Georgia State University
ScholarWorks @ Georgia State University

Mathematics Theses

Department of Mathematics and Statistics

11-28-2007

Empirical Likelihood Confidence Intervals for Generalized Lorenz Curve

Nelly E. Belinga-Hill

EMPIRICAL LIKELIHOOD CONFIDENCE INTERVALS FOR GENERALIZED LORENZ CURVE

by

Nelly E. Belinga-Hill

Under the Direction of Dr. Gengsheng Qin

ABSTRACT

Lorenz curves are extensively used in economics to analyze income inequality metrics. In this thesis, we discuss confidence interval estimation methods for generalized Lorenz curve. We first obtain normal approximation (NA) and empirical likelihood (EL) based confidence intervals for generalized Lorenz curves. Then we perform simulation studies to compare coverage probabilities and lengths of the proposed EL-based confidence interval with the NA-based confidence interval for generalized Lorenz curve. Simulation results show that the EL-based confidence intervals have better coverage probabilities and shorter lengths than the NA-based intervals at 100p-th percentiles when p is greater than 0.50. Finally, two real examples on income are used to evaluate the applicability of these methods: the first example is the 2001 income data from the Panel Study of Income Dynamics (PSID) and the second example makes use of households' median income for the USA by counties for the years 1999 and 2006.

INDEX WORDS: Lorenz curve, Generalized Lorenz curve, Lorenz Ordinate, Confidence Intervals, Empirical Likelihood, Normal Approximation, Income data.

EMPIRICAL LIKELIHOOD CONFIDENCE INTERVALS FOR GENERALIZED LORENZ CURVE

by

Nelly E. Belinga-Hill

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

2007

EMPIRICAL LIKELIHOOD CONFIDENCE INTERVALS FOR GENERALIZED LORENZ CURVE

by

Nelly E. Belinga-Hill

Committee Chair: Dr. Gengsheng (Jeff) Qin

Committee: Dr. Yu-Sheng Hsu

Dr. Pulak Ghosh

Dr. Xu Zhang

Electronic Version Approval:

Office of Graduate Studies

College of Arts and Science

Georgia State University

December 2007

DEDICATION

To Nikao Hampton Belinga Hill, my son and my treasure.

ACKNOWLEDGEMENTS

My most appreciative gratitude goes to my thesis advisor Dr. Gengsheng Qin who has been able to patiently guide me through all the intricacies of empirical likelihood methods.

More thanks go to the other members of my committee Dr. Yu-Sheng Hsu, Dr. Pulak Ghosh, Dr. Xu Zhang, for taking the time to proofread and add valuable input to this thesis.

In addition, many thanks to my mentor Dr. Doris Derby as well as the entire Office of African American Student Services and Programs (OAASS&P) at GSU, whose visionary contributions have motivated me to strive through adverse times.

Many more thanks to my sisters and brothers: Steve Belinga, Gaelle Belinga, Cynthia Belinga, Maureen Belinga, Harry Belinga and to my friends Nnenna Nnanna, Rhoda Assumadu, Shona Campbell, and Melissa Chan-Leiba; without you I would not be here.

Extreme gratitude addressed to my father Gilles-Roger Belinga, in Cameroon, who is always setting an example of discipline and priority; then to my moms Florine Belinga and Peggy Terry for their nurturing spirits.

Last but not least, much appreciation to my husband Hamilton Hill for his unconditional support and encouragements.

TABLE OF CONTENT

LIST OF ABBREVIATIONS

CI:	**Confidence Interval**
NA:	**Normal Approximation**
EL:	**Empirical Likelihood**
MSE:	**Mean Square Error**
STDEV:	**Standard Deviation**
DEV :	**Deviation**
MCMC :	**Markov Chains Monte Carlo**
LC :	**Lorenz Curve**
GL :	**Generalized Lorenz**
PSID:	**The Panel Study of Income Dynamics**
HUD:	**Housing and Urban Development**

LIST OF TABLES

CHAPTER I: INTRODUCTION

Lorenz curve was named after Max Otto Lorenz (1905). Interests for Lorenz curves significantly rose however around the 1970s when Atkinson (1970) and Gastwirth (1971) presented quantitative measuring and inequality comparisons with the welfare economic implications of Lorenz curve. More contributions to Lorenz curves analysis were made by Sen (1973), Jakobsson (1976), Kakwani (1977), Goldie (1977), Marshall and Olkin (1979). Charles M. Beach and Russel Davidson (1983) took studies of Lorenz curves further as they derived the asymptotic joint variance-covariance structure for the Lorenz curve ordinates. They picked up from Shorack (1972) and Sendler (1979) derivation of the variance-covariance structure to offer a simpler tool to researchers. John A. Bishop, S. Chakraborti, and Paul D. Thistle (1989) continued in these tangent and proposed more results on analysis of generalized Lorenz ordinate that are useful for testing second-degree stochastic dominance. Recent development have been made by Mosler (1994, 2007), Arnold (1990) and Lambert (2001) whose findings have lead to numerous applications, particularly in reliability theory.

Let X be a positive random variable with cumulative distribution function $F(x)$. Gastwirth (1971) defined the Lorenz curve as the following function of $p \in (0,1]$:

$$\eta = \frac{1}{\mu}\int_0^{\xi_p} x dF(x),$$

where $\mu = \int_0^{\infty} x dF(x)$, and $\xi_p = F^{-1}(p)$ is the p-th quantile of *F*.

The generalized Lorenz curve is defined by

$$\theta = \int_0^{\xi_p} x dF(x).$$

In the analysis of income data, the distribution function $F(x)$ for the income is usually unknown. It is of interest to estimate Lorenz ordinates η and θ at a given p. Ryu and Slottje (1996) suggested an approach for the estimation of Lorenz curve by expanding the quantile function in terms of an exponential polynomial series and a sequence of Bernstein polynomial functions. Hikaru Hasegawa and Hideo Kozumi (2003) proposed an alternative method for estimating Lorenz curve by using Bayesian nonparametric approaches. They claim that their method is one the best of methods since it permits heteroscedasticity in individual incomes; however, it still needs to be evaluated with practical data.

Several econometrists have used Lorenz curves on actual datasets to evaluate welfare and poverty in given countries. For example, Pundarik Mukhopadhaya (2003) analyzed the changes in social welfare in Singapore by studying Labor Force Survey data from 1982 to 1999 published by the Manpower Research and Statistics of Singapore. He concluded that according to the generalized Lorenz dominance, 1999 ranks first on social welfare trends in Singapore. Another practical application of Lorenz curves can be found in the subject of famine and poverty evaluation, as Amartya Kumar Sen (1973) brought to light by describing the causes and effects of economic disparities with indexes such as Lorenz curves and Gini coefficients.

In this thesis, we focus on the construction of confidence intervals for the generalized Lorenz curve. We propose an empirical likelihood based confidence interval for the generalized Lorenz curve and compare it with the normal approximation based confidence interval. The thesis is organized as follows: In Chapter II, we review the normal approximation based interval for the generalized Lorenz curve. In Chapter III, we discuss the EL-based interval for the generalized Lorenz curve. In Chapter IV, we conduct simulation studies to evaluate the performances of these intervals. In Chapter V, we analyze two real data sets to compare the two methods. Finally, the conclusions are discussed in Chapter VI.

CHAPTER II: NORMAL APPROXIMATION BASED CONFIDENCE INTERVAL

In this chapter, normal approximation is used to construct confidence interval for the generalized Lorenz curve. We first need to find a suitable estimator for the generalized Lorenz curve.

Gastwirth (1971) defined the generalized Lorenz curve as

$$\theta = \int_0^{\xi_p} x dF(x) \qquad (2.1)$$

where $\xi_p = F^{-1}(p)$.

Let $X_1, X_2, \ldots, X_n$ be a random sample from $F(x)$, a consistent estimator for θ is

$$\hat{\theta} = \int_0^{\hat{\xi}_p} x d\hat{F}_n(x) = n^{-1} \sum_{i=1}^{n} X_i I(X_i \le \hat{\xi}_p) \qquad (2.2)$$

where $\hat{F}_n$ is the empirical distribution function of $X_1, X_2, \ldots, X_n$, $\hat{\xi}_p = \hat{F}_n^{-1}(p)$ is the p-th quantile of $\hat{F}_n$, and $I(X \le x)$ is the indicator function.

Zheng (2002) has shown that $\hat{\theta}$ is asymptotically normal with variance $\sigma_v^{\ 2}$, i.e.,

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i I(X_i \le \hat{\xi}_p) - \theta) \xrightarrow{d} N(0, \sigma_v^{\ 2}),$$

where $\sigma_v^2 = Var(X - \xi_P) I(X \le \xi_P)]$.

