SCB STATISTISKA CENTRALBYRÅN

1985-05-20

PROMEMORIOR FRÅN P/STM

NR 14

INCOME INEQUALITY MEASURES BASED ON SAMPLE SURVEYS

AV FREDRIK NYGÅRD OCH ARNE SANDSTRÖM

INLEDNING

TILL

Promemorior från P/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1978-1986. – Nr 1-24.

Efterföljare:

Promemorior från U/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1986. – Nr 25-28.

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

Promemorior från P/STM 1985:14. Income inequality measures based on sample surveys / Fredrik Nygård; Arne Sandström.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-PM-PSTM-1985-14

SCB STATISTICS SWEDEN

Statistical Research Unit 1985-05-20

INCOME INEQUALITY MEASURES BASED ON SAMPLE SURVEYS

Fredrik Nygård[1] Arne Sandström[2]

Invited Paper, 1985 Centennial Session of the International Statistical Institute, Amsterdam.

[1] Statistiska Institutionen vid Åbo Akademi, Fänriksgatan 3,
SF-20500 ÅBO, Finland

[2] Statistiska centralbyrån, P/STM
115 81 STOCKHOLM, Sverige

CONTENTS

INCOME INEQUALITY MEASURES BASED ON SAMPLE SURVEYS

Fredrik Nygård
Department of Statistics
Swedish University of Turku
SF-20500 Turku 50, Finland

Arne Sandström
Statistical Research Unit
Statistics Sweden
S -115 81 Stockholm, Sweden

1. INTRODUCTION

The interest in the size distribution of income, its shape in different countries, and evolution over time has increased substantially during the last decades. One reason is that many political steps are taken to promote equality between individuals and/or households - steps which typically involve redistribution of incomes by means of taxes and transfers.

As a consequence the concept 'income inequality' has entered the public consciousness, and the question of how to assess its magnitude properly has been much at issue in leading journals in economics and statistics.

In rough outline, the research devoted to make the concept 'income inequality' operational has adhered to one of two optional approaches. According to one approach, a measure of income inequality should be derived from a well-defined social welfare function, cf. e.g. Dalton (1920), Atkinson (1970), and Röthschild and Stiglitz (1970), whereas the other approach is to determine the properties an appropriate measure should possess and then derive its mathematical form, cf. e.g. Kolm (1976), Cowell (1980), and Shorrocks (1980, 1983).

However, much less work has been done with respect to the sampling properties of inequality measures. In the vast majority of papers on this subject, income inequality is firmly regarded as a population characteristic (parameter), ignoring the fact that this parameter frequently has to be estimated from sampled income data. Even if occasional attention has been given this topic - Mendershausen (1939) being one of the first - it was not until the 1970s it was brought to light more energetically.

The aim of this paper is to point out that income inequality measures can be estimated from sample surveys. In doing this, we shall merely consider one family of inequality parameters, and especially one of its members, and see how it can be estimated by some different approaches. The family under consideration is the Gini family, including the most well-known income inequality measure, viz. the Gini coefficient.

In Section 2 we define the Gini family of inequality parameters.A brief review of its large-sample properties is given in Section 3. Section 4 contains a discussion of three approaches in estimating the finite population parameters, viz. the fixed population, the model, and the auxiliary model approaches.Various variance estimators are discussed in Section 5 and some empirical results are given in Section 6. Decompositions of the members of the Gini family, by subgroups and income sources, are briefly pointed at in Section 7.

2. THE GINI FAMILY

The discussion in this paper will be confined to the Gini family of inequality measures. Before giving a formal definition of this family, we shall shortly review three related topics, viz. the Lorenz curve, Gini's mean difference, and the Gini coefficient.

"Plot along one axis cumulated per cents. of the population from poorest to richest, and along the other the per cent. of the total wealth held by these per cents. of the population." In these words Lorenz (1905) introduced a graphical method for displaying income data, today widely known as the Lorenz curve (LC). A main property of the LC is that, under quite reasonable assumptions (cf. Atkinson (1970)) it may be interpreted in terms of income inequality, viz. the closer the LC is to the diagonal in the Lorenz diagram, the lesser the inequality in the distribution. As a summary measure of this crucial distance we may use the area between the diagram and the LC, usually referred to as the Lorenz area (LA).

Closely related to the LA is the dispersion measure suggested by Gini (1912), viz. $G_N = N^{-2}\Sigma_i\Sigma_j|y_i-y_j|$, where $y_1,\ldots,y_N$ denotes the incomes in a finite population. Gini (1914) observed that G_N, Gini's mean difference, is related to the LA through $G_N=4\bar{y}_N LA$, where $\bar{y}_N=N^{-1}\Sigma_i\ y_i$, and proposed as a measure of concentration the ratio, R_N, between the LA and the largest possible LA, i.e. $R_N=2LA=G_N/(2\bar{y}_N)$. The Gini family of income inequality measures is basically a generalization of the parameter R_N, the Gini coefficient.

To give a formal representation of the ideas by use of the Lebesgue-Stiltjes integral, let $F(y)$ denote the distribution function (df) of a variate Y with the finite mean $\mu= \int_{-\infty}^{\infty} ydF(y) \neq 0$. Its first moment distribution, defined as $F_1(y)=$ $\mu^{-1} \int_{-\infty}^{y} tdF(t)$, represents the ordinate of the LC when plotted with the population shares $p=F(y)$ as abscissa. Using the inverse of the df, defined as $F^{-1}(p)=\inf_y\{y|F(y)\geq p\}$, $0<p\leq 1$, and $F^{-1}(0)= \inf_y\{y|F(y)>0\}$, the LC may be given the single equation representation

$$L(p) = \mu^{-1} \int_0^p F^{-1}(t)dt.$$

Thus, the Lorenz area is given by LA = $\int_0^1 (p-L(p))dp$, and more generally we may define weighted Lorenz areas (WLA) as

$$WLA = \int_0^1 W(p)(p-L(p))dp,$$

where W(p) is a suitably chosen 'weight function'. The Gini coefficient, R, is now obtained as a special case of the WLA, by putting W(p)=2, Further, under mild regularity conditions, see Nygård and Sandström (1981,p.206), the WLA may by rewritten as WLA = $\mu^{-1}\int_0^1 J(p)F^{-1}(p)dp$, where the function J(p) is derived from W(p) through $J(p)=U(p) - \int_0^1 U(p)dp$ with $U(p)=\int_0^p W(t)dt$. We will adopt this reformulation as the formal definition of the family of inequality parameters.

