SCB STATISTISKA CENTRALBYRÅN

1985-01-09

PROMEMORIOR FRÅN P/STM

NR 13

ESTIMATING GINI AND ENTROPY INEQUALITY PARAMETERS

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 1985:13. Estimating Gini and entropy inequality parameters / Fredrik Nygård; Arne Sandström.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-PM-PSTM-1985-13

STATISTICS SWEDEN Working Paper 1985-01-09

Statistical Research Unit

E S T I M A T I N G G I N I A N D E N T R O P Y

I N E Q U A L I T Y P A R A M E T E R S

Fredrik Nygård[1)] Arne Sandström[1)]

[1)] Fredrik Nygård is Research Assistent, Department of Statistics, Swedish University of Turku,SF-20500 Turku 50,Finland and Arne Sandström is Director, Statistical Research Unit, Statistics Sweden,S-115 81 Stockholm, Sweden. The research was supported by the Joint Committé of the Nordic Social Research Council.

A B S T R A C T

This paper examines two families of inequality parameters, frequently used as measures of income inequality, viz. the Gini family and the Generalized Entropy family. Computations in total surveys and estimation in sample surveys are discussed. The estimation procedures are made both under a fix population approach and under an auxiliary model approach. Various variance estimators are discussed and for the Gini coefficient the sampling distributions of the point estimator and the variance estimators from two small populations are compared.

K E Y W O R D S:

Inequality parameters, Gini family, Generalized Entropy family, Total surveys, Sample surveys, Variance estimation.

1 INTRODUCTION

When describing a set of data - or comparing two or more data sets - the variance is the most frequently used measure of dispersion.

Another way of describing variability has emerged from studies of the size distribution of income. In the case of income data, dispersion is often interpreted as reflecting "income inequality" and in order to assess its magnitude particular measures ("measures of income inequality") have been derived from assumptions ("criteria") on how a measure should respond to specific changes in the income distribution. An example of such a measure of income inequality is the well-known Gini coefficient.

These dispersion measures will here be called inequality parameters to point out that their field of application is not only restricted to income distributions. In fact, applications to e.g. trading balance, unemployment, consumption, and residential density are found in the literature and, in general, inequality parameters may be calculated for any quantitative data set.

In this paper we show how some commonly used inequality parameters may be computed in total surveys and estimated in sample surveys. To be more specific, we focus on two families of inequality parameters, viz. the Gini and the Generalized Entropy families.

The paper is organized in the following way: The inequality parameters are defined in Section 2 using statistical functionals. In Section 3 we discuss parameter computation in total surveys based on complete

or grouped data. Estimators, and variance estimators, based on probability samples are discussed in Section 4. In the Appendix the variance estimators are compared for the Gini coefficient under a simple random sampling design.

2 INEQUALITY PARAMETERS

In this section we will pick up two frequently used classes of inequality parameters and give their formal definitions by use of a functional approach. The first class is choosen because its relation to the well-known Lorenz Curve (LC). The area between the LC and the diagonal line in a Lorenz diagram is oftenly used as a measure of income inequality. This first class is defined as a weighted Lorenz area and will be called the Gini family, because it includes the Gini coefficient of income inequality as a member. The second class of parameters is the Generalized Entropy family, which is choosen because it has been proved that the members of this family are the only parameters that fulfill some special criteria imposed on inequality measures, see e.g. Cowell (1980), and that these members are the only parameters that can be decomposed in accordance with the proposals given by Shorrocks (1980), (1983).

2.1 Definitions by a functional approach

In defining the two families of inequality parameters it will prove convenient to represent all parameters as statistical functionals (or ratios of statistical functionals) by use of the Lebesgue-Stiltjes integral. Let the variate Y have a distribution function (df) $F_Y(y)$ with $E(Y) = \mu_Y \neq 0, < \infty$. In terms of a statistical functional μ_Y can be written as

$$T_\mu(F) = \int_{-\infty}^{\infty} y dF_Y(y). \tag{2.1}$$

In a total survey of a finite population, cf. Section 3, with the finite population df F_N, (2.1) becomes

$$T_\mu(F_N) = \int_{-\infty}^{\infty} y dF_N(y) = N^{-1} \sum_{i=1}^{N} y_i = \bar{y}_N \tag{2.2}$$

and an estimate of (2.2) based on a sample survey is obtained by (i) estimating F_N and (ii) changing F_N for its estimate, say $\hat{F}_N$, i.e.

$$T_\mu(\hat{F}_N) = \int_{-\infty}^{\infty} y d\hat{F}_N(y). \tag{2.3}$$

The last procedure is discussed in Section 4.

The inequality parameters that we will discuss here are all relative measures of dispersion, i.e. they are scale invariant. The two families of parameters that we consider are

+ <u>the Gini family</u>:

$$I_G(F) = T_G(F)/T_\mu(F), \tag{2.4}$$

where $T_G(F) = \int_{-\infty}^{\infty} J(F(y)) y dF(y)$
and $J(\cdot)$ is a smooth function.

+ the Generalized Entropy family:

$$I_{E,c}(F) = \frac{1}{c(c-1)} \{ T_c(F) / T_\mu(F)^c - 1 \}, \qquad c \neq 0,1 \qquad (2.5a)$$

where $T_c(F) = \int_{-\infty}^{\infty} y^c dF(y)$,

with the limiting value, see e.g. Shorrocks (1982), when $c \to 0$ and $c \to 1$:

$$I_{E,c}(F) = (-1)^{1-c} T_c(F)/T_\mu(F)^c, \qquad c = 0,1 \qquad (2.5b)$$

where $T_c(F) = \int_{-\infty}^{\infty} y^c \log(y/T_\mu(F)) dF(y)$.

The $J(\cdot)$ - function of the Gini family is sometimes referred to as a weight function since the parameters of this family may be interpreted as weighted Lorenz areas.

In Table 2.1 some examples of parameters belonging to the above families are given.

- - - - - - - - -

Table 2.1 in here

- - - - - - - - -

Since the main objective of this report is on estimation we have no intention to discuss the relevance of any members of the two families. That is a question for the user of income inequality measures.

2.2 *Some useful results*

The following two propositions can prove helpful when analysing inequality parameters, e.g. in variance estimation. The reformulations of the parameters proposed here assume that F is continuous.

PROPOSITION 2.1 Assume F to be continuous and let $W(\cdot)$ be a monotone non-decreasing function and assume $\int_0^1 |W(p)F^{-1}(p)|dp$ to exist, where $F^{-1}(p) = \inf_x \{x|F(x)\geq p\}$, $0<p\leq 1$, and $F^{-1}(0) = \inf_x\{x|F(x)>0\}$. If we assume that $\int_0^1 W(p)dp < \infty$ then

$$\int_0^1 W(p)F^{-1}(p)dp = \frac{1}{2}\int_0^1\int_0^1 |W(p) - W(q)|\cdot|F^{-1}(p) - F^{-1}(q)|dpdq +$$

$$+ \mu\int_0^1 W(p)dp, \qquad (2.6)$$

where $\mu = \int_0^1 F^{-1}(p)dp < \infty$.

The proof follows by changing order of integration in the first term on the right hand side.

REMARK 2.1 If $W(\cdot)$ is monotone non-increasing the Proposition is valid if we change sign on the right hand sid of (2.6).

REMARK 2.2 Note the following special cases of the Proposition:

i) $W(p) = c$, constant, is trivial.

ii) $\sigma^2 = \int_0^1 (F^{-1}(p) - \mu)^2 dp = \int_0^1 W(p)F^{-1}(p)dp$, with $W(p) = F^{-1}(p) - \mu$.

By (2.6) σ^2 can be written as $\int_0^1\int_0^1 \frac{1}{2}(F^{-1}(p) - F^{-1}(q))^2 dpdq$, where

$\frac{1}{2}(F^{-1}(p) - F^{-1}(q))^2$ is the symmetric kernel corresponding to the

U-statistic equal to the sample variance.

iii) The central moment μ_k can by (2.6) be written as

$$\mu_k = \sum_{i=1}^{k-1} \frac{1}{2} \mu^{i-1} \int_0^1 \int_0^1 |(F^{-1}(p) - \mu)^{k-i} - (F^{-1}(q) - \mu)^{k-i}| \cdot |F^{-1}(p) - F^{-1}(q)| \, dpdq + \mu^k.$$

iv) If W(p) is the function corresponding to a linear function of order statistic, usually denoted J(p), then (2.6) is obvious.

If W(p) = J(p) as in Remark 2.2 iv) and J(p) is a power function then the following Proposition can be used to rewrite $\int_0^1 J(p)F^{-1}(p)dp$.

PROPOSITION 2.2 Let J(p) be a power function in p, power $r \geq 1$, and assume

$\int_0^1 |J(p)F^{-1}(p)|dp$ to exist, where F is continuous and F^{-1} is

defined as in Proposition 2.1. Then

$$\int_0^1 J(p)F^{-1}(p)dp = \int_0^1 \int_0^1 D(p)|F^{-1}(p) - F^{-1}(q)|dpdq, \qquad (2.7a)$$

where D(p) is a function in p of power r-1 and the following relation between J(p) and D(p) holds

$$J'(p) = (2p-1)D'(p) + 4D(p). \qquad (2.7b)$$

The proof is straightforward.

REMARK 2.3 The parameters belonging to the Gini family can either be rewritten according to (2.6) or to (2.7a). As an example, take the Gini coefficient where J(p) = 2p-1. Then by (2.6) we have, since $\int_0^1 (2p-1)\, dp = 0$,

$$R = \frac{1}{\mu} \int_0^1 \int_0^1 |p-q| \cdot |F^{-1}(p) - F^{-1}(q)|dpdq.$$

J'(p) = 2 and hence, by (2.7b), $D(p) = \frac{1}{2}$,

which gives, by (2.7a),

$$R = \frac{1}{2\mu} \int_0^1 \int_0^1 |F^{-1}(p) - F^{-1}(q)| \, dpdq.$$

In Table 2.2 the parameters of Table 2.1 are rewritten according to Propositions 2.1 and 2.2.

- - - - - - - - - -
Table 2.2 in here
- - - - - - - - - -

3 TOTAL SURVEYS

3.1 Calculations in Total Surveys

The computation of the inequality parameter in a finite population is, in view of the functional approach, straightforward. The finite population df F_N is defined as

$$F_N(y) = N^{-1} \sum_{i=1}^{N} I_{\{y_i \leq y\}} , \tag{3.1}$$

where $I_{\{\cdot\}}$ is the indicator function taking on the value 1 when the event $\{\cdot\}$ occure and the value 0 otherwise.

REMARK 3.1 The data set in the finite population, $\underset{\sim}{y}_N = (y_1,\ldots,y_N)$, is a fixed vector.

REMARK 3.2 If the observations in $\underset{\sim}{y}_N$ are arranged in non-decreasing order, i.e. $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(N)}$ then we can write (3.1) as

$$F_N(y) = \begin{cases} 0 & \text{if } y < y_{(1)} \\ i/N & \text{if } y_{(i-r+1)} \leq y < y_{(i+1)} \text{ and } y_{(i-r)} < y_{(1-r+1)} = \\ & \qquad = y_{(i-r+2)} = \cdots = y_{(i)} < y_{(i+1)} \\ 1 & \text{if } y_{(N)} \leq y. \end{cases}$$

Both this definition and (3.1) include the possibility of ties.