Therefore, a (1-α) normal approximation (NA) based confidence interval for θ can be constructed as follows:

$$\left(\hat{\theta} - z_{1-\frac{\alpha}{2}} \hat{\sigma}_v \big/ \sqrt{n}, \hat{\theta} + z_{1-\frac{\alpha}{2}} \hat{\sigma}_v \big/ \sqrt{n} \right)$$

where $z_{1-\frac{\alpha}{2}}$ is the (1-α/2)-th quantile of the standard normal distribution, and

$$\hat{\sigma}_v^2 = \frac{1}{n} \sum_{i=1}^{n} \left[(X_i - \hat{\xi}_p) I(X_i \le \hat{\xi}_p) - \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\xi}_p) I(X_i \le \hat{\xi}_p) \right]^2$$

is a consistent estimate for $\sigma_v^{\ 2}$.

CHAPTER III: EMPIRICAL LIKELIHOOD BASED CONFIDENCE INTERVAL

Empirical likelihood (EL), introduced by Owen (1988, 1990), is a prevailing nonparametric method. Some advantages of the EL method are as follows: it has better small sample performance than the normal approximation. It is also range preserving and transformation respecting. Wu and Rao (2006), Claeskens et al. (2003), DiCiccio and Romano (1989), Hall (1990) and Tsao (2001) have proposed ways to improve the accuracy of Empirical Likelihood based methods. In this chapter we will use empirical likelihood method to construct confidence interval for the generalized Lorenz curve.

From the definition of generalized Lorenz curve, we observe that

$$E[XI(X \le \xi_p)] - \theta = 0 .$$

So the generalized Lorenz ordinate θ is the mean of random variable X truncated at ξ_p. Based on observed data, we can define the empirical likelihood for θ as follows:

$$\tilde{L}_1(\theta) = \sup\left\{ \prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i V_i = 0 \right\},$$

where $\boldsymbol{p} = (p_1, \ldots p_n)$ is a probability vector and $V_i = X_i I(X_i \le \xi_p) - \theta$.

Since the population quantile is unknown, replacing V_i by $\hat{V}_i = X_i I(X_i \le \hat{\xi}_p) - \theta$, we obtain an estimated empirical likelihood for the generalized Lorenz ordinate θ:

$$L_1(\theta) = \sup\left\{ \prod_{i=1}^{n} p_i : \sum p_i = 1, \sum_{i=1}^{n} p_i \hat{V}_i = 0 \right\} \quad . \qquad (3.1)$$

By the Lagrange multiplier, we get

$$p_i = \frac{1}{n}\left\{1 + t\hat{V}_i\right\}^{-1}, \; i=1,\ldots..,n$$

where t is the solution to

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{V}_i}{1 + t\hat{V}_i} = 0 .$$

We note that $\prod_{i=1}^{n} p_i$ subject to $\sum_{i=1}^{n} p_i = 1$ attains its maximum n^{-n} at $p_i=n^{-1}$. Therefore the empirical ratio for θ will be

$$R_1(\theta) = \prod_{i=1}^{n} np_i = \prod_{i=1}^{n} \left\{1 + t\hat{V}_i\right\}^{-1} . \tag{3.2}$$

The corresponding empirical log-likelihood ratio is

$$l_1(\theta) = -2\log R_1(\theta) = 2\sum_{i=1}^{n} \log\left\{1 + t\hat{V}_i\right\} . \tag{3.3}$$

Qin (2006) established the following theorem:

Theorem: If $E(X^2) < \infty,$ and θ_0 is the true value of θ, then the limiting distribution of $l_1(\theta_0)$ is a scaled chi-square distribution with degree of freedom 1, that is,

$$r_1 l_1(\theta_0) \xrightarrow{L} \chi_1^2 ,$$

where the scale constant $r_1 = \sigma_p^2 / \sigma_v^2$ with

$$\sigma_p^2 = Var[XI(X \le \xi_P)],$$

$$\sigma_v^2 = Var[(X - \xi_P) I(X \le \xi_P)].$$

The scale constant r_1 is still unknown, but it can be consistently estimated by

$$\hat{r}_1 = \hat{\sigma}_p^2 / \hat{\sigma}_v^2 ,$$

where

$$\hat{\sigma}_p^2 = \frac{1}{n}\sum_{i=1}^{n}\left[X_i I(X_i \le \hat{\xi}_p) - \frac{1}{n}\sum_{i=1}^{n} X_i I(X_i \le \hat{\xi}_p)\right]^2 ,$$

$$\hat{\sigma}_v^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(X_i - \hat{\xi}_p) I(X_i \le \hat{\xi}_p) - \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\xi}_p) I(X_i \le \hat{\xi}_p)\right]^2 .$$

Therefore, a (1-α)-th empirical likelihood based confidence intervals for θ can be constructed as follows:

$$\{\theta : \hat{r}_1 l_1(\theta) \le \chi_{1,1-\alpha}^2\} \tag{3.4}$$

where $\chi_{1,1-\alpha}^2$ is the (1-α)-th quantile of chi-square distribution with degree of freedom 1.

CHAPTER IV: A SIMULATION STUDY

In this chapter, we perform a simulation study to evaluate the estimation methods for the generalized Lorenz curve. In order to assess the accuracy of the point estimator, the BIAS and Mean Square Error (MSE) of the estimates are calculated. We also compare the NA-based confidence interval with the EL-based confidence interval for the generalized Lorenz curve in terms of coverage probabilities and interval lengths.

In the simulation study, the population income distribution *F(x)* is assumed to be a *Weibull* distribution with parameters *(a, b)* where *a* is the shape parameter and *b* is the scale parameter. In the study, we choose *(a, b)* = (1, 1) and (1, 2) respectively. m=10,000 random samples of size *n=50, 100, 150, 200, 500* are generated from *Weibull (a,b)* . Using the simulated random samples, the BIAS and MSE of the estimates for the generalized Lorenz ordinates are calculated at the 100*p*-th percentile of the income distribution. In the study, *p* is taken to be 0.95, 0.90, 0.75, 0.5, 0.25, 0.1, and 0.05 respectively. We also calculate the coverage probabilities and interval lengths of 90% and 95% confidence intervals for the generalized Lorenz curve by using the NA approach and the EL approach presented in Chapter II and Chapter III.

The S-Plus code for the simulation study is presented in Appendix C. The results of the simulation study are reported in Tables 5-14 in Appendix A. From these Tables, we make the following observations:

1. Both the BIAS and MSE of the estimates for the generalized ordinates are very close to 0. As the sample sizes increase, these BIAS and MSE get closer to 0. Hence, the proposed estimator is a good point estimator for the generalized Lorenz curve.

2. The coverage probabilities of the EL-based intervals are closer to the nominal confidence levels than those of the NA-based intervals at the 100p-th percentiles in most cases considered here, particularly when $p \geq 0.50$. However, at the 100p-th percentiles with $p < 0.5$ the NA-based intervals may have better coverage probabilities than the EL-based intervals.

3. When Analyzing the lengths of all confidence intervals obtained for both methods, we observe that the lengths of the 95% and 90% EL-based confidence intervals are shorter than the NA-based confidence intervals when $p \geq 0.50$. When $p < 0.50$ we experience cases when NA interval lengths are smaller, and other cases when EL interval lengths are smaller.

In conclusion, we recommend that the EL-based confidence interval for the generalized Lorenz curve when $p \geq 0.50$. The NA-based confidence intervals can still be used when $p < 0.50$.

CHAPTER V: REAL DATA EXAMPLES

EXAMPLE 1: PSID Family 'Income Plus' Files

The Panel Study of Income Dynamics (PSID) is a longitudinal survey of men, women, and children, and families in the U.S. Since 1968, the PSID has conducted studies at the University of Michigan's Survey Research Center. It has annually collected information on U.S. families and to date, approximately 37,500 individuals have been interviewed. 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 (2002) explains that another originality of the PSID data comes from the fact that they initially collected data in order to study dynamics of poverty; as a result too many low income and Black households were included in the samples. However, by 2001, they have included more income variety and 2,043 Latino (Mexican, Cuban, and Puerto Rican) households to help correct for omissions in representing post-1968 immigrants.

In this thesis, we used the data from the PSID Family 'Income Plus' Files 1994-2001. This data can be found in a SAS or SPSS data format in http://simba.isr.umich.edu/Zips/zipSupp.aspx#income94- . Hasegawa and Kozumi (2003) used this PSID data for year 1997 to apply Bayesian nonparametric methods to the estimation of the Lorenz Curve and inequality measures. We prefer to use the 2001 income data instead since it is the most recent data available. We focus on the variable labeled: FAMINC01 which represents the total family income for the year 2000. The sample consists of 7,406 individuals.

