Ann Inst Stat Math (2013) 65:1–21
DOI 10.1007/s10463-012-0355-z

Empirical likelihood-based inferences for the Lorenz curve

Gengsheng Qin · Baoying Yang · Nelly E. Belinga-Hall

Received: 22 November 2010 / Revised: 2 November 2011 / Published online: 11 March 2012

Abstract In this paper, we discuss empirical likelihood-based inferences for the Lorenz curve. The profile empirical likelihood ratio statistics for the Lorenz ordinate are defined under the simple random sampling and the stratified random sampling designs. It is shown that the limiting distributions of the profile empirical likelihood ratio statistics are scaled Chi-square distributions with one degree of freedom. We also derive the limiting processes of the associated empirical likelihood-based Lorenz processes. Hybrid bootstrap and empirical likelihood intervals for the Lorenz ordinate are proposed based on the newly developed empirical likelihood theory. Extensive simulation studies are conducted to compare the relative performances of various confidence intervals for Lorenz ordinates in terms of coverage probability and average interval length. The finite sample performances of the empirical likelihood-based confidence bands are also illustrated in simulation studies. Finally, a real example is used to illustrate the application of the recommended intervals.

Keywords Bootstrap · Confidence interval/band · Empirical likelihood · Income distribution · Lorenz curve

G. Qin (✉) · N. E. Belinga-Hall
Department of Mathematics and Statistics, Georgia State University, 30 Pryor Street,
Atlanta, GA 30303, USA
e-mail: matgjq@langate.gsu.edu

B. Yang
College of Mathematics, Southwest Jiaotong University, Chengdu 610031, China

1 Introduction

The Lorenz (1905) curve was introduced to investigate the problem of measuring concentrations of wealth in a population. It plots the percentage of total income earned by various portions of the population when the population is ordered by the size of their incomes (Csörgö et al. 1998; Gastwirth 1971). Let X be a non-negative random variable with a cumulative distribution function $F(x)$. Assume that $F(x)$ is differentiable. Gastwirth (1971) provided a general definition of Lorenz curve as follows:

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

where $\mu = \int_0^\infty x\mathrm{d}F(x)$ is the mean of F, and $\xi_t = F^{-1}(t)$ is the tth quantile of F. 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 tth fraction of incomes.

The Lorenz curve has been widely utilized in economics and social sciences. It provides a way for the partial ordering of income distributions (Atkinson 1970). Economists have used Lorenz dominance to analyze income and earning inequality (Doiron and Barrett 1996; Sen 1973). The Lorenz curve analysis have been applied in other areas such as industrial concentration (Hart 1971), reliability (Gail and Gastwirth 1978), and medical and health services research (Chang and Halfon 1997; Hallas and Støvring 2006).

Since income distribution F 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 independent copies of X. The Lorenz curve can be empirically estimated by

$$\widehat{\eta}(t) = \frac{1}{\widehat{\mu}}\int_0^{\widehat{\xi}_t} x\ \mathrm{d}F_n(x), \tag{2}$$

where $\widehat{\mu}$ is the sample mean, F_n is the empirical distribution function of the sample, and $\widehat{\xi}_t = \inf\{y : F_n(y) \geq t\}$ is the tth sample quantile.

The asymptotic theory for $\widehat{\eta}(t)$ has been developed in Beach and Davidson (1983). In addition, Zheng (2002) showed that the empirical Lorenz ordinates are asymptotically and normally distributed when samples are not simple random. Goldie (1977) and Csörgö et al. (1986) derived the limiting Gaussian processes of the Lorenz process $\{\widehat{\eta}(t) : t \in [0, 1]\}$. These asymptotic theories can be used to make inference for the Lorenz curve. However, the existing normal approximation-based inferential methods have poor performance when the population distribution F is skewed and t falls in the tails of the Lorenz curve.

Empirical likelihood (EL), introduced by Owen (1988, 1990), is a powerful nonparametric method and its advantages over the normal approximation-based methods have been well-recognized (Hall and La Scala 1990). Over last two decades, empirical likelihood has been widely applied in many areas such as survey sampling, medical studies, and econometrics. For survey sampling, Chen and Qin (1993) proposed EL-based inferences for finite populations with auxiliary information. Zhong and Rao (2000) and Wu (2004) extended Chen and Qin (1993) approach to data in complex surveys. In the area of health care, Zhou et al. (2006) developed EL-based inferences in censored cost regression models and showed that the EL-based method outperforms the existing method. The other extreme is the skewness of zero costs for some healthy individuals. Chen and Qin (2003) developed EL-based inferences for data containing observations that are zero. Finally, we observe that most income data are skewed or highly skewed data in economics study. Belinga-Hill (2007) considered the interval estimation for the generalized Lorenz curve. Through some simulation studies, she showed that the EL method has better performance than that of the normal approximation method. This motivates us to develop new EL-based methods to make inferences for the Lorenz curve.

The paper is organized as follows. In Sect. 2, we first define the profile EL ratio statistic for the Lorenz ordinate under simple random sampling design. Secondly, the asymptotic distribution of the statistic is shown to be a scaled Chi-square distribution. Finally, we also derive the limiting process of the EL-based Lorenz process. In Sect. 3, we extend the EL theory for the Lorenz curve to the stratified random sample. In Sect. 4, we propose various confidence intervals for the Lorenz ordinate; we also propose the EL-based confidence bands for the Lorenz curve based on the limiting Lorenz process. Simulation studies are conducted to evaluate the small sample performances of these intervals and confidence bands. In Sect. 5, we apply the proposed intervals to a real income data set. Finally, the proof of the EL theorems are given in Appendix.

2 Empirical likelihood for the Lorenz curve with simple random sample

Let $\{X_1, \ldots, X_n\}$ be a simple random sample drawn from the population of X with c.d.f. F. N is the population size. For a fixed $t \in (0, 1)$, the Lorenz ordinate $\eta(t)$ satisfies $E[X(I(X \le \xi_t) - \eta(t))] = 0$. So, the empirical likelihood for $\eta(t)$ can be defined as follows:

$$\widetilde{L}_1(\eta(t)) = \sup_{\mathbf{p}} \left\{ \prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i D_i(t) = 0 \right\}, \tag{3}$$

where $\mathbf{p} = (p_1, \ldots, p_n)$ is a probability vector, $D_i(t) = X_i[I(X_i \le \xi_t) - \eta(t)], i = 1, \ldots, n$. Since $D_i(t)$ in (3) depends on unknown population quantile ξ_t, we substitute ξ_t with its consistent estimator $\widehat{\xi}_t = X_{([nt])}$, where $X_{([nt])}$ is the $[nt]$th ordered value of X_is. Then, we get the profile empirical likelihood for $\eta(t)$:

$$L_1(\eta(t)) = \sup_{\mathbf{p}} \left\{ \prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \widehat{D}_i(t) = 0 \right\}, \tag{4}$$

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

A unique maximum for $\mathbf{p}$ in (4) exists if $\eta(t)$ is inside the convex hull of $\{X_1[I(X_1 \le \widehat{\xi}_t) - \eta(t)], \ldots, X_n[I(X_n \le \widehat{\xi}_t) - \eta(t)]\}$. By Lagrange multiplier method, the supremum occurs at $p_i = \frac{1}{n}\{1 + \nu(t)\widehat{D}_i(t)\}^{-1}, i = 1, \ldots, n$, where $\nu(t)$ is the solution to

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{D}_i(t)}{1+\nu(t)\widehat{D}_i(t)}=0. \tag{5}$$

Note that $\prod_{i=1}^{n} p_i$, subject to $\sum_{i=1}^{n} p_i = 1, p_i \geq 0, i = 1, 2, \ldots, n$, attains its maximum n^{-n} at $p_i = n^{-1}$. So, the profile empirical likelihood ratio for $\eta(t)$ can be defined as

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

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

$$l_1(\eta(t))=-2\log R_1(\eta(t))=2\sum_{i=1}^{n}\log\{1+\nu(t)\widehat{D}_i(t)\}. \tag{6}$$

The following result, proved in Appendix, gives the limiting distribution of $l_1(\eta(t))$.

Theorem 1 *If $E(X^2) < \infty, n/N \longrightarrow 0$ 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 $l_1(\eta(t_0))$ is a scaled Chi-square distribution with degree of freedom 1. That is, $r_1 l_1(\eta(t_0)) \xrightarrow{\mathcal{L}} \chi_1^2$, where the scale constant $r_1 = s_p^2(t_0)/s_d^2(t_0)$ with $s_p^2(t_0) = \int_0^\infty \{x[I(x \leq \xi_{t_0}) - \eta(t_0)]\}^2 \mathrm{d}F(x)$, $s_d^2(t_0) = \int_0^\infty [(x-\xi_{t_0})I(x \leq \xi_{t_0}) - x\eta(t_0)]^2 \mathrm{d}F(x) - (t_0\xi_{t_0})^2$.*

Theorem 1 can be used to make inference for the Lorenz ordinate $\eta(t)$ at a fixed t. However, to make inference for the full Lorenz curve, we need the following theorem:

Theorem 2 *If $E(X^2) < \infty$, and $n/N \longrightarrow 0$, then the EL-based Lorenz process $\{l_1(\eta(t)) : t \in [0, 1]\}$ converges to $J^2(t)/s_p^2(t)$ in distribution, where*

$$J(t)=\frac{1}{\mu}\left[\int_0^{\xi_t}B(F(x))\mathrm{d}x-\eta(t)\int_0^{\infty}B(F(x))\mathrm{d}x\right], \tag{7}$$

B is the Brownian bridge on $[0, 1]$, and $J(t)$ is a Gaussian process with mean zero and the following covariance function:

$$\mathrm{Cov}(J(s), J(t)) = \mu^{-2}[\sigma_1(s,t)+\eta(s)\eta(t)\sigma_1(1,1)-\eta(t)\sigma_1(s,1)-\eta(s)\sigma_1(t,1)],$$

with $\sigma_1(s,t) = \int_0^{\xi_s}\int_0^{\xi_t}(F(x \wedge y) - F(x)F(y))\mathrm{d}x\mathrm{d}y$.

