Journal of Official Statistics
Vol. 1, No. 4, 1985, pp. 399–412
© Statistics Sweden

The Estimation of the Gini and the Entropy Inequality Parameters in Finite Populations

Fredrik Nygård[1] *and Arne Sandström*[2]

Abstract: 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 estimations in sample surveys are discussed. The estimation procedures are made both under a fix population approach and under an auxiliary model approach. A number of variance estimators are discussed.

Key words: Inequality parameters; Gini family; Generalized Entropy family; finite populations; 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.

We will call these dispersion measures *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.

The discussion in this paper is restricted to the case of *finite* populations and shows how some commonly used inequality parameters, viz the Gini and the Generalized Entropy families, may be *computed* in total surveys and *estimated* in sample surveys. The pure model approach is not discussed here; the reader is referred to e.g. Nygård and Sandström (1981).

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

A review of the sampling properties of the Gini family is given in Nygård and Sandström (1985 b).

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 from finite populations 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 illuminate two frequently used classes of inequality parameters and formally define them by using a functional approach. The first class is related to the well-known Lorenz Curve, because its members may be interpreted as weighted Lorenz areas. This class will be called the *Gini family*, since it includes the Gini coefficient. The second class of parameters is the *Generalized Entropy family*, consisting of members fulfilling some special criteria imposed on inequality measures, see e.g. Cowell (1980).

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*) $F(y)$ with $E(Y) = \mu \neq 0, < \infty$. In terms of a statistical functional, μ 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

the Gini family:

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

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

and $J(\cdot)$, sometimes referred to as "the weight function," is bounded and continuous.

the Generalized Entropy family:

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

where $T_c(F) = \int_{-\infty}^{\infty} y^c dF(y)$,
with the limiting value, see e.g. Shorrocks (1982), when $c \to 0$ or $c \to 1$:

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

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

In Table 1 some examples of parameters belonging to the above families are given.

Table 1. Some inequality parameters belonging to the Gini family and to the Generalized Entropy family

1. The Gini family	
Weight function, $J(p)$	Name[1]
$2p-1$	R, The Gini coefficient
$1-3(1-p)^2$	M, Mehran's measure
$\frac{1}{2}(3p^2-1)$	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	$E_2 = V^2/2$, V is the coefficient of variation[4]

[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.

The arithmetic mean in the finite population is given by (2.2). If there is $N^1 \leq N$ distinct values of y then we define the probability function at $y_{(i)}$, where $y_{(1)} < y_{(2)} <, \ldots, < y_{(N^1)}$, as

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

With use of (3.2) for unordered distinct values y_i we get

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

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

$$T_G(F_N) = \int_{-\infty}^{\infty} J(F_N(y))\, y\, dF_N(y)$$

$$= \sum_{i=1}^{N^1} J(F_N(y_i))\; y_i f_N(y_i), \tag{3.3a}$$

and if no tied y-values are present, we can rewrite (3.3a) as a linear function of the ordered data set

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

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 $\{\cdot\}$ occurs 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. so the computation is straightforward when the observations are rank-ordered.

The weight function for the finite population Gini coefficient (cf. Table 1) is, in terms of (3.3a), $J(F_N(y_i)) = 2F_N(y_i) - 1$ and, of (3.3b), $2\frac{i}{N} - 1$. For a non-negative variable $R_N \in [1/N, 1]$. The usual definition found in the literature, cf. Nygård and Sandström (1981), is based on its relation to Gini's mean difference $G = E(|X-Y|)$, which is $R = G/2\mu$. In the finite population case, a J-function corresponding to (3.3a) will be $2F_N(y_i) - 1 - f_N(y_i)$. In the case of (3.3b), we have a J-function equal to $2\frac{i}{N} - 1 - \frac{1}{N}$. The term $-\frac{1}{N}$ will be called the Gini finite population correction (Gfpc). In the

Table 2. Expressions for some parameters of the Gini family corrected for finite populations

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	$\frac{3}{2N^3\bar{y}_N}\sum_{i=1}^{N} i^2y_{(i)} - \frac{3}{2N^3\bar{y}_N}\Sigma iy_{(i)} - \frac{(N-1)(N+1)}{2N^2}$	$[0, (1-\frac{1}{N})(1-\frac{1}{2N})]$

Table 3. 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 $E_{cN} \in [0, \frac{N^{c-1}-1}{c(c-1)}]$

non-negative case and including the Gfpc-term $R_N \in [0, 1-\frac{1}{N}]$. 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 op.cit.), 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 the finite population corrected parameters of the Gini family. In Table 2 explicit expressions for some members belonging to the Gini family are given and in Table 3 we have explicit expressions for parameters of the Generalized Entropy family.

3.2. *Calculations from Grouped Data*

In practice, we frequently have to deal with situations in which we do not have access to the complete data and 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 has specific assumptions regarding the behaviour of the distribution function $F_N(y)$ within the different groups – a vast number of suggestions are found in the litera-

ture (for references see Nygård and Sandström (1981), p. 113, Dagum (1983), MacDonald (1984)). According to other related methods, the parameter calculation is based on interpolation/extrapolation techniques (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 population's parameter value 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, $i=1,\ldots,k$, $\Sigma_i N_i = N$, $\Sigma_i N_i \bar{y}_i = N\bar{y}_N$.

In this situation, the standard textbook method of calculating the Gini and Entropy parameters in Table 2 and 3 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 sense a sound procedure, since the 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. also Gastwirth (1975)) that the maximum is obtained by placing $(1-\lambda_i)N_i$ of the observations in interval 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 3.2. The minimum parameter value occurs when all observations equal the group mean, and the maximum value occurs when the observations are placed at the group boundaries. This is an immediate consequence of the parameter value increasing when an initially high income increases at the expense of a corresponding decrease in a lower income.

For this method to function, the original observations must be at least somewhat evenly distributed within the interval. They should not lie on the boundaries or the interval mean $\bar{y}_i$. Then, from the original distribution, we may always derive (without violating the assumptions of given interval means) a hypothetical distribution by increasing high incomes with corresponding reductions in lower incomes. In this hypothetical distribution, $\lambda_i N_i$ of the points are situated at the upper boundary a_i and $(1-\lambda_i)N_i$ at the lower boundary a_{i-1}.

This will increase the parameter value, and hence the inequality parameter will exceed (strictly: never less than) the value associated with the original incomes. Similarly, we may from a hypothetical distribution where all incomes equal their group mean derive the original parameter value by increasing some incomes at the expense of others. Again, this increases the parameter value and so the original value can never be lower than the value obtained by replacing actual incomes by group means.

Formulas for the lower bound and maximum bias, which added to the lower bound gives the upper bound, are presented in Table 4 for the Gini and Generalized Entropy parameters.

Table 4. Lower bounds and the maximum bias[1] of these bounds for parameters of the Gini and

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} $c\neq 0,1$	$\frac{1}{c(c-1)}\frac{1}{N}\sum_{i=1}^{k} N_i\left\{\left(\frac{\bar{y}_i}{\bar{y}_N}\right)^c-1\right\}$

[1] *The upper bound is obtained by addition of the maximum bias to the lower bound.*

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

REMARK 3.3. 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 3.4. 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).

