the distribution and for sample sizes that match

macroeconomic type inequality analysis. Results are robust to different assumptions about the shape of the null distributions. While irregularities arising from the right tail have long been documented, we find that left tail irregularities are equally important in explaining the failure of standard inference methods. On recalling that permutation tests are difficult to invert, our results underscore the usefulness of the Fieller-type method for evidence-based policy. An empirical analysis of regional and international income per capita inequality reinforces this result, and casts a new light on traditional controversies in the growth literature.

Fieller's approach is frequently applied in medical research and to a lesser extent in applied economics despite its solid theoretical foundations (Srivastava 1986; Willan and O'Brien 1996; Johannesson et al. 1996; Laska et al. 1997). This could be due to the seemingly counter-intuitive non-standard confidence sets it produces which economists often find hard to interpret. Consequently, many applied researchers encountering the estimation of ratios avoid using it and opt to use methods that yield closed intervals regardless of theoretical validity. This paper illustrates serious empirical and policy flaws that may result from such practices in inequality analysis.

Author Contributions All authors contributed to the study. They all read and approved the final manuscript.

Funding This work was supported by the William Dow Chair of Political Economy (McGill University), the Bank of Canada (Research Fellowship), the Toulouse School of Economics (Pierre-de-Fermat Chair of excellence), the Universitad Carlos III de Madrid (Banco Santander de Madrid Chair of excellence), the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, the Fonds de recherche sur la société et la culture (Québec), and by the French government under the "France 2030" investment plan managed by the French National Research Agency (reference: ANR-17-EURE-0020, ANR-17-CE41-0007, ANR-19-FRAL-0006) and from Excellence Initiative of Aix-Marseille University - A*MIDEX.

Data Availibility Statement The data that support the findings of this study are available from the corresponding author upon request.

Declarations

Financial and non-financial interests The authors have no relevant financial or non-financial interests to disclose.

Ethical Approval Not applicable

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

References

Andrews, D.W., Cheng, X.: Maximum likelihood estimation and uniform inference with sporadic identification failure. J. Econom. **173**(1), 36–56 (2013)

Andrews, I., Mikusheva, A.: Maximum likelihood inference in weakly identified dynamic stochastic general equilibrium models. Quant. Econ. **6**(1), 123–152 (2015)
Andrews, I., Stock, J.H., Sun, L.: Weak instruments in instrumental variables regression: Theory and practice. Ann. Rev. Econ. **11**, 727–753 (2019)
Bahadur, R.R., Savage, L.J.: The nonexistence of certain statistical procedures in nonparametric problems. Ann. Math. Stat. **27**(4), 1115–1122 (1956)
Barro, R.J.: Convergence and modernization revisited. Technical report, National Bureau of Economic Research (2012)
Barro, R.J., Sala-i Martin, X.: Convergence. J. Polit. Econ. **100**(2), 223–251 (1992)
Beaulieu, M.-C., Dufour, J.-M., Khalaf, L.: Identification-robust estimation and testing of the zero-beta CAPM. Rev. Econ. Stud. **80**(3), 892–924 (2013)
Bennett, B.: On a multivariate version of Fieller's theorem. J. R. Stat. Soc. Ser. B Methodol. **21**(1), 59–62 (1959)
Bernard, J.-T., Chu, B., Khalaf, L., Voia, M., et al.: Non-standard confidence sets for ratios and tipping points with applications to dynamic panel data. Ann. Econ. Stat. **134**, 79–108 (2019)
Bernard, J.-T., Idoudi, N., Khalaf, L., Yélou, C.: Finite sample inference methods for dynamic energy demand models. J. Appl. Econom. **22**(7), 1211–1226 (2007)
Bernat, G.A., Jr.: Convergence in state per capita personal income, 1950–99. Surv. Curr. Bus. **81**(6), 36–48 (2001)
Bertanha, M., Moreira, M.J.: Impossible inference in econometrics: Theory and applications. J. Econ. **218**(2), 247–270 (2020)
Bolduc, D., Khalaf, L., Yélou, C.: Identification robust confidence set methods for inference on parameter ratios with application to discrete choice models. J. Econ. **157**(2), 317–327 (2010)
Brachmann, K., Stich, A., Trede, M.: Evaluating parametric income distribution models. Allg. Stat. Arch. **80**, 285–298 (1995)
Cowell, F.A.: Measurement of inequality. In Handbook of income distribution, vol. 1, pp. 87–166. Elsevier, Amsterdam (2000)
Cowell, F.A., Flachaire, E.: Income distribution and inequality measurement: The problem of extreme values. J. Econom. **141**(2), 1044–1072 (2007)
Cowell, F.A., Flachaire, E.: Statistical methods for distributional analysis. In Handbook of income distribution, vol. 2, pp. 359–465. Elsevier, Amsterdam (2015)
Cowell, F.A., Flachaire, E., et al.: Inequality measurement and the rich: why inequality increased more than we thought. (2018)
Cox, D.R.: Fieller's theorem and a generalization. Biometrika **54**(3–4), 567–572 (1967)
Davidson, R., Flachaire, E.: Asymptotic and bootstrap inference for inequality and poverty measures. J. Econom. **141**(1), 141–166 (2007)
Deaton, A.: Covid-19 and global income inequality. Technical report, National Bureau of Economic Research (2021)
Dufour, J.-M.: Some impossibility theorems in econometrics with applications to structural and dynamic models. Econom. **65**(6), 1365–1387 (1997)
Dufour, J.-M., Flachaire, E., Khalaf, L.: Permutation tests for comparing inequality measures. J. Bus. Econ. Stat. **37**(3), 457–470 (2019)
Dufour, J.-M., Flachaire, E., Khalaf, L., Zalghout, A.: Confidence sets for inequality measures: Fieller-type methods. In: Green, W., Khalaf, L., Makdissi, P., Sickles, R., Veall, M., Voia, M. (eds.) Productivity and inequality. Springer, Cham (2018)
Dufour, J.-M., Hsiao, C.: Identification. In The new palgrave dictionary of economics, The new palgrave dictionary of econometrics. Palgrave MacMillan, 2nd edition. (2008)
Dufour, J.-M., Jasiak, J.: Finite sample limited information inference methods for structural equations and models with generated regressors. Int. Econ. Rev. **42**(3), 815–844 (2001)
Dufour, J.-M., Taamouti, M.: Projection-based statistical inference in linear structural models with possibly weak instruments. Econom. **73**(4), 1351–1365 (2005)
Dufour, J.-M., Taamouti, M.: Further results on projection-based inference in iv regressions with weak, collinear or missing instruments. J. Econom. **139**(1), 133–153 (2007)
Fieller, E.C.: Some problems in interval estimation. J. R. Stat. Soc. Ser. B Methodol. **16**(2), 175–185 (1954)
Foster, J.E., Ok, E.A.: Lorenz dominance and the variance of logarithms. Econom. **67**(4), 901–907 (1999)
Franz, V.H.: Ratios: A short guide to confidence limits and proper use. arXiv:0710.2024 (2007)
Gleser, L.J., Hwang, J.T.: The nonexistence of 100 (1-α)% confidence sets of finite expected diameter in errors-in-variables and related models. Ann. Stat. **15**(4), 1351–1362 (1987)
Higgins, M.J., Levy, D., Young, A.T.: Growth and convergence across the United States: Evidence from county-level data. Rev. Econ. Stat. **88**(4), 671–681 (2006)