3 Empirical likelihood for the Lorenz curve with stratified random sample

Stratified random sampling is an often used sample design in collecting economic data. Suppose a population X with c.d.f. $F(x)$ is divided into H independent strata and the jth stratum X_j has the c.d.f $F_j(x)$, for $j = 1, \ldots, H$. Assume that the population size of the jth stratum is N_j, $N = \sum_{j=1}^{H} N_j$ is the whole population size. Let $X_{j1}, \ldots, X_{jn_j}$ be a simple random sample from the jth stratum X_j, for $j = 1, \ldots, H$. Then the stratified random sample has a sample size of $n = \sum_{j=1}^{H} n_j$. Furthermore, we assume that n_j/N_j is small for each j, and sample size of each stratum is proportional to its population size, i.e., $\frac{N_j}{N} = \frac{n_j}{n}$, $j = 1, \ldots, H$. This set-up for the stratified random sample has been used in Zheng (2002) for testing Lorenz curves with non-simple random samples.

Under the stratified random sample design, the Lorenz ordinate $\eta(t)$ is

$$\eta(t) = \frac{1}{\mu} \sum_{j=1}^{H} \frac{N_j}{N} \int_0^{\xi_t} x \mathrm{d}F_j(x) = \frac{1}{\mu} \sum_{j=1}^{H} \frac{N_j}{N} E[X_j I(X_j \le \xi_t)],$$

where $\mu = \int_0^{\infty} x \mathrm{d}F(x) = \sum_{j=1}^{H} \frac{N_j}{N} \int_0^{\infty} x \mathrm{d}F_j(x) = \sum_{j=1}^{H} \frac{N_j}{N} \mu_j$, and $\mu_j = E(X_j)$, $j = 1, \ldots, H$. Therefore,

$$\sum_{j=1}^{H} \frac{N_j}{N} [E(X_j I(X_j \le \xi_t)) - \mu_j \eta(t)] = 0. \tag{8}$$

Let $\mathbf{p_j} = (p_{j1}, \ldots, p_{jn_j})$ be a probability vector, for $j = 1, \ldots, H$. Using (8), the profile empirical likelihood for $\eta(t)$ can be defined as follows:

$$L_2(\eta(t)) = \sup \left\{ \prod_{j=1}^{H} \prod_{k=1}^{n_j} p_{jk} : \sum_{k=1}^{n_j} p_{jk} = 1, j = 1, \ldots, H, \right. \\ \left. \times \sum_{j=1}^{H} \frac{N_j}{N} \sum_{k=1}^{n_j} p_{jk} \widehat{D}_{jk}(t) = 0 \right\}, \tag{9}$$

where $\widehat{D}_{jk}(t) = X_{jk} I(X_{jk} \le \widehat{\xi}_t) - \bar{X}_j \eta(t), k = 1, \ldots, n_j$, $\bar{X}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} X_{jk}$, and $\widehat{\xi}_t = X_{([nt])}$ is the $[nt]$th order statistic of the stratified random sample.

To solve (9), a two-step procedure (see also Zhong and Rao 2000) is proposed here. The first step is to find the profile EL for $\widetilde{\eta}_1(t), \widetilde{\eta}_2(t), \ldots, \widetilde{\eta}_H(t)$:

$$L_{21}(\widetilde{\eta}_1(t), \ldots, \widetilde{\eta}_H(t)) = \sup \left\{ \prod_{j=1}^{H} \prod_{k=1}^{n_j} p_{jk} : \sum_{k=1}^{n_j} p_{jk} = 1, \right. \\ \left. \times \sum_{k=1}^{n_j} p_{jk} \widetilde{D}_{jk}(t) = 0, j = 1, \ldots, H \right\}$$

where $\widetilde{D}_{jk}(t) = X_{jk} I(X_{jk} \le \widehat{\xi}_t) - \widetilde{\eta}_j(t),\ k = 1, \ldots, n_j$, and $\widetilde{\eta}_j(t) = E[X_j I(X_j \le \xi_t)],\ j = 1, \ldots, H$.

Using the Lagrange multiplier method, the supremum occurs at $p_{jk} = \frac{1}{n_j}\frac{1}{1+\nu_j(t)\widetilde{D}_{jk}(t)}$, where $\nu_1(t), \nu_2(t), \ldots, \nu_H(t)$ are Lagrange multipliers determined by estimating functions

$$\frac{1}{n_j}\sum_{k=1}^{n_j}\frac{\widetilde{D}_{jk}(t)}{1+\nu_j(t)\widetilde{D}_{jk}(t)} = 0, \quad j = 1, 2, \ldots, H. \tag{10}$$

Then we get that

$$\log L_{21}(\widetilde{\eta}_1(t), \ldots, \widetilde{\eta}_H(t)) = -\sum_{j=1}^{H}\sum_{k=1}^{n_j}\log(1+\nu_j(t)\widetilde{D}_{jk}(t)) - \sum_{j=1}^{H} n_j \log n_j.$$

The second step is to solve the following optimal problem to find the profile empirical likelihood for $\eta(t)$:

$$L_2(\eta(t)) = \sup_{\widetilde{\eta}_j(t),\ j=1,\ldots,H}\left\{L_{21}(\widetilde{\eta}_1(t), \ldots, \widetilde{\eta}_H(t)) : \sum_{j=1}^{H}\frac{N_j}{N}\widetilde{\eta}_j(t) = \sum_{j=1}^{H}\frac{N_j}{N}\bar{X}_j\eta(t)\right\}.$$

We use the Lagrange multiplier method once again. Let

$$LL_2 = -\sum_{j=1}^{H}\sum_{k=1}^{n_j}\log(1+\nu_j(t)\widetilde{D}_{jk}(t)) - n\widetilde{\nu}(t)\sum_{j=1}^{H}\frac{N_j}{N}(\widetilde{\eta}_j(t) - \bar{X}_j\eta(t)).$$

By (10) and $\frac{1}{n_j}\sum_{k=1}^{n_j}\frac{1}{1+\nu_j(t)\widetilde{D}_{jk}(t)} = 1$, $j = 1, \ldots, H$, solving $\frac{\partial LL_2}{\partial \eta_j(t)} = 0$, we get that $\nu_j(t) = \frac{n}{n_j}\frac{N_j}{N}\widetilde{\nu}(t) = \widetilde{\nu}(t)$, and

$$\log L_2(\eta(t)) = -\sum_{j=1}^{H}\sum_{k=1}^{n_j}\log(1+\widetilde{\nu}(t)\widetilde{D}_{jk}(t)) - \sum_{j=1}^{H} n_j \log n_j,$$

where $\widetilde{\nu}(t), \widetilde{\eta}_1(t), \widetilde{\eta}_2(t), \ldots, \widetilde{\eta}_H(t)$ satisfy the following system of equations:

$$\frac{1}{n_j}\sum_{k=1}^{n_j}\frac{\widetilde{D}_{jk}(t)}{1+\widetilde{\nu}(t)\widetilde{D}_{jk}(t)} = 0, \quad j = 1, 2, \ldots, H. \tag{11}$$

$$\eta(t) = \frac{\sum_{j=1}^{H}\frac{N_j}{N}\widetilde{\eta}_j(t)}{\sum_{j=1}^{H}\frac{N_j}{N}\bar{X}_j}. \tag{12}$$

Note that $\prod_{j=1}^{H}\prod_{k=1}^{n_j} p_{jk}$, under the constraint $\sum_{k=1}^{n_j} p_{jk} = 1$ for all j, k, attains its maximum at $p_{jk} = \frac{N}{N_j n} = n_j^{-1}$. So, the profile empirical log-likelihood ratio for $\eta(t)$ is:

$$l_2(\eta(t)) = 2\sum_{j=1}^{H}\sum_{k=1}^{n_j}\log(1+\widetilde{\nu}(t)\widetilde{D}_{jk}(t)). \tag{13}$$

To find the limiting distribution of $l_2(\eta(t))$, we need the Bahadur's representation for the sample quantile with stratified samples (see Appendix). Under a series of conditions, Francisco and Fuller (1991) proved the validity of the Bahadur's representation. These conditions are not presented here but they are needed in the following theorem. We refer readers to see Francisco and Fuller's article for details.

