Computational Engineering and Technology Innovations

www.ceti.reapress.com

Comput. Eng. Technol. Innov. Vol. 2, No. 2 (2025) 107–147.

Paper Type: Original Article

Interval Estimation for Robust Versions of Lorenz Curve Integrating Empirical Likelihood Framework With Influence Functions

W.A.R De Mel [1,*], T. L. S. M.Sathsarani [1]

[1]Department of Mathematics , University of Ruhuna, Matara, Sri Lanka; ajith@maths.ruh.ac.lk; sasini.maduranga22@gmail.com

Citation:

Received: 14 April 2024 Revised: 21 May 2024 Accepted:02 October 2024	De mel, W. A. R., & Sathsarani. (2025). Interval Estimation for Robust Versions of Lorenz Curve Integrating Empirical Likelihood Framework With Influence Functions.Computational Engineering and Technology Innovations, 2(2), 107-147.

Abstract

Real-world data often pose significant challenges for statistical inference due to issues like outliers, violation of classical assumptions, and inherent heterogeneity. Traditional parametric methods, heavily reliant on assumptions such as normality and homoscedasticity, frequently yield unreliable results when applied to skewed distributions, exemplified by income data with extremely high earners. Empirical likelihood methods have gained prominence to overcome these limitations as they reduce dependence on strict distributional assumptions. Further robustness is achieved by incorporating influence functions into empirical likelihood frameworks, explicitly quantifying estimator sensitivity to data contamination, and improving inference reliability. The Lorenz curve is a crucial economic measure of income and wealth disparity, but is sensitive to skewness and outliers. This study introduces robust quantile-based estimators for the Lorenz curve by integrating empirical likelihood and influence function methodologies. Confidence intervals are constructed through empirical likelihood optimization using a Lagrange multiplier framework and are evaluated against three bootstrap methods. Extensive simulations reflecting realistic income distributions demonstrate that the influence function-based empirical likelihood method significantly outperforms traditional empirical likelihood and other statistical approaches. Its superior robustness arises primarily from effectively mitigating the distortive effects of extreme observations. Consequently, this approach ensures reliable inference even under challenging real-world conditions characterized by data abnormalities. Nevertheless, the proposed methods present limitations, notably computational complexity and sensitivity to the assumptions underlying influence functions. Future research should expand its application across socioeconomic contexts, improve computational efficiency, and deepen theoretical and empirical exploration. This framework equips researchers and policymakers with precise tools to measure and analyze income inequality.

Keywords: Empirical likelihood, Influence function, Lagrange multiplier approach, Lorenz curve, Quantile inequality measures.

1|Introduction

Real-world data commonly present substantial challenges for statistical inference due to the presence of outliers, violation of classical assumptions, and inherent heterogeneity within samples. Traditional parametric methods

Corresponding Author: ajith@maths.ruh.ac.lk

https://doi.org/10.48314/ceti.vi.50

Licensee System Analytics. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0).

typically assume conditions such as normality, independence, and homoscedasticity, which are frequently violated in practice, particularly with skewed and complex distributions. Income data exemplify such scenarios due to significant skewness arising from a small number of extremely high earners, compromising the reliability of traditional mean-based statistical measures.

Empirical likelihood methods have emerged as robust, non-parametric approaches that overcome these limitations by minimizing reliance on strict distributional assumptions. Originally introduced by Owen [Athur(1988)], empirical likelihood methods combine the flexibility of non-parametric inference with likelihood-based frameworks, facilitating the construction of accurate confidence intervals and reliable hypothesis tests without restrictive assumptions. Influence function-based empirical likelihood further enhances robustness by explicitly quantifying estimator sensitivity to data contamination, thus ensuring greater reliability when analyzing skewed and irregular data.

The robustness provided by influence function-based empirical likelihood has proven effective across multiple disciplines, such as survival analysis in bio-statistics, social inequality assessments in sociology, environmental impact studies in environmental science, and particularly income inequality analysis in econometrics. The Lorenz curve, introduced by Max Lorenz [Lorenz(1905)], is widely adopted in economics to visualize and quantify income and wealth disparities. However, traditional Lorenz curve estimations are typically compromised due to their sensitivity to skewness and extreme values, highlighting the necessity for robust, quantile-based methods. Building upon the empirical likelihood framework, Yang and Qin [Yang(2012)] refined interval estimation techniques specifically tailored for Lorenz curve analysis, enhancing the theoretical foundation through the chi-square asymptotic properties of the empirical log-likelihood ratio statistic. Similarly, foundational contributions from Gastwirth [Gastwirth(1971)] established robust statistical properties for direct estimation of Lorenz curves, and recent advancements by Prendergast [Prendergast(2016)] further highlighted the benefits of quantile-based Lorenz curve estimators to mitigate sensitivity to extreme data points.

Motivated by these methodological developments, this study proposes robust estimators for the Lorenz curve utilizing quantile-based measures and influence functions. The robustness of the proposed estimators is enhanced by integrating influence function-based techniques, enabling more reliable inference in the presence of data irregularities. This paper introduces robust quantile-based Lorenz curve estimators, constructs confidence intervals using empirical likelihood optimization via the Lagrange multiplier framework, and evaluates the performance of these intervals against three alternative bootstrap methods. Through extensive simulation studies reflecting realistic income distributions, the proposed methodology demonstrates significant improvements in robustness and accuracy, providing researchers and policymakers with advanced tools for precise and dependable assessments of income inequality.

Empirical Likelihood. Empirical likelihood, first proposed by Owen [Athur(1988)], combines the flexibility of non-parametric methods with the inferential rigor of likelihood-based approaches. It has become a prominent statistical methodology for constructing confidence regions, hypothesis testing, and regression analyses directly from empirical data. The key advantage of empirical likelihood lies in its adaptability, effectively accommodating diverse statistical contexts, including univariate and multivariate analyses, continuous and categorical data types, and complex dependencies or interactions among variables.

Empirical likelihood constructs inference by assigning probabilities to observed data points while adhering to minimal distributional assumptions. Specifically, probabilities are optimized under constraints reflecting known moment conditions or theoretical expectations about parameters of interest. Maximizing the empirical likelihood function under these constraints provides robust parameter estimates, facilitating valid confidence intervals and hypothesis tests without relying on restrictive parametric assumptions.

Integrating empirical likelihood with Lorenz curve estimation enhances the robustness and accuracy of income inequality measurements, addressing limitations of traditional mean-based methods. By offering reliable inference under data irregularities such as skewness, outliers, and heterogeneity, empirical likelihood empowers researchers and policymakers to derive more meaningful insights into socioeconomic disparities and effectively inform related policy decisions.

Empirical Distribution. Let $x_1, x_2..., x_n$ be an independent and identically distributed simple random sample drawn from a common cumulative distribution function $F(x)$.

Then the Empirical distribution function is defined as,

$$F_n = \frac{1}{n}\sum_{i=0}^{n} I_{\{x_i \leq t\}} \tag{1}$$

where t is a given point mass and

$$I_{\{x_i \leq t\}} = \begin{cases} 1 & \{x_i \leq t\} \\ 0 & otherwise \end{cases}$$

is an indicator function.

The value of F_n at any given point t is equal to the proportion of observations in the sample that are less than or equal to that point. The empirical distribution function $F_n(t)$ provides an unbiased estimate of the cumulative distribution function $F(t)$. For a given t, the indicator $I_{\{X_i \leq t\}}$ is a Bernoulli random variable with

parameter $P = F(t)$. Therefore, $n \cdot F_n(t)$ is a Binomial random variable with expected value $n \cdot F(t)$ and variance $n \cdot F(t) \cdot (1 - F(t))$. Thus, empirical distribution is the discrete variable's distribution function.

The expected value and variance of the empirical distribution function can be calculated as follows, and it is evident that $F_n(t)$ is the unbiased estimator of $F(t)$.

$$\mathbb{E}(F_n(t)) = \frac{1}{n}\sum_{i=1}^{n} P\{X_i \leq t\} = F(t)$$

$$\mathrm{Var}(F_n(t)) = \frac{1}{n^2}\sum_{i=1}^{n} P\{X_i \leq t\} \cdot (1 - P\{X_i \leq t\}) = \frac{1}{n} F(t)\,(1 - F(t))$$

To create empirical quantiles, sort the n number of data in ascending order to obtain a sequence of order statistics, $x(1) \leq x(2) \leq ... \leq x(n)$. Then the p^{th} quantile (x_p) can be obtained as follows,

$$x_p = \begin{cases} x_{(n+1)\,p}; & \text{if } (n+1)p \text{ is integer} \\ \frac{x_{\lfloor (n+1)\,p \rfloor} + x_{\lfloor (n+1)\,p \rfloor + 1}}{2}; & \text{otherwise} \end{cases} \tag{2}$$

where, $x_{(n+1)\,p}$ is the $(n+1)\,p^{th}$ order statistic and $\lfloor (n+1)\,p \rfloor$ is the greatest integer less than or equal to $(n+1)\,p$.

Empirical Likelihood Function. Let $X_1, X_2, ..., X_n$ are independent and identically distributed random variables with unknown cumulative distribution function $F(x)$. Then the empirical likelihood function is given by,

$$L(F) = \prod_{i=1}^{n} p_i = \prod_{i=1}^{n} (F(x_i) - F(x_{i-1})) \tag{3}$$

where, $\{p_i\}$ is the probability of the set $\{x_i\}$ under distribution F also referred as weights. x_i is the observed value of X_i.

Hence empirical log-likelihood function is defined as,

$$L^*(F) = \ln\left(L(F)\right) = \ln\left(\prod_{i=1}^{n} p_i\right) = \sum_{i=1}^{n} \ln(p_i) \tag{4}$$

$L(F)$ is maximized by the empirical distribution function $F_n(x)$ which put equal weight $1/n$ on each observations (i.e.:- $p_i = 1/n$).

The maximum empirical likelihood estimator (MELE) is given by,

$$MELE = max\, L(F) = \prod_{i=1}^{n} max\, p_i = \prod_{i=1}^{n} \left(\frac{1}{n}\right) = \left(\frac{1}{n}\right)^n \tag{5}$$

Suppose $\theta = T(F)$ be a parameter which is any functional of distribution function F. The empirical likelihood ratio is the cornerstone for inferring about $\theta = T(F)$ such as interval estimation or hypothesis testing based on the empirical likelihood framework.

Let $\{x_1, x_2, \ldots, x_n\}$ be a sample from an unknown distribution F. The empirical likelihood is based on assigning probabilities $\{p_1, p_2, \ldots, p_n\}$ to each data point, subject to:

$$p_i \geq 0 \quad \text{for all } i, \quad \sum_{i=1}^{n} p_i = 1 \quad \text{(probability constraint)}.$$

For a parameter θ and a function $f(x_i, \theta)$, we aim to satisfy the moment constraint:

$$\sum_{i=1}^{n} p_i f(x_i, \theta) \ = 0 \text{ such that } \ \mathbb{E}\Big[f(x_i, \theta)\Big] = 0.$$

Then empirical likelihood ratio is given by,

$$R(\theta) = \frac{L(\tilde{p}_1, \ldots, \tilde{p}_n)}{MELE} = \frac{L(\tilde{p}_1, \ldots, \tilde{p}_n)}{\prod_{i=1}^{n} \left(\frac{1}{n}\right)}. \tag{6}$$

where,

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0.$$

$\tilde{p}_i$ are called the constrained maximum empirical likelihood estimators for p_i under H_0, and they are the solution to the optimization problem of the empirical log-likelihood function;

$$\max_{\{p_i\}} \sum_{i=0}^{n} \ln p_i.$$

corresponding to the above-mentioned constraints.

The term $-2 \cdot \ln\left(R(\theta)\right)$ follows a chi-square distribution asymptotically with one degree of freedom. The $100(1-\alpha)\%$ level of the empirical likelihood confidence interval for $\theta = T(F)$ is constructed by solving the equation;

$$\theta : -2\ln(R(\theta)) \leq \chi^2_{(1,1-\alpha)}$$

where $\chi^2_{(1,1-\alpha)}$ is the critical value of the chi-square distribution with one degree of freedom. Moreover, the testing hypothesis; $H_0 : \theta = \theta_0$ about parameter $\theta = T(F)$ could be done by checking if:

$$-2\ln(R(\theta_0)) \leq \chi^2_{(1,1-\alpha)}$$

If the inequality holds, do not reject H_0 ; otherwise, reject it.

Income Distribution and Data Explanation. Generally, income data refers to earnings or financial resources that have been made by individuals, households, businesses, or other entities from various sources such as wages, salaries, business profits, investments, rents, and government transfers over a specific time period; typically measured annually for statistical or analytical purposes. Income data is critical for understanding societal economic dynamics and guiding decision-making processes aimed at increasing social welfare and economic success. Furthermore, income data frequently demonstrates positive skewness, which means that the distribution is skewed towards higher income values. The following figures (1) and (2) show that the income data is distributed with positive skewness and also has outliers.

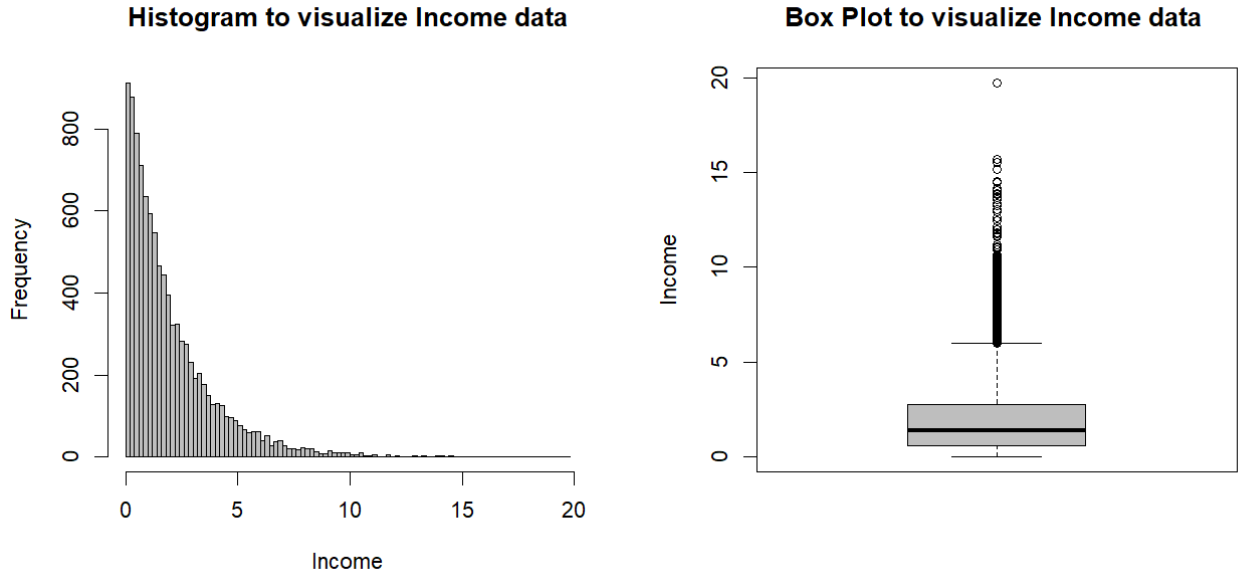


FIGURE 1. Visualizing behavior of income data using histogram and boxplot

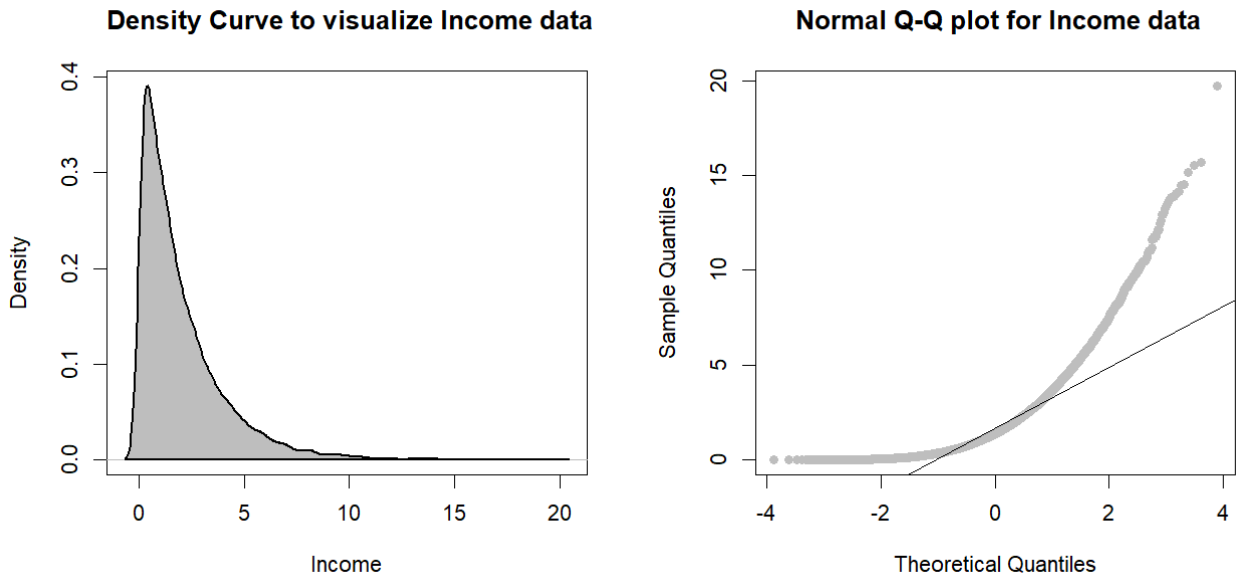


FIGURE 2. Visualizing behavior of income data using density curve and Q-Q plot

Income distribution refers to how the total income earned within a society or country is divided among its members or specific groups of the population. Simply, income distribution refers to the dispersion of income among individuals or households, providing the level of economic inequality, social fairness, and access to resources in a society. Understanding income distribution is crucial for policymakers seeking to design effective strategies for promoting equitable growth, comprehending consumer behavior and purchasing power, reducing poverty, forecasting economic trends and market conditions, and fostering social mobility. By examining the distributional effects of various economic policies and interventions, the economist can provide valuable insights into their potential impacts on different income groups and overall societal welfare. The income distribution is influenced by various factors such as wage levels, employment opportunities, access to education, government policies, globalization and trade, demographics, and technological changes. It can be improved by using several measures such as progressive taxation, providing support such as unemployment benefits, healthcare, and education to low-income groups, ensuring a fair baseline income for workers, and improving access to education and training to bridge skill gaps and provide better job opportunities.

