U-Statistics and Their Asymptotic Results for Some Inequality and Poverty Measures (Long Version)

Kuan Xu

Department of Economics

Dalhousie University

Halifax, Nova Scotia

Canada B3H 3J5

Email: kuan.xu@dal.ca

Current Version: June 2006

Abstract

U-statistics form a general class of statistics which have certain important features in common. This class arises as a generalization of the sample mean and the sample variance and typically members of the class are asymptotically normal with good consistency properties. The class encompasses some widely-used income inequality and poverty measures, in particular the variance, the Gini index, the poverty rate, average poverty gap ratios, the Foster-Greer-Thorbecke index, the Sen index and its modified form. This paper illustrates how these measures come together within the class of U-statistics, and thereby why U-statistics are useful in econometrics.

JEL Codes: C100, I300

Keywords, U-statistics, inequality, poverty, and measures

Acknowledgment: I wish to thank anonymous referees, Esfandiar Maasoumi (the Editor), Gordon Fisher, Yulia Kotlyarova, Stéphane Mussard and participants at the Statistics Colloquium, Department of Mathematics and Statistics, Dalhousie University, for helpful suggestions and comments. Further comments are welcome.

1 Introduction

Sound income inequality and poverty measures and reliable statistical procedures have become increasingly important in policy making. When economists measure income inequality or poverty on the basis of sample data, they need to select appropriate statistical methods for each chosen measure. There are generally two broad categories of methods in the literature: asymptotic-theory-based and simulation-based methods.

In the first category, various methods have been proposed. Halmos (1946) initiated the discussion of U-statistics. Hoeffding (1948) generalized the results of U-statistics and, based on this class, discussed the Gini index as a function of U-statistics. This approach was revived by Glasser (1962) and Gastwirth (1972) for both the Gini index and Lorenz curves. Then, Gail and Gastwirth (1978) and Sandstrom, Wretman, and Walden (1988) considered statistical inference for the Gini index along similar lines.[1] More recently, U-statistics have been used primarily for the Sen index of poverty intensity and its various extensions by Bishop, Formby and Zheng (1997, 1998, 2001), and Zheng, Formby, Smith and Chow (2000).[2] Obviously, this approach can be further extended. As noted by Xu and Osberg (2002), the Sen and modified Sen indices of poverty intensity share a similar mathematical structure, which ensures that U-statistics are applicable to the modified Sen index within a

[1] Along somewhat different paths, Nygård and Sandström (1981) and Aaberge (1982) discussed the issues of statistical inference for the Gini index. For example, Nygård and Sandström (1981) used the approach of Sendler (1979).

[2] Differing from the approach of Bishop, Chakraborti, and Thistle (1990), the use of U-statistics avoids the need to employ a finite number of quantiles or order statistics to compute income inequality and poverty measures.

more general framework.[3]

In the second category of methods, statistical inference is made on the basis of simulation. Yitzhaki (1991) and Karagiannis and Kovacivic (2000) propose the jackknife for the Gini index. Xu (1998) and Osberg and Xu (2000) advocate the bootstrap for the modified Sen index of poverty intensity and its components. Biewen (2002) provides a comprehensive review on the bootstrap for the family of generalized entropy measures, Atkinson indices, the coefficient of variation, the logarithmic variance, Kolm indices, Maasoumi-Zandvakili-Shorrocks mobility indices, Prais mobility indices, and the Foster-Greer-Thorbecke poverty index. While Biewen does not consider the Gini index and the Sen and modified Sen indices, he does mention the usefulness of U-statistics for the Gini index.[4]

Can U-statistics be used for all of these widely-used income inequality and poverty measures? This paper shows that U-statistics can be applied to the variance, the Gini index, the poverty rate, mean poverty gap ratios, the Foster-Greer-Thorbecke (FGT) index, the Sen index, and the modified Sen index. This generalization is achieved because the sample counterparts of these inequality and poverty measures can be expressed either as U-statistics themselves or as functions of U-statistics.[5] Within the framework of U-statistics, this paper therefore provides the suitable estimators for these important income inequality and poverty measures and develops the asymptotic

[3] Anderson (2004) notes that some statistical procedures with point-wise estimation and comparison of underlying distributions are biased. This observation further justifies the use of U-statistics.

[4] Ogwang (2000, 2004) proposes a simplified approach for the Gini index based on the jackknife.

[5] Within this broader category, it should be noted that a number of authors propose simplified methods for computing the Gini index [see Giles (2004)].

distributions for these estimators.

The remainder of the paper is organized as follows. Section 2 introduces the basic notation and definitions of various inequality and poverty measures. Section 3 explains U-statistics and their application to these measures. Finally, concluding remarks are given in Section 4.

2 Inequality and Poverty Measures

Let F_y and f_y be the probability distribution function and probability density function, respectively, for income y with the support $[0, +\infty)$.[6] Let $0 < z < +\infty$ be the poverty line. Let the indicator function be: $I(A) = 1$ if A is true; $I(A) = 0$ otherwise. The poverty gap ratio of the population is defined as

$$x = \left(\frac{z - y}{z}\right) I(y < z). \tag{1}$$

The poverty gap ratio of the non-poor is zero. The poverty gap ratio of the poor is therefore $x_p = \{x | 0 < x \leq 1\}$.

The variance and the Gini index are the simplest measures of income inequality.[7] The variance is defined as

$$\sigma_y^2 = \int_0^{+\infty} \left(y - \mu_y\right)^2 dF_y(y) \tag{2}$$

[6]Both F_y and f_y can accommodate either discrete distributions, or continuous distributions or a combination of the two.

[7]The variance of logarithms is another inequality measure, the estimator of which can be viewed as a U-statistic. But this measure is not desirable, as pointed by Sen (1973), because it violates the principle of transfer. Foster and Ok (1999) also find that the variance of logarithms is not only inconsistent with the Lorenz dominance criterion but is also capable of making very serious errors.

where $\mu_y = \int_0^{+\infty} y dF_y(y)$ is the mean of income y. This measure possesses good theoretical properties—absolute inequality invariance, symmetry, the principle of the transfers, the principle of population, and subgroup decomposability [see Chakravarty (2001a, 2001b)] and, hence, is used widely [see Sen (1973) and Chakravarty (1990)].

The Gini index, which is defined in equation (5) below, is probably the most widely-used measure of income inequality. This measure can be defined in various ways [see Yitzhaki(1998) and Xu (2003)]. The relative mean difference approach links the Gini index directly to U-statistics [see Hoeffding (1948)] while the normative approach links the Gini index directly to the Sen and modified Sen indices [Xu and Osberg (2002)].

The absolute mean difference is defined as the mean difference between any two variates of the same distribution function F_y:

$$\Delta_y = E|y_i - y_j| \tag{3}$$

where E is the mathematical expectation operator and y_i and y_j are the variates from the same distribution F_y. The relative mean difference is the mean-scaled absolute mean difference:

$$\frac{\Delta_y}{\mu_y} = \frac{E|y_i - y_j|}{\mu_y}. \tag{4}$$

The Gini index is defined as half of the relative mean difference:

$$G_y = \frac{\Delta_y}{2\mu_y}. \tag{5}$$

Defining an inequality or poverty measure with respect to a class of underlying social welfare functions is called the normative approach to income inequality or poverty. Given the Gini social welfare function

$$W_G(y) = 2\int_0^{+\infty} y(1 - F_y(y))dF_y(y) \tag{6}$$

and the corresponding equally-distributed-equivalent income (EDEI),[8]

$$\Xi_G(y) = \frac{W_G(y)}{W_G(1)} = \frac{W_G(y)}{1} = 2\int_0^{+\infty} y(1 - F_y(y))dF_y(y), \tag{7}$$

the Gini index can then be defined using $\Xi_G(y)$ as

$$G_y = \frac{\mu_y - \Xi_G(y)}{\mu_y}, \tag{8}$$

which implies

$$\Xi_G(y) = \mu_y(1 - G_y). \tag{9}$$

The most popular poverty measure is the poverty rate or headcount ratio:

$$H = P(y < z) = F_y(z) = \int_0^{+\infty} I(y < z)dF_y(y), \tag{10}$$

which indicates the proportion of the population whose incomes fall below the poverty line z. The other two often-cited poverty measures are the mean

