South African Statistical Journal
Vol. 55, No. 1, 15–28
https://doi.org/10.37920/sasj.2021.55.1.2

ESTIMATING THE GINI INDEX FOR HEAVY-TAILED INCOME DISTRIBUTIONS

Amina Bari
The Center University of Tindouf, Tindouf, Algeria
e-mail: *bari.amina93@gmail.com*

Abdelaziz Rassoul[1]
National Higher School of Hydraulics, Blida, Algeria
e-mail: *a.rassoul@ensh.dz*

Hamid Ould Rouis
University of Blida 1, LRDSI Laboratory, Blida, Algeria
e-mail: *houldrouis@hotmail.com*

In the present paper, we define and study one of the most popular indices which measures the inequality of capital incomes, known as the Gini index. We construct a semiparametric estimator for the Gini index in case of heavy-tailed income distributions and we establish its asymptotic distribution and derive bounds of confidence. We explore the performance of the confidence bounds in a simulation study and draw conclusions about capital incomes in some income distributions.

Key words: Extreme quantile, Gini index, Heavy-tailed incomes, Income inequality.

1. Introduction and Motivation

The last decade has seen considerable use and development of statistical theory for inferring the dominance of one distribution (of income, wealth, wages, etc.) over another. The results thus provide the statistical framework within which to assess the progressivity of taxes and benefits, and the changes in inequality of income, or in the ranking of individuals with respect to income, which they may cause. The results can also be applied to the impact of a tax and benefit system (or of other socio-economic phenomena) on poverty indices when such poverty indices depend on estimated population quantiles. They furthermore encompass as special cases most of the previous statistical inference results for the measurement of inequality and social welfare.

There are many ways of measuring inequality, all of which have some intuitive or mathematical appeal (Cowell, 1985). However, many apparently sensible measures behave in perverse fashions. Numerous indices exist for measuring the degree of inequality in the distribution of income and wealth. They range from simple measures like the share of aggregate earnings received by each quintile to more complex measures such as the Gini, Theil (1967), Atkinson and generalized entropy indices (see Atkinson, 1970). All have different mathematical constructions, which can lead to different

[1] Corresponding author.
MSC2020 subject classifications. 62G05, 62G20, 62G32, 62P20.

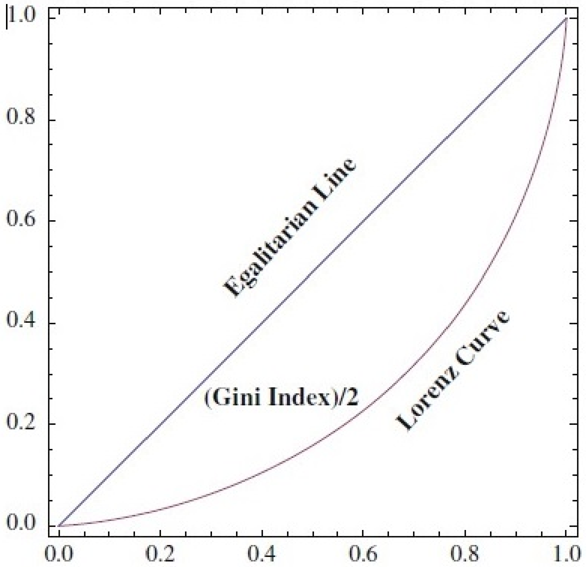

Figure 1. Egalitarian line $y = u$, Lorenz curve $y = L(u)$, and Gini index.

assessments concerning the degree of inequality. In our study, the main measure of inequality used as a proxy to show the distribution of income is the Gini coefficient.

The Gini index is the most popular and important inequality measure. This index has a long history, dating back to Gini (1914), if not earlier. In particular, the Gini index has been widely used by economists and sociologists to measure economic inequality. Measures inspired by the index have been employed to assess the equality of opportunity and estimate income mobility. Naturally, numerous modifications and extensions of the classical Gini index have been proposed during the past 100 years, depending on one's needs and/or point of view.

The Gini index is based on the area between the egalitarian line and the Lorenz curve. This quantity is multiplied by 2, in order to have a range of values in the interval $[0, 1]$. The Gini index was developed by the Italian statistician, demographer and sociologist Corrado Gini (Gini, 1914).

Note that the Lorenz curve can be considered to be a cumulative distribution function on $[0, 1]$ (Lorenz, 1905; Gastwirth, 1972; Kovacevic and Binder, 1997; Cowell, 1977; Langel and Tillé, 2013). We can exploit this fact and employ the moments of the Lorenz curve to develop new measures of inequality.

The Gini index has several possible interpretations and alternative ways in which it can be expressed. Perhaps the most popular description of this measure is one related to the area between the population Lorenz curve and the egalitarian line.

Figure 1 represents the egalitarian line $y = u$, the Lorenz curve $y = L(u)$, and the Gini index for a hypothetical distribution. Consequently, if the Gini index is $G = 0$ we have perfect equality (all incomes identical) and $G = 1$ corresponds to perfect inequality.

The existing literature has intensively studied various estimators of G and the associate inference theory. We cite here (Davidson, 2009; Qin et al., 2010; Yitzhaki, 1983; Kpanzou et al., 2013, 2017).

