5-2-2018

Influence Function-Based Empirical Likelihood And Generalized Confidence Intervals For Lorenz Curve

Yuyin Shi

INFLUENCE FUNCTION-BASED EMPIRICAL LIKELIHOOD AND GENERALIZED CONFIDENCE INTERVALS FOR LORENZ CURVE

by

YUYIN SHI

Under the Direction of Gengsheng Qin, PhD

ABSTRACT

This thesis aims to solve confidence interval estimation problems for Lorenz curve. First, we propose new nonparametric confidence intervals with influence function-based empirical likelihood method. It is shown that the limiting distributions of log-empirical likelihood ratios are standard Chi-square distributions. Then the "exact" parametric intervals based on generalized pivotal quantities for Lorenz ordinates are also developed. Extensive simulation studies are conducted to evaluate the finite sample performance of the proposed methods. Finally, our methods are applied on real income data sets.

INDEX WORDS: Empirical likelihood; Influence function; Generalized pivotal quantities; Lorenz ordinates.

INFLUENCE FUNCTION-BASED EMPIRICAL LIKELIHOOD AND GENERALIZED CONFIDENCE INTERVALS FOR LORENZ CURVE

by

YUYIN SHI

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2018

INFLUENCE FUNCTION-BASED EMPIRICAL LIKELIHOOD AND GENERALIZED CONFIDENCE INTERVALS FOR LORENZ CURVE

by

YUYIN SHI

Committee Chair: Gengsheng Qin

Committee: Qi Xin

Xiaoyi Min

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2018

DEDICATION

This thesis is dedicated to Georgia State University.

ACKNOWLEDGEMENTS

I would like to give my deep thanks and sincere gratitude to everyone for helping me complete this thesis.

First, I want to show my gratitude to my academic advisor, Dr. Gengsheng Qin, who has supported me throughout my thesis with his patience and knowledge. Dr. Qin gave me lots of suggestions and guidance while I was doing this thesis research, from the most basic language modification to the mathematical proof. His insight and knowledge always inspired me and I truly learned a lot from him.

I'd like to give my special thanks to my committee members, Dr. Xin Qi and Dr. Xiaoyi Min, for reviewing my thesis and their insightful comments and encouragement.

My gratitude also goes to Dr. Guantao Chen and Dr. Yichuan Zhao for their letters of recommendation to the Ph.D. programs I applied. To Dr. Jing Zhang, who gave wonderful lectures on different topics on statistics and biostatistics.

Special thanks to all the faculty, staff and students in the Department of Mathematics and Statistics at Georgia State University.

Last but not least, I would like to thank my parents, my uncle, my aunt, my boyfriend, my cousins and all my friends who helped me when I faced difficulties.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

- ACI - Asymptotic Confidence Interval
- BCI - Bootstrap Confidence Interval
- CDF - Cumulative Distribution Function
- CI - Confidence Interval
- CvM - Cramer-von Mises
- EL - Empirical Likelihood
- GCI - Generalized Confidence Interval
- GPQ - Generalized Pivotal Quantity
- GSU - Georgia State University
- IFEL - Influence Function-based Empirical Likelihood
- IFJEL - Influence Function-based Jackknife Empirical Likelihood
- i.i.d. - independent identically distributed
- JEL - independent identically distributed
- KS - Kolmogorov-Smirnov
- MLE - Maximum Likelihood Estimator
- NA - Normal Approximation
- PEL - Profile Empirical Likelihood
- PSID - Panel Study of Income Dynamics

CHAPTER 1

INTRODUCTION

Much attention has been given to the rising income polarization in the U.S. and across the world, thus the estimation accuracy of the increasing income inequality is crucially important when government makes economic policy decisions. One widely used tool to measure the income distribution and income inequality is the Lorenz curve, which shows the percentage of the total income that the bottom $(100 * t)\%$ $(t \in [0, 1])$ of households have. The Lorenz curve is illustrated in Figure 1, the line at the 45° angle shows perfect equality of income through all the households, while the Lorenz curve describes the inequality. The further away the Lorenz curve is from the diagonal, the more unequal is the income distribution.

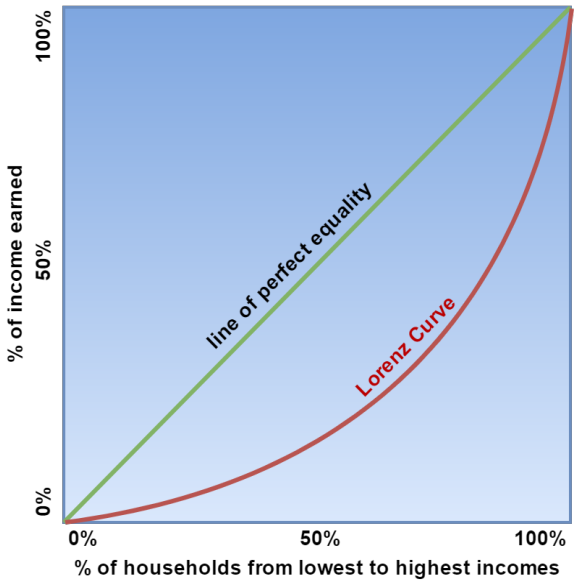

Figure 1.1. Lorenz Curve

Let X be a non-negative random variable with a cumulative distribution function $F(x)$,

i.e., $F(x)$ represents the proportion of the population whose income is less than or equal to x. Assuming that $F(x)$ is differentiable, Gastwirth (1971)[14] provided a definition of Lorenz curve as below:

$$\eta(t) = \frac{1}{\mu}\int_0^{\xi_t} x dF(x), \quad t \in [0,1], \tag{1.1}$$

where $\mu = \int_0^\infty x dF(x)$ is the mean of F, and $\xi_t = F^{-1}(t) = inf\{x : F(x) \geq t\}$ is the t-th quantile of F. For a fixed $t \in [0,1]$, the Lorenz ordinate $\eta(t)$ is the ratio of, the mean income of the lowest t-th fraction of households, and, the mean income of total households.

Lorenz curve has been primarily utilized in economics and social sciences. Atkinson (1970) [1] provided a theorem related to the social welfare function and the Lorenz curve, Doiron (1996) [9] used Lorenz dominance to analyze income and earning inequality. Besides economics, Lorenz curve is also widely used in other disciplines including medical and health research (Chang and Halfon 1997 [4]), industrial concentration (Smith 1947 [25]) and reliability (Gail and Gastwirth 1978 [13]).

Since income distribution $F(x)$ is rarely known in practice, the Lorenz curve has to be estimated from the income data. Let $X_1, X_2, \ldots, X_n$ be a simple random sample drawn from the population X, the empirical estimate for $\eta(t)$ is defined as

$$\hat{\eta}(t) = \frac{1}{\hat{\mu}}\int_0^{\hat{\xi}_t} x d\hat{F}_n(x), \quad t \in [0,1], \tag{1.2}$$

where $\hat{F}_n(x)$ is the empirical distribution function of $X_1, X_2, \ldots, X_n$, $\hat{\mu}$ is the sample mean, and $\hat{\xi}_t$ is the t-th sample quantile. One approach to make inferences on Lorenz curve is the normal approximation (NA) method (Beach and Davidson 1983 [2], Beach and Richmond 1985 [3], Csörgö and Zitikis 1996b [7]). However, the NA-based confidence intervals may have poor performances due to the skewness of the real income data. Bootstrap, introduced by Efron (1981[10], 1982a[11]), is another powerful statistical method to construct confidence intervals (Diciccio and Efron 1996[8]). There are still some limitations: 1) bootstrap method can be time consuming; 2) the sampling method used in generating bootstrap sample would also contribute to the sampling bias, according to Haukoos and Lewis (2005) [16].

Empirical likelihood (EL), introduced by Owen (1988) [20], has been shown to have diverse advantages over the normal approximation and bootstrap method (Hall and La Scala 1990 [15]). For example, we can use EL method to construct a confidence interval without choosing an underlying distribution; the EL-based method is also able to construct confidence interval without variance estimation. As mentioned by Wood et al.(1996)[30], the EL ratio statistic, under mild conditions, converges in distribution to a chi-square distribution. Thus, EL-based method may have advantages in developing statistical inferences with skewed data. EL has been widely used in many fields, such as survey sampling (Chen and Qin 1993 [5]), health care (Zhou et al. 2006[31]) and medical diagnostics (Qin and Zhou 2006 [23]). Recently, Qin, Yang and Belinga (2013) [22] developed a plug-in EL method to make inferences for the Lorenz curve. However, the limiting distribution of their EL ratio statistic is a scaled chi-square distribution, which requires heavy computation of the scale constant. Moreover the performances of their EL intervals are not stable due to the plug-in estimate of the scale constant. In order to alleviate the computational burden and obtain more consistent confidence intervals, we propose a new influence function-based empirical likelihood (IFEL) method to make inferences for the Lorenz curve. At the same time, the influence function-based Jackknife empirical likelihood (IFJEL) is implemented to be compared with IFEL method. Jackknife empirical likelihood (JEL) method was proposed by Jing, Yuan and Zhou (2009)[17] and the general idea of the JEL is to construct a jackknife pseudo-sample which is assumed to be asymptotically independent. Through simulation studies, we found that the proposed IFEL and IFJEL have good coverage probabilities in most cases except for those when t is large. Thus, it motivated us to propose the "exact" parametric intervals based on generalized pivotal quantities (GPQ), which has good performances in all the cases. In the GPQ method, we introduce two distributions which are the most commonly used parametric distributions for modeling income data: the Pareto distribution and the Lognormal distribution. The Pareto law for income distributions was developed and had been verified to hold universally by Pareto in 1897. The Pareto distribution fits the data fairly well toward the right tail of income and wealth distributions[29]. If we consider the

entire range of income, there is evidence that the income of $97\% - 99\%$ of the population is distributed log-normally in economics studies[6]. Therefore, the fit may be better from the Lognormal distribution. Our GPQ-based approach is built on the concepts of generalized inferential procedures (Shafiei, Saboori and Doostparast, 2016 [24]) when the underlying income distribution is a Pareto distribution or a Lognormal distribution.

The rest of the thesis is organized as follows. In Chapter 2, we review the profile empirical likelihood (PEL) for Lorenz ordinate and propose the influence function-based empirical likelihood (IFEL) for Lorenz curve. Moreover, we implement the influence function-based Jackknife empirical likelihood (IFJEL) method. In Chapter 3, we briefly explain the methodology of generalized confidence interval (GCI) and the construction of confidence intervals for the Lorenz curve under Pareto and Lognormal distributions. In Chapter 4, we present various confidence intervals including normal approximation-based intervals and bootstrap-based intervals. Simulation studies are conducted to compare the performances of these proposed intervals. In Chapter 5, we use the income data from the Panel Study of Income Dynamics and median income data of twenty occupations in the U.S. in 1950 to illustrate the proposed intervals. In Chapter 6, there are some discussions and conclusions. The proof of the main theorem for the Lorenz curve is given in the Appendix.

CHAPTER 2

EMPIRICAL LIKELIHOOD-BASED METHODS

2.1 Influence Function-based Empirical Likelihood for the Lorenz Curve

Consider $\{X_1, X_2, \cdots, X_n\}$ as a simple random sample from the population of X with c.d.f. F. For a fixed $t \in (0,1)$, the Lorenz ordinate $\eta(t)$ must satisfies $E[X(I(X \leq \xi_t) - \eta(t))] = 0$. Thus, we can define the empirical likelihood for $\eta(t)$ as follows:

$$\tilde{L}_1(\eta(t)) = \sup_{\mathbf{p}} \left\{ \prod_{i=1}^{n} p_i \ : \quad \sum_{i=1}^{n} p_i = 1, \ \sum_{i=1}^{n} p_i D_i(t) = 0 \right\}, \tag{2.1}$$

where $\mathbf{p} = (p_1, p_2, \cdots, p_n)$ is a probability vector, $D_i(t) = X_i[I(X_i \leq \xi_t) - \eta(t)]$, $i = 1, 2, \cdots, n$. Since the population quantile is unknown, we use the $[nt]$-th ordered value of X_i's to represent ξ_t, let's say $\hat{\xi}_t = X_{[nt]}$, we get the profile empirical likelihood (PEL) for $\eta(t)$:

$$L_1(\eta(t)) = \sup_{\mathbf{p}} \left\{ \prod_{i=1}^{n} p_i \ : \quad \sum_{i=1}^{n} p_i = 1, \ \sum_{i=1}^{n} p_i \hat{D}_i(t) = 0 \right\}, \tag{2.2}$$

where $\hat{D}_i(t) = X_i[I(X_i \leq \hat{\xi}_t) - \eta(t)], \ i = 1, \cdots, n$.