[8]For the discrete distribution $W_G(y) = \int_0^{+\infty} y(1 - F_y(y))dF_y(y) = \frac{1}{n^2}\sum_{i=1}^{n}(2n - 2i + 1)y_i$. The term $\frac{1}{n^2}\sum_{i=1}^{n}(2n-2i+1)y_i$ can be rewritten as $\frac{2}{n}\sum_{i=1}^{n}(1 - \frac{i}{n} + \frac{1}{2n})y_i$, in which $(1 - \frac{i}{n} + \frac{1}{2n})$ is a discrete representation of the rank-based weight in the continuous case $(1 - F_y(y))$.

poverty gap ratio of the poor:

$$\mu_{x_p} = \frac{\int_0^{+\infty} I(y < z)\frac{z-y}{z}dF_y(y)}{\int_0^{+\infty} I(y < z)dF_y(y)} \tag{11}$$

and the mean poverty gap ratio of the population:

$$\mu_x = H\mu_{x_p} = \int_0^{+\infty} I(y < z)\frac{z-y}{z}dF_y(y). \tag{12}$$

The former measures the depth of poverty among the poor while the latter gauges the depth of poverty of the whole population.

The poverty rate and mean poverty gap ratios are criticized by Sen (1976) because each alone cannot capture all important dimensions (incidence, depth and inequality) of poverty. It is also worth noting the FGT index of poverty proposed by Foster, Greer, and Thorbecke (1984) is closely related to H and μ_{x_p}. The FGT index of poverty with order α is defined as

$$FGT_\alpha = \int_0^{+\infty} I(y < z)\left(\frac{z-y}{z}\right)^\alpha dF_y(y). \tag{13}$$

The parameter α can be set to 0, 1, 2, 3, etc. and the higher the value of α the higher the degree of poverty aversion that is imposed on the FGT index. If $\alpha = 0$, then $FGT_0 = H$. If $\alpha = 1$, $FGT_1 = H\mu_{x_p} = \mu_x$. When $\alpha \geq 2$, FGT_α can measure the degree of inequality. It is its additive decomposability that makes the FGT index attractive to applied researchers.

The Sen index of poverty intensity (S), which incorporates the incidence, depth and inequality of poverty simultaneously, can be defined according to

Figure 2 in Xu and Osberg (2002, p. 148):[9]

$$S = 2\int_0^{+\infty} I(y<z)\left(\frac{z-y}{z}\right)\left(\frac{1-F_y(y)}{F_y(z)}\right)dF_y(y). \tag{14}$$

The Sen index can be viewed as the product of three poverty measures—H (incidence), μ_{x_p}(depth), and $(1-G_{x_p})$ (inequality):

$$S = H\cdot\mu_{x_p}\cdot(1-G_{x_p}) \tag{15}$$

as shown in Xu and Osberg (2002).[10] In view of equation (9), it can be shown

[9]When the income distribution is discrete, given that q y_i's out of n y_i's are less than z,

$$S = \frac{2}{q}\sum_{i=1}^{q}\left(\frac{z-y_i}{z}\right)\left(\frac{1-\frac{i}{n}+\frac{1}{2n}}{\frac{q}{n}}\right) = \frac{1}{q^2}\sum_{i=1}^{q}\left(\frac{z-y_i}{y_i}\right)(2n-2i+1).$$

Note that $\left(\frac{1-\frac{i}{n}+\frac{1}{2n}}{\frac{q}{n}}\right)$ corresponds to $\left(\frac{1-F_y(y)}{F_y(z)}\right)$ in the continuous case.

[10]Note that Bishop, Formby, and Zheng (1997) have derived asymptotic statistical results for

$$S = H\left[\mu_{x_p}+(1-\mu_{x_p})G_{y_p}\left(\frac{q}{q+1}\right)\right]$$

where G_{y_p} is the Gini index of incomes of the poor (y_p). This paper focuses on the large sample version of the above

$$S = H\left[\mu_{x_p}+(1-\mu_{x_p})G_{y_p}\right]$$

where as $n\to\infty$, $q/(q+1)\to 1$. Xu and Osberg (2002) have noted that if G_{x_p} is computed based on x_p arranged in non-increasing order, then

$$S = H\cdot\mu_{x_p}\cdot(1-G_{x_p});$$

otherwise, we have

$$S = H\cdot\mu_{x_p}\cdot(1+G_{x_p}).$$

A similar argument can be made for

$$S_m = H\cdot\mu_{x_p}\cdot(1-G_x).$$

that

$$
\begin{aligned}
S &= H \cdot \mu_{x_p} \cdot (1 - G_{x_p}) \\
&= H \cdot \Xi_G(x_p) \\
&= \int_0^{+\infty} I(y<z)dF_y(y) \frac{2\int_0^{+\infty} I(y<z)\frac{z-y}{z}\left(\frac{1-F_y(y)}{F_y(z)}\right)dF_y(y)}{\int_0^{+\infty} I(y<z)dF_y(y)} \\
&= 2\int_0^{+\infty} I(y<z)\frac{z-y}{z}\left(\frac{1-F_y(y)}{F_y(z)}\right)dF_y(y).
\end{aligned} \tag{16}
$$

The modified Sen index can be defined according to Figure 3 in Xu and Osberg (2002, p. 149):[11]

$$
S_m = 2\int_0^{+\infty} I(y<z)\left(\frac{z-y}{z}\right)(1-F_y(y))\,dF_y(y), \tag{17}
$$

which is the product of three poverty measures—H (incidence), μ_{x_p}(depth), and $(1 - G_x)$ (inequality):[12]

$$
S_m = H \cdot \mu_{x_p} \cdot (1 - G_x) \tag{18}
$$

as shown in Xu and Osberg (2002). Because $H\mu_{x_p} = \mu_x$ and in view of

[11] When the income distribution is discrete,

$$
S_m = \frac{2}{n}\sum_{i=1}^{q}\left(\frac{z-y_i}{z}\right)\left(1-\frac{i}{n}+\frac{1}{2n}\right) = \frac{1}{n^2}\sum_{i=1}^{q}\left(\frac{z-y_i}{z}\right)(2n-2i+1).
$$

Note that $\left(1-\frac{i}{n}+\frac{1}{2n}\right)$ corresponds to $(1 - F_y(y))$ in the continuous case. Both are the rank-based weights.

[12] When the income distribution is discrete, given that q y_i's out of n y_i's are less than z,

$$
S_m = \frac{2}{n}\sum_{i=1}^{q}\left(\frac{z-y_i}{z}\right)\left(1-\frac{i}{n}+\frac{1}{2n}\right) = \frac{1}{n^2}\sum_{i=1}^{q}\left(\frac{z-y_i}{y_i}\right)(2n-2i+1).
$$

Note that $\left(1-\frac{i}{n}+\frac{1}{2n}\right)$ corresponds to $(1 - F_y(y))$.

equation (9), it can be shown that

$$\begin{aligned} S_m &= H \cdot \mu_{x_p} \cdot (1 - G_x) \\ &= \mu_x \cdot (1 - G_x) \\ &= \Xi_G(x) \\ &= 2\int_0^{+\infty} I(y < z)\tfrac{y-z}{z}(1 - F_y(y))dF_y(y). \end{aligned} \tag{19}$$

3 Statistical Inference Using U-Statistics

In this section, the links between U-statistics and the inequality and poverty measures are examined on the basis of one-sample U-statistics.[13]

Consider a generic estimable parameter (θ) of the population distribution function F_y:

$$\theta = \int \cdots \int \varphi(y_1, \ldots, y_m) dF_y(y_1) \cdots dF_y(y_m) \tag{20}$$

where $\varphi(y_1, \ldots, y_m)$ is a symmetric function of m independent identically distributed (i.i.d.) random variables, called the kernel for θ.[14] The smallest integer m is called the order of θ. Then the corresponding estimator (U) of the parameter θ, called a U-statistic, is defined as the function of an i.i.d.