More specifically, let $X \geq 0$ denote the income variable with distribution function $F(x)$, and the corresponding quantile function $\mathbb{Q}\left(t\right)$ for $0 < t < 1$, with Lorenz curve L_X. A formula for its Gini index, $G(X)$ or simply G if the random variable is known from the context, is

$$G = 2\int_0^1 \left[u - L_X(u)\right] du = 1 - 2\int_0^1 L_X(u)du, \tag{1}$$

where $u = F\left(x\right)$ is a cumulative distribution function (CDF) of a non-negative income with positive expectation $\mu = E(X)$ and $L_X\left(p\right)$ is Lorenz function defined by

$$L_X\left(p\right) := \frac{1}{\mu}\int_0^p \mathbb{Q}\left(t\right) dt. \tag{2}$$

Using (1) and (2), it follows that we can also rewrite the Gini index as

$$G = 1 - \frac{2}{\mu}\int_0^1 \int_0^p \mathbb{Q}\left(t\right) dtdp.$$

Inequality measures are often underestimated using sample data. It has been noted that the sample Lorenz curve often exhibits less inequality than does the population Lorenz curve. This fact suggests that the sample curve is a positively biased estimate of the population curve. If we have a sample $X_1, X_2, ..., X_n$ of size n from a distribution $F_X(x)$, recall that the corresponding sample Lorenz curve is defined to be a linear interpolation of the points $(0,0)$ and $(j/n, \sum_{i=1}^{j} X_i / \sum_{i=1}^{n} X_i)$, $j = 1, 2, ..., n$. As usual, denote the sample Lorenz curve by $L_n(u)$.

Replacing the population quantile function $\mathbb{Q}(s)$ by its empirical counterpart $\mathbb{Q}_n(s)$, which is equal to the ith order statistic $X_{i,n}$ for all $s \in ((i-1)/n, i/n]$, and for all $i = 1, \ldots, n$, where $X_{1,n} \leq X_{2,n} \leq .... \leq X_{n,n}$ are the order statistics based on the sample $X_1, X_2, ..., X_n$. Also, the empirical estimator of the mean is $\mu_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. We arrive at the 'traditional' Gini estimator

$$\hat{G}_n = \frac{2}{n^2 \mu_n}\sum_{i=1}^{n}\left(i - \frac{1}{2}\right) X_{i,n} - 1. \tag{3}$$

Of course, the empirical Gini index $\hat{G}_n$ can be rewritten in many other ways, such as the ratio of two L-statistics or the ratio of two U-statistics, which are perhaps more familiar to the reader, but formula (3) is best suited in the context of the present discussion.

The asymptotic theory for the empirical Gini index has been known at least since Hoeffding's paper (Hoeffding, 1948) on U-statistics. Indeed, the Gini index has been one of the most popular examples for illustrating the classes of L- and U-statistics. For this reason, Beach and Davidson (1983) have developed an asymptotic theory for the traditional Gini estimator, assuming that the underlying i.i.d. random variables $X_1, X_2, ..., X_n$ have finite $(2+\epsilon)$th moments for some $\epsilon > 0$ as small as desired.

The latter moment assumption plays a crucial role. To illustrate the performance of G_n, we draw samples from the Pareto distribution

$$1 - F(x) = x^{-1/\gamma}, \quad x > 1,$$

for some $\gamma > 0$, which is called the tail index. When $\gamma > 1$, then G is not defined. When $\gamma < 0.5$, then $E[X^{2+\epsilon}] < \infty$ for some $\epsilon > 0$, and so we can use the available estimator of G.

Therefore, from now on we restrict ourselves to only those γ that are in the interval $(0.5, 1)$.

The present research has been motivated by the need for a better understanding of the distribution and inequality of capital incomes, which in many cases appear to be heavy-tailed. Since there are many individuals with no capital income, we restrict our attention to only those with positive capital incomes.

In mathematical terms, a heavy-tailed income distribution of a random variable X is regularly varying at infinity with index $(-1/\gamma) < 0$ if

$$1 - F(x) = x^{-1/\gamma}\mathbb{L}(x), \qquad \text{for every } x > 0, \tag{4}$$

for some $\gamma > 0$ and a slowly varying function $\mathbb{L} : (0, \infty) \to (0, \infty)$, i.e., $\mathbb{L}(\lambda x) / \mathbb{L}(x) \to 1$ as $x \to \infty$, for all $\lambda > 0$. The parameter γ is referred to as the tail index of F. Its estimation is of fundamental importance to the applications of extreme value theory (see for example the monographs by Hill, 1975; Beirlant and Teugels, 1989; Matthys and Beirlant, 2003; Beirlant et al., 2004; de Haan and Ferreira, 2006; and the references therein). This class includes a number of popular income distributions such as the Pareto, generalized Pareto, Burr, Fréchet and Student t, which are known to be appropriate models for fitting large incomes. In the remainder of this paper, we restrict ourselves to this class of distributions. Moreover, we focus our paper on the case $\gamma \in (1/2, 1)$ to ensure that the Gini index is finite, and in that case the results of Beach and Davidson (1983) cannot be applied, the second moment of X being infinite.

The rest of this paper is organized as follows. In Section 2, we construct an alternative estimator of the Gini index and we construct confidence bounds using this estimator. In Section 3 we illustrate the performance of the new estimator and compare it with the empirical estimator for some heavy-tailed models. The proof of the main results are postponed to Section 4.