Theorem 3 *Under the conditions of Theorem 3 in* Francisco and Fuller (1991)*, if* $j = 1, \ldots, H$*, and* $\eta(t_0) = \frac{\sum_{j=1}^{H}\frac{N_j}{N}E[X_jI(X_j\le\xi_{t_0})]}{\sum_{j=1}^{H}\frac{N_j}{N}E(X_j)}$ *for a given* $t_0 \in (0,1)$*, then the limiting distribution of* $l_2(\eta(t_0))$ *is a scaled Chi-square distribution with degree of freedom* 1*. That is,* $r_2 l_2(\eta(t_0)) \xrightarrow{\mathcal{L}} \chi_1^2$*, where* $r_2 = \frac{\psi_p^2(t_0)}{\psi_d^2(t_0)}$ *with* $\psi_p^2(t_0) = \sum_{j=1}^{H}\rho_j Var[X_jI(X_j\le\xi_{t_0})]$, $\psi_d^2(t_0) = \sum_{j=1}^{H}\rho_j Var[(X_j-\xi_{t_0})I(X_j\le\xi_{t_0})]$.

Remark 1 To calculate $l_2(\eta(t_0))$, existing Splus/R functions like nlmin(g, x, …) can be used to solve the system of equations (11)–(12) for a given $\eta(t_0)$. We can also use the following algorithm to compute it: (i) Choose an initial value of $\widetilde{\nu}(t_0) = 0$, (ii) solve (11) for $\widetilde{\eta}_1(t_0), \ldots, \widetilde{\eta}_H(t_0)$, and then compute $\eta_{new}(t_0) = \frac{\sum_{j=1}^{H}\frac{N_j}{N}\widetilde{\eta}_j(t_0)}{\sum_{j=1}^{H}\frac{N_j}{N}\bar{X}_j}$, (iii) if $|\eta_{new}(t_0) - \eta(t_0)| > \epsilon$ (ϵ is a pre-selected small value, e.g., $\epsilon = 0.001$), then update $\widetilde{\nu}(t_0)$ and go to step (ii), otherwise use (4) to find the value of $l_2(\eta(t_0))$.

Theorem 4 *Assume the same conditions as those in Theorem* 3*. Then the EL-based Lorenz process* $\{l_2(\eta(t)) : t \in [0,1]\}$ *converges to* $W^2(t)/\psi_p^2(t)$ *in distribution, where* $W(t)$ *is a Gaussian process with mean zero and the following covariance function:*

$$\text{Cov}(W(t), W(s)) = \sum_{j=1}^{H}\rho_j\text{Cov}((X_j-\xi_t)I(X_j\le\xi_t), (X_j-\xi_s)I(X_j\le\xi_s)).$$

4 Confidence intervals/bands for the Lorenz curve and simulation studies

In this section, we construct confidence intervals for the Lorenz ordinate $\eta(t_0)$ with fixed $t_0 \in (0,1)$ and confidence bands for the Lorenz curve under simple random sampling design, respectively. Simulation studies are also conducted to evaluate finite sample performances of these intervals and bands.

4.1 Confidence intervals for the Lorenz ordinate

4.1.1 Normal approximation and bootstrap-based intervals for $\eta(t_0)$

It is well known that the estimate $\widehat{\eta}(t_0)$ is asymptotically normal with variance $s_d^2(t_0)$ (see Zheng 2002). i.e., $\sqrt{n}(\widehat{\eta}(t_0)-\eta(t_0)) \longrightarrow \mathcal{N}(0, s_d^2(t_0))$. So, a $(1-\alpha)$ level normal approximation (NA)-based confidence interval for $\eta(t_0)$ can be constructed as follows:

$$(l_1, u_1) = (\widehat{\eta}(t_0) - z_{1-\frac{\alpha}{2}}\widehat{s}_d(t_0)/\sqrt{n},\ \widehat{\eta}(t_0) + z_{1-\frac{\alpha}{2}}\widehat{s}_d(t_0)/\sqrt{n}),$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution, and $\widehat{s}_d^2(t_0) = \int_0^\infty[(x-\widehat{\xi}_{t_0})I(x \le \widehat{\xi}_{t_0}) - x\widehat{\eta}(t_0)]^2 \mathrm{d}F_n(x) - (t_0\widehat{\xi}_{t_0})^2$.

Since the estimation of the asymptotic variance of $\widehat{\eta}(t_0)$ involves in estimation of unknown quantile and Lorenz ordinate as well as distribution function, the NA-based interval may have poor small sample performance particularly when the underlying income distribution is skewed or have outliers. Instead, bootstrap-based methods could be useful alternatives for the interval estimation of Lorenz ordinates. Let $\{X_1^*, \ldots, X_n^*\}$ be a bootstrap re-sample from the original data. The bootstrap version of $\widehat{\eta}(t_0)$ is:

$$\widehat{\eta}^*(t_0) = \frac{\sum_{i=1}^n X_i^* I(X_i^* \le \widehat{\xi}_{t_0}^*)}{\sum_{i=1}^n X_i^*}, \quad \text{where } \widehat{\xi}_{t_0}^* \text{ is the } t_0\text{th sample quantile of } X_i^{*'}s.$$

After repeatedly drawing bootstrap re-samples from the original data, we generate K bootstrap copies of $\widehat{\eta}(t_0)$: $\{\widehat{\eta}_k^*(t_0) : k = 1, 2, \ldots, K\}$, where $K \ge 200$ is recommended. The asymptotic variance of $\widehat{\eta}(t_0)$ can be estimated by

$$V^* = \frac{1}{K-1}\sum_{k=1}^K(\widehat{\eta}_k^*(t_0) - \bar{\eta}^*(t_0))^2, \quad \text{where } \bar{\eta}^*(t_0) = \frac{1}{K}\sum_{k=1}^K \widehat{\eta}_k^*(t_0).$$

Based on this bootstrap variance estimate, two $(1-\alpha)$ level confidence intervals (called BV1 and BV2) for $\eta(t_0)$ can be constructed as follows:

(i) BV1 interval: $(l_2, u_2) = (\widehat{\eta}(t_0) - z_{1-\alpha/2}V^{*\frac{1}{2}},\ \widehat{\eta}(t_0) + z_{1-\alpha/2}V^{*\frac{1}{2}})$.

(ii) BV2 interval: $(l_3, u_3) = (\bar{\eta}^*(t_0) - z_{1-\alpha/2}V^{*\frac{1}{2}},\ \bar{\eta}^*(t_0) + z_{1-\alpha/2}V^{*\frac{1}{2}})$.

Another confidence interval for $\eta(t_0)$ is the *bootstrap bias correction and acceleration* (BCa) interval defined as follows:

(iii) BCa interval: $(l_4, u_4) = (\widehat{\eta}^*_{([K\beta_1])}(t_0),\ \widehat{\eta}^*_{([K\beta_2])}(t_0))$, where

$$\beta_1 = \Phi\left(q + \frac{q + z_{\alpha/2}}{1 - p(q + z_{\alpha/2})}\right), \quad \beta_2 = \Phi\left(q + \frac{q + z_{1-\alpha/2}}{1 - p(q + z_{1-\alpha/2})}\right),$$

with $p = \frac{1}{6}\sum_{j=1}^n \varphi_j^3/(\sum_{j=1}^n \varphi_j^2)^{\frac{3}{2}}, q = \Phi^{-1}(\frac{1}{K}\sum_{k=1}^K I(\widehat{\eta}_k^*(t_0) \le \widehat{\eta}(t_0))), \varphi_j = \widehat{\eta}_{(.)}(t_0) - \widehat{\eta}_{(-j)}(t_0)$, and $\widehat{\eta}_{(-j)}(t_0)$ is the $\widehat{\eta}(t_0)$ computed by deleting the jth observation in original data, and $\widehat{\eta}_{(.)}(t_0) = \frac{1}{n}\sum_{j=1}^n \widehat{\eta}_{(-j)}(t_0)$.

4.1.2 Empirical likelihood-based confidence intervals for $\eta(t_0)$

By Theorem 1, the EL interval for $\eta(t_0)$ can be constructed as follows.

$$(l_5, u_5) = \{\eta(t_0) : \widehat{r}_1 l_1(\eta(t_0)) \le \chi^2_{1,1-\alpha}\},$$

where $\chi^2_{1,1-\alpha}$ is the $(1-\alpha)$th quantile of Chi-square distribution with degree of freedom one, and $\widehat{r}_1 = \widehat{s^2_p}(t_0)/\widehat{s^2_d}(t_0)$ is a plug-in estimate for r_1 with $\widehat{s}^2_p(t_0) = \int_0^\infty \{x[I(x \leq \widehat{\xi}_{t_0}) - \widehat{\eta}(t_0)]\}^2 \mathrm{d}F_n(x), \widehat{s}^2_d(t_0) = \int_0^\infty [(x - \widehat{\xi}_{t_0})I(x \leq \widehat{\xi}_{t_0}) - x\widehat{\eta}(t_0)]^2 \mathrm{d}F_n(x) - (t_0\widehat{\xi}_{t_0})^2$.