[13] For a general introduction to the U-statistics, see Hoeffding (1948), Randles and Wolfe (1979), Serfling (1980), Lee (1990), and Bishop, Formby, and Zheng (1998).

[14] When the kernel $\varphi^*(\cdot)$ is not symmetric, it can be modified to be symmetric by using $\varphi(y_1, \ldots, y_m) = \frac{1}{m!}\sum_{\beta \in \mathcal{B}} \varphi^*(y_1, \ldots, y_m)$ where the summation is over $\mathcal{B} = \{\beta | \beta$ is a permutation of the integers $1, \ldots, m\}$.

sample $\{y_1, y_2, \ldots, y_n\}$ from F_y:

$$U = \frac{1}{\begin{pmatrix} n \\ m \end{pmatrix}} \sum_{\alpha \in \mathcal{A}} \varphi(y_{\alpha_1}, \ldots, y_{\alpha_m}) \tag{21}$$

where $\mathcal{A}$ is the collection of all $\begin{pmatrix} n \\ m \end{pmatrix}$ unordered subsets of m integers chosen without replacement from the set $\{1, 2, \ldots, n\}$ and α is any one of those unordered subsets. It can be shown that the expected value of the U-statistic U is θ; that is, $E(U) = \theta$.

The sample proportion, a U-statistic for the estimable parameter of order 1 ($\theta_1 = F_y(z)$) with the symmetric kernel for $\varphi_1(y) = I(y < z)$, is given by

$$U_1 = \widehat{F}_y(z) = \frac{1}{\begin{pmatrix} n \\ 1 \end{pmatrix}} \sum_{i=1}^{n} I(y_i < z) = \frac{1}{n} \sum_{i=1}^{n} I(y_i < z) = \int_0^{+\infty} I(y < z) d\widehat{F}_y(y),$$

where $\widehat{F}_y$ is the empirical counterpart of F_y. U_1 is an unbiased estimator for $\theta_1 = F_y(z) = \int_0^{+\infty} I(y < z) dF_y(y)$. U_1 can be viewed as the estimator of the poverty rate $\widehat{H}$. The sample mean, another U-statistic for the estimable parameter of order 1 ($\theta_2 = \mu_y$) with the symmetric kernel for the $\varphi_2(y) = y$, is given by

$$U_2 = \widehat{\mu}_y = \frac{1}{\begin{pmatrix} n \\ 1 \end{pmatrix}} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{i=1}^{n} y_i = \int_0^{+\infty} y d\widehat{F}_y(y), \tag{22}$$

which is an unbiased estimator for $\theta_2 = \int_0^{+\infty} y dF_y(y)$. This U-statistic has a number of commonly seen examples such as the sample mean income ($\widehat{\mu}_y$), the sample mean poverty gap ratio of the poor ($\widehat{\mu}_{x_p}$), and the sample mean poverty gap ratio of the population ($\widehat{\mu}_x$). The sample variance, another U-statistic for the estimable parameter of order 2 ($\theta_3 = \sigma_y^2$) with the symmetric kernel $\varphi_3(y_1, y_2) = \frac{1}{2}(y_1 - y_2)^2$,[15] is defined as[16]

$$U_3 = \widehat{\sigma}_y^2 = \frac{1}{\begin{pmatrix} n \\ 2 \end{pmatrix}} \sum_{i<j} \frac{1}{2} (y_i - y_j)^2 = \frac{1}{(n-1)} \left(\sum_{i=1}^{n} y_i^2 - n\widehat{\mu}_y^2 \right)$$

$$= \int_0^{+\infty} (y - \mu_y)^2 d\widehat{F}_y(y), \qquad (23)$$

[15] Note that this kernel is made to be symmetric from a more intuitive but nonsymmetric one. For $\sigma_y^2 = E(y^2) - E(y)E(y)$, a more intuitive but nonsymmetric kernel is either $\varphi_1 = y_1^2 - y_1 y_2$ or $\varphi_2 = y_2^2 - y_2 y_1$. To make it symmetrical, the new kernel is the average of two nonsymmetric kernels.

$$\varphi_3(y_1, y_2) = \frac{1}{2}(\varphi_1 + \varphi_2) = \frac{1}{2}[(y_1^2 - y_1 y_2) + (y_2^2 - y_2 y_1)] = \frac{1}{2}(y_1 - y_2)^2.$$

[16] Note that $\sum_{i<j}$ represents the sum of all cases where $1 \leq i < j \leq n$ while $\sum_{i=1}^{n}$ represents the sum from $i = 1$ to $i = n$. Therefore,

$$\begin{aligned}
\widehat{\sigma}_y^2 &= \frac{1}{\begin{pmatrix} n \\ 2 \end{pmatrix}} \sum_{i<j} \tfrac{1}{2} (y_i - y_j)^2 \\
&= \tfrac{2}{n(n-1)} \sum_{i<j} \tfrac{1}{2} (y_i - y_j)^2 \\
&= \tfrac{1}{n(n-1)} \sum_{i<j} (y_i - y_j)^2 \\
&= \tfrac{1}{n(n-1)} \sum_{i<j} \left(y_i^2 + y_j^2 - 2y_i y_j\right) \\
&= \tfrac{1}{n(n-1)} \left[n \sum_{i=1}^{n} y_i^2 - n \left(\sum_{i=1}^{n} y_i\right)^2 \right] \\
&= \tfrac{1}{(n-1)} \sum_{i=1}^{n} \left[y_i^2 - \tfrac{n}{n^2} \left(\sum_{i=1}^{n} y_i\right)^2 \right] \\
&= \tfrac{1}{(n-1)} \left(\sum_{i=1}^{n} y_i^2 - n\widehat{\mu}_y^2 \right).
\end{aligned}$$

which is an unbiased estimator for $\theta_3 = \int_0^{+\infty}(y-\mu_y)^2 dF_y(y)$. The population variance is a measure of income inequality.[17]

It is now useful to introduce the variance of a generic U-statistic U. First, let the conditional expectation of the kernel function of order m be

$$\varphi_c(y_1, y_2, \ldots, y_c) = E[\varphi(y_1, y_2, \ldots, y_c, Y_{c+1}, \ldots, Y_m)] \tag{24}$$

on the basis of c ($c < m$) out of m i.i.d. random variables (while $Y_{c+1}, \ldots, Y_m$ are not conditioned on) and its variance be

$$\zeta_c = Var[\varphi_c(y_1, y_2, \ldots, y_c)] = E[\varphi_c^2(y_1, y_2, \ldots, y_c)] - \theta_c^2, \tag{25}$$

where θ_c is the mean of $\varphi_c(y_1, y_2, \ldots, y_c)$.[18] Second, it can be shown that the variance of the generic U-statistic, $Var(U)$, for the estimable parameter θ of order m is given by

$$Var(U) = \begin{pmatrix} n \\ m \end{pmatrix}^{-1} \sum_{i=1}^{m} \begin{pmatrix} m \\ i \end{pmatrix} \begin{pmatrix} n-m \\ m-i \end{pmatrix} \zeta_i. \tag{26}$$

The terms in $\begin{pmatrix} n \\ m \end{pmatrix}^{-1} \begin{pmatrix} m \\ i \end{pmatrix} \begin{pmatrix} n-m \\ m-i \end{pmatrix}$ convey useful information: there are $\begin{pmatrix} n \\ m \end{pmatrix}$ ways selecting m out of n elements; then there are $\begin{pmatrix} m \\ i \end{pmatrix}$ ways

[17]The sample absolute mean difference is also a common example and will be introduced later. The coefficient of variation is a function of two U-statistics—the sample mean and variance.