2. Main results

The idea behind the new estimator of G is to estimate the quantile function $\mathbb{Q}$ in the definition of the Gini index by the empirical quantile function for $s < 1 - k/n$, and by an extrapolated quantile function from the heavy-tail assumption for $s \geq 1 - k/n$. We next define an alternative estimator for the mean of a heavy-tailed distribution. Indeed, recall that, the mean μ can be rewritten as

$$\mu = \int_0^1 \mathbb{Q}(s)ds = \int_0^{1-k/n} \mathbb{Q}(s)ds + \int_0^{k/n} \mathbb{Q}(1-s)ds = \mu_1 + \mu_2.$$

We formulate the mean estimator for a heavy-tailed income distribution satisfying (4) as follows:

$$\hat{\mu}_{n,k} = \int_0^{1-k/n} \mathbb{Q}_n(s)ds + \left(\frac{k}{n}\right)\frac{X_{n-k,n}}{1 - \widehat{\gamma}^H_{n,k}},$$

where $\widehat{\gamma}^H_{n,k}$ is the Hill estimator of the tail index γ (Hill, 1975):

$$\widehat{\gamma}^H_{n,k} = \frac{1}{k}\sum_{i=1}^{k} i\left(\log X_{n-i+1,n} - \log X_{n-i,n}\right).$$

Note that to estimate μ_2 we use a Weissman-type estimator for $\mathbb{Q}$, (Weissman, 1978):

$$\widehat{\mathbb{Q}}(1-s) := X_{n-k,n}\,(k/n)^{\widehat{\gamma}^H_{n,k}}\, s^{-\widehat{\gamma}^H_{n,k}}, \; s \to 0.$$

The Hill estimator has been extensively studied in the literature for an intermediate sequence k, i.e. a sequence such that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$.

Finally, we obtain a semi-parametric estimator of the Gini index for a heavy-tailed income distribution:

$$\hat{G}_{n,k} = 1 - \frac{2}{\hat{\mu}_{n,k}} \int_0^1 \int_0^t \mathbb{Q}_n(s) ds dt.$$

Asymptotic normality for $\hat{G}_{n,k}$ is obviously related to that of $\widehat{\gamma}^H_{n,k}$. As usual in the extreme value framework, to prove such a type of result, we need a second-order condition on the tail quantile function $\mathbb{U}$, defined as

$$\mathbb{U}(z) = \inf\{y : F(y) \geq 1 - 1/z\}, \quad z > 1.$$

We say that the function $\mathbb{U}$ satisfies the second-order regular variation condition with second-order parameter $\rho \leq 0$ if there exists a function $A(t)$ which does not change its sign in a neighborhood of infinity, and is such that, for every $x > 0$,

$$\lim_{t\to\infty} \frac{\log \mathbb{U}(tx) - \log \mathbb{U}(t) - \gamma \log(x)}{A(t)} = \frac{x^\rho - 1}{\rho}. \tag{5}$$

When $\rho = 0$, the ratio on the right-hand side of (5) should be interpreted as $\log(x)$. As an example of heavy-tailed income distributions satisfying the second-order condition, we have the so-called and frequently used Hall's model (see Hall, 1982; Hall and Welsh, 1985), which is a class of cdfs such that

$$\mathbb{U}(t) = ct^\gamma \left(1 + dA(t)/\rho + o\left(t^\rho\right)\right) \text{ as } t \to \infty,$$

where $\gamma > 0$, $\rho \leq 0$, $c > 0$, and $d \in \mathbb{R}^*$. For statistical inference concerning the second-order parameter ρ we refer, for example, to de Haan and Stadtmüller (1996), Peng and Qi (2004), Gomes et al. (2005), and Gomes and Pestana (2007).

First, the family includes many of the most popular distributions used in the analysis of income, see for example Arnold and Sarabia (2018), wealth and risk analysis, as special or limiting cases. This subclass of heavy-tailed distributions contains the Pareto, Burr, Fréchet, and Student t distributions. This family has several advantages for practical use.

Theorem 1. *Assume that the cdf F satisfies condition* (5) *with* $\gamma \in (1/2, 1)$. *Then for any sequence of integers $k = k_n \to \infty$ such that $k/n \to 0$ and $k^{1/2}A(n/k) \to 0$ when $n \to \infty$, on a suitable probability space, and with Brownian bridges $\mathcal{B}_n(s)$ appropriately constructed, we have that*

$$\begin{aligned}\frac{\sqrt{n}(\hat{G}_{n,k} - G)}{\sqrt{k/n}\mathbb{Q}(1-k/n)} &= -\int_0^{1-k/n} \frac{v(s)\mathcal{B}_n(s)}{\sqrt{k/n}\mathbb{Q}(1-k/n)} d\mathbb{Q}(s) \\ &\quad + \frac{\gamma^2 v(1-k/n)}{(1-\gamma)^2}\sqrt{\frac{n}{k}}\mathcal{B}_n\left(1-\frac{k}{n}\right) - \frac{\gamma v(1-k/n)}{(1-\gamma)^2}\sqrt{\frac{n}{k}}\int_{1-k/n}^1 \frac{\mathcal{B}_n(s)}{1-s} ds + o_p(1)\end{aligned} \tag{6}$$

as $n \to \infty$, where

$$v(s) = \frac{2}{\mu^2}\int_0^s \int_0^t \mathbb{Q}(s) ds.$$

The proof of Theorem 1 is complex and deferred to Section 4. From the statistical inference point of view, the following corollary is our main result.