We report a portion of the original data below:

Table 1: "Total Family Income in 2000"

ID01	FIPS _STN	PSID _STN	FAMINC01	TXHW01	TRHW01	TXOFM01	TROFM01	SSEC01
1	12	9	15087	15	0	0	0	15072
2	12	9	40700	13800	9500	0	0	17400
3	37	32	3894	3000	894	0	0	0
4	37	32	34000	34000	0	0	0	0
5	37	32	10800	0	10800	0	0	0
6	37	32	4880	0	1848	0	0	3032
7	6	4	9864	0	840	0	0	9024
8	28	23	1593	1593	0	0	0	0
9	37	32	78521	78521	0	0	0	0
10	12	9	18400	1360	8244	0	0	8796
11	12	9	20001	6001	2000	0	0	12000
12	19	14	61287	60600	0	0	0	687
13	32	27	55100	55100	0	0	0	0
14	47	41	22400	20000	0	0	0	2400
15	41	36	29976	10200	9536	8000	2240	0
16	26	21	43095	24895	7400	0	0	10800
17	10	7	5396	500	0	0	0	4896
18	5	3	13011	1593	10680	0	0	738
19	49	43	9331	8593	0	0	0	738
20	6	4	18792	5400	0	0	0	13392
21	12	9	21719	2594	13909	5000	0	216
22	4	2	30000	30000	0	0	0	0
23	45	39	13288	700	0	0	0	12588
24	37	32	7530	6642	0	0	888	0
25	37	32	738	0	0	0	0	738
26	39	34	23840	1000	7000	0	0	15840
27	45	39	28400	9200	6000	0	0	13200
28	17	12	50049	50049	0	0	0	0
29	28	23	25500	25500	0	0	0	0
30	37	32	24000	0	15000	0	0	9000
31	13	10	6820	0	100	0	0	6720
32	51	45	115600	106000	9600	0	0	0
33	19	14	15564	6000	0	0	0	9564
34	6	4	247800	247800	0	0	0	0
35	17	12	68876	31400	13764	0	0	23712
36	41	36	25000	13000	12000	0	0	0
37	17	12	85300	85300	0	0	0	0
38	17	12	75200	75200	0	0	0	0
39	21	16	14172	3000	0	0	0	11172
40	37	32	7290	2496	1578	0	0	3216

LABELS:

ID01="2001 INTERVIEW NUMBER"

FIPS_STN="FIPS STATE NUMERIC CODE"

PSID_STN="PSID STATE CODE"

FAMINC01="TOTAL FAMILY INCOME 2000"

TXHW01="TAXABLE INCOME HEAD AND WIFE 2000"

TRHW01="TRANSFER INCOME OF HEAD AND WIFE 2000"

TXOFM01="TAXABLE INCOME OTHER FAMILY UNIT MEMBERS"

TROFM01="TRANSFER INCOME OTHER FAMILY UNIT MEMBER"

SSEC01="SOCIAL SECURITY INCOME 2000"

Summary statistics for this data can be found in the table below. The first set of column report summary statistics when negative incomes are transformed. These negative variables arose from a business loss or from living on liquidated assets such as farms or businesses. We transformed the data by adding the absolute value of the minimum income value to the whole data. The second set of columns report summary statistics of the same data when negative incomes are erased.

Table 2: Summary Statistics for "Total Family Income in 2000"

NEGATIVE VALUES TRANSFORMED		NEGATIVE VALUES ERASED	
Mean	$119,075.61	Mean	$59,334.915
Median	$102,203.50	Median	$42,460
Standard Deviation	77,831.62	Standard Deviation	77,817.218
Minimum	0	Minimum	0
Maximum	$2,172,248.00	Maximum	$2,112,300
Count	7406	Count	7387
Percentiles		Percentiles	
5%	$66,448.00	5%	$6,649.2
10%	$70,992.00	10%	$11,250.00
20%	$78,664.60	20%	$18,888.80
25%	$82,541.00	25%	$22,720.50
30%	$85,948.00	30%	$26,000.00
40%	$93,747.20	40%	$33,954.40
50%	$102,227.00	50%	$42,460.00
60%	$112,050.80	60%	$52,160.00
70%	$124,945.00	70%	$65,000.00
75%	$132,678.00	75%	$72,877.00
80%	$141,628.80	80%	$81,800.00
90%	$173,948.00	90%	$114,057.20
95%	$212,480.00	95%	$152,805.10
99%	$391,559.84	99%	$332,227.94

We upload the original data in S-PLUS to compare the 95% and 90% NA-based confidence intervals with the EL-based confidence intervals for the generalized Lorenz curve for incomes in 2000. Results are presented in Appendix B.1. We observe that the lengths of EL-based intervals are shorter than those of NA intervals for all the percentiles used even when p=0.25 or smaller.

EXAMPLE 2: Section 8 Housing Median Income Data

In this thesis we also used a data from the Housing and Urban Development (HUD) programs which are more commonly known as section 8. This is a Housing Choice Voucher Program dedicated to sponsoring subsidized housing for low-income families and individuals. The data used represents households' median income for the USA by counties for the years 1999 and 2006.

Historically, Federal housing assistance programs began during the Great Depression. In the 1960s and 1970s, the federal government created subsidy programs to help low income families pay their rent. In 1961, housing authorities selected eligible families from their waiting list, placed them in housing and determined the rent that tenants would have to pay. The housing authority would then sign a lease with the private landlord and pay the difference between the tenant's rent and the market rate for the same size unit. Housing authorities agreed to perform regular building maintenance.

Section 8 is attributed to families based on a set of rules. Eligible families pay 30% of their income while living in the apartment. The local housing authority pays the owner the remaining rent, subject to a cap referred to as "Fair Market Rent" (FMR) which is determined by HUD. Median Family Income Estimates (MFI) serve as estimates as the basis for a family to qualify to section 8 housing. HUD updates the MFI by using American Community Survey (ACS) income datasets.

The original data includes 4,764 variables. Table 2 below presents a portion of the data, the data in its entirety can be retrieved from :

http://www.huduser.org/DATASETS/il/il06/index.html.

Table 3: Households Median Income by USA counties for 1999 and 2006 (partial data)

State	State	County_Town_Name	County	Metro_Area_Name	CBSASub	County_Name	median1999	median2006
AL	1	Autauga County	1	Montgomery, AL MSA	METRO33860M33860	Autauga County	45182	55900
AL	1	Baldwin County	3	Baldwin County, AL	NCNTY01003N01003	Baldwin County	47030	58100
AL	1	Barbour County	5	Barbour County, AL	NCNTY01005N01005	Barbour County	31877	38700
AL	1	Bibb County	7	Birmingham-Hoover,	METRO13820M13820	Bibb County	46422	57400
AL	1	Blount County	9	Birmingham-Hoover,	METRO13820M13820	Blount County	46422	57400
AL	1	Bullock County	11	Bullock County, AL	NCNTY01011N01011	Bullock County	24003	29700
AL	1	Butler County	13	Butler County, AL	NCNTY01013N01013	Butler County	30911	38300
AL	1	Calhoun County	15	Anniston-Oxford, AL MSA	METRO11500M11500	Calhoun County	39907	49500
AL	1	Chambers County	17	Chambers County, AL	NCNTY01017N01017	Chambers County	36598	45300
AL	1	Cherokee County	19	Cherokee County, AL	NCNTY01019N01019	Cherokee County	36920	45400
AL	1	Chilton County	21	Chilton County,	METRO13820N01021	Chilton County	39503	49000
AL	1	Choctaw County	23	Choctaw County, AL	NCNTY01023N01023	Choctaw County	31870	39100
AL	1	Clarke County	25	Clarke County, AL	NCNTY01025N01025	Clarke County	34548	42600
AL	1	Clay County	27	Clay County, AL	NCNTY01027N01027	Clay County	34026	42200
AL	1	Cleburne County	29	Cleburne County, AL	NCNTY01029N01029	Cleburne County	35579	44300
AL	1	Coffee County	31	Coffee County, AL	NCNTY01031N01031	Coffee County	39664	48900
AL	1	Colbert County	33	Florence-Muscle Shoals,	METRO22520M22520	Colbert County	40652	50000
AL	1	Conecuh County	35	Conecuh County, AL	NCNTY01035N01035	Conecuh County	31424	38300
AL	1	Coosa County	37	Coosa County, AL	NCNTY01037N01037	Coosa County	36088	44400
AL	1	Covington County	39	Covington County, AL	NCNTY01039N01039	Covington County	33197	40800
AL	1	Crenshaw County	41	Crenshaw County, AL	NCNTY01041N01041	Crenshaw County	31724	38500
AL	1	Cullman County	43	Cullman County, AL	NCNTY01043N01043	Cullman County	39342	48400
AL	1	Dale County	45	Dale County, AL	NCNTY01045N01045	Dale County	37806	46800
AL	1	Dallas County	47	Dallas County, AL	NCNTY01047N01047	Dallas County	29906	37400
AL	1	DeKalb County	49	DeKalb County, AL	NCNTY01049N01049	DeKalb County	35802	44300
AL	1	Elmore County	51	Montgomery, AL MSA	METRO33860M33860	Elmore County	45182	55900
AL	1	Escambia County	53	Escambia County, AL	NCNTY01053N01053	Escambia County	36086	44300
AL	1	Etowah County	55	Gadsden, AL MSA	METRO23460M23460	Etowah County	38698	47400
AL	1	Fayette County	57	Fayette County, AL	NCNTY01057N01057	Fayette County	35289	43700

Summary statistics for the data are presented in Table 4.