A unique maximum for $\mathbf{p}$ in (2.2) exists if $\eta(t)$ is inside the convex hull of $\{X_1[I(X_1 \leq \hat{\xi}_t) - \eta(t)], \ ..., \ X_n[I(X_n \leq \hat{\xi}_t) - \eta(t)]\}$. By Lagrange multiplier method, the supremum occurs at $p_i = \frac{1}{n}\left\{1 + \nu(t)\hat{D}_i(t)\right\}^{-1}, i = 1, \cdots, n$, where $\nu(t)$ is the solution to

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{D}_i(t)}{1 + \nu(t)\hat{D}_i(t)} = 0. \tag{2.3}$$

Note that $\prod_{i=1}^{n} p_i$, subject to $\sum_{i=1}^{n} p_i = 1, p_i \geq 0, i = 1, 2, ..., n$, attains its maximum

n^{-n} at $p_i = n^{-1}$. So, the profile empirical likelihood ratio for $\eta(t)$ can be defined as

$$R_1(\eta(t)) = \prod_{i=1}^{n}(np_i) = \prod_{i=1}^{n}\{1 + \nu(t)\hat{D}_i(t)\}^{-1} .$$

The corresponding profile empirical log-likelihood ratio for $\eta(t)$ is:

$$l_1(\eta(t)) = -2\log R_1(\eta(t)) = 2\sum_{i=1}^{n}\log\{1 + \nu(t)\hat{D}_i(t)\} . \tag{2.4}$$

The EL interval for $\eta(t)$ is:

$$\{\eta(t) : rl_1(\eta(t)) \leq \chi^2_{1,1-\alpha}\}, \tag{2.5}$$

where r is the scale constant and $r = s_p^2(t)/s_d^2(t)$ with $s_p^2(t) = \int_0^\infty \{x[I(x \leq \xi_t) - \eta(t)]\}^2 dF(x)$, $s_d^2(t) = \int_0^\infty [(x - \xi_t)I(x \leq \xi_t) - x\eta(t)]^2 dF(x) - (t\xi_t)^2$.

Based on the theories in Qin et al. (2013) [22], we can derive:

$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{D}_i(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i[I(X \leq \hat{\xi}_t) - \eta(t)])$

$= \sqrt{n}\{\frac{1}{n}\sum_{i=1}^{n}[(X_i - \xi_t)I(X_i \leq \xi_t) + t_0\xi_t - X_i\eta(t)]\} + o_p(1)$

$= \sqrt{n}\{\frac{1}{n}\sum_{i=1}^{n} g(X_i, \eta(t))\} + o_p(1)$

where $g(X_i, \eta(t))$ is called the influence function.

Next, we define the empirical likelihood based on influence function for $\eta(t)$ as:

$$L_{IF}(\eta(t)) = \sup_{\boldsymbol{p}}\left\{\prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i = 1, \sum_{i-1}^{n} p_i\hat{g}(X_i, \eta(t)) = 0\right\}. \tag{2.6}$$

where $\hat{g}(X_i, \eta(t)) = (X_i - \hat{\xi}_t)I(X_i \leq \hat{\xi}_t) + t\hat{\xi}_t - X_i\eta(t)$.

And the EL ratio based on the estimated influence function can be defined as:

$$R_{IF}(\eta(t)) = \prod_{i=1}^{n}(np_i) = \prod_{i=1}^{n}\{1 + \nu_{IF}(t)\hat{g}(X_i, \eta(t))\}^{-1}, \tag{2.7}$$

where ν_{IF} is the solution to:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{g}(X_i,\eta(t))}{1+\nu_{IF}(t)\hat{g}(X_i,\eta(t))}=0. \tag{2.8}$$

The corresponding influence function-based empirical log-likelihood ratio for $\eta(t)$ is:

$$l_{IF}(\eta(t))=-2logR_{IF}(\eta(t))=2\sum_{i=1}^{n}log\{1+\nu_{IF}(t)\hat{g}(X_i,\eta(t))\}, \tag{2.9}$$

Then the following result gives the limiting distribution of the empirical likelihood based on influence function.

Theorem 1. *If $E(X^2)<\infty$ and $\eta(t_0)=E[XI(X\leq\xi_{t_0})]/E(X)$ for a given $t=t_0\in(0,1)$, then the limiting distribution of $l_{IF}(\eta(t_0))$ is a standard chi-square distribution, i.e., $l_{IF}(\eta(t_0))\to\chi^2_1$ as $n\to\infty$.*

This theorem makes it much easier to construct confidence intervals for the Lorenz ordinates compared with the PEL method. Based on Theorem 1, a $(1-\alpha)$ level asymptotic confidence interval for Lorenz ordinate $\eta(t)$ can be constructed as $R_{IF}=\{\eta(t):l_{IF}(\eta(t))\leq\chi^2_{1,1-\alpha}\}$.

2.2 Influence Function-based Jackknife Empirical Likelihood for the Lorenz Curve

Furthermore, we implement the influence function-based Jackknife empirical likelihood (IFJEL) method. Recall the influence function $g(X_i,\eta(t_0))$ has been defined in Section 2.1. Let

$$\hat{V}_{-i}=\frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\hat{g}_{-i}(X_j,\eta(t_0)),i=1,2,\cdots,n,j=1,2,\cdots,n. \tag{2.10}$$

where $\hat{g}_{-i}(X_j,\eta(t_0))=(X_j-\hat{\xi}_{t,-i})I(X_j\leq\hat{\xi}_{t,-i})+t\hat{\xi}_{t,-i}-X_j\eta(t)$, $\hat{\xi}_{t,-i}$ is the $\hat{\xi}_t$ based on $\{X_1,\cdots,X_{i-1},X_{i+1},\cdots,X_n\}$.

$$\hat{W}_i = \sum_{j=1}^{n} \hat{g}(X_j, \eta(t_0)) - (n-1)\hat{V}_{-i}, i = 1, 2, \cdots, n, j = 1, 2, \cdots, n. \tag{2.11}$$

Here we use $\hat{W}_i$ to substitute $\hat{g}(X_i, \eta(t_0))$ in (2.6) (2.7), (2.8) and (2.9), which gives the log influence function-based jackknife empirical likelihood ratio as

$$l_{IFJEL}(\eta(t_0)) = -2logR_{IFJEL}(\eta(t_0)) = 2\sum_{i=1}^{n} log\{1 + \nu_{IFJEL}(t)\hat{W}_i\}, \tag{2.12}$$

$l_{IFJEL}(\eta(t_0))$ can be proved to follow a standard chi-square distribution based on Jing, Yuan and Zhou's paper (2009)[17]. Similar to the procedure in Section 2.1, we can construct the $(1-\alpha)\%$ IFJEL confidence interval by $R_{IFJEL} = \{\eta(t) : l_{IFJEL}(\eta(t)) \leq \chi^2_{1,1-\alpha}\}$.

CHAPTER 3

GENERALIZED INFERENTIAL PROCEDURES FOR LORENZ CURVE UNDER PARETO AND LOGNORMAL DISTRIBUTIONS

Tsui and Weerahandi (1989)[26] introduced the concept of generalized p values and generalized test variables, which are useful for developing hypothesis tests in situations involving nuisance parameters. Subsequently, the concept of a generalized pivotal quantity (GPQ) was introduced by Weerahandi (1993) [28]. In order to define a GPQ, let $\mathbf{X}$ be a random sample whose distribution depends on a parameter of interest θ, and a nuisance parameter δ. Let x denote the observed value of $\mathbf{X}$, then $Q(\mathbf{X};\mathrm{x},\theta,\delta)$ is called a generalized pivotal quantity (GPQ) for θ, given the following two conditions:

(i) Given the observed value x, the distribution of $Q(\mathbf{X};\mathbf{x},\theta,\delta)$ is free of any unknown parameter;

(ii) The observed value of $Q(\mathbf{X};\mathbf{x},\theta,\delta)$ at $\mathbf{X}=\mathbf{x}$, *i.e.*, $Q(\mathbf{x};\mathbf{x},\theta,\delta)$, is equal to θ.

From this definition, the percentiles of $Q(\mathbf{X};\mathbf{x},\theta,\delta)$ are considered as the GCIs for θ. It's easy to see that $(Q_{\frac{\alpha}{2}}(\mathbf{X};\mathbf{x},\theta,\delta),\ Q_{1-\frac{\alpha}{2}}(\mathbf{X};\mathbf{x},\theta,\delta))$ is the equi-tail two-sided $100(1-\alpha)\%$ confidence interval for θ.

3.1 Lorenz Curve under Pareto Distribution

The Pareto distribution in the shape-scale form is defined as

$$F(x;\beta,\lambda)=P(X\leq x)=1-(\frac{\beta}{x})^{\frac{1}{\lambda}},x\geq\beta, \tag{3.1}$$

where λ and β are the shape and the scale parameters, respectively. Moreover, Malik (1970) [18] derived distributions of the maximum likelihood estimators (MLEs) of the parameters in the Pareto distribution based on a random sample of size n. Let $\mathbf{X}=\{X_1,\cdots,X_n\}$ be a random sample from the Pareto distribution with the CDF (3.1). Then the MLEs of β and

λ are given by

$$\hat{\beta} = \hat{\beta}(\mathbf{X}) = X_{(1)} \qquad and \qquad \hat{\lambda} = \hat{\lambda}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} ln(\frac{X_i}{X_{(1)}}), \tag{3.2}$$

where $X_{(1)}$ denotes the first order statistic among $X_1, \cdots, X_n$. Let

$$Z_1 = \frac{1}{\lambda} ln(\frac{\hat{\beta}}{\beta}) \qquad and \qquad Z_2 = \frac{\hat{\lambda}}{\lambda}. \tag{3.3}$$

Z_1 and Z_2 are independent random variables and $2nZ_1$ follows Chi-square distribution with 2 degrees of freedom, $2nZ_2$ follows Chi-square distribution with $2(n-1)$ degrees of freedom.

When the income distribution is a Pareto distribution, the Lorenz curve is

$$\eta(t;\lambda,\beta) = \begin{cases} 1-(1-t)^{1-\lambda} & \text{if } 0<\lambda<1, 0<t<1, \\ 1 & \text{if } 0<\lambda<1, t=1. \end{cases} \tag{3.4}$$

Let

$$Q_\lambda^* = \frac{\hat{\lambda}_0}{Z_2}, \tag{3.5}$$

where Z_2 is defined by (3.3) and $\hat{\lambda}_0$ denotes the observed value of $\hat{\lambda}$. It is easy to see that the distribution of Q_λ^* is free of the unknown parameter λ, and $2nZ_2 \sim \chi^2_{2(n-1)}$. Moreover, the observed value of Q_λ^* is λ. Therefore Q_λ^* is a GPQ for λ, and a GPQ for $\eta(t;\lambda,\beta)$ is

$$Q_\eta^{'}(\mathbf{X};\mathbf{x},t,\beta) = \begin{cases} 1-(1-t)^{1-Q_\lambda^*} & \text{if } 0<Q_\lambda^*<1, 0<t<1, \\ 1 & \text{if } 0<Q_\lambda^*<1, t=1. \end{cases} \tag{3.6}$$

Since the CDF of $Q_\eta^{'}(\mathbf{X};\mathbf{x},t,\beta)$ does not have a closed form, the following Monte Carlo method is needed to find GPQ-based CIs for $\eta(t;\lambda,\beta)$.

We summarize the procedure in the following steps:

For a given random sample $\{x_1, x_2, \ldots, x_n\}$ from a pareto distribution, compute the value of $\hat{\lambda}_0$.