4. Sample Surveys

4.1. *The Fix Population Approach*

Let $y_1, y_2,\ldots,y_N$ be values associated with the units of a finite and identifiable population of size N. The population universe is defined as a label set $U = \{1, 2,\ldots, N\}$. A sample s is a subset of U and a sampling experiment will yield a sample $s \subset U$ according to a probability distribution $P(s)$. $\{P(s), s \subset U\}$ is called the sampling design. $p_n = n/N$ is called the sampling fraction, $0<p_n<1$, where n is the fixed sample size. The inclusion indicator is defined as $I_{\{i\in s\}} = 1$ if $i\in s$ and 0 otherwise, and $E(I_{\{i\in s\}}) = P(i\in s) = \pi_i$ is the first order inclusion probability. In a similar way, higher order inclusion probabilities may be defined. If we are summing over the sample s we write either $\sum_{i\in s}$ or $\sum_{i=1}^{n}$ and we use the same kind of notation when summing over the whole population.

the Generalized Entropy family when calculated from grouped data

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} $c\neq 0,1$	$\frac{1}{c(c-1)}\frac{1}{N}\sum_{i=1}^{k} N_i\{(1-\lambda_i)\,(\frac{a_{i-1}}{\bar{y}_i})^c+\lambda_i(\frac{a_i}{\bar{y}_i})^c-(\frac{y_i}{\bar{y}_N})^c\}$

By the functional representation of the inequality parameters introduced in Section 2, we have only 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

$$\hat{F}_N(y) = \hat{N}_s^{-1}\sum_{i\in s} I_{\{y_i\leq y\}}/\pi_i,\quad \forall y, \tag{4.1}$$

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

REMARK 4.1. The estimator (4.1) is a Hájek 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 (4.1) 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 Hájek 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 5 for the parameters under consideration. The estimation procedure in the Gini case has to be done in two steps: first the data are arranged in increasing order such that $y_{j_1}\leq y_{j_2}\leq,\ldots,\leq y_{j_n}$, $j_i\varepsilon s$, and then step two is a straightforward computation.

Table 5. Point estimators of the finite population inequality parameters under the fix population approach

Family	Parameter	Estimator	
Gini	Gini coefficient, R_N	$\hat{R}_N = \dfrac{2\sum_{i\in s}\left(P_{s(i)} + \dfrac{1}{2\pi_i}\right) y_i/\pi_i}{P_n \sum_{i\in s} y_i/\pi_i} - 1$	$P_{s(i)} = \sum_{i\in s} I_{\{y_j<y_i\}}/\pi_j$ $P_n = \sum_{i\in s} 1/\pi_j = \hat{N}_s$
	Mehran's measure, M_N	$\hat{M}_N = \dfrac{6\sum_{i\in s}\left(P_{s(i)} + \dfrac{1}{2\pi_i}\right) y_i/\pi_i}{P_n \sum_{i\in s} y_i/\pi_i} - \dfrac{3\sum_{i\in s}\left(P_{s(i)}^2 + P_{s(i)} \cdot \dfrac{1}{\pi_i} + \dfrac{1}{3\pi_i^2}\right) y_i/\pi_i}{P_n^2 \sum_{i\in s} y_i/\pi_i} - 2$	
	Piesch's measure, P_N	$\hat{P}_N = \dfrac{3\sum_{i\in s}\left(P_{s(i)}^2 + P_{s(i)}\dfrac{1}{\pi_i} + \dfrac{1}{3\pi_i^2}\right) y_i/\pi_i}{2P_n^2 \sum_{i\in s} y_i/\pi_i} - \dfrac{1}{2}$	
Generalized Entropy	E_{0N}	$\hat{E}_{0N} = -\log \hat{N}_s + \log\left\{\sum_{i\in s} y_i/\pi_i\right\} - \hat{N}_s^{-1} \sum_{i\in s} \pi_i^{-1} \cdot \log y_i$	
	E_{1N}	$\hat{E}_{1N} = \log \hat{N}_s - \log\left\{\sum_{i\in s} y_i/\pi_i\right\} + \dfrac{\sum_{i\in s} \dfrac{y_i}{\pi_i} \log y_i}{\sum_{i\in s} y_i/\pi_i}$	
	E_{cN}, $c \neq 0,1$	$\hat{E}_{cN} = \dfrac{\hat{N}_s^{c-1} \sum_{i\in s} y_i^c/\pi_i}{c(c-1)\left(\sum_{i\in s} y_i/\pi_i\right)^c} - \dfrac{1}{c(c-1)}$	

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. The two variance estimators proposed for the Gini family are both obtained by analogy with the approximate variance of a ratio estimator, based on a first order Taylor approximation. The first estimator (method i) will be called a "Taylor estimator" and the second (method ii) a "Ratio estimator." The difference lies in the fact that, for a fixed value on y_i, the J-function depends on the sample s: in the Taylor estimator we take account for this stochasticity, but not in the Ratio estimator where the J-function is considered as a constant (see Nygård and Sandström (1985a)).

The Taylor estimator of $V(\hat{R}_N)$ is given explicitly in Table 6. It has the disadvantage

Table 6. Variance estimator of the estimator of the Gini coefficient using the linear terms in the Taylor expansion (method i)

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

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

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

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

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

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

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

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

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

Table 6 (cont).

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

$$+\sum_{i\neq j\neq k\neq l\in s}\sum\sum\sum\frac{\pi_{ijkl}-\pi_{ij}\pi_{kl}}{\pi_{ijkl}}\cdot\frac{I_{[y_j<y_i]}}{\pi_j}\cdot\frac{I_{[y_l<y_k]}}{\pi_l}\cdot\frac{y_i}{\pi_i}\cdot\frac{y_k}{\pi_k},$$

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

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

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

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

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

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

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

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

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

of including up to the fourth order inclusion probabilities, because of the stochasticity of the J-function. The variance estimators of the estimated Mehran's and Piesch's parameters will include up to the sixth order inclusion probabilities! Simpler, but cruder, variance estimators can be obtained to all estimators belonging to the Gini family by use of the Ratio estimator. An estimator of $V(I_G(\hat{F}_N))$ is then

$$\hat{V}(I_G(\hat{F}_N))=\frac{1}{\hat{t}_y^2}\frac{1}{2}\sum_{i,}\sum_{j\in s}\frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}}\left(\frac{\hat{\hat{Z}}_i}{\pi_i}-\frac{\hat{\hat{Z}}_j}{\pi_j}\right)^2, \tag{4.2}$$

where $\hat{t}_y=\sum_{i\in s}y_i/\pi_i$ and $\hat{\hat{Z}}_i=(J(\hat{F}_N(y_i))-I_G(\hat{F}_N))y_i$. 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 case of the Generalized Entropy family the two methods are identical. The estimators are given in Table 7.

Explicit expressions of the two variance estimators for the Gini coefficient with an estimator based on the asymptotic variance (4.5b) are compared in the simple random sampling case in Nygård and Sandström (1985a).

