SCB STATISTISKA CENTRALBYRÅN

1985-07-05

PROMEMORIOR FRÅN P/STM

NR 17

VARIANCE ESTIMATORS OF THE GINI COEFFICIENT

- PROBABILITY SAMPLING

AV ARNE SANDSTRÖM, JAN WRETMAN OCH BERTIL WALDÉN

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 1985:17. Variance estimators of the Gini coefficient – probability sampling / Arne Sandström m.fl.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-PM-PSTM-1985-17

SCB STATISTISKA CENTRALBYRÅN

1985-07-05

PROMEMORIOR FRÅN P/STM

NR 17

VARIANCE ESTIMATORS OF THE GINI COEFFICIENT

- PROBABILITY SAMPLING

AV ARNE SANDSTRÖM, JAN WRETMAN OCH BERTIL WALDÉN

1985-07-05

VARIANCE ESTIMATORS OF THE GINI COEFFICIENT - PROBABILITY SAMPLING

Arne Sandström, Jan H. Wretman, and Bertil Waldén[1)]

ABSTRACT: The most well-known measure of income inequality, the Gini coefficient, is a ratio statistic. We have studied the approximate sampling distribution of the estimator under a stratified random sampling design. Variance estimators are proposed and compared. Explicit formulas are also given for more complex designs. Some empirical results from Sweden in 1982 are also included.

KEY-WORDS: Gini coefficient, Variance estimators, Simulations, Probability sampling.

1) Sandström and Wretman are Directors, Statistical Research Unit, Statistics Sweden, S-115 81 Stockholm, Sweden, and Waldén is BA, University of Linköping, S-581 83 Linköping. The authors are grateful to Fredrik Nygård, Department of Statistics, Swedish University of Turku, Finland, for valuable comments on an earlier draft.

1. INTRODUCTION

Sandström et al. (1985) studied the sampling distributions of the estimated Gini coefficient and compared four variance estimators under simple random sampling, both with and without replacement. This paper complements the previous one in that we have studied the sampling distributions and the variance estimators under a more general framework, viz. probability sampling. In particular, we use real income data, taken from a Swedish survey "Income Distribution of Households", to conduct the simulations.

The Monte Carlo Study is discussed in Section 2, and in Section 3 we discuss the estimation of the Gini coefficient, the inference problem, and the simulation results. Various variance estimators are introduced in Section 4 accompanied by results from the simulations. Some empirical results and conclusions are given in Section 5.

2. THE MONTE CARLO STUDY

Annual income distribution surveys have been carried out by Statistics Sweden since 1972. The surveys consist of around ten thousand households. Beginning with the 1975 survey, the design was stratified sampling with rotating panels. Each panel, consisting of about half the sample, was part of the surveys for two consecutive years. In this study we have used panel no. 9 (used 1982 and 1983) as the parent population.

The annual surveys are based on a stratified sample of individuals (16 strata). Since there is no way of knowing what types of households these individuals belong to, household classification information is gathered through a questionnaire. As the number of strata at the individual level is 16, the possible number of households is 136, because two individuals may belong to two different strata. The only information on the data file concerning stratum affiliation was the inclusion probabilities. To make the sampling procedure easier and to guarantee each stratum having at least a few hundred objects, the households were reclassified into seven strata. In Table 2.1 we give the number of households in our parent population (which is equal to the number of individuals and households in panel no. 9) and the estimated number of individuals in the total Swedish population.

Table 2.1 The number of objects (individuals/households) in the parent population and estimated number of individuals in the Swedish population.

Stratum h	#individuals in panel no. 9 and households in our parent population n'_h	Estimated number of individuals in Sweden $\hat{N}'_h$
1	939	170 739
2	1 172	106 377
3	1 058	1 439 647
4	826	1 143 073
5	488	298 525
6	577	1 891 367
7	352	1 386 009
	5 412	6 435 737

From our parent population of N = 5412 households, samples of size n = 300 were drawn according to three different designs, of which two were stratified samples (with simple random sampling - srs - without replacement within each stratum) and the third was srs without replacement. The number of sampled units from each stratum was proportional to: the estimated number of individuals in Sweden belonging to the stratum (Simulation 1, S1), and the number of households in the parent population belonging to the stratum (Simulation 2, S2). Hence, if n_{h1} and n_{h2} are the number of sampled units within stratum h, h = 1,..,7, in S1 and S2, respectively, then

$$n_{h1} = 300 \frac{\hat{N}'_h}{\sum \hat{N}'_h}$$

and

$$n_{h2} = 300 \frac{n'_h}{\sum n'_h}$$

The simulation based on the srs design is denoted S3. The number of replicates in each simulation was 500. In Table 2.2 we give the total number of sampled units within each stratum in S1 and S2 with the inclusion probabilities.

Table 2.2 The number of sampled units within each stratum in simulations S1 and S2 together with the inclusion probabilities.

Stratum	# sampled units		inclusion probabilities	
	S1	S2	S1	S2
1	8	47	0.0085	0.0501
2	5	63	0.0043	0.0538
3	67	63	0.0633	0.0595
4	53	46	0.0642	0.0557
5	14	27	0.0287	0.0553
6	88	32	0.1525	0.0555
7	65	22	0.1847	0.0625
	300	300		

3. THE GINI COEFFICIENT AND THE PROBLEM OF INFERENCE

We will begin with the definition of the Gini coefficient, then discuss various approaches to inference and conclude with the simulation results.