[18] Note that c represents any c of m observations. Fraser (1957, p. 224–225) and Randles and Wolfe (1979, p. 64-65) provide an intuitive explanation on ζ_c.

of selecting i out of m elements; and $\begin{pmatrix} n-m \\ m-i \end{pmatrix}$ ways of selecting $(m-i)$ out of the remaining $(n-m)$ elements.

The above definition of the variance of a generic U-statistic can be used to find the variances for the following three U-statistics—the sample proportion, the sample mean and the sample variance. For the sample proportion, $m = 1$, $\varphi_1(y_1) = I(y < z)$, y_1 is known or conditioned on. and

$$Var(\widehat{F}_y(z)) = \begin{pmatrix} n \\ 1 \end{pmatrix}^{-1} \sum_{i=1}^{1} \begin{pmatrix} 1 \\ i \end{pmatrix} \begin{pmatrix} n-1 \\ 1-i \end{pmatrix} \zeta_i = \frac{1}{n} Var(I(y < z))$$

$$= \frac{1}{n} \left\{ \int_0^{+\infty} [I(y < z)]^2 \, dF_y(y) - (F_y(z))^2 \right\} = \frac{F_y(z)(1 - F_y(z))}{n}. \tag{27}$$

For the sample mean, $m = 1$, $\varphi_1(y_1) = y_1$, y_1 is known or conditioned on, and

$$Var(\widehat{\mu}_y) = \begin{pmatrix} n \\ 1 \end{pmatrix}^{-1} \sum_{i=1}^{1} \begin{pmatrix} 1 \\ i \end{pmatrix} \begin{pmatrix} n-1 \\ 1-i \end{pmatrix} \zeta_i = \frac{1}{n} Var(y_1) = \frac{\sigma_y^2}{n}. \tag{28}$$

For the sample variance, $m = 2$ and a few steps must be taken. Given that y_1 is known,

$$\varphi_1(y_1) = E\left[\frac{1}{2}(y_1 - Y_2)^2\right] = \frac{1}{2}[\sigma_y^2 + (y_1 - \mu_y)^2]. \tag{29}$$

When both y_1 and y_2 are known,

$$\varphi_2(y_1, y_2) = \frac{1}{2}(y_1 - y_2). \tag{30}$$

Here, the derivation of $\varphi_1(y_1)$ is based on $\sigma_y^2 = E(Y_2^2) - \mu_y^2$. From the above,

$$\zeta_1 = Var\left\{\frac{1}{2}[\sigma_y^2 + (y_1 - \mu_y)^2]\right\} = \frac{1}{4}(\mu_4 - \sigma_y^4), \tag{31}$$

and

$$\zeta_2 = Var\left[\frac{1}{2}(y_1 - y_2)^2\right] = \frac{1}{2}(\mu_4 + \sigma_y^4), \tag{32}$$

where μ_4 is the 4th raw moment of F_y.[19] Substituting ζ_1 and ζ_2 into the following expression yields

$$\begin{aligned}
Var(\widehat{\sigma}_y^2) &= \begin{pmatrix} n \\ 2 \end{pmatrix}^{-1} \sum_{i=1}^{2} \begin{pmatrix} 2 \\ i \end{pmatrix} \begin{pmatrix} n-2 \\ 2-i \end{pmatrix} \zeta_i & (33) \\
&= \begin{pmatrix} n \\ 2 \end{pmatrix}^{-1} (2(n-2)\zeta_1 + \zeta_2) & (34) \\
&= \frac{4\zeta_1}{n} + \frac{2\zeta_2}{n(n-1)} - \frac{4\zeta_1}{n(n-1)} & (35) \\
&= \frac{\mu_4 - \sigma_y^4}{n} + \frac{2\sigma_y^4}{n(n-1)} & (36) \\
&= \frac{\mu_4 - \sigma_y^4}{n} + O(n^{-2}). & (37)
\end{aligned}$$

As can be seen from the above, the asymptotic variance of the sample variance is $\frac{\mu_4 - \sigma_y^4}{n}$ as $n \to \infty$.

The above results can be generalized to the case of s U-statistics. The joint limiting distribution of U_i of order m_i, for $i = 1, 2, \ldots, s$, is a multivariate normal distribution. If $F_y(y)$ is continuous and has a finite variance,

[19] The detailed derivation of the above results can be found in Serfling (1980, p. 182, Example A).

which implies $E[\varphi_i(y_1, \ldots, y_{m_i})]^2$ exists, then, as $n \to \infty$, the joint distribution of

$$\left[\sqrt{n}\left(U_1 - \theta_1\right), \sqrt{n}\left(U_2 - \theta_2\right), \ldots, \sqrt{n}\left(U_s - \theta_s\right)\right] \tag{38}$$

converges to the multivariate normal distribution with mean zero and variance-covariance matrix $\left\{m_i m_j \zeta_{ij}\right\}$ with $i, j = 1, 2, \ldots, s$ and

$$\zeta_{ij} = E\left[\varphi_i(y_1, \ldots, y_{m_i}) \cdot \varphi_j(y_1, \ldots, y_{m_i}, y_{m_j+1}, \ldots, y_{2m_j - m_i})\right] - \theta_i \theta_j \tag{39}$$

with $m_i \leq m_j$.[20]

To make sense of the joint limiting distribution with a concrete example, let $s = 2$ and let the two U-statistics be the sample mean and sample absolute

[20] $E\left[\varphi_i(y_1, \ldots, y_{m_i}) \cdot \varphi_i(y_1, \ldots, y_{m_i}, y_{m_j+1}, \ldots, y_{2m_j - m_i})\right]$ is a conditional variance if y_c, $c = 1, 2, \ldots, m_i$, are known. According to Hoeffding (1948, page 304, equations 6.1, 6.2, and 6.3) and Lee (1990, pages 11–12, Theorem 2 and its proof), this can be illustrated by the following. Note that

$$\int \cdots \int \varphi_i(y_1, \ldots, y_{m_i}) \prod_{i=c+1}^{m_i} dF_y(y_i) = \varphi_c(y_1, \ldots, y_c)$$

and

$$\int \cdots \int \varphi_i(y_1, \ldots, y_c, y_{m_j+1}, \ldots, y_{2m_j - c}) \prod_{i=m_j+1}^{2m_j - c} dF_y(y_i) = \varphi_c(y_1, \ldots, y_c).$$

Hence,

$$\begin{aligned}
&E\left[\varphi_i(y_1, \ldots, y_{m_i}) \cdot \varphi_i(y_1, \ldots, y_c, y_{m_j+1}, \ldots, y_{2m_j - c})\right] \\
&= \int \cdots \int \varphi_i(y_1, \ldots, y_{m_i}) \cdot \varphi_i(y_1, \ldots, y_c, y_{m_j+1}, \ldots, y_{2m_j - c}) \textstyle\prod_{i=1}^{2m_j - c} dF_y(y_i) \\
&= \int \cdots \int \left\{\int \cdots \int \varphi_i(y_1, \ldots, y_{m_i}) \textstyle\prod_{i=c+1}^{m_i} dF_y(y_i)\right\} \times \\
&\left\{\int \cdots \int \varphi_i(y_1, \ldots, y_c, y_{m_j+1}, \ldots, y_{2m_j - c}) \textstyle\prod_{i=m_j+1}^{2m_j - c} dF_y(y_i)\right\} \textstyle\prod_{i=1}^{c} dF_y(y_i) \\
&= \int \cdots \int \varphi_c^2(y_1, \ldots, y_c) \textstyle\prod_{i=1}^{c} dF_y(y_i) \\
&= E\left[\varphi_c^2(y_1, \ldots, y_c)\right].
\end{aligned}$$

mean difference:[21]

$$U_1 = \widehat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{40}$$

is the estimator for $\theta_1 = \mu_y$ and

$$U_2 = \widehat{\Delta}_y = \frac{2}{n\,(n-1)}\sum_{i<j} |y_i - y_j| \tag{41}$$

is the estimator for $\theta_2 = \Delta_y$. Here $m_1 = 1$ and $m_2 = 2$; $\varphi_1(y_1) = y_1$ and $\varphi_2(y_1, y_2) = |y_1 - y_2|$. Hence,