Table 7. Variance estimators of the estimated members of the Generalized Entropy family using the linear terms in the Taylor expansion

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

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

$$\hat{V}(\hat{E}_{cN}) = \frac{\hat{N}^{2(c-1)}\hat{t}_u^2}{(c-1)^2\hat{t}^{2(c+1)}}\hat{V}(\hat{t}_y) + \frac{\hat{N}^{2(c-1)}}{c^2(c-1)^2\hat{t}_y^{2c}}\hat{V}(\hat{t}_u) + \frac{\hat{N}^{2(c-2)}\hat{t}_u^2}{c^2\hat{t}_y^{2c}}\hat{V}(\hat{N})$$
$c \neq 0,1$
$$- 2\frac{\hat{N}^{2(c-1)}\hat{t}_u}{c(c-1)^2\hat{t}_y^{2c+1}}\hat{\mathrm{Cov}}(\hat{t}_y, \hat{t}_u) - 2\frac{\hat{N}^{2c-3}\hat{t}_u^2}{c(c-1)\hat{t}_y^{2c+1}}\hat{\mathrm{Cov}}(\hat{t}_y, \hat{N}) +$$
$$+ 2\frac{\hat{N}^{2c-3}\hat{t}_u}{c^2(c-1)\hat{t}_y^{2c}}\hat{\mathrm{Cov}}(\hat{t}_u, \hat{N})$$

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

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

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

and

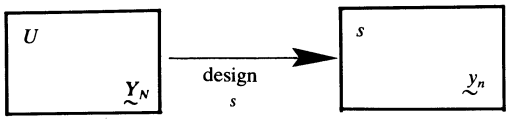
$$\hat{\mathrm{Cov}}(\hat{t}_y, \hat{N}) = \frac{1}{2}\sum_{i,j\in s}\sum \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)\left(\frac{1}{\pi_i} - \frac{1}{\pi_j}\right)$$

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 chosen to the sample are 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 (i.i.d.) as Y with continuous cumulative *df* $F_Y(y)$. The two approaches are illustrated by Fig. 1.

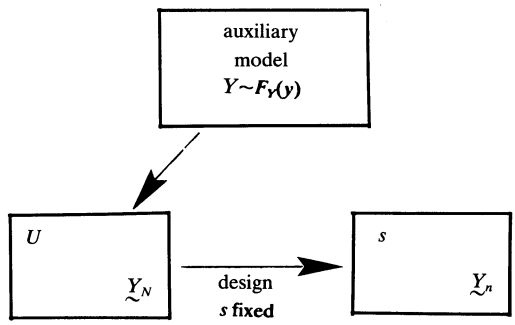
Fig. 1. Illustration of the fix population approach (a) and the auxiliary model approach (b)

(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 $\underset{\sim}{Y}_N$

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. Royall (1971) states that for a given sample s, the probability of coverage is the same as the probability that the interval includes the random variate $T(F_N)$ when the generation of Y-values from the model is "repeated." The obtained asymptotic results can also be used as the basis for large sample inference in the fix population approach. Now consider 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 $p_t = n_t/N_t \to p$, $0 < p < 1$, 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 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 concerns the corresponding estimator 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

$$F_{n_t}^*(y) = n_t^{-1} \sum_{i \in s_t} \frac{w_{it}}{\bar{w}_t} I_{\{Y_i \leq y\}}, \tag{4.3}$$

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

REMARK 4.4. If the weights are equal to some positive constant, then $F_{n_t}^*(y)$ coincide with the 'ordinary' empirical df and if $w_{it} = \pi_{it}^{-1}$, where π_{it} denotes known inclusion proba-

bilities, then (4.3) is similar to (4.1). The only difference is that in (4.3) s_t is fixed and Y_i is stochastic with the reversed relation in (4.1).

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

$$\max_{i \in s_t} (w_{it}/w_t)^2 \leq d^2 < \infty, \quad \forall\, t. \tag{4.4}$$

REMARK 4.5. When the weights equal some positive constant then (4.4) 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 $p_t = n_t / N_t$ which is assumed to converge towards a constant p, $0 < p < 1$, so the assumption mainly states that the design may not be such that $\min_{i \in s_t} \pi_{it} \to 0$ as $t \to \infty$.

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

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

Sandström (1983) shows that, if J is bounded and continuous, $\sigma_G^2 > 0$ and under Assumption 4.1, that

$$\frac{n_t^{1/2} \{I_G(F_{n_t}^*) - I_G(F_{N_t})\}}{\{1 - p_t + v_t^2\}^{1/2}} \xrightarrow{\mathcal{L}} U \sim N(0, \mu^{-2} \sigma_G^2), \tag{4.5a}$$

where

$$\sigma_G^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{\min(F(y), F(x)) - F(y)F(x)\} J_1(F(y)) J_1(F(x)) dy dx, \tag{4.5b}$$

and $J_1(F) = J(F) - I_G(F)$. In Nygård and Sandström (1985a) a similar result is shown to hold for the members of the Generalized Entropy family, viz

THEOREM 4.1. 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 $F \in \mathbf{F}$, $\mathbf{F} = \{F; |I_{E,c}(F)| < \infty\}$. 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 $0 < \sigma_c^2 < \infty$,

$$\frac{n_t^{1/2} \{I_{E,c}(F_{n_t}^*) - I_{E,c}(F_{N_t})\}}{\{1 - p_t + v_t^2\}^{1/2}} \xrightarrow{\mathcal{L}} U \sim N(0, \sigma_c^2), \tag{4.6}$$

where σ_c^2 equals

$$c=0: \sigma_0^2 = V(\log Y) + \frac{1}{\mu^2} V(Y) - 2\frac{1}{\mu} \mathrm{Cov}(\log Y, Y), \tag{4.7a}$$

where $\mu = E(Y)$.

$$c = 1: \sigma_1^2 = \frac{1}{\mu^2} V(Y \log Y) + \frac{(\mu_l + \mu)^2}{\mu^4} V(Y) - 2\frac{(\mu_l + \mu)}{\mu^3} \mathrm{Cov}(Y, Y \log Y), \tag{4.7b}$$

where $\mu_l = E(Y \log Y)$.

$$c \neq 0,1: \sigma_c^2 = \frac{1}{c^2 (c-1)^2 \mu^{2c}} V(Y^c) + \left(\frac{\mu_c'}{(c-1)\mu^{c+1}}\right)^2 V(Y) - 2\frac{\mu_c'}{c(c-1)^2 \mu^{2c+1}} \mathrm{Cov}(Y^c, Y), \tag{4.7c}$$

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

4.4. *Grouped Sample Data*

In Section 3.2, we discussed the calculation of the inequality parameters from grouped data. Estimates based on sample data are subject to sampling variation. For the Gini coefficient, R, let R_L be the lower bound given in Table 4 and $R_U = R_L$ + bias be the upper bound with the bias term as in Table 4. In a recent paper, Gastwirth et al. (1984) gave the joint asymptotic distribution of $\hat{R}_L$ and $\hat{R}_U$ based on a srs design.