1. Generate a random value of $2nZ_2$ from $\chi^2_{2(n-1)}$, compute $Q^*_\lambda = \frac{2n\hat{\lambda}_0}{2nZ_2}$, and then plug Q^*_λ in (3.6) to obtain $Q^{'}_\eta$.

2. Repeat step 1 for m (m is recommended to be 5000 or bigger) times, so we can get m copies of $Q^{'}_\eta$.

3. Sort the m copies of $Q^{'}_\eta$ in ascending order, and denote the ordered values as $Q^{'}_\eta[1]$, $Q^{'}_\eta[2]$,$\cdots$, $Q^{'}_\eta[m]$, where $[a]$ means the greatest integer less than or equal to a.

Therefore the $100(1-\alpha)\%$ generalized confidence interval (GCI) for $\eta(t;\lambda,\beta)$ is $[Q^{'}_\eta[m*\alpha/2], Q^{'}_\eta[m*(1-\alpha/2)]]$.

3.2 Lorenz Curve under Lognormal Distribution

A random variable X has a Lognormal distribution if its logarithm $log(X)$ has a normal distribution, i.e. $Y = log(X) \sim N(\mu, \sigma^2)$, and the MLE of μ and σ^2 is given by

$$\hat{\mu} = \bar{Y} = \frac{\sum_{i=1}^n Y_i}{n} \qquad and \qquad \hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}, \tag{3.7}$$

where $Y_i = log(X_i)$, $\hat{\mu}$ and $\hat{\sigma}^2$ are independent. And $U^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

When the underlying distribution is a Lognormal distribution, the Lorenz curve is

$$\eta(t;\mu,\sigma) = \phi(\phi^{-1}(t) - \sigma) \qquad if \qquad 0 < t < 1, \tag{3.8}$$

where ϕ denotes the probability density of the standard normal distribution.

Let

$$Q^*_{\sigma^2} = \frac{s^2(n-1)}{U^2}, \tag{3.9}$$

where s^2 denotes the observed value of $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$ and $U^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. The distribution of $Q^*_{\sigma^2}$ is free of unknown parameter σ^2 and the observed value of $Q^*_{\sigma^2}$ is σ^2. Therefore, $Q^*_{\sigma^2}$ is a GPQ for σ^2 and the GPQ for $\eta(t;\mu,\sigma)$ is

$$Q^*_\eta(\mathbf{X};\mathbf{x},t,\sigma) = \phi(\phi^{-1}(t) - \sqrt{Q^*_{\sigma^2}}) \qquad if \qquad 0 < t < 1, \tag{3.10}$$

Similar to the case of Pareto distribution, we have the following steps to find GPQ-based CIs for $\eta(t;\mu,\sigma)$.

For a given random sample $\{x_1,\ x_2,\ \ldots\ ,\ x_n\}$ from a Lognormal distribution, compute the value of s^2.

1. Generate a random value of U^2 from χ^2_{n-1}, compute $Q^*_{\sigma^2} = \frac{s^2(n-1)}{U^2}$, and then plug $Q^*_{\sigma^2}$ in (3.10) to obtain Q^*_η.

2. Repeat step 1 for m times, so we can get m copies $\{Q^*_{\eta,1}, Q^*_{\eta,2}, \cdots, Q^*_{\eta,m}\}$ of Q^*_η.

3. Sort the m copies of Q^*_η in ascending order, and denote the ordered values as $Q^*_\eta[1]$, $Q^*_\eta[2]$,$\cdots$, $Q^*_\eta[m]$.

The $100(1-\alpha)\%$ generalized confidence interval (GCI) for $\eta(t;\mu,\sigma)$ is $[Q^*_\eta[m*\alpha/2], Q^*_\eta[m*(1-\alpha/2)]]$.

CHAPTER 4

SIMULATION STUDIES

In this part, intensive simulation studies are conducted to evaluate finite sample performance of the proposed methods.

4.1 Influence Function-based Empirical Likelihood Intervals

Based on Theorem 1, the $100(1-\alpha)\%$ confidence interval is defined as $R_{IF} = \{\eta(t) : l_{IF}(\eta(t)) \leq \chi^2_{1,1-\alpha}\}$. By the influence function-based empirical likelihood (IFEL) approach, the coverage probability is calculated using the following procedures:

1. Generate $\{x_1, x_2, \cdots, x_n\}$ from an underlying distribution.

2. Calculate $\eta(t_0)$ for fixed t_0, solve equation (2.8) for $\nu_{IF}(t_0)$ and get the value of log-likelihood $l_{IF}(\eta(t_0)) = -2logR_{IF}(\eta(t_0)) = 2\sum_{i=1}^{n} log\{1 + \nu_{IF}(t)\hat{g}(X_i, \eta(t_0))\}$.

3. Set the initial value of $(\eta(t), \nu_{IF}(t))$ as $(\eta(t_0) - 0.1, 0)$. Then obtain the solution η_1 for $\eta(t)$ by solving the nonlinear equations (2.8) and

$$l_{IF}(\eta(t)) - \chi^2_{1,1-\alpha} = 0 \tag{4.1}$$

where $l_{IF}(\eta(t)) = 2\sum_{i=1}^{n} log\{1 + \nu_{IF}(t)\hat{g}(X_i, \eta(t))\}$. Set $(\eta(t_0) + 0.1, 0)$ as the initial value of $(\eta(t), \nu_{IF}(t))$ and solve the nonlinear equations again, get solution η_2. The interval length is calculated by $\eta_2 - \eta_1$.

4. Repeat 1-3 for B (a large number, e.g. B=5000) times, then calculate the coverage probability of the IFEL interval:

$$\frac{1}{B}\sum_{b=1}^{B} I(l_{IF,b}(\eta(t_0)) \leq \chi^2_{1,1-\alpha}), \tag{4.2}$$

and the average length of the confidence interval is

$$\frac{1}{B}\sum_{b=1}^{B}(\eta_{2,b}-\eta_{1,b}), \tag{4.3}$$

where $l_{IF,b}(\eta(t_0))$, $\eta_{1,b}$ and $\eta_{2,b}$ are the values of $l_{IF}(\eta(t_0))$, η_1 and η_2 based on b-th simulated sample, respectively.

Similarly, we can calculate the coverage probability and average length of IFJEL interval.

4.2 Asymptotic Confidence Intervals

One of the most popular methods to construct a confidence interval for an unknown parameter is the normal approximation. Since the MLE's $(\hat{\lambda},\hat{\beta})$ of Lorenz curve under Pareto distributions is invariant. Under the proper regularity assumptions, as $n\to\infty$, we have the following asymptotical result,

$$\sqrt{n}(\eta(t;\hat{\lambda},\hat{\beta})-\eta(t;\lambda,\beta)) \xrightarrow{d} N(0,(\nabla\eta(t;\lambda,\beta))^{\intercal}I(\Theta)^{-1}(\nabla\eta(t;\lambda,\beta))), \quad as \quad n\to\infty, \tag{4.4}$$

where $\hat{\Theta}=(\hat{\lambda},\hat{\beta})^{\intercal}$ has an asymptotic bivariate normal distribution with mean $\Theta=(\lambda,\beta)^{\intercal}$ and the covariance matrix determined by the Fisher information matrix $I(\Theta)$, $\nabla\eta$ is the first derivative of $\eta(t;\lambda,\beta)$.

There is a consistent estimator for Fisher information matrix $I(\Theta)$, which is proposed by Meilijson (1989) [19]. Let $l_i(x_i,\Theta)$ represent the single observation log-likelihood function and

$$s_i(x_i,\Theta)=\frac{\partial l_i(x_i,\Theta)}{\partial\Theta} \tag{4.5}$$

be the score function. Then the empirical Fisher information matrix $H(x,\Theta)$ can be defined as,

$$H(x,\Theta)=\frac{1}{n}\sum_{i=1}^{n}s_i(x_i,\Theta)s_i(x_i,\Theta)^{\intercal}-\frac{1}{n^2}S(x,\Theta)S(x,\Theta)^{\intercal} \tag{4.6}$$

where $S(x,\Theta)=\sum_{i=1}^{n}s_i(x_i,\Theta)$. To construct CI for a Lorenz ordinate, $I_n(\Theta)=nI(\Theta)$ needs

to be estimated by $\hat{I}_n(\Theta) = nH(x, \hat{\Theta})$. Therefore, an asymptotic $100(1-\alpha)\%$ confidence interval (ACI) for $\eta(t; \lambda, \beta)$ is

$$ACI = (\eta(t; \hat{\lambda}, \hat{\beta}) - z_{1-\alpha/2}\hat{\sigma}, \eta(t; \hat{\lambda}, \hat{\beta}) + z_{1-\alpha/2}\hat{\sigma}) \tag{4.7}$$

where $\hat{\sigma} = ((\nabla\eta(t; \hat{\lambda}, \hat{\beta}))^{\intercal}(nH(x, \hat{\Theta}))^{-1}(\nabla\eta(t; \hat{\lambda}, \hat{\beta})))^{1/2}$, $\nabla\eta(t; \hat{\lambda}, \hat{\beta})$ is the matrix gradient of $\eta(t; \lambda, \beta)$ at $\Theta = \hat{\Theta}$ and z_α is the α-th quantile of the standard normal distribution. When the underlying income distribution is a Lognormal distribution, we use the same way to find the ACI for $\eta(t; \mu, \sigma)$.

4.3 Bootstrap Confidence Intervals

The empirical bootstrap is a statistical technique popularized by Efron (1993)[12]. The key idea is to perform computations on the data itself to estimate the specific statistics from the same data. The bootstrap setup is as follows:

1. Generate $\{x_1, x_2, \cdots, x_n\}$ from a Pareto distribution with parameters λ and β, compute the MLE's $\hat{\lambda}$ and $\hat{\beta}$.

2. Generate bootstrap sample $x^* = \{x_1^*, x_2^*, \cdots, x_n^*\}$ from $\{x_1, x_2, \cdots, x_n\}$, compute the bootstrap copies $\hat{\lambda}^*$ and $\hat{\beta}^*$ of $\hat{\lambda}$ and $\hat{\beta}$ based on the bootstrap sample, then plug them in the Lorenz curve $\hat{\eta}^* = \eta^*(t, \hat{\lambda}^*, \hat{\beta}^*)$.

3. Repeat steps 1-2 for B (a large number, e.g. B=5000) times and get B bootstrap copies of $\hat{\eta}^*$. Let $\hat{\eta}_1^*, \hat{\eta}_2^*, \cdots, \hat{\eta}_B^*$ denote the bootstrap copies in ascending order, the $(1-\alpha)$-th percentile bootstrap confidence interval (BCI) for Lorenz ordinate $\eta(t)$ is given by

$$BCI = (\hat{\eta}^*_{[B\alpha/2]}, \hat{\eta}^*_{[B(1-\alpha/2)]}) \tag{4.8}$$

where $\hat{\eta}^*_{[B\alpha]}$ is the $[B\alpha]$-th value in the ordered list of the B replications of $\hat{\eta}^*$.

When the underlying distribution is a Lognormal distribution, we use the same algorithm to construct the confidence interval for Lorenz curve.

4.4 Simulation results

When generating samples, we should notice that most income distributions are positively skewed, so the choice of underlying distribution F should be a positively skewed distribution. In this part, we choose Pareto distribution with the shape parameter $\frac{1}{\lambda} = \frac{1}{15}$ and the scale parameter $\beta = 0.9$, and Lognormal distribution with mean $\mu = 0.05$ and standard deviation $\sigma = 0.5$. The sample size n is chosen to be 100, 200, 300 and 400. B and m are set as 5000. Let $t_0 = 0.1 + 0.05k$, $k = 0, 1, \cdots, 16$, we calculate the coverage probabilities and average interval lengths of 95% level confidence intervals for $\eta(t_0)$.