Table 4: Summary Statistics for Households Median Income by USA counties for 1999 and 2006

1999		***2006***	
Mean	44,537.771	Mean	53,942.107
Median	43,180	Median	51,900
Standard Deviation	11,288.31365	Standard Deviation	14,115.15486
Minimum	12,293	Minimum	14,600
Maximum	94,229	Maximum	116,300
Count	4,764	Count	4,764
Percentiles		Percentiles	
5%	$29,483.90	5%	$34,800.00
10%	$32,287.80	10%	$38,800.00
20%	$36,041.00	20%	$43,300.00
25%	$36,826.00	25%	$36,826.00
30%	$37,878.00	30%	$45,800.00
40%	$40,523.20	40%	$48,800.00
50%	$43,180.00	50%	$51,900.00
60%	$45,664.20	60%	$55,400.00
70%	$49,414.00	70%	$60,300.00
75%	$51,060.00	75%	$62,600.00
80%	$53,090.00	80%	$64,240.00
90%	$59,651.00	90%	$71,900.00
95%	$66,460.00	95%	$82,000.00
99%	$74,611.00	99%	$90,300.00

We upload the original data in S-PLUS to compare the 95 % and 90% NA-based confidence intervals with the EL-based confidence intervals for the generalized Lorenz curve for median incomes in 1999 and in 2006. Results are presented in Appendix B.2. We observe that the lengths of EL-based intervals for the generalized Lorenz are shorter than those of NA-based intervals when p=0.5 or higher. When p=0.25 the lengths of NA-based confidence intervals are much smaller than those of the EL-based confidence intervals.

Based on our simulation study, we would like to use the EL-based confidence intervals for the generalized Lorenz curve when $p \geq 0.50$ and the NA-based confidence intervals for the generalized Lorenz curve when $p < 0.50$ in these two applications.

CHAPTER VI: DISCUSSION AND CONCLUSIONS

In this thesis, we have compared the normal approximation and the empirical likelihood based confidence intervals for the generalized Lorenz curve. From the simulation study we have observed that the coverage probability of EL-based intervals are much closer to the nominal confidence levels at 100p-th percentiles when $p \geq 0.50$. However, when $p=0.10$ and below, the coverage accuracy of the NA-based intervals may outperform the EL-based intervals.

Wu and Rao (2006) explained that NA-based intervals are simple but usually not the best in terms of coverage probabilities. Another disadvantage of NA-based interval lies in the fact that it may have poor performance when the underlying distribution is skewed. In economic studies, the income distributions are often skewed. We need to assess the performance of NA-based interval before its use. From our simulation results and analysis for the real examples, we recommend the use of EL-based confidence intervals for the generalized Lorenz curve when $p \geq 0.50$. The NA-based confidence intervals can still be used when $p < 0.50$.

Further studies will be concentrated on construction of confidence intervals for the Lorenz curve using empirical likelihood method.

REFERENCES

2006 Catalog of Federal Domestic Assistance (CFDA) of the United States http://en.wikipedia.org/wiki/Federal_assistance_in_the_United_States .

Allen, Arnold O. (1990). *Probability, Statistics, and Queuing Theory with Computer Science Applications, 2nd ed*. Academic Press.

Atkinson, Anthony B. (1970). On the Measurement of Inequality. *Journal of Economic Theory*, 2, 244-63.

Charles M. Beach and Russel Davidson (1983). Distribution-Free Statistical Inference with Lorenz Curves and Income Shares. *The Review of Economic Studies*, **50**, 723-735.

John A. Bishop, S. Chakraborti, and Paul D. Thistle (1989). Asymptotically Distribution-Free Statistical Inference for Generalized Lorenz Curves. *The Review of Econometrics and Statistics*, **71**, 725-727.

Claeskens, G., Jing, B.-Y., Peng, L. & Zhou, W. (2003). Empirical Likelihood Confidence Regions for Comparison Distributions and ROC Curves, *The Canadian Journal of Statistics*, **31**, 173-190.

DiCiccio and Romano (1989). Empirical Likelihood in Barlett-Correctable. *The Annals of Statistics*, **19**, No. 2, 1053-1061.

El Barmi (1996). Empirical Likelihood Test For or Against a Set of Inequality Constraints. *Journal of statistical planning and inference*, **55**, n^{o}2, pp. 191-204.

Duncan, Greg J., Bjorn Gustafsson, Richard Hauser, Guenther Schmaus, Stephen Jenkins, Hans Messinger, Ruud Muffels, Brian Nolan, Jean-Claude Ray, and Wolfgang Voges. 1995. “Poverty and Social-Assistance Dynamics in the United States, Canada and Western Europe” in *Poverty, Inequality and the Future of Social Policy: Western States in the New World Order*, Edited by Katherine McFate, Roger Lawson and William J. Wilson. New York: Russell Sage.

Gastwirth, J. L. (1971). A General Definition of Lorenz Curve. *Econometrica*, **39**, 1037-1039

HUD FY2006 Income Limits http://www.huduser.org/DATASETS/il/il06/index.html .

Kakwani, N.C., Podder, N., (1973). On The Estimation Of Lorenz Curves From Grouped Observation. *International Economic Review,* **14**, 278-292

Kim, Y.-S., Loup, T., Lupton, J., Stafford, F.P., 2000. Notes on the “Income Plus” files:1994-1997 family income and components files. Documentation, the Panel Study of Income Dynamics (http://www.isr.umich.edu/src/psid/income94-97/y-pls-notes.htm)

C. Kleiber (1999). On The Lorenz order within Parametric Families of Distributions. *The Indian Journal of Statistics,* **61**, 514-517.

Hall, P. and la Scala, b. (1990). Methodology and Algorithms of Empirical Likelihood. *International Statist. Review.* **58**, 2, 109-107.
Hikaru Hasegawa, Hideo Kozumi (2003). Estimation of Lorenz curves: A Bayesian nonparametric approach. *Journal of Econometrics,* **115**, 277-291.

Hill, Martha. (1992). *The Panel Study of Income Dynamics*. Beverly Hills: Sage Publications

Lambert, P.J., (1993). The Distribution and Redistribution of Income: A Mathematical analysis, 2nd Edition. Manchester University Press, Manchester.

Lorenz, M. C. (1905). Method of Measuring the Concentration of Wealth. *J. Ame. Statist.*, **9**, 209-219.

W. Marshall and I. Olkin, (1979).Inequalities: Theory of Majorization and its Applications. *Academic Press*, New York.

K. Mosler (1994). Majorization in Economic Disparity Measures. Linear Algebra and Its Applications, **91**,-114.

K. Mosler, G. Koshevov (2007). Multivariate Lorenz dominance based on zonoids). ASTA: Advances in Statistical Analysis, **91**, 57-76

PUNDARIK MUKHOPADHAYA (2003) The Ordinal And Cardinal Judgment Of Social Welfare Changes In Singapore, 1982?99,*The Developing Economies*, **41** (1), 65–87.

Nikitin, Ya. Yu., Tchirina, A. V. (1996) Bahadur Efficiency and Local Optimality of a Test for the Exponential Distribution Based on Gini Statistic. *J. Ital. Statist. Soc.,* **5,** 163-175**.**

Owen, A. (1990). Empirical likelihood radio confidence regions. *Annals of Statistics,* **18**, 90-120.

Owen, A. (2001). Empirical likelihood. *Chapman & Hall/CRC*, New York.

Qin, J. Lawless (1994). Empirical Likelihood and General Estimating Equations, *Ann. Statist.,* **22,** 300-325.

Hang K. Ryu, Daniel J. Slottje (1996). Two Flexible Functional Form Approaches For Approximating the Lorenz Curve. *Journal of Econometrics***,72**, 251-274.

Sen, Amartya, (1973) *On Economic Inequality*, New York, Norton.

Shorack Galen R. (1972) Functions of Order Statistics. *Ann. Math. Statist.,***43,** 412-127.

Slottje, D. J. (1989). *The Structure of Earnings and the Measurement of Income Inequality in the US.* North-Holland, Amsterdam.