$$m_1^2\zeta_{11} = E\left[y_1^2\right] - \theta_1^2 = \zeta(\theta_1) \tag{42}$$

is the variance of $\sqrt{n}\,(U_1 - \theta_1)$,

$$m_2^2\zeta_{22} = 4\left\{E\left[|y_1 - y_2|^2\right] - \theta_2^2\right\} = 4\zeta(\theta_2) \tag{43}$$

is the variance of $\sqrt{n}\,(U_2 - \theta_2)$, and

$$\begin{aligned} m_1 m_2 \zeta_{12} &= 2\left[E\left(y_1|y_1 - y_2|\right) - \theta_1\theta_2\right] \\ &= 2\left[\int\int y_1|y_1 - y_2| dF_y(y_1) dF_y(y_2) - \theta_1\theta_2\right] \\ &= 2\zeta\left(\theta_1, \theta_2\right) \end{aligned} \tag{44}$$

is the covariance between $\sqrt{n}\,(U_1 - \theta_1)$ and $\sqrt{n}\,(U_2 - \theta_2)$.

[21] For simplicity, we freely redefine U_i, $i = 1, 2, 3, \ldots$ from time to time.

The consistent estimators for $\zeta\left(\theta_1\right)$, $\zeta\left(\theta_2\right)$, and $\zeta\left(\theta_1, \theta_2\right)$ are[22]

$$\widehat{\zeta}\left(\theta_1\right) = \frac{1}{(n-1)}\left(\sum_{i=1}^{n} y_i^2 - nU_1^2\right), \tag{45}$$

$$\widehat{\zeta}\left(\theta_2\right) = \frac{2}{n\left(n-1\right)\left(n-2\right)}$$

$$\times \sum_{i<j<k} \left\{|y_i - y_j||y_i - y_k| + |y_j - y_i||y_j - y_k| + |y_k - y_i||y_k - y_j|\right\} - U_2^2, \tag{46}$$

and

$$\widehat{\zeta}\left(\theta_1, \theta_2\right) = \frac{1}{n(n-1)} \sum_{i<j} (y_i + y_j)|y_i - y_j| - U_1 U_2. \tag{47}$$

One of the important applications of the asymptotic distribution of U-statistics is to establish the asymptotic distribution of the sample Gini index. Based on the knowledge that the sample Gini index of incomes y is half of the relative mean difference which is the ratio of the sample absolute mean difference to the sample mean, as pointed by Hoeffding (1948), the sample Gini index, $\widehat{G}_y = \frac{\widehat{\Delta}_y}{2\widehat{\mu}_y}$, has an asymptotic normal distribution. More precisely, as $n \to \infty$, $\sqrt{n}(\widehat{G}_y - \frac{\Delta_y}{2\mu_y})$ converges to a normal distribution with mean 0 and variance:

$$\frac{\Delta_y^2}{4\mu_y^4}\zeta(\mu_y) - \frac{\Delta_y}{\mu_y^3}\zeta(\mu_y, \Delta_y) + \frac{1}{\mu_y^2}\zeta(\Delta_y). \tag{48}$$

where the population parameters for μ_y and Δ_y in the ζ functions are used replace θ_1 and θ_2.[23]

The key application of the asymptotic distribution of U-statistics pre-

[22] See Bishop et al. (1997).

[23] Nygård and Sandström (1981, p. 384) give an estimator of the asymptotic variance of $\widehat{G}_y$ as

$$\widehat{Var}(\widehat{G}_y) =$$

sented in this paper is to establish the asymptotic distributions of the sample Sen and modified Sen indices and their components. This will be more involved. In the Sen index $S = H \cdot \mu_{x_p} \cdot \left[1 - G_{x_p}\right]$ and the modified Sen index $S_m = H \cdot \mu_{x_p} \cdot [1 - G_x]$, the Gini index of poverty gap ratios of the population G_x and that of the poor G_{x_p} differ from the Gini index of incomes G_y. It can be shown that[24]

$$G_x = \frac{1}{2\mu_x z^3} \int_0^{+\infty} \int_0^{+\infty} I(y_1 < z) I(y_2 < z) |y_1 - y_2| dF_y(y_1) dF_y(y_2) \tag{49}$$

and

$$G_{x_p} = \frac{1}{2\mu_{x_p} z^3 [F_y(z)^2]} \int_0^{+\infty} \int_0^{+\infty} I(y_1 < z) I(y_2 < z) |y_1 - y_2| dF_y(y_1) dF_y(y_2). \tag{50}$$

That is, the sample estimators of H and μ_{x_p} are U-statistics and the sample counterparts of G_x and G_{x_p} are functions of U-statistics.

It is now necessary to find the estimators for H, μ_{x_p}, G_{x_p}, G_x, S and S_m and their asymptotic distributions. Given the sample of size n from F_y, $\{y_1, y_2, \ldots, y_n\}$, the U-statistic

$$U_1 = \frac{1}{n} \sum_{i=1}^{n} I(y_i < z) = \widehat{H} \tag{51}$$

$$\sum_{i=1}^{n-1} \sum_{j=1}^{n-1} \left[min\left(\frac{i}{n}, \frac{j}{n}\right) - \left(\frac{i}{n}\right)\left(\frac{j}{n}\right) \right] \left(\frac{2i-1}{n} - 1 - \widehat{G}_y\right) \left(\frac{2j-1}{n} - 1 - \widehat{G}_y\right) (y_{i+1} - y_i)(y_{j+1} - y_j).$$

[24] See Appendix A.

is an estimator for H. The U-statistic

$$U_2 = \frac{1}{n}\sum_{i=1}^{n} y_i I(y_i < z) = \widehat{\mu}_{y<z}. \tag{52}$$

is an estimator for $\mu_{y<z}$. Then the estimator for μ_x is given by

$$\widehat{\mu}_x = U_1\left(1 - \frac{U_2}{zU_1}\right). \tag{53}$$

The estimator for μ_{x_p} is given by

$$\widehat{\mu}_{x_p} = 1 - \frac{U_2}{zU_1}. \tag{54}$$

The sample absolute mean difference for $y < z$ is given by

$$U_3 = \frac{1}{(n(n-1))}\sum_{i=1}^{n}\sum_{j=1}^{n} |y_i - y_j| I(y_i < z) I(y_j < z). \tag{55}$$

According to equation (49), the estimator of G_x is a function of the U-statistics U_1, U_2, and U_3,

$$\widehat{G}_x = \frac{U_3}{2z^3 U_1\left(1 - \frac{U_2}{zU_1}\right)} \tag{56}$$

Based on equation (50), the estimator of G_{x_p} is also a function of the U-statistics U_1, U_2, and U_3,

$$\widehat{G}_{x_p} = \frac{U_3}{2z^3 U_1^2\left(1 - \frac{U_2}{zU_1}\right)}. \tag{57}$$