Figures $4.1 - 4.8$ show the simulation results. First, we observe that the coverage probabilities of the GCIs are very close to the nominal level in all the cases, and GCIs perform much better than ACIs and BCIs when sample size is small. IFEL and IFJEL also have good performances except for cases when t_0 falls in the upper tail of Lorenz curve, and they both outperform EL with better coverage probabilities. As sample size increases, the coverage probabilities of IFEL and IFJEL confidence intervals are in better agreement with the nominal level. Second, when we look at the average lengths of all the confidence intervals, GCIs and BCIs have similar average lengths and they are the shortest among all the confidence intervals. We also observe that IFEL and IFJEL confidence intervals have comparable average lengths, while EL confidence intervals has the longest average lengths. Overall, we recommend the GCI for the Lorenz curve when underlying distribution is a Pareto distribution or a Lognormal distribution, and recommend the IFEL confidence interval and IFJEL confidence interval for Lorenz curve when underlying distribution is unknown.

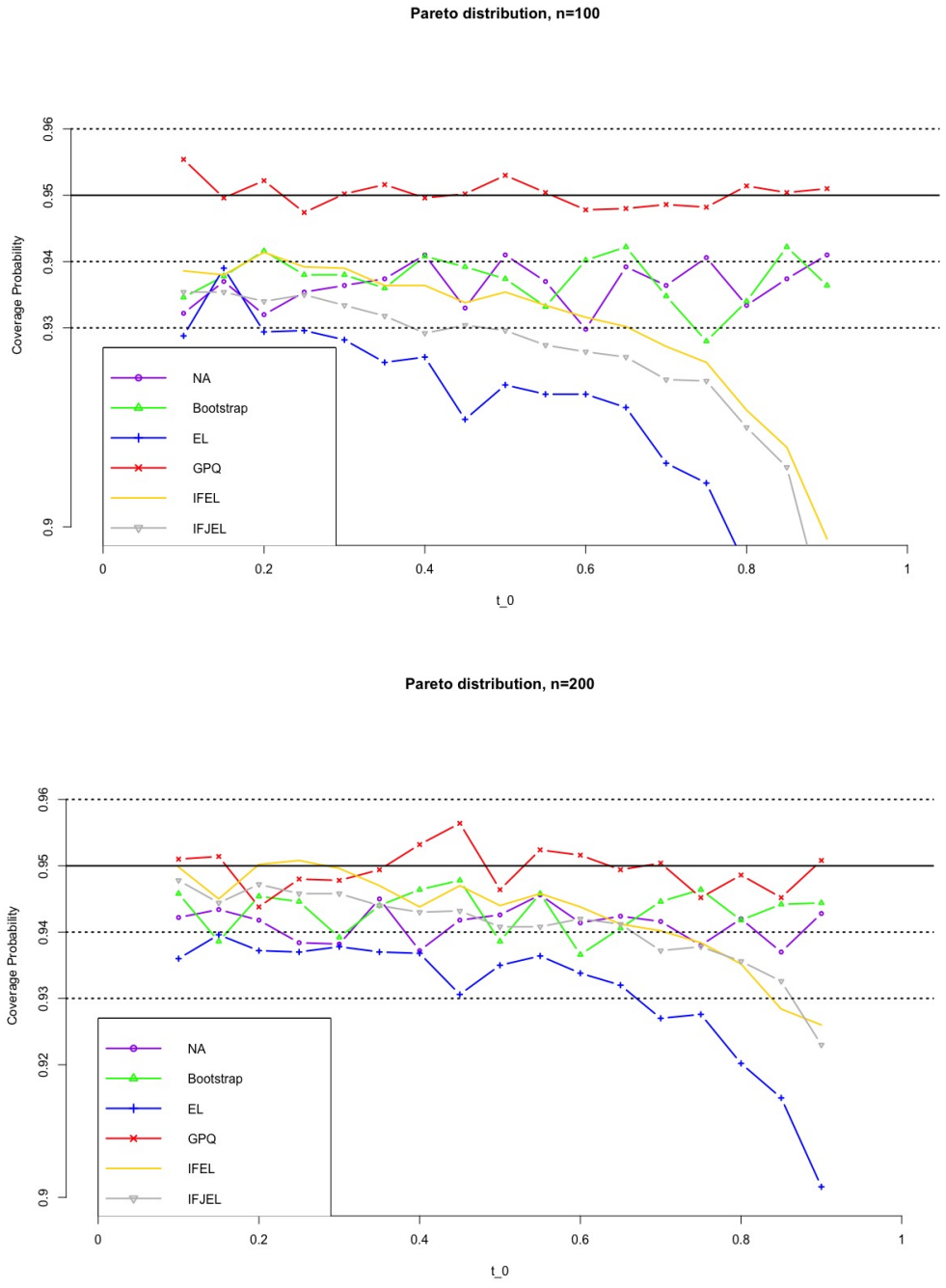

Figure 4.1. Coverage probabilities of the 95% confidence intervals for the Lorenz curve under Pareto distribution(n=100,200)

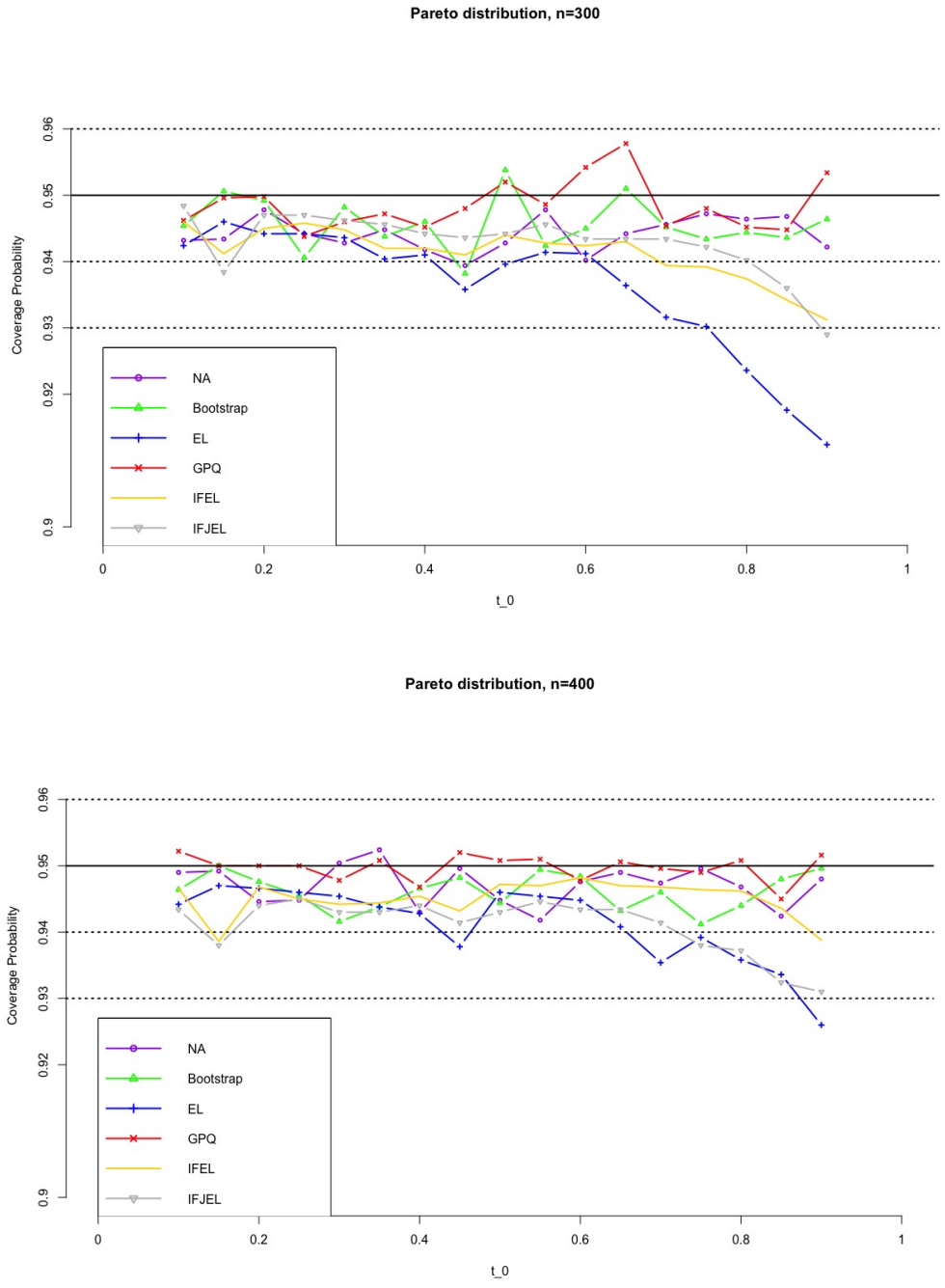

Figure 4.2. Coverage probabilities of the 95% confidence intervals for the Lorenz curve under Pareto distribution(n=300,400)

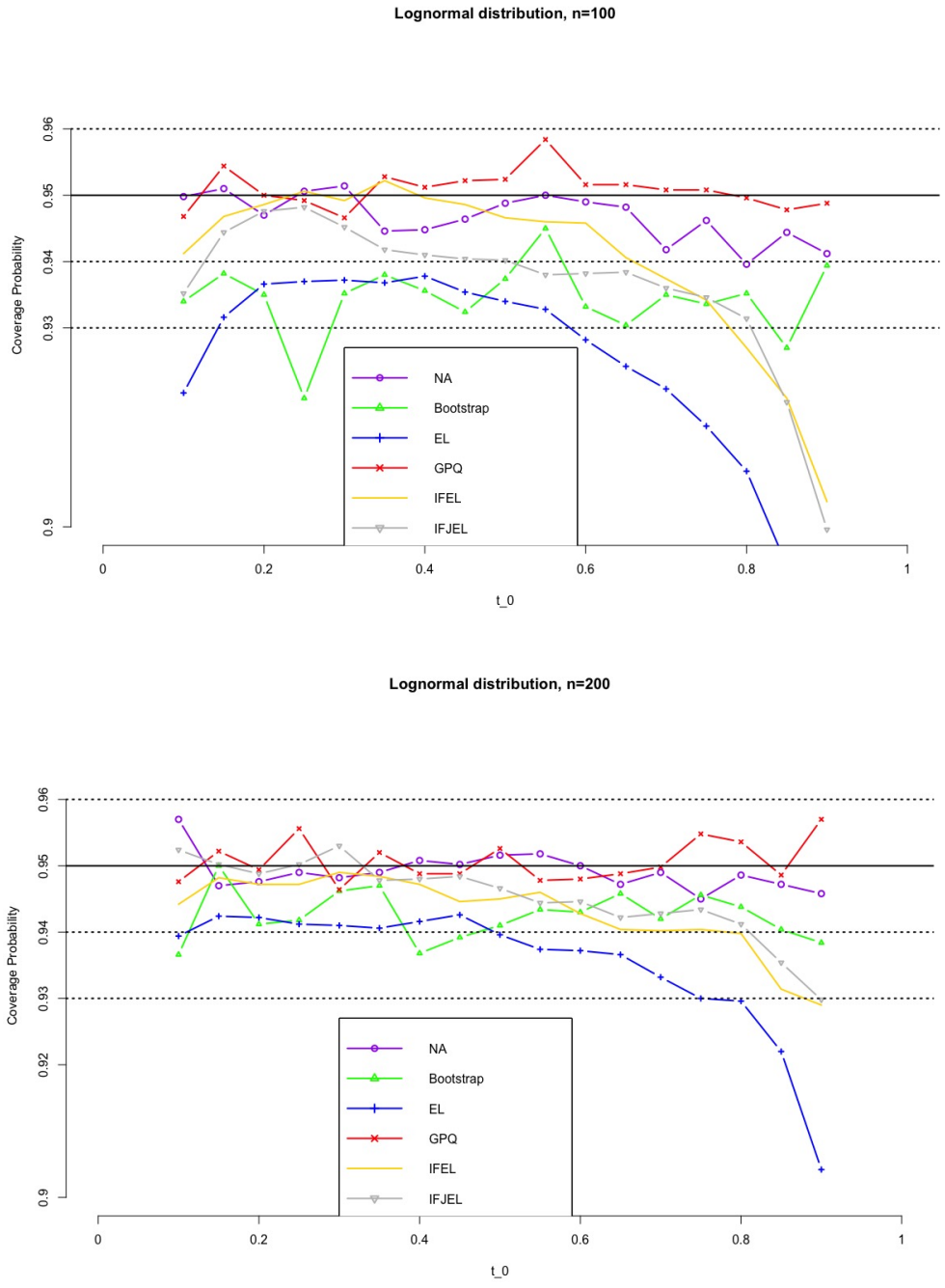

Figure 4.3. Coverage probabilities of the 95% confidence intervals for the Lorenz curve under Lognormal distribution(n=100,200)