Statistical Distribution for Income Data.

Pareto-II Distribution. The Pareto-II distribution, also known as the Lomax distribution [Lomax(1954)], adds a scale parameter to the classical Pareto distribution, making it more versatile and starting at$x = 0$. Its heavy tail effectively models extreme values, making it ideal for analyzing income data and identifying income inequality trends.

The probability density function of the log-normal distribution is given by,

$$f(y;\mu,\sigma) = \begin{cases} \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(y)-\mu)^2}{2\sigma^2}\right), & y > 0, \end{cases} \tag{11}$$

The probability density function of the Pareto-II distribution is given by,

$$f(x;\alpha,\sigma) = \begin{cases} \frac{\alpha}{\sigma}\left(1+\frac{x}{\sigma}\right)^{-(\alpha+1)} & ; \quad x \geq 0, \\ 0 \quad ; & x < 0 \end{cases} \tag{7}$$

and its cumulative distribution function is given by,

$$F(x;\alpha,\sigma) = \begin{cases} 1-\left(1+\frac{x}{\sigma}\right)^{-\alpha} & ; \quad x \geq 0, \\ 0 \quad ; & x < 0 \end{cases} \tag{8}$$

where $\alpha > 0$ is the shape parameter that controls the tail behavior and $\sigma > 0$ is the scale parameter that determines the location of the distribution.

Weibull Distribution. The Weibull distribution is a flexible probability distribution used for reliability, manufacturing, survival, and modeling extreme events [Lai(2006)]. Its shape and scale parameters influence its shape and location, making it suitable for income data with right-skewed characteristics.

Probability Density Function :

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^{k}} & ; \quad x \geq 0, \\ 0 \quad ; & x < 0 \end{cases} \tag{9}$$

Cumulative Distribution Function :

$$F(x;\lambda,k) = \begin{cases} 1-e^{-\left(\frac{x}{\lambda}\right)^{k}} & x \geq 0, \\ 0 & x < 0 \end{cases} \tag{10}$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter.

Log-Normal Distribution. Log-normally distributed random variables follow a normal distribution for positive values, used in modeling future price stocks, component lifetimes, pollutant concentration, and rainfall amounts due to the multiplicative nature [Torrent(1978)].

A random variable Y is log-normally distributed if:

$$\ln(Y) \sim N(\mu, \sigma^2)$$

where, $\ln(Y)$ is the natural logarithm of Y, μ is the mean, and $\sigma > 0$ is the standard deviation of the logarithm of Y.

The probability density function of the log-normal distribution is given by,

$$f(y;\mu,\sigma) = \begin{cases} \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(y)-\mu)^2}{2\sigma^2}\right), & y > 0, \\ 0, & y \leq 0 \end{cases} \tag{11}$$

The cumulative distribution function is expressed in terms of the standard normal cumulative distribution function Φ:

$$F(y;\mu,\sigma) = \Phi\left(\frac{\ln(y)-\mu}{\sigma}\right), y > 0 \tag{12}$$

where

$$\Phi(Y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{t^2}{2}}\, dt$$

Dagum Distribution. The Dagum distribution was introduced by economist Camilo Dagum in the 1970s. It is a continuous probability distribution used to model positive data, which provides intuitive insights into inequality measures and can handle both the lower and upper parts of income distributions [kleiber(2008)]. The Dagum distribution is characterized by three positive parameters: $a > 0$, one of the shape parameters that controls the inequality in the lower tail, $p > 0$, the second shape parameter that governs the disparity in the upper tail, and $b > 0$, the scale parameter.

Probability Density Function :

$$f(x; a, b, p) = \frac{ap}{x} \cdot \left(\frac{\left(\frac{x}{b}\right)^a}{1 + \left(\frac{x}{b}\right)^a} \right)^{p+1}, \quad x > 0 \tag{13}$$

Cumulative Distribution Function :

$$F(x; a, b, p) = \left(1 + \left(\frac{b}{x}\right)^a\right)^{-p}, \quad x > 0 \tag{14}$$

The Lorenz Curve. The Lorenz curve is a key cornerstone in economics for illustrating the disparities of income or wealth within a population. Named after its inventor, the American economist Max O. Lorenz in 1905, it provides a visual representation that graphically depicts the degree of economic disparity in an entire nation [Kleiber(2008)]. The curve is created by arranging individuals or households in an ascending order based on their income or wealth. The cumulative proportion of the population is represented on the horizontal axis, while the cumulative percentage of total income or wealth received by this population is shown on the vertical axis. The resulting curve illustrates how the distribution of income or wealth deviates from perfect equality, with the diagonal line representing perfect equality and any deviation from this line indicating varying levels of inequality. There are two points that the Lorenz curve has to go through. The first point is the origin (0,0), since 0% of the population earns 0% of the total income. At the second point (1,1), 100% of the population must have 100% of the income. If everyone had the same income, then there would be perfect equality; hence, a line with an angle of 45° (diagonal line) represents a perfect (equal) income distribution while the curve reflects the real distribution of income. That means the Lorenz curve must lie beneath the equality line.

According to Figure 3, it is shown that the equality line reads that the proportion of below 90% of households would gain 90% of the income while the richest 10% of households would gain 10% of the total income. The Lorenz curve reads that the proportion of below 90% of households hold 55% of the total income while the richest 10% of households gain 45% of the total income. The Lorenz curve can move closer or further away

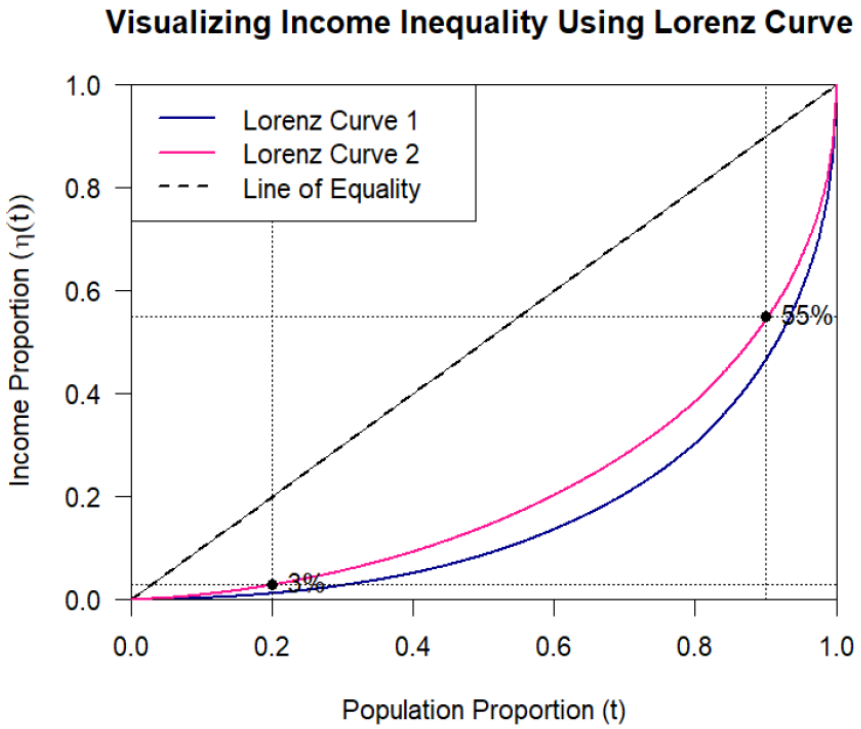


FIGURE 3. Graphical Representation of the Lorenz Curve

from the equality line. It has moved closer to the equality line, indicating that inequality in that economy has reduced, as well as further away from the equality line, indicating that inequality in that economy has grown.

Mathematical Representation of the Lorenz Curve. Let $\{x_1, x_2, ..., x_n\}$ be a simple random sample drawn from the population of X with cumulative distribution function $F(x)$; where X is a non-negative random variable. Assuming that $F(x)$ is differentiable, in general, the mathematical representation of the Lorenz ordinate at the lowest t^{th} fraction of income is defined by,

$$\eta(t) = \frac{1}{\mu} \int_0^{\epsilon_t} x \, dF(x), \quad t \in [0, 1] \tag{15}$$

where

$\mu = \int_0^{\infty} x \, dF(x)$: The mean of F , which denotes the average standard of living in the total population.

$\epsilon_t = F^{-1}(t)$: The t^{th} quantile of F. $(i.e. : F(\epsilon_t) = t)$

t : The given cumulative proportion of the population

For a fixed $t \in [0, 1]$, the Lorenz ordinate $\eta(t)$ is the fraction of the total income owned by holders of the lowest t^{th} fraction of income.

$$x = F^{-1}(p) \Longrightarrow F(x) = p \Longrightarrow dF(x) = dp \Longrightarrow f(x)dx = dp$$

when $x = 0$,

$$0 = F^{-1}(p) \Longrightarrow F(0) = p \Longrightarrow \text{as } F(0) = 0 \Longrightarrow p = 0$$

when $x = \infty$,

$$\infty = F^{-1}(p) \Longrightarrow F(\infty) = p \Longrightarrow \text{as } \quad F(\infty) = 1 \Longrightarrow p = 1$$

Therefore another expression for $\mu = \int_0^\infty x\, dF(x)$ can be obtained as,

$$\mu = \int_0^1 F^{-1}(p)\, dp = \int_0^1 q(p)\, dp$$

where, $x = F^{-1}(p) = q(p)$ is the living standard level (quantile function) below that can find a proportion p of the population.

when $x = 0 \Longrightarrow F(0) = p \Longrightarrow p = 0 \quad \text{as } F(0) = 0$

when $x = \epsilon_t \Longrightarrow F(\epsilon_t) = p \Longrightarrow p = t \quad \text{as } F(\epsilon_t) = t$

then,

$$\int_0^{\epsilon_t} x\, dF(x) = \int_0^{\epsilon_t} x\, f(x)dx = \int_0^t F^{-1}(p)\, dp = \int_0^t q(p)\, dp$$

From the general definition of Lorenz ordinate,

$$\eta(t) = \frac{1}{\mu}\int_0^{\epsilon_t} x\, dF(x) = \frac{\int_0^t q(p)\, dp}{\int_0^1 q(p)\, dp} = \frac{t \cdot \mu_t}{\mu}, \quad t \in [0,1]$$

here the numerator sums the earnings of the bottom t proportion of the population, whereas the denominator sums all incomes throughout the entire population. Thus, $\eta(t)$ represents the cumulative percentage of total income held by a cumulative proportion t of the population when individuals are ordered in increasing income stages.

The Lorenz curve needs to be estimated because the income distribution $F(x)$ is rarely known in practice. If $x_1, x_2, ..., x_n$ are independent copies of X, the empirical estimation of the Lorenz curve can be done by,

$$\hat{\eta(t)} = \frac{1}{\hat{\mu}}\int_0^{\hat{\epsilon_t}} x\, dF_n(x) = \frac{t\,\hat{\mu_t}}{\hat{\mu}} \tag{16}$$

where,

$\hat{\mu}$:- Sample mean

$\hat{\mu_t}$:- Sample mean of incomes of bottom t proportion

$\hat{\epsilon_t}$:- t^{th} sample quantile

$F_n(X)$:- Empirical distribution {unbiased estimator for $F(x)$}

The Lorenz curve has the following well-known properties.

- Lorenz curve is not defined when $\mu = 0$ or $\mu = \infty$.
- It is a continuous and convex function.
- $\frac{d\,\eta(t)}{dt} = \frac{q(t)}{\mu} = \frac{\epsilon_t}{\mu}$; $\epsilon_t = F^{-1}(t)$ is the tangent drawn to the Lorenz curve at ϵ_t. For instance, $\frac{q(0.5)}{\mu}$ is given the slope drawn to the curve at the median point of the population.

- The mean income of the population is found at that percentile at which the slope of the Lorenz curve equals one.

$$\left(\text{ i.e:- } \frac{q(t)}{\mu} = 1 \Longrightarrow q(t) = \mu \right)$$

Robust Versions of the Lorenz Curve. Robust statistics are characterized by their ability to produce accurate results in situations of data abnormalities, such as outliers or noise, or model assumption violations. They often rely on resistant measures, which are summary statistics that are less impacted by extreme values or outliers in the data. For example, resistant metrics include the median (instead of the mean) for the central tendency and the interquartile range (instead of the standard deviation) for dispersion. Robust methods are designed to reduce the influence of deviations from ideal conditions, such as normality, ensuring reliable results in real-world scenarios. The concept of robust statistics was organized in the 18^{th} century with the development of procedures for rejecting outliers. Later on, estimators to reduce outliers were created. A few years later, Huber (1964) and Hampel (1974) [Hampel(1971)] developed a comprehensive theory of robust statistics by introducing generalized estimators (Huber's M-estimators) that are robust to outliers. Hence, in recent years, many researchers have attempted to robustify classical statistical methods and the development of a comprehensive theory of robustness. The data related to income or wealth is also normal for outliers or inconsistent observations. For instance, the cumulative share of income held by a small portion of the population might be significantly impacted by a single outlier representing a very high earner. Thus, it can significantly influence inequality measures such as the Lorenz curve.

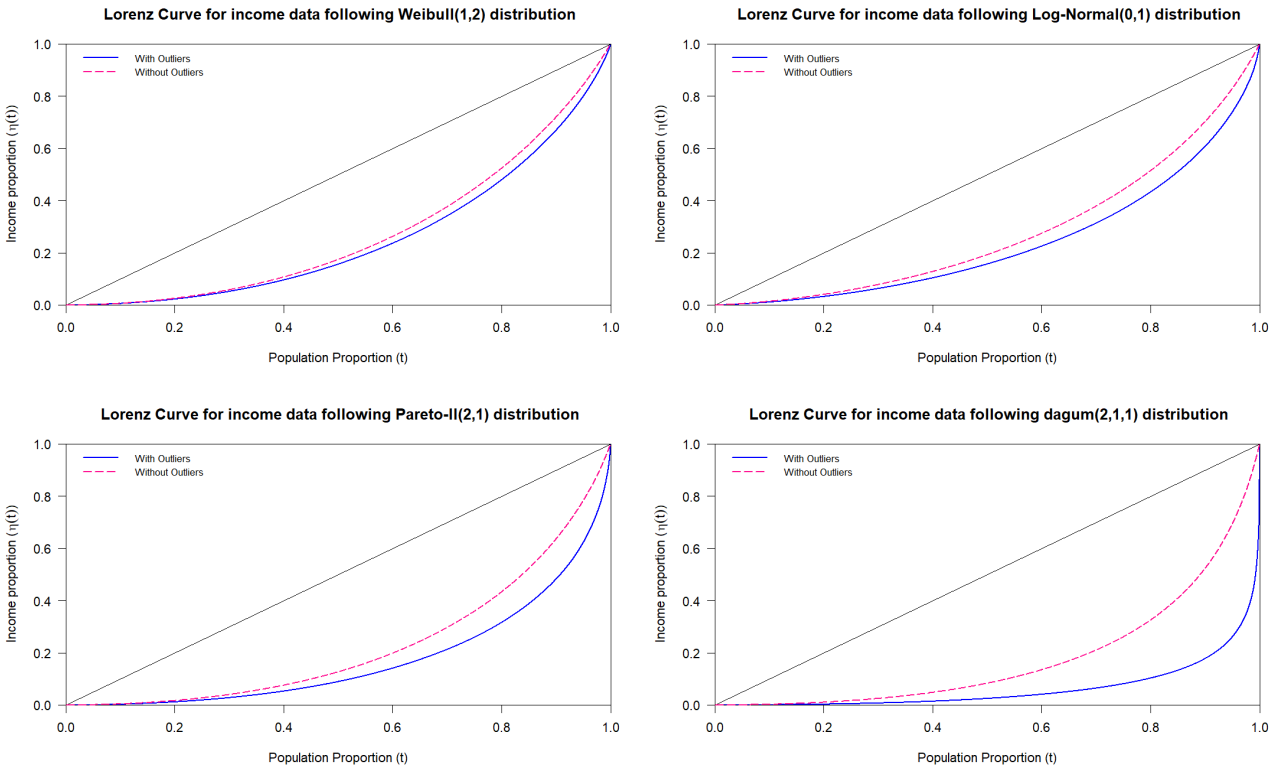


FIGURE 4. Impact of the outliers for the Lorenz curve

Figure (4) shows the effect of outliers in income data on the Lorenz curve by highlighting that without outliers, the gap between the equality line and the curve becomes less, reducing the inequality. The study in [Joshep(2014)] showed that substituting the median for the mean in the inequality coefficient's denominator will increase its sensitivity to changes in income inequality, particularly in detecting shifts affecting the majority of the population. Furthermore, the impact of means versus medians on poverty indicators is explored in [Thomas(2010)] and with

The three robust quantile inequality curves are obtained by replacing terms associated with the definition of the general Lorenz ordinate,

$$\eta(t) = \frac{1}{\mu}\int_0^{\epsilon_t} x\, dF(x) = \frac{\int_0^t q(p)\, dp}{\int_0^1 q(p)\, dp} = \frac{t\cdot\mu_t}{\mu}, \quad t \in [0,1]$$

Initially, the average proportion of t of the total population,

$$\int_0^{\epsilon_t} x\, dF(x) = \int_0^t F^{-1}(p)\, dp = \int_0^t q(p)\, dp = \mu_t$$

is replaced by its median value $\epsilon_{t/2}$. It implies that the proportion $t/2$ of the population has a lower income than $\epsilon_{t/2} = q(t/2)$. Furthermore, the term μ is replaced by three different quantiles;

- Population median. (i.e. :- $\epsilon_{0.5} = q(0.5)$)
- $(1-\frac{t}{2})^{th}$ quantile. (i.e. :- $\epsilon_{1-\frac{t}{2}} = q(1-\frac{t}{2})$)
- Average of $(\frac{t}{2})^{th}$ quantile and $(1-\frac{t}{2})^{th}$ quantile. (i.e. :- $\frac{\epsilon_{\frac{t}{2}}+\epsilon_{1-\frac{t}{2}}}{2}$)

Then, three robust versions of the Lorenz curve are denoted by [Prendergast(2016)],

$$\eta_1(t) = \frac{t\cdot\epsilon_{t/2}}{\epsilon_{0.5}} \tag{17}$$

$$\eta_2(t) = \frac{t\cdot\epsilon_{t/2}}{\epsilon_{1-t/2}} \tag{18}$$

$$\eta_3(t) = \frac{t\cdot\epsilon_{t/2}}{\frac{\epsilon_{t/2}+\epsilon_{1-t/2}}{2}} = \frac{2}{\frac{1}{t}+\frac{1}{\eta_2(t)}} \tag{19}$$

where $t \in [0,1]$ is a considered proportion of population. The multiplication of each quantile inequality version by the corresponding proportion of the population serves a specific purpose, ensuring that it appropriately reflects the cumulative distribution of wealth or income.