Tsao (2001). A Small Sample Calibration for the Empirical Likelihood. *Statistics and Probability Letters*, **54**, N. 1, pp. 41-45.

Wu C and Rao JNK (2006). Pseudo-Empirical Likelihood Ratio Confidence Intervals For Complex Surveys. *The Canadian Journal of Statistics*, **34**, 359-375.

Zheng (1999). Statistical Inferences for Testing Marginal Rank and (Generalized) Lorenz Dominances. *Southern Economical Journa*l, **65**(3), 557-570.

Zheng (2002). Testing Lorenz Curves with Non-Simple Random Samples. *Econometrica*,**70**(3), 1235-1243.

APPENDIX

APPENDIX A: SIMULATION TABLES

APPENDIX A.1: SIMULATION TABLES FOR WEIBULL(1,1)

Table 5 : Weibull Distribution, BIAS and MSE of the Estimate for the generalized Lorenz ordinate								
Weibull(a =1, b=1)								
Sample size	Estimate errors	p=0.95	p=0.90	p=0.75	p=0.50	p=0.25	p=0.10	p=0.05
n=50	BIAS	-0.02107	0.00931	-0.00655	0.00456	0.00562	0.00100	0.00116
	MSE	0.01385	0.01114	0.00476	0.00118	0.00018	9.7758E-06	3.73821E-06
n=100	BIAS	0.00285	0.00379	0.00391	0.00246	0.00134	0.00053	0.00025
	MSE	0.00702	0.00538	0.00245	0.00059	0.00006	4.1076E-06	6.19499E-07
n=150	BIAS	-0.00750	0.00326	-0.00224	0.00138	0.00175	0.00032	0.00035
	MSE	0.00455	0.00359	0.00161	0.00038	0.00005	2.6586E-06	5.38101E-07
n=200	BIAS	0.00109	0.00190	0.00142	0.00119	0.00066	0.00023	0.00013
	MSE	0.00344	0.00279	0.00121	0.00029	0.00003	2.0275E-06	2.65744E-07
n=500	BIAS	0.00135	0.00073	0.00067	0.00063	0.00020	0.00009	0.00005
	MSE	0.00051621	0.00036345	0.00049	0.00011	0.00001	6.9534E-07	1.46568E-09

Table 6: Weibull Distribution, Coverage Probability of the 95% CI
for the generalized Lorenz ordinate

Weibull(a =1, b=1)

Sample Size	Method	p=0.95	p=0.90	p=0.75	p=0.50	p=0.25	p=0.10	p=0.05
		Coverage Probability	Coverage Probability	Coverage Probability	Coverage Probability	Coverage Probability	Coverage Probability	Coverage Probability
n=50	EL	0.9848	0.9367	0.9600	0.9346	0.9592	0.8858	0.7338
	NA	0.9330	0.9235	0.9386	0.8969	0.8918	0.6329	0.9148
n=100	EL	0.9458	0.9449	0.9458	0.9378	0.9343	0.9201	0.9004
	NA	0.9661	0.8636	0.9116	0.8744	0.8259	0.8649	0.9553
n=150	EL	0.9455	0.9471	0.9480	0.9442	0.9320	0.9242	0.8667
	NA	0.9002	0.9669	0.9723	0.9258	0.9337	0.8470	0.8212
n=200	EL	0.9497	0.9445	0.9449	0.9461	0.9431	0.9327	0.9169
	NA	0.9487	0.8937	0.9447	0.9400	0.8531	0.7392	0.9537
n=500	EL	0.9502	0.9491	0.9489	0.9484	0.9503	0.9414	0.9342
	NA	0.9488	0.9447	0.9192	0.9401	0.9400	0.9316	0.9813

Table 7: Weibull Distribution, Length of the 95% CI
for the generalized Lorenz ordinate

Weibull(a =1, b=1)

Sample Size	Method	p=0.95	p=0.90	p=0.75	p=0.50	p=0.25	p=0.10	p=0.05
		Length	Length	Length	Length	Length	Length	Length
n=50	EL	0.41898	0.37429	0.26623	0.13456	0.06618	0.03864	0.03113
	NA	0.45516	0.53790	0.33525	0.14470	0.05316	0.01357	0.00317
n=100	EL	0.30130	0.28384	0.19480	0.09381	0.04259	0.02009	0.01820
	NA	0.31261	0.28545	0.24530	0.10012	0.04508	0.01179	0.00258
n=150	EL	0.25381	0.23582	0.16688	0.07662	0.03046	0.01432	0.00859
	NA	0.25915	0.26353	0.17512	0.07634	0.01770	0.00813	0.00163
n=200	EL	0.20564	0.20596	0.12776	0.06663	0.02531	0.01155	0.00779
	NA	0.23134	0.21524	0.14791	0.07213	0.02340	0.00706	0.00324
n=500	EL	0.14648	0.12750	0.08590	0.03990	0.01311	0.00365	0.00092
	NA	0.14882	0.13451	0.09594	0.04669	0.01399	0.00638	0.00359

Table 8: Weibull Distribution, Coverage Probability of the 90% CI
for the generalized Lorenz ordinate

Weibull(a =1, b=1)

Sample Size	Method	p=0.95	p=0.90	p=0.75	p=0.50	p=0.25	p=0.10	p=0.05
		Coverage Probability	Coverage Probability	Coverage Probability	Coverage Probability	Coverage Probability	Coverage Probability	Coverage Probability
n=50	EL	0.9578	0.8857	0.9148	0.8838	0.8340	0.8374	0.6736
	NA	0.8823	0.8595	0.8889	0.8335	0.9295	0.5480	0.8713
n=100	EL	0.8939	0.8951	0.8971	0.8869	0.8851	0.8704	0.8554
	NA	0.9265	0.7940	0.8454	0.8069	0.7472	0.7966	0.9269
n=150	EL	0.8977	0.8971	0.9004	0.8909	0.8844	0.8740	0.8085
	NA	0.8363	0.8819	0.8878	0.8641	0.8771	0.7797	0.7617
n=200	EL	0.9084	0.8949	0.8968	0.8980	0.8964	0.8827	0.8705
	NA	0.8937	0.8262	0.8943	0.8881	0.7781	0.6520	0.9239
n=500	EL	0.9034	0.9096	0.8981	0.8996	0.9004	0.8919	0.8839
	NA	0.9013	0.8964	0.8592	0.8867	0.8868	0.8757	0.9594

Table 9: Weibull Distribution, Length of the 90% CI
for the generalized Lorenz ordinate

Weibull(a =1, b=1)

Sample Size	Method	p=0.95	p=0.90	p=0.75	p=0.50	p=0.25	p=0.10	p=0.05
		Length	Length	Length	Length	Length	Length	Length
n=50	EL	0.34268	0.30258	0.21432	0.10597	0.04863	0.03106	0.01839
	NA	0.38198	0.45142	0.28135	0.12143	0.04461	0.01139	0.00266
n=100	EL	0.25146	0.22560	0.15821	0.07507	0.03104	0.01541	0.01562
	NA	0.25286	0.23956	0.20586	0.08402	0.03783	0.00989	0.00217
n=150	EL	0.21160	0.18764	0.13402	0.06161	0.02247	0.01122	0.00720
	NA	0.21748	0.22116	0.14697	0.06407	0.01486	0.00682	0.00137
n=200	EL	0.17257	0.16278	0.10722	0.05414	0.01905	0.00917	0.00664
	NA	0.18373	0.18064	0.11684	0.06053	0.01963	0.00593	0.00272
n=500	EL	0.12293	0.10701	0.07209	0.03348	0.01100	0.00306	0.00077
	NA	0.12314	0.10904	0.07544	0.03789	0.01129	0.00492	0.00282

APPENDIX A.2: SIMULATION TABLES FOR WEIBULL (1, 2)

Table 10 : Weibull Distribution, BIAS and MSE of the Estimate for the generalized Lorenz ordinate

Weibull(a =1, b=2)

Sample size	Estimate errors	p=0.95	p=0.90	p=0.75	p=0.50	p=0.25	p=0.10	p=0.05
n=50	BIAS	-0.04099	0.0180037	-0.012707	0.0095577	0.0112159	0.00196691	0.002414067
	MSE	0.0564755	0.0429248	0.0200568	0.004824	0.000721468	3.8837E-05	1.52E-05
n=100	BIAS	0.0100205	0.0102114	0.0079249	0.0048624	0.002478528	0.00093146	0.000505389
	MSE	0.029035	0.0218353	0.0101645	0.0022861	0.000257103	1.70E-05	2.40E-06
n=150	BIAS	-0.015419	0.004997	-0.003823	0.0029761	0.00363945	0.00067638	0.000702251
	MSE	0.0184317	0.0142784	0.0064864	0.0016162	0.000187111	1.10E-05	2.14E-06
n=200	BIAS	0.0039354	0.0039521	0.0038949	0.0031396	0.001292246	0.00049365	0.000251136
	MSE	0.0138585	0.0109202	0.0050013	0.0011698	0.000120377	7.64E-06	1.01E-06
n=500	BIAS	0.0031587	0.0015305	0.001564	0.0012272	0.000359996	0.00018649	8.55E-05
	MSE	0.0055718	0.0042333	0.0019646	0.0004621	4.87E-05	2.96E-06	3.48E-07