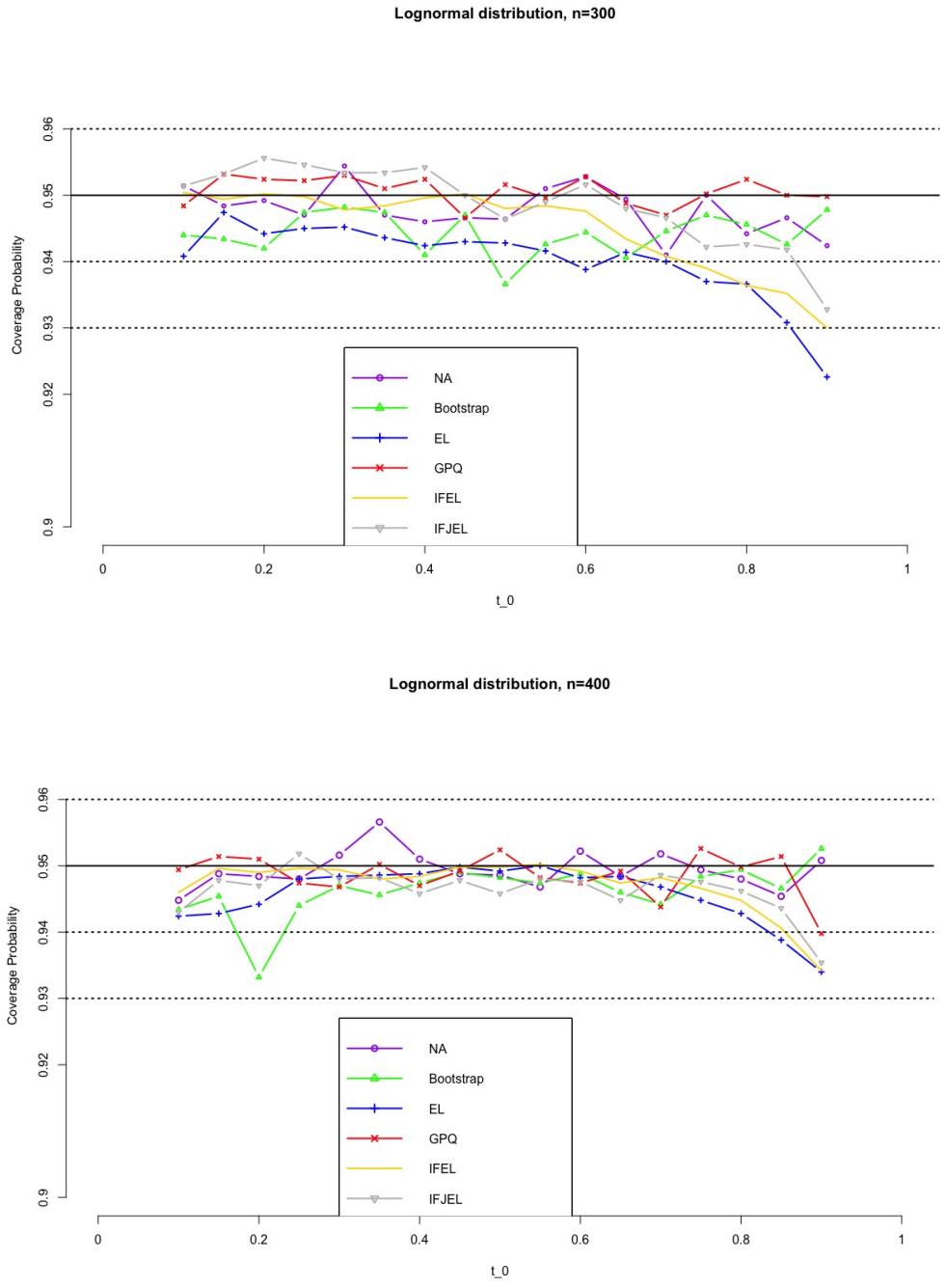

Figure 4.4. Coverage probabilities of the 95% confidence intervals for the Lorenz curve under Lognormal distribution(n=300,400)

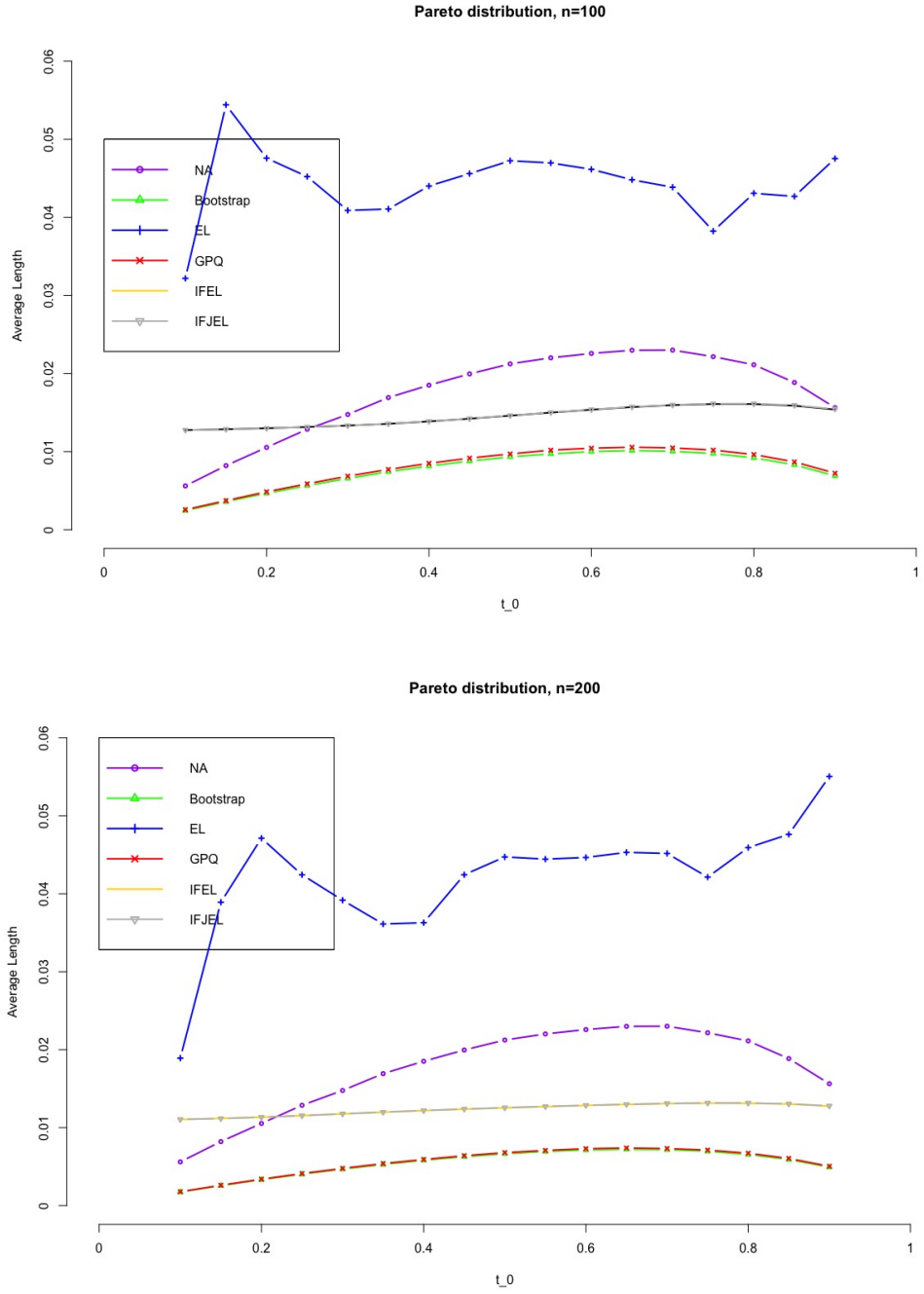

Figure 4.5. Average lengths of 95% confidence intervals for the Lorenz curve under Pareto distribution(n=100,200)(Note: IFEL almost overlaps IFJEL as their average lengths are very close, same for the GPQ and Boostrap.)

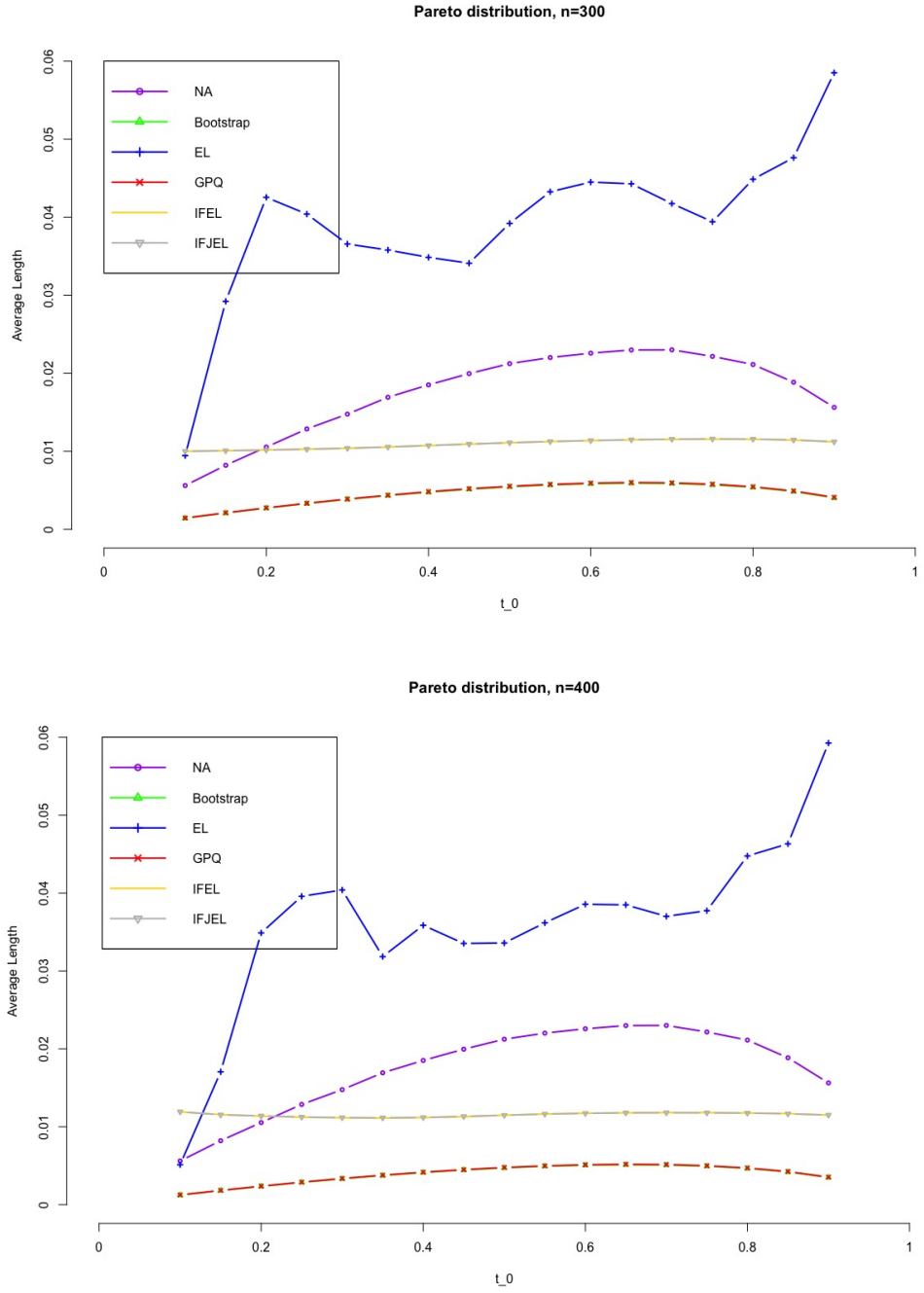

Figure 4.6. Average lengths of 95% confidence intervals for the Lorenz curve under Pareto distribution(n=300,400)(Note: IFEL almost overlaps IFJEL as their average lengths are very close, same for the GPQ and Boostrap.)