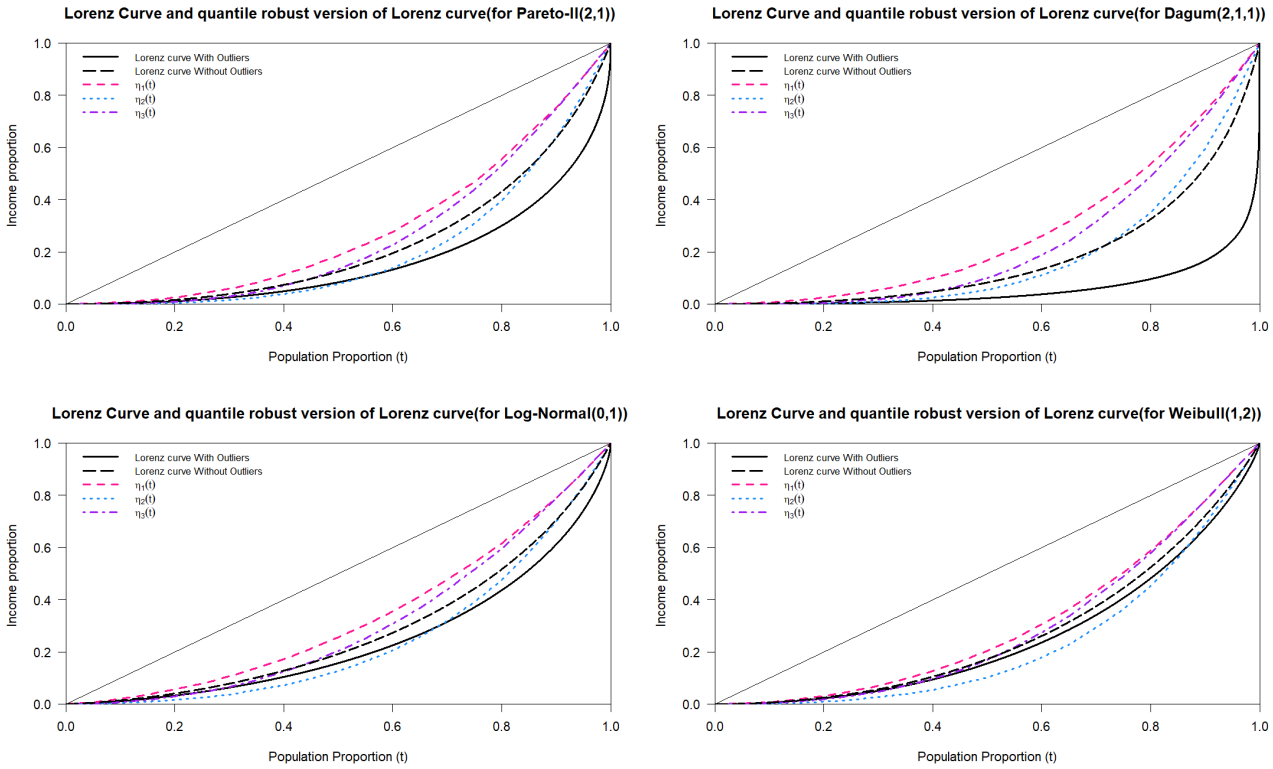


FIGURE 5. Visualization of three robust quantile versions for income data simulated from different statistical distributions.

Influence Function. The influence function is a key concept in robust statistics. It assesses the sensitivity of a statistical estimator to small changes in the data [shi(2020)]. By understanding how an estimator responds to data contamination or outliers, the influence function helps assess the robustness and reliability of statistical methods. The Influence function quantifies how the estimator T changes when a small amount of contamination is introduced at a specific point x.

Let, $\mathcal{F}' = \{F(x)|F'(x) = f(x)$ exists and it is strictly positive$\}$

As previously mentioned t^{th} quantile is defined as, $\epsilon_t = F^{-1}(t) = q(t)$ and then the quantile density is defined by,

$$Q(t) = \frac{\partial q(t)}{\partial t} = \frac{\partial \epsilon_t}{\partial t} = \frac{1}{F'(\epsilon_t)} = \frac{1}{f(\epsilon_t)} \tag{20}$$

The influence function of the t^{th} quantile function $\epsilon_t = F^{-1}(t)$ with contamination point z is given by,

$$IF(\, z;\, \epsilon_t,\, F\,) = \begin{cases} (t-1)\, Q(t) & \text{if } z < \epsilon_t \\ 0 & \text{if } z = \epsilon_t \\ t\, Q(t) & \text{if } z > \epsilon_t \end{cases} \tag{21}$$

For any given value x such that $F(x) = p$;

$$IF(\, z;\, x,\, F\,) = (p-1)\, Q(p)\, I(z < x) + 0\, I(z = x) + p\, Q(p)\, I(z > x) \tag{22}$$

$$= Q(p)\, \{\, (p-1)\, I(z < x) + p\, I(z > x)\, \} \tag{23}$$

$$\begin{aligned}\mathbb{E}[IF(\,z;\,x,\,F\,)] = Q(p)\,(\,(p-1)\,\mathbb{E}[I(z<x)] + p\,\mathbb{E}[I(z>x)]\,) \\ &= Q(p)\,\{\,(p-1)\,Pr(z<x) + p\,Pr(z>x)\,\} \\ &= Q(p)\,\{\,(p-1)\,F(x) + p\,(1-F(x))\,\} \\ &= Q(p)\,\{\,p-p\,\} \;=\; 0\end{aligned}$$

The expected value of the influence function of the quantile proved the well-known property that the expectation of the influence function of any functional is zero.

$$\begin{aligned}\mathrm{Var}[IF(\,z;\,x,\,F\,)] &= \mathbb{E}[(IF(\,z;\,x,\,F\,) - \mathbb{E}[IF(\,z;\,x,\,F\,)])^2] \\ &= \mathbb{E}\Big[(IF(\,z;\,x,\,F)^2\Big] \\ &= Q^2(p)\,\mathbb{E}\Big[(p-1)^2 I(z<x) + p^2 I(z>x)\Big] \\ &= Q^2(p)\Big[(p-1)^2\,F(x) + p^2\,(1-F(x))\Big] \\ &= Q^2(p)\Big[(p-1)^2\,p + p^2\,(1-p)\Big] \;=\; Q^2(p)\Big[p\,(1-p)\Big]\end{aligned}$$

However, the quantile inequality curves have finite influence functions, provided only that $F \in \mathcal{F}$ by [Prendergast(2016)], where $\mathcal{F}$ is the class of all cumulative distribution functions F with $F(0) = 0$. The paper [Cowell(2002)] shows that the influence function of the Lorenz curve at the point p is unbounded, implying that even minor contamination can cause a significant bias in the estimate.

Methodology

Influence Function For Robust Versions of Lorenz Curve. Since the robust versions of the Lorenz curve $\eta_j(t)$ for a given proportion t of the population are defined in terms of two different quantile measures, the derivation of their corresponding influence function can be obtained in terms of the influence function of each quantile as follows.

$$IF(z;\eta_j(t),F) = t\left\{\frac{IF(z;\epsilon_{t/2},F)}{d_j(t)} - \frac{\epsilon_{t/2}\cdot IF(z;d_j(t),F)}{d_j^2(t)}\right\} \tag{24}$$

where,

$$\eta_j(t) = \frac{t\cdot\epsilon_{t/2}}{d_j(t)} \quad \text{for}\, j = 1,2,3$$

$$d_1(t) = \epsilon_{0.5}, \quad d_2(t) = \epsilon_{1-t/2}, \quad d_3(t) = \frac{\epsilon_{t/2}+\epsilon_{1-t/2}}{2}$$

$$IF(\,z;\,\epsilon_{t/2},\,F\,) = Q(\frac{t}{2})\left[\frac{t}{2} - I(z<\epsilon_{t/2})\right]$$

The above stated result for $IF(\,z;\,\epsilon_{t/2},\,F\,)$ is obtained by further simplification of the expression in equation(22) such that;

$$\begin{aligned}IF(\,z;\,x,\,F\,) &= Q(p)\left[\,(p-1)\,I(z<x) + p\,I(z>x)\,\right] \\ &= Q(p)\left[\,(p\,I(z<x) - I(z<x)) \;+\; p \;-\; p\,I(z=x) - p\,I(z<x)\}\,\right] \\ &= Q(p)\left[\,p \;-\; I(z<x)\,\right]\end{aligned}$$

Hence,

$$IF(\,z;\,d_1(t),\,F\,) = IF(\,z;\,\epsilon_{0.5},\,F\,) = Q(0.5)\left[0.5 - I(z < \epsilon_{0.5})\right],$$

$$\begin{aligned} IF(\,z;\,d_2(t),\,F\,) =& IF(\,z;\,\epsilon_{1-\frac{t}{2}},\,F\,) \\ =& Q\left(1-\frac{t}{2}\right)\,\left[\,\left(1-\frac{t}{2}\right) - I\left(z < \epsilon_{1-t/2}\right)\,\right] \end{aligned}$$

and $IF(\,z;\,d_3(t),\,F\,)$ is utilized by

$$IF(z; d_3(t), F) = \frac{IF(z; \epsilon_{t/2}, F) + IF(z; \epsilon_{1-t/2}, F)}{2}.$$

For instance,

$$\begin{aligned} IF(z; \eta_1(t), F) =& t\left\{\frac{IF(z; \epsilon_{t/2}, F)}{\epsilon_{0.5}} - \frac{\epsilon_{t/2}\cdot IF(z; \epsilon_{0.5}, F)}{\epsilon_{0.5}^2}\right\} \\ =& t\left\{\frac{Q(\frac{t}{2})\left[\frac{t}{2} - I(z < \epsilon_{t/2})\right]}{\epsilon_{0.5}} - \frac{\epsilon_{t/2}\cdot Q(0.5)\left[0.5 - I(z < \epsilon_{0.5})\right]}{\epsilon_{0.5}^2}\right\} \end{aligned}$$

and then,

$$\begin{aligned} \mathbb{E}[IF(z; \eta_1(t), F)] =& t\left\{\frac{Q(\frac{t}{2})\left[\frac{t}{2} - \mathbb{E}[I(z < \epsilon_{t/2}])\right]}{\epsilon_{0.5}} - \frac{\epsilon_{t/2}\cdot Q(0.5)\left[0.5 - \mathbb{E}[I(z < \epsilon_{0.5})]\right]}{\epsilon_{0.5}^2}\right\} \\ =& t\left\{\frac{Q(\frac{t}{2})\left[\frac{t}{2} - F(\epsilon_{t/2})\right]}{\epsilon_{0.5}} - \frac{\epsilon_{t/2}\cdot Q(0.5)\left[0.5 - F(\epsilon_{0.5})\right]}{\epsilon_{0.5}^2}\right\} \\ =& t\left\{\frac{Q(\frac{t}{2})\left[\frac{t}{2} - \frac{t}{2})\right]}{\epsilon_{0.5}} - \frac{\epsilon_{t/2}\cdot Q(0.5)\left[0.5 - 0.5\right]}{\epsilon_{0.5}^2}\right\} = 0 \end{aligned}$$

Using similar techniques and processes, it could be proved that

$$\mathbb{E}[IF(z; \eta_2(t), F)] = \mathbb{E}[IF(z; \eta_3(t), F)] = 0$$

So, these results proved that the well-known property of the influence function, which is that the expected value is zero, is also well-suited for the influence function of robust versions of the Lorenz curve.

Quantile Density Function. The quantile density $Q(t)$ is defined as:

$$Q(t) = \frac{\partial q(t)}{\partial t} = \frac{\partial \epsilon_t}{\partial t} = \frac{1}{F'(\epsilon_t)} = \frac{1}{f(\epsilon_t)}$$

Example : For Log-Normal distribution. The cumulative distribution function of Lognormal distribution $F(y)$ is defined as,

$$F(y) = \Phi\left(\frac{\ln(y) - \mu}{\sigma}\right)$$

Then the quantile y (inverse of $F(y)$) for the given proportion t is given by,

$$y = \exp\left(\mu + \sigma\Phi^{-1}(t)\right)$$

where Φ^{-1} is the quantile of the standard normal distribution.

According to the density function of Log-Normal distribution,

$$f(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(y) - \mu)^2}{2\sigma^2}\right)$$

The quantile density function is then,

$$Q(t) = \frac{1}{F'(y)} = \frac{1}{f(y)} = y \cdot \sigma\sqrt{2\pi} \cdot \exp\left(\frac{(\ln(y) - \mu)^2}{2\sigma^2}\right) \tag{25}$$

Example : For Pareto-II distribution. According to the cumulative distribution function of Pareto-II distribution $F(x)$, the quantile x for the given proportion t (i.e. $F(x) = t$) could be obtained by starting,

$$\begin{aligned} t =& 1 - \left(1 + \frac{x}{\sigma}\right)^{-\alpha} \quad \Longrightarrow \quad \left(1 + \frac{x}{\sigma}\right) = (1-t)^{-\frac{1}{\alpha}} \\ \Longrightarrow \quad & x = \sigma \cdot \left((1-t)^{-\frac{1}{\alpha}} - 1\right) \end{aligned}$$

since,

$$f(x) = \frac{\partial F(x)}{\partial x} = \frac{\sigma}{\alpha} \cdot \left(1 + \frac{x}{\sigma}\right)^{-\alpha-1}, \quad x \geq 0$$

Then quantile density $Q(t)$ could be derived as,

$$Q(t) = \frac{\partial x}{\partial t}$$

$$\begin{aligned} \Longrightarrow \quad & x' = \sigma \cdot \frac{\partial}{\partial t}\left((1-t)^{-\frac{1}{\alpha}} - 1\right) \quad \Longrightarrow \quad x' = \sigma \cdot \left(-\frac{1}{\alpha} \cdot (1-t)^{-\frac{1}{\alpha}-1} \cdot (-1)\right) \\ \Longrightarrow \quad & x' = \frac{\sigma}{\alpha} \cdot (1-t)^{-\frac{1}{\alpha}-1} \end{aligned}$$

Hence final quantile density expression is,

$$Q(t) = \frac{\sigma}{\alpha} \cdot (1-t)^{-\frac{1}{\alpha}-1} \tag{26}$$

Example : For Weibull distribution. By solving the cumulative distribution function of Weibull distribution $F(x) = t$,

$$\begin{aligned} t = 1 - e^{-\left(\frac{x}{\lambda}\right)^k} \quad & \Longrightarrow \quad e^{-\left(\frac{x}{\lambda}\right)^k} = 1 - t \\ \Longrightarrow \quad \left(\frac{x}{\lambda}\right)^k = -\ln(1-t) \quad & \Longrightarrow \quad x = \lambda \cdot (-\ln(1-t))^{1/k} \end{aligned}$$

Then the quantile density function is obtained as:

$$\implies \quad x' = \lambda \cdot \frac{\partial}{\partial t}\left(-\ln(1-t)\right)^{1/k} \quad \implies \quad x' = \lambda \cdot \frac{1}{k} \cdot \left(-\ln(1-t)\right)^{\frac{1}{k}-1} \cdot \frac{\partial}{\partial t}\left(-\ln(1-t)\right)$$
$$\implies \quad x' = \lambda \cdot \frac{1}{k} \cdot \left(-\ln(1-t)\right)^{\frac{1}{k}-1} \cdot \frac{1}{1-t}$$

simplifying,

$$Q(t) = \frac{\lambda \cdot \left(-\ln(1-t)\right)^{\frac{1}{k}-1}}{k \cdot (1-t)} \tag{27}$$

Example : For Dagum distribution. The quantile $y = F^{-1}(t)$ is obtained by $F(y) = t$:

$$\implies \quad t = \left(1 + \left(\frac{b}{y}\right)^a\right)^{-p} \quad \implies \quad \left(1 + \left(\frac{b}{y}\right)^a\right) = t^{-1/p}$$
$$\implies \quad \left(\frac{b}{y}\right)^a = t^{-1/p} - 1 \quad \implies \quad y = \frac{b}{\left(t^{-1/p} - 1\right)^{1/a}}$$

By using the definition of the quantile density function,

$$\implies \quad y' = b \cdot \frac{\partial}{\partial t}\left(\left(t^{-1/p} - 1\right)^{-1/a}\right) \quad \implies \quad y' = b \cdot \left(-\frac{1}{a}\right) \cdot \left(t^{-1/p} - 1\right)^{-\frac{1}{a}-1} \cdot \frac{\partial}{\partial t}\left(t^{-1/p} - 1\right)$$
$$\implies \quad y' = \frac{b}{a \cdot p} \cdot \left(t^{-1/p} - 1\right)^{-\frac{1}{a}-1} \cdot t^{-1/p-1}$$

Simplifying the final expression is given as,

$$Q(t) = \frac{b \cdot \left(t^{-1/p} - 1\right)^{-\frac{1}{a}-1} \cdot t^{-1/p-1}}{a \cdot p} \tag{28}$$

Influence Function-Based Empirical Likelihood Confidence Intervals Construction For Robust Versions of Lorenz Curve. Let $x_1, x_2, \ldots, x_n$ be an independent and identically distributed simple random sample drawn from a common cumulative distribution function $F(x)$ with size n, and they are sorted in ascending order.