DEFINITION 2.1 The <u>Gini family</u> of inequality parameters is defined as

$$I(F) = T_J(F)/T_\mu(F), \tag{2.1}$$

where $T_J(F) = \int_0^1 J(p)F^{-1}(p)dp$, $T_\mu(F)=\mu= \int_0^1 F^{-1}(p)dp$, and J(p) is a smooth function.

The traditional Gini coefficient is obtained from (2.1) by selecting J(p) = 2p-1, i.e. $R=\mu^{-1} \int_0^1 (2p-1)F^{-1}(p)dp$. By (2.1), the finite population Gini coefficient equals $R_N=(2/N^2\bar{y}_N)\ \Sigma_i\ iy_{i:N} - 1 - N^{-1}$, where $y_{1:N}\leqslant y_{2:N}\leqslant\dots\leqslant y_{N:N}$ denote ranked incomes. Since $\int_0^1(2p-1)F^{-1}(p)dp$ may be rewritten as G/2, where G = $\int_0^1\int_0^1|F^{-1}(p)-F^{-1}(q)|dpdq$ denotes Gini's mean difference in the general case, the relation $R=G/(2\mu)$ is obvious.

This relation between R and G, and the fact that the numerator $T_J(F)$ in (2.1) is a parameter corresponding to an L-statistic, have both influenced the development of the large-sample results of the statistic corresponding to the parameter R.

3. A REVIEW OF SAMPLING PROPERTIES

The early discussion by Lorenz and Gini was in terms of non-negative quantities, and it was not until the work by Wold (1935) that the LC and the Gini coefficient were defined for quantities taking on values on the whole real axis.

Let $Y_1,Y_2,\dots,Y_n$ be independent and identically distributed (iid) as the random variable Y with df F(y) and the Lorenz curve L(p). Wold (op.cit.) showed that the sample LC, $L_n(p)$, converges uniformly to L(p) as $n\to\infty$ and that the sample LA converges to the LA based on F. Let $G_n = n^{-2}\Sigma_i\Sigma_j|y_i-y_j|$ be Gini's mean difference, based on a random sample of size n. The sample variance, $Var(G_n)$, was first given by Nair (1936) and later corrected and proved in a simpler way by Lomnicki (1952).

Glasser (1962) gave an alternative expression of Var(G_n) and used a first-order Taylor approximation to estimate both Var(G_n) and Var(R_n), where R_n is the sample Gini coefficient. Using another method of estimation, based on expressing the variable as a series of polynomials in F, Sillitto (1969) was able to estimate Var(G_n) from a sample.

Hoeffding (1948) showed that G_n belongs to the class of U-statistics and hence is asymptotically normal. It was also shown that if $y \in [0,\infty[$, then R_n is asymptotically normal as well.

To obtain the large-sample properties of $R_n = G_n/(2\bar{Y}_n)$, where $\bar{Y}_n$ is the sample mean, one usually uses the first-order Taylor approximation of R_n about G and μ. The resulting approximation is

$$(R_n - R) \approx \mu^{-1}\{\tfrac{1}{2} G_n - R\bar{Y}_n\}, \qquad (3.1)$$

where the right-hand side of (3.1) is a linear combination of two asymptotically normally distributed statistics.

G can also be written as $T_J(F)$ in (2.1) with $J(p) = 2(2p-1)$. Let the empirical df $F_n(y)$ be defined as $n^{-1}\Sigma_i I\{Y_i \leq y\}$, where $I\{\cdot\}$ is an indicator function. Changing F for F_n in $T_J(F)$, defined by (2.1), we obtain

$$T_J(F_n) = \int_0^1 J(p)F_n^{-1}(p)\, dp. \qquad (3.2)$$

In the case of the Gini coefficient, (2.1) and (3.2) imply that $R_n = (2/n^2\bar{y}_n) \times \Sigma_i\, i y_{i:n} - 1 - n^{-1}$, where $y_{1:n} \leq y_{2:n} \leq \ldots \leq y_{n:n}$. In a similar way we have $T_\mu(F_n) = \bar{y}_n = n^{-1}\Sigma_i y_i$.

Jung (1955) showed that if J is bounded and has at least four bounded derivatives, then $E(T_J(F_n)) = T_J(F) + O(n^{-1})$ and $n\mathrm{Var}(T_J(F)) = \sigma_1^2 + O(n^{-2})$, where

$$\sigma_1^2 = \int_0^1\int_0^1 \{\min(p,q) - pq\}\, J(p)J(q)dF^{-1}(p)dF^{-1}(q). \qquad (3.3)$$

If $J(\cdot)$ is continuous in $[0,1]$ then the theorem on uniform convergence of the Bernstein polynomial, see e.g. Feller (1966,p.221), may be applied to show that $E(T_J(F_n)) \rightarrow T_J(F)$ uniformly in probability.

The asymptotic normality of the L-estimate (3.2) has been proved by several authors with various methods and restrictions on the J-function, see e.g. Serfling (1980) and David (1981) for reviews: A first proof was given by Chernhoff, Gastwirth and Johns (1967) using a transformation and characteristic functions. Moore (1968) Taylor-expanded J to obtain normality, and Stigler (1974) used Hájek's projection lemma. Shorack (1972) used an invariance principle for the empirical process to prove normality, and Sendler (1979) used Shorack's approach and the Taylor approximation (3.1) to obtain asymptotic normality for R_n, i.e.