The arithmetic mean in the finite population is given by (2.2). If there is $N' \leq N$ distinct values of y then we define the probability function at $y_{(i)}$ as

$$f_N(y_{(i)}) = F_N(y_{(i)}) - F_N(y_{(i-1)}). \tag{3.2}$$

With use of (3.2) we get

$$T_\mu(F_N) = \bar{y}_N = \sum_{i=1}^{N'} y_i f_N(y_i).$$

The Gini family is defined by $I_G(F_N) = T_G(F_N)/T_\mu(F_N)$, where

$$T_G(F_N) = \int_{-\infty}^{\infty} J(F_N(y))y\,dF_N(y) = \sum_{i=1}^{N'} J(F_N(y_i))y_i f_N(y_i) \tag{3.3a}$$

and if no tied y-values are present we can rewrite (3.3a) as a linear function of the ordered data set (with use of $F_N(\cdot)$ given in Remark 3.2)

$$T_G(F_N) = N^{-1} \sum_{i=1}^{N} J(i/N)y_{(i)}, \tag{3.3b}$$

so the computation is straightforward when the observations are rank-ordered.

REMARK 3.3 The weight function for the finite population Gini coefficient is according to Table 2.1, for (3.3a) $J(F_N(y_i)) = 2F_N(y_i) - 1$ and for (3.3b) $2\frac{i}{N} - 1$. For a non-negative variate $R_N \in [1/N,1]$. The usual definition found in the literature, cf. e.g. Nygård and Sandström (1981), is based on Proposition 2.2. With use of the result in Remark 2.3 the J-function corresponding to (3.3a) will be $2F_N(y_i) - 1 - f_N(y_i)$, where $f_N(y_i)$ is defined by (3.2). In the case of rank-ordered data we get $2\frac{i}{N} - 1 - \frac{1}{N}$. The term $-\frac{1}{N}$ will be called the Gini finite population correction (Gfpc). In the non-negative case with R_N including the Gfpc-term we have $R_N \in [0,1 - \frac{1}{N}]$. There are at least three reasons for making this correction, viz. i) the lower bound of the parameter is zero for non-negative data (the RANGE criterion in e.g. Nygård and Sandström (1981)), ii) the REPLIC criterion is fulfilled, cf. op. cit., and iii) the bias in the sample estimator $T_G(\hat{F}_N)$ is decreased.

In the sequel we will use finite population corrected parameters of the Gini family, see Remark 3.3. In Table 3.1 explicit expressions for some members belonging to the Gini family are given and in Table 3.2 we have explicit expressions for parameters of the Generalized Entropy family.

- - - - - - - - -
Table 3.1 in here
- - - - - - - - -

- - - - - - - - -
Table 3.2 in here
- - - - - - - - -

3.2 Calculations from grouped data

In practice we frequently have to deal with situations in which we - instead of having access to the complete data - are provided only with data in condensed form (frequency tables etc.).

In this section we address the problem of how to calculate parameters of the Gini and Generalized Entropy family in these cases.

One method of calculating parameters from grouped data starts out from some specific assumption regarding the behaviour of the distribution function $F_N(y)$ within the different groups - a vast amount of suggestions are found in the literature (for references see e.g. Nygård and Sandström (1981), p.113, Dagum (1983), MacDonald (1984). According to other related methods the parameter calculation is based on some interpolation/extrapolation technique (cf. Gastwirth and Glauberman (1976), Kakwani (1980), Cowell and Mehta (1982)).

In contrast to these methods, the approach reported in this section is basically 'non-parametric' (cf. Gastwirth (1975)) in that it provides lower and upper bounds for the parameter value inherent in the population without any distributional assumptions on the complete data.

We start out by assuming that the available information about the distribution is given in a frequency table with the range divided into k intervals with boundaries

$]\, a_{i-1},\, a_i]$, $a_{i-1} < a_i$, $i=1,\ldots,k$, where $a_0 \geq 0$ and $a_k < \infty$.

Let N_i and $\bar{y}_i$ denote the frequency and mean respectively, within group i, $i=1,\ldots,k$, $\sum_i N_i = N$, $\sum_i N_i \bar{y}_i = N\bar{y}_N$.

In this situation the standard textbook method of calculating the Gini and Entropy parameters of Table 3.1 and 3.2 substitutes the group means $\bar{y}_i$ into the calculation formulas - implicitly assuming that all observations within each group equal the group mean. Actually, this is in a very precise sence a sound procedure, since it may readily be seen that substitution of group means into the complete data formulas minimizes the Gini and Entropy parameters subject to the restriction of fixed means. As a consequence, the resulting parameter values are negatively biased as the corresponding complete data parameter in general will exceed the calculated value. An upper bound for this bias may be found by maximization of the parameter values subject to given group means and boundaries. It turns out (cf. Gastwirth (1975)) that the maximum is obtained by placing $(1-\lambda_i)N_i$ of the observations in group i at the lower boundary a_{i-1} and the remaining $\lambda_i N_i$ observations at the upper boundary a_i, where

$$\lambda_i = (\bar{y}_i - a_{i-1})/(a_i - a_{i-1})$$

is derived from the restriction of a fixed group mean.

REMARK 3.4 That the minimum parameter value occurs when all observations equal the group mean and the maximum value when the observations are placed at the group boundaries is actually an immediate consequence of the fact that the parameters under consideration satisfy the principle of transfers i.e. the parameter value increases if an amount $\Delta > 0$ is "transfered" from y_p to y_r, $y_p \leq y_r$.

Formulas for the lower bound and maximum bias, which added to the lower bound gives the upper bound, are presented in Table 3.3 for the Gini and Generalized Entropy parameters.

- - - - - - - - - - - - - - -

Table 3.3 in here

- - - - - - - - - - - - - - -

REMARK 3.5 Note that lower parameter bounds in the case of a decile type frequency table with $N_i = N/k$, $i=1,\ldots,k$, simply are obtained by substituting k for N and $\bar{y}_i$ for y_i in the complete data formulas.

REMARK 3.6 Upper bounds for the parameters of the Gini family may also be derived in the case of unknown boundary points, a_i, $i=1,\ldots,k$. See Mehran (1975), Nygård and Sandström (1981).

REMARK 3.7 The parameter bounds may readily be sharpened by introducing additional assumptions on the distribution within the separate groups. See e.g. Gastwirth (1972), (1975) for an application to the case when data has a decreasing density in some interval.

REMARK 3.8 Upper and lower parameter bounds may also prove useful when considering optional boundary points for data presentation. Optimal methods for grouping, when the purpose is to calculate parameters of the Gini family, are found in Aghevli and Mehran (1981).

REMARK 3.9 Note that the 'non-parametric' parameter bounds derived from grouped data not should be confused with 'confidence' statements about the true parameter value when data is obtained through sampling. The sampling case, in which the expressions of Table 3.3 give bounds on the parameter estimator, will be addressed in Section 4. See also Beach and Davidson (1983) for a discussion of the estimation problem when only grouped sample data is available.

4 SAMPLE SURVEYS

4.1 The fix population approach

Assume a finite and identifiable population of size N. The identifiability assumption makes it possible to uniquely label the population units from 1 to N. We also assume that the label of each unit is known, which implies that we can define a label set $U = \{1,2,\ldots,N\}$ of the population universe. With the jth unit, $j \in U$, we associate some number y_j, which can be seen as a reslut of measuring unit j (the y_j can be a vector of numbers).

A sample s is a subset of U, i.e. $s = \{j_i | j_i \in U,\ i=1,2,\ldots,n(s)\}$, where n(s) is the sample size which may depend on s. 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 and observed. $\{P(s),\ s \subset U\}$ is called the sampling design (plan). In the sequel we only consider fixed size designs, i.e. n(s) = n, where the sampling procedure is taken without replacement. $f_n = n/N$ is called the sampling fraction, $0 < f_n < 1$ ($f_n = 1$ implies a total survey, see Section 3).

The inclusion probability of first order of unit i is defined as $\pi_i = P(i \in s)$ and the second-order inclusion probability of units i and j as $\pi_{ij} = P(i,j \in s)$, $i \neq j$. Higher order inclusion probabilities can be defined in a similar way. For a fixed size design $\sum_{i \in U} \pi_i = n$. Let us define the inclusion indicator, which we will have much use of, as

$$I_{\{i \in s\}} = \begin{cases} 1 & \text{if } i \in s \\ 0 & \text{otherwise .} \end{cases} \qquad (4.1)$$

If we are summing over the sample s we write either $\sum_{i\in s}$ or $\sum_{i=1}^{n}$ depending on the situation and in a similar way when summing over the whole population (cf. above). Note that s in $\sum_{i\in s}(\cdot)$ is stochastic but by use of (4.1) we can rewrite the sum in the following way:

$\sum_{i\in s}(\cdot) = \sum_{i\in U} I_{\{i\in s\}}(\cdot)$, where U is constant.

The expectation of the inclusion indicator is

$$E(I_{\{i\in s\}}) = \pi_i, \ \forall i\in U. \tag{4.2}$$

An unbiased estimator of the population size is $\hat{N}_s = \sum_{i\in s}\pi_i^{-1}$. This is simply proved by use of (4.2)

$$E(\hat{N}_s) = E(\sum_{i\in U} I_{\{i\in s\}}\pi_i^{-1}) = \sum_{i\in U} E(I_{\{i\in s\}})\pi_i^{-1} = N.$$

For simple random sampling (srs) the first order inclusion probability is $\pi_i = n/N$, $\forall i$, and hence $\sum_{i\in s}\pi_i^{-1} = N$.

By the functional representation of the inequality parameters introduced in section 2 we only have to estimate the finite population df F_N to obtain point estimates. The following definition gives an estimator of the df F_N.

DEFINITION 4.1 An estimator of the finite population df F_N is

$$\hat{F}_N(y) = \hat{N}_s^{-1} \sum_{i\in s} I_{\{y_i\leq y\}}/\pi_i, \quad \forall y, \tag{4.3}$$

where $\hat{N}_s = \sum_{i\in s}\pi_i^{-1}$.

REMARK 4.1 The estimator (4.3) is a Hajek estimator which is a modification of the Horvitz-Thompson (HT-) type estimator. The estimator is biased since it is a ratio of two HT-estimators. If $\hat{N}_s$ is changed for N, the correct population size, then the estimator (4.3) would be unbiased, but it will not have all the properties of a df since $\hat{F}_N(\infty) \gtrless 1$ depending on the ratio $\hat{N}_s/N$.

DEFINITION 4.2 A Hajek estimator of the finite population inequality parameter $I(F_N)$ based on a design $\{P(s), s \subset U\}$ is $I(\hat{F}_N)$, where $\hat{F}_N$ is defined in Definition 4.1.

Explicit estimation expressions are given in Table 4.1 for the parameters under consideration. The estimation procedure in the Gini case has to be done in two steps: i) data is arranged in increasing order such that $y_{j_1} \leq y_{j_2} \leq \ldots \leq y_{j_n}$, $j_i \in s$, and then ii) straightforward computation.

- - - - - - - - -
Table 4.1 in here
- - - - - - - - -

REMARK 4.2 Even if we assume $\hat{N}_s \approx N$, and having approximately unbiased estimators of F_N, the estimators of the inequality parameters are biased since they are ratios.

REMARK 4.3 The expression for the Gini coefficient given by Brewer (1981) is based on a reformulation of R_N. Different reformulations of R_N are given in Nygård and Sandström (1981).