Hybrid bootstrap and empirical likelihood (HBEL) approach has been introduced in statistical literature to produce confidence intervals for unknown parameters (see Chen et al. 2003). The EL theory developed in Sect. 2 can be employed to construct HBEL intervals for $\eta(t_0)$. We summarize the procedure in the following steps:

1. Draw a bootstrap sample of size n, X_i^*s, with replacement from the sample X_is.
2. Calculate the bootstrap versions of $\widehat{s}^2_p(t_0), \widehat{s}^2_d(t_0)$ and $l_1(\eta(t_0))$, respectively: $\widehat{s}^{*2}_p(t_0) = \int_0^\infty \{x[I(x \leq \widehat{\xi}^*_{t_0}) - \widehat{\eta}^*(t_0)]\}^2 \mathrm{d}F^*_n(x), \widehat{s}^{*2}_d(t_0) = \int_0^\infty [(x - \widehat{\xi}^*_{t_0})I(x \leq \widehat{\xi}^*_{t_0}) - x\widehat{\eta}^*(t_0)]^2 \mathrm{d}F^*_n(x) - (t_0\widehat{\xi}^*_{t_0})^2, l^*_1(\widehat{\eta}(t_0)) = 2\sum_{i=1}^n \log\{1 + \nu^*\widehat{D}^*_i(t_0)\}$, where F^*_n is the empirical distribution of X^*_i's, $\widehat{\xi}^*_{t_0}$ is the t_0th quantile of F^*_n, $\widehat{D}^*_i(t_0) = X^*_i[I(X^*_i \leq \widehat{\xi}^*_{t_0}) - \widehat{\eta}(t_0)]$, and ν^* is the solution to $\frac{1}{n}\sum_{i=1}^n \frac{\widehat{D}^*_i(t_0)}{1+\nu^*\widehat{D}^*_i(t_0)} = 0$.
3. Repeat the first two steps K times to obtain three sets of bootstrap replications: $\{\widehat{s}^{*2}_{p,k}(t_0) : k = 1, \ldots, K\}, \{\widehat{s}^{*2}_{d,k}(t_0) : k = 1, \ldots, K\}$, $\{l^*_{1,k}(\widehat{\eta}(t_0)) : k = 1, \ldots, K\}$ (it is recommended that $K \geq 200$; in this paper, we take $K = 300$).

Two new HBEL intervals for $\eta(t_0)$ are defined as follows.

(i) HBEL1 interval: $(l_6, u_6) = \{\eta(t_0) : l_1(\eta(t_0)) \leq l^*_{1,([K(1-\alpha)])}(\widehat{\eta}(t_0))\}$, where $l^*_{1,(k)}(\widehat{\eta}(t_0))$ is the kth ordered value of $l^*_{1,k}(\widehat{\eta}(t_0))$'s.

(ii) HBEL2 interval: $(l_7, u_7) = \{\eta(t_0) : \widehat{r}_1 l_1(\eta(t_0)) \leq L^*_{([K(1-\alpha)])}(\widehat{\eta}(t_0))\}$, where $\widehat{r}_1 = \widehat{s}^2_p(t_0)/\widehat{s}^2_d(t_0)$, and $L^*_{(k)}$ is the kth ordered value of $\{L^*_k = \frac{\widehat{s}^{*2}_{p,k}}{\widehat{s}^{*2}_{d,k}} l^*_{1,k}(\widehat{\eta}(t_0)) : k = 1, \ldots, K\}$.

4.2 Empirical likelihood confidence band for the Lorenz curve

Based on Theorem 2, we can construct an asymptotic $100(1-\alpha)\%$ confidence band for $\eta(t)$ on $[0, 1]$: $\Re = \{(t, \eta(t)) : l_1(\eta(t)) \leq c_\alpha$ for $t \in [0, 1]\}$, where c_α is the upper αth quantile of the limiting process $\{J^2(t)/s^2_p(t) : t \in [0, 1]\}$ of the EL-based Lorenz process $\{l_1(\eta(t)) : t \in [0, 1]\}$. To find c_α, we suggest using the bootstrap distribution of $l^*_1(\widehat{\eta}(t))$ to approximate the distribution of $l_1(\eta(t))$ (see also Hall and Owen 1993). Similar to the procedure used in the previous section, we can generate a large number of bootstrap copies $\{l^*_{1,k}(\widehat{\eta}(t)) : k = 1, \ldots, K\}$. Then c_α is approximately the upper αth quantile of $\{\sup_{t\in[0,1]} l^*_{1,1}(\widehat{\eta}(t)), \ldots, \sup_{t\in[0,1]} l^*_{1,K}(\widehat{\eta}(t))\}$.

4.3 Simulation studies

Seven confidence intervals for the Lorenz ordinate $\eta(t_0)$ have been proposed in Sect. 4.1. In order to compare finite sample performances of these intervals, simulation studies are conducted in this section.

In the simulation studies, the Weibull distribution with the shape parameter $a = 1$ and the scale parameter $b = 2$ is chosen to be the underlying income distribution

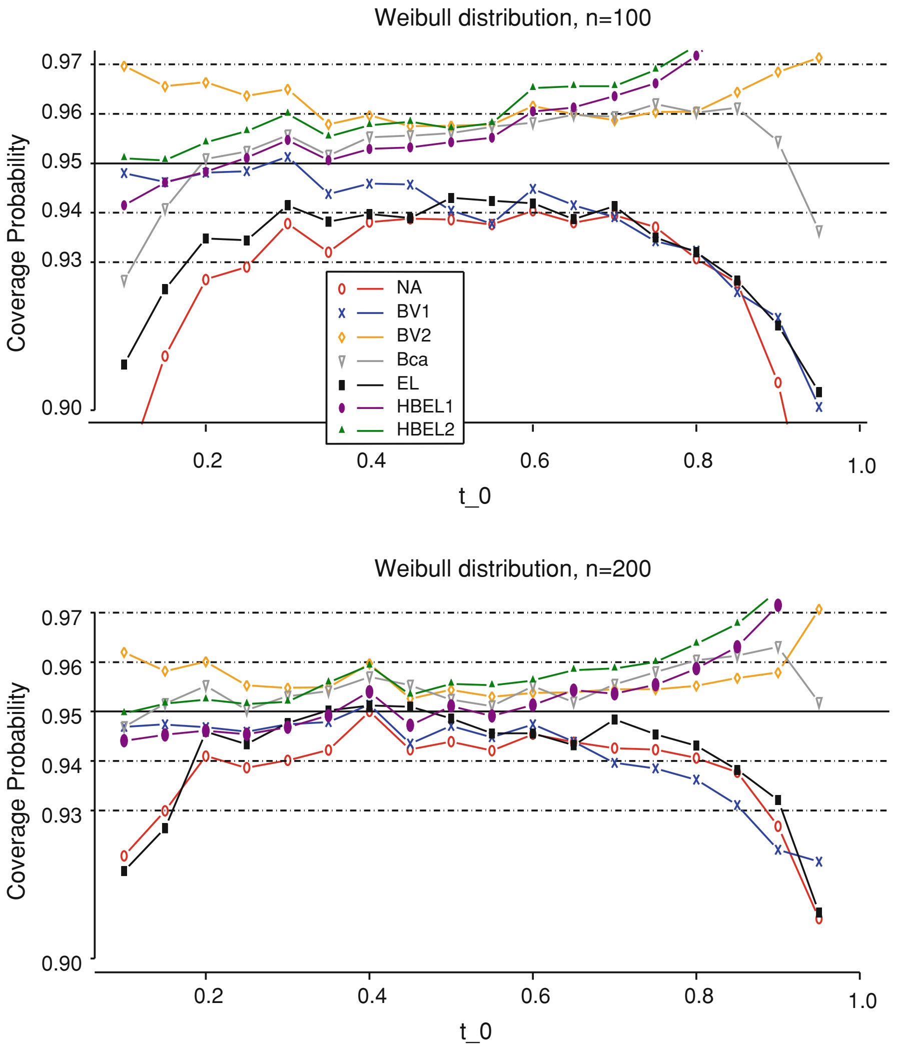

Fig. 1 Coverage probabilities of the 95% confidence intervals for the Lorenz ordinates

$F(x)$, respectively. The sample size n is chosen to be 100, 200, respectively. Under these simulation settings, we generate 10,000 random samples of size n from $F(x)$, and calculate the coverage probabilities and average interval lengths of 95% level confidence intervals for $\eta(t_0)$ at different t_0 using the simulated data. In the computation of bootstrap and HBEL intervals, we draw $K = 300$ bootstrap re-samples from the original samples.

Figures 1 and 2 display the simulation results. From these figures, we observe that the coverage probabilities of all the intervals are closer to the nominal level as sample size increases. When t_0 falls in both the lower and the upper tails of the Lorenz curve, the coverage probabilities of the NA intervals are much lower than the nominal level