As a rule of thumb, Champernowne (cf. Cowell and Mehta (1982, p. 289)) proposed that $R_* = (2/3)R_L + (1/3)R_U$ would yield a good approximation to R. From the joint distribution of $\hat{R}_L$ and $\hat{R}_U$ the asymptotic distribution of $\hat{R}_*$ is also obtained in Gastwirth et al. (1984). In op. cit., the joint asymptotic distribution of estimated lower and upper bounds of the members of Gastwirth's (1975) class of inequality parameters is also obtained. This class includes, among others, Theil's and Atkinson's inequality parameters.

5. References

Atkinson, A. B. (1970): On the Measurement of Inequality. Journal of Economic Theory, 2, pp. 244–263.

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

Cowell, F. A. (1980): On the Structure of Additive Inequality Measures. Review of Economic Studies, 47, pp. 521–531.

Cowell, F. A. and Mehta, F. (1982): The Estimation and Interpolation of Inequality Measures. Review of Economic Studies, 49, pp. 273–290.

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

Gastwirth, J. L. (1975): The Estimation of a Family of Inequality Measures. Journal of Econometrics, 3, pp. 61–70.

Gastwirth, J. L. and Glauberman, M. (1976): The Interpolation of the Lorenz Curve and the Gini Index from Grouped Data. Econometrica, 44, pp. 479–483.

Gastwirth, J. L., Nayak, T. K., and Krieger, A. M. (1984): Large Sample Theory for the Bounds on the Gini and Related Indices of Inequality Estimated from Grouped Data. Research Report, Departments of Statistics, George Washington University and University of Pennsylvania.

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, pp. 1768–1773.

MacDonald, J. (1984): Some Generalized Functions for the Size Distribution of Income. Econometrica, 52, pp. 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, pp. 64–66.

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

Nygård, F. and Sandström, A. (1985 a): Estimating Gini and Entropy Inequality Parameters. Promemorior från P/STM, Nr 13, 85–01–09, Statistics Sweden.

Nygård, F. and Sandström, A. (1985 b): Income Inequality Measures Based on Sample Surveys. Invited paper, 45th Session of the International Statistical Institute, Amsterdam.

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, Dep. of Statistics, University of Stockholm.

Shorrocks, A. F. (1982): Inequality Decomposition by Factor Components. Econometrica, 50, pp. 193–211.

Received February 1985
Revised July 1985

Journal of Official Statistics
Vol. 1, No. 4, 1985, pp. 413–426
© Statistics Sweden

The Metainformation System: Its Structure and Role in the Statistical Information System

Anton Klas[1]

Abstract: The article begins with a history of metadata, a new form of data which has appeared as a result of the recent development of information systems. The concepts of data and metadata are analyzed. Data is defined as a logical assertion in which a particular attribute value is assigned to a given entity by means of an attribute. Metadata is also defined as a logical assertion; metadata, however, is not related to socioeconomic reality but to the properties of the information system and its individual components. Metadata forms the heart of the metainformation system: it is arranged into catalogues, and the catalogues are conveniently interlinked in the metainformation system. The concluding part of the article discusses the problems arising in updating the catalogues and the requirements put upon the computer implementation of the metainformation system.

Key words: Data; metadata; information system; metainformation system; entities; attributes; basic and complementary components of the name of an attribute; identifying, classifying and indicating functions of an attribute; catalogues; dictionaries; directories.

0. Summary

The article begins with a history of metadata, a new form of data which has appeared as a result of the recent development of information systems. In addition to the new requirements put upon information systems, the demand for the regularization of terminology, classification and data processing, the rise of metainformation systems was also influenced by increasing automation.

The next part analyzes the concepts of data and metadata. Data is defined as a logical assertion in which a particular attribute value is assigned to a given entity by means of an attribute. Three basic functions of attributes are distinguished: identifying, classifying and indicating functions. The names of attributes contain a basic and a complementary component. The structure of a data item is illustrated in Table 1 on page 416.

Metadata is also defined as a logical assertion; unlike data, however, metadata is related not to the socioeconomic reality but rather to the properties of the information system and its individual components. The structure of metadata is illustrated in Table 2 on page 418.

Metadata forms the heart of the metainformation system. The metadata is systematically arranged into catalogues within this system. Among the most important catalogues are the catalogue of indicators, catalogue of classifications, catalogue of special metadata, catalogue of registers, catalogue of algorithms,

[1] Institute of Socioeconomic Information and Automation in Management, VUSEI-AR, Dubravska 3, 842 21 Bratislava, Czechoslovakia.

directory and a data dictionary (thesaurus). Their contents are illustrated in Fig. 1 on page 420.

The headings of the main catalogues are also given in Fig. 2–5. The catalogues are conveniently interlinked in the metainformation system. The contents of catalogues, their number and their interlinkage depend on the particular requirements which the metainformation is to fill. Fig. 6 on page 423 illustrates schematically how the user's requirements for the information system are met by means of the catalogues of the metainformation system.

The final part of this article deals with problems in updating catalogues and with the requirements imposed on the computer implementation of the metainformation system.

1. Introduction

The concepts of metadata and metainformation system have been introduced in statistical literature only recently. Thus it will be useful to discuss briefly the origin and development of these concepts.

In the past, developments in the individual branches of statistics were quite independent. Often different names were used for the same data items or the same name was used for different data items. In addition to the confusion this nomenclature difficulty caused, new requirements arose for the management and development of statistical information systems.

Automation has contributed to the solution of these difficulties, but has brought about a number of other problems. The volume of data processed has increased, as have the demands on their storage and retrieval. Mastering the large data files and their processing has gone far beyond the powers of human memory. The volume of requirements from various users has substantially increased as well. The need for efficiency in automated processing led to separate storage and processing of many individual data components which, before automation, were processed as an integrated whole. This made the problems in the area of data processing, retrieval and interpretation even more complex.

All these developments made it necessary to create better processing conditions and to provide better documentation. Thus steps have been taken to regularize the data report structure and organization on storage media, introduce a uniform terminology, and to unify the data content and interpret it without ambiguity.

To make these modifications more effective, it was necessary to create documentation records. Documentation had existed before automation, but in many diverse and incompatible forms. After the transition to automated processing, these diversities markedly reduced the efficiency of statistical data processing. The extension and unification of documentation records gave rise to a comprehensive system of descriptions of the content of information systems and of modes of surveying, processing, storing and providing data. Such data, which was in fact data on data and on other components of information systems, has been termed metadata, and the system providing this type of data has been termed a metainformation system.

After a certain volume of documentation records was reached, relatively independent components began to be created in the form of catalogues, dictionaries, directories, etc. Further development also necessitated documentation of the processing methods, applicable software, survey modes, etc. The metainformation system has become an important tool for statisticians as well as for other users of information systems.

The development and use of metainformation systems have also brought about a differentiation of its various functions. The information function, i.e. the task of providing statisticians, designers, administrators and other users of information systems with an

outline of the contents of those information systems and of how they can be used, has naturally become the primary function of the metainformation system. Of the other functions, especially relatively recent ones, we mention the integrating function. The metainformation system gives a survey of one or more information systems, but it also creates an efficient means for their mutual cooperation. This function is of particular value in extensive nationwide information systems, the components of which are often kept in different forms by different institutions in different places. In centrally planned economies there is, moreover, a further requirement: to integrate the information system with other management tools, and especially with a national economic plan.