3.1 The Gini Coefficient

Let F denote a distribution function (df), by which we mean a real-valued function defined on $(-\infty, \infty)$ that is nondecreasing, right continuous and satisfies $F(-\infty) = 0$ and $F(\infty) = 1$. Gini's mean difference, G, associated with F is defined, in terms of the Lebesgue-Stieltjes integral, as

$$G = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |x-y| dF(x) dF(y), \tag{3.1}$$

and the Gini coefficient, R, associated with F is defined as

$$R = \frac{G}{2\mu} \tag{3.2}$$

$$\mu = \int_{-\infty}^{\infty} y dF(y) \neq 0. \tag{3.3}$$

Let F_N denote the finite population df and let $y_1, \ldots, y_N$ denote the values associated with the units of the finite population (of size N). $F_N(y)$ is then the proportion of units such that $y_k \leq y$. Let $y_{1:N} \leq y_{2:N} \leq \ldots \leq y_{N:N}$ denote the values arranged in nondecreasing order. Inserting F_N for F in (3.1)-(3.3) we obtain, cf. Nygård and Sandström (1981),

$$G_N = \frac{1}{N^2} \sum_{k=1}^{N} \sum_{l=1}^{N} |y_k - y_l| =$$

$$= \frac{2}{N} \sum_{i=1}^{N} (2\frac{i}{N} - 1 - \frac{1}{N}) y_{i:N}, \tag{3.4}$$

$$\mu_N = \bar{y}_N = \frac{1}{N} \sum_{k=1}^{N} y_k,$$

and

$$R_N = \frac{G_N}{2\bar{y}_N} = \frac{\sum_{i=1}^{N} (2\frac{i}{N} - \frac{1}{N}) y_{i:N}}{\sum_{k=1}^{N} y_k} - 1. \tag{3.5}$$

REMARK 3.1 G in (3.1), can be rewritten as $2\int_{-\infty}^{\infty} (2F(y)-1)\, y dF(y)$, and the identity between this formulation and (3.4) when $F = F_N$ is seen by use of the inverse df, cf. Nygård and Sandström (1985b).

The quantity R_N defined by (3.5) is a finite population parameter and it is essentially a ratio between two finite population totals $\sum_{k=1}^{N} w_k y_k$ and $\sum_{k=1}^{N} y_k$, where the ordered w_k's are $w_{i:N} = 2\frac{i}{N} - \frac{1}{N}$. The estimation of R_N is not straightforward, because the w_k-values of the sampled units will remain unknown and have to be estimated.

As indicated above, it is the finite population parameter R_N that we are interested in making inferences about. However, in some situations it may well be a corresponding model parameter that we are interested in. In the next section we will point out different approaches to

making inferences about R or R_N, and we also give explicit point estimates.

3.2 The Problem of Inference and Explicit Estimators

We will shortly consider three different approaches to inference, each on its own assumptions.

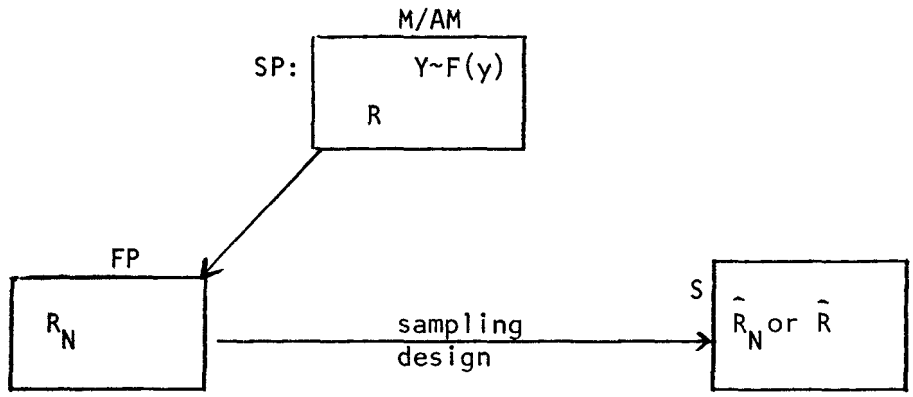
First, we have the Model (M) approach: We assume a superpopulation model with an unknown parameter R and consider the sample, irrespective of the sampling design, as consisting of observations that are independent and identically distributed (iid) as $Y \sim F(y)$.

Second, we have the Auxiliary Model (AM) approach: The superpopulation model is used as a means to get asymptotic results about the estimator of R_N. In this approach, R_N is a stochastic variable (and R is a model parameter) and any confidence statements about R_N are as by Royall (1971), cf. also Cassel et al. (1977, p. 126). The values in the finite population constitute a vector $\underset{\sim}{y}_N = (y_1, y_2, \ldots, y_N)$, considered a random outcome of a stochastic vector $\underset{\sim}{Y}_N = (Y_1, Y_2, \ldots, Y_N)$, where the Y_i's are iid as $Y \sim F(y)$. The sample is assumed to be fixed so the only stochastic element in this approach is the random nature of the finite population vector $\underset{\sim}{Y}_N$. Two subapproaches can be used, the first is based on work by Hoem and Funck-Jensen (1982) and says that if the design is noninformative then it can be ignored. The estimation procedure will be as in the M approach. The second subapproach maintains that one should incorporate the effects of the sampling design in the estimation. In doing so, the point estimates of R_N will be exactly the same as under the fixed and Finite Population (FP) approach, our third approach to inference.

In the FP approach, the vector $\underset{\sim}{y}_N = (y_1, y_2, \ldots, y_N)$ is considered to be fixed and the only stochastic element in this approach is the randomization of the sample. Inference about the FP parameter R_N is based on large sample considerations, i.e. by use of asymptotic results in the M or AM approaches.

The three approaches are summarized in Table 3.1.