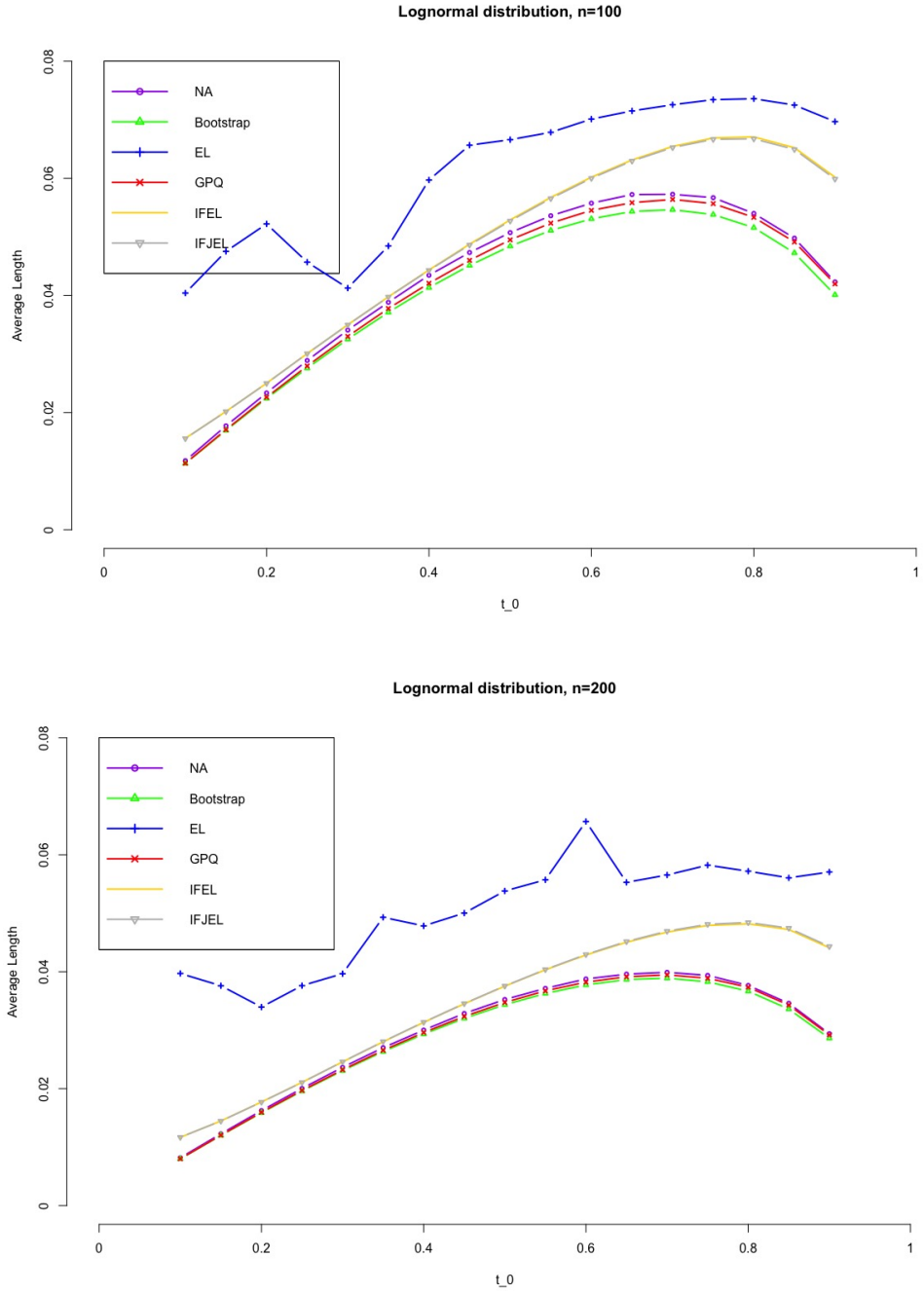

Figure 4.7. Average lengths of 95% confidence intervals for the Lorenz curve under Lognormal distribution(n=100,200)(Note: IFEL almost overlaps IFJEL as their average lengths are very close, same for the GPQ, Boostrap and NA.)

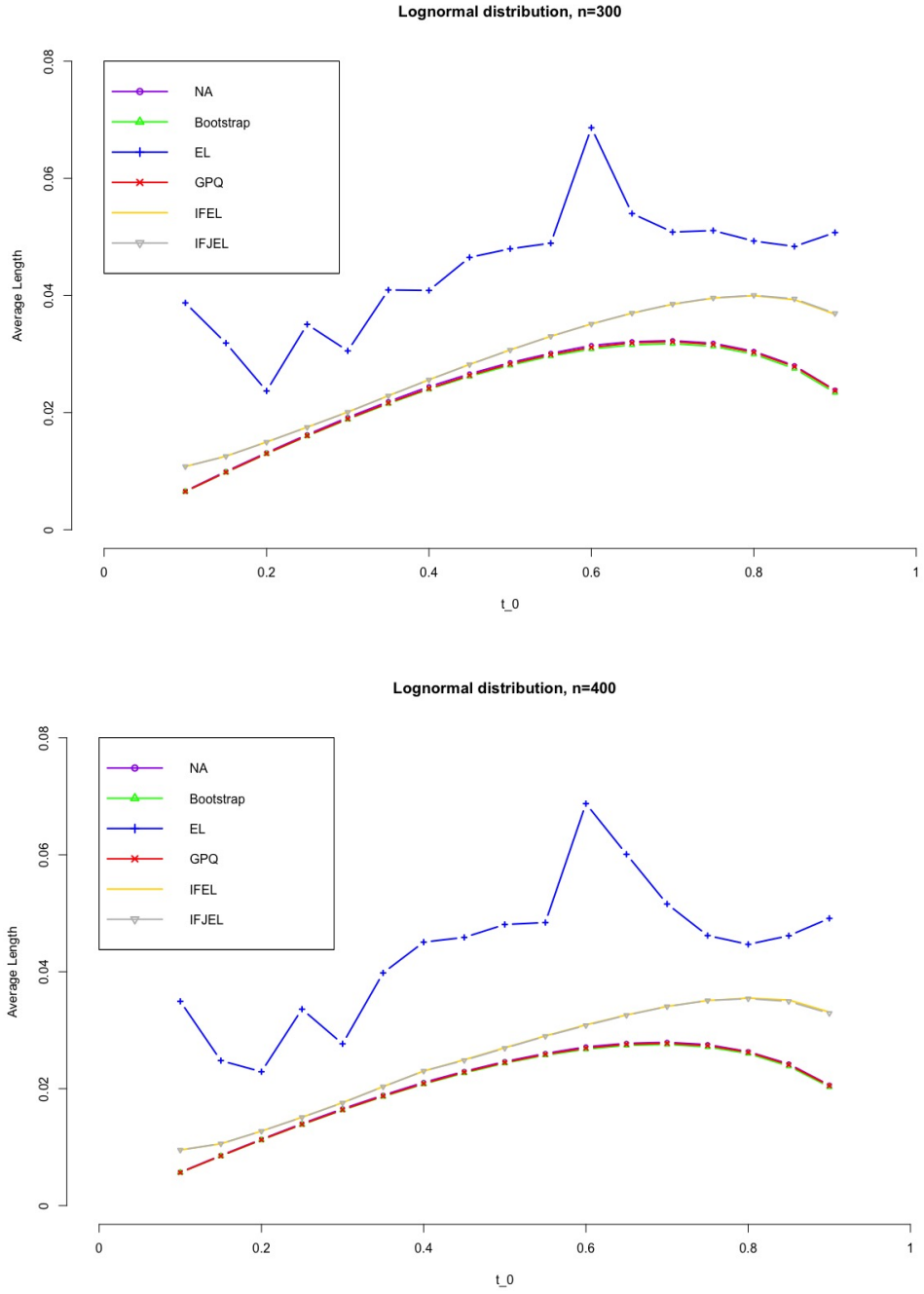

Figure 4.8. Average lengths of 95% confidence intervals for the Lorenz curve under Lognormal distribution(n=300,400)(Note: IFEL almost overlaps IFJEL as their average lengths are very close, same for the GPQ, Boostrap and NA.)

CHAPTER 5

REAL DATA EXAMPLES

5.1 Public Income Data from the Panel Study of Income Dynamics

In this section, we apply the proposed methods to make inferences for Lorenz curve with a real income dataset. Income inequality is a significant economic problem all over the world and the United States have the highest rising of income inequality among most developed countries (Weeks 2007 [27]). By constructing confidence intervals for Lorenz ordinates at different t, we can discuss how the income inequality is in the United States and have a general view of the inequality.

This income dataset is obtained from the database - The Panel Study of Income Dynamics (PSID) - from the University of Michigan (PSID 2017 [21]). It is part of the 2015 PSID Main Family Data, and it contains three variables: ER60001 - Release Number, ER60002 - 2015 Family Interview (ID) Number and ER65349 - Total Family Income-2014. The income reported here was collected in 2015 for tax year 2014. Please note that this variable can contain negative values. Negative value indicates a net loss, which in waves prior to 1994. These losses occur as a result of business or farm loss. Positive values mean actual income amounts and zero means there is no family income in 2014. There are in total 9048 households in this dataset.

A brief summary of this income data is shown in Table 5.1:

Table 5.1. Summary of 2015 PSID Family Data - Income

Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
-22000	24000	49310	69540	90210	5250000

The histogram (Figure 5.1) shows that this income data is severely positively skewed.

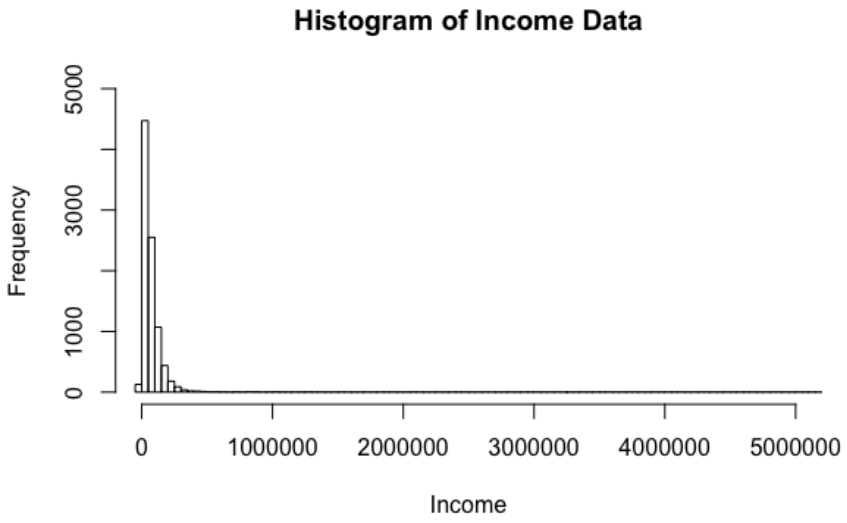

Figure 5.1. Histogram of Income Data

Here we utilize the Pareto distribution and Lognormal distribution to fit this income data. In order to test the goodness-of-fit, the popular goodness-of-fit index named Kolmogorov-Smirnov (K-S) statistic is used. We consider the null hypothesis $H_0 : X \sim Pareto(\beta, \lambda)$ and $H_0 : X \sim Lognormal(\mu, \sigma^2)$, the p-values for Pareto distribution and Lognormal distribution are both far less than 0.05. These p-values indicate that the income data does not follow a Pareto distribution nor a Lognormal distribution. So we can only use nonparametric methods, i.e. IFEL, IFJEL and EL methods to construct confidence intervals for Lorenz curve.