$$\sqrt{n}(R_n - R) \xrightarrow{\mathcal{L}} U \sim N(0, \mu^{-2}\sigma_2^2), \qquad (3.4)$$

where σ_2^2 corresponds to (3.3) with the J-function changed for $J_1(p) = \{J(p)-R\}$ in accordance with the approximation (3.1). A strongly consistent variance estimator, see Sendler (1979), is obtained if we write $\mu^{-2}\sigma_2^2 = T_\mu(F)^{-2}T_V(F)$, i.e. as a product of two statistical functionals, and change F for F_n. Boos (1979) proved the asymptotic normality of $T_J(F_n)$ by use of a stochastic Frechet differential and Yang (1977, 1981) extended the results of Stigler. Yang proved that, if h(y,x) is a real function of y and x and is of bounded variation, then $T(F_n)=n^{-1}\Sigma_i\ J(i/n)h(y_{i:n}, x_{[i:n]})$ is asymptotically normal, where $x_{[i:n]}$ is the concomitant to the order statistic $y_{i:n}$. This result is useful, together with the approximation (3.1), when considering decomposed Gini coefficients (ref. in Section 7).Goldie (1977) discussed the convergence of the empirical LC and the asymptotic normality of $\sqrt{n}(R_n-R)$. Beach and Davidson (1983) derived the full (asymptotic) variance-covariance structure of points on the empirical LC, based on simple random sampling (srs).

Glasser (1962) used a simple random sample of the size n=15 from 163 residential properties to illustrate the Taylor approximation approach in estimating $Var(G_n)$ and $Var(R_n)$. As income surveys are usually based on sample surveys that are more complex than srs, Love and Wolfson (1976) compared Glasser's approach to a balanced repeated replication (brr) approach on Canadian income data - a survey using a multistage, stratified, cluster design - and estimated $Var(R_n)$. The design effects, as measured by $\{\hat{Var}_{brr}(R_n)/\hat{Var}_{srs}(R_n)\}^{1/2}$, were between 1.0 to 1.8 in various subgroups. Nygård (1981) used Finnish income data and constructed a finite population from which a Monte Carlo study was conducted, cf. also Nygård and Sandström (1981). One thousand replicates of a srs design, with sample sizes n=500 and n=1 000, were taken and the following results were obtained for the Gini coefficient (=0.3258):

Sampling distribution of R_n (1 000 replicates)

n	Mean	Min	Max	Variance
500	0.3252	0.2856	0.4004	213×10^{-6}
1 000	0.3259	0.3010	0.3686	112×10^{-6}

The Gini coefficient can algebraically be reformulated in several ways, see Nygård and Sandström (1981). Two such expressions were used independently by Brewer (1981) and Sandström (1982) to give explicit expressions of R_n based on probability samples. Brewer used a jackknife procedure to estimate $Var(R_n)$. To take the sample design into account, Sandström (1983) used an auxiliary model approach in estimating the finite population Gini coefficient and decomposed

Gini coefficients, and derived explicit variance expression. Nygård and Sandström (1985) gave estimates of Var(R_n) both under a fixed population approach and the auxiliary model approach. These variance estimates were compared in a Monte Carlo study, under the srs design, both with and without replacement, in Sandström, Wretman, and Waldén (1985).

4. SAMPLE SURVEYS

Surveys on size distribution of incomes are usually based on samples, i.e. we select a part of a finite population of units on which to base statements about a universe. The universe may either be (i) a fixed and finite population or(ii) an infinite population of which the finite population is a random part. If the purpose is to make inference about a finite population parameter then the first approach is the common one, but if the purpose is to estimate the inequality parameter in some underlying process that generates the inequality inherent in the finite population then the second approach is applicable. A third approach is to use the infinite population as an auxiliary model in making inference about a finite population (universe) parameter. In Figure 4.1 the three approaches are illustrated.

Assume a finite and identifiable population of size N. We uniquely label the population units from 1 to N and assume that the label of each unit is known, which implies that we can define a label set U={1,2,...,N} of the population. With the jth unit, $j \in U$, we associate some number, say y_j, which can be seen as a result of measuring unit j. In the fixed and finite population approach these numbers constitute a vector $\underset{\sim}{y}_N=(y_1,y_2,\ldots,y_N)$. In the infinite population (model) approach and the auxiliary model approach this vector is considered as a random outcome of a stochastic vector $\underset{\sim}{Y}_N=(Y_1,Y_2,\ldots,Y_N)$, where the Y_i 's are, for example, assumed as independent and identically distributed as Y with a continuous df F(y).

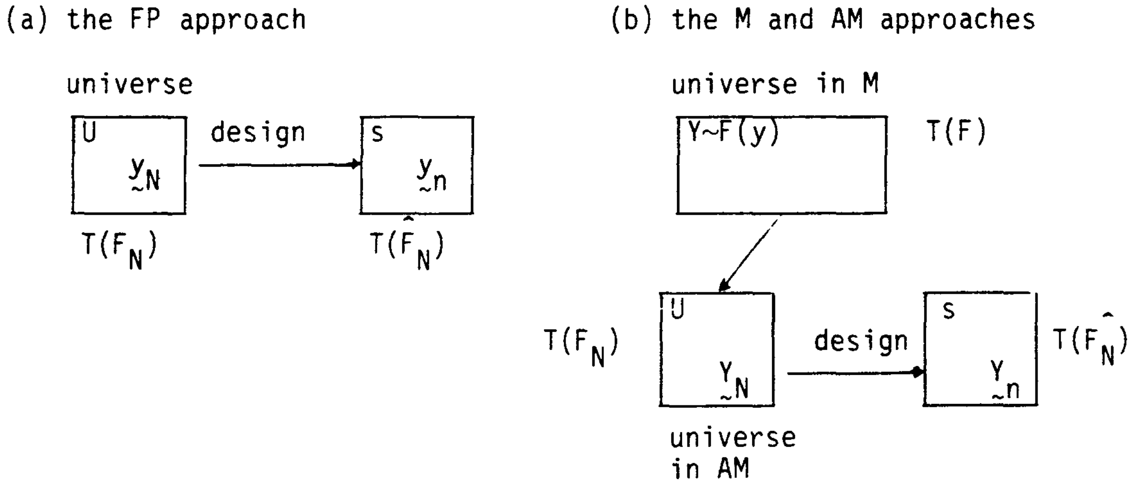
A sample s, of the fixed size n, is a subset of U, i.e. $s=\{j_i|j_i \in U, i=1,2,\ldots,n\}$, and a sampling experiment will yield a sample $s \subset U$ according to a probability distribution P(s), where P(s) denotes the probability with which s is choosen. $\{P(s), s \subset U\}$ is called the sampling design. The stochastic element of the fixed and finite population (FP) approach is the randomization of the sample s. In the model (M) and the auxiliary model (AM) approaches, s is assumed to be fixed, and the stochastic element in these approaches is the randomization of the finite population vector $\underset{\sim}{Y}_N$.