Table 3.1 Three approaches in making inference about a Gini coefficient. AM = Auxiliary Model approach, FP = Finite Population approach, M = Model approach, S = sample, SP = superpopulation



Approach	Parameter/variable to be estimated	Inference (95% 'confidence' interval)
M	R, parameter	$\hat{R} \pm 1.96\ \sigma_M$ estimation made from S, (iid), taken from SP
AM	R_N, stochastic variable	$\hat{R}_N \pm 1.96\ \sigma_{AM}$ i. ignoring of the sampling design; estimation as under M ii. taking account for the sample design; confidence statements according to Royall (1971)
FP	R_N, parameter	$\hat{R}_N \pm 1.96\ \sigma_{FP}$ "Large sample" considerations

Variance estimators of σ^2_M, σ^2_{AM}, and σ^2_{FP} are discussed in Section 4. Our personal views are that it is only the FP approach and, perhaps, the AM approach (taking the sampling design into account) that are of primary interest if the finite population Gini coefficient is to be estimated.

Point estimators of the Gini coefficient are given in Table 3.2, cf, Nygård and Sandström (1985a), (1985b), where π_i deonotes the inclusion probability of unit i, $i \in s$, and s denotes the sample of fixed size n.

Table 3.2 Point estimates of the Gini coefficient (from Nygård and Sandström (1985b)).

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

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

3.3 The Monte Carlo Studies of the Sampling Distributions

The Gini coefficient in our parent population (N = 5412) is $R_N = 0.293$. and the arithmetic means of the 500 point estimates based on samples of size n = 300 in the three simulations are, together with the relative means,

	Mean of $\hat{R}_N$	Mean of $\hat{R}_N/R_N$	Notes
S1:	0.2839	0.968	allocated according to the Swedish population
S2:	0.2859	0.974	proportional allocation
S3:	0.2922	0.996	srs

In Figure 3.1 the three observed sampling distributions are illustrated. As seen from this figure and from the results of Table 3.3, S1 and S2 gave rise to the "most" symmetrical distributions.

Table 3.3 Coefficients of skewness and kurtosis, and the minimum and the maximum value of $\hat{R}_N$ in the observed sampling distributions

Simulation	Coefficent of Skewness	Coefficent of Kurtosis	Min $\hat{R}_N$	Max $\hat{R}_N$
S1	0.049	0.417	0.2413	0.3308
S2	0.144	-0.165	0.2545	0.3281
S3	0.682	1.027	0.2558	0.3425

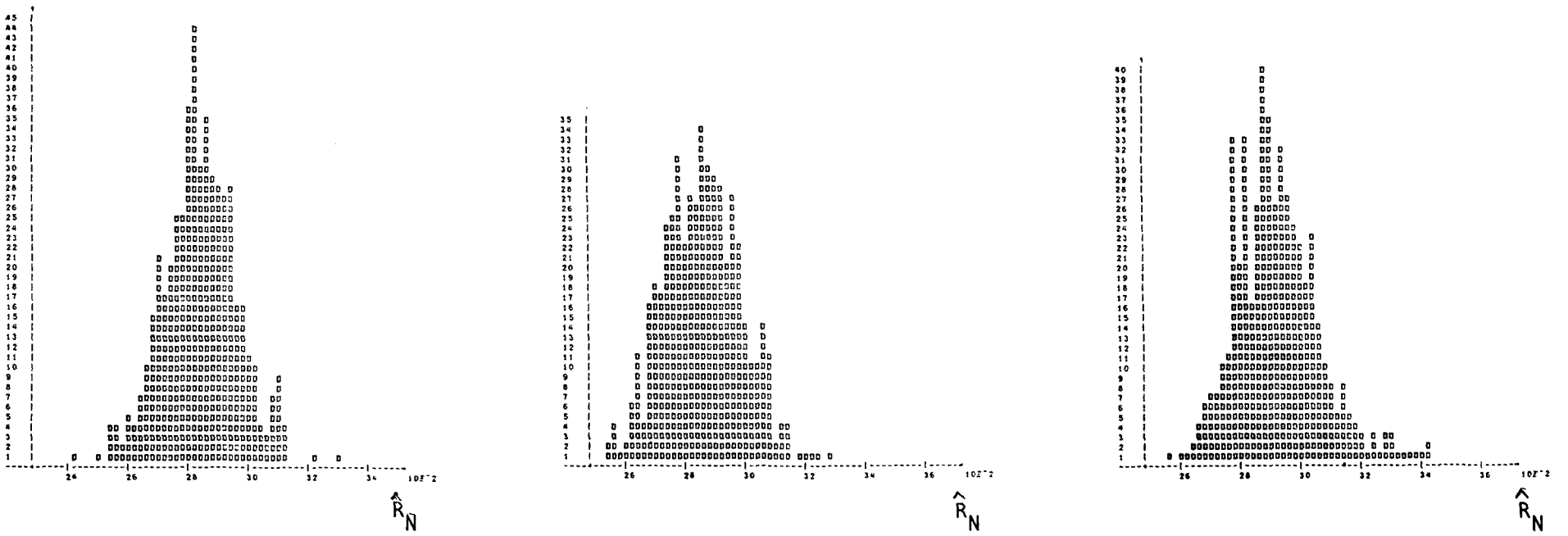
$\hat{R}_N$ based on S3, i.e. under the design srs (without replacement), can be seen as an estimate of R in the M approach. Taken as an estimate of R_N it is "better" than the other two where we have taken the design into consideration. On the other hand, the shape of the sampling distribution indicates that the estimators based on S1 and S2 give better confidence intervals for the Gini coefficient. In the srs-case, the observed sampling distribution are quite symmetrical even for small samples, see Sandström et al. (1985).