Since the necessity for building metainformation systems is a common problem faced by several statistical offices, this issue has also become the subject of research via the international project "Statistical Computing Project – SCP." The common efforts of workers from Czechoslovakia, Bulgaria, the Netherlands and Poland have resulted in a manual called "Users' Guide to Metainformation Systems in Statistical Offices."

2. The Concept of Statistical Data

If we inspect any completed statistical form or questionnaire, at first glance we can distinguish two kinds of designations in it, one exemplified by preprinted names and the other by records filled in by the reporting unit. On a closer look at the preprinted designations, we see that they consist of several kinds of names denoting various aspects of reality or having different meanings. In principle, there are two kinds of designations:

(1) designation of the class of objects for which the report is given, e.g. an enterprise, respondent, etc. The individual concrete objects will hereafter be called entities;

(2) designations of properties or relations of entities on which the reporting unit is to report. Such properties or relations represent the attributes of the given object.

The items to be filled in by the reporting unit in a questionnaire are either designations of the particular entity, as for example an enterprise named "Tesla", "Robotron", "Volvo" etc., or designations of values of individual attributes, e.g. the number of employees, "2 650."

By filling in the individual items of a form or a questionnaire we generate statistical data. From a logical point of view, an item of statistical data represents an assertion, expressed in a system of characters, setting the values of the relevant attributes to the given entity. From a grammatical point of view, a data item represents a sentence.

Symbolically, a data item can be expressed by the relation (1):

$$a_{ij} = A_j(e_i), \tag{1}$$

where e_i is the entity of reality ($i = 1,2,\ldots,k$), A_j is the attribute of the entity ($j = 1,2,\ldots,n$) and a_{ij} is the value of the attribute.

From a mathematical and logical point of view, the attribute represents a function which assigns the attribute values to the entities of a given class. This function may be defined in a different mode, e.g. by an instruction determining how to calculate the value of production, number of employees, labor productivity, etc. For example,

(1) electrotechnical industry = branch ("Robotron");

(2) 350 000 000 = fixed assets in crowns ("Tesla").

(Here the proper entity is distinguished from its name by quotation marks.) In the examples above the attributes are represented by the designations *branch* and *fixed assets*. The individual entities are represented by the designations of enterprises, as *Robotron* and *Tesla*.

The attribute values are represented by the designations *electrotechnical industry* and *350 000 000*. These examples can be simply expressed in the following way: "The Robotron enterprise belongs to the electrotechnical branch of industry" and "The Tesla enterprise has fixed assets of the value of 350 000 000 crowns."

In statistical practice some attributes often have different modifications; for example, labor productivity can be expressed in man-years, man-months, man-hours, on the basis of net production, in units, in kind, in value units, etc.

The attributes in the information system fill three basic functions:

(1) an identifying function; this arises from the fact that the identifying attribute values, which unambiguously distinguish each entity from all others, are assigned to the particular entities;

(2) a grouping function; this allows us to define a population of entities or various subpopulations of entities on the basis of the attribute values which satisfy certain conditions. For example, a population of enterprises is formed by organizations in which the attribute values correspond to the essentials defining the enterprise according to the valid legal regulations; or a subpopulation of employees can be defined having a salary greater than 3 000 crowns. When we consider the attribute values allowing us to classify all entities of the given population, we speak of a classifying function. For example, enterprises can be classified according to regions, ministries, branches, etc.;

(3) an indicating function; this contains important facts about the entities which are necessary for describing and managing socioeconomic reality. The data representing the indicating function are usually called indicators in statistics.

The names of attributes can consist of several components. In principle, they can be divided into:

a) a basic component denoting the primary meaning of the attribute; for example, labor productivity;

b) a complementary component which states more exactly the attribute meaning. It contains information on the units of measurement, the period to which the attribute value applies, whether this is an actual, planned or estimated value, etc. An example of data structure with its usual functions is given in Table 1.

Table 1. An Example of Data Structure

Entity in reality	Attribute of Entities						
	Identifying	Classifying			Indicating		
	A_1 (designation of entity)	A_2 (kind of entity)	A_3 (branch)	A_4 (region)	A_5 (production in million crowns)	A_6 (fixed assets in million crowns)	...
1st entity	"INDUSTRA"	state enterprise	Industry	West	9.5	90	...
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
*i*th entity	"AGRONA"	agriculture cooperative	Agriculture	East	7.2	81	...
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
*k*th entity	"TRANSITA"	transport cooperative	Transport	Central	4.7	62	...

The data fulfilling the identifying and classifying functions is usually kept in a special form called a register. This data is relatively constant. The data fulfilling the indicating function is usually stored separately. The two types of data are linked by means of code lists allowing the retrieval of indicating data according to different classifying aspects (e.g. production data for all enterprises in the branch of agriculture).

3. The Concept of Statistical Metadata

Returning to the statistical form or questionnaire, we see that the attributes dealt with are represented by a preprinted text. Until the form is filled in, it contains only the attributes without concrete values. The forms or enclosed directions contain instructions for correct completion of the preprinted text. These instructions indicate a uniform way of assigning the values of individual attributes (e.g. how to calculate the value of production, determine the number of employees, etc.) for the given entity (e.g. an enterprise). They also contain more detailed explanations of the meaning of the data required, how to calculate it, the period to which it is related, the population of entities for which the data should be completed, etc.

As can be seen, we are dealing here with a type of data different from that representing socioeconomic reality. In order to eliminate confusion, the first type of data is referred to as metadata and the second type, that representing socioeconomic reality, is called object data. The metadata, like the object data, are also assertions from a logical point of view. They are, however, not assertions about socioeconomic reality but are rather assertions about the information system, its elements and attributes. The following assertion can be referred to as metadata: "A data item about labor productivity in the information system is surveyed for the enterprises belonging to some branch of industry."

A metadata item can be expressed formally by the relation (2):

$$f_{rs} = F_s(A_r) \tag{2}$$

where A_r is the entity at the level of metadata ($r = 1,2,\ldots,n,\ldots,p$), F_s is the attribute of the entity at the level of metadata ($s = 1,2,\ldots,q$), and f_{rs} is the value of the attribute.

The entities at the level of metadata are represented by the designations of attributes of the entities of socioeconomic reality. We are not interested in the particular object data but in the designations of attributes which are common to a whole class of object data. A time series for production, for example, may contain many particular values, but they all concern the same attribute, i.e. production. The metadata item will then reflect the fact that the information system contains a time series of data for production.

The repertoire of entities at the level of metadata is not confined to the object data attributes only; it also involves attributes of other components of the information system, such as forms and other data media, data storage locations, information system software, modes of data surveying and processing, etc. This fact is expressed symbolically in relation (2) above.

As with data, in metadata, too, the attribute F_s represents a function assigning the attribute value to the relevant entity. These functions are usually defined in various ways, e.g. by methodological instructions, rules, etc. The attribute values f_{rs} are represented, for example, by a list of all branches, regions, ministries, etc., by the formulas for calculating the individual object data items, by the data on the survey frequency, by the definitions of designations of the object data, etc.