Below, let T(·) denote a statistical functional. In the FP approach, $T(F_N)$ is the parameter under study and $T(\hat{F}_N)$ is a stochastic variable, based on an estimate $\hat{F}_N$ of F_N, while in both the M and the AM approaches T(F) is a parameter and $T(F_N)$ and $T(\hat{F}_N)$ are stochastic variables. It is notable that in the AM approach we are interested in $T(F_N)$, a stochastic variable, which means that any confidence statements about $T(F_N)$ based on $T(\hat{F}_N)$ is of Royall-type, cf. Royall (1971), i.e. for a given sample s the probability of coverage means the probability that the interval includes the stochastic variable $T(F_N)$ when the generating of Y-values from the auxiliary model is "repeated".

The finite population df $F_N(y)$ is by the FP approach defined as $F_N(y) = N^{-1} \Sigma_U I\{y_i<y\}$ and by the M and AM approaches as $F_N(y) = N^{-1}\Sigma_U I\{Y_i<y\}$, where $I\{\cdot\}$ is an indicator function. For a fixed y, $I\{\cdot\}$ is constant in the FP but a stochastic variable in the M and AM, and if $Y_1, Y_2,\ldots,Y_n$ are iid so are $I\{Y_1<y\}, I\{Y_2<y\},\ldots, I\{Y_n<y\}$. The inverse $F_N^{-1}(p)$ is defined analogously, cf. Section 2.

The Gini family of income inequality measures in the finite population is defined as $I(F_N) = T_J(F_N)/T_\mu(F_N)$, where $T_J(F_N) = \int_0^1 J(p)F_N^{-1}(p)dp$, $T_\mu(F_N) = \int_0^1 F_N^{-1}(p)dp$, and F_N is defined as above depending on the approach. As an example, in the FP we have $T_\mu(F_N) = \bar{y}_N = N^{-1}\Sigma_U y_i$ and in the M and AM, $T_\mu(F_N) = \bar{Y}_N = = N^{-1}\Sigma_U Y_i$.

Figure 4.1 Illustrations of the three approaches in making inference about an unknown inequality parameter based on a selected part of a finite population (FP = finite population, M = model, AM = auxiliary model).


(a) the FP approach

(b) the M and AM approaches

Before we give explicit point estimates of $I(F_N)$, we shall introduce some useful notations and define two estimates of F_N. Let $I\{i\in s\}$ be the indicator function with the expectation $\pi_i = P(i\in s)$, the inclusion probability of first order of unit i.

The following definition gives a Håjek estimator of the FP df F_N.

DEFINITION 4.1 A Håjek estimator of the FP df $F_N(y)$ is

$$\hat{F}_N(y) = \hat{N}^{-1}\Sigma_s I\{y_i<y\}/\pi_i, \ \forall y, \qquad (4.1)$$

where $\hat{N} = \Sigma_s \pi_i^{-1}$.

The stochastic element in (4.1) is s, and the two sums may be rewritten as $\Sigma_s I\{y_i<y\}/\pi_i = \Sigma_U I\{i\in s\} I\{y_i<y\}/\pi_i$ and $\Sigma_s \pi_i^{-1} = \Sigma_U I\{i\in s\}/\pi_i$, respectively. The expectation of the numerator and the denominator are $\Sigma_U I\{y_i<y\}$ and N, respectively. Although both parts are unbiased, the ratio (4.1) is generally biased. In simple random sampling $\hat{N}$ = N and hence (4.1) is unbiased in this case.

Under the M approach, where s is assumed fixed, the design is usually ignored. It has been shown, see e.g. Hoem and Funck-Jensen (1982), that if the design is non-informative then the outcomes of the sample may be regarded as iid and hence $\hat{F}_N(y) = n^{-1}\Sigma_s I\{Y_i<y\}, \forall y$. On the other hand, in the AM approach we will take account for the design to have estimates at the same level as in the FP approach. To do this we consider a sequence of populations $U_t=\{1,2,..,N_t\}$ such that $N_t\to\infty$ as $t\to\infty$. For a fixed t we denote the sample of size n_t by s_t and assume that $n_t\to\infty$ as $t\to\infty$ and that the sampling fraction $f_t=n_t/N_t\to f$, $0<f<1$. When t increases we get new subsets of U_t such that s_t is not necessarily a subset

Table 4.1 Point estimates of the Gini coefficient, R.

Approach	Point estimates
FP and AM	$\hat{R}_N = \dfrac{\Sigma_s(2P_i + \pi_i^{-1})y_i/\pi_i}{\hat{N}\Sigma_s y_i/\pi_i} -1$ $P_i = \sum_{j\in s} I\{y_j < y_i\}/\pi_j$ $\hat{N} = \sum_{j\in s} \pi_j^{-1}$
M	$R_n = \dfrac{\Sigma_i(2Q_i + 1)y_i}{n^2\bar{y}_n} - 1$ $Q_i = \Sigma_i I\{y_j < y_i\}$ $\bar{y}_n = n^{-1}\Sigma_i y_i$ if $y_1<y_2<...<y_n$ then $Q_i = i - 1$

Note: If the sampling design is srs the $\hat{R}_N$ in the FP and AM approaches is identical with R_n in the M approach.

of s_{t+1}. The first order inclusion probability is, in a similar way, denoted by π_{it}. The following definition gives an estimator of the AM df $F_N(y)$, corresponding to (4.1), cf. Koul (1970) and Sandström (1983).