Figure 3.1 The approximated sampling distributions of the Gini coefficient based on 500 replicates of samples of size n = 300 from a parent population of the size N = 5412.

S1: Allocation proportional to the number of individuals in the Swedish population; srs, without replacement, within strata.

S2: Proportional allocation; srs, without replacement, within strata

S3: srs, without replacement



4 VARIANCE ESTIMATION

In making inference about any Gini coefficient, e.g. R or R_N, we have to estimate its variance. Under the three approaches, summarized in Table 3.1, there are three,more or less different,variances to be estimated, viz., σ^2_M, σ^2_{AM}, and σ^2_{FP}.

4.1 Variance Estimators

Under the M approach a consistent estimator $\hat{\sigma}^2_M$ is given by Sendler (1979). This is a special case of our estimator $\hat{\sigma}^2_a$ below.

Since the estimator of the Gini coefficient is a "complex" statistic, it is not possible to estimate its variance by traditional methods of unbiased variance estimation. In this case we have to rely on approximate variance estimation technique. In Sandström et al (1985) four such estimators were considered:

i) One method uses a variance estimation formula obtained by the same proces: as the well-known formula for estimating the approximate variance of a ratio estimator ($\hat{\sigma}^2_r$), based on a first-order Taylor approximation. This estimator is proposed in Nygård and Sandström (1985a) as a rough estimator because the weights w_k (see Section 3.1) are considered not sample-dependent. Sandström et al. (1985) show that $\hat{\sigma}^2_r$ overestimates the true variance by a factor of 10-360 depending on the shape of the parent populations. Similar results were obtained in the present study.

ii) In Nygård and Sandström (1985a) another variance estimator, based on the same first-order Taylor approximation as in i), was proposed. The Taylor estimator ($\hat{\sigma}^2_t$) takes account of the sample dependence in the weights w_k. An explicit variance formula is given in op. cit . One disadvantage of $\hat{\sigma}^2_t$ is that the general probability sampling design implies inclusion probabilities up to the fourth order to be included. When n and N are large this estimator coincides with the third estimator, see below. As this estimator does not seem to have any advantages over the next two estimators (cf. Sandström et al. (1985)) and because we wanted to keep away from computational problems, this variance estimator was not computed in the simulation study.

iii) The third method is motivated by the AM approach, and the resulting variance estimation formula is a consistent estimator $(\hat{\sigma}_a^2)$ of the asymptotic expression for the "AM-expected squared error" $\mathcal{E}(\hat{R}_N - R_N)^2$, where $\mathcal{E}$ denotes expectation with respect to the assumed AM, for a fixed sample s and where the inclusion probabilities are deterministic weights. For details, see Sandström (1983) and Nygård and Sandström (1985a). The formula, under a fixed probability sample, is given in Table 4.1.

iv) The forth method is based on a jackknife technique $(\hat{\sigma}_j^2)$. One observation at a time is deleted from the sample. Each time we calculate $\hat{R}_N^{(j)}$, analogous to $\hat{R}_N$ (with the inclusion probabilities), based on the remaining n-1 observations and deleting the jth observation, j=1,2,..n. The variance estimation formula is:

$$\hat{\sigma}_j^2 = (n-1)n^{-1} \sum_{j=1}^{n} (\hat{R}_N^{(j)} - \hat{R}_N^{(\cdot)})^2 ,$$

where $\hat{R}_N^{(\cdot)} = n^{-1} \sum_{j=1}^{n} \hat{R}_N^{(j)}$.

The estimators $\hat{\sigma}_r^2$, $\hat{\sigma}_t^2$, and $\hat{\sigma}_a^2$ are given for srs, without replacement, in Nygård and Sandström (1985a) and in Sandström et al. (1985). Glasser (1962) used method ii) $(\hat{\sigma}_t^2)$ to estimate the variance based on srs with replacement and Love and Wolfson (1976) compared Glasser's approach to a balanced repeated replication approach in estimating the variance when the sampling design was more complex than srs.

In Table 4.2 we have summarized some possible choices of variance estimators for the three approaches summarized in Table 3.1.

Table 4.1 Explicit expression for the asymptotic variance estimator, $\hat{\sigma}_a^2$.

$$\hat{\sigma}_a^2 = \frac{(1-\hat{f}+\hat{v}^2)}{n\hat{\bar{y}}_N^2}\,\hat{\sigma}_1^2,$$

where $\hat{f} = \frac{n}{\hat{N}}$

$$\hat{N} = \sum_{i=1}^{n} \pi_i^{-1}$$

$$\hat{v}^2 = \frac{\hat{f}}{\hat{N}} \sum_{i=1}^{n} (\pi_i^{-1} - \hat{f}^{-1})^2 \quad \text{, the squared coefficient of } 1/\pi_i,\ i=1,..n$$

$$\hat{\bar{y}}_N = \hat{N}^{-1} \sum_{i=1}^{n} y_i/\pi_i,$$

and

$$\hat{\sigma}_1^2 = (\hat{b}\Delta_o - a\Delta_1)(\hat{b}\Delta_1 - a\Delta_2) - (\hat{b}\Delta_1 - a\Delta_2)^2 -$$