Examples:

Names of all branches	=	List	(Branch)
Formulas for calculating the indicators in labor productivity	=	Methodological rule of calculation	(Labor productivity)
Monthly	=	Methodological rule on the frequency of survey	(Value of production)
The individual definition of the concept of "fixed assets"	=	Methodological rule defining the content of indicator	(Fixed assets)

The examples above can be simply expressed in this way: "The data item on a branch comprises these branches: agriculture, building industry, transport industry, etc." "The data item on labor productivity is calculated according to the formulas adduced in the methodological rule." "The data item on the value of production is surveyed monthly." "The data item on fixed assets has the meaning specified in the pertinent methodological instruction."

The examples above also indicate that in relation (2) the entities at the level of metadata are represented by the designations which played the role of attributes in relation (1). Thus, if $r = 1,2,\ldots,n$, the entities in relation (2) are identical with the attributes in relation (1).

The metadata and its structure are illustrated in Table 2. It can be seen that the first column of Table 2 is identical with the heading of Table 1, with the exception of the identifying attribute A_1.

4. The Metainformation System and Its Functions

The subject of the metainformation system is the information system. The main objective of

Table 2. The Metadata and Its Structure

Entity in metainformation system	Attribute				
	F_1 (list of values)	F_2 (frequency of survey)	... ...	F_{s-1} (algorithm)	F_s (place of storage)
Type of entity A_2	list of entity types	if a change occurs	...	–	Designation Data file
Branch A_3	list of branches	if a change occurs	...	–	Designation Data file
Region A_4	list of regions	if a change occurs	...	–	Designation Data file
Production A_5	–	monthly	...	mode of calculation	Designation Data file
Fixed assets A_6	–	biannually	...	mode of calculation	Designation Data file
. . .	. . .	. . .	. . .	. . .	. . .

the metainformation system is to create the conditions for:

- the efficient performance of the functions of the information system;
- the cooperation of several information systems;
- the design and economy of the information system;
- the management and improvement of the information system.

The metadata is the means of representing the elements and properties of the information system. It is classified and stored by the metainformation system in special forms which are called catalogues. The catalogue is a systematic ordering of metadata according to its type and/or according to the function it fulfills in the efficient running of the information system. The main elements of the metadata given in the catalogue are:

- name of entity (name of indicator, form, program, etc.);
- attributes of entity (frequency of survey, mode of calculation, place of storage);
- codes for linking with other catalogues or with data and other components of the information system.

Summarizing the above, we can say that the metainformation system:

(1) describes the individual components of the information system as well as the links among them via metadata;
(2) unifies the mode of that description within the given information system;
(3) unifies the description of several information systems needed to respond efficiently to users' demands;
(4) creates a uniform dictionary of the information language and its interpretation;
(5) systematizes the metadata and stores it in properly structured catalogues;
(6) creates a means for interlinking the catalogues into a system;
(7) creates a means for linking the system of catalogues with the corresponding components of the information system, particularly the statistical data files and registers;
(8) creates a means for analyzing and improving the information system.

5. Catalogues – The Basic Form of Metadata Holding

If we disregard the other components of the information system and focus our attention only on the data component (analyzed in more detail above), the basic system of catalogues for the data component can be derived from Tables 1 and 2. By their suitable combination and adaptation we obtain Fig. 1 on page 420, illustrating the structure of the contents of the metainformation system as well as the corresponding catalogues.

As illustrated in Fig. 1, among the most important catalogues related to the data component are:

(1) catalogues containing designations of the individual attributes of entities of socioeconomic reality; this group includes also the catalogues of indicators and the catalogues of classifications;
(2) catalogues of special metadata containing the surveying characteristics for the object data; for example, the catalogue of forms;
(3) catalogues containing the description of mode of calculation; these include, e.g., the catalogue of algorithms;
(4) catalogues containing data on the types of entities of socioeconomic reality for which the data is surveyed; for example, the catalogue of registers;
(5) a catalogue containing the interpretation of all designations; this catalogue is called the data dictionary. Such a catalogue, in conjunction with a permuted index of elements of names, also acts as a thesaurus;

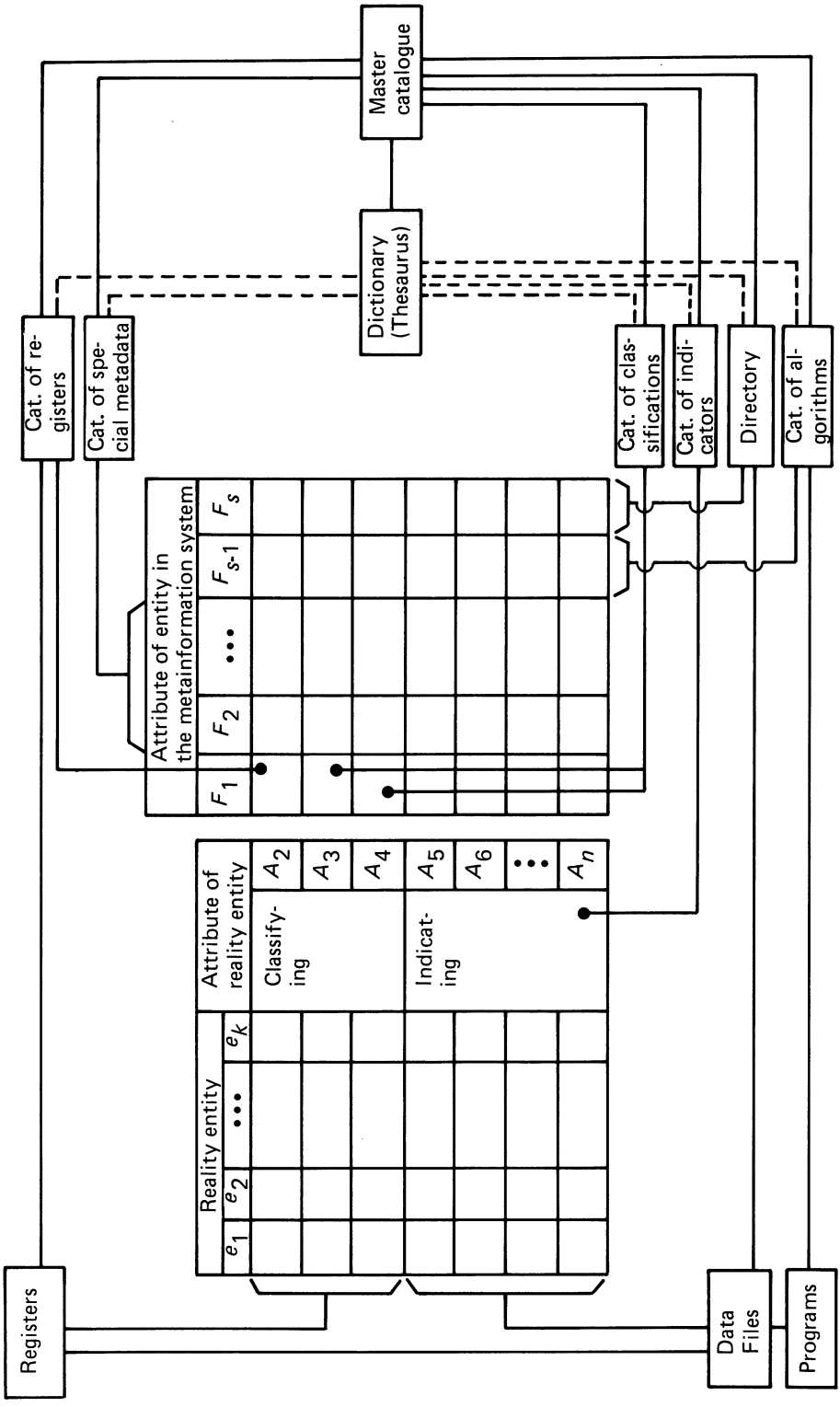


Fig. 1. Important Catalogues Related to the Data Component

(6) a catalogue containing data on storage location and on data access; this is called the data directory;
(7) a catalogue containing a survey of all catalogues, their content and structure; this is called the master catalogue.