4.2 Variance estimators

Both the procedure of estimating the finite population df F_N and the structure of the parameters to be estimated imply that the resulting estimators are ratio-estimators. Hence both the numerator and the denominator are stochastic. In estimating the variances of the estimators directly, and not using subsample procedures, we can make use of a frequently used approximation method, viz. a method based on a first order Taylor approximation technique. To illustrate this let t_y and t_x be the totals of y and x, respectively, and let the Horvitz-Thompson (HT-) estimators be $\hat{t}_y$ and $\hat{t}_x$, respectively. Define a ratio $r = t_y/t_x$ and its HT-type estimator by $\hat{r} = \hat{t}_y/\hat{t}_x = f(\hat{t}_y,\hat{t}_x)$. We Taylor expand $\hat{r} = f(\hat{T}_y,\hat{t}_x)$ about t_y and t_x as

$$\hat{r} - r = f(\hat{t}_y,\hat{t}_x) - f(t_y,t_x) =$$

$$= t_x^{-1}(\hat{t}_y - t_y) - (t_y/t_x^2)(\hat{t}_x - t_x) \qquad (4.4a)$$

$$= (\hat{t}_y - r\hat{t}_x)/t_x = \hat{t}_z/t_x, \qquad (4.4b)$$

where $\hat{t}_z = \sum_{i \in s} z_i/\pi_i$ and $z_i = y_i - rx_i$. According to (4.4a) the variance may be written as

$$V(\hat{r}) = t_x^{-2}V(\hat{t}_y) + (t_y/t_x^2)^2V(\hat{t}_x) - 2(t_y/t_x^3)\mathrm{Cov}(\hat{t}_y,\hat{t}_x) \qquad (4.4c)$$

and according to (4.4b)

$$V(\hat{r}) = t_x^{-2}V(\hat{t}_z). \qquad (4.4d)$$

The two variances are by definition identical. In the case of the Gini coefficient the numerator in the ratio (4.6) below will have a weight depending on the sample s. We will therefore use the Taylor approximation (4.4c) and take account for the stochasticity in the weights (method i)) when estimating the variance. We will also use a rough variance estimator according to the ratio variance (4.4d) (method ii)) and not take account for the stochasticity in the weights in the numerator.

The following Proposition gives an explicit formulation of the Yates-Grundy-Sen estimator of the covariance of two HT-estimators from a bivariate sample.

PROPOSITION 4.1 Assume a bivariate sample s with data (x_i,y_i), $i \in s$, and let $\hat{t}_y$ and $\hat{t}_x$ be the HT-estimators of the totals $t_y = \sum_{i \in U} y_i$ and $t_x = \sum_{i \in U} x_i$, respectively. Then the Yates-Grundy-Sen estimator of the covariance between $\hat{t}_y$ and $\hat{t}_x$ is

$$\hat{Cov}(\hat{t}_y,\hat{t}_x) = \frac{1}{2} \sum_{i,j \in s} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right) \left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right) \qquad (4.5)$$

Proof: The proof is similar to the proof of the Yates-Grundy-Sen variance estimator, see e.g. Cochran (1977 - pp.260-261).

REMARK 4.4 If y = x then, of course, $\hat{Cov}(\hat{t}_y,\hat{t}_x) = \hat{V}(\hat{t}_y)$, i.e. the ordinary Yates-Gurndy-Sen variance estimator.

REMARK 4.5 The covariance estimator of $Cov(\hat{t}_y, \hat{N})$, where $\hat{N}$ is the population size estimator, is

$\hat{Cov}(\hat{t}_y,\hat{N}) = \frac{1}{2}\sum_{i,j\in s}\sum \frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)\left(\frac{1}{\pi_i}-\frac{1}{\pi_j}\right)$. The variance estimator of $V(\hat{N})$ is given by the covariance estimator with $y_i = 1$, $\forall i\in s$.

REMARK 4.6 In Section 4.3 it will be shown, under an auxiliary model approach, that the estimators are asymptotically normal. Resting on large sample arguments we can construct approximately confidence intervals for the finite population parameters.

4.2.1 The Gini Family

The variances of the estimators belonging to the Gini family can either be estimated by method i) ("Taylor variance") or by method ii) ("Ratio variance").

First, we use method i) on the Gini coefficient. The ratio estimator corresponding to R_N is

$$\hat{R}_N = 2\frac{\hat{t}_{wy}}{N\,\bar{t}_y} - 1 = f(\hat{t}_{wy},\hat{t}_y,\hat{N}), \tag{4.6}$$

where $\hat{N} = \sum_{i\in s}\pi_i^{-1}$

$$\hat{t}_y = \sum_{i\in s} y_i/\pi_i$$

$$\hat{t}_{wy} = \sum_{i\in s} P_{si}y_i/\pi_i$$

and $P_{si} = \sum_{j\in s} I_{\{y_j<y_i\}}/\pi_j + \frac{1}{2\pi_i} = P_{s(i)} + \frac{1}{2\pi_i}$, $P_{s(i)} = \sum_{j\in s} I_{\{y_j<y_i\}}/\pi_j$.

Table 4.2 in here

The variance estimator of $V(\hat{R}_N)$ is given explicitly, be method i), in Table 4.2. It has the disadvantage of including up to the fourth order inclusion probabilities which is due to the stochastic weights.

The two other estimators, Mehran's and Piesch's, will include up to the sixth order inclusion probabilities! But, if we can estimate $V(\hat{R}_N)$ then we can make use of Proposition 2.1 or 2.2 to overestimate $V(\hat{M}_N)$ and $V(\hat{P}_N)$, the variances of Mehran's and Piesch's estimators, respectively.

Simpler, but cruder, variance estimators can be obtained to all estimators belonging to the Gini family by use of method ii). Let the finite population parameters be defined, in general, by

$$I_G(F_N) = \frac{1}{\bar{y}_N} \frac{1}{N} \sum_{i \in U} J(F_N(y_i))y_i, \quad \text{cf. (2.4)},$$

where the specific parameters R_N, M_N amd P_N depends on $J(\cdot)$, with corrections made for finite populations.

An estimator of $V(I_G(\hat{F}_N))$, using method ii), is

$$\hat{V}(I_G(\hat{F}_N)) = \frac{1}{\hat{t}_y^2} \frac{1}{2} \sum_{i,j \in s} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{\hat{\hat{z}}_i}{\pi_i} - \frac{\hat{\hat{z}}_j}{\pi_j}\right)^2, \qquad (4.7)$$

where $\hat{\hat{z}}_i = (J(\hat{F}_N(y_i)) - I_G(\hat{F}_N))y_i$. As an example, take the Gini coefficient where $\hat{\hat{z}}_i = (2\hat{F}_N(y_i) - \hat{f}_N(y_i) - 1 - \hat{R}_N)y_i$.

In the Appendix the two variance estimators for the Gini coefficient together with the estimator based on the asymptotic variance (4.11) are compared in the srs case. In the illustrations given in the Appendix we have also included a jacknife estimator.

4.2.2 The Generalized Entropy Family

In the case of the Gini family the simplest variance estimators were obtained using method ii). However, in the case of the Generalized Entropy family the two methods will give identical estimators. The estimators are given in Table 4.3.

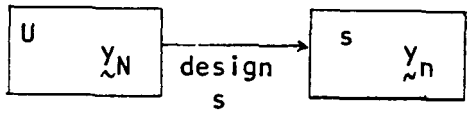
- - - - - - - - -
Table 4.3 in here
- - - - - - - - -

4.3 An auxiliary model approach

In the fix population approach the sample s was obtained according to a sampling design from the finite population U and the stochastic element in this procedure is the randomization of the sample $s \subset U$. Another way of interpreting a sample s from a finite population U is as follows: Assume the sample s to be fixed, i.e. the subset s of labels from U and the corresponding units in the finite population that is choosen to the sample is fixed. The vector of inclusion probabilities associated with the sample s and the design is considered as a vector of deterministic weights. We introduce an auxiliary model in such a way that the finite population vector $\underset{\sim}{y}_N = (y_1, y_2, \ldots, y_N)$ is regarded as selected from a set of population vectors $\underset{\sim}{Y}_N = (Y_1, Y_2, \ldots, Y_N)$, where $Y_1, Y_2, \ldots, Y_N$ are independent and identically distributed (IID) as Y with continuous cumulative df $F_Y(y)$. The two approaches are illustrated by Figure 4.1.

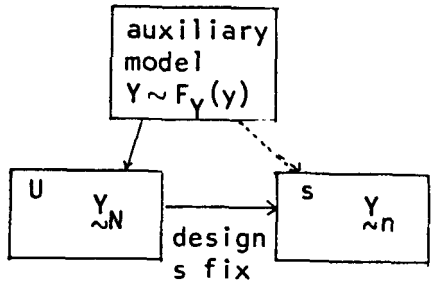
FIGURE 4.1 Illustration of the fix population approach (a) and the auxiliary model approach (b).

(a) the fix population approach



Stochastic element: the randomization of the sample s

(b) the auxiliary model approach



Stochastic element: the randomization of the finite population vector $\underset{\sim}{Y}_N$

The dotted arrow in Figure 4.1(b) illustrates the fact that, since the sample s is fixed, the randomization of the sampled vector $\underset{\sim}{Y}_n$, which is a subset of $\underset{\sim}{Y}_N$, can be considered as made directly from the auxiliary model to the sample.

Let $T(F)$, $T(F_N)$ and $T(F_n)$ be the model parameter, the finite population variable and the sample variable, respectively. In the fix population approach $T(F_N)$ was a parameter but under the auxiliary model it is a stochastic variate. It will be seen that we may obtain asymptotic results for a statistic on the form $\sqrt{n}(T(F_n) - T(F_N))$. Any confidence statement in this case is of Royall-type. cf. Royall (1971), i.e. for a given sample s the probability of coverage gives the probability that the interval includes the random variate $T(F_N)$ when the generating of Y-values from the model is 'repeated'. The obtained asymptotic results can also be used as bases for large sample inference in the fix population approach.

Consider now a sequence of populations $U_t = \{1,2,\ldots,N_t\}$ such that $N_t \to \infty$ as $t \to \infty$. For a fixed t we denote the sample by s_t with sample size n_t and assume that $n_t \to \infty$ so that the sampling fraction $f_t = n_t/N_t \to f$, $0 < f < 1$, as $t \to \infty$. When t increases we get new subsets of U_t such that s_t is not necessarily a subset of s_{t+1} (in other words we have a triangular array). In a similar way we denote the first order inclusion probability by π_{it} and the second order inclusion probability by π_{ijt}.

By Definition 4.1 we have an estimator of the finite population df F_N under the fix population approach. The next definition is the corresponding one under the auxiliary model approach, cf. Koul (1970) and Sandström (1983).

DEFINITION 4.3 Let $w_{it} \geq 0$ be bounded ($\forall t$) deterministic weights, $i \in U_t$, and $\bar{w}_t = n_t^{-1} \sum_{i \in s_t} w_{it} \neq 0$. The weighted empirical distribution function (wedf) is given by

$$F^*_{n_t}(y) = n_t^{-1} \sum_{i \in s_t} \frac{w_{it}}{\bar{w}_t} I_{\{Y_i \leq y\}}, \tag{4.8}$$

where $Y_1, Y_2, \ldots, Y_{n_t}$ are IID as Y with continuous cumulative df $F_Y(y)$ and $I_{\{Y_i \leq y\}}$ is an IID indicator function.

REMARK 4.7 If the weights are equal to some positive constant, then $F^*_{n_t}(y)$ coincide with the 'ordinary' empirical df and if $w_{it} = \pi_{it}^{-1}$, $\pi_{it} > 0$ and known inclusion probabilities, then (4.8) is similar to (4.3), the only difference is that in (4.8) s_t is fixed and Y_i is stochastic with the reversed relation in (4.3).