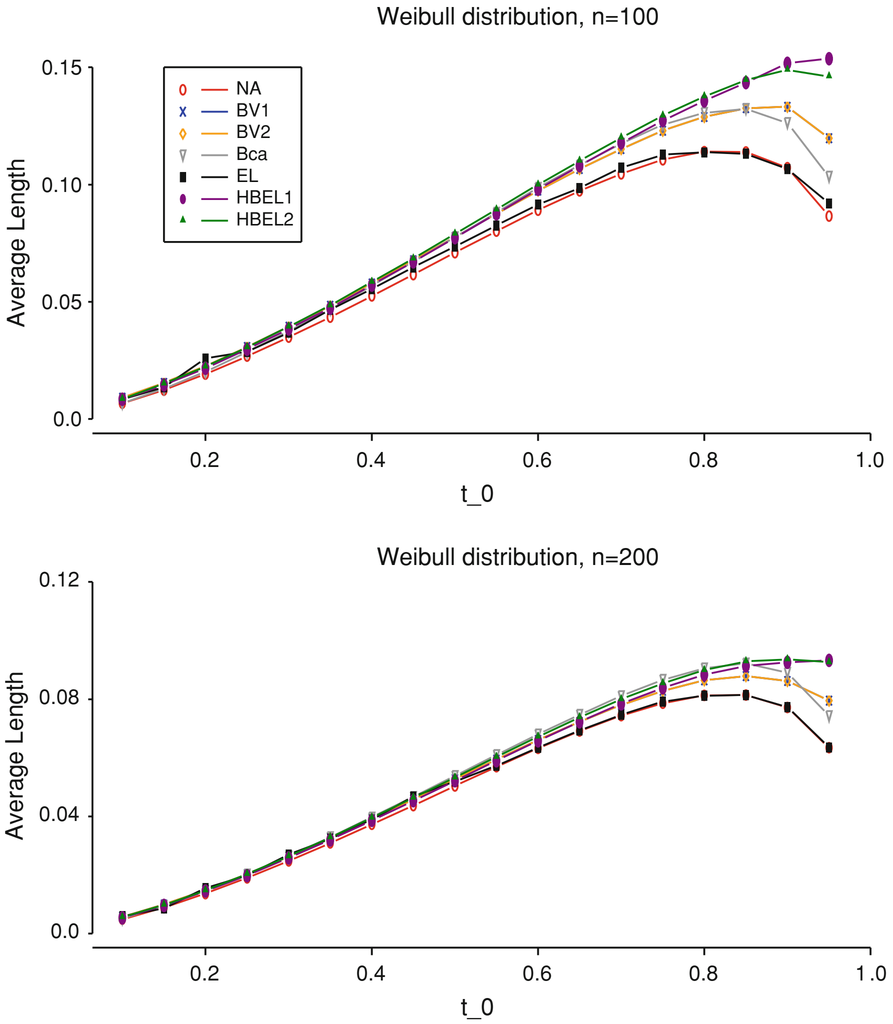

Fig. 2 Average length of the 95% confidence intervals for the Lorenz ordinates

particularly when sample size is small ($n = 100$). The EL intervals have coverage probabilities closer to the nominal level than those of the NA intervals. However, the performances of the EL intervals are not stable due to the plug-in estimate of the scale constant. The HBEL1, HBEL2, BCa, and BV2 intervals outperform the other intervals in most cases considered here. All the intervals have similar lengths except in the right tails where the EL and NA intervals have shorter length. Therefore, we recommend the use of the HBEL1, HBEL2, BCa, and BV2 intervals for the Lorenz ordinate when income data are skewed data.

To illustrate finite sample performances of the EL-based confidence band defined in Sect. 4.2, we plot the asymptotic $100(1-\alpha)\%$ confidence bands for $\eta(t)$ on $[0, 1]$. To calculate the critical value c_α, we draw $K = 5{,}000$ bootstrap re-samples from the

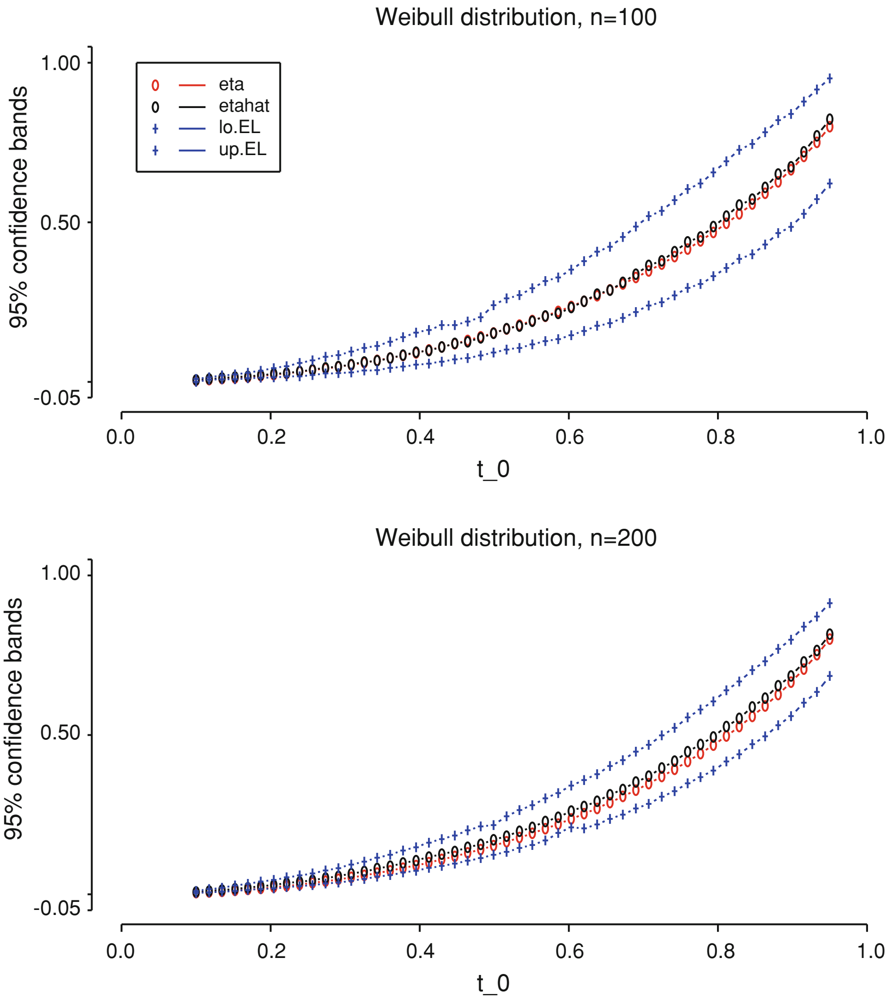

Fig. 3 The 95% empirical likelihood-based confidence bands for the Lorenz curve. eta = true value of $\eta(t)$, etahat = estimated value of $\eta(t)$, lo EL = lower confidence band, up EL = upper confidence band

original samples. Figure 3 displays the confidence bands. We can see that the bands cover the true Lorenz curve almost everywhere and the widths of the bands decrease as sample size increases.

5 Application to real data

The Panel Study of Income Dynamics (PSID) is a longitudinal survey of men, women, children, and families in the USA. Since 1968, the PSID has been conducted at the University of Michigan Survey Research Center. It has annually collected information on US families and to date, approximately 37,500 individuals have been interviewed.

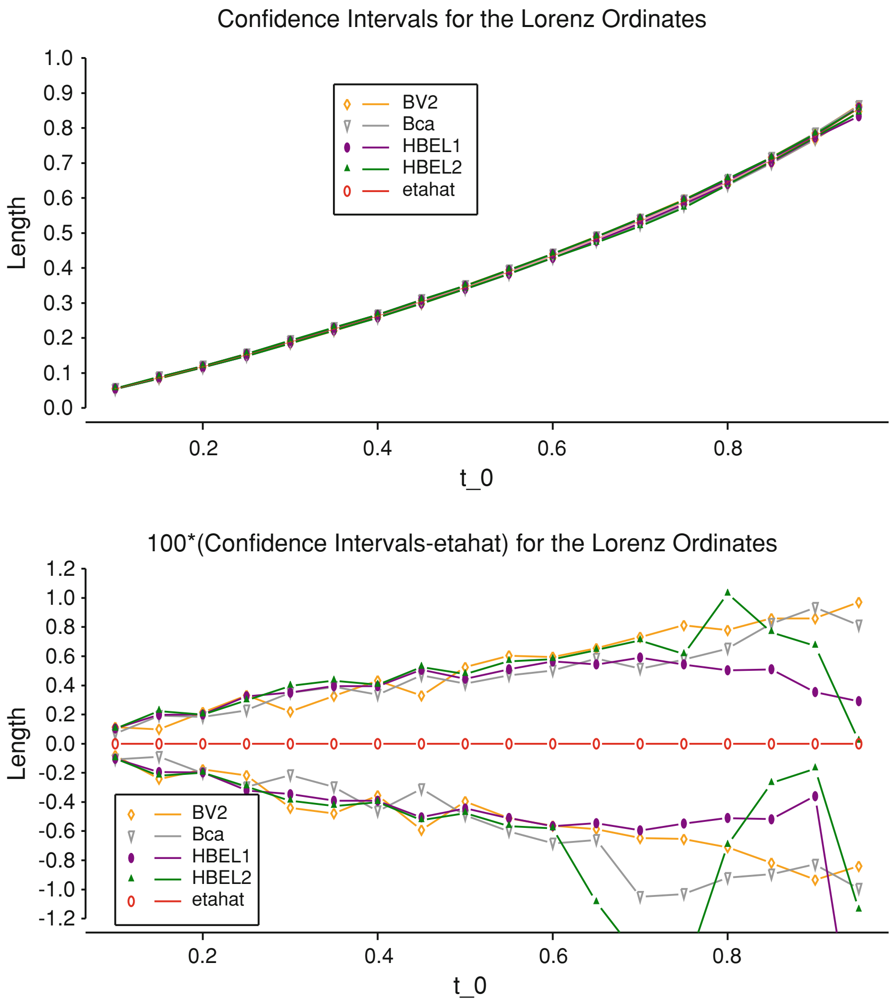

Fig. 4 95% confidence intervals for the Lorenz ordinates in real data example

The PSID User Guide notes that one commendable aspect of their data lies in the fact that adults are followed as they grow older, and children are observed as they become adults and form families of their own. Hill (1992) explained that another feature of the PSID data came from the fact that they initially collected data in order to study the dynamics of poverty; as a result too many low income and Black households were included in the samples. In this paper, we apply our recommended methods to the total family income data of the year 2000. The sample consists of 7,406 families. The non-parametric estimates for the Lorenz ordinates with their 95% confidence intervals are presented in Fig. 4. From this figure, we can see that the proposed intervals have short interval length for any fixed t. The lower panel in the figure is the enlarged graph for $100 * (\text{lower/upper confidence bounds} - \widehat{\eta}(t))$. From the graph, it can be seen that HBEL1 interval is more stable and has generally shorter interval length.