Corollary 1. *Under the conditions of Theorem 1, we have*

$$\frac{\sqrt{n}(\hat{G}_{n,k}-G)}{\sigma(\gamma)\sqrt{k/n}\mathbb{Q}(1-k/n)}\xrightarrow{\mathcal{L}}\mathcal{N}(0,1),$$

as $n\longrightarrow\infty$, where

$$\sigma^2(\gamma)=\frac{\nu^2(1)\gamma^4}{(1-\gamma)^4(2\gamma-1)}.$$

3. Simulation study

To discuss practical implementation of Theorem 1, we first fix a significance level $\alpha\in(0,1)$ and use the classical notation $z_{\alpha/2}$ for the $(1-\alpha/2)$-level quantile of the standard normal distribution $\mathcal{N}(0,1)$. Given a realization of the random variables $X_1,\ldots,X_n$ (e.g. claim amounts), which follow a cdf F satisfying the conditions of Theorem 1, we construct a $(1-\alpha)$-level confidence interval for $\mathbb{G}$ as follows. First, we need to choose an appropriate number k of extreme values. Since Hill's estimator has in general a substantial variance for small k and a considerable bias for large k, we search for a k that balances between the two shortcomings, which is indeed a well-known hurdle when estimating the tail index.

To resolve this issue, several procedures have been suggested in the literature, and we refer to, e.g., Dekkers and de Haan (1993), Drees and Kaufmann (1998), Danielsson et al. (2001), Cheng and Peng (2001), Neves and Fraga Alves (2004), Gomes et al. (2009), and the references therein.

In our current study we employ the method of Cheng-Peng for deciding on an appropriate value k^* of k. We note that, the optimal value of k that minimizes the absolute value of the leading coverage error term of Hill estimator, this fraction k depends on the sign of second-order regular variation, for more detail, see Cheng and Peng (2001). Having computed Hill's estimator and consequently determined $X_{n-k^*:n}$, we then compute the corresponding values of $\hat{G}_{n,k}$ and $\sigma^2(\widehat{\gamma}_n)$, and denote them by $\hat{G}_{n,k^*}$ and $\sigma^{2*}(\widehat{\gamma}_n)$, respectively. Finally, using Theorem 1 we arrive at the following $(1-\alpha)$-confidence interval for G:

$$\hat{G}_{n,k^*}\pm z_{\alpha/2}\frac{(k^*/n)^{1/2}X_{n-k^*:n}\sigma^*(\widehat{\gamma}_n)}{\sqrt{n}}.$$

To illustrate the performance of this confidence interval, we carried out a small scale simulation study based on the Pareto cdf $F(x)=1-x^{-1/\gamma},x\geq1$, and the Fréchet cdf $F(x)=\exp(-x^{-1/\gamma}),x\geq0$ with the tail index γ set to $2/3$ and $3/4$, in which case we have fewer than two finite moments.

For the first part, we generated 500 independent replicates from the selected parent distribution of three samples of sizes $n=500$, 1000, and 2000. For every simulated sample, we obtained estimates $\hat{G}_{n,k}$. Then we calculated the arithmetic averages over the values from the 500 repetitions, with the absolute error and root mean squared error (rmse) of the new estimator $\hat{G}_{n,k}$ reported in Table 1 for the Pareto model, and in Table 2 for the Fréchet model.

In the tables we also report 95%-confidence intervals with their lower (lcb) and upper bounds (ucb), coverage probabilities (covpr), and lengths.

The major observations from our simulations results presented in Table 1 and Table 2 are summarized as follows: (1) The error and rmse decrease as the sample size is increased for all cases. (2)

Table 1. Simulation and confidence bounds of the estimator of the Gini index for the Pareto distribution.

	$\gamma = \frac{2}{3}, G = 0.500003$							
n	k^*	$\hat{G}_{n,k}$	**error**	**rmse**	**lcb**	**ucb**	**covpr**	**length**
500	26	0.47928	0.02074	0.01927	0.23462	0.76538	0.91782	0.53076
1000	70	0.48718	0.01288	0.01542	0.25881	0.74116	0.93526	0.48231
2000	103	0.50969	0.00969	0.01002	0.29499	0.70501	0.95236	0.41002
	$\gamma = \frac{3}{4}, G = 0.6000003$							
n	k^*	$\hat{G}_{n,k}$	**error**	**rmse**	**lcb**	**ucb**	**covpr**	**length**
500	27	0.58429	0.01572	0.01723	0.20304	0.99696	0.92671	0.79392
1000	51	0.58975	0.01025	0.01358	0.26944	0.92145	0.93056	0.66112
2000	102	0.59183	0.00817	0.000904	0.336381	0.83621	0.94748	0.47241

Table 2. Simulation and confidence bounds of the estimator of the Gini index for the Fréchet distribution.

	$\gamma = \frac{2}{3}, G = 0.58693$							
n	k^*	$\hat{G}_{n,k}$	**error**	**rmse**	**lcb**	**ucb**	**covpr**	**length**
500	26	0.5711	0.01481	0.11025	0.21351	0.97412	0.84811	0.76061
1000	52	0.59892	0.01198	0.07581	0.23069	0.96715	0.90124	0.73646
2000	103	0.58028	0.00665	0.03159	0.26062	0.89995	0.94201	0.63934
	$\gamma = \frac{3}{4}, G = 0.67979$							
n	k^*	$\hat{G}_{n,k}$	**error**	**rmse**	**lcb**	**ucb**	**covpr**	**length**
500	26	0.66812	0.01167	0.10231	0.35501	0.98124	0.86220	0.65623
1000	55	0.68541	0.00892	0.07814	0.374 79	0.996 03	0.89532	0.62124
2000	104	0.68101	0.00128	0.04215	0.378 75	0.98327	0.91202	0.60452

Table 3. Results of comparison bias and mse between $\hat{G}_n$ and $\hat{G}_{n,k}$ for Pareto model.