Then the Empirical distribution function is defined as,

$$F_n = \frac{1}{n}\sum_{i=0}^{n} I_{\{x_i \leq x^*\}}$$

where x^* is a given point mass and

$$I_{\{x_i \leq x^*\}} = \begin{cases} 1 & \{x_i \leq x^*\} \\ 0 & otherwise \end{cases}$$

is an indicator function. The Empirical likelihood function for three robust versions $j = 1, 2, 3$ with the given proportion t of the whole population is given as,

$$L(\eta_j(t)) = L(p_1, p_2, ..., p_n) = \prod_{i=1}^{n} p_i \tag{29}$$

where p_i are the probabilities assigned to each observation $i = 1, 2, ..., n$. Then, the empirical likelihood ratio function for robust defined as,

$$\begin{aligned} R(\eta_j(t)) = & \frac{\max_{H_0} L(\eta_j(t))}{\max_{H_0 \cup H_1} L(\eta_j(t))} \\ = & \frac{\max_{H_0} L(p_1, \ldots, p_n)}{\max_{H_0 \cup H_1} L(p_1, \ldots, p_n)} \; = \; \frac{\prod_{i=1}^{n} \tilde{p}_i}{\prod_{i=1}^{n} \frac{1}{n}} \\ = & \prod_{i=1}^{n} n \cdot \tilde{p}_i \; ; \quad i = 1, 2, \ldots, n \text{ and } j = 1, 2, 3. \end{aligned} \tag{30}$$

where H_0 is the null hypothesis and H_1 is the alternative hypothesis corresponding to the study.

$\max_{H_0} L(\eta_j(t)) = \max_{H_0} L(p_1, \ldots, p_n)$ is the maximum empirical likelihood function for j^{th} quantile robust version under relevant null hypothesis.

$\max_{H_0 \cup H_1} L(\eta_j(t)) = \max_{H_0 \cup H_1} L(p_1, \ldots, p_n)$ is the overall($H_0 \cup H_1$) maximum empirical likelihood function for j^{th} quantile robust version. This attains its maximum by giving equal weight $1/n$ to each observation (i.e., $p_i = 1/n$).

$\tilde{p}_i = p_i(\eta_j(t))$ are called constrained maximum empirical likelihood estimators for p_i under the corresponding null hypothesis, and they are solutions to the optimization problem of the empirical log-likelihood function,

$$\max\{p_i\} \; \log \left(\prod_{i=1}^{n} P_i \right) \; = \; \max\{p_i\} \sum_{i=0}^{n} \log p_i \tag{31}$$

subject to the probability constraints

$$p_i \geq 0 \; , \; \sum_{i=1}^{n} p_i = 1 \tag{32}$$

and specific moment constraints,

$$\sum_{i=1}^{n} p_i \, f(x_i \, , \, \eta_j(t)) = 0$$

where, $f(x_i \, , \, \eta_j(t))$ is a function of the data and the robust version $\eta_j(t)$ satisfying the property that,

$$\mathbb{E}\Big[f(x_i \, , \, \eta_j(t)) \Big] = 0$$

The influence function of the $\eta_j(t)$ is a function of itself satisfying the property and hence the relevant constraint is,

$$\sum_{i=1}^{n} p_i \, f(x_i \, , \, \eta_j(t)) = \sum_{i=1}^{n} p_i \, \psi_{ij}(t) = 0 \tag{33}$$

where $\psi_{ij}(t) = \; IF(z_i; \eta_j(t), F_n)$ is the influence function of η_j under empirical distribution F_n. The Lagrange multiplier technique is used to solve the above optimization problem. For Lagrange multipliers λ_j and $\phi_j(t)$, the Lagrange function is defined by,

$$\tau_j(t) = \sum_{i=0}^{n} \log p_i + \lambda_j \left(\sum_{i=1}^{n} p_i - 1 \right) + \phi_j(t) \left(\sum_{i=1}^{n} p_i \, \psi_{ij}(t) \right) \tag{34}$$

by taking partial derivative of $\tau_j(t)$ with respect to p_i ,

$$\frac{\partial \tau_j(t)}{\partial p_i} = \sum_{i=1}^{n} \Big[\frac{1}{p_i} + \lambda_j + \phi_j(t)\,\psi_{ij}(t)\Big]. \tag{35}$$

Let $\frac{\partial \tau_j(t)}{\partial p_i} = 0$,

$$\implies \sum_{i=1}^{n} \Big[\frac{1}{p_i} + \lambda_j + \phi_j(t)\,\psi_{ij}(t)\Big] = 0$$

$$\implies \quad \frac{1}{p_i} + \lambda_j + \phi_j(t)\,\psi_{ij}(t) = 0 \tag{36}$$

$$\implies \quad p_i = \frac{-1}{\lambda_j + \phi_j(t)\,\psi_{ij}(t)} \tag{37}$$

From equation(36),

$$1 \;+\; \lambda_j\, p_i \;+\; \phi_j(t)\, p_i\, \psi_{ij}(t) = 0$$

$$\sum_{i=1}^{n} \Big[1 \;+\; \lambda_j\, p_i \;+\; \phi_j(t)\, p_i\, \psi_{ij}(t)\Big] = 0$$

$$\implies \quad n + \lambda_j = 0 \quad \implies \quad \lambda_j = -n$$

by substituting for λ_j and dividing both numerator and denominator by n,

$$\implies \quad p_i = \frac{1/n}{1 - \frac{\phi_j(t)}{n}\,\psi_{ij}(t)}$$

Let $\frac{-\phi_j(t)}{n} = v_j(t)$ and then solution for p_i can be obtained as,

$$p_i(\eta_j(t)) = \frac{1}{n\;(1 + v_j(t)\,\psi_{ij}(t))} \tag{38}$$

As $\sum_{i=1}^{n} p_i\,\psi_{ij}(t) = 0$, the value of $v_j(t)$ is the unique solution to the equation,

$$\frac{1}{n}\Big\{\; \sum_{j=1}^{n} \frac{\psi_{ij}(t)}{1 + v_j(t)\,\psi_{ij}(t)} \Big\} = 0 \tag{39}$$

Since $\psi_{ij}(t)$ depends on unknown population quantiles, $\epsilon_{t/2},\ d_1(t),\ d_2(t)$ and $d_3(t)$ substituting them with their consistent estimators $x([np])$, the observation related term $\hat{\psi}_{ij}(t)$ could be obtained. $x([np])$ is the $[np]^{th}$ ordered value of $x_i's$ for associated proportion p of population for quantiles $\epsilon_{t/2},\ d_1(t),\ d_2(t)$ and $d_3(t)$. Then, for inequality curves at proportion t, the estimated empirical likelihood ratio, $\hat{R}(\eta_j(t))$, and the estimated empirical log-likelihood ratio, $\hat{l}(\eta_j(t))$, are defined as follows,

$$\hat{R}(\eta_j(t)) = \prod_{i=1}^{n} \{1 + v_j(t)\,\hat{\psi}_{ij}(t)\}^{-1} \tag{40}$$

$$\hat{l}(\eta_j(t)) = -2 \cdot \ln\{\hat{R}(\eta_j(t))\} = 2 \cdot \sum_{i=1}^{n} \ln\{1 + v_j(t)\,\hat{\psi}_{ij}(t)\} \tag{41}$$

Lemma 1. *If* $\mathbb{E}(X^2) < \infty$ *and for any given* $t \in (0,1)$ *then,*

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\psi}_{ij}(t) - \psi_{ij}(t)\right]^2 = o_p(1)$$

Lemma 2. I. $\max_i \left|\hat{\psi}_{ij}(t)\right| = o_p\left(\sqrt{n}\right)$, $t \in [0,1]$, *where* $i = 1, ..., n$ *and* $j = 1, 2, 3$.

II. $\frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_{ij}^2(t) = \sigma^2 + o_p(1)$.

III. $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\psi}_{ij}(t) \xrightarrow{L} \mathcal{N}\left(0, \sigma^2\right)$.

If $\mathbb{E}(X^2) < \infty$ and let $\eta_j(t)$ be the j^{th} robust version at any given $t \in (0,1)$. Thus, $\hat{l}(\eta_j(t))$ possesses a limiting distribution that is a standard chi-square distribution with one degree of freedom.

Using the convergence property,

$$\hat{l}(\eta_j(t)) \xrightarrow{L} \chi_1^2 \quad \text{as} \quad n \to \infty$$

the $(1-\alpha)\ 100\,\%$ empirical likelihood confidence interval based on influence function for $\eta_j(t)$ with the critical value of the Chi-square distribution at the α significance level $(\chi^2_{1,1-\alpha})$ can be derived as,

$$(l_j, u_j) = \left[\eta_j(t) : \hat{l}(\eta_j(t)) \leq \chi^2_{1,1-\alpha}\ ,\ \text{for } t \in [0,1]\right] \tag{42}$$

Bootstrap Confidence Intervals - Method I. Bootstrapping is a statistical method that involves collecting random samples multiple times to create a sampling distribution from a single sample. It is used to find confidence intervals for statistics without making more assumptions and to examine rare events or tail distributions where theoretical assumptions are difficult to justify.

Let $\{x_1^*, x_2^*, ..., x_n^*\}$ be the randomly drawn n observations with replacement from the main sample containing data $\{x_1, x_2, ..., x_n\}$ of size n that are taken from the original data. By repeating this process several times, bootstrap resamples can be obtained. Each bootstrap sample might include duplicate observations from the original dataset due to replacement. Based on the definition of the bootstrap approach for the general Lorenz ordinate at a given t cumulative population proportion [Qin(2013)], the corresponding definition for each robust version could be derived as follows.

$$\eta_1^m(t) = \frac{\sum_{i=1}^{n} x_i^*\, I(x_i^* \leq \epsilon^*_{t/2})}{\sum_{i=1}^{n} x_i^*\, I(x_i^* \leq \epsilon^*_{0.5})} \tag{43}$$

$$\eta_2^m(t) = \frac{\sum_{i=1}^{n} x_i^*\, I(x_i^* \leq \epsilon^*_{t/2})}{\sum_{i=1}^{n} x_i^*\, I(x_i^* \leq \epsilon^*_{1-t/2})} \tag{44}$$

$$\eta_3^m(t) = \frac{\sum_{i=1}^{n} x_i^*\, I(x_i^* \leq \epsilon^*_{t/2})}{\frac{\sum_{i=1}^{n} x_i^*\, I(x_i^* \leq \epsilon^*_{t/2}) + \sum_{i=1}^{n} x_i^*\, I(x_i^* \leq \epsilon^*_{1-t/2})}{2}} \tag{45}$$

$$= \frac{2 \cdot \sum_{i=1}^{n} x_i^*\, I(x_i^* \leq \epsilon^*_{t/2})}{\sum_{i=1}^{n} x_i^*\, I(x_i^* \leq \epsilon^*_{t/2}) \ + \ \sum_{i=1}^{n} x_i^*\, I(x_i^* \leq \epsilon^*_{1-t/2})} \tag{46}$$

where, $\{\eta_j^m(t) : m = 1, 2, \ldots, K \ \text{ and } \ j = 1, 2, 3\}$ is the j^{th} robust version of the Lorenz curve defined over the n number of observations in the bootstrap re-sample m and ϵ_x^* is the x^{th} sample quantile of bootstrap re-observed data x_i^* s such that $x = 0.5,\ t/2$ and $1 - t/2$. After that, repeatedly drawing K number of bootstrap resamples, the asymptotic variance of each robust version of the Lorenz curve $\eta_j^*(t)$, $j = 1, 2, 3$ with K bootstrap copies could be estimated by,

$$\mathrm{Var}(\eta_j^*(t)) = \frac{1}{K-1}\sum_{m=1}^{K}\left[\eta_j^m(t) - \bar{\eta_j^m}(t)\right]^2 \tag{47}$$

where,

$$\bar{\eta_j^m}(t) = \frac{1}{K}\sum_{m=1}^{K}\eta_j^m(t). \tag{48}$$

Hence, $(1-\alpha)100\%$ level bootstrap confidence intervals for each robust version are determined by,

$$\left(\bar{\eta_j^m}(t) - z_{1-\alpha/2}\sqrt{\mathrm{Var}(\eta_j^*(t))}\ ,\ \bar{\eta_j^m}(t) + z_{1-\alpha/2}\sqrt{\mathrm{Var}(\eta^*(t))}\right).$$

Bootstrap Confidence Intervals - Method II. Based on bootstrap variance, another type of $(1-\alpha)100\%$ level bootstrap confidence intervals for each robust version can be derived as follows:

$$\left(\hat{\eta}_j(t) - z_{1-\alpha/2}\sqrt{\mathrm{Var}(\eta_j^*(t))}\ ,\ \hat{\eta}_j(t) + z_{1-\alpha/2}\sqrt{\mathrm{Var}(\eta^*(t))}\right).$$

where, $\hat{\eta}_j(t), j = 1,2,3$ is the j^{th} estimated robust ordinate from main sample data $\{x_1, x_2, ..., x_n\}$ at given proportion t.

Influence Function-based Hybrid Empirical Likelihood-Bootstrap Confidence Intervals. In here need to derive $\hat{l}^*(\eta_j(t))$, $v_j^*(t)$ and $\hat{\psi}_{ij}^*(t)$ which are bootstrap versions of $\hat{l}(\eta_j(t))$, $v_j(t)$, $\hat{\psi}_{ij}(t)$ described in section (2.3). Repeating over K number of bootstrap resamples, set of bootstrap replications $\{\hat{l}_m^*(\eta_j(t)),\ m = 1,2,...,K\}$ and after sorting them, $[K(1-\alpha)]^{th}$ ordered value of $\hat{l}_m^*(\eta_j(t))$ for given t $\left(i.e:\ \hat{l}^*_{[K(1-\alpha)]}(\eta_j(t))\right)$ are obtained.

The hybrid empirical likelihood-bootstrap confidence intervals for each robust version are derived as,

$$(l_j^*, u_j^*) = \left[\eta_j(t) : \hat{l}(\eta_j(t)) \le \hat{l}^*_{[K(1-\alpha)]}(\eta_j(t)),\ \text{for } t \in [0,1]\right] \tag{49}$$

Results & Discussion

Simulation Studies. Since income data frequently show skewness and heterogeneity, some statistical distributions are well-suited for modeling income data due to their heavy-tail nature and positive skewness. To construct the confidence intervals, the sample size n is chosen to be 100 and 300 for each data simulated from four key distributions: Pareto-II(shape = 2, scale = 1), Weibull(shape = 1, scale = 2), Log-Normal(mean = 0, sd = 1), and Dagum(scale = 2, shape1 = 1, shape2 = 1), which are then treated as observed data in the empirical likelihood framework.

Simulated data sets from each distribution are particularly useful in this context because they represent realistic income or wealth distribution scenarios, allowing testing of the method's robustness under varying levels of skewness and inequality. The 95% confidence intervals for each quantile inequality curve $\eta_j(t)$, $j = 1,2,3$ at different $t \in [0,1]$ were constructed using data-driven methods from each distribution. Under these simulation settings, 100 random samples of sizes 100 and 300 were generated, and the average length of the intervals and coverage probability were calculated through these samples.

Algorithm 1 : Influence function-based empirical likelihood method. **Inputs** : Sample size (n), Sorted sample data, Selected cumulative population proportions (t), Significance level (α), Critical value for the chi-square distribution ($\chi^2_{1,1-\alpha}$), Numerical tolerance (ε)

Process :

- Loop over each t to compute $Q(t), \hat{\psi}_{ij}(t)$ and solve for $v_j(t)$.
- Use root finding Secant method to find both bounds using the condition $\hat{l}(\eta_j(t)) \leq \chi^2_{1,1-\alpha}$
  - Start with initial guesses x_0 and x_1
  - Iterate using the formula

$$x_{new} = x_1 - f(x_1)\frac{(x_1 - x_0)}{f(x_1) - f(x_0)}$$

  - Stop when $|\, f(x_{new})\,| < \varepsilon$.

Output : $\Big\{\textbf{lower bound} = l_j \quad , \quad \textbf{upper bound} = u_j\Big\}$.

Algorithm 2 : Bootstrap - I method. **Inputs** : Sorted main sample data, Number of bootstrap samples (K), n, α, t, Standard normal critical value ($z_{\alpha/2}$)

Process :

- Draw K number of bootstrap samples.
- Loop over each bootstrap sample and t to compute $\eta_j^{\bar{m}}(t)$ and $\mathrm{Var}(\eta^*(t))$

Output :

$$\Big\{\textbf{lower bound} = \eta_j^{\bar{m}}(t) - z_{1-\alpha/2}\sqrt{\mathrm{Var}(\eta_j^*(t))}\,,\ \textbf{upper bound} = \eta_j^{\bar{m}}(t) + z_{1-\alpha/2}\sqrt{\mathrm{Var}(\eta^*(t))}\Big\}$$

Algorithm 3 : Bootstrap - II method. **Inputs** : Sorted main sample data, Number of bootstrap samples (K), n, α, t, Standard normal critical value ($z_{\alpha/2}$)

Process :

- Draw K number of bootstrap samples.
- Compute estimated quantile ordinate using main sample data ($\hat{\eta}_j(t)$)
- Loop over each bootstrap sample and t to compute $\eta_j^{\bar{m}}(t)$ and $\mathrm{Var}(\eta^*(t))$

Output :

$$\Big\{\textbf{lower bound} = \hat{\eta}_j(t) - z_{1-\alpha/2}\sqrt{\mathrm{Var}(\eta_j^*(t))}\,,\ \textbf{upper bound} = \hat{\eta}_j(t) + z_{1-\alpha/2}\sqrt{\mathrm{Var}(\eta^*(t))}\Big\}$$

Algorithm 4 : Influence function-based hybrid empirical likelihood-bootstrap method. **Inputs** : Sorted main sample data, K, n, α, t

Process :

- Draw K number of bootstrap samples.
- Loop over each bootstrap sample and t to compute $\hat{l}^*(\eta_j(t))$, $v_j^*(t)$ and $\hat{\psi}_{ij}^*(t)$.
- Obtain $\hat{l}^*_{[K(1-\alpha)]}(\eta_j(t))$.
- Choose candidate $\eta_j(t)$ satisfying $\hat{l}(\eta_j(t)) \leq \hat{l}^*_{[K(1-\alpha)]}(\eta_j(t))$.