Since $\widehat{S}$ ($\widehat{S}_m$) is a function of $\widehat{H}$, $\widehat{\mu}_{x_p}$, and $\widehat{G}_{x_p}$ (or $\widehat{G}_x$), combining equations (51), (54) and (50) [or (49)] yields

$$\widehat{S} = \frac{2z^3U_1^2 - 2z^2U_1U_2 - U_3}{2z^3U_1} \tag{58}$$

and

$$\widehat{S}_m = \frac{2z^3U_1 - 2z^2U_2 - U_3}{2z^3}. \tag{59}$$

To understand the consistency of the above estimators, note that they are continuous functions of U-statistics without involving n. Also, assume that these functions have their second order partial derivatives in the neighborhood of the true parameters θ_1, θ_2, and θ_3. Under these conditions, these estimators are consistent and have an asymptotic joint distribution [see Hoeffding (1948), Theorem 7.5]: the U-statistics U_1, U_2, and U_3 are consistent estimators for

$$\theta_1 = \int_0^{+\infty} I(y < z) dF_y(y), \tag{60}$$

$$\theta_2 = \int_0^{+\infty} I(y < z) y dF_y(y), \tag{61}$$

and

$$\theta_3 = \int_0^{+\infty}\int_0^{+\infty} I(y_1 < z) I(y_2 < z) |y_1 - y_2| dF_y(y_1) dF_y(y_2), \tag{62}$$

respectively. If F_y is continuous and has a finite variance, then, as $n \to \infty$, the joint distribution of

$$\sqrt{n}\left(\mathbf{U} - \boldsymbol{\theta}\right) = \left[\sqrt{n}\left(U_1 - \theta_1\right), \sqrt{n}\left(U_2 - \theta_2\right), \sqrt{n}\left(U_3 - \theta_3\right)\right]^\top \tag{63}$$

converges to a multivariate normal distribution function with mean zero and variance-covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \theta_1(1-\theta_1) & \theta_2(1-\theta_1) & 2\theta_3(1-\theta_1) \\ \theta_2(1-\theta_1) & \zeta(\theta_2) & 2\zeta(\theta_2,\theta_3) \\ 2\theta_3(1-\theta_1) & 2\zeta(\theta_2,\theta_3) & 4\zeta(\theta_3) \end{bmatrix} \tag{64}$$

where

$$\zeta(\theta_2) = \int_0^{+\infty} I(y<z)y^2 dF_y(y) - \theta_2^2, \tag{65}$$

$$\zeta(\theta_3) = \int_0^{+\infty} I(y_1<z)\left(\int_0^{+\infty} I(y_2<z)|y_1-y_2|dF_y(y_2)\right)^2 dF_y(y_1) - \theta_3^2, \tag{66}$$

and

$$\zeta(\theta_2,\theta_3) = \int_0^{+\infty}\int_0^{+\infty} I(y_1<z)I(y<z_2)y_1|y_1-y_2|dF_y(y_1)dF_y(y_2) - \theta_2\theta_3, \tag{67}$$

respectively [see Hoeffding's Theorem 7.1 (1948) and Bishop, Formby, and Zheng (1997)].

Given that the estimators of H, μ_{x_p}, G_{x_p}, G_x, S and S_m are functions of U-statistics—U_1, U_2, and U_3—and that $\sqrt{n}\,(\mathbf{U}-\boldsymbol{\theta}) \xrightarrow{a} \mathcal{N}(\mathbf{0},\mathbf{\Sigma})$, we can find the limiting distributions for the following two vectors of the estimators

$$\widehat{\boldsymbol{\alpha}} = [\widehat{H}, \widehat{\mu}_{x_p}, \widehat{G}_{x_p}, \widehat{S}]^\top \tag{68}$$

and

$$\widehat{\boldsymbol{\alpha}}_m = [\widehat{H}, \widehat{\mu}_{x_p}, \widehat{G}_x, \widehat{S}_m]^\top \tag{69}$$

for the Sen index and its components and the modified Sen index and its components, respectively. The functions $h_1(\mathbf{w}) = w_1$, $h_2(\mathbf{w}) = 1 - \frac{w_2}{zw_1}$, $h_3(\mathbf{w}) = \frac{w_3}{2z^3 w_1^2\left(1-\frac{w_2}{zw_1}\right)}$, $h_{m3}(\mathbf{w}) = \frac{w_3}{2z^3 w_1\left(1-\frac{w_2}{zw_1}\right)}$, $h_4(\mathbf{w}) = \frac{2z^3w_1^2-2z^2w_1w_2-w_3}{2z^3w_1}$, and $h_{m4}(\mathbf{w}) = \frac{2z^3w_1-2z^2w_2-w_3}{2z^3}$ can be used to define $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\alpha}}_m$:

$$\widehat{\boldsymbol{\alpha}} = \left[\widehat{H}, \widehat{\mu}_{x_p}, \widehat{G}_{x_p}, \widehat{S}\right]^\top = \mathbf{H}(\mathbf{U}) = [h_1(\mathbf{U}), h_2(\mathbf{U}), h_3(\mathbf{U}), h_4(\mathbf{U})]^\top \tag{70}$$

and

$$\widehat{\boldsymbol{\alpha}}_m = \left[\widehat{H}, \widehat{\mu}_{x_p}, \widehat{G}_x, \widehat{S}_m\right]^\top = \mathbf{H}_m(\mathbf{U}) = [h_1(\mathbf{U}), h_2(\mathbf{U}), h_{m3}(\mathbf{U}), h_{m4}(\mathbf{U})]^\top . \tag{71}$$

Similarly, these functions can be used to define $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_m$:

$$\boldsymbol{\alpha} = \left[H, \mu_{x_p}, G_{x_p}, S\right]^\top = \mathbf{H}(\boldsymbol{\theta}) = [h_1(\boldsymbol{\theta}), h_2(\boldsymbol{\theta}), h_3(\boldsymbol{\theta}), h_4(\boldsymbol{\theta})]^\top \tag{72}$$

and

$$\boldsymbol{\alpha}_m = \left[H, \mu_{x_p}, G_x, S_m\right]^\top = \mathbf{H}_m(\boldsymbol{\theta}) = [h_1(\boldsymbol{\theta}), h_2(\boldsymbol{\theta}), h_{m3}(\boldsymbol{\theta}), h_{m4}(\boldsymbol{\theta})]^\top . \tag{73}$$

Define $\mathbf{T} = \frac{\partial \mathbf{H}}{\partial \mathbf{w}}|_{\mathbf{w}=\boldsymbol{\theta}}$ or

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{\theta_2}{z\theta_1^2} & -\frac{1}{z\theta_1} & 0 \\ \frac{\theta_3(\theta_2-2z\theta_1)}{2z^2\theta_1^2(z\theta_1-\theta_2)^2} & \frac{\theta_3}{2z^2\theta_1(z\theta_1-\theta_2)^2} & \frac{1}{2z^2\theta_1(z\theta_1-\theta_2)} \\ \frac{2z^3\theta_1^2+\theta_3}{2z^3\theta_1^2} & -\frac{1}{z} & -\frac{1}{2z^3\theta_1} \end{bmatrix}. \tag{74}$$

Define $\mathbf{T}_m = \frac{\partial \mathbf{H}_m}{\partial \mathbf{w}}|_{\mathbf{w}=\boldsymbol{\theta}}$ or

$$\mathbf{T}_m = \begin{bmatrix} 1 & 0 & 0 \\ \frac{\theta_2}{z\theta_1^2} & -\frac{1}{z\theta_1} & 0 \\ \frac{-\theta_3}{2z(z\theta_1-\theta_2)^2} & \frac{\theta_3}{2z^2(z\theta_1-\theta_2)^2} & \frac{1}{2z^2(z\theta_1-\theta_2)} \\ 1 & -\frac{1}{z} & -\frac{1}{2z^3} \end{bmatrix}. \tag{75}$$

As $n \to \infty$, the joint distribution of

$$\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \tag{76}$$

converges to a multivariate normal distribution with mean zero and variance-covariance matrix

$$\boldsymbol{\Omega} = \mathbf{T}\boldsymbol{\Sigma}\mathbf{T}^\top. \tag{77}$$

Similarly, as $n \to \infty$, the joint distribution of

$$\sqrt{n}(\widehat{\boldsymbol{\alpha}}_m - \boldsymbol{\alpha}_m) \tag{78}$$

converges to a multivariate normal distribution function with mean zero and variance-covariance matrix

$$\boldsymbol{\Omega}_m = \mathbf{T}_m\boldsymbol{\Sigma}\mathbf{T}_m^\top. \tag{79}$$