$$- \frac{1}{2}\{a^2 S_2 - 2a\hat{b}S_1 + \hat{b}^2 S_3\},$$

where

$$a = 2$$

$$\hat{b} = 1 + \hat{R}_N$$

$$\Delta_o = y_{(n)} - y_{(1)}$$

$$y_{(1)} \le y_{(2)} \le \cdots \le y_{(n)}$$

$$\Delta_1 = y_{(n)} - \hat{\bar{y}}_N$$

$$\Delta_2 = y_{(n)} - \hat{\bar{y}}_N(\hat{R}_N - 1).$$

Let furthermore

$$\Delta_{oi} = y_{(i)} - y_{(1)}, \qquad \text{i.e. } \Delta_{on} = \Delta_o$$

$$\Delta_{1i} = F_i y_{(i)} - \frac{1}{\hat{N}} \sum_{\{y_k \le y_{(i)}\}} y_k/\pi_k \quad , \qquad \text{i.e.} \Delta_{1n} = \Delta_1$$

$$\Delta_{2i} = F_i^2\, y_{(i)} - \sum_{\{y_k \le y_{(i)}\}} (F_k^2 - F_{k-1}^2) y_k \quad , \quad \text{i.e. } \Delta_{2n} = \Delta_2$$

$$\text{and } F_i = \hat{N}^{-1} \sum_{j=1}^{n} I\{y_j \le y_i\}/\pi_j .$$

Now S_1, S_2, and S_3 are, with all summations taken over $i=1,2,..,n$,

$$S_1 = \Delta_1^2 - \Delta_2\Delta_o + 2y_{(n)}(\Delta_o - \Delta_1) - 2\Sigma F_1^2(\Delta_{oi} - \Delta_{oi-1})y_i +$$

$$+ 2\hat{N}^{-2}\Sigma\Delta_{oi-1}y_i/\pi_i^2 - 4\hat{N}^{-1}\Sigma F_i\,\Delta_{oi-1}y_i/\pi_i +$$

$$+ 2\Sigma F_i(\Delta_{1i} - \Delta_{1i-1})y_i + 2\hat{N}^{-1}\Sigma\Delta_{1i-1}y_i/\pi_i$$

$$S_2 = 2y_{(n)}(\Delta_1 - \Delta_2) + 2\Sigma F_i(\Delta_{2i} - \Delta_{2i-1})y_i - 2\Sigma F_i^2(\Delta_{1i} - \Delta_{1i-1})y_i +$$

$$+ 2\hat{N}^{-2}\Sigma\Delta_{1i-1}\,y_i/\pi_i^2 + 2\hat{N}^{-1}\Sigma\Delta_{2i-1}y_i/\pi_i - 4\hat{N}^{-1}\Sigma F_i\Delta_{1i-1}y_i/\pi_i$$

$$S_3 = 2y_{(n)}(\Delta_o - \Delta_1) - 2\Sigma F_i(\Delta_{oi} - \Delta_{oi-1})y_i -$$

$$- 2\hat{N}^{-1}\Sigma\Delta_{oi-1}\,y_i/\pi_i + 2\Sigma(\Delta_{1i} - \Delta_{1i-1})y_i .$$

Note: This variance estimator, $\hat{\sigma}_a^2$, is a consistent estimator of the variance

$$\sigma^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{\min(F(y),F(x)) - F(y)F(x)\}\{2F(y)-1-R\}\{2F(x)-1-R\}dx\,dy,$$

cf. Sendler (1979).

All partial sums are easily computed if the observations are arranged in nondecreasing order.

The relation between F_i and P_i in Table 3.2 is

$$F_i = \hat{N}^{-1}P_i + \hat{N}^{-1}\sum_{j=1}^{n} I\{y_j = y_i\}/\pi_j$$

Table 4.2 Some possible variance estimators in the three inference approaches discussed in Section 3.2. Cf. Table 3.1

Approach	Variance	Some possible estimators
M	σ^2_M	$\hat{\sigma}^2_a$, srs with replacement cf. Sendler (1979)
AM	σ^2_{AM}	i) neglecting the sampling design: as under M ii) taking account of the sampling design: $\hat{\sigma}^2_a$ (method iii), Table 4.1) $\hat{\sigma}^2_j$ (-"- iv))
FP	σ^2_{FP}	$\hat{\sigma}^2_a$ (method iii), Table 4.1) $\hat{\sigma}^2_j$ (-"- iv)) $\hat{\sigma}^2_r$ (-"- i)) and $\hat{\sigma}^2_t$ (method ii), not considered here because it includes up to fourth order inclusion probabilities)

4.2 The Monte Carlo Studies of the Variance Estimators

In each simulation (S1, S2, and S3) we computed the variance among the 500 estimates $\hat{R}_N$ based on a sample size of $n = 300$. In each sample, we computed the variance estimators $\hat{\sigma}_a^2$, $\hat{\sigma}_j^2$, and $\hat{\sigma}_r^2$ to study their sampling distributions and the coverage rate of confidence interval of the type $\hat{R}_N \pm 1.96\ \hat{\sigma}$..

The standard deviation of the 500 observed values of $\hat{R}_N(\hat{\sigma})$ with the square root of the arithmetric means of the 500 variance estimators are

	$\hat{\sigma}$	$\hat{\sigma}_a$	$\hat{\sigma}_j$	$\hat{\sigma}_r$
S1:	0.0123	0.0123	0.0128	0.0422
S2:	0.0129	0.0124	0.0128	0.0423
S3:	0.0139	0.0139	0.0146	0.0435

To compare the bias of the various estimators relative to $\hat{\sigma}$ we have looked at the following data:

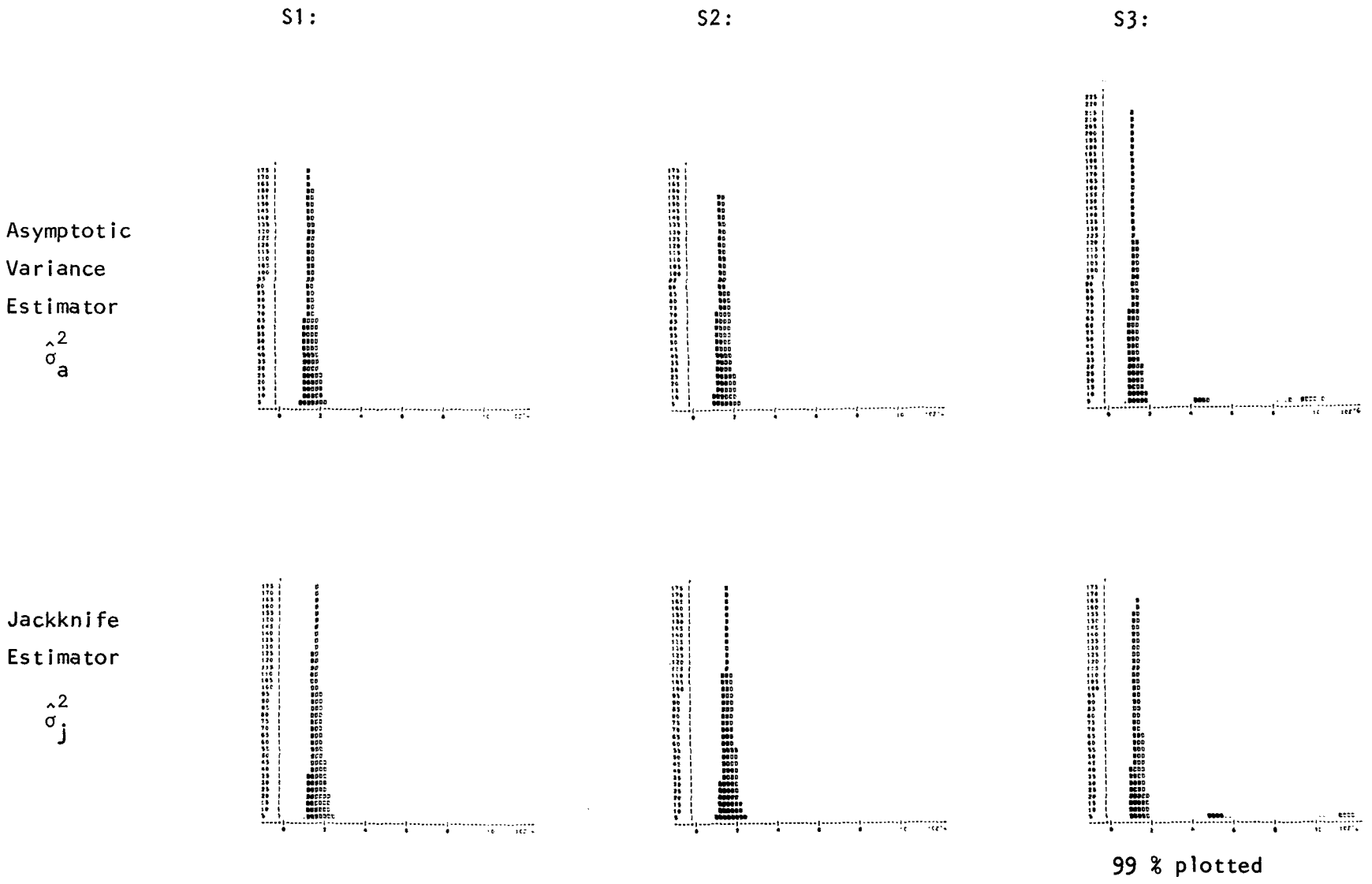
	$\hat{\sigma}_a/\hat{\sigma}$	$\hat{\sigma}_j/\hat{\sigma}$	$\hat{\sigma}_r/\hat{\sigma}$
S1:	1.00	1.04	3.43
S2:	0.96	0.99	3.28
S3:	1.00	1.05	3.13

As is shown by the above data the estimator $\hat{\sigma}_r^2$ is, as mentioned earlier, a very rough estimator. It overestimates $\hat{\sigma}^2$ by a factor of 10. Because of this we will not discuss it any further.

The design effect, as compared with the srs-design, can be measured as $\hat{\sigma}_j/\hat{\sigma}_{j,srs}$ and $\hat{\sigma}_a/\hat{\sigma}_{a,srs}$, where for example $\hat{\sigma}_{j,srs}^2$ is the jackknife variance estimator under srs:

	$\hat{\sigma}_a/\hat{\sigma}_{a,srs}$	$\hat{\sigma}_j/\hat{\sigma}_{j,srs}$
S1:	0.88	0.88
S2:	0.89	0.88

Figure 4.1 The approximated sampling distributions of $\hat{\sigma}_a^2$ and $\hat{\sigma}_j^2$, based on 500 replicates of samples of size n = 300 from a parent population of the size N = 5412.



Once again the close resemblance in the results of $\hat{\sigma}^2_a$ and $\hat{\sigma}^2_j$ is shown.

Both the visual display of the approximated sampling distributions in Figure 4.1 and the summarizing measures of these distributions in Table 4.3 shows that $\hat{\sigma}^2_a$ and $\hat{\sigma}^2_j$ are quite similar in their performance. The distribution of $\hat{\sigma}^2_a$ is more symmetrical than that of $\hat{\sigma}^2_j$ and the estimates of $\hat{\sigma}^2_j$ are more spread out (wider range).