Output :

$$\Big\{\textbf{lower bound} = min\Big(\eta_j(t)\Big)\,,\ \textbf{upper bound} = max\Big(\eta_j(t)\Big)\Big\}$$

According to Figures (6), (7), (8), and (9), the confidence intervals cover the quantile inequality curves almost everywhere and show that for each data point simulated from four key distributions. They provide a

According to Figures (6), (7), (8), and (9), the confidence intervals cover the quantile inequality curves almost everywhere and show that for each data point simulated from four key distributions. They provide a slightly similar behavior and a more even spread compared to each other. The curves and confidence intervals highlight differences in how each version captures extreme values by reflecting robustness in the presence of data irregularities. As the cumulative proportiont of the population increases, the curve starts to capture fewer people with very high incomes that contribute disproportionately to total income by increasing the variability of the estimates. This is an advancement in making inferences using interval estimation rather than point estimation. The interval estimation is more informative and accounts for variability and uncertainty, providing greater insight to make decisions based on real-world data.

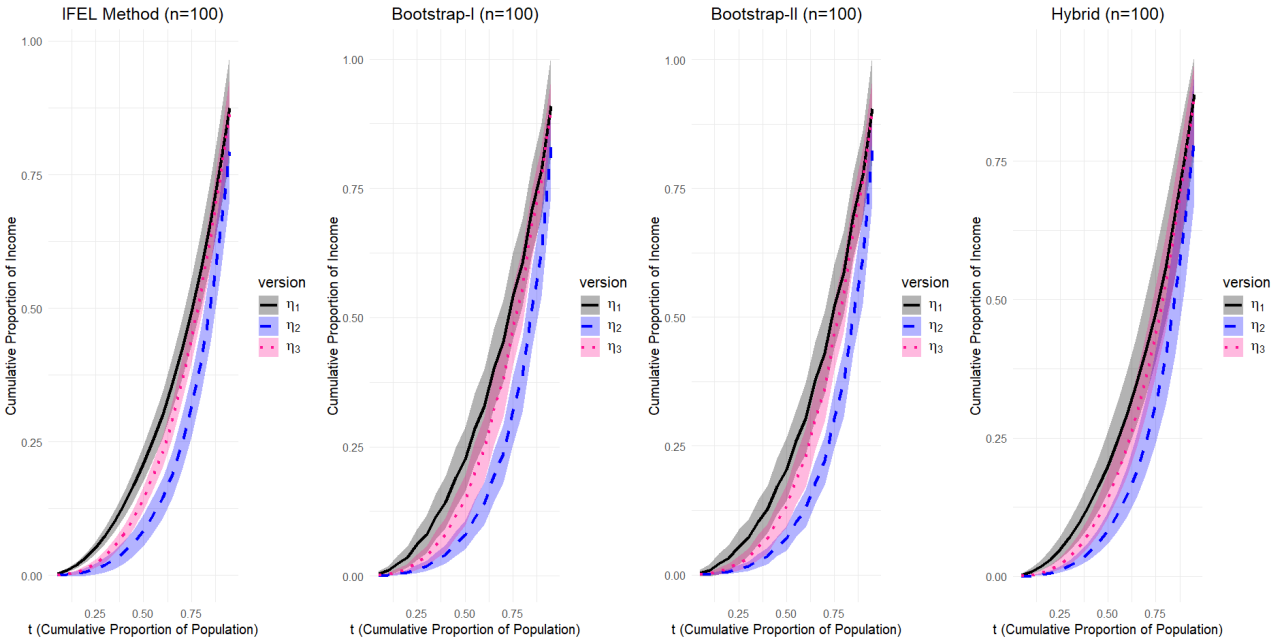


Figure 6. 95% Confidence Intervals for each Robust version [for simulated data from Pareto-II(2,1) with sample size = 100]

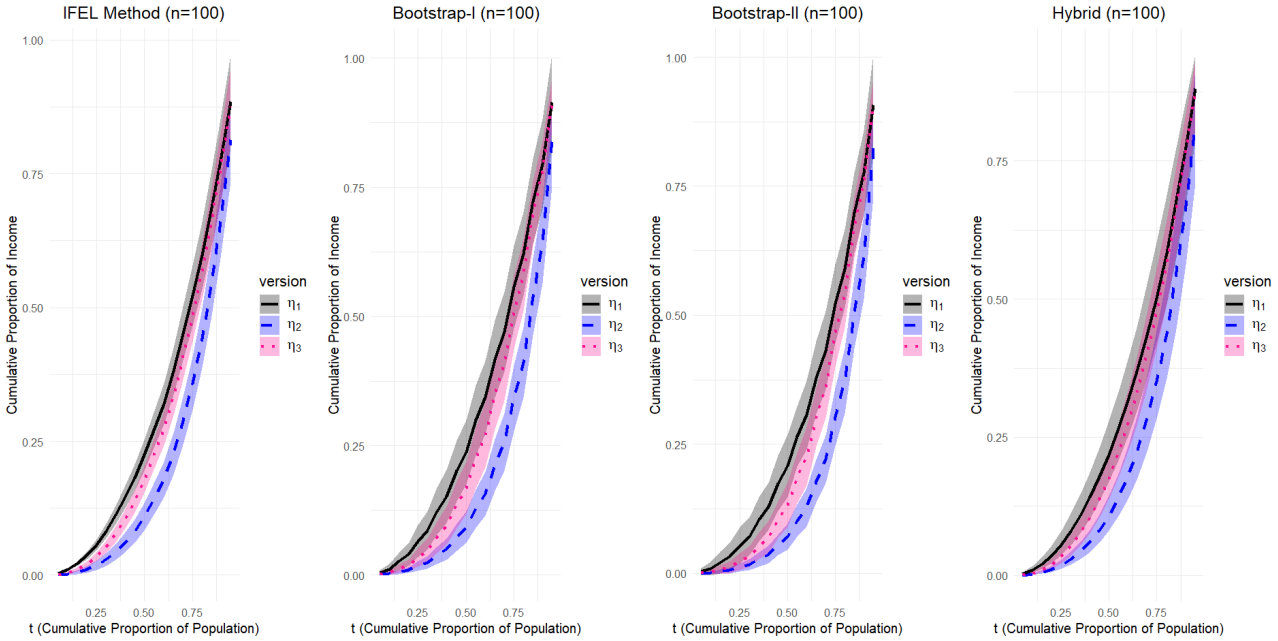


FIGURE 7. 95% Confidence Intervals for each Robust version [for simulated data from Weibull(1,2) with sample size = 100]

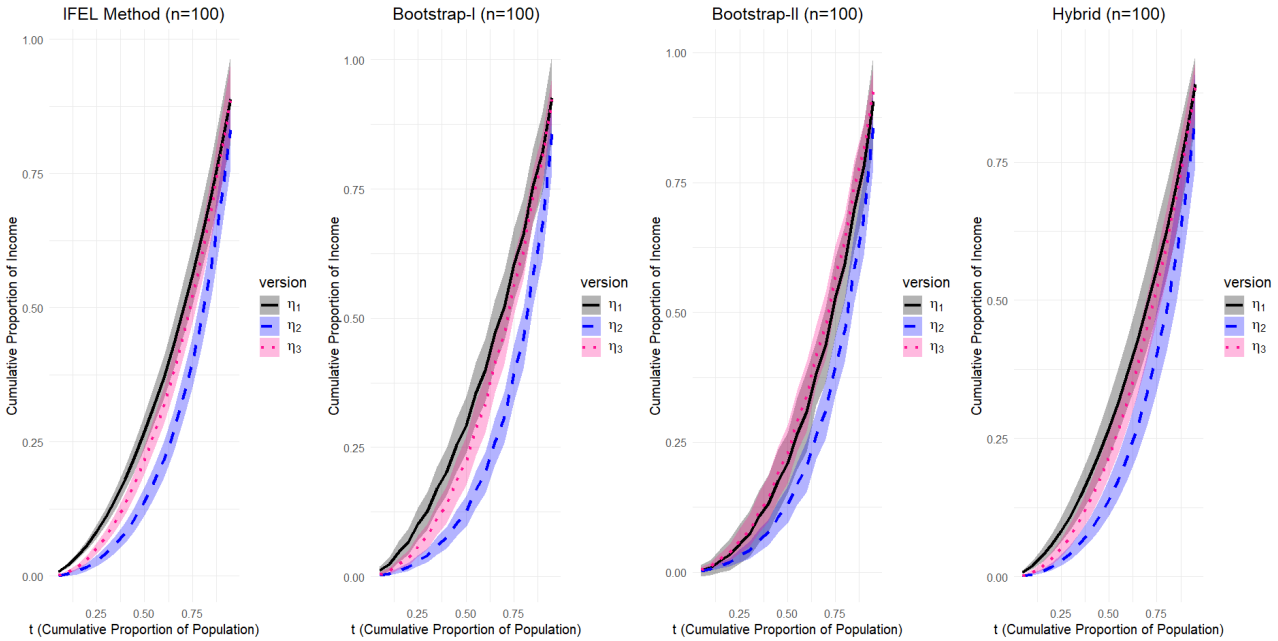


FIGURE 8. 95% Confidence Intervals for each Robust version [for simulated data from LogNormal(0,1) with sample size = 100]

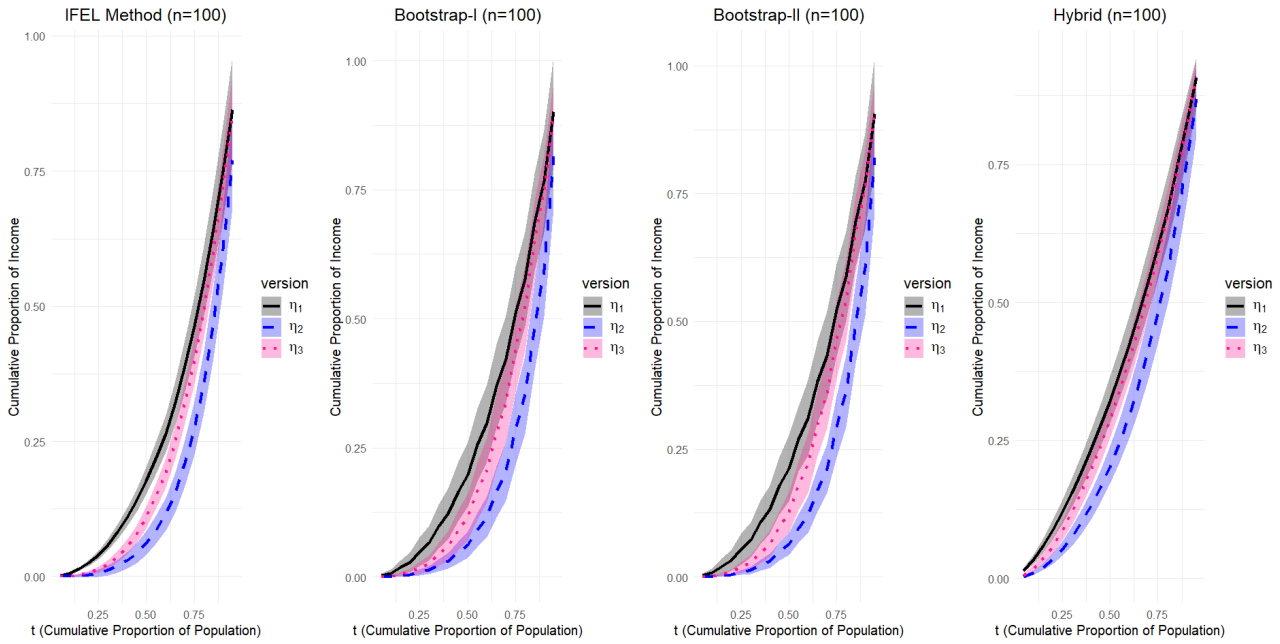


FIGURE 9. 95% Confidence Intervals for each Robust version [for simulated data from Dagum(2,1,1) with sample size = 100]

According to the law of large numbers, as the sample size grows, the sample statistics converge progressively closer to the actual population parameters. The estimation of confidence intervals for each robust version also uses this idea, which reduces the variability of random sampling and produces more accurate and condensed interval estimates. Large samples capture more of the variability of the population, including extreme values to better represent the entire distribution. In summary, increasing the sample size reduces the uncertainty in the estimation process, leading to tighter confidence intervals. Figures (10), (11), (12), and (13) below show that 95% confidence intervals for each robust version are derived from four methods with increased sample size (n = 300). Also, the figures clearly show the underlying behavior that occurred when the sample size was increased.

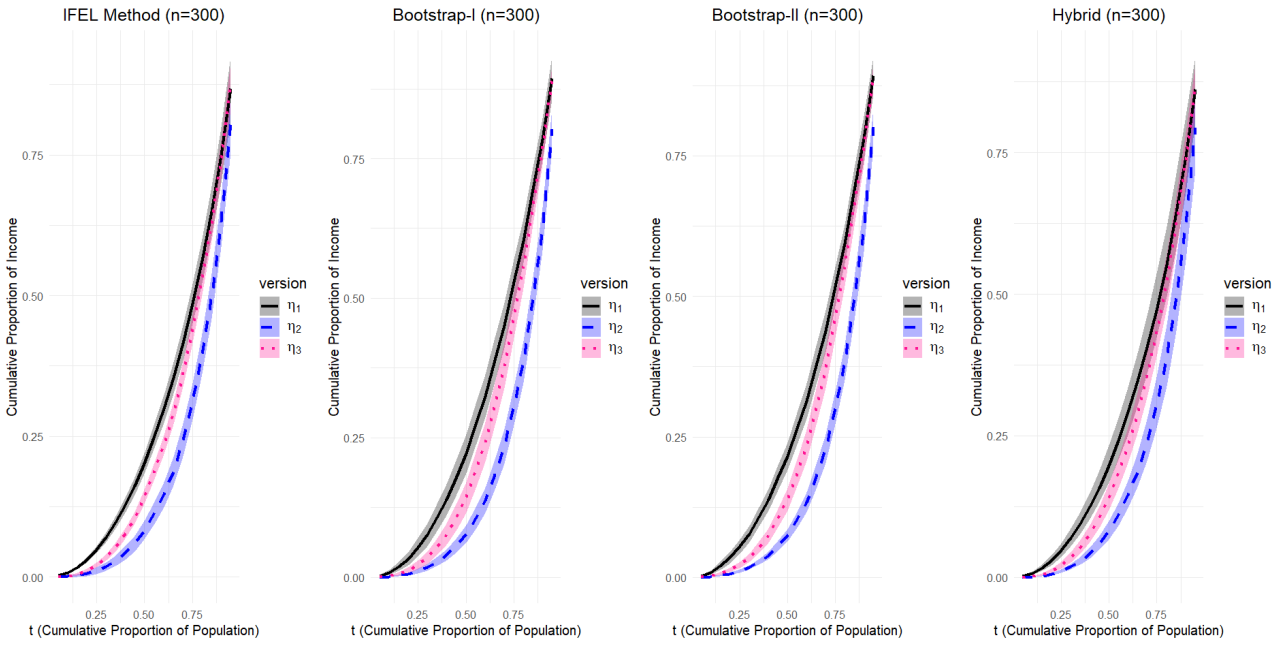


FIGURE 10. 95% Confidence Intervals for each Robust version [for simulated data from Pareto-II(2,1) with sample size = 300]

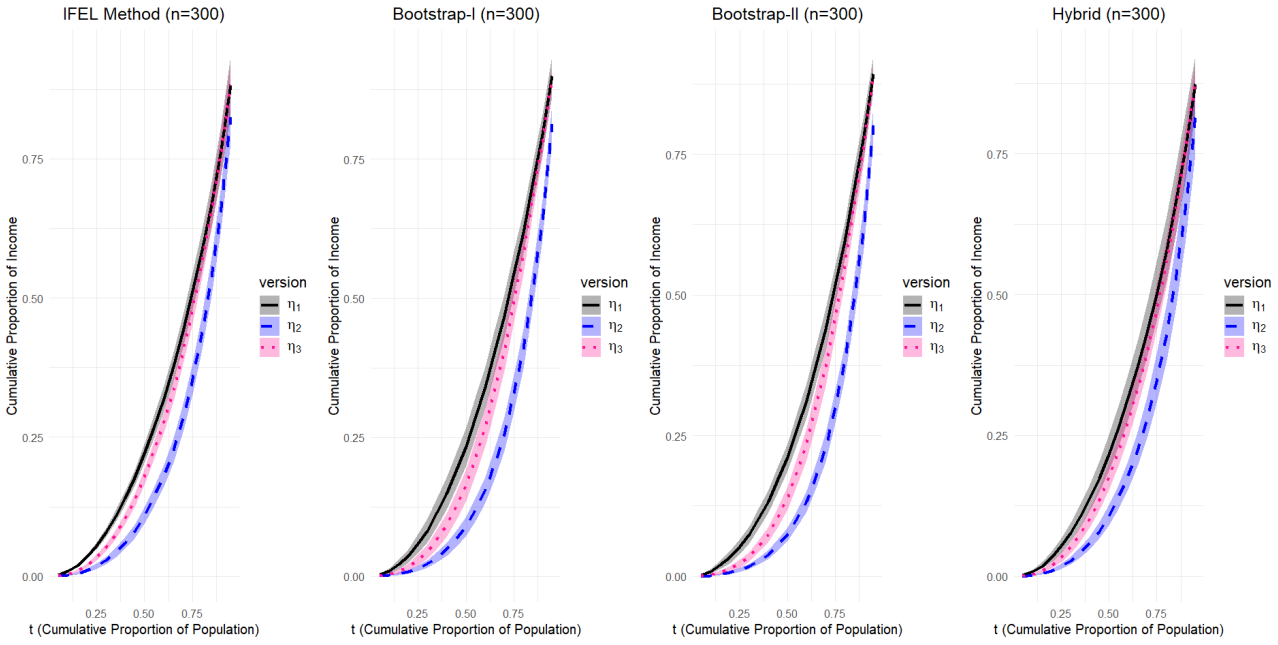


FIGURE 11. 95% Confidence Intervals for each Robust version [for simulated data from Weibull(1,2) with sample size = 300]