Johannesson, M., Jönsson, B., Karlsson, G.: Outcome measurement in economic evaluation. Health Econ. **5**(4), 279–296 (1996)
Kleibergen, F.: Testing parameters in GMM without assuming that they are identified. Econom. **73**(4), 1103–1123 (2005)
Koschat, M.A., et al.: A characterization of the Fieller solution. Ann. Stat. **15**(1), 462–468 (1987)
Laska, E.M., Meisner, M., Siegel, C.: Statistical inference for cost-effectiveness ratios. Health Econ. **6**(3), 229–242 (1997)
Maasoumi, E., Jeong, J.H.: The trend and the measurement of world inequality over extended periods of accounting. Econ. Lett. **19**(3), 295–301 (1985)
McCann, P.: Perceptions of regional inequality and the geography of discontent: Insights from the UK. Reg. Stud. **54**(2), 256–267 (2020)
Milanovic, B.: Worlds apart. Princeton University Press, In Worlds Apart (2011)
Olea, J.L.M., Stock, J.H., Watson, M.W.: Inference in structural vector autoregressions identified with an external instrument. J. Econom. **225**(1), 74–87 (2021)
Quah, D.T.: Empirics for economic growth and convergence. Eur. Econ. Rev. **40**(6), 1353–1375 (1996)
Ram, R.: International income inequality: 1970 and 1978. Econ. Lett. **4**(2), 187–190 (1979)
Rebelo, S.: Long-run policy analysis and long-run growth. J. Polit. Econ. **99**(3), 500–521 (1991)
Romer, P.M.: The origins of endogenous growth. J. Econ. Perspect. **8**(1), 3–22 (1994)
Sala-i Martin, X.X.: The classical approach to convergence analysis. Econ. J. **106**(437), 1019–1036 (1996)
Schluter, C.: On the problem of inference for inequality measures for heavy-tailed distributions. Econ. J. **15**(1), 125–153 (2012)
Shorrocks, A.F.: The class of additively decomposable inequality measures. Econometrica **48**(3), 613–625 (1980)
Srivastava, M.: Multivariate bioassay, combination of bioassays, and Fieller's theorem. Biometrics **42**(1), 131–141 (1986)
Theil, H.: Economics and information theory. North-Holland, Amsterdam (1967)
Theil, H.: World income inequality and its components. Econ. Lett. **2**(1), 99–102 (1979)
Willan, A.R., O'Brien, B.J.: Confidence intervals for cost-effectiveness ratios: An application of Fieller's theorem. Health Econ. **5**(4), 297–305 (1996)
Young, A.T., Higgins, M.J., Levy, D.: Sigma convergence versus beta convergence: Evidence from US county-level data. J. Money Credit Bank. **40**(5), 1083–1093 (2008)
Zerbe, G.O.: On Fieller's theorem and the general linear model. American Stat. **32**(3), 103–105 (1978)

Authors and Affiliations

Jean-Marie Dufour[1,2] · Emmanuel Flachaire[3] · Lynda Khalaf[4] · Abdallah Zalghout[5]

Jean-Marie Dufour
jean-marie.dufour@mcgill.ca

Emmanuel Flachaire
emmanuel.flachaire@univ-amu.fr

Abdallah Zalghout
zalghouta@macewan.ca

1 William Dow Professor of Economics, McGill University, Centre interuniversitaire de recherche en analyse des organisations (CIRANO), and Centre interuniversitaire de recherche en économie quantitative (CIREQ), Quebec, Canada

2 Department of Economics, McGill University, 855 Sherbrooke Street West, Montréal, Québec H3A 2T7, Canada

3 Aix-Marseille Université, CNRS, AMSE, 5-9 bd Maurice Bourdet, 13001 Marseille, France

4 Economics Department, Carleton University, Ottawa, ON K1S 5B6, Canada

5 Department of Anthropology, Economics and Political Science, MacEwan University, Edmonton, AB T5J 4S2, Canada