$\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_m$ must be estimated. To do so, one must estimate θ_1, θ_2, θ_3, $\zeta(\theta_2)$, $\zeta(\theta_3)$, and $\zeta(\theta_2, \theta_3)$ by U_1, U_2, U_3, $\widehat{\zeta}(\theta_2)$, $\widehat{\zeta}(\theta_3)$, and $\widehat{\zeta}(\theta_2, \theta_3)$. U_1, U_2, and U_3 are given by equations (51), (52), and (55), respectively. $\widehat{\zeta}(\theta_2)$, $\widehat{\zeta}(\theta_3)$, and $\widehat{\zeta}(\theta_2, \theta_3)$ are given, respectively, by:

$$\widehat{\zeta}(\theta_2) = \frac{1}{(n-1)} \left(\sum_{i=1}^{n} y_i^2 I(y_i < z) - nU_2^2 \right), \tag{80}$$

$$\widehat{\zeta}(\theta_3) = \frac{2}{n\,(n-1)\,(n-2)} \sum_{i<j<k} \{|y_i - y_j||y_i - y_k| + |y_j - y_i||y_j - y_k| + |y_k - y_i||y_k - y_j|\}$$
$$\times I(y_i < z) I(y_j < z) I(y_k < z) - U_3^2, \tag{81}$$

and

$$\widehat{\zeta}(\theta_2, \theta_3) = \frac{1}{n\,(n-1)} \sum_{i<j} (y_i + y_j)|y_i - y_j| I(y_i < z) I(y_j < z) - U_2 U_3. \tag{82}$$

This completes the explanation on why the U-statistics can be used to handle the statistical inferential issues for the Sen and modified Sen indices and their components.

4 Concluding Remarks

Considering the fact that U-statistics are not introduced in conventional econometric textbooks, this paper advocates the use of U-statistics for income inequality and poverty measures with special focus on the variance, the absolute/relative mean difference, the Gini index, the poverty rate, the mean poverty gap ratios, the Foster-Greer-Thorbecke (FGT) index, the Sen index, and the modified Sen index.

The framework and general results for the U-statistics illustrated in this paper are useful for establishing statistical procedures for widely-used income inequality and poverty measures. The U-statistics approach for income inequality and poverty measures represents a more attractive alternative to the approach that depends primarily on a limited number of quantiles or order statistics because the U-statistics approach uses all, rather than parts of, the sample information.

Although the variances of some U-statistics may appear to be complex, they are merely complex functions of conditional expectations. The literature on U-statistics also suggests that these quantities can often be estimated by the bootstrap method.

References

[1] Aaberge, Rolf (1982). "On the Problem of Measuring Inequality," (In Norwegian) *Rapp* 82/9, Central Bureau of Statistics, Oslo, Norway.

[2] Anderson, Gordon (2004). "Distributional Overlap, A Simple Multi-Dimensional Measure and Test of Alienation and Social Cohesion: The Case of Single Parent and Pensioner Households in the United Kingdom," Mimeo, University of Toronto, Canada.

[3] Biewen, Martin (2002). "Bootstrap Inference for Inquality, Mobility, and Poverty Measurement," *Journal of Econometrics*, vol. 108, no. 2, pp. 317-342.

[4] Bishop, John A., S. Chakraborti, and Paul D. Thistle (1990) "An Asymptotically Distribution-Free Test for Sen's Welfare Index," *Oxford Bulletin of Economics and Statistics*, vol. 52, no. 1, pp. 105–113.

[5] Bishop, John A., John P. Formby, and Buhong Zheng (1997). "Statistical Inference and the Sen Index of Poverty," *International Economic Review*, vol. 38, no. 2, pp. 381–387.

[6] Bishop, John A., John P. Formby, and Buhong Zheng (1998). "Inference Tests for Gini-Based Tax Progressivity Indexes," *Journal of Business & Economic Statistics,* vol. 16, no. 3, pp. 322–330.

[7] Bishop, John A., John P. Formby, and Buhong Zheng (2001). "Sen Measures of Poverty in the United States: Cash versus Comprehensive Incomes in the 1990s," *Pacific-Economic Review*, vol. 6, no. 2, pp. 193–210.

[8] Chakravarty, Satya R. (1990). *Ethical Social Index Numbers* , Berlin: Springer-Verlag, 1990.

[9] Chakravarty, Satya R. (2001a). "Why Measuring Inequality by the Variance Makes Sense from a Theoretical Point of View," *Journal of Income Distribution*, vol. 10, no. 3-4, pp. 82–96.

[10] Chakravarty, Satya R. (2001b). "The variance as a subgroup decomposable measure of inequality," *Social Indicators Research*, vol. 53, no. 1, pp. 79–95.

[11] Foster, James E. and Efe A. Ok (1999). "Lorenz Dominance and the Variance of Logarithms," *Econometrica*, vol. 67, no. 4, pp. 901–907.

[12] Foster, James, Joel Greer, and Erik Thorbecke (1984), "A Class of Decomposable Poverty Measures," *Econometrica*, vol.52, no.3, pp. 761–766.

[13] Fraser, D. A. S. (1957). *Nonparametric Methods in Statistics*, New York: John Wiley & Sons, Inc.

[14] Gail, M., and J. L. Gastwirth (1978). "A Scale-Free Goodness-of-Fit Test for the Exponential Distribution Based on the Gini Statistics," *Journal of the Royal Statistical Society*, Series B, 40, pp. 350–357.

[15] Gastwirth, Joseph L. (1972). "The Estimation of the Lorenz Curve and the Gini Index," *Review of Economics and Statistics*, vol. 54, no. 3, pp. 306–316.

[16] Gastwirth, Joseph L. (1974). "Large Sample Theory of Income Inequality," *Econometrica*, vol. 42, no. 1, pp. 191–196.

[17] Gastwirth, Joseph L. (1976). Errata for "Large Sample Theory of Income Inequality," *Econometrica* , vol. 44, no. 4, p. 840.

[18] Giles, David E. A. (2004). "Calculating a Standard Error for the Gini Coefficient: Some Further Results," *Oxford Bulletin of Economics and Statistics*, vol. 66, no. 3, pp. 425–433.

[19] Glasser, G. J. (1962), "Variance Formulas for the Mean Difference and Coefficient of Concentration," *Journal of American Statistical Association*, vol. 57, pp. 648–654.

[20] Halmos, Paul R. (1948). "The Theory of Unbiased Estimation," *Annals of Mathematical Statistics*, vol. 17, no. 1, pp. 34–43.

[21] Hoeffding, W. (1948). "A Class of Statistics with Asymptotically Normal Distribution," *Annals of Mathematical Statistics*, vol. 19, no. 3, pp. 293–325.

[22] Karagiannis, Elias and Milorad Kovacivic (2000). "A Method to Calculate the Jackknife Estimator for the Gini Coefficient," *Oxford Bulletin of Economics and Statistics*, vol. 62, no. 1, pp. 119–121.

[23] Lee, A. J. (1990). *U-Statistics: Theory and Practice*, New York: Marcel Dekker, Inc.

[24] Nygård, Fredrik and Arne Sandström (1981). *Measuring Income Inequality*, Stockholm, Sweden: Almqvist & Wiksell International.

[25] Ogwang, Tomson (2000). "A Convenient Method of Computing the Gini Index and Its Standard Error," *Oxford Bulletin of Economics and Statistics*, vol. 62, no. 1, pp. 123–129.

[26] Ogwang, Tomson (2004). "Calculating a Standard Error for the Gini Coefficient: Some Further Results: Reply," *Oxford Bulletin of Economics and Statistics*, vol. 66, no. 3, pp. 435–437.

[27] Osberg, Lars (2000). "Poverty in Canada and the United States: Measurement, Trends, and Implications," Presidential Address, *Canadian Journal of Economics*, vol. 33, no. 4, pp. 847–877.

[28] Osberg, Lars and Kuan Xu (1999). "Poverty Intensity — How Well Do Canadian Provinces Compare?" *Canadian Public Policy*, vol. 25, no. 2, pp. 1–17.