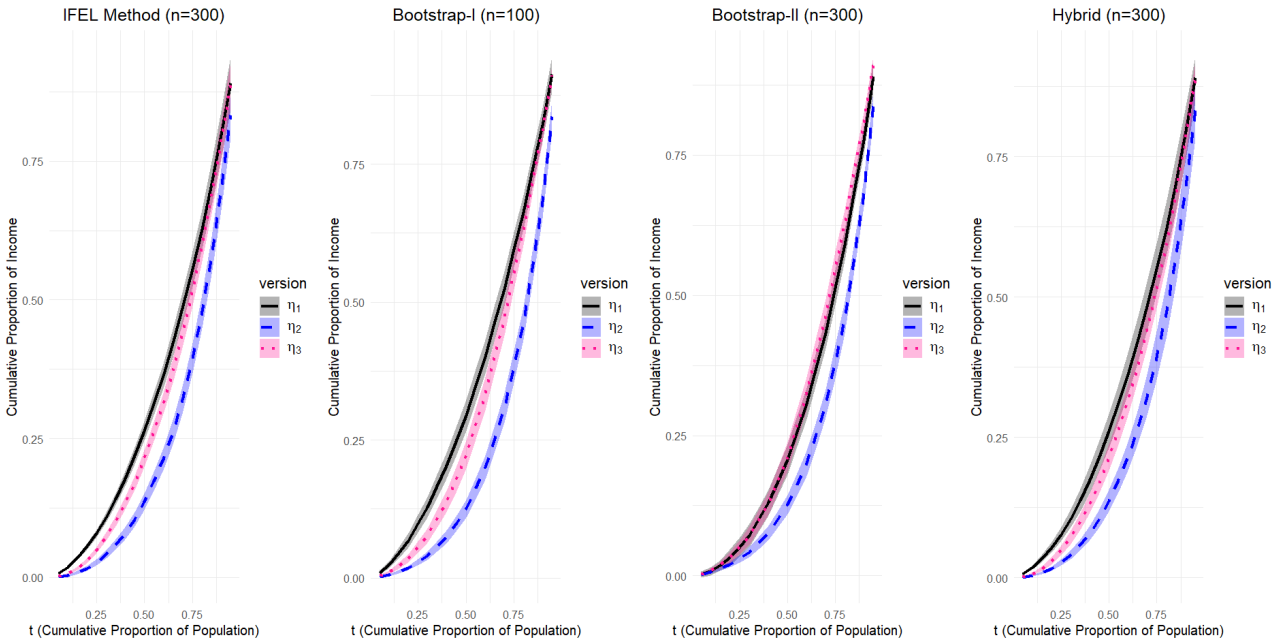


FIGURE 12. 95% Confidence Intervals for each Robust version [for simulated data from Log-Normal(0,1) with sample size = 300]

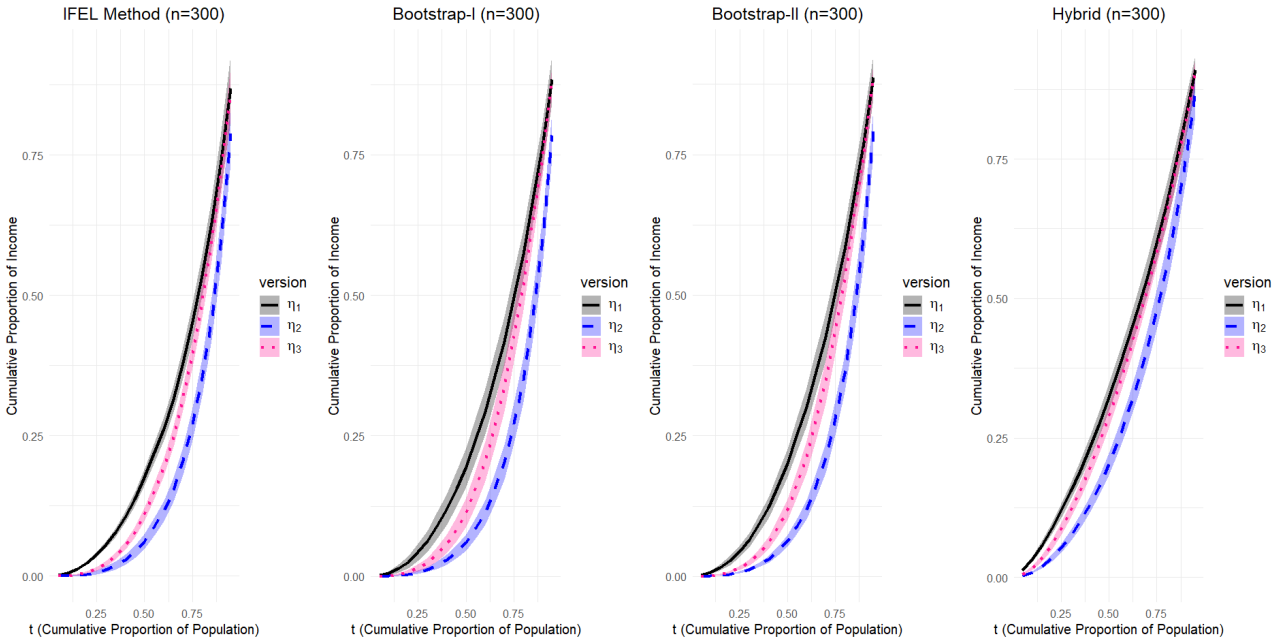
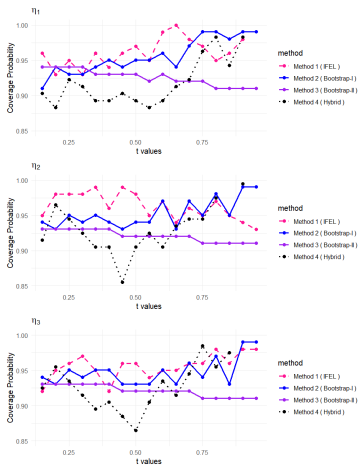


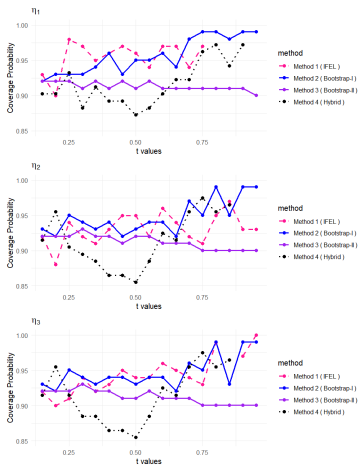
FIGURE 13. 95% Confidence Intervals for each Robust version [for simulated data from Dagum(2,1,1) with sample size = 300]

The coverage probability refers to the actual proportion of the time that the true value of the robust versions of the Lorenz curve lies within the constructed confidence interval over repeated sampling or simulations. The confidence intervals derived from each method provide better coverage, achieving the nominal level. The achievement of nominal coverage indicates that the interval estimation methods are accurate and reliable for

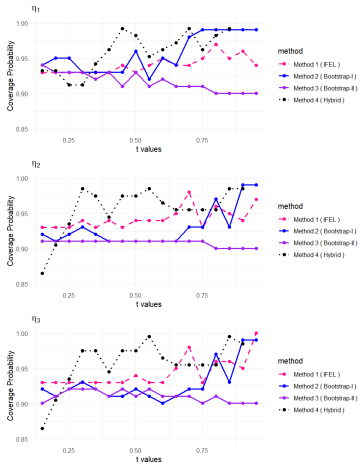
the considered data. Also, they indicate that the selected sample size is sufficient to estimate the confidence intervals and the methods are well-calibrated and perform as intended. Better coverage could also be obtained by increasing the number of simulation settings. In particular, the hybrid empirical likelihood-bootstrap technique provides better performance by increasing the number of bootstrap samples (K 200), and hereK = 100 is used. Additionally, approaches are effective for moderate or small sample sizes (n = 100 andn = 300). Compared to influence function-based empirical likelihood, which uses influence functions to better manage data abnormalities, bootstrap techniques I and II are less reliable at capturing the long-tail behavior with extreme values. Additionally, when the underlying distribution is unknown, the empirical likelihood approaches based on impact functions have an advantage.



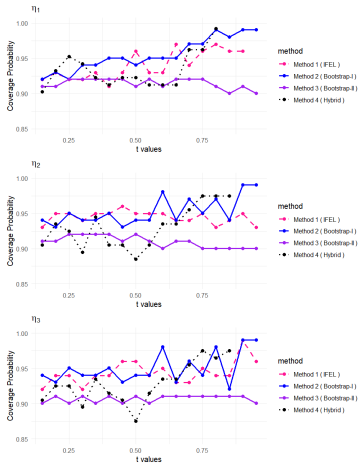
(a) For data simulated from Pareto-II(2,1)



(b) For data simulated from Weibull(1,2)



(c) For data simulated from LogNormal(0,1)

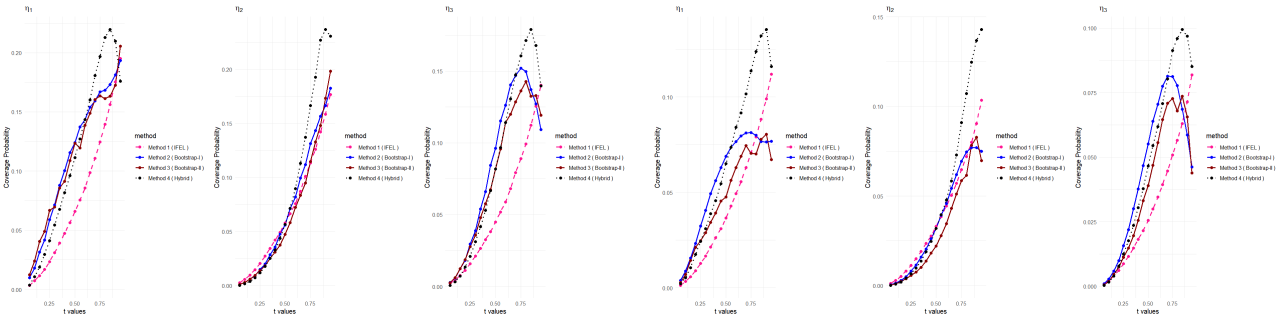


(d) For data simulated from Dagum(2,1,1)

FIGURE 14. Coverage probability for each robust version

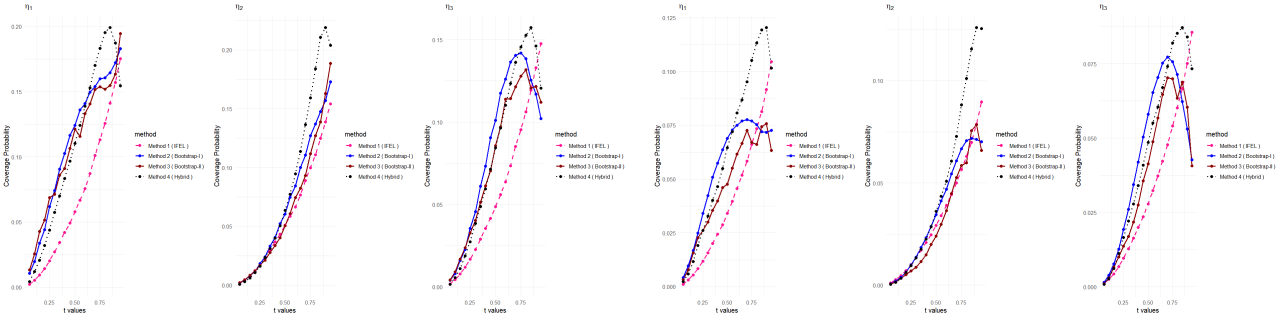
The average length is referred to as the average width of confidence intervals computed over repeated samples. A confidence interval's average length indicates how accurate the estimation was. For every approach, the average length of the confidence intervals increased witht . This is to be expected since, as the cumulative population proportion rises, so does unpredictability and uncertainty. In the lower values oft , a large proportion of the population contributes to the cumulative income due to low-income holders. The high density of data

points reduces variability. Fewer data points representing high-income individuals remain at the extreme ends, leading to increased uncertainty. These are the reasons for increasing the length of the intervals whent increases. In applications like policy-making or economic analysis, shorter intervals ensure that decisions are based on more accurate and dependable estimates. The absence of curvature at the ends of the influence function-based empirical likelihood confidence intervals compared to bootstrap-based intervals highlights the differences in their underlying mechanics. Unlike bootstrap methods, which rely on repeated resampling and can suffer from sparse representation in the tails, influence function-based empirical likelihood directly uses the data's statistical properties. This avoids the reduced variability at the ends caused by limited resampled observations.



(a) For Pareto-II(2,1) with $n = 100$

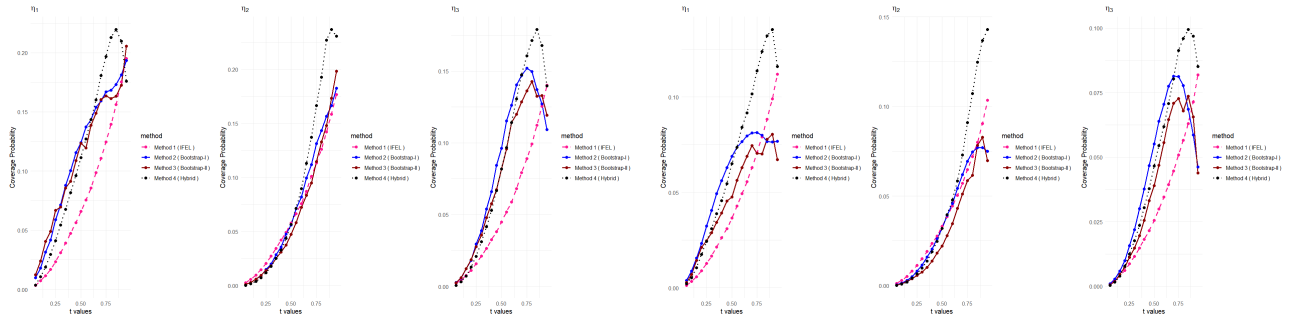
(b) For Pareto-II(2,1) with $n = 300$



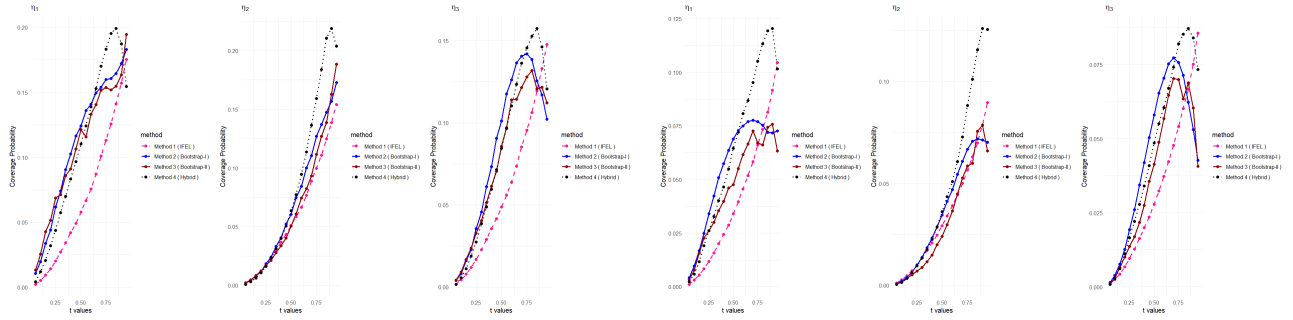
(c)For Weibull(1,2) with $n = 100$

(d) For Weibull(1,2) with $n = 300$

FIGURE 15. Average Length for each robust version for data simulated from different distributions.



(e) For LogNormal(0,1) with $n = 100$ (f) For LogNormal(0,1) with $n = 300$



(g) For Dagum(2,1,1) with $n = 100$ (h)For Dagum(2,1,1) with $n = 300$

FIGURE 16. Average Length for each robust version for data simulated from different distributions.

Real Data Application. The Panel Study of Income Dynamics (PSID) is the longest-running longitudinal household survey, dating back to 1968. It includes more than 18,000 people from 5,000 families in the United States and their descendants, providing comprehensive data across generations. The survey collects data on employment, income, wealth, expenditures, health, marriage, childbearing, child development, and education. The most recent publication is the 2021 data collection (released in October 2023), which allows researchers to examine social and economic consequences across generations. In this study, the data contained in the variable "ER81775 - TOTAL FAMILY INCOME 2020" of the 2021 PSID dataset are used as actual data. The selected data set includes 9,207 interview cases, which record household income levels for the 2020 tax year, including the taxable income of the reference person and the spouse, other taxable income, transfer income, and social security income.

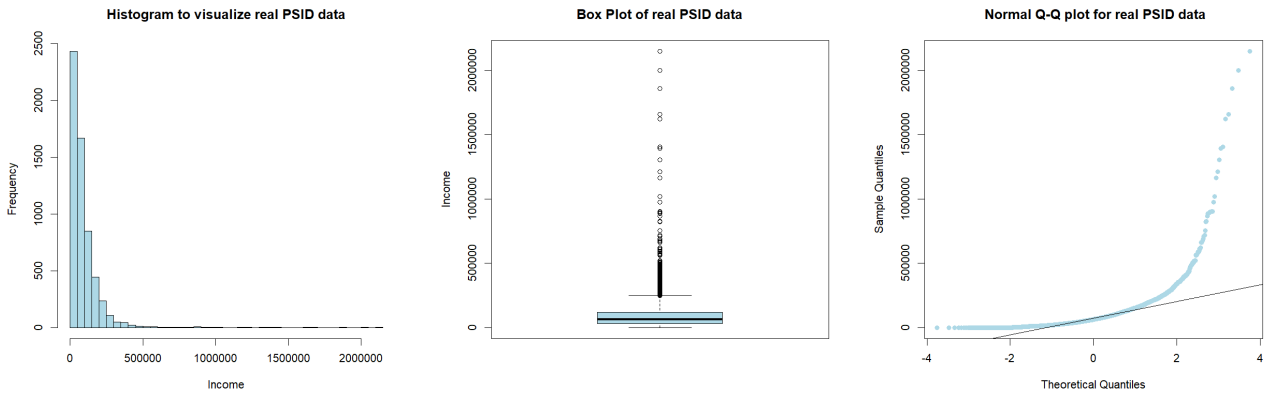


FIGURE 17. Behavior of Total Family Income Data for the 2020 Tax Year.