ASSUMPTION 4.1 The weights w_{it} are defined as above with $\bar{w}_t \neq 0$. We assume that

$$\max_{i \in s_t} \left(\frac{w_{it}}{\bar{w}_t}\right)^2 \leq d^2 < \infty, \quad \forall t. \tag{4.9}$$

REMARK 4.8 When the weights equal some positive constant then (4.9) is always fulfilled. This is the case of simple random sampling and proportional stratified random sampling designs ($w_{it} = \pi_{it}^{-1} = N/n$). With other designs, $w_{it} = \pi_{it}^{-1}$, the assumption states that $(n_t / \sum_{i \in s_t} \pi_{it}^{-1}) \cdot (\min_{i \in s_t} \pi_{it})^{-1}$ is bounded. The first factor is an estimate of the sample fraction $f_t = n_t/N_t$ which is assumed to converge towards a constant f, $0 < f < 1$, so the assumption mainly states that the design may not be such that $\min_{i \in s} \pi_{it} \to 0$ as $t \to \infty$.

Let v_t^2 be the squared coefficient of variation of the weights, i.e.

$$v_t^2 = s_{wt}^2 / \bar{w}_t^2 \text{ and } s_{wt}^2 = n_t^{-1} \sum_{i \in s_t} (w_{it} - \bar{w}_t)^2.$$

For the Gini family we have the following asymptotic result.

THEOREM 4.1 Let $I_G(F)$ be defined as in (2.4) and assume
$F \in \mathcal{F}$, $\mathcal{F} = \{F; |\int J(F(y))ydF(y)| < \infty\}$.
Assume that the function J is bounded and continuous. Then if $\sigma_G^2 > 0$ and under Assumption 4.1

$$\frac{n_t^{1/2} \{I_G(F^*_{n_t}) - I_G(F_{N_t})\}}{\{1 - f_t + v_t^2\}^{1/2}} \xrightarrow{\mathcal{L}} U \sim N(0, \mu^{-2}\sigma_G^2), \tag{4.10}$$

where $f_t = n_t/N_t$, v_t is the coefficient of variation of the weights and $T(F^*_{n_t})$ and $T(F_{N_t})$ are stochastic functionals defined as $T(F)$ with

F changed for $F^{*}_{n_t}$ and F_{N_t}, respectively. $\mu = T_{\mu}(F)$ is defined by (2.1). The asymptotic variance is

$$\sigma_G^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{\min(F_Y(y),F_Y(x)) - F_Y(y)F_Y(x)\} \cdot J_1(F_Y(y))J_1(F_Y(x))dydx,$$

where $J_1(F) = J(F)-I_G(F)$. (4.11)

Proof: See Theorem 5.2 in Sandström (1983).

REMARK 4.9 The asymptotic normality of the statistic on the left-hand side of (4.10) gives us a basis for confidence statements in the fix population approach.

REMARK 4.10 Under the auxiliary model approach the finite population correction (fpc) includes v_t^2, the squared coefficient of variation of the weights w_{it}.

For the parameters belonging to the Generalized Entropy fimily we can proceed as in Sandström (1983) using stochastic differentials of functionals to obtain asymptotic distributions. By this procedure and the use of Proposition 4.2 we will obtain the results stated in Theorem 4.2. We start with Proposition 4.2.

PROPOSITION 4.2 Let $g(\cdot)$ and $h(\cdot)$ be two real functions, both of bounded variation and assume $E|g(X)|$ and $E|h(X)|$ to exist and be bounded. If $X_1,X_2,\ldots,X_n$ are IID as X with continuous df $F_X(x)$ then

$$\mathrm{Cov}(h(X),g(X)) = \int_{-\infty}^{\infty}\int\{\min(F_X(x), F_X(y)) - F_X(x)F_X(y)\}\, dg(y)dh(x). \tag{4.12a}$$

If $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$ are IDD as (X,Y) with bivariate continuous cumulative df $F_{XY}(x,y)$ then

$$\mathrm{Cov}(h(X),g(Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{F_{XY}(x,y) - F_X(x)F_Y(y)\} \; dg(y)dh(x), \quad (4.12b)$$

where $F_X(x) = F_{XY}(x,\infty)$ and $F_Y(y) = F_{XY}(\infty,y)$.

Proof. The main assumption is that of bounded variation of the real functions $g(\cdot)$ and $h(\cdot)$. The proof is similar to that of Lehmann (1966) which he attributes to Hoeffding 1940.

Remark 4.11 If $h(\cdot) = g(\cdot)$ we get the variance $V(h(X))$ corresponding to (4.12a).

THEOREM 4.2 Let $I_{E,c}(F)$ be defined by (2.5a) when $c\neq 0,1$ and by (2.5b) when $c = 0,1$ and assume $F \in \mathcal{F}$, $\mathcal{F} = \{F; |I_{E,c}(F)| < \infty\}$. Assume $E|\log Y|^2, E|Y\log Y|^2$, and $E|Y^c|^2$ to exist and to be finite. Then under Assumption 4.1, provided that $0 < \sigma_c^2 < \infty$,

$$\frac{n_t^{1/2}\{I_{E,c}(F^*_{n_t}) - I_{E,c}(F_N)\}}{\{1 - f_t + v_t^2\}^{1/2}} \xrightarrow{\mathcal{L}} U \sim N(0,\sigma_c^2), \quad (4.13)$$

where f_t and v_t are defined as in Theorem 4.1 and σ_c^2 equals

$$c=0: \; \sigma_0^2 = V(\log Y) + \frac{1}{\mu_y^2} V(Y) - 2\frac{1}{\mu_y}\mathrm{Cov}(\log Y, Y), \quad (4.14a)$$

where $\mu_y = E(Y)$.

$$c=1:\ \sigma_1^2 = \frac{1}{\mu_y^2} V(Y\log Y) + \frac{(\mu_\ell+\mu_y)^2}{\mu_y^4} V(Y) - 2\frac{(\mu_\ell+\mu_y)}{\mu_y^3} Cov(Y,Y\log Y), \tag{4.14b}$$

where $\mu_\ell = E(Y\log Y)$.

$$c\neq 0,1:\ \sigma_c^2 = \frac{1}{c^2(c-1)^2\mu_y^{2c}} V(Y^c) + \left(\frac{\mu_c'}{(c-1)\mu_y^{c+1}}\right)^2 V(Y) - 2\frac{\mu_c'}{c(c-1)^2\mu_y^{2c+1}} Cov(Y^c,Y), \tag{4.14c}$$

where $\mu_c' = E(Y^c)$.

Proof. We only give a sketch of the proof for the case $c=0$. In this case the functional equals $T(F) = I_{E,0}(F) = \log T_2(F) - T_1(F)$, where $T_1(F) = \int_0^\infty \log y dF(y)$ and $T_2(F) = \int_0^\infty y dF(y)$. The stochastic differential equals

$$T_F'(F_{n_t}^* - F) = \int_0^\infty (F_{n_t}^*(y) - F(y))d\log y - \frac{1}{\mu_y}\int_0^\infty (F_{n_t}^*(y) - F(y))dy.$$

Let the remainder term be $R_{1n_t}^* = T(F_{n_t}^*) - T(F) - T_F'(F_{n_t}^* - F)$. If we can show that $(n_t/c_t)^{1/2} R_{1n_t}^* \xrightarrow{P} 0$, where $c_t = 1+v_t^2$, then the asymptotic distribution of $(n_t/c_t)^{1/2}(T(F_{n_t}^*) - T(F))$ is equivalent to, if any, that of $(n_t/c_t)^{1/2} T_F'(F_{n_t}^* - F)$. One way to show this is to show that

$$n_t^{1/2} \sup_{0\leq\lambda\leq 1} \left| \frac{d^2 T(F_\lambda)}{d\lambda^2} \right| \xrightarrow{P} 0, \tag{4.15}$$

where $F_\lambda = F + \lambda(F_{n_t}^* - F)$, see Serfling (1980 -p.216). Let q be a positive bounded function on $[0,1]$ such that $q(t)=q(1-t)$, $0 \leq t \leq 1/2$ and increasing in t and $t(1-t)q^{-4}(t)$ is an integrable

function. Let $\int_0^\infty q(F(y))dy < \infty$. Define a q-norm on (0,1) as $||G-F||_{q(F)} = \sup_y |(G(y)-F(y))/q(F(y))|$. Then by Lemma 5.3 in Sandström (1983) it follows that $(n_t/c_t)^{1/2}||F^*_{n_t}-F||_{q(F)}$ is stochastically bounded. By this result it is easily shown that (4.15) is fulfilled. By the same argument as in Sandström (1983), i.e. by use of Assumption 4.1 and the Central Limit Theorem for triangular arrays, cf. Lehmann (1976 - p.352), it is readily seen that

$$(n_t/c_t)^{1/2}T'_F(F^*_{n_t}-F) \xrightarrow{\mathcal{L}} U \sim N(0,\sigma_0^2),$$

where, by use of Proposition 4.2, σ_0^2 equals (4.14a). It is now easily checked that (4.13) holds for c=0, cf. Theorem 5.2 in Sandström (1983). The cases c=1 and C≠0,1 follows in similar ways.

APPENDIX

COMPARISON OF VARIANCE ESTIMATORS FOR THE GINI COEFFICIENT

We will compare the two variance esitmators for the Gini coefficient given in Section 4, where the first is based on method i) ("Taylor estimator" in Table 4.2) and the second on method ii) ("Ratio estimator", formula (4.7)). To simplify the comparison we only consider simple random sampling (srs), without replacement. A third variance estimator can be obtained from the asymptotic variance (4.11). A consistent estimator of (4.11) is given in Sandström (1983). Explicit formulas for the three variance estimators are given in Table A.1.

When N and n are large ($n \to \infty$, $N \to \infty$, $f_n = n/N \to f$, $0 < f < 1$) the Taylor estimator and the formula based on the asymptotic variance estimator are identical. The three variance formulas in this case are given in Table A.2.

REMARK A.1 The variance estimator given by Glasser (1962) is similar to our method i).

To illustrate[1] the behaviour of the three variance estimators we will compare their sampling distributions from srs without replacement from two small parent distributions, both of size N = 11. Population 1 (P1) is a "symmetric" population and population 2 (P2) is a "skew". The sample sizes are in both cases n = 5, i.e. the total number of samples is $\binom{11}{5} = 462$. The population values are

[1] We wish to thank Mr. Bertil Waldén for the computations which were performed on the IBM 370 at Statistics Sweden.

P1: 20,40,45,47,49,50,51,53,55,60,80 and

P2: 20,21,22,23,24,25,30,40,50,60,80.

The population distributions are depicted in Figure A.1a). The arithmetic means and Gini coefficients are $\bar{y}_{N1} = 50$, $\bar{y}_{N2} = 395/11 \doteq 35.91$ and $R_{N1} = 424/3025 \doteq 0.1402$, $R_{N2} = 232/869 \doteq 0.2670$, respectively.

REMARK A.2 If we had not used the Gini coefficient with its Gfpc-term, cf. Remark 3.3, the inequality parameters would had been $R_{N1} \doteq 0.2311$ and $R_{N2} \doteq 0.3579$, respectively.

In figure A1 b) the sampling distributions of the 462 possible estimators of R_N are plotted. As one can see the sampling distribution from P1 is more symmetric than that from P2, cf. Table A.3. The relative Bias of $\hat{R}_N$ as an estimator of R_N is for P1 -11.2% and for P2 -15.0%. This bias can be decreased through an expectation-correction, viz. by defining $R_{Nec} = (N/N-1)R_N$ and $\hat{R}_{Nec} = (n/n-1)\hat{R}_N$. These correted definitions are often found in the literature. The relative bias of $\hat{R}_{Nec}$ is for P1 +0.9% and for P2 -3.0%.