The results of S3 (srs, without replacement) are astounding: the distributions of the variance estimators are positively skewed and trimodal. This is also seen in Figure 4.2 where we have plotted $\hat{R}_N$ against $\hat{\sigma}^2_a$ in the three simulations. The same relationship holds for $\hat{\sigma}^2_j$.

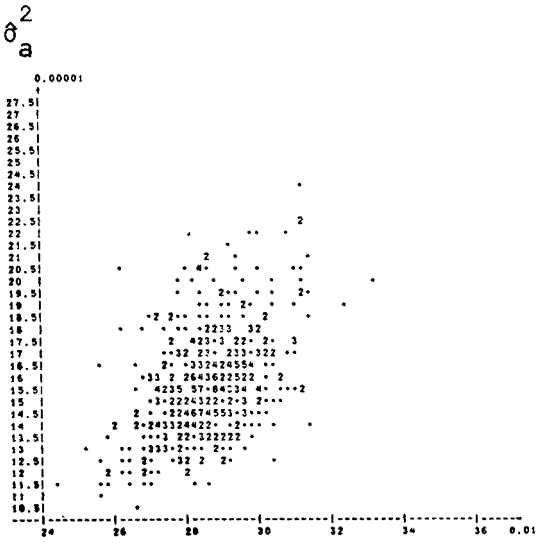
Table 4.3 Coefficients of skewness and kurtosis, in the observed sampling distributions of $\hat{\sigma}^2_a$ and $\hat{\sigma}^2_j$, and the minimum and the maximum values of $\hat{\sigma}_a$ and $\hat{\sigma}_j$.

	Coefficient of skewness		kurtosis		Min $\hat{\sigma}$		Max $\hat{\sigma}$	
	$\hat{\sigma}^2_a$	$\hat{\sigma}^2_j$	$\hat{\sigma}^2_a$	$\hat{\sigma}^2_j$	$\hat{\sigma}_a$	$\hat{\sigma}_j$	$\hat{\sigma}_a$	$\hat{\sigma}_j$
S1:	0.538	0.576	0.365	0.372	0.0100	0.0104	0.0153	0.0159
S2:	0.353	0.415	-0.012	0.153	0.0101	0.0104	0.0150	0.0156
S3:	3.145	3.178	8.688	8.887	0.0092	0.0095	0.0323	0.0350

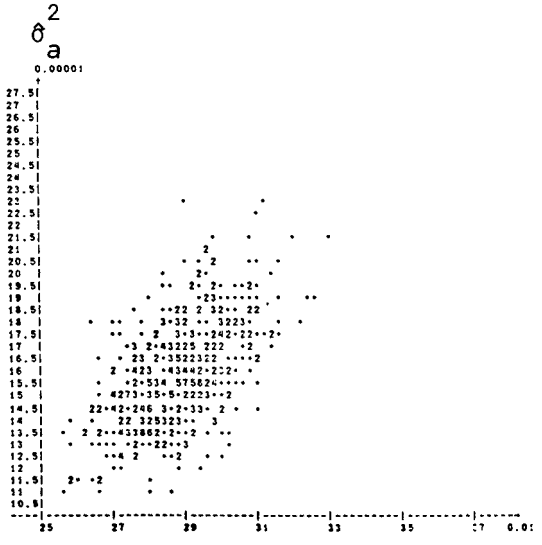
Figure 4.2 The estimators $\hat{R}_N$ plotted against the variance estimators $\hat{\sigma}^2_a$; 500 replicates of samples of size n = 300 from a parent population of the size N = 5412.

Interpretation:		
	1 *OBSERVATION*	: •
	2→9	: 2→9
	10→19	: *O*
	20→49	: ⊛
	50→99	: ▯
	>100	: ⊟

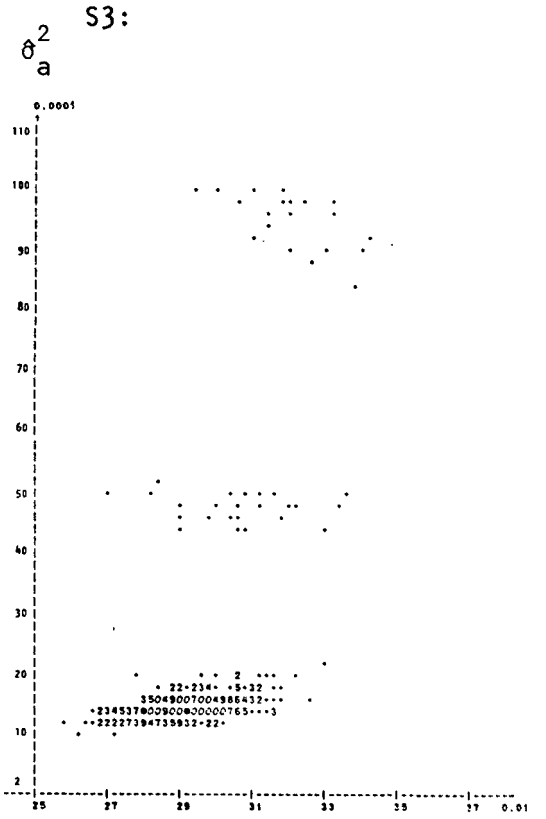
S1:



S2:



S3:



The correlations between the estimator $\hat{R}_N$ and the variance estimators are:

	$\rho_{\hat{R}_N, \hat{\sigma}^2}$	
	$\hat{\sigma}^2_a$	$\hat{\sigma}^2_j$
S1	0.47	0.46
S2	0.58	0.58
S3	0.57	0.57

The resemblance between $\hat{\sigma}^2_a$ and $\hat{\sigma}^2_j$ is also seen by the correlation between them (S1:) 0.99, (S2:) 0.99, and (S3:) 1.00. In the same way the rate of coverage (nominal: 95 %) is quite similar and, as one can believe from the above results, the rate will be highest in S3, because of the skewed distribution of the variance estimators:

	Rate of Coverage (nominal: 95%)	
	Variance estimator	
	$\hat{\sigma}^2_a$	$\hat{\sigma}^2_j$
S1	87.4	88.6
S2	87.8	89.4
S3	94.0	94.8

5. CONCLUSIONS AND SOME EMPIRICAL RESULTS

5.1 Conclusions

From the above results we may draw the following conclusions (in inference about R_N):

i) When the sampling design is stratified random sampling we believe that it is better to take account of the design than just to use srs-methods <u>even</u> if the latter gives point estimates closer to the true value and a rate of coverage closer to the nominal value. This is rested upon the distributions of the variance estimators.