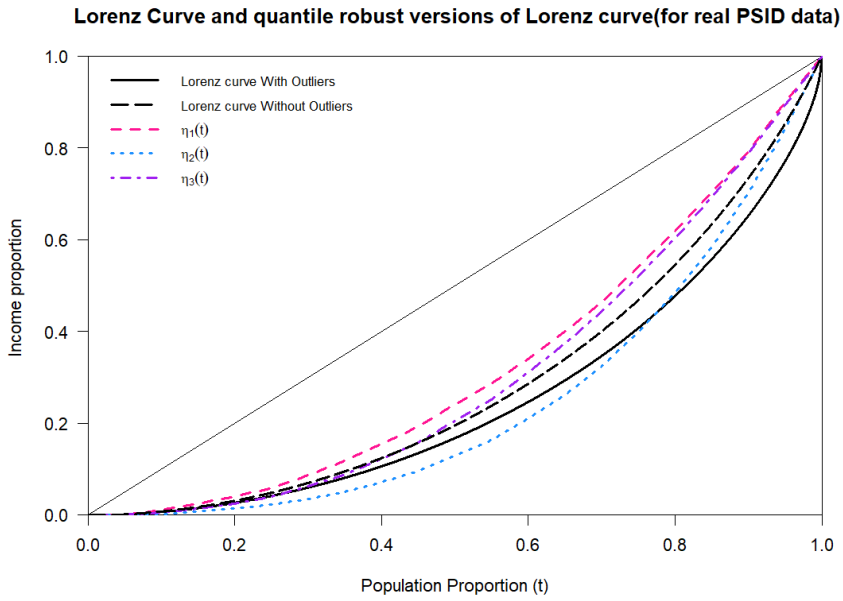


FIGURE 18. Visualization of Lorenz curve and three robust versions for total family income data.

The figures below illustrate recommended 95% confidence intervals for each robust version of the Lorenz curve, their average lengths, and coverage probabilities computed using actual data with 100 sample sizes based on the stated theoretical point of view.

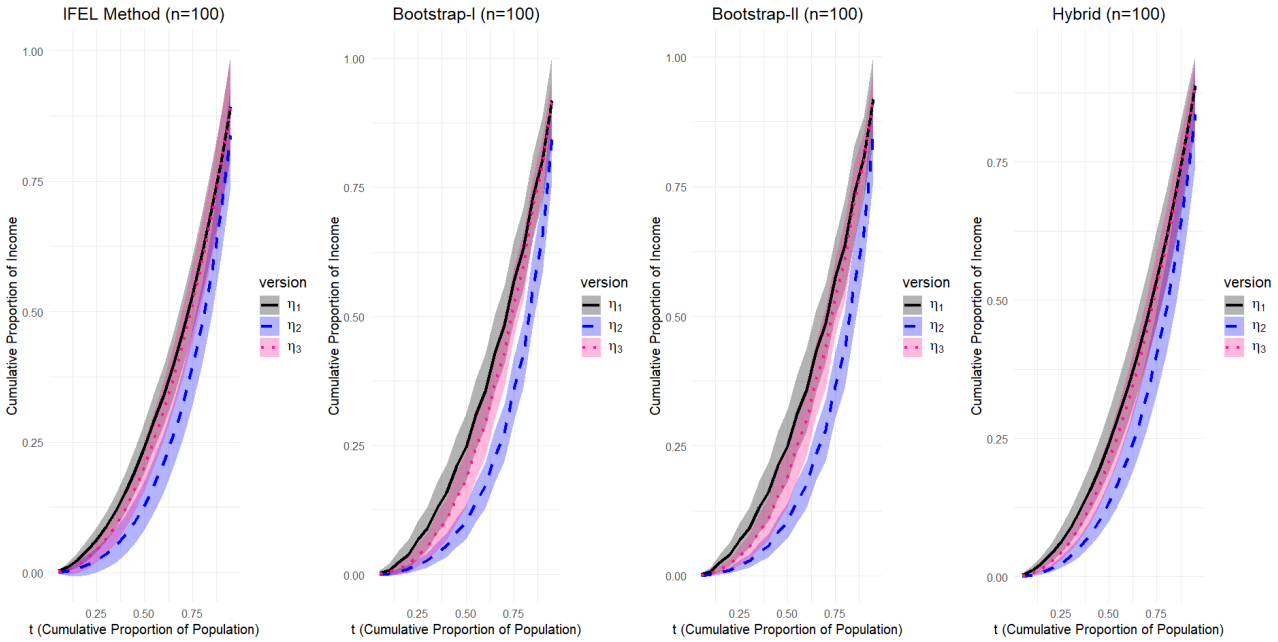


FIGURE 19. 95% confidence intervals for PSID total family income data (sample size = 100).

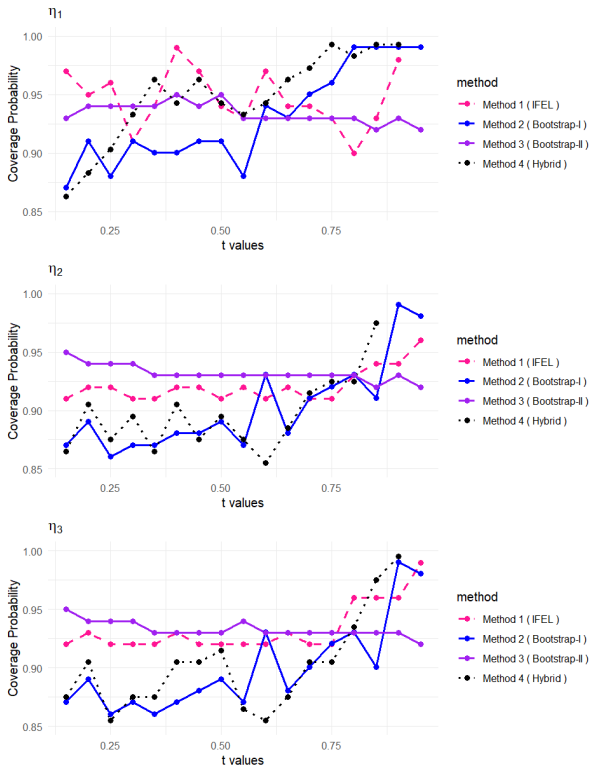


FIGURE 20. Coverage probabilities of each robust version for PSID total family income data (sample size = 100).

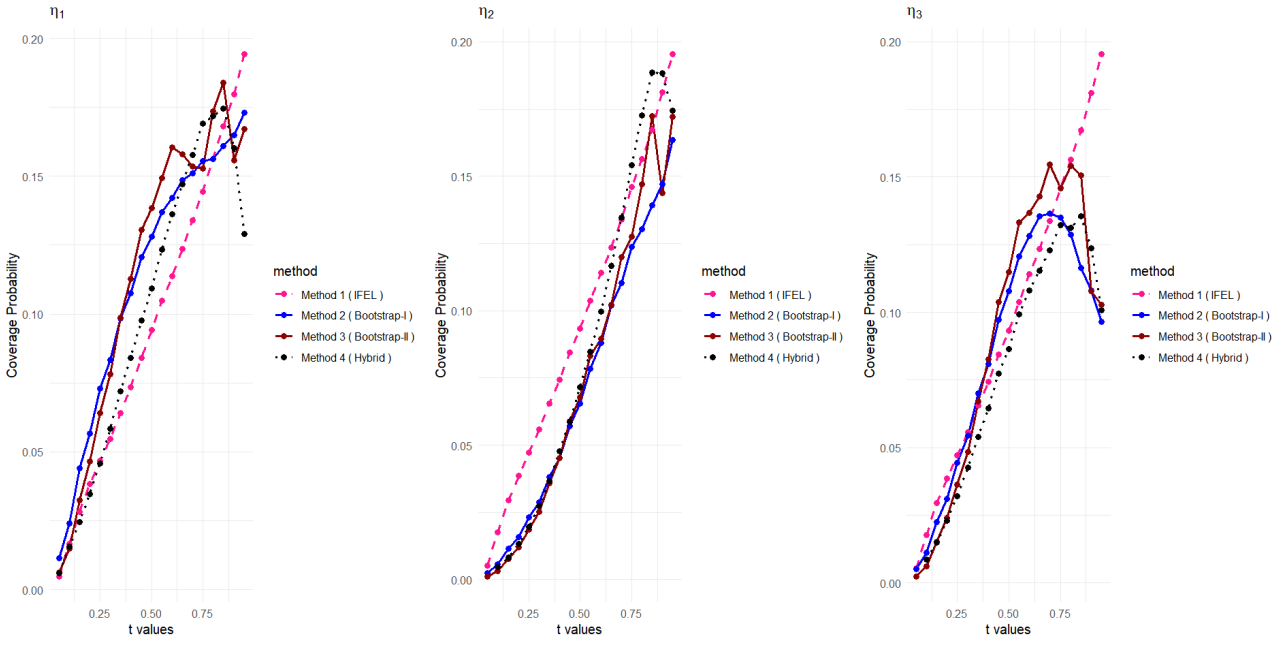


FIGURE 21. Average lengths of each robust version for PSID total family income data (sample size = 100).

This study introduced a robust quantile-based empirical likelihood approach incorporating influence functions to accurately estimate the Lorenz curve, a critical measure of income and wealth inequality. By addressing common issues prevalent in economic data, such as skewness, outliers, and heterogeneity, the proposed method substantially improves the reliability and robustness of statistical inference compared to traditional empirical likelihood and bootstrap alternatives. Through comprehensive simulation studies mimicking realistic economic datasets, the influence function-based empirical likelihood method demonstrated superior performance in constructing confidence intervals for the Lorenz curve. This advantage arises primarily from the ability of influence functions to effectively mitigate the distortive effects caused by extreme observations and data contamination.

While this method presents significant advantages, it also encounters limitations, particularly in computational intensity and dependence on assumptions associated with chosen influence functions. To further enhance the practical applicability and efficiency of this approach, future research should explore computational optimization strategies, extend the methodology to a wider range of socioeconomic measures, and rigorously evaluate its theoretical properties and empirical effectiveness across various economic contexts. Ultimately, the robust empirical likelihood framework proposed in this paper equips researchers and policymakers with a powerful analytical tool for accurately assessing income inequality, thereby enabling more informed and effective socioeconomic decision-making.Acknowledgments

The authors would like to express their sincere gratitude to the editors and anonymous reviewers for their invaluable comments and constructive feedback, which significantly contributed to the enhancement of this paper.Author Contribution

R. Ghasem Pour: software, and writing. W.A.R De Mel: conceptualization. methodology,and editing. All authors have read and agreed to the published version of the manuscript.Funding

The authors declare that no external funding or support was received for the research presented in this paper, including administrative, technical, or in-kind contributions. Data Availability All data supporting the reported findings in this research paper are provided within the manuscript.Conflicts of Interest

The authors declare that there is no conflict of interest concerning the reported research findings. Funders played no role in the study's design, in the collection, analysis, or interpretation of the data, in the writing of the manuscript, or in the decision to publish the results.Appendix Section

Proof of Lemma 1:

Let's consider terms related to the first robust version $j = 1$;

$$\psi_{i1}(t) = t \left\{ Q(^{t}_{\overline{2}}) \frac{\left[\frac{t}{2} - I(z < \epsilon_{t/2})\right]}{\epsilon_{0.5}} - \frac{\epsilon_{t/2} \cdot Q(0.5)\left[0.5 - I(z < \epsilon_{0.5})\right]}{\epsilon_{0.5}^2} \right\}$$

$$\hat{\psi}_{i1}(t) = t\left\{\frac{Q(\frac{t}{2})\left[\frac{t}{2} - I(z < \hat{\epsilon}_{t/2})\right]}{\hat{\epsilon}_{0.5}} - \frac{\hat{\epsilon}_{t/2}\cdot Q(0.5)\left[0.5 - I(z < \hat{\epsilon}_{0.5})\right]}{\hat{\epsilon}^2_{0.5}}\right\}$$

The difference between $\hat{\psi}_{i1}(t)$ and $\psi_{i1}(t)$ can be expressed as

$$\hat{\psi}_{i1}(t) - \psi_{i1}(t) = A_{i1} + B_{i1}$$

where

$$A_{i1} = t\,Q(t/2)\left\{\frac{[\frac{t}{2} - I(z < \hat{\epsilon}_{t/2})]}{\hat{\epsilon}_{0.5}} - \frac{[\frac{t}{2} - I(z < \epsilon_{t/2})]}{\epsilon_{0.5}}\right\}$$

$$B_{i1} = t\,Q(0.5)\left\{\frac{\epsilon_{t/2}[0.5 - I(z < \epsilon_{0.5})]}{\epsilon^2_{0.5}} - \frac{\hat{\epsilon}_{t/2}[0.5 - I(z < \hat{\epsilon}_{0.5})]}{\hat{\epsilon}^2_{0.5}}\right\}$$

using the inequality $(a+b)^2 \leq 2\,(a^2+b^2)$

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\psi}_{i1}(t) - \psi_{ij}(t)\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[A_{i1} + B_{i1}\right]^2 \leq \frac{2}{n}\sum_{i=1}^{n}A^2_{i1} + \frac{2}{n}\sum_{i=1}^{n}B^2_{i1} \quad (50)$$

$$\tfrac{2}{n}\,\Sigma_{i=1}^{n}\,A^2_{i1} = \tfrac{2}{n}\,\Sigma_{i=1}^{n}\left\{t^2\,Q^2(\tfrac{t}{2})\left(\frac{\left[\frac{t}{2} - I(z < \hat{\epsilon}_{t/2})\right]^2}{\hat{\epsilon}^2_{0.5}} + \frac{\left[\frac{t}{2} - I(z < \epsilon_{t/2})\right]^2}{\epsilon^2_{0.5}} - 2\frac{\left[\frac{t}{2} - I(z < \hat{\epsilon}_{t/2})\right]\left[\frac{t}{2} - I(z < \epsilon_{t/2})\right]}{\hat{\epsilon}_{0.5}\,\epsilon_{0.5}}\right)\right\}$$

rearranging terms,

$$\frac{tI(z < \hat{\epsilon}_{t/2})}{\hat{\epsilon}_{0.5}\epsilon_{0.5}} - \frac{tI(z < \hat{\epsilon}_{t/2})}{\hat{\epsilon}^2_{0.5}} = \frac{tI(z < \hat{\epsilon}_{t/2})}{\hat{\epsilon}^2_{0.5}\,\hat{\epsilon}_{0.5}}[\hat{\epsilon}_{0.5} - \epsilon_{0.5}]$$

$$\frac{tI(z < \epsilon_{t/2})}{\hat{\epsilon}_{0.5}\epsilon_{0.5}} - \frac{tI(z < \epsilon_{t/2})}{\epsilon^2_{0.5}} = \frac{-tI(z < \epsilon_{t/2})}{\epsilon^2_{0.5}\,\hat{\epsilon}_{0.5}}[\hat{\epsilon}_{0.5} - \epsilon_{0.5}]$$

$$\frac{I^2(z < \hat{\epsilon}_{t/2})}{\hat{\epsilon}^2_{0.5}} + \frac{I^2(z < \epsilon_{t/2})}{\epsilon^2_{0.5}} - 2\frac{I(z < \epsilon_{t/2})I(z < \hat{\epsilon}_{t/2})}{\epsilon_{0.5}\,\hat{\epsilon}_{0.5}} = \left[\frac{I(z < \hat{\epsilon}_{t/2})}{\hat{\epsilon}_{0.5}} - \frac{I(z < \epsilon_{t/2})}{\epsilon_{0.5}}\right]^2$$

$$\frac{t^2/4}{\hat{\epsilon}_{0.5}^2}+\frac{t^2/4}{\epsilon_{0.5}^2}-\frac{t^2/2}{\hat{\epsilon}_{0.5}\,\epsilon_{0.5}}\;=\;\frac{t^2}{4}\left[\frac{1}{\hat{\epsilon}_{0.5}}-\frac{1}{\epsilon_{0.5}}\right]^2$$