Table 11: Weibull Distribution, Coverage Probability of the 95% CI for the generalized Lorenz ordinate

Weibull(a =1, b=2)

Sample Size	Method	p=0.95	p=0.90	p=0.75	p=0.50	p=0.25	p=0.10	p=0.05
		Coverage Probability	Coverage Probability	Coverage Probability	Coverage Probability	Coverage Probability	Coverage Probability	Coverage Probability
n=50	EL	0.9377	0.9355	0.9441	0.9348	0.8902	0.8904	0.7291
	NA	0.9044	0.7775	0.8886	0.9254	0.9185	0.9138	0.7492
n=100	EL	0.9529	0.9475	0.9518	0.9428	0.9380	0.9158	0.8885
	NA	0.9351	0.9604	0.9426	0.9290	0.9288	0.9663	0.9191
n=150	EL	0.9448	0.9422	0.9511	0.9485	0.9331	0.9445	0.8537
	NA	0.9166	0.9404	0.9370	0.8282	0.9309	0.9266	0.7443
n=200	EL	0.9443	0.9514	0.9418	0.9464	0.9386	0.9296	0.9167
	NA	0.9701	0.9210	0.9664	0.9204	0.8021	0.9370	0.9349
n=500	EL	0.9528	0.9515	0.9492	0.9536	0.9462	0.9502	0.9374
	NA	0.9501	0.9450	0.9320	0.9498	0.9427	0.9779	0.9750

Table 12: Weibull Distribution Distribution, Length of the 95% CI for the generalized Lorenz ordinate

Weibull(a =1, b=2)

Sample Size	Method	p=0.95	p=0.90	p=0.75	p=0.50	p=0.25	p=0.10	p=0.05
		Length	Length	Length	Length	Length	Length	Length
n=50	EL	0.88620	0.75578	0.50012	0.25871	0.09665	0.05067	0.09995
	NA	1.05675	0.95893	0.52625	0.39449	0.14536	0.00948	0.00491
n=100	EL	0.62606	0.45043	0.36968	0.19124	0.06506	0.02932	0.02010
	NA	0.83085	0.54311	0.50003	0.20966	0.05922	0.01218	0.00445
n=150	EL	0.51162	0.44703	0.30985	0.14603	0.05160	0.02159	0.01143
	NA	0.58433	0.47949	0.32568	0.16353	0.04186	0.00797	0.00656
n=200	EL	0.44382	0.38786	0.27044	0.14008	0.04422	0.01826	0.01030
	NA	0.52673	0.46364	0.27958	0.14406	0.04230	0.01110	0.00521
n=500	EL	0.27690	0.24854	0.17142	0.08196	0.02857	0.00881	0.00499
	NA	0.28490	0.26400	0.17116	0.09294	0.03138	0.00766	0.00446

Table 13: Weibull Distribution, Coverage Probability of the 90% CI
for the generalized Lorenz ordinate

Weibull(a =1, b=2)

Sample Size	Method	p=0.95	p=0.90	p=0.75	p=0.50	p=0.25	p=0.10	p=0.05
		Coverage Probability	Coverage Probability	Coverage Probability	Coverage Probability	Coverage Probability	Coverage Probability	Coverage Probability
n=50	EL	0.8783	0.8895	0.8938	0.8822	0.8301	0.8488	0.6683
	NA	0.8638	0.6893	0.8170	0.8687	0.8753	0.8772	0.6906
n=100	EL	0.9043	0.9041	0.8927	0.8907	0.8861	0.8679	0.8499
	NA	0.8888	0.9127	0.9179	0.8732	0.8787	0.9393	0.8855
n=150	EL	0.8923	0.8973	0.9005	0.8973	0.8759	0.9057	0.7961
	NA	0.8538	0.8889	0.8796	0.7467	0.8733	0.8793	0.6754
n=200	EL	0.8944	0.9033	0.8891	0.8964	0.8908	0.8810	0.8714
	NA	0.8325	0.8609	0.9279	0.8602	0.7199	0.8924	0.8978
n=500	EL	0.9040	0.9032	0.8958	0.9038	0.8985	0.9017	0.8915
	NA	0.8996	0.8927	0.8761	0.9034	0.8895	0.9552	0.9485

Table 14: Weibull Distribution Distribution, Length of the 90% CI
for the generalized Lorenz ordinate

Weibull(a =1, b=2)

Sample Size	Method	p=0.95	p=0.90	p=0.75	p=0.50	p=0.25	p=0.10	p=0.05
		Length	Length	Length	Length	Length	Length	Length
n=50	EL	0.71125	0.60425	0.40718	0.20859	0.07441	0.03931	0.15821
	NA	0.88685	0.80476	0.44164	0.33107	0.12199	0.00796	0.00412
n=100	EL	0.49633	0.37801	0.29504	0.15555	0.05120	0.02295	0.01645
	NA	0.69727	0.43042	0.41964	0.17596	0.04970	0.01022	0.00373
n=150	EL	0.40534	0.35592	0.24693	0.12255	0.03976	0.01625	0.00895
	NA	0.49039	0.40240	0.27332	0.12982	0.04118	0.00669	0.00550
n=200	EL	0.35204	0.30902	0.21458	0.11385	0.03547	0.01366	0.00797
	NA	0.44204	0.38910	0.23463	0.11756	0.03550	0.00932	0.00437
n=500	EL	0.22917	0.19723	0.13742	0.06878	0.02413	0.00647	0.00397
	NA	0.23238	0.22156	0.14364	0.07269	0.02633	0.00643	0.00375

APPENDIX B: REAL DATA TABLES

APPENDIX B.1: REAL DATA 1: PSID Family 'Income Plus' Files: 2001

Table 15 : Real Example * PSID Family 'Income Plus' Files: 2001 *
Confidence Interval for the generalized Lorenz ordinate (after transforming negative income values)

Method	CI	p=0.95	p=0.90	p=0.75	p=0.50	p=0.25	p=0.10	p=0.05
95%								
NA	Lowerbound	101,020.7448	91,704.2928	69,490.9314	40,770.3948	17,943.6007	6,502.0802	3,099.0990
	Upperbound	102,832.9176	93,246.2829	70,576.1939	41,365.5239	18,204.7612	6,619.1512	3,185.8753
	Length	1,812.1728	1,541.9901	1,085.2625	595.1291	261.1606	117.0710	86.7763
EL	Lowerbound	101,020.7435	91,704.2928	69,490.9320	56,886.8886	17,943.5368	6,461.6020	3,215.7154
	Upperbound	102,832.9163	93,246.2829	70,576.1920	57,141.2642	18,204.6973	6,734.2621	3,692.5957
	Length	1,812.1728	1,541.9901	1,085.2600	254.3756	261.1605	272.6601	476.8804
90%								
Method	CI	p=0.95	p=0.90	p=0.75	p=0.50	p=0.25	p=0.10	p=0.05
NA	Lowerbound	101,166.4195	91,828.2484	69,578.1722	40,818.2353	17,964.5945	6,511.4912	3,106.0747
	Upperbound	102,687.2429	93,122.3273	70,488.9531	41,317.6834	18,183.7674	6,609.7402	3,178.8996
	Length	1,520.8234	1,294.0789	910.7810	499.4481	219.1729	98.2491	72.8250
EL	Lowerbound	101,166.4183	91,828.2491	69,578.1724	54,347.3752	17,964.1023	6,499.5795	3,227.9671
	Upperbound	102,687.3702	93,122.3261	70,488.9518	54,846.3892	18,183.7646	6,511.4889	3,626.0203
	Length	1,520.9519	1,294.0770	910.7794	499.0140	219.6623	11.9094	398.0532
Estimate for θ		101,927	92,475	70,034	41,068	18,074	6,561	3,142

Table 16 : Real Example * PSID Family 'Income Plus' Files: 2001 *
Confidence Interval for the generalized Lorenz ordinate (after erasing negative income values)