	$\gamma = \frac{2}{3}$				$\gamma = \frac{3}{4}$			
	$\hat{G}_n$		$\hat{G}_{n,k}$		$\hat{G}$		$\hat{G}_{n,k}$	
n	**bias**	**mse**	**bias**	**mse**	**bias**	**mse**	**bias**	**mse**
500	0.2370	0.0642	0.0556	0.0263	0.2066	0.0509	0.0670	0.00141
1000	0.1898	0.0368	0.0356	0.0123	0.1668	0.0291	0.0049	0.00077
2000	0.1257	0.0225	0.0328	0.0016	0.1393	0.0199	0.0026	0.00043

Table 4. Results of comparison bias and rmse between $\hat{G}_n$ and $\hat{G}_{n,k}$ for Fréchet model.

	$\gamma = \frac{2}{3}$				$\gamma = \frac{3}{4}$			
	$\hat{G}_n$		$\hat{G}_{n,k}$		$\hat{G}$		$\hat{G}_{n,k}$	
n	**bias**	**mse**	**bias**	**mse**	**bias**	**mse**	**bias**	**mse**
500	0.0387	0.0154	0.0197	0.00165	0.0455	0.0072	0.0111	0.00061
1000	0.0295	0.0148	0.0106	0.00138	0.0448	0.0028	0.0041	0.00014
2000	0.0168	0.0129	0.0102	0.00108	0.0331	0.0014	0.0027	0.000049

In terms of coverage probability, we find acceptable results. These results show that the coverage probability increases as the sample size is increased. (3) In terms of average length of confidence intervals, that of our interval estimators decreases when the sample size is increased.

The second part of our simulation study consists of a numerical comparison between the absolute bias and the mean square error (mse) of $\hat{G}_n$ and $\hat{G}_{n,k}$, for two models (Pareto and Fréchet) with two values of tail index ($\gamma = 2/3$ and $\gamma = 3/4$). We vary the common size n of the sample. For each size, we generated 500 independent replicates. Our overall results are taken as the empirical means of the results obtained through the 500 repetitions. To determine the optimal number of upper order statistics (which we denote by k^*) used in the computation of $\hat{\gamma}^H_{n,k}$, we apply the algorithm of Cheng and Peng (2001). The simulation results are summarised in Table 3 for the Pareto model and in Table 4 for the Fréchet model (where abs bias and mse respectively stand for the absolute value of the bias and the mean squared error of the estimation).

The results presented in Table 3 and Table 4, which represents the comparison between our proposed estimator $\hat{G}_{n,k}$ and the traditional estimator $\hat{G}_n$ in terms of bias and mse, show the performance of our estimator. The bias and mse of our estimator are smaller in all cases in comparison with the bias and mse of the traditional estimator. Furthermore, the values of the bias and mse decrease as the size of the sample is increased. In light of these results, we see that, from the point of view of the bias and the mse, the estimation accuracy increases when the size of the sample is increased.

4. Proofs

Proof of Theorem 1. Let $U_i = F(X_i)$ for $i = 1, 2, ..., n$. Then $U_1, U_2, ..., U_n$ is a sequence of i.i.d. random variables following the uniform distribution on $[0, 1]$. The following result shows that the empirical and quantile processes based on the sequence $U_1, U_2, ..., U_n$ can be approximated by a series of Brownian bridges; see Csörgő et al. (1986). These Brownian bridges are the same as on the right-hand side of equation (6) in Theorem 1. Let $\alpha_n(s)$ be the uniform empirical process defined by

$$\alpha_n(s) = \sqrt{n}\left(H_n(s) - s\right), 0 \leq s \leq 1,$$

where $H_n(s) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{U_i \leq s\}}$, and let $\beta_n(s)$ be the uniform quantile process defined by

$$\beta_n(s) = \sqrt{n}\left(H_n^{-1}(s) - s\right), 0 \leq s \leq 1.$$

Under a Skorokhod-type construction, there exists a sequence of Brownian bridges $\mathcal{B}_1, \mathcal{B}_2, ...$ such that, when $n \to \infty$, we have (cf. Csörgő et al., 1986)

$$\sup_{0 \le s \le 1-1/n} n^{\nu_1} \frac{|\alpha_n(s) - \beta_n(s)|}{(1-s)^{1/2-\nu_1}} = O_P(1) \text{ for any } 0 \le \nu_1 \le \frac{1}{4},$$

and

$$\sup_{0 \le s \le 1-1/n} n^{\nu_2} \frac{|\mathcal{B}_n(s) + \beta_n(s)|}{(1-s)^{1/2-\nu_2}} = O_P(1) \text{ for any } 0 \le \nu_1 \le \frac{1}{2}.$$

We start the proof of Theorem 1 by the calculation of the following difference

$$\begin{aligned}
\hat{G}_{n,k} - G &= \left(1 - \frac{2}{\hat{\mu}_n}\int_0^1 \int_0^t \mathbb{Q}_n(s)dsdt\right) - \left(1 - \frac{2}{\mu}\int_0^1 \int_0^t \mathbb{Q}(s)dsdt\right) \\
&= -\frac{2}{\hat{\mu}_n}\int_0^1 \int_0^t \mathbb{Q}_n(s)dsdt + \frac{2}{\mu}\int_0^1 \int_0^t \mathbb{Q}(s)dsdt \\
&= -\frac{2}{\hat{\mu}_n}\int_0^1 \int_0^t \mathbb{Q}_n(s)dsdt + \frac{2}{\hat{\mu}_n}\int_0^1 \int_0^t \mathbb{Q}(s)dsdt \\
&\quad + \frac{2}{\mu}\int_0^1 \int_0^t \mathbb{Q}(s)dsdt - \frac{2}{\hat{\mu}_n}\int_0^1 \int_0^t \mathbb{Q}(s)dsdt.
\end{aligned}$$