In addition to the three variance estimators discussed above we have for illustrative purposes included the sampling distribution of a jacknife estimator, computed as $((n-1)/n) \sum_{i=1}^{n} (\hat{R}_N^i - \bar{\hat{R}}_N^{\cdot})^2$, where $\hat{R}_N^i$ is an estimator of the Gini coefficient excluding the ith observation and $\bar{\hat{R}}_N^{\cdot} = n^{-1} \sum_{i=1}^{n} \hat{R}_N^i$.

For both populations, the "Ratio estimator" of the variance overestimates the true value by a factor 12-14 while the "Taylor estimator" slightly

overestimates the variance in the symmetric case and underestimates it in the skew case. The Asymptotic estimator underestimates the variance while the Jacknife overestimates it.

When both the sample size and the population size increases one could guess that the differences in the variance estimators would decreas and that the sampling distribution of $\hat{R}_N$ will be more symmetrical. More work will be done in investigating the sampling properties of the variance estimators.

- - - - - - - - - - - - - - - - - -

Table A.1 in here

Table A.2 in here

Figure A.1 in here

Table A.3 in here

Figyre A.2 in here

Table A.4 in here

- - - - - - - - - - - - - - - - - -

REFERENCES:

AGHEVLI, B.B. and MEHRAN, F. (1981). Optimal Grouping of Income Distribution Data. Journal of the American Statistical Association, 76, 22-26.

ATKINSON, A.B. (1970). On the Measurement of Inequality. Journal of Economic Theory, 2, 244-263.

BEACH, C.M. and DAVIDSON, R. (1983). Distribution-Free Statistical Inference with Lorenz Curves and Income Shares. Review of Economic Studies, 50, 723-735.

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.

COCHRAN, W.G. (1977). Sampling Techniques. 3rd ed., John Wiley & Sons, New York.

COWELL, F.A. (1980). On the Structure of Additive Inequality Measures. Review of Economic Studies, 47, 521-531.

COWELL, F.A. and MEHTA, F. (1982). The Estimation and Interpolation of Inequality Measures. Review of Economic Studies, 49, 273-290.

DAGUM, C. (1983). Income Distribution Models. Entry in Encyclopedia of Statistical Sciences (eds.: Kotz, Johnson, Read), 4, 27-34, John Wiley & Sons, New York.

GASTWIRTH, J.L. (1972). The Estimation of the Lorenz Curve and Gini Index. Review of Economics and Statistics, 54, 306-316.

_ (1975). The Estimation of a Family of Inequality Measures. Journal of Econometrics, 3, 61-70.

GASTWIRTH, J.L: and GLAUBERMAN, M. (1976). The Interpolation of the Lorenz Curve and the Gini Index from Grouped Data. Econometrica, 44, 479-483.

GLASSER, G.J. (1962). Variance Formulas for the Mean Difference and Coefficient of Concentration. Journal of the American Statistical Association, 57, 648-654.

KAKWANI, N.C. (1980). Income Inequality and Poverty, Methods of Estimation and Policy Applications. Oxford University Press, New York.

KOUL, H.L. (1970). Some Convergence Theorems for Ranks and Weighted Empirical Cumulatives. The Annals of Mathematical Statistics, 41, 1768-1773.

LEHMANN, E.L. (1966). Some Concepts of Dependence. The Annals of Mathematical Statistics, 37, 1137-1153.

LEHMANN, E.L. (1976). Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco.

MACDONALD, J. (1984). Some Generalized Functions for the Size Distribution of Income. Econometrica, 52, 647-663.

MEHRAN, F. (1975). Bounds on the Gini Index Based on Observed Points of the Lorenz Curve. Journal of the American Statistical Association, 70, 64-66.

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

PIESCH, W. (1975). Statistische Konzentrationsmasse. J.C.B. Mohr (Paul Siebeck), Tübingen.

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

SANDSTRÖM, A. (1983). Estimating Income Inequality, Large Sample Inference in Finite Populations. Research Report 1983:5, Department of Statistics, University of Stockholm.

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

SHORROCKS, A.F. (1980). The Class of Additively Decomposable Inequality Measures. Econometrica, 48, 613-625.

- (1982). Inequality Decomposition by Factor Components. Econometrica, 50, 193-211.

- (1983). Inequality Decomposition by Population Subgroups. Discussion paper, University of Essex (forthcoming in Econometrica).

Table 2.1 Some inequality parameters belonging to the Gini family and to the Generalized Entropy family.

1. THE GINI FAMILY	
Weight function, J(u)	Name[1]
$2u - 1$	R, the Gini coefficient
$1 - 3(1 - u)^2$	M, Mehran's measure
$\frac{1}{2}(3u^2 - 1)$	P, Piesch's measure[2]

2. THE GENERALIZED ENTROPY FAMILY[3]	
c	Name
0	E_0, Theil's 2nd measure
1	E_1, Theil's 1st measure
2	E_2[4] $= V^2/2$, V is the coefficient of variation

NOTES: 1) The following relation holds between the parameters in the Gini family: $M = 3R - 2P$

2) This parameter belongs to a general class defined by Piesch (1975 - p.131)

3) The generalized entropy family is related to Atkinson's family of measures, see Atkinson (1970).

4) E_2 is also labelled Hirschman's index.

TABLE 2.2 Reformulations of some inequality parameters of income inequality.

THE GINI FAMILY		
Parameter , J(F)	according to (2.6)	according to (2.7a)[1]
Gini, 2F-1	$\frac{1}{\mu}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\lvert F_X(x)-F_X(y)\rvert\cdot\lvert x-y\rvert dF_X(y)dF_X(x)$	$\frac{1}{2\mu}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\lvert x-y\rvert dF_X(x)dF_X(y)$
Mehran, $1-3(1-F)^2$	$\frac{3}{2\mu}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\lvert 2(F_X(x)-F_X(y))-(F_X^2(x)-F_X^2(y))\rvert\cdot\lvert x-y\rvert dF_X(y)dF_X(x)$	$\frac{1}{\mu}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\frac{5}{4}-F_X(x))\lvert x-y\rvert dF_X(y)dF_X(x)$
Piesch, $\frac{3}{2}F^2-\frac{1}{2}$	$\frac{3}{2\mu}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\lvert F_X^2(x)-F_X^2(y)\rvert\cdot\lvert x-y\rvert dF_X(y)dF_X(x)$	$\frac{1}{\mu}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\frac{1}{2}F_X(x)+\frac{1}{8})\lvert x-y\rvert dF_X(y)dF_X(x)$
THE GENERALIZED ENTROPY FAMILY		
Parameter	according to (2.6)	
E_0	- -	
E_1	$\frac{1}{2\mu}\int_0^{\infty}\int_0^{\infty}\lvert\log\frac{x}{y}\rvert\times\lvert x-y\rvert dF_X(y)dF_X(x)+T_2$	
E_c, $c\neq 0,1$	$\frac{1}{c(c-1)}\sum_{i=1}^{c}\frac{1}{\mu^{c+1-i}}\int_0^{\infty}\int_0^{\infty}\lvert x^{c-i}-y^{c-i}\rvert\times\lvert x-y\rvert dF_X(y)dF_X(x)$	

Notes: 1) The relation between Gini's mean difference and the Gini coefficient is obvious. All parameters belonging to the Gini family are weighted Gini's mean differences.

Table 3.1 Expressions for some parameters of the Gini family corrected for finite populations, cf. Remark 3.3.

PARAMETER	FORMULA	RANGE (NON-NEGATIVE DATA)
Gini, R_N	$\frac{2}{N^2\bar{y}_N} \sum_{i=1}^{N} iy_{(i)} - 1 - \frac{1}{N}$	$[0, 1 - \frac{1}{N}]$
Mehran, M_N	$\frac{6}{N^2\bar{y}_N}(1 + \frac{1}{2N}) \sum_{i=1}^{N} iy_{(i)} - \frac{3}{N^3\bar{y}_N} \sum_{i=1}^{N} i^2y_{(i)} - \frac{(N+1)(2N+1)}{N^2}$	$[0, (1-\frac{1}{N})(1 + \frac{1}{N})]$
Piesch, P_N	$\frac{3}{2N^3\bar{y}_N} \sum_{i=1}^{N} i^2y_{(i)} - \frac{3}{2N^3\bar{y}_N} \sum iy_{(i)} - \frac{(N-1)(N+1)}{2N^2}$	$[0, (1-\frac{1}{N})(1-\frac{1}{2N})]$

Table 3.2 Expressions for the parameters of the Generalized Entropy family.

PARAMETER	FORMULA	RANGE (POSITIVE DATA)
E_{0N}	$-\frac{1}{N}\sum_{i=1}^{N}\log(\frac{y_i}{\bar{y}_N})$	if $y \in [0,\infty[$ then $E_{0N} \in [0,\infty[$
E_{1N}	$\frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\bar{y}_N}\log(\frac{y_i}{\bar{y}_N})$	if $y \in [0,\infty[$ then $E_{1N} \in [0,\log N]$
E_{cN} , $c\neq 0,1$	$\frac{1}{c(c-1)}\cdot\frac{1}{N}\sum_{i=1}^{N}\{(\frac{y_i}{\bar{y}_N})^c-1\}$	if $y \in [0,\infty[$ then $E_{cN} \in [0,\frac{N^{c-1}-1}{c(c-1)}]$

Table 3.3. Lower bounds and the maximum bias[1)] of these bounds for parameters of the Gini and the Generalized Entropy family when calculated from grouped data.

Parameter	Lower bound [2)]
GINI FAMILY	
R_N	$\frac{1}{N^2\bar{y}_N} \sum_{i=1}^{k} N_i(2Q_i + N_i)\bar{y}_i - 1$
M_N	$\frac{1}{N^3\bar{y}_N} \sum_{i=1}^{k} N_i\{3N(2Q_i+N_i) - 3Q_i(Q_i+N_i) - N_i^2\}\bar{y}_i - 2$
P_N	$\frac{1}{2N^3\bar{y}_N} \sum_{i=1}^{k} N_i\{3Q_i(Q_i+N_i) + N_i^2\}\bar{y}_i - \frac{1}{2}$
GENERALIZED ENTROPY FAMILY	
E_{0N}	$\frac{1}{N} \sum_{i=1}^{k} N_i \log\{\frac{\bar{y}_N}{\bar{y}_i}\}$
E_{1N}	$\frac{1}{N} \sum_{i=1}^{k} N_i \frac{\bar{y}_i}{\bar{y}_N} \log\{\frac{\bar{y}_i}{\bar{y}_N}\}$
E_{cN} $c\neq 0,1$	$\frac{1}{c(c-1)} \frac{1}{N} \sum_{i=1}^{k} N_i \{(\frac{\bar{y}_i}{\bar{y}_N})^c - 1\}$

Table 3.3. (cont.)