Method	CI	p=0.95	p=0.90	p=0.75	p=0.50	p=0.25	p=0.10	p=0.05
95%								
NA	Lowerbound	44268.5930	37894.9745	24706.1227	10942.4414	3068.7481	606.321	175.0999
	Upperbound	46078.3576	39431.8417	25785.0233	11527.7159	3309.7075	686.474	210.9377
	Length	1809.7646	1536.8672	1078.9006	585.2745	240.9594	80.153	35.8378
EL	Lowerbound	44268.5930	37894.9745	24706.1227	10942.4414	3068.7481	606.321	175.6548
	Upperbound	46075.3215	39431.8417	25672.2312	11527.8653	3309.7075	686.474	175.6743
	Length	1806.7285	1536.8672	966.1085	585.4238	240.9594	80.153	0.0195
90%								
NA	Lowerbound	44,414.074	38,018.518	24,792.852	10,989.490	3,088.118	612.764	177.981
	Upperbound	45,932.876	39,308.298	25,698.294	11,480.668	3,290.338	680.031	208.057
	Length	1,518.802	1,289.780	905.442	491.178	202.220	67.267	30.076
EL	Lowerbound	44,414.074	38,018.518	24,792.852	10,990.349	2,883.091	612.764	151.805
	Upperbound	45,931.948	39,308.297	25,698.295	11,480.668	3,088.118	680.031	178.409
	Length	1,517.874	1,289.778	905.443	490.319	205.027	67.267	26.604
Estimate for θ		45,173	38,663	25,246	11,235	3,189	646	193

APPENDIX B.2: REAL DATA 2: Section 8 Housing Median Income Data

Table 17 : Real Example * Section 8 Housing Median Income Data for year 1999 *
Confidence Interval for the generalized Lorenz ordinate

Method	CI	p=0.95	p=0.90	p=0.75	p=0.50	p=0.25	p=0.10	p=0.05
95%								
NA	Lowerbound	40,752.8795	37,572.2040	29,489.2625	18,053.4394	7,853.2409	2,626.9833	1,084.6721
	Upperbound	41,361.1242	38,127.5807	29,940.8642	18,364.8537	8,053.1336	2,766.4932	1,195.6520
	Length	608.2447	555.3767	451.6017	311.4143	199.8927	139.5098	110.9798
EL	Lowerbound	40,752.7028	37,572.2040	29,489.2628	18,053.4386	7,510.1757	1,819.7511	345.9789
	Upperbound	41,361.1457	38,127.5809	29,940.8610	18,278.4669	8,129.3951	2,766.4273	972.8386
	Length	608.4429	555.3769	451.5982	225.0283	619.2193	946.6762	626.8597
90%								
Method	CI	p=0.95	p=0.90	p=0.75	p=0.50	p=0.25	p=0.10	p=0.05
NA	Lowerbound	40,801.7744	37,616.8489	29,525.5653	18,078.4729	7,869.3096	2,638.1981	1,093.5934
	Upperbound	41,312.2294	38,082.9357	29,904.5614	18,339.8201	8,037.0648	2,755.2784	1,186.7307
	Length	510.4551	466.0868	378.9961	261.3471	167.7552	117.0803	93.1372
EL	Lowerbound	40,801.9807	37,616.8489	29,525.5653	17956.30614	7,560.1324	2,020.6458	501.5115
	Upperbound	41,312.5360	38,082.5733	29,904.5614	18339.82064	8,110.4835	2,238.6382	960.7740
	Length	510.5553	465.7243	378.9961	383.5145	550.3511	217.9924	459.2625
Estimate for θ		41,057	37,850	29,715	18,209	7,953	2,697	1,140

Table 18 : Real Example * Section 8 Housing Median Income Data for year 2006 *
Confidence Interval for the generalized Lorenz ordinate

95%

Method	CI	p=0.95	p=0.90	p=0.75	p=0.50	p=0.25	p=0.10	p=0.05
NA	Lowerbound	49,528.5877	45,413.2162	36,004.4579	21,348.1935	9,537.6314	3,504.9984	1,294.8857
	Upperbound	50,292.2351	46,097.2372	36,575.9367	21,727.4992	9,788.3971	3,675.6901	1,424.1739
	Length	763.6474	684.0209	571.4788	379.3056	250.7657	170.6916	129.2882
EL	Lowerbound	49,526.5812	45,413.2159	35,983.8957	20,197.2436	9,569.9835	3,433.7177	1,294.8857
	Upperbound	50,292.2351	46,094.4083	36,575.9367	21,727.4986	9,808.4715	3,505.6269	1,424.1739
	Length	765.6539	681.1923	592.0410	1,530.2551	238.4881	71.9092	129.2882

90%

Method	CI	p=0.95	p=0.90	p=0.75	p=0.50	p=0.25	p=0.10	p=0.05
NA	Lowerbound	49,589.9749	45,468.2025	36,050.3973	21,378.6847	9,557.7896	3,518.7198	1,305.2787
	Upperbound	50,230.8480	46,042.2509	36,529.9974	21,697.0080	9,768.2389	3,661.9687	1,413.7809
	Length	640.8731	574.0485	479.6001	318.3233	210.4492	143.2489	108.5021
EL	Lowerbound	49,589.9749	45,468.2025	36,050.5354	19,148.4061	9,587.3217	3,425.5899	1,305.2787
	Upperbound	50,230.7866	45,952.5094	36,529.9990	21,877.3285	9,789.6583	3,519.1642	1,413.7809
	Length	640.8118	484.3070	479.4636	2,728.9224	202.3367	93.5743	108.5021
Estimate for θ		49,910	45,755	36,290	21,538	9,663	3,590	1,360