Then

$$\begin{aligned}
\frac{\sqrt{n}(\hat{G}_{n,k} - G)}{\sqrt{k/n}\mathbb{Q}(1-k/n)} &= -\frac{2}{\hat{\mu}_n}\left(\int_0^1 \sqrt{n}\frac{\int_0^t [\mathbb{Q}_n(s) - \mathbb{Q}(s)]\, dsdt}{\sqrt{k/n}\mathbb{Q}(1-k/n)}\right) \\
&\quad + \frac{2}{\mu\hat{\mu}_n}\int_0^1 \frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sqrt{k/n}\mathbb{Q}(1-k/n)}\int_0^t \mathbb{Q}(s)dsdt = I_1 + I_2.
\end{aligned}$$

Since I_1 is an integral over $[0,1]$, we split it into the sum of two terms, I_{11} and I_{12}, which are the same integrals but over the intervals $[0, 1-k/n]$ and $[1-k/n, 1]$, respectively. A similar split is applied on I_2, which results in $I_2 = I_{21} + I_{22}$. We shall prove that $I_{12} = o_P(1)$ and $I_{22} = o_P(1)$ when $n \to \infty$. We shall next show in several steps that $I_{11} = T_{n,1} + o_P(1)$ and $I_{21} = T_{n,2} + T_{n,3} + o_P(1)$ when $n \to \infty$. This will conclude the proof of Theorem 1. Hence, from now on we deal with the process A_n, which may be rewritten as

$$A_n(t) = \int_t^{1-k/n} [\mathbb{Q}_n(s) - \mathbb{Q}(s)]\, ds, \tag{7}$$

which is an integral of the general quantile process $\mathbb{Q}_n - \mathbb{Q}$. To reduce it to an integral of the general empirical process $F_n - F$, we employ the (general) Vervaat process (see, e.g., Zitikis, 1998)

$$V_n(t) = \int_0^t (\mathbb{Q}_n(s) - \mathbb{Q}(s))ds + \int_{-\infty}^{\mathbb{Q}(t)} (F_n(x) - F(x))dx. \tag{8}$$

The process $V_n(t)$ satisfies the boundary conditions $V_n(0) = 0$ and $V_n(1) = 0$, is non-negative for all $t \in [0,1]$, and such that

$$\sqrt{n}V_n(t) \le |e_n(t)||\mathbb{Q}_n(t) - \mathbb{Q}(t)|. \tag{9}$$

Hence, upon recalling that $e_n(t) = \sqrt{n}(F_n(\mathbb{Q}(t)) - t)$, we conclude from (9) that the difference between the quantities

$$\sqrt{n}\int_0^t (\mathbb{Q}_n(s) - \mathbb{Q}(s))ds \tag{10}$$

and

$$-\sqrt{n}\int_{-\infty}^{\mathbb{Q}(t)} (F_n(x) - F(x))dx \tag{11}$$

tends to zero when $n \longrightarrow \infty$ whenever $\mathbb{Q}_n(t)$ converges to $\mathbb{Q}(t)$, which holds because F is continuous and strictly increasing. This is the main idea of employing the Vervaat process in the present proof, as it allows us to replace quantity (10) by (11), which is much easier to tackle. We have the following equation

$$A_n(t) = -\int_{Q(t)}^{Q(1-k/n)} (F_n(x) - F(x))\,dx + V_n(1-k/n) - V_n(t)$$

which we apply on the right-hand sides of (7) and (8). By changing the variable of integration, we get

$$A_n(t) = -\int_t^{1-k/n} \frac{e_n(s)}{\sqrt{n}} d\mathbb{Q}(s) + V_n(1-k/n) - V_n(t)$$

and

$$\int_0^t (\mathbb{Q}_n(s) - \mathbb{Q}(s))ds = -\int_0^t \frac{e_n(s)}{\sqrt{n}} d\mathbb{Q}(s) + V_n(t).$$

Then

$$I_{11} = \frac{2}{\mu}\int_0^{1-k/n} \frac{\int_0^t e_n(s)d\mathbb{Q}(s)}{\sqrt{k/n}\mathbb{Q}(1-k/n)}dt - \frac{2}{\mu}\int_0^{1-k/n} \frac{\sqrt{n}V_n(t)}{\sqrt{k/n}\mathbb{Q}(1-k/n)}dt.$$

Taking into account that

$$\int_0^{1-k/n} \frac{\sqrt{n}V_n(t)}{\sqrt{k/n}\mathbb{Q}(1-k/n)}dt = o_p(1),$$

when $n \longrightarrow \infty$, we have

$$I_{11} = \frac{2}{\mu}\int_0^{1-k/n} \frac{\int_0^t e_n(s)d\mathbb{Q}(s)}{\sqrt{k/n}\mathbb{Q}(1-k/n)}dt + o_p(1). \tag{12}$$

Here we replace e_n by $\mathcal{B}_n$ in the expressions for (12). Namely, when $n \longrightarrow \infty$, by the use of the Fubini theorem, we obtain

$$I_{11} = \int_0^{1-k/n} \frac{\mathcal{B}_n(s)v(s)}{\sqrt{k/n}\mathbb{Q}(1-k/n)} d\mathbb{Q}(s) + o_p(1)T_{n,1} + o_p(1).$$