The non-parametric estimates for the Lorenz ordinates with their 95% confidence intervals are presented in Table 5.2. We can see that IFEL and IFJEL intervals are very similar in most cases. Based on our simulation results, we suggest to use IFEL and IFJEL intervals. For example, when $t = 0.9$, the 95% IFEL confidence interval for Lorenz curve is (0.6456; 0.6654), it means the ratio of the mean income of the lowest 90% households and the mean income of total households is greater than 0.6456, but smaller than 0.6654.

Table 5.2. 95% level confidence intervals for Lorenz ordinates

	IFEL	IFJEL	EL
t	Confidence interval	Confidence interval	Confidence interval
0.10	(0.0032,0.0110)	(0.0002,0.0080)	(0.0043,0.0116)
0.20	(0.0270,0.0306)	(0.0240,0.0276)	(0.0339,0.0392)
0.30	(0.0588,0.0652)	(0.0558,0.0622)	(0.0739,0.0850)
0.40	(0.1065,0.1146)	(0.1035,0.1116)	(0.1334,0.1439)
0.50	(0.1640,0.1776)	(0.1610,0.1746)	(0.2080,0.2194)
0.60	(0.2458,0.2582)	(0.2428,0.2552)	(0.2958,0.3147)
0.70	(0.3477,0.3597)	(0.3447,0.3567)	(0.4186,0.4313)
0.80	(0.4750,0.4900)	(0.4720,0.4870)	(0.5551,0.5705)
0.90	(0.6456,0.6654)	(0.6426,0.6624)	(0.7207,0.7391)

5.2 Median Income Data of Twenty Occupations in the U. S. in 1950

The second data set is about the median income of the twenty occupations in the United States Census of Population, 1950, Occupational Characteristics (Special Report,P-E No. 1B)(Shafiei, Saboori and Doostparast, 2016 [24]). The data set is presented in Table 5.3. The p-value of K-S test for Pareto distribution with parameters β and λ is 0.9905, the p-value of K-S test for Lognormal distribution with parameters μ and σ^2 is far less than 0.05. These p-values suggest that the median income data follows a Pareto distribution, so we use the Pareto distribution to model the median income data.

We recommend to use GCI for Lorenz curve as it has best performances in our simulation studies. Figure 5.2 shows the coverage probabilities of the 95% GCIs for the Lorenz curve under Pareto distribution when sample size is 20, which is the same size as the median income data. We can observe that the coverage probabilities are all very close to the nominal level 0.95. Table 5.4 displays various 95% level confidence intervals for Lorenz curve. For example, when $t = 0.1$, the 95% GCI for Lorenz curve is (0.0485, 0.0799), it means the ratio of the mean income of the lowest 10% occupations and the mean income of total occupations is greater than 0.0485, but smaller than 0.0799.

Table 5.3. Median income (by 1949 in dollars) of 20 occupations in the United States.

Occupation	Income	Occupation	Income
Accountants and auditors	3977	Mail-carriers	3480
Architects	5509	Plumbers and pipe fitters	3353
Authors, editors, and reporters	4303	Motormen, street, subway, and elevated rai1way	3424
Chemists	4091	Teachers (n.e.c.)	3456
Dentists	6448	Insurance agents and brokers	3771
Engineers, civil	4590	Electricians	3447
Lawyers and judges	6284	Locomotive engineers	4648
Physicians and surgeons	8302	Machinists and job setters, metal	3303
College presidents, professors, and instructors (n.e.c.)	4366	Managers, officials, and proprietors (n.e.c.) - self-employed - wholesale and retail trade	3806
Managers, officials, and proprietors (n.e.c.) - self-employed - manufacturing	4700		

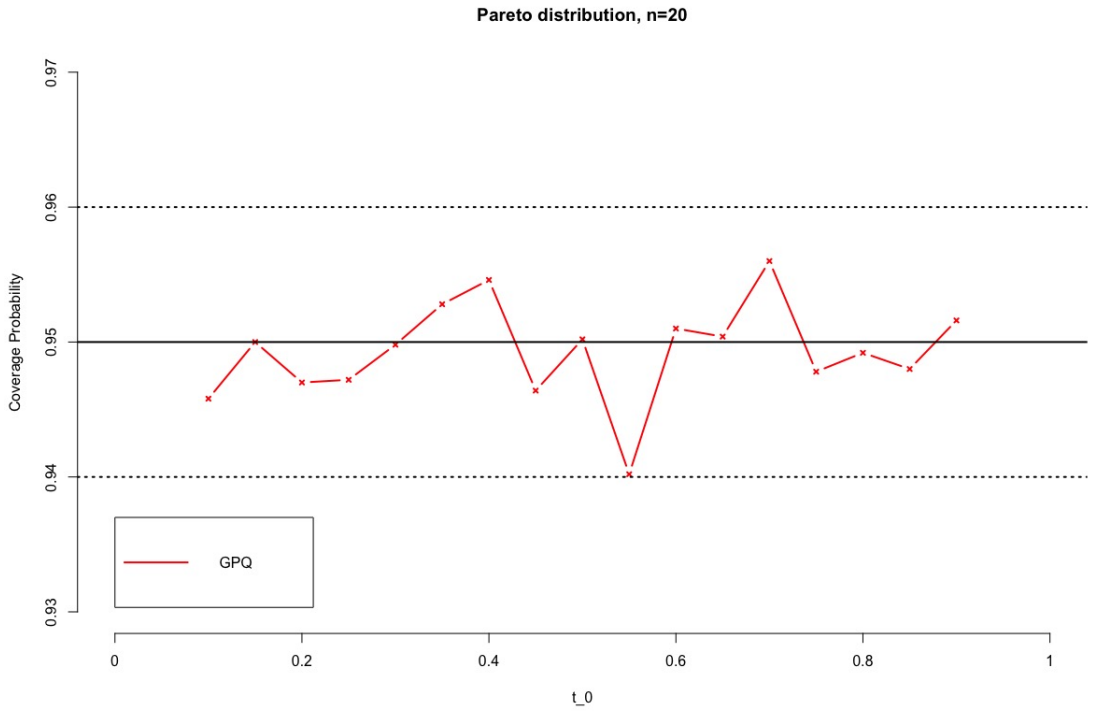

Figure 5.2. Coverage probabilities of the 95% GCIs for the Lorenz curve under Pareto distribution(n=20)

Table 5.4. Length of 95% CI for median income data of 20 occupations in the United States

t	Method	Confidence interval	t	Method	Confidence interval
0.1	GCI	(0.0485,0.0799)	0.7	GCI	(0.4360,0.6145)
	ACI	(0.0566,0.0864)		ACI	(0.4930,0.6500)
	BCI	(0.0589,0.0837)		BCI	(0.5019,0.6319)
	IFEL	(0.0620,0.0800)		IFEL	(0.5364,0.6330)
	IFJEL	(0.0499,0.0499)		IFJEL	(0.5376,0.6334)
	EL	(0.0319,0.0397)		EL	(0.5458,0.6419)
0.3	GCI	(0.1604,0.2458)	0.9	GCI	(0.6669,0.8381)
	ACI	(0.1798,0.2642)		ACI	(0.7330,0.8715)
	BCI	(0.1852,0.2572)		BCI	(0.7387,0.8530)
	IFEL	(0.1979,0.2518)		IFEL	(0.7915,0.8563)
	IFJEL	(0.1978,0.2524)		IFJEL	(0.7943,0.8587)
	EL	(0.1811,0.1836)		EL	(0.8000,0.8634)
0.5	GCI	(0.2831,0.4218)			
	ACI	(0.3214,0.4508)			
	BCI	(0.3298,0.4375)			
	IFEL	(0.3486,0.4303)			
	IFJEL	(0.3374,0.3818)			
	EL	(0.3498,0.3569)			

CHAPTER 6

CONCLUSIONS

In this thesis, an empirical likelihood method based on influence function is proposed and used to construct confidence intervals for the Lorenz ordinates. The Jackknife empirical likelihood method based on influence function is carried out. At the same time, we develop the generalized confidence intervals for Lorenz ordinates under Pareto and Lognormal distributions. Since Pareto and the lognormal distributions play a central role as probabilistic models for the distributions of various phenomena in different fields, the confidence intervals derived in this thesis should be of practical interest.

Simulation results show good coverage probabilities of influence function-based empirical likelihood confidence intervals and influence function-based Jackknife empirical likelihood confidence intervals in the lower tails when the sample size is large. Moreover, the coverage probabilities of the generalized confidence intervals are in good agreement with the nominal level for all the cases considered. They have pretty good performances even for the small samples, compared with the bootstrap confidence intervals and asymptotic confidence intervals. The real data examples show the confidence intervals for Lorenz ordinates at different t's, which gives us a general view of the income inequality and how severe it is in the United States. It also gives us an idea when to use our proposed methods. In sum, if the underlying income distribution is a Pareto distribution or a Lognormal distribution, generalized confidence interval could give us a better way to make inferences on Lorenz curve. If it is hard to know the underlying distribution or to fit a Pareto or a Lognormal distribution to the data, influence function-based empirical likelihood and influence function-based Jackknife empirical likelihood methods will be very useful in constructing confidence intervals for the Lorenz Curve.

REFERENCES

[1] A. B. Atkinson. On the measurement of inequality. *Journal of Economic Theory*, 2:244–263, 1970.

[2] C. M. Beach and R. Davidson. Distribution-free statistical inference with lorenz curves and income shares. *Review of Economic Studies*, 50:723–735, 1983.

[3] C. M. Beach and J. Richmond. Joint confidence intervals for income shares and lorenz curves. *International Economic Review*, 26:439–450, 1985.

[4] R. Chang and N. Halfon. Graphical distribution of pediatricians in the united states: An analysis of the fifty states and washington, dc. *Pediatrics*, 100:172–179, 1997.

[5] J. H. Chen and J. Qin. Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80:107–116, 1993.

[6] F. Clementi and M. Gallegati. Pareto's law of income distribution: Evidence for germany, the united kingdom, and the united states. In Chatterjee A., Yarlagadda S., and Chakrabarti B.K., editors, *Econophysics of Wealth Distributions*. Springer, Milano, 2005.

[7] M. Csörgö and R. Zitikis. Strassens lil for the lorenz curve. *Journal of Multivariate Analysis*, 59:1–12, 1996.

[8] J. T. DiCiccio and B. Efron. Bootstrap confidence intervals. *Statistical Science*, 11(3):189–228, 1996.

[9] D. J. Doiron and G. F. Barrett. Inequality in male and female earnings: the roles of hours and earnings. *Review of Economics and Statistics*, 78:410–420, 1996.

[10] B. Efron. Nonparametric standard errors and confidence intervals (with discussion). *Canadian Journal of Statsitics*, 9:139–172, 1981.

[11] B. Efron. *The Jackknife, the Boostrap, and Other Resampling Plans*. Philadelphia, Pa.: Society for Industrial and Applied Mathematics, 1982.

[12] B. Efron. *An Introduction to the Bootstrap*. Chapman: Hall/CRC, 1993.

[13] M. H. Gail and J. L. Gastwirth. A scale-free goodness-of-fit test for the exponential distribution based on the lorenz curve. *Journal of the American Statistical Association*, 73:229–243, 1978.

[14] J. L. Gastwirth. A general definition of lorenz curve. *Econometrica*, 39:1037–1039, 1971.

[15] P. Hall and B. La Scala. Methodology and algorithms of empirical likelihood. *International Statistical Review*, 58:109–127, 1977.

[16] J. S. Haukoos and R. J. Lewis. Advanced statistics: bootstrapping confidence intervals for statistics with 'difficult' distributions. *Academic Emergency Medicine*, 12:360–365, 2005.

[17] B. Jing, J. Yuan, and W. Zhou. Jackknife empirical likelihood. *Journal of the American Statistical Association*, 104:1224–1232, 2009.

[18] H. J. Malik. Estimation of the parameters of the pareto distribution. *Metrika*, 16:126–132, 1970.

[19] I. Meilijson. A fast improvement to the em algorithm on its own terms. *Journal of the Royal Statistical Society, Series B*, 51:127–138, 1970.