Parameter	Maximum bias
GINI FAMILY	
R_N	$\frac{1}{N^2\bar{y}_N} \sum_{i=1}^{k} N_i^2 \lambda_i (1-\lambda_i)(a_i - a_{i-1})$
M_N	$\frac{1}{N^3\bar{y}_N} \sum_{i=1}^{k} N_i^3 \lambda_i (1-\lambda_i)(\frac{3N}{N_i} - 2 + \lambda_i)(a_i - a_{i-1})$
P_N	$\frac{1}{2N^3\bar{y}_N} \sum_{i=1}^{k} N_i^3 \lambda_i (1-\lambda_i)(2-\lambda_i)(a_i - a_{i-1})$
GENERALIZED ENTROPY FAMILY	
E_{0N}	$\frac{1}{N} \sum_{i=1}^{k} N_i (\log\bar{y}_i - (1-\lambda_i)\log a_{i-1} - \lambda_i \log a_i)$
E_{1N}	$\frac{1}{N} \sum_{i=1}^{k} N_i ((1-\lambda_i) a_{i-1} \log a_{i-1} + \lambda_i a_i \log a_i - \bar{y}_i \log\bar{y}_i)$
E_{cN} $c \neq 0,1$	$\frac{1}{c(c-1)} \frac{1}{N} \sum_{i=1}^{k} N_i ((1-\lambda_i)(\frac{a_{i-1}}{\bar{y}_i})^c + \lambda_i (\frac{a_i}{\bar{y}_i})^c - (\frac{y_i}{\bar{y}_N})^c)$

1) The upper bound is obtained by addition of the maximum bias to the lower bound.

2) Q_i is defined through $Q_i = \sum_{j=1}^{i-1} N_j$.

Table 4.1 Point estimators of the finite population inequality parameters under the fix population approach.

FAMILY	PARAMETER	ESTIMATOR
GINI	Gini coefficient, R_N	$\hat{R}_N = \dfrac{2\sum_{i\in s}(P_{s(i)}+\frac{1}{2\pi_i})\frac{y_i}{\pi_i}}{P_n\sum_{i\in s} y_i/\pi_i} - 1$
	Mehran's measure, M_N	$\hat{M}_N = \dfrac{6\sum_{i\in s}(P_{s(i)}+\frac{1}{2\pi_i})y_i/\pi_i}{P_n\sum_{i\in s} y_i/\pi_i} - \dfrac{3\sum_{i\in s}(P^2_{s(i)}+P_{s(i)}\cdot\frac{1}{\pi_i}+\frac{1}{3\pi_i^2})\frac{y_i}{\pi_i}}{P_n^2\sum_{i\in s} y_i/\pi_i} - 2$
	Piesch's measure, P_N	$\hat{P}_N = \dfrac{3\sum_{i\in s}(P^2_{s(i)}+P_{s(i)}\frac{1}{\pi_i}+\frac{1}{3\pi_i^2})y_i/\pi_i}{2P_n^2\sum_{i\in s} y_i/\pi_i} - \frac{1}{2}$
		$P_{s(i)} = \sum_{j\in s} I_{\{y_j < y_i\}}/\pi_j$ $P_n = \sum_{j\in s} 1/\pi_j = \hat{N}_s$
GENERALIZED ENTROPY	E_{0N}	$\hat{E}_{0N} = -\log\hat{N}_s + \log\{\sum_{i\in s} y_i/\pi_i\} - \hat{N}_s^{-1}\sum_{i\in s}\pi_i^{-1}\cdot\log y_i$
	E_{1N}	$\hat{E}_{1N} = \log\hat{N}_s - \log\{\sum_{i\in s} y_i/\pi_i\} + \dfrac{\sum_{i\in s}\frac{y_i}{\pi_i}\log y_i}{\sum_{i\in s} y_i/\pi_i}$
	E_{cN}, $c\neq 0,1$	$\hat{E}_{cN} = \dfrac{\hat{N}_s^{c-1}\sum_{i\in s} y_i^c/\pi_i}{c(c-1)(\sum_{i\in s} y_i/\pi_i)^c} - \dfrac{1}{c(c-1)}$

Table 4.2 Variance estimator of the estimator of the Gini coefficient using the linear terms in the Taylor expansion (method i).

$$\hat{V}(\hat{R}_N) = \frac{4}{\hat{N}^2 \hat{t}_y^2} \hat{V}(\hat{t}_{wy}) + \frac{4\hat{t}_{wy}^2}{\hat{N}^2 \hat{t}_y^4} \hat{V}(\hat{t}_y) + \frac{4\hat{t}_{wy}^2}{\hat{N}^4 \hat{t}_y^2} \hat{V}(\hat{N}) - \frac{8\hat{t}_{wy}}{\hat{N}^2 \hat{t}_y^3} \hat{Cov}(\hat{t}_{wy}, \hat{t}_y) -$$

$$- \frac{8\hat{t}_{wy}}{\hat{N}^3 \hat{t}_y^2} \hat{Cov}(\hat{t}_{wy}, \hat{N}) + \frac{8\hat{t}_{wy}^2}{\hat{N}^3 \hat{t}_y^3} \hat{Cov}(\hat{t}_y, \hat{N}),$$

where $$\hat{V}(\hat{t}_{wy}) = \frac{1}{8} \sum_{i,j \in s} \sum \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i^2} - \frac{y_j}{\pi_j^2} \right)^2 + \sum\sum_{i \neq j \in s} (1-\pi_{ij}) \frac{I_{[y_j<y_i]}}{\pi_j^2} \cdot \frac{y_i^2}{\pi_j^2} +$$

$$+ \sum\sum_{i \neq j \in s} (1-\pi_i) \frac{I_{[y_j<y_i]}}{\pi_j} \cdot \frac{y_i^2}{\pi_i^3} +$$

$$+ \sum\sum_{i \neq j \in s} (1-\pi_j) \frac{I_{[y_j<y_i]}}{\pi_j} \cdot \frac{y_i}{\pi_i} \cdot \frac{y_j}{\pi_j^2} +$$

$$+ \sum\sum\sum_{i \neq j \neq k \in s} \frac{\pi_{ijk} - \pi_{ij}\pi_{ki}}{\pi_{ijk}} \cdot \frac{I_{[y_j<y_i]}}{\pi_j} \cdot \frac{I_{[y_i<y_k]}}{\pi_i} \cdot \frac{y_i}{\pi_i} \cdot \frac{y_k}{\pi_k} +$$

$$+ \sum\sum\sum_{i \neq j \neq k \in s} \frac{\pi_{ijk} - \pi_{ij}\pi_{kj}}{\pi_{ijk}} \cdot \frac{I_{[y_j<y_i]}}{\pi_j} \cdot \frac{I_{[y_j<y_k]}}{\pi_j} \cdot \frac{y_i}{\pi_i} \cdot \frac{y_k}{\pi_k} +$$

$$+ \sum\sum\sum_{i \neq j \neq k \in s} \frac{\pi_{ijk} - \pi_{ij}\pi_{ik}}{\pi_{ijk}} \cdot \frac{I_{[y_j<y_i]}}{\pi_j} \cdot \frac{I_{[y_k<y_i]}}{\pi_k} \cdot \frac{y_i^2}{\pi_i^2} +$$

$$+ \sum\sum\sum_{i \neq j \neq k \in s} \frac{\pi_{ijk} - \pi_{ij}\pi_{jk}}{\pi_{ijk}} \cdot \frac{I_{[y_j<y_i]}}{\pi_j} \cdot \frac{I_{[y_k<y_j]}}{\pi_k} \cdot \frac{y_i}{\pi_i} \cdot \frac{y_j}{\pi_j} +$$

Table 4.2, cont

$$+ \sum_{i\neq j\neq k\in s}\sum\sum \frac{\pi_{ijk}-\pi_{ij}\pi_k}{\pi_{ijk}} \cdot \frac{I_{[y_j<y_i]}}{\pi_j} \cdot \frac{y_i}{\pi_i} \cdot \frac{y_k}{\pi_k^2} +$$

$$+ \sum_{i\neq j\neq k\neq \ell\in s}\sum\sum\sum \frac{\pi_{ijk\ell}-\pi_{ij}\pi_{k\ell}}{\pi_{ijk\ell}} \cdot \frac{I_{[y_j<y_i]}}{\pi_j} \cdot \frac{I_{[y_\ell<y_k]}}{\pi_\ell} \cdot \frac{y_i}{\pi_i} \cdot \frac{y_k}{\pi_k} \quad ,$$

$$\hat{V}(\hat{t}_y) = \frac{1}{2} \sum_{i,j\in s}\sum \frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}} \cdot \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

$$\hat{V}(\hat{N}) = \frac{1}{2} \sum_{i,j\in s}\sum \frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}} \cdot \left(\frac{1}{\pi_i} - \frac{1}{\pi_j}\right)^2$$

$$\hat{Cov}(\hat{t}_{wy},\hat{t}_y) = \frac{1}{2} \sum_{i\in s} (1-\pi_i)\frac{y_i^2}{\pi_i^3} + \frac{1}{2}\sum_{i\neq j\in s}\sum \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}} \cdot \frac{y_i}{\pi_i^2} \cdot \frac{y_j}{\pi_j} +$$

$$+ \sum_{i\neq j\in s}\sum (1-\pi_i) \frac{I_{[y_j<y_i]}}{\pi_j} \frac{y_i^2}{\pi_i^2} + \sum_{i\neq j\in s}\sum (1-\pi_j) \cdot \frac{I_{[y_j<y_i]}}{\pi_j} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} +$$

$$+ \sum_{i\neq j\neq k\in s}\sum\sum \frac{(\pi_{ijk}-\pi_{ij}\pi_k)}{\pi_{ijk}} \frac{I_{[y_j<y_i]}}{\pi_j} \frac{y_i}{\pi_i} \frac{y_k}{\pi_k}$$

Table 4.2, cont.

$$\hat{\mathrm{Cov}}\,(\hat{t}_{wy},\hat{N}) = \frac{1}{4}\sum_{i,j\in s}\frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}}\left(\frac{y_i}{\pi_i^2}-\frac{y_j}{\pi_j^2}\right)\left(\frac{1}{\pi_i}-\frac{1}{\pi_j}\right) +$$

$$+\sum_{i,j\in s}(1-\pi_i)\,\frac{I_{[y_j<y_i]}}{\pi_j}\,\frac{y_i}{\pi_i}\left(\frac{1}{\pi_i}-\frac{1}{\pi_j}\right) +$$

$$+\sum_{i\neq j\neq k\in s}\frac{(\pi_k\pi_{ij}-\pi_{ijk})}{\pi_{ijk}}\,\frac{I_{[y_j<y_i]}}{\pi_j}\,\frac{y_i}{\pi_i}\left(\frac{1}{\pi_j}-\frac{1}{\pi_k}\right)$$

$$\hat{\mathrm{Cov}}(\hat{t}_y,\hat{N}) = \frac{1}{2}\sum_{i,j\in s}\frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)\left(\frac{1}{\pi_i}-\frac{1}{\pi_j}\right),\ \text{cf. Remark 4.5}$$