In a similar way, first writing I_{21} in terms of the empirical and Vervaat processes,

$$I_{21} = \frac{2}{\mu\hat{\mu}_n}\int_0^{1-k/n} \frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sqrt{k/n}\mathbb{Q}(1-k/n)}\int_0^t \mathbb{Q}(s)dsdt.$$

With the results of Peng (2001), Necir et al. (2010), there exist a sequence of Brownian bridge processes $\{\mathcal{B}_n(s), 0 \leq s \leq 1\}_{n\geq 1}$ such that, for any n large enough, we have

$$\frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sqrt{k/n}\mathbb{Q}(1-k/n)} \overset{d}{=} -\int_0^{1-k/n} \frac{e_n(s)}{\sqrt{k/n}\mathbb{Q}(1-k/n)} d\mathbb{Q}(s) + \frac{\gamma^2}{(1-\gamma)^2}\left\{\sqrt{\frac{n}{k}}\mathcal{B}_n\left(1-\frac{k}{n}\right)\right\} - \frac{\gamma}{(1-\gamma)^2}\sqrt{\frac{n}{k}}\int_{1-k/n}^1 \frac{\mathcal{B}_n(s)}{1-s}ds + o_P(1).$$

Then,

$$I_{21} \overset{d}{=} \int_0^{1-k/n} \frac{v(s)e_n(s)}{\sqrt{k/n}\mathbb{Q}(1-k/n)} d\mathbb{Q}(s) + \frac{\gamma^2 v(1-k/n)}{(1-\gamma)^2}\left\{\sqrt{\frac{n}{k}}\mathcal{B}_n\left(1-\frac{k}{n}\right)\right\} - \frac{\gamma v(1-k/n)}{(1-\gamma)^2}\sqrt{\frac{n}{k}}\int_{1-k/n}^1 \frac{\mathcal{B}_n(s)}{1-s}ds + o_P(1).$$

We can easily show that

$$\int_0^{1-k/n} \frac{v(s)e_n(s)}{\sqrt{k/n}\mathbb{Q}(1-k/n)} d\mathbb{Q}(s) = o_P(1).$$

Then,

$$\begin{aligned} I_{21} &\overset{d}{=} \frac{\gamma^2 v(1-k/n)}{(1-\gamma)^2}\left\{\sqrt{\frac{n}{k}}\mathcal{B}_n\left(1-\frac{k}{n}\right)\right\} - \frac{\gamma v(1-k/n)}{(1-\gamma)^2}\sqrt{\frac{n}{k}}\int_{1-k/n}^1 \frac{B_n(s)}{1-s}ds + o_P(1) \\ &= T_{n,2} + T_{n,3}. \end{aligned}$$

Finally,

$$\frac{\sqrt{n}(\hat{G}_{n,k} - G)}{\sqrt{k/n}\mathbb{Q}(1-k/n)} = \sum_{i=1}^{3} T_{n,i} + o_p(1),$$

where

$$\begin{aligned} T_{n,1} &= -\int_0^{1-k/n} \frac{\mathcal{B}_n(s)v(s)}{\sqrt{k/n}\mathbb{Q}(1-k/n)} d\mathbb{Q}(s), \\ T_{n,2} &= \frac{\gamma^2 v(1-k/n)}{(1-\gamma)^2}\sqrt{\frac{k}{n}}\mathcal{B}_n\left(1-\frac{k}{n}\right), \\ T_{n,3} &= -\frac{\gamma v(1-k/n)}{(1-\gamma)^2}\sqrt{\frac{k}{n}}\int_{1-k/n}^1 \frac{\mathcal{B}_n(s)}{1-s}ds. \end{aligned}$$ ■

Proof of corollary 1. Without the remainder term $o_P(1)$, the right-hand side of equation (6) is a mean-zero normal random variable, whose variance converges to $\sigma^2(\gamma)$ when $n \to \infty$, as the following

$$\begin{aligned} &E[T_{n,1}^2] \to \frac{2\gamma v^2(1)}{2\gamma - 1}, \quad E[T_{n,2}^2] \to \frac{\gamma^4 v^2(1)}{(1-\gamma)^4}, \quad E[T_{n,3}^2] \to \frac{\gamma^2 v^2(1)}{(1-\gamma)^4}, \\ &E[T_{n,1}T_{n,2}] \to \frac{\gamma^2 v^2(1)}{(1-\gamma)^2}, \quad E[T_{n,1}T_{n,3}] \to \frac{\gamma v^2(1)}{(1-\gamma)^2}, \quad E[T_{n,2}T_{n,3}] \to \frac{\gamma^3 v^2(1)}{(1-\gamma)^4}. \end{aligned}$$ ■

Acknowledgements. We thank the Editor-in-Chief of SASJ, an Associate Editor, and the two reviewers for their constructive comments, which significantly improved the quality of this manuscript.

References

Arnold, B. C. and Sarabia, J. M. (2018). Inequality measures. *In Majorization and the Lorenz Order with Applications in Applied Mathematics and Economics*. Springer, Cham.

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*, **2**, 244–263.

Beach, C. M. and Davidson, R. (1983). Distribution-free statistical inference with Lorenz curves and income shares. *Review of Economic Studies*, **50**, 723–735.

Beirlant, J., Goegebeur, Y., Teugels, J., and Segers, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley, Chichester.