6 Appendix: Proof of Theorems

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

$$\text{(i) } \frac{1}{n}\sum\nolimits_{i=1}^{n} \widehat{D}_i^2(t_0) \xrightarrow{p} s_p^2(t_0), \quad \text{(ii). } \frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n} \widehat{D}_i(t_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, s_d^2(t_0)).$$

Proof (i) For a given t_0, let $D_i(t_0) = X_i(I(X_i \le \xi_{t_0}) - \eta(t_0))$. By the law of large number, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} D_i^2(t_0) &= \frac{1}{n}\sum_{i=1}^{n}[X_i(I(X_i \le \xi_{t_0}) - \eta(t_0))]^2 \xrightarrow{p} E[X(I(X \le \xi_{t_0}) - \eta(t_0))]^2 \\ &= \int_0^\infty \{x[I(x \le \xi_{t_0}) - \eta(t_0)]\}^2 \mathrm{d}F(x) =: s_p^2(t_0).\end{aligned}$$

So, we only need to prove that $I_1(t_0) \equiv \left|\frac{1}{n}\sum_{i=1}^{n} \widehat{D}_i^2(t_0) - \frac{1}{n}\sum_{i=1}^{n} D_i^2(t_0)\right| = o_p(1)$. By the strong consistency of the sample quantile $\widehat{\xi}_{t_0}$, we have that $|I(X_i \le \widehat{\xi}_{t_0}) - I(X_i \le \xi_{t_0})| \xrightarrow{p} 0$, for $i = 1, 2, \ldots, n$. From $\frac{1}{n}\sum_{i=1}^{n} |X_i|^2 \longrightarrow E(X^2) < \infty$ a.s., it follows that

$$\begin{aligned}|I_1(t_0)| &\le \frac{1}{n}\sum_{i=1}^{n} |\widehat{D}_i(t_0) + D_i(t_0)||\widehat{D}_i(t_0) - D_i(t_0)| \\ &\le \frac{2}{n}\sum_{i=1}^{n} |X_i||X_i||I(X_i \le \widehat{\xi}_{t_0}) - I(X_i \le \xi_{t_0})| \xrightarrow{p} 0.\end{aligned}$$

(ii) From the Bahadur representation for the sample quantile $\widehat{\xi}_{t_0}$, following (2.3) and (2.5) in Zheng (2002), i.e., $\widehat{\xi}_t - \xi_t = \frac{t - \frac{1}{n}\sum_{i=1}^{n} I(X_i \le \xi_t)}{f(\xi_t)} + o_p(n^{-\frac{1}{2}})$, and

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} X_i[I(X_i \le \widehat{\xi}_t) - \eta(t)] &= \frac{1}{n}\sum_{i=1}^{n}[(X_i - \xi_t)I(X_i \le \xi_t) + t\xi_t - X_i\eta(t)] \\ &\quad + o_p(n^{-\frac{1}{2}}),\end{aligned}$$

we get that

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \widehat{D}_i(t_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[X_i(I(X_i \le \widehat{\xi}_{t_0}) - \eta(t_0))] \\ &= \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \xi_{t_0})I(X_i \le \xi_{t_0}) + t_0\xi_{t_0} - \frac{1}{n}\sum_{i=1}^{n} X_i\eta(t_0)\right] + o_p(1).\end{aligned}$$

From $E[(X-\xi_{t_0})I(X\le\xi_{t_0})-X\eta(t_0)]=-t_0\xi_{t_0}$, and

$$\begin{aligned}&\mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(X_i-\xi_{t_0})I(X_i\le\xi_{t_0})-X_i\eta(t_0)]\right)\\&=\left(1-\frac{n}{N}\right)\mathrm{Var}[(x-\xi_{t_0})I(x\le\xi_{t_0})-x\eta(t_0)]\\&\to\int_0^{\infty}[(x-\xi_{t_0})I(x\le\xi_{t_0})-x\eta(t_0)]^2\mathrm{d}F(x)-(t_0\xi_{t_0})^2=:s_d^2(t_0),\end{aligned}$$

it follows that $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{D}_i(t_0)\xrightarrow{\mathcal{L}}\mathcal{N}(0,s_d^2(t_0))$. □

The Proof of Theorem 1. Using Lemma 1(i) and the similar method used in Owen (1990), we can prove that $|\nu|=O_p(n^{-1/2})$. By $E|X|^2<\infty$, we have $\max_i|\widehat{D}_i(t_0)|\le C\max_i|X_i|=o(n^{1/2})\ \ a.s.$. Then, applying Taylor's expansion to (6), we have

$$l_1(\eta(t_0))=2\sum_{i=1}^{n}\log\{1+\nu\widehat{D}_i(t_0)\}=2\sum_{i=1}^{n}(\nu\widehat{D}_i(t_0)-\frac{1}{2}(\nu\widehat{D}_i(t_0))^2)+r_{1n}\quad(14)$$

with $|r_{1n}|\le C\sum_{i=1}^{n}|\nu\widehat{D}_i(t_0)|^3\le C|\nu|^3\max_i|\widehat{D}_i(t_0)|\sum_i\widehat{D}_i^2(t_0)=o_p(1)$. From (5), and

$$\begin{aligned}\sum_{i=1}^{n}\frac{\widehat{D}_i(t_0)}{1+\nu\widehat{D}_i(t_0)}&=\sum_{i=1}^{n}\widehat{D}_i(t_0)\left[1-\nu\widehat{D}_i(t_0)+\frac{(\nu\widehat{D}_i(t_0))^2}{1+\nu\widehat{D}_i(t_0)}\right]\\&=\sum_{i=1}^{n}\widehat{D}_i(t_0)-\left(\sum_{i=1}^{n}\widehat{D}_i^2(t_0)\right)\nu+\sum_{i=1}^{n}\frac{\widehat{D}_i(t_0)(\nu\widehat{D}_i(t_0))^2}{1+\nu\widehat{D}_i(t_0)},\end{aligned}$$

it follows that $\nu=\frac{\sum_{i=1}^{n}\widehat{D}_i(t_0)}{\sum_{i=1}^{n}\widehat{D}_i^2(t_0)}+o_p(n^{-1/2})$. Furthermore, we can get that

$$\sum_{i=1}^{n}\nu\widehat{D}_i(t_0)=\sum_{i=1}^{n}(\nu\widehat{D}_i(t_0))^2+o_p(1).\quad(15)$$

Therefore, by Lemma 1, we get that

$$\begin{aligned}r_1l_1(\eta(t_0))&=\frac{s_p^2(t_0)}{s_d^2(t_0)}\sum_{i=1}^{n}(\nu\widehat{D}_i(t_0))^2+o_p(1)\\&=\frac{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{D}_i(t_0)\right)^2}{s_d^2(t_0)}\frac{s_p^2(t_0)}{\frac{1}{n}\sum_{i=1}^{n}\widehat{D}_i^2(t_0)}+o_p(1)=\chi_1^2+o_p(1).\end{aligned}$$

□

Lemma 2 *Under the conditions in Theorem 2, we have*

(i). $\frac{1}{n}\sum_{i=1}^{n}\widehat{D}_i^2(t)\xrightarrow{p}s_p^2(t)$, uniformly on [0, 1]. (ii). $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{D}_i(t)\xrightarrow{\mathcal{L}}J(t)$,

where $J(t)$ is a Gaussian process defined in Theorem 2.

Proof (i) Since $\{X_iI(X_i\le\xi_t):t\in[0,1]\}$ and $\{X_i^2I(X_i\le\xi_t):t\in[0,1]\}, i=1,\ldots,n$, are two sequences of manageable processes with integrable envelops $\{|X_i|, i=1,2,\ldots,n\}$ and $\{X_i^2, i=1,\ldots,n\}$, respectively, by the uniform law of large number (Pollard 1990), we get that

$$\frac{1}{n}\sum_{i=1}^{n}D_i^2(t)\xrightarrow{p}s_p^2(t),\ \text{uniformly on } [0, 1]. \tag{16}$$

Note that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\widehat{D}_i^2(t)-\frac{1}{n}\sum_{i=1}^{n}D_i^2(t)\right|=o_p(1),\ \text{uniformly on } [0, 1]. \tag{17}$$

Then, Lemma 2(i) follows immediately from (16) and (17).