DEFINITION 4.2 Let $w_{it}>0$ be bounded $(\forall t)$ deterministic weights, $i\in U_t$, and $\bar{w}_t = n_t^{-1}\Sigma_s w_{it} \neq 0$. A weighted empirical df is defined by

$$\hat{F}_{N_t}(y) = n_t^{-1}\Sigma_s \frac{w_{it}}{\bar{w}_t} I\{Y_i<y\},\ \forall y, \tag{4.2}$$

where $Y_1,Y_2,..,Y_{n_t}$ are iid as Y with continuous df $F(y)$ and $I\{Y_i<y\}$ is an iid indicator function.

If the weights equal some positive constant, then $\hat{F}_{N_t}(y)$ coincides with the 'ordinary' empirical df, and if $w_{it} = \pi_{it}^{-1}$ then (4.2) is similar to (4.1), the only difference being that in (4.2) s_t is fixed and Y_i is stochastic, while the reverse relation is found in (4.1). The only assumption we make on the weights is that $\max_s(w_{it}/\bar{w}_t)^2<d^2<\infty$, $\forall t$. If $w_{it} = \pi_{it}^{-1}$, then this assumption mainly states that the design may not be such that $\min_s\pi_{it}\to 0$ as $t\to\infty$, see Sandström (1983). In Table 4.1 point estimates I (F_N) of the Gini coefficient are given.

Under the M approach the asymptotic normality of $\sqrt{n}(R_n-R)$ is given by (3.4). With the consistent variance estimator of Sendler (1979) we also have a base for large sample estimation of R_N under the FP approach when the design is srs. This result is easily generalized to probability samples for the members of the Gini family, as is done for the Gini coefficient in Sandström (1983), i.e.

$$\frac{n_t^{1/2}\{I(\hat{F}_{N_t}) - I(F)\}}{\{1 + v_t^2\}^{1/2}} \xrightarrow{\mathcal{L}} U \sim N(0,\mu^{-2}\sigma_2^2)\ , \tag{4.3}$$

where $v_t^2 = s_w^2/\bar{w}_t^2$, $s_w^2 = n_t^{-1}\Sigma_s(w_{it} - \bar{w}_t)^2$ and w_{it} is given by Definition 4.2. Especially when we have probability samples, then $w_{it} = \pi_{it}^{-1}$, i.e. $\bar{w}_t = \hat{N}/n_t$ and $s_w^2 = n_t^{-1}\Sigma_s(\pi_{it}^{-1} - \hat{N}/n_t)^2$. The variance σ_2^2 in (4.3) equals (3.3) with $J(p)$ changed for $J_1(p) = \{J(p) - I(F)\}$. A consistent variance estimator to $\mu^{-2}\sigma_2^2$, based on the estimate $\hat{F}_{N_t}(y)$ in (4.2), is also given in the same paper.

In the AM approach, Sandström (1983) has shown that $\sqrt{n}_t(\hat{R}_{N_t}-R_{N_t})/(1-f_t+v_t^2)^{1/2}$ is asymptotically normal. This result is also easily generalized to all members of the Gini family, i.e.

$$\frac{n_t^{1/2}\{I(\hat{F}_{N_t})-I(F_{N_t})\}}{\{1 - f_t + v_t^2\}^{1/2}} \xrightarrow{\mathcal{L}} U \sim N(0,\mu^{-2}\sigma_2^2), \qquad (4.4)$$

where σ_2^2 is as in (4.3). Note that if the design is srs then $v_t^2 = 0$.

5. VARIANCE ESTIMATORS

According to (4.4) the asymptotic variance of the AM estimator $I(\hat{F}_{N_t})$, based on the weighted empirical df, is given by

$$\sigma_{AM}^2 = n_t^{-1}(1-f_t + v_t^2)\ \mu^{-2}\sigma_2^2 \qquad (5.1)$$

In the M case, in line with (3.4), the corresponding variance equals

$$\sigma_M^2 = n^{-1}\ \mu^{-1}\sigma_2^2 \qquad (5.2)$$

Thus, in the M and AM approaches, the precision of the estimates depends crucially on the magnitude of

$$\sigma_2^2 = \int_0^1\int_0^1 (\min(p,q)-pq)(J(p)-I(F))(J(q)-I(F))dF^{-1}(p)dF^{-1}(q).$$

Consistent estimates, $\hat{\sigma}_{AM}^2$ and $\hat{\sigma}_M^2$, are obtained by substituting the empirical df's $\hat{F}_{N_t}(y)$ and $\hat{F}_N(y)$, respectively, for $F(y)$ in the calculation of $\mu = T_\mu(F)$ and σ_2^2. Explicit expressions for the resulting estimators in the case of the Gini coefficient are given in Sandström (1983,p.181), and Nygård and Sandström (1981, p.384).

In the FP approach, with a general sampling design, the variance of the point estimator $I(\hat{F}_N)$ may be derived by a Taylor approximation similar to (3.1) giving

$$\sigma_{FP}^2 = \mu^{-2}\ \mathrm{Var}(T_J(\hat{F}_N) - I(F)T_\mu(\hat{F}_N)) \ . \qquad (5.3)$$

In evaluating σ_{FP}^2 it should be noted that $T_J(\hat{F}_N) = \int_0^1 J(p)\hat{F}_N^{-1}(p)dp$ is subject to sample-dependent random variations both through the inverse $\hat{F}_N^{-1}(p)$ and the weight function $J(p)$ (for details se Nygård and Sandström (1985)). As a consequence, the explicit variance expressions are quite excessive and awkward to adopt in practice without simplifying assumptions. The variance σ_{FP}^2 and its estimate $\hat{\sigma}_{FP}^2$ in the case of the Gini coefficient, for instance, involve inclusion probabilities up to the fourth order, cf. Nygård and Sandström (1985).

An alternative method of obtaining variance estimates in the FP approach is to use some subsampling technique involving systematic deletion of observations from the sample. As an illustration, consider the case of deleting one observation at a time and let $I(\hat{F}_N^{(i)})$ denote the point estimate based on n-1 observations, with the ith observation deleted. A variance estimator of jackknife type is then given by

$$\hat{\sigma}_J^2 = (n-1)n^{-1} \Sigma_i (I(\hat{F}_N^{(i)}) - I(\hat{F}_N^{(\cdot)}))^2, \qquad (5.4)$$

where $I(\hat{F}_N^{(\cdot)})$ denotes the mean of the estimates $I(\hat{F}_N^{(i)})$.