The catalogues may under certain conditions have different numbers of subcatalogues. For instance, a catalogue of special metadata can be divided into the catalogue of forms, catalogue of reporting units, catalogue of survey units (the units which are the subject of the survey), etc. The catalogue of indicators can also be divided further into the catalogue of time series and the catalogue of the internal structure of indicators (e.g. according to the types of goods, products, qualifications of employees, etc.). In illustration we adduce in the following figures the headings of some typical catalogues: Fig. 2. Catalogue of Indicators, Fig. 3. Catalogue of Classifications, Fig. 4. Catalogue of Forms, and Fig. 5. Master Catalogue.

Identifying code of indicator	Classifier of indicator	Period of survey indicator	Measurement unit	Plan or Reality	Time characteristics of indicator	Type of evaluation
560106	*FAA001*	*annual*	*crowns*	*reality*	*by the end of the year*	*purchasing price*
continued						
Identifying code of the form	Branch statistics containing the form	Insert of the form	Division of the form	Sequence number of indicator of the form	Key words of indicator	Full name of indicator
521	*fixed assets*	*1*	*1*	*7*	*fixed assets value*	*value of the fixed assets*

Fig. 2. Catalogue of Indicators

Identifying code of classification	Name of classification	Valid since	Code of the classfication item	Name of the classification item
12	*classification of regions and districts*	*1978*	*5200*	*West Slovakian region*

Fig. 3. Catalogue of Classifications

Identifying code of the form	Branch statistics containing the form	Period of the form	Number of type indicators on the form	Number of indicator values on the form	Code of methodological instruction	Full name of the form
521	*fixed assets*	*annual*	*137*	*386*	*24 ZP*	*Annual form on the fixed assets*

Fig. 4. Catalogue of Forms

Name of catalogue	Identification of user	Authorized access	Primary key	Secondary key	Logical structure of the catalogue record
catalogue of indicators	*name of user*	*degree of authorized access to data*	*code of indicator*	*code of the form*	*prerequisites of the catalogue of indicators specified in its heading*

continued (Implementation characteristics of the given catalogue fulfilling the function of a directory)

Medium	Number of medium	Organization of data storage	Maximum number of blocks	Actual number of blocks	Size of blocks (characters)	Percentage of blocks filled	Type of code	Length of record (characters)
disc	*75*	*indexed sequential*	*520*	*480*	*1693*	*100*	*binary*	*54*

Fig. 5. Master Catalogue

It is clear that one catalogue is not sufficient to fulfill the functions of a metainformation system: more are necessary. Consequently, they must be built so that they can be interlinked, and a system must be created which allows their appropriate cooperation. Thus the specification of the contents of catalogues and their number is of the utmost importance. This determines the possibilities and modes of interlinking the catalogues, as well as what type of questions can be answered in the future. The system of mutually interlinked catalogues forms the structure of the metainformation system.

To build such a system, the metainformation system uses names chosen according to approved nomenclature, various types of internal code lists serving to identify metadata, its files, and its mutual links as well as links with other components of the information system, particularly the data component.

If the catalogues have the proper contents and are adequately interlinked, they can provide the user with important information on the content of the information system and can directly satisfy his or her requirements if they are directly connected with the object data. Meeting any user's request presumes:

(1) correct interpretation of the request, i.e. that the data requested by the user is designated correctly and has the required meaning;
(2) proper assignment of the appropriate catalogues and relations among them.

The requests are interpreted with the help of a data dictionary (thesaurus). With this, too, one can examine whether the user's request is consonant with the designation of the data he or she is asking for. The function of assigning the particular catalogues to the user's requests is carried out by a master catalogue. Its function can be formally expressed by the relation (3):

appropriate catalogue = master catalogue (specification of the user's requirements) (3)

6. Handling of User Requests

To illustrate how a user's request is handled we give an example which at the same time will show how the catalogues are interlinked. Let us suppose that the user asks for the following data item:

"Time series of the value of fixed assets at the end of the year in purchase price