Beirlant, J. and Teugels, J. (1989). Asymptotic normality of Hill's estimator. *In Extreme Value Theory*. Springer, New York, NY.

Cheng, S. and Peng, L. (2001). Confidence intervals for the tail index. *Bernoulli*, **7**, 751–760.

Cowell, F. A. (1977). *Measuring Inequality*. Phillip Allan, Oxford.

Cowell, F. A. (1985). Measures of distributional change: An axiomatic approach. *Review of Economic Studies*, **52**, 135–151.

Csörgő, M., Csörgő, S., Horváth, L., and Mason, D. M. (1986). Weighted empirical and quantile processes. *Annals of Probability*, **14**, 31–85.

Danielsson, J., de Haan, L., Peng, L., and de Vries, C. G. (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate Analysis*, **76**, 226–248.

Davidson, R. (2009). Reliable inference for the Gini index. *Journal of Econometrics*, **150**, 30–40.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer, New York, NY.

de Haan, L. and Stadtmüller, U. (1996). Generalized regular variation of second order. *Journal of the Australian Mathematical Society: Series A*, **61**, 381–395.

Dekkers, A. L. M. and de Haan, L. (1993). Optimal choice of sample fraction in extreme-value estimation. *Journal of Multivariate Analysis*, **47**, 173–195.

Drees, H. and Kaufmann, E. (1998). Selection of the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and Their Applications*, **75**, 149–195.

Gastwirth, J. (1972). The estimation of the Lorenz curve and Gini index. *Review of Economics and Statistics*, **54**, 306–316.

Gini, C. (1914). On the measurement of concentration and variability of characters. *Metron*, **63**, 3–38.

Giorgi, M. J. and Gigliarano, C. (2016). The Gini concentration index: A review of the inference literature. *Journal of Economic Surveys*, **31**, 1–19.

Gomes, M. I., Figueiredo, F., and Mendonça, S. (2005). Asymptotically best linear unbiased tail estimators under a second-order regular variation condition. *Journal of Statistical Planning and Inference*, **134**, 409–433.

GOMES, M. I. AND PESTANA, D. (2007). A simple second-order reduced bias' tail index estimator. *Journal of Statistical Computation and Simulation*, **77**, 487–504.

GOMES, M. I., PESTANA, D., AND CAEIRO, F. (2009). A note on the asymptotic variance at optimal levels of a bias-corrected Hill estimator. *Statistics & Probability Letters*, **79**, 295–303.

HALL, P. (1982). On some simple estimates of an exponent of regular variation. *Journal of the Royal Statistical Society: Series B*, **44**, 37–42.

HALL, P. AND WELSH, A. H. (1985). Adaptative estimates of parameters of regular variation. *Annals of Statistics*, **13**, 331–341.

HILL, B. M. (1975). A simple approach to inference about the tail of a distribution. *Annals of Statistics*, **3**, 1136–1174.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, **19**, 293–325.

KOVACEVIC, M. AND BINDER, D. (1997). Variance estimation for measures of income inequality and polarization: The estimation equations approach. *Journal of Official Statistics*, **13**, 41–58.

KPANZOU, T. A., DE WET, T., AND LO, G. S. (2017). Measuring inequality: Application of semi-parametric methods to real life data. *African Journal of Applied Statistics*, **4**, 157–164.

KPANZOU, T. A., DE WET, T., AND NEETHLING, A. (2013). Semi-parametric estimation of inequality measures. *South African Statistical Journal*, **47**, 33–48.

LANGEL, M. AND TILLÉ, Y. (2013). Variance estimation of the Gini index: Revisiting a result several times published. *Journal of the Royal Statistical Society: Series A*, **62**, 521–540.

LORENZ, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, **62**, 209–219.

MATTHYS, G. AND BEIRLANT, J. (2003). Estimating the extreme value index and high quantiles with exponential regression models. *Statistica Sinica*, **13**, 853–880.

NECIR, A., RASSOUL, A., AND ZITIKIS, R. (2010). Estimating the conditional tail expectation in the case of heavy-tailed losses. *Journal of Probability and Statistics*, **2010**, 1–17.

NEVES, C. AND FRAGA ALVES, M. I. (2004). Reissand Thomas' automatic selection of the number of extremes. *Computational Statistics & Data Analysis*, **47**, 689–704.

PENG, L. (2001). Estimating the mean of a heavy tailed distribution. *Statistics & Probability Letters*, **52**, 255–264.

PENG, L. AND QI, Y. (2004). Estimating the first and second-order parameters of a heavy-tailed distribution. *Australian & New Zealand Journal of Statistics*, **46**, 305–312.

QIN, Y., RAO, J. N. K., AND WU, C. (2010). Empirical likelihood confidence intervals for the Gini measure of income inequality. *Economic Modelling*, **27**, 1429–1435.

THEIL, H. (1967). Economics and information theory. *In Studies in Mathematical and Managerial Economics*, volume 7. North Holland Publishing Company, Amsterdam.

WEISSMAN, I. (1978). Estimation of parameters and large quantiles based on the *k* largest observations. *Journal of the American Statistical Association*, **73**, 812–815.

YITZHAKI, S. (1983). On the extension of the Gini index. *International Economic Review*, **24**, 617–628.

Zitikis, R. (1998). The Vervaat process. *In Asymptotic Methods in Probability and Statistics*. Elsevier, Amsterdam, 667–694.

Manuscript received 2020-05-15, revised 2020-12-16, accepted 2021-01-04.