[20] A. Owen. Empirical likelihood ratio confidence intervals for single functional. *Biometrika*, 75:237–249, 1988.

[21] PSID. Panel Study of Income Dynamics, public use dataset. Produced and distributed by the Institute for Social Research, University of Michigan, Ann Arbor, MI. 2017.

[22] G. S. Qin, B. Y. Yang, and N. E. Belohnga-Hill. Empirical likelihood-based inferences for the lorenz curve. *Annals of the Institute of Statistical Mathematics*, 65:1–21, 2013.

[23] G. S. Qin and X. H. Zhou. Empirical likelihood inference for the area under the roc curve. *Biometrics*, 62:613–622, 2006.

[24] S. Shafiei, H. Saboori, and M. Doostparast. Generalized inferential procedures for generalized lorenz curves under the pareto distribution. *Journal of Statistical Computation and Simulation*, 87:267–279, 2016.

[25] G.C. Smith JR. Lorenz curve analysis of industrial decentralization. *Journal of the American Statistical Association*, 42:591–596, 1947.

[26] K. W. Tsui and S. Weerahandi. Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of the American Statistical Association*, 84:602–607, 1989.

[27] J. Weeks. Inequality trends in some developed oecd countries. *United Nations DESA Woking Paper*, 6:1–14, 2007.

[28] S. Weerahandi. Generalized confidence intervals. *Journal of the American Statistical Association*, 88:899–905, 1993.

[29] Wikipedia. Pareto principle, 2018. [Online at `https://en.wikipedia.org/wiki/Pareto_principle`; accessed 20-Feb-2018].

[30] A.T.A. Wood, K.A. Do, and N.M. Broom. Sequential linearization of empirical likelihood constraints with application to u-statistics. *Journal of Computational and Graphical Statistics*, 5(4):365–385, 1996.

[31] X. H. Zhou, G. S. Qin, H. Z. Lin, and G. Li. Inferences in censored cost regression models with empirical likelihood. *Statistica Sinica*, 16:1213–1232, 2006.

Appendix A

PROOF OF THEOREMS

Lemma 1. *Under the conditions in Theorem 1, we have*

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}(X_i,\eta(t_0))-g(X_i,\eta(t_0)))^2=o_p(1)$$

Proof of Lemma 1.

For a given t_0, let $g(X_i,\eta(t_0))=(X_i-\xi_{t_0})I(X_i\le\xi_{t_0})+t_0\xi_{t_0}-X_i\eta(t_0)$. The difference between $\hat{g}(X_i,\eta(t_0))$ and $g(X_i,\eta(t_0))$ can be expressed as

$$\hat{g}(X_i,\eta(t_0))-g(X_i,\eta(t_0))=A_i+B_i$$

where

$$A_i=(X_i-\hat{\xi}_{t_0})I(X_i\le\hat{\xi}_{t_0})-(X_i-\xi_{t_0})I(X_i\le\xi_{t_0})$$

$$B_i=t_0(\hat{\xi}_{t_0}-\xi_{t_0})$$

Applying the inequality $(a+b)^2\le 2(a^2+b^2)$, we obtain that

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}(X_i,\eta(t_0))-g(X_i,\eta(t_0)))^2=\frac{1}{n}\sum_{i=1}^{n}(A_i+B_i)^2\le\frac{2}{n}\sum_{i=1}^{n}A_i^2+\frac{2}{n}\sum_{i=1}^{n}B_i^2$$

Therefore, we only need to prove that the sample means of A_i^2 and B_i^2 converge to zero in probability. The proofs will be presented in **(A)** and **(B)** below.

(A) $\frac{1}{n}\sum_{i=1}^{n}A_i^2=o_p(1)$

$$\frac{1}{n}\sum_{i=1}^{n} A_i^2 = \frac{1}{n}\sum_{i=1}^{n}[(X_i - \hat{\xi}_{t_0})I(X_i \le \hat{\xi}_{t_0}) - (X_i - \xi_{t_0})I(X_i \le \xi_{t_0})]^2 =$$
$$\frac{1}{n}\sum_{i=1}^{n}[X_i^2 I(X_i \le \hat{\xi}_{t_0})(I(X_i \le \hat{\xi}_{t_0}) - I(X_i \le \xi_{t_0})) - 2X_i\hat{\xi}_{t_0} I(X_i \le \hat{\xi}_{t_0})(I(X_i \le \hat{\xi}_{t_0})$$
$$- I(X_i \le \xi_{t_0})) + \hat{\xi}_{t_0} I^2(X_i \le \hat{\xi}_{t_0})(\hat{\xi}_{t_0} I(X_i \le \hat{\xi}_{t_0}) - \xi_{t_0} I(X_i \le \xi_{t_0})$$
$$+ 2X_i\xi_{t_0} I(X_i \le \xi_{t_0})(I(X_i \le \hat{\xi}_{t_0}) - I(X_i \le \xi_{t_0})) - X_i^2 I(X_i \le \xi_{t_0})(I(X_i \le \hat{\xi}_{t_0})$$
$$- I(X_i \le \xi_{t_0})) - \xi_{t_0} I(X_i \le \xi_{t_0})(\hat{\xi}_{t_0} I(X_i \le \hat{\xi}_{t_0}) - \xi_{t_0} I(X_i \le \xi_{t_0}))]$$

By the strong consistency of the sample quantile $\hat{\xi}_{t_0}$, we obtain that $|I(X_i \le \hat{\xi}_{t_0}) - I(X_i \le \xi_{t_0})| \xrightarrow{p} 0$, $|\hat{\xi}_{t_0} I(X_i \le \hat{\xi}_{t_0}) - \xi_{t_0} I(X_i \le \xi_{t_0})| \xrightarrow{p} 0$, for $i = 1, 2, \cdots, n$. From $\frac{1}{n}\sum_{i=1}^{n} |X_i|^2 \to E(X^2) < \infty$ a.s., it follows that

$$\frac{1}{n}\sum_{i=1}^{n} A_i^2 = o_p(1)$$

(B) $\frac{1}{n}\sum_{i=1}^{n} B_i^2 = o_p(1)$

It's obvious that $|\hat{\xi}_{t_0} - \xi_{t_0}| \xrightarrow{p} 0$, thus $\frac{1}{n}\sum_{i=1}^{n} B_i^2 = o_p(1)$. □

Lemma 2.

(i) $\max_i |\hat{g}(X_i, \eta(t_0))| = o_p(\sqrt{n})$.

(ii) $\frac{1}{n}\sum_{i=1}^{n} \hat{g}^2(X_i, \eta(t_0)) = \sigma^2 + o_p(1)$.

(iii) $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{g}(X_i, \eta(t_0)) \xrightarrow{d} N(0, \sigma^2)$.

Proof of Lemma 2.

(i) Since $g(X_i, \eta(t_0))$ are i.i.d. random variables with zero mean and finite variance σ^2, $\max_i |g(X_i, \eta(t_0))| = o_p(\sqrt{n})$. We obtain that

$$\max_i |\hat{g}(X_i, \eta(t_0))| \le \max_i |\hat{g}(X_i, \eta(t_0)) - g(X_i, \eta(t_0))| + \max_i |g(X_i, \eta(t_0))| = o_p(\sqrt{n})$$

(ii) Similar to proof of Lemma 1, we can prove that

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i,\eta(t_0))(\hat{g}(X_i,\eta(t_0)) - g(X_i,\eta(t_0))) = o_p(1). \tag{A.1}$$

Then from Lemma 1, $(A.1)$, $(A.2)$ and $\frac{1}{n}\sum_{i=1}^{n} g^2(X_i,\eta(t_0)) = \sigma^2$, we have that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \hat{g}^2(X_i,\eta(t_0)) = \ & \frac{1}{n}\sum_{i=1}^{n}(\hat{g}(X_i,\eta(t_0)) - g(X_i,\eta(t_0)))^2 + \frac{1}{n}\sum_{i=1}^{n} g(X_i,\eta(t_0))^2 \\ & + \frac{2}{n}\sum_{i=1}^{n} g(X_i,\eta(t_0))(\hat{g}(X_i,\eta(t_0)) - g(X_i,\eta(t_0))) = \sigma^2 + o_p(1).\end{aligned} \tag{A.2}$$

Lemma 2(ii) is proved.

(iii) Similar to proof of Lemma 1, we can obtain that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{g}(X_i,\eta(t_0)) - g(X_i,\eta(t_0))) = o_p(1). \tag{A.3}$$

Then we have

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{g}(X_i,\eta(t_0)) \ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{g}(X_i,\eta(t_0)) - g(X_i,\eta(t_0))) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(X_i,\eta(t_0)) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(X_i,\eta(t_0)) + o_p(1)\end{aligned} \tag{A.4}$$

From $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(X_i,\eta(t_0)) \xrightarrow{d} N(0,\sigma^2)$, Lemma 2(iii) is proved. □

Proof of Theorem 1.

Using Lemma 2 and Lemmas in Owen (1990), we have $|\nu_{IF}| = O_p(n^{-\frac{1}{2}})$. Applying Taylor expansion to (2.9), we can obtain

$$\begin{aligned} l_{IF}(\eta(t_0)) \ &= \ 2\sum_{i=1}^{n} log\{1 + \nu_{IF}(t_0)\hat{g}(X_i,\eta(t_0))\} \\ &= \ 2\sum_{i=1}^{n}(\nu_{IF}(t_0)\hat{g}(X_i,\eta(t_0)) - 12(\nu_{IF}(t_0)\hat{g}(X_i,\eta(t_0))^2) + r_{1n}\end{aligned}$$

with $|r_{1n}| \leq C\sum_{i=1}^{n}|\nu_{IF}(t_0)\hat{g}(X_i,\eta(t_0))|^3 \leq C|\nu_{IF}(t_0)|^3\max_i|\hat{g}(X_i,\eta(t_0))|\sum_{i=1}^{n}\hat{g}^2(X_i,\eta(t_0)) = o_p(1)$. From (2.8),

$$\begin{aligned}\sum_{i=1}^{n}\frac{\hat{g}(X_i,\eta(t_0))}{1+\nu_{IF}(t_0)\hat{g}(X_i,\eta(t_0))} = & \sum_{i=1}^{n}\hat{g}(X_i,\eta(t_0))[1-\nu_{IF}(t_0)\hat{g}(X_i,\eta(t_0))+\frac{(\nu_{IF}(t_0)\hat{g}(X_i,\eta(t_0)))^2}{1+\nu_{IF}(t_0)\hat{g}(X_i,\eta(t_0))}] = \\ & \sum_{i=1}^{n}\hat{g}(X_i,\eta(t_0)) - \nu_{IF}(t_0)\sum_{i=1}^{n}\hat{g}^2(X_i,\eta(t_0)) + \\ & \sum_{i=1}^{n}\frac{\hat{g}(X_i,\eta(t_0))(\nu_{IF}(t_0)\hat{g}(X_i,\eta(t_0)))^2}{1+\nu_{IF}(t_0)\hat{g}(X_i,\eta(t_0))} = 0,\end{aligned}$$

it follows that $\nu_{IF}(t_0) = \frac{\sum_{i=1}^{n}\hat{g}(X_i,\eta(t_0))}{\sum_{i=1}^{n}\hat{g}^2(X_i,\eta(t_0))} + o_p(n^{-\frac{1}{2}})$. Further, we have that $\sum_{i=1}^{n}\nu_{IF}(t_0)\hat{g}(X_i,\eta(t_0)) = \sum_{i=1}^{n}(\nu_{IF}(t_0)\hat{g}(X_i,\eta(t_0)))^2 + o_p(1)$.

Therefore, by Lemma 1, we obtain that

$$\begin{aligned} l_{IF}(\eta(t_0)) &= \sum_{i=1}^{n}(\nu_{IF}(t_0)\hat{g}(X_i,\eta(t_0)))^2 + o_p(1) \\ &= \frac{(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{g}(X_i,\eta(t_0)))^2}{\frac{1}{n}\sum_{i=1}^{n}\hat{g}^2(X_i,\eta(t_0))} + o_p(1) = \chi_1^2 + o_p(1). \ \square \end{aligned}$$