6. EMPIRICAL RESULTS

The variances of the point estimates of the finite population parameter $I(F_N)$ are in the M, AM, and FP approaches derived from first-order Taylor approximations. Consequently their accuracy depend on how well the approximations succeed in capturing the true variance in the sampling distribution. Moreover, even if the Taylor approximation is precise, the variance estimators $\hat{\sigma}_M^2$ and $\hat{\sigma}_{AM}^2$ are based on asymptotic considerations, and it is not clear how they behave in samples of small or moderate size. The variance estimator $\hat{\sigma}_{FP}^2$ escapes this kind of objections, but is on the other hand difficult to calculate in practice due to its dependence on higher-order inclusion probabilities. The jackknife estimator $\hat{\sigma}_J^2$ is questionable, as it lacks sufficient theoretical justification. As a consequence, the validity of the different variance estimators is in part elusive.

To illustrate the case of simple random sampling without replacement from a small population, Sandström, Wretman and Waldén (1985) considered two sets of populations, one consisting of four symmetric, and one consisting of three positively skewed populations. From the populations, all of the size N = 11 and within the two sets differing only with respect to location, the sampling distribution of the estimated Gini coefficient, based on n = 5 observations, was derived. As variance estimates $\hat{\sigma}_{AM}^2$, $\hat{\sigma}_{FP}^2$, and $\hat{\sigma}_J^2$ were used. A fourth variance estimator $\hat{\sigma}_{FP*}^2$, derived from (5.3), was included in an attempt to simplify $\hat{\sigma}_{FP}^2$. It assumes that the random variation in $T_J(\hat{F}_N)$ due to the weight function may be ignored, implying that only inclusion probabilities up to the second order matter. In Table 6.1 a summary of the results is given.

In the symmetric population set P1-P4 the precision of $\hat{R}_N$ decreases when the population mean gets close to zero. The same holds for the skewed populations P5-P7. In fact, four of the sampling distributions (P3, P4, P6, P7) are remarkably ill-conditioned. In addition, the four variance estimators clearly tend to

Table 6.1 Exact sampling distribution of $\hat{R}_N$ and the variance estimates. SRS (n=5) from seven small populations (N=11).

		P1	P2	P3	P4	P5	P6	P7
Population parameters	$\bar{Y}_N$	50	15.1	10.1	0.1	35.91	11.01	0.21
	G_N	14.02	14.02	14.02	14.02	19.18	19.18	19.18
	R_N	0.1402	0.4641	0.6939	70.08	0.2670	0.8708	45.85
Sampling distribution	$E(\hat{R}_N)$	0.1245	0.4606	0.9571	0.0915	0.2270	0.8211	-0.2265
	$V(\hat{R}_N)$	0.0029	0.0635	1.581	250.6	0.0036	1.418	287.3
	$E(\hat{\sigma}^2_{FP})$	0.0033	0.1678	148.6	990 850	0.0026	77.08	1 286 000
	$E(\hat{\sigma}^2_{FP*})$	0.0416	0.1188	145.0	990 760	0.0452	76.90	1 286 000
	$E(\hat{\sigma}^2_{AM})$	0.0026	0.1517	119.9	792 630	0.0012	61.71	1 029 000
	$E(\hat{\sigma}^2_{J})$	0.0049	0.1809	163.5	7 486	0.0087	12.45	1 245

exaggerate the dispersion in the sampling distribution. The deviating behavior of $\hat{\sigma}^2_{FP*}$ in P1 and P5, and of $\hat{\sigma}^2_J$ in P4 and P7 is also quite startling.

In order to demonstrate the performance of the estimators in larger populations, Sandström, Wretman and Walden (1985) conducted a Monte Carlo study, using logistic, uniform, normal,lognormal, Pareto, and standard Weibull parent distributions to construct finite populations of the size N=10000. Adopting a srs design with replacement and sample sizes of n=5, 10, 20, and 100, the sampling distribution of $\hat{R}_N$ was approximated form 500 replicates. In Table 6.2 an excerpt from the case n=100 is given.

As compared with the small population case, the point estimates $\hat{R}_N$ are now more well-behaved, even if a slight tendency to underestimation is apparent. The relative precision of $\hat{R}_N$, as measured by the coefficient of variation among the 500 replicates, is however still quite poor, ranging from 0.048 (uniform parent population) to 0.213 (Pareto), Yet, the variance estimates $\hat{\sigma}^2_{AM}$, $\hat{\sigma}^2_{FP}$, and $\hat{\sigma}^2_J$ are on the average fairly close to the observed sampling variances, and in this sense they seem to capture the dispersion. On the other hand, the crude design estimates $\hat{\sigma}^2_{FP*}$ are marked by huge overshooting, except in the Pareto case.

To inspect the consequences of passing from the srs case to a somewhat more complex sampling design, a simulation study was conducted. Using one panel

Table 6.2 Approximated sampling distributions, based on 500 replicates, of $\hat{R}_N$ and the variance estimates. SRS (n=100) from six parent populations (N=10 000).

		UNIFORM	LOGISTIC	NORMAL	LOGNORMAL	WEIBULL	PARETO
Population parameters	$\bar{y}_N$	5	5	5	244.7	2	3
	σ_N	0.57	1.81	1	320.8	0.48	-
	G_N	0.666	1.998	1.127	253.8	1.172	2.932
	R_N	0.0666	0.1998	0.1127	0.5185	0.2929	0.4886
Approximated sampling distribution	$\hat{E}(\hat{R}_N)$	0.0659	0.1994	0.1115	0.5053	0.2907	0.4520
	$\hat{V}(\hat{R}_N)$	10	363	65	1390	404	9310
	$\hat{E}(\hat{\sigma}^2_{FP})$ 10^{-6}x	10	334	70	1332	381	4913
	$\hat{E}(\hat{\sigma}^2_{FP*})$	3282	3152	3233	4073	3022	8844
	$\hat{E}(\hat{\sigma}^2_{AM})$	10	332	69	1235	383	4403
	$\hat{E}(\hat{\sigma}^2_J)$	10	336	71	1547	390	10050

(N=5412 households) of the Swedish income distribution survey carried out by Statistics Sweden in 1982 as the parent population, 500 replicates of a stratified sample (srs within strata) were drawn. Based on a sample size of n=300 each sample was allocated among seven strata according to the relative size of each stratum within the whole Swedish population. Due to its intractability, the variance estimator $\hat{\sigma}^2_{FP}$ was excluded. Table 6.3 summarizes the simulation results.