[29] Osberg, Lars and Kuan Xu (2000). "International Comparison of Poverty Intensity: Index Decomposition and Bootstrap Inference," *Journal of Human Resources*, vol. 35, no. 1, pp. 51–81; errata noted in *Journal of Human Resources*, vol. 35, no. 3.

[30] Randles, Ronald H. and Douglas A. Wolfe (1979). *Introduction to The Theory of Nonparametric Statistics*, New York: John Wiley & Sons.

[31] Sandstrom, A., J. H. Wretman, and B. Walden (1988). "Variance Estimators of the Gini Coefficient—Probability Sampling," *Journal of Business and Economic Statistics*, vol. 6, no. 1, pp. 113–119.

[32] Sen, Amartya (1973). *On Economic Inequality*, Oxford: Clarendon.

[33] Sen, Amartya (1976). "Poverty: An Ordinal Approach to Measurement," *Econometrica*, vol. 44, no.2, pp. 219–231.

[34] Sendler, W. (1979). "On Statistical Inference in Concentration Measurement," *Metrika*, vol. 26. no. 1, pp. 109–122.

[35] Serfiling, Robert J. (1980). *Approximation Theorems of Mathematical Statistics*, New York: John Wiley & Sons.

[36] Xu, Kuan (1998). "The Statistical Inference for the Sen-Shorrocks-Thon index of Poverty Intensity," *Journal of Income Distribution*, vol. 8, no. 1, pp. 143–152.

[37] Xu, Kuan (2003). "How Has the Literature on the Gini Index Evolved in the Past 80 Years?," *China Economic Quarterly*, vol. 2, no. 4, pp. 757-778.

[38] Xu, Kuan and Lars Osberg (2002). "The Social Welfare Implications, Decomposability, and Geometry of the Sen Family of Poverty Indices," *Canadian Journal of Economics*, vol. 35, no. 1, pp. 138-152.

[39] Yitzhaki, Shlomo (1998). "More Than A Dozen Alternative Ways of Spelling Gini," *Research in Economic Inequality*, vol. 8, pp. 13–30.

[40] Yitzhaki, Shlomo (1991). "Calculating Jackknife Variance Estimators for Parameters of the Gini Method," *Journal of Business and Economic Statistics*, vol. 9, pp. 235-239.

[41] Zheng, Buhong, John P. Formby, W. James Smith, Victor K. Chow (2000). "Inequality Orderings, Normalized Stochastic Dominance, and

Statistical Inference," *Journal of Business and Economic Statistics* , vol. 18, no. 4, pp. 479–488.

Appendix A. Definitions of G_x and G_{x_p}

From the statistical point of view, the Gini index of income inequality can also be defined as half of the relative mean difference

$$G_y = \frac{1}{2\mu_y} \int_0^{+\infty} \int_0^{+\infty} |y_1 - y_2| dF_y(y_1) dF_y(y_2) \tag{83}$$

or

$$G_y = \frac{1}{2\mu_y} \int_0^{+\infty} \int_0^{+\infty} |y_1 - y_2| f_y(y_1) f_y(y_2) dy_1 dy_2 \tag{84}$$

where y_1 and y_2 are two variates from the same distribution function F_y. The Gini index of the poverty gap ratios of the population G_x and that of the poor G_{x_p} differ from the Gini index of incomes G_y. The poverty gap ratio is a function of income; that is, $x = g(y) = \frac{z-y}{z}$ for $y < z$ and $x = 0$ for $y \geq z$. The Gini index of poverty gap ratios of the population should be defined on the probability density function of x, f_x while that of the poor should be defined on the probability density function of x, $f_{x|x>0} = \frac{f_x}{F_y(z)}$. The support for f_y is $[0, +\infty)$ and that for f_x and $f_{x|x>0}$ is $[0, z)$. To find f_x and $f_{x|x>0}$, note that $x = g(y) = \max\{0, \frac{z-y}{z}\}$ and $\overset{*}{y} = g^{-1}(x) = z(1-x)$. Thus, $f_x(x) = f_y(y)|\frac{\partial x}{\partial y}| = \frac{1}{z} f_y(y)$ and $f_{x|x>0}(x) = \frac{1}{zF_y(z)} f_y(y)$ for $y \in [0, z]$, which corresponds with $x \in [1, 0]$. G_x and G_{x_p} can be defined as

$$G_x = \frac{1}{2\mu_x} \int_0^1 \int_0^1 |x_1 - x_2| f_x(x_1) f_x(x_2) dx_1 dx_2 \tag{85}$$

and

$$G_{x_p} = \frac{1}{2\mu_{x_p}} \int_0^1 \int_0^1 |x_1 - x_2| f_{x|x>0}(x_1) f_{x|x>0}(x_2) dx_1 dx_2, \tag{86}$$

respectively. Substituting $x = g(y) = \max\{0, \frac{z-y}{z}\}$, $f_x(x) = \frac{1}{z} f_y(y)$, and

$f_{x|x>0}(x) = \frac{1}{zF_y(z)} f_y(y)$ into the above expressions and changing the limits yields

$$G_x = \frac{1}{2\mu_x} \int_0^z \int_0^z \frac{1}{z}|y_1 - y_2| \frac{1}{z} f_y(y_1) \frac{1}{z} f_y(y_2) dy_1 dy_2, \tag{87}$$

and

$$G_{x_p} = \frac{1}{2\mu_{x_p}} \int_0^z \int_0^z \frac{1}{z}|y_1 - y_2| \frac{1}{zF_y(z)} f_y(y_1) \frac{1}{zF_y(z)} f_y(y_2) dy_1 dy_2, \tag{88}$$

respectively. After some rearrangements, these expressions become

$$G_x = \frac{1}{2\mu_x z^3} \int_0^{+\infty} \int_0^{+\infty} I(y_1 < z) I(y_2 < z) |y_1 - y_2| dF_y(y_1) dF_y(y_2) \tag{89}$$

and

$$G_{x_p} = \frac{1}{2\mu_{x_p} z^3 [F_y(z)^2]} \int_0^{+\infty} \int_0^{+\infty} I(y_1 < z) I(y_2 < z) |y_1 - y_2| dF_y(y_1) dF_y(y_2), \tag{90}$$

respectively. The above can be verified using simple numerical examples.[25]

---

[25] The two different ways of computing G_x and G_{x_p} can be illustrated by using the following data. Let $\mathbf{y} = [1/2, 3/2, 2, 4]'$ and $z = 2$. Then, $\mu_y = 2$, $\overset{*}{\mathbf{y}} = [1/2, 3/2, 2, 2]'$, $\mathbf{x} = [3/4, 1/4, 0, 0]'$and $\mathbf{x}_p = [3/4, 1/4]'$. From the above, $\mu_x = 1/4$ and $\mu_{x_p} = 1/2$. Using the data set, the two approaches,

$$G_x = \frac{1}{2(\frac{1}{4})4^2} \left(\frac{2}{4} + \frac{3}{4} + \frac{3}{4} + \frac{2}{4} + \frac{1}{4} + \frac{1}{4} + \frac{3}{4} + \frac{1}{4} + \frac{3}{4} + \frac{1}{4} \right) = \frac{5}{8}$$

and

$$G_x = \frac{1}{2(\frac{1}{4})2^3(\frac{1}{2})^2 4^2} \left(\frac{2}{2} + \frac{3}{2} + \frac{3}{2} + \frac{2}{2} + \frac{1}{2} + \frac{1}{2} + \frac{3}{2} + \frac{1}{2} + \frac{3}{2} + \frac{1}{2} \right) = \frac{5}{8}$$

give the same answer. Similarly, using the data set, the two approaches,

$$G_{x_p} = \frac{1}{2(\frac{1}{2})2^2} \left(\frac{2}{4} + \frac{2}{4} \right) = \frac{1}{4}$$

and

$$G_{x_p} = \frac{1}{2(\frac{1}{2})2^3(\frac{1}{2})^2 2^2}\,(1+1) = \frac{1}{4}$$

give the same answer.