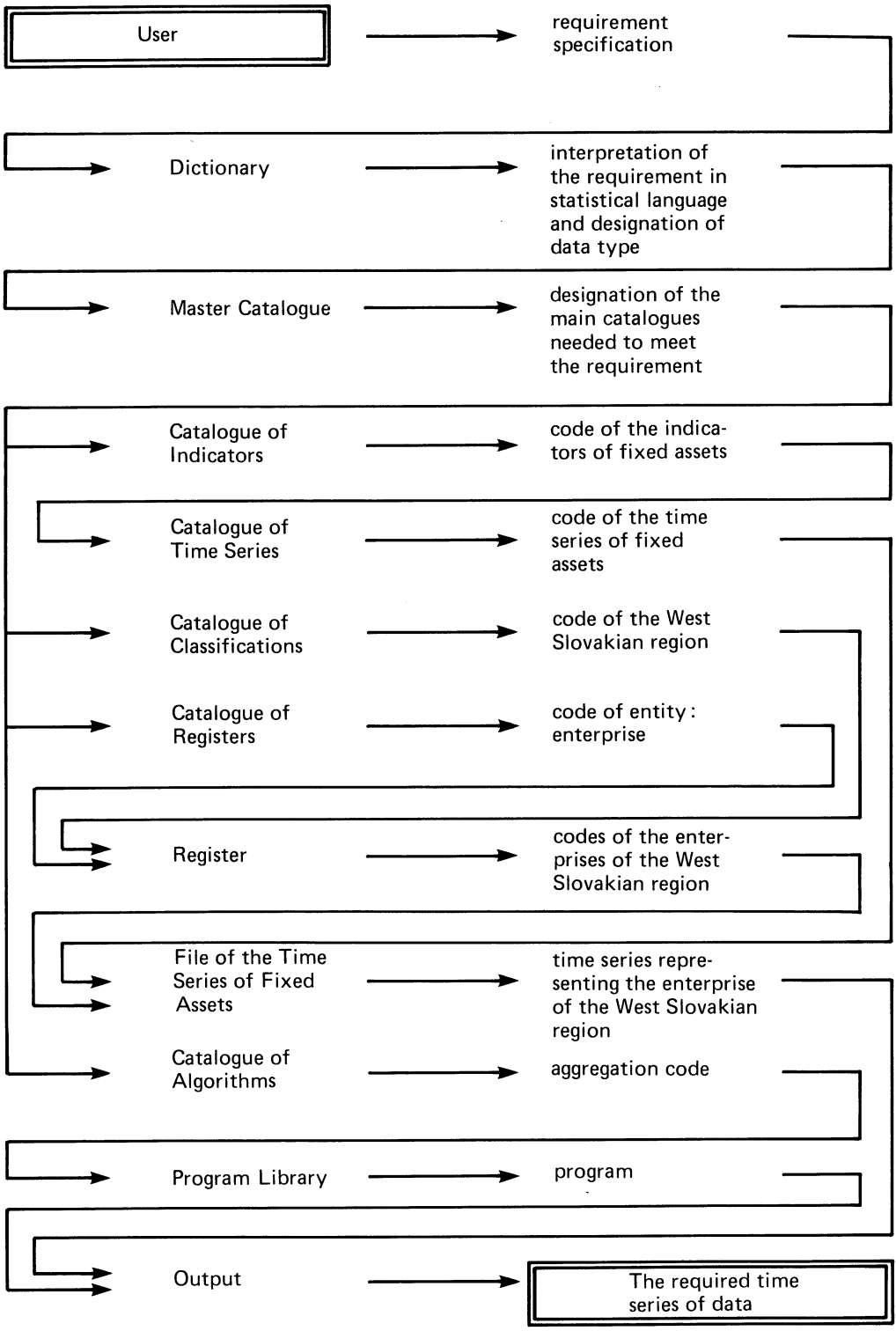


Fig. 6 Handling of User's Request

expressed in crowns for the years 1950–1980, surveyed for an enterprise and aggregated for the West Slovakian region."

1st step: Determination of key words and examination of their designation and meaning in the data dictionary.

The request as formulated contains the following key words which should be examined:

- time series,
- fixed assets,
- at the end of the year,
- purchase price,
- West Slovakian region,
- aggregation,
- enterprise.

After examining the designations and meaning of the above key words, the data dictionary provides information concerning the type of data to be obtained. Here we are dealing with the following types of data: indicator, classifier, entity (enterprise) and algorithm.

2nd step: The assignment of the appropriate catalogues by means of the master catalogue.

The following catalogues correspond to the types of data in our case:

- catalogue of indicators,
- catalogue of classifications,
- catalogue of registers,
- catalogue of algorithms.

3rd step: By searching in the appropriate catalogues for the corresponding designations stated in the user's request, we either obtain the codes providing information on the existence of data and its location, or we obtain directly the data item needed (if the catalogues are linked with the statistical data files). The whole procedure is illustrated in Fig. 6 on the preceding page. For clarity we adduce only those items from each catalogue which are directly connected with this example.

As indicated by the above examples, the metadata contains the facts on the contents and tools of the information system, thus creating a background for its more efficient utilization. By its ability to create such a picture on not only one but several information systems, the metadata system allows the use of a substantially larger group of data and tools than could be provided by any one information system. In addition, the metadata contains some information also valuable for the management and improvement of the information system.

The data reflects the state of socioeconomic reality; the metadata captures the designations by which our requirements for data and its processing are formulated. This difference in the function of data and metadata is reflected also in the function of the metainformation system. While the main function of the information system is to process the data, from its capture to its final form of presentation, the fundamental function of the metainformation system is to process user requests, from their initial interpretation until an answer is obtained. To fulfill this function, the metainformation system creates not only the foundations of a statistical language but also the principles by which the users' requirements are met in the information system.

7. Updating the Catalogues of a Metainformation System

Updating the metadata catalogues is a pressing current problem, and how it is solved will, to a considerable degree, affect the future of the metainformation system. Several projects have already failed in this respect, that is as the result of unsolved problems in the area of updating.

For the successful solution of the updating problem it is necessary:

(1) to entrust the preparation of supporting materials concerning updating to those units which are responsible for surveying the corresponding data and which must already inform the administrator of the metadata catalogue of all changes;

(2) to identify the catalogues which play a key role in updating and are linked to all the catalogues affected by updating; the catalogue of forms, for instance, can play this key role;
(3) to create software suitable for the automatic preparation of changes from the key catalogue to all other catalogues, including automated checks and corrections.

Experience in this sphere indicates that such a solution is possible and will create favorable conditions for ensuring the economic operation of the metainformation system.

8. Some Experience in Metainformation System Implementation in Czechoslovakia

The requirements for the implementation of the metainformation system do not differ greatly from those placed on the implementation of any other information system. As the manner of implementation is highly dependent on the hardware and software environment, we shall describe some experience gained in Czechoslovakia in the implementation of the metainformation system on the EC 1055 computer using the SOFIS programming system.

When implementing the catalogue system on the computer, two characteristic modes of operation with data were used: batch mode and interactive mode. If we are dealing with a user without knowledge of the metainformation system, the first mode of operation is used. When we are dealing with a professional user who understands the structure and implementation of the metainformation system, he or she can direct the computer operation by questions and suggestions and can help obtain the answer more rapidly. The SOFIS DBMS programming system is used to keep, use, maintain, operate and protect the metadata base, and in the interactive mode of operation the SOFIS DIAGEN programming system is used. Operation in batch mode is ensured by tailormade software, using COBOL and FORTRAN.

9. Bibliography

Dörnyei, J. (1983): The Role of Metainformation in Statistical Integration: A Subjective Essay. Invited paper, 44th ISI Session, Madrid.

Kent, W. (1978): Data and Reality. North Holland.

Klas, A. (1978): Structuring of the Computerized Information Systems from the Aspect of Integration. Seminar ISIS, CES (SEM. 10) 22, Bratislava.

Klas, A. (1981): Time Series Oriented Data Bases in Czechoslovak Statistics. Invited paper, 43rd ISI Session, Buenos Aires.

Kühn, J. (1983): Daten und Systemdokumentation in Informationssystem des Statistischen Bundesamtes für die BRD. Invited paper, 44th ISI Session, Madrid.

Senko, M. E. (1975): Conceptual Schemes, Abstract Data Structures, Enterprise Statistics. International Computing Symposium, North Holland.

Soltés, D. (1984): The Metainformation System and the Conceptual Level of Statistical Data Modelling. Statistical Journal of the United Nations, 2, pp. 97–108.

Statistical Computing Project (1981): Analysis and Evaluation. Present State of the METIS in Participating Countries. SCP (MI)1.

Statistical Computing Project (1983): Description of Results Achieved at the Development and Utilization of Metainformation Systems (MS) in Statistics. UNDP, Economic Commission for Europe, SCP (MI) WP. 51, Bratislava, July 15.

Statistical Computing Project (1984): Users Guide to Metainformation Systems in Statistical Offices. SCP(MI) WP. 106, draft

prepared by the Joint Group of the Meta-information System, Bratislava.

Sundgren, B. (1980): Meta-Information in Statistical Agencies. Statistics Sweden, Stockholm.

Yasin, E. G. (1974): Ekonomicheskaya Informaciya. Statistika, Moscow.

Received June 1985
Revised October 1985