Table 4.3. Variance estimators of the estimated members of the Generalized Entropy family using the linear terms in the Taylor expansion

$$\hat{V}(\hat{E}_{0N}) = \frac{1}{\hat{t}_y^2}\hat{V}(\hat{t}_y) + \frac{1}{N^2}\hat{V}(\hat{t}_z) + \left(\frac{1}{N} - \frac{\hat{t}_z}{N^2}\right)^2 \hat{V}(\hat{N}) - \frac{2}{N\hat{t}_y}\hat{Cov}(\hat{t}_y,\hat{t}_z) -$$

$$- 2\left(\frac{1}{N} - \frac{\hat{t}_z}{N^2}\right)\frac{1}{\hat{t}_y}\hat{Cov}(\hat{t}_y,\hat{N}) + 2\frac{1}{N}\left(\frac{1}{N} - \frac{\hat{t}_z}{N^2}\right)\hat{Cov}(\hat{t}_z,\hat{N})$$

$$\hat{V}(\hat{E}_{1N}) = \left(\frac{\hat{t}_v - \hat{t}_y}{\hat{t}_y^2}\right)^2 \hat{V}(\hat{t}_y) + \frac{1}{\hat{t}_y^2}\hat{V}(\hat{t}_v) + \frac{1}{N^2}\hat{V}(\hat{N}) -$$

$$- 2\left(\frac{\hat{t}_v - \hat{t}_y}{\hat{t}_y^2}\right)\frac{1}{\hat{t}_y}\hat{Cov}(\hat{t}_y,\hat{t}_v) - 2\left(\frac{\hat{t}_v - \hat{t}_y}{\hat{t}_y^2}\right)\frac{1}{N}\hat{Cov}(\hat{t}_y,\hat{N}) + 2\frac{1}{\hat{t}_y}\frac{1}{N}\hat{Cov}(\hat{t}_v,\hat{N})$$

$$\hat{V}(\hat{E}_{cN}) \quad c \neq 0,1 \qquad = \frac{\hat{N}^{2(c-1)}\hat{t}_u^2}{(c-1)^2\hat{t}^{2(c+1)}}\hat{V}(\hat{t}_y) + \frac{\hat{N}^{2(c-1)}}{c^2(c-1)^2\hat{t}_y^{2c}}\hat{V}(\hat{t}_u) + \frac{\hat{N}^{2(c-2)}\hat{t}_u^2}{c^2\,\hat{t}_y^{2c}}\hat{V}(\hat{N}) -$$

$$- 2\frac{\hat{N}^{2(c-1)}\hat{t}_u}{c(c-1)^2\hat{t}_y^{2c+1}}\hat{Cov}(\hat{t}_y,\hat{t}_u) - 2\frac{\hat{N}^{2c-3}\hat{t}_u^2}{c(c-1)\hat{t}_y^{2c+1}}\hat{Cov}(\hat{t}_y,\hat{N}) +$$

$$+ 2\frac{\hat{N}^{2c-3}\hat{t}_u}{c^2(c-1)\hat{t}_y^{2c}}\hat{Cov}(\hat{t}_u,\hat{N})$$

where $\hat{V}(\hat{t}_y) = \frac{1}{2}\sum\sum_{i,j\in s}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$

and $\hat{V}(\hat{t}_z)$, $\hat{V}(\hat{t}_v)$ and $\hat{V}(\hat{t}_u)$ are obtained by changing y_i for $z_i = \log y_i$, $v_i = y_i \log y_i$ and $u_i = y_i^c$, respectively. The covariance estimators are given by (4.5) and Remark 4.5.

Table A.1 Variance estimators for the Gini coefficient.

$$s_y^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y}_n)^2 , \qquad \bar{y}_n = n^{-1} \sum_{i \in s} y_i$$

$$\hat{R}_y = \frac{2}{n^2 \bar{y}_n} \sum_{i \in s} i y_{(i)} - 1 , \qquad \hat{R}_z = \frac{2}{n^2 \bar{z}_n} \sum_{i \in s} i z_{(i)} - 1, \qquad z_i = y_i^2$$

$$\hat{B}_y = \frac{2}{n^2 \bar{y}_n} \sum_{i \in s} i y_{(i)} - 1 - \frac{1}{n} , \qquad \bar{z}_n = n^{-1} \sum_{i \in s} y_i^2$$

$$T_i = \sum_{j=1}^{i} j y_{(j)} , \qquad T_0 \equiv 0$$

$$f_n = n/N$$

$$A_n = \frac{1}{n^3} \sum_{i \in s} i^2 y_{(i)}^2 \qquad B_n = \frac{1}{n^3} \sum_{i \in s} T_{i-1} y_{(i)}$$

1. "Ratio estimator" (from (4.7))

$$n \hat{\sigma}^2_{\hat{R},r} = \frac{(1-f_n)}{\bar{y}_n^2} \{4(\frac{n}{n-1}) A_n - (s_y^2 + \bar{y}_n^2(\frac{n}{n-1}))[2(\hat{R}_y+1)(\hat{R}_z+1)-(\hat{R}_y+1)^2]\}$$

2. "Taylor estimator" (from Table 4.2)

$$c_1 = \frac{(N-2)n(n-1)}{(N-1)(n-2)(n-3)} \qquad c_2 = \frac{4Nn^2-6(N+n)n+6n}{(N-1)(n-2)(n-3)}$$

$$c_3 = 2 - 3c_1 + \frac{c_2}{n} + \frac{2}{n-2} + \frac{n}{N-1} \qquad c_4 = \frac{5}{2} - 4c_1 + \frac{c_2}{n} + \frac{3}{n-2} + \frac{n}{N-1}$$

$$n \hat{\sigma}^2_{\hat{R},t} = \frac{(1-f_n)}{\bar{y}_n^2} \{4c_1 A_n + 16 c_1 B_n +$$

$$+ s_y^2 [(\hat{B}_y+1)^2 - 2(\frac{n-1}{n-2})(\hat{B}_y+1)(\hat{R}_z+1) + 2\frac{(2n-1)}{n(n-2)}(\hat{B}_y+1) + \frac{1}{n^2} + 2\frac{(n-1)}{n^2} c_3(\hat{R}_z+1) - 4\frac{(n-1)}{n^3} c_4] +$$

$$\bar{y}_n^2 [\frac{4n}{(n-2)}(\hat{B}_y-1)(\hat{R}_y-1) - c_2(\hat{R}_y+1)^2 - 2(\frac{n}{n-2})(\hat{B}_y+1)(\hat{R}_z+1) - 2(\frac{n+1}{n-2})(\hat{B}_y+1) - 4\frac{c_2}{n^2} +$$

$$+ 4\frac{c_2}{n}(\hat{R}_y+1) - \frac{4}{n-2}(\hat{R}_y+1) + \frac{2}{n}(\frac{n+4}{n-2}) - 8\frac{c_1}{n} + 2\frac{1}{n} c_3(\hat{R}_z+1) - 4\frac{1}{n^2} c_4]\}$$

Table A.1 Cont.

3. "Asymptotic estimator" (from (4.11))

$$n\hat{\sigma}^2_{R,a} = \frac{(1-f_n)}{\bar{y}_n^2}\{4A_n + 16\,B_n +$$

$$+ s_y^2\left[\left(\frac{n-1}{n}\right)(\hat{R}_y+1)^2 - 2\left(\frac{n-1}{n}\right)(\hat{R}_y+1)(\hat{R}_z+1) + 2\left(\frac{n-1}{n^2}\right)(\hat{R}_y+1) + 2\left(\frac{n-1}{n^2}\right)(\hat{R}_z+1)\right] -$$

$$- \bar{y}_n^2\left[2(\hat{R}_y+1)(\hat{R}_z+1) + 2\left(\frac{n-3}{n}\right)(\hat{R}_y+1) + 4\,\frac{(n+1)}{n^2} - 2\frac{1}{n}\,(\hat{R}_z+1)\right]\}$$

Table A.2 Approximated variance estimators for the Gini coefficient when N and n are large.

Notation: See Table A.1.

$n \to \infty$, $N \to \infty$, $\hat{B}_y = \hat{R}_y$

$$C_n = 4A_n - (s_y^2 + \bar{y}_n^2)\,[2(\hat{R}_y+1)(\hat{R}_z+1) - (\hat{R}_y+1)^2]$$

$$D_n = 16B_n - \bar{y}_n^2\,[(\hat{R}_y+1)^2 + 2(\hat{R}_y+1)]$$

1. "Ratio estimator"

$$n\hat{\sigma}^2_{\hat{R},r} = \frac{(1-f_n)}{\bar{y}_n^2}\, C_n$$

2. "Taylor estimator"

$$n\hat{\sigma}^2_{\hat{R},t} = \frac{(1-f_n)}{\bar{y}_n^2}\,\{C_n + D_n\}$$

3. "Asymptotic estimator"

$$n\hat{\sigma}^2_{\hat{R},a} = \frac{(1-f_n)}{\bar{y}_n^2}\,\{C_n + D_n\}$$

TABLE A. 3 Properties of the two sampling distributions of $\hat{R}_N$ from P1 and P2, respectively. In both cases N = 11 and n = 5, i.e. both sampling distributions consists of 462 possible values $\hat{R}_N$.

	P1	P2
R_N	0.1402	0.2670
Sampling distribution		
$\bar{\hat{R}}_N$	0.1245	0.2270
Bias	-0.0157	-0.0399
$Var(\hat{R}_N)$	0.002931	0.003649
$\gamma_1(\hat{R}_N)$ 1)	0.0927	-0.9130
$\gamma_2(\hat{R}_N)$ 2)	-0.8745	0.4994
$CV(\hat{R}_N)$ 3)	0.4349	0.2661
Max $\hat{R}_N$	0.2286	0.3133
Min $\hat{R}_N$	0.0224	0.0348
Range	0.2062	0.2785

Notes: 1) $\gamma_1(\hat{R}_N) = \sum_{i=1}^{462} (\hat{R}_i - \bar{\hat{R}}_N)^3 / (Var(\hat{R}_N))^{3/2}$ is the coefficient of skewness

2) $\gamma_2(\hat{R}_N) = \sum_{i=1}^{462} (\hat{R}_i - \bar{\hat{R}}_N)^4 / (Var(\hat{R}_N))^2 - 3$ is the coefficient of kurtosis

3) $CV(\hat{R}_N) = (Var(\hat{R}_N))^{1/2} / \bar{\hat{R}}_N$ is the coefficient of variation.

TABLE A.4 Properties of the sampling distributions of $\hat{Var}(\hat{R}_N)$ from P1 and P2. Four variance estimators are compared, viz. the Ratio, Taylor, Asymptotic and Jacknife estimator.

		P1	P2
$Var(\hat{R}_N)$		0.002931	0.003649
Sampling distributions	Estimators		
$\overline{\hat{Var}(\hat{R}_N)}$	Ratio	0.041634	0.045173
	Taylor	0.003323	0.002627
	Asymptotic	0.002573	0.001215
	Jacknife	0.004981	0.008721
Relative Bias = $\frac{\overline{\hat{Var}(\hat{R}_N)}}{Var(\hat{R}_N)}$	Ratio	14.2047	12.3796
	Taylor	1.1336	0.7198
	Asymptotic	0.8779	0.3330
	Jacknife	1.6995	2.3899
$Var(\hat{Var}(\hat{R}_N))$	Ratio	0.00002529	0.00003897
	Taylor	0.00000721	0.00001495
	Asymptotic	0.00000648	0.00000128
	Jacknife	0.00001313	0.00010225
$\gamma_1(\hat{Var}(\hat{R}_N))$ [1]	Ratio	0.0656	0.3720
	Taylor	0.2045	1.3505
	Asymptotic	0.2616	1.5338
	Jacknife	0.0367	2.0851
$\gamma_2(\hat{Var}(\hat{R}_N))$ [2]	Ratio	-1.1878	-0.8068
	Taylor	-1.5622	1.4334
	Asymptotic	-1.8118	1.8872
	Jacknife	-1.5079	4.1708
$CV(\hat{Var}(\hat{R}_N))$ [3]	Ratio	0.1208	0.1382
	Taylor	0.8082	1.4720
	Asymptotic	0.9889	0.9301
	Jacknife	0.7275	1.1596

Table A.4, cont.