We suggest that whenever the design is more "complex" than srs one should take account of this in the estimation procedure.

ii) Of the variance estimators, $\hat{\sigma}_a^2$ and $\hat{\sigma}_j^2$ seem to be quite similar. The observed sampling distributions of the variance estimates, suggest that $\hat{\sigma}_a^2$ might be preferred to $\hat{\sigma}_j^2$. The computational formula of $\hat{\sigma}_a^2$ in Table 4.1 may seen quite laborious, but once the income data has been arranged in nondecreasing order both $\hat{R}_N$ and $\hat{\sigma}_a^2$ are arrived at simultaneously after a single computation.

Because we believe that in most situations it is R_N that we are interested in the following approximated 100 $(1-\alpha)$ per cent confidence interval can be constructed (for details, see Table 4.1):

$$\hat{R}_N \pm z_{\alpha/2} \underbrace{\frac{(1-\hat{f}+\hat{v}^2)^{\frac{1}{2}}\,\hat{\sigma}_1}{n^{\frac{1}{2}}\,\hat{\bar{y}}_N}}_{\hat{\sigma}_a}, \qquad (5.1)$$

where $z_{\alpha/2}$ is the solution of $\phi(z_{\alpha/2}) - \phi(-z_{\alpha/2}) = 1-\alpha$ and $\phi(\cdot)$ denotes the standard normal distribution function.

The interval (5.1) has two interpretations depending on the choice of inference approach: Under FP (5.1) has the interpretation of an ordinary confidence interval but under AM the interpretation is of Royall-type, cf. the discussion in Section 3.2.

5.2 Some Empirical Results

Our study is based on half the sample in the 1982 Swedish Survey "Income Distribution of Households". In order to estimate the Gini coefficient in disposable income among Swedish households in 1982 we used the whole sample of n = 10 234 households. Both the asymptotic and the jackknife variance estimators ($\hat{\sigma}_a^2$ and $\hat{\sigma}_j^2$) were used, but as stated above, we believe that $\hat{\sigma}_a^2$ should be preferred. The study includes both the Gini coefficient of the disposable income/household (R_{Nh}) and the disposable income/consumption unit (R_{Nc}) as calculated in the survey. The results are summarized in Table 5.1.

Table S.1 Income inequality among Swedish households in 1982

Sample size, $n = 10234$

Estimated total number of households, $\hat{N} = 4\ 389\ 211$

$\hat{f} = n/\hat{N} = 0.0023$

$\hat{v}^2 = 2.3447$ (squared coefficient of variations of π_i^{-1})

$1 - \hat{f} + \hat{v}^2 = 3.3424$

Disposable income/household	Approximated 95% confidence interval	Coefficient of variation, cv $(\hat{R}_N)$
$\hat{R}_{Nh} = 0.3215$		
$\hat{\sigma}_j^2 = 9.8663 \times 10^{-6}$	0.3215 ± 0.0062	0.0120
$\hat{\sigma}_a^2 = 14.9371 \times 10^{-6}$	0.3215 ± 0.0076	0.0098
$\hat{\sigma}_1^2 = 252\ 854\ 032.5$		
$\hat{\bar{y}}_N = 74\ 354.7$		
Disposable income/consumption unit		
$\hat{R}_{Nc} = 0.2099$		
$\hat{\sigma}_j^2 = 7.4252 \times 10^{-6}$	0.2099 ± 0.0053	0.0184
$\hat{\sigma}_a^2 = 17.9172 \times 10^{-6}$	0.2099 ± 0.0076	0.0130
$\hat{\sigma}_1^2 = 117\ 369\ 372.9$		
$\hat{\bar{y}}_N = 50\ 692.1$		

REFERENCES

CASSEL, C.M., SÄRNDAL, C.E., and WRETMAN, J.H. (1977): Foundations of Inference in Survey Sampling, John Wiley & Sons, New York

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

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

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

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

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

- (1985b): Income Inequality Measures Based on Sample Surveys, Invited Paper, International Statistical Institute, Amsterdam, Memo No. 14, Statistical Research Unit, Statistics Sweden, 1985-05-20

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

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

SANDSTRÖM, A., WRETMAN, J.H., and WALDEN. B. (1985): Variance Estimators of the Gini Coefficient, Simple Random Sampling, Memo No. 16, Statistical Research Unit, Statistics Sweden

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

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)

14 Income inequality measures based on sample surveys. Fredrik Nygård och Arne Sandström (1985-05-20)

15 Granskning och evalvering av surveymodeller, tiden före 1960. Gösta Forsman (1985-05-30)

16 Variance estimators of the Gini coefficient - simple random sampling. Arne Sandström, Jan Wretman och Bertil Waldén (Memo, Februari 1985)

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

☆Garnisonstryckeriet Stockholm 1985☆