Table 6.3 Approximated sampling distributions, based on 500 replicates, of $\hat{R}_N$ and variance estimates. Stratified sampling (n=300) form Swedish 1982 income data (N=5412).

Population	Approximated sampling distribution				
R_N	$\hat{E}(\hat{R}_N)$	$\hat{V}(\hat{R}_N)$	$\hat{E}(\hat{\sigma}^2_{FP*})$	$\hat{E}(\hat{\sigma}^2_{AM})$	$\hat{E}(\hat{\sigma}^2_J)$
0.2925	0.2847	150	1432	152	163
		$\underbrace{\qquad\qquad}_{x10^{-6}}$			

The approximated sampling distribution of $\hat{R}_N$, with a coefficient of skewness equal to -0.074 and a kurtosis amounting to -0.019 turns out to be quite close to a normal distribution. The rather large bias is a somewhat surprising flaw in the sampling distribution, but may probably be attributed to an abortive allocation of the sample. The variance estimates $\hat{\sigma}^2_{AM}$ and $\hat{\sigma}^2_J$ seem in this case too to be successful in capturing the sampling dispersion, whereas $\hat{\sigma}^2_{FP*}$ once again results in large overestimates.

To get an idea of the sampling error involved in estimating the finite population Gini coefficient from large samples, the variance estimates $\hat{\sigma}^2_{AM}$ and $\hat{\sigma}^2_J$ were, in the light of the above results, applied to the income data of the 1982 Swedish income distribution survey. From this stratified sample, of the size n=10234 households, point estimates were obtained as $\hat{R}_N$=0.3215 in the case of disposable income and as $\hat{R}_N$=0.2099 in the case of disposable income per consumption unit. Relying on the corresponding variance estimates $\hat{\sigma}^2_{AM}$ and $\hat{\sigma}^2_J$, we get the following approximate 95 % confidence intervals for R_N:

	Disposable income/household	Disposable income/consumption unit
AM	0.3215 ± 0.0075	0.2099 ± 0.0075
J	0.3215 ± 0.0062	0.2099 ± 0.0053

7. DECOMPOSITIONS

In many papers on income inequality, the determinants of inequality have been discussed and the contribution of various components of total inequality has been subjected to measuring efforts. One method of analysis is to decompose income inequality by either income determining characteristics (subgroups) or by income sources. A general question is then: How much of the total inequality is *attributable to the variability in various income determinants and how much to* various income sources?

This question produces two different types of decomposition rules. According to the first, we have to subdivide the population into mutually exclusive subgroups and according to the second, we have to consider total income as a sum of different income sources.

Decompositions of the Gini coefficient and some members of the Gini family are discussed in e.g. Nygård and Sandström (1981) and in Sandström (1983) asymptotic results, similar to those above, are obtained.

BIBLIOGRAPHY

ATKINSON,A.B.(1970): On the Measurement of Inequality, J.Econ. Theory, Vol.2

BEACH,C.M. and DAVIDSON,R.(1983):Distribution-Free Statistical Inference with Lorenz Curves and Income Shares, Rev. Econ. Stud., Vol. 50

BOOS,D.D. (1979): A Differential for L-Statistics, Ann. Statist., Vol. 7

BREWER,K.R.W.(1981): The Analytical Use of Unequal Probability Samples: A Case Study, Invited Paper, 43rd Session of the International Statistical Institute,Buenos Aires.

CHERNOFF,H., GASTWIRTH,J.L., and JOHNS,M.V. Jr.(1967): Asymptotic distribution of linear combinations of order statistics, with applications to estimation, Ann.Math.Stat.,Vol.38

COWELL,F.A.(1980): On the Structure of Additive Inequality Measures, Rev.Econ. Stud.,Vol.47

DALTON,H.(1920): The Measurement of the Inequality of Income, Econ.J., Vol.30

DAVID,H.A.(1981):Order Statistics, 2nd. ed., John Wiley & Sons, New York

FELLER,W.(1966): An Introduction to Probability Theory and Its Applications. Vol.II, John Wiley & Sons, New York

GINI,C.(1912):Variabilità e mutabilità, Bologna

-- (1914): Sulla misura della concentrazione e della variabilita del caratteri, Atti del R. Istituto Veneto, Bd. 73, 1913-1914

GLASSER,G.J.(1962):Variance Formulas for the Mean Difference and Coefficient of Concentration, JASA, Vol.57

GOLDIE,C.M.(1977):Convergence Theorems for Empirical Lorenz Curves and their Inverses, Adv. Appl. Prob., Vol.9

HOEFFDING,W.(1948): A Class of Statistics with Asymptotically Normal Distribution, Ann. Math. Statist., Vol.19

HOEM,J.M. and Funck -JENSEN,U.(1982): Multistate life table methodology: A Probabilistic Critique, in K.C.Land and A.Rogers(eds.):Multidimensional Mathematical Demography, Academic Press New York

JUNG,J.(1955):On linear estimates defined by a continuous weight function, Arkiv för matematik, Band 3, nr 15

KOLM,S.C.(1976): Unequal Inequalities I and II, J. Econ. Theory, Vol. 12

KOUL,H.L.(1970): Some Convergence Theorems for Ranks and Weighted Empirical Cumulatives, Ann. Math. Statist., Vol. 41

LOMNICKI,Z.A.(1952): The Standard Error of Gini's Mean Difference, Ann. Math. Statist.,Vol.23

LORENZ,M.O.(1905): Methods for Measuring Concentration of Wealth, JASA, New Series No. 70

LOVE,R. and WOLFSON,M.C.(1976): Income Inequality: Statistical Methodology and Canadian Illustrations, Statistics Canada, Catalogue 13-559, Occasional, March