Max $\hat{Var}(\hat{R}_N)$	Ratio	0.051493	0.059357
	Taylor	0.008007	0.015855
	Asymptotic	0.006791	0.005528
	Jacknife	0.011405	0.051416
Min $\hat{Var}(\hat{R}_N)$	Ratio	0.033464	0.034128
	Taylor	0.000008	-0.002716
	Asymptotic	0.000009	0.000046
	Jacknife	0.000044	0.000083
Range	Ratio	0.018029	0.025229
	Taylor	0.007999	0.018571
	Asymptotic	0.006782	0.005482
	Jacknife	0.011360	0.051333
Level of Coverage (95 % CI)	Ratio	100 %	100 %
	Taylor	72.727 %	59.091 %
	Asymptotic	65.152 %	64.719 %
	Jacknife	72.727 %	87.879 %

Notes: See Table A.3

FIGURE A.1 The two parent populations, both of size N = 11, and the sampling distributions of $\hat{R}_N$, the estimated Gini coefficient from simple random sampling without replacement, n = 5.

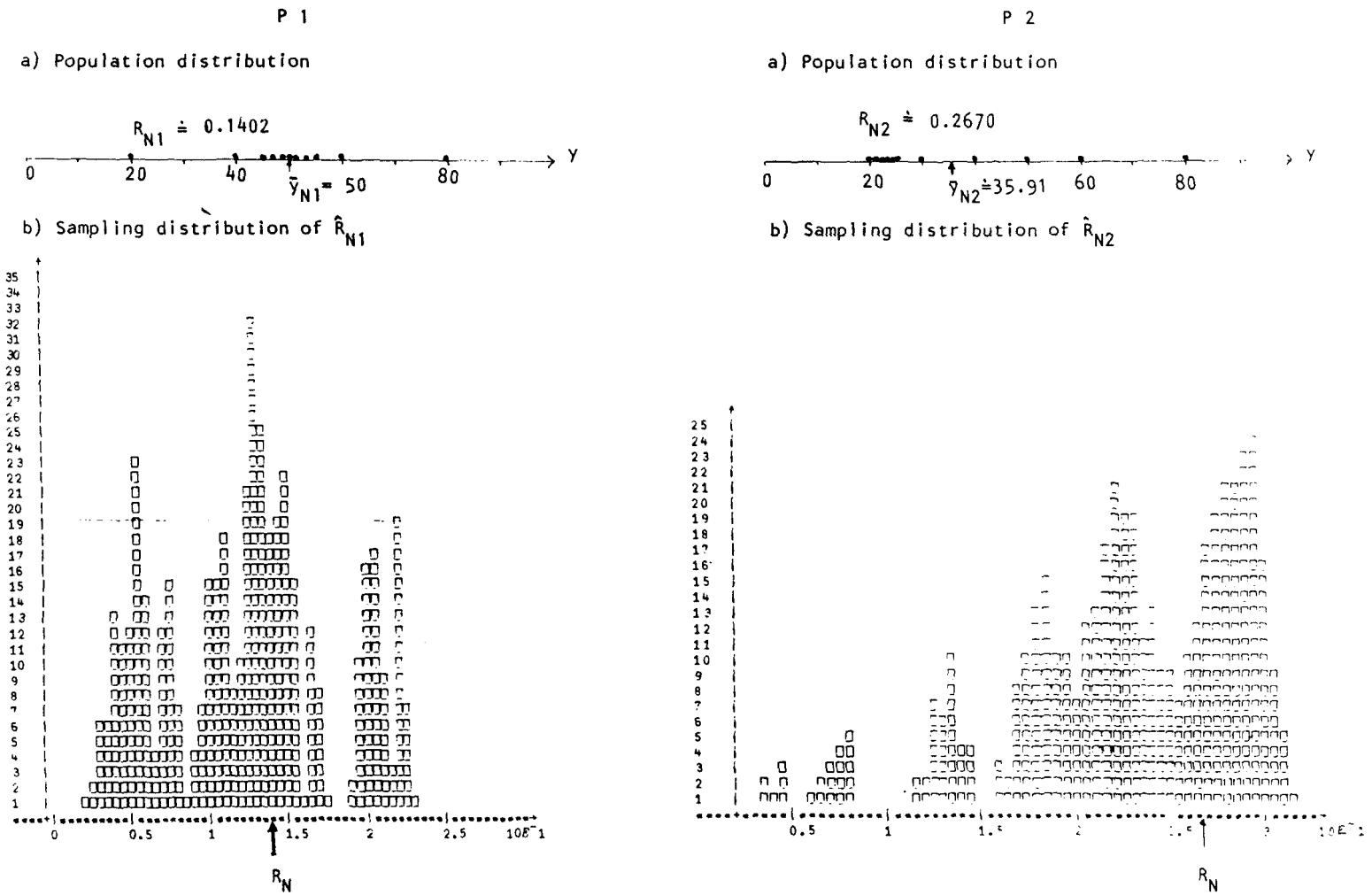
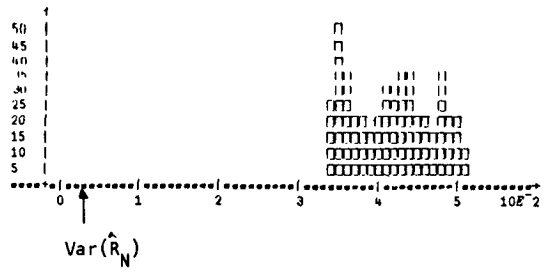


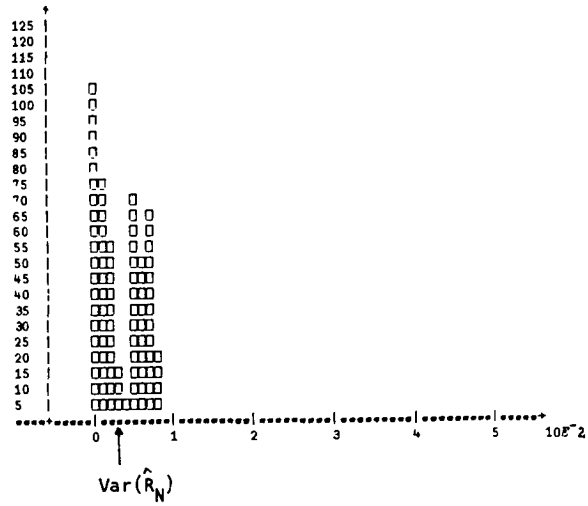
FIGURE A.2 Sampling distributions of variance estimators for the estimated Gini coefficient

P 1

a) Ratio estimator

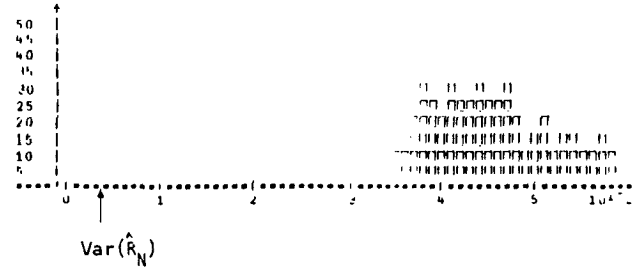


b) Taylor estimator



P 2

a) Ratio estimator



b) Taylor estimator

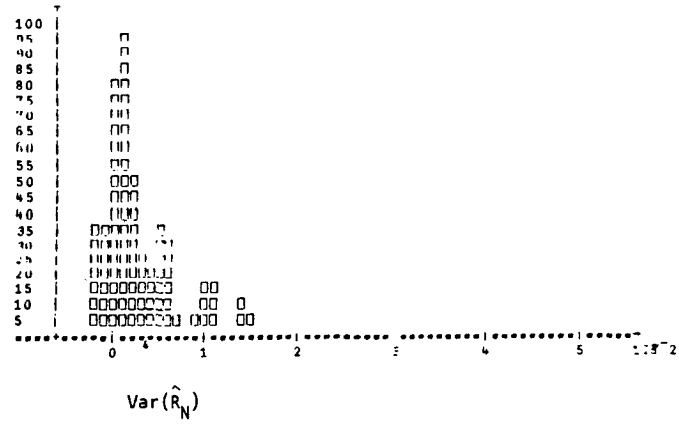
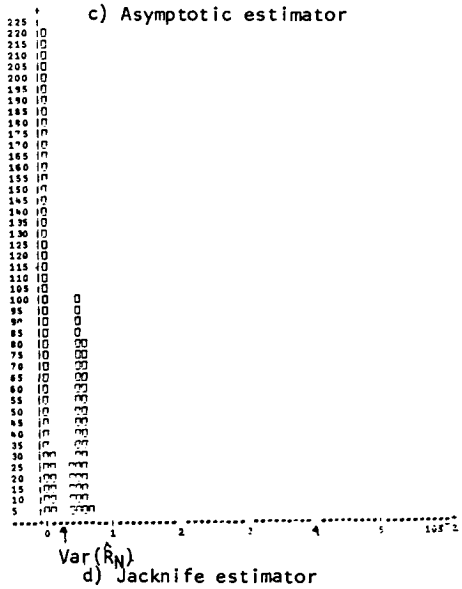
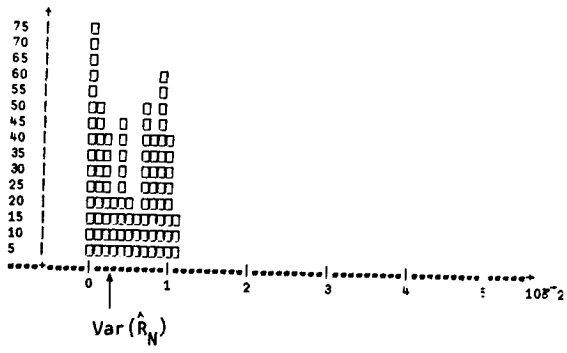


Figure A.2, cont.

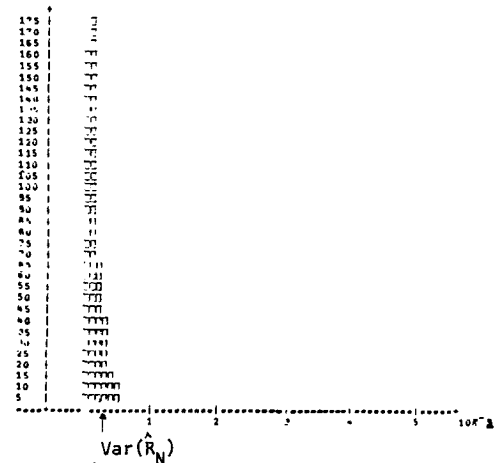
c) Asymptotic estimator



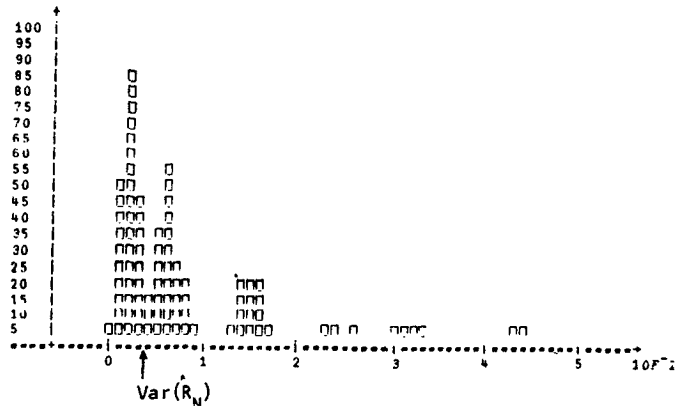
d) Jacknife estimator



c) Asymptotic estimator



d) Jacknife estimator



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 prognoser 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 Dalen (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)

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

☆Garnisonstryckeriet Stockholm 1985☆