By the strong consistency of the sample quantiles,

$$\left|\hat{\epsilon}_{0.5}-\epsilon_{0.5}\right|\xrightarrow{p}0\quad,\quad\left|\frac{1}{\hat{\epsilon}_{0.5}}-\frac{1}{\epsilon_{0.5}}\right|\xrightarrow{p}0\quad,\quad\left|\frac{I(z<\hat{\epsilon}_{t/2})}{\hat{\epsilon}_{0.5}}-\frac{I(z<\epsilon_{t/2})}{\epsilon_{0.5}}\right|\xrightarrow{p}0$$

and it implies that,

$$\frac{2}{n}\sum_{i=1}^{n}A_{i1}^2=o_p(1)\tag{51}$$

$$\begin{aligned}\tfrac{2}{n}\textstyle\sum_{i=1}^{n}B_{i1}^2=\tfrac{2}{n}\textstyle\sum_{i=1}^{n}\Bigg\{&t^2\,Q^2(0.5)\Bigg(\frac{\hat{\epsilon}_{t/2}^2\left[0.5^2-I(z<\hat{\epsilon}_{0.5})+I^2(z<\hat{\epsilon}_{0.5})\right]}{\hat{\epsilon}_{0.5}^4}\\&+\frac{\epsilon_{t/2}^2\left[0.5^2-I(z<\epsilon_{0.5})+I^2(z<\epsilon_{0.5})\right]}{\epsilon_{0.5}^4}\\&+\frac{2\hat{\epsilon}_{t/2}\epsilon_{t/2}\left[0.5^2-0.5I(z<\hat{\epsilon}_{0.5})-0.5I(z<\epsilon_{0.5})+I(z<\hat{\epsilon}_{0.5})\,I(z<\epsilon_{0.5})\right]}{\hat{\epsilon}_{0.5}^2\,{\epsilon_{0.5}}^2}\Bigg)\Bigg\}\end{aligned}$$

rearranging similar terms,

$$\frac{\hat{\epsilon}_{t/2}\,\epsilon_{t/2}\,I(z<\hat{\epsilon}_{0.5})}{\hat{\epsilon}_{0.5}^2\,{\epsilon_{0.5}}^2}-\frac{\hat{\epsilon}_{t/2}^2\,I(z<\hat{\epsilon}_{0.5})}{\hat{\epsilon}_{0.5}^4}\;=\;\frac{-\hat{\epsilon}_{t/2}\,I(z<\hat{\epsilon}_{0.5})}{\hat{\epsilon}_{0.5}^2}\left[\frac{\hat{\epsilon}_{t/2}}{\hat{\epsilon}_{0.5}^2}-\frac{\epsilon_{t/2}}{{\epsilon_{0.5}}^2}\right]$$

$$\frac{\hat{\epsilon}_{t/2}\,\epsilon_{t/2}\,I(z<\epsilon_{0.5})}{\hat{\epsilon}_{0.5}^2\,{\epsilon_{0.5}}^2}-\frac{\epsilon_{t/2}^2\,I(z<\epsilon_{0.5})}{{\epsilon_{0.5}}^4}\;=\;\frac{\epsilon_{t/2}\,I(z<\epsilon_{0.5})}{{\epsilon_{0.5}}^2}\left[\frac{\hat{\epsilon}_{t/2}}{\hat{\epsilon}_{0.5}^2}-\frac{\epsilon_{t/2}}{{\epsilon_{0.5}}^2}\right]$$

$$\begin{aligned}&\frac{\hat{\epsilon}_{t/2}^2\,I^2(z<\hat{\epsilon}_{0.5})}{\hat{\epsilon}_{0.5}^4}+\frac{\epsilon_{t/2}^2\,I^2(z<\epsilon_{0.5})}{{\epsilon_{0.5}}^4}-2\frac{\hat{\epsilon}_{t/2}\,\epsilon_{t/2}\,I(z<\hat{\epsilon}_{0.5})\,I(z<\epsilon_{0.5})}{\hat{\epsilon}_{0.5}^2\,{\epsilon_{0.5}}^2}\\&=\left[\frac{\hat{\epsilon}_{t/2}\,I(z<\hat{\epsilon}_{0.5})}{\hat{\epsilon}_{0.5}^2}-\frac{\epsilon_{t/2}\,I(z<\epsilon_{0.5})}{{\epsilon_{0.5}}^2}\right]^2\end{aligned}$$

$$\frac{\hat{\epsilon}_{t/2}^2}{\hat{\epsilon}_{0.5}^4} + \frac{\epsilon_{t/2}^2}{{\epsilon_{0.5}}^4} - 2\frac{\hat{\epsilon}_{t/2}\,\epsilon_{t/2}}{\hat{\epsilon}_{0.5}^2\,{\epsilon_{0.5}}^2} = \left[\frac{\hat{\epsilon}_{t/2}}{\hat{\epsilon}_{0.5}^2} - \frac{\epsilon_{t/2}}{{\epsilon_{0.5}}^2}\right]^2$$

According to the property of sample quantile consistency, it is obvious that;

$$\left[\frac{\hat{\epsilon}_{t/2}}{\hat{\epsilon}_{0.5}^2} - \frac{\epsilon_{t/2}}{{\epsilon_{0.5}}^2}\right] \xrightarrow{p} 0 \quad \text{and} \quad \left[\frac{\hat{\epsilon}_{t/2}\,I(z<\hat{\epsilon}_{0.5})}{\hat{\epsilon}_{0.5}^2} - \frac{\epsilon_{t/2}\,I(z<\epsilon_{0.5})}{{\epsilon_{0.5}}^2}\right] \xrightarrow{p} 0$$

hence,

$$\frac{2}{n}\sum_{i=1}^{n} B_{i1}^2 = o_p(1) \tag{52}$$

from (A.1),(A.2),(A.3) and (A.4) :- $\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\psi}_{i1}(t) - \psi_{i1}(t)\right]^2 = o_p(1).$ □

Furthermore, a similar procedure is applied to prove Lemma 1 for the other two robust versions when $j = 2, 3$.

Proof of Lemma 2(I) :

As $\psi_{ij}(t) = IF(z_i; \eta_j(t), F_n)$ are independent and identically distributed random variables with zero mean and finite variance (say σ^2), standard order statistics results imply that the maximum absolute value of $\psi_{ij}(t)$ grows at most at the rate of $o_p(\sqrt{n})$.

$$i.e. : \max_i |\psi_{ij}(t)| = o_p\left(\sqrt{n}\right)$$

This follows a well-known result in probability theory that the maximum of n independent and identically distributed random variables with finite variance behaves like $o_p\left(\sqrt{n}\right)$.

$$\max_i \left|\hat{\psi}_{ij}(t)\right| \leq \max_i \left|\hat{\psi}_{ij}(t) - \psi_{ij}(t)\right| + \max_i |\psi_{ij}(t)| \tag{53}$$

$$\max_i \left|\hat{\psi}_{ij}(t) - \psi_{ij}(t)\right| = o_p(1)$$

which is controlled by Lemma 1. Hence the proof that,

$$\max_i |\psi_{ij}(t)| = o_p\left(\sqrt{n}\right).$$

□

Proof of Lemma 2(II) :

By decomposing the term $\hat{\psi}_{ij}(t)$,

$$\hat{\psi}_{ij}(t) = \psi_{ij}(t) + \left(\hat{\psi}_{ij}(t) - \psi_{ij}(t)\right)$$

and expanding square,

$$\hat{\psi}_{ij}(t)^2 = \psi_{ij}^2(t) + 2\psi_{ij}\left(\hat{\psi}_{ij}(t) - \psi_{ij}(t)\right) + \left(\hat{\psi}_{ij}(t) - \psi_{ij}(t)\right)^2$$

the term $\frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_{ij}^2(t)$ can be expressed as,

$$\implies \frac{1}{n}\sum_{i=1}^{n}\psi_{ij}^2(t) + \frac{2}{n}\sum_{i=1}^{n}\psi_{ij}\left(\hat{\psi}_{ij}(t) - \psi_{ij}(t)\right) + \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\psi}_{ij}(t) - \psi_{ij}(t)\right)^2$$

since $\mathrm{Var}[\psi_{ij}(t)] = \mathbb{E}[\psi_{ij}^2(t)] - (\mathbb{E}[\psi_{ij}(t)])^2 = \sigma^2$ and $\mathbb{E}[\psi_{ij}(t)] = 0$,

$$\mathbb{E}[\psi_{ij}^2(t)] = \frac{1}{n}\sum_{i=1}^{n}\psi_{ij}^2(t) = \sigma^2$$

similar to the proof of Lemma 1 it could be obtained that,

$$\frac{2}{n}\sum_{i=1}^{n}\psi_{ij}\left(\hat{\psi}_{ij}(t) - \psi_{ij}(t)\right) = o_p(1)$$

and clearly from Lemma 1,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\psi}_{ij}(t) - \psi_{ij}(t)\right)^2 = o_p(1)$$

.

hence the proof;

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_{ij}(t)^2 = \sigma^2 + o_p(1).$$

□

Proof of Lemma 2(III) :

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\psi}_{ij}(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\hat{\psi}_{ij}(t) - \psi_{ij}(t)\right) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{ij}(t)$$

Since $\psi_{ij}(t) = IF(z_i; \eta_j(t), F_n)$ are independent and identically distributed random variables with zero mean and finite variance σ^2, by the central limit theorem;

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{ij}(t)\xrightarrow{L}\mathcal{N}\left(0,\sigma^2\right)$$

and similar to the proof of Lemma 1, it is easy to show that;

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\hat{\psi}_{ij}(t)-\psi_{ij}(t)\right)=o_p(1)$$

combining the results obtained,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\psi}_{ij}(t)\xrightarrow{L}\mathcal{N}\left(0,\sigma^2\right).$$

□

Proof of Theorem 1 :

From [Athur(1990)], the important result is that;

$$|v_j(t)|=O_p(n^{-1/2})$$

It arises from empirical likelihood constraints and plays a key role in approximating the empirical log-likelihood ratio function. Expanding the empirical log-likelihood ratio function using the Taylor series,

$$\begin{aligned}\hat{l}(\eta_j(t))&=-2\cdot\ln\{\hat{R}(\eta_j(t)\}=2\cdot\sum_{i=1}^{n}\ln\{1+v_j(t)\,\hat{\psi}_{ij}(t)\}\\&=2\cdot\sum_{i=1}^{n}\left[v_j(t)\,\hat{\psi}_{ij}(t)-\frac{1}{2}\left(v_j(t)\,\hat{\psi}_{ij}(t)\right)^2\right]+K_{ij}\end{aligned}$$

here $K_{ij}=O\left(\left[v_j(t)\,\hat{\psi}_{ij}(t)\right]^3\right)$ represents the higher order remainder term, and the bounding of the term could be obtained as follows.

$$|K_{ij}|\leq r\sum_{i=1}^{n}\left|v_j(t)\,\hat{\psi}_{ij}(t)\right|^3\leq r\,|v_j(t)|^3\max_i\left|\hat{\psi}_{ij}(t)\right|\sum_{i=1}^{n}\hat{\psi}_{ij}^2(t)$$

where r is an arbitrary constant term. From Lemma 2,

$$\max_i\left|\hat{\psi}_{ij}(t)\right|=o_p\left(\sqrt{n}\right)$$

squaring and summing over n,

$$\implies \sum_{i=1}^{n} \hat{\psi}_{ij}^2(t) = O_p(n)$$

$$|K_{ij}| \leq r \cdot o_p(n^{-3/2}) \cdot o_p(n^{1/2}) \cdot o_p(n) = o_p(1)$$

hence,

$$\hat{l}(\eta_j(t)) = 2 \cdot \sum_{i=1}^{n} \left[v_j(t)\, \hat{\psi}_{ij}(t) - \frac{1}{2} \left(v_j(t)\, \hat{\psi}_{ij}(t) \right)^2 \right] + o_p(1) \tag{54}$$

Then by equation,

$$\frac{1}{n} \left\{ \sum_{j=1}^{n} \frac{\hat{\psi}_{ij}(t)}{1 + v_j(t)\, \hat{\psi}_{ij}(t)} \right\} = 0$$

expanding denominator using Taylor series expansion,

$$\left(1 + v_j(t)\, \hat{\psi}_{ij}(t) \right)^{-1} \simeq 1 - v_j(t)\, \hat{\psi}_{ij}(t) + \frac{\left(v_j(t)\, \hat{\psi}_{ij}(t) \right)^2}{1 + v_j(t)\, \hat{\psi}_{ij}(t)}$$

therefore,

$$\sum_{j=1}^{n} \left[\hat{\psi}_{ij}(t) - v_j(t)\, \hat{\psi}_{ij}^2(t) + \frac{\left(v_j(t)^2\, \hat{\psi}_{ij}^3(t) \right)}{1 + v_j(t)\, \hat{\psi}_{ij}(t)} \right] = 0$$

As in $|v_j(t)| = O_p(n^{-1/2}) \implies |v_j(t)|^2 = O_p(n^{-1})$, the last term is negligible and could be ignored asymptotically.

then solving for $v_j(t)$,

$$v_j(t) = \frac{\Sigma_{j=1}^{n}\, \hat{\psi}_{ij}(t)}{\Sigma_{j=1}^{n}\, \hat{\psi}_{ij}^2(t)} + o_p(n^{-1/2})$$

considering the term,

$$\sum_{j=1}^{n} v_j(t)\, \hat{\psi}_{ij}(t) = v_j(t) \sum_{j=1}^{n} \hat{\psi}_{ij}(t) = \frac{\left[\Sigma_{j=1}^{n}\, \hat{\psi}_{ij}(t) \right]^2}{\Sigma_{j=1}^{n}\, \hat{\psi}_{ij}^2(t)}$$

$$\implies \sum_{j=1}^{n} v_j(t)\, \hat{\psi}_{ij}(t) = \sum_{j=1}^{n} \left[v_j(t)\, \hat{\psi}_{ij}(t) \right]^2$$

$$\begin{aligned}\hat{l}(\eta_j(t)) &= 2 \cdot \sum_{j=1}^{n} v_j(t)\,\hat{\psi}_{ij}(t) - \sum_{j=1}^{n}\left[v_j(t)\,\hat{\psi}_{ij}(t)\right]^2 + o_p(1)\\ &= \sum_{j=1}^{n}\left[v_j(t)\,\hat{\psi}_{ij}(t)\right]^2 + o_p(1)\\ &= \frac{\left[\sum_{j=1}^{n}\hat{\psi}_{ij}(t)\right]^2}{\sum_{j=1}^{n}\hat{\psi}_{ij}^2(t)} + o_p(1)\\ &= \frac{\left[\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\hat{\psi}_{ij}(t)\right]^2}{\frac{1}{n}\sum_{j=1}^{n}\hat{\psi}_{ij}^2(t)} + o_p(1)\end{aligned}$$

According to Lemma 2 (II) and (III);

$$\frac{\left[\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\hat{\psi}_{ij}(t)\right]^2}{\frac{1}{n}\sum_{j=1}^{n}\hat{\psi}_{ij}^2(t)} \xrightarrow{L} \chi_1^2$$

therefore,

$$\hat{l}(\eta_j(t)) \quad \xrightarrow{L} \quad \chi_1^2.$$

□

References

[Kampke(2012)] Kämpke, T and Radermacher, FJ (2012) *Lorenz curves history, state and some recent results.*

[Athur(1988)] ART B. OWEN (1988) *Empirical likelihood ratio confidence intervals for a single functional, Biometrika*, Biometrika, vol. 75, no. 2, pp. 237–249.

[Athur(1990)] ART B. OWEN (1990) *Empirical likelihood ratio confidence regions*, The annals of statistics, vol. 18, no. 1, pp. 90–120.

[Lorenz(1905)] Lorenz, Max O (1905) *Methods of measuring the concentration of wealth*, Publications of the American statistical association, vol. 9, no. 70, pp. 209–219.

[Joshep(2014)] Joseph L Gastwirth (2014) *Median-based measures of inequality: Reassessing the increase in income inequality in the US and Sweden*, Statistical Journal of the IAOS, vol. 30, no. 4, pp. 311–320.

[Thomas(2010)] Thomas Kampke (2010) *The use of mean values vs. medians in inequality analysis*, Journal of economic and social measurement, vol. 35, no. 1, pp. 43.

[Prendergast(2016)] Prendergast, Luke A and Staudte, Robert G (2016) *Quantile versions of the Lorenz curve.*

[Cowell(2002)] Cowell, Frank A and Victoria-Feser, Maria-Pia (2002) *Welfare rankings in the presence of contaminated data*, Econometrica, vol. 70, no. 3, pp. 1221–1233.

[Gastwirth(1971)] Gastwirth, Joseph L (1971) *A general definition of the Lorenz curve*, Econometrica: Journal of the Econometric Society, pp. 1037–1039.

[Prisner(2009)] Prisner, Erich (2009) *Income distribution, Lorenz curve, and Gini index: a complex calculus project for the classroom*, Franklin College, Switzerland.

[Gastwirth(1972)] Gastwirth, Joseph L (1972) *The estimation of the Lorenz curve and Gini index*, The review of economics and statistics, pp. 306–316.

[Lubrano(2017)] Lubrano, Michel (2017) *The econometrics of inequality and poverty Chapter 4: Lorenz curves, the Gini coefficient and parametric distributions.*

[Qin(2013)] Qin, Gengsheng and Yang, Baoying and Belinga-Hall, Nelly E (2013) *Empirical likelihood-based inferences for the Lorenz curve*, Annals of the Institute of Statistical Mathematics, vol. 65, pp. 1–21.

[Belz(2019)] Belz, Enora (2019) *Estimating inequality measures from quantile data.*

[Hampel(1971)] Hampel, Frank R (1971) *A general qualitative definition of robustness*, The annals of mathematical statistics, vol. 42, no. 6, pp. 1887–1896.

[Yang(2012)] Yang, BY and Qin, GS and Belinga-Hill, Nelly E (2012) *Non-parametric inferences for the generalized Lorenz curve*, Sci Sin Math, vol. 42, no. 3, pp. 235–250.

[Kleiber(2008)] Kleiber, Christian (2008) *"The Lorenz curve in economics and econometrics"* in *Advances on income inequality and concentration measures* , pp. 247–264.

[liu(2023)] Liu, Pangpang and Zhao, Yichuan (2023) *A review of recent advances in empirical likelihood*, Wiley Interdisciplinary Reviews: Computational Statistics, vol. 15, no. 3, pp. e1599.

[Lomax(1954)] Lomax, Kenneth S (1954) *Business failures: Another example of the analysis of failure data*, Journal of the American statistical association, vol. 49, no. 268, pp. 847–852.

[Lai(2006)] Lai, Chin-Diew and Murthy, DN and Xie, Min (2006) *"Weibull distributions and their applications"* in *Springer Handbooks*, pp. 63–78.

[Torrent(1978)] Torrent, Roberto (1978) *The log-normal distribution: A better fitness for the results of mechanical testing of materials*, Matériaux et Constructions, vol. 11, pp. 235–245.

[kleiber(2008)] Kleiber, Christian (2008) *"A guide to the Dagum distributions"* in *Modeling income distributions and Lorenz curves*, pp. 97–117.

[shi(2020)] Shi, Yuyin and Liu, Bing and Qin, Gengsheng (2020) *Influence function-based empirical likelihood and generalized confidence intervals for the Lorenz curve*, Statistical Methods & Applications, vol. 29, pp. 427–446.

[ratnasingam(2024)] Ratnasingam, Suthakaran and Wallace, Spencer and Amani, Imran and Romero, Jade (2024) *Nonparametric confidence intervals for generalized Lorenz curve using modified empirical likelihood*, Computational Statistics, vol. 39, no. 6, pp. 3073–3090.

[vejdani(2023)] Vejdani, Mahdiyeh and Roknabadi, Abdolhamid Rezaei and Fakoor, Vahid and Jomhoori, Sarah (2023) *Empirical Likelihood Confidence Intervals for Lorenz Curve with Length-Biased Data*, Iranian Journal of Science, vol. 47, no. 5, pp. 1617–1631.