APPENDIX C: R/Splus CODES FOR SIMULATIONS

```
#The population income distribution has a p.d.f f(x)=weibull(a,b)
#The Weibull distribution with shape parameter a and scale parameter b
#The cumulative distribution function is F(x) = 1 - exp(- (x/b)^a) on x >= 0,
#the mean is E(X) = b*Gamma(1 + 1/a)
#and  Var(X) = b^2*(Gamma(1 + 2/a)-(Gamma(1 + 1/a))^2)

#step1: estimation of true generalized Lorenz ordinate “theta” and estimation
of true Lorenz ordinate “eta”

solveNonlinear <- function(f, y0, x, ...)
{
  # solve f(x) = y0
  # x is vector of initial guesses, same length as y0
  # ... are additional arguments to nlmin (not to f)
  g <- function(x, y0, f) sum((f(x) - y0)^2)
  g$y0 <- y0   # set g's default value for y0
  g$f <- f     # set g's default value for f
  nlmin(g, x, ...)
}


lowerNAg1<-0
upperNAg1<-0
lengthNAg1<-0

lowerNAg2<-0
upperNAg2<-0
lengthNAg2<-0

lowerELg1<-0
upperELg1<-0
lengthELg1<-0

lowerELg2<-0
upperELg2<-0
lengthELg2<-0

coverageNAg1<-0
coverageELg1<-0
coverageNAg2<-0
coverageELg2<-0

biastheta<-0
MSEtheta<-0

thetahat<-0
Thetahati<-0
etahat<-0
rhat1<-0
rhat2<-0
x<-0

alpha1<-0.05
alpha2<-0.1

m<-10000                      # number of iterations
n<-500                        # sample size 50,100,150,200,500
```

```
a<-1                                    # shape parameter of weibull
distribution

b<-1                                    # scale parameter of weibull distribution

p<-0.95                  # percentiles 0.95, 0.90, 0.75 0.5, 0.25, 0.1 0.05

mu<-b*gamma(1+1/a)
kxi<-qweibull(p, a, b)

f<-function(y,p=0.95,a=1,b=1){dweibull(y,a,b)*y}

theta<- integrate(f,0,qweibull(p,a,b))$integral # generalized Lorenz ordinate
eta<-theta/mu          # Lorenz ordinate

#step2: estimation of sample generalized Lorenz ordinate “thetahat” and
estimation of sample Lorenz ordinate “etahat”

for(i in 1:m)
{
      x<-rweibull(n,a,b)      # generating a random sample of size n
      kxihat<-quantile(x,p)               # the p-th sample quantile

      Thetahati<-x*(x <= kxihat)    # Truncated X by the sample quantile

      thetahat[i]<- mean(Thetahati)         # estimator for theta

      Truncatedx<-(x-kxihat)*(x <=kxihat) # Truncated (x-kxihat)

      sigmap<-var(Thetahati)
      sigmav<-var(Truncatedx)

      # step3: Normal Approximation of CI for the generalized Lorenz ordinate

      marginoferrorV1<-qnorm(1-alpha1/2)*sqrt(sigmav)/sqrt(n)
      lowerNAg1<-thetahat- marginoferrorV1
      upperNAg1<-thetahat+ marginoferrorV1

      marginoferrorV2<-qnorm(1-alpha2/2)*sqrt(sigmav)/sqrt(n)
      lowerNAg2<-thetahat- marginoferrorV2
      upperNAg2<-thetahat+ marginoferrorV2

      lengthNAg1<- upperNAg1- lowerNAg1
      lengthNAg2<- upperNAg2- lowerNAg2

      coverageNAg1<-(lowerNAg1<= theta)*(theta <= upperNAg1)
      coverageNAg2<-(lowerNAg2<= theta)*(theta <= upperNAg2)


      # step4: Empirical likelihood of CI for the generalized Lorenz ordinate


      # length of CI for alpha 1

      # X[1]: theta  X[2]:  lambda


      g1 <- function(X) c( mean((Thetahati-X[1])/(1 + X[2]*(Thetahati-
      X[1]))), 2*rhat1*sum( log( abs(1 + X[2]*(Thetahati-X[1]))))-qchisq(1-
      alpha1,1))

      lowerELg1[i]<-solveNonlinear(g1, c(0,0), c(mean(lowerNAg1),0))$x[1]
```

```
      upperELg1[i]<-solveNonlinear(g1, c(0,0), c(mean(upperNAg1),0))$x[1]
      lengthELg1<-upperELg1-lowerELg1

      # Length of the CI for alpha 2

      g2 <- function(X) c( mean((Thetahati-X[1])/(1 + X[2]*(Thetahati-
      X[1]))), 2*rhat1*sum( log( abs(1 + X[2]*(Thetahati-X[1]))))-qchisq(1-
      alpha2,1))

      lowerELg2[i]<-solveNonlinear(g2, c(0,0), c(mean(lowerNAg2),0))$x[1]
      upperELg2[i]<-solveNonlinear(g2, c(0,0), c(mean(upperNAg2),0))$x[1]
      lengthELg2<-upperELg2-lowerELg2


      # Coverage of CI

      Vihat<-Thetahati-theta
      rhat1<-sigmap/sigmav                # scale constant

      funclambda<-function(lam)mean(Vihat/(1+lam*Vihat))

      lambda<-solveNonlinear(funclambda, c(0), c(0.01))$x

      l1theta<-2*sum(log(abs(1+lambda*Vihat)))

      coverageELg1[i]<-(rhat1*l1theta <= qchisq(1-alpha1,1))*1

      coverageELg2[i]<-(rhat1*l1theta <= qchisq(1-alpha2,1))*1


# step5: bias, MSE calculations and printing results

biastheta[i]<- thetahat[i]-theta

MSEtheta[i] <- (thetahat[i]-theta)^2

}

NAgcoverage1<-mean(sort(coverageNAg1))
ELgcoverage1<-mean(sort(coverageELg1))

NAgcoverage2<-mean(sort(coverageNAg2))
ELgcoverage2<-mean(sort(coverageELg2))

cat("n=", n, "a=", a, "b=", b, "p=", p, "kxip=", kxi, "\n")

cat("theta=", theta, "\n")
cat("thetahat=", mean(thetahat), "\n")

cat("biastheta=", mean(biastheta),  "\n")
cat("MSEtheta =", mean(MSEtheta) , "\n")

cat("coverage of level 0.95 EL CI for theta is:", ELgcoverage1, "\n")
cat("coverage of level 0.95 NA CI for theta is:", NAgcoverage1, "\n")

cat("coverage of level 0.90 EL CI for theta is:", ELgcoverage2, "\n")
cat("coverage of level 0.90 NA CI for theta is:", NAgcoverage2, "\n")


cat("length of 0.95 EL CI for theta :", mean(lengthELg1), "\n")
cat("length of 0.95 NA CI for theta :", mean(lengthNAg1), "\n")
```

```
cat("length of 0.90 EL CI for theta :", mean(lengthELg2), "\n")
cat("length of 0.90 NA CI for theta :", mean(lengthNAg2), "\n")
```

APPENDIX D: R/S-PLUS CODE FOR REAL EXAMPLES

```
solveNonlinear <- function(f, y0, x, ...)
{
  # solve f(x) = y0
  # x is vector of initial guesses, same length as y0
  # ... are additional arguments to nlmin (not to f)
  g <- function(x, y0, f) sum((f(x) - y0)^2)
  g$y0 <- y0   # set g's default value for y0
  g$f <- f     # set g's default value for f
  nlmin(g, x, ...)
}

#step1: reading the data

lowerNAg1<-0
upperNAg1<-0
lengthNAg1<-0

lowerNAg2<-0
upperNAg2<-0
lengthNAg2<-0

lowerELg1<-0
upperELg1<-0
lengthELg1<-0

lowerELg2<-0
upperELg2<-0
lengthELg2<-0


thetahat<-0
Thetahati<-0
etahat<-0
x<-0
kxihat<-0

alpha1<-0.05
alpha2<-0.1

data<-read.table("K:/Thesis Research/data/data2.txt", header=T)$Income

#x<- data-min(data) # use this shift in the presence of negative values

x<- data          # use this instead if there are no negative values

n<-length(x)                                # sample size


meanx<-mean(x)                           # mean of the data
stdvx<-stdev(x)                          # stdev of the data

p<-0.05           # percentiles 0.95, 0.90, 0.75 0.5, 0.25, 0.1 0.05


#step2: estimation of sample generalized Lorenz ordinate “thetahat” and
estimation of sample Lorenz ordinate “etahat”
```

```
kxihat<-quantile(x,p)               # the p-th sample quantile

Thetahati<-x*(x <= kxihat)    # Truncated X by the sample quantile

thetahat<- mean(Thetahati)  # estimator for generalized Lorenz ordinate

Truncatedx<-(x-kxihat)*(x <=kxihat) # Truncated (x-kxihat)

sigmap<-var(Thetahati)
sigmav<-var(Truncatedx)
```

```
# step3: Normal Approximation of CI for the generalized Lorenz ordinate

      marginoferrorV1<-qnorm(1-alpha1/2)*sqrt(sigmav)/sqrt(n)
      lowerNAg1<-thetahat- marginoferrorV1
      upperNAg1<-thetahat+ marginoferrorV1
      lengthNAg1<- upperNAg1-lowerNAg1

      marginoferrorV2<-qnorm(1-alpha2/2)*sqrt(sigmav)/sqrt(n)
      lowerNAg2<-thetahat- marginoferrorV2
      upperNAg2<-thetahat+ marginoferrorV2
      lengthNAg2<- upperNAg2-lowerNAg2


# step4: Empirical likelihood of CI for the generalized Lorenz ordinate

      # X[1]: theta  X[2]:  lambda

      rhat1<-sigmap/sigmav                # scale constant

      # length of CI for alpha 1

      g1 <- function(X) c( mean((Thetahati-X[1])/(1 + X[2]*(Thetahati-
      X[1]))), 2*rhat1*sum( log( abs(1 + X[2]*(Thetahati-X[1]))))-qchisq(1-
      alpha1,1))

      lowerELg1<-solveNonlinear(g1, c(0,0), c(lowerNAg1, 0.001))$x[1]
      upperELg1<-solveNonlinear(g1, c(0,0), c(upperNAg1, 0.001))$x[1]
      lengthELg1<-upperELg1- lowerELg1

      # length of CI for alpha 2

      g2 <- function(X) c( mean((Thetahati-X[1])/(1 + X[2]*(Thetahati-
      X[1]))), 2*rhat1*sum( log( abs(1 + X[2]*(Thetahati-X[1]))))-qchisq(1-
      alpha2,1))

      lowerELg2<-solveNonlinear(g2, c(0,0), c(lowerNAg2, 0.001))$x[1]
      upperELg2<-solveNonlinear(g2, c(0,0), c(upperNAg2, 0.001))$x[1]
      lengthELg2<-upperELg2- lowerELg2
```

```
# step5: printing results for the normal approximation

      cat("n=", n, "p=", p, "kxi=", kxihat, "\n")

      cat("lowerbound of 95% NA CI for theta:", lowerNAg1, "\n")
      cat("upperbound of 95% NA CI for theta:", upperNAg1, "\n")
      cat("length of of 95% NA CI for theta :", lengthNAg1, "\n")

      cat("lowerbound of 95% EL CI for theta:", lowerELg1, "\n")
      cat("upperbound of 95% EL CI for theta:", upperELg1, "\n")
      cat("length of of 95% EL CI for theta :", lengthELg1, "\n")


      cat("thetahat", thetahat, "\n")

      cat("-------------------------------------------------",  "\n")


      cat("lowerbound of 90% NA CI for theta:", lowerNAg2, "\n")
      cat("upperbound of 90% NA CI for theta:", upperNAg2, "\n")
      cat("length of 90% NA CI for theta :", lengthNAg2, "\n")

      cat("lowerbound of 90% EL CI for theta:", lowerELg2, "\n")
      cat("upperbound of 90% EL CI for theta:", upperELg2, "\n")
      cat("length of 90% EL CI for theta :", lengthELg2, "\n")

      cat("thetahat", thetahat, "\n")
```