(ii) From Csörgö et al. (1986), we get that

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{D}_i(t)&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i[I(X_i\le\widehat{\xi}_t)-\eta(t)]\\&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(X_i-\xi_t)I(X_i\le\xi_t)+t\xi_t-X_i\eta(t)]+o_p(1)\\&\xrightarrow{\mathcal{L}}J(t)=\frac{1}{\mu}\left[\int_0^{\xi_t}B(F(x))dx-\eta(t)\int_0^{\infty}B(F(x))\mathrm{d}x\right],\end{aligned}$$

where $J(t)$ is the Gaussian process defined in Theorem 2. □

The Proof of Theorem 2. By Lemma 2, following the same lines as the proof of Theorem 1, we have that

$$\begin{aligned}l_1(\eta(t))&=\sum_{i=1}^{n}(\nu\widehat{D}_i(t))^2+o_p(1)\\&=\frac{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{D}_i(t)\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\widehat{D}_i^2(t)}+o_p(1)\xrightarrow{\mathcal{L}}J^2(t)/s_p^2(t),\end{aligned}$$

where $o_p(1)$ is uniformly on [0,1]. □

Lemma 3 *Under the conditions in Theorem 3, we have*

(i) $\frac{1}{n_j}\sum_{k=1}^{n_j}\widetilde{D}_{jk}^2(t_0)\xrightarrow{p} Var[X_j(I(X_j\le\xi_{t_0})],\ j=1,2,\ldots,H.$

(ii) $\frac{1}{\sqrt{n}}\sum_{j=1}^{H}\sum_{k=1}^{n_j}\widetilde{D}_{jk}(t_0)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\psi_d^2(t_0)).$

Proof (i) Let $D_{jk}(t_0)=X_{jk}I(X_{jk}\le\xi_{t_0})-E[X_jI(X_j\le\xi_{t_0})]$. By the law of large number, we have that $\frac{1}{n_j}\sum_{k=1}^{n_j}X_{jk}I(X_{jk}\le\xi_{t_0})\xrightarrow{p}E[X_jI(X_j\le\xi_{t_0})]$, and

$$\frac{1}{n_j}\sum_{k=1}^{n_j}D_{jk}^2(t_0)=\frac{1}{n_j}\sum_{k=1}^{n_j}\{X_{jk}I(X_{jk}\le\xi_{t_0})-E[X_jI(X_j\le\xi_{t_0})]\}^2$$
$$\xrightarrow{p}\mathrm{Var}[X_jI(X_j\le\xi_{t_0})]. \tag{18}$$

Lemma (i) follows from (18) and $\left|\frac{1}{n_j}\sum_{k=1}^{n_j}\widetilde{D}_{jk}^2(t_0)-\frac{1}{n_j}\sum_{k=1}^{n_j}D_{jk}^2(t_0)\right|=o_p(1)$.

(ii) Under the conditions of Theorem 3 in Francisco and Fuller (1991), the following Bahadur's representation for the sample quantile $\widehat{\xi}_{t_0}$ with stratified random sample is still valid (see (2.9) and (2.10) in Zheng (2002)), i.e.,

$$\widehat{\xi}_t-\xi_t=\frac{t-\sum_{j=1}^{H}\frac{N_j}{N}\left[\frac{1}{n_j}\sum_{k=1}^{n_j}I(X_{jk}\le\xi_t)\right]}{f(\xi_t)}+o_p(n^{-\frac{1}{2}}),$$
$$\sum_{j=1}^{H}\frac{N_j}{N}\frac{1}{n_j}\sum_{k=1}^{n_j}X_{jk}I(X_{jk}\le\widehat{\xi}_{t_0})$$
$$=\sum_{j=1}^{H}\frac{N_j}{N}\frac{1}{n_j}\sum_{k=1}^{n_j}(X_{jk}-\xi_{t_0})I(X_{jk}\le\xi_{t_0})+t_0\xi_{t_0}+o_p(n^{-1/2}).$$

From $\frac{1}{n}\sum_{j=1}^{H}\sum_{k=1}^{n_j}\widetilde{D}_{jk}(t_0)=\sum_{j=1}^{H}\frac{N_j}{N}\frac{1}{n\,j}\sum_{k=1}^{n_j}\widetilde{D}_{jk}(t_0)$, we get that

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{H}\sum_{k=1}^{n_j}\widetilde{D}_{jk}(t_0)=\sqrt{n}\sum_{j=1}^{H}\frac{N_j}{N}\left(\frac{1}{n_j}\sum_{k=1}^{n_j}X_{jk}I(X_{jk}\le\widehat{\xi}_{t_0})-E[X_jI(X_j\le\xi_{t_0})]\right)$$
$$=\sqrt{n}\sum_{j=1}^{H}\frac{N_j}{N}\left(\frac{1}{n_j}\sum_{k=1}^{n_j}(X_{jk}-\xi_{t_0})I(X_{jk}\le\xi_{t_0})\right.$$
$$\left.-E[X_jI(X_j\le\xi_{t_0})]+t_0\xi_{t_0}\right)+o_p(1).$$

From $E(\frac{1}{n_j}\sum_{k=1}^{n_j}X_{jk}I(X_{jk}\le\xi_{t_0}))=E[X_jI(X_j\le\xi_{t_0})]$, $E(\frac{1}{n_j}\sum_{k=1}^{n_j}I(X_{jk}\le\xi_{t_0})\xi_{t_0})=F_j(\xi_{t_0})\xi_{t_0}$, $E(\sum_{j=1}^{H}\frac{N_j}{N}\frac{1}{n_j}\sum_{k=1}^{n_j}I(X_{jk}\le\xi_{t_0})\xi_{t_0})=t_0\xi_{t_0}$, and

$$\begin{aligned}
&\mathrm{Var}\left(\sqrt{n}\sum_{j=1}^{H}\frac{N_j}{N}\frac{1}{n_j}\sum_{k=1}^{n_j}\widetilde{D}_{jk}(t_0)\right)\\
&\quad=\mathrm{Var}\left(\sqrt{n}\sum_{j=1}^{H}\frac{N_j}{N}\frac{1}{n_j}\sum_{k=1}^{n_j}(X_{jk}-\xi_{t_0})I(X_{jk}\le\xi_{t_0})\right)\\
&\quad=\sum_{j=1}^{H}\left(\frac{N_j}{N}\right)^2\frac{n}{n_j}\frac{N_j-n_j}{N_j}\mathrm{Var}[(X_j-\xi_{t_0})I(X_j\le\xi_{t_0})]\\
&\quad\to\sum_{j=1}^{H}\rho_j\mathrm{Var}[(X_j-\xi_{t_0})I(X_j\le\xi_{t_0})]=:\psi_d^2(t_0),
\end{aligned}$$

it follows that $\frac{1}{\sqrt{n}}\sum_{j=1}^{H}\sum_{k=1}^{n_j}\widetilde{D}_{jk}(t_0)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\psi_d^2(t_0))$. □

The Proof of Theorem 3. Using Lemma 3(i) and the similar method used in Owen (1990), we can prove that $|\widetilde{\nu}(t_0)|=O_p(n^{-1/2})$. By $E|X_j|^2<\infty$, $j=1,2,\ldots,H$, we have $\max_k|X_{jk}|=o(n_j^{1/2})=o(n^{1/2})$ *a.s.* for $j=1,2,\ldots,H$. Hence, $\max_k|\widetilde{D}_{jk}(t_0)|\le C\max_k|X_{jk}|=o(n^{1/2})$ *a.s.*, $j=1,2,\ldots,H$. Using Taylor's expansion to (4), we get

$$\begin{aligned}
l_2(\eta(t_0))&=2\sum_{j=1}^{H}\sum_{k=1}^{n_j}\log(1+\widetilde{\nu}(t_0)\widetilde{D}_{jk}(t_0))\\
&=2\sum_{j=1}^{H}\sum_{k=1}^{n_j}\left[\widetilde{\nu}(t_0)\widetilde{D}_{jk}(t_0)-\frac{1}{2}(\widetilde{\nu}(t_0)\widetilde{D}_{jk}(t_0))^2\right]+r_{2n},\qquad(19)
\end{aligned}$$

where

$$|r_{2n}|\le C\sum_{j=1}^{H}\sum_{k=1}^{n_j}\left|\widetilde{\nu}(t_0)\widetilde{D}_{jk}(t_0)\right|^3\le Cn^{-3/2}\max_{j,k}|\widetilde{D}_{jk}(t_0)|\sum_{j=1}^{H}\sum_{k=1}^{n_j}\widetilde{D}^2_{jk}(t_0)=o_p(1).$$

By (11) and Lemma 3(i), similar to the proof of (15), we have

$$\sum_{k=1}^{n_j}\widetilde{\nu}(t_0)\widetilde{D}_{jk}(t_0)=\sum_{k=1}^{n_j}(\widetilde{\nu}(t_0)\widetilde{D}_{jk}(t_0))^2+o_p(1),\ \text{ for }\ j=1,2,\ldots,H.$$

Hence, $l_2(\eta(t_0))=\sum_{j=1}^{H}\sum_{k=1}^{n_j}\widetilde{\nu}(t_0)\widetilde{D}_{jk}(t_0)+o_p(1)$. By (11), and

$$\begin{aligned}
0&=\frac{1}{n_j}\sum_{k=1}^{n_j}\frac{\widetilde{D}_{jk}(t_0)}{1+\widetilde{\nu}(t_0)\widetilde{D}_{jk}(t_0)}\\
&=\frac{1}{n_j}\sum_{k=1}^{n_j}(X_{jk}I(X_{jk}\le\widehat{\xi}_{t_0})-\widetilde{\eta}_j(t_0))-\frac{\widetilde{\nu}(t_0)}{n_j}\sum_{k=1}^{n_j}\widetilde{D}^2_{jk}(t_0)+o_p(n_j^{-1/2}),
\end{aligned}$$

we get that

$$\begin{aligned}\widetilde{\nu}(t_0)\sum_{j=1}^{H}\frac{N_j}{N}\frac{1}{n_j}\sum_{k=1}^{n_j}\widetilde{D}_{jk}^2(t_0) &= \sum_{j=1}^{H}\frac{N_j}{N}\frac{1}{n_j}\sum_{k=1}^{n_j}(X_{jk}I(X_{jk}\le\widehat{\xi}_{t_0})-\widetilde{\eta}_j(t_0))\\ &\quad+o_p(n^{-1/2}),\\ &= \sum_{j=1}^{H}\frac{N_j}{N}\frac{1}{n_j}\sum_{k=1}^{n_j}\widetilde{D}_{jk}(t_0)+o_p(n^{-1/2}) \end{aligned}\tag{20}$$