MENDERSHAUSEN,H.(1939): On the Measurement of the Degree of Inequality of Income Distributions, Cowles Commission for Research in Economics, Univ. of Chicago

MOORE,D.S.(1968): An Elementary Proof of Asymptotic Normality of Linear Functions of Order Statistics, Ann. Math. Statist., Vol. 39

NAIR,U.S.(1936): The Standard Error of Gini's Mean Difference, Biometrika, Vol. 28

NYGÅRD,F.(1981): Mätning av inkomstojämnhet - en studie av ett statistiskt operationaliseringsproblem. Medd. från ekonomisk-statsvetenskapliga fakulteten vid Åbo Akademi, Statistiska Institutionen, Ser.A:163, Turku (in Swedish)

NYGÅRD,F. and SANDSTRÖM,A.(1981):Measuring Income Inequality, Almqvist & Wiksell International, Stockholm

-- (1985): Estimating Gini and Entropy Inequality Parameters, Memo No.13, Statistical Research Unit, Statistics Sweden, 1985-01-09

ROTHSCHILD,M. and STIGLITZ,J.E.(1970):Increasing Risk:I. A Definition, J.Econ.Theory,Vol.2

ROYALL,R.M.(1971): Linear Regression Models in Finite Population Sampling Theory in V.R. Godambe and D.A.Sprott (eds):Foundations of Statistical Inference, Holt,Rinehart and Winston of Canada, Toronto, Montreal

SANDSTRÖM,A.(1982):Estimating the Gini Coefficient, Dept. of Statist.,Univ. of Stockholm, Research Report 1982:18

-- (1983):Estimating Income Inequality, Large Sample Inference in Finite Populations, Dept. of Statist., Univ. of Stockholm, Research Report 1983:5

SANDSTRÖM,A.,WRETMAN,J.H.,and WALDEN,B.(1985): Variance Estimators of the Gini Coefficient, Simple Random Sample, Statistical Research Unit, Statistics Sweden, Memo February 1985

SENDLER,W.(1979): On Statistical Inference in Concentration Measurement, Metrika Vol. 26

SERFLING,R.J.(1980): Approximation Theorems of Mathematical Statistics, Wiley & Sons, New York

SHORACK,G.R.(1972): Functions of Order Statistics, Ann. Math. Statist., Vol. 43

SHORROCKS,A.F.(1980): The Class of Additively Decomposable Inequality Measures, Econometrica Vol. 48

-- (1983): Inequality Decomposition by Population Subgroups, Discussion Paper, Univ. of Essex

SILLITTO,G.P.(1969): Derivation of approximants to the inverse distribution function of a continuous univariate population from the order statistics of a sample, Biometrika, Vol. 56

STIGLER,S.M.(1974): Linear Function of Order Statistics with Smooth Weight Functions, Ann. Statist., Vol. 2

WOLD,H.(1935): A Study on the Mean Difference, Concentration Curves and Concentration Ratio, Metron, Vol. 12

YANG,S.S.(1977): Linear Functions of Concomitants of Order Statistics, Tech. Report No. 7, Sept. 1977, Dept of Math., MIT Cambridge, Mass.

-- (1981): Linear Functions of Concomitants of Order Statistics, With Application to Nonparametric Estimation of a Regression Function, JASA, Vol. 76.

SUMMARY

By relating the Gini coefficient to a general family of inequality measures, the Gini family, methods of estimating inequality parameters from samples are reviewed. Three approaches to making inferences about unknown inequality parameters are discussed and some large-sample results are presented. Alternative variance estimates are presented and compared in the case of the Gini coefficient. Two different ways of decomposing the members of the Gini family are briefly discussed.

RESUMÉ

En relatant le coefficient de GINI à une famille générale de measures d'inégalité, la famille GINI, des méthodes d'estimation de paramètres d'inégalité sont examinés.

Trois manières d'aborder l'inférence concernant des paramètres d'inégalité sont discutées et des résultats basés sur de grands échantillons sont présentés. Plusieurs estimateurs de variance sont présentés et comparés dans le cas du coefficient de GINI. Deux façons différentes de décomposer les membres de la famille GINI sont discutés brièvement.

Tidigare nummer av Promemorior från P/STM:

NR

1 Bayesianska idéer vid planeringen av sample surveys. Lars Lyberg (1978-11-01)

2 Litteraturförteckning över artiklar om kontingenstabeller. Anders Andersson (1978-11-07)

3 En presentation av Box-Jenkins metod för analys och prognos av tidsserier. Åke Holmén (1979-12-20)

4 Handledning i AID-analys. Anders Norberg (1980-10-22)

5 Utredning angående statistisk analysverksamhet vid SCB: Slutrapport. P/STM, Analysprojektet (1980-10-31)

6 Metoder för evalvering av noggrannheten i SCBs statistik. En översikt. Jörgen Dalén (1981-03-02)

7 Effektiva strategier för estimation av förändringar och nivåer vid föränderlig population. Gösta Forsman och Tomas Garås (1982-11-01)

8 How large must the sample size be? Nominal confidence levels versus actual coverage probabilities in simple random sampling. Jörgen Dalén (1983-02-14)

9 Regression analysis and ratio analysis for domains. A randomization theory approach. Eva Elvers, Carl Erik Särndal, Jan Wretman och Göran Örnberg (1983-06-20)

10 Current survey research at Statistics Sweden. Lars Lyberg, Bengt Swensson och Jan Håkan Wretman (1983-09-01)

11 Utjämningsmetoder vid nivåkorrigering av tidsserier med tillämpning på nationalräkenskapsdata. Lars-Otto Sjöberg (1984-01-11)

12 Regressionsanalys för f d statistikstuderande. Harry Lütjohann (1984-02-01)

13 Estimating Gini and Entropy inequality parameters. Fredrik Nygård och Arne Sandström (1985-01-09)

Kvarvarande exemplar av ovanstående promemorior kan rekvireras från Elseliv Lindfors, P/STM, SCB, 115 81 Stockholm, eller per telefon 08 7834178

☆ Garnisonstryckeriet Stockholm 1985 ☆