Observing that $\eta(t_0)=\frac{\sum_{j=1}^{H}\frac{N_j}{N}\widetilde{\eta}_j(t_0)}{\sum_{j=1}^{H}\frac{N_j}{N}\bar{X}_j}$, it follows from (20) and $\frac{N_j}{N}=\frac{n_j}{n}$ that

$$\begin{aligned}\widetilde{\nu}(t_0) &= \frac{\sum_{j=1}^{H}\frac{N_j}{N}\frac{1}{n_j}\sum_{k=1}^{n_j}X_{jk}(I(X_{jk}\le\widehat{\xi}_{t_0})-\eta(t_0))}{\sum_{j=1}^{H}\frac{N_j}{N}\frac{1}{n_j}\sum_{k=1}^{n_j}\widetilde{D}_{jk}^2(t_0)}+o_p(n^{-1/2})\\ &= \frac{\frac{1}{n}\sum_{j=1}^{H}\sum_{k=1}^{n_j}\widetilde{D}_{jk}(t_0)}{\sum_{j=1}^{H}\frac{N_j}{N}\frac{1}{n_j}\sum_{k=1}^{n_j}\widetilde{D}_{jk}^2(t_0)}+o_p(n^{-1/2}).\end{aligned}$$

Therefore,

$$l_2(\eta(t_0))=\frac{\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{H}\sum_{k=1}^{n_j}\widetilde{D}_{jk}(t_0)\right)^2}{\sum_{j=1}^{H}\frac{N_j}{N}\frac{1}{n_j}\sum_{k=1}^{n_j}\widetilde{D}_{jk}^2(t_0)}+o_p(1).\tag{21}$$

By Lemma 3(i)–(ii), and (21), we get that

$$\begin{aligned}r_2l_2(\eta_0(t_0)) &= \left[\frac{\frac{1}{\sqrt{n}}\sum_{j=1}^{H}\sum_{k=1}^{n_j}\widetilde{D}_{jk}(t_0)}{\psi_d(t_0)}\right]^2\frac{\psi_p^2(t_0)}{\sum_{j=1}^{H}\frac{n_j}{n}\frac{1}{n_j}\sum_{k=1}^{n_j}\widetilde{D}_{jk}^2(t_0)}+o_p(1)\\ &= \chi_1^2+o_p(1).\end{aligned}$$

□

Lemma 4 *Under the conditions in Theorem 4, we have*

(i) $\frac{1}{n_j}\sum_{k=1}^{n_j}\widetilde{D}_{jk}^2(t)\xrightarrow{p}\mathrm{Var}(X_j(I(X_j\le\xi_t))$, *uniformly on* [0,1], *for* $j=1,\ldots,H$.

(ii) $\frac{1}{\sqrt{n}}\sum_{j=1}^{H}\sum_{k=1}^{n_j}\widetilde{D}_{jk}(t)\xrightarrow{\mathcal{L}}W(t)$,

where $W(t)$ is a Gaussian process defined in Theorem 4.

Proof The proof of Lemma 4(i) is similar to those of Lemma 3(i) and Lemma 2(i), hence omitted here.

For the proof of Lemma 4(ii), using the similar methods to the proofs of Lemma 3(ii) and Lemma 2(ii), we get that

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{j=1}^{H}\sum_{k=1}^{n_j}\widetilde{D}_{jk}(t)\\&=\sqrt{n}\sum_{j=1}^{H}\frac{N_j}{N}\left(\frac{1}{n_j}\sum_{k=1}^{n_j}(X_{jk}-\xi_t)I(X_{jk}\le\xi_t)-E[X_jI(X_j\le\xi_t)]+t\xi_t\right)\\&\quad+o_p(1)\xrightarrow{\mathcal{L}}W(t),\end{aligned}$$

where $W(t)$ is a Gaussian process with mean zero and the following covariance function:

$$\begin{aligned}&\mathrm{Cov}\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{H}\sum_{k=1}^{n_j}\widetilde{D}_{jk}(t),\quad\frac{1}{\sqrt{n}}\sum_{j=1}^{H}\sum_{k=1}^{n_j}\widetilde{D}_{jk}(s)\right)\\&=\mathrm{Cov}\left(\sqrt{n}\sum_{j=1}^{H}\frac{N_j}{N}\frac{1}{n_j}\sum_{k=1}^{n_j}(X_{jk}-\xi_t)I(X_{jk}\le\xi_t),\right.\\&\quad\left.\times\sqrt{n}\sum_{j=1}^{H}\frac{N_j}{N}\frac{1}{n_j}\sum_{k=1}^{n_j}(X_{jk}-\xi_s)I(X_{jk}\le\xi_s)\right)\\&=\sum_{j=1}^{H}\left(\frac{N_j}{N}\right)^2\frac{n}{n_j}\left(1-\frac{n_j}{N_j}\right)\mathrm{Cov}((X_j-\xi_t)I(X_j\le\xi_t),(X_j-\xi_s)I(X_j\le\xi_s))\\&\quad\to\sum_{j=1}^{H}\rho_j\mathrm{Cov}((X_j-\xi_t)I(X_j\le\xi_t),(X_j-\xi_s)I(X_j\le\xi_s))\\&\quad=:\mathrm{Cov}(W(t),W(s)).\end{aligned}$$

□

The Proof of Theorem 4. Using Lemma 4 and the similar methods in the proof of Theorem 2, we can easily get Theorem 4. □

Acknowledgments We would like to thank the associate editor and referees for their valuable comments which led to substantial improvements in this article. Dr. Gengsheng Qin's work is supported in part by US NSA Grant H98230-12-1-0228.

References

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

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

Belinga-Hill, N. (2007). Empirical likelihood confidence intervals for generalized Lorenz curve. *Master thesis at Georgia State University, 38*, Atlanta, GA, USA.

Chang, R. K. R., Halfon, N. (1997). Graphical distribution of pediatricians in the United States: An analysis of the fifty states and Washington, DC. *Pediatrics, 100*, 172–179.

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

Chen, S. X., Qin, J. (2003). Empirical likelihood-based confidence intervals data with possible zero observations. *Statistics & Probability Letters, 65*, 29–37.

Chen, S. X., Leung, H. Y., Qin, J. (2003). Information recovery in a study with surrogate endpoints. *Journal of the American Statistical Association, 98*, 1052–1062.

Csörgö, M., Csörgö, S., Horvath, L. (1986). *An asymptotic theory for empirical reliability and concentration process* (Vol. 33). Springer, New York.

Csörgö, M., Gastwirth, J., Zitikis, R. (1998). Asymptotic confidence bands for the Lorenz and bonferroni curves based on the empirical Lorenz curve. *Journal of statistical planning and inference, 74*, 65–91.

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

Francisco, C., Fuller, W. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics, 19*, 454–469.

Gail, M. H., Gastwirth, J. L. (1978). 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.

Gastwirth, J. L. (1971). A general definition of Lorenz curve. *Econometrica, 39*, 1037–1039.

Goldie, C. M. (1977). Convergence theorems for empirical Lorenz curves and their inverses. *Advances in Applied Probability, 9*, 765–791.

Hall, P., La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *International Statistical Review, 58*, 109–127.

Hall, P., Owen, A. B. (1993). Empirical likelihood confidence bands in density estimation. *Journal of Computational and Graphical Statistics, 2*, 273–289.

Hallas, J., Støvring, H. (2006). Templates for analysis of individual-level prescription data. *Basic and Clinical Pharmacology and Toxicology, 98*, 260–265.

Hart, P. E. (1971). Entropy and other measures of concentration. *Journal of the Royal Statistical Society: Series A, 134*, 73–89.

Hill, M. (1992). *The Panel Study of Income Dynamics: a user's guide*. Newbury Park, California/London, England: Sage Publications.

Lorenz, M. C. (1905). Methods of measuring the concentration of wealth. *Journal of the American Statistical Association, 9*, 209–219.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for single functional. *Biometrika, 75*, 237–249.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics, 18*, 90–120.

Owen, A. B. (2001). *Empirical Likelihood*. Noca Raton: Chapman & Hall/CRC.

Pollard, D. (1990). *Empirical processes: Theory and applications*. NSF-CBMS Regional Conference Series in Probability and Statistics (Vol. 2). Hayward, CA: Institute of Mathematical Statistics.

Sen, A. (1973). *On Economic inequality*. New York: Norton.

Wu, C. (2004). Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica, 14*, 1057–1067.

Zheng, B. (2002). Testing Lorenz curves with non-simple random samples. *Econometrica, 70*, 1235–1243.

Zhong, B., Rao., J. N. K. (2000). Empirical likelihood inference under stratified random sampling using auxiliary population information. *Biometrika, 87*, 929–938.

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